Statistics and Its Interface Volume 5 (2012) 137–148

Constructing human phenome-interactome networks for the prioritization of candidate genes

Yong Chen[*,†], Wangshu Zhang[*], Mingxin Gan[‡] and Rui Jiang[*,§]

Although remarkable success has been achieved by traditional gene-mapping methods in locating genes associated with inherited human diseases, the resulting chromosomal regions are usually large, containing tens or even hundreds of genes. Therefore, it is indispensable to develop computational methods for the identification of genes that are truly responsible for diseases from candidate genes. To tackle this problem, several methods have been proposed to use both a phenotype similarity profile (phenome) and a protein-protein interaction network (interactome) for the prioritization of candidate genes. The use of the phenome broadens the scope of applications of these methods for identifying disease-associated genes, and the use of the interactome provides a reliable measure of functional similarities between genes. These two data sources, together with carefully designed computational models, result in computational methods with superior performance in the prioritization of candidate genes for a given query disease of interest. In this paper, we review recent achievements of such computational methods that rely on the integration of the phenome and the interactome to prioritize candidate genes. We also summarize how similar methods can be readily used in identifying microRNAs that are potentially involved in complex diseases and discovering drugs that may target on disease-associated proteins. Finally, we discuss future prospects and challenges for the integration of multiple genomic data sources to systematically discover genes that underlie human diseases.

Keywords and phrases: Diseases, Genes, Phenome, Interactome, Prioritization, Protein-protein interaction.

1. INTRODUCTION

Knowledge about genes that are associated with human inherited diseases will facilitate the understanding of pathogenesis of human diseases and further benefit the prevention, diagnosis, and therapy of these diseases. The identification of such disease-associated genes therefore becomes a fundamental problem in human and medical genetics. Traditionally, disease-associated genes are identified via statistical methods such as family-based linkage analysis and population-based association studies [1]. However, these methods can only associate a query disease of interest with chromosomal regions that typically contain tens or even hundreds of genes [2]. Consequently, it becomes indispensable to develop computational methods to aid the discovery of genes that are truly responsible for the query disease from a long list of candidate genes.

In the past few years, several methods have been proposed to tackle this problem from the perspective of prioritizing candidate genes. For example, according to the "guilt-by-direct-association" principle, the prioritization is enabled by ranking candidate genes in a susceptibility region according to their relevance to genes that are already known to be associated with the query disease under investigation (i.e., seed genes). Based on this principle, a wide variety of information, including protein sequences [3, 4], gene expression profiles [4–6], functional annotations [6–9], literature descriptions [4, 5, 10], protein-protein interactions (PPI) [5, 6, 11, 12], and many others [13], has been used to facilitate the prioritization of candidate genes.

Nevertheless, the requirement of the presence of a set of seed genes limits the scope of applications of the methods that are based on the guilt-by-direct-association principle, because genetic bases of about half of known human diseases are completely unknown [14], making these methods not applicable to these diseases. To overcome this limitation, a number of recent studies have suggested the "guilt-by-indirect-association" principle, which resorts to the modular nature of human inherited diseases [13, 15–19] and utilizes similarities between disease phenotypes to infer genes that are truly associated with diseases [20–24].

The basic assumption of the guilt-by-indirect-association principle is that phenotypic overlap between two diseases implies genetic overlap between the diseases, and thus genes associated with the diseases would be similar in their functions [25, 26]. The phenotypic overlap between two diseases is typically quantified using a similarity score that can be calculated using text mining techniques [19, 20], and the resulting pairwise similarity profile between every two disease is often referred to as the *phenome*. The functional similarity

[*]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.
[†]Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China.
[‡]School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China.
[§]Corresponding author.

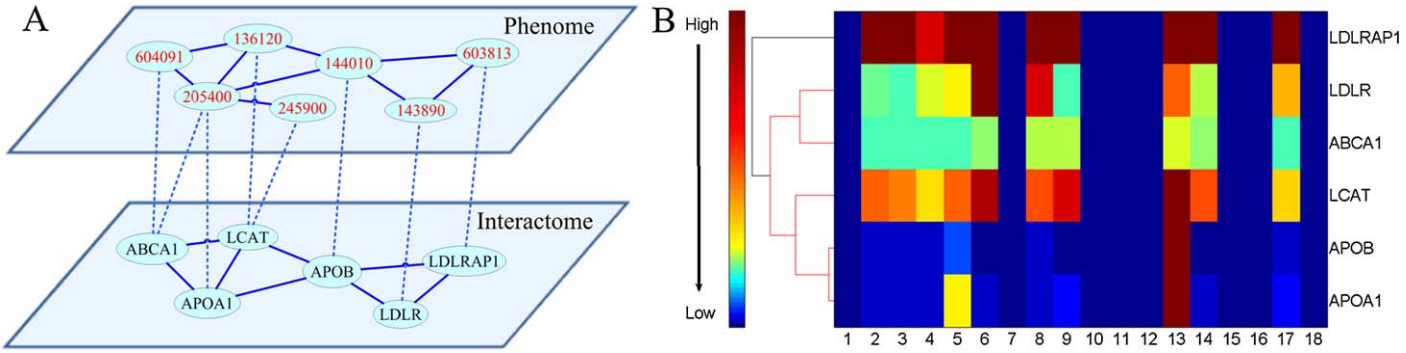

Figure 1. The phenome-interactome network. *A: Red numbers in ellipses indicate accession IDs of diseases in the Online Mendelian Inheritance in Man (OMIM) database, and black words represent gene symbols. An edge between two diseases indicates that the connected diseases share significant phenotypic overlap. An edge between two genes indicates that the corresponding proteins have physical interaction. A dashed line between a disease and a gene indicates a known association between the disease and the gene. 136120: Fish-eye disease, 144010: Hypercholesterolemia, autosomal dominant, type b, 143890: Hypercholesterolemia, autosomal dominant, 604091: Hypoalphalipoproteinemia, primary, 603813: Hypercholesterolemia, autosomal recessive, 205400: Tangier disease, 245900: Lecithin: cholesterol acyltransferase deficiency. B: The expression profile of the genes in 18 human tissues (collected from GeneCards [37]). 1: Bladder, 2: Brain, 3: Bone Marrow, 4: Kidney, 5: Heart, 6: Lung, 7: SalivaryGland, 8: Prostate, 9: Thymus, 10: Whole Blood, 11: Colon, 12: Cervix, 13: Liver, 14: Spleen, 15: Breast, 16: Ovary, 17: Pancreas, 18: Skin. (Color online)*

between two genes is typically quantified using the relatedness of the genes in a PPI network, which is usually referred to as the *interactome*. Furthermore, the combination of a phenome, an interactome, and known associations between diseases and genes is typically referred to as a *phenome-interactome network*. An example of the mapping between the phenome and the interactome in a phenome-interactome network is shown in Figure 1. In the figure, seven example diseases are of high phenotype similarity scores, and genes associated with these diseases genes have physical interactions. Moreover, the diseases of high similarity and the genes having interactions have similar graph structures in the phenome and the interactome, respectively. In addition, the expression values of the genes in 18 normal tissues are also highly correlated. Therefore, the mapping between the phenome and the interactome is essential for systematically exploring not only the molecular complexity of a query disease, but also the molecular relationships between apparently distinct phenotypes [20, 22, 26].

