Building with a scaffold: emerging strategies for high- to low-level cellular modeling

Trey Ideker[1] and Douglas Lauffenburger[2]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Massachusetts Institute of Technology, Biological Engineering Division, Cambridge, MA 02139, USA

Computational cellular models are becoming crucial for the analysis of complex biological systems. An important new paradigm for cellular modeling involves building a comprehensive scaffold of molecular interactions and then mining this scaffold to reveal a hierarchy of signaling, regulatory and metabolic pathways. We review the important trends that make this approach feasible and describe how they are spurring the development of models at multiple levels of abstraction. Pathway maps can be extracted from the scaffold using 'high-level' computational models, which identify the key components, interactions and influences required for more detailed 'low-level' models. Large-scale experimental measurements validate high-level models, whereas targeted experimental manipulations and measurements test low-level models.

The Human Genome Project has taught us the immense power of systematic biology for understanding gene function. Consider that we, as biologists, are often confronted with a short stretch of DNA corresponding to an unknown gene, typically isolated from a clone library or a gel electrophoresis experiment. Given that the Human Genome Project has deposited the complete sequence of all genes into a publicly accessible database, we can use software tools, such as BLAST [1], to search this genome database for sequences that are similar to that of the unknown gene. Some of these similar sequences are likely to correspond to genes or proteins with known functions, and, by association, we can infer that the function of our novel sequence is related. Starting from an initial query sequence, a genome database search efficiently yields information about how this sequence is positioned in a greater functional context.

In the post-genomic era, focus is now shifting from understanding the function of individual proteins to understanding how the many proteins interact together in a complex web of signaling, regulatory, structural and metabolic pathways in the cell. In contrast to the systematic methods of genome sequencing, however, efforts to identify and characterize pathways typically proceed in a molecule-directed fashion, beginning with an initial protein of interest and trying to establish other proteins that are involved in the same pathway. For example, the initial protein might be used as a 'bait' in genetic experiments such as a synthetic lethal screen [2]. These approaches implicate additional proteins that are possibly involved in the same pathway, which can themselves be used as baits in future genetic and biochemical experiments.

Although molecule-directed approaches have been successful in assembling most of the knowledge we have about pathways to date, they are associated with several inherent difficulties. The first is the time required: accurate models of pathway function emerge only after evidence is accumulated over many years by many researchers and laboratories. Second, these approaches do not directly reveal how multiple pathways influence each other, or reveal this cross-talk only accidentally. Third, despite encouraging efforts to construct consolidated pathway databases [3] (http://stke.sciencemag.org/; http://www.afcs.org/), the vast amount of information on the various intracellular pathways remains decentralized, buried in the primary literature.

Just as systematic sequencing projects led to a revolution in mapping genes and genomes, might it therefore be feasible to adopt a systematic approach to mapping pathways? Indeed, several emerging experimental and computational trends indicate how such a systems approach might work. In all of these cases, the key preliminary step involves building a comprehensive scaffold of molecular interactions that broadly covers many aspects of cellular function and physiological responses. Although this step constitutes a sizable initial investment, the molecular interaction scaffold provides a broad foundation for more directed studies to follow. Just as we can use BLAST to query a genome database to identify sequences of interest, so new pathway discovery and search tools are enabling systems biologists to query the molecular interaction scaffold to identify and map pathways of interest in a systematic fashion.

Descending from the scaffold to high- and low-level pathway models

A good way to visualize this pathway mapping procedure is as a descent through a series of models at increasing levels of detail and decreasing levels of abstraction (Box 1). At the highest level of abstraction, the goal is to analyse the

Corresponding author: Trey Ideker (trey@wi.mit.edu).

 0167-7799/03/$ - see front matter doi:10.1016/S0167-7799(03)00115-X

Box 1. Diverse spectrum of high- to low-level computational modeling approaches

Computational models of cellular processes span a wide range of levels of abstraction (Fig. I). At the highest level, statistical data-mining approaches correlate dependent with independent variables, elucidating model components and their potential interrelationships. At a somewhat lower level, Bayesian networks expand on these relationships by modeling conditional dependencies of 'child' nodes on their 'parents' in the network, whereas Boolean- and fuzzy-logic models dictate logical rules governing these dependencies. Finally, at a relatively detailed level, Markov chains allow probabilistic production, loss and interconversion among molecular species and states, and complex systems of differential equations explicitly model physico-chemical reaction rates, binding constants and diffusion and transport coefficients.

