Cell. Mol. Life Sci. 64 (2007) 1739 – 1751
1420-682X/07/141739-13
DOI 10.1007/s00018-007-7053-7

Cellular and Molecular Life Sciences

Network analysis approach for biology

C. K. Kwoh[a,*] and P. Y. Ng[a,b]

[a] Nanyang Technological University, Blk N4 #2A-32, School of Computer Engineering, Nanyang Avenue, Singapore 639798 (Singapore), Fax: +65 6792 6559, e-mail: asckkwoh@ntu.edu.sg
[b] Institute of Molecular and Cell Biology, 61 biopolis Drive, Proteos, Singapore 138673 (Singapore), Fax: +65 67791117, e-mail: mcblab66@imcb.a-star.edu.sg

Online First 6 April 2007

Abstract. The biological system is a complex physicochemical system consisting of numerous dynamic networks of biochemical reactions and signaling interactions between cellular components. This complexity makes it virtually unanalyzable by traditional methods. Hence, biological networks have been developed as a platform for integrating information from high- to low-throughput experiments for analysis of biological systems. The network analysis approach is vital for successful quantitative modeling of biological systems. The numerous online pathway databases vary widely in coverage and representation of biological processes. An integrated network-based information system for querying, visualization and analysis promised successful integration of data on a large scale. Such integrated systems will greatly facilitate the understanding of biological interactions and experimental verification.

Keywords. Pathway, pathway database, biological networks, information system, networks biology, drug discovery, integration.

Introduction

Biologists have always been interested in understanding the fundamental chemical features of biological processes and in ascertaining all aspects of the components through experimental design. This approach will continue to be a major aspect of basic biological research, and very much of that of modern biology. The goal is to reduce biological phenomena to the behavior of molecules as interactions between molecules to determine of the function of enormously complex machinery, both in isolation and when surrounded by other cells. Moreover, biologists are also increasingly interested in a systems-level overview where relationships among system components and processes can be investigated. This desire to gain detailed understanding of the components of a biological organism inevitably leads to a quest for better tools for visualing of how these components interact with each other in the environment in which the organism or phenomenon is embedded.

Themain interest of molecular biologists is to understand interactions among the various systems of a cell, including DNA, RNA and protein and how these interactions are regulated. Researchers have uncovered a multitude of biological facts, such as protein properties and genome sequences. But this alone is not sufficient to interpret biological systems and understand their robustness, which is one of the fundamental properties of living systems at different levels [1]. Cell, tissues, organs, organisms or any other biological systems defined by evolution are essentially complex physicochemical systems. They consist of numerous dynamic networks of biochemical reactions and signaling interactions between active cellular components. This cellular complexity has made it difficult to build a complete understanding of cellular machinery to achieve a specific purpose [2].

As advancement in accurate, quantitative experimental approaches will doubtless continue [3–6], insights into the functioning of biological systems must be augmented with information from other sources due

[*] Corresponding author.

to the intrinsic complexity of biological systems. Such efforts will require a combination of experimental and computational approaches in understanding biology as complex systems: systems biology and biological networks.

Systems biology is an academic field that seeks to integrate high-throughput biological studies to understand how biological systems function. It is aimed at interpreting and contextualizong large, diverse sets of biological data and elucidating the mechanisms underlying complex biological processes through an integrated perspective, eventually generating the ability to develop an understandable model of the whole system [7–9].

In contrast to molecular biology, systems biology does not seek to break a system down into all of its parts and study one part of the process at a time; it commonly uses controlled theoretic approaches in the hope of being able to reassemble all the parts into a whole. Some systems biologists have argued that the reductionist approach to biology must always fail, either because of nature's redundancy and complexity, or because we have not understood all the parts of the processes. In principle, all the information necessary to define the structure of the biological system of interest should be provided by its genome sequence. However, a master global reaction network could still be formulated to represent the complete repertoire of possible biochemical reaction systems within the cell based on our current resources and understanding [7–9].

Biological networks, also know as pathways, are systems that begin with the knowledge of known genes and proteins in an organism, and then use either high-throughput techniques such as microarrays to measure the changes in all messenger mRNAs (mRNA), or proteomics methods to measure changes in protein concentration, in response to a given perturbation [10, 11]. A crucial part of this process is to model the inherent stochastic nature of the system [7–9]. This information on functional molecular interactions [12], is also known as pathway databases facilitates a variety of analysis and simulation techniques to enrich our understanding of cellular systems [13].

Although the biological networks and systems biology approaches are very similar, biological networks are based more on the 'interactome' (dynamic network of biochemical reactions and signaling interactions among active proteins). Hence, they rely more heavily on systemic network analysis and other data-mining techniques compared with systems biology, which emphasizes statistical learning. There are several advantages to viewing biological interactions globally as a network over viewing these interactions as statistical learning based on binary datasets. First, confidence levels for individual interactions can be increased by analysing of networks. Previously unknown set of biological interactions can also be discovered by analysis. These often result in uncovering new interactions that may unexpectedly link diverse cellular processes or indicate crosstalk between cellular compartments [14]. The resultant molecular interaction map provides a tremendously useful framework for annotation of new knowledge. Perhaps more important, it suggests new interpretations and frames questions for productive experimentation.

In recent years, laboratory techniques such as automated DNA sequencing, global gene expression measurements, proteomics and metabonomics techniques developed for the study of molecular biology and systems biology have generated immense amounts of data. However, such data are poorly utilized because of the lack of adequate methods for interpretation in the context of biological function. To address the issue, it is useful to review the current landscape of biological networks, pathway data and work on data integration and propose techniques for efficient analysis and modeling for understanding biological function [15].

Coverage and reliability of experimental data

Historically, studies of cellular complexity as a manifestation of the enormous diversity of molecules and reaction processes needed to carry out cellular functions for growth, division, differentiation and response to extracellular factors are done through study of the functions of genes by analysis of mutant phenotypes, genetic interactions, biochemical activities, and inference by homology to other proteins of known function, and physical interactions with other proteins. These highly complex processes are controlled and regulated mainly by interactions among proteins, DNA, RNA and other compounds. Such interactions are the building blocks of the biological networks. Earlier experimental strategies were targeted at protein interactions, as most cell biological functions are mediated through proteins and protein interactions. Now, however, it has diversified [16].

