SCIENCE CHINA
Information Sciences

• Supplementary File •

Logistic regression algorithm to identify candidate disease genes based on reliable protein-protein interaction network

Xiujuan Lei[1*] & Wenxiang Zhang[1]

[1]*School of Computer Science, Shaanxi Normal University, Xian* 710119*, China*

Appendix A LR-RPN algorithm

In this section, LR-RPN algorithm is described in detail. Firstly, we introduce some preliminaries in section A.1. Next, we present the process of constructing a reliable PPI network and how to find disease genes by logistic regression algorithm in section A.2 and A.3, respectively.

Appendix A.1 Preliminaries

Let $G = \{g_1, g_2, \cdots, g_N\}$ represent the set of all human genes, and N is defined as the totally number of human genes. Given another set $P = \{g_1, g_2, \cdots, g_m\}$ ($P \subseteq G$) represents unknown genes (candidate disease genes) which means that we do not know whether they are related to the certain disease. Next, $U = G/P = \{g_{m+1}, g_{m+2}, \cdots, g_N\}$ represents disease genes which are related to diseases. In other words, the candidate disease genes in P are unlabeled.

In order to facilitate the understanding, we define $D = \{D_1, D_2, \cdots, D_M\}$ to represent the set of human diseases where D_i is a data set composed of genes that are known associated with the ith disease. For a specific disease D_k, let $X^k = \{x_1^k, x_2^k, \cdots, x_N^k\}$ be label informations (labels, *i.e.*, the value is set either 1 or 0) defined on training data, where $x_i^k = 1$ represents g_i is associated with the kth disease (D_k), and $x_i^k = 0$ otherwise.

Identifying candidate disease genes can convert to a positive-unlabelled (PU) learning task under these notations [1, 2]. On the basis of the logistic regression, we can get the conditional probability $P(x_i^k = 1|\phi)$ for each unknown gene. Here ϕ represents prior information, such as protein complexes, tissue expressions and the internal connection between reliable PPI network and so on. We will introduce it in detail in later sections.

Appendix A.2 Constructing a reliable PPI network

Appendix A.2.1 *Connecting PPI network and keywords*

In order to consider the topological similarity of the PPI network and the properties of the protein individual characteristic, we add keywords into PPI network to construct a new heterogeneous network. A similar constructed method has been used in [3].

Formally, the heterogeneous network can be represented as an undirected graph $H = \{V, E\}$, where V and E represents the set of nodes and edges, respectively. Here, V is divided into two types: protein and protein, protein and keyword. In other words, an edge only connects two proteins or a protein and a keyword. The detailed definition of H is as follows:

$$H = \begin{bmatrix} H_P & H_{KP} \\ H_{PK} & 0 \end{bmatrix}_{(N+K)\times(N+K)} \tag{1}$$

$$H_P(i,j) = \begin{cases} 1, & \text{if there is an edge linking Proteins } i \text{ and } j \text{ in PPI} \\ 0, & \text{if there is no edge linking Proteins } i \text{ and } j \text{ in PPI} \end{cases} \tag{2}$$

$$H_{PK}(i,j) = \begin{cases} 1, & \text{if there is an edge linking Protein } i \text{ and Keyword } j \\ 0, & \text{if there is no edge linking Protein } i \text{ and Keyword } j \end{cases} \tag{3}$$

* Corresponding author (email: xjlei@snnu.edu.cn)

$$H_{KP}(i,j) = \begin{cases} 1, & \text{if there is an edge linking Keyword } i \text{ and Protein } j \\ 0, & \text{if there is no edge linking Keyword } i \text{ and Protein } j \end{cases} \tag{4}$$

We add keywords into PPI network, which has two advantages. Firstly, the association between protein and keyword can consolidate useful interaction in the PPI network [3]. Secondly, undirected proteins, which maybe have a certain biological relationship, can be connected through keywords.

Appendix A.2.2 *Random walking on the heterogeneous network*

The RWR [4] algorithm mimics a random walker that starts on seed nodes and then randomly moves to a neighbor or returns to seed nodes with a probability γ. Li *et al.* [5] applied improved RWR algorithm on the heterogeneous network (RWRH). Here we adopt it to get a probability matrix.

Firstly, we construct the transition matrix M based on heterogeneous network H, which can be defined as follows:

$$M = \begin{bmatrix} \lambda M_P & (1-\lambda)M_{KP} \\ (1-\lambda)M_{PK} & 0 \end{bmatrix}_{(N+K)\times(N+K)} \tag{5}$$

where λ is used to control the probability of jumping between proteins and keywords. M_P is intra-subnetwork transition matrixes. M_{KP} and M_{PK} are inter-subnetwork matrixes. The detailed definitions of M_P,M_{KP} and M_{PK} are as follows:

$$M_P(i,j) = \begin{cases} \frac{H_P(i,j)}{\sum_k H_P(i,k)} & \text{,if } \sum_k H_P(i,k) > 0 \\ 0 & \text{,otherwise} \end{cases} \tag{6}$$

$$M_{PK}(i,j) = \begin{cases} \frac{H_{PK}(i,j)}{\sum_k H_{PK}(i,k)} & \text{,if } \sum_k H_{PK}(i,k) > 0 \\ 0 & \text{,otherwise} \end{cases} \tag{7}$$

$$M_{KP}(i,j) = \begin{cases} \frac{H_{KP}(i,j)}{\sum_k H_{KP}(i,k)} & \text{,if } \sum_k H_{KP}(i,k) > 0 \\ 0 & \text{,otherwise} \end{cases} \tag{8}$$

The initial probability matrix for our heterogeneous network is defined as follows:

$$P(0) = \begin{bmatrix} (1-\eta) * P_{PPI}(0) & 0 \\ 0 & \eta * P_{Keywords}(0) \end{bmatrix}_{(N+K)\times(N+K)} \tag{9}$$

where the initial probability matrix $P(0)$ for our heterogeneous network is a diagonal unit matrix of $|N+K|$ rows and $|N+K|$ columns. N and K is defined as the number of genes in PPI network and the number of keywords terms, respectively. Besides, $P_{PPI}(0)$ and $P_{Keywords}(0)$ are unit diagonal matrixes, which are N-dimensional and K-dimensional matrixes, respectively. We can see from equation (9) that if seed node belongs to PPI, we give it initial probability $1-\eta$, otherwise η. The parameter $\eta \in (0,1)$ is used to determine the initial probability for the node of PPI and Keywords.

