CHAPTER 11

Cancer Gene Prediction Using a Network Approach

Xuebing Wu and Shao Li

CONTENTS

11.1 INTRODUCTION

Cancer is a genetic disease (Vogelstein and Kinzler 2004). Decades of research in molecular genetics have identified a number of important genes responsible for the genesis of various types of cancer (Futreal et al. 2004) and drugs targeting these mutated cancer genes have brought dramatic therapeutic advances and substantially improve and prolong the lives of cancer patients (Huang and Harari 1999). However, cancer is extremely complex and heterogeneous. It has been suggested that 5% to 10% of the human genes probably contributed to oncogenesis (Strausberg, Simpson, and Wooster 2003), while current experimentally validated cancer genes only cover 1% of the human genome (Futreal et al. 2004), suggesting

that there are still hundreds or even thousands of cancer genes that remain to be identified. For example, in breast cancer, known susceptibility genes, including BRCA1 (Miki et al. 1994) and BRCA2 (Wooster et al. 1995), can only explain less than 5% of the total breast cancer incidence and less than 25% of the familial risk (Oldenburg et al. 2007). The same challenge is also faced by other types of cancer and other complex diseases, such as diabetes (Frayling 2007) and many brain diseases (Burmeister, McInnis, and Zollner 2008; Folstein and Rosen-Sheidley 2001). There is a long way to go from changes in genetic sequence to visible clinical phenotypes. The complex molecular interaction networks, together with environmental factors, further lower the penetrance of a single causal gene and complicate the relationship between genes and diseases. This high complexity and low penetrance might explain why so many disease genes remain unidentified.

Traditional gene mapping approaches, such as linkage analysis and association studies, have limited resolution to localize the causal genes in the genome, and the resultant region often contains hundreds of candidate genes (Altshuler, Daly, and Lander 2008). The functional testing and validation of causative genes are time consuming and laborious. The priority of candidate genes is usually determined by expert judgment based on the gene's known functions (Pharoah et al. 2007), which are often biased and limited by the scope of the expert. Alternatively, with the increasing availability of genome-wide sequence, genomics, proteomics, and epigenomics data, computational methods are exploited to predict and prioritize disease genes (Oti and Brunner 2007; Zhu, Gerstein, and Snyder 2007), significantly reducing the number of candidate genes for further testing. Computational prediction and prioritization is complementary to genetic mapping, in terms of integrating existing knowledge on disease biology and relatively unbiased whole genome measurements.

More recently, large-scale molecular interaction network data have become available, and it turns out to be particularly powerful for disease gene prediction when used alone (Kohler et al. 2008; Oti et al. 2006) or combined with other data sources (Karni, Soreq, and Sharan 2009; Lage et al. 2007; Mani et al. 2008; Wu et al. 2008). Molecular interaction networks depict the basic skeleton of cellular processes, and network analysis has the ability to model the complex interactions among multiple genes and their higher-level organizations (Barabasi and Oltvai 2004; Han 2008; Zhu, Gerstein, and Snyder 2007). In this chapter, we will focus on network-based approaches for cancer gene prediction. Many of the methods discussed here are designed for general disease instead of cancer. Nonetheless, they can be applied to predict cancer genes as a special case, and most of these network-based methods have been demonstrated by applying them to various types of cancer.

11.2 MOLECULAR NETWORKS AND HUMAN DISEASES

Before going into the details of network-based gene prioritization methods, we will briefly introduce some basic concepts about molecular networks, the data sources and tools for building networks, and the working principles for network approaches in predicting disease genes.

Network is a simple but efficient abstraction of biological systems (Barabasi and Oltvai 2004). Nodes/vertices in a molecular network represent biomolecules, such as genes, proteins, and metabolites. Edges/links between nodes indicate physical or functional interactions, including transcriptional binding, protein-protein interaction, genetic interaction (such as synthetic

lethal), biochemical reactions, and many others. An edge on a network (if it happens in the cell) shows that two molecules are functionally related with each other, and the distance on a network is correlated with functional similarity (Sharan, Ulitsky, and Shamir 2007). Network/graph theory provides multiple definitions and tools to measure the distance/proximity between two nodes on a network, which makes network analysis particularly suitable to the quantitative modeling of gene-gene and gene-disease relationships (see Box 11.1 for basic graph concepts).

BOX 11.1 BASIC GRAPH CONCEPTS

A **graph** is a pair G(V,E), where V is a set of nodes (or vertices) and E is a set of edges (or links, or interactions) connecting pairs of nodes. On molecular interaction networks, the nodes represent molecules such as genes or proteins, and the edges represent interactions such as protein-protein interaction, transcriptional binding between protein and DNA.

A graph can be represented by an **adjacent matrix** A, where $A_{ij} = 1$ if there is an edge between nodes i and j; otherwise $A_{ij} = 0$.

A **path** from node A to B is a sequence of nodes started with A and ended with B, such that from each of its nodes there is an edge to the next node in the sequence.

The **length** of a path is the number of edges in the path.

The **distance** of two nodes is usually defined as the length of the short path between the nodes. More complex definitions of graph distance are discussed in the main text.

The ***k*th-order** neighbor of a node is the nodes whose distance from it is k.

The **centrality** of a node measures how centrally a node is located in a given graph. Three commonly used centrality measure are degree, betweenness, closeness, and eigenvector centrality.

The **degree** of a node is the number of edges it is connected with.

The **eigenvector centrality** is a weighted version of the degree centrality, such that x_i of node i is proportional to the sum of the centralities of its neighbors:

$$x_i = \lambda^{-1} \sum_{j=1}^{n} A_{ij} x_j$$

Let the vector $x = (x_1, x_2, \ldots, x_n)$ be the centralities of the nodes then we have

$$\lambda x = Ax$$

where x is an eigenvector of the adjacency matrix A with eigenvalue λ. Theoretical results show that there is only one eigenvector x with all centrality values non-negative and this is the unique eigenvector that corresponds to the largest eigenvalue λ. Eigenvector centrality assigns each node a centrality that not only depends on the quantity of its connections, but also on their qualities.

The **closeness** of a node measures the centrality of a node based on how close it is to other nodes in the network. It can be calculated by inverting the sum of the distances from it to other nodes in the network.

The **betweenness** of a node is the number of shortest paths between other nodes that run through the node in interest. Betweenness centrality characterizes the control of a node over the information flow of the network.