Large-scale experiments have the potential to discover previously unknown functional connections among components of the cell, and thus promise to rapidly expand our knowledge of biology. The high-throughput automated strategies used to generate protein interaction maps can be classified into classical and reverse proteomics. The distinction between these two approaches is similar to the difference between forward and reverse genetics. In classical proteomics, the starting material is generally the organism of interest. Protein complexes are isolated and then analyzed, and complete genome sequences are used to identify complex components. In reverse proteomics, the starting point is the DNA sequence of the genome of an organism. First, the transcriptome (complete set of transcripts) and proteome (complete set of proteins) are predicted *in silico*. Subsequently, this information is used to generate reagents for their analysis [17]. Even though all of these approaches may be used to predict protein interactions, their directions and goals are different. Yeast two-hybrid and mass spectrometry techniques aim to detect physical binding between proteins, whereas genetic interactions, mRNA co-expression and *in silico* methods seek to predict functional associations, for example, between a transcriptional regulator and the pathway it controls. In many cases, however, such functional associations do take the form of physical binding [18, 19].

Data quality is of paramount importance in this knowledge expansion. Although technologies have accelerated the pace of discovery for biomolecular interactions, experimental interaction data obtained using different methods remain dismal. Large-scale techniques also have not shown enough internal consistency to warrant complete acceptance of the resulting data. Experiments have to be repeated before achieving high enough data quality for a particular method [20]. For instance, over 80,000 protein-protein interactions were detected in yeast by six high throughput experimental methods, but only 2400 of these interactions were supported by more than one method. Such a low overlap limits the applicability of a direct comparison between high-throughput interaction datasets of different experimental origin. The discrepancies between the interacting partners identified in high-throughput studies and those identified in small-scale experiments highlight the need for caution when interpreting results from high-throughput studies [21–23]. Therefore, enhancing the confidence of interactions by assessment and minimization of false negatives and positives will be the key issue in interpreting the results of high-throughput experimental technologies [24–26] and iterative query (hypothesis) generation, with confidence and retrospective analysis for query refinement and additional experiment designs.

Many statistical methods have been developed for enhancing the confidence of interactions derived from low confidence data and for analyzing the general parameters of the interaction datasets. One of the more direct and effective methods is to combined evidence of interaction between two or more datasets generated from different high-throughput techniques but to consider only interactions supported by agreement of two or three of any of the methods shown. However, there is normally a trade-off between coverage and accuracy when such filtering methods are used [21]. Protein microarrays, biological knowledge and biological networks (pathways) can be used as complementary methods to increase and improve both coverage and confidence in detected or predicted interactions. With the development of confidence measures and existence of a correlation distance, datasets with different interactions can be merged. Pathway information, high-throughput proteomics and genetic data can also be merged for better biological investigation. The presence of a genetic interaction between proteins close together in a network would suggest that these proteins have a high probability of being in the same complex. Proteins that are separated by one or two physical interaction links and connected by genetic interaction, together with their bridging proteins, are likely related and belong to a single complex. Subnetworks built from high-confidence interactions suggest intercomplex connections that would otherwise be obscured by low-confidence, spurious interactions. Static network topology is not sufficient to define protein function; incorporating time-dependent expression data is important for understanding pathway function. Merging expression data with proteomics data enables the function of the protein to be inferred by identifying protein complexes and pathways via clustering of expression profiles [27–29]. This filtering method will increase the confidence of the information. Even with such filtering methods to validate interactions, real success is only achievable based on iterative expert intervention with knowledge of manually curated physical protein interactions extracted from original small-scale experiments and the literature.

Network analysis approach for biology

The main motivation for building pathway databases at various detail levels managed by information systems is to facilitate merging of information from physical interactions and the literature, and qualitative and quantitative modeling of biological systems using software on powerful computers, in short using pathway data to answer biological questions. A wide range of techniques have been developed that use such pathway data to answer specific biological questions.

The overview of a framework for the network analysis approach to studying biology is shown in Figure 1. The system incorporates preprocessing to recognize and analyze quires of interest called functional snapshots.

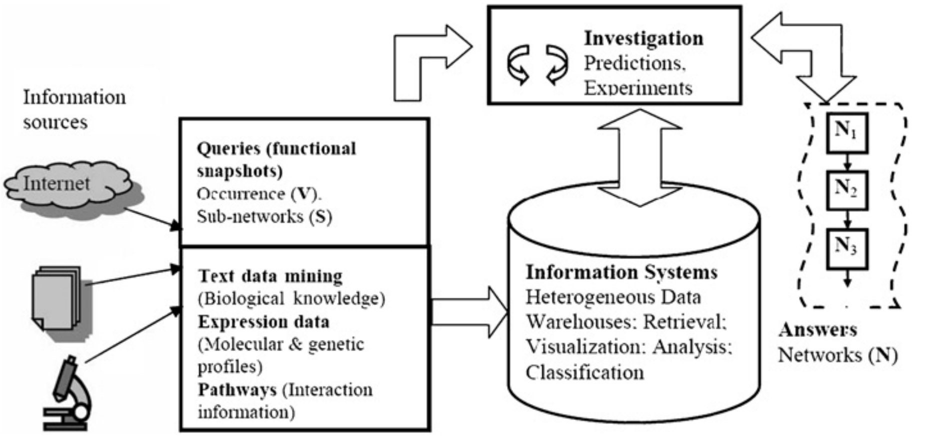


Figure 1. A framework for Network Analysis Approach to answer biological questions.

These glimpses of information can range from co-occurrence of biological entities (**V**) to more complex subnetworks (**S**) from data. The data may represent processed information such as profiles from microarray experiments that have gone through clustering [30] and feature selection processes, or textural input such as that extracted by natural language processing or entered by biologists (or investigators). The system will than use a certain scoring scheme to determine the best candidate list of corresponding networks (**N**) from the knowledge databases. Experimental data from model organisms, cell lines and human tissues can be uploaded and mapped onto the networks. New hypotheses can be made about the pathways connecting the protein of interest. This in turn can be used to guide analyses of massive data with prior/tacit knowledge, and to provide reliability indicators to the user for further actions, such as treatment or drug design with additional experiments or investigations, emphazing the iterative approach. Such systems will be able to support construction of computational models to explain cellular processes at various levels and to make testable predictions of behavior. Experimental results are compared with these predictions and used to refine the model and to design additional experiment.

Figure 2 shows collection of biological knowledge from a variety of sources. These sources include molecular interaction information, such as that in pathway databases, and is the cleanest form of information. Raw experimental data, such as molecular and genetic profiles representing experiments from microarray, ELISA and other proteomics investigations must be transformed into understandable and computationally tractable protein interaction networks with bioinformatics support in several key processing steps [24]. Biological knowledge of protein, gene and other relevant interactions obtained by text mining the extensive published literature in the MEDLINE database must be processed, prepared and stored in an information system. The end result is a network representation of the published data as support for pathways in multi-resolution networks in accordance with basic standards [31, 32]. The information systems are primarily for data and knowledge storage besides these primary requirements, they should support query, visualization, and analysis and classification. Mapping network abstractions and information from distributed data sources is essential for information tracing and to support investigative queries of the data sources.

