RC23104 (W0402-027) February 5, 2004
Chemistry

IBM Research Report

Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand

J. Jeremy Rice, Gustavo Stolovitzky
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Abstract

Pathway reconstruction is a fundamental task in systems biology toward an ultimate goal of full scale *in silico* simulations. The data for such reconstructions is mostly lacking, but collection is underway for some model organisms. However, biological specificity may limit the ability to extrapolate findings. High throughput data and methods may alleviate these problems, but only coarse or limited reconstructions are now possible. Inclusion of multiple data sources may improve the situation but remains a challenge.

The *in silico* goal and its data requirements

Reverse engineering cellular pathways is a central theme in systems biology [1,2]. By reverse engineering, we mean the inference of signaling, metabolic or gene regulatory pathways from experimental data. Ideally, enough experimentation would provide sufficient detail to develop *in silico* models as concise representations of biological systems. The models may serve as integration tools where necessary components must be assembled and function together to recapitulate behaviors of the real system. Taken a step farther, models can predict and provide insight into how complex behaviors emerge from simpler interactions. Models may not need to be fully detailed to achieve these goals. To varying degrees, all models are abstractions, and even simple or qualitative models can play very useful roles as important ways to test hypotheses and best direct limited experimental resources to collect the most salient data [1,3]. However, we suggest that predictive models with mechanistic detail best demonstrate that a complex system is well understood, and this is ultimately the goal of systems biology.

Currently, however, we are only beginning to have enough data to construct rather rudimentary simulations of a handful of cellular processes. For example, in a well studied system such as cell cycle, current models [4] are parsimonious descriptions containing mainly key players with speculative features to fill in gaps in the current knowledge. The point here is the paucity of data and not to degrade the value of this work. One must make best use of the data at hand, and more complex models can be built on simpler beginnings. The metabolic processes in *E. coli* is another well-characterized system for which large-scale models have been developed [5-7]. These models are certainly integrative in that large amounts of data must be culled from many sources. In some case, valuable insights are gained from topological [8] or semi-quantitative methods such as Petri Nets [9]. However, more quantitative models require that rate or equilibrium constants be measured or extracted from the literature. Amalgamating data from many sources carries costs, especially if in vitro data must be extrapolated to in vivo conditions. The work of Palsson and coworkers [6,7] and others [10] shows that additional constraints such as optimal growth rates provide the ability to refine parameters sufficiently to produce predictive metabolic models . Furthermore constraint-based approaches point out that simply modeling all metabolic pathways is insufficient because all pathways are not typically active at all times; hence, no real world instantiation of the system is well represented. Indeed, there is increasing work to understand the role that gene regulation plays in controlling metabolic processes [7,10].

The metabolic models can take advantage of a long history of research and well populated pathway-based databases. For example, WIT [11], KEGG [12] and Ecocyte [13] are some of the pioneering databases to store metabolic pathways including genomic, enzymatic, proteomic, and functional information (for relevant websites, see Table 1). To improve utility, many existing databases are being expanded to include new biomolecular information (i.e., carbohydrate structures in KEGG [12]) and larger numbers of organisms (i.e., 158 organisms are represented in MetaCyc [14]). However, many other areas of biology are yet to be systematically characterized and stored. Developing biological databases itself is a demanding task given the need to efficiently handle heterogeneous and often inconsistent or incomplete data [15]. Karp has pointed out that most biological data exist in flat files or widely-used relational databases, whereas object-oriented databases may be a much more effective method of storing and retrieving biomolecular data [16] . On the content side, large-scale projects are underway to systematically collect more consistent datasets. The E-cell project is gathering vast amounts of data on *E. coli* to improve the ability to simulate this organism. The Alliance for Cellular Signaling has been collecting data for two model cells (cardiac myocytes and B-cells), and more recently human macrophages, in order to have a more complete picture of the cellular signal systems. Simulation of the signaling pathways is goal for the later stages of this project.

