BRIEFINGS IN BIOINFORMATICS. VOL 7. NO 3. 243–255 doi:10.1093/bib/bbl022

Graph-based methods for analysing networks in cell biology

Tero Aittokallio and Benno Schwikowski

Submitted: 4th May 2006; Accepted: 27th June 2006

Abstract

Availability of large-scale experimental data for cell biology is enabling computational methods to systematically model the behaviour of cellular networks. This review surveys the recent advances in the field of graph-driven methods for analysing complex cellular networks. The methods are outlined on three levels of increasing complexity, ranging from methods that can characterize global or local structural properties of networks to methods that can detect groups of interconnected nodes, called motifs or clusters, potentially involved in common elementary biological functions. We also briefly summarize recent approaches to data integration and network inference through graph-based formalisms. Finally, we highlight some challenges in the field and offer our personal view of the key future trends and developments in graph-based analysis of large-scale datasets.

Keywords: *graph algorithms; data integration; cellular networks; protein–protein interactions; transcriptional regulatory networks; network modularity*

INTRODUCTION

Recent advances in large-scale experimental technologies have resulted in an accumulation of experimental data that reflect the interplay between biomolecules on a global scale. Due to the complexity of the control mechanisms involved, and the large number of possible interactions, there is a great need for computer-assisted tools to manage, query and interpret the experimental observations with formal network models. In their most basic abstraction level, cellular networks can be represented as mathematical graphs, using nodes to represent cellular components, and edges to represent their various types of interactions [1]. For instance, protein–protein interaction (PPI) networks are conveniently modelled by undirected graphs, where the nodes are proteins and two nodes are connected by an undirected edge if the corresponding proteins physically bind. In contrast, transcriptional regulatory networks can be modelled as directed weighted graphs, where the weights of directed edges capture the degree of the regulatory effect of the transcription factors (source nodes) to their regulated genes (sink nodes). Metabolic networks generally require more complex representations, such as hypergraphs, as reactions in metabolic networks generally convert multiple reaction inputs into multiple outputs with the help of other components. An alternative, reduced representation for a metabolic network, is a weighted bipartite graph, where two types of nodes are used to represent reactions and compounds, respectively, and the edges connect nodes of different types, representing either substrate or product relationships.

The representation of complex cellular networks as graphs has made it possible to systematically investigate the topology and function of these networks using well-understood graph-theoretical concepts that can be used to predict the structural and dynamical properties of the underlying network. Such predictions can suggest new biological hypotheses regarding, for instance, unexplored new interactions of the global network or the function of individual cellular components that are testable

Corresponding author. Tero Aittokallio, Systems Biology Group, Institut Pasteur, 25-28 Rue du Dr Roux, FR-75724 Paris, France. Tel: +33 1 4061 3784; Fax: +33 1 4061 3704; E-mail: teanai@pasteur.fr

Tero Aittokallio is a postdoctoral researcher in the Systems Biology Group at Institut Pasteur. His research interests include mathematical modelling and machine learning methods, especially with applications to data mining and data integration problems. **Benno Schwikowski** leads the Systems Biology Group at Institut Pasteur in Paris. He is interested in developing measurement technology and experimental–computational approaches to problems in cell biology and infectious disease.

Table I: Graph-based approaches to cellular network analysis covered in this article

Section header	Network topology	Interaction patterns	Network decomposition
Description level	Global structural properties	Local structural connectivity	Hierarchical functional organization
Basic concepts	Scale-free topology	Subgraphs	Modules
	Degree distribution	Centrality	Motifs
	Clustering coefficient	Pathways	Clusters
Specific aim	Characterization of large-scale attributes of cellular networks	Analysis of elementary interaction patterns of cellular mechanisms	Classification into groups of functionally related biomolecules

The methods are presented on three levels of increasing complexity, along with their basic concepts and specific aims in the cellular network analysis. Global structural attributes try to characterize the behaviour of the cell as a whole, whereas local network analyses aim at discovering such individual interaction patterns that may carry significant information about their roles in cellular mechanisms. Dissection of hierarchical organization of the cell through motif searches or network clustering seeks to partition the complex networks into functionally organized hierarchy of interconnected groups that are involved in common cellular functions.

with subsequent experimentation. Even a simplistic dynamical system originating from small Boolean network models, where nodes represent discrete biological entities (i.e. mRNA or protein), that can be thought to be either on or off and edges their Boolean relationships ('genotypes'), can give rise to a multitude of designable dynamical outputs ('phenotypes') [2]. Mathematical modelling also enables an iterative process of network reconstruction, where model simulations and predictions are closely coupled with new experiments chosen systematically to maximize their information content for subsequent model adjustments, providing increasingly more accurate descriptions of the network properties [3]. The topological relations underlying graph-based methods can also convey structure to putative pathways. This helps avoiding approaches that test many known sets of molecules without causal interactions [4]. Furthermore, graph formalisms may provide powerful tools for 'omics' data integration to address fundamental biological questions at the systems level [5*].

This review describes network analysis approaches in which the concept of a graph is a key component, together with a large collection of recently introduced methods and available tools. This excludes some related problems, such as hierarchical clustering or phylogenetic footprinting, where typically only the result of the computational analysis is represented in the form of a specific graph, such as a dendrogram [6] or a phylogenetic tree [7]. As substantial efforts have recently been devoted to develop graph-based methods for a wide range of computational and biological tasks, only representative examples of different approaches can be surveyed here, with an emphasis on methods related to concrete biological questions, rather than computational issues. The selected methods are presented in the broader context of network analysis, summarizing some of the basic concepts and themes such as scale-free networks, pathways and modules (Table 1). The order of sections roughly reflects the increasing demands placed for the type and amount of data the methods require and their applicability to address more targeted problems in cell biology. Accordingly, the methods reviewed range from very elementary measures that characterize the global topological structure and require only general assumptions about the underlying network model to recent software systems available for integrating multiple types of cellular data within a graph-based framework that can be applied to solve concrete biological problems.

