Theor Appl Genet (2013) 126:1991–2002
DOI 10.1007/s00122-013-2112-y

ORIGINAL PAPER

Evaluation of linkage disequilibrium in wheat with an L1-regularized sparse Markov network

Gota Morota · Daniel Gianola

Received: 8 February 2013 / Accepted: 26 April 2013 / Published online: 10 May 2013

Abstract Linkage disequilibrium (LD) is defined as a stochastic dependence between alleles at two or more loci. Although understanding LD is important in the study of the genetics of many species, little attention has been paid on how a covariance structure between many loci distributed across the genome should be represented. Given that biological systems at the cellular level often involve gene networks, it is appealing to evaluate LD from a network perspective, i.e., as a set of associated loci involved in a complex system. We applied a Markov network (MN) to study LD using data on 1,279 markers derived from 599 wheat inbred lines. The MN attempts to account for association between two markers, conditionally on the remaining markers in the network model. In this study, the recovery of the structure of a LD network was done through two variants of pseudo-likelihoods subject to an L1 penalty on the MN parameters. It is shown that, while the L1-regularized Markov network preserves features of a Bayesian network (BN), the nodes in the resulting networks have fewer links. The resulting sparse network, encoding conditional independencies, provides a clearer picture of association than marginal LD metrics, and a sparse graph eases interpretation markedly, since it includes a smaller number of edges than a BN. Thus, an L1-regularized sparse Markov network seems appealing for representing conditional LD with high-dimensional genomic data, where variables, e.g., single nucleotide polymorphism markers, are expected to be sparsely connected.

Communicated by M. Frisch.

G. Morota (✉) · D. Gianola
Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA
e-mail: morota@ansci.wisc.edu

D. Gianola
Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706, USA

D. Gianola
Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA

Introduction

Linkage disequilibrium (LD) is defined as a stochastic dependence between alleles at two or more loci. Characterizing the pattern and extent of LD is important for successful fine-scale mapping of quantitative trait loci (Meuwissen and Goddard 2000) and for genome-enabled prediction in quantitative genetics (Meuwissen et al. 2001). It may also provide insights into the evolutionary history that a population has undergone. A large amount of polymorphic loci has been identified in recent years, thanks to efficient high-throughput genotyping technologies. With this type of genomic data, the resolution of LD structure can be investigated more deeply and with greater precision.

Probabilistic graphical models (Borgelt et al. 2009) provide a framework for assessing conditional dependencies among variables graphically, as well as a means for interpreting the underlying association structure. A Bayesian network (BN) (Neapolitan 2003) is one of the most popular such models used for constructing association networks based on discrete random variables, and it was applied recently in an evaluation of LD among single nucleotide polymorphisms (SNPs) having an effect on milk yield of Holstein cattle (Morota et al. 2012). As a probabilistic graphical model, a BN shares similarities with path analysis and structural equation models (Wright 1921a; Haavelmo 1943). In a BN, each node represents a random

variable, and a node is conditionally independent of its non-descendants given its parents. A BN can greatly simplify the calculation of joint probabilities and, sometimes, it can represent causal relationships involving a target node of interest.

A main challenge comes when high dimensional data needs to be fed to the BN. Due to its inability to produce a sparse graph, a BN often detects very weak genetic signals and produces a complex network that is not amenable to easy graphical interpretation. This limits the use of BN with highly dimensional data, such as genome-wide SNPs derived from high-throughput genotypes, so one is often forced to reduce the number of nodes to be fed, as done in Morota et al. (2012). However, if a main purpose is to detect SNPs that are in moderate or high LD for use in subsequent studies, it would seem reasonable to "kill" weak associations, while highlighting those SNPs in noticeable LD. Given that LD is expected to decline rapidly as the physical distance between two loci increases, and that pairs of loci on different chromosomes rarely show high LD, it seems reasonable to assume sparsity in a network involving a large number of loci.

The edges in a network may be either directed or undirected. A directed network has the potential of conveying causal relationships, with "parental" nodes influencing patterns or occurrence of "children" nodes. In contrast, nodes are connected by edges without arrows in an undirected network. The choice between oriented and unoriented networks is commonly based on whether an underlying network encodes an asymmetric or a symmetric relationship among a set of nodes. In the context of LD, the directionality of an underlying network is unclear since associations among loci are typically assumed bi-directional. Therefore, undirected networks may be sensible in the study of LD.

The r^2 metric, i.e., the squared correlation between alleles at two SNP loci, is the most commonly used measure for quantifying LD in population genetics (Hill and Robertson 1968). An alternative is provided by what is called a "relevance network" (Butte et al. 2000). This type of network offers a simple way of measuring associations and has been used in RNA expression analysis. In the context of LD, the rule applied in a relevance network is that two loci are connected if and only if the absolute value of the pairwise correlation exceeds a pre-defined threshold. However, a disadvantage is that both the relevance network and r^2 represent a marginal dependence structure only, implying that pairwise association values are unconditional on other loci, as explained later. Given that complicated biological systems at the cellular level involve gene networks (Sharan and Ideker 2006), ignoring loci beyond the pair in question may end up capturing superficial genetic associations only. Therefore, it may be worthwhile to use methods that also account for association between two loci, conditionally on the remaining loci in the data. Such methods should be suited for high-dimensional genomic data as well.

This study illustrates an application of a novel method for construction of a LD network in wheat that has the following three properties: (1) provides an efficient way of constructing a LD network from high-dimensional data by introducing sparsity, (2) represents the resulting LD network via nodes that are connected by undirected links, and (3) evaluates associations between loci conditionally on the remaining loci, as opposed to the r^2 metric or the relevance network. This article is structured as follows. Section Methods describes the wheat data, providing an overview of a Markov network (MN), and describes the two methods used for constructing sparse binary Markov networks for detecting conditional LD. Section Results gives the results, and concluding remarks are presented in Discussion.

Methods

Data

A wheat data set collected through the CIMMYT Global Wheat Breeding Program including 599 inbred lines, was used. The trait considered in this analysis was grain yield in the first (out of 4) environment represented in the data set. Each line was genotyped using 1,447 Diversity Array Technology (DArT) binary markers generated by Triticarte Pty. Ltd. Markers with a minor allele frequency (MAF) lower than 0.05 were discarded, and missing genotypes were imputed via random sampling of genotypes using probabilities corresponding to the observed genotype frequencies at a locus. This editing stage resulted in 1,279 markers coded as 0 or 1. More details about the data are in Crossa et al. (2007); de los Campos et al. (2009); Crossa et al. (2010).

Overview of Markov networks

The problem of characterizing associations among genotypes at various loci can be casted as one of estimating a MN structure among loci. A MN, also known as a Markov random field, is an undirected graph $G = (V, E)$, where V and E are sets of nodes and edges, respectively. As opposed to a BN, the MN does not encode causal relationships. A MN is more suited than a BN for stating soft constraints between random variables when a clear directionality cannot be assumed, and it is convenient to express an "affinity", instead of a causal relationship (Koller and Friedman 2009; Bishop 2006). Nodes in the MN represent

discrete random variables $\mathbf{x}^T = (x_1, ..., x_p)$, and edges convey their pairwise relationships.

