Bioinformatics, 2017, 1–8
doi: 10.1093/bioinformatics/btx339
Advance Access Publication Date: 24 May 2017
Original Paper

OXFORD

Genome analysis

A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization

Xiong Li

School of Software, East China Jiaotong University, Nanchang 330013, China

Associate Editor: John Hancock

Received on April 2, 2017; revised on May 6, 2017; editorial decision on May 20, 2017; accepted on May 20, 2017

Abstract

Motivation: The existing epistasis analysis approaches have been criticized mainly for their: (i) ignoring heterogeneity during epistasis analysis; (ii) high computational costs; and (iii) volatility of performances and results. Therefore, they will not perform well in general, leading to lack of reproducibility and low power in complex disease association studies. In this work, a fast scheme is proposed to accelerate exhaustive searching based on multi-objective optimization named ESMO for concurrently analyzing heterogeneity and epistasis phenomena. In ESMO, mutual entropy and Bayesian network approaches are combined for evaluating epistatic SNP combinations. In order to be compatible with heterogeneity of complex diseases, we designed an adaptive framework based on non-dominant sort and top *k* selection algorithm with improved time complexity *O(k*M*N)*. Moreover, ESMO is accelerated by strategies such as trading space for time, calculation sharing and parallel computing. Finally, ESMO is nonparametric and model-free.

Results: We compared ESMO with other recent or classic methods using different evaluating measures. The experimental results show that our method not only can quickly handle epistasis, but also can effectively detect heterogeneity of complex population structures.

Availability and implementation: https://github.com/XiongLi2016/ESMO/tree/master/ESMO-common-master.

Contact: lx_hncs@163.com

1 Introduction

A central goal for epidemiologists is to understand how the DNA sequence variations influence the progression of complex diseases and predict common complex diseases such as diabetes and cancers and so on. Because the calculation on millions of single nucleotide polymorphisms (SNP) suffered an enormous challenge of computing resources, most of traditional genome-wide association studies adopted single variable strategy to simplify diseases models for saving costs. However, this kind of strategy yields perceived problems such as lack of reproducibility and 'missing heritability' (Urbanowicz *et al.*, 2013).

One major reason for the problems is that the development and progression of complex diseases involve multiple SNPs which may be epistatic. Single variable methods aimed at major effects do not work well for analyzing epistasis (Moore *et al.*, 2010). Recently, lots of multi-locus epistasis methods have been designed and can be generally categorized as exhaustive search, heuristic search and machining learning methods (Jing and Shen, 2015; Tuo *et al.*, 2016; Xie *et al.*, 2012). Exhaustive methods which evaluate the association strength of all possible multi-locus epistatic combinations on disease state have been criticized for their huge computational burden. MDR is the most representative method of exhaustive search, in which multi-locus genotype predictors are effectively induced from *n* dimensions to one dimension (Moore *et al.*, 2006). The cross-validation and permutation testing of MDR are the key steps for the best model selection. MDR 2.0 beta 8.4, the latest implementation of MDR methodology, can be used for detecting, characterizing and

visualizing epistasis in a elegant manner. Although MDR 2.0 is powerful and efficient in a parallel mode, it confronts two major drawbacks: (i) embedded cross-validation and permutation in exhaustive search consume lots of computation resources; (ii) insufficient power in heterogeneous datasets. An approach KNN-MDR basically follows the same pipeline of MDR by modifying majority vote within a set of the K nearest neighbors of the tested individual (Abo and Farnir, 2017). For heuristic search, FHSA-SED (Tuo *et al.*, 2016) and MACOED (Jing and Shen, 2015) combine two objectives to heuristically search two-locus epistatic combination spaces. Based on swarm intelligent algorithm and multi-objective optimization, FHSA-SED and MACOED perform better than previous heuristic approaches. However, they also meet several challenges. Firstly, they are specific for two-locus epistasis. Taking MACOED as an example, the heuristic pheromones of n-1 loci is inefficient for constructing n loci epistasis in pure epistasis analysis, especially for high-order epistasis with huge combination space. This is because for pure epistasis model, the genotypes in n-1 loci may not shown any association with the disease states. Secondly, their performance still needs to be further improved. And, tuning the parameters (such as the number of iterations, the size of swarm and so on) which significantly influence the power and computational costs is a hard work. Finally, due to the randomness of the swarm intelligent algorithm, although running on the same dataset, the results may be inconsistent, leading to clinical researchers to lose confidence in the bioinformatics approach. Machining learning based methods such as logistic regression (Wu *et al.*, 2009) and Bayesian network (Jiang *et al.*, 2011) are like a black box which is hard to interpret the relationship between epistasis and complex diseases.

