Engineering, **2013, 5, 181-188**
http://dx.doi.org/10.4236/eng.2013.510B039 Published Online October 2013 (http://www.scirp.org/journal/eng)

Semi-Global Inference in Phenotype-Protein Network

Siliang Xia[1], Guangri Quan[2], Yongbo Zhao[3], Xuhui Jia[4]
[1]Institute of Architecture of Application Systems, University of Stuttgart, Stuttgart, Germany
[2]School of Computer Science, Harbin Institute of Technology at Weihai, Weihai, China
[3]Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China
[4]Department of Computer Science, The University of Hong Kong, Hong Kong, China
Email: xiasiliang.hit@gmail.com

Received May 2013

ABSTRACT

Discovering genetic basis of diseases is an important goal and a challenging problem in bioinformatics research. Inspired by network-based global inference approach, Semi-global inference method is proposed to capture the complex associations between phenotypes and genes. The proposed method integrates phenotype similarities and protein-protein interactions, and it establishes the profile vectors of phenotypes and proteins. Then the relevance between each candidate gene and the target phenotype is evaluated. Candidate genes are then ranked according to relevance mark and genes that are potentially associated with target disease are identified based on this ranking. The model selects nodes in integrated phenotype-protein network for inference, by exploiting Phenotype Similarity Threshold (PST), which throws lights on selection of similar phenotypes for gene prediction problem. Different vector relevance metrics for computing the relevance marks of candidate genes are discussed. The performance of the model is evaluated on Online Mendelian Inheritance in Man (OMIM) data sets and experimental evaluation shows high performance of proposed Semi-global method outperforms existing global inference methods.

Keywords: Diseases Gene Prioritization; Phenotype-Protein Network; Semi-Global Inference; Phenotype Similarity Threshold

1. Introduction

It is challenging for biomedical research to figure out the genetic basis of diseases. Traditional biology researchers adopt linkage analysis and association studies [1] to discover disease genes, which firstly locate disease genes in a chromosome region. However, the resolution of this approach is low and further analysis of candidate genes in a large genomic region is an expensive task, which prevents gene identification even after a region has been detected.

Many studies have tried to discover disease genes with computational methods. Some work related was based on annotations [2-4], or based on sequences [5]. But, the methods rely on functional annotations are limited because only a small part of genes in the genome have been annotated currently and methods based on sequencing is an expensive task. Moreover, they treated disease genes as separate and independent, however, biological processes are not realized by a single molecule, but rather by the complex interactions of proteins, and the breakdown in protein interaction networks could result in diseases [6]. Moreover, some research indicates that phenotypically similar diseases are caused by functionally related genes [7], and the proteins coded by these functionally related genes usually have direct or indirect interactions [8]. From this perspective, disease genes could then be investigated through the interaction networks of disease proteins.

Recently, researchers took advantage of the computing method to build biological network to help explore the relationship among biological information in multiple granularity, and network approach in biology is proposed and under active research [9], which also facilitates disease gene discovery. A wide range of methods are proposed based on network methods for disease gene prioritization [10-16]. A method utilizing Bayesian predictor and ranking of protein complexes linked to human diseases is proposed by Kasper Lage *et al*. to predict genes of human's inherited phenotypes [13]. Xuebing Wu *et al*. proposed network-based global inference approach [14]. These methods achieve some accomplishments in disease gene prioritization, which primarily relies on analysis of the topological properties of PPI networks and the expectation that the products of genes that are associated with similar diseases interact heavily with each other.

Motivated by these existing network based approaches, we propose a network based Semi-global inference model for disease gene prioritization, which selects diseases in

integrated phenotype-protein network for building profile vectors of candidate genes and target disease, by exploiting Phenotype Similarity Threshold (PST). The model evaluates the relevance between candidate genes and the given target phenotype. Candidate genes are then ranked according to relevance marks. Genes that are potentially associated with target disease are prioritized based on this ranking. To evaluate the effectiveness of the model, the proposed model is tested on known phenotype and gene pairs from OMIM. Our research has three contributions:

- Semi-global inference method with Phenotype Similarity Threshold (PST) is proposed to prioritize candidate disease genes. The experimental result shows the proposed Semi-global method outperforms existing global inference method.
- Phenotype Similarity Threshold (PST) is defined to make a difference between high similarities and low similarities and to distinguish between diseases closely related to target disease and diseases less related, which specifies phenotypes in the network to be considered and exploited for inference. Two methods (S-PST, D-PST) to get PST are introduced and compared.
- Performance of proposed model with different vector relevance metrics (Pearson correlation coefficient, Euclidean distance and Cosine similarity) are evaluated and compared. We show that Semi-global inference works well with Euclidean distance and Cosine similarity.