In a MN, the absence of an edge between two nodes, x_j and x_k, implies conditional independence, given all other nodes (Koller and Friedman 2009). Also, if there is a link between two nodes, this probabilistic connection holds regardless of the presence of other links in the network. The "pairwise" conditional independence property can be represented as

$$p(x_j, x_k|\mathbf{x}_{-j,-k}) = p(x_j|\mathbf{x}_{-j,-k})p(x_j|\mathbf{x}_{-j,-k})$$

where $\mathbf{x}_{-j,-k}$ denotes vector $\mathbf{x}$ with random variables x_j and x_k removed.

A MN does not possess directed links, as opposed to a BN, so factorizing the joint distribution as a product of conditional probability distributions of nodes given by parents does not hold, so special care needs to be taken. This is achieved by decomposing the joint distribution as products of functions of the nodes in "cliques". A clique is a subset of nodes in a graph such that every pair of nodes within a clique is connected by some edge; a maximum clique is defined as the clique having the largest number of nodes (Bishop 2006) and there can be more than one maximum clique. Knowing the maximum clique of a graph is sufficient for factorizing the joint distribution, since any subset of the nodes in the maximum clique are redundant. Let C be a set of cliques in a certain graph structure G, and X_c be the set of variables in clique c. The representation of the joint distribution is:

$$p(X) = \frac{1}{Z}\prod_{C\in G}\phi_c(X_c) \quad (1)$$

where ϕ is called a "clique potential" and Z is a normalizing constant defined by

$$Z = \sum_c \prod_{C\in G}\phi_c(X_c).$$

Clique potentials are positive functions that do not necessarily represent probabilities or conditional probabilities. Hence, to guarantee that the sum of probabilities is equal to 1, we need to explicitly divide by a normalization factor (Z), to ensure that $p(X)$ behaves as a probability distribution. Equation (1) is given by the Hammersley–Clifford theorem (Hammersley and Clifford 1971; Clifford 1990) stating that if $p(X)$ is a positive distribution over $X = x_i, \ldots, x_n$, and if G is a Markov network over X, then $p(X)$ factorizes over cliques in G (Koller and Friedman 2009); this is called a Gibbs distribution. Here, a positive distribution means that for all outcomes $(X = x_i, \ldots, x_n)$ such that $X \neq \emptyset$ (empty set), we have $p(X) > 0$. If joint distribution $p(X)$ is a Gibbs distribution relative to to the MN G, it can be decomposed into a product over cliques in the network structure (Koller and Friedman 2009; Newton 1999).

Pseudo-likelihood based regression with p regularization parameters

The objective here is to reconstruct a network from the covariance structure among markers. The data for each individual consists of a random vector $\mathbf{x}^T = (x_1, ..., x_p)^T$ with p binary markers, such that element $x_j \in (0, 1)$ denotes a binary genotype at locus j $(1 \leq j \leq p)$. We assume that the distribution of random vectors is governed by an unknown MN. For simplicity, we focus on a pairwise MN in this analysis, which is a special case of Markov networks where cliques are over a single node or pairs of nodes (Koller and Friedman 2009). Therefore, in a pairwise MN, we factorize equation (1) such that only a set of node potentials ϕ (x_j) or a set of link potentials ϕ (x_j, x_k) get involved in a model. The joint distribution is given by multiplication of these node and link potential functions.

A pairwise MN for binary variables is also known as an Ising model, which derives from the statistical mechanics literature (Koller and Friedman 2009; Hastie et al. 2009). It models a system of interacting atoms, where each atom is a binary-valued random variable $x_j \in (-1, 1)$; these values describe the direction of the atom spin. In practice, it is common to work in terms of log-linear models, and in this framework, equation (1) is represented as

$$p(X) = \frac{1}{Z}\exp\left[\sum_{q=1}^{k}\theta_q\phi_q(X_q)\right] \quad (2)$$

where $(X_1, ..., X_k)$ are cliques in the MN; $(\phi_1(X_1), ..., \phi_k(X_k))$ are sets of clique potentials, and $(\theta_1, ..., \theta_k)$ are parameters (weights) of the log-linear models, one for each clique. This model allows a compact representation of various distributions, ensuring that all probabilities are positive (Koller and Friedman 2009). In a pairwise MN, we limit the maximum clique size up to 'two' in Eq. (2), so only 'main effects' of each node and all 'pairwise interactions' between nodes are included. Fitting this log-linear model is equivalent to a model selection problem addressing which pairs of nodes (clique size of two) should be included in the model. If the clique of size two corresponding to the jth and kth nodes is included, these nodes will be connected by a link in the network.

Our aim is to estimate the underlying LD network from the p binary markers spanning across the wheat genome, with their joint distribution specified by some unknown MN. In the context of a LD network, each node represents a marker, and lack (presence) of an edge between two markers suggests conditional independence (dependence), given all other markers. We estimate the MN parameters

represented in a $\Theta_{p\times p}$ matrix, by maximizing a log-likelihood. For a given data point (inbred line), the joint distribution of marker vector $\mathbf{x}$ in the log-linear model has the form (Hastie et al. 2009)

$$f(x_1,\ldots,x_p)=\frac{1}{\Psi(\Theta)}\exp\left(\sum_{j=1}^{p}\theta_{j,j}x_j+\sum_{1\le j<k\le p}\theta_{j,k}x_jx_k\right) \tag{3}$$

The first term in the exponential function accounts for the 'main effect' of binary marker x_j (node potential) and the second term accounts for 'interaction effects' between pairs of binary markers x_j and x_k (link potential), and $\theta_{j,j}$ or $\theta_{j,k}$ are elements of Θ. In the denominator, $\Psi(\Theta)$ is the normalization constant (partition function), that is

$$Z=\Psi(\Theta)=\sum_{x\in 0,1}\exp\left(\sum_{j=1}^{p}\theta_{j,j}x_j+\sum_{1\le j<k\le p}\theta_{j,k}x_jx_k\right)$$

This partition function guarantees that the sum of all probabilities adds up to one over the sample space. A difficulty is that computation of the partition function requires evaluation of many terms in the outer sum. For instance, for p variables each having k states, evaluation of the partition function involves a sum over k^p elements; in the wheat data, $k=2$, $p=1{,}279$, resulting in $2^{1{,}279}$ terms. This, in general, makes computation of the partition function itself and of its derivatives infeasible. A number of authors have proposed approximation methods to overcome this challenge.

