Chapter 9
Mining Epistatic Interactions from High-Dimensional Data Sets

Xia Jiang[1], Shyam Visweswaran[1], and Richard E. Neapolitan[2]

[1] Department of Biomedical Informatics, University of Pittsburgh
[2] Department of Computer Science, Northeastern Illinois University

Abstract. Genetic epidemiologists strive to determine the genetic profile of diseases. Two or more genes can interact to have a causal effect on disease even when little or no such effect can be observed statistically for one or even both of the genes individually. This is in contrast to Mendelian diseases like cystic fibrosis, which are associated with variation at a single genetic locus. This gene-gene interaction is called epistasis. To uncover this dark matter of genetic risk it would be pivotal to be able to discover epistatic relationships from data. The recent availability of high-dimensional data sets affords us unprecedented opportunity to make headway in accomplishing this. However, there are two central barriers to successfully identifying genetic interactions using such data sets. First, it is difficult to detect epistatic interactions statistically using parametric statistical methods such as logistic regression due to the sparseness of the data and the non-linearity of the relationships. Second, the number of candidate models in a high-dimensional data set is forbiddingly large. This paper describes recent research addressing these two barriers. To address the first barrier, the primary author and colleagues developed a specialized Bayesian network model for representing the relationship between features and disease, and a Bayesian network scoring criterion tailored to this model. This research is summarized in Section 2. To address the second barrier the primary author and colleagues developed an enhancement of Greedy Equivalent Search. This research is discussed in Section 3. Background is provided in Section 1.

1 Introduction

Genetic epidemiologists strive to determine the genetic profile of diseases. For example, the $\epsilon 4$ allele of the APOE gene has been established as a risk factor for late-onset Alzheimer's disease (Coon et al., 2007; Papassotiropoulos et al., 2006; Corder et al., 1993). However, often genes do not affect phenotype according to the simple rules developed by Mendel (Bateson, 1909). Rather two or more genes can interact to have a causal effect on phenotype even when little or no such effect can be observed statistically for one or even both of the genes individually. For example, (Rieman et al., 2007) found that the GAB2 gene seems to be statistically relevant to Alzheimer's disease when the ϵ_4 allele of the

D.E. Holmes, L.C. Jain (Eds.): Data Mining: Found. & Intell. Paradigms, ISRL 25, pp. 187–209.
springerlink.com

APOE gene is present, but GAB2 by itself exhibits no statistical relevance to the disease. This is in contrast to Mendelian diseases like cystic fibrosis, which are associated with variation at a single genetic locus.

This gene-gene interaction is called *epistasis*. Much of the genetic risk of many common diseases remains unknown and is believed to be due to epistasis. This is referred to as the *dark matter* of genetic risk (Galvin et al., 2010). To uncover this dark matter of genetic risk it would be pivotal to be able to discover epistatic relationships from data.

The *dimension* of a data set is the number of attributes in the data set. The recent availability of high-dimensional data sets affords us unprecedented opportunity to learn the etiology of disease from data. For example, the advent of high-throughput technologies has enabled *genome-wide association studies* (*GWAS* or *GWA* studies) (Wang et al., 1998; Matsuzaki et al., 2004), which involve sampling in cases and controls around 500,000 genetic loci. The government has invested heavily in studies that produce these high-dimensional data sets, and the initial results of these studies have been gratifying in that they have suggested a number of previously unsuspected etiologic pathways (Manolio, 2009). However, analysis of these data sets has not yet yielded the level of disease-associated feature discoveries originally anticipated (Wade, 2010). This could well be do to the difficulty with discovering epistatic interactions using such data sets.

There are two central barriers to successfully identifying genetic interactions using high-dimensional sets. First, it is difficult to detect epistatic interactions statistically using parametric statistical methods such as logistic regression due to the sparseness of the data and the non-linearity of the relationships (Velez et al., 2007). So we need to develop efficacious methods for evaluating candidate epistatic models. Second, the number of candidate models in a high-dimensional data set is forbiddingly large. For example, if we only examined all 1, 2, 3, and 4 loci models when there are 100,000 loci, we would need to examine about 4.17×10^{18} models. Since we do not have the computational power to investigate so many models, we need efficient algorithms that enable us to only investigate the most promising ones.

This paper describes recent research addressing these two barriers. Since the research described here pertains to data sets containing many different possible causal features including both genetic and environmental factors, we will refer to the risk factors simply as *features*. The phenotype need not be disease status (e.g. it could be height or math ability); however, for the sake of focus we will use disease terminology throughout. To address the first barrier just identified, the primary author and colleagues developed a specialized Bayesian network model for representing the relationship between features and disease, and a Bayesian network scoring criterion tailored to this model (Jiang et al., 2010a). This research is summarized in Section 3. To address the second barrier the primary author and colleagues developed an enhancement of Greedy Equivalent Search (Chickering, 2003) called Multiple Beam Search (Jiang et al., 2010b). This research is discussed in Section 4. First, we provide some background.

2 Background

We review epistasis, GWAS, a well-known epistatic learning method called multifactor dimensionality reduction, and Bayesian networks.

2.1 Epistasis

Biologically, *epistasis* refers to gene-gene interaction when the action of one gene is modified by one or several other genes. Statistically, epistasis refers to interaction between genetic variants at multiple loci in which the net effect on disease from the combination of genotypes at the different loci is not accurately predicted by a combination of the individual genotype effects. In general, the individual loci may exhibit no marginal effects.

Example 1. *Suppose we have two loci G_1 and G_2, disease D, and the alleles of G_1 are A and a, whereas those of G_2 are B and b. Suppose further that we have the probabilities (relative frequencies in the population) in the following table:*

	AA (.25)	Aa (.5)	aa (.25)
BB (.25)	0.0	0.1	0.0
Bb (.5)	0.1	0.0	0.1
bb (.25)	0.0	0.1	0.0

The entries in the table denote, for example, that

$$P(D = yes|G_1 = Aa, G_2 = BB) = 0.1.$$

The heading AA (.25) means 25% of the individuals in the population have genotype AA. We also assume that G_1 and G_2 mix independently in the population (no linkage). We then have the following (we do not show the random variables G_i for brevity):

$$\begin{aligned} P(D = yes)|AA) &= P(D = yes|AA, BB)P(BB) + \\ &\quad P(D = yes|AA, Bb)P(Bb) + P(D = yes|AA, bb)P(bb) \\ &= 0.0 \times .25 + 0.1 \times 0.5 + 0.0 \times .25 = .05 \end{aligned}$$

$$\begin{aligned} P(D = yes)|Aa) &= P(D = yes|Aa, BB)P(BB) + \\ &\quad P(D = yes|Aa, Bb)P(Bb) + P(D = yes|Aa, bb)P(bb) \\ &= 0.1 \times .25 + 0.0 \times 0.5 + 0.1 \times .25 = .05 \end{aligned}$$

$$\begin{aligned} P(D = yes)|aa) &= P(D = yes|aa, BB)P(BB) + \\ &\quad P(D = yes|aa, Bb)P(Bb) + P(D = yes|aa, bb)P(bb) \\ &= 0.0 \times .25 + 0.1 \times 0.5 + 0.0 \times .25 = .05 \end{aligned}$$

So if we look at G_1 alone no statistical correlation with D will be observed. The same is true if we look at G_2 alone. However, as can be seen from the above table, the combinations AABb, AaBB, Aabb, and aaBb make disease D more probable.

