Prediction and validation of disease genes using HeteSim Scores

Xiangxiang Zeng, *Member, IEEE,* Yuanlu Liao, Yuansheng Liu, and Quan Zou*, *Member, IEEE*

Abstract—Deciphering the gene disease association is an important goal in biomedical research. In this paper, we use a novel relevance measure, called HeteSim, to prioritize candidate disease genes. Two methods based on heterogeneous networks constructed using protein-protein interaction, gene-phenotype associations, and phenotype-phenotype similarity, are presented. In HeteSim_MultiPath (HSMP), HeteSim scores of different paths are combined with a constant that dampens the contributions of longer paths. In HeteSim_SVM (HSSVM), HeteSim scores are combined with a machine learning method. The 3-fold experiments show that our non-machine learning method HSMP performs better than the existing non-machine learning methods, our machine learning method HSSVM obtains similar accuracy with the best existing machine learning method CATAPULT. From the analysis of the top 10 predicted genes for different diseases, we found that HSSVM avoid the disadvantage of the existing machine learning based methods, which always predict similar genes for different diseases. The data sets and Matlab code for the two methods are freely available for download at http://datamining.xmu.edu.cn/~xzeng/dgassociations/klk/WebRoot/index.jsp.

Index Terms—disease gene prediction, HeteSim, multipath methods, HSMP, HSSVM.

1 Introduction

THE recognition of disease genes has long been an important goal of biomedical research, which may contribute to the improvement of medical care and the understanding of gene functions, interactions, and pathways. Traditional gene-mapping approaches, such as linkage analysis and association studies [1], have made a large contribution to this, but these methods have some disadvantages. Linkage analysis can associate disease traits with specific genomic regions, but these regions often contain tens or even hundreds of genes. Sequencing all the candidate genes in a particular region is still a time-consuming and expensive task. Although association studies work well when applied to a set of carefully selected functional candidate genes, the selection of functional candidates is not straightforward and often limited by specialized knowledge.

With the understanding that phenotypically similar diseases are often caused by functionally related genes, network-based approaches were proposed for prioritizing gene-disease associations [2]. More recently, Wu *et al.* [3] construct a gene-phenotype heterogeneous network, which is composed of a PPI network from HPRD, a disease gene-phenotype associations network obtained from the Online Mendelian Inheritance in Man (OMIM) database [4], [5], [6], and a phenotype-phenotype similarity dataset calculated through text mining [7], to infer human disease genes. Various computational methods based on networks have been proposed for prioritizing gene-disease associations, for example, CIPHER [3], Random Walk [8], Diffusion Kernel [8], PRINCE [9], and RWRH [10]. Inspired by social network analyses, Singh-Blom *et al.* [11] introduce the Katz method, which has been successfully applied for link prediction in social networks [12], into the disease genes prediction problem. When the gene-disease association problem is viewed as a supervised learning problem, machine learning methods such as ProDiGe [13] and CATAPULT [11] are proposed to prioritize candidate disease genes.

Although many fruitful network-based algorithmic approaches have been developed for prioritizing gene-disease associations, most of these methods simply view objects in gene-phenotype heterogeneous networks as the same type and do not consider the different semantic meanings behind the paths. For example, Katz and CATAPULT use only walk count to find the similarity between objects. This approach tends to cause gene objects that have higher number of the associated phenotypes always to be identified as disease genes.

The HeteSim [14] is a path-based measure to calculate relevance between objects in heterogeneous network. It can capture effectively the subtle semantics of paths, which is meaningful for calculating the relevance between nodes in heterogeneous networks. In addition, the pair-wise random walk used in this method can be used to calculate the relatedness of different-typed nodes as well as same-typed nodes. An example of comparing walk count and HeteSim is illustrated in Fig. 1. The example shows that the walk count between a and c more than the walk count between b and c. The walk count tends to evaluate nodes with higher degree have higher similarity than others. Thus, Katz and CATAPULT measure that $a - c$ has a higher similar score than $b - c$. However, we found that each of the connections starting from a possess less meaning than the connections starting from b. We believe that the connections between b and c stronger than the connections between a and c. The similarity calculated by the HeteSim measure seems to be a more reasonable result.

- *X. Zeng and Y. Liao are with Department of Computer Science, Xiamen Unviersity, Xiamen 361005, Fujian, China. (e-mail: xzeng@xmu.edu.cn, yuanluliao@stu.xmu.edu.cn)*
- *Y. Liu is with School of Software, Dalian University of Technology, Dalian 116024, China.*
- *Q. Zou* is with School of Computer Science and Technology, Tianjin University, Tianjin 300072, China. (e-mail: zouquan@xmu.edu.cn)*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520947, IEEE/ACM Transactions on Computational Biology and Bioinformatics

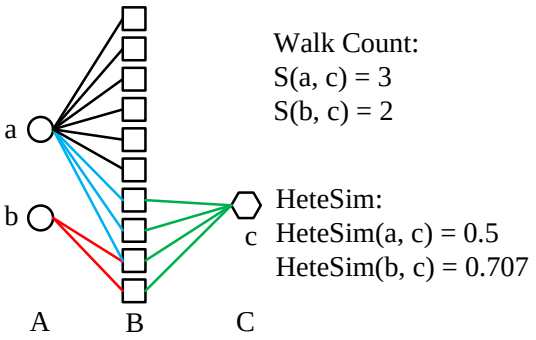


Fig. 1. Example of heterogeneous network for comparing walk count and HeteSim measure. Circles, squares, and sexangle denote three different types A, B, and C, respectively.

Here we propose two novel multipath methods, HeteSim_MultiPath (HSMP) and HeteSim_SVM (HSSVM), based on the HeteSim measure. HSMP uses the HeteSim measure to calculate the similarity between nodes in heterogeneous networks. Then, the HeteSim scores of different paths are combined with a constant that dampens contributions from longer paths. HSSVM also uses the HeteSim measure to calculate similarity. However, HSSVM uses a machine learning method instead of a constant to combine HeteSim scores. A Positive Unlabeled (PU) learning method was chosen to learn the different weights of the different paths because only positive and unlabeled examples were considered here.

