Using Biological Networks in Protein Function Prediction and Gene Expression Analysis

Limsoon Wong
School of Computing
National University of Singapore
13 Computing Drive, Singapore 117417
Email: wongls@comp.nus.edu.sg

Abstract

While sequence homology search has been the main work horse in protein function prediction, it is not applicable to a significant portion of novel proteins that do not have informative homologs in sequence databases. Similarly, while statistical tests and learning algorithms based purely on gene expression profiles have been popular for analyzing disease samples, critical issues remain in the understanding of diseases based on the differentially expressed genes suggested by these methods. In the past decade, a large number of databases providing information on various types of biological networks have become available. These databases make it possible to tackle these and other biological problems in novel ways. This paper presents a review on biological network databases and on approaches to protein function prediction and gene expression profile analysis that are based on biological networks.

1 Introduction

Present-day biomedical researchers are confronted by vast amounts of data from genome sequencing, microscopy, high-throughput analytical techniques for DNA, RNA, and proteins, and a host of other new experimental technologies. Coupled with the advances in computing power, this flow of information should enable scientists to model and understand biological systems in novel ways. The appearance of many databases containing information on biological networks, which are critical to understanding the function of genes and proteins in a more holistic way, is particularly exciting. Indeed, many different types of biological network information have been used for analyzing biological data over the past decade.

These networks can be roughly categorized into the following three types:

- Databases of natural biological pathways—e.g., metabolic networks and gene regulation networks.
- Databases of unorganized individual interactions—e.g., protein interaction networks.
- Artificial networks derived from relationships of biological entities—e.g., co-expression networks and Medline abstract co-occurrence networks.

Physical protein interactions constitute a major aspect of all cellular processes. Consequently, analysis of protein interaction networks is expected to produce several types of useful information such as protein

function [72,90], protein complexes [10,39,93] and functional modules [20,21,83]. In particular, there are three main groups of approaches for prediction of protein function using protein interaction networks. The first group [13,17,28,70] predicts the function of an unknown protein based on what functions are over represented among its direct and indirect interaction partners. The second group [4,9,40,55] predicts the function of an unknown protein by assigning to it the function of a protein whose neighbours in the protein interactome have functions that are most similar to the neighbours of this unknown protein. The third group of approaches [6,55,69] clusters proteins based on the similarity of their neighbourhood in the protein interaction network and infer that proteins in the same cluster should have similar functions. All three groups of approaches have their basis on the fact that proteins interact to perform their respective function and, therefore, the function of a protein should correlate with the functions of neighbouring proteins in the protein interaction network [60,91].

The possibility of using gene expression profiling by microarrays for diagnostic and prognostic purposes has also generated much excitement and research in the last ten years. Nevertheless, a number of issues persist such as how to rectify batch effects (i.e., non-biological variations) [44], how to handle missing values [81] and, most importantly, how to identify genes that are meaningful in explaining the difference in disease phenotypes [75]. There are three main groups of approaches, that make use of biological pathways (e.g., enzymatic pathways, gene regulatory pathways, and protein interaction networks), for improving gene selection and for transitioning from the selected genes to the understanding of the sequences of causative molecular events. The first group are the overlap analysis methods [18, 37, 94], which test the significance of the intersection of differentially expressed genes with a biological pathway. The second group are the direct group analysis methods [23, 38, 79], which test whether a biological pathway is differentially expressed as a whole. The third group are the network-based analysis methods [15, 73, 77], which zoom into a subnetwork of a biological pathway and test whether the subnetwork is differentially expressed. All of these approaches have their basis on the fact that every disease phenotype has some underlying biological causes. Therefore, it is reasonable to analyse the gene expression profiles of disease phenotype with respect to the biological contexts provided by biological pathways and protein interaction networks.

This paper is organized as follows. Representative databases [32, 34, 35, 62, 76] of the first type of biological networks—i.e., natural biological pathways—are presented in Section 2. Their consistency and comprehensiveness, as well as their unification for more effective use, are discussed. Representative databases [5,7,31,63,68] of the second type of (unorganized) networks—i.e., protein interaction networks—are presented in Section 3. The noise that is present in them and approaches for dealing with this noise are discussed. The third type of networks are used in many types of analysis, e.g., protein function prediction [12] and disease-gene association studies [30,54]. However, these are diverse and there are few major databases capturing them. Hence we do not describe them further. Then, in Section 4, the three groups of approaches [4, 6, 9, 13, 17, 28, 40, 55, 69, 70] for prediction of protein function using biological networks are presented. After that, in Section 5, the three groups of approaches [15, 18, 23, 37, 38, 73, 77, 79, 94] for improving the reliability of gene selection using biological networks are described. Finally, in Section 6, we briefly discuss some other uses that biological network data have been put to.

2 Biological Pathway Databases

The major biological pathway databases include those that are curated by a single lab (e.g., KEGG, BIOCYC), by a community of collaborating labs (e.g., WikiPathways, Reactome), and by commercial companies (e.g., Ingenuity, Molecular Connections), as well as those that are derived by an integration of these databases (e.g., Pathway Commons, PathwayAPI):

- KEGG PATHWAY [34], accessible at `http://www.genome.jp/kegg`, contains about 380 pathway maps for metabolism, genetic information processing, environmental information processing, and other cellular processes that are curated manually from over 120,000 published articles.[1] The content of the database can be downloaded in XML format, as well as accessed using an API (application programming interface).

- BIOCYC [35], accessible at `http://biocyc.org`, is a set of more than 1,000 databases. Each database in this collection describes the genome and metabolic pathways of a single organism. The databases are categorized into tiers. Tier 1 are curated manually. Tier 2 are generated based on reviewed predictions by the Pathologic software [59]. Tier 3 are generated based on unreviewed predictions by the Pathologic software. The content of BIOCYC can be downloaded in BioPAX, SBML, and other formats, as well as accessed using an API.

