J. Biomedical Science and Engineering, 2008, 1, 52-58

Scientific Research Publishing

A combinatorial analysis of genetic data for Crohn's disease

Weidong Mao[1] & Jeonghwa Lee[2]

[1]Department of Mathematics & Computer Science, Virginia State University, Petersburg, VA 23806, USA. [2]Department of Computer Science, Shippensburg University, Shippensburg, PA 17257, USA. Correspondence should be addressed to Weidong Mao (wmao@vsu.edu) or Jeonghwa Lee(jlee@ship.edu).

ABSTRACT

The both environmental and genetic factors have roles in the development of some diseases. Complex diseases, such as Crohn'sdisease or Type II diabetes, are caused by a combination of environmental factors and mutations in multiple genes. Patients who have been diagnosed with such diseases cannot easily be treated. However, many diseases can be avoided if people at high risk change their living style, one example being their diet. But how can we tell their susceptibility to diseases before symptoms are found and help them make informed decisions about their health? With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. This paper uses a combinatorial method to analyze the genetic data for Crohn's disease and search disease-associated factors for given case/control samples. An optimum random forest based method has been applied to publicly available genotype data on Crohn's disease for association study and achieved a promising result.

Keywords: Genetic factor; Crohn's disease; Random forest

1. INTRODUCTION

Crohn's disease (also known as regional enteritis) is a chronic, episodic, inflammatory condition of the gastrointestinal tract characterized by transmural inflammation (affecting the entire wall of the involved bowel) and skip lesions (areas of inflammation with areas of normal lining in between). Crohn's disease is a type of inflammatory bowel disease (IBD) and can affect any part of the gastrointestinal tract from mouth to anus. As a result, the symptoms of Crohn's disease can vary among affected individuals. The exact cause of Crohn's disease is unknown. However, research shows that the inflammation seen in the people with Crohn's disease involves several factors: the genes the patient has inherited, the immune system itself, and the environment [1]. In other words, genetic factor has been invoked in the pathogenesis of the disease.

Although the Crohn's disease cannot easily be treated, it can be avoided if people at high risk change their living style, such as their diet. But how can we tell the susceptibility of people to the disease before symptoms are found and help them make informed decisions about their health? With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. Assessing the association between DNA variants and disease has been used widely to identify regions of the genome and candidate genes that contribute to disease [2].

99.9% of one individual's DNA sequences are identical to that of another person. Over 80% of this 0.1% difference will be Single Nucleotide Polymorphisms (SNP) and they promise to significantly advance our ability to understand and treat human disease. A SNP is a single base substitution of one nucleotide with another. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. It is important to study SNPs because they represent genetic differences among human beings. Genome-wide association studies require knowledge about common genetic variations and the ability to genotype a sufficiently comprehensive set of variants in a large patient sample [3]. High-throughput SNP genotyping technologies make massive genotype data, with a large number of individuals, publicly available. Accessibility of genetic data makes genome-wide association studies for complex diseases possible.

Success stories when dealing with diseases caused by a single SNP or gene, sometimes called monogenic diseases have been reported [4]. However, most complex diseases, such as psychiatric disorders, are characterized by a non-mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other [5]. A fundamental issue in the analysis of SNP data is to define the unit of genetic function that influences disease risk. Is it a single SNP, a regulatory motif, an encoded protein subunit, a combination of SNPs in a combination of

genes, an interacting protein complex, a metabolic or a physiological pathway [6]? In general, it may be impossible to associate a single SNP or gene with a disease because a disease may be caused by completely different modifications of alternative pathways, and each gene only makes a small contribution. This makes the identification of genetic factors difficult. Multi-SNP interaction analysis is more reliable but it is computationally infeasible. An exhaustive search among multi-SNP combination is computationally infeasible even for a small number of SNPs. Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease.

It is important to search for informative SNPs among a huge number of SNPs. These informative SNPs are assumed to be associated with genetic diseases. Tag SNPs generated by the multiple linear regression based method [7] are good informative SNPs, but they are reconstruction-oriented instead of disease-oriented. Although the combinatorial search method [8] for finding disease-associated multi-SNP combinations has a better result, the exhaustive search is still very slow.

Multivariate adaptive regression spline models [9, 10] are used to detect associations between diseases and SNPs with some degree of success. However, the number of selected predictors is limited, and the type of possible interactions must be specified in advance. Multifactor dimensionality reduction methods [11, 12] are developed specifically to find gene-gene interactions among SNPs, but they are not applicable to a large set of SNPs.

