ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Enhancement of a stochastic Markov blanket framework with ant colony optimization, to uncover epistasis in genetic association studies

Christine Sinoquet[1] and Clément Niel[1] *

1 - LS2N, UMR CNRS 6004, Université de Nantes
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex, France

Abstract.
In association genetics, many studies rely on univariate statistical tests to reveal genotype-phenotype relationships, and are thus prone to miss the situations of epistasis (interaction between genes). We designed SMMB (Multiple Stochastic Markov blankets), and SMMB-ACO, a variant combined with ant colony optimization, to detect epistasis. We compare our proposals with three other methods. SMMB-ACO outperforms the other methods for 50% of simulated datasets. On real datasets, the detection ability of SMMB-ACO is close to that of the best approach, which is a slow method, and SMMB-ACO is the fastest algorithm behind a much less performing method.

1 Introduction

In spite of a certain amount of progress, our knowledge of the genetic architecture of complex phenotypes (*e.g.*, diseases) is still very limited. All studies involving univariate statistical tests are prone to miss the situations of epistasis, where genes (or Single Nucleotide Polymorphisms (SNPs)), interact together to determine some studied phenotype. A recent review has been dedicated to epistasis detection [1]. Exhaustive strategies are either constrained to the exploration of 2-way interactions or are not scalable. Machine learning has contributed to this line of investigation through various proposals: ensemble learning techniques based on random forests, metaheuristics designed for combinatorial optimization and Bayesian network-based methods. Even though various approaches have been put forward to identify epistatic interactions, their common flaw is the lack of detection power, especially when epistatic interactions involve SNPs with no of feeble marginal effect on the phenotype. This situation is called "pure" epistasis hereafter.

In this work, we explore the Markov blanket approach, to tackle epistasis detection, and we introduce SMMB, an innovative hybrid approach which combines Markov blanket construction with stochastic and ensemble features. Moreover, our second proposal, SMMB-ACO, guides the stochastic sampling procedure by incorporating ant colony optimization.

*C. Niel was supported by the Research project GRIOTE (Pays de la Loire region, France) and the EGID LabEx (Lille, France). The software development and the realization of experiments were performed in part at the CCIPL (Centre de Calcul Intensif des Pays de la Loire, Nantes, France). The authors wish to thank G. Rocheleau for his help on adaptive permutations.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Section 2 presents an essential property related to the Markov blanket (MB) concept and briefly explains the shortcomings of existing MB-based algorithms when dealing with epistasis detection. Section 3 introduces SMMB. The variant SMMB-ACO is presented in Section 4. Experimental results and discussion are presented in Section 5.

2 The Markov blanket concept

In a Bayesian network, the Markov blanket of a target variable T, $MB(T)$, is the minimal set of variables that can render T independent from all other variables in $\mathbf{V}$, that do not belong to $MB(T)$: $\forall\ X \in \mathbf{V}\ X \perp\!\!\!\perp T \mid \mathbf{MB}$. To grow efficiently an optimal MB from the empty set, several proposals were made following the pioneer algorithm IAMB [2], with variations around the design and interleaving of the forward phase, to admit candidate SNPs into the MB under construction, and of the backward phase, to discard false positives. The conditional test above mentioned is one of the essential ingredients of these methods. In this line, DASSO-MB was designed to tackle epistasis detection [3]. The main flaws reported for these methods are the lack of detection power, non-scalability in high-dimensional settings, poor performances for imperfect data (not verifying the faithfulness property required by several algorithms), and impossibility to detect pure epistasis when adding variables one at a time. Indeed, at first iteration, performing tests of independence conditional on the *empty* MB biases the whole process since a variable marginally dependent with the target variable is included. Therefore, "pure" epistasis cannot be detected. SMMB palliates this issue by incorporating groups of variables instead.

3 Description of the SMMB algorithm

The data provided to SMMB consists of D, a matrix describing the p variables of set $\mathbf{V}$, for each of n observations, together with T, a vector of size n representing the target variable. In association studies, D describes p SNPs for each of n subjects (affected and unaffected) and T is a vector of phenotypes (affected / unaffected status). SMMB outputs a Markov blanket for target variable T.

