FEBS 29327 FEBS Letters 579 (2005) 1815–1820

Minireview

Pathway information for systems biology

Michael P. Cary, Gary D. Bader, Chris Sander

Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 460, New York, NY 10021, USA

Accepted 1 February 2005

Available online 9 February 2005

Edited by Robert Russell and Giulio Superti-Furga

Abstract Pathway information is vital for successful quantitative modeling of biological systems. The almost 170 online pathway databases vary widely in coverage and representation of biological processes, making their use extremely difficult. Future pathway information systems for querying, visualization and analysis must support standard exchange formats to successfully integrate data on a large scale. Such integrated systems will greatly facilitate the constructive cycle of computational model building and experimental verification that lies at the heart of systems biology.

Keywords: Pathway data integration; Pathway database; Standard exchange format; Ontology; Information system

1. Introduction

To understand biological processes, we must integrate new observations with existing knowledge to create testable models that can be iteratively refined. This will only be successful if the vast amounts of data gathered by large-scale profiling of biological features, such as mRNA transcripts and proteins, can be efficiently integrated with data from the literature and databases for visualization and analysis.

One major source for computable data about biological processes are databases that capture information on the functional interactions of molecular species [1]. These "pathway" databases facilitate a variety of analysis and simulation techniques that can enrich our understanding of cellular systems.

While recent dramatic growth in the number of pathway databases is a great boon to biologists, it also presents several important challenges. Almost 170 "pathway" databases exist, which differ widely in form and content. This multiplicity of information sources can be daunting to researchers who simply wish to find information about genes or pathways of interest. The lack of uniform data models and data access methods makes pathway data integration extremely difficult, both mechanistically and semantically.

To address these issues, it is useful to review the current landscape of pathway data and techniques for data integration, and then to extrapolate the shape of desirable pathway information systems which flexibly and efficiently facilitate the analysis and modeling of biological systems.

2. Surveying the pathway data landscape

One abstraction that biologists have found extremely useful in their efforts to describe and understand the inner workings of cellular biology is the notion of a biomolecular network, often called a pathway. A pathway is a set of interactions, or functional relationships, between the physical and/or genetic [2] components of the cell which operate in concert to carry out a biological process. Despite tremendous variety in the cellular processes described as pathways, several pathway representation patterns are prevalent in current practice. In the Pathway Resource List, a catalog of almost 170 pathway databases (see http://cbio.mskcc.org/prl), we use these patterns to group pathway databases into four major, slightly overlapping categories: metabolic, signaling, protein interaction, and gene regulation. A description of the major features of these categories provides an overview of the current pathway data landscape.

Metabolic pathway databases generally contain detailed data models that represent a pathway as a series of biochemical reactions, focusing mainly on the chemical modifications made to the small molecule substrates of enzymes (Fig. 1A). Many metabolic pathways have been mapped to the molecular level of detail since the 1950s or earlier and metabolic pathway databases are the earliest and perhaps the best-known. Metabolic databases generally do not represent higher order cellular processes, such as gene regulation.

Metabolic databases predominantly contain prokaryotic pathways, about which rich datasets have been collected. A few metabolic pathway databases, for example KEGG [3], the BioCyc database family [4] and others [5], map pathways from well-studied organisms onto other organisms via functional annotations, such as Enzyme Commission numbers [6], and orthology relationships, but these approaches are imperfect and the resulting pathways often contain a number of gaps, i.e., missing steps in a chain of biochemical reactions. Gap-filling algorithms attempt to address this problem [7].

Signaling pathways propagate information from one part or sub-process of the cell to another, often via a series of protein covalent modifications, such as protein phosphorylation. Dysregulation of biological processes by aberrant signaling pathways causes many common diseases, such as cancer and

E-mail address: pathways_febs@cbio.mskcc.org.

