DRUG DISCOVERY TODAY BIOSILICO

Biological networks and analysis of experimental data in drug discovery

Yuri Nikolsky, Tatiana Nikolskaya and Andrej Bugrim

Cellular life can be represented and studied as the 'interactome' – a dynamic network of biochemical reactions and signaling interactions between active proteins. Systemic networks analysis can be used for the integration and functional interpretation of high-throughput experimental data, which are abundant in drug discovery but currently poorly utilized. The composition and topology of complex networks are closely associated with vital cellular functions, which have important implications for life science research. Here we outline recent advances in the field, available tools and applications of network analysis in drug discovery.

Over the past several years there has been a paradigm shift in life science research as a result of the unprecedented advances in several laboratory techniques, such as automated DNA sequencing, global gene expression measurements, and proteomics and metabonomics techniques. The high throughput data collectively referred to as 'OMICs' data are ubiquitous throughout the drug discovery pipeline from target identification and validation to the development and testing of drug candidates. However, OMICs data are poorly utilized because of the lack of adequate methods for interpretation in the context of disease and biological function. Although bioinformatics has developed robust statistical solutions for evaluation of the significance and clustering of data points, statistics alone does not explain 'the underlying biology'.

The complexity of our own biology requires a system-wide approach to data analysis, which can be defined as the integration of 'OMICs' data using computational methods [1]. It is clear from the field that the identification of the 'part list' of all the genes and proteins is insufficient to understand the whole. It is the assembly of these parts (the general schema, the modules and elements) and the dynamics of changes in response to stimuli that is truly the key to understanding life, form and function [2,3]. The assembly of 'cellular machinery' is most effectively presented as the 'interactome', the network of interconnected signaling, regulatory and biochemical networks with proteins as the nodes and physical protein–protein interactions as edges [4,5]. As in many fields of science, technology and social life, the topology and dynamics of complex networks can be studied by graph theory [5]. The information about protein interactions has been collated from the vast amount of published experimental data annotated and assembled in interactions databases. Network data analysis tools are now commercially available and robust enough for simultaneous processing of multiple 'whole genome' data files, such as human expression microarrays. Recently, the interpretation of experimental 'OMICS' datasets in the context of accumulated knowledge on human functional networks was described as the first step in studying complex systems [2,6]. Now, we can consider the building of the basic framework, databases and logistics needed to accomplish this. Networks-centered data analysis is well underway in the major pharmaceutical companies. In this review, we will

Yuri Nikolsky*
Tatiana Nikolskaya
Andrej Bugrim
GeneGo,
500 Renaissance Drive, #106,
St. Joseph,
MI 49085,
USA
*e-mail: yuri@genego.com

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03420-3

describe the state of the field, present the available tools and suggest the applications of networks analysis throughout the drug discovery pipeline.

Protein interactions as building blocks for biological networks

There are multiple ways to elucidate protein–protein interactions. One approach is to screen experimental literature, using text mining algorithms, for co-occurrence (therefore, association) of gene/protein symbols and names in the same text. Typically, Natural Language Processing (NLP) and other text-mining algorithms are used for the automated 'mining' of abstracts and titles of PubMed articles [7–9]. It is generally believed (pers. commun. from pharmaceutical researchers), and supported by comparative studies, that up to a half of NLP associations do not correspond to experimentally verified protein interactions [9,10], although over 60% of shown interactions can be elicited by automated text mining [11]. The reliability of NLP-derived associations can be enhanced by the compilation of field-specific synonyms dictionaries, using longer word strings for searching and full-text articles to query against, and statistical methods (reviewed in [9]) In a recent study, the NLP engine MedScan was used to extract 280,000 functional relations including 20,000 protein interaction facts between human proteins from full text articles with a precision of 91% for 361 randomly extracted protein interactions [12].

The interactions can also be derived from high-throughput experimentation. For example, the yeast 2-hybrid (Y2H) screen test identifies protein interactions in yeast cells [13]. A widely used wet laboratory technique, Y2H was scaled-up for global mapping of protein interactions in yeast [14], fly *D. melanogaster* [15] and worm *C. elegans* [16] and became the technology base for several tools and discovery companies such as Curagen (www.curagen.com) and Hybrigenics (www.hybrigenics.fr). However, Y2H-derived interactions are known for a high (over 50%) level of false positives and false negative interactions [17,18]. In a recent study, over 70,000 interactions for 6,231 human proteins were predicted assuming the interactions between these proteins' orthologs in yeast, worm and fly [19]. The accuracy of predicted interactions remains questionable. Although it was assessed computationally based on relative correspondence of interacting protein pairs to gene ontology processes, the interactions were not directly compared with any high-confidence experimental set now available [20,21]. The interactions can also be deduced from condition-specific co-occurrence of gene expression based on the assumption that interacting proteins must be expressed in sync [22], especially when encoded by the homologous genes [23]. Abundant and readily obtainable even from small cell populations, it is often stated that co-expression-based clustering will become the major approach for determining tissue-, disease- and treatment-specific interactions. However, the overall confidence in co-expression-derived interactions in yeast is about 50% (47% anti-correlation for novel interactions) [24,25]. Another method, co-immunoprecipitation (Co-IP) consists of the affinity precipitation of protein complexes in mild conditions using antibodies to one of the complex's subunits, followed by mass-spectrometry or western blot analysis. Unlike the discussed above methods, Co-IP enables direct and quantitative detection of interactions between active proteins, and so it is a true proteomics method. Co-IP was used in simultaneously published studies of the yeast interactome [26,27]. The other, less commonly used experimental and computational methods include protein arrays, fusion proteins, neighbor genes in operons (for prokaryotic proteins), paralogous verification method (PVM), co-localization, synthetic lethality screens and phage display; each method with its merits and biases (reviewed in [28,29]). The overall confidence in interactions, defined as the intersection between interacting pairs obtained using different methods, remains dismal. For instance, over 80,000 protein–protein interactions were detected in *S.cerevisiae* by six high-throughput (HT) experimental methods, but only 2,400 of these interactions were supported by more than one method [30]. Such low overlap limits the applicability of a direct comparison between HT interactions datasets of different experimental origin. Recently, statistical methods were developed for enhancing the confidence of interactions derived from low confidence data and for analyzing the general parameters of interaction datasets [31,32]. Y2H and Co-IP yeast protein interaction data applied in yeast were extensively compared for experimental biases and correlation [31]. Although only 6% of Y2H interactions were confirmed by the Co-IP method, the authors managed to develop a statistical regression model for prediction of biological relevance and confidence of HT interactions based on sub-network analysis [31]. In another study, graph-theoretical statistics were used for comparative analysis of the interaction datasets in yeast [32]. The parameters and algorithms were realized in the publicly available tool TopNet for comparison of biological sub-networks of different origin (http://networks.gersteinlab.org/genome/interactions/networks/core.html).

