

Synthetic cell biology

Adam P Arkin

Synthesis of data into formal models of cellular function is rapidly becoming a necessary industry. The complexity of the interactions among cellular constituents and the quantity of data about these interactions hinders the ability to predict how cells will respond to perturbation and how they can be engineered for industrial or medical purposes. Models provide a systematic framework to describe and analyze these complex systems. In the past few years, models have begun to have an impact on mainstream biology by creating deeper insight into the design rules of cellular signal processing, providing a basis for rational engineering of cells, and for resolving debates about the root causes of certain cellular behaviors. This review covers some of the recent work and challenges in developing these 'synthetic cell' models and their growing practical applications.

Addresses

Howard Hughes Medical Institute, Departments of Bioengineering and Chemistry, University of California, Berkeley, Physical Biosciences Division, EO Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 3-144 Berkeley, CA 94720, USA; e-mail: aparkin@lbl.gov

Current Opinion in Biotechnology 2001, 12:638–644

0958-1669/01/\$ – see front matter

© 2001 Elsevier Science Ltd. All rights reserved.

Introduction

Molecular biology has entered a stage of maturity that requires its transformation into an engineering discipline. The wealth of data on cellular components and their interactions will promote an understanding of cellular behavior that is sufficient for prediction, control and redesign. Diagrams tracing all the interactions, activities, locations and expression times of the proteins, metabolites and nucleic acids involved have become so dense with lines and annotations that reasoning about their functions has become almost impossible.

Given the complexity and quantity of this information, some believe that developing a rational engineering framework for cellular systems may be untenable for the present. An entire industry in 'irrational' engineering of cells has grown up using combinatorial methods in chemistry, genetic engineering and high-throughput screening technology. This approach has proven powerful in certain cases of strain improvement, drug discovery and natural product synthesis [1–5]. What is gained in expediency, however, is often lost in insight. It is often unclear why the pathways and chemicals discovered produce the effects they do, and it is difficult to generalize these results to other systems. Further, the more complicated the target function (e.g. the biosynthetic route to a product or the inhibition, without side-effect, of pathways associated with a disease), the less likely it is that a solution will be found using combinatorial or forced evolutionary methods. The drive for understanding cellular function and the related ability to accurately

diagnose cellular state, together with the economic pull for rational design of metabolic/biosynthetic pathways and molecular strategies for disease treatment, call for the development of computable models.

Models summarize current knowledge and hypotheses about missing information. Depending on the amount of data available and the questions being addressed, models are more or less detailed and abstract. Models containing detailed statements about a process are easier to falsify than abstract models and require more physical detail. A validated physical model is the most predictive and useful for understanding points of control in cellular networks and for designing new functions within them. It is also the most computable type of model.

There have been several reviews of cellular model simulation and engineering published in the past two years (e.g. [6–14]). Many of these provide excellent detailed descriptions of various modeling strategies and applications. Most have focused on metabolism or gene expression as these have the most data and immediate economic impact. A dedicated review of signal transduction models is lacking largely owing to the paucity of such models. They do exist, however, and their small number attests to the difficulty in analysis of this type of pathway. The goal of this article is to bring together the various approaches to modeling biological pathways with a focus on signal transduction and, thus, to provide a touch-point summary of the process and application of cellular model building and analysis.

From data to models

Data are the precursor to any model. The minimal basis of a cellular network model is a list of the molecular players and a list of the 'influences' of one set of players on another or on a lumped cell behavior (such as growth). Molecular players and their interactions have traditionally been discovered through painstaking genetic and biochemical experiments. Technologies like yeast two-hybrid screening, co-immunoprecipitation, surface plasmon resonance, and fluorescence resonance energy transfer (FRET) experiments report on direct interactions between pairs of molecules. Other techniques yield more indirect measures of the interactions among molecules.