A cross-validation method was introduced to evaluate the performance of the two proposed methods, two recently proposed methods (Katz and CATAPULT), and two state-of-the-art methods (PRINCE and ProDiGe). We found that HSMP and HSSVM not only outperformed PRINCE and ProDiGe, they were also slightly better than Katz and CATAPULT in terms of the top predictions. We also evaluated the average overlap ratio and average number of the associated phenotypes with phenotype of the predicted top 10 genes and compared them with other methods. We found that the top 10 disease genes predicted by HSMP and HSSVM had lower overlap ratios and lower number of the associated phenotypes than the top 10 disease genes predicted by the other methods, showing that the two new methods are reasonable and credible.

2 METHODS

2.1 Datasets

In the subsection, we briefly introduce the three networks are employed to construct heterogeneous network in our experiments.

Gene-gene interaction network: Two different networks HumanNet [15] and HPRD network [16] are used. The HumanNet is a functional gene network for human genes and others orthologous genes of yeast, worm, and fly. It was constructed by Lee *et al.* in 2011, and contains a variety of sources of information such as diverse expression, protein interaction, and gene co-expression. There are $16,243$ genes and $476,399$ non-zero functional linkages in HumanNet. HPRD network is a much sparser protein-protein interaction network. There are $41,327$ protein-protein interactions among $30,047$ protein entries. For any two distinct proteins, their corresponding protein-coding genes are connected if their interact with each other in the HPRD network. The two networks are also used to identify associations of genes with diseases in [11].

Gene-phenotype association network: The gene-phenotype associations were collected from 9 different species: Human (Hs), Plant (At), Worm (Ce), Fruit fly (Dm), Mice (Mm), Yeast (Sc), Escherichia coli (Ec), Zebrafish (Dr), and Chicken (Gg). The data set were collected by Singh-Blom *et al.* from different literature and public databases. It was downloaded from literature [11]. The network contains $16,153$ phenotypes, and $362,987$ gene-phenotype associations.

Phenotype similarity network: The phenotype similarities for human were derived solely from the MimMiner [7], which is a text-mining approach to evaluate the similarities between human phenotypes from the OMIM database [6]. According to the analysis of Vanunu *et al.* [17], a logistic transformation $L(x) = \frac{1}{1+exp(cx+d)}$ was applied to adopt to the process, where x represents the weights between phenotypes, c is the parameter tuned by cross validation, and $d = \log(9999)$. The phenotype similarities for the other species are simply set to zero.

2.2 Construction of the heterogeneous network

We construct the heterogeneous network by connecting the gene interaction network and phenotype similarity network utilizing the bipartite graph of the gene-phenotype association network. The schema of the heterogeneous network is illustrated in Fig. 2. The network then contains 10 types of objects. As an abbreviation, we use Ge to denote the type of gene, and corresponding binomial name to denote different species, e.g., Hs for human and Dm for fruit fly. A path $\mathcal{P}$ is defined at the object type level, and is denoted in the form of $A_1 \rightarrow \cdots \rightarrow A_i \rightarrow \cdots \rightarrow A_{l+1}$, where A_i represent the object type. We list two examples of path in Figs. 3(a) and 3(b).

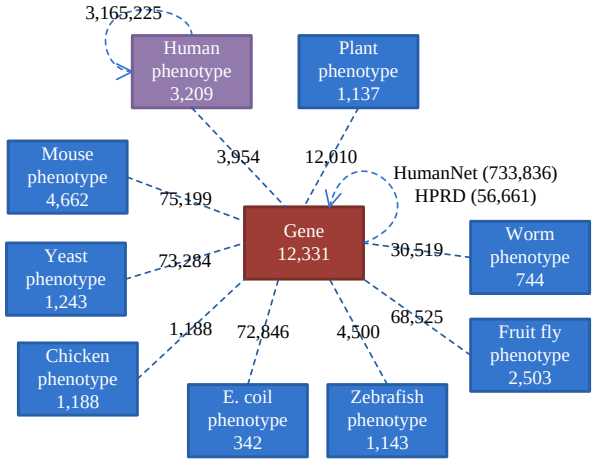


Fig. 2. Illustration of the heterogeneous network schema. Squares represent object types and the dash lines represent the associations. The numbers in the square represent the total number of entity in corresponding object types. The numbers on the dash lines are the number of associations between entities of two different object types.

Suppose that matrixes $\mathbf{G}$, $\mathbf{Q}$ and $\mathbf{P}$ are adjacency matrix for gene-gene interaction network, phenotype similarity

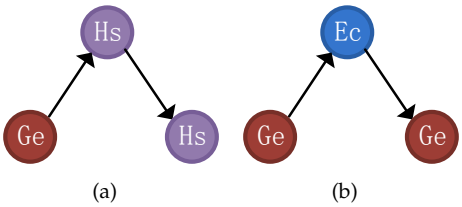


Fig. 3. The path examples: (a) GeHsHs; (b) GeEcGe.

network and gene-phenotype association network, respectively. Therefore, the adjacency matrix of the heterogeneous network can be expressed as

$$\mathbf{H} = \begin{bmatrix} \mathbf{G} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{Q} \end{bmatrix},$$

where $\mathbf{P}^T$ is the transpose of $\mathbf{P}$. Let $\mathbf{Q}_{Hs}$ and $\mathbf{Q}_s$ denote the adjacency matrix of phenotype similarity network of Hs and other species, respectively. There is no phenotype similarity for other species. Thus, we can obtained $\mathbf{Q}_s = [\mathbf{0}]$ and

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{Hs} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_s \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{Hs} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

2.3 Model description

2.3.1 HeteSim measure

The HeteSim measure is a path-constrained measure that can be used to calculate the relatedness of heterogeneous objects with same or different types in a uniform framework. It has been proven that HeteSim has some good properties [14], such as self-maximum and symmetric, and has shown its potential to mining valuable information in heterogeneous network. Therefore, the HeteSim measure was used to calculate the relatedness between genes and human phenotypes.

