IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 31, 2018, accepted January 16, 2019, date of publication January 24, 2019, date of current version February 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894676

A Review of Ant Colony Optimization Based Methods for Detecting Epistatic Interactions

JUNLIANG SHANG[1,2], (Member, IEEE), XUAN WANG[2], XIAOYANG WU[3], YINGXIA SUN[2], QIAN DING[2], JIN-XING LIU[2], (Member, IEEE), AND HONGHAI ZHANG[3]

[1]School of Statistics, Qufu Normal University, Qufu 273165, China
[2]School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China
[3]College of Life Science, Qufu Normal University, Qufu 273165, China

Corresponding author: Honghai Zhang (zhanghonghai67@126.com)

This work was supported in part by the China Postdoctoral Science Foundation under Grant 2018M642635, and in part by the National Science Foundation of China under Grant 61502272, Grant 31872242, Grant 61872220, and Grant 61572284.

ABSTRACT Detection of epistatic interactions, which are referred to as nonlinear interactive effects of single nucleotide polymorphisms (SNPs), is increasingly being recognized as an important route in capturing the underlying genetic causes of complex diseases. Its methodological and computational challenges have been well understood, and many methods also have been proposed from different perspectives. Among them ant colony optimization (ACO)-based methods are promising due to their controllable time complexities, heuristic positive feedback search, and high detection power. Nevertheless, there is no comprehensive overview of them so far. This paper, therefore, provides a systematic review of 25 ACO-based epistasis detection methods. First, the generic ACO algorithm, as well as how it is applied to detect epistatic interactions, is briefly described. Then, an in-depth review of ACO-based methods for detecting epistatic interactions is discussed from four aspects, including path selection strategies, pheromone updating rules, fitness functions, and two-stage designs. Finally, this paper analyzes the strengths and limitations of involved methods, provides guidelines for applying them, and gives several views on the future directions of epistasis detection methods.

INDEX TERMS Ant colony optimization (ACO), epistatic interactions, single nucleotide polymorphisms (SNPs), heuristic information, genome-wide association studies (GWAS).

I. INTRODUCTION

Genome-Wide Association Studies (GWAS) have become routine strategies in investigating the underlying genetic mechanisms of complex diseases during the past decade. Some promising methods of GWAS have been proposed [1], resulting in hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) speculated to associate with complex diseases being reported. Nevertheless, these SNPs can only explain a small proportion of genetic causes of complex diseases, and the mystery of ''missing heritability'' needs to be further unraveled [1]–[3]. This is mainly because the proposed GWAS methods normally assume SNPs having independent effects on the phenotype and therefore use the single-SNP tests, in which SNPs are tested one by one for association with the phenotype [4]. Although they can successfully identify multiple SNPs, only additive effects of these SNPs are considered, ignoring the nonlinear interactive effects of SNPs on disease susceptibility [5].

It has been widely accepted that the detection of epistatic interactions, also known as epistasis, SNP-SNP interactions, or gene-gene interactions, which are referred to as the nonlinear interactive effects of SNPs, is a compelling step forward from GWAS to better unravel the mystery of ''missing heritability'' [6]–[10]. Hence, a vast number of methods for the detection of epistasis have been proposed from different perspectives in recent years [11]. Among them, the most direct and simplest way to detect epistatic interactions is by exhaustively searching all combinations of SNPs within the data sets. Multifactor dimension reduction (MDR) [12] is the typical representative of exhaustive search methods, which has enjoyed great popularity in applications [13]. Its main idea is to reduce high dimensional SNP combination data to a one-dimensional variable by pooling multiple combination genotypes into high-risk and low-risk groups. Exhaustive search methods usually perform well on two-order interactions or small scale data sets.

Higher-order interactions or large scale data sets are not scalable due to high computational burden. Filtering methods select a subset of candidate SNPs, rather than the whole data set like exhaustive search methods, for the subsequent epistasis detection. Besides the apparent advantage of speed, they sometimes have greater detection power than exhaustive search methods because of much reduced multiple tests [7], [14], [15]. One of the most famous filter in GWAS is the Relief [16], [17], as well as its extensions and modifications [18], [19], which is capable of capturing SNP dependencies even in the absence of marginal effects by estimating SNP weights based on whether the nearest neighbor of a randomly selected sample from the same class and the one from the different class have the same or different genotypes. However, filtering methods could miss interacting SNPs in subsets since they are very sensitive to filters and thresholds. Recently, many swarm intelligence methods have been proposed to infer epistatic interactions [20]–[47]. These methods mimic collective behaviors of organisms which can jointly perform many complex tasks though each individual is very limited in its capability, and adopt some heuristics to avoid exhaustive search in initial data sets. Among them, ant colony optimization (ACO) based methods are promising due to their controllable time complexities, heuristic positive feedback search, and high detection power. Therefore several ACO based methods have been presented [23]–[47] and it is necessary to provide an in-depth review of them to analyze their strengths and limitations, provide guidelines for applying them, and give several views on future directions of the detection of epistatic interactions.

In fact, many reviews of epistasis detection methods have been reported [1], [3]–[11], [48]–[62]. They systematically summarized background, challenges, methods, and directions of epistasis detection, and therefore are essential for researchers to develop new methods and for users to properly apply them. However, though many promising ACO based methods have been presented for detecting epistatic interactions, there is no comprehensive overview of them yet except two summarizing them briefly in only one chapter [4], [11]. Niel *et al.* [4] presented the main strategies proposed to detect epistatic interactions. They classified ACO based methods into the group of non-exhaustive combinatorial optimization approaches, and discussed their operating principles, concluding that the positive feedback effect is an interesting feature of the algorithm and its main limitation is that many parameters require fine tuning. Uppu *et al.* [11] reviewed 7 groups of epistasis detection methods, including exhaustive search methods, random forests, neural networks, support vector machines, regression models, Bayesian approaches, and ACO approaches. For the group of ACO approaches, they provided a brief overview of 7 extensions and modifications of the generic ACO algorithm for detecting epistatic interactions [26], [31], [35], [37], [40], [43], [46].

The aim of this paper is to summarize 25 variations of ACO algorithm for detecting epistatic interactions. Firstly, the generic ACO algorithm and its application to detect epistatic interactions are described. Then, all ACO based methods for detecting epistatic interactions are discussed by categories, which are classified according to their core modifications, including path selection strategies, pheromone updating rules, fitness functions, and two-stage designs. Finally, their strengths and limitations are analyzed to provide guidelines for applying them, and several clues are given for future directions of methods for detecting epistatic interactions.

II. ANT COLONY OPTIMIZATION (ACO) ALGORITHM

A. BIOLOGICAL MODEL OF ACO

ACO algorithm, proposed by Dorigo *et al.* [63], [64], takes inspiration from the foraging behavior of some biological ant species. These ants explore an optimal path from nest to a food source by communicating with each other indirectly through releasing and perceiving pheromones along the path. The pheromones on paths gradually evaporate as time passes. The subsequent ants perceive the presence of pheromones and are more likely to follow the paths with higher pheromones, thereby creating a positive feedback and eventually most of the ants, if not all, are able to transport food to their nest in the optimal path. Figure 1.A shows the foraging behavior of biological ants. Specifically, at the beginning, ants randomly select paths to search for a food source because there are no pheromones on both paths (Figure 1.A1); later, ants on the short path firstly arrive at the food source and release pheromones on path that have passed (Figure 1.A2); despite pheromones continuously evaporate on both paths, the short path has relatively higher pheromone levels and the return ants select the short path with a higher probability (Figure 1.A3); as the number of iterations increases, the proportion of pheromones on the short path to those on the long path increases accordingly, resulting in more and more ants selecting the short paths (Figure 1.A4).

B. SNP DATA FOR EPISTASIS DETECTION

Mathematically, SNP sequences are generally mapped into two numerical matrices (Figure 1.B): one being SNP data matrix, and one being sample labels matrix. For the SNP data matrix, a row represents genotypes of a sample and a column represents a SNP. Genotypes of a sample are usually coded as 0, 1, 2, 3, corresponding to missing data, homozygous common genotype (e.g., *AA*), heterozygous genotype (e.g., *Aa* and *aA*), and homozygous minor genotype (e.g., *aa*) [2], [20], [65]. The sample labels matrix has only one column listing the binary phenotype of each sample, where 0 denotes control and 1 denotes case. There are, of course, other numerical representation forms of SNP sequences. For example, these two matrices can be merged into one, where the first or the last column saves the sample labels; the transposes of related matrices are sometimes used; genotypes and sample labels are coded as other numbers. In essential, these representation forms are consistent with the introduction one.