Random forest model has been explored in disease association studies [13], but it was applied on simulated case-control data in which the interacting model among SNPs and the number of associated SNPs are specified, thus making the association model simple and the association is relatively easier to detect. For real data, such as Crohn's disease [14], multi-SNP interaction is much more complex , which involves more SNPs.

In Section 2 of this paper, we propose an optimum random forest model for searching the disease-associated multi-SNP combination for given case-control data. In the optimum random forest model, we generate a forest for each variable (e.g. SNP) instead of randomly selecting some variables to grow the classification tree. We can find the best classifier (a combination of SNPs which includes the SNP) for each SNP, and then we may have M classifiers if the length of the genotype is M. We rank classifiers according to their prediction rate, and the SNP with a higher prediction rates is more disease-associated.

The association of multi-SNP combination can be measured by the disease susceptibility prediction rate. In Section 3 we address the disease susceptibility prediction problem [15, 16, 17, 18]. The goal of disease susceptibility prediction is to assess accumulated information targeted to predicting susceptibility to complex diseases with significantly high accuracy and statistical power. The problem is based on the association study we described above. The Disease-associated multi-SNP combination found in association studies can be used to predict the susceptibility to diseases. On the other side, the prediction results can be used to evaluate the accuracy of the association studies. A higher prediction rate means the higher reliability of the association studies.

The proposed method is applied to analyze the genetic data of the Crohn's disease. We find the disease-associated multi-SNP combination and apply it to predict the susceptibility. The accuracy of the prediction is higher than that of all previously known methods. It can be also applied in disease prevention and control in the near future. For example, after training the available case-control genome data, we can find those significant SNPs which are well associated with the disease. When a patient comes, and we obtain his/her genetic data, we don't need to check the whole sequence, but only disease-associated SNPs instead. This will save much money and time for diagnosis and can be done before the onset of diseases. Therefore, treatment could start earlier to prevent or delay the occurrence of the disease.

2. DISEASE ASSOCIATION SEARCH FOR CROHN'S DISEASE

In this section we first give an overview of the random forest tree and classification tree, then we will describe the genetic model. Next we propose the optimum random forest algorithm to search Tag SNPs.

2.1. Classification Trees and Random Forest

In machine learning, a Random Forest is a classifier that consists of many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number $m<<M$ is specified such that at each node, m variables are selected randomly out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning [19].

A different bootstrap sample from the original data is used to construct a tree. Therefore, about one-third of the cases are left out of the bootstrap sample and not used in the construction of the tree. Cross-validation is not required because the one-third **oob (out-of-bag)** data is used to get an unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance. After each tree is built, we compute the proximities of each terminal node.

In every classification tree in the forest, put down the **oob** samples and make prediction the classification of the **oob** samples. In such way we can compute the importance score for variables in each tree based

on the number of votes cast for the correct class. All variables can be ranked and those important variables can be found in this way.

Random forest is a sophisticated method in data mining to solve classification problems, and it can be used efficiently in disease association studies to find most disease-associated variables such as SNPs that may be responsible for diseases.

2.2. Genetic Model

Recent work has suggested that SNPs in human population are not inherited independently; rather, sets of adjacent SNPs are present on alleles in a block pattern, so called **haplotype**. Many haplotype blocks in human have been transmitted through many generations without recombination. This means although a block may contain many SNPs, it takes only a few SNPs to identify or to tag each haplotype in the block. A genome-wide haplotype would comprise half of a diploid genome, including one allele from each allelic gene pair. The **genotype** is the descriptor of the genome which is the set of physical DNA molecules inherited from the organism's parents. A pair of haplotype consists of a genotype.

SNPs are bi-allelic and can be referred as 0 for majority allele and 1, otherwise. If alleles on both haplotypes are the same, then the corresponding genotype is homogeneous, and can be represented as 0 or 1. If the two alleles on the two haplotypes are different, the genotype is heterozygous, represented as 2.

In **Figure 1**, there are four chromosomes, we assume the first two chromosomes belong to one person and the other two chromosomes belong to another person. We can find on most sites the four chromosomes are identical, but on some sites they are different, nucleotides on these sites are SNP. The haplotype is the concatenation of SNPs and a genotype is composed of two haplotypes.

