Bioinformatics, 31(5), 2015, 634–641
doi: 10.1093/bioinformatics/btu702
Advance Access Publication Date: 22 October 2014
Original Paper

OXFORD

Genome analysis

MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies

Peng-Jie Jing[1,2] and Hong-Bin Shen[1,2,*]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University and [2]Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 4, 2014; revised on October 12, 2014; accepted on October 19, 2014

Abstract

Motivation: The existing methods for genetic-interaction detection in genome-wide association studies are designed from different paradigms, and their performances vary considerably for different disease models. One important reason for this variability is that their construction is based on a single-correlation model between SNPs and disease. Due to potential model preference and disease complexity, a single-objective method will therefore not work well in general, resulting in low power and a high false-positive rate.

Method: In this work, we present a multi-objective heuristic optimization methodology named MACOED for detecting genetic interactions. In MACOED, we combine both logistical regression and Bayesian network methods, which are from opposing schools of statistics. The combination of these two evaluation objectives proved to be complementary, resulting in higher power with a lower false-positive rate than observed for optimizing either objective independently. To solve the space and time complexity for high-dimension problems, a memory-based multi-objective ant colony optimization algorithm is designed in MACOED that is able to retain non-dominated solutions found in past iterations.

Results: We compared MACOED with other recent algorithms using both simulated and real datasets. The experimental results demonstrate that our method outperforms others in both detection power and computational feasibility for large datasets.

Availability and implementation: Codes and datasets are available at: www.csbio.sjtu.edu.cn/bioinf/MACOED/.

Contact: hbshen@sjtu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the development of genome-wide high-density single nucleotide polymorphism (SNP) genotyping technology, genome-wide association studies (GWAS) that aim to explore associations between SNPs and disease in a population play an increasingly important role in identifying the causes of disease (Churchill *et al.*, 2004; Fontanesi *et al.*, 2012). Although great success has been achieved in identifying single-locus SNPs that are responsible for disease such as Mendelian diseases (Moore *et al.*, 2010), searching for susceptible loci for complex diseases, such as diabetes, hypertension and some psychiatric disorders, has proven to be more problematic. One reason for this is that many diseases are influenced by multi-locus SNPs that interact with each other in a process known as epistasis and thereby complicating matters. Epistasis plays an essential role in human complex disease, and methods designed

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

for single-locus detection do not work for epistasis detection due to these interactions that cannot be modeled (Moore *et al.*, 2010).

Recently, a number of multi-locus approaches have been proposed to detect epistasis. These approaches can generally be classified into three categories: (i) exhaustive search, (ii) stochastic search and (iii) machine-learning approaches (Shang *et al.*, 2011; Xie *et al.*, 2012).

Exhaustive search enumerates all possible single-locus and multi-locus SNP combinations to identify the causative SNP or SNPs. A representative method of exhaustive search is the multifactor-dimensionality reduction algorithm, in which multi-locus genotypes are partitioned into two groups termed high-risk and low-risk, thereby reducing the original high-dimensional genotype-prediction problem to one dimension. Next, an exhaustive search is then conducted with these newly formed one-dimensional variables, and the model that predicts disease status the best by cross-validation and permutation testing will be selected (Ritchie *et al.*, 2001). Wan *et al.* (2010) proposed another exhaustive search method named Boolean operation-based screening and testing (BOOST), which screens data using an upper bound of the likelihood followed by an exhaustive search. Although exhaustive search is the most straightforward idea that has been helpful in many complex diseases, the recent high-dimensional datasets (containing up to one million SNPs) that lead to the so-called 'small sample size problem' [e.g. the small N (number of samples), large P (number of SNPs) problem] prohibit its practical application in the real-world GWAS field because implementation requires tremendous computing resources and takes an unusually long time.

Stochastic methods detect epistasis via random sampling that can greatly speeds up the process. Bayesian epistasis association mapping (BEAM) is one example. This method detects suspected SNPs and their interactions via a Bayesian partitioning model and computes the posterior probabilities of the candidates belonging to true-associated SNPs and epistasis via Markov Chain Monte Carlo sampling. BEAM significantly outperformed existing methods as tested on an age-related macular degeneration GWAS dataset (Zhang and Liu, 2007). However, stochastic methods are criticized for using random elements in each iteration, which results in a dramatic loss of power as the search space expands exponentially (Shang *et al.*, 2011).

With the development of machine-learning technologies, an increasing number of statistical learning methods have been used in the GWAS field, such as regression-based algorithms (Wu *et al.*, 2009), Bayesian networks (BNs) (Jiang *et al.*, 2011) and others. The regression-based method is regarded as the most natural first-line approach for modeling and testing genetic effects (Van Steen, 2012). North *et al.* applied the logistic regression approach to case–control association studies involving two causative loci (North *et al.*, 2005). Jiang *et al.* (2011) tried a different solution which constructed a BN to model the association between SNPs and disease that was subsequently evaluated by a series of scoring criteria. However, these traditional statistical machine-learning methods still suffer from the 'small sample size problem' and generally perform poorly with real-world GWAS datasets. To solve this problem, other machine-learning-based methods introduced heuristic information to speed up the process, such as AntEpiSeeker (Wang *et al.*, 2010). Wang *et al.* developed the AntEpiSeeker method using heuristic search based on the ant colony optimization (ACO) algorithm and performed better than other methods on the rheumatoid arthritis dataset from the Wellcome Trust Case Control Consortium (Burton *et al.*, 2007). Machine-learning-based methods are generally acknowledged to have variable strengths and weakness due to their formation from different aspects.