Until now, widely used large-scale human gene/protein networks have been generated mainly by four approaches: high throughput technology for large-scale screen of genetic interaction or protein-protein interaction, manual curation of high-quality interaction data from published small-scale experiment results, automatic text mining to extract gene interactions from the published literature, and computational prediction by integrating multiple genomics data. Generally, high-throughput technology such as yeast-2-hybrid (Fields and Song 1989; Fields and Sternglanz 1994) can yield relatively unbiased protein interaction data, but the false positive rate can reach 50% (Sprinzak, Sattath, and Margalit 2003; von Mering et al. 2002). In addition, though the interactome (a full list of interactions) for species like yeast (Ito et al. 2001), worm (S. Li et al. 2004), and fly (Giot et al. 2003) have been extensively mapped using high-throughput technology, data generated in this way for human (Ghavidel, Cagney, and Emili 2005; Rual et al. 2005) composes only a small part of the known human interactome data. On the other hand, the most reliable experimental data comes from manual curation of interaction data reported by traditional small-scale experiments, and most of these data has been included in manually curated databases such as HPRD (Peri et al. 2003), BIND (Bader and Hogue 2003), and BioGRID (Breitkreutz et al. 2008). Occasionally traditional pathway-based databases are also used, including KEGG (Kanehisa and Goto 2000) and Reactome (Vastrik et al. 2007). Despite the intensive effort in mapping the human protein network, the current human interactome is far from complete (Hart, Ramani, and Marcotte 2006). Automatic literature mining techniques have also been developed to identify putative interacting relationships between human genes/proteins described in the published biomedical literature, such as the GENEWAYS system (Rzhetsky et al. 2004). Literature mining also has the advantage that is allows the construction of context-specific networks, such as the prostate cancer specific gene network (Ozgur et al. 2008) and angiogenesis network (S. Li, Wu, and Zhang 2006). In the LMMA (S. Li, Wu, and Zhang 2006) approach, we have also shown that the systematic integration of microarray data significantly refines the literature mined network and yields more biological insights. Finally, multiple computational approaches (Franke et al. 2006; Jansen et al. 2003; Lage et al. 2007; Rhodes et al. 2005; Xia, Dong, and Han 2006) have been developed to predict a comprehensive human interactome map, usually by integrating a number of unbiased genome-wide annotation data, such as sequence, expression, functional annotation, known interaction data, and many others. Among these datasets, homologous mapping is commonly used to transfer protein interactions from other organisms to human by sequence conservation. Typical high-quality interaction databases for other organisms include: BioGrid (Breitkreutz et al. 2008), BIND (Bader and Hogue 2003), MIPS (Mewes et al. 2004), DIP (Salwinski et al. 2004), MINT (Chatr-aryamontri et al. 2007), and IntAct (Kerrien et al. 2007). STRING (von Mering et al. 2005) and OPHID (Brown and Jurisica 2005) are two of the widely used databases hosting predicted interactions.

With all these network data available, studies on model organisms have shown that central positions on the network implicate important roles in cellular processes. For example, in yeast, the number of partners of a gene is positively correlated with lethal phenotypes (Jeong et al. 2001). With the increasing availability of human protein interaction data, network analysis has also shed light on human diseases. For example, consistent with the

observation from yeast, human disease genes tend to have higher network centrality, such as higher degrees, compared to nonessential and nondisease genes (Feldman, Rzhetsky, and Vitkup 2008; Goh et al. 2007; J. Xu and Li 2006), and cancer genes are found to be even more central than other disease genes (Goh et al. 2007; Jonsson et al. 2006). Besides, consistent with the long-held assumption that genes that are closely related are more likely to cause the same or similar diseases, network analysis shows that genes causing the same or similar diseases are likely to interact directly or indirectly with each other (Lim et al. 2006; Oti et al. 2006; Oti and Brunner 2007; van Driel et al. 2006). For example, Lim et al. (2006) show that many ataxia-causing proteins share interacting partners and form a small tightly connected subnetwork. Recent genome-wide cancer mutation screen studies suggest that, though ~80 mutations can be found in a typical cancer, they tend to fall into a few functional pathways (Wood et al. 2007). The functional relatedness of genes causing similar diseases seems to be very general for human diseases, and network analysis provides powerful tools to fully exploit its potential in human disease study. Recently various network-based approaches have emerged to predict disease genes based on the observations described above, generally achieving much better performance than traditional disease gene prediction approaches.

11.3 NETWORK APPROACH FOR CANCER GENE PREDICTION

For clarity we first give the typical settings for a network-based disease gene prediction method (Figure 11.1). Given a list of *N* candidate genes which is assumed to contain at least one disease gene, the goal is to pick out the true disease gene or to rank it at top Mi, where *M* is much smaller than *N*. The candidate genes can be genes within a linkage interval having been associated with the disease under study. Or, if there is no genetic mapping

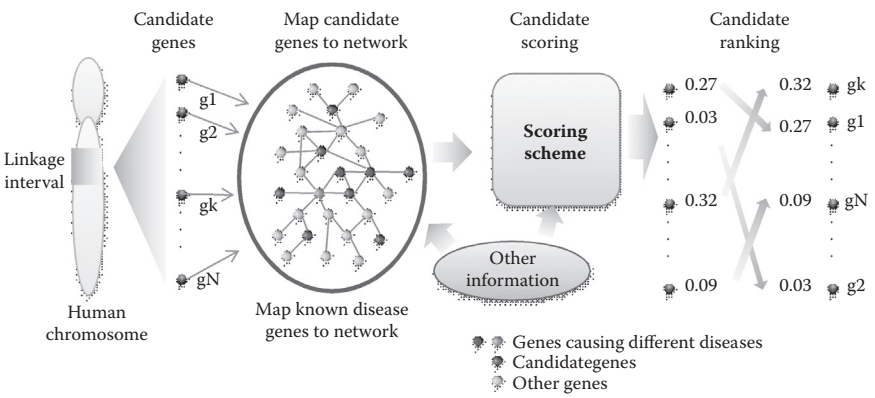

FIGURE 11.1 Sketch map of network-based candidate gene prioritization and prediction. A list of candidate genes such as those in a linkage interval or all the human genes are mapped onto a human gene/protein network, and if applicable, known disease genes and other information (such as sequence characteristics and mRNA expression) are also mapped onto the network. A scoring scheme is used to score each candidate gene based on current data and output a rank list of all candidate genes. Genes ranked above a certain position is predicted as disease-causative.

information, one can simply use the entire human genome as the candidate list. Next, all candidate genes are mapped to a human gene/protein network, the construction of which is described in the previous section. If applicable, known disease genes and other information are mapped to the network too. After that, a scoring scheme scores each candidate gene according to its relative position on the network and additional information. The score is assumed to reflect the probability of the candidate gene to cause the disease under study, given the observed data sources. Finally, all candidate genes are ranked according to the score, and the top 1 or top *M* genes are predicted to be disease causing. The predictability of this score or the performance of the proposed approach is often assessed by cross-validation with known gene-disease relationships (the ability to rediscover known disease genes).

The scoring scheme is the key to a disease gene prediction method. In the following section, we will review different scoring functions used by different methods. To be clearer, we group these methods by the basic principles underlying their scoring schemes (Table 11.1).

11.3.1 Prioritize by Network Proximity

The common principle underlying all methods in this category is "guilt-by-proximity," that is, genes that lie closer to each other on the network are more likely to lead to the same disease. If some genes are already known to be related with the disease under study, then basically one can use the inverse of the distance (proximity) to these disease genes as the score. Otherwise, distance between candidate pairs is used. The methods described below differ in the way they define the distance measure and how the distance is combined with other information to rank candidate genes.

