G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2013;11(4):200-210
http://dx.doi.org/10.5808/GI.2013.11.4.200

REVIEW ARTICLE

Review of Biological Network Data and Its Applications

Donghyeon Yu, MinSoo Kim, Guanghua Xiao, Tae Hyun Hwang*

Department of Clinical Sciences, Quantitative Biomedical Research Center,
University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

Studying biological networks, such as protein-protein interactions, is key to understanding complex biological activities. Various types of large-scale biological datasets have been collected and analyzed with high-throughput technologies, including DNA microarray, next-generation sequencing, and the two-hybrid screening system, for this purpose. In this review, we focus on network-based approaches that help in understanding biological systems and identifying biological functions. Accordingly, this paper covers two major topics in network biology: reconstruction of gene regulatory networks and network-based applications, including protein function prediction, disease gene prioritization, and network-based genome-wide association study.

Keywords: biological network, disease gene prioritization, gene regulatory networks, genome-wide association study, protein function prediction, protein-protein interaction

Introduction

Network representations have been used to describe interactions between entities of interest in various areas. In particular, network representations are useful to analyze and visualize complex biological activities. Global patterns in a large-scale complex system can be shown by representing the entities and their interactions with nodes and edges, respectively. For instance, Schwikowski *et al*. [1] created protein-protein interaction (PPI) networks to predict novel protein functions in yeast *Saccharomyces cerevisiae*. By using the network representations, it was found that 2,358 among 2,709 total interactions compose a single large network. Also, biological pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [2] and Reactome [3], provide numerous pathways that are represented by networks with nodes of molecules and directed edges of their actions. In addition, various mathematical properties and models of a network have been developed in graph theory. Several reviews [4-6] have illustrated the mathematical properties and topological characteristics of a network with biological systems.

The advance of high-throughput technologies, including DNA microarray [7], next-generation sequencing [8], and the two-hybrid screening system [9], has led to the large-scale datasets in genomics and proteomics, which are referred to as 'omics' data. These omics data have been collected and organized to identify biological functions. This paper focuses on biological network analysis related to omics data, such as gene expression levels and PPIs. We first report several major public interaction databases for the omics data and then introduce two major topics in network biology: reconstruction of gene regulatory networks (GRNs) and network-based applications, including protein function prediction, disease gene prioritization, and network-based genome-wide association study.

Network Resources

Experimental results from high-throughput technologies, such as the two-hybrid screening system for detecting interactions between biological molecules, have formed various types of network datasets that are released and updated in public databases on the web. These databases commonly enable web-based searches and provide raw datasets of pairs of molecules. In this review, we report 11

Received October 15, 2013; Revised November 20, 2013; Accepted November 21, 2013

***Corresponding author:** Tel: +1-214-648-4003, Fax: +1-214-648-1663, E-mail: TaeHyun.Hwang@UTSouthwestern.edu

useful public databases for PPIs and transcriptional regulatory interactions (Table 1) [10-20]. For PPIs, BioGRID [13], MIPS [11], and STRING [16] are the most frequently used to predict protein functions for PPIs. BioGRID provides 496,761 non-redundant PPIs for various species, ranging from yeast *S. cerevisiae* to human. On the other hand, MIPS not only provides mammalian PPIs but also offers the functional catalogs that contain descriptions of protein functions. Unlike BioGRID and MIPS, STRING contains identified and predicted functional interactions of proteins with functional similarity scores (i.e., STRING offers weighted networks). For transcriptional regulatory interactions, transcriptional regulatory element database (TRED) [19] offers GRNs and transcriptional factors for three species; human, mouse, and rat. On the other hand, RegulonDB [20] contains both experimental datasets and computational prediction results of transcriptional regulatory interactions for the *Escherichia coli* K-12 organism.

Statistical Reconstruction of the Gene Regulatory Network from Gene Expression Data

Biological networks are generally constructed using known interactions identified from previous experiments. To integrate these separate pieces of information in the literature, text mining methods [21-24] have been proposed and used in the majority of public databases explained in the previous section. Although most biological networks are based on identified interactions under many different conditions and properties, GRNs can be constructed from gene expression datasets from a user's experimental datasets that implicitly contain gene regulation information in specific conditions (e.g., disease-specific, tissue-specific, or drug-specific GRNs). Most recently, the Encyclopedia of DNA Elements (ENCODE) project [25] produced numerous RNA-sequencing (RNA-seq) datasets that can provide gene