Meinshausen and Bühlmann (2006) proposed a pseudo-likelihood based technique for finding a high-dimensional Gaussian network structure via regressing each variable (e.g., a marker) on the rest of the variables. An application in the context of gene regulatory networks can be found in Krämer et al. (2009). In this framework, construction of a LD network can be casted as a p-regressions problem, and the network structure is recovered from the sparsity pattern of the estimated regression coefficients. The least-squares estimates of the coefficients of the linear regression of marker j on all other markers are

$$\begin{aligned}\hat{\beta}^{-j}&=\text{argmin}\sum_{i=1}^{n}||X_j-X_{-j}\beta^{-j}||^2\\&=(X_{-j}^TX_{-j})^{-1}X_{-j}^TX_j\end{aligned} \tag{4}$$

where $\hat{\beta}^{-j}=(\hat{\beta}_1^j,\ldots\hat{\beta}_{j-1}^j,\hat{\beta}_{j+1}^j\cdots\hat{\beta}_p^j)$ is the $(p-1)\times 1$ vector of estimates of the regression of marker j on all other markers, X_j is the $n\times 1$ vector of genotypes for the jth marker, and X_{-j} is the resulting $n\times(p-1)$ genotype matrix after removing the jth marker from $X_{n\times p}$. Subsequently, all $\hat{\beta}^{-j}$ $(j=1,2,\ldots,p)$ are combined together to form a $(p\times p)$ matrix $\hat{\Theta}$ of estimates, obtained by placing $\hat{\beta}^{-j}$ in the jth row of Θ, with its $\Theta_{j,j}$ element set to zero. Thus, the "total" objective function is comprised of the sum of the log-likelihoods from the p regressions. For example, with p markers, the estimated MN parameter $\hat{\Theta}$ in (3) takes the form:

$$\hat{\Theta}=\begin{bmatrix}0&\hat{\beta}_2^{-1},&\ldots,&\hat{\beta}_{p-1}^{-1}&\hat{\beta}_p^{-1}\\\hat{\beta}_1^{-2}&0&\ldots,&\hat{\beta}_{p-1}^{-2}&\hat{\beta}_p^{-2}\\\vdots&\ldots,&0&\ldots,&\vdots\\\hat{\beta}_1^{-(p-1)}&\ldots,&\hat{\beta}_{p-2}^{-(p-1)}&0&\hat{\beta}_p^{-(p-1)}\\\hat{\beta}_1^{-p}&\ldots,&\hat{\beta}_{p-2}^{-p}&\hat{\beta}_{p-1}^{-p}&0\end{bmatrix}$$

The procedure given above involves performing a series of p regressions, where p is the total number of markers considered. In the $n<<p$ setting that is common in genomic data ($n=599$, $p=1{,}279$ in the wheat data), however, this regression framework above is inappropriate and a regularized regression approach is needed. Meinshausen and Bühlmann (2006) employed the least absolute shrinkage and selection operator (Lasso) (Tibshirani 1996), which is a common choice for high-dimensional regression models, because it can introduce sparsity, i.e., each marker would eventually have a small number of edges.

The analogy between the method of Meinshausen and Bühlmann (2006) and a sparse binary MN was proposed by Ravikumar et al. (2010), but using a generalized linear model instead. The pseudo-likelihood based on the 'local' conditional likelihood associated with each binary marker can be represented as:

$$l(\Theta)=\prod_{i=1}^{n}\prod_{j=1}^{p}\pi_{i,j}^{x_{i,j}}(1-\pi_{i,j})^{1-x_{i,j}} \tag{5}$$

where $\pi_{i,j}$ is the conditional probability that $x_{i,j}=1$, given all other markers (Ravikumar et al. 2010; Guo et al. 2010); i denotes individual (wheat line) and j indicates a given marker. Using a logistic link function,

$$\begin{aligned}\pi_{i,j}&=\text{P}(x_{i,j}=1|x_{i,k},k\neq j;\theta_{j,k},1\le k\le p)\\&=\frac{\exp(\theta_{j,j}+\sum_{k\neq j}\theta_{j,k}x_{i,k})}{1+\exp(\theta_{j,j}+\sum_{k\neq j}\theta_{j,k}x_{i,k})}\end{aligned}$$

This is a logistic regression where the jth marker is the response variable and the remaining markers are covariates. Equation (5) can be rewritten as

$$\begin{aligned}l(\Theta)&=\prod_{i=1}^{n}\prod_{j=1}^{p}\left(\frac{\pi_{i,j}}{1-\pi_{i,j}}\right)^{x_{i,j}}(1-\pi_{i,j})\\&=\prod_{i=1}^{n}\prod_{j=1}^{p}\left(\exp\left(\theta_{j,j}+\sum_{k\neq j}\theta_{j,k}x_{i,k}\right)\right)^{x_{i,j}}\\&\left(1+\exp\left(\theta_{j,j}+\sum_{k\neq j}\theta_{j,k}x_{i,k}\right)\right)^{-1}\end{aligned} \tag{6}$$

Taking the logarithm of Eq. (6) gives

$$\log(l(\Theta)) = \sum_{j=1}^{p}\sum_{i=1}^{n}\left[x_{i,j}\left(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\right) - \log\left(1 + \exp\left(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\right)\right)\right]$$

Adding an L1 penalty term to the above equation, the penalized log-likelihood function to be optimized with respect to Θ becomes

$$\max \sum_{j=1}^{p}\left\{\sum_{i=1}^{n}\left[x_{i,j}\left(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\right) - \log\left\{1 + \exp\left(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\right)\right\}\right] - \lambda_j\sum_{k\neq j}|\theta_{j,k}|\right\}$$

where λ_j is the regularization parameter for the jth marker; note that $\theta_{j,j}(j = 1,\ldots,p)$ is not penalized. Ravikumar et al. (2010) performed p separate logistic regressions while imposing p different sparsity constraints through L1 regularization on each regression. This means that construction of a MN on a graph G is equivalent to recovering a neighborhood set for the jth variable, for all $j \in V$. The neighborhood set of the jth variable is defined as all variables (except the jth one) that are not shrunken to zero. The authors showed that if sample size n grows faster than $d^3\log(p)$, where d is the maximum neighborhood size in the network, this leads to asymptotically consistent estimates of parameters as well as to model selection.

The resulting matrix of neighborhood estimates $\hat{\Theta}$ is not necessary symmetric. Coefficients $\theta_{j,k}$ and $\theta_{k,j}$ may have a different value or sign, so an additional step is needed to induce symmetry. Ravikumar et al. (2010) proposed following two simple rules

$$\text{max rule} : \hat{\theta}_{j,k} = \begin{cases} \hat{\theta}_{j,k} & \text{if} \quad |\hat{\theta}_{j,k}| > |\hat{\theta}_{k,j}| \\ \hat{\theta}_{k,j} & \text{if} \quad |\hat{\theta}_{j,k}| \leq |\hat{\theta}_{k,j}| \end{cases}$$

$$\text{min rule} : \hat{\theta}_{j,k} = \begin{cases} \hat{\theta}_{j,k} & \text{if} \quad |\hat{\theta}_{j,k}| < |\hat{\theta}_{k,j}| \\ \hat{\theta}_{k,j} & \text{if} \quad |\hat{\theta}_{j,k}| \geq |\hat{\theta}_{k,j}| \end{cases} \tag{7}$$

In our analysis, the final Θ was constructed using a slightly modified version of the min rule in equation (7) according to Krämer et al. (2009). If the sign of $\hat{\theta}_{j,k}$ is not equal to that of $\hat{\theta}_{k,j}$, both elements are set to zero. Further, we took

$$\tilde{\Theta} = \sqrt{\hat{\Theta} \bullet \hat{\Theta}^T}$$

where $\hat{\Theta} \bullet \hat{\Theta}^T$ is the Hadamard (Schur) product, or element-by-element multiplication and the square root operator applies to all elements of the product. This is viewed as an 'OR' rule, i.e., markers x_j and x_k are considered independent if $\hat{\theta}_{j,k}$ or $\hat{\theta}_{k,j}$ are zero. Any element $\theta_{ij} > 1$ was forced to one. The MN parameter Θ was estimated by the following cyclic coordinate descent algorithm.