- WikiPathways [62], accessible at `http://www.wikipathways.org`, is curated by a community of collaborating labs in a Wikipedia-like setting. It has information on about 360 human pathways consisting of about 4,400 genes. Each pathway in WikiPathways is a wiki page that presents the pathway diagram, the component gene, protein, and metabolite lists. The main content of the database can be downloaded in the form of GPML, as well as accessed through a web service API.

- Reactome [32], accessible at `http://www.reactome.org`, is also curated manually by a community of collaborating labs. It contains a total of 13,197 pathways from 21 organisms, including 1,112 pathways from human. The main content of the database can be downloaded in BioPax, SBML, and other formats.

- Ingenuity Systems offers the Ingenuity® Knowledge Base and the associated IPA analysis software on a commercial basis. The knowledge base is a repository of biological interactions and other information. The content of the knowledge base can only be accessed using proprietary tools such as IPA and is typically returned to the user in the form of an image file. More information can be obtained at `www.ingenuity.com`.

- Molecular Connections offers NetPro™ on a commercial basis. This is a comprehensive database covering more than 320,000 protein-protein and protein-small molecule interactions in the biological pathways of 20 organisms. These interactions and other information are curated manually. Direct access by SQL queries and XML-format downloads are supported. More information can be obtained at `www.molecularconnections.com`.

- Pathway Commons, accessible at `www.pathwaycommons.org`, provides convenient access to a collection of publicly available pathways from multiple sources. The data from these multiple sources are made available by Pathway Commons in a common format. Pathway Commons does not perform any unification of the underlying pathways. That is, if the information of a pathway is contained in n source databases, Pathway Commons presents them as n separate pathways.

- PathwayAPI [76], accessible at `http://www.pathwayapi.com`, is database of over 450 unified human pathways consisting of over 60,000 interactions derived from an integration of KEGG, WikiPathways, and Ingenuity. In contrast to Pathway Commons, if the information of a pathway is contained in n source databases, PathwayAPI merges them into a single consistent unified pathway. The main content of PathwayAPI can be downloaded as a MySQL dump or as a CSV file; it can also be accessed in JSON format via an API.

[1] All statistics given on KEGG and other databases, unless mentioned otherwise, are based on information available on their respective websites on 13 February 2011.

As these biological pathway databases are generally curated manually, their content can be regarded as reliable. However, it is important to be aware of the following two issues before using these databases. Firstly, many biological pathways are curated only in some databases and not in other databases. That is, none of the databases is sufficiently comprehensive in terms of the number of biological pathways that they curate. Secondly, even when a biological pathway is curated in two databases, there is usually some disagreement between these two databases on this pathway. For example, a recent study [76] shows that, for a pathway that is as pervasive as the human apoptosis pathway, the agreement between KEGG, Ingenuity, and WikiPathways is a mere 32–46% in terms of gene overlap and an alarming 11-16% in terms of interaction overlap. The same study also shows that the agreement on many other pathways is no better. This lack of agreement can be partially attributed to the fact that the boundaries of many biological pathways are not clearly defined [24]. However, it also strongly suggests that the manual curation effort of these databases is not sufficiently comprehensive even at the individual pathway level.

The obvious solution to these two issues is to integrate these biological pathway databases. Despite impressive progress in broad-scale general data integration technologies in the past two decades [86], there are significant challenges that have to be overcome to achieve such a unified database, including incompatibility of access methods, incompatibility of data formats, incompatibility of molecular representations, and incompatibility in naming of pathways. There is a variety of approaches to deal with these four incompatibility problems. For example, Pathway Commons and PathCase [19] can be considered as taking the "aggregator" approach. In this approach, a common access method and data format are adopted or developed for a set of pathways imported from a collection of source databases. The aggregator approach does not perform any unification of the underlying pathways—viz., if n source databases each contain information on a particular pathway, that pathway is presented by the aggregator as n separate pathways. On the other hand, GenMapp [67], Cytoscape [71], and PathVisio [84] can be considered as taking the "converter" approach. Basically, these tools support the import and export of biological pathways in a variety of formats, even though these tools are designed mainly for exploring, visualizing, and editing biological pathways. Lastly, PathwayAPI [76] can be considered as taking the "full unification" approach. In this approach, pathways in difference source databases that are meant to represent the same pathway are merged and molecular objects mentioned in the different source pathways that are meant to represent the same objects are matched. This approach is technically more difficult than other approaches; but it has the advantage of presenting a more coherent comprehensive view of the pathways.

Among the four types of incompatibilities that are encountered when unified pathways are constructed, the resolution of the incompatibility in the naming of pathways offers an interesting lesson. There are three basic ways to detect whether two pathways in two databases are meant to represent the same biological pathway—viz., match by large overlap of genes, match by large overlap of interactions, and match by similarity of pathway names. We consulted a number of experts in computer science and biology on which choice to adopt when we were developing PathwayAPI. Almost without exception, it was thought that matching by largest overlap of interactions would give the best result, as it was feared that different databases would give the same pathway names that are very different. Unfortunately, matching pathways by largest overlap of interactions requires a threshold on the overlap. Too small a threshold, we get a large number of false-positive matches. Too large a threshold, we get a large number of false negatives. In fact, we tried a whole continuum of thresholds and did not find a good compromise. Fortunately, it turns out that different databases actually do not give very different names to the same pathways. Thus a strategy based on approximate longest substring match of pathway names works well in practice [76].

3 Protein Interaction Databases

Although many interactions of genes and proteins have been organized into natural biological pathways, not all known interactions can yet be put into the context of a natural biological pathway. This gives rise to protein interaction databases, which focus on capturing pairwise interaction information but generally do not seek to organize these pairwise interactions into functional groups or pathways. Nevertheless, such protein interaction databases are useful in many applications because they cover far more interactions than those found in natural biological pathway databases.

