nature genetics

Integrative systems biology

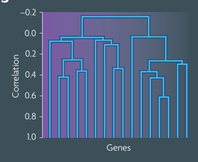
Trey Ideker and Sourav Bandyopadhyay

Systems biology is the study of biology through systematic perturbation, global read-out of the multifaceted response and integration of these data to formulate predictive models[1]. Here, we highlight the key steps in the systems biology approach, with a focus on how global data sets are assembled into models of system structure and function. Techniques for model assembly span many layers of abstraction, including statistical mining, alignment across data sets, probabilistic inference, differential equations and data visualization. These integrative approaches chart the key components and interactions of biological systems over scales ranging from single pathways to whole cells to entire populations of individuals. Major applications of systems biology to biomedical research are to identify genetic risk factors for disease, allow for model-based personalized diagnostics and treatment regimens and suggest new avenues for drug discovery.

Model assembly by data synthesis and integration

Statistical mining

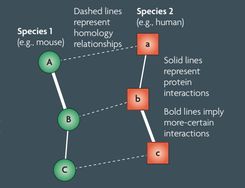
Data filtering and clustering
The most basic form of data integration is to identify statistical overlap between data sets. Another statistical method is clustering, which groups molecules with similar profiles. 'Co-clustering' uses multiple integrated data sets (e.g., mRNA expression & protein networks).
Tools: GenePattern, Cluster, JTreeview, MeV



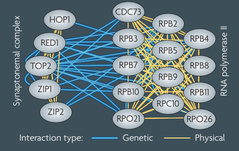
Functional enrichment
Outstanding measurements define gene sets which can be integrated with databases of known gene functions or pathways, indicating statistically enriched categories.
Tools: DAVID, BiNGO, GSEA

Data alignment

Cross-species
Molecular sequences, states or interactions are aligned across species to identify conserved and diverged clusters.
Tools: BLAST, NetworkBLAST (right), IsoRankN

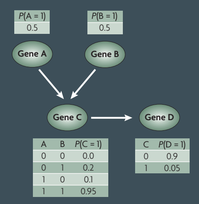


Cross-data type
Alignment is also performed across multiple data types, such as mRNA versus protein profiles or networks of physical versus genetic interactions (right)[2]. Molecular causes (e.g., genetic perturbations) are connected to effects (e.g., expression changes) through physical interaction paths[3].



Probabilistic and mathematical modeling

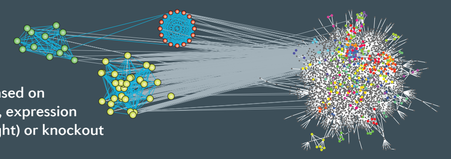
Probabilistic inference
Classification methods (e.g. logistic regression) weigh many different measurements to learn functional links between proteins or other properties[4]. Bayesian networks use changing molecular states over perturbations or time to identify direct causal relationships among genes and can incorporate other data through network 'priors'[5].
Tools: Arachne, Banjo



Network dynamics and fluxes
Information flow through pathways is modeled through differential equations or biophysical simulations, which predict biological outcomes and are fit to measurements and reaction kinetics[3,6]. Tools: SBW, Cell Designer, Copasi

Network visualization

Data projection
Sets network visuals, e.g., node and edge colors, shapes and sizes, based on biological data, e.g., expression levels, functions (right) or knockout phenotypes.

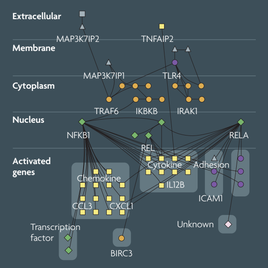


Network layout
Force-directed layouts seek to minimize edge crossings through a physical simulation with edges as springs and with nodes as electrically charged particles.

Attribute-directed layouts use data attributes to attract nodes or edges to a layout region (right, proteins arranged by cellular compartment using Cerebral[7]).

Hierarchical layouts partition nodes into layers, e.g., master regulators above a set of regulated genes.