11.3.1.1 Proximity to Known Disease Genes of the Same Disease

Roughly about half of the diseases in the OMIM database (McKusick 2007) have at least one gene known to be involved in the particular disease. For these diseases, the most straightforward way to score and rank candidate genes is to use the proximity to known disease genes as the measure of the disease causing probability. If a candidate is more closely related with a known disease gene, it is more likely to be a disease gene too; therefore, it should get a higher score. If multiple disease genes are already known, then the final score will be the sum of scores across all known disease genes. This procedure can be viewed as a propagation of disease signal: known disease gene serves as the source of disease signal and this signal is propagated along paths on the network to other nodes, and the signal gradually damps as it travels to more distant nodes. Now the problem is how to define distance between two nodes in a network. Three types of distance measure can be found in disease gene finding approaches: direct neighbor, shortest path length, and global distance defined by diffusion kernel or random walk.

11.3.1.1.1 Direct Neighbor In this type of measure, nodes that are directly connected have a distance of 1; otherwise they have a distance of infinity. Approaches employing this measure are actually doing neighbor counting: candidates with more neighbors causing the

TABLE 11.1 A Summary of Network-Based Disease Gene Prediction Methods

Method	Disease Tested	Network Data Sources
Proximity-based		
Direct neighbor		
Oti et al. 2006	General	HPRD, DIP
CPS (George et al. 2006)	General	OPHID
Aragues et al. 2008	Cancer	HPRD, DIP, MIPS, MINT, BioGrid, IntAct
Furney et al. 2008a	Cancer	DIP, MIPS
ENDEAVOUR (Aerts et al. 2006)	General	BIND
Shortest path		
Krauthammer et al. 2004	Alzheimer's disease	Literature mining by GENEWAYS
Liu et al. 2006	Alzheimer's disease	Inferred from multiple dataset
Radivojac et al. 2008	General	HPRD, OPHID
Prioritizer (Franke et al. 2006)	General	Inferred from multiple dataset
Diffusion kernel		
Kohler et al. 2008	General	HPRD, BIND, BioGrid, STRING, DIP, IntAct
Chen et al. 2009	General	HPRD, BIND, BioGrid
Similarity-assisted		
Ala et al. 2008	General	Coexpression
Miozzi et al. 2008	General	Coexpression
Lage et al. 2007	General	MINT, BIND, IntAct, KEGG, Reactome
CIPHER (Wu et al. 2008)	General	HPRD, OPHID
AlignPI (Wu et al. 2009)	General	HPRD
Centrality-based		
Ozgur et al. 2008	Prostate cancer	Literature mining by GIN (Ozgur et al. 2008)
Ortutay and Vihinen 2009	Immunodeficiency	HPRD
Gudivada et al. 2008	Cardiovascular disease	Genomic-phenomic Semantic web
Others		
Mani et al. 2008	Cancer	B-cell interactome, Co-expression
Karni et al. 2009	General	HPRD

disease are more likely to be related to the disease. For example, Oti et al.(2006) predict candidate genes as those that directly interact with known causative genes of the same disease, and they validate this method against 289 diseases with at least two known disease genes in OMIM. Though the performances vary for different protein network datasets, all are much better than random selection. By applying this method to diseases with both known genes and uncharacterized loci, they are able to predict 300 novel disease candidate genes, of which 10% are confirmed by literature evidence outside OMIM. The same strategy is used in the CPS method in the study of George et al. (2006). When benchmarking with protein interaction data from OPHID, the method has a sensitivity of 0.42, a specificity of 1.0. In another study on cancer gene prediction, Aragues, Sander, and Oliva (2008) define the cancer linker degree (CLD) of a gene as the number of its neighbors that are known to be involved in cancer. They find that CLD of a gene is a good indicator of the probability of being a cancer gene.

Similar results are obtained by Furney et al. (2008a). By integrating protein interaction data with protein sequence conservation, protein domain, gene structure, and regulatory data, Furney et al. train Bayesian classifiers to prioritize proto-oncogenes and tumor suppressor genes. For protein interaction data, they use the number of interactions and the number of interactions with cancer genes, assuming that cancer genes have higher degree and are more likely to interact with other cancer genes. The study by Furney et al. is a typical data integration strategy for gene prioritization. First, a number of data sources/evidences are collected for each candidate gene, and then some machine learning algorithms are used to integrate these features and generate ranking scores. Often data sources are explored in a relatively simple fashion. Another example is provided by Aerts et al. (2006). In this study, up to 12 data sources, including protein interaction data in the database BIND (Bader and Hogue 2003), are used separately to calculate the similarity between training genes (known disease genes) and candidate genes, yielding 12 ranking lists. A rank aggregation algorithm based on order statistics is used to combine these rank lists into a single rank. Again, only direct neighbors are considered for protein interaction data, but instead of neighbor counting, Aerts et al. use the number of common neighbors as the similarity score between known disease genes and candidate genes.

11.3.1.1.2 Shortest Path Length The direct neighbor strategy has some limitations. It is quite possible that two functionally related genes do not interact directly with each other. For example, they may function in different steps of a signaling cascade, yet still lead to the same disease (Brunner and van Driel 2004; Wood et al. 2007). The direct neighbor strategy is more likely to be true for cases where two genes function in the same protein complex (Lage et al. 2007), instead of a pathway. To make use of indirect interactions, one can take higher-order neighborhoods into consideration. The shortest path length measure of distance considers the influence between nodes that are reachable. The length of the shortest path between two biomolecules in molecular interaction networks are assumed to be related with the speed of information communication and/or the strength of the functional association between the two molecules. Thus, the shortest path length is a good measure of functional relatedness, as demonstrated by its correlation with functionally similarity (based on Gene Ontology) (Sharan, Ulitsky, and Shamir 2007). One of the pioneering works to apply shortest path analysis to gene prioritization is from the Rzhetsky group, with a method called Molecular Triangulation (Krauthammer et al. 2004). They use an automatic literature mining system to construct a network around four Alzheimer's disease (AD) genes, and then calculate the shortest path length between all other nodes to these four seed genes. The statistical significance of the distance serves as the final score. The method performs well in predicting additional AD gene candidates identified manually by an expert. This approach was later extended by Liu et al. (2006) by applying shortest path length scoring on a brain-specific gene network, and based on the same four AD seed genes, they were able to rank 37 AD associated genes within the top 46 high-scoring genes.

Like the direct neighbor approach, shortest path analysis has also been used in data integration methods to transform protein interaction data into feature sets. Radivojac et al.

(2008) integrate human protein interaction network, protein sequence, function, physicochemical and structural properties to train support vector machines that are able to predict gene-disease associations with relatively high accuracy. Protein network data are used to calculate the distance between candidate proteins and disease causing proteins, which serves as one important feature for the classifier. A case study for leukemia is given in this study. The training set contains 80 genes associated with leukemia, which are manually curated from OMIM, Swiss-Prot (Boeckmann et al. 2003), and HPRD. Cross-validation shows an accuracy of 77.5% and 15 novel genes are predicted to be associated with leukemia. The authors are able to find from the published literature strong association for 8 of the 15 predictions. One limitation of this approach is that the SVM requires at least 10 known disease-related genes to train the model and to predict novel disease genes.

