BIOINFORMATICS

Vol. 26 ECCB 2010, pages i632–i637
doi:10.1093/bioinformatics/btq392

Modeling associations between genetic markers using Bayesian networks

Edwin Villanueva and Carlos Dias Maciel*

Electrical Engineering Department, Sao Carlos School of Engineering, University of Sao Paulo, Sao Carlos, Sao Paulo, Brazil

ABSTRACT

Motivation: Understanding the patterns of association between polymorphisms at different loci in a population (linkage disequilibrium, LD) is of fundamental importance in various genetic studies. Many coefficients were proposed for measuring the degree of LD, but they provide only a static view of the current LD structure. Generative models (GMs) were proposed to go beyond these measures, giving not only a description of the actual LD structure but also a tool to help understanding the process that generated such structure. GMs based in coalescent theory have been the most appealing because they link LD to evolutionary factors. Nevertheless, the inference and parameter estimation of such models is still computationally challenging.

Results: We present a more practical method to build GM that describe LD. The method is based on learning weighted Bayesian network structures from haplotype data, extracting equivalence structure classes and using them to model LD. The results obtained in public data from the HapMap database showed that the method is a promising tool for modeling LD. The associations represented by the learned models are correlated with the traditional measure of LD D'. The method was able to represent LD blocks found by standard tools. The granularity of the association blocks and the readability of the models can be controlled in the method. The results suggest that the causality information gained by our method can be useful to tell about the conservability of the genetic markers and to guide the selection of subset of representative markers.

Availability: The implementation of the method is available upon request by email.

Contact: maciel@sc.usp.br

1 INTRODUCTION

The detection of linkage disequilibrium (LD), the non-random association of alleles at different loci in a population, and the assessing of its intensity, extent and distribution is a fundamental step in many genetic studies. In association studies, for example, the search for the locus (or loci) responsible of a particular trait or disease is narrowed to regions of high LD (LD blocks) where genotyped genetic markers were observed to be associated with the studied phenotype (Mueller, 2004). In population genetic, LD patterns has been widely used to study the evolutionary and demographic processes in a variety of animal and plant populations, such as admixtures, migration and natural selection (Tishkoff *et al.*, 1996; Zhang *et al.*, 2004). LD information was also useful to learn more about the architecture of the human genome and its biology of recombination (Pritchard and Przeworski, 2001).

A variety of coefficients have been proposed to quantify the intensity of LD (see Mueller, 2004 for a review). Pairwise LD measures were the first ones reported for this purpose, which measure the overall allele association between two loci. Popular examples of such measures are D' and r^2 (Hedrick, 1987). Subsequently, multi-locus LD coefficients were proposed to measure simultaneous allele associations among multiple loci. Classical examples are the I_A index and the coefficient H (Mueller, 2004; Sabatti and Risch, 2002). Recently, information theory was used to develop new LD coefficients. Some examples are: the coefficent ε, based in entropy (Nothnagel *et al.*, 2002); the normalized mutual information (MIR) coefficient (Zhang *et al.*, 2009); and the normalized relative entropy (ER) coefficient (Liu and Lin, 2005). This active search of LD coefficients has been accompanied with the development of various tools that display LD measures in a comprehensive way. Examples of those tools are: GOLDsurfer (Pettersson *et al.*, 2004), Haploview (Barrett *et al.*, 2005) and LdCompare (Hao *et al.*, 2007). This remarkable interest in developing new measures and tools for studying LD can be explained by the dramatic increase of public genotype data [from the HapMap project (The International HapMap Consortium, 2003), for example], which at the same time enabled association studies in a whole-genome scale that presented new statistical and computational challenges due to the vast quantity of data collected.

Although some of the aforementioned LD measures and tools were useful in characterizing LD at various genomic regions of several populations, they are limited to provide a static view of the LD structure in the studied region. In an attempt to go beyond these measures some generative models (GMs) were proposed (Hudson, 2001; Li and Stephens, 2003; Maniatis *et al.*, 2002; McVean *et al.*, 2002). GMs are useful because they model the process that generate the observed data, providing the machinery to do inferences and simulations of yet unobserved situations or to help understanding the underlying generative process. Models based in coalescent theory (Kingman, 2000) have been the most appealing GMs because they relate LD to evolutionary factors, such as recombination rate, migration and mutation. However, the parameter estimation and inference in such models is still computationally challenging and only applicable to short regions (Nicolas *et al.*, 2006). A more practical alternative to model LD is to learn probabilistic graphical models (PGMs) directly from the observed genotype or haplotype data, as proposed by Thomas and Camp (2004). PGM are a type of GMs that have proven to be useful in modeling a variety of complex real-world problems mainly due to the following reasons: intuitive interpretability of the models, their ability to encode the joint probability distribution with a reduced number of parameters (Heckerman *et al.*, 1995) and the existence of efficient methods to learn PGM from data and to make inferences. Because the learning