Logistic Lasso regression via a cyclic coordinate descent algorithm: In a linear regression setting, the response variable is the jth marker, and the predictors are the remaining $j - 1$ markers. In ordinary least-squares, the solution is as (4), but in a Lasso setting, with the jth marker as response variable, one needs to find

$$\hat{\theta}_{-j} = \text{argmin}_{\theta}\sum_{i=1}^{n}\left(x_{i,j} - \sum_{k\neq j}^{p}\theta_{j,k}x_{i,k}\right)^2 + \lambda(\theta)$$

where $x_{i,j}$ be the genotype for response marker j observed in individual i, which takes values 0 or 1, and λ is the regularization parameter applied to all regressions on markers other than j. The algorithm applied in this study was cyclic coordinate descent (CCD) (Friedman et al. 2010).

The λ parameter plays a central role on the degree of graph sparsity, and it needs to be chosen for each of the p regressions. CCD first searches for the smallest $\lambda_{\max}$ that shrinks every coefficient to zero. Then it produces a decreasing sequence of values from $\lambda_{\max}$ to $\lambda_{\min}$. In this study, $\lambda_{\min}$ was chosen such that $\lambda_{\min} = \epsilon\lambda_{\max}$, with ϵ set to 0.01. Computation was carried out in the R environment via the glmnet R package (Friedman et al. 2010).

A major advantage of this method is that computation of the partition function is not needed, and it leads to efficient estimation of the edge set E. The disadvantage is that it does not optimize the log-likelihood jointly, so that p different regularization parameters need to be tuned. This p individual Lasso logistic regressions method only finds the elements of $\theta_{i,\ j}$ that are present or absent, instead of estimating Θ fully. Thus, this method can be viewed as an approximation to full maximum penalized likelihood (Hastie et al. 2009).

Pseudo-likelihood based regression with a single regularization parameter

Other methods, instead of trying to capture a pattern of zeros in Θ via separate p regressions, aim to optimize jointly over Θ. Under the assumption that the marker genotypes follow a Gaussian distribution, this can be achieved because the objective function has a closed form (Friedman et al. 2008; Peng et al. 2009). For instance, the Graphical Lasso (Friedman et al. 2008) produces a sparse MN in an appealing manner. It gives penalized maximum likelihood estimates of Θ, and it has been shown that the CCD algorithm can be incorporated efficiently. An illustration of this method with a gene regulatory network is in Menéndez et al. (2010).

In contrast, in binary Markov networks, optimizing the corresponding log-likelihood is very challenging because the likelihood (partition function) has no closed form. Various methods have been proposed to approximate the partition function (Guo et al. 2010; Lin et al. 2009; Kolar and Xing 2008; Höfling and Tibshirani 2009; Wang et al. 2011), and attempts have been made to achieve exact minimization of a binary-valued L1-penalized log-likelihood (Lin et al. 2009; Höfling and Tibshirani 2009; Lee et al. 2006). Höfling and Tibshirani (2009) attempted to directly extend the Graphical Lasso to a binary response; however, they discovered that computing was much slower than for a Gaussian Graphical Lasso. Instead, they suggested to approximate the full penalized log-likelihood through repeated iteration of pseudo-likelihood functions (Besag 1975), which is analogous to Ravikumar et al (2010). A main difference with Ravikumar et al (2010) is that one does not perform p independent logistic regressions, but optimizes jointly over all elements of Θ at the same time; a single regularization parameter is needed to control the overall sparsity of the resulting network. Furthermore, this ensures symmetry of $\hat{\Theta}$, so no extra steps are required to produce a matrix that is symmetric.

The log-likelihood associated with equation (3) for all n observations is given by

$$\begin{aligned} l(\Theta) &= \sum_{i=1}^{n} \log f(x_{i1}, \ldots, x_{ip}) \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \theta_{j,j} x_{ij} + \sum_{1 \le j < k \le p} \theta_{j,k} x_{ij} x_{ik} \right) \\ &\quad - n \log(\Psi(\Theta)) \end{aligned} \tag{8}$$

Now, adding the L1 penalty to equation (8) yields the objective function (Höfling and Tibshirani 2009)

$$\sum_{i=1}^{n} \log f(x_1, \ldots, x_p) - n||S \bullet \Theta||_1 \tag{9}$$

where $S = 2R - diag(R)$, R is a $p \times p$ lower triangular matrix of containing the penalty parameters, $\bullet$ is the Hadamard product, and $||S \bullet \Theta||_1$ is the penalty term in the form of an L1 norm. This implies that all off-diagonal elements have the same entry λ, without penalty for diagonal elements. A local quadratic Taylor expansion of the log-likelihood around $\Theta^{(m)}$, where m is a step of the algorithm yields (Höfling and Tibshirani 2009)

$$\begin{aligned} f_{\Theta^{(m)}}(\Theta) = C + \sum_{j \ge k} \frac{\partial l}{\partial \theta_{jk}} (\theta_{jk} - \theta_{jk}^{(m)}) + \frac{1}{2} \frac{\partial^2 l}{(\partial \theta_{jk})^2} (\theta_{jk} \\ - \theta_{jk}^{(m)})^2 - n||S \bullet \Theta||_1 \end{aligned} \tag{10}$$

where C is a constant; $\frac{\partial l}{\partial \theta_{jk}}$ is the first derivative of $l(\Theta)$ with respect to θ_{jk} is the second derivative of $l(\Theta)$, employed to form a diagonal Hessian which allows less expensive computation. If equation (10) is set to zero, the solution is soft thresholding (Hastie et al. 2009; Höfling and Tibshirani 2009) because the Hessian is diagonal, leading to

$$\hat{\theta}_{jk} = \text{sign}(\tilde{\theta}_{jk}) \left(|\tilde{\theta}_{jk}| - \frac{s_{jk}}{\frac{\partial^2 l}{(\partial \theta_{jk})^2}} \right)_+$$

where s_{jk} is the appropriate element of S, and $\tilde{\theta}_{jk}$ is the solution from the unpenalized version of $l(\Theta)$ (obtained with the Newton–Raphson method) and defined as

$$\tilde{\theta}_{jk} = \theta_{jk}^{(m)} - \left(\frac{\partial^2 l}{(\partial \theta_{jk})^2} \right)^{-1} \left(\frac{\partial l}{\partial \theta_{jk}} \right)$$

The soft thresholding operator $\left(|\tilde{\theta}_{jk}| - s_{jk} / \frac{\partial^2 l}{(\partial \theta_{jk})^2} \right)_+$ returns $|\tilde{\theta}_{jk}| - s_{jk} / \frac{\partial^2 l}{(\partial \theta_{jk})^2}$ if $|\tilde{\theta}_{jk}| > s_{jk} / \frac{\partial^2 l}{(\partial \theta_{jk})^2}$, and zero otherwise. Therefore, it shrinks $|\tilde{\theta}_{jk}|$ by the amount in the second term, or sets the amount to zero. The value at the next iteration, $\Theta^{(m+1)}$, can be found by performing a backtracking line search between $\hat{\Theta}$ and $\Theta^{(m)}$ which decides a search direction and how far to move along that direction. The pseudocode for the algorithm is shown in Table 1. The computation was carried out here via the BMN R package (Höfling and Tibshirani 2009).