11.3.1.1.3 Global Distance Measure The problem with shortest path length is that it considers only one of the shortest paths, ignoring the contribution of other shortest paths and other paths with longer length. Most of the time there will be more than one path and even more than one shortest path between two nodes, and the existence of these paths showing additional relatedness between two genes. Another defect is that the shortest path length lacks resolution: the lengths are integers and the longest path in a biological network is typically very small, due to the small world property of biological networks (Jeong et al. 2000; Watts and Strogatz 1998). The so-called global distance measure, mainly diffusion-type distance measure overcomes these drawbacks by considering the topology of the entire network (see illustrations in Kohler et al. 2008). The diffusion kernel K of a graph G is defined as $K = e^{-\beta L}$, where β controls the magnitude of the diffusion. The matrix L is the Laplacian of the graph, defined as D − A, where A is the adjacency matrix of the interaction graph and D is a diagonal matrix containing the nodes' degrees. The inverse Laplacian takes into account all powers of diffusion and thus incorporates all paths along the network. Kohler et al. (2008) propose using the following scoring function to quantify the association between a candidate gene j and a disease:

$$S_j = \sum_i K_{ij}$$

where i represents known disease genes. By applying this approach and another similar random walk approach to an assembled human protein-protein interaction network, they show that methods based on global distance measure significantly outperform those based on local distance measure and non-network approaches. This result is consistently observed for monogenic disorders, polygenic disorders, and cancer. Similar random walk algorithms have been widely used in social- and Web-network analysis to find important nodes (persons or web pages) on the network, such as the PageRank algorithm (Brin and Page 1998) used by Google to rank web pages. By fixing known disease genes as root nodes, some of these algorithms have recently been exploited to prioritize disease genes based on protein network (Chen, Aronow, and Jegga 2009).

11.3.1.2 Proximity of Candidate Gene Pairs: Enabling de Novo Discovery

All the approaches discussed above require at least one disease gene known to cause the disease under study, which covers only about half of human diseases. For genetically unrecognized diseases, these methods do not work. We call methods that do not rely on known disease genes of the same disease de novo methods. To enable de novo prediction, one has to add some other disease-specific information, such as disease similarity, to use genes causing similar disease as a surrogate. We will discuss this type of information later. Here we introduce another method, called Prioritizer (Franke et al. 2006), which does not rely on such phenotype information. Prioritizer assumes the disease-specific information is provided when the candidate genes are available, for example, from a linkage locus associated with the disease. Prioritizer takes at least two genomic regions as input, each containing many candidate genes. Each of the regions is supposed to contain at least one gene causing the disease under study. Assuming the two disease genes should be close to each other on the network, the scoring scheme is designed such that a candidate gene has a higher score if it has smaller distance to genes in another region. A permutation test is introduced to correct the topology differences and yield a p-value based on which all candidate genes are prioritized. Theoretically Prioritizer can be used in de novo discovery of disease genes when multiple genetic regions are given, and this is demonstrated by a case study on breast cancer. Ten 100-gene artificial loci are constructed around 10 known breast cancer genes, and Prioritizer is able to rank 2 to 4 of the 10 breast cancer genes in the top 10 of each locus, when using different gene networks. When the candidate genes in a region are fixed to some known disease genes, this method is essentially the shortest path analysis discussed in the above section. Another method employing this principle is CPS (George et al. 2006), which predicts genes directly interacting with genes from another locus as disease genes.

11.3.2 Phenotype Similarity-Assisted Methods

A natural generalization of the "guilt-by-proximity" principle is that genes causing similar (instead of the same) diseases are likely to be closely related. The additional information provided by similar diseases enables de novo prediction of causative genes for diseases without known causative genes, and will also improve the performance for those with known causative genes. Then two questions remain to be addressed: (1) how to define and compute the similarity between diseases, and (2) how to incorporate disease similarity into disease gene prediction approaches.

11.3.2.1 Calculating and Validating Phenotypic Similarity

A disease can be represented by a set of terms describing its clinical symptoms, namely phenotypes. The phenotypic similarity between two diseases quantifies the overlap or semantic similarity between two sets of terms (Brunner and van Driel 2004; Oti and Brunner 2007). Four different approaches (Care et al. 2009; Lage et al. 2007; Robinson et al. 2008; van Driel et al. 2006) have been proposed to calculate the phenotypic similarity for diseases in OMIM.

van Driel et al. (2006) use a text mining technique to map OMIM disease records to a set of standardized terms, that is, terms defined in MeSH (Medical Subject Headings;

Lowe and Barnett 1994), and then a vector is created for each disease, with each element in the vector representing the number of times the term occurs in the disease record. After adjusting for the hierarchical relationship between different terms, the relative frequency of terms, and size of each record, each disease is represented by a high dimension feature vector of weighted MeSH terms. The similarity between any two diseases is then calculated as the cosine of the angle between the two vectors. Essentially, the method amounts to detecting standardized terms that are (1) common to the description of both diseases, and (2) do not occur too frequently among all diseases such that they are informative about the disease under study. Lage et al. (2007) propose a similar method, and the major difference is, instead of using MeSH as the standard vocabulary, they use UMLS (the unified medical language system; Bodenreider 2004), a more general system containing MeSH and several other vocabularies.

Different approaches have been proposed to evaluate the quality of the calculated disease similarity data. Instead of directly assessing the quality of disease similarity, van Driel et al. (2006) correlated the similarity with the functional relatedness between disease genes. They find that the genes that lead to more similar diseases are more likely to have similar protein sequences, more likely to interact with each other, and more likely to share Pfam domains and Gene Ontology annotations, thus demonstrating that the phenotypic similarity reflects real biological knowledge. Lage et al. (2007) directly evaluate the phenotypic similarity score by comparing it with a putative golden positive set of ~7000 disease pairs. A disease pair is included in this set if one disease is referred to in the text record of another disease. One hundred disease pairs randomly selected from the putative golden positive set are subject to expert OMIM curators' evaluation, and over 90% of them are judged to have a high degree of phenotypic overlap. The phenotypic similarity is then evaluated by calculating the percentage of disease pairs attaining at least a given similarity threshold present in the putative golden positive set. Recently, Care et al. (2009) proposed to use a more stringent golden putative set by only accepting disease pairs with reciprocal references, resulting in a set of about 4000 disease pairs. However, this set has not been evaluated by expert OMIM curators. Interestingly, based on this stringent disease pair set, Care et al. find that the mapping of free text to standard vocabulary is not necessary, as a simple word counting method outperforms the UMLS based method. However, if the disease ID is also counted as terms, the evaluation procedure will prefer the word counting method, thus the comparison is biased. Further studies are needed to exclude this bias and show whether simple word counting is also more powerful than MeSH and other standard vocabularies. All these three phenotypic similarity datasets (Care et al. 2009; Lage et al. 2007; van Driel et al. 2006) have been used in disease gene prioritization, and all show significant improvement compared to methods that do not employ phenotypic similarity data.