Based on $P(0), P(t)$ and transition matrix, the probability vector at step $t+1$ can be described as follows:

$$P(t+1) = (1-\gamma)M^T P(t) + \gamma P(0) \tag{10}$$

where the parameter $\gamma \in (0,1)$ is the restart probability, and $P(t)$ represents a matrix in which the *ith* iteration holds the probability of finding the random walker. At each iteration, the random walker has a probability of γ to get back the seed nodes.

After some steps, the $P(t+1)$ will reach a stationary distribution when the difference between $P(t+1)$ and $P(t)$ is less than 10^{-6}.

Appendix A.2.3 *Reconstructing a reliable PPI network*

In section 2.2.2, we could get a $|N+K| \times |N+K|$ probability matrix. Because our ultimate goal is to reconstruct the PPI network, we only extract the first N rows and the first N columns of the final probability matrix, and then normalize by rows. The final matrix is defined as Z.

Because the difference between the probability vectors of different nodes is very small in the final matrix Z, we employ the PPI network reconstruction method in [6] to magnify the difference between them in next step. First, we construct a median vector H from the matrix Z, where $H(j)$ is the median of the *ith* row of Z. Next, we calculate the $|N| \times |N|$ offset matrix Θ, where $\Theta_{ij} = Z_{ij} - H_i$. Finally, the topological similarity matrix of Z is calculated by the Pearson correlation coefficient. Empirically, we define a topological similarity matrix C about Θ, where $C_{ij} = pcc(\Theta_{1\sim|N|,i}, \Theta_{1\sim|N|,j})$.

In the end, a reliable PPI network is constructed from the topological similarity matrix C by connecting pairs of nodes, whose similarities between nodes in C are more than a certain threshold.

Appendix A.3 Prioritization of candidate disease genes

Appendix A.3.1 *Prior label estimation*

It is well known that the labeling informations of samples are needed in any machine learning algorithm. Obviously, we can assign 1 or 0 to the known disease genes according to known gene datasets. However, other genes need to be given a prior probability to determine if it is assigned 1 or 0.

Being inspired by [7], we take protein complexes and tissue expression terms as prior information. If a gene g_i encodes a protein in a certain kind of tissue or protein complex, its prior probability, which is used to measure the possibility of correlation between g_i and D_k, is calculated as follows:

$$p_i = max\left(p_i^c, p_i^t\right) \tag{11}$$

$$p_i^c = \frac{A^c}{B^c} \tag{12}$$

$$p_i^t = \frac{A^t}{B^t} \tag{13}$$

where p_i^c and p_i^t are prior probabilities of g_i, which can be calculated by protein complexes and tissue expression terms respectively. In formula (5), A^c is the number of D_k-related genes in a certain protein complex that contains g_i, and B^c is the number of all genes in corresponding protein complex. Similarly, A^t and B^t also have the similar meaning to A^c and B^c respecting to tissue expression terms. If g_i belongs to various protein complexes and tissue expression terms, the maximum values of p_i^c and p_i^t are chosen. In addition, there is a special case that if g_i does not encode protein in any protein complex and tissue expression term, let $p_i = M/F$ be its prior probability, in which M is the number of D_k-related genes and F is the number of genes in G. Next, the prior label for g_i can be reckoned as follows:

$$t = rand(0, 1) \tag{14}$$

$$x_i^k = \begin{cases} 1, & t \leqslant p_i \\ 0, & t > p_i \end{cases} \tag{15}$$

where $rand(a, b)$ is a function, which can generate a random number that obeys the standard uniform distribution between a and b.

Appendix A.3.2 *Prioritizing candidate disease genes based on logistic regression*

Logistic regression algorithm is a classical classification algorithm representing the conditional probability distribution. According to most of assumptions [7–9], the conditional posterior probability distribution of a specific gene g_i, associated with a specific disease D_k, can be formulated as follows:

$$p(x_i^k = 1|s_i^k, w) = \frac{exp(w^T s_i^k)}{exp(w^T s_i^k) + 1} \tag{16}$$

$$p(x_i^k = 0|s_i^k, w) = \frac{1}{exp(w^T s_i^k) + 1} \tag{17}$$

where s_i^k is input and represents multiple heterogeneous biological feature vector. If x_i^k is equal to 1, $p(x_i^k|s_i^k, w)$ represents a posterior probability that g_i is associated with D_k, x_i is equal to 0 otherwise. w is the weight vector. In order to define the multiple heterogeneous biological feature vector, we take g_i as an example, and its feature value vector s_i^k can be defined as follows:

$$s_i^k = \{1, s_{i1}^k, s_{i2}^k, s_{i3}^k, s_{i4}^k, s_{i5}^k, s_{i6}^k, s_{i7}^k, s_{i8}^k\} \tag{18}$$

where s_{i1}^k is the number of genes associated with D_k in direct neighbors of g_i. s_{i2}^k is equal to $NN - s_{i1}^k$, where NN is the number of direct neighbors of g_i. s_{i3}^k is the number of genes associated with D_k in a certain protein complex which contains g_i. s_{i4}^k is equal to $NC - s_{i3}^k$, where NC is the number of genes in corresponding protein complex. s_{i5}^k and NT have similar definition as s_{i3}^k and NC, here we just simply replace the protein complex with the tissue in their definitions. s_{i6}^k is equal to $NT - s_{i5}^k$. s_{i7}^k is the similarity score between g_i and D_k using Wang's measure in *GoSemSim*, which is an R package for semantic similarity computation among gene clusters, sets of GO terms, gene products and GO terms [10,11]. Firstly, the similarity of g_i and D_k is defined as follows:

$$s_{i7}^k = \max_{1 \leqslant j \leqslant o} \left(\frac{MF_{geneSim(g_i, dg_j^k)} + CC_{geneSim(g_i, dg_i^k)} + BP_{geneSim(g_i, dg_j^k)}}{3} \right) \tag{19}$$

where dg_j^k belongs to the set of D_k, in other words, dg_j^k is a gene associated with D_k. $MF_{geneSim(g_i, dg_j^k)}$, $CC_{geneSim(g_i, dg_j^k)}$ and $BP_{geneSim(g_i, dg_j^k)}$ is the semantic similarity values between g_i and dg_j^k based on MF, CC and BP by *geneSim* function in *GoSemSim*, respectively. s_{i8}^k is equal to $1 - s_{i7}^k$.

