Front. Electr. Electron. Eng. China 2010, 5(2): 107–118
DOI 10.1007/s11460-010-0018-x

RESEARCH ARTICLE

Wangshu ZHANG, Yong CHEN, Rui JIANG

Comparative study of network-based prioritization of protein domains associated with human complex diseases

Abstract Domains are basic structural and functional unit of proteins, and, thus, exploring associations between protein domains and human inherited diseases will greatly improve our understanding of the pathogenesis of human complex diseases and further benefit the medical prevention, diagnosis and treatment of these diseases. Based on the assumption that deleterious nonsynonymous single nucleotide polymorphisms (nsSNPs) underlying human complex diseases may actually change structures of protein domains, affect functions of corresponding proteins, and finally result in these diseases, we compile a dataset that contains 1174 associations between 433 protein domains and 848 human disease phenotypes. With this dataset, we compare two approaches (guilt-by-association and correlation coefficient) that use a domain-domain interaction network and a phenotype similarity network to prioritize associations between candidate domains and human disease phenotypes. We implement these methods with three distance measures (direct neighbor, shortest path with Gaussian kernel, and diffusion kernel), demonstrate the effectiveness of these methods using three large-scale leave-one-out cross-validation experiments (random control, simulated linkage interval, and whole-genome scan), and evaluate the performance of these methods in terms of three criteria (mean rank ratio, precision, and AUC score). Results show that both methods can effectively prioritize domains that are associated with human diseases at the top of the candidate list, while the correlation coefficient approach can achieve slightly higher performance in most cases. Finally, taking the advantage that the correlation coefficient method does not require known disease-domain associations, we calculate a genome-wide landscape of associations between 4036 protein domains and 5080 human disease phenotypes using this method and offer a freely accessible web interface for this landscape.

Keywords protein domains, disease phenotypes, prioritization, guilt-by-association, correlation coefficient

Received January 27, 2010; accepted March 5, 2010

Wangshu ZHANG, Yong CHEN, Rui JIANG (✉)
MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China
E-mail: ruijiang@tsinghua.edu.cn

Yong CHEN
School of Sciences, University of Jinan, Jinan 250014, China

1 Introduction

The past decade has witnessed a golden era for the development of human genetic association studies that focus on the identification of genetic risk factors underlying human inherited diseases [1–3]. Of all researches in this area, the problem of identifying genes and their products that are responsible for specific diseases has long been concerned and extensively explored [4–14]. Since a protein is typically composed of several structural domains [15], it is reasonable to assume that harmful genetic variants such as deleterious nonsynonymous single nucleotide polymorphisms (nsSNPs) that are responsible for a specific disease may actually change structures of some protein domains, affect functions of corresponding proteins, and further result in the disease. Hence, revealing the relationships between protein domains that contain deleterious genetic variants and human diseases with which the variants are associated will greatly improve our understanding of genetic mechanisms of the diseases and further benefit medical prevention, diagnosis, and treatment of these diseases.

Analogous to the prioritization of disease genes [16–21], the problem of discovering protein domains that are associated with a specific disease phenotype can be formulated as a prioritization problem, in which one first scores the possibility of association between a candidate domain and a specific disease phenotype of interest and then ranks a list of candidates according to their scores. This general one-class novelty learning formulation, together with the explosive expansion of domain-domain

interaction databases [22–25] and pair-wise similarity measurements of human disease phenotypes [26], suggests the following methods for prioritizing protein domains that are associated with human complex diseases. First, according to the "guilt-by-association" principle [27] that has been successfully used in the identification of disease genes [28,29], one assumes that a candidate domain is likely to be associated with a disease if the candidate has similar properties as domains that are known to be associated with the disease. According to this assumption, a set of domains that are associated with a specific disease are defined as seed domains, and the possible association of a candidate domain to the disease is measured by the total distance from the candidate to all seed domains in some domain-domain interaction network. Second, according to the "correlation coefficient" approach [4], one assumes that phenotypically similar diseases are often caused by functionally related domains, and, thus, a positive correlation should exist between a domain-domain relatedness profile and a phenotype-phenotype similarity profile. As a result, there is no need of defining seed domains, and candidate domains are prioritized by the correlation coefficient of corresponding relationships in the domain-domain interaction network and those in the phenotype-phenotype similarity network.

In this paper, we implement a proof-of-concept analysis and comparison on the performance of the above two prioritization approaches, say, the guilt-by-association approach and the correlation coefficient approach. For both approaches, a set of known disease-domain associations are obtained from the UniProt database [30] and the Pfam database [31]. In addition, a domain-domain interaction network is compiled from the DOMINE database [22]. Furthermore, for the correlation coefficient approach, an additional phenotype-phenotype similarity network is obtained from Van Driel et al.'s result [26]. For both methods, we use three distance measures: direct neighbors (DN), shortest path with Gaussian kernel (SG), and diffusion kernel (DK), to calculate the distance between two domains in the given domain-domain interaction network.

We apply three leave-one-out cross-validation experiments (random control, simulated linkage interval, and whole-genome scan) to validate the above approaches, and we evaluate the performance of each approach in terms of three widely used criteria (mean rank ratio, precision, and the area under the receiver operating characteristic (ROC) curve (AUC) score). Results show that both approaches can successfully recover the associations between protein domains and human disease phenotypes, though the correlation coefficient approach performs better in the random control and the simulated linkage interval experiments, and the guilt-by-association approach achieves higher performance in the whole-genome scan. We further verify that the performance of our approaches does rely on the information contained in the domain-domain interaction network by shuffling the original network. Finally, taking the advantage that the correlation coefficient method does not require known disease-domain associations, we calculate a genome-wide landscape of the associations between 4036 protein domains and 5080 human disease phenotypes and offer a freely accessible web interface for this landscape.

2 Materials and methods

2.1 Data sources

2.1.1 Domain-domain interaction network

We ground our inference of disease-domain associations on a domain-domain interaction network that is extracted from the DOMINE database [22]. DOMINE is one of the most widely-used databases of known and predicted domain-domain interactions. As for February 2008, this database contains a total of 20513 domain-domain interactions, out of which 4349 (gold-standard positives) are inferred from Protein Data Bank (PDB) entries (the union of the sets of interactions from iPfam [25] and 3did [32,33]), and 17781 are predicted by at least one computational approach of eight different approaches using Pfam domain annotations. Of the 17781 predictions, 3143 interactions are high-confidence predictions (predicted by ME [34] or at least two different approaches), 729 interactions are medium-confidence predictions (hetero-domain interactions in which both domains belong to the same biological process in the gene ontology), and the remaining 13909 are low-confidence predictions [22]. Among all the interactions, the PDB part possesses the highest reliability with the lowest false positive rate, followed by the high-confidence interactions, and then the medium and low-confidence interactions. On the other hand, the coverage of network only composing PDB entries is the lowest (2285 no-self-loop interactions out of altogether 4349 interactions among 1971 non-isolated domains out of altogether 2948 domains, with averagely 1.1593 interactions per domain), while that of the entire DOMINE network is the highest (18239 no-self-loop interactions out of altogether 20513 interactions among 3499 non-isolated domains out of altogether 4036 domains, with averagely 5.2126 interactions per domain). For this dilemma, we have to seek for a compromise between the reliability and the coverage. Consequently, we use the PDB + high-confidence part of the DOMINE database, which has the medium rate of false interactions and also the medium coverage (3960 no-self-loop interactions out of altogether 6163 interactions among 2282 non-isolated domains out of altogether 3055 domains, with averagely 1.7353 interactions per domain). With this definition, all its 2282 domains with a total of 3960 no-self-loop interactions are utilized as the domain-domain interaction network.

