C

C_{60}

► Fullerenes for Drug Delivery

Cancer

► Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites

Cancer Modeling

► Models for Tumor Growth

Capacitive MEMS Switches

Dimitrios Peroulis
School of Electrical and Computer Engineering, Birck Nanotechnology Center, Purdue University, West Lafayette, IN, USA

Synonyms

Electrostatic RF MEMS switches; Micromechanical switches; RF MEMS switches

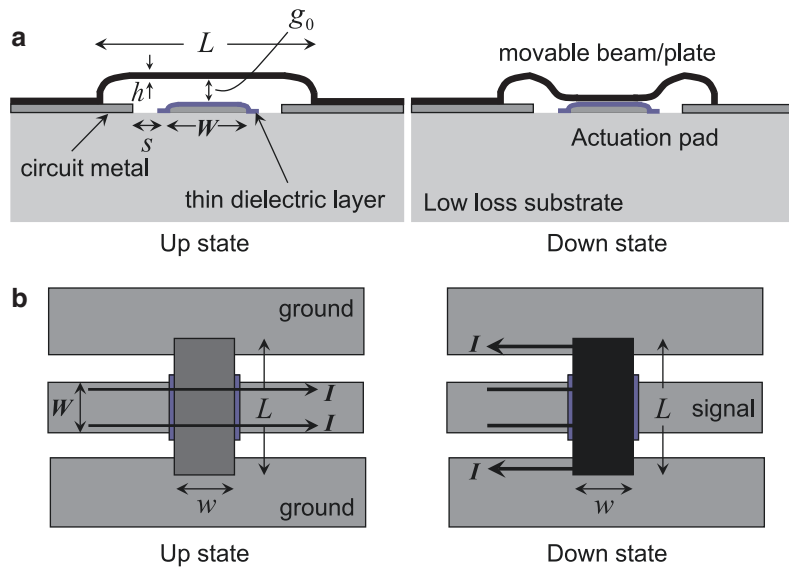
Definition

Capacitive micro-electro-mechanical systems (MEMS) switches are a special type of micromachined switches that control radio frequency (RF) signal paths in microwave and millimeter-wave circuits through mechanical motion and contact.

Overview

Capacitive and direct current (dc)-contact MEMS switches are among the most important micromachined devices for high-frequency applications due to their near-ideal RF performance. Dc-contact switches function similarly to conventional relays: micromachined beams or plates move under the influence of an appropriately applied force (e.g., electrostatic force) to open or close a metal-to-metal contact. While micromachined beams or plates are also utilized in capacitive switches, these switches rely on metal-to-dielectric contacts to implement their on and off states. Capacitive switches are particularly attractive for demanding high-frequency communications, electronic warfare, and radar systems due to their ultralow loss (<0.1–0.2 dB up to 40 GHz), high isolation (>20–50 dB for frequencies beyond 10 GHz), very high linearity (>66 dBm third-order intercept point), and near-zero power consumption (∼tens of nJ per

B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1



Capacitive MEMS Switches, Fig. 1 (**a**) Side-view and (**b**) top-view schematics of a typical shunt capacitive MEMS switch. Both the *up* and *down* states are shown

switching cycle and zero quiescent power for electrostatically actuated switches). When compared to solid-state switches, capacitive switches are relatively slow devices with speeds ranging in the tens to hundreds of microseconds range. This speed is primarily limited by switch inertia and squeeze film damping. Their relatively large lateral dimensions of tens or hundreds of μm allow capacitive switches to handle several 100 mW of RF power. Long-term operation, however, can only be achieved if they are hermetically sealed in order to avoid contamination- and humidity-induced failure. Hermetically sealed capacitive switches have successfully switched over 100 billion cycles at room temperature and under low RF power conditions (20 dBm). Despite the aforementioned RF advantages, capacitive switches are currently not available commercially and are not widely utilized in defense or communication systems. This is primarily due to the facts that (1) high-yield manufacturing processes are not widely available yet and (2) their main failure modes such as dielectric charging, dc/RF gas discharge and metal creep and the physics behind them have not been adequately understood and addressed today.

Switch Structure and Actuation Mechanisms

Figure 1 shows a typical capacitive MEMS switch [1]. This is a shunt switch configuration, and is the dominant capacitive switch configuration in the literature today. The signal travels down the center conductor, and if the switch closes, will return along the outside conductors. It is, however, possible to design geometries for series configurations. Their characteristics, nevertheless, are very similar to the ones found in shunt switches. Consequently, this entry focuses primarily on shunt capacitive switches.

The switch in Fig. 1 consists of a movable beam or plate that is anchored to the switch substrate. While typically the term "plate" may better characterize the geometry of Fig. 1 for typical lateral switch dimensions, it is common in the RF MEMS literature to refer to this geometry as a fixed-fixed beam. Hence the term "beam" is adopted here to describe such capacitive switch geometries. Cantilever beams are also possible particularly for series switch configurations. The fixed-fixed beam anchors are typically metallic and are connected to the RF line. For example, they can be connected to the ground planes of a

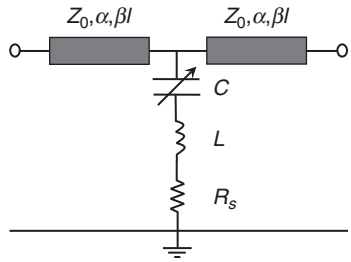
coplanar waveguide line as shown in Fig. 1b. The movable beam is typically composed of a thin-film (thickness h of 0.5–2 μm) metal such as gold, aluminum, nickel, or molybdenum. It is also possible that the beam is comprised of multiple layers including thin-film dielectrics such as silicon nitride or silicon dioxide and metals. One or more metallic pads are placed underneath the beam. In the simplest case, a single metallic pad is placed under the beam as shown in Fig. 1b. In this case, this pad is the center conductor of the coplanar waveguide line. This pad is usually covered by a thin (0.1–0.3 μm) dielectric layer such as silicon nitride, silicon dioxide, or more recently amorphous [2] or ultrananocrystalline [3] diamond.

The beam's length and width are determined by the required down-state switch capacitance as explained in the following section. This results in length and width in the tens to hundreds of μm for 5–40 GHz capacitive switches. The thickness of the dielectric layer also impacts the down-state capacitance. While high capacitance is in general required, thicknesses lower than 0.1 μm are hard to achieve in practice due to the need to handle high electric fields (dc and/or RF) across this dielectric. The gap (g_0) between the bottom surface of the movable beam and the top surface of the dielectric layer is determined by the need to minimize the up-state capacitance. A low up-state capacitance is necessary for low insertion loss. Capacitive switches in the 5–40 GHz range typically have gaps in the 2–5 μm range.

The switch has two states of operation. In the up state, the RF signal goes through the signal line almost unaffected by the movable beam. The up state is also called "zero biased" or "on state". In the down state the RF signal does not go through the RF line because it is reflected (see following section). The down state is also called the "biased" or "off state." An actuation mechanism is required to move the switch beam between these two states. Several possible actuation schemes exist, including electrostatic, electrothermal, magnetostatic, and piezoelectric. Table 1 summarizes the main advantages and drawbacks of each actuation scheme. The vast majority of reported capacitive switches are electrostatically actuated. They use the same actuation principle as the one originally proposed when the first capacitive switch was invented and reduced to practice in 1994 [4, 5]. In this original scheme a dc actuation voltage is applied between the movable beam and the actuation pad underneath it (Fig. 1). The beam is attracted due to the generated electrostatic field and collapses on the switch dielectric layer. Despite the need to generate a high dc voltage (30–100 V), which can be readily accomplished using a dc-dc converter, electrostatic switches exhibit the most desirable electromechanical characteristics, including the fastest possible response,

Capacitive MEMS Switches, Table 1 Main advantages and drawbacks of common actuation schemes for capacitive RF MEMS switches

Actuation mechanism	Advantages	Drawbacks
Electrostatic	Zero quiescent power consumption, easy biasing circuit, fast transient response (tens of microseconds)	Need to generate high voltage, high voltage may lead to charging and breakdown issues
Magnetostatic	Low voltage, high contact pressure, potentially low power with latching mechanism	Low quiescent power consumption requires latching, slower than electrostatic due to increased switch size
Electrothermal	Low voltage, high contact pressure, size comparable to electrostatic schemes	High quiescent power consumption (mW), slow response time (tens to hundreds of milliseconds)
Piezoelectric	Same as electrostatic by with low voltage	Difficult to achieve high-quality piezoelectric layer and integrate it with RF circuit at low temperatures



Switch Impedance

$$Z_s = R_s + j\omega L + \frac{1}{j\omega C},\ C = \begin{cases} C_{UP} \\ C_{DOWN} \end{cases}$$

$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$

Typical Values		
	X-band	K-band
C_{UP}/C_{DOWN}	0.1/6 pF	0.04/3 pF
L	4–80 pH	6–50 pH
R_s	0.1–0.3 Ω	0.1–0.3 Ω

$$Z_s = \begin{cases} 1/j\omega C & for \quad f << f_0 \\ R_s & for \quad f = f_0 \\ j\omega L & for \quad f >> f_0 \end{cases}$$

Capacitive MEMS Switches, Fig. 2 Lumped-element equivalent circuit of the capacitive switch shown in Fig. 1. Typical values are provided for the lumped components of this circuit

zero quiescent power consumption, and the easiest possible biasing circuits. The resulting high fields though may lead to (gas and solid) dielectric charging and their associated reliability issues. More detailed discussion can be found in the last section.

RF Performance

Figure 2 shows a simple but physically meaningful and accurate lumped-element equivalent circuit of the switch shown in Fig. 1. It also includes typical values for all the equivalent circuit components [1]. The parameters α and β in this figure represent the attenuation constant and propagation constant of the transmission line respectively. The up- and down-state capacitance values are the most critical ones in this equivalent circuit. The up-state capacitance can often be accurately calculated by a typical quasi-static expression

$$C_{UP} = C_{PP} + C_{ff} = \frac{\varepsilon_0 A}{g_0 + \frac{t_d}{\varepsilon_r}} + C_{ff}$$

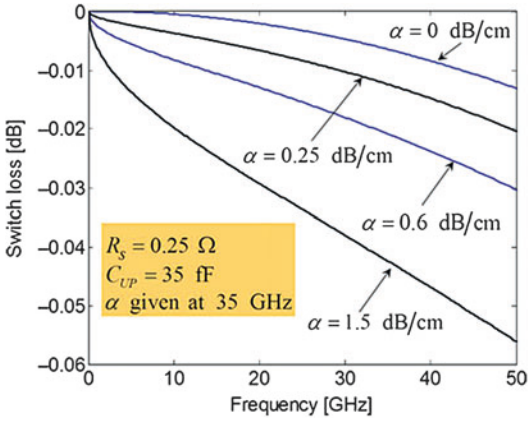
where C_{PP} and C_{ff} are the parallel-plate and fringing-field capacitances, respectively, A is the RF area of the switch ($A = Ww$ in Fig. 1), g_0 is the initial switch height, t_d is the dielectric layer thickness, and ε_r is this layer's dielectric constant. For typical switch geometries, the fringing-field capacitance could reach 25–50 % of the parallel plate capacitance. If improved accuracy is needed, a full-wave simulation is performed to estimate the switch up-state capacitance. The up-state capacitance must be sufficiently small to minimize the up-state insertion loss. Assuming a well-designed switch where the contributions from L and R_s can be ignored in the up state, the switch up-state reflection coefficient can be calculated from

$$S_{11} = \frac{-j\omega C_{UP} Z_0}{2 + j\omega C_{UP} Z_0}$$

where Z_0 is the characteristic impedance of the transmission line (typically 50 Ω). For example, an up-state capacitance of 70 fF results in $S_{11} < -10$ dB up to approximately 30 GHz. The up-state switch ohmic loss is the other critical up-state characteristic. The total ohmic loss of the switch can be calculated as

$$Loss = 1 - |S_{11}|^2 - |S_{21}|^2$$

This depends on (a) the attenuation α (dB/cm) of the transmission line underneath the movable beam and (b) on the switch series resistance R_s.



Capacitive MEMS Switches, Fig. 3 Simulated ohmic loss for a typical shunt capacitive switch. The attenuation *a* (dB/cm) depends on the transmission line characteristics

Well-designed capacitive switches can exhibit a total loss of less than 0.1 dB up to 40 GHz. Figure 3 shows numerical values for this loss for typical switch characteristics.

The switch down-state capacitance is more complicated to calculate because the switch beam may not be perfectly flat against the dielectric layer. Even in good designs this may not be possible due to the roughness of the layers involved. A model that is often used to capture the nonideal down-state switch capacitance is [1]

$$C_{DN} = \frac{\varepsilon_0 A}{2}\left(\frac{1}{r + \frac{t_d}{\varepsilon_r}} + \frac{\varepsilon_r}{t_d}\right)$$

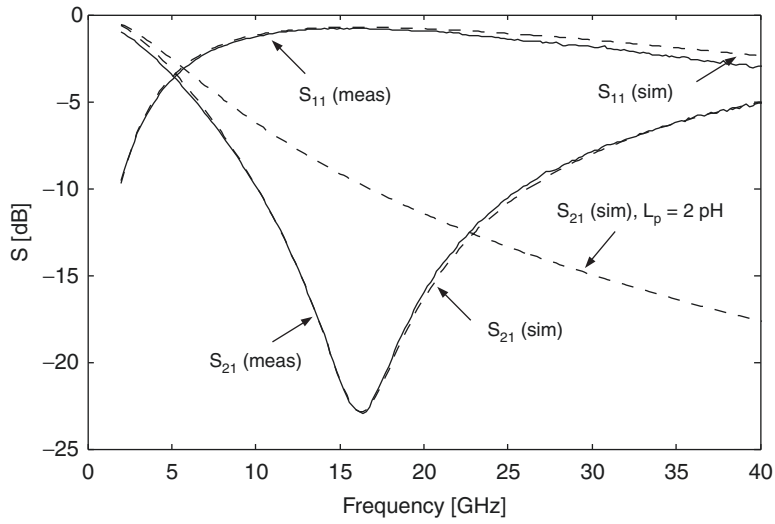
where r is the roughness amplitude. The down-state fringing-field capacitance is not included in this equation because it is typically not significant (<5 % of the parallel-plate capacitance) due to the small dielectric layer thickness. As shown by the equation above, the experimentally achieved down-state capacitance can vary greatly depending on the true contact area and the dielectric layer characteristics including its roughness. In practice, it is difficult to avoid a 30–50 % degradation of the down-state capacitance compared to the theoretical parallel-plate value. A high down-state capacitance (2–5 pF) is typically required in order to achieve an acceptable isolation (>20–50 dB) level at the desired frequency. One way to achieve this is to decrease the dielectric layer thickness. However, this dielectric layer needs to sustain very high electric fields (50–150 V/μm) across its thickness. Given typical fabrication process limitations that prohibit high-temperature growth processes for the dielectric layer (please see fabrication section), dielectric layers thinner than 0.1–0.2 μm become impractical. A second way to increase the down-state capacitance is to increase the dielectric constant. For instance, barium-strontium-titanate (BST) or strontium-titanate-oxide (STO) films with dielectric constants up to 400 can be employed. Besides additional fabrication complexities, these films have not been thoroughly studied in MEMS switches and may exhibit unacceptably high dielectric charging. Most switches typically employ some form of silicon dioxide or silicon nitride dielectric with dielectric constant in the 3.9–7.5 range. For switches with very low inductances (L < 1–2 pH), the switch isolation can be approximately calculated as

$$S_{21} = \frac{2}{2 + j\omega C_{DN} Z_0}$$

A more accurate calculation reveals that the switch inductance and series resistance also determine the total down-state isolation. In particular, its inductance cancels the down-state capacitance at the switch resonant frequency

$$f_0 = \frac{1}{2\pi\sqrt{LC_{DN}}}$$

This frequency can be adjusted by controlling the switch physical geometries. Switch inductances in the range of 1–100 pH can be readily achieved [1]. However, higher switch inductance typically results in a higher switch series resistance. Typical switch resistance values range in the 0.1–2 Ω range. The series resistance is the primary limiting factor of the switch isolation at its resonant frequency. At that frequency the switch isolation can be calculated as



Capacitive MEMS Switches, Fig. 4 Simulated and measured down-state scattering parameters of a capacitive MEMS switch (After Ref. [6] with permission)

$$S_{21} = \frac{2R_S}{2R_S + Z_0} \approx \frac{2R_S}{Z_0} \text{ at } f = f_0$$

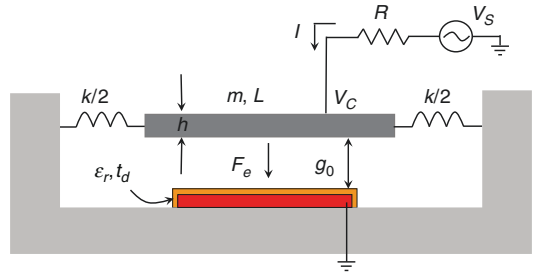
Figure 4 shows measured and simulated results of a typical capacitive switch [6] with $C_{DN} = 1.1\,\text{pF}$, $L = 87\,\text{pH}$, $R_s = 1.95\,\Omega$. This figure also shows the expected performance when the inductance is reduced to $L = 2\,\text{pH}$.

Electromechanical Considerations: Static Behavior

Figure 5 shows a simple but physically meaningful one-dimensional electromechanical model of the switch geometry of Fig. 1. The beam is modeled as a spring-mass system with a spring constant k. This spring constant depends on (a) the beam geometry, (b) the electrostatic force distribution on the beam, and (c) the residual stress of its structural film. The residual stress σ (MPa) is due to the fabrication process and depends on the exact deposition conditions. Typically a tensile stress (σ(MPa)) is needed in order to avoid buckling. The spring constant can be expressed as

$$k = k_1 + k_2$$

where k_1 depends on the first two factors and k_2 depends on the residual stress. The exact values



Capacitive MEMS Switches, Fig. 5 One-dimensional electromechanical model of a capacitive RF MEMS switch

can be calculated based on the specific switch design. For example, for the fixed-fixed beam of Fig. 1 and assuming that $W = L/3$ and that the electrostatic attractive force is uniformly distributed along the beam section directly above the coplanar waveguide center conductor, the spring constant can be calculated as [1]

$$\begin{aligned} k &= k_1 + k_2 \\ &= 32Ew\left(\tfrac{h}{L}\right)^3\left(\frac{27}{49}\right) + 8\sigma(1-\nu)w\left(\frac{h}{L}\right)\left(\frac{3}{5}\right) \end{aligned}$$

where ν is the beam's Poisson's ratio. For usual beam geometries and fabrication processes with residual stress in the order of 10–50 MPa, the second term dominates the spring constant. Typical spring constant values range from 10 to 50 N/m in order to provide sufficiently high restoring

force and avoid stiction issues. While low spring-constant designs have been successfully demonstrated [7], special care needs to be taken in avoiding stiction and self-actuation due to high RF power [8]. It is also important to mention that the above spring constant calculations are based on small-deflection theory. A nonlinear spring constant may need to be derived if this condition is not satisfied.

The electrostatic force F_e on the switch beam can be calculated as

$$F_e = \frac{\partial W_e}{\partial g} = \frac{1}{2} V^2 \frac{\partial C(g)}{\partial g} \approx \frac{1}{2} \frac{\varepsilon_0 A V^2}{g^2}$$

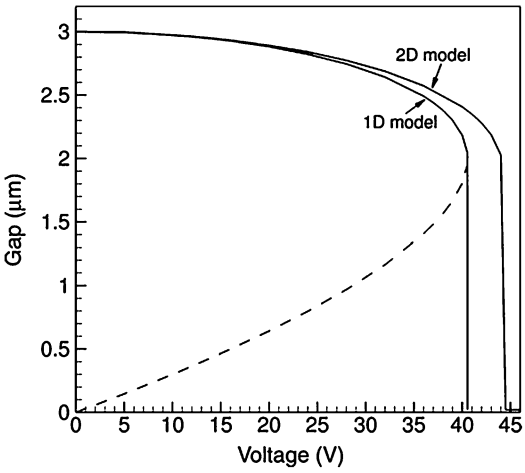
where W_e is the stored electrostatic energy, g is the switch gap between the beam and the actuation pad, and V is the applied electrostatic voltage. The last approximation is based on assuming a parallel-plate capacitance approximation and by ignoring the dielectric contribution. The static switch gap can be calculated by taking into account the static equilibrium of the forces applied on the beam

$$\frac{1}{2} \frac{\varepsilon_0 A V^2}{g^2} k(g_0 - g)$$

The above equation can be solved for the applied voltage as

$$V = \sqrt{\frac{2k}{\varepsilon_0 A} g^2 (g_0 - g)}$$

This equation is plotted in Fig. 6 for typical switch parameters. As Fig. 6 shows, there are two possible gaps for any given actuation voltage. This is not observed in practice and is a result of the unstable behavior of the beam. In particular, for small actuation voltages, the electrostatic force is increased proportionally to $\frac{1}{g^2}$. However, the restoring force is only increased proportionally to g. Hence, there is a critical gap beyond which the restoring force cannot hold the beam and the beam collapses on the dielectric surface. This critical gap can be found from the previous equation by taking its derivative and setting it up to zero



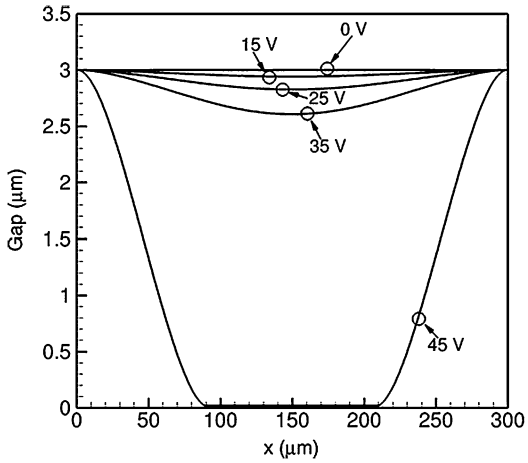
Capacitive MEMS Switches, Fig. 6 Simulated gap-voltage relationship for a capacitive MEMS switch with the following characteristics: $L = 300$ μm, $w = 100$ μm, $W = 120$ μm, $h = 1$ μm, $g_0 = 3$ μm, (Young's modulus) = 79 GPa, $\sigma = 10$ MPa. These results have been o btained with the PRISM center online simulation tool in MEMShub [13]

$$\frac{\partial V}{\partial g} = 0 \overset{g_c = \frac{2}{3} g_0}{\rightarrow}$$

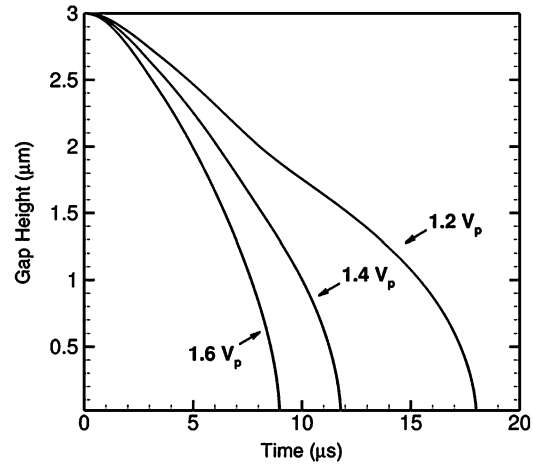
The voltage, therefore, required for actuating the beam, called the pull-in or pull-down voltage V_p, is given by

$$V_p = V(g_c) = \sqrt{\frac{8 k g_0^3}{27 \varepsilon_0 A}}$$

Figure 6 also shows the voltage-gap relationship as obtained by a two-dimensional beam model [13]. The gap plotted is between the center of the beam and the actuation pad. This curve is slightly different because the beam is not deformed as a perfectly flat object as assumed by the one-dimensional model. Figure 7 shows the actual deformed shape of the movable beam for voltages up to the pull-down voltage. Actuation voltages in the range of 30–100 V are typical in RF MEMS switches. Notice, however, that once the beam is actuated, the voltage required to hold the beam down (hold-down voltage V_h) is much lower because the gap between the beam and the actuation pad is much lower than g_0. It is hard to



Capacitive MEMS Switches, Fig. 7 Simulated shape of the movable beam of a capacitive MEMS switch for different bias voltages. The characteristics of the switch geometry are the same as in Fig. 6. These results have been obtained with the PRISM center online simulation tool in MEMShub [13]



Capacitive MEMS Switches, Fig. 8 Simulated switching time (pull-down) of a capacitive MEMS switch for different bias voltages. The characteristics of the switch geometry are the same as in Fig. 6. These results have been obtained with the PRISM center online simulation tool in MEMShub [13]

analytically calculate the hold-down voltage because it depends on many fabrication-dependent conditions such as the adhesion force between the beam and the dielectric layer. Typical hold-down voltages are in the range of 5–15 V [1]. Consequently, the gap-voltage relationship is strongly hysteretic. It is also worth mentioning that nonideal conditions such as an initial beam curvature and nonlinear bending are not included in this model. These effects can be captured by more complicated nonlinear beam models such as the ones presented in [9, 10].

Electromechanical Considerations: Dynamic Behavior

The dynamic behavior of the capacitive MEMS switch of Fig. 1 can be approximately captured by a one-dimensional model as described by the following equation

$$mg''(t) + bg(t) + kg(t) = F$$

where m is the switch mass, b is the damping coefficient, and f is the externally applied force. RF MEMS switches are typically packaged in an environment of 1 atm in order to avoid excessive ringing due to an underdamped response. As a result, the switch damping is dominated by squeeze-film damping as the gas under the beam is displaced during the switch motion. Due to the small gap g_0, it is in general difficult to accurately calculate the damping coefficient particularly in the near-contact region. This is further complicated by the possible existence of holes in the beam that aid its fabrication and substantially improve its switching speed. While there are several published approximations that can yield a reasonable approximation to the damping coefficient [1], accurate macro-models based primarily on rarefied gas dynamics only recently started becoming available [11, 12]. An equivalent way to characterize the switch damping is by its mechanical quality factor defined as

$$Q = \frac{k}{\omega_0 b}$$

where $\omega_0 = \sqrt{\frac{k}{m}}$ is the switch mechanical frequency. Typical switches show mechanical frequencies in the 20–100 kHz range and quality factors in the 0.5–2 range. Figures 8 (switch closure) and 9 (switch opening) illustrate dynamic

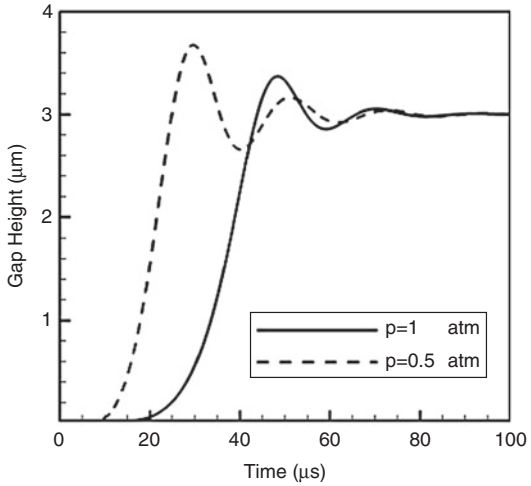
responses as calculated by two-dimensional beam models that accurately capture squeeze-film damping [13]. Notice that the displacement at the center of the beam is plotted in these graphs.

The switching speed can also be estimated based on the simple one-dimensional model. While it is difficult, in general, to derive an exact analytical solution, this model can provide reasonable approximations for common cases. For example, for $Q > 2$, the closing time can be estimated by [1]

$$t_s \approx 3.67 \frac{V_p}{V_s \omega_0}$$

where V_s is the applied voltage. Other limiting cases can be found in [1]. In general, closing times in the 5–50 μs range can be achieved. A similar range is typically possible for the release times.

The one-dimensional model can also be used to estimate the velocity and acceleration of the switch. Switching velocity in the 1–10 m/s range can be observed in the near-contact region. Due to its small mass, the switch acceleration can exceed 10^6 m/s^2 in the same region.

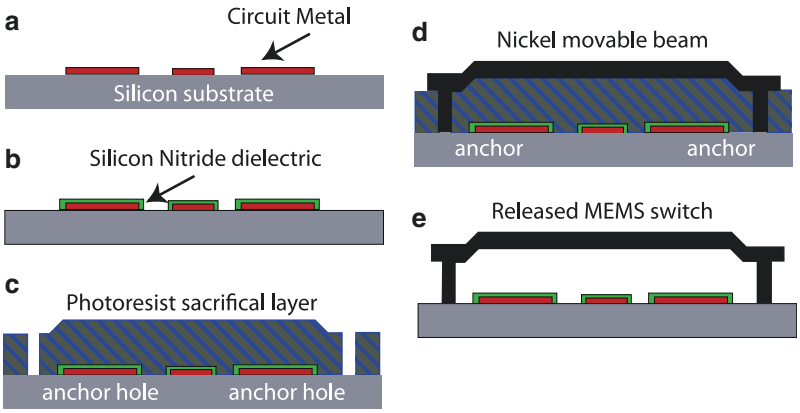


Capacitive MEMS Switches, Fig. 9 Simulated switching time (release) of a capacitive MEMS switch for different pressure levels. The characteristics of the switch geometry are the same as in Fig. 6. These results have been obtained with the PRISM center online simulation tool in MEMShub [13]

Fabrication Methods

Capacitive switches can be fabricated with conventional micromachining processes and require a small number of masks. Figure 10 illustrates the masks of a typical fabrication process.

- Step (a): The first mask defines the circuit metal of Fig. 1 after this metal layer is deposited on the substrate through evaporation, sputtering, or electroplating. Gold, aluminum, and copper are common metal choices. A smooth metal surface is particularly important directly underneath the switch beam in order to minimize local electric field enhancement.
- Step (b): The second mask defines the dielectric layer to cover the portion of the metal that



Capacitive MEMS Switches, Fig. 10 Simplified typical fabrication process for a capacitive MEMS switch

will be under the beam. Unless a high melting temperature metal has been deposited (e.g., tungsten), a relatively low-temperature process is required for the deposition and patterning of the dielectric layer in order to avoid damaging the circuit metal. Plasma-enhanced chemical vapor deposition (PECVD) is the most common process for depositing silicon nitride/oxide films. This is usually followed by a reactive ion etching (RIE) step that helps etching the unwanted dielectric layer parts.

- Step (c): The third mask is used to define the sacrificial layer of the switch. The sacrificial layer is the layer upon which the beam will be deposited, and this material needs to be removed at the end of the process to release the beam. The beam anchor points are defined in this step by selectively etching the sacrificial layer. This can be a dry- or a wet-etch step. More complicated processes may involve an additional planarization step before the next mask. The choice of the sacrificial layer material is critical as it controls many important parameters of the switch design, including the residual stress of the beam. Common material choices include photoresists, other photoconductive polymers, and polyimides. There is no sufficient understanding in the open literature of the exact processes that are involved in controlling the beam layer residual stress in the presence of a sacrificial layer. Important parameters though include, among others, the atomic structures of each film and the deposition temperatures of each film.
- Step (d): The fourth mask defines the actual beam layer. A variety of processes can be utilized including evaporation, sputtering, and electroplating of the beam layer(s). This step is also critical in determining the final residual stress of the beam.
- Step (e): The beam is finally released by etching (wet or dry) the sacrificial layer and by drying (if needed) the switch wafer. Etching of the sacrificial layer can also influence the beam residual stress particularly if a high-temperature process is necessary. If wafer drying is needed, this needs to be done carefully in order to avoid damage to the beam or causing stiction to the substrate. Stiction may occur if drying involves removing a liquid with high surface tension (e.g., water) underneath the beam. Such a liquid will pull the beam down as it evaporates. Special drying processes and equipment based on supercritical carbon dioxide have been successfully developed [14] and followed by many MEMS researchers.

The last step in the fabrication process is packaging, which is discussed in the following section.

Packaging

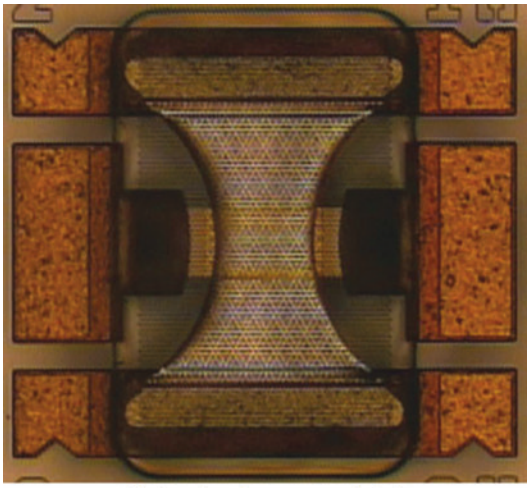
Hermetic packaging is required for capacitive RF MEMS switches to avoid any contamination- or humidity-induced early failure. While conventional hermetic packages exist, they are not well suited for capacitive RF MEMS switches or circuits. First, if switches need to be inserted in conventional hermetic packages, these switches will have to be diced first since several thousands of them can be simultaneously fabricated on a wafer. Dicing released switches is particularly dangerous for the switches and may considerably reduce the process yield. Second, conventional hermetic packages are expensive (∼tens of dollars/package) and not well-suited for cost-driven consumer applications. Third, they typically exhibit a relatively high insertion loss, which is often much higher than the switch itself (e.g., a DC-40 GHz package could exhibit 0.6 dB at 20 GHz [1]).

As a result, it is important to follow a cost-effective on-wafer hermetic packaging scheme. In this case, a wafer-scale package is first completed and then dicing follows. A wide variety of approaches have been developed so far to accomplish this. These approaches can be divided into three main categories:

- Two-wafer hermetic packages completed by fusion, glass-frit, thermocompression, eutectic, or anodic bonding. These packages are created by bonding two wafers together using one of the aforementioned approaches. The main advantage of these techniques is that they

result in excellent hermetic bonds. Their main drawback is that they may require high temperatures (300–1,000 °C depending on the technique) with the exception of low-temperature eutectic bonds (e.g., indium–gold bonds). In addition, these techniques tend to be relatively expensive since packaging cost usually accounts for 60–80 % of the total cost.

- Two-wafer quasi-hermetic packages completed by low-temperature polymer or solder-bump bonding. This technique is similar to the previous one, except that sealing is achieved by low-temperature bonding (room temperature −150 °C). A wide variety of polymers can be used for this. While low temperatures can be achieved, packages fabricated with such bonding typically exhibit very low but nonzero leak rates [1].
- Hermetic packages fabricated by on-wafer microencapsulation using micromachining techniques [15]. These techniques do not require a second wafer cap. Instead, every switch or switching circuit is encapsulated in a tiny package on its own wafer by a microfabricated technique. Typical temperatures in this process are in the 200–250 °C range. These techniques are ideally suited for the small size and high RF bandwidth of MEMS devices and typically result in low-cost fabrication. Particular attention needs to be paid though to ensure compatibility between the switch and package fabrication processes. Figure 11 shows a switch packaged with this technique.



Capacitive MEMS Switches, Fig. 11 The packaged memtronics switch (After Ref. [15] with permission)

Circuits and Applications

Capacitive MEMS switches may be employed in a number of circuits mostly for communication, radar, and electronic warfare systems [1]. Variable capacitors and impedance tuners for cell phones and other radios in mobile form factors constitute the most important applications in the commercial sector. MEMS variable capacitors (varactors) can be formed by connecting in parallel several capacitive MEMS switches and selectively activating them. Several companies including WiSpry and Cavendish Kinetics are pursuing this particularly due to the large cell phone market size.

The most important of the high-frequency circuit applications (>10 GHz) are high-isolation switching packets, true-time delay networks and phase shifters, reconfigurable impedance tuners for amplifiers and antennas, high-quality-factor reconfigurable filters, and tunable oscillators. Many of these circuits exploit the near-ideal RF performance of capacitive switches. Consequently, optimal performance can be usually achieved by employing a circuit- or sub-system-level package instead of a device-level package. Examples of several of these circuits can be found in [1].

Failure Mechanisms and Reliability

Capacitive MEMS switches suffer from high electric dc and/or RF fields through narrow gaps and dielectric layers. These fields can readily reach 5–50 V/μm in the up state and may increase further during actuation. Such fields may cause field emission and ionize the gas in the switch gap [16]. The long-term effects of field emission and

gas discharge are not known at this point. In addition, when such fields are applied across a thin-film solid dielectric, charges may get trapped in the dielectric layer leading to dielectric charging. A number of studies have been completed (see for example [17, 18]) focusing mostly on charges trapped in the bulk of the dielectric. However, surface charging of the solid dielectric that is also influenced by gas ionization is potentially more detrimental to the switch performance and is not well understood today. Charging phenomena are the leading cause of failure in capacitive switches today. Long-term drift of actuation voltage, stiction, and breakdown can be observed as a result of these charging issues.

Besides solid and gas dielectric charging, metal creep is another potential failure mechanism. Creep may be developed in the movable beam material if a switch is subjected to a constant stress. For example, if a switch is left in its down state for a long time (typically tens to thousands of hours), the beam material may creep resulting in a temporary or permanent change of the switch spring constant. Several recent papers show it is potentially an area of concern for capacitive MEMS switches [19, 20]. Creep at high temperature may be even a more significant area of concern.

Other possible failure modes include

- Beam buckling due to high temperatures. This may be caused during release process, normal high-temperature operation, or due to high RF currents through the movable beam under high RF power conditions.
- Self-actuation of the switch movable beam due to high RF power [8]. A high RF voltage may result in self-actuation because the attractive electrostatic force is proportional to the square of the switch voltage. This limits the switch power handling.
- Hot switching failure. When a capacitive switch needs to interrupt high RF currents or sustain high transient RF voltages, abrupt chemical changes may occur at its surfaces leading to premature wear and failure. This is related to the dielectric charging phenomena.
- Shock-induced failure. High shocks (>30,000–100,000 g) may result in beam fracture particularly if contact is achieved. Such events are rare in most applications.

Failures related to cycling-induced fatigue, crack generation, and fracture are not typically observed under normal operating conditions. However, they may become important at extreme temperatures particularly for movable beams based on thin-film metals. Despite the aforementioned failure modes, the best switches today have achieved over 100 billion cycles under typical laboratory conditions when driven by 30 kHz bipolar bias waveforms with approximately 35 V peak amplitude. These devices were hot-switched at a power level of 20 dBm at 35 GHz [21]. However, these cannot be considered typical results. Early failures are found in several wafer samples. Additional research is required to increase the observed reliability and limit early failures that are commonly due to poor fabrication process control.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ NEMS Piezoelectric Switches
- ▶ Piezoelectric MEMS Switch

References

1. Rebeiz, G.M.: RF MEMS Theory, Design, and Technology. Wiley, Hoboken (2003)
2. Webster, J.R., Dyck, C.W., Sullivan, J.P., Friedmann, T.A., Carton, A.J.: Performance of amorphous diamond RF MEMS capacitive switch. Electron. Lett. **40**(1), 43 (2004)
3. Goldsmith, C., Sumant, A., Auciello, O., Carlisle, J., Zeng, H., Hwang, J.C.M., Palego, C., Wang, W., Carpick, R., Adiga, V.P., Datta, A., Gudeman, C., O'Brien, S., Sampath, S.: Charging characteristics of ultra-nano-crystalline diamond in RF MEMS capacitive switches. In: Proceedings of the IEEE MTT-S International Microwave Digest, Anaheim, pp. 1246–1249, May 2010
4. Goldsmith, C.L., Kanack, B.M., Lin, T., Norvell, B.R., Pang, L.Y., Powers, B., Rhoads, C., Seymour, D.: Micromechanical microwave switching. US Patent 5,619,061, 31 Oct 1994

5. Goldsmith, C.L., Yao, Z., Eshelman, S., Denniston, D.: Performance of low-loss RF MEMS capacitive switches. IEEE Microwave Wireless Compon. Lett. **8**(8), 269–271 (1998)
6. Peroulis, D.: RF MEMS devices for multifunctional integrated circuits and antennas. PhD Dissertation, The University of Michigan, Ann Arbor (2003)
7. Peroulis, D., Pacheco, S.P., Sarabandi, K., Katehi, L.P.B.: Electromechanical considerations in developing low-voltage RF MEMS switches. IEEE T. Microw. Theory **51**(1), 259–270 (2003)
8. Peroulis, D., Pacheco, S.P., Katehi, L.P.B.: RF MEMS switches with enhanced power-handling capabilities. IEEE T. Microw. Theory **52**(1), 59–68 (2004)
9. Snow, M.: Comprehensive modeling of electrostatically actuated MEMS beams including uncertainty quantification. MSc thesis, Purdue University, West Lafayette (2010)
10. Younis, M.I., Abdel-Rahman, E.M., Nayfeh, A.: A - reduced-order model for electrically actuated microbeam-based MEMS. J. Microelectromech. Syst. **12**(5), 672–680 (2003)
11. Guo, X., Alexeenko, A.: Compact model of squeeze-film damping based on rarefied flow simulations. J. Micromech. Microeng. **19**(4), 045026 (2009)
12. Parkos, D., Raghunathan, N., Venkattraman, A., Alexeenko, A., Peroulis, D.: Near-contact damping model and dynamic response of micro-beams under high-g loads. In: Proceedings of the IEEE International Conference on Micro Electro Mechanical Systems (MEMS), Cancun, pp. 465–468, Jan 2011
13. Ayyaswamy, V., Alexeenko, A.: Coarse-grained model for RF MEMS device. MEMShub.org, http://memshub.org/resources/prismcg. Accessed Mar 2011
14. Tousimis: http://www.tousimis.com/. Accessed Mar 2011
15. Memtronics: http://www.memtronics.com/. Accessed Mar 2011
16. Garg, A., Ayyaswamy, V., Kovacs, A., Alexeenko, A., Peroulis, D.: Direct measurement of field emission current in E-static MEMS structures. In: Proceedings of the 24th IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2011), pp. 412–415, Jan 2011
17. Peng, Z., Yuan, X., Hwang, J.C.M., Forehand, D.I., Goldsmith, C.L.: Superposition model for dielectric charging of RF MEMS capacitive switches under bipolar control-voltage waveforms. IEEE T. Microw. Theory **55**(12), 2911–2918 (2007)
18. Papaioannou, G., Exarchos, M.-N., Theonas, V., Wang, G., Papapolymerou, J.: Temperature study of the dielectric polarization effects of capacitive RF MEMS switches. IEEE T. Microw. Theory **53**(11), 3467–3473 (2005)
19. Hsu, H.-H., Peroulis, D.: A viscoelastic-aware experimentally-derived model for analog RF MEMS varactors. In: Proceedings of the 23rd IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2010), Wanchai, pp. 783–786, Jan 2010
20. McLean, M., Brown, W.L., Vinci, R.P.: Temperature-dependent viscoelasticity in thin Au films and consequences for MEMS devices. IEEE/ASME J. Microelectromech. Syst. **19**(6), 1299–1308 (2010)
21. Goldsmith, C., Maciel, J., McKillop, J.: Demonstrating reliability. IEEE Microw. Mag. **8**(6), 56–60 (2007)

Capillarity Induced Folding

▶ Capillary Origami

Capillary Flow

Prashant R. Waghmare and Sushanta K. Mitra
Micro and Nano-scale Transport Laboratory, Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada

Synonyms

Passive pumping; Surface tension–driven flow

Definition

The fluid flow in an enclosed conduit due to simultaneous changes in the inherent surface energies of fluid and solid surface of the conduit.

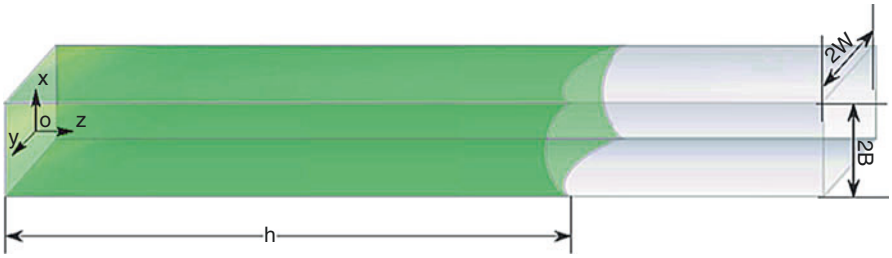
Introduction

Everything in the universe has its own state of energy, which is represented by possible combinations of 132 elements of periodic table. The simplest form of each element is an atom and each atom has three different components: proton, electron, and neutron. Further, each element has a fixed number of electrons, which arrange themselves in different shells, and the number of electrons in each shell can be determined by Bohr's theory. Several individual elements or atoms do not have sufficient number of electrons

in their outer shell and this makes it unstable and further the element tries to become stable by searching for required electrons. The finding of sufficient electrons at the outer shell results in the formation of a molecule or in bulk cluster of molecules. The surface or interface formation with this cluster of molecules creates an imbalance in the arrangement and orientation of the molecules in the cluster, mainly across the interface. This imbalance represents the energy of system. Every system in the universe tries to attain the minimum state of energy and, therefore, in the case of liquid in the air, it has been observed that the raindrops always attain the spherical shape. Liquid molecules have capability to orient themselves. Hence, they can form the shape of minimum energy but in the case of solid molecules, they try to minimize the state of energy by covering up the liquid if it comes in contact. In this case, the solid surfaces are not in equilibrium with the saturated vapor, i.e., they are not in a state of minimum energy. The moment at which conduit or channel with high energy level surfaces comes in contact with lower energy level liquid interface, the solid surface tries to envelope itself with the liquid to attain the minimum energy. In the case of the higher energy liquid in comparison with the solid surfaces, the liquid interface tries to minimize its surface. The prior case can be illustrated with the water (72 dyn/cm) in glass channel (~200–300 dyn/cm) and the mercury (~700–735 dyn/cm) with the same glass channel is an example of the latter case.

The recent developments in the microfabrication technologies allowed to fabricate features of sizes from micro- to nanoscales. The surface to volume ratio of the feature increases as the scale of the feature decreases which in turn makes surface forces dominant over other forces. Because of high surface forces, very high pressure is required to pump the fluid in microchannels. Hence, researchers have developed different nonmechanical pumping mechanism with the help of electrokinetic and/or megnetohydrodynamic approaches. Generation and actuation of electric and/or magnetic field are an additional burden on the system which increases both the fabrication process and cost of the device. Therefore, attempts are being made to develop a flow without any external means. The fluid flow can be achieved by controlling surface chemistry, fluid properties like surface tension, or by changing the geometries. Such transport of fluid is called autonomous flow or autonomous pumping which is an ideal transport mechanism for microfluidic applications. The dominance of surface forces at microscale plays a significant role in deciding the pumping approach in microfluidic devices. Hence, nowadays surface tension–driven flow in microfluidic devices has widely attracted the attention of researchers. As explained earlier, the capillary action is the interplay between the surface energies or surface tension between the fluid and solid surface in contact. Moreover, at microscale, due to very high surface to volume ratio, the possibility of available surface area is very high. One can easily pump the fluid with capillarity provided the fluid has lower surface energy than the solid surfaces. For optimum design and function of any microfluidic device which works on capillary flow principle, it is essential to predict its behavior in advance and it is therefore necessary to perform the theoretical analysis of capillary flow within the microchannels. The prediction of the temporal variations in the flow front position along capillary length is the ultimate goal of the theoretical analysis. Therefore, generally the analysis is performed to predict the flow front position, i.e., penetration depth for given working and operating conditions.

Over the last century, the capillary phenomenon has become a topic of interest due to its importance in several areas. First time in literature, Washburn proposed a closed form solution for the penetration depth in a channel of millimeter dimension. The closed form solution is derived by balancing the surface tension force to viscous force and it is observed that the penetration depth is proportional to the square root of the time. As explained earlier, capillary phenomenon is the change in the surface energy process and to encompass the concept of change in the surface energy, one can use the thermodynamic approach for analysis. The surface energy of solid is the



Capillary Flow, Fig. 1 Schematic of the microchannel of width 2B, depth 2 W considered for theoretical modeling [6]

topic of ongoing debate. Moreover, there are several effects like dynamic contact angle, inlet effects, reservoir effects, suspension flow, etc., which require tedious and cumbersome analysis. Attempts are also being made to present analysis with microscopic energy balance approach where different forces are accounted in terms of different forms of energy. On the other hand, hydrodynamic models, based on the conventional fluid mechanics principle, are easy to implement. Such models are mainly developed by two distinct approaches in the literature, namely, differential and integral approach. The moving fluid-air interface with the differential approach becomes computationally costly, whereas the integral approach with moving control volume provides a simple form of ordinary differential equation. In such modeling or analysis, the governing equation for flow front transport is obtained by balancing different forces like viscous, inertial, gravity, pressure forces, etc. The velocity-dependent terms like inertial and viscous terms of the governing equation are determined with the velocity profile across the channel. In the literature, the steady state assumption is applied from the very entrance of the channel neglecting the entrance length effect. This assumption has been widely adopted till date as done by Washburn where the time and length scale used for the validation of the theoretical model was big. Hence, the assumption of a steady state velocity profile holds true. Whereas, in the case of microfluidic channels the length and timescale is very small; hence, the Washburn prediction does not follow the observation as demonstrated by Saha and Mitra [1, 2]. The later part of this entry is dedicated to emphasize the importance of such microscale effects in the analysis. Main emphasis is given to the integral approach based on modeling due to the ease of its adaptability. Therefore, in the next section, the overview of the modeling of the capillary transport is discussed in brief.

Mathematical Modeling

Figure 1 shows the microchannel of width 2B and depth of 2 W is considered for the theoretical modeling. The momentum equation in integral form for homogeneous, incompressible, and Newtonian fluid can be written as [3]:

$$\sum F_z = \frac{\partial}{\partial t}\int_0^h \int_{-W}^{W} \int_{-B}^{B} \rho v_z dx dy dz + \int_{-W}^{W} \times \int_{-B}^{B} v_z(-\rho v_z) dx dy \tag{1}$$

where $\sum F_z$ refers to all forces present during the development of the fluid-air interface, ρ is the fluid density. Generally, forces present during the capillary transport are viscous (F_v), gravity $\left(F_g = 4\rho g h B W\right)$, pressure forces at the flow front (F_{pf}), and at the inlet (F_{pi}).

$$\sum F_z = \underbrace{F_v}_{\text{velocity dependent}} + \underbrace{F_g + F_{pf} + F_{pi}}_{\text{velovity independent}} \tag{2}$$

The Eq. 1 contains three velocity-dependent terms, namely, transient, convective, and viscous force which is generally determined as velocity profile, v_z, across the channel. The fully developed

flow assumption, i.e., Poiseuille flow assumption, is widely used in the literature neglecting the transience in the velocity profile [4, 5]. The consequence of such assumptions particularly at microscales will be discussed in detail in the later part.

The steady state velocity profile can be used to determine the velocity-dependent terms of the momentum equation. Moreover, remaining pressure force terms are calculated with available expressions (from the literature) for pressure fields at respective locations. The pressure force at the fluid-air interface can be determined by well-known Young-Laplace equation with fluid surface tension (σ) and equilibrium contact angle (θe). The approximated pressure field expression at the entrance of the microchannel is used widely to determine the pressure force at the entrance of the microchannel. Levin et al. [7], for the first time in the literature, claimed that atmospheric pressure cannot be used as entrance pressure at the inlet of the microchannel. The pressure field expression for circular capillary is derived by assuming a separate hemispherical control volume as fluid source other than the control volume considered within the microchannel. Further several researchers have used same expression for rectangular microchannels with an assumption of equivalent radius. In such analysis, with an equivalent radius assumption, the hemispherical control volume of equivalent radius of projected area of rectangular microchannel entrance is presented in Eq. 3.

$$p(o,t) = p_{atm} - \left\{ 1.11\rho\sqrt{BW}\frac{d^2h}{dt^2} + 1.58\rho\left(\frac{dh}{dt}\right)^2 + \frac{1.772\mu\, dh}{\sqrt{BW}\, dt} \right\} \tag{3}$$

The importance of an appropriate entrance pressure field expression for rectangular microchannel is also discussed in detail in later part of the study. Finally, determining all terms of Eq. 1 and rearranging as per order of differential operator, one can obtain the dimension form of the ordinary differential equation which governs the capillary transport in the microchannels. Further, nondimensional governing equation as shown in Eq. 4 can be obtained by performing nondimensional analysis with characteristic time, $t_0 = \frac{\rho(2B)^2}{12\mu}$ and characteristic length $h_0 = 2B$.

Capillary Flow, Table 1 Constants of the generalized nondimensional governing equation for a capillary flow in a microchannel with fully developed velocity profile [3]

Constants	Expressions
C_1	$\frac{0.55}{\sqrt{\gamma}}$
C_2	0.958
C_3	1
C_4	$0.295\sqrt{\gamma}$
C_5	$\frac{B_o}{144Oh^2}$
C_6	$\frac{\gamma - cos\,\theta_e}{72Oh^2}$

$$(h^* + C_1)\frac{d^2h^*}{dt^{*2}} + C_2\left(\frac{dh^*}{dt^*}\right)^2 + (C_3 + C_4h^*)\frac{dh^*}{dt^*} + C_5h^* + C_6 = 0 \tag{4}$$

The coefficients of Eq. 4 are tabulated in Table 1. Two nondimensional numbers are obtained, i.e., Bond number (Bo) and Ohnesorge number (Oh). The constants of this equation are functions of different nondimensional groups like Ohnesorge number (Oh), Bond number (Bo), and aspect ratio (γ). The Ohnesorge number represents the ratio of viscous to surface tension force, i.e., $Oh = \frac{\mu}{\sqrt{2B_{\rho\sigma}}}$, the Bond number dictates the ratio of gravity to surface tension force, i.e., $\left(B_O = \frac{\rho g(2B^2)}{\sigma}\right)$.

The solution of Eq. 4 predicts the penetration depth with capillary flow. The numerical [8] and analytical [3] solutions of Eq. 4 are available in the literature.

Revisiting the Assumptions for Microscale Applications

As mentioned earlier, the velocity-dependent terms of momentum equation are determined with an assumption of a steady state. It is assumed that at the very entrance of the microchannel the flow is fully developed. However, in reality, three different flow regimes can be observed in the capillary flow: entry regime, Poiseuille regime,

where the flow is fully developed, and the regime behind the fluid-air interface, i.e., surface tension regime. The steady state assumption, i.e., parabolic velocity profile assumption, is valid for steady state flow, whereas capillary flow is inherently a transient phenomenon. The parabolic velocity profile assumption is a reasonably good assumption for macroscale capillaries as shown by several researches [5, 7, 9]. Moreover, such assumptions are only valid in the case of a very high viscous fluid or very low Reynolds number flow which may not be true in every case [10]. Hence, it is important to consider and analyze such transience in the analysis at microscale. This can be tackled by considering the transient or developing velocity profile instead of the steady state velocity profile as explained in the following section.

The transient momentum equation in the direction of the flow for pressure-driven flow is:

$$\rho \frac{\partial v_z}{\partial t} = \mu \frac{\partial^2 v_z}{\partial x^2} + \frac{dp}{dz} \tag{5}$$

The velocity in Eq. 5 is a combination of steady and transient part of velocity as depicted in Eq. 6 [11]:

$$v_z(x,t) = v_{z\infty}(x) + v_{zt}(x,t) \tag{6}$$

where $v_{z\infty}(x)$ is the fully developed or steady state velocity, i.e.,

$$v_{z\infty}(x) = \frac{B^2}{2\mu}\frac{dp}{dz}\left[1 - \left(\frac{x}{B}\right)^2\right] \tag{7}$$

The transient part of the velocity can be obtained by separation of the variable method which can be given as [6]:

$$v_{zt}(x,t) = 2\sum_{n=1}^{\infty}(-1)^n\left[\frac{1}{B\mu\lambda_n^3}\frac{dp}{dz}\right]cos\,(\lambda_n x)exp \times \left(-v\lambda_n^2 t\right) \tag{8}$$

where v is the kinematic viscosity of the fluid and $\lambda_n = \frac{(2n-1)\pi}{2B}$. By combining Eqs. 5 and 7 the transient velocity profile $v_z(x, t)$ can be obtained by:

$$v_z(x,t) = \left\{\sum_{n=1}^{\infty}(-1)^n\frac{1}{B\mu}\left(\frac{2}{\lambda_n^3}\right)cos\,(\lambda_n x)exp \times \left(-v\lambda_n^2 t\right) + \frac{1}{2\mu}\left(B^2 - x^2\right)\right\}\frac{dp}{dz} \tag{9}$$

Further, the average velocity across the channel can be represented as:

$$v_z(t)_{avg} = \frac{B^2}{3\mu}\left[1 - \sum_{n=1}^{\infty}\frac{96}{(2n-1)^4\pi^2}exp \times \left(-\frac{(2n-1)^2\pi^2 vt}{4B^2}\right)\right]\frac{dp}{dz} \tag{10}$$

Finally, the velocity profile in terms of the penetration with transient velocity is:

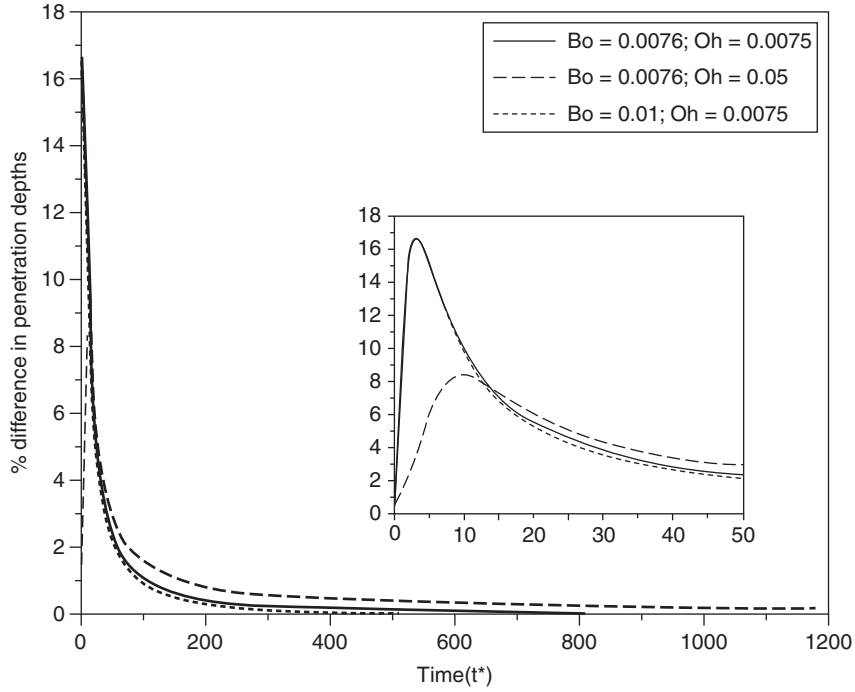
$$v_z(x,t) = \frac{B^2}{2\mu}\alpha_1\left\{\sum_{n=1}^{\infty}(-1)^n\left[-\frac{4}{(B\lambda_n)^3}\right]cos\,(\lambda_n x)exp\left(-v\lambda_n^2 t\right) + \left(1 - \frac{x^2}{B^2}\right) \times \left\{\frac{1}{\alpha_1\left[1 - \sum_{n=1}^{\infty}\beta_1 exp\left(-\lambda_n^2 vt\right)\right]}\right\}\frac{dh}{dt} \tag{11}$$

where

$$\alpha_1 = \frac{\left[(\phi)^4 - 4exp - \frac{\phi^2 t}{3}\right]}{\left[(\phi)^4 - 6exp - \frac{\phi^2 t}{3}\right]}$$

and $\phi = \lambda_n B$.

A similar approach can be followed as explained in the previous section and further the governing equation for capillary transport can be derived with the transient velocity profile provided in Eq. 11. Moreover, the difference in the penetration depth with both approaches, i.e., with the steady state and transient velocity profile, under different operating conditions can be compared. Figure 2 shows the difference in the penetration depth in such cases where the difference in the penetration depth is more at the beginning of

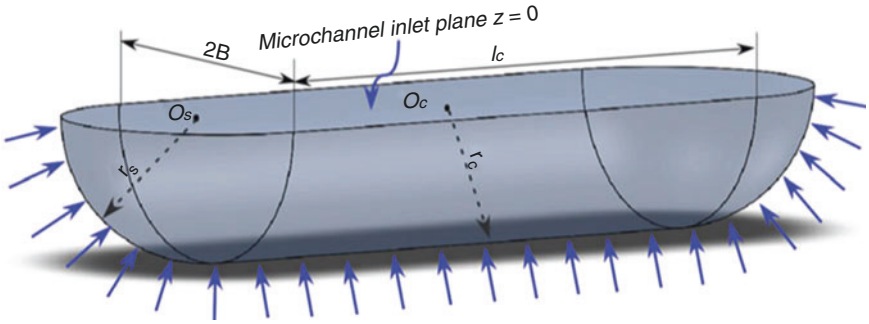


Capillary Flow, Fig. 2 Transient response in the difference in the penetration depths with the fully developed (steady state) and developing (unsteady) velocity profile under different conditions [6]

the filling process, as shown in the inset of Fig. 2. This difference in the penetration depth decreases as the flow progresses along the microchannel where the flow becomes a fully developed flow. This can also be explained with the help of boundary layer theory which is the effect of fluid viscosity. The boundary layer thickness increases as the viscosity of fluid increases because of the retardation of flow due to increase in the viscosity, whereas in the case of the fluid density, the effect is opposite to viscosity. Therefore, the difference in penetration depth with the high density fluid (Bo $= 0.01$) is higher than the difference with the high viscous fluid. It is evident from the analysis that the transience effect in the analysis has a significant impact on the filling process prediction, particularly at the beginning of the filling process. At microscale, such difference needs to be accounted prior to the design.

As discussed earlier, the pressure force at the entrance of the microchannel is determined with the help of the pressure field at the microchannel entrance. Several researchers [3–8] have adopted the pressure field expression with an equivalent radius assumption. Levin et al. [6] developed an entrance pressure field expression for circular capillary, assuming a hemispherical control volume as a separate control volume at the entrance which is responsible for a sink flow at the entrance of capillary and the pressure field. Moreover, a similar expression for rectangular capillaries is extended with an equivalent radius assumption. In such cases, the radius of circular capillary is replaced by the equivalent radius of projected area at the entrance of the channel. This is not a realistic representation for noncircular capillaries particularly for high aspect ratio microchannels where it is not appropriate to consider the hemispherical control volume for the sink flow or pressure field at the microchannel entrance. In the case of such geometries, the control volume needs to be considered as a combination of semicylinder and hemisphere as shown in Fig. 3.

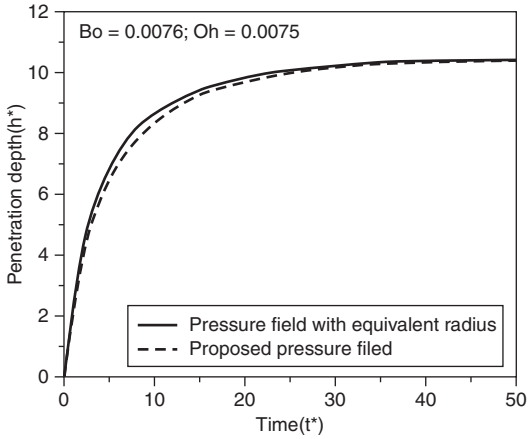
The detailed derivation of the pressure field expression with this control volume can be seen in [6] which is:



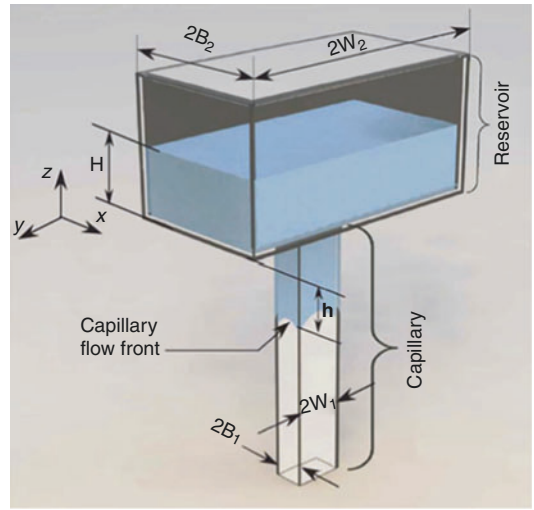
Capillary Flow, Fig. 3 The fluid volume from infinite reservoir considered as control volume for pressure field expression analysis in the case of rectangular microchannel. The *arrow* shows the direction of the fluid flow from the reservoir into the microchannel [6]

$$p(0,t) = p_{atm} - \rho B\left\{\left[\frac{4\gamma + 3(1-\gamma)}{24}\right]\left\{\pi\left[\frac{1}{2\pi} + \frac{2}{\pi^2} + \frac{6}{10}\right]\right\} + \left[1 - \frac{2}{\pi} ln\frac{R_\infty}{B}\right]\right\}\frac{d^2h}{dt^2}$$
$$+\rho\left\{\left[\frac{1(1-\gamma)}{\pi^2} - \frac{6}{5}\right] - \left[\frac{4\gamma + 3(1-\gamma)}{6}\right] \times \left[\frac{(2-\gamma)}{2\pi} - \frac{(1-\gamma)}{\pi^2}\right]\right\}\left(\frac{dh}{dt}\right)^2 - \frac{4\mu}{B} \times \left[(2-\gamma) + \frac{(1-\gamma)}{\pi}\right]\frac{dh}{dt} \quad (12)$$

where R_∞ represents the radial distance far away from the control volume in the reservoir, where the sink action, i.e., entrance pressure force, disappears. One can re-derive the governing Eq. 4, using pressure field expression presented in the Eq. 12, and determine the effect of such a pressure field on the analysis. Figure 4 shows the comparison of variations in the penetration depth with recently proposed pressure and with equivalent radius field expressions. The approximated pressure field overpredicts the penetration depth. The difference in the penetration depth with the proposed pressure field is significant, which shows that it is important to consider the proposed pressure field for a rectangular microchannel rather than an approximated pressure field. The transport with a capillary action is the balance among surface, viscous, and other body forces which retard the flow as it progresses. Hence, the capillary flow always attains a steady state which is generally termed as an equilibrium penetration depth in the literature. If the length of the channel is longer than that of the equilibrium penetration depth, then flow front cannot reach the outlet and, therefore, the assistance to the capillary flow is attempted in such cases. Passive or nonmechanical pumping approaches combined with the capillary flow serve this enhancement. The scaling analysis suggests that the gravity force is less dominant at microscale [12], but several researchers have demonstrated that gravity can be used as an assistance to the capillary flow [13–15]. Generally, the capillary flow analysis is performed with an assumption of infinite reservoir. Hence, the reservoir effect and the gravitational force from the reservoir are generally neglected in the analysis. To accommodate the entrance effect of finite size reservoir at the inlet of the microchannel in the theoretical modeling, the entrance pressure field is developed for the arrangements shown in Fig. 5. The rectangular microchannel with rectangular reservoir on the top of the microchannel is considered and the pressure field with the gravity and reservoir effect is developed in the flow [16]. Moreover, this pressure field is used to obtain the governing equation for capillary flow under the influence of gravity head from the reservoir.



Capillary Flow, Fig. 4 The comparison of variations in the penetration depth with equivalent radius and recently proposed pressure field expressions. Figure 4 shows the comparison of penetration depth for $\gamma = 0.9$ with the corresponding difference in the penetration depth [6]

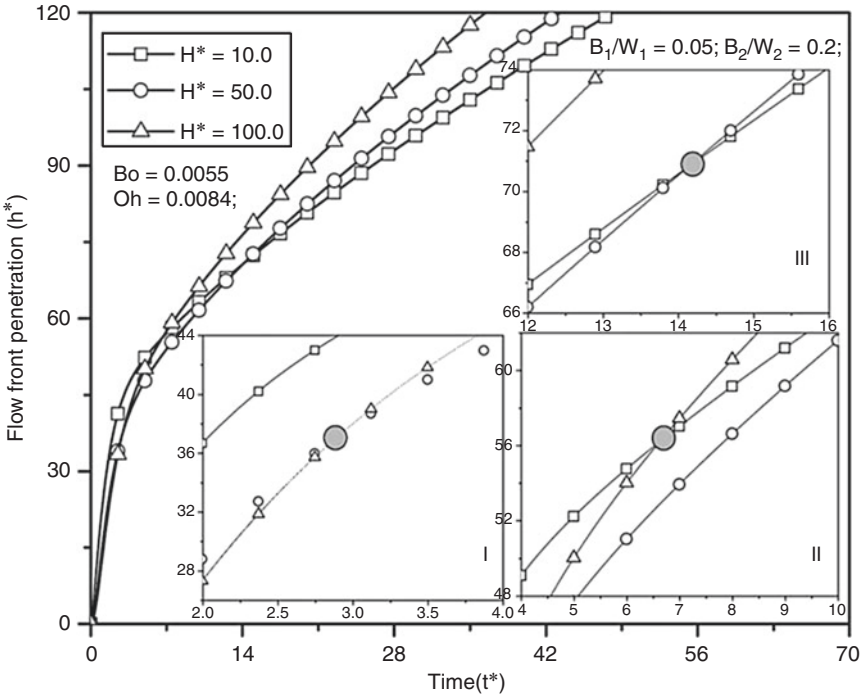


Capillary Flow, Fig. 5 Schematic of a gravity-assisted capillary flow in a vertically oriented capillary of width $2B_1$ and depth $2\ W_1$. The additional gravitational head from the fluid in a finite reservoir of size ($2B_2 \times 2\ W_2$) is assisting the capillary flow [16]

The reservoir with three different levels of fluid in the reservoirs (H^*), namely, 10, 50, and 100, is considered for the analysis. Figure 6 shows the variations in the penetration depth (h^*) with different operating conditions. This analysis represents the interplay between the surface tension force and gravity head from the reservoir. The capillary flow takes place in the channel which remains same for all three cases, whereas the level of the fluid from the reservoir increases from case I to case III. Thus, for three cases, the capillary effect is the same but the gravity head is different. At the beginning of the transport, the fluid from the reservoir offers less inertia to the fluid transport and the capillary force dominates over the gravity from the reservoir. Therefore, at the beginning of the transport, the penetration depth with a lower reservoir fluid level ($H^* = 10.0$) is higher than the other two penetration depths as shown in the inset I. Similarly, the penetration depth with the highest reservoir fluid level ($H^* = 100.0$) has the lowest penetration depth as compared to others.

Moreover, as the fluid progresses in the microchannel, the momentum from the reservoir fluid assists the capillary flow and the gravitational force, due to which the fluid from the reservoir becomes dominant over the capillary force within the microchannel. This results in transcendence among the penetration depths with a different gravity head. The penetration depth with the highest gravity head ($H^* = 100$) surpasses the penetration depth with gravity head $H^* = 50$ and $H = 10^*$ in inset I and II two, respectively. This can be attributed to as an interplay between the surface tension force, i.e., the capillarity and gravitational force from reservoir. In the case of microfluidic applications, the sizes of reservoir and microchannel are comparable to each other. Hence one cannot neglect the effect of the reservoir in such cases, particularly if it is surface tension–driven pumping. Further, one can assist the capillary flow with an appropriate arrangement of reservoir.

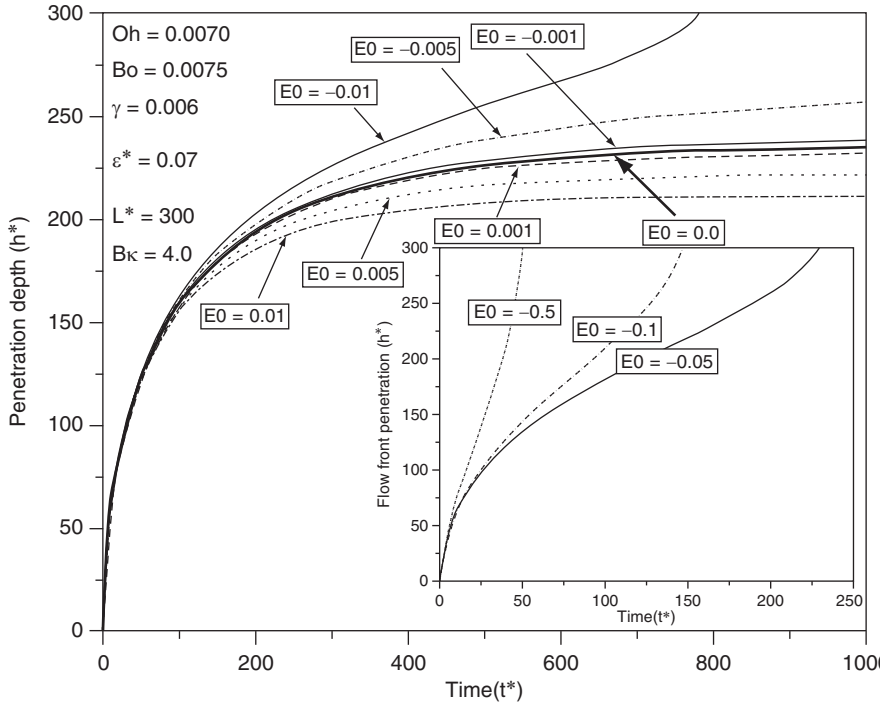
There are always certain limitations to the autonomous pumping which make them inadequate in long microchannels. Hence, it is important to enhance the pumping ability by other means. Further enhancement in the capillary flow can be achieved by coupling the capillary flow with the electroosmotic flow which is one of the electrokinetic pumping mechanisms. In most of the cases, the inner wall of a microchannel always has surface charges due to different



Capillary Flow, Fig. 6 Transient response of a flow front transport for different gravitational heads in the reservoir with Bo = 0.0055, Oh = 0.0084, $B_1/W_1 = 0.05$ $B_2/W_2 = 0.2$. (**I**) Flow front penetration rate for $H^* = 100$ surpasses the penetration rate for $H^* = 50$. (**II**) Flow front penetration rate for $H^* = 100$ surpasses the penetration rate for $H^* = 10$. (**III**) Flow front penetration rate for H = 50 surpasses the penetration rate for $H^* = 10$ [16]

mechanisms like ionization, dissociation of ions, isomorphic substitution, etc., [17]. These surface charges distribute ions of the electrolytes in a specific pattern when brought into contact with an electrolyte which is generally termed as the formation of electrical double layer (EDL). After applying the electric field across the channel, the movement of the ions takes place, which results in the movement of the fluid due to an electric field [18]. The electrolyte solution is transported with the capillary action; one can further assist the capillary flow. An additional body force due to electroosmotism is added to Eq. 2, which accommodates the additional effect of electroosmosis. Further one can analyze the interplay between the capillarity and electroosmotisms as presented in the recent studies [19]. Figure 7 shows the variation in the penetration depth of the capillary flow under the influence of electroosmotism. Through a nondimensional analysis, a new nondimensional number is proposed, i.e., Eo which represents the ratio between the surface tension force and electroosmotic force. The direction of the electroosmotic flow can be reversed by changing the electric field direction. Hence, negative and positive Eo numbers are observed in the analysis. The negative Eo numbers represent the change in the direction of the electric field as compared to positive Eo numbers. The pure capillary flow can be seen as Eo = 0.

The variation in the penetration depth under three different operating conditions is shown in Fig. 7. As observed in the pure capillary case (Eo = 0), the penetration depth attains the equilibrium penetration depth, whereas in the case of − Eo numbers, the equilibrium penetration depth increases with increment in the magnitude of − Eo numbers. This represents that in the case of electroosmotic flow with − Eo number, the capillary flow is assisted by electroosmotism. In this analysis, the nondimensional length of the



Capillary Flow, Fig. 7 Variation in the penetration depth for vertically oriented channel with water as electrolyte, where B = 100 μm, W = 400 μm, L = 75 mm, $\zeta = -75$ mV and constant contact angle is 27°. The inset shows the variation in the electric field within the electrolyte as the flow front progresses under different applied voltages [19]

microchannel (L^*) is considered as 300 and with –Eo = −0.01 the entire filling of the microchannel is observed. In the case of positive Eo numbers, it is observed that the electroosmotism acts in a opposite direction of the capillary flow. Hence, it retards the flow and this can be observed by the decrement in the equilibrium penetration depth with the increment in the positive Eo numbers.

For the enhancement in the capillary flow, a transport with additional gravity head and electroosmotic forces are considered. A generalized theoretical modeling for a gravity-assisted capillary flow with reservoir effects and electroosmotically assisted capillary flow is reported in brief. It is observed that even though the scaling among forces suggests that the gravitation force is negligible at microscale, the reported analysis infers that with a finite reservoir, an added advantage due to gravity can be a useful tool to transport the fluid at microscale. This added force for the capillary transport can be utilized without any additional burden in the design of the LOC device. The electroosmotically assisted capillary flow model suggests that in a combined flow the electrokinetic parameters have an important influence on the capillary flow. Such electrokinetic flow approaches can be coupled to enhance the capillary flow transport in the microchannel.

The wetting properties of the fluid decide the capability of pumping with a capillary flow. Therefore, it is important to know the precise magnitude of wetting properties like contact angle and surface tension of the working fluid. The microfluidics has become a promising option for biomedical application and inclusion of biomolecules is an unavoidable part in such applications. In most cases, the biomolecules are attached with the microbeads and transported to the desired locations. It is evident from the experimental analysis that the inclusion of microbeads changes the wetting behavior drastically [20]. Therefore, it is

necessary to consider the effect of microbeads in the fluid for the analysis. This can be done by considering the following expressions for surface tension and contact angle: density and viscosity which are functions of volume fraction of microbeads. Such correlations of the surface tension and contact angle are provided in [20]. Such expressions for the variation in the contact angle and surface tension with the volume fraction can be readily used in modeling transport processes of microbead suspensions in micro-capillaries, used in the microfluidic devices.

In passive pumping, particularly with the capillary flow, different aspects due to microscale effects like aspect ratio–dependent velocity profile, contact angle at four walls, fluid-air interface dynamics in the case of suspension flow, etc., need to be investigated in detail. Theoretically the concept of electroosmotically assisted capillary flow has been presented but the experimental demonstration of such phenomena is also an interesting area of research. Moreover, wetting of biomolecule suspensions under transient effects instead of the steady state is also needed to be studied. The experimental study of the flow behind the front and at the entrance of the microchannels is also an interesting study to perform.

Cross-References

- ▶ AC Electroosmosis: Basics and Lab-on-a-Chip Applications
- ▶ Electrowetting
- ▶ Micro/Nano Flow Characterization Techniques
- ▶ Micropumps
- ▶ Surface Tension Effects of Nanostructures
- ▶ Wetting Transitions

References

1. Saha, A., Mitra, S.K.: Numerical study of capillary flow in microchannels with alternate hydrophilic-hydrophobic bottom wall. J. Fluid Eng. Trans. ASME **131**, 061202 (2009)
2. Saha, A., Mitra, S.: Effect of dynamic contact angle in a volume of fluid (VOF) model for a microfluidic capillary flow. J. Colloid Interface Sci. **339**, 461–480 (2009)
3. Xiao, Y., Yang, F., Pitchumani, R.: A generalized flow analysis of capillary flows in channels. J. Colloid Interface Sci. **298**, 880–888 (2006)
4. Washburn, E.: The dynamics of capillary flow. Phys. Rev. **17**, 273 (1921)
5. Chakraborty, S.: Electroosmotically driven capillary transport of typical non-Newtonian biofluid in rectangular microchannels. Anal. Chim. Acta **605**, 175–184 (2007)
6. Waghmare, P.R., Mitra, S.K.: A comprehensive theoretical model of capillary transport in rectangular microchannels. Microfluid. Nanofluid. (2011). doi:10.1007/s10404-011-0848-8
7. Levin, S., Reed, P., Watson, J.: A theory of the rate of rise a liquid in a capillary. In: Kerker, M. (ed.) Colloid and Interface Science, p. 403. Academic, New York (1976)
8. Marwadi, A., Xiao, Y., Pitchumani, R.: Theoretical analysis of capillary-driven nanoparticulate slurry flow during a micromold filling process. Int. J. Multiph. Flow **34**, 227 (2008)
9. Dreyer, M., Delgado, A., Rath, H.: Fluid motion in capillary vanes under reduced gravity. Microgravity Sci. Technol. **4**, 203 (1993)
10. Bhattacharya, S., Gurung, D.: Derivation of governing equation describing time-dependent penetration length in channel flows driven by non-mechanical forces. Anal. Chim. Acta **666**, 51–54 (2010)
11. Keh, H., Tseng, H.: Transient electrokinetic flow in fine capillaries. J. Colloid Interface Sci. **242**, 450 (2001)
12. Nguyen, N., Werely, S.: Fundamentals and Applications of Microfluidics. Artech House, New York (2003)
13. Yamada, H., Yoshida, Y., Terada, N., Hagihara, T., Teasawa, A.: Fabrication of gravity-driven microfluidic device. Rev. Sci. Instrum. **79**, 124301 (2008)
14. Jong, W.R., Kuo, T.H., Ho, S.W., Chiu, H.H., Peng, S. H.: Flows in rectangular microchannels driven by capillary force and gravity. Int. Commun. Heat Mass Transf. **34**, 186–196 (2007)
15. Kung, C., Chui, C., Chen, C., Chang, C., Chu, C.: Blood flow driven by surface tension in a microchannel. Microfluid. Nanofluid. **6**, 693 (2009)
16. Waghmare, P.R., Mitra, S.K.: Finite reservoir effect on capillary flow of microbead suspension in rectangular microchannels. J. Colloid Interface Sci. **351**(2), 561–569 (2010)
17. Hunter, R.: Zeat Potential in Colloid Science, Principle and Applications, Principle and Applications. Academic, London (1981)
18. Israelachvili, J.N.: Intermolecular and Surface Forces. Academic, London (1998)
19. Waghmare, P.R., Mitra, S.K.: Modeling of combined electroosmotic and capillary flow in microchannels. Anal. Chim. Acta **663**, 117–126 (2010)
20. Waghmare, P.R., Mitra, S.K.: Contact angle hysteresis of microbead suspensions. Langmuir **26**, 17082–17089 (2010)

Capillary Origami

Supone Manakasettharn[1], J. Ashley Taylor[2] and Tom N. Krupenkin[2]
[1]National Nanotechnology Center (NANOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand
[2]Department of Mechanical Engineering, The University of Wisconsin-Madison, Madison, WI, USA

Synonyms

Capillarity induced folding; Elasto-capillary folding; Surface tension–powered self-assembly

Definition

Capillary origami is the folding of an elastic planar structure into a three-dimensional (3D) structure by capillary action between a liquid droplet/bubble and a structure surface.

Why Capillary Origami?

The fabrication of 3D structures is one of the major challenges for micro- and nano-fabrication. Folding of an elastic planar structure after patterning and release is one technique to fabricate a 3D structure using self-assembly. The term *origami* is taken from the Japanese art of paper folding; while the actuation of the folding is accomplished by using capillary forces of a fluid droplet, hence, the technique has been termed *capillary origami*. The combination of the folding process with capillary forces has resulted in a new technique for micro- and nano-fabrication.

History

The term *capillary origami* was first introduced in 2007 by Charlotte Py et al. to describe the folding of a polydimethylsiloxane (PDMS) sheet into a 3D structure by using capillary forces created by a water droplet [1]. As early as 1993, Syms and Yeatman demonstrated that 3D structures could be fabricated by folding surfaces using capillary forces produced by molten solder [2]. Later Richard R. A. Syms introduced the term *surface tension-powered self-assembly* to describe the technique [3, 4]. Both of these techniques are quite similar in that 3D structures can be produced by folding elastic thin films. Both use capillary forces for self-assembly. In the first example, various liquids such as water are used, while for the second study, molten metals such as solder were used, which then solidified fixing the 3D microstructures.

Principles

At the macroscale level, the influence of capillary forces is negligible compared to other forces such as gravity, electrostatic, or magnetic. Because capillary forces scale linearly with the characteristic size of the system, at submillimeter dimensions, capillary forces begin to dominate since the majority of other forces decrease much more rapidly than the first power of the length. For example, a human cannot walk on water because capillary forces produced at the water surface are much smaller than the gravitational force acting on a human, which scales as the cube of the length. On the other hand, the much smaller water strider can easily walk on water because capillary forces are large enough to balance the gravitational force produced by the water strider. For capillary origami, capillary forces need to be large enough to counteract the weight of the liquid droplet and the structural forces of the planar layer.

In terms of energy, for capillary origami, one needs to consider the interplay of three different energies: capillary energy, bending energy, and gravitational potential energy. For a two-dimensional (2D) model, the capillary energy per unit length of the interface (2D analog of the surface energy) is defined as $E_c = L'\gamma$, where L' is the length of the interfacial surface of the fluid and γ is the surface tension [5]. The bending energy per unit length is approximately $E_b = \frac{LB}{2R^2}$, where

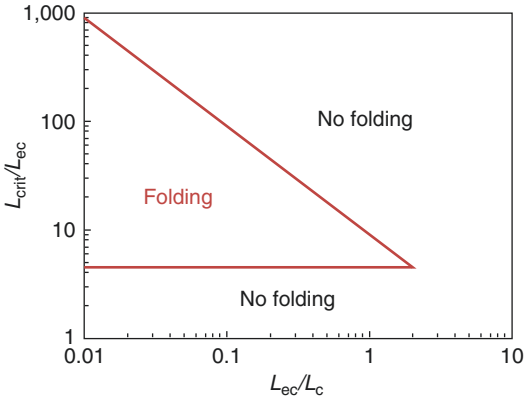
L is the length of the structure and R is the radius of curvature [6]. $B = \frac{Eh^3}{12(1-\nu^2)}$ is the bending rigidity of the structure, where E is Young's modulus, h is thickness of the layer, and ν is Poisson's ratio. If one only considers the mass of the fluid, assuming that it is much larger than the mass of the structure, then the gravitational potential energy per unit length is $E_g = \rho S g z$, where ρ is the density, S is the surface area, g is the constant of gravity, and z is the height of the center of mass. By neglecting the effect of gravity and assuming complete circular folding, [7] derived simplified criteria for folding considering the interplay between capillary and bending energies, which are expressed as

$$\sqrt{2\pi} < \frac{L_{\text{crit}}}{L_{\text{ec}}} < 2\sqrt{2\pi}\frac{L_{\text{c}}}{L_{\text{ec}}} \qquad (1)$$

where L_{crit} is the critical length of a structure for folding to occur, $L_{\text{c}} = \sqrt{\frac{\gamma}{\rho g}}$ is the capillary length [5], and $L_{\text{ec}} = \sqrt{\frac{B}{\gamma}}$ is the elasto-capillary length [8].

The simplified criteria for folding derived from Eq. 1 can be plotted as shown in Fig. 1. To fold a structure requires $\frac{L_{\text{ec}}}{L_{\text{c}}} < 2$ or $L_{\text{c}} > \frac{L_{\text{ec}}}{2}$ or $\gamma > \frac{\sqrt{B\rho g}}{2}$ indicating that the capillary length must be larger than half of the elasto-capillary length so that the capillary effect can overcome bending rigidity of the structure. The other requirement for folding is $\frac{L_{\text{crit}}}{L_{\text{ec}}} > \sqrt{2\pi} \cong 4.44$ or $L_{\text{crit}} > 4.44 L_{\text{ec}}$ confirming that the length of the structure should also be long enough for a liquid droplet to wet the surface to produce sufficient capillary forces to fold the structure. For the 3D structures, the folding criteria become more complex. In particular in 3D, the critical length also depends on the shape of the initial template such that $L_{\text{crit}} \cong 7L_{\text{ec}}$ for squares and $L_{\text{crit}} \cong 12L_{\text{ec}}$ for triangles [9]. Figure 2 shows examples of capillary origami structures of a pyramid, a cube, and a quasi-sphere obtained from folding triangle-, cross-, and flower-shaped PDMS sheets, respectively [10].

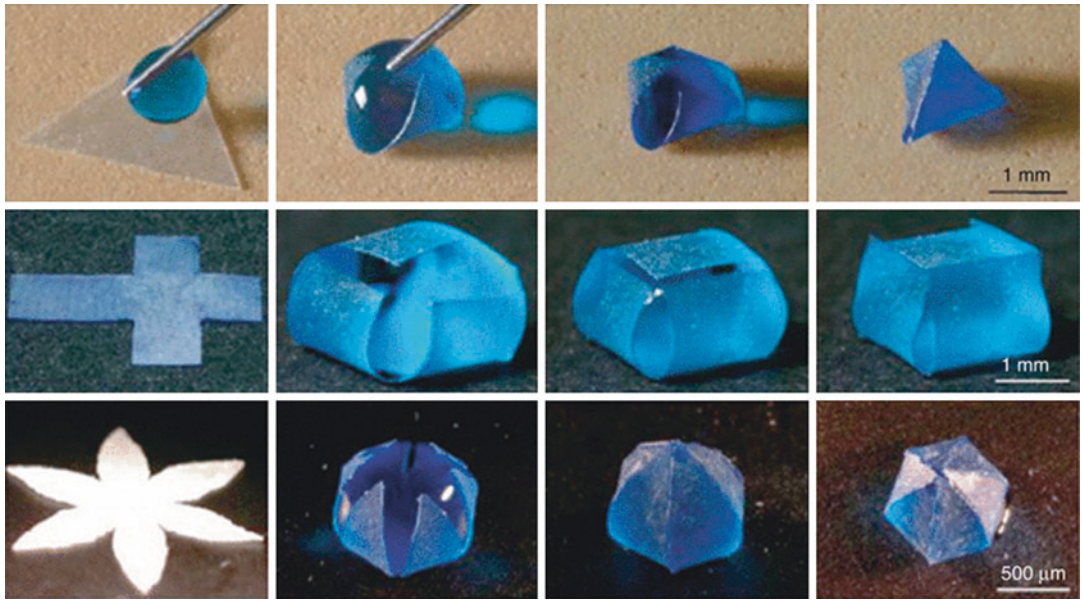
The concept of capillary origami has also been modeled at the nanoscale. Molecular dynamics simulations show the folding of planar graphene nanostructures by water nanodroplets [11]. Figure 3 summarizes the folding by presenting a phase-like diagram by plotting droplet radius (R_d) versus the ribbon width (w). From the graph, four different areas or phases can be identified corresponding to four different bending modes: nonfolding, sliding, zipping, and rolling. Figure 3 (left) shows the nonfolding phase where the nanoribbon end does not fold around the nanodroplet. The nonfolding phase is adjacent to the sliding phase where the nanoribbon end folds around the nanodroplet and then slides on the nanoribbon surface. The right top of the figure shows the zipping phase where the width of ribbon is several times greater than the droplet radius allowing the ribbon to fold around the droplet in the orthogonal direction. In the rolling phase ($w \leq \frac{1}{2}R_d$), the ribbon folds around the droplet and then continues wrapping in a scroll-like manner.



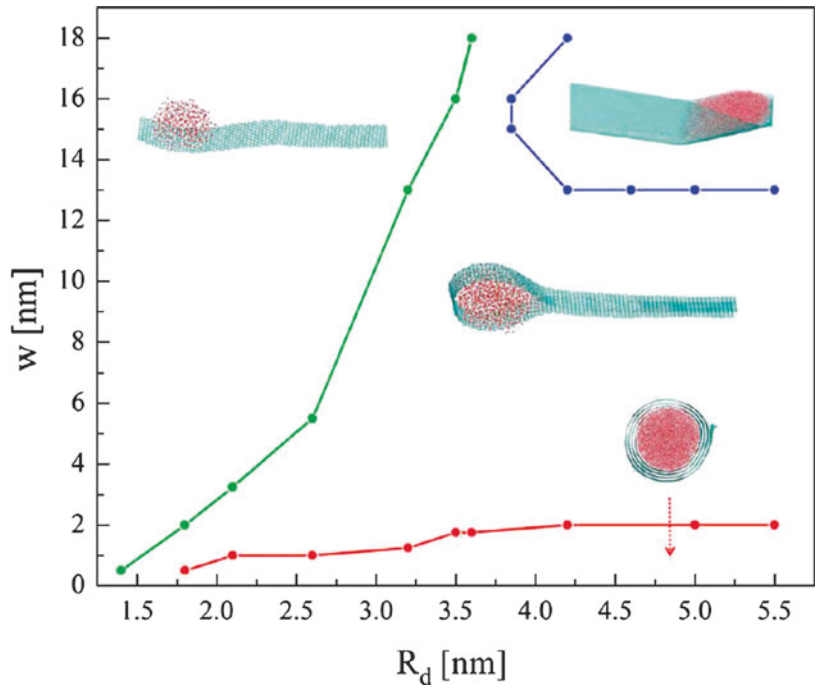
Capillary Origami, Fig. 1 Folding criteria plotted from Eq. 1 assuming complete circular folding and neglecting the effect of gravity

Applications

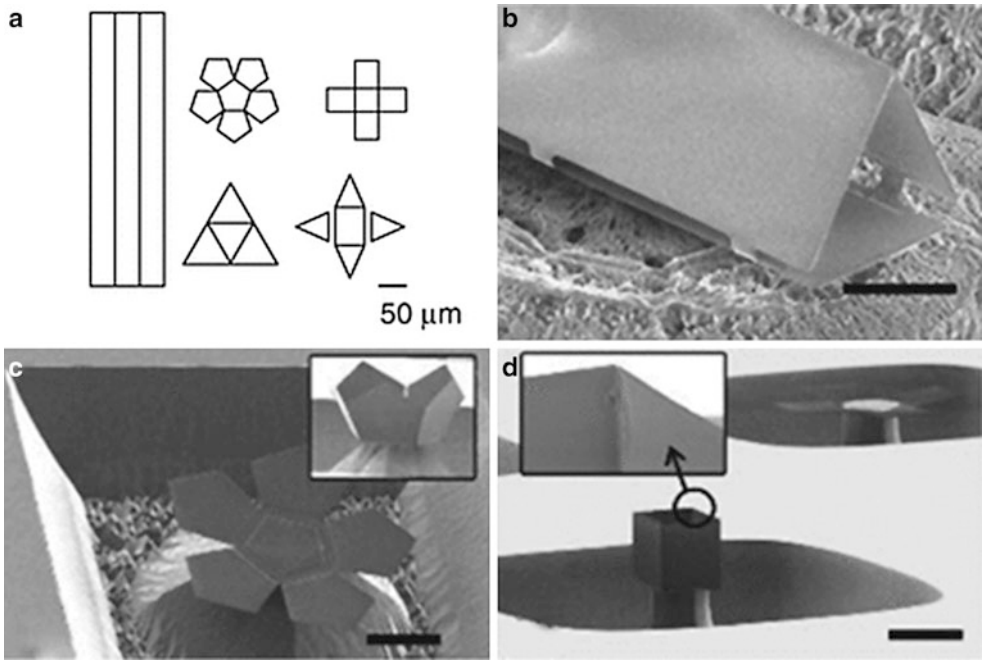
Capillary origami has been used to fabricate a number of 3D microstructures. Figure 4 illustrates the self-assembly of structures with various geometries. The initial planar templates are shown in Fig. 4a, and the folded final 3D microstructures are shown in Fig. 4b–d. The initial planar templates with lengths ranging from 50 to 100 μm and



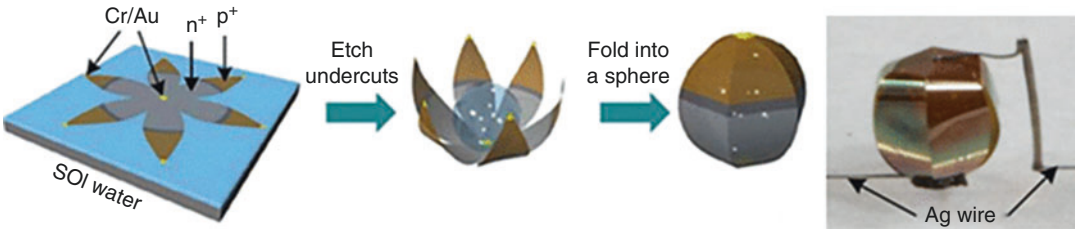
Capillary Origami, Fig. 2 Capillary origami 3D structures of a pyramid, a cube, and a quasi-sphere obtained by folding triangle-, cross-, and flower-shaped PDMS sheets, respectively, actuated with a water droplet (Reprinted with permission from Ref. [10]. Copyright 2007, American Institute of Physics)



Capillary Origami, Fig. 3 The phase diagram of a nanodroplet and graphene nanoribbon showing four different folding dynamics (Reprinted with permission from Ref. [11])



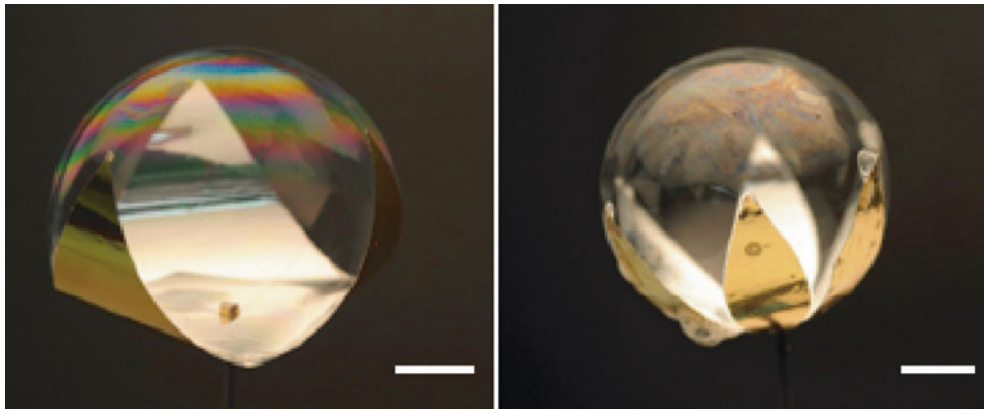
Capillary Origami, Fig. 4 (**a**) Schematics of initial templates. (**b**)–(**d**) SEM images of 3D microstructures after folding (scale bar: 50 µm) (Reprinted with permission from Ref. [12]. Copyright 2010, American Institute of Physics)



Capillary Origami, Fig. 5 Three schematics from left to right showing steps to fabricate a spherical-shaped silicon solar cell. The image at the far right shows the final spherical-shaped silicon solar cell (Images reprinted from Ref. [13] with permission)

a thickness of 1 µm were fabricated from silicon nitride thin films deposited and patterned by using standard micromachining processing typically used for integrated circuit and MEMS fabrication. Water droplets then were deposited on the templates to fold 3D microstructures [12]. Figure 5 shows another example of microfabrication of a quasi-spherical silicon solar cell based on capillary origami. After fabrication by conventional micromachining processing, the initial flower-shaped silicon template was folded into a sphere using a water droplet. Unlike conventional flat solar cells, this spherical solar cell enhanced light trapping and served as a passive tracking optical device, absorbing light from a wide range of incident angles [13].

Besides using liquid droplets, capillary origami structures can be constructed by using soap



Capillary Origami, Fig. 6 Capillary origami 3D structures formed from triangle- and flower-shaped templates using soap bubbles (scale bar: 2 cm) (Images reprinted from Ref. [14] with permission)

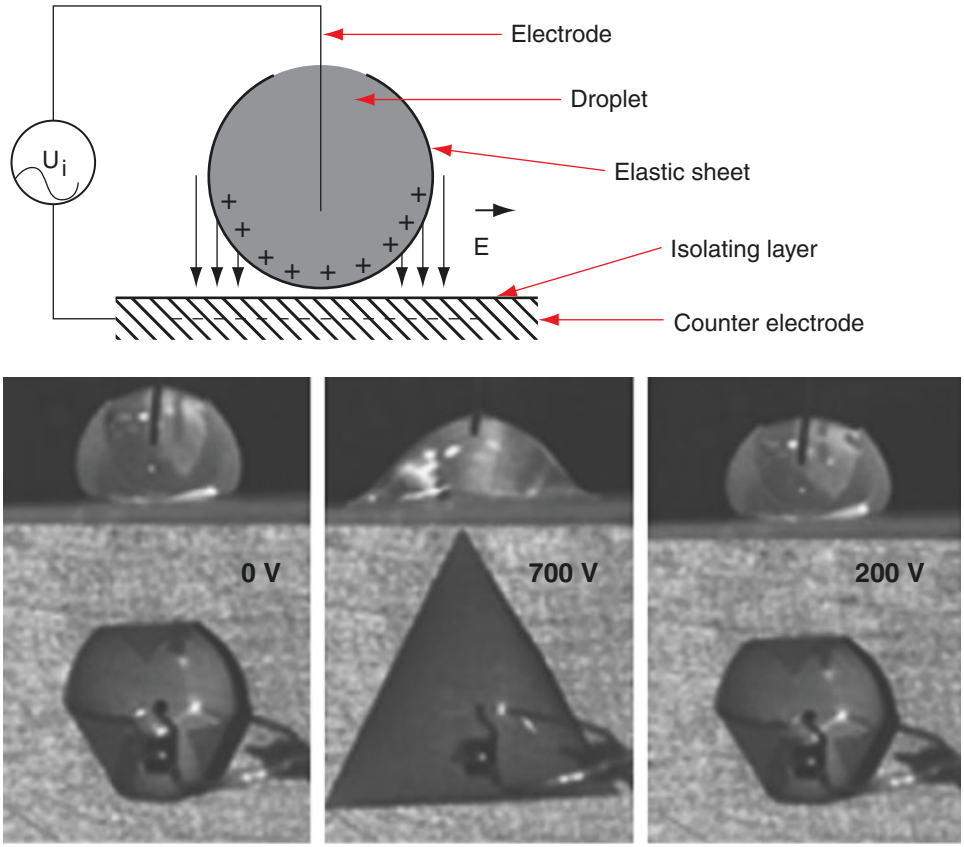
Capillary Origami, Fig. 7 The folding of an artificial flower when submerged in water (Reprinted with permission from Ref. [15]. Copyright 2009, American Institute of Physics)

bubbles as shown in Fig. 6. The weight of a soap bubble is much less than that of a liquid droplet especially for large droplets capable of covering centimeter-size structures when gravitational forces become significant. A soap bubble was shown to fold a centimeter-size elastic structure, which cannot be accomplished using a liquid droplet [14].

Petals of a flower also can be folded into a structure similar to capillary origami when submerged in water as shown in Fig. 7. The folding of the flower in water is accomplished by the interplay of elastic, capillary, and hydrostatic forces. During submersion, hydrostatic pressure pushes against the back of petals, and surface tension prevents water from penetrating through the spacing between petals resulting in trapping an air bubble inside a flower. The inside of the folded flower remains dry protected by the air bubble [15].

Structures formed by capillary origami also can be actuated by using electrostatic fields to reversibly fold and unfold then. For this application, we need to take into account the interplay of capillary, elastic, and electrostatic forces. As shown in Fig. 8, an electric field was applied between the droplet and the substrate. When the voltage was increased, the electrostatic force increased eventually overcoming capillary forces resulting in unfolding of the PDMS sheet. When the voltage was decreased below a certain threshold, the electrostatic force was no longer strong enough to prevent capillary forces from again folding the elastic sheet [16].

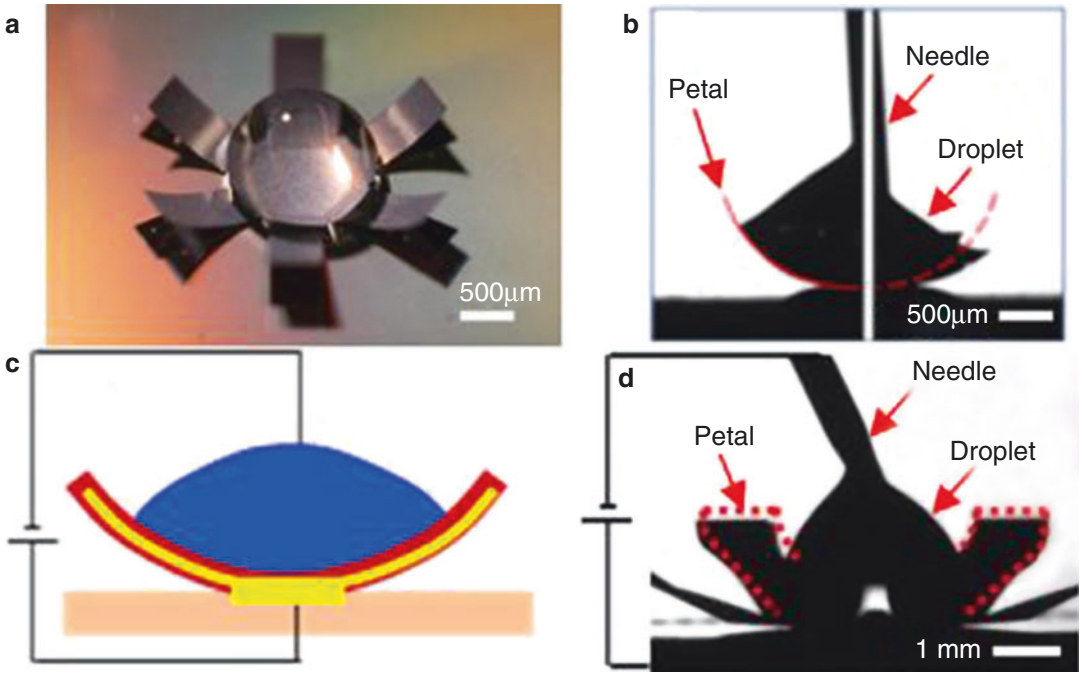
For flower-shaped polycrystalline silicon microstructures (called microflowers), the degree



Capillary Origami, Fig. 8 Capillary origami controlled by an electric field. The schematic of the experimental setup is shown at the far left, and three images to the right show the results of increasing voltage from 0 V to 700 V and decreasing voltage from 700 V to 200 V (Images from Ref. [16] – reproduced by permission of The Royal Society of Chemistry)

of folding was dynamically controlled by two different methods as shown in Fig. 9 [17]. The change in the bending angle was achieved by changing the volume of a liquid droplet by using an automated syringe pump. The petals reversibly bent and relaxed as the liquid was added and withdrawn. The electrowetting process (Fig. 9c) was also used to control bending. Capillary forces exerted by the liquid droplet on the petals can be changed by changing its contact angle by applying a voltage to the conducting liquid droplet placed on the petals. A thin dielectric layer of SiO_2 was thermally grown on the conductive Si petals to fabricate the electrowetting device. Both methods were used to reversibly actuate the petals without completely wrapping around the droplet, leaving part of the petal surface free to reflect incident light. The direction of the incoming light can then be dynamically controlled by the bending of the petal demonstrating an example of an optofluidic device.

Capillary origami is a simple and inexpensive method to fabricate 3D structures at the submillimeter scale. By using capillary forces, intricate and delicate 3D thin-film structures can easily be fabricated, which would be difficult to obtain by other methods. More applications exploiting the advantages of capillary origami itself or in combination with electric fields can readily be envisioned. Ultimately one expects to see more commercial products based on this versatile technique.



Capillary Origami, Fig. 9 (**a**) A microflower with a captured microdroplet. (**b**) Dependence of petal angle on liquid volume (only half of a frame is shown to facilitate comparison). As the amount of liquid captured by a microflower decreases, the petal angle also decreases. (**c**) Schematics of a process showing a droplet on the dielectric layer surrounding the conductive polycrystalline silicon petal of the microflower. (**d**) Electrowetting actuation of a microflower. The image shows petal position without applied voltage, and the red dash lines indicate petal positions with the voltage applied (Reprinted with permission from Ref. [17]. Copyright 2011, American Institute of Physics)

Cross-References

- ▶ Electrowetting
- ▶ Self-Assembly for Heterogeneous Integration of Microsystems
- ▶ Surface Tension Effects of Nanostructures

References

1. Py, C., Reverdy, P., Doppler, L., Bico, J., Roman, B., Baroud, C.N.: Capillary origami: spontaneous wrapping of a droplet with an elastic sheet. Phys. Rev. Lett. **98**, 156103 (2007)
2. Syms, R.R.A., Yeatman, E.M.: Self-assembly of three-dimensional microstructures using rotation by surface tension forces. Electron. Lett. **29**, 662–664 (1993)
3. Syms, R.R.A.: Surface tension powered self-assembly of 3-D micro-optomechanical structures. J. Microelectromech. Syst. **8**, 448–455 (1999)
4. Syms, R.R.A., Yeatman, E.M., Bright, V.M., Whitesides, G.M.: Surface tension-powered self-assembly of microstructures – the state-of-the-art. J. Microelectromech. Syst. **12**, 387–417 (2003)
5. Berthier, J.: Microdrops and Digital Microfluids. William Andrew Pub, New York (2008)
6. Timoshenko, S., Woinowsky-Krieger, S.: Theory of Plates and Shells, 2nd edn. McGraw-Hill, New York (1959)
7. de Langre, E., Baroud, C.N., Reverdy, P.: Energy criteria for elasto-capillary wrapping. J. Fluids Struct. **26**, 205–217 (2010)
8. Bico, J., Roman, B., Moulin, L., Boudaoud, A.: Adhesion: elastocapillary coalescence in wet hair. Nature **432**, 690 (2004)
9. Py, C., Reverdy, P., Doppler, L., Bico, J., Roman, B., Baroud, C.N.: Capillarity induced folding of elastic sheets. Eur. Phys. J. Spec. Top. **166**, 67–71 (2009)
10. Py, C., Reverdy, P., Doppler, L., Bico, J., Roman, B., Baroud, C.: Capillary origami. Phys. Fluids **19**, 091104 (2007)
11. Patra, N., Wang, B., Kral, P.: Nanodroplet activated and guided folding of graphene nanostructures. Nano Lett. **9**, 3766–3771 (2009)
12. van Honschoten, J.W., Berenschot, J.W., Ondarcuhu, T., Sanders, R.G.P., Sundaram, J., Elwenspoek, M., Tas, N.R.: Elastocapillary fabrication of

three-dimensional microstructures. Appl. Phys. Lett. **97**, 014103 (2010)
13. Guo, X., Li, H., Yeop Ahn, B., Duoss, E.B., Jimmy Hsia, K., Lewis, J.A., Nuzzo, R.G.: Two- and three-dimensional folding of thin film single-crystalline silicon for photovoltaic power applications. Proc. Natl. Acad. Sci. U. S. A. **106**, 20149–20154 (2009)
14. Roman, J., Bico, J.: Elasto-capillarity: deforming an elastic structure with a liquid droplet. J. Phys. Condens. Matter **22**, 493101 (2010)
15. Jung, S., Reis, P.M., James, J., Clanet, C., Bush, J.W. M.: Capillary origami in nature. Phys. Fluids **21**, 091110 (2009)
16. Pineirua, M., Bico, J., Roman, B.: Capillary origami controlled by an electric field. Soft Matter **6**, 4491–4496 (2010)
17. Manakasettharn, S., Taylor, J.A., Krupenkin, T.N.: Bio-inspired artificial iridophores based on capillary origami: fabrication and device characterization. Appl. Phys. Lett. **99**, 144102 (2011)

Carbon MEMS

Rodrigo Martinez-Duarte, Monsur Islam and Rucha Natu
Multiscale Manufacturing Laboratory, Department of Mechanical Engineering, Clemson University, Clemson, SC, USA

Synonyms

C-MEMS; Carbon microelectromechanical systems

Definition

A process used to fabricate miniaturized glass-like carbon structures through the carbonization of previously shaped organic precursors.

Overview

Carbon MEMS combines different patterning techniques with carbonization to derive miniaturized glass-like carbon features from an organic precursor. Since glass-like carbon is a brittle material, the carbonization of an already patterned precursor compares advantageously against subtractive manufacturing of glass-like carbon bulk. To date, the most common precursor patterning technique used in C-MEMS has been conventional photolithography, but many other techniques could be used. These cover a wide range: from relatively inexpensive like electrospinning, stamping, molding, polymer machining, and casting to next-generation lithography (NGL) techniques such as nanoimprint lithography (NIL) and electron beam lithography. The choice of precursor patterning technique is dictated by the quality, complexity, and final dimensions of the desired carbon part. Shrinkage of the precursor structure occurs during carbonization and must be accounted for. However, the existence of commercial high-quality precursors, like the negative photoresist SU-8, and availability of standardized fabrication tools make the fabrication process and the dimensional control highly reproducible.

Historical Background

Carbon can be found as a number of allotropes including diamond, lonsdaleite, buckminsterfullerenes, graphene, carbyne, graphite, carbon nanofoams, diamond-like carbon, amorphous carbon, and those carbons derived from the pyrolysis of organic materials, better known as glass-like carbons. Different precursors to obtain glass-like carbon may be used, including phenolic resins, polyfurfuryl alcohols, cellulose, polyvinyl chloride, and polyimides.

The first documented modern derivation of glass-like carbon from an organic resin (phenolic in this case) appears to be from 1915, when Weintraub and Miller in Massachusetts, USA, derived disks of a "very bright, shiny looking carbon with hardness equal or greater than 6 on the Mohs mineral scale." These disks were then used for better microphones in telephone transmitters. Their fabrication protocol featured slow heating of a hardened resin to a temperature close to 700 °C in about 1 week, followed by firing at temperatures from 800 to 1100 °C in just a few hours. The advantage of patterning an organic precursor, rather than carbon bulk, was realized

since then [1]. A sustained flow of publications on glass-like carbon did not begin until 1962, when Davidson, at the General Electric Co. in Kent, England, derived glass-like carbon from cellulose [2], and Yamada and Sato, at the Tokai Electrode Manufacturing Co. in Nagoya, Japan, published preliminary characterization results of carbon derived from organic polymers [3]. They named such carbon "glassy carbon," a term that was later registered as a trademark. Around the same time, in 1963, Lewis, Redfern, and Cowlard at the Plessey Company in the UK postulated the use of glass-like carbon, named "vitreous carbon" by the authors and later registered as trademark as well, as an ideal crucible material for semiconductors. Later that year, Redfern and Greens disclosed several production processes to derive this "vitreous carbon" in a patent [4]. In 1965, the advantages of glass-like carbon electrodes for voltammetry and analytical chemistry were characterized by Zittel and Miller, from Oak Ridge National Laboratory in the USA, using "glassy carbon" from the Tokai Electrode Manufacturing Co. [5]. In 1967, Cowlard and Lewis published a detailed description of the properties of "vitreous carbon," the fabrication process and its potential applications [6]. The decade of 1970 brought a significant interest on the use of glassy carbon as a material for different implants and biomedical instrumentation and also witnessed an explosion of the interest on glassy carbon by the analytical and electrochemistry communities, which still remains strong. In 1971, a structural model for glass-like carbon was postulated by Jenkins and Kawamura [7]. This model is up to this date the only one capable of explaining most of the experimental results obtained with glass-like carbon.

Carbon derived from organic polymers by pyrolysis in inert atmosphere has been historically known by three different names: "vitreous carbon," "glassy carbon," and glass-like carbon. Although highly referenced in implant-related publications during the 1970s, the term "vitreous carbon" started to fall in disuse by the end of that decade. "Vitreous carbon" is now better identified with reticulated vitreous carbon (RVC), a material introduced in the late 1970s by Chemotronics International Inc. from Ann Arbor, Michigan. The commercialization of Tokai's "glassy carbon" electrodes targeting the electrochemistry market made "glassy carbon" the term of preference for the electroanalytical chemistry community to refer to glass-like carbon. In 1995, the IUPAC (International Union of Pure and Applied Chemistry) defined glass-like carbon as the material derived by the pyrolysis of organic polymers and recommended that the terms "glassy carbon" and "vitreous carbon," which had been introduced as trademarks, should not be used as synonymous for glass-like carbon.

From the microfabrication standpoint, in 1983 Lyons et al. at AT&T Bell Laboratories published their work on the use of photodefined novolac resist patterns as precursors for carbon microstructures [8]. The drive behind this effort was finding an alternative to carbon films deposited by chemical vapor deposition. Miniaturized glass-like carbon 3D structures were not reported until the late 1990s by Schueller and coworkers at Harvard University. In their process, polydimethylsiloxane (PDMS) molds were fabricated using soft lithography and then used to pattern furfuryl alcohol-modified phenolic resins and phenol-formaldehyde resins, which were subsequently carbonized [9]. By 2000, Kostecki and colleagues were obtaining further results on the fabrication of planar carbon microelectrodes from positive photoresists to study the influence of their geometry in their electrochemical response [10]. Such work ignited the use of pyrolyzed photoresist films, or PPF, in fields such as electrochemistry. In 2002, the derivation of carbon from negative photoresists was reported by Singh et al. [11]. The obtained carbon showed higher electrical resistivity and vertical shrinkage than the one synthesized from positive resists. In 2004, structures with aspect ratios higher than 10 were reported by Wang et al. at the University of California, Irvine (UCI) [12, 13]. These authors coined the term Carbon MEMS. Since then, most of the work identified as Carbon MEMS has used SU-8 as carbon precursors and photolithography to shape 3D structures. Other precursors include resorcinol–formaldehyde gels [14] and cellulose. Starting in 2009, electrospinning of SU-8 and PAN (polyacrylonitrile) has been developed toward obtaining carbon fibers in the nanoscale [15, 16].

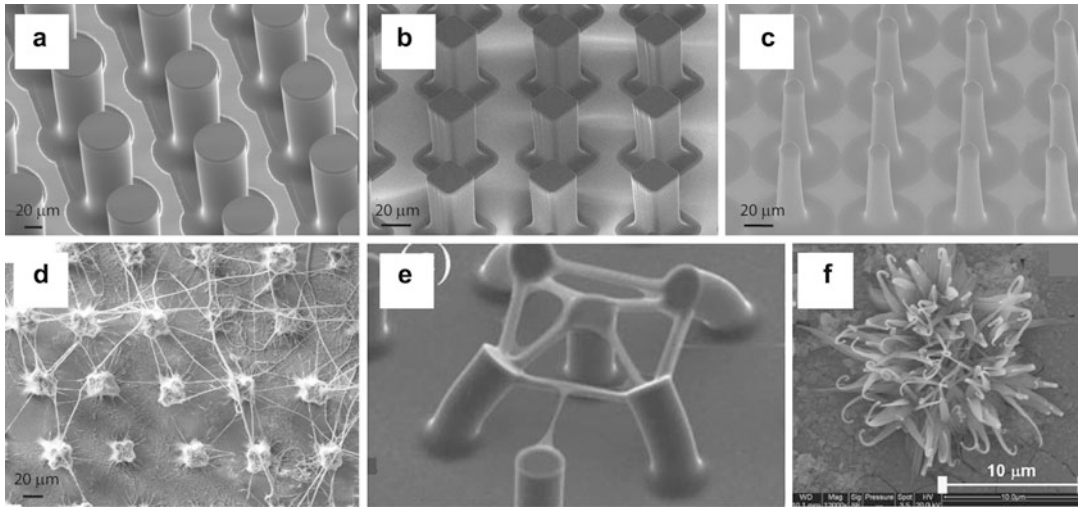
Material Properties of Carbon MEMS

The carbon obtained in a Carbon MEMS process resembles glass-like carbon. Although the choice of precursor and patterning technique has an impact on the properties of the resultant carbon, the following properties can serve as a starting reference. Although a consensus on the crystalline structure of glass-like carbon has not been reached to this date, the most widely accepted model [7] is the one that considers this type of carbon as made up of tangled and wrinkled aromatic ribbon molecules that are randomly cross-linked by carbon–carbon covalent bonds. The ribbon molecules form a networked structure, the unit of which is a stack of high strained aromatic ribbon molecules. Such structure of crystallites reflects the features of thermosetting resins structure which are commonly used as precursors for glass-like carbons. Glass-like carbon is widely considered as impermeable to gases and extremely inert, with a remarkable resistance to chemical attack from strong acids and other corrosive agents such as bromine. The material will react with oxygen at temperatures above few hundred degrees centigrade, making oxygen plasma a favored etching process. Glass-like carbon has a hardness of 6–7 on Mohs scale, a value comparable to that of quartz. Its density ranges from 1.4 to about 1.5 g cm^{-3}, compared to 2.3 g cm^{-3} for graphite. X-ray diffraction studies have shown that glass-like carbon presents an extremely small pore size, ~50 Å, of a closed nature, and that has an amorphous structure [31–38]. Glass-like carbon features a coefficient of thermal expansion of 2.2–3.2 × 10^{-6} K^{-1} which is similar to some borosilicate glasses. Its Young's modulus can vary between 10 and 40 GPa. An indicative electrical resistivity of glass-like carbon is 1 × 10^{-4} Ω m when the carbonization temperature is as high as 900 °C. The electrical resistivity drops further as the temperature increases over this value or the dwell time at temperatures below it increase [17, 18]. One of the reasons glass-like carbon is preferred by electrochemists is the fact that it has a wider electrochemical stability window than platinum and gold.

Photolithography as Toolbox for Carbon MEMS

The embracing of photolithography to fabricate the precursor photoresist structure has brought significant advantages and enabled rapid and notable developments. Photolithography refers to patterning with light, since the chemical composition of the photoresist changes upon being exposed to light of specific wavelengths, usually 365 nm for SU-8. In SU-8, light starts a cross-linking reaction in the matrix that makes the exposed section less soluble in a developer. The photolithography mask, a patterned stencil that enables the selective pass of light, must then be designed accordingly to the desired final topography: transparent areas in the mask must correspond to the topography to be fabricated. Traditional photolithography of a negative-tone photoresist involves a set of basic processing steps: spin coating a photoresist on a substrate, soft bake to evaporate the casting solvent, exposure through the mask to initiate cross-linking, postexposure bake to finalize cross-linking, and developing to obtain the exposed topography. Optimization of specific steps in traditional photolithography protocols has enabled the fabrication of a number of structures detailed below. The use of electrospinning has also emerged as an alternative to spin coating to deposit nanoscaled polymer fibers on a substrate, either as bulk or as single fibers (Fig. 1).

The fabrication of high-aspect ratio carbon structures, like those shown in Fig. 1a, b, was the first milestone of Carbon MEMS technology. SU-8 was an ideal choice of precursor to achieve this, given its low absorbance of light at wavelengths above 360 nm and the possibility to implement very thick layers in a single spin-coating step. Multistep photolithography has been used to fabricate dense arrays of carbon electrodes of different shapes to be used for capacitors [20, 21], heaters [22], glucose sensors [23], batteries [24–27], bioparticle manipulation using electrokinetics [28–36], and cell scaffolding [37]. By varying the exposure setup and dose, wires and suspended plates can also



Carbon MEMS, Fig. 1 (**a**) High-aspect ratio carbon electrodes on planar connecting leads. SU-8 precursor structure was fabricated using two-layer photolithography. (**b**) Further example of high-aspect ratio carbon structures. (**c**) Tapered carbon structures fabricated using backexposure during photolithography. (**d**) Example of SU-8 wires suspended between carbon posts. The fibers were made with electrospinning and later carbonized (Reprinted from [16] with permission from Elsevier). (**e**) Wire arrays made using electron beam lithography. Reprinted from [19] with permission from Elsevier. (**f**) Example of a fractal structure after carbonizing a sol–gel polycondensation of resorcinol and formaldehyde (Reprinted from [14], with permission from Elsevier)

be manufactured. The implementation of a photolithography process on a releasable substrate or on a sacrificial layer leads to free-standing SU-8 structures which were used as micromolds after carbonization [38]. Only the carbonization of free-standing SU-8 structures leads to a true isometric shrinkage. Carbonization of precursor structures anchored to a substrate yields carbon structures with a slight distortion on their base. UV backexposure of an SU-8 layer through a transparent substrate, fused silica, for example, leads to the development of tapered structures resembling cones (Fig. 1). Exposure dose can be tailored to specify the taper of the angle [39]. SU-8 wires suspended between two posts were made in 2006 using electron beam lithography. After carbonization, a number of carbon wire arrays were obtained (Fig. 1e) [19]. Suspended carbon wires like those shown in Fig. 1d have also been made using electrospinning in recent years [16, 40–42]. Electrospinning was also used to deposit a fiber mat on a substrate which was then patterned using photolithography. The fibrous structures were then carbonized to obtain highly porous carbon structures [15].