Hence, the feature value of all genes about a specific disease D_k can be grouped together in a matrix form as follows:

$$FF^k = \begin{bmatrix} 1 & s_{11}^k & s_{12}^k & s_{13}^k & s_{14}^k & s_{15}^k & s_{16}^k & s_{17}^k & s_{18}^k \\ 1 & s_{21}^k & s_{22}^k & s_{23}^k & s_{24}^k & s_{25}^k & s_{26}^k & s_{27}^k & s_{28}^k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_{m1}^k & s_{m2}^k & s_{m3}^k & s_{m4}^k & s_{m5}^k & s_{m6}^k & s_{m7}^k & s_{m8}^k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_{N1}^k & s_{N2}^k & s_{N3}^k & s_{N4}^k & s_{N5}^k & s_{N6}^k & s_{N7}^k & s_{N8}^k \end{bmatrix}_{N \times 9} \tag{20}$$

Besides, the corresponding weighted parameter is set as $w = (w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9)^T$. If weighted parameter is given and prior labels for all genes have been estimated, we can get all posterior probability that all unknown genes are related with D_k according to the above equation.

Appendix A.3.3 *Parameter estimation*

Some studies [7–9] have discussed how to estimate the weighted parameter w. The estimation of parameter w can be calculated based on the training set, where the known genes associated with D_k are marked as 1, and others are labeled based on the equation of (11)∽(15).

The likelihood can be written as:

$$\widehat{w} = arg\max_{w} \prod_{i=1}^{N} P(x_i^k | s_i^k, w) \tag{21}$$

marked as:

$$\widehat{w} = arg\max_{w} L(w) \tag{22}$$

where $L(w)$ is the log likelihood function. After taking into(16), (17) into (21) and some mathematical reasoning, we can obtain $L(w)$ as

$$L(w) = \sum_{N}^{i=1} \left[x_i w^T s_i^k - ln\left(1 + exp(w^T s_i^k)\right)\right] \tag{23}$$

The log likelihood (23) has been proved to be a concave function on the page 354 of [9]. In this study, it can be solved by using the standard matlab function $fminunc()$ which is used to calculate the minimum solution $-L(w)$. Here, the initial parameter of w is equal to zero.

Appendix A.4 Training 10 logistic regression model classifiers

Because the identification of disease genes is a class-imbalance classification problem, we use the under-sampling method. Besides, in order to avoid loss of some important information, we construct ten negative data sets, in which samples are randomly selected from all unknown genes. The number of samples in each negative data set is equal to twice the number of known genes on a special disease. Then we combine these 10 negative data sets and known genes data sets into 10 training data sets. In these training data sets, known genes are the same but unknown genes are different. Next, we train 10 logistic regression classifiers by these training sets. The final posterior probability, which indicates the degree of association between genes and disease, is equal to the average of the posterior probability from 10 logistic regression model classifiers.

However, the posterior probability does not always work well [7]. Therefore, we use the rank score. Firstly, we calculate the number of posterior probabilities of genes that is smaller than the posterior probability of current research gene, and then it divided by the number of genes is equal to the rank score. Obviously, the larger rank score represents that corresponding candidate genes have a high probability of being associated with a special disease.

Appendix B Experiment

Appendix B.1 Data Sources

In this study, multiple biological data sources are collected to test the performance of LR-RPN, which are presented as follows.

Firstly, the dataset of PPI network is obtained from the human protein reference database (HPRD) (Release 9) [12]. The HPRD database provides protein interaction data, where we can link two human genes if corresponding proteins interact together.

Secondly, the protein complexes are collected from the database of CORUM [13] and PCDq [14]. These human protein complexes contain at least one gene that can be mapped back to HPRD database.

Thirdly, the gene-disease of association data is obtained from Goh *et al.* [15] and OMIM database [16]. Goh *et al.* [15] have manually classified diseases in OMIM into 22 primary disease classes.

For the above datasets, we directly obtain them from Chen *et al.* [7]. They have been preprocessed in such a way that the PPI network contained 7311 human genes from HPRD database, the protein complexes have 2870 types containing 3881 human genes, and the gene-disease of association data contains 12 disease classes which are connected with 815 human genes.

In order to improve the accuracy of the identification of disease genes based on various biological data, we also integrate other data as follows.

The tissue expression data is collected from the HPRD (Release 9) [12]. There are 307 types of expression terms in this database with at least two proteins, which involve 1110 proteins in those expression terms. The original tissue expression file can be obtained in $AdditionalFile \setminus Tissue\_Expressions$. The keywords data is collected from the UniProt database [17]. We select 632 types of term in UniProt database with at least two proteins. The original keywords file can be obtained in $AdditionalFile \setminus Keywords$.

To validate the effectiveness of LR-RPN algorithm, we select 5 multifactorial diseases that are associated with multiple genomic regions from OMIM database [16] in Dec. 2017, which belong to cancer class according to Goh *et al.* [15]. According to the MIM records, all these 5 diseases are associated with great than or equal to 10 known valid causing genes locating different genomic regions according to the OMIM file of *morbidmap.xlsx*, and they also are used in [18, 19]. Detailed information is shown in Table B1.

Table B1 The detail information of five kinds of cancer disease

Phenotype name	Phenotype ID	No. of associated genes
Breast cancer	114480	23
Lung cancer	211980	18
Prostate Cancer	176807	18
Leukemia	601626	22
Colon cancer	114500	26

Appendix B.2 Evaluation criteria

In this study, the performance of LR-RPN is evaluated by leave-one-out cross validation (LOOCV) based on the rank score. LOOCV is a special case of k-fold cross-validation where k is equal to 1. In each round of experiments, only one gene, being associated with d_k, is deleted in training set and then add to the test set. The remaining known disease-related genes are used as a training set to assess whether the test gene is related to a certain disease. According to the leave-one-out cross-validation result, we plot the receiver operating characteristic (ROC) curve and compute the area under the curve (AUC) values based on rank score.

When we plot the ROC value, the positive control genes are those known associated with disease d_k. However, we need the extra operation for negative control genes. A similar approach has already been adopted in [7]. The negative control genes are randomly selected from those known gene samples, which are not associated with d_k. If the number of positive control genes is equal to s, we randomly extract s samples from unknown genes as negative control genes.

Based on rank score, the *precision* and *recall* are also employed to evaluate the performance of our method. *Precision* reflects the proportion of true positives in the top-k samples for a disease class. *Recall* is equals to true positives in the top-k samples divided by the total number of true positives in the test for a disease class. The values of *precision* and *recall* are different for different top-k samples. They can be defined as follows:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where TP is the number of predicted disease genes matched with known disease genes. FP is the number of predicted disease genes, which are not matched with known disease genes. FN is the number of known disease genes that are not located in predicted disease genes.