A number of studies have revealed that these approaches perform inconsistently with different disease models and that even the same approach will often vary when applied to different disease models. One important reason is that existing approaches were constructed based on a single correlation model or objective function for SNPs and disease. Considering the potential preference of the correlation model and the complexity of different disease models, a single-objective method will understandably not work well in general. In this study, we proposed a new multi-objective heuristic search approach based on the ACO algorithm called MACOED.

We conducted experiments on a wide range of synthetic datasets and achieved good power and time performance. We also compared MACOED with the representative methods, including AntEpiSeeker, BEAM and BOOST and found that our method showed improved power in detecting correct epistatic interactions with different disease models.

2 Materials

2.1 Simulated datasets

In this study, we present two different types of epistasis models commonly used in generating simulation datasets: the DME model (which displays both the interactive and marginal effects of the disease) and the DNME model (which displays only interactive and no marginal effects of the disease) (Shang *et al.*, 2011; Wan *et al.*, 2010; Xie *et al.*, 2012).

An epistasis model is usually defined by the penetrance table whose elements represent the probability of being affected with the disease given the genotype combination, denoted as $P(D|G_i)$ where D indicates disease and G_i indicates the ith genotype combination. The penetrance values are usually decided by three parameters: disease prevalence ($P(D)$), genetic heritability (h^2) and minor allele frequency (MAF). More details about the penetrance table and the definitions of the three parameters are provided in the supplementary information.

Despite being given the three user-specified parameters above, we still could not solve the penetrance table because too many free parameters exist. Thus, in DME epistasis models, researchers usually designate a penetrance function which represents the relationship between each conditional probability $P(D|G_i)$. Using this information, we can now numerically solve the penetrance table and generate the datasets accordingly. In this article, we considered three different penetrance functions used in DME epistasis models. For these three penetrance functions, we fix $P(D)$ and h^2 and vary MAF with four values (0.05, 0.1, 0.2 and 0.5). Then, through Equations (S8) and (S9), and the penetrance functions, we obtain 12 different penetrance tables. Details about the three DME models and corresponding 12 penetrance tables for different parameter settings are provided in the supplementary information. Here, we generated simulated datasets with solved DME penetrance tables using the software GAMETES_2.0 (Urbanowicz *et al.*, 2012).

In DNME epistasis models with user-specified $P(D)$, h^2 and MAF, we have no choice but to search the field of $P(D|G_i)$. If $P(D|G_i)$ satisfies the condition of displaying no marginal effects, a DNME penetrance table will be obtained. Using the software GAMETES_2.0, we can conveniently search the DNME penetrance tables and generate the corresponding simulated datasets. In our experiments, we searched 40 DNME penetrance tables with different combinations of $P(D)$, h^2 and MAF and generated their simulated

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

datasets. The details about 40 DNME penetrance tables are also provided in the supplementary information.

To study the effects of the algorithm, 100 datasets are generated for each penetrance table, with each dataset containing 100 SNPs and 1600 samples. Each dataset contains an equal number of cases and controls for both the DME and DNME models. In addition, to test the algorithm for different sizes, we generated another 100 datasets for each penetrance table, in which the SNP size varies from 20, 100, 500 and 1000 and the sample size varies from 200, 400, 800 and 1600. All these datasets can be downloaded at www.csbio.sjtu.edu.cn/bioinf/MACOED/.

2.2 Real GWAS dataset

Late-onset Alzheimer's disease (LOAD) is the most common form of Alzheimer's disease (AD), which is frequently diagnosed in people over 65 years of age. To date, the apolipoprotein E (APOE) $\varepsilon 4$ allele has been definitively determined to contribute to LOAD risk (Avramopoulos, 2009). The APOE gene has three common variants, $\varepsilon 2$, $\varepsilon 3$ and $\varepsilon 4$, of which $\varepsilon 2$ has some protective effects by inhibiting AD occurrence, while $\varepsilon 4$ has some induced effects by increasing incidence of LOAD. SNPs related to LOAD have been reported in the GWAS field (Reiman *et al.*, 2007). Reiman *et al.* found that 10 SNPs located in the GAB2 gene on chromosome 11q14.1 have an epistasis effect with APOE $\varepsilon 4$ in association with LOAD disease.

We applied the LOAD GWAS data from https://www.tgen.org/. After pre-processing, the LOAD dataset consists of 1411 samples. Of these, 861 were diagnosed with LOAD, and 550 were not. Each sample in this dataset contains the genotype information of 312 316 SNPs, APOE status and LOAD status. Here, we recode the APOE gene state with a binary variable where 1 represents the $\varepsilon 4$ variant and 0 represents the other three variants. An SNP locus was recoded as a quaternary variable considering the missing state.

3 Methods

In MACOED, two objectives are combined to evaluate the search results. For the first one, standard logistic regression is used to model the relationship between genotypes and phenotypes, and the Akaike information criterion (AIC) score, which represents the likelihood and complexity of the model, is designated as *Objective 1*. For the second one, we adopt a Bayesian perspective to model the genotype-phenotype association using a BN and the BN-derived K2 score as *Objective 2* to represent the fitness of the model and correlation relationship between SNP subsets and disease status. The above two objectives are designed from opposing schools of statistics to rate the associations, and our following results show that they are complementary to each other and result in a better performance on general datasets.