CHARACTERIZATION OF NETWORK TOPOLOGY

Perhaps the most general level of network analysis comes from global network measures that allow us to characterize and compare the given network topologies (i.e. the configuration of the nodes and their connecting edges). Global measures such as the degree distribution (the degree of a node is the number of edges it participates in) and the clustering coefficient (the number of edges connecting the neighbours of the node divided by the maximum number of such edges) have recently been thoroughly reviewed in the context of cellular networks [8**] and in proteomics [9]. It has been proposed that these quantitative graph concepts can efficiently capture the cellular network organization, providing insights into their evolution, function, stability and dynamic responses [10**]. For instance, several types of surveyed biological networks, such as PPI, gene regulation and metabolic networks, are thought

to display scale-free topologies (i.e. most nodes have only a few connections whereas some nodes are highly connected), characterized by a power-law degree distribution that decays slower than exponential. This particular type of network topology is also frequently observed in numerous non-biological networks and it can be generated by simple and elegant evolutionary models, where new nodes attach preferentially to sites that are already highly connected. Numerous improvements to this generic model include, for instance, iterative network duplication and integration to its original core, leading to hierarchical network topologies, which are characterized by non-constant clustering coefficient distribution [8, 10].

It should be noticed, however, that, in practice, the architecture of large-scale biological networks is determined with sampling methods, resulting in subnets of the true network, and only these partial networks can be applied to characterize the topology of the underlying, hidden network [11]. It has recently been recognized that it is possible to extrapolate from subnets to the properties of the whole network only if the degree distributions of the whole network and randomly sampled subnets share the same family of probability distributions [12]. While this is the case in specific classes of network graph models, including classical Erdös–Rényi and exponential random graphs, the condition is not satisfied for scale-free degree distributions. Accordingly, recent studies in interactome networks have revealed that the commonly accepted scale-free model for PPI networks may fail to fit the data [13]. Moreover, limited sampling alone may as well give rise to apparent scale-free topologies, irrespective of the original network topology [14]. These results suggest that interpretation of the global properties of the complete network structure based on the current—still limited—accuracy and coverage of the observed networks should be made with caution. Moreover, while the scale-free and hierarchical graph properties can efficiently characterize some large-scale attributes of networks, the local modularity and network clustering is likely to be the key concept in understanding most cellular mechanisms and functions.

GRAPH ANALYSIS OF INTERACTION PATTERNS

As an alternative to the study of global graph characteristics, elementary graph algorithms have been used to characterize local interconnectivity and more detailed relationships between nodes. Such graph methods can facilitate addressing fundamental biological concepts, such as essentiality and pathways, especially when additional biological information is incorporated into the analysis in addition to the primary data. For instance, while gene expression clustering traditionally makes the assumption that genes with similar expression profiles have similar functions in cells, a more targeted approach could aim at identifying the genes participating in a particular cellular pathway where not every components has a similar transcriptional profile [15*]. Once the network of interest has been represented as a graph, the conventional graph-driven analysis work-flow involves the following two steps: (i) applying suitable graph algorithms to compute the local graph properties, such as the number and complexity of given subgraphs, the shortest path length of indirectly connected nodes or the presence of central nodes of the network and (ii) evaluating the sensitivity and specificity of the model predictions using curated databases of known positive examples or random models of synthetic negative examples, respectively. We start by surveying the basic graph concepts used in network analysis together with corresponding recent work.

Subgraphs and centrality statistics

A subgraph represents a subset of nodes with a specific set of edges connecting them. As the number of distinct subgraphs grows exponentially with the number of nodes, efficient and scalable heuristics have been developed and applied for detecting the given subgraphs and their frequencies in large networks. In contrast to network motif searches, Przulj *et al.* [13] argued that it is equally important to understand the organization of infrequently observed subgraphs as the frequently observed ones. Graphlets are defined as small induced subgraphs, consisting of all edges of the original graph that connect a given group of nodes, regardless of whether or not they appear at significantly higher frequencies than expected in randomized networks. Since exhaustive searches become computationally infeasible even when applied to rather small networks, Przulj *et al.* [16] designed sampling heuristics for finding graphlets in high-confidence PPI networks that concentrate on specific parts of the graph, depending on the particular model (either geometric random graph model or more general sampling strategy).

Under the random graph model, it is also possible to calculate analytically the estimated distribution of different subgraphs with given number of nodes, edges and their specific global properties like degree distribution and clustering coefficient [17]. Such approximate analytical expressions will save substantial amounts of computing time when analysing e.g. lists of proteins with large undirected graphs representing their known functional relationships [18].

Centrality is a local quantitative measure of the position of a node relative to the other nodes, and can be used to estimate its relative importance or role in global network organization. Different flavors of centrality are based on the node's connectivity (degree centrality), its shortest paths to other nodes (closeness centrality) or the number of shortest paths going through the node (betweenness centrality). Estrada [19] recently showed that centrality measures based on graph spectral properties can distinguish essential proteins in PPI network of yeast *Saccharomyces cerevisiae* (essential genes are those upon which the cell depends for viability). In particular, the best performance in identifying essential proteins was obtained with a novel measure introduced to account for the participation of a given node in all subgraphs of the network (subgraph centrality), which gives more weight to smaller subgraphs. It was proposed that ranking proteins according to their centrality measures could offer a means to selecting possible targets for drug discovery [19]. A similar approach to characterize the importance of individual nodes, based on trees of shortest paths and concepts of 'bottleneck' nodes, demonstrated that 70% of the top 10 most frequent 'bottleneck' proteins were inviable and structural proteins that do not participate in cellular signaling [20⋆]. With degree centrality analyses in the metabolic networks of *Escherichia coli, S. cerevisiae* and *Staphylococcus aureus*, it was also demonstrated that most reactions identified as essential turned out to be those involving the production or consumption of low-degree metabolites [21].