*To whom correspondence should be addressed.

of PGM from data is an empirical approach, the learned models are not intended to describe LD in terms of evolutionary factors, but rather to represent the current LD structure in an accurate, compact and understandable way, which is important in genetic association studies. In the present article, we propose the use of Bayesian networks (BN) to model LD. BN (Pearl, 1988) are a type of PGM that encode a joint probability distribution via a directed acyclic graph (DAG), where nodes represent random variables and edges represent dependencies between them. We choose BN because we were interested in learning the causality of the associations, which we found to give additional information about the LD structure. Our approach is different from the previous works of Thomas and Camp (2004) and Thomas (2005), which propose the use of Markov networks to represent dependences between proximal loci and several runs of simulated annealing algorithm to learn optimal models. BN models are more naturally interpretable than Markov networks, since the components (factors) of the factorized joint probability distribution are associated to the model nodes instead of cliques. The learning of the BN structures is based on the K2GA algorithm (Larranaga *et al.*, 1996), which finds the optimal BN structure(s) through the combination of a global search [using a genetic algorithm] GA on the space of topological orderings and a local search (using the K2 greedy search method) on the subspace delimited by each ordering. The learning method computes the strength of each edge. The learned BN structures are grouped into equivalence structure classes (set of BN with the same set of dependence/independence relationships), which are represented by a partially DAG (PDAG). The identification of LD associations and LD blocks are performed in these models. The method was tested in public haplotype data from the HapMap database in three different segments located in ENCODE regions. The results showed the plausibility of the method in modeling LD, representing associations and LD blocks consistent with standard tools. The effects of pruning weak edges on the formation of association blocks are studied. The correlation of the traditional measure D' with the associations represented by the PDAGs is also studied.

2 APPROACH

2.1 Data

We consider that we have in hand haplotype data obtained from a given population. This data are formed by m haplotypes, each one consisting of N marker loci sorted by their physical location in the chromosome. Let x_{ij} denote the allele state of the j-th marker in the haplotype i, which can be one of the allele set $(1,2,...,r_j)$, where r_j is the number of alleles of marker j. The haplotypes may be experimentally determined or inferred from genotype data by a phasing algorithm [e.g. fastPHASE (Scheet and Stephens, 2006); BEAGLE (Browning and Browning, 2007)]. In the context of BN, the j-th marker is modeled by a random variable X_j. Thus, an haplotype $\boldsymbol{x}_i=(x_{i1},...,x_{iN})$ is regarded as a realization of a N-dimensional random variable $\boldsymbol{X}=(X_1,...,X_N)$.

2.2 Bayesian networks

Bayesian networks (Neapolitan, 2003; Pearl, 1988) are a type of PGM that can represent the conditional dependencies and independencies between a set of random variables $\boldsymbol{X}=(X_1,...,X_N)$ via a DAG. A BN is composed by a *model structure* S and a set of *model parameters* $\boldsymbol{\theta}$. The *model structure* is the qualitative part of a BN and is formed by a DAG $S=(V,E)$, where V is the set of nodes representing the random variables $\boldsymbol{X}$ and E is the set of edges representing the conditional dependencies between the random variables of $\boldsymbol{X}$. The *model parameters* are the quantitative part of a BN that in conjunction with the DAG define completely its joint probability distribution. The model parameters are formed by N parameter vectors $\boldsymbol{\theta}=(\theta_1,...,\theta_N)$ defining the conditional probability distributions of each variable, i.e. θ_i defines the conditional probability distribution of X_i, $P(x_i|pa_i,\theta_i)$, where x_i represents the instantiation of the variable X_i and pa_i represents the instantiation of the parent variables of X_i, Pa_i. All BN have the property that each variable is conditionally independent on its non-descendants given its parent variables (Markov property). For example, the BN $\{a\rightarrow b\rightarrow c\}$ and $\{a\leftarrow b\rightarrow c\}$ represent the same independences (a and c are independent given b) and, therefore, are indistinguishable. The BN $\{a\rightarrow b\leftarrow c\}$ differs from the previous because it represent diferent independencies (a and c are marginally independent and all other pairs are dependent). This property allows the computation of the joint probability of any instantiation of $\boldsymbol{X}$, $\boldsymbol{x}$, as the product of the local conditional probabilities: $P(\boldsymbol{x})=\prod_{i=1}^{N}P(x_i|pa_i,\theta_i)$.