In general, it is believed that only manually curated physical protein interactions extracted from original 'small scale' experimental literature can be used with sufficient confidence [28,29].

Dozens of the original and compilation academic protein–protein and protein–DNA interaction databases are available, covering high-throughput and 'small scale' experimental interactions experimentally and computed interactions. We have outlined some key database projects, pathways database and analytical tools in Table 1.

Architecture and composition of biological networks

Biological networks are presented as nodes (proteins, genes and compounds) connected by edges (protein–protein,

TABLE 1

Tools and databases for network analysis

Name	Description	URL address
Protein interaction databases		
BIND	A curated database of interactions, derived both from the literature and experimental datasets. 8,500 interactions are deduced from high-confidence small scale experiments from multiple species. BIND can be used for querying and as a browser [65]	http://bind.ca
DIP	A database of experimentally determined protein–protein interactions, mostly from yeast. Around 10% of DIP interactions are derived from high confidence small scale experiments [29,66]	http://dip.doe-mbi.ucla.edu/
HPRD	Human Protein Reference Database provides curated human-specific protein interactions; currently >22,000 interactions for >10,000 human proteins. It also contains 7 signaling maps. HPRD is used as a browser for interactions, protein annotations, motifs and domains [20]	http://www.hprd.org/
MetaCore database	A manually curated interactions database for >90% human proteins with known function [53,56]. Content of MetaCore (see below)	http://www.genego.com.
MINT and HomoMINT	A searchable interaction database with total of 40,000 interactions, mostly from yeast and fly. 70% of interactions are from lower-confidence Y2H screens. Only 3800 interactions include human proteins [67]	http://mint.bio.uniroma2.it/mint/
MIPS	A well-known searchable database on high-quality small scale experiments protein–protein interactions in yeast [68] and most recently mammals [21]. Several hundred human interactions	http://mips.gsf.de
PathArt database	A manually curated database of ~7,500 protein–protein and protein–compound interactions and pathways. Content of PathArt (see below)	http://jubilantbiosys.com
Pathway Analysis database	The mammalian interactions content of Pathway Analyst (see below). The number of interactions is not announced	http://www.ingenuity.com
ResNet database	Automatically extracted and manually validated database of human protein interactions (>30,000), transcriptional regulation (10,000), protein modifications (10,000) and functional regulations (350,000) [12]. Content of PathwayAssist (see below).	http://www.ariadnegenomics.com
STRING	A database of known and predicted protein interactions deduced from over 110 genomes, high-throughput experiments and gene co-expression [69]	http://string.embl.de
Pathways maps and process ontologies		
BIOCarta	A commercial collection of ~350 maps on human biology representing canonical pathways	http://www.biocarta.com/genes/index.asp
Gene Ontology	The most often referred to publicly available protein classification based on cellular processes developed by Gene Ontology Consortium [52]	http://www.geneontology.org
GenMAPP	Gene MicroArray Pathway Profiler is a database of GO-derived diagrams designed for viewing and analyzing gene expression data [17]	http://www.genmapp.org
KEGG	A well known database of generic metabolic maps for bacteria and eukaryotes. Recentky added some regulation maps. Software allows comparison of genome maps, graph comparison and path computation [70]	http://www.genome.jp/kegg/pathway.html
MetaCore, pathway module	A part of the commercial tool MetaCore™. The pathways module contains 350 interactive maps for >2,000 established pathways in human signaling, regulation and metabolism. HT data can be superimposed on the maps and networks built for any object	http://www.genego.com
Protein Lounge	A commercial package with ~300 human metabolic and signaling maps	http://www.proteinlounge.com
Network data mining suites		
MetaCore/ MetaDrug, GeneGo,Inc	An integrated analytical suite based on a manually curated database of human protein–protein and protein–DNA interactions. All types of HT data can be used for building networks. Medicinal chemistry module allows predicting human metabolism and toxicity for novel compounds. Networks are connected to functional processes, 350 proprietary metabolic and signaling maps. Web access or enterprise solution	http://www.genego.com
Pathway Analyst, Ingenuity, Inc.	An integrated analytical suite based on a manually curated database of literature-derived mammalian protein–protein interactions. Visualization on networks and analysis of HT data. Networks are connected to GO processes, 60 KEGG metabolic maps and Cell Signaling Inc.'s signaling maps. Web access, enterprise solution	http://www.ingenuity.com
PathArt, Jubilant Biosystems	A curated database of generic protein interactions, pathways and bioactive molecules supported by HT data parsers and visualization tools. Connectivity with ligand databases, GO categories. Web access	http://www.jubilantbiosys.com/pd.htm
PathwayAssist, Ariadne Genomics	A software tool for mapping the HT data on networks, maps and pathways. The source of interactions data are NLP mining of PubMed abstracts. PathwayAssist is bundled with Jubilant and Integrated Genomics pathways content. A desktop product	http://ariadnegenomics.com