Other methods to shape the precursor exist. For example, in 2009, carbon fractals and microspheres were obtained by pyrolyzing a sol–gel polycondensation of resorcinol with formaldehyde [14]. The amount of surfactant during sol–gel synthesis determined the size of the spheres and the eventual switch to fabricating fractal structures (Fig. 1f). Micromolding of parylene has also been used to fabricate high-aspect ratio carbon posts [43].

Shrinkage

Carbonization is the process by which solid residues with a high content of carbon are obtained from organic materials, usually by pyrolyzing them in an inert atmosphere. The degree of shrinkage and carbon yield, the ratio of the weight of carbon to the weight of the original polymer sample, varies depending on the choice

of carbon precursor, its degree of cross-linkage, and its shape. For example, SU-8 features a carbon yield of 35–60 % [39]. The variability seems to depend on the shape of the precursor micro- or nanostructure.

In the case of Carbon MEMS, carbonization usually takes place in a furnace under a nitrogen, vacuum of forming gas atmosphere under a flow around 2000 ml/min. The carbonization protocol features three stages: (1) a temperature ramp from room temperature to 200–300 °C at 10 °C/min, followed by a 30 min dwell (this step is to completely eliminate the solvent and allow for any residual oxygen to be evacuated from the chamber and prevent combustion of the polymer as the temperature is raised further), (2) a temperature ramp to 900–1000 °C at 10 °C/min with a 1–4 h dwell, and (3) a natural cooldown to room temperature by turning the furnace off. Carbonization is a complex process with many reactions taking place concurrently, including dehydrogenation, condensation, hydrogen transfer, and isomerization. The pyrolysis process of organic compounds can be divided into three major steps: (1) pre-carbonization, (2) carbonization, and (3) annealing. During pre-carbonization ($T < 300$ °C), molecules of solvent and unreacted monomer are eliminated from the polymeric precursor. The next step, carbonization, can be further divided into two substages: (a) from 300 to 500 °C, when heteroatoms such as oxygen and halogens are eliminated causing a rapid loss of mass while a network of conjugated carbon systems is formed and hydrogen atoms start being eliminated, and (b) from 500 to 1200 °C, where hydrogen, oxygen, and nitrogen atoms are almost completely eliminated and the aromatic network is forced to become interconnected. At this point, permeability decreases and density, hardness, Young's modulus, and electrical conductivity increase. The carbon content of the structures carbonized at 900 °C is expected to exceed a mass fraction of 90 % in weight. At $T \sim 1300$ °C, more than 99 % carbon can be found. Annealing is usually carried out at temperatures above 1200 °C to allow for the gradual elimination of any structural defects and evolution of any further impurities.

Applications

Several applications have been demonstrated using Carbon MEMS. Some examples include the development of 3D architectures for lithium-ion batteries [24, 25], carbon plates in PEM fuel cells derived from machined polyimide [44, 45], post arrays which were electrochemically activated or decorated with carbon nanotubes to achieve supercapacitors [20, 21], and the development of carbon microspheres and fractal-like structures for sensors and batteries [14]. Other sensor applications include post arrays for rapid quantification of glucose in low concentrations [23] and interdigitated arrays with very narrow gaps to implement redox-based detection of dopamine with an amplification factor up to 25 [46]. C-MEMS devices have also been used as substratum for cell growth [47] and as carbon scaffolds to induce stem cell differentiation [37]. Extended arrays of carbon posts have been integrated in flow-through microfluidic devices to implement high-throughput cell separation and manipulation using electric fields in a technique now known as carbon-electrode dielectrophoresis [28, 29, 48]. This technique has been used to purify viable bacteria from an antibiotic-treated sample [31], to implement a sample preparation step that increases the sensitivity of traditional PCR protocols [34], and to extract DNA from a sample [32]. The combination between centrifugal microfluidics and carbon-electrode DEP marked an important step toward a sample-to-answer diagnostic platform [30]. Particle transport using electric fields has also been implemented [36], as well as high-throughput electrical lysis of different cells [33]. Other applications include the use of Carbon MEMS techniques to fabricate robust, inexpensive carbon shapes for the micromolding of bulk metallic glasses [38, 49].

Future Directions of the Field

Although most of the Carbon MEMS work has been done using SU-8 photoresist and photolithography, there is a strong interest on using different precursors, including biopolymers [50].

Development of techniques for shaping these precursors in the micro- and nanoscale will be needed, i.e., extrusion-based additive manufacturing. Electrospinning has been gaining importance as a technique to fabricate precursor nanostructures and is expected to be developed further. A broader use of composites is also expected: the patterning of SU-8-CNTs composites has been published [21] and preliminary work has also been disclosed using SU-8- silver and SU-8-silica composites. The use of catalysts to lower the energy required for carbonization is also of high interest, as well as understanding the shrinking process.

References

1. Weintraub, E., Miller, L.B.: Microphone. US Patent 1,156,509 (1915)
2. Davidson, H.W.: The properties of G.E.C. Impermeable carbon. Nucl. Eng. **7**, 159–161 (1962)
3. Yamada, S., Sato, H.: Some physical properties of glassy carbon. Nature **193**, 261–262 (1962)
4. Redfern, B., Greens, N.: Bodies and shapes of carbonaceous materials and processes for their production. US Patent 3,109,712 (1963)
5. Zittel, H.E., Miller, F.J.: A glassy-carbon electrode for voltammetry. Anal. Chem. **37**, 200–203 (1965)
6. Cowlard, F.C., Lewis, J.C.: Vitreous carbon – a new form of carbon. J. Mater. Sci. **2**, 507–512 (1967)
7. Jenkins, G., Kawamura, K.: Structure of glassy carbon. Nature **231**, 175–176 (1971)
8. Lyons, A., Wilkins, C., Robbins, M.: Thin pinhole-free carbon films. Thin Solid Films **103**, 333–341 (1983)
9. Schueller, O.J.A., Brittain, S.T., Whitesides, G.M.: Fabrication of glassy carbon microstructures by pyrolysis of microfabricated polymeric precursors. Adv. Mater. **9**, 477–480 (1997)
10. Kostecki, R., Song, X.Y., Kinoshita, K.: Influence of geometry on the electrochemical response of carbon interdigitated microelectrodes. J. Electrochem. Soc. **147**, 1878–1881 (2000)
11. Singh, A., Jayaram, J., Madou, M., Akbar, S.: Pyrolysis of negative photoresists to fabricate carbon structures for microelectromechanical systems and electrochemical applications. J. Electrochem. Soc. **149**, E78–E83 (2002)
12. Wang, C., Taherabadi, L., Jia, G., Madou, M., Yeh, Y., Dunn, B.: C-MEMS for the manufacture of 3D microbatteries. Electrochem. Solid-State Lett. **7**, A435–A438 (2004)
13. Wang, C., Jia, G., Taherabadi, L.H., Madou, M.J.: A novel method for the fabrication of high-aspect ratio C-MEMS structures. J. Microelectromechan. Syst. **14**, 348–358 (2005)
14. Sharma, C.S., Kulkarni, M.M., Sharma, A., Madou, M.: Synthesis of carbon xerogel particles and fractal-like structures. Chem. Eng. Sci. **64**, 1536–1543 (2009)
15. Sharma, C.S., Sharma, A., Madou, M.: Multiscale carbon structures fabricated by direct micropatterning of electrospun mats of SU-8 photoresist nanofibers. Langmuir **26**, 2218–2222 (2010)
16. Sharma, C.S., Katepalli, H., Sharma, A., Madou, M.: Fabrication and electrical conductivity of suspended carbon nanofiber arrays. Carbon N. Y. **49**, 1727–1732 (2011)
17. Park, B.Y., Taherabadi, L., Wang, C., Zoval, J., Madou, M.J.: Electrical properties and shrinkage of carbonized photoresist films and the implications for carbon microelectromechanical systems devices in conductive media. J. Electrochem. Soc. **152**, J136–J143 (2005)
18. Mardegan, A., Kamath, R., Sharma, S., Scopece, P., Ugo, P., Madou, M.: Optimization of carbon electrodes derived from epoxy-based photoresist. J. Electrochem. Soc **160**, B132–B137 (2013)
19. Malladi, K., Wang, C., Madou, M.: Fabrication of suspended carbon microstructures by e-beam writer and pyrolysis. Carbon N. Y. **44**, 2602–2607 (2006)
20. Beidaghi, M., Chen, W., Wang, C.: Electrochemically activated carbon micro-electrode arrays for electrochemical micro-capacitors. J. Power Sources **196**, 2403–2409 (2011)
21. Chen, W., Beidaghi, M., Penmatsa, V., Bechtold, K., Kumari, L., Li, W.Z., Wang, C.: Integration of carbon nanotubes to C-MEMS for on-chip supercapacitors. IEEE Trans. Nanotechnol. **9**, 734–740 (2010)
22. Jeong, O.C., Konishi, S.: Three-dimensionally combined carbonized polymer sensor and heater. Micromechanics Sect. Sensors Actuators (SAMM), based Contrib. Revis. from Tech. Dig. IEEE 20th Int. Conf. Micro Electro Mech. Syst. (MEMS 2007) - MEMS 2007, IEEE 20th Int. Conf. M. 143, 97–105 (2008).
23. Xu, H., Malladi, K., Wang, C., Kulinsky, L., Song, M., Madou, M.: Carbon post-microarrays for glucose sensors. Biosens. Bioelectron. **23**, 1637–1644 (2008)
24. Min, H.-S., Park, B.Y., Taherabadi, L., Wang, C., Yeh, Y., Zaouk, R., Madou, M.J., Dunn, B.: Fabrication and properties of a carbon/polypyrrole three-dimensional microbattery. J. Power Sources **178**, 795–800 (2008)
25. Teixidor, G.T., Zaouk, R.B., Park, B.Y., Madou, M.J.: Fabrication and characterization of three-dimensional carbon electrodes for lithium-ion batteries. J. Power Sources **183**, 730–740 (2008)
26. Wang, C., Madou, M.: From MEMS to NEMS with carbon. Biosens. Bioelectron. **20**, 2181–2187 (2005)
27. Wang, C., Taherabadi, L., Jia, G., Kassegne, S., Zoval, J., Madou, M.: Carbon-MEMS architectures for 3D microbatteries. Proc. SPIE **5455**, 295–302 (2004)

28. Martinez-Duarte, R., Renaud, P., Madou, M.: A novel approach to dielectrophoresis using carbon electrodes. Electrophoresis **32**, 2385–2392 (2011)
29. Jaramillo, M.D.C., Torrents, E., Martinez-Duarte, R., Madou, M.J., Juarez, A.: On-line separation of bacterial cells by carbon-electrode dielectrophoresis. Electrophoresis **31**, 2921–2928 (2010)
30. Martinez-Duarte, R., Gorkin, R.A., Abi-Samra, K., Madou, M.J.: The integration of 3D carbon-electrode dielectrophoresis on a CD-like centrifugal microfluidic platform. Lab Chip **10**, 1030–1043 (2010)
31. Elitas, M., Martinez-Duarte, R., Dhar, N., McKinney, J.D., Renaud, P.: Dielectrophoresis-based purification of antibiotic-treated bacterial subpopulations. Lab Chip **14**, 1850–1857 (2014)
32. Martinez-Duarte, R., Camacho-Alanis, F., Renaud, P., Ros, A.: Dielectrophoresis of lambda-DNA using 3D carbon electrodes. Electrophoresis **34**, 1113–1122 (2013)
33. Mernier, G., Martinez-Duarte, R., Lehal, R., Radtke, F., Renaud, P.: Very high throughput electrical cell lysis and extraction of intracellular compounds using 3D carbon electrodes in lab-on-a-chip devices. Micromachines **3**, 574–581 (2012)
34. Jaramillo, M.D.C., Martínez-Duarte, R., Hüttener, M., Renaud, P., Torrents, E., Juárez, A.: Increasing PCR sensitivity by removal of polymerase inhibitors in environmental samples by using dielectrophoresis. Biosens. Bioelectron. **43**, 297–303 (2013)
35. Martinez-Duarte, R.: Carbon-electrode dielectrophoresis for bioparticle manipulation. ECS Trans. **61**, 11–22 (2014)
36. Rouabah, H.A., Park, B.Y., Zaouk, R.B., Morgan, H., Madou, M.J., Green, N.G.: Design and fabrication of an ac-electro-osmosis micropump with 3D high-aspect-ratio electrodes using only SU-8. J. Micromech. Microeng. **21**, 035018 (2011)
37. Amato, L., Heiskanen, A., Caviglia, C., Shah, F., Zór, K., Skolimowski, M., Madou, M., Gammelgaard, L., Hansen, R., Seiz, E.G., Ramos, M., Moreno, T.R., Martínez-Serrano, A., Keller, S.S., Emnéus, J.: Pyrolysed 3D-carbon scaffolds induce spontaneous differentiation of human neural stem cells and facilitate real-time dopamine detection. Adv. Funct. Mater. **24**, 7042–7052 (2014)
38. Schroers, J., Kumar, G., Madou, M., Martinez-Duarte, R.: Carbon molds for use in the fabrication of bulk metallic glass parts and molds. US 2012/0125071 A1 (2012)
39. Martinez-Duarte, R.: SU-8 photolithography as a toolbox for carbon MEMS. Micromachines **5**, 766–782 (2014)
40. Maitra, T., Sharma, S., Srivastava, A., Cho, Y.-K., Madou, M., Sharma, A.: Improved graphitization and electrical conductivity of suspended carbon nanofibers derived from carbon nanotube/polyacrylonitrile composites by directed electrospinning. Carbon N. Y. **50**, 1753–1761 (2012)
41. Sharma, S., Sharma, A., Cho, Y.-K., Madou, M.: Increased graphitization in electrospun single suspended carbon nanowires integrated with carbon-MEMS and carbon-NEMS platforms. ACS Appl. Mater. Interfaces **4**, 34–39 (2012)
42. Canton, G., Do, T., Kulinsky, L., Madou, M.: Improved conductivity of suspended carbon fibers through integration of C-MEMS and electromechanical spinning technologies. Carbon N. Y. **71**, 338–342 (2014)
43. Naka, K., Konishi, S.: Micro and nano structures of carbonised polymer through pyrolytic transformation from polymer structures. Micro Nano Lett. **1**, 79 (2006)
44. Park, B.Y., Madou, M.J.: Design, fabrication, and initial testing of a miniature PEM fuel cell with micro-scale pyrolyzed carbon fluidic plates. J. Power Sources **162**, 369–379 (2006)
45. Lin, P.-C., Park, B.Y., Madou, M.J.: Development and characterization of a miniature PEM fuel cell stack with carbon bipolar plates. J. Power Sources **176**, 207–214 (2008)
46. Heo, J.I., Shim, D.S., Teixidor, G.T., Oh, S., Madou, M.J., Shin, H.: Carbon interdigitated array nanoelectrodes for electrochemical applications. J. Electrochem. Soc. **158**, J76 (2011)
47. Teixidor, G.T., Gorkin, R.A., Tripathi, P.P., Bisht, G. S., Kulkarni, M., Maiti, T.K., Battacharyya, T.K., Subramaniam, J.R., Sharma, A., Park, B.Y., Madou, M.: Carbon microelectromechanical systems as a substratum for cell growth. Biomed. Mater. **3**, 034116 (2008)
48. Martinez-Duarte, R.: Carbon-electrode dielectrophoresis for bioparticle manipulation. ECS Trans. **61**, 11–22 (2014)
49. Martinez-Duarte, R.: Fabrication of Carbon Micro Molds. University of California, Irvine (2009)
50. Islam, M., Martinez-Duarte, R.: Additive manufacturing of carbides using renewable resources. In: Proceedings of the ASME 2015 IMECE. In press, Houston (2015)

Carbon Microelectromechanical Systems

► Carbon MEMS

Carbon Nanotube (CNT) Arrays

► Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Carbon Nanotube Materials

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Carbon Nanotube NEMS

Max Zenghui Wang
Department of Electrical Engineering and Computer Science, Case School of Engineering, Case Western Reserve University, Cleveland, OH, USA

Synonyms

CNT NEMS; CNT resonator; Nanotube adsorption sensor; Nanotube force sensor; Nanotube mass sensor; Nanotube resonant sensor; Nanotube resonator

Definition

Carbon nanotube nanoelectromechanical systems (NEMS) are devices that combine the electrical and mechanical degrees of freedom in carbon nanotubes, the one-dimensional (1D) form of crystalline nanocarbon. It holds the record of the smallest NEMS made to date, with the part in motion being a single molecule – an individual carbon nanotube. The most common type of carbon nanotube NEMS device is carbon nanotube resonator, which generally assumes the geometry of a suspended carbon nanotube field-effect transistor (FET), with source and drain electrodes connecting both ends of the nanotube and a third nearby gate electrode forming a capacitor with the nanotube. The suspended segment of the nanotube can be actuated electrostatically by applying a bias voltage (DC, AC, or both) between the nanotube and the gate electrode, and the mechanical motion of the device can in turn be detected electrically by monitoring the electronic transport through the nanotube FET. Combining the outstanding electrical and mechanical properties of its building material, carbon nanotube NEMS devices exhibit, in several aspects, superior performance currently unavailable in NEMS based on other nanomaterials. In particular, leveraging the ultrasmall mass in motion, NEMS resonators based on carbon nanotubes have demonstrated extreme sensitivity to external stimuli (e.g., mass, force). In addition, the atomically perfect crystalline surface of carbon nanotube makes it a unique platform for studying surface physics processes. Together with the ultrahigh sensitivity to adsorbed mass, exotic phenomena have been observed using carbon nanotube NEMS, such as phase transitions in the pseudo-1D system formed by the adsorbed gas atomic layer on the surface of an individual carbon nanotube.

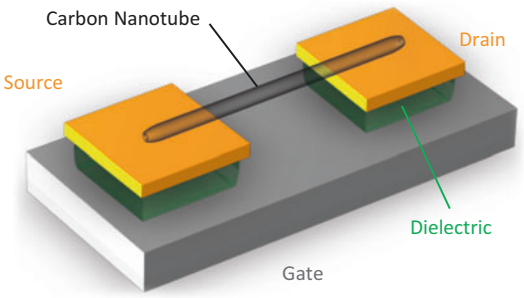
Introduction

Since its discovery, carbon nanotube (CNT) has been extensively studied for its outstanding physical properties. Structurally, CNT can be categorized into single-walled carbon nanotubes (SWCNT) and multi-walled carbon nanotubes (MWCNT). A SWCNT is a hollow cylinder whose entire surface (except the ends) is made of carbon atoms arranged in the graphitic hexagonal honeycomb lattice. It can be conceptually constructed by rolling a piece of graphene (a single layer of graphite) along one of the directions that allow the carbon atoms to seamlessly stitch to their counterparts on the opposite edge of the graphene sheet. Typical SWCNTs have diameter on the order of 1 nm. A MWCNT consists of multiple coaxial layers, with each layer being an individual SWCNT. Consequently, MWCNT generally has greater diameter and larger mass than SWCNT. Most carbon nanotube NEMS devices are constructed based on SWCNTs to take advantage of their small mass and diameter.

Nanoelectromechanical systems are devices that have both electronic and mechanical degrees of freedom, with one or more dimensions of the motional part being on the scale of 100 nm or less. The scaling down of the device dimensions can give rise to new physical phenomena and enhanced

functionalities. NEMS devices are often used for sensing, signal processing, and actuation at the nanoscale.

By having an individual carbon nanotube as the part in motion, CNT NEMS devices leverage the ultrasmall diameter/ultrahigh aspect ratio and ultralow mass of nanotubes, as well as their electronic properties (high current density, quantized and ballistic electronic transport, etc.) and mechanical (high flexibility, high elastic modulus, etc.) properties. While several NEMS based on multi-walled nanotubes utilize the interlayer motion to build linear [1] and rotational [2] bearings, the great majority of carbon nanotube NEMS devices are resonators [3–20] with the resonator body made entirely by a nanotube, with the first one [3] demonstrated in 2004. The rest of the article thus focuses on CNT NEMS resonators.



Carbon Nanotube NEMS, Fig. 1 Schematic of a carbon nanotube NEMS resonator/suspended-channel FET

Geometries/Structures

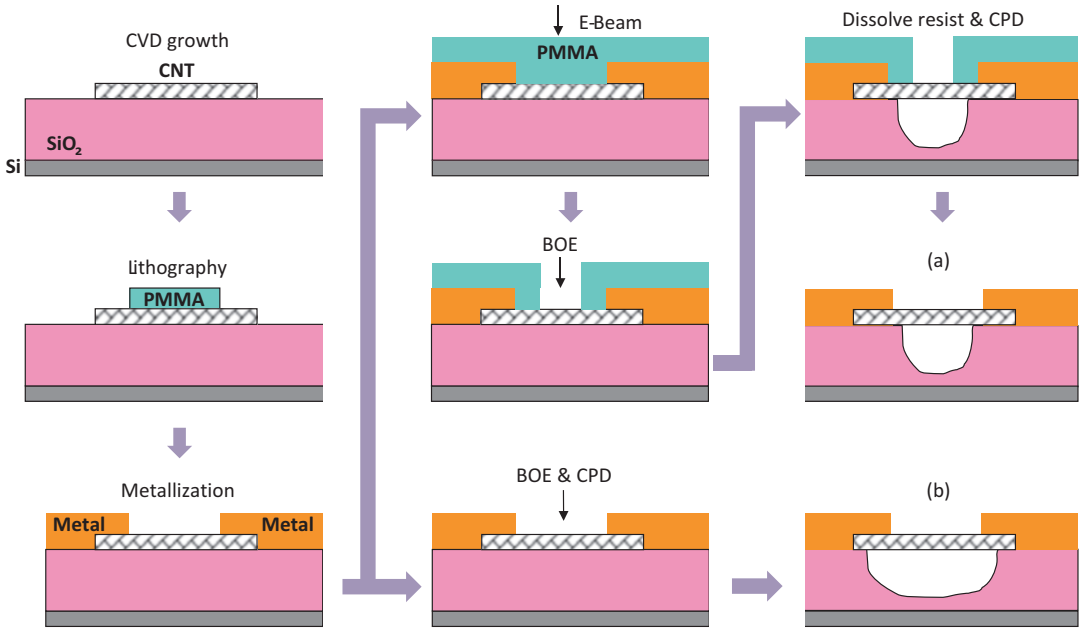
Most carbon nanotube NEMS resonators assume the geometries of a doubly clamped beam/string [3–18] and function as suspended-channel FETs (Fig. 1). The carbon nanotube is suspended over a microtrench and is mechanically anchored at both ends and electrically connected to the source and drain electrodes. In some cases, the electrodes themselves serve as the mechanical anchors for the nanotube (i.e., the nanotube is directly suspended between the two electrodes). The gate electrode is typically located underneath the nanotube (toward the bottom of the trench): a common configuration is to use the heavily doped Si back gate that extends the entire device die; alternatively, local gate at the bottom of the trench (through aligned deposition of metal layers) can also be used. Specially designed gate (e.g., side gate) has also been demonstrated in carbon nanotube NEMS [4].

A less common type of carbon nanotube NEMS takes the form of a singly clamped beam [19, 20]. To date, such structure has been realized using multi-walled carbon nanotubes (with the thinnest being double walled). Electrically, the nanotube is directly connected to one electrode (the clamp) and faces an opposing electrode which is not in contact with the nanotube. Additional gate electrodes along the length of the nanotube can also be used, even though they generally do not provide the canonical gating effect as in the FET structure (as the singly clamped structure does not facilitate electronic transport measurements). Compared with the doubly anchored suspended FET structure, the singly clamped beam structure presents some unique opportunities such as utilization of the field-emission properties of carbon nanotube, at the cost of forfeiting the access to carbon nanotube's electronic transport properties and the capability of electrostatically tuning the mechanical motion through gating, which are detailed in following sections.

Making Carbon Nanotube NEMS

There are a number of ways to fabricate doubly clamped carbon nanotube NEMS, which can be categorized into two main schemes: (i) grow/deposit carbon nanotube on the substrate first and then fabricate the suspended structure and (ii) fabricate the device structure first without carbon nanotube and then grow/transfer the nanotube onto the prefabricated structure.

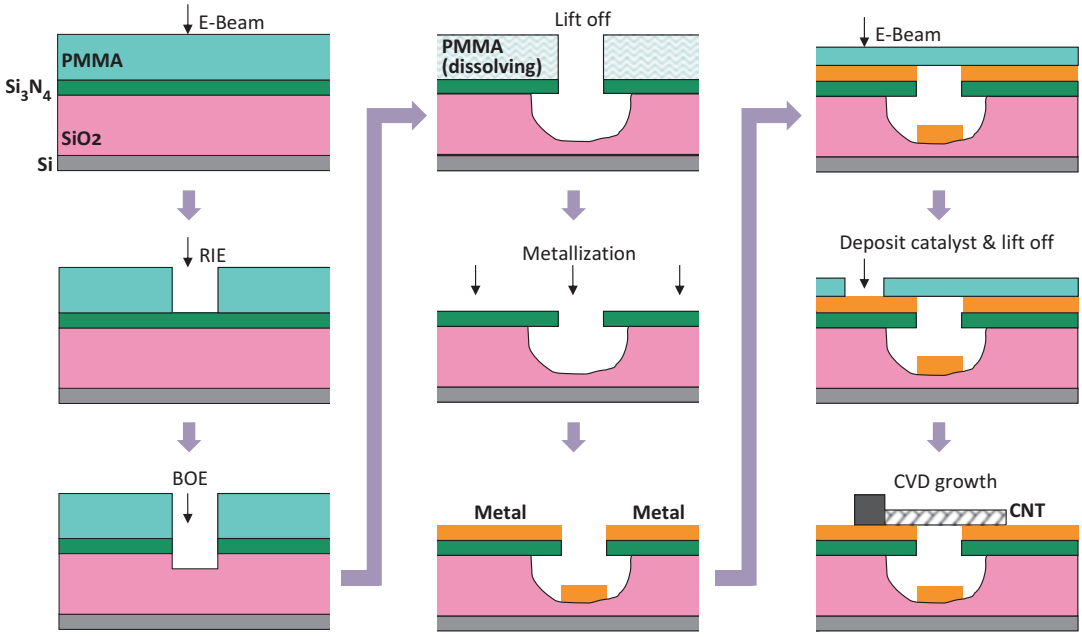
Typical process flow for making doubly clamped carbon nanotube NEMS resonator using scheme (i) is illustrated in Fig. 2. First, carbon nanotube is grown on a SiO_2-covered Si wafer using chemical vapor deposition (CVD). Metal electrodes are then patterned on the carbon nanotube (either aligned to a selective nanotube or



Carbon Nanotube NEMS, Fig. 2 Schematic of fabricating doubly clamped carbon nanotube NEMS resonator using scheme (i), with two major categories of resulting device geometry (**a**) and (**b**) shown

patterned randomly but in a large array) through photolithography, electron-beam lithography, or contact (stencil) mask followed by electron-beam or thermal evaporation. Non-suspended carbon nanotube FET is obtained upon completion of the above steps. To suspend the electrically contacted nanotubes, often a second layer of aligned lithography is performed to open etch windows at locations where suspended segment of carbon nanotube is desired. Such etch window may or may not enclose part of the electrodes (i.e., the inner edges of the electrodes may or may not be exposed in the etch window). In some cases, the metal contacts themselves are used as etch masks, and the second lithography step (for open etch windows) is omitted. Once the etch window is defined, an etch step is performed to remove the underlying oxide in the exposed area to suspend the carbon nanotube. Buffered oxide etch (BOE) is most commonly used, often followed by critical point drying (CPD) to prevent the nanotube from collapsing and sticking to the bottom of the etching-formed microtrench (due to the capillary force during the drying process), while vapor HF etch is a practical alternative.

Depending on the size and geometry of the etch window, two major categories of device geometry can result from this fabrication scheme: (a) the nanotube segment between the metal electrodes is partially suspended in the middle section and clamped to the SiO_2 substrate on both sides and (b) the entire nanotube segment between the metal electrodes is suspended, and the electrodes themselves, while also serving as mechanical clamps for the suspended nanotube, are also partially suspended. Each geometry has its own unique attributes. In the former case (a), the mechanical clamp through van der Waals interaction with the SiO_2 is often sufficiently strong and provides good mechanical clamping, but the electronic properties measured between the electrodes contain information from both suspended and unsuspended segments. In the latter case (b), the entire electronic signal comes from suspended nanotube, but the suspended metal layer can be floppy and exhibit its own resonances during measurement. Note that due to the nondirectional nature of the etch processes (BOE or vapor HF), it is not practical to obtain devices where the boundary of the etched microtrench exactly aligns



Carbon Nanotube NEMS, Fig. 3 Schematic of fabricating doubly clamped carbon nanotube NEMS resonator using scheme (ii)

with the edge of metal electrodes, a third geometry which presumably possesses the advantages of both geometries (a and b) mentioned above.

Typical fabrication process for building carbon nanotube NEMS using scheme (ii) is shown in Fig. 3. The first part is to fabricate the device structure without the nanotube. It can be generally considered as the procedure as in scheme (i) without the nanotube growth at the beginning. Nevertheless, due to the absence of carbon nanotube, a number of things can be done differently. First, directional etch (such as RIE) of SiO_2 can be used instead of BOE or vapor HF, which can result in vertical sidewalls of the microtrench which can be aligned to the edges of the metal electrodes (e.g., using the metal as an etch mask). Second, the choice of supporting material for the carbon nanotube is expanded. For example, a Si_3N_4 can be added on top of the SiO_2 to facilitate undercutting of SiO_2 (which can result in substantially high aspect ratio structures) while providing strong mechanical support to the otherwise floppy metal electrodes [5]. These features are desirable for creating suspended carbon nanotube NEMS in the following step. For example, the undercut will prevent the nanotube from sticking to the SiO_2 microtrench sidewall during growth, yielding a higher suspension rate; and the nitride support will prevent metal from collapsing during the high-temperature growth process while also serving as a stiff clamp once the nanotube is suspended between electrodes. Once the device structure is prepared, a final step is performed to deposit carbon nanotube across the microtrench (and between the electrodes), often through CVD growth [5, 6]. Other techniques, such as stamp transfer, can also be used [7].

Both fabrication schemes have their advantages and limitations. In scheme (i), individual nanotubes can be identified, even characterized (such as using electronic transport and AFM), before additional effort is invested to turn it into a NEMS device by suspending the nanotube. In addition, with proper procedures (such as CPD), the success rate for achieving individual suspended devices is reasonably high, while the effort in making each device is substantial (as devices are fabricated on individual bases). In contrast, in scheme (ii), the device structures can be prefabricated at wafer scale, but for CVD growth, it requires an additional lithography

step to pattern the catalyst (otherwise, the randomly grown nanotubes could electrically short the electrodes), and the yield of the growth step (chances of obtaining an individual suspended carbon nanotube across a microtrench and bridging a pair of electrodes) is far from unity. Nevertheless, the large-scale processes (patterning and growth) largely make up for the total device yield. While the CVD growth may impose limitations on the material choice for metal electrodes (to sustain during the high-temperature growth), the nanotube transfer technique [7] offers a versatile alternative, as long as the nanotube-metal contacts are properly treated (such as annealing or use freshly deposited metal) to ensure good electronic performance.

Besides the abovementioned differences in device structures and materials between the two fabrication schemes, another important contrast is the cleanness of the carbon nanotube. In scheme (i), the nanotube undergoes a number of wet processes and is directly exposed to various chemicals (resist, polymer, etchant, solvent, etc.), which could lead to contaminants and defects on the nanotube surface. However, this does not necessarily preclude the resulting carbon nanotube NEMS device to exhibit certain desired performances, such as ultrahigh sensitivity to admass. Consequently, atomic-level mass sensing has been demonstrated on carbon nanotube NEMS resonators fabricated using this scheme [8–10]. In contrast, in scheme (ii), the carbon nanotube does not undergo any additional processing and thus can maintain its as-grown atomically perfect surface. Therefore, devices fabricated using this scheme are used in applications where pristine nanotube surface is required, such as studying transport in ultraclean suspended nanotubes or using nanotube as a substrate for investigating monolayer gas adsorption on low-dimensional surfaces [5, 11].

Singly clamped nanotube NEMS structures are often made by mounting the carbon nanotube onto a predefined structure, which often involves steps that are not compatible with standard nanofabrication processes, such as attachment of nanotube inside SEM/TEM. The fabrication is often performed at the level of individual devices and is thus highly labor intensive.

Actuation of Device Motion

To date, most nanotube NEMS resonators are operated in vacuum (mTorr or below), with a few measured in low pressure (a few Torr). The most common actuation scheme for carbon nanotube NEMS is electrostatic actuation through a gate electrode, which prevails in the doubly clamped device structure. In this configuration, the nanotube and its gate form the two electrodes of a capacitor. When a voltage is applied across the capacitor (with 0 net electric charge), an attractive electrostatic force is generated between the two electrodes, with its magnitude equal to

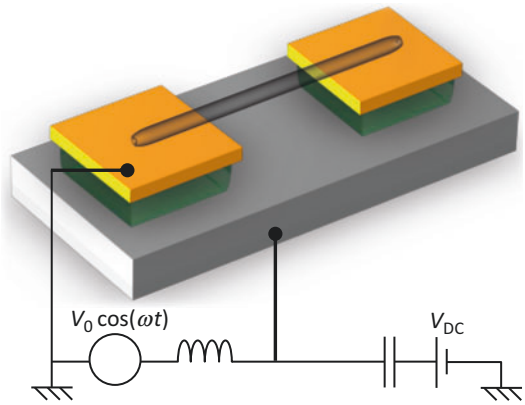
$$F = \frac{1}{2} C' V^2, \tag{1}$$

where C' is a derivative of the capacitance with respect to the distance between its two electrodes and V is applied voltage. Accordingly, the nanotube is attracted toward the gate when a finite gate voltage exists and is driven into oscillation when an AC voltage is applied. For carbon nanotube NEMS resonators, when the driving signal is at the frequency of the device's natural frequency, mechanical resonance with enhanced amplitude ensues.

It is worth noting that in most cases, a DC voltage is applied in addition to an AC voltage (Fig. 4), and there are two main reasons. First, Eq. 1 shows that the driving force is proportional to V^2. If a pure AC voltage $V = V_0\cos(\omega t)$ is applied, $F \propto V^2 \propto \cos^2(\omega t) = (1 + \cos(2\omega t))/2$, which has no $\cos(\omega t)$ term. This suggests that the device will be driven at 2ω instead; thus, the response will be more complicated than that of a simple harmonic resonator. Instead, when a DC component V_{DC} is included, $V = V_0\cos(\omega t) + V_{\mathrm{DC}}$, then

$$\begin{aligned} F \propto V^2 &= (V_0 \cos(\omega t) + V_{DC})^2 \\ &= V_{DC}^2 + 2V_{DC}V_0 \cos(\omega t) \\ &\quad + V_0^2 \cos^2(\omega t), \end{aligned} \tag{2}$$

which, in the case of $V_{\mathrm{DC}} \gg V_0$ (which is often the case), the driving force is mostly at 1ω. Second, Eq. 2 shows that there will be a static term ${V_{\mathrm{DC}}}^2$ in the force applied. Due to the ultra-flexibility of



Carbon Nanotube NEMS, Fig. 4 Electrostatic actuation of doubly clamped carbon nanotube NEMS resonators

carbon nanotube (in comparison with thicker devices with similar geometry, such as lithography-defined nanowires), this static force is capable of generating sufficient tension in the nanotube and significantly tuning its resonance frequency. It is not atypical for a resonance to be tuned by more than 200 % with just a few volts of DC gate voltage.

For singly clamped devices, a nearby electrode (similar to the gate electrode in the doubly clamped case) can be used to capacitively actuate the vibratory motion [19]. Resonant motion driven using electromagnetic wave has also been demonstrated [20]. Additionally, due to the ultrasmall mass of carbon nanotubes, it is possible to directly visualize their completely undriven thermomechanical resonances in electron microscopes as a position-dependent blurring of the nanotube image. This has been observed for both singly and doubly clamped devices.

Readout of Device Motion

Electronic readout prevails in doubly clamped carbon nanotube NEMS resonators. This readout scheme utilizes the fact that motion-induced on-tube charge fluctuation modulates the conductance of the CNT FET and thus generates an oscillation in the amplitude of the electronic current through the device. Specifically, the device displacement δz induces a change in the nanotube-gate capacitance $\delta C = C'\delta z$, which leads to an additional on-tube electric charge δQ, which further leads to a modulation in the device conductance δG. Specifically, the charge variation can be expressed as

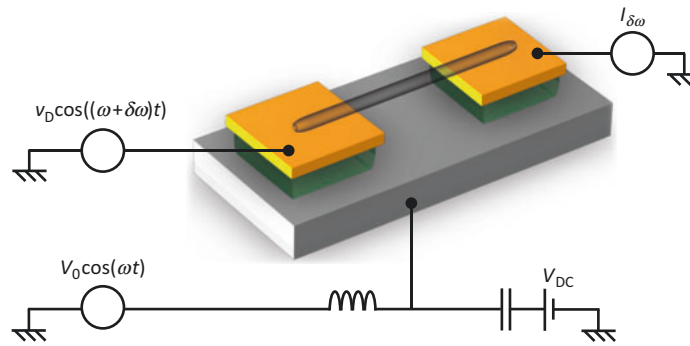
$$\delta Q = V_g \delta C + C \delta V_g. \quad (3)$$

The first term in Eq. 3 accounts for the motion-induced change in capacitance. When the device is mechanically on resonance, the motion amplitude increases substantially from the off-resonance background. Consequently, the first term surges at the device's resonance frequency. The second term represents the conventional gating effect in CNT FET devices and is present even in non-suspended devices. When the device is on resonance, the motion amplitude can be as large as several nm, while the vacuum gap to the gate is typically 100 s of nm. The change in gate capacitance (δC) is therefore only $2 \sim 3$ orders of magnitude smaller than the static gate capacitance of the nanotube, which is sometimes on the same order of magnitude as the $\delta V_g / V_g$ ratio. Consequently, the contribution of both terms in Eq. 3 ($V_g \delta C$ and $C \delta V_g$) can be on the same order of magnitude when the device is on resonance, and the mechanical resonance is often manifested as a sharp feature on top of a slowly varying background.

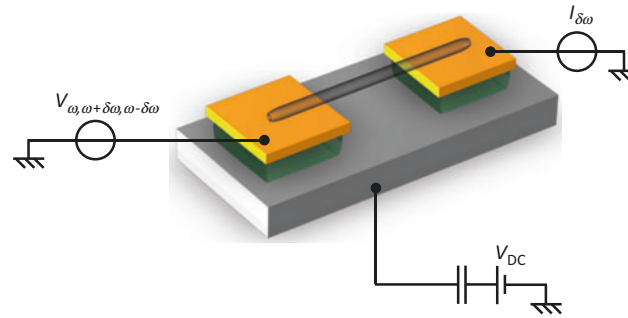
However, such current fluctuation – which is at the same frequency as the device motion – is often masked by the capacitive current generated by the electrostatic driving signal on the gate, which is also at the same frequency, but at a much larger amplitude, due to the large parasitic capacitance from the electrodes (including the wire bonding pads), which is typically several orders of magnitude larger than the nanotube's capacitance.

To overcome this challenge, frequency downmixing technique is often used to extract the motion-induced signal from the large capacitive background. There are two types of mixing techniques: one uses amplitude modulated (AM) signal and the other one uses frequency modulation (FM).

The basic operation principle of AM downmixing roots from a two-source mixing technique [3], which is illustrated in Fig. 5. In this setup, two



Carbon Nanotube NEMS, Fig. 5 Schematic of the two-source mixing setup for resonant motion readout



Carbon Nanotube NEMS, Fig. 6 Schematic of the one-source (*AM*) mixing setup for resonant motion readout

radio-frequency (RF) signals offset by a small frequency difference $\delta\omega$ (typically in the kHz range) are applied to the gate and drain of the CNT FET, respectively. The RF signal at frequency ω generates a nanotube motion at the same frequency and thus the modulation of device conductance: $\delta G = \delta G_0 \cos(\omega t)$. The drain bias applied is at a slightly different frequency: $V_{\text{Drain}} = V_{\text{D}} \cos((\omega + \delta\omega)t)$. The modulation in the current through the nanotube FET is therefore

$$\begin{aligned}\delta I &= V_{\text{Drain}} \times \delta G \\ &= \delta G_0 V_{\text{D}} \cos\left((\omega + \delta\omega)t\right) \cos\left(\omega t\right) \\ &= \delta G_0 V_{\text{D}} (\cos\left((2\omega + \delta\omega)t\right) + \cos\left(\delta\omega t\right))/2\,,\end{aligned} \tag{4}$$

which has a term at the difference frequency $\delta\omega$ and still carries the information of the device motion (through the term δG). In contrast, the capacitive current, though still much larger in amplitude, remains purely at frequency ω. Using a lock-in amplifier referenced at the difference frequency $\delta\omega$, it is possible to reject the electronic signal from all other frequencies and detect the device motion (at ω) by examining the down-mixed FET current at $\delta\omega$.

The AM mixing technique simplifies the two-source scheme by requiring only one RF signal source (Fig. 6). The output of the RF source is AM modulated at the lock-in reference frequency $\delta\omega$ and is applied to the drain electrode

$$V_D = (1 + A\cos\left(\delta\omega t\right))V_0 \cos\left(\omega t\right), \tag{5}$$

which has components at the frequencies ω, $\omega + \delta\omega$, and ω-$\delta\omega$. A is the modulation depth and ranges between 0 and 100 % (values close to 100 % are often used). The gate voltage only carries a DC component in this case. The AC component on the drain effectively creates an alternating bias voltage between the nanotube and the gate and drives the vibration of the nanotube. When considering the motion of the nanotube resonator, this is equivalent to applying these three AC components at the gate. Thus, the AM

scheme is equivalent to the two-source setup, with both the conductance modulation and the drain bias having components in ω, $\omega + \delta\omega$, and ω-$\delta\omega$. Consequently, the mixing current has low frequency components at DC, $\delta\omega$, and $2\delta\omega$. Typically, the $\delta\omega$ component is being measured by the lock-in amplifier.

In the AM readout scheme, even though the capacitive current from the contacts are removed through frequency down-mixing, the non-motional current (the second term in Eq. 3) has a finite term at $\delta\omega$, which presents a frequency-dependent background, and can sometimes hinder the observation of mechanical resonance. The FM readout scheme [12], a later-developed technique, removes this non-motional mixing current in addition to the capacitive current, allowing more efficient detection of mechanical resonance.

For singly clamped devices, direct electronic readout is unavailable as the nanotube itself does not typically form a closed path in the circuit. One can force the open circuit to close by using the nanotube as a field emitter by applying a high voltage across the tube and an opposing cathode and monitor the field-emission current as the nanotube vibrates. Such readout scheme can be destructive, and irreversible shortening of the nanotube due to the large current has been observed [20]. Alternatively, the vibrational amplitude can be estimated directly through imaging in electron microscope. While in such scheme quantitative measurement remains very challenging, imaging offers the opportunity of directly observing the mode shape of the individual resonances, similar to AFM measurements [13].

Frequency Tuning

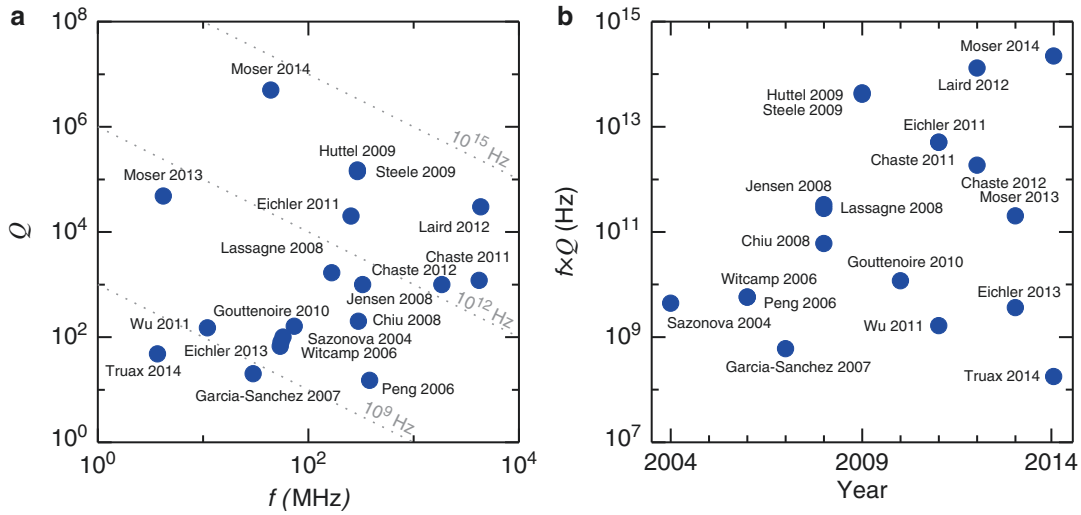
The resonance frequency of carbon nanotube NEMS resonator can often be effectively tuned via adjusting the DC gate voltage. There are two main effects involved: electrostatic tensioning, which leads to an increase in resonance frequency as V_g is increased, and capacitive softening, which has the opposite effect.

For doubly clamped devices, the attractive force resulting from the application of gate bias induces tension in the nanotube and thus increases its resonance frequency. The magnitude of the gate voltage affects the elastic regime in which the nanotube NEMS resonator operates [3]. At small gate bias, the flexural rigidity dominates, and the nanotube vibrates as a bending beam. As V_g increases, the flexural rigidity is overcome by the electrostatic force from the gate, and the nanotube operates as a hanging chain that cannot be elongated (i.e., the chain has negligible flexural rigidity but infinite extensional rigidity). As the gate voltage further increases, the extensional modulus becomes important, and the nanotube behaves like an elastic string. In these different regimes, the frequency has different dependences on the gate voltage.

Due to the large aspect ratio of carbon nanotube, it is common for the suspended segment in a doubly clamped device to exhibit finite slack (i.e., the length of the nanotube is longer than the distance between its two clamping points). It is worth noting that the amount of initial slack affects how the frequency increases with V_g in the different regimes, and such effect further depends on individual resonant modes [3, 14].

Capacitive softening roots from the fact that the nanotube is vibrating in a nonuniform electric field, and the finite field gradient reduces the restoring force toward the equilibrium position (a negative effective spring constant). With finite slack, the nanotube can thus have both in-plane and out-of-plane resonant motion, which exhibit different responses to the capacitive softening effect [4], as the observation of such effect requires the device motion be in the direction of the electric field gradient.

Unlike in doubly clamped devices, tension cannot be easily introduced in singly clamped nanotube NEMS resonators. Thus, such device typically operates in the bending regime, and the use of multi-walled nanotubes (with greater bending rigidity) can help achieve higher resonance frequencies. Further, the gate tuning effect, though it exists, is much smaller than in doubly clamped devices in which tension can be effectively introduced.



Carbon Nanotube NEMS, Fig. 7 Partial summary of (**a**) f_{res}, Q and (**b**) $f_{res} \times Q$ for carbon nanotube NEMS resonators studied between 2004 and 2014. *Dashed lines* in (**a**) represent contours of constant $f_{res} \times Q$

Device Performances

For NEMS resonators, two important device parameters are the resonance frequency f_{res} and quality factor Q. While f_{res} can be effectively turned by gate voltage, within the commonly used V_g range (on the order of ± 10 V), the f_{res} variation is generally within one order of magnitude. In contrast, other device parameters, such as dimension and initial tension, have much larger effects on f_{res}. Specifically, shorter devices have higher f_{res} (e.g., $f_{res} = 4.2$GHz for an 110-nm-long nanotube resonator [15], while $f_{res} = 4.2$ MHz for a 4-μm-long device made by the same group of researchers [16]), and devices with lower tension can have very low f_{res} (e.g., $f_{res} = 3.7$ MHz for a 1.89-μm-long nanotube suspended over a MEMS structure which can introduce slack/reduce tension in the nanotube in a controlled way [17]).

The quality factor of carbon nanotube resonators typically increases as temperature decreases: Q ranges between $10^1 \sim 10^3$ at room temperature, and increases to $10^1 \sim 10^6$ at cryogenic temperatures, due to reduction of phonon-phonon scattering and suppression of electron motion (in the Coulomb blockade regime), two mechanisms that dissipate energy during the nanotube motion. Advancement in transduction techniques has further enabled electronic readout of the completely undriven thermomechanical motion of carbon nanotube NEMS resonators, and by removing external driving and electrostatic noise from the gate electrode, researchers have observed mechanical resonance with intrinsic Q up to 5×10^6 [18]. Figure 7a provides a partial list of f_{res} and Q for some of the nanotube resonators demonstrated between 2004 and 2014, and Fig. 7b shows how $f_{res} \times Q$, one of the most important figures of merit for NEMS devices, evolves throughout the decade.

Applications

Carbon nanotube NEMS resonators are highly sensitive to external stimuli due to the small mass and high flexibility of carbon nanotubes and are thus used in a number of sensing applications.

The resonance frequency of a NEMS resonator is sensitive to any adsorption on its surface, which increases the mass in motion (and thus decreases f_{res}). On the first order (not considering the mode shape of individual resonance modes and the detailed location of adsorbed mass), the frequency

shift δf due to adsorption of additional mass δm for a resonator with mass m is

$$\delta f = \frac{\delta m}{2m} f_{\text{res}}. \tag{6}$$

Given the same frequency resolution, a smaller resonator mass m leads to a finer adsorption mass resolution δm. Given the ultrasmall mass of an individual carbon nanotube (especially single-walled carbon nanotube), it is natural that CNT NEMS resonators are used for ultimate mass sensing. Initial works have achieved mass sensitivity of $10^{-21} \sim 10^{-22}$g [8, 9, 19], sufficient for resolving individual Argon atoms. From Eq. 6, it is also clear that a higher f_{res} leads to a higher frequency-to-mass responsivity (f_{res} shift per adsorbed mass). Consequently, years after the initial experiments, with improvement in device fabrication and transduction scheme, and very importantly, by using a carbon nanotube resonator with a much higher resonance frequency (f_{res} ~2GHz, one order of magnitude higher than in previous works), mass sensing at single proton level (10^{-24}g) has been demonstrated [10].

In addition to mass sensing, carbon nanotube NEMS resonators have also been used for ultrasensitive force detection. In contrast to mass sensing, where high f_{res} (which means short nanotube with high tension) is desirable, to achieve more sensitive force detection, the opposite is preferred. This is because long nanotubes with little initial tension have ultralow spring constant, and thus, its frequency can be highly sensitive to external forces which modify the spring constant of the resonator system. Using a 4-μm-long nanotube resonator with f_{res} as low as 4.2 MHz, researchers have demonstrated force sensitivity of $10^{-20}\text{N/Hz}^{-1/2}$ at cryogenic temperatures [16].

Compared with other NEMS structures fabricated using top-down techniques (such as lithography and etching), carbon nanotube resonator is unique in that the surface of the motional part (the nanotube) can be atomically well defined, especially for CNT NEMS fabricated using the growth-on-prefabricated-structure scheme. This offers the opportunity of exploring surface phenomena on a highly curved and atomically perfect surface, such as phase transition in the adsorbed monolayer on the nanotube surface. Historically, graphite surface has been the canonical platform for studying surface adsorption and phase behavior in two-dimensional adsorption systems, thanks to its simple and well-defined surface structure (the honeycomb pattern formed by carbon atoms). The surface of a single-walled carbon nanotube inherits the advantages of the graphite surface and offers a few additional properties unavailable in conventional two-dimensional (2D) graphitic surfaces: first, it is highly curved and thus imposes an additional boundary condition for the adsorbed layer (along the circumferential direction), and due to this strong confinement, the resulting adsorption system approaches the 1D limit; second, the significant reduction of carbon atoms (compared with those on and under conventional 2D graphitic surface) leads to decreased adsorption potential energy and thus can shift the equilibria of phase transitions in the adsorption system. When a phase transition takes place in the adsorbed atomic layer (which is accompanied by a sudden change in its density, as in 3D phase transitions), the mass of the resonator undergoes an abrupt change, which is reflected in a sudden shift in resonance frequency. Experimentally, phase transitions have been observed for Ar, Kr, and ^{4}He, with several low-dimensional phases, including commensurate phase – a phase specific to adsorption systems, in which the arrangement of adsorbate atoms conforms to the underlying substrate atoms – being identified [5, 11].

Besides sensing applications, carbon nanotube NEMS resonators have also been demonstrated as active components in RF circuits. Both singly clamped [20] and doubly clamped [12] structures have been used. In one example, researchers utilize the field-emission current from a singly clamped nanotube resonator to function the NEMS device as the antenna, tuner, amplifier, and demodulator in a radio. In another case, using the FM mixing technique, researchers demonstrate digital demodulation function with a doubly clamped carbon nanotube NEMS resonator, with data transfer rate meeting the GSM specifications.

Cross-References

- Carbon Nanotube NEMS
- Carbon-Nanotubes
- Graphene NEMS
- Nanomechanical Properties of Nanostructures
- Nanomechanical Resonant Sensors and Fluid Interactions
- NEMS Mass Sensors
- NEMS Resonant Chemical Sensors
- NEMS Resonant Mass Sensors

References

1. Cumings, J., Zettl, A.: Low-friction nanoscale linear bearing realized from multiwall carbon nanotubes. Science **289**, 602–604 (2000)
2. Fennimore, A.M., Yuzvinsky, T.D., Han, W.Q., Fuhrer, M.S., Cumings, J., Zettl, A.: Rotational actuators based on carbon nanotubes. Nature **424**, 408–410 (2003)
3. Sazonova, V., Yaish, Y., Ustunel, H., Roundy, D., Arias, T.A., McEuen, P.L.: A tunable carbon nanotube electromechanical oscillator. Nature **431**, 284–287 (2004)
4. Wu, C.C., Zhong, Z.: Capacitive spring softening in single-walled carbon nanotube nanoelectromechanical resonators. Nano Lett. **11**, 1448–1451 (2011)
5. Wang, Z., Wei, J., Morse, P., Dash, J.G., Vilches, O.E., Cobden, D.H.: Phase transitions of adsorbed atoms on the surface of a carbon nanotube. Science **327**, 552–555 (2010)
6. Peng, H.B., Chang, C.W., Aloni, S., Yuzvinsky, T.D., Zettl, A.: Ultrahigh frequency nanotube resonators. Phys. Rev. Lett. **97**, 87203 (2006)
7. Wu, C.C., Liu, C.H., Zhong, Z.: One-step direct transfer of pristine single-walled carbon nanotubes for functional nanoelectronics. Nano Lett. **10**, 1032–1036 (2010)
8. Chiu, H.Y., Hung, P., Postma, H.W.C., Bockrath, M.: Atomic-scale mass sensing using carbon nanotube resonators. Nano Lett. **8**, 4342–4346 (2008)
9. Lassagne, B., Garcia-Sanchez, D., Aguasca, A., Bachtold, A.: Ultrasensitive mass sensing with a nanotube electromechanical resonator. Nano Lett. **8**, 3735–3738 (2008)
10. Chaste, J., Eichler, A., Moser, J., Ceballos, G., Rurali, R., Bachtold, A.: A nanomechanical mass sensor with yoctogram resolution. Nat. Nanotechnol. **7**, 301–304 (2012)
11. Lee, H.-C., Vilches, O.E., Wang, Z., Fredrickson, E., Morse, P., Roy, R., Dzyubenko, B., Cobden, D.H.: Kr and ^{4}He adsorption on individual suspended single-walled carbon nanotubes. J. Low Temp. Phys. **169**, 338–349 (2012)
12. Gouttenoire, V., Barois, T., Perisanu, S., Leclercq, J.-L., Purcell, S.T., Vincent, P., Ayari, A.: Digital and FM demodulation of a doubly clamped single-walled carbon-nanotube oscillator: towards a nanotube cell phone. Small **6**, 1060–1065 (2010)
13. Garcia-Sanchez, D., Paulo, A.S., Esplandiu, M.J., Perez-Murano, F., Forro, L., Aguasca, A., Bachtold, A.: Mechanical detection of carbon nanotube resonator vibrations. Phys. Rev. Lett. **99**, 85501 (2007)
14. Eichler, A., Moser, J., Chaste, J., Zdrojek, M., Wilson-Rae, I., Bachtold, A.: Nonlinear damping in mechanical resonators made from carbon nanotubes and graphene. Nat. Nanotechnol. **6**, 339–342 (2011)
15. Chaste, J., Sledzinska, M., Zdrojek, M., Moser, J., Bachtold, A.: High-frequency nanotube mechanical resonators. Appl. Phys. Lett. **99**, 213502 (2011)
16. Moser, J., Güttinger, J., Eichler, A., Esplandiu, M.J., Liu, D.E., Dykman, M.I., Bachtold, A.: Ultrasensitive force detection with a nanotube mechanical resonator. Nat. Nanotechnol. **8**, 493–496 (2013)
17. Truax, S., Lee, S.-W., Muoth, M., Hierold, C.: Axially tunable carbon nanotube resonators using co-integrated microactuators. Nano Lett. **14**(11), pp 6092–6096 (2014), doi:10.1021/nl501853w
18. Moser, J., Eichler, A., Güttinger, J., Dykman, M.I., Bachtold, A.: Nanotube mechanical resonators with quality factors of up to 5 million. Nat. Nanotechnol. **9**, 1007–1011 (2014), doi:10.1038/nnano.2014.234, advance online publication
19. Jensen, K., Kim, K., Zettl, A.: An atomic-resolution nanomechanical mass sensor. Nat. Nanotechnol. **3**, 533–537 (2008)
20. Jensen, K., Weldon, J., Garcia, H., Zettl, A.: Nanotube radio. Nano Lett. **7**, 3508–3511 (2007)

Carbon Nanotube-Metal Contact

Wenguang Zhu
Department of Physics and Astronomy, The University of Tennessee, Knoxville, TN, USA

Synonyms

Carbon nanotube-metal interface

Definition

Carbon nanotube-metal contacts are widely present in many carbon nanotube-based nanodevices, and their electronic structures may significantly

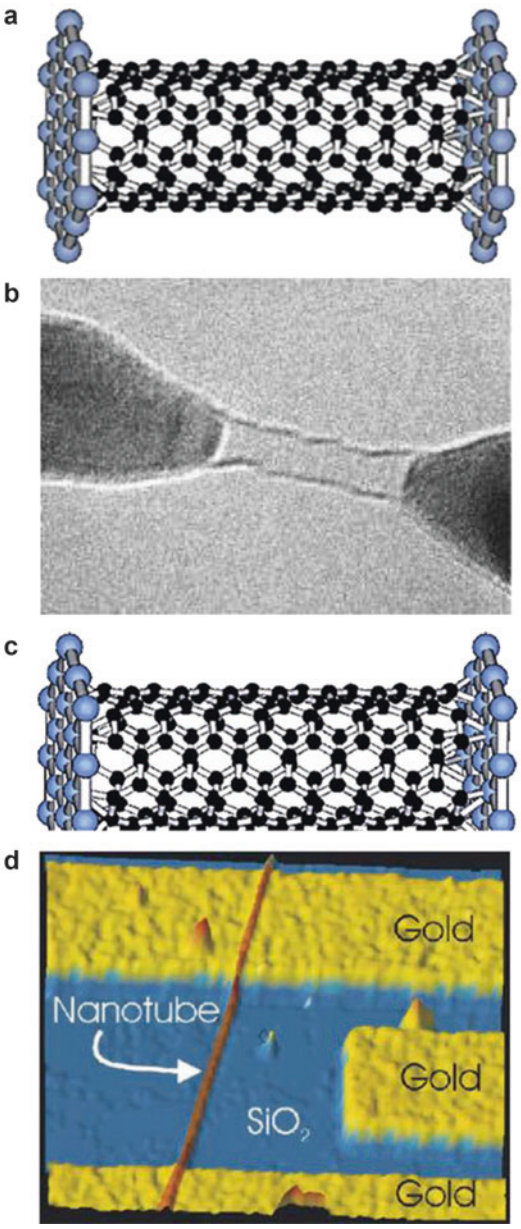
influence the operation and performance of carbon nanotube-based nanodevices.

Overview

Carbon nanotubes (CNTs) are quasi-one-dimensional materials with remarkable mechanical and electronic properties promising a wide range of applications from field-effect transistors (FETs) and chemical sensors to photodetectors and electroluminescent light emitters. In most of these CNT-based nanodevices, metals are present as electrodes in contact with the CNTs. Many factors including the CNT-metal contact geometry, microscopic atomic details at the interface, and the resulting electronic structure can play a significant role in determining the functionality and performance of the devices. For instance, it has been demonstrated that an individual semiconducting CNT can operate either as a conventional FET or an unconventional Schottky barrier transistor, depending on the properties of the metal-CNT contact. In general, the electrical transport characteristics of the CNT-metal systems are sensitive to the choice of metal element as the electrode.

CNT-Metal Contact Geometry

There are two types of interface geometries of CNT-metal contacts, i.e., end contact and side contact [1, 2]. The end-contact geometry refers to the cases where metals are merely in contact with the open ends of one-dimensional CNTs, as illustrated in Fig. 1a. This contact geometry can be naturally achieved in the catalytic CVD growth of CNTs, where CNTs sprout from catalytic metal particles with the CNT axis normal to the metal surface. Figure 1b shows a sample experimental image of an end contact between a single-wall CNT and two Co tips in an in situ electron microscopy setup. The side-contact geometry refers to the cases where metals are in contact with the sidewall of CNTs, as illustrated in Fig. 1c. This contact geometry occurs when a CNT lays on the surface of a flat metal substrate. In most of CNT-based nanodevices, such as CNT FETs, metal strips are deposited from above to cover the CNTs laying on the surface as to build



Carbon Nanotube-Metal Contact, Fig. 1 A schematic illustration of (**a**) an end contact and (**c**) a side contact between a CNT and a metal (From Ref. [3], Fig. 1). (**b**) A CNT forming end contacts with Co tips (From Ref. [4], Fig. 1). (**d**) A CNT forming side contacts on gold electrodes (From Ref. [2], Fig. 24)

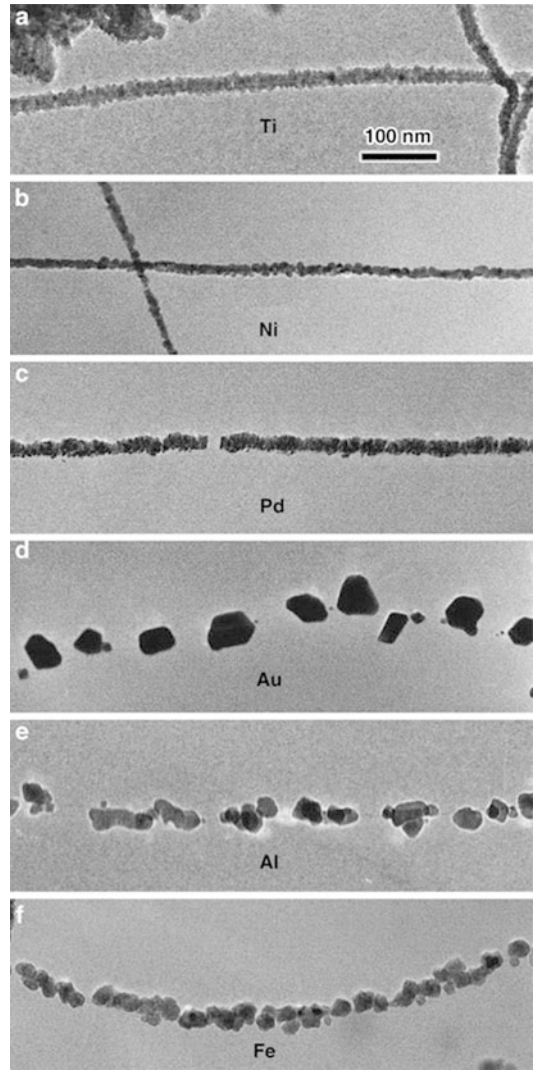
electrodes, fully covering sections of the CNTs, as shown in Fig. 1d. Among these two contact geometries, the side-contact geometry is more technologically relevant to CNT-based nanodevices.

Bonding and Wetting Properties of Metals on CNTs

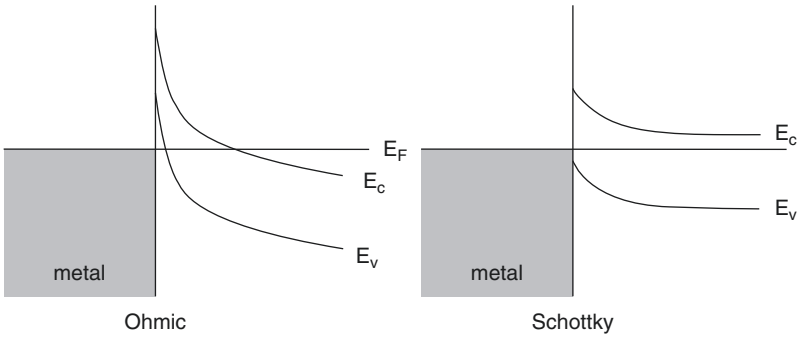
In the end-contact geometry, metals form strong covalent bonds with carbon atoms at the open ends of CNTs [5]. The bonding energy can be as high as, e.g., 7.6 eV for a single bond at a CNT–Co contact, according to density functional calculations. Due to the strong covalent nature of the bonding, large mismatch-induced strains or high tensile strength can be built up at the interface.

In the side-contact geometry, metals and CNTs form much weaker bonds due to the nearly chemically inert side walls of CNTs [5]. Single-wall CNTs are built up of a cylindrically closed sheet of graphene, in which carbon atoms arranged in a honeycomb structure form very stable *sp2*-hybridized covalent bonds with the *pz*-orbitals of carbon extending normal to the sidewalls. The interaction between metals and CNTs in the side-contact geometry is determined by the hybridization between the carbon *pz*-orbitals and the unbonded orbitals of the metals. Alkali and simple metals have binding energy around 1.5 eV per atom. Some transition metal atoms with unpaired *d* electrons, such as Sc, Ti, Co, Ni, Pd, Pt, form strong bonds with a binding energy around 2.0 eV per atom, whereas the transition metals with fully occupied *d* orbitals such as Cu, Au, Ag, and Zn have relatively weak binding with a binding energy less than 1.0 eV per atom. On the other hand, the binding energy of metal on CNTs also depends on the radius of CNTs. In general, the larger is the radius, the weaker the binding energy.

The wettability of metals on CNTs is critical to the electrical transport properties at CNT-metal contacts. In addition, CNTs can be used as templates to produce metallic nanowires with controllable radius by continuously coating the sidewalls of CNTs with metals. Experiments using different techniques such as electron beam evaporation, sputtering, and electrochemical approaches have achieved continuous coating of Ti and quasi-continuous coating of Ni and Pd on CNTs [5, 6]. Such metallic nanowires are ideal to be used as conducting interconnects in nanodevices. Metals such as Au, Al, Fe, Pb form isolated discrete clusters rather than a uniform coating layer on the surface of CNTs. Figure 2 shows sample TEM images of Ti, Ni, Pd, Au, and Fe coatings on CNTs [6]. The correlation between the wettability



Carbon Nanotube-Metal Contact, Fig. 2 TEM images of (**a**) Ti, (**b**) Ni, (**c**) Pd, (**d**) Au, (**e**) Al, and (**f**) Fe coatings on carbon nanotubes



Carbon Nanotube-Metal Contact, Fig. 3 A schematic illustration of the energy levels for an ohmic and a Schottky contact between a metal and a semiconducting CNT

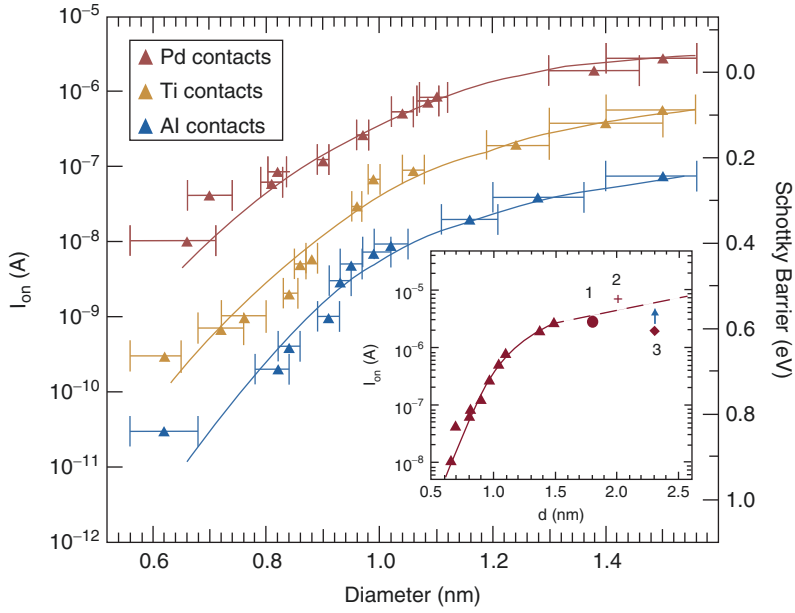
of these metals and their binding energies on CNTs is clear, i.e., metals with relatively strong binding energies with CNTs tend to form uniform coatings.

Electronic Structures of CNT-Metal Contacts

The electronic structure of CNT-metal contacts has a significant impact on the operation and performance of CNT-based nanodevices [2]. Due to the one-dimensional nature of CNTs and their special contact geometries, CNT-metal contacts exhibit some unusual features when compared to traditional planar contacts.

For metallic CNTs, as in contact with metals, ohmic contacts are normally formed at the interface, where no interface potential barrier exists, and the contact resistance is primarily determined by the wettability of the metal and the local atomic bonding and orbital hybridization at the interface [2]. Palladium is found to be optimal as electrodes to make ohmic contacts with metallic CNTs.

Semiconducting CNTs form either ohmic contacts or Schottky barriers at the interface with metals [2]. Figure 3 schematically illustrates the energy levels for an ohmic and a Schottky contact between a metal and a semiconducting CNT. A distinctive feature of CNT-metal contacts from traditional planar metal/semiconductor interfaces is that the height of the Schottky barrier formed at CNT-metal contacts strongly depends on the work function of the metal for a given semiconducting CNT [7, 8]. In general, at traditional planar metal/semiconductor interfaces, the Schottky barrier height shows very weak dependence on the metal work function due to the so-called Fermi-level pinning effect [9]. The strong dependence of the Schottky barrier on the metal work function in CNT-metal contacts is attributed to the reduced dimensionality of CNTs, which entirely changes the scaling of charge screening at the interface, making the depletion region decay rapidly in a direction normal to the interface and thus significantly weakening the Fermi-level pinning effect [7]. Experimental and theoretical work has shown that the interface Schottky barrier regions are much thinner in one dimension than those in three dimensions. In this case, charge carrier tunneling through the Schottky barriers becomes important. Because of the involvement of tunneling and thermionic emission in the carrier transport at the interfaces, the dependence of the on-current of CNT transistors on the Schottky barrier becomes very strong. Figure 4 shows experimental CNT-FET on-current and Schottky barrier height as a function of the CNT diameter for three different metal electrodes, Pd, Ti, and Al [8]. When the metal work functions are in the valence or conduction band of the semiconducting CNTs, ohmic contacts will likely be formed at the interface. Experimental measurements have shown that certain metals with high work functions, such as Pd, can produce nearly ohmic contacts with semiconducting CNTs. Ohmic contacts are more desirable in devices where contact resistance needs to be minimized. In addition to the metal work function, other factors, such as the



Carbon Nanotube-Metal Contact,
Fig. 4 Experimental CNT-FET on-current (*left axis*) and computed Schottky barrier height (*right axis*) as a function of the CNT diameter for three different metal electrodes, Pd, Ti, and Al

contact geometry and the chemical bonding at the interface also play important roles in the transport properties of CNT-metal contacts.

Cross-References

- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Carbon-Nanotubes
- ▶ CMOS-CNT Integration

References

1. Banhart, F.: Interactions between metals and carbon nanotubes: at the interface between old and new materials. Nanoscale **1**, 201–213 (2009)
2. Anantram, M.P., Léonard, F.: Physics of carbon nanotube electronic devices. Rep. Prog. Phys. **69**, 507–561 (2006)
3. Palacios, J.J., Pérez-Jiménez, A.J., Louis, E., SanFabián, E., Vergés, J.A.: Phys. Rev. Lett. **90**, 106801 (2003)
4. Rodríguez-Manzo, J.A., et al.: Small **5**, 2710–2715 (2009)
5. Ciraci, S., Dag, S., Yildirim, T., Gülseren, O., Senger, R. T.: Functionalized carbon nanotubes and device applications. J. Phys. Condens. Matter **16**, R901–R960 (2004)
6. Zhang, Y., Franklin, N.W., Chen, R.J., Dai, H.J.: Metal coating on suspended carbon nanotubes and its implication to metal-tube interaction. Chem. Phys. Lett. **331**, 35–41 (2000)
7. Léonard, F., Tersoff, J.: Role of fermi-level pinning in nanotube Schottky diodes. Phys. Rev. Lett. **84**, 4693–4696 (2000)
8. Chen, Z.H., Appenzeller, J., Knoch, J., Lin, Y.-M., Avouris, P.: The role of metal – nanotube contact in the performance of carbon nanotube field-effect transistors. Nano Lett. **5**, 1497–1502 (2005)
9. Tung, R.T.: Recent advances in Schottky barrier concepts. Mater. Sci. Eng. R **35**, 1–138 (2001)

Carbon Nanotube-Metal Interface

- ▶ Carbon Nanotube-Metal Contact

Carbon Nanotubes

- ▶ Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae

Carbon Nanotubes (CNTs)

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Physical Vapor Deposition

Carbon Nanotubes for Chip Interconnections

Gilbert Daniel Nessim
Chemistry Department, Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan, Israel

Synonyms

Carbon nanotubes for interconnects in integrated circuits; Carbon nanotubes for interconnects in microprocessors

Definition

Chip interconnections electrically connect various devices in a microprocessor. Today's established technology for interconnects is based on copper. However, it may be technically challenging to extend copper use to future interconnects in microprocessors with smaller lithographic dimensions due to materials properties limitations. Carbon nanotubes are currently investigated as a potential replacement for future integrated circuits (microprocessors). Although carbon nanotubes are a clear winner against copper in terms of materials properties, multiple fabrication challenges need to be overcome for carbon nanotubes to enter the semiconductor fab and replace copper for chip interconnections.

Motivation

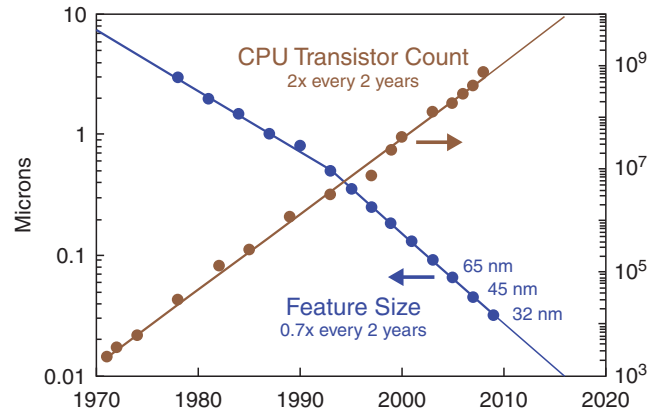
Following over 40 years of successful fulfillment of Moore's law, stating that the number of transistors in a chip doubles every 2 years, we have already moved from microelectronics to nanoelectronics [1]. Although the *"end of scaling"* has been predicted many times in the past, enormous technical challenges, especially quantum mechanical issues and billion-dollar lithography investments, are a serious threat to further miniaturization (Fig. 1).

Today's latest processors are manufactured using the 32-nm technology. To move toward the 22-nm node and beyond, issues such as lithographic limitations, leaking currents in ultra-thin dielectrics (only a few monolayers thick), insufficient power and thermal dissipation, and interconnect reliability must be resolved [1]. At the transistor level, the performance is negatively affected by increased off-state currents due to short channel effects, increased gate leakage due to tunneling through nanometer-thin dielectric layers, and increased overall gate capacitance due to decreasing gate pitch.

Although quantum mechanical tunneling and leakage currents may eventually stop further scaling, efficient heat removal from a chip is currently the biggest obstacle. In this respect, the many kilometers of copper interconnects present in today's chips are the main culprit for heat generation. For instance, in 2004, Magen et al. [2] showed that for a microprocessor fabricated with the 0.13-μm-node technology consisting of 77 million transistors, interconnects consumed more than 50 % of the total dynamic power. Given the increased length of interconnects, their reduced cross section, and the increased current densities circulating into the interconnects of our latest chips, the problem has been further exacerbated.

Additionally, copper interconnects are a major contributor to the total resistance-capacitance (RC) delay of the chip, can fail by electromigration, and need a liner to avoid diffusion into the silicon. Bottom line, the interconnect issue is so serious that the International Semiconductor Roadmap [3] (ITRS, an expert team assessing the semiconductor industry's future technology requirements for the next 15 years) indicates copper interconnects as a possible dealbreaker to further miniaturization for IC nodes beyond 22 nm.

Many technology options are currently under investigation to replace copper for interconnects. Among them, we can mention other metals (mainly silicides), wireless, plasmonics, and optical interconnects. Most notably, there has been an intense research effort on new nanotechnology materials such as carbon nanotubes, which, at



Carbon Nanotubes for Chip Interconnections, Fig. 1 Moore's Law: Transistor count has doubled while feature size has decreased by 0.7X every 2 years (Figure reprinted with permission from Kuhn [1])

the theoretical level, could solve all the above technical issues suffered by copper.

The plan of this section is to first introduce the reader to copper interconnects' fabrication and limitations. Next, we will compare copper to carbon nanotubes (CNTs) and detail possible models for implementation. An important paragraph will focus on the state-of-the-art of CNT fabrication, prior to concluding on the outstanding issues and outlook for future CNT-based chip interconnections.

Background on Copper Interconnects and Dual-Damascene Process

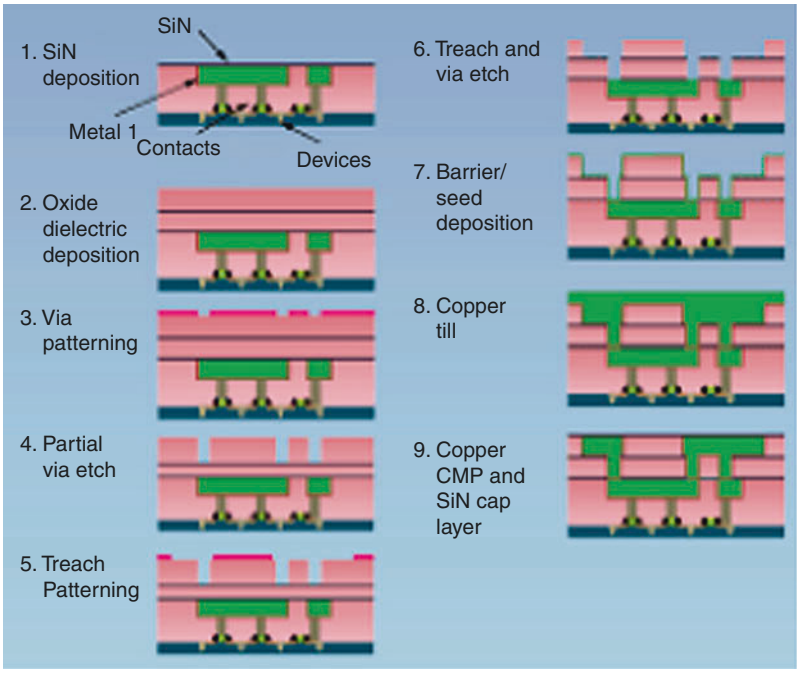
In 1997, IBM introduced the revolutionary *"dual-damascene"* process to fabricate copper interconnects and to replace aluminum interconnects, the industry standard at the time. Compared to aluminum, copper presents two major advantages: (1) 50 % lower resistivity (Cu ≈ 1.75 μm cm vs. Al ≈ 3.3 μm cm) and (2) higher current densities before failure by electromigration (up to 5×10^6 A/cm^2) [4, 5]. Although, as a material, copper was a clear winner against aluminum, fabrication challenges delayed its introduction. Historically, we may be at a similar juncture with carbon nanotubes compared to copper as we were with copper compared to aluminum in 1997: in spite of their superior materials properties, mainly fabrication issues are now preventing the introduction of nanotubes in the semiconductor industry to replace copper interconnects.

Copper diffuses into silicon, generating mid-gap states that significantly lower the minority carrier lifetime and which lead to leakage in diodes and bipolar transistors. Copper also diffuses through SiO_2 and low-k dielectrics, and therefore requires complete encapsulation in diffusion barriers. Since no dry etches were known for copper, IBM's bold innovation of polishing using chemical mechanical polishing (CMP), after electroplating the copper, was significantly at odds with the technological processes at that time in semiconductor fabrication.

The copper dual-damascene process consists of the following steps:

- Develop a pattern for wires or vias by patterned etching of the dielectric.
- Deposit a barrier layer (usually Ta) to prevent copper diffusion into silicon.
- Deposit a copper seed layer.
- Fill the vias with copper using electrodeposition.
- Remove excess copper using CMP.
- Repeat the process to lay the alternating layers of wires and vias which will form the complete wiring system of the chip (Fig. 2).

Typical microprocessor design follows a *"reverse scaling"* metallization scheme with multiple layers of interconnects labeled as local, intermediate, and global interconnects, with increasing width. Very thin local interconnects locally connect gates and transistors within a functional block and are usually found in the lower two



Carbon Nanotubes for Chip Interconnections, Fig. 2 Dual-damascene process of copper filling an interconnect via (Figure reprinted with permission from Jackson et al. [21])

metal layers. The wider and taller intermediate interconnects have lower resistance and provide clock and signal within a functional block up to 4 μm. Global interconnects are found at the top metal layers and provide power to all functions in addition to connecting functional blocks through clock and signal. They are usually longer than 4 μm (up to half of the chip perimeter) and exhibit very low resistance to minimize RC delay and voltage drop. Below are a typical cross section of an I.C. chip and a possible implementation using CNTs (Figs. 3 and 4).

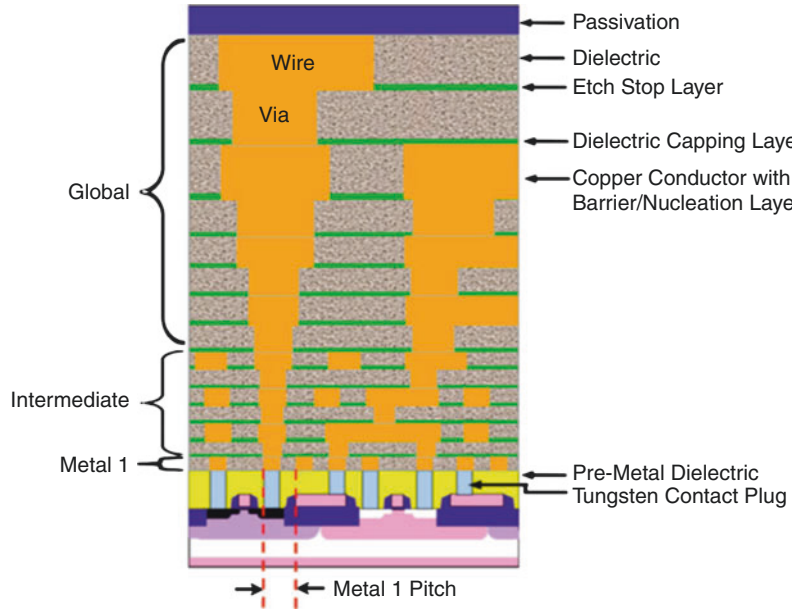
Limitations of Copper Interconnections

Copper interconnects have efficiently scaled down to the current 32-nm-node microprocessors, although this has required many technological advances to allow ever-shrinking copper cross sections to carry increasing currents without failure. However, we may be very close to smashing against a technical wall because of materials failure and related fabrication issues.

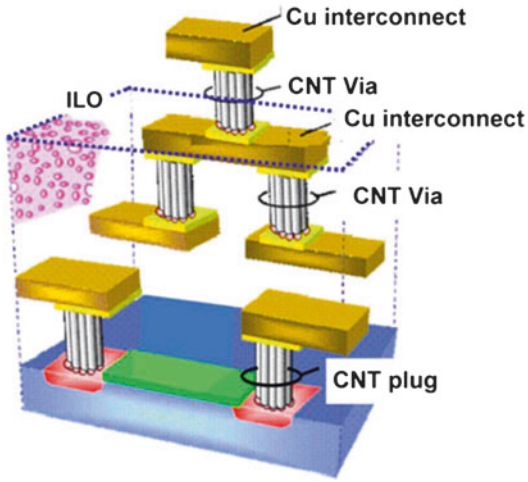
Alternative materials or technologies would require many changes in semiconductor fabrication and massive investments; thus the large semiconductor companies are doing the impossible to extend copper application to future nodes. It is clear that only when up against an insurmountable technical wall will the semiconductor industry switch to a new technology.

Electrical resistance is a major issue now that copper interconnect cross sections are comparable to the mean free path of electrons in copper (~40 nm in Cu at room temperature). Grain boundary and surface scattering are significant contributors to the increased resistance, especially now that we have reached nanoscale dimensions. At the microstructural level, the grain boundaries play an important role, hence, among other fabrication concerns, controlling the copper grain size during electrodeposition has allowed to limit the grain boundary scattering impact thus far.

The steep rise in interconnect resistance for smaller IC nodes is a major source of RC delays and directly affects the chip reliability by increasing the risk of electromigration failure, a major issue for further downscaling. Electromigration is the transport of material caused by the gradual movement of the copper ions due to the momentum transfer between conducting electrons and



Carbon Nanotubes for Chip Interconnections, Fig. 3 Typical cross sections of hierarchical scaling in current microprocessor (Figure reprinted with permission from the Semiconductor Industry Association [22])



Carbon Nanotubes for Chip Interconnections, Fig. 4 Schematic view of possible implementation of carbon nanotube via interconnects in lieu of copper (Figure reprinted with permission from Awano et al. [7])

diffusing metal atoms, which occurs for high current densities, which can create voids leading to open circuits. Given that downscaling leads to a reduction of the interconnects' cross section, the problem is amplified at subsequently smaller nodes. To compound the issue, the need for a resistive diffusion barrier layer, also called a liner (usually Ta), to avoid copper diffusion into silicon, further reduces the available conductive copper cross section, thus increasing the risk of electromigration failure, especially as the operating temperature rises.

In addition to the increased resistance and the electromigration failure risk, many other aspects of the dual-damascene process are becoming potential sources of failure as the node shrinks. Among the many integration concerns, we can mention materials issues such as interface adhesion between the different materials (copper, low-k dielectrics, etc.), liner effectiveness, metal voids, CMP interface defects, etc. Concurrently, there is a long list of process-related issues such as the need for etch/strip/clean processes (to avoid damage to low-k dielectric materials), atomic layer deposition (ALD) processes to deposit liners, copper plating and CMP techniques, etc.

A few interesting numerical estimates taken from the 2009 projections from ITRS, [3] provide the reader with the magnitude of the technical challenge to extend copper interconnect technology (Table 1).

As already mentioned, many alternative technologies are currently being investigated for replacement of copper as interconnect material that would require significant chip redesigns and

Carbon Nanotubes for Chip Interconnections, Table 1 Selected critical parameters for copper use as interconnects in future IC nodes (Data from the Semiconductor Industry Association [22])

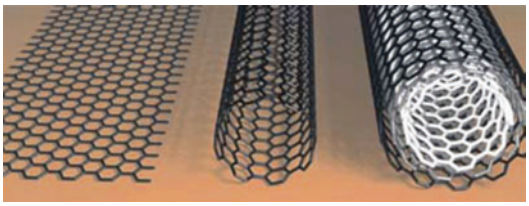
Year of production (estimated)	2010	2015	2020
MPU/ASIC metal 1½ pitch (nm) (contacted)	45	25	14
Total interconnect length (m/cm^2) – metal 1 and 5 intermediate levels, active wiring only	2,222	4,000	7,143
Barrier/cladding thickness (for Cu intermediate wiring) (nm)	3.3	1.9	1.1
Interconnect RC delay [ps] for a 1-mm Cu intermediate wire, assumes no scattering and an effective ρ of 2.2 μΩ-cm	1,132	3,128	9,206

new fabrication technologies. Some examples include optical interconnects, radio frequency (RF) interconnects, plasmonics, and 3-D interconnects (probably still copper). The interested reader can find more details on these alternative technologies in the review paper from Havemann et al. [6] (now a little dated) or in the latest ITRS report on interconnects [3].

The Case for Carbon Nanotube Interconnects

An interesting solution, which has been the subject of intense research in recent years, is to replace copper with carbon nanotubes. If a reliable and repeatable fabrication process consistent with Complementary Metal Oxide Semiconductor (CMOS) technology requirements could be developed, integration into existing chip architectures may not require significant process redesign (Fig. 5).

Carbon nanotubes, which can be visualized as rolled sheets of graphene, have been widely investigated as a promising new material for many electrical device applications [5, 7] (e.g., transistor (CNT-FET), interconnects) as they exhibit exceptional electrical, thermal, and mechanical



Carbon Nanotubes for Chip Interconnections, Fig. 5 Graphical representations of ideal graphene sheet, SWCNT, MWCNT (Figure reprinted with permission from Graham et al. [22])

properties [8]. When comparing materials properties, CNTs are a clear winner against copper. Studies show that CNTs are stable for current densities up to 10^9 A/cm^2, two orders of magnitude higher than copper. CNTs can exhibit multichannel ballistic conduction over distances of microns. Because of their higher chemical stability relative to copper, diffusion barriers (liners) are not needed for CNTs, thus allowing a larger conductive cross section compared to copper for the same technology node. Additionally, their mechanical tensile strength (100 times that of steel) and their high thermal conductivity (comparable to diamond) give CNTs an edge compared to copper. Finally, growing CNTs in high aspect ratio vias could allow the design of chips with higher interlayer spacing to reduce overall RC losses and to decrease chip-layer energy dissipation [9].

Before examining possible models of CNT-based interconnect architectures, it is important to clearly understand CNTs' electrical properties, which represent the most critical material limitation to resolve with respect to copper. The electronic band structures of single-wall CNTs (SWCNTs) and of graphene are very similar. For graphene and metallic SWCNTs, the valence band and the conduction band touch at specific points in the reciprocal space. For semiconducting SWCNTs, the conduction band and the valence band do not touch. Semiconducting SWCNTs have been extensively studied as channels in transistor devices while metallic SWCNTs have been considered for applications such as IC interconnects and field emission.

The resistance of a CNT contacted at both ends is the sum of three resistances [5, 10]:

$$R_{CNT} = R_Q + R_L + R_{CONTACT}$$

where R_Q is the quantum resistance, R_L is the scattering resistance, and $R_{CONTACT}$ is the contact resistance. We will now discuss these three resistances.

An ideal (defect-free) metallic SWCNT electrically contacted at both ends, in the absence of scattering or contact resistance, exhibits a resistance $R = 2R_Q \approx 13\,\text{k}\Omega$ as a SWCNT has two conduction channels. The quantum resistance $R_Q = 6.5\,\text{k}\Omega$ is due to the mismatch between the number of conduction channels in the nanotube and the macroscopic metallic contacts. The one-dimensional confinement of electrons, combined with the requirement for energy and momentum conservation, leads to ballistic conduction over distances in the order of a micron.

The scattering resistance is due to impurities or nanotube defects that reduce the electron mean free path, and depends on the length l of the SWCNT:

$$R_L = \frac{1}{2R_Q}\left(\frac{l}{l_0}\right)$$

For defect-free SWCNT lengths below a micron, we can neglect the scattering resistance.

The contact resistance, which results from connecting the SWCNT to a contact (usually metallic), depends strongly on the material in contact with the nanotube, and on the difference between their work functions. The work functions of multiwall CNTs (MWCNTs) and SWCNTs have been estimated to be 4.95 and 5.10 eV, respectively [10]. Palladium has been found to be one of the materials minimizing the contact resistance, better than titanium or platinum contacts (which exhibited nonohmic behavior when in contact with CNTs) [10].

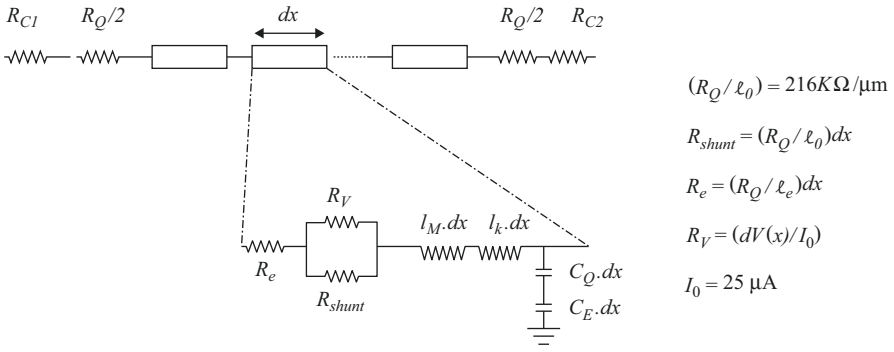
For interconnect applications, most often bundles of SWCNTs are considered. It is important to note that the coupling between adjacent SWCNTs is negligible since, for defect-free SWCNTs, the electrons would rather travel along the SWCNT axis (ballistic path) than across SWCNTs because of the large inter-CNT tunneling resistance (2–140 MΩ) [10]. Thus, the resistance of a bundle of SWCNTs can be viewed as a parallel circuit of the resistances of the individual SWCNTs. If we have n SWCNTs, the resistance of the bundle will be:

$$R_{SCWNT\ bundle} = \frac{R_{SCWNT}}{n}$$

The above overview related to SWCNTs. The electrical properties of MWCNTs have not been as extensively studied because of the additional complexities arising from their structure, as every shell has different electronic characteristics and chirality, in addition to interactions between the shells [11]. Geometrically, the interwall distance in a MWCNT is 0.34 nm, the same as the spacing between graphene sheets in graphite. What still has to be clarified is how the conductivity of a MWCNT varies with the number of walls.

Initially, it was thought that the conductance of a MWCNT occurred only through the most external wall, which seems to be the case at low bias and temperatures, where electronic transport is dominated by outer-shell conduction. However, theoretical models and experimental results indicate that shell-to-shell interactions can significantly lower the resistance of MWCNTs with many walls [5, 10].

One view is that the conductivity of a MWCNT with n walls is simply n times the conductivity of a SWCNT. Li et al. [12] experimentally measured an electrical resistance of only 34.4 Ω for a large MWCNT with outer diameter of 100 nm and inner diameter of 50 nm (= > 74 walls). This value was much lower than the one that could be calculated assuming all walls participated separately in the electrical conduction (i.e., calculated as the parallel of the resistances for each wall) showing that interwall coupling contributes to additional channels of conductance. Naeemi et al. [13] also assumed intercoupling between CNT walls in their models to increase the channels of conduction with increasing number of walls. However, their conductance was lower compared to that measured experimentally by Li's team.



Carbon Nanotubes for Chip Interconnections, Fig. 6 Equivalent circuit model of metallic SWCNT used in HSPICE simulations (Figure reprinted with permission from Naeemi and Meindl [14])

Bottom line: The conductivity of MWCNTs increases with the number of walls but the exact relationship has not yet been exactly clarified.

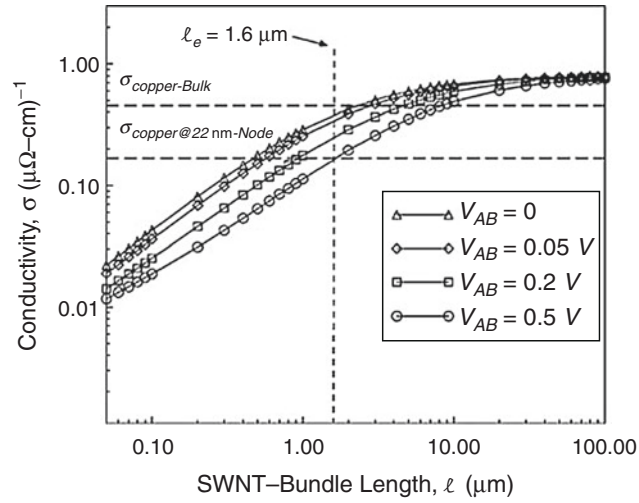
Models of CNTs as Interconnects

Various studies investigated replacing copper interconnects with bundles of CNTs (SWCNTs or MWCNTs) or with one large MWCNT. A first approach consists of using densely packed SWCNTs. Modeling SWCNTs as equivalent electrical circuits and using SPICE simulations, Naeemi et al. [14] showed that a target density of SWCNTs of at least 3.3×10^{13} CNTs/cm^2 was required. Currently achieved maximum densities for CNTs in vias barely reach 10^{12} CNTs/cm^2, which is still an order of magnitude smaller than required. Furthermore, since statistically only one-third of the SWCNTs grown are metallic (the other two-third are semiconducting), the conduction of the bundle will only occur in the metallic SWCNTs (Fig. 6).

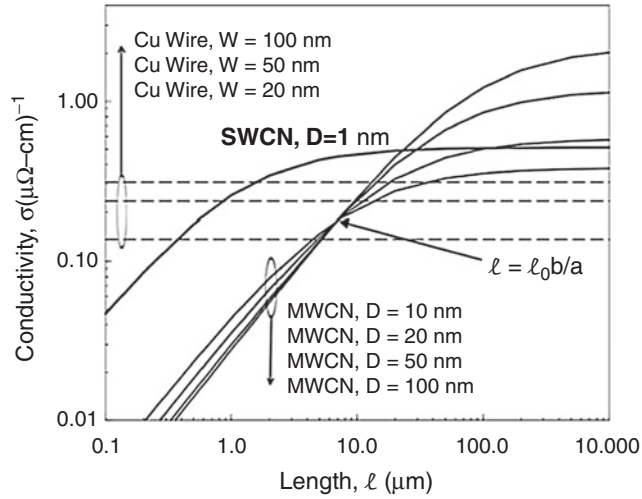
Naeemi et al. [14] also compared SWCNT bundles to copper as local, intermediate, and global interconnects. They showed that, in SWCNT bundles, resistance and kinetic inductance decreased linearly with the number of nanotubes in the bundle, while magnetic inductance changed very slowly. The resistance of a bundle of SWCNTs with sufficient metallic nanotubes was smaller than the resistance of copper wires, while capacitance was comparable. SWCNT bundles also fared better compared to copper in reducing power dissipation, delay, and crosstalk. For local interconnects, they quantified the improvements as 50 % reduction in capacitance, 48 % reduction of capacitance coupling between adjacent lines, and 20 % reduction in delay. For intermediate interconnects, the improvements were more marked, especially in terms of improved conductivities. For global interconnects, dense SWCNT bundles proved critical to improve bandwidth density (Fig. 7).

Using MWCNTs, which are all electrically conductive as they exhibit multiple channels of conduction (compared to only one-third metallic SWCNTs), could lower the resistivity of the bundle, although fewer of them can be packed in the same space because they usually have larger diameters (but also require a lower packing density compared to SWCNTs). In a different modeling study, Naeemi et al. [13] explored the suitability of MWCNTs as replacement for copper interconnects. They concluded that for long lengths (over 100 μm), MWCNTs have conductivities many times that of copper and even of SWCNT bundles. However, for short lengths (less than 10 μm), dense SWCNT bundles can exhibit a conductivity that is twice that of MWCNT bundles. Thus, for via applications, they recommended using dense bundles of SWCNTs or, alternatively, bundles of MWCNTs with small diameter (i.e., with few walls) (Fig. 8).

Using an individual MWCNT with large diameter could offer high conductivity due to the participation of multiple walls to significantly



Carbon Nanotubes for Chip Interconnections, Fig. 7 Conductivity of densely packed SWCNT bundles versus length for various bias voltages (Figure reproduced with permission from Naeemi and Meindl [14])



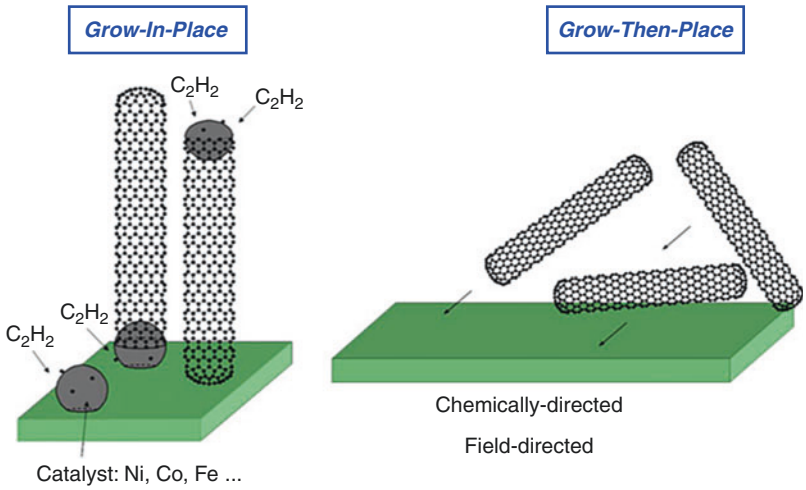
Carbon Nanotubes for Chip Interconnections, Fig. 8 Conductivity of MWCNTs with various diameters compared to Cu wires and dense bundles of SWCNTs (Figure reproduced with permission from Naeemi and Meindl [13])

increase the channels for conduction. However, as previously mentioned, the exact relationship between the number of walls and the conductance has yet to be clarified.

In conclusion, the choice of nanotubes may differ depending on the type of interconnect. For instance, dense bundles of SWCNTs or MWCNTs with few walls may be more suitable for small-section vertical vias, while dense bundles of larger MWCNTs may be more appropriate for long-range interconnects. The option of using a large MWCNT which fills all the space available needs to be further investigated. It is also plausible that hybrid systems of copper/SWCNTs/MWCNTs may be the best solution; for instance, small-section vertical vias may be replaced by dense SWCNT bundles, while larger long-range horizontal interconnects may still use copper, dense bundles of MWCNTs, or even metal-CNT composites.

Practical Implementation: Fabrication State of the Art and Outstanding Issues

Reliable and repeatable high-yield CNT fabrication compatible with CMOS standards is the main bottleneck in replacing copper in chip interconnections. Although hundreds of research teams have focused their efforts on nanotube growth and thousands of papers detailing growth recipes have been



Carbon Nanotubes for Chip Interconnections, Fig. 9 Pictorials comparing the "grow-in-place" and "grow-then-place" techniques (Reproduced with permission from Professor Carl V. Thompson [18])

published, surprisingly, very few have focused on the growth on conductive substrates at CMOS-compatible processing temperature [7, 10, 15, 16]. It is still a challenge to reliably and consistently synthesize CNTs on conductive layers at temperatures below 400–450 °C, the maximum temperature allowed in CMOS fabrication to avoid disrupting previous diffusion patterns. Furthermore, it is still difficult to precisely control CNT diameter and height, although chemical vapor deposition (CVD) from thin films of controlled thickness [17] or from nanoparticles [16] of controlled size has shown encouraging results.

To utilize carbon nanotubes in industrial applications, two main approaches have been considered: *"grow-in-place"* and *"grow-then-place"* [18] (Fig. 9).

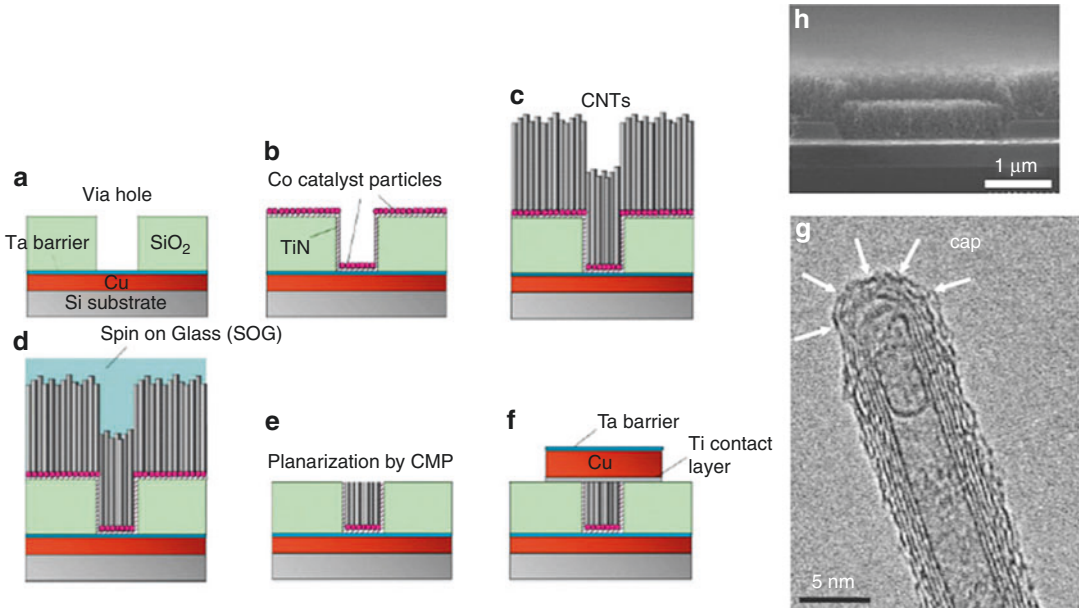
Grow-then-place: This technique consists of first preparing nanotubes and subsequently transferring them to a substrate. Arc discharge and laser ablation are the main techniques used to synthesize free-standing nanotubes. The nanotubes may be subsequently selected (e.g., separating SWCNTs or metallic SWCNTs) and purified prior to use. To transfer them to another substrate, CNTs are usually functionalized in a way that they will attach to pre-patterned areas of the substrate which will attract functionalized CNTs. An interesting technique for interconnect vias is based on using electrophoresis to push CNTs dispersed in a liquid solution into a matrix with pits (e.g., porous alumina matrix).

The advantages of this method are that it places no restrictions on the process or temperature used for CNT synthesis and allows to pretreat the CNTs (e.g., select, purify, functionalize). The major drawback is that no successful and repeatable technique to transfer the CNTs to the substrate has been developed to date. The challenge of resolving this issue appears too high to make this technique a candidate for the CMOS industry. However, free-standing, purified, CNTs are manufactured by many companies and sold for other applications (e.g., CNT-polymer composites).

Grow-in-place: this technique usually consists of preparing the sample with a catalyst present in the locations where the nanotubes will be synthesized. For instance, a thin catalyst film can be deposited using e-beam evaporation or sputtering; alternatively, nanoparticles can be deposited on a substrate. Synthesis is usually performed using thermal or assisted (e.g., plasma) CVD.

This method has several advantages: (1) good control of nanotube position (CNTs will grow where there are catalysts), (2) proven recipes to obtain crystalline CNTs (at least on insulating substrates), (3) proven capabilities to obtain carpets of vertically aligned CNTs, (4) physical contact with the substrate, (5) electrical contact with the substrate, and (6) CVD techniques are commonplace in the CMOS industry.

The major drawbacks are that (1) the processing temperature should be below 400–450 °C (CMOS-compatibility), thus putting



Carbon Nanotubes for Chip Interconnections, Fig. 10 Process to synthesize CNTs into pits, SEM cross section, and TEM showing crystalline MWCNT (Reprinted with permission from Yokoyama et al. [23])

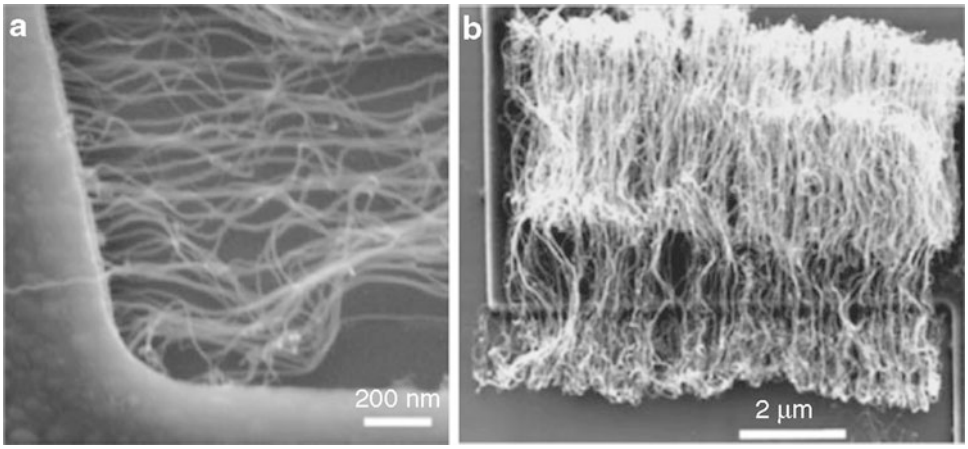
serious limits on the synthesis method and (2) the CNTs should be directly synthesized on the substrate of choice, usually a metallic layer to provide electrical contact. Although growing dense carpets of crystalline CNTs on insulating substrates such as alumina or silicon oxide has been achieved by many, CNT growth on metallic layers still remains a serious challenge. Interactions between the catalyst and the metallic substrate (e.g., alloying) are the major impediments for the successful growth of dense carpets of CNTs on metallic layers.

In addition to interesting results obtained from university research, good progress on the growth and characterization of carbon nanotubes for interconnects has been achieved by industrial laboratories, initially by Infineon and now by the Fujitsu laboratories. In 2002, Kreupl et al. [19] of Infineon showed that bundles of CNTs could be grown in pits of defined geometry.

More recently, Awano et al. [7] of Fujitsu grew bundles of MWCNTs in a 160 nm via at 450 °C and measured an electrical resistance of 34Ω (the CNT density observed was 3×10^{11} CNTs/cm^2). This follows a previous result obtained earlier by the same team where they grew MWCNTs into a 2 μm via at temperatures close to 400 °C with the lowest resistance measured of 0.6Ω after CMP and annealing in a hydrogen atmosphere [5] (Fig. 10).

Kreupl et al. [19] succeeded in growing a single MWCNT into a 25-nm hole and measured a high resistance of 20–30 kΩ. Given the difficulty of growing a single large MWCNT and the difficult task of making a precise electrical measurement, there may be room for further improvement if we could grow an individual, crystalline (defect-free) MWCNT with the maximum number of walls for a given external diameter, thus maximizing the number of channels of conduction.

Most of the effort on CNT synthesis to replace copper interconnects has focused on vertical growth of dense carpets of CNTs, which can be achieved by a high density of active catalyst dots. In contrast there have been fewer successful reports of horizontal growth, with less spectacular results. Many techniques have been used to achieve horizontal alignment among which we can mention high gas flow rates, electric fields, and epitaxial techniques to guide horizontal alignment of the nanotubes [10]. In 2010, Yan et al. [20] obtained an interesting horizontal



Carbon Nanotubes for Chip Interconnections, Fig. 11 Scanning electron microscope images of dense carpets of horizontally aligned CNTs grown using CVD (Figure reproduced from Yan et al. [20])

growth of bundled CNTs with a density of 5×10^{10} CNTs/cm^2, which is approaching what has been achieved for vertical CNT growth (although still over an order of magnitude lower compared to the best result for vertical CNT growth) (Fig. 11).

Although the experimental results obtained are encouraging, there are still numerous challenges that need to be resolved for CNTs to enter the semiconductor fab:

1. Increase the CNT areal density by one or two orders of magnitude. For SWCNTs, assuming all of them are metallic, a packing density of 10^{13}–10^{14} CNTs/cm^2 is required to compete with copper in terms of resistance, while for MWCNTs, the required packing density is lower and depends on the number of channels of conduction (i.e., number of walls). This will require, among other considerations, adequate catalyst and underlayer materials choice and deposition, possible surface pretreatment (e.g., plasma, reduction, and etching), maximum nucleation of active catalyst dots, and optimizing the CNT growth process. Although CNT areal density is an important issue, it may not be the dealbreaker.
2. Minimize the contact resistance between the CNTs and the substrate. To achieve maximum conductivity, the choice of the appropriate underlayer is critical; specific metals (e.g., Pd) and possibly silicides are good candidates. For MWCNTs, it is also important to ensure electrical contact with all the walls.
3. When using SWCNTs, synthesize only metallic SWCNTs (on average one-third of the SWCNTs grown) which are the ones participating in the electrical conduction. This is closely linked to the issue of chirality control, for which no solution has been proposed to date. Selective catalyst choice may provide an alternative avenue to synthesize a higher fraction of metallic SWCNTs.
4. Control growth direction of CNTs. This is especially challenging for horizontal interconnects where the directionality and the packing density achieved are still lagging compared to vertical growth of CNTs, despite some interesting progress in this area [10, 20].
5. Synthesize crystalline, defect-free CNTs to ensure maximum electrical conductivity in the nanotube. This is challenging, especially when combined with the requirement of growing CNTs at low temperature to achieve CMOS compatibility.
6. Synthesize CNTs at temperatures below 400–450 °C to ensure CMOS compatibility.
7. Repeatably yield the same CNT structures when the same process conditions are applied. This is a major issue since important variations in the structure and shape of the CNTs grown have been experimentally observed.

Conclusions and Outlook for CNTs as Chip Interconnections

In the past decade, the synthesis of CNTs and the understanding of their growth mechanisms have massively improved. However, for CNTs to enter the CMOS fab and replace copper, significant challenges still need to be resolved. In my opinion, the most significant challenge to overcome is developing a reliable and repeatable fabrication process consistent with CMOS conditions. To achieve that tall order, we need to improve our understanding of the CNT growth mechanisms. Although many simulation models and many experimentally-based insights have been achieved [10], there are still many questions related to CNT growth mechanisms that have not been fully answered. For instance:

- Which precursor gases favor CNT growth and which gases hinder CNT growth? What is the role of the gases in the resulting level of crystallinity of the CNTs grown? Could we pretreat the gases to improve the CNT yield or structure?
- What is the exact role of the catalyst? How does its materials properties and its lattice structure influence the resulting CNTs grown (in shape and structure)?
- What is the role of the underlayer (layer below the catalyst) and its interactions with the catalyst? Why is it so challenging to grow CNTs on metallic layers?

In addition to improving our mechanistic understanding of CNT growth, I believe that a parallel effort focused on developing better reactors is needed. Most researchers use standard CVD-based systems that were designed for a general purpose. A customized reactor, where the same growth conditions can be repeatably achieved with very small variations could provide the repeatability in results (CNT structures) that has eluded us so far.

Carbon nanotubes are already becoming a manufacturing reality in mechanical engineering applications (e.g., CNT-based composites) and many interesting results have been obtained to develop novel CNT structures for electrical applications. Although the jury is still out, if process repeatability could be achieved, we could hope, not only that CNTs will enter the CMOS fabs and replace copper for chip interconnections, but also that they will lead to innovative ventures requiring lower investments to develop integrated circuits with radically new architectural designs using carbon nanotubes as new building blocks.

Cross-References

- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes
- ▶ Chemical Vapor Deposition (CVD)
- ▶ CMOS-CNT Integration
- ▶ Nanotechnology
- ▶ Physical Vapor Deposition
- ▶ Synthesis of Carbon Nanotubes

References

1. Kuhn, K.J.: Moore's Law Past 32 nm: Future Challenges in Device Scaling. Intel Publication, Hillsboro (2009)
2. Magen, N., Kolodny, A., Weiser, U.: Interconnect-power dissipation in a microprocessor. In: Proceedings of the 2004 International Workshop, Paris, 1 Jan 2004
3. ITRS. International Technology Roadmap for Semiconductors – Interconnect 2009, International Sematech, Austin
4. Goel, A.K.: High-Speed VLSI Interconnections, 2nd edn. Wiley/IEEE, Hoboken (2007)
5. Nessim, G.D.: Carbon Nanotube Synthesis for Integrated Circuit Interconnects. Massachusetts Institute of Technology, Cambridge, MA (2009)
6. Havemann, R.H., Hutchby, J.A.: High-performance interconnects: an integration overview. Proc. IEEE **89**(5), 586–601 (2001)
7. Awano, Y., Sato, S., Nihei, M., Sakai, T., Ohno, Y., Mizutani, T.: Carbon nanotubes for VLSI: interconnect and transistor applications. Proc. IEEE **98**(12), 2015–2031 (2010)
8. Dresselhaus, M.S., Dresselhaus, G., Avouris, P. (eds.): Carbon Nanotubes: Synthesis, Structure, Properties, and Applications. Springer, Berlin (2001)
9. Chen, F., Joshi, A., Stojanović, V., Chandrakasan, A.: Scaling and evaluation of carbon nanotube interconnects for VLSI applications. In: Nanonets Symposium 07, Catania, 24–26 Sept 2007
10. Nessim, G.D.: Properties, synthesis, and growth mechanisms of carbon nanotubes with special focus on thermal chemical vapor deposition. Nanoscale **2**(8), 1306–1323 (2010)

11. Collins, P.G., Avouris, P.: Multishell conduction in multiwalled carbon nanotubes. Appl. Phys. **74**(3), 329–332 (2002)
12. Li, H.J., Lu, W.G., Li, J.J., Bai, X.D., Gu, C.Z.: Multichannel ballistic transport in multiwall carbon nanotubes. Phys. Rev. Lett. **95**(8), 086601 (2005)
13. Naeemi, A., Meindl, J.D.: Compact physical models for multiwall carbon-nanotube interconnects. IEEE Electron. Device Lett. **27**(5), 338–340 (2006)
14. Naeemi, A., Meindl, J.D.: Design and performance modeling for single-walled carbon nanotubes as local, semiglobal, and global interconnects in gigascale integrated systems. IEEE Trans. Electron Devices **54**(1), 26–37 (2007)
15. Nessim, G.D., Seita, M., O'Brien, K.P., Hart, A.J., Bonaparte, R.K., Mitchell, R.R., Thompson, C.V.: Low temperature synthesis of vertically aligned carbon nanotubes with ohmic contact to metallic substrates enabled by thermal decomposition of the carbon feedstock. Nano Lett. **9**(10), 3398–3405 (2009)
16. Awano, Y., Sato, S., Kondo, D., Ohfuti, M., Kawabata, A., Nihei, M., Yokoyama, N.: Carbon nanotube via interconnect technologies: size-classified catalyst nanoparticles and low-resistance ohmic contact formation. Phys. Status Solidi **203**(14), 3611–3616 (2006)
17. Nessim, G.D., Hart, A.J., Kim, J.S., Acquaviva, D., Oh, J.H., Morgan, C.D., Seita, M., Leib, J.S., Thompson, C.V.: Tuning of vertically-aligned carbon nanotube diameter and areal density through catalyst pre-treatment. Nano Lett. **8**(11), 3587–3593 (2008)
18. Thompson, C.V.: Carbon nanotubes as interconnects: emerging technology and potential reliability issues. In: 46th International Reliability Symposium; 2008: IEEE CFP08RPS-PRT, p. 368, 2008
19. Kreupl, F., Graham, A.P., Duesberg, G.S., Steinhogl, W., Liebau, M., Unger, E., Honlein, W.: Carbon nanotubes in interconnect applications. Microelectron. Eng. **64**(1–4), 399–408 (2002)
20. Yan, F., Zhang, C., Cott, D., Zhong, G., Robertson, J.: High-density growth of horizontally aligned carbon nanotubes for interconnects. Phys Status Solidi. **247** (11–12), 2669–2672 (2010)
21. Jackson, R.L., Broadbent, E., Cacouris, T., Harrus, A., Biberger, M., Patton, E., Walsh, T.: Processing and integration of copper interconnects. In: Solid State Technology. Novellus Systems, San Jose (1998)
22. Graham, A.P., Duesberg, G.S., Hoenlein, W., Kreupl, F., Liebau, M., Martin, R., Rajasekharan, B., Pamler, W., Seidel, R., Steinhoegl, W., Unger, E.: How do carbon nanotubes fit into the semiconductor roadmap? Appl. Phys. **80**, 1141–1151 (2005). Copyright 2005, Springer Berlin/Heidelberg
23. Yokoyama, D., Iwasaki, T., Yoshida, T., Kawarada, H., Sato, S., Hyakushima, T., Nihei, M., Awano, Y.: Low temperature grown carbon nanotube interconnects using inner shells by chemical mechanical polishing. Appl. Phys. Lett. **91**, 263101 (2007). Copyright 2007, American Institute of Physics

Carbon Nanotubes for Interconnects in Integrated Circuits

▶ Carbon Nanotubes for Chip Interconnections

Carbon Nanotubes for Interconnects in Microprocessors

▶ Carbon Nanotubes for Chip Interconnections

Carbon Nanowalls

▶ Chemical Vapor Deposition (CVD)

Carbon-Nanotubes

▶ Robot-Based Automation on the Nanoscale

Car–Parrinello Molecular Dynamics

Mauro Boero[1] and Atsushi Oshiyama[2]
[1]Institut de Physique et Chimie des Materiaux de Strasbourg (IPCMS), University of Strasbourg and CNRS, UMR 7504, Strasbourg, France
[2]Department of Applied Physics, The University of Tokyo, Tokyo, Japan

Synonyms

Ab initio molecular dynamics; CPMD; DFT-based molecular dynamics; First principles molecular dynamics

Definition

The Car–Parrinello molecular dynamics (CPMD) is an extension of the Lagrangian formalism of

classical molecular dynamics in which the model potential describing the interaction among atoms is replaced by the total energy functional of the system as provided by the Density Functional Theory (DFT). The electronic wavefunctions are explicitly introduced as new dynamical variables. The simultaneous Euler-Lagrange equations of motion for both sets of dynamical variables, atomic coordinates and electronic wavefunctions, avoid the explicit minimization of the DFT total energy at each step of the dynamics. Instead, they introduce a fictitious dynamics of the wavefunctions representing an adiabatic updating on-the-fly of the electronic structure along the atomic dynamics.

Introduction

The main target in atomic-scale simulations is to reproduce in a realistic way physical and chemical events occurring in materials. Specifically, the scope of First Principles Molecular Dynamics (FPMD) is to study a system of *interacting* nuclei and electrons by recreating it on a computer in a way as close as possible to nature and by simulating its dynamics over a *physical length of time* relevant to the properties of interest. The inherent complexity of the simulated systems, from solids to biological macromolecules, calls for methods able to go beyond the simple calculation of the electronic structure of a given set of coordinates $\mathbf{R}_I$ representing the positions of atoms. This is exactly the idea that started the entire field of Molecular Dynamics (MD).

From an historical point of view, the MD approach was introduced by Alder and Wainwright [1] in the late 1950s to study the interactions of hard spheres. Many important insights concerning the behavior of simple liquids emerged from their studies, but due to the limitations of the computational facilities and the pioneering stage of the MD, it was only in 1964 that the first dynamical simulation could be done. That milestone case focused on liquid Ar with the interatomic interaction modeled by a truncated Lennard-Jones potential [2]. In a nutshell, any MD method is an iterative numerical scheme for solving some equations of motion (EOM), representing the physical evolution of the system under study. Modeling the interaction of atoms with an analytic potential $V(\mathbf{R}_I)$, especially when chemical bonds evolve in time and they are broken or formed is a hard task not yet solved apart from a very limited class of chemical species. On the other hand, the electronic structure for a general many-body system can be determined with a computationally reasonable workload by means of the density functional theory (DFT), originally proposed in the early 1960s by Kohn, Hohenberg, and Sham [3, 4]. Its importance in the advancement of computational quantum chemistry and related fields was acknowledged by the Nobel Prize in Chemistry in 1998 awarded jointly to Walter Kohn and John A. Pople. Joining the two fields, MD and DFT, is exactly what the Car–Parrinello method is about, extending the range of both concepts [5, 6].