Biological specifity: The devil is in the details

Although some large and consistent data sets may be generated for specific cases, a quantitative understanding of a multitude of cell types and species will be required if systems biology is to fulfill its promises. Even if signaling pathways from model systems generalize to some extent, the devil is often in the details. For example, the ability to knock-out genes in mouse has made this species a preferred animal model. However, differences in electrophysiology (i.e. basal heart rate is about 600 beat/min vs. 60 for human) coupled with other differences in calcium handling and myofilament properties [17] have caused some researchers to suggest that mouse may be a poor model for human hearts [18]. Likewise considerable differences exist across different cardiac tissues (i.e., atrium versus ventricle) even within a given species [19]. These differences have necessitated separate cardiac cell models to represent species- and tissue-specific physiologies (i.e. rat ventricle: [20]; canine ventricle: [21]; and canine atrium: [22]). These examples suggest that care must be taken in extrapolating data across species and tissue when one desires to generate detailed and quantitative models. To some extent, the problem may be alleviated by automated, high throughput methods that can be used to collect species- and tissue-specific data sets. However current high throughput technologies present difficulties as well as opportunities in terms of reconstructing cellular pathways. This point will be addressed next.

High throughput data to the rescue?

The advent of gene expression array technology has provided the ability to capture a “snapshot” of the transcriptome, that is, to what level each gene in the genome is being expressed. However, one is often faced with an odd contradiction of insufficient data for reconstruction despite the flood of data from this and other high throughput sources.

Some have suggested that “high throughput methods sacrifice specificity for scale” [23]. For example, studies of yeast cell cycle have yielded nearly a thousand transcript profiles oscillating with some synchrony to the cell cycle [24,25]. Using this data alone, the task of deciphering the key regulatory genes from the genes being regulated becomes virtually impossible considering the large number of oscillating transcript levels coupled the noise inherent to this technology. The common technique of clustering expression profile only suggests co-regulated genes and not the causal relationships. Moreover, many cellular processes that occur post-translationally will not be detected by gene arrays; hence, one is trying to reconstruct interactions from a limited view of the whole panorama of cellular processes.

The current data can be characterized by several key limitations: small signal to noise ratios, insufficient time resolution, insufficient spatial resolution, and too few signals being measured. We consider each of these briefly. Several researchers have pointed to limitations of gene array data attributed to both inherent measurement noise and the variability of sample preparation [26,27]. Presumably repeated measurements may compensate for these limitations but the expense of gene arrays generally prevents scaling up to large numbers of measurements. Another limitation related to expense is that time-course data is often collected at coarse time intervals. This coarseness leads to obvious problems when we wish to characterize dynamic processes such as cascades in gene regulatory networks. Insufficient spatial resolution refers to the necessity of sampling multiple cells with gene array technologies. One cannot assume *a priori* that each cell in a population will be in a similar internal state because of tissue inhomogeneity or asynchrony as in the case of cell cycle [24,25]. Yet another difficulty resides with the object being measured. The most mature technology of gene arrays only measures mRNA levels and not translated protein levels. When comparisons have been made between the two signals, the correlation has been small [28]. This is a serious problem because many important cellular processes occur post-translationally, and high throughput technologies to measure these signals are still in early stages of development.

Despite the difficulties described above, strategies are being developed to press on with reverse engineering cellular pathways from high-throughput data sources. While this is a fairly nascent area of research, a wide variety of approaches have been proposed. In this short perspectives article, we are limited to mention only a small number of these efforts (for a comprehensive review, see [29]). We humorously say that systems biology is currently in state of “aiming low” and “mixing grapes.” The first description comes from an old joke that states that the best way to not burden oneself with failure is to aim low. In this vein, the systems biologist may seek to qualitatively reconstruct pathways instead of aiming at fully detailed kinetic models. Mixing grapes refers to attempts by winemakers to make a better wine by mixing varieties of grapes. The idea is to compensate for the weaknesses of a single grape type, but the question remains if two “wrong” grapes make a “right” wine. Likewise, a big push in systems biology is to combine disparate data sets to get a more complete picture of cellular function. While this approach makes intuitive sense, one may still question the value of combining disparate data sets, especially if the cost is great or the data sets may be contradictory.

We next consider several approaches that contain variations on the themes of "aiming low" and "mixing grapes".