Another reason is that the impact of heterogeneity is neglected in most of association studies. Heterogeneity refers to independent effects corresponding to subgroups of complex diseases (Urbanowicz *et al.*, 2013). The mixture of subgroups leads to epistatic patterns to be hard to be detected, partly because the sample size of each homogeneous subgroup is reduced and the epistatic signals are interfered with each other. A common strategy for handling heterogeneity is to pure its confounding effect by data stratification (Fenger *et al.*, 2008), resulting in the loss of power. Only a few approaches concurrently analyze the phenomena of epistasis and heterogeneity without resorting to some form of stratification (Urbanowicz *et al.*, 2013). MDR profiles heterogeneity by returning all underlying epistatic models which may correspond to different subtypes of complex diseases. However, the power on heterogeneous datasets still needs to be improved. The major reason is that the single objective may be partial or insufficient for profiling complex genetic structure of heterogeneity. An adaptive learning classifier systems (LCSs) are a rule-based method that integrate machine learning with evolutionary computing and other heuristics (Urbanowicz *et al.*, 2013). LCSs addresses heterogeneity by introducing multiple classification rules which also break single objective paradigm. However, it is not accessible for download.

Undoubtedly, the power of association studies is the first indicator to other measures such as computational costs. In most cases, exhaustive strategy is criticized for its uneconomical computation mode and its infeasibility on genome-wide. For example, the calculation complexity of exhaustive method is $O(n^k \times S)$, in which n is the number of SNPs and k is the epistasis order and S is the complexity for model selection. Obviously, in exhaustive search mode, every uneconomical operator will be significantly enlarged. For example, the embedded cross-validation operator is the main obstacles for MDR. Despite this, we insist to apply exhaustive strategy in this study because of three reasons. Firstly, exhaustive search ensures the stability and globalization of solutions. In addition, dimension reduction method can be applied for filtering genome-wide SNPs before searching. Finally, the high performance computing platform (e.g. GPU and cloud computing) further mitigates the computational burden of large scale calculation. For example, to efficiently exploit the whole computational capacity of modern clusters based on GPUs and Xeon Phi coprocessors, pairwise epistasic detection method is run on heterogeneous clusters using both types of accelerators (González-Domínguez *et al.*, 2015); epiSNP also has been modified for modern multi-core and highly parallel many-core processors to efficiently handle these large datasets, with exquisite parallelism scheme such as the serial optimizations, dynamic load balancing and so on (Weeks *et al.*, 2016).

In this study, we propose a fast evaluation scheme for concurrently handling epistasis and heterogeneity based multi-objective optimization in exhaustive search called ESMO. We compared ESMO with other methods, including MDR, FHSA-SED and MACOED both on pure and heterogeneous datasets. Experiments show that ESMO has practical meanings for epistasis and heterogeneity analysis.

2 Materials and methods

In ESMO, the model selection step based on multi-objective is embedded in k nesting loop, so that the number of objectives will significantly influence the costs of calculation. Here, we combine only two objectives to evaluate candidate epistatic models. The first objective based on mutual entropy is used to profile the relationship between disease state and each combination of k-epistatic SNPs. For the second objective, Bayesian network derived $K2$ score is adopted to select models fitting to samples from statistical perspective. It is also worth noting that ESMO is adaptive both on pure and heterogeneous datasets. It means that for pure epistasis datasets, these two objectives are consistent for the best model. For the heterogeneous datasets, these two objectives can profile different genetic models with a better performance.

In order to ease the burden of exhaustive search, a fast scheme including calculation sharing and trading space for time is proposed to accelerate the searching process. Note that, the fast scheme can be supported by parallel computing.

2.1 Mutual entropy and K2 score

Objective 1: For a case-control study, each sample is labeled as 0 (control) or 1 (case). In this work, the Shannon entropy $H(Y)$ can be used to quantify the uncertainty of the disease state Y in bits as Formula 1.

$$H(Y) = -\sum_{i=0}^{1} p(y^i) \log_2 p(y_i) \tag{1}$$

where $p(y_0)$ represents the possibility of controls and $p(y_1)$ represents the possibility of cases in population.