It is believed that epistasis may play an important role in susceptibility to many common diseases (Galvin et al, 2010). For example, Ritchie et al. (2001) found a statistically significant high-order interaction among four polymorphisms from three estrogen pathway genes (COMT, CYP1B1, and CYP1A1) relative to sporadic breast cancer, when no marginal effect was observed for any of the genes.

2.2 Detecting Epistasis

It is difficult to detect epistatic relationships statistically using parametric statistical methods such as logistic regression due to the sparseness of the data and the non-linearity of the relationships (Velez et al., 2007). As a result, non-parametric methods based on machine-learning have been developed. Such methods include combinatorial methods, set association analysis, genetic programming, neural networks and random forests (Heidema et al., 2006).

Combinatorial methods search over all possible combinations of loci to find combinations that are predictive of the phenotype. The combinatorial method *multifactor dimensionality reduction* (*MDR*) (Hahn et al., 2005) combines two or more variables into a single variable (hence leading to dimensionality reduction); this changes the representation space of the data and facilitates the detection of nonlinear interactions among the variables. MDR has been successfully applied in detecting epistatic interactions in complex human diseases such as sporadic breast cancer, cardiovascular disease, and type II diabetes (Ritchie et al., 2001; Coffey et al., 2004; Cho et al., 2004).

A combinatorial method must focus on a relatively small number of loci to be tractable. For example, if we examined all 1, 2, 3, and 4 loci subsets of 100,000 loci, we would need to examine about 4.17×10^{18} subsets. The successes of MDR were achieved by identifying relatively few relevant loci up front. For example, in the sporadic breast cancer discovery the focus was on five genes known to produce enzymes in the metabolism of estrogens.

2.3 High-Dimensional Data Sets

To uncover the dark matter barrier of genetic risk it would be pivotal to be able to discover epistatic relationships from data without identifying relevant loci up front. The recent availability of high-dimensional data sets affords us unprecedented opportunity to accomplish this. For example, the advent of high-throughput technologies has enabled *genome-wide association studies* (*GWAS* or *GWA* studies) (Wang et al., 1998; Matsuzaki et al., 2004). A *single nucleotide polymorphism* (*SNP*) is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a species. Usually, the less frequent allele must be present in 1% or more of the population for a site to

qualify as a SNP (Brooks, 1999). A GWAS involves sampling in an individual around 500,000 representative SNPs that capture much of the variation across the entire genome. The initial results of these studies has been gratifying. Over 150 risk loci have been identified in studies of more than 60 common diseases and traits. These associations have suggested previously unsuspected etiologic pathways (Manolio, 2009). For example, a GWAS in (Reiman et al, 2007) identified the GAB2 gene as a risk factor for Alzheimer's disease, and a GWAS in (Hunter et al., 2007) found four SNPs in intron 2 of FGFR2 highly associated with breast cancer. These initial GWAS successes were achieved by analyzing the association of each locus individually with the disease. Their success notwithstanding, analysis of high-dimensional data sets has not yet yielded the level of disease-associated feature discoveries originally anticipated (Wade, 2010). To realize the full potential of a GWAS and perhaps approach maximizing what we can discover from them, we need to analyze the effects of multiple loci on a disease (i.e. epistasis).

Realizing this, recently researchers have worked on developing methods for simultaneously analyzing the effects of multiple loci using high-dimensional data sets. *Lasso* is a shrinkage and selection method for linear regression (Tibshirani, 1996). It has been used successfully in problems where the number of predictors far exceeds the number of observations (Chen et al., 1998). So, researchers applied lasso to analyzing the multiple effects of loci on disease based on GWAS data (Wu et al., 2009; Wu et al. 2010). There are two difficulties with this procedure. First, as mentioned earlier, regression has difficulty handling a nonlinear epistatic relationship. Second, loci interactions with no marginal effects will not be detected at all unless we include terms for pairwise interactions. Wu et al. (2010) do this, but now we are faced with the combinatorial explosion problem discussed above. Another strategy involves using permutation tests (Zhang et al., 2009). These methods use standard statistic analysis to investigate different ensembles of two-loci analyses. Other methods include the use of ReliefF (Moore and White, 2007; Epstein and Haake, 2008), random forests (Meng et al. 2007), predictive rule inference (Wan et al., 2010), a variational Bayes algorithm (Longston et al. 2010), a Bayesian marker partition algorithm (Zhang and Liu, 2007), the Bayesian graphical method developed in (Verzilli et al., 2006), and the Markov blanket-base method discussed in (Han et al., 2009). The Bayesian graphical method does approximate model averaging using *Markov chain Monte Carlo* (*MCMC*) and is unlikely to scale up. The Markov blanket-based method uses a G^2 test and forward search investigating one locus at a time. Such as search would miss a loci-loci interaction that had no marginal effects.

2.4 Barriers to Learning Epistasis

As mentioned earlier, there are two central barriers to successfully identifying potential interactions in the etiology of disease using high-dimensional sets. First, we need to find an efficacious way to evaluate candidate models that can identify relationships like epistasis which exhibit little or no marginal effects. That is, even if we had the computational power to investigate all subsets of a

high-dimensional set, we would want to evaluate each subset using a method that has been shown to learn epistatic relationships well. Second, since we do not have the computational power to investigate all subsets, we need efficient algorithms that enable us to only investigate the most promising subsets. Not surprisingly, Evans et al. (2006) showed that as the marginal effects of two loci approach zero, the power to detect a two-loci interactions approaches zero unless we exhaustively investigate all two-loci interactions. None of the methods discussed above circumvent or solve this problem. Their evaluations were performed using high-dimensional data sets in which there were marginal effects.

As mentioned earlier, recent research by the primary author and colleagues addressing these barriers is discussed in Section 3 and Section 4. First, we provide further review.

2.5 MDR

Multifactor dimensionality analysis (*MDR*) (Hahn et al., 2003) is a well-known method for detecting epistasis. In Section 3 we include MDR in a comparison of the performance of two methods. So we review MDR using an example taken from (Velez et al., 2007).

