Internet Mathematics Vol. 7, No. 4: 274–298

Using Biological Networks in Protein Function Prediction and Gene Expression Analysis

Limsoon Wong

Abstract. While sequence homology search has been the main workhorse in protein function prediction, it is not applicable to a significant portion of novel proteins that do not have informative homologues 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 of biological network databases and 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

ISSN: 1542-7951 print

with 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 that 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 [Wu and Lonardi 08, Sharan et al. 07], protein complexes [Chua et al. 08, King et al. 04, Yu 10], and functional modules [Gao et al. 09, Enright et al. 02, Ulitsky and Shamir 07]. In particular, there are three main groups of approaches for prediction of protein function using protein interaction networks. The first group [Schwikowski 00, Hishigaki et al. 01, Deng 03, Chua et al. 07b] predicts the function of an unknown protein based on what functions are overrepresented among its direct and indirect interaction partners. The second group [Milenkovic and Przulj 08, Chen et al. 07, Kirac and Ozsoyoglu 08, Bogdanov and Singh 10] predicts the function of an unknown protein by assigning to it the function of a protein whose neighbors in the protein interactome have functions that are most similar to the neighbors of this unknown protein. The third group of approaches [Brun et al. 03, Samanta and Liang 03, Milenkovic and Przulj 08] clusters proteins based on the similarity of their neighborhoods in the protein interaction network and assume that proteins in the same cluster should have similar functions. All three groups of approaches have their basis in the fact that proteins interact to perform their respective functions, and therefore, the function of a protein should correlate with the functions of neighboring proteins in the protein interaction network [Pandey et al. 10, Yook et al. 04].

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., nonbiological variations) [Leek et al. 10], how to handle missing

values [Tsiporkova and Boeva 07], and most importantly, how to identify genes that are meaningful in explaining differences in disease phenotypes [Soh et al. 07]. 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 consists of the overlap analysis methods [Doniger et al. 03, Zeeberg et al. 03, Khatri and Draghici 05], which test the significance of the intersection of differentially expressed genes with a biological pathway. The second group consists of the direct group analysis methods [Goeman et al. 04, Kim and Volsky 05, Subramanian et al. 05], which test whether a biological pathway is differentially expressed as a whole. The third group consists of the network-based analysis methods [Sohler et al. 04, Sivachenko et al. 07, Chuang et al. 07], which zoom into a subnetwork of a biological pathway and test whether the subnetwork is differentially expressed. All of these approaches have their basis in the fact that every disease phenotype has some underlying biological causes. Therefore, it is reasonable to analyze 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 [Kanehisa et al. 10, Karp et al. 05, Pico et al. 08, Joshi-Tope et al. 05, Soh et al. 10] 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 [Chatraryamontri et al. 07, Breitkreutz et al. 08, Salwinski et al. 04, Prasad et al. 09, Jensen 09] 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 [Chua et al. 07a] and disease–gene association studies [Ideker and Sharan 08, Matsunaga et al. 10]. 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 [Schwikowski 00, Hishigaki et al. 01, Brun et al. 03, Samanta and Liang 03, Deng 03, Chua et al. 07b, Chen et al. 07, Kirac and Ozsoyoglu 08, Bogdanov and Singh 10, Milenkovic and Przulj 08] for prediction of protein function using biological networks are presented. After that, in Section 5, the three groups of approaches [Doniger et al. 03, Zeeberg et al. 03, Khatri and Draghici 05, Goeman et al. 04, Kim and Volsky 05, Subramanian et al. 05, Sohler et al. 04, Sivachenko et al. 07, Chuang et al. 07] for improving the reliability of gene selection using biological networks are described. Finally, in Section 6,

we briefly discuss some other uses to which biological network data have been put.