In Section 2, we briefly introduce the background of network based candidate gene prioritization by describing the problem formally and discussing the related work and their limitations. Section 3 presents Semi-global inference model and explains strategies of PST to select nodes in phenotype-protein network. Section 4 shows experimental results of proposed Semi-global inference model with variation of relevance metrics and PST，and comprehensively compares the performance of proposed model against an existing global inference method. In Section 5, we draw some conclusions and point out further work.

2. Background

2.1. Network Based Candidate Gene Prioritization

Here is a brief description of network-based disease gene prioritization problem referring to [17]: given target disease d, the input to the candidate disease gene prioritization problem consists of two sets of genes, known set K and candidate set C. The known set K contains prior knowledge of the disease d, e.g., it is the set of genes known to be associated with d and diseases similar to d. Each gene $g \in K$ is associated with a similarity score $\sigma(g, d)$, indicating the known degree of association between g and d. The candidate set C contains candidate genes, one or more of which is potentially associated with target disease d (e.g., these genes might be in the linkage interval of d that is identified by association studies). The purpose of network based disease prioritization is to use a PPI network $G = (V, E)$, to compute a score $\varphi(v, D)$ for each gene $g \in C$ that represents the likelihood of g to be associated with d.

The PPI network $G = (V, E)$ consists of a set of gene products V and a set of undirected interactions E between these gene products, in which $uv \in E$ represents an interaction between $u \in V$ and $v \in V$. In this network, the set of interacting partners of a gene product $v \in V$ is defined as $N(v) = \{u \in V: uv \in E\}$.

Global prioritization methods use this network information to compute φ by propagating σ over G. Candidate genes with high relevance to target disease of interest are ranked in the top and are regarded as the disease genes.

2.2. Related Work

Xuebing Wu *et al*. have proposed network-based global inference approach called CIPHER algorithm [14], in which Pearson correlation coefficient is adopted to evaluate the relevance between candidate genes and the target disease. Another global inference method is proposed based on a network propagation algorithm to formulate constraints on the prioritization function [16].

Although these existing global network based methods to some extent throw lights on disease gene prioritization problems, they have some drawbacks and limitations. Research of Xuebing Wu *et al*. is based on the assumption of the linear correlation between profiles of phenotypes and disease genes, which shows some bias against genes whose related proteins have few interactions with other peers [14]. Moreover, as reported in literatures, network based global inference methods, favor genes whose products are highly connected in the network and perform poorly in identifying loosely connected disease genes, due to centrality of target disease genes [17] and incomplete and noisy nature of the PPI data [18].

In global inference method, all the diseases in the phenotype similarity network are exploited to generate a prediction, including less related diseases to profile a target disease, which fails to take into consideration that more similar diseases may play more important roles in inference. No work has been done for disease gene prioritization using only parts of diseases in phenotype network, and nodes selection strategy has not been explored. Secondly, phenotype similarities vary. A target disease has different phenotype similarities to other diseases in the network. No selection criteria is made to treat roles of diseases differently in phenotype network,

no methods make a difference between high similarities and low similarities, which might be considered to determine which related diseases to refer in gene prioritization problems.

Our research aims at exploring the uncovered areas mentioned and overcoming limitations of global inference methods. We propose Semi-global inference method by exploiting PST as the criteria to select phenotypes in network for inference, which is the essential difference between proposed Semi-global model and existing global inference methods.

3. Methodology

In this section we present the mathematical model and show the general framework of gene prioritization algorithm of Semi-global inference. Furthermore, we explain how Phenotype Similarity Threshold is exploited for nodes selection in phenotype network, which is the core of Semi-global inference model.

It is important to note that the purpose here is to infer functional associations between genes from functional and physical interactions between their products. For this reason, any reference to interactions between genes in this paper refers to the interactions between their products. Meanwhile, disease gene prioritization is inferred from phenotypically similar diseases, term disease and term phenotypes deliver identical conception in this paper.

3.1. Mathematical Model

- Undirected graph

$$GPhenotyp =< P, E_p > \quad (1)$$

is defined as phenotype similarity network;
$P = \{p_1, p_2, p_3 \ldots p_m\}$ is a subset of all the phenotypes, $|P| = m$; $E_p = \{S_{jk} \mid p_j, p_k \in P\}$ and the element S_{jk} is the similarity of phenotypes p_i, p_k.