The case-control sample populations consist of N individuals who are represented in genotype with M SNPs. Each SNP attains one of the three values 0, 1, or 2. The sample G is an (0, 1, 2)-valued N x M matrix, where each row corresponds to an individual, each column corresponds to a SNP.

The sample G has 2 classes, case and control, and M variables, and each of them represents a SNP. To construct a classification tree, we split the sample S into 3 child sub-samples, depending on the value (0, 1, 2) of the variable (SNP) on the splitting site (loci). In fact we can construct a binary tree (split sample according to homozygous or heterozygous), but there is no way to tell the difference between major allele (1) and minor allele (0). In order to distinguish them we split the sample into 3 sub-samples instead of 2. We grow the tree to the largest possible extent. The construction of the classification tree for case-control sample is illustrated in **Figure 2**. In the first level, we split the sample (30 genotypes, 14 cases and 16 controls) into 3 sub-samples (17, 8, 5) at loci 5 (the 5th SNP). In the second level, the first sub-sample splits at loci 9 and the second sub-sample splits at loci 7. No splitting is required for the third sub-sample because it is a terminal node with only one class. In the third level, the only split node splits at loci 3. The relationship of a leaf to the tree on which it grows can be described by the hierarchy of splits of branches (starting from the trunk) leading to the last branch from which the leaf hangs. The collection of split site is a Multi-SNPs combination (*MSC*), which can be viewed as a classification tree. In this example, $MSC = \{5, 9, 7, 3\}$ and $m = 4$, which is a collection of 4 SNPs, represented as their loci.

2.3. Searching for Disease Associated Multi-SNPs

To fully understand the basis of complex diseases, it

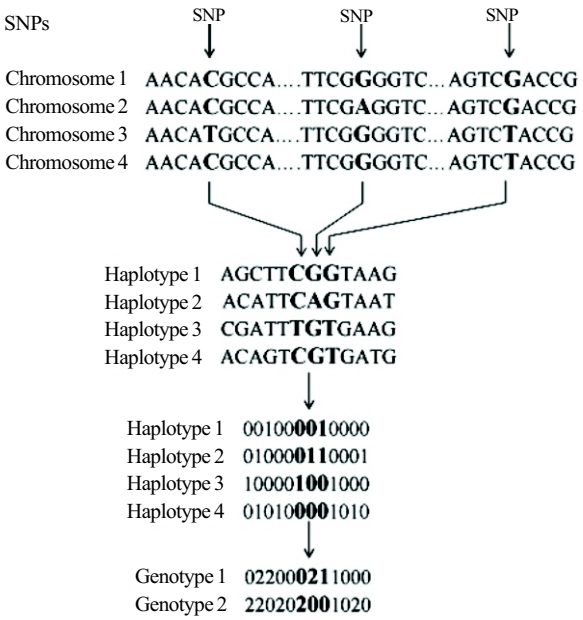

Figure 1. SNP, haplotype and genotype.

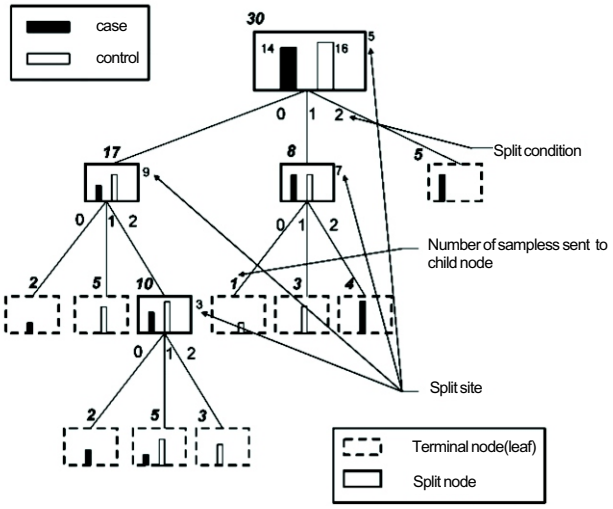

Figure 2. Classification tree for case-control sample.

is important to identify the critical genetic factors involved, which is a combination of multiple SNPs. For a given sample G, S is the set of all SNPs (denoted by loci) for the sample, and a multi-SNPs combination (MSC) is a subset of S. In disease associations, we need to find a MSC which consists of a combination of SNPs that are well associated with the disease. To find such MSC, we need first rank all SNPs according to their association degree (measured as weight) with diseases. Based on the sorting, we can find the n most disease associated SNPs for a given threshold n.