Recent advances in high-throughput molecular biology and measurement have led to maps of molecular networks that are stunning in their complexity. For example, there have been several large-scale yeast two-hybrid screens for the detection of protein–protein interactions in organisms ranging from the T7 phage [15] to *Caenorhabditis elegans* [16]. Here, the molecular players are determined either from direct experiment or from predictions with genome annotation tools. To date, the largest scale studies of this

Table 1

Example databases in support of simulation.

Name (details)	Descriptors*	Website
BBID (Biological Biochemical Image Database)	sgmE	http://bbid.grc.nia.nih.gov/
BIND (Biomolecular Interaction Network Database)	sED	http://www.bind.ca/
Brenda (a comprehensive enzyme information system)	mEPM	http://www.brenda.uni-koeln.de/
BRITE (Biomolecular Relations in Information Transmission and Expression)	sgmEP	http://www.genome.ad.jp/brite/brite.html
CSNDB (Cell Signaling Networks Database)	sE	http://geo.nih.gov/jp/csndb/
DIP (Database of Interacting Proteins)	smEPD	http://dip.doe-mbi.ucla.edu/
EcoCyc/Metacyc (Encyclopedia of <i>E. coli</i> Genes and Metabolism)	mED	http://ecocyc.org/
EMP (Enzymes and Metabolic Pathways Database)	mEPM	http://www.empproject.com/
GeNet	sE	http://www.csa.ru:81/Inst/gorb_dep/inbios/genet/genet.htm
GeneNet (information on gene networks)	smgE	http://www.mgs.bionet.nsc.ru/mgs/systems/genenet/
Kegg (Kyoto Encyclopedia of Genes and Genomes)	sgmED	http://www.genome.ad.jp/kegg/kegg.html
SPAD (Signaling Pathway Database)	sgE	http://www.grt.kyushu-u.ac.jp/eny-doc/
Transfac/Transpath	sgED	http://transfac.gbf.de/TRANSFAC/
UM-BBD (University of Minnesota Biocatalysis/Biodegradation Database)	mE	http://umbbd.ahc.umn.edu/
WIT (supports the curation of function assignments made to genes and the development of metabolic models)	mEPM	http://wit.mcs.anl.gov/WIT2/

*The descriptors are as follows: g, genetic pathways; m, metabolic pathways; s, signal transduction pathways; D, can be downloaded; E, edited content; P, primary data; M, mechanistic/kinetic data available.

sort have been carried out in *Saccharomyces cerevisiae*. Uetz *et al.* [17^{*}] found 957 interactions among 1004 proteins. Ito *et al.* [18] found 4549 two-hybrid interactions among 3278 proteins. Interestingly, the two studies found few overlapping interactions. The experimental protocols used by each group differed in several ways and the screening was far from saturated. In addition, two-hybrid screening methods have substantial rates of false-positives and false-negatives.

Newer experimental technologies have been developed for indirectly deducing interactions among molecules. These include the multiple alignment of DNA regions upstream of genes clustered by their expression patterns [19], statistical analysis of concomitant variation and temporal sequencing [20], and perturbation/response modeling [21,22,23^{*},24,25]. By whatever method interactions are deduced, direct or indirect, the results are reported in the literature in non-standard formats. There have been some recent encouraging results using natural language processing (NLP) techniques to extract protein names and interactions directly from the online literature in order to predict cellular networks [26–28]. As applications of NLP technologies to molecular biology are still in their early days, they currently have high false-positive and false-negative rates and severe limitations on the type and complexity of interaction they can capture. Nevertheless,

they show exceptional promise in aiding the early stages of model development.

Less common are experimental methods for assigning mechanism and obtaining physical constants for all the known interactions and processes; these are the most difficult to obtain even *in vitro*. *In vivo* quantitative measurements are few and far between, with the notable exception of NMR-based flux measurements [29]. In a model with many interactions, the probability of finding any quantitative data on all of the interactions is minuscule. When such data exist, the conditions for each experimental measurement and even the strain of organism may be different. This dearth of information, notable in this data-rich age, is the major impediment to detailed predictive models of cellular networks.