2.1.2 Phenotype similarity network

The phenotype similarity network is obtained from an earlier work of Van Driel et al. [26], in which the pair-wise relationships between 5080 human genetic disease phenotypes from the Online Mendelian Inheritance in Man (OMIM) database are mapped. The anatomy (A) and the disease (C) sections of the medical subject headings (MeSH) vocabulary are used to extract terms from OMIM, providing a standard way to represent the OMIM records as corresponding phenotype feature vectors. Each phenotype is 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. For each pair of phenotype, their similarity score is calculated by the cosine of their feature vector angle [35]. The reliability of the phenotype similarity score has been tested [26], showing that these phenotype similarities are positively correlated with a number of measures of gene functions. The final phenotype network contains pair-wise similarity scores for 5080 OMIM phenotypes, covering the majority of recorded human phenotypes.

2.1.3 Domain-disease associations

We define that a domain is associated with a disease phenotype of interest if the domain contains at least one deleterious nsSNP that is associated with the phenotype; therefore, the associations between protein domains and disease phenotypes are obtained by bridging the associations between phenotypes and deleterious nsSNPs as well as relationships between protein and domains.

Associations between phenotypes and nsSNPs are obtained from the UniProt database [30], where nsSNPs are classified into three categories: disease, polymorphism, and unclassified. In version 57.12 (released on December 15th, 2009) of this database, 22992 nsSNPs belong to the disease category; 36179 belong to the polymorphism category; and the rest 1993 nsSNPs are unclassified. For an nsSNP belonging to the disease category, the entry identification (ID) of the specific disease in the OMIM database is also offered. We further calculate the percentage of disease-related SNPs located in the domain region in the whole protein region, and of all the 22992 disease-related nsSNPs, 16366 are located in the domain region, with the percentage 71.18%. In our study, associations between phenotypes and nsSNPs are constructed using nsSNPs that belong to the disease category.

Relations between human proteins and domains are obtained from the Pfam database [31]. This database is a large collection of protein domain families containing two levels of quality: a manually curated, high-quality collection of domain families (Pfam-A), and an un-annotated, low-quality collection of domain families (Pfam-B). In version 24.0 of the Pfam-A collection (released in October 2009), 11912 domain families that cover more than 75.15% of known proteins are collected.

With the above two data sources, we collect from the UniProt and the Pfam databases 1174 associations between 433 domains and 848 diseases, with 625 diseases being associated with one domain and 223 diseases being associated with two or more domains.

2.2 Guilt-by-association approach

The "guilt-by-association" approach [27–29] is constructed based on the assumption that domains that are close in a domain-domain interaction network are likely to be associated with the same disease. On the basis of this assumption, a set of domains that have associations with a certain disease phenotype are defined as seed domains, and the total distance of a candidate domain to the set of seed domains is utilized as the score for prioritization. The scheme of the guilt-by-association approach is shown in Fig. 1(a).

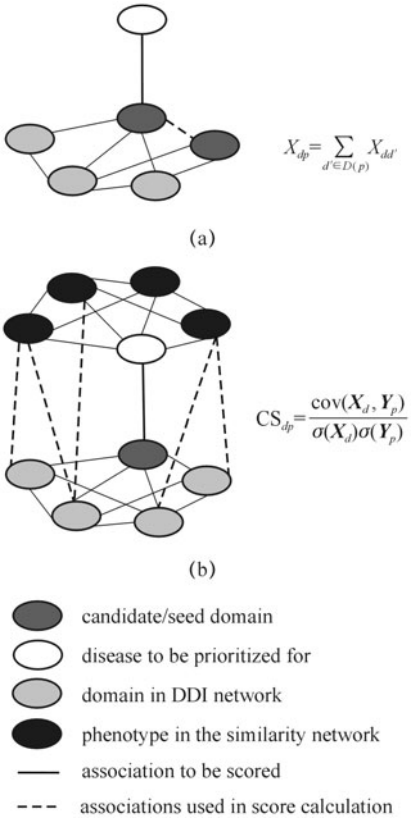

Fig. 1 Schemes of two approaches. (a) Guilt-by-association approach; (b) correlation coefficient approach

We employ three distance measures with this approach: 1) DN, 2) SG, and 3) DK. The direct neighbor distance DN(u,v) between two domains u and v is defined 1 if the two domains are adjacent in the domain-domain interaction network and $+\infty$ if the two domains are not

adjacent. The shortest path distance SP(u,v) between two domains u and v is defined as the length of the shortest path between the two domains. With the use of Gaussian kernel, the distance measure SG(u,v) is obtained as

$$\mathrm{SG}(u,v) = \mathrm{e}^{-\beta(\mathrm{SP}(u,v))^2},$$

where β is a free parameter, and we set it to 1 in our studies. The diffusion kernel $K = \{k_{uv}\}_{n\times n}$ of a domain-domain interaction network with n domains is defined as $K = \mathrm{e}^{\gamma L}$, where $0 < \gamma < 1$ is a free parameter that controls the magnitude of diffusion, and we set it to 0.05 in our analysis. The matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ is the Laplacian of the network, where $\boldsymbol{D}$ is a diagonal matrix containing node degrees, and $\boldsymbol{A}$ is the adjacency matrix of the domain-domain interaction network. With the diffusion kernel being calculated, we define the diffusion kernel distance between two domains u and v as

$$\mathrm{DK}(u,v) = k_{uv}.$$

2.3 Correlation coefficient approach

A limitation of the guilt-by-association approach is that it can only prioritize candidate domains for a disease phenotype that has at least one seed domain known to be associated with the phenotype. Therefore, this approach cannot be applied to disease phenotypes that are lacking of any associated domains. To overcome this limitation, we follow the CIPHER method proposed by Ref. [4] to obtain concordance scores for candidate domains by correlating human phenotype similarity profile and domain proximity profile and further prioritize the candidates according to their score. CIPHER is a tool for prioritizing disease genes. This method utilizes a simple linear regression model that integrates human protein-protein interactions, disease phenotype similarities, and known gene-phenotype associations to capture the relationships between phenotypes and genotypes [4]. Following this notion, we assume a linear relationship between the disease phenotype similarity profile of the disease of interest and the domain proximity profile of a candidate domain, and we define the correlation coefficient between the domain proximity profile and the phenotype similarity profile as a score to measure the possibility of association between the disease phenotype and the candidate domain.

