Hindawi
Complexity
Volume 2017, Article ID 5024867, 10 pages
https://doi.org/10.1155/2017/5024867

Research Article

FAACOSE: A Fast Adaptive Ant Colony Optimization Algorithm for Detecting SNP Epistasis

Lin Yuan,[1] Chang-An Yuan,[2] and De-Shuang Huang[1]

[1]*Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China*
[2]*Science Computing and Intelligent Information Processing of Guang Xi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning, Guangxi 530001, China*

Correspondence should be addressed to De-Shuang Huang; dshuang@tongji.edu.cn

Received 31 March 2017; Accepted 24 July 2017; Published 7 September 2017

Academic Editor: Jianxin Wang

The epistasis is prevalent in the SNP interactions. Some of the existing methods are focused on constructing models for two SNPs. Other methods only find the SNPs in consideration of one-objective function. In this paper, we present a unified fast framework integrating adaptive ant colony optimization algorithm with multiobjective functions for detecting SNP epistasis in GWAS datasets. We compared our method with other existing methods using synthetic datasets and applied the proposed method to Late-Onset Alzheimer's Disease dataset. Our experimental results show that the proposed method outperforms other methods in epistasis detection, and the result of real dataset contributes to the research of mechanism underlying the disease.

1. Introduction

Accompanied by the rapid development of genomics and gene chip technology, Genome-Wide Association Studies (GWAS) predicted massive genetic variations related to complex traits [1, 2]. Although this method has achieved great success. It can only explain a small part of the mechanism under the complex diseases known as "missing heritability" [3]. That is to say, marginal genetic effects of GWAS identified single nucleotide polymorphisms (SNPs) account for small part of pathogenic causes. For single-locus SNPs related disease [4], GWAS can identify SNPs that are responsible for disease trait. However, complex diseases are often due to the small and complex effects of large SNPs, such as type 2 diabetes [5], prostate cancer, and rheumatoid arthritis (RA) [6]. More and more studies have shown that epistasis exists in SNPs interaction. Many SNPs will interact with each other in the process of affecting the disease traits [7]. Some SNPs will affect the disease and dominate the effect of others. The relationship of one SNP repressing the effect of another SNP is known as epistasis. In many complex human diseases, the effect of epistasis among complex human diseases is unclear. The proposed methods for SNP related disease may have poor performance due to failure to identify epistasis.

During the past decade, a lot of approaches have been proposed to detect epistasis. Some methods focus on the interaction between two certain SNPs. Zhang et al. [8] proposed a Bayesian partition method for epistatic eQTL modules. Kang et al. [9] proposed four different models to measure epistasis effect between two loci and suggest a statistical strategy to infer the hierarchical relationships. Recently, Lin et al. [10] reported forty-five SNP-SNP interaction models by considering the inheritance modes and model structures. Though these methods have been successful in studying epistasis between two SNPs. The GWAS data is high dimension data which contains hundreds of thousands or even million SNPs; at the same time, GWAS data only contains dozens or hundreds of individual sample data, for example, the small number sample data and the high dimension features; it needs vast amounts of time to identify the interaction between each pair of SNPs [11–13]. The computational burden is out of bounds.

More and more machine learning methods are applied to research epistasis. Many methods were proposed to model

epistasis effect from the perspective of the overall data. Moore et al. [14] applied regression method to identify the relationship between gene expression and epistasis effect. Michael et al. [15] applied Bayesian networks to identify the epistasis effect network from the original SNPs data. Although these methods solved some problems, they still did not show significant effects with the large scale Genome-Wide Association Study datasets owing to the same "high-dimensional small sample size problem." With the rapid development of multiobjective optimization method and machine learning discipline, ant colony optimization (ACO) algorithm was applied to epistasis research. Wang et al. [16] proposed AntEpiSeeker; AntEpiSeeker combines heuristic search with the ant colony optimization to identify SNPs which dominate other SNPs. Experimental results on real rheumatoid arthritis dataset show that AntEpiSeeker is better than other methods. The drawback of this method is that other methods show different performance on different disease models. Zhang and Liu [17] developed the Bayesian inference method which identifies the epistatic interactions in case-control studies. However, the BEAM method needs a lot of time in GWAS dataset. In this paper we extend SNP epistasis study to a fast adaptive ant colony optimization algorithm for detecting SNP epistasis. We search SNP epistasis with two-objective functions and fast adaptive ant colony optimization.

The experiments on several simulated datasets show the good performance of our method. We also compare our method with the benchmark methods, including BEAM, generic ACO, and AntEpiSeeker. Experimental results show that our method has better performance in GWAS datasets containing epistasis effect among SNPs.

2. Methods

2.1. Ant Colony Optimization. In the research of artificial intelligence and large scale problem solving, the ant colony optimization (ACO) algorithm is inspired by the ants food search behaviour in nature. Assume that the food search paths constitute a graph; the ant colony optimization algorithm can reduce time of search paths through graphs [18]. This algorithm with other ant colony optimization algorithms is kind of swarm intelligence methods, and it is member of metaheuristic optimizations. Marco Dorigo proposed the ant colony optimization algorithm in 1992 in his Ph.D. thesis. In the GWAS datasets, the datasets often contain tens of hundreds to millions of SNPs. It is not feasible to identify the relationship of every pair of SNPs within an acceptable time. ACO algorithm was used here to reduce the complexity of exhaustive search. In kingdom of insects, in the process of finding food, ants look like they are walking randomly, and in the back and forth path of searching for food, the ants will leave pheromones on the path. If the path is found by other ants, other ants tend to follow the path but not walk randomly; going further, if they find food through this path, they will also leave pheromones; the pheromone value on this path is enhanced. Subject to other factors in nature, pheromone value starts to evaporate and the path's attractive strength starts to decrease. The longer the path is, the more the time the ants are looking for food. As a comparison, the time the ants take to walk through the short path is greatly shortened, and pheromone values will be larger on shorter paths than longer paths. Pheromone evaporation results in dynamic changes in the path. Path dynamic changes can avoid the convergence of solutions to a locally optimal solution. If there is no pheromone values evaporation, the food search path selected by first ants would tend to be the only path or the most attractive path. This phenomenon will lead to limitation of the solution space. The mechanism of pheromone evaporation in ant colony is unclear, but pheromone evaporation is a very important application in artificial intelligence systems. Though the ant colony optimization algorithm has achieved great success in application [19–21].