Definition 1 (Transition probability matrix). [14] A and B are two object types in a heterogeneous network, $(\mathbf{W}_{AB})_{n\times m}$ is an adjacent matrix between type A and B. The transition probability matrix of $A \to B$ can be expressed as

$$\mathbf{T}_{AB}(i,j) = \frac{\mathbf{W}_{AB}(i,j)}{\sum_{k=1}^{m} \mathbf{W}_{AB}(k,j)}.$$

In other word, $\mathbf{T}_{AB}$ is a normalized matrix of $\mathbf{W}_{AB}$ along the row vector.

Definition 2 (Reachable probability matrix). [14] Given a heterogeneous network, a reachable probability matrix for path $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$ is defined as

$$\mathbf{R}_{\mathcal{P}} = \mathbf{T}_{A_1A_2}\mathbf{T}_{A_2A_3}\cdots\mathbf{T}_{A_lA_{l+1}}.$$

Given a path $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$ and two entities $a \in A_1$ and $b \in A_{l+1}$. The HeteSim score between a and b measures the cosine of the probability distributions of a and b will meet at the middle type node when a follows along the path and b goes against the path. When the length l of path $\mathcal{P}$ is even, a and b will meet at the middle type node $A_{\frac{l}{2}+1}$. The path $\mathcal{P}$ can be divided into two equal-length parts as $\mathcal{P} = (\mathcal{P}_L\mathcal{P}_R)$, where $\mathcal{P}_L = A_1A_2\cdots A_{mid-1}A_{mid}$ and $\mathcal{P}_R = A_{mid}A_{mid+1}\cdots A_{l+1}$, $mid = \frac{l}{2}+1$. However, in the case of l is odd, a and b will never meet at the same objects. Path decomposition approach, which add a middle type object between $A_{\frac{l+1}{2}}$ and $A_{\frac{l+1}{2}+1}$, are adopted by Shi *et al.* [14]. This method will further increase complexity of calculation. To overcome this disadvantages, we consider the following two splitting cases: (1) $\mathcal{P}_L = A_1A_2\cdots A_{mid-1}A_{mid}$ and $\mathcal{P}_R = A_{mid}A_{mid+1}\cdots A_{l+1}$; (2) $\mathcal{P}_L = A_1A_2\cdots A_{mid}A_{mid+1}$ and $\mathcal{P}_R = A_{mid+1}A_{mid+2}\cdots A_{l+1}$; where $mid = \frac{l+1}{2}$. The final HeteSim value is the average of two HeteSim values, which is generate by the above two different cases. Finally, the HeteSim score between a and b based on the path $\mathcal{P}$ is calculated as follow:

$$HeteSim(a,b|\mathcal{P}) = \frac{\mathbf{R}_{\mathcal{P}_L}(a,:)(\mathbf{R}_{\mathcal{P}_R^{-1}}(b,:))^T}{\|\mathbf{R}_{\mathcal{P}_L}(a,:)\|_2 \times \|\mathbf{R}_{\mathcal{P}_R^{-1}}(b,:)\|_2},$$

where $\mathcal{P}_R^{-1}$ is the reverse path of $\mathcal{P}_R$.

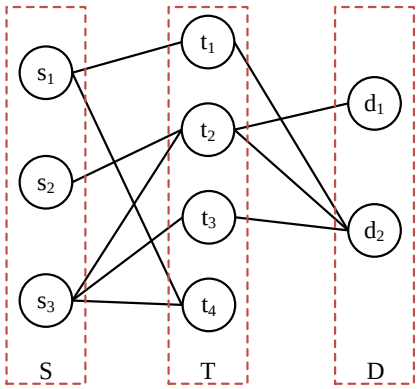


Fig. 4. Illustration for computing HeteSim.

Assume we are given a heterogeneous network as shown in Fig. 4 composing three object types. We simply show the procedure of the calculation of HeteSim scores between s_3 and d_1, d_2 under the path $\mathcal{P} = (STD)$. We can obtain the adjacency matrix $\mathbf{W}_{ST}$ and $\mathbf{W}_{DT}$ are:

$$\begin{array}{c} \mathbf{W}_{ST} \\ s_1 \\ s_2 \\ s_3 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{W}_{DT} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

After normalized the above two matrixes along the row vector, the transition probability matrix of $S \to T$ and $D \to T$ are

$$\mathbf{T}_{ST} = \begin{bmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0 & 0.3333 & 0.3333 & 0.3333 \end{bmatrix},$$

and

$$\mathbf{T}_{DT} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.3333 & 0.3333 & 0.3333 & 0 \end{bmatrix},$$

respectively. In this instance, we divide $\mathcal{P} = (STD)$ as two parts $\mathcal{P}_L = (ST)$ and $\mathcal{P}_R = (TD)$. The reachable probability matrix for path $\mathcal{P}_L$ and $\mathcal{P}_R^{-1}$ are equivalent their corresponding transition probability matrix, i.e., $\mathbf{R}_{\mathcal{P}_L} = \mathbf{T}_{ST}$

and $\mathbf{R}_{\mathcal{P}_R^{-1}} = \mathbf{T}_{DT}$. Finally, we can obtain two HeteSim scores

$$HeteSim(s_3, d_1|\mathcal{P}) = \frac{\mathbf{T}_{ST}(3,:)(\mathbf{T}_{DT}(1,:))^T}{\|\mathbf{T}_{ST}(3,:)\|_2 \times \|\mathbf{T}_{DT}(1,:)\|_2} = 0.5774,$$

and

$$HeteSim(s_3, d_2|\mathcal{P}) = \frac{\mathbf{T}_{ST}(3,:)(\mathbf{T}_{DT}(2,:))^T}{\|\mathbf{T}_{ST}(3,:)\|_2 \times \|\mathbf{T}_{DT}(2,:)\|_2} = 0.3849.$$

2.3.2 HeteSim_MultiPath (HSMP) method

In a heterogeneous network, there are different paths connect two objects. For example, a gene and a human phenotype can be connected via "gene - human phenotype" path, "gene - gene - human phenotype" path, and so on. Different paths have different semantic meanings, e.g., "gene - gene - human phenotype" path shows that if a gene is associated with a human phenotype, then other genes similar to the gene will be potential associated with the human phenotype; "gene - human phenotype - human phenotype" path shows that if a human phenotype associated with a gene, then other human phenotypes similar to the human phenotype will be potential associated with the gene. Therefore, it is significant to consider different paths in the procedure of similarity calculation. We then introduce a systematic approach to measure the similarity between objects in the biological heterogeneous network.