Network inference using high throughput data

As described earlier, current data limit our ability to produce fully detailed kinetic models of cellular processes. Data issues aside, the task of detailed kinetic reconstructions based on time-series data alone is extremely difficult indeed. Even from a theoretical perspective, the underdetermined nature of the problem (more unknowns than equations) implies that a unique solution is not generally possible because an infinite number of reconstructed systems are consistent with any given set of time-series data. To deal with this non-uniqueness, the solution space is often limited by *a priori*, and often reasonable, assumptions such as linearity [30-32], sparseness [31,32], or predetermined model structures such as reactions limited in the number of possible reactants and substrates [33]. While these methods may hold promise in the future, unfortunately the limitations of existing data render these approaches as mostly theoretical exercises except for very small systems with high quality data [32]. If one accepts that detailed large-scale kinetic reconstructions are not generally feasible yet, then the next logical step is to aim lower by considering reconstructions that are coarser or less detailed (see Fig. 1). The question is then how low to aim. A logical place to start is a diagram or map of the chemical connections in a system without all the information to fully understand the dynamics of the system [1]. To organize our discussion, we next consider two hierarchical levels at which cellular pathways can be described.

Inferring Network Topology - In this level we are only concerned with identifying the interaction between nodes (genes, proteins, metabolites, etc) in the system. The goal is the generation of a diagram of non-directional connections between all interacting nodes (See Fig. 1A). For example, many have sought to develop large-scale maps of protein-protein interactions derived from various sources. Two-hybrid studies have produced genome-wide interaction maps for *E. coli* bacteriophage T7 [34], yeast [35], drosophila [36] and *C. elegans* [37]. Although this approach can be comprehensive in regard to being genome wide, many interactions are not reproducible (a potential source of false negatives) and putative interactions occur between unlikely protein combinations (a potential source of false positives). Noting such problems, the most recent studies of this type proposes computational methods to better assess the confidence of the putative interactions [36,37].

Another approach to constructing large-scale connection maps is by mining databases. Specific databases of protein interactions are being developed, the largest of which are DIP [38] and BIND [39]. These databases combine data from many high throughput experiments along with data from other sources, such as published literature. Other methods have sought to mine MEDLINE/PubMed abstracts that are considered to contain concise records of peer-reviewed published results. The simplest methods, often called "guilt by association", seek to find co-occurrence of genes or protein names in abstracts [40] or even smaller structures such as sentences or phrases [41]. This approach assumes that co-occurrences are indicative of functional links, although an obvious limitation is that negative relations (e.g., A does not regulate B) are counted as positive associations. To overcome this problem, other natural language processing method involves syntactic parsing of the language in the abstracts to determine the nature of the interactions (i.e. [42,43]). There are obvious computation costs in these approaches, and the considerable complexity in human language will likely render any machine-based method imperfect. Even with limitations, such methods will likely be required to make knowledge in the extant literature accessible to machine-based analyses [13]. For example, PreBIND used support vector machines to help select abstracts likely to contain useful biomolecular interactions to "backfill" the BIND database [44].

Along other lines, investigators have attempted to identify topological links by analyzing the dynamic behavior of networks. Pioneering work in this area shows that metabolic network topologies can be derived from correlation of time-series measurements of species concentrations [45]. The method is further refined to better identify connections in non-linear systems using mutual information instead of correlation [46]. In another method, pair-wise correlation of gene expression data is used to predict functional connections that could then be combined into "relevance networks" of linked genes [47]. Other methods may seek to use some combination of data sources, although this may not be completely straightforward. For example, discrepancies have been reported between yeast two-hybrid interaction data and gene expression profiles; some long-lasting complexes such as the 20S proteasome correlate well, while transiently interacting proteins tend to correlate poorly [48]. Part of the discrepancy may lay in the noise that plagues some of the employed methods such as the yeast-two hybrid technology. However, noisy datasets can often be combined with other complementary data to produce more reliable results [36,48,49]. Hence, although one may want to use multiple data sources to gain a richer picture of cellular function, in some cases, the results is a smaller but more accurate characterization.