Abstract, high-level models generally represent qualitative features of a system, whereas detailed, low-level models most typically represent quantitative aspects. However, abstraction and quantitation need not be mutually exclusive concepts. For example, a high-level model might include the quantitative probability that gene *A* will undergo an expression change in response to perturbation of *Z*. Similarly, a low-level analysis might produce the qualitative network structure giving the best fit to an observed set of mRNA levels for *A*, *B* and *C* in response to *Z* perturbation.

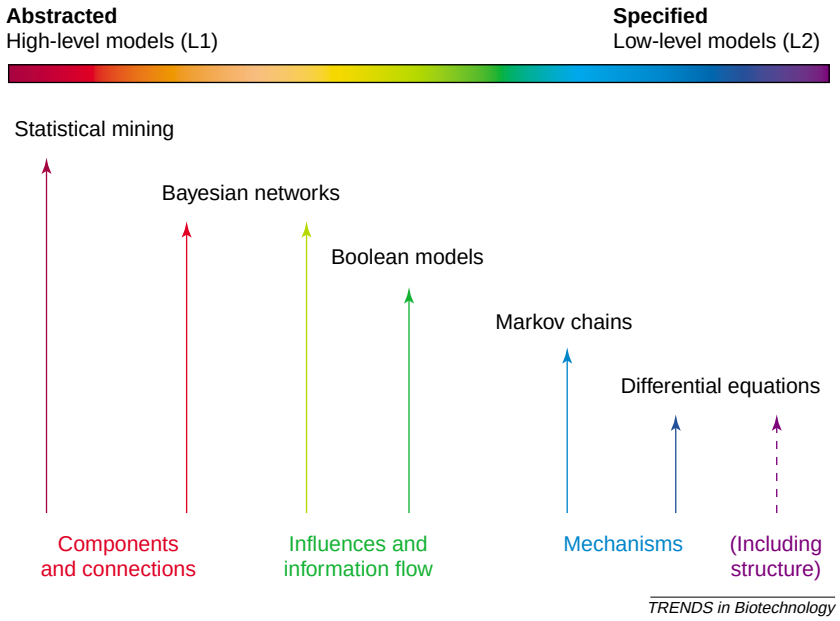

Fig. I. A diverse spectrum of high- to low-level computational modeling approaches.

complete interaction scaffold to extract the basic components and connectivity of the pathway of interest. We term these connectivity-driven models 'High Level' or 'Level One' (L1) pathway models. At a more detailed level, we might wish to supplement the overall pathway connectivity with an indication of how biological information flows from one component to another in the pathway and, at a lower level still, with the abundances, kinetics and binding affinities of pathway components and interactions. We term this broad, more detailed class of models 'Low Level' or 'Level Two' (L2) models. The Systems Biology Markup Language (SBML) [4], under development as a modeling and interchange format for biological pathways, uses a similar nomenclature for pathway models at different levels of abstraction. Of course, specifying these levels is by nature somewhat arbitrary; for purposes of this article, our intention is merely to distinguish between models at one end of the spectrum from those at the other. Clearly, a major challenge facing the modeling community at large is how to move among models at the various levels.

We now discuss in further detail the trends in experimental and computational molecular biology that are making pathway mapping feasible on a large scale. Based on these trends, we propose a strategy for extracting high-level models from the interaction scaffold, for moving between high- and low-level pathway models, and for designing experiments to advance these models most effectively.

Systematic experiments for characterizing networks and states

Signaling and regulatory pathways consist of some number of components, such as genes, proteins and small molecules, wired together in a complex network of intermolecular interactions. Recent technological developments are enabling us to define and interrogate these pathways more directly and systematically than ever before, using two complementary approaches. First, it is now possible systematically to measure the molecular interactions themselves, by screening for protein–protein, protein–DNA and small-molecule interactions (Table 1; first column). Several methods are available for measuring protein–protein interactions at large scale, two of the most popular being the yeast two-hybrid system [5,6] and protein co-immunoprecipitation in conjunction with

Table 1. Two systematic ways to learn about pathways[a]