For solving the small sample size problem, we consider the identification of the associated SNPs as a heuristic optimization problem. In MACOED, we obtained the optimal solutions (named the non-dominated solutions) in terms of the two objectives using the Pareto optimal optimization technique based on the ant colony system (Chaharsooghi *et al.*, 2008).

To avoid the problem of randomness in the heuristic search, we designed the following heuristic-exhaustive two-step protocol: (i) a non-dominated SNP subset is generated by the multi-objective ACO optimization algorithm; and (ii) an exhaustive search of epistatic interactions is conducted with a χ^2 test within the derived non-dominated SNP subset in the first stage. It is also worth pointing out that the ACO optimization algorithm implemented in the first step is different from the traditional ACO approach, as a new memory strategy has been proposed to take into account correlations between successive iterations. Using the new ACO variant, the good solutions found in each step will have a higher chance of being retained.

3.1 Logistic regression and BN

Objective 1 is derived from logistic regression, which has been widely used in GWAS (Wu *et al.*, 2009). Based on the results of North *et al.* (2005), we construct the ADDINT logistic regression model with the phenotypes as dependent variables and the genotypes as independent variables. With logistic regression analysis, we can compute the maximized log-likelihood of the model, denoted as log *lik*, and the number of free parameters, denoted as *d*. Then we obtain the AIC score of the model as follows:

$$\text{AIC score} = -2\log lik + 2d \tag{1}$$

The AIC score reflects both the model's fitness to the dataset by log *lik* and the complexity by *d*. As the AIC score is designated as Objective 1 in MACOED, by comparing AIC scores of different SNPs in the model, the evidence for the effect of different SNPs on disease risk can be investigated. In this modeling approach, SNPs with low AIC score are considered as disease-correlated SNPs.

Objective 2 is derived from the BN, also known as the directed graphical model. In the GWAS BN, the genotypes and phenotypes are denoted as a set of nodes, and their conditional dependences are denoted as a set of edges. There are many BN structure learning methods, and on the basis of previous studies (Han *et al.*, 2012; Jiang *et al.*, 2011) in this article we choose the K2 score as follows:

$$\text{K2 score}_{\log} = \sum_{i=1}^{I}\left(\sum_{b=1}^{r_i+1}\log(b) - \sum_{j=1}^{J}\sum_{d=1}^{r_{ij}}\log(d)\right), \tag{2}$$

where I is the combinatorial number of SNP nodes with different values (if l-SNP nodes are connected to disease node y, the number of SNP nodes' combinations is 3^l as the possible value of an SNP node is 0, 1 or 2), J is the state number of disease node y (two for all samples), r_i is the number of cases with SNP nodes taking the ith combination and r_{ij} is the number of cases where the disease node takes the jth state and its parents take the ith combination.

The K2 score is a measure of the causative relationship between SNP nodes and disease nodes. Thus, we designate the K2 score as Objective 2 in the proposed MACOED, and the lower the logarithm score, the stronger the association between the SNP subset and the disease (more details about the derivation of the two objectives are in the supplementary information).

3.2 Pareto optimal approach

In the previous sections, we introduced two different modeling approaches and their corresponding score functions. In MACOED, we designate AIC score as Objective 1, K2 score as Objective 2 and treat an SNP subset from whole GWAS data as a candidate solution to the two objectives. For both objectives, the lower the score, the stronger the association between the SNP subset and the disease. Thus, the problem of detection of epistasis becomes the problem of finding the best solution with respect to the two objectives.

In the real world, a solution (an SNP subset) may have the best performance in one Objective while performing poorly in the other Objective as compared with other solutions. Thus, for a problem with more than one Objective function, there is usually no unique

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

optimal solution. The relationship between two solutions has only two possibilities: either one dominates the other or neither dominates. In terms of the two objectives in this article, a solution A_1 is said to dominate another solution A_2 if it satisfies both of the following conditions:

1. For both objectives, the value of $f_w(A_1)$ is not higher than $f_w(A_2)$.
2. The Objective $f_w(A_1)$ is strictly lower than $f_w(A_2)$ for at least one Objective.

where $f_w(\bullet)$ is the Objective function as previously defined, and $w \in \{1,2\}$ denotes the two Objective functions used in this article.

If the above two conditions are satisfied, solution A_1 is a non-dominated solution, and, accordingly, A_2 is a dominated solution. For all solutions according to both objectives, we can classify them into two sets: a non-dominated set and a dominated set. As all solutions in the non-dominated set have equal optimizations, the epistasis detection problem can be extended to find a non-dominated set of solutions, also known as the 'Pareto Optimal Set'. The mathematical equation for the example 2-SNP epistasis model can be formulated as follows:

$$\text{Minimize}\begin{cases} f_1(A_k) = \text{AIC score}(A_k) \\ f_2(A_k) = \text{K2 score}_{\log}(A_k) \end{cases}, \tag{3}$$

where A is the decision space, and $A_k \in A$.

Here, we use the non-dominated sort algorithm (Supplementary Fig. S2) to find the non-dominated set in the decision space A (Deb, 1999). The pseudo code of the algorithm and the time complexity are found in the supplementary information.

3.3 Ant colony optimization

To reduce the complexity of exhaustive search, we introduce a swarm intelligence optimization algorithm named the ACO algorithm, which has proved useful in GWAS.