The SMMB algorithm involves three main procedures. The top level procedure drives the construction of n_{mbs} suboptimal Markov blankets. Each such MB is learned from a subset $\mathbf{D_K}$ of K variables sampled from complete dataset $\mathbf{D}$. The construction of each MB is delegated to procedure **learnMB**. Once n_{mbs} MBs are built, or a maximum number of iterations (n_1) is reached, a consensus is built from all MBs constructed so far, and is further refined.

The stochastic procedure **learnMB** attempts to construct a suboptimal Markov blanket MB through a forward phase interleaved with backward phases. Its sketch is the following:

1. Initialize **MB** to the empty set.
2. Sample a subset $\mathbf{S_q}$ of q variables from subset $\mathbf{D_K}$.
3. For each of the $2^q - 1$ non-empty subsets of $\mathbf{S_q}$, compute $score_A$, a score of association with the target variable T, conditional on the current MB.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

4. Identify $\mathbf{S}$, the subset that maximizes $score_A$ over all subsets of $\mathbf{S_q}$. Assess statistical significance for conditional independence test $\mathbf{S} \perp\!\!\!\perp T \mid \mathbf{MB}$, using type I error threshold α'. If conditional independence is rejected, include $\mathbf{S}$ in $\mathbf{MB}$ and complete a full backward phase.
5. Repeat steps 2 to 4 until the non-empty Markov blanket remains unchanged, or a maximal number of iterations (n_2) is reached while $\mathbf{MB}$ is still empty.

At step 4, the backward phase seeks to discard false positives from the Markov blanket under construction. For this purpose, it iteratively examines each variable X of the current MB, to identify whether there exists a subset $\mathbf{S}$, $\mathbf{S} \subseteq \mathbf{MB}$, verifying: $X \perp\!\!\!\perp T \mid \mathbf{S}$. If statistical significance is assessed for conditional independence at type I error threshold α', X is discarded from $\mathbf{MB}$.

3.1 Score of association

Step 2 of procedure **learnMB** has to evaluate groups of variables that are candidate to inclusion in $\mathbf{MB}$. For this purpose, a score of association between a group of variables and the target variable, conditional on $\mathbf{MB}$, must be defined: $score_A(\mathbf{S}, T, \mathbf{MB})$. In the current version of the heuristic SMMB, this score is defined as follows:

$$score(\mathbf{S}, T, \mathbf{MB}) = \max_{X \in \mathbf{S}} \{score(X, T, \mathbf{MB} \cup (\mathbf{S} \setminus X))\},$$

where $score(X, T, \mathbf{C})$ is the statistic returned by the conditional G-test of independence between X and T, given set $\mathbf{C}$. Thus, a group of variables candidate to inclusion in $\mathbf{MB}$ is effectively considered as a whole, since variables are tested conditional on the MB enriched with the group but one variable.

3.2 Building the Markov blanket consensus

Once n_{mbs} suboptimal Markov Blankets (at the most) are built, a MB consensus is constructed. In the current version of SMMB, the consensus is initialized to the union of the suboptimal MBs. Then, this set is refined through a complete backward phase. This backward phase follows the same scheme as in procedure **learnMB**, with the notable exception that this time, correction for multiple testing is introduced. The correction is performed based on *adaptive* permutations, which are less time consuming than standard permutations.

4 SMMB with ant colony optimization

In the ACO (ant colony optimization) framework applied to Markov blanket construction, each ant is assigned a sample $\mathbf{D_K}$, and applies procedure **learnMB** on it. In SMMB, K variables (at top level), and q variables (in procedure **learnMB**) are repeatedly drawn from the uniform law. Enriching SMMB with an ACO feature allows to govern the sampling of variables based on probability distribution $\mathbb{P}$:

$$\forall\ X \in \mathbf{V},\ \mathbb{P}(X) = \frac{\tau(X)^\alpha \cdot \eta(X)^\beta}{\sum_{Y \in \mathbf{V}}\ \tau(Y)^\alpha \cdot \eta(Y)^\beta}, \tag{1}$$

where $\tau(X)$ is the pheromone rate for variable X. In the feature selection problem undelying epistasis detection, the pheromone rate $\tau(X)$ deposited by the ants

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

indicates the significance of X to contribute to interactions with other variables, to determine the target variable. $\eta(X)$ intends to integrate prior knowledge (*e.g.*, biological knowledge). Parameters α and β allow to adjust the relative weights between pheromome rate and prior knowledge.