Mathematically, let $Y_{pp'}$ denote the similarity score between a query disease phenotype p and another disease phenotype p'. We define the disease phenotype similarity profile for the disease phenotype p as the vector

$$\boldsymbol{Y}_p = (Y_{pp1}, Y_{pp2}, ..., Y_{ppm}),$$

that is, the similarities between the disease phenotype p and all other disease phenotypes $p_1, p_2, \ldots, p_m$, where m is the total number of disease phenotypes in the disease phenotype similarity database. Let $X_{dd'}$ denote the proximity (distance) of two domains d and d' in the domain-domain interaction network, calculated from one of the three methods (i.e., direct neighbor, shortest path, and diffusion kernel) described above. Let $D(p)$ denote the set of domains known to be associated with a disease phenotype p. We define X_{dp}, the proximity of a candidate domain d to a disease phenotype p, as the total distance from the domain d to all domains known to be associated with the disease phenotype p, that is,

$$X_{dp} = \sum_{d' \in D(p)} X_{dd'}.$$

Furthermore, we define $\boldsymbol{X}_d$, the domain proximity profile for a candidate domain d, as the vector composed of the proximities of the candidate domain d to all disease phenotypes, that is,

$$\boldsymbol{X}_d = (X_{dp1}, X_{dp2}, ..., X_{dpm}).$$

With these definitions, we define the concordance score of a domain d and a disease phenotype p as the Pearson's correlation coefficient of the domain proximity profile $\boldsymbol{X}_d$ and the disease phenotype similarity profile $\boldsymbol{Y}_p$, that is,

$$\mathrm{CS}_{dp} = \frac{\mathrm{cov}(\boldsymbol{X}_d, \boldsymbol{Y}_p)}{\sigma(\boldsymbol{X}_d)\sigma(\boldsymbol{Y}_p)}.$$

Obviously, this correlation coefficient between the domain proximity profile $\boldsymbol{X}_d$ and the disease phenotype similarity profile $\boldsymbol{Y}_p$ should be high if the similarities between the disease phenotype p and all other disease phenotypes can be explained by the proximities of the candidate domain to domains that are known to be associated with the other disease phenotypes. Consequently, we can use this correlation coefficient to prioritize a list of candidate domains. Following Ref. [4], we name this correlation coefficient the concordance score. The scheme of the correlation coefficient approach is shown in Fig. 1(b).

2.4 Validation methods and evaluation criteria

On the basis of the domain-domain interaction network and the known associations between protein domains and disease phenotypes, we would like to validate how well the above two approaches perform in recovering these known associations. We adopt three large-scale leave-one-out cross-validation experiments for this purpose.

First, in the validation of random controls, we prioritize domains that are known to be associated with disease phenotypes (i.e., disease domains) against randomly selected control domains. Specifically, in each run of the validation, we select an association between a domain and a disease phenotype, assume that the association is unknown, and prioritize the domain against a set of 99 randomly selected control domains. Note that for the guilt-by-association approach, seed domains are selected as all

domains that are associated with the disease phenotype, except for the domain under investigation. For this reason, we only focus on disease phenotypes that are associated with at least two domains in the validation procedure for the guilt-by-association approach, since we need to use at least one domain that is associated with a phenotype as the seed to calculate the closeness score. For the correlation coefficient approach, however, there is no such limitation, and all the phenotype-associated domains can be used in the validation procedure.

Second, in the validation of simulated linkage intervals, we prioritize domains that are known to be associated with disease phenotypes (i.e., disease domains) against domains that are located around the disease domains. Specifically, in each run of the validation, we select an association between a domain and a disease phenotype, assume that the association is unknown, and prioritize the domain against a set of control domains that are located in 10 Mbp upstream and downstream around this domain [36]. Our analysis shows that there are, on average, 53 candidate domains in the linkage intervals simulated in this way, where the minimum number of candidate domains in the linkage interval is four, and the maximum number is 395. As analyzed before, for the guilt-by-association approach, we only focus on disease phenotypes that are associated with at least two domains and select seed domains as all domains that are associated with the disease phenotype except for the domain under investigation, while for the correlation coefficient approach, we could further validate disease phenotypes that are associated with only one domain.

Third, in the validation of whole-genome scan, we prioritize domains that are known to be associated with disease phenotypes (i.e., disease domains) against all known domains. Specifically, in each run of the validation, we select an association between a domain and a disease phenotype, assume that the association is unknown, and prioritize the domain against a set of control domains that are collected in the Pfam database and included in the domain-domain interaction network. Again, for the guilt-by-association approach, we only focus on disease phenotypes that are associated with at least two domains, while for the correlation coefficient approach, we further validate disease phenotypes that are associated with only one domain.

In either of the above leave-one-out cross-validation experiments, we repeat the validation run for every known association between a disease phenotype and a domain, and we are able to obtain a number of ranking lists. We further normalize the ranks by dividing them with the total number of candidate domains in the rankling list to obtain rank ratios and derive three criteria to measure the performance of a prioritization method.

We first propose a criterion called the mean rank ratio, which is simply the average of rank ratios for all known disease domains in a cross-validation experiment. This criterion provides a summary of the ranks of all domains that are known to be associated with disease phenotypes, and the smaller the mean rank ratio, the better a method. The second criterion we propose to use is called precision. We consider a prediction as successful if the known disease domain is ranked at the top (with rank 1). Then, the proportion of successful predictions among all predictions is defined as the precision. Obviously, a high precision means a method with high prediction power. The third criterion we propose to use is AUC, which is the area under the ROC curve. Given a list of rank ratios and a predefined threshold, we define the sensitivity as the percentage of disease domains that are ranked above the threshold and the specificity as the percentage of control domains that are ranked below the threshold. Varying the threshold values, we are able to plot a receiver operating characteristic curve, which shows the relation between sensitivity and 1 – specificity. Calculating the AUC, we are able to obtain the AUC score, which provides an overall measure for the performance of the prioritization method.

3 Results

3.1 Comparison of proposed approaches

We first compare the performance of the guilt-by-association approach and the correlation coefficient approach through the three large-scale leave-one-out cross-validation experiments using the data compiled from the DOMINE [22], Pfam [31], and UniProt [30] databases. Briefly, we extract the PDB + high-confidence part of the DOMINE database and obtain a domain-domain interaction network, which has the medium rate of false interactions and also the medium coverage (3960 no-self-loop interactions out of a total of 6163 interactions among 2282 non-isolated domains out of a total of 3055 domains, with on average 1.7353 interactions per domain). We analyze the connectivity of this network and find 2282 non-isolated domains with 3960 no-self-looped interactions. Further analysis shows that the giant component (the largest connected component) of this network contains 3563 interactions among 1721 domains, while the second largest connected component contains only ten interactions among ten domains. There are further one component with seven domains, three components with six domains, three components with five domains, ten components with four domains, 51 components with three domains, and 159 components with two domains. Considering that the interactions in this network are of relatively high quality, we decide to use all the 2282 non-isolated domains with 3960 interactions as the domain-domain interaction network in our cross-validation experiments. We then collect from the UniProt [30] and the Pfam databases [31] 1174 associations between 433 domains and 848 phenotypes, in which 625 (74%) phenotypes are associated with

one domain and 223 (26%) phenotypes are associated with two or more domains.