A Brief Overview of Density Functional Theory: The CPMD Potential

The DFT is a formulation of the many-body quantum mechanics in terms of an electron density distribution, $\rho(\mathbf{x})$, which describes the ground state of a system composed of interacting electrons and point-like nuclei having positions $\{\mathbf{R}_I\}$ [7]. All along the text, atomic units will be used for simplicity. In practice, single-particle wavefunctions $\psi_i(\mathbf{x})$ are used to express the many-body mathematical function ρ $(\mathbf{x})$. The dramatic simplification, is the fact that not even the specific analytic form of the complex function $\psi_i(\mathbf{x})$ matters, but only its square modulus, so that the electron density reads

$$\rho(\mathbf{x}) = \sum_{i=1}^{N^{occ}} f_i |\psi_i(\mathbf{x})|^2 \qquad (1)$$

This expression is clearly a single Slater determinant constructed from wavefunctions representing all the N^{occ} occupied orbitals. The coefficients f_i are the (integer) occupation numbers, and they are equal to 1 in the case in which the spin is explicitly

considered (spin-unrestricted) or equal to 2 if the spin is neglected and energy levels are considered as doubly-occupied (spin-restricted). Furthermore, the wavefunctions $\psi_i(\mathbf{x})$ are subject to the orthonormality constraint

$$\int \psi_i^*(\mathrm{x})\psi_j(\mathrm{x})\, d^3x = \delta_{ij} \tag{2}$$

as in any quantum mechanics approach. The Kohn-Sham (KS) DFT total energy of the system in its ground state is then written as

$$E^{KS}[\{\psi_i\}] = E_k[\{\psi_i\}] + E_H[\rho] + E_{xc}[\rho] + E_{eI}[\rho] + E_{II} \tag{3}$$

In Eq. 3, the first three terms on the right-hand side (E_k, E_H, E_{xc}) describe all the electron–electron interactions, the fourth term (E_{eI}) refers to the electron–nucleus interaction, and the fifth one (E_{II}) corresponds to the nucleus–nucleus interaction. More explicitly, E_k is the Schrödinger-like kinetic energy expressed in terms of the single-particle wavefunctions $\psi_i(\mathbf{x})$ as

$$E_k[\{\psi_i\}] = \sum_{i=1}^{N^{occ}} f_i \int \psi_i^*(\mathbf{x}) \left(-\frac{1}{2}\nabla^2\right) \psi_i(\mathbf{x})\, d^3x \tag{4}$$

It must be remarked that this expression for the kinetic energy does not depend on the density $\rho(\mathbf{x})$ but directly on the wavefunctions. The second term, E_H, is the Hartree energy, i.e., the Coulomb electrostatic interaction between two charge distributions

$$E_H[\rho] = \frac{1}{2}\int\int \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} d^3x\, d^3y \tag{5}$$

The exchange interaction and the electron correlations due to many-body effects are represented by the term $E_{xc}[\rho]$, whose exact analytical expression is unknown. There are good approximations derived from the homogeneous electron gas limit for the exchange interaction [7], the so-called local density approximation (LDA), whose name comes from the fact that a homogeneous distribution of interacting electrons is assumed, in which $\rho(\mathbf{x})$ depends just on the local point $\mathbf{x}$. Similarly, in the LDA version of the correlation energy [7], the explicit analytic form of the functional comes from a parameterization of the results of quantum Monte Carlo calculations. Due to the insufficiency of a simple LDA approximation for many real systems, nonlocal approximations, including the gradient of the density, are often adopted and the exchange-correlation functional becomes $E_{xc}[\rho,\ \nabla\rho]$. In practical applications, however, the gradient enters only with its modulus, thus adding only a modest computational cost. The electrostatic interaction between electrons and nuclei, is then

$$E_{eI}[\rho] = -\int \sum_{I=1}^{M} \frac{Z_I \rho(\mathbf{x})}{|\mathbf{x}\text{-}\mathbf{R}_I|} d^3x \tag{6}$$

where Z_I is the charge of the Ith nucleus. However, in practice, this expression "as is" is computationally expensive. In fact, two different length scales come into play: a small one for the core electrons, characterized by rapidly varying wavefunctions, especially in the region very close to the nucleus, and a longer one for the valence electrons that form chemical bonds and vary more smoothly. Clearly, the first one would dominate and add a computational workload that would make impractical simulations of large systems. To overcome this problem, one can observe that core electrons are generally inert and do not participate to chemical bonds. This crucial observation led to the use of pseudopotentials [6]. Namely, core electrons are eliminated and a potential describing the core–valence interaction is built by fitting to the all-electron solutions of the Schrödinger or Dirac equation for the single atom of the chemical species considered. In a pseudopotential (PP) approach, the electron–nucleus interaction is rewritten as

$$E_{eI}[\rho] = \int d^3x V_{ps}(\mathbf{x} - \mathrm{R}_I) \cdot \rho(\mathbf{x}) \tag{7}$$

Finally, the fifth and last term in right-hand side of Eq. 3 is simply the Coulomb interaction between two classical nuclei I and J and is written as

$$E_{II} = \sum_{I<J}^{M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \tag{8}$$

where Z_I and Z_J are the net valence charge in a PP approach. The total energy E^{KS} of the ground state of such a system of interacting electrons and nuclei is obtained by minimizing the KS functional with respect to the single-particle orbitals $\psi_i(\mathbf{x})$, which, in practice, means solving the KS Schrödinger-like equations

$$\frac{\delta E^{KS}}{\delta \psi_i^*} \equiv H^{KS}\psi_i(\mathbf{x}) = \varepsilon_i \psi_i(\mathbf{x}) \tag{9}$$

The Basis Set Issue

A somehow arbitrary issue is the proper definition of $\psi_i(\mathbf{x})$. The answer is the selection of a proper basis set on which orbitals can be expanded. One possible choice is a localized basis set expressing the one-electron wavefunctions as

$$\psi_i(\mathbf{x}) = \sum_{k=1}^{M} c_i^k \phi_k(\mathbf{x}; \{\mathbf{R}_I\}) \tag{10}$$

and the number of analytic functions used, M, is also an indicator of the computational cost of the quantum calculation, in the obvious sense that the larger the basis set, the higher the computational workload. One of the most popular basis sets is represented by Gaussian-type orbitals (GTO)

$$\phi_k(\mathbf{x} - \mathbf{R}_I) = \phi_k(\mathbf{r}) = N_k \cdot r_x^{k_x} r_y^{k_y} r_z^{k_z} \cdot \exp\left(-\alpha_k \cdot r^2\right) \tag{11}$$

where $\mathbf{r} = \mathbf{x} - \mathbf{R}_I$. When such a basis set is used, the constants $N_\mathbf{k}$, and $\alpha_\mathbf{k}$ are kept fixed during the electronic structure calculation, whereas the coefficients c_i^k are allowed to vary until they are fully optimized [8]. It must be remarked that orbitals expanded in a localized basis set depend on the atomic positions $\mathbf{R}_I$. As a consequence, in any calculation in which the forces acting on the ions are required, the explicit derivatives of these wavefunctions with respect to $\mathbf{R}_I$ must be computed, leading to non-Hellmann-Feynman force components known in the literature as Pulay forces [6, 8]. An alternative basis set rather popular in physics is represented by plane waves (PW)

$$\psi_i(\mathbf{x}) = \sum_{\mathrm{G}=0}^{\mathrm{G}^{\max}} c_i(\mathbf{G})\, e^{i\mathbf{G}\cdot\mathbf{x}} \tag{12}$$

where the sum is truncated at a suitable cut-off $\mathbf{G}^{\max}$. In this case, no dependence on the atomic coordinates and no arbitrariness in the increase in the number of basis functions exist.

First Principles Molecular Dynamics

Until the early 1980s, few applications of DFT went beyond the static calculations of the electronic structure. Nonetheless, finite temperature and entropy effects are two of the dominant features in matter and their role is often far from negligible. In this respect, the FPMD has represented a huge step forward. In this particular combination of DFT and MD, the interactions among atoms, instead of being described by an analytical function $V(\{\mathbf{R}_I\})$ of the atomic coordinates $\mathbf{R}_I$, is directly computed from the DFT total energy E^{KS}, which is simultaneously a function of the electron wavefunctions and the atomic coordinates. The Born–Oppenheimer (BO) approximation [9] allows to disentangle the motions of the electrons and the nuclei, and each time the nuclei $\mathbf{R}_I(t)$ are displaced from given positions at time t to new positions $\mathbf{R}_I(t + \Delta t)$ at a subsequent time $t + \Delta t$, an optimization of the electronic structure has to be performed. Then the forces acting on the nuclei are estimated from the gradient of E^{KS} with respect to the ionic position and the variables $\mathbf{R}_I(t)$ are updated to $\mathbf{R}_I(t + \Delta t)$ by solving via finite differences the Newton-like EOM

$$M_I \ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} \min_{\{\psi_i\}} E^{KS}[\{\psi_i\}, \{\mathbf{R}_I\}] \tag{13}$$

This iterative procedure assumes that the electronic structure is recomputed and the full diagonalization of the Hamiltonian is performed at each time step t along the discrete trajectory $\{\mathbf{R}_I(t)\}$.

Car–Parrinello Molecular Dynamics

An alternative to this scheme, which has represented a real breakthrough in first principles dynamical simulations, was proposed in 1985 by R. Car and M. Parrinello [5]. The scope (and driving force) was to overcome the two major efforts arising in FPMD: On one hand one has to integrate the equations of motion for the nuclear positions as in Eq. 13, which represent the long timescale part to the problem. On the other hand, one has to propagate dynamically the smooth time-evolving (ground state) electronic subsystem. The Car–Parrinello molecular dynamics (CPMD) is able to satisfy this second requirement in a numerically stable way and makes an acceptable compromise for the time step length of the nuclear motion. The formulation is an extension of a classical molecular dynamics Lagrangian in which the electronic degrees of freedom (wavefunctions) are added to the system, along with any other dynamical variable $q_\alpha(t)$, i.e., thermostats, barostats, reaction coordinates, etc.

$$\mathcal{L}^{CP} = \frac{1}{2}\sum_I M_I \dot{\mathbf{R}}_I^{\,2} + \sum_i \mu \int \left|\dot{\psi}_i(\mathbf{x})\right|^2 d^3x + \frac{1}{2}\sum_\alpha \eta_\alpha \dot{q}_\alpha^2 - E^{KS}[\rho, \{\mathbf{R}_I\}, q_\alpha] + \sum_{ij}\lambda_{ij}\left(\int d^3x \psi_i^*(\mathrm{x})\psi_j(\mathrm{x}) - \delta_{ij}\right) \quad (14)$$

The first term on the right-hand side of Eq. 14 is the kinetic energy of the nuclei, the second one the fictitious kinetic energy of the electrons representing the update of the wavefunctions during the dynamics, the third one the kinetic term of any further dynamical variable used in the sense specified above, the fourth one is the DFT total energy, and the last addendum is the orthonormality constraint for the wavefunctions. The kinetic energy for the electronic degrees of freedom is the main novelty of the CPMD approach: A strategy to update on-the-fly the wavefunctions when ions undergo a dynamical displacement, avoiding expensive iterative diagonalization required by the BO approach at each time step. The Euler–Lagrange EOM are as follows:

$$\mu\ddot{\psi}_i(\mathbf{x}) = -\frac{\delta E^{KS}}{\delta \psi_i^*} + \sum_j \lambda_{ij}\psi_j(\mathbf{x}) \quad (15)$$

$$M_I\ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} E^{KS} \quad (16)$$

$$\eta_\alpha \ddot{q}_\alpha = -\frac{\partial E^{KS}}{\partial q_\alpha} \quad (17)$$

The fictitious mass μ assigned to the orbitals $\psi_i(\mathbf{x})$ is the parameter that controls the speed of the updating of the wavefunctions with respect to the nuclear positions. For this reason, it determines the degree of adiabaticity of the two subsystems, electrons and nuclei.

It is straightforward to give a Hamiltonian, instead of a Lagrangian, formulation of the CPMD method, via a simple Legendre transform after defining the momenta

$$\pi_i(\mathbf{x}) = \frac{\delta \mathcal{L}^{CP}}{\delta \dot{\psi}_i}(\mathbf{x}) = \mu\dot{\psi}_i(\mathbf{x})$$

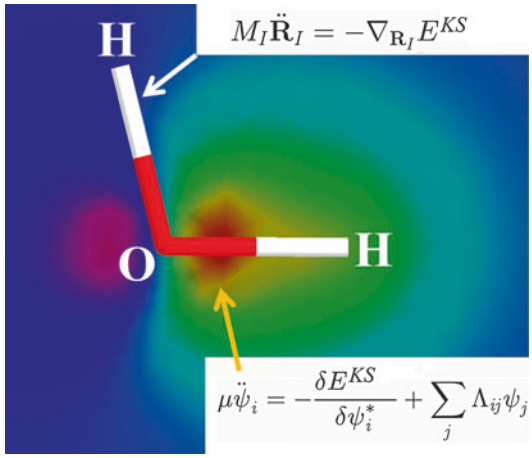
$$\pi_i^*(\mathbf{x}) = \frac{\delta \mathcal{L}^{CP}}{\delta \dot{\psi}_i(\mathbf{x})} = \mu\dot{\psi}_i^*(\mathbf{x}) \quad (18)$$

$$\mathbf{p}_i = \nabla_{\dot{\mathbf{R}}_I}\mathcal{L}^{CP} = M_I\dot{\mathbf{R}}_I \quad (19)$$

$$\dot{\zeta}_\alpha = \frac{\partial \mathcal{L}^{CP}}{\partial \dot{q}_\alpha} = \eta_\alpha \dot{q}_a \quad (20)$$

so that the Hamiltonian reads

$$H^{CP} = \sum_I \frac{\dot{\mathbf{p}}_I^{\,2}}{2M_I} + \sum_i \int \frac{\pi_i^*(\mathbf{x})\,\pi_i(\mathbf{x})}{\mu} d^3x + \sum_\alpha \frac{\dot{\zeta}_\alpha^2}{2\eta_\alpha} - E^{KS}[\rho, \{\mathbf{R}_I\}, q_\alpha] - \sum_{ij}\lambda_{ij}\left(\int \psi_i^*(\mathbf{x})\psi_{\mathrm{j}}(\mathbf{x})\, d^3x - \delta_{ij}\right) \quad (21)$$



Car–Parrinello Molecular Dynamics, Fig. 1 Atomic structure (sticks; *red* = O, *white* = H) of a water molecule and electronic wavefunction of an O–Hs-bond in terms of density map (*blue*: $|\psi_i(\mathbf{x})|^2 = 0$, *red*: $|\psi_i(\mathbf{x})|^2 =$ maximum). The two sets of dynamical variable evolve in time according to the equations of motions indicated, coupled via E^{KS}

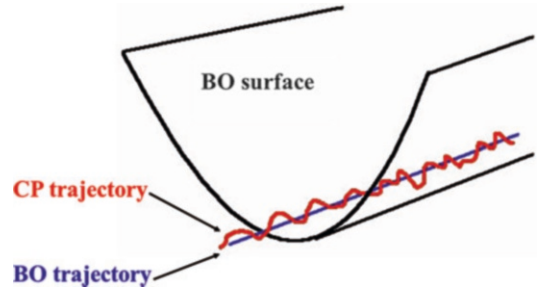
and the CPMD equations of motion (Fig. 1) will be given by the corresponding Hamilton EOMs.

A rigorous mathematical proof of this scheme has been given by Bornemann and Schütte [10], showing that the CPMD trajectory $\{\mathbf{R}^{\mathrm{CP}}(t)\}$ stays close to the BO one $\{\mathbf{R}^{\mathrm{BO}}(t)\}$ and the upper bound is proportional to the square root of the fictitious mass μ

$$\left|\mathrm{R}^{CP}(t) - \mathrm{R}^{BO}(t)\right| < C \cdot \sqrt{\mu} \qquad (22)$$

where C is a positive constant (Fig. 2).

The fact that the CPMD was a milestone step forward in realistic simulations of materials at various thermodynamics conditions can be easily seen by the number of publications in first principles molecular dynamics (FPMD) before and after 1985. Indeed, the original Car–Parrinello publication has more than 7,000 citations in 2014 (source: ISI Web of Science), and, to acknowledge the importance of the method, the international Physics and Astronomy Classification Scheme (PACS) introduced in 1996 a new identification number, 71.15.Pd, to classify Car–Parrinello related publications. Since then, the method has been applied to a wide variety of materials, ranging from solids, to liquids, and to biological systems [11, 12].



Car–Parrinello Molecular Dynamics, Fig. 2 Schematic representation of a Car–Parrinello trajectory (*red line*) with respect to a Born–Oppenheimer dynamics (*blue line*) one a given DFT-based potential energy surface

Numerical Details

Although it is not restriction neither of DFT [13] nor of CPMD [14], PWs are often used as a convenient basis set for the coding of CPMD, since they have several good properties: (i) Accuracy can be systematically improved in a fully variational way, (ii) PWs are independent from atomic positions (i.e., no Pulay forces [6]), (iii) PWs can be easily distributed in parallel processing. However, it must be observed that the fact that PWs are not localized can lead to inefficiencies for small clusters or surfaces placed in a large simulation cell. The equations of motion are discretized by finite differences, via a Verlet or velocity-Verlet algorithm [15]. The ionic variables $\mathbf{R}_I$(t) are updated at a rate Δt, while the electronic degrees of freedom are updated at a rate $\Delta t/\mu^{1/2}$, i.e.,

$$\begin{aligned}\frac{\mu}{\Delta t^2}\left[c_i(\mathbf{G}, t+\Delta t) + (\mathbf{G}, t-\Delta t) - (\mathbf{G}, t)\right] \\ = -\sum_{G'} \left\langle \mathbf{G}\middle| H^{CP} \middle|\mathbf{G}'\right\rangle c_i(\mathbf{G}') \\ + \sum_j \Lambda_{ij} c_j(\mathbf{G})\end{aligned} \qquad (23)$$

For most of the applications, Δt and μ fall in the range 3–5 au and 300–600 au, respectively. Of course, the (quantum) time scale of electrons is dominating in this kind of approaches and simulations times are of the order of few tens or, at very best, hundreds of ps. As far as the system size is

concerned, with N electrons and $\mathbf{G}^{\max}$ PWs, $\mathbf{G}^{\max}$ being integer, the scaling of the various parts composing the CPMD algorithm is $O(N\ \mathbf{G}^{\max})$ for the kinetic term, $O(N\ \mathbf{G}^{\max}\ \log\ \mathbf{G}^{\max})$ for the local potential and $O(N^2\ \mathbf{G}^{\max})$ for both the nonlocal term and wavefunctions orthogonalization procedure.

Second-Generation Car–Parrinello Molecular Dynamics

The inherent high computational cost associated to the electronic structure calculations has limited the affordable timescales for several years. Only the most advanced high performing computer platforms recently available have allowed to increase the system size to about a thousand of atoms and simulation times towards hundreds of picosecond. Yet, many phenomena still call for a substantial boost. These are, for instance, diffusion in solids or, in the case of glasses generation from the melt, a less rapid cooling rate suitable to avoid numerically induced structural problems. While linear scaling methods can be a viable way to access larger system sizes, they still have to face the problem of the simulation timescale. Moreover, the crossover point at which linear scaling methods become advantageous has remained fairly large, especially if high accuracy is needed. An interesting attempt at overcoming these limitations has been proposed in 2007 [16]. The basic idea is to join the advantages of both the BO approach and the CPMD; in a "nutshell"

	CPMD	BO
Conservation of constants of motion	**Good**	Convergence dependent
Electronic optimization	**Not needed**	Needed
Hamiltonian diagonalization	**Not needed**	Needed
Integration step Δt	Small	**Large**
Minimum of the BO surface	Approximate	**Exact**

These two approaches have nearly complementary features as sketched above. Following the CPMD formulation, it can be remarked that the Lagrangian formulation for the propagation of the wavefunctions is stable by construction, thus providing a reliable integration. This *stability* feature must then be preserved. Concerning the *efficiency*, large integration steps Δt are desirable and possibly a small, or better zero, deviation from the BO surface should be kept all along the dynamics to get a high *accuracy*. The mathematical result of this list of requirement resumes into a modified ionic EOM

$$M_I\ddot{\mathbf{R}}_I = -\frac{\partial \mathrm{E}_{NSC}}{\partial \mathbf{R}_I} + \sum_{i,j}\Lambda_{ij}\frac{\partial}{\partial \mathbf{R}_I}\left\langle \psi_i \middle| \psi_j \right\rangle - 2\frac{\partial \langle \psi_i|}{\partial \mathbf{R}_I}\left(\frac{\delta \mathrm{E}_{NSC}}{\delta \langle \psi_i|} - \sum_j \Lambda_{ij}|\psi_j\rangle\right) \tag{24}$$

While the first term in the right hand of the equation is clear, the rest seems a bit puzzling at a first glance. Indeed, in the original formulation [16], the selected basis set is not PW, but a localized basis set as in Eq. 10. Hence, the electronic wavefunctions depend also on the atomic coordinates $\mathbf{R}_I$ and the request of orthogonality at each step is released to save time, meaning that the scalar product $<\psi_i|\psi_j>$ is no longer vanishing. Analogously, the total energy E^{DFT} is not reoptimized as in full self-consistent BO procedures and for this reason is indicated as non-self-consistent energy E_{NSC}. Wavefunctions are propagated according to an algorithm which resembles the original CPMD formulation in the sense that second order EOMs are used, but a damping term (first-order derivative) is present which reminds a sort of steepest-descent algorithm typical of the BO dynamics. The net result is the electron dynamics,

$$\mu\frac{d^2}{dt^2}|\psi_i\rangle + \gamma\frac{d}{dt}|\psi_i\rangle = -\frac{\delta E_{NSC}}{\delta\langle\psi_i|} + \sum_j \Lambda_{ij}|\psi_i\rangle \tag{25}$$

which is then solved via a predictor–corrector scheme. With no pretention of completeness, the procedure can be summarized as follows. On a

first instance, in a localized basis set $\{|q>\}$, the electronic wavefunctions are expanded as

$$|\psi_i\rangle = \sum_{q=1}^{M} C_{iq}|q\rangle \tag{26}$$

on the M functions composing the localized basis. Then the NxM matrix of the expansion coefficient is written as

$$\mathbf{C} = [C_{iq}] = \begin{pmatrix} C_{11} & . & . & C_{1M} \\ . & . & & . \\ . & & . & . \\ C_{N1} & . & . & C_{NM} \end{pmatrix} \tag{27}$$

and the density matrix becomes $\mathbf{P} = \mathbf{C}\mathbf{C}^T = \mathbf{PSP}$. The MxM matrix indicated as $\mathbf{S}$ is just given by the expansion coefficient and its matrix elements have the usual form

$$S_{qq'} = \sum_{i=1}^{N} C_{iq}^{*} C_{iq} \tag{28}$$

Hence, the (DFT) total energy can be rewritten as $E^{\text{tot}}[\mathbf{C},\mathbf{R}_I]$ which can be used in a straightforward way to write a BO dynamics

$$M\ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} \min\{E^{\text{tot}}[\mathbf{C},\mathbf{R}_I]\} \quad \forall \mathbf{C}^T \cdot \mathbf{S} \cdot \mathbf{C} = \mathbf{1} \tag{29}$$

under the given constraint on C which resumes in an implicit orthogonality condition. However it must be kept into account that: (i) Diagonalization and minimization of E^{tot} are required in BO; (ii) Hellman-Feynman forces are just one component since Pulay forces due to the local basis set are present. Residual force components appear due to non-self-consistency (NSC) of the approach. To take into account all the points above, the basic strategy can be summarized in four major points:

1. Propagate the electronic variables in time according to the CP original idea of updating on-the-fly to avoid expensive full diagonalization operations
2. Use a good propagation algorithm $\mathbf{C}(t_n) = f(\mathbf{C}(t_{n\text{-}1}),\ldots,\mathbf{C}(t_{n\text{-}m}))$ depending on previous time steps $m \in [1,K]$ time steps
3. Select the appropriate number of steps K to keep $\mathbf{C}(t_n)$ as close as possible to the (electronic) ground state
4. Enforce convergence on the BO surface, correct this propagation CPMD-like afterwards

Point 3 corresponds to the first move in the numerical integration procedure and it can be identified as the "predictor" part directly deriving from a standard numerical integration of the CPMD type equations of motion. Point 5, instead, is the "corrector" needed afterwards to better converge the wavefunctions and to restore the neglected self-consistent loop. The use of not necessarily fully converged wavefunctions at the predictor propagation stage allows for large integration steps, thus resulting in a remarkable boost in the dynamics.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computer Modeling and Simulation of Materials
- ▶ Electronic Structure Calculations
- ▶ First Principles Calculations
- ▶ Molecular Dynamics Method
- ▶ Molecular Dynamics Simulations of Interactions Between Biological Molecules and Nanomaterials
- ▶ Molecular Dynamics Simulations of Nanobiomaterials
- ▶ Surface Electronic Structure

References

1. Alder, B.J., Wainwright, T.E.J.: Phase transition for a hard sphere system. Chem. Phys. **27**, 1208 (1957)
2. Rahman, A.: Correlation in the motion of atoms in liquid argon. Phys. Rev. **136**, A405 (1964)
3. Hohenberg, P., Kohn, W.: Inhomogeneous electron gas. Phys. Rev. **136**, B864 (1964)
4. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, A1133 (1965)
5. Car, R., Parrinello, M.: Unified approach for molecular dynamics and Density-Functional theory. Phys. Rev. Lett. **55**, 2471 (1985)
6. Marx, D., Hutter, J.: Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods. Cambridge

University Press, New York (2009). ISBN 978-0521898638
7. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, New York (1989). ISBN 0-19-504279
8. Hehre, W.J., Radom, L., Schleyer, P.V.R., Pople, J.A.: Ab Initio Molecular Orbital Theory. Wiley, New York (1986). ISBN 978-0471812418
9. Born, M., Oppenheimer, J.R.: Zur Quantentheorie der Molekeln. Ann. Phys. **84**, 457 (1927)
10. Bornemann, F.A., Schütte, C.: A mathematical investigation of the Car-Parrinello method. Numer. Mat. **78**, 359 (1998)
11. Boero, M., Tateno, M.: Quantum theoretical approaches to proteins and nucleic acids. In: Oxford Handbook of Nanoscience and Technology, Vol. 1: Basic Aspects, pp. 549–598. Oxford University Press, New York (2010). ISBN 978-0199533046
12. Boero, M.: Reactive simulations for biochemical processes. In: Atomic-Scale Modeling of Nanosystems and Nanostructured Materials. Lecture Notes in Physics, vol. 795, pp. 81–98. Springer, Berlin/Heidelberg (2010). ISBN 978-3-642-04650-6
13. Oshiyama, A., Iwata, J.: Large-scale electronic-structure calculations for nanomaterials in density functional theory. J. Phys. Conf. Ser. **302**, 012030 (2011)
14. Car, R., Parrinello, M.: The unified approach to density functional and molecular dynamics in real space. Solid State Commun **62**, 403 (1987)
15. Pang, T.: An Introduction to Computational Physics. Cambridge University Press, New York (2000). ISBN 978-1840858839
16. Kühne, T.D., Krack, M., Mohamed, F.R., Parrinello, M.: Efficient and accurate Car-Parrinello-like approach to Born-Oppenheimer molecular dynamics. Phys. Rev. Lett. **98**, 066401 (2007)

Carrier-Free Electrophoresis

▶ Micro Free-Flow Electrophoresis (μFFE)

Catalyst

▶ Chemical Vapor Deposition (CVD)
▶ Physical Vapor Deposition

Catalytic Bimetallic Nanorods

▶ Molecular Modeling on Artificial Molecular Motors

Catalytic Chemical Vapor Deposition (CCVD)

▶ Chemical Vapor Deposition (CVD)

Catalytic Janus Particle

▶ Molecular Modeling on Artificial Molecular Motors

Cathodic Arc Deposition

▶ Physical Vapor Deposition

Cavity Optomechanics

▶ Nano-optomechanical Systems (NOMS)

Cell Adhesion

▶ Bioadhesion
▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Cell Adhesion and Detachment

▶ Biological Breadboard Platform for Studies of Cellular Dynamics

Cell Micro-patterning

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Cell Migration

▶ Biological Breadboard Platform for Studies of Cellular Dynamics

Cell Patterning

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Cellular and Molecular Toxicity of Nanoparticles

► Cellular Mechanisms of Nanoparticle Toxicity

Cellular Dynamics

► Biological Breadboard Platform for Studies of Cellular Dynamics

Cellular Electronic Energy Transfer

► Micro/Nano Transport in Microbial Energy Harvesting

Cellular Imaging

► Electrical Impedance Tomography for Single-Cell Imaging

Cellular Mechanisms of Nanoparticle Toxicity

Francelyne Marano, Fernando Rodrigues-Lima, Jean-Marie Dupret, Armelle Baeza-Squiban and Sonja Boland
Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), Paris cedex 13, France

Synonyms

Cellular and molecular toxicity of nanoparticles

Definition

The interaction between nanoparticles and cell triggers a cascade of molecular events which could induce a toxicity and cell death. They are associated with the uptake of nanoparticles, their persistence at cellular level, and their ability to release free radicals and to induce an oxidative stress. The resulting activation of molecular pathways and transcription factors could lead to a pro-inflammatory response or, depending on the level of free radicals, apoptosis.

Background

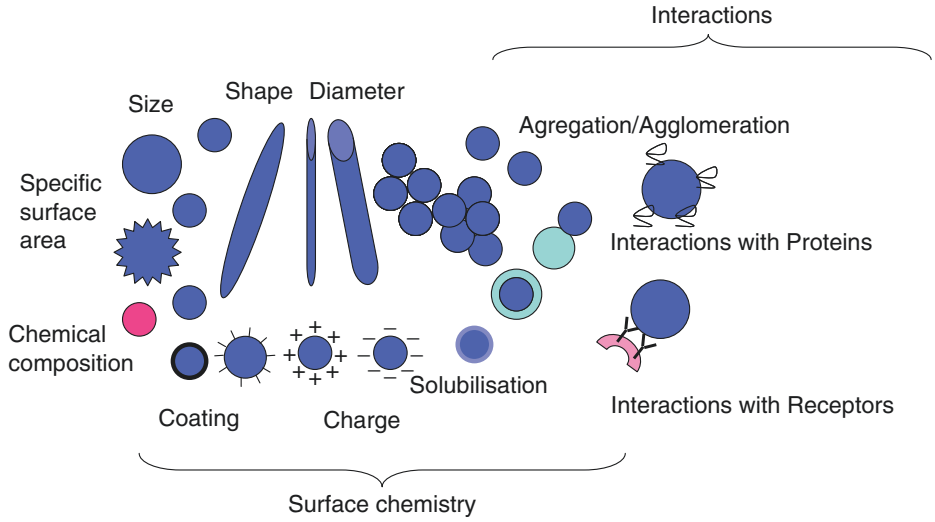
The last 5 years have shown an increasing number of papers on the mechanisms of nanoparticle (NP) cytotoxicity. What are the reasons? It is likely that the specific useful properties which appear at nanoscale can also lead to adverse effects. This hypothesis is strongly supported by in vivo and in vitro studies to compare the toxicity of NPs with their fine counterparts of the same chemical composition. These results have clearly demonstrated a higher toxicity of particles at nanoscale than at microscale. Moreover, it appears from experimental studies that solid nano-sized particles could be translocated beyond the respiratory tract and could induce a systemic response. The interstitial translocation of a same mass of particles is higher for ultrafine than fine particles after intratracheal instillation in rats [1]. Surface area, which is strongly increased for NPs compared to micro-particles of same chemical composition, and surface reactivity are considered as the principal indicators of NP reactivity. It was shown that a toxic response could be observed even to apparently nontoxic substances when the exposure occurred in the nanometer size range. All these observations have led to the development of a new field of toxicology, nanotoxicology [2]. However, the toxicological mechanisms which sustained the biological response are not yet clear and a matter of debates.

The concerns about the toxicity of engineered nanoparticles, which are increasingly used for industrial and medical applications, came also

from the knowledge on the toxicity of non-intentional atmospheric particles. Short-term epidemiological studies in Europe and North America have showed an association between cardiorespiratory morbidity and mortality and an increased concentration of atmospheric fine particles [3]. Moreover, the long-term epidemiological studies have also demonstrated an association between exposure to atmospheric particles (particulate matter or PM10 and 2.5) and increased cancer risk [3]. In parallel, in vitro and in vivo studies on fine and ultrafine airborne particles such as diesel exhaust particles, PM2.5, gave causal explanations to these adverse health effects (reviewed in Ref. [1]). They allow defining the molecular events induced by these particles in lung cells. The major event is a pro-inflammatory response which is characterized by the release of various cytokines (pro-inflammatory mediators), associated with the activation of transcription factors and signaling pathways. This was especially demonstrated for diesel exhaust particles (DEP), a major component of urban PM in Europe. These events are mostly induced by organic components of the DEP and are probably mediated by the generation of reactive oxygen species (ROS) during the metabolism of organic compounds (for a review, see Ref. [4]). These findings were used as a background for the researches on biological mechanisms induced by NPs considering that fine and ultrafine atmospheric particles have great similarities with NPs, especially diesel exhaust particles which are of nano size and aggregate after their release in the atmosphere.

It became rapidly obvious that the understanding of the cellular and molecular mechanisms leading to the biological effects of NPs was essential for the development of safe materials and accurate assays for risk assessment of engineered NPs [2], and several recent reviews were focused on demonstrated or hypothetic cellular mechanisms of these responses [4–6].

The first event, when NPs enter in contact with the human body by inhalation, oral and dermal exposure, or intravenous application, is their interaction in the biological fluids and the cellular microenvironment with biological molecules such as proteins thus forming a protein corona [7]. Consequently, NPs do not directly interact with the cell membrane but through the protein and/or lipids of the corona. NP-bound proteins may recognize and interact with membrane receptors or could bind nonspecifically to cellular membranes. Whatever these interactions, they seem to play a central role which could determine further biological responses. In particular, these interactions may drive the uptake of NPs by the first target cells at the level of the biological barriers such as immune cells (macrophages, dendritic cells, and neutrophils) or epithelial and endothelial cells. This uptake seems to be general for many NPs which are able to bind proteins at their surface, and the paradigm of "Trojan horse" was developed to explain this uptake and the further biological responses. One of the first responses is the direct or indirect production of ROS which is associated with the size, the chemical composition, and the surface reactivity of the NPs. This common response occurs for a large number of NPs even with different chemical patterns and different abilities to form agglomerates, and thus the paradigm of the central role of oxidative stress was developed [5]. These authors suggested that "although not all materials have electronic configurations or surface properties to allow spontaneous ROS generation, particle interactions with cellular components are capable of generating oxidative stress." Further activation of nuclear factors and specific genetic programs are associated with the level of ROS production leading to cell death by necrosis and apoptosis or adaptive responses such as pro-inflammatory responses, antioxidant enzyme activation, repair processes, effects on cell cycle control, and proliferation. Over the last years, numerous in vitro studies have confirmed this hypothesis leading to the development of assays using the detection of ROS or oxidative stress for the screening of NPs. However, new data during the last year have pointed out other specific effects of NPs which are not related to oxidative stress. For example, NPs can interact with membrane receptors, induce their aggregation, and mimic sustained physiological responses through specific signaling pathways in the target cells. This type of mechanism may contribute to the development of diseases but

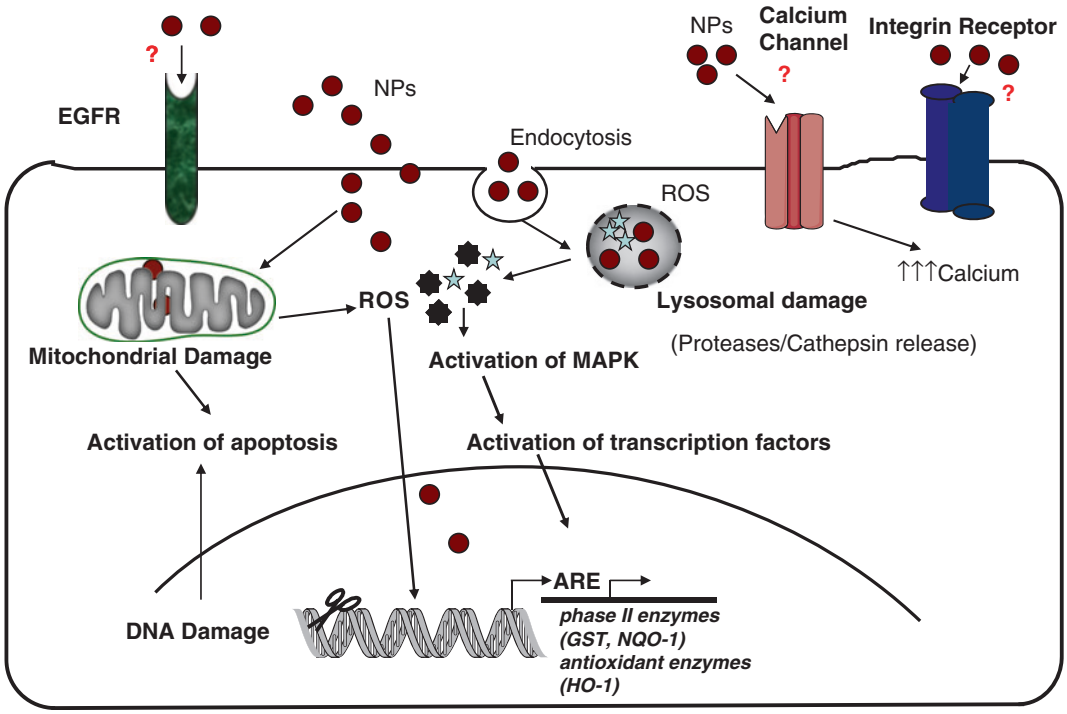


Cellular Mechanisms of Nanoparticle Toxicity, Fig. 1 Different physicochemical characteristics of the nanomaterials involved in their biological activity: size, surface area, shape, bulk chemical composition, surface chemistry including solubility as well as surface charge or coatings, and interactions between particles leading to agglomeration and aggregation as well as with proteins leading to "corona" or with receptors in the cell membrane

could also be of use to develop therapeutic strategies whereby NPs activate or block specific receptors.

Cellular Uptake of Nanoparticles and Their Fate at Cellular Level

The uptake of particles by specialized immune cells in human is a normal process which leads to their removal and contributes to the integrity of the body. However, depending on the level of the uptake, this process could induce an increasing release of inflammatory mediators and disturbance of the normal functions of phagocytes such as the clearance and the destruction of pathogens. One of the knowledge of the last fifty years on the effects of a sustained exposure to airborne particles, especially at occupational level, is the concept of overloading. If the mechanisms of clearance are not sufficient to eliminate the particles and if they are persistent, the particles could accumulate in the tissues, leading to a sustained inflammation and chronic pathologies. This was demonstrated not only for exposure to quartz, asbestos, coal, and mineral dusts but also for longtime exposure to heavy PM-polluted atmospheres such as in Mexico City [1]. These questions of uptake and persistence are fundamental for risk assessment evaluation of NPs. This may explain the number of papers published recently that analyze the mechanisms of uptake, the behavior, and the translocation of various NPs. So far, it appears that the response depends on several different parameters: the NP surface and its specific chemical composition resulting from the engineering processes; the capacity of NPs to form aggregates (strongly bonded or fused NPs) or agglomerates (collection of weakly bound NPs); and the methods used for dispersion and experimental preparation, which determine the ability of NPs to adsorb or not specific biological compounds such as proteins, to form the "corona" and to interact with biological membranes [7]. The amount and the structural/functional properties of the adsorbed proteins drive the interactions of these nanomaterials with the membranes and their uptake (Fig. 1). Recent studies have clearly identified a number of serum proteins such as albumin, IgG, IgM, IgA, apolipoprotein E, cytokines, or transferrin that bind to carbon black, titanium dioxide, acrylamide,



Cellular Mechanisms of Nanoparticle Toxicity, Fig. 2 A schematic representation of NP-triggered cellular pathways through membrane receptors, ROS production, and implication of oxidative stress in these responses. NPs could induce activation of EGF or integrin receptors can lead to apoptosis, inflammation, or proliferation. ROS produced by NPs in immediate cellular environment or inside the cells lead to activation of redox-dependent signaling pathways like MAPK and the activation of transcription factors, e.g., AP-1, NF-kB, or Nrf2. They migrate to the nucleus and modify gene expression of cytokines, phase 2 enzymes (glutathione S transferase or GST, quinone oxidoreductase 1 or NQO-1), and antioxidant enzymes (heme oxygenase 1 or HO-1). Oxidative stress could also result in the damage of different organelles like the mitochondria, lysosomes, and nucleus resulting to apoptosis. Accumulation of high intracellular calcium levels through a direct effect on calcium channel might also act as an alternative mechanism for the induction of these mechanisms (Adapted from Marano et al. [11])

or polystyrene NPs [8]. Among the identified proteins, several are ligands for cellular receptors and may contribute to the biological effects of NPs. For example, receptor aggregation induced by NPs could lead to cell signaling: coated gold NPs were able to bind and cross-link IgE-Fc epsilon receptors leading to degranulation and consequent release of chemical mediators [9].

On another hand, integrins such as $\alpha_5\beta_3$ are known to play a key role in cell signaling, and their activation by extracellular ligands can modulate biological processes such as matrix remodeling, angiogenesis, tissue differentiation, and cell migration. These receptors were recently demonstrated as important membrane targets for carbon NPs, and their activation induced lung epithelial cell proliferation which was due at least in part to β1-integrin activation [6].

As far as uptake process is concerned, it is likely that different cell types might have different uptake mechanisms, even for the same NPs. The possible pathways of cellular uptake were previously described by several authors (see Ref. [10]). It could occur through phagocytosis, macropinocytosis, clathrin-mediated endocytosis, non-clathrin- and non-caveolae-mediated endocytosis, caveolae-mediated endocytosis, or diffusion (Fig. 2). These mechanisms have been described for different NPs and may occur for the same NP depending on the cell type, the medium, and the level of aggregation. Therefore, uptake processes are considered as very complex and not

easy to measure. Dawson et al. [12] have postulated that the uptake depends mostly on the size: NPs less than 100 nm can enter the cells and less than 40 nm in the nucleus. It was also suggested that the size of the NPs determine caveolin- versus clathrin-dependent uptakes [13]. However, these oversimplified scenarios are refuted by obvious discrepancies in the recent literature about the optimal size, shape, and mechanisms of internalization of NPs.

The surface charge of the NPs could be an important factor for uptake since the negatively charged surface membrane could favor the positively charged NPs for higher internalization. However, negatively charged NPs were also shown to have enhanced uptake as compared to unfunctionalized NPs, perhaps by their possible interactions with proteins. Endocytosis of small NPs is energy dependent and associated with lipid rafts, dynamin, and F-actin mechanisms. Phagocytosis and macropinocytosis are mostly involved in the endocytosis of large particles (more than 500 nm) and also in the uptake of the aggregates or agglomerates of NPs which could be promoted by their opsonization in the biological fluids. Macropinocytosis (which is one kind of pinocytosis) is also an important mechanism for positively charged NPs and TiO_2 or carbon black aggregate internalization [14].

The behavior of the NPs after their uptake is another important question, but, surprisingly, as far as now, little is known about the intracellular fate of NPs. Most of the transmission electron microscopy (TEM) observations have shown the NPs in cytoplasmic vesicles limited by membranes. These vesicles could further be transported in the cytoplasm through the microtubule network. The bio-persistence of nanomaterials which are resistant to degradation in the endosomal compartment could be one of the factors of further toxicity and accumulation. However, several metal oxide NPs are toxic after dissolution in the cell. Indeed, the uptake of ZnO NPs into the lysosomal acidic medium accelerates their dissolution and the release of Zn^{2+} ions in the cytoplasm. Their excess could induce cytokine production and cytotoxicity and the initiation of acute inflammation at the level of the target organ such as the lung.

NPs such as TiO_2 or carbon black NPs were also observed free in the cytoplasm of cells [14]. Two explanations may be put forward. The first one is that NPs could directly enter by diffusion through the lipid bilayer. It has been shown that cationic NPs could pass though cell membranes by generating transient holes without membrane disruption [15]. Another possible explanation could be the release of NPs after rupture of endosomal compartment. It was described that cationic NPs, after binding to lipid groups on the cell surface membrane, could be endocytosed in vesicles and accumulated in the lysosomal compartment. Within, they are able to sequestrate protons which could lead to the activation of proton pumps and further rupture of the ion homeostasis and lysosomal accumulation of water. The subsequent lysosomal swelling and membrane rupture lead to the cytoplasmic release of NPs [16]. In proliferating cells, these cytoplasmic NPs, associated or not with microtubules, could enter in the nucleus during the mitosis, which could explain that nonsoluble NPs were observed in the nucleus [14]. More rarely, NPs were also observed within the mitochondrial matrix but, so far, no explanation was given to explain this organelle localization.

The Cellular Stress Induced by Nanoparticles and Its Biological Consequences

The last 10 years of research conducted on the mechanisms of toxicity of non-intentional as well as engineered NPs has led to the establishment of a consensus within the scientific community of toxicologists to consider the central role of oxidative stress in cellular responses to NPs leading to inflammation or apoptosis [5, 17] (Fig. 2). The concept of oxidative stress was developed for many years to explain dysfunctions leading to pathologies. Oxidative stress could occur when reactive oxygen species (ROS) are overproduced leading to an imbalance between ROS production and antioxidant defense capacity. Oxidative stress could also occur when the organism shows a deficiency in antioxidant systems and, especially,

in antioxidant enzymatic systems (superoxide dismutase, catalase, glutathione peroxidase). An increased concentration of ROS, exceeding the antioxidant capacity of the cells, can lead to oxidative damage at molecular or cellular level.

ROS have important cellular roles either by acting as second messengers for the activation of specific pathways and gene expressions or by causing cell death. In the hierarchical oxidative stress model in response to NPs, Nel et al. [5] propose that a minor level of oxidative stress leads to the activation of the antioxidant protection, whereas, at a higher level, cell membrane and organelle injuries could lead to cell death by apoptosis or necrosis, but specific signaling pathways and gene expression are involved at each step. The induction of oxidative stress by several NPs is due to their ability to produce ROS (e.g., TiO_2) or to lead to their production. The surface properties of NPs modulate the production of ROS, and the smaller they are, the higher is their surface area and their ability to react with biological components and to produce ROS. However, if this cellular induction appears to be general, all the NPs are not able to produce ROS, and the cellular increase of the latter could be an indirect effect of the uptake.

ROS interact nonspecifically with biological compounds, yet some macromolecules are more sensitive such as the unsaturated lipids, amino acids with a sulfhydril group (SH), and guanine sites in nucleic acids. When lipid bilayer is attacked by ROS, cascade peroxidation occurs leading to the disorganization of the membranes and of their functions (exchange, barriers, information). The most sensitive proteins contain methionine or cysteine residues, especially in their active site, and their oxidation could lead to modifications of their activity and even to their inactivation.

The adaptive cellular responses to NPs are associated with the modulation of different redox-sensitive cellular pathways. Tyrosine kinases and serine/threonine kinase such as mitogen-activated protein kinases or MAP kinases were especially studied (ERK, p38, and JNK) in association with several transcription factors such as NFκB. The free radical can degrade the NFκB inhibitor IκB by the activation of the cascades leading to its proteolysis. The activation of NFκB induces its translocation within the nucleus and its link to consensus sequences in the promoter of numerous genes leading to their transcription. This is also the case for other transcriptions factors such as AP1 and NrF2. The latter plays an essential role in the antioxidant response element (ARE)-mediated expression of phase 2 enzymes such as NQO1 (NADPH quinone oxidoreductase-1) and antioxidant enzymes such as heme oxygenase-1 (HO-1). Indeed, HO-1 was found to be activated by CeO_2 NP exposure of human bronchial cells via the p38-Nrf-2 signaling pathway. The ability of NPs to interact with these signaling pathways could partially explain their cytotoxicity. Recently, TiO_2 and SiO_2 NPs were demonstrated in vitro and in vivo to induce the release of IL1β and IL1α, two potent mediators of innate immunity, via the activation of the inflammasome, a large multiprotein complex containing caspase 1 which cleaves pro IL1-β in its active form. These results lead to consider that these NPs could induce a potent inflammatory response. However, the mechanisms leading to this activation are not yet clear.

Another important target of ROS produced by NPs is DNA. Oxidative damage of DNA could generate intrachain adducts and strand breakage. The bond between the base and deoxyribose could also be attacked leading to an abasic site, and the attack on the sugar could create a single-strand break. The genotoxicity of NPs begins to be studied and recent reviews pointed out the possible genotoxic mechanisms.

However, oxidative stress appears now not sufficient to explain all the biological effects of NPs. The role of epidermal growth factor receptor (EGFR) was investigated by the group of K. Unfried with the demonstration that carbon black NPs induce apoptosis and proliferation via specific signaling pathways both using EGFR [18]. Carbon black NPs could also impair phagosome transport and cause cytoskeletal dysfunctions with a transient increase of intracellular calcium not associated with the induction of ROS since antioxidants did not suppress the response,

which could be due to a direct effect on ion channels that control the calcium homeostasis in the cell [19]. Even if all the mechanisms are not completely demonstrated, it appears now that transmembrane receptors are implicated in NP-induced cell signaling and could lead to specific biological responses to NPs.

Nanoparticles and Cell Death

NPs have also been shown to induce either apoptotic or necrotic cell death in a variety of in vitro systems depending on the concentration and duration of exposure. This induction of cell death mechanism by NPs might act as a basis of different pathologies, and consequently it is important to understand NP-induced apoptosis pathways. Cells are able to undergo apoptosis through two major pathways, the extrinsic pathway with the activation of death receptors and the intrinsic pathway with the central role of mitochondria, its permeabilization, and the release of cytochrome *c* leading to the activation of apoptosome. Recently, the permeabilization of lysosomal membrane was also shown to initiate apoptosis with the release of catepsins and other hydrolases from the lysosomal lumen. The molecular pathways of apoptosis induction by carbon black and titanium dioxide NPs in human bronchial epithelial cells were recently studied. It was shown that the initial phase of apoptosis induction depends upon the chemical nature of the NPs. Carbon black NPs triggered the mitochondrial pathway, with the decrease of mitochondrial potential, the activation of bax (a proapoptotic protein of the Bcl2 family), and the release of cytochrome *c*, and the production of ROS is implicated in the downstream mitochondrial events. TiO_2 NPs induced lysosomal pathway with lipid peroxidation, lysosomal membrane destabilization, and cathepsin B release [20]. Lysosomal permeabilization has also been shown to be important in silica NP-induced apoptosis. These results point out the necessity of a careful characterization of the molecular mechanisms involved by NPs and not just describing at the final outcome.

Future Directions of Research

The interactions between nanomaterials and their biological target are essential to explain their biological effect, and the interest of the recent researches on the cellular mechanisms induced by NPs is to take into account the specificity of the cells and their microenvironment. The first step is the formation of the corona in biological fluids whose composition and affinity kinetics strongly depend on the characteristics of NPs and, especially, their size and surface reactivity. This coating of proteins influences the aggregation, the final size, and, finally, the uptake of NPs via the interaction with the membranes, their specific receptors, or lipid rafts. This could determine if the nanomaterial is bioavailable and if NPs induce or not adverse interactions. The central mechanism proposed to explain the biological response is the oxidative stress. However, this paradigm is debated because very similar oxidative stress effects observed in cellular models and induced by different particles could lead in vivo to different pathological effects. It is now obvious that oxidative stress is a common and nonspecific mechanism in toxicology and that the responses at the level of the cell depend on the perturbation of the redox balance with a few number of induced signaling pathways. The different biological responses could depend on the tissue specificity which could lead to different diseases observed after occupational or environmental exposure to well-known particles or fibers.

Recent studies have also shown that NPs could develop a response without a direct contact with the cells but after an induction of secreted factors, which is the "bystander effect." Small molecules such as purines could be increased at cytoplasmic level in response to NPs, transferred through the gap junctions within a tissue to activate specific receptors [10]. Moreover, NP-induced apoptosis was also demonstrated to be propagated through hydrogen peroxide-mediated bystander killing in an in vitro model of human intestinal epithelium. These specific responses could explain the differences observed in vivo. Finally, the

interactions of NPs with proteins, enzymes, cytokines, and growth factors, outside or inside the cell, lead to modifications of the functions of these proteins with a possible indirect pathological effect.

The large variety of engineered NPs on the market and under development makes these studies very complex. However, the development of safe nanomaterials depends on better knowledge of these specific interactions.

Cross-References

- ▶ Genotoxicity of Nanoparticles
- ▶ In Vivo Toxicity of Carbon Nanotubes
- ▶ Quantum Dot Toxicity
- ▶ Toxicology: Plants and Nanoparticles

References

1. Donaldson, K., Borm, P.: Particle Toxicology, p. 434. CRC Press, Boca Raton, Florida, USA (2007)
2. Oberdorster, G., Oberdorster, E., Oberdorster, J.: Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environ. Health Perspect. **113**, 823–839 (2005)
3. Brunekreef, B., Holgate, S.T.: Air pollution and health. Lancet **360**, 1233–1242 (2002)
4. Marano, F., Boland, S., Baeza-Squiban, A.: Particle-associated organics and proinflammatory signaling. In: Donaldson, K., Borm, P. (eds.) Particle Toxicology, pp. 211–226. CRC Press, Boca Raton, Florida, USA (2007)
5. Nel, A., Xia, T., Madler, L., Li, N.: Toxic potential of materials at the nanolevel. Science **311**, 622–627 (2006)
6. Unfried, K., Albrecht, C., Klotz, L.O., Mikecz, A.V., Grether-Beck, S., Schins, R.P.F.: Cellular responses to nanoparticles: target structures and mechanisms. Nanotoxicology **1**, 52–71 (2007)
7. Nel, A.E., Madler, L., Velegol, D., Xia, T., Hoek, E.M., Somasundaran, P., Klaessig, F., Castranova, V., Thompson, M.: Understanding biophysicochemical interactions at the nano-bio interface. Nat. Mater. **8**, 543–557 (2009)
8. Lynch, I., Salvati, A., Dawson, K.A.: Protein-nanoparticle interactions: what does the cell see? Nat. Nanotechnol. **4**, 546–547 (2009)
9. Huang, Y.F., Liu, H., Xiong, X., Chen, Y., Tan, W.: Nanoparticle-mediated IgE-receptor aggregation and signaling in RBL mast cells. J. Am. Chem. Soc. **131**, 17328–17334 (2009)
10. Bhabra, G., Sood, A., Fisher, B., Cartwright, L., Saunders, M., Evans, W.H., Surprenant, A., Lopez-Castejon, G., Mann, S., Davis, S.A., Hails, L.A., Ingham, E., Verkade, P., Lane, J., Heesom, K., Newson, R., Case, C.P.: Nanoparticles can cause DNA damage across a cellular barrier. Nat. Nanotechnol. **4**, 876–883 (2009)
11. Marano, F., Hussain, S., Rodriges-Lima, F., Baeza-Squiban, A., Boland, S.: Nanoparticles: molecular target and cell signaling. Arch. Toxicol. 85(7):733-41 (2011)
12. Dawson, K.A., Salvati, A., Lynch, I.: Nanotoxicology: nanoparticles reconstruct lipids. Nat. Nanotechnol. **4**, 84–85 (2009)
13. Rejman, J., Oberle, V., Zuhorn, I.S., Hoekstra, D.: Size-dependent internalization of particles via the pathways of clathrin- and caveolae-mediated endocytosis. Biochem. J. **377**, 159–169 (2004)
14. Hussain, S., Boland, S., Baeza-Squiban, A., Hamel, R., Thomassen, L.C., Martens, J.A., Billon-Galland, M.A., Fleury-Feith, J., Moisan, F., Pairon, J.C., Marano, F.: Oxidative stress and proinflammatory effects of carbon black and titanium dioxide nanoparticles: role of particle surface area and internalized amount. Toxicology **260**, 142–149 (2009)
15. Gratton, S.E., Ropp, P.A., Pohlhaus, P.D., Luft, J.C., Madden, V.J., Napier, M.E., Desimone, J.M.: The effect of particle design on cellular internalization pathways. Proc. Natl. Acad. Sci. U. S. A. **105**, 11613–11618 (2008)
16. Xia, T., Kovochich, M., Liong, M., Zink, J.I., Nel, A. E.: Cationic polystyrene nanosphere toxicity depends on cell-specific endocytic and mitochondrial injury pathways. ACS Nano **2**, 85–96 (2008)
17. Ayres, J.G., Borm, P., Cassee, F.R., Castranova, V., Donaldson, K., Ghio, A., Harrison, R.M., Hider, R., Kelly, F., Kooter, I.M., Marano, F., Maynard, R.L., Mudway, I., Nel, A., Sioutas, C., Smith, S., Baeza-Squiban, A., Cho, A., Duggan, S., Froines, J.: Evaluating the toxicity of airborne particulate matter and nanoparticles by measuring oxidative stress potential – a workshop report and consensus statement. Inhal. Toxicol. **20**, 75–99 (2008)
18. Sydlik, U., Bierhals, K., Soufi, M., Abel, J., Schins, R. P., Unfried, K.: Ultrafine carbon particles induce apoptosis and proliferation in rat lung epithelial cells via specific signaling pathways both using EGF-R. Am. J. Physiol. Lung Cell. Mol. Physiol. **291**, L725–L733 (2006)
19. Moller, W., Brown, D.M., Kreyling, W.G., Stone, V.: Ultrafine particles cause cytoskeletal dysfunctions in macrophages: role of intracellular calcium. Part. Fibre Toxicol. **2**, 7 (2005)
20. Hussain, S., Thomassen, L.C., Feracatu, I., Borot, M. C., Andreau, K., Fleury, J., Baeza-Squiban, A., Marano, F., Boland, S.: Carbon black and titanium oxide nanoparticles elicit distinct apoptosic pathways in bronchial epithelial cells. Part. Fibre.Toxicol. **7**(10), 1–17 (2010). Online 16 Apr

Cellular Toxicity

▶ Nanoparticle Cytotoxicity

Chaos

▶ Nonlinear and Parametric NEMS Resonators

Characterizations of Zinc Oxide Nanowires for Nanoelectronic Applications

▶ Fundamental Properties of Zinc Oxide Nanowires

Charge Transfer

▶ Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Charge Transfer on Self-Assembled Monolayer Molecules

▶ Charge Transport in Self-Assembled Monolayers

Charge Transport in Carbon-Based Nanoscaled Materials

▶ Electronic Transport in Carbon Nanomaterials

Charge Transport in Self-Assembled Monolayers

Jeong Young Park
Graduate School of EEWS (WCU), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

Synonyms

Charge transfer on self-assembled monolayer molecules

Definition

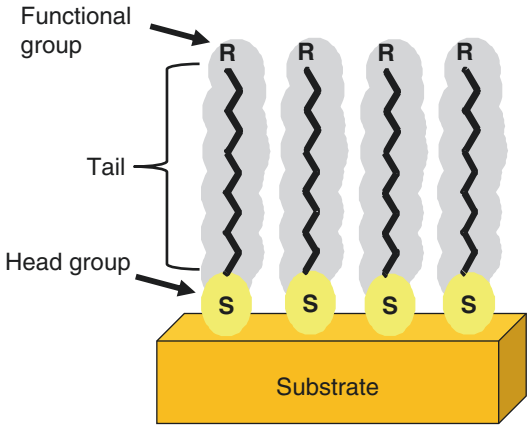
Charge transport in self-assembled monolayers (SAMs) is the transport of an electron or a hole through an organized molecule layer which is bound to a substrate.

Overview

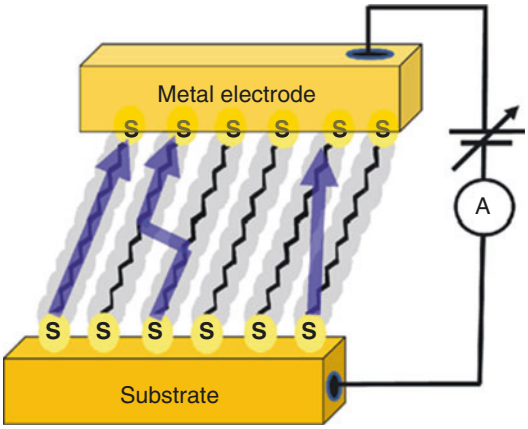
Charge Transport Through Organic Molecules

Significant studies on charge transport properties through organic molecules have been carried out in the general area of molecule-based and molecule-controlled electronic devices, often termed "molecular electronics" [1, 2]. Self-assembled monolayers (SAMs) are composed of an organized layer of amphiphilic molecules in which one end of the molecule, the "head group," shows a special affinity for a substrate [3]. SAMs also consist of a tail with a functional group at the terminal end, as seen in Fig. 1.

Charge transport of organic molecules is usually limited by hopping processes and is therefore dominated by surface ordering. Self-assembled monolayers are a good model system of molecular electronics due to the ordered surface structure. In order to measure charge transport in a self-assembled monolayer, the substrate surface should be metallic. For example, a gold surface



Charge Transport in Self-Assembled Monolayers, Fig. 1 Schematic of a self-assembled monolayer (*SAMs*) showing the head group that is bound to the substrate. SAMs consist of a tail with a functional group at the terminal



Charge Transport in Self-Assembled Monolayers, Fig. 2 Scheme of charge transport mechanisms through self-assembled monolayers. The dominant charge transport mechanism in a molecular junction involves "through-bond" (*TB*) tunneling, and "through space" (*TS*) as illustrated in the left and right transport channel, respectively

exhibits strong bonds with alkanethiol through S–H bonds. The other electrode should also be metallic for charge transport through the self-assembled monolayer. The measurement scheme of charge transport through a self-assembled monolayer that represents a conductor-molecule-conductor junction is shown in Fig. 2.

Charge Transport Mechanism

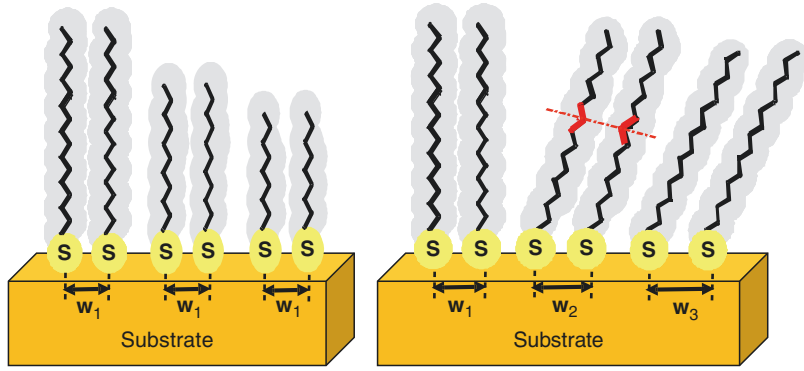
For insulating molecules, such as alkane chains, electron transport occurs via tunneling mechanisms. When such molecules are placed between electrodes, the junction resistance changes exponentially: $R = R_0 \exp(\beta s)$, with electrode separation s, where R_0 is the contact resistance and β a decay parameter. In most experiments, the separation s is the length of the alkane chain. However, length is not the only important parameter. Conformation and molecular orientation relative to the electrodes are also important. Other factors need to be considered as well, including energy positions of the highest occupied and lowest unoccupied molecular orbitals (HOMO, LUMO), electrode work function, and nature of the bonding to the electrodes.

Charge transport mechanisms through self-assembled monolayers consist mainly of three processes [4]. The dominant charge transport mechanism in a molecular junction is "through-bond" (TB) tunneling, where the current follows the bond overlaps along the molecules (as illustrated in the left transport channel of Fig. 2). Another contribution involves the charge transport from electrode to electrode, in which the molecule plays the role of a dielectric medium that is called "through space" (TS), as illustrated in the right transport channel of Fig. 2. The last contribution of charge transport pathway involves a chain-to-chain coupling as illustrated in the middle of Fig. 2. As the molecular chains tilt, the decrease of the electron tunneling distance leads to a lateral hop between the neighboring molecular chains.

Two Pathway Models

If electron transport was determined purely by tunneling through the alkane chains, one would expect the value of β to equal zero, since the tunneling distance is the same for all tilt angles. The nonzero value of β indicates the existence of either intermolecular charge transfer or variations in the S-Au bonding as a function of tilt that affect the conductivity in an exponential way with angle.

Slowinski et al. [4] proposed a two-pathway conductance model involving "through-bond" tunneling, and the "chain-to-chain" coupling.



Charge Transport in Self-Assembled Monolayers, Fig. 3 Scheme of the measurement of the junction resistance for two different situations: (*1*) decreasing of the alkane chain (*left* part), and (*2*) the tilting of the alkane chain while maintaining the same number of carbon atoms (*right* part), which will yield the resistance (*1*) per unit length of molecule or (*2*) tilting angle of the molecules, respectively

Assuming no effects due to changes in S-Au bonding, the first pathway is independent of tilt, while the second depends on the tilt angle. The tunneling current, thus, is given by

$$I_t = I_0 \exp(-\beta_{TB} d) + I_0 n_s \exp[-\beta_{TB}(d - d_{CC} \tan \Theta)] \times \exp(-\beta_{TS} d_{CC})$$

where I_t is the current at a specific tilt angle Θ, d is the length of the molecule, n_s is a statistical factor accounting for the number of pathways containing a single lateral hop as compared to those containing only through-bond hops, d is the diameter of the molecule chains, β_{TB} and β_{TS} are respectively through-bond and through-space decay constants. For example, in case of C16 alkanethiol molecule chains, $d_{CC} = 4.3$ Å, $d = 24$ Å, and $n_s = 16$, i.e., the number of carbon atoms in the molecule.

Decay Constant upon Shortening and Tilting of Molecules

The junction resistance is dependent on electrode spacing for two different situations: (1) shortening of the alkane chain [5] and maintaining the same width (w) between chains and (2) tilting of the alkane chain but changing w [6, 7]. These measurements will yield the resistance per unit length of molecule or tilting angle of the molecules, respectively. The conductance decay constant β has already been measured using SAMs with different chain lengths when the separation between electrodes decreases as a function of the alkane chain length (the left image of Fig. 3). The decay constant, β, upon tilting of molecules can be measured using deformation with an AFM tip and simultaneous measurement of current (the right image of Fig. 3). This methodology will be described in the next section.

Basic Methodology

Preparation of Self-Assembled Monolayer

The organic molecular films on various types of substrates (conducting, semiconducting, or insulating substrates) have been prepared using techniques such as the Langmuir-Blodgett technique, dipping the substrates into solution with molecules, drop casting, or spin-coating [8].

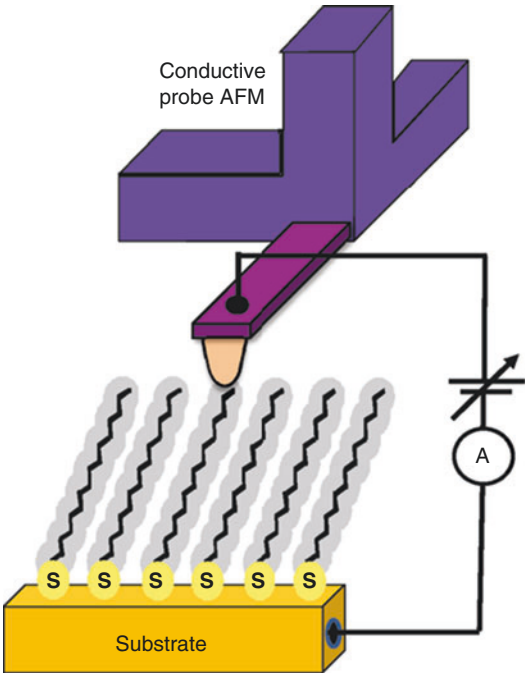
As one example, details on the preparation of an alkanethiol SAM will be described below. Gold substrates (200–300 nm of gold coating over 1–4 nm of chromium layer n glass) are prepared by butane flame annealing in air after cleaning in acetone, chloroform, methanol, and a piranha solution (1:3; H_2O_2:H_2SO_4). The resulting surface consisted of large grains with flat terraces of (111) orientation (sizes up to 400 nm) separated by monatomic steps. Flatness and cleanness were tested by the quality of the lattice-resolved images of the gold substrate.

Two types of hexadecanethiol (C16) self-assembled monolayer can be formed on Au (111): complete monolayers of the molecules and islands of molecules covering only a fraction of the substrate. In the first case, the film was produced by immersing the substrate in 1 mM ethanolic solution of C16 for about 24 h, followed by rinsing with absolute ethanol and drying in a stream of nitrogen to remove weakly bound molecules. Incomplete monolayers in the form of islands were prepared by immersing the substrate in a 5 μM ethanolic solution of C16 for approximately 60 s, followed by rinsing. Samples consisting of islands facilitate the determination of the thickness of the molecular film relative to the surrounding exposed gold substrate. The molecular order of the islands improves with storage time at ambient conditions.

Techniques to Measure Charge Transport in Self-Assembled Monolayers

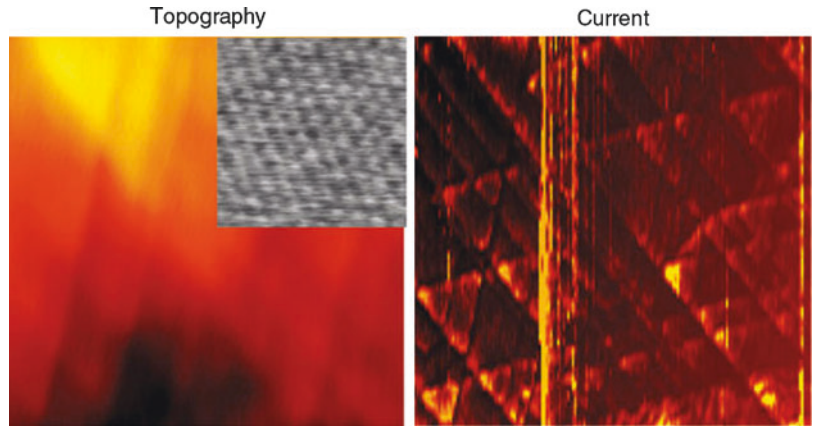
The current through a thiol SAM on a hanging Hg drop electrode can be measured in an electrochemical solution. The current was measured as a function of the monolayer thickness that can be tuned by two methods: by changing the number of carbons in the alkane chain and therefore its length; or, expansion of the Hg drop such that the monolayer surface coverage was reduced and the molecules increased their tilt angle with respect to the surface. Slowinski et al. determined the decay constants $\beta_{TB} = 0.91/Å$ and $\beta_{TS} = 1.31/Å$ by both a fit to their experimental data and by independent ab initio calculations. Mercury drop expansion experiments by Slowinsky et al. have shown a dependence of the current through the alkanethiol monolayers on surface concentration, prompting the authors to suggest the existence of additional pathways for charge transfer, like chain-to-chain tunneling.

Scanning tunneling microscopy and scanning tunneling spectroscopy have been used to reveal the atomic scale surface structure and charge transport properties of SAM layers [9, 10]. STM has been used to reveal various phases of surface structure and atomic scale defects, which could play a crucial role in the electrical transport.



Charge Transport in Self-Assembled Monolayers, Fig. 4 Scheme of conductance measurements of SAM with a conductive-probe atomic force microscopy (*CP-AFM*) system

Conductance measurements were performed with a conductive-probe atomic force microscopy (CP-AFM) system. The use of AFM with conducting tips provides the ability to vary the load on the nanocontact and also opens the way for exploring electron transfer as a function of molecular deformation. A junction is fabricated by placing a conducting AFM tip in contact with a metal-supported molecular film, such as a self-assembled monolayer (SAM) on Au, as shown in Fig. 4. The normal force feedback circuit of the AFM controls the mechanical load on the nanocontact while the current–voltage (*I–V*) characteristics are recorded. The possibility to control the load on the contact is an unusual characteristic of this kind of junction and provides the opportunity to establish a correlation between the mechanical deformation and electronic properties of organic molecules. The normal force exerted by the cantilever was kept constant during AFM imaging, while the current between tip and sample was recorded. It is crucial to carry out the experiment in



Charge Transport in Self-Assembled Monolayers, Fig. 5 AFM images (200 × 200 nm) of topography, and current images obtained simultaneously for a full monolayer of C16 on Au (111) surface. Lattice-resolved images of the film (*inset* in the *left* figure) reveal a lattice image of SAM (size: 2 × 2 nm)

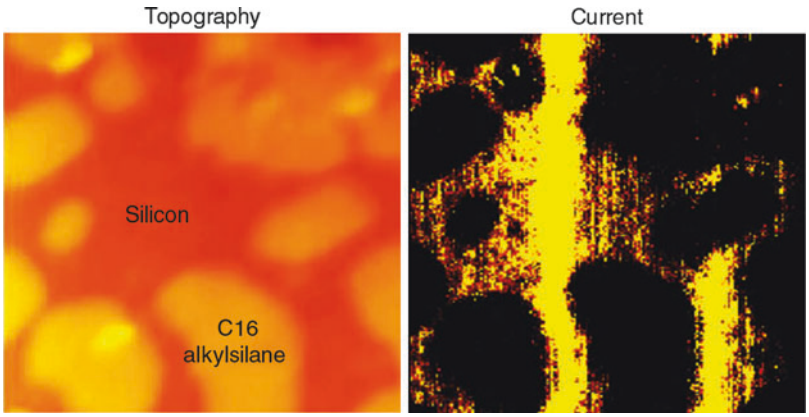
the low load regime so that there is no damage to the surface. This can be confirmed by inspection of the images with Ångstrom depth sensitivity as well as by the reproducibility of the current and adhesion measurements. If the measured conductance did not change at constant load and did not show time-dependent behavior in the elastic regime, the tip experiences minimal changes during subsequent contact measurements.

Key Research Findings

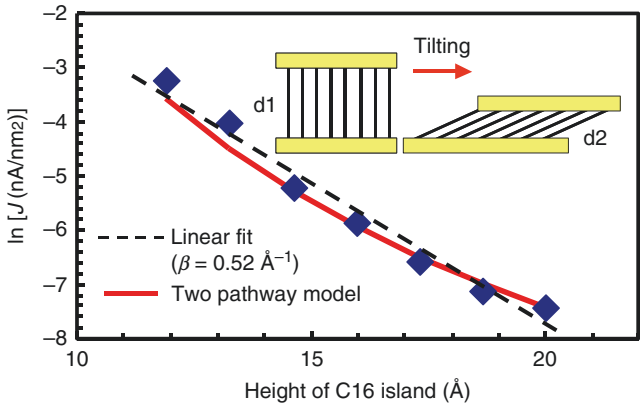
The molecular tilt induced by the pressure applied by the tip is one major factor that leads to increased film conductivity. By measuring the current between the conductive AFM tip and SAM as a function of the height of the molecules, the decay parameter (β) can be obtained [11]. Wold et al. studied the junction resistance as a function of load using AFM. The resistance was found to decrease with increasing load within two distinct power law scaling regimes [12]. Song et al. examined the dependence of the tunneling current through Au-alkanethiol-Au junctions on the tip-loading force [13]. It is found that the two-pathway model proposed by Slowinsky et al. can reasonably fit with the results, leading the authors to conclude that the tilt configuration of alkanethiol SAMs enhances the intermolecular charge transfer.

Figure 5 shows topography and current images obtained simultaneously for a full monolayer of C16 on an Au (111) surface. The topographic image reveals the commonly found structure of the gold film substrate, composed of triangular-shaped terraces separated by atomic steps. Lattice-resolved images of the film (inset in the left figure) reveal a ($\sqrt{3} \times \sqrt{3}$)-R30° periodicity of the molecules relative to the gold substrate. Qi et al. measured current–voltage (I–V) characteristics on the C16 alkanethiol sample for loads varying between −20 and 120 nN, and found that the current changes in a stepwise manner and the plateaus are associated with the discrete tilt angle of the molecules. A stepwise response of the SAM film to pressure has been observed previously in other properties such as film height and friction of alkanesilanes on mica and alkanethiols on gold.

In order to measure the thickness of the self-assembled monolayer upon molecular deformation, the SAM islands that partially cover the substrate can be used. The heights of the islands can be obtained from topographical AFM images, while charge transport properties of alkanesilane SAMs on silicon surface are measured using AFM with a conducting tip. In this manner, the load applied to the tip-sample contact can be varied while simultaneously measuring electric conductance. Figure 6 shows the topographic and current images, respectively, that were acquired simultaneously on hexadecylsilane islands on a silicon surface. The image size is 500 × 500 nm. The hexadecylsilane islands are 100–200 nm in diameter and have a height of 1.6 nm at the applied load of 0 nN (or effective total load of 20 nN). It is also



Charge Transport in Self-Assembled Monolayers, Fig. 6 AFM images (500 × 500 nm) of topographic, and current images, respectively, that were acquired simultaneously on hexadecylsilane SAM islands on silicon surface



Charge Transport in Self-Assembled Monolayers, Fig. 7 Semilog plot of current density (nA/nm^2) as a function of the height of the hexadecylsilane SAM islands on a silicon surface. A decay constant (β) = 0.52 ± 0.04 Å^{-1} was found for the current passing through the film as a function of tip-substrate separation

clear that the current measured on the alkanesilane island is much smaller than that measured on the silicon surface.

These changes were shown to correspond to the molecules adopting specific values of tilt angle relative to the surface, and explained as the result of methylene groups interlocking with neighboring alkane chains. In the case of complete monolayers of alkanethiol SAM, the junction resistance (R) was measured as a function of the applied load [6]. These data were converted to current versus electrode separation by assigning each step in the current to a specific molecular tilt angle, following the sequence established in previous experiments. It was found that ln(R) increases approximately linearly with tip-surface separation, with an average slope β = 0.57 (±0.03) Å^{-1}. Similar measurement of the decay parameter upon the molecular tilts was carried out with a scanning tunneling microscope and simultaneous sensing of forces. By measuring the current as a function of applied load, a tunneling decay constant β = 0.53 (±0.02) Å^{-1} was obtained [14].

In the case of hexadecylsilane molecules, the local conductance of hexadecylsilane SAM islands on a silicon surface was measured with conductive-probe AFM. A semilog plot of current density (nA/nm^2) was obtained as a function of the height of the hexadecylsilane SAM islands on a silicon surface, as shown in Fig. 7. A decay constant (β) = 0.52 ± 0.04 Å^{-1} was found for the current passing through the film as a function of tip-substrate separation [7]. Figure 7 shows the best fit of the two-pathway model with the experimental current measurement as a function of the heights of molecule islands by using the fitting parameters of β_{TB} and β_{TS} that are 0.9 and 1.1 Å^{-1}, respectively. The good fit indicates that

the two-path tunneling model is a valid model to describe this observation.

While saturated hydrocarbon chains mainly interact with each other via weak van der Waals forces, much stronger intermolecular π-π interactions can be present in organic films comprised of conjugated/hybrid molecules. This influences charge transport significantly [5, 15]. In a conductance AFM study of two SAM systems, Fang et al. revealed the role of π-π stacking on charge transport and nanotribological properties of SAM consisting of aromatic molecules [16]. The two model molecules chosen in this study are (4-mercaptophenyl) anthrylacetylene (MPAA) and (4-mercaptophenyl)-phenylacetylene (MPPA). In MPPA, the end group is a single benzene ring, while in MPAA it is changed to a three fused benzene ring structure. This structural difference induces different degrees of lattice ordering in these two molecular SAM systems. Lattice resolution is readily achieved in the MPAA SAM, but it is not possible for the MPPA SAM under the same imaging conditions, indicating the MPAA is lacking long-range order. However, it is important to note that even without long-range order, the stronger intermolecular π-π stacking in the MPAA SAM greatly facilitates charge transport, resulting in approximately one order of magnitude higher conductivity than in the MPPA SAM.

Future Directions for Research

In this contribution, the basic concept of and recent progress on charge transport studies of organic SAM films formed by saturated hydrocarbon molecules and conjugated molecules has been outlined. Several techniques, including AFM, STM, and hanging Hg drop electrode, are used to elucidate the charge transport properties of SAM layers. A number of molecular scale factors such as packing density, lattice ordering, molecular deformation, grain boundaries, annealing induced morphological evolution, and phase separation play important roles in determining charge transport through SAM films. High resolution offered by scanning probe microscopy (SPM) is a key element in identifying and studying microstructures (e.g., molecular tilt, lattice ordering, defects, vacancies, grain boundaries) in organic films and their effects on electronic properties. Other advanced surface characterization techniques, such as SAM with nano-electrodes, in combination with conductive-probe atomic force microscopy, and spectroscopic techniques such as ultraviolet photoemission spectroscopy (UPS) and inverse photoemission spectroscopy (IPES), could be promising venues to explore the correlation between microstructures and electronic properties of organic films.

Cross-References

- Atomic Force Microscopy
- Conduction Mechanisms in Organic Semiconductors
- Electrode–Organic Interface Physics
- Scanning Tunneling Microscopy
- Self-Assembly

References

1. Aviram, A., Ratner, M.A.: Molecular Electronics: Science and Technology. New York Academy of Sciences, New York (1998)
2. Reed, M.A., Zhou, C., Muller, C.J., Burgin, T.P., Tour, J.M.: Conductance of a molecular junction. Science **278**, 252–254 (1997)
3. Ulman, A.: An Introduction to Ultrathin Organic Films from Langmuir-Blodgett to Self-Assembly. Academic, Boston (1991)
4. Slowinski, K., Chamberlain, R.V., Miller, C.J., Majda, M.: Through-bond and chain-to-chain coupling. Two pathways in electron tunneling through liquid alkanethiol monolayers on mercury electrodes. J. Am. Chem. Soc. **119**, 11910–11919 (1997)
5. Salomon, A., et al.: Comparison of electronic transport measurements on organic molecules. Adv. Mater. **15**, 1881–1890 (2003). doi:10.1002/adma.200306091
6. Qi, Y.B., et al.: Mechanical and charge transport properties of alkanethiol self-assembled monolayers on a Au(111) surface: the role of molecular tilt. Langmuir **24**, 2219–2223 (2008). doi:10.1021/la703147q
7. Park, J.Y., Qi, Y.B., Ashby, P.D., Hendriksen, B.L.M., Salmeron, M.: Electrical transport and mechanical properties of alkylsilane self-assembled monolayers on silicon surfaces probed by atomic force microscopy. J. Chem. Phys. **130**, 114705 (2009)

8. Barrena, E., Ocal, C., Salmeron, M.: Molecular packing changes of alkanethiols monolayers on Au(111) under applied pressure. J. Chem. Phys. **113**, 2413–2418 (2000)
9. Bumm, L.A., Arnold, J.J., Dunbar, T.D., Allara, D.L., Weiss, P.S.: Electron transfer through organic molecules. J. Phys. Chem. B **103**, 8122–8127 (1999)
10. Xu, B.Q., Tao, N.J.J.: Measurement of single-molecule resistance by repeated formation of molecular junctions. Science **301**, 1221–1223 (2003)
11. Wang, W.Y., Lee, T., Reed, M.A.: Electron tunnelling in self-assembled monolayers. Rep. Prog. Phys. **68**, 523–544 (2005)
12. Wold, D.J., Haag, R., Rampi, M.A., Frisbie, C.D.: Distance dependence of electron tunneling through self-assembled monolayers measured by conducting probe atomic force microscopy: unsaturated versus saturated molecular junctions. J. Phys. Chem. B **106**, 2813–2816 (2002). doi:10.1021/jp013476t
13. Song, H., Lee, H., Lee, T.: Intermolecular chain-to-chain tunneling in metal-alkanethiol-metal junctions. J. Am. Chem. Soc. **129**, 3806 (2007)
14. Park, J.Y., Qi, Y.B., Ratera, I., Salmeron, M.: Noncontact to contact tunneling microscopy in self-assembled monolayers of alkylthiols on gold. J. Chem. Phys. **128**, 234701 (2008). doi:10.1063/1.2938085
15. Yamamoto, S.I., Ogawa, K.: The electrical conduction of conjugated molecular CAMs studied by a conductive atomic force microscopy. Surf. Sci. **600**, 4294–4300 (2006)
16. Fang, L., Park, J.Y., Ma, H., Jen, A.K.Y., Salmeron, M.: Atomic force microscopy study of the mechanical and electrical properties of monolayer films of molecules with aromatic end groups. Langmuir **23**, 11522–11525 (2007)

Chem-FET

► Nanostructure Field Effect Transistor Biosensors

Chemical Beam Epitaxial (CBE)

► Physical Vapor Deposition

Chemical Blankening

► Chemical Milling and Photochemical Milling

Chemical Dry Etching

► Dry Etching Processes

C

Chemical Milling and Photochemical Milling

Seajin Oh and Marc Madou
Department of Mechanical and Aerospace Engineering and Biomedical Engineering, University of California at Irvine, Irvine, CA, USA

Synonyms

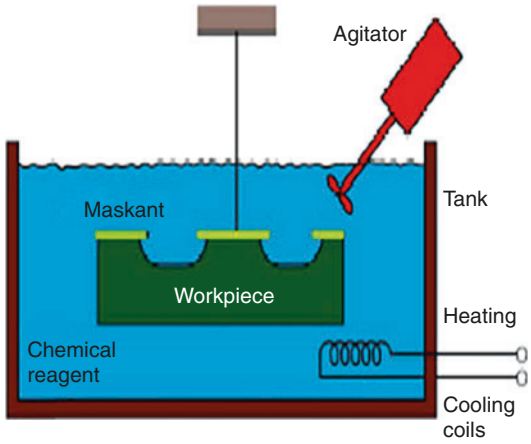
Chemical blankening; Photoetching; Photofabrication; Photomilling

Definition

Photochemical milling (PCM), also known as photochemical machining, is the process of fabricating high precision metal workpieces using photographically produced masks and etchants to corrosively remove unwanted parts. This process is called wet etching in MEMS fabrication techniques and can be also applied to nonmetal materials. Wet etching, when combined with nanolithography, is a useful process to fabricate detailed nanostructures by extremely controlled removal (Fig. 1).

Overview

Photochemical machining (PCM) produces three-dimensional features by wet chemical etching (Fig. 2). PCM yields burr-free and stress-free metal products and allows for the machining of a wide range of materials which would not be suitable for traditional metal working techniques. PCM is also known as photoetching,



Chemical Milling and Photochemical Milling, Fig. 1 (**a**) Schematic illustration of photochemical milling process

photomilling, photofabrication, or chemical blankening [1]. There is a special type of photochemical milling that uses light for initiating or accelerating the wet etching process in metal or semiconductor materials.

The combination of photoresists and wet etching enables the fabrication of very detailed structures with complex geometry or large arrays of variable etching profiles in thin (<2 mm) flat metal sheets. Photoresists are made of synthetic polymers having consistent properties. Liquid photoresist coats a thin film by dipping or spin casting which enables the production of detailed patterns, but often creates pinholes in the thin layer. Thick dry photoresist films applied by hot lamination have advantages of process simplicity and reliability although the materials are expensive.

The process of wet etching is based on the redox chemistry of etchant reduction and metal oxidation, which results in the formation of soluble metal-containing ions that diffuse away from the reaction metal surface. Many metals commonly used in manufacturing industry are etched readily in aqueous solutions comprising etchant (e.g., ferric chloride). Metal oxides and virtually all materials can be etched with a proper selection of etchant regardless of different etching rates.

Wet etching in the PCM process is isotropic where the etchant attacks both downward into the material and sideways under the edge of the resist layer and the ratio of the depth to the undercut is termed the etch factor (Fig. 3). In MEMS device fabrication, the undercut plays a key role in fabricating free-standing microstructure patterns (e.g., beams and cantilevers) that are necessary where microstructures have to be flexible, thermally isolated, small mass, double-sided contacts with gases or liquid surroundings. In most cases, the free-standing material is deposited as a film on a substrate surface, termed a sacrificial material. The etchant, possessing a lateral etching component, must be sufficiently selective not to attach the free-standing material. Further, novel nanolithography techniques enable to create submicrometer-scaled features.

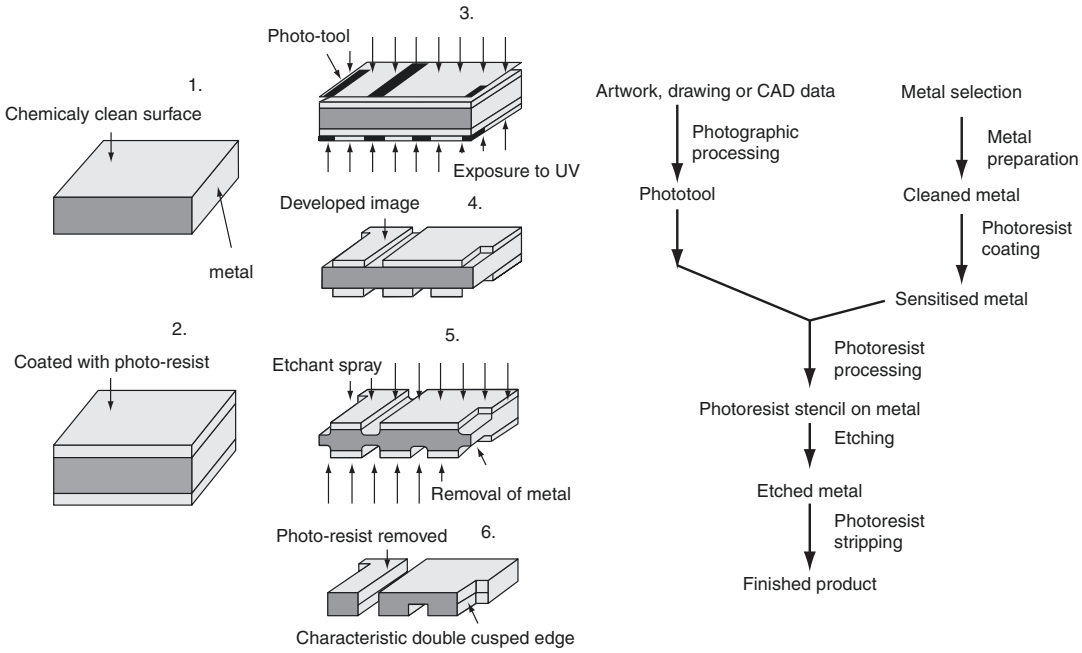
Basic Methodology

Photolithography

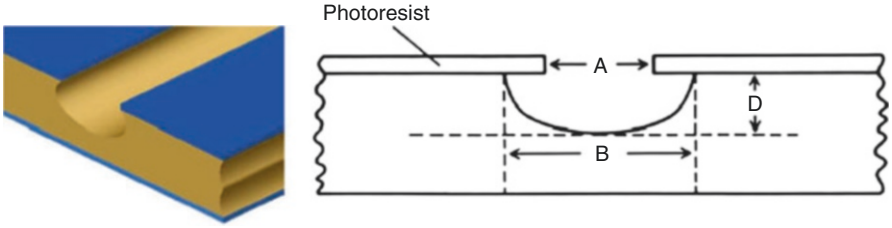
An image of the profile of the flat feature is generated by computer-aided design (CAD) and electronically transferred onto a photographic film to produce a phototool, a photolithographic mask as known in MEMS. The photoresist is exposed through the phototool from ultraviolet source. There are two types of photoresist – negative and positive. UV lights soften the positive film, and the exposed area is released in the developing solution. The negative photoresist film has reversed pattern developing characteristics. Wet resist is applied to a metal sheet by a dipping process while spin coating is commonly used in micromachining (Fig. 4).

Wet Etching

In wet chemical etching, the components of the solid are changed into soluble chemical components which are transported by diffusion or convection away from the surface into the bulk of the solution. The solvent molecules form a shell around the dissolved solid particles that are mobile in the liquid phase. The specific interactions of the components of the liquid with the solid determine the reaction rate, which is attributed to a greater etching selectivity of the solid



Chemical Milling and Photochemical Milling, Fig. 2 PCM process in metal sheet etching. *1*. Chemically clean the metal surface. *2*. Coat both sides of the plate with photoresist that adheres to the metal when exposed to UV light. *3*. Expose plate and phototool to ensure image transfer. *4*. Develop and create photomask image. *5*. Spray metal with etchant or dip it in hot acidic solution to etch all material other than part covered with photoresist. *6*. Rinse the plate to ensure photoresist and etchant removal Ref. [2]



Chemical Milling and Photochemical Milling, Fig. 3 Schematic illustration of etch detail through line openings in the patterned photoresist (*blue*). Etch factor = Depth of etch (*D*)/Undercut [½ (B − A)] Ref. [2]

than dry etching methods. Water is used as the solvent in most wet etching processes [3].

1. In a metal or a semiconductor, the dissolution of metal or semiconductor is accompanied by an electron transfer and obeys the laws of electrochemistry. The oxidation reaction of metal or semiconductor M releases metal ions into solution and produce electrons.

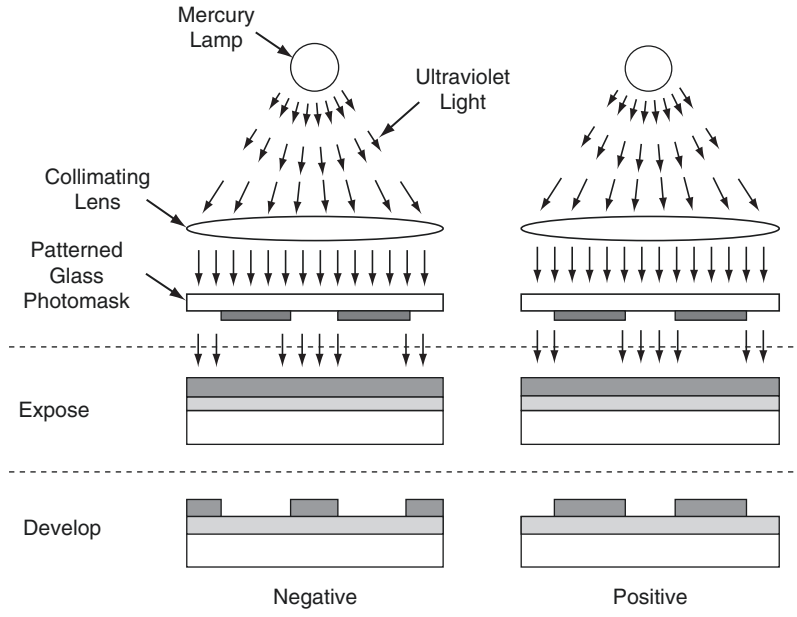
$$M = M^{n+} + ne^{-} \quad (\text{anodic partial process})$$

A secondary chemical redox process takes place to transfer the liberated electrons to an oxidizing agent OM in the solution.

$$OM + ne^{-} = OM^{n-} \quad (\text{cathodic partial process})$$

The anodic and cathodic partial processes results in etching metal.

$$M_{\text{in solid}} + OM_{\text{in solution}} = M^{n+} + OM^{n-}$$



Chemical Milling and Photochemical Milling, Fig. 4 Positive and negative photoresist Ref.[2]

The etch rate corresponds to the number of metal ions produced at the solid surface per time unit, which is proportional to the interchanged anodic partial current I over the surface A, I/A. Metals and semiconductors often dissolve as complexes (e.g., $[MY_x]^+$) where smaller molecules or ions (ligands) form a chemically bound primary shell around the central atom. The etch rate can be changed by varying the concentration of reactants, temperature, viscosity, and convection of the solution.

2. In dielectric materials, acid–base reactions take place if the material to be etched reacts with hydrogen or hydroxyl ions. The cations are solvated by water as a strong polar solvent and they can diffuse rapidly into the bulk of the solution. Etching processes are applicable with oxides and hydroxides of metals and semiconductors at low pH-values,

$$M_xO_y + 2yH^+ = xM^{2y/x+} + yH_2O$$
$$M(OH)_x + xH^+ = M^{x+} + xH_2O$$

M = metal or semiconductor (e.g., Cu, Si)

Similarly, at very high pH-values some metals and semiconductors form stable water dissolvable complexes that are easily solvated by water.

Acidic anions, such as chloride and fluoride, or neutral molecules react as ligands, a typical example of which is the etching of SiO_2 in HF-containing etchants.

$$SiO_2 + 6HF = H_2SiF_6 + H_2O$$

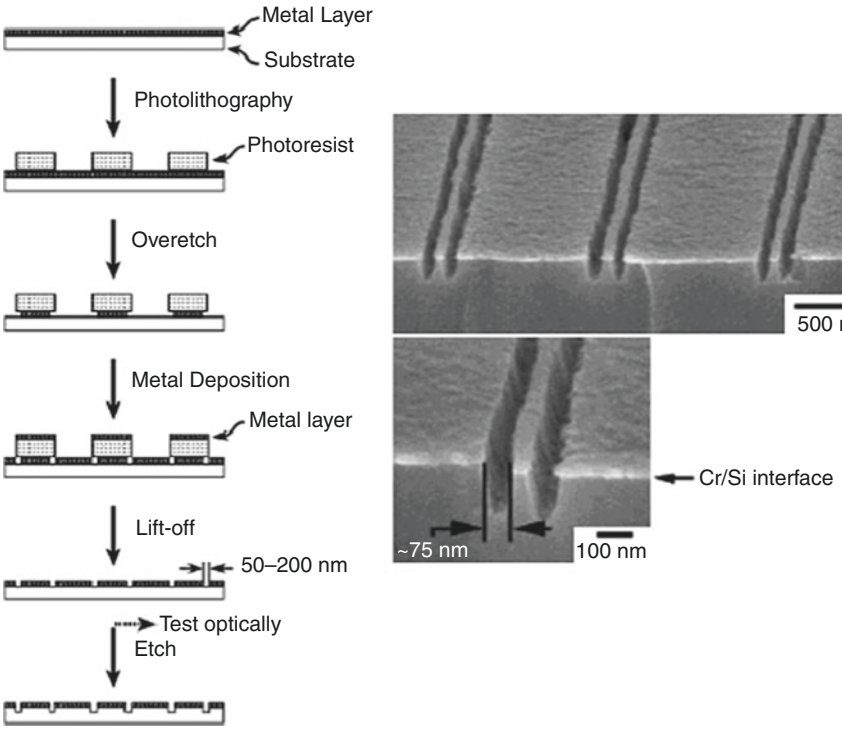
Salt-like film is also dissolved by complexing agents. For example, copper chloride is etchable in neutral KCl solutions.

$$CuCl + 2Cl- = [CuCl_3]^{2-}$$

Key Research Findings for Nanotechnology

Edge Lithography

Undercutting by isotropic wet etching is applied to transfer the edges of a photoresist pattern into a feature of the final pattern, as illustrated in Fig. 5. The process generates 50 nm scale trenches by controlled undercutting. Currently, patterning features <100 nm are possible by advanced lithography techniques – deep ultraviolet, electron beam writing, extreme ultraviolet, and x-ray photolithography – but are prohibitively



Chemical Milling and Photochemical Milling, Fig. 5 (*Left*) Schematic illustration of the process generating nanometer-scale lines by controlled undercutting. A pattern is produced in the photoresist by photolithography or soft lithography. An isotropic etch is applied to the substrate beneath the photoresist. Shallow undercutting of the base layer, followed by evaporation into the exposed areas, and lift-off generates ~50 ± 200 nm gaps in the thin film at the edges of the photoresist pattern. The pattern is then used as an optical filter. (*Right*) SEM image of a cross section of linear trenches at the edges of a 100 nm line transferred into a Si <100 > substrate. The trenches are 250 nm deep Ref. [4]

expensive. In contrast, edge lithography is a convenient, inexpensive technique for patterning features with nanometer-scale dimensions.

Wet Etching for Maskless Patterning

In the photochemical etching process, light exposure can increase the charge carrier density in near surface areas which enhances the anodical as well as the cathodical partial processes in wet etching. For example, defect electrons in the lower band left by light exposure support the release of cations and at the same time the released electrons in the conduction band are readily accepted by an oxidizing agent in the solution. In the same process, the intensive exposure with a focused beam enables direct pattern generation without a lithographic mask. This method is specified preferentially for patterning semiconductors deposited on a nonconducting substrate (e.g., GaN on sapphire) [5].

Future Directions for Research

Wet chemical etching is a simple, inexpensive, and well-understood process. The process plays a key role in the field of MEMS and nanotechnology. Continuous effort is being made to provide controllability, repeatability, and, most importantly, detail in fabricating microstructures. New etchant compositions have been developed to apply wet chemical etching to the materials constituting new devices. A representative example is selective removal of metal nitride films on sapphire for a

new version of a light-emitting diode. At the same time, great efforts have been made to develop new lithography techniques for nanoscale patterns such as proximal probe lithography, very thin to monolayer lithography, and soft lithography. PCM equipped with the emerging lithography techniques can cause the paradigm shift in the creation of nanoscaled features [6].

Cross-References

- ▶ Dry Etching Processes
- ▶ DUV Photolithography and Materials
- ▶ Electron Beam Lithography (EBL)
- ▶ EUV Lithography
- ▶ Nanoimprint Lithography
- ▶ Nanotechnology
- ▶ Stereolithography

References

1. Abate, K.: Photochemical etching of metals. Met. Finish. **100**(6A), 448–451 (2002)
2. Allen, D.M.: Photochemical machining: from manufacturing's best kept secret to a $6 billion rapid manufacturing process. CIRP J. Manuf. Syst. **53**, 559–572 (2005)
3. Köhler, J.M.: Wet chemical etching method. In: Etching in Microsystem Technology. Wiley, Weinheim (1999)
4. Love, J.C., Paul, K.E., Whitesides, G.M.: Fabrication of nanometer-scale features by controlled isotropic wet chemical etching. Adv. Mater. **13**(8), 604–607 (2001)
5. Bardwell, J.A., Webb, J.B., Tang, H., Fraser, J., Moisa, S.: Ultraviolet photoenhanced wet etching of GaN in K2S2O8 solution. J. Appl. Phys. **89**(7), 4142–4149 (2001)
6. Madou, M.: Fundamentals of Microfabrication, 2nd edn. CRC Press, Boca Raton (2002)

Chemical Modification

▶ Nanostructures for Surface Functionalization and Surface Properties

Chemical Solution Deposition

▶ Sol–Gel Method

Chemical Vapor Deposition (CVD)

Yoke Khin Yap and Dongyan Zhang
Department of Physics, Michigan Technological University, Houghton, MI, USA

Synonyms

Aerosol-assisted chemical vapor deposition (AACVD); Atmospheric pressure chemical vapor deposition (APCVD); Atomic layer chemical vapor deposition (ALCVD); Atomic layer deposition (ALD); Atomic layer epitaxial (ALE); Boron nitride nanotubes (BNNTs); Carbon nanotubes (CNTs); Carbon nanowalls; Catalyst; Catalytic chemical vapor deposition (CCVD); Chemical vapor deposition (CVD); Cold-wall thermal chemical vapor deposition; Dissociated adsorption; Double-walled carbon nanotubes (DWCNTs); Graphene; High-pressure carbon monoxide (HiPCO); Hot filament chemical vapor deposition (HFCVD); Hot-wall thermal chemical vapor deposition; Inductively coupled-plasma chemical vapor deposition (ICP-CVD); Low-pressure chemical vapor deposition (LPCVD); Metalorganic chemical vapor deposition (MOCVD); Multiwalled carbon nanotubes (MWCNTs); Nanobelts; Nanocombs; Nanoparticles; Nanotubes; Nanowires; Plasma-enhanced chemical vapor deposition (PECVD); Single-walled carbon nanotubes (SWCNTs); Thermal chemical vapor deposition; Ultrahigh vacuum chemical vapor deposition (UHVCVD); Vertically-aligned carbon nanotubes

Definition

Chemical vapor deposition (CVD) is referred to deposition process of thin films and nanostructures through chemical reactions of vapor-phase precursors. Since CVD can be conducted using high purity precursors, it likely leads to thin film and nanostructures with high purity. The use of vapor-phase precursors also enables better control on the composition and

doping of thin films and nanostructures. For example, Si thin films can be deposited by decomposition of silane gas (SiH_4) by plasma or heat as follows: SiH_4 (g) → Si (s) + $2H_2$ (g). Doping of Si films with boron will lead to p-type Si films and can be achieved by the addition of B_2H_6 gas.

Classification

Classification by Operating Pressures

Chemical reactions involved in a CVD technique can be initiated by many ways leading to the classification of various types of CVD approaches. For example, CVD can be classified according to the operating pressures as follows:

- *Atmospheric pressure CVD* (APCVD) is referred to CVD processes that are conducted at atmospheric pressure. The advantage of APCVD is a simple experimental setup without the need of a vacuum system. The potential drawback will be the undesired contamination.
- *Low-pressure CVD* (LPCVD) is referred to CVD processes at pressures below and close to atmospheric pressure. One of the purposes of reducing the operation pressure is to avoid undesired reactions between precursors. LPCVD can also improve film uniformity.
- *Ultrahigh vacuum CVD* (UHVCVD) is referred to LPCVD processes at a very low pressure, typically below ~10^{-8} torr.

There are *CVD* processes that operate at high pressures. See details in section "Classification by Excitation Techniques."

Classification by Excitation Techniques

In addition, CVD can be classified by the excitation techniques that initiate the chemical reactions. For example:

- *Plasma-enhanced CVD* (PECVD) is referred to CVD processes that employed plasmas to initiate the needed chemical reactions. In general, PECVD could reduce the growth temperatures as the chemical reactions in PECVD are not ignited by heat. There are many subclassification of PECVD depending on the type of AC potential used for plasma generation. For example, RF-PECVD and MW-PECVD employed radio frequency (RF, typically at 13.56 MHz) and microwave (MW, 2.45 GHz) potential to dissociate gases and produce the needed chemical reactions. Some PECVD are classified by the configuration of plasma generation. For instance, inductively coupled-plasma CVD (ICP-CVD) is actually RF-PECVD that uses an induction coil as the RF electrode outside the vacuum chamber and is sometimes called remote plasma CVD [1]. On the other hand, two RF plasmas can be used in a CVD system. For example, a dual RF-PECVD approach was demonstrated for the growth of vertically aligned carbon nanofoils/nanowalls and carbon nanotubes (CNTs) [2, 3].
- *Thermal CVD* is referred to CVD processes that employed heat to initiate the needed chemical reactions. The most common thermal CVD technique employed external furnace to control the growth temperatures of the entire reaction zone (e.g., vacuum quarts/glass chambers). This is sometimes called *hot-wall thermal CVD*. The advantages of this approach are including the potential of large-scale synthesis. In contract, there are several approaches to achieve the so-called *cold-wall thermal CVD*. For example, hot filaments (HF) are used to heat up the temperatures of the adjacent substrates while the needed chemical reactions are ignited with higher temperatures on the filaments. This approach is called *hot filament CVD* (HFCVD). Other heating approaches can also be used including IR lamps, laser beams, and passing current flows through a suspended Si chip [4].

Classification by the Precursor Type and Feeding Procedure

CVD can also be classified by the type of precursors or the procedures where precursors are introduced into the reaction chamber. Some of the examples are described as follows:

- *Aerosol-assisted CVD (AACVD)* involves the use of carrier gases (usually inert gases such as

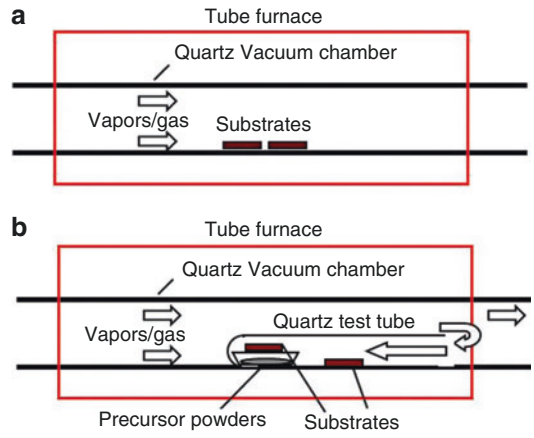
Ar, He, etc.) to transfer vapors of liquid-phase precursor into the reaction chamber in the form of aerosol. This approach enables the use of liquid precursors for the CVD process. The most well-known AACVD is *metalorganic CVD* (*MOCVD*) where metal organic solids are dissolved in organic solvents. For example, trimethylgallium [TMGa, $Ga(CH_3)_3$] is often used as the source of Ga to form GaN when it reacts with ammonia (NH_3) at high temperatures. Thus *AACVD* can be viewed as a subclassification of *thermal CVD*.

- *Atomic layer CVD* (*ALCVD*) is better known as *atomic layer deposition* (*ALD*) or *atomic layer epitaxial* (*ALE*) [5]. ALD allows conformal deposition of monolayer of binary (or ternary) compounds by introducing the two (or three) reacting precursors in a pulsed mode one after another. For example, monolayers of zinc sulfide (ZnS) can be deposited by first exposing the growth surface with $ZnCl_2$. Once chemisorption of $ZnCl_2$ is completed, the reaction chamber will be purged with an inert gas. Thereafter, hydrogen sulfide (H_2S) will be introduced to the chamber so that the following reaction will take place on the growth surface: $ZnCl_2$ (g) + H_2S (g) → ZnS (s) + HCl (g). Since the chemical process is self-limiting, only one monolayer (or less) of ZnS will be formed in each cycle. The advantage of *ALD* is conformal coating on the full surface of the sample and the precision of film thickness control. The major drawback is the extreme low deposition rate. *ALD* is usually conducted at low temperatures (200–400 °C) and may be viewed as a subclassification of *thermal CVD*. However, plasma and metal organic precursors are sometimes used in *ALD*.

Examples of CVD Approaches for Nanotechnology

Catalytic Chemical Vapor Deposition (CCVD)

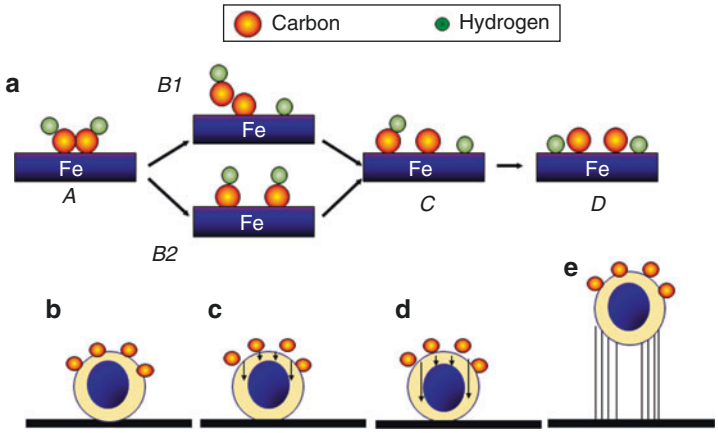
Catalytic chemical vapor deposition (*CCVD*) is the simplest and most popular technique for the synthesis of carbon nanotubes (CNTs), graphene, and nanowires (NWs). *CCVD* is simply a *thermal CVD* approach with the use of catalyst that induces the formation of nanomaterials. A typical experimental layout for *CCVD* is shown in Fig. 1. As shown, the *CCVD* system consists of a tube furnace and a quartz tube chamber where chemical reactions are taking place.



Chemical Vapor Deposition (CVD), Fig. 1 (**a**) Typical setup for CCVD. (**b**) Modified double-tube configuration

Synthesis of Vertically Aligned Carbon Nanotubes by CCVD

A typical experimental setup of *CCVD* for the synthesis of carbon nanotubes (CNTs) is shown in Fig. 1a. In this case, the substrates (usually oxidized Si) are first deposited with catalyst films such as Fe, Ni, or Co by electron beam evaporation, sputtering, pulsed laser deposition, etc. [6–8]. Then, these samples were then annealed at 600 °C in Ar, N_2, or H_2 for about 30 min. During the annealing process, these catalyst films will be converted into nanoparticles that will serve as the growth sites for CNTs. After the annealing, the growth temperatures (~600–800 °C) and growth ambient (usually Ar, H_2, or their mixtures) will be set prior to the introduction of hydrocarbon source gas (methane, CH_4; ethylene C_2H_4, acetylene, C_2H_2). These hydrocarbon gases will be decomposed on the surface of the catalyst nanoparticles through the chemical process of dissociative adsorption [6].

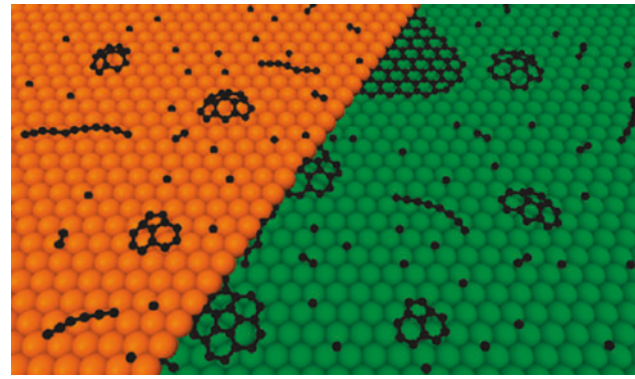


Chemical Vapor Deposition (CVD), Fig. 2 (**a**) Sequences of dissociative adsorption of C_2H_2 on Fe surface (see text for detailed description). The (**b**) decomposed carbon atoms (**c**) diffused into the subsurface of the solid-core Fe nanoparticle until (**d**) supersaturation and then (**e**) segregate as nanotubes

For example, dissociated adsorption of C_2H_2 is summarized in Fig. 2 [6]. Figure 2a shows the adsorption of a C_2H_2 molecule (step *A*) on the surface of the Fe nanoparticle. This will lead to either the breaking of C–H bond (step *B1*) to form C_2H and H fragments or the breaking of C=C bond (step *B2*) to form two C–H fragments. The catalytic function of the Fe nanoparticle is to reduce the energy required for decomposition by a charge transfer from hydrocarbon molecules to Fe. According to a first principle calculation, the dissociation energy of the first hydrogen atom from an isolated C_2H_2 (step *A* to *B1*) in vacuum can be reduced from 5.58 to 0.96 eV. On the other hand, the energy barrier between *A* and *B2* is 1.25 eV. The C–H bond breaking (step *B1*) is followed by C–C bond breaking (step *C*) with a potential barrier of 1.02 eV, whereas C=C bond breaking (step *B2*) is followed by C–H bond breaking (step *C*) with an energy barrier of 0.61 eV. Both modes (*A* to *B1* to *C or A* to *B2* to *C*) are possible and give one C–H fragment, one C, and one H. The decomposition of C_2H_2 is completed after the breaking of the last C–H bond (step *D*) with the need of a potential energy of 0.61 eV.

The decomposed carbon atoms (Fig. 2b) will then diffuse into the subsurface of the Fe nanoparticle (Fig. 1c). At typical growth temperature (650–800 °C), these nanoparticles are not melted even after considering the eutectic point of Fe-C phase. Since dissociative adsorption is an exothermic process, the near surface temperatures of the catalytic nanoparticles will be higher than the growth temperatures. Since the melting of nanoparticles starts from their surfaces, it is possible that the near surface region of the particles is melted. This will form the gas-liquid interface between carbon and Fe solid-core nanoparticles. Due to the high diffusion, rate of carbon in Fe melts; a Fe-C alloy will start to form. When these nanoparticles become supersaturated with carbon (Fig. 2d) to a value critical for growth at the solid-liquid interface, the excess carbon will segregate as carbon nanotubes (Fig. 2e). A tip-growth mode is illustrated in this figure where the nanoparticles remained at the tips of CNTs. A base-growth mode is also possible where the nanoparticles remained at the bases of CNTs. Thus, the diameters of CNTs depend to certain extent on the diameters of the nanoparticles. By controlling the thickness of the catalyst films and other parameters, *CCVD* enables the growth of vertically aligned (VA) single-, double-, and multi-walled CNTs [8].

In fact, one of the most impressive achievements in the growth of vertically aligned single-walled CNTs (VA-SWCNTs) is the so-called super growth [9]. This water-assisted approach was based on *CCVD* with the addition of 20–500 ppm of water vapors. These water vapors were introduced into the growth chamber by flowing carrier gas (Ar, or He with 40 % H_2) through a water bubbler. Ethylene was used as the carbon source along with various catalysts



Chemical Vapor Deposition (CVD), Fig. 3 Nucleation of graphene on Ni (111) surface (Reprinted with permission from J. Gao et al., J. Am. Chem. Soc. C **133**, 5009 (2011). Copyright (2011) American Chemical Society)

[Fe (1 nm), Al (10 nm)/Fe (1 nm), Al_2O_3 (10 nm)/Fe (1 nm), to Al_2O_3 (10 nm)/Co (1 nm)]. VA-SWCNTs with the length of several mm and diameter of 1–3 nm were reported. This approach also enabled the growth of double- and multi-walled CNTs [10] (Fig. 3).

The abovementioned *CCVD* approaches are achievable at relatively low temperatures and lead to VA-SWCNTs. Historically SWCNTs were grown by *CCVD* at higher temperatures in a powder form. In 1996, Dai et al. demonstrated the growth of SWCNTs at 1200 °C by using carbon monoxide (CO) as the carbon source gas and supported MoO_x powder as the catalyst [11]. In this case, powder form of SWCNTs was grown on the catalyst that was loaded on a quartz boat. Later, the growth temperature was being able to reduce to 1000 °C by using CH_4 as the source gas, resulting in the growth of randomly distributed SWCNTs on powders of supported metal-oxide catalysts [12]. The diameters of these SWCNTs are 1–6 nm. In addition, *CCVD* could also be modified into a high-pressure mode for large-scale synthesis of SWCNTs [13]. In this approach, liquid iron pentacarbonyl, $Fe(CO)_5$, was used as the catalyst and was introduced into the CVD chamber by CO gas. At operational pressures of 1–10 atm and temperatures of 800–1200 °C, $Fe(CO)_5$ will thermally be decomposed into iron clusters in gas phase and react with CO gas to produce SWCNTs. The yield of SWCNTs was found to increase with temperature and pressure. The average diameter of SWCNTs was decreased from ~1.0 nm at 1 atm to 0.8 nm at 10 atm. This approach is now known as high-pressure carbon monoxide (HiPCO) technique, one of the major techniques for large-scale synthesis of SWCNTs with small diameters. Finally, alcohol was also used as the carbon source for the synthesis of SWCNTs by *CCVD* [14]. In this case, ethanol vapors (5 torr) were supplied to the reaction chamber that contained Fe/Co catalyst supported with zeolite at 700–800 °C.

Synthesis of Graphene by CCVD

CCVD is also applicable for the synthesis of two-dimensional (2D) materials. For example, the growth of graphene by CCVD has been demonstrated using Ni [15, 16] and Cu plates [17]. Theory suggests that the decomposition of hydrocarbon gases on Ni or Cu are similar to the dissociative adsorption discussed for the growth of CNTs discussed earlier with differences on the atomic configuration [18]. The major differences are (1) carbon atoms dissolve into the bulk / sub-surface of the Ni plate and led to the formation of few layered graphene (FLG); (2) contrarily, the growth process is limited on the surface in the Cu case and result in the formation of mono layered graphene (MLG) [17]. Theory also suggests that the growth of graphene on Cu surface by methane (CH4) gas could happened with hydrocarbon radical instead of free carbon atoms [18].

The nucleation of 2D materials such as graphene is also different from that in 1D CNTs. for the case of CNTs, nucleation is confined on

the surface and subsurface regions of the zero dimensional metal particles. The nucleation of graphene involves a lot more surface diffusion processes [19]. As shown in Fig. 3, the process involves the transformation of 1D carbon chain to 2d sp2 carbon clusters/networks. Furthermore, it was suggested that graphene nucleation near a step edge is more preference to that on a terrace.

Synthesis of ZnO Nanostructures by CCVD

On the other hand, *CCVD* can be modified into a *double-tube* configuration as shown in Fig. 1b. This approach will enable the use of solid precursors to generate the needed growth vapors. For example, various ZnO nanostructures can be grown by such a setup by using ZnO and graphite powders as the precursors [20, 21]. As shown, these powders were loaded on a ceramic combustion boat which is contained at the end of a closed-end quartz tube. The substrates can be placed a distance away from the boat. The following reactions will occur at 1100 °C when oxygen gas (O_2) is introduced:

$$\mathrm{ZnO\,(s)\ +\ C\,(s)\ \rightarrow\ Zn\,(g)\ +\ CO\,(g)}$$

$$\mathrm{2Zn\,(g)\ +\ O_2(g)\ \rightarrow\ 2ZnO\,(s)}$$

Based on this approach, various ZnO nanostructures can be grown with and without the use of catalyst (gold films, Au), including nanotubes [20], nanowires, nanobelts, nanocombs [21], and nanosquids [22].

Synthesis of Boron Nitride Nanotubes by CCVD

The double-tube *CCVD* configuration in Fig. 1b was also used for the growth of boron nitride nanotubes (BNNTs) [23–25]. In this case, boron, magnesium oxide, and iron oxide were used as the precursors and were loaded on the ceramic boat. The possible chemical reaction at 1100–1200 °C is 4B (s) + MgO (s) + FeO (s) → $2B_2O_x$ (g) + MgO_y (s) + FeO_z (s), where x, y, and z are yet to be determined. When ammonia gas (NH_3) is introduced into the chamber, the generated boron oxide vapors (B_2O_x) will react with NH_3 to form BNNTs. The partially reduced MgO_y and FeO_z are the possible catalysts for the formation of BNNTs [23]. It is noted that the synthesis of BNNTs by these chemical processes usually requires temperatures above 1350 °C. The key success for the low-temperature growth discussed here is due to the so-called "growth vapor trapping" approach obtained by placing the bare substrates directly on the ceramic boat. These substrates trapped the growth vapors from the precursors and enhanced the nucleation rate of BNNTs at low temperatures. In recent experiments, Fe, Ni, or MgO films were coated on the substrates used as the catalysts [24, 25]. Such an approach leads to patterned growth of BNNTs. The possible chemical reactions are

$$\mathrm{B_2O_2(g)\ +\ MgO\,(s)\ +\ 2NH_3(g)\ \rightarrow\ 2BN\,(s,\ BNNTs)\ +\ MgO(s)\ +\ 2H_2O(g)\ +\ H_2(g),\ or}$$

$$\mathrm{B_2O_2(g)\ +\ Ni\ (s)\ or\ Fe\ (s)\ +\ 2NH_3(g)\ \rightarrow\ 2BN\ (s,\ BNNTs)\ +\ Ni\ (s)\ or\ Fe\ (s)\ +\ 2H_2O\ (g)\ +\ H_2(g)}$$

Acknowledgments Yoke Khin Yap acknowledges the support from the National Science Foundation (Award number DMR-1261910).