Reconstruction of the network

Two implementations of an L1-regularized MN based on Ravikumar et al. (2010) and Höfling and Tibshirani (2009) were used in this study. These two methods yield a Θ matrix derived from a sparse estimator containing many zeros. Since weak associations are shrunk toward zero, this sparse matrix does not require use of pre-assigned thresholds, or to conduct a series of multiple testings to assess whether an association is significant enough or not. The procedure of Ravikumar et al. (2010) consisted of

Table 1 L1-penalized pseudo-likelihood algorithm

1. Initialize $\Theta^{(0)} = \text{diag}(\text{logit}(\hat{p}^{(0)}))$ where $\hat{p}_j^{(0)} = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$
2. Set $m = 0$
3. while no convergence is achieved
   - Set up quadratic approximation to equation (9) $f_{\Theta^{(m)}}(\Theta)$ using $\Theta^{(m)}$
   - Solve for $\hat{\Theta}$ via the soft thresholding
   - Find $\Theta^{(m+1)}$ based on $\hat{\Theta}$ through a backtracking line search
   - Set $m = m + 1$
   - end

connecting the jth with the kth locus with an undirected edge if and only if the estimates of $\theta_{j,k}$ and $\theta_{k,j}$ are both not equal to zero. In other words, if $\hat{\Theta}_{j,k} = 0$, the corresponding nodes are not connected, suggesting that markers x_j and x_k are conditionally independent, given the other markers. Similarly, if $\hat{\Theta}_{j,k} \neq 0$, then there is a link between the two nodes and the markers x_j and x_k are conditionally dependent, given the other markers. The matrix entries can be considered as edge weights, i.e., zero means absence of an edge, and values between zero and one mean presence of an edge. The actual values of $\theta_{j,k}$ are of less concern here.

Subset selection and the reference models

In regression-based procedures, the regularization parameter determines the sparsity of the networks. To interpret how this parameter controls the overall sparseness, the L1-regularized MN based on CCD was applied to a subset of 30 markers having the largest effects on wheat yield. Firstly, CCD as in Ravikumar et al. (2010) was applied to the subset of 30 markers by specifying the λ values in advance and with cross-validation (CV). The former involved carrying out 30 separate logistic regressions with λ sequences that differ from other. We chose six different λ points corresponding to the 10th, 15th, 25th, 40th, 50th, and 55th λ values in decreasing order from a sequence of 65 values of the regularization parameter that were evenly spaced on the log scale, where the 1st and the 65th values were $\lambda_{\max}$ and $\lambda_{\min}$, respectively. These maximum and minimum tuning parameters may differ for each separate sparse logistic regression. In CV approach, rather than fixing λ as done previously, 30 CVs were performed for 30 separate L1-penalized logistic regressions. This means that each sparse logistic regression had a unique regularization parameter derived from the CV. We chose deviance as loss function to use for CV and the 'cv.glmnet' function was applied for this purpose. Subsequently, the method from Höfling and Tibshirani (2009) was fitted. The regularization parameter was chosen as $\lambda = \sqrt{\log(p)/n}$ where p is the number of markers and n is the number of data points, as proposed by Ravikumar et al. (2010). For our case, $\lambda = \sqrt{\log(30)/599} = 0.075$. Lastly, the full data (1,279 markers) were fed to both methods.

This subset was chosen by fitting the Bayesian Lasso (Park and Casella 2008; de los Campos et al. 2009) assuming the following joint prior distribution.

$$p(\beta, \sigma_\epsilon^2, \tau^2, \lambda) = p(\beta|\sigma_\epsilon^2, \tau^2) \cdot p(\sigma_\epsilon^2) \cdot p(\tau^2|\lambda) \cdot p(\lambda)$$

Specifically, the prior for each of the marker coefficients β_j ($j = 1, ..., p$) was a normal prior with mean zero and variance $\sigma_\epsilon^2\tau_j^2$; a scaled inverted Chi-square distribution with scale parameter $S_\epsilon = 0.5$ and degrees of freedom $df_\epsilon = 1$ was assigned to the residual variance σ_ϵ^2; the prior for the scale parameter τ_j^2 was an exponential distribution; and a beta distribution with two shape parameters $\alpha_1 = 1.2$ and $\alpha_2 = 1.2$ spanning the range [0,500] was chosen for λ. After discarding 30,000 samples as burn-in, 50,000 samples were used to compute the posterior means of the marker effects with a thinning rate of 10 using the BLR R package (Pérez et al. 2010).

For comparison purposes, and to obtain a 'reference model', the same subset of markers was fed to a BN and was studied with the r^2 metric. The algorithm used here for learning the BN was IAMB (Incremental Association Markov Blanket) (Tsamardinos et al. 2003). This method was used for inferring the network structure of SNP markers in Holstein cattle (Morota et al. 2012). It is based on a constraint-based algorithm and estimates conditional independencies through a series of hypothesis tests. The type I error rate was set to 0.05 for Pearson's χ^2 conditional independence tests. The R package bnlearn (Scutari 2010) was used to learn the BN structure.

Results

The result of applying CCD as in Ravikumar et al. (2010) to the subset of 30 markers selected by the Bayesian Lasso is shown in Fig. 1, as rendered by igraph (Csardi and Nepusz 2006). The nodes labeled 0–29 denote the top 30 markers and a lack of an edge between two nodes indicates conditional independence. As shown in Fig. 1, the larger λ values (e.g., the 10th and 15th values of the sequences) led to sparser networks. The number of edges detected in the networks examined were 2, 4, 11, 29, 46, and 48 for the λ values in decreasing order of magnitude.

Figure 2 is a network based on a tenfold CV. Although it is not straightforward to identify the λ values used since this involved 30 CVs for each of the 30 markers as response, the sparseness of this network (18 edges) resulted from the 25th and 40th λ in the sequence given above.

In Fig. 3, the result obtained with Höfling and Tibshirani (2009) is presented. This method, which requires only one regularization parameter $\lambda = 0.075$, captured four edges. These edges were also captured in Fig. 2, e.g., markers 10 and 19 were found to be conditionally dependent, given rest of the markers.

Figure 4 displays the resulting BN used as reference. It was more dense than the network constructed through the CV based L1-regularized binary graphical model (Fig. 2), and the number of edges was exactly twice, i.e., 36 edges.