In SMMB-ACO, the top level procedure runs n_{aco} iterations. Each of these iterations drives n_a ants. Thus $n_{aco} \times n_a$ suboptimal Markov blankets (at most) are constructed, to further build the MB consensus. Each time a score of association is computed for a given ant (see Section 3.1), the statistics for all conditional independence tests triggered by this computation are memorized by the ant. At the end of each of the n_{aco} iterations, cooperation between the ants is achieved by enriching a global memory with the ants' feedbacks. This global memory is used to update τ, the vector of pheromone rates required to update distribution $\mathbb{P}$ (Equation 1).

5 Experiments

SMMB and SMMB-ACO were implemented in C++, using OpenMP. The software packages are available at https://ls2n.fr/listelogicielsequipe/DUKe/128/ and https://ls2n.fr/listelogicielsequipe/DUKe/130/. The methods compared were run using six cores composed of biprocessors XEON 5462 2.66 GHz.

In this section, we first describe the experimental protocol. Then we present the comparisons of SMMB and SMMB-ACO with three different approaches, on simulated and real data sets.

5.1 Experimental settings

Simulated data sets We used the software program GAMETES [4] to simulate case-control data sets harbouring epistasis patterns under a disease model. We considered three situations: M1 and M2 model 2-way epistasis, whereas M3 is a 3-way epistasis model. For each model, we generated 100 data sets (100 SNPs, $2,000$ cases, $2,000$ controls). The SNPs in epistasis were assigned the same minor allele frequency (MAF). We varied this common MAF in $\{0.05, 0.1, 0.2, 0.5\}$.

Real data sets The genome-wide Rheumatoid Arthritis (RA) data set was provided by the Wellcome Trust Case Control Consortium (see Table 1). We ran ten executions of each stochastic method on each of the 23 human chromosomes, and on the whole genome.

Methods compared SMMB and SMMB-ACO were compared to BEAM [5], DASSO-MB [3] and AntEpiSeeker [6]. Each of these three methods shares a feature with our algorithms. BEAM relies on a Bayesian framework, this time using Monte-Carlo Markov Chains. DASSO-MB implements Markov blanket learning. AntEpiSeeker relies on ant colony optimization. The criterion retained

2,938 controls; 1,860 cases	
total number of SNPs for the 23 chromosomes	469,612
number of SNPs per chromosome	min = 5,754; max = 38,867; median = 21,477

Table 1: Characteristics of the Rheumatoid Arthritis (RA) data set.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

simulations	SMMB	$n_{mbs} = 100$; $n_1 = 1,000$; $K = 10$	$q = 3$
	SMMB-ACO	$n_{aco} = 34$; $n_a = 3$; $K = 10$; Parameters kept constant through all experiments: τ initialized to 100 for each variable; η initialized to 1 for each variable; $\alpha = \beta = 1$	$n_2 = 30$
RA data set 23 separate chromosomes	SMMB	$n_{mbs} = 40,000$; $n_1 = 100$; $K = 180$	$\alpha' = 0.05$
	SMMB-ACO	$n_{aco} = 10,000$; $n_a = 4$; $K = 180$	
RA data set whole genome	SMMB	$n_{mbs} = 50,000$; $n_1 = 10,000$; $K = 600$	
	SMMB-ACO	$n_{aco} = 8,333$; $n_a = 6$; $K = 600$	
simulations	BEAM	1,000 and 10,000 iterations in burning and stationary phases	
	DASSO-MB	$\alpha = 0.05$	
	AntEpiSeeker	450 iterations; 1,000 ants; $\alpha = 0.01$	
RA data set 23 separate chromosomes & whole genome	BEAM	10^6 and $10'$ iterations in burning and stationary phases	
	DASSO-MB	$\alpha = 0.05$	
	AntEpiSeeker	30,000 iterations; 50,000 ants; $\alpha = 0.01$	

Table 2: Parameter settings for the five methods compared.

for the comparison study on simulated data was the F-measure $= 2/(1/recall + 1/precision)$, with $recall = TP/(TP + FN)$ and $precision = TP/(TP + FP)$.

The parameter settings for the five methods are given in Table 2.

