Hindawi
Complexity
Volume 2017, Article ID 9194801, 8 pages
https://doi.org/10.1155/2017/9194801

Research Article

Identifying the Risky SNP of Osteoporosis with ID3-PEP Decision Tree Algorithm

Jincai Yang,[1] Huichao Gu,[1] Xingpeng Jiang,[1] Qingyang Huang,[2] Xiaohua Hu,[1] and Xianjun Shen[1]

[1]*School of Computer Science, Central China Normal University, Wuhan 430079, China*
[2]*School of Life Science, Central China Normal University, Wuhan 430079, China*

Correspondence should be addressed to Jincai Yang; jcyang@mail.ccnu.edu.cn

Received 31 March 2017; Revised 26 May 2017; Accepted 8 June 2017; Published 7 August 2017

Academic Editor: Fang-Xiang Wu

Copyright © 2017 Jincai Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the past 20 years, much progress has been made on the genetic analysis of osteoporosis. A number of genes and SNPs associated with osteoporosis have been found through GWAS method. In this paper, we intend to identify the suspected risky SNPs of osteoporosis with computational methods based on the known osteoporosis GWAS-associated SNPs. The process includes two steps. Firstly, we decided whether the genes associated with the suspected risky SNPs are associated with osteoporosis by using random walk algorithm on the PPI network of osteoporosis GWAS-associated genes and the genes associated with the suspected risky SNPs. In order to solve the overfitting problem in ID3 decision tree algorithm, we then classified the SNPs with positive results based on their features of position and function through a simplified classification decision tree which was constructed by ID3 decision tree algorithm with PEP (Pessimistic-Error Pruning). We verified the accuracy of the identification framework with the data set of GWAS-associated SNPs, and the result shows that this method is feasible. It provides a more convenient way to identify the suspected risky SNPs associated with osteoporosis.

1. Introduction

Osteoporosis is a type of systemic skeletal disease that is characterized by reduced bone mass and microarchitecture deterioration of bone tissues, thereby leading to the loss of strength and increased risk of fractures [1]. It is one of the age-related diseases with arteriosclerosis, hypertension, diabetes, and cancer. Currently, none of the medical methods is safe and effective to cure osteoporosis. Therefore, it is necessary to provide theoretical basis for developing a medical strategy to cure the disease from the pathogenesis of osteoporosis.

With the completion of the International HapMap Project and 1000 Genomes Project, about ten millions SNPs of human were annotated, among which more than 3 million are common SNPs. Genetic analysis has reached the stage of genome-wide association study (GWAS). The GWAS is applied to the study of 40 kinds of diseases that are related to more than 500 thousands SNPs [2].

Osteoporosis is a complex and polygenic disease of bone system with the heritability of bone mass is about 60–80% [3]. Much progress has been made on the genetic analysis of osteoporosis in the past 20 years and it has been found that a lot of genes and SNPs are associated with osteoporosis through GWAS [4, 5].

Computational biology refers to the development and application of data analysis, the theory of data method, mathematical modeling, and computer simulation technology, used in the study of biology, behavioral, and social group system of a discipline [6]. The rapid mass of biological data accumulation is unprecedented in the history of human science. Now, a variety of methods and tools of computational biology through the Internet have been successfully applied in every aspect in the field of biological research. They are powerful for post-GWAS studies [7] and could identify the potential and promising causal SNPs that require experimental tests for follow-up functional studies. Extensive work has

been done in this area in recent years. The performances were well validated through identifying numerous disease-associated SNPs for further study and revealing previously unknown mechanisms for complex diseases [8].

The method of computational biology can also be used to study and understand these osteoporosis-susceptible genes and the function of SNP. All the osteoporosis associated genes and SNPs (including linkage disequilibrium (LD) SNPs) sequence information were collected and aggregated from the national center for biological information (NCBI) database, and the effects of osteoporosis GWAS-associated lead SNPS and their linked SNPs to transcription factor (TF) binding affinity were studied through JASPAR database. At the same time, the osteoporosis GWAS-associated genes have also been analyzed with Protein-Protein Interaction (PPI) network analysis tool in the study of the osteoporosis GWAS-associated SNPs associated by the online PPI tool named String. Combining with GO and pathway analysis, we found that the hub proteins associated and the Wnt signaling pathway were related to the mesenchymal stem cell differentiation and hormone signaling that was related to the metabolism of osteoporosis [9]. Finally, it was found that the osteoporosis GWAS-associated SNPs in special region of genes had long-range interaction signal with other locus by analyzing the long-range interaction of osteoporosis associated SNPs on GWAS3D [10].

In the BIBM workshop paper [11], we utilized the known osteoporosis GWAS-associated SNPs and genes as the data set to identify the osteoporosis suspected risky SNPs. The process for identification was achieved by computation method. In this extension, we made some improvements on the paper. Firstly, we had achieved graphical description for the SNPs identification process. We added a flow chart for the paper to describe the process of identification method that made the method more intuitive. Secondly, we used ID3 decision tree algorithm with PEP method instead of ID3 decision tree algorithm in the second part of the method. We made the improvement to solve the overfitting problem in ID3 decision tree algorithm; we used the C4.5 algorithm to make a comparison with our ID3-PEP algorithm. Finally, we added type 2 diabetes (T2D) GWAS-associated SNPs and genes as the negative data set based on osteoporosis GWAS-associated SNPs and genes to verify the accuracy of the method comprehensively.