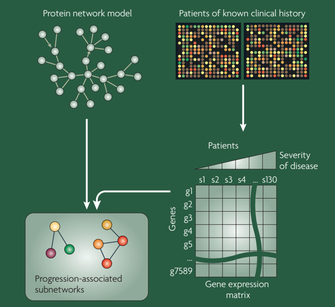
Tools: Cytoscape, Osprey, VisAnt, Pajek

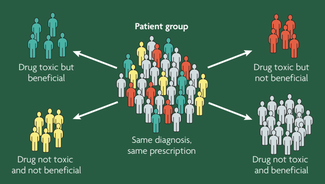


Pharmaceutical and clinical endpoints

A primary application of systems biology is to better understand and treat human disease. Network models will be central to next-generation drug development and patient management tools.

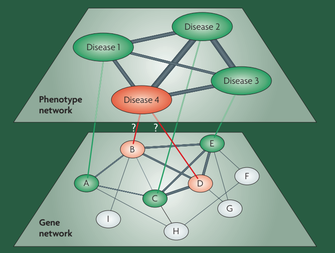
Systems level biomarkers
Biomarkers are classically viewed as individual genetic or protein variants associated with disease risk. In complex diseases, network models integrating a variety of risk factors can improve predictive value (right)[8]. Integrative biomarkers can be developed through superposition of mRNA and microRNA[9] or use of metabolomic profiles[10]. Network level biomarkers have also been used to predict the developmental origin of tissues[11].



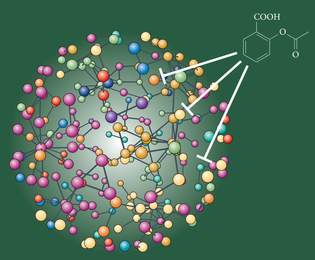


Personalized medicine
Differences in patients' response to a drug are due in part to genetic variation. By integrating datatypes, including genotype, gene expression and protein interaction, network models may improve personalized treatment regimens[12,13].

Genetic risk factors
Genetic association studies link genomic loci to disease risk but may fall short of pinpointing the causal gene or variant. Integration of different datatypes can prioritize functionally plausible candidate genes within the implicated locus[14,15]. Network models can aggregate information from multiple loci into association scores for pathways[16,17].

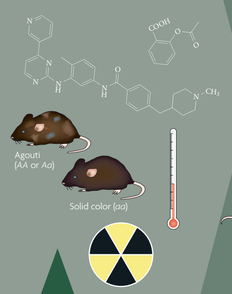


Drug target identification
Network models inform drug development by predicting new gene targets for treatment of disease. They also identify pathways affected by an existing drug, either to predict the mode of action or to suggest drug repurposing in which drugs developed for one purpose can be reused to treat related diseases[18–20].



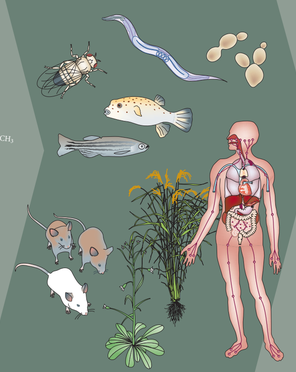
Perturbations

- Chemicals/ small molecules
- Genetic mutations/ RNAi
- Natural variation
- Changing environments
- Shifting time



Biological system

This system under perturbation can range in scale from molecular processes to cells, tissues, single organisms or up to populations of individuals.