The major protein interaction databases include MINT, BioGRID, DIP, HPRD, and STRING:

- MINT [7], accessible at `http://mint.bio.uniroma2.it/mint`, contains 90,695 physical protein interactions curated from the literature.
- BioGRID [5], accessible at `http://www.thebiogrid.org`, contains 193,484 physical protein interactions and 177,348 genetic interactions curated from the literature. It is especially complete for yeast protein interaction data.
- DIP [68], accessible at `http://dip.doe-mbi.ucla.edu`, contains 71,276 protein interactions curated from the literature. It focuses on model organisms (e.g., yeast, fruit fly, *E. coli*, *C. elegans*) and has less data on other organisms.
- HPRD [63], accessible at `http://www.hprd.org`, contains about 40,000 protein interactions curated from the literature. It focuses on human protein interactions.
- STRING [31], accessible at `http://string-db.org`, is a database of known (by copying from MINT, BioGRID, DIP, HPRD, etc.) and predicted protein interactions. It covers the interactions of about 2.59 million proteins from 630 organisms. There is an important caveat: A large fraction of protein interactions in STRING are predicted ones; these predicted interactions may not be reliable.

Protein interactions are often viewed as a form of binary relationships—i.e., interact or not. Nonetheless, it is important to be aware of the following two issues before using them. Firstly, protein interactions in these databases vary in reliability. Protein interaction data are generated by experiments such as co-immunoprecipitation, synthetic lethal screening, tandem affinity purification, and two-hybrids [58]. Some of these experimental methods—e.g., two-hybrids—are highly susceptible to noise and may have a high false-positive rates [78, 85]. Secondly, some of these experimental methods—e.g., tandem affinity purification—identify groups of proteins that interact together to form a complex, though the proteins within each group may not be directly interacting [22]. Nevertheless, treating proteins captured by a bait protein as interacting with the bait protein does not seem to have a negative effect on important applications such as inferring protein function [13] and identifying protein complexes [50].

For some analysis, it is crucial to use a subset of protein interaction data that are more reliable. Consequently, much efforts have been invested to develop solutions to this problem [14]. An obvious idea for ranking the reliability of protein interactions is based on the sharing of a common cellular localization or a common cellular role, as a pair of interacting proteins is generally expected to be localized to the same cellular component or to have a common cellular role [57, 78]. The main shortcoming of this approach is that protein functional annotations and cellular localization information are often incomplete. Besides, even if two proteins share a common cellular localization or a common functional role, there is still a chance that they do not interact in real life.

Another early idea is based on the reproducibility and nonrandomness of the observation of an interaction [11, 26]. Obviously, an interaction that is observed in two or more separate experiments is more reliable than one that is observed in just one experiment. The main shortcoming of this approach is that it requires multiple independent interaction experiments to be performed on the proteins to confirm the reliability of their interactions. As such, if the additional experimental data are not available, which is often the case, this method cannot be used.

As the additional information required by these approaches are often unavailable, a new class of reliability indices that are based solely on the topology of the neighborhood of an interacting pair of proteins in the interactome has been developed [8, 14]. One of the most important early examples of this idea is the Czekanowski-Dice distance [6], defined as $\mathrm{CD}_{u,v} = 2|N_{u,v}|/(|N_u| + |N_v|)$, where $N_{u,v}$ is the set of interaction partners shared by proteins u and v, and N_u and N_v are respectively the set of interaction partners of u and v. Two proteins that have many interaction partners in common must share some physical or biochemical characteristics that allow them to bind to these common interaction partners. Consequently, they are also more likely to share a common cellular role or a common cellular function or to belong to the same protein complex. This makes them more likely to interact. Therefore, a reliability index for a pair of reported interacting proteins can be formulated in terms of the proportion of interaction partners that two proteins have in common, as in $\mathrm{CD}_{u,v}$.

It is possible to combine all three approaches. Suppose there is some additional information—such as functional annotations or multiple experiments—to estimate the reliability $r_{u,v}$ of an interaction between protein u and v according to the first two approaches. Assuming independence, the probability of u and v having a common interaction partner w is $r_{u,w}r_{w,v}$. Then $\mathrm{CD}_{u,v}$ incorporating this information is $\mathrm{CD}_{u,v} = 2\sum_{w\in N_{u,v}} r_{u,w}r_{w,v}/(\sum_{w\in N_u} r_{u,w} + \sum_{w\in N_v} r_{w,v})$. Another refinement is to add a damping term λ to the denominator because $\mathrm{CD}_{u,v}$ has large fluctuations when u and v have too few neighbours. A third refinement is to use an iteration process similar to an expectation maximization; to wit, let $\mathrm{CD}^i_{u,v}$ be the $\mathrm{CD}_{u,v}$ value computed in the ith iteration, then $\mathrm{CD}^{i+1}_{u,v} = 2\sum_{w\in N_{u,v}} \mathrm{CD}^i_{u,w}\mathrm{CD}^i_{w,v}/(\sum_{w\in N_u} \mathrm{CD}^i_{u,w} + \sum_{w\in N_v} \mathrm{CD}^i_{w,v} + \lambda)$ and setting $\mathrm{CD}^0_{u,v} = r_{u,v}$. The rationale for this iteration process is an intuitive one. Assuming that we accept the Czekanowski-Dice distance as a good model of the reliability of a protein interaction, then $\mathrm{CD}^1_{u,v}$ is a more accurate estimate than $\mathrm{CD}^0_{u,v} = r_{u,v}$. So substituting it for $r_{u,v}$ in the formula should give us a more accurate $\mathrm{CD}^2_{u,v}$, and so on.