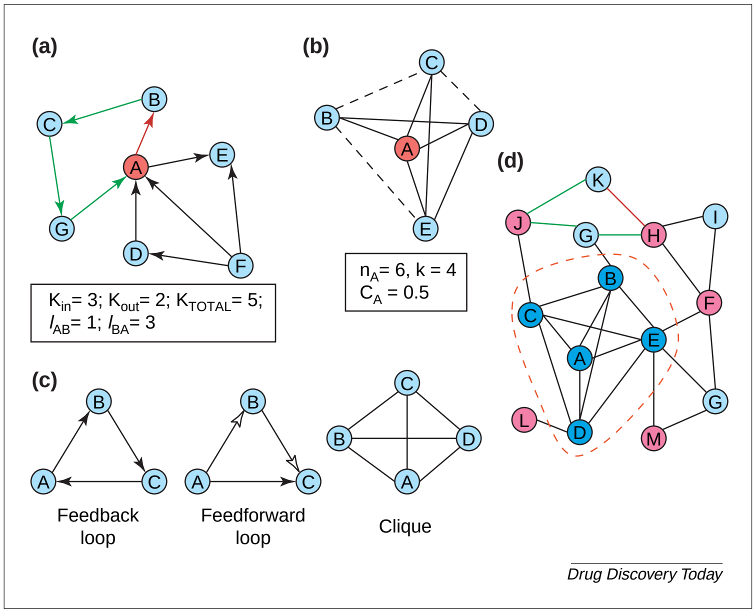


FIGURE 1

Network architecture and analysis. (a) A directed network has three types of node degree: K_{in}, K_{out}, and K_{total} (shown for node A). The shortest paths between two nodes depend on direction, e.g.: A→B, red; B→A, green. **(b)** Calculation of a clustering coefficient for node A (solid lines show actual interactions, dashed lines other possible connections). **(c)** Several types of network motifs (clique is a pattern where every node connected to every other node); **(d)** When analyzing expression data within networks, both similarity of expression and connectivity are taken into account: nodes A to E represent a potential condition-specific module – they have similar expression patterns and are connected into a tight cluster. Nodes F, H, J, L and M also share an expression pattern, but are not well connected, hence are less likely to constitute a functional pathway.

protein–gene, protein–compound interactions and metabolic reactions). Depending on the type of underlying data and the interaction mechanism, the edges are either directed or undirected. For instance, protein binding interactions derived from Y2H assays are undirected, while most of the physical interactions extracted from full text articles have one direction (e.g. protein A activates protein B, but not vice versa) There are several major parameters the networks can be described and compared with (reviewed in [32]; see Figure 1).

(1) Average degree (K) is the average number of edges per node. In directed networks one can distinguish incoming degree (Kin), outgoing degree (Kout) and total degree.

(2) Average clustering coefficient (C) is the average ratio of the actual number of links between the node's neighbors and the maximum possible number of links between them; The clustering coefficient for the node i can be calculated as Ci= 2ni/k(k-1), where ni is the actual number of links connecting k neighbors of the node to each other Figure 1b.

(3) Shortest path l AB for the pair of nodes is the minimum number of network edges that need to be passed to travel from A to B. On a directed graph the shortest path from A to B may be different from the path from B to A as shown on Figure 1a. Characteristic path length (L) is the average length of 'shortest paths' for all pairs of nodes on the graph.

(4) Diameter (D) is the longest distance between a pair of nodes on the graph. The 'default' random network theory states that pairs of nodes are connected with equal probability and the degrees follow a Poisson distribution. This implies that it is very unlikely for any node to have significantly more edges than average [32]. The analysis of the yeast interactome (the best studied organism in terms of interactions) revealed that the networks are remarkably non-random and the distribution of edges is very heterogeneous, with a few highly connected nodes (hubs) and the majority of nodes with very few edges. Such topology is defined as scale-free, meaning that the node connectivity obeys a power law: P(k) ~ k–g, where and P(k) is the fraction of nodes in the network with exactly k links [33]. Interestingly, the hubs are predominantly connected to low-degree nodes, a feature that gives biological networks the property of robustness. A removal of even a substantial fraction of nodes still leaves the network connected [34]. At the level of global architecture, networks of different origin (e.g. metabolic, regulatory, protein interactions, networks for different organisms) share the same properties [15,33,35,36]. Taken together, the metabolic reactions and signaling interactions form a large cluster linked via molecular nodes shared among many cellular processes [18]. This runs contrary to a 'traditional' model of small and relatively independent linear pathways.

Network modules and substructure search

The key property of biological networks is their modular nature [4]. According to modular theory, various kinds of cellular functionality are provided by relatively small, transient but tightly connected networks of molecules (5–25 nodes) that are engaged in performing specific functions. 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. For example, our analysis of the network of a subset of 35,000 experimentally proven human signaling interactions in the MetaCore™ database revealed approximately 2 billion linear 5-step network paths that were all physically possible. It is clear that only few of these paths are realized in any cell and at any particularly time as active pathways.

Different approaches have been offered for automated parsing of large networks into modules. One set of methods identifies the modules using various clustering algorithms. These include the Monte Carlo optimization method for finding tightly connected clusters of nodes [37], clustering based on shortest-paths-length distribution [38] and unsupervised graph clustering [39]. It was shown that some clusters identified using approaches such as these correspond either to known protein complexes or to metabolic pathways [37]. Another approach consists of parsing networks into motifs, defined as fairly simple subgraphs that share certain structural and functional features, such as a feedback or feed-forward loop [40] (Figure 1c). The number of different motifs in the network is calculated and then compared with the number of the same motifs in a randomly connected network. Those motifs in which the network is enriched when compared to the random network may represent potential functional modules. Such motifs were identified in regulatory networks of *E. coli* [41] and yeast [42]. It should be noted that performance of these algorithms is usually judged by how well they can recall known functional units or processes. On this account, all these algorithms are prone to a high level of false-positives (i.e. recalling modules that do not correspond to any known pathways.

Conditionally active functional modules can also be elucidated by the analysis of high-throughput molecular data (e.g. gene expression, protein abundance, metabolic profiles) in the context of networks. One straightforward approach relies on statistical clustering of gene expression data followed by mapping the resulting clusters onto the networks obtained from independent sources [43]. The advantage of this approach is the prioritization of gene clusters based on the number of links to the network. The drawback is that the statistics-derived clusters are inherently artificial and can be connected to multiple networks and cellular processes. In another method, the network clustering algorithms, such as superparamagnetic clustering, are used to identify tightly connected sets of nodes. The expression data helps to assign weights to the edges and nodes; the combined distance is then computed based on both expression profiles and the network distance between gene products [44]. Other methods include simulated annealing [45] and probabilistic graphical models [46]. Essentially, analysis of molecular data within the context of interaction networks reveals genes that share a similar pattern of expression and at the same time are closely connected on the network (Figure 1d). Another important way of finding putative functional pathways is comparison of networks derived from different data sources. For example, a heuristic graph comparison algorithm was developed for finding functionally related enzymes clusters (FRECS) across bacterial species [47] and between protein and gene expression networks [48]. Another algorithm allows one to identify common interaction pathways by inter-species alignment of protein interaction networks; for example between yeast *S. cerevisiae* and bacterium *H. pylori* [49].