Although there are many statistical methods to detect the most disease associated SNPs, such as odds ratio or risk rates, the result is not satisfactory. We decide to use the random forest to find them.

2.4. Optimum Random Forest

We randomly generate a group of $MSCs$ for each SNP. The size of the MSC should be much less than the size of set S ($m << M$). Each MSC can be represented as a tree and all trees make the forest F. All trees (or $MSCs$) of the forest F_i ($i=1, 2, ..., M$) must include the i^{th} SNP and the other (m-1) SNPs can be randomly chosen from S except the i^{th} SNP. In this way, the M forests cover all SNPs in S.

We grow a classification tree for every MSC in each forest F_i. We run all the testing samples down these trees to get the classifier for each sample in the training set, then we can get a classification rate for each tree in F_i. The MSC_i is the representative for the forest F_i and the MSC_i has the highest classification rate among all trees in F_i. Each member (SNP) of the MSC_i is assigned a weight $w_{i,j}$ ($j \in MSC$) based on the classification rate. The weights for SNPs in the same MSC are the same. We can find M $MSCs$ for the M forests. If a SNP is not a member of MSC_i, then $w_{i,j} = 0$.

The weight for each SNP W_j (j = 1, 2, …, M) in M is the sum of weights from all $MSCs$.

$$W_j = \sum_{i=1}^{M} w_{i,j} \qquad (1)$$

In the general random forest (GRF) algorithm, the MSC is selected completely randomly and $m << M$. It may miss some important SNPs if they are not chosen for any MSC. In our optimum random forest (ORF) algorithm, this scenario is avoided because we generate at least one MSC for each SNP. On the other hand, in GRF, the classifier (forest) consists of trees where there is a correlation between any two trees in the forest, and the correlation will decrease the rate of the classifier. But in ORF, we generate a forest by randomly choosing MSC and samples for each tree and the prediction for testing samples is in this forest only, which is completely independent from the other trees. In this way, we extinguish the correlation among trees.

All SNPs are sorted according to their cumulative weights. The most disease-associated SNP is the one with the highest weight. The contribution to diseases of each SNP is quantified by its weight, but in GRF there is no way tell the difference of contribution among SNPs. The GRF can only tell the difference among classifiers (trees).

3. DISEASE SUSCEPTIBILITY PREDICTION

In this section we first describe the input and the output of prediction algorithms and then show how to apply the optimum random forest to the disease susceptibility prediction.

Data sets have n genotypes and each has m SNPs. The input for a prediction algorithm includes:

(G1) Training genotype set $g_i = (g_{i,j})$, $i = 0, 1, ..., n$, $j = 1, ... m$, $g_{i,j} \in \{0,1,2\}$

(G2) Disease status $s(g_i) \in \{0,1\}$, indicating if g_i, $i = 0, 1, ..., n$, is in case (*1*) or in control (*0*), and

(G3) Testing genotype g_t without any disease status.

We will refer to the parts (G1-G2) of the input as the training set and to the part (G3) as the test set. The output of prediction algorithms is the disease status of the genotype $s(g_t)$.

We use leave-one-out cross-validation to measure the quality of the algorithm. In the leave-one-out cross-validation, the disease status of each genotype in the data set is predicted while the rest of the data is regarded as the training set.

We describe several universal prediction methods below. These methods are adaptations of general computer-intelligence classifying techniques.

Closest Genotype Neighbor (CN). For the test genotype g_t, find the closest (with respect to Hamming distance) genotype g_i in the training set, and set the status $s(g_t)$ equals to $s(g_i)$.

Support Vector Machine Algorithm (SVM). Support Vector Machine (SVM) is a generation learning system based on recent advances in statistical learning theory. SVMs deliver a state-of-the-art performance in real-world applications and have been used in case/control studies [18, 20]. There are some SVM softwares available and we decide to use libsvm-2.71 [19] with the following radial basis function:

$$exp(-\tau^{※} | u-v |^{2})$$

General Random Forest (GRF). We use Leo Breiman and Adele Cutler's original implementation of RF version [19]. This version of RF handles unbalanced data to predict accurately. RF tries to perform a regression on the specified variables to produce the suitable model. RF uses bootstrapping to produce random trees and it has its own cross-validation technique to validate the model for prediction/classification.