These refinements have been shown to significantly improve $\mathrm{CD}_{u,v}$ and other related topology-based reliability indices for protein interactions. In particular, a recent study [49] used the DIP yeast data set for assessment. 54.7% of the interacting protein pairs reported in DIP are co-localized. Since proteins in general can only interact when they are co-localized, this suggests up to 45.3% noise in the data set. After these pairs were ranked using the iterated version of $\mathrm{CD}_{u,v}$, about 90% of the top 30% of interacting pairs are co-localized.

Nevertheless, the performance of $\mathrm{CD}_{u,v}$ and related indices deteriorates when the input interaction network is sparse, due to the lower number of direct and indirect interactions in such networks [14]. A recent idea [42, 92] to overcome this problem is using a larger interaction neighbourhood via a manifold embedding. Here, a protein-protein similarity matrix is first computed based on the shortest distance—in terms of number of hops in the initial protein interaction network—between each pair of proteins. Then multi-dimensional scaling is applied to this similarity matrix to embed each protein into a low-dimensional space. After that, a graph is defined by connecting proteins that are close to each other in this low-dimensional space. Finally, an index like $\mathrm{CD}_{u,v}$ is applied to this graph to estimate the likelihood that proteins u and v interact. Experiments have confirmed that, for sparse protein interaction networks, this additional step of manifold-embedding has led to much better performance [92].

Besides noise dealt with by the approaches above, protein interaction assays are also plagued by false negatives. The detection of false negatives is considerably more difficult because new protein interactions have to be predicted. A variety of approaches have been reviewed in earlier papers [14], including gene-fusion events [53], interacting domains [25], interacting motifs [45], co-evolution of proteins or residues [33], topology of protein-protein interaction networks [61], and machine learning from multiple information types [65]. Incidentally, it is possible to use topology-based indices like $\mathrm{CD}_{u,v}$ for predicting new interactions [87]—one can predict two proteins u and v to interact if the value of $\mathrm{CD}_{u,v}$ is sufficiently high.

4 Protein Function Prediction Using Biological Networks

Proteins are important building blocks that contribute to key processes within cells. The elucidation of mechanisms underlying protein functionality is an important pursuit and remains a challenging task in computational biology [27, 41]. Sequence similarity search methods like BLAST [3] are the primary tools for this problem. However, a non-negligible proportion of protein sequences do not have identifiable informative homologs in current databases. Therefore, a variety of new bioinformatics methods have been developed for inferring protein function using "guilt by association" of other functional properties to complement sequence similarity search [27].

In particular, many approaches have been proposed to use protein interaction networks for protein function prediction [72]. These approaches can be roughly divided into three groups. The first group [13, 17, 28, 70] is based on the hypothesis that proteins having similar functions are topologically close in the protein interaction network. This is a reasonable hypothesis because a pair of proteins that participate in the same cellular processes or localize to the same cellular compartment are many folds more likely to interact than a random pair of proteins [49]. The second group [4, 9, 40] is based on the hypothesis that proteins with similar function have interaction neighbourhoods that are similar. This is also a reasonable hypothesis because, when the proteins in the neighbourhood of a protein have similar functions as the proteins in the neighbourhood of another protein, the two proteins are likely to operate in similar environments and have similar properties. The third group [6, 55, 69] clusters proteins based on similarity of certain features—in particular their neighbourhood in the protein interaction network—and hypothesizes that such groups of proteins are functionally coherent. This is also a reasonable hypothesis and, as we shall see later, corresponds to the "if and only if" form of the other two hypotheses.

An early example of the first group is the "majority vote" method which assigns a protein with the function that is over represented among its interaction partners [28, 70]. Another example is to apply global optimization techniques—e.g., Markov random fields—to transfer the function of a protein from its neighbour and also to propagate predictions so that the function of proteins without characterized neighbors can be predicted [17].

A shortcoming of these methods is that the function predicted for a protein is generally taken from proteins that directly interact with it. Even those global optimization methods that propagate a function from a protein u that is several hops away to a protein v essentially force the whole chain of proteins connecting u and v to have that function. Yet it has been observed that, while 48% of yeast proteins share some function with their immediate interaction partners in BioGrid, 69% share some function with their indirect interaction partners [11]. Hence, at least with respect to yeast proteins, these methods' sensitivity is limited to 48%. Another shortcoming of these methods is that they generally do not take into account the reliability of the protein interaction network used. For example, the majority vote method gives all the interaction partners of the unknown protein equal vote, regardless of the reliability

of those interactions. This affects the precision of these methods.

A more recent example of the first group—the fsweight method [11]—overcomes these shortcomings by weighted voting of both direct and indirect neighbours. This method defines the functional similarity weight $S_{FS}(u, v)$ between two proteins u and v based on the size of the intersection of their interaction neighbourhoods. $S_{FS}(u, v)$ is a variation of $\mathrm{CD}_{u,v}$ where the size of the intersection is defined taking into account the reliability of individual interactions and giving equal weight the interaction neighbourhoods of u and v. Then a direct neighbour v of a protein u having function a votes for function a with weight $S_{FS}(u, v)$. Similarly, an indirect neighbour v' of a protein u having function a votes for function a with weight $S_{FS}(u, v')$. The function a that receives a total amount of votes exceeding a threshold is assigned as a function of protein u. Experiments have shown that this fsweight method has good recall and precision. For example, in a study [13] based on the BioGrid yeast protein interaction network, out of about 100 biological processes considered, fsweight achieved ROC score of 0.8 for about 80 of these biological processes and 0.9 for about 60 of these biological processes. Cross validation experiments have also shown that this method can provide substantial number of high-quality predictions that cannot be inferred from sequence homology [13].

