A Fast Algorithm for Learning Epistatic Genomic Relationships

Xia Jiang, PhD[1], Richard E. Neapolitan, PhD[3], M. Michael Barmada[2], PhD, Shyam Visweswaran, MD, PhD[1], Gregory F. Cooper, MD, PhD[1]
[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
[2]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA
[3]Department of Computer Science, Northeastern Illinois University, Chicago, IL

Abstract

Genetic epidemiologists strive to determine the genetic profile of diseases. Epistasis is the interaction between two or more genes to affect phenotype. Due to the often non-linearity of the interaction, it is difficult to detect statistical patterns of epistasis. Combinatorial methods for detecting epistasis investigate a subset of combinations of genes without employing a search strategy. Therefore, they do not scale to handling the high-dimensional data found in genome-wide association studies (GWAS). We represent genome-phenome interactions using a Bayesian network rule, which is a specialized Bayesian network. We develop an efficient search algorithm to learn from data a high scoring rule that may contain two or more interacting genes. Our experimental results using synthetic data indicate that this algorithm detects interacting genes as well as a Bayesian network combinatorial method, and it is much faster. Our results also indicate that the algorithm can successfully learn genome-phenome relationships using a real GWAS dataset.

Introduction

Genetic epidemiologists endeavor to determine the genetic profile of diseases. As an example, the $\varepsilon 4$ allele of the APOE gene has been established as a risk factor for late-onset Alzheimer's disease (LOAD) [1]. Genes do not always affect phenotype according to simple Mendelian inheritance. Rather several genes may have a joint effect on phenotype even though individually one or more have little effect [2]. For example, results in [3] indicate that the GAB2 gene is statistically relevant to LOAD when the APOE $\varepsilon 4$ allele is present, but GAB2 alone has no association with LOAD. *Epistasis* is the interaction between two or more genes to affect phenotype. Biologically, epistasis refers to interactions between biomolecules occurring in an organism. Statistically, epistasis refers to interactions between multiple loci such that the net affect on phenotype cannot be predicted by simply combining the effects of the individual loci. The individual loci may exhibit no marginal effects. The phenotype discussed here is the presence of a disease. An epistatic relationship in which each of the interacting loci exhibits no marginal effect on the disease cannot be learned using single-locus methods. So, methods for analyzing combinations of genes have been developed. Parametric methods include logistic regression [4] and nonparametric ones include combinatorial methods, genetic programming, neural networks, and random forests [5].

Combinatorial methods investigate a subset of combinations of loci without employing a search strategy. An example of a combinatorial method is Multifactor dimensionality reduction (MDR) [2,6]. MDR combines two or more variables into a single variable, thereby leading to dimensionality reduction. MDR has been applied to detecting epistatic interactions in several domains including breast cancer [3] and cardiovascular disease [7] using datasets with relatively few genetic loci. For example, the breast cancer study investigated 10 loci.

A common type of genetic variation is the single nucleotide polymorphism (SNP). To study the underlying genetic variants of common diseases, genome-wide association studies (GWAS) that simultaneously assay hundreds of thousands of SNPs are being increasingly used. Using existing methods, it is difficult to analyze epistasis using a GWAS dataset. For example, consider a combinatorial method. If we only investigated all 1, 2, 3 and 4-SNP combinations when there are 500,000 SNPs, we would need to investigate 2.604×10^{21} combinations. Given this difficulty, so far the data obtained from a GWAS have usually been analyzed using single-locus methods [3, 8].

Much of the genetic risk of many common diseases is unknown. This is called the *dark matter* of genetic risk, and a great deal of it is believed to be due to hard-to-detect genetic interactions [9]. The advent of GWAS data sets affords us unprecedented opportunity to discover these interactions. So the analysis of multi-locus interactions using GWAS data sets is a vital problem. Recently, penalized linear regression (PLR) has been applied to this problem [10]. Park and Hastie [11] compared PLR to MDR.