Most Reliable 2 SNP Prediction (MR2) [17]. This method chooses a pair of adjacent SNPs (site of s_i and s_{i+1}) to predict the disease status of the test genotype g_t by voting among genotypes from the training set which have the same SNP values as g_t at the chosen sites s_i and s_{i+1}. They choose the 2 adja-

cent SNPs with the highest prediction rate in the training set.

LP-based Prediction Algorithm (LP). This method assumes that certain haplotypes are susceptible to the disease while others are resistant to the disease. The genotype susceptibility is then assumed to be a sum of susceptibilities of its two haplotypes.

We want to assign a positive weight to susceptible haplotypes and a negative weight to resistant haplotypes such that for any control genotype the sum of weights of its haplotypes is negative and for any case genotype it is positive. We would also like to maximize the confidence of our weight assignment which can be measured by the absolute values of the genotype weights. In other words, we would like to maximize the sum of absolute values of weights over all genotypes.

This method is based on a graph $X = \{H, G\}$, where the vertices H correspond to distinct haplotypes and the edges G correspond to genotypes connecting its two haplotypes. The density of X is increased by dropping SNPs which do not collapse edges with an opposite status. The linear program assigns weights to haplotypes that, for any non-diseased genotype, the sum of weights of its haplotypes is less than 0.5 and greater than 0.5 otherwise. We maximize the sum of absolute values of weights over all genotypes. The status of the testing genotype is predicted as sum of its endpoints [15].

Optimum Random Forest (ORF). In the training set, the optimum random forest algorithm we described above is used to sort all SNPs, and find out the m most disease associated SNPs for a given threshold m. The m most disease associated SNPs (Tag SNPs) are used to build the optimum random forest to test the left-out sample. In leave-one-out test, since the training set is different after leaving one sample out, we may have different Tag SNPs for different training sets. The m variables (SNPs) are used to grow many different classification trees by permuting the order of the splitting site (Note that the tree {3, 9, 5} is different from the tree {5, 9, 3}). We may use the m Tag SNPs to grow many (say, 500) trees and choose the best tree (classifier) to predict the disease status of the testing genotype. The best tree has the highest average prediction rate (over 1000 trials) in the training set. Then we run the testing genotype down the best tree to get its disease status. The Optimum Random Forest algorithm is illustrated in **Figure 3**.

4. RESULTS & DISCUSSION

In this section we first describe the genetic data of the Crohn's disease and then discuss our experimental results.

4.1. Data Set

The genetic data is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios [14]. All offspring belong to the case population, while almost all parents belong to the control population. In the entire data, there are 144 case and 243 control individuals. The missing genotype data and haplotypes have been inferred using the 2SNP phasing method [21].

4.2. Measures of Prediction Quality

To measure the quality of prediction methods, we need to measure the deviation between the true disease status and the result of predicted susceptibility, which can be regarded as measurement error. We will present the basic measures used in epidemiology to quantify the accuracy of our methods.

The basic measures are:

Sensitivity: the proportion of persons who have the disease and who are correctly identified as cases.

Specificity: the proportion of people who do not

Input: Training genotype set $G^{N,M}$, N: the number of samples, M: the number of SNPs
Disease status of $G^{N,M}$, $s^{N,M}$,
The threshold m,
Testing genotype g_t.

Sorting the M SNPs, find the MSC with the m most disease-associated SNPs
For $i = 1$ to 500,
  Permute the order of MSC, generate a tree T_i,
    For $j = 1$ to 1000,
    Randomly generate a bootstrapped sample S_j from G,
    Run S_j down the tree T_i to get the classification tree,
    Predict testing sample G'_j ($G'_j = G - S_j$) to get the prediction rate $p_{i,j}$,
  Compute the average prediction rate $\bar{p}_i$ for T_i,
Find the best tree T_b which has the highest $\bar{p}$,
Run g_t down the best tree T_b to get the disease status.

Output: Disease status of the test genotype $s(g_t)$.

Figure 3. Optimum Random Forest Algorithm.

have the disease and who are correctly classified as controls.

The definitions of these two measures of validity are illustrated in **Table1**.

In this table:

a = True positive, people with the disease who test positive

b = False positive, people without the disease who test positive

c = False negative, people with the disease who test negative

d = True negative, people without the disease who test negative

From **Table1**, we can compute Sensitivity (accuracy in classification of cases, Specificity (accuracy in classification of controls) and accuracy:

$$Sensitivity = \frac{a}{a+c} \quad (2)$$

$$Specificity = \frac{d}{b+d} \quad (3)$$

$$Accuracy = \frac{a+d}{a+b+c+d} \quad (4)$$

Sensitivity is the ability to correctly detect a disease. Specificity is the ability to avoid calling normal as disease. Accuracy is the percent of the population that are correctly predicted.

