The dynamics of molecular networks: applications to therapeutic discovery

Roland Somogyi and Larry D. Greller

Recent advances in biomedical research, genomics and bioinformatics have given the pharmaceutical and biotechnology industries new promises and expectations: providing effective cures for complex diseases, discovering and prioritizing drug targets more efficiently, eliminating toxic and ineffective compounds earlier and delivering the right drug therapy to the appropriate patients. Ultimately, the biomedical information generated today must be transformed into integrated predictive models that can be consulted for decision-making in drug discovery, efficacy and toxicity screening and individualized therapy. Here we describe how models that capture different aspects of network dynamics can be generated and applied in disease pathway identification, drug screening, diagnostics and individualized therapy.

Roland Somogyi and
***Larry D. Greller**
Molecular Mining Corporation
128 Ontario St
Kingston
Ontario K7L 2Y4
Canada
tel: +1 613 547 9752
fax: +1 613 547 6835
*e-mail: lgreller@molecularmining.com
web: www.molecularmining.com

▼ Most diseases cannot be explained by one genetic mutation, the action of a single gene-product or pathway. Genes interact in complex ways through multiprotein complexes, feedback regulation, and networks of interacting pathways. Sequence variations in multiple genes can yield variants of proteins that significantly alter disease susceptibility and drug response (phenotype). The vast amount of data on molecular structure, interaction and activity information being acquired in the biomedical research domain brings forth a picture of intricate molecular networks that underlie biological function, complex phenotypes and diseases.

Much of the high-value information for therapeutic target identification, prioritization and manipulation resides in the combinatorial, and often non-linear, interactions between the many components organized into networks (whether regulatory, transcriptional, metabolic, signaling, inflammatory, immunological, or physiological, and so on). In building and refining useful dynamic network models, it will become increasingly important to identify, and then incorporate, the information derived from the explosion of experimental data being generated by the large-scale gene- and protein-expression technologies. As more temporal data becomes available, and the recognition of its use for drug discovery becomes more widespread, data-driven dynamic modeling will play a pivotal role in the reduction of data to useful information concerning:

- Stimulus–response interactions;
- Prediction of new targets based on pathway context;
- Potential use of combinatorial therapies;
- Pathway responses including the understanding of reactive or compensatory behavior;
- Stress and toxic response mechanisms;
- Off-target effects of therapeutic compounds;
- Pharmacodynamics;
- Characterization of disease states by dynamical behavior;
- Gene- and protein-expression signatures for diagnostics; and
- Design of optimized time-dependent dosing regimens.

The development of high-throughput molecular assay technologies, as well as breakthroughs in information processing and storage technologies provides integrated views of biological and medical information. Scientists are on the verge of applying genome-wide RNA- and protein-expression measurement technologies to generate detailed profiles of the molecular activities that underlie biological function in health, disease and response to drugs and therapies. We now have the complete sequence of the human genome, and are building databases of human genetic variation, haplotypes, single nucleotide polymorphisms (SNPs), gene expression, and their association with disease, therapeutics and

1359-6446/01/$ – see front matter PII: S1359-6446(01)02096-7

Table 1. Examples of network models predicting biological outcomes across a wide range of applications

Data	Results	Applications
NCI60 data – comparing baseline gene expression to drug sensitivity in 60 cancer cell lines	Higher-order and non-linear rules are required for predicting drug response from baseline gene expression data (Fig. 1)	Individualized medicine and drug target discovery: selecting effective drug treatments based on expression information of gene tumor biopsies; genes predictive of drug response could also be targets themselves
Cancer and non-cancer tissue samples	Identification of sets of 8–10 genes required for robust and accurate distinction of cancer and non-cancer tissue (Fig. 2)	Drug screening: robust, expression-based assays for effective drug response
SNP profiles of individuals with cardiovascular disease	Sets of up to eight SNPs provide more accurate separation of diseased and non-diseased individuals (Fig. 3)	Diagnostics, clinical trials and individualized medicine: SNP-based stratification of cardiovascular disease, disease susceptibility screening
Blood coagulation cascade	Detailed modeling of the coagulation cascade enables in-depth exploration of drug-target dynamics and response to treatment (Fig. 4)	*In silico* drug target discovery: efficient prioritization of drug targets based on optimal behavior in model
Gene expression time-series in CNS development and response to injury	Reverse engineering of gene network from activity data; model can be used to generate network interaction graph, and to recreate and explore detailed behavior of network dynamics (Fig. 5)	*De novo* discovery of gene function, *in silico* drug target discovery: novel gene functions and interactions can be directly postulated through reverse engineering of networks; network model dynamics can be explored in detail to determine optimal sets of perturbation points for therapeutic intervention

Abbreviations: NCI60, National Cancer Institute database of drug sensitivity information from 60 cells lines; SNP, single nucleotide polymorphism.

drug responses[1–4]. Biomedical research results, including biomolecular interactions, pathways and molecular disease associations, are increasingly available in databases enabling systematic data mining. Moreover, databases of clinical outcomes are being coupled to gene sequence and activity profiles, providing the ultimate links for enabling individualized medicine.

Thus, major medical breakthroughs in providing genuine cures for diseases will depend on an increased understanding of the dynamics of biological function in complex diseases, also taking into account individual variation. How do complex molecular networks change over time, and which combinations of links are crucial for targeting, to therapeutically guide the system from a diseased state to a healthy or non-detrimental state? Computational molecular networks are needed to capture the information from experimental exploration of gene function, gene activity and genetic variation (see Table 1 for examples). These models can be built from known interactions and pathways or inferred ('reverse engineered') from activity data for the discovery of unbiased novel gene interactions and functions. Profiles of genetic variation would then be used to build the individualized models.

Well-developed models to guide experimentation

Ultimately, effective computational models of gene or molecular networks will permit detailed, *in silico* experimentation into how genes interact and could be perturbed for the targeted manipulation of the phenotype, with phenotypes being represented as network attractors[5,6]. Such *in silico* studies will provide key sets of results on the basis of which 'wet-lab' experiments can be designed to validate and finalize the development of complex molecular therapeutics. Although all experimentation could, in principle, be carried out in a wet-lab setting (except for cases in which particular molecular perturbation technologies do not yet exist), the number of experiments required to explore the combinatorial interactions of many genes would be truly enormous and would exceed any reasonable science budget or research timeline. We will, therefore, rely on well-developed models of genetic and molecular networks for detailed *in silico* exploratory experimentation

for drug target and drug-action pathway discovery. The insights gained from these analyses will allow us to minimize costly wet-lab experimentation towards the validation of the most promising and valuable hypotheses.