Cross-References

- ▶ Atomic Layer Deposition
- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Carbon-Nanotubes
- ▶ Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
- ▶ Synthesis of Carbon Nanotubes

References

1. Tsu, D.V., Lucovsky, G., Dvidson, B.N.: Effects of the nearest neighbors and the alloy matrix on SiH stretching vibrations in the amorphous SiOr:H (0<r<2) alloy system. Phys. Rev. B. **40**, 1795–1805 (1989)
2. Menda, J., et al.: A Dual-RF-Plasma Approach for Controlling the Graphitic Order and Diameters of Vertically-Aligned Multiwall Carbon Nanotubes. Appl. Phys. Lett. **87**, 173106 (2005) (3 pp)

3. Hirao, T., et al.: Formation of vertically aligned carbon nanotubes by dual-RF-plasma chemical vapor deposition. Jpn. J. Appl. Phys. **40**, L631–L634 (2001)
4. van Laake, L., Hart, A.J., Slocum, A.H.: Suspended heated silicon platform for rapid thermal control of surface reactions with application to carbon nanotube synthesis. Rev. Sci. Instrum. **78**, 083901 (2007) (9 pp)
5. Leskelä, M., Ritala, M.: Atomic layer deposition chemistry: Recent developments and future challenges. Angew. Chem. Int. Ed. **42**, 5548–5554 (2003)
6. Kayastha, V.K., et al.: Controlling dissociative adsorption for effective growth of carbon nanotubes. Appl. Phys. Lett. **85**, 3265–3267 (2004)
7. Kayastha, V.K., et al.: High-density vertically aligned multiwalled carbon nanotubes with tubular structures. Appl. Phys. Lett. **86**, 253105 (2005) (3 pp)
8. Kayastha, V.K., et al.: Synthesis of vertically aligned single- and double- walled carbon nanotubes without etching agents. J. Phys. Chem. C **111**, 10158–10161 (2007)
9. Hata, K., et al.: Water-assisted highly efficient synthesis of impurity-free single-walled carbon nanotubes. Science **306**, 1362–1364 (2004)
10. Yamada, T., et al.: Size-selective growth of double-walled carbon nanotube forests from engineered iron catalysts. Nat. Nanotechnol. **1**, 131–136 (2006)
11. Dai, H., et al.: Single-wall nanotubes produced by metal-catalyzed disproportionation of carbon monoxide. Chem. Phys. Lett. **260**, 471–475 (1996)
12. Kong, J., Cassell, A.M., Dai, H.: Chemical vapor deposition of methane for single-walled carbon nanotubes. Chem. Phys. Lett. **292**, 567–574 (1998)
13. Nikolaev, P., et al.: Gas-phase catalytic growth of single-walled carbon nanotubes from carbon monoxide. Chem. Phys. Lett. **313**, 91–97 (1999)
14. Maruyama, S., et al.: Low-temperature synthesis of high-purity single-walled1 carbon nanotubes from alcohol. Chem. Phys. Lett. **360**, 229–234 (2002)
15. Obraztsov, A.N., Obraztsova, E.A., Tyurnina, A.V., Zolotukhin, A.A.: Chemical vapor deposition if thin graphite films of nanometer thickness. Carbon **45**, 2017–2021 (2007)
16. Kim, K.S. et al.: Large-scale pattern growth of graphene films for stretchable trasparent electrodes. Nature **457**, 706–710 (2009)
17. Li, X.S. et al.: Large-area synthesis of high-quality and uniform graphene films on copper foils. Science **324**, 1312–1314 (2009)
18. Zhang. W., Wu, P., Li, Z., Yang, J.: First-principles thermodynamics of graphene growth on Cu surfaces. J. Phys. Chem. C **115**, 17782–17787 (2011)
19. Gao, J., Yop, J., Zhao, J., Yakobson, B.I., Ding, F.: Graphene nucleation on transition metal surface: Structure transformation and role of the metal step edge. J. Am. Chem. Soc. **133**, 5009–5015 (2011)
20. Mensah, S.L., et al.: Formation of single crystalline ZnO nanotubes without catalysts and templates. Appl. Phys. Lett. **90**, 113108 (2007)
21. Mensah, S.L., et al.: Selective growth of pure and long ZnO nanowires by controlled vapor concentration gradients. J. Phys. Chem. C **111**, 16092–16095 (2007)
22. Mensah, S.L., et al.: ZnO nnosquids: banching nnowires from nnotubes and nnorods. J. Nanosci. Nanotechnol. **8**, 233–236 (2008)
23. Lee, C.H., et al.: Effective growth of boron nitride nanotubes by thermal chemical vapor deposition. Nanotechnology **19**, 455605 (2008)
24. Lee, C.H., et al.: Patterned growth of boron nitride nanotubes by catalytic chemical vapor deposition. Chem. Mater. **22**, 1782–1787 (2010)
25. Wang, J., Lee, C.H., Yap, Y.K.: Recent advancements in boron nitride nanotubes. Nanoscale. **2**, 2028–2034 (2010)

Chemical Vapor Machining (CVM)

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Chemical-Assisted Ion Beam Etching (CAIBE)

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Chemistry of Carbon Nanotubes

▶ Functionalization of Carbon Nanotubes

Chemotherapeutic

▶ Use of Nanotechnology in Pregnancy

Chitosan

▶ Chitosan Nanoparticles

Chitosan Nanoparticles

Burcu Aslan[1], Hee Dong Han[2,4], Gabriel Lopez-Berestein[1,3,4,5] and Anil K. Sood[2,3,4,5]
[1]Department of Experimental Therapeutics, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[2]Gynecologic Oncology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[3]Cancer Biology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[4]Center for RNA Interference and Non–coding RNA, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[5]The Department of Nanomedicine and Bioengineering, UTHealth, Houston, TX, USA

Synonyms

Chitosan; Drug delivery system; Nanoparticles

Definition

Chitosan nanoparticles are biodegradable, nontoxic carriers for nucleotides and drugs with the potential for broad applications in human disease.

Overview

Characteristics of Chitosan

Chitosan is a natural cationic polysaccharide composed of randomly distributed *N*-acetyl-D-glucosamine and β-(1,4)-linked D-glucosamine. Chitosan can be chemically synthesized via alkaline deacetylation from chitin, which is the principal component of the protective cuticles of crustaceans [1]. Chitosan is biodegradable in vivo by enzymes such as lysozyme, which is endogenous and nontoxic [2]. In addition, biodegradation of chitosan is highly associated with the degree of deacetylation. These properties render chitosan particularly attractive for clinical and biological applications as a highly biocompatible material with low toxicity and immunogenicity.

Several techniques have been designed to assemble chitosan nanoparticles (CH-NPs) as a drug delivery system including emulsions, ionotropic gelation, micelles, and spray drying. A variety of therapeutic agents can be loaded into CH-NPs with high efficiency, which can then be injected intravenously, intraperitoneally, or intrathecally.

Chemical Modification of Chitosan

The abundant amine and hydroxyl groups present in chitosan offer a unique opportunity to attach targeting ligands or imaging agents. Numerous derivatives of chitosan have been designed and tailored to improve the physicochemical and adhesive properties of nanoparticles such as size, shape, charge, density, and solubility. Quaternized chitosan, *N*,*N*,*N*-trimethyl chitosan, thiolated chitosan, carboxyalkyl chitosan, sugar-bearing chitosan, bile acid-modified chitosan, and cyclodextrin-linked chitosan are among the modifications frequently utilized in chitosan-based drug delivery systems. Each modification offers unique properties and characteristics. For instance, trimethyl chitosan is soluble over a wide pH range and enhances the condensation capacity of plasmid DNA at neutral pH due to fixed positive charges on its backbone. Thiolation of chitosan provides free sulphydryl groups on its side chains and forms disulfide bonds with cysteine-rich subdomains of mucoglycoproteins on cell membranes and increases cellular uptake. In addition, both modifications have been used as nonviral carrier systems to combine the advantages of trimethyl chitosan and thiolated chitosan while minimizing their shortcomings [3].

Hydrophobic moieties can also be attached to chitosan to facilitate the incorporation of insoluble drugs, i.e., hydrophobic glycol chitosan nanoparticles. Chemical conjugation of hydrophobic 5β-cholanic acid to the hydrophilic glycol chitosan backbone allows for the incorporation of the water-insoluble drug camptothecin [4].

Key Research Findings

Preparation of Chitosan Nanoparticles

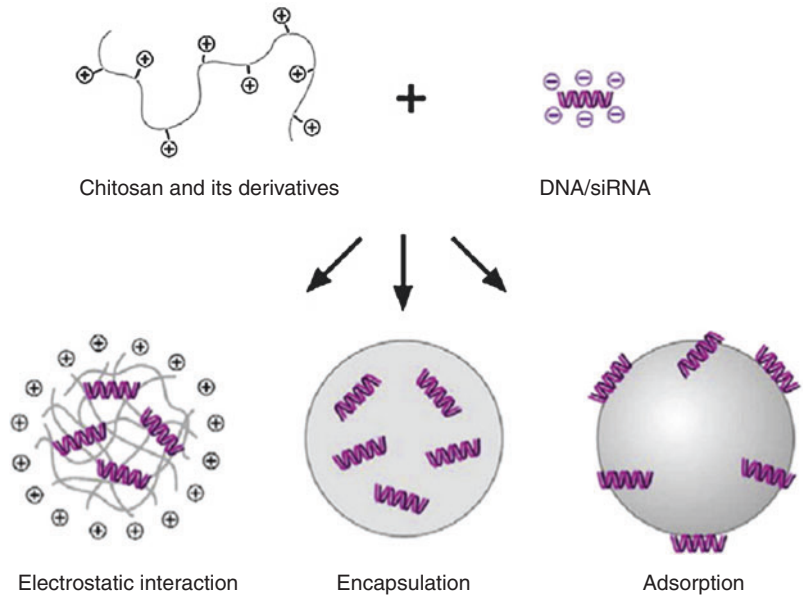
The amino groups of chitosan backbone can interact with anionic molecules such as tripolyphosphate (TPP). The ionic cross-linking of chitosan is advantageous since the method is easy and mostly performed under mild conditions without using organic solvents. Ionotropic gelation of chitosan using TPP for the incorporation of low molecular drugs, proteins, DNA/siRNA have been demonstrated [5]. CH-NPs are rapidly formed through ionic interactions between the negatively charged phosphates of TPP and positively charged amino groups of chitosan.

Drug Delivery

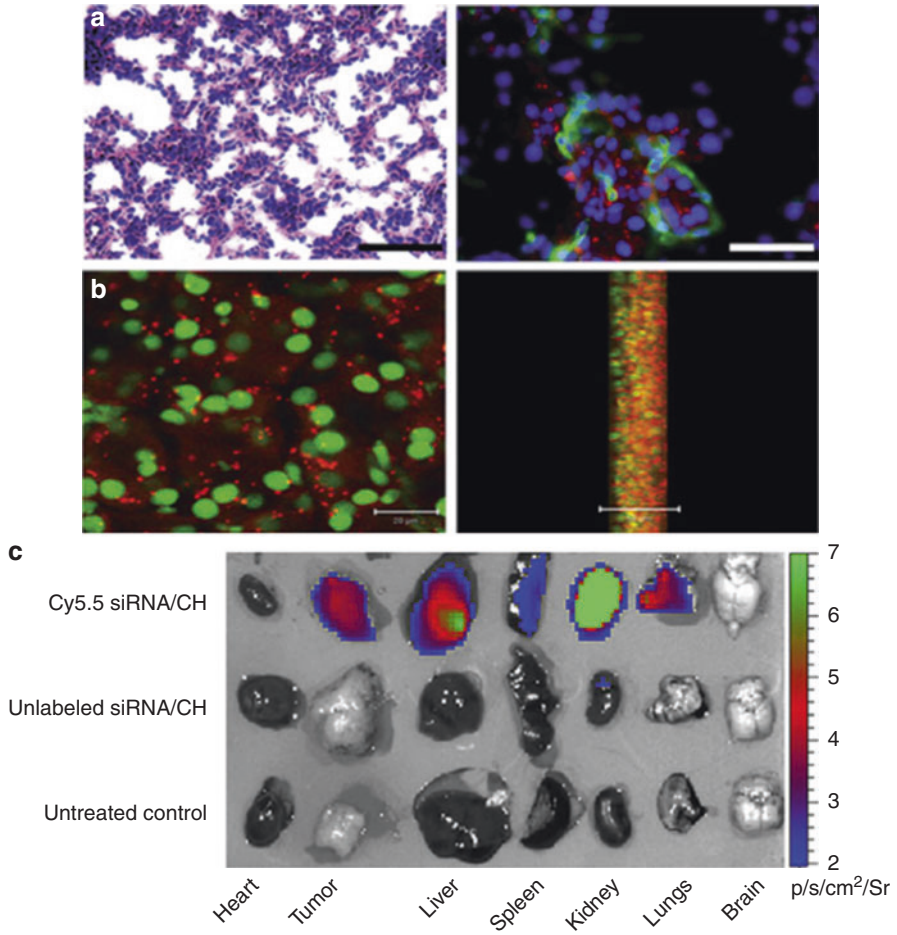
Ringsdorf first reported the concept of polymer–drug conjugates for delivering small molecule drugs [6]. The concept of polymer–drug conjugates allows chemical conjugation of a drug using a biodegradable spacer. The spacer is usually stable in the bloodstream, but cleaved at the target site by hydrolysis or enzymatic degradation. Based on this concept, several polymer–drug conjugates have been developed such as glycol chitosan conjugated with doxorubicin, which forms self-assembled nanoparticles in an aqueous condition. A paclitaxel-chitosan conjugate that can be cleaved at physiological conditions was developed for oral delivery of paclitaxel. *N*-succinyl-chitosan-mitomycin conjugates demonstrated high antitumor efficacy against a variety of murine tumor models of leukemia, melanoma, and primary and metastatic liver tumors.

Gene Delivery

Recently, plasmid DNA, siRNA, and oligonucleotide-loaded CH-NPs have been used for targeted gene silencing. Positively charged chitosan can easily form polyelectrolyte complexes with negatively charged nucleotides based on electrostatic interaction, incorporation, or adsorption, as illustrated in Fig. 1 [7]. The positive charge on the surface of CH-NPs is desirable to prevent aggregation due to electrostatic repulsion and increase binding efficiency with the negatively charged cell membrane by enhancing electrostatic interactions. However, other cellular uptake mechanisms such as clathrin-mediated endocytosis, caveolae-mediated endocytosis, and macropinocytosis may also be involved [8]. Moreover, the configuration of chitosan is modified under acidic pH by triggering the opening of tight junctions. Acidic pH results in an increase in the number



Chitosan Nanoparticles, Fig. 1 Preparation of chitosan-based DNA/siRNA nanoparticles based on different mechanisms [7]



Chitosan Nanoparticles, Fig. 2 (**a**) Fluorescent siRNA distribution in tumor tissue. Hematoxylin and eosin, original magnification 2003 (*left*); tumor tissues were stained with anti-CD31 (*green*) antibody to detect endothelial cells (*right*). The scale bar represents 50 mm. (**b**) Fluorescent siRNA distribution in tumor tissue. Sections (8 mm thick) were stained with Sytox green and examined with confocal microscopy (scale bar represents 20 mm) (*left*); lateral view (*right*). Photographs taken every 1 mm were stacked and examined from the lateral view. Nuclei were labeled with Sytox green and fluorescent siRNA (*red*) was seen throughout the section. At all time points, punctated emissions of the siRNA were noted in the perinuclear regions of individual cells, and siRNA was seen in >80 % of fields examined. (**c**) Shows fluorescence intensity overlaid on white light images of different mouse organs and tumor

of protonated amines on the chitosan leading to a further disruption on membrane organization. It has also been reported that chitosan can swirl across the membrane lipid bilayer and facilitate the cellular uptake of the polyplex due to increase in mole fraction of chitosan, leading to reduction in the polymeric chains which results in decreased molecular weight [7]. In addition, Lu et al. have recently reported the biodistribution of intravenously administered siRNA-CH-NPs in tumor-bearing mice. They demonstrated that CH-NPs allowed for a higher localization of siRNA in tumor tissues compared to other organs (Fig. 2) [9].

Electrostatic interactions between protonated amines of chitosan and negative charge of DNA or siRNA leads to spontaneous formation of highly compact encapsulation of either DNA or siRNA into CH-NPs [10]. Gel electrophoresis has been used to assess hydrogen bonding and hydrophobic interactions between chitosan and DNA [11]. However, a major limitation of siRNA

delivery is its rapid degradation in plasma and cytoplasm. It has been reported that stable chitosan/siRNA complexes can protect siRNA degradation in circulation of CH-NPs in bloodstream to overcome extracellular and intracellular barriers. On the other hand, disassembly is also needed to allow release of siRNA. This emphasizes the importance of an appropriate balance between protection and release of siRNA for biological functionality [12].

Positively charged CH-NPs can bind to negatively charged cell surfaces with high affinity. CH-NPs are known to overcome endosomal escape via its "proton sponge effect." Once these nanoparticles penetrate into an acidifying lysosomal compartment, the unsaturated amino groups of chitosan distrain protons that are delivered by proton pumps (vATPase) which is called the "proton sponge mechanism." Subsequent lysosomal swelling and rupture leads to endolysosomal escape of nanoparticles [13].

Targeted Delivery

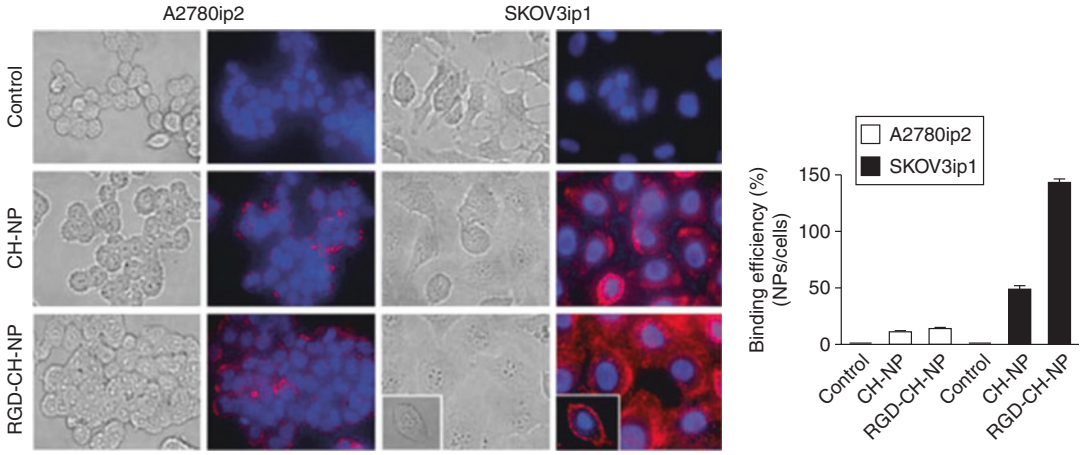
An ideal delivery system should lead to enhanced concentrations of therapeutic payloads at disease sites, and minimize potential non-desirable off-target effects, and ultimately raise the therapeutic index. Differences between tumor and normal tissue microenvironment and architecture, such as vascularization, overexpressed receptors, pH, temperature, ionic strength, and metabolites, can be exploited for selective targeting.

Targeted delivery systems have been designed to increase and/or facilitate uptake into target cells, and to protect therapeutic payloads. Recent work comparing non-targeted and targeted nanoparticles have shown that the primary role of the targeting ligands is to enhance selective binding efficiency to receptor in cell surface and cellular uptake into target cells, and to minimize accumulation in normal tissues. The addition of targeting ligands that provide specific ligand-receptor binding on nanoparticle-cell surface interactions can play a vital role in the ultimate location of nanoparticles. For example, nanoparticles decorated with specific moieties such as peptides, proteins, or antibodies can be targeted to cancer cells via cell-surface receptor proteins such as transferrin or folate receptors (known to be increased on a wide range of cancer cells). These targeting ligands enable nanoparticles to bind on to cell-surface receptors and penetrate cells by receptor-mediated endocytosis.

Target selective ligand-labeled CH-NPs can enhance receptor-mediated endocytosis. Various receptors on the tumor cell surface have been established as a target-binding site to achieve selective delivery. The overexpression of transferrin and folate in certain tumors has been exploited to deliver CH-NPs conjugated with these receptor's ligands [14, 15]. Another example is the $\alpha v\beta 3$ integrin, which is overexpressed in a wide range of tumors, and is largely absent in normal tissues. Han et al. [16] have recently reported that the administration of RGD peptide-labeled CH-NPs led to increased tumor delivery of siRNA-CH-NP and enhanced antitumor activity in ovarian carcinoma models (Figs. 3 and 4).

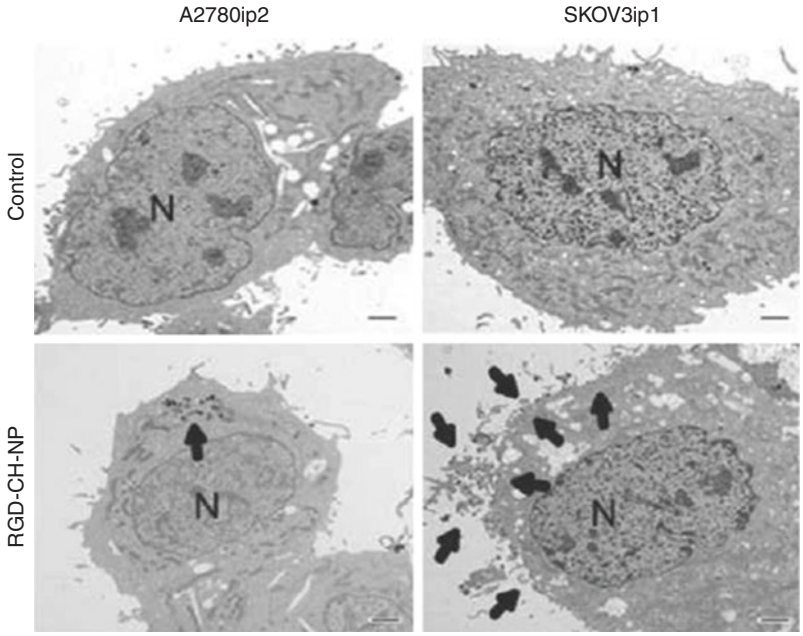
Chitosan-Based Environmental-Responsive Particles

Physiological alterations such as pH, temperature, ionic strength, and metabolites in the microenvironment of tumor have gained increased interest in terms of targeted therapy. Potential differences in these parameters between tumor and normal tissue can be used for enhanced targeting. A novel, pH responsive NIPAAm/CH-NPs containing camptothecin and paclitaxel was successfully used to enhance tumor uptake and antitumor activity [17]. On the other hand, thermosensitive CH-g-poly(*N*-vinylcaprolactam) composite has been developed by an ionic cross-linking method and incorporated with 5-FU (5-fluorouracil). This study showed that 40 % of drug is released from the particles when the temperature is above a lower critical solution temperature of 38 °C while only 5 % of drug is released below 38 °C, which confirms the drug release mechanism of this polymeric carrier system is based on temperature. These nanoparticles were more toxic to cancer cells while devoid of toxicity to normal cells, leading to enhance antitumor efficacy [18].



Chitosan Nanoparticles, Fig. 3 Binding of Alexa555 siRNA/RGD-CH-NPs and Alexa555 siRNA-CH-NP in SKOV3ip1 or A2780ip2 cells by fluorescence microscopy

Chitosan Nanoparticles, Fig. 4 Binding of siRNA/RGD-CH-NPs in SKOV3ip1 or A2780ip2 cells by transmission electron microscopy against ovarian cancer cells in vitro



Chitosan-Based Magnetic Nanoparticles

Chitosan-based magnetic nanoparticles have been developed for magnetic resonance (MR) imaging via passive and tissue-specific targeting. The low-oxidizing ferromagnetic materials are the most commonly used compounds for nanoparticle formulations and provide a stable magnetic response. The accumulation of magnetic nanoparticles when injected intravenously can be induced when the tissue or organ is subjected to an external high-gradient magnetic field [19]. Drugs, DNA plasmids, or bioactive molecules are released into target tissues and effectively taken up by tumor cells after accumulation of these carriers. On the other hand, magnetic nanoparticles such as iron oxide nanoparticles

(Fe_3O_4) are applied to oscillating magnetic fields, which results in the generation of heat and holds potential for rapid heating of tumor tissue.

When magnetic nanoparticles are administered systemically, they are rapidly coated with plasma proteins. These particles are taken up by the reticuloendothelial system, leading to decreased circulation time. Magnetic nanoparticles coated with hydrophilic materials such as chitosan could provide longer circulation time. In addition, chitosan-coated magnetic nanoparticles, which are used in magnetic resonance imaging, can also be functionalized and used as theranostic carriers due to amino and hydroxyl groups of chitosan.

Application of Chitosan Nanoparticles for SiRNA Delivery

Recently, siRNA targeted to EZH2 (a critical component of the polycomb repressive complex 2 [PRC2]), loaded CH-NPs were shown to enhance the delivery of siRNA to tumors, leading to downregulation of the target protein and subsequently enhanced antitumor activity. Lu et al. demonstrated target gene silencing using EZH2 siRNA loaded CH-NPs, leading to increased antitumor efficacy in animal tumor models. Moreover, Zhang et al. reported the use of intrathecal administration of siRNA targeted against specific muscarinic receptor subtypes loaded in CH-NPs in a rat model of pain. CH-NP/siRNA was distributed to the spinal cord and the dorsal root ganglion [20]. The administration of Chitosan-M_2-siRNA caused a large reduction in the inhibitory effect of muscarine on the rat paw withdrawal threshold from a heat stimulus. These studies support the use of CH-NPs as a delivery system for siRNA into neuronal tissues in vivo.

Future Directions for Research

Chitosan nanoparticles offer a unique potential for clinical and biological applications due to low immunogenicity, low toxicity, and high biocompatibility. In addition to its advantages such as protonated amine groups, chitosan can increase binding efficiency with cells because of electrostatic interactions. Therefore, CH-NPs may be used for broad applications in human disease. Moreover, chitosan allow modifications that will exploit the inherent physicochemical properties by conjugation of selective ligands. The highly desirable specific targeting of drugs has been elusive to date; however, CH-NPs can bring us closer to this goal.

Acknowledgments Portions of this work were supported by the NIH (CA 110793, 109298, P50 CA083639, P50 CA098258, CA128797, RC2GM092599, U54 CA151668), the Ovarian Cancer Research Fund, Inc. (Program Project Development Grant), the DOD (OC073399, W81XWH-10-1-0158, BC085265), the Zarrow Foundation, the Marcus Foundation, the Kim Medlin Fund, the Laura and John Arnold Foundation, the Estate of C. G. Johnson, Jr., the RGK Foundation, and the Betty Anne Asche Murray Distinguished Professorship.

Cross-References

- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Nanomedicine
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles

References

1. Muzzarelli, R.A.A.: Chitin, pp. 1–37. Pergamon, Elmsford (1977)
2. Khor, E.: Chitin: Fulfilling a Biomaterials Promise. Elsevier, Oxford (2001)
3. Zhao, X., Yin, L., Ding, J., Tang, C., Gu, S., Yin, C., Mao, Y.: Thiolated trimethyl chitosan nanocomplexes as gene carriers with high in vitro and in vivo transfection efficiency. J. Control. Release **144**, 46–54 (2010)
4. Min, K.H., Park, K., Kim, Y., Bae, S.M., Lee, S., Jo, H. G., Park, R.W., In-San Kim, I.S., Jeong, S.Y., Kim, K., Kwon, I.C.: Hydrophobically modified glycol chitosan nanoparticles-encapsulated camptothecin enhance the drug stability and tumor targeting in cancer therapy. J. Control. Release **127**, 208–218 (2008)
5. Park, J.H., Saravanakumar, G., Kim, K., Kwon, I.C.: Targeted delivery of low molecular drugs using chitosan and its derivatives. Adv. Drug Deliv. Rev. **62**, 28–41 (2010)
6. Ringsdorf, H.: Structure and properties of pharmacologically active polymers. J. Polym. Sci. Polym. Symp. **51**, 135–153 (1975)

7. Lai, W.F., Lin, M.C.M.: Nucleic acid delivery with chitosan and its derivatives. J. Control. Release **134**, 158–168 (2009)
8. Nam, H.Y., Kwon, S.M., Chung, H., Lee, S.Y., Kwon, S.H., Jeon, H., Kim, Y., Park, J.H., Kim, J., Her, S., Oh, Y.K., Kwon, I.C., Kim, K., Jeong, S.Y.: Cellular uptake mechanism and intracellular fate of hydrophobically modified glycol chitosan nanoparticles. J. Control. Release **135**, 259–267 (2010)
9. Lu, C., Han, H.D., Mangala, L.S., Ali-Fehmi, R., Newton, C.S., Ozbun, L., Armaiz-Pena, G.N., Hu, W., Stone, R.L., Munkarah, A., Ravoori, M.K., Shahzad, M.M.K., Lee, J.W., Mora, E., Langley, R. R., Carroll, A.R., Matsuo, K., Spannuth, W.A., Schmandt, R., Jennings, N.J., Goodman, B.W., Jaffe, R.B., Nick, A.M., Kim, H.S., Guven, E.O., Chen, Y. H., Li, L.Y., Hsu, M.C., Coleman, R.L., Calin, G.A., Denkbas, E.B., Lim, J.Y., Lee, J.S., Kundra, V., Birrer, M.J., Hung, M.C., Lopez-Berestein, G., Sood, A.K.: Regulation of tumor angiogenesis by EZH2. Cancer Cell **18**, 185–197 (2010)
10. Mao, S., Sun, W., Kissel, T.: Chitosan-based formulations for delivery of DNA and siRNA. Adv. Drug Deliv. Rev. **62**, 12–27 (2010)
11. Messai, I., Lamalle, D., Munier, S., Verrier, B., Ataman-Onal, Y., Delair, T.: Poly(D, Llactic acid) and chitosan complexes: interactions with plasmid DNA. Colloids Surf. A Physicochem. Eng. Asp. **255**, 65–72 (2005)
12. Liu, X., Howard, K.A., Dong, M., Andersen, M.O., Rahbek, U.L., Johnsen, M.G., Hansen, O.C., Besenbacher, F., Kjems, J.: The influence of polymeric properties on chitosan/siRNA nanoparticle formulation and gene silencing. Biomaterials **28**, 1280–1288 (2007)
13. Nel, A.E., Mädler, L., Velegol, D., Xia, T., Hoek, E.M. V., Somasundaran, P., Klaessig, F., Castranova, C., Thompson, M.: Understanding biophysicochemical interactions at the nano–bio interface. Nat. Mater. **8**, 543–557 (2009)
14. Mao, H.Q., Roy, K., Troung-Le, V.L., Janes, K.A., Lin, K.Y., Wang, Y., August, J.T., Leong, K.W.: Chitosan–DNA nanoparticles as gene carriers: synthesis, characterization and transfection efficiency. J. Control. Release **70**(3), 399–421 (2001)
15. Fernandes, J.C., Wang, H., Jreyssaty, C., Benderdour, M., Lavigne, P., Qiu, X., Winnik, F.M., Zhang, X., Dai, K., Shi, Q.: Bone-protective effects of nonviral gene therapy with Folate–Chitosan DNA nanoparticle containing Interleukin-1 receptor antagonist gene in rats with adjuvant-induced arthritis. Mol. Ther. **16**(7), 1243–1251 (2008)
16. Han, H.D., Mangala, L.S., Lee, J.W., Shahzad, M.M. K., Kim, H.S., Shen, D., Nam, E.J., Mora, E.M., Stone, R.L., Lu, C., Lee, S.J., Roh, J.W., Nick, A.M., Lopez-Berestein, G., Sood, A.K.: Targeted gene silencing using RGD-Labeled chitosan nanoparticles. Clin. Cancer Res. **16**(15), 3910–3922 (2010)
17. Li, F., Wu, H., Zhang, H., Gu, C.H., Yang, Q.: Antitumor drug paclitaxel loaded pH-sensitive nanoparticles targeting tumor extracellular pH. Carbohydr. Polym. **77**(4), 773–778 (2009)
18. Rejinold, N.S., Chennazhi, K.P., Nair, S.V., Tamura, H., Jayakumar, R.: Carbohydr. Polym. (2010). doi:10.1016/j.carbpol.2010.08.052
19. Veiseh, O., Gunn, J.W., Zhang, M.: Design and fabrication of magnetic nanoparticles for targeted drug delivery and imaging. Adv. Drug Deliv. Rev. **62**, 284–304 (2010)
20. Zhang, H.M., Chen, S.R., Cai, Y.Q., Richardson, T.E., Driver, L.C., Lopez-Berestein, G., Pan, H.L.: Signaling mechanisms mediating muscarinic enhancement of GABAergic synaptic transmission in the spinal cord. Neuroscience **158**, 1577–1588 (2009)

Clamping Loss

▶ Anchor Loss in MEMS/NEMS

Clinical Adhesives

▶ Bioadhesives

Cluster

▶ Synthesis of Subnanometric Metal Nanoparticles

C-MEMS

▶ Carbon MEMS

CMOS (Complementary Metal-Oxide-Semiconductor)

▶ CMOS MEMS Fabrication Technologies

CMOS MEMS Biosensors

Michael S.-C. Lu
Department of Electrical Engineering, Institute of Electronics Engineering, and Institute of NanoEngineering and MicroSystems, National Tsing Hua University, Taiwan, Republic of China

Synonyms

Integrated biosensors

Definition

CMOS MEMS biosensors are miniaturized biosensors fabricated on CMOS (complementary metal-oxide semiconductor) chips by the MEMS (microelectromechanical systems) technology.

Overview

A biosensor is a device designed to detect a biochemical molecule such as a particular deoxyribonucleic acid (DNA) sequence or particular protein. Many biosensors are affinity-based, meaning that they use an immobilized capture probe that selectively binds to the target molecule being sensed. Most biosensors require a label attached to the target, and the amount of detected label is assumed to correspond to the number of bound targets. Labels can be fluorophores, magnetic beads, gold nanoparticles, enzymes, or anything else allowing convenient binding and detection; however, labeling a biomolecule can change the associated binding properties, especially for protein targets.

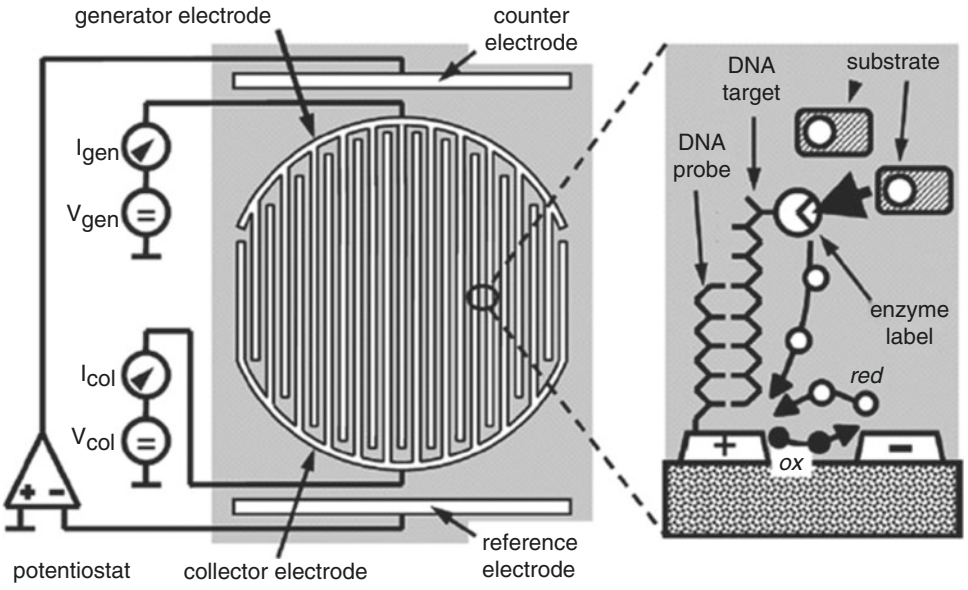
An electrical biosensor is capable of detecting a binding event by producing an electrical current and/or a voltage. Conventional optical detection methods require external instruments that are expensive and not amenable to miniaturization. Electrical bioassays hold great promise for numerous decentralized clinical applications ranging from emergency-room screening to point-of-care diagnostics due to their low cost, high sensitivity, specificity, speed, and portability. Miniaturization of an electrical biosensor can be achieved through microfabrication – widely known as the MEMS technology; in addition, the sensing circuits can be embedded on the same chip through integrated-circuit (IC) processes, among which the CMOS technology is the most popular choice for implementing various analog and digital circuits. A fully integrated CMOS MEMS biosensor array is capable of providing real-time high-throughput detection of multiple samples. CMOS biosensors can be implemented based on the electrochemical, impedimetric, ion-sensitive, magnetic, optical, and micromechanical approaches, which require different MEMS processes to construct the sensing interfaces. Some of the methods require labeling and some are label-free for bio-signal transduction. More details are provided in the following sections.

Sensing Principles and Key Research Findings

Electrochemical Biosensors

Electrochemical biosensors have been the subject of basic as well as applied research for many years. In 1970, Dr. Leland C. Clark demonstrated that glucose could be measured in whole blood with the presence of the glucose oxidase enzyme. Commercial glucose sensors based on electrochemical detection have been developed since then. Electrochemical biosensors typically depend on the presence of a suitable enzyme in the biorecognition layer to catalyze reaction of electroactive substances. Affinity-based electrochemical sensors use enzymes as labels that bind to antibodies, antigens, or oligonucleotides with a specific sequence.

Electrochemical biosensors can employ potentiometric, amperometric, and impedimetric sensing principles to convert the chemical information into a measurable electrical signal. In potentiometric devices, ion selective electrodes (ISE) are commonly used to transduce a biorecognition event into a potential signal between the working and the reference electrodes whose value, depending on the concentration of the analyte,



CMOS MEMS Biosensors, Fig. 1 Schematic of the CMOS interdigitated microelectrode for electrochemical DNA detection. On the *right*: schematic illustration of the redox-cycling process (Schienle et al. © 2004 IEEE)

can be predicted by the Nernst equation. The current flowing through the electrode is equal to or near zero.

Amperometric biosensors operate by applying a constant potential and monitoring the current associated with the reduction or oxidation of an electroactive species. The sensors can work in three- or four-electrode configurations. The former case consists of a reference, a working, and a counter electrode. The four-electrode setup has an additional working electrode such that oxidation and reduction take place at anode and cathode simultaneously. Working electrodes are normally made of noble metals which are critical to the operation of a sensor. Since these metals are not CMOS compatible, the electrodes must be formed after fabrication of CMOS circuits. Silver/silver chloride (Ag/AgCl) is commonly used as the reference electrode in many electrochemical biosensors. Since not all particles oxidized at the anode reach the cathode, a potentiostat, whose input and output are connected to a reference and to a counter electrode, respectively, is required to provide the difference current to the electrolyte and regulate the potential of the electrolyte to a constant value.

Interdigitated microelectrodes have been adopted in many electrochemical biosensors for quantitative analysis. The width and spacing of microelectrodes are reduced by microfabrication. The main advantage is the enhanced redox current due to fast redox recycling, as the chemical products produced at one side of the electrodes are readily collected at the other side of adjacent electrodes and regenerated to the original states. The relationship between the produced redox current and the analyte concentration has been derived by Aoki et al. [1].

Electrochemical DNA hybridization biosensors rely on the conversion of the DNA base-pair recognition event into an electrical signal. Schienle et al. [2] reported a CMOS electrochemical DNA sensor array with each sensing element consisting of interdigitated gold electrodes separated by 1 μm. As depicted in Fig. 1, probe molecules were immobilized on the gold surface through thiol coupling and the target molecules were tagged by an enzyme label (alkaline phosphatase). A chemical substrate (para-aminophenyl phosphate) was applied to the chip after hybridization. The enzyme label on the matched DNA strands cleaved the phosphate group and

generated an electroactive compound (para-aminophenol), which was subsequently oxidized and reduced as the indicator of successful DNA hybridization. Levine et al. [3] also reported a CMOS electrochemical DNA sensor array which was operated based on conventional cyclic voltammetry (CV). During the measurement the electrode potential was scanned up and down in order to produce the redox reactions associated with the ferrocene labels attached to DNA strands. In addition to DNA detection, CMOS electrochemical sensors have been realized to allow high-throughput detection of dopamine and catecholamine release from adrenal chromaffin cells [4], since the release of neurotransmitters from secretory vesicles of biological cells is closely related to the function of a nervous system.

Impedimetric Biosensors

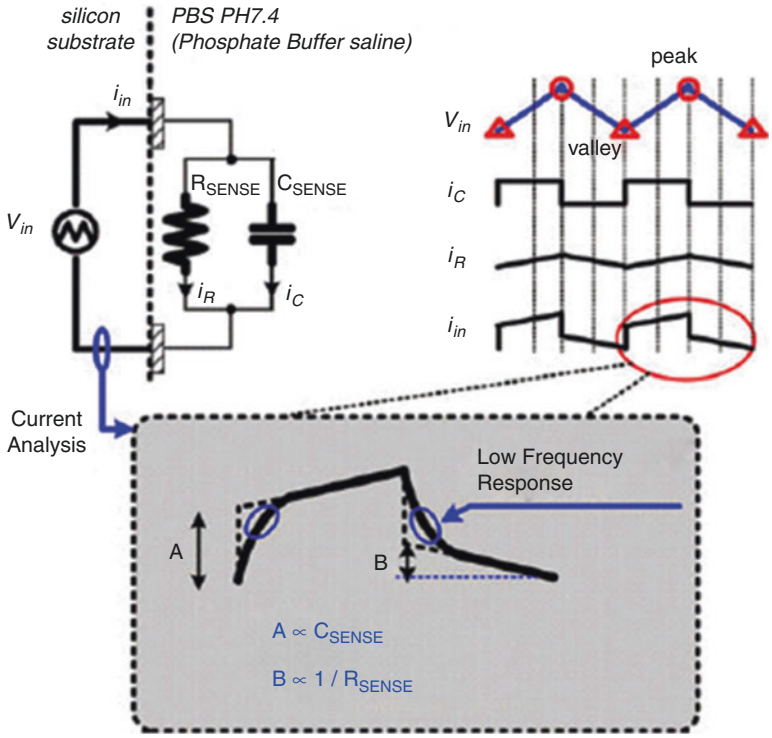
Changes in the electrical properties of a sensing interface (e.g., capacitance, resistance) can occur when a target biomolecule interacts with a probe-functionalized surface. A conventional impedimetric biosensor measures the electrical impedance of an electrode-solution interface in a.c. steady state with constant d.c. bias conditions. This approach, known as electrochemical impedance spectroscopy (EIS), is accomplished by imposing a small sinusoidal voltage over a range of frequencies and measuring the resulting current. The current–voltage ratio gives the impedance, which consists of both energy dissipation (resistor) and energy storage (capacitor) elements. Results obtained by EIS are often graphically represented by a Bode plot or a Nyquist plot. EIS reveals information about the reaction mechanism of an electrochemical process since different reaction steps can dominate at certain frequencies. Impedimetric biosensors can detect a variety of target analytes by simply varying the probe used. Changes in the impedance can be correlated to DNA hybridization, antigen–antibody reaction, or be used to detect biological cells.

The interface impedance is commonly represented by an equivalent circuit model for analysis. E. Warburg [5] first proposed that the interface impedance can be represented by a polarization resistance in series with a polarization capacitor. The interface capacitance possesses a frequency dependency and is commonly represented by the Gouy-Chapman-Stern model as the series combination of the double-layer capacitance (Helmholtz capacitance) and the diffuse layer capacitance (Gouy-Chapman capacitance).

In addition to the frequency-domain measuring method, interface impedance changes can be measured by the potentiostatic step method where small potential steps are applied to the working electrode and the transient current responses, as determined by the time constant of the interface resistance and capacitance, are measured accordingly. Lee et al. [6] reported a fully integrated CMOS impedimetric sensor array for label-free detection of DNA hybridization. The changes in the reactive capacitance and the charge-transfer resistance on the gold sensing electrodes were extracted by applying a triangular voltage waveform and monitoring the produced currents. As illustrated in Fig. 2, the current flowing through the capacitor is associated with the slope of the applied triangular wave, while the current flowing through the resistor is in proportion to the magnitude of the triangular wave. The reported detection limit was 10 nanomolar (nM).

It is important to distinguish the differences between faradaic and non-faradaic biosensors for impedance detection. In electrochemical terminology, a faradaic process involves charge transfer across an interface, while a transient current can flow without charge transfer in a non-faradaic process by charging a capacitor. The faradaic EIS requires the addition of a redox species which is alternately oxidized and reduced by the transfer of charges to and from the metal electrode. In contrast, no additional reagent is required for non-faradaic impedance spectroscopy. The associated impedance change is predominantly capacitive with the charge transfer resistance being omitted.

Miniaturized CMOS capacitive sensors have been developed for numerous biosensing applications. As the sensor size is small, monolithic integration provides the benefit of enhancing the signal-to-noise ratio by reducing the parasitic capacitance observed at the sensing node, which



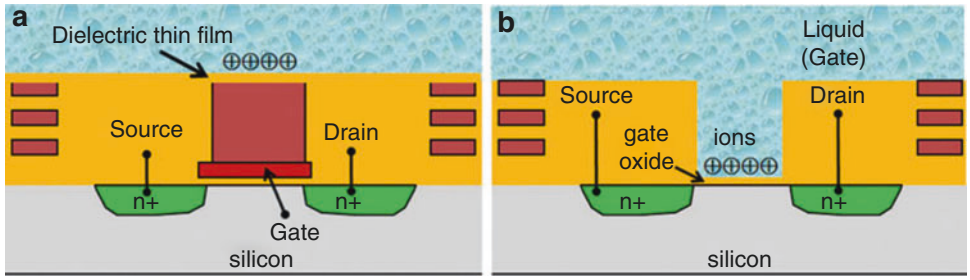
CMOS MEMS Biosensors, Fig. 2 Schematic of the impedance extraction method (Adopted by Lee et al. © 2010 Elsevier)

would otherwise negatively impact the detection limit during direct capacitance measurement. Stagni et al. [7] reported an 8 × 16 CMOS DNA sensor array where the bindings of complementary DNA strands on gold microelectrodes reduced the dielectric constant of the electrode-analyte impedance and the associated change was used to modify the charging and discharging transients of the detection circuit. Capacitive sensitivity can be enhanced by use of interdigitated microelectrodes with the minimum gap defined by the adopted CMOS process. Lu et al. [8] reported very sensitive detection of the neurotransmitter dopamine in the sub-femtomolar (fM) range. CMOS capacitive sensors have also been used for monitoring cellular activity since the morphological and physiological states of biological cells have correlations with their electrical properties [9].

The impedance change of a biosensing event can be made purely resistive with additional chemical modifications. Li et al. [10] reported a DNA array detection method in which the binding of immobilized DNA probes functionalized with gold nanoparticles produced a conductivity change between adjacent metal electrodes, which were made of gold metal in order to withstand the chemical in the clean processing. The detected signal can be further enhanced by additional silver deposition. Silicon dioxide was used as the underlying material under the gap for immobilization of the DNA probes. Detection limit of the target DNA concentration was 1 picomolar (pM).

Ion-Sensitive Field Effect Transistors (ISFET)

ISFETs were first developed in the early 1970s and have been utilized in various biosensing applications which depend on the type of receptor used for analyte recognition or how a signal is generated. Immunologically modified FETs and DNA-modified FETs detect the change of surface charges by monitoring the associated current–voltage relationship. Cell-based FETs detect the potential changes produced by biological cells due to the flow of ions across the cell membranes upon stimulations. For applications where it is required to build an ISFET array to provide detection of multiple samples at different locations,



CMOS MEMS Biosensors, Fig. 3 (**a**) Schematic of the floating-gate ISFET structure. (**b**) Schematic of the open-gate ISFET structure

monolithic integration is then preferred in order to reduce wiring complexity and noise interference.

CMOS ISFET sensors can be built by using a floating-gate or an open-gate structure as depicted in Fig. 3. The floating-gate structure is easier to fabricate as it requires minimal post-processing after completion of a conventional CMOS process. Silicon dioxide or silicon nitride in the CMOS passivation thin films can be used as the material for surface functionalization. The produced signal due to charges on sensor surface is capacitively coupled to the floating gate through a relatively thick dielectric layer which reduces the signal-coupling efficiency. The gate in an open-gate ISFET is removed and replaced by the aqueous solution whose potential is commonly set via a reference electrode. The exposed gate oxide is used for surface functionalization. Since the gate oxide thickness is only tens of angstroms in a conventional sub-μm CMOS process, sensitivity is thus significantly enhanced as compared to a floating-gate ISFET.

Accumulated charges on the ISFET surface modulate the threshold voltage of a MOS transistor, leading to a channel current change under fixed voltage biases of the drain, source, and gate terminals. The threshold voltage of an open-gate ISFET is expressed as:

$$V_{th} = E_{ref} - \Psi + \chi_{sol} - \frac{\phi_{si}}{q} - \frac{Q_{ox} + Q_{ss} + Q_B}{C_{ox}} + 2\phi_F$$

where E_{ref} is the reference electrode potential, Ψ is a chemical input parameter as a function of solution pH value, χ_{sol} is the surface dipole potential of the solvent, ϕsi is the work function of silicon, q is the electron charge, C_{ox} is the gate oxide capacitance per unit area, Q_{ox} and Q_{ss} are the charges in the oxide and at the oxide-silicon interface, Q_B is the depletion charges in silicon and ϕ_F is the difference between the Fermi potential of the substrate and intrinsic silicon.

Kim et al. [11] reported the use of p-type ISFET fabricated in a standard CMOS process for detection of DNA immobilization and hybridization. Gold was deposited by post-processing as the gate material for immobilizing DNA due to its chemical affinity with thiols. The channel current increased during hybridization due to the negative charges present in the phosphate groups of DNA strands. Li et al. [12] presented a post-CMOS fabrication method to make open-gate ISFETs and demonstrated ultrasensitive dopamine detection in the fM range.

Magnetic Biosensors

Some of the magnetic biosensors require the use of magnetic beads as labels attached to the samples in order to induce a measurable electrical signal when specific binding on sensor surface occurs. Magnetic beads are made of small ferromagnetic or ferrimagnetic nanoparticles that exhibit a unique quality referred to as superparamagnetism in the presence of an externally applied magnetic field. This phenomenon, as discovered by Louis Néel (Nobel Physics Prize winner in 1970), has been used in numerous applications such as magnetic data storage and magnetic resonance imaging (MRI).

Several methods can be used to electronically detect the existence of magnetic beads, such as the GMR (giant magnetoresistance) effect discovered by the 2007 Nobel Physics Prize winners Albert Fert and Peter Grünberg. The effect appears in thin film structures composed of alternating ferromagnetic and nonmagnetic layers. A significant change in the electrical resistance is observed depending on whether the magnetization of adjacent ferromagnetic layers is in a parallel alignment (the low-resistance state) due to the applied magnetic field or an anti-parallel alignment (the high-resistance state) in the absence of the magnetic field.

Han et al. [13] reported a CMOS DNA sensor array that adopted the spin valve structure to observe the GMR effect. Magnetic thin films of nanometers in thickness were deposited and patterned after the conventional CMOS process. The biotinylated analyte DNA was captured by complementary probes immobilized on the sensor surface. Then streptavidin-coated magnetic labels were added and produced specific binding to the hybridized DNA. The stray magnetic field of the magnetic labels was detected as a resistance change in the sensor. In general, signal modulation is required in the sensing scheme in order to separate the true bio-signals from the false ones caused by drifts or ionic solution interference. Other than the GMR principle, detection of magnetic beads can be achieved based on the Hall effect of a CMOS sensor. As discovered by Edwin Hall in 1879, the effect produces an electric field perpendicular to the magnetic induction vector and the original current direction. Detection of a single magnetic bead has been demonstrated [14].

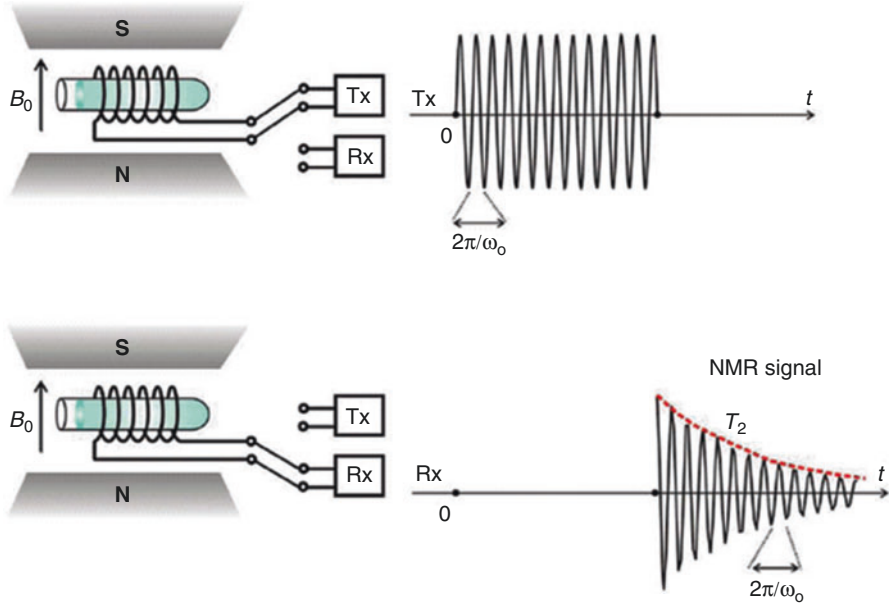
To eliminate the needs of externally applied magnetic fields and post-CMOS fabrication for a magnetoresistive biosensor, Wang et al. [15] presented an inductive approach that can detect existence of a single magnetic bead on a CMOS chip. The sensing scheme used a highly stable integrated oscillator with an on-chip LC resonator. An a.c. electrical current through the on-chip inductor produced a magnetic field that polarized the magnetic particles present in its vicinity, leading to an increased effective inductance and therefore a reduced oscillation frequency. As the frequency shift due to a single micron-size magnetic bead is typically a few parts per million (ppm) of the resonant frequency, the sensing oscillator needs to have small phase noises at small offset frequencies to achieve a stable frequency behavior.

A miniaturized CMOS NMR (nuclear magnetic resonance) system has been reported by Sun et al. [16] for applications in biomolecular sensing. NMR was first discovered by Isidor Rabi who was awarded Nobel Prize in Physics in 1944. Magnetic nuclei, like ^{1}H and ^{31}P, could absorb radio frequency (RF) energy when placed in a magnetic field of a strength specific to the identity of the nuclei. The nucleus is described as being in resonance when the absorption occurs. Different atomic nuclei within a molecule exhibit different resonant frequencies for the same magnetic field strength. Essential chemical and structural information about a molecule can thus be studied by observing such magnetic resonant frequencies.

The reported CMOS NMR system consists of a magnet (static magnetic field), an RF coil surrounding a sample, and an RF transceiver linked to the coil as shown in Fig. 4. The RF magnetic field produced through the coil at the right frequency ω_o can excite nuclei spins within the sample. Once the RF excitation is stopped and the receiver is connected to the coil, the detected NMR signal displays an exponential relaxation in the precession of the net magnetic moment. Both the resonant frequency and the relaxation's characteristic time are specific for the sample to be studied.

Optical Biosensors

Optical detection methods such as luminometry and fluorometry can be utilized to make CMOS-based biosensors. Luminometric methods, such as luciferase-based assays, involve the detection and quantification of light emission as a result of a chemical reaction. Luciferase is a generic term for the class of oxidative enzymes used in bioluminescence. Such methods have been used to detect pathogens and proteins, perform gene expression and regulation studies, and sequence DNA. Fluorometric methods require an excitation source to stimulate photoemission of the fluorescent-tagged species and optical filters to separate the generated photoemission from the



CMOS MEMS Biosensors, Fig. 4 Operation of the CMOS NMR system (Sun et al. © 2009 IEEE)

high background interference. Luminometry is more amenable to miniaturization and integration than the fluorometric methods since no filters or excitation sources are required.

Eltoukhy et al. [17] reported a CMOS photodetector array which was directly integrated with a fiber-optic faceplate with immobilized luminescent reporters/probes for DNA synthesis by pyrosequencing. Pyrosequencing is a "sequencing by synthesis" technique which involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. Synthesis of the complementary strand is achieved with one base pair at a time by monitoring the activity of the DNA synthesizing enzyme with another chemiluminescent enzyme. The challenge of such a detection system is to achieve high sensitivity despite the presence of relatively high dark currents. The reported sensor array was able to detect emission rates below 10^{-6} lux over a long integration time (30 s).

Huang et al. [18] reported a CMOS microarray chip that leveraged low-cost integration of solid-state circuits for fluorescence-based diagnostics. The featured time-gated fluorescence detection was able to significantly reduce interferences from the external excitation source, eliminating the need for additional optical filters. Direct immobilization of DNA capture strands on the CMOS chip allows placement of optical detectors in close vicinity to the fluorescent labels, leading to improved collection efficiency and providing imaging resolution limited by pixel dimensions rather than diffraction optics.

In addition to the aforementioned approaches, a method based on detection of the incident light intensity using a normal light source has been developed [19]. The technique uses gold nanoparticles with silver enhancement to induce opacity on top of CMOS photodetectors when specific bindings occur. It does not require an external optical scanner and a specific light source as needed in the fluorescence-based method.

Micromechanical Cantilevers

Microcantilever-based biosensors have attracted considerable attention as a means of label-free detection of biomolecules. Intermolecular forces arising from adsorption of small molecules are known to induce surface stress. The specific binding between ligands and receptors on the surface of a microcantilever beam thus produces physical

bending of the beam. Optical detection of the beam deflection due to hybridization of complementary oligonucleotides has been demonstrated [20]; however, the method requires external instruments that are not amenable to monolithic integration.

Piezoresistive detection is a viable alternative for CMOS integration because it is compatible with aqueous media. The piezoresistive effect is associated with the resistivity change of a semiconductor subject to a mechanical strain. By placing a MOS transistor into the base of a cantilever, modulation of the channel current underneath the gate region can be measured as a result of adsorption-induced surface stress [21]. Sensing resolution of the static beam deflection is limited by the flicker noise – the dominant source of noise in MOS transistors at low frequencies.

Summary

A miniaturized biosensing platform can be achieved through monolithic integration of sensing devices and detection circuits by the CMOS MEMS technology. Sensing devices that can be directly fabricated on a CMOS chip are introduced, including those based on electrochemical, impedimetric, ion-sensitive, magnetic, optical, and micromechanical approaches. Some of the methods require labeling (e.g., magnetic beads, nanoparticles) and some are label-free for bio-signal transduction. Some approaches use the devices (e.g., transistors) or the materials (e.g., metal electrodes) in a CMOS process for sensing, such that sensor performances can be enhanced in accordance with the scaling of CMOS technologies. Arrays of sensors can be conveniently fabricated on a CMOS chip such that sensing resolution and accuracy can be enhanced through statistical analysis of the collected data.

Cross-References

- ▶ Biosensors
- ▶ Nanogap Biosensors

References

1. Aoki, K., Morita, M., Niwa, O., Tabei, H.: Quantitative analysis of reversible diffusion-controlled currents of redox soluable species at interdigitated array electrodes under steady-state conditions. J. Electroanal. Chem. **256**, 269–282 (1988)
2. Schienle, M., Paulus, C., Frey, A., Hofmann, F., Holzapfl, B., Schindler-Bauer, P., Thewes, R.: A fully electronic DNA sensor with 128 positions and in-pixel A/D conversion. IEEE J. Solid State Circuits **39**, 2438–2445 (2004)
3. Levine, P.M., Gong, P., Levicky, R., Shepard, K.L.: Active CMOS sensor array for electrochemical biomolecular detection. IEEE J. Solid State Circuits **43**, 1859–1871 (2008)
4. Ayers, S., Berberian, K., Gillis, K.D., Lindau, M., Minch, B.A.: Post-CMOS fabrication of working electrodes for on-chip recordings of transmitter release. IEEE Trans. Biomed. Circuits Syst. **4**, 86–92 (2010)
5. Warburg, E.: Ueber das Verhalten sogenannter unpolarisbarer Elektroden gegen Wechselstrom. Ann. Phys. Chem. **67**, 493–499 (1899)
6. Lee, K., Lee, J., Sohn, M., Lee, B., Choi, S., Kim, S. K., Yoon, J., Cho, G.: One-chip electronic detection of DNA hybridization using precision impedance-based CMOS array sensor. Biosens. Bioelectron. **26**, 1373–1379 (2010)
7. Stagni, C., Guiducci, C., Benini, L., Riccò, B., Carrara, S., Samorí, B., Paulus, C., Schienle, M., Augustyniak, M., Thewes, R.: CMOS DNA sensor array with integrated A/D conversion based on label-free capacitance measurement. IEEE J. Solid State Circuits **41**, 2956–2963 (2006)
8. Lu, M.S.-C., Chen, Y.C., Huang, P.C.: 5 × 5 CMOS capacitive sensor array for detection of the neurotransmitter dopamine. Biosens. Bioelectron. **26**, 1093–1097 (2010)
9. Ghafar-Zadeh, E., Sawan, M., Chodavarapu, V.P., Hosseini-Nia, T.: Bacteria growth monitoring through a differential CMOS capacitive sensor. IEEE Trans. Biomed. Circuits Syst. **4**, 232–238 (2010)
10. Li, J., Xue, M., Lu, Z., Zhang, Z., Feng, C., Chan, M.: A high-density conduction-based micro-DNA identification array fabricated with a CMOS compatible process. IEEE Trans. Electron Devices **50**, 2165–2170 (2003)
11. Kim, D.S., Jeong, Y.T., Park, H.J., Shin, J.K., Choi, P., Lee, J.H., Lim, G.: An FET-type charge sensor for highly sensitive detection of DNA sequence. Biosens. Bioelectron. **20**, 69–74 (2004)
12. Li, D.C., Yang, P.H., Lu, M.S.-C.: CMOS open-gate ion-sensitive field-effect transistors for ultrasensitive dopamine detection. IEEE Trans. Electron Devices **57**, 2761–2767 (2010)
13. Han, S., Yu, H., Murmann, B., Pourmand, N., Wang, S.X.: A high-density magnetoresistive biosensor array with drift-compensation mechanism. In: IEEE International Solid-State Circuits Conference (ISSCC)

Digest of Technical Papers, pp. 168–169, San Francisco, 11–15 Feb 2007
14. Besse, P., Boero, G., Demierre, M., Pott, V., Popovic, R.: Detection of a single magnetic microbead using a miniaturized silicon Hall sensor. Appl. Phys. Lett. **80**, 4199–4201 (2002)
15. Wang, H., Chen, Y., Hassibi, A., Scherer, A., Hajimiri, A.: A frequency-shift CMOS magnetic biosensor array with single-bead sensitivity and no external magnet. In: IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, pp. 438–439 (2009)
16. Sun, N., Liu, Y., Lee, H., Weissleder, R., Ham, D.: CMOS RF biosensor utilizing nuclear magnetic resonance. IEEE J. Solid State Circuits **44**, 1629–1643 (2009)
17. Eltoukhy, H., Salama, K., El Gamal, A.: A 0.18-μm CMOS bioluminescence detection lab-on-chip. IEEE J. Solid State Circuits **41**, 651–662 (2006)
18. Huang, T.D., Sorgenfrei, S., Gong, P., Levicky, R., Shepard, K.L.: A 0.18-μm CMOS array sensor for integrated time-resolved fluorescence detection. IEEE J. Solid State Circuits **44**, 1644–1654 (2009)
19. Xu, C., Li, J., Wang, Y., Cheng, L., Lu, Z., Chan, M.: A CMOS-compatible DNA microarray using optical detection together with a highly sensitive nanometallic particle protocol. IEEE Electron Device Lett. **26**, 240–242 (2005)
20. Fritz, J., Baller, M.K., Lang, H.P., Rothuizen, H., Vettiger, P., Meyer, E., Güntherodt, H.-J., Gerber, C., Gimzewski, J.K.: Translating biomolecular recognition into nanomechanics. Science **288**, 316–318 (2000)
21. Shekhawat, G., Tark, S.H., Dravid, V.P.: MOSFET-embedded microcantilevers for measuring deflection in biomolecular sensors. Science **311**, 1592–1595 (2006)

CMOS MEMS Fabrication Technologies

Hongwei Qu[1] and Huikai Xie[2]
[1]Department of Electrical and Computer Engineering, Oakland University, Rochester, MI, USA
[2]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

Synonyms

CMOS (complementary metal-oxide-semiconductor); CMOS-MEMS; Integration; MEMS (micro-electro-mechanical systems)

Definition

CMOS-MEMS are micromachined systems in which MEMS devices are integrated with CMOS circuitry on a single chip to enable miniaturization and performance improvement. CMOS-MEMS also refers to microfabrication technologies that are compatible with CMOS fabrication processes.

Overview

Microelectromechanical systems (MEMS) leverage semiconductor fabrication technologies to manufacture various miniature sensors and actuators. Due to their low cost and small size as well as their much improved reliability, MEMS devices have been widely used even in our daily life, e.g., MEMS accelerometers for automobiles' airbags, MEMS gyroscopes for electronic stability program (ESP) in automobile braking systems, MEMS tire pressure sensors; digital micromirror device (DMD)-enabled portable projectors, MEMS inkjet printers, MEMS resonators as frequency references, etc. Moreover, smart cell phones are now equipped with MEMS gyroscopes and accelerometers for motion actuated functions. They are also installed with surface-mounted MEMS microphones for even smaller size. The worldwide MEMS market reached 6.5 billion US dollars in 2010 [1].

Continuous miniaturization, expanded functionalities, lower cost, and improved performance are the ultimate goals of MEMS. The nature of MEMS strongly suggests direct integration of mechanical structures with electronics whose fabrication is dominated by CMOS technologies. In the last couple of decades, great efforts have been made in the integration of MEMS structures with ICs on a single CMOS substrate. The pioneering work for CMOS-MEMS transducers was done by H. Baltes and his coworkers at the Swiss Federal Institute of Technology Zurich (ETH) [2]. They employed both wet bulk silicon micromachining and surface micromachining techniques in the fabrication of integrated CMOS-MEMS devices. With the great advances in IC and MEMS

technologies, the current focus of CMOS-MEMS integration technology is on the modification and standardization of CMOS technology to accommodate MEMS technology. One of the best-known commercial CMOS-MEMS devices is the digital micromirror device (DMD) manufactured by Texas Instruments. Recently, some CMOS foundries, such as TSMC, X-Fab and Global Foundries, have begun to offer CMOS-MEMS services for research and product developments.

This entry summarizes a variety of CMOS-MEMS technologies and devices that have been developed. Particular materials needed in associated CMOS-MEMS will also be introduced. Typical MEMS devices, including inertial sensors, resonators, and actuators, are exemplified in featuring the respective technologies.

Classification of CMOS-MEMS Technologies

MEMS can be integrated with CMOS electronics in many different ways. One common way to categorize CMOS-MEMS technologies is from the perspective of manufacturing processes. Based on the process sequence, CMOS-MEMS technologies can be classified into three categories: Pre-CMOS, Intra-CMOS, and Post-CMOS [3]. Due to its popularity and accessibility, post-CMOS will be described in more detail in this entry.

Pre-CMOS

It is widely accepted that pre-CMOS technologies are represented by the modular integration process originally developed at Sandia National Laboratories. As suggested by the name, in pre-CMOS technology, MEMS structures are pre-defined and embedded in a recess in silicon wafer; and the recess is then filled with oxide or other dielectrics. The wafer is then planarized prior to the following process steps for CMOS electronics [4]. In this "MEMS first" process, although MEMS structures are pre-defined, a wet etch after the completion of the standard CMOS processes is required to release the pre-defined MEMS structures. Due to the involvement of photolithography process needed for patterning the MEMS in the recess, the thickness of the MEMS structures is constrained by the lithographical limit.

Other methods for the formation of MEMS structures in pre-CMOS technologies, including wafer bonding and thinning for epitaxial and SOI wafers in which MEMS are prefabricated, have also been reported in the fabrication of a variety of MEMS devices.

Inter-CMOS

In early 1990s, Analog Devices, Inc (ADI) specifically developed a MEMS technology based on its BiCMOS process. This "iMEMS" technology, originally dedicated to manufacturing CMOS-MEMS accelerometers and gyroscopes, is an intermediate-CMOS-MEMS, or Inter-CMOS-MEMS technology in which the CMOS process steps are mixed with additional polysilicon thin-film deposition and micromachining steps to form the sensor structures [5]. Infinion's pressure sensors are also fabricated using this kind of Inter-CMOS-MEMS technology. To reduce the residual stress in structural polysilicon, high temperature annealing is normally required in the Inter-CMOS-MEMS, which could pose a potential risk to CMOS interconnect and active layers. Thus, the thermal budget should be carefully designed. Moreover, it is almost impractical to perform intermediate CMOS and MEMS processed in separate foundries due to possible contaminations in the wafer transfer and processes. Therefore, a dedicated foundry used for both CMOS and MEMS is necessary for Inter-CMOS-MEMS technology, which may not take full advantages of mainstream technologies in either area.

Limitations of Pre- and Inter-CMOS-MEMS

Since surface micromachining and polysilicon are typically used, most of the Pre- and

Inter-CMOS-MEMS technologies suffer from the limitations of thin-film structural materials. (1) Structural curling and cost associated with stress compensation: Due to the residual stress in the deposited thin films in the device, polysilicon structures often exhibit curling after release, resulting in reduced sensitivity, lower mechanical robustness, and increased temperature dependence. Although stress compensation can be realized via multiple controlled process steps, the associated cost is quite high. (2) Small size and/or mass: The curling of thin-film structures in turn limits the size of the overall microstructure. The mass is further reduced due to the small thickness of thin-film polysilicon. For inertial sensors, the smaller the structure mass, the lower performance of the device. (3) Parasitics: In a surface micromachined polysilicon accelerometer, depending on polysilicon wiring path, the parasitic impedance may considerably lower the static and dynamic performance of the device. (4) Cost and suboptimal processes of the dedicated foundry needed: The dedicated foundry combining CMOS and MEMS fabrications needed for pre- and inter-CMOS-MEMS is normally expensive and suboptimal for either fabrication. It is against the modern trends in which flexible accessibility of optimal and cost-effective processes are preferable.

Post-CMOS

Post-CMOS-MEMS refer to the CMOS-MEMS processes in which all MEMS process steps are performed after the completion of the CMOS fabrication. The advantages of post-CMOS-MEMS over Pre- and Inter-CMOS-MEMS include process flexibility and accessibility and low cost. In contrast to both Pre-CMOS-MEMS and Inter-CMOS-MEMS, for Post-CMOS-MEMS technology, the fabrications of CMOS circuitry and MEMS structures are performed independently. The flexibility of foundry access makes it possible to take advantages of both advanced CMOS technologies and optimal MEMS fabrication. This is particularly attractive to research community in exploration of state-of-the-art in MEMS. Some design rules may need to be changed to accommodate MEMS structure design in the CMOS design stage. Meanwhile, post-CMOS microfabrication should be carefully designed, particularly considering the thermal budget, so as not to affect the on-chip CMOS electronics.

According to how MEMS structures are formed, post-CMOS-MEMS technologies fall into two categories: additive and subtractive. In additive post-CMOS-MEMS, structural materials are deposited on a CMOS substrate. In subtractive post-CMOS-MEMS, MEMS structures are created by selectively etching CMOS layers. Apparently, additive post-CMOS-MEMS methods require more stringent material compatibility with the CMOS technologies used. Thus, they are less utilized than subtractive post-CMOS-MEMS. The following introduction will focus more on subtractive post-CMOS-MEMS.

Additive MEMS Structures on CMOS Substrate

In additive post-CMOS-MEMS, metals, dielectrics, or polymers are deposited and patterned to form MEMS structures normally on top of the CMOS layers. Some commercial MEMS products are fabricated using additive post-CMOS-MEMS approaches. In this category, the best-known product is probably the digital mirror device (DMD), the core of the digital light processing (DLP) technology developed by Texas Instruments. In a DMD, tilting mirror plates and their driving electrodes are fabricated directly on top of CMOS circuits. Three sputtered aluminum layers are used to form the top mirror plate and the two parallel-plate electrodes for electrostatic actuation, respectively. The driving electrodes are addressed via CMOS memory cell. To release the mirror plate and top electrodes in the post-CMOS-MEMS fabrication of the mirrors, deep-UV hardened photoresist is used as the sacrificial layer.

In some circumstances where CMOS protection is well designed, electroplating can also be used to grow microstructures on top of CMOS electronics. Other structural and sacrificial materials, such as polycrystalline SiGe and Ge, have

CMOS MEMS Fabrication Technologies, Table 1 Representative thin-film deposition additive post-CMOS-MEMS technologies

Authors and references	Institute	Structural material	Sacrificial material	Interconnect material	Year
Hornbeck [7]	Texas instruments	Al	Photoresist	Al	1989 (invented in 1987)
Yun et al. [8]	UC-Berkeley	Polysilicon	SiO_2	W/TiN	1992
Franke et al. [9]	UC-Berkeley	Poly-SiGe	Ge or SiO_2	Al	1999
Sedky, Van Hoof et al. [10]	IMEC	Poly-SiGe	Ge	Al	1998

been used to create CMOS-MEMS as well [6]. Additive post-CMOS-MEMS processes, along with their respective materials, are summarized in the following table (Table 1).

In addition to the approaches of forming MEMS structures on top of the CMOS substrate by thin-film deposition, wafer bonding provides another method to directly integrate MEMS structures on CMOS substrate [11]. For instance, a prefabricated polysilicon capacitive acceleration sensor wafer is bonded to a CMOS wafer with read-out electronics. In a wafer-bonded piezoresistive accelerometer, the micromachined bulk silicon proof mass was sandwiched by a bottom glass cap and a top CMOS chip on which the conditioning circuit was integrated. SOI-CMOS-MEMS has also been attempted for monolithic integration of electronics with bulk MEMS structures. With 3-dimensional packaging enabled by technological breakthroughs such as through-silicon vias (TSVs), this integration method promises to be further developed in manufacturing complex microsystems. MEMS suppliers, including STMicroelectronics and InvenSense, have adopted wafer-to-wafer or chip-to-wafer bonding CMOS-MEMS integration.

Subtractive Post-CMOS-MEMS

In these devices, MEMS structures are formed from built-in CMOS thin-film stacks including metals and SiO_2, or from the silicon substrate. These materials are patterned and removed partially by wet or dry etching to release the MEMS structures. This section describes the thin-film and bulk CMOS-MEMS formed by such subtractive processes.

Subtractive CMOS-MEMS by Wet Etching

The first generation of CMOS-MEMS sensors was fabricated using a post-CMOS subtractive process in which silicon substrate was completely or partially removed using a wet etching method, leaving behind thin-film or bulk MEMS structures [2]. For thermal sensors in which beams or membranes consisting of dielectric layers, the substrate silicon is normally etched away completely to obtain thermally isolated structures. The silicon dioxide membrane can act as an intrinsic etch stop layer in backside silicon anisotropic wet etch using KOH, ethylene diamine-pyrocatechol (EDP), or Tetramethylammonium hydroxide (TMAH). A high-Q RF MEMS filter with an inter-metal dielectric layer as structural material was reported by IBM. A medical tactile sensor array was also reported in which the aluminum sacrificial layer was etched from the backside of the wafer after the CMOS substrate was etched through [12].

The silicon substrate can also be included in the MEMS structures using a wet etch process. The first method is to perform a time-controlled backside etch with a well-calibrated etching rate. A uniform single crystal silicon membrane with a desired thickness can be created. This method has been widely used in industry for fabrication of large volume products such as integrated pressure sensors. In cases where the silicon membrane thickness is not critical, even mechanical processing such as grinding can be used to create the backside cavity.

The second method involves the utilization of an automatic etch stop technique to create silicon membranes or MEMS structures. In this case, an anisotropic etch stops at the electrochemically

CMOS MEMS Fabrication Technologies, Table 2 CMOS-MEMS devices enabled by subtractive process wet etching

Authors and references	Institutions	Year	Device	Device structure	Etching method
Wise et al. [16]	U. of Michigan	1979	Pressure sensor	Silicon diaphragm	Backside EDP etching
Wise et al. [14]	U. of Michigan	1985	Neuron probe array	CMOS thin films and Si substrate	EDP etching, p++ etching stop
Yoon and Wise [17]	U. of Michigan	1990	Mass flow sensor	CMOS thin films	Backside, SiO_2 etching stop
Baltes et al. [2]	ETH Zurich	1996	Thermal capacitor	CMOS thin films	Front side etching
Haberli et al. [18]	ETH Zurich	1996	Pressure sensor	CMOS thin films	Front side etching of aluminum as sacrificial layer
Schneider et al. [19]	ETH Zurich	1997	Thermal sensor	CMOS thin films and suspended substrate	PN junction electrochemical etch stop
Akiyama et al. [20]	U. of Neuchatel, ETH Zurich	2000	AFM probe	CMOS thin films and silicon substrate	N-well electrochemical etch stop
Schaufelbuhl et al. [21]	ETH Zurich	2001	Infrared imager	CMOS thin films	Backside KOH
Verd et al. [22]	U. of Barcelona	2006	RF MEMS	CMOS thin films	Front side SiO_2 etching

biased p-n junction formed between the n-well and p-type substrate in CMOS [13]. Although the electrochemical electrode design and implementation are complicated, this process can be specifically used in the fabrication of highly sensitive pressure/force and thermal sensors. The anisotropic etch stop can also occur at highly doped ***p*** regions in the substrate. This method has been used in fabrication of many suspended structures including neural probes [14]. Note that the p++ doping process may not be available in a standard CMOS process. In the case where only a small portion of the silicon substrate needs to be removed to reduce the circuit-substrate coupling, a wet silicon etch can be performed from the front side. In wet silicon etching, either silicon nitride or additional polymers or both can be used to protect the front CMOS and pads.

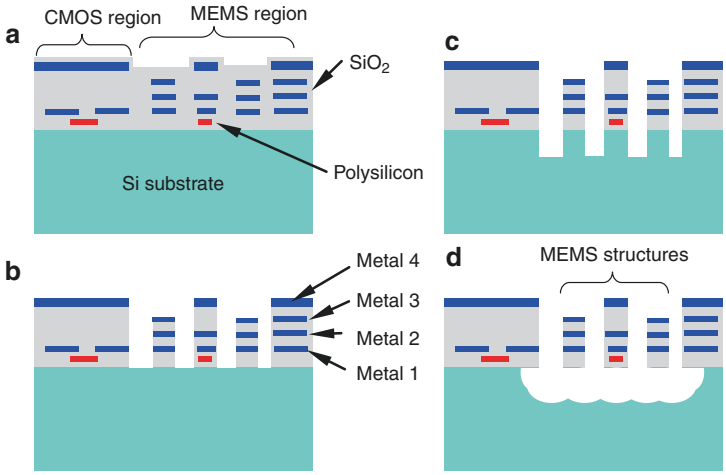
Polymers sensitive to analytes can be coated on finished CMOS-MEMS structures for chemical and biological sensing. For example, the first CMOS-MEMS electronic nose was demonstrated by forming polymer-coated CMOS thin-film cantilevers on a CMOS chip [15].

Table 2 summarizes some representative devices that were fabricated using wet etching when this technology was dominant in post-CMOS micromachining. Bibliographies of these efforts can be found in the above citations in this section.

Subtractive Post-CMOS-MEMS by Dry Etching

Dry etching processes have quickly become popular in microfabrication for both MEMS research and industry. Particularly, the deep reactive ion etching (DRIE) technology, or Bosch process, has revolutionized subtractive post-CMOS microfabrication [23]. This section describes thin-film and bulk CMOS-MEMS devices fabricated using dry etching processes.

Most dry etching processes are based on plasma processes, such as reactive ion etch (RIE) and DRIE, while etchants in the vapor phase can also be used for dry etching. For example, vapor XeF_2 provides good isotropic etching of silicon, which has been used for releasing CMOS thin-film MEMS structures [24]. The combination of RIE and DRIE, performed from the front or back side, or both sides, has been used to fabricate a large variety of CMOS-MEMS devices. Depending on the structural materials and etching methods employed, subtractive post-CMOS can be divided into two types: thin-film processes and bulk processes.

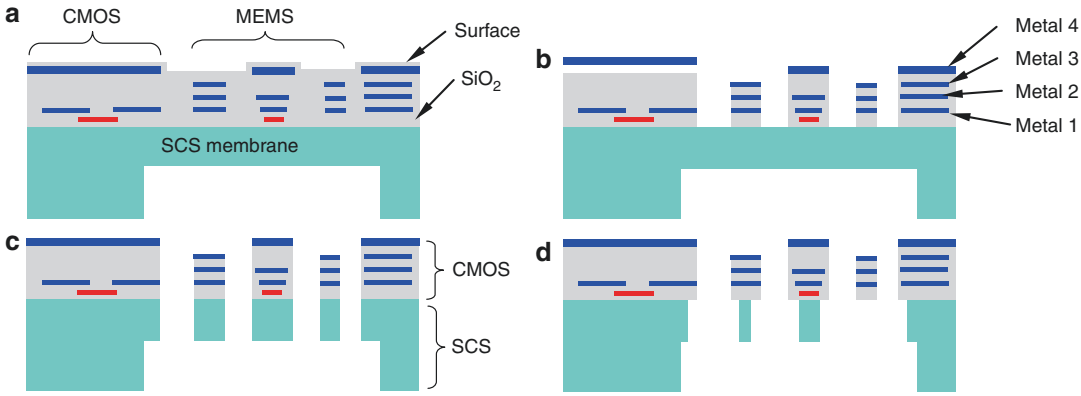


CMOS MEMS Fabrication Technologies,
Fig. 1 CMU post-CMOS fabrication process for MEMS structures made of CMOS thin films

Thin-Film Post-CMOS-MEMS Dry Processes In thin-film processes, structural materials are composed of CMOS thin films. Figure 1 depicts the process flow of a thin-film post-CMOS-MEMS process, which was originally developed at Carnegie Mellon University [25]. A sequenced process consisting of an isotropic SiO_2 etching, a silicon DRIE and an isotropic Si RIE releases the MEMS structure. In these process steps, the top metal layer acts as a mask to form the MEMS structures and to protect the CMOS circuitry, as seen in Fig. 1a, b. Anisotropic and isotropic silicon etching complete the process flow, as seen in Fig. 1c, d. Various inertial sensors have been fabricated using this thin-film technology. In all these inertial sensors, mechanical springs and proof masses are formed by the multiple-layer CMOS stacks consisting of SiO_2 and metals. The sensing capacitance is formed from sidewall capacitance between comb fingers. The multiple CMOS metal layers inside the comb fingers and other mechanical structures allow very flexible electrical wiring, facilitating different sensing schemes including vertical comb-drive sensing. Akustica, Inc. has commercialized digital microphones using a modified version of this process. Other sensors have also been demonstrated using similar thin-film technology, such as humidity sensors and chemical sensors.

All these thin-film post-CMOS dry etching processes have excellent CMOS compatibility and accessibility as well as design flexibility. However, a major issue is the large vertical curling and lateral buckling of suspended MEMS structures, which is caused by the residual stress in the stacked thin-film CMOS layers. Although structural curling can be tolerated for some small devices such as RF MEMS, for devices such as inertial sensors that need relatively large size, the impact of structural curling can be severe.

Bulk CMOS-MEMS Dry Process In order to overcome the structural curling and to increase the mass, flatness, and robustness of MEMS structures, single crystal silicon (SCS) may be included underneath the CMOS thin-film stacks. The SCS silicon structures are formed directly from the silicon substrate using DRIE. Figure 2 illustrates the process flow [26]. The process starts with the backside silicon DRIE to define the MEMS structure thickness by leaving a 10–100 μm-thick SCS membrane (Fig. 2a). Next, the same anisotropic SiO_2 etch as in the thin-film process is performed on the front side of wafer (chip) to expose the SCS to be removed (Fig. 2b). The following step differs from the thin-film process in that an anisotropic DRIE, instead of isotropic etch, finalizes the structure release by etching through the remaining SCS diaphragm, as shown in Fig. 2c. With the SCS underneath the CMOS interconnect layers included, large and flat MEMS microstructures can be obtained. If necessary, an optional time-controlled isotropic silicon etch can be added. This step will undercut the SCS



CMOS MEMS Fabrication Technologies, Fig. 2 DRIE bulk CMOS-MEMS process flow

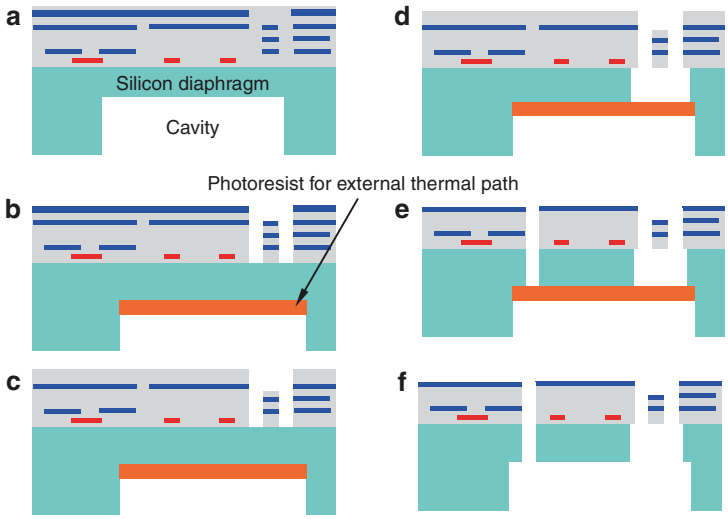
underneath the designed narrow CMOS stacks to create thin-film structures (Fig. 2d). This step is particularly useful for fabricating capacitive inertial sensors. It can be used to form the electrical isolation structures between sensing fingers and silicon substrate.

The DRIE CMOS-MEMS technology has shown great advantages in the fabrication of relatively large MEMS devices such as micromirrors. Large flat mirror can be obtained by including portion of silicon substrate underneath the aluminum mirror surface. A CMOS-MEMS gyroscope with a low noise floor of 0.02°/s/sqrtHZ has also been demonstrated using this technology [27].

By attaching SCS underneath the CMOS stack comb fingers, the sensing capacitance can be considerably increased for larger signal-to-noise ratio (SNR). Although CMOS thin films are still used in some microstructures for electrical isolation, the length of the thin-film portion is minimal to reduce the temperature effect. Compared to the thin-film dry CMOS-MEMS process, a backside silicon DRIE step is added. This requires an additional backside lithography step to define the region for MEMS structures. The maximum thickness of the MEMS structures is limited by the aspect ratio that the silicon DRIE can achieve.

An Improved Bulk CMOS-MEMS Process The bulk CMOS-MEMS process depicted in Fig. 2 is useful in fabrication of many devices where SCS structures are desired to improve both mechanical and electrical performance of the devices. However, for some devices, very fine structures are formed in step (c) in Fig. 2; so the damage caused by the step (d) to these fine structures may be severe. This is particularly true for the fabrication of capacitive inertial sensors where narrow-gap sensing comb fingers are needed. For instance, in performing the isotropic silicon undercut to form the narrow CMOS beams for electrical isolation, the SCS in the comb fingers is also undercut. The sensing gap increases due to the undesired undercut greatly reduce sensitivity and signal-to-noise ratio (SNR). If the undercut occurs in mechanical structures such as suspension springs, the characteristics of the device will also be affected. Another issue is related to the thermal effect in the plasma etch for the SCS undercut. Upon completion of the silicon undercut, the greatly reduced thermal conductance from the isolated structure to the substrate can cause a temperature rise on the released structures. Slight over-etch is often necessary to accommodate process variations, but this will generate a large temperature rise on the suspended structures which in turn dramatically increases the SCS etching rate, resulting in uncontrollable and damaging results [28].

A modified dry bulk CMOS-MEMS process has been demonstrated to effectively address the issues caused by the undesired SCS



CMOS MEMS Fabrication Technologies, Fig. 3 The modified bulk CMOS-MEMS process for separate etching of CMOS beams and SCS microstructures. Backside photoresist coating effectively reduces temperature in the device release, reducing deleterious non-uniform etching

undercut [28]. In the refined process illustrated in Fig. 3, the etching of the CMOS connection beams is performed separately from the etching of the microstructures where SCS is needed. The top metal layer is specifically used to define the connection beams. After their formation, the top metal layer is removed using a plasma or a wet etch. Then other microstructures are exposed after a SiO_2 etch. The direct etch-through of the remaining silicon on the microstructures will complete the release process. To reduce the thermal effect described above, a thick photoresist layer is patterned on the backside of the cavity. In the release step, the applied photoresist provides a thermal path that reduces the temperature rise on the etched-through structures. The removal of the photoresist using O_2 plasma etching completes the entire microfabrication process. Owing to the monolithic integration and large proof mass enabled by the inclusion of SCS, bulk CMOS-MEMS inertial sensors have demonstrated better performance than their thin-film counterparts [29]. The photoresist coating can also be replaced by sputtering a layer of metal such as aluminum.

Combined Wet/Dry Processes

In addition to the integration methods described above, efforts have been continuously made to integrate CMOS with MEMS using the combination of different microfabrication technologies. By combing silicon anisotropic wet etch with DRIE, some sophisticated surface and bulk MEMS structures such as bridges and cantilever arrays can be created. A multi-sensor system was demonstrated using a combined etch process [30]. In the accelerometers reported in [31], isotropic wet etching is used to remove metal layers in CMOS thin stacks to create parallel-plate-like vertical capacitors for gap-closing sensing. A silicon RIE follows to release the MEMS devices and break the coupling between the sensing thin films and the substrate. Sensitivities are largely increased with the gap-closing sensing compared to comb-finger sensing.

Summary

CMOS-MEMS technologies have been placed in pre-CMOS, intra-CMOS, and post-CMOS categories. Both pre-CMOS and intra-CMOS have issues such as dedicated foundries with suboptimal and less cost-effective processes. So it is normally impractical for academic research community to access these dedicated facilities. Post-CMOS provides excellent CMOS compatibility, foundry accessibility, and design flexibility, and the cost is also relatively low. While the

process standardization and industrialization of CMOS-MEMS technologies are in continuous progress, innovative processing technologies have opened up new pathways for integration. Wafer bonding–based integration has blurred the boundary between pre- and post-CMOS-MEMS integrations. SOI-CMOS-MEMS have also been aggressively explored. The technologies involved in this new exploration have emerged as enabling means for three-dimensional and systems-in-package integrations. More recently, technologies to co-fabricate many subsystems including nano-systems are being pursued enthusiastically.

CMOS-MEMS integration will continually evolve with the emergence of new fabrication technologies and new materials. While some companies have demonstrated promising CMOS-MEMS products, more joint efforts from research community and industries are needed for new process transfer and standardization to allow large volume fabrication of new products.

Cross-References

- ▶ Integration
- ▶ MEMS
- ▶ Nanofabrication
- ▶ Sensors

References

1. Johnson, R.C.: MEMS market projected to hit double-digit growth, again. www.eetimes.com. Accessed 14 July 2011
2. Baltes, H., Paul, O., et al.: IC MEMS microtransducers. In: Proceedings of International Electronic Device Meeting IEDM '96, San Francisco, pp. 521–524 (1996)
3. Baltes, H., Brand, O., Fedder, G.K., Hierold, C., Korvink, J.G., Tabata, O.: CMOS-MEMS: Advanced Micro and Nanosystems, 1st edn. Wiley, Weinheim (2005)
4. Smith, J.H., Montague, S., Sniegowski, J.J., Murray, J. R., McWhorter, P.J., Smith J.H.: Embedded micromechanical devices for the monolithic integration of MEMS with CMOS. In: Proceedings of International Electronic Device Meeting, IEDM '95, Washington, DC, pp. 609–612 (1995)
5. Kuehnel, W., Sherman, S.: A surface micromachined silicon accelerometer with on-chip detection circuitry. Sens. Actu. A Phys. **45**, 7–16 (1994)
6. Franke, A.E., Heck, J.M., King, T.J., Howe, R.T.: Polycrystalline silicon-germanium films for integrated microsystems. J. Micromech. Sys. **12**, 160–171 (2003)
7. Hornbeck, L.: Deformable-mirror spatial light modulators and applications. SPIE Crit. Rev. **1150**, 86–102 (1989)
8. Yun, W., Howe, R.T., Gray, P.R.: Surface micromachined, digitally force-balanced accelerometer with integrated CMOS detection circuitry. Technical Digest of Solid State Sensors and Actuators Workshop, Hilton Head Island, pp. 126–131 (1992)
9. Franke, A.E., Bilic, D., Chang, D.T., Jones, P.T., King, T.J., Howe, R.T., Johnson, G.C.: Post-CMOS integration of germanium microstructures. In: The 12th IEEE International Conference on Micro Electro Mechanical Systems, Orlando, pp. 630–637 (1999)
10. Sedky, S., Fiorini, P., Caymax, M., Loreti, S., Baert, K., Hermans, L., Mertens, R.: Structural and mechanical properties of polycrystalline silicon germanium for micromachining applications. J. MEMS **7**, 365–372 (1998)
11. Fedder, G.K., Howe, R.T., Liu, T.J., Quevy, E.P.: Technologies for cofabricating MEMS and electronics. Proc. IEEE **96**, 306–322 (2008)
12. Salo, T., Vancura, T., Brand, O., Baltes, H.: CMOS-based sealed membranes for medical tactile sensor arrays. In: Proceedings of International Conference on Micro Electro Mechanical Systems, Kyoto, pp. 590–593 (2003)
13. Muller, T., Brandl, M., Brand, O., Baltes, H.T.: An industrial CMOS process family adapted for the fabrication of smart silicon sensors. Sen. Actu. A Phys. **84**, 126–133 (2000)
14. Najafi, K., Wise, K.D., Mochizuki, T.K.: A high-yield IC-compatible multichannel recording array. IEEE T. Electron. Dev. **32**, 1206–1211 (1985)
15. Baltes, H., Koll, A., Lange, D.H.: The CMOS MEMS nose-fact or fiction? In: Proceedings of IEEE International Symposium on Industrial Electronics ISIE '97, Guimaraes, vol. 1, pp. SS152–SS157 (1997)
16. Borky, J.M., Wise, K.D.: Integrated signal conditioning for silicon pressure sensors. IEEE T. Electron. Dev. **ED-27**, 927–930 (1979)
17. Yoon, E., Wise, K.D.: A multi-element monolithic mass flowmeter with on-chip CMOS readout electronics. Technical Digest of Solid State Sensors and Actuators Workshop, Hilton Head, pp. 161–164 (1990)
18. Haberli, A., Paul, O., Malcovati, P., Faccio, M., Maloberti, F., Baltes, H.: CMOS integration of a thermal pressure sensor system. In: IEEE International Symposium on Circuits and Systems, ISCAS '96, Atlanta, vol. 1, pp. 377–380 (1996)
19. Schneider, M., Muller, T., Haberli, A., Hornung, M., Baltes, H.: Integrated micromachined decoupled CMOS chip on chip. In: Proceedings of 10th IEEE

International Workshop on MEMS, Nagoya, pp. 512–517 (1997)
20. Akiyama, T., Akiyama, T., Staufer, U., de Rooij, N.F., Lange, D., Hagleitner, C., Brand, O., Baltes, H., Tonin, A., Hidber, H.R.: Integrated atomic force microscopy array probe with metal-oxide-semiconductor field effect transistor stress sensor, thermal bimorph actuator, and on-chip complementary metal-oxide-semiconductor electronics. J. Vac. Sci. Technol. B **18**, 2669–2675 (2000)
21. Schaufelbuhl, A., Schneeberger, N., Munch, U., Waelti, M., Paul, O., Brand, O., Baltes, H., Menolfi, C., Huang, Q., Doering, E., Loepfe, M.: Uncooled low-cost thermal imager based on micromachined CMOS integrated sensor array. J. MEMS **10**, 503–510 (2001)
22. Verd, J., Uranga, A., Teva, J., Lopez, J.L., Torres, F., Esteve, J., Abadal, G., Perez-Murano, F., Barniol, N.: Integrated CMOS-MEMS with on-chip readout electronics for high-frequency applications. IEEE Electron. Dev. Lett. **27**, 495–497 (2006)
23. Laermer, F., Schilp, A.: Method of anisotropically etching silicon. US Patent 5,501,893, Robert Bosch Gmbh (1992)
24. Kruglick, E.J.J., Warneke, B.A., Pister, K.S.: CMOS 3-axis accelerometers with integrated amplifier. In: Proceedings of International Conference on Micro Electro Mechanical System, MEMS-98, pp. 631–636. Heidelberg. (1998)
25. Fedder, G.K., Santhanam, S., Reed, M.L., Eagle, S.C., Guillou, D.F., Lu, M.S.C., Carley, L.R.: Laminated high-aspect-ratio microstructures in a conventional CMOS process. Proceedings of International Conference on Micro Electro Mechanical Systems, MEMS-96, San Diego, pp. 13–18 (1996)
26. Xie, H., Erdmann, L., Zhu, X., Gabriel, K.J., Fedder, G. K.: Post-CMOS processing for high-aspect-ratio integrated silicon microstructures. J. Microelectromech. Syst. **11**, 93–101 (2002)
27. Xie, H., Fedder, G.K.: Fabrication, characterization, and analysis of a DRIE CMOS-MEMS gyroscope. IEEE Sens. J. **3**, 622–631 (2003)
28. Qu, H., Xie, H.: Process development for CMOS-MEMS sensors with robust electrically isolated bulk silicon microstructures. J. Microelectromech. Syst. **16**, 1152–1161 (2007)
29. Qu, H., Fang, D., Xie, H.: A monolithic CMOS-MEMS 3-axis accelerometer with a low-noise, low-power dual-chopper amplifier. IEEE Sens. J. **8**, 1511–1518 (2008)
30. Hagleitner, C., Lange, D., Hierlemann, A., Brand, O., Baltes, H.: CMOS single-chip gas detection system comprising capacitive, calorimetric and mass-sensitive microsensors. IEEE J. Solid-St. Circc **37**, 1867–1878 (2002)
31. Tsai, M.H., Sun, C.M., Liu, Y.C., Wang, C.W., Fang, W.L.: Design and application of a metal wet-etching post-process for the improvement of CMOS-MEMS capacitive sensors. J. Micromech. Microeng. **19**, 105017 (2009)

CMOS-CNT Integration

Huikai Xie and Ying Zhou
Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

C

Synonyms

Monolithic integration of carbon nanotubes

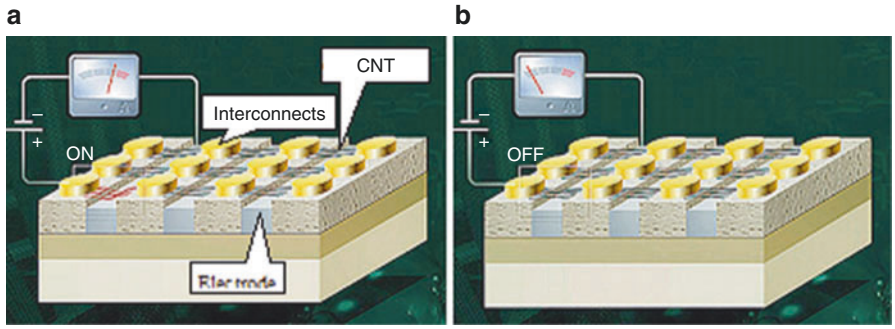
Definition

Carbon nanotubes can be directly grown on CMOS substrate without degrading the performance of CMOS electronics.

Introduction

With numerous outstanding electrical, mechanical, and chemical properties, carbon nanotubes (CNTs) have been explored for various applications with great success. As CMOS circuits possess powerful interfacing, signal amplification, conditioning, and processing capabilities, it is also highly desired to integrate CNTs with CMOS. CNTs may be either used as part of CMOS electronics or as sensing elements to form functioning nano-electromechanical systems (NEMS). Figure 1 shows a nanotube random-access memory (NRAM) which uses CNT ribbons as switches [1].

Many sensors based on CMOS-CNT hybrid systems have also been demonstrated, including mechanical, thermal, and chemical sensors [3, 4]. The integration of CMOS circuits with CNT sensors can increase signal-to-noise ratio and dynamic range, lower power consumption, and provide various controls and automations. Other efforts have been made to use multiwall carbon nanotubes (MWNTs) as CMOS interconnect for high frequency applications [5], or to apply CNT-based nano-electromechanical



CMOS-CNT Integration, Fig. 1 The structure of NRAM at (**a**) on and (**b**) off states [2]

switches for leakage reduction in CMOS logic and memory circuits [6].

However, monolithic integration of CMOS and CNTs is still very challenging. Most CMOS-CNT systems have been realized either by a two-chip solution or low-throughput CNT manipulations. In this entry, CMOS-CNT integration approaches are reviewed, with a particular focus on a localized heating CNT synthesis method that can grow CNTs on foundry CMOS.

CNT Synthesis

There are three main methods for carbon nanotube synthesis: arc-discharge [7], laser ablation [8], and chemical vapor deposition (CVD) [9]. The first two methods involve evaporation of solid-state carbon precursors and condensation of carbon atoms to form nanotubes, where high annealing temperature, typically over 1000 °C, is required to remove defects and thus produce high-quality nanotubes. However, they tend to produce a mixture of nanotubes and other by-products such as catalytic metals, so the nanotubes must be selectively separated from the by-products. This requires post-growth purification and manipulation.

In contrast, the CVD method employs a hydrocarbon gas as the carbon source and involves heating metal catalysts in a tube furnace to synthesize nanotubes. Nanotubes can grow either on the top (tip growth) or from the bottom (base growth). The diameters and locations of the grown CNTs can be controlled via catalyst size and catalyst patterning, and the orientation can be guided by an external electric field. Suitable catalysts that have been reported include Fe, Co, Mo, and Ni [10]. Compared to the arc-discharge and laser ablation methods, CVD uses much lower growth temperature, but it is still too high for direct CNT growth on CMOS substrates.

In addition, during or after CNT growth, electrical contacts need to be formed for functional CNT-based devices. It is reported that Mo provides good ohmic contacts with nanotubes and shows excellent conductivity after growth, with resistance ranging from 20 kΩ to 1 MΩ per tube [11]. Several other metals, such as palladium, gold, titanium, tantalum, and tungsten, have also been investigated as possible electrode materials.

CMOS-CNT Integration

To integrate CNT on CMOS, there are several factors that must be taken into account: temperature budget, material compatibility, CNT type, CNT quality, and contamination. Depending on when CNTs are made, CMOS-CNT integration technology can be categorized as follows:

- Pre-CMOS: CMOS processes will be performed after CNTs are synthesized in place
- Intra-CMOS: CNT growth steps are inserted into CMOS fabrication steps
- Post-CMOS: CNTs are introduced after all CMOS processes have been done

For pre-CMOS, CNTs must go through standard CMOS process steps. This is very difficult to realize. There are temperature constraints, material compatibility, and contamination issues. There is no report about pre-CMOS CNT integration yet. For intra-CMOS, CNTs will be introduced at a later stage in the CMOS fabrication sequence, so it is easier to protect CNTs than in the pre-CMOS case. But temperature and contamination issues still must be considered. Post-CMOS, on the hand, completely eliminates CMOS contamination issues. It has potential to achieve mass production and low cost, but the temperature remains a limiting factor.

Intra-CMOS CNT Integration

Intra-CMOS (High Temperature CNTs)

Thermal CVD has been used to grow CNTs directly on CMOS substrate. For example, Tseng et al. demonstrated, for the first time, a process that monolithically integrates SWNTs with n-channel metal oxide semiconductor (NMOS) FET in a CVD furnace at 875 °C [12]. However, the high synthesis temperature (typically 800–1000 °C for SWNT growth) may damage the aluminum metallization layers and change the characteristics of the on-chip transistors as well. Ghavanini et al. assessed the deterioration level of CMOS transistors with certain CNT CVD synthesis conditions applied, and they reported that one PMOS transistor lost its functions after the thermal CVD treatment (610 °C, 22 min) [13]. As a result, the integrated circuits in Tseng's thermal CVD CNT synthesis can only consist of NMOS and use n+ polysilicon and molybdenum as interconnects, which make it incompatible with foundry CMOS processes.

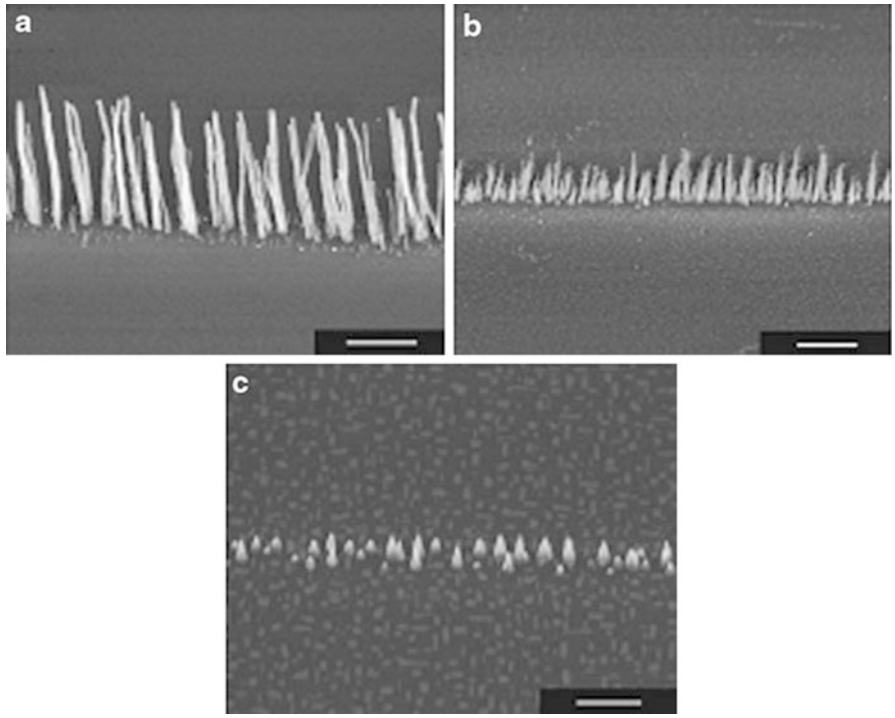
Intra-CMOS (Low Temperature CNTs)

Some other attempts have been made to develop low temperature growth using various CVD methods. Hofmann et al. reported vertically aligned carbon nanotubes grown at temperature as low as 120 °C by plasma-enhanced chemical vapor deposition (PECVD) [14]. However, the decrease in growth temperature jeopardizes both the quality and yield of the CNTs, as shown in Fig. 2. The synthesized products are actually defect-rich carbon nanofibers rather than MWNTs or SWNTs.

Intra-CMOS (Localized Heating)

Localized heating: To accommodate both the high temperature requirement (800–1000 °C) for high-quality SWNT synthesis and the temperature limitation of CMOS processing (<450 °C), CNT synthesis based on localized heating has drawn great interest recently. Englander et al. demonstrated, for the first time, the localized synthesis of silicon nanowires and carbon nanotubes based on resistive heating [15]. The fabrication processes are shown in Fig. 3. Operated inside a room temperature chamber, the suspended micro-electromechanical system (MEMS) structures serve as resistive heaters to provide high temperature at predefined regions for optimal nanotube growth, leaving the rest of the chip area at low temperature. Using the localized heating concept, direct integration of nanotubes at specific areas can be potentially achieved in a CMOS compatible manner, and there is no need for additional assembly steps. However, the devices typically have large sizes and their fabrication processes are not fully compatible with the standard foundry CMOS processes. Although this concept has solved the temperature incompatibility problem between CNT synthesis and CMOS circuit protection, the fabrication processes of microheater structures still have to be well designed to fit into standard CMOS foundry processes and the resistor materials must be selected to meet the CMOS compatibility criteria.

Using the localized heating technique described above, on-chip growth using CMOS micro-hotplates was demonstrated by Haque et al. [16]. As shown in Fig. 4, tungsten was used for both the micro-hotplates (as the heating source) and interdigitated electrodes for nanotubes contacts. MWNTs have been successfully synthesized on the membrane, and simultaneously connected to CMOS circuits through tungsten metallization. Although tungsten can survive the high temperature growth process, and has high connectivity and conductivity, Franklin et al. reported that no SWMTs were found to grow from catalyst particles on the



CMOS-CNT Integration, Fig. 2 SEM images of vertically aligned CNFs grown by PECVD deposition at (**a**) 500 °C, (**b**) 270 °C, and (**c**) 120 °C (scale bars: (**a**) and (**b**) 1 μm and (**c**) 500 nm) [14]

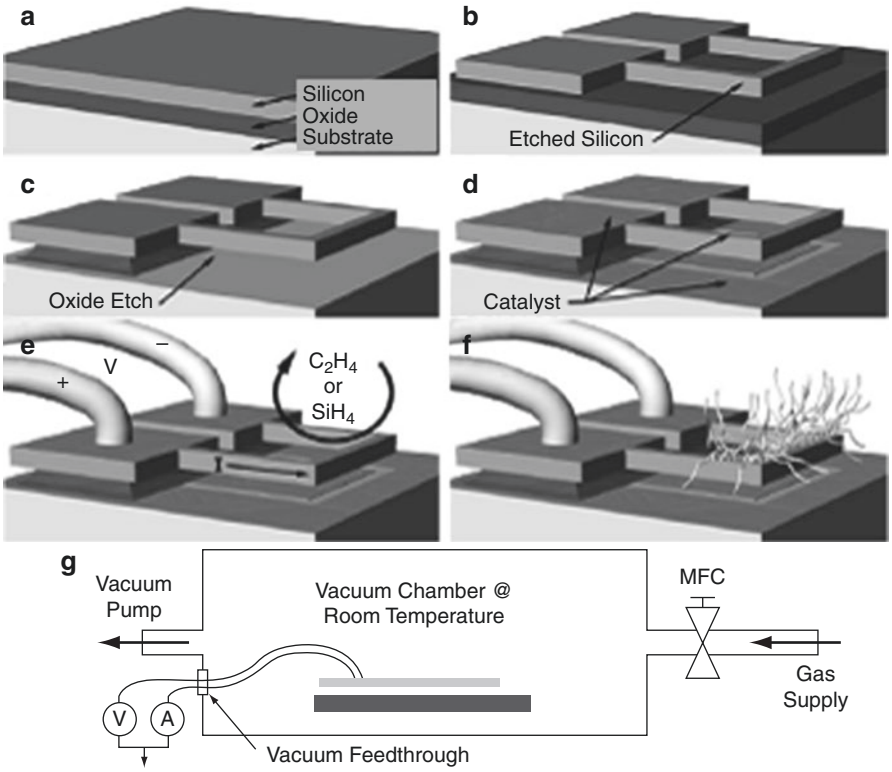
tungsten electrodes, presumably due to the high catalytic activity of tungsten toward hydrocarbons [11]. Further, although the monolithic integration has been achieved, the utilization of tungsten, a refractory metal, as interconnect metal is limited in foundry CMOS, especially for mixed-signal CMOS processes. Moreover, this approach requires a backside bulk micromachining process and is limited to SOI CMOS substrates.

Post-CMOS CNT Integration

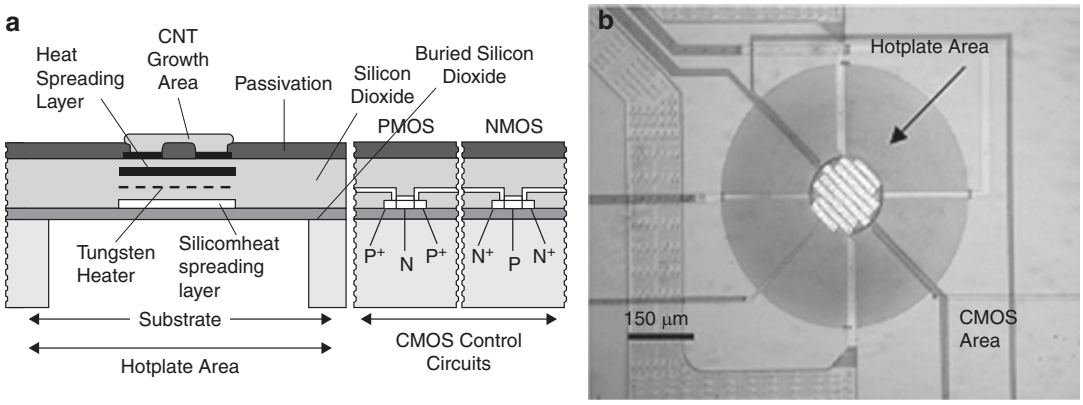
Post-CMOS (CNT Transfer and Assembly)

To overcome the temperature limitation, one possible solution is to grow nanotubes at high temperature first and then transfer them to the desired locations on CMOS substrates at low temperature. However, handling, maneuvering, and integrating these nanostructures with CMOS chips/wafers to form a complete system are very challenging. In the early stage, an atomic force microscope (AFM) tip was used to manipulate and position nanotubes into a predetermined location under the guide of scanning electron microscope (SEM) imaging [17]. Although this nanorobotic manipulation realized precise control over both the type and location of CNTs, its low throughput makes large scale assembly prohibitive.

Other post-growth CNT assembly methods include surface functionalization [18], liquid-crystalline processing [19], dielectrophoresis (DEP) [20], and large scale transfer of aligned nanotubes grown on quartz [21]. A 1 GHz CMOS circuit with CNT interconnects has been demonstrated using a DEP-assisted assembly technique [5]. The fabrication process flow and the assembled MWNT interconnect are shown in Fig. 5. The DEP process provides the capability of precisely positioning the nanotubes in a noncontact manner, which minimizes the parasitic capacitances and allows the circuits to operate at more than 1 GHz. However, to immobilize the DEP-trapped CNTs in place and to improve the electrical contact between CNTs and the



CMOS-CNT Integration, Fig. 3 Fabrication process and localized heating concept [15]

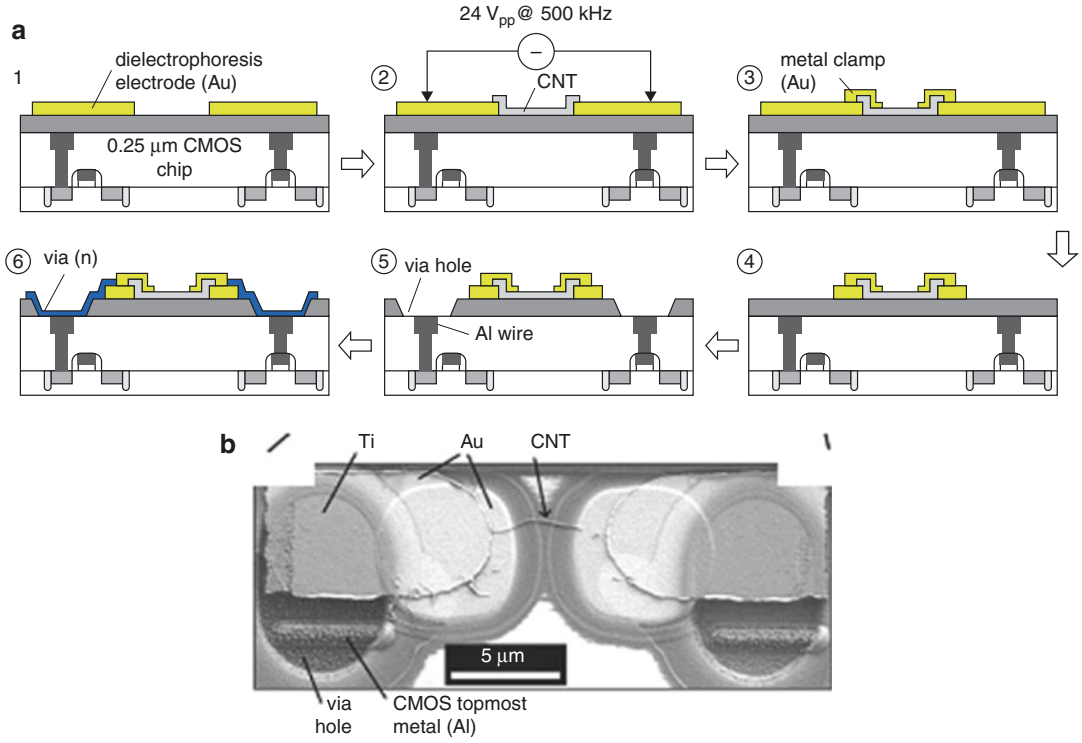


CMOS-CNT Integration, Fig. 4 (**a**) Schematic of the cross-sectional layout of the chip. (**b**) Optical image of the device top view showing the tungsten interdigitated electrodes on top of the membranes, heater radius = 75 µm, membrane radius = 280 µm [16]

electrodes, metal clamps must be selectively deposited at both ends of the CNTs (Fig. 5a, step 3). The process complexity and low yield (~8 %, due to the MWNT DEP assembly limitation) are still the major concerns.

Post-CMOS (Localized Heating)

Monolithic CMOS-CNT integration is desirable to fully utilize the potentials of nanotubes for emerging nanotechnology applications, but the approaches introduced above still cannot

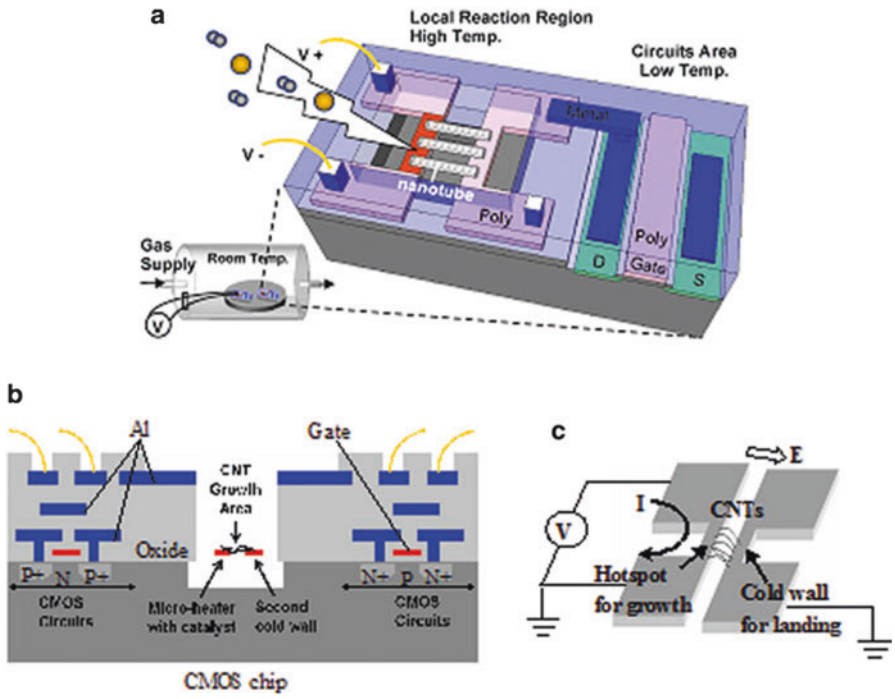


CMOS-CNT Integration, Fig. 5 (**a**) Process flow to integrate MWNT interconnects on CMOS substrate. (**b**) SEM image of one MWNT interconnect (wire and via) [5]

meet all the requirements and realize complete compatibility with CMOS processes. To solve the problem, a simple and scalable monolithic CMOS-CNT integration technique using a novel maskless post-CMOS surface micromachining processing has been proposed. This approach is fully compatible with commercial foundry CMOS processes and has no specific requirements on the type of metallization layers and substrates.

As illustrated in Fig. 6, the basic idea of the monolithic integration approach is to use maskless post-CMOS MEMS processing to form micro-cavities for thermal isolation and use the gate polysilicon to form resistors for localized heating as well as the nanotube-to-CMOS interconnect. The microheaters, made of the gate polysilicon, are deposited and patterned along with the gates of the transistors in the standard CMOS foundry processes. One of the top metal layers (i.e., the metal-3 layer as shown in Fig. 6b) is also patterned during the CMOS fabrication. It is used as an etching mask in the following post-CMOS microfabrication process for creating the micro-cavities. Finally, the polysilicon microheaters are exposed and suspended in a micro-cavity on a CMOS substrate. The circuits are covered under the metallization and passivation layers, as illustrated in Fig. 6b. Unlike the traditional thermal CVD synthesis in which the whole chamber is heated to above 800 °C, the CVD chamber is kept at room temperature all the time, with only the microheaters activated to provide the local high temperature for CNT growth (Fig. 6a, the red part represents the hot microheater).

The top view of a microheater design is shown in Fig. 6c. There are two polysilicon bridges: one as the microheater for generating high temperature to initiate CNT growth and the other for CNT landing. With the cold wall grounded, an E-field perpendicular to the surface of the two bridges will be induced during CNT growth. Activated by localized heating, the nanotubes will start to grow from the hotspot



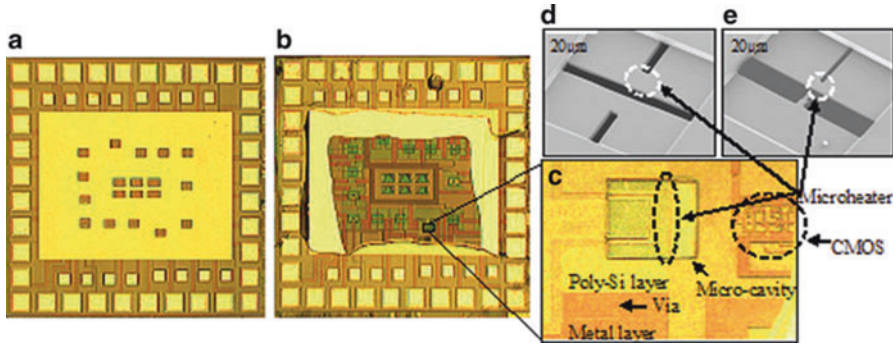
CMOS-CNT Integration, Fig. 6 (**a**) The 3D schematic showing the concept of the CMOS-integrated CNTs. The CVD chamber is kept at room temperature all the time. The *red* part represents the hot microheater that has been activated for high temperature nanotube synthesis. (**b**) Cross-sectional view of the device. (**c**) The schematic 3D microheater showing the local synthesis from the hotspot and self-assembly on the cold landing wall under the local electric field

(i.e., the center of the microheater) and will eventually reach the secondary cold bridge under the guide of the local E-field. Since both the microheater bridge and the landing bridge are made of the gate polysilicon layer and have been interconnected with the metal layers in CMOS foundry process, the as-grown CNTs can be electrically connected to the CMOS circuitry on the same chip without any post-growth clamping or connection steps.

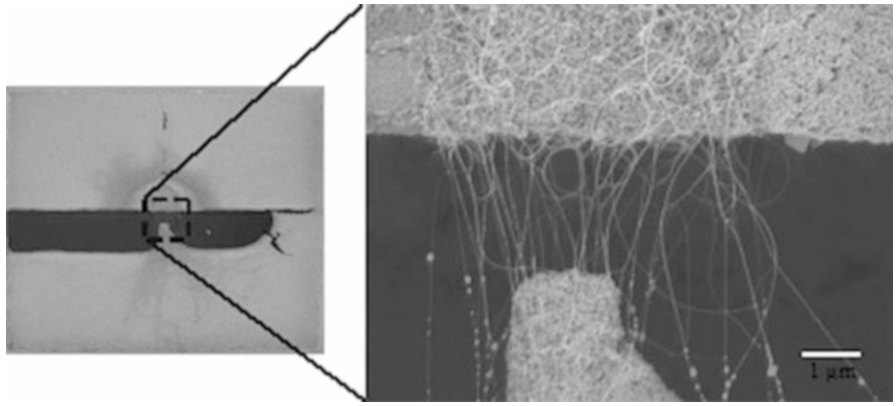
This technology has been verified at the chip level. The CMOS chips were fabricated in the AMI 0.5 μm 3-metal CMOS process. Optical microscope images of a CMOS chip before and after MEMs fabrication are shown in Fig. 7a, b. The total chip area is 1.5 × 1.5 mm^2, including test circuits and 13 embedded microheaters. SEMs of two microheaters are shown in Fig. 7d, e, with resistances of 97 and 117 Ω, respectively. At about 2.5 V, red glowing was observed for the design in Fig. 7e. This voltage was also used for the CNT growth.

Figure 8 shows one device with successful CNT growth, where individual suspended carbon nanotubes were grown from the 3 × 3 μm microheater shown in Fig. 7e and landed on the near polysilicon tip. The overall resistance of the CNTs is measured between the microheater and the cold polysilicon wall at room temperature. The typical resistances of in situ synthesized CNTs range from 5 to15 MΩ. The resistance variation from device to device is mainly due to the variation of the CNT quantity grown on each microheater. Junction effects of Schottky contacts were observed for self-assembled polysilicon/CNTs/polysilicon heterojunctions.