In the case of GWAS, the detecting space comprises all SNPs and their combinations. The ants traverse routes and construct solutions which consist of any possible SNP combinations according to the pheromone values and transfer rules. Pheromone values are stored as a matrix τ, whose dimensionality is the number of SNPs in the entire GWAS dataset, and whose element represents the interaction between its column-label SNP and row-label SNP in association with the disease. At the starting point of MACOED, the pheromone value between every pair of SNPs is initialized to a fixed value τ_0, meaning that we treat the interaction among each SNP and the association between the SNPs and disease with equal possibility.

The transfer rule indicates that an ant k selecting SNP j from SNP i and consists of two parts, as described in Equations (4) and (5):

$$p_k(i,j) = \begin{cases} R \ \text{ if } (q \leq T_0) \\ 1 \ \text{ when } j = \text{rand}(U_k(i)) \text{ if} (q > T_0) \end{cases}, \tag{4}$$

where $p_k(i,j)$ is the probability that an ant selects SNP j followed by SNP i, q is a number generated randomly with uniform distribution in (0,1) and T_0 is a threshold to balance the convergence speed and to avoid being trapped into a locally optimal solution. $U_k(i)$ is the set of neighbor nodes of SNP node i that have not yet been visited by ant k, and rand($U_k(i)$) denotes the ant k selecting a random SNP from the set $U_k(i)$. R represents the selection strategy of the next SNP to be added by considering their pheromones' distribution. The selection probability distribution of R is depicted as follows:

$$R = \begin{cases} \dfrac{{\tau_{ij}}^{\delta}{\eta_j}^{\beta}}{\sum_{u \in U_k(i)} \tau_{iu}^{\delta}\eta_u^{\beta}} & \text{if}(j \in U_k(i)) \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where τ_{ij} is pheromone value between SNP i and SNP j, η_j is some form of prior information on SNP j and δ and β are parameters determining the weights of pheromone value and prior information on the SNPs, respectively. In our work, we let η equal to 1, indicating that we treat each locus equally before the optimization phase, and, similarly, we let δ equal to 1.

The ACO algorithm obtains the optimal solutions through positive feedback formed by the pheromone iteration. As described previously, all non-dominated solutions have the equivalent and highest association strength with disease, and all other solutions must be omitted. Thus, the pheromone value will be updated as:

$$\tau_{ij} \leftarrow (1\text{-}\rho)\tau_{ij} + \rho\Delta\tau_{ij}, \tag{6}$$

where ρ is the evaporate coefficient between 0 and 1, and $\Delta\tau_{ij}$ is the changing pheromone value between SNP i and SNP j as calculated below:

$$\Delta\tau_{ij} = \begin{cases} \lambda, \text{ if } A_k \in \text{non-dominated solution set} \\ 0, \text{otherwise} \end{cases}, \tag{7}$$

where λ is a weight coefficient for non-dominated solutions.

In MACOED, we implement a variant of ACO in order to save the optimal solutions in each iteration and speeding up the convergence. In traditional ACO, each iteration is independent of each other, so solutions generated in the previous iterations will be ignored in the current iteration; such 'zero-order' working mode will lose some of the good solutions already discovered in past iterations. Instead of completely discarding the solutions from previous iterations, we design a memory-based strategy to retain the non-dominated solutions from the last iteration instead and compare them to the detected solutions of new ants in the current iteration to decide the final non-dominated solutions, which will continue to be used in the next iteration. The merit of this strategy is that good solutions generated in any of the iterations will not be lost, yielding a more accurate way for epistasis searching. All non-dominated solutions in the last iteration will be fed into the second stage of MACOED after the number of ACO iterations reaches the user-specified number.

3.4 Pearson's χ^2 test

In the previous 'screening' stage, we have obtained a set of non-dominated solutions after the ACO combined with the Pareto optimal approach. In the next 'cleaning' stage, an exhaustive search of epistatic interactions with Pearson's χ^2 test is conducted within the selected non-dominated solutions. The χ^2 test is the most frequently used approach and has been integrated into many GWAS software packages (Balding, 2006; Wang *et al.*, 2010; Zhang *et al.*, 2012). The main virtue of the χ^2 test is its simple yet powerful computation and its identification of associated SNPs without considering the disease model.

In epistasis detection, the χ^2 test is based on the contingency table. The null hypothesis is that the SNP dataset has no association with the disease, and the sampling distribution of the test statistic is a χ^2 distribution when the null hypothesis is true. The alternative hypothesis that the SNP dataset has a certain extant association

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

with the disease is accepted when the P-value of the test statistic is below a user-defined significance level α_0. However, facing the problem of increased type I error in the presence of multiple testing, we implemented the conservative Bonferroni correction. Thus, SNP subsets with P-value below the Bonferroni-corrected significance level $\alpha = \alpha_0/C_m^l$ will be reported by the MACOED algorithm, where m is the SNP number of the GWAS dataset and l is the user-specified interaction size.

The elements discussed above are synthesized to comprise the proposed MACOED algorithm; the pseudo code of our algorithm is given in Supplementary Figure S3. Figure 1 shows an example explaining of the MACOED algorithm for epistasis learning of 2-SNP interactions with a total SNP number of 5.

3.5 Evaluation criteria

To evaluate the performance of the epistasis-learning algorithm, the power is traditionally defined as follows:

$$\text{Power} = \frac{\#(S)}{100}, \tag{8}$$

where $\#(S)$ means the number of datasets in which the disease-associated SNPs are successfully identified among all 100 datasets generated by the same parameters and penetrance table.