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 [Kanehisa et al. 10], 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 [Karp et al. 05], 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. Those in tier 1 are curated manually. Tier 2 databases are generated on the basis of reviewed predictions by the Pathologic software [Paley and Karp 02]. Tier 3 databases are generated on the basis of 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 [Pico et al. 08], 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 [Joshi-Tope et al. 05], accessible at http://www.reactome.org, is also curated manually by a community of collaborating labs. It contains a

[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.

total of 13,197 pathways from 21 organisms, including 1,112 human pathways. 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 be accessed only 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 [Soh et al. 10], accessible at http://www.pathwayapi.com, is a 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.

Since these biological pathway databases are generally curated manually, their content can be regarded as reliable. However, it is important that one be aware of the following two issues before using these databases. Firstly, many biological pathways are curated only in some databases and not in others. 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 regarding that pathway. For example, a recent study [Soh et al. 10] shows that for a pathway that is as pervasive as that for human apoptosis, 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 [Green 06]. 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 [Wong 02], 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 [Elliott et al. 08] 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. That is, 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 [Salomonis et al. 07], Cytoscape [Shannon et al. 03], and PathVisio [van Iersel et al. 08] 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 [Soh et al. 10] can be considered as taking the "full unification" approach. In this approach, pathways in different 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: 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 as to 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, since it was feared that different databases would give very different names to the same pathway.

Unfortunately, matching pathways by largest overlap of interactions requires a threshold on the overlap. With too small a threshold, we get a large number of false-positive matches, while too large a threshold leads to 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 [Soh et al. 10].

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 [Chatraryamontri et al. 07], accessible at http://mint.bio.uniroma2.it/mint, contains 90,695 physical protein interactions curated from the literature.
- BioGRID [Breitkreutz et al. 08], 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 [Salwinski et al. 04], 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 [Prasad et al. 09], accessible at http://www.hprd.org, contains about 40,000 protein interactions curated from the literature. It focuses on human protein interactions.
- STRING [Jensen 09], 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., interaction or no interaction. 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 [Ng and Tan 04]. Some of these experimental methods, e.g., two-hybrids, are highly susceptible to noise and may have high false-positive rates [von Mering et al. 02, Sprinzak et al. 03].

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 [Gavin et al. 06]. 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 [Chua et al. 07b] and identifying protein complexes [Liu et al. 09].

For some analysis, it is crucial to use a subset of protein interaction data that are more reliable. Consequently, much effort has been invested in developing solutions to this problem [Chua and Wong 08]. 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, since a pair of interacting proteins are generally expected to be localized to the same cellular component or to have a common cellular role [Sprinzak et al. 03, Nabieva et al. 05]. The main shortcoming of this approach is that protein functional annotations and cellular localization information are often incomplete. Moreover, 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 [Chua et al. 06, Hart et al. 07]. 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.

Since the additional information required by these approaches is 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 [Chen et al. 06, Chua and Wong 08]. One of the most important early examples of this idea is the Czekanowski–Dice distance [Brun et al. 03], 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 sets 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 the value of $\mathrm{CD}_{u,v}$ incorporating this information is

$$\mathrm{CD}_{u,v} = \frac{2\sum_{w\in N_{u,v}} r_{u,w}r_{w,v}}{\sum_{w\in N_u} r_{u,w} + \sum_{w\in N_v} r_{w,v}}.$$

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 neighbors. 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

$$\mathrm{CD}^{i+1}_{u,v} = \frac{2\sum_{w\in N_{u,v}} \mathrm{CD}^i_{u,w}\,\mathrm{CD}^i_{w,v}}{\sum_{w\in N_u} \mathrm{CD}^i_{u,w} + \sum_{w\in N_v} \mathrm{CD}^i_{w,v} + \lambda},$$

and $\mathrm{CD}^0_{u,v} = r_{u,v}$.