0014-5793/$30.00
doi:10.1016/j.febslet.2005.02.005

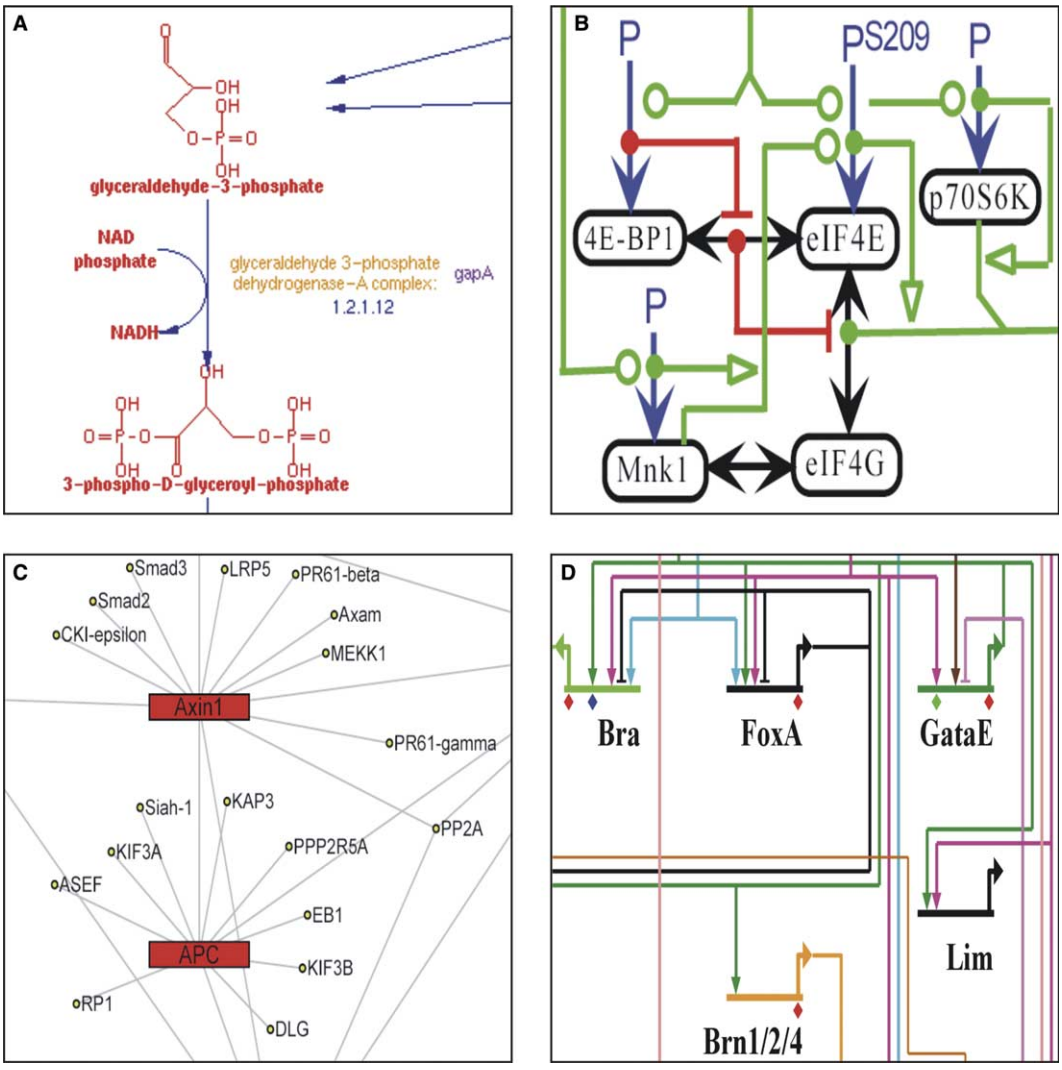

Fig. 1. Common alternative representations of pathway data. (A) Section of the glycolysis 1 pathway diagram from EcoCyc [50], drawn in high detail mode, showing a single biochemical reaction. Blue arrows depict biochemical conversion of substrates to products. The conversion arrows are labeled with the catalyzing enzyme using gold text. (B) Section of a molecular interaction map from the eMIM resource [51] showing regulation of hypoxia-responsive genes. Diagram shows phosphorylation events (blue arrows originating in blue letter P's; phosphorylation sites, if known, are abbreviated in superscript, e.g., S209 = serine 209), inhibitory relationships (red flat-headed arrows), enzymatic stimulation of events (green lines ending in open circles), binding interactions (black double-headed arrows), and non-specific stimulation of events (green arrows). Proteins are shown in black ovals, nodes (filled circles) placed on lines represent the products of processes; e.g., the node on the binding interaction arrow between eIF4E and eIF4G represents the eIF4E:eIF4G complex. (C) Section of the WNT pathway diagram from HPRD [21]. Proteins identified as important components of the pathway are shown as red boxes, other proteins are depicted as small yellow circles. Protein–protein interactions are drawn as edges between proteins. (D) Section of the endomesoderm gene network in the BioTapestry network viewer (see http://www.biotapestry.org). Genes are shown as short, thick horizontal lines. Gene products are represented as short vertical arrows originating at genes and ending in right angles. Activating and inhibitory relationships are shown as normal and flat-headed arrows, respectively, drawn from gene products to regulated genes.

diabetes [8,9]. Though not as well-established as metabolic pathway databases, signaling pathway databases are being actively constructed by a number of groups.

As many signaling pathways are present only in multi-cellular organisms, signaling databases tend to focus on eukaryotes. These organisms are much more complex and less well-studied than some bacteria and their signaling pathways appear to be more diverse than metabolic pathways. Accordingly, signaling pathway databases tend to use higher level abstractions compared to metabolic databases (Fig. 1B). For example, CSNDB [10], TRANSPATH [11] and others [12,13], often forego detailed description of the biochemical reactions involved in signaling and instead use generic concepts of activation and inhibition.

Protein interaction databases contain by far the largest number of interactions of any type of pathway database. Large amounts of protein interactions (protein–protein, protein–DNA, etc.) are generated by various large-scale experimental methods, unlike metabolic and signaling pathway data, which are generated primarily by traditional small-scale experimental techniques [14]. A well-known problem with most high throughput methods of detecting molecular interactions is the high rate of false positive results they generate [15]. Protein interactions detected by these methods should therefore be treated with less confidence until they have been verified by repeated observations or orthogonal experiments [16], and storing experimental evidence for each interaction is important for most protein interaction databases.