Predicting network outcomes using direct inference

Although the genome contains the blueprint for molecular networks, at present, only the complex cellular machinery within the context of the organism's momentary functional state can decipher this code into the structures and functional processes of the organism. For this reason, accurate measurement of the molecular activity profiles and their dynamics, together with genomic sequence information, are required to obtain a useful, in-depth understanding of molecular networks. At the lowest level, genetic variation determines differences in the outcomes, that is, the phenotypes. At a higher level, the impact of external influences on an individual, together with their genetic constitution, will determine network outcome with respect to health and disease. The importance of the temporal component in these complexities has been long appreciated. For example:

> 'Even [Ernst] Mayr, however, does not clearly express the most fundamental and basic characteristic of phenotypes, namely that they change in time. The phenotype is the name given to the results of the activities of genes. In a very simple organism, these activities may be carried out in a relatively short period, resulting in the formation, for instance, of a certain number of proteins. In more complex organisms, these proteins themselves interact with one another and with other substances so that it is a long and complex sequence of processes that the genes set in motion. But in either case a time duration, whether short or long, is an essential component of a phenotype.'[7]

Even when we make functional predictions based on sequence or structural data, we always imply dynamics, because all biological function requires action over time.

There are many available and potential sources of data and analysis methods that will allow predictions of insightful, straightforward network outcomes. In the simplest case, a dataset combining a set of genetic or phenotypic variables (input) associated with a phenotypic outcome (output) provides an opportunity to predict the outcome. The minimum requirement for this to be effective is that the input variables represent an association with the output variables that is unlikely to occur just by chance. In mechanistic terms, such an association could be the product of the input variable connecting to the output effect directly through a causal molecular interaction, or indirectly, through an associated, non-causative, 'marker' effect. Both might be useful, certainly for diagnostic purposes. However, in terms of drug target discovery, it is necessary that the predictive variables bear a causal relationship to the outcome.

Prediction of drug response from gene expression

Drug response can be predicted from baseline gene expression data according to higher-order and non-linear correlations. For example, in the case of the NCI60 study in which cancer drug–response data were compared with basal gene expression measurements[8], many associations that predicted drug response from gene expression were computationally discovered (Kotlyar, M., unpublished). Often these predictive relationships take complex, non-linear shapes involving more than one gene. For example, in the prediction of the fluorodopan response (Fig. 1a), we discovered a relationship akin to the logical 'Exclusive OR' (for two findings being compared that have opposite labels, e.g. low or high, both do not have the same label). If one of the genes was high in expression, and another was low in expression, or vice versa, this rule predicts insensitivity to the drug. Highly drug-sensitive cell lines were predicted only if both genes exhibited either high expression or low expression. If only one of the genes exhibited high or low expression, there was no predictive power, that is, no separation of high and low sensitivity cell lines in projections on the individual gene axes. Overall, most of the genes implicated in predicting drug response were part of gene pairs; singleton genes did not perform as well in predictions (Fig. 1b). These observations suggest a network nature of regulation, in which genes interact in complex, often non-linear ways to produce an outcome. By choosing data-mining methods that take these types of relationships into account, we can infer fundamental network, input–output relationships. Rules linking the baseline expression of sets of genes to cancer drug response could be applied to individualized therapy, for example, gene expression patterns assayed from tumor biopsies could provide invaluable information for the selection of a drug appropriate for an individual patient.

More robust drug screening using higher-order gene sets

In some cases, there can be a gradation of predictive power of the input variables. Although a single input variable might provide partial predictive power over the outcome, additional variables, either individually or in combination, can result in a more in-depth or robust prediction. For example, in an analysis of gene expression data, genes and gene sets have been identified that can distinguish cancer

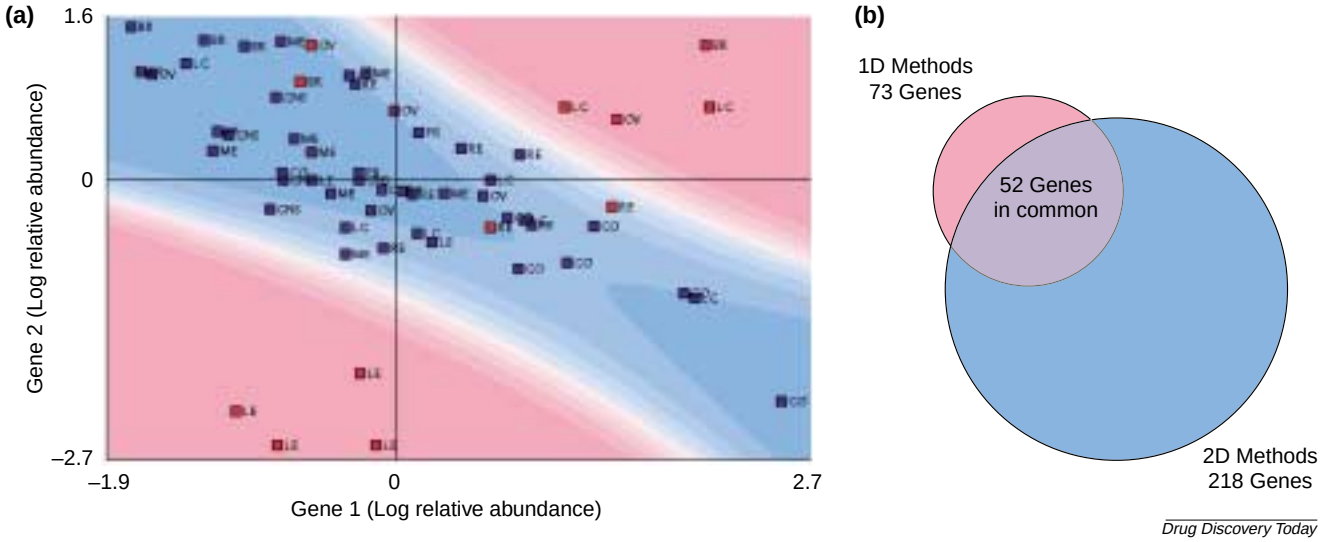


Figure 1. Analysis of gene expression–drug sensitivity data for 60 cancer cell lines. **(a)** Each point represents a cell line, with its location specified by the relative expression of two genes (x and y coordinates). The points are coloured to represent the cell line's relative rate of growth-inhibition response to fluorodopan. Squares show the experimentally measured values for specific cell lines. Red squares represent high growth inhibition (i.e. –log (GI_{50}) at least one standard deviation larger than the –log (GI_{50}) averaged across all 60 cell lines. Remaining squares are colored blue. The coloring of the background shows the decision surface generated by the two fitted distributions: red areas are classified as high response and the blue areas as low response. The letters beside each point indicate the tissue type of a cell line. Gene 1 expression alone is uncorrelated with sensitivity because a positive value (on a logarithmic scale) can correspond to either high or low sensitivity, and a negative value (on a logarithmic scale) can correspond to either high or low sensitivity; the same is true of gene 2. However, when either negative or positive values co-occur for genes 1 and 2, sensitivity is high. When expression of gene 1 and gene 2 have opposite values (i.e. negative and positive), sensitivity is low. We, therefore, obtain a rule for the correlation of the pair of genes 1 and 2 with fluorodopan sensitivity. Abbreviations (cancer tissues): BR, breast; CNS, central nervous system; CO, colon; LC, lung; LE, leukemia; ME, melanoma; OV, ovarian; PR, prostate; RE, renal. **(b)** Venn diagram comparing statistically significant results from one-dimensional (1D) and two-dimensional (2D) methods. 1D methods identified a much smaller number of genes than 2D methods: 73 compared with 218. The majority of the genes found by 1D methods was also identified by 2D methods (52 of 73).

from non-cancer tissue samples [collaboration between Avalon Pharmaceuticals, (Gaithersburg, MD, USA) and Molecular Mining Corporation]. Although individual genes provide accurate separation, it was found that, when exposed to noise, only longer gene-sets of 8–10 genes provided robust predictions in which accuracy was not compromised (Fig. 2). Such genes are currently being used in anti-cancer drug screening by Avalon Pharmaceuticals. By identifying predictive gene sets, we achieve the benefits of networked control, namely the robustness and redundancy inherent in distributed control over several genes.