For a fair comparison, we apply both the guilt-by-association approach and the correlation coefficient method to the 223 disease phenotypes that are associated with two or more domains and calculate the mean rank ratios, precisions, and AUC scores of both methods, as shown in Table 1 and upper panels of Figs. 2–4. In Table 1, the domain-domain interaction network includes the PDB + high-confidence part of the DOMINE database. GBA represents guilt-by-association approach, CC represents correlation coefficient approach. The results for random controls are mean (standard derivation) of ten runs.

First, we can see from these results that both methods can successfully recover the associations between protein domains and human disease phenotypes. For example, in the cross-validation for random controls, the mean rank ratios are less than 10%; the precisions are greater than 50%; and the AUC scores are greater than 90%. In the cross-validation for linkage intervals, the mean rank ratios are less than 20%; the precisions are greater than 14%; and the AUC scores are greater than 80%. In the cross-validation for whole-genome scan, the mean rank ratios are less than 15%; the precisions are greater than 55%; and the AUC scores are greater than 85%. We, therefore, conclude that both the guilt-by-association and correlation coefficient approaches are effective in the identification of protein domains that are associated with human disease phenotypes.

Second, we conjecture from these results that the correlation coefficient approach can achieve higher performance than the guilt-by-association method in most cases. For example, in the cross-validations of both random controls and simulated linkage intervals, the correlation coefficient approach can achieve smaller mean rank ratios, higher precisions, and larger AUC scores. When looking at the ROC curve (Figs. 2–4), we can see clearly that the curve of the correlation coefficient approach climbs much faster towards the upper left corner (0, 1) of the plot than does that of the guilt-by-association method.

Third, we conclude from these results that both the shortest path measure with Gaussian kernel and the diffusion kernel measure are significantly better than the direct neighbor measure, and the difference between these two measures is not obvious, though the diffusion kernel method seems slightly better. One possible explanation of this phenomenon lies in that the direct neighbor measure is a local measure which only uses interaction information of domains that are directly connected, while the shortest path and the diffusion kernel measures can use domain-domain interaction information of all connected parts of the whole network. Another reason may be that the direct neighbor measure produces much more identical ranks than that of the shortest path measure which produces only a few identical ranks, while the diffusion kernel measure almost never produces identical ranks. Since in our problem, the final rank list of candidate domains is calculated by the weighted mean of their original ranks, if too many identical ranks exist, then the rank of the true candidate domain we intend to obtain may be "pulled" down. For example, the disease Retinitis Pigmentosa is associated with eight domains: 7tm_1, Pkinase_Tyr, PDEase_I, fn3, Ion_trans, Homeobox, Laminin_EGF, and GAF. When we use 7tm_1 as the candidate domain, the direct neighbor measure prioritizes all the eight domains at rank 1 and thus leads to a final rank of 4.5. On the other hand, the shortest path measure prioritizes both of 7tm_1 and Pkinase_Tyr at rank 1 and leads to a final rank of 1.5, while the diffusion kernel measure only prioritizes 7tm_1 at rank 1 and leads to a final rank of 1. This scenario is quite common when one prioritizes candidate domains for phenotypes that have two or more seed domains, and it is possible that this characteristic of the diffusion kernel measure contributes to its higher performance.

Finally, the guilt-by-association method requires a set of seed domains and thus can only be applied to prioritize candidate domains for disease phenotypes with which

Table 1 Leave-one-out cross-validation results for 223 phenotypes associated with two or more domains, based on PDB + high-confidence part of domain-domain interactions in DOMINE

criterion	distance measure	random control		linkage interval		whole-genome scan	
		GBA	CC	GBA	CC	GBA	CC
mean rank ratio	DN	0.2951 (0.0002)	0.1979 (0.0011)	0.2853	0.1965	0.3187	0.1819
	SG	0.1180 (0.0008)	0.1065 (0.0009)	0.2246	0.1893	0.1102	0.1324
	DK	**0.0944** (0.0008)	**0.0793** (0.0011)	**0.1843**	**0.1581**	**0.0858**	**0.1279**
precision	DN	0.4523 (0.0118)	0.5597 (0.0070)	0.1300	0.2333	0.5592	**0.5883**
	SG	0.4576 (0.0178)	0.4475 (0.0120)	0.1367	0.2567	0.5628	0.4590
	DK	**0.5100** (0.0157)	**0.6189** (0.0069)	**0.1467**	**0.2667**	**0.6193**	0.4772
AUC	DN	0.7142 (0.0001)	0.8052 (0.0007)	0.7440	0.8131	0.6827	0.8135
	SG	0.8931 (0.0009)	0.9061 (0.0010)	0.7705	0.8206	0.8900	0.8644
	DK	**0.9150** (0.0008)	**0.9311** (0.0012)	**0.8250**	**0.8802**	**0.9119**	**0.8690**