Example 2. *Suppose we are investigating whether* SNP_1 *and* SNP_2 *are correlated with disease* D*. Suppose further that we obtain the data depicted in Figure 1 (a). The number of individuals with the disease appears on the left in each cell, whereas the number without it appears on the right. For example, of those individuals who have* $SNP_1 = 0$*,* 49 *have the disease and* 44 *do not have the disease. We see from Figure2 (a) that neither SNP by itself seems correlated with the disease.*

Using MDR we investigate whether the SNPs together are correlated with the disease. First, we take the cross product of all values of the SNPs as shown on the left in Figure 1 (b). For each combination of values of the SNPs we determine how many individuals have the disease and how many do not, as also shown in that figure. Let $\#D$ *be the number who have the disease and* $\#noD$ *be the number who do not have the disease. For a given SNP combination we call the combination high-risk (HR) if* $\frac{\#D}{\#noD} > T$*, where* T *is a threshold. Ordinarily,* $T = 1$ *if in total we have the same number of individuals with the disease as without the disease. Using this threshold, the number of high-risk individual for each SNP combination is the number labeled HR on the left in Figure 1 (b).*

Next we create a new binary variable $SNP_1 \times SNP_2$ *whose value is* HR *if the SNP combination is one of the high-risk SNPs and whose value is* LR *(low-risk) if the SNP combination is one of the low-risk SNPs. We then compute the total number of individuals who have value* HR *and have the disease, have value* HR *and do not have the disease, have value* LR *and have the disease, and have value* LR *and do not have the disease. These totals appear on the right in Figure 1 (b). For example, consider those individuals who have value* HR *and have the disease. We obtain the total as follows:*

$$46 + 49 + 46 + 59 = 200.$$

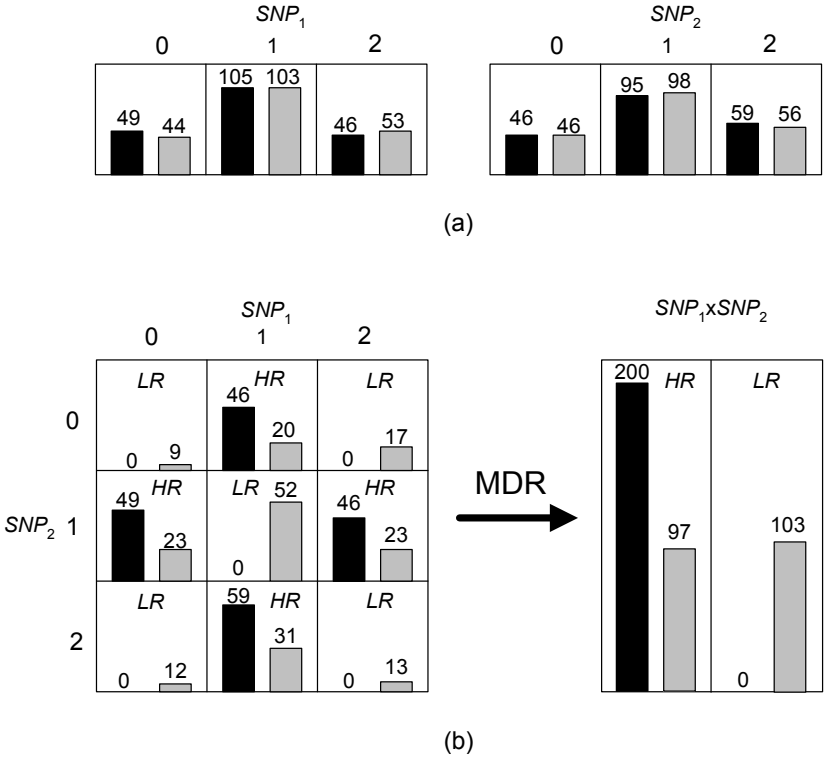

Fig. 1. The number of individuals with the disease appears on the left in each cell, whereas the number without it appears on the right. (a) shows the numbers for each value of each SNP individually. (b) shows the the number for the cross product of the values of the SNPs and the numbers for the binary-valued variable $SNP_1 \times SNP_2$ obtained using MDR. The values of this variable are HR (high-risk) and LR (low-risk).

We see from the right part of Figure 1 (b) that the disease appears to be correlated with the cross product of the SNPs.

We illustrated MDR for the case where we are investigating the correlation of a 2-SNP combination with a disease. Clearly, the method extends to three or more SNPs. If we are investigating n SNPs and considering k-SNP combinations with a disease, we investigate all $\binom{n}{k}$ combinations and choose the combination that appears best according to some criterion. (Velez et al., 2007) use the following *classification error* as the criterion:

$$\frac{(\#\text{ individuals with HR and no disease}) + (\#\text{ individuals with LR and disease})}{\#\text{ individuals}}.$$

For example, for the SNPs illustrated in Figure 1 the classification error is

$$\frac{97+0}{400} = 0.2425.$$

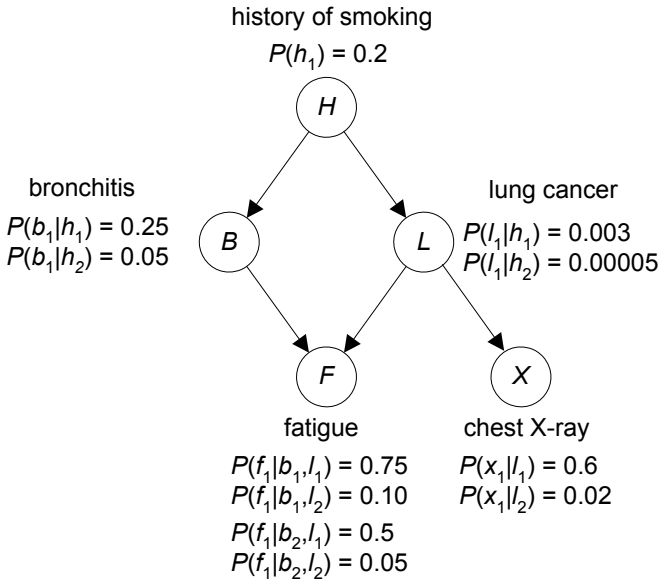

Fig. 2. A BN modeling lung disorders. This BN is intentionally simple to illustrate concepts; it is not intended to be clinically complete.

(Velez et al., 2007) determine a likely SNP combination using 10-fold cross validation. That is, they break the data set into 10 partitions, they use 9 partitions to learn a k-SNP combination (the one with the smallest classification error), and then using the learned SNP combination they determine the prediction error for the remaining partition. They repeat this procedure for all possible choices of the 9 learning partitions, and then they take the average. This process is repeated for all $1, 2, ..., k$ SNP combinations that are computationally feasible. A model is chosen for each number of SNPs. A final model is chosen from this set of models based on minimizing the average prediction error and maximizing cross-validation consistency.

If the number of individuals with the disease is not the same as the number of individuals without the disease, some adjustments need to be made. (Velez et al., 2007) discuss using *balanced accuracy* and an *adjusted threshold* to handle this situation.

