The application of systems biology to drug discovery

Carolyn R Cho[1], Mark Labow[2], Mischa Reinhardt[3], Jan van Oostrum[4] and Manuel C Peitsch[4]

Recent advances in the 'omics' technologies, scientific computing and mathematical modeling of biological processes have started to fundamentally impact the way we approach drug discovery. Recent years have witnessed the development of genome-scale functional screens, large collections of reagents, protein microarrays, databases and algorithms for data and text mining. Taken together, they enable the unprecedented descriptions of complex biological systems, which are testable by mathematical modeling and simulation. While the methods and tools are advancing, it is their iterative and combinatorial application that defines the systems biology approach.

Addresses
[1] Department of Systems Biology, Genome and Proteome Sciences, Novartis Institutes of BioMedical Research, Cambridge MA 02139, USA
[2] Department of Platform Chemistry and Biology, Genome and Proteome Sciences, Novartis Institutes of BioMedical Research, Cambridge MA 02139, USA
[3] Department of Scientific Computing and Statistics, Genome and Proteome Sciences, Novartis Institutes of BioMedical Research, Novartis AG, CH-4002 Basel, Switzerland
[4] Department of Systems Biology, Genome and Proteome Sciences, Novartis Institutes of BioMedical Research, Novartis AG, CH-4002 Basel, Switzerland

Corresponding author: Peitsch, Manuel C (manuel.peitsch@novartis.com)

Current Opinion in Chemical Biology 2006, **10**:294–302

This review comes from a themed issue on
Next-generation therapeutics
Edited by Clifton E Barry III and Alex Matter

Available online 5th July 2006

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2006.06.025

Introduction

Drug discovery is a complex undertaking facing many challenges [1•], not the least of which is a high attrition rate as many promising candidates prove ineffective or toxic in the clinic owing to a poor understanding of the diseases, and thus the biological systems, they target. Therefore, it is broadly agreed that to increase the productivity of drug discovery one needs a far deeper understanding of the molecular mechanisms of diseases, taking into account the full biological context of the drug target and moving beyond individual genes and proteins [2–5]. Systems biology, and especially the elucidation and dynamic analysis of cellular signaling pathways, provides a new grammar [2], or framework, for drug discovery.

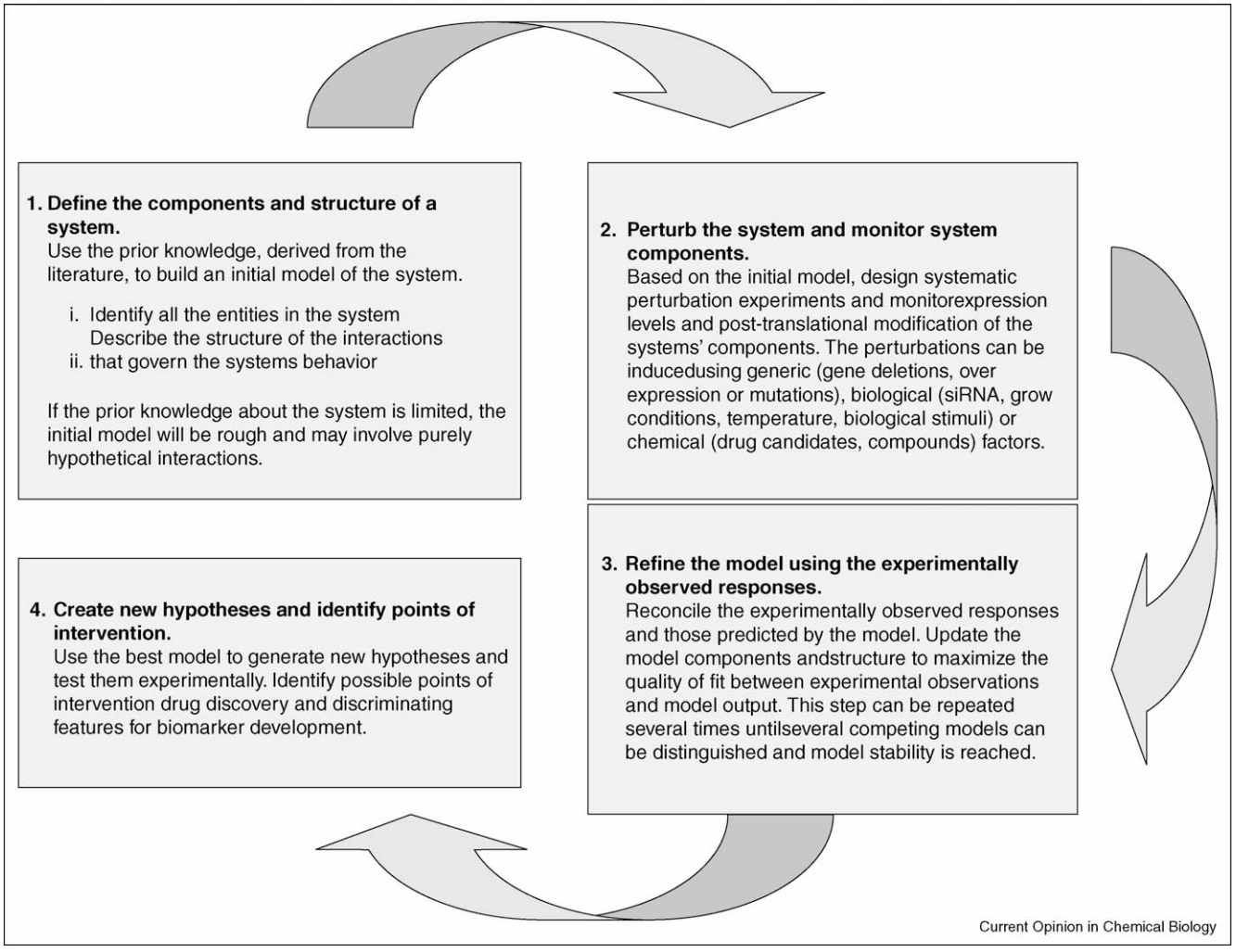
Systems biology is the 'systematic' interrogation of the biological processes within the complex, physiological milieu in which they function. Insight into the combined behavior of these many, diverse, interacting components is achieved through the integration of experimental, mathematical and computational sciences in an iterative approach (Figure 1). Through this contextual understanding of the molecular mechanisms of disease, a systems approach has the potential to further facilitate the identification and validation of the therapeutic modulation of regulatory and metabolic networks and hence help identify targets and biomarkers, as well as 'off-target' and side effects of drug candidates [3–5].