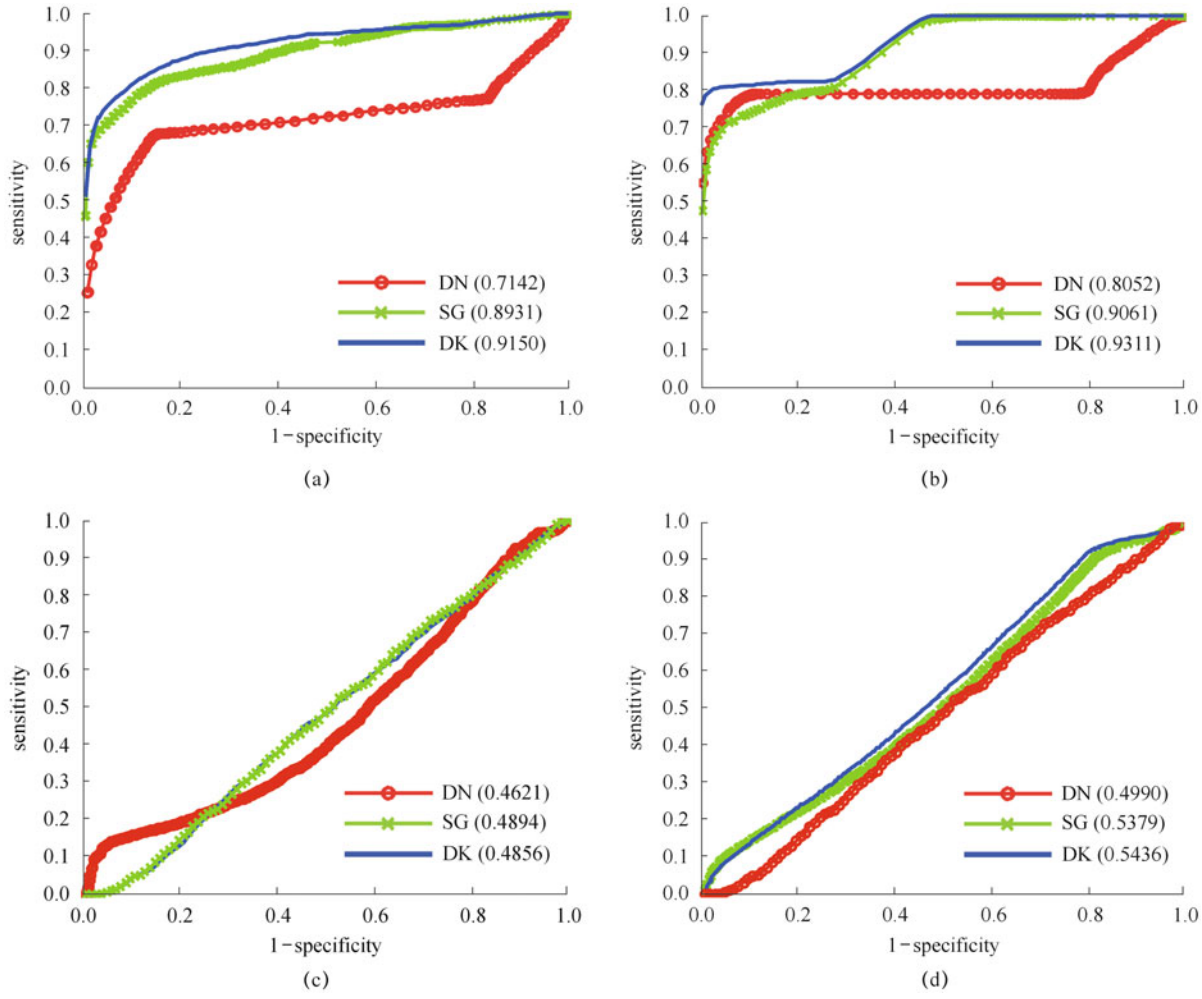

Fig. 2 ROC curves and AUC scores for random controls. (a) Results of guilt-by-association approach on PDB + high-confidence interactions in DOMINE database; (b) results of correlation coefficient approach on PDB + high-confidence interactions in DOMINE database; (c) results of guilt-by-association approach on shuffled PDB + high-confidence interactions in DOMINE database; (d) results of correlation coefficient approach on shuffled PDB + high-confidence interactions in DOMINE database

some domains are known to be associated. The correlation coefficient approach, however, does not have such restriction and can be applied to prioritize candidate domains for disease phenotypes with which no domains are known to be associated. With this understanding, we further perform the three leave-one-out cross-validation experiments with the use of the 625 disease phenotypes that are associated with only one domain and present the results in Table 2. In Table 2, the domain-domain interaction network includes the PDB + high-confidence part of the DOMINE database. The results for random controls are mean (standard derivation) of ten runs.

In general, these results are consistent with our previous analysis. For example, we can see from the table that both the shortest path measure and the diffusion kernel measure can achieve better performance than the direct neighbor measure, while the difference between these two measures is not obvious.

3.2 Robustness of proposed approaches

Although the above validation results suggest that the proposed approaches can successfully prioritize candidate domains and put the domain that is truly associated with the query disease phenotype at the top of the ranking list, it is necessary to validate whether the correct prioritization of disease domains is due to the connectivity information that is included in the domain-domain interaction network. For this purpose, we can artificially destroy informative interactions in the network and see what results can be obtained. It should be expected that both the mean rank ratios and the AUC scores should be around 50%, together with very low precisions. With this understanding, we shuffle interactions among domains while fixing the degree (number of direct neighbors) of every domain. Then we repeat each of the leave-one-out cross-validation experiments using the shuffled network, which contains no