2. Material and Method

We identified the suspected risky SNPs associated with osteoporosis by algorithm based on the analysis of osteoporosis GWAS-associated SNPs with the method mentioned above [9]. It is assumed that the SNPs that are similar to the osteoporosis GWAS-associated SNPs are possible risky SNPs associated with osteoporosis. The identification process of the suspected risky SNPs includes two steps in general. Firstly, we constructed a Protein-Protein Interaction (PPI) network based on the Protein-Protein Interaction analysis of the osteoporosis GWAS-associated genes and the genes associated with suspected risky SNPs and identify whether the genes associated with the suspected risky SNPs are associated with osteoporosis through random walk algorithm based on Markov chain. By the algorithm, we also selected the suspected risky SNPs whose associated genes are identified to be associated with osteoporosis. We then classified those SNPs based on their characteristics of function and their loci features by a classification decision tree, and the decision tree was constructed by ID3 decision tree algorithm with Pessimistic-Error Pruning. Figure 1 describes the process to identify the osteoporosis risky SNPs.

2.1. The Identification of Genes Associated with Suspected Risky SNPs. According to the modular property of the genetic diseases, many scholars have proposed prioritization algorithms to predict the disease-causing genes based on the PPI, Human Disease Network, and DISEASOME recently [12–16]. Similarly, we obtained the scores of the genes associated with the suspected risky SNPs through the random walk algorithm based on the PPI of the osteoporosis GWAS-associated genes and the genes associated with suspected risky SNPs. Then, the result was acquired by setting up a threshold k, and the genes associated with suspected risky SNPs are probably the osteoporosis associated genes if their scores are greater than k.

2.2. The Random Walk Algorithm Based on Markov Chain. Kohler proposed a method for the problem of candidate-gene prioritization by random walk algorithm based on the global network distance of PPI. The results indicate that the algorithm is more effective than the local network distance algorithm [17]. The random walk algorithm was applied to Protein-Protein Interaction network of all associated genes.

An undirected graph $G = (V, E)$ is defined for the Protein-Protein Interaction network of all associated genes. In the undirected graph G, V is the set of vertices for the interactors of the network. And V is defined as $V = \{v_1, v_2, \ldots, v_n\}$; E is the set of edges; and E is defined as $E = \{\langle v_i, v_j\rangle \mid v_i,\ v_j \in V\}$. Every edge in the set of edges corresponds to two nodes of the set of vertices for the interaction between the interactors. Moreover, it is assumed that a random process meets the condition of Markov chain. The random process should be as follows:

(a) The probability distribution of time $t+1$ is only related to the state of time t, and it is not related to the state before time t.

(b) The state transition is not related to the value of t from the time t to time $t + 1$. Therefore, the Markov chain model is defined as

$$(S, P, Q)\,. \tag{1}$$

S is a nonempty set that consists of all the possible states of the system. It is a state space that can be a limited and denumerable set or a nonempty set. $P = [P_{ij}]_{n\times n}$ is the state transfer-probability matrix, P_{ij} is the probability that the system is in the state i at time t to the state j at time $t + 1$. N is the number of system states. $Q = \{q_0, q_1, \ldots, q_{n-1}\}$ is the initial probability distribution of the system, q_i is the

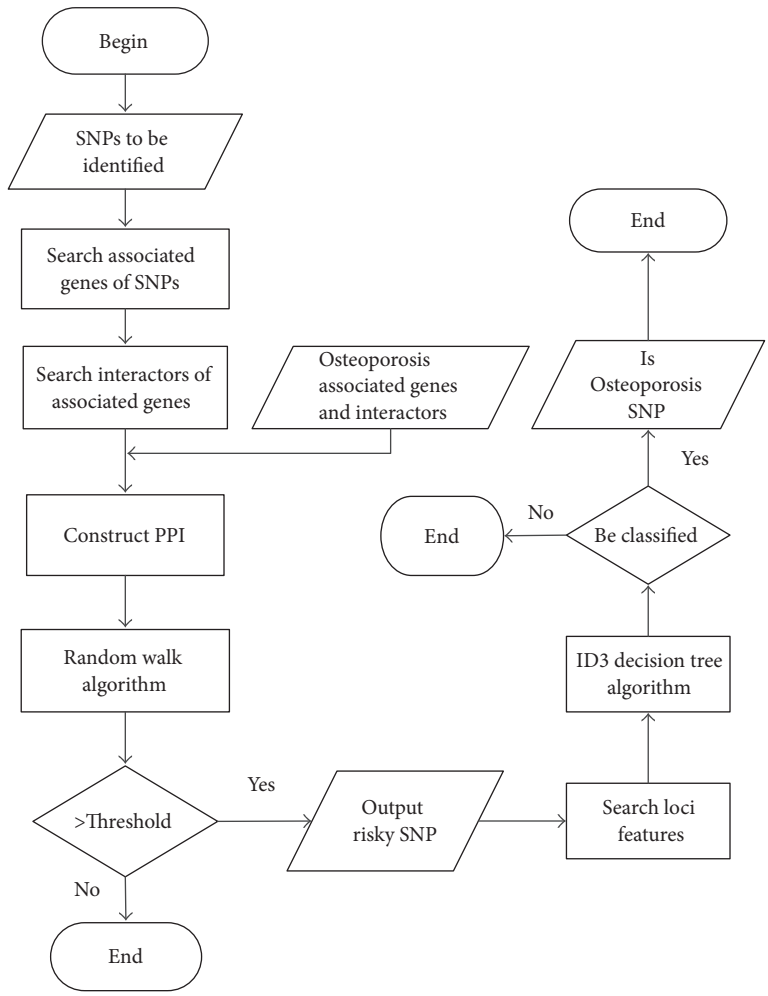


Figure 1: Process to identify the suspected risky SNPs associated with osteoporosis.

probability that the system in state i at the initial time, and $\sum_{i=0}^{N} q_i = 1$.

