SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

February 2011 Vol.54 No.2: 139–145
doi: 10.1007/s11427-010-4117-y

Identification of a combination of SNPs associated with Graves' disease using swarm intelligence

WEI Bin[1,2,3], PENG QinKe[1,2,3*], ZHANG QuanWei[1,2,3] & LI ChenYao[1,2,3]

[1]*State Key Laboratory for Manufacturing Systems Engineering and School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;*
[2]*MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China;*
[3]*Electronic and Information School of Xi'an Jiaotong University, Xi'an 710049, China*

Received July 26, 2010; accepted November 3, 2010

Graves' disease, the production of thyroid-stimulating hormone receptor-stimulating antibodies leading to hyperthyroidism, is one of the most common forms of human autoimmune disease. It is widely agreed that complex diseases are not controlled simply by an individual gene or DNA variation but by their combination. Single nucleotide polymorphisms (SNPs), which are the most common form of DNA variation, have great potential as a medical diagnostic tool. In this paper, the *P*-value is used as a SNP pre-selection criterion, and a wrapper algorithm with binary particle swarm optimization is used to find the rule for discriminating between affected and control subjects. We analyzed the association between combinations of SNPs and Graves' disease by investigating 108 SNPs in 384 cases and 652 controls. We evaluated our method by differentiating between cases and controls in a five-fold cross validation test, and it achieved a 72.9% prediction accuracy with a combination of 17 SNPs. The experimental results showed that SNPs, even those with a high *P*-value, have a greater effect on Graves' disease when acting in a combination.

Graves' disease, single nucleotide polymorphism, case-control study, swarm intelligence

Citation: Wei B, Peng Q K, Zhang Q W, *et al*. Identification of a combination of SNPs associated with Graves' disease using swarm intelligence. Sci China Life Sci, 2011, 54: 139−145, doi: 10.1007/s11427-010-4117-y

Graves' disease (GD) is one of the most common forms of human autoimmune disease, with an estimated frequency of up to 1.3% (0.5% clinical and 0.7% subclinical) in the United States [1] and 0.25%–1.09% in China [2]. GD is caused by the production of thyroid-stimulating hormone receptor-stimulating antibodies leading to hyperthyroidism [3]. DNA variations in the human genome, which alter physiological pathways, are considered the primary risk factors for many diseases. Single-nucleotide polymorphisms (SNPs) are the most common form of DNA variation [4]. It is estimated that there are approximately 12 million SNPs in the human genome [5]. Although most are neutral, certain SNPs can affect phenotypes such as height, skin color, resistance to infection and susceptibility to disease [6].

Understanding the mechanisms underlying diseases will contribute to the development of future therapies. In the past few decades, association studies have identified some associations between genetic variants and diseases or human phenotypes [7,8], but the evaluation of combinations of SNPs has been less commonly addressed. Studies show that while individual SNPs may make a very small contribution to disease, combinations of SNPs can be strongly associated with complex diseases [9]. Similarly, GD is modestly affected by single SNPs, but might be greatly affected by a combination of SNPs. One of the major goals of association

*Corresponding author (email: qkpeng@xjtu.edu.cn)

studies is to identify the combinations of SNPs that lead to higher disease risk. The aim of this paper is to identify the rules for classifying the case (GD) and control (non-GD) groups. Because of the combinatorial explosion of the number of subsets, exhaustive enumeration was impractical for this study.

Recently, there have been several reports examining the effects of SNPs on disease [10,11]. Wan *et al*. [9] developed an alternative learning approach to select the important SNPs for disease. Kusiak and Shah [12] identified the SNPs that optimally predict the risk of disease by minimizing the classification error. Xie *et al*. [13] presented an adaptation of the decision forest pattern recognition algorithm for esophageal cancer association studies. Yasuyuki *et al*. [14,15] evaluated SNPs to predict the risk of allergic asthma. However, due to the non-deterministic polynomial-time-hard nature of this problem, these methods typically suffer from inefficiency and inaccuracy problems. In addition, the *P*-value has been seldom used in most of the previous studies.