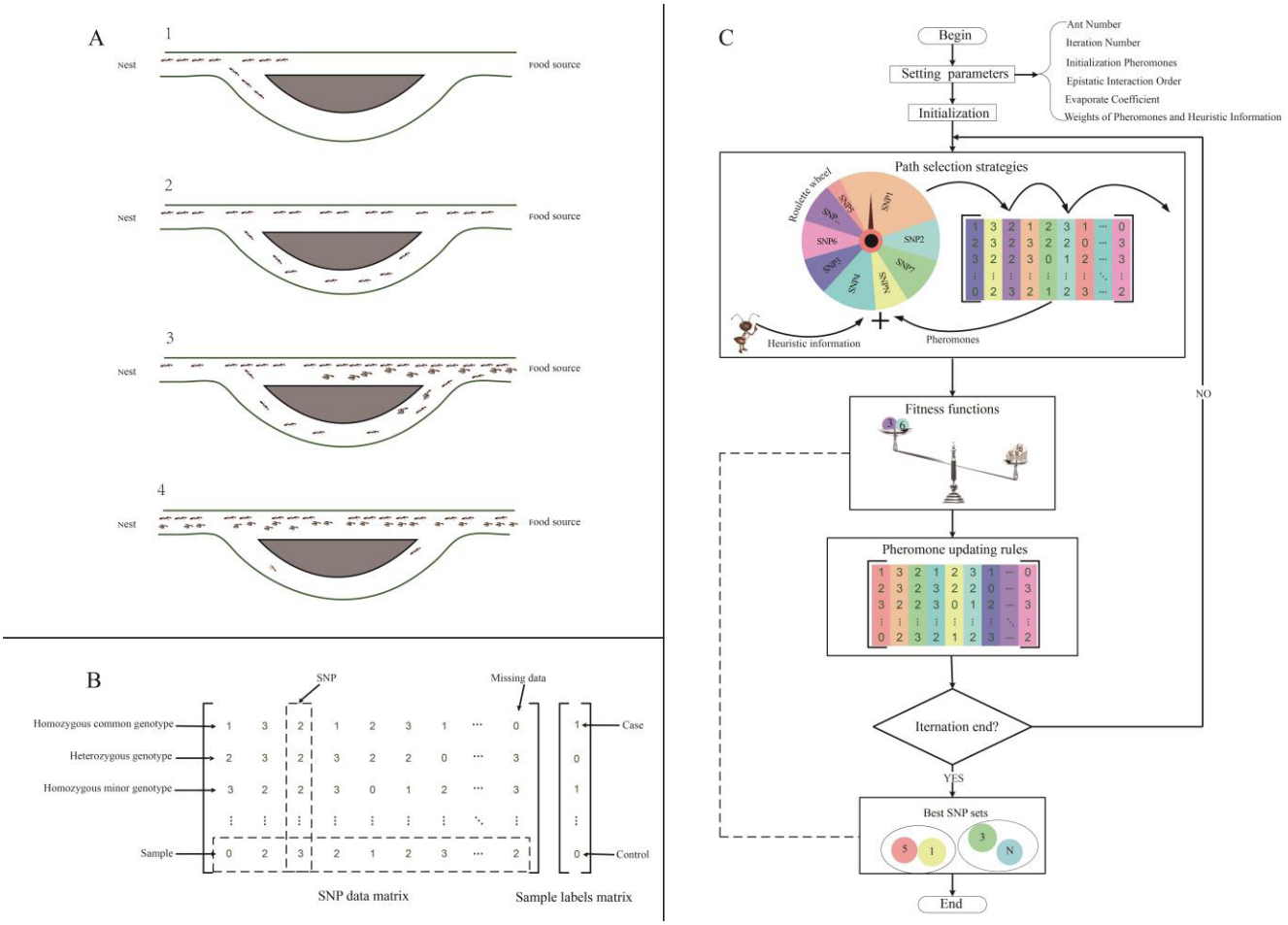


FIGURE 1. **The generic ACO algorithm for epistasis detection. Figure 1.A is the foraging behavior of biological ants. Figure 1.B is the SNP data matrix and the sample labels matrix. Figure 1.C is the flow chart of generic ACO algorithm on epistasis detection.**

Based on this SNP mapping, epistasis detection can be mathematically described as finding multiple SNP combinations to predict the phenotype as high as possible, where SNPs in each combination have nonlinear interactive effects rather than additive effects to the phenotype, and SNPs between combinations are not allowed to overlap with each other.

C. MATHEMATICAL MODELING OF ACO ALGORITHM ON EPISTASIS DETECTION

ACO algorithm is a particularly appropriate strategy for epistasis detection because of its simplicity and parallelization. Figure 1.C shows its mathematical modeling on epistasis detection, from which it is seen that path selection strategies, fitness functions, and pheromone updating rules are its highlights.

In the ACO algorithm, unlike biological ants, the artificial ants exchange information and select paths via a probability function determined by pheromones and heuristic information. Suppose the algorithm uses I ants and T iterations to infer K-order epistatic interactions from the data that having N SNPs genotyped with M samples. The path here is represented by a set of K SNPs. Specifically, the probability of ant i at iteration t selecting SNP X into its path is defined as

$$P_X^i(t) = \begin{cases} \dfrac{(\tau_X(t))^\alpha \cdot (\eta_X)^\beta}{\sum\limits_{j \notin Tabu_i(t); j \in [1,N]} (\tau_j(t))^\alpha \cdot (\eta_j)^\beta} & X \notin Tabu_i(t) \\ 0 & X \in Tabu_i(t), \end{cases} \tag{1}$$

where $i \in [1, I]$, $t \in [1, T]$, $X \in [1, N]$, $\tau_X(t)$ is the pheromones of SNP X at iteration t, η_X is the heuristic information of SNP X, $Tabu_i(t)$ is the already selected SNPs in ant i's path at iteration t, α and β are parameters that determine the weights of pheromones and heuristic information, respectively. Though heuristic information is a crucial component, it is difficult to get heuristic information since no prior knowledge is usually available. Therefore, it is normally set to 1. Besides, both α and β are normally set to 1, and initial pheromones of all SNPs are set to a constant τ_0.

Using such a roulette wheel selection strategy, each ant at iteration t can select a set of K SNPs. This SNP set is then evaluated by an employed fitness function to quantify its association effect to the phenotype. For the SNP set $S^i(t)$

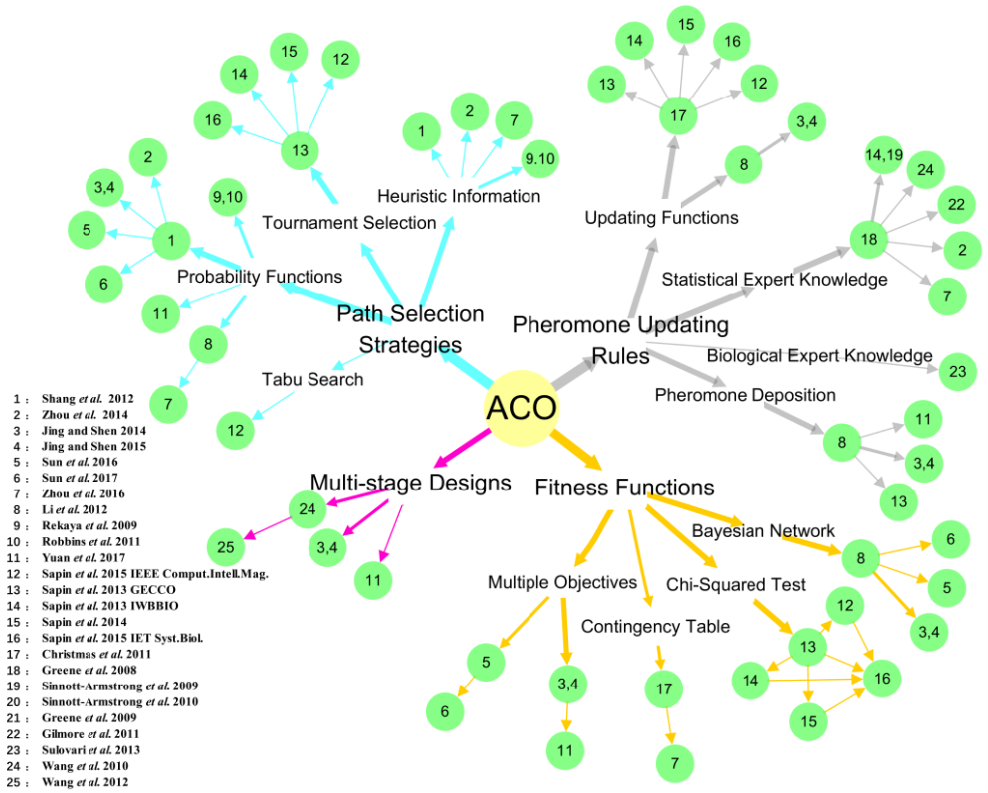


FIGURE 2. A full picture of discussed ACO based methods for detecting epistasis interactions.

that selected by ant i at iteration t, its fitness function value is denoted as $f\left(S^i(t)\right)$. Here, the higher the fitness function value, the stronger the association between the SNP set and the phenotype.