Joint entropy can be applied to measure the joint uncertainty of k SNPs. Let X be a biallelic SNP with three kinds of genotypes aa (0), aA (1) and AA(2), then the joint entropy of k SNPs can be defined as Formula 2.

$$H(X_1, \ldots, X_k) = -\sum_{x_1=0}^{2} \cdots \sum_{x_k=0}^{2} p(x_1, \ldots x_k) \log_2 p(x_1, \ldots x_k) \tag{2}$$

To quantify the information contribution of a k order epistatic combination to disease state Y (or vice versa), mutual entropy can be calculated as Formula 3.

$$I(Y|X_1,\ldots X_k) = H(Y) + H(X_1,\ldots X_k) - H(Y,X_1,\ldots X_k) \quad (3)$$

where $I(Y|X_1,\ldots,X_k)$ denotes the uncertainty reduction of the disease state when the k-epistatic combination is observed. We set the *object 1* score to be reciprocal of $I(Y|X_1,\ldots,X_k)$. Consequently, SNPs with low score are considered to be epistatic.

Objective 2: A Bayesian network (BN) is a probabilistic directed graphic model consisted of a set of nodes (random variables) and edges (conditional dependence) widely used in previous studies (Jing and Shen, 2015; Skwark *et al.*, 2017; Zeng *et al.*, 2016). Given the Markov condition, the joint probability distribution for a $k+1$ nodes (k SNP and a disease state) can be calculated as Formula 4.

$$\mathrm{p}(X_1, X_2, \ldots, X_{k+1}) = \prod_{i=1}^{k+1} P(X_i|\pi(X_i)) \quad (4)$$

where $\pi(X_i)$ represents the set of parents nodes of X_i. If $\pi(X_i) = \varnothing$, $P(X_i|\pi(X_i))$ is a marginal distribution. Attributed to Markov condition, for a epistatic model, only edges from a SNP node to disease state need to be considered. A k-epistatic BN model can be seen in Figure 1. There are a total of C_n^k combinations in exhaustive search, in which n is the number of all SNPs within samples.

Like the *K2* score in (Jing and Shen, 2015), *K2* score is defined as Formula 5 when the prior distribution is assumed to be a Dirichlet distribution $D[\alpha_{11}\ldots\alpha_{ij}]$. In this study, assume that no prior knowledge about the relationship between SNPs and disease state, then we set $\alpha_{ij} = 1$.

$$K2 = \sum_{i=1}^{I}\left(\sum_{b=1}^{r_i+1}\log(b) - \sum_{j=1}^{2}\sum_{d=1}^{r_{ij}}\log(d)\right) \quad (5)$$

where I is the total number of combinations and in this study, $I = 3^k$. $r_i$is the frequency of ith genotype combination in all samples and r_{ij} refers to the number of ith genotype combination in samples with jth state. Although *K2* score is defined to be some form of k-locus epistasis in (Jing and Shen, 2015; Xie *et al.*, 2012), MACOED and FHSA-SED based on swarm intelligent algorithm are only effective for 2-locus epistasis detection. In this work, k can be 2, 3 and so on if computing resources allowed.

2.2 An adaptive framework for heterogeneity and epistasis analysis

Two objectives are introduced to select top models in previous sections. In ESMO, candidate epistatic combinations with lower scores on these two objectives have stronger associations with diseases state. Ideally, there is only one best model for pure datasets, while heterogeneous datasets may exist multiple non-dominant solutions corresponding to different epistatic models. Therefore, detecting epistasis and profiling heterogeneity becomes the problem of sorting solutions according to these two objectives.

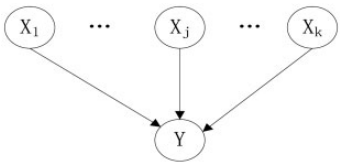

Fig. 1. A BN model between k-epistatic SNPs and disease state

The non-dominant sorting algorithm proposed in NSGA-II (Deb *et al.*, 2002) is to find all the solutions on an optimal plane since the objectives conflict with each other and its time complexity of is $O(M*N^2)$, in which M is the number of objectives and N is the size of solutions. Obviously, for a exhaustive strategy, it is unsuitable. In this study, we firstly design an adaptive framework to concurrently handle heterogeneity and epistasis based on fusing non-dominant sort and top k selection algorithm whose time complexity is $O(k*M*N)$. The pseudo code of the algorithm is listed as follows:

Non-dominant sort and top k selection algorithm

Input: the decision space A and the parameter k.
Output: k solutions set K.
selected = 1;
While(selected < =k)
  P = ∅;
  For each objective $A_i \in A$
    p_i = Find the solution with the smallest score in A_i
  and save its label into the P;
  End for
  If all $p_i \in P$ are equal
    Add the p_i solution into K;
    selected = selected + 1;
    Delete the p_i solution from decision space A;
  Else
    Add all these w different solutions into K;
    selected = selected + w;
    Delete the w solutions from decision space A;
  End if
  Return the first k solutions from K;
End while

The core ideas of the algorithm are: (i) if a solution has the best scores on all objectives, it certainly has to be considered as a epistatic combination; (ii) if a solution has the best score at least on one objective, it cannot be dominated by any other solutions.