2.3 Structural learning of BN

The induction of BN structures from data is a NP-hard problem (Chickering *et al.*, 2004). We need a heuristic method to find the structure(s) that best reflect the dependency/independence relations contained in the data. To this end, we take the idea of the K2GA algorithm (Larranaga *et al.*, 1996) to learn model structures from haplotype data, as described below.

The main idea of K2GA is to find optimal BN structures by searching on the space of *topological orderings* using a combination of global search (using a GA) and a local search [using the greedy search K2 heuristic (Cooper and Herskovits, 1992)]. A *topological ordering* (TO) is an ordering of the system variables $\boldsymbol{X}$ such that $\forall i,j$, if X_j comes before X_i in the ordering, namely $(...,X_j,...,X_i,...)$, then X_i can only have node X_j as a parent node. A TO, therefore, delimits a subspace of structures. A GA is used to evolve a population of TOs. Each TO in the evolving population is evaluated by the K2 heuristic. Such evaluation consists in finding the best structure within the subspace of structures spanned by the TO. The data provided to K2 are: the TO, a dataset D containing the observations of the variables $\boldsymbol{X}$, and an upper bound u on the number of parents a node may have. For each node X_i, $(i=1,...,N)$, K2 iteratively attempts to add up to u edges from $Anc(X_i)$, the ancestors of X_i according to the TO. In each iteration one node is selected from $Anc(X_i)$ and added to the set of parent nodes of X_i, Pa_i. The selected parent is one whose addition most increases the structure score. K2 stops to adding parents to the node X_i when u parents were added or when no more ancestors in $Anc(X_i)$ can increase the structure score. The posterior probability of the structure given the data (Heckerman and Chickering, 1995), known as BDe metric, is used as the structure score, which is computed as follows:

$$P(S|D)=\prod_{i=1}^{N}g(i,Pa_i), \qquad (1)$$

$$g(i,Pa_i)=\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$$

where r_i is the number of states of X_i, q_i is the number of different instantiations that the parents Pa_i can take; N_{ijk} is the number of cases in D where X_i takes its k-th state and its parents Pa_i take their j-th state; N_{ij} is the sum of N_{ijk} over k; α_{ijk} is a prior on N_{ijk}, calculated as α/r_iq_i when a uniform prior is considered with equivalent sample size α; and α_{ij} is the sum over k of α_{ijk}. When the logarithm of the structure score [Equation (1)] is applied, this reduces to a sum of *node scores*, $\log(g(i,Pa_i))$, which only depends on the node X_i and its parents Pa_i. This modularization of the structure score is advantageous to K2, which selects the set of parents for each node X_i based in the maximization of the node score instead of the maximization of the whole structure score.

The strength of any edge is measured as the increase on the structure score that the edge provokes when it is added to the structure. For example, if a new parent node X_j is added to the set of parents of X_i, i.e. an edge $X_j \rightarrow X_i$, the edge strength is computed as $\log(g(i,Pa_i \cup X_j)) - \log(g(i,Pa_i))$. This strength is equivalent to the logarithm of the Bayes factor.