In this paper, we propose a wrapper algorithm with binary particle swarm optimization to analyze the association between combinations of SNPs and GD, whose classifier is a two-layer linear classifier (TLC). Binary particle swarm optimization (BPSO) is used as a feature selection method to identify the combination of SNPs, and the TLC (optimized by particle swarm optimization (PSO)) is used to predict the susceptibility to disease. In addition, the *P*-value was used as the pre-selection criterion. In this study, we investigated 108 SNPs in major GD-related pathways in 384 cases and 652 controls. We achieved a 72.9% prediction accuracy with a combination of 17 SNPs. Experimental results demonstrate the feasibility of incorporating our method into a case-control study.

1 Materials and methods

1.1 Data

The dataset used in this paper was supplied by Dr. Dumitru Brinza [16]. The dataset was derived from 330 kb of human DNA sequence containing the genes *CD28*, *CTLA4* and *ICOS*, which are related to GD [17]. A total of 108 SNPs were genotyped in 384 cases and 652 controls [17].

1.2 Particle swarm optimization

PSO is an iterative optimization algorithm inspired by the observation of collective behaviors in animals (e.g., bird flocking) [18]. In PSO, each candidate solution to an optimization problem is represented by one particle. Each particle i is described by its position x_i and velocity v_i. The algorithm starts with random initialization of the particles. Then, the particles change their positions according to their velocities, which update in each iteration. Given that p_i is the best position found by particle i in all the preceding iterations and p_g is the best position found so far by the entire swarm, the velocity and position of particle i in bit j will be updated according to the following formulae:

$$v_{ij}(t+1) = v_{ij}(t) + c_1 r_1 (p_{ij}(t) - x_{ij}(t)) + c_2 r_2 (p_{gj}(t) - x_{ij}(t)), \quad (1)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1), \quad (2)$$

where r_1 and r_2 are random numbers between 0 and 1, and c_1 and c_2 define the degree of influence of p_i and p_g on the particle's velocity. The velocity v_{ij} is bounded within a range of $[-V_{\max}, V_{\max}]$ to prevent the particle from flying out of the solution space.

1.3 Binary particle swarm optimization

Because many optimization problems are set in a discrete space, Kennedy and Eberhart [19] extended the PSO to a BPSO in 1997. In BPSO, a particle moves in a state space restricted to 0 or 1 in each bit, where v_{ij} represents the probability of the bit x_{ij} taking the value 1. Therefore, v_{ij} must be constrained to the interval $[0.0, 1.0]$. A logistic transformation $S(v_{ij})$ can be used to accomplish this modification, and the position update function is defined as follows:

$$x_{ij}(t+1) = \begin{cases} 1, & rand() < \dfrac{1}{1+e^{-v_{ij}}}, \\ 0, & \text{otherwise}, \end{cases} \quad (3)$$

where $\dfrac{1}{1+e^{-v_{ij}}} \in (0,1)$ and *rand*() is a random number selected from a uniform distribution in [0.0, 1.0].

1.4 BPSO-TLC with a *P*-value filter algorithm

The proposed algorithm consists of two parts: the *P*-value filter approach and the BPSO-TLC wrapper approach.

First, the *P*-value is used to filter the sub-dataset. The *P*-value filter is a preprocessing method that is used to decrease the size of the feature space. The smaller the *P*-value of a SNP, the more relevant it is to GD. Thus, if the *P*-value of any SNP is smaller than the threshold value, the SNP is selected. Otherwise, the SNP is removed.

Next, BPSO-TLC is used to optimize the performance of the selected SNPs. Because combining the best individual SNPs may not yield the best set, BPSO-TLC is used to find the 'optimal' SNP subset. BPSO-TLC contains two main steps. The first step entails the selection of a set of SNPs by BPSO (particles in BPSO represent the features mask, where a bit with value '1' indicates that the SNP is selected and '0' indicates that it is not selected). Then, in the second step, the selected SNPs are passed to the TLC classifier to

acquire a fitness value for each particle of BPSO, while the product of the genotype and its corresponding *P*-value is the input for the TLC. The procedure for the TLC classifier is shown in Figure 1.