The network of interacting networks allows interplay between the behavior, structure and function of biological system to be identified and quantified. The cell can be approached from the bottom up, moving from molecules to motifs and modules, or from top to bottom, moving to organism-specific modules and molecules. Structure, topology and network topology must be tightly interlinked, thus providing a more integrated approaches for analysis.

Biological networks databases

Despite tremendous variety in the cellular processes described as pathways, several pathway representation patterns have become standards in current practice. There are many structured databases that store interaction data collected from high-throughput experiments. These databases have been made publicly available on the Internet. They contain results of experimental data, diverse inference by computational methods and a great deal of manually curated information. The on-line resource center Pathguides (see http://www.pathguide.org/) [16] contains information about 219 biological pathway resources. These databases have been grouped into four major, slightly overlapping categories: protein interactions, metabol-

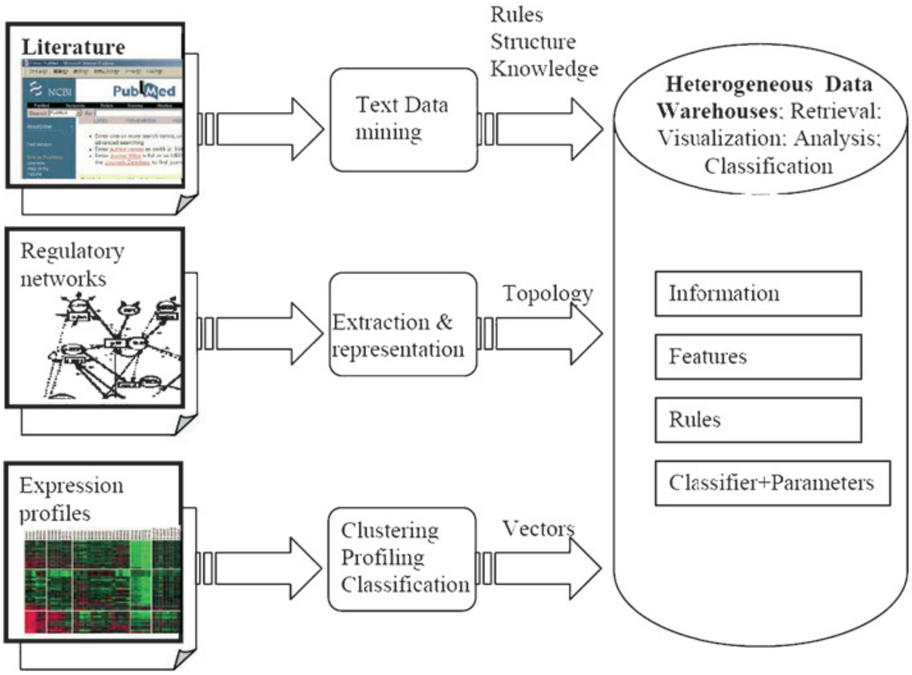


Figure 2. Knowledge processing, where data and text mining and pathways information are analysed, extracted, content understood and represented in a structural indexed database warehouse and stored.

ic pathways, signaling pathways, and transcription factors/gene regulatory networks. It also has a specific category called Pathway Diagrams.

Pathway Diagrams

In these databases, interaction is represented by its two protein partners, sometimes accompanied by basic annotations or cross-references to other databases. Some databases have further identified the most reliable subsets of interaction data. Such developments are crucial for further standardization of interaction datasets and data-exchange formats as well as for integration of the database with other bioinformatics resources [13, 33–35].

Biological networks, also know as pathways, with their non-random nature, are associated with the biological function of nodes (**V**) and edges (**E**), where E_i is $\{V_j, V_k\}$. A network is a set of interactions, or functional relationships, between the physical and/or genetic components of the cell [36], which operate in concert to carry out a biological process. Despite tremendous variety in the cellular processes described as networks, several network representation patterns are prevalent in current practice. We use these patterns to categorize network databases into four major categories: protein networks, genetic networks, metabolic networks and signaling networks. A description of the major features will provide better insight to these categories.

Protein networks

Proteins are traditionally identified on the basis of their function, which they exert based on three tertiary structures. The tertiary structure of the protein is stabilized by key residues acting as hubs in the network of interaction. Protein structure networks exhibit small-world, single-scale and, to some degree, scale-free properties when amino acids are represented by nodes, whereas edges are used to represent spatial proximity. The average shortest path-lengths are highly correlated with the residue fluctuations, providing a link between the spatial arrangement of the residues and protein dynamics. Such networks will provide a priority list of candidate residues that are most likely to affect the stability of the target protein [37, 38].

In post-genomic science, proteins are recognized as elements in complex protein interaction networks, as opposed to having a single function. In a large-scale Y2H screen, a fairly small set of highly connected proteins and domains shapes the topology of the underlying network [10, 11]. Protein networks can be generated by combining pairwise interactions as predicted by the conserved co-occurrence of their genes during expression. By quantifying the correlations between connectivities of interacting nodes and comparing them with a randomized network, the links between highly connected proteins are systematically suppressed, whereas those between highly connected and less-connected pairs of proteins are favored. This pattern decreases the likelihood of crosstalk between the different functional modules of the cell and increases the overall robustness of the network by localizing the effects of deleterious perturbations [39–41].

Genetic networks

Gene expression networks possess a more complicated hierarchical structure. The transcription regulatory network comprise a set of direct and indirect regulatory interactions between the genes. The topology of gene expression networks indicate the presence of regulatory hubs, while at the local level, network substructures such as motifs and modules are identified. Network topologies derived from gene expression data might be used to characterize entire cellular states using the global pattern of gene co-expression events instead of the frequently used fold change measure [42]. Such networks would also be useful for identifying of biological nexuses and providing new priority lists for pharmaceutical modulation, based on their importance for network structure.

Metabolic networks

Metabolic networks contain three major, fully connected subsets: a substrate, a product and a third composite subnetwork. The largest fully connected subset of every metabolic network is scale free and contains less than one-third of the nodes. Its average length approximates quite well the average path length of the whole network. The small-world architecture of metabolic networks was selected by evolution to minimize the transition times between metabolic states, while the unevenly distributed number of enzymatic reactions per metabolite ensured the stability of the system to random mutation. The restricted list of 'hubs' provides the key nodes that control the behavior of the whole system [43, 44].

Signaling networks

Signaling pathways propagate information from one part or subprocess of the cell to another via a series of protein covalent modifications, such as protein phosphorylation. Any aberrant signaling in the pathways could cause dysregulation of biological processes, resulting in diseases such as cancer and diabetes [45]. Due to the complexity of such pathways, only a relatively few networks are currently being constructed. 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 harder to study than bacteria. These types of networks are more diverse than metabolic pathways and tend to use higher-level abstractions compared to metabolic databases

The modular architecture of biological networks has unraveled new perspectives in the interaction and control of biological entities. This evidence shows that the cellular phenotypes observed at the macroscopic level depend on the collective characteristics of the underlying networks [38, 46], and thus must be studied/investigated simultaneously across scales.