Paths and pathways

In the theory of directed graphs, a path is a chain of distinct nodes, connected by directed edges, without branches or cycles. Such pathways in cellular network graphs can represent, for instance, a transformation path from a nutrient to an end product in a metabolic network, or a chain of post-translational modifications from the sensing of a signal to its intended target in a signal transduction network [10⋆⋆]. Pathway redundancy (the presence of multiple paths between the same pair of nodes) is an important local property that is thought to be one of the reasons for the robustness of many cellular networks. Betweenness centrality can be used to measure the effect of node perturbations on pathway redundancy, whereas path lengths characterize the response times under perturbations. With shortest paths and centrality-based predictions in the *S. cerevisiae* PPI and metabolic networks, respectively, the existence of alternate paths that bypass viable proteins can be demonstrated, whereas lethality corresponds to the lack of alternative pathways in the perturbed network [20⋆, 22]. Besides the various commercial software packages for pathway analysis there exist also freely available tools for some specific graph queries, such as finding shortest paths between two specified seed nodes on degree-weighted metabolic networks [23] or searching for linear paths that are similar to query pathways in terms of their composition and interaction patterns on a given PPI network [24].

The relatively high degree of noise inherent in the interactions data in current PPI databases can make pathway modelling very challenging. Integration of prior biological knowledge, such as Gene Ontology (GO), can be used to make the process of inferring models more robust by providing complementary information on protein function. GO terms and their relationships are encoded in the form of directed acyclic graph (DAG). Guo *et al.* [25] recently assessed the capability of both GO graph structure-based and information content-based similarity measures on DAG to evaluate the PPIs involved in human regulatory pathways. They also showed how the functional similarity of proteins within known pathways decays rapidly as their path length increases. While most of the analysis methods designed for PPI networks consider unweighted graphs, where each pairwise interaction is considered equally important, Scott *et al.* [26] recently presented linear-time algorithms for finding paths and more general graph structures such as trees that can also consider different reliability scores for PPIs. By exploiting a powerful randomized graph algorithm, called color coding, they efficiently recovered several known *S. cerevisiae* signaling pathways such as MAPK, and showed that in general the pathways

they detected score higher than those found in randomized networks. In addition to known pathways, they also predicted (by unsupervised learning) novel putative pathways in the PPI network that are functionally enriched (i.e. share significant number of common GO annotations) [26].

NETWORK DECOMPOSITION INTO FUNCTIONAL MODULES

The decomposition of large networks into distinct components, or modules, has come to be regarded as a major approach to deal with the complexity of large cellular networks [27–29]. This topic has witnessed great progress lately, and only representative examples of different approaches are presented here. In cellular networks, a module refers to a group of physically or functionally connected biomolecules (nodes in graphs) that work together to achieve the desired cellular function [8**]. To investigate the modularity of interaction networks, tools and measures have been developed that can not only identify whether a given network is modular or not, but also detect the modules and their relationships in the network. By subsequently contrasting the found interaction patterns with other large-scale functional genomics data, it is possible to generate concrete hypotheses for the underlying mechanisms governing e.g. the signaling and regulatory pathways in a systematic and integrative fashion. For instance, interaction data together with mRNA expression data can be used to identify active subgraphs, that is, connected regions of the network that show significant changes in expression over particular subsets of experimental conditions [30].

Motifs

Motifs are subgraphs of complex networks that occur significantly more frequently in the given network than expected by chance alone [29]. Consequently, the basic steps of motif analyses are (i) estimating the frequencies of each subgraph in the observed network, (ii) grouping them into subgraph classes consisting of isomorphic subgraphs (topologically equivalent motifs) and (iii) determining which subgraph classes are displayed at much higher frequencies than in their random counterparts (under a specified random graph model). While analytical calculations from random models can assist in the last step, exhaustive enumeration of all subgraphs with a given number of nodes in the observed network is impossible in practice. Kashtan *et al*. [31] therefore developed a probabilistic algorithm that allows estimation of subgraph densities, and thereby detection of network motifs, at a time complexity that is asymptotically independent of the network size. The algorithm is based on a subgraph importance sampling strategy, instead of standard Monte Carlo sampling. They noticed that, network motifs could be detected already with a small number of samples in a wide variety of biological networks, such as the transcriptional regulatory network of *E. coli* [31]. Recently, efficient alternatives together with graphical user interfaces have also been implemented to facilitate fast network motif detection and visualization in large network graphs [32, 33].

Many of the methodologies recently introduced in network analysis are inspired by established approaches from sequence analysis. The concepts utilized in both fields include approximate similarity, motifs and alignments. As network motifs represent a higher-order biological structure than protein sequences, graph-based methods can be used to improve the homology detection of standard sequence-based algorithms, such as PSI-BLAST, by exploiting relationships between proteins and their sequence motif-based features in a bipartite graph representing protein-motif network [34]. The definition of network motifs can be enriched by concepts from probability theory. The motivation is that if the network evolution involves elements of randomness and the currently available interaction data is imperfect, then functionally related subgraphs do not need to be exactly identical. Accordingly, Berg and Lässig [35] devised a local graph alignment algorithm, which is conceptually similar to sequence alignment methodologies. The algorithm is based on a scoring function measuring the statistical significance for families of mutually similar, but not necessarily identical, subgraphs. They applied the algorithm to the gene regulatory network of *E. coli* [35].

Motifs have increasingly been found in a number of complex biological and non-biological networks, and the observed over-representation have been interpreted as manifestations of functional constraints and design principles that have shaped network architecture at the local level. Significance of motifs is typically assessed statistically by comparing the distribution of subgraphs in an observed network with that found in a particular computer-generated

sample of randomized networks that destroy the structure of the network while preserving the number of nodes, edges and their degree distribution. It can be argued what kind of random model provides the most appropriate randomization, and especially whether it is realistic to assume that the edges in the randomized network are connected between the nodes globally at random and without any preference [36]. However, the principal application of network motif discovery should not originate from a rigorous statistical testing of a suitable null hypothesis, but from the possibility to reduce the complexity of large networks to smaller number of more homogeneous components. Analogously with gene expression cluster analysis, where statistical testing is also difficult because of the lack of an established null model, network decomposition may be used as a tool to identify biologically significant modules, irrespective of their statistical significance.