Obviously, using only one criterion to evaluate a statistical approach is one-sided. In the GWAS problem, we treat the true associated SNP combinations as positives and the true unassociated SNP combinations as negatives, leading to an imbalanced problem of a small number of positives and a large number of negatives. Taking the dataset used in our experiment as an example, the dataset has 100 SNPs and thus has $C_{100}^2 = 4950$ different combinations when considering 2-SNP correlations. In these combinations, we only have one true disease-associated SNP combination indicating one positive and all other 4949 combinations are negatives. Using the traditional criteria, the more associated SNPs an algorithm reports, the greater its chance of obtaining a high power.

In order to strictly evaluate a method on its performance as balanced by outputted true and false-positive rates, we introduce a set of more appropriate criteria named precision, recall and F-measure, which are calculated based on the confusion matrix (Fig. 2) and commonly used in the pattern recognition field. The recall (also known as sensitivity) is the number of true positives in the output divided by the total number of true positives in all datasets (Equation 9). The precision is the number of true positives in the output divided by the total number of outputted values and is used to reflect the false-positive rate of an algorithm (Equation 10). In the GWAS problem, a high recall means the algorithm can return most true associated SNP combinations without considering the number of returned SNP combinations; while a high precision means the true associated SNP combinations account for a high proportion of the returned SNP combinations. The criterion F-measure is the harmonic mean of precision and recall, which is a synthesized measure combining both precision and recall (Equation 11).

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{10}$$

$$F\text{-measure} = \frac{2}{1/\text{precision} + 1/\text{recall}}, \tag{11}$$

where TP, FN and FP are defined in Figure 2.

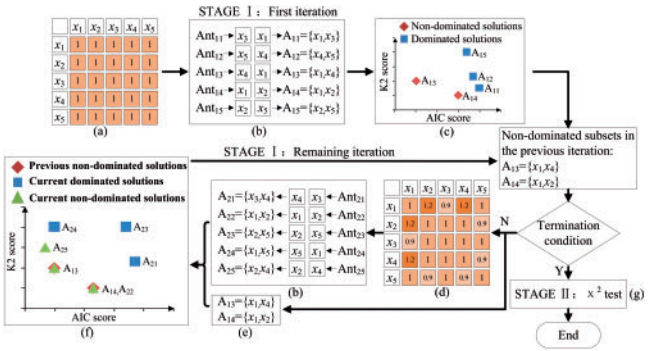

Fig. 1. Flowchart of steps involved in implementation of the MACOED algorithm: (a) the pheromone matrix is initialized with all values equal to 1, indicating that we treat each SNP equally. (b) $\text{Ant}_{tk}(\text{A}_{tk})$ denotes the kth ant in the t_{th} iteration. Ant_{tk} seeks the SNP subsets and constructs solutions according to the transfer rule. (c) The non-dominated sort algorithm is used to divide the solutions into two parts. (d) The pheromone matrix is updated according to the sort program result; here we set $\rho = 0.9$, $\lambda = 3$. The deep color denotes the strong interaction between the row SNP and the column SNP and their association strength with disease. The light color denotes the opposite trend. (e) The designed memory strategy keeps non-dominated solutions in the previous iteration that will be compared with current solutions. (f) The non-dominated solutions from the previous iteration and the current constructed solutions are combined and sorted by the non-dominated sort algorithm to obtain the current step's solution set. (g) An exhaustive search with χ^2 test is conducted in the non-dominated solutions remaining in the last iteration, where the SNP sets with P-values below the Bonferroni-corrected significance level will be outputted

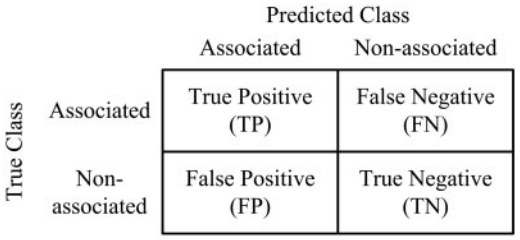

Fig. 2. Confusion matrix used to evaluate the epistasis learning approach

4 Experiments and results

4.1 Experiments on simulated datasets

In this section, the MACOED algorithm was tested in terms of different aspects, and the section is organized as follows.

First, in order to demonstrate the superiority of combining multiple objectives over a single objective, a naive exhaustive multi-objective epistasis-detecting method is conducted to compare with exhaustive single-objective method.

Second, the performance of the MACOED algorithm is compared with some recently proposed methods, including AntEpiSeeker, BEAM and BOOST. The software of the naive exhaustive multi-objective epistasis-detecting method and MACOED can be downloaded from www.csbio.sjtu.edu.cn/bioinf/MACOED/.

4.1.1 Multi-objective versus single-objective

In the exhaustive single-objective method, all SNP combinations are evaluated by the score function, and the output ranks all SNP combinations where the association with disease decreases as the score increases. The output of the exhaustive multi-objective method is a set of non-dominated solutions that are treated as having the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

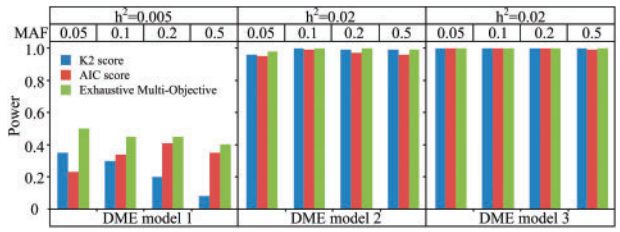

Fig. 3. Power performance comparisons between the single- and multi-objective exhaustive searching methods on three DME models

same optimum. To be unbiased, we set the number of outputted top SNPs from a single-objective to be the same as the number of SNPs in the non-dominated set of the multi-objective method.