The classification rule is defined as follows:

$$\text{Labels} = \begin{cases} +1, & \sum_{i=1}^{n}(1-\lambda_i)\times w_i \times s_{ki} > 0, \\ -1, & \text{otherwise}, \end{cases} \quad (4)$$

where s_{ki} is the genotype (1 and 2 represent sites homozygous for the major and minor alleles, respectively, and 3 represents the heterozygous sites) of the ith SNP in the kth sample; λ_i is the *P*-value of the ith SNP; w_i is the weight value (the value of the PSO particle in bit i) of the ith SNP.

The details of the BPSO-TLC are as follows:

Algorithm BPSO-TLC:

Input:

Training samples $D = [D_{\text{case}}, D_{\text{control}}]$;

Class labels $y = [y_1, y_2, ..., y_l]$, $y_i \in \{-1, +1\}$ (−1 represents the cases and +1 represents the controls).

Initialize:

Generate *M* velocity and position vectors randomly for PSO;

Generate *N* velocity and position vectors randomly for BPSO;

Repeat until the termination criteria are met (the maximum number of iterations);

Obtain the feature subset from the BPSO and *P*-value filtering method;

Train the TLC classifier:

Calculate the value of the function (4);

Calculate the fitness value (the prediction accuracy) of PSO;

Update the PSO velocity and position vectors according to (1) and (2), respectively.

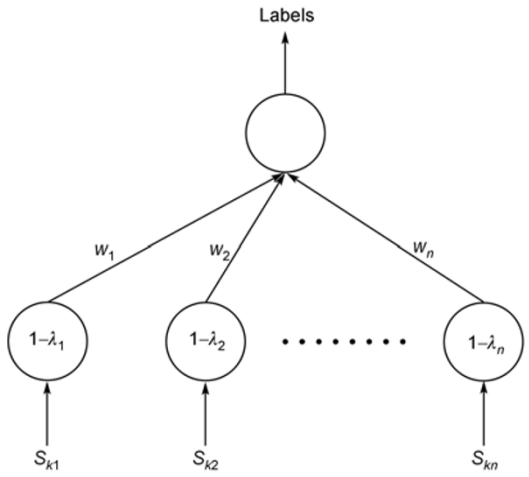

Figure 1 TLC classifier.

Test the TLC classifier:

Calculate the fitness value of BPSO;

Update BPSO velocity and position vectors according to (1) and (3), respectively.

Output:

The best BPSO particle and corresponding weight vector (PSO particle).

1.5 Cross-validation

The *k*-fold method [20] was employed in the experiments, with the value of *k* set to five. For five-fold cross-validation, the whole dataset was divided into five subsets with approximately equal size. Then, the classifier was trained five times—each time, one subset was used as the testing data to validate the classifier.

2 Results

Based on previous research [21,22], the parameters of PSO and BPSO are as follows:

$V_{\max}$		4
c_1		2
c_2		2
Number of particles	PSO	100
	BPSO	50
Maximum number of iterations	PSO	100
	BPSO	200

The evaluation criteria include sensitivity (Sn), specificity (Sp) and accuracy (Acc), which are defined as follows:

$$Sn = \frac{TP}{TP + FN} \quad (5)$$

$$Sp = \frac{TN}{TN + FP} \quad (6)$$

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

where *TN*, *TP*, *FN*, and *FP* refer to true negative (TN), number of controls that were correctly classified as control; true positive (TP), number of cases that were correctly classified as case; false negative (FN), number of cases that were wrongly classified as control; and false positive (FP), number of controls that were wrongly classified as case, respectively.

We performed our algorithm to identify the combination of SNPs most highly associated with GD (Table 1). The accuracy of classification was 70.6% with the entire set of 108 SNPs, and better results were obtained when the *P*-value filter was used. However, the SNP subset with the lowest *P*-value (P<0.001) did not yield the best result. This might be attributable to the different effects of a single SNP and a combination of SNPs. That is, the *P*-value filter method

Table 1 Performance of BPSO-TLC with a *P*-value filter[a)]