2.6 Bayesian Networks

The epistasis discovery method presented in Sections 3 and 4 is based on Bayesian networks; so, we review them next.

Bayesian networks (Neapolitan, 2004; Koller and Friedman, 2009) are increasingly being used for modeling and knowledge discovery in many domains including bioinformatics (Neapolitan, 2009). A *Bayesian network* (*BN*) consists of a directed acyclic graph (DAG) G whose nodes are random variables and the

conditional probability distribution of each node given its parents in G. Figure 2 shows a BN. In that BN h_1, for example, means an individual has a smoking history, whereas h_2 means the individual does not.

Using a Bayesian network inference algorithm we can compute the probability of nodes of interest based on the values of other nodes. For example, for the Bayesian network in Figure 2, we could compute the probability that a patient has bronchitis or lung cancer given that the patient has a smoking history and positive chest X-ray.

Methods have been developed for learning both the parameters in a BN and the structure (DAG) from data. The task of learning a unique DAG model from data is called *model selection*. In the *constraint-based approach* (Spirtes et al., 1993, 2000) to model selection, we try to learn a DAG from the conditional independencies that the data suggest are present in the generative probability distribution . In a *score-based approach* (Neapolitan, 2004), we assign a score to a DAG based on how well the DAG fits the data. A straightforward score, called the *Bayesian score*, is the probability of the given the DAG. This score for discrete random variables is as follows (Cooper and Herskovits, 1992):

$$\begin{aligned} score_{Bayes}(G : Data) &= P(Data|G) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=1}^{r_i} a_{ijk})}{\Gamma(\sum_{k=1}^{r_i} a_{ijk} + \sum_{k=1}^{r_i} s_{ijk})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + s_{ijk})}{\Gamma(a_{ijk})} \end{aligned} \quad (1)$$

where r_i is the number of states of X_i, q_i is the number of different instantiations of the parents of X_i, a_{ijk} is the ascertained prior belief concerning the number of times X_i took its kth value when the parents of X_i had their jth instantiation, and s_{ijk} is the number of times in the data that X_i took its kth value when the parents of X_i had their jth instantiation.

The Bayesian score assumes that our prior belief concerning each of the probability distributions in the network is represented by a Dirichlet distribution. When using the $Dir(\theta_X; a_1, a_2, \ldots, a_r)$ distribution to represent our belief concerning the unknown probability distribution θ_X of random variable X, due to cogent arguments such as the one in (Zabell, 1982), it has become standard to represent prior ignorance as to the value of θ_X by setting all parameter equal to 1. That is, $a_1 = a_2 = \cdots = a_r = 1$. The parameters $\{a_{ij1}, a_{ij2}, \ldots a_{ijr_i}\}$ in Equation 1 are Dirichlet parameters representing our belief about θ_{ij}, which is the conditional probability distribution of X_i given that the parents of X are in their jth instantiation. To represent prior ignorance as to all conditional probabilities in the network, Cooper and Herskovits (1992) set $a_{ijk} = 1$ for all i, j, and k; they called this the K2 score.

Heckerman et al. (1995) noted a problem with setting all a_{ijk} to 1, namely that equivalent DAGs could end up with different Bayesian scores. For example, the DAGs $X \rightarrow Y$ and $X \leftarrow Y$ could obtain different scores. Heckerman et al. (1995) proved that this does not happen if we use a *prior equivalent sample* size α in the DAG. When using a prior equivalent sample size we specify the same prior sample size α at each node. If we want to use a prior equivalent sample

size and represent a prior uniform distribution for each variable in the network, for all i, j, and k we set

$$a_{ijk} = \frac{\alpha}{r_i q_i}.$$

When we determine the values of a_{ijk} in this manner, we call the Bayesian score the *Bayesian Dirichlet equivalence uniform (BDeu)* score.

Another popular way of scoring is to use the *Minimum Description Length* (*MDL*) Principle (Rissanen, 1978), which is based on information theory and says that the best model of a collection of data is the one that minimizes the sum of the encoding lengths of the data and the model itself. To apply this principle to scoring DAGs, we must determine the number of bits needed to encode a DAG G and the number of bits needed to encode the data given the DAG. Suzucki (1999) developed the following well-known MDL score:

$$score_{MDL}(G : Data) = \sum_{i=1}^{n} \frac{d_i}{2} \log_2 m - m \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} P(x_{ik}, pa_{ij}) \log_2 \frac{P(x_{ik}, pa_{ij})}{P(x_{ik})P(pa_{ij})}, \tag{2}$$

where n is the number of nodes in G, d_i is the number of parameters stored for the ith node in G, m is the number of data items, r_i is the number of states of X_i, x_{ik} is the kth state of X_i, q_i is the number of instantiations of the parents of X_i, pa_{ij} is the jth instantiation of the parents of X_i in G, and the probabilities are computed using the data. The first term is the number of bits required to encode the DAG model (called the *DAG penalty*), and the second term concerns the number of bits needed to encode the data given the model. Lam and Bacchus (1994) developed a similar MDL score.

Another score based on information theory is the *Minimum Message Length Score* (*MML*) (Korb and Nicholson, 2003).

If the number of variables is not small, the number of candidate DAGs is forbiddingly large. Furthermore, the BN structure learning problem has been shown to be NP-hard (Chickering, 1996). So heuristic algorithms have been developed to search over the space of DAGs during learning (Neapolitan, 2004). When the number of variables is large relative to the number of variables, many of the highest scoring DAGs can have similar scores (Heckerman, 1996). In this case approximate model averaging using MCMC may obtain better results than model selection (Hoeting et al., 1999).

3 Discovering Epistasis Using Bayesian Networks

First we describe a specialized Bayesian network model for representing epistatic interactions; then we discuss an MDL score tailored to this model; finally we provide experimental results evaluating the performance of this score.

3.1 A Bayesian Network Model for Epistatic Interactions

Consider all DAGs containing features and a phenotype D, where D is a leaf. We are not representing the relationship between gene expression levels. Rather

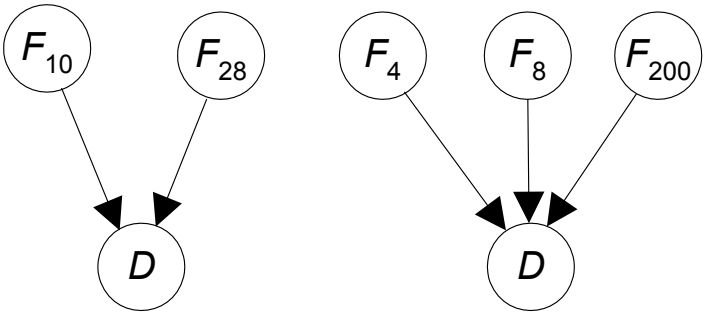

Fig. 3. Example DDAGs

we are representing the statistical dependence of the phenotype on alleles of the genes. So there is no need for edges between features, and we need only consider DAGs where the only edges are ones to D. Call such models *direct DAGs* (*DDAGs*). Figure 3 shows DDAGs. The size and complexity of the search space have been immensely reduced. Even so, there are still 2^n models where n is the number of features. In practice a limit is put on the number of parents.