An early example of the second group is LaMoFinder [9]. It first discovers network motifs [2] from a protein interaction network. A "network motif" is a frequently occurring connection pattern in the network—for example, the "triangle" motif represents the topology of "A interacts with B, B interacts with C, C interacts with A", where A, B, C are placeholders for proteins to be mapped. After that, the placeholders in these network motifs are labelled with functions of proteins in subnetworks that are mapped to them, thereby determining the various biological contexts in which such a network motif occurs. When the subnetwork of a protein of unknown function and its functionally labelled neighbourhood is aligned to such a network motif, its function can be inferred from the vertex that this protein is mapped to in the network motif.

A limitation of LaMoFinder [9] is that it works only for proteins in subnetworks that can be mapped to such network motifs. A generalization that avoids this limitation is to find the best pairwise graph alignment of the functionally labelled subgraph rooted at the unknown protein to functionally labelled subgraphs rooted at other nodes in the protein interaction network [40].

Both LaMoFinder [9] and this refinement [40] rely on topological match of subnetworks. Thus their performance is affected in less reliable protein interaction networks which have more false interactions and missing interactions. A recent idea [4] to overcome this shortcoming uses probabilitistic technique to define and match network patterns as follows. First, the "affinity" of a protein u to another protein v in the interaction network is defined as the steady-state probability $p_{u,v}$ of random-walks from the first protein to the second protein. The affinity of a protein v to a function a is defined as $S_f^v(a) = \sum_u p_{u,v}$, where $u \neq v$ ranges over proteins having function a. The vector S_f^v is then normalized to give the functional profile of the protein v. The function of an unknown protein is predicted by a weighted voting of the k proteins who are its nearest neighbours in terms of functional profile. It has been shown [4] that this method has very good recall and precision, and outperforms the fsweight method when the protein interaction network is sparse.

Some early examples of the third group are PRODISTIN [6] and a closely related method [69]. These methods cluster several proteins into the same group if these proteins have a significantly larger number of common interaction partners than what is expected from a random network. Then, by assuming that proteins in the same group are functionally coherent, an unknown protein in a group can be assigned functions that are common among other members of the group. Interestingly, such methods [6, 69] share the same hypothesis as the first group of methods [28, 70]. To see this, we first note that, by construction, members of the same group are generally interaction partners of each other. Thus, assigning to a protein

a function that is common among other members of the group is akin to a majority vote of interaction partners of the protein. At the same time, the PRODISTIN type of methods [6, 69] uses the hypothesis in a stronger way than the majority-vote type of methods [28, 70] because, as implied by the clustering step, the former further requires that the interaction partners from which the votes are taken are also interaction partners of each other.

A more recent idea of the third group [55] first clusters proteins based on the similarity of the network motifs (called graphlets by the authors) that they participate in. A protein of unknown function can then be assigned the same function as other annotated proteins in the same cluster. Interestingly, such a method [55] shares the same hypothesis as the second group of methods typified by LaMoFinder [9]. To see this, we first note that, by construction, members of the same cluster are mapped to the same network motifs. Thus, assigning to a protein the same function as other proteins in the same cluster is akin to inferring function from other proteins aligned to the same network motifs. However, this method [55] uses the hypothesis in a stronger way than LaMoFinder [9] because, as implied by the clustering step, the former further requires that the interaction partners from which the function is taken also have the same network motifs as each other.

Finally, we should briefly mention an idea that is related to—and can be considered as a generalization of—the second group of methods. This is the idea of comparing and aligning the biological networks of two different species [36, 56]. Here, after the network alignment is performed, the alignment can be analysed and used to infer protein functions based on shared topology in the aligned (sub-)networks of the two species [43].

5 Microarray Analysis Using Biological Networks

Many approaches [46,47,51,52,82,96] have been proposed for the inference of differentially expressed genes that are useful for diagnosis of diseases and prognosis of treatment response. However, the statistical significance of the selected genes and the reproducibility of the resulting diagnosis system have a high degree of uncertainty. In particular, many of these methods produce gene lists that do not have significant overlap when they are applied to different data sets of the same disease phenotypes [95]. Furthermore, the transition from the selected genes to the understanding of the sequences of causative molecular events is unclear [75].

In order to qualitatively improve the statistical power of microarray analysis methods and the reliability of the results, additional dimensions present in the problem have to be brought into consideration. For example, each disease generally has an underlying cause. So there should be a unifying biological theme—a biological pathway or a subnetwork of protein interactions—for genes that are truly associated with the disease. Hence the uncertainty in the reliability of the selected genes can be reduced by considering the molecular functions and the biological processes associated with the genes. Such a unifying biological theme is also a basis for inferring the underlying cause of the disease phenotype. There are a number of approaches to analyze gene expression data with respect to biological context. These approaches can be roughly divided into three groups [75]. The first group comprises the overlap analysis methods [18,37,94]. The second group comprises the direct group analysis methods [23, 38, 79]. The third group comprises the network-based analysis methods [15, 73, 77].

The overlap analysis methods [18, 37, 94] share a common principle. They basically first determine a list of differentially expressed genes. This list of genes is then intersected with each biological pathway and the statistical significance of the overlap is computed, e.g., by a hypergeometric test. The differentially expressed genes that are in a statistically significant intersection with a pathway are declared candidate

biomarkers and causal factors of the disease phenotypes. ORA [37] is an example of this group of methods.

A shortcoming of these methods is that the starting list of differentially expressed genes is defined using some test statistics with arbitrary thresholds. Different test statistics and different thresholds result in a different list of differentially expressed genes. As a result, the outcome of the whole procedure is not stable. Another shortcoming is that a real causal gene that is not differentially expressed can never be suggested by these methods. Note that it is not uncommon for a real causal gene underlying a disease phenotype to be not differentially expressed. For example, suppose a gene A upregulates both genes B and C in normal people. Suppose also that genes A, B, and C are observed to be highly expressed in normal samples, and that only gene A is observed to be highly expressed in disease samples. Then only genes B and C are differentially expressed and have a chance to be suggested by these methods. In such as a situation, since A is not suggested by these methods as important, we need to postulate mutations in B and C in order to explain their differential expression. However, a more likely explanation is that A has a mutation that does not change its expression but changes its ability to upregulate B and C.