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 improve $\mathrm{CD}_{u,v}$ and other related topology-based reliability indices for protein interactions significantly. In particular, a recent study [Liu et al. 08] used the DIP yeast data set for assessment and showed that 54.7% of the interacting protein pairs reported in DIP are co-localized. Since proteins in general can interact only when they are co-localized, this suggests a level of noise in the data set of up to 45.3%. After these pairs were ranked using the iterated version of $\mathrm{CD}_{u,v}$, about 90% of the top 30% of interacting pairs were determined to be 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 [Chua and Wong 08]. A recent idea [Kuchaiev et al. 09, You et al. 10] to overcome this problem is to use a larger interaction neighborhood 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 multidimensional 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 such as $\mathrm{CD}_{u,v}$ 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 [You et al. 10].

Besides noise dealt with by the approaches mentioned 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 [Chua and Wong 08], including gene-fusion events [Marcotte et al. 99], interacting domains [Han et al. 04], interacting motifs [Li et al. 06], co-evolution of proteins or residues [Juan et al. 08], topology of protein–protein interaction networks [Pei and Zhang 05], and machine learning from multiple information types [Qiu and Noble 08]. Incidentally, it is possible to use topology-based indices such as $\mathrm{CD}_{u,v}$ for predicting new interactions [Wong and Liu 10]—one can predict that two proteins u and v 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 [Hawkins and Kihara 07, Koh et al. 09]. Sequence similarity search methods such as BLAST [Altschul et al. 90] are the primary tools for this problem. However, a nonnegligible proportion of protein sequences do not have identifiable informative homologues 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 [Hawkins and Kihara 07].

In particular, many approaches have been proposed to use protein interaction networks for protein function prediction [Sharan et al. 07]. These approaches can be roughly divided into three groups.

The first group [Schwikowski 00, Hishigaki et al. 01, Deng 03, Chua et al. 07b] 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 times more likely to interact than a random pair of proteins [Liu et al. 08].

The second group [Chen et al. 07, Kirac and Ozsoyoglu 08, Bogdanov and Singh 10] is based on the hypothesis that proteins with similar function have interaction neighborhoods that are similar. This is also a reasonable hypothesis, because when the proteins in the neighborhood of a protein have similar functions to those of the proteins in the neighborhood of another protein, the two proteins are likely to operate in similar environments and have similar properties.

The third group [Samanta and Liang 03, Milenkovic and Przulj 08, Brun et al. 03] clusters proteins based on similarity of certain features—in particular their neighborhood 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 a function that is overrepresented among its interaction partners [Schwikowski 00, Hishigaki et al. 01]. Another example is to apply global optimization techniques—e.g., Markov random fields—to transfer the function of a protein from its neighbor and also to propagate predictions so that the function of proteins without characterized neighbors can be predicted [Deng 03].

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 [Chua et al. 06]. 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 an 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 [Chua et al. 06]—overcomes these shortcomings by weighted voting of both direct and indirect neighbors. This method defines the functional similarity weight $S_{\mathrm{FS}}(u, v)$ between two proteins u and v based on the size of the intersection of their interaction neighborhoods. The weight $S_{\mathrm{FS}}(u, v)$ is a variation of $\mathrm{CD}_{u,v}$ in which the size of the intersection is defined with the reliability of individual interactions taken into account and with equal weight given to the interaction neighborhoods of u and v.

Then a direct neighbor v of a protein u having function a votes for function a with weight $S_{\mathrm{FS}}(u, v)$. Similarly, an indirect neighbor v' of a protein u having function a votes for function a with weight $S_{\mathrm{FS}}(u, v')$. The function a that receives a total number 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 [Chua et al. 07b] based on the BioGrid yeast protein interaction network, out of about 100 biological processes considered, FSweight achieved an 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 a substantial number of high-quality predictions that cannot be inferred from sequence homology [Chua et al. 07b].

An early example of the second group is LaMoFinder [Chen et al. 07]. It first discovers network motifs [Alon 07] 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 labeled 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 labeled neighborhood are aligned to such a network motif, its function can be inferred from the vertex to which this protein is mapped in the network motif.