While all ants having selected and measured their paths at iteration t, the pheromones of each SNP X are updated according to the following formula,

$$\tau_X(t+1) = (1-\rho)\cdot\tau_X(t) + \Delta\tau_X(t), \tag{2}$$

where ρ is the evaporate coefficient between 0 and 1, $\Delta\tau_X(t)$ is the additional pheromones of SNP X contributed by ants whose paths contain the SNP X during the t iteration cycle, which can be written as

$$\Delta\tau_X(t) = \sum_{X\in S^j(t); j\in[1,I]} f\left(S^j(t)\right). \tag{3}$$

After completing the iteration process, several shortest paths, that is, best SNP sets, with their fitness function values being greater than a given threshold, are reported as epistatic interactions.

III. ACO BASED METHODS FOR DETECTING EPISTATIC INTERACTIONS

In total 25 ACO based methods for detecting epistatic interactions are discussed from four aspects, including path selection strategies, pheromone updating rules, fitness functions, and two-stage designs [23]–[47]. A full picture of these methods is shown in Figure 2, each leaf of which represents an ACO based method. It should be noted that most of them have more than one main modification and thus they could be discussed in multiple sections.

A. PATH SELECTION STRATEGIES

1) PROBABILITY FUNCTIONS

Shang *et al.* [23] presented AntMiner for the detection of epistatic interactions with different orders simultaneously. Besides the probability function of formula (1), AntMiner also introduced another probability function to avoid falling into local optimal solution, which was defined as

$$P_X^i(t) = \begin{cases} 1 & X = rand\left(N - Tabu_i(t)\right) \\ 0 & otherwise, \end{cases} \tag{4}$$

where $N - Tabu_i(t)$ is the feasible set in which SNPs are not yet visited by ant i at iteration t, $rand(\cdot)$ is a function to uniformly select a SNP from the given set. For each ant, these two probability functions are chosen randomly with a user-defined threshold. Specifically, the ant first generates a random value; if the value is greater than the threshold, the ant selects the probability function of formula (1), otherwise, formula (4). These two probability functions and the random selection mechanism increases the diversity of the search, and also has been adopted by other ACO methods, including

epiACO(Z) [24], MACOED [25], [26], IACO [27], and epiACO(S) [28].

Zhou *et al.* [29] developed a modified ACO method for detecting epistatic interactions, and implemented it on a hadoop cluster utilizing Google's MapReduce framework. Similarly, they provided another probability function to greedily find the best SNP, which was defined as

$$P^i_{X\to Y}(t)=\begin{cases}1 & Y=\arg\max_{j\notin Tabu_i(t);j\in[1,N]}\left\{\left(\tau_{X\to j}(t)\right)^{\alpha}\cdot\left(\eta_{X\to j}\right)^{\beta}\right\}\\ 0 & otherwise,\end{cases}\tag{5}$$

where $P^i_{X\to Y}(t)$ is the probability of ant i from SNP X to SNP Y at iteration t, $\tau_{X\to j}(t)$ is the pheromones on the path from SNP X to SNP j at iteration t, and $\eta_{X\to j}$ is heuristic information of the path from SNP X to SNP j. It is seen that in this method the pheromones are deposited on connections between SNPs rather than on SNPs, detailed descriptions of which can be seen later. This probability function is based on the greedy search and always selects the best SNP. To increase the diversity of the search, the probability function of formula (1) is also considered in the method. Besides this method, another method [30] also used such path selection strategy to select path.

Rekaya *et al.* [31], [32] provided a probability function with two-layer pheromones, which was defined as

$$P^i_X(t)=\begin{cases}\dfrac{(\tau_X(t))^{\alpha}\cdot(\tau 2_X(t))^{\alpha'}\cdot(\eta_X)^{\beta}}{\sum_{j\notin Tabu_i(t);j\in[1,N]}\left(\tau_j(t)\right)^{\alpha}\cdot\left(\tau 2_j(t)\right)^{\alpha'}\cdot\left(\eta_j\right)^{\beta}} & X\notin Tabu_i(t)\\ 0 & X\in Tabu_i(t),\end{cases}\tag{6}$$

where $\tau_X(t)$ and $\tau 2_X(t)$ are the first and the second layer pheromones of SNP X at iteration t. This two-layer pheromones design is to overcome the limitation of pheromones of a SNP depending on not only its association strength to the phenotype but also times that ants selected.

Yuan *et al.* [33] presented a framework FAACOSE with fast adaptive ACO algorithm to detect epistatic interactions. For the modified algorithm, the weight of pheromones α in the formula (1) was adaptively adjusted according to the following formula,

$$\alpha(t+n)=\begin{cases}g_1\alpha(t) & \alpha(t)\le\alpha_m\\ \alpha_m & otherwise,\end{cases}\tag{7}$$

where g_1 is a user-specified parameter greater than 1, n is a predefined iteration window and $\alpha_m \le 5$. If the optimal SNP set is not changed after n iterations, α should be adjusted according to formula (7). The authors said that with the increase of α, FAACOSE can jump out of local optimal solution and had ability to search for global optimal solution [33]. However, an example of two SNPs with their pheromones being $(2, 3)$, selection probabilities are $(0.4, 0.6)$, $(0.308, 0.692)$ and $(0.229, 0.771)$ corresponding to α being 1, 2, and 3, respectively, resulting in that with the increase of α, the selection probability of the SNP with pheromones being 3 becomes higher, much easier for FAACOSE to fall into local optimal solution.

2) TABU SEARCH

Sapin *et al.* [34] incorporated tabu search into path selection strategy to prevent the modified ACO algorithm from continually selecting SNPs displaying strong main effects. The modification of this ACO_Tabu method can be described as follows. First, let the modified ACO algorithm run *num* iterations and the SNP *snp*1 with the highest pheromones is picked out, where *num* is a user specified parameter. Second, all the 2-SNP combinations of *snp*1 and the remaining SNPs are evaluated and the best combinations are recorded. Third, the SNP *snp*1 is removed from the data set. This process is repeated until end of the run, which allows ACO_Tabu to concentrate on SNP combinations displaying small marginal effects.

3) HEURISTIC INFORMATION

The major innovation of AntMiner [23] is that heuristic information was incorporated into path selection strategy to direct the search, which was defined as

$$\eta_X = SMUC(X, C) + SURF(X, C),\tag{8}$$

where $SMUC(X, C)$ is the SymMetrical UnCertainty (SMUC) score between SNP X and the phenotype C, $SURF(X, C)$ is the Spatially Uniform ReliefF (SURF) score between X and C. SMUC [66] is the normalized mutual information that can effectively measure the dependence between a SNP and the phenotype, which was defined as

$$SMUC(X, C) = 2\cdot\frac{MI(X; C)}{H(X)+H(C)},\tag{9}$$

where $H(X)$ is the entropy of X, $MI(X; C)$ is the mutual information between X and C. SURF [19], as a member of Relief family, is able to capture interacting SNPs even in the absence of marginal effects. Because SMUC focuses on SNPs with strong main effects and SURF focuses on interacting SNPs with weak or even no main effects, combination of them is promising in finding various types of SNP effects to the phenotype.

Similarly, Zhou *et al.* [24] developed an epistasis detection method epiACO by incorporating heuristic information into path selection strategy. Since two methods that respectively proposed by Zhou *et al.* [24] and Sun *el al.* [28] have the same name epiACO, in order to distinguish them, they are referred to as epiACO(Z) and epiACO(S) here. For epiACO(Z), the heuristic information was defined as

$$\eta_X = TuRF(X, C),\tag{10}$$

where $TuRF(X, C)$ is the Tuned ReliefF (TuRF) [18] score between SNP X and the phenotype C. The TuRF, as another member of Relief family, is capable of capturing SNPs that predict the phenotype primarily through interactions with other SNPs. Though requiring more computational costs,

the TuRF improves its performance when the data contain a large number of noise SNPs by iterating the ReliefF and deleting SNPs with the lowest ReliefF scores at each iteration.

For the method proposed by Zhou *et al.* [29], its heuristic information, they called it distance, was got by the following formula,

$$\eta_{X\to j} = \frac{2}{f_1(\{X,j\})+1} \tag{11}$$

where $f_1(\{X,j\})$ is the fitness function value of SNP combination $\{X,j\}$. This fitness function will be descripted in detail later.

4) TOURNAMENT SELECTION

For most of discussed methods, the roulette wheel selection strategy is used for ants selecting SNPs into their paths. However, for large scale data sets, especially those in GWAS, the roulette wheel selection strategy slows down the convergence speed since each SNP only owns a very small segment of the roulette wheel, resulting in this strategy tending to a random selection. In order to overcome this limitation, Sapin *et al*. [34]–[38] adopted tournament selection strategy to guide ants selecting SNPs. Firstly, *nts* SNPs are randomly selected to form a tournament, where *nts* is the tournament size specified by users, which can be adjusted to alter the convergence speed. Secondly, one SNP with the highest pheromones in the tournament is selected as part of its path. Thirdly, this process repeats K times to form a path, i.e., a SNP set. The tournament selection strategy incorporates elements of random selection and a bias towards those SNPs with higher pheromones. This biased random nature ensures a balance between the exploration and exploitation [37], [38]. Compared with the roulette wheel selection strategy, the tournament selection strategy has several merits, including lower time complexity, easy to parallelize, not easy to fall into local optimal solution, not need to sort, and so on, and was proven to be better for high dimension problems [67].