The GA used to evolve TOs starts with an initial population of λ TOs (called individuals in the context of the GA) generated randomly. Then, each individual is evaluated with K2, which returns the score of the best structure as the fitness of the individual. The following operators are then sequentially performed to create the next generation (the evaluated population is referred as the current generation). (i) *Selection*: ρ pairs of individuals are selected from the current generation to form a mating pool. The selection is based on the individual's fitness using the roulette wheel algorithm. (ii) *Crossover*: produces two new individuals (offspring) for each pair of the mating pool. The order-based crossover operator OX2 (Larranaga *et al.*, 1996) is used for this purpose. This operator is only applied when a random number generated uniformly in the interval $[0,1]$ is less than μ, the *crossover rate* parameter, otherwise the resulting offspring is replicated from the mating individuals. (iii) *Mutation*: this is applied to each offspring individual resulting of the previous step. The displacement mutation (DM) operator (Larranaga *et al.*, 1996) is used for this purpose. Like crossover, this operator is only applied when a random number generated uniformly in the interval $[0,1]$ is less than ν, the *mutation rate* parameter, otherwise the offspring individual is left unchanged. (iv) *Scoring*: evaluates the resulting offspring individuals after mutation with the K2 heuristic. (v) *Replacement*: creates the next generation by replacing the worst α individuals of the current generation with the best α individuals of the offspring population, provided that the replacing individuals are better than the replaced individuals. The new generation is set as the current generation and a new loop is started. The algorithm terminates when a predefined number of generations is reached.

2.4 Obtaining BN structure equivalence classes for modeling LD

As pointed by Chickering (2002), it is more appropriate to learn equivalence classes of network structures than single structures. An equivalence class represents a group of structures that encode the same set of dependence/independence relationships.

We took advantage on the ability of the above learning method in generating several optimal BN structures (structures with the same maximum fitness found along the evolutionary process) to get equivalence classes. All the optimal BN structures are analyzed and grouped according to their topological connection, i.e. structures in one group only differ in edge directions. Each group is represented by a PDAG constructed by superimposing all the DAGs of the group, as shown in Figure 1. The resulting PDAGs can serve to characterize LD. Two genetic markers are associated if they are not marginally independent, which imply that the corresponding nodes are connected in a PDAG by a directed path or by a common ancestor node. In the example of Figure 1 all the pair of nodes with the same color are associated. Additionally, the pairs (3, 4) and (3, 5) are also associated. We define an *association block* in a PDAG as the set of the largest number of consecutive markers (consecutive in terms of its physical location) that are associated with each other. In Figure 1, two association blocks can be found, identified with different colors.

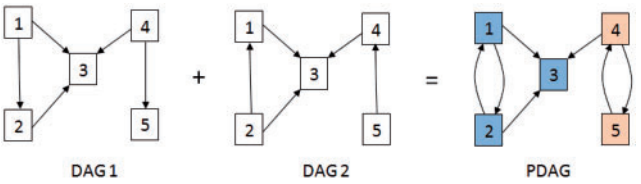

Fig. 1. Example illustrating the construction of a PDAG from two optimal DAGs. The nodes with the same color are associated with each other forming an association block.

3 RESULTS

The described method was implemented in the C++ programming language and tested in a 64-bit 2-core (2.1 GHz) computer with 4 GB of RAM, running a Linux operating system. The graph drawing was performed using the Graph Visualization Software (www.graphviz.org).

To test the ability of the proposed approach in characterizing LD, we used public data registered in the HapMap database (The International HapMap Consortium, 2003). We chose three genomic segment located in ENCODE[1] regions of different chromosomes and different populations, as shown in Table 1. ENCODE regions were chosen due to their high density of genotyped SNPs markers and to the availability of existing studies in such regions (The International HapMap Consortium, 2005). The segments were selected increasing their size and complexity of the LD patterns (Figs 4, 5 and 6), aiming to test the method in different situations. The datasets were obtained from the HapMap repository, version III release 2, via the Genome Browser web application (http://hapmap.ncbi.nlm.nih.gov) using the option 'Download Phased Haplotype Data'. The SNP markers with minor allele frequency (MAF) < 0.03 were discarded, since they have little polymorphism.

The following set of parameters were introduced to the learning method in all tests: population size $\lambda = 20$, couples in the mating pool $\rho = 10$, crossover rate $\mu = 0.95$, mutation rate $\nu = 0.05$, number of replacement individuals in each generation $\alpha = 10$, number of generations = 300, maximum number of parents a node may have $u = \#\text{SNPs} - 1$. For clarity and space, we first detail the results for the ENm010_CHB+JPT dataset and then present summarized results for the other datasets.

