Bioinformatics, 36(16), 2020, 4389–4398
doi: 10.1093/bioinformatics/btaa215
Advance Access Publication Date: 30 March 2020
Original Paper

Genome analysis

Multipopulation harmony search algorithm for the detection of high-order SNP interactions

Shouheng Tuo*, Haiyan Liu and Hao Chen

School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on September 5, 2019; revised on January 1, 2020; editorial decision on March 23, 2020; accepted on March 24, 2020

Abstract

Motivation: Recently, multiobjective swarm intelligence optimization (SIO) algorithms have attracted considerable attention as disease model-free methods for detecting high-order single nucleotide polymorphism (SNP) interactions. However, a strict Pareto optimal set may filter out some of the SNP combinations associated with disease status. Furthermore, the lack of heuristic factors for finding SNP interactions and the preference for discrimination approaches to disease models are considerable challenges for SIO.

In this study, we propose a multipopulation harmony search (HS) algorithm dedicated to the detection of high-order SNP interactions (MP-HS-DHSI). This method consists of three stages. In the first stage, HS with multipopulation (multiharmony memories) is used to discover a set of candidate high-order SNP combinations having an association with disease status. In HS, multiple criteria [Bayesian network-based K2-score, Jensen–Shannon divergence, likelihood ratio and normalized distance with joint entropy (ND-JE)] are adopted by four harmony memories to improve the ability to discriminate diverse disease models. A novel evaluation criterion named ND-JE is proposed to guide HS to explore clues for high-order SNP interactions. In the second and third stages, the *G*-test statistical method and multifactor dimensionality reduction are employed to verify the authenticity of the candidate solutions, respectively.

Results: We compared MP-HS-DHSI with four state-of-the-art SIO algorithms for detecting high-order SNP interactions for 20 simulation disease models and a real dataset of age-related macular degeneration. The experimental results revealed that our proposed method can accelerate the search speed efficiently and enhance the discrimination ability of diverse epistasis models.

Availability and implementation: https://github.com/shouhengtuo/MP-HS-DHSI.

Contact: tuo_sh@126.com

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

1 Introduction

Recently, the cost of genome-wide sequencing has been greatly reduced with the rapid development of genome sequencing technology and genome-wide data. This cost reduction provides a valuable opportunity for genome-wide association studies (GWASs) that aim to search genomes for small variations, such as single nucleotide polymorphisms (SNPs). SNPs are the most common genetic variation in the DNA sequences of humans. In a person's genome, there are >3 million SNPs. Complex diseases are widely considered to result from genetic variations, especially the interaction of multiple SNPs (called high-order SNP interaction). High-order SNP interaction represents a combination of multiple SNPs that influence complex diseases linearly or nonlinearly, and detection is of great importance in identifying the pathogenic causes of complex diseases in humans.

Over the past decade, thousands of SNPs associated with diseases have been identified, focusing mainly on individual SNPs isolated based on their contribution to disease status. However, the findings fail to effectively explain the causes of complex diseases. Numerous researchers have begun to conduct methodological studies for detecting high-order SNP interactions (Goudey *et al.*, 2015; Guo *et al.*, 2014; Gyenesei *et al.*, 2012; Wan *et al.*, 2010; Wang *et al.*, 2015; Yang *et al.*, 2009; Zhang and Liu, 2007).

The methodologies used to detect SNP interactions include determining how to accurately distinguish the relationship between high-order SNP combinations and disease status and how to quickly explore high-order SNP interactions throughout the genome. The greatest challenge is developing a fast search method for finding disease-causing SNP combinations without evaluating all feasible SNP combinations.

Exhaustive search is simple and reliable, but very high computational costs are required to evaluate the associations of all feasible SNP combinations, which are astronomical in number. To solve this problem, search techniques such as high-performance computing

(Goudey *et al.*, 2015; Guo *et al.*, 2014; Gyenesei *et al.*, 2012; Yung *et al.*, 2011) and random search (Yang *et al.*, 2009; Zhang and Liu, 2007) have been proposed to speed up the detection of high-order SNP interactions. High-performance computing is employed to evaluate the associations of all k-order SNP combinations by using a high-performance computer system, such as parallel computing and cloud computing. However, when k is >3, the detection of k-order SNP interactions from genomic data with hundreds of thousands of SNPs is still unable to achieve high performance due to the enormous computational burden. Although traditional stochastic search algorithms can discover some k-order SNP interactions, these algorithms are still insufficient for the detection of high-order SNP interactions by using posterior association probabilities of individual loci, especially for the detection of complex disease models with low or no marginal effects (Moore *et al.*, 2010).

Recently, swarm intelligence optimization (SIO) algorithms, such as genetic algorithms (GAs) (Mitchell, 1996), ant colony optimization (Blum, 2005), differential evolution (Storn and Price, 1995), particle swarm optimization (Eberhart and Kennedy, 1995) and harmony search (HS) (Geem *et al.*, 2001), have received much attention for their use in the detection of high-order SNP interactions because SIO has the following advantages over traditional optimization algorithms (e.g. gradient descent).

1. SIO is a global optimization method because it does not depend on the initial search point and can escape from a local optimum when the population is trapped in a local search.
2. The objective function of SIO is not restricted to continuity and differentiability but may be expressed in any form.
3. SIO is powerful for exploring a complex search space and has the ability to discover the global optimal solution rapidly via learning and communication between individuals in a population.

The harmony search algorithm (HSA) is a simple SIO algorithm that can easily implement the algorithm code. This algorithm is effective and efficient in solving combinatorial optimization problems and has been widely applied to solve complex NP-hard problems (Alia, 2017).

In this work, our goal is to design and develop a fast and effective HSA to speed up the detection of high-order SNP interactions from hundreds of thousands of SNPs. To achieve this objective, we have specified the following aims:

1. Propose a novel HSA (named MP-HS-DHSI) with multiple populations (four harmony memories: HM1, HM2, HM3 and HM4) for enhancing the global exploration power, in which four objective functions (K2-score, JS-score, LR-score and JE-score) are employed to evaluate the fitness value of candidate solutions in HM1–HM4. The multipopulation is intended to improve the diversity of populations to prevent the search from being trapped in a local optimum.
2. Employ a three-stage selection mechanism to decrease the computational burden. In the first stage, HS is used to choose candidate solutions (k-order SNP combinations) that have a stronger association with disease status. In the second stage, the G-test statistical method is adopted to test significant associations between candidate SNP combinations and disease status. In the third stage, multifactor dimensionality reduction (MDR) with 10-fold cross-validation is employed to choose solutions with high prediction accuracy for disease status.
3. Employ four complementary scoring functions (objective functions) that aim to discriminate various disease models and overcome preferences for specific types of disease models.
4. Investigate the arithmetic speed, detection power and precision of the proposed MP-HS-DHSI algorithm for disease models with marginal effects (DMEs) and disease models with no marginal effects (DNMEs).
5. Investigate the efficiency and effectiveness of the proposed algorithm for two real datasets: age-related macular degeneration (AMD) and breast cancer (BC).

In MP-HS-DHSI, the process of detection consists of three stages:

First stage. Searching candidate solutions. In the first stage, an HSA with multiple populations (four HMs) is proposed to explore candidate SNP combinations that have a greater association with disease status than other SNP combinations. Four complementary and lightweight objective functions (also called score functions) are employed to calculate the associations between SNP combinations and disease status. The main role of the first stage is to reduce the computational burden using HSA, where multiple populations are used for improving the global exploration ability of HSA, and four objective functions are dedicated to enhancing the discrimination ability to diverse disease models and avoid preference for specific disease models.