B. PHEROMONE UPDATING RULES

1) PHEROMONE DEPOSITION

For most of discussed methods here, pheromones are normally deposited on SNPs, except the methods [25], [26], [29], [30], [33], which deposit pheromones on connections between SNPs. Specifically, pheromones are stored as a square matrix, whose dimensionality is equal to the SNP number N, to reflect association strengths between two-SNP combinations and the phenotype. This means that formulas (1), (2) and (3) should be slightly adjusted: $P^i_{X\to Y}(t)$, $\tau_{X\to Y}$, $\tau_{X\to j}$, $\eta_{X\to j}$, and $\Delta\tau_{X\to Y}(t)$ instead of $P^i_X(t)$, τ_X, τ_j, η_X, and $\Delta\tau_X(t)$ respectively. Though this modification seems simple and easy to understand, there are in total $N(N-1)/2$ elements in square matrix that need to be deposited by ants. In real applications, especially GWAS, this astronomical number leads the matrix extremely sparse, thus making it difficult to converge to optimal solution steadily. Sapin *et al*. [38] gave a more detailed explanation. The detection of epistatic interactions can be described as a subset problem since there is no concept of order between SNPs. The ACO algorithm can deal with not only the ordering problems but also the subset problems. They differ in the way that pheromones are deposited. For ordering problems, the connections between elements receive pheromones, whereas the components themselves receive pheromones in subset problems.

2) UPDATING FUNCTIONS

The MACOED [25], [26] and the method [30] employed a variant of formula (2) to update pheromones,

$$\tau_X(t+1) = (1-\rho)\cdot\tau_X(t) + \rho\cdot\Delta\tau_X(t). \tag{12}$$

As a matter of fact, this variant, as well as another frequently used variant in methods [34]–[39], which was defined as

$$\tau_X(t+1) = (1-\rho)\cdot(\tau_X(t) + \Delta\tau_X(t)), \tag{13}$$

is equivalent to formula (2), just a range adjustment of additional pheromones.

Recently, Sun *el al.* [28] developed a method epiACO(S) to identify epistatic interactions. Besides that path selection strategy like AntMiner [23], both fitness function and memory based strategy like IACO [27], its highlight is the pheromone updating rule,

$$\tau_X(t+1) = (1-\rho)\cdot\tau_X(t) + \Delta\tau_X(t) + \Delta\tau^*_X(t), \tag{14}$$

where $\Delta\tau^*_X(t)$ is reward pheromones for the SNPs that belong to the candidate SNP sets at each iteration, defined as

$$\Delta\tau^*_X(t) = \sum_{X\in S^j(t); j\in[1,I]; S^j(t)\in CS(t)} f_{Svalue}\left(S^j(t)\right), \tag{15}$$

where $CS(t)$ is the candidate SNP sets at iteration t, f_{Svalue} is the fitness function.

Rekaya *et al.* [31], [32] provided a path selection strategy with two-layer pheromones. The first layer pheromones are updated according to formula (2). The second layer pheromones are updated using the following formula,

$$\tau 2_X(t+1) = \frac{t\cdot\tau 2_X(t) + \Delta\tau 2_X(t)}{t+ns}, \tag{16}$$

where $\Delta\tau 2_X(t)$ is the change in pheromones of SNP X based on the sum of fitness function values of all SNP sets containing genotypes of SNP X; ns is times of SNP X being selected at iteration t.

For MACOED [25], [26], two fitness functions ($f_{K2\log}$ and f_{AIC}) and a Pareto optimality approach [68] were used to update pheromones. Based on the Pareto optimality approach, a SNP set $S^i(t)$ is considered to dominate another SNP set $S^j(t)$, or $S^i(t)$ is a non-dominated solution and accordingly $S^j(t)$ is a dominated solution, only if they satisfy the following two conditions,

$$\begin{cases} f_{K2\log}\left(S^i(t)\right) \le f_{K2\log}\left(S^j(t)\right) \,\&\, f_{AIC}\left(S^i(t)\right) \le f_{AIC}\left(S^j(t)\right) \\ f_{K2\log}\left(S^i(t)\right) < f_{K2\log}\left(S^j(t)\right) \,|\, f_{AIC}\left(S^i(t)\right) < f_{AIC}\left(S^j(t)\right). \end{cases} \tag{17}$$

Therefore, at each iteration, all SNP sets that collected by ants can be divided into two groups: a non-dominated group and a dominated group. Pheromones are only updated by SNP sets in the non-dominated group,

$$\Delta\tau_{X\to Y}(t) = \sum_{\{X,Y\}\in S^k(t);k\in[1,I];S^k(t)\in NDG} \delta, \quad (18)$$

where X and Y are neighbor SNPs in the set, δ is a user-specified weight for SNP sets in the non-dominated group *NDG*. This Pareto optimality approach and the multiple fitness functions strategy were also employed by FAACOSE [33].

For FAACOSE [33], it adopted an adaptive evaporate coefficient of pheromones to optimize the pheromone updating rule. The evaporate coefficient ρ is used to balance the effects of deposited and additional pheromones. The small ρ leads to deposited pheromones dominating the search process and therefore the method is easily to fall into local optimal solution, while the large ρ leads to additional pheromones dominating the search process and therefore the method tends to be a stochastic one. The adaptive evaporate coefficient of FAACOSE was defined as

$$\rho(t+n) = \begin{cases} g_2\rho(t) & \rho(t) \le \rho_m \\ \rho_m & otherwise, \end{cases} \quad (19)$$

where g_2 is a user-specified parameter greater than 1, n is the predefined iteration window and ρ_m is the control value of evaporate coefficient. If the optimal SNP set is not changed after n iterations, the weight of pheromones should be adjusted.

3) STATISTICAL EXPERT KNOWLEDGE

Except being incorporated into path selection strategies to guide the ACO search, heuristic information, known as expert knowledge, can also be integrated into pheromone updating rules.

Greene *et al.* [40] presented an expert knowledge guided ACO method for inferring epistatic interactions, and showed that it was successful when expert knowledge was supplied through the pheromone updating rule to prevent single SNPs dominating the search space. More specifically, expert knowledge is included as additional pheromones $\Delta\tau_X(t)$ of SNP X contributed by ants whose paths contain the SNP X at iteration t according to the following formula,

$$\Delta\tau_X(t) = \sum_{X\in S^j(t), j\in[1,I]} \left(MDR\left(S^j(t)\right)\right)^a \cdot (E_X)^b, \quad (20)$$

where $S^j(t)$ is the SNP set that selected by ant j at iteration t, $MDR\left(S^j(t)\right)$ is the MDR classification accuracy of $S^j(t)$, E_X is the expert knowledge of SNP X, Both a and b are coefficients that determine the weights of $MDR\left(S^j(t)\right)$ and E_X. In the method, TuRF [18] scores were used as the expert knowledge. i.e.,

$$E_X = TuRF(X, C). \quad (21)$$

For simplicity, this method is called TuRF-ACO in subsequent discussions for short. Later, the TuRF-ACO was implemented on graphics processing units (GPUs) [41], [42] to further improve its performance. The TuRF-ACO reduces computation costs by testing small portions of the data effectively but this presents a tradeoff between computation time and the portion of the search space. By using GPUs, the portion of the search space can be increased while significantly reducing computation costs.

Greene *et al.* [43] modified TuRF-ACO by using selection probabilities rather than TuRF scores as its expert knowledge. Frist, SNPs are sorted according to their ascending TuRF scores. Second, TuRF scores of SNPs are normalized by Min-Max normalization method,

$$F_{r_X} = \frac{E_{r_X} - E_1}{E_N - E_1}. \quad (22)$$

where $r_X \in [1, N]$ is the index of SNP X in the sorted list. Third, an exponential probability selection function is developed to transform normalized TuRF scores into selection probabilities,

$$EF\left(E_{r_X}\right) = \frac{\theta^{-F_{r_X}}}{\sum_{j=1}^{N}\theta^{-F_j}}, \quad (23)$$

where $\theta \in (0, 1]$ is a user-adjustable parameter, which facilitates users to control scaling of TuRF scores to selection probabilities. The lower the value of θ, the more likely that SNPs with high expert knowledge scores are selected over those with low scores. In a word, E_X in formula (20) is replaced by $EF\left(E_{r_X}\right)$ to improve the performance of TuRF-ACO.