The travelling salesperson problem (TSP) is a problem with some cities and physical distances between each pair of cities. The question is what is the shortest possible path where travelling salesperson visits each city once and finally returns to the origin city? Suppose there are n cities; there are $(n-1)!/2$ solutions to the problem. The feasible solutions will increase exponentially when the number of city increases, making the computation impractical. Obviously, it is an NP-hard problem of combinatorial optimizations.

Suppose that m ants are randomly placed in n cities, the kth ant in the ith city; the probability if ant chooses the next city j is

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}(t)\,\eta_{ij}^{\beta}(t)}{\sum_{o \in \text{candidate}_k} \tau_{io}^{\alpha}(t)\,\eta_{io}^{\beta}(t)}, & j \in \text{candidate}_k, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

$$\eta_{ij}(t) = \frac{1}{d_{ij}},$$

where $\tau_{ij}(t)$ indicates the surplus information on path ij in moment t. $\eta_{ij}(t)$ indicates the heuristic function. d_{ij} indicates the physical distance between city i and city j. tabu_k indicates the cities set which indicates ant k has visited. candidate_k indicates the set of cities which ant k can visit next.

Over time, after n moments, the ants complete a cycle; the information of each path should be adjusted according to

$$\begin{aligned} \tau_{ij}(t+n) &= (1-\rho)\,\tau_{ij}(t) + \Delta\tau_{ij}, \\ \Delta\tau_{ij} &= \sum_{k=1}^{m} \Delta\tau_{ij}^{k}, \end{aligned} \tag{2}$$

where $\Delta\tau_{ij}$ indicates information increment of path ij after this cycle.

$$\Delta\tau_{ij} = \begin{cases} \dfrac{Q}{l_k}, & ij \in L_k \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where L_k indicates ant k's paths in this cycle. l_k indicates the path length of ant k in this cycle. The parameters needed to be determined are $\alpha, \beta, \rho, m, Q$; the number of ants is less than or equal to city number; Q is a large suitable

number. ACO is always used in large scale data problems. However, slowness is still a bottleneck in the application of the ant colony algorithm for large scale search optimization problems. Pheromone update strategy is one of the keys to determine the convergence rate.

In the process of applying ant colony optimization to specific problems, the search space should be as large as possible. At the same time, ACO should consider time efficiency. ACO should balance the optimal solutions and solve speed. On the basis of previous studies [22–24]. We only consider pheromone evaporation factor ρ and pheromone importance factor α. In (2), ρ is used to balance the effects of old pheromone value and current pheromone value. When ρ is too small, the residual pheromone value is too much and leads to local minimum solution. We adopt adaptive ρ, when the algorithm does not improve the current optimal solution within n iterations.

$$\rho(t+n) = \begin{cases} g\rho(t), & \rho(t) \le \rho_{\max} \\ \rho_{\max}, & \text{otherwise,} \end{cases} \tag{4}$$

where $\rho_{\max}$ equals 0.85 in practice. g equals 1.02 as tune parameter. When the pheromone value reaches the critical value, the pheromone importance factor begins to play a role. With the increase of pheromone importance factor α, the algorithm will jump out of the local optimal solution and has ability to search for global optimal solution.

$$\alpha(t+n) = \begin{cases} g_1\alpha(t), & \alpha(t) \le \alpha_{\max} \\ \alpha_{\max}, & \text{otherwise,} \end{cases} \tag{5}$$

where g_1 is a constant larger than one and $\alpha_{\max}$ is less than or equal to five. In the process of calculation, first, we follow the standard ant colony optimization algorithm for N iterations. N is predefined number. If the current optimal solution is not improved after N iterations, update the parameters according to formulas (4) and (5). Then update all pheromone value according to (2).

Given pheromone values and transfer rules, we can use the ant colony optimization algorithm to find a group of SNPs which affect the disease. Assume there are P SNPs in the global Genome-Wide Association Studies dataset, we can construct a p-dimensional symmetric matrix M to store every ant's pheromone value. The element m_{ij} of matrix M denotes the interaction which is related to disease between ith SNP and jth SNP. At the beginning of our method, every element of matrix M is assigned to a constant value m_0; equivalent value shows the epistasis in every pair of SNPs and there is equal possibility relationship between the SNPs and disease.

At the final pheromone iteration, the ACO algorithm will obtain the optimal solutions through forward selection strategy. The advantage of ACO algorithm in this paper is that the result contains nondominated solutions which have the potentially equivalent possibility and potentially highest related strength with disease and omit dominated solutions.

The disadvantages of traditional ant colony optimization algorithm are long search time and tendency to fall into the local optimal solution. The drawback of this working mode is that the current pheromone evaporation factor and pheromone importance factor are predefined. As an improved strategy, we extended the "dynamic adaptive strategy" to ant colony optimization. The advantage of this strategy is the fast convergence rate and searching for global optimization solution. Compared with traditional ACO, the new strategy can provide more accurate result.

2.2. Two-Objective Function Optimization. The results of ant colony optimization need to be evaluated. We combine two-objective methods to assess the final epistasis results. In general, one of two-objective functions combines Akaike Information Criterion (AIC) score and logistic regression function to measure relationship between phenotypic trait and genotype data; Akaike Information Criterion indicates the effectiveness and complexity of the model [25, 26]. In our method, on the basis of the standard logistic regression, following the North et al. [27] strategy, we use ADDINT logistic regression model to search the relationship between disease and SNP nodes. The second objective function uses frequency measurement based on mutual information theory to model the relationship between genotype data and phenotypic trait from the perspective of information theory. The second objective function used to represent the selected SNP subsets can explain how much information is about the disease trait. Our proposed method obtains information from data rather than a lot of priori information. The above two-objective functions are designed from the different perspective to measure the quality of the search results, and the simulation data experiment results show that our two-objective functions have a better performance than other methods on simulated and real biological datasets.