Inferring Qualitative Connections - In this level we include not only associations between cellular entities but also the causal relations of such associations, such as which entities serve as input to others. The goal of this level is to create a diagram of *directional* connections (arrows) from input to output nodes (See Fig. 1B). The issue of causality becomes critically important for reconstructing biological networks as many levels of causal connections may exist. In the category, most methods seek to identify a qualitative indicator of how the input affects the output (i.e. a positive or negative arrow). Researcher have proposed methods that infer connectivities from the estimations of the Jacobian matrix for metabolic [50], signaling and genetic networks [51]. Ross and coworkers have proposed method based on propagated perturbations of chemical species can reconstruct causal sequences of reactions from synthetic [52] and experimental data [53]. To reconstruct gene regulatory systems, methods include fuzzy logic analysis of facilitator/repressor groups in the yeast cell cycle [54] and reconstruction of binary networks (e.g., [55]). However, the wide application of such methods is often limited because the continuous nature of many biological systems prevents easy abstractions into coarser signals. Recently there has been considerable work using Bayesian network inference. Examples include inferring gene regulation using gene expression data from yeast cell cycle [56] or using data from synthetic gene networks [57].

Of the methods discussed, Bayesian networks have found the widest usage and, hence, are described in more detail. Friedman *et al.* [56] first used Bayesian inference methods to analyze gene expression data. Others have published refinements or variations of this general approach (e.g., [58]). Proponents of the Bayesian approach point to several important advantages including the ability to handle noisy and incomplete data sets; easy incorporation of *a priori* knowledge of the network structure with new data; the ability to accommodate hidden variables, and a quantitative output that can be scored against new observations. Some properties of the Bayesian approach are not well suited to biological systems. For example, in principle, Bayesian networks can handle continuous value variables whereas, in practice, data such as mRNA levels must be discretized to allow for computation of joint probabilities between input variables. The optimal discretization method is not obvious and must balance more a faithful representation of the input data (many fine bins) versus a better estimation of joint probabilities (fewer large bins). Another problem may arise if feedback loops exist in the biological system because the inferred Bayesian networks must be acyclic and hence cannot represent loops. In theory, this can be solved with dynamic Bayesian networks that can “unroll” loops, but in practice, a bottleneck might arise because of the amount of data needed to pursue this approach. As we will discuss momentarily, current approaches already have trouble constraining time-invariant Bayesian networks, let alone dynamic Bayesian networks.

In the preceding paragraph we suggest that Bayesian networks have some important advantages and disadvantages. We also see differences in what is possible in theory versus what is practically possible with the data available today. Let us consider these points in a bit more detail. Some initial results with Bayesian networks were less than spectacular. For example, the work of Hartemink *et al.* showed that even three node networks were hard to reconstruct in the yeast galactose metabolic pathway [58]; however, much of the trouble may lie with the quality of the experimental data and not the method *per se*. Later work by the same group showed better results using synthetic gene networks where the researchers had better control of the data quality and quantity [57]. Still, there was trouble in the reconstruction of connections, especially in the case of many connections converging on a node. While the discussion above focuses on Bayesian inference methods, many of the issues discussed generalize to other reconstruction methods as well. Many studies have employed synthetic data or have considered restrictive sizes or types of connections in networks in order to improve the quality of the reconstruction.

Athough studies of synthetic networks are important to test and understand methods, the eventual goal is, of course, to reconstruct networks from real data. In an elegant work [32], researchers were able to reconstruct much of a nine-gene sub-network in a DNA repair pathway in *E. coli* by controlled perturbations of a subset of the member genes. This work made the assumption of sparseness of the pathway connections and included a robust experimental design that kept low levels of noise in the real-time PCR measurement of the transcript levels. In other recent studies [23,59] genome-wide yeast expression data and preliminary clustering was used to determine likely functional modules, i.e. the sets of genes working together to perform a particular function. In addition, other data sources (candidate regulatory genes [59] and yeast two-hybrid data [23]) were combined to predict the functional modules. Hence, better results are found for understanding the regulation of sets of genes (versus single genes in initial attempts), more and perhaps better quality data, and the use of complementary data types in addition to expression data alone. These differences bring us to the next section.