Here, we focus on selected recent advances in the disciplines of systems biology (Box 1) that are relevant to drug discovery.

Experimental methods

Experimental approaches in systems biology are generally aimed at identifying the components of a system and their interactions, and monitoring the effect of perturbations on these components. Recent advances in proteomics, genomics and metabolomics [6,7] and their integration [8] are radically transforming the drug discovery process. For instance, the identification of protein network components and the characterization of their post-translational modifications has recently reached new levels of scale and complexity as exemplified by the analysis of the insulin receptor substrate 1 serine/threonine phosphorylation sites [9] and the interactome analysis of the human TNFα/NFκB network members [10], and the ErbB/EGF receptors [11,12]. This has been enabled not only by the rapidly maturing MS-based proteomics methods [9–11] but also by protein arrays [12], which have much progressed over the past few years [13,14,15•,16]. Although protein forward arrays have been successfully applied in a number of settings [12–14,15•], including to the profiling receptor tyrosine kinase activation [14], reverse protein arrays are arguably the most broadly applicable technology. They are based on the principle that complex protein mixtures (such as a cell lysates) are spotted in an array format and probed with selected antibodies in a multiplexed manner. Reverse arrays have been used to analyze cell lines for potential biomarkers [17], profile molecular pathways in cells excised from cancer tissue by laser capture micro-dissection [15•,18,19], and detect auto-antibodies in serum

Figure 1



Process diagram of systems biology.

[16,20•], which opens new opportunities for their application to screening in a clinical setting [20•]. Furthermore, reverse arrays are particularly well suited to monitor dynamic network responses (Figure 2) to compound-induced perturbations. This compound-based systems response profiling [5], or structure pathway activity relationship (Figure 3) has the potential to provide early indications about possible off-target and toxic effects of drug candidates. The availability of highly specific antibodies is, however, a prerequisite for reverse arrays. Whereas a reasonable fraction of commercially available antibodies appears to be suitable following validation, the generalization of reverse arrays will depend on the generation of broad collections of high-quality antibodies [15•]. For instance, the Human Protein Atlas Initiative (http://www.proteinatlas.org), part of the Human Antibody Initiative of the Human Proteome Organization (HUPO), has started a systematic approach to the generation and validation of antibodies [21,22]. It is, however, noteworthy that molecular recognition is not limited to immunoglobulin domains, and proteomics may soon benefit from newly engineered protein-binding domains such as designed ankyrin-repeat proteins (DARPins) [23]. While protein arrays provide data for the average content of a sample (e.g., cell lysates), recent developments in flow cytometry-based single-cell proteomics can further complement these technologies in that a limited number of signaling events and surface markers can be measured simultaneously in the same cell and hence enable discrimination between various cell populations in rare samples and biopsies [24].

Beyond monitoring the effect of perturbations on predefined network components, systems biology relies on the systematic analysis of gene function in signaling pathways and cellular processes. This has recently been

Box 1 Systems biology is interdisciplinary.

Experimental sciences: Direct systemic interrogation. Large scale genomics, proteomics and metabolite measurements are used to monitor gene translation, protein expression, signaling events and metabolite fluxes induced by the systematic perturbation of a biological system by biological, genetic or chemical factors. Furthermore, large-scale experimental methods are used to identify the nodes of signaling networks, through the comprehensive identification of interaction partners and protein modifications (e.g., phosphorylations). The comparison of samples from a disease state with those of matched normal donors and the creation of animal models of disease (through the knock-out of selected genes and/or the introduction of specific mutations) represents a well established approach to study a system through the analysis of naturally occurring perturbations.

Data analysis and pathway informatics: Analysis and understanding of data in the context of larger systems. Statistical and computational methods are used to analyze and interpret large experimental data sets with the aim to identify molecular species significantly affected by the experimental conditions (perturbations). Databases and software tools are used to store, manage and visualize expression data in the context of cellular network and pathway information. This information is assembled from literature by hand or using mining techniques.

Literature mining: Discovery, extraction and synthesis of current knowledge. Entity recognition (identifying substances), information extraction (identifying relationships between biological entities) and natural language processing (combination of syntax and semantics analysis to extract relationships from complex sentences) are used to extract known facts such as protein–protein interactions, protein phosphorylation, regulatory relationships between molecular entities and genotype–phenotype relationships from the scientific literature. Text mining enables the inference of relationships between extracted entities and facts that are not formally enunciated in a sentence.

Mathematical modeling: Extrapolation and prediction to test understanding. Mathematical modeling and simulation is used to identify assumptions (iterating with literature mining) and gaps in understanding (iterating with additional analysis or data), and to generate new and experimentally verifiable hypotheses, thereby closing the iteration loop.

made possible by the development of cell-based genome-scale approaches, where over 20 000 individual genes can be interrogated in highly multiplexed experiment. These technologies aim to quantify the effects of either the overexpression of individual proteins using full length cDNAs [25,26] (Figure 4), or the inhibition of gene expression by RNAi molecules [27–29] (Figure 4). They are enabled by the creation of genome-scale collections of reagents [25,30,31••] and their optimization through computational approaches [32]. These genome-scale screens produce systems-level activity data for each gene, providing rich databases that can be mined to predict the physiological and biochemical function of individual proteins as well as the regulatory networks and pathways underpinning distinct phenotypes. In one early set of experiments performed in mammalian cells, data from several cDNA overexpression screens were used to predict the biochemical activity of a gene of previously unknown function [25]. In a later example, this approach was applied to identify genes and pathways capable of inducing nuclear TORC accumulation [26]. The genome-scale RNAi-based screen was developed in Drosophila melanogaster cell cultures and used to examine the role of 21 000 genes in cell growth [33]. This approach has since been applied to a number of signal transduction and cell-based processes in *D. melanogaster* and mammalian cell systems [34–36] as well as *in vivo* using the nematode *Caenorhabditis elegans* [34,37]. Furthermore, similar approaches have recently been used to globally assess the role of non-coding RNAs on pathway function in mammalian cells [38]. Although high-throughput functional systems level analysis provides quantitative data about gene activity and functional interactions, it expected that new technologies and areas of investigation, such as microRNA biology [39,40], will emerge and further complement our understanding of biological systems.