Table 1. Public network resources

Database	Type	Species	URL	Reference
DroID	Protein interaction	*Drosophila*	http://www.droidb.org	[10]
MIPS	Protein interaction/functional catalog	Mammal	http://mips.helmholtz-muenchen.de	[11]
HPRD	Protein interaction	Human	http://www.hprd.org	[12]
BioGRID (GRID)	Protein interaction	No restriction	http://thebiogrid.org	[13, 14]
DIP	Protein interaction	No restriction	http://dip.doe-mbi.ucla.edu/dip	[15]
STRING	Protein interaction	No restriction	http://string-db.org	[16]
MINT	Protein interaction	No restriction	http://mint.bio.uniroma2.it/mint	[17]
IntAct	Protein interaction	No restriction	http://www.ebi.ac.uk/intact	[18]
Reactome	Pathway/protein Interaction	No restriction	http://www.reactome.org	[3]
TRED	Transcriptional regulatory	Human/mouse/rat	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home	[19]
RegulonDB	Transcriptional regulatory	*Escherichia coli* K-12	http://regulondb.ccg.unam.mx/	[20]

Table 2. Reconstruction methods for gene regulatory network

Method	Approach	Implementation	URL
SPACE	Gaussian graphical model	C, R	http://cran.r-project.org/web/packages/space
Graphical Lasso	Gaussian graphical model	Fortran, R	http://cran.r-project.org/web/packages/glasso
CLIME	Gaussian graphical model	R	http://cran.r-project.org/web/packages/clime
GeneNet	Gaussian graphical model	R	http://cran.r-project.org/web/packages/GeneNet
B-Course	Bayesian network	Java	http://b-course.cs.helsinki.fi
BNT	Bayesian network	Matlab	http://code.google.com/p/bnt/
Werhli *et al.*'s BN	Bayesian network	Matlab	http://www.bioss.ac.uk/people/adriano/comparison/comparison.html#software
WGCNA	Correlation	C, R	http://cran.r-project.org/web/packages/WGCNA
Relevance network	Information theory	Java	http://www.newatlantictech.com/products.html
ARACNE	Information theory	C++, Java	http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE
CLR	Information theory	C, Matlab	http://m3d.mssm.edu/network_inference.html
GTRNetwork	Information theory	Matlab	http://www.biomedcentral.com/1471-2105/12/233
NARROMI	Information theory	Matlab	http://csb.shu.edu.cn/narromi.htm

expression levels and chromatin immunoprecipitation-sequencing (ChIP-seq) datasets that directly contain information about transcription factors (TFs). Integrative methods for reconstructing GRNs [26, 27] have been developed with the ENCODE [25] and modENCODE [28] project datasets. But, we focus on statistical approaches to reconstruct GRNs with gene expression datasets. Readers can refer to [26, 27] to read the details of reconstructing GRNs with ENCODE project datasets.

Many statistical approaches that infer networks from gene expression datasets have been developed. In this section, we briefly introduce four approaches to reconstruct GRNs: Gaussian graphical model, Bayesian network, correlation network, and information theory. Table 2 summarizes the methods described in this section with implementation languages and available URLs. A comparison study of several methods in these approaches has been published [29]. We remark that this review covers recent methods of the Gaussian graphical model and information theory approaches.

Gaussian graphical model

To cover the basic principle behind the Gaussian graphic model, let $G = (V, E)$ be an undirected graph with a set V of nodes and a set E of edges and $X = (X^1, X^2, \cdots, X^p) = (X_1, X_2, \cdots, X_n)^T$ be an $n \times p$ dimensional design matrix, where X^j denotes a j-th variable and X_i denotes an i-th sample. In the Gaussian graphical model, it is assumed that an observation is from a p-dimensional multivariate normal distribution with mean zero and covariance matrix Σ (i.e., $X_i \sim N(0, \Sigma)$ for $i=1, 2, \cdots, n$). From this normality assumption, X^i and X^j are conditionally independent if $(\Sigma^{-1})_{ij} = 0$ [30]. With this property, the Gaussian graphical model represents a conditional dependency between two variables into an edge (i.e., $(i, j) \in E$ if $(\Sigma^{-1})_{ij} \neq 0$). Thus, we can obtain a network structure by estimating the inverse of covariance matrix (Σ^{-1}), which is called a precision matrix. It is known that the maximum likelihood estimator (MLE) of the precision matrix is an inverse matrix of the sample covariance matrix. However, for high-dimensional data ($p > n$), the MLE of the precision matrix can not be obtained from the sample covariance matrix, since the sample covariance matrix is singular.

To resolve this problem and obtain the sparse solution, l_1-regularized methods have been developed. These methods can be categorized into four types: regression-based [31, 32], penalized likelihood [33-38], empirical Bayes [39], and constrained l_1 minimization [40]. We briefly introduce a recently developed example for each type of method. First, the sparse partial correlation estimation (SPACE) method [32] jointly solves p regression problems with l_1 norm penalty on partial correlations. SPACE performs well in the detection of hub nodes that have many connections with other nodes. Unlike regression based-methods, the penalized likelihood-based methods directly maximize a likelihood function with positive definite constraints and l_1 norm penalty on elements of the precision matrix. This maximization problem, proposed by [33], is more difficult to solve than the regression problem. To efficiently solve this problem, Friedman *et al*. [38] developed the graphical lasso algorithm that is motivated from [34, 35] and faster than other existing methods. In the empirical Bayes-based methods, Schafer and Strimmer [39] proposed the GeneNet method, based on multiple testing procedure on the partial correlations estimated by the Moore-Penrose pseudo inverse and the bootstrap method. Finally, Cai *et al*. [40] recently proposed constrained l_1-minimization for inverse matrix estimation (CLIME), which directly minimizes the l_1-norm of the precision matrix with a constraint for relaxation of the precision matrix condition.