Molecular state measurements (nodes)

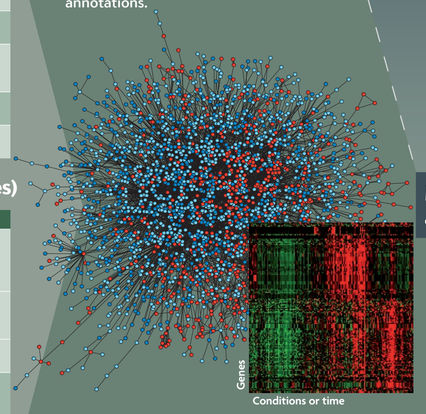
	Large-scale dataset types	Technologies
Genome	Whole-genome DNA sequences, SNPs and CNVs	DNA sequencing, genotyping microarrays
Epigenome	Chromatin modifications and structure	ChIP-seq, methyl-seq, DHS-chip
Transcriptome	Transcript abundances, translation rate and microRNAs	DNA microarrays, RNA-seq, CAGE, GRO-seq, ribosome profiling
Proteome	Protein abundances and modifications	NMR, mass spectrometry, multiparameter FACS
Metabolome	Metabolite profiling	Mass spectrometry, liquid chromatography

Molecular interaction measurements (edges)

	Data types	Technologies
Physical	Protein–protein	Immuno-precipitation (IP), co-affinity purification, yeast two-hybrid, protein arrays, kinase-substrate measurements
	Protein–DNA, protein–RNA	Genome-wide chromatin immuno-precipitation (ChIP), DNA binding arrays
	Protein–small molecule, reaction fluxes	Isotope labeling, mass spectrometry
Genetic and functional	Synthetic lethality, epistasis	Synthetic genetic arrays (SGA) combinatorial RNAi, population genetics
	Cause–effect relationships	Genetic perturbation (gene knockout, RNAi) followed by phenotyping (microarrays, cellular imaging); *trans* eQTLs (expression quantitative trait loci)

Molecular databases

New measurements are stored alongside existing data, including functional annotations.

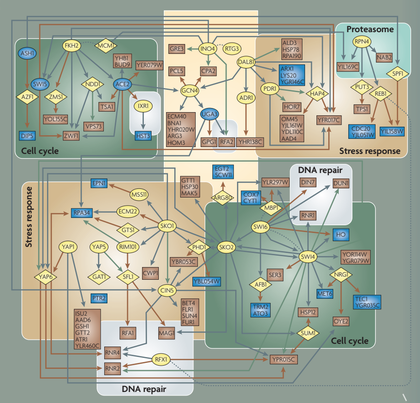


	Annotations	States	Interactions
Selected molecular databases	GO, KEGG, REACTOME, NetPath, UCSC Genome Browser	Genbank, GEO, ArrayExpress, Proteome Commons	BioGRID, HPRD, IntAct, TRANSFAC, STRING, iHOP, functionalnet.org

Model assembly

Library of network models

The critical task of model assembly extracts and integrates the diverse datasets stored in databases into network models that are descriptive, predictive and executable.



Model refinement and validation

An iterative process by which cellular models are refined based on the goodness-of-fit between predictions and data, giving rise to further experimentation.

References

1. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343 (2001).
2. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
3. Herrgard, M. J., Covert, M. W. & Palsson, B. O. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.* **13**, 2423–2434 (2003).
4. Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
5. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. (MIT Press, Cambridge, MA, 2009).
6. Janes, K. A. *et al.* A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).
7. Barsky, A., Gardy, J. L., Hancock, R. E. & Munzner, T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23**, 1040–1042 (2007).
8. Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
9. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).
10. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457**, 910–914 (2009).
11. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).
12. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006).
13. Dressman, H. K. *et al.* An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J. Clin. Oncol.* **25**, 517–525 (2007).
14. Iossifov, I., Zheng, T., Baron, M., Gilliam, T. C. & Rzhetsky, A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res.* **18**, 1150–1162 (2008).
15. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
16. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
17. Hannum, G. *et al.* Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet.* **5**, e1000782 (2009).
18. Campillos, M. *et al.* Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
19. Gardiner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
20. Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).

Acknowledgements
The authors are at the Departments of Medicine and Bioengineering, University of California, San Diego, California, USA. This work was supported by the National Institutes of Health (GM070743, RR018627) and the National Science Foundation (IIS0803937).

Edited by Orli Bahcall; copy edited by Erica Schultz; designed by Simon Fenwick.