The second reference benchmark was based on the pairwise correlations among the top 30 markers, as measured by the r^2 metric. The heatmap in Fig. 5 shows the abundant

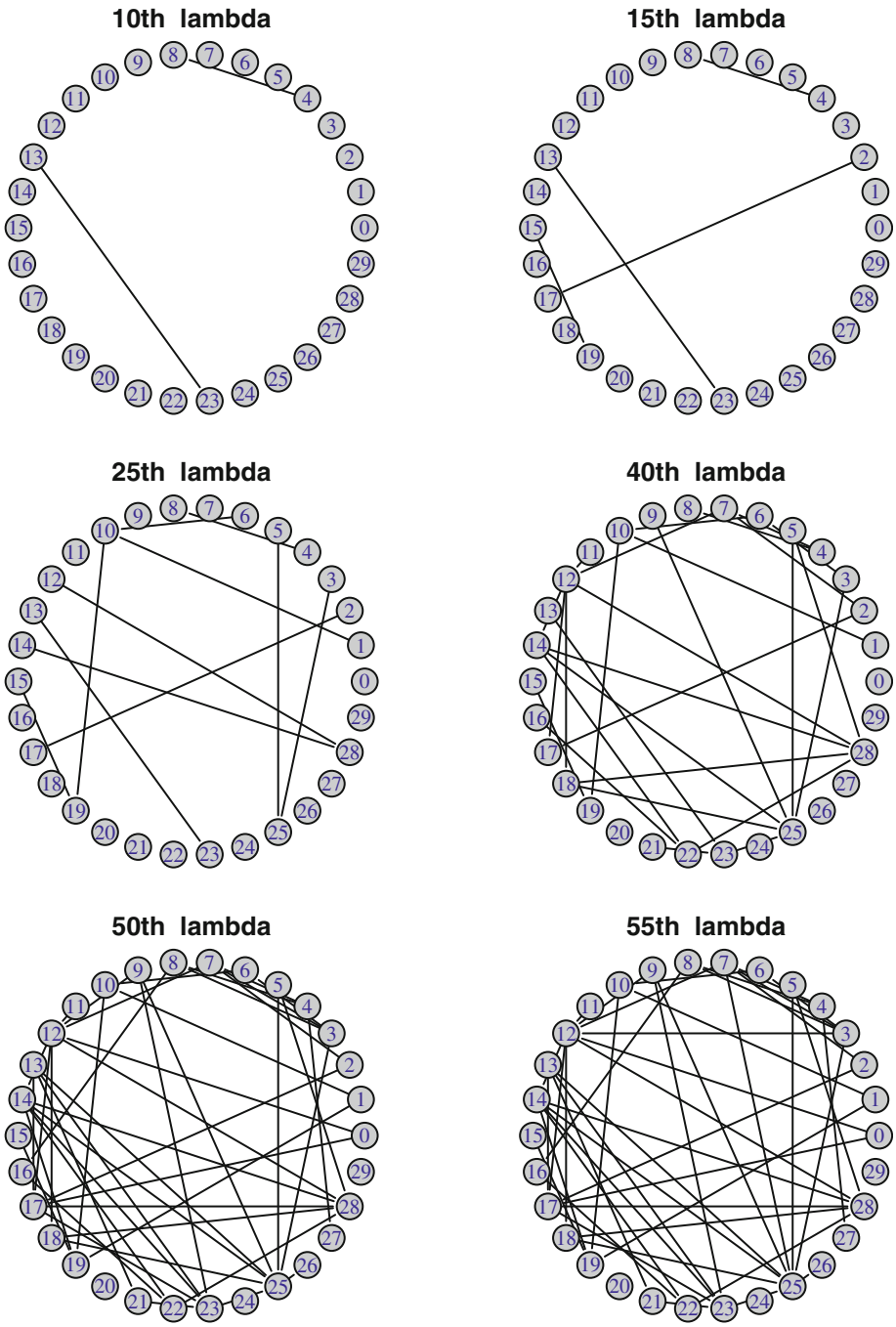

Fig. 1 LD networks with six different λ values obtained from a sequence of 65 values that were equally spaced in a log-scale. A node corresponds to a marker locus

pairwise LD present in the inbred wheat lines analyzed in this study. Comparison between the L1-regularized Markov networks and this heat map brings up a clear picture of conditional dependencies. For example, four pairs of loci captured by two variants of the L1-regularized Markov networks involved markers (4–8), (6–10), (10–19), and (13–23). Their extent of LD, as measured by the r^2 metric, was 0.93, 0.42, 0.55, 0.12, respectively. It is worthwhile noting that the pairwise correlation between loci (13–23) was lower than for other pairs. Since there are other marker pairs that showed higher LD (> 0.4) with the r^2 metric, this may suggest that the degree of LD between pair (13–23) is comparable to that of the three other pairs of loci, (4–8, 6–10,10–19), if one conditions on the remaining markers.

The full data (1,279 markers) were also fed to the CV based of Ravikumar et al. (2010), and to the $\sqrt{\log(p)/n}$ based procedure of Höfling and Tibshirani (2009). In total, with the CV-based method, 7,118 edges were identified out of the lower triangular of the $\hat{\Theta}$ matrix, with a total of 817,281 elements. The degree of sparseness was $7118/817281 \approx 0.0087$, which indicates that most edges

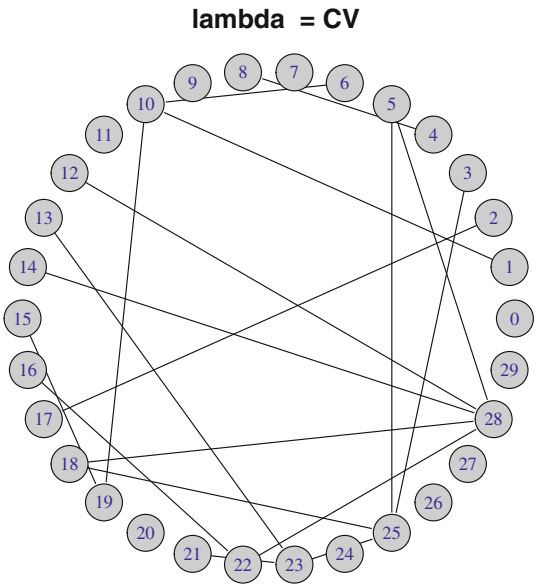

Fig. 2 L1-regularized LD network learned by the method of Ravikumar et al. with λ chosen by CV. Nodes denote 30 marker loci