The HSMP method uses the HeteSim measure to calculate the similarity between objects in heterogeneous networks. The HeteSim scores of different paths are combined with a constant that dampens contributions from longer paths. Because of the HeteSim measure is based on the path-based relevance framework, it can capture effectively the subtle semantics of search paths. Consequently, we combined the HeteSim score of different paths with a constant β to dampen the contributions from longer paths. After a comprehensive searching, we found that the parameter $\beta = 1$ achieves best performance. Thus, $\beta = 1$ are selected for further association prediction in our experiments. Consequently, the HSMP score $S(a, b)$ measures the similarity between object is defined as

$$S(a,b) = \sum_{l=2}^{\infty} \left(\beta^{l-1} \times \sum_{\mathcal{P}_i \in \Psi_l} HeteSim(a, b|\mathcal{P}_i) \right),$$

where a and b are entities of object type A and B, respectively, Ψ_l is the set of path from object type A to B with path length l. Usually, a short path may contribute more than a long path. Therefore, only paths with lengths less than five are considered in our experiments. Let $\text{Sp} \in \{\text{Hs}, \text{At}, \text{Ce}, \text{Dm}, \text{Mm}, \text{Sc}, \text{Ec}, \text{Dr}, \text{Gg}\}$. All paths, which are used to measure the similarity between gene and phenotype, are list in Tab. 1. There are total 43 paths. Given a gene g and a human phenotype p (p isn't associated with g), the similarity score is

$$\begin{aligned} S(g,p) = & \beta * \big(HeteSim(g,p|\text{GeGeHs}) \\ & + HeteSim(g,p|\text{GeHsHs})\big) \\ & + \beta^2 * \big(HeteSim(g,p|\text{GeGeGeHs}) \\ & + HeteSim(g,p|\text{GeGeHsHs}) \\ & + HeteSim(g,p|\text{GeHsHsHs}) \\ & + HeteSim(g,p|\text{GeSpGeHS})\big) + \cdots \end{aligned}$$

2.3.3 HeteSim_SVM (HSSVM) method

The HSSVM method also uses the HeteSim measure to calculate the similarity between objects in heterogeneous networks. However, unlike HSMP, it uses a machine learning method to combine the HeteSim scores instead of a constant. Different paths make different contributions to the relevance score; therefore, the weight of the contribution of each path to the score is determined through a machine learning method.

A PU learning method was used to determine the different weights of different paths. The association between genes and human phenotypes in heterogeneous networks implies that the relation has been verified previously. However, if gene g and human phenotype p are not associated in the network, it cannot be assumed that g is not the disease gene of p. In other words, only partial positive associations are recorded in a network, not negative associations. Further, although a large numbers of gene-human phenotype pairs are unlabeled, most of them are negative associations.

In the HSSVM method, the associations between genes and human phenotypes were used as the positive set, and gene-phenotype pairs for which no associations existed were used as the unlabeled set. The HeteSim scores were used for each feature based on 66 constrained paths that were used to construct 66 features for each gene-phenotype pair. The 66 paths are listed in Table 2.

TABLE 1
Paths with length less than five.

Path scheme	Pathway	Number
GeGeHs	gene → gene → human phenotype	1
GeHsHs	gene → human phenotype → human phenotype	1
GeGeGeHs	gene → gene → gene → human phenotype	1
GeGeHsHs	gene → gene → human phenotype → human phenotype	1
GeHsHsHs	gene → human phenotype → human phenotype → human phenotype	1
GeSpGeHs	gene → all phenotypes → gene → human phenotype	9
GeGeGeGeHs	gene → gene → gene → gene → human phenotype	1
GeHsHsHsHs	gene → human phenotype → human phenotype → human phenotype → human phenotype	1
GeSpGeHsHs	gene → all phenotypes → gene → human phenotype → human phenotype	9
GeGeSpGeHs	gene → gene → all phenotypes → gene → human phenotype	9
GeSpSpGeHs	gene → all phenotypes → all phenotypes → gene → human phenotype	9

TABLE 2
Constrained paths and the corresponding features

Feature id	Paths	Numbers
$1-4$	GeGeHs, GeGeGeHs, GeGeGeGeHs, GeGeGeGeGeHs	4
$5-11$	GeHsHs, GeHsGeHsHs, GeHsGeGeHsHs, GeGeHsGeHsHs, GeHsGeGeGeHsHs, GeGeHsGeGeHsHs, GeGeGeHsGeHsHs	7
$12-65$	GeSpGeHs, GeSpGeGeHs, GeGeSpGeHs, GeSpGeGeGeHs, GeGeSpGeGeHs, GeGeGeSpGeHs	$6 \times 9 = 54$
66	No use, ones vector	1

Negative datas are not exist in the datasets. Therefore, a set of examples from an unlabeled set should be selected randomly and the random sample should be used as the negative data set. The negative and positive sets are then used to train a supervised classifier, in this paper, we trained a biased support vector machine (biased SVM) model. Because the true negatives are unknown, using the biased SVM, which penalizes the mistakes on positives heavier than negatives, is reasonable. Then we use the trained biased SVM model to predict all other data, and output a score for each gene-phenotype pair, where the score reflect confidence that the gene-phenotype is a positive pair.