Clusters

An alternative approach to the identification of functional modules in complex networks is discovering similarly or densely connected subgraphs of nodes (clusters), which are potentially involved in common cellular functions or protein complexes [37]. As in expression clustering, the application of graph clustering is based on the assumption that a group of functionally related nodes are likely to highly interact with each other while being more separate from the rest of the network. The challenges of clustering network graphs are similar to those in the cluster analysis of gene expression data [6]. In particular, the results of most methods are highly sensitive to their parameters and to data quality, and the predicted clusters can vary from one method to another, especially when the boundaries and connections between the modules are not clear-cut. This seems to be the case at least in the PPI network of *S. cerevisiae* [38]. Moreover, it should be noted that modules are generally not isolated components of the networks, but they share nodes, links and even functions with other modules as well [8★★]. Such hierarchical organization of modules into smaller, perhaps overlapping and functionally more coherent modules should be considered when designing network clustering algorithms. The functional homogeneity of the nodes in a cluster with known annotations can be assessed against the cumulative hypergeometric distribution that represents the null model of random function label assignments [20★].

Highly connected clusters

Most algorithms for determining highly connected clusters in PPI networks yield disjoint modules [39]. For instance, King *et al.* [40] partitioned the nodes of a given graph into distinct clusters, depending on their neighbouring interactions, with a cost-based local search algorithm that resembles the tabu-search heuristic (i.e. it updates a list of already explored clusters that are forbidden in later iteration steps). Clusters with either low functional homogeneity, cluster size or edge density were filtered out. After optimizing the filtering cut-off values according to the cluster properties of known *S. cerevisiae* protein complexes from MIPS database, their methods could accurately detect the known and predict new protein complexes [40]. Other local properties such as centrality measures can be used for clustering purposes as well. A recent algorithm by Dunn *et al.* [41], for example, divides the network into clusters by removing the edges with the highest betweenness centralities, then recalculating the betweenness and repeating until a fixed number of edges have been removed. They applied the clustering method to a set of human and *S. cerevisiae* PPIs, and found out that the protein clusters with significant enrichment for GO functional annotations included groups of proteins known to cooperate in cell metabolism [41].

Overlapping clusters

Corresponding to the fact that proteins frequently have multiple functions, some clustering approaches, such as the local search strategy by Farutin *et al.* [42], also allow overlapping clusters. Like in motif analysis, the score for an individual cluster in the PPI network graph is assessed against a null model of random graph that preserves the expected node degrees. They also derived analytical expressions that allow for efficient statistical testing [18]. It was observed that many of the clusters on human PPI network are enriched for groups of proteins without clear orthologues in lower organisms, suggesting functionally coherent modules [42]. Pereira-Leal *et al.* [43] used the line graph of the network graph (where nodes represent an interaction between two proteins and edges represent shared interactors between interactions) to produce an overlapping graph partitioning of the original PPI network of *S. cerevisiae*. Recently, Adamcsek *et al.* [44] provided a program for locating and visualizing overlapping, densely interconnected groups of nodes in a given

undirected graph. The program interprets as motifs all the *k*-clique percolation clusters in the network (all nodes that can be reached via chains of adjacent *k*-cliques). Larger values of *k* provide smaller groups resulting in higher edge densities. Edge weights can additionally be used to filter out low-confidence connections in the graphs [44].

Distance-based clusters

Another approach to decompose biological networks into modules applies standard clustering algorithms on vectors of nodes' attributes, such as their shortest path distances to other nodes [45]. As the output then typically consists of groups of similarly linked nodes, the approach can be seen as complementary to the above clustering strategies that aim at detecting highly connected subgraphs. To discover hierarchical relationships between modules of different sizes in PPI graphs, Arnau *et al*. [46] explored the use of hierarchical clustering of proteins in conjunction with the pairwise path distances between the nodes. They considered the problem of lacking resolution caused by the 'small world' property (relatively short—and frequently identical—path length between any two nodes) by defining a new similarity measure on the basis of the stability of node pair assignments among alternative clustering solutions from resampled node sets. As ties in such bootstrapped distances are rare, standard hierarchical clustering algorithms yield clusters with a higher resolution. The clusters obtained in *S. cerevisiae* PPI data were validated using GO annotations and compared with those refined from gene expression microarray data [46]. A similar approach was also applied to decompose metabolic network of *E. coli* into functional modules, based on the global connectivity structure of the corresponding reaction graph [47].

Supervised clustering

Provided that the eventual aim of module analysis is function prediction, it can be argued that supervised clustering (or classification), rather than unsupervised clustering methods, should be employed. In the context of cellular networks, classification aims at constructing a discriminant rule (classifier) that can accurately predict the functional class of an unknown node based on the annotation of neighbouring nodes and connections between them. To this end, Tsuda and Noble [48] considered a binary classification problem, and calculated pairwise distances on undirected graphs with a locally constrained diffusion kernel. They demonstrated a good protein function prediction with a support vector machine (SVM) classifier from *S. cerevisiae* PPI and metabolic networks. Supervised clustering methods in function prediction are challenged by their notorious dependence on the quality of the training examples [49]. As fully curated databases are rarely available, especially for less-studied organisms, the applicability of such methods is still limited. Therefore, an intermediate method between the two extremes of supervised and unsupervised clustering may be preferable. The protein function prediction algorithm by Nabieva *et al*. [50] suggests such an approach that exploits both global and local properties of the network graphs. They demonstrated better predictions than previous methods in cross-validation testing on the unweighted *S. cerevisiae* PPI graph. More importantly, they showed that the performance could be substantially improved further by weighting the edges of the interaction network according to information from multiple data sources and types [50].