As these interaction data models grow more complex, other biological information can be added. This information includes background information, and gene and proteome expression levels. None of these networks are independent, as they are interconnected, forming a network of networks that is responsible for the behavior of the cell. Therefore, the major challenge will be to integrate theoretical and experimental approaches to map out, understand and model in quantifiable terms the topological and dynamic properties of various networks.

Some popular databases

Table 1 lists popular websites in the on-line resource centre Pathguides (http://www.pathguide.org/) as indexed by Google.

Network defination and theories

Basic network nomenclature

The behavior of most complex systems emerges from the coordinated activity of many components interacting pairwise each other. These components can be reduced to a series of nodes (**V**) that are connected to each other by links or edges (**E**), with each link representing the interactions between two components. The nodes connected edges together form the basic components of networks. Physical interactions between cellular molecules are easily conceptualized using the node-link nomenclature. Nevertheless, more complex functional interactions can also be considered within this representation. Depending on the type of underlying data and interaction mechanism, edges can have associated weights and directions (directed or undirected) [24]. One of our implementations of a Web-based network for the study of cytokines in the most popular signaling pathways – COPE:Cytokines and Cells Online Pathfinder Encyclopaedia (http://www.copewithcytokines.de/) is shown in Figure 3. The data model uses a quaternary relation (node1, node2, nature of interaction, catalyst or environment). The main objective is to provide pathway visualization functionalities for intercytokine relationships, as well as for other types of relationships, with other cells for a specific cytokine(s) of interest. A natural language processor is first used to extract information from Web pages that concerns the cytokine(s) of interest. The results obtained are then further processed and displayed graphically to the

Table 1. List of popular websites in the on-line resource centre Pathguides as indexed by Google.

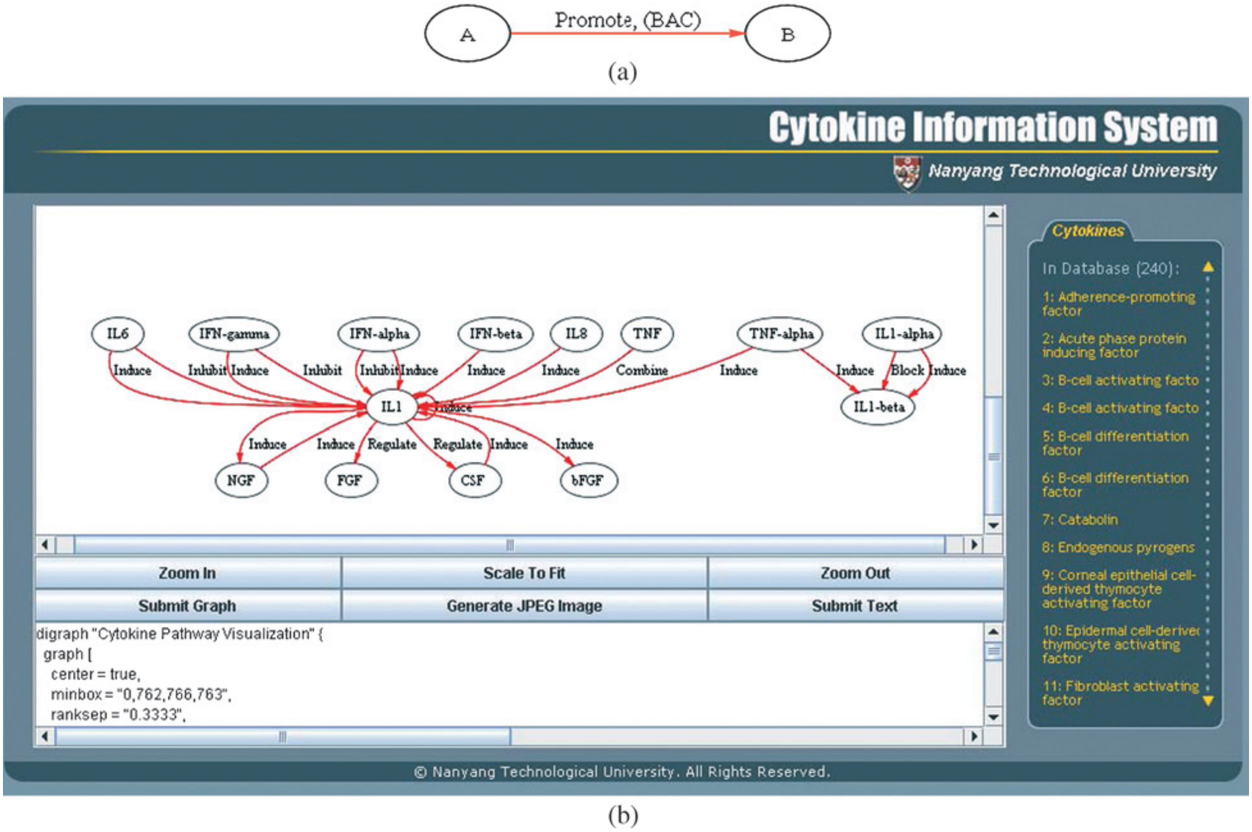
Hits	Resource	URL
	Protein-Protein Interactions	
492	DIP – Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/
435	AfCS – Alliance for Cellular Signaling Molecule Pages Database	http://www.signaling-gateway.org/
283	HPRD – Human Protein Reference Database	http://www.hprd.org/
	Metabolic Pathways	
4690	KEGG – Kyoto Encyclopedia of Genes and Genomes	http://www.genome.ad.jp/kegg/
485	BRENDA – Comprehensive Enzyme Information System	http://www.brenda.uni-koeln.de/
430	PharmGKB – The Pharmacogenetics and Pharmacogenomics Knowledge Base	http://www.pharmgkb.org/
	Signaling Pathways	
2830	COPE – Cytokines Online Pathfinder Encyclopedia	http://www.copewithcytokines.de/
435	AfCS – Alliance for Cellular Signaling Molecule Pages Database	http://www.signaling-gateway.org/
262	CST – Cell Signaling Technology Pathway Database	http://www.cellsignal.com/
	Pathway Diagrams	
4690	KEGG – Kyoto Encyclopedia of Genes and Genomes	http://www.genome.ad.jp/kegg/
430	PharmGKB – The Pharmacogenetics and Pharmacogenomics Knowledge Base	http://www.pharmgkb.org/
403	MPB – Metabolic Pathways of Biochemistry	http://www.gwu.edu/~mpb/
	Transcription Factors / Gene Regulatory Networks	
305	TRANSFAC – Transcription Factor Database	http://www.gene-regulation.com/
219	ooTFD – Object Oriented Transcription Factors Database	http://www.ifti.org/
145	PRODORIC – Prokaryotic database of gene regulation	http://prodoric.tu-bs.de/

user. Useful information such as the type of reaction and catalyst involved, if any, are also displayed. In addition, the system also offers functionalities for graphical manipulation of the visualized pathways. The system has been shown to provide a better overview, and hence improved learning to readers who are new to this field by virtue of accurate inputs obtained from the natural language processing module [47].