Based on the above theory model, the random walk on graphs is defined as an iterative walk's transition from its current node to a randomly selected neighbor starting at given source node [17]. The random walk is defined as

$$P^{t+1} = (1 - \alpha) P^t W + \alpha P^0. \quad (2)$$

P^t is a vector in which the ith element holds the probability of being at node i at time step t. α is a constant between 0 and 1 that it is the restart of the walk in every step at the node i with probability α, and $\alpha \in (0, 1]$ [17]. P^0 is a row vector of $1 \times n$ which is the initial state of the system, and n is the element number of V. The value of known elements of P^0 is equal, and the sum of them is 1. And the value of other elements is 0. W is the transition probability matrix which is defined as

$$W = D^{-1} A. \quad (3)$$

A is an adjacency matrix of undirected graph V. Every element a_{ij} of A is defined as follows: if there is interaction between v_i and v_j in the network, the element $a_{ij} = 1$; otherwise, $a_{ij} = 0$ the formula is defined as

$$a_{ij} = \begin{cases} 1, & \langle v_i, v_j \rangle \in E \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

D is a diagonal matrix. Each element d_{ij} of D is defined as follows: if $i = j$ then it should have $d_{ij} = d_{ii}$; otherwise $d_{ij} = 0$. d_{ij} is the degree of v_i in the network. The formula is defined as

$$d_{ij} = \begin{cases} \sum_{k=1}^{n} a_{ik}, & i = j \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

The transition probability matrix W is also a row-normalized adjacency matrix of the graph. Formula (2) meets the state of stationary distribution of Markov train model obviously, so the central point of random walk algorithm is evaluating the stationary distribution state of the probability of the nodes in the undirected network G which consists of PPI. Firstly, the transition probability matrix W should be

obtained and the initial value is set for P^0. Then, process t times iteration based on formula (2) until $\lim(p^{t+1} - p^t) = 0$, p^{t+1} is a convergence vector. A threshold is set for the probability value, and if the probability value of the nodes (or genes) is greater than the threshold, they are osteoporosis associated genes.

2.3. Classify the Suspected Risky SNPs by ID3 Decision Tree Algorithm. ID3 decision tree algorithm is a classification algorithm for tree structure [18, 19]. The goal of the algorithm is to predict target variable based on multiple input variables and deduce a classification rule with decision tree form from a group of irregular samples. We assume that all input characteristic elements have a limited discrete domain and need an individual characteristic element as a category. The nonleaf nodes of a classification decision tree classify the samples by characteristics of samples, and each leaf node of the tree is a class or classes of probability distribution. Therefore, we chose decision tree to classify the SNP based on the condition of the training set and algorithm characteristics.

SNPs located within the promoter or distant enhancer region of genes may alter the binding of TFs with DNA and subsequently regulate gene expression [20]. The suspected risky SNPs are classified by ID3 decision tree algorithm based on four features of significant position on genes, mapping on putative enhancer region, mapping on distal interaction, and the region where the SNPs are located [21].

The decision tree algorithm chooses the attribute with the maximum information gain after it is split, and the algorithm searches the decision-space by way of top-down greedy algorithm. S is defined as the training set of SNPs with their loci features, and the training set is divided into n classes. That is, $C = \{S_1, S_2, \ldots, S_n\}$. The number of the training instances in ith class is defined as $|S_i| = C_i$. The number of the training instances in S is $|S|$. The probability that a training instance belongs to the ith class is $P(S_i)$. And a formula is defined as

$$P\left(S_i\right) = \frac{C_i}{|S|}. \tag{6}$$

For the training set S, $H(S)$ is defined as the information entropy of C, and we have the formula

$$H\left(S\right) = -\sum_{i=1}^{n} P\left(S_i\right) \log_2 P\left(S_i\right). \tag{7}$$

The greater the value of information entropy $H(S)$ is, the smaller the degree of uncertainty for the division of C is. The attribute T is selected as the test attribute which is the loci features of the training set SNPs, and the value set for attribute T is $T = \{t_1, t_2, \ldots, t_m\}$. The probability of the attribute belongs to ith class when $T = t_j$ can be formulated as

$$P\left(S_i \mid T = t_j\right) = \frac{C_{ij}}{|S|}. \tag{8}$$

C_{ij} is the number of training instances which belongs to ith class.

When the attribute $T = t_j$, a formula is used to define the conditional entropy of the attribute T as

$$H\left(X_j\right) = -\sum_{i=1}^{n} P\left(S_i \mid T = t_j\right) \log_2 P\left(S_i \mid T = t_j\right). \tag{9}$$

X_j is the training instances set of training set S.

The information entropy of attribute T is defined as

$$\mathrm{IG}\left(T\right) = H\left(S\right) - \sum_{j=1}^{m} P\left(S_i \mid T = t_j\right) H\left(X_j\right). \tag{10}$$

We built a top-down decision tree and classified the training instances by choosing the attribute with the maximum information entropy based on the formulas above.