The discovery of relationships between diseases and genes through a phenome-interacteome network is not trivial, because the network is usually far from complete, and noise often exists. To tackle these difficulties, several methods have been proposed. For example, Lage et al. proposed a Bayesian model to integrate a phenotype similarity profile and a PPI network [20]. Wu et al. developed a regression model called CIPHER to explain phenotype similarities using gene proximities in a PPI network [22]. Wu et al. also proposed a method called AlignPI to align the phenotype similarity network against the PPI network [21]. Li and Patra utilized a random walk model called RWRH to simulate the stationary distribution of the strength of associations for genes [23]. Vanunu et al. proposed a network propagation method called PRINCE to mimic the sharing of disease status among genes [24]. These methods have exhibited state-of-the-art performance in finding genes that are associated with diseases.

In this paper, we first present a review of existing methods for constructing phenome-interactome networks. Then, we analyze principles of existing methods for discovering disease-associated genes using phenome-interactome networks. Finally, we analyze potential applications of phenome-interactome networks in the identification of microRNAs that are potentially involved in complex diseases and the discovery of drugs that may target on disease-associated proteins.

2. CONSTRUCTION OF A PHENOME-INTERACTOME NETWORK

A phenome-interactome network is usually constructed by integrating a phenotype similarity profile, a PPI network, and known associations between diseases and genes.

2.1 Phenotype similarity profiles

Clinical traits of human disease phenotypes have been recognized and collected in databases such as the Online Mendelian Inheritance in Man (OMIM) [14]. In the recent version of this database (Released on January 23, 2011), 6,675 human disease phenotypes have been collected, and descriptions of clinical traits of each of these diseases have been provided. Therefore, to calculate similarity scores between disease phenotypes, one needs to first extract clinical characteristics of the diseases automatically from OMIM

record using text analysis techniques. Currently, there have been three methods for this purpose, based on 1) the Medical Subject Headings (MeSH), 2) the Unified Medical Language System (UMLS), and 3) the Human Phenotype Ontology (HPO).

van Driel et al. proposed to use the anatomy (A) and the disease (C) sections of the medical subject headings vocabulary (MeSH) to extract terms from the OMIM database, thus providing a standard way of presenting the OMIM records as corresponding phenotype feature vectors [19]. As a result, each disease phenotype was characterized by a vector of standardized and weighted phenotypic feature terms mapped from corresponding OMIM records in the full text (TX) and clinical synopsis (CS) fields. Then, for each pair of disease phenotypes, a similarity score was calculated as the cosine of the angle between their corresponding feature vectors. The reliability of the phenotype similarity score was tested, showing that these similarities were positively correlated with a number of measures of gene functions [19]. The final phenotype similarity network contained pairwise similarity scores for 5,080 OMIM records, covering a majority of recorded human disease phenotypes.

Lage et al. proposed to quantify each OMIM records using a phenotype vector that was composed of weighted medical terms by parsing the clinical synopsis and mapping text into the metathesaurus (MTH) concepts of the Unified Medical Language System (UMLS) [20]. Then, the phenotype vector was normalized and projected onto a medical term space. In this space, the cosine of the angle between each normalized vector pair was calculated to quantify the pairwise phenotypic overlap between records. Implementing the above process for all combinations of vector pairs, a matrix of pairwise phenotypic similarity scores between all OMIM records was finally obtained. To demonstrate the reliability of the resulting phenotype similarity scores, the authors fitted a calibration curve of the scores against the overlaps between the OMIM record pairs. With this curve, the authors showed a direct correlation between the phenotype similarity scores of records measured by the text-mining scheme and the probability that the records had been independently evaluated to have a phenotypic overlap by the OMIM curators. The constructed phenotype vectors and scoring scheme were therefore verified to indeed produce a reliable measure of phenotypic overlap between OMIM records.

Although it is obvious that accurate and clear clinical descriptions of a disease should have positive contribution to the understanding of molecular pathophysiology of the disease, the terms that clinicians have used to describe phenotypic manifestations have been evolving in an uncoordinated manner and thus are hard to be processed by computers automatically. To overcome this drawback, the Human Phenotype Ontology (HPO) was constructed by using ontological concepts to represent clinical attributes of human diseases in the form of a directed acyclic graph [27, 28]. The latest version of HPO contained over 9,500 terms and about 50,000 annotations of these terms, describing phenotypic features of 4,779 human diseases collected in the OMIM database [29]. HPO was originally constructed using data from OMIM, whereby synonyms were merged and semantic links were created between the terms to generate the ontological structure. Every term in HPO described a distinct phenotypic abnormality, and all the terms of HPO were arranged in a hierarchical structure representing subclass relationships. Phenotypic information in the form of ontology could be captured to exploit the semantic relationships between terms. A most used method was PhenExplorer, which was available at the web site of HPO [27]. This method measured the specificity of a term as the information content (IC), which was defined as the negative natural logarithm of the frequency of occurrence of the term, e.g., $-\log p(t)$. The similarity between two terms t_1 and t_2 could then be calculated as the IC of their most informative common ancestor, as

$$\mathrm{sim}(t_1, t_2) = \max_{a \in A(t_1, t_2)} \{-\log p(a)\},$$

where $A(t_1, t_2)$ denoted the set of all common-ancestor terms t_1 and t_2. For individual diseases that were annotated to have multiple phenotypic features, the similarity between diseases d_1 and d_2 was defined as

$$\mathrm{sim}(d_1, d_2) = \frac{1}{2|D_1|} \sum_{s \in D_1} \max_{t \in D_2} \mathrm{sim}(s, t) + \frac{1}{2|D_2|} \sum_{s \in D_2} \max_{t \in D_1} \mathrm{sim}(s, t),$$

where D_1 and D_2 denoted the sets of annotations for disease d_1 and d_2, respectively. Compared with the other two strategies for inferring similarities among diseases from OMIM or UMLS, an advantage of HPO was that the terms and structure of the ontology were based on medical knowledge rather than on text-mining systems. A pre-computed matrix containing pairwise similarity scores between $4,779$ human diseases was provided in the web site of HPO (http://www.human-phenotype-ontology.org).