Biological implications of network topology

The non-random nature of biological networks is associated with the biological functions of nodes and edges. Recently, several studies in yeast revealed correlations between the topology and composition of a network and important biological properties of nodes [17,18,50,51]. The well connected hubs (defined here as the top quartile of all nodes in terms of the number of edges) are largely represented by evolutionary conserved proteins because the interactions impose certain structural constraints on sequence evolution [50]. In both *S. cerevisiae* and *C. elegans*, a significant negative correlation was shown between the number of interactions and the relative evolutionary rate [50]. Recently, it was revealed that the number of interactions a protein has positively correlates with its relative importance in yeast [18,51]. Essential and 'marginally essential' (relative importance of a non-essential gene to a cell) genes tend to be hubs with short characteristic path length to the neighbors [51]. Essential proteins tend to be more closely connected to each other. Furthermore, essential proteins tend to be the more promiscuous transcription factors and target genes that are regulated by fewer transcription factors. Many of these targets are 'housekeeping' genes with high expression levels and less expression fluctuation [51]. It was also noted that soluble proteins possess more interactions than membrane proteins [32]. As mentioned above, the links between highly-connected and low-connected pairs of proteins defines the specific topology of a network. In yeast, the direct links between highly connected hubs are suppressed and the interactions between hubs and low-connected node pairs are favored. Such topology probably prevents cross-talk between the functional modules-subnetworks [17]. The findings may have substantial implications for the practice of drug discovery in terms of target prioritization and identification of multi-gene/multi-protein biomarkers.

Mapping and interpretation of experimental data on the networks

Biological networks are the most suitable tool for functional mining of large, inherently noisy experimental datasets such as microarray and SAGE expression patterns, proteomics and metabonomic profiles. There is an important distinction between networks and the other methods available for HT data analysis (such as statistical clustering, linking to pathway databases, process ontologies, pathway maps, cross-species comparisons etc.). Unlike other methods, networks' edges provide primary information about physical connectivity between proteins, their subunits, DNA sequences and compounds. The complete set of interactions which assembles into networks, defines 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. Obviously, only a fraction of all possible interactions is activated at any given condition as only some of the genes are expressed in tissues at a time and only a fraction of the cellular protein pool is active. The subset of activated (or repressed) genes and proteins is captured by OMICs experiments, such as global gene expression profiles, proteomics or metabonomics profiles – the functional snapshots of cellular response. Analyzed separately, these datasets cannot explain the whole picture. There are many levels of information flow between a gene and an active protein it encodes, including gene expression, mRNA processing, protein trafficking, posttranslational modifications, folding and assembly into active complexes (Figure 2). Eventually, active proteins perform certain cellular functions (such as a metabolic transformation of malonyl into acetyl-CoA in this example), which can be presented as one-step interactions in the space of thousands of metabolic transformations regulated at multiple levels from the cell membrane receptors to transcription factors. The 'intersection' of the experimental data with the interactions content on the networks (derived from experimental literature) provides 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 or fatty acid metabolism. The networks can be interpreted in terms of these higher level processes, and the mechanism of an effect can be unraveled. This is achieved by linking the network objects to Gene Ontology (GO) [52] and other process ontologies, metabolic and signaling maps (Figure 2; Table 1). The networks can be scored and prioritized based on statistical 'relevance' to the functional processes and maps or relative saturation with the uploaded data [53]. The latter can be defined as a proportion of the nodes with the data (for example the overexpressed genes) to the total number of nodes on the networks and measured with z-score. Experimental adjustment can be done by choosing tissue, disease and experiment specific interactions, removing and adding specific interactions mechanisms, linking orthologous genes from other species, etc. A network can also be connected to outside databases and HT data analyzing software. The outcome of such systemic analysis can be new hypotheses on the critical bottlenecks in the disease pathways (potential drug targets) or conservative interactions modules supported by HT data (possible biomarkers) (Figure 2).

Therefore, networks represent a flexible and powerful analytical tool for comparison and cross-validation of