The need to organize as much experimental data as possible in a systematic manner has led to several excellent databases of molecular properties, interactions and pathways. These range from highly edited and curated databases, such as EcoCyc, to databases of primary data such as the Database of Interacting Proteins. Table 1 lists some of these databases and their properties. These databases provide an essential infrastructure for future modeling efforts, although they are difficult to compile and maintain. They also suffer from a plethora of non-standard

formats for storage, query and transport. Some of these problems are being addressed by the development of XML-based data transport standards (e.g. <http://xml.coverpages.org/xml.html#applications>), but standards are still in flux. When journals start requiring that data generated in support of a paper be submitted in standard, machine-readable format, many of these problems will be ameliorated.

Classes of models

Graphical models

Because of the heterogeneity in data type, quality and availability, cellular network modelers have had to develop several different model classes that can operate at different levels of abstraction. The most common models are graphical models (i.e. cartoons) of the process. Cartoons graphically depict each biological component connected to others with arrows indicating their interaction. There is little standard nomenclature for cartoons, although at least two formal graphical annotations have been suggested recently [30*,31]. But even Kurt Kohn's wonderful summary cartoon of the mammalian cell cycle [30*], wherein the conventions for representing interactions and species are outlined in some detail, contains abstractions and missing information. One example in this system is the representation of the important tumor suppressor p53. The diagram shows this single protein with 27 sites to which phosphate can be added or removed by various specific enzymes. The implication is that this protein, theoretically, can be in any of 2^{27} (= 34,217,728) possible phosphorylation states and each of these states has a different possible Gibbs free energy and different interaction kinetics with all other molecules in the system. This is a problem not only for this model type, but spells trouble for any detailed mechanistic model of this system. Nevertheless, the graphical model summarizes a great deal of the current information about a pathway and facilitates the formation of hypotheses about network function as well as pointing out some of the difficulties involved in understanding the network.

Qualitative models

Qualitative models are the first form of a model beyond a cartoon that can be analyzed automatically. They range from simple graphs to logical and statistical models. For example, Jeong and colleagues [32,33] used only the yeast protein-protein interaction data cited above to conclude that the statistical properties of the graph implied a particular stability of network function to most 'deletions' in the graph. This conclusion was strengthened by correlation of the number of interactions per protein with phenotypes of knockout mutants collected from the literature. For more dynamical and specific questions, logical models are often used when mechanistic data are lacking. Boolean, fuzzy logical or rule-based systems have been developed to approach the simulation of complex networks; Thieffry and Thomas [10] review their pioneering work in this field. Many groups have used this paradigm for modeling genetic and developmental systems. Lee *et al.* [34] used fuzzy logic (a generalization of Boolean logic) as a supplement to

kinetic models to include uncertain information necessary for fitting the kinetics of metabolic enzymes. Trelease *et al.* [35] used a general qualitative simulation tool, QSIM, to simulate the effect of exogenous gene activations in the NF κ B network. All these models require expert insight to codify the high-level rules in a consistent and accurate fashion. Because so much interpretation of the data is necessary before a model is made, there is an increased danger of building-in a desired answer.

When perturbation response or time-series data are available, statistical influence models become feasible. Linear [36], neural network-like [37], and Bayesian models [20] have all been used to deduce both the topology of gene expression networks and their dynamics. The amount of data necessary to fit these models often prohibits their use. Statistical influence models are not precisely causal models in that they are fits of the model structure to indirect data on interactions. Interpretations of control in these models must be cautious.

Mechanistic models

With enough data, more mechanistic models can be developed. Cybernetic [38] and power law [39] formalisms assert a causal structure, but employ generic nonlinear functions numerically fit to kinetic data and possibly constrained by optimal conditions. Such models form the basis of a large class of metabolic control analyses and dynamic simulations. More detailed models require that chemical or physical mechanisms be asserted for each interaction. For example, McAdams and Arkin [40] propose that, because of the small concentrations of the molecules involved, gene expression must be a stochastic process of a particular sort. They followed the implications of the theory in an integrated model of the λ phage lysis/lysogeny decision and showed that the decision is fundamentally non-deterministic [41]. Physical models have the largest data requirement, are the most difficult to falsify, and, in principle, are the most predictive.