Bayesian network

A Bayesian network is a probabilistic framework for representing a directed acyclic graph (DAG) structure, while Gaussian graphical models consider undirected graph structures. With the structure of DAG, the joint probabilities can be explicitly calculated from a simple conditional distribution. To obtain the network structure, Bayesian network methods generally select the graph structure by following three steps. Step 1—a candidate DAG structure is chosen among all possible DAG structures. Step 2—the posterior probability of the candidate DAG structure, given the gene expression data, is calculated. Step 3—repeat steps 1 and 2 until the highest posterior probability is obtained. Thus, searching candidate DAG structures becomes a critical part of the Bayesian network methods when the number of nodes increases. To avoid this problem, many heuristics are applied to improve efficiency [41-43]. For instance, B-Course [41], which is a web-based application, uses a combination of stochastic and greedy search, and BNT [42] and Werhli's Bayesian network [43] adopt Markov Chain Monte Carlo (MCMC) methods to find the best network structure for given gene expression data.

Correlation network

Correlation networks represent pairwise correlations between two nodes into edges. Although the construction of correlation networks can be straightforward to obtain from Pearson's correlations, it leads to a problem—that all the edges in the network are connected, since these correlations are generally non-zero. Thus, it is required that meaningless relationships in the correlation network be removed to focus

on important edges that highly correlate nodes. To remedy this problem, hard thresholds [44] or soft thresholds [45] are applied to Pearson's correlations. By using threshold cutoffs, correlation network methods can capture biologically meaningful relationships. Moreover, this representation, based on correlations, is useful not only to identify modules of genes but also to interpret regulation interactions. "WGCNA" [46], an R implemented package, is an example that makes use of correlation networks. WGCNA provides the construction of unweighted (or weighted) networks with hard and soft threshold schemes [44, 45]. In addition, WGCNA offers various functions to analyze the network, including module detection, calculation of topological properties, and visualization.

Information theory

Information theory-based methods construct the GRN, based on information theoretic scores, such as the mutual information (MI), to measure dependencies between variables [47-50]. Unlike correlations, the MI does not assume the linearity and the continuity of variables. Thus, information theory-based methods generally outperform other methods that are based on correlation coefficients when the true network structure contains non-linearity dependencies. We first review two popular methods that reconstruct GRNs based on MI values and then introduce recently developed methods.

The relevance network method [47] uses the MI values to determine the edges in the GRN. To select significant edges, the relevance network proposes a threshold rule based on the distribution of the permuted MI values. The maximum value of the average of the permuted MI values is considered the threshold value. On the other hand, ARACNE [48, 49] additionally takes into account data processing inequality (DPI), in which the MI value of an indirect interaction is less than or equal to each MI value of the direct interactions for all triplets of nodes that only have two direct interactions. This DPI enforces that ARACNE [48, 49] discards indirect interactions and some direct interactions with small MI values.

To improve the accuracy of the information theory-based methods, the context likelihood of relatedness (CLR) algorithm [50] and the gene expression and transcription factor activity-based relevance network (GTRNetwork) algorithm [51] have been developed. These methods focus on the transcriptional regulatory interactions between known TFs and their target genes. Instead of using the MI values directly, the CLR algorithm proposes a likelihood value based on z-scores derived from the empirical distribution of the MI values in order to adjust random noises in the MI values. In addition to the CLR algorithm, the GTRNetwork algorithm additionally considers transcription factor activities (TFAs) between TFs and target genes as a hidden layer. The GTRNetwork algorithm first estimates changes of TFAs with known TF-gene networks and then identifies transcriptional regulatory interactions between the estimated TFAs and genes with the CLR algorithm. More recently, Zhang *et al*. [52] described the transcription procedure with TFAs by using the ordinary differential equation. Unlike the GTRNetwork algorithm, the NARROMI algorithm considers the gene expression data for candidate TFs as the TFAs and solves the minimization of the absolute errors between inferred and observed expressions with l_1-norm penalty on the strength coefficients, which is referred to as recursive optimization in [52]. The NARROMI algorithm proposes the linear combination of the MI values and the absolute values of estimated strength coefficients as the final scores to construct the GRN.