Figure 3 presents the performance comparisons of different methods on three DME models, and Figure 4 shows the results on DNME models. Clearly, the results show that the multi-objective method outperforms all single-objective methods in both the DME and DNME models, demonstrating the multi-objective's superiority over the single-objective. In terms of the two single objective functions, it shows that the performance of the K2 score and AIC score are closer in DME models, indicating that the Bayesian model and logistic model can fit the DME models well from different perspectives. However, in DNME models the performance of the K2 score is more robust than the AIC score. The reason for this discrepancy may be that we consider all possible distributions of disease given the associated SNPs to have equal likelihood in the Bayesian models, while in the logistic models we only consider the ADDINT model which might lose power because the disease model displays no marginal effects.

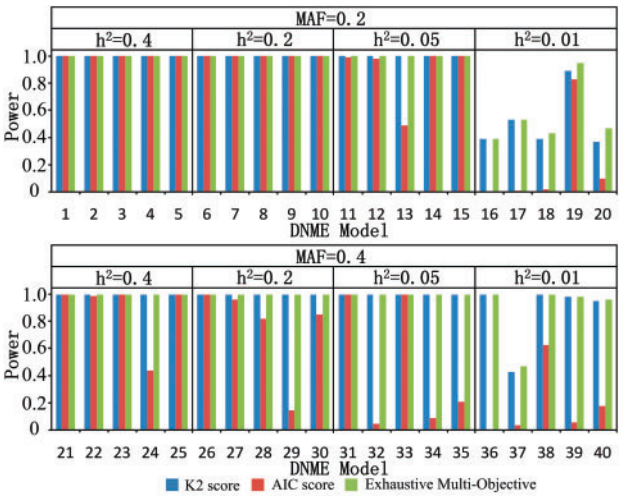

Fig. 4. Power performance comparisons between the single- and multi-objective exhaustive searching methods on DNME models with eight different sets of parameters

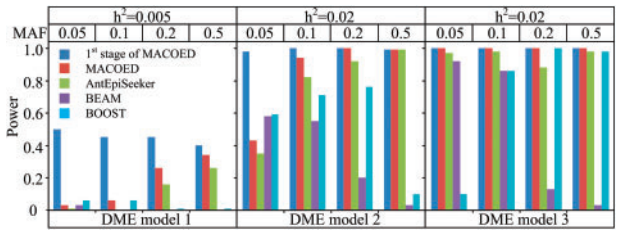

Fig. 5. Power performance comparisons between the MACOED and other comparative methods on three DME models

4.1.2 MACOED versus other comparative methods

We also compared our algorithm with some commonly used algorithms, including AntEpiSeeker, BEAM and BOOST. In this experiment, we set the significance threshold *P*-value as 0.1 after Bonferroni correction; thus, the unadjusted *P*-value for 2-locus epistasis detection in 100 SNPs is $0.1/C^2_{100} = 2.0202 \times 10^{-5}$.

We set the parameters of the three comparative methods according to the author's recommendations. In our method, we specified several parameters including τ_0, T_0, ρ, λ, *dim_epi*, *num_ant* and *max_iter* (Supplementary Fig. S3). For a 2-locus epistasis detecting experiment, we set $\tau_0 = 1$ and $dim\_epi = 2$. According to the previous study (Wang *et al.*, 2010), a large T_0, ρ and λ should be adopted for a small number of SNPs in GWAS datasets (denoted as *m*) and small values for a large *m*. The values of *num_ant* and *max_iter* are also determined by *m*, where a large *num_ant* and *max_iter* should be adopted correspondingly to a large *m*.

Because the key idea of our method is multi-objective based on the ACO algorithm, we output the intermediate non-dominated solutions in the screening stage and evaluate them with the same power definition. Figure 5 presents the performance comparisons on three DME models, and Supplementary Figure S4 presents the performance comparisons on DNME models. As shown in Figure 5, the results of the MACOED algorithm exhibit increased power on the DME models when compared with other methods in most sets of parameters, with the exception of $MAF = 0.05$ on the first and second DME models. This is because tiny h^2 and *MAF* values may make the χ^2 test in the second stage of MACOED perform poorly, where the contingency table might have some empty cells.

It is also interesting to observe from Figure 5 that the power of intermediate results generated by the first stage of MACOED has a better performance than the other methods in all settings of DME models, denoting that the intermediate results are worth studying, despite not returning correspondingly significant levels. Supplementary Figure S4 shows the same trend with Figure 5 that the MACOED has a better performance in high h^2 and *MAF*s.

These results demonstrate that, in DNME models, the performance of both the intermediate and final outputs of MACOED are very comparable with the outputs of BOOST in the case of $h^2 = 0.4$, 0.2 and 0.1, while the power of MACOED is a bit lower than BOOST on most models when $h^2 = 0.05$. This is because the DNME models only display interactive with no marginal effects, while BOOST's mathematical model only considers the interactive with no marginal effects, thus fitting this dataset perfectly. In addition, BOOST adopts an approximated χ^2 test with four degrees of freedom for 2-locus interactions, while MACOED takes a χ^2 test with eight degrees of freedom. This may cause BOOST to report more associated SNPs than other methods and achieve a higher power accordingly.