Figure 2 shows the evolution of the best fitness and average fitness of the BN structures learned from the ENm010_CHB+JPT

[1]Encyclopedia of DNA Elements (ENCODE Project Consortium, 2004)

Table 1. Datasets used to test the method

Dataset name	ENCODE region	Chromosome band	Genomic segment (kp)	HapMap population	Number of SNPs (MAF > 0.03)	Number of haplotypes
ENm010_CHB+JPT	ENm010	7p15.2	27 070–27 126	CHB+JPT	15	340
ENr131_CEU	ENr131	2q37.1	235 065–235 122	CEU	31	234
ENr321_YRI	ENr321	8q24.11	118 797–118 895	YRI	47	230

These datasets were obtained from the HapMap repository (version III, release 2) using the Genome Browser web application (http://hapmap.ncbi.nlm.nih.gov). Only SNPs with MAF > 0.03 were considered.

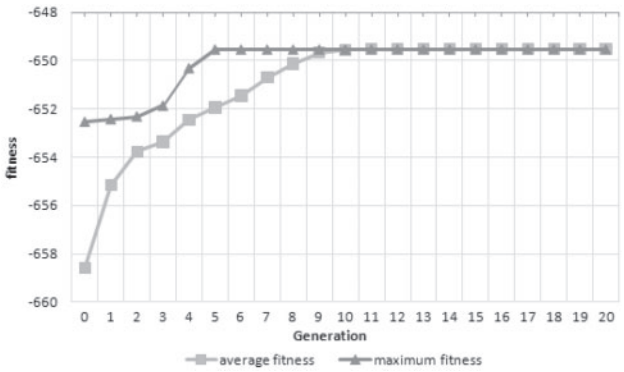

Fig. 2. Evolution of the maximum and average fitness of BN structures learned from the ENm010_CHB+JPT dataset.

dataset. Only the first 20 generations are shown, since the curves remain unchanged in subsequent generations. As can be observed, the first optimal BN structure is found at the fifth generation. Six generations later the entire population converged to only optimal BN structures (the average fitness stabilizes at its maximum value). This can be considered a fast convergence, since it is reached in approximately the first 4% of the total number of generations. We use a predefined high number of generations as a stopping criterion because we were interested in exploring the topological diversity of optimal BN structures to get equivalence classes (PDAGs).

All the optimal learned DAGs (a total of 5800 DAGs from generations 11 to 300) were analyzed. Twelve different DAGs were identified in that set, which were grouped into four different PDAGs, each having 31 edges. One of these PDAGs is shown in Figure 3a. The other PDAGs are similar to this figure, differing only in the location of three edges. For each of these PDAGs was determined the set of pairwise marker associations. No difference was found between these four sets of pairwise associations. As can be observed, two association blocks were identified in the PDAG, which are consistent with the two LD blocks (Block1 and Block 2) found in the triangular LD plot (Fig. 4) by the Haploview tool using the *Strong LD Spine* block definition (Barrett *et al.*, 2005).

With the aim to knowing how the associations and association blocks are affected with the elimination of weak edges, we perform two experiments in the PDAG of Figure 3a. In the first experiment, the edges with strengths lower than the first quartile (the lower quartile) were eliminated. The resulting PDAG is shown in Figure 3b with 23 edges. There was no alteration either in the set of pairwise associations or in the association blocks (these blocks are also indicated with line segments at the top of the LD plot of Figure 4).