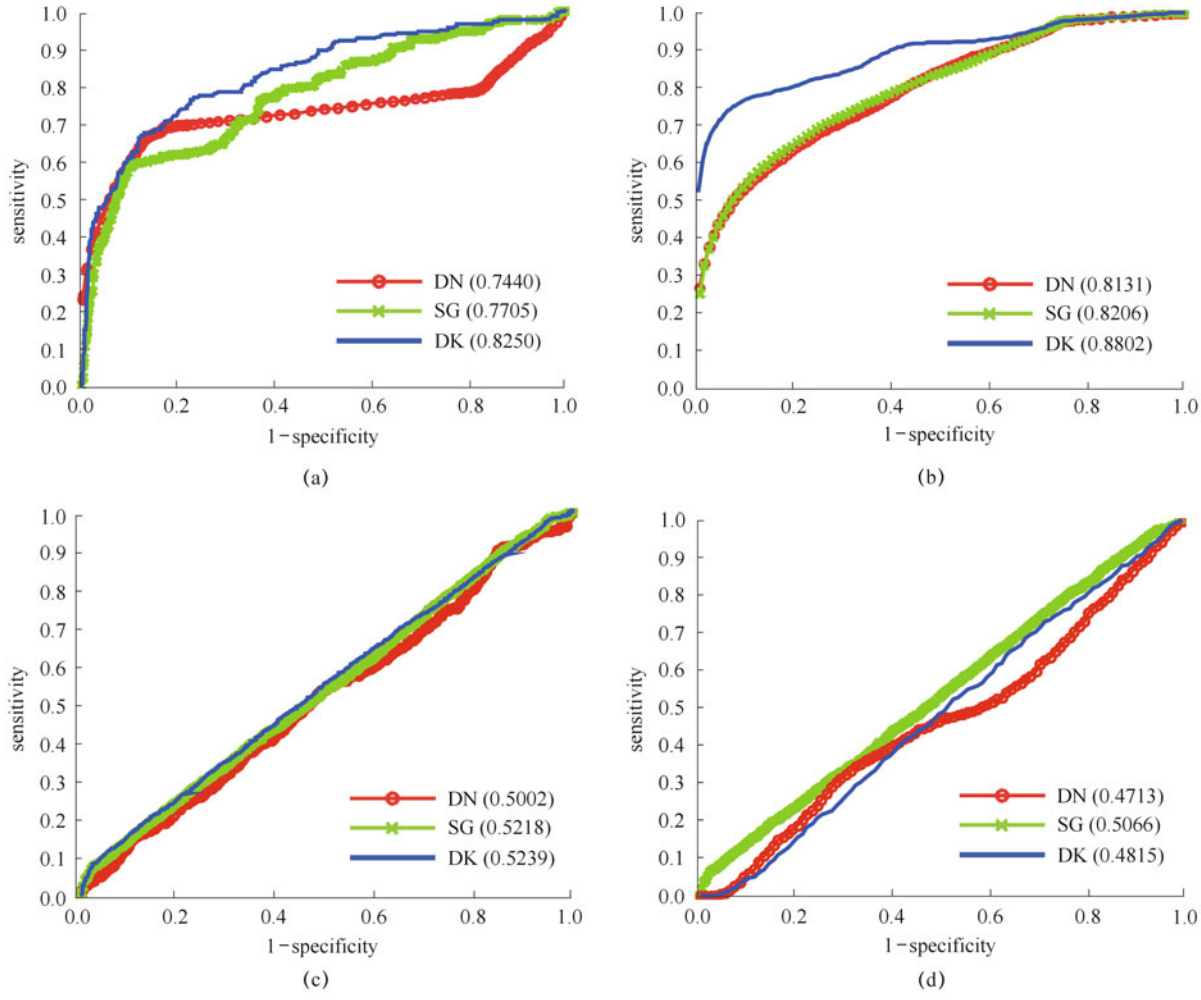

Fig. 3 ROC curves and AUC scores for linkage intervals. (a) Results of guilt-by-association approach on PDB + high-confidence interactions in DOMINE database; (b) results of correlation coefficient approach on PDB + high-confidence interactions in DOMINE database; (c) results of guilt-by-association approach on shuffled PDB + high-confidence interactions in DOMINE database; (d) results of correlation coefficient approach on shuffled PDB + high-confidence interactions in DOMINE database

informative interactions among domains. As we expected, the mean rank ratios are, in general, about 50% (data not shown); the precisions are, in general, no more than 0.07 (data not shown); the AUC scores are, in general, about 50% (lower panels of Figs. 2–4). From these results, we conclude that the successful prioritization of candidate domains is indeed due to the informative interactions among domains that are included in the domain-domain interaction network.

We notice that the domain-domain interaction network composed of both high-confidence interactions and the predicted predictions is quite different from the network composed of only the high-confidence interactions (PDB and interactions predicted by at least two computational methods). For example, the average degree of the high-confidence network is only 1.7353, while that of the entire network is 5.2126. Obviously, many predicted interactions may be actually noise and thus badly affect the prioritization of disease domains. It is, therefore, necessary to validate the robustness of the proposed approaches. For this purpose, we implement the same validation process based on the domain-domain interaction network composed of all interactions in the DOMINE database and present the results in Tables 3 and 4. In Tables 3 and 4, the domain-domain interaction network includes all interactions in the DOMINE database. The results for random controls are mean (standard derivation) of ten runs. From these tables, we can see that the performances of both methods using the entire DOMINE network are generally a little inferior to those using the PDB + high-confidence part of DOMINE network. We then conjecture from these results that both proposed methods are robust to the possible noise in the domain-domain interaction network.

Further, we notice that the parameter β in the shortest

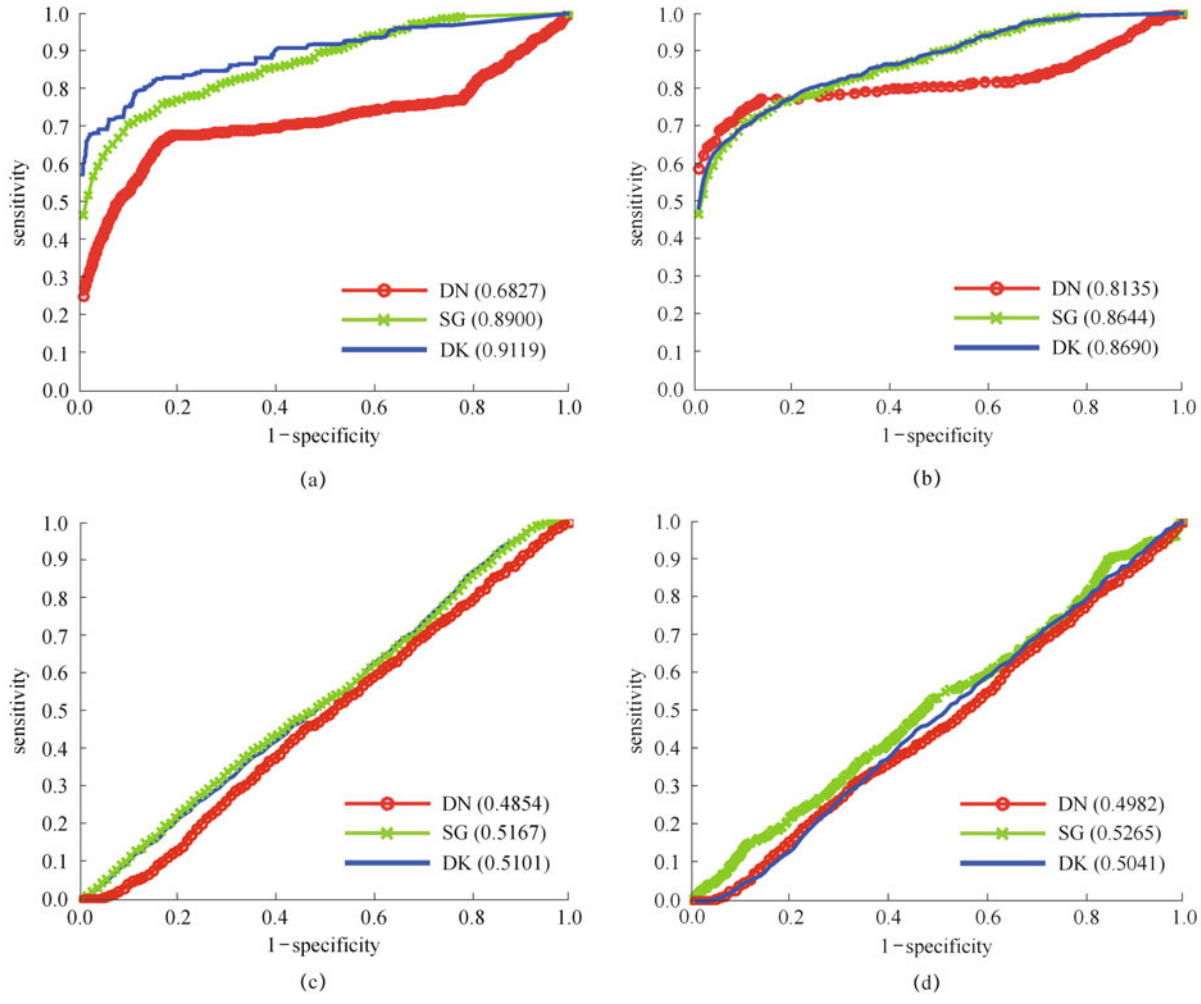

Fig. 4 ROC curves and AUC scores for whole-genome scan. (a) Results of guilt-by-association approach on PDB + high-confidence interactions in DOMINE database; (b) results of correlation coefficient approach on PDB + high-confidence interactions in DOMINE database; (c) results of guilt-by-association approach on shuffled PDB + high-confidence interactions in DOMINE database; (d) results of correlation coefficient approach on shuffled PDB + high-confidence interactions in DOMINE database