2.2 Protein-protein interaction networks

An interactome is a collection of all genes and interactions between the genes. Typically, an interactome is constructed using a PPI network. There have been a few PPI networks with diverse coverage and quality. For example, the Human Protein Reference Database (HPRD) contains human protein-protein interactions that have been manually extracted from the literature by expert biologists [30]. In release 8 of this database, 36,634 interactions between 9,470 human genes have been collected. The Biological General Repository for Interaction Datasets (BioGRID) contains protein and genetic interactions of major model organism species [31]. In version 2.0.63 of this database, 29,558

interactions between 9,043 human genes have been collected. The Biomolecular Interaction Network Database (BIND) contains both high-throughput and manually curated interactions between biological molecules [32]. There have been 14,955 interactions between 6,089 human genes collected in this database. The IntAct molecular interaction database (IntAct) contains protein-protein interactions derived from literature [33]. There have been 30,030 interactions between 6,775 human genes collected in this database. The Molecular INTeraction database (MINT) contains information about physical interactions between proteins [34]. There have been 15,902 interactions between 7,200 human proteins collected in this database.

2.3 Known associations between diseases and genes

Known associations between diseases and genes can be extracted from several databases, including OMIM [14], LocusLink [35], HGMD (The Human Gene Mutation Database) [36] and GeneCards [37]. There have also been bioinformatics tools developed to facilitate the retrieval of these databases. For example, BioMart [38] is a convenient tool for extracting associations between diseases and genes from the OMIM database.

3. PRIORITIZATION OF CANDIDATE GENES

There have been a few methods that use the phenome-interactome network with different probabilistic models for the prioritization of candidate genes [12, 20–24, 39]. In the following sections, we will briefly review these methods.

3.1 Bayesian prediction methods

Grounded on a widely acceptable assumption that mutations existing in several genes of a functional module could lead to phenotypes with overlapping clinical manifestations, Lage et al. proposed the use of a Bayesian prediction method to perform a large-scale prioritization of protein complexes (comprising of gene products) that were implicated in many categories of human diseases [20]. This method first created a phenome-interactome network by integrating quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score, and obtained a benchmark set composed of 963 genes and 1,404 distinct phenotypes. They then utilized a five-fold cross-validation experiment to validate the proposed Bayesian predictor on the benchmark set that was composed of known disease-associated genes and their neighbors in linkage intervals. Specifically, for each phenotype in the benchmark set, they assigned proteins corresponding to candidate genes (including one known disease-associated gene and several genes located nearby) to protein complexes, ranked these complexes based on the phenotypes assigned to their members by text mining, and then computed for each candidate in a critical interval, indicating the posterior probability that the protein was related to the disease of interest. According to their validation experiments, among all predictions in which the top-ranked proteins were scored above 0.1, about 45% of known disease-associated proteins could be successfully recovered, while among all predictions in which the top-ranked proteins were scored above 0.9, about 65% of known disease-associated proteins could be successfully recovered, demonstrating that high-scoring candidates were very likely to be correct. Using this trained Bayesian predictor, Lage et al. further implemented a prediction towards diseases for which there were no confirmed disease-associated genes.

As was mentioned, the success of their method was mainly owing to the integration of high-confidence protein interaction data with a phenotype similarity scheme. In contrast, the most common failure of their method obstructing the correct identification of the disease genes lied in the lack of interaction partners involved in similar phenotypes.

3.2 Regression-based methods

In order to capture the relationship between phenotype similarities of diseases and functional similarities of genes in a systematic way, Wu et al. proposed a regression model called CIPHER to explain relationships between phenotype similarities of diseases using network proximities of gene products, and then calculated a global concordance score to measure the strength of association between a candidate gene and a given query disease of interest [22].

In detail, the CIPHER model assumed that the phenotype similarity between a pair of diseases had a linear relationship with the overall proximity between genes that were associated with the diseases. With this assumption, they proposed the following regression model

$$S_{pp'} = C_p + \sum_{g \in G(p)} \sum_{g' \in G(p')} \beta_{pg} \exp(-L_{gg'}^2),$$

where $S_{pp'}$ was the similarity score between a query phenotype p and another phenotype p', and $L_{gg'}$ was the topological distance between genes g and g' on the protein network. $G(p)$ denoted all disease genes associated with the phenotype p. The Gaussian kernel $\exp(-L_{gg'}^2)$ was used to transfer distances between genes to similarities between genes. C_p was a constant, and β_{pg} was the coefficient of this regression model.

In this work, functional similarity between a pair of genes was quantified using the network proximity between the corresponding proteins in a PPI network. In detail, CIPHER adopted two similarity measures, depending on how indirect interaction was considered. The first measure was called SP (meaning shortest path), in which $L_{gg'}$ was defined as the length of the shortest path between proteins corresponding to genes g and g' in the underlying PPI network. The second measure was called DN (meaning direct neighbor), in which $L_{gg'}$ was set to $+\infty$ if the corresponding proteins of the two genes did not have direct interaction in the PPI network.

To quantify the association between a phenotype and a gene, CIPHER defined the closeness of gene g to phenotype p' as the summation of similarities between gene g to all genes that were associated with the disease phenotype p' as

$$\Phi_{gp'} = \sum_{g' \in G(p')} \exp(-L^2_{gg'}).$$

Then, with the phenotype similarity profile of a given disease phenotype p defined as $S_p = (S_{pp_1}, \ldots, S_{pp_n})$ and the gene closeness profile defined as $\Phi_g = (\Phi_{gp_1}, \ldots, \Phi_{gp_n})$, the strength of association between the disease phenotype p and the gene g was calculated as the concordance score

$$CS_{pg} = \frac{\mathrm{cov}(S_p, \Phi_g)}{\sigma(S_p)\sigma(\Phi_g)},$$

with cov and σ being covariance and standard deviation, respectively.

It was demonstrated that CIPHER can achieve superior performance to the Bayesian prediction approach. This method was applicable to genetically uncharacterized phenotypes, effective in the genome-wide scan of disease genes, and also extensible to explore cooperativity among genes in complex diseases [22]. A better regression strategy was still desired to overcome the limitations of noise and incompleteness of biological data.

3.3 Alignment-based methods

In modeling complex biological systems, graph topology could reveal the basic properties of connectivity, robustness, modularity, hierarchical structure, and other properties. Since similarities between diseases showed overlap with similarities between genes that were associated with the diseases, the topological structures between the phenome and the interactome might also be similar. With this consideration, Wu et al. proposed an approach called AlignPI to directly align the phenome network with the interactome network using the network alignment technology [21]. This method worked as follows. First, input networks were assembled into a weighted alignment graph, and a likelihood ratio model was used to score subnetworks in the alignment graph. Second, the scoring model compared the compatibility of a subnetwork with a desired structure (linear path or clique) against the likelihood of the subnetwork, given that input networks were randomly constructed. Finally, an algorithm was used to search exhaustively over the alignment graph to identify subnetworks with high scores. It was reported that AlignPI could achieve higher performance than CIPHER. AlignPI also scaled well to the whole genome, as demonstrated by prioritizing 6,154 genes across 37 chromosome regions for Crohn's disease (CD).