Bayesian networks are a leading architecture for representing uncertainty in artificial intelligence [12]. We represent the relationships between SNPs and disease status using a Bayesian network, and we develop an efficient search algorithm for learning the Bayesian network containing the SNPs associated with the disease from data. We present experimental results of using both synthetic data and a GWAS dataset. In the experiments with synthetic data, our algorithm detects interacting SNPs as well as does a combinatorial Bayesian network method, but requires much less time. When analyzing a real Alzheimer's disease dataset [3], the algorithm finds SNPs that are consistent with those identified by the original investigators [3]. Those investigators identified SNPs using the same dataset, prior domain knowledge, and a single-locus search method. It took our algorithm 4.1 hours to do this study. We estimate that it would take the combinatorial method about 3.71 years.

Method

Bayesian Networks. Suppose we have a joint probability distribution *P* of the random variables in some set *V* and a directed acyclic graph (DAG) $G = (V, E)$, where *E* denotes the set of arcs among the variables in *V*. We say that (*G*, *P*) satisfies the *Markov condition* if for each variable $X \in V$, *X* is conditionally independent of the set of all its nondescendents given the set of all its parents. If (*G*,*P*) satisfies the Markov condition, we call (*G*,*P*) a *Bayesian network* (*BN*) [12]. In a BN the joint probability distribution of the variables equals the product of the conditional probability distributions of each variable given its parents in *G*, whenever these conditional distributions exist. That is, if our variables are $X_1, X_2, \ldots, X_n$, and PA_i is the set of parents of X_i, then

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid PA_i).$$

Methods for learning both the structure (DAG) and parameters in BNs from data have been developed [12]. One method for learning the DAG is to score all DAGs using a scoring criterion. One such score is the *Bayesian score* [13], which is the probability of the data given the DAG. Other scores include those based on the minimum description length principle [14], namely the *minimum description length* (*MDL*) score [15] and the minimum message length (*MML*) score [16]. A *consistent scoring criterion* for BNs assigns the highest score to a concise DAG containing the generative distribution when the dataset is sufficiently large. The scores just mentioned are all consistent.

To learn a DAG from data we can score all DAGs using one of these scores and then choose the highest scoring DAG. However, if the number of variables is not small, the number of candidate DAGs is forbiddingly large. So heuristic algorithms have been developed to search over the space of DAGs [12].

A BN learning algorithm called *Greedy Equivalent Search* (*GES*) [17] can learn the most concise DAG representing a probability distribution under the assumptions that the scoring criterion is consistent and that the probability distribution admits a faithful DAG representation and satisfies the composition property [12]. Briefly, the algorithm starts with the empty DAG and greedily adds the edge to the DAG that increases the score the most until no edge increases the score. Then it greedily deletes the edge from the DAG such that the deletion increases the score the most until no deletion decreases the score.

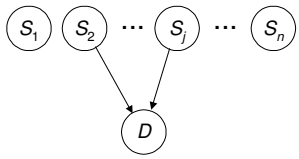

Figure 1. A DDAG.

The MBS Algorithm. A BN model for the relationship between *n* SNPs and a disease variable *D* is shown in Figure 1. This specialized BN structure can be viewed as a BN rule that predicts the disease status given SNPs as predictors. This model assumes there are *n* SNPs and one or more SNPs may have edges to *D*. We call such models *Direct DAGs* (*DDAGs*). This model was used in [18] to learn epistatic relationships using a combinatorial method.

In general, the GES algorithm could not detect epistatic interactions by searching over DDAGs. Suppose that two SNPs predict disease variable *D* strongly, but neither SNP alone predicts *D*. Such an epistatic relationship does not satisfy the composition property which the GES algorithm requires in this type of situation to find the two predictive SNPs.

Our approach to this problem is to do greedy search starting with every SNP rather than just the single SNP that increases the score of the empty DAG most. In this way every SNP pair will be investigated. We conjecture that many (but not all) forms of epistasis will be detected by extending greedy search from a single SNP to every SNP. Such search is tractable

even when the number of SNPs is in the thousands. The algorithm follows ($score(A_i)$ is the score of the model that has edges from the SNPs in set A_i to D):

```
for each SNP SNP_i
   A_i = { SNP_i };
   score_i = score(A_i);
   do
      if adding any SNP to A_i increases score_i
         then add SNP to A_i that increases score_i most;
         score_i = score(A_i);
   while adding some SNP increases score_i;
   do
      if deleting any SNP from A_i increases score_i
         then delete SNP from A_i that increases score_i most;
         score_i = score(A_i);
   while deleting some SNP increases score_i;
endfor;
report the k highest scoring models;
```

We call this algorithm *Multiple Beam Search* (*MBS*). Its worst-case time complexity is $O(n^3)$, where n is the total number of SNPs. In practice we would add at most m SNPs in the first step, where m is a parameter, resulting in a time complexity of $O(mn^2)$.