Network-Based Applications

From the advances of high-throughput technologies, large-scale networks have been identified and are available from various public databases, as summarized in Table 1. In this section, we focus on network-based applications, especially how to use previously identified network resources in order to obtain meaningful biological interpretations. We first introduce three basic concepts in graph theory [53] to give an overview of the basic concepts of the network-based methods. First, a *neighborhood* of a node is a set of nodes connected to the node. Second, the *distance* between two nodes is defined as a length of the shortest path between them if the path exists; otherwise, it is set to infinity. Finally, an *adjacency matrix* is a binary and symmetric matrix such that its *ij*-th element is equal to 1 if there is an edge from a node *i* to a node *j*; otherwise, it is 0. In some cases, a weighted adjacency matrix can be used to represent the strengths of edges that usually fall between 0 and 1. With these basic concepts, we introduce three network-based applications: protein function prediction, disease gene prioritization, and genome-wide association study.

Protein function prediction

With the results of sequencing genomes, the efforts to predict protein functions have been focused on the functionalities of genomic annotations. In the initial stage, the prediction of protein function begins with the sequence homologies to annotated proteins [54-57]. These methods have been successful, but 70% of proteins still remain unannotated [58]. Accordingly, various types of methods have been developed to characterize unannotated proteins. In this review, we introduce four approaches in protein

function prediction, based on direct interactions in the network: neighborhood-based, graph-based, Bayesian, and Kernel-based approaches. The methods are summarized in Table 3 [10, 11, 13-16, 59-83].

In neighborhood-based approach, the proposed methods [1, 62, 63, 65] commonly consider the number of edges connected to annotated proteins in a neighborhood. The neighborhood counting method [1] only takes into account the frequencies of annotated proteins in the neighborhood and chooses the top three functions, with the calculated frequencies ranked in descending order for each protein. Other methods also have derived their own scores, such as χ^2 statistics [62] and functional similarities [63, 65], based on the annotated protein information in the neighborhood. These methods choose a function with the highest score as a predicted function for each protein.

The graph-based approach is similar to the neighborhood-based approach, but the graph-based approach focuses more on how to label the unannotated proteins with graph theoretical properties, such as the distance and the adjacency matrix. The label assignment models have been proposed with the adjacency matrix [66] and the weighted adjacency matrix [67], respectively. Since these assignment problems are computationally intractable, the heuristic methods, such as the simple threshold rule or simulated annealing [84], have been applied. To avoid these computational problems, several propagation-flavored methods have been developed. First, the label propagation methods [69, 70] obtain the optimal assignments and the optimal combination of the weight matrices that reflects different types of networks. Second, the functional flow method [68] iteratively spreads the flow from the annotated protein to the unannotated proteins by connected edges. The functional flow score, defined as an amount of the flow, is the criterion of the prediction. Third, the sequential linear neighborhood propagation method [71] sequentially updates unlabeled proteins according to their shortest path distance to the set of labeled proteins. Finally, the neighbor relativity coefficient (NRC) method [72] derives the NRC score by integrating various graph topological properties, such as the shortest path distance, path connectivity, and common neighbors.

The Bayesian approach takes into account the posterior probabilities of binary label random variables to obtain the prediction from the observed network and annotated proteins. Markov random field (MRF)-based [85] methods [74, 75] have been proposed and modified to Bayesian MRF [77] recently. To predict protein functions, these methods commonly derive the marginal posterior probability of the binary label variable given other variables and then estimate the posterior probability by Gibbs sampling. In addition to MRF-based models, other probabilistic models [78-80] have been developed under hierarchical structures from gene ontology [64]—which provides gene product annotation data that are characterized into three categories: biological processes, cellular components, and molecular functions—with some models, such as the hierarchical binomial neighborhood model [79]. In particular, Jiang *et al*. [80] considered the auto-probit model with a weighted network information

Table 3. Network-based protein function prediction methods

Approach	Method	Data resource	Reference
Neighborhood-based	Neighbor counting	YPD [59], MIPS [11], CuraGen [60], Ito *et al*. [61]	Schwikowski *et al*. [1]
	χ^2 statistic	YPD [59], MIPS [11], CuraGen [60], Ito *et al*. [61]	Hishigaki *et al*. [62]
	Functional similarity	MIPS [11], GRID [14]	Chua *et al*. [63]
	Protein similarity/ functional similarity	BioGRID [13], GO [64]	Chi and Hou [65]
Graph-based	Label assignment	YPD [59], MIPS [11], CuraGen [60], Ito *et al*. [61]	Vazquez *et al*. [66]
	Label assignment	GRID [14], GO [64]	Karaoz *et al*. [67]
	Functional flow	GRID [14], MIPS [11], GO [64]	Nabieva *et al*. [68]
	Label propagation	MIPS [11]	Tsuda *et al*. [69]
	Label propagation	MIPS [11], GO [64]	Mostafavi *et al*. [70]
	SLNP	GRID [14], MIPS [11], GO [64]	Wang and Li [71]
	NRC	DIP [15], GO [64]	Moosavi *et al*. [72]
Bayesian	MRF	YPD [59], MIPS [11], SGD [73], GO [64]	Deng *et al*. [74, 75]
	BMRF	Collins *et al*. [76], GO [64]	Kourmpetis *et al*. [77]
	Posterior probability	GRID [14], MIPS [11], SGD [73]	Nariai *et al*. [78]
	HBNM	BioGRID [13], GO [64]	Jiang *et al*. [79]
	Auto-probit model	STRING [16], GO [64]	Jiang *et al*. [80]
Kernel-based	SVM	MIPS [11]	Lanckriet *et al*. [81]
	KLR	MIPS [11], SGD [73], GO [64]	Lee *et al*. [82]
	FCML	BioGRID [13], MIPS [11], GO [64]	Wang *et al*. [83]