CURRENT CHALLENGES AND FUTURE TRENDS

Functional modules across multiple data sources

As the high-throughput assays are inherently noisy and biased in their nature, and each single data source or type can describe only a limited scope of a system, it is evident that integrative analysis of data from such measurements will be essential in order to fully understand the system's behaviour on a global scale [51⋆]. In many biological applications, it is beneficial to perform the network analysis in a truly integrated manner, simultaneously rather than sequentially, like when validating the results against external data sources. Graph-based frameworks can also be used in such an integrative analysis of data from different sources. The composition of data sources required depends naturally on the specific biological goals of the study. In the analysis of transcriptional regulatory networks, for instance, clustering becomes a problem of dissecting genes into regulatory modules (sets of coexpressed genes regulated by common transcription factors). It has been shown that the identification of regulatory modules can be improved by combining gene expression data (inferred e.g. from microarrays) with the knowledge

of transcription factor binding to the DNA motifs (extracted e.g. from chromatin immunoprecipitation ChIPs) [52–54]. Recently, Tanay *et al*. [55] identified modules across diverse genome-wide data sources and types, including gene expression, protein interactions, growth phenotype data and transcription factor binding. By modelling genomic information as properties of a weighted bipartite graph in the yeast system, they defined clusters of genes with a common behaviour across a set of the experiments. This provides a general data processing and integration framework for revealing a detailed view of both the global and local organization of a molecular network even in higher organisms [55].

In interaction networks, the detection of modules, motifs or clusters has also been performed on multiple graphs simultaneously using efficient algorithms for exact or approximate pattern mining across a set of graphs constructed from same data type [56, 57]. Towards integrated graph analysis of heterogeneous genome-wide data sources, Yeger-Lotem *et al*. [58] developed algorithms for detecting composite network motifs with two or more types of interactions and applied them to a combined data set of PPIs and transcription-regulation interactions in *S. cerevisiae*. Similarly, Moon *et al*. [59] built a unified network model for both protein–protein and domain–domain interactions to detect network motifs between proteins and their domains by applying a colored vertex graph model. The module searches can also be extended to incorporate more than one species in order to elucidate the evolution of cellular machinery or to predict more reliably the protein functions. Sharan and Ideker [60**] recently reviewed the computational approaches available to comparative biological network analysis, that is, contrasting two or more interaction networks representing different species, conditions, interaction types or time points. In particular, network integration can assist in predicting protein interactions or uncovering protein modules that are supported by interactions of different types. Besides interaction data sets, Chen and Xu [61] encoded into their functional linkage graph also other genome-scale data types, including microarray gene expression profiles, and used them simultaneously when annotating *S. cerevisiae* proteins into multiple GO categories. Future studies involving a blend of multiple experimental and computational approaches will hopefully provide added insights into the biological roles of network motifs and clusters [62–64].

Software tools for graph-based network analysis

The availability of genome-scale data sets has increased the need for software tools that can integrate, construct, analyze and visualize the high-dimensional data effectively. Several such software packages developed for these challenging tasks along with their specific functionalities were recently listed by Joyce and Palsson [5*]. Publicly available software systems that use graph-based data integrating visual frameworks for networks include e.g. Cytoscape together with its recent plug-ins [65–67], Osprey [68], GiGA [69], megNet [70], VisANT [71], BioPIXIE [72], Pointillist [73, 74], PIANA [75] and PathSys [76]. An important component of such systems is the possibility to visualize the graphs under analysis. This can be regarded as a fundamental tool in explorative network analysis; even if one wants to address only a very specific question within the given network graph, it may be helpful to visualize the result to discern possible flaws or follow-up questions. Recently introduced graph drawing tools include e.g. WebInterViewer [77], CADLIVE [78] and PATIKAweb [79]. By meeting the challenges of automated construction and simultaneous visualization of multiple pathways, such software tools can be of great help in relating the selected node sets and their interconnections to the underlying biological significance.

Bioconductor project incorporates also open source tools to support computational analysis of graphical data structures (http://www.bioconductor.org/). The available packages implement not only algorithms for efficient graph visualization (AT&T Graphviz), but also the C++ Boost Graph Library for basic graph algorithms (RBGL package). At present, procedures that can be interfaced in the R environment include minimum spanning tree construction, shortest path finding, depth-first search, topological sorting, edge-connectivity measurement and connected component decomposition [80]. The number of packages available within Bioconductor grows rapidly as many authors make their R source codes freely available for academic use [81–83]. These advanced graph algorithms and post-processing tools can be used also in conjunction with more specific, freely available software packages such as GenePattern

(http://www.broad.mit.edu/genepattern). One major challenge of computational network analysis deals with the selection of model types appropriate for analysing data from different experimental approaches [84**]. In particular, accurate modelling and integration of protein interactions measured with yeast two-hybrid and affinity purification/mass spectrometry can be very critical e.g. for understanding the physical properties and functional operation of local protein complexes [82].

Network graph reconstruction by reverse engineering

A number of computational approaches have been tried to reconstruct the underlying global network structure or even the causal regulatory relationships between the nodes from the experimental data sets. This challenging problem is often referred to as network inference or reverse engineering [85, 86]. For instance, several works have dealt with gene regulatory networks inference from gene expression microarray data alone. In such a hypothetical network, the nodes conventionally correspond to both the particular gene and the protein it encodes, and the edges to the statistical relations between the genes. Bayesian network offers a convenient probabilistic model, where nodes represent gene expression levels as random variables, edges represent their conditional dependence relations and the corresponding DAG the joint probability distributions of the observed expression patterns. However, it has been recognized that these data are sufficient for reconstruction of only relatively small networks and that even in idealized situations the estimated models contain many false edges because the expression data alone cannot unambiguously distinguish the underlying target network [87]. Further challenges are faced when applying the inference algorithms to limited quantities of experimentally collected noisy data from real biological systems [88]. Suggested solutions to tackle these problems include the usage of gene network motifs [87], network pruning methods [88] or reduced network models [89]. One way to further refine these hypothetical models is to conduct an automated design of new experiments, enabling both iterative model building and candidate model discrimination [90, 91]. Such reverse-engineered gene networks could be of great medical significance, for instance, in identification of drug targets [92].