Second stage. Testing with G-test. In the first stage, when the search algorithm ends, the SNP combinations that remain in four HMs will be treated directly as candidate solutions as a rough choice. To further decrease the number of SNP combinations, the G-test statistic method is adopted to choose the SNP combinations that have a significant association with disease status (P-value < 0.01/the total number of feasible k-order SNP combinations) in the second stage.

Third stage. Determination with MDR. To further verify the SNP combinations that are chosen in the second stage, MDR with 10-fold cross-validation is utilized to calculate the classification accuracy and filter certain SNP combinations with low classification accuracy.

The flowchart of MP-HS-DHSI is shown in Supplementary Figure S1.

2 Materials and methods

2.1 Concepts and terms

Let a set of SNP variables $X = \{x_1, x_2, \ , x_N\}$ indicate N SNP markers for n individuals (samples) and $Y = \{y_1, y_2, \ldots, y_J\}$ denote the disease variable [J is the number of disease states; $J = 2$, in this work, $y_j = 0$ for control and $y_j = 1$ for case ($j = 1, 2$)]. The homozygous major allele, heterozygous allele and homozygous minor allele in the sample dataset are defined as 0, 1 and 2, respectively. For a k-order SNP combination, there are $I = 3^k$ genotype combinations. n_i is the number of samples (including case and control) in the dataset with SNP loci having the value of the ith genotype combination, and n_{ij} represents the number of samples with the ith genotype combination that are actually associated with disease state y_j.

Definition (high-order SNP interaction). Let $S_k = \{x_{s_1}, x_{s_2}, \ldots, x_{s_k}\}(1 < k < N, x_{s_i} \in X)$ be a set with k SNP loci. $f(S_k, Y)$ is a score function for evaluating the association between S_k and disease state Y. A k-order SNP combination S_k is said to be synergistically associated with Y if and only if $\forall S' \subset S_k \wedge f(S_k, Y) \succ f(S', Y)$($\succ$ is a binocular operator for comparing the association strength with disease) and is said to be strongly associated with Y if $f(S_k, Y) > \theta$ (θ is a threshold value). A k-order SNP combination S_k is called a k-order SNP interaction if and only if it is truly a disease-causing SNP combination with Y.

Mathematical model. To detect genome-wide k-order SNP interactions, we establish the optimization model as follows:

$$\max_X \ f(X, Y), X = (x_{s_1}, x_{s_2}, \ldots x_{s_k}), \tag{1}$$

where $s_i (i = 1, 2, \ldots, k)$ is the index of SNP locus x_{s_i} and $f(X, Y)$ denotes the objective function for calculating the association between SNP combination X and disease status Y.

This model is a very complex combinatorial optimization problem, and the number of k-order SNP combinations for data with n SNPs is equal to $C_n^k \propto n^k$. Clearly, it is impractical to examine and evaluate the associations of all feasible k-order SNP combinations

from a dataset with hundreds of thousands of SNPs using an exhaustive search algorithm. To address this complex problem, we propose a novel HSA for accelerating the detection of high-order SNP interactions.

2.2 Harmony search algorithm

HSA is a population-based metaheuristic algorithm whose idea is to achieve the cognition of unknown complex problems by exchanging information and learning between individuals in a group. This algorithm is inspired by the process of music creation by jazz musicians, where musicians improvise their instruments' pitches searching for a perfect state of harmony (Das *et al.*, 2011; Geem *et al.*, 2001). The HSA has been widely applied in combination optimization problems with large scale, such as portfolio optimization (Tuo, 2016, 2018; Tuo and He, 2019) and the knapsack problem (Kong *et al.*, 2015).

In an HSA, a candidate solution $X = (x_1, x_2, \ldots, x_N)$ is referred to as a harmony. A set of candidate solutions is named a HM, which is similar to the memory of a Tabu search (TS) algorithm and the population of a GA. The number of harmonies in an HM is called HMS. An HM is a matrix of order HMS×N or an augmented matrix of order HMS × $(N+1)$ (Zhang and Geem, 2019) as follows:

$$\mathrm{HM} = \left[\begin{array}{c|c} X^1 & f(X^1) \\ X^2 & f(X^2) \\ \vdots & \vdots \\ X^{\mathrm{HMS}} & f(X^{\mathrm{HMS}}) \end{array}\right] = \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_N^1 & f(X^1) \\ x_1^2 & x_2^2 & \cdots & x_N^2 & f(X^2) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1^{\mathrm{HMS}} & x_2^{\mathrm{HMS}} & \cdots & x_N^{\mathrm{HMS}} & f(X^{\mathrm{HMS}}) \end{array}\right],$$

where X^i ($i = 1, 2, \ldots$, HMS) is the ith harmony in HM and $f(X^i)$ denotes the value of the objective function.

The steps of the standard HSA are introduced in Supplementary File S1.

Detecting genome-wide high-order SNP interactions is a classic combinatorial optimization problem. Generally, the sequencing data of SNPs have the characteristics of high dimensionality and small sample sizes. A very high computational burden is encountered due to the combinatorial explosion of SNPs and multiple tests. To address this problem, we propose a fast HSA (named MP-HS-DHSI) for discovering candidate SNP combinations.

2.3 MP-HS-DHSI algorithm

In MP-HS-DHSI, multipopulation (four HMs) and multiple objective functions (four criteria for evaluating the association between SNP combinations and disease status) are adopted to improve the search ability of HSA, where four HMs respectively employ BN K2-score, JS-score, LR-score and ND-JE-score as objective functions for evaluating the association between SNP combination and disease status. In the proposed MP-HS-DHSI algorithm, multipopulation is intended to promote the global search ability for a high-dimensional genome-wide search space; the use of multicriteria is intended to enhance the discrimination ability for various disease models and avoid the preference for special disease models. The MP-HS-DHSI flow chart is introduced in Supplementary Figure S1.

In the MP-HS-DHSI, each harmony in an HM is regarded as an SNP combination among all k-order SNP combinations. A k-order SNP combination consists of k SNPs that are distinct from one another. The method aims to find k-order SNP combinations associated with disease status, and all of the SNPs in an SNP combination work together for the disease.

2.4 Complementary and lightweight scoring functions

Another very important challenge is authenticating whether a k-order SNP combination is a true epistasis interaction or a false one. Existing methods, such as logistic regression, mutual information and machine learning, either have preferences for the types of disease or require complicated computations. Logistic regression, which is a parametric model, is time consuming to use to fit parameters based on iteration and is limited in identifying k-order SNP interactions involving many loci since the number of combination

Algorithm 1. MP-HS-DHSI

Inputs:
Data matrix with m samples (includes case samples and control samples) and N SNPs.
MaxFEs: maximum number of evaluations of associations between SNP combinations and disease status, used as the terminal condition for HS.
k: the number of SNPs in an SNP combination.
Parameters: **HMCR, PAR**.
Output: a set of k-order SNP combinations that have a strong association with disease status.

Step 1. Randomly initialize four harmony memories (HM_1, HM_2, HM_3, and HM_4).
Step 2. Calculate the four scores of each harmony X in the four HMs.
[K2-score, JS-score, LR-score, ND-JE] = $f(X,Y)$.
Step 3. Generate new harmony X^{new} as **algorithm 2.**
Step 4. Update HM1, HM2, HM3 and HM4 using X^{new} as **algorithm 3**.
Step 5. Check the stopping criterion. If the stopping criterion (MaxFEs) is met, the computation is terminated. Otherwise, Steps 3 and 4 are repeated.