Another computational tool called MCDGPA used a graph partition method to prioritize disease genes in three steps: module partition, genes prioritization in each disease-associated module, and rank fusion for the global ranking [39]. MCDGPA was tested on a prostate cancer and breast cancer network, significantly improving previous algorithms in terms of cross-validation and disease-associated gene prediction. In addition, the improvement was robust to the selection of gene prioritization methods, suggesting that MCDGPA was a general framework that allowed integrating many previous gene prioritization methods and improving predictive accuracy.

It is usually difficult to perform an optimal partition of the underlying network in this category of methods, because the number of possible partitions is huge, and the optimality of a partition relies only on genes in the partition and thus is not a global criterion. To overcome this limitation, a class of diffusion-based methods has been proposed recently, as will be introduced in the following section.

3.4 Diffusion-based methods

A random walk model was introduced by Kohler et al. to use only the interactome to facilitate the prioritization of candidate genes [12]. In this model, the random walk process on a graph was defined as iterative transitions of a walker from its current location to a neighboring location that was selected at random, when starting from a given source node. This process could further be modified to allow a restart at every time step with a certain probability. Formally, let s be the starting source node, and γ the restart probability. The random walk with a restart was defined as $p^{(t+1)} = (1-\gamma)Wp^{(t)} + \gamma p^{(0)}$, where W was the column-normalized adjacency matrix of the graph, and $p^{(t)}$ was a vector in which the i-th element was the probability of being at node i at time step t. The stationary probability $p^{(\infty)}$ could then be used to rank candidate genes.

The above random walk strategy was then extended by Li and Patra to include the phenome, resulting a method called RWRH (Random Walk with Restart on Heterogenous network) [23]. The basic idea of their method was to construct a heterogenous network that included both the phenome and the interactome, and then performed the random walk with restart method on the heterogenous network. A recent work compared this model with other seven computational methods and confirmed the outstanding performance of this approach [40].

Recently, Vanunu et al. proposed a method called PRINCE to associate genes and protein complexes with diseases via a mechanism called network propagation [24]. This method was based on formulating constraints on the prioritization function that related to its smoothness over network and usage of prior information. This function was also exploited to predict the association of not only genes but also protein complex with a disease of interest. Overall the method generalized the network-based approaches by considering the network signal in a global manner and going beyond single genes to the modules that were affected in a given disease. It was claimed that PRINCE could achieve an even higher performance than RWRH and CIPHER. Although PRINCE got promising results, there were still several limitations to be acknowledged. First, the application

of PRINCE is restricted to such diseases that are phenotypically similar to diseases with known causal genes, because this method belongs to the supervised learning category and thus needs to use the causal genes as the training data. Second, this method highly depends on accurate and comprehensive protein-protein interaction data, but it is obvious that available PPI datasets are both noisy and far from complete. Last, the method is slower than other strategies such as the random walk.

3.5 Network flow-based methods

Although both the random walk and the network propagation models had achieved great successes in prioritizing candidate genes, these methods were usually computationally intensive [41]. To overcome this limitation, Chen et al. proposed the first combinatorial approach called MAXIF (MAXimum Information Flow) for prioritizing candidate disease genes [41]. This method first constructed a flow network by integrating the phenome, the interactome, and known associations between diseases and genes. Then, it calculated the strength of association between a query disease and a candidate gene as the amount of information that could flow from the disease to the gene, and further prioritized candidate genes according to their association scores. Validation results showed that MAXIF could achieve higher performance than other computational methods, including RWRH and PRINCE. Besides, MAXIF was also more computationally economy than both RWRH and PRINCE.

Chen et al. further introduced two interesting applications of MAXIF. In the first application, they studied the problem of identifying diseases with which a given query gene might be associated, and they showed that the MAXIF method was effective in solving this problem. In the second application, they explored the possibility of inferring driver genes in cancer studies. It was known that copy number aberrations (CNAs) had great influence on many biological processes involved in many diseases. Particularly, this typical type of genomic variation had been found frequently in cancers, probably due to genomic instability. Typically, genes located in a copy number aberration region could be classified into two categories according to their contributions in a cancer: "driver genes" that were causally implicated in oncogenesis and "passenger genes" that had no contribution to the development of the cancer. Nevertheless, how to identify driver genes from a copy number aberration region was still an open problem that appeals for computational methods. In their studies, Chen et al. showed the capability of MAXIF in predicting driver genes by a case study on a copy number aberration data set of melanoma. In 50 regions of copy number aberration, MAXIF successfully predicted 47 possible driver genes, among which 2 had been validated as driver genes in literature [42]. Chen et al. also analyzed the modularity properties of other predicted driver genes.

3.6 Validation methods and evaluation criteria

Typically, the capability of a method in uncovering genes that are known to be associated with certain diseases (i.e., disease genes) from a set of candidate genes can be validated through leave-one-out cross-validation experiments. Depending on the strategy for selecting candidate genes, there are often three cross-validation schemes. First, in the validation against random genes, one takes a known association between a gene and a disease in each run, assumes that the association is unknown, and prioritizes the gene against a set of control genes that are selected at random from all genes in the interactome. The number of control genes is usually set to 99. Second, in the validation against a linkage interval, one selects control genes in each validation run as all genes that are located within the 10M bp region centered around the disease gene under consideration. Third, in the validation against the whole genome, one selects control genes as all genes in the interactome.

It is possible that a disease is associated with multiple genes. This situation is common for complex diseases. Intuitively, the inclusion of known relationships between a query disease and all its associated genes may facilitate the identification of novel genes that are associated with the disease. To eliminate such a confounding factor, one can perform *ab initio* predictions to examine the capability of a method in discovering genes that are associated with a disease whose genetic basis is completely unknown. Specifically, in an *ab initio* prediction, one considers a known association between a gene and a disease, assumes that the association is unknown, and prioritizes the gene against a set of control genes. In this procedure, one should also remove all known associations between the disease and other genes. Similar to the leave-one-out cross-validation experiments, one can also use three control sets, random genes, a linkage interval, and the whole genome.