In order to avoid the bad impact of high dimension small size sample problem, the identification of disease-associated SNPs is known as a heuristic optimization problem. In our proposed method, proposed method yields optimal solutions which is nondominated solutions; the proposed two-objective functions method actually is kind of multiobjective optimization; the proposed method uses ant colony optimization to search for optimal solution [28].

Our proposed fast adaptive ACO framework contains two stages. In the first stage, we use modified ACO optimization algorithm with two-objective functions to search for nondominated SNP subset. After generating the nondominated SNP subset, we apply Fisher exact test [29, 30] to the dataset containing nondominated SNP generated in the algorithm first stage. The Fisher exact test will be used to identify the relationship between disease and SNPs.

2.2.1. AIC Score. The Akaike Information Criterion (AIC) is used to measure quality of dataset statistical models. AIC is from information theory, and it estimates loss of information when a statistical model is used to express the data generation process. The mechanism of Akaike Information Criterion is that it deals with the trade-off between the goodness of fit of the model and the complexity of the model. Based on the nature of the AIC, we construct AIC model from the perspective of GWAS dataset. The goal of our method is to measure the relationship between the genotype data of

genome and phenotype disease trait. Logistic regression is widely used to quantitatively analyze the correlation between dependent variable and independent variable. Based on above methods, we construct AIC score model containing logistic regression and gradient penalty function. Logistic regression can compute the maximized log-likelihood of the model; k is used to express the number of free parameters. AIC score deals with the trade-off between the fitness effect of the model and the complexity of the model. We follow Jing and Shen [28] strategy:

$$\text{AICscore} = 2k - 2\log lik, \tag{6}$$

where k denotes the number of free parameters.

2.2.2. Explanation Score. In GWAS research, the relationship between two loci and disease, in SNP research, each locus has three values, 0, 1, and 2; 0 means major allele homozygous, 1 means heterozygote, and 2 means minor allele homozygous [31]. For two loci, there are nine cases of their combination; the disease related SNP locus often changes when the disease occurs. In the case of double locus combination, x_i means the number of ith combinations of two SNP loci, Y means case or control state, y_1 means state case, and y_2 means state control. The potential interrelationships of two discrete random variables X and Y are defined as $H(X;Y)$; the relationship between locus combination and disease is measured based on the information of locus frequency. $H(X;Y)$ is described as below:

$$H(X;Y) = \sum_{i=1}^{I} \left(\left| x_{iy_1} - x_{iy_2} \right| \right), \tag{7}$$

where I means the total number of locus combinations. To avoid unbalanced sample, the size affects score. For example, if data size of case is larger than control, we extract the same size of control data from case samples randomly. To avoid the impact of randomness, we extract sample several times and average the results. The large value H means the potential association probability between disease and SNPs is large. Equation can also be applied to more than two locus combinations. We name this score explain score.

2.3. Pareto Optimality for SNP Epistasis Detection. Pareto optimality defines such a situation. Pareto optimality is proposed to solve the following questions where it is impossible to make all objective function values of multiobjective optimization optimal values [32, 33]. Pareto optimality is first applied to the area of income distribution and economy. Now Pareto optimality has been extended to engineering and multiobjective optimization research. On the basis of previous proposed methods, the modified ant colony optimization algorithm with first objective function and second objective function, the first objective function is AIC score with logistic regression and related parameters; the second objective function is explain score. For the first objective functions, the lower score of the objective function indicates the high potential relationship between disease phenotype trait and SNPs [34]. For the second objective functions, the higher score of the objective function indicates the high potential relationship between the disease phenotype trait and SNPs. The target of fast two-stage ant colony optimization algorithm is to find the epistasis effect among SNPs and extract real SNP subset with respect to the above proposed methods.

In the real GWAS datasets, an identified SNP subset may perform the best compared with other method solutions in terms of one-objective function, but SNP subset may perform poorly in terms of another objective function. Thus, the target is how to select better SNP subset with respect to both objective functions. In practical application, rare subset performs better than other solutions while satisfying both conditions. Thus, for a framework with two-objective functions, it is hard and even impossible to calculate the global optimal solution. On the basis of previous studies [28, 34, 35], we adopt Pareto optimality to find the practical optimal solution. We first compare the two solutions, in terms of GWAS SNP subset, a solution named S_1, and another solution named S_2; comparing S_1 and S_2 only have two consequences; one result is one solution dominates the other; another result is S_1 does not dominate S_2; in turn, the solution S_2 does not dominate S_1. Based on the mind of Pareto optimality, we consider S_1 dominates S_2 if they satisfy the following two conditions. The first condition is the value of $f_e(S_1)$ is not higher than $f_e(S_2)$ for those two-objective functions. The second condition is the objective function $f_e(S_1)$ is lower than $f_e(S_2)$ for at least one-objective function. The function f_e denotes the objective function: modified AIC score objective function and explain score objective function. The e equal to one denotes the first objective function; the e equal to two denotes the second objective function. If solutions S_1 and S_2 satisfy the above two conditions, we say solution S_1 is a nondominated solution; in turn, we say solution S_2 is a dominated solution. Based on above Pareto optimality approach and two-objective functions, all solutions can be divided into two kinds; one is nondominated set and another is dominated set. Finally, nondominated sets contain many solutions and all the solutions from our proposed method with respect to two-objective functions; now our goal is to find a nondominated set which is the best under certain conditions.

Next, we will use the judgment rule mentioned earlier to sort the solutions of nondominated sets to find the optimal nondominated set. Specifically, in the first case, $f_1(S_2)$ is larger than $f_1(S_1)$; at the same time, $f_e(S_2)$ is larger than $f_2(S_1)$. In the second case, $f_1(S_2)$ equals $f_1(S_1)$; at the same time, $f_2(S_2)$ is larger than $f_2(S_1)$. In the third case, $f_1(S_2)$ is larger than $f_1(S_1)$; at the same time, $f_2(S_2)$ equals $f_2(S_1)$.