The direct group analysis methods [23, 38, 79] work on a different principle to avoid the shortcomings above. They do not start from differentially expressed genes. Instead, they start from each individual biological pathway and test whether the pathway is differentially expressed as a whole. This is done by comparing the distributions of expression values of genes on the biological pathway with the distributions of expression values of all the other genes, e.g., by a weighted Kolmogorov-Smirnov test. FCS [23] and GSEA [79] are examples of this group of methods. These direct group analysis methods are able to detect more subtle changes in gene expression profiles. For example, if the majority of genes on the biological pathway have small but correlated expression level changes, they can still result in a high statistical significance of the biological pathway under a direct group analysis method, even though the whole group is likely to be missed by all the overlap analysis methods.

A shortcoming of the direct group analysis methods is that they work on a whole-pathway basis and, thus, they can miss a large pathway when a small subnetwork in that pathway is responsible for the disease phenotype. Continuing with our earlier example, suppose the pathway has a second branch involving 30 other genes besides the branch containing the upregulation of B and C by A. Suppose these 30 other genes are not differentially expressed while, as before, B and C are differentially expressed and A is highly expressed. Due to the predominant of the 30 other non-differentially expressed genes in the pathway, the whole pathway may not be considered differentially expressed by these direct group analysis methods.

The network-based analysis methods [15,73,74,77] are the latest development in gene expression analysis. Instead of considering a whole biological pathway, these methods try to identify subnetworks that are significantly differentially expressed. An early example of this approach is NEA [73]. For each regulator in a biological pathway, NEA considers it and all its targets in the pathway as a group, which is then evaluated in a GSEA-like manner. This splitting into separate regulatory groups pinpoints the transcriptional regulators whose targets exhibit consistent differential expression pattern, leading to a more precise hypothesis that explains the disease phenotype. A shortcoming of NEA is that it considers only immediate regulator-regulatee relationship in a biological pathway. In particular, it may not be able to detect a linear chain of genes that are differentially expressed in a pathway, even though a biologist would consider such a linear chain highly suggestive of the underlying cause of the disease phenotype.

The latest addition to this family of methods is SNet [74], which overcomes this shortcoming as follows. SNet first maps the genes that are highly expressed—but not necessarily differentially expressed—in most samples of the disease phenotype in question to biological pathways or protein interaction networks. Other genes and proteins in these pathways and networks are discarded. Each remaining connected component is considered to be a candidate subnetwork. A score is then computed for each subnetwork s for each sample i based on the genes —in the subnetwork s—that are highly expressed in that sample i. A t-statistics

is then computed between the scores of each subnetwork s in samples having the disease phenotype and the scores of subnetwork s in other samples. The obtained t-statistics is compared to a null distribution obtained by permutating class labels to decide whether the subnetwork s is significant. Experiments have shown that SNet produces subnetworks that are both much more substantial in size and much more consistent cross independent data sets of the same disease phenotypes than other methods [74]. In particular, it was tested on four diseases [74]. For each disease, there were two independent data sets obtained on different microarray platforms. SNet was run independently on the two independent data sets and the genes it selected were intersected. In each disease, it selected less than 100 genes in each of the two independent data sets and achieved 51.2% to 93.0% agreement between the two independent data sets.

6 Final Remarks

In more recent years, biological networks have been put to many more interesting uses. Some of these interesting analysis enabled by biological network data are briefly mentioned below:

- Protein complex discovery. Proteins often perform function by aggregating into complexes to perform sophisticated biological tasks. This has motivated approaches to identify protein complexes computationally from protein-protein interactions data. Most of these approaches are based on the hypothesis that proteins within a complex should have more interactions with each other than with proteins outside the complex [1, 10, 20, 50, 64, 93]. However, these algorithms have relative low sensitivity and precision because, when mapped to protein interaction networks, perhaps due to the noise and incompleteness of protein interaction network data, many protein complexes appear to have low density [87].

- Countering pathogen drug resistance. There is a need to address the emergence of drug-resistant pathogens, e.g., *M. tuberculosis*. It has been proposed that a systems-level analysis of the biological pathways and protein interactions in these pathogens is critical to gaining insights into their routes to drug resistance [66,87]. A pioneering idea [66] in this direction uses protein interaction networks to identify possible paths between known drug targets and known mechanisms for drug resistance such as efflux pumps and cytochromes-like enzymes; then gene expression experiments can be performed to reveal which of these paths are activated; after that, analysis can be made to identify druggable proteins in these paths to serve as "co-targets" to deactivate the drug-resistance mechanisms in the pathogen. Another idea [29, 87] is to identify a minimum number of proteins whose simultaneous inhibition can disrupt a maximum number of pathways.

- Epistatic interaction detection. Genome-wide association study is an effort to examine the association between phenotype and genotype. As many diseases have complex underlying mechanisms, analysis at the level of single SNP (single-nucleotide polymorphism) is not sufficient. There is thus intense interest to explore the interactions of multiple SNPs—the so-called epistatic interactions—to uncover more significant associations [16]. Exhaustively considering all the possible SNP combinations is not feasible. One promising idea that has recently emerged is to restrict the search to SNPs in the loci of genes that are within the same biological pathways and are proximate in a protein interaction network [80].

- Disease gene identification. This has long been a major research effort. The availability of networks based on protein interactions, known gene-phenotype associations, disease phenotype similarities, and other forms of associations has open new avenues for infering gene-phenotype associations.

For example, causative genes for diseases that are phenotypically similar have been observed in the same biological module or are tightly linked in a protein interaction subnetwork [30, 88]. A recent idea [48, 89] in this direction is to formulate some scores that correlate the distance between genes in protein interaction networks to the similarity of disease phenotypes to which the genes are associated. Variations of this idea include using networks derived from hyperlinks between OMIM pages that describe genes and diseases [54], instead of protein interaction networks.