There are three measures commonly used to evaluate the performance of a prioritization method. Taking the cross-validation against random genes as an example, after each validation run, one is able to obtain a ranked list and calculate rank ratios of genes by dividing their ranks with the number of genes in the list. For a set of validation runs, one can then calculate the following measure. First, one can calculate the proportion of top ranking disease genes and obtain a measure named the precision (PRE). Second, one can calculate the mean rank ratio (MRR) of all disease genes as the average of rank ratios of all disease genes in the validation runs. Third, given a threshold of rank ratio, one can calculate the sensitivity as the fraction of disease genes ranked above the threshold and the specificity as the fraction of control genes ranked below the threshold. Varying the threshold value from 0.0 to 1.0, one is able to draw a receiver operating characteristic (ROC) curve and further calculate the area under this curve (AUC). Obviously, larger

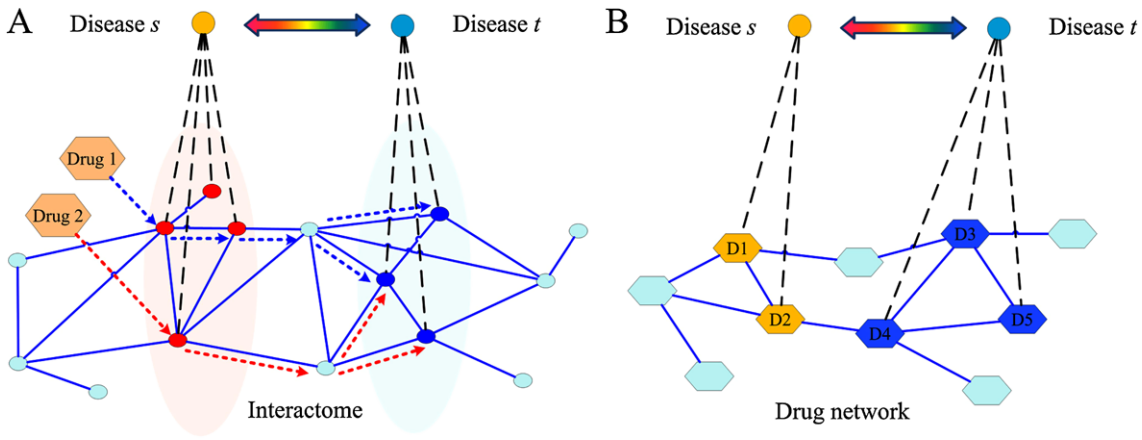

Figure 2. Drug effects through the interactome, the phenome, and the drug network. *A: drug 1 and drug 2 of disease s may show effects on genes associated with disease t through the interactome (dotted arrows show the affected paths). B: the drugs (D1, D2) of disease s may show structure similarity to the drugs (D3, D4, D5) of disease t.*

PRE/AUC values and smaller MRR value indicate higher performance of a prioritization method.

4. APPLICATIONS OF NETWORK-BASED KNOWLEDGE OF DISEASES

4.1 Micro RNAs involved in diseases

MicroRNAs (miRNAs) are small RNAs that are about 22 nucleotides long and are involved in the control of gene expression. Several studies have found that miRNAs play important roles in the development and progression of human diseases, and thus are critical for the prognosis, diagnosis and evaluation of treatment responses for these diseases [43–47]. Over 1,000 miRNAs have been proposed in two manually curated databases, the miR2Disease [48] and the Human MicroRNA Disease Database (HMDD) [49], providing a comprehensive resource of experimentally verified miRNA-disease associations. A computational approach has also been proposed to infer potential miRNA-disease associations based on a phenome-miRNAome network that is similar to the phenome-interactome network [50]. The key step is therefore to construct a network of functionally related miRNAs. Jiang et al. assumed that two miRNAs were functionally related if the overlap between their target genes was statistically significant and proposed to use a p-value from Fisher's exact test to evaluate such overlap [50].

Certainly, there are several limitations existing in the above method. First, the known experimentally verified miRNA-disease associations are insufficient. Second, the network of functionally related miRNAs is constructed based on the perspective that two miRNAs are functionally related if the number of their shared target genes is statistically significant. In reality, however, two miRNAs might be functionally related when their target genes reside in the same cellular pathways or functional modules [45, 51], rather than overlap significantly. Therefore, integrating other information sources such as functional annotations of genes and known interactions between proteins might be helpful. Third, although the above method has achieved positive results, it is still a long way to predict possible association of miRNAs and human diseases. For example, the database of known miRNAs is far from complete, and the relationship between miRNAs is not rigorously defined yet. Considering these facts, the methods used in gene prioritization should be used carefully in the identification of potential associations between miRNAs and human diseases.

4.2 Drug discovery

For many diseases, there have been increasing pieces of evidence supporting the fact that genes associated with the diseases often have dense connections in some biological interaction network [52]. This perspective has been shifting the paradigm of drug discovery from methods that focus on individual genes towards approaches that are based on biological networks and/or pathways [53–55]. It is also believed that biological systems, such as disease states, are generally resistant to perturbations and are able to maintain their functions through different mechanisms such as redundancy and diversity [52, 56]. Therefore, the selection process for new putative drug targets should also consider the network positioning [57]. As is shown in figure 2A, the drugs that treat a disease may have potential affects on another similar disease. Mapping the phenome and the interactome to achieve the selection of these nodes will allow the consideration of the robustness of the system, which is not possible in approaches based on individual genes [52, 58, 59].

As shown in figure 2B, the drug network can also be mapped to the phenome network to discover new drugs for query diseases. In this sense, the methods used in the prioritization of candidate genes can be similarly used to prioritize