After successful synthesis of carbon nanotubes, the influence of the localized heating on nearby CMOS circuits was evaluated. Simple circuits, such as inverters, were tested and proved working properly. There was no change to the rising and falling time after the CNT growth. The dc electrical characteristics of individual



CMOS-CNT Integration, Fig. 7 (**a**) The CMOS chip photograph (1.5 × 1.5 mm^2) after foundry process; (**b**) The CMOS chip photograph after post-CMOS process (before final DRIE step); (**c**) Close-up optical image of one microheater and nearby circuit. CMOS circuit area, although visible, is protected under silicon dioxide layer. Only the microheater and secondary cold wall within the micro-cavity are exposed to synthesis gases. Polysilicon heater and metal wire are connected by via. (**d**) and (**e**) Closed-up SEM images of two microheaters



CMOS-CNT Integration, Fig. 8 Localized synthesis of carbon nanotubes grown from the 3 × 3 μm microheater, suspended across the trench and landed on the secondary polysilicon tip

transistors had no considerable change after CNT growth, demonstrating the CMOS compatibility of this integration approach.

Summary

CMOS-CNT integration has been demonstrated by using both intra- and post- CMOS processes. Several methods have been developed to overcome the temperature conflict between CNT growth and CMOS, including using high temperature refractory metals for interconnect, low temperature CVD, transferring/assembling CNTs prepared off site, and localized heating. Among these techniques, localized heating is very promising. Truly monolithic CNT-CMOS integration has been demonstrated on foundry CMOS substrate by employing MEMS and localized heating. This post-CMOS microfabrication is maskless, and the CNT growth does not affect the characteristics of the transistors on the same chip.

Cross-References

- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Synthesis of Carbon Nanotubes

References

1. Zhang, W., Jha, M., Shang, L.: NATURE: A hybrid nanotube/CMOS dynamically reconfigurable architecture. Design Automation Conference, 2006 43 rd ACM/IEEE, pp. 711–716 (2006)
2. Nantero, I.: "NRAM®," in http://www.nantero.com/mission.html, 2000–2009
3. Agarwal, V., Chen, C.-L., Dokmeci, M. R., Sonkusale, S.: A CMOS integrated thermal sensor based on single-walled carbon nanotubes. IEEE Sensors 2008 Conference, pp. 748–751 (2008)
4. Cho, T.S., Lee, K.-J., Kong, J., Chandrakasan, A.P.: A 32-uW 1.83-kS/s carbon nanotube chemical sensor system. IEEE J. Soild-State Circuits **44**, 659–669 (2009)
5. Close, G.F., Yasuda, S., Paul, B., Fujita, S., Wong, H.-S.P.: A 1 GHz integrated circuit with carbon nanotube interconnects and silicon transistors. Nano Lett. **8**, 706–709 (2008)
6. Chakraborty, R.S., Narasimhan, S., Bhunia, S.: Hybridization of CMOS with CNT-based nano-electromechanical switch for low leakage and Robust circuit design. IEEE Trans. Circuits Syst. **54**, 2480–2488 (2007)
7. Journet, C., Maser, W.K., Bernier, P., Loiseau, A., LamydelaChapelle, M., Lefrant, S., Deniard, P., Lee, R., Fischerk, J.E.: Large-scale production of single-walled carbon nanotubes by the electric-arc technique. Nature **388**, 756–758 (1997)
8. Guo, T., Nikolaev, P., Thess, A., Colbert, D.T., Smalley, R.E.: Catalytic growth of single-walled nanotubes by laser vaporization. Chem. Phys. Lett. **243**, 49–54 (1995)
9. Cassell, A.M., Raymakers, J.A., Kong, J., Dai, H.: Large scale CVD synthesis of single-walled carbon nanotubes. J. Phys. Chem. B **103**, 6484–6492 (1999)
10. Meyyappan, M.: Carbon Nanotubes: Science and Applications. CRC Press, New York (2005)
11. Franklin, N.R., Wang, Q., Tombler, T.W., Javey, A., Shim, M., Dai, H.: Integration of suspended carbon nanotube arrays into electronic devices and electromechanical systems. Appl. Phys. Lett. **81**, 913–915 (2002)
12. Tseng, Y.-C., Xuan, P., Javey, A., Malloy, R., Wang, Q., Bokor, J., Dai, H.: Monolithic integration of carbon nanotube devices with silicon MOS technology. Nano Lett. **4**, 123–127 (2004)
13. Ghavanini, F.A., Poche, H.L., Berg, J., Saleem, A.M., Kabir, M.S., Lundgren, P., Enoksson, P.: Compatibility assessment of CVD growth of carbon nanofibers on bulk CMOS devices. Nano Lett. **8**, 2437–2441 (2008)
14. Hofmann, S., Ducati, C., Robertson, J., Kleinsorge, B.: Low-temperature growth of carbon nanotubes by plasma-enhanced chemical vapor deposition. Appl. Phys. Lett. **83**, 135–137 (2003)
15. Englander, O., Christensen, D., Lin, L.: Local synthesis of silicon nanowires and carbon nanotubes on microbridges. Appl. Phys. Lett. **82**, 4797–4799 (2003)
16. Haque, M.S., Teo, K.B.K., Rupensinghe, N.L., Ali, S. Z., Haneef, I., Maeng, S., Park, J., Udrea, F., Milne, W. I.: On-chip deposition of carbon nanotubes using CMOS microhotplates. Nanotechnology **19**, 025607 (2008)
17. Huang, X.M.H., Caldwell, R., Huang, L., Jun, S.C., Huang, M., Sfeir, M.Y., O'Brien, S.P., Hone, J.: Controlled placement of individual carbon nanotubes. Nano Lett. **5**, 1515–1518 (2005)
18. Liu, J., Casavant, M.J., Cox, M., Walters, D.A., Boul, P., Lu, W., Rimberg, A.J., Smith, K.A., Colbert, D.T., Smalley, R.E.: Controlled deposition of individual single-walled carbon nanotubes on chemically functionalized templates. Chem. Phys. Lett. **303**, 125–129 (1999)
19. Ko, H., Tsukruk, V.V.: Liquid-crystalline processing of highly oriented carbon nanotube arrays for thin-film transistors. Nano Lett. **6**, 1443–1448 (2006)
20. Schwamb, T., Schirmer, N.C., Burg, B.R., Poulikakos, D.: Fountain-pen controlled dielectrophoresis for carbon nanotube-integration in device assembly. Appl. Phys. Lett. **93**, 193104 (2008)
21. Ryu, K., Badmaev, A., Wang, C., Lin, A., Patil, N., Gomez, L., Kumar, A., Mitra, S., Wong, H.-S.P., Zhou, C.: CMOS-analogous wafer-scale nanotube-on-insulator approach for submicrometer devices and integrated circuits using aligned nanotubes. Nano Lett. **9**, 189–197 (2009)

CMOS-MEMS

► CMOS MEMS Fabrication Technologies

CMOS-MEMS Resonators

Sheng-Shian Li
Institute of NanoEngineering and MicroSystems, National Tsing Hua University, Hsinchu, Taiwan

Synonyms

MEMS resonators integrated with circuits; MEMS/IC integrated resonator circuits; Micromechanical resonators fabricated using CMOS-MEMS technologies

Definition

Based upon a strict definition, a CMOS-MEMS resonator is fabricated using a CMOS foundry orientated process to realize MEMS/IC integration for resonator applications, such as oscillators and filters. The integrated CMOS-MEMS fabrication platform is a key technology to reduce the form factor, enhance the performance, increase the functionality, and facilitate circuit integration for portable sensing devices and Internet of Things (IoTs). The integrated CMOS-MEMS circuits greatly reduce the parasitic stray capacitances at the MEMS-circuit interface so that it not only improves the system responses at high frequencies but also reduces the power consumption of the electronic circuits. This platform benefits the future portable/wearable electronic products for smaller size and longer standby time.

Principle of Operation

CMOS-MEMS resonators are mainly formed by the CMOS back-end-of-line (BEOL) materials, including interlayer dielectrics (ILD), e.g., SiO_2, and conductive metals, e.g., AlCu or Cu [1]. The target of this technology is to facilitate MEMS/IC monolithic integration without the need of additionally complex processes, such as heterogeneous integration using wafer bonding or 3D IC. Figure 1 presents two configurations of CMOS technology where the BEOL is highlighted. As can be seen, the BEOL materials will be used as MEMS structures of resonators, while the transistor circuits are naturally integrated with the resonators, thus leading to a much smaller form factor and less parasitics as compared to the current technologies (e.g., two-chip or heterogeneous integration solution). Since the capacitive transducers inherently exist, they are used as the major driving and sensing mechanism for CMOS-MEMS resonators [1–4] although the piezoresistive readout [5] can also be implemented once the polysilicon material (cf. Fig. 1) serves as the piezoresistors for motion detection. In this entry, only the pure capacitively transduced CMOS-MEMS resonator technology, i.e., the mainstream transduction in CMOS-MEMS, is introduced.

Resonator Structure

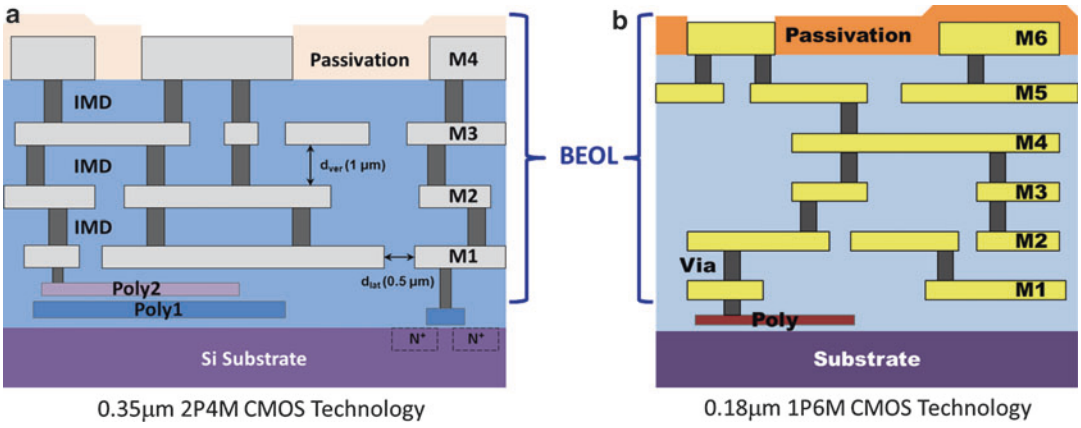
Based on various maskless post-CMOS release processes (i.e., no additional lithography step is required), the CMOS-MEMS resonator structures can be categorized into (a) mere-metal, (b) metal-rich composite, (c) and oxide-rich composite configurations, as shown in Fig. 2. The implemented processes resulting in structures in Fig. 2 will be discussed in the next section. With these structures, the CMOS-MEMS resonators exhibit different performance in terms of quality factor (Q) and thermal stability (i.e., temperature coefficient of frequency, TC_f).

Resonance Frequency

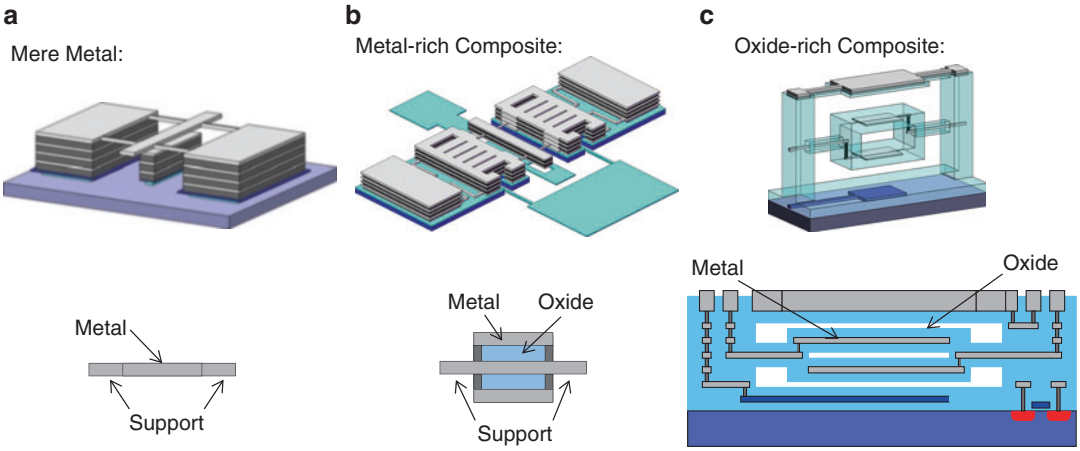
To determine the resonance frequency of the CMOS-MEMS composite beam resonators indicated in Fig. 2b and c, the Euler-Bernoulli approach can be used to attain a first-order approximation of the resonance frequency (f_{nom}) using a formula given by

$$f_{nom} = \frac{1}{2\pi}\sqrt{\frac{k_m}{m_{re}}} = \frac{1}{2\pi}(\beta_1 L_r)^2\sqrt{\frac{\sum(E_i I_i)}{\sum(\rho_i A_i)}}\frac{1}{L_r^2} \quad (1)$$

where k_m and m_{re} are the mechanical stiffness and effective mass, respectively, of the beam resonator, L_r is the length of the beam, β_1 represents the frequency parameter of a fundamental mode of the beam, i represents the corresponding CMOS structural materials (e.g., metal and SiO_2), and E_i, ρ_i, A_i, and I_i are the Young's modulus, density, cross-sectional area, and moment of inertia of each structural layer, respectively. If the mere-metal resonator is designed, Eq. 1 can be greatly simplified by using only one uniform structural material. Precise resonance frequencies can be simulated by the finite-element analysis once more complicated or accurate structures of the CMOS-MEMS resonators are considered, as shown in Fig. 3.



CMOS-MEMS Resonators, Fig. 1 CMOS configurations. (**a**) 0.35 µm 2-poly-4-metal (2P4M) CMOS technology node. (**b**) 0.18 µm 1-poly-6-metal (1P6M) CMOS technology node



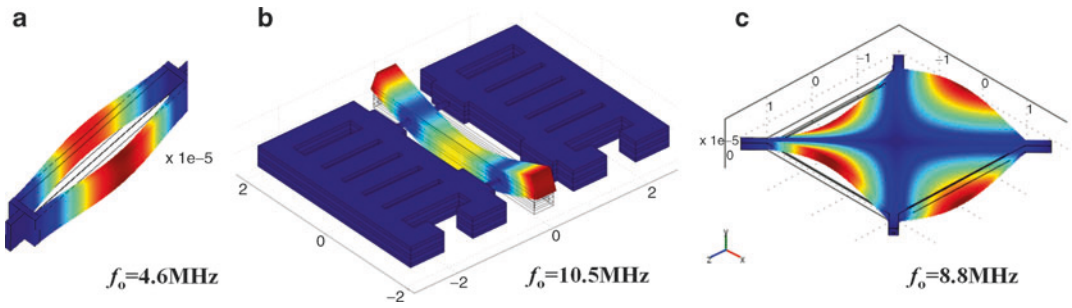
CMOS-MEMS Resonators, Fig. 2 Typical structures of CMOS-MEMS resonators in their perspective and cross-sectional views. (**a**) Mere-metal resonator. (**b**) Metal-rich composite resonator where metal layers enclose oxide. (**c**) Oxide-rich composite resonator where oxide layers enclose metal

Transduction and Overall Transfer Function

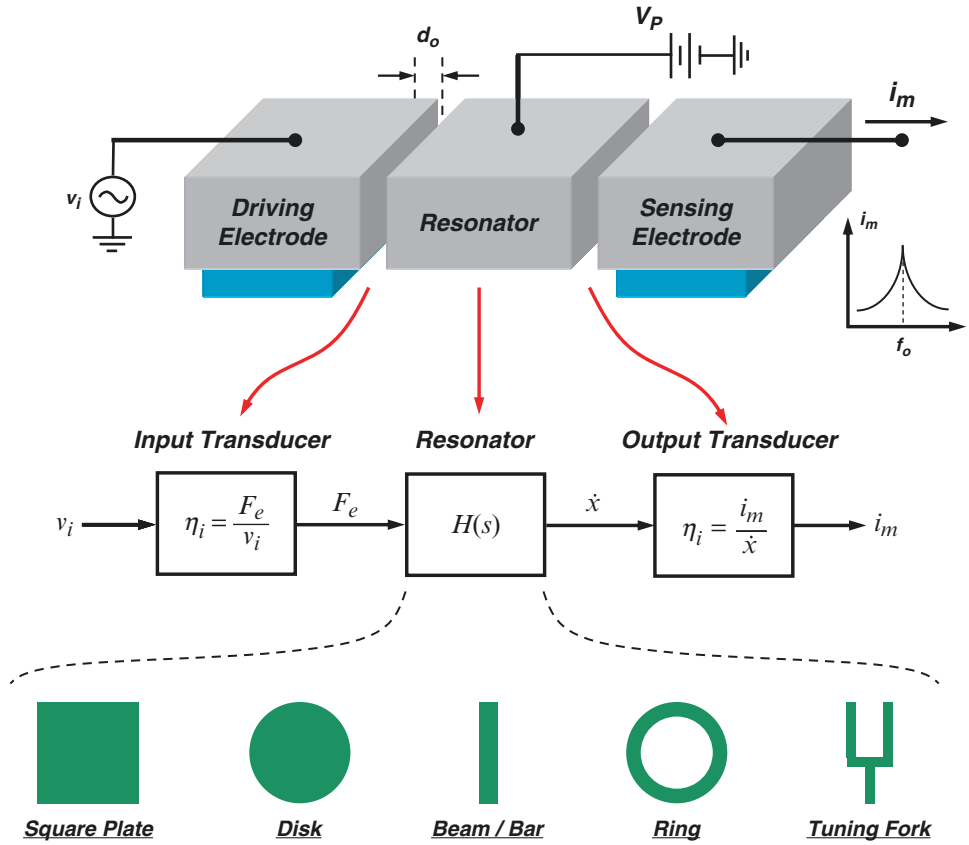
After the resonance frequency is set through Eq. 1 based on the geometrical dimensions and material properties of a given resonator, one has to determine the capacitive transduction efficiency to obtain the most important parameter, the motional impedance R_m, and overall transmission of the CMOS-MEMS resonator. A block diagram shown in Fig. 4 can be used to facilitate the understanding of the CMOS-MEMS resonator which is modeled in an electrical domain by linking the mechanical transfer function $H(s)$ with the input/output electromechanical coupling coefficients, η_i and η_o. The overall admittance transfer function $Y = i_m/v_i$ of the CMOS-MEMS resonator can be expressed by

$$Y(s) = \frac{i_m(s)}{v_i(s)} = \left(\frac{F_e}{v_i}\right) \times \left(\frac{\dot{x}}{F_e}\right) \times \left(\frac{i_m}{\dot{x}}\right) = \eta_i \times H(s) \times \eta_o \quad (2)$$

where v_i and i_m represent the small-signal input driving voltage and output motional current of the resonator, F_e is the electrostatic force generated by v_i to drive the resonator into resonance, and $\dot{x}$

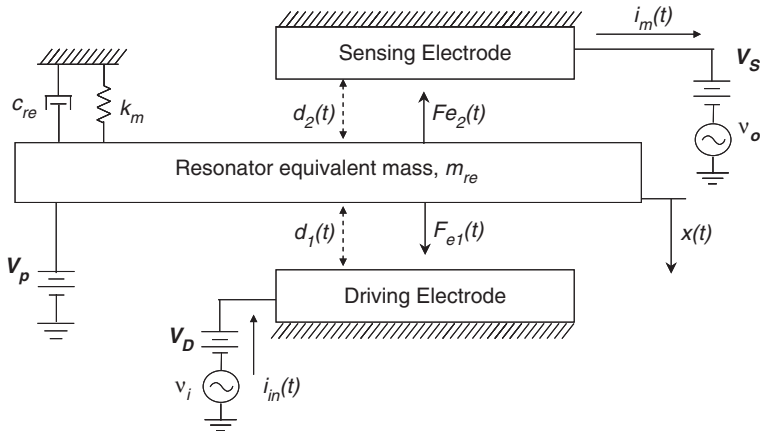


CMOS-MEMS Resonators, Fig. 3 Finite-element simulated modal frequencies and corresponding mode shapes of CMOS-MEMS resonators. (**a**) Tuning fork. (**b**) Free-free beam. (**c**) Flexural mode plate



CMOS-MEMS Resonators, Fig. 4 Conceptual schematic and block-diagram representation of a generalized capacitive MEMS resonator

represents the velocity of the resonator due to vibration. This is why η_i and η_o are treated as the electromechanical coupling coefficients where η_i converts the "electrical signal v_i" into the "mechanical force F_e" while η_i transforms the "mechanical velocity $\dot{x}$" into the "electrical current i_m". The mechanical transfer function $H(s)$ in Fig. 4 represents a general high-Q biquad response (cf. inset of Fig. 4) of the vibrating resonators, some of which are illustrated at the



CMOS-MEMS Resonators, Fig. 5 Equivalent lumped mechanical model for a capacitive MEMS resonator

bottom. Each type of resonators in Fig. 4 can be modeled by a simple mass-spring-damper system, as shown in Fig. 5, where the capacitive transducer gaps and voltage differences between the electrodes and the resonator are assumed to be different to preserve generality.

To derive the equivalent mechanical transfer function, $H(s)$, the equation of motion for the second-order mechanical system is used, given by

$$m_{re}\ddot{x} + c_{re}\dot{x} + k_m x = F_e \tag{3}$$

where x is the position of the mass away from its equilibrium position, m_{re} is the equivalent mass, c_{re} is the equivalent damping, k_m is the equivalent linear spring constant of the resonator, and F_e is the external driving force, respectively. Each element in Eq. 3 is assumed to be linear.

The force-to-velocity transfer function $H(s)$ can then be derived as

$$H(s) = \frac{\dot{x}}{F_e} = \frac{sX(s)}{F_e(s)} = \frac{1/m_{re}}{s^2 + \frac{c_{re}}{m_{re}}s + \frac{k_m}{m_{re}}}. \tag{4}$$

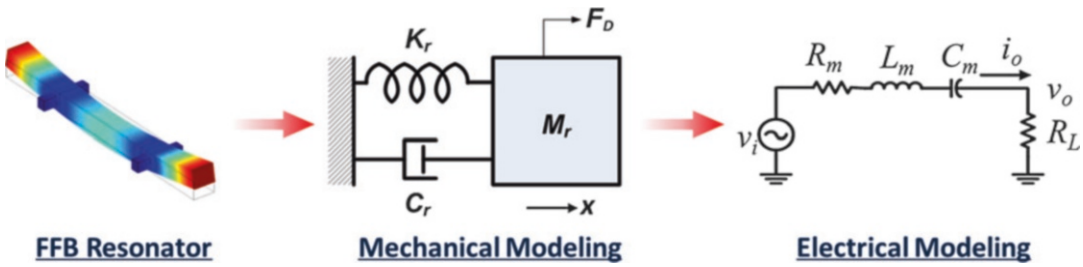
In Fig. 5, V_D, V_S, and V_P are the DC bias voltages applied into the input electrode, output electrode, and resonator structure, respectively. Based on the derivation of the capacitive transducers, the first-order electrostatic force and the output motional current of the resonator are given by

$$\begin{aligned} F_e &= -(V_P - V_D)\frac{\partial C_1}{\partial x}v_i \\ &= (V_P - V_D)\frac{\varepsilon_o A_1}{d_1^2}v_i \text{ and } \eta_i \equiv \frac{F_e}{v_i} \\ &= -(V_P - V_D)\frac{\partial C_1}{\partial x} \end{aligned} \tag{5}$$

$$\begin{aligned} i_m &= -(V_P - V_S)\frac{\partial C_2}{\partial x}\dot{x} \\ &= (V_P - V_S)\frac{\varepsilon_o A_2}{d_2^2}\dot{x} \text{ and } \eta_o \equiv \frac{i_m}{\dot{x}} \\ &= -(V_P - V_S)\frac{\partial C_2}{\partial x} \end{aligned} \tag{6}$$

where ε_o is the air permittivity; C_1 and C_2 are capacitors of the drive/sense ports, respectively; A_1 and A_2 are capacitive transducer areas of the drive/sense ports, respectively; and d_1 and d_2 are the capacitive transducer gaps of the drive/sense ports, respectively.

To simply describe the modeling procedure, Fig. 6 shows the CMOS-MEMS resonator, such as a free-free beam resonator, first modeled in mechanical domain as a mass-spring-damper (*MKC*) system. Then the *MKC* system is further transformed into a real *RLC* equivalent circuit in electrical domain through the electromechanical analogy. Based on Eqs. 2, 3, 4, 5, and 6, an equivalent series *RLC* model of an electromechanical resonator shown in Fig. 6 can be obtained with the motional elements expressed as



CMOS-MEMS Resonators, Fig. 6 Equivalent lumped electrical model for a capacitive MEMS resonator

$$L_m = \frac{m_{re}}{\eta_i \times \eta_o}, C_m = \frac{\eta_i \times \eta_o}{k_m}, R_m = \frac{c_{re}}{\eta_i \times \eta_o}. \tag{7}$$

For most capacitively transduced micromechanical resonators, their larger-than-conventional impedance (i.e., motional impedance R_m) is the main issue for system integration. Assuming the gap size d_o and overlap area A of the input and output transducers are symmetric, the motional impedance R_m in Eq. 7 can be further simplified as

$$R_m = \frac{c_{re}}{\eta_e^2} \cong \frac{\sqrt{k_m m_{re}} d_o^4}{Q \varepsilon_o^2 A^2 V_P^2} \tag{8}$$

where $\eta_e = \eta_i = \eta_o$, $d_o = d_1 = d_2$, $A = A_1 = A_2$, and Q is the quality factor of the resonator.

It is evident that to reduce R_m of a capacitive resonator, the stiffness, mass, gap size, and overlap area must be carefully designed under a V_P-constrained condition (high bias V_P is not desirable). Several design approaches can be adopted to reduce R_m, including (i) the gap size d_o should be as small as possible, (ii) the transducer overlap area A should be as large as possible, and (iii) the Q factor of the resonator should be as high as possible.

Thermal Stability

The nominal mechanical resonance frequency of the composite beam resonator (cf. Fig. 2b) composed of three structural materials, including metal (i.e., aluminum alloy), tungsten, and oxide, can be expressed as a combination of the resonance frequency of each portion for a given composite beam by an algebraic manipulation of Eq. 9 and given by [6]

$$f_{nom}{}^2 = \frac{m_{\text{metal}}}{m_r} f_{\text{metal}}{}^2 + \frac{m_{\text{tungsten}}}{m_r} f_{\text{tungsten}}{}^2 + \frac{m_{\text{oxide}}}{m_r} f_{\text{oxide}}{}^2 \tag{9}$$

where m_{metal}, m_{tungsten}, and m_{oxide} are the portion of mass of the composite beam for metal, tungsten, and oxide constituents, respectively, and f_{metal}, f_{tungsten}, and f_{oxide} are the mechanical resonance frequencies of beams for metal, tungsten, and oxide constituents, respectively. As a result, the linear temperature coefficient of frequency (TC_f) of the composite free-free beam resonator can be derived by taking the derivative of Eq. 9 with respect to temperature and then simplified as

$$TC_{f1} = \frac{TC_{f1,\ \text{metal}} + A \cdot TC_{f1,\ \text{tungsten}} + B \cdot TC_{f1,\ \text{oxide}}}{1 + A + B} \tag{10}$$

where $A = \frac{m_{\text{tungsten}} \cdot f^2_{\text{tungsten}}}{m_{\text{metal}} \cdot f^2_{\text{metal}}} = \frac{E_{\text{tungsten}} \cdot I_{\text{tungsten}}}{E_{\text{metal}} \cdot I_{\text{metal}}}$ and $B = \frac{m_{\text{oxide}} \cdot f^2_{\text{oxide}}}{m_{\text{metal}} \cdot f^2_{\text{metal}}} = \frac{E_{\text{oxide}} \cdot I_{\text{oxide}}}{E_{\text{metal}} \cdot I_{\text{metal}}}$ and where $TC_{f1,\text{metal}}$, $TC_{f1,\text{tungsten}}$, and $TC_{f1,\text{oxide}}$ are the linear TC_fs of metal, tungsten, and oxide, respectively; E_{metal}, E_{tungsten}, and E_{oxide} are Young's moduli of metal, tungsten, and oxide, respectively; and I_{metal}, I_{tungsten}, and I_{oxide} are the moments of inertia of metal, tungsten, and oxide, respectively. In addition, the linear TC_f of each material in Eq. 10 can be expressed as [6]

$$TC_{f1,\ \text{mat}} = \frac{TC_{E1,\ \text{mat}} + \alpha_{1,\ \text{mat}}}{2} \tag{11}$$

where $TC_{E1,\,mat}$ and $\alpha_{1,\,mat}$ are the linear TC_E and the coefficient of thermal expansion (CTE), respectively, of each structural material. In Eq. 11, $TC_{E1,\,mat}$ is the major parameter to induce the resonance frequency drift as temperature changes (i.e., $TC_{f1,\,mat}$), indicating the structural oxide with positive TC_E is effective for passive temperature compensation of CMOS-MEMS resonators. To approach the ultimate goal of zero TC_f, thermal stability of the CMOS-MEMS composite resonators can be further improved by manipulating the ratio A and ratio B in Eq. 10, respectively. As an extreme case when both A and B are close to zero, it represents that TC_f of such a resonator is exactly the same with that of metal-type resonators, while structures with A of zero and infinite B belong to oxide-type resonators exhibiting positive TC_f.

Methods of Fabrication

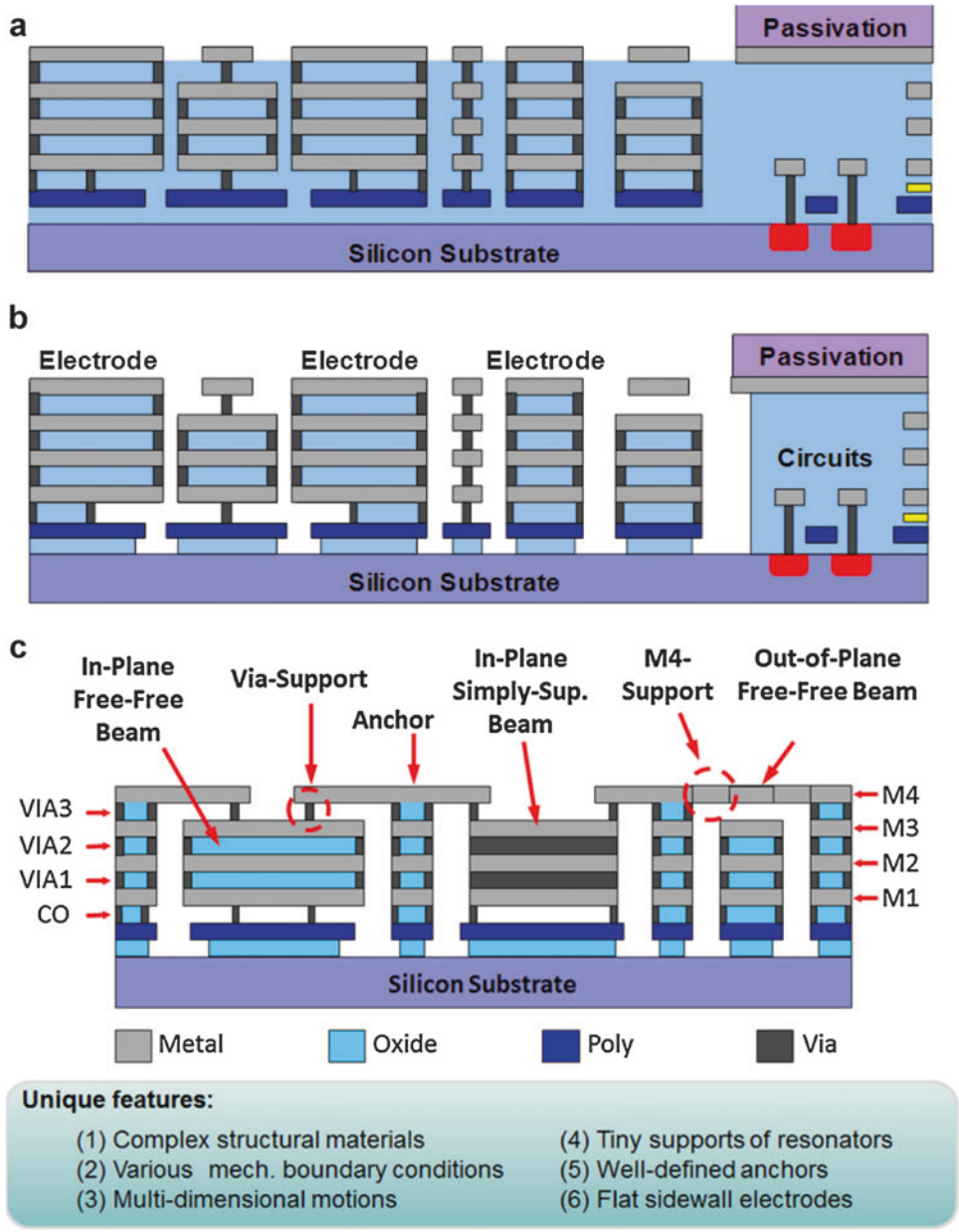
Several CMOS-MEMS resonator platforms [1–4] co-fabricating mechanical resonators and their amplifier circuits for MEMS/IC integration have been developed in the past decades. These platforms can be mainly categorized into two different release methods – oxide removal and metal removal maskless post-CMOS processes, implemented in both 0.35 μm 2-poly-4-metal (2P4M) and 0.18 μm 1-poly-6-metal (1P6M) CMOS technologies. In the following contents, their pros and cons will be discussed.

Oxide Removal Release Process in a 0.35 μm CMOS

To fabricate resonators using the platform illustrated in Fig. 7, chips are manufactured utilizing a standard 0.35 μm 2-poly-4-metal CMOS service from the Taiwan Semiconductor Manufacturing Company (TSMC) with a cross-sectional view shown in Fig. 7a, including two polysilicon layers underneath four metal layers [1]. CMOS circuit areas are masked by the passivation layer mostly comprised of silicon nitride while the MEMS regions (i.e., openings) expose sacrificial oxide to the release etchant. A commercial SiO_2 etchant, silox vapox III, with very high selectivity to metal layers, vias (i.e., tungsten), and contacts (i.e., tungsten) is utilized to release the resonator structures as depicted in Fig. 7b without the help of critical point dryers while the transistor circuits are still protected by the passivation layer. Figure 7c indicates that CMOS-MEMS resonators fabricated using this platform specifically possess several unique features including (i) complex structural materials which can be made of metal-oxide composite, metal composite, and mere-metal structures; (ii) various mechanical boundary conditions for resonators, such as fixed, pinned-pinned boundary, and free-free boundary designs; (iii) multidimensional displacements of resonators capable of in-plane and out-of-plane motions with respect to the substrate surface; (iv) standard CMOS vias (VIA) and contacts (CO) acting as tiny supports of resonators to effectively isolate the vibrating energy from resonators to their anchors; (v) well-defined anchors without undercut issue which is often seen in conventional CMOS-MEMS [7–11] or SOI process [12]; and (vi) better transducer efficiency attained by utilizing via-connected walls to form a flat sidewall electrodes, all of which offer a variety of flexible options suited for sensor and RF applications. Figure 8 presents the fabricated resonators, such as beam and comb-drive resonators. Note that the minimal electrode-to-resonator gap spacing is limited by the feature size of the 0.35 μm CMOS technology node such that the motional impedance of the fabricated resonators is in several MΩs due to their low electromechanical coupling coefficient.

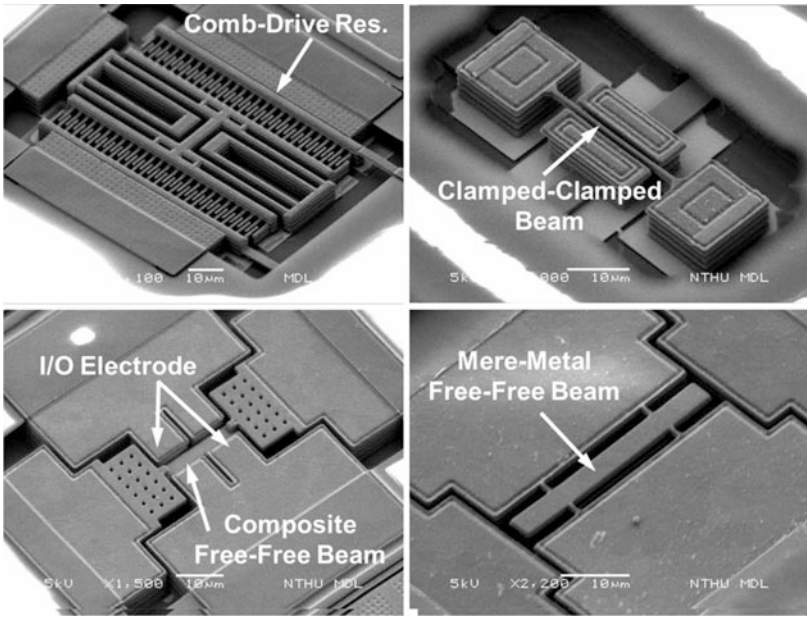
Oxide Removal Release Process in a 0.18 μm CMOS

To address the issue of high motional impedance in the previous platform, one can transfer this oxide removal process of section "Oxide Removal Release Process in a 0.35 μm CMOS" into a 0.18 μm CMOS as indicated in Fig. 9a [2]. The cross-sectional view of CMOS 0.18 μm BEOL and SEM of the fabricated resonator are shown in Fig. 9b, c, respectively. The electrode-to-resonator gap spacing is reduced



CMOS-MEMS Resonators, Fig. 7 Fabrication process flow for a CMOS-MEMS resonator in a 0.35 μm 2P4M technology under the oxide removal post process. (**a**) Chip prepared by the CMOS foundry. (**b**) Device is released by a buffer HF solution. (**c**) Various configurations of CMOS-MEMS resonators are realized in the platform

through the minimal feature size of the 0.18 μm CMOS (1.8X smaller than the 0.35 μm CMOS), while the transducer area can be increased by the six-metal stacking (four-metal stacking in the 0.35 μm CMOS). The combined merit of the gap and area leads to much lower motional impedance as compared to the 0.35 μm CMOS-MEMS resonators. The use of the 0.18 μm 1P6M platform implies the advanced technology nodes would lower the motional impedance, reduce the DC bias voltage, and attain higher resonance frequencies.



CMOS-MEMS Resonators, Fig. 8 SEM photos of the fabricated resonators using the oxide removal release process in a 0.35 μm CMOS technology

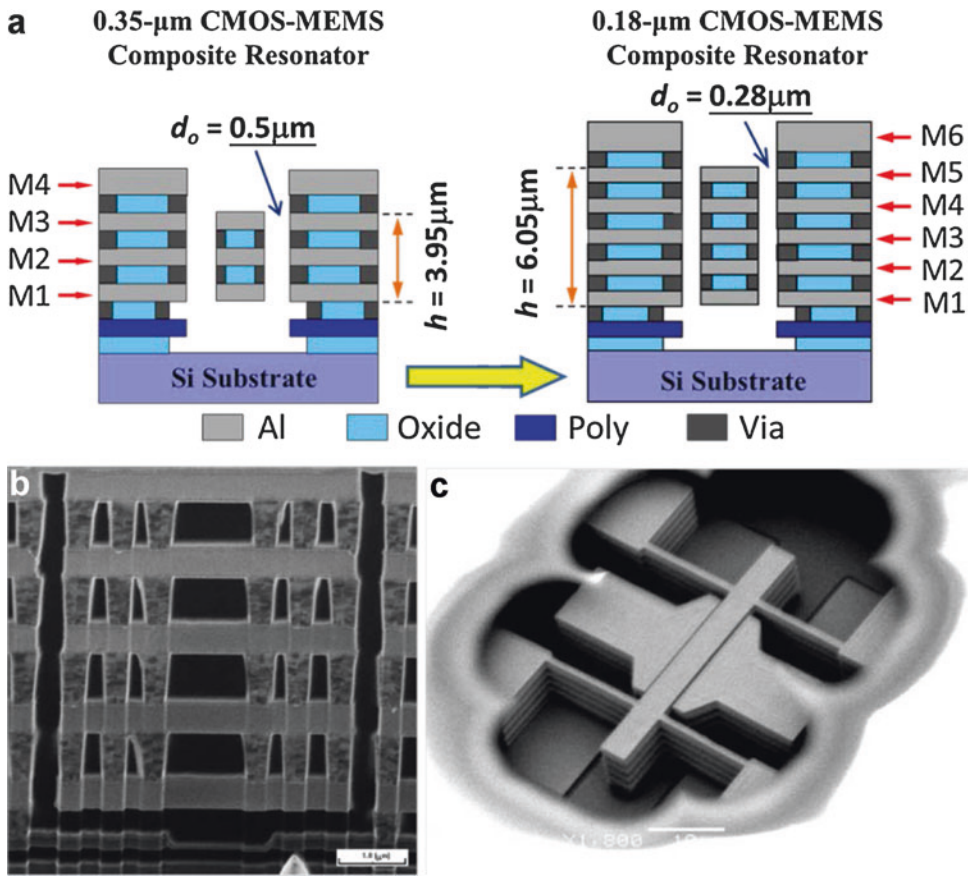
Metal Removal Release Process in a 0.35 μm CMOS

Although the oxide removal process only requires one maskless wet etching step, the metal-rich feature of the aforementioned resonators places a bottleneck on quality factor Q since the structural metal (AlCu) is often treated as a lossy acoustic material. Q of metal-rich resonators is often lower than 2,000, which is not sufficient for practical oscillator and filter implementations [3]. To address the low-Q issue in the oxide removal process, a metal removal post process has been developed as an alternative to form low-loss oxide-rich resonator structures to enable high Q.

As indicated in Fig. 10a, the device is first prepared by the CMOS foundry using a standard 0.35 μm 2P4M CMOS process. Then the etching solution containing H_2SO_4 and H_2O_2 is utilized to remove metal layers. This maskless etching technique provides excellent selectivity between the metal and dielectric layers; therefore, the issues from time-based etching in sections "Oxide Removal Release Process in a 0.35 μm CMOS" and "Oxide Removal Release Process in a 0.18 μm CMOS" can be greatly alleviated. Finally, reactive ion etching (RIE) is used to remove the passivation layer for the later wire bonding. Figure 10b presents a resonator SEM photo and its cross-sectional view, showing the metal layers are embedded inside the main SiO_2 structure. Q of the oxide-rich resonator is greater than 6,000, which is 6X higher than the metal-rich counterparts in sections "Oxide Removal Release Process in a 0.35 μm CMOS" and "Oxide Removal Release Process in a 0.18 μm CMOS."

Metal Removal Release Process in a 0.18 μm CMOS

The release approach in section "Metal Removal Release Process in a 0.35 μm CMOS" has also been transferred into a 0.18 μm CMOS node as shown in Fig. 11 to gain the smaller feature size of the advanced CMOS for low motional impedance R_m [4]. To fabricate the resonators, the chips are manufactured using a standard TSMC 0.18 μm 1P6M CMOS foundry process. Then a metal wet etchant, comprising H_2SO_4 and H_2O_2, is utilized to remove the sacrificial metals, hence providing lateral air-gap spacing d_o of 0.38 μm. Finally, the RIE is utilized to open bond pads for later wire bonding. As can be seen in the cross-sectional view of the resonator in Fig. 11b, the resonator structure consists of more than 97 % SiO_2. As shown in Fig. 11c, such oxide-rich resonators



CMOS-MEMS Resonators, Fig. 9 CMOS-MEMS resonator in a 0.18 μm 1P6M CMOS-MEMS platform. (**a**) Technology transformation. (**b**) Cross-sectional view of 0.18 μm BEOL. (**c**) Free-free beam resonator realized in a 0.18 μm CMOS-MEMS platform

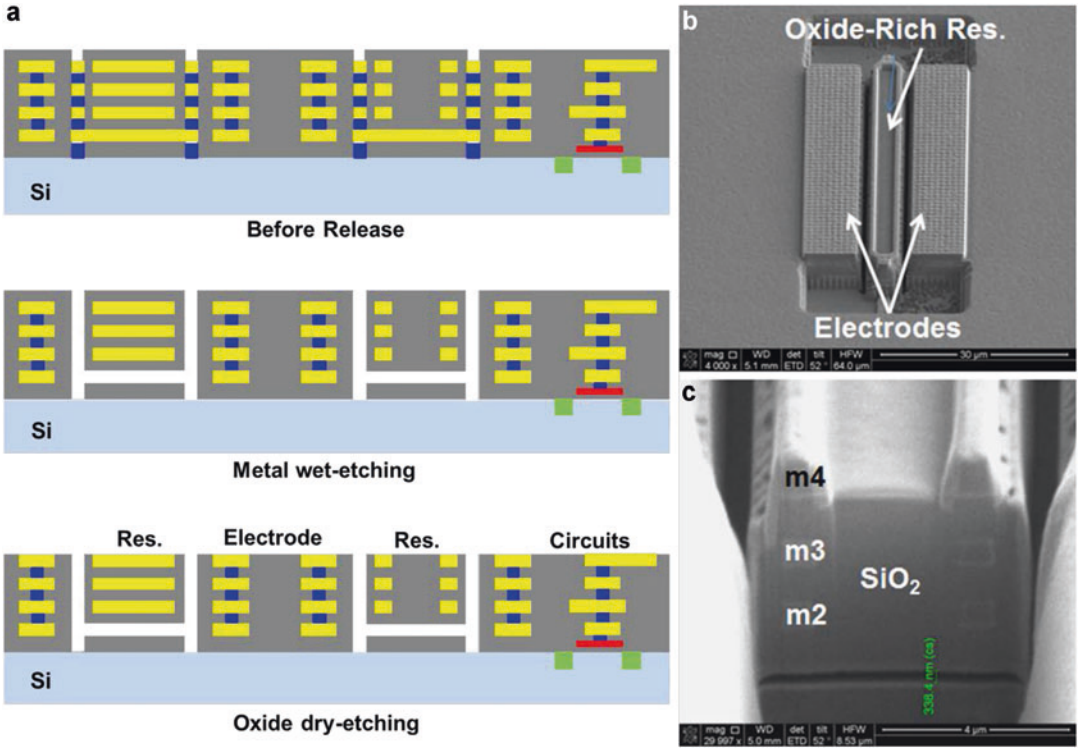
possess Q more than 10,000, which is comparable to commercial single-crystal silicon (SCS) or polysilicon-based resonators. Even tested in air, Q of the resonators is still greater than 6,000, an evidence of less squeeze film damping effect and suitable for resonant sensor applications.

Main Research Accomplishment

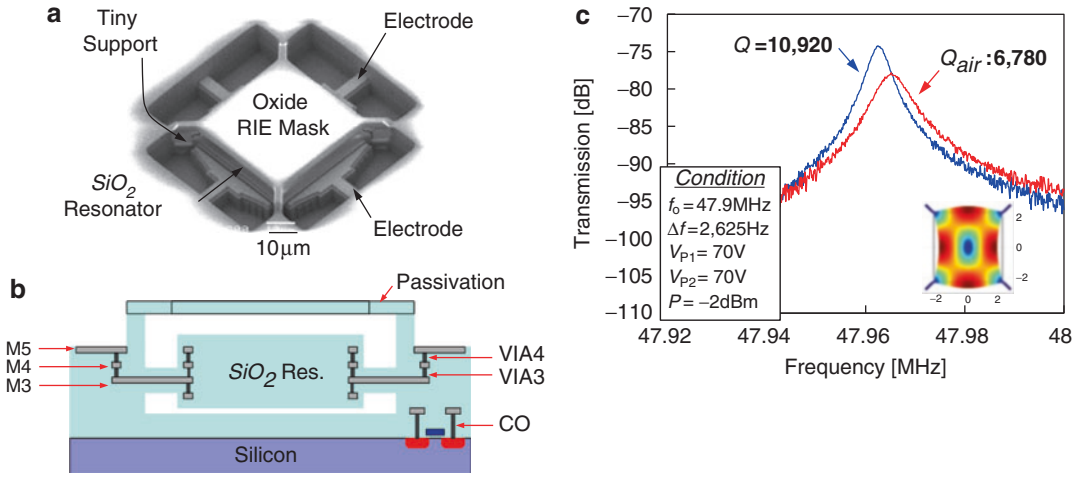
Motional Impedance R_m Reduction

The motional impedance governed by Eq. 8 is strongly dependent on the electrode-to-resonator gap spacing d_o. Figure 12 presents the performance comparison of similar free-free beam designs implemented in the 0.35 μm 2P4M and 0.18 μm 1P6M platforms, showing the advanced technology node (0.18 μm) would significantly lower the motional impedance R_m due to its smaller gap spacing (1.8X) and larger transduction area (1.6X) as compared to the 0.35 μm 2P4M technology.

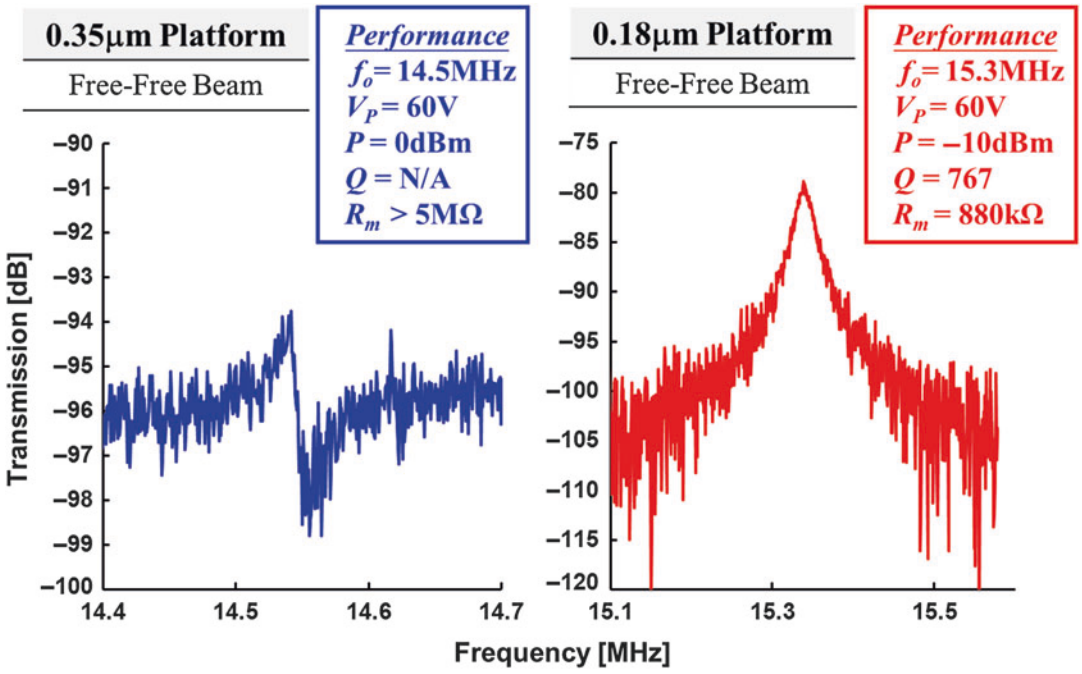
In addition to the use of the advanced technology node, a mechanically coupled array design [13] offers an efficient approach to reduce the motional impedance R_m which is inversely proportional to the number of the constituent resonators in an array device. The more resonators mechanically coupled as an array, the smaller the motional impedance. Since the coupled array



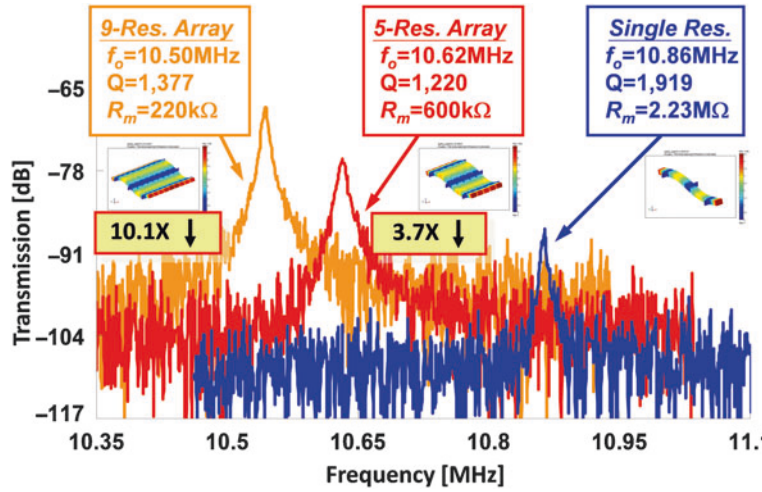
CMOS-MEMS Resonators, Fig. 10 CMOS-MEMS resonator using the metal removal post process in a 0.35 μm 2P4M CMOS-MEMS platform



CMOS-MEMS Resonators, Fig. 11 (**a**) SEM, (**b**) cross-sectional configuration, and (**c**) frequency responses of a CMOS-MEMS resonator fabricated in a 0.18 μm 1P6M technology under the metal removal post process

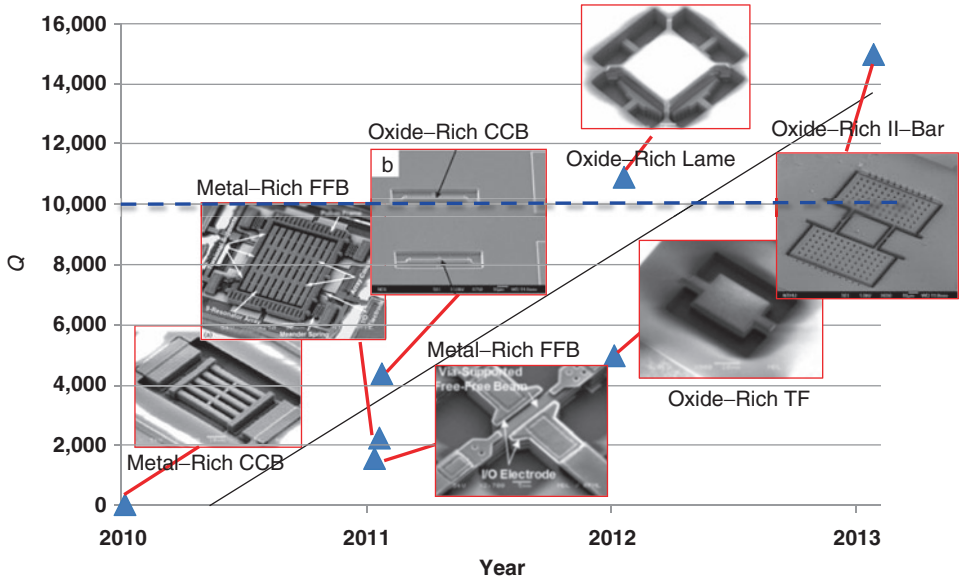


CMOS-MEMS Resonators, Fig. 12 Measured frequency characteristics of free-free beam resonators using the (**a**) 0.35 µm 2P4M and (**b**) 0.18 µm 1P6M CMOS-MEMS platforms



CMOS-MEMS Resonators, Fig. 13 Comparison of a single resonator and a nine-resonator array under the same DC bias and input power

becomes a multi-degree-of-freedom mechanical system, a high-velocity coupling scheme is used to accentuate the desired mode shape and reject the spurious modes. As shown in Fig. 13, the proposed nine-resonator and five-resonator arrays [13] have been experimentally characterized to have 10.1X and 3.7X smaller motional impedances, respectively, as compared to a single resonator. In addition, the resonator array benefits from the large transduction area to effectively reduce the required DC bias voltage while maintaining reasonable motional impedance.



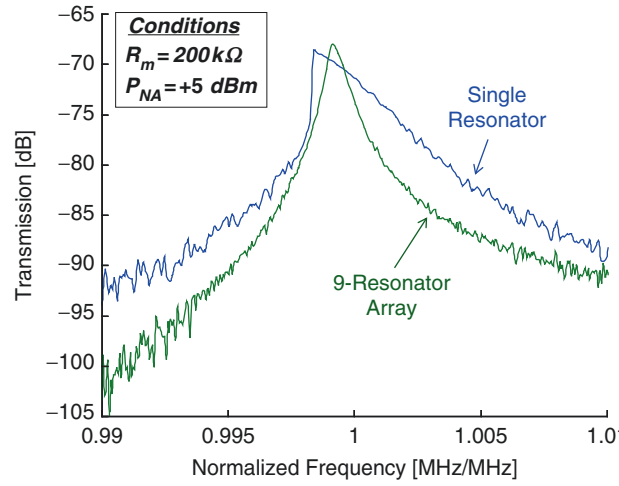
CMOS-MEMS Resonators, Fig. 14 Progress on quality factor of the CMOS-MEMS resonators developed in our group

Quality Factor *Q* Improvement

As compared to the single-crystal or polysilicon-based MEMS resonators, Q of the CMOS-MEMS resonators becomes a main issue to be used in timing reference and spectral processing applications. Due to the lossy nature of the metal materials in the CMOS back-end-of-line (BEOL), CMOS-MEMS metal-rich composite resonators often suffer Q value limited to 1,000. Several approaches are proposed to resolve this bottleneck, including (i) adopting proper mechanical boundary condition and implement nodal support design (i.e., support at the motionless locations of the resonator to avoid acoustic loss), such as the use of a free-free beam resonator with nodal supports instead of a clamped-clamped beam version; (ii) utilizing high-Q structural material in BEOL, such as SiO_2-rich composite structure in Fig. 10, to attain high Q; and (iii) taking advantage of high-Q bulk mode instead of low-Q flexural mode, as illustrated in Fig. 11. Figure 14 finally presents the progress on quality factor of our developed CMOS-MEMS resonators in the past 5 years. As can be seen, Q greater than 10,000 has been realized by the combination of the nodal support design, oxide-rich structural material, and bulk-mode operation, thus greatly benefiting the future oscillator and filter performance.

Power Handling P_o Enhancement

Upon high-input power levels, capacitive resonators often suffer the Duffing nonlinearity which resulted from both capacitive and mechanical nonlinearities, hence degrading output power and phase noise performance of oscillators implemented using those resonators. The mechanically coupled array design [13] described in section "Motional Impedance R_m Reduction" not only reduces their motional impedance R_m but greatly enhances the power handling capability P_o since the effective stiffness of the N-resonator array ideally is N times higher than that of a single resonator, thus contributing N times larger power handling ability. Figure 15 presents the frequency characteristics of single resonator and nine-resonator array designs under the same input power and motional impedance, where the former falls into the nonlinear regime, while the latter still preserves the linear vibration with a symmetrical frequency response and high Q.



CMOS-MEMS Resonators, Fig. 15 Comparison of the single resonator and nine-resonator array under the same motional impedance and input power

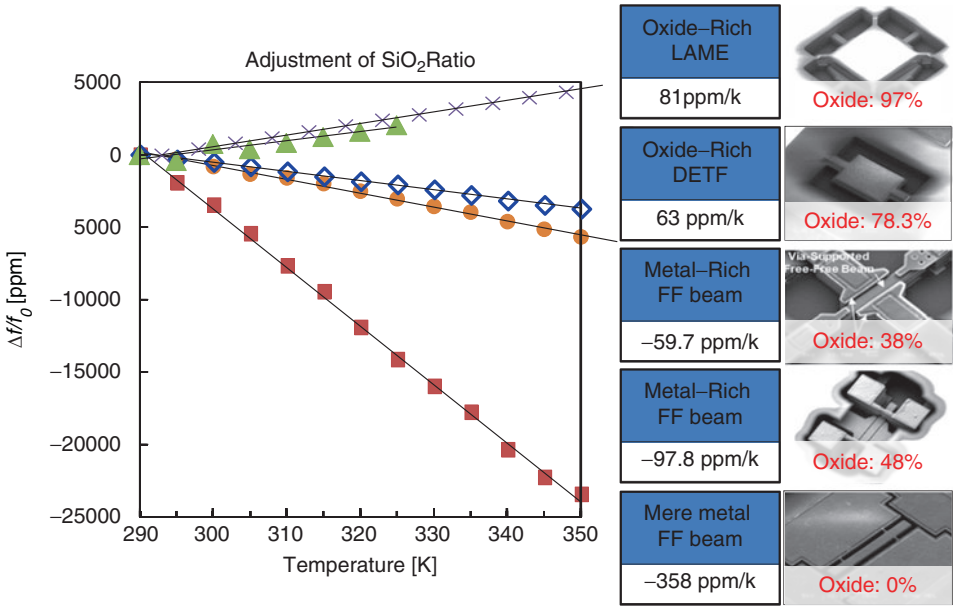
Temperature Compensation

The CMOS-MEMS resonators mostly composed of metals are sensitive to temperature variation due to the very negative temperature coefficients of Young's modulus (TC_Es) of their constituent metal materials. To improve the thermal stability of those CMOS-MEMS resonators, the silicon dioxide (SiO_2) of the CMOS BEOL possessing positive TC_E offers a simple and effective temperature compensation scheme where metal-oxide composite structures can be utilized to build resonators capable of improving their thermal stability [1]. For example, the oxide constituent ratio of the CMOS-MEMS resonators can be adjusted as indicated in Fig. 2 where the mere-metal, metal-rich, and oxide-rich structures would lead to very negative, slightly negative, and very positive temperature coefficients of frequency (TC_fs). Figure 16 presents the comparison of fractional frequency change versus temperature measurements for mere-metal, metal-rich composite, and oxide-rich composite beam resonators, indicating the near-zero TC_f is plausible once the ratio between SiO_2 and metal is optimized, such as manipulation of A and B in Eq. 10 to enable $TC_f \sim 0$. Thus, the use of SiO_2 as part of the composite structures in CMOS-MEMS resonators brings an easy and effective temperature compensation scheme [14].

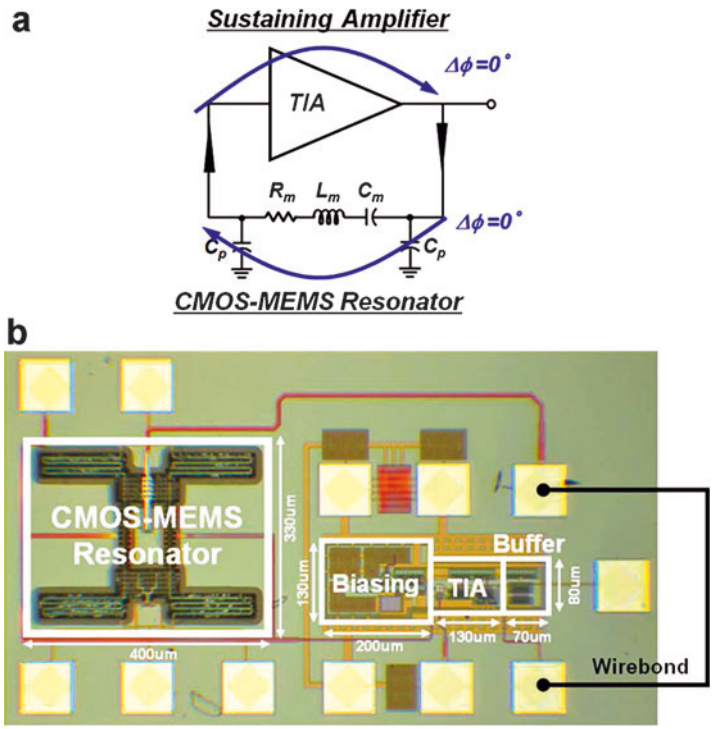
Examples of Application

Oscillator Implementation

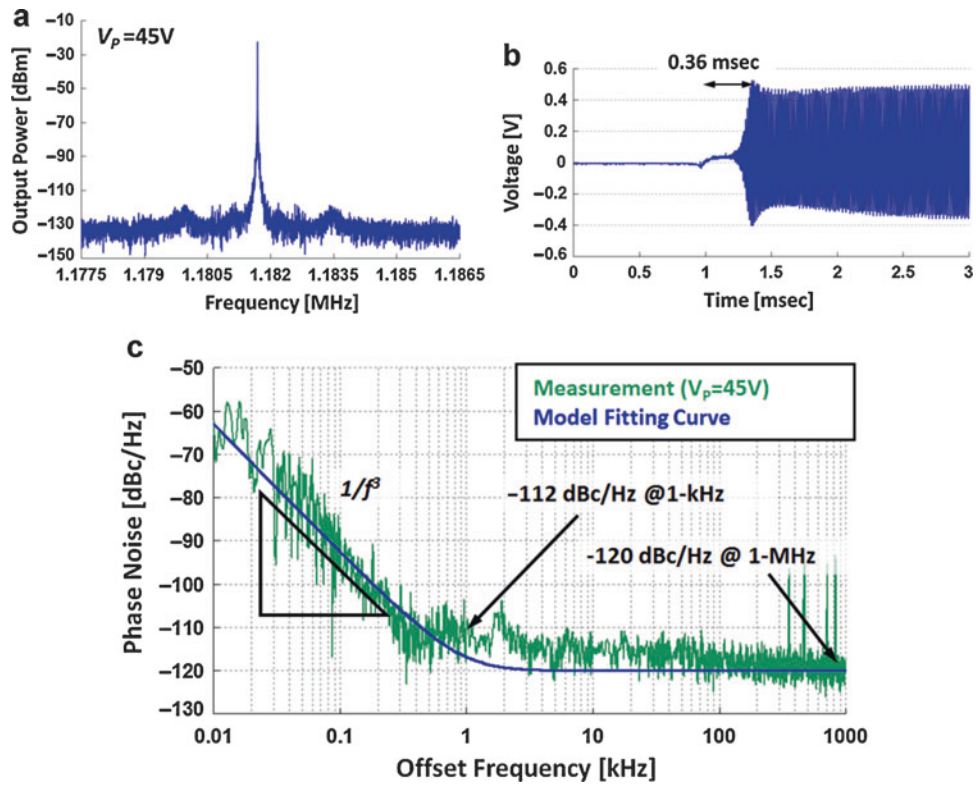
After addressing the major bottlenecks and their corresponding solutions to enhance the overall performance of the CMOS-MEMS resonators, the oscillator implementation is of great importance for timing, frequency synthesizing, and sensing applications. As depicted in Fig. 17a, a micromechanical resonator and its sustaining amplifier can form a closed loop to enable oscillation once the Barkhausen criteria (i.e., loop gain >1 and loop phase $= 0°$) are satisfied. Using a physical implementation shown in Fig. 17b as an example, a single-chip CMOS-MEMS oscillator has been successfully demonstrated in vacuum with the measured frequency-domain output spectrum and time-domain waveform shown in Fig. 18a, b, respectively. The measured phase noise performance in Fig. 18c is comparable to the silicon-based oscillators. Even in air (i.e., with



CMOS-MEMS Resonators, Fig. 16 Temperature coefficients of frequency can be adjusted by the constituent ratio of SiO_2 in CMOS-MEMS resonators



CMOS-MEMS Resonators, Fig. 17 (**a**) Top-level schematic and (**b**) optical view of the monolithic CMOS-MEMS oscillator

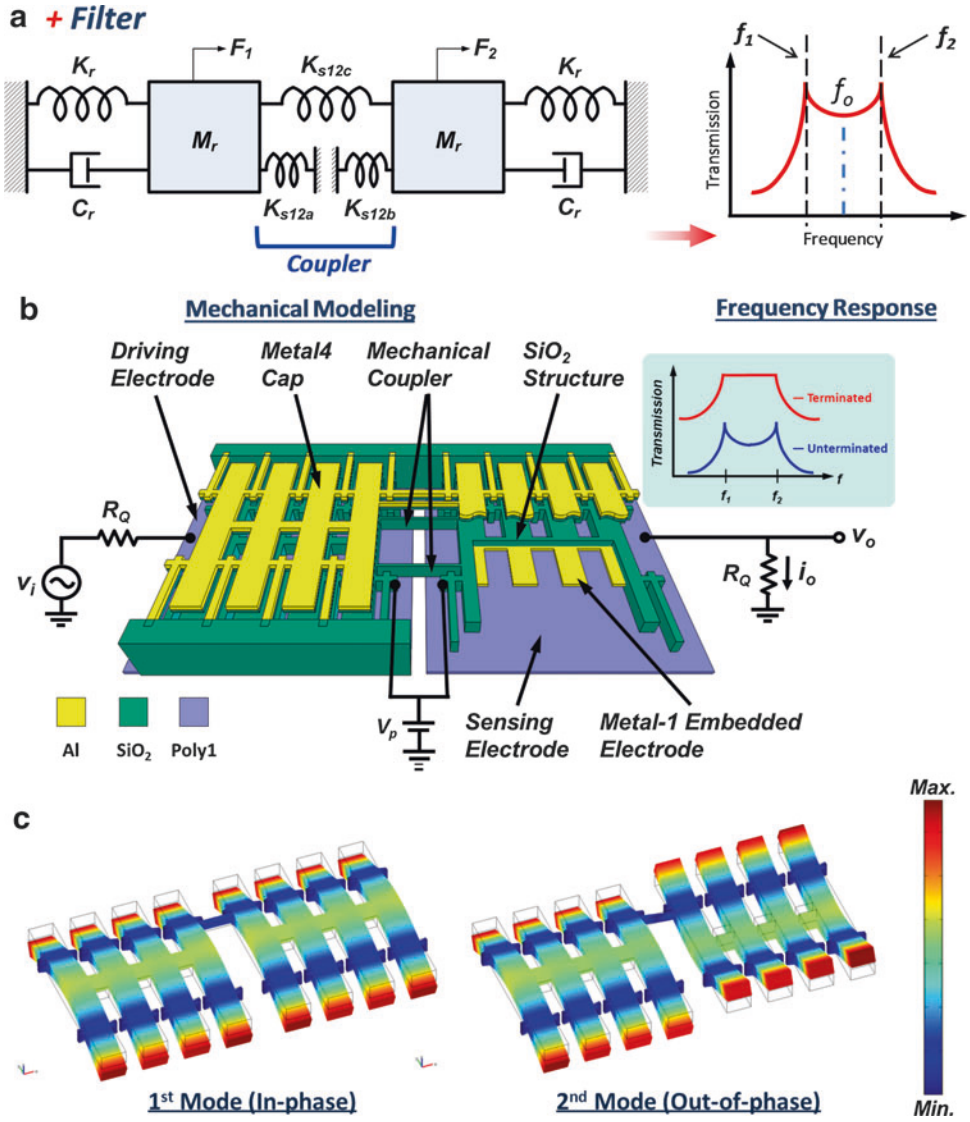


CMOS-MEMS Resonators, Fig. 18 (**a**) Frequency-domain response, (**b**) time-domain waveform, and (**c**) phase noise performance of the CMOS-MEMS oscillator

large squeeze film damping), the oscillator is still functional, indicating great potential to be used in gas, chemical, and mass sensing applications based on the frequency-shift mechanism. The abovementioned approaches to improve quality factor and power handling of the CMOS-MEMS resonators would eventually benefit the close-to-carrier and far-from-carrier phase noise performance, respectively, of the implemented oscillators. In addition, the reduction of the motional impedance described in the previous section would also lead to the overall phase noise reduction for the CMOS-MEMS integrated oscillators.

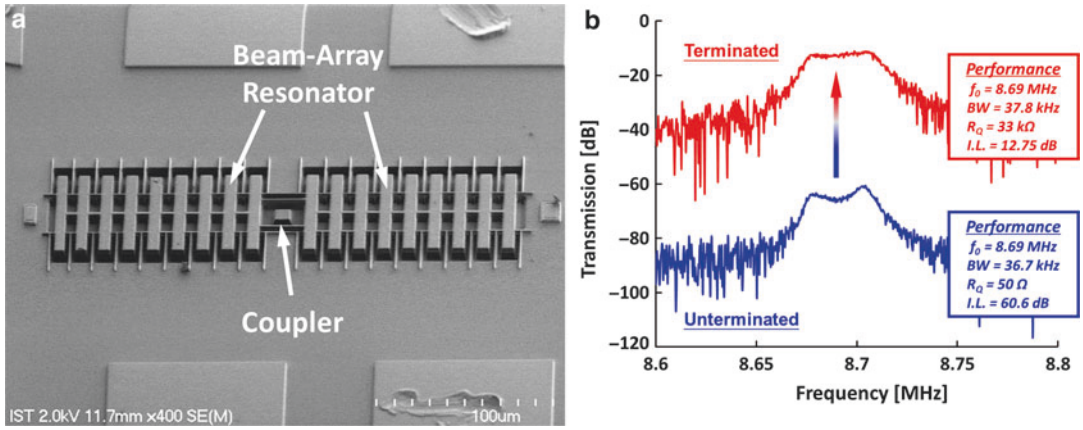
Filter Implementation

Another significant building block for frequency control and wireless communication is the frequency selection element. The developed CMOS-MEMS resonators can serve as the fundamental units for a band-pass filter. Since the resonance frequency of the CMOS-MEMS resonators is limited, it is best to be used in sensor front end with much lower operation frequencies. Figure 19a illustrates the concept of a fourth-order band-pass filter realized by two mechanical resonators, each of which is modeled by a mass-spring-damper system, coupled by a mechanical spring system to form a desired passband of the filter. The center frequency of the filter is determined by the resonance frequency of the constituent resonators, while the bandwidth is mainly set by the spring constant of the mechanical coupler. To simplify the filter design, the constituent resonators should be identical and the coupling element is designed in its quarter-wavelength configuration, allowing the center frequency of the filter to be set at the resonance frequency of



CMOS-MEMS Resonators, Fig. 19 (**a**) Equivalent mechanical circuit model, (**b**) perspective-view schematic, and (**c**) simulated filter modes of a mechanically coupled CMOS-MEMS band-pass filter

the resonators (f_o) while the frequencies (f_1 and f_2) of the two filter modes possess equal distance (i.e., separation) to the filter center frequency (f_o). Figure 19b presents the physical implementation of a CMOS-MEMS filter comprising two CMOS-MEMS resonator arrays mechanically coupled through a quarter-wavelength mechanical link to enable a narrow-bandwidth band-pass filter for channel selection [15]. The in-phase and out-of-phase filter modes shown in Fig. 19c form a passband of the filter. The fabrication process is similar to that of the CMOS-MEMS resonators in the previous section using the metal removal release process in a 0.35 μm CMOS technology [3], and the SEM view of the fabricated filter is shown in Fig. 20a. After filter termination, a flat passband and real insertion loss can be obtained as shown in Fig. 20b.



CMOS-MEMS Resonators, Fig. 20 (**a**) SEM view and (**b**) un-terminated and terminated frequency responses of a mechanically coupled CMOS-MEMS filter

Cross-references

► CMOS-MEMS

References

1. Chen, W.-C., Fang, W., Li, S.-S.: A generalized CMOS-MEMS platform for micromechanical resonators monolithically integrated with circuits. J. Micromech. Microeng. **21**(6), 065012 (2011)
2. Li, C.-S., Hou, L.-J., Li, S.-S.: Advanced CMOS-MEMS resonator platform. IEEE Electron Device Lett. **33**(2), 272–274 (2012)
3. Liu, Y.-C., Tsai, M.-H., Chen, W.-C., Li, M.-H., Li, S.-S., Fang, W.: Temperature-compensated CMOS-MEMS oxide resonators. IEEE/ASME J. Microelectromech. Syst. **22**(5), 1054–1065 (2013)
4. Chen, W.-C., Li, M.-H., Liu, Y.-C., Fang, W., Li, S.-S.: A fully-differential CMOS-MEMS DETF oxide resonator with Q > 4,800 and positive TCF. IEEE Electron Device Lett. **33**(5), 721–723 (2012)
5. Li, C.-S., Li, M.-H., Chin, C.-H., Li, S.-S.: Differentially piezoresistive sensing for CMOS-MEMS resonators. IEEE/ASME J. Microelectromech. Syst. **22**(6), 1361–1372 (2013)
6. Melamud, R., Chandorkar, S.A., Kim, B., Lee, H.K., Salvia, J.C., Bahl, G., Hopcroft, M.A., Kenny, T.W.: Temperature-insensitive composite micromechanical resonators. IEEE/ASME J. Microelectromech. Syst. **18**(6), 1409–1419 (2009)
7. Fedder, G.K., Santhanam, S., Reed, M.L., Eagle, S.C., Guillou, D.F., Lu, M.S.-C., Carley, L.R.: Laminated high-aspect-ratio microstructures in a conventional CMOS process. Sensors Actuators A **57**, 103–110 (1996)
8. Lo, C.-C., Chen, F., Fedder, G.K.: Integrated HF CMOS-MEMS square-frame resonators with on-chip electronics and electrothermal narrow gap mechanism. In: Technical Digest, Transducers'05, pp. 2074–2077. Seoul (2005)
9. Uranga, A., Teva, J., Verd, J., Lopez, J.L., Torres, F., Esteve, J., Abadal, G., Perez-Murano, F., Barniol, N.: Fully CMOS integrated low voltage 100 MHz MEMS resonator. IEEE Electron. Lett. **41**(24), 1327–1328 (2005)
10. Verd, J., Uranga, A., Teva, J., Lopez, J.L., Torres, F., Esteve, J., Abadal, G., Perez-Murano, F., Barniol, N.: Integrated CMOS-MEMS with on chip read-out electronics for high frequency applications. IEEE Electron Device Lett. **27**(6), 495–497 (2006)
11. Teva, J., Abadal, G., Uranga, A., Verd, J., Torres, F., Lopez, J.L., Esteve, J., Pérez-Murano, F., Barniol, N.: From VHF to UHF CMOS-MEMS monolithically integrated resonators. In: Technical Digest, 21st IEEE International Conference on Micro Electro Mechanical Systems (MEMS'08), pp. 82–85. Tucson. (2008)
12. Pourkamali, S., Hao, Z., Ayazi, F.: VHF single crystal silicon elliptic bulk-mode capacitive disk resonators, part II: implementation and characterization. IEEE/ASME J. Microelectromech. Syst. **13**(6), 1054–1062 (2004)
13. Li, M.-H., Chen, W.-C., Li, S.-S.: Mechanically-coupled CMOS-MEMS free-free beam resonator arrays with enhanced power handling capability. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **59**(3), 346–357 (2012)
14. Chen, W.-C., Fang, W., Li, S.-S.: High-Q integrated CMOS-MEMS resonators with deep-submicron gaps and quasi-linear frequency tuning. IEEE/ASME J. Microelectromech. Syst. **21**(3), 688–701 (2012)
15. Chen, C.-Y., Li, M.-H., Li, C.-S., Li, S.-S.: Design and characterization of mechanically-coupled CMOS-MEMS filters for channel-select applications. Sensors Actuators A Phys. **216**, 394–404 (2014)

CNT assembly

▶ CNT Handling and Integration

CNT Biosensor

▶ Nanostructure Field Effect Transistor Biosensors

CNT Handling and Integration

Zhan Yang
Jiangsu Key Laboratory for Advanced Robotics Technologies and Collaborative Innovation Center of Suzhou, Nano Science and Technology, Soochow University, 215021, Suzhou, China

Synonyms

CNT assembly; Nano Manipulation; Nanorobotics

Definition

Manipulation matter at nanoscale is one of the main purposes of nanotechnology. In with the publication of the pioneering work, "Engines of Creation" in 1986, atom-by-atom assembly became less a dream and more a search for a practical technology to implement it. Researchers were getting more adept at using Atomic Force and Scanning Tunneling Microscopes (STMs), culminating with the now famous writing of the IBM logo in 1989 with 35 xenon atoms. The team, led by Andreas Heinrich of IBM Research-Almaden (California), has successfully stored one magnetic bit of data with just 12 atoms of iron and a full byte of data in 96 atoms. This represents a storage density that is at least 100 times denser than the largest hard drive platters or flash memory chips in 2012.

In the recent two decades, nanorobotic manipulation was developing rapidly. Nanorobotic manipulation was applied to atomic force microscope, scanning electron microscope, transmission electron microscope towards to nanomanufacture nanoassembly and nanofabrication.

The AFM consists of a cantilever with a sharp tip (probe) at its end that is used to scan the specimen surface. The cantilever is typically silicon or silicon nitride with a tip radius of curvature on the order of nanometers. When the tip is brought into proximity of a sample surface, forces between the tip and the sample lead to a deflection of the cantilever according to Hooke's law. Depending on the situation, forces that are measured in AFM include mechanical contact force, van der Waals forces, capillary forces, chemical bonding, electrostatic forces, magnetic forces (see magnetic force microscope, MFM), Casimir forces, solvation forces, etc. Along with the force, additional quantities may simultaneously be measured through the use of specialized types of probes (see scanning thermal microscopy, scanning Joule expansion microscopy, photothermal microspectroscopy, etc.). Typically, the deflection is measured using a laser spot reflected from the top surface of the cantilever into an array of photodiodes. Other methods that are used include optical interferometry, capacitive sensing, or piezoresistive AFM cantilevers. These cantilevers are fabricated with piezoresistive elements that act as a strain gauge. Using a Wheatstone bridge, strain in the AFM cantilever due to deflection can be measured, but this method is not as sensitive as laser deflection or interferometry. Prof. N. Xi's team developed a real-time, force feedback nanomanipulation system.

Scanning electron microscope is a real-time nanoscale observation equipment. With the benefit of big specimen chamber, a complex nanomanipulation system with multiple-DOFs could be constructed. Recently, the environmental scanning electron microscope (ESEM) with nanorobotic manipulation was employed to characterize the biological property of cells.

Prof. Fukuda's team is a leader of nanorobotic manipulation in FESEM from the beginning of

this century. The carbon nanotube was picked-up, assembled, measured, and fabricated by multiple DOFs nanomanipulator.

Dr. M. Rizuan and Dr. Y. Shen constructed a nanomanipulation system in order to characterize the biological property of cells with low pressure and different humidity. Two types of experiments have been conducted, i.e., mechanical properties of individual yeast cells and forces on cell-substrate area. Different kinds of spring constants (0.02 N/m, 0.09 N/m, and 0.7 N/m) and tip's shapes (sharp, flat, and needle-like) of the cantilevers have been used during the experiments. From the analysis, the compressed forces needed to penetrate the cell wall of the yeast cells using sharp and flat tips (0.02 N/m for both) which are in the range of 87–278 nN and 57–207 nN, respectively. Locality mechanical measurement has been performed using needle-like tip (0.09 N/m) on single-cell and mother-daughter cells where the elastic properties of the cells are in the range of 1.32–3.95 MPa.

Nanorobotic manipulation is a rapid developing technology on nanofabrication and nanoassembly. It is powerful with nanomaterial measurement and manipulation. With the control theory, the automation is employed to the nanorobotic manipulation system. A high efficacy nanorobotic manipulation center will be applied to nanomaterial, nanosensor, and biological samples.

Introduction

This entry is focused on the study of CNT interaction with environment and the application of CNT for pH and temperature sensing.

CNT Handling

Carbon nanotubes (CNTs) [1] have been widely used as nanobuilding blocks to assemble nanostructures and nanodevices. Therefore, the interaction between a CNT and an environment is important. In the present study, the interaction forces, which consisted mainly of van der Waals [2] forces, between a single multiwalled carbon nanotube tip and a gold surface were evaluated under a scanning electron microscope by means of nanorobotic manipulation. The influence of electron beam irradiation was also investigated. It was found that electron beam irradiation can increase the interaction forces by producing contamination at the contact area. The van der Waals forces were calculated theoretically, and the theoretical values were close to the experimental values. In the following, the measurement of van der Waals force between a CNT tip and a gold surface is introduced first, and then a method of manipulating CNTs in 3D is presented.

Measurement of van der Waals Force Between a CNT Tip and Gold Surface

At the nanoscale, van der Waals force is one of the predominant surface forces that also include the electrostatic force, the Casimir force, etc. The measurement of nanoscale surface forces can help characterize the properties of nanomaterials and also be useful for the assembly of nanostructures and nanodevices. Carbon nanotubes (CNTs) as typical nanomaterials have been used for atomic force microscope (AFM) probes, nanotweezers, nanoposition sensors, and three-dimensional nanostructures. The interaction between a CNT and a substrate, including the van der Waals forces, is important for the assembly of these mechatronic applications. The deformation of a CNT placed on a substrate by surface van der Waals forces will significantly modify the idealized geometry of the free nanotube. The interaction has been investigated by AFM and molecular-mechanics simulations. The joint interaction forces with two CNT tips by chemical bonds were also investigated. Cumings and Zettl calculated the attractive van der Waals forces between the inner and outer layer. They demonstrated a lowfrictional nanoscale linear bearing realized from multiwalled carbon nanotubes (MWNTs). Zheng et al. showed through theoretical calculations that the extruded core of a MWNT oscillating on the outer shell caused the excess van der Waals interaction energy and

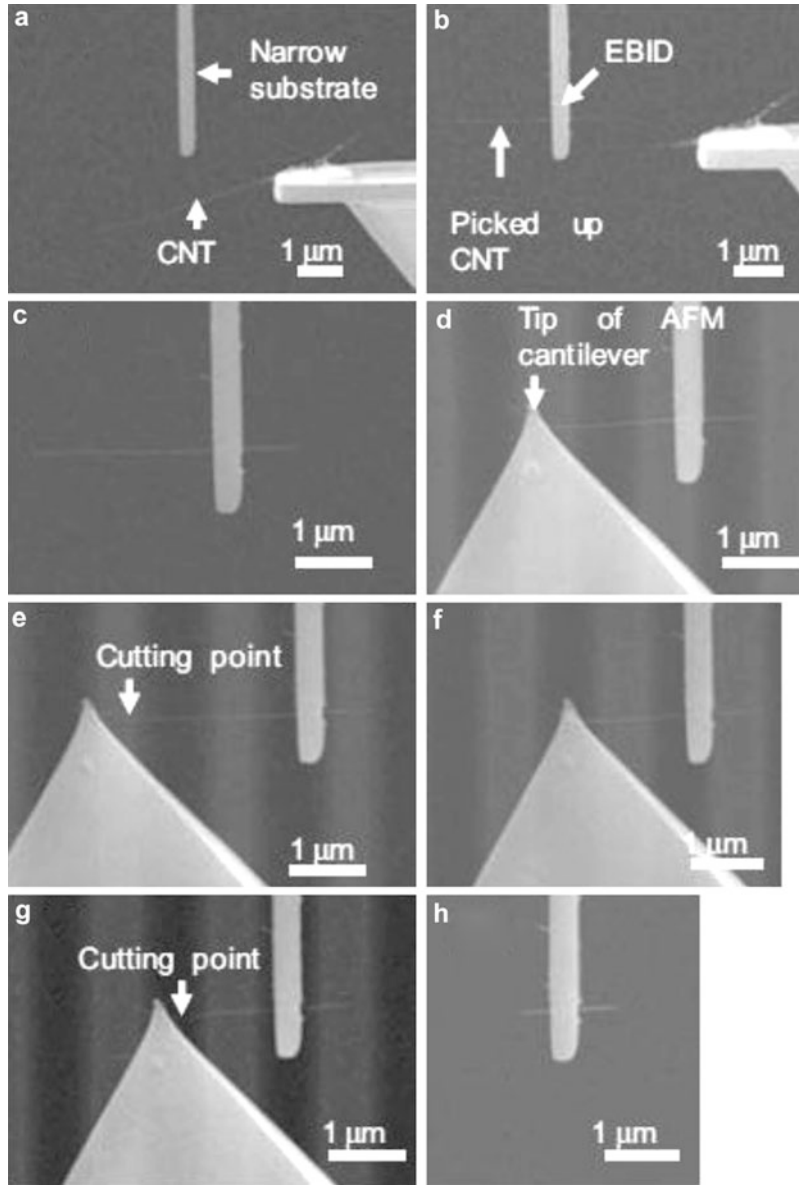
found that the oscillation frequency can be significantly higher than 1 GHz [3]. The effects of surface forces in a three-terminal nanorelay were investigated by Jonsson et al. [4]. They showed that van der Waals forces have a significant impact on the characteristics of the relay. For the calculation or analysis of van der Waals forces, a parameter called the Hamaker constant is an important value.

In previous studies, the Hamaker constant between a CNT and a surface was obtained by approximate calculation [5] or by original experiments [6]. Akita et al. obtained the Hamaker constant of 60×10^{-20} J for the sidewall of a CNT adhered on a silicon material surface [6]. For a CNT and a gold surface, Walkeajärvi et al. just used a typical value of 10×10^{-20} J for their calculation. For two CNTs, a previous report used the value of graphite to replace it. The van der Waals forces between a CNT tip and a gold surface are not clearly understood experimentally and need to be investigated. In particular, the interaction force is considered to be changed by irradiation of the electron beam during the observation in an electron microscope, because electron beam irradiation produces contamination on the observation area.

From an engineering viewpoint, it is important to control the interaction forces between CNT probes and metallic surfaces for nanodevice fabrication and assembly. This entry reports on the measurement of the van der Waals forces between a CNT tip and a gold surface under a scanning electron microscope (SEM) based on nanorobotic manipulation. The influence of electron beam irradiation was also investigated. The van der Waals forces were calculated theoretically and compared with the experimental values.

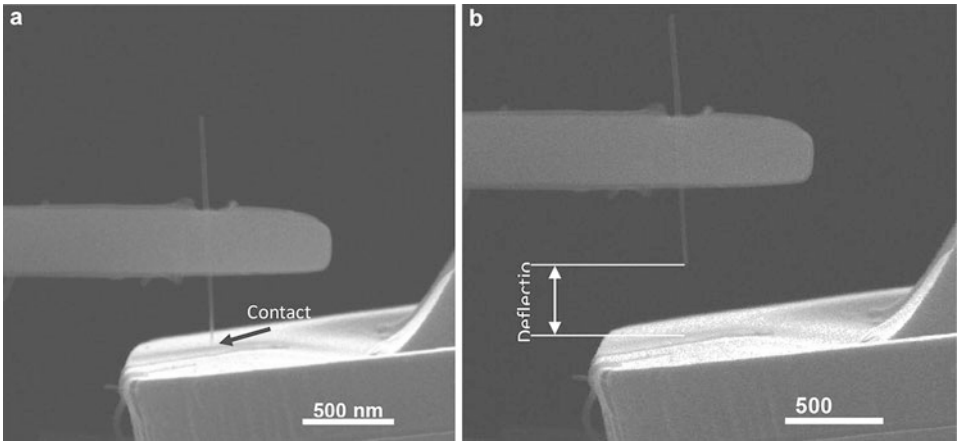
For the measurement of the interaction forces between the CNT tip and the surface of a gold substrate, a single MWNT probe was first prepared as shown in Fig. 1. The MWNT, typically 20–50 nm in diameter, was synthesized by the standard arc-discharge method. Figure 1a shows a single MWNT with diameter of 30 nm that was picked up from the CNT bundle by a nanorobotic manipulator under a SEM (JEOL, JSM-6500 F). The MWNT was fixed at the narrow substrate with electron beam–induced deposition (EBID) as shown in Fig. 1b. The deposition is caused by the dissociation of molecules adsorbed to a surface by high-energy electrons. In the present experiments, tungsten hexacarbonyl was used as precursor by filling into a glass tube and introducing it into the vicinity of the sample. The magnified image of the narrow substrate with MWNT is shown in Fig. 1c. It was etched to a width of 350 nm by a focused ion beam (FIB) process. The MWNT probe should be prepared with no impurities and straight with short length, because the bending phenomenon occurs easily and makes it difficult to measure the vertical interaction forces. Some methods such as electron beam–induced fabrication assisted with oxygen gas or current-induced fabrication can be used to adjust the CNT length. Here, the current-induced fabrication process was used. One end of the CNT was touched at the tip of the AFM cantilever as shown in Fig. 1d. The part of the CNT near the cantilever tip evaporated by joule heating when the electric current applied was 7–20 mA. The length of the MWNT probe was adjusted by using the same process as shown in Fig. 1d, e by the first process. Figure 1f, g show the second process. Finally, the MWNT probe is fabricated as shown in Fig. 1h.

In the present work, the interaction force measurement between the CNT probe and the gold surface of an AFM cantilever was performed experimentally. The AFM cantilever was coated by a gold layer with thickness of 40 nm. The interaction forces were measured from the deflection of the AFM cantilever. The CNT probe and AFM cantilever were connected to the ground to avoid the occurrence of electrostatic forces. The SEM was kept at low magnification (less than 10,000 times) during the operation to avoid contamination. Other type forces caused by chemical bonding can be negligible, because the CNT probe and the surface were kept clean under vacuum conditions. Therefore, the interaction force between the CNT tip and the surface of the AFM cantilever is considered to consist mainly of van der Waals forces. In the experiment, the AFM cantilever tip was moved to the

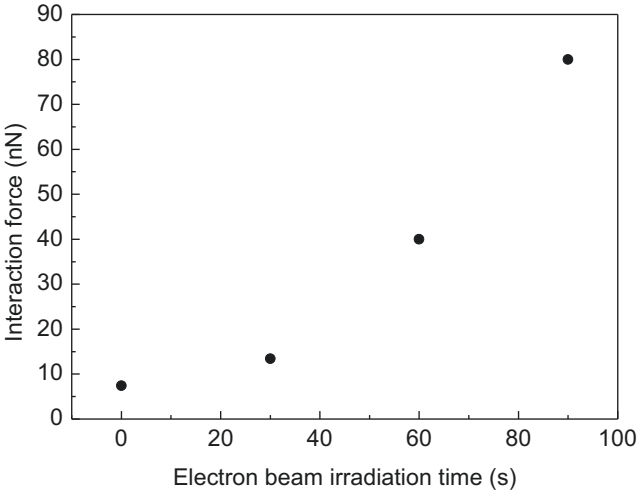


CNT Handling and Integration, Fig. 1 A multiwalled carbon nanotube was placed on an etched substrate. The length of the CNT was adjusted by the current-induced fabrication process

CNT tip by a nanorobotic manipulator, until they touched each other (Fig. 2a). Then the CNT probe was moved in the inverted direction until release of the adhesion by van der Waals forces. The position of the AFM cantilever was controlled in step motions, each step of moving was about 2 nm. When the CNT probe was released from the AFM cantilever surface, the AFM cantilever was stopped in a certain distance. The AFM cantilever was released where the attraction is less than the tension caused by the deformation of the cantilever. The deflection of the AFM cantilever was 370 nm in the present experiment (Fig. 2b). From this, the

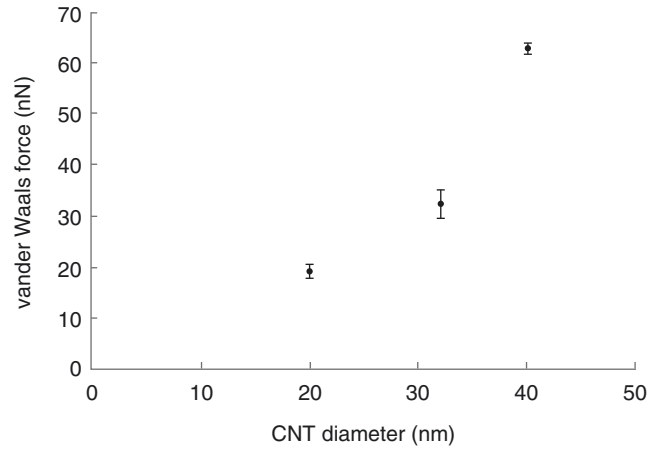


CNT Handling and Integration, Fig. 2 The tip of a single CNT was adhered to an AFM cantilever surface. After that, the AFM cantilever was pulled back by a nanomanipulator. The interaction force between the carbon nanotube tip and the surface can be obtained from the deflection of cantilever



CNT Handling and Integration, Fig. 3 The relationship between the interaction force and electron beam irradiation time

interaction forces from the deflection of the AFM cantilever were calculated according to $F = \delta d$, where d is the deflection and is the elastic constant of the AFM cantilever, is 0.02 N/m (Olympus Corp., OMCL-TR400). Therefore, the interaction force between the CNT tip and the gold surface was calculated as 7.4 nN. The interaction force measurement was also performed under electron beam irradiation to affect the influence of interaction forces. The interaction forces were measured after irradiation for different times. After the above experiment to measure the van der Waals interaction forces, the contact area, where the CNT tip was adhered on the gold surface, was magnified to 30,000 times. The acceleration voltage of the SEM was 15 kV and the beam current was 30 pA. The irradiation periods were set at 30, 60, and 90 s. The interaction forces were obtained as 13.4, 40, and 80 nN, respectively, as shown in Fig. 3. The results show that electron beam



CNT Handling and Integration, Fig. 4 Three teams experiment results with error bar of van der Waals force

irradiation can increase the interaction forces between the CNT tip and the surface. The increment of interaction force is almost proportional to the electron beam irradiation time period.

The increase of interaction force induced by electron beam irradiation is considered to result from the contamination produced at the contact area. In the experiment, the precursors come from the vacuum pump oils in the sample chamber; hence the contamination should be mainly amorphous carbon. This suggests that electron beam irradiation can be used to control the interaction forces in the nanonewton range for the assembly of CNT nanobuilding blocks.

We assume that the interaction between CNT tip and the surface is nonretarded and additive. The interaction energy can be described by Lennard-Jones's pair potential as the following equation

$$W = \frac{r^2\pi^2 C\rho_1\rho_2}{12D^2} = \frac{r^2 A}{12D^2} \tag{1}$$

where r is the radius of CNT, and C is the coefficient in the atom-atom pair potential. ρ_1 and ρ_2 are the numbers of atoms per unit volume in the two materials, and D the distance between the CNT tip and the surface, as shown in Fig. 5. A is the Hamaker constant.

Then the van der Waals forces are obtained by taking the derivative of the pair potential with respect to the distance,

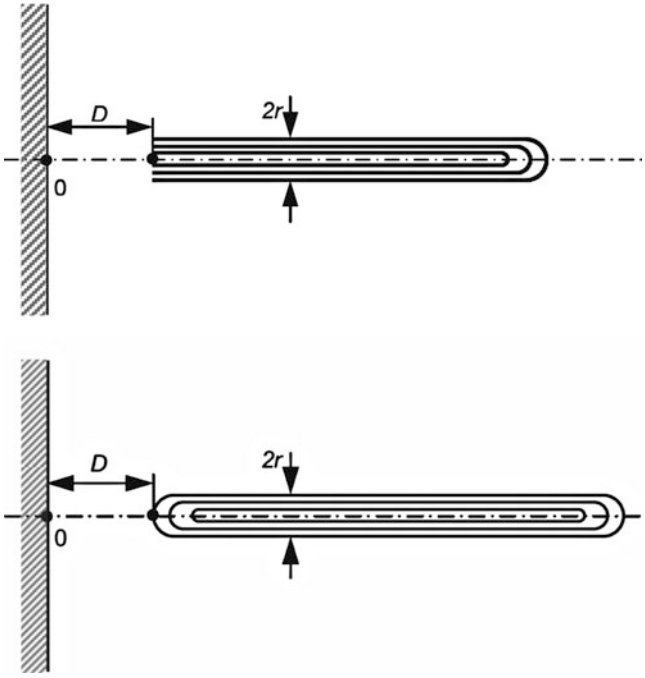
$$F = W' = \frac{r^2 A}{6D^3} \tag{2}$$

To calculate F, the Hamarker constant $A_{\mathrm{CNT}-Au}$ between the CNT and the gold surface interacting across vacuum is needed. The unknown $A_{\mathrm{CNT}-Au}$ can be obtained approximately by using the combining relation as the nonretarded Hamaker constant for carbon nanotube and gold surface [3], which is given by

$$A_{\mathrm{CNT}-Au} = \sqrt{A_{\mathrm{CNT}} A_{Au}}\,, \tag{3}$$

where A_{Au} and A_{CNT} are $5.47 \times 10^{-19} J$ [2] and $0.284 \times 10^{-19} J$ [5] from previous report, respectively. Thus, the $A_{\mathrm{CNT}-Au}$ is obtained as $1.246 \times 10^{-19} J$ from Eq. 3.

If the CNT is considered as a column, the van der Waals forces between the CNT tip and the gold surface are calculated as 77, 197, and 308 nN at the CNT diameters 20, 32, and 40 nm as shown in Fig. 4. It is simply assumed that the minimum distance D is about 0.3 nm, and the r is used as the radius of the CNT in Fig. 5 The calculated result is so much larger than the value of experimental result. Therefore, this mode which considers the CNT as a column cannot precisely get the van der Waals forces.



CNT Handling and Integration, Fig. 5 Schematic drawing of a CNT adhered to a surface with a tiny separation D. The CNT is $2r$ in diameter with open cap and closed cap

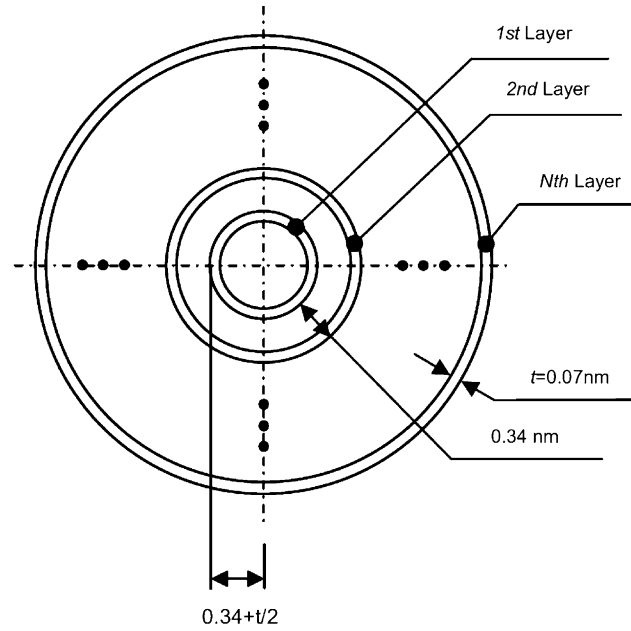
We suggest that the model of CNT can be considered as a multilayered cylinder as shown in Fig. 6. The cross-sectional area of the CNT is less than that calculated with the column model. Each layered cross-sectional area is described by following equations,

$$\begin{aligned} A_1 &= \pi(0.34 + t/2)^2 - \pi(0.34 - t/2)^2 \\ A_2 &= \pi(0.34 + t/2 + 0.34)^2 - \pi(0.34 - t/2 + 0.34)^2 \\ &\vdots \\ A_N &= \pi[0.34 + t/2 + 0.34(N-1)]^2 - \pi[0.34 - t/2 + 0.34(N-1)]^2 \end{aligned} \tag{4}$$

And the total cross-sectional area is,

$$A = \sum_{n=0}^{n=N-1} \left[\pi(0.34 + t/2 + 0.34n)^2 - \pi(0.34 - t/2 + 0.34n)^2\right] \tag{5}$$

Here, t is 0.07 nm, the diameter of carbon molecular. Then the total area is calculated as 148 nm^2. The distance between two layers is 0.34 nm that has clearly observed by a TEM [7]. N is the number of layers. The relationship between the number of layers and the radius of CNT can be approximated by $0.34N \approx r$ in our case. We obtained that the layer's number of the CNT is about 29, 47, and 59 at the CNT diameters at 20, 32, and 40 nm. Finally, the van der Waals forces were obtained theoretically as 15.9, 41.3, and 64.8 nN at the CNT diameters at 20, 32, and 40 nm from the Eq. 2. This result is very close to the experimental data. It demonstrates that the cylinder model is good or accurate enough for calculation of van der Waals force between CNTs and the surface. For more precise calculation, the CNT should be considered as a unit of carbon atoms. In that case, the sum of the forces between all individual atoms and the substrate needs to be calculated. Therefore, the shapes of CNT tip should be also taken into more careful consideration. Those are our future works.



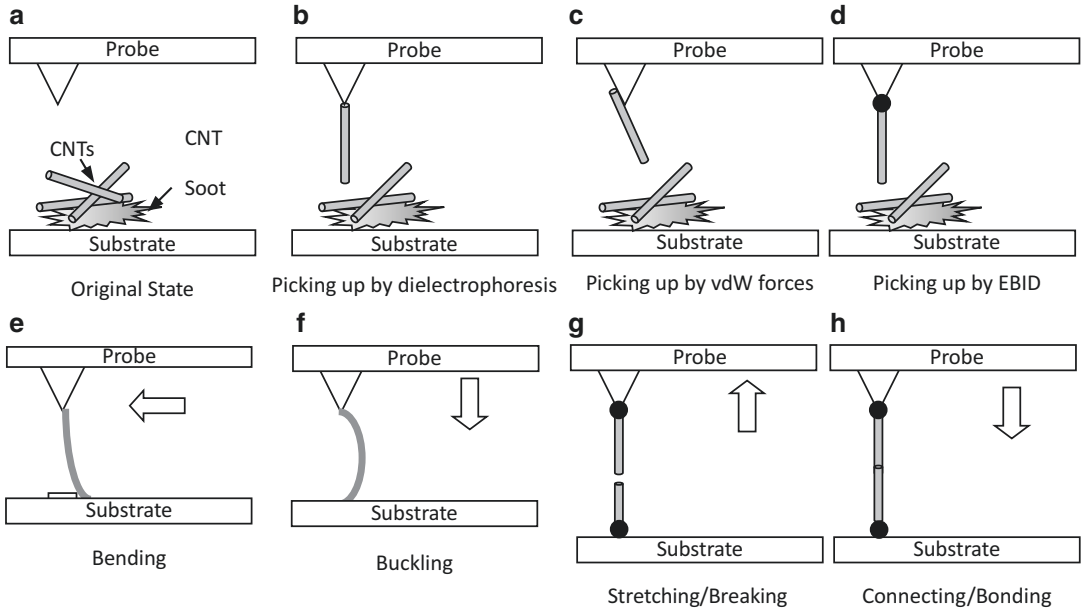
CNT Handling and Integration, Fig. 6 Schematic drawing of a CNT described by multilayered cylinder model. The distance between layers of CNT is 0.34 nm. The t is the diameter of carbon molecule with 0.07 nm

In summary, a single MWNT was actually prepared with the nanorobotic manipulation, electron beam–induced deposition, and current-induced fabrication. The interaction forces between the MWNT tip and the gold surface were investigated and measured with an AFM cantilever, which were mainly consisted of van der Waals forces. The method proposed is easy to operate and have a large measurement range. It can be used for evaluation of connection strength during assembling of nanodevices. We also found that the electron beam irradiation can increase the interaction forces. The increase of interaction force induced by electron beam irradiation might be produced by the contamination. We suggest that the electron beam irradiation should be used to adjust the interaction forces in nanonewton scale during the assembly of CNT.

For calculating the van der Waals forces, the CNT was described by a multilayered cylinder model. The result of the experiment is very close to theoretical calculation. That model can be used to estimate the nanometer order van der Waals forces.

3D CNT Handling

Manipulations of CNTs in 3D space are very important techniques for assembling them into structures and devices. The basic step for this is to pick up a single tube from nanotube soot (Fig. 7a). This has been shown first by using dielectrophoresis through nanorobotic manipulations (Fig. 7b). The interaction between a tube and the atomic flat surface of AFM cantilever tip has been shown to be strong enough for picking up a tube onto the tip (Fig. 7c). By using electron beam–induced deposition (EBID), it is possible to pick up a nanotube onto a probe (Fig. 7d). For handling a tube, weak connection between the tube and the probe is desired. Bending and buckling a CNT as shown in Fig. 7e, f are important for the in-situ property characterizations of a nanotube, which is a simple way to get the Young's modulus of a nanotube without damaging the tube if performed in elastic range and hence can be used for the selection of a tube with desired properties. Plastic bending or buckling can generate intramolecular kink junctions of CNTs. Combined bending and buckling with shape fixing with EBID can be used for the shape modifications of a nanotube. Stretching a nanotube between two probes or a probe and a substrate has brought out several interesting results. The first demonstration of 3D nanomanipulations of nanotubes took this one as an example to show



CNT Handling and Integration, Fig. 7 3D manipulations of CNTs. Basic technique for such manipulations is to pick up an individual tube from CNT soot (**a**) or oriented array. (**b**) shows a free-standing nanotube is picked up by dielectrophoresis generated by nonuniform electric field between the probe and substrate, (**c**) and (**d**) show the same manipulation by contacting a tube with the probe surface or fixing (e.g., with EBID) a tube to the tip. Vertical manipulations of nanotubes also include bending (**e**), buckling (**f**), stretching/breaking (**g**), and connecting/bonding (**h**). This family is open for new strategies

the breaking mechanism of a MWNT and the tensile strength of CNTs. By breaking a MWNT in a controlled manner, some interesting nanodevices have been fabricated. This technique – destructive fabrication – has been presented to get sharpened and layered structures of nanotubes and to improve the length control of a nanotube. Bearing motion has also been observed in an incompletely broken MWNT. The reverse process, namely the connection of broken tubes, has been demonstrated recently, and the mechanism is revealed as rebonding of unclosed dangling bonds at the ends of a broken tube. Based on this interesting phenomenon, mechanochemical nanorobotic manipulations have been presented.

3D nanorobotic manipulations have opened a new route for the assembly of nanotubes into nanodevices.

In the following, we will focus on our recent advances in this field.

CNT Integration

A CNT integration method by mechanical engineering is nanorobotic manipulation [8–10] which has advantages such as real-time controllability, nanometer lever positioning resolution, and three-dimensional fabrications. The automation nanofabrication has a potential for auto nanostructure assembly fabrication and measurement. Nanomanipulation allows for the detection and manipulation of tiny entities such as single molecules, nanotubes, cells, viruses, proteins, and DNA molecules.

Fabrication of a pH Sensor Nanoprobe with Tungsten Oxide and Platinum Nanowires Based on Nanorobotic Manipulation

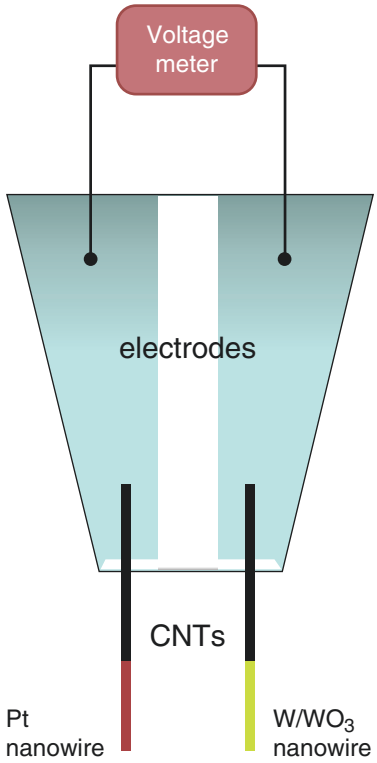
A novel solid-state pH sensor nanoprobe based on a tungsten oxide nanowire work electrode and a platinum nanowire counter electrode is reported. A cantilever is used for the electrodes

of the device and etched by a focused ion beam (FIB). Two multiple-walled carbon nanotubes (MWCNTs) are assembled to the separated electrodes of the cantilever. A tungsten probe is etched by an FIB to 300 nm in diameter and 25.4 μm in length. Then the probe is coated by parylene and the tip cut to expose the tungsten. A tungsten nanowire is used for the work electrode, and a platinum nanowire is used for the counter electrode. The nanowires are fabricated via the field emission method from carbon nanotube tips by introducing hexacarbonyltungsten ($W(CO)_6$) and trimethylcyclopentadienylplatinum ($CpPtMe_3$), respectively, inside a field emission electron microscope. The nanoprobe pH sensor is tested in acidic (pH 4.0, pH 6.1, and pH 6.a9) and alkaline (pH 8.0) buffer solutions and different responses are found. The sensitivity of the pH sensor nanoprobe was measured as −35.04 mv/pH.

Chemical detection of single cells and their local environment has attracted much interest recently. The pH value is a promising parameter to describe the living condition of single cell. For examples in animal cells, pH controls the activity of key enzymes, protein synthesis, DNA and RNA synthesis, cell cycle, and ion conductance by membranes. The extracellular fluid pH is usually $7.35 \sim 7.45$, and in vascular smooth muscle cell (VSMC) the pH value is usually $7.0 \sim 7.2$. In some pathological conditions, such as hypoxia, acidosis, or alkalosis, pH can be fluctuated between 6.5 and 8.1, of VSMC pH corresponding generate phase when large fluctuations. And the oxygen changes primer pH to vascular effects of very complex, it may be by changes in intracellular Ca ion release and re-uptake, the activity of calcium channels on the plasma membrane or to adjust the sensitivity of the calcium ions in the contraction protein vasoconstriction state. The pH sensors towards biological cell require a resolution of 1 pH. The measurement of the pH value of a single cell is usually performed by visualizing the color change of indicators, chemical probes, H^+ sensitive microelectrodes, and nuclear magnetic resonance (NMR) spectrometry.

Traditionally, the measuring methods for pH values fall roughly into four categories: indicator reagents, pH test strips, metal electrode methods (hydrogen electrode, quinhydrone electrode, and antimony electrode method), and glass electrode methods. In recent decades, new techniques of pH detection have been developed such as optical-fiber-based pH sensors [11], mass-sensitive pH sensors [12], metal oxide pH sensors [13], conducting polymer pH sensors [14], nanoconstructed cantilever-based pH sensors [15], ISFET-based pH sensors [16], and pH-image sensors [17].

To detect the pH distribution for cytoplasm and understand the relation between pH and enzymes and protein in detail within a single cell, which has a scale of several to several tens of micrometers, many nanoscale pH sensors have been developed. Zhang et al. reported a solid-state pH sensor based on WO_3-modified vertically aligned MWCNTs, which use a WO_3-CNT array for the pH detection electrode. This sensor has sensitivity close to −40 mV/pH [18]. Fenster et al. reported a single tungsten nanowire pH- sensitive electrode, which has a sensitivity of −61.1 mV/pH [19]. The previous two pH sensors contain nanoscale tungsten pH sensitive electrodes and macroscale stranded hydrogen electrodes which are impossible to apply to a single cell. We have designed double nanometal electrodes which can avoid the size problem in cell pH detection (Fig. 8). We aimed to fabricate the pH sensor by using electrodes by which sensing of the proton concentration becomes possible if either the electrochemical equilibrium involves protons in the reaction or if a separation of the electrolyte is achieved by a proton-specific ionophore. The probe was fabricated by the platinum and tungsten oxide nanowire electrodes grown via field emission technique. Because of the scale of the nanowires, nanoscale resolution positioning and manipulation was required. To this end a nanorobotic manipulation system was employed. Figure 8 shows the concept of the proposed device. In this research, a platinum and tungsten dual nanowire electrode pH sensor has been fabricated. A nanoprobe pH sensor evaluation



CNT Handling and Integration, Fig. 8 Concept draws of nanowire probe pH sensor

system has been constructed, and the pH response of this sensor has been tested in acidic buffer solution (pH = 4, 6.1, 6.9) and alkaline buffer solution (pH = 8).

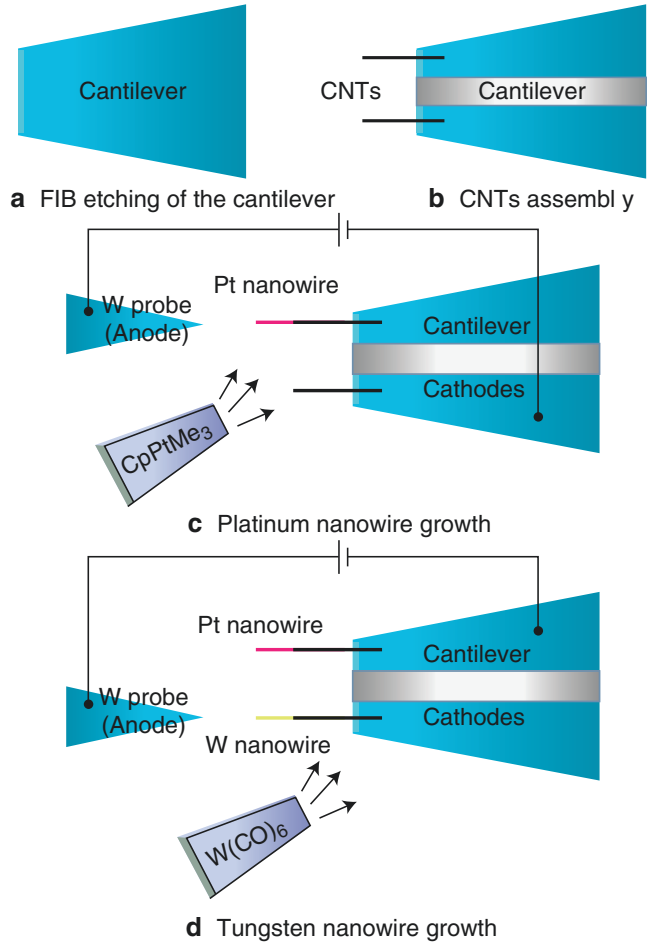
During the field emission nanowire growth, an MWCNT is employed for the cathode and a tungsten probe for the anode. Unit 1 and unit 4 were used to assemble the MWCNT emitters and set the tungsten probes as the anode.

The tungsten probe is etched by an FIB and then coated with parylene. Then the tip of the tungsten probe is opened using the FIB. The tungsten was set to manipulator 4 for anode. The carbon nanotubes were picked up by an AFM cantilever. An AFM cantilever (Olympus, OMCL-TR400PB-1) is set on the manipulator as the end-effector to pick up the MWCNT from the carbon bulk of MWCNTs which were synthesized by the arc-discharge method. The MWCNT is set up as the emitter (cathode) which has a low field emission voltage.

The fabrication procedure is shown in Fig. 9. A PZT cantilever was used as the substrate. The cantilever is etched by an FIB for separated electrodes (a). An FESEM was introduced for real-time observation and two units of manipulators inside the specimen chamber of FESEM were used to fabricate the device. An AFM cantilever was assembled to the manipulator to pick up CNTs from the carbon bulk. The CNTs were cut to 1.7 μm length by current, and fabricated to PZT cantilever (b). A tungsten probe was used for the anode in field emission. The probe was etched by an FIB and then coated with parylene. The FIB was then used to cut the tip of the probe to expose the conductive material to concentrate the electro field. The etched tungsten probe was positioned to the counter of the CNTs. For field emission of CNTs, the emission characteristics should be measured before growth. The trimethylcyclopentadienylplatinum precursor was inserted into the specimen chamber of the FESEM and a constant current applied to CNT 1 and etched tungsten probe to grow the platinum nanowire (c). Using the precursor hexacarbonyl tungsten and applying a constant current to the etched tungsten probe the CNT 2 resulted in the growth of the tungsten nanowire (d).

A tungsten nanoprobe (MicroManipulator Inc, MODEL 7B) was used as the anode for nanowire growth via field emission. To decrease the branching of the nanowire, the applied field needs to be concentrated. The cathode for the field emission is an MWCNT which is tens of micrometeres in diameter. The tip radius of the tungsten probe is nearly 1 μm as shown in Fig. 10a and needs to be reduced. The FIB technique is used to etch the tungsten tip to a 300 nm thick plane which is shown in Fig. 10b. The probe is then turned clockwise through 90°, and the plane etched to form a probe 300 × 300 nm in cross section (Fig. 10c). A parylene coating unit (Dix-coating, DACX-Lab) was employed to coat a 300–400 nm thickness of the polymer (Fig. 10d). The tip of the probe was etched by an FIB to remove the polymer as shown in Fig. 10e.

A dynamic force microscope (DFM) cantilever (KEYENCE OP-75041 DFM/SS) was used for the substrate CNTs. The cantilever comprises

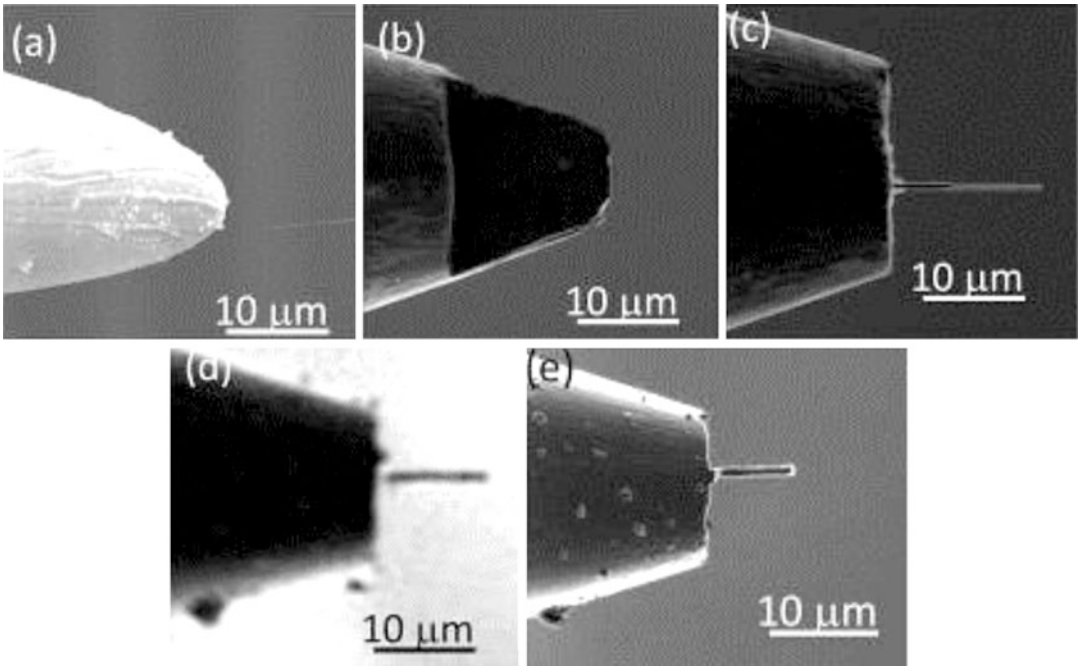


CNT Handling and Integration, Fig. 9 Concept draws of nanowire probe pH sensor

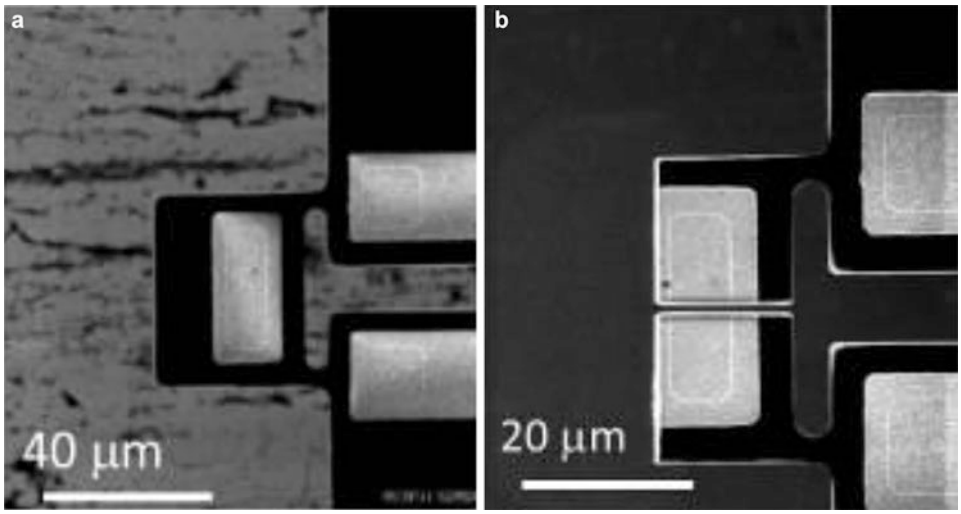
three layers. The bottom layer and the upper layer are insulators and the middle layer is alumina. We cut the edge of the cantilever and then separated the cantilever with FIB etching. The gap of the electrodes is 700 nm. Figure 11a, b, respectively, shows the cantilever before and after etching.