Bringing it all together: Modules and integrative approaches

An important theme in systems biology has been to look for functional modules that have been conserved and reused. The idea of breaking biological systems into small functional blocks has obvious appeal; the parts can be divided and conquered so that the most complex of machines become readily understood in terms of block diagrams or sets of subroutines. Clearly some conserved modules exist such as the ribosome and the tricarboxylic acid cycle. One method to search for modules involves looking for higher-order structures or recurring sub-networks (often termed "motifs") in metabolic [60] or gene regulatory networks [61]. Another approach mentioned earlier is clustering expression profiles to produce groups of genes that appear to be co-regulated that should ideally reveal the functional modules. However, this assumption does not appear to generalize to all functional groups under all conditions, as some functional groups show well-correlated expression profiles whereas others do not. In [62], the low correlation of genes observed within some functional groups was attributed to the fact that some of these genes belong to multiple functional classes. In another analyses in *E. coli*, 99 cases where found where one reaction existed in multiple pathways in EcoCyc [13]. These observations above suggest potential pitfalls with anticipating too much functional modularity in terms of biology being neatly partitioned into non-overlapping modules. Moreover, the tissue- or species-specific differences mentioned earlier may prevent simplistic transfer of modules from one biological system to another. It remains to be seen if biology is as modular as the system biologist might like it to be.

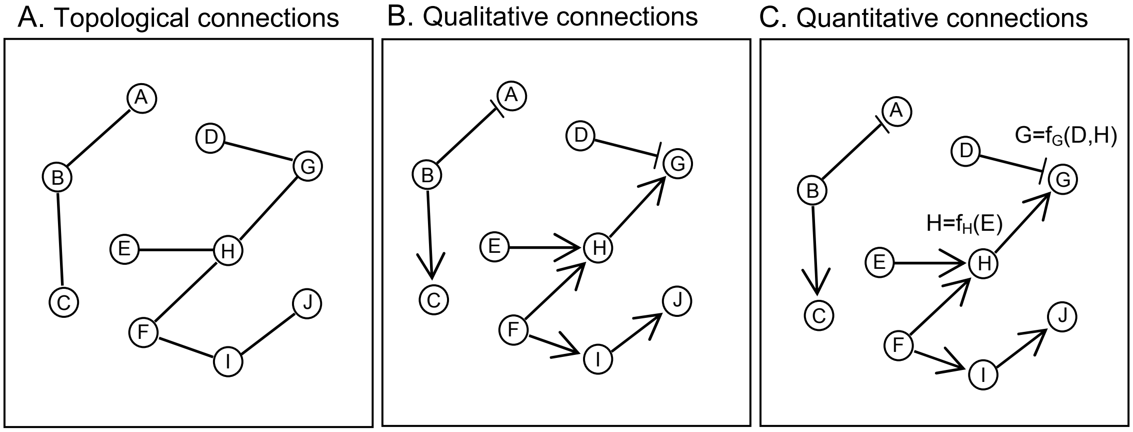
Biological modules may turn out be more interconnected and overlapping than independent in many systems. In addition, the experiences with pathway reconstruction suggest that the combinations of data source produce a more accurate if not more complete characterization of the system under study. These observations point to an eventual need to develop large-scale, predictive models based on a multitude of data sources. For example, metabolic models may combine data from many sources into a quantitative set of equations that can make predictions amenable to experimental verification [6,7,10]. In another system, cardiac models can bridge data at multiple levels (i.e., molecular, cellular, organ, etc.) and their corresponding characteristic timescales [2,63]. In this system, modeling efforts at the single cell level in the heart [64] suggested a mechanism of increased contraction force that was later confirmed in experimental studies of whole heart [65]. The ability to make predictions that are later experimentally verified is often considered a key validation of the utility of the biological models.