Appendix B.3 Experimental results

In this section, we first evaluate the effects of different forms of feature value in original PPI network. Next, the effects of λ and η are analyzed for identifying disease genes result. Then, we compare the performance of prioritizing disease genes between the origin PPI network and the reliable PPI network, which aims is to prove the reliability of reconstructed PPI network. Besides, we compare our algorithm with some other methods, which are: (1) the DIR algorithm [20], (2) the RWR algorithm [4], (3) the MRF algorithm [21] and (4) logistic regression (F3PC) [7]. Finally, we conduct case studies to verify the effectiveness of LR-RPN in the identificiation of disease genes.

Appendix B.3.1 *Effects of different forms of feature values*

To analyze the effects of different heterogeneous data on prioritizing disease genes, we define three types of feature value vectors for g_i as follows.

$$Feature1 = \{1, s_{i1}^k, s_{i2}^k\}$$

$$Feature2 = \{1, s_{i1}^k, s_{i2}^k, s_{i3}^k, s_{i4}^k\}$$

$$Feature3 = \{1, s_{i1}^k, s_{i2}^k, s_{i3}^k, s_{i4}^k, s_{i5}^k, s_{i6}^k\}$$

$$Feature4 = \{1, s_{i1}^k, s_{i2}^k, s_{i3}^k, s_{i4}^k, s_{i5}^k, s_{i6}^k, s_{i7}^k, s_{i8}^k\}$$

$$Feature5 = \{1, s_{i1}^k, s_{i3}^k, s_{i5}^k, s_{i7}^k\}$$

$$Feature6 = \{1, \frac{s_{i1}^k}{s_{i1}^k + s_{i2}^k}, \frac{s_{i3}^k}{s_{i3}^k + s_{i4}^k}, \frac{s_{i5}^k}{s_{i5}^k + s_{i6}^k}, s_{i7}^k\}$$

- $Feature1$: According to the definition of s_i^k in section 2.3.2, we can see that $Feature1$ only considers one factor, which is the direct neighbor of g_i in origin PPI network (coming from HPRD database). This method is similar with Chen *et al.* [22] methods.
- $Feature2$: Similarly, we can see that $Feature2$ considers two heterogeneous factors, which are the direct neighbors of g_i in origin PPI network and protein complex.
- $Feature3$: We enhance heterogeneous factor comparing with $Feature1$ and $Feature2$. The vector can capture the PPI, protein complex and tissue information to prioritize disease genes.
- $Feature4$: We further enhance heterogeneous factor for $Feature4$. It contains PPI network, protein complex, tissue expression and the semantic similarity of genes based on Go terms. It can capture more heterogeneous information than other feature vectors.

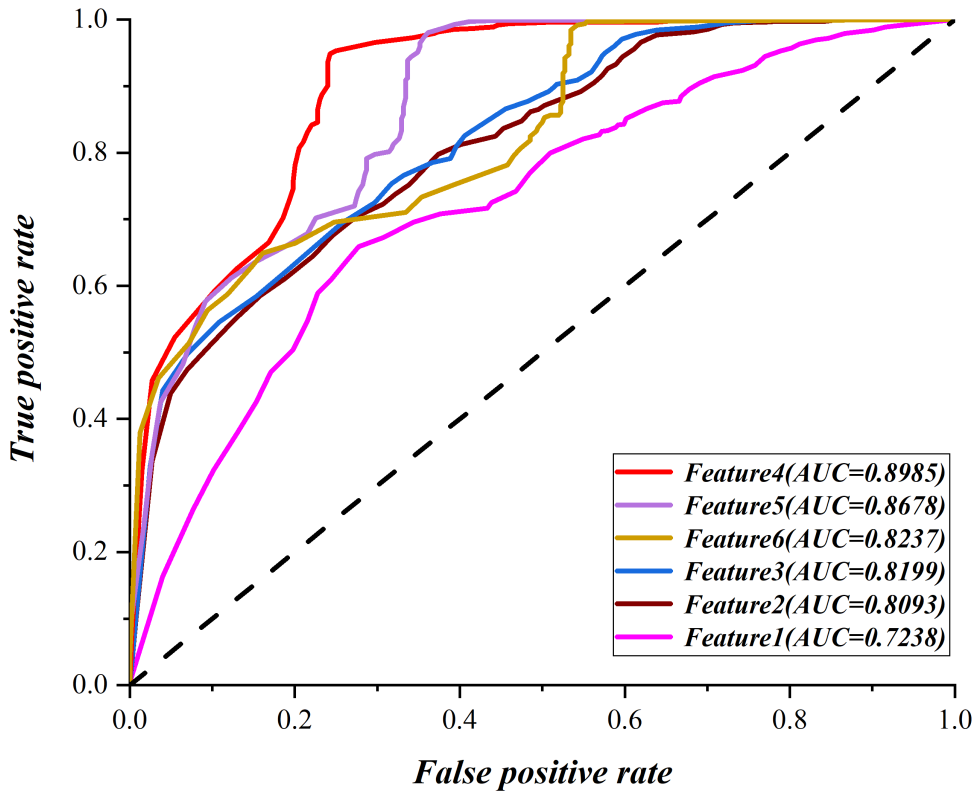


Figure B1 Performance of different feature vectors to logistic regression for prioritizing disease genes

- $Feature5$: This feature only considers disease factors for different biological data.
- $Feature6$: This feature is not only considering disease factors but also the non-disease factor for different biological data. However, it combines the disease and non-disease factors into a comprehensive factor in each of the biological data.

In subsection 2.3.2, the attribute name of every feature has been explained. Figure B1 shows the AUC performance of the six feature vectors to logistic regression for prioritizing disease genes in original PPI network. Obviously, we can see that the values of AUC increase as the heterogeneous feature vectors increase. It proves that integrating multiple heterogeneous data can enhance the performance of prioritizing disease genes. The detailed result data can be obtained in $AdditionFile \setminus Different\_Feature\_Result$.