from STRING [16], and their auto-probit model also reflects the uncertainty of the annotation [64].

The Kernel-based approach considers the protein function prediction problem as a classification problem. To reflect network information into the classification state, the network information is converted into a kernel matrix. Lanckriet *et al*. [81] proposed the kernel-based support vector machine (SVM) method, which incorporates heterogeneous types of data, such as amino acid sequence, gene expression data, and PPI network data, by converting these data into kernel matrices. The SVM method can be reformulated as semi-definite programming (SDP) [86] with kernels. Although the SVM method performs well, this method becomes slow when the dimension increases, caused by the computational complexity of the SDP. To remedy this problem, Lee *et al*. [82] proposed the kernel-based logistic regression (KLR) method by combing MRF-based methods [74, 75] with the diffusion kernel [87]. The KLR can contain multiple functions and various types of datasets at once. It has been shown that the KLR method is faster than the SVM and is comparable to the SVM in prediction accuracy [82]. Recently, Wang *et al*. [83] proposed the function-function correlated multi-label method, which treats all function categories in the prediction at once, while other methods only consider one function at a time, except for KLR.

Disease gene prioritization

Disease-gene association studies have focused on identifying relationships between disease phenotypes and genes to reveal and understand human disease mechanisms. Although traditional approaches, including linkage analysis and association studies, have been successful, the specified genomic intervals often contain tens or even hundreds of genes. It may not be possible by experiments to identify the correct causative genes of all the genes that lie on the specified intervals. To reduce experimental costs and efforts, the prioritization of candidate genes becomes one of the major tasks in disease-gene association studies.

Taking into consideration that genes corresponding to similar disease phenotypes tend to be neighbors in either a PPI network [88, 89] or a pathway [90], several network-based disease gene prioritization methods have been proposed recently. The network-based disease gene prioritization methods [91-94] define different similarity scores between the disease and genes, based on either functional annotations or PPI networks, to rank candidate genes. The random walk method [91] considers the random walk on graphs and uses a diffusion kernel matrix [87] derived from a PPI network as the steady-state transition probability matrix. The random walk method then defines the similarity score as the sum of elements of the diffusion kernel corresponding to known disease genes for each candidate gene. CIPHER [92] adopts a regression model for similarities between phenotypes and considers the Gaussian kernel for closeness between genes. The Pearson correlation coefficient between the similarity of phenotypes and the closeness between genes is used as a similarity score. On the other hand, PRINCE [93] and MINProp [94] are based on the iterative label propagation algorithm [95]. PRINCE defines the prioritization function, which combines functional similarities from the network information with prior information for disease phenotypes as the similarity score. The prioritization function is obtained by iteratively applying a label propagation algorithm. MINProp first defines two heterogeneous networks, such as a gene network and a disease network, and then additionally defines bipartite networks between two heterogeneous networks from known disease-gene associations. After initializing the labels of the disease genes, MINProp iteratively propagates label information through three networks until convergence occurs. Finally, the converged label scores are used as the similarity scores. From the comparison study in [93, 94], PRINCE and MINProp outperform the random walk and CIPHER methods, respectively. There is no comparison study between PRINCE and MINProp, but MINProp is more general than PRINCE, since MINProp can be applied to three or more heterogeneous networks. We report these disease gene prioritization methods with their data resources in Table 4 [12, 13, 15-18, 91-94, 96-99].

Genome-wide association study (GWAS)

The GWAS measures DNA sequence variations in the human genome to identify associations between genetic variants and diseases (or phenotypes). To measure genetic variations, the single- nucleotide polymorphism (SNP),