- Undirected graph

$$GProtein =< G, E_g > \quad (2)$$

is defined as protein interaction network;
$G = \{g_1, g_2, g_3 \ldots g_n\}$ is a subset of all the proteins, $|G| = n$; $E_g = \{I_{jk} \mid g_j, g_k \in G\}$ and the element I_{jk} denotes the interaction of proteins g_i, g_k.

- Given a phenotype $p_i \in P$, $\forall g_k \in G$, set

$$CAssociation_j = \{< p_j, g_k > \mid g_k \in G \wedge g_k \text{ is associated with } p_j\} \quad (3)$$

is defined as association set of p_j; each element $< p_j, g_k >$ in $CAssociation_j$ is an association of p_j; set

$$CAssociation = \bigcup_{j=1}^{m} CAssociation_j \quad (4)$$

is defined as global association set, which contains all phenotype-protein associations.

- Given phenotype similarity network *GPhenotype*, protein interaction network *GProtein* and global association set *CAssociation*, set

$$N =< GPhenotype, GProtein, CAssociation > \quad (5)$$

is phenotype-protein network.

- Given a phenotype $p_j \in P$ and a protein $g_k \in G$, $\forall p_x \in P \forall g_v \in G$,

$$f(g_k, p_x) = Max\{I_{ky} \mid < p_x, g_y > \in CAssociation_x\} \quad (6)$$

denotes one dimension of the profile vector of protein g_k.

- Phenotype Similarity Threshold (PST) is a manually set similarity value that satisfies

$$Min\{s_{jk} \mid p_j, p_k \in P\} \le PST \le Max\{s_{jk} \mid p_j, p_k \in P\} \quad (7)$$

- Given a phenotype $p_i \in P$, set

$$CRP_j = \{p_k \mid p_k \in P \wedge s_{jk} \ge PST\} \quad (8)$$

contains the phenotypes that have similarities higher or equals to PST with p_j. Each element in CRP_j is defined as a Closely Related Phenotype of p_i.

- Given a phenotype $p_j \in P$, $\forall 1 \le r \le m$, if $s_{ji_r} \ge PST$, then S_{ji_r} is used as a dimension in profile vector of p_i; vector

$$\boldsymbol{p_j} = (S_{ji_1}, S_{ji_2}, S_{ji_3} \ldots S_{ji_r}) \quad (9)$$

characterizes the profile of p_j in phenotype similarity network, in which $i_1 < i_2 \ldots < i_r$. Means only the similarities of Closely Related Phenotypes (higher than PST) are used to build the profile vector of a target phenotype of interest.

- Given a phenotype $p_j \in P$ and a protein $g_k \in G$, $\forall 1 \le r \le m$, if $p_{ji_r} \in CRP_j$, then $f(g_k, p_{i_r})$ is used as a dimension of vector of g_k; vector

$$\boldsymbol{g_k} = (f(g_k, p_{i_1}), f(g_k, p_{i_2}), f(g_k, p_{i_3}) \ldots f(g_k, p_{i_r})) \quad (10)$$

characterizes the profile of g_k in Protein Network, in which $i_1 < i_2 \ldots < i_r$.

Given a phenotype $p_j \in P$ and a protein $g_k \in G_j$, let $\varphi(p_j, g_k)$ denote a relevance metric of vector $\boldsymbol{p_j}$ and vector $\boldsymbol{g_k}$. Three different metrics are defined, which characterize the correlation between profile vectors of protein g_k and phenotype p_j and thus indicate the relevance of candidate protein g_k and target phenotype p_i. Let φ_1 denote Euclidean distance of two vectors,

$$\varphi_1(p_j, g_k) = \left(\sum_{i=1}^{m} (S_{ji} - f(g_k, d_i))^2\right)^{\frac{1}{2}} \quad (11)$$

Cosine similarity is a metric measuring the included angle of two vectors, which is denoted as φ_2,

$$\varphi_2(p_j, g_k) = \cos(p_j, g_k) = \frac{p_j \cdot g_k}{|\, p_j \,||\, g_k \,|} \quad (12)$$

Pearson correlation coefficient indicates linear correlation between two vectors, which is denoted as φ_3,

$$\varphi_3(p_j, g_k) = \frac{\mathrm{cov}(p_j, g_k)}{\sigma(p_j)\sigma(g_k)} \quad (13)$$

3.2. Phenotype Similarity Threshold (PST)

According to the biological assumption that phenotypically similar diseases are caused by functionally related genes [7], the proposed Semi-global inference model takes into consideration only phenotypes that are highly similar to target disease, with similarities higher than PST. We use only those Closely Related Phenotypes (refer to (8)) of p_j and exploit corresponding similarities to characterize the target phenotype. Therefore, In (9) and (10), given a phenotype p_j, the dimensions of profile vector $\boldsymbol{p_j}$ are determined by the number of phenotypes in CRP_j, the dimensions of profile vector of candidate genes are reduced correspondingly.