However, the overfitting problem could not be avoided if there were many noise samples in the training set, because of a complicated classification decision tree constructed by ID3 decision tree algorithm with a fair amount of noise samples in the training set. To solve the problem, a PEP (Pessimistic-Error Pruning) algorithm was exerted on the ID3 decision tree classification algorithm. PEP is the most accurate top-down pruning strategy which deals with the pruning problem without separating the training set.

We define a decision tree T which grows on a large scale based on the training set of SNPs with their loci features. T_1 is a nonleaf node set, T_2 is a leaf node set, and T_3 is for all nodes of T. The formula is $T_3 = T_1 \cup T_2$.

Before pruning, we define $r(t)$ as the error rate of node t in the decision tree. The formula is

$$r\left(t\right) = \frac{e\left(t\right)}{n\left(t\right)}. \tag{11}$$

$n(t)$ is the number of samples in node t, and $e(t)$ is the number of samples that does not belong to node t actually.

We define T_t as a subtree of the decision tree T, and t is the root node of T_t. So the error rate of the subtree T_t is

$$r\left(T_t\right) = \frac{\sum_{s \in S_t} e\left(s\right)}{\sum_{s \in S_t} n\left(s\right)}. \tag{12}$$

S_t is the leaf node set of subtree T_t, and we define $S_t = \{s_1, s_2, \ldots, s_n\}$.

Apparently, the formula for error rate of the subtree T_t is binomial distribution. We define a continuity correction factor $r'(t)$ in order to make the binomial distribution approach the normal distribution. And the formula is

$$r'\left(t\right) = \frac{e\left(t\right) + 1/2}{n\left(t\right)}. \tag{13}$$

Therefore, we deduce the continuity correction factor for the subtree T_t. The formula is

$$r'\left(T_t\right) = \frac{\sum_{s \in S_t} \left[e\left(s\right) + 1/2\right]}{\sum_{s \in S_t} n\left(s\right)} = \frac{\sum_{s \in S_t} e\left(s\right) + \left|S_t\right|/2}{\sum_{s \in S_t} n\left(s\right)}. \tag{14}$$

In order to simplify the formula, we define $e'(t)$ as the error sample number instead of error rate. So the error sample number of node t in the decision tree T is

$$e'(t) = e(t) + \frac{1}{2}. \tag{15}$$

Therefore, the error sample number of the subtree T_t is

$$e'(T_t) = \sum_{s \in S_t} e(s) + \frac{|S_t|}{2}. \tag{16}$$

Similarly, the formula for the error sample number of subtree T_t is binomial distribution. And the standard deviation for $e'(T_t)$ is defined as

$$\mathrm{SE}\left(e'(T_t)\right) = \left[\frac{e'(T_t) \times \left(n(t) - e'(T_t)\right)}{n(t)}\right]^{1/2}. \tag{17}$$

Finally, we deduce from formulas above that the subtree T_t will be cut if the node t meets the condition:

$$e'(t) \leq e'(T_t) + \mathrm{SE}\left(e'(T_t)\right). \tag{18}$$

The process of the PEP algorithm is as follows:

Algorithm: PEP
Begin
Input: decision tree T before pruned
Output: decision tree T after pruned
(1) Get the nonleaf node set T_1 of the decision tree T
(2) For $k = 1$ to length (T_1)
(3) Do get a subtree T_t whose root node is $t[k]$ $(t[k] \in T_1)$
(4) If $(e'(t) \leq e'(T_t) + SE(e'(T_t)))$
(5) Then delete T_t
(6) Else $k++$
(7) End
End

We classified the suspected risky SNPs effectively based on their loci characteristics and studied their functions according the ID3 decision tree algorithm and PEP.

3. Results

By the end of 2014, nine GWAS and nine meta-analyses had reported 107 genes and 129 SNPs (lead SNP) that were associated with BMD, osteoporosis, or fractures with a significant threshold of 5×10^{-8}. 222 SNPs linked to osteoporosis GWAS-associated lead SNPs had also been identified by using LD information in the Caucasians population via HapMap website [9]. Moreover, we obtained 107 known osteoporosis GWAS-associated genes which showed significant connectivity among proteins. And there were interactions between osteoporosis GWAS-associated genes and interactors. We used the common Protein-Protein Interaction databases, such as Human Protein Interaction database (HPID) and General Repository for Interaction Data (GRID), to find the interactors which had interactions with the osteoporosis GWAS-associated genes and their interactions. Then, we obtained the interaction network graph by Cytoscape v3.4.0. Figure 2 is the PPI of osteoporosis GWAS-associated genes.

Table 1: Part of the classification of training set.

SNP	bda	td	Enhancer	Gene region	Class
rs7524102	Y	Y	Y	Intergenic	C
rs34920465	Y	Y	Y	Control region	D
rs6426749	Y	N	Y	Control region	G
rs1430742	N	N	N	Cording sequence	B
rs6929137	Y	Y	Y	Missense	A
rs479336	Y	Y	N	Cording sequence	K
rs11898505	Y	N	Y	Intergenic	F
rs17040773	Y	Y	Y	Cording sequence	E
rs344081	Y	N	Y	Cording sequence	H
rs6909279	Y	Y	N	Intergenic	I

(a) The first column is part of osteoporosis GWAS-associated SNPs; (b) the column of "bda," "td," and "enhancer" means whether the SNP is on significant TFs binding affinity, mapping on distal interaction, and mapping on putative enhancer region; (c) the last column is the category the SNP belong to.