Predicting disease using higher-order SNP motifs

On occasion, one input variable might predict one output variable according to a simple relationship. This is the case in monogenic disease, in which one gene is solely responsible for the disorder, or a disease in which one phenotypic marker bears a direct relationship to the outcome. However, most prevalent diseases are caused by a complex combination of genetic and environmental factors. In this case, it is necessary to accurately capture the sets of variables either directly responsible or indirectly associated with the underlying cause. The identification of these associations is not trivial, neither in terms of selecting the appropriate variables to be assayed, nor in terms of reliably determining the combinatorial relationships between the predictive variables that robustly determine the outcome. There are subtle but significant differences in the predictive power of cumulative predictive variables versus genuinely combinatorial variables in terms of information content and predictive content.

To maximally exploit combinatorial predictive power, searches cannot just be conducted by concatenating the best individual predictors; one must begin by examining full sets right from the start. For example, in a case of sets of gene variants (SNPs) relating to cardiovascular disease outcome (Fig. 3), the separation of diseased versus normal samples becomes more accurate as we look at higher-order SNP motifs, up to eight SNPs per motif. Individual SNPs within larger sets usually have only low odds-ratios (Fig. 3 legend) by themselves. The interactions of the functional variations for which these SNPs are responsible could be further investigated in targeted experimentation on these genes.

In silico exploration of drug targets

Modeling of integrated network dynamics based on known interactions

In silico models are gaining acceptance as advanced commercial methodologies that can deliver important insights and value to pharmaceuticals research organizations (e.g. http://www.physiome.com; http://www.entelos.com)[9,10]. The interactions of molecular species or clearly defined phenotypic variables have been studied in depth for select pathways (e.g. http://www.hort.purdue.edu/cfpesp/models/), providing at minimum an interaction diagram between these variables and, at best, a valuable estimate of the quantitative interaction strengths. This information can be used to construct descriptive mathematical models, from some spanning metabolism, gene regulation and signalling, to realistic models of cell physiology, organ function and even disease behavior. Such models, usually based on differential equations, can capture known interaction parameters. In some biological cases there are good physical reasons to believe that the concentrations of transcripts, gene products, effector proteins and metabolites are so low that the continuum approximation assumptions upon which differential equation models rely would be suspect. Even here, there still are potent applicable *in silico* modeling genre available, for example, Boolean models[11–13] or the stochastic dynamics models advanced by McAdams and Arkin[14,15]. When not all the parameter values are known, estimates can be made. Ultimately, the model predictions must be compared with experimental observations to determine the appropriateness of the model architecture and parameter-value estimates.

For example, mathematical modeling was used with a pharmaceutical development program at the former SmithKline Beecham (SB; now GlaxoSmithKline, King of Prussia, PA,

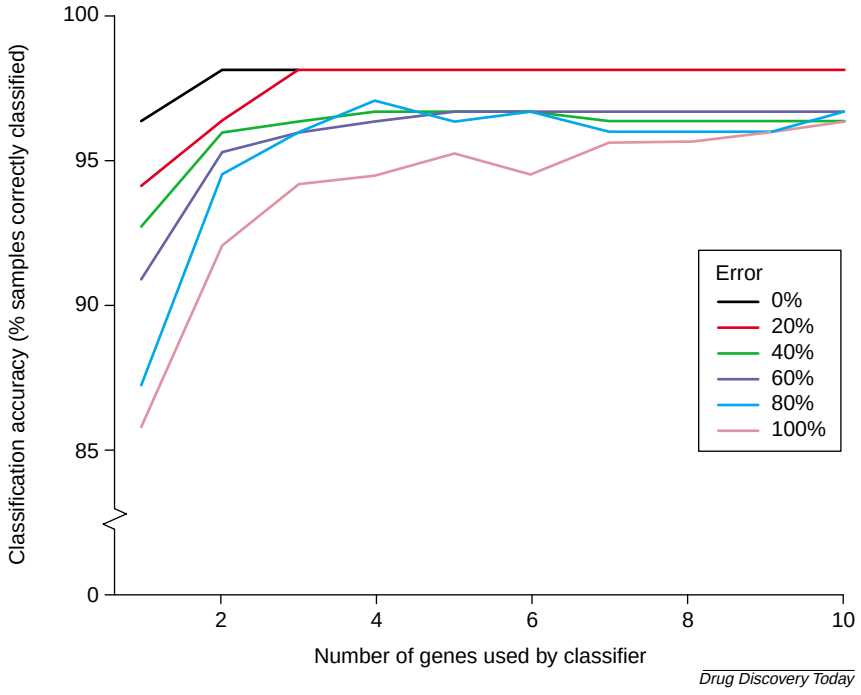


Figure 2. Robustness of predictive gene sets identifying cancer. Under 'noiseless' conditions, maximal accuracy in correctly classifying cancer can be achieved by a combination of two genes. However, as measurement noise increases to 100% (one standard deviation of the measurement noise in the gene expression assay), sets of 8–10 genes are required to accurately predict cancer samples.

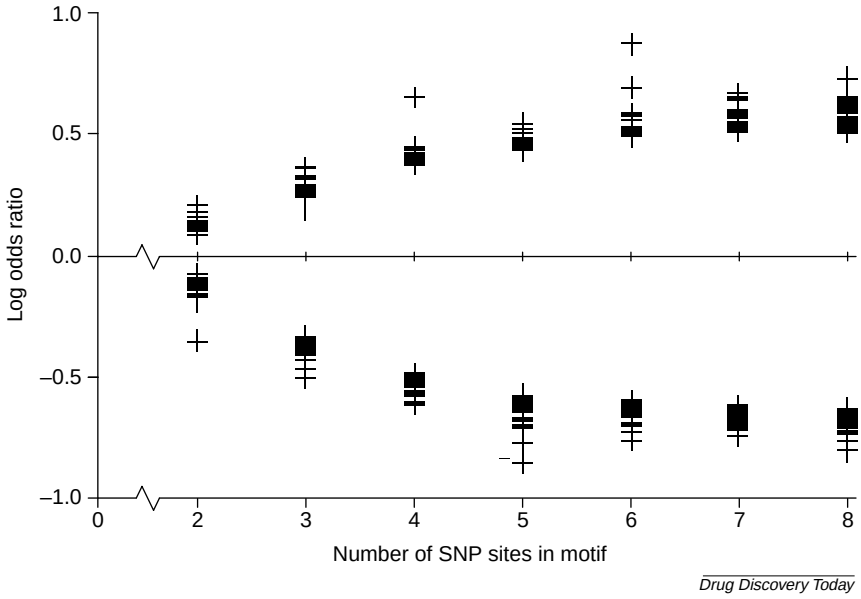


Figure 3. Higher order single nucleotide polymorphism (SNP) motifs provide the best separation of diseased from normal samples. The accuracy of separating cardiovascular disease samples from normal samples is reflected in the 'odds ratio' where: odds ratio = (present in cases/absent in cases)/(present in controls/absent in controls). Good separation is given for log odds-ratio $\gg 0$ if the motif is enriched in the disease cases, and log odds-ratio $\ll 0$ if the motif is enhanced in the control samples with respect to the diseased samples.