3.3. Semi-Global Inference

Based on the mathematical model above, here we give the computation framework of proposed semi-global inference method, which consists of two algorithms to prioritize candidate disease genes.

Algorithm 1 *Relevance Mark Calculation* calculates the relevance mark for a given pair of target phenotype p_j and candidate protein g_k. Algorithm 2 *Disease Gene Prioritization* takes a target phenotype as the input and evaluates relevance mark for all candidate proteins in linkage interval, then prioritizes the candidate proteins based on their relevance marks. Proteins with high relevance mark are regarded highly related to target phenotype and thus genes associate with these top ranked proteins are the underlying causing genes of target disease, as the predictive result of Semi-global inference model.

In practice, each of metrics (11) or (12) (13) are tested respectively in Algorithm 1 as relevance evaluation of candidate proteins. Algorithm 2 is invoked to prioritize candidate genes for all phenotypes we are interested in.

4. Results

In this section, we comprehensively evaluate the performance of proposed Semi-global inference model with different setting of metrics and PSTs. Then we compare proposed model to global inference method.

4.1. Datasets

To evaluate the proposed model, data sets needed are listed as follows: Phenotype set and quantified similarities between each pair of phenotypes. Protein set and quantified protein interaction between each pair of proteins. Set of known pairs (associations) of phenotypes and associated proteins, which serves as the validation set.

Phenotype set and their linkage intervals are obtained from Online Mendelian Inheritance in Man (OMIM) Morbid Map [19], which provides a publicly accessible and comprehensive database of genotype-phenotype relationship in humans; phenotype similarities come from the research of van Driel *et al*. [20]; quantified protein interaction marks are extracted from STRING database [21] to build PPI network; chromosome mapping of proteins are extracted from Ensembl database [22]; validation set can be built from phenotype-protein network, by extracting the phenotype-gene mapping from OMIM Morbid Map and gene-protein mapping from bioDBnet database [23] and mapping phenotype network to PPI network.

Those phenotypes that can not be mapped to proteins are removed, due to lack of known associated genes or incomplete information of proteins coded by genes in the linkage interval. We finally get 1897 phenotypes and 84652 proteins in total, while only 156584 protein-protein interactions are available. Those missing PPI records are regarded as zero. 2549 known phenotype-gene pairs are maintained for evaluation.

4.2. Experimental Setting

We apply leave-one-out cross-validation in order to evaluate the performance of different methods in terms of accuracy of disease gene prioritization. For each disease of interest, we conduct following experiment:

- We remove all associations of this target disease from global association set (refer to (4)).
- All the genes in the linkage interval are regarded as candidate genes to be prioritized. On average, there are 750 candidate genes in the linkage interval of a disease.
- In practice, we exploit **Position Parameter** λ to get PST: Phenotype similarities are sorted in an array in ascending order, then PST is assigned as the value retrieved from the array with index of (*array size* * λ), so PST is determined by assigning λ a value from zero to one. It is important to note that when $\lambda = 0$, all the nodes in phenotype network are considered in inference. In this case, Semi-global model degenerates into global inference. Thus, global inference method is a case of proposed Semi-global model when $\lambda = 0$. We conduct experiment with two methods to get PST:

Static method (S-PST). All the phenotype similarities are sorted in one array. PST is a global static value for all target diseases during the experiment.

Dynamic method (D-PST). PST is retrieved from a smaller phenotype similarity set containing only the similarities related to current target disease. Different PSTs are gained for prediction of different target diseases, according to the similarity range of that target disease.

- We conduct the experiment with each combination of relevance metrics and PST methods.
- In order to systematically compare the performance of proposed model, we use following evaluation criteria:

Average Rank. Average rank in proposed model of known disease genes.

Fold Enrichment. Ability to enrich known disease genes over random selection [13].

Distribution of Cases. Percentage of the test cases ranked within top 1%, top 5% and top 10%.