We find that ESMO searches for each objective a set of best solutions separately, which indicates that the non-dominant sort and top k selection algorithm only searches for two separate points rather than the Pareto plane in the decision space. Here, we adhere to this strategy for the following reasons:

1. For the exhaustive scheme, searching all non-dominant solutions in the whole plane needs considerable computing sources. Therefore, we just focus on several representative solutions.
2. When the epistatic combination space is large, there may be many non-dominant solutions in the whole plane, which could result in high level of false positive. To overcome false positive ratio, clean stage must be adopted such as MACOED.
3. In spite of heterogeneity, the number of subgroups will not be too much. It means that several representative solutions may be enough.

2.3 A fast scheme for accelerating exhaustive search

As mentioned before, every uneconomical operator will be amplified in exhaustive method. We believe that the major reason for the

Table 1. Configuration of datasets simulation

Data group ID	Replication	No. of SNPS	HP	*k*	MAF
DG1	2	100	1.0	2	(0.2, 0.2)
DG2	2	100	1.0	2	(0.3, 0.3)
DG3	2	100	1.0	2	(0.4, 0.4)
DG4	1	100	(50%,50%)	(2,2)	(0.2, 0.2) (0.3, 0.3)
DG5	1	100	(60%,40%)	(2,2)	(0.2, 0.2) (0.3, 0.3)
DG6	2	1000	1.0	2	(0.2, 0.2)
DG7	2	100	1.0	3	(0.2, 0.2, 0.2)
DG8	2	100	1.0	3	(0.2, 0.3, 0.2)
DG9	2	100	1.0	3	(0.2, 0.3, 0.3)
DG10	1	100	(50%,50%)	(3, 3)	(0.2, 0.2, 0.2) (0.2, 0.3, 0.3)
DG11	1	100	(60%,40%)	(3, 3)	(0.2, 0.2, 0.2) (0.2, 0.3, 0.2)
DG12	1	100	(50%,50%)	(3, 3)	(0.2, 0.2, 0.2) (0.2, 0.3, 0.2)
DG13	2	1000	1.0	3	(0.2, 0.2, 0.2)
DG14	1	100	(60%,20%,50%)	(3,3,3)	(0.2, 0.2, 0.2) (0.3, 0.3, 0.3) (0.4, 0.4, 0.4)
DG15	1	100	(40%,30%,30%)	(3,3,3)	(0.3, 0.3, 0.4) (0.4, 0.4, 0.4) (0.3, 0.4, 0.4)

infeasibility of MDR in larger scale dataset is the huge computation costs caused by cross-validation and permutation. In this study, we use multiple objectives and non-dominant solution searching schemes to ensure the consistency of epistatic models, which are faster than cross-validation and permutation. In addition, we have applied other strategies to save computational costs such as computing share, trading space for time and parallel mode.

Computing share: The open-source ESMO is implemented on MATLAB 2014a. We found that counting the frequencies of genotypes is common between entropy calculation and possibility calculation of BN. Consequently, we extract this common operator to share its computing results.

Trading space for time: From the Formula 5, the logarithm form calculation is the most frequent operator in multi-objective optimization. For example, in exhaustive search, 10^6 calculations of multi-objective will lead to about $8*10^8$ calculations for factorial values. Actually, we found that the vast majority of factorial calculations are repeated. Therefore, we can calculate all unique factorial values before exhaustive search, so that computing Formula 5 only needs to locate the factorial value in the corresponding location. For a dataset with *M* samples, we only need to store *M* factorial values before exhaustive search. This strategy significantly saves the running time cost for ESMO.

Parallel mode: For a large scale dataset or higher order epistasis detection, parallel computing is one effective way to accelerate computing. One of the major feature of the exhaustive search is easy to be parallelism. In this work, we just naively use the simplest 'parfor' structure of MATLAB to illustrate this feature. It is important to remember that a wide variety of parallel techniques is faster than 'parfor'. Note that if there is no parallel computing devices, 'parfor' will degenerate into a serial program like 'for' loops.

3 Datasets and evaluation measures

3.1 Simulated datasets and case study

Simulated datasets: In this study, both pure and heterogeneous datasets generated by the tool GAMETES_2.1 (Urbanowicz *et al.*, 2012) are applied to evaluate different approaches. GAMETES_2.1 can simulate epistatic datasets with a friendly graphic user interface, and customer can set parameters such as MAF, heterogeneity proportion (HP), epistatic order (*k*), number of SNPs and so on) to determine the architecture of disease models.