P-value range	Number of SNPs for each *P*-value range	Sn (%)	Sp (%)	Acc (%)	Number of SNPs for each combination
<1	108	66.0	78.0	70.6	19
<0.9	107	64.1	82.9	71.2	21
<0.8	105	64.4	82.1	71.1	20
<0.7	96	65.7	78.4	70.4	23
<0.6	93	65.4	78.3	70.3	16
<0.5	86	70.8	74.8	72.3	19
<0.4	81	69.4	77.5	72.4	22
<0.3	77	70.7	76.3	72.9	17
<0.2	68	70.2	76.3	72.6	18
<0.1	63	65.7	81.5	71.7	16
<0.05	58	74.9	61.9	70.2	20
<0.01	46	66.0	79.7	71.2	14
<0.005	42	68.6	71.5	69.8	15
<0.001	34	66.7	72.3	68.9	8

a) Sn=sensitivity; Sp=specificity; Acc=accuracy.

can reduce the search space from a large number of all possible combinations to a manageable one; however, the interdependence among SNPs is ignored. Figure 2 shows the classification accuracy and SNPs selected with different *P*-values.

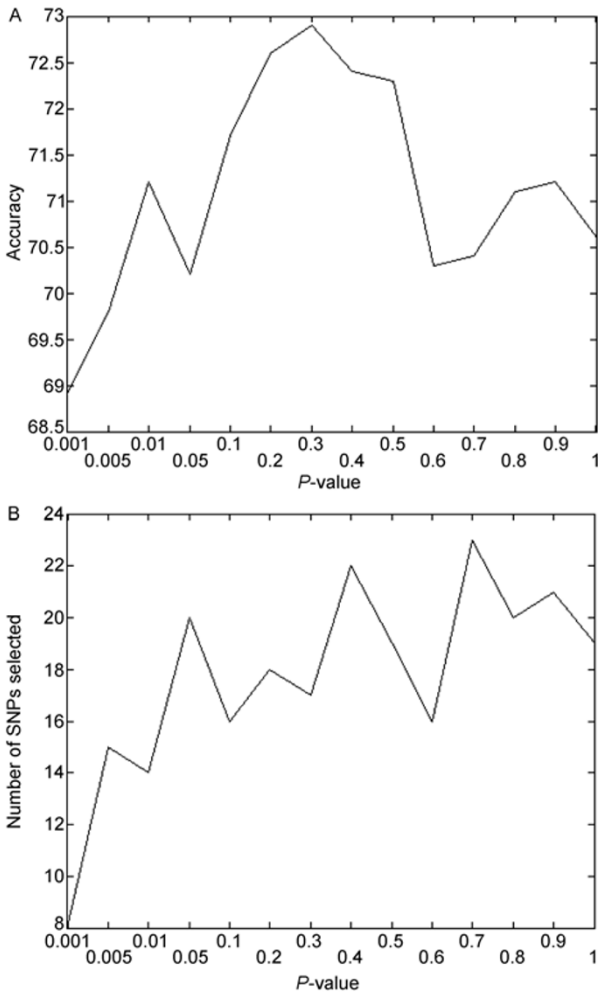

Figure 2 Classification accuracies and SNPs selected with different *P*-values. A, *P*-value range vs. accuracy. B, *P*-value range vs. number of SNPs selected.

The highest prediction accuracy obtained by our method was 72.9%, with a combination of 17 SNPs, each with $P<0.3$. The details of these SNPs are shown in Table 2. Figure 3 shows the number of iterations vs. prediction accuracy and the number of iterations vs. SNPs selected. The number of SNPs selected converges at the later stages; however, the prediction accuracy keeps improving. The reason for this is that the actual SNPs selected can be different even if the total number is the same. The prediction accuracy of our method is higher than that of some previous studies, for example, 69% in Listgarten *et al.*'s study [23] or 67.5% in Uhmn *et al.*'s study [24]. Although these previous studies used different disease samples, and it is, therefore, difficult to compare the prediction accuracy directly, the

Table 2 Details of the selected SNPs

SNP name	*P*-value	Allele	Odds ratio	Weight
CTAF322	0.002	CT	1.33	4.0251
CTAF343	0.000025	CT	1.61	–3.6333
Rs1863800	0.000045	CT	1.47	4.3863
CTAF450_1	0.00027	CT	1.41	–10.0000
CTAF450_4	0.0001	AT	1.44	–10.0000
MH30	0.000025	GC	1.49	–4.0715
MH13_1	0.011	GA	1.28	7.8697
CT55	0.00063	TC	1.39	1.9708
CT60	0.0000016	GA	1.56	–1.8483
JO37_2	0.00074	GA	1.38	–2.2434
JO36	0.00095	GA	1.39	3.6163
JO34	0.00012	GA	1.47	–1.2719
JO18	0.00072	TC	1.38	10.0000
JO13	0.00057	TC	1.42	5.4304
JO3	0.00047	CA	1.39	–4.0771
CTBC053	0.0012	TC	1.42	–5.9207
CTIC065	0.27	CT	1.13	6.3882

performance of our algorithm is acceptable.