4.3. Experiment with Variation of PST and Relevance Metrics

Proposed model with Euclidean distance shows a rapid increase of average rank with the increase of λ, though the performance is always poorer than that of model with the other two relevance metrics. The model exhibits a high average rank with high PST (high Position Parameter λ) using S-PST, in spite of relevance metrics adopted.

For model with Euclidean distance and Cosine similarity, fold enrichment gets higher along with the increase of λ. On the other hand, **Figure 1** to **Figure 4** show that proposed model with Cosine similarity gets higher performance than the other two relevance metrics. Moreover, the trend of the performance with increasing λ shows that the model gains better performance when highly similar diseases are referred to profile target disease and candidate genes, in which the profile vectors consist of only a few dimensions and only small part of nodes (eg. diseases holding top 5% highest similarities in whole phenotype network in S-PST and diseases holding

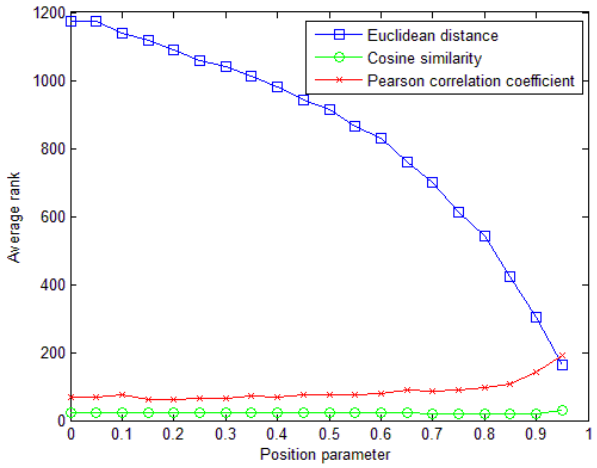

Figure 1. Average rank to compare the performance of proposed model using S-PST with different relevance metrics.

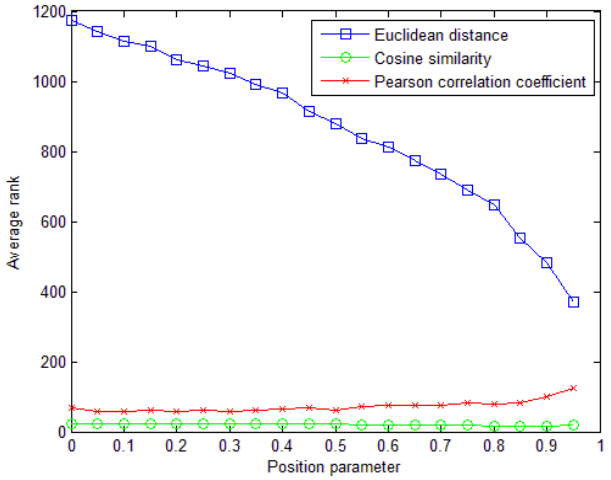

Figure 2. Average rank to compare the performance of proposed model using D-PST with different relevance metrics.

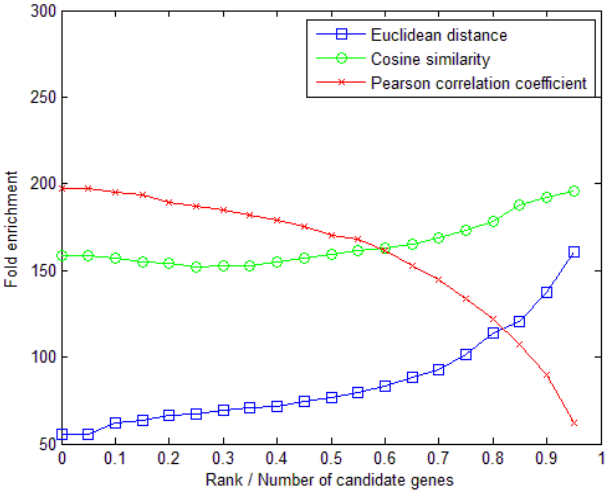

Figure 3. Fold enrichment to compare the performance of S-PST model with different metrics.