In short, given the holistic information in biological networks, the possibilities are immense.

Acknowledgements

This work is supported in part by a Singapore Ministry of Education Tier 2 grant MOE2009-T2-2-004, a Singapore Agency for Science Technology & Research grant SERC-102-1010-0030, and a Singapore National Research Foundation grant NRF-G-CRP-2007-04-082(d).

References

[1] B. Adamcsek et al. CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.

[2] U. Alon. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[4] P. Bogdanov and A. Singh. Molecular function prediction using neighbourhood features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(2):208–217, 2010.

[5] B. J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bahler, V. Wood, K. Dolinski, and M. Tyers. The BioGRID interaction database: 2008 update. *Nucleic Acids Research*, 36(Database Issue):D637–D640, 2008.

[6] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6, 2003.

[7] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. MINT: The Molecular INTeraction database. *Nucleic Acids Research*, 35:D572–D574, 2007.

[8] J. Chen, H. N. Chua, W. Hsu, M.-L. Lee, S.-K. Ng, R. Saito, W.-K. Sung, and L. Wong. Increasing confidence of protein-protein interactomes. In *Proceedings of 17th International Conference on Genome Informatics*, pages 284–297, Yokohama, Japan, 2006.

[9] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng. Labeling network motifs in protein interactomes for protein function prediction. In *Proceedings of 23rd IEEE International Conference on Data Engineering*, pages 546–555, Istanbul, Turkey, 2007.

[10] H. N. Chua, K. Ning, W.-K. Sung, H. W. Leong, and L. Wong. Using indirect protein-protein interactions for protein complex prediction. *Journal of Bioinformatics and Computational Biology*, 6(3):435–466, 2008.

[11] H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[12] H. N. Chua, W.-K. Sung, and L. Wong. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 23(24):3364–3373, 2007.

[13] H. N. Chua, W.-K. Sung, and L. Wong. Using indirect protein interactions for the prediction of gene ontology functions. *BMC Bioinformatics*, 8(Suppl 4):S8, 2007.

[14] H. N. Chua and L. Wong. Increasing the reliability of protein interactomes. *Drug Discovery Today*, 13(15/16):652–658, 2008.

[15] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3:140, 2007.

[16] H. J. Cordell. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[17] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10:947–960, 2003.

[18] S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin. MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, 2003.

[19] B. Elliott, M. Kirac, A. Cakmak, G. Yavas, S. Mayes, E. Cheng, Y. Wang, C. Gupta, G. Ozsoyoglu, and Z. M. Ozsoyoglu. PathCase: Pathways database system. *Bioinformatics*, 24(21):2526–2533, 2008.

[20] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[21] L. Gao, P. G. Sun, and J. Song. Clustering algorithms for detecting functional modules in protein interaction networks. *Journal of Bioinformatics and Computational Biology*, 7(1):217–242, 2009.

[22] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[23] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[24] M. L. K. P. Green. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, 34(13):3687–3697, 2006.

[25] D. S. Han, H. S. Kim, W. H. Jang, S. D. Lee, and J. K. Suh. PreSPI: A domain combination based prediction system for protein-protein interaction. *Nucleic Acids Research*, 32(21):6312–6320, 2004.

[26] G. T. Hart, I. Lee, and E. M. Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8(1):236, 2007.

[27] T. Hawkins and D. Kihara. Functional prediction of uncharacterized proteins. *Journal of Bioinformatics and Computational Biology*, 5(1):1–30, 2007.

[28] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):525–531, 2001.

[29] F. Hormozdian, R. Salari, V. Bafna, and S. C. Sahinalp. Protein-protein interaction network evaluation for identifying potential drug targets. *Journal of Computational Biology*, 17(5):669–684, 2010.

[30] T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 18:644–652, 2008.

[31] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database Issue):D412–D416, 2009.

[32] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database Issue):D428–D432, 2005.

[33] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci USA*, 105(3):934–939, 2008.

[34] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database Issue):D355–D360, 2010.

[35] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 19:6083–6089, 2005.

[36] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker. PathBLAST: A tool for alignment of protein interaction networks. *Nucleic Acids Research*, 32(Suppl 2):W83–W88, 2004.

[37] P. Khatri and S. Draghici. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[38] S. Y. Kim and D. J. Volsky. PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 8(6):144, 2005.

[39] A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.

[40] M. Kirac and G. Ozsoyoglu. Protein function prediction based on patterns in biological networks. In *Proceedings of 12th Annual International Conference on Research in Computational Molecular Biology*, pages 197–213, Singapore, 2008.

[41] C. H. Koh, S. Lin, G. Jedd, and L. Wong. Sirius PSB: A generic system for analysis of biological sequences. *Journal of Bioinformatics and Computational Biology*, 7(6):973–990, 2009.

[42] O. Kuchaiev, M. Rasajski, D. J. Highham, and N. Przulj. Geometric de-nosing of protein-protein interaction networks. *PLoS Computational Biology*, 5(8):e1000454, 2009.

[43] O. Kuchaiev, T. Milenkovic, V. Memisevic, W. Hayes, and N. Przulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, 7(50):1341–1354, 2010.

[44] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

[45] H. Li, J. Li, and L. Wong. Discovering motif pairs at interaction sites from sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996, 2006.

[46] J. Li, H. Liu, J. R. Downing, A. E.-J. Yeoh, and L. Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19:71–78, 2003.

[47] J. Li, H. Liu, and L. Wong. Mean-entropy discretized features are effective for classifying high-dimentional biomedical data. In *Proceedings of 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, pages 17–24, Washington, DC, 2003.

