Author’s response to reviews

Title: epiACO - A Method for Identifying Epistasis Based on Ant Colony Optimization Algorithm

Authors:
Yingxia Sun (sunyingxia026@163.com)
Junliang Shang (shangjunliang110@163.com)
Jin-Xing Liu (sdcavell@126.com)
Shengjun Li (qfnulsj@163.com)
Chun-Hou Zheng (zhengch99@126.com)

Version: 1 Date: 10 Mar 2017

Author’s response to reviews:
Dear Dominik Heider,

Thank you very much for your time on our manuscript.
And also, we would thank reviewers for their constructive comments and efforts on the manuscript. Our point-by-point replies listed below.

We have revised the manuscript according to reviewers' and your comments and suggestions. The revised manuscript has been edited again by a professional scientific writer. The main changes have been highlighted with "tracked changes".

The revised manuscript has been resubmitted to your honorable journal. We hope these changes lead to the acceptance of manuscript, and look forward to hearing from you soon.
Thanks again.

Yours sincerely
Junliang Shang

Replies to Reviewer #1
The authors introduce a new method of ant colony optimization called epiACO for the task of identifying epistatic interactions. The authors claim that the problem of identifying epistatic interactions can be converted into detecting the SNP combinations with a larger value of a fitness function (S-value) introduced by them.

1. However, it remains unclear why an ant colony optimization algorithm is necessary for this task. The authors should discuss why a common standard conjugate gradient algorithm (e.g. used for large data sets in NeRV [Venna et al., 2010]) cannot solve this problem.

√ Thank you very much for your comments.

Though many works have been done for identifying epistatic interactions, due to their methodological and computational challenges, the algorithmic development is still ongoing. Current methods can be mainly classified into three categories based on their search strategies: exhaustive search [1-3], stochastic search [4-6], and heuristic search [7-12]. Heuristic search methods usually obtain solutions at a substantially reduced time cost, based on their introducing heuristic information and prior knowledge of biological data. Among heuristic search methods, ant colony optimization (ACO) algorithm appears promising and some related works have been reported, including AntEpiSeeker, AntMiner, MACOED, and IACO. [Section 1 in manuscript]

The aim of this study is to develop an epistasis detection method based on ACO algorithm, which is comparable and sometimes superior to existing methods, especially ACO based methods. This does not mean that methods based on other principles are not suitable for identifying epistasis, and their performance are worse than epiACO. We believe that each method has its own merits and limitations, and the more the methods based on different principles being proposed, the faster this problem being solved. For instance, the common standard conjugate gradient algorithm [13] and the genetic algorithm [14] are important optimization algorithms, and can undoubtedly deal with this problem. [Conclusions in manuscript]

2. In the first chapter of the paper, a many uncited claims are made which are not verifiable by the reader, please provide citations for the following lines (page 1: 41-41,53-54; page 2: 1-2,3-4,7-8,18-20,28 )

√ Thank you very much.
We have rewritten the first chapter in the revised manuscript, and the mentioned citations have been provided. [Section 1 in manuscript]

3. It is not obvious why for comparison the methods AntMiner, IACO, AntEpiSeeker and MACOED are chosen and why other methods are disregarded (e.g. SNPRuler). Please provide detailed explanations or compare with all described methods.

\[\sqrt{\text{Thank you very much for your comments.}}\]

In the revised manuscript, performance of epiACO is analyzed by comparison with eight typical 2-SNP epistasis detection methods, i.e., epiMODE, TEAM, BOOST, SNPRuler, AntEpiSeeker, AntMiner, MACOED, and IACO. Among them, BOOST and TEAM are exhaustive search methods, epiMODE is a stochastic search method, others are heuristic search methods. In particular, AntEpiSeeker, AntMiner, MACOED, and IACO are all ACO based methods, which will be discussed in more detail than others since they are from the same family, as well as epiACO. [Section 3.3 in manuscript]

4. In the methods chapter 2.1, I suggest the reading of [Martens et al., 2011] about ACO and correct the text accordingly. Also please explain why the constants \( \theta \) and \( \zeta \) are necessary and how a user could choose them appropriately. Also explain the number of ants chosen.

\[\sqrt{\text{We accept your comments. Thank you.}}\]

We have read the article of Martens et al. [15] carefully and have correct the text accordingly in the revised manuscript. [Section 2.1 in manuscript]