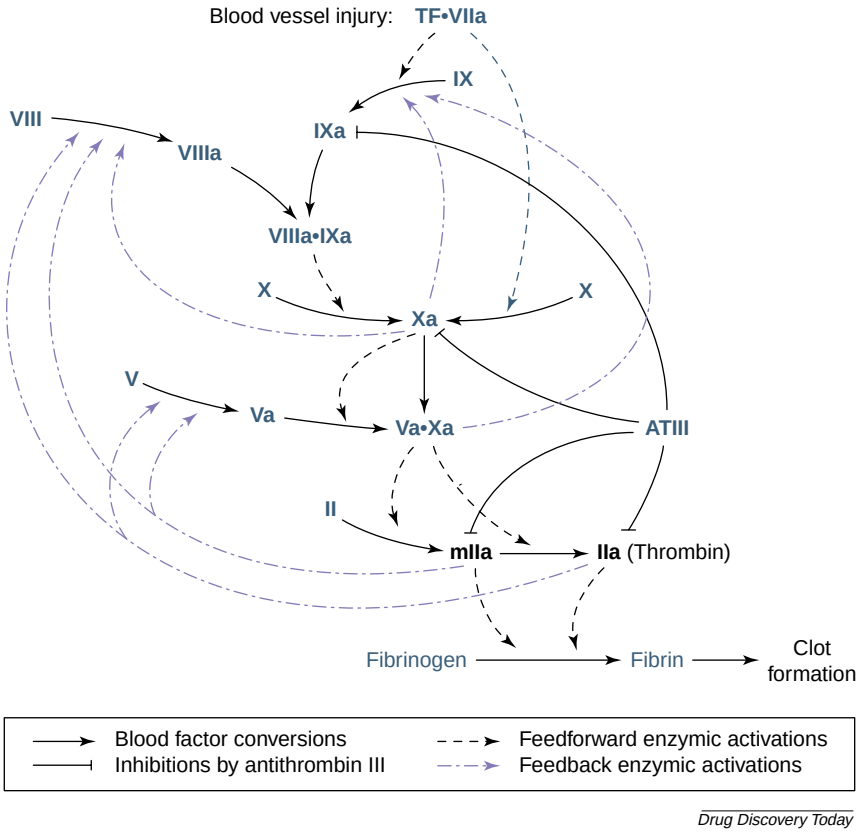


Figure 4. Coagulation cascade biochemical network, from tissue factor to thrombin. The figure shows the network of blood-factor enzyme activations leading to the production of the clot-formation enzyme, thrombin. The cascading succession of enzyme activations are initiated by the exposure of tissue factor•VIIa to plasma upon blood vessel injury. The catalytically active forms are indicated by the traditional 'a' suffix. The standard textbook notation for the molecular factors is employed[17,18].

USA) to study the dynamical behavior of the biochemical network, anti-Factor IX, in the context of the blood coagulation cascade. A fully humanized anti-Factor IX antibody entered clinical trials under this program[16]. To address certain research questions motivated by the program, an SB team employed a network mathematical model: a customized modification of the Jones and Mann model[17] for the experimentally reconstituted kinetics of the coagulation cascade from formation of tissue factor•VIIa, as initiated by blood vessel injury, to the production of thrombin[18]. The model provided a reaction-kinetics basis for understanding the limited prolongation of time-to-clot when an upstream network component, Factor IX, is inhibited. It explained how the observed limited anticoagulation is a kinetics property of the target (Factor IX) because of its position and role in the overall network. The model also predicted kinetics-based limited anticoagulation when Factor VIII (a cofactor of activated Factor IX) is selected as an alternative target. Fig. 4 shows a schematic diagram of the coagulation cascade enzyme-activation network. Results of the mathematical model shown diagrammatically in Fig. 5 predicts the important limited anticoagulation effect when the Factor VIIIa–Factor IX complex is inhibited, even by 100%, in contrast to the massive anticoagulation possible with heparin. The complicated basis for these dramatically different kinds of outcomes versus dosing was readily understood by studying the dynamics of the reaction network.

De novo exploration of gene function

Generating models of network dynamics by reverse engineering

High-throughput measurement technologies are generating large amounts of molecular activity data (e.g. RNA and protein expression), which can be profiled in terms of responses to drugs, targeted gene perturbations, and for the characterization of temporal responses[19–21]. These large datasets contain much implicit information on the structure of molecular networks. However, this information is hidden, and requires advanced algorithms and mathematical methods to enable the reverse engineering of these networks[22], which promise the discovery of hitherto unknown gene interactions and pathways.

The case of the idealized Boolean 'gene' networks using the REVEAL™ algorithm has shown that it is possible to fully reverse-engineer networks[23]. Using a combination of data mining and visualization techniques, Arkin and Ross were able to reverse-engineer the essential glycolytic reactions from metabolite activity data in an *in vitro* assay[24]. For the case of yeast steady-state gene expression data in response to gene knockouts (at a single time-point), Onami and colleagues successfully applied the difference-based regulation finding (DBRF) method[21,25] for the reconstruction of a directed network graph. They were able to successfully infer major steps of the mitogen-activated protein kinase (MAPK) pathway, and identified several additional candidate regulatory steps.

Direct reverse-engineering of quantitative network dynamics from expression data was demonstrated by D'haeseleer and colleagues[26], using CNS developmental

and injury response-time series for 70 genes. Remarkably, a linear model was found to be adequate in providing a network solution that was able to recreate the training data with great accuracy (Figs 6,7). Moreover, the matrix of gene–gene interaction coefficients was found to be sparse, in agreement with the general biological expectation that gene networks are far from 'fully wired'. Furthermore, within the graph of 70 interacting genes, strong connections were found between genes from the γ-aminobutyric acid (GABA) signaling family, including GABA receptors and the GABA synthesizing enzyme, glutamic acid decarboxylase (GAD), which appear biologically plausible (genes within a signaling family are expected to regulate their expression through positive and negative feedbacks; Fig. 7). Although this model was shown to be a mathematically accurate platform for calculating the dynamics of expression for this training set, behaviors of this model for time series outside this training set have not been tested. Ultimately, such models need to be explored *in silico* through perturbation studies, which in turn need to be validated in additional laboratory experiments to provide confidence in the model.