There is always a balance between top-down and bottom-up models. No model is fully bottom-up. Abstractions can both clarify the sources of control in a network and indicate where more data are necessary. There is always the problem of unknown players and unknown and uncharacterized interactions in the network. A formal model that can be represented in mathematical form has the advantage of being a precise statement of the current understanding and can be formally proved or disproved and checked for consistency.

Basic analysis and simulation of models

Once a model has been formulated, there are several standard approaches for analyzing its properties. If the model is dynamic, then simulation is the most common approach. Steady-state analyses like bifurcation theory, stoichiometric network analysis, flux-balance analysis, and sensitivity analysis are also commonly used. These give more detailed insights into system control and can indicate where the

Table 2

Example simulation programs.

Name	Descriptors*	Website
Gepasi/Copasi	fkFW	http://gepasi.dbs.aber.ac.uk/softw/gepasi.html
BioSim	qWMU	http://www.molgen.mpg.de/~biosim/BioSim/BioSimHome.html
Jarnac	krfbFWS	http://members.tripod.co.uk/sauro/Jarnac.htm
DBSOLVE	kbFWD	http://websites.ntl.com/~igor.goryanin/
MCELL	rsU	http://www.mcell.cnl.salk.edu/
Virtual Cell	ksDFWMU	http://www.nrcam.uchc.edu/
E-Cell	kWUS	http://www.e-cell.org/
Neuron	ksFWMUS	http://neuron.duke.edu/
Genesis	ksUS	http://www.bbb.caltech.edu/GENESIS/genesis.html
Plas	kfbFW	http://correio.cc.fc.ul.pt/~aenf/plas.html
Ingeneue	qkFMWUS	http://www.ingeneue.org/
DynaFit	kfW	http://www.biokin.com/dynafit/
Stochsim	rS	http://www.zoo.cam.ac.uk/comp-cell/StochSim.html
T7 Simulator	kUS	http://virus.molsci.org/t7/
Molecularizer/Stochastirator	krUS	http://opnsrbcio.molsci.org/alpha/comps/sim.html

All packages have facilities for chemical kinetic simulation of one sort or another. Some are better designed for metabolic systems, others for electrochemical systems, and still others for genetic systems. *The descriptors are as follows: b, bifurcation analyses and steady-state calculation; f, flux balance/metabolic control and related analyses;

k, deterministic kinetic simulation; q, qualitative simulation; r, stochastic process models; s, spatial processes; D, database connectivity; F, fitting, sensitivity and optimization code; M, runs on Macintosh; S, source code available; U, runs on linux or Unix; W, runs on windows.

model is sensitive to parameters or missing data. Any analysis of a model generally goes beyond these formal methodologies (discussed below).

Several different tools have been developed to simulate and analyze models of cellular systems (see Table 2). More general tools, such as Mathematica and MATLAB, are also commonly used for simulation. Because these two programs are ubiquitous, they provide good facilities for transferring models among different researchers. Platform-independent model specification languages are also under development; for example, SBML (<http://www.cds.caltech.edu/erato/sbml/docs/index.html>) and CellML (<http://www.cellml.org/>) are currently being developed in a cooperative and community fashion.

Applications

Models and their analysis have many purposes. In recent years we have seen the development of interesting uses. The small selection of papers described below serves to demonstrate how models can be useful in organizing thoughts and testing hypotheses.