	(1) Directly observe the interactions	(2) Observe states induced by interactions
	Protein–DNA interactions	**Gene expression**
Methods	Chromatin immunoprecipitation followed by microarray analysis	DNA microarrays; SAGE
Databases	TRANSFAC (http://transfac.gbf.de/TRANSFAC/) BIND (http://www.bind.ca/)	GEO (http://www.ncbi.nlm.nih.gov/geo/) ArrayExpress (http://www.ebi.ac.uk/microarray/ArrayExpress/)
	Protein–protein interactions	**Protein levels, locations, modifications**
Methods	Two hybrid system Co-immunoprecipitation followed by mass spectrometry	Mass spectrometry; 2D PAGE Protein tagging followed by fluorescence microscopy; Protein arrays
Databases	BIND (http://www.bind.ca/) DIP (http://dip.doe-mbi.ucla.edu/) BRITE (http://www.genome.ad.jp/brite/) MIPS (http://mips.gsf.de/)	SWISS-2DPAGE (http://us.expasy.org/ch2d/) TRIPLES (http://ygac.med.yale.edu/triples/) Scansite (http://scansite.mit.edu/)
	Metabolic interactions and reactions	**Metabolite and drug levels**
Methods	No truly systematic measurements, although protein arrays show promise	Mass spectrometry; two-dimensional NMR Current challenge is to determine the molecular identities of all distinct compounds detected
Databases	MetaCyc (http://biocyc.org/metacyc/) KEGG (http://www.genome.ad.jp/kegg/) Klotho (http://www.biocheminfo.org/klotho/)	Public repositories of metabolic profiles not widely available, although data exchange standards for expression profiles (e.g. MAGE-ML) might support metabolic data in future.

[a]This table is provided as a representative sample of methods and databases, not as a comprehensive listing. We apologize in advance to those whose work was omitted because of space considerations.

tandem mass spectrometry [7,8]. Although the vast majority of protein interactions have been generated for the budding yeast *Saccharomyces cerevisiae*, protein interactions are becoming available for a variety of other species including *Helicobacter pylori* [9] and *Caenorhabditis elegans* [10], and are catalogued in public databases such as BIND [11] and DIP™ [12]. A current drawback of these high-throughput measurements is an associated high error rate [13]. As we will discuss, one way to address this problem is to integrate several complementary datasets together (e.g. two-hybrid interactions with coIP data or gene expression profiles) to reinforce the common signal.

Protein–DNA interactions, as commonly occur between transcription factors and their DNA binding sites, constitute another interaction type that can now be measured with high throughput. Recently, Lee *et al.* [14] used the technique of chromatin immunoprecipitation to characterize the complete set of promoter regions bound under nominal conditions for each of >100 transcription factors in yeast, revealing >5000 novel protein–DNA interactions in that organism. Additional types of pathway interactions, such as those between proteins and small molecules (carbohydrates, lipids, drugs, hormones and other metabolites), are difficult to measure on a large scale, although protein array technology [15–17] might enable high-throughput measurements of protein–small-molecule interactions in the near future.

In addition to characterizing molecular interactions, a second major way to interrogate pathways is to systematically measure the molecular and cellular states induced by the interaction wiring (Table 1; second column). For example, global changes in gene expression are measured with DNA microarrays [18], whereas changes in protein abundance [19], protein phosphorylation state [20] and metabolite concentration [21] can be quantified with mass spectrometry, nuclear magnetic resonance and other advanced techniques. Measurements made by DNA microarrays are currently the most comprehensive (every mRNA species is detected), high throughput (a single technician can assay several conditions per week), well characterized (experimental error is appreciable but understood) and cost-effective (whole-genome microarrays are purchased commercially for US$50–1000, depending on the organism). However, continued advances in protein labeling and separation technology are making measurement of protein abundance and phosphorylation state almost as feasible, with the primary barrier being the expense and expertise required to set up and manage a mass spectrometry facility. Measurement of metabolite concentrations, an endeavor otherwise known as metabonomics [22], is currently limited not by detection (thousands of peaks, each representing a different molecular species, are found in a typical nuclear magnetic resonance spectrum) but by identification (matching each peak with a chemical structure is difficult). Clearly, measuring changes in cellular state at the protein and metabolic levels will be crucial if we are to gain insight into not only regulatory pathways but also those pertaining to the cell's signaling and metabolic circuitry.