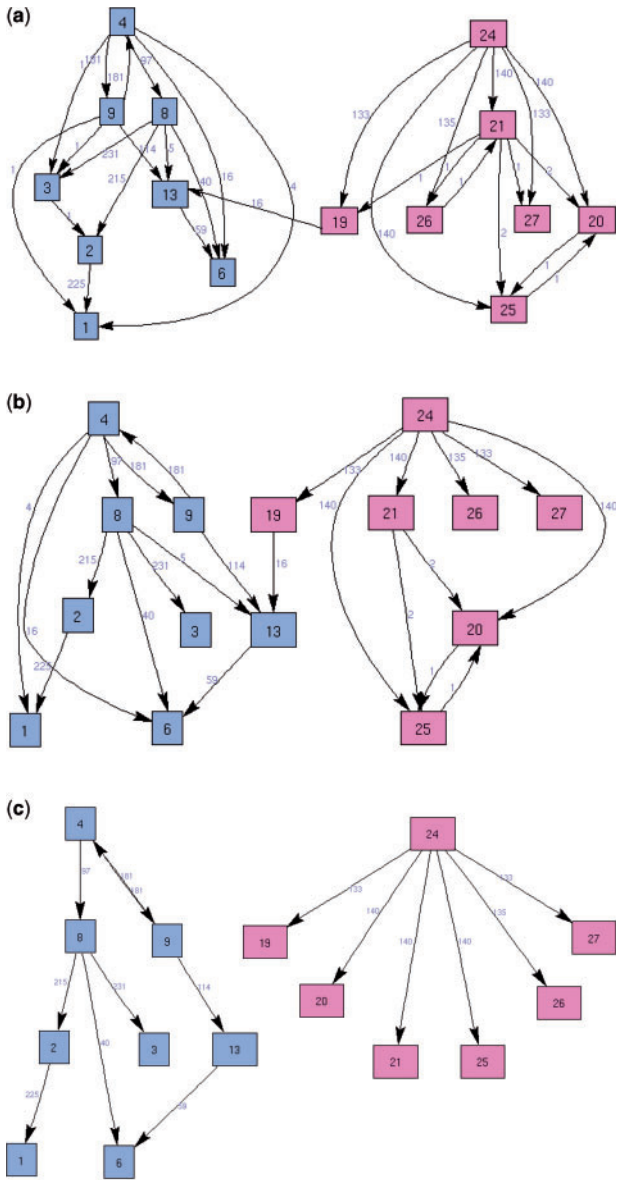

Fig. 3. (**a**) A PDAG (original) learned from ENm010_CHB+JPT dataset. PDAGs resulting of pruning edges at the first quartile (**b**) and at the second quartile (**c**) of (a). The node numbers are the codes assigned by the Haploview tool in the segment. The edge labels are the edge strengths.

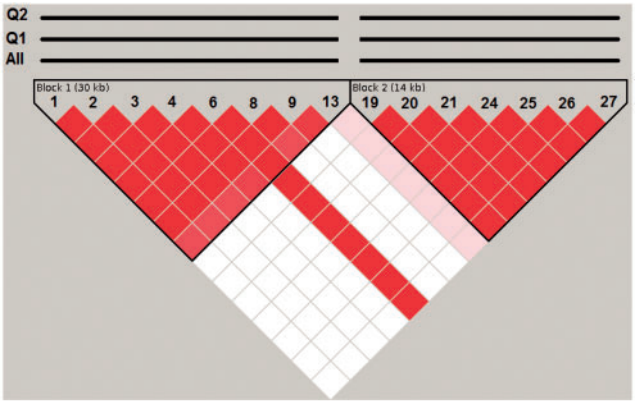

Fig. 4. LD plot for the genomic segment of the ENm010_CHB+JPT dataset. The segmeted lines All, Q1 and Q2 at the top represent the association blocks found in the original PDAG, the PDAG with the first quartile of edges removed and the PDAG with the second quartile of edges removed, respectively.

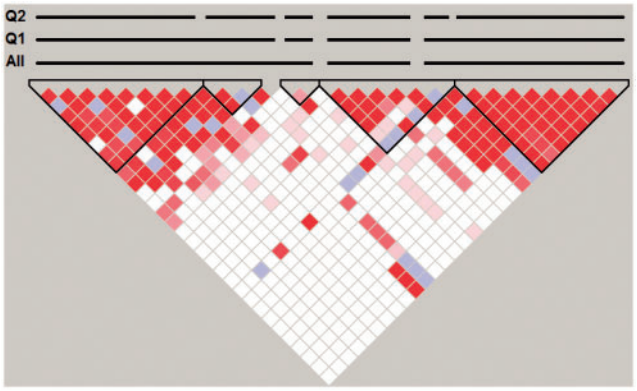

Fig. 5. LD plot for the genomic segment of the ENr131_CEU dataset.

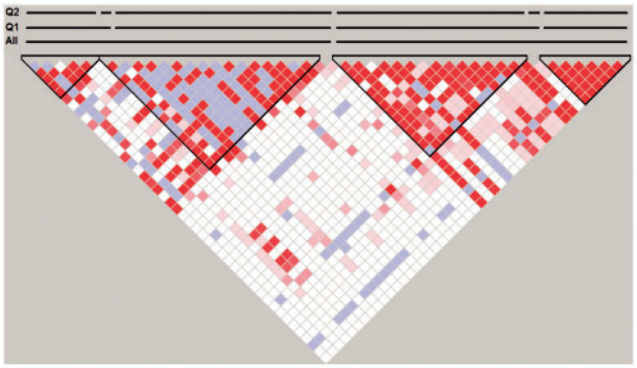

Fig. 6. LD plot for the genomic segment of the ENr321_YRI dataset.