Demonstrating a design property of a network

Quite often a network is so complex or so odd in structure that it is of interest to understand what properties of its design are necessary for cellular function. In a semimechanistic

model of the gap and pair rule (genes in *Drosophila melanogaster* that determine segment polarity) von Dassow *et al.* [42**] showed that the structure of the network is both sufficient to explain a great deal of the observed cellular patterning and, moreover, that the network behavior is robust to parameter variation. To achieve this robustness the authors had to add hypothetical (but reasonable) additions to the known network, thus demonstrating another use for models: the ability to formally propose and justify new mechanistic hypotheses and predict new network elements. This evokes earlier results by Barkai and Leibler [43] who demonstrated robustness in exact adaptation in the *Escherichia coli* chemotactic pathways. Endy *et al.* [44**], in one of the first nearly whole genome models of an organism, model the entire life-cycle of the T7 phage during infection of *E. coli* [45]. They explore the effect of genome organization on the efficiency of phage growth and find that the qualitative behavior of the model was largely insensitive to genome organization. Specific 'shuffled' genomes were found that allowed the computational phage to grow better than wild-type under specified conditions. These predictions were not borne out experimentally, but the experiment did not entirely match the 'genotype' of the model. Even so, models can be most useful when they fail in unexpected ways. Modifications to the model to bring it in line with experiment provide hypotheses that can then be experimentally tested.

Another interesting example of a generic design property in a signal transduction model is given by the investigation of Levchenko *et al.* [45] into the role of scaffolding proteins in mitogen-activated protein kinase signaling. They showed that the combinatorics of protein binding to the scaffold molecule can either amplify or reduce activity of the pathway depending on the relative concentration of the scaffold itself.

Developing an understanding of endogenous control

Once more mechanistic models become available, it will be possible to make precise statements about control in the networks. The most popular form of control analysis is metabolic control analysis, which has been reviewed elsewhere [12]. One of the best examples of elucidating control with simulation, phase plane analysis and bifurcation theory is the ongoing work by Tyson on the yeast cell cycle. Chen *et al.* [46•] summarize and extend this work to elucidate the control of different phases of mitosis and explain the impact of 50 different mutants on these decisions. The control of cytokine trafficking in immune cells has been modeled by several groups. Recently, Fallon and Lauffenburger [47•] used a kinetic model to explain, among other things, a counter-intuitive result that the increased potency of an interleukin-2 analog does not derive from tighter receptor binding, but instead results from differential binding to receptor subunits that are alternately sorted.

Developing a strategy for control or design

Models can be used to test design ideas for engineering networks in cells. Elowitz and Leibler [48] and Gardner, Cantor and Collins [49] used very simple models to support the design of a genetic oscillator and a switch in *E. coli*. Models can also be used to test designs for the control of cellular networks. Endy and Yin [50] used their T7 model to propose a pharmaceutical strategy for preventing both T7 propagation and the development of drug resistance through mutation.

Proving necessity and/or sufficiency

Given an observed cell behavior, models can be used to prove necessity of a given regulatory motif or the sufficiency of known interactions to produce the phenomenon. Yi *et al.* [51] demonstrate the necessity of at least one integral feedback loop to explain robust adaptation in the chemotactic signal transduction pathways; they identify the loop in *E. coli* with the receptor methylation/demethylation system. Qi, Groves and Chakraborty [52••] demonstrate the sufficiency of membrane energetics, protein diffusion, and receptor-binding kinetics to generate a particular dynamic pattern of protein location at the synapse between two immune cells. The model is striking in that the cytoskeleton is thought to play an active role in organization of the supramolecular structures. In rare cases, models can also be used to show insufficiency of a reasonable model to explain observed behavior (e.g. see the discussion in [44••]).

Explaining the contradictory or exotic behavior of complex networks

Contradictions in the literature about any particular pathway abound. In one example, the behavior of several different mutants in the *E. coli* chemotaxis pathway seemed to behave contrary to the logical organization of the network. Abouhamad *et al.* [53••] used a model of chemotaxis that had been validated on over 65 mutant strains of *E. coli* to help design experiments to successfully explain the paradoxes. In a series of simplified kinetic models for eukaryotic signal transduction pathways, such as the epidermal growth factor receptor signaling pathway, Bhalla and Iyengar [54] explore how these pathways could generate sustained activation, and bistable and pulsatile behaviors.