Gilmore *et al.* [44] developed other three expert knowledge scaling strategies for TuRF-ACO, which were described as linear fitness (*LF*), linear rank (*LR*), and exponential rank (*ER*),

$$LF\left(E_{r_X}\right) = \frac{F_{r_X}}{\sum_{j=1}^{N}F_j}, \quad (24)$$

$$LR\left(E_{r_X}\right) = \frac{L_{r_X}}{\sum_{j=1}^{N}L_j}, \quad (25)$$

$$ER\left(E_{r_X}\right) = \frac{\theta^{-L_{r_X}}}{\sum_{j=1}^{N}\theta^{-L_j}}, \quad (26)$$

where L_{r_X} is also a normalized value of E_{r_X}, defined as

$$L_{r_X} = \frac{r_X - 1}{N - 1}. \quad (27)$$

It is seen that formula (22) respects the interval between TuRF scores, but formula (27) only uses the ranking of the SNPs.

In epiACO(Z) [24], pheromones of each SNP were updated according to both the MDR classification accuracy and

B-statistic [69] score of involved SNP sets. The pheromone updating rule is similar with the formula (20) except using $B\left(S^j(t)\right)$ instead of E_X, where $B\left(S^j(t)\right)$ is the B-statistic score of SNP set $S^j(t)$. The B-statistic uses a mixture distribution to accommodate the possibilities that SNPs in the controls may or may not be in linkage equilibrium, which is more powerful than the standard *chi*-squared statistic for measuring epistatic interactions [69]. Similarly, Zhou *et al.* [29] also updated pheromones of each SNP using the formula (20) except using $SMUC(X, C) + SURF(X, C)$ instead of E_X.

4) BIOLOGICAL EXPERT KNOWLEDGE

Sulovari *et al.* [45] presented an expert knowledge guided ACO method for detecting epistatic interactions by including biological expert knowledge as additional pheromones. The bright spot of the method is that it generates biological expert knowledge from a network of gene-gene interactions produced by a literature mining software, Pathway Studio [70]. The Pathway Studio is developed for navigation and analysis of biological pathways. In the method, for a given real SNP data set, all genes corresponding to these SNPs are queried by Pathway Studio, result of which is a network of gene-gene interactions. The number of connections for each gene is counted as expert knowledge scores of its corresponding SNPs under the hypothesis that SNPs belonging to genes with many interactions are more important to predict the phenotype than those with less interactions. This method is a preliminary exploration of how expert knowledge can be obtained from real biological information, which is a trend for further epistasis detection. Nevertheless, the hypothesis is not always correct, especially for SNPs displaying strong interactive effects with weak or even no marginal effects. Besides, many SNPs are in the inter-genic regions, which cannot be handled by the method.

C. FITNESS FUNCTIONS

1) CONTINGENCY TABLE

Christmas *et al.* [39] used the basic ACO algorithm to discover epistatic interactions associated with type 2 diabetes, and demonstrated that it was both accurate and computationally tractable on large scale data sets. They introduced two implementations based on different SNP data matrices derived from Fig.1B. For each SNP, three genotypes are recoded according to their descending frequencies: the most common genotype with the highest frequency being coded as 0, other two uncommon genotypes being coded as 1 in the first SNP data matrix and being coded as 1 and 2 in the second SNP data matrix, respectively. For ant i at iteration t, once a path $S^i(t)$, that is, a set of SNPs or SNP/uncommon genotypes in two implementations, has been selected, samples are divided into four classes. Controls and cases that possess the uncommon genotypes for all SNPs in the set are true positives (TP) and false positives (FP) respectively, Other controls and cases are false negatives (FN) and true negatives (TN) accordingly. Hence for each path a contingency table can be constructed, based on which three fitness functions (f_1, f_2, f_3) were proposed,

$$f_1\left(S^i(t)\right) = \frac{2\left(TP \cdot TN - FN \cdot FP\right)}{N^2}, \tag{28}$$

$$f_2\left(S^i(t)\right) = \frac{TP^2}{FN \cdot FP} - 1, \tag{29}$$

$$f_3\left(S^i(t)\right) = \frac{TP}{TP + FN}. \tag{30}$$

2) CHI-SQUARED TEST

Sapin *et al*. [35] used a*chi*-squared test based fitness function to measure relationship between a combination of two SNPs and the phenotype. First, for each of 9 genotypes in the combination, positive samples are those having such genotype and others are negative samples. Therefore, the numbers of positive and negative controls, as well as the numbers of positive and negative cases, are determined. Second, their expected values are calculated. Third, a *chi*-squared score is then computed with these observed and expected values. Fourth, the largest one among 9 *chi*-squared scores is considered as the final score. In the first step, each combination genotype is in fact the logical "AND" between genotypes of involved SNPs. Then, 4 logical operations ("AND", "OR", "AND NOT", and "XOR") between two SNPs is considered [34], [36]. For each combination genotype, 4 logical interactions are tested to discriminate positives and negatives in the first step. Finally, the largest one among 4×9 *chi*-squared scores is considered as the final score. Recently, Sapin *et al*. [37] further modified ACO algorithm by using the decision tree or the contingency table to obtain a fitness function score. For the decision tree, each ant at each iteration selects a set with 4 SNPs, which are used to create decision trees: the first SNP being the root with its three branches (homozygous common genotype, heterozygous genotype, and homozygous minor genotype) linking other three SNPs in turn, so producing 9 leaves which are associated with a phenotypic variable. Hence, 2^9 decision trees are created for each ant at each iteration, each of which classifies samples into 4 categories, yielding a *chi*-squared score. Finally, the largest one among 512 scores is considered as the fitness function score. For the contingency table, two SNPs are selected by each ant at each iteration, resulting in 9 combination genotypes, each of which is assigned a phenotypic variable. Similarly, the largest one among these 2^9*chi*-squared scores is considered as the fitness function score. In their newly paper [38], these models, including logical operation, decision tree, and contingency table, were described in more detail.

3) BAYESIAN NETWORK

Li *et al.* [30] used a Bayesian network to represent association relationship between a SNP set and the phenotype. The Bayesian network is a two-layer probabilistic graphical model, where one layer consists of a set of SNP nodes and

another of a disease node. Their conditional dependences are denoted as a set of edges in a directed acyclic graph. On the basis of previous studies [71], [72], a logarithm form of K2 score function derived from Bayesian scoring criteria was chosen as its fitness function,

$$f_{K2\log}\left(S^i(t)\right)=\sum_{j=1}^{3^K}\left(\sum_{d=1}^{s_j+1}\log(d)-\sum_{l=0}^{1}\sum_{e=1}^{s_{jl}}\log(e)\right), \quad (31)$$

where K is the number of SNPs in the set $S^i(t)$, s_j is the number of samples that have the $j-th$ combinatorial genotype, s_{jl} is the number of samples that have the l phenotype (0 denotes control and 1 denotes case) and the $j-th$ combinatorial genotype. The $f_{K2\log}$ score is a measure of the causative relationship between the SNP set and the phenotype, and the smaller the score the stronger the association. Its reciprocal form was employed to satisfy the pheromone updating rule, which also has been adopted by other ACO based methods, including MACOED [25], [26], IACO [27], and epiACO(S) [28].

4) MULTIPLE OBJECTIVES

Considering that potential model preference and disease models complexity, a detection method with only one fitness function may not always work well. Jing and Shen [25], [26] hence presented an ACO based method MACOED with two fitness functions, or evaluation objectives, to detect 2-order epistatic interactions. The first is the logarithm form of K2 score function and the second is the Akaike Information Criterion (AIC) [73] score function of ADDitive INTeractive logistic regression (ADDINT) [74] model. The ADDINT model, which represents association relationship between two SNPs and the phenotype, can be written as

$$\log\frac{c}{1-c}=\lambda_0+\lambda_1X+\lambda_2Y+\lambda_3XY, \quad (32)$$

where c is the probability of a population with the given genotype being a case, X and Y are SNPs in $S^i(t)$. With this ADDINT model, the AIC score, which deals with the tradeoff between the goodness of fit and the complexity of the model, can be computed,

$$f_{AIC}\left(S^i(t)\right)=2\left(\mu-\log lik\right), \quad (33)$$

where $\log lik$ is the maximized log-likelihood of the ADDINT model, and μ denotes the number of free parameters. For the AIC score function, SNP sets with lower scores are much more likely to be epistatic interactions. Jing and Shen claimed that the combination of these two fitness functions are complementary, resulting in better performance than those using each independently.

The FAACOSE [33] also used two fitness functions to infer epistatic interactions: one being the AIC score function, and another being the explain score function, which was defined as,

$$f_{Exp}\left(S^i(t)\right)=\sum_{j=1}^{3^K}\left|s_{j1}-s_{j0}\right|, \quad (34)$$

where K is the number of SNPs in the set $S^i(t)$, producing 3^K combination genotypes, s_{j1} and s_{j0} are respective numbers of cases and controls that have $j-th$ combination genotypes.