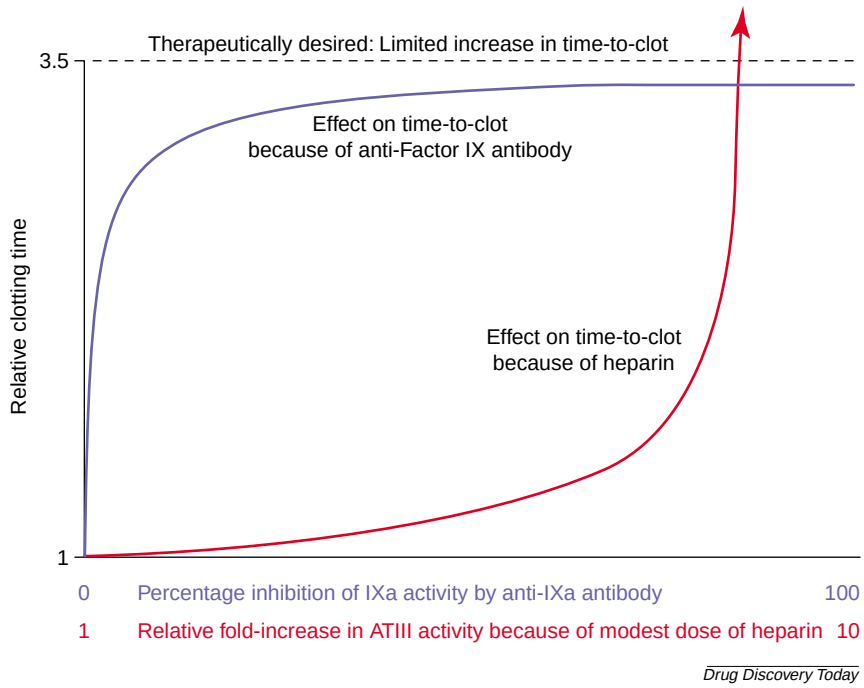


Figure 5. Results from the network mathematical model predict experimental outcomes: inhibiting the VIIIa•IXa complex (e.g. by anti-Factor IXa antibody) causes therapeutically desired limited anti-coagulation, whereas heparin dosing can cause unlimited anti-coagulation (because of the accelerated inactivation of Xa and thrombin in addition to IXa by heparin-activated antithrombin III). The ordinate represents anti-coagulation effect in terms of prolongation of time-to-clot relative to the standard uninhibited state of the reaction scheme. The abscissa represents, respectively, either (1) the percent inhibition (0–100%) of the catalytic activity of the VIIIa–IXa complex by anti-Factor IX antibody, or (2) the relative fold-increase in antithrombin III (ATIII) catalytic activity (1–10-fold) because of modestly increasing the dose of heparin. Results were obtained by numerical integration of the non-linear ordinary differential equations obtained by standard methods from the detailed chemical reaction scheme defining the network[17,18].

Understanding and prediction of observed phenomena

> 'A proven tool in the quantitative analysis of complicated systems, whether living or not, is mathematical modeling. Simply put, mathematical modeling is really a way to formulate hypotheses in a framework that allows their full implications to be recognized. Mathematical equations are merely the language by which the hypotheses are described and their implications communicated. When systems are simple, the mathematical language can be discarded as superfluous. However, the history of many scientific disciplines indicates that discarding it for complicated systems leaves one a club short in the bag; that is, examination of hypotheses for such systems will be incomplete without use of the full range of tools. As Maddox[27] writes '…the neglect of quantitative considerations [in molecular biology] may well be a recipe for overlooking problems inherently of great importance'[28]

It has been clear for a long time that both qualitative and quantitative predictions are crucial in biology and medicine, in theory and in practice. Dynamical models of temporal biological behavior[29–43] can provide the conceptual, quantitative, and explanatory linkages between the observed natural phenomena and their understanding, prediction and, on occasion, beneficial manipulation through the design of therapeutic interventions. It could even be argued that the drive to extract high value from large-scale temporal data will necessitate the reverse-engineering of network models. Network models reverse-engineered from the massively accumulating data could become the primary practical means by which huge dimensionality and astronomical combinatorial possibilities are reduced to manageable numbers of key organizing

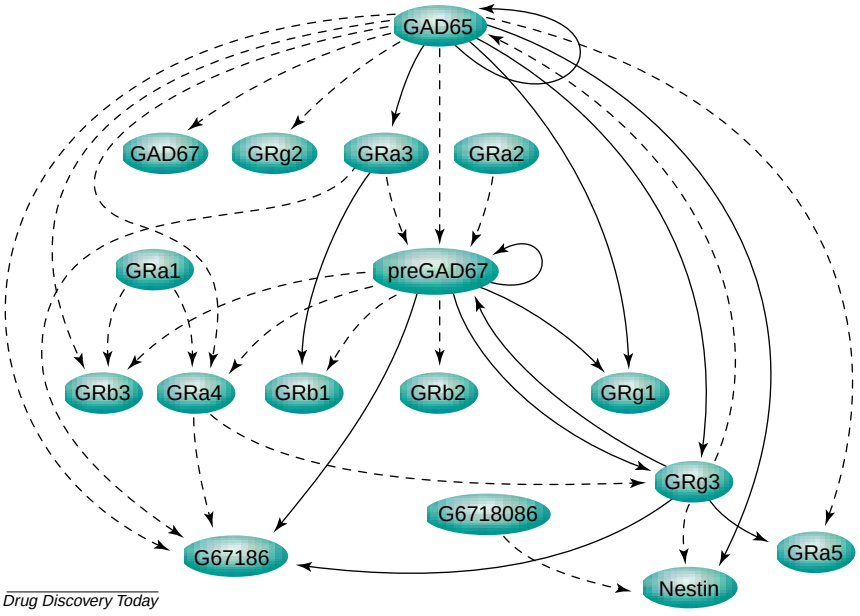


Figure 6. Reverse-engineering and linear modeling of a CNS genetic network. This figure shows a hypothetical gene interaction diagram for the γ-aminobutyric acid (GABA) signaling family inferred from developmental gene expression data (spinal cord and hippocampus data). Although individual proposed interactions have not yet been experimentally verified, the predicted high connectivity within this gene family appears biologically plausible. The positive feedback interaction of the glutamic acid decarboxylase (GAD) species has been proposed independently in another study[46]. Solid lines correspond to positive interactions, broken lines suggest inhibitory relationships[26].