With good quantitative data and relative long histories of development, metabolic and cardiac models are clearly special cases. As argued before, we still generally lack the knowledge to build highly detailed models of many biological systems. However, systems biology seems to be on the path to rectify this situation. Current tools such as GenMapp [66] and Cytoscape [13] already allow for gene expression data to be combined and analysed with pathway and other biological data. Compiling pathways and characterizing their dynamic properties is the obvious place to begin the process of the development of simulation-based models. Others are working to improve the infrastructure for integrative modeling in systems biology. Specific simulation platforms are in development like the E-Cell [5], Gepasi [67], and the Virtual Cell Project [68]. Markup languages based on XML [63,69] and associated tools called System Biology Workbench have been developed to ease the exchange of biological data and models. For example, the pathways in the KEGG data have recently been released in XML format [37]. The development of ontologies to organize genomic and proteomic data (such the Gene Ontology Consortium [70]) have proven extremely useful as standardized data resources. These resources can be used to validate new data and automated processing methods (i.e., see [36]). An automated system has been developed to build ontologies of regulatory networks by extracting relationships from the literature using natural language processing [71]. Other ontologies for the exchange of physiological and anatomical information are just beginning to be developed and deployed [63]. The hope is that standardized tools, data and exchange methods will facilitate the development of detailed, quantitative simulations that capture the dynamic nature of biology more effectively than mostly static pathway maps.

Conclusions-

Despite the availability of genome-wide high-throughput data, we are still far from having all the information needed for large-scale, kinetic simulation of cellular dynamics. Hence, as a logical first step, current pathway reconstruction methods are directed at more static descriptions of the connections between cellular components. Early results generally show that reconstruction improves by the integration of multiple data types. Approaches that propose to identify functional modules are enticing, although problems may arise when modules are formed by cellular components that belong to more that one functional class. The necessity to combine disparate data sets combined with the desire to extract maximum information for existing data is driving the development of new methods and analysis tools. Similarly, some large-scale efforts are underway to collect the necessary data to better reconstruct pathways in at least some cell types. It remains to be seen how well these will generalize, because critical species- and tissue- specific differences are often found despite more general modules or motifs being reused in biology. The full task of pathway reconstruction and eventual quantitative modeling will require effective tools, data, and data management techniques that will probably keep systems biology practitioners busy for the foreseeable future. But the expected result, namely the organization of the vast amounts of data into predictive models of cellular function, is definitely worth the challenge.

Figure 1


A. Topological connections

B. Qualitative connections

C. Quantitative connections

Examples -protein-protein interactions from two hybrid methods [34-37]; pair-wise correlation of metabolic species [45] or gene expression [47] profiles; co-occurrences of gene/protein names in Medline abstracts [40,41].

Examples – estimation of Jacobian matrices [50,51]; perturbations analysis of metabolic species [52,53]; Bayesian inference networks [56-58]; fuzzy logic analysis of facilitator/ repressor groups [54] .

Examples – flux-balance and constraint-based metabolic models [6,7,10]; linear models of gene regulation based on perturbations [32] or singular value decomposition and regression [31].

Figure 1 – Examples of hierarchical levels of biological reconstructions are shown. The examples shown are illustrative only and do not represent any particular system. The nodes can represent genes, proteins, metabolites or other biological entities. The connections represent physical interactions (i.e., protein-protein binding or substrate-product linkages) or functional relationships (i.e., expression of gene E activates expression of gene H). **A**. At the level of topological connections, the goal is a diagram of non-directional connections between all interacting nodes. **B**. At the level of qualitative connections, the goal is to create a diagram of directional connections showing causality and an indication of how input nodes affect the output nodes. As is typically shown, arrows indicate positive or activating connections, and lines terminated with perpendicular segments indicate negative or inhibitory connections. **C**. At the level of quantitative connections, one also wants a quantitative function of how the inputs control the output. For an example, $H = f_H(E)$ would describe how the expression level of gene E would affect the expression level of gene H. Similarly, $G = f_G(D,H)$ would explicitly describe the co-dependence of the expression of gene G with respect to expression levels of D and H. Note that such information is not conveyed at the level qualitative connections. For example, one cannot know the relative influence of D and H on G (i.e., the net result on G by raising both D and H) from the information in panel B alone.