The result was verified by 10-fold cross-validation based on the data set of osteoporosis GWAS-associated genes and SNPs. We divided the data set of 129 osteoporosis GWAS-associated lead SNPs and 222 SNPs linked with them into 10 samples. One sample was then randomly chosen and saved as the validation set to verify the model from the 10 samples, and the other 9 samples were saved as training set. The verification process was repeated 10 times so that each sample was the validation set once, and the accuracy was calculated every time. A 10-fold cross-validation was completed by the process above.

We set a threshold k ($k > 10^{-3}$) as a result of the validation. The recall was calculated, which was the true positive result to positive result ratio. The 10-fold cross-validation was repeated for ten times and the average recall rate of every validation was calculated. The result was shown in Figure 3.

The classification result was also verified by 10-fold cross-validation. The osteoporosis GWAS-associated SNPs were used as the data set. The SNPs of training set were classified based on their loci features. Part of classification of the training set was shown in Table 1. We classified the SNPs of validation set through ID3 decision tree algorithm and recorded the accuracy of classification, which was the proportion of classification accurate samples to all the samples.

Then, the process of validation was repeated for ten times and calculated the average accuracy rate and average classification reliability. The result was shown in Figure 4.

We also used genome-wide association studies (GWAS) of type 2 diabetes (T2D) data as negative data to verify our method [22]. 50 lead SNPs of T2D were obtained with

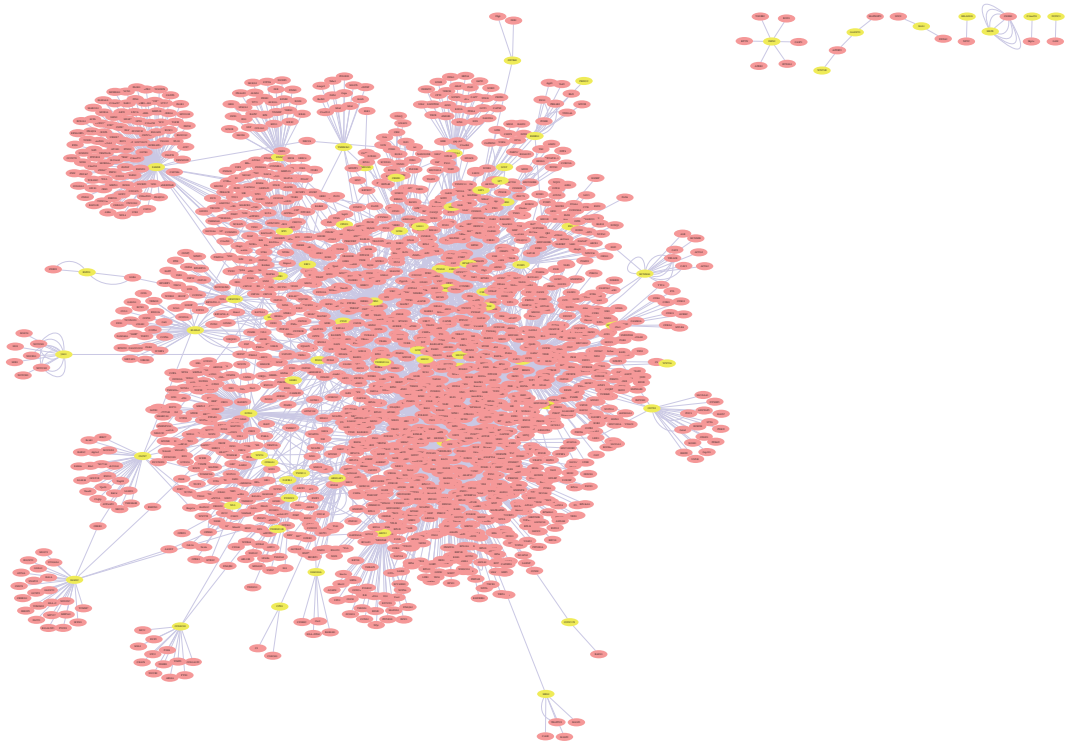


Figure 2: PPI of osteoporosis GWAS-associated genes (the pink nodes indicated those which had interactions with the osteoporosis GWAS-associated genes, and the yellow nodes indicated the osteoporosis GWAS-associated genes).

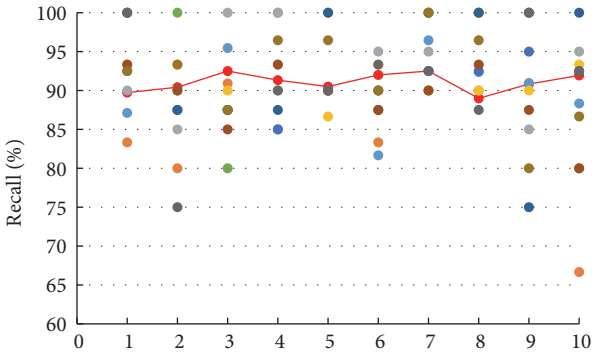


Figure 3: Result of random walk (the ten colors of the points indicated ten 10-fold cross-validation, and the same color of points indicated the validation process. The points connected by a line were the average recall value of ten experiments. The x-axis was the 10-step verification of the 10-fold cross-validation process).