Of the various types of pathway databases, protein interaction databases tend to be the least detailed (Fig. 1C), although they often have broad organism coverage [17,18]. GRID [19], for example, stores only the that fact that an interaction was observed between two proteins in at least one experiment. Some databases add additional detail, such as binding sites or, if known, the functional consequences of an interaction on the participants [20,21].

Gene regulation databases currently tend to focus on the relationships between transcription factors and the genes they regulate (Fig. 1D). These databases also have broad organism coverage and share features with both signaling and protein interaction databases, as they collect protein–DNA interactions [22] and regulatory (activation and inhibition) events [23]. Some genetic regulatory databases incorporate protein–DNA binding data from high-throughput assays, such as chromatin immunoprecipitation followed by cDNA microarray analysis (ChIP^2) [24]. This transcription factor-DNA binding information only indicates that the prerequisite of classical gene regulation, transcription factor binding upstream of a regulated gene, can occur – it does not provide information on the functional consequences, if any, of a DNA–protein binding interaction. Other aspects of gene regulation, such as control of alternative splicing, post-transcriptional regulation of protein expression and regulation of the degradation of gene products, are currently rarely covered in gene regulation databases.

Though the attributes 'metabolic', 'signaling', 'protein interaction', and 'gene regulation' serve as useful distinctions for discussing pathway data, these categories arise from experimental capabilities, research trends and common abstractions of biological relationships, and do not correspond closely to physical or chemical features of cellular biology. Furthermore, this classification scheme is not logically disjoint, universally accepted, nor all inclusive. Some databases span multiple categories, such as Reactome [25], which we might classify as both a metabolic and a signaling database. Others do not fit into these categories, such as those that store genetic interactions [26], and databases that store literature co-cited gene name links [27] or more detailed literature extracted links [28]. While these databases may not be universally considered pathway databases, they contain valuable functional links between genes, many of which are not available in other pathway databases. While integration of these diverse data sets is challenging, an inclusive definition of pathway data is necessary to cover existing knowledge and to generate flexible and accurate input for model building in systems biology.