A limitation of LaMoFinder 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 labeled subgraph rooted at the unknown protein to functionally labeled subgraphs rooted at other nodes in the protein interaction network [Kirac and Ozsoyoglu 08].

Both LaMoFinder and this refinement rely on a topological matching of subnetworks. Thus their performance is affected in less-reliable protein interaction networks that have more false interactions and missing interactions. A recent idea [Bogdanov and Singh 10] to overcome this shortcoming uses a probabilistic 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 that are its nearest neighbors in terms of functional profile. It has been shown [Bogdanov and Singh 10] 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 [Brun et al. 03] and a closely related method [Samanta and Liang 03]. 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 [Brun et al. 03, Samanta and Liang 03] share the same hypothesis as the first group of methods [Schwikowski 00, Hishigaki et al. 01]. 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 [Brun et al. 03, Samanta and Liang 03] uses the hypothesis in a stronger way than the majority-vote type of methods [Schwikowski 00, Hishigaki et al. 01], because as implied by the clustering step, the former further requires that the interaction

partners from which the votes are taken be also interaction partners of each other.

A more recent idea of the third group [Milenkovic and Przulj 08] first clusters proteins based on the similarity of the network motifs (called graphlets by the authors) in which they participate. A protein of unknown function can then be assigned the same function as other annotated proteins in the same cluster. Interestingly, such a method [Milenkovic and Przulj 08] shares the same hypothesis as the second group of methods typified by LaMoFinder [Chen et al. 07]. 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 [Milenkovic and Przulj 08] uses the hypothesis in a stronger way than LaMoFinder, 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 a generalization of—the second group of methods. This is the idea of comparing and aligning the biological networks of two different species [Milenkovic et al. 10, Kelley 04]. Here, after the network alignment is performed, the alignment can be analyzed and used to infer protein functions based on shared topology in the aligned (sub)networks of the two species [Kuchaiev et al. 10].

5. Microarray Analysis Using Biological Networks

Many approaches [Tusher et al. 01, Liu et al. 10, Zhao and Wang 10, Li et al. 03b, Li et al. 03a, Liu et al. 05] have been proposed for the inference of differentially expressed genes that are useful in the diagnosis of diseases and prognosis of treatment responses. 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 [Zhang et al. 09]. Furthermore, the transition from the selected genes to an understanding of the sequences of causative molecular events is unclear [Soh et al. 07].

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 exist a number of approaches to analyze gene expression data with respect to biological context. These approaches can be roughly divided into three groups [Soh et al. 07]. The first group comprises the overlap analysis methods [Doniger et al. 03, Zeeberg et al. 03, Khatri and Draghici 05]. The second group comprises the direct group analysis methods [Goeman et al. 04, Kim and Volsky 05, Subramanian et al. 05]. The third group comprises the network-based analysis methods [Sohler et al. 04, Sivachenko et al. 07, Chuang et al. 07].

The overlap analysis methods [Doniger et al. 03, Zeeberg et al. 03, Khatri and Draghici 05] 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 [Khatri and Draghici 05] 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 a situation, since A is not suggested by these methods as being 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 [Goeman et al. 04, Kim and Volsky 05, Subramanian et al. 05] 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 [Goeman et al. 04] and GSEA [Subramanian et al. 05] 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 predominance of the 30 other nondifferentially 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 [Sohler et al. 04, Sivachenko et al. 07, Chuang et al. 07, Soh et al. 11] 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 [Sivachenko et al. 07]. 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 a consistent differential expression pattern, leading to a more precise hypothesis that explains the disease phenotype. A shortcoming of NEA is that it considers only the immediate regulator–regulated 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 [Soh et al. 11], 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-statistic 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-statistic is compared to a null distribution obtained by permuting 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 across independent data sets of the same disease phenotypes than other methods; in particular, it was tested on four diseases [Soh et al. 11]. 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 fewer 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

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

Protein complex discovery. Proteins often perform a function by aggregating into complexes to perform sophisticated biological tasks. This has motivated approaches to identify protein complexes computationally from protein–protein interaction 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 [Enright et al. 02, Przulj and Wigle 03, Adamcsek et al. 06, Chua et al. 08, Liu et al. 09, Yu 10]. However, these algorithms have relatively 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 [Wong and Liu 10].