The MBS algorithm is effective for handling epistatic interactions in which we have a group A of k SNPs interacting, each of them by itself is probabilistically independent of the disease, there is a probabilistic dependence between the disease and at least one pair of the interacting SNPs, and each of the other k - 2 SNPs in A predicts the disease given the pair.

Experiments

First, we compared the performances of a BN combinatorial method and MBS using synthetic datasets. Then we used MBS to analyze a real dataset.

Synthetic Datasets. A dataset developed in [6] was used in our experiment concerning synthetic data. This dataset was created as follows. The developers created 70 different probabilistic relationships in which 2 SNPs combined are correlated with the disease, but neither SNP is individually correlated. The relationships represented various degrees of penetrance, heritability, and minor allele frequency. Supplementary Table 1 to [6] shows the details of the 70 models. Datasets were then developed having a case-control ratio of 1:1. To create one dataset they fixed the model. Based on the model, they then developed data concerning the two SNPs that were related to the disease in the model, 18 other unrelated SNPs, and the disease. For each of the 70 models, 100 datasets were developed, making a total of 7000 datasets. They followed this procedure for dataset sizes of 200, 400, 800, and 1600.

Real Dataset. Studies indicate that the apoplipoprotein E (APOE) gene is associated with many cases of late-onset Alzheimer's disease (LOAD) [1]. The APOE gene has three common variants $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$. The least risk is associated with the $\varepsilon 2$ allele, while each copy of the $\varepsilon 4$ allele increases risk.

Reiman et al. [3] investigated the association of 312,316 SNPs separately in APOE $\varepsilon 4$ carriers and in APOE $\varepsilon 4$ noncarriers. A discovery cohort and two replication cohorts were used in the study. Within the discovery subgroup consisting of APOE $\varepsilon 4$ carriers, 10 of the 25 SNPs exhibiting the greatest association with LOAD (contingency test p-value 9×10^{-8} to 1×10^{-7}) were located in the GRB-associated binding protein 2 (GAB2) gene on chromosome 11q14.1. Associations with LOAD for 6 of these SNPs were confirmed in the two replication cohorts. Combined data from all three cohorts exhibited significant association between LOAD and all 10 GAB2 SNPs in APOE $\varepsilon 4$ carriers. These 10 SNPs were not significantly associated with LOAD in the APOE $\varepsilon 4$ non-carriers. The researchers also provided immunohistochemical validation for the relevance of GAB2 to the neuropathology of LOAD.

In our second experiment, we used the combined dataset consisting of all three cohorts investigated in [3]. This dataset contains data on 1411 subjects.

All experiments were run on a PC running Windows XP with a 2.19 GHz processor and 1 GB of RAM.

Results. We analyzed the synthetic data using the following methods: 1) a Bayesian network combinatorial method, which we call *BayCom*, and which scored all 1-SNP, 2-SNP, 3-SNP, and 4-SNP DDAGs; 2) MBS with a maximum of m = 4 SNPs added in the first step. Candidate models were scored with the MML score. This score has previously been used successfully in causal discovery [13].

Table 1 shows the number of times the correct model scored highest over all 7000 datasets. Important detection measures include *recall*, *precision*, and the *overlap coefficient*. In the current context they are as follows. Let S be the set of SNPs in the correct model and T be the set of SNPs in the highest scoring model. Then *recall* = $\#(S \cap T)/\#(S)$, *precision* = $\#(S \cap T)/\#(T)$, and *overlap coef.* = $\#(S \cap T)/\#(S \cup T)$, where # returns

the number of items in a set. Table 2 shows the average values of these measures. MBS performed as well as BayCom in terms of accuracy and the other measures. Table 3 shows the running times. MBS was up to 28 times faster than BayCom.