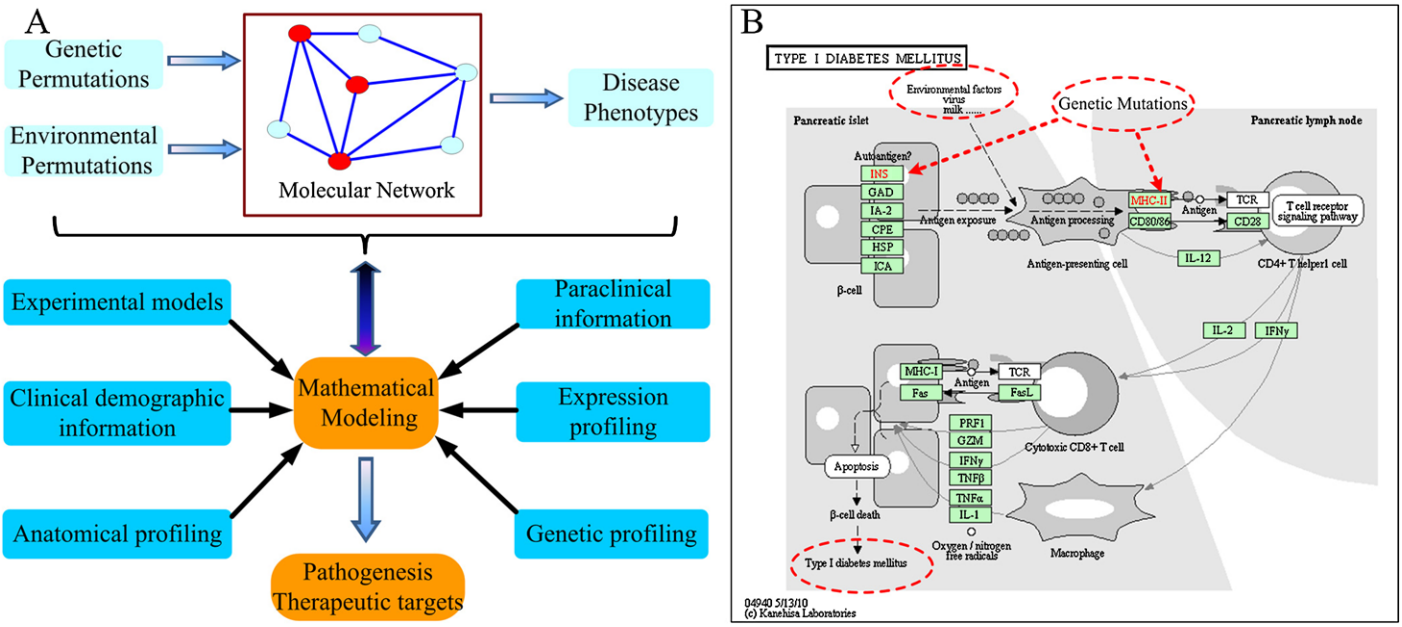

Figure 3. Analysis of the molecular network underlying complex diseases (the disease pathway). *A: Illustration of the integration of information leading to the discovery of the molecular network underlying complex diseases. B: an example disease pathway (hsa04940) constructed for Type I Diabetes Mellitus (downloaded from the KEGG database).*

candidate drugs. For this purpose, the drug network and known drug-disease associations are needed. For instance, PROMISCUOUS has been designed to collect a comprehensive set of drugs, including withdrawn or experimental drugs annotated with drug-protein, and protein-protein relationships compiled from public resources via text and data mining [60]. Based on the similarities of ligand-set, drug networks can be calculated with either an approach called Similarity Ensemble Approach (SEA) or a method derived from Bayesian statistics [51, 61]. Furthermore, measures of structural similarity for drugs as well as known side-effects can also be connected to protein-protein interactions to establish and analyze networks responsible for multi-pharmacology [60].

4.3 Beyond the inference of disease genes

The ultimate goal of identifying genetic factors that are responsible for complex diseases is to identify and characterize biological pathways and processes critical to the disorder [62]. As shown in figure 3, diseases are viewed as perturbed states of the molecular system, and multiple databases can be integrated and modeled to construct disease pathways. Large scale relational databases collect experimental data from different profiling platforms that interrogate DNA, RNA, proteins, post-translational modifications and metabolites, among others, for mining and analysis. Together with clinical and paraclinical records, realistic mathematical and statistical models can be developed to accurately reproduce the molecular network underlying disease [62–64]. Extensive training of such models should allow the development of increasingly accurate predictions of differential diagnosis, treatment outcome and drug-associated toxicity. The construction of disease pathways is a challenging topic in computational systems biology after disease gene prioritization.

5. CONCLUSIONS AND DISCUSSION

In this paper, we have reviewed recent computational methods that rely on a phenotype similarity profile (phenome) and a PPI network (interactome) to prioritize candidate genes. The use of the phenotype similarity profile broadens the scope of applications of computational methods for identifying disease-associated genes, and the use of the PPI network provides a simplified yet systematic measure of functional similarities between genes. These two data sources, together with carefully designed probabilistic models, result in computational methods with superior performance in the prioritization of candidate genes for a given query disease of interest. However, the disadvantage of relying on a single PPI network to estimate functional similarities between genes is also obvious. It is known that PPI networks are noisy and far from complete, and thus relying on a single PPI network can only cover a limited number of known human genes. A possible way to overcome the second limitation is to incorporate multiple PPI networks into current methods that use a single PPI network. For example, Zhang et al. have proposed a Bayesian regression approach that can be used with either a single PPI network or multiple networks to prioritize candidate genes [65]. When used with

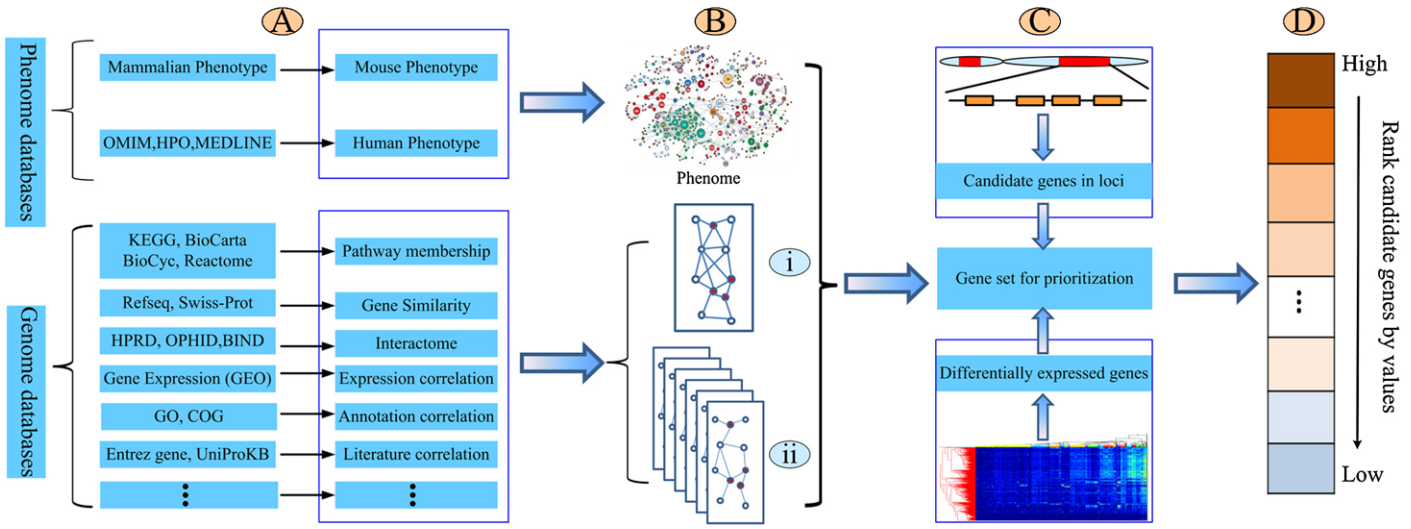

Figure 4. A typical workflow of integrating multiple data sources to the prioritization of candidate genes. *A: Genome and phenome knowledge sources considered to create different relationships among diseases/genes. B: Similarities between diseases are calculated, and a phenome network is constructed as a weighted graph. Similarities between genes can be calculated in two ways. (i) The relationships of gene pairs in all databases are combined as one final relationship and then a combined functional network is constructed. (ii) The relationship of a gene pair in each database is calculated individually and multiple genotype networks are constructed. C: the genes collected from linkage analysis or differentially expressed genes from microarray experiments are used as the test gene set. D: candidate genes are ranked by using the calculated values outputted by computational tools.*

a single PPI network, this method outperforms CIPHER in terms of several statistical criteria. When used with multiple PPI networks, this method can cover much more genes while keeping the superior performance. Moreover, as a simple yet effective method to integrate multiple PPI networks into a single prioritization model, the Bayesian regression approach sheds light on integrating multiple data sources into the prioritization of disease genes in the future.