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 insight into their routes to drug resistance [Raman and Chandra 08, Wong and Liu 10]. A pioneering idea [Raman

and Chandra 08] 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 cytochrome-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 [Hormozdian et al. 10, Wong and Liu 10] 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. Since many diseases have complex underlying mechanisms, analysis at the level of a single SNP (single-nucleotide polymorphism) is insufficient. There is thus intense interest in exploring the interactions of multiple SNPs—the so-called epistatic interactions—to uncover more significant associations [Cordell 02]. Exhaustively considering all the possible SNP combinations is infeasible. 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 [Sun and Kardia 10].

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 opened new avenues for inferring 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 [Wood et al. 07, Ideker and Sharan 08]. A recent idea [Wu et al. 08, Li and Patra 10] 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 [Matsunaga et al. 10] instead of protein interaction networks.

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

Acknowledgments. 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

[Adamcsek et al. 06] B. Adamcsek et al. "CFinder: Locating Cliques and Overlapping Modules in Biological Networks." *Bioinformatics* 22:8 (2006), 1021–1023.

[Alon 07] U. Alon. "Network Motifs: Theory and Experimental Approaches." *Nature Reviews Genetics* 8:6 (2007), 450–461.

[Altschul et al. 90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (1990), 403–410.

[Bogdanov and Singh 10] P. Bogdanov and A. Singh. "Molecular Function Prediction Using Neighborhood Features." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7:2 (2010), 208–217.

[Breitkreutz et al. 08] B. J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, et al. "The BioGRID Interaction Database: 2008 Update." *Nucleic Acids Research* 36 (Database Issue) (2008), D637–D640.

[Brun et al. 03] 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 (2003), R6.

[Chatraryamontri et al. 07] A. Chatraryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. "MINT: The Molecular INTeraction Database." *Nucleic Acids Research* 35 (2007), D572–D574.

[Chen et al. 06] 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*, pp. 284–297, Yokohama, Japan, 2006.

[Chen et al. 07] 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*, pp. 546–555, Istanbul, Turkey, 2007.

[Chua and Wong 08] H. N. Chua and L. Wong. "Increasing the Reliability of Protein Interactomes." *Drug Discovery Today* 13: 15/16 (2008), 652–658.

[Chua et al. 06] H. N. Chua, W.-K. Sung, and L. Wong. "Exploiting Indirect Neighbors and Topological Weight to Predict Protein Function from Protein–Protein Interactions." *Bioinformatics* 22:13 (2006), 1623–1630.

[Chua et al. 07a] 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 (2007), 3364–3373.

[Chua et al. 07b] 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) (2007), S8.

[Chua et al. 08] 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 (2008), 435–466, 2008.

[Chuang et al. 07] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. "Network-Based Classification of Breast Cancer Metastasis." *Molecular Systems Biology* 3 (2007), 140.

[Cordell 02] 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 (2002), 2463–2468.

[Deng 03] 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 (2003), 947–960.

[Doniger et al. 03] 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* 4:11 (2003), R7.

[Elliott et al. 08] B. Elliott, M. Kirac, A. Cakmak, G. Yavas, S. Mayes, et al. "PathCase: Pathways Database System." *Bioinformatics* 24:21 (2008), 2526–2533.

[Enright et al. 02] 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 (2002), 1575–1584.

[Gao et al. 09] L. Gao, P. G. Sun, and J. Song. "Clustering Algorithms for Detecting Functional Modules in Protein Interaction Networks." *Journal of Bioinformatics and Computational Biology* 7L1 (009), 217–242.