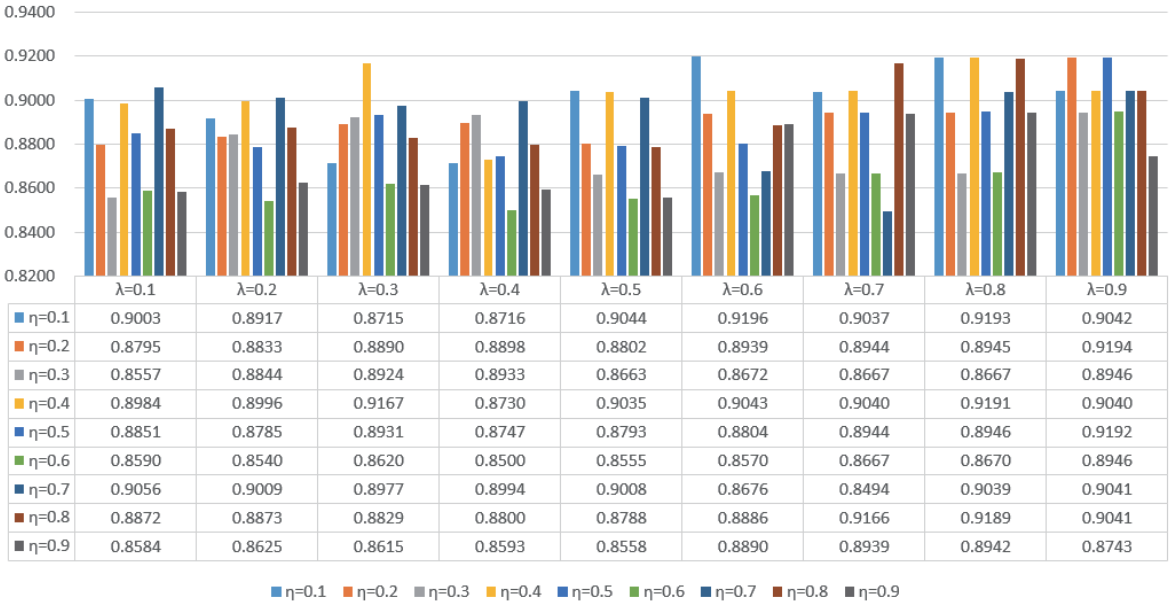
Appendix B.3.2 *Effects of parameters*

In the process of constructing reliable PPI network, there are three parameters which are γ, λ and η. As we all known, the parameter of γ is the restart probability. It has been demonstrated that γ only has a slight effect on the experimental result [14]. Here, we assign γ to the value of 0.7, based on the previous studies [4]. From Figure B1, we can see that $Feature4$ has the best performance for logistic regression, so we select $Feature4$ as the feature vector for next evaluation.

To analyse the effects of two parameters, we set various values for them ranging from 0.1 to 0.9. Then we run LR-RPN in disease class and experiment result is shown in Figure B2. The AUC value ranges from 0.8494 to 0.9196. These results indicate that LR-RPN has better stability and accuracy in prioritizing candidate disease genes for disease class. We plot the distribution of AUC values showed in Figure B3. Besides, we can see the overall effect of two parameters from the Figure B3(b). In general, according to Figure B2 and Figure B3, we can observe the influence of parameters on the final result from different perspectives, we can find that it is optimal when $\lambda = 0.6$ and $\eta = 0.1$. The data sources, which contain reconstructing reliable PPI network and corresponding result, can be obtained in the $AdditionalFile \setminus RePPI\_Keywords$ and $AdditionalFile \setminus RePPI\_Result\_Keywords$.

Appendix B.3.3 *Effects of different networks*

To further highlight the importance of the reliable PPI network, we compare the logistic regression algorithm in original PPI network (LR-OPN), reconstructed PPI network by random walk with restart on PPI network (LR-REPN), and reliable



	λ=0.1	λ=0.2	λ=0.3	λ=0.4	λ=0.5	λ=0.6	λ=0.7	λ=0.8	λ=0.9
η=0.1	0.9003	0.8917	0.8715	0.8716	0.9044	0.9196	0.9037	0.9193	0.9042
η=0.2	0.8795	0.8833	0.8890	0.8898	0.8802	0.8939	0.8944	0.8945	0.9194
η=0.3	0.8557	0.8844	0.8924	0.8933	0.8663	0.8672	0.8667	0.8667	0.8946
η=0.4	0.8984	0.8996	0.9167	0.8730	0.9035	0.9043	0.9040	0.9191	0.9040
η=0.5	0.8851	0.8785	0.8931	0.8747	0.8793	0.8804	0.8944	0.8946	0.9192
η=0.6	0.8590	0.8540	0.8620	0.8500	0.8555	0.8570	0.8667	0.8670	0.8946
η=0.7	0.9056	0.9009	0.8977	0.8994	0.9008	0.8676	0.8494	0.9039	0.9041
η=0.8	0.8872	0.8873	0.8829	0.8800	0.8788	0.8886	0.9166	0.9189	0.9041
η=0.9	0.8584	0.8625	0.8615	0.8593	0.8558	0.8890	0.8939	0.8942	0.8743

Figure B2 Effect of value and based on whole genome

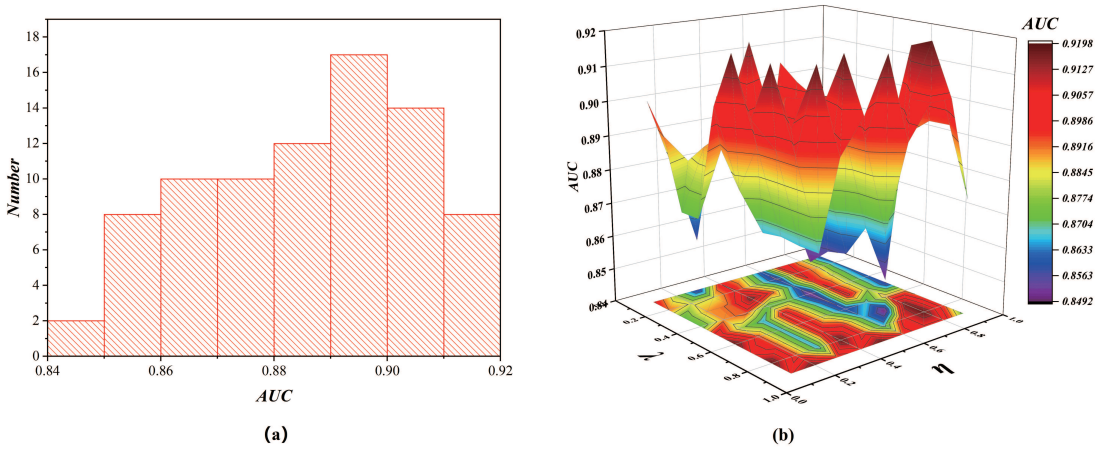


Figure B3 The distribution map of AUC values

PPI network by walking on heterogeneous network (LR-RPN). Here λ and η is set as 0.1 and 0.7 for LR-RPN, respectively. Besides, the situations of LR-REPN and LR-RPN are the same, when the parameters λ and η of LR-RPN is taken as 1 and 0, respectively.