To verify the effectiveness of our algorithm, we compared its performance with that of other current methods (Table 3). It is clear that our algorithm outperforms other methods. The accuracy obtained by our method was 72.9%, whereas C4.5, RandForest, NB and SVM obtained only 61.2%, 62.3%, 58.4% and 66.4%, respectively.

For further comparison, we used two published methods, IBPSO and HPG [21,22] on our GD dataset. The parameters used were as in [21,22], respectively. Table 4 shows the sensitivity, specificity and accuracy obtained by these algorithms. It can be seen that our algorithm yielded a higher accuracy (72.9%) than that obtained by IBPSO [21] or HPG [22] (70.5% and 72.2%, respectively).

Although a combination of 17 SNPs contributed most highly to the prediction accuracy, the importance of the individual SNPs varies. Table 2 shows the weights (importance) of those 17 SNPs. Those same values are shown in Figure 4 after normalization. CTAF450_1, CTAF450_4 and JO34 strongly correlate with GD.

The frequencies of the three genotypes of the 17 selected SNPs in the case and control datasets are shown in Figure 5. The frequencies of genotype 1 (homozygosity for the major allele) at SNPs MH13_1, CT60, JO36 and JO18 1 are much higher in the control group than in the case group. Therefore, these SNPs may exert protective effects against GD. The frequency of genotype 2 (homozygosity for the minor allele) at CTAF343 is clearly different between the two groups. The frequencies of genotype 3 (heterozygosity) at JO37_2, JO13, JO3, CTBC053 and CT_C065 are much higher in the case group than their corresponding control group frequencies. When the genotype frequencies in cases are greater than those in the controls, that combination of SNPs is regarded as a risk factor; the opposite implies a protective factor.

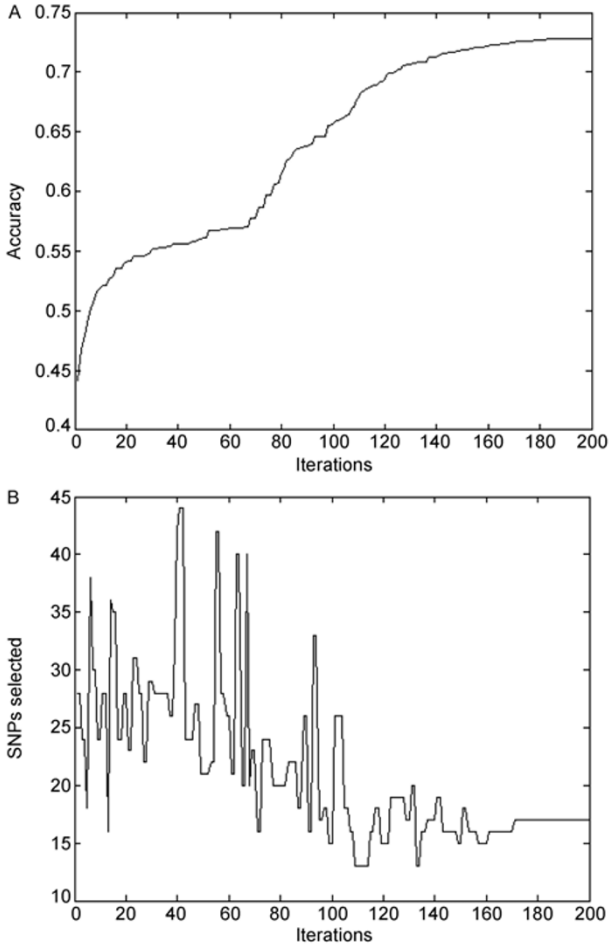

Figure 3 A, Number of iterations vs. prediction accuracy. B, Number of iterations vs. the number of SNPs selected.