In Table 1, there are 15 groups, and each group contains 100 datasets. And there are 800 samples in each dataset. In this study, we simulated two sizes of datasets: 100 and 1000. The replication denotes the number of times the dataset was repeatedly and randomly generated with the same parameters. For example, the replication of DG1 is 2, then with the same parameters there are two randomly groups DG1_1 and DG1_2. When HP equals to 1.0, it means that this kind of dataset is pure. In the pure dataset, *k* is the order of epistasis and the values of MAF in each parentheses correspond to *k* loci, respectively. Otherwise, the value of HP in parentheses refers to the proportion of each epistatic model in heterogeneous dataset. In this case, the *k* and MAF of heterogeneous dataset correspond to different disease model. Taken DG10 as an example, (60%,40%) means that 60% of the data from one model (3-locus epistasis and MAF are 0.2, 0.2 and 0.2)and 40% from the other model (3-locus epistasis and MAF are 0.2, 0.3 and 0.3) (Urbanowicz *et al.*, 2012). Note that we only use two epistatic models labeled H1 and H2 to simulate heterogeneous datasets. DG14 and DG15 are more complex situations for heterogeneity, in which there are three different 3-locus diseases models mixed in populations.

Case study: A breast cancer dataset consisted with 10 000 samples (5000 cases and 5000 controls) and each sample has 23 SNPs collected from 6 genes (COMT, CYP19A1, ESR1, PGR, SHBG and STS) (Yang *et al.*, 2013). These genes have been proved to be important in steroid hormone metabolism and signaling (Gabriel *et al.*, 2013).

3.2 Evaluation measures and parameters setting

To evaluate the performance of epistasis and heterogeneity detection, we use measures such as the power, loss, running time and speedup.

Table 2. The parameter setting of all methods

Algorithm	Parameters
MDR	Search type: Exhaustive; For 2-locus (3-locus) epistasis, attribute count is 2 (3).
MACOED	tau0 = 1;T0 = 0.8;rou = 0.9;lambda = 2; Ant number = 100; maximum iterations = 100 (4500) for 100 (1000) SNPs
FHSA-SED	All defaults, except maximum iterations = 4500 (100000) for 100 (1000) SNPs
ESMO	k = 3

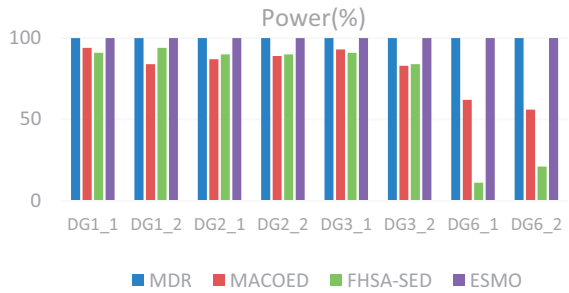

Fig. 2. The power comparison on 2-locus epistasis

Power: The power is traditionally defined as Formula 6.

$$Power = \frac{n}{N} \tag{6}$$

where n indicates the frequency of correctly recognizing the real functional loci in a group. Each group contains a fixed number N of datasets, and in this study N = 100.

Loss: To fairly compare with other methods which is not specific for heterogeneous dataset, the loss is defined as Formula 7.

$$\text{Loss} = \frac{l}{N} \tag{7}$$

where l is the number of datasets in which none of any epistatic functional loci has been detected.

Running time: We count the seconds of all methods (MDR, MACOED, FHSA-SED and ESMO) to evaluate their efficiency. All these methods are running on our workstation (Windows 10, Intel i7-6800K 6 cores and 16G RAM) and the parameters are listed in Table 2.

Speedup: The speedup is the ratio of sequential calculation running time (Ts) to parallel calculation running time (Tp) as Formula 8.

$$Speedup = \frac{Ts}{Tp} \tag{8}$$

4 Experiments and results

4.1 Experiments on pure datasets

We simulated two sizes (100 and 1000) of pure datasets such as DG1-3, DG6-9 and DG13. In the following subsections, the power, running time and speedup results are listed to show the performance of ESMO for epistasis detection.

4.1.1 ESMO versus other methods for 2-locus epistasis on pure datasets

Figure 2 depicts the results of the power on datasets with 100SNPs (DG1-3) and datasets with 1000 SNPs (DG6). The results show that, for pure epistasis analysis, the performance of MDR is very comparable with ESMO and both of them reach 100%. Obviously, the main reason for this is that ESMO and MDR search epistatic combination space exhaustively. Consequently, if the criteria is fit for profiling epistasis, the best model, namely real functional loci, will be precisely detected.