2.4. Fisher Exact Test for Experimental Results. Fisher exact test is used in contingency tables to get a statistical significance [36–38]. Although in practice it is used in small size sample, it is can also be used in all sample sizes. Ronald Fisher first proposed this method and Fisher exact test is one kind of exact tests.

In terms of our GWAS datasets research article, on the basis of unified framework which contains fast adaptive ant colony optimization (ACO) algorithm, Akaike Information Criterion (AIC) score, explain score, and Pareto optimality, we can obtain the final result which is a nondominated SNP set; in this section, we will use Fisher exact test to exhaustively

search for the epistasis effect. Fisher exact test is based on hypergeometric distribution; the P value in the Fisher exact test is accurate for all individual samples. Fisher exact test is used on the basis of contingency table. The null hypothesis is that the identified SNP subset and disease are not associated. The alternative hypothesis is that SNP subset affects the expression of the disease when the Fisher exact test's P value is significant, when P value is less than predetermined value such as 0.05 or smaller value. Our proposed method will identify significance SNP subsets.

2.5. Power Test. In previous section, we introduce each part of our proposed fast adaptive ant colony optimization algorithm for detecting SNP epistasis. Our proposed unified framework contains fast adaptive ant colony optimization algorithm, Akaike Information Criterion (AIC) score, explain score, Pareto optimality, and modified Fisher exact test. In this section, we introduce how to verify the significance of the results. We construct 100 datasets according to the same parameters. Then we use the traditional power test to measure the effect of methods. The power test is defined as follows:

$$\text{Power} = \frac{|SD|}{100}, \tag{8}$$

where $|SD|$ denotes the number of disease related datasets which were correctly selected from 100 datasets. Only using the single test criterion may not clearly show the quality of results. We use precision recall standard to measure true positive rate and false positive rate. Precision recall criteria have been widely used in classification model evaluation model [39, 40]. In pattern recognition and information retrieval with binary classification, precision, also called positive predictive value, is the fraction of retrieved instances that are relevant; while recall, also known as sensitivity, is the fraction of relevant instances that are retrieved [26]. Both precision and recall are therefore based on an understanding and measure of relevance. We use precision recall criteria to determine whether the classification results are good or bad. The precision recall criteria can avoid the imbalance problem of precision recall numbers. In our research, the number of precision and recall always differs greatly. In terms of the SNP epistasis research, precision is also known as positive predictive value, equivalent to the true disease related SNP subsets; recall is also known as sensitivity or negative, equivalent to the true disease unrelated SNP subsets. If we use only one judgment criterion, thus false positive rate, single indicator cannot make the real result clear. We use false positive rate and true positive rate to measure the real result. This is why we use precision and recall. We also use F_1 score (also F score or F measure) to measure the precision recall test accuracy. The precision and recall will be introduced next with confusion matrix (Figure 1).

$$\begin{aligned} \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ F_1 &= \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{9}$$

		Predicted class: Associated	Predicted class: Nonassociated
True class	Associated	True positive (TP)	False negative (FN)
True class	Nonassociated	False positive (FP)	True negative (TN)

Figure 1: Precision recall explanation matrix.

The precision, also known as specificity, denotes true positive number ratio in the result through the number of true positives divided by the sum of true positive number and false positive number; precision is often used to report false positive rate of an algorithm's false positive rate. The recall, also known as sensitivity, denotes true positive ration in the sum of true positives and false negative. In terms of SNPs selection problem, the larger the recall value is, the larger the number of real true disease-related SNP combinations can be found. Simultaneously, the larger the precision value, the larger the number of real true disease-related SNP combinations account for a high proportion of the identified SNP combinations. The criterion F measure is the harmonic mean of precision and recall, which is a synthesized measure combining both precision and recall [41].

3. Simulation Experiments

3.1. Compared with One-Objective Function. In this section, we use simulation data to compare our proposed method with other existing methods. In order to avoid data favor caused by the model, we adopt BEAM package to generate simulation datasets [17]. Data was simulated following three genetic models: (1) additive model, (2) epistatic interactions with multiplicative effects, and (3) epistatic interactions with threshold effects. In order to introduce our experiments, the additive model is referred to as ADDME. The model about epistatic interactions with multiplicative effects is referred to as EIME. The epistatic interactions with threshold effects are referred to as EITEME. In the next section, we will use the short name to indicate the corresponding data model.

Because our method is two-objective-based SNP epistasis search method, first, we compared our proposed method with existing single objective-based exhaustive SNP epistasis search method to demonstrate the effectiveness of two-objective function SNP epistasis subset search method. Second, we compare our proposed method with recently proposed method BEAM [17], generic ACO algorithm, and AntEpiSeeker [16]. In the one-objective function SNP epistasis search method, the objective function is used to score every SNP combinations; in general, the score for every SNP combination is not the same. Based on the nature of the method, low score indicates the association between SNP combination and disease is relatively small; high score indicates the association between SNP combination and disease is relatively large. Then the one-objective function ranks all SNP combinations based on the scores. However, the two-objective-based SNP epistasis search method is to find a set of nondominated results, and every nondominated SNP