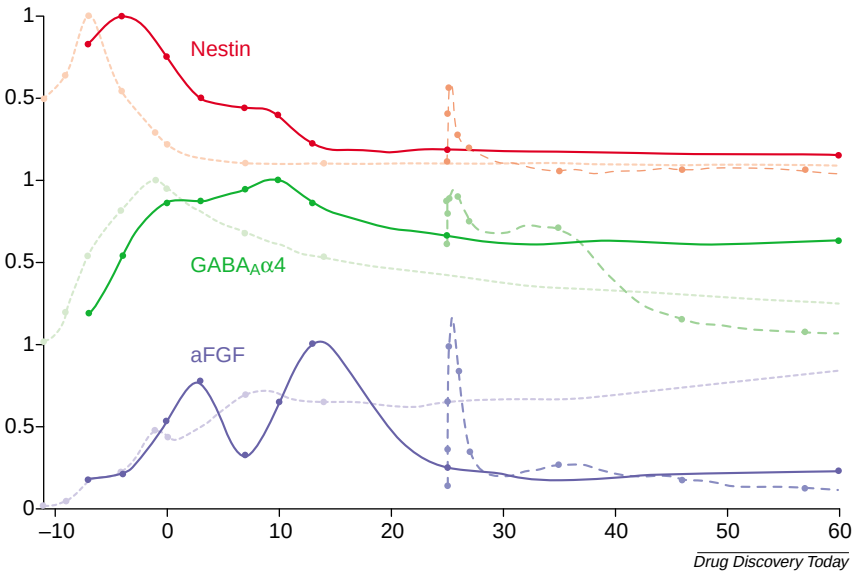


Figure 7. Experimental gene expression data. The points represent development and injury, and the lines represent simulation using a linear model. The model faithfully reproduces the time series of the training datasets. The dotted lines represent the spinal cord, starting 11 days before birth. The solid lines represent the development of the hippocampus, starting seven days before birth. The dashed lines represent the hippocampus kainate injury, starting at postnatal day 25. Abbreviations: $GABA_A\alpha4$, γ-aminobutyric acid receptor subtype A α4; aFGF, alpha fibroblast growth factor.

relationships among the components that matter the most. Ultimately, as precise training and validation datasets for analysis and inference become routinely available, computational models of molecular networks, at various levels of organization, will become the platform on which much biological and medical exploration will be carried out.

These *in silico* platforms will allow cost-efficient experimentation and hypothesis exploration, computationally uncovering the behavior of molecular species and combinatorial interactions that would be difficult, impossible, or too expensive to carry out in wet-lab settings. Within these model frameworks, we have the opportunity to investigate the global dynamics of these networks in terms of their possible dynamic outcomes, known as their attractors, directly corresponding to biological phenotypes.

An example from idealized Boolean 'gene' network models illustrates the principle of such an exploration. It is relatively straightforward to obtain an exhaustive description of the complete dynamics of a Boolean network [A. Wuensche, Discrete Dynamics Lab (DDLab); http://www.ddlab.com][44], as shown in Fig. 8 for a 12-element network (elements with sub-indices collapsed into one element in the analysis, resulting in 12 effective elements). The wiring diagram (Fig. 8a) and rules (not shown; see Ref. 6) completely determine the network trajectory and attractor, that is, network state transitions from one discrete time-point to the next into a repeating state-cycle (Fig. 8b). This trajectory is mapped out in the central graph, the nodes representing states, with the numbering corresponding to the time points in the detailed trajectory. This graph illustrates the important point that the example trajectory is but one of many alternative trajectories leading to the attractor, together defining a

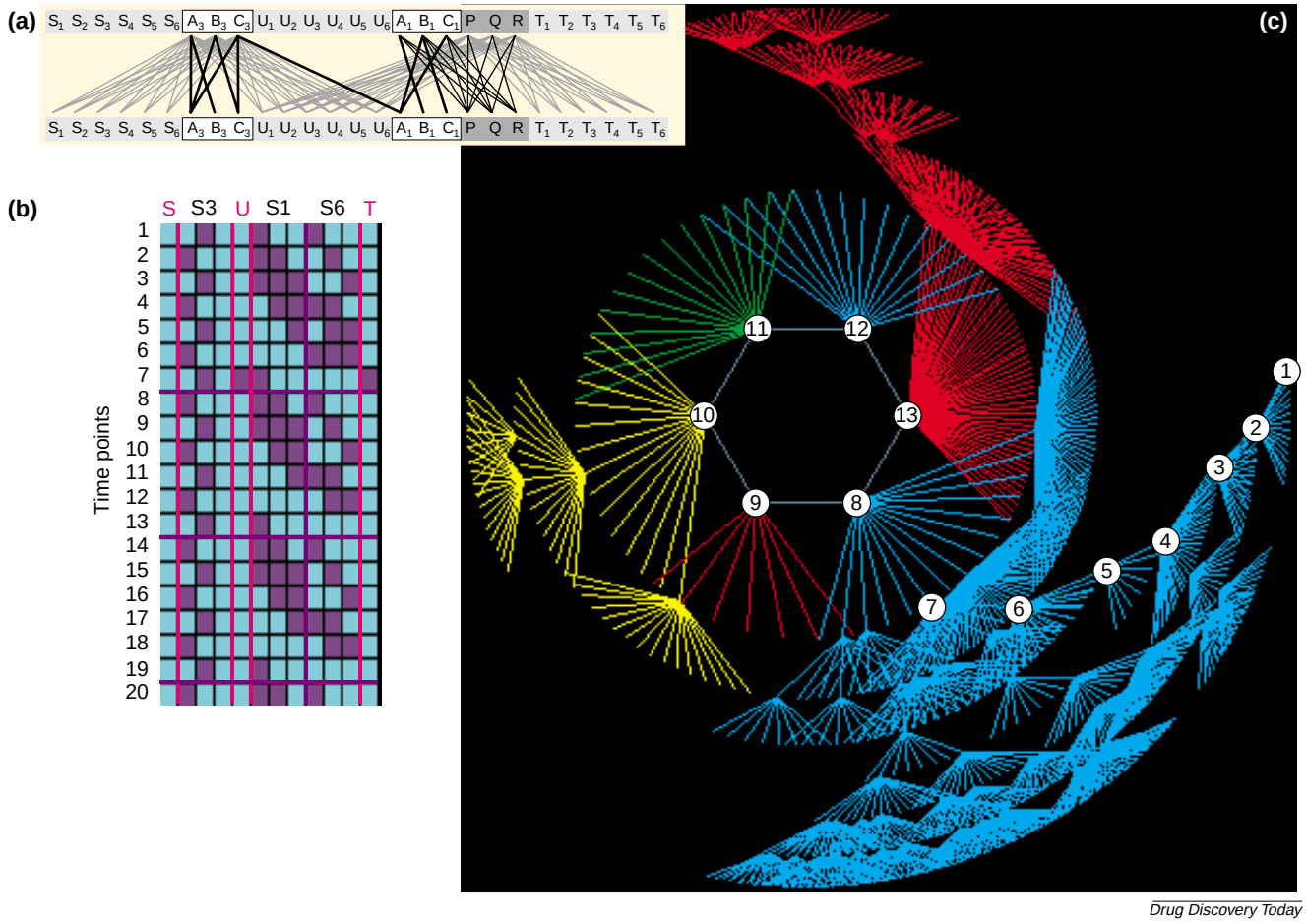
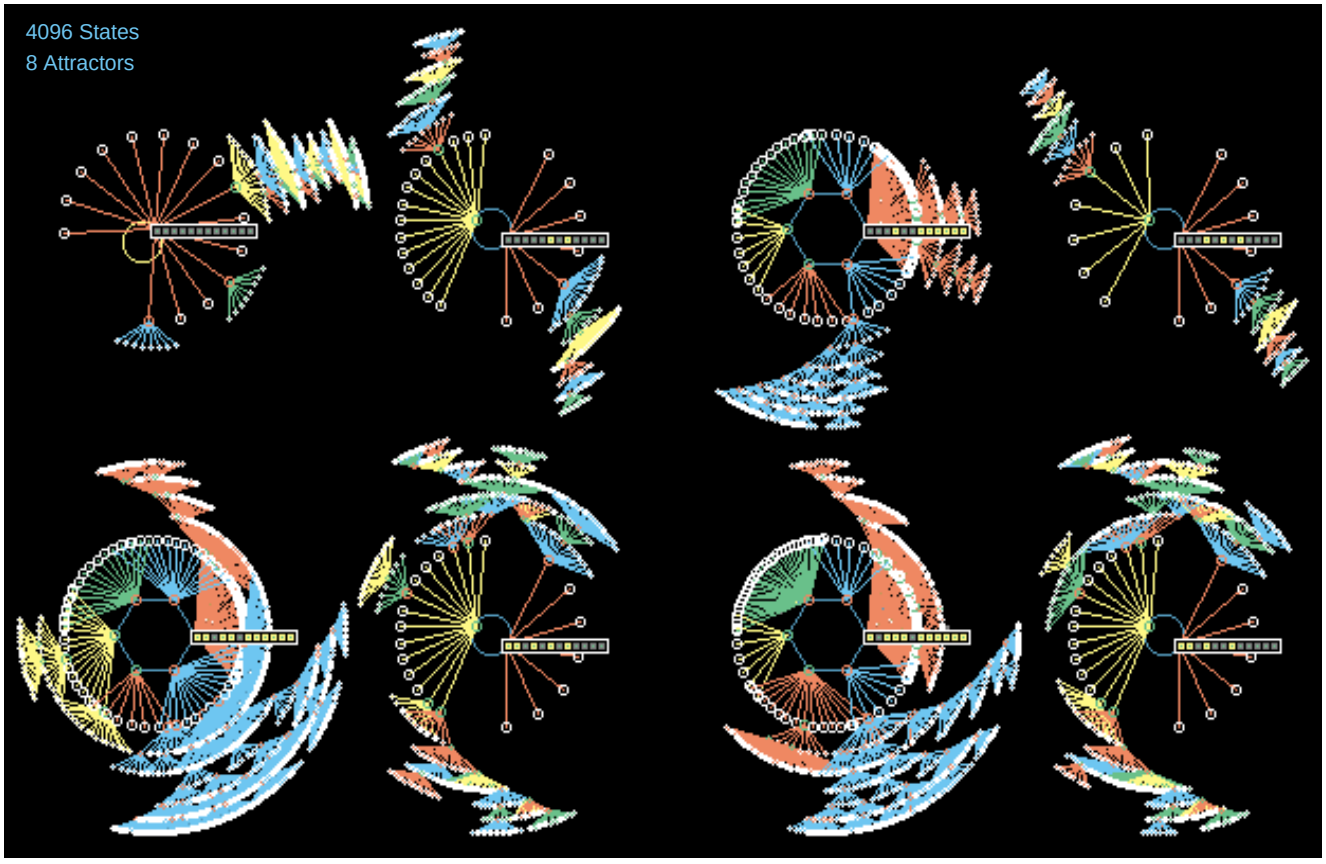