Figure 2 also reflects the volatility of swarm intelligent algorithm based methods FHSA-SED and MACOED. The power of FHSA-SED is better than MACOED for 100 SNPs dataset, on average. However, for DG6, MACOED achieves far higher power than FHSA-SED. Note that, the performance of FHSA-SED and MACOED also depends on the parameters. It means that in some cases of parameters settings FHSA-SED is better than MACOED, but in other situations, MACOED may be better than FHSA-SED.

From the results of Table 3, whether or not parallel technology is used, the running time of ESMO is the lowest for 100 SNPs datasets. Although MDR and ESMO-parfor both are parallelism, the running time cost of MDR is more than 10 times that of ESMO-parfor. It is interesting that the sequential algorithm FHSA-SED is also significantly faster than MDR. However, FHSA-SED is at the expense of the power. For 1000 SNP datasets, the running time of MDR and FHSA-SED are similar, and we can expect that as the size of datasets increases, the advantages of parallelism will gradually show. And, ESMO only needs about 1400 seconds to detect 2-locus epistasis on 1000 SNPs datasets, even with a simple parallel structure 'parfor'.

Although ESMO-unparfor achieves 100% power, it takes a relative longer running time than MDR and FHSA-SED. From the results of the power and running time in Figure 2 and Table 3, we find that MACOED and FHSA may be hard to directly applied to bigger scale of epistasis analysis.

4.1.2 ESMO versus MDR for 3-locus epistasis on pure datasets

Because FHSA-SED and MACOED are specific for 2-locus epistasis analysis, in this section, we only compare ESMO with MDR. For pure datasets with 100 SNPs, we find that both MDR and ESMO can completely recognize the true functional 3-locus epistasis, reaching the power 100% as shown in Table 4. However, the running time cost of MDR is about 2 times of ESMO-parfor. For ESMO-unparfor, the seconds of handling 3-locus on 100 SNPs datasets is about 2400, which is not listed here.

For a 3-locus epistasis, ESMO-parfor consumes about 743819 seconds (about 8 days) to handle 100 datasets and each dataset contains 800 samples and 1000 SNPs. ESMO also achieve 100% precision for detecting 3-locus epistasis. Here, we do not list the results of MDR, because we rough estimate it may take 500 hours to handle 100 datasets.

4.1.3 The speedup

The time complexity of ESMO is $O(n^k * S)$, in which n is the number of SNPs and k is the epistasis order and S denotes the cost of model selection. In this study, we directly use 'parfor' structure to accelerate the outermost loop of k-nesting loop. The results show that the speedups are bigger than 2.5 on a computing platform with

Table 3. The results of running time on pure datasets

Datasets	MDR (s)	MACOED (s)	FHSA-SED (s)	ESMO-parfor (s)	ESMO-unparfor (s)
DG1_1	537	2375	177	44	67
DG1_2	540	2019	177	45	67
DG2_1	531	2005	177	44	68
DG2_2	548	2013	176	43	67
DG3_1	539	2007	177	44	67
DG3_2	549	2004	177	44	67
DG6_1	4456	121100	4585	1402	6058
DG6_2	4561	121356	4430	1413	6071

Table 4. The 3-locus epistasis detection results

Datasets	MDR		ESMO-parfor	
	Power (%)	Running time (s)	Power (%)	Running time (s)
DG7_1	100	1613	100	785
DG7_2	100	1621	100	782
DG8_1	100	1624	100	786
DG8_2	100	1615	100	781
DG9_1	100	1615	100	783
DG9_2	100	1787	100	782
DG13_1	#	More than 500 hours	100	743819
DG13_2	#	More than 500 hours	100	744821

Note: '#' means that no results available, because it consumes too much time.

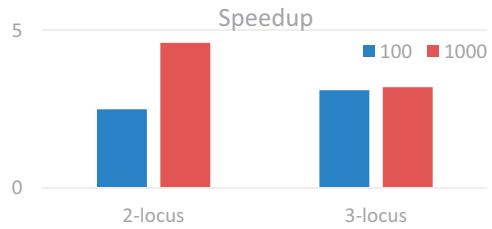

Fig. 3. The speedup of ESMO

6 cores in Figure 3. As the scale of datasets increases, the increment of 2-locus analysis is significant. However, the increment on 3-locus is marginal. The main reason for this is that the most time-consuming part of the calculation has changed from 2-locus to 3-locus analysis. It means that as the order of epistasis grows, the proportion of parts that cannot be parallelized is getting higher. In general, exhaustive method is easy to be parallelism.