When represented as interaction graphs, genome-scale data offer the possibility of performing various types of analyses. At the most abstract level, a simplified network can only provide very basic biological knowledge. This type of graph has allowed biologists to visually examine the graph and map onto it various types of information, such as functional annotation, cellular localization and expression level. This mapped information allows definition of functional context, from which the biological role of individual proteins or genes can be inferred [22, 23]. Networks generated using such data cannot be analyzed mathematically. At a more sophistical level, the power of graph theory is exploited by performing various analyses that yield useful insights for each different type of complex network [15]. These binary data are generated using differential and quantified values from the experimental data. The protein interactions are converted into binary data using the matrix mode, which puts edges between all possible pairs of interactions in the same protein complex. The use of the matrix model facilitates the search for possible network motifs, modules and pathways found in complex protein networks. There are three basic types of binary data, each generated from different experimental techniques. (i) Binary interaction is based on pairwise, direct and transient associations. It is a direct measurement of physical interaction generated from two-hybrid systems and co-IP. (ii) Protein complex is based on characterization of protein complexes. It is an indirect measurement of physical interaction. Such data are generation by affinity purification-mass spectrometry (MS) approaches such as TAP and HMS-PCI. (iii) Co-expression is based on the characterization of expression patterns. This is an indirect measurement of interaction generated using microarrays and protein arrays.

Network modules

Biological networks can be broken down into groups of interacting molecules or modules. They are the elementary units of cellular networks. Modular theory states that various kinds of cellular functionality are provided by relatively small, transient but tightly connected networks of molecules that are engaged in performing specific functions. In a highly clustered


A
Promote, (BAC)
B
(a)
Cytokine Information System
Nanyang Technological University
IL6
IFN-gamma
IFN-alpha
IFN-beta
IL8
TNF
TNF-alpha
IL1-alpha
Induce
Inhibit
Induce
Inhibit
Inhibit
Induce
Induce
Induce
Combine
Induce
Induce
Block
Induce
IL1
Induce
IL1-beta
Induce
Induce
Regulate
Regulate
Induce
Induce
NGF
FGF
CSF
bFGF
Zoom In
Scale To Fit
Zoom Out
Submit Graph
Generate JPEG Image
Submit Text
Cytokines
In Database (240):
1: Adherence-promoting factor
2: Acute phase protein inducing factor
3: B-cell activating facto
4: B-cell activating facto
5: B-cell differentiation factor
6: B-cell differentiation factor
7: Catabolin
8: Endogenous pyrogens
9: Corneal epithelial cell-derived thymocyte activating factor
10: Epidermal cell-derive thymocyte activating factor
11: Fibroblast activating factor
© Nanyang Technological University. All Rights Reserved.
(b)

```
digraph "Cytokine Pathway Visualization" {
 graph [
  center = true,
  minbox = "0,762,766,763",
  ranksep = "0.3333",
```

Figure 3. Cytokine Information System and Pathway Visualization: (*a*) is a basic quaternary-attribute edge; (*b*) is a screen snapshot of the implemented Cytokine Information System.

network, functional modules can be isolated by identifying various modules of highly interlinked groups of nodes. At a local level, modules reveal specific patterns of interconnections which characterize a given network. However, not all modules are equally significant in real networks. Each real network is characterized by its own set of distinct motifs (types of module), the identification of which provides information about the typical local interconnection patterns in the network [15, 48].

In order to identify and understand the modules and their relationships in a given network, tools have been developed to identify the modularity of the networks. Figure 4 displays a snapshot of a subnetwork from PathwayBlast, software developed by the authors for matching networks in KEGG, the most popular pathway diagram [KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg/)]. Note that marking the genes IL1 and STK3 in the query, hilights intermediate node CASP (caspase 14, apoptosis-related cysteine peptidase) between the genes. The system identified and uses the IL1→ CASP→STK3 relationship for ranking the list of probable networks. Identification of such modules is a non-trivial problem, as complex networks can be parsed into subsets in many different ways, potentially generating billions of combinations. The hierarchical modularity has indicated that modules do not have a characteristic size. The network is as likely to be partitioned in a set of clusters of 10–20 components as into fewer, but larger modules. Each network is characterized by its own set of distinct motifs, the identification of which provides information about the typical local interconnection patterns in the network [49, 50]. Analysis of a network of subset of 35,000 experimentally proved human signaling interactions revealed about 2 billion linear five-step networks paths that were all physically possible. It is clear that only few of these paths are realized in any cell and at any particular time as active pathways. The convergent evolution that is seen in the transcription-regulatory network of diverse species towards the same motif types further indicate that motifs are indeed of direct biological relevance [18, 24–26, 51]. Proteins or genes of interest can be identified with biological network analysis by determining the related nodes, modules and pathways in a given condition.