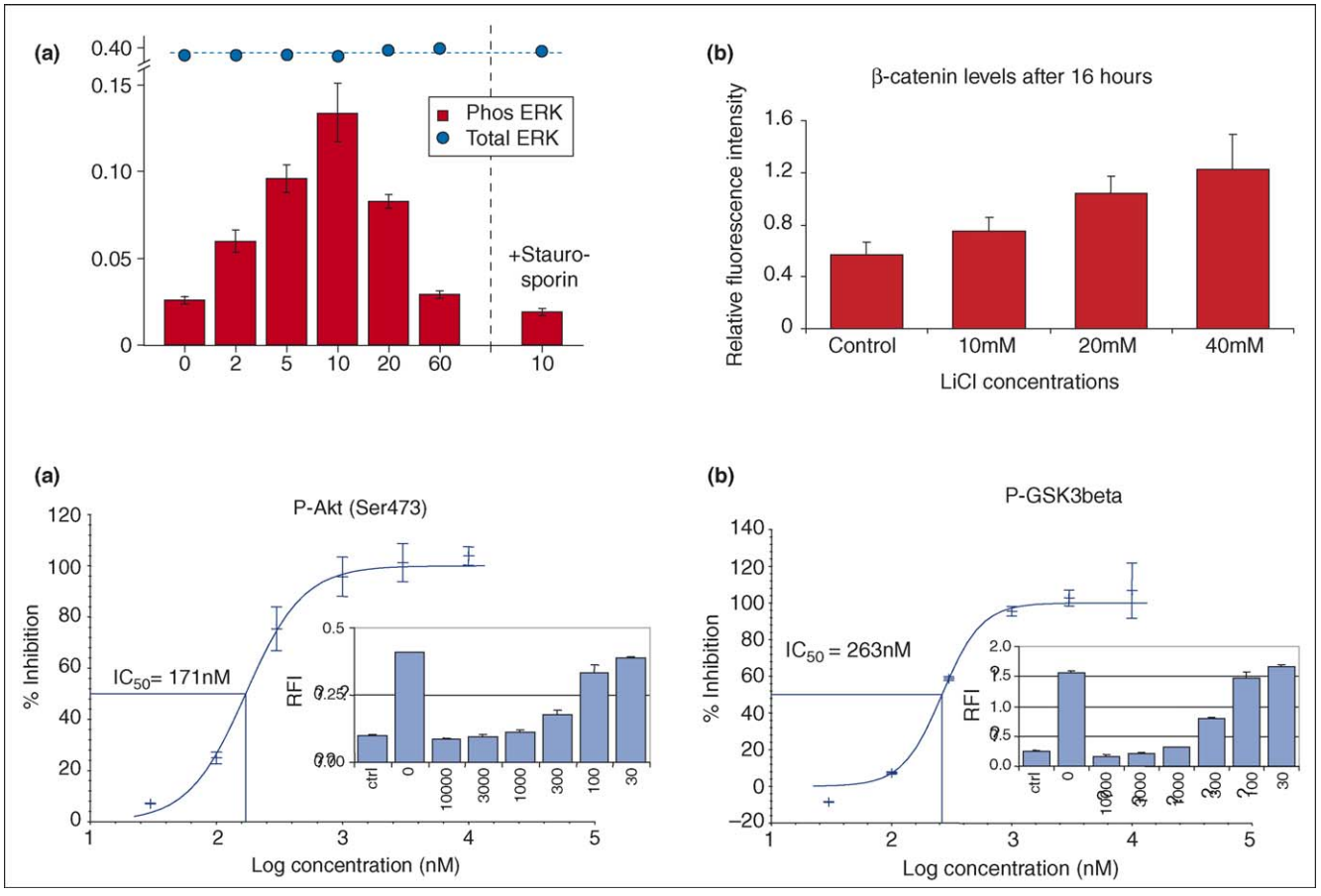
Data mining and pathway informatics

The evolution of these genomic and proteomic methods has necessitated the development of new algorithms to analyze the resulting data in the context of drug discovery [41]. In particular, integrating data from different experiments is a challenge that is being successfully addressed. For example, a Bayesian inference of sub-networks from a set of 300 microarray experiments has been used to uncover a number of pathways [42], and methods have been developed to overlay gene expression data with genome-wide transcription factor location data obtained by ChIP-on-Chip experiments [43], leading to the identification of previously unknown regulatory networks using data obtained from rapamycin treated *S. cerevisiae* [43]. Besides methods that infer pathways and regulatory networks, there is a growing number (>150) of databases [3] — of which KEGG [44] (http://www.genome.jp/keg/kegg2.html) is probably the best known — that collect such information from scientific publications using literature mining and manual curation. Such databases can be used to overlay functional and pathway information onto rank-ordered gene lists derived from differential expression experiments [45••]. This approach — called gene set enrichment analysis (GSEA) — removes the undue bias of selecting individual up-regulated genes by focusing on entire sets of genes [45••]. GSEA led to the discovery of several other disease-relevant pathways including in cancer [46], Huntington's disease [47] and myoblast differentiation [48], with potential implication for drug and biomarker discovery. By further combining signature-based predictions across several pathways, one can identify coordinated patterns of pathway deregulations. Such patterns were shown to distinguish specific cancers and tumor subtypes [49•] and to reflect the biology and outcome of specific cancers. Furthermore, in cell lines, these patterns predict their sensitivity to therapeutic agents [49•] and may help reposition or extend the application of existing drugs.