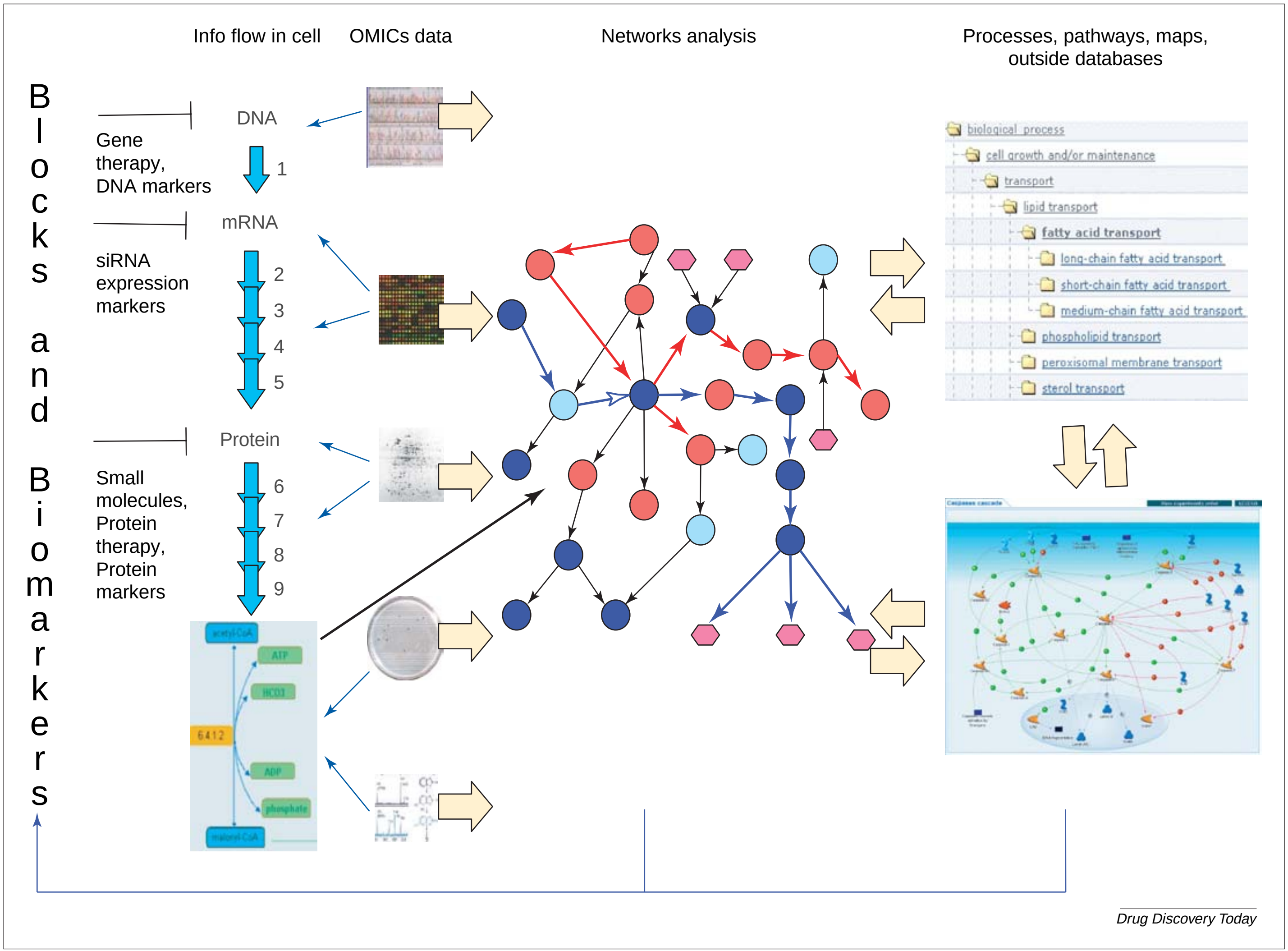


FIGURE 2

General schema of network analysis of 'omics' data. Different types of high-throughput experimental data can be linked to the tables of human protein interactions and enzymatic reactions and visualized as dynamic networks. The networks are then scored and prioritized based on relative relevance of the nodes to functional processes (Gene Ontology, or GO), or static maps of canonical pathways. The software tools can be used for elucidation of network modules and novel pathways as the novel hypotheses for therapeutic targets and biomarkers data.

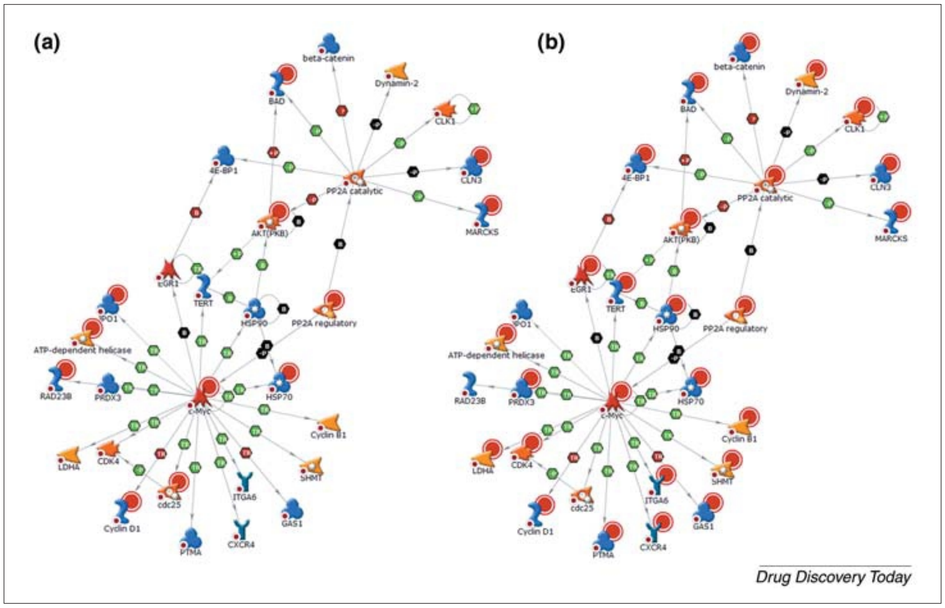


FIGURE 3

The patterns of gene expression in mammary gland epithelium from a non-invasive (a) and invasive (b) breast cancer mapped onto the same network. Gene expression was measured by the SAGE method The network was built using the 'shortest path' algorithm in a standard version of MetaCore (GeneGo, Inc.). The objects marked with large red circles indicate gene overexpression.