Table 1- Relevant websites

E-cell project	www.e-cell.org
Berhnard Palsson - In Silico Organisms	gcrg.ucsd.edu/organisms
Alliance for Cellular Signaling	www.signaling-gateway.org
KEGG	www.genome.ad.jp/kegg
EcoCyc	ecocyc.org
MetaCyc	metacyc.org
MEDLINE/PubMed	www.ncbi.nlm.nih.gov/entrez
GenMapp	www.GenMapp.org
Cytoscape	www.cytoscape.org
Virtual Cell Project	www.nrcam.uchc.edu
Gepasi	www.gepasi.org
Go Consortium	www.geneontology.org
SBML	www.sbml.org
Systems Biology Workbench Project	www.sbw-sbml.org/the_project.html
CellML	www.cellml.org

References-

1 Kitano, H. (2002) Computational systems biology. *Nature* 420 (6912), 206-210.
2 Noble, D. (2002) Modeling the heart--from genes to cells to the whole organ. *Science* 295 (5560), 1678-1682.
3 Ideker, T.E. *et al.* (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac Symp Biocomput*, 305-316.
4 Novak, B. and Tyson, J.J. (2003) Modelling the controls of the eukaryotic cell cycle. *Biochem Soc Trans* 31 (Pt 6), 1526-1529.
5 Tomita, M. *et al.* (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15 (1), 72-84.
6 Famili, I. *et al.* (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A* 100 (23), 13134-13139.
7 Reed, J.L. *et al.* (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol* 4 (9), R54.1465-6914
8 Dandekar, T. *et al.* (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems* 70 (3), 255-270.
9 Zevedei-Oancea, I. and Schuster, S. (2003) Topological analysis of metabolic networks based on Petri net theory. *In Silico Biol* 3 (3), 323-345.
10 Klipp, E. *et al.* (2002) Prediction of temporal gene expression. Metabolic opimization by re-distribution of enzyme activities. *Eur J Biochem* 269 (22), 5406-5413.
11 Overbeek, R. *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* 28 (1), 123-125.
12 Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32 (1), D277-280.

13 Karp, P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science* 293 (5537), 2040-2044.

14 Krieger, C.J. *et al.* (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 32 (1), D438-442.

15 Nelson, M.R. *et al.* (2003) Designing databases to store biological information. *Biosilico* 1 (4), 134-142

16 Karp, P.D. (2003) What database management system(s) should be employed in bioinformatics applications? *OMICS* 7 (1), 35-36.

17 Stull, L.B. *et al.* (2002) Physiological determinants of contractile force generation and calcium handling in mouse myocardium. *J Mol Cell Cardiol* 34 (10), 1367-1376.

18 Kass, D.A. *et al.* (1998) Murine cardiac function: a cautionary tail. *Circ Res* 82 (4), 519-522.

19 Schram, G. *et al.* (2002) Differential distribution of cardiac ion channel expression as a basis for regional specialization in electrical function. *Circ Res* 90 (9), 939-950.

20 Pandit, S.V. *et al.* (2001) A mathematical model of action potential heterogeneity in adult rat left ventricular myocytes. *Biophys J* 81 (6), 3029-3051.

21 Winslow, R.L. *et al.* (1999) Mechanisms of altered excitation-contraction coupling in canine tachycardia-induced heart failure, II: model studies. *Circ Res* 84 (5), 571-586.

22 Kneller, J. *et al.* (2002) Time-dependent transients in an ionically based mathematical model of the canine atrial action potential. *Am J Physiol Heart Circ Physiol* 282 (4), H1437-1451.

23 Troyanskaya, O.G. *et al.* (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A* 100 (14), 8348-8353.

24 Cho, R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2 (1), 65-73.

25 Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9 (12), 3273-3297.

26 Lee, M.L. *et al.* (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 97 (18), 9834-9839.

27 Tu, Y. *et al.* (2002) Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci U S A* 99 (22), 14031-14036.

28 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292 (5518), 929-934.

29 van Someren, E.P. *et al.* (2002) Genetic network modeling. *Pharmacogenomics* 3 (4), 507-525.