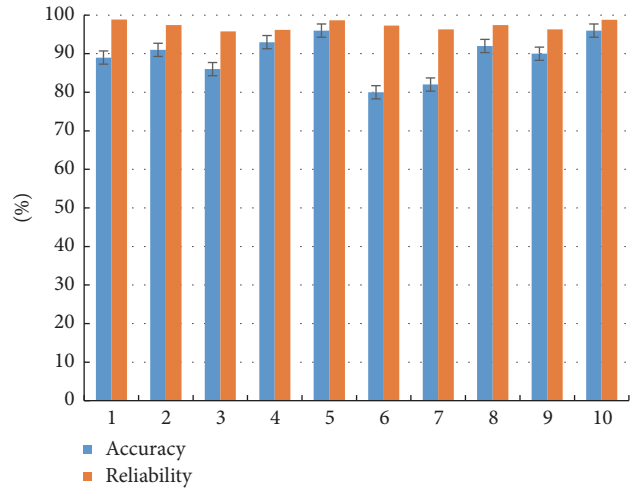


Figure 4: Result of ID3 decision tree (the blue credibility refers to the average accuracy values of 10-fold cross-validation, and the orange credibility refers to the average reliability value).

their position features and associated genes. We searched the interactors of the associate genes from the PPI database and constructed the PPI network with the known osteoporosis GWAS-associated genes. The random walk algorithm was used on the PPI network.

We then used PEP for ID3 decision tree to construct a simplified classification decision tree. We combined the two steps of the risky SNPs identification method and verified the method by 10-fold cross-validation. Finally, we found that not only was the computation efficiency improved, but also the accuracy rate of the result by using ID3 decision tree algorithm with PEP in the identification method was higher. The improvement is due to the fact that we had cut the subtrees which were constructed by the noise samples and solved the overfitting problem. While we defined ID3

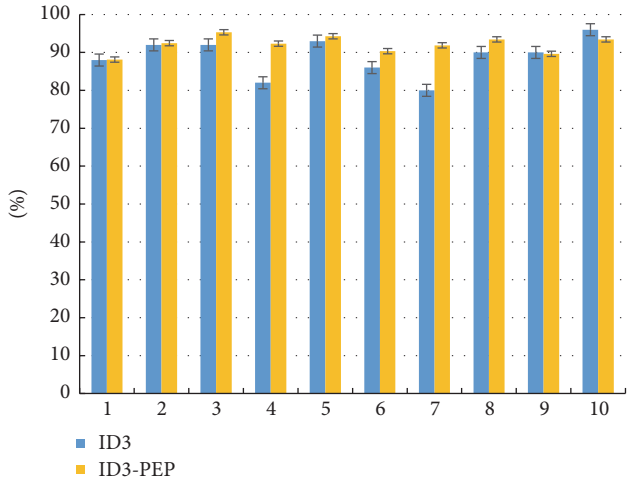


Figure 5: Comparison of two classification algorithm (the blue credibility refers to the classification accuracy by ID3 algorithm, and the yellow credibility refers to the classification accuracy by ID3 decision tree algorithm with PEP).

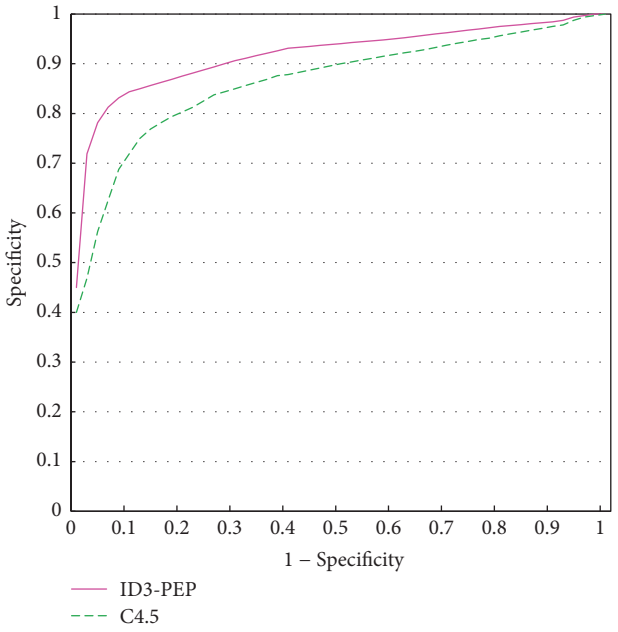


Figure 6: The comparison of ID3-PEP and C4.5.

decision tree algorithm with PEP in the identification method as ID3-PEP and ID3 decision tree algorithm as ID3, the result comparison of these two classification algorithm in the identification method was described by Figure 5. According to the result, we concluded that the ID3-PEP in the identification method was more stable than ID3 algorithm, and it had better effect for the classification problem.

C4.5 is the optimization of ID3. They have the same way to learn training set and build a classification decision tree, but the difference of them is the way of choosing split attribute. C4.5 algorithm chooses the maximum attribute with information gain ratio to split. In order to solve the problem of overfitting in ID3 decision tree algorithm, C4.5 algorithm needs to scan the data set and rank them in every step. This calculation method and process of the algorithm have low operational efficiency. ID3-PEP algorithm solved the problem and was more accurate than C4.5. We made a comparison of these two algorithms through ROC curve, which is shown in Figure 6. Result shows that ID3-PEP is better than C4.5 in our classification.