More recently, the Human Phenotype Ontology (HPO) was created to standardize the annotation of OMIM diseases (Robinson et al. 2008). Ontology is a special type of standard vocabulary that is particularly suited for knowledge representation and computation, and the usefulness of ontology in biology is evidenced by the great success of Gene Ontology (GO) (Ashburner et al. 2000). GO annotation is now widely accepted as the representation of gene functions, and various methods have been developed to calculate the

functional similarity between genes using GO annotations (T. Xu, Du, and Zhou 2008). Most of these methods can be applied to calculating disease phenotypic similarity using HPO, because HPO is designed to have the same structure as GO (Robinson et al. 2008). In fact, along with the publication of HPO, Robinson et al. (2008) also applied a classic approach (Lord et al. 2003) for calculating gene functional similarity to generate a HPO-based phenotypic similarity for 727 OMIM diseases, which have been classified into one of 21 physiological disorder classes (Goh et al. 2007). Though this similarity data has not been evaluated using the same methods as the UMLS-based disease similarity data, the HPO-based disease similarity network shows a pattern consistent with the physiological disorder classes. Compared to MeSH and UMLS, HPO has several potential advantages for computational phenotype analysis. First, HPO is specifically designed for the needs of describing human hereditary diseases and their phenotypes, and second, as demonstrated by GO, the ontology framework may be more powerful in knowledge representation and computation. Finally, instead of annotating diseases using automatic text mining, HPO experts have manually annotated almost all the OMIM diseases. It is expected that the phenotypic similarity calculated based on HPO will provide more strong support for disease gene prioritization, though so far no such study is available.

11.3.2.2 Modeling with Molecular Network and Phenotype Similarity

The hypothesis underlying most if not all similarity-assisted methods is that similar diseases are caused by functionally related genes. Methods of this type differ in the way to model such correlation and how they incorporate phenotypic similarity information into the model.

11.3.2.2.1 Group Diseases by Similarity The simplest way to exploit phenotypic similarity perhaps is to treat diseases showing a certain level of similarity as the same disease, thus more known disease genes are available for model training or seed propagation. For example, van Driel et al. (2006) have shown that, for the MeSH-based similarity score, biologically meaningful relationships were mostly detected in disease pairs with a similarity score equal to or greater than 0.4. Ala et al. (2008) use this phenotype similarity data, and group diseases according to this threshold. They then employ essentially a neighbor counting strategy, together with a human-mouse conserved coexpression network, to predict disease genes. A similar procedure has been applied to a different dataset (Miozzi et al. 2008).

11.3.2.2.2 Weighted Neighbor Counting Lage et al. (2007) propose a Bayesian model to systematically integrate the UMLS-based similarity score with a weighted human protein-protein interaction network. Basically, for each candidate gene, all the direct neighbors are annotated with, if any, diseases associated with them, and weighted by the similarity to the disease under study. At the same time, all the edges are weighted with a confidence score. Based on these observed data and a uniform priori, the posterior probability of the candidate gene to be associated with the disease under study is obtained via the Bayesian formula. This is essentially a weighted version of neighbor counting: the neighbors of the gene under consideration are weighted by the confidence of the edges (protein-protein

interactions), and the similarity between the disease they lead to and the disease under study. The more reliably a gene is connected to neighbors associated with diseases similar to the disease under study, the more likely the gene is involved in the disease. When applying this approach to 669 genetic loci with known disease genes, they are able to rank the disease gene as the top candidate in 298 loci, significantly outperforming all other methods compared in this study. As the first study to incorporate phenome-wide disease similarity information into disease gene prioritization, it clearly demonstrates the benefits of phenotype data. They then apply the method to 870 genetic loci without the known causative genes and predict a list of 113 candidates for 91 loci, 24 of which are likely to be true predictions according to the recently published literature.

11.3.2.2.3 Prioritize by Interactome-Phenome Correlation Using the same type of data (phenotypic similarity and protein networks), we have proposed a novel method, CIPHER (Correlating protein Interaction network and PHEnotype network to pRedict disease genes), with drastically different formulation (Wu et al. 2008). We choose to directly model the correlation between disease phenotypic similarity and gene functional relatedness, and use the correlation to prioritize candidate genes. We hypothesize that the phenotypic similarity between any two diseases p and p' can be explained by the proximity of their disease genes on the network:

$$S_{p,p'} = C_p + \sum_i \beta_{p,i} \sum_{i'} \exp\left(-L_{i,i'}^2\right)$$

or

$$S_{p,p'} = C_p + \sum_i \beta_{p,i} \Phi_{i,p'}$$

where C_p and $\beta_{p,i}$ are constants for a fixed disease p, and i and i' indicate disease genes of p and p', respectively. $L_{i,i'}$ is the graph distance between gene i and i', which is transformed into proximity by a Gaussian kernel function. The distance measure can be any of the direct neighbor (CIPHER-DN), shortest path (CIPHER-SP), or diffusion kernels.

$$\Phi_{i,p'} = \sum_{i'} \exp\left(-L_{i,i'}^2\right)$$

is defined as the proximity between gene i and disease p' by summing the gene proximity over all known disease genes of p'. This is the classical measure used in shortest path analysis to prioritize candidate genes (Franke et al. 2006; Krauthammer 2004), which do not rely on the phenotypic similarity information. Instead, we choose to evaluate the ability of gene-disease proximity in explaining the disease similarity for a pair of gene and disease (i, p). We create a phenome-wide vector for each gene i: $\mathbf{\Phi}_i = (\Phi_{i,p'})$, and each disease p: $S_p = (S_{p,p'})$, with p' varying for all human diseases. Then we use the correlation

TABLE 11.2 The Ranks and Percentages of Known Breast Cancer Susceptibility Genes in Genome-Wide de Novo Prioritization

Known Breast Cancer Gene	Rank by CIPHER-SP	Rank by CIPHER-DN
BRCA1	1	2
AR	3	3
ATM	19	4
CHEK2	66	19
BRCA2	139	49
STK11	150	21
RAD51	174	36
PTEN	188	24
BARD1	196	41
TP53	287	45
RB1CC1	798	6360
NCOA3	973	343
PIK3CA	1644	367
PPM1D	1946	7318
CASP8	4978	2397
TGF1	7116	3502

between these two vectors as the final score for association between gene *i* and disease *p*. The usefulness of this score is validated by both systematic large-scale cross-validation, and a case study for breast cancer. We have shown that the proposed CIPHER approach can accurately pinpoint the true disease genes from linkage loci or from the whole genome. Further analysis shows that CIPHER is robust to noise in the phenotype similarity data and the protein network data. Without any modification, CIPHER can be applied to de novo discovery, that is, to diseases without known disease genes (without mapped locus or with mapped but uncharacterized loci).