In the revised version, according to your comment, the constant has been deleted since it is not the necessary parameter of epiACO. is a specified constant that used to control the additional increment of pheromones, and the recommended range is \([0.2, 0.5]\). [Section 2.3.2 in manuscript]

How to set parameters appropriately is a great challenge for the family of swarm intelligence algorithms, like ACO algorithm. In general, larger ant number and iteration number lead to higher detection power, with more computational cost. In order to balance result accuracy and computational cost, multiple groups of parameter settings are used and the most appropriate one among them is selected in practice. [Conclusions in manuscript]
5. In the results section the evaluation criterion is not plausible. Please explain why you do not measure for example sensitivity and specificity or other common measures (accuracy, balanced Accuracy) instead of introducing a new quality measure. If it is a common measure for these kind of problems, please provide citations.

√ Thank you for your comments.

Detection power is one of the generally accepted and widely used evaluation measure in the field of identifying epistasis [4, 6-8, 16-18]. In this study, we directly use the detection power proposed by previous studies [4, 6-8, 16-18], which is defined as the proportion of data sets in which the epistasis models are perfectly identified with no false positives. [Section 3.1 in manuscript]

Other evaluation measures, for example, sensitivity, specificity, accuracy, balanced accuracy, and so on, should be used to carry out a broader performance analysis, which is a direction, and inspire us to continue work in the future. [Conclusions in manuscript]

6. Looking at the figures, I assume that all methods and models used are deterministic yielding always to the same result. If they are stochastic (e.g. ACO hybrids), please provide an estimation of the variance of the power measure.

√ For obtaining accurate results, each method runs 20 times with different random seeds on each data set of each model in the revised manuscript, which can ensure that the method has not been biased by its initial starting conditions. The mean detection power and the mean running time with their respective standard deviations are recorded in Table 2. [Section 3.3 in manuscript]

7. Please provide a software package on a site in English where the software package can be downloaded. Please state the programming language and operation system for the package.

√ We accept your comments.

The Matlab version of epiACO running on Microsoft Windows is available online at: https://sourceforge.net/projects/epiaco/files/epiACO.rar/download. [Section 1 in manuscript]

8. Please provide in the discussion and reference, where the self-adaptation parameter $q_0$ is introduced in the epiACO algorithm.
Thank you very much for your suggestions.

In order to control the rate of convergence, avoid falling into locally optimal solution, and improve the detection power of models displaying no marginal effects, two path selection strategies, that is, probabilistic path selection strategy and stochastic path selection strategy, are provided to adaptively guide ant behaviors of exploration and exploitation. is a threshold, which defined as the ratio of current iteration number to the total iteration number. The threshold of allows epiACO to cover a wider search space while the iteration number is small and to converge on promising regions of the search space while the iteration number turns to large, resulting in high detection power not only in ME models but also in NME models. [Section 2.3.1, Conclusions in manuscript]

Replies to Reviewer #2

The manuscript describes an ACO approach to the problem of epistasis discovery in genome-wide association data. The description of previous work and of the method itself is reasonably good, however, there are significant issues with the presentation of the work which mean that the paper cannot be published in its current form:

1. There are a large number of English language errors that should be corrected in any revised manuscript. This occasionally obscures the meaning of the authors and so addressing this would undoubtedly improve the quality of the manuscript.

Thank you very much. The revised manuscript has been edited again by a professional scientific writer.

2. The presentation of the results needs work. Figures 2-4 present the results of the approach for each epistasis model. The sections on the x-axis contains a set of parameter settings which are modified within a given range, but the parameters themselves are not labeled, nor any rationale for the range of each parameters given. It is not clear why line graphs have been used in each case. The Figure 5 sets out the performance of each algorithm in power and computational time terms, but it is difficult to determine from this which is the best approach. AntMiner for example clearly takes the longest to run, but is also the most powerful, whereas the proposed approach appears to offer only modest improvements over the other ACO approaches that don't incorporate heuristic information. A table of results here would provide a much stronger
comparison of the algorithms and would allow a direct comparison of the approaches for each epistasis model.

✓ Thank you very much for your suggestions.

Figures 2-4 in the original manuscript has been deleted, as well as their analyses, in this revised manuscript since they are not the necessary for the study. How to set parameters appropriately is a great challenge for the family of swarm intelligence algorithms, like ACO algorithm. In general, larger ant number and iteration number lead to higher detection power, with more computational cost. In order to balance result accuracy and computational cost, multiple groups of parameter settings are used and the most appropriate one among them is selected in practice.