4.3. Results and Discussion

The normalized weights of 103 SNPs are shown in **Figure 4**. SNPs with higher weights are more associated with the disease.

In **Table 2** we compare the optimum random forest (ORF) method with the other 5 methods we described in Section 3. The best accuracy is achieved by ORF - 74.4%. From the results we can find that the ORF has the best result since we select the most disease-associated multi-SNPs to build the random forest for prediction. Because these SNPs are well associated with the disease, the random forest may produce a good classifier to reflect the association.

Table1. Classification contingency table.

		True Status	
		+	-
Classified	+	a	b
Status	-	c	d

Table 2. The comparison of the prediction rates of 6 prediction methods.

Measures	Prediction Methods					
	CN	SVM	GRF	MR2	LP	ORF
Sensitivity	45.5	20.8	34.0	30.6	37.5	70.1
Specificity	63.3	88.8	85.2	85.2	88.5	76.9
Accuracy	54.6	63.6	66.1	65.5	69.5	74.4

Figure 5 shows the receiver operating characteristics (ROC) curve for 6 methods. A ROC curve represents the tradeoffs between sensitivity and specificity. The ROC curve also illustrates the advantage of ORF over all previous methods.

If the size of *MSC* is *m*, and the total number of SNPs is *M*, to get a good classifier, then *m* should be much less than *M*. The prediction rate depends on the size of *MSC*, as shown in **Figure 6**. In our experiment, we found that the best size of *MSC* is 19.

5. CONCLUSION

In this paper, we discuss the potential of applying ran-

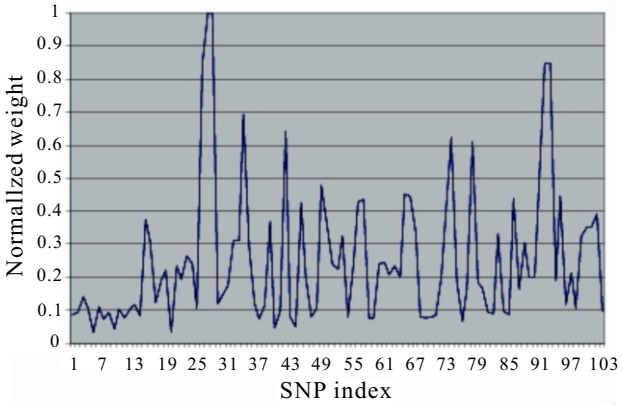

Figure 4. Normalized weights for 103 SNPs.

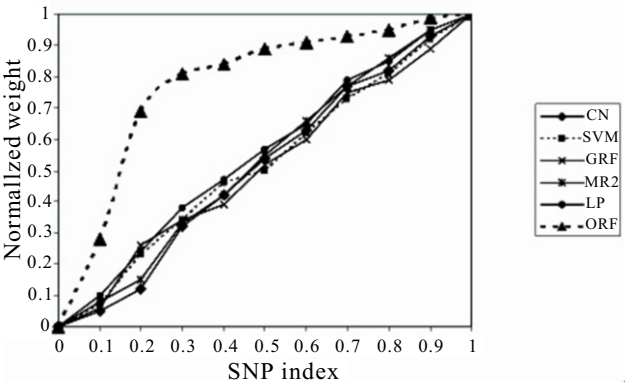

Figure 5. ROC curve for 6 prediction methods.

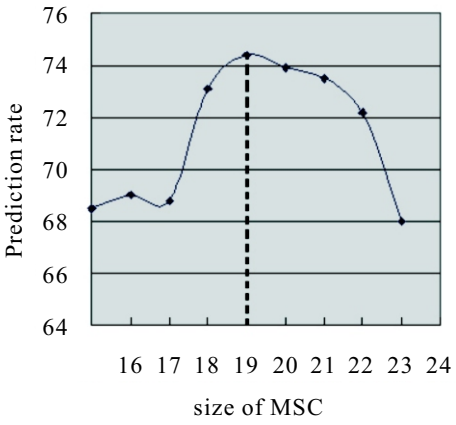

Figure 6. Best *MSC* size.

dom forest on disease association studies. The proposed genetic susceptibility prediction method based on the optimum random forest is shown to have a high prediction rate and the multi-SNPs being selected to build the random forest are well associated with diseases. Actually the cause of complex diseases is the combination of the environmental, genetic factors and some other factors such as infection and races. In our future work we are going to analyze the interactive contribution of these factors for the development of complex diseases. Our next project is going to find the relationship between the genetic factor and race in the development of Type 2 Diabetes. The integrated software will be available soon for public use.