To improve the stability and accuracy of the method, a bootstrap procedure [18], [19] was adopted. The selection of an unlabeled set and training of the supervised classifier are repeated 30 times in our study. The average score of each repetition is used as the final HSSVM score. Each gene-phenotype pair score reflects the confidence of whether a pair should be connected. All obtained scores are then ranked in descending order and the final ranking is used for prioritization.

3 Results and Discussion

3.1 Comparison of HSMP and HSSVM with other methods

To demonstrate the effectiveness of HSMP and HSSVM, we compared our methods with four other methods: PRINCE, ProDiGe, Katz, and CATAPULT. PRINCE is a state-of-the-art method that was designed to share information across diseases. It was first proposed by [9] and is often used as a reference for comparisons with other novel methods. ProDiGe is a SVM-based method that uses a large number of information sources to calculate the relevance score of each gene-phenotype pair. The Matlab code of this method was obtained from [13]. Katz is a recently proposed method inspired by social network analyses. This method uses the number of walks of different lengths between two objects as the similarity of these objects. CATAPULT, another recently proposed method, uses the PU learning method to learn the contribution of each path to the similarities of objects in heterogeneous networks and combines them with the learned contribution weight.

3.2 Effectiveness measure using cross-validation

In this study, we used the 3-fold cross-validation method used previously by [13] and [11]. First all known gene-human phenotype associations were split into three sets of the same size randomly. In the disease gene prioritization experiment, one set was set aside as the test set and the other two were used as known information. In each experiment, we hide the test set and use the known information as the training data. The experiment was repeated three times so that each set was hidden once and each hidden gene-phenotype pair obtained a predict relevance score. According to the predict relevance scores of hidden pair (g, p), we obtain the rank of the gene g in the list associated with phenotype p. The cumulative distribution function (CDF) of the rank scores of genes in hidden gene-phenotype pairs were used to evaluate the effectiveness of HSMP and HSSVM. The CDF represents the number of hidden pairs that were ranked in the top k. Because $12,331$ genes were used, the $rank(k)$ will always be between 1 and $12,331$.

Gene associations from HumanNet were used first to construct the gene-phenotype heterogeneous network. The results are shown in Fig. 5(a). Because only the top ranked genes are meaningful and can be used for further analysis, only the top 100 results are shown. The results show that the HSMP and HSSVM methods performed better than other methods, except CATAPULT. For HSMP, 13.38% hidden genes were ranked in the top 100, while for Katz, only 12.09% were in the top 100. The HSSVM results were similar to those of CATAPULT. For HSSVM, 15.20% hidden genes were ranked in the top 100, which was better than the performance of the methods that do not use machine learning. Both our methods performed much better than PRINCE and ProDiGe.

In experiment above, PRINCE method use HPRD gene network only and ProDiGe method can only use part of datasets. To further test the effectiveness of HSMP and HSSVM in predicting disease genes, an additional gene-phenotype heterogeneous network was constructed using HPRD. The prediction results, which are shown in Fig. 5(b), are similar to the results based on the gene-phenotype heterogeneous network (Fig. 5(a)). The results of the HSSVM method were again similar to the CATAPULT results and HSMP method better than the Katz results for genes below the top 60.

3.3 Performance of the methods on phenotype sets with a single known gene and phenotype sets with many known genes

The results of the comparisons between the performances of the different methods on phenotype sets with a single known gene and phenotype sets with many known genes are shown in Fig. 6. The performance of HSSVM was worse than that of CATAPULT for phenotypes with a single known gene; i.e., only 5.26% of the genes in the test set were in the

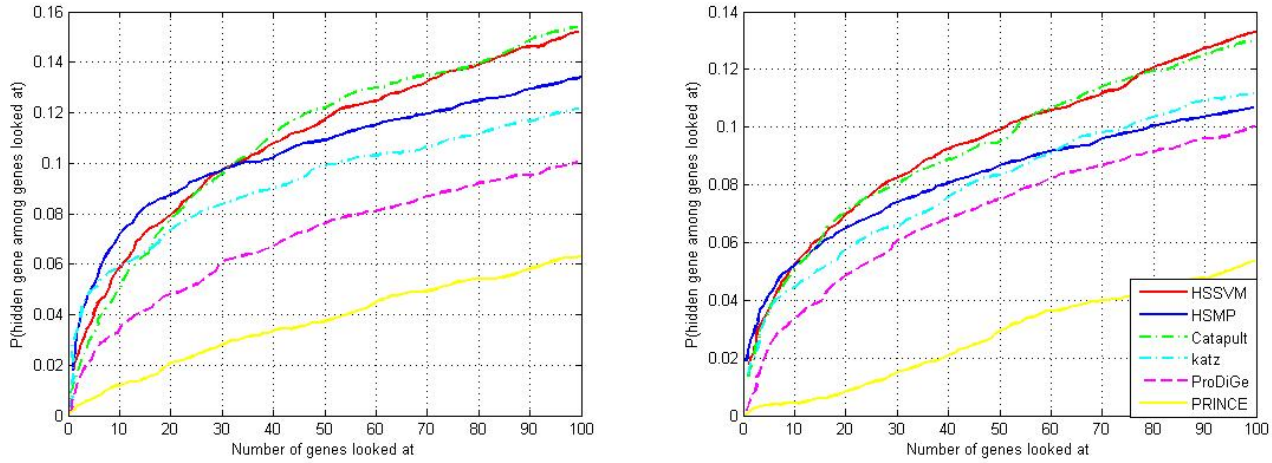
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520947, IEEE/ACM Transactions on Computational Biology and Bioinformatics

top 100, whereas 6.8% of the genes predicted by CATAPULT were in the top 100. The performance of HSMP was better than that of Katz, although only 0.52% more of the genes in the test data set were in the top 100 compared with the percentage of genes in the top 100 of Katz. As for the result on phenotype sets with many known genes (Fig. 6(b)), HSMP predicted 37.74% of the genes and HSSVM predicted 43.48% of the genes in the top 100, while Katz predicted only 34.14% and CATAPULT predicted 39.59%.