Literature mining

The scientific literature (which includes patents) is where the key knowledge and facts relevant to systems biology

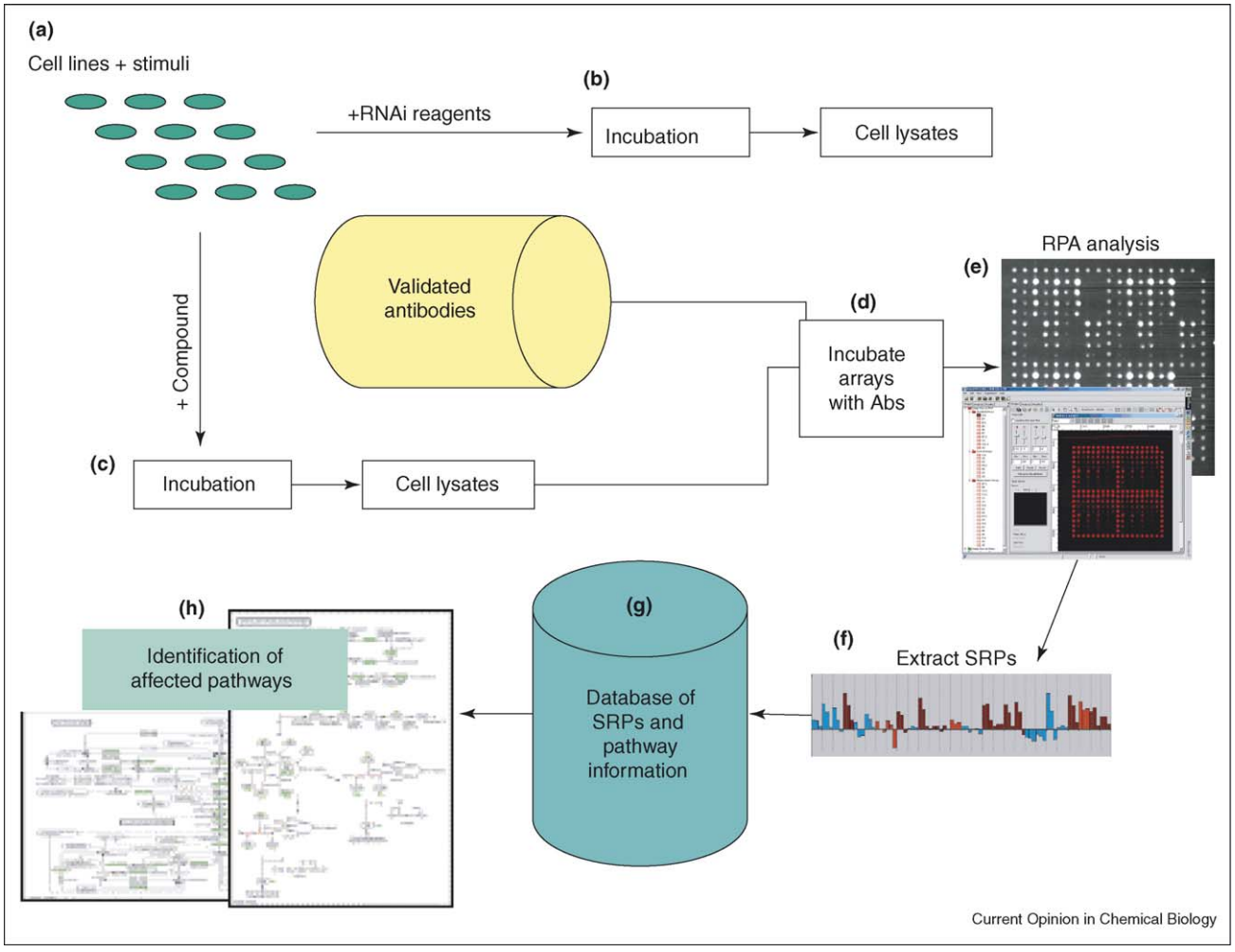
Figure 2



Sample applications of reverse protein arrays. **(a)** Monitoring phosphorylation events: Jurkat cells were treated with OKT3 and α-CD28 antibodies for the indicated period (x-axis), lysed and analyzed on reverse protein arrays using an α-ERK (blue circles) and an α-phospho-ERK antibody (red bars). This demonstrates that transient and rapid phosphorylation events can be measured using reverse protein arrays. **(b)** Monitoring other signaling events: untreated and LiCl-treated Jurkat cells were incubated for 16 h with the indicated concentrations of LiCl, which mimics Wnt pathway/β-catenin signaling by inhibiting GSK3β. Following lysis of the cells, the β-catenin levels were assessed using a specific antibody. The inhibition of GSK3β-mediated phosphorylation of β-catenin blocks its degradation, leading to the observed accumulation of β-catenin. **(c,d)** Monitoring the downstream effect of cell signaling inhibitors. Starved A431 cells were stimulated with insulin and co-treated with increasing concentrations of an inhibitor of the IGF1- receptor tyrosine kinase. After 30 min of treatment, the cells were lysed and the phosphorylation levels of Akt and GSK3β were monitored with antibodies specific for (c) Ser473-Phospho-Akt and (d) Ser9-phospho-GSK3β. By plotting the percent inhibition versus inhibitor concentration, one can derive IC_{50}-like data from such experiments. RFI is the relative fluorescence intensity.

are stored and reported [50•]. This resource is, however, growing and diversifying at a staggering pace. As a consequence, computational tools designed to efficiently extract entities and their relationships (biological facts) will play a pivotal role in systems biology [51,52•]. Indeed, model building starts with the identification of the components of a system and how they interact (Figure 1), facts that are then formalized in a diagram from which a mathematical model will be derived [53]. The best approach to identify protein and gene entities in text is to use a carefully curated list of synonyms [54] and recently developed methods for synonym extraction [55] and terminology disambiguation [56–58]. The consecutive extraction of the interactions between these entities relies on entity co-occurrence analysis [59–61] and natural language processing [62–67]. These methods have been successfully applied to the reconstruction of networks [67,68•] relevant to drug discovery and to the analysis of biological data [68•,69•] and bioinformatics databases [70,71•] in the context of literature-derived information and networks. Furthermore, combining text mining and statistical approaches enables scientists to extract significant information about research trends, emerging fields from patents [72], and infer (potential) new pathway/target-disease relationships from the scientific literature [67,73]. Recent advances in high-performance GRID-based text mining [74] of full text scientific articles [75] opens new doors for the discovery of scientific

Figure 3



Possible application of reverse protein arrays in systems biology: structure pathway activity relationship (SPAR). **(a)** Selected cell lines are treated with appropriate combinations of activating stimuli and treated with either **(b)** si/shRNA **(c)** or test compounds. Treated cells are sampled in a time-dependent manner and lysed before being spotted on reverse protein arrays. The arrays are **(d)** incubated with pre-defined antibodies and **(e)** measurements are taken. The systems response profiles (SRPs) are deduced from **(f)** the fluorescence intensities and **(g)** stored in a database along with pathway information. The treatment with siRNAs allows identification of SRPs caused by well-targeted network perturbations, which can serve as the reference set against which SRPs caused by drug candidates can be compared. **(h)** Thereby off-target effects can be deduced.