Computational methods have also been used in assisting the completion of the existing PPI networks by prioritizing the interactions, either observed or missing, that warrant further experimental confirmation. For instance, Yu *et al.* [93] first searched for defective cliques in the incomplete network graphs (nearly complete groups of pairwise interacting nodes), and then they predicted new interaction that can complete these cliques. Albert and Albert [94] showed that machine-learning algorithms can achieve success rates between 20 and 40% for predicting the correct interaction partner of a protein based solely on the presence of conserved interaction motifs within the given network. Ultimately, however, integrated usage of multiple large-scale data types together with local and global topological properties will likely be essential for effective prediction of networks and their functions [51*, 84**]. Towards such integrative approaches, several groups have recently combined multiple heterogeneous data sources to construct global models of gene regulatory networks or PPIs [95–97]. Qi *et al.* [98*] showed that in supervised protein interaction prediction, some of the most important features are actually derived from indirect information sources, such as gene expression measurements. Both indirect statistical relations and direct physical interactions can also be used when predicting or interpreting genetic interactions, observed by comparing phenotypic variations, which are involved in many complex human diseases [99, 100]. However, while most studies have concentrated on snapshots of interactions under particular conditions, it is likely that only by coupling interactions from several functional and temporal states we can reveal truly significant dynamic reconstructions of cellular networks in the future.

CONCLUSION

The large-scale data on biomolecular interactions that is becoming available at an increasing rate enables a glimpse into complex cellular networks. Mathematical graphs are a straightforward way to represent this information, and graph-based models can exploit global and local characteristics of these networks relevant to cell biology. Most current research activities concern the dissection of networks into functional modules, a principal approach attempting to bridge the gap between our very detailed understanding of network components in isolation and the 'emergent' behaviour of the

network as a whole, which is frequently the phenotype of interest on a cellular level. Approaches developed for DNA and protein sequence analysis, such as multiple alignment and statistical over-representation of parts, are being carried over to address these problems. Network graphs have the advantage that they are very simple to reason about, and correspond by and large to the information that is globally available today on the network level. However, while binary relation information does represent a critical aspect of interaction networks, many biological processes appear to require more detailed models. Therefore, we expect that one of the main directions in the development of graph-based methods will be their extension to other types of large-scale data from existing and new experimental technologies. This may eventually prove mathematical models of large-scale data sets valuable in medical problems, such as identifying the key players and their relationships responsible for multifactorial behaviour in human disease networks.

Key Points

- Regardless whether motif searches or network clustering is used for network decomposition, the resulting modules should not be considered as isolated components, but they can interact and frequently overlap with each other.
- Supervised methods that can be adjusted to the needs of the specific biological problem and data sources, without requiring large sets of curated training examples, appear suitable for analysing large-scale network data.
- Several recent works underscore the benefits gained from fully integrated analysis, where the local and global structural and functional properties of the network extracted from different data sources are modelled together.

References

* Articles of particular interest published within the period and scope of the review.

** Articles of extreme interest published within the period and scope of the review.

1. Carter GW. Inferring network interactions within a cell. *Brief Bioinform* 2005;**6**:380–9.
2. Nochomovitz YD, Li H. Highly designable phenotypes and mutational buffers emerge from a systematic mapping between network topology and dynamic output. *Proc Natl Acad Sci* 2006;**103**:4180–5.
3. Papin JA, Hunter T, Palsson BO, *et al*. Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* 2005;**6**:99–111.
4. Curtisa RK, Orešič M, Vidal-Puiga A. Pathways to the analysis of microarray data. *Trends Biotechnol* 2005;**23**: 429–35.
5. *Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 2006;**7**: 198–210.
   This review presents recent work with regards to studying biology at the systems level. The authors summarize several laboratory technologies that can be used to produce genome-scale data for varying types of cellular components and the recently developed data integration techniques to address important biological questions.
6. D'haeseleer P. How does gene expression clustering work? *Nat Biotechnol* 2005;**23**:1499–501.
7. Kunin V, Goldovsky L, Darzentas N, *et al*. The net of life: Reconstructing the microbial phylogenetic network. *Genome Res* 2005;**15**:954–9.
8. **Barabási1 AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
   This review provides an excellent overview of the current network tools that can be used to understand the cell's functional organization and evolution, ranging from large-scale attributes, such as degree distribution and clustering coefficient, through specific network models to motifs and motif clusters. The authors also discuss the impact of network robustness and temporal aspects of interactions on the network behaviour.
9. Grindrod P, Kibble M. Review of uses of network and graph theory concepts within proteomics. *Expert Rev Proteomics* 2004;**1**:229–38.
10. **Albert R. Scale-free networks in cell biology. *J Cell Sci* 2005;**118**:4947–57.
    The author describes how graph representation and graph concepts can be used to analyze the structure of cellular networks and how these attributes can provide insights into their biological function and dynamic responses. Several models and properties of specific cellular networks are overviewed together with their biological interpretation.
11. Lappe M, Holm L. Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol* 2004;**22**: 98–103.
12. Stumpf MP, Wiuf C, May RM. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA* 2005;**102**:4221–4.
13. Przulj N, Corneil DG, Jurisica I. Modeling interactome: scale-free or geometric? *Bioinformatics* 2004;**20**:3508–15.
14. Han J-DJ, Dupuy D, Bertin N, *et al*. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 2005;**23**:839–44.
15. *Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci USA* 2002;**99**:12783–8.
    The authors carried out a novel pathway analysis to identify 'transitive genes' between two given genes from the same biological process of the Gene Ontology. They first constructed a correlation-weighted undirected graph from large-scale yeast microarray expression data, and then showed that the function of unknown genes can be predicted from known genes lying on the same shortest path, perhaps without correlated expression profiles, in a more precise manner than with the conventional hierarchical clustering algorithm.
16. Przulj N, Corneil DG, Jurisica I. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics* 2006;**22**:974–80.
17. Vazquez A, Dobrin R, Sergi D, *et al*. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci USA* 2004;**101**:17940–5.