To more strictly compare and evaluate different methods, we adopt the aforementioned new criteria to reevaluate the outputs of each method. Table 1 presents the precision, recall, and *F*-measure of all three DME models with parameter settings of $MAF = 0.2$ for all compared methods. The results of all parameter settings of the DME models and DNME models can be seen in Supplementary Tables S4 and S5.

As seen from Table 1, our MACOED method outperforms other comparative methods not only in criterion recall but also in precision, thereby resulting in a superior performance in overall *F*-measure. When $MAF = 0.2$, MACOED has the highest recall in

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 1. Different criterion performances on DME models of *MAF* = 0.2 for MACOED and other comparative methods

Model	Method	Recall	Precision	*F*-measure
DME Model 1	Stage I of MACOED	0.46	0.04	0.08
	MACOED	**0.26**	**0.74**	**0.39**
	AntEpiSeeker	0.16	0.7	0.26
	BEAM	0	0	0
	BOOST	0.01	0.01	0.01
DME Model 2	Stage I of MACOED	1	0.89	0.94
	MACOED	**1**	**0.96**	**0.98**
	AntEpiSeeker	0.92	0.94	0.93
	BEAM	0.2	0.12	0.15
	BOOST	0.76	0.51	0.61
DME Model 3	Stage I of MACOED	**1**	**1**	**1**
	MACOED	**1**	**1**	**1**
	AntEpiSeeker	0.88	0.99	0.93
	BEAM	0.13	0.32	0.18
	BOOST	1	0.63	0.77

Model 1 is a 2-locus multiplicative model, Model 2 is a 2-locus threshold model, and Model 3 is a 2-locus concrete model. Refer to supplementary information for more results of different parameter settings.

the DME models 2 and 3, and the highest precision in all three DME models, indicating that it has the lowest false-positive rate (Type I error). Although the first stage of MACOED achieves the highest recall in the DME model 1, it has low precision and low *F*-measure, indicating that it improves the recall at the cost of increasing the false-positive rate. However, the results of first stage of MACOED in the other two DME models achieve relatively good performances in all criteria and hence yield a closer *F*-measure with MACOED.

In all 12 different parameter settings of the three DME models, MACOED has the highest *F*-measure in nine of them, and the increase from the second highest *F*-measure to the highest *F*-measure ranges from 0.02 to 0.13 (Supplementary Table S4). BEAM has a lower *F*-measure in most settings except for achieving the highest when *MAF* = 0.05 of DME Model 2, possibly due to the data simulation method and its stochastic protocol. BOOST achieves a high recall especially in cases of high *MAF*s. However, its precision is lower when compared with MACOED, which will thus yield a lower *F*-measure. The results demonstrate that BOOST reports more disease-associated SNPs, resulting in a high sensitivity but with a higher false-positive rate. AntEpiSeeker has a good performance in both recall and precision, resulting in a high *F*-measure, due to its minimizing false-positives procedure (Wang *et al.*, 2010).

In the 40 different parameter settings of DNME models, MACOED achieves the highest *F*-measure in 36, and the remaining four are achieved by BOOST (Supplementary Table S5). These remaining four parameter settings are all in $h^2 = 0.01$. These results imply that MACOED may perform poorly with low heritability in DNME models. In summary, most of the results prove that the multi-objective method is an effective method for detection of epistasis.

We also developed a comparative test between MACOED and the random feature method. As the mean length of outputted SNPs is 4 in MACOED, we randomly select four SNPs in each iteration in the random feature method and this process iterates 50 times (the same as the *max_iter* in MACOED). Then we combine the selected SNPs in each iteration and obtain the mean evaluation criteria for all models: Recall = 0.71, Precision = 2×10^{-4}, and *F*-measure = 4×10^{-4}. The comparative trial results indicate that the random feature method tends to report too many false positives, rendering the results meaningless.

In addition, in the comparison experiments among different methods, we compare the mean running time and power for one model (considering both the DME and DNME models) between MACOED, AntEpiSeeker, BEAM and BOOST on different sample sizes *N* and SNP sizes *M*. All experiments were performed on a computation platform using a Windows system with 8G RAM and i7-37700 CPU. The results are provided in Supplementary Figure S5. As can be seen, when *M* increases, the running-time curves of MACOED, AntEpiSeeker and BEAM are between linear and quadratic. Surprisingly, BOOST is the fastest method despite its categorization as an exhaustive method. This is due to its efficient filtering step that may also lose some good solutions. The power of MACOED outperforms all three other methods on small SNP sizes and is comparable with BOOST on large SNP sizes. On the other hand, the running time of all methods increase linearly as *N* increases. The power of MACOED outperforms the three other methods in this case. These results demonstrate that MACOED is time-efficient and capable of achieving better power than the compared approaches.

4.2 Experiments on a real GWAS dataset

We performed 2-locus epistasis detection using our MACOED algorithm in SNPs from each separated chromosome, combining APOE gene state and parameters set as suggested in the previous section. All outputted epistases of our MACOED algorithm contain the APOE gene and the other SNPs scattered in all chromosomes, most of which come from GAB2 with Bonferroni-corrected *P*-value = 0.1. Seven of our detected SNPs, which come from chromosome 11, are in the 10 SNPs reported by Reiman *et al.*

Interestingly, MACOED also identifies some new significant genes that interact with the APOE gene. For example, the rs6486084 and rs4347364 from chromosome 11 are reported by MACOED but have not yet biologically validated. All the mined new knowledge about epistatic interactions for LOAD is summarized in Supplementary Table S6. The results suggest a basis for further experimental validation and demonstrate that our method has practicality for using in detecting gene–gene interactions on the real GWAS datasets.