Figure 8. The dynamics of Boolean networks (illustrations prepared using the DDLab software[44] showing **(a)** wiring, **(b)** trajectory and **(c)** basins of attraction[6]. **(a)** Wiring diagram of binary gene-interactions: the network model comprises hypothetical regulatory gene modules (S1, S3, S6) and dependent modules of structural genes (S, U and T – share identical wiring and rules). **(b)** Trajectory: as determined by the wiring and rules, the network passes through a series of states from a given starting state, finally arriving at a repeating attractor pattern, a six-state cycle in this case. Dark gray = on; light gray = off; time points numbered at left. Modules S, U and T have been collapsed into a single element for simplicity. **(c)** Basins of attraction: the states of the trajectory shown in **(b)** are displayed as a series of connected points (labeled by time points) arriving in a cyclic graph resembling the attractor. The additional nodes in the graph resemble other states which also lead to the same attractor, hence the term 'basin of attraction'. All states in this graph merge into a single attractor.

complete 'basin of attraction'. This basin is one of a total of eight basins of attraction (Fig. 9), containing all 4096 (i.e. 2^{12}) states of the network.

A meaningful biological analogy lies in basins of attraction of genetic networks; these essentially correspond to stable phenotypes such as cells or whole organisms, which can be achieved at the end of a developmental process or perturbation response[5,6]. These attractors are robust to minor perturbations, that is, if one of the six states within the attractor cycle is transmuted into one of the many other states in the basin of attraction, the network will fall right back into the original cycles after a few time-steps. Following the example above, Fig. 9 would then characterize all possible phenotypes for a genetically determined network. Pathological phenotypes can also be viewed as attractors that are reached after alterations in external inputs (state changes) or genetic mutations (network wiring and rules). This view is particularly relevant to complex diseases such as cancer, the causes of which are known to involve complex, multigenic alterations in cell growth-control signaling. In summary, the secret to systematic bioengineering and therapeutics is to understand which network connections are the key controllers that form the boundaries between the individual attractors. Wet-lab and clinical experimentation would then be focused on the validation of a small number of carefully chosen model predictions of crucial variables that determine the attractors or phenotypes encapsulated by the network.



Drug Discovery Today

Figure 9. Complete basin of attraction fields covering all 4096 states of example network.

Successful network models based on incomplete data

Phenomenological models are always incomplete. They are deliberately oversimplified condensations of the current state of knowledge concerning the system of interest. However, therein lies the models' potential utility – abstract representations of the decisive relationships linking the most important entities, encoded in a manner that can be mathematically and computationally manipulated, studied and understood.

Why should such models be successful at all with respect to understanding and prediction of behavior? Perhaps part of the answer resides in the physical reality of the networks themselves. What we usually think of as a network is a coherent system that comprises units interacting in some sort of orchestrated and regulated fashion such that the emergent behavior of the whole (i.e. the network) is recognizable and characterizable. Once some of the real physical behavior is recognized, the system can be described or modeled at a level of detail appropriate to the system behavior, while ignoring the details of the system's constituent components.

To make an analogy: the physical reality of a liquid-to-solid phase transition can be modeled and understood rather well without resorting to all the details concerning the behaviors of the individual constituent molecules. This works because the phase transition is a collective, albeit co-operative, behavior manifested by the many individual constituents across large length-scales (relative to the size of the constituents) and across long time-scales (relative to the constituents' vibrations). The real physical behavior, as well as the model of it, is robust and relatively insensitive to the behavior of any particular constituent. (This is not to say, of course, that the properties of the constituents are unimportant to the system; for example, the van der Waals forces between individual molecules are important for determining the temperature at which a liquid solidifies.)