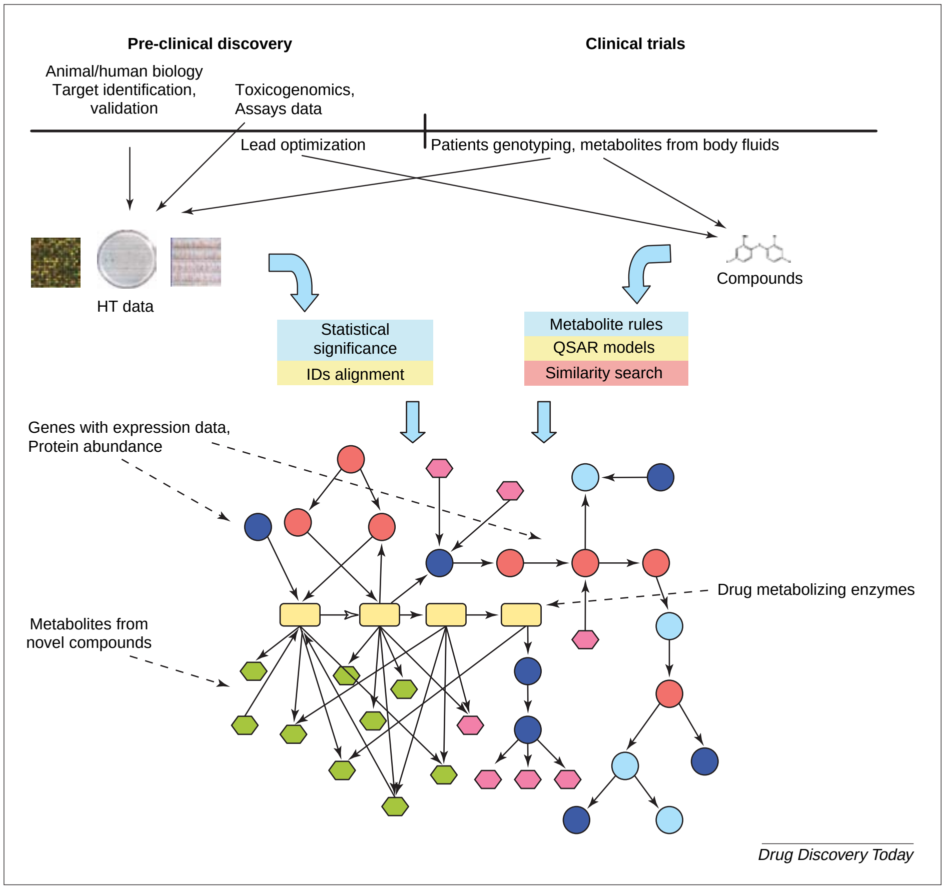

Pre-clinical discovery
Clinical trials
Animal/human biology
Target identification, validation
Toxicogenomics, Assays data
Lead optimization
Patients genotyping, metabolites from body fluids
HT data
Compounds
Statistical significance
IDs alignment
Metabolite rules
QSAR models
Similarity search
Genes with expression data, Protein abundance
Metabolites from novel compounds
Drug metabolizing enzymes
Drug Discovery Today

FIGURE 4

Network analysis in drug discovery pipeline. Biological high-throughput data from different stages of pre-clinical discovery and clinical trials can be uploaded, used for building networks, and analyzed. Human metabolites for novel small molecule compounds can be predicted *in silico* or uploaded directly. All data types can be visualized on the *same* networks.

different types of datasets associated with a condition, such as a disease or a drug treatment). In fact, any experimental or literature-derived datasets with recognizable gene or protein identities (such as LocusLink, Unigene, SwissProt, RefSeq, OMIM) can be visualized, mapped and compared against each other on the same network. For example, one can directly compare the list of genes determined from genetic analyses with data from gene expression arrays from a patient in clinical trials or a knockout mouse. When the same data type and experimental platform is used, the conditional networks can be compared in great detail for common and different sub-networks and patterns. Such fine mapping can be performed to compare the tissue and cell type specific response, different time points, drug dosage; different patients from the same cohort, etc. For instance, we have compared SAGE gene expression patterns from mammary gland duct epithelium of two breast cancer patients, one from pre-invasive DSIC stage, another with invasive cancer (original data from [54]). Both datasets were used for building the initial networks, and then visualized separately. One of the top-scoring networks included the major cell proliferation activator oncogene c-Myc [55] (Figure 3). One can see that the expression pattern for invasive cancer (B) features many more upregulated genes in the immediate vicinity of c-Myc. Modern integrated network analytical suites are well equipped with a range of tools and algorithms for such analyses (Table 1).

Applications of network analysis in drug discovery

Networks analysis is broadly applicable throughout the drug discovery and development pipeline, both on the biology and the chemistry side. Any type of data which can be linked to a gene, a protein or a compound, can be recognized by the input parsers, visualized and analyzed on the networks. Therefore, almost any pre-clinical HT experiment, as well as patient DNA or metabolic tests from clinical trials (Figure 4) can be included in network analyses. Most importantly, all these different datasets can be processed on the same network backbone [56]. Therefore, networks represent the universal platform for data integration

BOX 1

Applications of network analysis in drug discovery

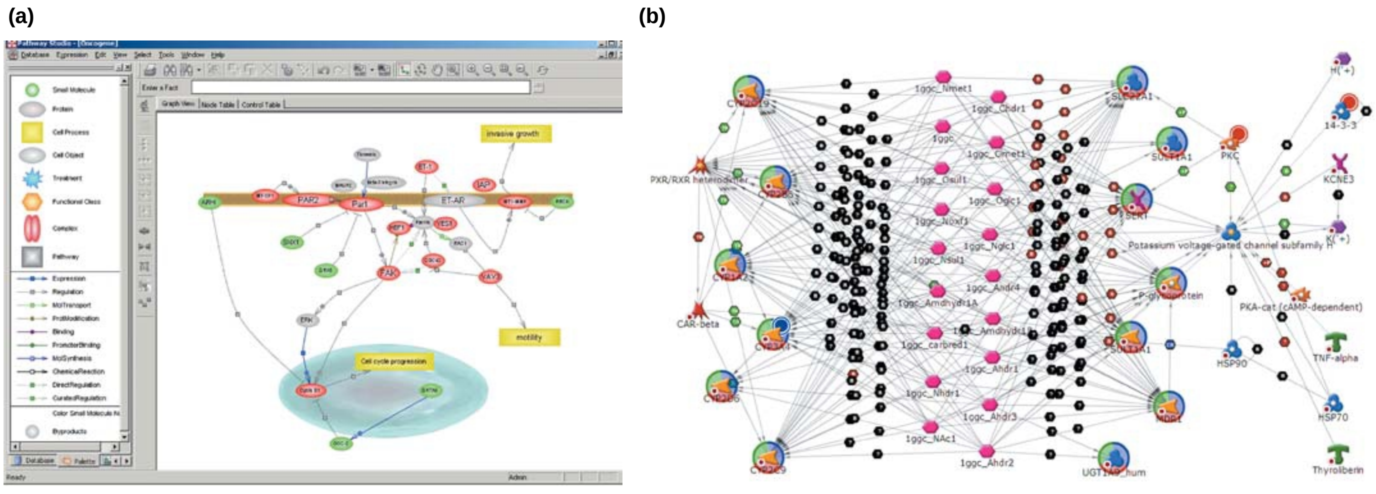
- **Target identification.** Experimental data from model organisms, cell lines and human tissues can be uploaded and mapped on networks. New hypotheses can be made on the pathways connecting the proteins of interest
- **Target validation and prioritization.** Data cross-referencing on the same networks, maps and pathways
- *Disease biomarkers*. The biomarkers can be identified as 'signature networks' – condition-specific conserved sets of nodes supported by differential gene expression and protein abundance data
- **Toxicity biomarkers.** Same as above, with signature networks derived from toxicogenomics data – typically rat or mouse liver arrays from drug-treated animals
- **Pharmacogenomics/haplotyping.** The networks modules can be used as a mean for haplotyping SNPs associated with a particular condition
- **Lead optimization and selection of drug candidates.** New compounds and their metabolites from pre-clinical studies can be mapped on tissue and disease specific metabolic and regulatory networks via structure similarity searches with metabolites and ligands included in the database.
- **Clinical studies.** The patients data (specific DNA sequences, expression microarrays, metabolites from body fluids) can be mapped onto networks and compared with pre-clinical data and published experiments
- **New indications for marketed drugs.** Secondary indications are an important part of follow-up development for bioactive compounds. New therapeutic areas can be suggested by analysis of tissue-specific, disease-specific networks from animals and humans treated with the drug
- **Post-market monitoring.** The patients' data (usually metabolites from body fluids) can be stored in a database and monitored using networks built during clinical and pre-clinical studies