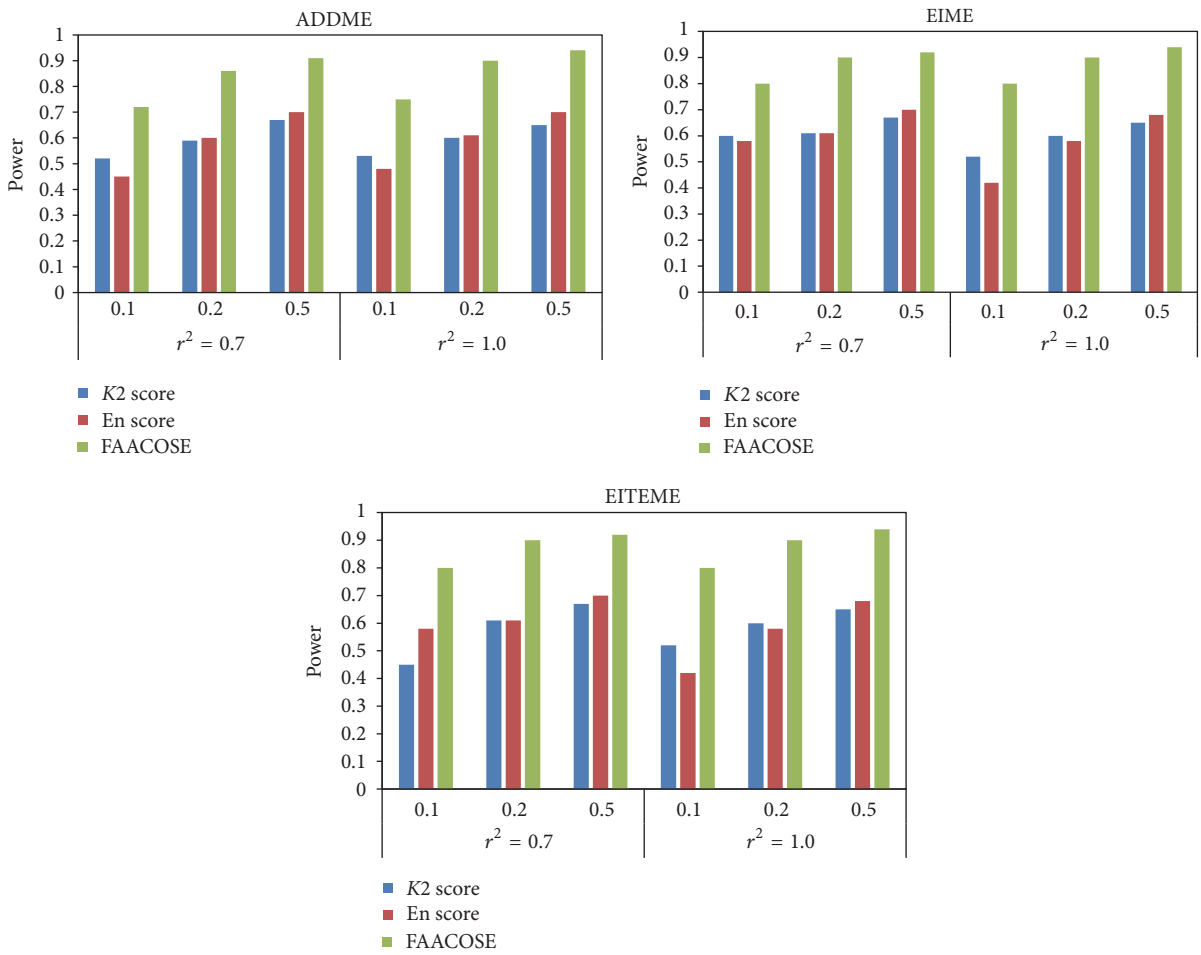

Figure 2: Power test comparisons between one-objective and two-objective methods on three different model with MAF value 0.1, 0.2, and 0.5.

epistasis results' score is the same. To ensure fairness, for the one-objective function, we collect the same number as two-objective-based SNP epistasis search method from the top of one-objective-based SNP rank. The comparing results show that the two-objective-based SNP epistasis search method is better than one-objective-based SNP epistasis search method in three simulation data models. In terms of two single objective-based SNP epistasis search methods, the results of one-objective-based SNP epistasis search methods are similar with the other one-objective-based SNP epistasis search methods. The simulation data experiment results show the effectiveness of two-objective-based SNP epistasis search method, and the poor experimental results show the insufficiency of one-objective functions. The experiment results are shown in Figure 2. The abscissa of Figure 2 is minor allele frequency (MAF) which is assigned 0.1, 0.2, and 0.5. We generate the simulate dataset and study the parameter setting following many previous studies [17, 42–44]. For each simulate dataset of parameter combination, we generated 100 datasets which contain 2,000 experimental samples (1,000 case samples and 1,000 control samples) and 1000 SNPs were simulated. We evaluate the algorithm performance through calculating the ratio of real number identified following the significance level 0.01 which is adjusted after Bonferroni correction. The parameter λ was set to 0.3 for ADDME and 0.2 for EIME and EITEME. The parameter range of linkage disequilibrium between SNPs is r^2 from 0.7 to 1.

3.2. Compared with Benchmark Methods. After comparing with single objective function. We compare our proposed method with existing method. The performance of our proposed method was evaluated by comparison with benchmark methods [45]. In many previous studies, the authors have already discussed the parameter settings problem. In this section, we set the parameters according to the existing strategy. We evaluated performance of FAACOSE by comparing with two recent methods, BEAM, generic ACO algorithm, and the AntEpiSeeker; we use BEAM package and previous parameter strategy to generate simulate dataset. Be aware of the fact that the generic ACO algorithm could not select larger size SNP set. We use simulated dataset introduced in Section 3.1. We evaluate the algorithm performance through calculating the ratio of real number identified following the significance level 0.01 which is adjusted after Bonferroni correction. We generate simulate datasets following three genetic models: ADDME, EIME, and EITEME. Other parameters