Table 2. Structural learning results for the three analyzed datasets

Dataset	Convergence generation	# PDAGs	Processing time (seg)
ENm010_CHB+JPT	11	4	7
ENr131_CEU	43	4	143
ENr321_YRI	166	3	735

In the second experiment, the edges with strengths lower than the second quartile (the worst half of edges) were eliminated. The resulting PDAG is shown in Figure 3c. In this case the number of pairwise associations fell from 63 to 49, but the association blocks remained unchanged. This preservation of the association blocks with the removal of significant number of edges can be due to the high compaction of the LD blocks.

A summary of results of executing the method in all datasets can be found in Table 2. It can be observed that the convergence generation increases with the number of markers, as well as the processing time. This is an expected result, since the search space is bigger and more complex. Like the results in the ENm010_CHB+JPT dataset, the four PDAGs obtained from the ENr131_CEU dataset and the three PDAGs obtained from the ENr321_YRI dataset represent the same set of pairwise associations and the same association blocks. These blocks are illustrated as segmented lines with the name 'All' at the top of Figures 5 and 6, respectively. As can be noticed, the association blocks represented in the PDAGs considering all edges tend to be generous, grouping large quantity of nodes into few blocks, including in such blocks markers with moderate to low D' values. When the lower quartile of edges is removed, the association blocks are splitted into more compact blocks (line segments Q1). When the second lower quartile of edges is removed the association blocks (line segments Q2) tend to be more segmented and similar to that found by the *Strong LD Spine* block definition.

It is interesting to know the relationship between the learned associations represented by the PDAGs and the classical measure of LD D'. For this end, Figures 7a, b and c shows the distributions of the associations with respect to 10 equally spaced intervals of D' for the three datasets, respectively. It is possible to observe a clear trend in the three datasets to represent associations with high D' values (in the interval [0.9–1.0]). When edges are pruned at first quartile, the distributions of the represented associations are almost unchanged. When the second quartile of edges are removed the association distributions are moderately altered, predominantly in the lower intervals of D'. This means that weak edges are important components of weak D' associations, and *vice versa*.

The learned PDAGs showed another interesting information: the markers most central in the blocks tend to have more outgoing edges (e.g. markers 4 and 24 in all PDAGs of Fig. 3) and the markers on the block boundaries tend to have predominantly incoming edges (markers 1, 13, 19 and 27 in all PDAGs of Fig. 3). An explanation for this behavior is that genetic markers near the center of the blocks are highly conserved in the population (had minimal recombination), which is reflected in the learned models by their tendency to be more causative than dependent nodes. This suggests that the causal information gained by BN models can be useful to tell about the conservability of the genetic markers in the analyzed population. The learned causal information can also be useful to guide the selection of an informative subset of 'tags' markers, since nodes that have predominantly outgoing edges can explain the allele status of their dependent markers, being good candidates as 'tag' markers.

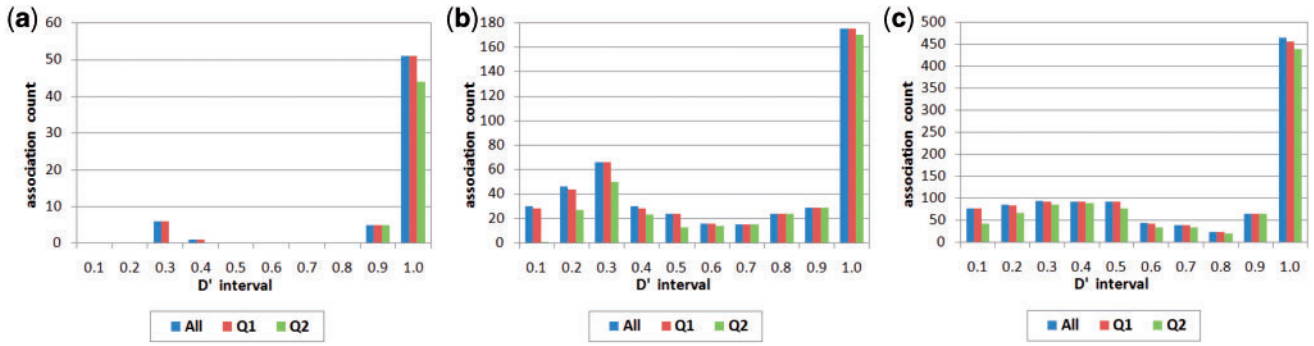

Fig. 7. Distributions of the PDAG associations with respect to 10 equally spaced intervals of D' for the three analyzed datasets: (**a**) ENm010_CHB+JPT; (**b**) ENr131_CEU; (**c**) ENr321_YRI.