Table 2 Leave-one-out cross-validation results of correlation coefficient approach for 625 phenotypes associated with only one domain, based on PDB + high-confidence part of domain-domain interactions in DOMINE

criterion	distance measure	random control	linkage interval	whole-genome scan
mean rank ratio	DN	0.2894 (0.0008)	0.2087	0.2825
	SG	0.2103 (0.0011)	0.1974	0.2366
	DK	**0.2072** (0.0010)	**0.1973**	**0.2320**
precision	DN	0.2534 (0.0103)	0.1524	0.2576
	SG	**0.2688** (0.0088)	0.1712	**0.2656**
	DK	0.2539 (0.0067)	**0.2095**	0.2544
AUC	DN	0.7156 (0.0016)	0.7959	0.7136
	SG	0.7973 (0.0012)	0.8068	0.7630
	DK	**0.8005** (0.0010)	**0.8074**	**0.7676**

Table 3 Leave-one-out cross-validation results for 223 phenotypes associated with two or more domains, based on all domain-domain interactions in DOMINE

criterion	distance measure	random control		linkage interval		whole-genome scan	
		GBA	CC	GBA	CC	GBA	CC
mean rank ratio	DN	0.2923 (0.0004)	0.2169 (0.0009)	0.2829	0.1775	0.2853	0.1829
	SG	0.1075 (0.0010)	0.1236 (0.0015)	0.2321	0.1790	0.0974	0.1251
	DK	**0.0943** (0.0007)	**0.1228** (0.0010)	**0.1959**	**0.1726**	**0.0853**	**0.1179**
precision	DN	0.4062 (0.0101)	0.4715 (0.0081)	0.1151	0.2619	0.4921	**0.5194**
	SG	0.4013 (0.0104)	**0.3967** (0.0106)	0.1456	0.2500	0.4806	0.4532
	DK	**0.5165** (0.0119)	0.3963 (0.0082)	**0.1959**	**0.2778**	**0.6173**	0.4331
AUC	DN	0.7223 (0.0012)	0.7745 (0.0006)	0.6833	0.8292	0.7172	0.8125
	SG	0.9273 (0.0009)	0.9219 (0.0007)	0.7791	0.8273	0.9032	0.8716
	DK	**0.9483** (0.0016)	**0.9234** (0.0005)	**0.8103**	**0.8338**	**0.9108**	**0.8793**

Table 4 Leave-one-out cross-validation results of correlation coefficient approach for 625 phenotypes associated with only one domain, based on all domain-domain interactions in DOMINE

criterion	distance measure	random control	linkage interval	whole-genome scan
mean rank ratio	DN	0.2917 (0.0008)	0.2357	0.2850
	SG	0.2334 (0.0011)	**0.2188**	0.2345
	DK	**0.2283** (0.0017)	0.2270	**0.2289**
precision	DN	**0.2528** (0.0079)	0.0865	0.2604
	SG	0.2430 (0.0065)	0.1315	0.2474
	DK	0.2474 (0.0048)	**0.1972**	**0.2617**
AUC	DN	0.7277 (0.0043)	0.7703	0.7112
	SG	0.8085 (0.0039)	0.7872	0.7638
	DK	**0.8148** (0.0024)	**0.7790**	**0.7693**

path measure with Gaussian kernel and the parameter γ in the diffusion kernel are free parameters that need to be pre-determined. However, in the above cross-validation experiments, we set them as 1 and 0.05, respectively, to seek for the simplicity. It is necessary to show whether the prioritization methods are sensitive to these parameters. For this purpose, we select several values across the range of these parameters, perform the cross-validation experiments, and see how the criteria change accordingly. We take the prioritization results with the best performance in the cross-validation experiments (correlation coefficient method and random control validation in Table 1) as an example to illustrate the influence of β. Since this parameter ranges from 0 to $+\infty$, we perform a grid search of this parameter by changing it from 0.01 to 100 with step 0.01 and see how the criteria change. We find that the peak performance is obtained at $\beta = 0.8$ (mean rank ratio = 0.0996, precision = 0.4614, and AUC score = 0.9085), and the results are very similar to those at the default value $\beta =$ 1.0 (mean rank ratio = 0.1065, precision = 0.4475, and AUC score = 0.9061). We then conjecture that the prioritization methods are not sensitive to this free parameter. Similarly, we find that the prioritization methods are not sensitive to the free parameter γ (data not shown).

Finally, in order to test the influence of the size of seed associations on the prioritization results, we use 100%, 90%, 80%, 70%, 60%, and 50% of the original seed associations, respectively, and we repeat the leave-one-out validation processes. We only calculate the performance using the correlation coefficient approach based on diffusion kernel measure and choose the PDB + high-confidence part of the DOMINE database as the domain-domain interaction network. Results show that with the percentage of used seed associations decreasing from 100% to 50%, performance in terms of mean rank ratio, precision and AUC score also decrease slightly, despite a few exceptions. For example, in the cross-validation for random controls, the decreases of mean rank ratios are no more than 8.28%; the decreases of precisions are no more than 10.04%; and the decreases of AUC scores are no more than 8.33%. In the cross-validation for linkage intervals, the decreases of mean rank ratios are no more than 7.69%; the decrease of precisions are no more than 9.02%; and the decreases of AUC scores are no more than 8.11%. In the cross-validation for whole-genome scan, the decreases of

mean rank ratios are no more than 6.83%; the decreases of precisions are no more than 10.05%; and the decreases of AUC scores are no more than 9.33%. From these results, we conclude that the prioritization methods are not sensitive to the size of seed associations.

3.3 Landscape of domain-phenotype associations

With the above validation results demonstrating the possibility of recovering the associations between protein domains and disease phenotypes, we further apply the correlation coefficient method to all available protein domains and human disease phenotypes and predicted a genome-wide landscape of the associations between protein domains and human disease phenotypes. There are a total of 5080 phenotypes in the phenotype similarity network and 4036 protein domains in the domain-domain interaction network (the entire DOMINE network). For each phenotype, we perform a prioritization of all domains with the use of the correlation coefficient approach (with the diffusion kernel measure). The prioritization results, together with a freely accessible web interface, are provided at http://bioinfo.au.tsinghua.edu.cn/member/wzhang/domain. All domains on the webpage are linked to the DOMINE database, from which further information can be obtained.

4 Discussion

In this paper, we study the problem of identifying associations between protein domains and human disease phenotypes from the viewpoint of one class novelty learning. We propose a correlation coefficient approach to prioritize a list of candidate domains and compare the performance of this approach with a method designed according to the guilt-by-association principle. We validate the performance of the proposed approach through extensive large-scale cross-validation experiments and demonstrate the superior performance of this approach.