The CNTs were picked up from the carbon bulk by the AFM cantilever controlled by the nanorobotic manipulator. As shown in Fig. 12, the CNTs were assembled to electrodes of the PZT cantilever individually. The CNT 1 is 1.75 μm in length by 20 nm in diameter, and the CNT 2 is 1.78 μm in length by 31 nm in diameter as shown in Fig. 13a. Figure 13a, b shows the platinum nanowire growth on the tip of CNT 1. The platinum nanowire was grown from the MWCNT emitter by using the CpPtMe3 (STREM CHEMICALS purity 99 %) precursor. The pressure of the FESEM specimen chamber was about 5×10^{-3} Pa. A constant field emission current 1 μm was applied for 300 s with the gap at 0.62 μm. The gap of CNT 1 and the etched tungsten probe is set using the nanorobotic manipulator (a). The platinum nanowire is 209 nm in length (b). Figure 13c, d shows the tungsten nanowire growth on the tip of CNT 2. The tungsten nanowire grew from the MWCNT emitter by intruding $W(CO)_6$ (ALDRICH, 97 %) precursor. The pressure of the FESEM specimen chamber was again about 5×10^{-3} Pa. A constant field emission current of 500 nA was applied for 60 s at a gap of 0.70 μm. The gap of CNT 2 and etched tungsten probe is set by the nanorobotic manipulator (c). The platinum nanowire is 907 nm in length (d).

We constructed a pH sensor evaluation system as shown in Fig. 14. The buffer solutions were

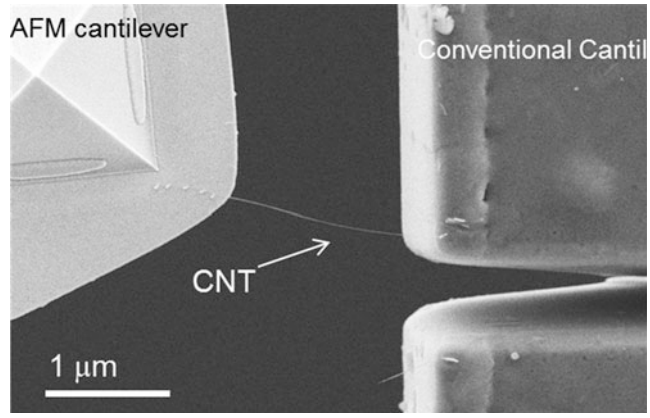


CNT Handling and Integration, Fig. 10 Fabrication of anode by FIB etching. (**a**) a tungsten probe tip (**b**) FIB etching to a plate (**c**) turned 90° and FIB etched to a needle. (**d**) etched tungsten needle coved with a parylene layer. (**e**) FIB etching the tip of the needle in order to cut the parylene layer at the tip area



CNT Handling and Integration, Fig. 11 Fabrication of the cantilever for output electrodes by FIB

introduced into the flow system. Next, a nanoprobe pH sensor was inserted to detect the electropotential together with a reference pH sensor for pH detection. A voltmeter which was constructed using a pA meter and a large resistance was employed to measure the electro potential. The characteristics were evaluated as the electro potential and pH values. Because of liquid



CNT Handling and Integration, Fig. 12 Fabrication of the cantilever for output electrodes by FIB

surface tension, the nanoprobe pH sensor has a risk of bending and being damaged. However, several buffer solutions are required for testing which increases the risk of damage to the nanoprobe pH sensor. It was necessary to provide a solution to this problem, and for this purpose we used continuous flow measurements. A container with an inlet and outlet was fabricated. Initially, buffer solutions of pH = 4 were flowed into the container and then the pH sensor nanoprobe was inserted into the solution. A reference pH sensor was placed close to the pH sensor nanoprobe in the container to monitor the real-time pH value. The electromotive force of the tungsten oxide and platinum nanowires was measured using a voltmeter. For the second step, a buffer solution of pH = 6.9 was flowed into the container, mixing with the previously introduced solution. The pH value changed to 6.1 as measured by the reference pH sensor. The electromotive force of the nanowires was measured. We continued to flow in the buffer solution of pH value 6.9 until the pH value of the container changed to 6.9. The electromotive force of nanowires was measure in this solution. A buffer solution of pH value of 9 was used to flow into the container, and when the pH value changed to 8, we again measured the electromotive force of the nanowires.

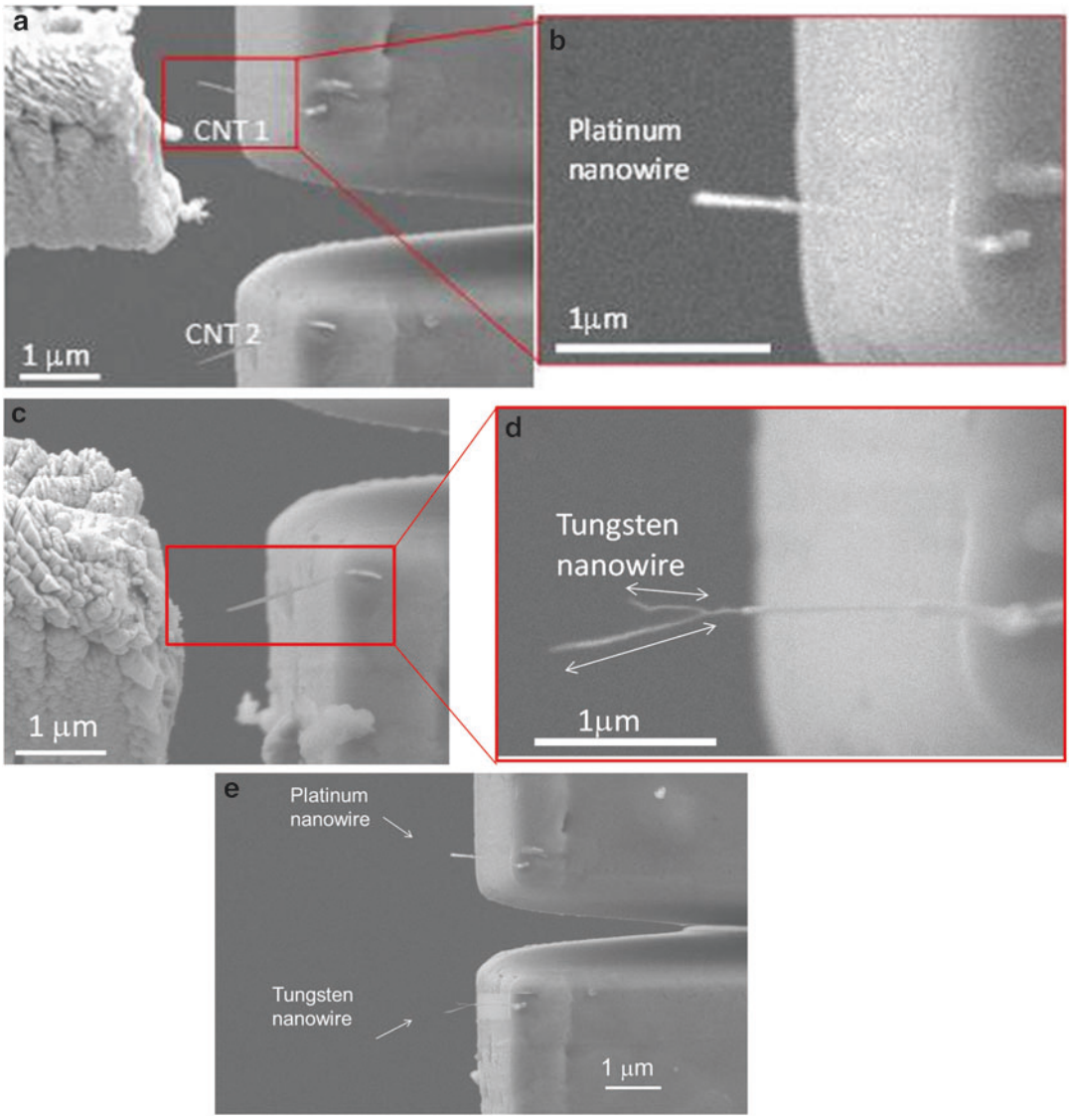
In order to prove that the nanoprobe pH sensor could be used to detect hydrogen ions based on the tungsten and tungsten oxide nanowire and the platinum nanowire and to discover the pH response of the CNTs, we fabricated double CNT electrodes and tested their pH response. Two CNTs were picked up from CNT bulks and assembled to FIB etched separated electrodes of a cantilever. The double CNTs device was tested with the buffer solution flow system using solutions with pH values of 4, 6.4, and 8. The electromotive force of the CNTs was a constant value of 0.018 mV.

The term pH value was coined to describe the solution pressure $\mathrm{pH} = -\lg\left(\mathrm{cH/mol\ dm^{-3}}\right)$, in 1909 by Søren P. L. Sørensen pondus Hydrogenii or potentia Hydrogenii, of hydrogen ions in aqueous solutions. International Union of Pure and Applied Chemistry (IUPAC) defines the quantity pH in terms of the molality basis activity αH of hydrogen ions in solution:

$$pH = -\lg[a(H^+)] = -\lg\left[\frac{m(H^+)\gamma_m(H^+)}{m^0}\right] \quad (6)$$

Where $\alpha(H^+)$ is the activity of hydrogen ion H^+ (hydrogen 1+) in aqueous solution, $m(H^+)$, $\gamma_m(H^+)$ is the activity coefficient of H^+ (molality basis) at molality $m(H^+)$, and $m^0 = 1\ \mathrm{mol/kg}$ is the standard molality.

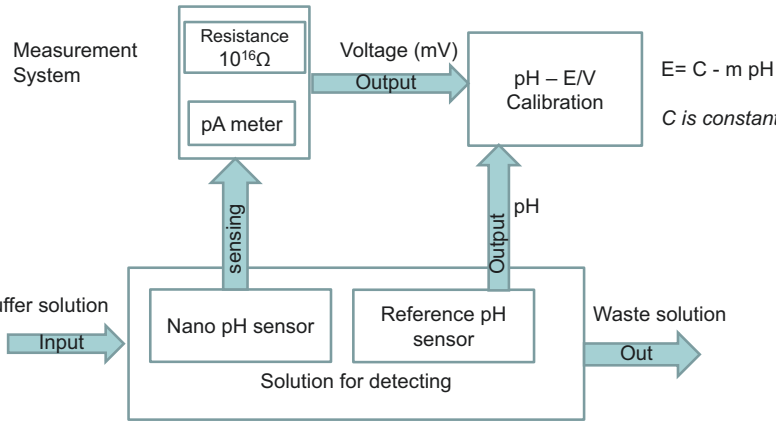
The nanoprobe pH sensor was evaluated using acidic and alkaline buffer solutions with pH 4 and 9 respectively, based on the solution flow system for a series of five times. The results reveal that the nanoprobe pH sensor can detect the pH for buffer solutions of acid (pH 4. 01, pH 6.1, and pH 6.9) and alkali (pH 8) (Fig. 15). Sensor 1 was used to test the buffer solutions at pH 4 and pH 9 to confirm that the sensor has a pH response in acidic and alkaline media. The open circuit potentials



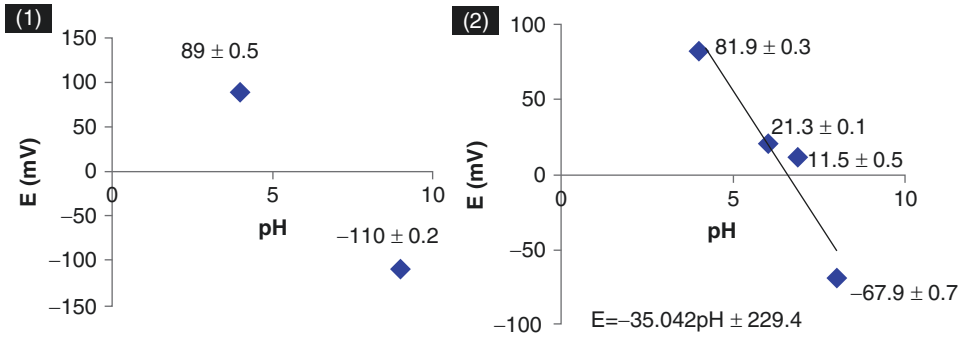
CNT Handling and Integration, Fig. 13 (**a**, **b**) show typical growth of the platinum nanowire and (**c**, **d**) show the tungsten nanowire by field emission which was used for the work and the counter electrode. (**e**) shows the assembled pH probe sensor

(OCP) were 89.1 ± 0.5 and −110.0 ± 0.2 mV, respectively. The results of sensor 1 show that the pH sensor has different electromotive forces under acidic and alkaline conditions. Sensor 2 was used to test the pH sensitivity of the tungsten oxide and platinum nanowire pH sensor. The OCP values in buffer solutions of pH 4.0, 6.1, 6.9, and 8.0 were 81.9 ± 0.3, 21.3 ± 0.1, 11.5 ± 0.5, and −67.9 ± 0.7 mV, respectively. By calculation, the sensitivity of the nanowire pH sensor was −35.04 mV pH^{-1}. The error of the sensor is 0.1 mV level. The resolution of open circuit potentials are 0.01 mV and significant figure should be 0.1 mV. The experimental pH accuracy is 0.8 pH.

The performance of pH sensors is usually characterized by measuring the OCP of the electrodes in solutions with various pH values.



CNT Handling and Integration, Fig. 14 Block diagram for the nano pH sensor evaluation



CNT Handling and Integration, Fig. 15 pH sensors evaluation in the buffer solutions

A single-phase interaction electrode may similarly be envisaged, though no example of a pH electrode involving oxygen-deficit phases has been provided [18, 19]. By omitting hydration, the surface mechanism can be expressed as follows:

$$WO_3 + 2nH^+ + 2ne^- \Leftrightarrow WO_{3-n} + nH_2O \quad (7)$$

According to the Nernst equation for the equilibrium case, the electrode potential can be stated:

$$\begin{aligned} E &= E^0 - \frac{R \cdot T}{z \cdot F} \ln \frac{\alpha_{(WO_{3-n})}}{\alpha_{(WO_3)} \cdot \alpha_{(H^+)}^{2n}} \\ &= E^0 - \frac{R \cdot T}{z \cdot F} \ln \frac{\alpha_{(WO_{3-n})}}{\alpha_{(WO_3)}} - \frac{R \cdot T}{F} \ln \frac{1}{\alpha_{(H^+)}} \\ &= E^0 - \frac{2.303 \cdot R \cdot T}{F} pH \\ &= E^0 + mpH \end{aligned} \quad (8)$$

Where E is the measured potential, E^0 is the standard potential which is zero, R is the gas constant (8.314 J K^{-1} mol^{-1}), T is the absolute temperature (K), z is the signed ionic charge (mol), and F is the Faraday constant (96487.3415 C mol^{-1}). At room temperature $(T = 298\ K)$, during the reaction the $\frac{\alpha_{(WO_{3-n})}}{\alpha_{(WO_3)}}$ approximate to 1 [20], the slope m should be

$$\begin{aligned} m &= -2.303 \cdot R \cdot T/F = -0.0591V/pH \\ &= -59.1mV/pH \end{aligned} \quad (9)$$

At the counter electrode, the platinum nanowire surface reaction can be described as follow [20]:

$$2H_2O \Leftrightarrow O_2 + 4H^+ + 4e^- \quad (10)$$

Here, platinum plays a role as catalyst, so the performance of this pH sensor is

−59.1 mV/pH. The sensitivity of the pH sensor nanoprobe which we fabricated was −35.04 mV/pH. The reason that this sensitivity had a bias with respect to that from the Nernst equation is due to the fact that the tungsten nanowire is not of a constant thickness. And the method of tungsten growth is field emission. The growth tungsten plays a role as field emission tip which is high temperature up to 1000 K. The high temperature leads to the decrease of sensitive of tungsten/tungsten trioxide nanowire hydrogen ion sensitive. The tungsten nanowire grown via field emission has a polycrystalline structure and the lattice direction was random. Various tungsten oxides covered the surface of the tungsten nanowire. The thickness of the tungsten trioxide is not constant, which leads to the observed bias in the value of the sensitivity.

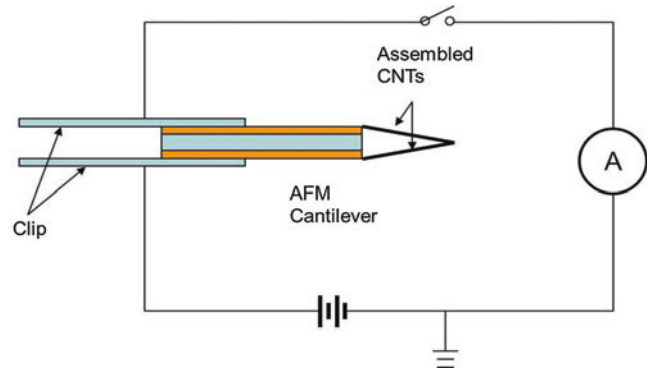
We aim to detect the pH value and pH distribution inside a cell. Common cell has a size of 5–50 μm. We design an electrochemistry pH sensor with a potential to inject to the cell and measure the pH distribution. The issue is how to make an integrated probe type pH sensor. The previous research had designed the pH sensor with a sensitive nanowire work electrode, however the counter electrode is several centimeter in diameter. It is impossible to insert into a cell and measure the pH distribution inside the cell. This is the "size problem." Here, we integrated the nanowork electrode and counter electrode into a single device with a gap $1 \sim 2\mu m$ in between. The proposed pH sensor has to insert cell and detection the pH.

The advantages of the pH sensor are as follows. First, the design of this pH sensor is an integrated probe type sensor; to the author's knowledge this is the first design of a nanoscale integrated probe type pH sensor. Second, in the fabrication the nanowires were used for work electrode and counter electrode and fabricated on one probe. This sensor has a potential to detect pH value inside the cell. Second, the sensitivity of pH sensor is near to the theory calculation. And with the calibration and calculation the resolution of this pH sensor is 0.8 pH.

CNT Probe Thermal Sensor Based on Nanorobotic Manipulation

This part reports a thermal sensor assembled by two multiwalled carbon nanotubes (MWCNTs) inside a field emission scanning electron microscope (FESEM) which constructed a probe tip in order to detect the local environment of single cell. An atomic force microscope (AFM) cantilever was used for substrate which is composed by Si_3N_4 and double side was covered by gold layer. Two MWCNTs were assembled to both side of AFM cantilever individually by employing a nanorobotic manipulation. A second AFM cantilever was used as an end-effector to manipulate the MWCNTs to touch each other, and used electron beam induced deposition (EBID) to bonding the two MWCNTs. The MWCNTs probe thermal sensor was evaluated inside a thermostated container at 25–60 °C. The experimental results show the positive characteristics of the temperature coefficient of resistance (TCR).

Nanomechanical and nanoelectromechanical devices have been widely investigated in recent years. The precise and controllable nanofabrication methods are important to develop the novel type of nanodevices. One-dimensional nanostructure such as nanowires and nanotubes are basic building blocks for NEMS. Since carbon nanotubes (CNT) have been discovered many applications have been developed such as mass sensor, nanoradio, and nanomotor. Nanowire is a solid one-dimensional nanostructure which includes metallic nanowires, semi conductive nanowires, and nonmetal nanowires which have applications as nanoneedles, nanohotwire transistor, and nanogenerator. One challenge is to assemble and apply a nanosensor to detect the local environment of single cell. The single cell is the basic element of life and it is the converging point of meter, energy and information. Because of the size of the cell, nanodevice is needed for measurement and manipulation. A thermometer is a device that measures temperature or temperature gradient using a variety of different principles of mechanical, thermojunctive, thermoresistive, and radiative (infrared and optical pyrometers). To use as a sensor, the extremely small size of the CNT can



CNT Handling and Integration, Fig. 16 Concept drawing of the MWCNTs probe thermal sensor

provide accurate measurement at nanoscale size and reduces the possibility of disturbing the local environment. In addition, the small size sensor implies very low power consumption. To use as a temperature sensor, CNT can have the smallest time constant and provides the extremely rapid time response with the measured object's temperature.

C. K. M. Fung. et al. reported a CNT-based thermal sensor. With two Au electrodes, the bundled CNTs were assembled in between on the substrate. C. Y. Kuo et.al. fabricated lateral growth of CNTs between two electrodes thermal sensor. The CNTs grow on the substrate between electrodes. F. Arai, et al. reported a single CNT thermal sensor with a CNT assembled on four electrodes. The former sensors have advantages of smaller size, lower energy cost, and suitable to measure the specimens within 2D surface. Due to the restriction of the architecture design, these sensors can only measure the specimens with the size larger than 1 mm scale.

We proposed a thermal sensor assembled by two multiwalled carbon nanotubes (MWCNTs) inside a field emission scanning electron microscope (FESEM). The advantage of the MWCNTs probe sensor is to detect the local environment of single cell with high accuracy as shown in Fig. 16. The basic idea is to detect temperatures by the thermal coefficient of resistance. The biological cells are commonly alive and cultured at temperature for 35–37 °C. When the temperature goes up to 43 °C most cells will be exterminated. We aim to evaluate the thermal sensor in the range of 30–60 °C.

A nanowire probe is designed by temperature coefficient of resistance (TCR) method. The electrical resistance of a conductor is dependent upon collisional processes within the wire; the resistance could be expected to increase with temperature since there will be more collisions. An intuitive approach to temperature dependence leads one to expect a fractional change in resistance which is proportional to the temperature change:

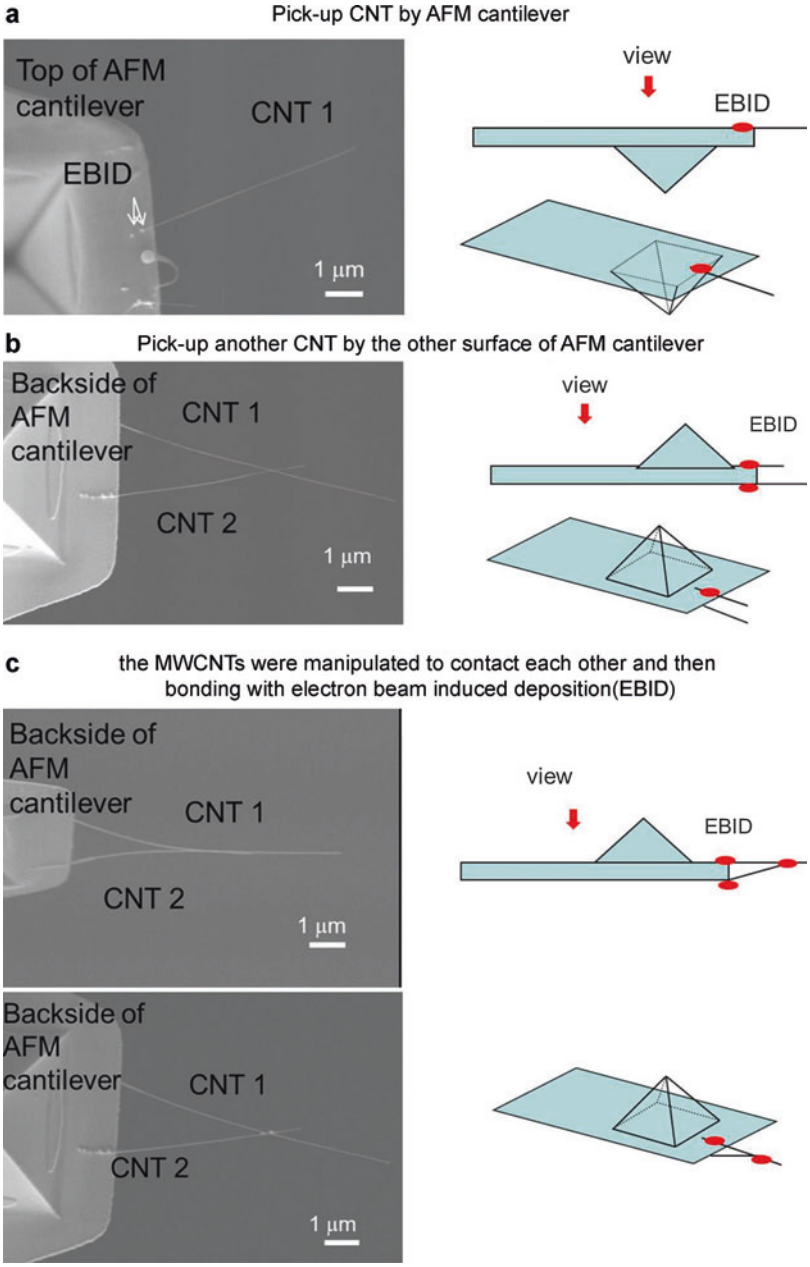
$$\frac{\Delta R}{R_0} = \alpha \Delta T \tag{11}$$

Here ΔR is the resistance change. R_0 is the initial resistance at 0 K. α is the temperature coefficient of resistance. ΔT is the temperature change.

Based on temperature coefficient of resistance, we designed a probe type thermal sensor by CNT. The basic design is the resistance of CNT change since temperature change. An AFM cantilever was used for substrate which is constructed by Si_3N_4 and coated by Au layer both sides. CNTs were assembled to the surface of AFM cantilever. Then use another cantilever to force the CNTs to touch each other.

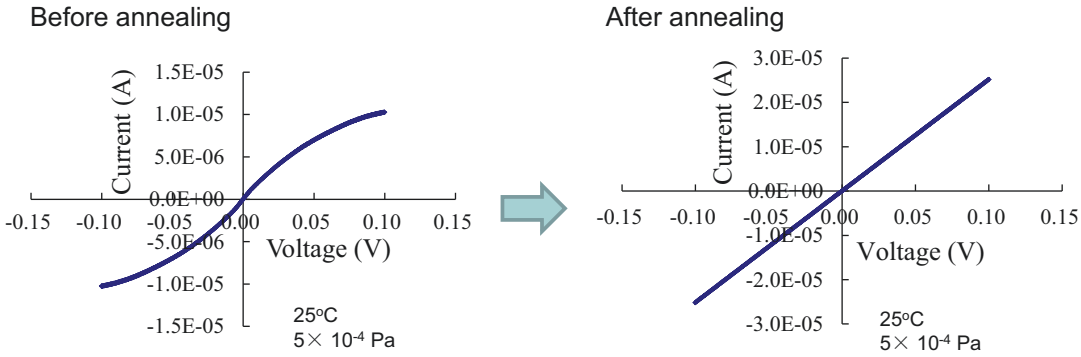
An AFM cantilever (Olympus, OMCL-TR400PB-1) is used as an end-effector to pick up the MWCNT. An AFM cantilever is constructed by Si_3N_4 and coated by Au layer on two sides. We use AFM cantilever as robot hand for MWCNT pick up and assembly. The MWCNT probe is fabricated from its bulks which are synthesized by arc-discharge method.

First, the AFM cantilever was manipulated to pick up a CNT from CNT bulk, which was

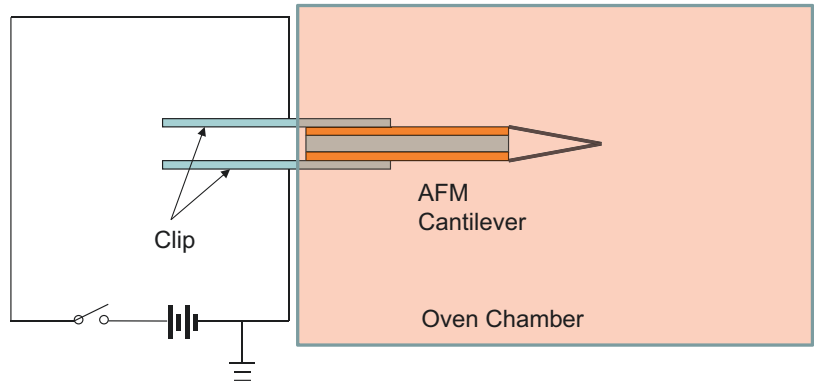


CNT Handling and Integration, Fig. 17 Assembly of MWCNTs probe thermal sensor. (**a**). Pick up CNT by AFM cantilever. (**b**) Pick up another CNT by the other surface of AFM cantilever. (**c**). The MWCNTs were manipulated to contact each other and then bonding with electron beam induced deposition (*EBID*)

synthesized by arc-charge method. The top surface of cantilever was used to pick up CNT 1. The cantilever was moved to below the CNT and went up in order to use Au surface to touch and bonding the CNT with van der Waals force. Electron beam induced deposition points were used to reinforce the bonding of CNT 1 and Au substrate as shown in Fig. 17a. Second step, we turned the cantilever over. And we assembled CNT 2 to the back surface of the Au surface as shown in Fig. 17b. Third step, we used another cantilever to manipulate CNT 1 in order to touch CNT 2. Then, an EBID was deposited on the CNT's contact point to improve the bonding as shown in Fig. 17c. In total four MWCNTs probe thermal sensors were fabricated. The MWCNTs were synthesized by



CNT Handling and Integration, Fig. 18 Annealing of CNT



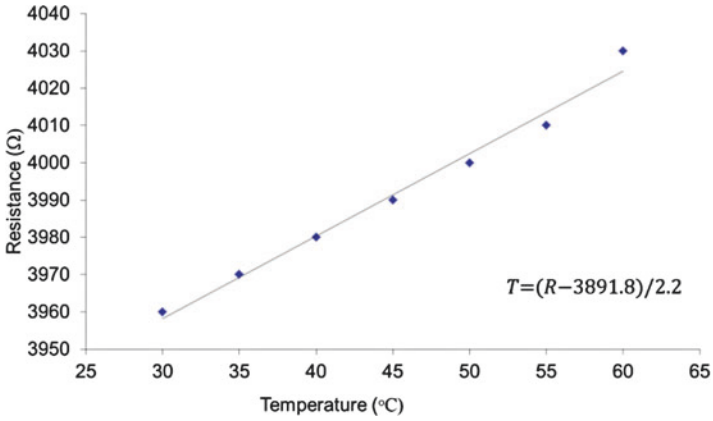
CNT Handling and Integration, Fig. 19 Schematic of evaluation of thermal sensor

arc-charge method. Defects of MWCNTs are issues of mechanical and electrical applications. Annealing is a method to solve this problem. We applied sweep voltage −1–1 V to anneal the CNT. The Fig. 18 shows the anneal I-V curve. Figure 19 shows the schematic of thermal sensor evaluation. A probe thermal sensor was placed to the holder which is colored yellow. The two sides of AFM cantilever were connected to positive and negative electrodes individually. The probe sensor was set into an oven and tested for I-V curve from 30 °C to 60 °C for every five degrees.

The oven was set to 30 °C and maintained at this temperature for 2 h. Then we tested the I-V response of the MWCNT probe thermal sensor. The resistance was 3960 Ω with the standard error of 1.5 Ω. For next step, the oven temperature was increased from 5°C to 35 °C in 1 h and maintained 2 h. The resistance was 3970 Ω with the standard error of 7.4 Ω. The test points at 40 °C, 45 °C, 50 °C, 55 °C and 60 °C was set-up. The tested resistance was 3980 Ω, 3990 Ω, 4000 Ω, 4010 Ω, and 4030 Ω with standard error 1.0 Ω, 1.9 Ω, 1.3 Ω 11.1 Ω, and 1.0 Ω.

The result of CNT annealing was shown in the Fig. 20. The I-V curve had been returned to linear. The defects of CNTs were repaired. The mechanism can be described like this, the CNT were cover some of amorphous carbon. When sweep current is applied to the CNT the remote joule heating [21] investigated to the amorphous carbons and rebuilt the defects of bonding at CNTs. When the charge carriers flow pass the CNT, the vibration of lattices of CNT will absorb the charge carrier energy and create a phonon.

The experimental results show that temperature coefficient of resistance of CNT was linear in Fig. 20. The relationship is temper = resistance −3891.8/2.2. And the standard error was at 10 Ω level. The metallic MWCNT usually has a



CNT Handling and Integration, Fig. 20 Evaluation result

resistance of several k Ω. Here the increase in temperature was set to 5 °C because of the oven's limitations. According to the Eqs. 5 and 6, the temperature coefficient of resistance is calculated as 0.05 %.

The total resistance thermal sensor

$$R_{\text{total}} = R_{\text{CNT}} + R_{c1} + R_{c2} + R_{c3} \tag{12}$$

Where, R_{total} is the total resistance of nanothermal sensor. R_{c1} is contact resistance of CNT 1 and Au substrate. R_{c2} is the contact resistance of CNT 2 and Au substrate. R_{c3} is the contact resistance of two CNTs. And the resistance of CNT can be calculated by the following equation.

$$R_{\text{CNT}} = \rho_{NT} \times \frac{L}{\pi\left(\frac{d}{2}\right)^2} \tag{13}$$

Where, R_{CNT} is the resistance of CNT. ρ_{NT} is the resistivity of the CNT. L is the length of the CNT. d is the diameter of the CNT. In this experiment, CNT 1 is approximately 3 μm in length and 37 nm in diameter. The CNT 2 is approximately 3 μm in length and 42 nm in diameter. The ρ_{NT} is $3.8 \times 10^{-5}\Omega\text{cm}$ [22]. By the calculation the $R_{\text{CNT}} = 1.8\,\text{k}\Omega$.

Chun Lan et al. discussed the metal carbon contact resistance and resistance of CNT. The MWCNT resistivity is 0.33 $k\Omega/\mu m$ and specific contact résistance of CNT and Au surface is 4.4 $k\Omega\mu m$.

$$R_{\text{CNT}} = 33\frac{k\Omega}{\mu m} \times 6\mu m \approx 1.98k\Omega \tag{14}$$

$$R_{c1} + R_{c2} = \frac{4.4k\Omega\mu m}{2.5\mu m} = 1.76\ k\Omega \tag{15}$$

So

$$R_{c3} \approx 210\ \Omega\ \sim 390\Omega \tag{16}$$

The calculation results of CNT resistance meet the reference data and the theoretical calculation. The MWCNT can be considered as a graphene cylinder. The single crystal graphite shows the linear thermal response when the temperature is higher than 200 K. The evaluation result of nanothermal sensor is linear.

In summary, this section presented the CNT integration and applications and fabrication of a functional pH sensor with a double metallic nanowire. A cantilever was used for the electrodes of the device and etched by an FIB. Two CNTs were assembled to the separated electrodes of the cantilever. A tungsten probe was etched by an FIB to 300 nm in diameter and 25.4 μm in length. Then the probe was coated by parylene and the tip cut to expose the tungsten. A tungsten nanowire with 907 nm length and a platinum nanowire with 209 nm length grew from the tip of the CNTs via field emission by introducing hexacarbonyltungsten and trimethylcyclopentadienylplatinum, respectively, inside a field

emission electron microscope. The acid and alkali response of the pH sensor was tested; the sensor was able to detect the pH value, and the performance of the pH sensor was −35.04 mv/pH. We hope that these findings will be rapidly adopted in the field of biochemical detection.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Nanorobotic Assembly
- ▶ Nanorobotics
- ▶ Scanning Electron Microscopy

References

1. Iijima, S.: Helical microtubules of graphitic carbon. Nature **354**(6348), 56–58 (1991)
2. Visser, J.: Van der Waals and other cohesive forces affecting powder fluidization. Powder Technology. **58**, 1–10 (1989)
3. Zheng, Q.S., Liu, J.Z., Jiang, Q.: Excess van der Waals interaction energy of a multiwalled carbon nanotube with an extruded core and the induced core oscillation. Phys. Rev. B **65** (2002)
4. Jonsson, L.M., Nord, T., Kinaret, J.M., Viefers, S.: Effects of surface forces and phonon dissipation in a three-terminal nanorelay. J. Appl. Phys. **96** , 629–635 (2004)
5. Dong, L.X., Arai, F., Fukuda, T.: Nanoassembly of carbon nanotubes through mechanochemical nanorobotic manipulations. Jpn. J. Appl. Phys. Part 1 **42**, 295–298 (2003)
6. Akita, S., Nishijima, H., Nakayama, Y.: Influence of stiffness of carbon-nanotube probes in atomic force microscopy. J. Phys. D-Appl. Phys. **33** , 2673–2677 (2000)
7. Xia, Y.N., Whitesides, G.M.: Soft lithography. Annu. Rev. Mater. Sci. **28** , 153–184 (1998)
8. Lixin, D., Arai, F., Fukuda, T.: Destructive constructions of nanostructures with carbon nanotubes through nanorobotic manipulation. IEEE/ASME Trans. Mechatron. **9**, 350–357 (2004)
9. Nakajima, M., Arai, F., Fukuda, T.: In situ measurement of young's modulus of carbon nanotubes inside a TEM through a hybrid nanorobotic manipulation system. IEEE Trans. Nanotechnol. **5**, 243–248 (2006)
10. Heping, C., Ning, X., Guangyong, L., Jiangbo, Z., Prokos, M.: Planning and control for automated nanorobotic assembly. In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on, pp. 169–174. (2005)
11. Deboux, B.J.C., Lewis, E., Scully, P.J., Edwards, R.: A novel technique for optical fiber pH sensing based on methylene blue adsorption. J. Lightwave Technol. **13**, 1407–1414 (1995)
12. Ruan, C., Zeng, K., Grimes, C.A.: A mass-sensitive pH sensor based on a stimuli-responsive polymer. Anal. Chim. Acta **497**, 123–131 (2003)
13. Yao, M.W.S., Madou, M.: A pH electrode based on melt-oxidized iridium oxide. J. Electrochem. Soc. **148**, 29–36 (2001)
14. Talaie, A., Lee, J.Y., Lee, Y.K., Jang, J., Romagnoli, J. A., Taguchi, T., et al.: Dynamic sensing using intelligent composite: an investigation to development of new pH sensors and electrochromic devices. Thin Solid Films **363**, 163–166 (2000)
15. Bashir, R., Hilt, J.Z., Elibol, O., Gupta, A., Peppas, N. A.: Micromechanical cantilever as an ultrasensitive pH microsensor. Appl. Phys. Lett. **81**, 3091–3093 (2002)
16. Bergveld, P.: Thirty years of ISFETOLOGY: what happened in the past 30 years and what may happen in the next 30 years. Sensors Actuators B Chem. **88**, 1–20 (2003)
17. Nomura, Y.K.S., Hiraishi, N., Nakajima, M., Nikaido, T.: Application of semiconductor pH-imaging sensor to the imaging of extracted human carious dentin. Anal. Sci. **17**, 539–541 (2001)
18. Fenster, C., Smith, A.J., Abts, A., Milenkovic, S., Hassel, A.W.: Single tungsten nanowires as pH sensitive electrodes. Electrochem. Commun. **10**, 1125–1128 (2008)
19. Zhang, W.-D., Xu, B.: A solid-state pH sensor based on WO_3-modified vertically aligned multiwalled carbon nanotubes. Electrochem. Commun. **11**, 1038–1041 (2009)
20. Kurzweil, P.: Metal oxides and ion-exchanging surfaces as ph sensors in liquids: state-of-the-art and outlook. Sensors **9** , 4955–4985 (2009)
21. Baloch, K.H., Voskanian, N., Bronsgeest, M., Cumings, J.: Remote Joule heating by a carbon nanotube. Nat. Nanotechnol. **7**, 316–319 (2012)
22. Primak, W., Fuchs, L.H.: Electrical conductivities of natural graphite crystals. Phys. Rev. **95**, 22–30 (1954)

CNT NEMS

- ▶ Carbon Nanotube NEMS

CNT Resonator

- ▶ Carbon Nanotube NEMS

CNT-FET

▶ Nanostructure Field Effect Transistor Biosensors

Coarse-Grained and Hybrid Simulations of Nanostructures

Richard Gowers and Paola Carbone
School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester, UK

Synonyms

Mesoscopic simulations; Multiscale simulations

Definition

In computational chemistry coarse-grained (CG) models are defined as molecular models where some details (i.e., degrees of freedom) of the original chemical structure have been removed. The resulting models are a coarser description of the chemical systems compared with the original ones and can then be used to perform either molecular dynamics or Monte Carlo simulations [1]. The reduction of the models' degrees of freedom enables the simulation of systems whose size is comparable with that of the experimental ones and the timescale spanned by these simulations can reach microseconds.

Overview

Computer modeling is a powerful technique to gain molecular level details of chemical systems under different physical conditions and enables to relate macroscopic observations with changes in the chemical and physical state of the system. However, all modeling techniques rely on computer hardware, and therefore their use is limited by the available computer power. State-of-art simulations can nowadays reach the size of few millions of atoms, but if standard high-performance computers are used, the system size usually does not exceed few hundred thousand particles. Indeed, during a molecular simulation, the number of interatomic interactions that must be computed every iteration is proportional to N^2, where N is the total number of system particles [1]. This heavy use of the CPUs limits not only the size of molecular models but also the timescale the system can be simulated for.

One way to circumvent this problem is to reduce the number of interacting particles (N) in the systems, simplifying the models and modifying the original interacting parameters to include in an implicit way the neglecting details. The simplest example of such coarse graining is the development of united-atom (UA) force fields [2] where the hydrogen atoms and the aliphatic carbon to which they are covalently bonded are modeled as a single entity. The assumption underlying the development of such force fields is that the physics of the model is not affected by neglecting the explicit interactions involving the aliphatic hydrogen atoms. A similar decision on how many and which atomistic details can be neglected in a molecular model (procedure known as mapping scheme) is the key decision that must be carefully made every time a new coarse-grained model is developed if some of the system chemical and physical features have to be maintained. It has been indeed shown that mapping schemes which retain different features of the original molecule perform differently depending on the property analyzed [3].

The model simplification done in the UA force fields can therefore be carried on at much larger scale: large group of atoms can be lumped up in single super-atoms (or beads) and even entire colloidal particles can be modeled as single rigid bodies.

A very broad distinction can be made between those CG models aiming at preserving some chemical details of the original system and those which instead preserve only the shape of large

molecular aggregates. The former can reproduce with a certain degree of accuracy the system enthalpy, while the latter reproduce the only entropic effects, and it is typically employed in modeling colloidal systems [3–6].

Procedures

Since different features of the atomistic model can be used as target properties, several procedures to develop the effective interactions between the beads have been proposed. Some of the most popular methods used in materials science are briefly introduced below.

Structural Based Model

The target property to reproduce is the structure of the atomistic model. The CG non-bonded part of the force field is therefore refined until it reproduces structural correlation functions such as the radial distribution function (RDF) of the atomistic system. The quality of the agreement is quantified using a merit function of the form

$$\chi^2 = \sum \left(y_{\text{target}} - y_{CG} \right)^2 \qquad (1)$$

where y_{target} and y_{CG} are the correlation functions calculated from the atomistic and CG model, respectively. The sum is over all coordinates and simulations. The minimization of χ is a nonlinear inverse problem for which no analytical expression is available and where linearization is unsuccessful. An iterative procedure is therefore run and the CG force field parameters are consequently adjusted. Two iterative procedures are usually employed: one uses Monte Carlo simulations [7] and the other one molecular dynamics [8]. In both cases, the new CG model is built upon the atomistic one and the existence of a one-to-one correspondence between the two models enables an easy interchange between the two model resolutions [3].

Thermodynamic Models

The aim of these types of models is to reproduce some thermodynamic properties of the atomistic system. The CG potential is usually modeled using a Mie (generalized Lennard-Jones) function

$$V(r) = C_\varepsilon \left\{ \left(\frac{\sigma}{r} \right)^n - \left(\frac{\sigma}{r} \right)^m \right\} \qquad (2)$$

where n and m are the exponents controlling the softness of the potential, σ is the particle diameter, ε is the potential well depth, and C is chosen in such a way that the minimum of the potential corresponds to $-\varepsilon$.

The key parameters to control and iteratively optimize are ε and σ. In some procedure the exponents n and m can also be varied. The target properties for the CG model include partition coefficients [9], density, and interfacial tension [10]. Another more systematic approach to obtain the sought force field parameters involves the solution of a molecular-based equation of state (EoS) which takes advantage of the fact that for some molecular fluids the SAFT EoS can be solved [11].

Force Matching

In this type of method [12, 13] the CG interactions are again parameterized using the underlying atomistic interactions, but in this case the difference to be minimized is that between the atomistic and CG forces and the minimization is solved using a variational approach

$$\chi^2 = \frac{1}{3N} \left\langle \sum_{I=1}^{N} \left(F_{\text{target}} - F_{CG} \right)^2 \right\rangle \qquad (3)$$

where the angular brackets denote average over the trajectory, F_{target} is the net force on the atomistic site I, F_{CG} is the force on the same CG site, and N is the number of sites (beads) in the CG model. Within this type of CG model a one-to-one correspondence between the atomistic and CG resolution is again established, and reintroduction of atoms in the mesoscopic model is possible.

Excess Entropy Model

In this case the function which needs to be minimized is the relative entropy (S_{rel}) which is defined as the difference between the distributions of configurations generated by the atomistic (P_{at})

and CG (P_{CG}) models. Using the Kullback–Leibler divergence formalist, one can define S_{rel} as

$$S_{\text{rel}}(U) = \int P_{\text{at}} \frac{P_{\text{at}}}{P_{CG}(U)} dR \tag{4}$$

where U is the CG potential and the average is evaluated over the CG configurational space, but weighted according to the atomistic probability distribution, P_{at}. S_{rel} provides a variational framework for determining the approximate CG potential (U) that reproduces target atomistic distributions [14, 15].

Dissipative Particle Dynamics

The techniques presented above despite simplifying the molecular model still retain a quite high degree of chemical specificity. If however much larger systems must be simulated, a mesoscopic approach such as dissipative particle dynamics (DPD) can be employed. Such method can sample large conformational space in relatively short time, and therefore systems of the order of hundreds of nanometers can be simulated [16, 17]. In this method, each bead represents a much larger amount of material than the previously discussed methods, such as an entire molecule or a few molecules of solvent. Because each DPD particle represents such large portion of the chemical system, the standard analytical functions used to model the particle-particle non-bonded interactions (Eq. 2) cannot be used. Therefore, in DPD the forces are expressed as the sum of three contributions

$$F_{ij} = F_{ij}^{C}(r_{ij}) + F_{ij}^{D}(v_{ij}) + F_{ij}^{R}(\theta) \tag{5}$$

where F_{ij}^{C} represents a conservative repulsive force between particles and it is notably weaker than the standard short distance forces (Eq. 2) allowing particles to pass through each other (the so-called soft potential). Such weak short-range interaction is a necessary consequence of the much larger mapping scheme used in this modeling technique. F_{ij}^{D} is a dispersive force between particles which is a function of their relative velocity and represents the effect of fluid viscosity. Finally, F_{ij}^{R} is a random force, a function of a Gaussian random number θ, which replicates the thermal and vibrational energy of the system.

Because all of these forces, including the random force, are applied between pairs of particles, momentum is conserved throughout the system. DPD simulations are tuned to replicate hydrodynamic properties by adjusting the strength of each of these three forces. DPD has been used to model problems which are otherwise out of reach using particle-based methods, including modeling self-assembly and phase diagram of complex fluids.

Mean Field Theory

A radically different approach to coarse graining is that which simplifies the molecular systems at such level that its discrete nature (the fact that is made by individual atoms) is replaced by a continuum mean field. In this context the molecular system is modeled using a grid of points on which an effective field, representing the averaged interaction caused by the presence of all the system particles within a cutoff distance, which is identified with the mesh of the grid, acts. In its basic form, to obtain the field (H_{eff}), the free energy of the system is minimized with respect to the distribution of all possible configurations of the system, P. Since both P and H_{eff} are not known, an iterative process is used until self- consistency is reached [18]:

$$P = \frac{\exp\left(\frac{-W}{k_B T}\right)}{\sum_N \exp\left(\frac{-W}{k_B T}\right)} \tag{6}$$

$$W = \frac{\partial H_{\text{eff}}}{\partial P} \tag{7}$$

where k_B and T are the Boltzmann constant and the temperature, respectively, W is the intermolecular potential, and the summation is run over all the N possible states of the system.

Applications in Materials Science

The procedure briefly presented above can be used in a variety of contexts in both materials science and biology. All have been indeed used

to model synthetic and biopolymers, complex and simple liquids, surfactants, and mixtures of them [3–6]. Below, three examples of such applications in materials science are presented.

Predicting Self-Assembly Properties for Amphiphilic Copolymers

Amphiphilic copolymers are polymer chain formed by more than one type of monomer each with a different polarity. When dissolved in solvents, they self-assemble in a variety of morphologies which are function of the relative chemical affinity of the solvent and monomers and the chain microstructure (the monomer sequence within the chain). Such self-assembled nanometric structure can then be used in a variety of applications from drug and gene delivery agents to emulsion stabilizers [19]. The a priori prediction of their phase diagram is very difficult to achieve, and modeling can help in avoiding the synthesis of many different materials. Using an atomistic model for such simulations is however not possible; therefore CG models have become a popular choice. Mainly using DPD [20] and thermodynamics model [21], it has been recently possible to follow at the self-assembly process and build entire phase diagrams for such systems [20].

Predicting the Long-Time Dynamics of Ionic Liquids

Ionic liquids (ILs) are high molecular weight ion pairs formed by organic cations and bulky counterions. They are liquid at room temperature and below so that they can be exploited in a considerable number of extraction and other chemical processes. Knowing the dynamics of ILs at low temperature can therefore improve the design of the ideal solvent for a specific extraction process; however such experiments are very difficult to carry out since the IL's viscosity increases very rapidly lowering the temperature [22]. CG modeling can help in gaining such data as the dynamics of the model is sped up compared with the experimental one by a factor which depends on the CG scheme used. Using CG models it has been possible to show that certain type of ILs with long aliphatic chains can self-assemble [23] and that at low temperatures they present an increased dynamic heterogeneity which is due to an increasing number of slow ions, while the amount of fast ions stays almost constant with temperature [24].

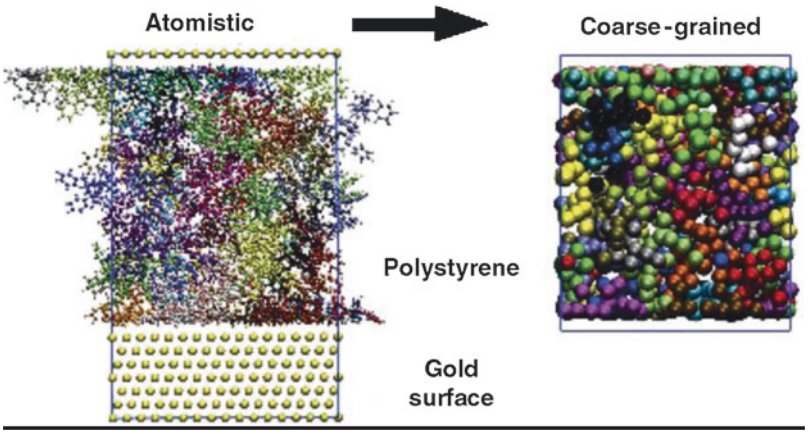
Calculating Interphase Thickness of Polymer Films on Solid Surfaces

When in contact with a solid surface, polymers exhibit complex behavior and understanding and characterizing this behavior is important in the design and development of nanostructures. Such systems are difficult to model as interactions at both small and long length scales need to be included.

A common approach to studying these systems therefore is to propagate information from the smaller length scales into the larger-scale models, and such an approach has been used for modeling polystyrene in contact with a gold surface [25]. Using data from atomistic-scale simulations of small amounts of the nanostructure, a coarser model was constructed (see Fig. 1). This model had reduced the number of particles in the polymer by means of coarse graining and also modeled the gold surface implicitly. This meant that each particle in the polymer interacted with the surface through a single function of normal distance instead of being a sum of interactions with many particles in the surface. This mesoscale model enabled the simulations of much larger volumes and led to the identification of an interphase thickness in which the polymer presents structural properties different than those in bulk.

Future Directions

Coarse-graining techniques have become powerful and well-established tools to access large length and timescales. The use of such techniques however is not without any drawbacks. In certain cases the degrees of freedom coarse-grained away in the mesoscale models represent crucial details without which the behavior of the models no longer faithfully recreate the phenomena that are to be studied. This is, for example, the case in molecular systems where short-range directional non-bonded interactions such as hydrogen bonds are present [3].



Coarse-Grained and Hybrid Simulations of Nanostructures, Fig. 1 An example of the reduction in detail possible when using coarse graining.

To circumvent this limitation typical of traditional coarse-graining techniques, the possibility of mixing both coarser and finer levels of detail in a single simulation in so-called hybrid-scale simulations has been recently explored. These two levels of detail are present simultaneously in the model allowing specific parts of the system to be modeled at a higher level of detail while allowing other parts to use a computationally cheaper model [26].

Two main approaches can be identified for such hybrid-scale simulations. Indeed one can mix either CG and atomistic force fields or particle-based (CG or atomistic) and field models. Both approaches are briefly explained below.

Hybrid Atomistic/Coarse-Grained Models

In models where atoms and CG beads are simultaneously present, in order to calculate the interactions between pairs of atoms or beads, preexisting standard force fields can be used [27]. However, interactions between atoms and beads, the so-called cross terms, must also be evaluated. These cross term interactions could be parameterized using one of the previously described coarse-graining methods; however this is an undesirable option because it would greatly aggravate the work required to derive the model and increase the number of parameters within the model. Instead, a commonly used solution is the use of virtual sites which are geometrical points identified in the finer (atomistic) resolved regions of the model which collect the interactions coming from the model coarser parts. If the CG model is derived from the atomistic one, as it is in the case of structural based or force matching models, the CG mapping scheme is built on the chain atomic structure and each bead center of mass corresponds either to a specific atom or to the center of mass of all the atoms lumped up in it. Thus, it is easy to identify in the atomistic resolved part of the model the positions of the virtual beads and use them as pinning position to collect any mixed interactions. The VS can therefore be seen as coarse-grained representation of a group of atoms, placed at their center of mass, which facilitate their interaction with coarse-grained beads. When a coarse-grained particle interacts with the atoms of a certain group, it is with the virtual site representation of that group. To ensure that all particles can interact with each other, all atoms must be mapped to such a virtual site.

Once all such pairwise interactions have been evaluated, the net force on each virtual site is transferred onto the constituent atoms. This is done on a mass-weighted basis, so that heavier atoms get a larger share of the force from the virtual site. This process ensures that Newton's third law is obeyed and, when used in molecular dynamics simulations, that linear momentum is conserved:

$$F_{\text{atom}} = F_{vs} \cdot \frac{m_{\text{atom}}}{m_{vs}} \tag{8}$$

where $\boldsymbol{F}$ refers to the net force and m the mass of the particle. Using this method, non-bonded

interactions no longer cross between the different resolutions and the coarse-grained force field can be used to describe and parameterize these interactions. This means that once an atomistic and coarse-grained force field is chosen, it is also possible to start creating hybrid-scale models.

Using virtual sites, models which mix both atoms and coarse-grained beads can be constructed. In these models selected chemical motifs are kept at a fine level of detail, while all other sections are coarse-grained. The choice of where to leave fine levels of detail is left to intuition, typically dipoles, including hydrogen bonding sites, or areas where steric factors are of great importance are kept at fine resolution. This approach allows a detailed description of some parts of the model to be combined with a less computationally intensive description of other parts of the model [27, 28]. The Hamiltonian for the simulation system can be defined as

$$H = K_{\text{atoms}} + K_{CG} + V_{\text{bonds}} + V_{nb(cg-cg)} + V_{nb(aa-aa)} + V_{nb(vs-cg)} \quad (9)$$

where K refers to the contribution from kinetic energy and V refers to potential energy between particles.

Adaptive Resolution Models

If what needs to be modeled at high resolution is a particular recurring chemical motif in the molecular system, then the model resolution can be static, which means that during the construction of the model, one decides which chemical moieties are modeled with the different resolutions and this model choice is kept for the entire simulation. However, in some cases it is more suited to define a geometrical region within the simulation box which is modeled with a different resolution. In this case, the total number of degrees of freedom of the model will change over the time depending on how many chemical species enter or leave that region (or regions) [29]. Usually systems are constructed to have a region of interest containing molecules modeled at atomic details, surrounded by a region of coarse-grained molecules. As simulations of such systems progress, the molecules are free to move between these two regions of different resolution, and a method to allow the transition between the two model resolutions is therefore required. To handle such transition, a buffer region (known as healing region) between the two levels of detail is used which allows the molecules to gradually change from one level of detail to another. In this healing region, the models switch from using one force field to the other as a function of their position within the healing region

$$V = \lambda V_{AA} + (1 - \lambda) V_{CG} \quad (10)$$

where λ refers to a switching function between the two force fields. Great care must be taken to ensure that there is no discontinuity in the force acting on the particles, as this would create mass transfer between different regions of the simulation box. Transitioning from the coarse region into the fine detail region is particularly difficult as it involves reintroducing the degrees of freedom that have previously been discarded.

Hybrid Particle Field Models

As the computational cost of both molecular dynamics and Monte Carlo simulation methods scales as $O(N^2)$, increasingly larger simulations become too demanding to produce results in a timely manner. An alternative to this particle-based approach is hybrid models where the particles surrounding a specific atoms or CG bead is replaced by a continuum field and the interparticle non-bonded interactions are calculated with respect to that external field. This method scales as $O(N)$, making the simulation of extremely large systems possible.

This approach is becoming popular in modeling polymer blends or solutions where the polymer chains self-assemble creating specific morphologies. A commonly used field method for simulating such polymer nanostructures is the single-chain mean field theory (SCMF) [30]. In this, single polymer chains or large molecules are separated and each determines its non-bonded interaction with respect to an external

field rather than having to interact with every other particle in every other molecule. To construct the mean field, self-consistent field (SCF) theory is used, where the particle density of a given chemical species (ϕ_A) is measured at all points on a discrete grid. The force caused by the mean field for a particle of species A at a given position is then given as

$$F_A(r) = -k_B T \sum_{A'} \chi_{AA'} \frac{\partial \phi_{A'}(r)}{\partial r} - \frac{1}{\kappa}\left(\sum_A \frac{\partial \phi_A(r)}{\partial r} - 1\right) \quad (11)$$

where k_B and T are the Boltzmann constant and temperature. The term $\chi_{AA'}$ is a Flory-Huggins mixing parameter between two species and must be defined between all combinations of species and κ represents the compressibility of the system. The first term defines how favorably different species can mix, while the second term tries to balance the particle density. This field varies both spatially and temporally and must be updated as the simulation progresses. However the variation of the field is slow with respect to the intramolecular forces, meaning that several particle moves can be done before the field must be updated. Within this approach while the intermolecular interactions are calculated via the SCMF theory, the intramolecular interactions (i.e., bonds, angles, and torsions) are handled at the same level of precision as conventional particle-based simulations. Initial work on this method used Monte Carlo to perform the particle moves [31] as the use of Monte Carlo allows the possible conformational space of the molecules to be sampled effectively. More recently molecular dynamics has also been used [32]. Since the calculation of the interactions with a discrete grid is computationally quick to perform, with each particle only interacting with eight different points rather than all their neighbors, this method is also particularly well suited to the increasing parallel nature of the hardware available, with some software able to achieve hyper-parallelism (achieving a speedup greater than n using n processors) [32].

Cross-References

- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Mesoscopic Modeling

C

References

1. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Oxford University Press, Oxford (2002)
2. Ponder, J.W., Case, D.A.: Force fields for protein simulations. Adv. Protein. Chem. **66**, 27–85 (2003)
3. Karimi-Varzaneh, H., van der Vegt, N.F.A., Müller-Plathe, F., Carbone, P.: How good are coarse-grained polymer models? A comparison for atactic polystyrene. ChemPhysChem **13**, 3428 (2012)
4. Voth, G.A.: Coarse-Graining of Condensed Phase and Biomolecular Systems, CRC edn. Taylor and Francis, Boca Raton (2008)
5. Carbone, P., Avendaño, C.: Coarse-grained methods for polymeric materials: Enthalpy and entropy driven models. Wiley Interdiscip. Rev.: Comput. Mol. Sci. **4**(1), 62–70 (2014)
6. Noid, W.G.: Perspective: Coarse-grained models for biomolecular systems. J. Chem. Phys. **7**, 139 (2013)
7. Lyubartsev, A.P., Laaksonen, A.: Calculation of the effective interaction potentials from radial distribution functions – a reverse Monte-Carlo approach. Phys. Rev. E **52**(4), 3730–3737 (1995)
8. Baschnagel, J., Binder, K., Doruker, P., Gusev, A.A., Hahn, O., Kremer, K., Mattice, W.L., Müller-Plathe, F., Murat, M., Paul, W., et al.: Bridging the gap between atomistic and coarse-grained models of polymers: Status and perspectives. Adv. Polym. Sci. **152**, 41–156 (2000)
9. Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., de Vries, A.H.: The MARTINI force field: Coarse grained model for biomolecular simulations. J. Phys. Chem. B **111**(27), 7812–7824 (2007)
10. Shinoda, W., Devane, R., Klein, M.L.: Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. Mol. Simul. **33**(1–2), 27–36 (2007)
11. Avendaño, C., Lafitte, T., Galindo, A., Adjiman, C.S., Jackson, G., Müller, E.A.: SAFT-gamma force field for the simulation of molecular fluids. 1. A single-site coarse grained model of Carbon Dioxide. J. Phys. Chem. B **115**(38), 11154–11169 (2011)
12. Ercolessi, F., Adams, J.B.: Interatomic potentials from 1st-principles calculations – the force-matching method. Europhys. Lett. **26**(8), 583–588 (1994)
13. Noid, W.G., Chu, J.W., Ayton, G.S., Krishna, V., Izvekov, S., Voth, G.A., Das, A., Andersen, H.C.: The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. J. Chem. Phys. **128**(24), 244114 (2008)

14. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79 (1951)
15. Shell, M.S.: The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J. Chem. Phys. **129**, 144108 (2008)
16. Groot, R.D., Warren, P.B.: Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. J. Chem. Phys. **107**, 4423–4435 (1997)
17. Espanol, P.: Dissipative particle dynamics. In: Handbook of Materials Modeling, pp. 2503–2512. Springer, Dordrecht (2005)
18. Altland, A., Simons, B.D.: Condensed Matter Field Theory, 2nd edn. Cambridge University Press, Leiden (2010)
19. Alexandridis, P., Lindman, B.: Amphiphilic Block Copolymers Self-Assembly and Applications. Elsevier, Amsterdam (2000). ISBN 978-0-444-82441-7
20. Ortiz, V., Nielsen, S.O., Discher, D.E., Klein, M.L., Lipowsky, R., Shillcock, J.: Dissipative particle dynamics simulations of polymersomes. J. Phys. Chem. B **109**, 17708–17714 (2005)
21. Nawaz, S., Carbone, P.: Coarse-graining poly(ethylene oxide)-poly(propylene oxide)-poly(ethylene oxide) (PEO-PPO-PEO) block copolymers using the MARTINI force field. J. Phys. Chem. B **118**, 1648–1659 (2014)
22. Rogers, R.D., Seddon, K.R.: Ionic Liquids, Science **302**, 792–793 (2003)
23. Wang, Y.T., Voth, G.A.: Unique spatial heterogeneity in ionic liquids. J. Am. Chem. Soc. **127**, 12192–12193 (2005)
24. Karimi-Varzaneh, H., Muller-Plathe, F., Balasubramanian, S., Carbone, P.: Studying long-time dynamics of imidazolium-based ionic liquids with a systematically coarse-grained model. Phys. Chem. Chem. Phys. **12**, 4714–4724 (2010)
25. Johnston, K., Harmandaris, V.: Hierarchical multiscale modeling of polymer–solid interfaces: atomistic to coarse-grained description and structural and conformational properties of polystyrene–gold systems. Macromolecules **46**, 5741 (2013)
26. Abrams, F.C., Delle Site, L., Kremer, K.: Dual-resolution coarse-grained simulation of the bisphenol-A-polycarbonate/nickel interface. Phys. Rev. E **67**, 21807 (2003)
27. Rzepiela, A.J., Louhivuori, M., Peter, C., Marrink, S. J.: Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. Phys. Chem. Chem. Phys. **13**, 10437–10448 (2011)
28. Di Pasquale, N., Marchisio, D., Carbone, P.: Mixing atoms and coarse-grained beads in modelling polymer melts. J. Chem. Phys. **137**, 164111 (2012)
29. Praprotnik, M., Delle Site, L., Kremer, K.: Multiscale simulation of soft matter: from scale bridging to adaptive resolution. Annu. Rev. Phys. Chem. **59**, 545–571 (2008)
30. Müller, M., de Pablo, J.J.: Computational approaches for the dynamics of structure formation in self-assembling polymeric materials. Annu. Rev. Mater. Res. **43**, 1–34 (2013)
31. Daoulas, K.C., Müller, M., de Pablo, J.J., Nealey, P.F., Smith, G.D.: Morphology of multi-component polymer systems: single chain in mean field simulation studies. Soft Matter **2**, 573–583 (2006)
32. Milano, G., Kawakatsu, T.: Hybrid particle-field molecular dynamics simulations for dense polymer systems. J. Chem. Phys. **130**, 214106 (2009)

Coarse-Grained Molecular Dynamics

► Dissipative Particle Dynamics, Overview

Cochlea Implant

► Bioinspired CMOS Cochlea

Cold Field Electron Emission from Nanostructured Materials

► Field Electron Emission from Nanomaterials

Cold-Wall Thermal Chemical Vapor Deposition

► Chemical Vapor Deposition (CVD)

Compliant Mechanisms

Larry L. Howell
Department of Mechanical Engineering, Brigham Young University, Provo, UT, USA

Synonyms

Compliant systems; Flexures; Flexure mechanisms

Definition

Compliant mechanisms gain their motion from the deflection of elastic members.

Main Text

Compliant mechanisms offer an opportunity to achieve complex motions within the limitations of micro- and nano-fabrication. Because compliant mechanisms gain their motion from the constrained bending of flexible parts, they can achieve complex motion from simple topologies. Traditional mechanisms use rigid parts connected at articulating joints (such as hinges, axles, or bearings), which usually requires assembly of components and results in friction at the connecting surfaces [1–3]. Because traditional bearings are not practical and lubrication is problematic, friction and wear present major difficulties.

Nature provides an example of how to effectively address problems with motion at small scales. Most moving components in nature are flexible instead of stiff, and the motion comes from bending the flexible parts instead of rigid parts connected with hinges (for example, consider hearts, elephant trunks, and bee wings). The smaller the specimen, the more likely it is to use the deflection of flexible components to obtain its motion. And so it is with man-made systems as well, the smaller the device, the greater the advantages for using compliance [1].

Advantages of Compliant Mechanisms

Some of the advantages of compliant mechanisms at the micro- and nanoscales include the following:

Can be made from one layer of material. Compliant mechanisms can be fabricated from a single layer. This makes them compatible with many common microelectromechanical system (MEMS) fabrication methods, such as surface micromachining, bulk micromachining, and LIGA. For example, consider the folded beam suspension shown in Fig. 1. This device is often used as a suspension element in MEMS systems. It offers a simple approach for constrained linear motion, and also integrates a return spring function. The device can achieve large deflections with reasonable off-axis stiffness. The compliant mechanism makes it possible to do these functions with a single layer of material.

No assembly required. Compliant mechanisms that gain all of their motion from the deflection of flexible components are "fully compliant mechanisms," where devices that combine both traditional and compliant elements are called "partially compliant mechanisms." Fully compliant mechanism can usually be fabricated without assembly of different components.

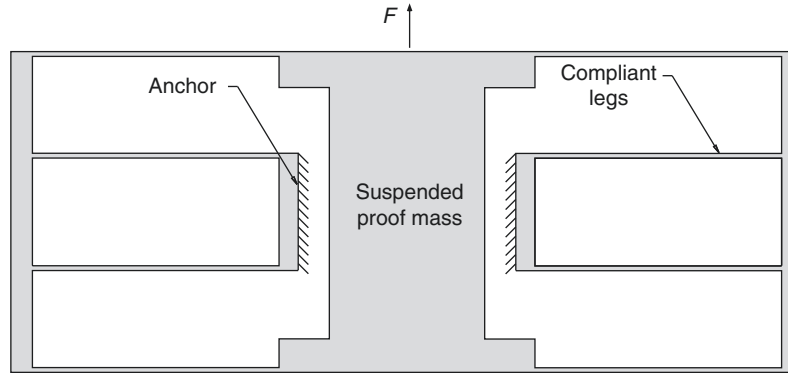
Small footprint. Some compliant mechanisms can also be designed to have a small footprint on the substrate on which they are built. Various strategies can be used to decrease the size of a mechanism. Figure 2 shows a thermal actuator that uses multiple layers to achieve a small footprint.

Friction-free motion. Because compliant mechanisms gain their motion from deflection of flexible members rather than from traditional articulating joints, it is possible to reduce or eliminate the friction associated with rubbing surfaces. This results in reduced wear and eliminates the need for lubrication, as described next.

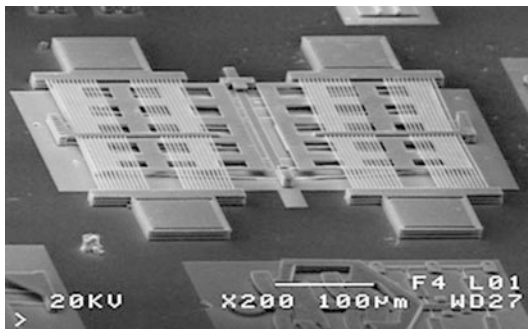
Wear-free motion. Wear can be particularly problematic at small scales, and the elimination of friction can result in the elimination of wear at the connecting surfaces of joints. For devices that are intended to undergo many cycles of motion, eliminating friction can dramatically increase the life of the system.

No need for lubrication. Another consequence of eliminating friction is that lubricants are not needed for the motion. This is particularly important at small scales where lubrication can be problematic.

High precision. Flexures have long been used in high precision instruments because of the repeatability of their motion. Some reasons for compliant mechanisms' precision are the



Compliant Mechanisms, Fig. 1 A folded-beam suspension is an example of a widely used compliant mechanism in microelectromechanical systems (MEMS) applications

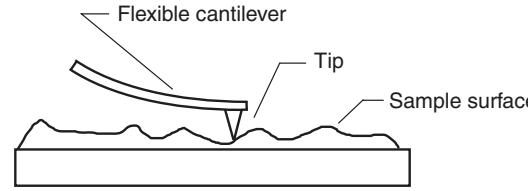


Compliant Mechanisms, Fig. 2 This scanning electron micrograph shows a thermal actuator that uses multiple layers of compliant elements to achieve large amplification with a small footprint

backlash-free motion inherent in compliant mechanisms and the wear-free and friction-free motion described above. The cantilever associated with an atomic force microscope (Fig. 3) is an example application.

Integrated functions. Like similar systems in nature, compliant mechanisms have the ability to integrate multiple functions into few components. For example, compliant mechanisms often provide both the motion function and a return-spring function. Thermal actuators are another example of integration of functions, as described later.

High reliability. The combination of highly constrained motion of compliant mechanisms, the relative purity of materials used in micro/nanofabrication, and wear-free motion result in high reliability of compliant mechanisms at the micro/nanoscale.



Compliant Mechanisms, Fig. 3 The cantilever of an atomic force microscope (AFM) is an example of compliance employed in high-precision instruments

Challenges of Compliant Mechanisms

Compliant mechanisms have many advantages, but they also have some significant challenges. A few of these are discussed below [1].

Limited rotation. One clear drawback of compliant mechanisms is the general inability to undergo continuous rotation. Also, if a fully compliant mechanism is constructed from a single layer of material, then special care has to be taken to ensure that moving segments of the compliant mechanism do not collide with other segments of the same mechanism.

Dependence on material properties. The performance of compliant mechanisms is highly dependent on the material properties, which are not always well known.

Nonlinear motions. The deflections experienced by compliant mechanisms often extend beyond the range of linearized beam equations. This can make their analysis and design more complicated.

Fatigue analysis. Because most compliant mechanisms undergo repeated loading, it is

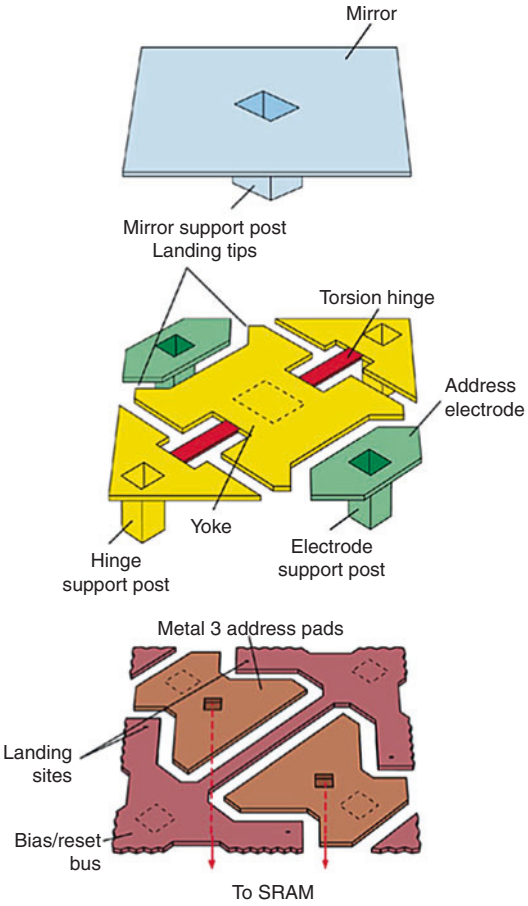
important to consider the fatigue life of the device. Interestingly, because of the types of materials used and their purity, many MEMS compliant mechanisms will either fail on their first loading cycle or will have infinite fatigue life. Because of the low inertia of MEMS devices, it is often easy to quickly test a MEMS device to many millions of cycles. Factors such as stress concentrations, the operating temperature, and other environment conditions can affect the fatigue life.

Difficult design. Integration of functions into fewer components, nonlinear displacements, dependence on material properties, the need to avoid self-collisions during motion, and designing for appropriate fatigue life, all combine to make the design of compliant mechanisms nontrivial and often difficult.

Example Applications of Compliant Mechanisms

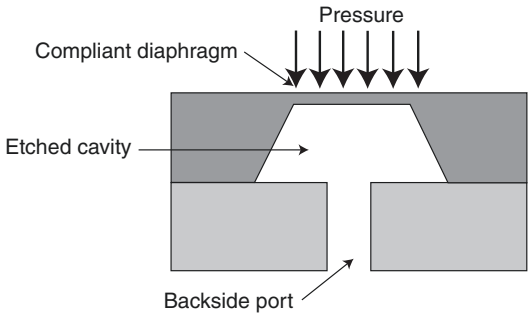
Examples of MEMS compliant mechanisms are shown here to further illustrate their properties and to demonstrate a few applications.

Digital Micromirrors. One of the most visible commercially available microelectromechanical systems is Texas Instruments' Digital Micromirror Device (DMD™), which is used in applications such as portable projectors. The DMD is a rectangular array of moving micromirrors that is combined with a light source, optics, and electronics to project high-quality color images. Figure 4 shows the architecture of a single DMD pixel. A 16-μm-square aluminum mirror is rigidly attached to a platform (the "yoke"). Flexible torsion hinges are used to connect the yoke to rigid posts. An applied voltage creates an electrostatic force that causes the mirror to rotate about the torsion hinges. When tilted in the on position, the mirror directs light from the light source to the projection optics and the pixel appears bright. When the mirror is tilted in the off position, the light is directed away from the projection optics and the pixel appears dark. The micromirrors can be combined in an array on a chip, and each micromirror is associated with the pixel of a projected image. The torsion hinges use compliance to obtain motion while avoiding rubbing parts that cause friction and wear. The hinges can be deflected thousands of times per second and infinite fatigue life is essential.

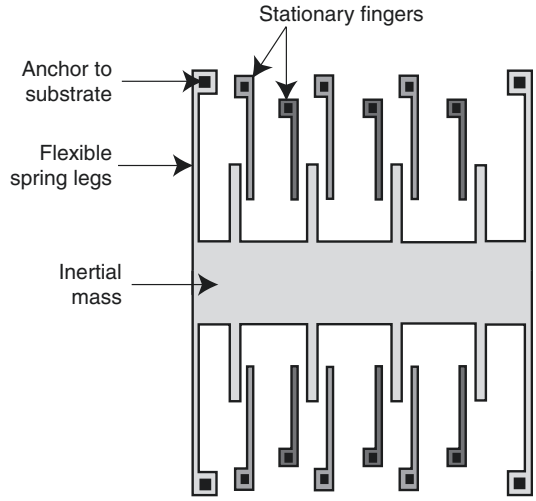


Compliant Mechanisms, Fig. 4 Texas Instrument's Digital Micromirror Device (DMD™) uses compliant torsion hinges to facilitate mirror motion (Illustration courtesy of Texas Instruments)

Piezoresistive pressure sensors. A sensor is a device that responds to a physical input (such as motion, radiation, heat, pressure, magnetic field), and transmits a resulting signal that is usually used for detection, measurement, or control. Advantages of MEMS sensors are their size and their ability to be more closely integrated with their associated electronics. Piezoresistive sensing methods are among the most commonly employed sensing methods in MEMS. Piezoresistance is the change in resistivity caused



Compliant Mechanisms, Fig. 5 The strain on a compliant diaphragm of a piezoresistive pressure sensors results in a detectable change in resistance, which is correlated with the pressure



Compliant Mechanisms, Fig. 6 This accelerometer makes use of compliant legs that deflect under inertial loads. The deflection results in a detectable change in capacitance and is correlated with the corresponding acceleration

by mechanical stresses applied to a material. Bulk micromachined pressure sensors have been commercially available since the 1970s. A typical design is illustrated in Fig. 5. A cavity is etched to create a compliant diaphragm that deflects under pressure. Piezoresistive elements on the diaphragm change resistance as the pressure increases; this change in resistance is measured and is correlated with the corresponding pressure.

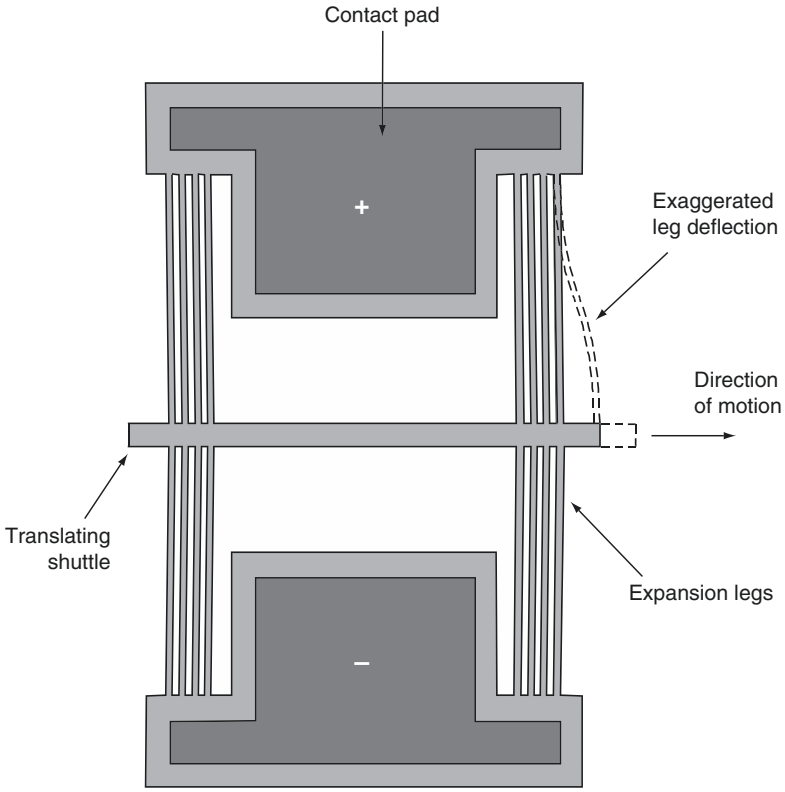
Capacitive acceleration sensors. Accelerometers are another example of commercially successful MEMS sensors. Applications include automotive airbag safety systems, mobile electronics, hard drive protection, gaming, and others. Figure 6 illustrates an example of a surface micromachined capacitive accelerometer. Acceleration causes a displacement of the inertial mass connected to the compliant suspension, and the capacitance change between the comb fingers is detected.

Thermal actuators. A change in temperature causes an object to undergo a change in length, where the change is proportional to the material's coefficient of thermal expansion [4]. This length change is usually too small to be useful in most actuation purposes. Therefore, compliant mechanisms can be used to amplify the displacement of thermal actuators. Figure 7 illustrates an example of using compliant mechanisms to amplify thermal expansion in microactuators. Figure 8 shows a scanning electron micrograph of a thermomechanical in-plane microactuator (TIM) illustrated in Fig. 7. It consists of thin legs connecting both sides of a center shuttle. The leg ends not connected to the shuttle are anchored to bond pads on the substrate and are fabricated at a slight angle to bias motion in the desired direction. As voltage is applied across the bond pads, electric current flows through the thin legs. The legs have a small cross-sectional area and thus have a high electrical resistance, which causes the legs to heat up as the current passes through them. The shuttle moves forward to accommodate the resulting thermal expansion. Advantages of this device include its ability to obtain high deflections and large forces, as well as its ability to provide a wide range of output forces by changing the number of legs in the design.

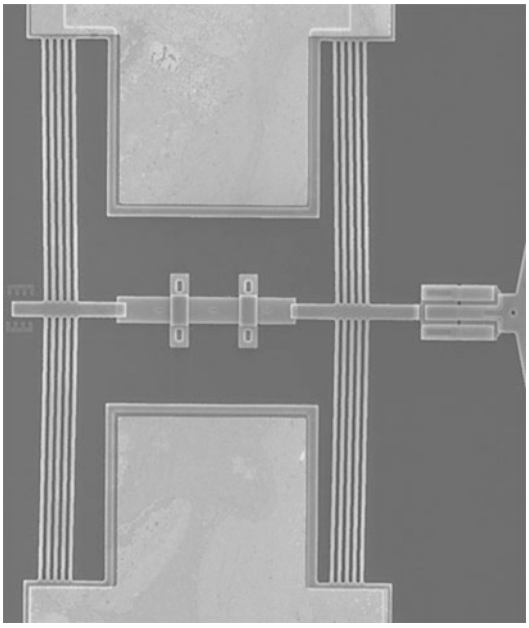
Analysis and Design of Compliant Mechanisms

Multiple approaches are available for the analysis and design of compliant mechanisms. Three of the most developed approaches are described below.

Finite element analysis. Finite element methods are the most powerful and general methods available to analyze compliant mechanisms.



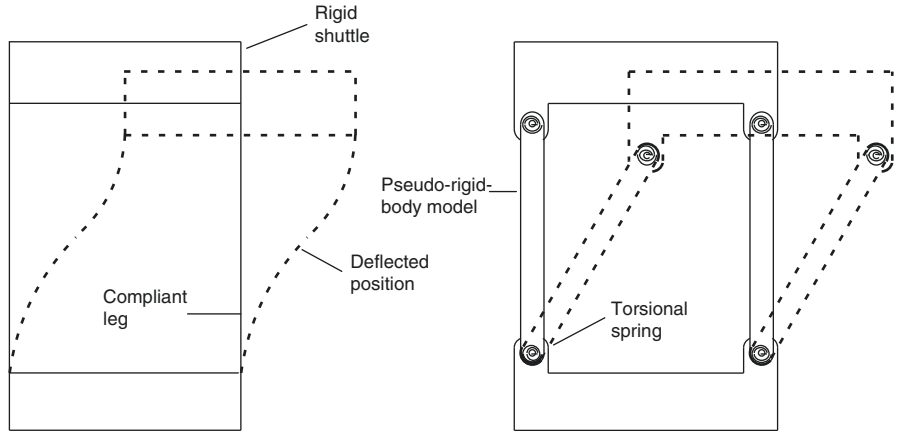
Compliant Mechanisms, Fig. 7 A schematic of a thermomechanical in-plane microactuor (TIM) that uses compliant expansion legs to amplify the motion caused by thermal expansion



Compliant Mechanisms, Fig. 8 A scanning electron micrograph of a thermal actuator illustrated in Fig. 7

Commercial software is currently available that has the capability of analyzing the large, nonlinear deflections often associated with compliant mechanisms. The general nature of the method makes it applicable for a wide range of geometries, materials, and applications. Increasingly powerful computational hardware has made it possible to analyze even very complex compliant mechanisms. It is also possible to use finite element methods in the design of compliant mechanisms, particularly once a preliminary design has been determined. But in the early phases of design, other methods (or hybrid methods) are often preferred so that many design iterations can be quickly analyzed.

Pseudo-rigid-body model. The pseudo-rigid-body model is used to model compliant mechanisms as traditional rigid-body mechanisms, which opens up the possibility of using the design and analysis methods developed for rigid-body mechanisms in the design of compliant mechanisms [1]. With the



Compliant Mechanisms, Fig. 9 The pseudo-rigid-body model of the compliant parallel-guiding mechanism consists of appropriately located pin joints and torsional springs (this device is a building block of other devices, such as the folded-beam suspension)

pseudo-rigid-body model approach, flexible parts are modeled as rigid links connected at appropriately placed pins, with springs to represent the compliant mechanism's resistance to motion. Extensive work has been done to develop pseudo-rigid-body models for a wide range of geometries and loading conditions. Consider a simple example. The micromechanism shown in Fig. 9 has a rigid shuttle that is guided by two flexible legs. (Note that the folded-beam suspension in Fig. 1 has four of these devices connected in series and parallel.) The pseudo-rigid-body model of the mechanism models the flexible legs as rigid links connected at pin joints with torsional springs. Using appropriately located joints and appropriately sized springs, this model is very accurate well into the nonlinear range. For example, if the flexible legs are single-walled carbon nanotubes, comparisons to molecular simulations have shown the pseudo-rigid-body model to provide accurate results [5]. The advantages of the pseudo-rigid-body model are realized during the early phases of design where many design iterations can be quickly evaluated, traditional mechanism design approaches can be employed, and motions can be easily visualized.

Topology optimization. Suppose that all that is known about a design is the desired performance and design domain. Topology optimization shows promise for designing compliant mechanisms under such conditions. The advantage is that very little prior knowledge about the resulting compliant mechanism is needed, and any biases of the designer are eliminated [6]. Topology optimization is often integrated with finite element methods to consider many possible ways of distributing material with the design domain. This has the potential to find designs that would not otherwise be discovered by other methods. Infinite possible topologies are possible and finite element methods can be employed to evaluate the different possibilities. The resolution of the design domain mesh can be a limiting factor, but once a desirable topology is identified, it can be further refined using other approaches.

Conclusion

Compliant mechanisms provide significant benefits for micro- and nano-motion applications. They can be compatible with many fabrication methods, do not require assembly, have friction-free and wear-free motion, provide high precision and high reliability, and they can integrate multiple functions into fewer components. The major challenges associated with compliant mechanisms come from the difficulty associated with their design, limited rotation, and the need to ensure adequate fatigue life. It is likely that compliant mechanisms will see increasing use in micro- and

nano-mechanical systems as more people understand their advantages and have tools available for their development.

Cross-References

- ▶ AFM
- ▶ Basic MEMS Actuators
- ▶ Biomimetics
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Insect Flight and Micro Air Vehicles (MAVs)
- ▶ MEMS on Flexible Substrates
- ▶ Nanogrippers
- ▶ Piezoresistivity
- ▶ Thermal Actuators

References

1. Howell, L.L.: Compliant Mechanisms. Wiley, New York (2001)
2. Lobontiu, N.: Compliant Mechanisms: Design of Flexure Hinges. CRC Press, Boca Raton (2003)
3. Smith, S.T.: Flexures: Elements of Elastic Mechanisms. Taylor & Francis, London (2000)
4. Howell, L.L., McLain, T.W., Baker, M.S., Lott, C.D.: Techniques in the design of thermomechanical microactuators. In: Leondes, C.T. (ed.) MEMS/NEMS Handbook, Techniques and Applications, pp. 187–200. Springer, New York (2006)
5. Howell, L.L., DiBiasio, C.M., Cullinan, M.A., Panas, R., Culpepper, M.L.: A pseudo-rigid-body model for large deflections of fixed-clamped carbon nanotubes. J. Mech. Robot. **2**, 034501 (2010)
6. Frecker, M.I., Ananthasuresh, G.K., Nishiwaki, S., Kikuchi, N., Kota, S.: Topological synthesis of compliant mechanisms using multi-criteria optimization. J. Mech. Des. **119**, 238–245 (1997)

Compliant Systems

- ▶ Compliant Mechanisms

Composite Materials

- ▶ Theory of Optical Metamaterials

Computational Chemistry for Drug Discovery

Giulia Palermo[1,2] and Marco De Vivo[1]
[1]Department of Drug Discovery and Development - CompuNet, Istituto Italiano di Tecnologia, Genoa, Italy
[2]Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Synonyms

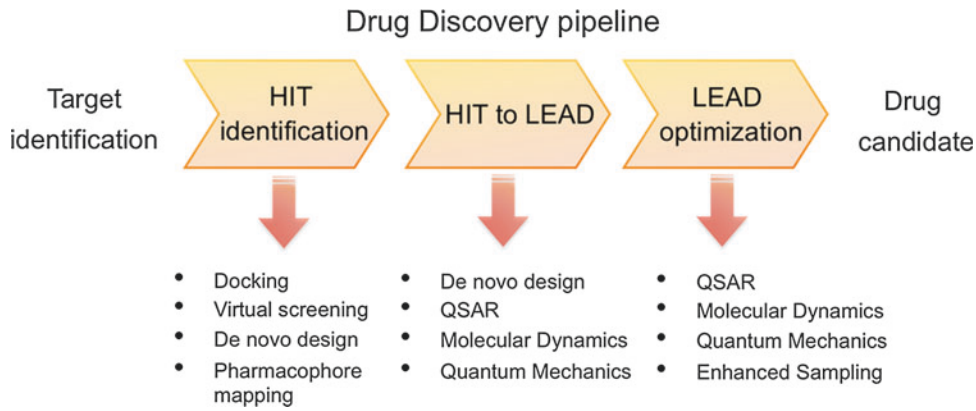
Drug design; Molecular modeling

Definition

Computational chemistry uses physics-based algorithms and computers to simulate chemical events and calculate chemical properties of atoms and molecules. In drug design and discovery, diverse computational chemistry approaches are used to calculate and predict events, such as the drug binding to its target and the chemical properties for designing potential new drugs.

Overview

Computational methods are nowadays routinely used to accelerate the long and costly drug discovery process. Typically, once the drug discovery target is selected, drug discovery activities are divided into those for (1) the *hit identification* phase, in which the aim is the identification of chemical compounds with a promising activity toward the target; (2) the *lead generation* phase, in which hit compounds are improved in potency against the target; and, finally, (3) the *lead optimization* phase, in which lead compounds are optimized, generating drug-like molecules ultimately able to exert their beneficial pharmacological effect in patients (Fig. 1). Computations can help in all these drug discovery activities, from drug target identification, commonly a receptor



Computational Chemistry for Drug Discovery, Fig. 1 Drug discovery pipeline, typically formed by three phases, namely, "hit identification," "hit to lead," and "lead optimization." Different computational approaches can be applied to each phase of the drug discovery pipeline

or an enzyme, to the design and optimization of a new drug-like compound. While computational methods for target identification rely mainly on computational sciences such as bioinformatics and computational genomics, different computational approaches are used once the target has been identified and the search for small molecule inhibitors has commenced, starting from the hit identification phase, moving toward the lead generation and optimization phases (Fig. 1). At that point, routinely applied computational chemistry approaches include methods for structure-based drug design (SBDD), when structural data of the target protein are available, and ligand-based drug design (LBDD), when structural information of the target is missing or not fully reliable. Overall, these methods facilitate the identification of promising chemical scaffolds that interfere favorably with the target's function, producing a positive pharmacological effect. Experimental biochemical and pharmacological data on the new compounds, such as their in vitro inhibitory potency and in vivo efficacy, can be used to check the computational predictions while also forming the basis upon which better models can be constructed, leading to the design of superior compounds [1].

The impact of computational chemistry on drug discovery has been intensified in the last few decades by the rapid development of faster architectures and better algorithms for time-affordable high-level computations. Several theoretical methods that were once prohibitive for effective drug discovery are now increasingly used for hit identification and lead generation. Recently, for example, CPU-intensive and GPU-based free energy perturbation (FEP) calculations have been applied to accurately estimate the binding free energy of closely related chemical analogs, generating very promising results for lead generation and optimization [2]. Also, classical molecular dynamics (MD) is currently proposed as a practical computational tool for studying the energetics and kinetics of a ligand binding to a target protein. This is relevant for lead optimization, representing a new frontier in computationally driven drug discovery. Recently, in fact, compounds have been evaluated not only for their ability to bind tightly to the target but also for their capacity to remain bound to the target for a long time (i.e., considering k_{on} and k_{off} of binding), increasing the chances of efficacy in vivo. Finally, in the last decade, quantum mechanics (QM) has become ever more accessible for performing SBDD for lead generation and optimization. For example, QM and hybrid QM/MM methods are increasingly used to study the interaction of covalent inhibitors with a drug discovery target.

It is challenging to tailor a perfect fit between a new compound and its target in order to generate potent inhibitory effects (i.e., high affinity), which is the goal during the lead generation phase.

However, there are several other challenges during the lead optimization phase. A potent inhibitor is not a drug. There are other physicochemical variables that dictate the pharmacokinetics (PK) of each compound, affecting their drug-likeness and, ultimately, their efficacy and safety in vivo. Absorption, distribution, metabolism, excretion, and toxicity (ADMET) are key parameters that need to be optimized to generate a drug candidate with good chances of success in clinical trials. ADMET prediction at early stages of the drug discovery process is key to preventing, or at least limiting, later failures in costly clinical trials. In this respect, computational methods for chemometrics and quantitative structure–activity relationship (QSAR) approaches play a prominent role in creating predictive models for selecting and prioritizing compounds, typically during the lead optimization phase.

Thus, each computational chemistry method can impact and accelerate a given phase of the drug discovery process, from docking and MD for hit identification and lead generation to QSAR for ADMET optimization (see Fig. 1). Detailed methodological descriptions of each method can be found in many excellent review articles and books that focus on the theoretical background of computational chemistry [3–5]. This essay aims instead to comprehensively outline the applicability of the computational chemistry methods and approaches used nowadays to accelerate drug design and discovery, with particular emphasis on SBDD. The point is to show how each computational approach suits better a certain phase of the drug discovery pipeline, from SBDD for the hit identification and lead generation phases to QSAR methods for lead optimization, where drug-like properties are tuned to generate a promising drug candidate. The everyday use of once-prohibitive computational methods, such as MD- and QM-based methods for SBDD, will also be highlighted.

Computational Methods for SBDD

Computational approaches to structure-based drug design (SBDD) rely on knowledge of the target protein structure, which is usually provided by high-resolution X-ray crystallography or NMR data. Through a detailed analysis of the interaction between the target structure and the (new) ligands, SBDD approaches allow informed decisions to be made in designing more potent (i.e., with high affinity for the target) and selective compounds. Therefore, these methods are mostly used for hit identification and during the hit-to-lead phase (Fig. 1). A wide range of computational chemistry approaches can be applied to SBDD, including force field-based methods such as molecular docking calculations, classical MD- or Monte Carlo (MC)-based simulations, and more sophisticated QM-based methods [1, 6]. Ultimately, the new rationally designed compounds must be experimentally evaluated to verify whether they are potent inhibitors. The experimental test of each compound demonstrates the interdisciplinary nature of effective drug discovery. It is also essential for establishing and evaluating the predictive power of computational approaches in identifying and generating promising drug candidates.

Force Field-Based Approaches for SBDD

Computational approaches based on molecular mechanics (MM) allow the energy and several properties of molecular systems to be computed [3]. The basic functional form of a force field describes the potential energy of the system, which is determined as the sum of different contributions that are parameterized to reproduce experimental or ab initio data. The interactions are divided into bonded interactions and nonbonded interactions. The typical force field equation (Chart 1) contains the terms for the bonded (highlighted in red) and nonbonded (highlighted in blue) interactions. In detail, the bonded terms describe the chemical bonds between two neighboring atoms, bond angles between three atoms, and dihedral angles between four atoms. Improper dihedral–angle terms can be additionally applied to maintain planar or tetrahedral conformations. The functional form of the bond and angle terms is

Computational Chemistry for Drug Discovery, Chart 1 Typical force field equation

$$V = \sum_{\text{bonds}} K_R(\mathrm{R} - \mathrm{R}_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{\mathrm{R}_{ij}^{12}} - \frac{B_{ij}}{\mathrm{R}_{ij}^{6}} + \frac{q_i q_j}{\varepsilon \mathrm{R}_{ij}}\right]$$

bonded interactions

non-bonded interactions

Lennard-Jones term (van der Waals interactions)

Coulomb term (electrostatic interactions)

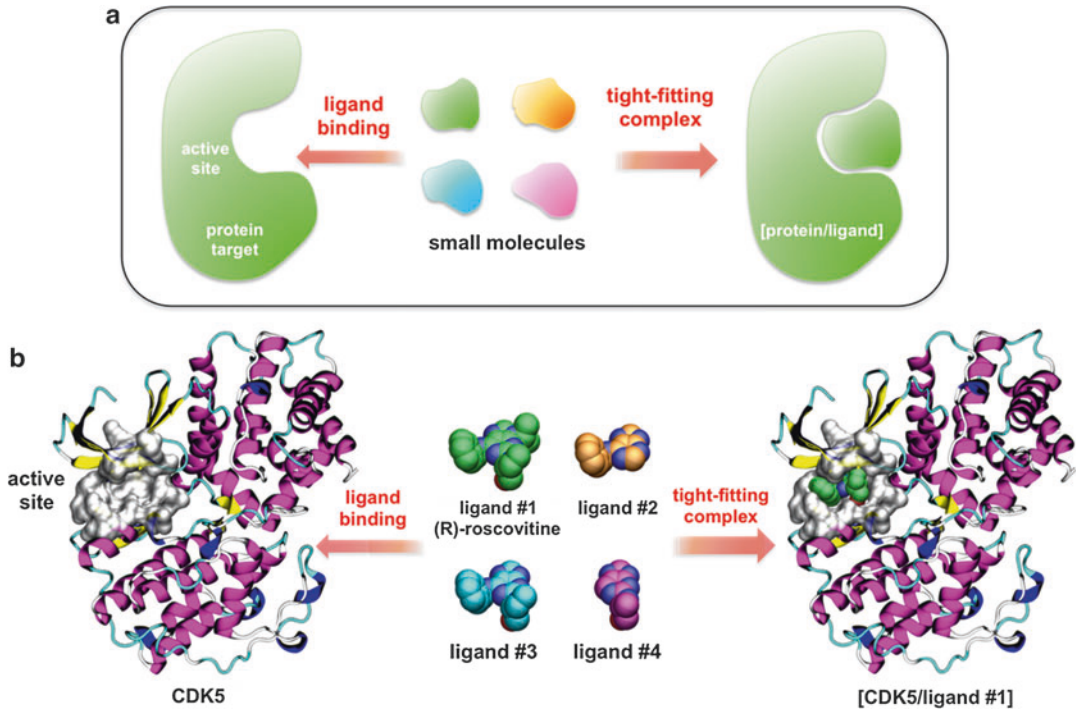
$$V_{vdW} = \sum_{i<j}\left[\frac{A_{ij}}{\mathrm{R}_{ij}^{12}} - \frac{B_{ij}}{\mathrm{R}_{ij}^{6}}\right] \quad + \quad V_{el} = \sum_{i<j}\left[\frac{q_i q_j}{\varepsilon \mathrm{R}_{ij}}\right]$$

quadratic, while the dihedral term uses a trigonometric function. In the typical force field equation, **R** is the distance between two atoms i and j that are bound together via a covalent bond, θ is the bond angle, $\mathbf{R}_{\text{eq}}$ and θ_{eq} refer to equilibrium bond lengths and angles, and K_R and K_θ are the vibrational constants. V_n is the torsional barrier corresponding to the nth barrier of a given torsional angle with phase γ. The last term of the typical force field equation refers to the nonbonded interactions, which are composed of a Lennard-Jones term for the van der Waals interactions and a Coulomb term for the electrostatic interactions between atoms i and j.

Force fields like AMBER, OPLS, GROMOS, and CHARMM are extensively used to study standard biomolecular systems such as protein, DNA, and RNA. Corrections and extensions are recurrently developed and reported for treatment of carbohydrates, lipids, ions, and other molecules of relevant biological interest, including drugs. Particularly relevant for drug design are the generalized AMBER force field (GAFF) and CHARMM general force field (CgFF), which provide suitable parameters for computations of drug-like organic molecules. The main advantage of using force field-based approaches is their relatively low computational cost compared to QM-based methods. For this reason, force field-based methods are by far the most applied in computational drug discovery. These include molecular docking and virtual screening of millions of novel compounds to identify new inhibitors, as well as de novo design and classical molecular dynamics simulations (see below).

Virtual Screening and Molecular Docking for Drug Discovery

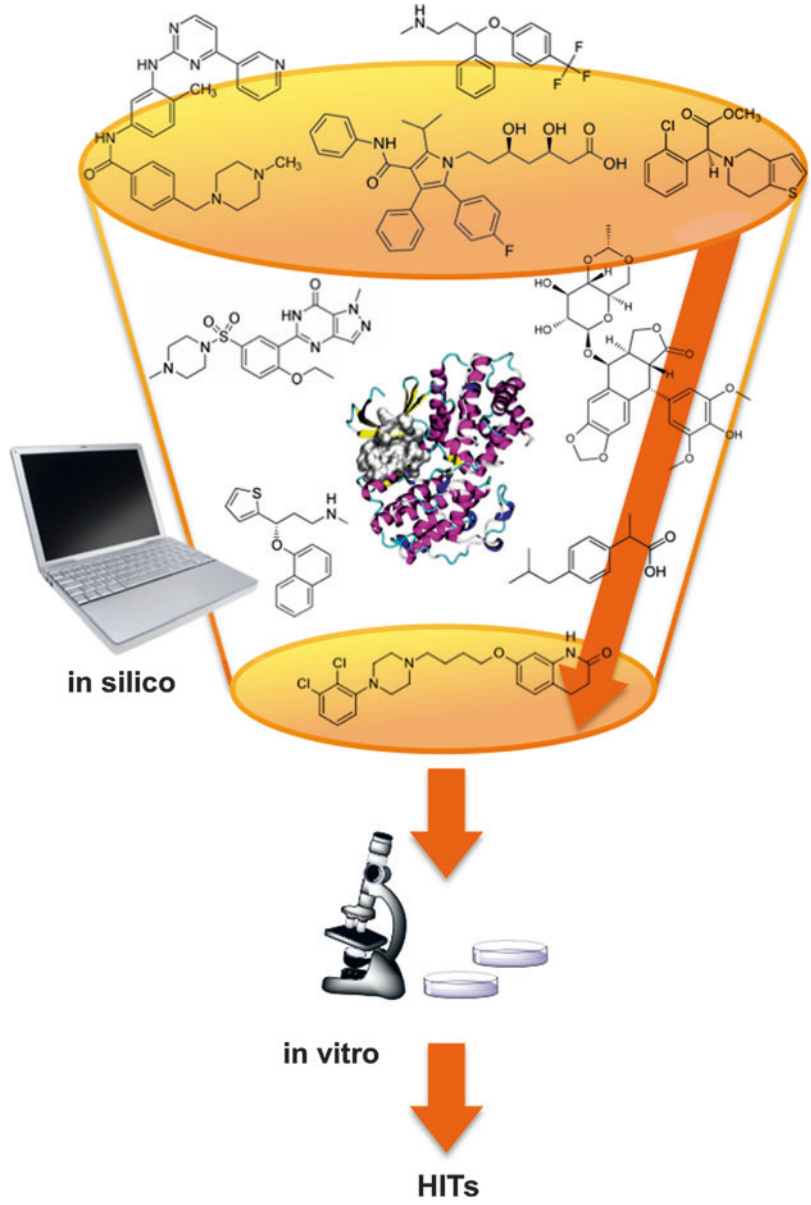
Virtual screening (VS) and molecular docking are broadly applied approaches in computational SBDD [1]. The ultimate goal of these methods is to identify compounds that dock well inside a targeted pocket (Fig. 2). These computational approaches are thus mostly used to identify promising hits and generate new leads. A compound able to fit into a target pocket would likely interfere with the target function, representing a promising chemical scaffold for the design of a new drug. GLIDE, GOLD, AUTODOCK, DOCK, and ICM-Dock are some of the more popular docking software. They are all used to increase the chances of identifying novel bioactive molecules from among the many small molecules that can be (virtually) screened. Starting from large compound collections, which can contain millions of different small molecules, VS protocols filter those collections and reduce them into smaller sets of active compounds with promising predicted activity against the target (Fig. 3). Usually, a library of hundreds of thousands or millions of virtual compounds is screened against a selected target. Then, a more sophisticated docking protocol, for example, a protocol that includes some receptor flexibility, can be used to rescore the best few thousand scored compounds. It is usually advisable to visually inspect the highly scored compounds, looking at each pose. This step reduces the risk of false-positive results, often induced by highly scored compounds that bind by adopting unlikely conformations or bind far from the pocket region of interest. After visual inspection, a few compounds are selected for



Computational Chemistry for Drug Discovery, Fig. 2 (**a**) Scheme of the ligand binding of a small molecule to the target proteins. (**b**) Binding of the tight-fitting ligand #1 [(R)-roscovitine] to the cyclin-dependent kinase-5 (CDK5) [23]. Other ligands (#2 to #4) are also shown. CDK5 is an attractive pharmacological target for drug discovery. CDK5 deregulation is implicated in many neurodegenerative processes such as Alzheimer's and Parkinson's diseases and amyotrophic lateral sclerosis. CDK5 is depicted with ribbons. The protein active site is represented in molecular surface, while a series of small ligand analogs with different chemical features – ligands #1–4 – are shown in space-filling representation. Different colors highlight the different characteristics of the four ligands

in vitro testing. VS is thus typically used for hit identification during drug discovery, where the goal is to identify new chemical scaffolds with promising, i.e., weak, inhibitory activity against the target. Hits are then used as a starting point to develop leads, i.e., more potent inhibitors (Fig. 1). Usually, hit compounds are experimentally identified as inhibitors characterized by IC_{50} (i.e., the concentration of inhibitor required to block 50 % of the target activity) around ~10–100 or higher microM activity, while lead compounds have IC_{50} in the low nanoM range.

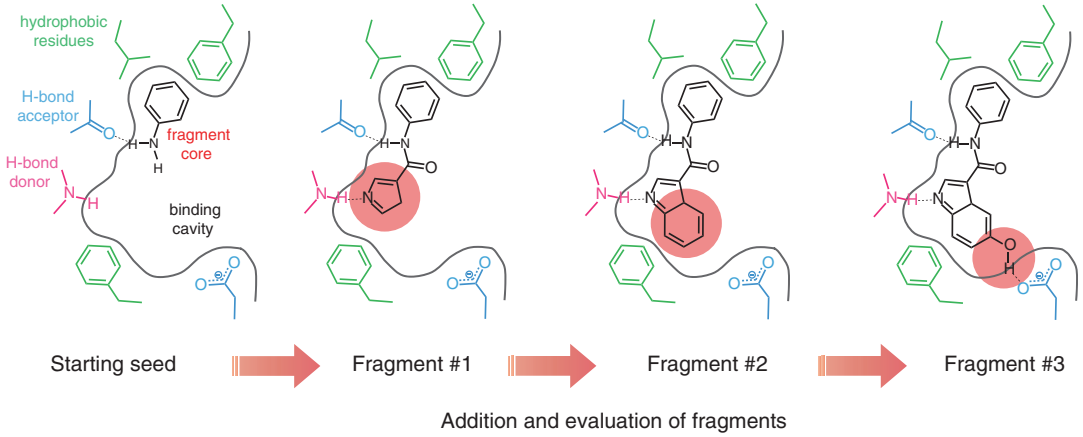
To predict and evaluate ligand–target interactions, molecular docking must correctly pose the ligand into the catalytic site and associate an energy score with each pose. There are two main steps during each docking calculation: the posing of the ligand into the pocket and the scoring of the final pose. To identify the possible binding poses, a conformational sampling of the ligand in the target active site is performed for each docking run. Each ligand pose is generated by exhaustive conformational searches performed by systematic or deterministic searches of rotatable bonds or by mapping the target/ligand geometric complementarities. More sophisticated search methods employ stochastic searches such as MC sampling procedures or evolutionary algorithms. Then, each final pose is scored, according to the chosen scoring function, and in VS, compounds are ranked according to their relative final energy score. Thus, a good docking pose is characterized by a favorable score, which should reflect and match experimental data, if available. The scoring step is thus as crucial as the posing step, where the ligand is accommodated into the binding site.



Computational Chemistry for Drug Discovery, Fig. 3 A typical virtual screening protocol starts with a large compound collection, which is virtually screened against a given target. Then, the most promising ligands are tested in vitro, leading to the identification of hit compounds (characterized by a concentration for 50 % target inhibition (IC_{50}) of ~100–10 μM). Ultimately, hit compounds will be transformed into lead molecules, which are more potent inhibitors that will be also tested in vivo for efficacy and safety

Although posing and scoring are often considered as independent problems, they are actually interconnected. This becomes clear when we consider that the primary criterion for the conformational selection of the bound ligand is its energy score, which is used to rank and select ligand poses among all the possible binding modes. Nowadays, there are several functions for scoring ligands, from force field-based scoring functions to knowledge-based potentials. Each scoring function has its own peculiarities and drawbacks. Care is therefore required in their selection and initial benchmarking, before embarking in a VS campaign. At times, multiple scoring functions are used to generate a consensus score, as the average or linear combination of single scoring function.

The challenges in achieving a correct score in VS and docking calculations include a proper treatment of electrostatics, as well as an exact



Computational Chemistry for Drug Discovery, Fig. 4 The de novo design of ligands is performed by building each compound within the active site. On the basis of an X-ray structure of the target protein, compounds are built within the binding cavity, using a number of fragments that can be joined together to create a new drug-like compound. Selected interaction centers are shown in green (hydrophobic residues), blue (H-bond acceptors), and magenta (H-bond donors). A starting fragment core (seed) is then grown into a bigger compound that can favorably interact with the target

evaluation of the desolvation energy of the ligand and binding pockets, and the explicit inclusion of entropic effects upon ligand binding. Recently, there has been intense investigation of the role of buried waters in the protein's pockets, which could act as a bridge for better ligand binding. A further challenge is that these methods consider a quasi-static complementarity between the ligand and its target. The restriction of target flexibility allows a limited induced fit upon binding but can potentially cause the miss of an important fraction of good hits/inhibitors during VS. To overcome this, recent docking algorithms and protocols allow multiple conformations of a few active-site residues that can be affected upon ligand binding, while other approaches try to solve the same issue via MD simulations or MC search [7]. Although rigid docking calculations are still preferred for VS of large compound libraries, flexible docking protocols are increasingly used nowadays, showing better predictivity than more static approaches.

Computational De Novo Design of Drug-Like Molecules

Computational de novo drug design means designing novel compounds from scratch, building them directly inside the binding site of the target macromolecule, piece by piece. As with docking and VS, the final goal is to explore chemical diversity and identify hit compounds that can fit well into the cavity of interest and likely interfere with the target function. De novo design programs include BOMB, BUILDER, GenStar, CONCERT, and several others, each with slightly different features and specifications [8]. The design process builds the new ligand directly within the binding pocket from an initial seed that can be as small as a hydrogen atom (Fig. 4). Then, ligand-growing and ligand-linking approaches are mainly used to transform one, or more, small fragments into a single drug-like compound with good predicted affinity for the target. With the ligand-growing method, functional groups are added to chemical handles of the core fragment, which is placed into the active site. Functional groups are usually available from an exhaustive list containing all sorts of aromatic and nonaromatic rings, aliphatic chains of varying length, etc. The user can therefore cherry pick the most interesting functional groups from the list and add those to the chemical handles of the core, generating new compounds. In contrast, ligand-linking methods use functional groups to

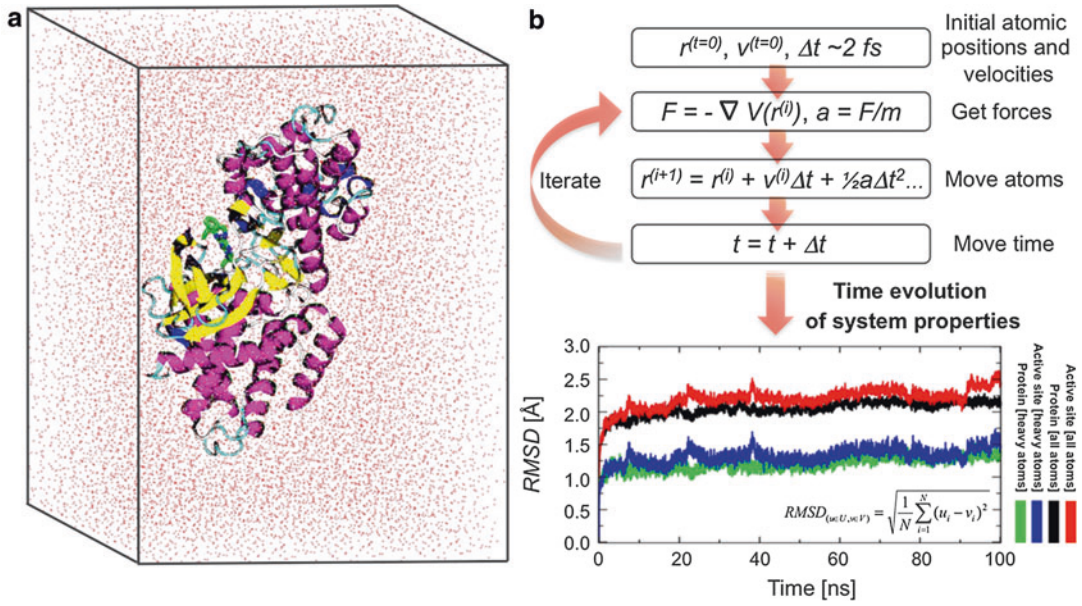
connect two or more fragments found or placed in the pocket, forming a single drug-like compound. In both procedures, the ligand construction is carried out by adding typical drug fragments that, once linked to the starting chemical core, are evaluated by considering several favorable conformers, usually followed by a geometry optimization of the newly formed compound within the target structure. Then, the obtained target/ligand complexes are scored in the same way and with the same limitations as scoring functions for VS and molecular docking.

The main advantage of de novo approaches, compared to docking and VS, is the possibility of exploring a virtually infinite search space and expanding chemical diversity. Drug-like molecules have been estimated to be in the order of 10^{60}–10^{100}. In this extremely large chemical space, the de novo design method allows the user to focus on chemical structures of interest, expanding on what is found in commercial libraries and converging on promising scaffolds that can be optimized ad hoc. The main risk is combinatorial explosion, due to the many combinations of functional groups and new compounds that could be constructed. In addition, the chemical space is limited by the synthesizability of the newly designed compounds, which might easily become too complex to be made in the lab. To overcome this, synthesizability constraints are often included in protocols for de novo computational design. Close collaboration between computational and synthetic chemists is also recommended. Ultimately, de novo molecular design methods help the hit identification phase by widening the chemical horizon for novel chemical scaffolds [2].

Molecular Dynamics for SBDD

Classical molecular dynamics (MD) simulations compute the time evolution of a system by applying Newtonian mechanics [3]. The method relies on force fields, which have been appropriately parameterized to describe biomolecular systems, such as the AMBER, OPLS-AA, and GROMOS force fields (see section above). These classical force fields provide a good description of the energetics of macromolecules such as proteins and nucleic acids. Also, in classical MD simulations of biological systems, there can be an explicit description of the water environment, with realistic molecular systems that can easily be constituted by a few hundred thousand atoms or more. In this respect, widely applied force fields for water molecules in biomolecular systems are the TIPnP models (Fig. 5).

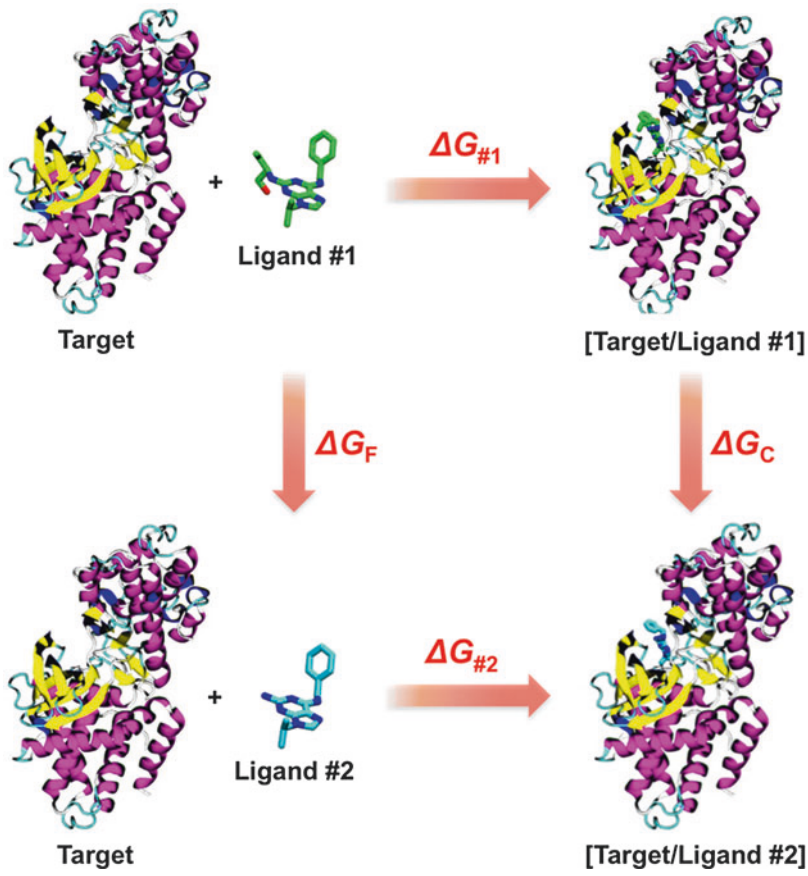
In drug design, classical MD simulations are mainly used to accurately calculate the binding free energy of the ligand to its target, explicitly taking into account the flexibility of the system, temperature, and entropic effects. MD is, however, much more computationally expensive than VS and docking. Therefore, MD is mostly recommended for, and so far applied to, lead optimization, where a promising compound is improved through small chemical modifications. MD simulations can suggest favorable chemical modifications through a detailed, although computationally demanding, understanding of the ligand–target interactions. MD simulations of realistic models of pharmaceutically relevant targets include the fatty acid amide hydrolase (FAAH) embedded into an explicit membrane environment [9, 10], the muscarinic acetylcholine receptors (mAChRs) and other GPCRs model systems bound to drugs, voltage-gated ion channels, and protein kinases, among others [11]. These types of studies elucidate the dynamical behavior of the ligand/target complex, highlighting key interactions that are likely responsible for drug binding and, ultimately, drug efficacy. In addition, the ability to simulate drug binding and unbinding events permits researchers to evaluate the residence time of the ligand into the binding pocket, which directly relates to the k_{on} and k_{off} of binding. The correct balance of these parameters is quite important in assuring a stable binding, which is at the basis of drug efficacy. Finally, the use of short MD runs as a post-processing and rescoring tool for docking results has been shown to improve the original docking score [7]. In this case, the docking binding pose relaxes, after introducing full flexibility into the system through MD, generating an optimized ligand/target complex that



Computational Chemistry for Drug Discovery, Fig. 5 (**a**) Typical MD model system, which shows a protein in explicit solvent. In this example, CDK5 is in complex with (R)-roscovitine, embedded in explicit water molecules [23]. (**b**) Scheme of MD simulation steps, which can provide information on the time evolution of certain selected system properties – e.g., the root mean square deviation (RMSD) expressed in Å of the overall protein and its active site, compared to the initial X-ray structure

leads to more accurate energy scores. Indeed, MD has lately been proposed, and preliminarily used, as a tool for screening medium-sized compound libraries (a few hundred or so compounds), with a protocol for dynamic docking that is expected to replace static VS in the near future [11].

Many biochemical events, such as diffusion events in the cell, can happen at long timescales spanning from micro- or milliseconds to seconds or even much longer. Those events are usually termed "rare" events. Often, these rare events require the overcoming of energetics barriers, which (in the case of ligand binding) are related to large protein conformational changes or ligand diffusion. That is, the biological phenomenon of ligand binding often consists of simulating a rare event. These simulations, and the accurate calculation of the associated free energy, require enhanced sampling methods, which can efficiently explore the free energy landscape of the rare biochemical event under investigation. These methods include umbrella sampling, steered MD, accelerated MD, thermodynamic integration, metadynamics, and others. They are often used in computational biophysics to study a ligand's binding/unbinding mechanism and other rare events [3]. A conventional method for the characterization of chemical events in computational biophysics is thermodynamics integration used to sample rare events by defining a path between the initial and final state. Then, thermodynamic parameters characterizing the system are changed gradually so that, at each stage, the system is in "constrained equilibrium." More simply, in the case of ligand binding, a suitable reaction coordinate can be chosen to discriminate between the bound and unbound states. Then, forces acting on this coordinate are monitored. The integration of the averaged forces along the binding path provides a good estimate of the free energy of the binding process. This method has been widely applied to numerous issues in biochemistry and biophysics,



Computational Chemistry for Drug Discovery, Fig. 6 Schematic representation of a thermodynamic cycle for the evaluation of the relative free binding energies of ligand #1 and ligand #2 to CDK5 [23]

including enzymatic chemical reactions [12–15], because it offers a rigorous evaluation of the free energy of the investigated process. With thermodynamic integration, one is essentially limited to compute the free energy landscape of one or two collective variables. In a more complex case, a more advanced method, e.g., metadynamics, is needed. Metadynamics allows the free energy landscape to be mapped by adding an external potential to the simulation, overcoming energetics barriers, and returning a thorough description of the explored free energy space. To restrict the virtually immense searchable configuration space, which is dictated by the degrees of freedom of the model systems, one must specify a restricted number of collective variables (i.e., the coordinates of interest) that properly describe the chemical event under study. In this way, for example, the path of the ligand, from the solvent to the target active site, can be simulated and examined, providing useful information on the dynamics and energetics of binding [16].

The free energy perturbation (FEP) method is a final example of free energy calculation methods that have successfully impacted SBDD. FEP calculations have demonstrated predictive power in calculating relative free binding energies of close analogs of active chemical scaffolds [2]. In this case, perturbations of each compound, which involve force field parameters, convert the initial ligand #1 into ligand #2, through a thermodynamic cycle (Fig. 6). The binding free energy difference between the ligands *#1* and *#2* is calculated as $\Delta\Delta G_{\text{bind}} = \Delta G_{\#1} - \Delta G_{\#2} = \Delta G_F - \Delta G_C$, where ΔG_F and ΔG_C are the energies associated with the unbound ligands in solution and with the target/ligand complexes, respectively. This allows the energetics of binding

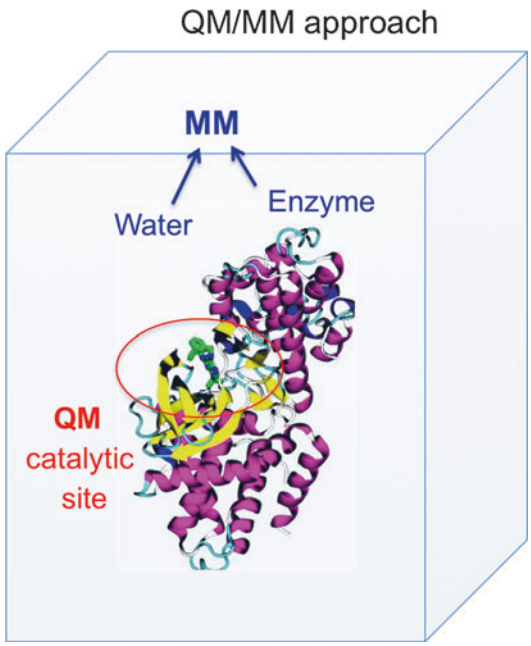
for the ligand *#1* and *#2* to be calculated, taking into full account the cost of the ligand's hydration. FEP theory has been broadly used for computational drug design purposes in combination with a Monte Carlo (MC) conformational search [17]. In this approach, a stochastic algorithm is used to randomly change the ligand within the target's active site. These random changes can also involve translational or rotational degrees of freedom of the ligand and of the residue side chains of the target. Whether a step is accepted or rejected is decided based on the Metropolis criterion, which generally accepts steps that lower the overall energy and occasionally accepts steps that increase energy. This prevents the system from getting stuck in local energy minima, allowing hopping over the energy barriers. As a drawback, an MC search lacks information about the timescale of the binding process.