Extracting L1 models from the interaction scaffold

To arrive at an L1 model of a pathway or cellular process of interest, data on molecular interactions and states are integrated in a multi-tiered pathway mapping strategy (Fig. 1). First, the global molecular interaction scaffold is constructed from systematic measurements of protein–protein interactions, protein–DNA interactions and/or metabolic reactions. In the case of budding yeast, a maximal set might include 14 941 protein–protein interactions (catalogued in the DIP™ database), 5631 protein–DNA interactions (a combination of TRANSFAC® [23]

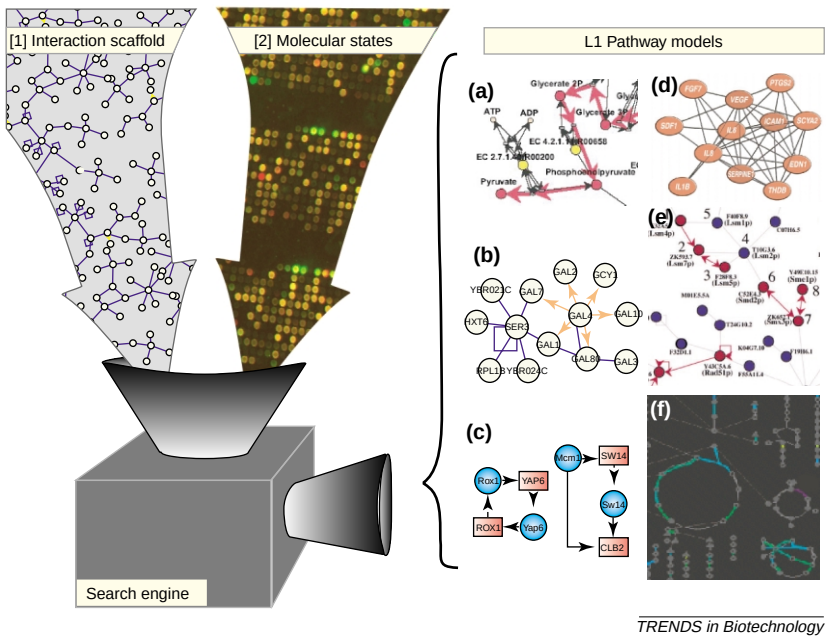

Fig. 1. Obtaining the Level 1 (L1) model. New pathway search engines generate L1 models by integrating two types of inputs: (1) an interaction database containing comprehensive information about the scaffold of known molecular interactions for a particular species or cell type; and (2) profiles of molecular and cellular states measured globally in response to particular conditions or perturbations. Given these inputs, the search engine seeks to identify modules of connected proteins in the scaffold whose states have been altered by perturbation. These 'activated networks' become prime candidates for further verification and modeling as important signaling and compensatory mechanisms controlling the cellular perturbation response. A range of recently reported methods (a–f; see text) subscribe to this general pathway discovery framework.

and Ref. [14]) and 599 enzymatic reactions (MetaCyc [24]). Second, this scaffold is filtered against changes in mRNA expression, protein expression and/or post-translational modifications recorded in response to different cellular perturbations. Networks within the interaction scaffold with mRNA or protein states that are significantly activated by perturbation are identified and mapped according to a computational search engine. The interaction pathways and complexes making up each 'activated network' in the scaffold constitute L1 models, which are then prime candidates for further verification and modeling as important signaling and compensatory mechanisms controlling the cellular perturbation response. The key advance of these searches is that, by integrating two complementary global datasets, it is possible to condense and partition the enormous quantity of data into a small number of relevant pieces suitable for lower level modeling.

Many instances of this general scheme have been reported in recent literature. For example, several groups [25–27] have used 'co-clustering' approaches to identify groups of proteins that are both expressed differently under similar sets of conditions and closely connected by the same network of interactions in the scaffold (Fig. 1a [27]; Fig. 1b [26]). In many cases, these 'expression-activated networks' correspond to well-known protein complexes, regulatory pathways or metabolic reaction pathways.

Other methods [14,28] use probabilistic approaches to match changes in gene expression with the transcription factors that are most likely to regulate them directly (Fig. 1c [14]). These methods start with a cluster of differently expressed genes and incrementally choose a small set of transcription factors that, by virtue of their levels and/or interactions in the scaffold, can maximally predict the observed levels of differential expression in the cluster. New transcription factors are added only if they lead to a sufficient increase in predictive power over the transcription factors already in the model.

Pathway searches can be performed using a wide range of scaffold types and search methods. Jenssen *et al.* [29] began with an interaction scaffold based not on physical interactions between proteins or proteins and DNA but on gene associations mined from journal abstracts indexed in PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) (Fig. 1d). Two proteins appearing together in the same abstract were linked by a direct interaction in the scaffold. Once again, by identifying connected regions of this scaffold in which genes were also co-expressed over one or more experiments, the group identified several 'literature clusters' of genes associated with B-cell activation. Matthews *et al.* [30] searched the yeast protein–protein interaction scaffold based not on gene expression data but on homology against a second such scaffold from *C. elegans*. The resulting sets of 'interologs' contained only those protein interactions that were present in both species (Fig. 1e).