5.2 Results

Figure 1 shows that SMMB performs slightly better than the other methods on model M1. SMMB also performs better than the other methods for model M2 and MAFs equal to 0.10 and 0.50. SMMB-ACO apart, SMMB and BEAM are always in the top 2 methods for model M3. Above all, we observe that in half of the simulations, SMMB-ACO is one of the two best and quasi similarly performing methods. Besides, fastest to slowest, on simulations, one encounters DASSO-MB, SMMB-ACO / SMMB, BEAM and AntEpiSeeker (see Table 3).

In the real data sets, a pattern of epistasis is reported by SMMB or SMMB-

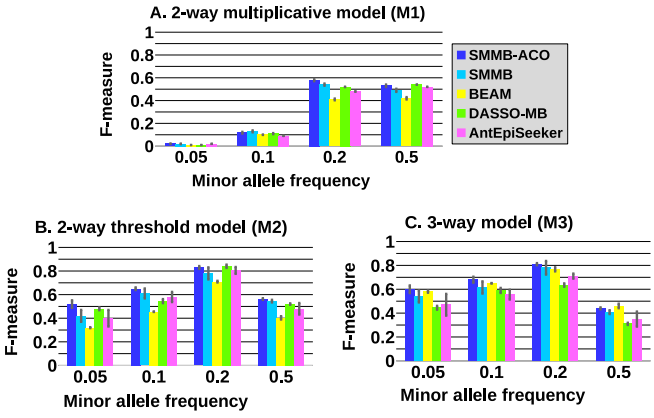

Fig. 1: Comparison of performances on simulated data.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

ACO if one of the sub-optimal MBs produced is retrieved in the final refined consensus. It was necessary to run several executions of SMMB, SMMB-ACO and AntEpiSeeker to obtain 7 putative 2-way epistatic interactions (results not shown). Each run of stochastic software BEAM yielded these 7 interactions. The deterministic algorithm DASSO-MB only retrieved 5 of these interactions. Table 3 shows a decrease in running times from SMMB to SMMB-ACO. On the whole genome set, 90% of the runs of SMMB-ACO output at least 6 of the 7 interactions (versus 80% for AntEpiSeeker which is slower than SMMB-ACO) (results not shown). Given this performance of 90%, and since SMMB-ACO is at least 2.8 times as fast as BEAM, it is affordable to launch several runs of SMMB-ACO. To note, we have observed that the construction of the MB consensus consumes between 50% and 60% of the total running time.

	SMMB	SMMB-ACO	BEAM	DASSO-MB	AntEpiSeeker
simulations	30s	30s	93s	5s	469s
23 separate chromosomes (total)	34h	13h	59h	12h	69h
whole genome	23h	19h	53h	17h	47h

Table 3: Comparison of running times.

6 Conclusion

SMMB was designed to cope with pure epistasis and imperfect data in high-dimensional settings. Incorporating feedback on the Markov blanket learning process *via* the ACO-related probability distribution allowed to enhance the performance of SMMB. Besides, in contrast to all other Markov blanket learning algorithms, correction for multiple testing based on adaptive permutations is a crucial ingredient in SMMB. In spite of the computational cost entailed, we showed that SMMB and SMMB-ACO are the top fastest in the panel of methods compared, including at the genome scale. In addition, we showed that SMMB-ACO frequently outperforms the other methods on the simulated data sets. On real data sets, the detection ability of SMMB-ACO is rather close to the optimum of the top-ranked method, a rather slow method. These promising results support continued effort to enhance the SMMB-based approach.

References

[1] C. Niel, C. Sinoquet, C. Dina and G. Rocheleau, A survey about methods dedicated to epistasis detection, *Frontiers in Genetics*, 6:285, 2015

[2] I. Tsamardinos, C. Aliferis and A. Statnikov, Algorithms for large scale Markov blanket discovery. In *proceedings of the 16th International FLAIRS Conference*, 376-380, 2003.

[3] B. Han, M. Park and X.-W. Chen, A Markov blanket-based method for detecting causal SNPs in GWAS, *BMC Bioinformatics*, 11(Suppl.3):S5, 2010.

[4] R. Urbanowicz, J. Kiralis, N. Sinnott-Armstrong, T. Heberling, J. Fisher and J. Moore, GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures, *BioData Mining*, 5:16, 2012.

[5] Y. Zhang, J. S. Liu, Bayesian inference of epistatic interactions in case-control studies, *Nature Genetics*, 39:1167-1173, 2007.

[6] 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 Research Notes*, 3:117, 2010.