There are many other excellent models with different properties and goals than those described here, and it is unfortunate that there is not space to discuss them further. As different researchers build more models about related systems, there will be a drive to combine them. The models will cover a wide range of physical phenomena from electrochemical to mechanical systems and cross scales from atomic dimensions to hundreds of microns. Aside from standardizing model specifications, there are deep theoretical and computational challenges to analyzing and simulating these multivariate, multiscale and hybrid systems that have proven refractory for decades. The examples above demonstrate that these difficulties need not prevent modeling from being a powerful adjuvant for the exploration of network function.

Conclusions

Modeling is becoming a common and powerful support for understanding cellular behavior. Paradoxically, although new measurement technologies are uncovering increasingly complex networks of chemical and physical interactions, there are relatively few systems with enough quality data to create detailed models of cellular function. The applications of these models are of sufficient importance to demand that this void be filled. The challenge is to design the computation/experiment cycle to efficiently deduce and characterize signaling pathways in single cells. This will require the application of many different experimental and computational technologies. Data from microarrays, protein mass spectroscopy, capillary and high-pressure chromatographies, high-end fluorescence microscopy, and other techniques must be combined so that models of sufficient accuracy can be built and validated on standardized datasets. Such coordinated marshaling of researchers and resources towards a shared goal is a common model for industry, but this multilaboratory approach is new for the academic environment. Large government-funded projects like the Alliance for Cellular Signaling (<http://www.cellularsignaling.org/>), a consortium of researchers dedicated to the quantitative study of G-protein-coupled signal transduction in cardiomyocytes and B cells, or private organizations like the Institute for Systems Biology (<http://www.systemsbiology.org/home.html>) are the new

great experiments in bringing a cooperative approach to academic biology. These efforts will have to be carefully nurtured if this first great push to synthesize data accumulated over half a century on the molecular basis of cellular function is to succeed.

Acknowledgements

I gratefully acknowledge the Howard Hughes Medical Institute (HHMI), Defense Advanced Research Project Agency (DARPA) and National Institutes of Health for support of research particular to this review. I also wish to thank Antoinette Czerwinski, Alex Gilman, Janet Jacobsen and Denise Wolf for editorial and critical readings of the manuscript.