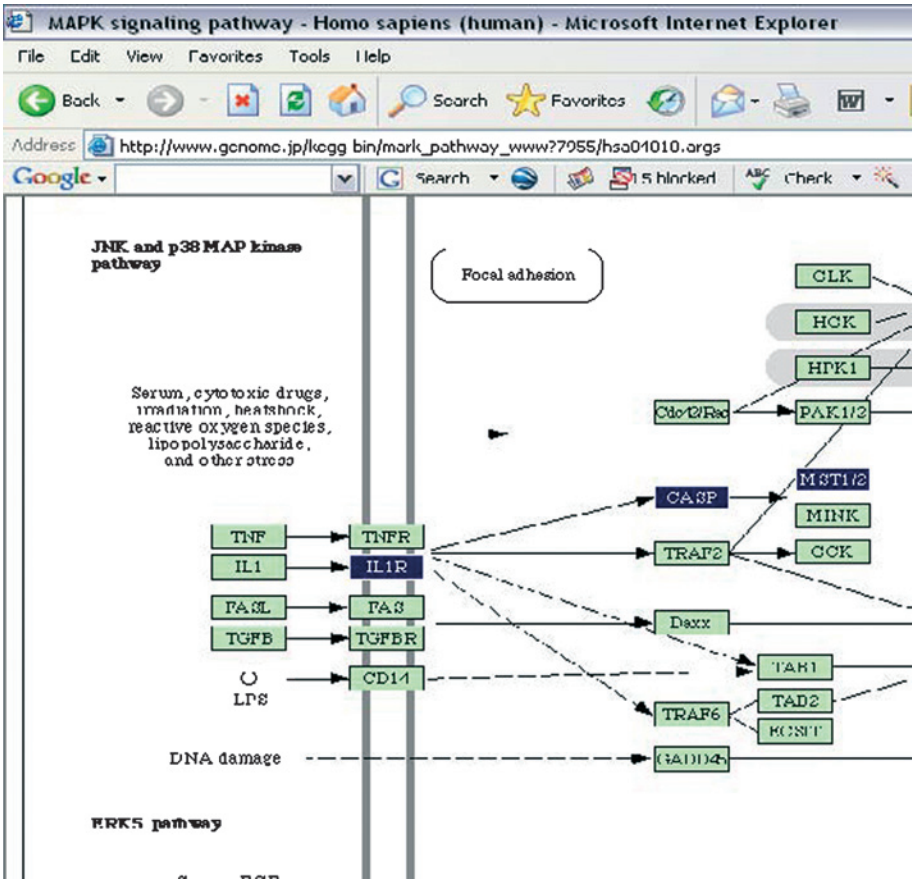

MAPK signaling pathway - Homo sapiens (human) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.genome.jp/kegg bin/mark_pathway_www?7955/hsa01010.args
Google
Search
5 blocked
Check
JNK and p38 MAP kinase pathway
Focal adhesion
Serum, cytotoxic drugs, irradiation, heatshock, reactive oxygen species, lipopolysaccharide, and other stress
CLK
HCK
HPK1
Cdc42/Rac
PAK1/2
MST1/2
CASP
MINK
TNF
TNFR
TRAF2
CCK
IL1
IL1R
FASL
FAS
Daxx
TGFB
TGFBR
TAK1
LPS
CD14
TAB2
TRAF6
ECSIT
DNA damage
GADD45
ERK5 pathway

Figure 4. PathwayBlast is a software development by authors for matching of networks with given snapshot of sub-network, note that given the 2 genes IL1 and STK3 are marked in the query, this intermediate node CASP (caspase 14, apoptosis-related cysteine peptidase) between the 2 genes is also marked. The system identified and uses the IL1→CASP→STK3 relationship (coloured in the graph) for ranking the list of probable networks.

Discussion: networks analysis of biological system

In this current stage of information explosion, massive scale experimental datasets are only meaningful when scrupulously interpreted in the context of biological processes. This interpretation, relying on the biological knowledge base, visualization, analysis and techniques of knowledge management, aided by the computation power in information systems has emerged as a crucial step in understanding the dynamic nature of cellular interactions.

In nature, the dynamics of cellular components is constantly balancing and changing depending on the state of the cell, such as undergoing differentiation,, division and apoptosis, or its present environment, such as response to external biological signal, physical/ chemical stress and virus/bacteria attacks. Therefore, at any given time, only a fraction of all possible interactions are activated and only a fraction of the cellular protein pool is active. At different time and under different conditions, different subsets of genes and proteins will be activated or repressed [14]. These functional snapshots of cellular response can be captured by high-throughput experiments, such as global gene expression, proteomics and metabonomics profiles.

The biological system is a multi-level highly complex network of information flowing between a gene and an active protein it encodes – a network of networks – to the physiological effect. This information flows from gene expression, mRNA processing, protein transport, post-translational modifications, folding and assembly into active complexes [17]. Eventually, active proteins perform certain cellular functions, which can be represented as a one-step interaction in the space of thousands of metabolic transformations regulated at multiple levels from cell membrane receptors to transcription factors. This information is tightly regulated within the cell.

Therefore, when analyzed separately, datasets obtained from these experiments cannot explain the whole picture. Intersection of the experimental data with the interaction content of the networks has to be used in order to provide the closest possible view of the activated molecular machinery in a cell. As all objects on the networks are annotated, they can be associated with one or more cellular functions, such as apoptosis, DNA repair, cell cycle checkpoints and fatty acid metabolism [52].

Biological networks are not the only method available for analysis of high-throughput data such as microarray and expression, proteomic and metabomic profiles. Other methods, such as statistical clustering, linking to pathway databases, process ontologies, pathway maps and cross-species comparisons, have also been used to reduce the number of variables in data analysis [10, 11, 53–55]. However, biological networks are the most suitable platform for functional mining of large, inherently noisy experimental datasets by their ability to provide primary information about physical connectivity between proteins, their subunits, DNA sequences and compounds via the networks' edges.

Any experimental or literature-derived datasets with recognizable gene or protein identities can be visualized, mapped and compared against each other in the same network. Experimental adjustment such as tissue type, different time points, disease and experiment-specific interaction mechanisms, and linking orthologous genes from other species are used to further enhance the common and different features between two or more different parameters. Such approaches have been widely used to identify gene(s) and protein(s) responsible for diseases [56]. The biological networks when assembled from a complete set of interactions will represented the potential of a cell to form multi-step pathways, signaling cascades and protein complexes representing the core machinery of cellular life in health and disease.

This paper has presented a summary of the current landscape of biological networks and examples of data visualization, and discussed the shape of desirable features to further facilitate biological investigation and experimental verification. Next, we will touch on the potential benefits of the network analysis approach.

Potential applications of network analysis

Drug discovery

Throughout the centuries, drug discovery has been successful through trial and error. Treatments with positive effects were retained, and unsuccessful remedies were discarded. With the advent of knowledge came new regulations. New drugs with positive effects are required to provide explanations for those effects, which requires scientifically rigorous testing of the biological mechanisms. Pharmaceutical researchers are under pressure to identify novel, promising therapeutic areas and targets quickly, while keeping costs under control. Identification of the 'right hit(s)' has became a critical part of the process because of the cost of the drug discovery. Compounding this situation is the fact that the pharmaceutical industry faces a further challenge of sustaining current and historical growth rates. In this environment, it is imperative to make highly qualified decisions about which targets to select for further development [57–61]. With the successful completion of the human genome project, biotech and pharmaceutical industries have recently turned to mapping of the human proteome, hoping that it will provide a faster and economical path in drug discovery. The development of the drug is a long and difficult process that involves numerous steps: target identification and validation; lead compound screening and design; compound optimization; compound purification and synthesis; and clinical trials [56, 60].