Obviously, we can see clearly from Figure B4 that the AUC value of LR-RPN is better than the AUC value of LR-OPN and LR-REPN. Besides, the AUC value of LR-REPN is better than the AUC value of LR-OPN. It proves that building a reliable PPI network by PPI and keywords is helpful for the identification of disease genes.

Appendix B.4 Compare with Previous Algorithms

To investigate the efficiency of LR-RPN (γ=0.7, λ=0.6 and η=0.1), four previous algorithms were introduced for comparison, which include: (1) the DIR algorithm [20], (2) the RWR algorithm [4], (3) the MRF algorithm [21] and (4) logistic regression (F3PC) [7]. As shown in Table B2, we can see LR-RPN is much better than other algorithms in terms of AUC. The AUC value of LR-RPN is 0.912, which is 8.2%, 19.1%, 19.6%, 20.1% than the F3PC algorithm, the MRF algorithm, the DIR algorithm and the RWR algorithm, respectively. The detailed ROC curve is displayed in Figure B5.

Next, we also evaluate LR-RPN and other methods by using *precision* and *recall*. A detailed description of *precision* and *recall* is located in section 3.2. For each algorithm, the *precision* and *recall* of top-k positions are calculated, which can help us to understand the local characteristics of these algorithms. The range of k is 5 to 815. From the results shown in Figure B6, we can see that *precision* and *recall* of LR-RPN are far superior to F3PC, RWR, DIR and MRF in Figure B6(a)

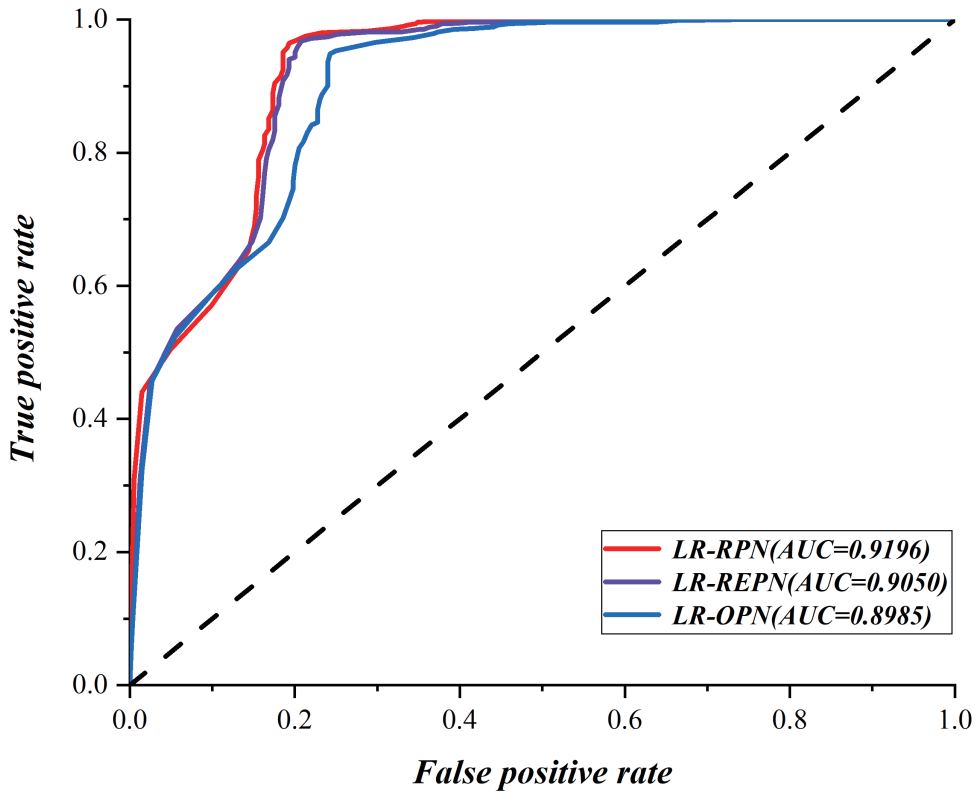


Figure B4 **ROC curves of cross-validation results of LR-RPN, LR-REPN, and LR-OPN**

Table B2 The performance of different algorithms

Method	LR-RPN	F3PC	MRF	DIR	RWR
AUC	0.9196	0.8300	0.7210	0.7160	0.7110

and Figure B6(b), respectively.

Finally, we compare the performance of each algorithm on 12 disease classes as shown in Figure B7. Obviously, it can be seen from Figure B7 that LR-RPN has the excellent performance for AUC value than other algorithms in each of 12 disease classes, except immunological class. The highest AUC proves that integrating multiple data is beneficial for prioritizing disease genes.

Appendix B.4.1 *Assessment by predicting new disease genes*

To further validate the effectiveness of LR-RPN for prioritizing new disease genes, we perform case studies here for 5 multifactorial cancer diseases. When the parameter γ, λ and η is set as 0.7, 0.1 and 0.5 respectively, the performance of LR-RPN is the best. So we only focus on the prediction of LR-RPN (γ=0.7, λ=0.6 and η=0.1). The genes are verified by literature and their PMID is given. Then, we only selecte Breast cancer (MIM: 114480) as the case study for LR-RPN (γ=0.7, λ=0.6 and η=0.1) and other cancer disease for case study can be download in $AdditionalFile \setminus CaseStudy$.