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. Khosla C, Zawada RJ: **Generation of polyketide libraries via combinatorial biosynthesis.** *Trends Biotechnol* 1996, 14:335-341.
2. Hutchinson CR: **Combinatorial biosynthesis for new drug discovery.** *Curr Opin Microbiol* 1998, 1:319-329.
3. Bull AT, Ward AC, Goodfellow M: **Search and discovery strategies for biotechnology: the paradigm shift.** *Microbiol Mol Biol Rev* 2000, 64:573-606.
4. Xue Y, Sherman DH: **Biosynthesis and combinatorial biosynthesis of pikromycin-related macrolides in *Streptomyces venezuelae*.** *Metab Eng* 2001, 3:15-26.
5. Rohlin L, Oh M, Liao JC: **Microbial pathway engineering for industrial processes: evolution, combinatorial biosynthesis and rational design.** *Curr Opin Microbiol* 2001, 4:330-335.
6. Bailey JE: **Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities.** *Biotechnol Prog* 1998, 14:8-20.
7. Zhu C: **Kinetics and mechanics of cell adhesion.** *J Biomech* 2000, 33:23-33.
8. Gombert AK, Nielsen J: **Mathematical modelling of metabolism.** *Curr Opin Biotechnol* 2000, 11:180-186.
9. van Helden J, Naim A, Mancuso R, Eldridge M, Wernisch L, Gilbert D, Wodak SJ: **Representing and analysing molecular and cellular function using the computer.** *Biol Chem* 2000, 381:921-935.
10. Thieffry D, Thomas R: **Qualitative analysis of gene networks.** In *Pacific Symposium on Biocomputing* 1998:77-88.
11. Smolen P, Baxter DA, Byrne JH: **Modeling transcriptional control in gene networks — methods, recent results, and future directions.** *Bull Math Biol* 2000, 62:247-292.
12. Schilling CH, Schuster S, Palsson BO, Heinrich R: **Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era.** *Biotechnol Prog* 1999, 15:296-303.
13. McAdams HH, Arkin A: **Simulation of prokaryotic genetic circuits.** *Annu Rev Biophys Biomol Struct* 1998, 27:199-224.
14. Christensen B, Nielsen J: **Metabolic network analysis. A powerful tool in metabolic engineering.** *Adv Biochem Eng Biotechnol* 2000, 66:209-231.
15. Bartel PL, Roedelkin JA, SenGupta D, Fields S: **A protein linkage map of *Escherichia coli* bacteriophage T7.** *Nat Genet* 1996, 12:72-77.
16. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M: **Protein interaction mapping in *C. elegans* using proteins involved in vulval development.** *Science* 2000, 287:116-122.
17. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, 403:623-627.
18. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, 98:4569-4574.
19. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, 22:281-285.
20. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, 7:601-620.
21. D'Haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, 16:707-726.
22. Ideker TE, Thorsson V, Karp RM: **Discovery of regulatory interactions through perturbation: inference and experimental design.** In *Pacific Symposium on Biocomputing* 2000, 292:305-316.
23. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, 292:929-934.
- This is an excellent example of using data-mining and data-fusion tools to develop statistical models of a pathway.
24. Arkin AP, Ross J: **Statistical construction of chemical mechanisms from measured time-series.** *J Phys Chem* 1995, 99:970-979.
25. Arkin AP, Shen P, Ross J: **A test case of correlation metric construction of reaction pathways from measurements.** *Science* 1997, 277:1275-1279.
26. Park JC, Kim HS, Kim JJ: **Bidirectional incremental parsing for automatic pathway identification with combinatorial grammar.** In *Pacific Symposium on Biocomputing* 2001:396-407.
27. Ng SK, Wong M: **Toward routine automatic pathway discovery from on-line scientific text abstracts.** *Genome Inform Ser Workshop Genome Inform* 1999, 10:104-112.
28. Ono T, Hishigaki H, Tanigami A, Takagi T: **Automated extraction of information on protein–protein interactions from the biological literature.** *Bioinformatics* 2001, 17:155-161.
29. Shulman RG, Rothman DL: **¹³C NMR of intermediary metabolism: implications for systemic physiology.** *Annu Rev Physiol* 2001, 63:15-48.
30. Kohn KW: **Molecular interaction map of the mammalian cell cycle control and DNA repair systems.** *Mol Biol Cell* 1999, 10:2703-2734.
- A tour de force summary of eukaryotic cell-cycle control with a graphical model. It sets the current standard in diagrammatic representation of networks.
31. Pirson I, Fortemaison N, Jacobs C, Dremier S, Dumont JE, Maenhaut C: **The visual display of regulatory information and networks.** *Trends Cell Biol* 2000, 10:404-408.
32. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, 411:41-42.
33. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, 407:651-654.
34. Lee B, Yen J, Yang L, Liao JC: **Incorporating qualitative knowledge in enzyme kinetic models using fuzzy logic.** *Biotechnol Bioeng* 1999, 62:722-729.
35. Trelease RB, Henderson RA, Park JB: **A qualitative process system for modeling NF- κ B and AP-1 gene regulation in immune cell biology research.** *Artif Intell Med* 1999, 17:303-321.
36. D'Haeseleer P, Wen X, Fuhrman S, Somogyi R: **Linear modeling of mRNA expression levels during CNS development and injury.** In *Pacific Symposium on Biocomputing* 1999, 163:41-52.
37. Mjolsness E, Sharp DH, Reinitz J: **A connectionist model of development.** *J Theor Biol* 1991, 152:429-453.
38. Varner J, Ramkrishna D: **Metabolic engineering from a cybernetic perspective: aspartate family of amino acids.** *Metab Eng* 1999, 1:88-116.
39. Voit EO, Radvovoyevitch T: **Biochemical systems analysis of genome-wide expression data.** *Bioinformatics* 2000, 16:1023-1037.
40. McAdams HH, Arkin A: **Stochastic mechanisms in gene expression.** *Proc Natl Acad Sci USA* 1997, 94:814-819.