Several software tools are now available for visualizing interaction scaffolds (Osprey, http://biodata.mshri.on.ca/; PIMRider®, http://pim.hybrigenics.com/; GenoMax™, http://www.informaxinc.com/; Cytoscape, http://www.cytoscape.org/; Pathway Tools, http://bioinformatics.ai.sri.com/ptools/). The Cytoscape framework [31] provides network visualization, layout and annotation, and clusters the network against expression data to generate activated (L1) network models. The PathwayTools component of the

MetaCyc metabolic pathway database [32] can superimpose enzyme expression levels on the map of biochemical reactions for a species, giving a good indication of which reaction pathways are most affected over a panel of growth conditions profiled by microarray (Fig. 1f).

Because DNA microarray technology is currently much more widespread than technologies for protein or metabolite profiling, the vast majority of these approaches have used gene expression profiling as the primary state measurement. Of course, pathway-mapping methods based on mRNA profiling alone capture just one facet of a much larger and more complex cellular response. As it becomes possible to measure cellular state at the protein and small-molecule level, we expect that algorithms similar to those described above will emerge. Currently, omitting this information from the analysis means that key signaling pathways are not mapped or are, at best, fragmented.

From L1 to L2 models

At the other end of the level spectrum (L2 models), one wishes to build models with predictive capability for cell behavior that are physicochemical in nature and based on low level molecular detail. For instance, consider an altered DNA sequence leading to a modified protein structure, which in turn yields an altered rate constant in a cell signaling process governing cell proliferation, differentiation or migration. Can we predict how that sequence change would propagate to a change in cell function? Several recent proof-of-principle demonstrations provide this predictive power to a limited degree. For example, an increase in neutrophil-precursor-cell proliferation, brought about by modifying a single amino acid residue in the mitogenic cytokine known as granulocyte colony-stimulating factor (GCSF), was successfully predicted *a priori* by a combination of molecular and cellular computational modeling [33]. Likewise, the effects of epidermal growth factor receptor ligand expression on embryonic tissue patterning in fly development have also been predicted [34]. Highly detailed physicochemical models (reaction kinetics and transport phenomena) of intracellular signaling networks are also beginning to emerge [35,36], although attempts to directly predict cell behavioral or tissue physiological functions from quantitative signaling pathway activities have not yet been undertaken.

In these studies, the crucial molecular network connectivities on which the models were built had previously been well established by many person-years of qualitative and quantitative biochemistry, molecular cell biology and developmental genetics. Thus, these systems were already ripe for low-level modeling. However, there are relatively few well-documented systems for which low-level computational modeling can be effectively pursued. A major goal and future challenge of systems biology must be to increase the throughput with which interesting and important biological problems can be brought to such a state.

Therefore, what steps can be productively taken between the L1 'interaction scaffold' modeling efforts described earlier and the L2 'physicochemical' modeling efforts noted here? What types of modeling approaches might lie between them? Some relevant lessons can be gleaned from circuit design in electrical engineering. Significantly, large-scale digital circuits are not comprehensively modeled at their low-level solid-state physical properties but instead are typically simulated across multiple layers of computational hierarchy. For example, Verilog® [37] is used to build L1 models that perform logic simulations of large digital circuits; its components are logic gates, memory units, timers, counters and clocks. At the next lower level, tools for circuit layout are used to specify the precise two- and three-dimensional geometry of logic gates on the silicon wafer. At a lower level still, tools such as Spice [38] simulate the analog behavior underlying each digital component; its components are basic electrical units such as resistors, capacitors, transistors and batteries. Finally, software packages such as Cadence® (http://www.cadence.com/) combine many of these functionalities into a single package, effectively bridging the gap between digital simulations, analog simulations and layout.