Table 3 Performance of our algorithm compared with that of other classifier methods[a)]

Method	Sn (%)	Sp (%)	Acc (%)
C4.5	61.6	60.4	61.2
RandForest	68.9	50.9	62.3
NB	64.4	48.1	58.4
SVM	64.8	69.1	66.4
Our algorithm	70.7	76.3	72.9

a) Sn=sensitivity; Sp=specificity; Acc=accuracy.

3 Conclusion

The large number of SNPs makes disease association stud-

Table 4 Performance of our algorithm compared with two other published methods[a)]

Method	Sn (%)	Sp (%)	Acc (%)
IBPSO	64.0	81.8	70.5
HPG	68.1	79.1	72.2
Our algorithm	70.7	76.3	72.9

a) Sn=sensitivity; Sp=specificity; Acc=accuracy.

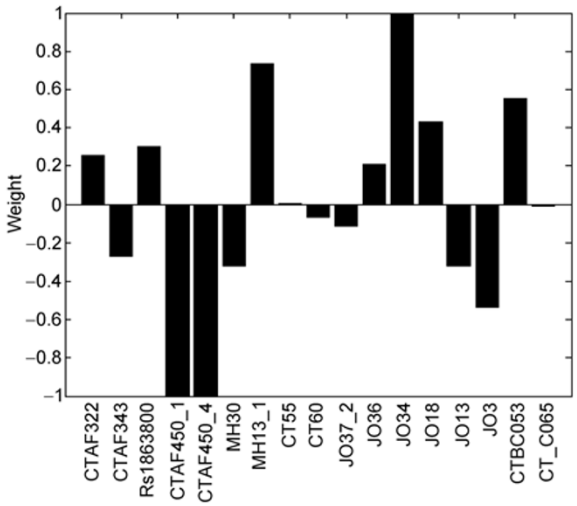

Figure 4 Weights (importance) of the 17 selected SNPs after normalization.

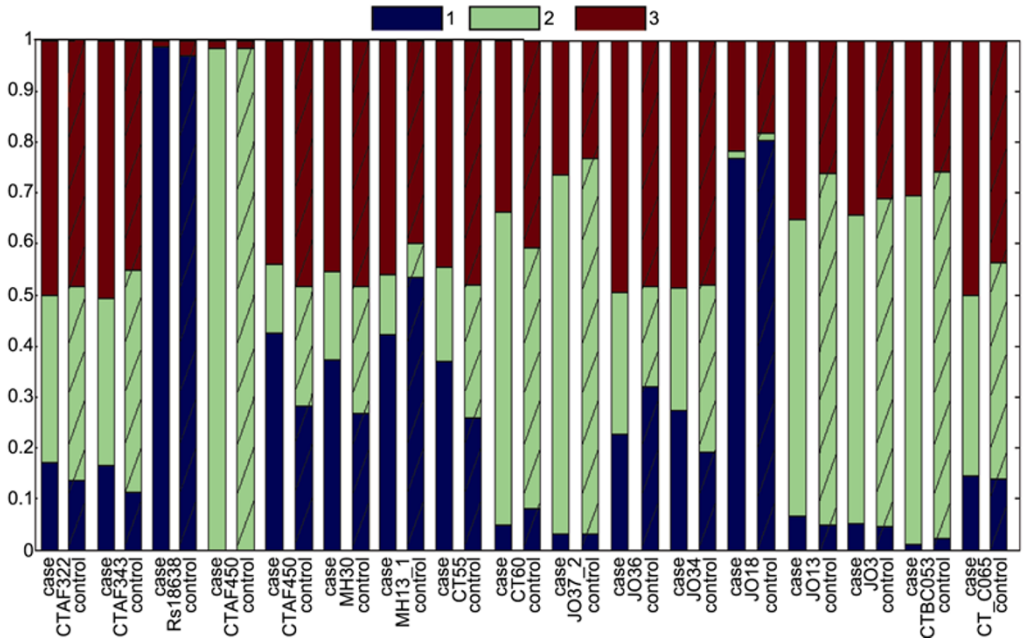

Figure 5 Frequencies of the three genotypes in the 17 selected SNPs. 1 represents homozygosity for the major allele; 2 represents homozygosity for the minor allele; 3 represents heterozygosity. The data for each SNP are shown in pairs of bars; hatched bars indicate controls and non-hatched bars indicate cases.