3.2 The BNMBL Score

Jiang et al. (2010a) developed an MDL score tailored to DDAGs. Each parameter in a DAG model is a fraction with precision $1/m$, where m is the number of data items. So it takes $\log_2 m$ to store each parameter. However, as discussed in (Friedman and Yakhini, 1996), the high order bits are not very useful. So we can use only $\frac{1}{2}\log_2 m$ bits to store each parameter. In this way we arrive at the DAG penalty in Equation 2. Suppose now that k SNPs have edges to D in a given DDAG. That is, D has k parents. Since each SNP has three possible values, there are 3^k different values of these parents. The expected value of the number of data items that assume particular values of the parents is therefore $m/3^k$. If we approximate the precision for each parameter needed for D by this expected value, our DAG penalty for a DDAG is

$$\frac{3^k}{2}\log_2\frac{m}{3^k} + \frac{2k}{2}\log_2 m. \tag{3}$$

When this DAG penalty is used in the MDL score, Jiang et al. (2010a) called the score the *Bayesian network minimum bit length* (*BNMBL*) score.

3.3 Experiments

Jiang et al. (2010a) compared the performance of BNMBL using both simulated and real data sets. Results obtained using each of the data sets are discussed next.

3.3.1 Simulated Data Sets

(Velez et al., 2007) developed a set of simulated data sets concerning epistatic models. These data sets, which are available at http://discovery.dartmouth.edu/epistatic_data/ VelezData.zip, were developed as follows. First, they created 70 different probabilistic relationships (models) in which 2 SNPs combined are correlated with the disease, but neither SNP is individually correlated. The relationships represented various degree of *penetrance, heritability*, and *minor allele frequency*. *Penetrance* is the probability that an individual will have the disease given that the individual has a genotype that is associated with the disease. *Heritability* is the proportion of the disease variation due to genetic factors. The *minor allele frequency* is the relative frequency of the less frequent allele in a locus associated with the disease. Supplementary Table 1 to (Velez et al., 2007) shows the details of the 70 models.

Data sets were then developed having a case-control ratio (ratio of individuals with the disease to those without the disease) of 1:1. To create one data set 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 data sets were developed, making

Table 1. Columns MDR and BNMBL show the powers for MDR and BNMBL for models 55-59. The row labeled "Total" is the sum of the powers over the five models.

Size of Data Set	Model	MDR	BNMBL
200	55	3	7
	56	3	4
	57	3	5
	58	3	7
	59	3	3
	Total (200)	**15**	**26**
400	55	8	8
	56	7	9
	57	11	9
	58	15	27
	59	8	7
	Total (400)	**49**	**60**
800	55	26	30
	56	22	36
	57	25	29
	58	49	67
	59	18	24
	Total (800)	**140**	**186**
1600	55	66	81
	56	59	83
	57	68	81
	58	88	96
	59	49	63
	Total (1600)	**330**	**404**

a total of 7000 data sets. They followed this procedure for data set sizes of 200, 400, 800, and 1600. The task of a learning algorithm is to learn the 2-SNP model used to create each data set from that data set.

Method. The performances of MDR and BNMBL were compared using the simulated data sets just discussed. Jiang et al. (2010a) used MDR v. 1.25, which is available at www.epistasis.org, to run MDR, and developed their own implementation of BNMBL.

We say that a method correctly learns the model generating the data if it scores that model highest out of all $\binom{20}{2} = 190$ 2-SNP models. For a given model, let *power* to be the number of times the method correctly learned the model generating the data out of the 100 data sets generated for that model. The powers of MDR and BNMBL were compared using the data sets concerning the hardest-to-detect models and using all the data sets.

Results. Velez et al. (2007) showed that MDR has the lowest detection sensitivity for models 55-59 in Supplementary Table 1 to (Velez et al., 2007). These models have the weakest broad-sense heritability (0.01) and a minor allele frequency of 0.2. Table 1 shows the power for MDR and BNMBL for these 5 models. BNMBL outperformed MDR in 16 of the experiments involving the most difficult models, whereas MDR outperformed BNMBL only 2 times.

The first three columns of Table 2 shows the sums of the powers over all 70 models. The last two column shows p-values. These p-values were computed as follows. The Wilcoxon two-sample paired signed rank test was used to compare the powers of MDR and BNMBL over all 70 models. If we let the null hypothesis be that the medians of the power are equal, and the alternative hypothesis be that the median power of BNMBL is greater than that of MDR, then p is the level at which we can reject the null hypothesis. We see from Table 2 that BNMBL significantly out-performed MDR for all data set sizes.

Looking again at Table 1, we see that when we have a relatively large amount of data (1600 data items), BNMBL correctly identified $404 - 330 = 74$ more difficult models than MDR. From Table 2 we see that when there are 1600 data items BNMBL correctly identified $6883 - 6792 = 91$ more total models than MDR. The majority of the improvement obtained by using BNMBL concerns the more difficult models. Arguably, real epistatic relationships are more often represented by such difficult models.

Table 2. The columns labeled MDR,and BNMBL show the sums of the powers over all 70 data sets for each of the methods. The other column shows p-values. See the text for their description.

n	**MDR**	**BNMBL**	p
200	4904	5016	0.009
400	5796	5909	0.004
800	6408	6517	0.003
1600	6792	6883	0.012

Table 3. Mean running times in seconds for MDR and BNMBL

n	**MDR**	**BNMBL**
200	119.8	0.020
400	146.6	0.031
800	207.9	0.050
1600	241.7	0.097

Table 3 shows the mean running times in seconds obtained by averaging the running times over the data sets generated from all 70 genetic models. MDR is several orders of magnitude slower than BNMBL. The superior running time of BNMBL is due largely to its ability to use the entire data set for computing the score of each model, while MDR performs multi-fold cross-validation to score the models.

3.3.2 Real Data Set

It is well-known that the apoplipoprotein E (APOE) gene is associated with many cases of LOAD, which is characterized by dementia onset after age 60 (Coon et al., 2007; Papassotiropoulos et al., 2006; Corder et al., 1993). 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.

Coon et al. (2007) performed a genome-wide association study, which investigated over 300,000 SNPs, using 1086 LOAD cases and controls to determine the odds ratio (OR) associated with genes relative to LOAD. Only SNP rs4420638 on chromosome 19, which is located 14 kilobase pairs distal to and in linkage disequilibrium with the APOE gene, significantly distinguished between LOAD cases and controls.