One might approach biomolecular networks in an analogous manner. Starting with L1 models representing the key components – DNA, mRNA and/or proteins – for a system of interest, a next step is to model their potential influences on one another's activities and on the cell, tissue or organ function of concern, even in the absence of detailed physicochemical information about the nature by which these influences are conducted. Bayesian network models appear to be very promising for this purpose; these have been applied with great success to gene regulatory networks [39,40] and, more recently, have been proposed for application to protein signaling networks [41]. This modeling framework provides insight into which of the topologically available interactions actually appear to be influential in the operational activity of the network. This consequently enables us to focus on subsets of network components for more in-depth physicochemical experimental measurement. Additional data on network

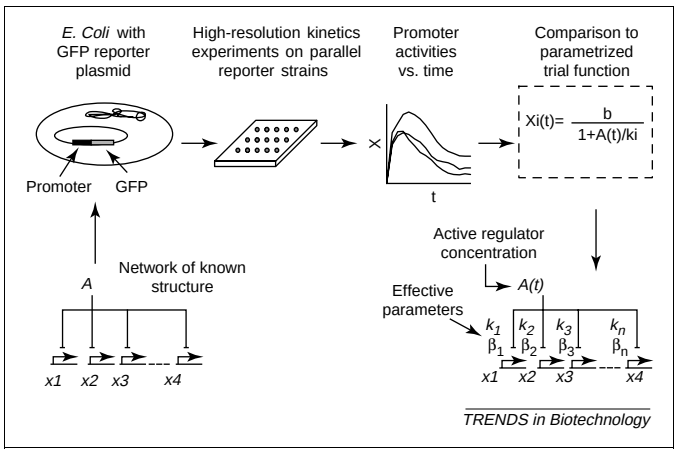

Fig. 2. One approach for bridging Level 1 and Level 2 models. Transcriptional regulatory networks in *Escherichia coli* are typically represented as high-level (Level 1) models, representing protein–DNA interactions between transcription factors and gene promoter regions. Ronen *et al.* [43] have described an integrated experimental–computational procedure for defining these models in greater detail by assigning kinetic parameters (numbers on the interactions) that capture the quantitative dynamics of the network. Reproduced with permission from Ref. [43]. (©2002 National Academy of Sciences, USA).

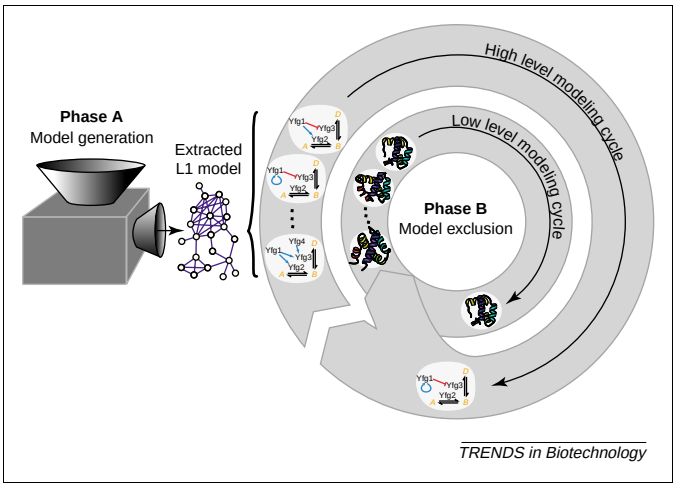

Fig. 3. Driving the modeling cycle. Level 1 (L1) models are extracted from the molecular interaction scaffold by systematic phase A experiments. In phase B, directed experiments are performed to verify particular models while excluding others. In general, models at a more abstract level are consistent with many possible models at the next level of detail, so that a well-supported L1 model might be instantiated as its corresponding set of Level 2 models in additional cycles of phase B experiments.

component states (for instance, in response to multiple driving conditions such as stimuli and treatments) then permit component–component and component–outcome influences to be cast in an even-more-explicit model framework, such as a Boolean algebra formalism. In an intriguing application of Boolean modeling, a signaling and cell cycle control circuit (including both mRNA and protein components) was used to model the simultaneous impacts of growth factor and extracellular matrix stimulation on cell proliferation [42]. Thus, logic rules for network component influences can be elucidated and used to interpret and predict cell behaviors. Once this degree of cause and effect can be determined, the substantial effort needed to produce an ultimate, physicochemical model is easily achieved. Ronen *et al.* [43] have provided an excellent example of progressing from an identified L1 model (in this case, governing gene expression in the *Escherichia coli* SOS DNA repair system) to a physicochemical model with detailed reaction kinetic parameters (Fig. 2).