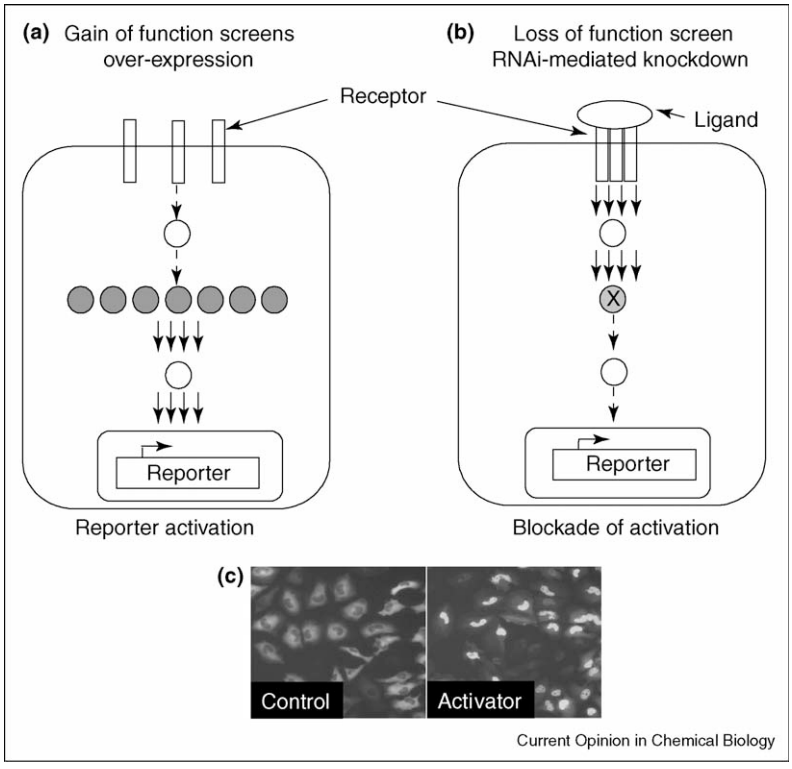
facts and interpretations so fundamental to systems biology. Finally, integrating text with biological, chemical and clinical data through text mining and entity-associated rules will enable systems biologists to quickly navigate between scientific data domains that are currently kept in disconnected databases [76].

Mathematical modeling

The advances in experimental approaches and in data and literature mining have also accelerated progress in the development and application of modeling approaches [53,77]. The most widely applied modeling method is the deterministic biochemical reaction description. The formalism, analysis and application that has been reviewed extensively [78•–80•] has matured to the extent that an annotation standard has begun to emerge [50•,81]. Emerging graphical ontology standards [82,83] will greatly aid in harmonizing the academic and commercial software tools that are available. In drug discovery, this modeling method has been successfully applied in pharmacokinetics/pharmacodynamics and dose-response modeling [84,85], and wider application is anticipated (http://www.fda.gov/oc/initiatives/criticalpath/stanski/stanski.html).

One drawback of the deterministic reaction approach is its lack of scalability. Genomic and proteomic approaches are aimed at identifying signaling networks of tens or more of molecules. The size, range of reaction parameters over many orders of magnitude, and extent of unknowns make these models intractable to computational methods.

Figure 4



Large-scale genetic screens. Large curated collections of gene-specific cDNAs and RNAi enable highly multiplexed and systematic gain and loss of function screens. The diagram outlines a hypothetical signal transduction pathway, initiated by a ligand-dependent membrane bound receptor, which after activation by a ligand leads to the activation of a specific signal transduction pathway. This pathway culminates in a downstream event such as transcriptional induction, which can be monitored directly or, as most often used in high-throughout genetic screens, via an enzymatic or fluorescent reporter gene. **(a)** Gain of function screens are performed by selectively over-expressing a critical pathway component (grey circles) and are expected to drive the equilibrium of a pathway forward, thus activating the reporter. Conversely, **(b)** loss of function screens are based on the elimination of a critical component (**X**) with an RNAi reagent that will prevent the activation of the pathway by a stimulus and thus the reporter gene. The phenotypic response, such as the nuclear transport of regulatory molecules, can be quantified using **(c)** a microscopy-based method. In this experiment, the effects of over 7000 individual genes on the nuclear import of the TORC1 CREB co-activator were quantified. This allowed the identification of a variety of genes and pathways capable of inducing TORC1 translocation and CREB activation [21].

New methods such as combinatorial reaction generation [86], and linear programming [87,88] address this need by providing automated methods for handling the complexity of large chemical reaction networks.

Other methods are addressing the more fundamental issue of the limitation of the deterministic approximation, itself. Stochastic representations account for the effects of small populations [89]. Rule-based computational approaches move completely away from a deterministic description [90–92]. These methods provide a molecule-centric description and can therefore be a natural bridge between data-driven inference and predictive models [93], as are functional inference methods [94]. These methods are highly scalable and easy to simulate; however, it is an open question as to how well they will be able to address questions involving complex non-linear dynamics.

Finally, there are many systems biology models that do not resemble reaction networks (for a range of examples visit http://www.cellml.org). In drug development, the most successful of these models are cardiac electrophysiology (EP) models that have been applied to safety assessment [95] and, most recently, to provide mechanistic rationale supporting the January 2006 US FDA approval of ranolazine for chronic angina [96] (http://www.fda.gov/bbs/topics/news/2006/NEW01306.html).

Conclusions

Although the methods and tools are advancing within each discipline, it is their iterative, combinatorial application that defines the systems biology approach. We believe that the discovery and understanding of complex disease mechanisms and therapeutic modalities will increasingly require this approach. This will have a profound impact on the systematic creation of large

collections of reagents (such as antibodies, RNAi and cDNA), detection methods and laboratory technology, computer science and organizational design. Indeed, a more widespread collaboration between mathematicians, computer scientists, physicians and experimental scientists will probably improve drug discovery in the next decade. This will certainly impact research organizations and the skills needed in research teams, and as a consequence calls for a broader scientific education of drug discovery scientists, bridging the classical disciplines of biology, mathematics, chemistry and medicine in an unprecedented manner. In these ways, systems biology promises to impact drug discovery significantly and improve the success rate of discovering crucial medicines.