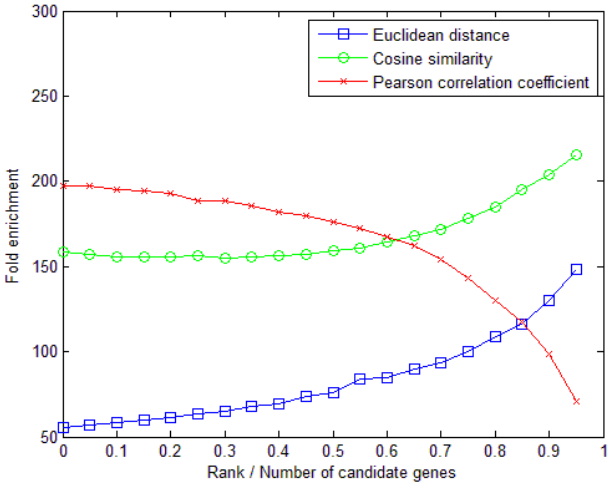

Figure 4. Fold enrichment to compare the performance of S-PST model with different metrics.

top 5% highest similarities to the target disease in D-PST) are exploited. Therefore it indicates the strategy that referring only part diseases in proposed Semi-global model works well with these two relevance metrics (especially with Euclidean distance) and nodes selection with PST and dimension reduction of profile vectors achieves performance improvement.

Model with Pearson correlation coefficient reaches its best performance when $\lambda = 0$ (global inference method) and shows a decline with increase of λ. Therefore, proposed Semi-global inference does not increase the performance if Pearson correlation is adopted as the relevance metric.

4.4. Comparison to Global Inference Method

Here we discuss the cases when D-PST is exploited with a certain λ assigned to get the relative high performance using different relevance metrics and compare them to global inference method using CIPHER algorithm [14] with the same relevance metric.

Table 1 and **Table 2** demonstrate that Semi-global model with D-PST and high λ outperforms global inference method using same relevance metrics. Especially for Euclidean distance, when λ is assigned with a high value, Semi-global model shows much higher performance than global inference.

D-PST and Cosine similarity work as the best combination, with which proposed model reaches a high performance with fold enrichment being 217.62 when $\lambda = 0.96$ and average rank being 16.29 when $\lambda = 0.92$. In this configuration, Semi-global model takes into account top 4% most similar diseases of the target disease for inference, outperforms the highest fold enrichment of 197.60 and average rank of 22.31 in global inference method.

Table 3 shows that with same relevance metrics, more known disease genes are ranked within top 1%, top 5% and top 10% in proposed Semi-global model using D-PST than that in global inference method. It also shows D-PST and Cosine similarity in proposed model achieves better performance than other combinations of PST methods and relevance metrics.

Then, we compare distribution and accumulation of test cases between proposed Semi-global model and global method when they reach their respective high performance with particular experimental settings.

Figures 5 and **6** are general views about distribution and accumulation of test cases of proposed Semi-global model and global inference method with particular settings to reach their high performance which are Semi-global model using D-PST with Euclidean distance and $\lambda = 0.995$, Semi-global model using D-PST with Cosine similarity and $\lambda = 0.96$ and CIPHER algorithm with Pearson correlation coefficient as a representative of global inference. Semi-global model using D-PST and Cosine similarity not only gets better performance in terms of average rank and fold enrichment than global inference, but also generates a more desirable distribution of test cases. It ranks more than 75% cases within top 1%, and the accumulated ratio of test cases is higher than global inference method.

Table 1. Fold enrichment to compare performance of proposed Semi-global model to a global inference method.[1]

	ED	CS	PCC
Global inference (CIPHER)	55.44	159.71	197.60
Semi-global inference (D-PST)	177.03 ($\lambda = 0.991$)	217.62 ($\lambda = 0.96$)	197.60 ($\lambda = 0$)

Table 2. Average rank to compare performance of proposed Semi-global model to a global inference method.[1]

	ED	CS	PCC
Global inference (CIPHER)	1172.60	22.31	67.94
Semi-global inference (D-PST)	81.46 ($\lambda = 0.995$)	16.29 ($\lambda = 0.92$)	56.43 ($\lambda = 0.1$)

[1]ED = Euclidean distance, CS = Cosine similarity, PCC = Pearson correlation coefficient.

Table 3. Percentage of the known disease genes ranked within top 1%, top 5% and top 10% in proposed Semi-global model and a global inference model.[1,2]

		Top 1%	Top 5%	Top 10%
Global inference (CIPHER)	ED	0.21	0.25	0.26
	CS	0.64	0.93	0.97
	PCC	0.72	0.91	0.94
Semi-global inference (D-PST)	ED ($\lambda = 0.995$)	0.59	0.82	0.85
	CS ($\lambda = 0.96$)	0.75	0.94	0.98
	PCC ($\lambda = 0$)	…	…	…

5. Conclusions

In this paper, a Semi-global inference model with PST is proposed for disease gene prioritization, which applies profile vectors in phenotype-protein network to characterize target disease and candidate genes. The model is evaluated comprehensively on OMIM dataset and the experimental result shows proposed Semi-global model outperforms existing global inference method.