Reiman et al. (2007) investigated the association of these same 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. See (Reiman et al., 2007) for the

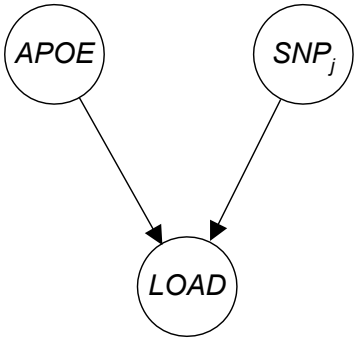

Fig. 4. A Bayesian networks in which the disease node (LOAD) has precisely two-parents, one being the APOE gene and the other being a SNP.

Table 4. The 28 highest scoring SNPs according to BNMBL. The column labeled GAB2 contains "yes" if the SNP is located on GAB2; the column labeled "Reiman" contains "yes" if the SNP is one of the 10 high scoring GAB2 SNPS discovered in (Reiman et al., 2007).

	SNP	$Score_{BNMBL}$	Chromosome	GAB2	Reiman
1	rs2517509	0.13226	6		
2	rs1007837	0.13096	11	yes	yes
3	rs12162084	0.13042	16		
4	rs7097398	0.13032	10		
5	rs901104	0.13019	11	yes	yes
6	rs7115850	0.13018	11	yes	yes
7	rs7817227	0.13009	8		
8	rs2122339	0.13002	4		
9	rs10793294	0.12997	11	yes	yes
10	rs4291702	0.12992	11	yes	yes
11	rs6784615	0.12986	3		
12	rs4945261	0.12963	11	yes	yes
13	rs2373115	0.12956	11	yes	yes
14	rs10754339	0.12932	1		
15	rs17126808	0.12932	8		
16	rs7581004	0.12929	2		
17	rs475093	0.12921	1		
18	rs2450130	0.12906	11	yes	
19	rs898717	0.12888	10		
20	rs473367	0.12884	9		
21	rs8025054	0.12873	15		
22	rs2739771	0.12863	15		
23	rs826470	0.12862	5		
24	rs9645940	0.12853	13		
25	rs17330779	0.12847	7		
26	rs6833943	0.12830	4		
27	rs2510038	0.12824	11	yes	yes
28	rs12472928	0.12818	2		

details of the cohorts. 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. These 10 SNPs were not significantly associated with LOAD in the APOE $\varepsilon 4$ noncarriers.

Reiman et al. (2007) also provided immunohistochemical validation for the relevance of GAB2 to the neuropathology of LOAD.

Method. Jiang et al. (2010a) investigated all 3-node Bayesian network models in which the disease node (LOAD) has precisely two-parents, one being the APOE gene and the other being one of the 312,260 SNPs investigated in (Reiman et al., 2007). Figure 4 shows one such a model. These models were scored with BNMBL using the combined data set consisting of all three cohorts described in (Reiman et al., 2007). The combined data set contains data on 1411 subjects. Of these subjects, 861 are LOAD cases, and 644 are APOE $\varepsilon 4$ carriers.

Results. Table 4 shows the 28 highest scoring SNPs. Since all DAG models have the same complexity, there is no need to include the term for DAG complexity in the score. So $Score_{BNMBL}$ consists only of term encoding the data without the minus sign, which means higher scores are better. We see that 7 of the top 13 SNPs were among the 10 SNPs discovered in (Reiman et al., 2007) and 9 of the top 27 SNPs are located in GAB2. The remaining high scoring SNPs are scattered among various chromosomes.

The results obtained using BNMBL substantiate the results in (Reiman et al., 2007), namely that GAB2 is associated with LOAD in APOE $\varepsilon 4$ carriers. This outcome demonstrates that BNMBL is a promising tool for learning real epistatic interactions.

An advantage of using BNMBL for knowledge discovery in this domain is that there is no need analyze the statistical relevance of a SNP separately under different conditions (e.g. first in all subjects, then in $\varepsilon 4$ carriers, and finally in $\varepsilon 4$ noncarriers). Rather we just score all relevant models using BNMBL.

4 Efficient Search

The second barrier to learning epistatic relationships from high-dimensional data sets is that we do not have the computational power to investigate very many subsets of the loci. So we need efficient algorithms that enable us to only investigate promising subsets.

Greedy Equivalent Search (*GES*) (Chickering, 2003) is an efficient Bayesian network learning algorithm that will 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. See (Neapolitan, 2004) for a complete discussion of these assumptions and the algorithm. Briefly, the algorithm starts with the empty DAG and in sequence greedily adds the edge to the DAG that increases the score the most until no edge increases the score. Then in sequence it greedily deletes the edge from the DAG until no edge decreases the score. It is not hard to see that if there are n variables, the worst-case time complexity of the algorithm is $\theta(n^2)$.

An initial strategy might be to try to learn the interacting SNPs by using the GES algorithm to search all DDAGs. However, a moment's reflection reveals that this could not in general work. Suppose we have an epistatic interaction between two SNPs and D such that each SNP is marginally independent of D

and all other SNPs are also independent of D. Suppose further that we have a data set so large that the generative distribution is represented exactly in the data set. In this case the GES algorithm would learn the correct DAG if its assumptions were met. However, in the first step of the algorithm all SNPs will score the same because they are all independent of D, none of them will increase the score and the algorithm will halt.

The problem is the epistatic interactions do not satisfy the composition property which is necessary to the GES algorithm. Jiang et al. (2010b) ameliorated this problem by initially expanding each of the SNPs using greedy search rather than initially starting only with the one that increases the score the most. In this way, we will definitely investigate every 2-SNP combination. If an epistatic interaction is occurring, two of the SNPs involved in the interaction may score high. Once we identify these two, often we should also often find possible 3rd and 4th and so on SNPs involved in the interaction. The algorithm follows. In this algorithm by $score(A_i)$ we mean the score of the model that has edges from the SNPs in A_i to D.

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

We call this algorithm *Multiple Beam Search* (*MBS*). It clearly requires $\theta(n^3)$ time in the worst case where n is the number of SNPs. However, in practice if the data set is large, we would add at most m SNPs in the first step, where m is a parameter. So the time complexity would be $\theta(mn^2)$.

This technique would not work if there is a dependence between k SNPs and D, but every proper subset of the k SNPs is marginally independent of D. The MBS algorithm is effective for handling the situation in which we have k SNPs interacting, each of them is marginally independent of the disease, and there is a dependence between the disease and at least one pair of the interacting SNPs. A reasonable conjecture is that many but certainly not all epistatic interactions satisfy this condition.

4.1 Experiments

Jiang el al. (2010b, 2010c) evaluated MBS using a simulated data set and two real GWAS data sets. The results are discussed next.