Sun *et al.* [27] proposed IACO for detecting SNP-SNP interactions based on the fitness function f_{Svalue}, which combined both Bayesian network and mutual information, and was defined as,

$$f_{Svalue}\left(S^i(t)\right)=\frac{MI\left(S^i(t);C\right)}{f_{K2\log}\left(S^i(t)\right)}, \quad (35)$$

where $f_{K2\log}$ is the Bayesian network score, and $MI\left(S^t(t);C\right)$ is the mutual information value between the SNP set $S^t(t)$ and the phenotype C, which can be defined as

$$MI\left(S^i(t);C\right)=H\left(S^t(t)\right)+H(C)-H\left(S^t(t),C\right), \quad (36)$$

where $H\left(S^i(t)\right)$ and $H(C)$ are the entropies of $S^t(t)$ and C respectively, $H\left(S^t(t),C\right)$ is the joint entropy between $S^t(t)$ and C, which was defined as,

$$H\left(S^t(t),C\right)=-\sum_{j=1}^{3^K}\sum_{l=0}^{1}\left(\frac{s_{jl}}{N}\cdot\log\frac{s_{jl}}{N}\right) \quad (37)$$

5) TWO-STAGE DESIGNS

Wang *et al.* [46] proposed AntEpiSeeker to identify epistatic interactions, which based on the two-stage design. In the first stage, the generic ACO algorithm was used, which results in a highly suspected SNP set determined by *chi*-squared scores, and a reduced SNP set determined by their pheromones. In the second stage, AntEpiSeeker conducted an exhaustive search of epistatic interactions within the highly suspected SNP set, as well as the reduced SNP set. Later, They [47] further extended AntEpiSeeker as AntEpiSeeker2.0 to infer epistasis associated pathways based on a natural use of ACO pheromones. Pheromones of pre-determined pathways can be estimated by the average pheromones of their top 25% or 50% associated SNPs ranked by SNP pheromones. Here, associated SNPs of a pathway are those that located between 1kb upstream and downstream of its involved genes. Then these pathways are ranked according to their descending pheromones, and top ones are more likely to be associated with the detected epistatic interactions. As far as we know, AntEpiSeeker2.0 is the first methods to provide inference of epistasis associated pathways.

The MACOED [25], [26] is also a two-stage method for detecting genetic interactions. In the screening stage, two fitness functions were combined with a memory based ACO algorithm to search for candidate SNP sets. In the cleaning stage, an exhaustive search of epistatic interactions was conducted within the candidate SNP sets of the last iteration in the first stage. The Pearson's *chi*-squared test p-value after Bonferroni correction was used to quantify the association strength between a candidate SNP set and the phenotype.

The FAACOSE [33] in the first stage used the fast adaptive ACO algorithm, AIC score function, explain score function, Pareto optimality approach, and memory based strategy

TABLE 1. Properties overview of aco based methods for detecting epistatic interactions.

Reference	Name	Fitness function	Order	Implementation and URL	Applications
Shang *et al.* 2012 [23]	AntMiner	*Chi*-squared test	1-*K* Simultaneously	Matlab https://sourceforge.net/projects/antminer/files/	Age-related Macular Degeneration
Zhou *et al.* 2014 [24]	epiACO(Z)[#1]	B-statistic	1-*K* Simultaneously	/	Type 1 Diabetes
Jing and Shen 2014 [25] Jing and Shen 2015 [26]	MACOED	$f_{K2\log} + f_{AIC}$ *Chi*-squared test	2	Matlab/C++ http://www.csbio.sjtu.edu.cn/bioinf/MACOED/	Late-Onset Alzheimer's Disease
Sun *et al.* 2016 [27]	IACO	f_{Svalue}	*K*	Matlab https://sourceforge.net/projects/iaco1/files/	Age-related Macular Degeneration
Sun *et al.* 2017 [28]	epiACO(S)[#1]	f_{Svalue}	*K*	Matlab http://sourceforge.net/projects/epiaco1/files/	Age-related Macular Degeneration
Zhou *et al.* 2016 [29]	/	f_1	Random Length Simultaneously	/	Rheumatoid Arthritis
Li *et al.* 2012 [30]	/	$f_{K2\log}$	1-4	/	/
Rekaya *et al.* 2009 [31] Robbins *et al.* 2011 [32]	ACA	Permutation test	2	/	/
Yuan *et al.* 2017 [33]	FAACOSE	$f_{AIC} + f_{Exp}$ Fisher Exact test	2	/	Late-Onset Alzheimer's Disease
Sapin *et al.*2015 [34]	ACO_Tabu	*Chi*-squared test	2	/	Type 2 diabetes Type 1 Diabetes Rheumatoid Arthritis Inflammatory Bowel Disease
Sapin *et al.* 2013 [35]	/	*Chi*-squared test	2	/	Type 2 diabetes
Sapin *et al.* 2013 [36]	/	Permutation test	2	/	Type 2 diabetes
Sapin *et al.* 2014 [37]	/	*Chi*-squared test	Decision tree: 4 Contingency table: 2	/	Type 2 diabetes
Sapin *et al.* 2015 [38]	/	*Chi*-squared test	Contingency table: 2,3 Logical interaction: 3 Decision tree: 4	/	Type 2 diabetes
Christmas *et al.* 2011 [39]	/	f_1 f_2 f_3	Random length Simultaneously	/	Type 2 diabetes
Greene *et al.* 2008 [40]	TuRF-ACO[#2]	MDR accuracy	2	C++, MDR packages, http://epistasis.org/	/
Sinnott-Armstrong *et al.* 2009 [41] Sinnott-Armstrong *et al.* 2010 [42]	TuRF-ACO[#2]	MDR accuracy	2	Python, PyCUDA for GPUs http://www.multifactordimensionalityreduction.org/	/
Greene *et al.* 2009 [43] Gilmore *et al.*2011 [44] Sulovari *et al.* 2013 [45]	TuRF-ACO[#2]	MDR accuracy	2	Java, MDR packages, http://epistasis.org/	/
Wang *et al.* 2010 [46]	AntEpiSeeker	*Chi*-squared test	2[#3]	C++ http://nce.ads.uga.edu/~romdhane/AntEpiSeeker/index.html	Rheumatoid Arthritis
Wang *et al.* 2012 [47]	AntEpiSeeker 2.0	*Chi*-squared test	2[#3]	C++ http://lambchop.ads.uga.edu/antepiseeker2/	Rheumatoid Arthritis

[#1]: These two methods have the same name epiACO, in order to distinguish them, they are referred to as epiACO(Z) and epiACO(S) here.
[#2]: This method and its extensions are called TuRF-ACO for short. They are not actually named at all in their original papers.
[#3]: The methods can theoretically detect *K*-order epistatic interactions. Nevertheless, their implementations only focus on 2-order.

to obtain candidate SNPs. In the second stage, a Fisher exact test was employed to exhaustively identify epistatic interactions.

IV. DISCUSSION AND CONCLUSIONS

Detection of epistatic interactions is particularly important to better unravel the genetic basis of complex diseases. Many methods therefore have been proposed, among which, ACO based ones are promising due to their controllable time complexities, heuristic positive feedback search, and high detection power. Nevertheless, there is no systematical review of them so far. In this paper, the generic ACO algorithm and its application to detect epistatic interactions are firstly described. Then, a full picture of the evolvement and improvement of ACO based methods is provided, and 25 methods are detailed discussed from 4 aspects, including path selection strategies, pheromone updating rules, fitness functions, and two-stage designs. Finally, their strengths and limitations are analyzed to provide guidelines for applying them, and to give several clues for future directions of ACO based epistasis detection methods, even swarm optimization based ones or other machine learning and data mining ones [7].

In summary, none of them is perfect in all scenarios and each has its own merits and limitations. For heuristic information methods, heuristic information that applies on either path selection strategies or pheromone updating rules indeed increases power, however, getting heuristic information sometimes dramatically increases the computational burden, for instance, the Relief family methods. Besides, it is hard to properly obtain biological or statistical heuristic information since no prior knowledge is usually available while given a specific disease data set. For multiple fitness functions methods, their fitness functions are always claimed to be complementary to each other, resulting in better performance than those using each independently. But finding several fitness functions that are proved to be complementary from theoretic and practical standpoints is a challenge. Also, these methods might be criticized for the complexity of multiple fitness functions computation. For MDR based methods, though MDR is one of the most popular methods for detecting epistatic interactions, its time complexity is high. The use of MDR strategy obviously increases the computational burden. Furthermore, these studies should test their performance using more epistasis models in simulation data sets and should be applied to real applications. For two-stage

methods, though significantly reducing time costs, they are often problematic for the discovery of epistatic interactions since during the screening stage SNPs must be filtered, leading to the exclusion of SNPs with weak or even no marginal effects, however, some of which are indeed causative ones.

Though providing empirical comparison and independent evaluation in terms of several criteria based on different testing data sets is necessary and could reinforce the statement by the authors, this review only focuses on algorithmic and mathematic details. This is for several reasons. First, most of these methods did not provide software or even source codes. Second, though some have software, they are implemented by different programming languages, for instance, Matlab, C++, Java, and Python, resulting in comparison of them on the same running platform and criteria being unfair. Third, some methods [29], [41], [42] were implemented on GPUs or a hadoop cluster. It might be inappropriate for them to compare with others.