Phenotype Similarity Threshold (PST) is proposed and Closely Related Phenotypes are defined. It is adopted as a criterion to select diseases in phenotype network to profile the target disease. Thus, by considering only highly similar diseases, proposed PST has significance in nodes

[2]Last row in **Table 3** is blank because proposed model with Pearson correlation coefficient reaches best performance when $\lambda = 0$, which degenerates into global inference.

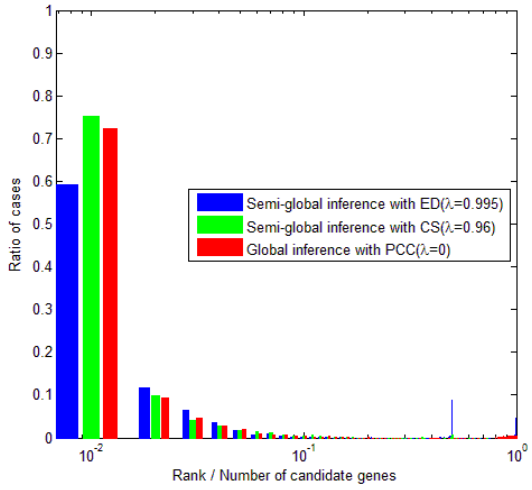

Figure 5. Comparison of distribution of test cases between proposed Semi-global model using D-PST and a global inference method.[1]

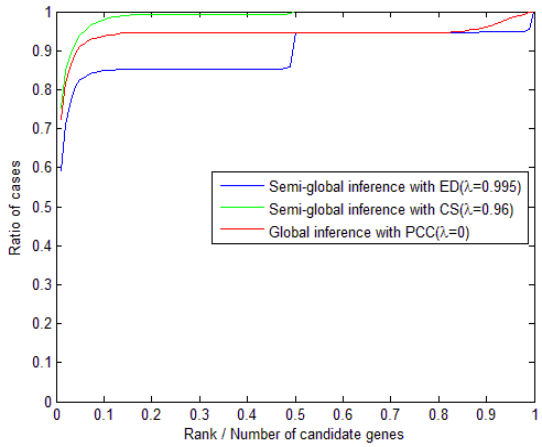

Figure 6. Comparison of accumulation of test cases between proposed Semi-global model using D-PST and a global inference method.[1]

selection in phenotype-protein network for gene prioritization problem, which as a trial demonstrates a novel understanding of the well accepted belief that phenotypically similar diseases are caused by functionally related genes.

Effect of different relevance metrics of profile vectors, different methods and variation of PST on the proposed model are discussed. The proposed model with Cosine similarity as relevance metric shows higher performance than model using other two metrics. Moreover, proposed model achieves performance improvement along with the increase of PST when Cosine similarity and Euclidean distance are adopted as relevance metrics. We have also shown proposed Semi-global model using D-PST exhibits higher average rank, fold enrichment and more admirable distribution than global method.

Further research includes configurations of Semi-global model (proper PST, Position Parameter and relevance metric) to achieve better performance, sensitivity of proposed model to noise of PPI data, and the issue of bias occurs in global inference.

6. Acknowledgements

This research is supported by National Science Foundation of China 60973077.

REFERENCES

[1] D. Botstein and N. Risch, "Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease," *Nature Genetics*, Vol. 33, 2003, pp. 228-237. http://dx.doi.org/10.1038/ng1090

[2] F. S. Turner, D. R. Clutterbuck and C. Semple, "Pocus: Mining Genomic Sequence Annotation to Predict Disease Genes," *Genome Biology*, Vol. 4, 2003, p. R75. http://dx.doi.org/10.1186/gb-2003-4-11-r75

[3] J. Chen, C. Shen and A. Sivachenko, "Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data," *Pacific Symposium on Biocomputing*, Vol. 11, 2006, pp. 367-378.

[4] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders," *Nucleic Acids Research*, Vol. 33, Database Issue, 2005.

[5] E. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. Pickard, "Speeding Disease Gene Discovery by Sequence Based Candidate Prioritization," *BMC Bioinformatics*, Vol. 6, 2005, p. 55. http://dx.doi.org/10.1186/1471-2105-6-55

[6] L. Sam, Y. Liu, J. Li, C. Friedman and Y. A. Lussier, "Discovery of Protein Interaction Networks Shared by Diseases," *Pacific Symposium on Biocomputing*, Vol. 12, 2007, pp. 76-87.