and analysis, which has always been the major objective of bioinformatics technology. Network analysis of complex human diseases is a very young area. In one recent study, the networks automatically generated from literature interactions were applied to the elucidation of specific modules around the genes involved in Alzheimer disease, and the scoring procedure for disease-relevant protein nodes was developed [57]. Some applications of network analysis in drug discovery are listed in Box 1.

Now, we will consider three straightforward applications of network analysis in more detail. In the first case, PathwayAssist (Ariadne Genomics Inc.) was used for generating hypotheses for novel therapeutic targets for ovarian cancer, the most lethal type of gynecologic cancer in the Western world. The list of 1191 differentially expressed genes including ones that are involved in cell growth, differentiation, adhesion, apoptosis and migration were identified by profiling 37 advanced stage papillary serous primary carcinomas [58]. Microarray data were imported into PathwayAssist and a signaling pathway associated with tumor cell migration, spread and invasion was identified as being activated in advanced ovarian cancer (Figure 5a). New pathway hypotheses generated in this study may be useful for elucidation of novel multi-component disease markers and therapeutic targets (Ariadne Genomics, privileged communication) In the second case, network analysis in MetaCore was used for generation of small 'signature networks' characteristic for microarray gene expression response in breast cancer cell line MCF-7 in response to treatment with estrogen and tamoxifen [59]. These small modules consisted of two major hubs – transcriptional factors uniquely linked to a dozen of cell cycle genes. Signature networks were shown to be conserved between the studies on different microarray platforms and, therefore, offered as condition-specific multi-component biomarkers [59]. In the third example (Figure 5b), we used networks for evaluation of toxicity and human metabolism of acetaminophen (APAP). The structure was processed in MetaDrug [60] using metabolic cleavage rules and models, and the resulted metabolites were displayed on the networks connected with the metabolizing enzymes. On the same network, we displayed microarray gene expression data from livers of the rats intoxicated with high dose of APAP [61]. The resulted networks can be used as a tool for elucidation of the effected signaling and metabolic pathways.

Future development

Analysis of biological networks is a young field and most of its applications are still in their infancy although we major developments in this area within the next several years. Qualitative network analysis will be enriched with quantitative modeling and simulation methods. Semi-quantitative techniques such as flux analysis [62] and extreme pathway analysis [63] are already being applied to the reconstruction of signaling and metabolic networks. There are also large computational frameworks such as 'Virtual Cell' [64] that attempt to run simulations on the simplified 'whole cell' level. In addition, network analysis will be substantially scaled up to accommodate the large sets of disease-related molecular data, such as gene expression profiling of hundreds and thousands of patients and the combination of different types of data within network analysis will become routine. At present there are few research groups that combine different types of HT techniques to study particular conditions. These methods query different levels of cell organization (e.g. gene expression, protein abundance, metabolite concentration) and networks will provide a framework for the efficient concurrent analysis of these data. Finally, depositories of higher-level knowledge for particular human diseases will be created. As large molecular datasets will be processed with the help of network analysis, a growing set of reconstructed pathways and networks will emerge. These will be highly specific 'signature networks' for particular conditions, for example, a disease subtype, treatment response, toxin action, and so on.



Drug Discovery Today

FIGURE 5

Applications of network analysis in drug development. (a) Schematic representation of potential signaling pathways from the proteins involved in ovarian cancer as described in [58] and recreated in the latest version of PathwayAssist (Ariadne Genomics, Inc.). Red represents the genes that are upregulated in cancer compared with normal ovarian epithelium, green indicates the genes that are downregulated in cancer specimens, and gray shaded symbols represent genes that did not show a significant difference between cancer and normal specimens. Expression values are taken from [58]. **(b)** Prediction of potential human metabolism and toxicity for novel compounds. APAP metabolites (pink hexagons) and their interactions with enzymes and signaling proteins as predicted and visualized by MetaDrug™ (GeneGo, Inc.) [60]. The metabolites will interact with (be processed by) several Cytochrome P450 enzymes and transferases. Metabolites also interact with PXR/RXR – transcriptional regulators of cytochromes. Microarray expression data from APAP-treated [61] rat livers is superimposed on the network. The results indicate downregulation of CYP3A4 (marked with blue circle, left) and upregulation of some genes linked to HERG channel (red circles, right). Observed downregulation of CYP3A4 provides independent corroboration of interactions between APAP metabolites and PXR/RXR [60].