All these methods and approaches can provide detailed mechanistic insights for discriminating between a few selected compounds that have successfully passed previous screening tests. Importantly, the continual development of more powerful processors and algorithms is allowing systems of increasing size and complexity to be studied. Protein folding, drug binding, membrane transport, and large conformational changes of proteins can nowadays be observed by performing long-timescale simulations, up to tens of microseconds, and more. For example, access to longer timescales has encouraged the study of binding kinetics. Calculations of the association and dissociation constants for the target/ligand complex – i.e., K_{on} and K_{off}, respectively – provide important information on the (ir)reversibility of the drug binding process, helping in the lead optimization phase. Overall, *state-of-the-art* MD simulations are nowadays a powerful SBDD tool.

Quantum Mechanics for SBDD

Quantum mechanics (QM) methods are the most advanced and computationally demanding methodologies, which are used to characterize the structure, dynamics, reactivity, and energetics of molecules. QM methods can be mainly classified as ab initio and density functional theory (DFT) methods. Ab initio QM methods aim to solve the Schrödinger equation, dealing with the wave function of the system. A variety of ab initio schemes have been proposed, from the simplest Hartree–Fock (HF) theory to the more sophisticated Møller–Plesset (MP) perturbation theory that includes correlation effects and applies a perturbation to the HF solution. In DFT, the central idea, proposed by the theorem of Hohenberg and Kohn in 1964, is that the ground-state energy of a system of interacting electrons is a unique functional of its electron density. Nowadays, calculations at the DFT level of theory are the method of choice for studying the electronic structure of realistic biochemical model systems, due to a rather balanced equilibrium between accuracy and computational affordability. DFT-based approaches in drug design are used to look at system properties that strongly depend on electronic structure, such as electrostatic potential maps used to characterize the binding site of the target or QM calculations to define the most probable protonation state of key residues located in the binding site. QM-derived scoring functions for better affinity evaluations in docking calculations is another field of application of QM for drug design. QM is also crucial for developing MM force fields to describe nonstandard residues, such as drug-like molecules [18].

The most common SBDD use of QM is the study of the catalytic and inhibitory mechanisms of pharmaceutically relevant enzymes. An atomic-level understanding of the inhibitory mechanism and its energetics can provide important information for improving existing drugs and designing new and more potent inhibitors. Indeed, QM-based investigation of enzymatic reaction mechanisms are linked to drug discovery through the design of inhibitors that resemble the structure and physicochemical properties of the enzymatic transition state (TS). This is because QM methods allow the study of bond-forming/breaking reactions, which characterize enzymatic catalysis and covalent inhibition. During new bond-forming events, polarization and charge transfer effects



Computational Chemistry for Drug Discovery, Fig. 7 QM/MM approach. A small part of the enzyme is treated at the QM level, while the remaining part of the system, i.e., water and protein, is treated at the MM level of theory

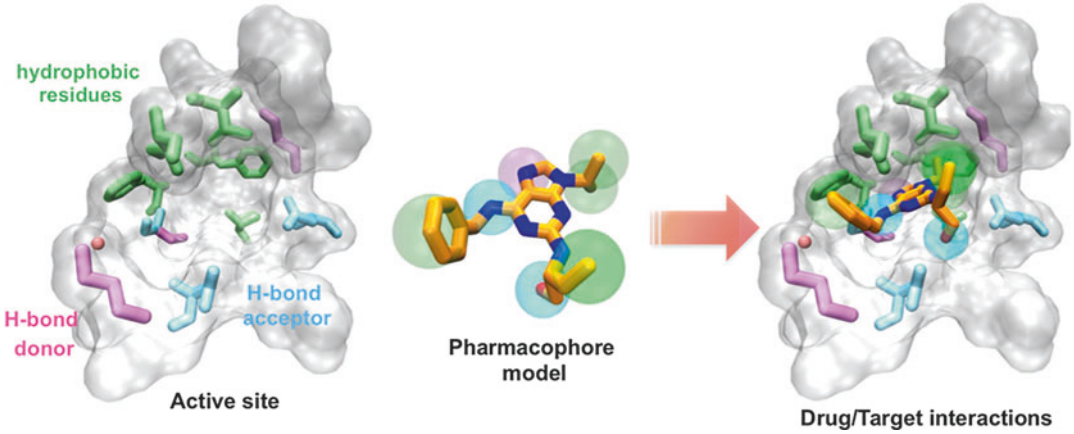
between the ligand and the active-site residues can be captured fully only via an explicit description of the electronic structure. Thus, QM-based methodologies allow the catalytic and inhibitory mechanisms to be deciphered in great detail. This information can ultimately facilitate and guide the design of better inhibitors, such as potent TS analogs [18].

Nevertheless, the applicability of most QM methods is limited to studying relatively small molecular systems in the gas phase, while biological systems in a realistic solution environment usually comprise thousands of atoms (~20,000 to ~100,000 atoms c.a.). This picture becomes even more complicated when dealing with membrane proteins, where the lipid bilayer can exert a role for drug binding. In this case, the model system can consist of ~250,000 atoms or more, making the use of QM prohibitive [15, 19]. To overcome this, QM methods are usually paired with classical MM approaches in the hybrid QM/MM scheme (Fig. 7). This allows the crucial part of the system in which the reaction takes place (the protein active site and the ligand) to be treated at the QM level, while the rest of the system is described with a classical MM force field. Since its first appearance in 1976, the QM/MM approach has become increasingly popular in studying enzymatic reaction mechanisms [20]. It has been of paramount importance in modern computational chemistry for allowing the multiscale modeling of complex chemical systems. The scientific impact of the QM/MM approach was acknowledged with the 2013 Nobel Prize in Chemistry.

For studying biological systems, the QM/MM scheme has been successfully coupled to ab initio MD methods, such as Car–Parrinello and Born–Oppenheimer MD. In this way, model systems of a few hundred atoms (i.e., the protein active site and the ligand) are simulated at their actual temperature, thus including entropic effects, during picosecond timescale trajectories. Of several different schemes for first-principles-based dynamics, the Car–Parrinello method has been widely used to study numerous pharmaceutically relevant targets, including phosphatases, hydrolases, and lactamases [6, 10]. Ab initio MD permits an atomic-level understanding of catalytic and inhibitory mechanisms in biological systems, providing results that can be directly compared with experiments. The practical impact of these high-level computations seems more applicable to the lead optimization phase, while a much broader applicability of QM-based methods is foreseen in SBDD over the next decade [18, 20].

Ligand-Based Drug Design Approaches

Ligand-based drug design (LBDD) approaches are applied when the three-dimensional structure of the target is unknown or when it is not fully reliable (i.e., homology models of the target are based on poor sequence identity) [1]. In this case, the experimental activity data of active compounds are generally used to construct pharmacophore models. These models include an ensemble of steric and electronic features



Computational Chemistry for Drug Discovery, Fig. 8 An example of a pharmacophore model of (R)-roscovitine, which tightly binds CDK5 [23]. Crucial interacting residues of CDK5 are shown in *green* (hydrophobic residues), *blue* (H-bond acceptors), and *magenta* (H-bond donors). The regions of the ligand that match those types of interaction are indicated with the same color

(usually indicated as descriptors) that are likely to be those mainly responsible for activity (Fig. 8). Typical descriptors are physicochemical features such as molecular weight, geometry, surface accessible area, aromaticity index, electronegativity, polarizability, and solvation properties. In general, a proper set of descriptors should cover a broad chemical diversity space. In this way, the more relevant descriptors for building predictive pharmacophores are likely to be recognized. These are then used to help identify new active compounds, for example, in a virtual library of drug-like molecules. The same principles are valid for molecular similarity methods, where similitudes in geometrical or physicochemical properties among active compounds are used to identify new active compounds from among the many that can be virtually screened. A similarity coefficient, such as the Tanimoto or Euclidean coefficients, is then used to identify whether a set of new compounds is likely to interact with a given target [5].

Geometrical and electronic descriptors are used to build a quantitative structure–activity relationship (QSAR), which results in a mathematical model able to predict the biological activity of new compounds [5]. For example, activity data of active molecules can be used to extract mathematical models for early prediction of activity or, more often, metabolic properties that could generate toxicity problems. Conventional 3D QSAR models consider the ensemble of conformations, orientations, tautomers, stereoisomers, and protonation states of the initial ligand set. More exotic multidimensional 4D or 5D QSAR models have also been developed, which take into account all energy contributions of ligand binding, including solvation energy and conformational entropy. Usually, linear regression and principal component analysis can be combined for a partial least square analysis that directly relates the reduced set of descriptors to the biological activity. However, a nonlinear relation is often present between the biological activity and the descriptors used. Therefore, nonlinear regression models using machine-learning algorithms have recently been introduced in QSAR. Of these, artificial neural networks algorithms are used to discover the relationship between descriptors and biological activity via an iterative process [21].

Absorption, distribution, metabolism, excretion, and toxicity (ADMET) are the key parameters that can be optimized with QSAR approaches to get a drug candidate with a proper drug-like pharmacokinetic (PK) profile. ADMET and PK

properties are critical to the success of any drug discovery program. A poor drug-like PK profile can result in the rapid metabolism and elimination of the compound from the body. A drug-like compound should be characterized by certain physicochemical properties, which are loosely recapitulated in Lipinski's rule of five. This says that an optimal orally bioavailable drug should have a molecular weight less than 500, less than 5 H-bond donor sites, and less than 10 H-bond acceptor sites, while the log of the octanol/water partition coefficient (logP), a measure of hydrophobicity, should be below 5. Despite its practical utility for quickly evaluating drug-likeness, it is nowadays clear that this rule can only estimate the compound's probable success by highlighting potentially problematic aspects of its physicochemical and structural profile [22]. In fact, many approved drugs, such as large macrolide antibiotics (MW $\sim$ 1000), differ in one or more descriptors from Lipinski's rule. Other important features can heavily affect the PK profile. For ideal absorption and distribution, an aqueous solubility above 10^{-6} M is advised (logS > -6). Transport properties, such as membrane permeability and brain/blood partitioning (log BB), for allowing blood–brain barrier penetration must also be taken into account in certain drug discovery projects. Additionally, interactions with influx/efflux transporter proteins or metabolic enzymes can greatly affect the final therapeutic effect and should be considered in the early phases of a drug discovery program. That is, for a given molecular structure, properties such as solubility, membrane permeability, partition coefficients, blood–brain barrier penetration, plasma protein binding, and metabolite formation are computed and compared against a database of known drugs. Computational tools and methods for predicting ADMET properties and producing predictive statistical models are thus valuable. These statistical models are trained on experimental ADMET data obtained from several tested compounds. These computational tools are therefore crucial for selecting, prioritizing, and optimizing promising lead compounds.

Conclusions and Perspective

This essay describes some key aspects of the most effective computational chemistry methods for accelerating drug discovery, showing how different challenges can be tackled with these methods. Each computational method is suited to a particular drug discovery phase (see Fig. 1). Indeed, over the last two decades, it has become increasingly evident that computational methods can accelerate the identification, design, and optimization of small molecules as promising drug candidates. SBDD approaches suit better the hit identification and lead generation phase, where new active small molecules are designed, identified, and transformed into potent inhibitors. QSAR methods are usually better for lead optimization, where drug-like properties are tuned to increase efficacy and lower attrition in drug development. In addition to conventional methods (such as virtual screening and QSAR), methods that were once prohibitive for effective drug design, such as molecular dynamics simulations and quantum-mechanics-based methods, are nowadays frequently used in all phases of the drug discovery pipeline, broadening drug discovery's computational arsenal. In fact, the advent of faster algorithms and better computer hardware will allow more extensive use of molecular dynamics and first-principles-based simulations. We therefore envisage that computational chemistry methods will have an even greater impact on drug discovery in the future years.

References

1. Jorgensen, W.L.: The many roles of computation in drug discovery. Science **303**, 1813–1818 (2004)
2. Jorgensen, W.L.: Efficient drug lead discovery and optimization. Acc. Chem. Res. **42**, 724–733 (2009)
3. Frenkel, D., Smit, B.: Understanding Molecular Simulation. Academic, San Diego (2002)
4. Jensen, F.: An Introduction to Computational Chemistry. Wiley, Chichester, England (1998)
5. Leach, A.R.: Molecular Modelling: Principles and Applications, 2nd edn. Prentice Hall, New York (2001)
6. Carloni, P., Rothlisberger, U., Parrinello, M.: The role and perspective of ab initio molecular dynamics in the

study of biological systems. Acc. Chem. Res. **35**, 455–464 (2002)
7. Nichols, S.E., Baron, R., Ivetac, A., McCammon, J.A.: Predictive power of molecular dynamics receptor structures in virtual screening. J. Chem. Inf. Model. **51**, 1439–1446 (2011)
8. Schneider, G., Fechner, U.: Computer-based de novo design of drug-like molecules. Nat. Rev. Drug Discov. **4**, 649–663 (2005)
9. Palermo, G., Branduardi, D., Masetti, M., Lodola, A., Mor, M., Piomelli, D., Cavalli, A., De Vivo, M.: Covalent inhibitors of fatty acid amide hydrolase: a rationale for the activity of piperidine and piperazine aryl ureas. J. Med. Chem. **54**, 6612–6623 (2011)
10. Palermo, G., Rothlisberger, U., Cavalli, A., Vivo, M. D.: Computational insights into function and inhibition of fatty acid amide hydrolase. Eur. J. Med. Chem. **91**, 15–26 (2015)
11. Borhani, D.W., Shaw, D.E.: The future of molecular dynamics simulations in drug discovery. J. Comput. Aided Mol. Des. **26**, 15–26 (2012)
12. De Vivo, M., Dal Peraro, M., Klein, M.L.: Phosphodiester cleavage in ribonuclease H occurs via an associative two-metal-aided catalytic mechanism. J. Am. Chem. Soc. **130**, 10955–10962 (2008)
13. De Vivo, M., Ensing, B., Dal Peraro, M., Gomez, G. A., Christianson, D.W., Klein, M.L.: Proton shuttles and phosphatase activity in soluble epoxide hydrolase. J. Am. Chem. Soc. **129**, 387–394 (2007)
14. De Vivo, M., Ensing, B., Klein, M.L.: Computational study of phosphatase activity in soluble epoxide hydrolase: high efficiency through a water bridge mediated proton shuttle. J. Am. Chem. Soc. **127**, 11226–11227 (2005)
15. Palermo, G., Campomanes, P., Cavalli, A., Rothlisberger, U., De Vivo, M.: Anandamide hydrolysis in FAAH reveals a dual strategy for efficient enzyme-assisted amide bond cleavage via nitrogen inversion. J. Phys. Chem. B.**119**(3), 789–801 (2015)
16. Ensing, B., De Vivo, M., Liu, Z.W., Moore, P., Klein, M.L.: Metadynamics as a tool for exploring free energy landscapes of chemical reactions. Acc. Chem. Res. **39**, 73–81 (2006)
17. Jorgensen, W.L., Thomas, L.L.: Perspective on free-energy perturbation calculations for chemical equilibria. J. Chem. Theory Comput. **4**, 869–876 (2008)
18. De Vivo, M.: Bridging quantum mechanics and structure-based drug design. Front. Biosci. Landmrk. **16**, 1619–1633 (2011)
19. Palermo, G., Stenta, M., Cavalli, A., Dal Peraro, M., De Vivo, M.: Molecular simulations highlight the role of metals in catalysis and inhibition of type II Topoisomerase. J. Chem. Theory Comput. **9**, 857–862 (2013)
20. Lodola, A., De Vivo, M.: The increasing role of QM/MM in drug discovery. Adv. Protein Chem. Struct. Biol. **87**, 337–662 (2012)
21. Nantasenamat, C., Isarankura-Na-Ayudhya, C., Prachayasittikul, V.: Advances in computational methods to predict the biological activity of compounds. Expert Opin. Drug Discov. **5**, 633–654 (2010)
22. Lipinski, C.: Chris Lipinski. Interview by Peter Kirkpatrick. Nat. Rev. Drug Discov. **11**, 900–901 (2012)
23. Mapelli, M., Massimiliano, L., Crovace, C., Seeliger, M.A., Tsai, L.H., Meijer, L., Musacchio, A.: Mechanism of CDK5/p25 binding by CDK inhibitors. J. Med. Chem. **48**, 671–679 (2005)

Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering

A. T. Conlisk
Department of Mechanical Engineering, The Ohio State University, Columbus, OH, USA

Synonyms

Microscale fluid mechanics; Nanoscale fluid mechanics

Definition

Because of the small scale of the fluid conduits, electric fields must often be used to transport fluids especially at the nanoscale. This means that the fluids must be electrically conducting, and so microfluidics and nanofluidics require the user to be knowledgeable in fluid mechanics, heat and mass transfer, electrostatics, electrokinetics, electrochemistry, and, if biomolecules are involved, molecular biology.

Introduction

The term microfluidics refers generally to internal flow in a tube or channel whose smallest dimension is under 100 μm. Nanofluidics refers to the

same phenomenon in a conduit whose smallest dimension is less than 100 nm.

Microchannels and nanochannels have large surface-to-volume ratio, so that surface properties become enormously important. In fully developed channel flow, the pressure drop $\Delta p \sim \frac{1}{h^3}$, where h is the small dimension, and so the pressure drop is prohibitively large for a nanoscale channel. Thus a solvent fluid such as water, proteins and other biomolecules, and other colloidal particles are most often transported electrokinetically. This means that the art of designing micro- and nanodevices requires a significant amount of knowledge of fluid flow and mass transfer (biofluids are usually multicomponent mixtures) and often heat transfer, electrostatics, electrokinetics, electrochemistry, and molecular biology. Details of the character of these fields of study can be found in the book by the author [1].

The study of micro-/nanofluidics requires knowledge of all of the abovementioned fields and so has a unifying effect. Moreover, nanofluidics opens the door to the discovery of the structure and conformation of biomaterials such as proteins and polysaccharides through molecular simulation.

In dealing with devices with small-scale features, microscale and below, there are three activities that normally comprise the design process; these are the following:

- Modeling: computational and theoretical
- Fabrication
- Experimental methods

Modeling is often done prior to the fabrication process as a guide as to what can be done. Experimental methods are usually used to assess the performance of a device, among other purposes.

Surface-to-Volume Ratio

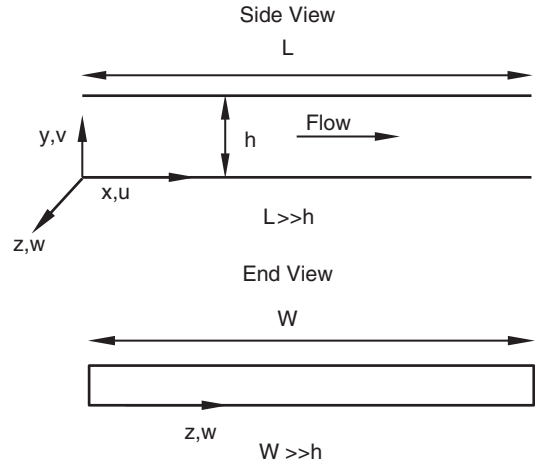
Consider a channel of rectangular cross section having dimensions in the (x, y, z) coordinate system of (L, h, W) with the primary direction of fluid motion being in the x direction. Then the surface-to-volume ratio is given

$$\frac{S}{V} = 2\left(\frac{1}{L} + \frac{1}{h} + \frac{1}{W}\right) = 6m^{-1} \tag{1}$$

for a channel having all three dimensions $L = h = W = 1m$. On the other hand, for a channel having dimensions (10 μm, 10^{-2} μm, 10 μm),

$$\frac{S}{V} \sim 2 \times 10^{8} m^{-1} \tag{2}$$

This means that surface properties become very important at the microscale and nanoscale and surfaces are routinely engineered to achieve a desired objective. In most devices the nanoscale features interface directly with microscale features. A typical channel geometry is depicted in Fig. 1.



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 1 Geometry of a typical channel. In applications $h \ll W, L$, where W is the width of the channel and L is the length in the primary flow direction. u, v, and w are the fluid velocities in the x, y, and z directions

Fluid Mechanics

Micro- and nanofluidics generally involve the flow of electrically conducting fluids, *electrolyte solutions*, that are assumed to be incompressible, having a constant density. Generally, the flows are internal, bounded on each side by walls, and are assumed to be fully developed. In this case, referring to Fig. 1, the governing equation for the velocity u in a channel is given by

$$\mu \frac{\partial^2 u}{\partial y^2} = \frac{\partial p}{\partial x} - B_x \tag{3}$$

where p is the pressure and B_x is a body force. The no-slip condition is applied at each wall: $u = 0$ at $y = 0, h$.

Mass Transfer

The molar flux of species A for a dilute electrically conducting mixture is

$$\vec{N}_A = -D_{AB}\nabla_{CA} + m_A z_A c_A \vec{E} + c_A \vec{V} \tag{4}$$

Here D_{AB} is the diffusion coefficient, R is the universal gas constant, T is the temperature, $z_A m_A$ is called the ionic mobility with $m_A = \frac{FD_{AB}}{RT}$, z_A is the valence, $F = 96500 \frac{Coul}{mole}$ is Faraday's constant, and $\vec{E}$ is the electric field. Equation 4 is called the Nernst-Planck equation, and the electric field term in the flux equation is called *electrical migration*. The boundary condition of interest here is that the solid walls in Fig. 1 are impermeable to species A, or $N_{A_y} = 0$ at $y = 0, h$.

In the absence of a velocity field $\vec{V}$ and in one dimension, Eq. 4 can be integrated to give

$$c_A = c_{A0} e^{-zA\phi} \tag{5}$$

and this is termed the *Boltzmann distribution* for the concentration of species A.

The Electric Field

An electric field is set up around any charged body and is defined as the force per charge on a surface. Electrical charges are either positive or negative, and like charges repel and opposite charges attract. For two bodies of charge q and q', the *electric field* is defined by

$$E = \frac{F}{q'} = \frac{q}{4\pi\epsilon_e r^2} \quad \frac{N}{C} \tag{6}$$

and is directed outward from the body of charge q and toward the body having a charge q' if $q > 0$, and the electric field is in the opposite direction if $q < 0$. In general, the electric field is a vector. This formula is called *Coulomb's Law*, and ϵ_e is called the *electrical permittivity.* The electrical permittivity is a transport property like the viscosity and thermal conductivity of a fluid.

The electric field due to a flat wall having a surface charge density σ in $\frac{Coulomb}{m^2}$ on one side is directed normal to the surface and has magnitude

$$E = \frac{\sigma}{2\epsilon_e} \tag{7}$$

A wire is characterized as having a line charge density, and if charges are distributed over a volume, a volume charge density is defined and called ρ_e in $\frac{Coulomb}{m^3}$.

The *electrical potential* is defined as the work done in moving a unit of charge, and mathematically

$$\phi = -\int_a^b \vec{E} \bullet d\vec{s} \tag{8}$$

The units of the electric potential are $\frac{Nm}{C} = 1\,Volt = 1\,V$. This formula is similar to the formula for mechanical work given by

$$W = -\int_a^b \vec{F} \bullet \vec{ds} \tag{9}$$

In differential form the electrical potential is given by

$$\vec{E} = -\nabla\phi \tag{10}$$

For a single charge *Gauss's Law* is given by

$$\int\int_S \epsilon_e \vec{E} \bullet d\vec{A} = q \tag{11}$$

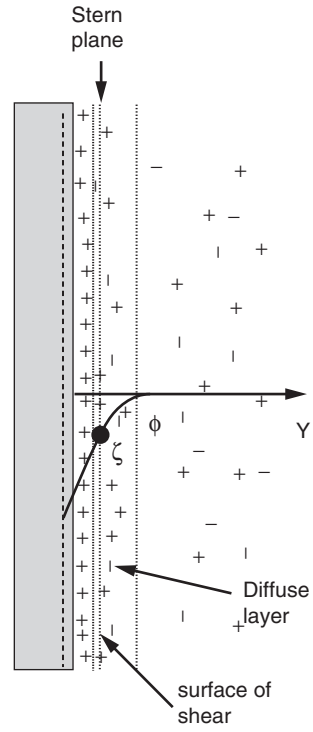
For a volume that contains a continuous distribution of charge, ρ_e, summing over all the charges using the definition of the integral, Gauss's Law becomes

$$\int\int_S \epsilon_e \vec{E} \bullet d\vec{A} = \int\int\int_{\mathcal{V}} \rho_e d\mathcal{V} \tag{12}$$

Using Eq. 12 and the differential form of the definition of the electrical potential, it follows that

$$\nabla^2\phi = -\frac{\rho_e}{\epsilon_e} \tag{13}$$

This is a *Poisson equation* for the potential given the volume charge density. The combination of Eqs. 4 and 13 is called the *Poisson-Nernst-Planck* system of equations.



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 2 The electric double layer (EDL) consists of a layer of counterions pinned to the wall, the Stern layer, and a diffuse layer of mobile ions outside that layer. The wall is shown as being negatively charged and the ζ potential is defined as the electrical potential at the Stern plane

Electrochemistry

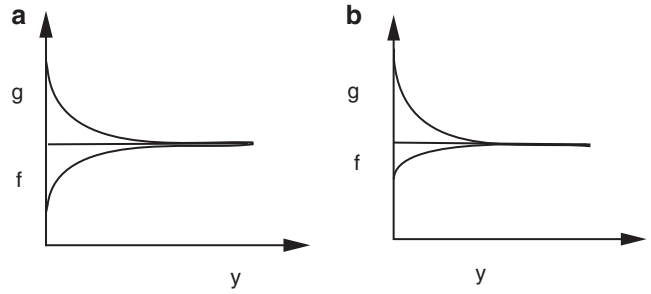
Electrochemistry may be broadly defined as the study of the electrical properties of chemical and biological material [2]. In particular much of electrochemistry pertinent to micro- and nanofluidics involves the study of the behavior of *ionic solutions* and the *electrical double layer* (*EDL*). Electrochemistry of electrodes is important to understand the operation of a *battery.*

An ionic or electrolyte solution is a mixture of ions, or charged species immersed in a solvent, often water. It is the charged nature of ionic solutions that allows the fluid to move under the action of an electric field, provided by electrodes placed upstream and downstream of a channel in a *nanopore membrane*. The term *membrane* is used to mean a thin sheet of porous material that allows fluid to flow in channels that make up the porous part of the membrane. Those channels are often like those channels depicted in Fig. 1.

Because the surface-to-volume ratio is so large in a nanoscale channel, the properties of the surface are extremely important. Fluid can be moved by an electric field if the surfaces of a channel are charged. If the surface is negatively charged, a surplus of positive ions will arrange themselves near the wall. This is shown in Fig. 2. It is because of this excess charge that allows fluid to be transported by an externally applied electric field.

The nominal length scale associated with the EDL is the *Debye length* defined by

$$\lambda = \frac{\sqrt{\epsilon_e RT}}{FI^{1/2}} \tag{14}$$



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 3 (**a**) Debye-Hückel [5] picture of the electric double layer. Here *g* denotes the cation mole fraction and *f* denotes the anion mole fraction. The Debye-Hückel model assumes the cation and anion wall mole fractions are symmetric about a mean value which occurs for low surface charge densities. (**b**) Gouy-Chapman model [3, 4] of the EDL allows many more counterions than coions to collect near the charged surface and is valid at higher surface charge densities. From [1]

where F is Faraday's constant, ϵ_e is the electrical permittivity of the medium, I is the *ionic strength*, $I = \sum_i z_i^2 c_i$, c_i is the concentrations of the electrolyte constituents at some reference location, R is the universal gas constant, z_i is the valence of species i, and T is the temperature.

The ion distribution within the EDL can be described by using the number density, concentration, or mole fraction. Engineers usually prefer the dimensionless mole fraction, whereas chemists usually use concentration or number density.

There are two views of the ion distribution within the electrical double layer that are generally thought to be valid and have been verified by numerical solutions of the governing equations (see the section on electrokinetic phenomena below). The Gouy-Chapman [3, 4] model of the electric double layer allows counterions to collect near the surface in much greater numbers than coions. This model as numerical solutions suggest [1] occurs at higher surface charge densities. The Debye-Hückel picture assumes that coions and counterions collect near the surface in roughly equal amounts, above and below a mean value. These pictures are depicted in Fig. 3.

Molecular Biology

The nanoscale is the scale of biology since many proteins and other biomolecules have nanoscale dimensions. Many of the applications of nanofluidics such as rapid molecular analysis, drug delivery, and biochemical sensing have a biological entity as an integral part of their operation. Moreover, using nanofluidic tools, DNA sequencing is now possible. The book by Alberts [6] is a useful tool for learning molecular biology.

Nucleic acids are polymers consisting of nucleotides. Those based on a sugar called *ribose* are called ribonucleic acids (RNA), and those based on *deoxyribose* are called deoxyribonucleic acids (DNA). RNA is single stranded, while DNA is usually double stranded although single-stranded DNA (ss-DNA) does exist. Nucleotides contain five-carbon sugars attached to one or more phosphate groups (a phosphorus central atom surrounded by four oxygens) and a base which can be either adenine (A), cytosine (C), guanine (G), or thymine (T). Two nucleotides connected by a hydrogen bond are called a base pair (bp). *Protein synthesis* begins at a gene on a particular strand of a DNA molecule in a cell [6].

There are seven basic types of proteins classified according to their function, although different authors use different terms to describe each class; see, for example, Alberts [6], panel 5-1. *Enzymes* are catalysts in biological reactions within the cell. For example, the immune system responds to foreign bacteria and viruses by producing *antibodies* that destroy or bind to the antigen, the foreign agent. The antigen is the catalyst, or

Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering,

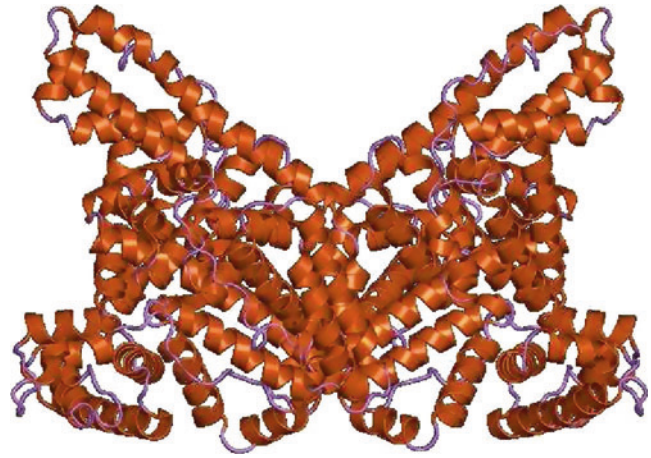


Fig. 4 Ribbon view of the protein albumin depicting its folding pattern, the *secondary structure* of a protein. From the European Bioinformatics Institute, public domain www.ebi.ac.uk

reaction enhancer, for inducing the immune response: the production of the antibodies. Proteins are responsible for many of the essential functions of the body, including moving material into and out of cells, regulating metabolism, managing temperature and *pH*, and muscle operation, among other functions.

Proteins are large and complex molecules, polymers made up of a total of 20 amino acids and held together by peptide bonds. The 20 amino acids have *side chains* that can be *basic*, *acidic*, *polar*, or *nonpolar*. Because they are so large, they cannot be described easily in a single chemical formula or picture. Thus molecular biologists depict proteins and other macromolecules in distinct levels of structure. The *primary structure* is the amino acid sequence, the order in which the 20 amino acids appear. The *secondary structure* depicts the folding properties of a protein as depicted in Fig. 4. Proteins are further described by more complex folding of the secondary structure (*tertiary structure*) and a *quaternary structure* if the protein has more than one backbone.

Proteins are usually negatively charged, and thus nanopore membranes for rapid molecular analysis can be used to separate different types of proteins and other biomolecules based on different values of size and charge. Biomolecules are what is termed soft material in that they are porous and deform under stress. Indeed, recent measurements of the conformation of albumin show that it may take the shape of a wedge, looking like a piece of pie. With the explosive growth of computer capability, conformations of biomolecules are actually being computed using molecular simulation tools like molecular dynamics and Monte Carlo schemes.

Ion channels are natural conical nanopores whose walls are made of proteins that play a crucial role in the transport of biofluids to and from cells. The basic units of all living organisms are cells. In order to keep the cells functioning properly, there needs to be a continuous flux of ions in and out of the cell and the cell components. The cell and many of its components are surrounded by a plasma membrane which provides selective transfer of ions. The membrane is made up of a double layer of lipid molecules (lipid bilayer) in which proteins are embedded. Ion channels are of two categories: carrier and pore. The *carrier* protein channel is based on the binding of the transport ion to a larger macromolecule, which brings it through the channel. A *pore* ion channel is a narrow, water-filled tunnel, permeable to the few ions and molecules small enough to fit through the tunnel (approximately 10 A in diameter).

Electrokinetic Phenomena

As the scale of the channels in a nanopore membrane becomes smaller, pressure, the normal

means for driving fluids through pipes and channels at macroscale (Fig. 1), becomes very difficult [1] since the pressure drop required scales as h^{-3} where h is the (nanoscale) channel height. Since in many applications the fluids used are electrically conducting, electric fields can be used to effectively pump fluid. Moreover electrically charged particles can move relative to the bulk fluid motion, and thus species of particles can be separated.

These *electrokinetic phenomena* are generally grouped into four classes [1]:

1. Electroosmosis (electroosmotic flow): the bulk motion of a fluid caused by an electric field
2. Electrophoresis: the motion of a charged particle in an otherwise motionless fluid or the motion of a charged particle relative to a bulk motion
3. Streaming potential or streaming current: the potential induced by a pressure gradient at zero current flow of an electrolyte mixture
4. Sedimentation potential: the electric field induced when charged particles move relative to a liquid under a gravitational or centrifugal or other force field

By far the two most important of these phenomena are electroosmosis and electrophoresis, and for the purposes of the theme of this entry, electroosmosis is discussed exclusively.

The dimensionless form of the streamwise momentum equation in the fully developed flow region in the absence of a pressure gradient is

$$\epsilon^2 \frac{\partial^2 u}{\partial y^2} = -\beta \sum_i z_i X_i \tag{15}$$

and the Poisson equation for the potential in dimensionless form is

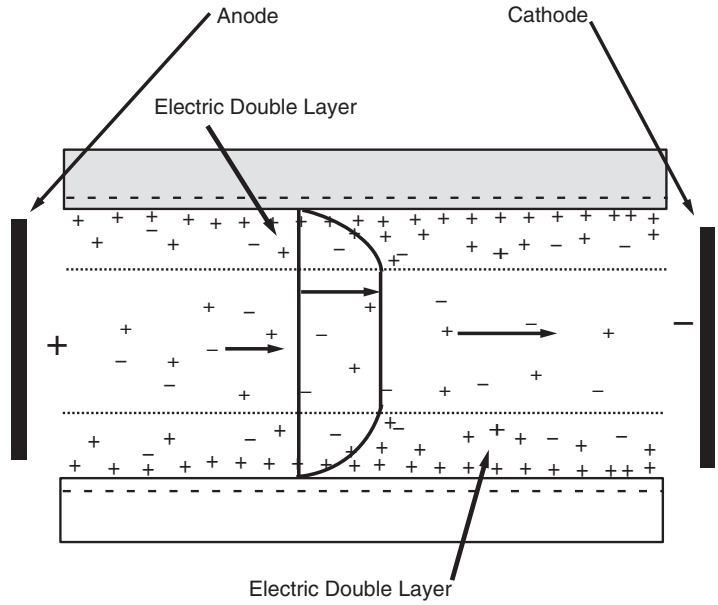
$$\epsilon^2 \frac{\partial^2 \phi}{\partial y^2} = -\beta \sum_i z_i X_i \tag{16}$$

where the partial derivatives in this one-dimensional fully developed analysis are really total derivatives, $\epsilon = \frac{\lambda}{h}$ and $\beta = \frac{c}{I}$ where c is the total concentration including the solvent and I is the ionic strength. Here X_i is the mole fraction, but if the electrolyte concentrations are scaled on the ionic strength, $I = \sum_i z_i c_i$, $\beta = 1$. It is seen from Eq. 15 that the combination of the electrodes that create an electric field and the excess charge in the electrical double layers produces the electrical force that balances the viscous force causing the electrolyte to move, and this is depicted in Fig. 5.

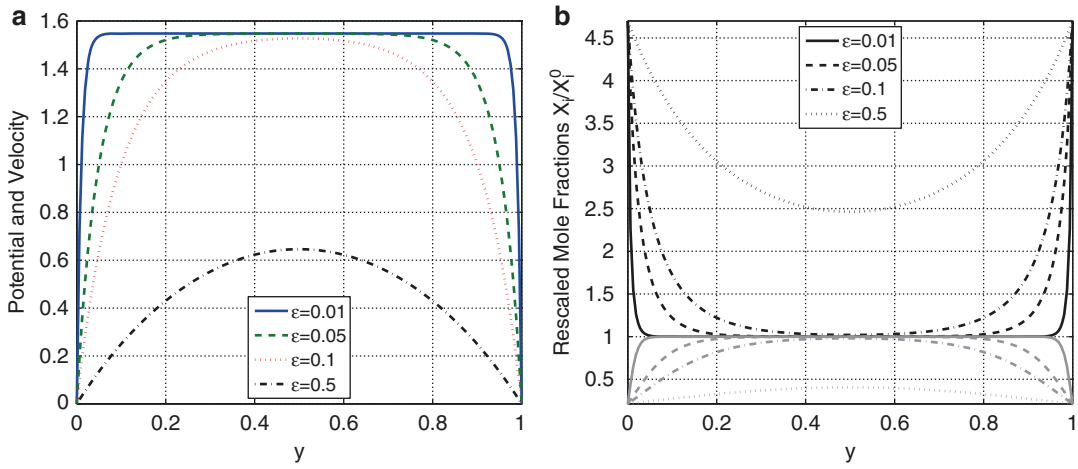
The fluid velocity satisfies the no-slip condition at the wall, and the electric potential satisfies $\phi(0) = \phi(1) = 0$. Then both the equations and the boundary conditions are identical and on a dimensionless basis $u(y) = \phi(y)$. In reality, the potential does not vanish at the wall, but if a Dirichlet boundary condition holds and the potential satisfies $\phi = \zeta$ at $y = 0{,}1$ and thus $u = \phi - \zeta$, where ζ is the *dimensionless* ζ-potential at the wall. Results for the potential and velocity and the concentrations scaled on the ionic strength are presented in Fig. 6 [1]. Note that for $\epsilon << 1$, the fluid velocity is constant away from the walls of the channel, unlike the Poiseuille flow of pressure-driven flow.

Dimensional Analysis

The equivalence of the dimensionless velocity and the dimensionless potential is important because the measurements of the velocity are equivalent to the measurements of the electric potential. Moreover, it is noted that a Debye length of $\lambda = 1$ nm in a $h = 100$ nm channel gives the same value of $\epsilon = 0.01$ as a $\lambda = 100$ nm Debye length in a $h = 10$ μm channel so that microscale measurements can be validated by nanoscale computations and, conversely, nanoscale computations can be validated by miscoscale experiments. Note that this similarity analysis does not apply for unsteady flow since there is no time derivative in the potential equation.



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 5 The combination of electrodes in the regions upstream and downstream of a charged channel or membrane, usually fluid reservoirs, causes electroosmotic flow



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 6 (**a**) Potential, velocity, and rescaled mole fractions for a 1:1 electrolyte for various values of ϵ. Here the dimensional potential on both walls is $\zeta^* = -40mV$. In (**b**) the mole fractions are rescaled based on the upstream reservoir mole fractions as $\frac{X_i}{X_i^0}$. The cations are plotted in black lines and the anions are plotted in gray lines. From [1]

Closure

Physics of fluids at the nanoscale is dominated by the large surface-to-volume ratio inherent at this length scale, and thus the surfaces of a channel or membrane become extremely important. Indeed, the pressure drop across a rectangular channel in a nanopore membrane, $\Delta p \sim \frac{1}{h^3}$, is prohibitively large for the efficient operation of a nanofluidic device if $h = O(nm)$. Thus fluid, charged biomaterials such as proteins, and colloidal particles are most often

transported electrokinetically at the nanoscale. The art of designing micro- and nanofluidic devices therefore requires a significant amount of knowledge of fluid flow, mass transfer and sometimes heat transfer, electrostatics, electrokinetics, electrochemistry, and molecular biology.

The common thread is micro-/nanofluidics, which plays the role of unifying and integrating these fields. In particular, nanofluidics opens the door to reveal the structure and behavior of flows around nanoparticles and the conformation of proteins and other biomolecules using molecular simulations.

Cross-References

- ▶ Applications of Nanofluidics
- ▶ Electrokinetic Fluid Flow in Nanostructures
- ▶ Nanochannels for Nanofluidics: Fabrication Aspects
- ▶ Rapid Electrokinetic Patterning and Its Applications

References

1. Conlisk, A.T.: Essentials of Micro- and Nanofluidics with Application to the Biological and Chemical Sciences. Cambridge University Press, Cambridge, UK (2013)
2. Bockris, J.O.M., Reddy, A.K.N.: Modern Electrochemistry. Ionics, vol. 1, 2nd edn. Plenum Press, New York (1998)
3. Gouy, G.: About the electric charge on the surface of an electrolyte. J. Phys. A **9**, 457–468 (1910)
4. Chapman, D.L.: A contribution to the theory of electrocapillarity. Phil. Mag. **25**, 475–481 (1913)
5. Debye, P., Huckel, E.: The interionic attraction theory of deviations from ideal behavior in solution. Z. Phys. **24**, 185 (1923)
6. Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Essential Cell Biology. Garland Publishing, New York (1998)

Computational Studies

- ▶ Surface Modeling of Ceramic Biomaterials

Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Leonid V. Zhigilei[1], Alexey N. Volkov[2] and Avinash M. Dongare[3]
[1]Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA, USA
[2]Department of Mechanical Engineering, University of Alabama, Tuscaloosa, AL, USA
[3]Department of Materials Science and Engineering, and Institute of Materials Science, University of Connecticut, Storrs, CT, USA

Synonyms

Carbon nanotube materials; Computer modeling and simulation of materials; Dislocation dynamics; Mesoscopic modeling; Molecular dynamics method; Multiscale modeling; Nanocrystalline materials; Nanofibrous materials and composites

Definitions

Nanomaterials (or nanostructured materials, nanocomposites) are materials with characteristic size of structural elements on the order or less than several hundreds of nanometers at least in one dimension. Examples of nanomaterials include nanocrystalline materials, nanofiber, nanotube, and nanoparticle-reinforced nanocomposites, as well as multilayered systems with submicron thickness of the layers.

Atomistic modeling is based on atoms as elementary units in the models, thus providing the atomic-level resolution in the computational studies of material structure and properties. The main classical atomistic methods in materials research are (1) molecular dynamics technique that yields "atomic movies" of the dynamic material behavior through the integration of the equations of motion of atoms and molecules, (2) Metropolis Monte Carlo method that enables evaluation of

the equilibrium properties through the ensemble averaging over a sequence of random atomic configurations generated according to the desired statistical-mechanics distribution, and (3) kinetic Monte Carlo method that provides a computationally efficient way to study systems where the structural evolution is defined by a finite number of thermally activated elementary processes.

Mesoscopic modeling is a relatively new area of computational materials science that considers material behavior at time and length scales intermediate between the atomistic and continuum levels. Mesoscopic models are system/phenomenon specific and adopt coarse-grained representations of the material structure, with elementary units in the models designed to provide a computationally efficient representation of individual crystal defects or other elements of micro/nanostructure. Examples of the mesoscopic models are coarse-grained models for molecular systems, discrete dislocation dynamics model for crystal plasticity, mesoscopic models for nanofibrous materials, cellular automata, and kinetic Monte Carlo Potts models for simulation of microstructural evolution in polycrystalline materials.

Computer Modeling of Nanomaterials

Rapid advances in synthesis of nanostructured materials combined with reports of their enhanced or unique properties have created, over the last decades, a new active area of materials research. Due to the nanoscopic size of the structural elements in nanomaterials, the interfacial regions, which represent an insignificant volume fraction in traditional materials with coarse microstructures, start to play the dominant role in defining the physical and mechanical properties of nanostructured materials. This implies that the behavior of nanomaterials cannot be understood and predicted by simply applying scaling arguments from the structure–property relationships developed for conventional polycrystalline, multiphase, and composite materials. New models and constitutive relations, therefore, are needed for an adequate description of the behavior and properties of nanomaterials.

Computer modeling is playing a prominent role in the development of the theoretical understanding of the connections between the atomic-level structure and the effective (macroscopic) properties of nanomaterials. Atomistic modeling has been at the forefront of computational investigation of nanomaterials and has revealed a wealth of information on structure and properties of individual structural elements (various nanolayers, nanoparticles, nanofibers, nanowires, and nanotubes) as well as the characteristics of the interfacial regions and modification of the material properties at the nanoscale. Due to the limitations on the time and length scales, inherent to atomistic models, it is often difficult to perform simulations for systems that include a number of structural elements that is sufficiently large to provide a reliable description of the macroscopic properties of the nanostructured materials. An emerging key component of the computer modeling of nanomaterials is, therefore, the development of novel mesoscopic simulation techniques capable of describing the collective behavior of large groups of the elements of the nanostructures and providing the missing link between the atomistic and continuum (macroscopic) descriptions. The capabilities and limitations of the atomistic and mesoscopic computational models used in investigations of the behavior and properties of nanomaterials are briefly discussed and illustrated by examples of recent applications below.

Atomistic Modeling

In atomistic models [1, 2], the individual atoms are considered as elementary units, thus providing the atomic-level resolution in the description of the material behavior and properties. In classical atomistic models, the electrons are not present explicitly but are introduced through the interatomic potential, $U\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$, that describes the dependence of the potential energy of a system of N atoms on the positions $\vec{r}_i$ of the atoms. It is assumed that the electrons adjust to changes in atomic positions much faster than the atomic nuclei move (Born-Oppenheimer approximation) and the potential energy of a system of

interacting atoms is uniquely defined by the atomic positions.

The interatomic potentials are commonly described by analytic functions designed and parameterized by fitting to available experimental data (e.g., equilibrium geometry of stable phases, density, cohesive energy, elastic moduli, vibrational frequencies, characteristics of the phase transitions, etc.). The interatomic potentials can also be evaluated through direct quantum mechanics–based electronic structure calculations in so-called first-principle (*ab initio*) simulation techniques. The *ab initio* simulations, however, are computationally expensive and are largely limited to relatively small systems consisting of tens to thousands of atoms. The availability of reliable and easy-to-compute interatomic potential functions is one of the main conditions for the expansion of the area of applicability of atomistic techniques to realistic quantitative analysis of the behavior and properties of nanostructured materials.

The three atomistic computational techniques commonly used in materials research are

1. Metropolis Monte Carlo method – the equilibrium properties of a system are obtained via ensemble averaging over a sequence of random atomic configurations, sampled with probability distribution characteristic for a given statistical mechanics ensemble. This is accomplished by setting up a random walk through the configurational space with specially designed choice of probabilities of going from one state to another. In the area of nanomaterials, the application of the method is largely limited to investigations of the equilibrium shapes of individual elements of nanostructure (e.g., nanoparticles) and surface structure/composition (e.g., surface reconstruction and compositional segregation [3]).
2. Kinetic Monte Carlo method – the evolution of a nanostructure can be obtained by performing atomic rearrangements governed by predefined transition rates between the states, with time increments formulated so that they relate to the microscopic kinetics of the system. Kinetic Monte Carlo is effective when the structural and/or compositional changes in a nanostructure are defined by a relatively small number of thermally activated elementary processes, e.g., when surface diffusion is responsible for the evolution of shapes of small crystallites [4] or growth of two-dimensional fractal-dendritic islands [5].
3. Molecular dynamics method – provides the complete information on the time evolution of a system of interacting atoms through the numerical integration of the equations of motion for all atoms in the system. This method is widely used in computational investigations of nanomaterials and is discussed in more detail below.

Molecular Dynamics Technique

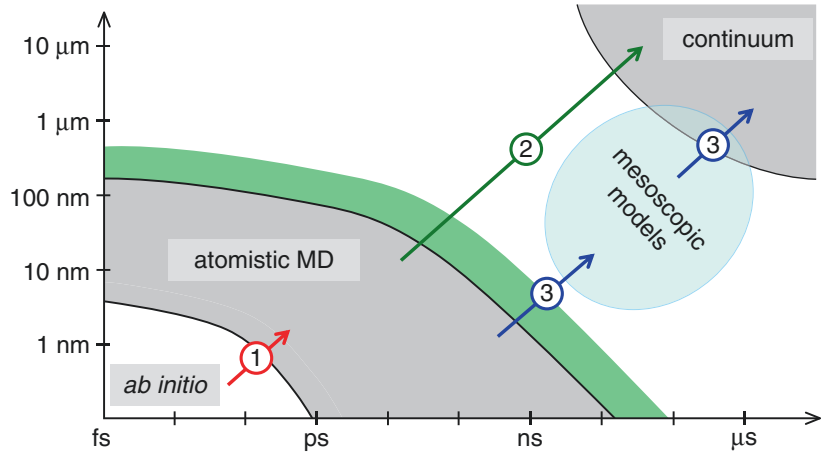
Molecular dynamics (MD) is a computer simulation technique that allows one to follow the evolution of a system of N particles (atoms in the case of atomistic modeling) in time by solving classical equations of motion for all particles in the system,

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i, \quad i = 1, 2, \ldots, N \qquad (1)$$

where m_i and $\vec{r}_i$ are the mass and position of a particle i and $\vec{F}_i$ is the force acting on this particle due to the interaction with other particles in the system. The force acting on the ith particle at a given time is defined by the gradient of the interparticle interaction potential $U\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$ that, in general, is a function of the positions of all the particles:

$$\vec{F}_i = -\vec{\nabla}_i U\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right) \qquad (2)$$

Once the initial conditions (initial positions and velocities of all particles in the system) and the interaction potential are defined, the equations of motion, Eq. 1, can be solved numerically. The result of the solution is the trajectories (positions and velocities) of all the particles as a function of time, $\vec{r}_i(t)$, $\vec{v}_i(t)$, which is the only direct output of a MD simulation. From the trajectories of all



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 1 Schematic representation of the time- and length-scale domains of fist-principle (*ab initio*) electronic structure calculations, classical atomistic MD, and continuum modeling of materials. The domain of continuum modeling can be different for different materials and corresponds to the time and length scales at which the effect of the micro/nanostructure can be averaged over to yield the effective material properties. The *arrows* show the connections between the computational methods used in multiscale modeling of materials: The *red arrow* #1 corresponds to the use of quantum mechanics–based electronic structure calculations to design interatomic potentials for classical MD simulations or to verify/correct the predictions of the classical atomistic simulations; the *green arrow* #2 corresponds to the direct use of the predictions of large-scale atomistic simulations of nanostructured materials for the design of continuum-level constitutive relations describing the material behavior and properties; and the two *blue arrows* #3 show a two-step path from atomistic to continuum material description through an intermediate mesoscopic modeling

particles in the system, however, one can easily calculate (sometimes by sampling the ensemble of initial conditions) the spatial and time evolution of structural and thermodynamic parameters of the system. For example, a detailed atomic-level analysis of the development of the defect structures or phase transformations can be performed and related to changes in temperature and pressure in the system (see examples below).

The main strength of the MD method is that only details of the interatomic interactions need to be specified and no assumptions are made about the character of the processes under study. This is an important advantage that makes MD to be capable of discovering new physical phenomena or processes in the course of "computer experiments." Moreover, unlike in real experiments, the analysis of fast nonequilibrium processes in MD simulations can be performed with unlimited atomic-level resolution, providing complete information on the phenomena of interest.

The predictive power of the MD method, however, comes at a price of a high computational cost of the simulations, leading to severe limitations on time and length scales accessible for MD simulations, as shown schematically in Fig. 1. Although the record length-scale MD simulations have been demonstrated for systems containing more than 10^{12} atoms (corresponds to cubic samples on the order of 10 µm in size) with the use of hundreds of thousands of processors on one of the world's largest supercomputers [6], most of the systems studied in large-scale MD simulations do not exceed hundreds of nanometers even in simulations performed with computationally efficient parallel algorithms (shown by a green area extending the scales accessible for MD simulations in Fig. 1). Similarly, although the record long time scales of up to hundreds of microseconds have been reported for simulations of protein folding performed through distributed computing [7], the duration of most of the simulations in the area of materials research does not exceed tens of nanoseconds.

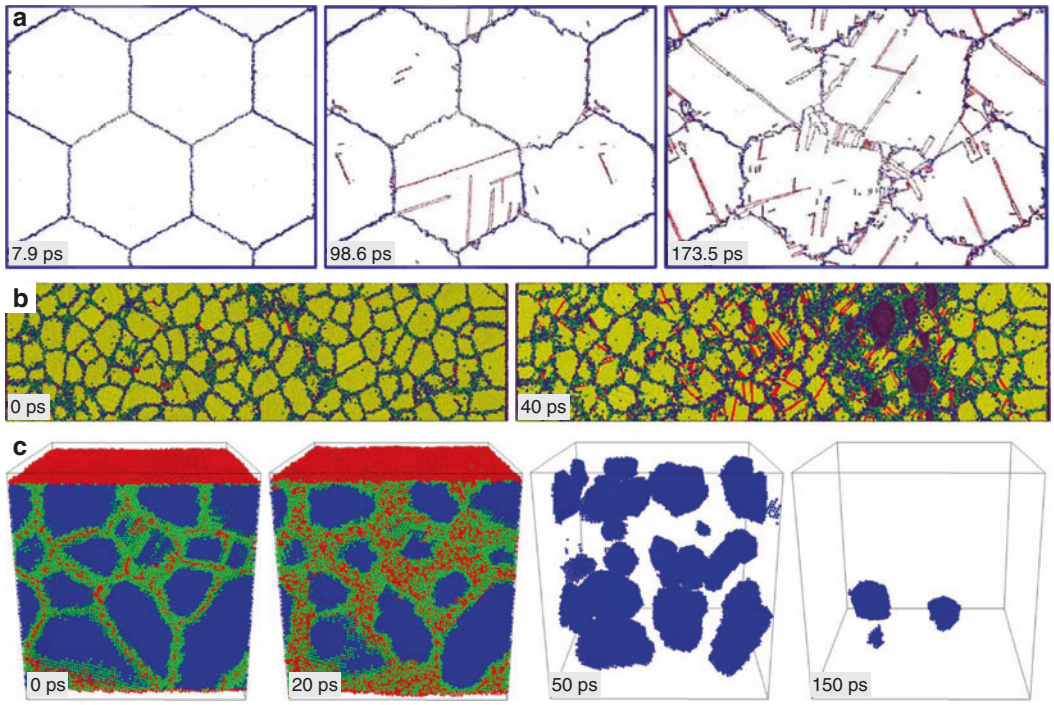
Molecular Dynamics Simulations of Nanomaterials

Both the advantages and limitations of the MD method, briefly discussed above, have important implications for simulations of nanomaterials. The transition to the nanoscale size of the structural features can drastically change the material response to the external thermal, mechanical, or electromagnetic stimuli, making it necessary to develop new structure-property relationships based on new mechanisms operating at the nanoscale. The MD method is in a unique position to provide a complete microscopic description of the atomic dynamics under various conditions without making any *a priori* assumptions on the mechanisms and processes defining the material behavior and properties.

On the other hand, the limitations on the time and length scales accessible to MD simulations make it difficult to directly predict the macroscopic material properties that are essentially the result of a homogenization of the processes occurring at the scale of the elements of the nanostructure. Most of the MD simulations have been aimed at investigation of the behavior of individual structural elements (nanofibers, nanoparticles, interfacial regions in multiphase systems, grain boundaries, etc.). The results of these simulations, while important for the mechanistic understanding of the elementary processes at the nanoscale, are often insufficient for making a direct connection to the macroscopic behavior and properties of nanomaterials.

With the fast growth of the available computing resources, however, there have been an increasing number of reports on MD simulations of systems that include multiple elements of nanostructures. A notable class of nanomaterials actively investigated in MD simulations is nanocrystalline materials – a new generation of advanced polycrystalline materials with sub-micron size of the grains. With a number of atoms on the order of several hundred thousands and more, it is possible to simulate a system consisting of tens of nanograins and to investigate the effective properties of the material (i.e., to make a direct link between the atomistic and continuum descriptions, as shown schematically by the green arrow #2 in Fig. 1). MD simulations of nanocrystalline materials addressing the mechanical [8, 9] and thermal transport [10] properties as well as the kinetics and mechanisms of phase transformations [11, 12] have been reported, with several examples illustrated in Fig. 2. In the first example, Fig. 2a, atomic-level analysis of the dislocation activity and grain-boundary processes occurring during mechanical deformation of an aluminum nanocrystalline system consisting of columnar grains is performed and the important role of mechanical twinning in the deformation behavior of the nanocrystalline material revealed [8]. In the second example, Fig. 2b, the processes of void nucleation, growth, and coalescence in the ductile failure of nanocrystalline copper subjected to an impact loading are investigated, providing important pieces of information necessary for the development of a predictive analytical model of the dynamic failure of nanocrystalline materials [9]. The third example, Fig. 2c, illustrates the effect of nanocrystalline structure on the mechanisms and kinetics of short-pulse laser melting of thin gold films. It is shown that the initiation of melting at grain boundaries can steer the melting process along the path where the melting continues below the equilibrium melting temperature and the crystalline regions shrink and disappear under conditions of substantial undercooling [12].

The brute-force approach to the atomistic modeling of nanocrystalline materials (increase in the number of atoms in the system) has its limits in addressing the complex collective processes that involve many grains and may occur at a micrometer length scale and above. Further progress in this area may come through the development of concurrent multiscale approaches based on the use of different resolutions in the description of the intragranular and grain boundary regions in a well-integrated computational model. An example of a multiscale approach is provided in Ref. [13], where scale-dependent constitutive equations are designed for a generalized finite element method (FEM) so that the atomistic MD equations of motion are reproduced in the regions where the FEM mesh is refined down to atomic level. This and other multiscale approaches can help to focus computational efforts on the important regions of the system



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 2 Snapshots from atomistic MD simulations of nanocrystalline materials: (**a**) mechanical deformation of nanocrystalline Al (only atoms in the twin boundaries left behind by partial dislocations and atoms in disordered regions are shown by *red* and *blue* colors, respectively) [8]; (**b**) spallation of nanocrystalline Cu due to the reflection of a shock wave from a surface of the sample (atoms that have local fcc, hcp, and disordered structure are shown by *yellow*, *red*, and *green/blue* colors, respectively) [9]; and (**c**) laser melting of a nanocrystalline Au film irradiated with a 200 fs laser pulse at a fluence close to the melting threshold (atoms that have local fcc surroundings are colored *blue*, atoms in the liquid regions are *red* and *green*, in the snapshots for 50 and 150 ps the liquid regions are blanked to expose the remaining crystalline regions) [12]

where the critical atomic-scale processes take place. The practical applications of the multiscale methodology so far, however, have been largely limited to investigations of individual elements of material microstructure (crack tips, interfaces, and dislocation reactions), with the regions represented with coarse-grained resolution serving the purpose of adaptive boundary conditions. The perspective of the concurrent multiscale modeling of nanocrystalline materials remains unclear due to the close coupling between the intragranular and grain boundary processes. To enable the multiscale modeling of dynamic processes in nanocrystalline materials, the design of advanced computational descriptions of the coarse-grained parts of the model is needed so that the plastic deformation and thermal dissipation could be adequately described without switching to fully atomistic modeling.

Mesoscopic Modeling

A principal challenge in computer modeling of nanomaterials is presented by the gap between the atomistic description of individual structural elements and the macroscopic properties defined by the collective behavior of large groups of the structural elements. Apart from a small number of exceptions (e.g., simulations of nanocrystalline materials briefly discussed above), the direct analysis of the effective properties of nanostructured

materials is still out of reach for atomistic simulations. Moreover, it is often difficult to translate the large amounts of data typically generated in atomistic simulations into key physical parameters that define the macroscopic material behavior. This difficulty can be approached through the development of mesoscopic computational models capable of representing the material behavior at time and length scales intermediate between the atomistic and continuum levels (prefix *meso* comes from the Greek word μέσος, which means middle or intermediate).

The mesoscopic models provide a "stepping stone" for bridging the gap between the atomistic and continuum descriptions of the material structure, as schematically shown by the blue arrows #3 in Fig. 1. Mesoscopic models are typically designed and parameterized based on the results of atomistic simulations or experimental measurements that provide information on the internal properties and interactions between the characteristic structural elements in the material of interest. The mesoscopic simulations can be performed for systems that include multiple elements of micro/nanostructure, thus enabling a reliable homogenization of the structural features to yield the effective macroscopic material properties. The general strategy in the development of a coarse-grained mesoscopic description of the material dynamics and properties includes the following steps.

1. Identifying the collective degrees of freedom *relevant for the phenomenon under study* (the focus on different properties of the same material may affect the choice of the structural elements of the model);
2. Designing, based on the results of atomic-level simulations and/or experimental data, a set of rules (or a mesoscopic force field) that governs the dynamics of the collective degrees of freedom;
3. Adding a set of rules describing the changes in the properties of the dynamic elements in response to the local mechanical stresses and thermodynamic conditions.

While the atomistic and continuum simulation techniques are well established and extensively used, the mesoscopic modeling is still in the early development stage. The central problem of the mesoscopic modeling is the design of the mesoscopic force field or, in general, the "equations of motion" that describe the dynamics of the collective degrees of freedom in the mesoscopic system. In principle, the equations for the collective degrees of freedom can be derived from the equations of motion of atoms and formulated as the so-called generalized Langevin equation by applying the Mori-Zwanzig projection operators [14, 15]. While formally exact, this equation is difficult to apply to real systems, since a number of parameters of this equation should be evaluated by considering the so-called projection dynamics that is different from the real microscopic dynamics of the corresponding atomistic system [16]. Moreover, the generalized Langevin equation is a stochastic integrodifferential equation with memory effects, which is difficult to apply to large-scale mesoscopic simulations. Therefore, apart from polymer systems, where the Mori-Zwanzig formalism is shown to be a practical tool for constructing mesoscopic computational models [16–18], the mesoscopic force fields for various nanomaterials are usually designed based on semiempirical approaches. There is no universal mesoscopic technique or methodology in the development of such semiempirical force fields, and the current state of the art in mesoscopic simulations is characterized by the development of system-/phenomenon-specific mesoscopic models.

The mesoscopic models used in materials modeling can be roughly divided into two general categories: (1) the models based on lumping together groups of atoms into larger dynamic units or particles and (2) the models that represent the material microstructure and its evolution due to thermodynamic driving forces or mechanical loading at the level of individual crystal defects. The basic ideas underlying these two general classes of mesoscopic models are briefly discussed below.

The models where groups of atoms are combined into coarse-grained computational particles are practical for materials with well-defined structural hierarchy (that allows for a natural choice of

the coarse-grained particles) and a relatively weak coupling between the *internal* atomic motions inside the coarse-grained particles and the *collective* motions of the particles. In contrast to atomic-level models, the atomic structure of the structural elements represented by the coarse-grained particles is not explicitly represented in this type of mesoscopic models. On the other hand, in contrast to continuum models, the coarse-grained particles allow one to explicitly reproduce the nanostructure of the material. Notable examples of mesoscopic models of this type are coarse-grained models for molecular systems including polymers [19–21] and mesoscopic models for carbon nanotubes and nanofibrous materials [22–26]. The individual molecules (or mers in polymer molecules) and nanotube/nanofiber segments are chosen as the dynamic units in these models. The collective dynamic degrees of freedom that correspond to the motion of the "mesoparticles" are explicitly accounted for in mesoscopic models, while the internal degrees of freedom are either neglected or described by a small number of internal state variables. The description of the internal states of the mesoparticles and the energy exchange between the dynamic degrees of freedom and the internal state variables becomes important for simulations of nonequilibrium phenomena that involve fast energy deposition from an external source, heat transfer, or dissipation of mechanical energy.

Another group of mesoscopic models is aimed at a computationally efficient description of the evolution of the defect structures in crystalline materials. The mesoscopic models from this group include the discrete dislocation dynamics model for simulation of crystal plasticity [27–29] and a broad class of methods designed for simulation of grain growth, recrystallization, and associated microstructural evolution (e.g., phase field models, cellular automata, and kinetic Monte Carlo Potts models) [27, 28, 30]. Despite the apparent diversity of the physical principles and computational algorithms adopted in different models listed above, the common characteristic of these models is the focus on a realistic description of the behavior and properties of individual crystal defects (grain boundaries and dislocations), their interactions with each other, and the collective evolution of the totality of crystal defects responsible for the changes in the microstructure.

Two examples of mesoscopic models (one for each of the two types of the models discussed above) and their relevance to the investigation of nanomaterials are considered in more detail next.

Discrete Dislocation Dynamics

The purpose of the discrete dislocation dynamics (DD) is to describe the plastic deformation in crystalline materials, which is largely defined by the motions, interactions, and multiplication of dislocations. Dislocations are linear crystal defects that generate long-range elastic strain fields in the surrounding elastic solid. The elastic strain field is accounting for ~90 % of the dislocation energy and is responsible for the interactions of dislocations among themselves and with other crystal defects. The collective behavior of dislocations in the course of plastic deformation is defined by these long-range interactions as well as by a large number of local reactions (annihilation, formation of glissile junctions or sessile dislocation segments such as Lomer or Hirth locks) occurring when the anelastic core regions of the dislocation lines come into contact with each other. The basic idea of the DD model is to solve the dynamics of the dislocation lines in elastic continuum and to include information about the local reactions. The elementary unit in the discrete dislocation dynamics method is, therefore, a segment of a dislocation.

The continuous dislocation lines are discretized into segments, and the total force acting on each segment in the dislocation slip plane is calculated. The total force includes the contributions from the external force, the internal force due to the interaction with other dislocations and crystal defects that generate elastic fields, the "self force" that can be represented by a "line tension" force for small curvature of the dislocation, the Peierls force that acts like a friction resisting the

dislocation motion, and the "image" force related to the stress relaxation in the vicinity of external or internal surfaces. Once the total forces and the associated resolved shear stresses, τ^*, acting on the dislocation segments are calculated, the segments can be displaced in a finite difference time integration algorithm applied to the equations connecting the dislocation velocity, v, and the resolved shear stress, e.g., [28]

$$v = A\left(\frac{\tau^*}{\tau_0}\right)^m \exp\left(-\frac{\Delta U}{kT}\right), \tag{3}$$

when the displacement of a dislocation segment is controlled by thermally activated events (ΔU is the activation energy for dislocation motion, m is the stress exponent, and τ_0 is the stress normalization constant) or

$$v = \tau^* b/B, \tag{4}$$

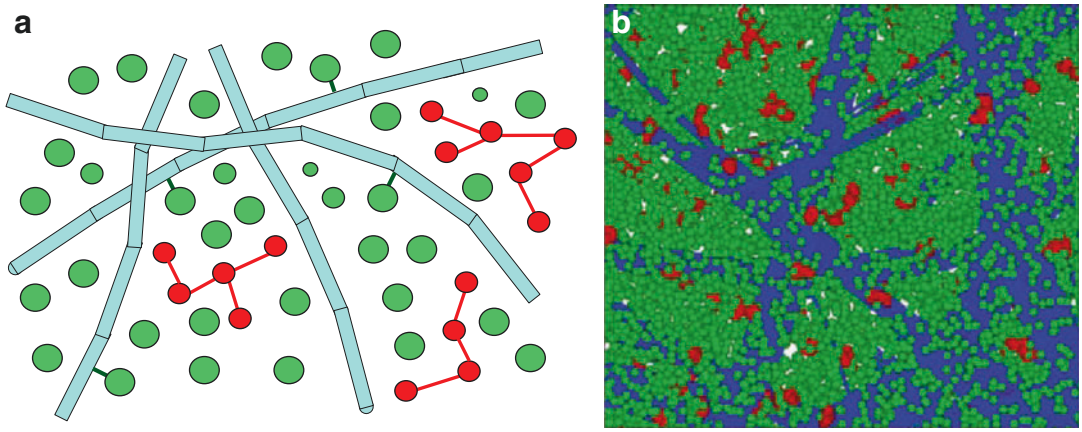
that corresponds to the Newtonian motion equation accounting for the atomic and electron drag force during the dislocation "free flight" between the obstacles (B is the effective drag coefficient and b is the magnitude of the Burgers vector of the dislocation).

Most of the applications of the DD model have been aimed at the investigation of the plastic deformation and hardening of single crystals (increase in dislocation density as a result of multiplication of dislocations present in the initial system). The extension of the DD modeling to nanomaterials is a challenging task as it requires an enhancement of the technique with a realistic description of the interactions between the dislocations and grain boundaries and/or interfaces as well as an incorporation of other mechanisms of plasticity (e.g., grain boundary sliding and twinning in nanocrystalline materials). There have only been several initial studies reporting the results of DD simulations of nanoscale metallic multilayered composites, e.g., [31]. Due to the complexity of the plastic deformation mechanisms and the importance of anelastic short-range interactions among the crystal defects in nanomaterials, the development of novel hybrid computational methods combining the DD technique with other mesoscopic methods is likely to be required for realistic modeling of plastic deformation in this class of materials.

Mesoscopic Model for Nanofibrous Materials

The design of new nanofibrous materials and composites is an area of materials research that is currently experiencing a rapid growth. The interest in this class of materials is fueled by a broad range of potential applications, ranging from fabrication of flexible/stretchable electronic and acoustic devices to the design of advanced nanocomposite materials with improved mechanical properties and thermal stability. The behavior and properties of nanofibrous materials are defined by the collective dynamics of the nanofibers and, in the case of nanocomposites, their interactions with the matrix. Depending on the structure of the material and the phenomenon of interest, the number of nanofibers that has to be included in the simulation in order to ensure a reliable prediction of the effective macroscopic properties can range from several hundreds to millions. The direct atomic-level simulation of systems consisting of large groups of nanofibers (the path shown by the green arrow #2 in Fig. 1) is beyond the capabilities of modern computing facilities. Thus, an alternative two-step path from atomistic investigation of individual structural elements and interfacial properties to the continuum material description through an intermediate mesoscopic modeling (blue arrows #3 in Fig. 1) appears to be the most viable approach to modeling of nanofibrous materials. An example of a mesoscopic computational model recently designed and parameterized for carbon nanotube (CNT)-based materials is briefly discussed below.

The mesoscopic model for fibrous materials and organic matrix nanocomposites adopts a coarse-grained description of the nanocomposite constituents (nanofibers and matrix molecules), as schematically illustrated in Fig. 3. The individual CNTs are represented as chains of stretchable cylindrical segments [22], and the organic matrix is modeled by a combination of the



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 3 Schematic representation of the basic components of the dynamic mesoscopic model of a CNT-based nanocomposite material (**a**) and a corresponding molecular-level view of a part of the system where a network of CNT bundles (*blue* color) is embedded into an organic matrix (*green* and *red* color) (**b**)

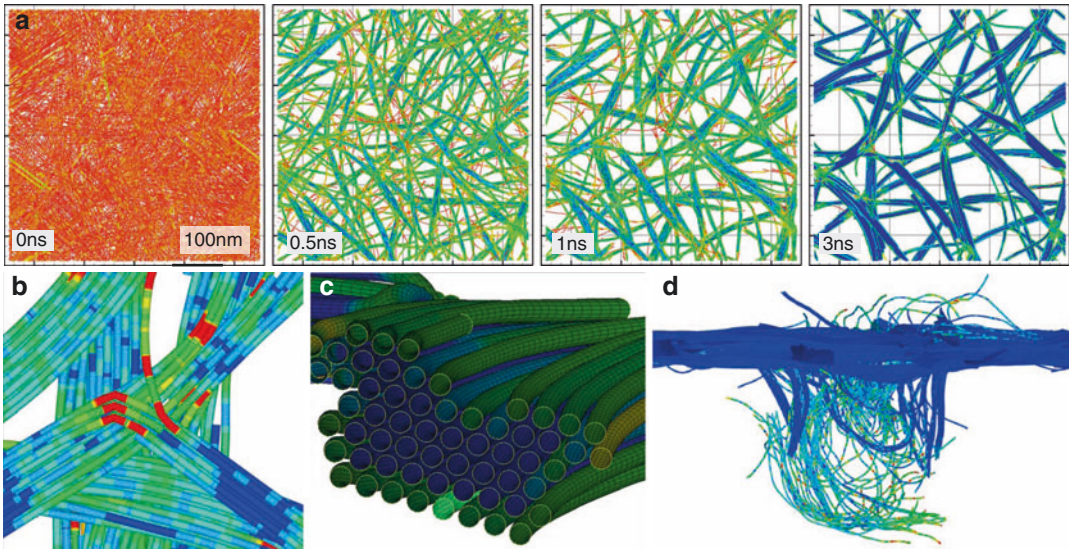
conventional "bead-and-spring" model commonly used in polymer modeling [19, 20] and the "breathing sphere" model developed for simulation of simple molecular solids [21] and polymer solutions [32].

The degrees of freedom, for which equations of motion are solved in dynamic simulations or Metropolis Monte Carlo moves are performed in simulations aimed at finding the equilibrium structures, are the positions of nodes defining the segments, the positions of the molecular units, and the radii of the spherical particles in the breathing sphere molecules. The potential energy of the system can be written as

$$U = U_{T(\text{int})} + U_{T-T} + U_{M-M} + U_{M(\text{int})} + U_{M-T} \quad (5)$$

where $U_{T(\text{int})}$ is the potential that describes the internal strain energy associated with stretching and bending of individual CNTs, U_{T-T} is the energy of intertube interactions, U_{M-M} is the energy of chemical and nonbonding interactions in the molecular matrix, $U_{M(\text{int})}$ is the internal breathing potential for the matrix units, and U_{M-T} is the energy of matrix–CNT interaction that can include both nonbonding van der Waals interactions and chemical bonding. The internal CNT potential $U_{T(\text{int})}$ is parameterized based on the results of atomistic simulations [22] and accounts for the transition to the anharmonic regime of stretching (nonlinear stress–strain dependence), fracture of nanotubes under tension, and bending buckling [33]. The intertube interaction term U_{T-T} is calculated based on the tubular potential method that allows for a computationally efficient and accurate representation of van der Waals interactions between CNT segments of arbitrary lengths and orientation [23]. The general procedure used in the formulation of the tubular potential is not limited to CNTs or graphitic structures. The tubular potential (and the mesoscopic model in general) can be parameterized for a diverse range of systems consisting of various types of nano- and microtubular elements, such as nanotubes, nanorodes, and microfibers.

Simulations performed with the mesoscopic model demonstrate that the model is capable of simulating the dynamic behavior of systems consisting of thousands of CNTs on a timescale extending up to tens of nanoseconds. In particular, simulations performed for systems composed of randomly distributed and oriented CNTs predict spontaneous self-assembly of CNTs into continuous networks of bundles with partial hexagonal ordering of CNTs in the bundles, Fig. 4a–c [23, 33]. The bending buckling of CNTs (e.g., see



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 4 Snapshots from mesoscopic simulations of systems consisting of (10,10) single-walled carbon nanotubes: (**a**) spontaneous self-organization of CNTs into a continuous network of CNT bundles (CNT segments are colored according to the local intertube interaction energy) [23]; (**b**) an enlarged view of a structural element of the CNT network (CNT segments colored according to the local radii of curvature and the *red* color marks the segments adjacent to buckling kinks) [33]; (**c**) a cross section of a typical bundle showing a hexagonal arrangement of CNTs in the bundle [23]; (d) snapshot from a simulation of a high-velocity impact of a spherical projectile on a free-standing thin CNT film [34]

Fig. 4b) is found to be an important factor responsible for the stability of the network structures formed by defect-free CNTs [33]. The structures produced in the simulations are similar to the structures of CNT films and buckypaper observed in experiments. Note that an atomic-level simulation of a system similar to the one shown in the left panel of Fig. 4 would require $\sim 2.5 \times 10^9$ atoms, making such simulation unfeasible.

Beyond the structural analysis of CNT materials, the development of the mesoscopic model opens up opportunities for investigation of a broad range of important phenomena. In particular, the dynamic nature of the model makes it possible to perform simulations of the processes occurring under conditions of fast mechanical loading (blast/impact resistance, response to the shock loading, etc.), as illustrated by a snapshot from a simulation of a high-velocity impact of a spherical projectile on a free-standing thin CNT film shown in Fig. 4d [34]. With a proper parameterization, the mesoscopic model can also be adopted for calculation of electrical and thermal transport properties of complex nanofibrous materials [35, 36].

Future Research Directions

The examples of application of the atomistic and mesoscopic computational techniques, briefly discussed above, demonstrate the ability of computer modeling to provide insights into the complex processes that define the behavior and properties of nanostructured materials. The fast advancement of experimental methods capable of probing nanostructured materials with high spatial and temporal resolution is an important factor that allows for verification of computational predictions and stimulates the improvement of the computational models. With further innovative development of computational methodology and the steady growth of the available computing resources, one can expect that both atomistic and

mesoscopic modeling will continue to play an increasingly important role in nanomaterials research.

In the area of atomistic simulations, the development of new improved interatomic potentials (often with the help of *ab initio* electronic structure calculations, red arrow #1 in Fig. 1) makes material-specific computational predictions more accurate and enables simulations of complex multicomponent and multiphase systems. Further progress can be expected in two directions that are already actively pursued: (1) large-scale MD simulations of the fast dynamic phenomena in nanocrystalline materials (high strain rate mechanical deformation, shock loading, impact resistance, response to fast heating, etc.) and (2) detailed investigation of the atomic structure and properties of individual structural elements in various nanomaterials (grain boundaries and interfaces, nanotubes, nanowires, and nanoparticles of various shapes). The information obtained in large-scale atomistic simulations of nanocrystalline materials can be used to formulate theoretical models translating the atomic-level picture of material behavior to the constitutive relations describing the dependence of the mechanical and thermal properties of these materials on the grain size distribution and characteristics of nanotexture (green arrow #2 in Fig. 1).

The results of the detailed analysis of the structural elements of the nanocomposite materials can be used in the design and parameterization of mesoscopic models, where the elementary units treated in the models correspond to building blocks of the nanostructure (elements of grain boundaries, segments of dislocations, etc.) or groups of atoms that have some distinct properties (belong to a molecule, a mer unit of a polymer chain, a nanotube, a nanoparticle in nanocomposite material, etc.). The design of novel system-specific mesoscopic models capable of bridging the gap between the atomistic modeling of structural elements of nanostructured materials and the continuum models (blue arrows #3 in Fig. 1) is likely to become an important trend in the computational investigation of nanomaterials. To achieve a realistic description of complex processes occurring in nanomaterials, the description of the elementary units of the mesoscopic models should become more flexible and sophisticated. In particular, an adequate description of the energy dissipation in nanomaterials can only be achieved if the energy exchange between the atomic degrees of freedom, excluded in the mesoscopic models, and the coarse-grained dynamic degrees of freedom is accounted for [34]. A realistic representation of the dependence of the properties of the mesoscopic units of the models on local thermodynamic conditions can also be critical in modeling of a broad range of phenomena.

In general, the optimum strategy in investigation of nanomaterials is to use a well-integrated multiscale computational approach combining the *ab initio* and atomistic analysis of the constituents of nanostructure with mesoscopic modeling of the collective dynamics and kinetics of the structural evolution of the material and leading to the improved theoretical understanding of the factors controlling the effective material properties. It is the improved understanding of the connections between the processes occurring at different time and length scales that is likely to be the key factor defining the pace of progress in the area of computational design of new nanocomposite materials.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Active Carbon Nanotube-Polymer Composites
- ▶ Carbon-Nanotubes
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Modeling Thermal Properties of Carbon Nanostructure Composites
- ▶ Molecular Modeling
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Plasticity Theory at Small Scales
- ▶ Reactive Empirical Bond-Order Potentials
- ▶ Self-Assembly of Nanostructures

References

1. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Clarendon, Oxford (1987)
2. Frenkel, D., Smit, B.: Understanding Molecular Simulation: From Algorithms to Applications. Academic, San Diego (1996)
3. Kelires, P.C., Tersoff, J.: Equilibrium alloy properties by direct simulation: oscillatory segregation at the Si-Ge(100) 2×1 surface. Phys. Rev. Lett. **63**, 1164–1167 (1989)
4. Combe, N., Jensen, P., Pimpinelli, A.: Changing shapes in the nanoworld. Phys. Rev. Lett. **85**, 110–113 (2000)
5. Liu, H., Lin, Z., Zhigilei, L.V., Reinke, P.: Fractal structures in fullerene layers: simulation of the growth process. J. Phys. Chem. C **112**, 4687–4695 (2008)
6. Germann, T.C., Kadau, K.: Trillion-atom molecular dynamics becomes a reality. Int. J. Mod. Phys. C **19**, 1315–1319 (2008)
7. http://folding.stanford.edu/
8. Yamakov, V., Wolf, D., Phillpot, S.R., Mukherjee, A. K., Gleiter, H.: Dislocation processes in the deformation of nanocrystalline aluminium by molecular-dynamics simulation. Nat. Mater. **1**, 45–49 (2002)
9. Dongare, A.M., Rajendran, A.M., LaMattina, B., Zikry, M.A., Brenner, D.W.: Atomic scale studies of spall behavior in nanocrystalline Cu. J. Appl. Phys. **108**, 113518 (2010)
10. Ju, S., Liang, X.: Investigation of argon nanocrystalline thermal conductivity by molecular dynamics simulation. J. Appl. Phys. **108**, 104307 (2010)
11. Xiao, S., Hu, W., Yang, J.: Melting behaviors of nanocrystalline Ag. J. Phys. Chem. B **109**, 20339–20342 (2005)
12. Lin, Z., Bringa, E.M., Leveugle, E., Zhigilei, L.V.: Molecular dynamics simulation of laser melting of nanocrystalline Au. J. Phys. Chem. C **114**, 5686–5699 (2010)
13. Rudd, R.E., Broughton, J.Q.: Coarse-grained molecular dynamics and the atomic limit of finite elements. Phys. Rev. B **58**, R5893–R5896 (1998)
14. Zwanzig, R.: Memory effects in irreversible thermodynamics. Phys. Rev. **124**, 983–992 (1961)
15. Mori, H.: Transport, collective motion, and Brownian motion. Prog. Theor. Phys. **33**, 423–455 (1965)
16. Hijón, C., Español, P., Vanden-Eijnden, E., Delgado-Buscalioni, R.: Mori-Zwanzig formalism as a practical computational tool. Faraday Discuss. **144**, 301–322 (2010)
17. Akkermans, R.L.C., Briels, W.J.: Coarse-grained dynamics of one chain in a polymer melt. J. Chem. Phys. **113**, 6409–6422 (2000)
18. Li, Z., Bian, X., Caswell, B., Karniadakis, G.E.: Construction of dissipative particle dynamics models for complex fluids via the Mori–Zwanzig formulation. Soft Matter. **10**, 8659–8672 (2014)
19. Peter, C., Kremer, K.: Multiscale simulation of soft matter systems. Faraday Discuss. **144**, 9–24 (2010)
20. Colbourn, E.A. (ed.): Computer Simulation of Polymers. Longman Scientific and Technical, Harlow (1994)
21. Zhigilei, L.V., Leveugle, E., Garrison, B.J., Yingling, Y.G., Zeifman, M.I.: Computer simulations of laser ablation of molecular substrates. Chem. Rev. **103**, 321–348 (2003)
22. Zhigilei, L.V., Wei, C., Srivastava, D.: Mesoscopic model for dynamic simulations of carbon nanotubes. Phys. Rev. B **71**, 165417 (2005)
23. Volkov, A.N., Zhigilei, L.V.: Mesoscopic interaction potential for carbon nanotubes of arbitrary length and orientation. J. Phys. Chem. C **114**, 5513–5531 (2010)
24. Buehler, M.J.: Mesoscale modeling of mechanics of carbon nanotubes: self-assembly, self-folding, and fracture. J. Mater. Res. **21**, 2855–2869 (2006)
25. Ostanin, I., Ballarini, R., Potyondy, D., Dumitrică, T.: A distinct element method for large scale simulations of carbon nanotube assemblies. J. Mech. Phys. Solids **61**, 762–782 (2013)
26. Torabi, H., Radhakrishnan, H., Mesarovic, S.D.J.: Micromechanics of collective buckling in CNT turfs. J. Mech. Phys. Solids **72**, 144–160 (2014)
27. Raabe, D.: Computational Materials Science: The Simulation of Materials Microstructures and Properties. Wiley-VCH, Weinheim/New York (1998)
28. Kirchner, H.O., Kubin, L.P., Pontikis, V. (eds.): Computer simulation in materials science. Nano/meso/macroscopic space and time scales. Kluwer Academic, Dordrecht/Boston/London (1996)
29. Groh, S., Zbib, H.M.: Advances in discrete dislocations dynamics and multiscale modeling. J. Eng. Mater. Technol. **131**, 041209 (2009)
30. Holm, E.A., Battaile, C.C.: The computer simulation of microstructural evolution. J Miner. Met. Mater. Soc. **53**, 20–23 (2001)
31. Akasheh, F., Zbib, H.M., Hirth, J.P., Hoagland, R.G., Misra, A.: Dislocation dynamics analysis of dislocation intersections in nanoscale metallic multilayered composites. J. Appl. Phys. **101**, 084314 (2007)
32. Leveugle, E., Zhigilei, L.V.: Molecular dynamics simulation study of the ejection and transport of polymer molecules in matrix-assisted pulsed laser evaporation. J. Appl. Phys. **102**, 074914 (2007)
33. Volkov, A.N., Zhigilei, L.V.: Structural stability of carbon nanotube films: the role of bending buckling. ACS Nano **4**, 6187–6195 (2010)
34. Jacobs, W.M., Nicholson, D.A., Zemer, H., Volkov, A. N., Zhigilei, L.V.: Acoustic energy dissipation and thermalization in carbon nanotubes: atomistic modeling and mesoscopic description. Phys. Rev. B **86**, 165414 (2012)
35. Volkov, A.N., Zhigilei, L.V.: Scaling laws and mesoscopic modeling of thermal conductivity in carbon nanotube materials. Phys. Rev. Lett. **104**, 215902 (2010)
36. Volkov, A.N., Zhigilei, L.V.: Heat conduction in carbon nanotube materials: strong effect of intrinsic thermal conductivity of carbon nanotubes. Appl. Phys. Lett. **101**, 043113 (2012)

Computational Systems Bioinformatics for RNAi

Zheng Yin, Yubo Fan and Stephen TC Wong
Center for Bioengineering and Informatics, Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX, USA

Synonyms

Automatic data analysis workflow for RNAi; Systems level data mining for RNAi

Definition

Computational systems bioinformatics for RNAi screening and therapeutics is defined as complete computational workflow applicable to the hypothesis generation from large-scale data from image-based RNAi screenings as well as the improvement of RNAi-based therapeutics; the workflow includes automatic image analysis compatible to large-scale cell image data, together with unbiased statistical analysis and gene function annotation.

Introduction

RNA interference (RNAi) defines the phenomenon of small RNA molecules binding to its complementary sequence in certain messenger RNA, recruiting a specific protein complex to dissect the whole mRNA and thus silencing the expression of the corresponding gene. It is a highly conserved system within living cells to quantitatively control the activity of genes. In 1998, Fire et al. first clarified the causality of this phenomenon and named it as RNAi [1], and the following decade saw RNAi evolving into a powerful tool for gene function study. In 2006, Fire and Mello were awarded Nobel Prize in Physiology or Medicine; and by 2007, scientific papers using high-throughput screening based on RNAi kept piling up while clinical trials of RNAi-based therapeutics on various diseases raised the expectation on a trend of soon-to-come "super drugs." Unfortunately, by late 2009 nearly all the first trend clinical trials have been terminated, meanwhile, researchers are struggling to effectively quantify the high-content information obtained from RNAi-based screening experiments.

Although facing obstacles, it is still believed that the combination of nanotechnology and systems biology would restore and amplify the glory of RNAi on both research and therapeutic areas. This entry will summarize the challenges facing RNAi-based therapeutics – especially the difficulty of delivery and some possible solution through chemoinformatics in nanometer scale; also difficulties facing RNAi-based high-content screening will be reviewed with suggestions on possible solutions.

RNAi-Based Therapeutics: How-to and What-to Deliver

Currently, various RNAi-based therapeutics are being tested in clinical trials but before the application becomes clinical, several problems have first to be solved. One of the critical challenges is how to specifically and effectively deliver the objects into the targeted cells. Serious side-effects in patients could be caused by off-target effects or immune response as reported in literatures [2, 3].

Apparently, the therapeutic goal is to achieve RNAi therapy, that is, systemically administered nucleic acids must survive in circulation long enough to reach their target tissue, enter the desired cells, "escape" their endosome or delivery packaging, and finally become incorporated into the RNA-induced silencing complex (RISC) – a towering task, and surprisingly, researchers have advanced a number of plausible solutions in recent years, including the use of specialized nanoparticle filled with an RNAi-based cancer therapy to target human cancer cells and silence the target gene. However, what making the task even more difficult for therapeutics is: everything now happens in vivo.

Chemoinformatics Solutions

Libraries have been built based on natural products and combinatorial chemistry with millions of compounds to date. The compound library is used to screen and locate small molecules to bind a particular protein, RNA or DNA. Virtual screening utilizing molecule fragments even can design unknown compounds with reasonable binding affinity to desired targets. However, these large-scale screening compounds tend to bind multiple targets (off-target effects) even after optimizations [4]. Also, it is virtually impossible to derive a solid algorithm or theory to screen and locate all bindings and inhibitions and their combinations every protein currently know. The lack of genome-scale coverage is a clear disadvantage of a compound based assay. Along with the lack of specificity comes the challenge of on-target potency. Even after a target has been identified the drug-like compounds have to be further engineered to increase efficiency and decrease off-target activities.

To a large extent, chemoinformatic approaches, where the target of a compound can be predicted by in silico alignment, modeling algorithms, and virtual screening, provide a more efficient way to screen compounds to the desired targets. A successful application has been reported for the inhibition of SARS protease [5]. The chemoinformatics approach excels at optimization around a small number of well defined targets, while RNAi approaches can more readily identify unknown pathways and phenotypes with no prior knowledge of the target.

Data Analysis for RNAi HCS: Challenge from Millions of Cells

Large volumes of datasets generated from RNAi HCS of RNAi prohibits manual or even semi-manual analysis; thus, automated data analysis is desperately needed [6, 7]. In the context of genome-wide RNAi HCS using cultured cells from Drosophila, a series of automated methods on cell image processing [8], online phenotype discovery [9, 10], cell classification, and gene function annotation [11] have been developed. All these methods are integrated into an automated data analysis pipeline, G-CellIQ (Genomic Cellular Image Quantitator), to support genome-wide RNAi HCS.

A lot of decisions need to be made en route to a genome-scale RNAi screens, including the selections of appropriate animal models, reagents for igniting RNAi, screening formats, and type of readouts (see [12] for a review). The focus here is arrayed high-content RNAi screening, a systematic screen with reagents spotted in 384- or 96-well plates and where each gene or gene group is knocked down individually. The readout of HCS is obtained through microscopy, which captures multiple phenotypic features simultaneously [13]. Compared with other screening approaches, RNAi HCS can offer broader insight into cellular physiology and provide informative and continuous phenotypic data generated by RNAi. However, the automated processing and analysis of the magnitude of image data presents great challenges to applying such findings at the genome level.

Image Processing: Cell Segmentation and Quantification

The scale of datasets has always been a huge obstacle. For example, in [6], approximately 17000 overlapping cells were segmented semi-manually across 10 months from an HCS targeting around 200 genes or gene combinations. However, such low-dimensional screens are not genome-wide, as they can only cover 1–2 % of the genome. For automatic data analysis, the following work needs to be accomplished.

1. Preprocessing: where empty images, incomplete cells, and other artifacts are identified and discarded
2. Cell segmentation: where dense or overlapping cells are segmented accurately
3. Feature extraction: where informative features are extracted to quantify cell morphology

Cell segmentation is the cornerstone for the whole data analysis workflow. While existing image analysis methods can handle the processing of standard images, they are limited in their scope

and capability to handle genome-wide RNAi HCS analysis. Thresholding methods basically set a cutoff on the intensity of pixels and classify them into background and foreground (cell), and they may fail due to uneven background and illumination levels. Rule-based correction on over- and under-segmentation starts from relatively simple segmentation methods (like watershed), and use a series of heuristic rules, like distance between neighborhood nuclei and properties of putative cell boundaries, such methods suffer from difficulties in devising rules to merge the cell cytoplasm.

Phenotype Identification, Validation, and Classification

Given the quantified features describing cell morphology, the following work is essential to address the biological function of morphological profiles.

1. Define biologically meaningful phenotypes to compose cell populations based on single cell morphology.
2. Model existing phenotypes and identify novel phenotypes online to continuously generate new data.
3. Assign cellular phenotypes into different sub-phenotypes to address morphological changes caused by RNAi treatment.

Statistical tests, artificial neural networks [6], Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [14], genetic algorithms, and various other methods have been used to select or extract informative subsets of features and model certain phenotypes. However, phenotypes are usually defined a priori from pilot datasets. Human intervention is currently necessary for image-based datasets of genetic or chemical perturbations where the dynamic range of cellular phenotypes cannot be predicted before data collection. Failing to accurately measure phenotypic variations will cause concomitant classification errors and mislead functional analysis; and it is impossible to perform manual analysis during the screening process where millions of images are acquired. Thus, the ability of these screens to identify new phenotypes is greatly limited [9].

Statistical Analysis and Gene Function Annotation

RNAi HCS inherits various statistical questions from traditional high-throughput screening, and the properties of image readouts raise specific challenges.

1. Summarizing gene function scores from quantified morphological change
2. Data triage and normalization based on readout from positive and negative control wells
3. Repeatability test and consolidation of scores from biological replicates
4. Cluster analysis, visualization, and biological interpretation of results

A comprehensive review [15] summarizes different dataset generated by RNAi screens and traditional small molecule screens. It also reviews statistical analysis methods applicable to most of problems outlined above. Quantitative morphological signatures (QMS) [6] represent the efforts on interpreting RNAi HCS datasets: use the similarity score to a panel of existing phenotypes to explore the broad phenotypic space. However, problems remain open when the discriminative ability of such scores is confounded by multiple phenotypes following a single RNAi treatment, so repeatability tests can become more complicated [6]. The use of publicly available databases on drugs and disease-related biological processes to interpret RNAi HCS datasets also remains unresolved.

G-CellIQ: An Integrated Automated Data Analysis Tool for RNAi High-Content Screening

Computational Architecture of G-CellIQ

G-CellIQ (**G**enomic **CELL**ular **I**maging **Q**uantitator) is developed to process large volumes of digital images generated from large-scale HCS studies. The workflow for G-CellIQ can be simplified as "three modules handling three databases." Images generated from HCS are stored in a Raw image database; the Image processing/cell morphology quantification

module segments each image into single cells and creates a quantified cell database; the Phenotype modeling and cell classification module compares each cell's morphology to a panel of "reference" phenotypes (which can be defined both manually and automatically) and generates a morphology score; the annotation of gene function module then summarize single cell scores into scores for cell populations, images and wells, and the consolidated scores for involved genes form a single gene function profile database.

Image Processing and Cell Morphology Quantification

A segmentation method consisting of nuclear segmentation, cell body segmentation, and over-segmentation correction [8] is used in G-CELLIQ. Each cell body is described by 211 morphological features; automatic image quality control is applied to filter images where the signal from certain channels is extremely dark or bright or cells are located at the edge of an image.

Nuclear segmentation cell nuclei are first separated from background using a binarization method which implements an adaptive thresholding [8]. To segment clustered nuclei, the nuclei shape and intensity information are integrated into a combined image and processed with Gaussian filter and the nuclei centers are detected as local maxima in the gradient vector field (GVF); after that, nuclei are segmented using the marker-controlled seeded-watershed method.

Cell body segmentation Preliminary cell body segmentation is done using an adaptive thresholding algorithm. Due to the large size of HCS dataset, seeded-watershed method is used to segment the touching cell bodies. Results from nuclei segmentation are used as the seed information [16].

Over-segmentation correction is necessary when there are multiple nuclei existing within cells. For cell segments smaller than a given size threshold, its neighboring cell segments sharing the longest common boundary are determined and the intensity variation in a rectangular region across the common boundary of touching cell segments are calculated. If the intensity variation was smaller than a given threshold, the corresponding cell segments are merged.

Feature extraction The detailed shape and boundary information of nuclei and cell bodies is obtained through the proposed segmentation method. To capture the geometric and appearance properties, 211 morphology features belonging to five categories were extracted following [11].

Online Phenotype Discovery, Phenotype Modeling, and Cell Classification

Expert opinion is implemented to identify a panel of reference phenotypes while candidate groups of informative features are selected to describe typical phenotypes. SVM-RFE and GA-SVM method are used to select the feature sets with the best performance for cross validation. A series of SVM classifiers are trained to differentiate the reference phenotypes from all others, and a continuous value rather than binary class label is used as the output of SVM to indicate each cell's morphological similarity to typical phenotypes.

A novel method is designed to do online phenotype discovery [9, 10]; it is based on online phenotype modeling and iterative phenotype merging. For the modeling part, each existing phenotype is modeled through a Gaussian Mixture Model (GMM), and each model is continuously updated according to information of newly incorporated cells with a minimum classification error (MCE) method. For the merging part, the newly generated cell population is iteratively combined with each existing phenotype, one at a time. Then an improved gap statistics method is used to identify the number of possible phenotypes in the combination. Through cluster analysis, some of the cells in the new populations are assigned into the same cluster as samples from existing phenotypes, and those cells are merged by existing phenotypes to help update the phenotype models. On the other hand, some cells are never merged and remain as candidate novel phenotype for validation by statistical tests [9, 10].

Annotation of Gene Function

After assigning each cell a score vector corresponding to its similarity to reference

phenotypes, all cells in control condition are pooled to model a baseline for cell morphology. All morphology scores are normalized to the Z-score relative to this control baseline. The scores for the qualified cells are then averaged to form scores for each well. In order to select repeatable wells from those undergoing the same dsRNA treatment, a series of repeatability tests are applied to scores for different well. The weighted average of the scores is then calculated for the repeatable wells and generated scores for each treatment condition (TC). Similar procedure consolidates the score from biological replicate TCs to form a score vector for each gene. Hierarchical clustering is implemented group genes with similar function scores into the same phenocluster.

RNAi HCS Applying G-CellIQ

Since more than 75 % of human disease genes have *Drosophila* orthologs, the pursuit of genes involved in normal fly morphogenesis and migration is expected to reveal mechanisms conserved in humans [17]. An example of using G-CellIQ in the context of RNAi HCS using cultured *Drosophila* cell lines is presented next.

Regulatory of cell shape change. A genome-scale RNAi screening is carried out for regulators of *Drosophila* cell shape. Drosophila Kc187 cell lines are utilized in the screen and wild-type cells have hemocyte-like properties. Using dsRNA to target and inhibit the activity of specific genes/proteins, the role of individual genes in regulating morphology can be systematically determined.

Work in [9, 10] relies on part of the genome dataset to target the group of kinase-phosphatases in order to develop and validate online phenotype discovery methods. A panel of five existing phenotypes is set by expert labeling and online phenotype discovery. In order to address the level of penetrance in this dataset, the phenotypic scores for single cells are sorted according to similarity to wild-type cells, and if a certain well has significantly less (at least one standard-deviation) wild-type cells than control wells, wild-type cells from this well are removed to reveal the phenotypic change relating to RNAi treatment.

Roles of Rho family small GTPases in development and cancer. Three automated and *quantitative* genome-wide screens for dsRNAs that induce the loss of the Rho-induced cytoskeletal structures (lamellipodia, filopodia, and stress fibers) are performed to identify putative Rho protein effectors. Candidate effectors identified in such image-based screens are readily validated in the context of the whole organism using the large number of mutant fly lines coupled with the vast arrays of in vivo techniques available to fly biologists [18].

Following image segmentation, feature extraction, classification of single cells, and scoring of individual wells, an image descriptor is assigned to each gene. Hierarchical clustering can then be used to cluster the image descriptor and identify groups of genes when targeted by RNAi result in quantitatively similar morphologies. Previous studies reported in [6] demonstrated that clustering results identifies groups of functionally related genes that operate in similar signaling pathways. These groups of genes are termed as "Phenoclusters" [19]. To validate the hypothesis generation ability of G-CELLIQ, 32 dsRNAs/wells are randomly selected from a dataset where individual kinases and phosphatases were inhibited by RNAi. Each dsRNA/well was assigned an image descriptor, and hierarchical clustering was used to group genes/wells. dsRNAs in this analysis clustered into two broad groups. One group of 19 conditions included 10/10 control conditions, as well as dsRNAs targeting the Insulin receptor (InR). Strikingly, the other large cluster of 13 conditions included 3/3 dsRNAs previously identified in a genome-wide screen for regulators of MAPK/ERK activation downstream of the EGF/EGFR activity [20]. These results demonstrate that automated high-throughput imaging can discriminate distinct morphologies and be used to model functional relationships between signaling molecules.

Cross-References

▶ RNAi in Biomedicine and Drug Delivery

References

1. Fire, A., et al.: Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature **391**(6669), 806–811 (1998)
2. Hornung, V., et al.: Sequence-specific potent induction of IFN-[alpha] by short interfering RNA in plasmacytoid dendritic cells through TLR7. Nat. Med. **11**(3), 263–270 (2005)
3. Grimm, D., et al.: Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. Nature **441**(7092), 537–541 (2006)
4. Copeland, R.A., Pompliano, D.L., Meek, T.D.: Drug-target residence time and its implications for lead optimization. Nat. Rev. Drug Discov. **5**(9), 730–739 (2006)
5. Plewczynski, D., et al.: In silico prediction of SARS protease inhibitors by virtual high throughput screening. Chem. Biol. Drug Des. **69**(4), 269–279 (2007)
6. Bakal, C., et al.: Quantitative morphological signatures define local signaling networks regulating cell morphology. Science **316**, 1753–1756 (2007)
7. Zhou, X., Wong, S.T.C.: Computational systems bioinformatics and bioimaging for pathway analysis and drug screening. Proc. IEEE **96**(8), 1310–1331 (2008)
8. Li, F.H., et al.: High content image analysis for human H4 neuroglioma cells exposed to CuO nanoparticles. BMC Biotechnol. **7**, 66 (2007)
9. Yin, Z., et al.: Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens. BMC Bioinform. **9**(1), 264 (2008)
10. Yin, Z., et al.: Online phenotype discovery based on minimum classification error model. Pattern Recogn. **42**(4), 509–522 (2009)
11. Wang, J., et al.: Cellular phenotype recognition for high-content RNA interference genome-wide screening. J. Mol. Screen. **13**(1), 29–39 (2008)
12. Perrimon, N., Mathey-Prevot, B.: Applications of high-throughput RNAi screens to problems in cell and developmental biology. Genetics **175**, 7–16 (2007)
13. Carpenter, A.E., Sabatini, D.M.: Systematic genome-wide screens of gene function. Nat. Rev. Genet. **5**(1), 11–22 (2004)
14. Loo, L., Wu, L., Altshuler, S.: Image based multivariate profiling of drug responses from single cells. Nat. Methods **4**(5), 445–453 (2007)
15. Birmingham, A., et al.: Statistical methods for analysis of high-throughput RNA interference screens. Nat. Methods **6**(8), 569–575 (2009)
16. Yan, P., et al.: Automatic segmentation of RNAi fluorescent cellular images with interaction model. IEEE Trans. Inf. Technol. Biomed. **12**(1), 109–117 (2008)
17. Reiter, L.T., et al.: A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. Genome Res. **11**, 1114–1125 (2001)
18. Bier, E.: Drosophila, the golden bug, emerges as a tool for human genetics. Nat. Rev. Genet. **6**(1), 9–23 (2005)
19. Piano, F., et al.: Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. Curr. Biol. **12**(22), 1959–1964 (2002)
20. Friedman, A., Perrimon, N.: Functional genomic RNAi screen for novel regulators of RTK/ERK signaling. Nature **444**, 230–234 (2006)

Computer Modeling and Simulation of Materials

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Computer Modeling of Electrochemical Systems

▶ Electrochemical Interfaces for Energy Storage and Conversion

Concentration Polarization

▶ Concentration Polarization at Micro-/Nanofluidic Interfaces

Concentration Polarization at Micro-/Nanofluidic Interfaces

Vishal V. R. Nandigana and N. R. Aluru
Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana – Champaign, Urbana, IL, USA

Synonyms

Concentration polarization; Micro-/Nanofluidic devices; Nonlinear electrokinetic transport

Definition

Concentration polarization (CP) is a complex phenomenon observed at the interfaces of micro/nanofluidic devices due to the formation of significant concentration gradients in the electrolyte solution resulting in accumulation and depletion of ions near the interfaces.

This chapter provides an overview of the underlying theory and physics that is predominantly observed on the integration of microfluidic channels with nanofluidic devices. Nonlinear electrokinetic transport and concentration polarization phenomenon are discussed in detail along with recent advancements in utilizing this phenomenon for designing novel devices.

Electrical Double Layer (EDL) and Electroosmotic Flow

A solid in contact with an aqueous solution acquires a surface charge (σ_s) due to the dissociation of ionizable groups on the solid walls. The fixed surface charge on the solid surface in contact with the liquid develops a region of counterions (ions with charges opposite to the solid surface) in the liquid to maintain the electroneutrality at the solid–liquid interface. This screening region is denoted as the electrical double layer (EDL) or Debye length (DL). For instance, in the case of silica channel with *KCl* electrolyte solution, the dissociation of silanol groups would make the channel negatively charged and affect the distribution of K^+ counterions in the solution. The layer where the ions get strongly attracted toward the channel surface due to the electrostatic force is called the inner layer with a typical thickness of one ion diameter. The outer Helmholtz plane separating the liquid and the diffusive layer constitute the liquid side part of the EDL. The ionic species in the diffusive layer are influenced by the local electrostatic potential, and the species distribution at equilibrium can be described by the Boltzmann equation. The thickness of the diffusive layer spans between 1 and 100 nm. The EDL thickness (λ_D) is given by [1]:

$$\lambda_D = \left(\frac{\varepsilon_0 \varepsilon_r RT}{F^2 \sum_{i=1}^{m} z_i^2 c_0} \right) 1/2 \qquad (1)$$

where F is Faraday's constant, z_i is the valence of ionic species i, c_0 is the bulk concentration of the electrolyte solution, ε_0 is the permittivity of free space, ε_r is the relative permittivity of the medium, m is the total number of ionic species, R is the universal gas constant, and T is the absolute temperature.

The thickness of the EDL plays a significant role in the transport of miniaturized devices. In microfluidic channels, the fluid transport is often controlled by electric fields, as it eliminates the use of external mechanical devices [1]. The electric field aids better control when compared to using pressure-driven techniques. Furthermore, the electric fields overcome the high pressures needed to transport the fluid at such length scales as the pressure follows a power–law relation with respect to the height (h) of the channel. The electric field acts on the charged counterions present at the interface of solution and stationary charged wall (i.e., at the EDL regions), resulting in the motion of the fluid which is referred as electroosmotic flow (EOF) [1]. As the EDL thickness in these channels is much smaller compared to their height $\left(\frac{h}{\lambda_D} \gg 1\right)$, the fluid flow has a plug-like flow characteristic. However, recent advancements in the fabrication technology [2] have motivated researchers around the globe to investigate the transport phenomenon in channel sizes of the order of few hundreds of nanometers. Transport in these devices is referred to as "nanofluidics." The electrical double layer in these devices spans much of the diameter or channel height leading to many interesting transport phenomena compared to its microscopic counterpart. The electroosmotic velocity no longer follows a plug-like flow characteristic but follows a Poiseuille-like (parabolic) characteristic as the electrokinetic body force is not just confined to a thin layer adjacent to the channel surface. Along with the aforementioned difference, the micro and nanofluidic systems also exhibit a different ion transport characteristic which is discussed below. In nanofluidic systems, as the EDL thickness

becomes comparable to the channel height $\left(\frac{h}{\lambda_D} \approx 1\right)$, there is a predominant transport of the counterions inside the channel, thus enabling the channel to be ion-selective [1]. These features are not observed in the microfluidic channels, as the counter-ionic space charge is confined to a very thin layer adjacent to the surface and the region away from the surface is essentially quasi-electroneutral (i.e., both co-ions and counterions are present away from the surface). Along with the EDL, the surface charge also plays a prominent role in controlling the transport inside the nanofluidic systems [2]. Similar ion-selective phenomenon was also observed in the intraparticle and intraskeleton mesopores of particulate and in membrane science [3].

Owing to the differences in the electrokinetic transport phenomena between the micro- and nanofluidic devices, the integration of these two devices paves way to complex physics. The models and the underlying theory developed to understand the electrokinetic transport in such systems are elaborated in section "Electrokinetic Theory for Micro/Nanochannels." A detailed discussion on the concentration polarization phenomenon and its applications are presented in section "Concentration Polarization." Finally, a brief summary is presented in section "Summary."

Electrokinetic Theory for Micro/Nanochannels

A complete set of equations for modeling the electrokinetic transport and to account for the EDL effects in micro/nanofluidic channels are presented. To understand the electrokinetic transport, space charge model developed by Gross et al. [4] is used extensively in the literature. The model solves the classical Poisson–Nernst–Planck (PNP) equations, which describe the electrochemical transport and the incompressible Navier–Stokes along with the continuity equations are solved to describe the movement of the fluid flow. These coupled systems of equations are more intensive, mathematically complicated, and computationally expensive. Though many linearized approximations were proposed to this model to study the electrokinetic transport [2], the governing equations of the complete nonlinear space charge model is discussed in this chapter.

In electrokinetic flows, the total flux is contributed by three terms: a diffusive component resulting from the concentration gradient, an electrophoretic component arising due to the potential gradient, and a convective component originating from the fluid flow. The total flux of each species in the solution is given by

$$\boldsymbol{\Gamma}_i = -D_i \nabla c_i - \Omega_i z_i F c_i \nabla \phi + c_i \boldsymbol{u} \tag{2}$$

where $\boldsymbol{\Gamma}_i$ is the flux vector, D_i is the diffusion coefficient, Ω_i is the ionic mobility, c_i is the concentration of the i^{th} species, $\boldsymbol{u}$ is the velocity vector of the fluid flow, and ϕ is the electrical potential. Note that the ionic mobility is related to the diffusion coefficient by Einstein's relation, $\Omega_i = \frac{D_i}{RT}$ [5]. The electrical potential distribution is calculated by solving the Poisson equation,

$$\nabla \cdot (\in_r \nabla \phi) = -\frac{\rho_e}{\in_0} \tag{3}$$

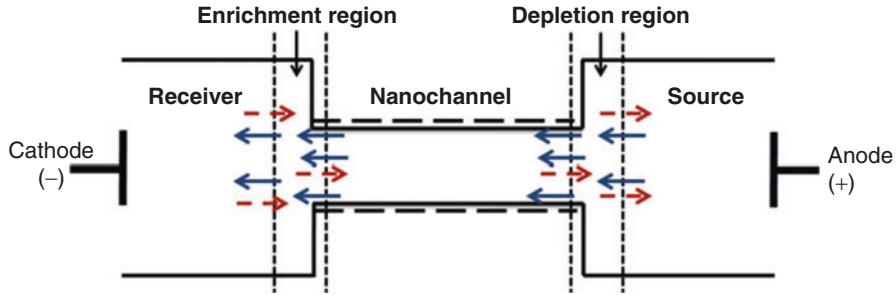
where ρ_e is the net space charge density of the ions defined as

$$\rho_e = F\left(\sum_{i=1}^{m} z_i c_i\right) \tag{4}$$

The mass transfer of each buffer species is given by the Nernst–Planck equation,

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot \Gamma_{\mathrm{i}} \tag{5}$$

Equations 3, 5, and 2 are the classical Poisson–Nernst–Planck (PNP) equations, which describe the electrochemical transport. The incompressible Navier–Stokes and the continuity equations are considered to describe the movement of the fluid flow through the channel, i.e.,



Concentration Polarization at Micro-/Nanofluidic Interfaces, Fig. 1 Schematic illustration of ion-enrichment and ion-depletion effect in cation-selective micro/nanofluidic channel. The *solid arrows* indicate the flux of cations and the *dotted arrows* indicate the flux of anions. At the nanochannel–anodic junction, both the anions and cations are depleted, while there is an enhancement of both the ions at the cathode–nanochannel junction

$$\rho\left(\frac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\nabla\boldsymbol{u}\right)=-\nabla p+\mu\nabla^2\boldsymbol{u}+\rho_e\boldsymbol{E} \quad (6)$$

$$\nabla\cdot\boldsymbol{u}=0 \quad (7)$$

where $\boldsymbol{u}$ is the velocity vector, p is the pressure, ρ and μ are the density and the viscosity of the fluid, respectively, and $\boldsymbol{E}=-\nabla\phi$ is the electric field. $\rho_e\boldsymbol{E}$ is the electrostatic body force acting on the fluid due to the space charge density and the applied electric field. Elaborate details on other simplified models are discussed in the review article of Schoch et al. [2].

Concentration Polarization

Concentration polarization (CP) is a complex phenomenon observed at the interface regions of micro/nanofluidic devices due to the formation of significant concentration gradients in the electrolyte solution near the interfaces causing accumulation of ions on the cathodic side and depletion of ions on the anodic side for a negatively charged nanochannel surface. This phenomenon was also observed in the field of colloid science and in membrane science which was extensively studied for over 40 years, and the early works of CP phenomenon is comprehensively reviewed by Rubinstein et al. [6]. The pioneering works from Rubinstein and his coworkers had revealed electrokinetic instabilities [7] in the concentration polarization regions leading to the breakdown of limiting current and resulting in the overlimiting conductance regimes in the ion exchange membranes. Such complex phenomenon could not be postulated using the classical equilibrium model of EDL [6]. All the underlying CP physics observed near the interfaces of micro/nanochannels are summarized in the following subsections.

Enrichment/Depletion Effects

In micro/nanofluidic devices, Pu et al. [8] first experimentally observed the CP effects near the interfaces and provided a simple model to explain the accumulation and depletion physics which is summarized below. For a negatively charged nanochannel, the EDL would be positively charged. For an overlapped EDL, as discussed in section "Electrical Double Layer (EDL) and Electroosmotic Flow," the nanochannel becomes ion-selective, resulting in higher cation concentration than anions. Thus, the flux of cations is higher compared to the anions in the nanochannel. With the application of positive potential at the source microchannel or reservoir (see Fig. 1), the cations move from the source (anode) reservoir to the receiving (cathode) reservoir end, while the anions move in the opposite direction through the nanochannel. At the cathodic side, the anion flux from the ends of reservoir to the nanochannel junction is higher compared to the anion flux from the junctions to the nanochannel as the anions are repelled by the negatively charged nanochannel. This difference in fluxes causes an accumulation

of anions at the cathode–nanochannel junction. The cation flux from the nanochannel to the cathode junction is greater than from the cathode junction to the reservoir as the cations have to balance the anions present at this junction. This results in an accumulation of cations as well at the nanochannel–cathode junction. At the anodic side, the anion flux from the nanochannel to the anode junction cannot balance the anion flux from the anode–nanochannel junction to the reservoirs due to the limited anions passing through the nanochannel. This results in a depletion of anions at this junction. The cation flux from the reservoir to the anode–nanochannel junction is less than the cation flux entering the nanochannel as the cations are attracted by the positively charged nanochannel. This in turn leads to the depletion of cations at this junction. To summarize, for a negatively charged nanochannel, both the cations and anions accumulate at the cathodic interface and are depleted at the anodic interface. The phenomenon is reversed for a positively charged nanochannel surface. A schematic diagram highlighting the accumulation and the depletion physics for a negatively charged nanochannel is displayed in Fig. 1.

Nonlinear Electroosmosis

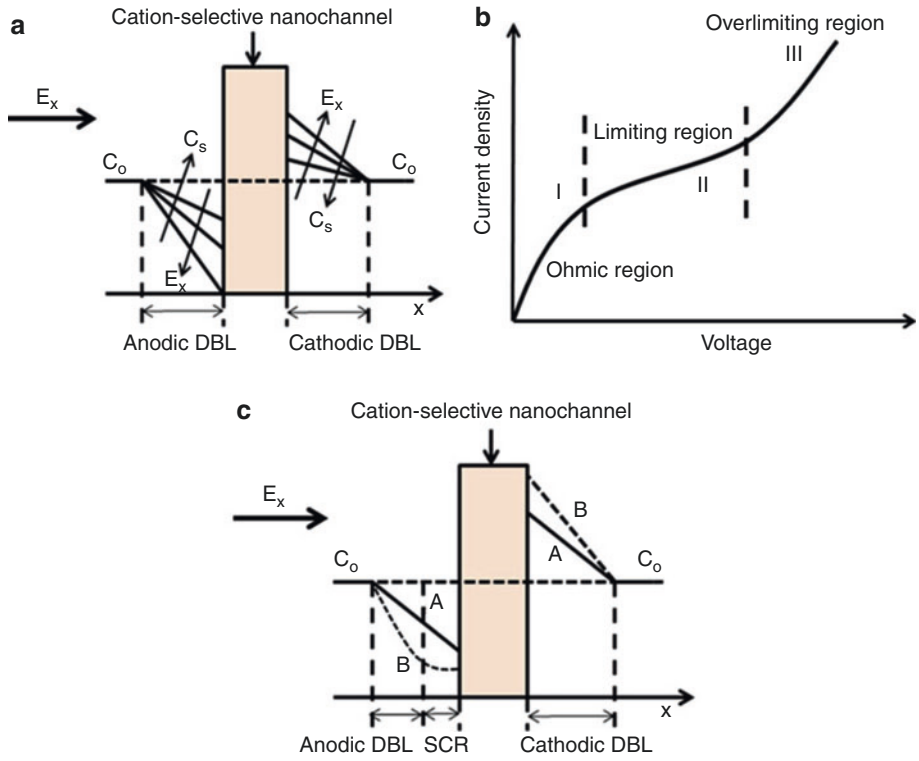
As discussed in the previous section, the integration of micro/nanofluidic devices leads to the accumulation and depletion of ions near the interfaces. Several numerical and experimental studies were carried out to gain a better understanding of the physics at these interfaces. The studies revealed complex and interesting physics at the depletion interface compared to the enrichment side. Rubinstein et al. [6] theoretically predicted the presence of space charges at the depletion region under large electric fields. The presence of the induced space charges near the depletion interface results in a nonequilibrium electrical double layer outside the nanochannel. The induced space charges under the action of the external electric field lead to nonlinear electroosmosis or otherwise known as electroosmosis of the second kind. The electroosmotic flow of the second kind was found to be directly proportional to the square of the applied electric field. Furthermore, the induced space charges also result in the generation of vortices at the depletion interface along with inducing large pressure and voltage gradients at this junction. Jin et al. [9] also reported similar physics from their extensive numerical study. In the case of a flat ion exchange membrane, Rubinstein et al. [7] derived a 2D nonequilibrium electroosmotic slip (u_s) for an applied voltage (V) using the linear stability analysis to impose strong vortex field near the membrane:

$$u_s = -\frac{1}{8}V^2 \frac{\overline{\frac{\partial^2 c}{\partial x \partial y}}}{\overline{\frac{\partial c}{\partial y}}} \tag{8}$$

where x and y are the axes parallel and perpendicular to the ion-exchange membrane, respectively. Experiments performed by Kim et al. [10] also reveal the nonequilibrium EOF near the micro/nanofluidic junctions. The application of electric field on the surface of particles also results in such induced space charges which spread over a larger region than the primary EDL resulting in highly chaotic flow patterns. A recent review by Höltzel and Tallarek [3] provide a detailed discussion on the polarization effects around membranes, packed beds, and glass monoliths. Recent advancements by Rubinstein and Zaltzman [11], however, revealed that the extended space charge region was not a part of the EDL, but develop from the counterion concentration minimum zone with the co-ions expelled under the action of the electric field. They further claim that the space charges would be present in the system even without equilibrium EDL. Their analysis included the study of the space charge dynamics in concentration polarization regions using 1D and three-layer models of EDL. From the understanding of these extended space charge layers, another important phenomenon, namely, the nonlinear current characteristics in micro/nanochannels is addressed below.

Nonlinear Current–Voltage Characteristics

There has been growing interest in the development of micro/nanofluidic devices as ionic filters



Concentration Polarization at Micro-/Nanofluidic Interfaces, Fig. 2 (**a**) Schematic distribution of ionic concentration in equilibrium concentration polarization under axial electric field (E_x). The local electroneutrality is maintained at both enrichment (cathodic interface) and depletion (anodic interface) diffusion boundary layers (DBL). The concentration gradients become steeper with the decrease in the ionic strengths (c_s) and at higher electric fields, (**b**) displays the nonlinear current–voltage characteristics for an ion-selective micro/nanochannel, and (**c**) shows the nonequilibrium concentration distribution due to the induced space charge region (SCR) (shown as *dotted lines* (*B*)) in the depleted region under very large electric fields

and nanofluidic batteries to control both ionic and molecular transport in aqueous solutions [2]. As discussed before, when the diameter/height of a charged channel scales comparable to the EDL, there is a predominant transport of the counterions inside the channel. Thus, the transport of electrical current inside the nanochannel is primarily due to the counterions. This feature enables the micro/nanochannel to be used as an ion exchange membrane. However, understanding the passage of ionic currents through such ion-selective solids is the most fundamental physical problem that has stimulated extensive research in this field for over a decade. Furthermore, the concentration polarization physics near the interfaces play a pivotal role in understanding the current–voltage characteristics. Figure 2a shows that the concentration gradients (CP regions) near the interfaces become steeper with the decrease in the ionic strength (c_s) and at higher electric fields ($\boldsymbol{E}_x$). At low electric fields, the current increases linearly with the applied voltage following the Ohm's law (region I in Fig. 2b).

However, at higher electric fields, the ion concentration in the depleted CP zone (i.e., near the anodic interface region) reaches toward zero and the classical Levich analysis [5] predicts a diffusion-limited current saturation according to which a saturation of current density occurs at a constant level described as the "limiting-current density" (region II in Fig. 2b).

The ionic current can be calculated considering the Fick's first law:

$$I = nFAD\frac{dc}{dx} \tag{9}$$

where n is the number of electrons transferred per molecule, D is the diffusion coefficient, and A is the electrode surface area. The concentration gradient is generally approximated by a linear variation (5),

$$I = nFAD\frac{c_0 - c(x=0)}{\delta} \tag{10}$$

where c_0 is the bulk electrolyte solution and δ is the diffusion boundary layer (DBL) thickness at the solid–liquid interface and $c(x=0)$ represents the concentration at the anodic (depletion) solid–liquid interface. δ typically ranges between 10 and 400 μm in ion exchange membranes [6], while it depends on the microchannel length in the case of micro/nanochannels [10].