[7] G. Jiminez-Sanchez, *et al*., "Human Disease Genes," *Nature*, Vol. 409, 2001, pp. 853-854 http://dx.doi.org/10.1038/35057050

[8] M. Oti and H. G. Brunner, "The Modular Nature of Genetic Diseases," *Clinical Genetics*, Vol. 71, 2007, pp. 1-11. http://dx.doi.org/10.1111/j.1399-0004.2006.00708.x

[9] J. H. Jing-Dong, "Understanding Biological Functions through Molecular Networks," *Cell Research*, Vol. 18, 2008, pp. 224-237. http://dx.doi.org/10.1111/j.1399-0004.2006.00708.x

[10] J. Chen, B. Aronow and A. Jegga, "Disease Candidate Gene Identification and Prioritization Using Protein Interaction Networks," *BMC Bioinformatics*, Vol. 10, No. 1, 2009, p. 73. http://dx.doi.org/10.1186/1471-2105-10-73

[11] M. Oti, B. Snel, M. A. Huynen and H. G. Brunner, "Predicting Disease Genes Using Protein-Protein Interactions," *Journal of Medical Genetics*, Vol. 43, 2006, pp. 691-698. http://dx.doi.org/10.1136/jmg.2006.041376

[12] S. Navlakha and C. Kingsford, "The Power of Protein

Interaction Networks for Associating Genes with Diseases," *Bioinformatics*, Vol. 26, 2010, pp. 1057-1063. http://dx.doi.org/10.1093/bioinformatics/btq076

[13] K. Lage, E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau and S. Brunak, "A Human Phenome-Interactome Network of Protein Complexes Implicated in Genetic Disorders," *Nature Biotechnology*, Vol. 25, 2007, pp. 309-316. http://dx.doi.org/10.1038/nbt1295

[14] X. B. Wu, R. Jiang, M. Q Zhang and S. Li, "Network-Based Global Inference of Human Disease Genes," *Molecular Systems Biology*, Vol. 4, 2008, p. 189. http://dx.doi.org/10.1038/msb.2008.27

[15] S. Kohler, S. Bauer, D. Horn and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," *The American Journal of Human Genetics*, Vol. 82, No. 4, 2008, pp. 949-958. http://dx.doi.org/10.1016/j.ajhg.2008.02.013

[16] Y. Li and J. C. Patra, "Genome-Wide Inferring Gene-Phenotype Relationship by Walking on the Heterogeneous Network," *Bioinformatics*, Vol. 26, No. 9, 2010, pp. 1219-1224. http://dx.doi.org/10.1093/bioinformatics/btq108

[17] S. Erten and M. Koyut¨urk, "Role of Centrality in Network-Based Prioritization of Disease Genes," *Proceedings of the* 8*th European Conf. Evolutionary Computation, Machine Learning, and Data Mining in Bioinformatics* (*EVOBIO*'10), Vol. LNCS 6023, 2010, pp. 13-25.

[18] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt and M. Gerstein, "Bridging Structural Biology and Genomics: Assessing Protein Interaction Data with Known Complexes," *Trends in Genetics*, Vol. 18, No. 10, 2002, pp. 529-536. http://dx.doi.org/10.1016/S0168-9525(02)02763-4

[19] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), World Wide Web URL: http://omim.org/

[20] M. A. van Driel, J. Bruggeman, G. Vriend, *et al.*, "A Text-Mining Analysis of the Human Phenome," *European Journal of Human Genetics*, Vol. 14, 2006, pp. 535-542. http://dx.doi.org/10.1038/sj.ejhg.5201585

[21] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, *et al*., "The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored," *Nucleic Acids Research*, Vol. 39, 2011, pp. D561-D568. http://dx.doi.org/10.1093/nar/gkq973

[22] E. Birney, D. Andrews, M. Caccamo, Y. Chen, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, *et al*., "Ensembl 2006," *Nucleic Acids Research*, Vol. 34, 2006, pp. D556-D561. http://dx.doi.org/10.1093/nar/gkj133

[23] U. Mudunuri, A. Che, M. Yi and R. M. Stephens, "bioDBnet: The Biological Database Network," *Bioinformatics*, Vol. 25, No. 4, 2009, pp. 555-556. http://dx.doi.org/10.1093/bioinformatics/btn654