[Gavin et al. 06] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, et al. "Proteome Survey Reveals Modularity of the Yeast Cell Machinery." *Nature* 440:7084 (2006), 631–636.

[Goeman et al. 04] 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 (2004), 93–99.

[Green 06] M. L. K. P. Green. "The Outcomes of Pathway Database Computations Depend on Pathway Ontology." *Nucleic Acids Research* 34:13 (2006), 3687–3697.

[Han et al. 04] 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 (2004), 6312–6320.

[Hart et al. 07] 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 (2007), 236.

[Hawkins and Kihara 07] T. Hawkins and D. Kihara. "Functional Prediction of Uncharacterized Proteins." *Journal of Bioinformatics and Computational Biology* 5:1 (2007), 1–30.

[Hishigaki et al. 01] 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 (2001), 525–531.

[Hormozdian et al. 10] 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 (2010), 669–684.

[Ideker and Sharan 08] T. Ideker and R. Sharan. "Protein Networks in Disease." *Genome Research* 18 (2008), 644–652.

[Jensen 09] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, et al. "STRING 8—A Global View on Proteins and Their Functional Interactions in 630 Organisms." *Nucleic Acids Research* 37 (Database Issue) (2009), D412–D416.

[Joshi-Tope et al. 05] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, et al. "Reactome: A Knowledgebase of Biological Pathways." *Nucleic Acids Research* 33 (Database Issue) (2005), D428–D432.

[Juan et al. 08] 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 (2008), 934–939.

[Kanehisa et al. 10] 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) (2010), D355–D360.

[Karp et al. 05] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, et al. "Expansion of the BioCyc Collection of Pathway/Genome Databases to 160 Genomes." *Nucleic Acids Research* 19 (2005), 6083–6089.

[Kelley 04] 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) (2004), W83–W88.

[Khatri and Draghici 05] P. Khatri and S. Draghici. "Ontological Analysis of Gene Wxpression Data: Current Tools, Limitations, and Open Problems." *Bioinformatics* 21:18 (2005), 3587–3595.

[Kim and Volsky 05] S. Y. Kim and D. J. Volsky. "PAGE: Parametric Analysis of Gene Set Enrichment." *BMC Bioinformatics* 8:6 (2005), 144.

[King et al. 04] A. D. King, N. Przulj, and I. Jurisica. "Protein Complex Prediction via Cost-Based Clustering." *Bioinformatics* 20:17 (2004), 3013–3020.

[Kirac and Ozsoyoglu 08] 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*, pp. 197–213, Singapore, 2008.

[Koh et al. 09] 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 (2009), 973–990.

[Kuchaiev et al. 09] O. Kuchaiev, M. Rasajski, D. J. Highham, and N. Przulj. "Geometric De-nosing of Protein–Protein Interaction Networks." *PLoS Computational Biology* 5:8 (2009), e1000454.

[Kuchaiev et al. 10] 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 (2010), 1341–1354.

[Leek et al. 10] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, et al. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." *Nature Reviews Genetics* 11 (2010), 733–739.

[Li and Patra 10] Y. Li and J. C. Patra. "Genome-Wide Inferring Gene-Phenotype Relationship by Walking on the Heterogeneous Network." *Bioinformatics* 26:9 (2010), 1219–1224.

[Li et al. 03a] 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 (2003), 71–78.

[Li et al. 03b] 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*, pp. 17–24, Washington, DC, 2003.

[Li et al. 06] H. Li, J. Li, and L. Wong. "Discovering Motif Pairs at Interaction Sites from Sequences on a Proteome-Wide Scale." *Bioinformatics* 22:8 (2006), 989–996.

[Liu et al. 05] H. Liu, J. Li, and L. Wong. "Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data." *Bioinformatics* 21:16 (2005), 3377–3384.

[Liu et al. 08] 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)*, pp. 138–149, 2008.