A case study for breast cancer is presented to demonstrate CIPHER's ability in de novo discovery of breast cancer genes. Sixteen known breast cancer genes are treated as non-breast cancer genes and then the whole human genome is prioritized by CIPHER. When using a shortest path length measure of distance (CIPHER-SP), the well-characterized breast cancer gene BRCA1 is ranked at the top, and 10 of the 16 genes are ranked in the top 300, roughly the top 1% of the human genome (Table 11.2). In addition, among the top 10% of the prioritized human genome the same de novo prioritization identifies 15 genes that have been suggested recently to be novel breast cancer genes, including AKT1, ranked at 27, a novel oncogene, and recently a transforming mutation was identified in human breast, colorectal, and ovarian cancers (Carpten et al. 2007). The case study also shows that, though direct neighbor distance measure (CIPHER-DN) works better on ranking known breast cancer genes than CIPHER-SP, it fails to assign ranks to many of the novel susceptibility genes.

All the advantages of CIPHER enable us to perform genome-wide candidate gene prioritization for almost all human diseases, leading to a comprehensive genetic landscape of

human diseases. Automatic clustering and enrichment analysis of the landscape reveal the modularity of human disease and gene relationships (Wu et al. 2008).

11.3.2.2.4 Network Alignment To fully explore the modularity of the human disease genetic landscape, Wu, Liu, and Jiang (2009) borrow ideas from the study of conservation in protein networks (Sharan et al. 2005), or network alignment. Sharan et al. propose a local alignment technique to identify conserved modules between two or more protein interaction networks. To apply this technique, Wu, Liu, and Jiang (2009) created a human disease network by linking diseases with a phenotypic similarity score larger than a given threshold, resulting in a human disease similarity network. Then they used the network alignment technique to compare the human disease network and human protein network, and identify 39 disease modules together with corresponding gene modules, or bimodules. Examining the functions of genes and categories of diseases, they show that these bimodules represent disease families and their common pathways. After validating the bimodule identification methods, they propose to use it for disease gene prediction. Essentially, they predict a candidate gene to cause a disease if it is linked to the disease in a bimodule. This approach, named AlignPI (Wu, Liu, and Jiang 2009), attains similar performance with CIPHER.

11.3.3 Prioritize by Network Centrality

The working principle for methods in this category is totally different from those discussed above. Here we assume that genes with higher centrality on a network are more likely to cause disease. To be more informative, the network is often specially designed.

11.3.3.1 Centrality in a Context-Specific Gene Network

Ozgur et al. (2008) introduce a sophisticated automatic literature mining approach to construct a disease-specific gene interaction network, in their example, a prostate caner network. Hypothesizing that genes with high centrality in a disease-specific network are likely to be related to the disease, they used several network centrality measures to rank genes in the prostate cancer network, and found that two measures, degree and eigenvector, were highly informative of known prostate cancer genes. Specifically, 19 of the top 20 genes returned by the approach have supportive evidence from either OMIM or PGDB (Prostate Gene DataBase; L.C. Li et al. 2003), a curated database of genes related to prostate cancer. One limitation of the approach is that, similar to the Molecular Triangulation approach (Krauthammer et al. 2004), it relies on a list of seed genes (genes known to be involved in the disease) to construct the network, yet to what extent the choice of seed genes influences the results is not discussed. In a second study, Ortutay and Vihinen (2009) create a human immune gene interaction network by linking curated immune genes with interaction data from HPRD, and use multiple centralities, including degree and closeness, to prioritize candidate genes for immunodeficiencies.

11.3.3.2 Centrality in a Genomic-Phenomic Network

So far we have focused on networks whose nodes are genes or proteins. There are also other network approaches using more complicated networks. For example, Gudivada et al. (2008)

create a network of various concepts, with edges representing the association between genes and Gene Ontology annotations, pathways, mouse phenotypes, and human clinical features, therefore establishing a semantic web of integrated genomic and phenomic knowledge. Assuming that disease-causing genes tend to play functionally important roles and share similar biochemical characteristics with genes causing diseases with similar clinical features, the authors use a Google-like search and ranking algorithm (Mukherjea 2005) to prioritize candidate genes. The efficiency of the proposed approach is tested in prioritizing candidate genes for cardiovascular diseases.

11.3.4 Other Methods

Here we discuss several interesting and promising approaches that do not fall into the above categories. These methods are interesting because they do not rely on known disease gene or disease similarity, yet still are able to find the causal gene based on the genome-wide secondary response.

Mani et al. (2008) propose a method called Interactome Dysregulation Enrichment Analysis (IDEA) to predict oncogenes. Using interactome and microarray data, they first identify dysregulated interactions, that is, gene pairs with annotated interaction but significantly changed correlation according to gene expression profiling of normal and tumor samples. Then genes with an unusually high number of dysregulated interactions in its neighborhood are predicted as oncogenes. The assumption is that genes implicated in cancer initiation and progression will show dysregulated interactions with their molecular partners. In three B-cell tumor phenotypes, the method correctly identifies the known genes in the top 20 candidates out of about 8000 genes. The IDEA method exploits direct neighbors only. As demonstrated by other examples discussed in previous sections, shortest path-based analysis might yield higher coverage and more novel predictions that are not so obvious from protein interaction data.

A more sophisticated network-based approach has been proposed to solve a problem with similar settings. With the protein interaction network available, Karni, Soreq, and Sharan (2009) attempted to predict the causal gene from expression profile data assumed to be perturbed by the gene. They first identified a set of disease-related genes whose expression is changed in the disease state, then based on a parsimonious assumption, an algorithm sought the smallest set of genes that could best explain the expression changes of the disease-related genes in terms of probable pathways leading from the causal to the affected genes in the network. Experiments with both simulated and real knock-out data show that the proposed approach attains very high accuracy. Further validations on expression data from different types of cancer show high accuracy in pinpointing known oncogenes. For example, using expression profiles for a subset of acute leukemias involving chromosomal translocation of the mixed leukemia gene (MLL), the algorithm correctly assigns MLL an average rank of 1.5, out of 168 genes in the neighboring region. When applying the algorithm to four breast cancer datasets, the major causal genes BRCA1 and BRCA2 are ranked very high. They are also able to show that the algorithm outperforms a naive algorithm that ranks disease-associated genes according to their shortest path length in the network to the directly affected genes.

11.4 DISCUSSION

Five years after the first network-based candidate gene prioritization method (Krauthammer et al. 2004), now there are more than 20 available in the published literature (Table 11.1), calling for a comprehensive comparison among them. Unfortunately, a systematic and rigorous direct comparison is very difficult and seldom occurs in the literature, mostly because different methods use different types of data sources, and are trained and tested on customized datasets which are often unavailable to others. For methods running with the same type of data sources, one can re-implement different methods proposed by other groups, and compare them using one dataset that is probably not the original dataset on which most methods were tested. Such a comparing scheme is only feasible for comparing methods that are easy to implement. For example, Kohler et al. (2008) implemented four algorithms that are purely network based and compared their performance, showing the superiority of global distance measures. For situations where methods are not easy to implement, researchers often compare self-reported performances along with the original publications. Self-reported performances are often transformed into so-called (average) fold enrichment, that is, the average fold of enriching the true disease genes among a short top list, compared to random selection (Lage et al. 2007; Wu et al. 2008). According to this criterion, disease similarity-assisted methods significantly outperform previous methods, and we are able to show that CIPHER works even better, especially for higher recall. The problem with the fold enrichment criterion is that it is influenced by the total number of candidate genes and the size of the top list, while these numbers often vary across different methods. For comprehensive comparison, a community-wide effort is needed, to establish a publicly available data platform, including widely used different data sources, a training dataset of known gene-disease associations, and a blinded test dataset. Such efforts have recently been performed in a related field, the mouse gene function prediction (Pena-Castillo et al. 2008).