[48] Y. Li and J. C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

[49] G. Liu, J. Li, and L. Wong. Assessing and predicting protein interactions using both local and global network topological metrics. In *Proc. 19th Intl Conf on Genome Informatics (GIW)*, pages 138–149, 2008.

[50] G. Liu, L. Wong, and H. N. Chua. Complex discovery from weighted PPI networks. *Bioinformatics*, 25(15):1891–1897, 2009.

[51] H. Liu, J. Li, and L. Wong. Use of extreme patient samples for outcome prediction from gene expression data. *Bioinformatics*, 21(16):3377–3384, 2005.

[52] Z. Liu et al. A multi-strategy approach to informative gene identification from gene expression data. *Journal of Bioinformatics and Computational Biology*, 8(1):19–38, 2010.

[53] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[54] T. Matsunaga, S. Kuwata, and M. Muramatsu. Computational gene knockout reveals transdisease-transgene association structure. *Journal of Bioinformatics and Computational Biology*, 8(5):843–866, 2010.

[55] T. Milenkovic and N. Przulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008.

[56] T. Milenkovic, W. L. Ng, W. Hayes, and N. Przulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121–137, 2010.

[57] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl. 1):i302–i310, 2005.

[58] S.-K. Ng and S.-H. Tan. Discovering protein-protein interactions. *Journal of Bioinformatics and Computational Biology*, 1(4):711–741, 2004.

[59] S. M. Paley and P. D. Karp. Evaluation of comoputational metabolic pathway predictions for *helicobacter pylori*. *Bioinformatics*, 18(5):705–714, 2002.

[60] J. Pandey, M. Koyuturk, and A. Grama. Functional characterization and topological modularity of molecular interaction networks. *BMC Bioinformatics*, 11(Suppl 1):S35, 2010.

[61] P. Pei and A. Zhang. A topological measurement for weighted protein interaction network. In *Proceedings of 4th International Computational Systems Bioinformatics Conference*, pages 268–278, Stanford, CA, 2005.

[62] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7):1403–1407, 2008.

[63] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, A. B. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database - 2009 update. *Nucleic Acids Research*, 37:D767–D772, 2009.

[64] N. Przulj and D. Wigle. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348, 2003.

[65] J. Qiu and W. S. Noble. Predicting co-complex protein pairs from heterogeneous data. *PLoS Computational Biology*, 4(4):e1000054, 2008.

[66] K. Raman and N. Chandra. *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. *BMC Microbiol.*, 8:234, 2008.

[67] N. Salomonis, K. Hanspers A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. M. Stuart, B. R. Conklin, and A. R. Pico. GenMAPP 2: New features and resources for pathway analysis. *BMC Bioinformatics*, 8:217, 2007.

[68] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database Issue):D449–D451, 2004.

[69] M. P. Samanta and S. Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA*, 100(22):12579–12583, 2003.

[70] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, 2000.

[71] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[72] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(8):1–13, 2007.

[73] A. Y. Sivachenko, A. Yuryev, N. Daraselia, and I. Mazo. Molecular networks in microarray analysis. *Journal of Bioinformatics and Computational Biology*, 5(2b):429–546, 2007.

[74] D. Soh. *Understanding Pathways*. PhD thesis, Department of Computing, Imperial College London, 180 Queens Gate, London SW7 2AZ, 2010.

[75] D. Soh, D. Dong, Y. Guo, and L. Wong. Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments. *ACM SIGKDD Explorations*, 9(1):3–14, 2007.

[76] D. Soh, D. Dong, Y. Guo, and L. Wong. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics*, 11:449, 2010.

[77] F. Sohler, D. Hanisch, and R. Zimmer. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521, 2004.

[78] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003.

[79] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA*, 102(43):15545–15550, 2005.

[80] Y. V. Sun and S. L. R. Kardia. Identification of epistatic effects using a protein-protein interaction database. *Human Molecular Genetics*, 19(22):4345–4352, 2010.

[81] E. Tsiporkova and V. Boeva. Two-pass imputation algorithm for missing value estimation in gene expression time series. *Journal of Bioinformatics and Computational Biology*, 5(5):1005–1022, 2007.

[82] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116–5121, 2001.

[83] I. Ulitsky and R. Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1:8, 2007.

[84] M. P. van Iersel, T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin, and C. Evelo. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, 9:399, 2008.

[85] C. von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[86] L. Wong. Technologies for integrating biological data. *Briefings in Bioinformatics*, 3(4):389–404, 2002.

[87] L. Wong and G. Liu. Protein interactome analysis for countering pathogen drug resistance. *Journal of Computer Science and Technology*, 25(1):124–130, 2010.

[88] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, et al. newblock The genomic landscapes of human breast and colorectal cancers. *Science*, 318:1108–1113, 2007.

[89] X. Wu, R. Jiang, M. Q. Zhang, and S. Li. Network-based global inference of human disease genes. *Molecular Systems Biology*, 4:189, 2008.

[90] Y. Wu and S. Lonardi. A linear-time algorithm for predicting functional annotations from PPI networks. *Journal of Bioinformatics and Computational Biology*, 6(6):1049–1065, 2008.

[91] S.H. Yook, Z. Oltvai, and A.-L. Barabasi. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942, 2004.

[92] Z.-H. You, Y.-K. Lei, J. Gui, D.-S. Huang, and X. Zhou. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, 26(21):2744–2751, 2010.

[93] L. Yu, L. Gao, and K. Li. A method based on local density and random walks for complex detection in protein interaction networks. *Journal of Bioinformatics and Computational Biology*, 8(Suppl. 1):47–62, 2010.

[94] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, et al. GoMiner: A resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, 2003.

[95] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang, C. Wang, and Z. Guo. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662–1668, 2009.

[96] Y. Zhao and G. Wang. Additive risk analysis of microarray gene expression data via correlation principal component regression. *Journal of Bioinformatics and Computational Biology*, 8(4):645–659, 2010.