4 CONCLUSIONS

In this article a novel application of Bayesian networks to model associations between genetic markers was proposed. The method is based on learning optimal BN structures (weighted) from haplotype data, extracting equivalence structure classes (PDAGs) and using them to model LD. The results obtained in public data from the HapMap database showed that our approach is a promising tool for modeling LD. All the association blocks represented in the learned PDAGs were consistent with LD blocks found by standard tools. It was shown that by pruning weak edges is possible to control the granularity of the association blocks and the clarity and interpretability of the models. The associations represented by the PDAGs were shown in correlation with the traditional measure of LD D'. It was suggested that the causality information learned by our approach can be useful to infer about the conservability of the genetic markers and to guide the selection of informative 'tags' markers. Our current developments are in improving the efficiency of the method and parallel implementations for extending greatly the number of markers that can be considered up to whole-genome scales.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nıvel Superior (CAPES), the Brazilian government agency for the development of human resources.

Conflict of Interest: none declared.

REFERENCES

Barrett,J. *et al*. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Chickering,D. (2002) Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.*, **2**, 445–498.

Chickering,D. *et al*. (2004) Large-sample learning of Bayesian networks is NP-Hard. *J. Mach. Learn. Res.*, **5**, 1287–1330.

Cooper,G.F. and Herskovits,E.A. (1992) A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, **9**, 309–347.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia of DNA elements) Project. *Science*, **306**, 636–640.

Hao,K. *et al*. (2007) LdCompare: rapid computation of single- and multiple-marker r(2) and genetic coverage. *Bioinformatics*, **23**, 252–254.

Heckerman,D. and Chickering,D.M. (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.*, **20**, 197–243.

Heckerman,D. *et al*. (1995) Real-world applications of Bayesian networks. *Commun. ACM*, **38**, 24–26.

Hedrick,P. (1987) Gametic disequilibrium measures - proceed with caution. *Genetics*, **117**, 331–341.

Hudson,R. (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.

Kingman,J. (2000) Origins of the coalescent: 1974-1982. *Genetics*, **156**, 1461–1463.

Larranaga,P. *et al*. (1996) Learning Bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE Trans. Syst. Man Cybernet. A Syst. Hum.*, **26**, 487–493.

Li,N. and Stephens,M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233.

Liu,Z. and Lin,S. (2005) Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet. Epidemiol.*, **29**, 353–364.

Maniatis,N. *et al*. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.

McVean,G. *et al*. (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Mueller,J. (2004) Linkage disequilibrium for different scales and applications. *Brief. Bioinform.*, **5**, 355–364.

Neapolitan,R.E. (2003) *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ.

Nicolas,P. *et al*. (2006) A model-based approach to selection of tag SNPs. *BMC Bioinformatics*, **7**, Article 303.

Nothnagel,M. *et al*. (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum. Hered.*, **54**, 186–198.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Pettersson,F. *et al*. (2004) GOLDsurfer: three dimensional display of linkage disequilibrium. *Bioinformatics*, **20**, 3241–3243.

Pritchard,J. and Przeworski,M. (2001) Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.*, **69**, 1–14.

Sabatti,C. and Risch,N. (2002) Homozygosity and linkage disequilibrium. *Genetics*, **160**, 1707–1719.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Thomas,A. (2005) Characterizing allelic associations from unphased diploid data by graphical modeling. *Genet. Epidemiol.*, **29**, 23–35.

Thomas,A. and Camp,N. (2004) Graphical modeling of the joint distribution of alleles at associated loci. *Am. J. Hum. Genet.*, **74**, 1088–1101.

Tishkoff,S. *et al*. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, **271**, 1380–1387.

Zhang,W. *et al*. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl Acad. Sci. USA*, **101**, 18075–18080.

Zhang,L. *et al*. (2009) A multilocus linkage disequilibrium measure based on mutual information theory and its applications. *Genetica*, **137**, 355–364.