4. Discussion and Conclusion

Since SNP plays a key role in the process of pathology and susceptibility of osteoporosis [23], it is necessary to find the unknown risky SNPs. Using the data set of known osteoporosis GWAS-associated SNPs and genes [8], we identified the genes of suspected risky SNPs associated with osteoporosis by random walk algorithm on the PPI network constructed by osteoporosis GWAS-associated genes and the genes associated with suspected risky SNPs. The suspected risky SNPs were classified based on the features of their loci position and function. We used 10-fold cross-validation to verify our method.

The result of the experiment above showed that the identification method for risky SNPs of osteoporosis was correct and effective. Our method efficiently achieved the process of identifying osteoporosis suspected risky SNPs.

However, there is still a need to perfect the identification method. First of all, we need to search the loci features of suspected risky SNPs associated with osteoporosis and the interactors of associated genes manually. The training set for our method is the known osteoporosis GWAS-associated SNPs, which is not large enough to identify the risky SNPs accurately. Therefore, further research is needed. Firstly, a workflow can be constructed to improve the identification process, aiming to automatically identify the suspected risky SNPs' features. In order to improve the accuracy of our method, more features of the SNPs should be examined, such as the conservation of SNPs and the influence of the SNPs on miRNA binding site. Finally, we use our method to predict risky SNPs associated with osteoporosis by constructing the PPI network of all the human genes.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper and the received funding did not lead to any conflicts of interest regarding the publication of this manuscript.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grants nos. 61532008 and 31371275), the National Social Science Foundation of China (no. 14BYY093), and the Fundamental Research Funds for the Central Universities (no. CCNU17TS0003).

References

[1] F. Rivadeneira, U. Styrkársdottir, K. Estrada, B. V. Halldórsson, Y. H. Hsu, J. B. Richards et al., "Twenty bonemineral-density loci identified by large-scale meta-analysis of genome-wide association studies," *Nature Genetics*, vol. 41, no. 11, pp. 1199-206, 2009.

[2] D. Welter, J. MacArthur, J. Morales et al., "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1001–D1006, 2014.

[3] Q.-Y. Huang, R. R. Recker, and H.-W. Deng, "Searching for osteoporosis genes in the post-genome era: Progress and challenges," *Osteoporosis International*, vol. 14, no. 9, pp. 701–715, 2003.

[4] Q. Huang and A. W. C. Kung, "Genetics of osteoporosis," *Molecular Genetics and Metabolism*, vol. 88, no. 4, pp. 295–306, 2006.

[5] J. B. Richards, H. F. Zheng, and T. D. Spector, "Genetics of osteoporosis from genome-wide association studies: advances and challenges," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 576–588, 2012.

[6] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak, "VISTA: computational tools for comparative genomics," *Nucleic Acids Research*, vol. 32, pp. W273–W279, 2004.

[7] Q. Y. Huang, "Genetic study of complex diseases in the post-GWAS era," *Journal of Genetics and Genomics*, vol. 42, no. 3, pp. 87–98, 2015.

[8] S. L. Edwards, J. Beesley, J. D. French, and A. M. Dunning, "Beyond GWASs: illuminating the dark road from association to function," *American Journal of Human Genetics*, vol. 93, no. 5, pp. 779–797, 2013.

[9] L. Qin, Y. Liu, Y. Wang et al., "Computational characterization of osteoporosis associated SNPs and genes identified by genome-wide association studies," *Plos One*, vol. 11, no. 3, Article ID e0150070, pp. 1–14, 2016.

[10] M. J. Li, L. Y. Wang, Z. Xia, P. C. Sham, and J. Wang, "GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications," *Nucleic acids research*, vol. 41, pp. W150–W158, 2013.

[11] J. Yang, H. Gu, X. Jiang, Q. Huang, X. Hu, and X. Shen, "Walking in the PPI network to predict the risky SNP of osteoporosis with decision tree algorithm," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM '16)*, pp. 1283–1287, Shenzhen, China, 2016.

[12] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[13] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, M. A. Durbin, and R. E. Handsaker, "An integrated map of genetic variation from 1, 092 human genomes," *Nature*, vol. 467l, no. 7311, pp. 52–58, 2010.

[14] K. Lage, E. O. Karlberg, Z. M. Storling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[15] 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.

[16] R. K. Nibbe, S. A. Chowdhury, M. Koyuturk, R. Ewing, and M. R. Chance, "Protein-protein interaction networks and subnetworks in the biology of disease," *Systems Biology and Medicine*, vol. 3, no. 3, pp. 357–367, 2010.

[17] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[18] M. J. Blow, D. J. McCulley, Z. Li et al., "ChIP-seq identification of weakly conserved heart enhancers," *Nature Genetics*, vol. 42, no. 9, pp. 806–812, 2010.

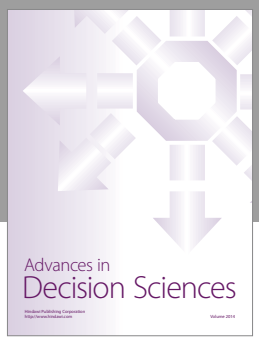
[19] J. R. Quinlan, "Generating production rules from decision trees," in *Proceedings of the IJCAI-87*, Milan, Italy, 1987.