REFERENCE

[1]National Digestive Diseases InformationClearinghouse (NDDIC), http://digestive.niddk.nih.gov/ddiseases/pubs/crohns.

[2]Cardon, L.R. & Bell, J.I. Association Study Designs for Complex Diseases. *Nature Reviews: Genetics* 2001, 2:91-98.

[3]Hirschhorn, J.N. & Daly, M.J. Genome-wide Association Studies for Common Diseases and Complex Diseases. *Nature Reviews: Genetics* 2005, 6:95-108.

[4]Merikangas, KR. & Risch, N. Will the Genomics Revolution Revolutionize Psychiatry. *American Journal of Psychiatry*, 2003, 160: 625-635.

[5]Botstein, D. & Risch, N. Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease. *Nature Genetics* 2003, 33: 228-237.

[6]Clark, A.G., Boerwinkle E., Hixson J. & Sing C.F. Determinants of the success of whole-genome association testing. *Genome Res.* 2005, 15:1463-1467.

[7]He, J. & Zelikovsky, A. Tag SNP Selection Based on Multivariate Linear Regression. *Proc. of International Conference on Computational Science* 2006, LNCS 3992:750-757.

[8]Brinza, D., He, J. & Zelikovsky, A. Combinatorial Search Methods for Multi-SNP Disease Association. *Proc. of International Conference of the IEEE Engineering in Medicine and Biology* 2006, pages 5802-5805.

[9]Cook N.R., Zee R.Y. & Ridker P.M. Tree and Spline Based Association Analysis of gene-gene interaction models for ischemic stroke. *Stat Med* 2004, 23(9):439-453.

[10]York T.P. & Eaves L.J. Common Disease Analysis using Multivariate Adaptive Regression Splines (MARS): Genetic Analysis Workshop 12 simulated sequence data. *Genetic Epidemiology* 2001, 21 (S1):649-654.

[11]Ritchie M.D., Hahn L.W., Roodi N., Bailey L.R., Dupont W.D., Parl F.F. & Moore J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*. 2001, 69: 138-147.

[12]Hahn L.W., Ritchie M.D. & Moore J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003, 19:376-382.

[13]Lunetta, K., Hayward, L., Segal, J. & Van Eerdewegh, P. Screening Large-scale Association Study Data: Exploiting Interactions Using Random Forests□, *BMC Genetics* 2004, pages 5:32.

[14]Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E. High resolution haplotype structure in the human genome. *Nature Genetics* 2001, 29:229-232.

[15]Mao, W., He, J., Brinza, D. & Zelikovsky, A. A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases. *Proc. International Conference of the IEEE Engineering In Medicine and Biology Society* 2005, pages 224-227.

[16]Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. & Zelikovsky, A. Genotype Susceptibility and Integrated Risk Factors for Complex Diseases. *Proc. IEEE International Conference on Granular Computing* 2006, pages 754-757.

[17]Kimmel, G. & Shamir R. A Block-Free Hidden Markov Model for Genotypes and Its Application to Disease Association. *J. of Computational Biology* 2005, 12(10): 1243-1260.

[18]Listgarten, J., Damaraju, S., Poulin B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner,R. & Zanke, B. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. *Clinical Cancer Research* 2004, 10:2725-2737.

[19]Breiman, L. & Cutler, A. *http://stat.berkeley.edu/breiman.*

[20]Waddell, M., Page, D., Zhan, F., Barlogie, B. & Shaughnessy, J., Predicting Cancer Susceptibility from SingleNucleotide Polymorphism Data: A Case Study in Multiple Myeloma. *Proc. of the 5th international workshop on Bioinformatics* 2005, pages 21-28.

[20]Chang, C. and Lin, C. *http://www.csie.ntu.edu.tw/libsvm.*

[21]Brinza, D. & Zelikovsky, A. 2SNP: Scalable Phasing Based on 2-SNP Haplotypes. *Bioinformatics* 2006, 22(3):371-373.