Algorithm 2. Improvise the new harmony X^{new}

(1) $\mathbf{R} = \mathbf{r} \in \{1,2,3,4\}$ // randomly select an HM from {HM1, HM2, HM3, HM4} to optimize.
For i = 1→k
  If rand(0,1) < HMCR
    $x_i^{new} = HM_R(a,i), a \in \{1,2,\ldots,HMS\}$
    If rand(0,1) < PAR
      $x_i^{new} = HM_R(a_1,i) + r_i \times (HM_R(a_2,i) - HM_R(a_3,i)),$
      $a_j \in \{1,2,\ldots,HMS\}, j = 1,2,3$
    EndIf
    $x_i^{new} = round(x_i^{new})$
  Else
    $x_i^{new} = r \in \{1,2,\ldots,N\}$
  EndIf
  // avoid generating repeated SNPs in X^{new}
  While $x_i^{new} \in \{x_1^{new}, x_2^{new}, \ldots, x_{i-1}^{new}\}$
    $x_i^{new} = r \in \{1,2,\ldots,N\}$
  EndWhile
EndFor
// avoid generating repeated SNP combinations in HMR
If $X^{new} \in HM_R$
  $J = i \in \{1,2,\ldots,k\}$
  $x_j^{new} = r \in \{1,2,\ldots,N\}$
  If $X^{new} \in HM_R$, Goto (1) **EndIf**
EndIf

$HM_R(a,i)$ denotes harmony x_i^a in $\mathrm{HM_R}$.

Algorithm 3. Update HM1, HM2, HM3 and HM4 with X^{new}

Inputs: X^{new}, HM1, HM2, HM3 and HM4
Outputs: HM1, HM2, HM3 and HM4

1. [K2-score$^{\text{new}}$, JS-score$^{\text{new}}$, LR-score$^{\text{new}}$, ND-JE$^{\text{new}}$] = $f(X^{new})$.
2. **If** K2-score$^{\text{new}}$ < K2-score$^{\text{HM1,worst}}$
3. $X^{HM_1,worst} \leftarrow X^{new}$
4. **If** JS-score$^{\text{new}}$ < JS-score$^{\text{HM2,worst}}$
5. $X^{HM_2,worst} \leftarrow X^{new}$
6. **If** LR-score$^{\text{new}}$ < LR-score$^{\text{HM3,worst}}$
7. $X^{HM_3,worst} \leftarrow X^{new}$
8. **If** ND-JE$^{\text{new}}$ < ND-JE$^{\text{HM4,worst}}$
9. $X^{HM_4,worst} \leftarrow X^{new}$
10. $X^{HM_i,worst}$ represents the worst harmony in HM_i (i=1,2,3,4).
11. K2-score$^{\text{HM1,worst}}$ is the K2-score of the worst harmony in HM_1.
12. JS-score$^{\text{HM2,worst}}$ is the JS-score of the worst harmony in HM_2.
13. LR-score$^{\text{HM1,worst}}$ is the LR-score of the worst harmony in HM_3.
14. ND-JE$^{\text{HM2,worst}}$ is the ND-JE of the worst harmony in HM_4.

terms grows exponentially (Visweswaran *et al.*, 2009). Mutual information is a lightweight method but has preferences for certain disease models. Machine learning methods, such as MDR, neural network work and random forest, which are nonparametric approaches, have received much attention for identifying the associations of high-order SNP combinations with disease status recently, but the very high computational burden has limited the usefulness of these methods to identify genome-wide associations.

To decrease the computational burden and overcome the preferences for specific types of disease models, in this work, we employ four evaluation functions [Bayesian network-based (BN) K2-score, Jensen–Shannon (JS) divergence, likelihood ratio (LR) and normalized distance with joint entropy (ND-JE)] to calculate the degree of association of SNP combinations with disease status. The four evaluation functions not only complement each other in identification but are also lightweight for calculating association. The four scoring functions work in reciprocal ways, and this approach is conducive to improving the discrimination performance for the disease-causing SNP combinations with complementary mechanisms.

2.4.1 BN K2-score

A BN is a statistical model that describes random variables and associations among those random variables using a directed acyclic graph (DAG) $G = (V, E)$, in which node set V consists of random variables and edges E represent conditional dependences between the two linked variables. A causal DAG contains edge $X \rightarrow Y$ only if X is a direct cause of Y (Neapolitan, 2004; Zhang and Liu, 2007). A BN model is a lightweight computing method and has high discrimination precision for evaluating the association between a k-order SNP combination and disease status. In this work, the BN-based K2-score is expressed as Equation (1).

$$\text{K2} - \text{score} = \prod_{i=1}^{I}\left(\frac{(J-1)!}{(n_i+J-1)!}\prod_{j=1}^{J} n_{ij}!\right). \quad (1)$$

For ease of calculation of the K2-score in Equation (1), Equation (1) can be translated using a natural logarithm as Equation (2).

$$\text{K2}-\text{score}_{\log} = \sum_{i=1}^{I}\left(\sum_{k}^{n_i+1}\log(k) - \sum_{j=1}^{J}\sum_{s=1}^{n_{ij}}\log(s)\right). \quad (2)$$

The lower the value of the K2-score is, the greater the association between an SNP combination and disease status.

2.4.2 LR-score

The LR is a composite indicator that reflects both sensitivity and specificity and can be used for a related measure to find the likelihood difference between a disease-causing SNP combination and an SNP combination that is not involved in the disease process (Bush *et al.*, 2008; Neyman and Pearson, 1928) as follows:

$$\text{LR} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}\ \ln\left(\frac{o_{ij}}{e_{ij}}\right) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\ \ln\left(\frac{n_{ij}}{e_{ij}}\right), \quad (3)$$

where o_{ij} and e_{ij} respectively represent the observed number and expected number of genotypes when a phenotype takes the ith disease state and an SNP combination takes the jth genotype. The expected number e_{ij} could be obtained based on the Hardy–Weinberg principle (Crow, 1999). Supplementary Table S1 gives an example of a contingency table for a 2-order SNP model.

2.4.3 JS-score

The JS divergence is derived from the Kullback–Leibler (KL) divergence that is an asymmetric divergence measure of two probability distributions (Lin, 1991). Compared to KL divergence, the JS divergence is a symmetrized divergence measure. This measure can be used to measure the SNP genotype deviation between case data and control data. For an SNP combination, let the genotype distributions for the case and control be P_{case} and P_{control}, respectively, and the JS divergence between P_{case} and P_{control} can be calculated as Equation (4).

$$\begin{aligned}\text{JS} &= 0.5\left(\sum_{i=1}^{I}\left(p_i^{\text{case}} \times \log\frac{2p_i^{\text{case}}}{p_i^{\text{case}}+p_i^{\text{control}}} + p_i^{\text{control}} \times \log\frac{2p_i^{\text{control}}}{p_i^{\text{case}}+p_i^{\text{control}}}\right)\right),\\ &= 0.5\left(\sum_{i=1}^{I}\sum_{j=1}^{2}\frac{n_{ij}}{n_i}\log\frac{2n_{ij}}{n_i}\right)\end{aligned} \quad (4)$$

where p_i^{case} and p_i^{control} represent the ratio of the ith genotype combination in the case and control samples, respectively.

The larger the value of the JS divergence is, the greater the divergence between genotypes in the case and genotypes in the control, which means that the association between an SNP combination and disease status is stronger.