3. Using pathway data to answer biological questions

The principal motivation for building pathway databases and information systems (Fig. 2) is to facilitate qualitative and quantitative modeling of biological systems, outside of the direct capacity of the human brain, using software on powerful computers. A wide range of techniques have been developed that use pathway data of varying detail to answer specific biological questions.

Questions such as 'What are the fundamental design patterns in the system?', 'What are the key relationships between system components?' and 'What are the physiological effects of system perturbation?' can be answered using quantitative and qualitative modeling. Quantitative modeling, such as representing a dynamic chemical process using a system of differential rate equations, requires highly detailed pathway information, such as kinetic constants, initial concentrations and clear connectivity of reactions. Some of this information is available in metabolic pathway databases and the literature [29].

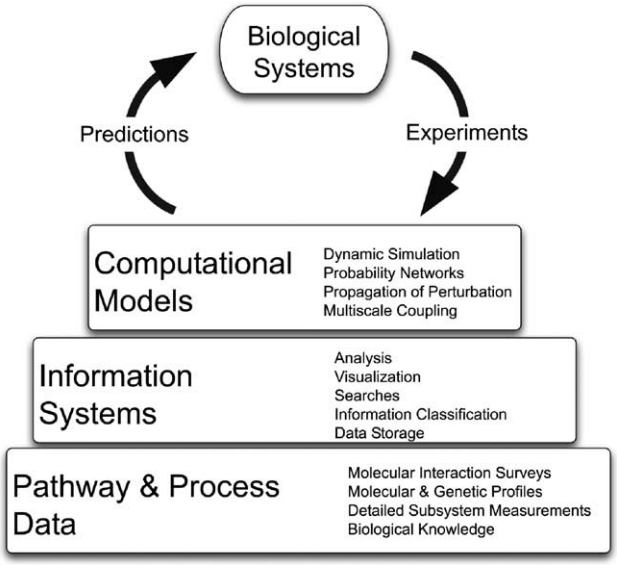

Fig. 2. The iterative biological system modeling method. Biological knowledge from a variety of sources, including molecular interaction surveys and molecular and genetic profiles, pathway databases and literature populate information systems that support data storage, querying, visualization and analysis. These information systems support the construction of computational models of cellular processes, which are used to make testable predictions of cellular behavior. Experimental results must be compared to these predictions and used for model refinement.

Qualitative models are easier to build because they require much less detailed knowledge of the system. Using only topography information [30] and/or qualitative information about reaction rates (e.g., fast or slow) [31], qualitative models can discover system properties not apparent in static pathway data.

Evolution-focused questions, such as 'Which biological processes are homologous?', can be answered using techniques that identify common functional motifs and design principles, e.g., through species comparison. For example, PathBLAST [32] can align protein–protein interaction networks and process algebra techniques [33] can be used to formally define process homology (bi-similarity).

Though many pathway databases only store interactions between genes, proteins and other cellular components, there is clear evidence for higher-order organization in these networks. Can we determine how networks are organized and create abstractions that serve as more effective descriptions of network features? A number of groups have tried to answer this question using only the molecular interaction network topology [34]. Molecular interaction networks have been found to cluster into regions that represent complexes [35] or processes [36]. Statistically over-represented motifs have also been found [37] and some of these have been thoroughly analyzed [38].

Basic questions, such as 'What is the function of my gene?' are still vitally important, since the majority of genes in most genomes have no known function. Examining genes in the network context can help answer this question. For example, a protein of unknown function connected to a set of proteins involved in the same biological process is likely to function in that process as well [39,40].

Less-detailed pathway data, such as proteomics-based protein–protein interactions, can be used to answer questions like 'What network patterns allow prediction of new interactions?' For example, statistically significant domain–domain correlations in a protein interaction network have been used to hypothesize that certain domains mediate binding interactions and to predict new interactions [41,42]. Machine learning techniques can also be used to predict protein–protein [43] or genetic interactions [44].