According to your suggestions, the mean detection power and the mean running time with their respective standard deviations are recorded in Table 2. It is seen that epiACO is comparable and sometimes superior to compared methods, especially ACO based methods. Specifically, epiACO outperforms AntEpiSeeker and IACO on all models in terms of detection power, and they also have the similar running time. epiACO has higher detection power than those of AntMiner and MACOED on ME models, but lower detection power than those of them on NME models. For MACOED, it is in fact the stochastic search method since its pheromone updating strategy is only used for more frequently detecting the epistasis that have been detected in the previous iterations, rather than to find better epistatic interactions. As it is known to all, stochastic search methods are good at identifying NME models, and hence its detection power on NME models are higher than that of epiACO. For AntMiner, heuristic information of SNPs is incorporated into ant-decision rules, which on the one hand improves detection power of NME models, and on the other hand significantly increases running time, hindering its widely application on large scale data sets, like those for GWAS. From above analysis, it seems that epiACO is promising in identifying epistasis and might be an alternative to existing ACO based methods. [Section 3.3 in manuscript]

3. ACO algorithms are stochastic algorithms and so require a number of runs to determine the mean performance of the algorithm using several random seeds. This is to ensure that the algorithm has not been biased by its initial starting conditions. Although it is made clear that 50
datasets are generated for each model, it is not clear if how many randomly-seeded runs have been conducted for each dataset in the algorithm comparisons.

We accept your suggestions.

For obtaining accurate results, each method runs 20 times with different random seeds on each data set of each model, which can ensure that the method has not been biased by its initial starting conditions. [Section 3.3 in manuscript]

4. The number of iterations and size of population for the ACO algorithm seem rather low. A relatively small number of iterations and population size are used for the experiments, and some justification should be provided for these parameter settings, particularly when the other algorithms might have benefited from differing parameter settings. Although I appreciate a full exploration of these is not possible for all algorithms, a justification for the range of parameter settings should be provided and some discussion of how these settings differ from those in the original papers for these algorithms should be entered into.

We accept your suggestions. Thank you very much.

In our study, parameters of each method are generally set as default. Only a few are modified according to suggestions in their respective user manual to balance result accuracy and computational cost. For epiMODE, the iteration number is set to 100. For TEAM, the permutation number is set to 100. For BOOST, the iteration threshold is set to 10. [Section 3.3 in manuscript]

For a fair comparison, parameter settings of the ACO based methods are the same. Specifically, the iteration number and the ant number are set to 25 and 200 respectively. the initial pheromone, the heuristic information, and the parameters and are all set to 1. The evaporation coefficient is set to 0.2. The constant is set to 0.3. [Section 3.3 in manuscript]

5. Finally, I'm surprised that the results for an evaporation rate of 0.8 are so good. One would expect that this formulation would converge very quickly on a set of sub-optimal results due to the large reduction in pheromone at each timestep. This does suggest that the ACO is not the core part of this method and that the other search aspects of the approach dominate. The authors should perhaps comment on this and their other results and conduct further analysis on the generated results.
Experiments and their analyses of epiACO on simulation data sets with different evaporation rates have been deleted in the revised manuscript, since they are not the necessary for the study. How to set parameters appropriately is a great challenge for the family of swarm intelligence algorithms, like ACO algorithm, which needs further study in detail and is a direction that inspire us to continue work in the future. [Conclusions in manuscript]

Replies to Editor

Editorial comment:

While the first reviewer already mentioned alternative methodology for solving the problem, i.e., gradient descent, I suggest to discuss also alternative computational optimization approaches, such as genetic algorithms (see e.g., https://www.ncbi.nlm.nih.gov/pubmed/21094663) etc. in the discussion.

Thank you very much for your comments.

The aim of this study is to develop an epistasis detection method based on ACO algorithm, which is comparable and sometimes superior to existing methods, especially ACO based methods. This does not mean that methods based on other principles are not suitable for identifying epistasis, and their performance are worse than epiACO. We believe that each method has its own merits and limitations, and the more the methods based on different principles being proposed, the faster this problem being solved. For instance, the common standard conjugate gradient algorithm [13] and the genetic algorithm [14] are important optimization algorithms, and can undoubtedly deal with this problem. [Conclusions in manuscript]

References