Table 5. Number of times the correct model scored highest out of 7000 data sets for MBS and Baycom

Size of Data Set	MBS	BayCom
200	4049	4049
400	5111	5111
800	5881	5881
1600	6463	6463

Table 6. Comparisons of average values of detection measures over 7000 data sets for MBS and Baycom

	Spatial Recall		Precision		Overlap Coefficient	
Data Set Size	MBS	BayCom	MBS	BayCom	MBS	BayCom
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	0.025

4.1.1 Simulated Data Sets

The simulated data sets developed in (Velez et al., 2007) (discussed in Section 3.3) were used in this evaluation which is taken from (Jiang et al., 2010b).

Method. The simulated data were analyzed using the following methods: 1) a Bayesian network combinatorial method, called *BayCom*, and which scores 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 mentioned in Section 2.6. This score has previously been used successfully in causal discovery (Korb and Nicholson, 2007).

Results. Table 5 shows the number of times the correct model scored highest over all 7000 data sets. 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 (# returns the number of items in a set)

$$recall = \frac{\#(S \cap T)}{\#(S)},$$

$$precision = \frac{\#(S \cap T)}{\#(T)},$$

$$overlapcoefficient = \frac{\#(S \cap T)}{\#(SUT)}.$$

Table 6 shows the average values of these measures over all 7000 data sets. MBS performed as well as BayCom in terms of accuracy and the other measures. Table 7 shows the running times. MBS was up to 28 times faster than BayCom.

Table 7. Average running times in seconds over 7000 data sets

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

Table 8. Occurrences of GAB2 and rs6094514 in high-scoring models when using MBS to analyze the LOAD data set

# 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

4.1.2 Real Data Set

The real data set analyzed using MBS was the LOAD data set introduced in (Reiman et al., 2007) (See Section 3.3).

Method. (Jiang et al, 2010b) analyzed the LOAD data set as follows. Using all 1411 cases, they pre-processed the data by scoring all DDAG models in which APOE and one of the 312,316 SNPs are each parents of LOAD (See Figure 4). They then selected the SNPs from the highest-scoring 1000 models. Next MBS was run using the data set consisting of APOE and these 1000 SNPs. They did not constrain APOE to be in the discovered models. At most $m = 3$ nodes were added in the first step of MBS. There were 4.175×1010 models under consideration. Of course MBS scored far fewer models.

Results. The 1000 highest scoring models encountered in the MBS search were recorded. APOE appeared in every one of these models, and a GAB2 SNP appeared in the top two models. Columns one and two in Table 8 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 (Reiman et al., 2007). 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. These results substantiate those in (Reiman et al., 2007), that GAB2 has an affect on LOAD. The 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 1, 2, and 3 SNP combinations involving only 200 SNPs in the LOAD data set were analyzed, the run time for BayCom was 1.04 hours. An extrapolation of this result indicates that it would take about 3.71 years to analyze all 1, 2, 3, and 4 combinations involving 1001 loci (1000 SNPs plus APOE).

An unexpected result was obtained. SNP rs6094514, which is an intron on the EYA2 gene on chromosome 20, often appeared along with GAB2 and LOAD. 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. (2010) found SNP rs6094515 to be associated with LOAD with a combined p-value of 8.54×10^{-6}. However, no prior literature shows that GAB2 and EYA2 may interact to affect LOAD, as the current 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 (WO/2008/131364) suggested that this SNP interacts with APOE to affect LOAD. The results discussed here support this association, but indicate no interaction with GAB2.

5 Discussion, Limitations, and Future Research

We showed a Bayesian network model for representing epistatic interactions called a DDAG and an MDL score called BNMBL designed specifically for this model. Using simulated data sets, BNMBL performed significantly better than MDR at identifying the two SNPs involved in an epistatic interaction. The BNMBL score performed well at identifying potential epistatic interactions from a real GWAS data set, as did the MML score.

We showed an algorithm called MBS that successfully identified potential epistatic interactions using a real GWAS data set. This algorithm requires quadratic time in terms of the number of SNPs from which we initiate beams (used as starting points for greedy search). If we initiated beams from all 500,000 SNPs in a given GWAS, quadratic time could take months. So in the study shown the data was pre-processed to identify the 1000 highest scoring individual SNPs from 2-parent models containing APOE and one of the 312,260 SNPs. Beams were then initiated from these 1000 SNPs and from APOE. In general, we would not suspect a gene such as APOE to be involved in the interaction. So our preprocessing would only involve scoring all 1-SNP models and choosing the 1000 highest scoring SNPs. If SNPs involved in an epistatic interaction exhibited absolutely no marginal effects, there is no reason they should appear in the top 1000 SNPs. So such an interaction would probably be missed. MBS has made progress in identifying epistasis when we either have a great deal of computational power or when at least one SNP shows a slight marginal effect. However,

further research is needed to investigate handling the situation where there are no marginal effects more efficiently.

After discovering candidate loci-phenotype relationships, researchers often report their significance using the Bonferroni correction or the False Discovery rate. However, some Bayesian statisticians (see e.g. (Neapolitan, 2008)) have argued that it is not reasonable to using these methods or any other method based on the number of hypotheses investigated, particularly in this type of domain. A simple example illustrating their argument is as follows: Suppose that one study investigates 100,000 SNPs while another investigates 500,000 SNPs. Suppose further that the data concerning a particular SNP and the disease is identical in the two studies. Due to the different corrections, that SNP could be reported as significant in one study but not the other. Yet the data concerning the SNP is identical in the two studies! It seems the only reason these corrections work at all is because they serve as surrogates for low prior probabilities. However, as the previous example illustrates, they can be very poor surrogates. It would be more consistent to ascertain prior probabilities that can be used uniformly across studies. Future research should investigate ascertaining prior probabilities in this domain, and reporting results using posterior probabilities rather than significance with a correction.

References

Bateson, W.: Mendel's Principles of Heredity. Cambridge University Press, New York (1909)

Brooks, A.J.: The Essence of SNPs. Gene. 234, 177–186 (1999)

Chen, S.S., et al.: Atomic Decomposition by Basis Pursuit. SIAM Journal on Scientific Computing 20, 33–61 (1998)

Chickering, M.: Learning Bayesian Networks is NP-Complete. In: Fisher, D., Lenz, H. (eds.) Learning from Data. Lecture Notes in Statistics, Springer, New York (1996)

Chickering, D.: Optimal Structure Identification with Greedy Search. The Journal of Machine Learning Research 3, 507–554 (2003)

Cho, Y.M., Ritchie, M.D., Moore, J.H., Moon, M.K., et al.: Multifactor Dimensionality Reduction Reveals a Two-Locus Interaction Associated with Type 2 Diabetes Mellitus. Diabetologia 47, 549–554 (2004)

Coffey, C.S., 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 5(49) (2004)

Coon, K.D., 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 68, 613–618 (2007)

Cooper, G.F., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning 9, 309–347 (1992)

Corder, E.H., et al.: Gene Dose of Apolipoprotein E type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. Science 261, 921–923 (1993)

Epstein, M.J., Haake, P.: Very Large Scale ReliefF for Genome-Wide Association Analysis. In: Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2008)