3.4 Analyses of the top 10 predicted genes

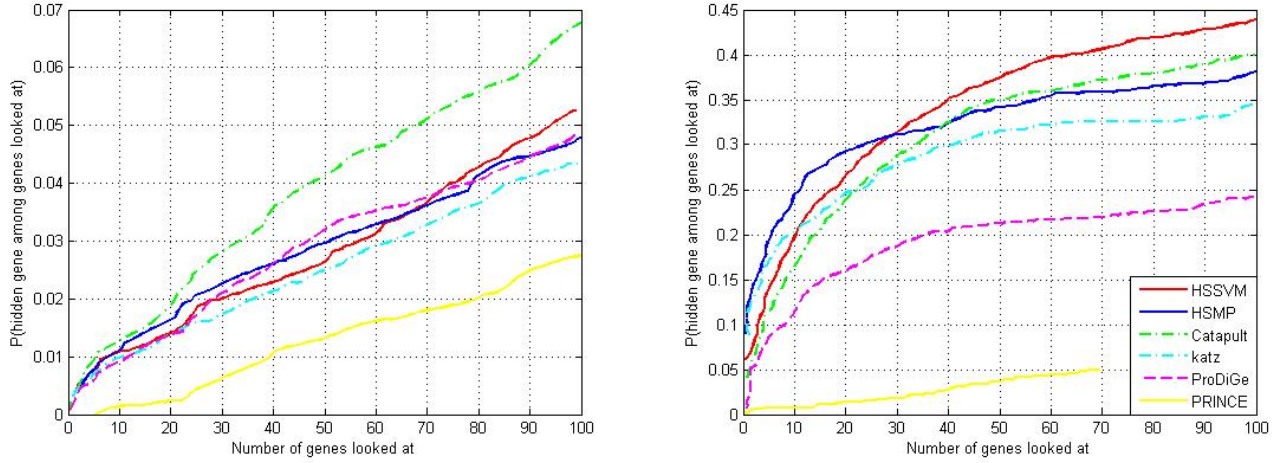
The top predictions have been reported to contain a very high degree of overlap [11]; therefore, we compared the top 10 predicted genes of the disease phenotypes. It must be said, though, this experiment use all known gene-phenotype association as the training data and use for predict unknown associations. The top 10 genes of 8 important diseases predicted by of HSMP and HSSVM are listed in Table 3 and 4. MYH11, for example, has been reported to be associated with leukemia through the inversion of a region on chromosome 16 and the formation of a CBFB-MYH11 chimera [20], and BRCA1 are known to be responsible for a large proportion of inherited predispositions to breast and ovarian cancer [21].

HSMP predicted that the DRD2 gene was associated with schizophrenia, which have been validated by [22], [23], [24]. Fan [25] also study the association of DRD2 gene polymorphisms with schizophrenia in a Chinese Han population. As for the Alzheimer's disease top ranked genes, there are many genes related to amyloid precursor protein, APP, in many ways, such as APLP2 and APLP1, which are homologs of APP, CTSB, also known as amyloid



(a) (b)

Fig. 5. Cumulative distribution function for the rank scores of genes in the test data sets under cross-validation. The ranks less than 100 are shown using HumanNet (Fig. 5(a)) and HPRD (Fig. 5(b)). The horizontal axis is the threshold k and the vertical axis is the ratio of true predictions in the top k.



(a) (b)

Fig. 6. Cumulative distribution function for the rank scores of genes in the test data sets under cross-validation: (a) phenotype sets with a single known gene; (b) phenotype sets with many known genes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520947, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 3
Top 10 genes predicted by HSMP

Leukemia	Alzheimer disease	Insulin resistance	Prostate cancer	Schizpophrenia	Breast cancer	Gastric cancer	Colorectal cancer
MIM: 601626	MIM: 104300	MIM: 125853	MIM: 176807	MIM: 181500	MIM: 114480	MIM: 137215	MIM: 114500
MYH11 (4629)	APLP2(334)	INS (3630)	RAD51 (5888)	DRD2 (1813)	BRCA1 (672)	NRAS (4893)	CTNNB1 (1499)
RUNX3 (864)	APLP1 (333)	AKT1 (207)	BRCA1 (672)	LINGO1 (84894)	RAD50 (10111)	EGFR (1956)	RAD51 (5888)
IMPDH2 (3615)	LRP1 (4035)	INSR (3643)	TP53 (7157)	RTN4 (57142)	MRE11A (4361)	HRAS (3265)	CCNA2 (890)
EP300 (2033)	APBB1 (322)	SLC2A2 (6514)	AKT2 (208)	CBS (875)	TOP3A (7156)	ERBB3 (2065)	MYBL2 (4065)
YBX3 (8531)	EPX (8288)	CREBBP (1387)	ATM (472)	SYN1 (6853)	DMC1 (11144)	COL4A5 (1287)	GSK3B (2932)
DKC1 (1736)	CTSB (1508)	IGF1R (3480)	MYC (4609)	AK2 (204)	BLM (641)	FGFR1 (2260)	TSG101 (7251)
KRAS (3845)	PXDN (7837)	EP300 (2033)	SIN3B (23309)	DDC (1644)	ATR (545)	BRAF (673)	CDK1 (983)
MLLT6 (4302)	APBA3 (9546)	MAFA (389692)	SIN3A (25942)	LMX1B (4010)	RAD52 (5893)	FGFR3 (2261)	AURKB (9212)
MYC (4609)	ABCB11 (8647)	GSK3B (2932)	TSG101 (7251)	HTR1B (3351)	SPO11 (23626)	RAF1 (5894)	KRAS (3845)
CBFA2T3 (863)	CALR (811)	AQP3 (360)	MXD1 (4084)	MTR (4548)	MUS81 (80198)	IRS1 (3667)	MAD2L1 (4085)

precursor protein secretase and LRP1, which is necessary for clearance of APP plaques. Associations between APP and Alzheimer can be found in [26]. A well-known fact, while HSSVM predicted that the ALOX5AP gene was associated with Alzheimer disease, which is similar to results reported previously by [27]. Many of the genes predicted by HSMP and HSSVM to be associated with various cancers (e.g TP53, RUNX3, KRAS, RAD50, RAD51, and RAD52) were identified among the top 10 predicted genes.