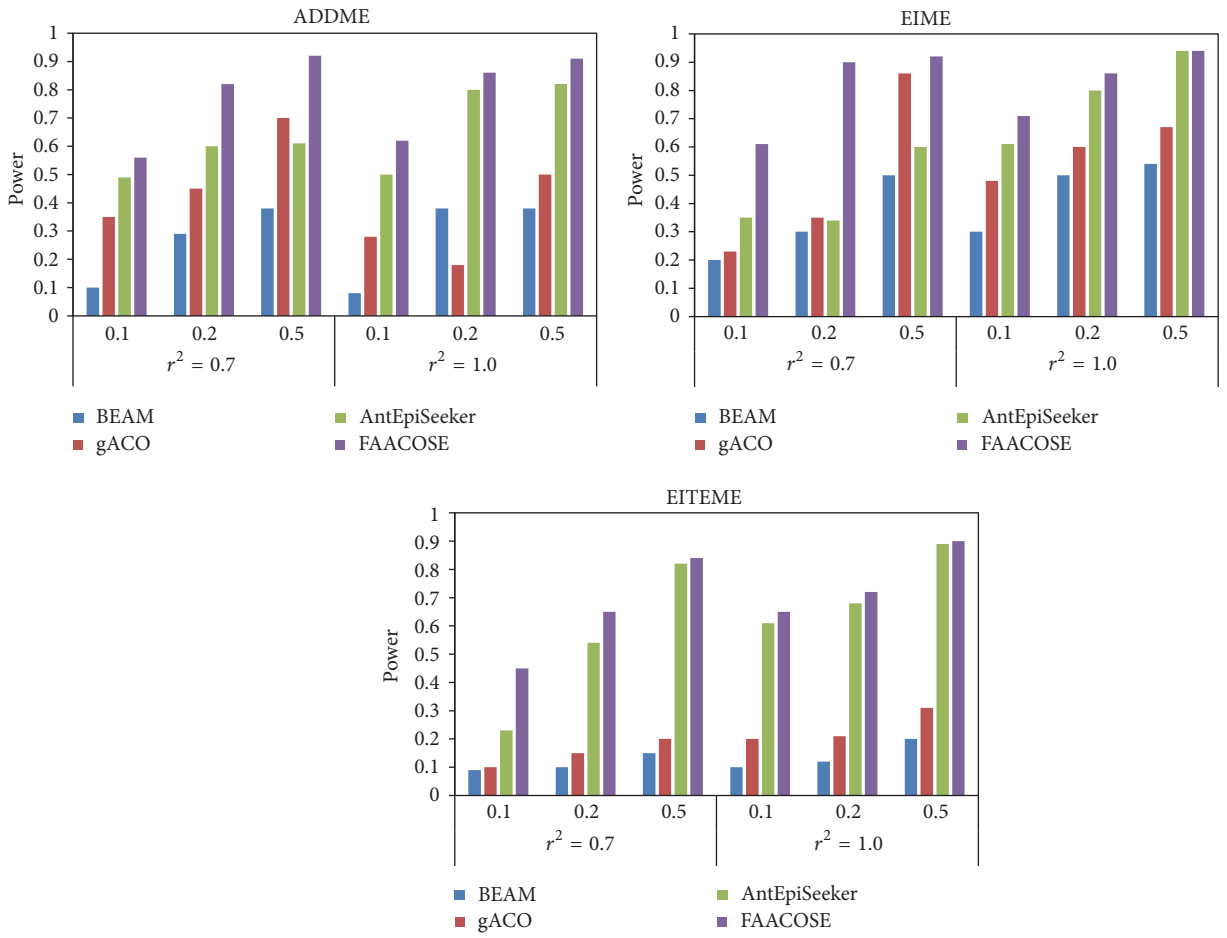

Figure 3: Power comparisons between existing methods and FAACOSE on three models.

for data simulation were the effective size λ, a measure of marginal effects as defined by Marchini et al. [42], linkage disequilibrium between SNPs measured by r^2, and minor allele frequencies (MAFs). λ was set to 0.3 for ADDME and 0.2 for EIME and EITEME. For r^2, two values (0.7 and 1.0) were used for each model. For MAFs, three values (0.1, 0.2, and 0.5) were considered. The parameters for BEAM were set as default. The parameter settings for AntEpiSeeker were large dataset size = 6, small dataset size = 3, count large = 150, count small = 300, epistasis model = 2, ant count = 1000, $\alpha = 1$, $\rho = 0.05$, and $\tau 0 = 100$ (also available in the software package documentation of AntEpiSeeker). The parameters of the generic ACO algorithm were set as ant count = 1000, $\alpha = 1$, $\rho = 0.05$, $\tau_0 = 100$, count (number of iterations) = 900, and epistasis model = 2. The comparison of detection power for BEAM, genetic ACO algorithm, and the AntEpiSeeker is presented in Figure 3. The results show that FAACOSE outperforms BEAM and the generic ACO in all parameter settings and is superior to AntEpiSeeker in most parameter settings.

In this section, we compare our proposed method with benchmark methods. First, we use power test to detect how many real SNP subsets can be found with our proposed method. Second, we use precision, recall, and F_1 score to evaluate the results. Precision denotes how many right SNP subsets in the total final identified SNP subsets. Recall denotes the number of right SNP subsets that are identified. F_1 score is an indicator used in statistics to measure the accuracy of two classification models. It takes into account the precision and recall of the classification model simultaneously. F_1 score can be seen as a weighted average of precision and recall, its maximum is 1, and minimum is 0.l. We show the results of FAACOSE with other methods on $r^2 = 0.7$ and MAF = 0.2 in Table 1.

The F_1 score of FAACOSE is better than other methods. We run the same experiment on datasets with different parameter combination. In all eighteen datasets FAACOSE has the highest F_1 score in fifteen of them. In real GWAS dataset experiment, the sample size of real dataset is huge. The efficiency of the method is also to be considered. The experimental results indicate that our proposed method is more effective method in real GWAS dataset. AntEpiSeeker is the most efficient algorithm among three methods. In different data samples, we compare run time of AntEpiSeeker and FAACOSE. And averaging the results, FAACOSE is faster 30% than AntEpiSeeker.

Table 1: F_1 score comparison between FAACOSE and other methods.