Table 4. Network-based disease gene prioritization methods

Method	Data resource	Reference
Random walk	OMIM [96], HPRD [12], BIND [97], BioGRID [13], IntAct [18], DIP [15], STRING [16]	Kohler *et al*. [91]
CIPHER	OMIM [96], HPRD [12], OPHID [98], BIND [97], MINT [17]	Wu *et al*. [92]
PRINCE	OMIM [96], GO [64], HPRD [12], GeneCards [99]	Vanunu *et al*. [93]
MINProp	OMIM [96], HPRD [12], OPHID [98], BIND [97], MINT [17]	Hwang and Kuang [94]

which represents a single base-pair change in the DNA sequence, is generally used as a marker of a genomic region in the GWAS. Generally, the GWAS conducts a comparison of the SNPs between case and control groups (i.e., disease and non-disease groups) by statistical methods, such as ANOVA and logistic regression, to identify significant SNPs related to the disease. Genetic risk factors revealed by a GWAS can be used as preventive markers or for therapeutic targets in curing the disease. There have been more than 11,000 SNPs discovered as candidate bio-markers in common diseases [100]. The large number of SNPs detectable in the human genome can, however, lead to multiple testing problems. To control the false positive errors (i.e., type I errors in the context of statistical testing procedures), the Bonferroni correction and false discovery rate [101] methods are commonly adopted. Although these multiple testing rules have been successful in the identification of significant single SNPs, these test procedures often fail to detect genomic regions that are weakly associated with the disease and still ignore the combined effects caused by the interactions between genes.

The network-based GWAS methods [102-104], summarized in Table 5 [12, 13, 17, 102-114], take into account both interactions between genes or proteins with association information available from a GWAS. The HumanNet method [115] combines functional gene networks derived from multiple network sources, such as the PPI network, and mRNA co-expression with the log odds ratio from GWAS data to prioritize disease genes. By combining functional gene networks and the information from GWAS data, HumanNet has higher power to detect disease-related genes. Unlike HumanNet, NIMMI [116] and dmGWAS [117] focus on identifying groups of genes associated with diseases, based on GWAS data and PPI network data. NIMMI constructs a PPI network with weights of genes by using a modified Google PageRank algorithm [118]. The weights of genes are then combined with the gene-based association p-values calculated from GWAS data. The subnetworks of genes are identified by the DAVID method [119, 120] with the network structure and weights of genes. On the other hand, dmGWAS only considers gene-based association p-values as the node weights and proposes a modified version of the dense module searching method [121] to prioritize candidate groups of genes.

Discussion

We have reviewed a number of methods related to two topics in network data analysis: network reconstruction and network-based application. Network reconstruction can be thought of as a reverse-engineering problem whose goal is to rebuild true structures or relationships from observations. In particular, we focused on statistical methods for building GRNs, including the Gaussian graphical model, correlation network, Bayesian network, and information theory-based methods. Most methods that we have reviewed consider the sparsity on the graph structure to select a small number of significant dependencies between nodes. This sparsity condition is adequate for the network in biological systems, since it reflects a scale-free feature, where several nodes have many edges but the majority of nodes only has three or four edges [5]. Although most methods in Gaussian graphical models are well studied in their theoretical properties, these methods have limitations when applied to biological data. Since these methods basically assume a Gaussian distribution, they are only applicable for continuous-type datasets, such as gene expression levels. To construct networks from other types of data, such as binary or counts, the Bayesian network and information theory-based methods are more attractive than correlation-based methods.

Second, we introduced various methods in three network-based applications. Most methods consider similarity measures between nodes and then use these measures to predict biological functions or prioritize candidate genes associated with diseases. As technologies in the experiments progress, the network-based methods can be improved and widely extended. For instance, even though the neighborhood counting method [1] in protein function prediction only considers the count of annotated functions in the neighborhood from a PPI network, the recently developed methods [72, 83] can contain not only neighborhood information but also functional similarities from multiple networks. In addition, module-assisted methods that focus on identifying a functional group of proteins are also

Table 5. Network-based genome-wide association study methods

Method	Data resource	Reference
HumanNet	NCBI [105], GEO [106], SMD [107], BioGRID [13], IntAct [18], BIND [97], MINT [17], HPRD [12], InterPro [108]	Lee *et al*. [102]
NIMMI	Height GWAS data ([109, 110], www.genome.gov/19518664), Crohn's disease GWAS data [111], BioGRID [13]	Akula *et al*. [103]
dmGWAS	Breast cancer study [112], pancreatic cancer study [113], PINA [114]	Jia *et al*. [104]

available and well summarized in [122]. Furthermore, the network-based tumor stratification (NBS) method has been developed recently [123]. This NBS method combines mutation profiles from The Cancer Genome Atlas (TCGA) projects [116, 117, 124, 125] and network information to obtain informative strata (e.g., subtypes of cancer). Although these network-based methods have been improved, these methods still lack accuracy compared with other methods [115]. Since the high-throughput data may contain many false positives [126], the network-based methods are affected in their accuracy. Although their performance depends on the quality of data, their effects are expected to decrease in the future as improvements are made in measurement accuracy.