References

1 Nicholson, J.K. and Wilson, I.D. (2003) Understanding ‘global, systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* 2, 668–676

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

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

4 Hartwell, L.H. *et al.* (1999) From molecular to modular cell biology. *Nature* 402(Suppl.), C47–C50

5 Murray, A.W. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826

6 Hood, L. and Perlmutter, R.M. (2004) The impact of systems approaches on biological problems in drug discovery. *Nat. Biotechnol.* 2, 1218–1219

7 Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling. *Genome Biology,* 3, RESEARCH0055

8 Chen, H. and Sharp, B.M. (2004) Content-rich biological network constructed by mining. *BMC Bioinformatics* 5, 147

9 Chaussabel, D. (2004) Biomedical literature mining: challenges and solutions in the ‘omics’ era. *Am. J. Pharmacogenomics* 4, 383–393

10 Blaschke, K. and Valencia, A. (2001) Can bibliographical pointers for known biological data be found automatically? Protein interactions as a case study. *Comp. Func. Genom* 2, 196–2006

11 Santos, C. *et al.* Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics* (in press)

12 Daraselia, N. *et al.* (2004) Extracting protein function information from MEDLINE using a full-sentence parser. In *Proc. 2nd European Workshop on Data Mining and Text Mining for Bioinformatics*, pp. 11–18, ECML/PKDD 2004 Committee (available online at www.informatik.hu-berlin.de/Forschung_Lehre/wm/ws04/#proceedings)

13 Chen, T.L. *et al.* (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. U. S. A.* 88, 9578–9582

14 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574

15 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster. Science* 302, 1727–1736

16 Li, S. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans. Science* 303, 540–543

17 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, R7

18 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

19 Lehner, B. and Frazer, A.G. (2004) A first-draft human protein-interactions map. *Genome Biol.* 5, R63

20 Peri, S. *et al.* (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32 Database issue, D497-501

21 Pagel, P. *et al.* The MIPS mammalian protein–protein interaction database. *Bioinformatics* (in press)

22 Highes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. Cell 109-126

23 van Noort, V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.* 19, 238–242

24 Kemmeren, P. *et al.* (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* 9, 1133–1143

25 Deane, C.M. *et al.* (2002) Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell.* Proteomics 1, 349–356

26 Ho, Y. *et al.* (2002) identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415, 180–183

27 Gavin, A-C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147

28 Navarro, J.D. and Pandey, A. (2004) Unraveling the human interactome: lessons from the yeast. *Drug Discov. Today* 3, 79–84

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

30 Von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403

31 Bader, J.S. *et al.* (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22, 78–85

32 Yu, H. *et al.* (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res*. 32, 328–337

33 Barabasi, A.L. and Oltavi, Z.N. (2004) Network Biology: understanding the cell's functional organization. *Nat. Rev. Genet*. 5, 101–113

34 Albert, R. *et al.* (2000) Error and attack tolerance of complex networks. *Nature* 406, 378–382

35 Ihmels, J. *et al.* (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat. Biotechnol*. 22, 86–92

36 Li, S. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science* 303, 540–543

37 Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A*. 100, 12123–12128

38 Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U. S. A*. 100, 1128–1133

39 Pereira-Leal, J.B. *et al.* (2004) Detection of functional modules from protein interaction networks. *Proteins* 54, 49–57

40 Milo, R. *et al.* (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 824–827

41 Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet*. 31, 64–68

42 Wuchty, S. *et al.* (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet*. 35, 176–179

43 Tornow, S. and Mewes, H.W. (2003) Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*. 31, 6283–6289

44 Hanisch, D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics* 18(Suppl. 1), S145–S154

45 Ideker, T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1), S233–S240

46 Segal, E. *et al.* (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* 19(Suppl. 1), I264–I272

47 Ogata, H. *et al.* (2000) Heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*. 28, 4021–4028

48 Nakaya, A. *et al.* (2001) Extraction of correlated gene clusters by multiple graph comparison. (Genome Inform. Ser. Workshop). *Genome Inform*. 12, 44–53

49 Kelley, B.P. *et al.* (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. U. S. A*. 100, 11394–11399

50 Frazer, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752

51 Yu, H. *et al.* (2004) Genomic analysis of essentiality within protein networks. *Trends Genet*. 20, 227–231

52 The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*. 25, 25–29

53 Nikolsky, Y. *et al.* (2005) Systems level network analysis of OMICs data. *Genet. Eng. News* 25, 28–29

54 Allinen, M. *et al.* (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6, 17–32

55 Nikolsky, Y. (2004) System-level analysis of SAGE and other high-throughput data in the context of functional networks. In *Proceedings SAGE 2004 Conference*, p. 36, SAGE

56 Ekins, S. *et al*. Systems biology: applications in drug discovery. In Drug Discovery Handbook. (Gad, S. ed.). *Wiley* (in press)

57 Kraufthammer, M. *et al.* (2004) Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A*. 101, 15148–15153

58 Donninger, H. *et al.* (2004) Whole genome expression profiling of advance stage papillary serous ovarian cancer reveals activated pathways. *Oncogene* 23, 8065–8077

59 Nikolsky, Y. *et al.* A novel method for generation of signature networks as biomarkers from complex high-throughput data. *Tox. Letters* (in press)

60 Ekins, S. *et al.* A novel method for visualizing nuclear hormone receptor networks relevant to drug metabolism. *Drug Metab. Dispos*. (in press)

61 Huang, Q. *et al.* (2004) Gene expression profiling reveals multiple toxicity endpoints induced by hepatotoxicants. *Mutat. Res*. 549, 147–168

62 Vo, T.D. *et al.* (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem*. 279, 39532–39540

63 Papin, J.A. and Palsson, B.O. (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J*. 87, 37–46

64 Slepchenko, B.M. *et al.* (2003) Quantitative cell biology with the Virtual Cell. *Trends Cell Biol*. 13, 570–576

65 Bader, G.D. *et al.* (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*. 31, 248–250

66 Salwinski, L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res*. 32, D449–451

67 Zanzoni, A. *et al.* (2002) MINT: a molecular interaction database. *FEBS Lett*. 513, 135–140

68 Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*. 30, 31–34

69 von Mering, C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. 31, 258–261

70 Kanehisa, M.A. *et al.* (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res*. 30, 42–46

Related articles in other Elsevier journals

Global properties of biological networks
Grigorov, M.G. (2005) *Drug Discov. Today* 10, 365–372

Comparison of network-based pathway analysis methods
Papin, J.A. *et al.* (2004) *Trends Biotechnol*. 22, 400–405

Reconstruction of microbial transcriptional regulatory networks
Herrgard, M.J. *et al.* (2004) *Curr. Opin. Biotechnol*. 15, 70–77

Building with a scaffold: emerging strategies for high- to low-level cellular modeling
Ideker, T. and Lauffenburger, D. (2003) *Trends Biotechnol*. 21, 255–262