Model	Method	Recall	Precision	F_1 score
ADDME	BEAM	0.29	0.15	0.20
	gACO	0.45	0.36	0.40
	AntEpiSeeker	0.6	0.55	0.57
	FAACOSE	0.82	0.74	0.78
EIME	BEAM	0.3	0.45	0.36
	gACO	0.35	0.32	0.33
	AntEpiSeeker	0.34	0.56	0.42
	FAACOSE	0.9	0.82	0.86
EITEME	BEAM	0.1	0.14	0.12
	gACO	0.15	0.20	0.17
	AntEpiSeeker	0.54	0.46	0.50
	FAACOSE	0.65	0.62	0.63

4. Application to Real SNP Dataset

Late-Onset Alzheimer's Disease (LOAD) is the most frequent form of Alzheimer's disease, which is frequently identified in people older than 65 years; the LOAD or AD is a kind of chronic neurodegenerative diseases which is frequently not obvious in the onset of the disease and slowly changes dementia over time. It is the cause of 60% to 70% of cases of dementia. The most common early symptom is difficulty in remembering recent events (short-term memory loss). As the disease advances, symptoms can include problems with language, disorientation (including easily getting lost), mood swings, loss of motivation, not managing self-care, and behavioural issues. LOAD is a multifactor genetic disease; its etiology and pathogenesis have not yet been fully understood. The apolipoprotein (APOE) gene is a definite risk factor for LOAD. The APOE gene has three forms. The $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$; the effect of $\varepsilon 2$ is positive; $\varepsilon 2$ can effectively prevent the occurrence of the disease. There has been research report that genetic variant $\varepsilon 4$ has induced effect on disease. Between 40 and 80% of people with AD possess at least one APOE $\varepsilon 4$ allele [46]. Previous studies have reported some significant SNPs in the field of Genome-Wide Association Studies [47]. Reference [47] reported that 10 SNPs in the area of GAB2 gene have an epistasis effect with APOE e4 in relation to Late-Onset Alzheimer's Disease. We applied our proposed method to the LOAD GWAS dataset from website https://www.tgen.org/ [47]. After data preprocessing, the real biological dataset contains 1368 samples [48, 49]. Of these, 836 samples were identified case studies; the remaining 532 samples were normal sample [50, 51]. Each sample of real biological dataset contains 309,316 SNPs with genotype information, APOE status, and LOAD status [52]. For the next calculation, we code the APOE gene state with a binary variable; the value 1 represents the $\varepsilon 4$ variant and in turn the value 0 represents the other three variants [53]. An SNP locus was coded as a quaternary variable considering the missing state. The high potential LOAD disease related SNP is shown in Table 2.

Table 2: The number of selected SNPs of FAACOSE in LOAD dataset.

SNP rs#			
rs7756992	rs611154	rs191840	rs7294919
rs1887922	rs304900	rs1999764	rs1385600
rs2373115	rs7101429	rs609812	rs613375
rs1007837	rs2510038	rs4945261	rs10793294
rs520227	rs191740	rs7924284	rs829465
rs602106	rs7174511	rs606889	rs602192

5. Discussions

In this paper, we proposed a novel ant colony optimization based fast search method for the discovery of epistasis interactions in large scale real GWAS dataset. FAACOSE was evaluated through comparison with existing three approaches on both simulated and real datasets. FAACOSE, which adopts a fast adaptive optimization procedure, is a modified algorithm derived from the generic ACO. And with two-objective function, to demonstrate the advantages of fast adaptive ant colony optimization algorithm, we also compared the performance of the FAACOSE with that of the generic ACO.

In future studies, we intend to find more powerful modeling approaches, ant colony optimization algorithm with faster convergence, objective functions which can better measure data structure of GWAS dataset, more efficient optimal SNP subset search, and identification strategies that can be combined and flexibly embedded into our SNP epistasis search framework to find more accurate SNP subset. With the rapid development of bioinformatics, more and more biological information related to disease is identified. More and more studies will consider prior knowledge. An important future research direction is that we will try to apply expert prior knowledge to GWAS dataset with our proposed method, that is, the fast adaptive ant colony optimization algorithm for detecting SNP epistasis. Expert prior knowledge can improve the power and efficiency of epistasis detection.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is partly supported by National Natural Science Foundation of China (Grant nos. 61520106006, 31571364, 61732012, 61532008, U1611265, 61672382, 61402334, 61472280, 61472173, 61572447, 61672203, 61472282, and 61373098) and China Postdoctoral Science Foundation (Grant nos. 2014M561513, 2015M580352, 2017M611619, and 2016M601646) Guangxi Bagui Scholars Program Special Fund.

References

[1] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[2] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genetics*, vol. 5, no. 6, Article ID e1000529, 2009.

[3] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[4] B. S. Shastry, "SNP alleles in human disease and evolution," *Journal of Human Genetics*, vol. 47, no. 11, pp. 561–566, 2002.

[5] B. Stubbs, D. Vancampfort, M. De Hert, and A. J. Mitchell, "The prevalence and predictors of type two diabetes mellitus in people with schizophrenia: a systematic review and comparative meta-analysis," *Acta Psychiatrica Scandinavica*, vol. 132, no. 2, pp. 144–157, 2015.

[6] K. P. Liao, "Cardiovascular disease in patients with rheumatoid arthritis," *Trends in Cardiovascular Medicine*, vol. 27, no. 2, pp. 136–140, 2017.

[7] Y. Mao, N. R. London, L. Ma, D. Dvorkin, and Y. Da, "Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model," *Physiological Genomics*, vol. 28, no. 1, pp. 46–52, 2006.

[8] W. Zhang, J. Zhu, E. E. Schadt, and J. S. Liu, "A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules," *PLoS Computational Biology*, vol. 6, no. 1, Article ID e1000642, 2010.

[9] M. Kang, C. Zhang, H.-W. Chun, C. Ding, C. Liu, and J. Gao, "EQTL epistasis: Detecting epistatic effects and inferring hierarchical relationships of genes in biological pathways," *Bioinformatics*, vol. 31, no. 5, pp. 656–664, 2015.

[10] H. Lin, D. Chen, P. Huang et al., "SNP interaction pattern identifier (SIPI): an intensive search for SNP–SNP interaction patterns," *Bioinformatics*, 2016.

[11] R. L. Prentice and L. Qi, "Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation," *Biostatistics*, vol. 7, no. 3, pp. 339–354, 2006.

[12] S.-P. Deng, L. Zhu, and D.-S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, vol. 16, no. 3, article no. S4, 2015.

[13] S.-P. Deng and D.-S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207–212, 2014.

[14] J. H. Moore, J. M. Lamb, N. J. Brown, and D. E. Vaughan, "A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 Levels," *Clinical Genetics*, vol. 62, no. 1, pp. 74–79, 2002.