Coverage versus leverage: the two phases of experimental design

The systematic pathway mapping approach has important implications for experimental design. Just as pathway modeling has multiple levels (L1 and L2), so experiments to specify and validate these models are performed in distinct modes or phases (A and B; see Fig. 3). The goal of the initial phase (phase A) is to perform experiments that are comprehensive in nature, to stimulate the pathway of interest broadly and to perturb all of its components. We might consider all perturbations of a certain type, such as all single gene knockouts or all drugs from a library, and then predict and experimentally observe the effects of each perturbation on the pathway in question. The result of this initial phase of experiments is an L1 pathway model. However, any single L1 model corresponds to many L2 models at the next lower level of detail. Thus, an ensuing experimental phase (phase B) aims to target specific components, interactions and other parameters, guided by the L1 model. In designing these directed experiments, the perturbations resulting in the most different simulation outcomes among the models are the most likely to distinguish between the models and thus to lead to the largest information gain about the biological system as a whole. In short, the initial experimental phase is concerned with generating a rich ensemble of model hypotheses, whereas subsequent phases focus on testing and distinguishing between these various models – that is, phase A experiments seek coverage and phase B experiments seek leverage.

For instance, in a recent study of metabolic and regulatory control [44], the yeast galactose-utilization pathway was stimulated by adding or withholding galactose from the growth medium and its components perturbed by deleting each galactose metabolic enzyme and regulatory gene in a series of knockout strains. In phase A experiments, gene-expression changes caused by each of these perturbations were measured using DNA microarrays. Although the observed changes were generally in good agreement with those predicted by the model, several observations were strikingly inconsistent; in these cases, new hypotheses were suggested to explain these discrepancies, and the validity of these hypotheses tested using a second phase of experimental perturbations directed to target the components in question more specifically.

A relatively crude but instructive example of phase B experiments is found in work by Palecek *et al.* [45]. Here, a mathematical model had been previously proposed (based on relatively simple kinetic, transport and mechanics processes) for how key molecular properties govern the speed of cell migration across a ligand-coated substratum [46]. This model had made non-intuitive theoretical predictions about the effects of cell receptor expression, substratum ligand density and receptor–ligand binding affinity. Palecek *et al.* used molecular cell biology and biochemistry techniques to enable quantitative variation and measurement of each of these parameters independently and systematically in an engineered cell line to give a rigorous and successful experimental verification of the component-to-system predictions. The capabilities of molecular genetics, pharmacology and biomaterials to vary molecular component properties systematically and quantitatively in cellular systems are now very strong; thus, once specific predictions are generated from L2 model hypotheses, there is no shortage of means to test them.

Perspectives

Where are pathway modeling efforts headed in the future? It is revealing that, outside biotechnology and the pharmaceutical industry, nearly every sector of manufacturing depends on computer simulation and modeling for product development. Circuit manufacturers rely on computer aided design (CAD) tools to model the wiring of transistors and other circuit components as well as their two- and three-dimensional layouts on the silicon wafer. Likewise, automotive engineers can estimate

how many miles per gallon to expect from the next model long before it is built on the assembly line, all through extensive CAD simulation [e.g. software from LMS International (http://www.lmsintl.com/) or Adams MSC (http://www.adams.com/)].

It is worth noting that, although instructive in their successes, these 'mainstream' computer modeling efforts have not been free from difficulty. For instance, even in disciplines in which modeling methodologies are well established, small discrepancies between the model and reality can compound to cause predictions that are grossly inaccurate. Moreover, as models become extremely large and complex, it becomes infeasible to evaluate their full range of inputs and resulting behaviors. Thus, although we will undoubtedly benefit from prior experience in mainstream engineering, we should also expect that substantial new research will be required to develop powerful CAD tools for biology.

Might pharmaceutical companies one day use computational modeling to simulate the effects of drugs on cells before proceeding to trials in human subjects? In the field of chemoinformatics, software tools are increasingly popular for predicting drug candidates that are likely to bind protein targets, and companies such as Physiome Sciences and Entelos market computer models of specific disease pathways for drug discovery and development, including clinical trial interpretation and design. However, the recent advances in systematic pathway mapping described here suggest a compelling use for computational tools at a different step in the drug development pipeline: assessing toxic side effects. Because the molecular interaction scaffold offers an unbiased, high-level view into many pathways, a search of this scaffold could indicate which compensatory pathways are perturbed by a drug in addition to its intended disease-pathway target. Drugs activating pathways associated with stress and other toxic effects could be eliminated from further consideration. Given that more than six out of every seven drug candidates that undergo human testing ultimately meet with failure [47], such software would act as a much-needed additional filter between high-throughput screening for drug candidates and the time-consuming, costly follow-up of human testing. In this regard, the emergence of comprehensive, high- to low-level modeling strategies will provide helpful impetus for the acceptance of computational modeling tools in both the pharmaceutical industry and in biology as a whole.