With reference [18] this paper is representative of a rich dataset for modeling. The analyses developed in this paper are an excellent example of the next generation of network bioinformatics.

41. Arkin A, Ross J, McAdams HH: **Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected *Escherichia coli* cells.** *Genetics* 1998, **149**:1633-1648.
42. von Dassow G, Meir E, Munro EM, Odell GM: **The segment polarity network is a robust developmental module.** *Nature* 2000, **406**:188-192.
This article is an excellent example of semiquantitative modeling of a complex signaling pathway. It is also representative of an ongoing discussion in the field about robustness of network behavior and whether or not there is a non-trivial general theory of this phenomenon. The application of a particular modeling tool developed for the purpose (Ingeneue) is demonstrated.
43. Barkai N, Leibler S: **Robustness in simple biochemical networks.** *Nature* 1997, **387**:913-917.
44. Endy D, You L, Yin J, Molineux IJ: **Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes.** *Proc Natl Acad Sci USA* 2000, **97**:5375-5380.
This paper has several points of interest. It presents nearly a whole genome model of the organism and models nearly the whole life-cycle. Furthermore this work explores a novel question in genome organization and shows how experiment and theory interact.
45. Levchenko A, Bruck J, Sternberg PW: **Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties.** *Proc Natl Acad Sci USA* 2000, **97**:5818-5823.
46. Chen KC, Csikasz-Nagy A, Gyorfy B, Val J, Novak B, Tyson JJ: **Kinetic analysis of a molecular model of the budding yeast cell cycle.** *Mol Biol Cell* 2000, **11**:369-391.
The latest description of one of the most developed models in the field and an excellent demonstration of model analysis and reasoning.
47. Fallon EM, Lauffenburger DA: **Computational model for effects of ligand/receptor binding properties on interleukin-2 trafficking dynamics and T cell proliferation response.** *Biotechnol Prog* 2000, **16**:905-916.
A detailed kinetic model of a eukaryotic signal transduction pathway that explains some counter-intuitive data. The paper also provides an excellent comparison with experimental results.
48. Elowitz MB, Leibler S: **A synthetic oscillatory network of transcriptional regulators.** *Nature* 2000, **403**:335-338.
49. Gardner TS, Cantor CR, Collins JJ: **Construction of a genetic toggle switch in *Escherichia coli*.** *Nature* 2000, **403**:339-342.
50. Endy D, Yin J: **Toward antiviral strategies that resist viral escape.** *Antimicrob Agents Chemother* 2000, **44**:1097-1099.
51. Yi TM, Huang Y, Simon MI, Doyle J: **Robust perfect adaptation in bacterial chemotaxis through integral feedback control.** *Proc Natl Acad Sci USA* 2000, **97**:4649-4653.
52. Qi SY, Groves JT, Chakraborty AK: **Synaptic pattern formation during cellular recognition.** *Proc Natl Acad Sci USA* 2001, **98**:6548-6553.
This detailed mechanistic model is an outstanding example of a 'sufficiency' argument. The model itself is interesting for its physical basis.
53. Abouhamad WN, Bray D, Schuster M, Boesch KC, Silversmith RE, Bourret RB: **Computer-aided resolution of an experimental paradox in bacterial chemotaxis.** *J Bacteriol* 1998, **180**:3757-3764.
This chemotaxis model is one of the most developed and powerful in the field and this paper demonstrates how it is used to drive experiments to explain complex cell-signaling behavior.
54. Bhalla US, Iyengar R: **Emergent properties of networks of biological signaling pathways.** *Science* 1999, **283**:381-387.