The success of the correlation coefficient approach is mainly due to the use of the similarity profiles between disease phenotypes. The principle behind this model is that the similarity between two disease phenotypes can be explained by the proximities of domains that are associated with the phenotypes via a linear functional relationship. Therefore, the qualities of both the phenotype similarity profiles and the proximities among domains are of great importance to this approach. As for the former, there are already several methods for calculating similarities between disease phenotypes [37–39], and how to integrate the results of these methods remains a challenge. As for the latter, how to develop effective computational methods to predict domain-domain interactions is still an open question. Besides the qualities of the data, a more comprehensive model that is capable of capturing the complicated functional relationship between the similarity profile and the proximity profile is desired. One possible approach is to explore this regression problem from the viewpoint of Gaussian process [40].

Acknowledgements This work was partly supported by the National Natural Science Foundation of China (Grant Nos. 60805010, 60928007, 60934004, and 10926027), Tsinghua National Laboratory for Information Science and Technology (TNLIST) Cross-discipline Foundation, the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 200800031009), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, China Postdoctoral Science Foundation (No. 20090450396), the Scientist Research Fund of Shandong Province (No. BS2009SW044), and the Doctor Research Fund from University of Jinan (No. XBS0914).

References

1. Glazier A M, Nadeau J H, Aitman T J. Finding genes that underlie complex traits. Science, 2002, 298(5602): 2345–2349
2. Bird T D. Genetic factors in Alzheimer's disease. The New England Journal of Medicine, 2005, 352(9): 862–864
3. Lander E S, Schork N J. Genetic dissection of complex traits. Science, 1994, 265(5181): 2037–2048
4. Wu X, Jiang R, Zhang M Q, Li S. Network-based global inference of human disease genes. Molecular Systems Biology, 2008, 4: 189
5. Goh K, Cusick M E, Valle D, Childs B, Vidal M, Barabási A L. The human disease network. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(21): 8685–8690
6. Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Molecular Biology and Evolution, 2008, 25(12): 2699–2707
7. Gohlke J M, Thomas R, Zhang Y, Rosenstein M C, Davis A P, Murphy C, Becker K G, Mattingly C J, Portier C J. Genetic and environmental pathways to complex diseases. BMC Systems Biology, 2009, 3: 46
8. Yu W, Clyne M, Khoury M J, Gwinn M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics, 2010, 26(1): 145–146
9. Ortutay C, Vihinen M. Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Research, 2009, 37(2): 622–628
10. Wu X, Liu Q, Jiang R. Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics, 2009, 25(1): 98–104
11. Ozgür A, Vu T, Erkan G, Radev D R. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics, 2008, 24(13): i277–i285
12. Ideker T, Sharan R. Protein networks in disease. Genome Research, 2008, 18(4): 644–652
13. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105 (11): 4323–4328

14. Kann M G. Protein interactions and disease: computational approaches to uncover the etiology of diseases. Briefings in Bioinformatics, 2007, 8(5): 333–346
15. Björkholm P, Sonnhammer E L. Comparative analysis and unification of domain-domain interaction networks. Bioinformatics, 2009, 25(22): 3020–3025
16. Adie E A, Adams R R, Evans K L, Porteous D J, Pickard B S. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics, 2005, 6: 55
17. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L C, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. Gene prioritization through genomic data fusion. Nature Biotechnology, 2006, 24(5): 537–544
18. Chen J, Bardes E E, Aronow B J, Jegga A G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research, 2009, 37(Web Server issue): W305–W311
19. Köhler S, Bauer S, Horn D, Robinson P N. Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics, 2008, 82(4): 949–958
20. Sun J, Jia P, Fanous A H, Webb B T, Van Den Oord E J, Chen X, Bukszar J, Kendler K S, Zhao Z. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. Bioinformatics, 2009, 25(19): 2595–2602
21. Tranchevent L C, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B D, Aerts S, Moreau Y. ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Research, 2008, 36(Web Server issue): W377–W384
22. Raghavachari B, Tasneem A, Przytycka T M, Jothi R. DOMINE: a database of protein domain interactions. Nucleic Acids Research, 2008, 36(Database issue): D656–D661
23. Ng S K, Zhang Z, Tan S H, Lin K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Research, 2003, 31(1): 251–254
24. Ng S K, Zhang Z, Tan S H, Radev D R. Integrative approach for computationally inferring protein domain interactions. Bioinformatics, 2003, 19(8): 923–929
25. Finn R D, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics, 2005, 21(3): 410–412
26. Van Driel M A, Bruggeman J, Vriend G, Brunner H G, Leunissen J A. A text-mining analysis of the human phenome. European Journal of Human Genetics, 2006, 14(5): 535–542
27. Altshuler D, Daly M, Kruglyak L. Guilt by association. Nature Genetics, 2000, 26(2): 135–137
28. Wang W, Zhang W, Jiang R, Luan Y. An approach to the discovery of associations of protein domains and complex diseases. In: Proceedings of the Seventh Asia Pacific Bioinformatics Conference. 2009, 908
29. Wang W. Statistical modeling for analysis of biological high-throughput data and its application. Dissertation for the Doctoral Degree. Jinan: Shandong University. 2009, 51–62
30. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek B E, Martin M J, McGarvey P, Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics, 2009, 10: 136
31. Finn R D, Tate J, Mistry J, Coggill P C, Sammut S J, Hotz H R, Ceric G, Forslund K, Eddy S R, Sonnhammer E L, Bateman A. The Pfam protein families database. Nucleic Acids Research, 2008, 36 (Database issue): D281–D288
32. Stein A, Panjkovich A, Aloy P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. Nucleic Acids Research, 2009, 37(Database issue): D300–D304
33. Stein A, Russell R B, Aloy P. 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Research, 2005, 33(Database issue): D413–D417
34. Lee H, Deng M, Sun F, Chen T. An integrated approach to the prediction of domain-domain interactions. BMC Bioinformatics, 2006, 7: 269
35. Brunner H G, Van Driel M A. From syndrome families to functional genomics. Nature Reviews Genetics, 2004, 5(7): 545–551
36. Rhead B, Karolchik D, Kuhn R M, Hinrichs A S, Zweig A S, Fujita P A, Diekhans M, Smith K E, Rosenbloom K R, Raney B J, Pohl A, Pheasant M, Meyer L R, Learned K, Hsu F, Hillman-Jackson J, Harte R A, Giardine B, Dreszer T R, Clawson H, Barber G P, Haussler D, Kent W J. The UCSC genome browser database: update 2010. Nucleic Acids Research, 2010, 38(Database issue): D613–D619
37. Robinson P N, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics, 2008, 83(5): 610–615
38. Lussier Y A, Liu Y. Computational approaches to phenotyping: high-throughput phenomics. Proceedings of the American Thoracic Society, 2007, 4(1): 18–25
39. Oti M, Huynen M A, Brunner H G. The biological coherence of human phenome databases. The American Journal of Human Genetics, 2009, 85(6): 801–808
40. Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning. Cambridge: MIT Press, 2006