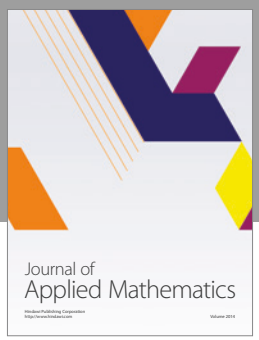
[20] L. A. Hindorff, P. Sethupathy, H. A. Junkins et al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 23, pp. 9362–9367, 2009.

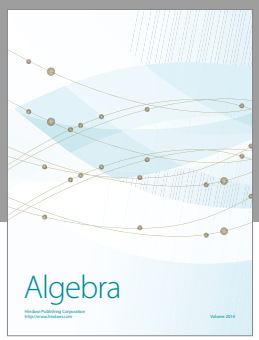
[21] A. Visel, E. M. Rubin, and L. A. Pennacchio, "Genomic views of distant-acting enhancers," *Nature*, vol. 461, no. 7261, pp. 199–205, 2009.

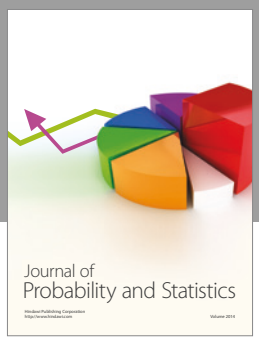
[22] M. Cheng, X. Liu, M. Yang, L. Han, A. Xu, and Q. Huang, "Computational analyses of type 2 diabetes-associated loci identified by genome-wide association studies," *Journal of Diabetes*, vol. 9, no. 4, pp. 362–377, 2016.

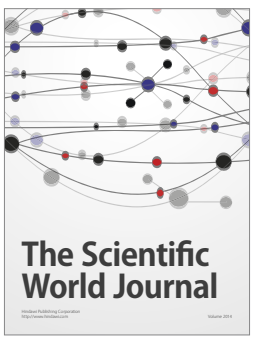
[23] E. T. Dermitzakis and A. G. Clark, "Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover," *Molecular Biology and Evolution*, vol. 19, no. 7, pp. 1114–1121, 2002.

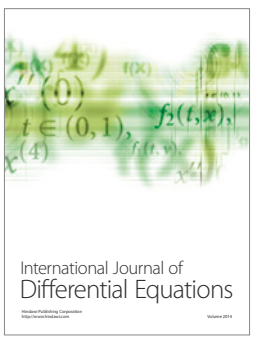


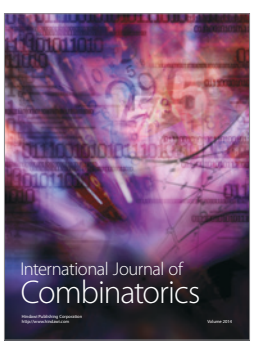


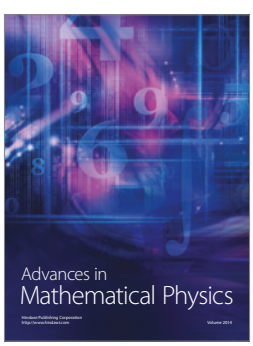


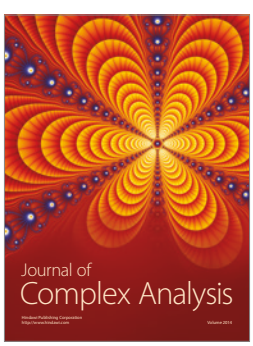


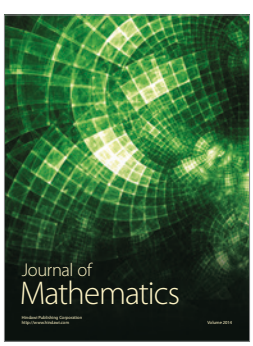


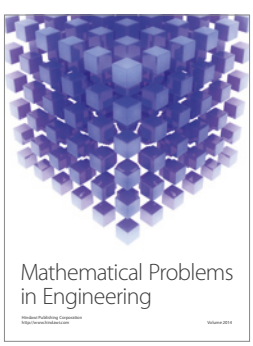


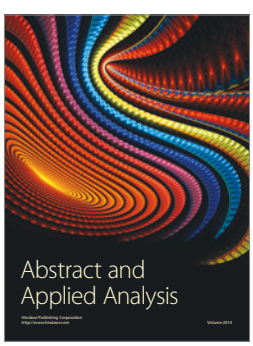


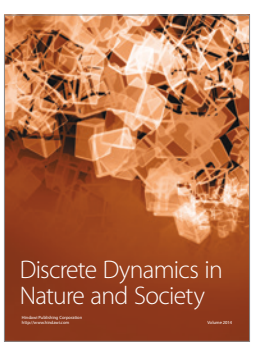


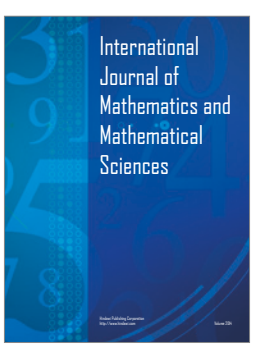


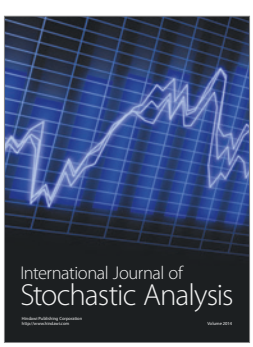












Hindawi

Submit your manuscripts at
https://www.hindawi.com