References

1. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000;18:1257-1261.
2. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27-30.
3. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;33:D428-D432.
4. Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief Bioinform* 2006;7: 243-255.
5. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5: 101-113.
6. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev* 2007;21: 1010-1024.
7. Heller MJ. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng* 2002;4:129-153.
8. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol* 2009;25:195-203.
9. Fields S, Sternglanz R. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet* 1994;10:286-292.
10. Murali T, Pacifico S, Yu J, Guest S, Roberts GG 3rd, Finley RL Jr. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res* 2011;39:D736-D743.
11. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, *et al*. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2000;28:37-40.
12. Prasad TS, Kandasamy K, Pandey A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol* 2009;577:67-79.
13. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;34:D535-D539.
14. Breitkreutz BJ, Stark C, Tyers M. The GRID: the general repository for interaction datasets. *Genome Biol* 2003;4:R23.
15. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 2004;32:D449-D451.
16. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013;41:D808-D815.
17. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;40:D857-D861.
18. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012;40:D841-D846.
19. Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 2007;35:D137-D140.
20. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, *et al*. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 2013;41:D203-D213.
21. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, *et al*. Text mining for biology: The way forward: opinions from leading scientists. *Genome Biol* 2008;9 Suppl 2:S7.
22. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol* 2008;9 Suppl 2:S4.
23. Krallinger M, Valencia A. Text-mining and information- retrieval services for molecular biology. *Genome Biol* 2005;6: 224.
24. Ananiadou S, Pyysalo S, Tsujii J, Kell DB. Event extraction for systems biology by text mining the literature. *Trends Biotechnol* 2010;28:381-390.
25. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306:636-640.
26. Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, *et al*. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* 2011;7:e1002190.
27. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, *et al*. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;489:91-100.
28. Washington NL, Stinson EO, Perry MD, Ruzanov P, Contrino S, Smith R, *et al*. The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details. *Database (Oxford)* 2011;2011:bar023.
29. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. *PLoS One* 2012;7:e29348.
30. Lauritzen SL. *Graphical Models*. Oxford: Clarendon Press, 1996.
31. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. *Ann Stat* 2006;34:1436-1462.
32. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *J Am Stat Assoc* 2009;

104:735-746.
33. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* 2007;94:19-35.
34. Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J Mach Learn Res* 2008; 9:485-516.
35. d'Aspremont A, Banerjee O, El Ghaoui L. First-order methods for sparse covariance selection. *SIAM J Matrix Anal Appl* 2008;30:56-66.
36. Fan J, Feng Y, Wu Y. Network exploration via the adaptive Lasso and Scad penalties. *Ann Appl Stat* 2009;3:521-541.
37. Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. *Electron J Stat* 2008;2: 494-515.
38. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008;9: 432-441.
39. Schäfer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 2005;21:754-764.
40. Cai T, Liu W, Luo X. A constrained l_1 minimization approach to sparse precision matrix estimation. *J Am Stat Assoc* 2011; 106:594-607.
41. Myllymäki P, Silander T, Tirri H, Uronen P. B-course: a web-based tool for Bayesian and causal data analysis. *Int J Artif Intell Tools* 2002;11:369-387.
42. Murphy KP. The bayes net toolbox for Matlab. *Comput Sci Stat* 2001;33:1024-1034.
43. Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics* 2006;22:2523-2531.
44. Carter SL, Brechbühler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 2004;20: 2242-2250.
45. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;4:Article17.
46. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9: 559.
47. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* 2000:418-429.
48. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet* 2005;37:382-390.
49. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, *et al*. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006;7 Suppl 1:S7.
50. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, *et al*. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 2007;5:e8.
51. Fu Y, Jarboe LR, Dickerson JA. Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics* 2011;12:233.
52. Zhang X, Liu K, Liu ZP, Duval B, Richer JM, Zhao XM, *et al*. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* 2013;29:106-113.
53. Kolaczyk ED. *Statistical Analysis of Network Data: Methods and Models*. New York: Springer, 2009.
54. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389-3402.
55. Andrade M, Casari G, de Daruvar A, Sander C, Schneider R, Tamames J, *et al*. Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Comput Appl Biosci* 1997;13:481-483.
56. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* 1998;281:949-968.
57. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 1988;85: 2444-2448.
58. Murali TM, Wu CJ, Kasif S. The art of gene function prediction. *Nat Biotechnol* 2006;24:1474-1475.
59. Costanzo MC, Hogan JD, Cusick ME, Davis BP, Fancher AM, Hodges PE, *et al*. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* 2000;28:73-76.
60. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, *et al*. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;403:623-627.
61. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, *et al*. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* 2000;97:1143-1147.
62. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein: protein interaction data. *Yeast* 2001;18:523-531.
63. Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 2006;22: 1623-1630.
64. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25:25-29.
65. Chi X, Hou J. An iterative approach of protein function prediction. *BMC Bioinformatics* 2011;12:437.
66. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 2003;21:697-700.

67. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, *et al*. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A* 2004;101:2888-2893.
68. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 2005;21 Suppl 1:i302-i310.
69. Tsuda K, Shin H, Schölkopf B. Fast protein classification with multiple networks. *Bioinformatics* 2005;21 Suppl 2:ii59-ii65.
70. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* 2008;9 Suppl 1:S4.
71. Wang J, Li Y. Sequential linear neighborhood propagation for semi-supervised protein function prediction. *J Bioinform Comput Biol* 2011;9:663-679.
72. Moosavi S, Rahgozar M, Rahimi A. Protein function prediction using neighbor relativity in protein-protein interaction network. *Comput Biol Chem* 2013;43:11-16.
73. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, *et al*. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res* 1998;26:73-79.
74. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. *J Comput Biol* 2003;10:947-960.
75. Deng M, Tu Z, Sun F, Chen T. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics* 2004;20:895-902.
76. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, *et al*. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteomics* 2007;6:439-450.
77. Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov Random Field analysis for protein function prediction based on network data. *PLoS One* 2010;5:e9293.
78. Nariai N, Kolaczyk ED, Kasif S. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS One* 2007;2:e337.
79. Jiang X, Nariai N, Steffen M, Kasif S, Kolaczyk ED. Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinformatics* 2008;9:350.
80. Jiang X, Gold D, Kolaczyk ED. Network-based auto-probit modeling for protein function prediction. *Biometrics* 2011; 67:958-966.
81. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. *Bioinformatics* 2004;20:2626-2635.
82. Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. *OMICS* 2006;10:40-55.
83. Wang H, Huang H, Ding C. Function-function correlated multi-label protein function prediction over interaction networks. *J Comput Biol* 2013;20:322-343.
84. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. *Science* 1983;220:671-680.
85. Li SZ. *Markov Random Field Modeling in Computer Vision*. Berlin: Springer-Verlag, 1995.
86. Vandenberghe L, Boyd S. Semidefinite programming. *SIAM Rev* 1996;38:49-95.
87. Kondor RI, Lafferty JD. Diffusion kernels on graphs and other discrete input spaces. In: ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning, 2002 Jul 8-12, Sydney. pp. 315-322.
88. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, *et al*. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 2006;38:285-293.
89. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual JF, *et al*. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 2006; 125:801-814.
90. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, *et al*. The genomic landscapes of human breast and colorectal cancers. *Science* 2007;318:1108-1113.
91. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;82:949-958.
92. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol* 2008;4:189.
93. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;6:e1000641.
94. Hwang T, Kuang R. A heterogeneous label propagation algorithm for disease gene discovery. In: Proceeding of the SIAM International Conference on Data Mining, 2010 Apr 29-May 1, Columbus, OH. pp. 583-594.
95. Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. *Adv Neural Inf Process Syst* 2004;16:321-328.
96. McKusick VA, Antonarakis SE. *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*. Baltimore: Johns Hopkins University Press, 1998.
97. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003;31: 248-250.
98. Brown KR, Jurisica I. Online predicted human interaction database. *Bioinformatics* 2005;21:2076-2082.
99. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet* 1997;13:163.
100. Ulfarsson MO, Solo V. Tuning parameter selection for underdetermined reduced-rank regression. *IEEE Signal Process Lett* 2013;20:881-884.
101. Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likelihood. *J R Stat Soc Series B Stat Methodol* 2013;75:531-552.
102. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 2011;21:1109-1121.
103. Akula N, Baranova A, Seto D, Solka J, Nalls MA, Singleton A,

et al. A network-based approach to prioritize results from genome-wide association studies. *PLoS One* 2011;6:e24220.
104. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics* 2011; 27:95-102.
105. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2009; 37:D5-D15.
106. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, *et al*. NCBI GEO: mining millions of expression profiles: database and tools. *Nucleic Acids Res* 2005;33:D562-D566.
107. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, *et al*. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res* 2003; 31:94-96.
108. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009;37:D211-D215.
109. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, Harris TB, *et al*. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J Am Geriatr Soc* 2000;48:1618-1625.
110. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
111. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661-678.
112. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, *et al*. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007;39:870-874.
113. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, *et al*. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* 2009;41:986-990.
114. Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP, Hautaniemi S. Integrated network analysis platform for protein-protein interactions. *Nat Methods* 2009;6:75-77.
115. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, *et al*. A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;10:221-227.
116. Cancer Genome Atals Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487:330-337.
117. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, *et al*. Integrated genomic characterization of endometrial carcinoma. *Nature* 2013;497:67-73.
118. Wang T, Zhu L. Consistent tuning parameter selection in high dimensional sparse linear regression. *J Multivar Anal* 2011;102:1141-1151.
119. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, *et al*. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;4:P3.
120. Fan J, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. *J R Stat Soc Series B Stat Methodol* 2013;75:603-680.
121. Sangalli LM, Ramsay JO, Ramsay TO. Spatial spline regression models. *J R Stat Soc Series B Stat Methodol* 2013; 75:681-703.
122. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007;3:88.
123. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods* 2013; 10:1108-1115.
124. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;474: 609-615.
125. Cancer Genome Atals Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61-70.
126. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, *et al*. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;417:399-403.