The breast cancer is etiologically and genetically heterogeneous, and histopathologically. Important genetic factors have been indicated by familial occurrence and bilateral involvement. As shown in Table 3, the first prediction **MID2** is analyzed in relation to breast cancer [23] (PMID: 26791755) in which the author indicated that MID2 may be a novel interventional target and prognostic marker in breast cancer. The prediction **HDAC1** is also analyzed in relation to breast cancer [24] (PMID: 28779562), and author indicated that the migration and proliferation of breast cancer cells via activation of Snail/IL-8 signals can be triggered by **HDAC1**. Gao Yun *et al.* [25] (PMID: 30390344) found the target genes of miR-93 were closely related to **RBBP7** by Gene Ontology enrichment analysis, and the serum levels of miR-93 were upregulated in breast cancer in the qPCR validation test. Yan Y *et al.* [26] (PMID: 26616021) demonstrated that the growth of breast cancer cells is reduced by **PIAS4** depletion. Cheng X W *et al.* [27] (PMID: 19858209) suggested the cell proliferation in

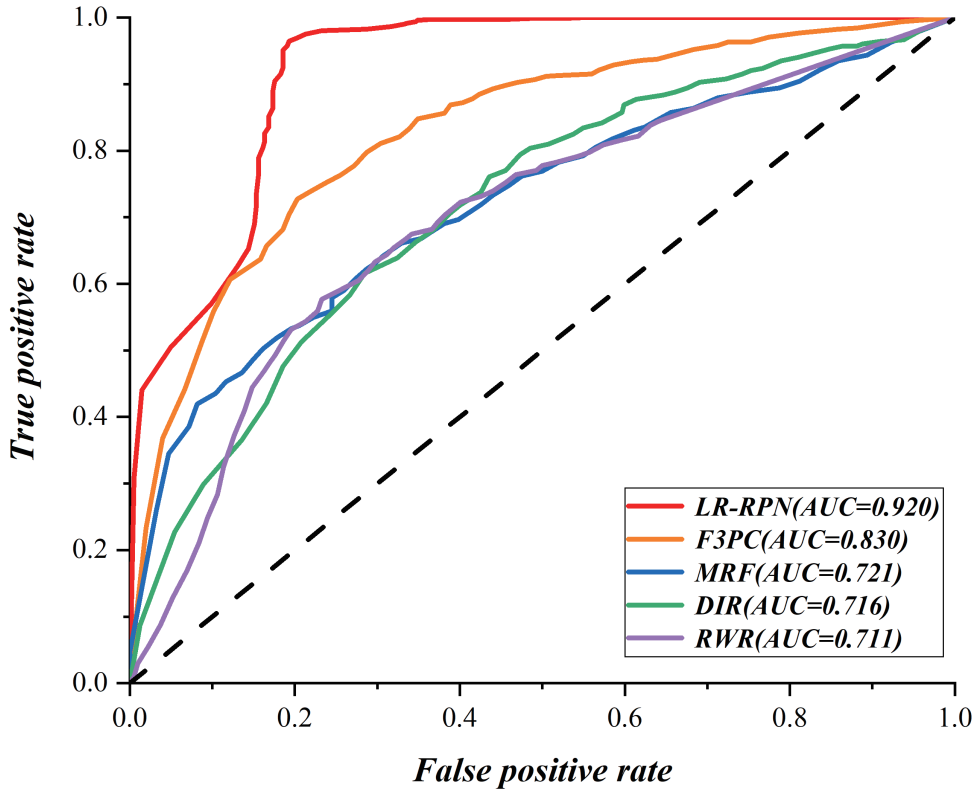


Figure B5 ROC curves of cross-validation results of different methods

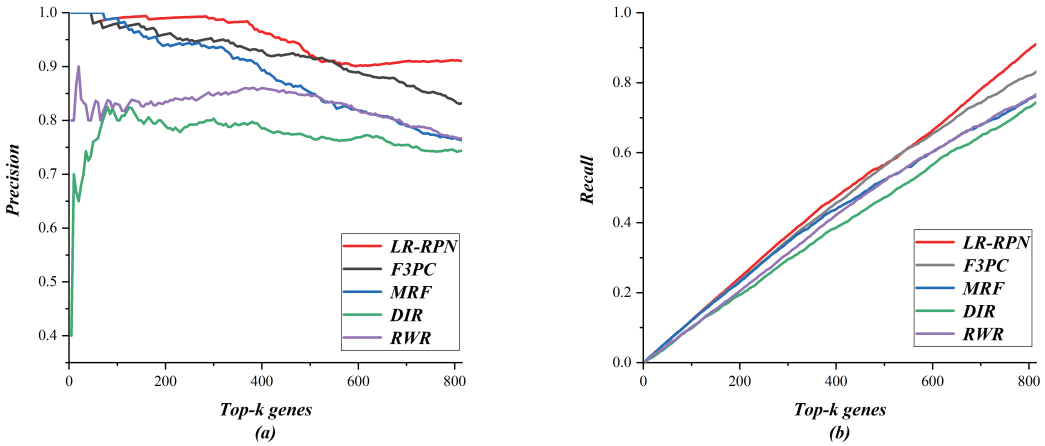
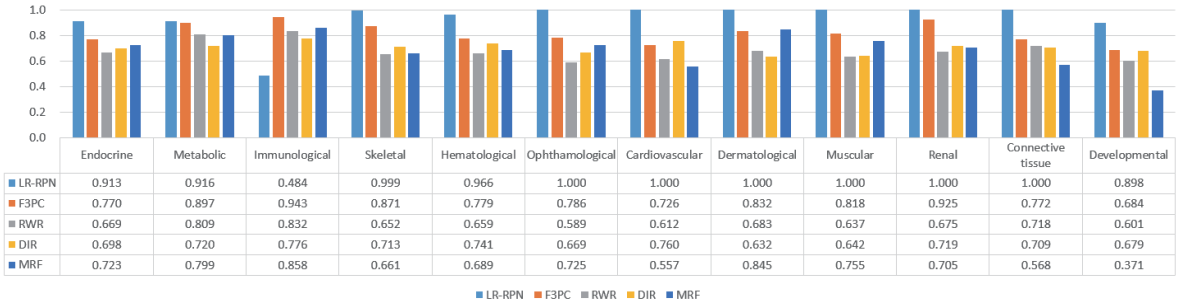


Figure B6 Average *precision* and *recall* on disease classes of test set at each top-k position. (a) average *precision* on all disease classes, (b) average *recall* on all disease classes.

MCF-7 breast cancer cells is promoted by depletion of GPS2 or SMRT by siRNA.