2.4.4 Normalized distance with joint entropy

Through many tests, we found that the disease-causing k-order SNP combination models with very low or no marginal effects usually have the following characteristics:

- For an individual SNP belonging to a k-order SNP combination without marginal effects, the genotype difference between the case and the control samples is very small or nonexistent.
- For a combination that involves certain disease SNPs, the genotype distributions of the case and control samples show little difference.

According to these characteristics, we develop a new score function, named ND-JE. This function is defined as follows:

$$\text{ND} - \text{JE} = {}^{\text{ND}(X)}\Big/_{\text{JE}(X_{\text{control}})}, \quad (5)$$

$$\mathrm{ND}(X) = \frac{\sum_{j=1}^{k} d(x_j)}{D(X)}, \tag{6}$$

$$d(x_j) = \sum_{s=0}^{2} \sqrt{(n_s^{j,\mathrm{control}} - n_s^{j,\mathrm{case}})^2}, \tag{7}$$

$$D(X) = \sum_{i=1}^{I} |n_i^{\mathrm{control}} - n_i^{\mathrm{case}}|, \tag{8}$$

$$\begin{aligned} \mathrm{JE}(X_{\mathrm{control}}) &= -\sum_{i=1}^{I} p_i^{\mathrm{control}} \log p_i^{\mathrm{control}} \\ &= -\sum_{i=1}^{I} \frac{n_i^{\mathrm{control}}}{n^{\mathrm{control}}} \log \frac{n_i^{\mathrm{control}}}{n^{\mathrm{control}}}, \end{aligned} \tag{9}$$

where $X = (x_1, x_2, \ldots, x_k)$ is a k-order SNP combination for all samples (including case and control samples); X_{control} indicates a k-order SNP combination for only control samples; $n_s^{j,\mathrm{control}}$ and $n_s^{j,\mathrm{case}}$ denote the numbers of control and case samples in the dataset, respectively, with the jth SNP locus taking the value of s (homozygous major allele 0, heterozygous allele 1 and homozygous minor allele 2); n_i^{control} and n_i^{case} represent the numbers of control and case samples in the dataset, respectively, with SNP combination X taking the value of the ith genotype combination; and n^{control} is the number of control samples.

In Equation (5), $d(x_j)$ ($j = 1,2,\ldots,k$) is used to evaluate the difference between the genotype distributions of the case and control at locus x_j. $D(X)$ aims to measure the distance between the case and control in k-order SNP combination $X = (x_1, x_2, \ldots, x_k)$. The smaller the value of $\mathrm{ND}(X)$ is, the larger the distribution difference (distance) between the case and control samples. The joint entropy (JE) of the control samples is employed to normalize the distance. The main goal of ND-JE is to find a clue to guide the HSA to detect disease-causing SNP combinations.

In the NHSA-DHSC algorithm (Tuo *et al.*, 2017), JE is proposed for guiding HS to discover clues for exploring disease-causing SNP combinations. In those simulation experiments, the NHSA-DHSC algorithm shows very good performance compared to other algorithms, but the generated simulation datasets do not follow the Hardy–Weinberg law. In this research, we find that the performance of JE is obviously degraded for data following the Hardy–Weinberg law. Compared with JE, the ND-JE possesses stronger robustness for various disease models.

2.4.5 Characteristics of the four evaluation functions

2.4.5.1 Lightweight. It can be clearly seen that the computational costs of the four score functions are very lightweight. Only the values of n_i and n_{ij} ($i = 1,2,\ldots,I$; $j = 1,2$) are required to count the individuals within the sample groups for a k-order SNP combination. The total calculation amount for the four scoring functions is not additive because n_i and n_{ij} can be used in the four objective functions repeatedly.

2.4.5.2 Complementary. Our previous studies (Tuo *et al.*, 2017) have found that the results were different when employing different objective functions. The BN-based K2-score has been widely used to evaluate the association. This measure has high power for detecting SNP interaction and is superior in discriminating certain disease models with low marginal effects for DMEs. However, for the interaction model with low minor allele frequencies and low genetic heritability (H^2), the BN-based K2-score has low performance for detecting high-order SNP interaction. The LR-score aims to discover the likelihood difference between a functional SNP combination and nonfunctional SNP combination by statistical theory. This method has good adaptability to unknown disease models. The JS-score is a symmetric divergence measure of two probability distributions, and its goal is to measure differences between the genotype distributions of the case and control based on information theory. The ND-JE in this work mainly aims to guide HSA to explore clues for disease-causing SNP combinations to accelerate HSA and reduce the computational burden caused by 'SNP combination explosion'. In Supplementary File S1, the four objective functions are compared using examples.

2.4.5.3 Heuristic. For swarm intelligence search algorithms, a heuristic factor is very important for finding clues about functional SNPs. In particular, for DNMEs, the objective function can determine the search efficiency. In Figure 1, we draw the landscape of K2-score, LR-score, JS-score and ND-JE for a DNME (Himmelstein *et al.*, 2011) (the dataset is from DNME-7 with five functional SNPs); the disease-causing SNP combination is SNP96, SNP97, SNP98, SNP99 and SNP100.

In Figure 1, the lower the score is, the stronger the association with disease. The scores are calculated per Algorithm 4. Here, we aim to determine which algorithms are able to distinguish between an SNP combination containing one or more functional SNPs and an SNP combination that does not contain functional SNPs.

From Figure 1a–c, we can see that the K2-score, JS-score and LR exhibit no obvious discrimination between an SNP combination including some of functional SNPs and that including no functional SNPs. Inversely, the scores of certain SNP combinations containing one or more functional SNPs are larger than the scores of those including no functional SNPs, which would force the algorithm away from the region to search and lose the optimal SNP combination.

However, the ND-JE shows this discrimination. We can see from Figure 1d that the more functional SNPs are included in the SNP combination, the lower the score is, which demonstrates that the ND-JE can provide clues for HS to explore high-order disease-causing SNP combinations during the search process.

2.5 *G*-test

In this first stage, SNP combinations that have a strong association with disease status are chosen as candidate SNP combinations.

In the second stage, the G-test statistical method (McDonald, 2014; Tuo *et al.*, 2016) is employed to test the significance level of candidate SNP combinations. This test is a LR test asymptotically similar to Pearson's χ^2 test, and it is superior to the approximation to the theoretical χ^2 distribution (Hoey, 2012). The formula for calculating the G-value is redefined by us in NHSA-DHSC (Hoey, 2012; Tuo *et al.*, 2016) as follows:

$$G = 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} p_{ij}$$

$$p_{ij} = \begin{cases} a \ln \dfrac{o_{ij}}{e_{ij}}, & \sum_{j=1}^{J} o_{ij} > \xi \\ 0, & \text{otherwise} \end{cases},$$

where o_{ij} is the observed number of the ith genotype when the disease status is y_j and e_{ij} is the corresponding expected number of the ith genotype when the disease status is y_j and can be calculated in terms of the Hardy–Weinberg principle (Crow, 1999).

2.6 Multifactor dimensionality reduction

MDR (Gola *et al.*, 2016; Namkung *et al.*, 2009; Ritchie *et al.*, 2001; Velez *et al.*, 2007; Yang *et al.*, 2017), as a nonparametric and model-free machine learning method, has received wide attention for use in detecting and characterizing nonlinear high-order SNP interactions in dichotomy classification. It is a very powerful and comprehensive method for the detection of epistasis. However, the MDR is a more expensive method than BN-based K2-score, JS-score and LR-score, limiting the usefulness of the MDR in directly detecting high-order SNP interaction at a genome-wide scale. To this end, we use MDR to determine a small number of candidate SNP combinations chosen from the second stage.