Study of diseases

Biological network analysis aims to describe and understand the operation of complex biological systems and eventually develop predictive models of human disease. The analysis of biological networks can define the elements of the system and characterize the flow of information that links these elements, and their networks, to any emergent biological process. Therefore, from a broad range of disease-relevant human biology, almost any high-throughput experiment, metabolic test and genetic profile data which can be linked to a gene, protein or compound can be recognized by the input parsers, visualized, analyzed and integrated into the network analyses. Most importantl, all these different datasets can be processed as networks. This data-driven dynamic modeling will play a pivotal role in the study of disease.

Networks thus represent the universal platform for data integration and analysis, which has always been the major objective of bioinformatics technology. Network analysis of complex human diseases can be broadly applied throughout drug discovery and its development pipeline. Patient data (specific DNA sequences, expression microarrays, metabolites from body fluids) can be mapped onto the networks and compares with preclinical data and published experiments. Such networks built during clinical and preclinical studies can also be used to monitor patients undergoing treatment after the clinical trial. These will be extremely useful for monitoring patients undergoing customized drug treatment in the future [59, 62, 63].

Target identification

Large datasets can be acquired for genomic and proteomic analyses. These various types of data can be integrated into a network model of how a particular biological system operates. The upstream target identification phase concentrates on discovering novel biological systems that are associated with a common disease through shared pathways or regulatory networks. Through genetic and environmental perturbations of system elements and comprehensive analyses of gene products, one can clarify the structure of the system network and delineate key nodal points (proteins). From this process, inferences can be drawn about the importance of individual proteins in the disease process. Proteins that are found to interact with members of a known disease process, and meet a more or less stringent set of criteria, thus provide good target candidates [24, 57, 62, 64]. Experimental data from model organisms, cell lines and human tissues can be uploaded and mapped onto the networks. New hypotheses can be made regarding the pathways connecting the protein of interest.

Biomarker for disease and toxicity

Drugs typically fail because they lack appropriate pharmaceutical properties. They either manifest deficiencies in absorption or pharmacokinetics of yield metabolites that have unfavorable effects. Biomarkers can be identified as 'signature networks', condition-specific conserved sets of nodes supported by differential gene expression and protein abundance data. Such signature networks can also be derived from toxic genomics data. With development of new spectroscopic tools that permit the simultaneous enumeration of thousands of metabolic products in biological fluids, the metabolic profiles assembled using these spectra could identify off-target toxicity profiles with patterns of accumulation of certain metabolites [57]. Such data can be integrated into relevant networks during the lead optimization phase.

The future

Despite recent advancements, analysis of biological networks and its applications are only in their infancy. As different technologies from various fields are required for generating and analyzing networks, further evolution of the networks will depended on the development of technologies surrounding them. New theoretical methods are required to characterize network topology into the dynamics of nodes, motifs, modules, pathways and biological functions. In order to increase our level of understand and knowledge, data collection abilities must be enhanced by development of highly sensitive tools for identifying and quantifying various types interactions within the biological networks and 'signature networks'.

The main advantage of biological network analysis for drug discovery in the ability to analyse different biological networks (systems) at different levels under different functional and temporal states. As technology progresses, network analysis will scaled up to accommodate large sets of disease-related molecular data, such as gene, proteomic and metabolic expression profiling and other new measurement of the biological system. More advanced mathematical and mapping models to integrate, investigate and analyze these large, complex sets of biological networks will provide an accurate and detailed understanding of biological pathways. These will subsequently form the basis of a better understanding of disease, tracking of its progression, and new drug discovery, development and application strategies.

When this happens, drug targets will shift from single proteins, to functional protein complexes, to whole networks. Together with new tools to prioritize biochemical experiments, targets and leads, single active molecule therapeutics will be replaced by molecular cocktails with components that target protein hubs in disease-associated molecular networks. Drug cocktails can further be specially customized for each different individual to maximize the drug effect and minimize side effects.

Acknowledgements. We thank my colleagues Dr Lim Ee Peng, Dr Vivekanand Gopalkrishnan, Dr He Yulan and Dr Salil K Bose and Dr Ng See-Kiong and Dr Li Jinyan of the Institute of Infocomm Research (Singapore) for useful discussions. We also thank Tan Shengyang and Tan Wenjie for implementing the software in their projects. The research was partly funded by the NTU-SGH Biomedical Engineering Initiative and SEED GRANT SPF 022–2005 from the Biomedical and Pharmaceutical Engineering Cluster, College of Engineering, Nanyang Technological University (NTU).

1 Kitano, H. (2006) Robustness from top to bottom. Nat. Genet. 38, 133.

2 Kitano, H. (2002) Computational systems biology. Nature 420, 206 – 210.

3 Yang, S., Ghanny, S., Wang, W., Galante, A., Dunn, W., Liu, F., Soteropoulos, P. and Zhu, H. (2006) Using DNA microarray to study human cytomegalovirus gene expression. J. Virol. Methods 131, 202 – 208.

4 Wang, Y., Hewitt, S. M., Liu, S., Zhou, X., Zhu, H., Zhou, C., Zhang, G., Quan, L., Bai, J., and Xu, N. (2006) Tissue microarray analysis of human FRAT1 expression and its correlation with the subcellular localisation of beta-catenin in ovarian tumours. Br. J. Cancer 94, 686 – 691.

5 Horiuchi, K.Y., Wang, Y., and Ma, H. (2006) Biochemical microarrays for studying chemical biology interaction: DiscoveryDot technology. Chem. Bio. Drug Dis. 67, 87 – 88.

6 Poetz, O., Schwenk, J.M., Kramer, S., Stoll, D., Templin, M.F., and Joos, T.O. (2005) Protein microarrays: catching the proteome. Mech. Ageing Dev. 126, 161 – 170.

7 Kitano, H. (2002) Systems biology: a brief overview. Science 295, 1662 – 1664.

8 Dhar, P.K., Zhu, H., and Mishra, S.K. (2004) Computational approach to systems biology: from fraction to integration and beyond. IEEE Trans. Nanobiosci. 3, 144 – 152.

9 Westerhoff, H.V. and Palsson, B.O. (2004) The evolution of molecular biology into systems biology. Nat. Biotechnol. 22, 1249 – 1252.

10 Zheng, Y., and Kwoh, C.K. (2005) Identifying simple discriminatory gene vectors with an information theory approach. In: IEEE Computational Systems Bioinformatics Conference CSB2005, IEEE Computer Society Press.

11 Zheng, Y., and Kwoh, C.K. (2005) Classifying eukaryotes with the discrete function learning algorithm. In: 3rd Asia-Pacific Bioinformatics Conference, APBC 2005.

12 Ideker, T., Galitski, T., and Hood, L. (2001) A new approach to decoding life: systems biology. Annu. Rev. Genomics Hum. Genet. 2, 343 – 372.