Finally, questions such as 'What biologically relevant patterns in molecular and genetic profiling data relate to disease?' are vitally important for clinical health research and require a large amount of pathway data to answer effectively. For example, transcriptionally active neighborhoods or regions in an integrated pathway network that correlate with disease state may indicate active pathway components that play a role in disease progression and provide leads for further study [45,46].

4. Pathway data integration for systems biology

The power of many pathway analysis techniques is proportional to the amount of input pathway data. For example, the activity centers algorithm [46] relies on connections between genes in order to detect regions of the interaction network that are up- or downregulated; missing connections could cause an important active region to go unnoticed by the algorithm. Thus, it is vital that as much pathway data as possible is available for the organism being studied.

The diversity among pathway databases makes this challenging. Differences in data models, data access methods, file formats and subtle semantic differences in shared terms create numerous difficulties for those attempting to gather and analyze data from multiple sources. Creation of a new data model is sometimes important for a particular group's research, the continued proliferation of new pathway databases, each with their own format, aggravates the data integration problem.

One way to overcome this challenge is to develop a widely supported pathway data standard. Data standards reduce the total number of translation operations needed to exchange data between multiple sources (from $n^2 - n$ to $2n$, where n is the total number of data suppliers and consumers). They also distribute the reduced translation burden more evenly between data consumers and data providers and facilitate collaboration and accessibility of pathway data to newcomers, thus promoting growth. Because of this, data standards are one of the few scalable data integration strategies (see [47] for a recent review).

Pathway data standards exist, but none cover all aspects of pathway data (Fig. 3). CellML [48] and SBML [49] both are designed to represent quantitative pathway simulation models that can be exchanged between simulation software packages. Since they do not contain data types for many concepts commonly represented in pathway databases, such as 'transport' or 'RNA', these formats are not well suited for data exchange between databases.

The Proteomics Standards Initiative's Molecular Interaction (PSI-MI) format [48], has been developed recently to exchange molecular interaction data between major protein–protein interaction databases. PSI-MI is developed in a practical leveled approach, following the lead of SBML, in which each level adds additional data representation capabilities. PSI-MI Level 1 is designed to represent proteomics protein–protein interaction data, including experimental method description. PSI-MI Level 2 expands this scope to include interactions involving small molecules, DNA and RNA. Though it is relatively new, a number of molecular interaction databases already support data export in the PSI-MI format (e.g., BIND [20], DIP [16], HPRD [21], IntAct [17], and MINT [18]).

To capture more of the pathway data that currently resides in databases, BioPAX (http://www.biopax.org) is being developed by various pathway database groups, also using a leveled approach. Because many less-detailed data types that exist in the pathway data space are difficult to represent in a highly detailed format, the BioPAX ontology allows representation of multiple levels of data resolution using an abstraction hierarchy. This feature is essential for capturing data from the disparate sources of pathway data in a convenient manner. BioPAX Level 1, released in mid-2004, is designed to represent metabolic pathway data and Level 2, near release, adds support for PSI-MI molecular interaction data. Future levels of the format will expand scope to include signaling pathways and genetic interactions.

5. Future directions

The ultimate aim of projects like PSI-MI and BioPAX is to enable effortless collection of pathway data so that it may be efficiently applied to answer biological questions. Ideally, biologists should never need to perform time-consuming data collection tasks in order to perform a particular analysis. Instead, they should be able to locate, retrieve and apply data of interest without worrying about data models, exchange formats, or integration methods.

To achieve this goal, data standards must become broadly adopted by pathway databases. This would enable a variety of large-scale data integration approaches, such as a centralized or distributed pathway data warehouse or a query engine able to retrieve data from multiple standards-compliant primary databases. Importantly, pathway data analysis tools must be built to interface with these integration systems to make pathway data retrieval painless.

Widely accepted data standards and integration infrastructure can also streamline one limiting factor for pathway database growth, namely pathway database curation through manual scientific literature mining. A unified, but distributed curation effort built on accepted curation and data validation tools, involving many biologists, may finally provide a cost-effective data entry solution that scales with exponential data growth. Journals could support this effort by making public pathway data deposition a precondition for publication, as many have done with sequence and structure data.