30 D'Haeseleer, P. *et al.* (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pac Symp Biocomput*, 41-52.

31 Yeung, M.K. *et al.* (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A* 99 (9), 6163-6168.

32 Gardner, T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301 (5629), 102-105.

33 Koza, J.R. *et al.* (2001) Reverse Engineering of Metabolic Pathways from Observed Data Using Genetic Programming. *Pacific Symposium on Biocomputing* 6, 434-445

34 Bartel, P.L. *et al.* (1996) A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet* 12 (1), 72-77.
35 Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403 (6770), 623-627.
36 Giot, L. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science* 302 (5651), 1727-1736.
37 Li, S. *et al.* (2004) A Map of the Interactome Network of the Metazoan C. elegans.
38 Xenarios, I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res* 28 (1), 289-291.
39 Bader, G.D. *et al.* (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 31 (1), 248-250.
40 Stapley, B.J. and Benoit, G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput*, 529-540.
41 Ding, J. *et al.* (2002) Mining MEDLINE: abstracts, sentences, or phrases? , 326-337.
42 Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, 60-67.
43 Rindflesch, T.C. *et al.* (1999) Mining molecular binding terminology from biomedical text. *Proc AMIA Symp*, 127-131.
44 Donaldson, I. *et al.* (2003) PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* 4 (1), 11.
45 Arkin, A. and Ross, J. (1995) Statistical construction of chemical reaction mechanism from measured time-series. *J. Phys. Chem.* 99, 970-979
46 Samoilov, M. *et al.* (2001) On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos* 11 (1), 108-114
47 Butte, A.J. and Kohane, I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput 2000*, 418-429.
48 Jansen, R. *et al.* (2002) Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics* 2 (2), 71-81.
49 Edwards, A.M. *et al.* (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* 18 (10), 529-536.
50 Chevalier, T. *et al.* (1993) Toward a systematic determination of complex reaction mechanisms. *J. Phys. Chem.* 97, 6776-6787
51 Kholodenko, B.N. *et al.* (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc Natl Acad Sci U S A* 99 (20), 12841-12846.
52 Vance, W. *et al.* (2002) Determination of causal connectivities of species in reaction networks. *Proc Natl Acad Sci U S A* 99 (9), 5816-5821.
53 Torralba, A.S. *et al.* (2003) Experimental test of a method for determining causal connectivities of species in reactions. *Proc Natl Acad Sci U S A* 100 (4), 1494-1498.
54 Woolf, P.J. and Wang, Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics* 3 (1), 9-15.
55 Liang, S. *et al.* (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, 18-29.

56 Friedman, N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J Comput Biol* 7 (3-4), 601-620.
57 Smith, V.A. *et al.* (2003) Influence of network topology and data collection on network inference. *Pac Symp Biocomput*, 164-175.
58 Hartemink, A.J. *et al.* (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput*, 422-433.
59 Segal, E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34 (2), 166-176.
60 Ravasz, E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297 (5586), 1551-1555.
61 Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 31 (1), 64-68.
62 Mateos, A. *et al.* (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res* 12 (11), 1703-1715.
63 Hunter, P.J. and Borg, T.K. (2003) Integration from proteins to organs: the Physiome Project. *Nat Rev Mol Cell Biol* 4 (3), 237-243.
64 Bluhm, W.F. *et al.* (1998) Mechanisms of length history-dependent tension in an ionic model of the cardiac myocyte. *Am J Physiol* 274 (3 Pt 2), H1032-1040.
65 Alvarez, B.V. *et al.* (1999) Mechanisms underlying the increase in force and Ca(2+) transient that follow stretch of cardiac muscle: a possible explanation of the Anrep effect. *Circ Res* 85 (8), 716-722.
66 Doniger, S.W. *et al.* (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 4 (1), R7.1465-6914
67 Mendes, P. and Kell, D. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. 869-883.
68 Slepchenko, B.M. *et al.* (2003) Quantitative cell biology with the Virtual Cell. *Trends Cell Biol* 13 (11), 570-576.
69 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19 (4), 524-531.
70 Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res* 11 (8), 1425-1433.
71 Rzhetsky, A. *et al.* (2000) A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics* 16 (12), 1120-1128.