Most of the methods discussed here are not designed particularly for cancer, though they can be applied to cancer without any modification. Here we discuss some cancer specific issues. Though these issues are not particularly related with network-based approaches, it will be important for us to realize their impact. First, prediction methods generally do not differentiate two types of cancer genes that are different in many aspects, thus fail to generate more testable hypotheses that could guide further experimental validation. Genes that can initialize tumorigenesis are traditionally divided into oncogene and tumor suppressor gene, though more recently stability gene has been proposed to be a further type of cancer gene (Vogelstein and Kinzler 2004). Study has shown that a classifier using protein conservation, gene sequence, protein domains, protein interactions, and regulatory data is able to differentiate oncogenes from tumor suppressor genes (Furney et al. 2008a). Specifically, they show that tumor suppressor genes have higher degree than oncogenes, while oncogene evolution appears to be more highly constrained (Furney et al. 2008b). Together, these results imply that oncogenes and tumor suppressor genes may be inherently different. Taking the difference into consideration may further improve the prediction of cancer genes. In addition, the experimental procedures to verify oncogenes and

tumor suppressor genes are different, computational prediction will facilitate the verification if it can tell oncogene from tumor suppressor gene. Another special issue for cancer gene prediction is that there are several cancer-specific genome-wide data sources which may greatly advance the prediction of cancer genes. For example, large-scale sequencing of the human cancer genome has identified thousands of genes carrying nonsilent mutations in breast or colon cancer samples (Sjoblom et al. 2006; Wood et al. 2007), while array-based techniques, such as array comparative genomic hybridization (aCGH; Pinkel et al. 1998) and representational oligonucleotide microarray analysis (ROMA; Lucito et al. 2003) have been developed to localize genes with altered copy numbers (amplified or deleted) in cancer samples. Combining candidate genes identified from the above large-scale screen and computational cancer gene prioritization methods will greatly facilitate the discovery of human cancer-causing genes.

With the development of high-throughput techniques in exploring the human cancer genome, and the increasing quality in large-scale detection of protein interactions, network-based cancer gene discovery will remain promising and continue to be an active research area. Progress in this area will also benefit from other network-based research, such as the network-based prediction of protein functions (Sharan, Ulitsky, and Shamir 2007), especially functions of cancer genes (Hu et al. 2007), and the discovery of novel drug targets for cancer (Campillos et al. 2008; Huang and Harari 1999), since the formulation are similar thus novel methods developed for one problem may also apply to the other. We expect that network analysis will provide both systems thinking and methodology advantages in our way to understand the complexity of life.

ACKNOWLEDGMENTS

This work is supported by the 863 project of China (No. 2006AA02Z311), the NSFC (Nos. 60934004 and 60721003) and NCET-07-0486.

REFERENCES

Aerts, S., Lambrechts, D., Maity, S. et al. 2006. Gene prioritization through genomic data fusion. *Nature Biotechnol* 24: 537–544.

Ala, U., Piro, R. M., Grassi, E. et al. 2008. Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol* 4: e1000043.

Altshuler, D., Daly, M. J., and Lander, E. S. 2008. Genetic mapping in human disease. *Science* 322: 881–888.

Aragues, R., Sander, C., and Oliva, B. 2008. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics* 9: 172.

Ashburner, M., Ball, C. A., Blake, J. A. et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* 25: 25–29.

Bader, G. D. and Hogue, C. W. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4: 2.

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

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32: D267–D270.

Boeckmann, B., Bairoch, A., Apweiler, R. et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31: 365–370.

Breitkreutz, B. J., Stark, C., Reguly, T. et al. 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* 36: D637–D640.

Brin, S and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In: *7th International World Wide Web Conference: April 14–18 1998*. Brisbane, Australia: Elsevier Science, pp. 107–117.

Brown, K. R. and Jurisica, I. 2005. Online predicted human interaction database. *Bioinformatics* 21: 2076–2082.

Brunner, H. G. and van Driel, M. A. 2004. From syndrome families to functional genomics. *Nature Rev Genet* 5: 545–551.

Burmeister, M., McInnis, M. G., and Zollner, S. 2008. Psychiatric genetics: progress amid controversy. *Nature Rev Genet* 9: 527–540.

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., and Bork, P. 2008. Drug target identification using side-effect similarity. *Science* 321: 263–266.

Care, M. A., Bradford, J. R., Needham, C. J., Bulpitt, A. J., and Westhead, D. R. 2009. Combining the interactome and deleterious SNP predictions to improve disease gene identification. *Hum Mutat* 30: 485–492.

Carpten, J. D., Faber, A. L., Horn, C. et al. 2007. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature* 448: 439–444.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M. et al. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 35: D572–D574.

Chen, J., Aronow, B. J., and Jegga, A. G. 2009. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 10: 73.

Feldman, I., Rzhetsky, A., and Vitkup, D. 2008. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci USA* 105: 4323–4328.

Fields, S. and Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340: 245–246.

Fields, S. and Sternglanz, R. 1994. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet* 10: 286–292.

Folstein, S. E. and Rosen-Sheidley, B. 2001. Genetics of autism: complex aetiology for a heterogeneous disorder. *Nature Rev Genet* 2: 943–955.

Franke, L., van, Bakel H., Fokkens, L. et al. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 78: 1011–1025.

Frayling, T. M. 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Rev Genet* 8: 657–662.

Furney, S. J., Calvo, B., Larranaga, P., Lozano, J. A., and Lopez-Bigas, N. 2008a. Prioritization of candidate cancer genes: an aid to oncogenomic studies. *Nucleic Acids Res* 36: e115.

Furney, S. J., Madden, S. F., Kisiel, T. A., Higgins, D. G., and Lopez-Bigas, N. 2008b. Distinct patterns in the regulation and evolution of human cancer genes. *In Silico Biol* 8: 33–46.

Futreal, P. A., Coin, L., Marshall, M. et al. 2004. A census of human cancer genes. *Nature Rev Cancer* 4: 177–183.

George, R. A., Liu, J. Y., Feng, L. L. et al. 2006. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* 34: e130.

Ghavidel, A., Cagney, G., and Emili, A. 2005. A skeleton of the human protein interactome. *Cell* 122: 830–832.

Giot, L., Bader, J. S., Brouwer, C. et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* 302: 1727–1736.

Goh, K. I., Cusick, M. E., Valle, D. et al. 2007. The human disease network. *Proc Natl Acad Sci USA* 104: 8685–8690.