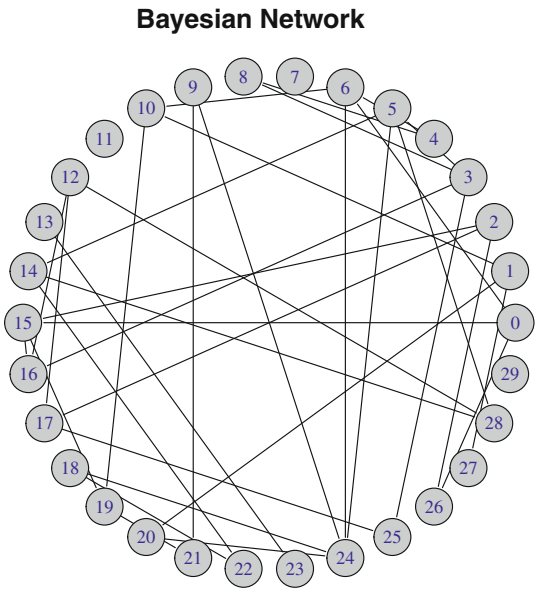

Fig. 4 LD network learned by a BN. Each node denotes a marker locus

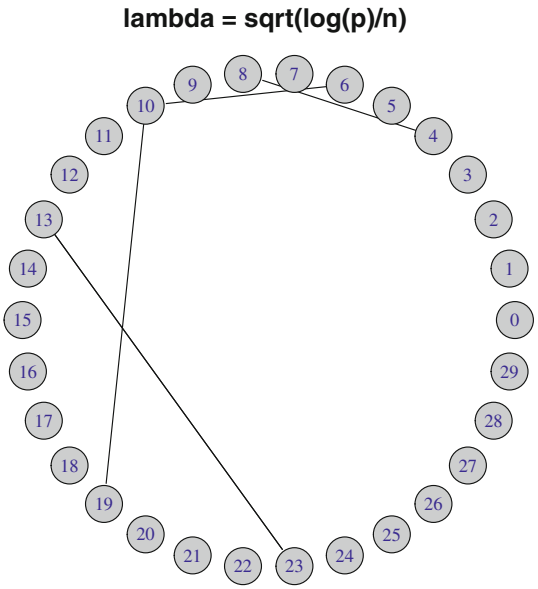

Fig. 3 L1-regularized LD network learned by Höfling and and Tibshirani's method with λ chosen as $\sqrt{log(p)/n}$ = 0.075, where $p = 30$, $n = 599$. Each node denotes a marker locus

were shrunk toward zero. The degree of sparseness was 764/817281 ≈ 0.0009 for Höfling and Tibshirani's method, i.e., even stronger than for the preceding procedure.

Discussion

Quantitative geneticists have exploited covariance structures among individuals (or genetic values) to evaluate relatedness (Fisher 1918; Wright 1921b; VanRaden 2008), estimate genetic parameters (heritability, genetic correlations) and derive predictions of breeding values for many years, e.g., (Henderson 1975). This has been achieved using two types of genetic data: pedigrees and, more recently, genetic markers.

On the other hand, less attention has been placed on how to define a covariance structure between alleles at loci distributed across the genome. Although several attempts have been made to correct the bias of the frequency dependent D' and r^2 metrics for LD, most of them still suffer from small sample size biases and depend on low minor allele frequency. Perhaps more importantly, these metrics only capture superficial marginal correlations. This study explored the possibility of employing graphical models as an alternative approach in the study of LD.

Markov networks were used for the purpose above. A MN is a graph with random variables having a distribution that factors according to an undirected graph structure $G = (V, E)$. It allows to convey a compact representation where associations between variables are symmetrical, as opposed to having a directional relationship. The joint distribution in a MN is parameterized in terms of a product of potentials over a set of cliques, and the network encodes a set of conditional independencies. Conversion of a BN to a MN, or vice versa, is possible by addition or deletion of edges and arrows, but this may produce a loss of conditional independence information. In other words, each of BN and MN are able to encode independence properties that the other cannot. This implies

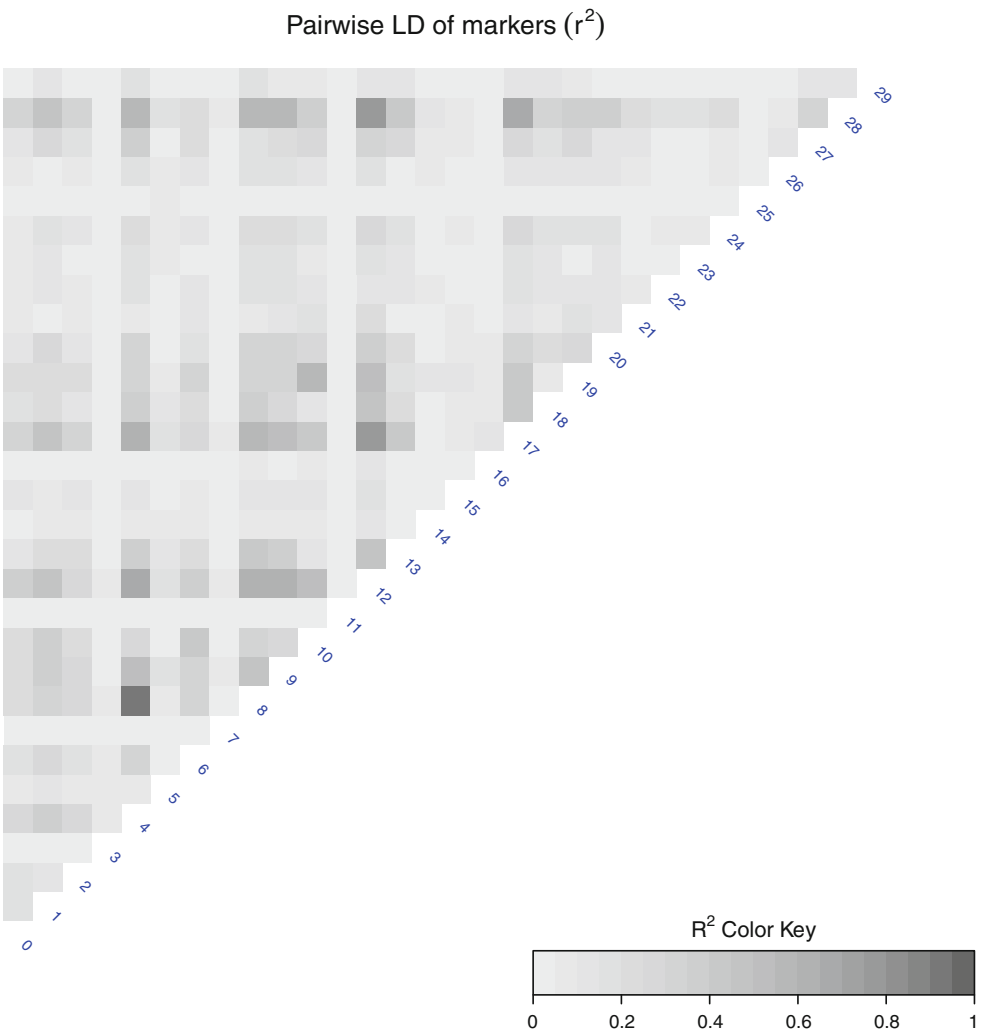

Fig. 5 LD among the top 30 markers using the r^2 metric

that it is important to decide which graphical representation is more suitable for representing the problem in question.

L1-regularized binary Markov networks were applied for reconstructing a LD network among binary markers in wheat. In the resulting binary MN, the LD associations were represented as undirected graphs, with edges connecting two loci if these were correlated, conditionally on all other markers fed to the networks. Lack of an edge indicated that two markers were conditionally independent, given the rest of the markers. The standard r^2 is viewed as a correlation between two loci but conditionally on the empty set $\emptyset$, with remaining markers ignored, i.e., pretending that the data involves just the two markers in question.