With public data sharing infrastructure, we can build software platforms that allow high-level and effective pathway

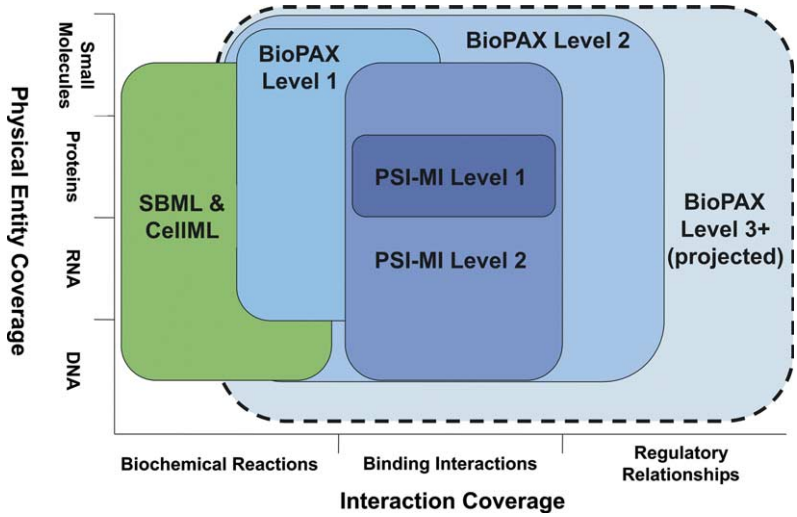

Fig. 3. Data coverage of pathway data formats. Pathway data space is represented two-dimensionally with physical entity classes vertically and interaction types horizontally. Approximate coverage of this space by each pathway data format is represented with colored boxes; database exchange formats are shown in blue, simulation model exchange formats are shown in green. Versions (e.g., level 1, 2, etc.) of a format that have different scope are drawn in separate boxes; formats with similar scope are shown in the same box. The dashed border indicates planned versions not yet available.

data manipulation and analysis using natural biological abstractions. When combined with the trend towards cheap, high-throughput cellular profiling technology, we can imagine a swift convergence on biological process understanding through iterative systems biology modeling methods.

References

[1] Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. Annu. Rev. Genomics Hum. Genet. 2, 343–372.
[2] Tong, A.H., et al. (2004) Global mapping of the yeast genetic interaction network. Science 303 (5659), 808–813.
[3] Kanehisa, M., et al. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res. 32 (Database issue), D277–D280.
[4] Romero, P., et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 6 (1), R2.
[5] Overbeek, R., et al. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. 28 (1), 123–125.
[6] Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999) Eur. J. Biochem. 264 (2), 610–650.
[7] Green, M.L. and Karp, P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics 5 (1), 76.
[8] Hahn, W.C. and Weinberg, R.A. (2002) Modelling the molecular circuitry of cancer. Nat. Rev. Cancer 2 (5), 331–341.
[9] Flyvbjerg, A., et al. (2004) The involvement of growth hormone (GH), insulin-like growth factors (IGFs) and vascular endothelial growth factor (VEGF) in diabetic kidney disease. Curr. Pharm. Des. 10 (27), 3385–3394.
[10] Takai-Igarashi, T., Nadaoka, Y. and Kaminuma, T. (1998) A database for cell signaling networks. J. Comput. Biol. 5 (4), 747–754.
[11] Schacherer, F., et al. (2001) The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. Bioinformatics 17 (11), 1053–1057.
[12] Fukuda, K. and Takagi, T. (2001) Knowledge representation of signal transduction pathways. Bioinformatics 17 (9), 829–837.
[13] Demir, E., et al. (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics 18 (7), 996–1003.
[14] Bader, G.D., et al. (2003) Functional genomics and proteomics: charting a multidimensional map of the yeast cell. Trends Cell Biol. 13 (7), 344–356.
[15] von Mering, C., et al. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417 (6887), 399–403.
[16] Salwinski, L., et al. (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res. 32 (Database issue), D449–D451.
[17] Hermjakob, H., et al. (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res. 32 (1), D452–D455.
[18] Zanzoni, A., et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett. 513 (1), 135–140.
[19] Breitkreutz, B.J., Stark, C. and Tyers, M. (2003) The GRID: the general repository for interaction datasets. Genome Biol. 4 (3), R23.
[20] Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res. 31 (1), 248–250.
[21] Peri, S., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13 (10), 2363–2371.
[22] Matys, V., et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31 (1), 374–378.
[23] Serov, V.N., Spirov, A.V. and Samsonova, M.G. (1998) Graphical interface to the genetic network database GeNet. Bioinformatics 14 (6), 546–547.
[24] Lee, T.I., et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298 (5594), 799–804.
[25] Joshi-Tope, G., et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33 (Database Issue), D428–D432.
[26] Mewes, H.W., et al. (2002) MIPS: a database for genomes and protein sequences, 30 (1), 31–34.
[27] Jenssen, T.K., et al. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28 (1), 21–28.
[28] Rzhetsky, A., et al. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J. Biomed. Inform. 37 (1), 43–53.