Appendix C Additional File

Additional file data association with LR-RPN can be downloaded, in github online, at https://github.com/zwx94/LR-RPN



	Endocrine	Metabolic	Immunological	Skeletal	Hematological	Ophthamological	Cardiovascular	Dermatological	Muscular	Renal	Connective tissue	Developmental
LR-RPN	0.913	0.916	0.484	0.999	0.966	1.000	1.000	1.000	1.000	1.000	1.000	0.898
F3PC	0.770	0.897	0.943	0.871	0.779	0.786	0.726	0.832	0.818	0.925	0.772	0.684
RWR	0.669	0.809	0.832	0.652	0.659	0.589	0.612	0.683	0.637	0.675	0.718	0.601
DIR	0.698	0.720	0.776	0.713	0.741	0.669	0.760	0.632	0.642	0.719	0.709	0.679
MRF	0.723	0.799	0.858	0.661	0.689	0.725	0.557	0.845	0.755	0.705	0.568	0.371

Figure B7 The AUC performance of different algorithm for 12 disease class

Table B3 Top-10 predicted causal genes of breast cancer

Gene Symbol	PMID	Evidence URL
MID2	26791755	https://www.ncbi.nlm.nih.gov/pubmed/?term=MID2+breast+cancer
MED26	unconfirmed	unconfirmed
COPS6	unconfirmed	unconfirmed
HDAC1	28779562	https://www.ncbi.nlm.nih.gov/pubmed/28779562
RBBP7	30390344	https://www.ncbi.nlm.nih.gov/pubmed/30390344
BANP	unconfirmed	unconfirmed
PIAS4	26616021	https://www.ncbi.nlm.nih.gov/pubmed/26616021
WASF2	unconfirmed	unconfirmed
CBFA2T2	unconfirmed	unconfirmed
GPS2	19858209	https://www.ncbi.nlm.nih.gov/pubmed/19858209

References

1 Mordelet Fantine, Vert Jean-Philippe. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. Bmc Bioinformatics, 2011, 12:389.
2 Chapelle O, Schlkopf B, Zien A. Semi-Supervised Learning. Journal of the Royal Statistical Society, 2013, 172(2):530-530.
3 Ata Sezin Kircali, Fang Yuan, Wu Min, et al. Disease Gene Classification with Metagraph Representations. Methods, 2017, 131: 83-92.
4 Köhler Sebastian, Bauer Sebastian, Horn Denise, et al. Walking the interactome for prioritization of candidate disease genes. American Journal of Human Genetics, 2008, 82(4):949-958.
5 Li Yongjin, Patra Jagdish C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 2010, 26(9):1219-1224.
6 Rakyan Vardhman K, Beyan Huriya, Down Thomas A, et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. Plos Genetics, 2011, 7(9).
7 Chen Bolin, Li Min, Wang JianXin, et al. A fast and high performance multiple data integration algorithm for identifying human disease genes. Bmc Medical Genomics, 2015, 8(Suppl 3): S2.
8 Shi Jinhong, Chen Bolin, Wu Fang-Xiang. Unifying protein inference and peptide identification with feedback to update consistency between peptides. Proteomics, 2013, 13(2):239-247.
9 Boyd SP, Vandenberghe L. Convex optimization. Cambridge University Press, New York, 2004
10 Wang James Z, Du Zhidian, Payattakool Rapeeporn, et al. A new method to measure the semantic similarity of GO terms. Bioinformatics, 2007, 23(10):1274-1281.
11 Yu Guangchuang, Li Fei, Qin Yide, et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products[J]. Bioinformatics, 2010, 26(7):976-978.
12 Keshava Prasad TS, Goel Renu, Kandasamy Kumaran, et al. Human Protein Reference Database. Nucleic Acids Res, 2009, 37(Database): D767-D772
13 Ruepp Andreas, Waegele Brigitte, Lechner Martin, et al. CORUM: the comprehensive resource of mammalian protein complexes - 2009. Nucleic Acids Res, 38(Database): D497-D501
14 Kikugawa Shingo, Nishikata Kensaku, Murakami Katsuhiko, et al. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset. BMC SYSTEMS BIOLOGY, 2012, 6(Suppl 2): S7.
15 Goh Kwang-Il, Cusick, Michael E, Valle David, et al. The human disease network. PROCEEDINGS OF THE

NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 2007, 104(21): 8685-8690
16 McKusick Victor A. Mendelian Inheritance in Man and its online version, OMIM. AMERICAN JOURNAL OF HUMAN GENETICS, 2007, 80(4): 588-604.
17 Bateman Alex, Martin Maria Jesus, O'Donovan Claire, et al. UniProt: a hub for protein information. Nucleic Acids Research, 2015, 43(D1): D1204-D212.
18 Zhu Jie, Qin Yufang, Liu Taigang, et al. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. BMC Bioinformatics. 2013, 14(Suppl 5):S5.
19 Tian Zhen, Guo Maozu, Wang Chunyu, et al. Constructing an integrated gene similarity network for the identification of disease genes. Journal of Biomedical Semantics, 2017, 8(Suppl 1): UNSP 32.
20 Chen Yixuan, Wang Wenhui, Zhou Yingyao, et al. In Silico Gene Prioritization by Integrating Multiple Data Sources. Plos One, 2011, 6(6): e21137.
21 Chen BoLin, Li Min, Wang JianXin, et al. Disease gene identification by using graph kernels and Markov random fields. Science China Life Sciences, 2014, 57(11):1054-1063.
22 Chen Bolin, Li Min, Wang Jianxin, et al. A logistic regression based algorithm for identifying human disease genes[C]. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, Belfast UK, 2015:197-200.
23 Wang L, Wu J H, Yuan J, et al. Midline2 is overexpressed and a prognostic indicator in human breast cancer and promotes breast cancer cell proliferation in vitro and in vivo. Frontiers of Medicine, 2016, 10(1): 41-51.
24 Tang Z H, Ding S J, Huang H L, et al. HDAC1 triggers the proliferation and migration of breast cancer cells via upregulation of interleukin-8. Biological Chemistry, 2017, 398(12): 1347-1356.
25 Gao Yun, Deng Kaifeng, Liu Xuexiang, et al. Molecular mechanism and role of microRNA-93 in human cancers: A study based on bioinformatics analysis, meta-analysis, and quantitative polymerase chain reaction validation. Journal of cellular biochemistry. 2018.
26 Yan Y, Ollila S, Wong I P L, et al. SUMOylation of AMPK alpha 1 by PIAS4 specifically regulates mTORC1 signalling. Nature Communications, 2015, 6.
27 Cheng X W, Kao H Y. G Protein Pathway Suppressor 2 (GPS2) Is a Transcriptional Corepressor Important for Estrogen Receptor alpha-mediated Transcriptional Regulation. Journal of Biological Chemistry, 2009, 284(52): 36395-36404.