[15] B. M. Michael, R. E. Neapolitan, X. Jiang, and V. Shyam, "Learning genetic epistasis using Bayesian network scoring criteria," *BMC Bioinformatics*, vol. 12, no. 1, 89 pages, 2011.

[16] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Research Notes*, vol. 3, article 117, 2010.

[17] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, 2007.

[18] M. Dorigo, M. Birattari, and C. Blum, "Ant colony optimization and swarm intelligence," *SpringerVerlag*, vol. 5217, no. 8, pp. 767–771, 2004.

[19] T. Stützle, M. López-Ibáñez, P. Pellegrini et al., "Parameter adaptation in ant colony optimization," *Autonomous Search*, vol. 9783642214349, pp. 191–215, 2012.

[20] C. Blum and M. Sampels, "An ant colony optimization algorithm for shop scheduling problems," *Journal of Mathematical Modelling and Algorithms*, vol. 3, no. 3, pp. 285–308, 2004.

[21] R. Musa, J.-P. Arnaout, and H. Jung, "Ant colony optimization algorithm to solve for the transportation problem of cross-docking network," *Computers and Industrial Engineering*, vol. 59, no. 1, pp. 85–92, 2010.

[22] G. N. Varela and M. C. Sinclair, "Ant colony optimisation for virtual-wavelength-path routing and wavelength allocation," in *Proceedings of the 1999 Congress on Evolutionary Computation (CEC '99)*, pp. 1809–1816, Washington, DC, USA, July 1999.

[23] K. M. Sim and W. H. Sun, "Ant colony optimization for routing and load-balancing: survey and new directions," *Systems Man & Cybernetics Part A Systems Humans IEEE Transactions on*, vol. 33, no. 5, pp. 560–572, 2003.

[24] S.-H. Ngo, X. Jiang, and S. Horiguchi, "Adaptive routing and wavelength assignment using ant-based algorithm," in *Proceedings of the 2004 12th IEEE International Conference on Networks, ICON 2004 - Unity in Diversity*, pp. 482–486, November 2004.

[25] S. I. Vrieze, "Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychological Methods*, vol. 17, no. 2, pp. 228–243, 2012.

[26] D.-S. Huang and J.-X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 12, pp. 2099–2115, 2008.

[27] B. V. North, D. Curtis, and P. C. Sham, "Application of logistic regression to case-control association studies involving two causative loci," *Human Heredity*, vol. 59, no. 2, pp. 79–87, 2005.

[28] P.-J. Jing and H.-B. Shen, "MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies," *Bioinformatics*, vol. 31, no. 5, pp. 634–641, 2015.

[29] N. Ryman, "CHIFISH: A computer program testing for genetic heterogeneity at multiple loci using chi-square and Fisher's exact test," *Molecular Ecology Notes*, vol. 6, no. 1, pp. 285–287, 2006.

[30] C. R. Mehta and N. R. Patel, "A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 427–434, 1983.

[31] B. Sobrino, M. Brión, and A. Carracedo, "SNPs in forensic genetics: A review on SNP typing methodologies," *Forensic Science International*, vol. 154, no. 2-3, pp. 181–194, 2005.

[32] O. Shoval, H. Sheftel, G. Shinar et al., "Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[33] D.-S. Huang and W. Jiang, "A general CPL-AdS methodology for fixing dynamic parameters in dual environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 5, pp. 1489–1500, 2012.

[34] L. Zhu, W.-L. Guo, S.-P. Deng, and D.-S. Huang, "ChIP-PIT: enhancing the analysis of chip-seq data using convex-relaxed pair-wise interaction tensor decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 55–63, 2016.

[35] C. Angione, G. Carapezza, J. Costanza, P. Lio, and G. Nicosia, "Pareto optimality in organelle energy metabolism analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1032–1044, 2013.

[36] R. A. Fisher, "On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, vol. 85, no. 1, p. 87, 1922.

[37] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, pp. 131–153, 1992.

[38] B. Wenzheng, C. Yuehui, and W. Dong, "Prediction of protein structure classes with flexible neural tree," *Bio-Medical Materials and Engineering*, vol. 24, no. 6, pp. 3797–3806, 2014.

[39] L. Zhu, Z.-H. You, D.-S. Huang, and B. Wang, "t-LSE: a novel robust geometric approach for modeling protein-protein interaction networks," *PLoS ONE*, vol. 8, no. 4, Article ID e58368, 2013.

[40] C.-H. Zheng, L. Zhang, V. T.-Y. Ng, C. K. Shiu, and D.-S. Huang, "Molecular pattern discovery based on penalized matrix decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1592–1603, 2011.

[41] D.-S. Huang and H.-J. Yu, "Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 457–467, 2013.

[42] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, vol. 37, no. 4, pp. 413–417, 2005.

[43] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics*, vol. 10, no. 1, article S65, 2009.

[44] J. Kruppa, A. Ziegler, and I. R. König, "Risk estimation and risk prediction using machine-learning methods," *Human Genetics*, vol. 131, no. 10, pp. 1639–1654, 2012.

[45] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, 2006.

[46] R. W. Mahley, K. H. Weisgraber, and Y. Huang, "Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5644–5651, 2006.

[47] E. M. Reiman, J. A. Webster, A. J. Myers et al., "GAB2 alleles modify Alzheimer's Risk in APOE $\varepsilon 4$ carriers," *Neuron*, vol. 54, no. 5, pp. 713–720, 2007.

[48] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[49] S.-P. Deng, L. Zhu, and D.-S. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 27–35, 2016.

[50] L. Zhu, S.-P. Deng, and D.-S. Huang, "A two-stage geometric method for pruning unreliable links in protein-protein networks," *IEEE Transactions on Nanobioscience*, vol. 14, no. 5, pp. 528–534, 2015.

[51] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein and Peptide Science*, vol. 15, no. 6, pp. 553–560, 2014.

[52] D.-S. Huang, *Systematic Theory of Neural Networks for Pat-tern Recognition*, Publishing House of Electronic Industry of China, May 1996.

[53] D.-S. Huang, "Radial basis probabilistic neural networks: model and application," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 7, pp. 1083–1101, 1999.