Certainly, besides the use of PPI networks, there are also many other genomic data available for estimating functional similarities between genes. These data sources include sequence similarity (BLAST) [66], literature (abstracts in EntrezGene) [67], functional annotation (Gene Ontology) [68], microarray expression (Atlas gene expression) [69], EST expression (EST data from Ensembl) [70], protein domains (InterPro) [71], protein-protein interactions (HPRD or BIND) [72], pathway membership (KEGG) [73], cis-regulatory modules (TOUCAN) [74], and transcriptional motifs (TRANSFAC) [75], and many others. The scientific question is then how to integrate these data sources to achieve a more precise estimation of functional similarities between genes. For this purpose, Guan et al. have proposed to construct a global functional network from the perspective of supervised learning through a Bayesian integration of diverse genetic and functional genomic data [76]. In their work, such a global functional network for the laboratory mouse is constructed, resulting in a network named MouseNet. Based on this network, people can do many valuable explorations such as querying biological processes or pathways participated by certain proteins, identifying disease-associated genes, analyzing the network topology and genome evolutionary history, and many others. In a more specific way, a general scheme of integrating multiple data sources for the prioritization of candidate genes is given in Figure 4.

The integration of multiple data sources can also be done from the viewpoint of unsupervised learning. For example, Ma et al. propose a method called CGI to integrate a PPI network and a gene expression profile for the prioritization of candidate genes [77]. Aerts et al. propose a method called Endeavour that uses order statistics to combine multiple ranking lists obtained using different data sources [4]. In the above ensemble strategies, different data sources should be given different weights in the decision of the combined ranking, since different data sources differ in their usefulness and suitability to rank candidate genes for a certain disease family. In most cases, the optimal parameters are designed by cross validation [78]. Therefore, an important topic is how to optimize the different contributions in an ensemble system.

ACKNOWLEDGEMENTS

This work was partly supported by the National Basic Research Program of China (2012CB316504), the National High Technology Research and Development Program of China (2012AA020401), the National Natural Science Foundation of China (61175002, 71101010, 60805010,

10926027), Tsinghua University Initiative Scientific Research Program, Tsinghua National Laboratory for Information Science and Technology (TNLIST) Cross discipline Foundation, and China Post-doctoral Science Foundation (20090450396, 2010003119). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Received 30 March 2011

REFERENCES

[1] Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33**(Suppl) 228–237.

[2] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265) 747–753.

[3] Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6** 55.

[4] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al. (2006). Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**(5) 537–544.

[5] van Driel, M. A., Cuelenaere, K., Kemmeren, P. P., Leunissen, J. A., and Brunner, H. G. (2003). A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet* **11**(1) 57–63.

[6] Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., et al. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* **78**(6) 1011–1025.

[7] Freudenberg, J. and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18**(Suppl 2) S110–115.

[8] Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat Genet* **31**(3) 316–319.

[9] Turner, F. S., Clutterbuck, D. R., and Semple, C. A. (2003). Pocus: Mining genomic sequence annotation to predict disease genes. *Genome Biol* **4**(11) R75.

[10] Gaulton, K. J., Mohlke, K. L., and Vision, T. J. (2007). A computational system to select candidate genes for complex human traits. *Bioinformatics* **23**(9) 1132–1140.

[11] Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *J Med Genet* **43**(8) 691–698.

[12] Kohler, S., Bauer, S., Horn, D., and Robinsonm, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* **82**(4) 949–958.

[13] Oti, M. and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clin Genet* **71**(1) 1–11.

[14] McKusick, V. A. (2007). Mendelian inheritance in man and its online version, omim. *Am J Hum Genet* **80**(4) 588–604.

[15] Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., et al. (2007). The human disease network. *Proc Natl Acad Sci USA* **104**(21) 8685–8690.

[16] Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., et al. (2006). A protein-protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell* **125**(4) 801–814.

[17] Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat Rev Genet* **8**(12) 921–931.

[18] Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science* **318**(5853) 1108–1113.

[19] 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**(5) 535–542.

[20] Lage, K., Karlberg, E. O., Storling, Z. M., Olason, P. I., Pedersen, A. G., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**(3) 309–316.

[21] Wu, X., Liu, Q., and Jiang, R. (2009). Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* **25**(1) 98–104.

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

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

[24] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* **6**(1) e1000641. MR2601382

[25] Thorisson, G. A., Muilu, J., and Brookes, A. J. (2009). Genotype-phenotype databases: Challenges and solutions for the post-genomic era. *Nat Rev Genet* **10**(1) 9–18.

[26] Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* **461**(7261) 218–223.

[27] Robinson, P. N., Kohler, S., Bauer, S., Seelow, D., Horn, D., et al. (2008). The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**(5) 610–615.

[28] Kohler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dolken, S., et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* **85**(4) 457–464.

[29] Robinson, P. N. and Mundlos, S. (2010). The human phenotype ontology. *Clin Genet* **77**(6) 525–534.

[30] Keshava Prasad, T. S., Goel, R., Kandasamy, K. Keerthikumar, S., Kumar, S., et al. (2009). Human protein reference database–2009 update. *Nucleic Acids Res* **37**(Database issue) D767–772.

[31] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., et al. (2006). Biogrid: A general repository for interaction datasets. *Nucleic Acids Res* **34**(Database issue) D535–539.

[32] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., et al. (2005). The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res* **33**(Database issue) D418–424.

[33] Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., et al. (2010). The intact molecular interaction database in 2010. *Nucleic Acids Res* **38**(Database issue) D525–531.

[34] Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., et al. (2010). Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**(Database issue) D532–539.

[35] Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. (2000). Introducing refseq and locuslink: Curated human genome resources at the ncbi. *Trends Genet* **16**(1) 44–47.

[36] Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., et al. (2009). The human gene mutation database: 2008 update. *Genome Med* **1**(1) 13.