From the analysis of top ten prediction of the 8 diseases predicted by PRODIGE and CATAPULT, we found that these methods based on machine learning show a very high degree of overlap, for example, ProDiGe ranks EXT1 in the top ten for six out of the eight diseases studied, and CATAPULT ranks TP53 in the top ten for five of the diseases [11]. Also, similarity methods based on path count like always view gene with higher associations as the disease gene, just as the example showed in Fig. 1, for example, the number of associated phenotypes by TP53 (CATAPULT predict) is 9.

For further evaluation, we propose Average Overlap Ratio of the Predicted Top 10 genes, AOR-T10 and Average Number of Associated Phenotypes of the Predicted Top 10 genes, ANAP-T10. The AOR-T10 were measured by calculating the number of genes predicted more than twice to be in the top 10, and the ANAP-T10 was calculated based on the number of the associated phenotypes of all the predicted top 10 genes. In each experiment, eight diseases were chosen randomly and then the methods were used to predict the top 10 genes associated with these diseases. These two experiments were repeated 100 times and we calculate the average of each result as the final AOR-T10 and ANAP-T10. The AOR-T10 result is shown in Table 5, and the ANAP-T10 result is shown in Table 6.

TABLE 5
AOR-T10 result of different methods

Katz	HSMP	CATAPULT	HSSVM
1.58%	1.14%	8.41%	1.86%

TABLE 6
ANAP-T10 result of different methods

Katz	HSMP	CATAPULT	HSSVM
0.8421	0.9639	3.3458	1.6330

Table 5 shows that, although the SVM method is used in HSSVM, the overlap ratio was similar to the overlap found by Katz, perhaps because HeteSim accurately evaluates the relevance between nodes, thereby ensuring that HSSVM do not always find nodes with higher associations. This result is also verified by the average degree results in Table 6.

Overall the accuracy of HSSVM was found to be similar to that of CATAPULT, while the AOR-T10 and ANAP-T10 result of HSSVM was much better than that of CATAPULT (Figs. 5 and 6, and Tabs. 5 and 6).

TABLE 4
Top 10 genes predicted by HSSVM

Leukemia	Alzheimer disease	Insulin resistance	Prostate cancer	Schizpophrenia	Breast cancer	Gastric cancer	Colorectal cancer
MIM: 601626	MIM: 104300	MIM: 125853	MIM: 176807	MIM: 181500	MIM: 114480	MIM: 137215	MIM: 114500
EGR2 (1959)	UROD (7389)	SHOX (6473)	PHOX2A (401)	MTHFD1 (4522)	LCA5 (167691)	EGFR (1956)	RAD51 (5888)
KRAS (3845)	PPOX (5498)	ADAR (103)	RAD51 (5888)	EFNB1 (1947)	PDGFRL (5157)	CTNNB1 (1499)	BARD1 (580)
DKC1 (1736)	BMP2 (650)	INS (3630)	BARD1 (580)	LMX1B (4010)	AURKA (6790)	MAP3K8 (1326)	HMMR (3161)
MYH11 (4629)	AGTR1 (185)	TERC (7012)	TSG101 (7251)	MTR (4548)	ODC1 (4953)	PPP2R1B (5519)	TSG101 (7251)
PDGFRA (5156)	AGT (183)	TRPV4 (59341)	PHB (5245)	FKBP5 (2289)	BRCA1 (672)	ARHGEF6 (9459)	PHB (5245)
EPOR (2057)	EPX (8288)	STAR (6770)	PPM1D (8493)	CBS (875)	PTPRJ (5795)	CASP8 (841)	PPM1D (8493)
RUNX3 (864)	ALOX5AP (241)	DHCR24 (1718)	HMMR (3161)	RTN4 (57142)	KLF6 (1316)	SLC26A4 (5172)	CTNNB1 (9821)
KITLG (4254)	CYP3A5 (1577)	INSR (3643)	RB1CC1 (9821)	KLHDC8B (200942)	CTNNB1 (1499)	COL4A5 (1287)	RB1CC1 (9821)
THPO (7066)	PRKCH (5583)	EFEMP1 (2202)	CYP2D6 (1565)	HABP2 (3206)	PLA2G2A (5320)	RAD51 (5888)	CDH1 (999)
NSD (64324)	PTGIS (5740)	ERBB2 (2064)	IL1B (3553)	MATR3 (9782)	EP300 (2033)	BRAF (673)	CASP8 (841)

4 Conclusion

Here, we used the HeteSim measure to calculate the relevance of different or same nodes types in a heterogeneous network. Two novel methods, HSMP and HSSVM that use the HeteSim measure, were developed. In HSMP, the HeteSim measure was used to calculate the similarity between nodes, after which the HeteSim scores of different paths were combined with a constant that dampens the contributions of longer paths. In HSSVM, instead of using a constant, the HeteSim measure and a machine learning method were combined to calculate the similarity between nodes. We evaluated various methods with cross-validation and found that HSMP and HSSVM were better than the state-of-the-art methods, PRINCE and ProDiGe. Our non-machine learning method HSMP performs better than the existing non-machine learning methods and our machine learning method HSSVM obtains similar accuracy with the best existing machine learning method CATAPULT. When compare with the AOR-T10 and ANAP-T10, we found HSSVM method obtain lower overlap ratios and lower associations than other machine learning methods. These results indicated that our two methods were reasonable and credible.