[29] Bhalla, U.S., Ram, P.T. and Iyengar, R. (2002) MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. Science 297 (5583), 1018–1023.
[30] Li, F., et al. (2004) The yeast cell-cycle network is robustly designed. Proc. Natl. Acad. Sci. USA 101 (14), 4781–4786.
[31] Ronen, M., et al. (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proc. Natl. Acad. Sci. USA 99 (16), 10555–10560.
[32] Kelley, B.P., et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. USA 100 (20), 11394–11399.
[33] Regev, A., Silverman, W. and Shapiro, E. (2001) Representation and simulation of biochemical processes using the pi-calculus process algebra. Pac. Symp. Biocomput., 459–470.
[34] Jeong, H., et al. (2001) Lethality and centrality in protein networks. Nature 411 (6833), 41–42.
[35] Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4 (1), 2.
[36] Pereira-Leal, J.B., Enright, A.J. and Ouzounis, C.A. (2004) Detection of functional modules from protein interaction networks. Proteins 54 (1), 49–57.
[37] Milo, R., et al. (2002) Network motifs: simple building blocks of complex networks. Science 298 (5594), 824–827.
[38] Mangan, S. and Alon, U. (2003) Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. USA 100 (21), 11980–11985.
[39] Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein–protein interactions in yeast. Nat. Biotechnol. 18 (12), 1257–1261.
[40] Lee, I., et al. (2004) A probabilistic functional network of yeast genes. Science 306 (5701), 1555–1558.
[41] Deng, M., et al. (2002) Inferring domain–domain interactions from protein–protein interactions. Genome Res. 12 (10), 1540–1548.
[42] Gomez, S.M. and Rzhetsky, A. (2002) Towards the prediction of complete protein–protein interaction networks. Pac. Symp. Biocomput., 413–424.
[43] Bock, J.R. and Gough, D.A. (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17 (5), 455–460.
[44] Wong, S.L., et al. (2004) Combining biological networks to predict genetic interactions. Proc. Natl. Acad. Sci. USA 101 (44), 15682–15687.
[45] Ideker, T., et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 (Suppl. 1), S233–S240.
[46] Pradines, J., et al. (2004) Detection of activity centers in cellular pathways using transcript profiling. J. Biopharm. Stat. 14 (3), 701–721.
[47] Stein, L.D. (2003) Integrating biological databases. Nat. Rev. Genet. 4 (5), 337–345.
[48] Lloyd, C.M., Halstead, M.D. and Nielsen, P.F. (2004) CellML: its future, present and past. Prog. Biophys. Mol. Biol. 85 (2–3), 433–450.
[49] 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.
[50] Keseler, I.M., et al. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res. 33 (Database Issue), D334–D337.
[51] Aladjem, M.I., et al. (2004) Molecular interaction maps – a diagrammatic graphical language for bioregulatory networks. Sci. STKE 2004 (222), pe8.