5 Discussion

To the best of our knowledge, we are the first to implement a multi-objective optimization framework based on swarm intelligence optimization in the GWAS field. The success of MACOED is primarily due to its complementary multi-objective and Pareto optimization, which increases sensitivity and minimizes false positives. On the other hand, our new memory-based strategy in the ACO algorithm also helps to increase the final power.

Although the results demonstrate that our method performs well on both simulated datasets and a real GWAS dataset, some limitations remain. The performance of logistic regression, a parametric method, partly depends on the mathematical model adopted in the association study. Based on the study by North *et al.*, different logistic models used in the first stage of MACOED may partly affect the results of non-dominated solutions (North *et al.*, 2005). Another criticism of logistic regression is that the number of parameters in models will increase exponentially when the number of loci

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

considered involved in epistasis increases, making the computation impractical.

On the other hand, we take an equal likelihood for all possible distributions in Bayesian model, which may lose power in some particular disease models. In addition, there is a number of criteria used to evaluate the BN models' fitness to the association study, e.g. BDeu, MDL (Jiang *et al.*, 2011). To balance the complexity and accuracy of our algorithm, we adopted a naive evaluation score, which may also partly affect the non-dominated solution results as well.

At the same time, the multi-objective algorithm may be criticized for the complexity of multi-objective computation and the naive non-dominated sorting used in Pareto optimal optimization. As the dimension of the GWAS dataset increases, this problem becomes more and more complex. Thus, in MACOED, we adopted the ACO algorithm and employed the iteration method to solve the large-dimension problem. Our previous experimental results demonstrate that MACOED is time-efficient. However, a few different optimization algorithms can be considered that would further reduce the complexity in future works, such as the non-dominated sorting genetic algorithm II (Deb *et al.*, 2002). We are also prepared to use the feature selection or filtering method to reduce the complexity of MACOED when applied in large GWAS datasets.

In future studies, we intend to find more powerful modeling approaches and corresponding score functions or appropriate efficient optimization strategies that can be combined and flexibly embedded into our framework to increase its ability. The other important future direction is that we will try to determine the best way to incorporate prior knowledge into MACOED when dealing with specific GWAS datasets. Currently, we have not used prior information in MACOED. However, useful prior information can improve the power and efficiency of epistasis detection as indicated by Greene *et al.* (2008).

Acknowledgement

We are grateful to Miss. Sara Walker for reading through the manuscript.

Funding

The National Natural Science Foundation of China [Nos. 61222306, 91130033, 61175024], Shanghai Science and Technology Commission [No. 11JC1404800] and a Foundation for the Author of National Excellent Doctoral Dissertation of PR China [No. 201048].

Conflict of interest: none declared.

References

Avramopoulos,D. (2009) Genetics of Alzheimer's disease: recent advances. *Genome Med.*, **1**, 34.

Balding,D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7**, 781–791.

Burton,P.R. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Chaharsooghi,S.K. *et al.* (2008) An effective ant colony optimization algorithm (ACO) for multi-objective resource allocation problem (MORAP). *Appl. Math. Comput.*, **200**, 167–177.

Churchill,G.A. *et al.* (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.*, **36**, 1133–1137.

Deb,K. (1999) Multi-objective genetic algorithms: problem difficulties and construction of test problems. *Evol. Comput.*, **7**, 205–230.

Deb,K. *et al.* (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, **6**, 182–197.

Fontanesi,L. *et al.* (2012) A genome wide association study for backfat thickness in Italian Large White pigs highlights new regions affecting fat deposition including neuronal genes. *BMC Genomics*, **13**, 583.

Greene,C.S. *et al.* (2008) Ant colony optimization for genome-wide genetic analysis. In D.,Marco *et al* (eds). *Ant Colony Optimization and Swarm Intelligence*. Springer, Heidelberg, pp. 37–47.

Han,B. *et al.* (2012) Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. *BMC Syst. Biol.*, **6**, S14.

Jiang,X. *et al.* (2011) Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics*, **12**, 89.

Moore,J.H. *et al.* (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**, 445–455.

North,B.V. *et al.* (2005) Application of logistic regression to case–control association studies involving two causative loci. *Hum. Hered.*, **59**, 79–87.

Reiman,E.M. *et al.* (2007) GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*, **54**, 713–720.

Ritchie,M.D. *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Shang,J. *et al.* (2011) Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, **12**, 475.

Urbanowicz,R.J. *et al.* (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining*, **5**, 1–14.

Van Steen,K. (2012) Travelling the world of gene–gene interactions. *Brief. Bioinform.*, **13**, 1–19.

Wan,X. *et al.* (2010) BOOST: a fast approach to detecting gene–gene interactions in genome-wide case–control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

Wang,Y. *et al.* (2010) AntEpiSeeker: detecting epistatic interactions for case–-control studies using a two-stage ant colony optimization algorithm. *BMC Res. Notes*, **3**, 117.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Xie,M. *et al.* (2012) Detecting genome-wide epistases based on the clustering of relatively frequent items. *Bioinformatics*, **28**, 5–12.

Zhang,X. *et al.* (2012) Mining genome-wide genetic markers. *PLoS Comput. Biol.*, **8**, e1002828.

Zhang,Y. and Liu,J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, **39**, 1167–1173.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016