Furthermore, especially for metabolic and gene networks, it seems plausible that evolution has selected networks whose behavior is buffered, or compensatory (or 'canalized', to use Waddington's term[45], with respect to changes in the network's boundary operating conditions because, after all, biological systems usually survive and function in fluctuating

environments. Assuming that real network behavior is coherent and inherently relatively insensitive to noisy fluctuations at its boundaries, and even buffered against fluctuations in the activities of its constituent components, then it follows that good models can adequately mimic network organized-behavior without necessarily incorporating overly detailed descriptions of all of its constituents.

Acknowledgements

The authors wish to thank colleagues at Molecular Mining Corporation: Alan Ableson and Max Kotlyar for providing materials for the manuscript, Joanne Luciano and Evan Steeg for valuable discussions and editorial suggestions. We thank Steve Horrigan of Avalon Pharmaceuticals for offering us the use of materials from a collaborative project.

References

1 Bishop, M.J., ed. (1999) *Genetics Databases*, Academic Press
2 Gray, I.C. *et al*. (2000) Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Genet.* 9, 2403–2408
3 Peet, N.P. and Bey, P. (2001) Pharmacogenomics: challenges and opportunities. *Drug Discov. Today* 6, 495–498
4 Roses, A. (2000) Pharmacogenetics and future drug development and delivery. *Lancet* 355, 1358–1361
5 Kauffman, S.A. (1993) *The Origins of Order*, Oxford University Press
6 Somogyi, R. and Sniegoski, C. (1996) Modeling complexity of genetic networks. *Complexity* 1, 45–63
7 Waddington, C.H., ed. (1968) *Towards a Theoretical Biology: an IUBS symposium*, Aldine Publishing
8 Scherf, U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244
9 Winslow, R.L. *et al.* (1998) Modeling the cellular basis of altered excitation–contraction coupling in heart failure. *Prog. Biophys. Mol. Biol.* 69, 497–514
10 Stokes, C.L. (2000) Biological systems modeling: powerful discipline. *Am. Inst. Chem. Eng. J.* 46, 430–433
11 Savageau, M.A. (1998) Rules for the evolution of gene circuitry. In *Pacific Symposium on Biocomputing* (Vol. 3), pp. 54–65
12 Thieffry, D. and Thomas, R. (1998) Qualitative analysis of gene networks. In *Proceedings of the Pacific Symposium on Biocomputing* (Vol. 3), pp. 18–29
13 Akutsu, T. *et al.* (2000) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* 16, 727–734
14 McAdams, H.H. and Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 94, 814–819
15 McAdams, H.H. and Arkin, A.P. (1998) Simulation of prokaryotic genetic networks. *Annu. Rev. Biophys. Biomol. Struct.* 27, 199–244
16 Feurstein, G.Z. *et al.* (1999) Antithrombic efficacy of a novel murine antihuman factor IX antibody in rats. *Arterioscler. Thromb. Vasc. Biol.* 19, 2554–2562
17 Jones, K.C. and Mann, K.G. (1994) A model for the tissue factor pathway to thrombin. II: A mathematical simulation. *J. Biol. Chem.* 269, 23367–23373
18 Cho, C. *et al.* (1999) Kinetically limited anticoagulation due to inhibition of factors VIII and IX: A mathematical modeling study. In *International Society Thrombosis and Haemostasis Book of Abstracts*, (Thrombosis and Haemostasis XVIIth Congress), Abstract 550
19 Wen, X. *et al.* (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. U. S. A.* 95, 334–339
20 D'haeseleer, P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726
21 Onami, S. *et al.* The DBRF method: gene network inference method and its application to yeast gene expression data. In *Foundations of Systems Biology* (Kitano, H., ed.) (in press)
22 Tobin, F.L. *et al.* (1999) Towards reconstruction of gene regulatory networks. In *Proceedings of International Conference on the Modeling and Simulation of Microsystems, (MSM 99)* (Vol. 2), Abstract T23.06
23 Liang, S. *et al.* (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Proceedings of the Pacific Symposium on Biocomputing* (Vol. 3), pp. 18–29
24 Arkin, A. *et al.* (1997) Deduction of a complex reaction mechanism from measured time series: verification of the theory of statistical construction. *Science* 277, 1275–1279
25 Kyoda, K.M. *et al.* (2000) A gene network inference method from continuous-value gene expression data of wild-type and mutants. *Genome Informatics* 11, 196–204
26 D'haeseleer, P. *et al.* (1999) Linear modeling of mRNA expression levels during CNS development and injury. In *Proceedings of the Pacific Symposium on Biocomputing* (Vol. 4), pp. 41–52
27 Maddox, J. (1992) Is molecular biology yet a science? *Nature* 355, 201
28 Lauffenburger, D.A. and Linderman, J.J. (1993) *Receptors: Models for Binding, Trafficking, and Signaling*, Oxford University Press
29 Rosen, R. (1970) *Dynamical System Theory in Biology*, Wiley
30 Davies, D.D. (1973) Rate Control of Biological Processes. *Symp. Soc. Exp. Biol.* 27, Cambridge University Press
31 Rubinow, S.I. (1973) *Mathematical Problems in the Biological Sciences*, SIAM
32 Murray, J.D. (1989) *Mathematical Biology*, Springer-Verlag
33 Segel, L.A. (1991) *Biological Kinetics*, Cambridge University Press
34 Keener, J. and Sneyd, J. (1998) *Mathematical Physiology*, Springer-Verlag
35 Kenakin, T. (1997) *Pharmacologic Analysis of Drug–Receptor Interactions*, Lippincott-Raven
36 Belair, J. *et al.*, eds (1995) *Dynamical Disease – Mathematical Analysis of Human Illness*, Amer. Inst. Physics
37 Asenchenkov, A. *et al.* (1994) *Disease Dynamics*, Birkhauser
38 Hoppensteadt, F.C. and Peskin, C. (1992) *Mathematics in Medicine and the Life Sciences*, Springer-Verlag
39 Guyton, A.C. and Hall, J.E. (1996) *Textbook of Medical Physiology*, W. B. Saunders
40 Fung, Y.C. (1997) *Biomechanics – Circulation*, Springer-Verlag
41 Martin, R.B. *et al.* (1998) *Skeletal Tissue Mechanics*, Springer–Verlag
42 Jacquez, J.A. (1996) *Compartmental Analysis in Biology and Medicine*, BioMedware
43 Harrison, L. (1993) *Kinetic Theory of Living Pattern*, Cambridge University Press
44 Wuensche, A. (1998) Genomic regulation modeled as a network with basins of attraction. In *Proceedings of the Pacific Symposium on Biocomputing* (Vol. 3), pp. 89–102
45 Waddington, C.H. (1942) Canalization of development and inheritance of acquired characters. *Nature* 150, 563
46 Somogyi, R. *et al.* (1995) Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. *J. Neurosci.* 15, 2575–2591