Evans, D.M., Marchini, J., Morris, A., Cardon, L.R.: Two-Stage Two-Locus Models in Genome-Wide Association. PLOS Genetics 2(9) (2006)

Friedman, N., Yakhini, Z.: On the Sample Complexity of Learning Bayesian Networks. In: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence, pp. 206–215 (1996)

Galvin, A., Ioannidis, J.P.A., Dragani, T.A.: Beyond Genome-Wide Association Studies: Genetic Heterogeneity and Individual Predisposition to Cancer. Trends in Genetics (3), 132–141 (2010)

Hahn, L.W., Ritchie, M.D., Moore, J.H.: Multifactor Dimensionality Reduction Software for Detecting Gene-Gene and Gene-Environment Interactions. Bioinformatics 19(3), 376–382 (2003)

Han, B., Park, M., Chen, X.: Markov Blanket-Based Method for Detecting Causal SNPs in GWAS. In: Proceeding of IEEE International Conference on Bioinformatics and Biomedicine (2009)

Heckerman, D.: A Tutorial on Learning with Bayesian Networks, Technical Report # MSR-TR-95-06. Microsoft Research, Redmond, WA (1996)

Heckerman, D., Geiger, D., Chickering, D.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, Technical Report MSR-TR-94-09. Microsoft Research, Redmond, Washington (1995)

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinksy, C.T.: Bayesian Model Averaging: A Tutorial. Statistical Science 14, 382–417 (1999)

Hunter, D.J., Kraft, P., Jacobs, K.B., et al.: A Genome-Wide Association Study Identifies Alleles in FGFR2 Associated With Risk of Sporadic Postmenopausal Breast Cancer. Nature Genetics 39, 870–874 (2007)

Jiang, X., Barmada, M.M., Visweswaran, S.: Identifying Genetic Interactions From Genome-Wide Data Using Bayesian Networks. Genetic Epidemiology 34(6), 575–581 (2010a)

Jiang, X., Neapolitan, R.E., Barmada, M.M., Visweswaran, S., Cooper, G.F. : A Fast Algorithm for Learning Epistatic Genomic Relationships. In: Accepted as Proceedings Eligible by AMIA 2010 (2010b)

Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)

Korb, K., Nicholson, A.E.: Bayesian Artificial Intelligence. Chapman & Hall/CRC, Boca Raton, FL (2003)

Lam, W., Bacchus, F.: Learning Bayesian Belief Networks: An approach based on the MDL Principle. In: Proceedings of 2nd Pacific Rim International Conference on Artificial Intelligence, pp. 1237–1243 (1992)

Logsdon, B.A., Hoffman, G.E., Mezey, J.G.: A Variational Bayes Algorithm for Fast and Accurate Multiple Locus Genome-Wide Association Analysis. BMC Bioinformatics 11(58) (2010)

Manolio, T.A., Collins, F.S.: The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy. Annual Review of Medicine 60, 443–456 (2009)

Matsuzaki, H., Dong, S., Loi, H., et al.: Genotyping over 100,000 SNPs On a Pair of Oligonucleotide Arrays. Nat. Methods 1, 109–111 (2004)

Meng, Y., et al.: Two-Stage Approach for Identifying Single-Nucleotide Polymorphisms Associated With Rheumatoid Arthritis Using Random Forests and Bayesian Networks. BMC Proc. 2007 1(suppl. 1), S56 (2007)

Moore, J.H., White, B.C.: Tuning reliefF for genome-wide genetic analysis. In: Marchiori, E., Moore, J.H., Rajapakse, J.C. (eds.) EvoBIO 2007. LNCS, vol. 4447, pp. 166–175. Springer, Heidelberg (2007)

Neapolitan, R.E.: Learning Bayesian Networks. Prentice Hall, Upper Saddle River (2004)
Neapolitan, R.E.: A Polemic for Bayesian Statistics. In: Holmes, D., Jain, L. (eds.) Innovations in Bayesian Networks. Springer, Heidelberg (2008)
Neapolitan, R.E.: Probabilistic Methods for Bioinformatics: with an Introduction to Bayesian Networks. Morgan Kaufmann, Burlington (2009)
Pappassotiropoulos, A., Fountoulakis, M., Dunckley, T., Stephan, D.A., Reiman, E.M.: Genetic Transcriptomics and Proteomics of Alzheimer's Disease. J. Clin. Psychiatry 67, 652–670 (2006)
Reiman, E.M., et al.: GAB2 Alleles Modify Alzheimer's Risk in APOE $\epsilon 4$ Carriers. Neuron 54, 713–720 (2007)
Ritchie, M.D., et al.: Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. Am. J. Hum. Genet. 69(1), 138–147 (2001)
Rissanen, J.: Modelling by Shortest Data Description. Automatica 14, 465–471 (1978)
Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. Springer, New York (1993); 2nd edn. MIT Press (2000)
Suzuki, J.: Learning Bayesian Belief Networks based on the Minimum Description length Principle: Basic Properties. IEICE Trans. on Fundamentals E82-A(9), 2237–2245 (1999)
Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso. J. Royal. Statist. Soc. B 58(1), 267–288 (1996)
Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H.: A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Dataset using Multifactor Dimensionality Reduction. Genetic Epidemiology 31, 306–315 (2007)
Verzilli, C.J., Stallard, N., Whittaker, J.C.: Bayesian Graphical Models for Genomewide Association Studies. The American Journal of Human Genetics 79, 100–112 (2006)
Wade, N.: A Decade Later, Genetic Map Yields Few New Cures. New York Times (June 12, 2010)
Wan, X., et al.: Predictive Rule Inference for Epistatic Interaction Detection in Genome-Wide Association Studies. Bioinformatics 26(1), 30–37 (2010)
Wang, D.G., Fan, J.B., Siao, C.J., et al.: Large-Scale Identification, Mapping, and Genotyping of Single Nucleotide Polymorphisms in the Human Genome. Science 80, 1077–1082 (1998)
Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K.: Genome-Wide Association Analysis by Lasso Penalized Logistic Regression. Genome Analysis 25, 714–721 (2009)
Wu, J., Devlin, B., Ringguist, S., Trucco, M., Roeder, K.: Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies. Genetic Epidemiology 34, 275–285 (2010)
Zabell, S.L.: W.E. Johnson's 'Sufficientness' Postulate. The Annals of Statistics 10(4) (1982)
Zhang, X., Pan, F., Xie, Y., Zou, F., Wang, W.: COE: A general approach for efficient genome-wide two-locus epistasis test in disease association study. In: Batzoglou, S. (ed.) RECOMB 2009. LNCS, vol. 5541, pp. 253–269. Springer, Heidelberg (2009)
Zhang, Y., Liu, J.S.: Bayesian Inference of Epistatic Interactions in Case Control Studies. Nature Genetics 39, 1167–1173 (2007)