In this work, we employed the balanced accuracy (BA) (Velez *et al.*, 2007) and predictive error rate (PER) (Yang *et al.*, 2017) as evaluation measures of the MDR classifier. The BA and PER are defined as follows:

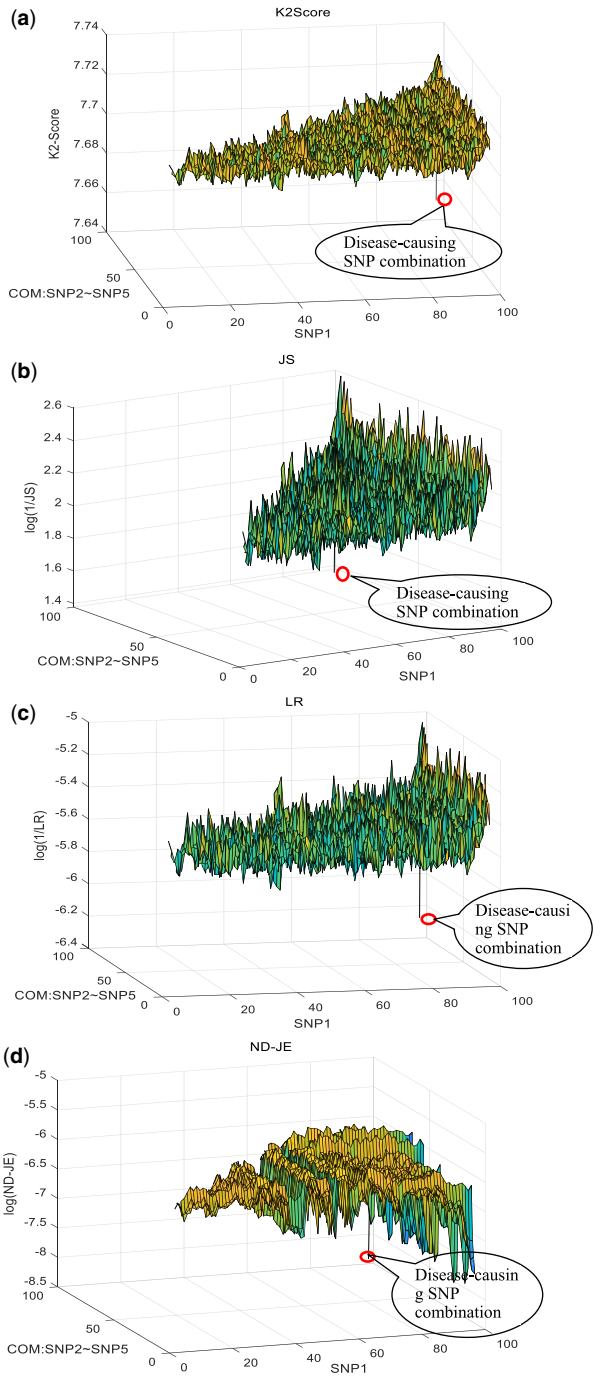

Fig. 1. The landscape of 5-order SNP combinations that are constructed in terms of Algorithm 4. (**a**) K2-score is utilized as objective function for evaluating the association between SNP combinations and disease status. (**b**) JS-score is utilized as objective function. (**c**) LR-score is utilized as objective function. (**d**) ND-JE is utilized as objective function. The point circled by the O is the disease-causing SNP combination (SNP96, SNP97, SNP98, SNP99 and SNP100)

$$\mathrm{BA} = 0.5\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right),$$

$$\mathrm{PER} = 0.5\left(\frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}\right).$$

Here, the output results of the MDR classifier are summarized in a 2×2 matrix as follows (Table 1).

Algorithm 4. Construct a matrix of 5-order SNP combinations

```
For I = 1→96
    For J = I+1 →97
        combination ← (I, J, J+1, J+2, J+3).
        Score(I, J) ← calculating the association of the combination
    EndFor
EndFor
```

Table 1. Confusion matrix of the MDR classifier

	High risk	Low risk
Cases	TP	FN
Controls	FP	TN

Note: TP is the number of correctly classified individuals in case samples by the MDR classifier. FP is the number of incorrectly classified individuals in control samples by the MDR classifier. TN is the number of correctly classified individuals in control samples by the MDR classifier. FN is the number of incorrectly classified individuals in case samples by the MDR classifier.

3 Simulation experiments

The performance of the proposed MP-HS-DHSI algorithm is evaluated for 12 DMEs, 8 DNMEs and 2 real datasets (AMD, BC). The algorithm is compared with four state-of-the-art swarm intelligence algorithms [CSE (Aflakparast *et al.*, 2014), MACOED (Jing and Shen, 2015), NHSA-DHSC (Tuo *et al.*, 2017) and epiACO (Sun *et al.*, 2017)], BEAM (Zhang and Liu, 2007) and exhaustive search (*G*-test is used as evaluation criteria for calculating the association between SNP combination and disease status). Supplementary Table S4 summarizes the parameters of the seven algorithms.

3.1 Evaluation criteria

Power is a measure of the capability of detecting disease-causing SNP combinations from genome data and is expressed as

$$\mathrm{Power} = \frac{S}{T},$$

where S is the number of found disease-causing SNP combinations from T datasets. Each dataset includes one disease-causing SNP combination. Here, the power is used to measure the search ability of the five intelligent optimization algorithms.

Note that the power in this work is mainly used to evaluate the search ability of the proposed method in the first stage. If the disease-causing SNP combination can be evaluated within the specified iterations (maximum number of objective functions evaluated, MaxFEs), the search is considered successful.

FEs stands for the number of evaluations of the associations between *k*-order SNP combinations and disease status until the disease-causing SNP combinations are found or the terminal condition of the algorithm is met.

runtime denotes the mean runtime that an algorithm takes to discover the disease-causing SNP combinations before the algorithm is terminated.

In this experiment, we aim to compare the speed of the algorithms by the FEs and runtime. FEs is used to measure the ability to reduce the computational burden. The value of FEs is less than M*axFEs* (M*axFEs* is far less than the number of feasible SNP combinations). The lower the value of FEs, the stronger the ability of the search algorithm to reduce the computational burden.

In the second stage, the significance level of *P*-value from *G*-test is set to $1/C_n^k$. In the third stage, the thresholds of BA and PER are set to 0.55 and 0.45, respectively.

For the simulated experiments, the functional SNP combinations are known before the search methods are run. To investigate the speed of the search algorithms, the search is terminated immediately when the disease-causing SNP combination is captured from the feasible SNP combinations during the search process. The lower the number of SNP combinations that are evaluated before the functional SNP combinations are found by an algorithm, the faster the search algorithm is.

3.2 Simulated datasets

3.2.1 Disease models with marginal effects

Twelve 2-order disease models (DME-1–DME-12) (the parameter settings are described in Supplementary Table S2) that have marginal effects and interaction effects were utilized (Himmelstein *et al.*, 2011; Jing and Shen, 2015; Tuo *et al.*, 2017). For each model, we generated two simulated 100 datasets with sample sizes of 800 (400 controls and 400 cases) and 100 datasets with 4000 (2000 controls and 2000 cases) using the software GAMETES_2.1 (Urbanowicz *et al.*, 2012).