[Liu et al. 09] G. Liu, L. Wong, and H. N. Chua. "Complex Discovery from Weighted PPI Networks." *Bioinformatics* 25:15 (2009), 1891–1897

[Liu et al. 10] Z. Liu et al. "A Multi-strategy Approach to Informative Gene Identification from Gene Expression Data." *Journal of Bioinformatics and Computational Biology* 8:1 (2010), 19–38.

[Marcotte et al. 99] 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 (1999), 751–753.

[Matsunaga et al. 10] T. Matsunaga, S. Kuwata, and M. Muramatsu. "Computational Gene Knockout Reveals Transdisease–Transgene Association Structure." *Journal of Bioinformatics and Computational Biology* 8:5 (2010), 843–866.

[Milenkovic and Przulj 08] T. Milenkovic and N. Przulj. "Uncovering Biological Network Function via Graphlet Degree Signatures." *Cancer Informatics* 6 (2008), 257–273.

[Milenkovic et al. 10] T. Milenkovic, W. L. Ng, W. Hayes, and N. Przulj. "Optimal Network Alignment with Graphlet Degree Vectors." *Cancer Informatics* 9 (2010), 121–137.

[Nabieva et al. 05] 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) (2005), i302–i310.

[Ng and Tan 04] S.-K. Ng and S.-H. Tan. "Discovering Protein–Protein Interactions." *Journal of Bioinformatics and Computational Biology* 1:4 (2004), 711–741.

[Paley and Karp 02] S. M. Paley and P. D. Karp. "Evaluation of Computational Metabolic Pathway Predictions for *Helicobacter pylori*." *Bioinformatics* 18:5 (2002), 705–714.

[Pandey et al. 10] J. Pandey, M. Koyuturk, and A. Grama. "Functional Characterization and Topological Modularity of Molecular Interaction Networks." *BMC Bioinformatics* 11 (Suppl. 1) (2010), S35.

[Pei and Zhang 05] P. Pei and A. Zhang. "A Topological Measurement for Weighted Protein Interaction Network." In *Proceedings of 4th International Computational Systems Bioinformatics Conference*, pp. 268–278, Stanford, CA, 2005.

[Pico et al. 08] 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 (2008), 1403–1407.

[Prasad et al. 09] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, et al. "Human Protein Reference Database: 2009 Update." *Nucleic Acids Research* 37 (2009), D767–D772.

[Przulj and Wigle 03] N. Przulj and D. Wigle. "Functional Topology in a Network of Protein Interactions." *Bioinformatics* 20:3 (2003), 340–348.

[Qiu and Noble 08] J. Qiu and W. S. Noble. "Predicting Co-complex Protein Pairs from Heterogeneous Data." *PLoS Computational Biology* 4:4 (2008), e1000054.

[Raman and Chandra 08] K. Raman and N. Chandra. "*Mycobacterium tuberculosis* Interactome Analysis Unravels Potential Pathways to Drug Resistance." *BMC Microbiol.*, 8:234, 2008.

[Salomonis et al. 07] N. Salomonis, K. Hanspers A. C. Zambon, K. Vranizan, S. C. Lawlor, et al. "GenMAPP 2: New Features and Resources for Pathway Analysis." *BMC Bioinformatics* 8 (2007), 217.

[Salwinski et al. 04] 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) (2004), D449–D451.

[Samanta and Liang 03] M. P. Samanta and S. Liang. "Predicting Protein Functions from Redundancies in Large-Scale Protein Interaction Networks." *Proc. Natl. Acad. Sci. USA* 100:22 (2003), 12579–12583.

[Schwikowski 00] B. Schwikowski, P. Uetz, and S. Fields. "A Network of Protein–Protein Interactions in Yeast." *Nature Biotechnology* 18:12 (2000), 1257–1261.

[Shannon et al. 03] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13:11 (2003), 2498–2504.