4.2 Experiments on heterogeneous datasets

To show the capability of ESMO to concurrently handle heterogeneity and epistasis, we use both the power and loss evaluation measures. Note that because MDR needs to manually count the results of each single dataset, we only extracted 10%, 20% and 30% of datasets in each group for convenience. Note that just two epistatic model (H1 and H2) are applied to simulate heterogeneous datasets by GAMETES.

4.2.1 ESMO versus MDR for 2-locus epistasis on heterogeneous datasets

As defined in Formula 7, the loss denotes the number of true 2-locus epistatic combinations missed. As shown in Table 5, we find that ESMO did not make any missing. For MDR, we investigate all top

Table 5. The loss of ESMO and MDR for 2-locus heterogeneity analysis

The proportion of entire datasets (%)	DG4		DG5	
	ESMO (%)	MDR (%)	ESMO (%)	MDR (%)
10	0	0	0	10
20	0	5	0	10
30	0	6.7	0	6.7

Table 6. The power of ESMO for detecting H1 and H2

Dataset	Power (%)
DG4	H1=94, H2=93
DG5	H1=100, H2=99

models and compare all these true epistatic models with top models. In some cases, none of true epistatic models has been recognized by MDR, resulting in the loss.

As shown in Table 6, even for exhaustive search, H1 and H2 cannot be fully recognized due to the mixture of different disease models. However, for a 2-locus heterogeneous datasets, ESMO can achieve the powers higher than 90%.

4.2.2 ESMO versus MDR for 3-locus epistasis on heterogeneous datasets

From Table 7, we can see that all the loss values of ESMO are no bigger than MDR. It is also interesting to find that the performance of both ESMO and MDR on DG11 are better than DG10 and DG12, respectively. It indicates that the heterogeneity proportions of heterogeneous datasets will confuse computational method to some degree. For example, the loss of MDR is nearly 50% on DG10 and DG12.

For DG14 and DG15, in these complex situations, the effects of ESMO and MDR are listed in Tables 7 and 8. For the perspective of the loss, as shown in Table 7, the loss of MDR is significantly higher than ESMO both on DG 14 and DG15. Note that the effects of ESMO and MDR are obviously weakened on DG 15 compared with DG14. This phenomenon is likely due to the changes on the HP of each subgroups.

Table 8 shows some kinds of relationship between the power and the HP. For the datasets DG10 and DG12 have the same HP (50%, 50%), which means that epistatic models H1 and H2 occupy the same proportion in the population. Then, the powers of H1 and H2 are similar. But for DG11 with the HP (60%, 40%), the powers of them are significantly different. The reason for this may be that one epistatic model with high proportion has a relative big sample

Table 7. The loss of ESMO and MDR for 3-locus epistasis analysis

The proportion of entire datasets (%)	DG10		DG11		DG12		DG14		DG15	
	ESMO (%)	MDR (%)	ESMO (%)	MDR (%)	ESMO (%)	MDR (%)	ESMO (%)	MDR (%)	ESMO (%)	MDR (%)
10	20	40	10	10	30	50	10	50	60	80
20	15	60	5	5	40	50	5	35	70	90
30	16.6	43.3	3.3	16.6	20	43.3	3.3	33.3	76.7	93.3

Table 8. The powers of ESMO for detecting all disease models

Dataset	Power (%)
DG10	H1=62,H2=53
DG11	H1=94,H2=16
DG12	H1=38,H2=42
DG14	H1=93,H2=1,H3=1
DG15	H1=21,H2=4,H3=4

size, so that this model could dominate the other epistatic model. The results of DG 14 and DG15 also confirm this view.

4.3 Experiments on a breast cancer dataset

We conduct heterogeneity and epistasis analysis on a breast cancer dataset both 2-locus and 3-locus. For the non-dominant sorting and top k selection algorithm, we also set the k to be 3. Firstly, after 2-locus epistatic analysis by ESMO, 3 kinds of 2-locus epistasis have been returned such as (rs3020314, rs2017591), (rs1514348,rs2017591) and (rs2077647,rs2017591). Taken a further statistical significance analysis, only (rs3020314, rs2017591) meets the significance level with p-value 0.00038895. With using MACOED and MDR, (rs3020314, rs2017591) is also considered to be strongly associated with breast cancer, which is consistent with the results of ESMO.

ESMO is also used to analyze 3-locus epistasis. (rs3020314, rs1514348, rs2017591), (rs10046, rs3020314, rs2017591) and (rs6269, rs3020314, rs2017591) are considered to be relative with breast cancer on these samples. However, MDR returns (rs3020314, rs9478249, rs2017591) considered to be the best model, which is slightly different with ESMO. The rs3020314 and rs1514348 returned by ESMO are on the estrogen receptor 1 (ESR1) which encodes an estrogen receptor, a ligand-activated transcription factor composed with several domains playing important role in hormone binding, DNA binding, and transcription activation. The results of ESMO mean that these two SNPs located in the same gene ESR1 may interact with rs2017591, resulting in the pathological processes of breast cancer.