Supplementary Table S5 summarizes the powers of seven algorithms for the 12 DME models. The results indicate that first NHSA-DHSC and first MP-HS-DHSI (in the first stage, the disease-causing SNP combinations are included in the set of candidate solutions that have not been tested by statistical methods) have obvious advantages over the other methods for datasets with 800 samples. However, the final results of MP-HS-DHSI are superior to those of NHSA-DHSC, which is because the NHSA-DHSC has a larger false negative rate than the proposed method, MP-HS-DHSI. Compared with exhaustive search, the MP-HS-DHSI is very close to it in terms of power. For datasets with sample sizes of 4000, the MP-HS-DHSI has the same power as exhaustive search except for DME-1–DME-3, and it is superior to other algorithms for all DME models.

Supplementary Table S6 shows the FEs of five swarm intelligence search algorithms and exhaustive search algorithm for the 12 DME models. The results display that, for the dataset with sample sizes of 800, the MP-HS-DHSI requires the fewest FEs to find the disease-causing SNP combinations except for models DME-1–DME-3, DME-5 and DME-11; for the dataset with sample sizes of 4000, the MP-HS-DHSI is superior to other algorithms except that the NHSA-DHSC is the superior for the DME-1 and DME-11. Supplementary Table S7 summarizes the runtime of seven algorithms for DME models. It shows that, for the datasets with sample sizes of 800, the NHSA-DHSC has shortest runtime for models DME-1–DME-3, DME-5–DME-6 and DME-9–DME-11, the proposed MP-HS-DHSI takes less time than the other algorithms. For the datasets, the MP-HS-DHSI takes the least time to find the disease-causing SNP combinations. For all datasets, the MP-HS-DHSI takes much less time than exhaustive search. In particular, for datasets with sample sizes of 4000, MP-HS-DHSI takes <1/10 of the exhaustive search and BEAM.

Summing up the above, the proposed method is effective and efficient for reducing the number of evaluations of SNP combinations on the DME models.

3.2.2 Disease models with no marginal effects

Eight *k*-order ($k = 3$, 4, 5) DNMEs (NM1–NM8) (see Supplementary Table S3 for the parameters of eight DNMEs) are employed to investigate the performance for the detection of high-order SNP interactions. For each DNME, there are 100 datasets with 1500 case samples and 1500 controls. The data of nonfunctional SNPs are generated randomly based on the Hardy–Weinberg equilibrium. Although NHSA-DHSC (Tuo *et al.*, 2017) has showed outstanding detection power for the detection of the eight DNME models, the simulation datasets of nonfunctional SNPs were generated randomly based on a uniform distribution that did not follow the Hardy–Weinberg law. Tables 2–4 depict the experimental results (power, runtime and FEs) of six intelligent search algorithms (MACOED can only detect 2-order SNP interactions) for DNMEs.

Table 2. The powers of four SIO algorithms for eight high-order DNMEs (NM1–NM8)

Model	NM1 (%)	NM2 (%)	NM3 (%)	NM4 (%)	NM5 (%)	NM6 (%)	NM7 (%)	NM8 (%)
CSE	67	64	73	23	67	12	1	3
NHSA-DHSC	51	10	92	3	22	93	2	5
epiACO	45	55	91	6	8	2	2	1
MP-HS-DHSI	98	65	100	100	67	100	83	33
BEAM	38	6	0	0	0	0	0	0
Exhaustive search	100	100	100	100	100	/	/	/

Note: '/' denotes that the run time is larger than 10 000 s. We were unable to complete the experiment in a limited time. NM represents the DNMEs.

Table 3. The runtime (s) of four SIO algorithms for eight high-order DNMEs

Model	CSE	NHSA-DHSC	epiACO	MP-HS-DHSI	Exhaustive search
NM1	147.2	114.8	56.2	**4.3**	202.7
NM2	137.8	218.2	47.4	**63.4**	202.7
NM3	758.6	95.2	160.6	**6.2**	4916
NM4	1053.2	601.7	162.3	**8.9**	4916
NM5	735.9	591.2	**286.4**	340	4916
NM6	6517.7	90.5	299.7	**12.5**	/
NM7	6606.9	1234	358.4	**43.3**	/
NM8	6567	1211	353.8	782	/

Note: '/' denotes that the run time is larger than 10 000 s; bold indicates the best result.

Table 4. The mean FEs of four SIO algorithms for eight high-order DNMEs

Model	CSE	NHSA-DHSC	epiACO	MP-HS-DHSI		Exhaustive search
				FEs	MP/Ex	
NM1	10 7530	77 272	62 785	**1872**	1.158%	161 700
NM2	110 743	81 001	56 960	**27 498**	17.006%	161 700
NM3	243 166	34 724	192 102	**2053**	0.052%	3 921 225
NM4	339 341	192 001	187 761	**2877**	0.073%	3 921 225
NM5	256 646	192 001	364 450	**101 338**	2.584%	3 921 225
NM6	374 692	30 613	375 100	**3389**	0.005%	75 287 520
NM7	374 523	375 001	374 429	**12 262**	0.016%	75 287 520
NM8	374 792	375 001	373 710	**265 701**	0.353%	75 287 520

Note: MP/Ex represents the ratio of the FEs required for the MP-HS-DHSI to the FEs required for an exhaustive search; bold indicates the best result.

As shown in Table 2, the MP-HS-DHSI has highest detection power in five nonexhaustive search algorithms and is much higher than CSE, NHSA-DHSC, epiACO and BEAM.

In terms of runtime (see Table 3), the MP-HS-DHSI takes much less time than other swarm intelligence search algorithms except for NM5 and NM8 (the epiACO takes less time than MP-HS-DHSI for the two models). In Table 4, the mean FEs required for MP-HS-DHSI is lowest among all compared algorithms; in particular, it is only 0.005% of the FEs required for exhaustive search on the NM6, which demonstrates that our method is effective for reducing the computation burden of detecting high-order SNP interactions.

3.2.3 AMD data

The proposed approach is employed to determine high-order SNP interactions for AMD data (Klein *et al.*, 2005) with 146 samples (96

cases and 50 controls) and 103 611 SNPs. There have been many studies on this AMD dataset, and the results can be taken as references for our proposed algorithm. In this experiment, certain candidate k-order ($k = 2, 3, 4, 5$) SNP combinations that have high association with AMD are output.

For these candidate k-order SNP combinations chosen in the first stage, the G-test and MDR are used to verify the interactions. A total of 170 2-order SNP combinations meet the significance level with G-test P-value 1E−7, the classification accuracy of MDR for each 2-order SNP combination is larger than 0.8 (see sheet SNP interaction in Supplementary File S2). Supplementary Figure S3 shows a 2-order SNP interaction network of the 170 2-order SNP combinations, in which each edge denotes a 2-order SNP combination that has a strong association with AMD. The results indicate that many SNPs interact with three SNPs: rs380390 (degree = 138 in Supplementary Fig. S3), rs1329428 (degree = 24 in Supplementary Fig. S3) and rs1363688 (degree = 7 in Supplementary Fig. S3). The SNPs rs380390 and rs1329428 have been widely reported to be associated with AMD (Guo *et al.*, 2019; Klein *et al.*, 2005; Lin and Lee, 2010; Shang *et al.*, 2015; Sun *et al.*, 2017; Tuo *et al.*, 2016, 2017; Zhang and Liu, 2007). Both of these SNPs are on gene *CFH* located in chromosome 1. They have been widely believed to be associated with AMD. rs1363688 (not in a gene) has also been reported to be associated with AMD (Guo *et al.*, 2014; Lin and Lee, 2010; Shang *et al.*, 2015; Sun *et al.*, 2017; Tuo, 2018; Tuo *et al.*, 2016, 2017). We suspect from Supplementary Figure S3 that these three SNPs may be drivers of AMD and that the cumulative effects of SNPs may be important to AMD.