Acknowledgements

We thank Mark S Boguski and Alan Buckler for their input and support during the writing of this review.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Schmid EF, Smith DA: **Keynote review: is declining innovation in the pharmaceutical industry a myth?** *Drug Discov Today* 2005, **10**:1031-1039.
This analysis, based on FDA data from 1945 to 2004, shows that the increasing R&D costs are not due to a decline in innovation of the pharmaceutical industry.

2. Fishman MC, Porter JA: **A new grammar for drug discovery**. *Nature* 2005, **437**:491-493.

3. Butcher EC, Berg EL, Kunkel EJ: **System biology in drug discovery**. *Nat Biotechnol* 2004, **22**:1253-1259.

4. Apic G, Ignjatovic T, Boyer S, Russell RB: **Illuminating drug discovery with biological pathways**. *FEBS Lett* 2005, **579**:1872-1877.

5. van der Greef J, McBurney RN: **Rescuing drug discovery: in vivo systems pathology and systems pharmacology**. *Nat Rev Drug Discov* 2005, **4**:961-967.

6. Lindon JC, Holmes E, Nicholson JK: **Metabonomics: systems biology in pharmaceutical research and development**. *Curr Opin Mol Ther* 2004, **6**:265-272.

7. Morris M, Watkins SM: **Focused metabolomic profiling in the drug development process: advances from lipid profiling**. *Curr Opin Chem Biol* 2005, **9**:407-412.

8. Saghatelian A, Cravatt BJ: **Global strategies to integrate the proteome and metabolome**. *Curr Opin Chem Biol* 2005, **9**:62-68.

9. Luo M, Reyna S, Wang L, Yi Z, Carroll C, Dong LQ, Langlais P, Weintraub ST, Mandarino LJ: **Identification of insulin receptor substrate 1 serine/threonine phosphorylation sites using mass spectrometry analysis: regulatory role of serine 1223**. *Endocrinology* 2005, **146**:4410-4416.

10. Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard V, Gagneur J, Ghidelli S *et al.*: **A physical and functional map of the human TNF-α/NF-κB signal transduction pathway**. *Nat Cell Biol* 2004, **6**:97-105.

11. Schulze WX, Deng L, Mann M: **Phosphotyrosine interactome of the ErbB-receptor kinase family**. *Mol Systems Biol* 2005, **1**:42-54.

12. Jones RB, Gordus A, Krall JA, MacBeath G: **A quantitative protein interaction network for the ErbB receptors using protein microarrays**. *Nature* 2006, **439**:168-174.

13. Hultschig C, Kreutzberger J, Seitz H, Konthur Z, Büssow K, Lehrach H: **Recent advances of protein microarrays**. *Curr Opin Chem Biol* 2006, **10**:4-10.

14. Nielsen UB, Cardone MH, Sinskey AJ, MacBeath G, Sorger PK: **Profiling receptor tyrosine kinase activation by using Ab microarrays**. *Proc Natl Acad Sci USA* 2003, **100**:9330-9335.

15. • Gulmann C, Sheehan KM, Kay EW, Liotta LA, Petricoin EF: **Array-based proteomics: mapping of protein circuitries for diagnostics, prognostics, and therapy guidance in cancer**. *J Pathol* 2006, **208**:595-606.
The authors review forward-phase and reverse-phase protein arrays in a number of applications.

16. Balboni I, Chan SM, Kattah M, Tenenbaum JD, Butte AJ, Utz PJ: **Multiplexed protein array platforms for analysis of autoimmune diseases**. *Annu Rev Immunol* 2006, **24**:391-418.

17. Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kouros-Mehr H, Bussey KJ, Lee JK, Espina V *et al.*: **Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays**. *Proc Natl Acad Sci USA* 2003, **100**:14229-14234.

18. Wulfkuhle JD, Aquino JA, Calvert VS, Fishman DA, Coukos G, Liotta LA, Petricoin EF: **Signal pathway profiling of ovarian cancer from human tissue specimens using reverse-phase protein microarrays**. *Proteomics* 2003, **3**:2085-2090.

19. Sheehan KM *et al.*: **Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma**. *Mol Cell Proteomics* 2005, **4**:346-355.

20. • Janzi M, Oedling J, Pan-Hammarström Q, Sundberg M, Lundeberg J, Uhlén M, Hammarström L, Nilsson P: **Serum microarrays for large scale screening of protein levels**. *Mol Cell Proteomics* 2005, **4**:1942-1947.
The authors analyzed over 2000 serum samples using reverse arrays and compared the data with clinical routine measurements. The results suggest that reverse protein arrays can be used to screen clinical serum samples.

21. Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C, Berglund L, Bergström K, Brumer H, Cerjan D *et al.*: **A human protein atlas for normal and cancer tissues based on antibody proteomics**. *Mol Cell Proteomics* 2005, **4**:1920-1932.

22. Nilsson P, Paavilainen L, Larsson K, Odling J, Sundberg M, Andersson AC, Kampf C, Persson A, Al-Khalili Szigyarto C, Ottosson J *et al.*: **Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling**. *Proteomics* 2005, **5**:4327-4337.

23. Binz HK, Amstutz P, Plückthun A: **Engineering novel binding proteins from nonimmunoglobin domains**. *Nat Biotechnol* 2005, **23**:1257-1268.

24. Irish JM, Kotecha N, Nolan GP: **Mapping normal and cancer cell signaling networks: towards single-cell proteomics**. *Nat Rev Cancer* 2006, **6**:146-155.

25. Iourgenko V, Zhang W, Mickanin C, Daly I, Jiang C, Hexham JM, Orth AP, Miraglia L, Meltzer J, Garza D *et al.*: **Identification of a family of cAMP response element-binding protein co-activators by genome-scale functional analysis in mammalian cells**. *Proc Natl Acad Sci USA* 2003, **100**:12147-12152.

26. Bittinger MA, McWhinnie E, Meltzer J, Iourgenko V, Latario B, Liu X, Chen CH, Song C, Garza D, Labow M: **Activation of cAMP response element-mediated gene expression by regulated nuclear transport of TORC proteins**. *Curr Biol* 2004, **14**:2156-2161.

27. Ito M, Kawano K, Miyagishi M, Taira K: **Genome-wide application of RNAi to the discovery of potential drug targets**. *FEBS Lett* 2005, **579**:5988-5995.

28. Wheeler DB, Carpenter AE, Sabatini DM: **Cell microarrays and RNA interference chip away at gene function**. *Nat Genet* 2005, **37(Suppl)**:S25-S30.