From Eq. 10 it is clear that the current I reaches a maximum value when the concentration at $c(x=0)=0$, resulting in a limiting/saturation current as predicted by the classical Levich theory.

$$I_{lim} = \frac{nFADc_0}{\delta} \tag{11}$$

However, experimental studies in micro/nanofluidic devices (also in membrane science) revealed ionic currents larger than the limiting value and this regime was termed as the overlimiting current regime (region III in Fig. 2b). Further, the limiting current region is termed as the limiting resistance region (in micro/nanofluidic devices) due to the large but finite limiting differential resistance as the current does not saturate to a limiting value but has a slope which is smaller than the ohmic region. The nonlinear current characteristics in micro/nanochannels are a subject of intensive discussions in the literature [3]. Earlier studies indicated that water dissociation effects leading to the generation of H^+ and OH^- ions were responsible for such overlimiting currents. Later, Rubinstein et al. [7] postulated that some mechanism of mixing should be present that destroys the DBL as lower δ leads to higher currents (from Eq. 11). Maletzk et al. [12] coated the surface of cation-exchange membranes with a gel which does not allow mixing. From these experiments, they observed a saturation of current with no enhancements in the current, thereby confirming the earlier postulation of Rubinstein. Further, the fluid flow in the overlimiting regime revealed strong fluctuations indicating convection close to the surface. This convection was first attributed to the gravitational buoyant forces due to the concentration and temperature gradients. However, later theories have argued that the convection was not due to the gravitational instability in CP zones. Dukhin et al. [13] suggested the mechanism for mixing to be electroconvection. The type of electroconvection present in the overlimiting regime was revealed as the electroosmotic flow of the second kind. As discussed in the previous section, the induced space charges in the depletion zone (see Fig. 2c) under the action of the electric field results in the EOF of second kind and this convective instability tends to destroy the DBL leading to overlimiting currents as shown in Fig. 2b. Experiments by Kim et al. [10] also revealed the nonlinear currents due to the nonequilibrium EOF in the ion-selective micro/nanochannels. Similar physics was also observed in the permselective membranes and ion-selective particles [3]. In spite of all these postulations, the physics behind the extended space charge layer still remains largely unclear and there are a lot of potential research opportunities to fully understand this complex physics.

Propagation of Concentration Polarization

In this section, the conditions and the scenarios which can lead to the propagation of CP in micro/nanofluidic devices are highlighted. Zangle et al. [14] highlighted the phenomena of concentration polarization propagation using a simplified model of charged species transport and validated the same by conducting experiments and by comparing with other experimental results. The CP phenomenon was found to be governed by a

type of Dukhin number, relating the bulk and the surface conductance. The inverse Dukhin number, for a symmetric electrolyte was specified as:

$$\frac{G_{bulk}}{G_\sigma} = \frac{Fhzc_0}{\sigma} \tag{12}$$

where G_{bulk} is the bulk conductance, G_σ is the surface conductance, c_0 is the concentration outside the EDL, and σ is the wall surface charge density. Zangle et al. postulated that CP depends on Dukhin number and not on the ratio of channel height to the Debye length $\left(\frac{h}{\lambda D}\right)$. Using their simplified model, they showed that both the enhancement and the depletion regimes at the interfaces of micro/nanofluidic channels propagate as shock waves under the following condition:

$$c^*_{o,r} h^*_n < \max\left(\upsilon^*_2, 2\upsilon^*_2 - 1\right) \tag{13}$$

where $c^*_{o,r} h^*_n = \frac{(\upsilon_1 z_1 - \upsilon_2 z_2) F h_n c_{o,r}}{-2\upsilon_1 \sigma}$ is an inverse Dukhin number describing the ratio of bulk to surface conductance as mentioned before. $\upsilon^*_2 = \frac{\upsilon_2 z_2 F_\eta}{\zeta_n \in}$ is the mobility of the co-ion nondimensionalized by the electroosmotic mobility. $c_{o,r}$ is the reservoir electrolyte concentration, h_n is the nanochannel height, υ_1 and υ_2 are the mobilities, and z_1 and z_2 are the valences of the positive and negative ionic species, respectively. ζ_n is the nanochannel zeta potential, $\in$ is the permittivity and η is the viscosity. Elaborate details of the model can be referred in [14]. From this model, they proposed a thumb rule to avoid propagation of CP and it was found that $c^*_{o,r} h^*_n \gg 1$. This condition for CP propagation was compared with 56 sets of experimental literature values and was found to give a sufficient first-hand prediction with regard to the concentration polarization propagation.

Though the model considers the effects of surface charge and the electrolyte concentration, the finite *Pe* effects which also play a critical role in the concentration polarization were not considered. Further, the experiments of Kim et al. [10] and the numerical studies performed by Jin et al. [9] also revealed that the applied potential also plays a pivotal role in the concentration polarization generation and propagation apart from the inverse Dukhin number. The shortcoming of the model was also highlighted by Zangle et al. in their work. Thus, still a clear and complete understanding of the CP regimes is yet to be reached and continuous efforts are being made to understand the physics at the micro/nanofluidic junctions to design advanced and novel devices.

Applications

In this section, various applications that have been developed utilizing the concentration polarization phenomenon are addressed. The applications range from preconcentrating biomolecules to fluid pumping and mixing and also in water desalination. A brief discussion of the aforementioned applications is presented below.

Preconcentration

Wang et al. [15] used the depletion region observed at the micro/nanojunction to preconcentrate proteins. The energy barrier created at the depletion region (due to the large voltage drop induced at this junction) prevents the entry of charged molecules into the nanochannel. This results in an increase in the concentration of the molecules near the depletion region. In their experiments, Wang et al. used two anodic microchannels which were independently controlled so that the direction of EOF can be aligned perpendicular to the axis of the nanopore. An increase of about 10^6–10^8 fold in the concentration of the protein was reported in their study. Over the past couple of years, similar preconcentration devices utilizing the CP effects were experimentally fabricated [3, 14]. Wang et al. [16] also presented an experimental approach to improve the binding kinetics and the immunoassay detection sensitivity using concentration polarization in micro/nanofluidic devices. The antigens were preconcentrated at the depletion region due to the strong electric field gradients resulting in the enhancement in the binding rates with the antibody beads.

Seawater Desalination

The phenomenon of concentration polarization witnessed in ion-selective membranes was successfully implemented to address the freshwater shortage issue by providing energy-efficient solution to water desalination. A microfluidic device was fabricated which provides 99 % salt rejection at 50 % recovery rate at a power consumption of less than 3.5 Wh/L [17]. The CP depletion layer acts as a barrier for any charged species and these species were diverted away from the desalted water using suitable pressure and voltage fields. Their design also ensured salt ions and other debris to be driven away from the membrane thereby preventing any membrane fouling which is often observed in other desalination techniques.

Mixing, Pumping, and Other Applications

As discussed earlier, the induced space charges observed at the depletion region of micro/nanochannel along with the large electric field gradients at this region result in strong vortices. Kim et al. [18] enhanced the mixing efficiency of microfluidic devices using the vortices created at this interface. Further, as the electroosmotic flow of the second kind observed at the depletion region is directly proportional to the square of the applied electric field, Kim et al. [19] was able to pump fluids using the nonequilibrium EOF and observed a fivefold increase in the volumetric flow rates compared to similar devices utilizing equilibrium EOF. Yossifon et al. [20] used an asymmetric microchannel in conjunction with the nanochannels. The application of forward and reverse bias (at the overlimiting regime) voltage led to an asymmetric space charge polarization which resulted in the rectification of the current. Such membranes have potential applications in selective species separation.

Summary

The origin and the underlying physics that is present at the interfaces of micro/nanofluidic devices were discussed. The complex phenomenon of concentration polarization (CP) leading to the enrichment and depletion of ions near the micro–nano junctions and the concepts of induced space charges and nonlinear electrokinetic transport were briefly discussed in this chapter. Further, the various controversies surrounding the physical mechanism for nonlinear current characteristics have been highlighted. The criteria for concentration polarization propagation and the various applications that have been developed utilizing the concentration polarization phenomenon were discussed. Though, a lot of extensive work has been carried out to understand the CP physics, a clear and complete understanding of the CP regimes and the induced space charge dynamics is yet to be reached and efforts need to be directed in this area to understand the physics at the micro/nanofluidic junctions to design novel devices.

Cross-References

- ▶ Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering
- ▶ Electrokinetic Fluid Flow in Nanostructures
- ▶ Integration of Nanostructures within Microfluidic Devices
- ▶ Surface-Modified Microfluidics and Nanofluidics

References

1. Karniadakis, G.E., Beskok, A., Aluru, N.R.: Microflows and Nanoflows: Fundamentals and Simulation. Springer, New York (2005)
2. Schoch, R.B., Han, J., Renaud, P.: Transport phenomena in nanofluidics. Rev. Mod. Phys. **80**, 839–883 (2008)
3. Höltzel, A., Tallarek, U.: Ionic conductance of nanopores in microscale analysis systems: where microfluidics meets nanofluidics. J. Sep. Sci. **30**, 1398–1419 (2007)
4. Gross, R.J., Osterle, J.F.: Membrane transport characteristics of ultrafine capillaries. J. Chem. Phys. **49**, 228–234 (1968)
5. Probstein, R.F.: Physiochemical Hydrodynamics: An Introduction. Wiley, New York (1994)
6. Rubinstein, I.: Electrodiffusion of Ions. SIAM, Philadelphia (1990)

7. Rubinstein, I., Zaltzman, B.: Electro–osmotic slip of the second kind and instability in concentration polarization at electrodialysis membranes. Math. Models Methods Appl. Sci. **11**, 263–300 (2001)
8. Pu, Q., Yun, J., Temkin, H., Liu, S.: Ion–enrichment and ion–depletion effect of nanochannel structures. Nano Lett. **4**, 1099–1103 (2004)
9. Jin, X., Joseph, S., Gatimu, E.N., Bohn, P.W., Aluru, N.R.: Induced electrokinetic transport in micro – nanofluidic interconnect devices. Langmuir **23**, 13209–13222 (2007)
10. Kim, S.J., Wang, Y.-C., Lee, J.H., Jang, H., Han, J.: Concentration polarization and nonlinear electrokinetic flow near a nanofluidic channel. Phys. Rev. Lett. **99**, 044501 (2007)
11. Rubinstein, I., Zaltzman, B.: Dynamics of extended space charge in concentration polarization. Phys. Rev. E. **81**, 061502 (2010)
12. Maletzki, F., Rösler, H.-W., Staude, E.: Ion transfer across electrodialysis membranes in the overlimiting current range: stationary voltage current characteristics and current noise power spectra under different conditions of free convection. J. Membr. Sci. **71**, 105–116 (1992)
13. Dukhin, S.S.: Electrokinetic phenomena of the second kind and their applications. Adv. Colloid Interface Sci. **35**, 173–196 (1991)
14. Zangle, T.A., Mani, A., Santiago, J.G.: Theory and experiments of concentration polarization and ion focusing at microchannel and nanochannel interfaces. Chem. Soc. Rev. **39**, 1014–1035 (2010)
15. Wang, Y.-C., Stevens, A.L., Han, J.: Million–fold preconcentration of proteins and peptides by nanofluidic filter. Anal. Chem. **77**, 4293–4299 (2005)
16. Wang, Y.C., Han, J.: Pre-binding dynamic range and sensitivity enhancement for immuno-sensors using nanofluidic preconcentrator. Lab Chip **8**, 392–394 (2008)
17. Kim, S.J., Ko, S.H., Kang, K.H., Han, J.: Direct seawater desalination by ion concentration polarization. Nat. Nanotechnol. **5**, 297–301 (2010)
18. Kim, D., Raj, A., Zhu, L., Masel, R.I., Shannon, M.A.: Non-equilibrium electrokinetic micro/nano fluidic mixer. Lab Chip **8**, 625–628 (2008)
19. Kim, S.J., Li, L.D., Han, J.: Amplified electrokinetic response by concentration polarization near nanofluidic channel. Langmuir **25**, 7759–7765 (2009)
20. Yossifon, G., Chang, Y.-C., Chang, H.-C.: Rectification, gating voltage and interchannel communication of nanoslot arrays due to asymmetric entrance space charge polarization. Phys. Rev. Lett. **103**, 154502 (2009)

Conductance Injection

► Dynamic Clamp

Conduction Mechanisms in Organic Semiconductors

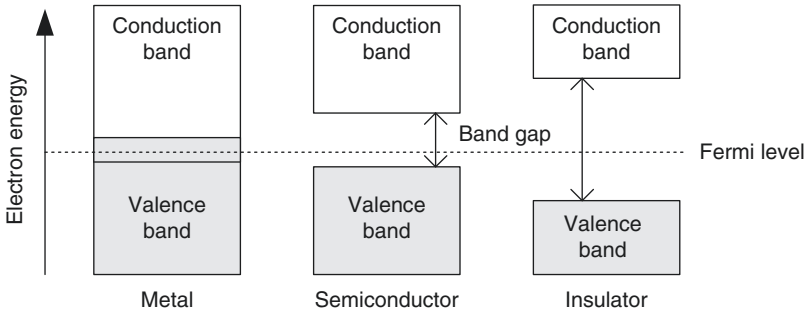
Weicong Li and Harry Kwok
Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Definition

Conduction mechanisms in organic semiconductors refer to the means by which electronic charges move through organic semiconductors under external stress particularly under the influence of an electrical field.

Overview

In order to understand the conduction mechanisms in organic semiconductors, it is necessary to first introduce the concept of *band theory*, which is well established in solid-state physics. Solids in general are made of atoms, each of which is composed of a positively charged nucleus surrounded by negatively charged electrons. In quantum mechanical terms, these electrons effectively reside in discrete energy states in orbits. When a large number of atoms (of order 10^{20} or more) are brought together to form a solid, the discrete energy states are so close together that energy bands begin to form. At the same time, there will be gaps between the energy bands which are known as the *band gaps*. Because of the presence of these energy gaps, there will be some energy bands that are almost fully occupied (known as the *valence bands*) and energy bands that are almost unoccupied (known as the *conduction bands*). Based on the band theory, solids are typically divided into the following three categories: *metals*, *semiconductors*, and *insulators*. In metals, there is an overlap between the energy bands so that the energy bands are partly filled by electrons at any temperature (even $T = 0$ K), while both the semiconductors and the insulators have fully filled



Conduction Mechanisms in Organic Semiconductors, Fig. 1 Band structures of metal, semiconductor, and insulator

valence bands and empty conduction bands at $T = 0$ K. To further study the distribution of electrons occupying the energy states in solids, it is also necessary to introduce the concept of Fermi level into the band theory, which represents the maximum energy of states that electrons can occupy at $T = 0$ K. Accordingly, all the allowed energy states below the Fermi level are occupied by electrons, and all the energy states above it are empty. When temperature is above 0 K, the probability that electrons occupy the state with energy E under thermodynamic equilibrium condition is given by Fermi-Dirac distribution function:

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \qquad (1)$$

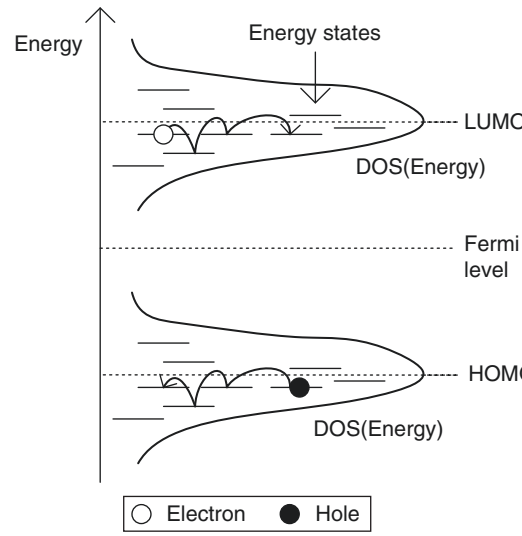
where E_F is the Fermi level, k is the Boltzmann constant, and T is the temperature in Kelvin. As mentioned earlier, in semiconductor and insulator, the valence band is fully occupied by electrons, and the conduction band is empty at $T = 0$ K. Therefore, one can infer that the Fermi level lies in the bandgap, between the valance and conduction bands. On the other hand, in metal, due to the fact that the energy bands are partly filled by electrons at any temperature, the Fermi level lies within the energy bands. The band structures of metal, semiconductor, and insulator, and the position of Fermi level in them are shown in Fig. 1. The distinction between the semiconductors and the insulators appears when temperature rises above 0 K. Because the band gap between conduction and valence bands in semiconductors is much narrower than that found in insulators, a fair amount of electrons can be thermally excited from the valence band to the conduction band in semiconductors at finite temperature, leading to measurable conductivity. This is not found in the insulators due to the larger band gaps even at room temperature, which lead to negligible probability of electrons occupying energy states in the conduction band, according to Eq. 1. The characteristic semiconductor band structure has allowed it to play an important role as the materials of choice in the prosperous electronic industry in the last few decades.

While inorganic semiconductors such as silicon dominated the electronic industry in the twentieth century, tremendous effort has been spent in the research and development of organic electronics in last decade due to the fact that organic semiconductors are usually easier and cheaper to form. Soluble organic materials, such as conjugated polymers, can be deposited in liquid phase (e.g., by printing and spin coating) onto large substrate areas at low processing temperature (below 100 °C). Due to this advantage, organic electronics are particularly attractive in the making of displays, sensors, light sources, photovoltaic panels, radio-frequency identification detectors (RFID), and in devices used in optical communications. As a consequence, research on the charge conduction mechanisms in organic semiconductors and devices is of significant importance.

In general, conduction mechanisms primarily describe how electronic charges (referred to as carriers) move inside the solids under the influence of an external electrical field. The process produces a current. At the macroscopic level, the current density J in solids produced by external electrical field is given by

$$J = env = en\mu F \tag{2}$$

where e is elementary charge of a single carrier, n the charge density, and v is the drift velocity. Furthermore, v can be expressed as the product of the charge mobility μ and the electrical field F. As can be seen in Eq. 2, a large current requires the presence of a substantial number of mobile charge carriers (electrons or holes). In organic semiconductors, mobile carriers are known to be produced from the distributed π-bonds, which are covalent chemical bonds resulting from the overlap of atomic orbitals. Thus, the limited current flow in many organic semiconductors are related to their irregular molecular structures which can result in low charge mobility in comparison to values found in silicon and other inorganic semiconductors. In addition, the more established conduction mechanisms based on band theory normally found in inorganic crystalline semiconductors are absent in the organic semiconductors. As mentioned earlier, band theory states that carriers could only exist and move in either the conduction bands or the valence bands because there are permitted energy states where carriers can reside and their movement between the energy states will produce a current. The use of "energy band" diagrams (see Fig. 1) to explain charge transport in organic semiconductors however is not possible. This is because of the presence of high densities of defects and trap states. Instead, charge transport in organic semiconductors is directly explained in terms of the energy (orbital) states which are termed either the *lowest unoccupied molecular orbital* (LUMO) or the *highest occupied molecular orbital* (HOMO). As such, LUMO and HOMO levels are not genuine energy bands and they are used merely to serve as references to demarcate ground state energy and the next activated state energy [1]. The distribution of these energy states known as the *density of states* (*DOS*) is usually considered to be Gaussian centered at the LUMO and HOMO (see Fig. 2). The width of the Gaussian energy states depends on both the regularity of the molecular structure and the impurities present in the organic semiconductor.

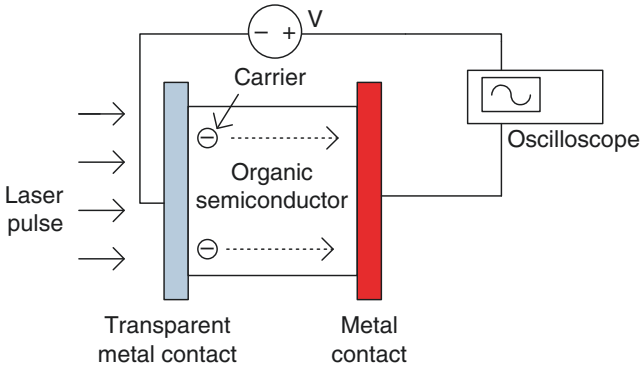


Conduction Mechanisms in Organic Semiconductors, Fig. 2 Density of (energy) states in an organic semiconductor

As expected, in most organic semiconductors both the carrier density n and the charge mobility μ are low and the value of the latter often depends on the strength of the electrical field F in contrast to what is observed in inorganic semiconductors. In some organic semiconductors, the molecular structures can be highly disordered and different conduction mechanisms are found to predominate depending on the associated manufacturing process.

Basic Methodology

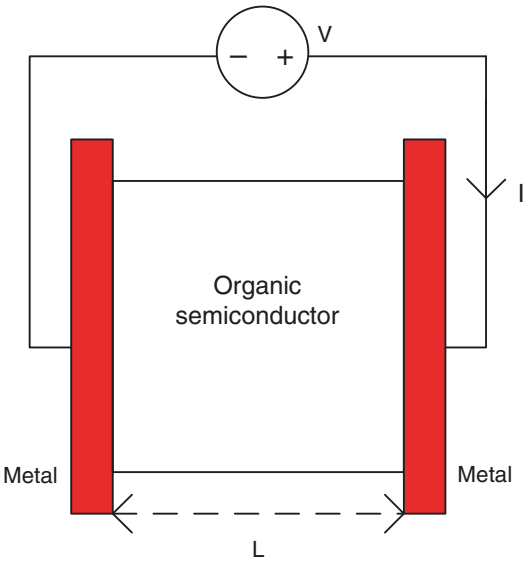
Many useful techniques have been proposed to study the conduction mechanisms in organic semiconductors including time-of-flight (TOF) experiment, space charge limited current (SCLC) measurement, and field-effect measurements using organic field-effect transistors (OFETs). These techniques, combined with the dependence on temperature, provide important information on mobility, trap concentration which are useful to assess the conduction mechanisms. Brief introductions to several different techniques commonly used are given here.



Conduction Mechanisms in Organic Semiconductors, Fig. 3 A schematic of the time-of-flight (TOF) experiment setup

Time-of-Flight (TOF) Experiment

As implied by its name, time-of-flight (TOF) experiment is the method of measuring the time it takes for one or a few carrier to travel a distance through the solid. When TOF experiment is used to study the conduction mechanisms in organic semiconductors (see Fig. 3), two metal electrodes (forming the contact) are deposited on the two ends of the organic semiconductor (one of the two is usually transparent). Initially, a few carriers are generated at one end near the transparent metal electrode using a short laser pulse with energy greater than the energy difference between HOMO and LUMO levels of organic semiconductor. The photo-generated carriers are then drifted toward the opposite end by an external electrical field generating a current pulse. By measuring the time delay of the current pulse, the velocity and the mobility of the charge carriers through the organic semiconductors can be computed.



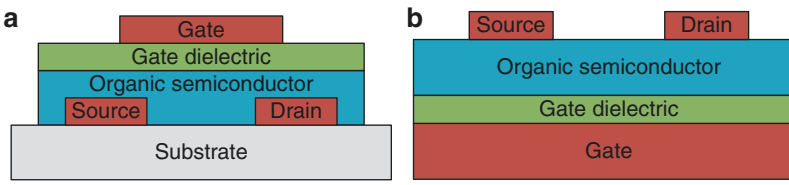
Conduction Mechanisms in Organic Semiconductors, Fig. 4 A simplified schematic of the setup used for space charge limited current (SCLC) measurement

Space Charge Limited Current (SCLC) Measurement

Due to the low mobility of carriers in organic semiconductors, the measured I–V characteristics usually deviate from Ohm's law (i.e., the linear relationship between current and voltage). This is illustrated in Fig. 4. In this case, if efficient charge injection from the metal electrode is achieved by choosing a suitable metal, the I–V characteristics will follow the space charge limited current equation as given by

$$I = \frac{9}{8}\theta\varepsilon_0\varepsilon_r\mu A\frac{V^2}{L^3} \qquad (3)$$

where θ is a parameter dependent on the traps present in the semiconductors, ε_0 the free space permittivity, ε_r the relative dielectric constant, μ drift mobility of injected charge carrier, A the cross section area of semiconductor, and L the distance between metal contacts. By measuring the I–V characteristics at different temperatures, one can determine the mobility and the trap density of traps. Thus, SCLC measurement is a very useful technique reflecting the conduction mechanisms in organic semiconductors.



Conduction Mechanisms in Organic Semiconductors, Fig. 5 Two device configurations for the OFETs: (**a**) top-gate, (**b**) bottom-gate

Measurement Based on the Organic Field-Effect Transistor (OFET)

Field-effect transistor (FET) is an electronic device widely used in active circuits. It consists of a semiconductor with a conducting channel, an isolated gate controlling charge flow in the channel, a gate dielectric between the semiconductor and the gate, as well as source and drain regions forming the output terminals. Organic field-effect transistor (OFET) is a field-effect transistor formed on an organic semiconductor. The device configuration can have a top-gate or a bottom-gate as shown in Fig. 5.

The basic operation principle of the OFETs is very simple. When a bias voltage is applied between gate and source electrodes, carriers are injected from the source into the organic semiconductor forming an extremely thin accumulation layer ($2 \sim 3$ nm) at the interface between organic semiconductor and the dielectric. The carriers conduct a current across the source and the drain regions, and the current depends on the gate voltage as well as the charge mobility which is also dependent on the gate voltage and the drain-to-source voltage. The operation of the OFETs therefore relies on carrier accumulation in the field-effect structure in contrast to the case of the inorganic FETs which rely on either charge depletion or inversion. Therefore, OFETs are efficient tools to investigate the interfacial conduction mechanisms, while TOF and SCLC measurements are mainly used to study the bulk conduction mechanisms in the organic semiconductors. For example, *I–V* characteristics of OFETs are usually analyzed to determine parameters such as the charge mobility and the threshold voltage both of which are closed related to the density of traps at the interface. In addition, spectroscopic techniques are sometimes used to probe the morphology of the organic semiconductor interface, to look for the potential relationship between regularity of molecular structure and conduction performance.

Key Research Findings

Band-Like Transport

For highly purified and ordered organic molecular crystals, it is possible that band-like charge transport similar to that of the inorganic semiconductors may occur. The main feature found in band-like charge transport is the fact that the temperature dependence of the charge mobility has the following form:

$$\mu(T) \propto T^{-n}, \text{ with } n = 1, 2, 3 \ldots \qquad (4)$$

In practice, n is usually positive which leads to increasing charge mobility when temperature decreases. In general, because the electrons are usually weakly delocalized even in the highly ordered organic crystals, the band widths of the HOMO and the LUMO are small compared to energy bands found in the inorganic semiconductors. As a result, room temperature charge mobilities observed in organic semiconductor crystals can only reach values in the range 1–20 cm^2/Vs [2]. Band-like charge transport has been observed in small-molecules and in single-crystal organic semiconductors (such as rubrene) formed by vapor deposition process. As a matter of fact, in the majority of organic semiconductors, traps and defects are formed during deposition which tends to destroy band-like properties.

Polaron Transport

A polaron is a quasiparticle composed of a charge carrier and its induced polarization field. In many organic materials, due to the low charge

mobilities, carriers tend to polarize their surrounding lattice. As a result, polarization fields are formed around the carriers, which can no longer be considered as "naked." Instead, the carriers will be localized in potential minima created by the so-called molecular deformations [1]. In other words, a charge is trapped by the deformation it induces. Such an entity is known as a "polaron." Polarons can move between molecules similar to the carriers except that they also carry the deformations along. In many disordered molecular organic semiconductors, deformations associated with the trapped charges can be considerable and conduction mechanisms characterizing polaron transport are under intensive research by many research groups across the world.

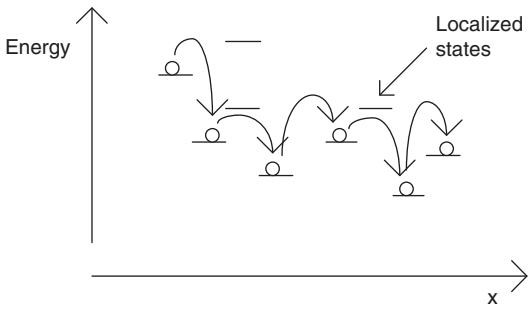
Variable Range Hopping (VRH) Transport

For most organic semiconductors such as polymers and oligomers, their molecular structures are highly disordered and have considerable densities of defects and traps. The energy band diagrams are no longer suitable to describe the densities of states as these energy states are now localized. Furthermore, band-like charge transport can no longer explain the observed low charge mobilities and the fact that their values increase with temperature (as opposed to what is observed in band-like charge transport). One of the widely accepted theories, known as the variable range hopping (VRH) transport, is proved to give a reasonable explanation by describing charge transport in terms of hopping of the charge carriers between localized states as shown in Fig. 6.

Hopping can be used to explain the lower mobility found in disordered organic semiconductors and instead of the power law dependence on temperature as in band-like charge transport, the temperature dependence in VRH charge transport exhibits temperature-dependent activation as well as dependence on the applied electric field as given by [3]

$$\mu(F,T) \propto \exp(-\Delta E/kT) * \exp\left(\beta\sqrt{F}/kT\right) \quad (5)$$

where μ is mobility, F is the electrical field, T is the temperature, ΔE is the activation energy, and β is a parameter related to disorder.



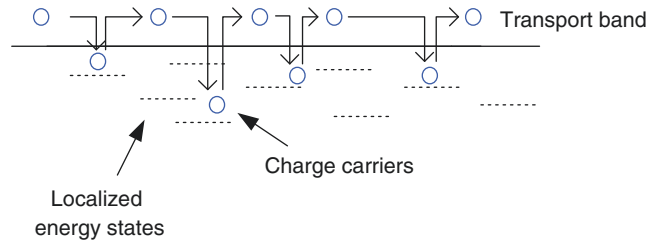
Conduction Mechanisms in Organic Semiconductors, Fig. 6 Hopping transport in organic semiconductors

Multiple Trap and Release (MTR) Transport

Charge transport in OFETs is affected by defects and impurities which exist in the intrinsic part of the organic semiconductors per se and can also be linked to an inferior semiconductor/dielectric interface. As a result, the performance of OFETs is sample-dependent, which is one of the major difficulties in characterizing the properties of OFETs. As mentioned earlier, VRH transport is more suitable to account for charge transport in highly disordered organic semiconductors and, in contrast, another well-established and widely accepted charge transport model known as the "multiple trap and release" (MTR) model is frequently applied to the relatively well-ordered organic semiconductors, such as small molecules and molecular crystals. The basic principle of MTR model includes two important components: (1) a transport band containing delocalized energy states whereby carriers can move freely and (2) the presence of a high density of localized energy states located in the vicinity of the edge of transport band acting as traps. During charge transport, carriers move freely in the transport band with a high probability of being trapped at the localized energy states and then subsequently thermally released into transport band again. The basic illustration of MTR transport process is shown in Fig. 7.

The effective mobility (μ_{eff}) in the MTR model is actually smaller than the "real" mobility (μ_0) in the transport band in the absence of localized energy states and is given by [4]

$$\mu_{\text{eff}} = \mu_0 \alpha \exp(-E_t/kT) \quad (6)$$



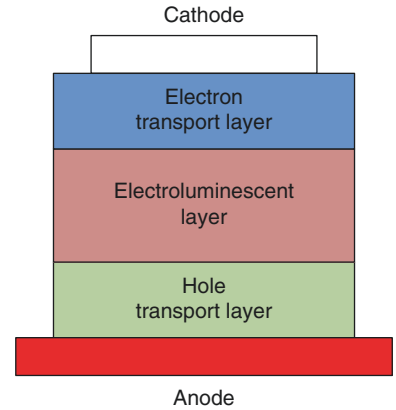
Conduction Mechanisms in Organic Semiconductors, Fig. 7 Multiple trap and release transport (*MTR*) model

where α is the ratio of the effective density of energy states at the edge of the transport band to the density of traps in the localized energy states, and E_t is the energetic distance between the edge of the transport band and the localized energy states. If the localized energy states are energetically dispersive, α and E_t must be recalculated according to the trap distribution.

In the study of OFETs, the MTR model is widely used to account for charge transport due to the fact that it offers a reasonable explanation on the gate voltage dependent mobility usually observed in OFETs. As mentioned above, unlike inorganic semiconductors, organic semiconductors usually have Gaussian density of states (DOS). When a bias is applied to the gate of an OFET, the Fermi level at the dielectric-semiconductor interface will be shifted toward the transport band so that a fair amount of localized energy states near the edge of the transport band will be filled when the Fermi level is located closer to the transport band. As a result, the mobility of the carriers in the MTR model is actually improved because of the reduced density of traps leading to a reduced value of E_t. Therefore, the gate-voltage dependent effective mobility of the carriers in the MTR transport model is given by [5]

$$\mu_{\text{eff}} = \mu_0 \frac{N_c}{N_{t0}} \left[\frac{C_i (V_G - V_T)}{q N_{t0}} \right]^{\frac{T_0}{T} - 1} \tag{7}$$

where μ_0 is the mobility of the carriers in the transport band, N_c the effective density of states at the edge of the transport band, N_{t0} the total density of traps, C_i the capacitance of the insulator per unit area, and T_0 is a characteristic temperature related to the distribution of the DOS.

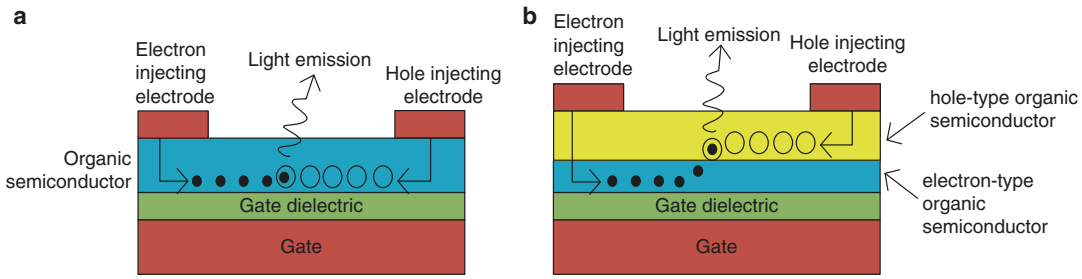


Conduction Mechanisms in Organic Semiconductors, Fig. 8 Schematic of a typical organic light-emitting diode (OLED)

Devices

Organic Light-Emitting Diode (OLED)

Organic light-emitting diode (OLED) is an electroluminescent diode composed of organic materials serving as the electroluminescent layer and charge transport layer. The typical OLED structure is shown in Fig. 8. During operation of an OLED, a bias voltage is applied between anode and cathode. Holes (electrons) are injected from the anode (cathode) into the electroluminescent layer through the hole (electron) transport layer. Because electrons and holes exist simultaneously in the same layer, there is a high probability that they recombine with each other due to electrostatic forces, leading to radiative emission. Therefore, efficient charge injection from both cathode and anode is requisite for the efficient operation of the OLED and current conduction is dominated by space charge limited current as introduced earlier.



Conduction Mechanisms in Organic Semiconductors, Fig. 9 Schematic illustrations of organic light-emitting field-effect transistors (OLEFETs) with light emission in: (**a**) the single-layer configuration and (**b**) the multilayer configuration

Organic Light-Emitting Field-Effect Transistor (OLEFET)

Organic light-emitting field-effect transistor (OLEFET) is a novel organic device combining the function of current conduction of an OFET with electroluminescence in an OLED. The operation of the OLEFET is actually the same as that of the OFET. However, if proper materials are chosen as the source and the drain to give efficient charge injection and under favorable voltage bias condition, electrons and holes can be injected and transported separately in the OFET channel(s). This type of charge transport is known as ambipolar charge transport, which is unique and only found in an organic field-effect transistor. Furthermore, in ambipolar charge transport if the electrons and holes are allowed to recombine radiatively in an emitter layer to give out light, this type of OFET with electroluminescence functionality is usually called OLEFET. Various device structures have been proposed to realize ambipolar charge transport and light emission in OLEFETs, and, in most cases, the proposed structures fall into two main categories as far as the charge layers are concerned. These are the single-layer OLEFET and multilayer OLEFET as shown in Fig. 9.

Cross-References

- ▶ Electrode–Organic Interface Physics
- ▶ Flexible Electronics
- ▶ Optical and Electronic Properties
- ▶ Surface Electronic Structure

References

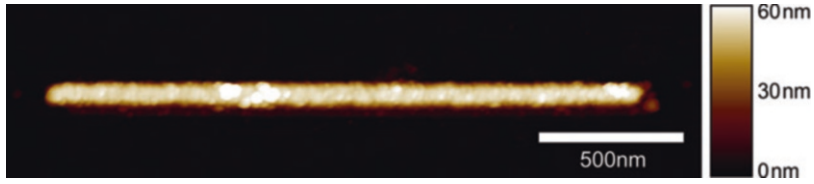
1. Kwok, H.L., Wu, Y.L., Sun, T.P.: Charge transport and optical effects in disordered organic semiconductors. In: Noginov, M.A., Dewar, G., McCall, M.W., Zheludev, N.I. (eds.) Tutorials in Complex Photonic Media, pp. 576–577. SPIE Press, Bellingham (2009)
2. Podzorov, V., Menard, E., Borissov, A., Kiryukhin, V., Rogers, J.A., Gershenson, M.E.: Intrinsic charge transport on the surface of organic semiconductors. Phys. Rev. Lett. **93**, 086602 (2004)
3. Brütting, W.: Physics of Organic Semiconductors. Wiley, Weinheim (2005)
4. Horowitz, G.: Organic field-effect transistors. Adv. Mater. **10**, 365–377 (1998)
5. Bao, Z., Locklin, J.: Organic Field-Effect Transistors. CRC Press, Boca Raton (2007)

Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy

J. Vogt[1], C. Huck[1], F. Neubrech[2] and A. Pucci[1]
[1]Kirchhoff Institute for Physics, Heidelberg University, Heidelberg, Germany
[2]4th Physics Institute, University of Stuttgart, Stuttgart, Germany

Synonyms

Electrical conductivity; Localized plasmon polaritons in the infrared; Nanoantennas



Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Fig. 1 AFM picture of a typical gold nanowire prepared by EBL on CaF_2. The measured average height of the wire is ca. 60 nm

Definition

Metal nanorods of a few micron length and a much smaller diameter feature strong plasmonic resonances in the infrared region and therefore act rather similar to radio antennas, but the spectral shape of such resonances is related to the conductivity of the nanoantenna material. In the scientific literature, analytic approaches exist which explain the relationship between the electronic conductivity and the resonance spectrum that is due to plasmon polaritons as mixed excitations from free electrons and photons.

Introduction

Metal nanowires are produced, for example, by electron-beam lithography (EBL), electrochemical, and wet chemical methods [1]. Usually they are inspected by scanning electron microscopy (SEM) or atomic force microscopy (AFM) in order to get geometric information; see Fig. 1. But, neither SEM nor AFM can deliver conductivity information of a nanoobject. Electrical measurements are not simple because of the necessary contacts and thus contact resistance comes into play. Nevertheless, because of the importance of the ohmic losses in electronics, many studies already investigated the change of resistivity with the size and the crystalline quality of interconnect structures [2]. The results demonstrate that the two main contributions to the resistivity increase with shrinking height and width of metal interconnects which are surface scattering and grain-boundary scattering. The Matthiessen rule according to which the several contributions add up to the total resistivity [3] was proven to be a valid approximation in most cases.

For plasmonic resonances of nanostructures, increased resistivity leads to higher damping and thus to less-efficient performance of antennas, sensors, and other devices. Vice versa, information on resistivity or conductivity, respectively, is included in the plasmonic resonance spectrum.

Infrared Conductivity of Metals

Below the onset of interband transitions, the infrared (IR) optical properties of metals are determined by the collective oscillations of free-charge carriers called plasmons. The circular frequency (ω)-dependent Drude dielectric function

$$\varepsilon(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega(\omega + i\omega_\tau)} \qquad (1)$$

is a good description of that behavior [4]. Effects from interband transitions on the background polarizability are included in ε_∞. This background permittivity can have values clearly above 1, which becomes important in the near infrared where the negative real part of the Drude term reaches the same order of magnitude. For gold and silver, ε_∞ becomes frequency dependent and complex in the visible, but for other metals like iron, chromium, and platinum, such behavior starts already in the mid-infrared, and the application of the models in this article is then restricted to the far IR below about 1000 cm^{-1} (in wave numbers). The Drude parameters ω_p and ω_τ describe the plasma frequency and the relaxation rate of the free charge carriers, respectively. Frequency-independent parameters (see Table 1) are in reasonable accord with the IR optical

Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Table 1 Drude parameters for selected metals at room temperature from the data collection by Ordal et al. [5]. The parameters are given as wave numbers $\varpi = \omega/(2\pi c)$ in cm^{-1} units. They were derived from fits to far and mid-IR spectra of polycrystalline bulk material (ε_∞ was set equal to one)

	ω_τ	ω_p
Gold	215	72,800
Silver	145	72,700
Aluminum	660	119,000

behavior of such metals where only s and p electrons contribute to the conductivity, for example, aluminum, silver, and gold (but not iron). With ω_p much higher than IR frequencies and $\omega_\tau << \omega_p$, the Drude dielectric function has a strong negative real part in the IR. For ω_τ spectrally located in the mid-IR (typical situation occurring for iron or defect-rich noble metals), the skin depth in absorbing region is

$$\delta \approx \frac{c}{\omega_p}\sqrt{\frac{2\omega_\tau}{\omega}}, \qquad (2)$$

but it is $\delta \approx c/\omega_p$ for $\omega_\tau << \omega$ in the reflecting region (c is the vacuum velocity of light). Typical values for skin depths of metals in the mid-IR are of the order of a few 10 nm and thus of the same order of magnitude as the diameters of typical plasmonic structures used in sensing applications. Any increase in resistivity may lead to a strong increase in the skin depth and thus significantly increases the ratio between absorption and reflection [4].

In the dc limit, the Drude parameters determine the dc resistivity $\rho_{dc} = \omega_\tau / \left(\varepsilon_v \omega_p^2\right)$ [3]; ε_v is the vacuum permittivity. Slight deviations from the IR optical parameters may occur due to an anisotropy of scattering [4] and the energy dependence of many body effects [6].

Electron Scattering

From a variety of studies, it is known that, as theoretically predicted, the contribution $\omega_{\tau s}$ of electronic surface scattering to resistivity increases directly proportional to the inverse wire diameter (with a proportionality factor depending on surface roughness and surface chemistry), e.g., [2] and references therein, and, according to Matthiessen's rule, [3] has to be added to the defect scattering rate $\omega_{\tau d}$. Therefore, the total electronic relaxation rate (inverse lifetime) is

$$\omega_\tau = \omega_{\tau e}(t,\omega) + \omega_{\tau p}(t) + \omega_{\tau s} + \omega_{\tau d} \qquad (3)$$

with $\omega_{\tau e}$ and $\omega_{\tau p}$ as the electron–electron and the electron–phonon scattering rate, respectively [2–4].

Above temperatures t of about 10 K, the temperature-dependent term in electron–electron scattering is small compared to electron–phonon scattering [3, 4]. The frequency-dependent term in electron–electron scattering is also negligible for metals with conduction carried by s and p electrons only [5, 6], like, for example, Au in the infrared where the dielectric function follows the simple Drude model with a constant ω_τ. The temperature dependence of ω_τ is almost due to the temperature-dependent number of phonons. With the Debye temperature θ_D, the Debye model for phonons leads to the Bloch–Grüneisen relation [7] for the scattering rate

$$\omega_{\tau p} \propto \left(\frac{t}{\theta_D}\right)^5 \int_0^{\theta_D/t} \frac{z^5}{(e^z - 1)(1 - e^{-z})} dz \qquad (4)$$

for a free electron gas in the dc limit. This relation predicts $\omega_\tau \propto t/\theta_D$ for $t >> \theta_D$, $\omega_\tau \approx \omega_{\tau d} + \omega_{\tau s}$ for $t << \theta_D$, and a transition region with $\omega_\tau \propto (t/\theta_D)^5$ (Bloch relation [3]), respectively. In the infrared, this relation is no more valid since additionally to the electron and phonon, also a photon with energy $\hbar\omega$ is involved, and the scattering rate

$$\omega_{\tau p} = \frac{1}{\tau_0}\left[\frac{2}{5} + 4\left(\frac{t}{\theta_D}\right)^5 \int_0^{\theta_D/t} \frac{z^4}{e^z - 1} dz\right] \qquad (5)$$

derived by Holstein is considered as a better approximation [8]. The rate $1/\tau_0$ depends on the kind of metal and could be approximated by

Holstein's bulk dc-scattering rate at sufficiently low temperature [4].

Granular Metal

Beyond the percolation threshold, the IR optical properties of conductive but granular metal films can be described by a Drude-type dielectric function ε^* with an effective plasma frequency $\omega_{p,eff}$. The analytic approach to the effective dielectric function ε_{eff} of inhomogeneous media that originally was developed by Bruggeman [9],

$$F\frac{\varepsilon-\varepsilon_{\text{eff}}}{\varepsilon+(D-1)\varepsilon_{\text{eff}}}+(1-F)\frac{\varepsilon_{\text{host}}-\varepsilon_{\text{eff}}}{\varepsilon_{\text{host}}+(D-1)\varepsilon_{\text{eff}}}=0, \tag{6}$$

with the spatial filling factor F, dimension D, and a host medium with ε_{host} can be used to estimate the effective Drude parameters. For a metal-island film $D = 2$ and in vacuum, without adsorbates or cover layers, $\varepsilon_{host} = 1$. With $1/\varepsilon \approx 0$ which is a valid approximation below a certain frequency, it follows $\varepsilon_{eff} = (2F - 1)\varepsilon$. Furthermore, with $\varepsilon_{eff}/F = \varepsilon^*$, a relation that accounts for equal absorption in both the models, the result is $\omega^2_{p,eff} = (2 - 1/F)\omega^2_p$. For $D = 3$ the same approximation leads to $\omega^2_{p,eff} = \omega^2_p(3F - 1)/(2F)$. For $D = 2$, the lowering of the squared-effective plasma frequency is more pronounced than for the $D = 3$ case. So, far-field spectra are especially sensitive to a morphology consisting of one layer of grains, which is mostly relevant for metal nanostructures [10]. The grainy structure of the wire in Fig. 1 could be clearly recognized. Upon annealing the morphology can be changed towards larger grains.

Infrared Plasmonic Resonances

Very similar to radio frequency (RF) antennas, resonantly excited plasmonic nanostructures act as optical antennas that concentrate energy of electromagnetic radiation to a confined volume of sub-wavelength scale. For example, metallic nanorods ("nanoantennas") with μm-sized lengths L show plasmonic resonances in the IR spectral range. Related to the not-negligible skin depth δ, the simple $\lambda/2$-dipole behavior known from RF antennas, where the relationship between L and the resonant wavelength λ_{res} is given by $2\ L = \lambda_{res}$, is not valid for nanoantennas at optical (including IR) frequencies [1]. The modified relation for cylindrical rods

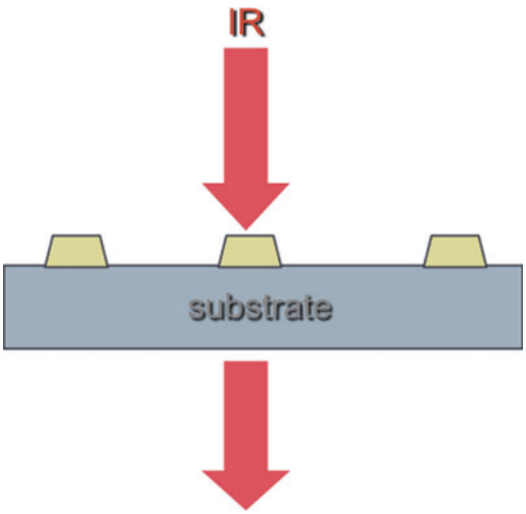
$$2L = \lambda_{\text{eff}} = n_2\left(\lambda_{\text{res}}/\lambda_{\text{p}}\right) + n_1 \tag{7}$$

(for the radius $R << L$) where λ_p denotes the plasma wavelength of the antenna's material better describes the optical antenna behavior in the case $\omega_\tau << \omega_{res}$ and makes clear that $\lambda_{res} \sim 1/\omega_p$ [11]. Thus granularity effects the resonance position; see the lowered effective plasma frequency above. The coefficients n_1 and n_2 depend on R and on the dielectric properties ε_∞ and ε_s, the dielectric constant of the nonabsorbing surrounding medium at the resonance frequency:

$$n_1 \approx R\left\{\left[2\pi\left(13.74 - 0.12\cdot\frac{\varepsilon_\infty + 141.04\varepsilon_s}{\varepsilon_s}\right)\right] - 4\right\},$$
$$n_2 \approx \frac{2\pi R}{\varepsilon_s}\cdot 0.12\cdot\sqrt{\varepsilon_\infty + 141.04\varepsilon_s}. \tag{8}$$

The metal was described as a free-electron gas according to the Drude model with negligible relaxation rate ω_τ compared to photon frequencies ω. Such conditions are fulfilled for perfect, crystalline gold (clearly also silver and copper) nanorods in the mid-IR, but not in case of strong defect scattering when δ becomes frequency dependent.

Plasmonic extinction in the IR (e.g., as measured in normal transmittance geometry, see Fig. 2, with polarization along the long-nanowire axis and the nanowire on a substrate that without nanowire is used for the reference spectrum [1]) contains an important contribution from absorption. This contribution becomes smaller (compared to scattering) for larger R (or larger height h and width w) of the antenna. As a rough approximation, these effects can be described within the radiation-corrected



Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Fig. 2 Scheme of the IR transmittance measurement at normal incidence of light. For studies of individual wires, the focal plane should be as small as possible but should include the full wire. The reference measurement has to be performed on the identical bare substrate with the same IR optics. The polarization of the IR light is important

quasistatic approximation that allows first-order estimates of resonances also for particles substantially larger than 1 % of wavelength if the incident light is uniform across the object, i.e., for IR antennas at normal incidence of light up to few 10 nm in diameter [12, 13]. In this approximation, for the fundamental resonance and $\omega_\tau << \omega$, the scattering cross section

$$\sigma_{\mathrm{sca}} = \omega_{\mathrm{p}}^2 \frac{Vr}{n_{\mathrm{s}} c} \times \frac{\omega^4 T_{\mathrm{L}}}{\left[\left(\omega_{\mathrm{res}}^2 - \omega^2\right)^2 + \omega^2 \left(\omega_\tau + \omega^2 T_{\mathrm{L}}\right)^2\right]} \tag{9}$$

and the absorption cross section

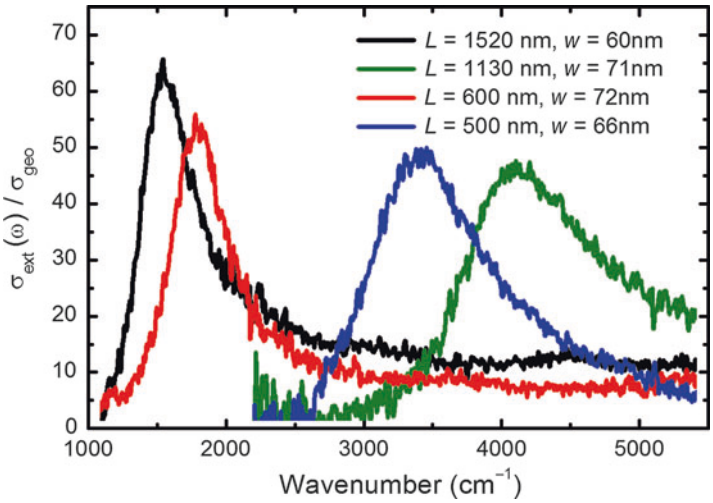
$$\sigma_{\mathrm{abs}} = \omega_{\mathrm{p}}^2 \frac{Vr}{n_{\mathrm{s}} c} \times \frac{\omega^2 \omega_\tau}{\left[\left(\omega_{\mathrm{res}}^2 - \omega^2\right)^2 + \omega^2 \left(\omega_\tau + \omega^2 T_{\mathrm{L}}\right)^2\right]} \tag{10}$$

differ in their asymmetric frequency-dependent spectral shape; notice the different exponents of ω in the two numerators. $T_{\mathrm{L}} = Vrn_{\mathrm{s}}\,\omega_{\mathrm{p}}{}^2/(6\pi c^3)$ is the Larmor-time parameter (with the vacuum velocity c of light) depending on the volume V (where the electrons are excited in the particle), the refractive index n_{s} of the surrounding medium, and the local-field ratio r [14]. For a spheroid in the quasistatic approximation, $r = \varepsilon_{\mathrm{s}}/[\varepsilon_{\mathrm{s}}+F(\varepsilon_\infty-\varepsilon_{\mathrm{s}})]$ with F as the depolarization factor [15]. For a sphere, for example, $r = 3\varepsilon_{\mathrm{s}}/(\varepsilon_\infty + 2\varepsilon_{\mathrm{s}})$ for an infinitely long needle $r = 1$ which is a reasonable approximation also for nanowires. For a plasmonic dipole of an infinitesimally small volume, T_{L} becomes also infinitesimally small and the damping of the oscillator gets equal to the electronic scattering rate ω_τ. However, related to the scattering of light in case of larger objects with finite V, the radiation damping rate $\omega_{\mathrm{rad}} = \omega^2 T_{\mathrm{L}}$ has to be included in the denominators of both the quasistatic expressions σ_{abs} and σ_{sca}. Typical values of ω_{rad} at $\omega = \omega_{\mathrm{res}}$ for IR nanoantennas from gold (with h and w similar to 50 nm, $n_{\mathrm{s}} \geq 1$) are of the order of the relaxation rate and thus are not negligible.

In experiments, resonances may appear much broader compared to the expectation based on the damping rates if the symmetry of the cylindrical shape is broken due to production errors. Then, for example, the even-order resonances (at 2 ω_{res}, usually not dipole active [1]) can be excited also by normal incidence of light. Further (Gaussian) broadening appears due to random parameter variations in a nanowire ensemble. In ensembles with interwire distances of the order of the resonance wavelength and below, nanowire interaction is an important issue in [16] not presented here.

Infrared Spectroscopy of Metallic Nanowires

With IR microspectroscopy (intense light sources are preferred), the far-field extinction related to a nanowire's plasmonic resonances can be detected well if a small aperture with the wire in its center is used [1, 16]. The IR spectra benefit from a carefully measured reference spectrum from the bare (identical) substrate. The extinction cross section $\sigma_{\mathrm{ex}} = \sigma_{\mathrm{abs}} + \sigma_{\mathrm{sca}}$ of a single nanoparticle



Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Fig. 3 Individual nanowire's extinction (normalized to the geometrical shadow) measured at room temperature in normal transmittance for an electric field parallel to the long axis of the wire. The transmittance is normalized to the IR spectrum of the ZnS substrate without any wire. The geometrical parameters (length L and width w) of the electron-beam lithographically produced gold nanostripes are given in the figure. The stripes' height was about 60 nm

(see examples in Fig. 3) on a thick transparent substrate (with refractive index $n_{\text{substrate}}$) can be estimated from the relative transmittance spectrum $T_{\text{rel}}(\omega)$ at normal incidence of light via the relation $\sigma_{\text{ext}} = A_0 \cdot (1 - T_{\text{rel}}) \cdot (n_{\text{substrate}} + 1)/2$ where A_0 is the focal area (with one particle in the center) [1, 16]. If the preconditions mentioned above are fulfilled, a fit based on the Eqs. 8 and 9 can deliver the parameters ω_τ and ω_{res} (and thus λ_{res} that via Eqs. 7 and 8 can be compared to the value expected from the ideal plasma frequency).

Conclusion

Bringing together well-established knowledge on metal optical properties, metallic resistivity, and optical antennas, the strong relationship between electrical conductivity parameters and the fundamental plasmonic resonance of nanowires in the infrared is clarified.

Cross-References

- ▶ AFM
- ▶ Gold Nanorods
- ▶ Effective Media
- ▶ Electron Beam Lithography (EBL)
- ▶ Local Surface Plasmon Resonance (LSPR)
- ▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods
- ▶ Plasmonics
- ▶ Spectromicroscopy

References

1. Pucci, A., Neubrech, F., Aizpurua, J., Cornelius, T., de la Chapelle, M.L.: Electromagnetic nanowire resonances for field-enhanced spectroscopy. In: Wang, Z. (ed.) One-Dimensional Nanostructures, pp. 175–216. Springer, New York (2008)
2. Steinhögl, W., Schindler, G., Steinlesberger, G., Engelhardt, M.: Size-dependent resistivity of metallic wires in the mesoscopic range. Phys. Rev. B **66**, 075414 (2002)
3. Ashcroft, N.W., Mermin, N.D.: Solid State Physics. Sounders College Publishing, Orlando (1976)
4. Abeles, F.: Optical properties of metals. In: Abeles, F. (ed.) Optical Properties of Solids, pp. 93–162. North Holland, Amsterdam (1972)
5. Ordal, M.A., Bell, R.J., Alexander Jr., R.W., Long, L. L., Querry, M.R.: Optical properties of fourteen metals in the infrared and far infrared: Al, Co, Cu, Au, Fe, Pb, Mo, Ni, Pd, Pt, Ag, Ti, V, and W. Appl. Optics **24**, 4493–4499 (1985)

6. Young, C.-Y.: The frequency and temperature dependence of the optical effective mass of conduction electrons in simple metals. J. Phys. Chem. Solids **30**, 2765–2769 (1969)
7. Ziman, J.M.: Principles of the Theory of Solids. Cambridge University Press, Cambridge (1979)
8. McKay, J.A., Rayne, J.A.: Temperature dependence of the infrared absorptivity of the noble metals. Phys. Rev. B **13**, 673–685 (1976)
9. Bittar, A.: The Bruggeman Effective Medium Theory Applied to the Optical Properties of Inhomogeneous Materials. Physics and Engineering Laboratory, Lower Hutt (1984)
10. Zhang, X., Stroud, D.: Optical and electrical properties of thin films. Phys. Rev. B **52**, 2131–2137 (1995)
11. Novotny, L.: Effective wavelength scaling for optical antennas. Phys. Rev. Lett. **98**, 266802 (2007)
12. Sarid, D., Challener, W.A.: Modern Introduction to Surface Plasmons. Cambridge University Press, Cambridge (2010)
13. Kats, M.A., Yu, N., Genevet, P., Gaburro, Z., Capasso, F.: Effect of radiation damping on the spectral response of plasmonic components. Opt. Express **19**, 21748–21753 (2011)
14. Doyle, W.T.: Electrodynamic response of metal spheres. J. Opt. Soc. Am. A **2**, 1031–1034 (1985)
15. Pelton, M., Bryant, G.: Introduction to Metal-Nanoparticle Plasmonics. Wiley, Hoboken (2013)
16. Weber, D., Pucci, A.: Antenna interaction in the infrared. In: de la Chapelle, M.L., Pucci, A. (eds.) Nanoantenna: Plasmon -Enhanced Spectroscopies for Biotechnological Applications, pp. 175–194. Pan Stanford Publishing, Singapore (2013)

Confocal Laser Scanning Microscopy

Reinhold Wannemacher
Madrid Institute for Advanced Studies, IMDEA Nanociencia, Madrid, Spain

Synonyms

Confocal scanning optical microscopy (CSOM); Laser scanning confocal microscopy

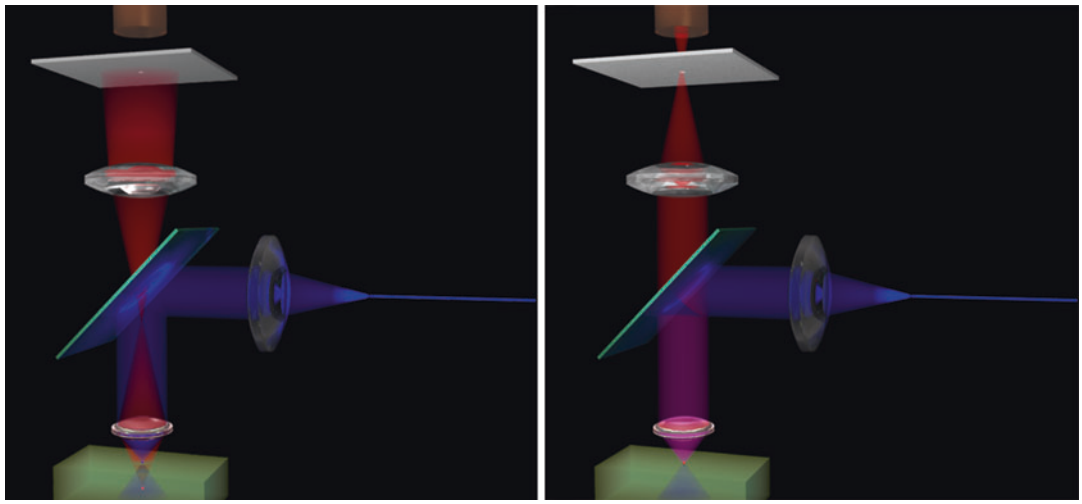
Definition

A confocal laser scanning microscope (CLSM) images a point light source used for excitation onto the sample via the objective lens and images the excited focal volume onto a point detector using reflected, transmitted, emitted, or scattered light. In contrast to conventional microscopes, this scheme permits strong rejection of out-of-focus light and optical sectioning of the sample. In order to obtain an image, the focal volume must be scanned relative to the sample. Scanning can be performed in the lateral as well as in the axial directions, and three-dimensional images of the sample can be generated in this way.

Operating Principle

Figure 1 illustrates the working principle of a fluorescence confocal microscope. A point light source, here the end of an optical fiber carrying the excitation light, is imaged onto the sample by the objective lens via a beam splitter. Emitted light from the focal spot is imaged through the beam splitter onto a pinhole via an auxiliary lens. The light is registered by the detector, because it passes through the pinhole (left-hand side of Fig. 1). Fluorescence from a fluorophore at an out-of-focus position in the excitation cone, on the other hand, arrives at the screen defocused. Therefore, only a small fraction of it passes through the pinhole and reaches the detector. The effect is understated by the schematic figure and actually much stronger in a real microscope because of the short focal length of the objective lens. The optical sectioning capability is the core of confocal microscopy and allows to render the object three-dimensionally under different angles by appropriate software once a stack of images at different depths has been acquired. In addition, the lateral resolution is slightly improved in confocal microscopy, compared to conventional microscopy, when the pinhole is small. On the other hand, an image can be acquired in this way only by serial scanning of the sample or of the excitation beam (for technical improvements in this respect see section "Confocal Microscopy Involving Modified Illumination").

Because a CLSM operates with light, it may be used to image many different physical quantities. These may be simply reflected or transmitted intensity (brightfield confocal microscopy) or the



Confocal Laser Scanning Microscopy, Fig. 1 Principle of confocal laser scanning microscopy, demonstrating the strong rejection of out-of-focus light

intensity of fluorescence excited in the sample (fluorescence confocal microscopy). Other options include polarization and phase of reflected or transmitted light, as well as the intensity, wavelength, lifetime, time correlation, or recovery after photobleaching of fluorescence from the sample or intensity, wavelength, and polarization of inelastically (Raman) scattered light. Moreover, a nonlinear response of the sample to the optical excitation near the laser focus, based, for example, on multiphoton excitation, second harmonic generation, or stimulated scattering processes may be used for confocal microscopy. Some of these options will be discussed in section "Variants of Confocal Laser Scanning Microscopy."

Optical sectioning and three-dimensional image acquisition being the essential feature of confocal optical microscopy, it is worth mentioning here that an alternative (diffraction-limited) brightfield optical microscopy technique with similar capability is digital holographic microscopy, although this technique is far less widely known and used. Here, the object is reconstructed from an intensity camera image and no scanning is necessary. Moreover, a phase image is obtained in addition to an amplitude image. Significant improvements in object reconstruction algorithms have been made in recent years, and lens-based versions with external reference beam as well as lensless versions have been demonstrated. Lens-based instruments with external reference beam are commercially available.

Basic Theory of the Confocal Microscope

The spatial resolution of modern high-quality conventional as well as confocal microscopes is limited by diffraction. This means that within the design spectral range of the objective lens aberrations, such as spherical aberration, astigmatism, coma, field curvature, distortion, and chromatic aberration, have a significantly smaller impact on the resolution of the microscope than diffraction. An important exception to this statement arises from aberrations due to refraction, when sample regions well inside refracting samples have to be imaged.

Most modern microscope objectives are now infinity corrected, that means they are corrected for forming an image at infinity. A tube lens is in this case required to form a real image at finite distance. The limitations by diffraction are, however, in any case dominated by the objective lens and not by the tube lens. This is due to the dependence of the diffraction limit on the opening angle of the rays contributing to the image, which is much smaller for the tube lens than for the objective lens.

Point Spread Function of the Confocal Microscope

The three-dimensional intensity distribution in the image space corresponding to a single-point object, demagnified by the magnification of the optical system, is called the (intensity) point spread function (PSF) of the lens. The PSF of a confocal microscope with an infinitesimally small pinhole is given by [1, 2]:

$$\mathrm{PSF}_{\mathrm{CF}}(x,y,z) = \mathrm{PSF}_{\mathrm{ill}}(x,y,z) \cdot \mathrm{PSF}_{\mathrm{det}}(x,y,z) \tag{1}$$

Here, $\mathrm{PSF}_{\mathrm{CF}}$, $\mathrm{PSF}_{\mathrm{ill}}$, and $\mathrm{PSF}_{\mathrm{det}}$ represent the point spread functions for the confocal imaging and the illumination and detection paths, respectively. Neglecting the contribution from the tube lens, as well as aberrations of the objective lens, the latter two functions are simply the point spread functions of a simple lens, which, in the paraxial and scalar approximation, can be calculated by means of the Huygens-Fresnel principle as [3]

$$\mathrm{PSF}(x,y,z) = |h(x,y,z)|^2 \tag{2}$$

with

$$h\left(\vec{r}\right) = \frac{\mathrm{C}'}{\lambda}\int\int_{A}\frac{\mathrm{e}^{iks}}{s}\mathrm{dA} \approx \frac{\mathrm{C}}{\lambda}\int\int_{\Omega} e^{-ik\vec{q}\cdot\vec{r}}\mathrm{d}\Omega. \tag{3}$$

Here, $h\left(\vec{r}\right)$ represents the scalar complex amplitude of the field in the image space at a position $\vec{r} = (x,y,z)$ relative to the location of the geometric focus; λ is the wavelength, $k = 2\pi/\lambda$: s is the distance between the point P at position $\vec{r}$ in the image space and a point Q at position $f\,\vec{q}$ in the pupil A of the lens of focal length f: $\vec{q}$ is a unit vector in the direction of Q: Ω is the solid angle subtended by the aperture of the lens as seen from the origin at the geometric focus, and C and C′ are constants. Equation 3 assumes homogeneous illumination of the lens. Inhomogeneous illumination can be taken into account by multiplying the integrand with a corresponding pupil function (compare section "Apodization"). For a confocal microscope operating in reflection or transmission modes $\mathrm{PSF}_{\mathrm{ill}}(\mathrm{x,\ y,\ z}) \approx \mathrm{PSF}_{\mathrm{det}}(\mathrm{x,\ y,\ z})$, and both functions are then identical to the one given in Eq. 2. This results in

$$\mathrm{PSF}_{\mathrm{CF}}(x,y,z) = |h(x,y,z)|^4. \tag{4}$$

Experimentally, this function would be observed when a point object is scanned through the focus of the instrument in both cases. It should be kept in mind here that the paraxial approximation is contrary to the actual typical situation in optical microscopy. The paraxial approximation, nevertheless, works surprisingly well even for N.A. ≈ 0.5 ($\theta_0 = 30°$ in the case of a dry lens, see below) and gives reasonable estimates of the diffraction limit even for higher numerical aperture objective lenses. Deficiencies of the scalar approximation will be discussed in section "Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light."

Figure 2 displays the three-dimensional intensity point spread function of a conventional (a) and a confocal (b) microscope calculated according to Eqs. 1 and 3. The numerical aperture of the objective lens is N.A. = 0.5. z is the coordinate along the axis of the lens and r the coordinate perpendicular to the optical axis, measured in wavelengths. The drastic reduction in side lobes for a confocal microscope is immediately evident. This also leads to drastic reduction in laser speckle in the case of coherent illumination.

Single-Point Resolution in the Focal Plane

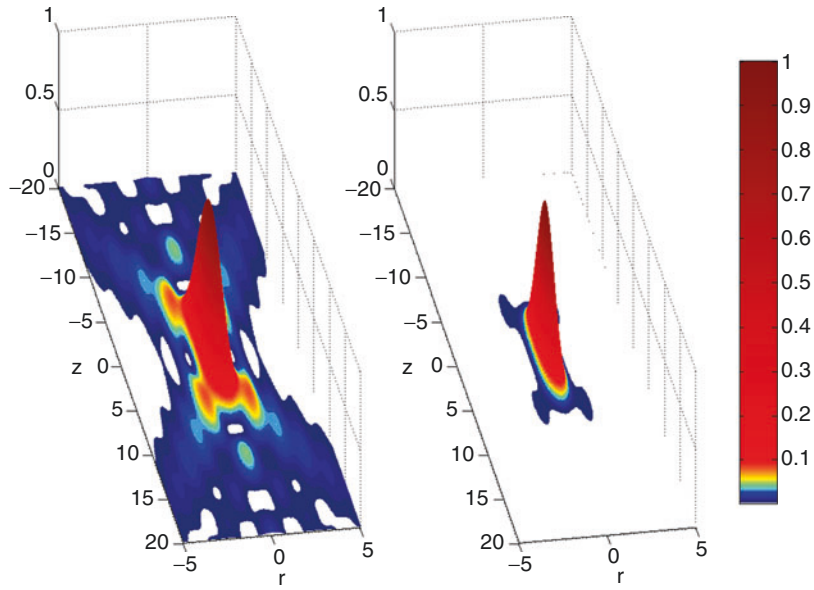
In the focal plane ($z = 0$), the PSF of the conventional microscope, as calculated from Eq. 3, coincides with the well-known Airy pattern

$$\mathrm{PSF} = \left(\frac{2J_1(\mathrm{v})}{\mathrm{v}}\right)^2 \tag{5}$$

with

$$\mathrm{v} = kr \cdot n\sin\theta_0 = kr \cdot \mathrm{N.A.} \tag{5a}$$

Here, r is the distance from the axis; θ_0 is the angle of a marginal ray passing through the aperture



Confocal Laser Scanning Microscopy,
Fig. 2 Intensity point spread function of a conventional (**a**) and a confocal (**b**) microscope in the scalar and paraxial approximations. The numerical aperture of the objective lens is N.A. = 0.5. z is the coordinate along the axis of the lens and r is the lateral dimension, both measured in wavelengths

toward the geometric focus, relative to the optical axis; and n is the refractive index in the image space. From this expression, the lateral full width at half maximum (FWHM) of the PSF of the conventional microscope

$$\mathrm{FWHM} = \frac{0.51\lambda}{\mathrm{N.A.}} \tag{5b}$$

is easily calculated. Equation 3 yields

$$\mathrm{PSF_{CF}} = \left(\frac{2\mathrm{J}_1(\mathrm{v})}{\mathrm{v}}\right)^4 \tag{5c}$$

for the PSF of the confocal microscope in reflection mode, and therefore

$$\mathrm{FWHM_{CF}} = \frac{0.37\lambda}{\mathrm{N.A.}}. \tag{5d}$$

Equations 5b and 5d demonstrate the enhancement in lateral (single-point) resolution for a confocal microscope relative to a conventional one, in the case when the pinhole is closed completely. As an example, for $\lambda = 488$ nm, N.A.=0.9, FWHM = 277 nm, and $\mathrm{FWHM_{CF}} = 201$ nm.

Single-Point Resolution on the Axis: Depth Response

Similarly, the PSF of the conventional microscope on the optical axis is calculated from Eq. 3 as

$$\mathrm{PSF}(u) = \left(\frac{\sin(u/4)}{u/4}\right)^2 \tag{6}$$

with

$$u = nkz\sin^2\theta_0. \tag{6a}$$

It turns out that the definition

$$u = 4kz\sin^2(\theta_0/2) = 2nkz(1 - \cos\theta_0) \tag{6b}$$

which is equivalent to Eq. 6a in the paraxial approximation is more appropriate at higher numerical apertures. The corresponding FWHM of the depth response is therefore

$$\mathrm{FWHM} = \frac{0.89\lambda}{n(1 - \cos\theta_0)} \tag{6c}$$

for the conventional microscope. Correspondingly, for the confocal microscope

$$\mathrm{PSF}_{\mathrm{CF}}(u) = \left(\frac{\sin(u/4)}{u/4}\right)^4 \tag{6d}$$

and

$$\mathrm{FWHM}_{\mathrm{CF}} = \frac{0.64\lambda}{n(1-\cos\theta_0)}. \tag{6e}$$

For the parameters used in the example above, $\lambda = 488$ nm, N.A. $= 0.9$, FWHM $= 770$ nm, and $\mathrm{FWHM}_{\mathrm{CF}} = 554$ nm. The FWHM of the point spread functions on the optical axis of the microscope is therefore about three times larger than the lateral FWHM for the conventional as well the confocal microscope, although the precise value depends on the numerical aperture and although the scalar theory used here is not applicable at large numerical aperture (see section "Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light").

V(z)

A clearer demonstration of the different optical sectioning capabilities of the conventional and confocal microscopes is obtained with planar objects instead of point objects. The depth response of a confocal microscope operating in reflection is often characterized by axially scanning a mirror through the focus position and registering the light intensity behind the pinhole during the scan. The corresponding amplitude function is called $V(z)$, an expression coined originally for the acoustic microscope, which is also a confocal instrument and to which the same scalar theory is applicable. Because the image of the illuminating infinitesimal pinhole is moving by a distance of $2z$ when the mirror moves by a distance z, the intensity detected behind the pinhole is derived from Eq. 6b, by replacing u by $2u$:

$$\begin{aligned} I(z) &= |V(z)|^2 = \left|\frac{\sin(u/2)}{u/2}\right|^2 \\ &= \left|\frac{\sin(nkz(1-\cos\theta_0))}{nkz(1-\cos\theta_0)}\right|^2 \end{aligned}. \tag{7}$$

This equation predicts a central maximum of width

$$\mathrm{FWHM}_{\mathrm{CF}} = \frac{0.44\lambda}{n(1-\cos\theta_0)} \tag{8}$$

which, in paraxial approximation $\theta_0 << 1$ for a dry lens, becomes

$$\mathrm{FWHM}_{\mathrm{CF}} = \frac{0.89\lambda}{\mathrm{N.A.}^2}. \tag{9}$$

Equation 7 also implies symmetric side lobes for negative and positive defocus. As an example, for a dry lens of N.A. $= 0.8$ and an operating wavelength of 488 nm, Eq. 8 yields FWHM $= 537$ nm. In a conventional microscope, on the other hand, the signal received by a large area detector would be independent of the position of the mirror.

The simple theory presented so far predicts a symmetric $V(z)$ function. Whereas the width of the main peak of the $V(z)$ is typically very close to measurements performed with real lenses, the aberrations present in any real objective lens typically lead to deviations as far as the side maxima are concerned and in particular to asymmetry in $V(z)$ for positive and negative defocus. This may be used for quantitative characterization of, for example, the amount of spherical aberration present in the optical system. Interferometric versions of confocal microscopy, however, have been more traditionally used for this purpose.

Two-Point Resolution: Rayleigh and Sparrow Criteria

The single-point resolution of the confocal microscope, as given by the PSF discussed above, is in most cases not the relevant quantity to judge the resolution of the instrument, because what is really desired is the capability to resolve certain details of a microscopic object consisting of various parts. It is therefore important to quantify the two-point resolution of the instrument, which means the capability to resolve two-point objects close to each other. There is some arbitrariness in this definition, because it depends on the subjective judgment, under which conditions two-point objects are resolved in an image. Only the lateral two-point resolution will be discussed here.

The Rayleigh criterion defines two-point sources as resolved, if the image of the second

point lies at the first zero of the image of the first one or at a larger distance. This leads to a lateral resolution

$$d_{\mathrm{R}} = \frac{0.61\lambda}{\mathrm{N.A.}} \tag{10}$$

for the conventional microscope and

$$d_{\mathrm{R,CF}} = \frac{0.56\lambda}{\mathrm{N.A.}} \tag{11}$$

for the confocal microscope, just 8 % less than for the conventional microscope. In the case of coherent sources, d_{R} depends on the phase difference of the sources: whereas two out-of-phase coherent sources can be clearly resolved, because there will be a zero of intensity halfway between the images of the two sources, the sources cannot be resolved, if they are in phase with each other, because the maximum will lie in the middle between the images of the two individual sources.

There is another criterion for the two-point resolution, which is more generally applicable, because it does not refer to a zero of the response function. The Sparrow criterion states that two sources are considered to be resolved, when the intensity halfway between the two images is the same as the one at the individual image locations. For incoherent illumination, this results in

$$d_{\mathrm{S}} = \frac{0.51\lambda}{\mathrm{N.A.}} \tag{12}$$

for *both* the conventional and confocal microscopes.

Coherence in Brightfield and Fluorescence Microscopy

The imaging of extended objects differs significantly for the conventional and confocal microscopes, respectively. For a CLSM operating in brightfield (reflection or transmission) mode, the imaging is spatially coherent, because the illumination generates a spatially coherent field distribution in the focus of the objective lens, as given by the complex amplitude point spread function of the objective lens. For a CLSM in reflection mode, this field distribution has to be multiplied by the reflectance R of the sample, and this weighted field distribution is then imaged onto the pinhole implying convolution with the combined amplitude point spread function of the objective and pinhole relay lenses. Assuming imaging of the object onto the pinhole by the same objective lens that is used for excitation and neglecting contributions to the point spread function from the pinhole relay lens in the second imaging step, the intensity behind the infinitesimally small pinhole can therefore be written as the convolution

$$I = \left| \iiint h(x,y,z) R(x,y,z) h(-x,-y,-z) \mathrm{d}x\mathrm{d}y\mathrm{d}z \right|^2 . \tag{13}$$

Here, $h(x,y,z)$ is the amplitude PSF of the objective lens (compare Eq. 3). For the case of an even PSF, Eq. 13 is equivalent to

$$I = \left| h^2 * R \right|^2 \tag{14}$$

that means the signal is given by the absolute square of the convolution of the amplitude PSF of the confocal microscope, $h^2\left(\vec{r}\right)$, with the local amplitude reflectivity of the sample.

In the case of a conventional microscope, on the other hand, the illumination is approximately incoherent and therefore

$$I = |h|^2 * |R|^2 . \tag{15}$$

In reality, for the conventional microscope, imaging is partially coherent, because the emission from each emitting point on the illumination source is imaged, due to diffraction at the condenser aperture, into a finite spatial region, which is occupied by a coherent field due to that point emitter, and the regions in the image space corresponding to neighboring, incoherently emitting points on the source partially overlap on the sample. This is true for critical illumination, where the spatial region would be given by the PSF of the condenser, as well as for Köhler illumination, where the spatial region is the whole illuminated region of the sample.

Fluorescence Confocal Microscopy

Fluorescence imaging is incoherent. Assuming that the fluorescence intensity is proportional to the excitation intensity, the signal obtained in a confocal microscope with an infinitesimally small pinhole is

$$I = \left(|h(\lambda)|^2 |h(\beta\lambda)|^2\right) * f \qquad (16)$$

where it has again been assumed that, as in standard commercial confocal microscopes, the same objective lens is used for excitation and fluorescence imaging, respectively. Here, f represents the distribution of fluorescent centers in the sample, λ the excitation wavelength, $\beta\lambda$ the fluorescence wavelength, and β the ratio of both wavelengths, the Stokes ratio. In the case of several different types of emitters, f would have to be weighted according to the spectral contribution of each emitter to the detected signal, which depends on the filters employed in detection for rejection of the excitation and also on the wavelength-dependent sensitivity of the detector. For a point emitter placed at the focus, the convolution with a δ function just yields the first two terms on the right-hand side of Eq. 16. For a conventional microscope, on the other hand, because the whole sample is illuminated, the single-point resolution is only determined by the intensity PSF of the objective lens at the fluorescence wavelength, and the excitation wavelength is irrelevant.

Effects of Finite Pinhole Size

A finite size of the pinhole is obviously required in order to obtain a measurable signal. This will reduce the lateral resolution as well as the optical sectioning capability. It can be shown [2] that for a single-point object, the lateral resolution is almost unaffected by the size of the pinhole, if

$$v_P = \frac{2\pi r_p}{\lambda M} \sin\alpha \leq 0.5 \qquad (17)$$

where r_P is the radius of the pinhole and M the magnification of the lens. The maximum value allowed by Eq. 17 therefore sets a reasonable value for the pinhole size, if optimum lateral resolution and reasonable signal are desired. As an example, in the case of a 100×/0.8 N.A. objective lens and $\lambda = 514$ nm, a critical diameter of the pinhole of 10.2 µm is calculated from Eq. 17. The depth discrimination, as measured by moving a mirror through focus, on the other hand, is less affected and is essentially unaltered if

$$v_p \leq 2.5 \qquad (18)$$

It is obvious from these equations that the pinhole size must be adapted when the objective lens is changed.

Apodization

Equation 3 assumed rectangular apodization, that is, homogeneous illumination of the lens pupil and neglects reflection losses at the lens. For a given objective lens, rectangular apodization yields the smallest FWHM of the focus in the focal plane, at the expense of larger side maxima, compared to pupil functions falling off toward the edge of the objective lens. It can be achieved only approximately with Gaussian laser beams and is then equivalent to loss of a large fraction of the power of the excitation beam. Gaussian apodization, on the other hand, increases the FWHM, but reduces the side maxima.

Deconvolution

As described by Eq. 16, the image acquisition process in confocal fluorescence microscopy can be modeled as a convolution of the spatially dependent fluorescence of the sample with a point spread function (PSF) of the imaging system. This is true also for conventional non-confocal fluorescence microscopy. In addition, random noise is superimposed on the image. Deconvolution with the PSF would naturally seem the appropriate way to determine the true fluorescence distribution in both cases. Applied to confocal images, the resolution may be improved. In the case of conventional microscopy, optical sectioning and removal of

out-of-focus blur may be achieved by post-processing instead of employing hardware in the optical setup.

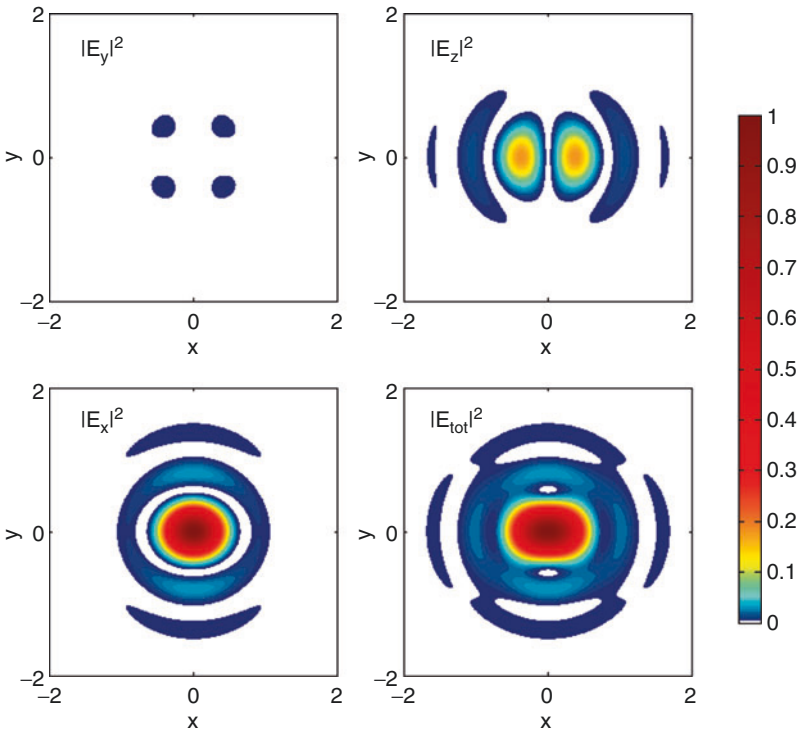
In general, the 3D PSF, necessary for this procedure, can be obtained experimentally or analytically. In the experimental methods, images of one or more point-like objects are collected. The problem with this technique lies in the poor signal-to-noise ratio obtainable with very small objects and the fact that the PSF may vary depending on the sample. In analytical calculations of the PSF, aberrations of the optical system are often partially taken into account, whereas, on the other hand, the scalar approximation is most often used and the effects of the vector character of light (compare section "Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light") are neglected.

Many 3D deconvolution methods are currently employed and some are available in commercial and noncommercial software packages [4]. The simplest class are *neighboring* methods, in which out-of-focus blur is removed by subtraction of neighboring (filtered) images within a stack. This method does not sufficiently remove noise. In contrast to that, *linear* methods apply deconvolution to the whole stack of images at once. Examples are inverse filtering, Wiener filtering, the linear least squares, and the Tikhonov filtering techniques. The last three methods do not restore high-frequency object components beyond the bandwidth of the PSF, and inverse filtering suffers from noise amplification. All methods are very sensitive to error in the PSF. Therefore, constrained iterative nonlinear algorithms are often employed, in which, starting from a guess for the true object, an error is minimized under certain constraints (positiveness of the image, finite support of the sample, etc.). In cases of strong noise in the image, *statistical* iterative methods (like the maximum likelihood method) are favored. A computational alternative are *blind deconvolution* methods, in which the PSF of the optical system and the "true object" are simultaneously determined in a converging iteration. These methods are, however, computationally demanding and sensitive to noise, and solutions may be nonunique.

Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light

In view of many more recent developments in confocal microscopy, it appears useful to shortly discuss deviations from the simple scalar theory. These deviations become increasingly important with increasing numerical aperture of the objective lens. In the case of linear polarization of the excitation beam, cylindrical symmetry is lost. The field in the focal plane and exactly on axis is then polarized in the direction of the excitation, but, away from the axis, it is elliptically polarized with a longitudinal component pointing in the direction of the optical axis. The intensity PSF becomes elongated and approximately elliptical, with the major axis of the ellipse in the direction of the excitation. Both effects increase with increasing numerical aperture. In the case of linearly polarized excitation, the two-point resolution of an ordinary optical microscope equipped with a well-corrected high numerical aperture objective lens (as well as that of a similar confocal microscope) therefore depends on the orientation of the line connecting the two points relative to the incoming polarization (as well as on the orientation of the dipolar point reflectors or absorbers/emitters). Figure 3 shows the PSF for numerical aperture N.A. = 0.95 and the absolute squares of all electric field components in the focal plane. In the scalar approximation, the lines of constant intensity would, of course, be circles, which is clearly not the case in the figure appearing in the lower-right corner of Fig. 3. Moreover, the maxima of the longitudinal field occur on both wings of the main maximum, along the direction of the incoming polarization (x direction), and the maximum absolute square of E_z is approximately 20 % of that of E_x.

The vector diffraction problem was first solved by Richards and Wolf [5]. A somewhat more physical treatment employs expansion of the field in the image space into vector multipoles centered at the focus [6]. This latter approach is also of interest for matching the focal field distribution to the fields of a dipole via the amplitude distribution and the polarization in the pupil plane. In this way, the coupling of the field to single atoms or molecules can be significantly enhanced, which is of



Confocal Laser Scanning Microscopy, Fig. 3 PSF of a well-corrected microscope objective of a high numerical aperture lens (N.A. = 0.95) for the case of linear polarization of the incoming beam, calculated using the Debye-Wolf integral. A constant pupil function has been assumed here and, correspondingly, reflection losses in the lens have been neglected. z is the coordinate along the axis of the lens and x, y are the lateral coordinates, all measured in wavelengths. The incoming polarization is along the x axis

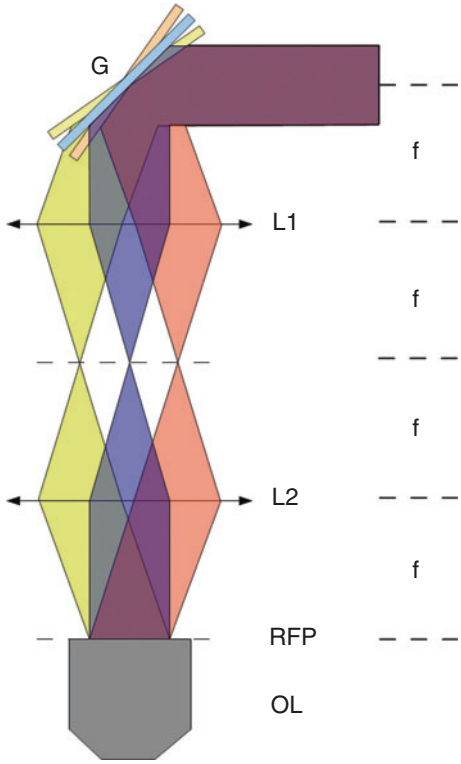
interest, for example, for quantum optical applications. The vectorial approach is in general also required to treat focusing of other distributions of intensity, phase, and polarization in the pupil plane of the lens. Examples relevant for applications include radially polarized excitation, producing a longitudinally polarized focus, azimuthally polarized excitation, which leads to a doughnut-shaped intensity distribution in the focal plane, or combinations of these distributions with scalar vortices, that means helical phase fronts. A longitudinal focus is essential for tip-enhanced Raman microscopy (TERS, compare section "Confocal Raman Microscopy," ▶ Scanning Near-Field Optical Microscopy), which combines a near-field technique with confocal imaging. Other distributions are relevant for optical tweezers.

Instrumental Details

Scanning Techniques

Whereas scanning the sample is an option, particularly in laboratory setups, many commercial confocal microscopes employ lateral scanning of the laser beam and sample scanning in the vertical direction. Beam scanning is typically achieved using mirrors mounted on galvanometer motors, which allow to vary the angle at which the beam passes the rear focal plane of the objective lens. An example for a telecentric 4f system that allows to vary this angle without displacing the beam in the rear focal plane is shown in Fig. 4.

Recently, resonant galvanometer-based beam scanning systems have become commercially available, which employ torsion spring-based sinusoidal oscillations of the scan mirror with frequencies in the kilohertz range with open-loop operation for the fast scan axis. This permits frame rates on the order of 30 frames per second and in this way allows to study fast processes, such as diffusion in biological cells or to avoid blurring of the image due to movement of organs in in vivo studies. Alternative fast-scanning confocal microscopes are discussed in section "Variants of Confocal Laser Scanning Microscopy."



Confocal Laser Scanning Microscopy, Fig. 4 Telecentric lens system minimizing beam walk off. *f* focal length of lenses L1 and L2, *G*: scan mirror mounted on galvanometer scanner, *RFP* rear focal plane of objective lens OL

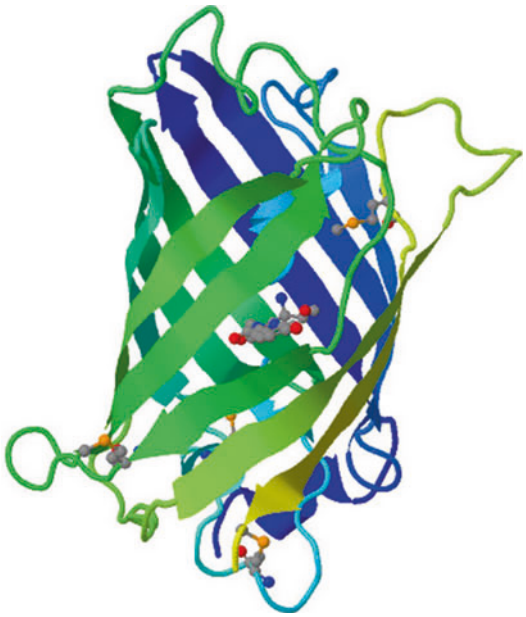
Excitation Sources and Beam Delivery

Ion lasers, HeNe lasers, diode-pumped solid-state lasers, and diode lasers have all been used as continuous wave light sources in confocal laser scanning microscopy. Argon ion lasers provide a choice of several excitation wavelengths and are therefore still popular in spite of their low efficiency. In order to combine several laser beams, dichroic beam splitters are employed. In many cases, a software-controlled acousto-optic tunable filter (AOTF) selects the excitation wavelength (s) of choice from this combined beam. The AOTF is based on the diffraction of light from ultrasonic waves generated in a birefringent crystal by an ultrasonic transducer. Incident and diffracted waves propagate as ordinary and extraordinary wave in the crystal, respectively, or vice versa and are therefore polarized perpendicular to each other. This allows convenient rejection of the incident light by a polarizer. Because of momentum conservation, the difference in the optical wave vectors of both beams must be equal to the acoustic wave vector. This means that the wavelength of the diffracted beam is controlled by the ultrasonic frequency. In a collinear AOTF, both optical beams and the acoustic wave propagate in the same direction, independent of the optical wavelength.

The optical output of the AOTF is typically fed into an optical fiber, which delivers the beam to the input optics of the confocal scan head of the microscope. In brightfield reflection confocal microscopy, the beam splitter which directs the excitation light toward the objective lens (compare Fig. 1) induces a considerable loss for the excitation as well as for the detected light. This loss is minimized by a 50/50 beam splitter. In fiber-based confocal microscopes, instead of beam splitters, 2 × 2 fiber couplers are typically employed. An improved version would make use of optical circulators, but these are presently not widely available for wavelengths in the visible range. In fluorescence confocal microscopy, the detected wavelength differs from the excitation wavelength and therefore dichroic dielectric beam splitters are used which are highly reflecting (transmitting) at the excitation (detection) wavelength and minimize losses in this way.

Detection

Standard detectors in confocal laser scanning microscopes are photomultipliers, which are in many cases operated in the analogue mode, which means by measuring the anode current and integrating over the pixel dwell time. Photon counting, on the other hand, is typically employed in fluorescence correlation and fluorescence lifetime microscopy (compare section "Variants of Confocal Laser Scanning Microscopy"), where usually avalanche photodiodes with high quantum efficiency and fast response times replace photomultipliers as detectors. Confocal



Confocal Laser Scanning Microscopy, Fig. 5 Tertiary structure of the green fluorescent protein [8]. The fluorescent chromophore, composed of three amino acids is located in the center of the beta-barrel protein cage, length about 4 nm, which prevents quenching of the fluorescence by water

microscopes which allow to spectrally disperse the light passing the confocal aperture often employ a charge-coupled camera (CCD) attached to a spectrograph to register the spectra. Back-illuminated Peltier or liquid nitrogen cooled CCDs provide high quantum efficiency (above 90 % over a wide spectral range) and low background noise, which is important for single molecule detection or when working with less photostable fluorescent probes.

Fluorescent Probes

Many samples are autofluorescent and therefore allow fluorescent imaging without having to introduce additional fluorescent probes. In biological samples, however, autofluorescence is typically weak, and therefore the sample frequently had to be stained with appropriate dyes. The latter, however, are often highly phototoxic in living cells. An important development in light microscopy, including confocal laser scanning microscopy, started in the year 1994 when the green fluorescent protein (GFP, see Fig. 5) from the jellyfish *Aequorea victoria* was genetically expressed in bacteria making them fluorescent at room temperature. In the same way, it is now generally possible to label proteins of interest in biological cells with fluorescent proteins by genetic manipulation, which can be achieved, for example, by injection of a virus vector. The number of fluorescent proteins used in the field has exploded by now and they are widely used in optical microscopy because of their considerably reduced phototoxicity, brightness, and photostability [7]. Genetically modified fluorescent proteins from *Aequorea victoria* now cover the spectral range from the deep blue to yellow and others derived from Anthozoa species (corals and anemones), as well as other sources, span the entire visible spectrum. The tertiary structure and size of these fluorescent proteins are very similar to those derived from *Aequorea victoria*, although the amino acid sequences are quite different. Red-emitting species with correspondingly longer excitation wavelengths are of particular interest because of reduced autofluorescence, deeper penetration, and better resistance toward high excitation density of biological tissue in this spectral region. Research is ongoing to improve brightness and photostability, reduce oligomerization and pH sensitivity, improve the appropriateness for fusion tagging, and reduce the time required for maturation of the protein in living organisms.

Variants of Confocal Laser Scanning Microscopy

Confocal Microscopy Involving Modified Illumination

Slit-Scanning Confocal Microscopes

Scanning a line focus, generated, for example, by a cylindrical lens, over the sample and imaging this line focus onto a slit aperture parallel to the image of the line focus still provides the optical

sectioning capability of the confocal microscope, because the slit rejects out-of-focus light. At the same time, the frame rate is significantly increased, because scanning is necessary only in one direction. A spatially sensitive detector must be used to resolve light passing through different positions along the exit slit. This may be achieved by imaging the exit slit onto a one-dimensional detector array, read out synchronously with the scan, or by scanning an image of the exit slit, synchronously with the scan, across a two-dimensional detector, such as a CCD camera, forming a confocal image in this way. Disadvantages over the single-point scanning technique include reduced lateral resolution in the direction of the line focus, enhanced out-of-focus background, and, in the case of coherent illumination, increased laser speckle.

Spinning Disk Confocal Microscopes

Instead of scanning a single-point focus across the sample, multiple focal spots may be simultaneously generated and imaged each onto a confocal aperture. A white light version of such a confocal microscope based on a spinning Nipkow disk was introduced by Petran and Hadravsky already in the 1960s and later improved by Xiao, Corle, and Kino. The disk contains pinholes arranged in a spiral pattern, which are slightly displaced such that the whole sample is illuminated after the disk has rotated by a certain angle. In more recent versions, light from each focal spot on the sample passes the same pinhole in the disk that was used for excitation. A two-dimensional detector, such as the eye of the observer or a CCD camera, is used to register a confocal image while the disk is spinning. Nipkow disk-based confocal microscopes are now commercially available from several manufacturers and provide fast confocal imaging, but at the cost of reduced flexibility, because the pinhole size cannot be varied and because the beams cannot be steered at will, as it is necessary, for example, in some experiments involving photobleaching. Moreover, cross-talk between the different focal spots may occur. Another version of confocal microscopy with multiple focal spots, *swept field confocal microscopy*, leaves the pinhole array stationary and sweeps the image of this array over the sample. By switching between different pinhole arrays, the pinhole size can be varied.

The light efficiency of Nipkow disk-based confocal microscopes may be significantly improved by adding a microlens array, mounted on a second disk, which is spinning on the same axis as the Nipkow disk and placed on top of the latter. Each microlens focuses incoming light onto one of the pinholes of the Nipkow disk. A dichroic beam splitter between both disks may be used to direct the detected fluorescence onto a camera. Another option is to use slit-shaped apertures on the Nipkow disk, which results in the same advantages and disadvantages as already described in section "Slit-Scanning Confocal Microscopes." For a review of applications of spinning disk microscopes in life science, see reference [9].

Chromatic Brightfield Confocal Microscopy

A chromatic confocal microscope operating in reflection deliberately introduces chromatic aberrations into the imaging system. Scanning in the vertical direction is then replaced by simultaneous detection of different spectral components which encode the depth information, because the depth of the focus depends on the wavelength. A complete stack of images can be acquired in this way in a single two-dimensional mechanical scan of the sample or of the excitation beam. Broadband excitation may be provided by a white light lamp or by a femtosecond laser-generated supercontinuum.

Structured Illumination Microscopy (SIM)

The SIM technique does not employ any pinhole and can be used with white light, but is related to confocal microscopy in its optical sectioning capability [10]. Optical sectioning is achieved by acquiring a sequence of images of the sample with structured illumination. The simplest case of structured illumination is thereby produced by placing a grid of fully transparent and fully opaque stripes of equal width (one half the period L of the grid) into the illumination path and projecting this grid onto the sample. Only sample

structures that are in focus will lead to significant variations of the image when the grid is displaced along the direction of periodicity, because only in focus the image of the grid within the sample is sharp. After acquiring three images with the grid displaced by 0, L/3, and 2 L/3, the optical section can be calculated (in the simplest version of the SIM algorithm) as the root mean square of the three differences between the three images. More sophisticated deconvolution algorithms are available and many other structures for illumination can be used. Movement of a grid illumination pattern across the sample may be replaced by the generation of arbitrary patterns by digital mirror devices (DMD) based on microelectromechanical systems (MEMS) or on spatial light intensity modulators (SLM), based on liquid crystals.

Whereas optical sectioning can be achieved more easily in this way than in a standard confocal system with a single-point focus, there are also some problems related to this approach. SIM works badly in strongly scattering samples, because small differences on a large background have to be determined. This is related to a significant loss in bit resolution and, hence, dynamic range in the final image. Moreover, the optical sectioning capability of SIM is slightly worse than for the standard single focus confocal microscope.

Another version of SIM employs either a grid pattern or random aperture arrays on a spinning disk, which both allow a large throughput of the light from the illumination source of the order of 50 %. Because of cross-talk between the light transmitted through neighboring apertures, the image acquired through the disk will contain a part that is not in focus. This part has to be subtracted from the image. A corresponding conventional image, to be subtracted from the partly confocal image, may be acquired by tilting the disk slightly, reflecting a second light source from the rear side of the disk toward the microscope objective and registering the corresponding image using a second camera. This procedure allows rapid optical sectioning and, hence, in vivo imaging of biological samples with very good signal-to-noise ratio with a comparatively simple instrument. A similar approach of subtracting the conventional image is based on a DMD, instead of a spinning disk, and was termed *programmable array microscope*.

Versions of structured illumination microscopy providing a moderate degree of super-resolution are based on Moiré patterns produced by projecting a high spatial frequency grid onto the sample (high-resolution SIM, HR-SIM). The Moiré pattern arises, because the observed signal is the product of the spatial distribution of the excitation with the concentration of the fluorophore and therefore contains spatial frequencies equal to differences between sample spatial frequencies and the one of the grid. The grid pattern must not only be shifted, but also be rotated, in order to be able to calculate the image. The method is able to increase the resolution by a factor of two beyond the diffraction limit.

Confocal Microscopy Beyond Brightfield and Standard Fluorescence

Confocal Raman Microscopy

In the Raman spectroscopy mode, the inelastically scattered light from the sample is detected, where the frequency shift toward lower (Stokes signal) or higher photon energy (anti-Stokes signal) coincides with an internal vibration of the sample. Raman scattering is typically very weak, and strong rejection of elastically scattered laser light is necessary. Historically, triple monochromators were often used for this purpose and still yield the highest spectral resolution, but in recent years dielectric long-pass filters with ultra-sharp transmission edges and high rejection factors have become available. This simplifies the instruments significantly, increases light efficiency, and in combination with slit-scanning techniques (see section "Slit-Scanning Confocal Microscopes") allows rapid multispectral confocal Raman imaging with acquisition times per pixel in the millisecond range.

Coherent anti-Stokes Raman scattering (CARS) and *stimulated Raman scattering (SRS)* microscopies are nonlinear variants of Raman

microscopy based on stimulated Raman scattering. Both techniques require two short-pulse lasers operating at different frequencies ν_1 and ν_2. The overlapping beams are focused by the microscope objective into the sample. When the difference in the optical frequencies is tuned to the frequency of a characteristic vibrational frequency ν_v of the sample ($\nu_1 - \nu_2 = \nu_v$), it will excite this vibration and at the same time generate anti-Stokes Raman scattered light at a frequency $2\nu_1 - \nu_2 = \nu_1 + \nu_v$ (CARS) or weakly deplete the pump and enhance the Stokes beams (SRS). Stimulated Raman scattering can be several orders of magnitude stronger than spontaneous Raman scattering and therefore allows rapid label-free imaging of a particular molecule in the sample. At least one of the two lasers necessary for CARS or SRS has to be tunable. This requirement maybe fulfilled, for example, by a Ti:sapphire laser or an optical parametric oscillator. Video rate in vivo SRS microscopy has been recently demonstrated and offers a number of advantages over CARS microscopy.

Another variant of confocal Raman microscopy, tip-enhanced Raman scattering (TERS) employs optical near fields of a sharp tip in order to increase the spatial resolution in Raman scattering over the diffraction limit. Confocal excitation and detection thereby reduce elastically scattered background in this setup.

Multiphoton Microscopy

Fluorescence excitation may in general be based on a linear process, in which a single photon excites the emitter into the excited state at energy E_1 from which it fluoresces, or on nonlinear processes, in which the emitter simultaneously absorbs n photons of energy E_1/n. Nonlinear processes are usually much less likely to occur than linear ones, with the probability strongly decreasing with the number of photons. Therefore, pulsed lasers are used for excitation, in which the optical power is concentrated in short pulses of widths in the femtosecond to picosecond range and the intensity during the pulse is very high. In addition, focusing the beam to a submicron spot leads, of course, to strong additional enhancement of the excitation probability.

Multiphoton microscopy has several important advantages over optical microscopy employing linear excitation. First, it is inherently confocal, in the sense that out-of-focus contributions to the detected signal are very small and therefore a confocal pinhole is not required (or the existing pinhole can be opened fully without losing the optical sectioning capability). This is advantageous, in particular for strongly scattering samples, because it yields a higher detection efficiency. Second, the photon energy used for excitation is only one half (for two-photon excitation) or one third (for three-photon excitation) of that used in the linear case, and the excitation wavelength therefore typically lies in the near infrared or even further in the infrared. Because the scattering in biological tissue and other inhomogeneous materials with inhomogeneities on a scale of the wavelength or below decreases strongly with the wavelength (proportional to λ^{-4} for very small inhomogeneities), the penetration depth is considerably larger for multiphoton excitation compared to single-photon excitation of the same fluorophore. This means that three-dimensional imaging deep into tissue becomes possible. Third, photo-induced damage to the sample and the fluorophore is reduced, also because of the lower photon energy.

Instead of making use of multiple photon excitation of a fluorescent chromophore, it is also possible in some samples to detect light due to second harmonic generation (SHG) and to use that for label-free imaging. Other label-free and, in addition, chemically specific, variants of multiphoton microscopy are CARS microscopy and SRS microscopy, as described in section "Confocal Raman Microscopy."

Fluorescence Lifetime Imaging (FLIM)

The lifetime of a fluorophore may vary depending on local pH, oxygen, or ion concentrations, or on intermolecular interactions, for example, fluorescence resonance energy transfer (FRET, see section "Fluorescence Resonance Energy Transfer

(FRET)"). On the other hand, within some limits, it does not respond to the intensity of the excitation light, the fluorophore concentration, or photobleaching. The fluorescence lifetime is therefore a useful physical quantity that can be used for imaging and quantitative analysis of local pH, ion concentrations, or intermolecular interactions. Fluorescence lifetimes are typically in the range of a few picoseconds, when dominated by non-radiative processes, to several tens of nanoseconds, when limited by the radiative transition rate. In some cases it is useful to employ fluorophores with very long lifetimes, in particular when strong autofluorescence is present. Performing a lifetime measurement at each position of the excitation laser focus within the sample and representing the corresponding values as a gray scale or color value from a lookup table yield a confocal lifetime image. At the same time, a standard fluorescence intensity image can be obtained.

A common method of measuring fluorescence lifetime is time-correlated single-photon counting (TCSPC). Here, a correlator, triggered by the short pulse of the exciting laser measures arrival times of fluorescence photons, typically detected by an avalanche photodiode with a short response time, and generates a histogram of delays. Fitting an exponential function to the histogram yields the fluorescence lifetime as the decay time of the exponential. Complications may arise when the decays are actually non-exponential. In the case of relatively long lifetimes in the nanosecond range, instead of TCSPC, a gated image intensifier may be used to measure the number of photons falling into a time window defined by the gate. Both techniques operate in the time domain. When the lifetime is comparatively long, it is also possible to modulate the laser pulses in the MHz range and detect the phase shift of the corresponding modulation in detected fluorescence intensity, which depends in a simple way on the fluorescence lifetime. TCSPC, however, is the most flexible way of measurement, because it is not restricted to long lifetimes and allows to analyze non-exponential decays as well.

Fluorescence Resonance Energy Transfer (FRET)

FRET [11] is a resonant non-radiative energy transfer from a donor to an acceptor fluorophore due to the dipolar interaction. As the dipolar interaction energy is proportional to the third power of the donor-acceptor distance R and because the probability for FRET to occur involves the square of an off-diagonal matrix element of the dipolar interaction, it falls off as R^6 and depends on the relative orientation between the molecules and the spectral overlap between the emission spectrum of the donor and the absorption spectrum of the acceptor. The probability is highest for parallel orientation of the donor and acceptor transition dipoles. Because of the steep fall, off FRET can only occur if R is sufficiently small, that means, if the donor and the acceptor are sufficiently close to each other. Because the critical distance is only 1–10 nm, typically 4–6 nm, in all practical cases, the FRET mechanism provides a *molecular ruler* for measuring the donor-acceptor distance and in this way provides an indirect mechanism for studying structure on the nanometer scale, which cannot directly be resolved by diffraction-limited optical microscopy (superresolution microscopy, compare section "Super-Resolution," might, however, in the future partly supersede FRET studies). This is of particular interest for protein-protein and intra-protein interactions. For this purpose, the proteins of interest have to be labeled by fluorophores with overlapping emission and absorption spectra. In many cases fluorescent proteins (compare section "Fluorescent Probes") are used for this purpose. A sensitive measure for FRET, which can be used for imaging, is the ratio of intensities of the donor and acceptor fluorescence peaks.

A problem with FRET confocal microscopy is that the signal-to-noise ratio is often very poor. Therefore, in many cases only the occurrence or absence of FRET is detected. The signal-to-noise ratio may be improved by measuring the donor lifetime, instead of the acceptor/donor fluorescence intensity ratio. This imaging option is usually known as FLIM-FRET (*fluorescence lifetime imaging-fluorescence resonance energy transfer*).

Fluorescence Recovery After Photobleaching (FRAP)

This microscopy technique, most often performed in a confocal setup, bleaches the fluorescence of a certain sample region, often an intracellular organelle, and registers the recovery of fluorescence due to diffusion of the fluorescently labeled molecules. It may therefore be used to investigate the mobility of the target molecule within the surrounding structures.

Fluorescence Correlation Spectroscopy (FCS)

In contrast to the previous techniques, FCS is usually used not as an imaging technique, but with a fixed position of the confocal volume within the sample. The laser (usually a continuous wave laser) thereby excites fluorescent particles within this volume, and particle movement in and out of the volume produces fluorescence intensity fluctuations. The autocorrelation function of these fluctuations provides information about the concentration, diffusion coefficient, and the mass of the particles. The diffusion coefficient depends on the viscosity of the medium via the Einstein-Smoluchowski relation. It may also depend on interactions of the particles with a microstructured environment.

A variant of FCS is fluorescence cross-correlation spectroscopy, in which the fluorescence intensity fluctuations of two fluorophores, labeling two different molecules, are measured simultaneously in two different channels. If the two molecules are bound in a dimer, the fluctuations will be highly correlated. Otherwise, no cross correlation is expected. The degree of cross correlation then is a measure of how many of the two different species of molecules are bound to each other.

Confocal Microscopy Sensitive to Phase

Interferometric versions of confocal laser scanning microscopy have been reported relatively early in the literature and were based on Mach-Zehnder or Michelson interferometers, for measurements in transmission or reflection, respectively [1]. In this way, it is possible to measure object topography or refractive index variations with interferometric precision. Interferometric confocal microscopes have the advantage over conventional interferometric microscopes that the shape of the wave front of the reference beam and are irrelevant because only phase and amplitude at the pinhole is important. This means that the requirement for matching optics, otherwise necessary in interferometric microscopy, is strongly relaxed.

Often two beam splitters and two detectors, each with its own pinhole, are employed in the detection beam path of interferometric confocal microscopes based on Michelson interferometers. This allows to separate the conventional confocal signal and the pure interference signal as the sum and difference of the outputs of the two detectors. This is based on the fact that two beams at the inputs of a symmetric lossless beam splitter are combined with a $+\pi/2$, $-\pi/2$ phase shift relative to each other at the two outputs of the beam splitter, respectively. A non-interferometric variant of phase-sensitive confocal microscopy, *differential phase-contrast confocal microscopy*, is obtained by omitting the mirror generating the reference beam in the Michelson interferometer and obscuring one half of each of the relay lenses in a complimentary manner [2].

The sensitivity of interferometric confocal microscopes may be enhanced by using a spectrally shifted reference beam (*heterodyne interferometric confocal microscopy*). Topographic resolution of about 0.01 nm using such a heterodyne interferometric confocal microscope has been reported. In addition, the optical sectioning capability of the confocal microscope and the corresponding dependence of the signal on defocus can be used to avoid phase unwrapping ambiguities.

Another variant of interferometric confocal microscopy is called *4π microscopy*. Here, two opposing objective lenses are used for excitation and/or detection. This increases the available numerical aperture and therefore enhances the resolution, in particular, in the axial direction. In addition, because of the coherent excitation, the

technique is inherently interferometric due to the interference of the two counterpropagating excitation beams in the focal region. This generates a standing wave interference pattern which modulates the main lobe of the focus in the axial direction and in this way allows to improve the axial resolution by a factor of about 4.5. The side lobes, due to interference, within the confocal main lobe may be effectively suppressed in the case of two-photon excitation. This suppression is due to the nonlinearity of the excitation process and can be additionally enhanced, if the sample fluorescence is also detected through both objective lenses in an interferometric setup. Because the position of maxima of the PSF for excitation and detection depends on the phase of both beams used for excitation and detection, precise control of the phase of both beams is required.

It should be mentioned here that a non-interferometric technique for phase measurement in optical microscopy is based on the so-called transport-of-intensity equation which can be derived from the paraxial time-dependent wave equation and relates the intensity and phase of a paraxial monochromatic wave to its longitudinal intensity derivative. The technique requires, however, the measurement of very small changes of intensity as a function of small defocus and therefore requires high bit resolution and long integration times.

Superresolution

Historically, the Abbe diffraction limit had been an insurmountable barrier in optical microscopy for many years. Many structures of interest, on the other hand, are significantly smaller than this limit and, particularly in biology, there has been a strong desire to significantly improve the resolution. In recent years, the diffraction limit has been broken in numerous ways, and this is based on some of the most exciting modern advancements in optics, which is still, although very old, a rapidly developing area of physics. The conceptually simplest way to achieve sub-Abbe resolution is by way of deconvolution (compare section "Deconvolution"). This does not, however, allow to achieve resolution enhancements of an order of magnitude, as urgently desired in many cases. One route to satisfy this demand employs non-propagating near fields. These near fields carry all optical information on length scales below the Abbe limit. The fact that these near fields are lost in far-field optical imaging can be viewed as the reason for the limited resolution of standard (including confocal) optical microscopes. Near-field related techniques (compare "▶ Scanning Near-Field Optical Microscopy") will not be discussed here (as an exception, compare TERS, section "Confocal Raman Microscopy"), although some of them may be combined with confocal imaging, for example, *solid immersion lens microscopy* (SIL) or *total internal reflection microscopy* (TIRF). Other routes, however, have opened the way to far-field optical nanoscopy in recent years. They are based on prior knowledge about the sample (PALM/STORM) or on optical nonlinearity (STED, SSIM) [12]. STED uses a specially designed optical excitation and is a scanning microscopy technique. PALM/STORM is non-confocal, because parallel detection using a CCD camera is used. These techniques have reached lateral resolution in the range of 20 nm, and ongoing research attempts to further push the resolution toward the molecular level (1–5 nm). Commercial instruments based on these techniques are increasingly becoming available [13]. The techniques will be shortly described, because of their potential as far-field microscopy techniques to partly supersede standard confocal fluorescence laser scanning microscopy.

Superresolution by Prior Knowledge (Profilometry, PALM/STORM)

In cases where it is known that one and only one sharp interface between two homogeneous materials or one and only one point emitter or reflector in the focus is present, the position of the interface in the z direction or the three-dimensional position of the point emitter/reflector can be determined

with a precision that is far beyond the Abbe limit and is limited only by the total amount of photons detected.

The simplest such technique consists in profiling a surface using a standard non-interferometric CLSM. By adjusting the z position to the wing of the $V(z)$ function, very slight changes in the topography of the sample surface can be measured, if the material and, hence, the reflectivity do not vary. Sub-nanometer depth resolution has been achieved in this way, averaged over the diffraction-limited lateral size of the confocal spot. As discussed in section "Confocal Microscopy Sensitive to Phase," the technique may be combined with interferometry to increase the depth resolution to about 0.01 nm.

In a principally similar way, *photoactivated localization* microscopy (PALM) and *stochastic optical reconstruction microscopy* (STORM) determine the position of single-point emitters with nanometric precision. Because it must be avoided to have more than one molecule in the focal volume and such a sparse distribution of fluorescent emitters would not allow sub-diffraction-limited resolution, it is necessary to separate the contributions from many individual molecules in some way. *Temporal* separation is key to current single molecule-based superresolution techniques. Other options, such as spectral separation, or more sophisticated schemes have also been discussed, however.

Current techniques use *photoactivable* molecules, which can be statistically turned on by another light source. Irreversible or reversible processes may be employed to turn off an activated subset. Determination of the positions of the activated molecules and multiple repetition then results in a software-generated image with nanometric resolution. Typically, thousands of images must be acquired and, correspondingly, imaging is slow, with a trade-off between spatial and temporal resolution. Effective frame rates of reconstructed images of about 1/(3 min) at a Nyquist-limited resolution of about 50 nm have been demonstrated for frames of 50 × 50 pixels. The obtainable resolution depends on the brightness of the fluorophores in the "on" state, as well as on the achievable contrast between "on" and "off" states. Fluorescent proteins as well as fluorescent dyes are currently in use [14].

PALM has hitherto mostly been used in total internal reflection configuration and is then restricted to essentially two-dimensional imaging. Lateral resolution down to ≈20 nm has been achieved. A three-dimensional STORM technique has achieved an image resolution of 20–30 nm in the lateral dimensions and 50–60 nm in the axial dimension in an illuminated sample volume of a few micrometers in thickness. An interferometric variant of PALM employing self-interference of fluorescent photons has improved the axial resolution to <20 nm for optically thin samples.

It may finally be worth noting that, whereas photoswitching is considered by some authors as a kind of optical nonlinearity, an optically nonlinear response is not essential to the breaking of the diffraction limit in the far field in the case of single molecule-based techniques. For example, sub-diffraction imaging due to binding and detachment of diffusing probes has been demonstrated, which obviously does not involve any optical nonlinearity. Moreover, photoswitching and temporal selection of molecules in general represent conceptually only one, although currently the standard and most successful option to ensure that only one molecule is detected within the PSF of the objective lens. It may therefore be said that in this case the breaking of the diffraction barrier is based on the knowledge, however obtained, to have only one molecule within the PSF and on the facts that the center-of-mass position of a diffraction-limited spot can be determined to a much higher precision than given by the size of this spot and that this position corresponds to the position of the fluorescent molecule.

Superresolution Due to Nonlinearity (SSIM, STED)

Some examples of confocal imaging methods employing optical nonlinearity have already

been discussed in sections "Confocal Raman Microscopy"(CARS, SRS), and "Multiphoton Microscopy." By reducing the size of the PSF for photons of the same wavelength, optical nonlinearity is in principle able to enhance the resolution. In multiphoton microscopy, however, because long wavelength photons are used for excitation of fluorophores, the resolution is actually worse than that of its linear counterpart.

Saturated structured illumination microscopy, SSIM, on the other hand, provides resolution enhancement without fundamental limit. This nonlinear variant of SIM (compare section "Structured Illumination Microscopy (SIM)") employs a nonlinear dependence of the fluorescence intensity on the excitation density, which is possible, for example, by saturation of the excited state. In this case, the effective illumination generally contains higher spatial frequency components leading to increased resolution. Optical resolution beyond 50 nm has been reported using this technique.

Stimulated emission depletion microscopy (STED) selectively de-excites fluorescing molecules within the PSF of the confocal microscope by stimulated emission, using a second laser. This laser beam passes a spiral phase mask, which generates a zero of intensity in the center of the beam cross section. Stimulated emission and corresponding deexcitation of fluorophores then occur everywhere within the PSF, except close to the axis of the beam, and the size of the region, from which fluorescence that can be collected depends on the intensity of the deexciting laser. Lateral resolution of about 20 nm has been reported using this setup. If STED is combined with a 4Pi setup, fluorescing spherical focal volumes of 40–45 nm diameter can be generated. STED is currently able to operate considerably faster than PALM. Video-rate STED microscopy with about 60 nm lateral resolution has been reported with a frame rate of 28 frames/s.

Cross-References

- ▶ Optical Techniques for Nanostructure Characterization
- ▶ Scanning Near-Field Optical Microscopy

C

References

1. Corle, R.C., Kino, G.S.: Confocal Scanning Optical Microscopy and Related Imaging Systems. Academic, San Diego (1996)
2. Wilson, T.: Confocal Microscopy. Academic, London (1990)
3. Born, M., Wolf, E.: Principles of Optics, 6th edn. Pergamon, Oxford (1980)
4. Sarder, P., Nehorai, A.: Deconvolution methods for 3-D fluorescence microscopy images. IEEE Signal Proc. Mag. **23**, 32–45 (2006)
5. Richards, B., Wolf, E.: Electromagnetic diffraction in optical systems. II. Structure of the image field in an aplanatic system. Proc. Roy. Soc. A **253**, 358–379 (1959)
6. Sheppard, C.J.R., Török, P.: Efficient calculation of electromagnetic diffraction in optical systems using a multipole expansion. J. Mod. Opt. **44**, 803–818 (1997)
7. Chudakov, D.M., Matz, M.V., Lukyanov, S., Lukyanov, K.A.: Fluorescent proteins and their applications in imaging living cells and tissues. Physiol. Rev. **90**, 1103–1163 (2010)
8. Ormo, M., Cubitt, A.B.., Kallio, K., Gross, L.A., Tsien, R.Y., Remington, S.J.: Crystal structure of the Aequorea victoria green fluorescent protein. Science **273**, 1392–1395 (1996). Image from the RCSB PDB (www.pdb.org) of PDB ID 1EMA (http://dx.doi.org/%2010.1021/ja1010652)
9. Gräf, R., Rietdorf, J., Zimmermann, T.: Live cell spinning disk microscopy. Adv. Biochem. Eng. Biotechnol. **95**, 57–75 (2005)
10. Langhorst, M.F., Schaffer, J., Goetze, B.: Structure brings clarity: structured illumination microscopy in cell biology. Biotechnol. J. **4**, 858–865 (2009)
11. Piston, D.W., Kremers, G.J.: Fluorescent protein FRET: the good, the bad and the ugly. Trends Biochem. Sci. **32**, 407–414 (2007)
12. Hell, S.W.: Far-field optical nanoscopy. Science **316**, 1153–1158 (2007)
13. Chi, K.R.: Ever-increasing resolution. Nature **462**, 675–678 (2009)
14. Heilemann, M., Dedecker, P., Hofkens, J., Sauer, M.: Photoswitches: key molecules for subdiffraction-resolution fluorescence imaging and molecular quantification. Laser Photon. Rev. **3**, 180–202 (2009)

Confocal Scanning Optical Microscopy (CSOM)

▶ Confocal Laser Scanning Microscopy

Conformal Electronics

▶ Flexible Electronics

Contact Angle

▶ Surface Engineering, Tailored Wettability, and Applications

Contour-Mode Resonators

▶ Laterally Vibrating Piezoelectric Resonators

Contraception

▶ Use of Nanotechnology in Pregnancy

Contrast Enhancement

▶ Magnetic Nanoparticles for Biomedical Applications

Cooling of Electronic Components

▶ Microchannel Flows as Heat Sinks

Core/Shell Nanostructures

▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods

Coupled Mode Theory

▶ Harnessing Disorder at the Nanoscale

Coupling

▶ Nonlinear and Parametric NEMS Resonators

Coupling Clamp

▶ Dynamic Clamp

CPMD

▶ Car–Parrinello Molecular Dynamics

Creep

► Nanomechanical Properties of Nanostructures

Crystallite

► Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Cuticle

► Arthropod Strain Sensors

Cylindrical Gold Nanoparticles

► Gold Nanorods