Acknowledgements

We are indebted to J. Lauffenburger, J. Doyle and O. Ozier for inspiring discussions and for very helpful comments on the manuscript. T.I. was generously supported by a research fellowship from Pfizer; D.L. was supported by grants from DARPA and NIGMS.

References

1 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
2 Guarente, L. (1993) Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet.* 9, 362–366
3 Karp, P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science* 293, 2040–2044
4 Hucka, M. *et al.* (2002) The ERATO systems biology workbench: enabling interaction and exchange between software tools for computational biology. *Pac. Symp. Biocomput.* 1, 450–461
5 Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246
6 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
7 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
8 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
9 Rain, J.C. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–215
10 Walhout, A.J. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122
11 Bader, G.D. *et al.* (2001) BIND – the biomolecular interaction network database. *Nucleic Acids Res.* 29, 242–245
12 Xenarios, I. and Eisenberg, D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol.* 12, 334–339
13 Deane, C.M. *et al.* (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* 1, 349–356
14 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
15 MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763
16 Zhu, H. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science* 293, 2101–2105
17 Haab, B.B. *et al.* (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* 2 research0004.1-00041.13
18 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
19 Gygi, S.P. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999
20 Zhou, H. *et al.* (2001) A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* 19, 375–378
21 Griffin, J.L. *et al.* (2001) Choline containing metabolites during cell transfection: an insight into magnetic resonance spectroscopy detectable changes. *FEBS Lett.* 509, 263–266
22 Nicholson, J.K. *et al.* (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* 1, 153–161
23 Wingender, E. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* 29, 281–283
24 Karp, P.D. *et al.* (2002) The MetaCyc database. *Nucleic Acids Res.* 30, 59–61
25 Jansen, R. *et al.* (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.* 12, 37–46
26 Ideker, T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 (Suppl.), S233–S240
27 Hanisch, D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics* 18 (Suppl.), S145–S154
28 Pe'er, D. *et al.* (2002) Minreg: inferring an active regulator set. *Bioinformatics* 18 (Suppl.), S258–S267
29 Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28
30 Matthews, L.R. *et al.* (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res.* 11, 2120–2126
31 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (in press)
32 Karp, P.D. *et al.* (2002) The Pathway Tools software. *Bioinformatics* 18 (Suppl.), S225–S232
33 Sarkar, C.A. *et al.* (2002) Rational cytokine design for increased lifetime and enhanced potency using pH-activated 'histidine switching'. *Nat. Biotechnol.* 20, 908–913
34 Shvartsman, S.Y. *et al.* (2002) Modeling and computational analysis of EGF receptor-mediated cell communication in *Drosophila oogenesis*. *Development* 129, 2577–2589
35 Bhalla, U.S. *et al.* (2002) MAP kinase phosphatase as a locus of

flexibility in a mitogen-activated protein kinase signaling network. *Science* 297, 1018–1023
36 Schoeberl, B. *et al.* (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* 20, 370–375
37 Golze, U. (1996) *VLSI Chip Design with the Hardware Description Language VERILOG: an Introduction Based on a Large RISC Processor Design*, Springer-Verlag
38 Banzhaf, W. (1992) *Computer-Aided Circuit Analysis Using Pspice*, Prentice–Hall
39 Friedman, N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620
40 Hartemink, A.J. *et al.* (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 1, 422–433
41 Sachs, K. *et al.* (2002) Bayesian network approach to cell signaling pathway modeling. *Sci. STKE* 148, E38 http://www.stke.org/cgi/content/full/sigtrans;2002/148/pe38
42 Huang, S. and Ingber, D.E. (2000) Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* 261, 91–103
43 Ronen, M. *et al.* (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U. S. A.* 99, 10555–10560
44 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science* 292, 929–934
45 Palecek, S.P. *et al.* (1997) Integrin–ligand binding properties govern cell migration speed through cell–substratum adhesiveness. *Nature* 385, 537–540
46 DiMilla, P.A. *et al.* (1991) Mathematical model for the effects of adhesion and mechanics on cell migration speed. *Biophys. J.* 60, 15–37
47 Flanagan, A. *et al.* (2001) *A Revolution in R&D: How Genomics and Genetics are Transforming the Biopharmaceutical Industry*, Boston Consulting Group