It is interesting that rs2017591 is the most common SNP existing in all these epistatic combinations. The rs2017591 locates in gene STS (Chromosome X) which refers to the biosynthetic pathway for estrogen. If these epistatic combinations contribute to different subtypes of breast cancer, thereby rs2017591 is common pathogenic SNP between these subtypes.

5 Discussion and conclusion

For epistasis analysis, as the order increases, the combination space expands exponentially. It is a great challenge for computational methods to search the true epistasis genotypes in such a huge space. Two main difficulties hindered us from recognizing the functional epistasis: complex disease model and limited calculation resources. Although the time complexity of swarm intelligent algorithm may be lower than exhaustive method in theory, in practice it is very hard to tune the parameters settings and the volatility of solutions is unsuitable for clinic research. More importantly, it is very challenging for swarm intelligent algorithm to pick up k SNPs step by step, because in pure epistasis the subset of k-locus has no advantages than others. It means that no useful information can be provided by the subsets. Therefore, it is difficult to converge to the best solution, especially when the k is big.

In this study, we adopt the exhaustive search like MDR, ensure the power as the first place. The ESMO has 3 major features: (i) Concurrently handle epistasis and heterogeneity based on multi-objective optimization. Here, mutual entropy and Bayesian network derived scores are combined to profile epistatic models from different perspectives; (ii) Exhaustive search has been criticized by their huge computation burden. In this work, unlike MDR with embedded cross-validation, multi-objective applied to select model is far faster. In addition, a well-designed non-dominant sorting and top k selection algorithm reduces the time complexity from $O(M * N^2)$ to $O(k * M * N)$. More importantly, calculation sharing and trading space for time are also applied to accelerate ESMO; 3) Our method ESMO ensures the globalization of solutions, which can significantly improve the confidence of clinical researchers for bioinformatics.

Although the results of 2-locus and 3-locus epistasis analysis show that ESMO is powerful than other methods, there are still some work that needs to be improved. First of all, in nature, there may be $k > 3$ SNPs involved in some kinds of complex diseases. As the number of SNPs and the epistasis order k increase, the cost of computing resources grows exponentially. Two ways to ease this are dimensions reduction and high performance computing platform.

On the other hand, prior knowledge should be borrowed to quicken the process of search. In addition, prior knowledge may directly reduce the k-locus epistasis to be lower, resulting in narrowed combination space. At the same time, it also enriches biological meanings to computational approaches.

Acknowledgements

We are grateful to the anonymous reviewers whose suggestions and comments contributed to the significant improvement of this paper.

Funding

This paper is partially supported by the National Natural Science Foundation of China (Serial No.61602174), the Jiangxi Provincial natural science fund (No. 20161BAB212052 and 20151BAB217011) and the Scientific and Technological Research Project of Education Department in Jiangxi Province (GJJ150496).

Conflict of Interest: none declared.

References

Abo,A.S. and Farnir,F. (2017) KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies. *BMC Bioinformatics*, **18**, 184.

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

Fenger,M. *et al*. (2008) Analysis of heterogeneity and epistasis in physiological mixed populations by combined structural equation modelling and latent class analysis. *BMC Genet*., **9**, 43.

Gabriel,C.A. *et al*. (2013) Association of progesterone receptor gene (PGR) variants and breast cancer risk in African American women. *Breast Cancer Res. Treat*., **139**, 833–843.

González-Domínguez,J. *et al*. (2015) Parallel pairwise epistasis detection on heterogeneous computing architectures. *IEEE T Parall. Distr*., **27**, 2329–2340.

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

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.

Moore,J.H. *et al*. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol*., **241**, 252–261.

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

Skwark,M.J. *et al*. (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet*., **13**, e1006508.

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

Urbanowicz,R.J. *et al*. (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Min*., **5**, 16.

Urbanowicz,R.J. *et al*. (2013) Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach. *J. Am. Med. Inform. Assoc*., **20**, 603–612.

Weeks,N.T. *et al*. (2016) High-performance epistasis detection in quantitative trait GWAS. *Int. J. High Perform. C*, 1–16.

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

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

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

Zeng,Z. *et al*. (2016) Discovering causal interactions using Bayesian network scoring and information gain. *BMC Bioinformatics*, **17**, 221.