[37] Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., et al. (2003). Human gene-centric databases at the weizmann institute of science: Genecards, udb, crow 21 and horde. *Nucleic Acids Res* **31**(1) 142–146.

[38] Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., et al. (2009). Biomart–biological queries made easy. *BMC Genomics* **10** 22.

[39] Chen, X., Yan, G. Y., and Liao, X. P. (2010). A novel candidate disease genes prioritization method based on module partition and rank fusion. *OMICS* **14**(4) 337–356.
[40] Navlakha, S. and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**(8) 1057–1063.
[41] Chen, Y., Jiang, T., and Jiang, R. (2011). Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics* **27** i167–176.
[42] Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., et al. (2010). An integrated approach to uncover drivers of cancer. *Cell* **143**(6) 1005–1017.
[43] Calin, G. A., and Croce, C. M. (2006). Microrna signatures in human cancers. *Nat Rev Cancer* **6**(11) 857–866.
[44] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., et al. (2005). Microrna expression profiles classify human cancers. *Nature* **435**(7043) 834–838.
[45] Bartel, D. P. (2004). Micrornas: Genomics, biogenesis, mechanism, and function. *Cell* **116**(2) 281–297.
[46] Borel, C. and Antonarakis, S. E. (2008). Functional genetic variation of human mirnas and phenotypic consequences. *Mamm Genome* **19**(7–8) 503–509.
[47] Sethupathy, P. and Collins, F. S. (2008). Microrna target site polymorphisms and human disease. *Trends Genet* **24**(10) 489–497.
[48] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., et al. (2009). Mir2disease: A manually curated database for microrna deregulation in human disease. *Nucleic Acids Res* **37**(Database issue) D98–104.
[49] Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., et al. (2008). An analysis of human microrna and disease associations. *PLoS One* **3**(10) e3420.
[50] Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., et al. (2010). Prioritization of disease micrornas through a human phenome-micrornaome network. *BMC Syst Biol* **4**(Suppl 1) S2.
[51] Keiser, M. J. and Hert, J. (2009). Off-target networks derived from ligand set similarity. *Methods Mol Biol* **575** 195–205.
[52] Jia, J., Zhu, F., Ma, X., Cao, Z., Li, Y., et al. (2009). Mechanisms of drug combinations: Interaction and network perspectives. *Nat Rev Drug Discov* **8**(2) 111–128.
[53] Russell, R. B. and Aloy, P. (2008). Targeting and tinkering with interaction networks. *Nat Chem Biol* **4**(11) 666–673.
[54] Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat Rev Genet* **12**(1) 56–68.
[55] Zanzoni, A., Soler-Lopez, M., and Aloy, P. (2009). A network medicine approach to human disease. *FEBS Lett* **583**(11) 1759–1765.
[56] Kitano, H. (2004). Biological robustness. *Nat Rev Genet* **5**(11) 826–837.
[57] Pache, R. A., Zanzoni, A., Naval, J., Mas, J. M., and Aloy, P. (2008). Towards a molecular characterisation of pathological pathways. *FEBS Lett* **582**(8) 1259–1265.
[58] Lalonde, R. L. and Honig, P. (2008). Clinical pharmacology in the era of biotherapeutics. *Clin Pharmacol Ther* **84**(5) 533–536.
[59] Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., et al. (2009). Predicting new molecular targets for known drugs. *Nature* **462**(7270) 175–181.
[60] Von Eichborn, J., Murgueitio, M. S., Dunkel, M., Koerner, S., Bourne, P. E., et al. (2011). Promiscuous: A database for network-based drug-repositioning. *Nucleic Acids Res* **39**(Database issue) D1060–1066.
[61] Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I., and Shoichet, B. K. (2008). Quantifying the relationships among drug classes. *J Chem Inf Model* **48**(4) 755–765.
[62] Oksenberg, J. R., Baranzini, S. E., Sawcer, S., and Hauser, S. L. (2008). The genetics of multiple sclerosis: Snps to pathways to pathogenesis. *Nat Rev Genet* **9**(7) 516–526.
[63] Nomura, D. K., Dix, M. M., and Cravatt, B. F. (2010). Activity-based protein profiling for biochemical pathway discovery in cancer. *Nat Rev Cancer* **10**(9) 630–638.
[64] Kool, J. and Berns, A. (2009). High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. *Nat Rev Cancer* **9**(6) 389–399.
[65] Zhang, W., Sun, F., and Jiang, R. (2011). Integrating multiple protein-protein interaction networks to prioritize disease genes: A bayesian regression approach. *BMC Bioinformatics* **12**(Suppl 1) S11.
[66] Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., et al. (2001). Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* **29**(14) 2994–3005.
[67] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Res* **39**(Database issue) D52–57.
[68] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., et al. (2000). Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat Genet* **25**(1) 25–29.
[69] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., et al. (2011). Ncbi geo: Archive for functional genomics data sets–10 years on. *Nucleic Acids Res* **39**(Database issue) D1005–1010.
[70] Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). Dbest–database for "expressed sequence tags". *Nat Genet* **4**(4) 332–333.
[71] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., et al. Interpro: The integrative protein signature database. *Nucleic Acids Res* **37**(Database issue) D211–215.
[72] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**(10) 2363–2371.
[73] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**(Database issue) D355–360.
[74] Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., et al. (2005). Toucan 2: The all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res* **33**(Web Server issue) W393–396.
[75] Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). Transfac: A database on transcription factors and their dna binding sites. *Nucleic Acids Res* **24**(1) 238–241.
[76] Guan, Y., Myers, C. L., Lu, R., Lemischka, I. R., Bult, C. J., et al. (2008). A genomewide functional network for the laboratory mouse. *PLoS Comput Biol* **4**(9) e1000165. MR2448486
[77] Ma, X., Lee, H., Wang, L. and Sun, F. (2007). Cgi: A new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* **23**(2) 215–221.
[78] Sun, J., Jia, P., Fanous, A. H., Webb, B. T., van den Oord, E. J., et al. (2009). A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. *Bioinformatics* **25**(19) 2595–6602.

Yong Chen
FIT 1-107, Tsinghua University
Beijing 100084
China
E-mail address: yongchen@tsinghua.edu.cn

Wangshu Zhang
FIT 1-107, Tsinghua University
Beijing 100084
China
E-mail address: zhangws08@mails.tsinghua.edu.cn

Mingxin Gan
School of Economics and Management
University of Science and Technology Beijing
Beijing 100083
China
E-mail address: ganmx@ustb.edu.cn

Rui Jiang
FIT 1-107, Tsinghua University
Beijing 100084
China
E-mail address: ruijiang@tsinghua.edu.cn
url: http://bioinfo.au.tsinghua.edu.cn/member/ruijiang