29. Bailey SN, Ali SM, Carpenter AE, Higgins CO, Sabatini DM: **Microarrays of lentiviruses for gene function screens in immortalized and primary cells**. *Nat Methods* 2006, **3**:117-122.

30. Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K *et al.*: **A resource for large-scale RNA-interference-based screens in mammals**. *Nature* 2004, **428**:427-431.

31. •• Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK *et al.*: **A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen**. *Cell* 2006, **124**:1283-1298.
Describes a library of 22 000 RNAi reagents that can be used in loss of function screens in human and murine systems.

32. Huesken D, Lange J, Mickanin C, Weiler J, Asselbergs F, Warner J, Meloon B, Engel S, Rosenberg A, Cohen D *et al.*: **Design of a genome-wide siRNA library using an artificial neural network**. *Nat Biotechnol* 2005, **23**:995-1001.

33. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Paro R, Perrimon N: **Heidelberg fly array consortium: genome-wide RNAi analysis of growth and viability in drosophila cells**. *Science* 2004, **303**:832-835.

34. Carpenter AE, Sabatini DM: **Systematic genome-wide screens of gene function**. *Nat Rev Genet* 2004, **5**:11-22.

35. Dasgupta R, Perrimon N: **Using RNAi to catch Drosophila genes in a web of interactions: insights into cancer research**. *Oncogene* 2004, **23**:8359-8365.

36. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B *et al.*: **A large-scale RNAi screen in human cells identifies new components of the p53 pathway**. *Nature* 2004, **428**:431-437.

37. Sieburth D, Ch'ng Q, Dybbs M, Tavazoie M, Kennedy S, Wang D, Dupuy D, Rual JF, Hill DE, Vidal M *et al.*: **Systematic analysis of genes required for synapse structure and function**. *Nature* 2005, **436**:510-517.

38. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG: **A strategy for probing the function of noncoding RNAs finds a repressor of NFAT**. *Science* 2005, **309**:1570-1573.

39. Bentwich I: **Prediction and validation of microRNAs and their targets**. *FEBS Lett* 2005, **579**:5904-5910.

40. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E *et al.*: **Identification of hundreds of conserved and nonconserved human microRNAs**. *Nat Genet* 2005, **37**:766-770.

41. Allison DB, Cui XQ, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nat Genet* 2006, **7**:55-65.

42. Pe'er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles**. *Bioinformatics* 2001, **17**:S215-S224.

43. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks**. *Nat Biotechnol* 2003, **21**:1337-1342.

44. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res* 2006, **34**:D354-D357.

45. •• Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al.*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**:267-273.
Describes gene set enrichment analysis and how this method detects modest but coordinated changes in the expression of groups of functionally related genes. The authors demonstrate how this method can be applied to the analysis of expression data obtained from biopsies.

46. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci USA* 2005, **102**:15545-15550.

47. Hodges A, Strand AD, Aragaki AK, Kuhn A, Sengstag T, Hughes G, Elliston LA, Hartog C, Goldstein DR, Thu D *et al.*: **Regional and cellular gene expression changes in human Huntington's disease brain**. *Hum Mol Genet* 2006, **15**:965-977.

48. Szustakowski JD, Lee JH, Marrese CA, Kosinski PA, Nimala NR, Kemp DM: **Identification of novel pathway regulation during myogenic differentiation**. *Genomics* 2006, **87**:129-138.

49. • Bild AH, Yao G, Chang JT, Wang QL, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A *et al.*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**. *Nature* 2006, **439**:353-357.
Demonstrates how the combination of patterns of pathway deregulations distinguishes between cancer types, reflects their biology and outcome and how they are predictive for treatment sensitivity in cell cultures.

50. • Oda K, Matsuoka Y, Funahashi A, Kitano H: **A comprehensive pathway map of epidermal growth factor receptor signaling**. *Mol Systems Biol* 2005, **1**:8-24.
This review presents a high-quality literature-based network reconstruction of EGF signaling.

51. Shatkay H, Edwards S, Wilbur WJ, Boguski M: **Genes, themes and microarrays: using information retrieval for large-scale gene analysis**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:317-328.

52. • Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery**. *Nat Rev Genet* 2006, **7**:119-129.
Using an example sentence, this review presents a good overview of literature mining for biologists describing the different techniques, what they can be used for and what their limitations are.

53. Rajasethupathy P, Vayttaden SJ, Bhalla US: **Systems modeling: a pathway to drug discovery**. *Curr Opin Chem Biol* 2005, **9**:400-406.

54. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J: **Prominer: rule-based protein and gene entity recognition**. *BMC Bioinformatics* 2005, **6**:S14.

55. Cohen AM, Hersh WR, Dubay C, Spackman K: **Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts**. *BMC Bioinformatics* 2005, **6**:103.

56. Chen L, Liu H, Friedman C: **Gene name ambiguity of eukaryotic nomenclatures**. *Bioinformatics* 2005, **21**:248-256.

57. Gaudan S, Kirsch H, Rebholz-Schuhmann D: **Resolving abbreviations to their senses in Medline**. *Bioinformatics* 2005, **21**:3658-3664.

58. Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA: **Thesaurus-based disambiguation of gene symbols**. *BMC Bioinformatics* 2005, **6**:149.

59. Cooper JW, Kershenbaum A: **Discovery of protein–protein interactions using a combination of linguistic, statistical and graphical information**. *BMC Bioinformatics* 2005, **6**:143.

60. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: **Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome**. *Genome Biol* 2005, **6**:R40.

61. Alako BT, Veldhoven A, van Baal S, Jelier R, Verhoeven S, Rullmann T, Polman J, Jenster G: **Copub mapper: mining Medline based on search term co-publication**. *BMC Bioinformatics* 2005, **6**:51.

62. Temkin JM, Gilder MR: **Extraction of protein interaction information from unstructured text using a context-free grammar**. *Bioinformatics* 2003, **19**:2046-2053.

63. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I: **Extracting human protein interactions from Medline using a full-sentence parser**. *Bioinformatics* 2004, **20**:604-611.

64. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ *et al.*: **Geneways: a system**

for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* 2004, **37**:43-53.

65. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: **Beyond the clause: extraction of phosphorylation information from Medline abstracts**. *Bioinformatics* 2005, **21**:i319-i327.

66. Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P: **Extraction of regulatory gene/protein networks from Medline**. *Bioinformatics* 2005, **22**:645-650.

67. Chen H, Sharp BM: **Content-rich biological network constructed by mining Pubmed abstracts**. *BMC Bioinformatics* 2004, **5**:147.

68. • Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A: **Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease**. *Proc Natl Acad Sci USA* 2004, **101**:15148-15153.
The authors integrated literature-based molecular networks and genetic linkage maps to find candidate disease genes.

69. • Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates**. *Nucleic Acids Res* 2005, **33**:1544-1552.
The authors combined tissue-expression data with disease–tissue relationships extracted from the literature to predict candidate disease genes.

70. Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J: **Analysis of genomic and proteomic data using advanced literature mining**. *J Proteome Res* 2003, **2**:405-412.

71. • Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P: **Systematic association of genes to phenotypes by genome and literature mining**. *PLoS Biol* 2005, **3**:e134.
The authors combined comparative prokaryote genome analysis with literature mining and uncovered genes that might play a role in infectious diseases.

72. Grandjean N, Charpiot B, Pena CA, Peitsch MC: **Competitive intelligence and patent analysis in drug discovery. Mining the competitive knowledge bases and patents**. *Drug Discov Today: Technol* 2005, **3**:211-215.

73. Natarajan J, Berrar D, Dubitzky W, Hack CJ, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG: **Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasivity of a glioblastoma cell line**. *BMC Bioinformatics* 2006, in press.

74. Natarajan J, Mulay N, DeSesa C, Hack CJ, Dubitzky W, Bremer EG: **A grid infrastructure for text mining of full text articles and creation of a knowledge base of gene relations**. *Lecture Notes in Bioinformatics* 2005, **3745**:101-108.

75. Natarajan J, Haines C, Berglund B, DeSesa C, Hack CJ, Dubitzky W, Bremer EG: **Getitfull – a tool for downloading and pre-processing full-text journal articles**. *Lecture Notes in Computer Science* 2006, **3869**:139-145.

76. Romacker M, Grandjean N: 002C Parisot P, Kreim O, Cronenberger D, Vachon T, Peitsch MC: **The Ultralink: an expert system for contextual hyperlinking in knowledge management**. In *Computer Applications in Pharmaceutical Research and Development.* Edited by Ekins S. Wiley & Sons; 2006:729-753.

77. Janes KA, Lauffenburger DA: **A biological approach to computational models of proteomic networks**. *Curr Opin Chem Biol* 2006, **10**:73-80.

78. • Kholodenko BN: **Cell-signalling dynamics in time and space**. *Nat Rev Mol Cell Biol* 2006, **7**:165-176.
Detailed review of principles, archetypical dynamical behaviors (feedback, bistable switch and oscillators), and some classical examples.

79. • Eungdamrong NJ, Iyengar R: **Computational approaches for modeling regulatory cellular networks**. *Trends Cell Biol* 2004, **14**:661-669.
This article uses a MAPK example to orient the reader through a review of methods, including a review of open access tools available at the time. For an updated list see http://www.sbml.org.

80. • Arnold S, Siemann-Herzberg M, Schmid J, Reuss M: **Model-based inference of gene expression dynamics from sequence information**. *Adv Biochem Eng Biotechnol* 2005, **100**:89-179.
Extended review for the expert who wishes to follow the model from building and validation through detailed analysis.

81. Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P *et al.*: **Minimum information requested in the annotation of biochemical models (MIRIAM)**. *Nat Biotechnol* 2005, **23**:1509-1515.

82. Karp PD, Mavrovouniotis ML: **Representing, analyzing, and synthesizing biochemical pathways**. *IEEE Expert* 1994, **9**:11-21.

83. Kitano H, Funahashi A, Matsuoka Y, Oda K: **Using process diagrams for the graphical representation of biological networks**. *Nat Biotechnol* 2005, **23**:961-966.

84. Chien JY, Friedrich S, Heathman MA, de Alwis DP, Sinha V: **Pharmacokinetics/pharmacodynamics and the stages of drug development: role of modeling and simulation**. *AAPS J* 2005, **7**:E544-E559.

85. Andersen ME, Thomas RS, Gaido KW, Conolly RB: **Dose-response modeling in reproductive toxicology in the systems biology era**. *Reprod Toxicol* 2005, **19**:327-337.

86. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS: **Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains**. *Bioinformatics* 2004, **20**:3289-3291.

87. Lin X, Floudas CA, Wang Y, Broach JR: **Theoretical and computational studies of the glucose signaling pathways in yeast using global gene expression data**. *Biotechnol Bioeng* 2003, **84**:864-886.

88. Araujo RP, Liotta LA: **A control theoretic paradigm for cell signaling networks: a simple complexity for a sensitive robustness**. *Curr Opin Chem Biol* 2006, **10**:81-87.

89. Ma'ayan A, Blitzer RD, Iyengar R: **Toward predictive models of mammalian cells**. *Annu Rev Biophys Biomol Struct* 2005, **34**:319-349.

90. Fisher J, Piterman N, Hubbard EJ, Stern MJ, Harel D: **Computational insights into Caenorhabditis elegans vulval development**. *Proc Natl Acad Sci USA* 2005, **102**:1951-1956.

91. Errampalli DD, Priami C, Quaglia P: **A formal language for computational systems biology**. *OMICS* 2004, **8**:370-380.

92. Fages F, Soliman S, Chabrier-Rivier N: **Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM**. *J Bio Phys Chem* 2004, **4**:64-73.

93. Mendoza L, Xenarios I: **A method for the generation of standardized qualitative dynamical systems of regulatory networks**. *Theor Biol Med Model* 2006, **3**:13.

94. Rice JJ, Stolovitzky G: **Making the most of it: pathway reconstruction and integrative simulation using the data at hand**. *Biosilico* 2004, **2**:70-77.

95. Bottino D, Penland RC, Stamps A, Traebert M, Dumotier B, Georgiva A, Helmlinger G: **Preclinical cardiac safety assessment of pharmaceutical compounds using an integrated systems-based computer model of the heart**. *Prog Biophys Mol Biol* 2006, **90**:414-443.

96. Fredj S, Sampson KJ, Liu H, Kass RS: **Molecular basis of ranolazine block of LQT-3 mutant sodium channels: evidence for site of action**. *Br J Pharmacol* 2006, **148**:16-24.