HSMP and HSSVM can be extended easily to other species when the necessary data become available. Recently, genome-wide association studies (GWAS) have been used to detect allelic variations that may affect susceptibility to complex diseases. We consider that HSMP and HSSVM have greet potential to identify disease-related single nucleotide polymorphisms (SNPs) from GWAS data. Further, the identification of microRNAs associated with diseases is known to very important for understanding the pathogenesis of diseases at the molecular level. The methods described here may provide the basis for designing specialized tools for disease prevention, diagnosis, and treatment. It may therefore be meaningful to extend HSMP and HSSVM to predict microRNA-disease associations.

The data sets and the Matlab code for the two methods described here are freely available at http://datamining.xmu.edu.cn/~xzeng/dgassociations/klk/WebRoot/index.jsp.

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, pp. 228–237, 2003.

[2] K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup *et al.*, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[3] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular systems biology*, vol. 4, no. 1, 2008.

[4] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, "Online Mendelian inheritance in man (OMIM)," *Human mutation*, vol. 15, no. 1, pp. 57–61, 2000.

[5] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 30, no. 1, pp. 52–55, 2002.

[6] 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, no. suppl 1, pp. D514–D517, 2005.

[7] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *European journal of human genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[8] S. Köhler, 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, pp. 949–958, 2008.

[9] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS computational biology*, vol. 6, no. 1, p. e1000641, 2010.

[10] Y. Li and J. C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.

[11] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of gene-disease associations using methods inspired by social network analyses," *PloS one*, vol. 8, no. 5, p. e58977, 2013.

[12] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[13] F. Mordelet and J.-P. Vert, "ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples," *BMC bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[14] C. Shi, X. Kong, Y. Huang, S. Y. Philip, and B. Wu, "HeteSim: A general framework for relevance measure in heterogeneous networks," *IEEE Transactions on Knowledge & Data Engineering*, no. 10, pp. 2479–2492, 2014.

[15] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, vol. 21, no. 7, pp. 1109–1121, 2011.

[16] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.

[17] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," Ph.D. dissertation, Publisher not identified, 2009.

[18] F. Mordelet and J.-P. Vert, "A bagging SVM to learn from positive and unlabeled examples," *arXiv preprint arXiv:1010.0772*, 2010.

[19] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *ICML*, vol. 2. Citeseer, 2002, pp. 387–394.

[20] M. M. Le Beau, R. A. Larson, M. A. Bitter, J. W. Vardiman, H. M. Golomb, and J. D. Rowley, "Association of an inversion of chromosome 16 with abnormal marrow eosinophils in acute myelomonocytic leukemia: a unique cytogenetic–clinicopathological association," *New England Journal of Medicine*, vol. 309, no. 11, pp. 630–636, 1983.

[21] D. Ford, D. F. Easton, D. T. Bishop, S. A. Narod, and D. E. Goldgar, "Risks of cancer in BRCA1-mutation carriers," *The Lancet*, vol. 343, no. 8899, pp. 692–695, 1994.

[22] C. Dubertret, L. Gouya, N. Hanoun, J.-C. Deybach, J. Adès, M. Hamon, and P. Gorwood, "The 3 region of the DRD2 gene is involved in genetic susceptibility to schizophrenia," *Schizophrenia research*, vol. 67, no. 1, pp. 75–85, 2004.

[23] Z. Zahari, L. K. Teh, R. Ismail, and S. M. Razali, "Influence of DRD2 polymorphisms on the clinical outcomes of patients with schizophrenia," *Psychiatric genetics*, vol. 21, no. 4, pp. 183–189, 2011.

[24] R. Kukreti, S. Tripathi, P. Bhatnagar, S. Gupta, C. Chauhan, S. Kubendran, Y. J. Reddy, S. Jain, and S. K. Brahmachari, "Association of DRD2 gene variant with schizophrenia," *Neuroscience letters*, vol. 392, no. 1, pp. 68–71, 2006.

[25] H. Fan, F. Zhang, Y. Xu, X. Huang, G. Sun, Y. Song, H. Long, and P. Liu, "An association study of DRD2 gene polymorphisms with schizophrenia in a Chinese Han population," *Neuroscience letters*, vol. 477, no. 2, pp. 53–56, 2010.

[26] H.-S. Hoe and G. William Rebeck, "Functional interactions of APP with the apoE receptor family," *Journal of neurochemistry*, vol. 106, no. 6, pp. 2263–2271, 2008.

[27] H. Manev and R. Manev, "5-Lipoxygenase (ALOX5) and FLAP (ALOX5AP) gene polymorphisms as factors in vascular pathology and Alzheimers disease," *Medical hypotheses*, vol. 66, no. 3, pp. 501–503, 2006.

Xiangxiang Zeng received the B.S. degree in automation from Hunan University, China, in 2005, the Ph.D. degree in system engineering from Huazhong University of Science and Technology, China, in 2011. From 2010 to 2011 he spent one year working in the group of natural computing in Seville University, Spain. Currently, he is an assistant professor in the Department of Computer Science, Xiamen University. His main research interests include natural computing, DNA computing, neural computing and automaton theory.

Yuanlu Liao is a Master student of the Department of Computer Science at Xiamen University. He received her BE in Computer Science from Jiangxi Normal University, Nanchang, China. His research interests include bioinformatics and data ming.

Yuansheng Liu is a PhD student of the School of Software at Dalian University of Technology. He received his BE and Msc degrees in Computer Science from Xiangtan University of China in 2012 and 2015. His research interest is bioinformatics.

Quan Zou is an Associate Professor of Computer Science at Xiamen University. He received his PH.D. from Harbin Institute of Technology, P.R.China in 2009. His research is in the areas of bioinformatics, machine learning and parallel computing. Now he is putting the focus on genome assembly, annotation and functional analysis from the next generation sequencing data with parallel computing methods. Several related works have been published by Briefings in Bioinformatics, Bioinformatics, PLOS ONE and IEEE/ACM Transactions on Computational Biology and Bioinformatics. He serves on many impacted journals and NSFC(National Natural Science Foundation of China)