There are several directions for further studies. First, how to set parameters appropriately when handling with a specific disease data set is a great challenge for ACO based methods. In order to balance the complexity and accuracy, recommended settings of these parameters should be given and their setting rules should be discussed in detail. Second, software packages are the bridges between computer scientists and geneticists. Current methods rarely provides software packages, even codes (Table 1), hindering their widely applications in the real world. Third, we should keep abreast of advances in ACO algorithm. Its improvements and applications on other fields should pay close attention to, some of which might be used in the epistasis detection field. Fourth, future studies should focus on using statistical and biological heuristic information in methods. For instance, the Relief family methods should be further studied to present new members that can efficiently and effectively provide statistical heuristic information. It is worthwhile to explore biological heuristic information from numerous databases, gene-gene interactions, protein-protein interactions, biochemical pathways, omics networks, as well as their integrated information. Fifth, with the parallel property of ACO algorithm, the future methods should be considered to be implemented on GPUs or a hadoop cluster. Sixth, the biological interpretations of results from these methods on real data sets need to be further investigated. Seventh, their abilities of search triples and higher order epistatic interactions should be discussed.

REFERENCES

[1] T. A. Manolio *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009.

[2] J. Shang, Y. Sun, J.-X. Liu, J. Xia, J. Zhang, and C.-H. Zheng, "CINOEDV: A co-information based method for detecting and visualizing n-order epistatic interactions," *BMC Bioinf.*, vol. 17, p. 214, May 2016.

[3] B. Maher, "Personal genomes: The case of the missing heritability," *Nature*, vol. 456, no. 7218, pp. 18–21, Nov. 2008.

[4] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, "A survey about methods dedicated to epistasis detection," *Frontiers Genet.*, vol. 6, p. 285, Sep. 2015.

[5] A. Upton, O. Trelles, J. A. Cornejo-Garcia, and J. R. Perkins, "Review: High-performance computing to detect epistasis in genome scale data sets," *Briefings Bioinf.*, vol. 17, no. 3, pp. 368–379, May 2016.

[6] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nature Rev. Genet.*, vol. 10, no. 6, pp. 392–404, Jun. 2009.

[7] W.-H. Wei, G. Hemani, and C. S. Haley, "Detecting epistasis in human complex traits," *Nature Rev. Genet.*, vol. 15, no. 11, pp. 722–733, Nov. 2014.

[8] K. V. Steen, "Travelling the world of gene-gene interactions," *Briefings Bioinf.*, vol. 13, no. 1, pp. 1–19, Jan. 2012.

[9] E. E. Eichler *et al.*, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Rev. Genet.*, vol. 11, no. 6, pp. 446–450, Jun. 2010.

[10] P. C. Phillips, "Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems," *Nature Rev. Genet.*, vol. 9, no. 11, pp. 855–867, Nov. 2008.

[11] S. Uppu, A. Krishna, and R. P. Gopalan, "A review on methods for detecting SNP interactions in high-dimensional genomic data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 2, pp. 599–612, Mar./Apr. 2018.

[12] M. D. Ritchie *et al.*, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Amer. J. Hum. Genet.*, vol. 69, no. 1, pp. 138–147, Jul. 2001.

[13] D. Gola, J. M. M. John, K. Van Steen, and I. R. Konig, "A roadmap to multifactor dimensionality reduction methods," *Briefings Bioinf.*, vol. 17, no. 2, pp. 293–308, Mar. 2016.

[14] J. P. Lewinger *et al.*, "Efficient two-step testing of gene-gene interactions in genome-wide association studies," *Genet. Epidemiol.*, vol. 37, no. 5, pp. 440–451, Jul. 2013.

[15] X. Sun, Q. Lu, S. Mukherjee, P. Crane, R. Elston, and M. Ritchie, "Analysis pipeline for the epistasis search—Statistical versus biological filtering," *Frontiers Genet.*, vol. 5, p. 106, Apr. 2014.

[16] I. Kononenko, *Estimating Attributes: Analysis and Extensions of RELIEF*. Berlin, Germany: Springer, 1994, pp. 171–182.

[17] K. Kira and L. A. Rendell, *A Practical Approach to Feature Selection*. New York, NY, USA: Elsevier, 1992, pp. 249–256.

[18] J. H. Moore and B. C. White, *Tuning ReliefF for Genome-Wide Genetic Analysis*. Berlin, Germany: Springer, 2007, pp. 166–175.

[19] C. S. Greene, N. M. Penrod, J. Kiralis, and J. H. Moore, "Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions," *BioData Mining*, vol. 2, no. 1, p. 5, Sep. 22, 2009.

[20] J. Shang, Y. Sun, S. Li, J.-X. Liu, C.-H. Zheng, and J. Zhang, "An improved opposition-based learning particle swarm optimization for the detection of SNP-SNP interactions," *BioMed. Res. Int.*, vol. 2015, Jan. 2015, Art. no. 524821.

[21] L. Y. Chuang, Y. D. Lin, H. W. Chang, and C. H. Yang, "An improved PSO algorithm for generating protective SNP barcodes in breast cancer," *PLoS ONE*, vol. 7, no. 5, p. e37018, May 2012.

[22] M. Aflakparast, H. Salimi, A. Gerami, M. P. Dube, S. Visweswaran, and A. Masoudi-Nejad, "Cuckoo search epistasis: A new method for exploring significant genetic interactions," *Heredity*, vol. 112, no. 6, pp. 666–674, Jun. 2014.

[23] J. Shang, J. Zhang, X. Lei, Y. Zhang, and B. Chen, "Incorporating heuristic information into ant colony optimization for epistasis detection," *Genes Genom.*, vol. 34, no. 3, pp. 321–327, Jun. 2012.

[24] Z. Zhou, G. Liu, L. Su, L. Han, and L. Yan, "A new epistasis detecting algorithm based on ant colony optimization," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2014, pp. 226–231.

[25] P. Jing and H. Shen, "A novel two-stage multi-objective ant colony optimization approach for epistasis learning," in *Proc. Chin. Conf. Pattern Recognit.*, 2014, pp. 528–535.

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

[27] Y. Sun, J. Shang, J. Liu, and S. Li, "An improved ant colony optimization algorithm for the detection of SNP-SNP interactions," in *Proc. Int. Conf. Intell. Comput.*, 2016, pp. 21–32.

[28] Y. Sun, J. Shang, J.-X. Liu, S. Li, and C.-H. Zheng, "epiACO—A method for identifying epistasis based on ant colony optimization algorithm," *Biodata Mining*, vol. 10, p. 23, Jul. 2017.

[29] Z. Zhou, G. Liu, and L. Su, "A new approach to detect epistasis utilizing parallel implementation of ant colony optimization by MapReduce framework," *Int. J. Comput. Math.*, vol. 93, no. 3, pp. 511–523, 2016.