In the experiment, four 3-order SNP combinations (see Supplementary Table S8) meet the significance level with G-test P-value 1E−8; the classification accuracy of MDR for these 3-order SNP combinations is >80%. Sixteen 4-order SNP combinations (See Supplementary Table S9) satisfy the significance level with P-value 1E−13 from G-test; the classification accuracy of these 16 4-order SNP combinations by MDR is >60%, but the accuracy of only one 4-order SNP combination (rs1038704, **rs380390**, rs9292651, **rs1363688**) is larger than 80%. As shown in Supplementary Tables S8 and S9, SNP '**rs380390**' is included in all 3- and 4-order SNP combinations, and both SNPs '**rs380390** and **rs1363688**' are included in all 4-order SNP combinations.

The top 15 SNPs with individual effects are summarized in Supplementary Table S10, in which **rs380390**, **rs1329428** and **rs1363688** are ranked 2, 4 and 12, respectively. The SNP rs3775652 is ranked first for individual effect but low marginal effect.

To elaborate on the interpretation of the identified SNP interactions in the real AMD data application, as detailed in Supplementary File S2, the values of linkage disequilibrium (LD, r^2) and the coefficient value for the identified SNP pairs are summarized. The χ^2 test P-value and classification accuracy of the support vector machine (SVM) are presented for the identified SNP combinations. The results indicate that the χ^2 test is invalid for high-order SNP combinations with small sample sizes, and the χ^2 test P-value is much less than the G-test P-value, meaning that the classification accuracy of the SVM is almost consistent with the classification accuracy of the MDR.

To further understand the identified high-order SNP interactions, we study the genotype combinations that tended to have a higher risk of BC, showing the BC % based on the genotype combinations (see sheet AMD in Supplementary File S2). The results indicate that for the SNP combinations (rs380390, rs10508731), higher-risk genotypes account for 83.33% of the case samples. With the 3-order SNP combination (rs380390, rs9296461, rs618499), the higher-risk genotypes account for 85.42% of the case samples. The higher-risk genotype combinations for the 4-order SNP combination (rs10503193, rs380390, rs2176324, rs1363688) account for 92.71% of the case samples. For detailed results, see sheet AMD in Supplementary File S2.xls.

3.2.4 BC dataset

We also applied the proposed method to detect high-order SNP interactions from a BC dataset that consisted of a candidate set of SNPs with 10 000 samples (Chuang *et al.*, 2012; Li, 2017; Li and Jiang, 2017; Yang *et al.*, 2013) (http://bioinfo.kmu.edu.tw). In the experiment, we try to detect 2-, 3-, 4- and 5-order SNP interactions. To elaborate on the interpretation of the identified SNP interactions for BC, the LD, χ^2 test and SVM methods are employed.

The results show that two 2-order SNP combinations (rs3020314, rs2017591) and (rs3020314, rs1514348) meet the significance level (see Supplementary Table S11) with the G-test P-value 0.0001 and χ^2 test P-value 1E−8. In both SNP pairs (rs3020314, rs2017591) and (rs3020314, rs1514348), there are six higher-risk BC genotype combinations, which account for 44.48% and 49.04% of case samples, respectively. For the 2-order SNP combination (rs3020314, rs2017591), the P-values from the G-test and χ^2 test are 5.577E−05 and 2.55E−11, respectively; this combination has been considered to have an association with BC (Chuang *et al.*, 2012; Li, 2017; Li and Jiang, 2017; Yang *et al.*, 2013). SNP combination (rs3020314, rs1514348) has also been reported by Li (2017) to be associated with BC. Both SNPs rs3020314 and rs1514348 are in the ESR1 gene, which is widely believed to be associated with BC susceptibility (Dunning *et al.*, 2009; Lipphardt *et al.*, 2013). rs2017591 is in the STS gene (located on the X Chromosome), which is related to the biosynthetic pathway for estrogen.

Three higher-risk 3-order SNP combinations (see Supplementary Table S12) are found, all of which contain SNPs rs3020314 and rs2017591 and meet the significance level with the G-test P-value 0.05 and χ^2 P-value 1E−7. In the three 3-order SNP combinations (rs3020314, rs2077647, rs2017591) contains 20 higher-risk genotype combinations that account for 54.58% of case samples, (rs3020314, rs1514348, rs2017591) contains 19 higher-risk genotype combinations that account for 48.64% of case samples, and (rs3020314, rs660149, rs2017591) contains 17 higher-risk genotype combinations that account for 54.34% of case samples. For detailed results, see sheet BC of Supplementary File S2.xls. The 3-order SNP combination (rs3020314, rs1514348, rs2017591) has been reported to be associated with BC by Li (2017). In Supplementary Table S13, the top six SNPs with individual effects are presented. In this experiment, no 4- or 5-order SNP combinations were found to meet the significance level of a G-test P-value of 0.05.

4 Discussion

In this work, we present a new SIO method dedicated to the detection of high-order SNP interactions. The main goal of the MP-HS-DHSI is to address two key difficulties—(i) combinatorial explosion of SNPs and (ii) diversity of disease epistasis models—and it is a great computational challenge to explore high-order SNP interactions in a superhigh-dimensional search space. For an SIO method, heuristic factors are very important for exploring the SNP interactions that have a strong association with disease status. However, the method is generally blind and completely random for the search if heuristic factors are nonexistent or cannot be found.

Thus far, SIO methods have achieved good results for detecting high-order SNP interactions with marginal effects, which is because the marginal effect of one or more SNPs provides important heuristic factors for the search of high-order SNP interactions (Jing and Shen, 2015; Tuo *et al.*, 2016; Yang *et al.*, 2013, 2017). However, to the best of the authors' knowledge, these methods do not work well for the detection of SNP interactions without marginal effects; because the adopted discrimination methods (such as BNs, information theory, logistical regression and machine learning) cannot obtain an effective heuristic factor, a disease-causing SNP combination is usually an isolated point in a landscape of SNP combinations.

In this paper, we propose an HS that employs multiple populations and multiple criteria (multiple objective functions) to improve the search ability and discrimination ability. Multiple criteria are intended to be suitable for discriminating diverse disease models. Multiple populations can search independently for enhancing the global search ability of HS. A novel ND-JE is presented. This measure is used as heuristic factor to guide the HS to find clues for

disease-causing SNP combinations. The HS decreases the time complexity from an exhaustive method $O(n^k)$ to $O(k \times MaxFEs)$.

Although the experimental results indicate that the proposed MP-HS-DHSI is effective and efficient compared with SIO methods, this method still cannot ensure the identification of true disease-causing SNP combinations. For some DNMEs, the performance of the MP-HS-DHSI is still not sufficient. With an increasing number of SNPs and epistasis order k,*MaxFEs* needs to be set to a large value for detecting SNP interactions. Experimental results from real AMD and BC show that the proposed algorithm can quickly find high-order SNP combinations related to complex diseases, and the SNPs with top significant individual effects also tend to have more significant high-order interactions than the SNPs with low significant individual effects (see Table S10–S13).