Gudivada, R. C., Qu, X. A., Chen, J. et al. 2008. Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge. *J Biomed Inform* 41: 717–429.

Han, J. D. 2008. Understanding biological functions through molecular networks. *Cell Res* 18: 224–237.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. 2006. How complete are current yeast and human protein-interaction networks? *Genome Biol* 7: 120.

Hu, P., Bader, G., Wigle, D. A., and Emili, A. 2007. Computational prediction of cancer-gene function. *Nature Rev Cancer* 7: 23–34.

Huang, S. M. and Harari, P. M. 1999. Epidermal growth factor receptor inhibition in cancer therapy: biology, rationale and preliminary clinical results. *Invest New Drugs* 17: 259–269.

Ito, T., Chiba, T., Ozawa, R. et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 98: 4569–4574.

Jansen, R., Yu, H., Greenbaum, D. et al. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302: 449–453.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. 2001. Lethality and centrality in protein networks. *Nature* 411: 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. 2000. The large-scale organization of metabolic networks. *Nature* 407: 651–654.

Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30.

Karni, S., Soreq, H., and Sharan, R. 2009. A network-based method for predicting disease-causing genes. *J Comput Biol* 16: 181–189.

Kerrien, S., Alam-Faruque, Y., Aranda, B. et al. 2007. IntAct: open source resource for molecular interaction data. *Nucleic Acids Res* 35: D561–D565.

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. 2008. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82: 949–958.

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. 2004. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci USA* 101: 15148–15153.

Lage, K., Karlberg, E. O., Storling, Z. M. et al. 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnol* 25: 309–316.

Li, L. C., Zhao, H., Shiina, H., Kane, C. J., and Dahiya, R. 2003. PGDB: a curated and integrated database of genes related to the prostate. *Nucleic Acids Res* 31: 291–293.

Li, S., Armstrong, C. M., Bertin, N. et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* 303: 540–543.

Li, S., Wu, L., and Zhang, Z. 2006. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics* 22: 2143–2150.

Lim, J., Hao, T., Shaw, C. et al. 2006. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125: 801–814.

Liu, B., Jiang, T., Ma, S. et al. 2006. Exploring candidate genes for human brain diseases from a brain-specific gene network. *Biochem Biophys Res Commun* 349: 1308–1314.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19: 1275–1283.

Lowe, H. J. and Barnett, G. O. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 271: 1103–1108.

Lucito, R., Healy, J., Alexander, J. et al. 2003. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* 13: 2291–2305.

Mani, K. M., Lefebvre, C., Wang, K. et al. 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol* 4: 169.

McKusick, V. A. 2007. Mendelian inheritance in man and its online version, OMIM. *Am J Hum Genet* 80: 588–604.

Mewes, H. W., Amid, C., Arnold, R. et al. 2004. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* 32: D41–D44.

Miki, Y., Swensen, J., Shattuck-Eidens, D. et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266: 66–71.

Miozzi, L., Piro, R. M., Rosa, F. et al. 2008. Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *PLoS One* 3: e2439.

Mukherjea, S. 2005. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. *Brief Bioinform* 6: 252–262.

Oldenburg, R. A., Meijers-Heijboer, H., Cornelisse, C. J., and Devilee, P. 2007. Genetic susceptibility for breast cancer: how many more genes to be found? *Crit Rev Oncol Hematol* 63: 125–149.

Ortutay, C. and Vihinen, M. 2009. Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res* 37: 622–628.

Oti, M. and Brunner, H. G. 2007. The modular nature of genetic diseases. *Clin Genet* 71: 1–11.

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. 2006. Predicting disease genes using protein-protein interactions. *J Med Genet* 43: 691–698.

Ozgur, A., Vu, T., Erkan, G., and Radev, D. R. 2008. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* 24: i277–i285.

Pena-Castillo, L., Tasan, M., Myers, C. L. et al. 2008. A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol* 9 (Suppl 1): S2.

Peri, S., Navarro, J. D., Amanchy, R. et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13: 2363–2371.

Pharoah, P. D., Tyrer, J., Dunning, A. M., Easton, D. F., and Ponder, B. A. 2007. Association between common variation in 120 candidate genes and breast cancer risk. *PLoS Genet* 3: e42.

Pinkel, D., Segraves, R., Sudar, D. et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet* 20: 207–211.

Radivojac, P., Peng, K., Clark, W. T. et al. 2008. An integrated approach to inferring gene-disease associations in humans. *Proteins* 72: 1030–1037.

Rhodes, D. R., Tomlins, S. A., Varambally, S. et al. 2005. Probabilistic model of the human protein-protein interaction network. *Nature Biotechnol* 23: 951–959.

Robinson, P. N., Kohler, S., Bauer, S. et al. 2008a. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 83: 610–615.

Rual, J. F., Venkatesan, K., Hao, T. et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437: 1173–1178.

Rzhetsky, A., Iossifov, I., Koike, T. et al. 2004. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* 37: 43–53.

Salwinski, L., Miller, C. S., Smith, A. J. et al. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32: D449–D451.

Sharan, R., Suthram, S., Kelley, R. M. et al. 2005. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* 102: 1974–1979.

Sharan, R., Ulitsky, I., and Shamir, R. 2007. Network-based prediction of protein function. *Mol Syst Biol* 3: 88.

Sjoblom, T., Jones, S., Wood, L. D. et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* 314: 268–274.

Sprinzak, E., Sattath, S., and Margalit, H. 2003. How reliable are experimental protein-protein interaction data? *J Mol Biol* 327: 919–923.

Strausberg, R. L., Simpson, A. J., and Wooster, R. 2003. Sequence-based cancer genomics: progress, lessons and opportunities. *Nature Rev Genet* 4: 409–418.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. 2006. A text-mining analysis of the human phenome. *Eur J Hum Genet* 14: 535–542.

Vastrik, I., D'Eustachio, P., Schmidt, E. et al. 2007. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8: R39.

Vogelstein, B. and Kinzler, K. W. 2004. Cancer genes and the pathways they control. *Nature Med* 10: 789–799.

von Mering, C., Jensen, L. J., Snel, B. et al. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 33: D433–D437.

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

Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of "small-world" networks. *Nature* 393: 440–442.

Wood, L. D., Parsons, D. W., Jones, S. et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318: 1108–1113.

Wooster, R., Bignell, G., Lancaster, J. et al. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378: 789–792.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. 2008. Network-based global inference of human disease genes. *Mol Syst Biol* 4: 189.

Wu, X., Liu, Q., and Jiang, R. 2009. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* 25: 98–104.

Xia, K., Dong, D., and Han, J. D. 2006. IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics* 7: 508.

Xu, J. and Li, Y. 2006. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 22: 2800–2805.

Xu, T., Du, L., and Zhou, Y. 2008. Evaluation of GO-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data. *BMC Bioinformatics* 9: 472.

Zhu, X., Gerstein, M., and Snyder, M. 2007. Getting connected: analysis and principles of biological networks. *Genes Dev* 21: 1010–1024.