[30] S.-S. Li, J. Chen, Q.-J. Jiao, L.-X. Yao, and H.-B. Shen, "A flexible novel approach to learn epistasis based on ant colony optimization," in *Proc. CCC*, 2012, pp. 7370–7375.
[31] R. Rekaya and K. Robbins, "Ant colony algorithm for analysis of gene interaction in high-dimensional association data," *Brazilian J. Animal Sci.*, vol. 38, pp. 93–97, Jul. 2009.
[32] K. Robbins, K. Bertrand, and R. Rekaya, "The use of the ant colony algorithm for the detection of marker associations in the presence of gene interactions," *Int. J. Bioinf. Res.*, vol. 3, no. 2, pp. 227–235, 2011.
[33] L. Yuan, C.-A. Yuan, and D.-S. Huang, "FAACOSE: A fast adaptive ant colony optimization algorithm for detecting SNP epistasis," *Complex*, vol. 2017, 2017, Art. no. 5024867.
[34] E. Sapin, E. Keedwell, and T. Frayling, "An ant colony optimization and tabu list approach to the detection of gene-gene interactions in genome-wide association studies," *IEEE Comput. Intell. Mag.*, vol. 10, no. 4, pp. 54–65, Nov. 2015.
[35] E. Sapin, E. Keedwell, and T. Frayling, "Subset-based ant colony optimisation for the discovery of gene-gene interactions in genome wide association studies," in *Proc. 15th Annu. Conf. Genet. Evol. Comput.*, 2013, pp. 295–302.
[36] E. Sapin, E. Keedwell, and T. Frayling, *Ant Colony Optimisation for Exploring Logical Gene-Gene Associations in Genome Wide Association Studies*. Exeter, U.K.: Univ. Exeter, 2013, pp. 449–456.
[37] E. Sapin, E. Keedwell, and T. Frayling, "Ant colony optimisation of decision trees for the detection of gene-gene interactions," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Belfast, U.K., Nov. 2014, pp. 57–61.
[38] E. Sapin, E. Keedwell, and T. Frayling, "Ant colony optimisation of decision tree and contingency table models for the discovery of gene-gene interactions," *IET Syst. Biol.*, vol. 9, no. 6, pp. 218–225, Dec. 2015.
[39] J. Christmas, E. Keedwell, T. M. Frayling, and J. R. B. Perry, "Ant colony optimisation to identify genetic variant association with type 2 diabetes," *Inf. Sci.*, vol. 181, no. 9, pp. 1609–1622, May 2011.
[40] C. S. Greene, B. C. White, and J. H. Moore, "Ant colony optimization for genome-wide genetic analysis," in *Ant Colony Optimization and Swarm Intelligence*. Berlin, Germany: Springer, 2008, pp. 37–47.
[41] N. A. Sinnott-Armstrong, C. S. Greene, and J. H. Moore, "Using evolutionary computing on consumer graphics hardware for epistasis analysis in human genetics," in *Proc. Genet. Evol. Comput. Conf. (GECCO)*, Montreal QC, Canada, 2009.
[42] N. A. Sinnott-Armstrong, C. S. Greene, and J. H. Moore, "Fast genome-wide epistasis analysis using ant colony optimization for multifactor dimensionality reduction analysis on graphics processing units," in *Proc. 2th Annu. Conf. Genet. Evol.*, 2010, pp. 215–216.
[43] C. S. Greene, J. M. Gilmore, J. Kiralis, P. C. Andrews, and J. H. Moore, "Optimal use of expert knowledge in ant colony optimization for the analysis of epistasis in human disease," in *Proc. Evol. Comput., Mach. Learn. Data Mining Bioinf.* Berlin, Germany: Springer, 2009, pp. 92–103.
[44] J. M. Gilmore, C. S. Greene, P. C. Andrews, J. Kiralis, and J. H. Moore, *An Analysis of New Expert Knowledge Scaling Methods for Biologically Inspired Computing*. Berlin, Germany: Springer, 2009, pp. 286–293.
[45] A. Sulovari, J. Kiralis, and J. H. Moore, "Optimal use of biological expert knowledge from literature mining in ant colony optimization for analysis of epistasis in human disease," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. EvoBIO*. Vienna, Austria. Berlin, Germany: Springer, 2013, pp. 129–140.
[46] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "AntEpiSeeker: Detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Res.Notes*, vol. 3, p. 117, Apr. 2010.
[47] Y. Wang, X. Liu, and R. Rekaya, "AntEpiSeeker2.0: Extending epistasis detection to epistasis-associated pathway inference using ant colony optimization," *Nature Precedings*. London, U.K.: Nature Publishing Group, 2012.
[48] H. J. Cordell, "Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans," *Proc. Hum. Mol. Genet.*, vol. 11, no. 20, pp. 2463–2468, Oct. 2002.
[49] J. H. Moore, "A global view of epistasis," *Nature Genet.*, vol. 37, no. 1, pp. 13–14, Jan. 2005.
[50] J. H. Moore and S. M. Williams, "Epistasis and its implications for personal genetics," *Amer. J. Hum. Genet.*, vol. 85, no. 3, pp. 309–320, Sep. 2009.
[51] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, no. 4, pp. 445–455, Feb. 2010.
[52] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genet.*, vol. 37, no. 4, pp. 413–417, Apr. 2005.
[53] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore, "Machine learning for detecting gene-gene interactions: A review," *Appl. Bioinf.*, vol. 5, no. 2, pp. 77–88, 2006.
[54] A. A. Motsinger-Reif, S. M. Dudek, L. W. Hahn, and M. D. Ritchie, "Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology," *Gene. Epidemiol.*, vol. 32, no. 4, pp. 325–340, May 2008.
[55] M. Garcia-Magarinos, I. Loper-de-UIIibarri, R. Cao, and A. Salas, "Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction," *Ann. Hum. Genet.*, vol. 73, no. 3, pp. 360–369, May 2009.
[56] Y. Wang, G. Liu, M. Feng, and L. Wong, "An empirical comparison of several recent epistatic interaction detection methods," *Bioinformatics*, vol. 27, no. 21, pp. 2936–2943, Nov. 2011.
[57] L. Chen *et al.*, "Comparative analysis of methods for detecting interacting loci," *BMC Genomics*, vol. 12, p. 234, Jul. 2011.
[58] C. C. M. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. Macrossan, "Methods for identifying SNP interactions: A review on variations of logic regression, random forest and Bayesian logistic regression," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 6, pp. 1580–1591, Nov./Dec. 2011.
[59] K. Van Steen, "Travelling the world of gene-gene interactions," *Briefings Bioinf.*, vol. 13, no. 1, pp. 1–19, Jan. 2012.
[60] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Briefings Funct. Genomics*, vol. 14, no. 2, pp. 143–155, Mar. 2015.
[61] X. Guo, N. Yu, F. Gu, X. Ding, J. Wang, and Y. Pan, "Genome-wide interaction-based association of human diseases—A survey," *Tsinghua Sci. Technol.*, vol. 19, no. 6, pp. 596–616, Dec. 2014.
[62] C. L. Koo *et al.*, "Software for detecting gene-gene interactions in genome wide association studies," *Biotechnol. Bioprocess. Eng.*, vol. 20, no. 4, pp. 662–676, Aug. 2015.
[63] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: Optimization by a colony of cooperating agents," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 26, no. 1, pp. 29–41, Feb. 1996.
[64] M. Dorigo and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 53–66, Apr. 1997.
[65] J. Shang, J. Zhang, Y. Sun, and Y. Zhang, "EpiMiner: A three-stage co-information based method for detecting and visualizing epistatic interactions," *Digit. Signal Process.*, vol. 24, pp. 1–13, Jan. 2014.
[66] J. R. Quevedo, A. Bahamonde, M. Perez-Enciso, and O. Luaces, "Disease liability prediction from large scale genotyping data using classifiers with a reject option," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 1, pp. 88–97, Jan./Feb. 2012.
[67] E. Sapin and E. C. Keedwell, *T-ACO Tournament Ant Colony Optimisation for High-Dimensional Problems*. Setúbal, Portugal: SciTePress, 2012, pp. 81–86.
[68] O. Shoval *et al.*, "Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, Jun. 2012.
[69] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genet.*, vol. 39, no. 9, pp. 1167–1173, Sep. 2007.
[70] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio—The analysis and navigation of molecular networks," *Bioinformatics*, vol. 19, no. 16, pp. 2155–2157, Nov. 2003.
[71] X. Jiang, M. M. Barmada, and S. Visweswaran, "Identifying genetic interactions in genome-wide data using Bayesian networks," *Genetic Epidemiol.*, vol. 34, no. 6, pp. 575–581, Sep. 2010.
[72] X. Jiang, R. E. Neapolitan, M. M. Barmada, and S. Visweswaran, "Learning genetic epistasis using Bayesian network scoring criteria," *BMC Bioinformatics*, vol. 12, no. 1, p. 89, Mar. 2011.
[73] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1973, pp. 610–624.
[74] B. V. North, D. Curtis, and P. C. Sham, "Application of logistic regression to case-control association studies involving two causative loci," *Hum. Heredity*, vol. 59, no. 2, pp. 79–87, 2005.

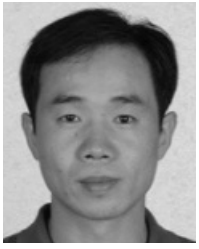
JUNLIANG SHANG received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, China, in 2007, 2010, and 2013, respectively. He is currently an Associate Professor with the School of Information Science and Engineering, Qufu Normal University, Rizhao, China. His research interests include bioinformatics and big data mining.

XUAN WANG received the B.S. degree in computer science and technology from Qufu Normal University, China, in 2017, where she is currently pursuing the master's degree in computer technology. Her research interest includes bioinformatics.

XIAOYANG WU received the B.S. degree in bioengineering and the master's degree in biology from Qufu Normal University, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in bioinformatics. His research interest includes bioinformatics.

YINGXIA SUN received the B.S. degree in computer science and technology from Qufu Normal University, China, in 2016, where she is currently pursuing the master's degree in computer science and technology. Her research interest includes bioinformatics.

QIAN DING received the B.S. degree in computer science and technology from Qufu Normal University, China, in 2018, where she is currently pursuing the master's degree in computer science and technology. Her research interest includes bioinformatics.



JIN-XING LIU received the B.S. degree in electronic information and electrical engineering from Shandong University, China, in 1997, the M.S. degree in control theory and control engineering from Qufu Normal University, Rizhao, China, in 2003, and the Ph.D. degree in computer simulation and control from the South China University of Technology, China, in 2008. From 2011 to 2015, he was with the Shenzhen Graduate School, Harbin Institute of Technology, as a Postdoctoral Research Fellow. He is currently a Professor with the School of Information Science and Engineering, Qufu Normal University. His research interests include pattern recognition, machine learning, and bioinformatics.

HONGHAI ZHANG received the B.S., M.S., and Ph.D. degrees from Northeast Forestry University, Harbin, China, in 1989, 1994, and 1997, respectively. He is currently the President of Qufu Normal University, Qufu, China. His research interests include evolutionary genomics and conservation biology of animals.