18. Pradines JR, Farutin V, Rowley S, *et al*. Analyzing protein lists with large networks: edge-count probabilities in random graphs with given expected degrees. *J Comput Biol* 2005;**12**:113–28.
19. Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 2006;**6**: 35–40.
20. ⋆Przulj N, Wigle DA, Jurisica I. Functional topology in a network of protein interactions. *Bioinformatics* 2004;**20**: 340–8.
This study presents a systematic graph-theory-based analysis of the PPI network of S. cerevisiae to construct computational models for describing and predicting the properties of lethal mutations and proteins participating in genetic interactions, functional groups, protein complexes and signaling pathways. These results are based on several graph-based methods, such as network clusters, hubs and shortest path analyses.
21. Samal A, Singh S, Giri V, *et al*. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* 2006;**7**:118.
22. Palumbo MC, Colosimo A, Giuliani A, *et al*. Functional essentiality from topology features in metabolic networks: A case study in yeast. *FEBS Letters* 2005;**579**: 4642–6.
23. Croes D, Couche F, Wodak SJ, *et al*. Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res* 2005;**33**:W326–30.
24. Shlomi T, Segal D, Ruppin E, *et al*. QPath: a method for querying pathways in a protein–protein interaction network. *BMC Bioinformatics* 2006;**7**:199.
25. Guo X, Liu R, Shriver CD, *et al*. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 2006;**22**: 967–73.
26. Scott J, Ideker T, Karp RM, *et al*. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol* 2006;**13**:133–44.
27. Hartwell LH, Hopfiled JJ, Leibler S, *et al*. From molecular to modular cell biology. *Nature* 1999;**402**(6761 Suppl): C47–52.
28. Lee TI, Rinaldi NJ, Robert F, *et al*. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002;**298**:799–804.
29. Milo R, Shen-Orr S, Itzkovitz S, *et al*. Network motifs: simple building blocks of complex networks. *Science* 2002; **298**:824–7.
30. Ideker T, Ozier O, Schwikowski B, *et al*. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002;**18**(Suppl 1):S233–40.
31. Kashtan N, Itzkovitz S, Milo R. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 2004;**20**: 1746–58.
32. Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics* 2006;**22**:1152–3.
33. Schreiber F, Schwobbermeyer H. MAVisto: a tool for the exploration of network motifs. *Bioinformatics* 2005;**21**: 3572–4.
34. Kuang R, Weston J, Noble WS, Leslie C. Motif-based protein ranking by network propagation. *Bioinformatics* 2005;**21**:3711–8.
35. Berg J, Lässig M. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci USA* 2004;**101**: 14689–94.
36. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, *et al*. Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science* 2004;**305**:1107.
37. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003;**100**:12123–8.
38. Han J-DJ, Bertin N, Hao T, *et al*. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004;**430**:88–93.
39. Brun C, Herrmann C, Guenoche A. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* 2004;**5**:95.
40. King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004;**20**:3013–20.
41. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* 2005;**6**:39.
42. Farutin V, Robison K, Lightcap E, *et al*. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins* 2006;**62**:800–18.
43. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. *Proteins* 2004;**54**:49–57.
44. Adamcsek B, Palla G, Farkas IJ, *et al*. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 2006;**22**:1021–3.
45. Rives AW, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci USA* 2003;**100**:1128–33.
46. Arnau V, Mars S, Marin I. Iterative cluster analysis of protein interaction data. *Bioinformatics* 2005;**21**:364–78.
47. Ma HW, Zhao XM, Yuan YJ, *et al*. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* 2004;**20**:1870–6.
48. Tsuda K, Noble WS. Learning kernels from biological networks by maximizing entropy. *Bioinformatics* 2004; **20**(Suppl 1):I326–33.
49. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opinion Microbiol* 2004;**7**:535–45.
50. Nabieva E, Jim K, Agarwal A, *et al*. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 2005;**21**(Suppl 1):i302–10.
51. ⋆Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform* 2005;**6**:34–43.
The author reviews recent computational methods that can be used to integrate gene expression microarray data with other large-scale data sources, such as sequence, interaction, localization and literature data. The emphasis is on increasing the accuracy of gene function prediction and reconstruction of biological networks.
52. Segal E, Shapira M, Regev A, *et al*. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003;**34**: 166–76.