The MP-HS-DHSI has made progress in detecting high-order SNP interactions based on SIO algorithms. There is still room for accelerating the search speed and improving the discrimination ability for a diversity of disease models. In the future, we should develop more effective heuristic factors for finding various disease models (Tuo *et al.*, 2019) and explore functional and biological information from gene databases, protein interaction databases and biological pathways (Shang *et al.*, 2019). An evaluation function with heuristic factors is an important way for SIO to discover genome-wide disease-causing SNP combinations. In addition, SIO can be used to analyze low-frequency and rare variants in GWAS (Tam *et al.*, 2019).

Funding

This work was partially supported by the Natural Science Foundation of China [61571341]; and the Ministry of Education of Humanities and Social Science Project of China [19YJCZH148].

Conflict of Interest: none declared.

References

Aflakparast,M. *et al.* (2014) Cuckoo search epitasis: a new method for exploring significant genetic interactions. *Heredity*, **112**, 666–674.

Alia,O.M. (2017) Dynamic relocation of mobile base station in wireless sensor networks using a cluster-based harmony search algorithm. *Inf. Sci.*, **385–386**, 76–95.

Blum,C. (2005) Ant colony optimization: introduction and recent trends. *Phys. Life Rev.*, **2**, 353–373.

Bush,W.S. *et al.* (2008) Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics*, **9**, 238.

Chuang,L.Y. *et al.* (2012) An improved PSO algorithm for generating protective SNP barcodes in breast cancer. *PLoS One*, **7**, e37018.

Crow,J.H. (1999) Weinberg and language impediments. *Genetics*, **152**, 821–825.

Das,S. *et al.* (2011) Exploratory power of the harmony search algorithm: analysis and improvements for global numerical optimization. *IEEE Trans. Syst. Man Cybern. B*, **41**, 89–106.

Dunning,A.M. *et al.* (2009) Association of ESR1 gene tagging SNPs with breast cancer risk. *Hum. Mol. Genet.*, **18**, 1131–1139.

Eberhart,R.C. and Kennedy,J. (1995) A new optimizer using particle swarm theory. In: *Proceedings of the Sixth International Symposium on Micromachine and Human Science*, Nagoya: IEEE, Japan, pp. 39–43.

Geem,Z.W. *et al.* (2001) A new heuristic optimization algorithm: harmony search. *Simulation*, **76**, 60–68.

Gola,D. *et al.* (2016) A roadmap to multifactor dimensionality reduction methods. *Brief. Bioinform.*, **17**, 293–308.

Goudey,B. *et al.* (2015) High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies. *Health Inform. Sci. Syst.*, **3**, S3.

Guo,X. *et al.* (2014) Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. *BMC Bioinformatics*, **15**, 102.

Guo,Y. *et al.* (2019) Epi-GTBN: an approach of epistasis mining based on genetic Tabu algorithm and Bayesian network. *BMC Bioinformatics*, **20**, 444.

Gyenesei,A. *et al.* (2012) High-throughput analysis of epistasis in genome-wide association studies with BiForce. *Bioinformatics*, **28**, 1957–1964.

Himmelstein,D.S. *et al.* (2011) Evolving hard problems: generating human genetics datasets with a complex etiology. *BioData Min.*, **4**, 21.

Hoey,J. (2012) The two-way likelihood ratio (G) test and comparison to two-way chi squared test. *Statistics*, **1–6**. http://arxiv.org/abs/1206.4881v2.

Jing,P.J. and Shen,H.B. (2015) MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics*, **31**, 634–641.

Klein,R.J. et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Kong,X.Y. *et al.* (2015) A simplified binary harmony search algorithm for large scale 0–1 knapsack problem. *Expert Syst. Appl.*, **42**, 5337–5355.

Li,X. (2017) A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization. *Bioinformatics*, **33**, 2829–2836.

Li,X. and Jiang,W. (2017) Method for generating multiple risky barcodes of complex diseases using ant colony algorithm. *Theor. Biol. Med. Model.*, **14**, 4.

Lin,J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theor.*, **37**, 145–151.

Lin,W.Y. and Lee,W.C. (2010) Incorporating prior knowledge to facilitate discoveries in a genome-wide association study on age-related macular degeneration. *BMC Res. Notes*, **3**, 26.

Lipphardt,M.F. *et al.* (2013) ESR1 single nucleotide polymorphisms predict breast cancer susceptibility in the central European Caucasian population. *Int. J. Clin. Exp. Med.*, **6**, 282–288.

McDonald,J.H. (2014) *G-test of Goodness-of-fit. Handbook of Biological Statistics*, 3rd edn. Sparky House Publishing, Baltimore, MD, pp. 53–58.

Mitchell,M. (1996) *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, ISBN 9780585030944.

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

Namkung,J. *et al.* (2009) New evaluation measures for multifactor dimensionality reduction classifiers in gene–gene interaction analysis. *Bioinformatics*, **25**, 338–345.

Neapolitan,R.E. (2004) *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, NJ.

Neyman,J. and Pearson,E.S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: part 1. *Biometrika*, **20A**, 175–240.

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.* (2015) An improved opposition-based learning particle swarm optimization for the detection of SNP–SNP interactions. *BioMed Res. Int.*, **2015**, 1–12.

Shang,J. *et al.* (2019) A review of ant colony optimization-based methods for detecting epistatic interactions. *IEEE Access*, **7**, 13497–13509.

Storn,R. and Price,K. (1995) *Differential Evolution—A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces*, Vol. **3**. ICSI, Berkeley, CA.

Sun,Y. *et al.* (2017) epiACO—a method for identifying epistasis based on ant colony optimization algorithm. *BioData Min.*, **10**, 23.

Tam,V. *et al.* (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*, **20**, 467–484.

Tuo,S.H. (2018) FDHE-IW: a fast approach for detecting high-order epistasis in genome-wide case–control studies. *Genes*, **9**, 435.

Tuo,S.H. and He,H. (2019) DEaf-MOPS/D: an improved differential evolution algorithm for solving complex multi-objective portfolio selection problems based on decomposition. *Econ. Comput. Econ. Cybern. Stud. Res.*, **53**, 151–167.

Tuo,S.H. *et al.* (2016) FHSA-SED: two-locus model detection for genome-wide association study with harmony search algorithm. *PLoS One*, **11**, e0150669.

Tuo,S.H. *et al.* (2017) Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations. *Sci. Rep.*, **7**, 11529.

Tuo,S.H. *et al.* (2019) A survey on swarm intelligence search methods dedicated to detection of high-order SNP interactions. *IEEE Access*, **7**, 162229–162244.

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

Velez,D.R. *et al.* (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.*, **31**, 306–315.

Visweswaran,S. *et al*. (2009) A Bayesian method for identifying genetic interactions[C]//AMIA Annual Symposium Proceedings. *Am. Med. Inform. Assoc*., **2009**, 673.

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,J. *et al*. (2015) A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC Genomics*, **16**, 1011.

Yang,C. *et al*. (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, **25**, 504–511.

Yang,C.H. *et al*. (2013) Evaluation of breast cancer susceptibility using improved genetic algorithms to generate genotype SNP barcodes. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **10**, 361–371.

Yang,C.H. *et al*. (2017) CMDR based differential evolution identify the epistatic interaction in genome-wide association studies. *Bioinformatics*, **33**, 2354–2362.

Yung,L.S. *et al*. (2011) GBOOST: a GPU-based tool for detecting gene–gene interactions in genome-wide case control studies. *Bioinformatics*, **27**, 1309–1310.

Zhang,T.H. and Geem,Z.W. (2019) Review of harmony search with respect to algorithm structure. *Swarm Evol. Comput*., **48**, 31–43.

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