ies difficult to conduct. In this paper, we have proposed a wrapper algorithm with a *P*-value filter to detect multiple disease-associated SNPs. For a single SNP, it is commonly agreed that the smaller its *P*-value, the more relevant it is to disease. However, this is not always true for combinations of SNPs. Our study shows that a SNP in a combination, whose *P*-value is smaller than an optimal threshold (not a very small value), has a greater effect on GD than the single SNP with the lowest *P*-value. In addition, we have demonstrated that our method is an effective tool for the identification of combinations of SNPs in association studies of GD.

This work was supported by the National Natural Science Foundation of China (Grant No. 60774086) and the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20090201110027). We thank Dr. Dumitru Brinza for providing the datasets.

1 Hollowell J G, Staehling N W, Flanders W D, *et al.* Serum TSH, T-4, and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III). J Clin Endocrinol Metab, 2002, 87: 489–499
2 Chen X, Wu W S, Chen G L, *et al*. The effect of salt iodization for 10 years on the prevalences of endemic goiter and hyperthyroidism. J Chin Endocrinol Metab, 2000, 18: 342–344
3 Wang P W, Liu R T, Juo S H H, *et al*. Cytotoxic T lymphocyte-associated molecule-4 polymorphism and relapse of Graves' hyperthyroidism after antithyroid withdrawal. J Clin Endocrinol Metab, 2004, 89: 169–173
4 Gibbs R A, Belmont J W, Hardenbol P, *et al*. The International HapMap Project. Nature, 2003, 426: 789–796
5 Wang T H, Wang H S. A genome-wide association study primer for clinicians. Taiwan J Obstet Gynecol, 2009, 48: 89–95
6 Xavier R J, Rioux J D. Genome-wide association studies: A new window into immune-mediated diseases. Nature Rev Immunol, 2008, 8: 631–643
7 Kraft P, Hunter D J. Genetic risk prediction—are we there yet? N Engl J Med, 2009, 360: 1701–1703
8 Moffatt M F, Kabesch M, Liang L M, *et al*. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature, 2007, 448: 470–473
9 Wan X, Yang C, Yang Q, *et al*. MegaSNPHunter: A learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. BMC Bioinformatics, 2009, 10: 13
10 Hampe J, Schreiber S, Krawczak M. Entropy-based SNP selection for genetic association studies. Hum Genet, 2003, 114: 36–43
11 Carlson C S, Eberle M A, Rieder M J, *et al*. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet, 2004, 74: 106–120
12 Kusiak A, Shah S. Cancer gene search with data-mining and genetic algorithms. Comput Biol Med, 2007, 37: 251–261
13 Xie Q, Ratnasinghe L D, Hong H, *et al*. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. BMC Bioinformatics, 2005, 6: S4
14 Tomida S, Hanai T, Koma N, *et al*. Artificial neural network predictive model for allergic disease using single nucleotide polymorphisms data. J Biosci Bioeng, 2002, 93: 470–478
15 Tomita Y, Tomida S, Hasegawa Y, *et al*. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics, 2004, 5: 120
16 Dumitru B, Alexander Z. Design and validation of methods searching for risk factor in genotype case-control studies. J Comput Biol, 2008, 15: 81–90
17 Ueda H, Howson J M M, Esposito L, *et al*. Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease.

Nature, 2003, 423: 506–511
18 Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of 1995 IEEE International Conference on Neural Networks, Perth, Australia, 1995. 4: 1942–1948
19 Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Piscataway, New Jersey, 1997. 4104–4108
20 Salzberg S L. On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Min Knowl Discov, 1997, 1: 317–328
21 Yeh C L, Chang H W, Tu C J, *et al*. Improved binary PSO for feature selection using gene expression data. Comput Biol Chem, 2008, 32: 29–37
22 Li S T, Wu X X, Tan M K. Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput, 2008: 1039–1048
23 Listgarten J, Damaraju S, Poulin B, *et al*. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res, 2004, 10: 2725–2737
24 Uhmn S, Kim D H, Ko Y W, *et al*. A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. Expert Syst, 2009, 26: 60–69