Size	MBS	BayCom
200	4049	4049
400	5111	5111
800	5881	5881
1600	6463	6463

Table 1. Number of times the correct model scored highest out of 7000 datasets for MBS and Baycom.

	Spatial Recall		Precision		Overlap Coef.	
Size	**MBS**	**BCom**	**MBS**	**BCom**	**MBS**	**BCom**
200	0.593	0.593	0.607	0.607	0.593	0.593
400	0.737	0.737	0.744	0.744	0.737	0.737
800	0.843	0.843	0.846	0.846	0.843	0.843
1600	0.925	0.925	0.926	0.926	0.925	0.925

Table 2. Comparisons of average values of detection measures over 7000 datasets for MBS and Baycom.

Size	MBS	BayCom
200	0.108	2.0
400	0.191	5.15
800	0.361	9.61
1600	0.629	18.0

Table 3. Average running times in sec. over 7000 datasets.

The real data set was analyzed as follows. Using all 1411 cases, we pre-processed the data by scoring all DDAG models in which APOE and one of the 312,316 SNPs are each parents of LOAD. We then selected the SNPs from the highest-scoring 1000 models. So SNPs showing even weak association with LOAD were selected. Next MBS was run using the dataset consisting of APOE and these 1000 SNPs. We did not constrain APOE to be in the discovered models. At most $m = 4$ nodes were added in the first step of MBS. There were 4.175×10^{10} models under consideration. MBS actually scored far fewer models.

# models in top 10 containing a GAB2 SNP	# models in top 100 containing a GAB2 SNP	# rs6094514 occurrences with GAB2 in top 10	# rs6094514 occurrences with GAB2 in top 100
6	36	6	33

Table 4. Occurrences of GAB2 and rs6094514 in high-scoring models when using MBS to analyze 1000 SNPs along with APOE.

We recorded the 1000 highest scoring models encountered in the MBS search. APOE appeared in every one of these models, and a GAB2 SNP appeared in the top two models. Columns one and two in Table 4 show the number of times a GAB2 SNP appeared respectively in the top 10 models and top 100 models. Of the 312,316 SNPs in the study, 16 are GAB2 SNPs. Seven of these 16 SNPs appeared in at least one of the 36 high-scoring models containing a GAB2 SNP.

All of these seven SNPs were among the 10 GAB2 SNPs identified in [3]. The probability of 36 or more of the top 100 models containing at least one of the 16 GAB2 SNPs by chance is 2.0806×10^{-106}. GAB2 SNPs never occurred together in a model. This pattern is plausible since each GAB2 SNP may represent the dependence between LOAD and GAB2, and therefore it could render LOAD independent of the other GAB2 SNPs. Our results substantiate those in [3], that GAB2 (or something in linkage disequilibrium with it) has an affect on LOAD. Our results do not indicate whether GAB2 influences LOAD by interacting with APOE since APOE appears in every high-scoring model.

The run time was 4.1 hours. When we analyzed 1, 2, and 3 SNP combinations involving only 200 SNPs in the LOAD dataset, the run time for BayCom was 1.04 hours. We extrapolated that it would take about 3.71 years to analyze all 1, 2, 3, and 4 combinations involving 1001 loci (1000 SNPs plus APOE).

We obtained an unexpected result. We noticed that SNP rs6094514, which is an intron on the EYA2 gene on chromosome 20, often appeared along with GAB2 and LOAD. So we investigated how often this occurred. The third and fourth columns in Table 4 show the numbers of such occurrences respectively in the top 10 and top 100 models. Among the top 100 models, SNP rs6094514 only occurred once without GAB2. As it turns out, prior research has associated this SNP with LOAD. In a cross-platform comparison of outputs from four GWAS, Shi et al. [16] found SNP rs6094515 to be associated with LOAD with a combined p-value of 8.54×10^{-6}. However, we know of no prior literature showing that GAB2 and EYA2 may interact to affect LOAD, as our results seem to suggest. MBS discovered this possibility because it is able to tractably investigate multi-loci interactions.