[Sharan et al. 07] R. Sharan, I. Ulitsky, and R. Shamir. "Network-Based Prediction of Protein Function." *Molecular Systems Biology* 3:8 (2007), 1–13.

[Sivachenko et al. 07] A. Y. Sivachenko, A. Yuryev, N. Daraselia, and I. Mazo. "Molecular Networks in Microarray Analysis." *Journal of Bioinformatics and Computational Biology* 5:2b (2007), 429–546.

[Soh et al. 07] 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 (2007), 3–14.

[Soh et al. 10] D. Soh, D. Dong, Y. Guo, and L. Wong. "Consistency, Comprehensiveness, and Compatibility of Pathway Databases." *BMC Bioinformatics* 11 (2010), 449.

[Soh et al. 11] D. Soh, D. Dong, Y. Guo, and L. Wong. "Finding Consistent Disease Subnetworks Across Microarray Datasets." *BMC Bioinformatics* 12(Suppl. 13) (2011), S15.

[Sohler et al. 04] F. Sohler, D. Hanisch, and R. Zimmer. "New Methods for Joint Analysis of Biological Networks and Expression Data." *Bioinformatics* 20:10 (2004), 1517–1521.

[Sprinzak et al. 03] E. Sprinzak, S. Sattath, and H. Margalit. "How Reliable Are Experimental Protein–Protein Interaction Data?" *Journal of Molecular Biology* 327:5 (2003), 919–923.

[Subramanian et al. 05] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, et al. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proc. Nat. Acad. Sci. USA* 102:43 (2005), 15545–15550.

[Sun and Kardia 10] Y. V. Sun and S. L. R. Kardia. "Identification of Epistatic Effects Using a Protein–Protein Interaction Database." *Human Molecular Genetics* 19:22 (2010), 4345–4352.

[Tsiporkova and Boeva 07] 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 (2007), 1005–1022.

[Tusher et al. 01] 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 (2001), 5116–5121.

[Ulitsky and Shamir 07] I. Ulitsky and R. Shamir. "Identification of Functional Modules Using Network Topology and High-Throughput Data." *BMC Systems Biology* 1 (2007), 8.

[van Iersel et al. 08] 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 (2008), 399.

[von Mering et al. 02] C. von Mering, R. Krause, B. Snel, et al. "Comparative Assessment of Large-Scale Data Sets of Protein–Protein Interactions." *Nature* 417:6887 (2002), 399–403.

[Wong 02] L. Wong. "Technologies for Integrating Biological Data." *Briefings in Bioinformatics* 3:4 (2002), 389–404.

[Wong and Liu 10] L. Wong and G. Liu. "Protein interactome Analysis for Countering Pathogen Drug Resistance." *Journal of Computer Science and Technology* 25:1 (2010), 124–130.

[Wood et al. 07] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, et al. "The Genomic Landscapes of Human Breast and Colorectal Cancers." *Science* 318 (2007), 1108–1113.

[Wu and Lonardi 08] Y. Wu and S. Lonardi. "A Linear-Time Algorithm for Predicting Functional Annotations from PPI Networks." *Journal of Bioinformatics and Computational Biology* 6:6 (2008), 1049–1065.

[Wu et al. 08] X. Wu, R. Jiang, M. Q. Zhang, and S. Li. "Network-Based Global Inference of Human Disease Genes." *Molecular Systems Biology* 4 (2008), 189.

[Yook et al. 04] S.H. Yook, Z. Oltvai, and A.-L. Barabasi. "Functional and Topological Characterization of Protein Interaction Networks." *Proteomics* 4 (2004), 928–942.

[You et al. 10] 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 (2010), 2744–2751.

[Yu 10] 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) (2010), 47–62.

[Zeeberg et al. 03] 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 (2003), R28.

[Zhang et al. 09] 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 (2009), 1662–1668.

[Zhao and Wang 10] 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 (2010), 645–659.

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

Received November 23, 2010; accepted March 9, 2011.