13 Cary, M.P., Bader, G.D., and Sander, C. (2005) Pathway information for systems biology. FEBS Lett. 579, 1815 – 1820.

14 Mayer, M.L., and Hieter, P. (2000) Protein networks-built by association. Nat. Biotechnol. 18, 1242 – 1243.

15 Barabasi, A.L., and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101 – 113.

16 Bader, G.D., Cary, M.P., and Sander, C. (2006) Pathguide: a pathway resource list. Nucleic Acids Res. 34, D504 – 506.

17 Walhout, A., and Vidal, M. (2001) Protein interaction maps for model organisms. Nat. Rev. Mol. Cell Biol. 2, 55 – 62.

18 Ge, H., Liu, Z., Church, G.M., and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet. 29, 482 – 486.

19 Ge, H., Walhout, A.J., and Vidal, M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet. 19, 551 – 560.

20 Bader, G.D., and Hogue, C.W. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. Nat. Biotechnol. 20, 991 – 997.

21 von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399 – 403.

22 von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33, D433 – 437.

23 von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31, 258 – 261.

24 Schachter, V. (2002) Protein-interaction networks: from experiments to analysis. Drug Discov Today 7, S48 – 54.

25 Schachter, V. (2002) Construction and prediction of protein-protein interaction maps. Ernst Schering Res. Found. Workshop, 191 – 220.

26 Schachter, V. (2002) Bioinformatics of large-scale protein interaction networks. Biotechniques Suppl, 16 – 18, 20 – 14, 26 – 17.

27 Bader, J.S., Chaudhuri, A., Rothberg, J.M., and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks. Nat. Biotechnol. 22, 78 – 85.

28 D'Haeseleer, P., Liang, S., and Somogyi, R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707 – 726.

29 Templin, M.F., Stoll, D., Bachmann, J., and Joos, T.O. (2004) Protein microarrays and multiplexed sandwich immunoassays: what beats the beads? Comb. Chem. High Throughput Screen, 7, 223 – 229.

30 Wuchty, S., Barabasi, A.L., and Ferdig, M.T. (2006) Stable evolutionary signal in a Yeast protein interaction network. BMC Evol. Biol. 6, 8.

31 XML Belief Network File Format.

32 BioPAX (http: www.biopax.org).

33 Salwinski, L., and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. Curr. Opin. Struct. Biol. 13, 377 – 382.

34 Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32, D449 – 451.

35 Salwinski, L., and Eisenberg, D. (2004) In silico simulation of biological network dynamics. Nat. Biotechnol. 22, 1017 – 1019.

36 Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. et al. (2004) Global mapping of the yeast genetic interaction network. Science 303, 808 – 813.

37 Greene, L.H., and Higman, V.A. (2003) Uncovering network systems within protein structures. J. Mol. Biol. 334, 781 – 791.

38 Atilgan, A.R., Akan, P., and Baysal, C. (2004) Small-world communication of residues and significance for protein dynamics. Biophys. J. 86, 85 – 91.

39 Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623 – 627.

40 Uetz, P., and Finley, R.L., Jr. (2005) From protein networks to biological systems. FEBS Lett. 579, 1821 – 1827.

41 Wuchty, S. (2002) Interaction and domain networks of yeast. Proteomics 2, 1715 – 1723.

42 Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004) Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14, 283 – 291.

43 Ravasz, E., Somera, A. L., Mongzu, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002) Hierarchical organization of modularity in metabolic networks. Science 297, 2552 – 1555.

44 Ma, H.W., and Zeng, A.P. (2003) The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 19, 1423 – 1430.

45 Hahn, W.C., and Weinberg, R.A. (2002) Modelling the molecular circuitry of cancer. Nat. Rev. Cancer 2, 331–341.

46 Yook, S.H., Oltvai, Z.N., and Barabasi, A.L. (2004) Functional and topological characterization of protein interaction networks. Proteomics 4, 928 – 942.

47 Tan, S.K., and Chee Keong (2005) Cytokine information system and pathway visualization. In: International Joint Conference of InCoB, AASBi and KSBI (BIOINFO2005), pp. 10 – 14, Pusan, Korea.

48 Wodak, S.J., and Mendez, R. (2004) Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. Curr. Opin. Struct. Biol. 14, 242 – 249.

49 Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004) Superfamilies of evolved and designed networks. Science 303, 1538 – 1542.

50 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002) Network motifs: simple building blocks of complex networks. Science 298, 824 – 827.

51 Gavin, A.C., Bosche, M., Krause, R., Graudi, P., Marzioch, M., Bauer, A., Schultz, J., Rick J. M., Michon, A. M., Cruciat, C. M. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141 – 147.

52 Wodak, S.J., and Janin, J. (2002) Structural basis of macromolecular recognition. Adv. Protein. Chem. 61, 9 – 73.

53 Zheng, Y.K. and Chee Keong (2005) A feature vector selection method for cancer classification. In: International Joint Conference of InCoB, AASBi and KSBI (BIOINFO2005) pp. 23 – 28, Pusan, Korea.

54 Zheng, Y., and Kwoh, C.K. (2006 (accepted)) Informative MicroRNA Expression Patterns for Cancer Classification. In: PAKDD2006 BioDM Workshop.

55 Zheng, Y., and Kwoh, C.K. (2006) Dynamic algorithm for inferring qualitative models of gene regulatory networks.International Journal of Data Mining and Bioinformatics 1, 111 – 137.

56 Somogyi, R., and Greller, L.D. (2001) The dynamics of molecular networks: applications to therapeutic discovery. Drug Discov. Today 6, 1267 – 1277.

57 Hood, L., and Perlmutter, R.M. (2004) The impact of systems approaches on biological problems in drug discovery. Nat. Biotechnol. 22, 1215 – 1217.

58 Voit, E.O. (2002) Metabolic modeling: a tool of drug discovery in the post-genomic era. Drug Discov. Today 7, 621 – 628.

59 Bredel, M., and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet 5, 262 – 275.

60 Butcher, E.C., Berg, E.L., and Kunkel, E.J. (2004) Systems biology in drug discovery. Nat. Biotechnol. 22, 1253 – 1259.

61 Butcher, E.C. (2005) Can cell systems biology rescue drug discovery? Nat. Rev. Drug Discov. 4, 461 – 467.

62 Flook, P.K., Yau, L. and Szalma, S. (2003) Target validation through high throughput proteomics analysis. DDT Target 2, 217 – 223.

63 Bredel, M., Bredel, C., and Sikic, B.I. (2004) Genomics-based hypothesis generation: a novel approach to unravelling drug resistance in brain tumours? Lancet Oncol. 5, 89 – 100.

64 Hood, L., Heath, J.R., Phelps, M.E., and Lin, B. (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306, 640 – 643.

To access this journal online:
http://www.birkhauser.ch/CMLS