Another result was that SNP rs473367 on chromosome 9 appeared in the 3rd and 4th models and in 22 of the top 99 models. It never appeared with GAB2. A previous study [20] suggested that this SNP interacts with APOE to affect LOAD. Our results support this association, but indicate no interaction with GAB2.

Discussion

We represented the relationships between SNPs and a disease using a BN rule, and we created the MBS algorithm for learning the BN rule containing the

SNPs associated with the disease from data. Results of experiments using synthetic data and real data showed the effectiveness of the algorithm in terms of learning and the efficiency of the algorithm in terms of run time. We substantiated previous results that GAB2 may affect LOAD. We obtained new results that EYA2 may interact with GAB2 to affect LOAD. This conjecture bears further investigation.

The MBS algorithm provides a way of learning epistatic relationships from GWAS datasets. It is effective in handling situations in which there are several interacting SNPS, most exhibit no marginal effect on the disease, and at least one pair of the SNPs exhibits a marginal effect. A limitation of the method is that it requires that at least one pair of interacting SNPs has a marginal effect. If this is not the case, then we would have to search every triplet or more of SNPs, which is far less computationally feasible. An open question is how many epistatic relationships have the property that at least one pair exhibits a marginal effect on the disease. Another limitation is that we must put a fairly small maximum (around 7) on the number of nodes investigated in the forward search because the time complexity of the BN score is exponential in terms of the number of parents. So high-order interactions will be missed.

Finally, MBS and methods like it are about discovery. The next question is what to do with the discovery. First, the significance (with Bonferroni correction) of the highest scoring models can be reported. If that significance is sufficiently high for a model, the model can be further investigated using additional data analysis and a study of its biological plausibility.

Acknowledgements

The research reported here was funded in part by grants R01-LM010020 and 5-T15-LM007059-23 from the National Library of Medicine.

References

1. Coon KD, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J. Clin. Psychiatry 2007;68:613-618.
2. Ritchie MD, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American Journal of Human Genetics 2001;69:38-147.
3. Reiman EM, et al. GAB2 alleles modify Alzheimer's risk in APOE $\varepsilon 4$ carriers. Neuron 2007;54:713-720.
4. Millstein J, Siegmund KD, Conti DV, Gauderman WJ. Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis, BMC Genetics 2005;6(Suppl. 1):S147.
5. Heidema A, Boer J, Nagelkerke N, Mariman E, van der AD, Feskens E. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genetics 2006;7(23).
6. Valez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 2007;31:306-315.
7. Coffey CS, et al. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. BMC Bioinformatics 2004;5(49).
8. Lambert JC, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nature Gen. 2009;41:1094-1099.
9. Park M, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics 2008; 9(1):30-50.
10. Galvin A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. Trends in Genetics; 2010;26(3):132-41.
11. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Genome Anal.;2009;25:714-721.
12. Neapolitan RE. Learning Bayesian Networks. (Prentice Hall, Upper Saddle River, NJ, 2004).
13. Cooper, GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 1992;9:309-347.
14. Rissanen J. Modeling by shortest data description. Automatica 1978;14:465-471.
15. Suzuki J. Learning Bayesian belief networks based on the minimum description length principle: basic properties. IEICE Transactions on Fundamentals 1999;E82-A:2237-2245.
16. Korb K, Nicholson AE. Bayesian Artificial Intelligence. (CRC, Boca Raton, FL, 2003).
17. Chickering D, Meek C. Finding optimal Bayesian networks. in Darwiche A, Friedman N (eds.): Uncertainty in Artificial Intelligence; Proceedings of the Eighteenth Conference (Morgan Kaufmann, San Mateo, CA, 2002).
18. Visweswaran S, Wong AI, Barmada M. A Bayesian method for identifying genetic interactions. AMIA 2009 Symposium Proceedings:673-677.
19. Shi H, et al. Analysis of genome-wide association study (GWAS) data looking for replicating signals in Alzheimer's disease (AD). Int J Mol Epidemiol Genet 2010;1(1):53-66.
20. (WO/2008/131364) NO EN TITLE. World Intellectual Property Organization. www.wipo.int /portal/index.html.en.