53. Bar-Joseph Z, Gerber GK, Lee TI, *et al*. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 2003;**21**:1337–42.
54. Liao JC, Boscolo R, Yang YL, *et al*. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci USA* 2003;**100**:15522–7.
55. Tanay A, Sharan R, Kupiec M, *et al*. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* 2004;**101**:2981–6.
56. Koyuturk M, Grama A, Szpankowski W. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics* 2004;**20**(Suppl 1):I200–7.
57. Hu H, Yan X, Huang Y, *et al*. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics* 2005;**21**(Suppl 1):i213–21.
58. Yeger-Lotem E, Sattath S, Kashtan N, *et al*. Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. *Proc Natl Acad Sci USA* 2004;**101**:5934–9.
59. Moon HS, Bhak J, Lee KH, *et al*. Architecture of basic building blocks in protein and domain structural interaction networks. *Bioinformatics* 2005;**21**:1479–86.
60. **Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol* 2006;**24**:427–33.
The authors survey the recent approaches to biological network comparison, in terms of network alignment, network integration and network querying, with applications to investigate cellular machinery and to predict protein function and interaction. They emphasize the analogy between the methodologies and concepts used in sequence and network comparison.
61. Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 2004;**32**:6414–24.
62. Mazurie A, Bottani S, Vergassola M. An evolutionary and functional assessment of regulatory network motifs. *Genome Biol* 2005;**6**:R35.
63. Middendorf M, Ziv E, Wiggins CH. Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *Proc Natl Acad Sci USA* 2005;**102**: 3192–7.
64. Zhang LV, King OD, Wong SL, *et al*. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *J Biol* 2005;**4**:6.
65. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
66. Reiss DJ, Avila-Campillo I, Thorsson V, *et al*. Tools enabling the elucidation of molecular pathways active in human disease: application to Hepatitis C virus infection. *BMC Bioinformatics* 2005;**6**:154.
67. Albrecht M, Huthmacher C, Tosatto SC, *et al*. Decomposing protein networks into domain-domain interactions. *Bioinformatics* 2005;**21**(Suppl 2):ii220–1.
68. Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biol* 2003;**4**:R22.
69. Breitling R, Amtmann A, Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics* 2004;**5**:100.
70. Gopalacharyulu PV, Lindfors E, Bounsaythip C, *et al*. Data integration and visualization system for enabling conceptual biology. *Bioinformatics* 2005;**21**(Suppl 1):i177–85.
71. Hu Z, Mellor J, Wu J, *et al*. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res* 2005;**33**:W352–7.
72. Myers CL, Robson D, Wible A, *et al*. Discovery of biological networks from diverse functional genomic data. *Genome Biol* 2005;**6**:R114.
73. Hwang D, Rust AG, Ramsey S, *et al*. A data integration methodology for systems biology. *Proc Natl Acad Sci USA* 2005;**102**:17296–301.
74. Hwang D, Smith JJ, Leslie DM, *et al*. A data integration methodology for systems biology: experimental verification. *Proc Natl Acad Sci USA* 2005;**102**:17302–7.
75. Aragues R, Jaeggi D, Oliva B. PIANA: protein interactions and network analysis. *Bioinformatics* 2006;**22**:1015–7.
76. Baitaluk M, Qian X, Godbole S, *et al*. PathSys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics* 2006;**7**:55.
77. Han K, Ju BH, Jung H. WebInterViewer: visualizing and analysing molecular interaction networks. *Nucleic Acids Res* 2004;**32**:W89–95.
78. Li W, Kurata H. A grid layout algorithm for automatic drawing of biochemical networks. *Bioinformatics* 2005;**21**: 2036–42.
79. Dogrusoz U, Erson EZ, Giral E, *et al*. PATIKAweb: a Web interface for analysing biological pathways through advanced querying and visualization. *Bioinformatics* 2006; **22**:374–5.
80. Carey VJ, Gentry J, Whalen E, *et al*. Network structures and algorithms in Bioconductor. *Bioinformatics* 2005;**21**: 135–6.
81. Balasubramanian R, LaFramboise T, Scholtens D, *et al*. A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics* 2004;**20**:3353–62.
82. Scholtens D, Vidal M, Gentleman R. Local modelling of global interactome networks. *Bioinformatics* 2005;**21**: 3548–57.
83. Zhu D, Hero AO, Cheng H, *et al*. Network constrained clustering for gene microarray data. *Bioinformatics* 2005;**21**: 4014–20.
84. **Vidal M. Interactome modelling. *FEBS Letters* 2005;**579**: 1834–8.
The mini-review focuses on early efforts at mapping a multicellular interactome network. The author emphasizes the importance of the technological development of high-throughput assays, their current limitations resulting in high false positive and negative rates, as well as the benefits gained from integrating data obtained from multiple distinct approaches and the challenges of modelling the constant changes in proteome through time and space.
85. D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 2000;**16**:707–26.
86. Hartemink AJ. Reverse engineering gene regulatory networks. *Nature Biotechnol* 2005;**23**:554–5.
87. Ott S, Hansen A, Kim SY, *et al*. Superiority of network motifs over optimal networks and an application to the

revelation of gene network evolution. *Bioinformatics* 2005; **21**:227–38.

88. Yu J, Smith VA, Wang PP, *et al*. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 2004;**20**: 3594–603.
89. Wagner A. Reconstructing pathways in large genetic networks from genetic perturbations. *J Comput Biol* 2004; **11**:53–60.
90. Yeang CH, Mak HC, McCuine S, *et al*. Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol* 2005;**6**:R62.
91. Kremling A, Fischer S, Gadkar K, *et al*. A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Res* 2004;**14**: 1773–85.
92. Di Bernardo D, Thompson MJ, Gardner TS, *et al*. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnol* 2005; **23**:377–83.
93. Yu H, Paccanaro A, Trifonov V, *et al*. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* 2006;**22**:823–9.
94. Albert I, Albert R. Conserved network motifs allow protein–protein interaction prediction. *Bioinformatics* 2004; **20**:3346–52.
95. Yeang CH, Ideker T, Jaakkola T, *et al*. Physical network models. *J Comput Biol* 2004;**11**:243–62.
96. Nariai N, Tamada Y, Imoto S, *et al*. Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics* 2005;**21**(Suppl 2):ii206–12.
97. Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. *Bioinformatics* 2005;**21**(Suppl 1):i38–46.
98. *Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 2006;**63**:490–500.
The work systematically compares different supervised machine-learning methods, including Random Forest (RF), Naïve Bayes, Decision Tree, Logistic Regression and SVM, to integrate direct and indirect data sources for the protein interaction prediction. They concluded that RF classifier was consistently good in three different problems: prediction of physical interaction, co-complex relationship and pathway co-membership. Also, the importance of different data sources depends on the particular prediction task.
99. Wong SL, Zhang LV, Tong AH, *et al*. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA* 2004;**101**:15682–7.
100. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnol* 2005; **23**:561–6.