We assumed that the underlying true LD network was sparse, and used an L1 penalty to try to recover it. We illustrated such recovery using different variations of L1-regularized regressions, applied to elements of the genotype matrix **X**, which were binary valued, in the inbred wheat lines considered. Due to the properties of the L1 norm penalty, this permitted to reconstruct sparse networks, with the regularization parameter λ playing a crucial role on the degree of sparseness of the resulting networks. The L1-regularization approach is applicable to the small n, large p setting, and yielded a sparse network, i.e., each marker was connected to only a small subset of other markers. This is highly desirable for purposes of model tractability and graphical interpretation of high-dimensional data. The approach avoids assigning pre-selected thresholds, or performing numerous statistical independence test to determine which edges should go into the model.

We considered two implementations of binary Markov networks: 1) Ravikumar et al. method (Ravikumar et al. 2010) regressed each marker on all others, and 2) Höfling and Tibshirani's method (Höfling and Tibshirani 2009) applied the same degree of sparseness, to estimate all regression coefficients jointly. The latter procedure produced a symmetric matrix that is easier to interpret than that of Ravikumar et al. (2010), which does not optimize a global likelihood and, hence, it does not ensure symmetry

of the matrix. Höfling and Tibshirani's method (Höfling and Tibshirani 2009) with $\lambda = \sqrt{\log(p)/n}$ as a choice seemed to produce stronger regularization than the method in Ravikumar et al. (2010). Further, all detected edges were a subset of those found with the method of Ravikumar et al. (2010), so a consistent result with respect to sparseness structure was seen. On the other hand, a standard BN picked up many more edges due to the inability of excluding weak associations between markers. It is worth noting that some edges detected by the BN were also captured in the networks constructed with the two L1-regularized MN. The resulting networks do not tell the degree of association between two nodes through a link, but we assumed that edges that were detected in the BN, but not in the two L1-regularized Markov networks had a weak association. Further, identifying whether non-zero elements in the matrix $\hat{\Theta}$ is due to a direct association between two loci, or to an indirect association via the rest of the loci was possible by comparing the resulting networks with heat maps stemming from pairwise correlations, such as those obtained with the r^2 metric. One possible application of the networks analyzed here might be selecting tag SNPs unconditionally, as well as conditionally, on other markers when the dimension of the data is high, e.g., data generated from next generation sequence technologies.

As shown by Lewontin (1988), there is no clear agreement as to which is the best metric for capturing non-random association between alleles at pairs of loci, i.e., LD. Further, all LD metrics, as well as the networks studied here, do not reveal the causes of multi-loci associations. However, there is still room for developing methods for characterizing a LD. A 'best' metric would be one that captures complex associations, and that reflects the underlying complex genetic architecture properly. Recently, Gianola et al. (2012) proposed indexes to measure association among genetic variables via statistical distances between distributions, based either on the Kullback–Leibler logarithmic distance, or on relative distance. A departure of distributions from stochastic independence is an indication of association, and this indexes allow to capture situations where loci are jointly dependent even though their correlation may be zero. Although our applied networks depend on a pairwise structure, we attempted to evaluate LD via probabilistic graphical models, to reflect the biological expectation that loci associate as a complex system. Extending to higher-order associations is feasible, e.g., Ding et al. (2011). Also the methods applied here are suitable for binary-valued variables only, so it is limited to assessing LD in inbred lines, where there are only two possible genotypes. Our approach differs from that of Thomas and Camp (2004), where they fitted a graphical model to haplotype data using simulated annealing search procedure. We used the more familiar regression scheme and considered sparsity in a network. Note that theoretically one can obtain a conditional variance or covariance from a marginal covariance matrix. While indeed this is an appealing procedure, it still requires use of arbitrary pre-assigned thresholds to assess whether an association is significant enough or not.

To summarize, commonly used metrics, such as r^2 or the relevance networks are limited because only marginal, pairwise, associations are measured. A BN is capable of further capturing conditional associations among relevant loci. The L1-regularized Markov networks studied here preserve this feature of BN, but also delete edges that lack a strong enough evidence of both unconditional and conditional association. Sparse networks provide a clearer picture of association, and a sparse graph eases interpretation markedly, because it includes a smaller number of edges than a BN. As shown here, L1-regularized binary Markov networks are suited for the $n << p$ setting, and these models are potentially valuable for studying conditional LD from high-dimensional genotype data, where variables are expected to be sparsely connected.

Acknowledgments The authors thank the anonymous reviewers for their valuable comments. This work was supported by the Wisconsin Agriculture Experiment Station and by a Hatch grant from the United States Department of Agriculture.

References

Besag J (1975) Statistical analysis of non-lattice data. In: Proceedings of the Twenty-First National Conference on artificial intelligence, pp 179–195

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Borgelt C, Steinbrecher M, Kruse RR (2009) Graphical models: representations for learning, reasoning and data mining. Wiley, New York

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci 97(22):12182–12186

Clifford P (1990) Markov random fields in statistics. In: Grimmett GR, Welsh DJA (eds) Disorder in physical systems. A volume in honour of John M. Hammersley. Oxford University Press, New York

Crossa J, Burgueño J, Dreisigacker S, Vargas M, Herrera-Foessel SA, Lillemo M, Singh RP, Trethowan R, Warburton M, Franco J, Reynolds M, Crouch JH, Ortiz R (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177(3): 1889–1913

Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724

Csardi G, Nepusz T (2006) The igraph software package for complex network research. Int J Complex Syst:1695. http://igraph.sf.net

de Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel KA, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

Ding S, Wahba G, Zhu X (2011) Learning higher-order graph structure with features by structure penalty. In: Proceedings of the 25th Annual Conference on neural information processing systems

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399–433

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Friedman J, Hastie T, Tibshirani R (2010) Regularized paths for generalized linear models via coordiante descent. J Stat Softw 33(1):1–22

Gianola D, Manfredi E, Simianer H (2012) On measures of association among genetic variables. Anim Genet 43:19–35

Guo J, Levina E, Michailidis G, Zhu J (2010) Joint structure estimation for categorical Markov Networks. Tech Rep Department of Statistics, University of Michigan, Ann Arbor

Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11:1–12

Hammersley JM, Clifford P (1971) Markov field on finite graphs and lattices. unpublished

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447

Hill WG, Robertson A (1968) Linkage disequilibrium in finite population. Theor Appl Genet 38:226–231

Höfling H, Tibshirani R (2009) Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. J Mach Learn Res 10(3):883–906

Kolar M, Xing EP (2008) Improved estimation of high-dimensional Ising models. http://arxiv.org/abs/0811.1239

Koller D, Friedman N (2009) Probabilistic graphical models: Principles and Techiniques. The MIT Press, London

Krämer N, Schäfer J, Boulesteix AL (2009) Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinforma 10:384

Lee SI, Ganapathi V, Koller D (2006) Efficient structure learning of Markov networks using L1 regularization. In: Proceeding of the Neural Information Processing Systems

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Lin Y, Zhu S, Lee DD, Taskar B (2009) Learning sparse Markov network structure via ensemble-of-trees models. In: Proceedings of the 12th Artificial Intelligence and Statistics, Florida

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable secletion with the lasso. Ann Stat 34(3):1436–1462

Menéndez P, Kourmpetis YAI, ter Braak CJF, van Eeuwijk FA (2010) Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge. PloS One 5(12):e14–147

Meuwissen THE, Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155(1):421–430

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Morota G, Valente BD, Rosa GJM, Weigel KA, Gianola D (2012) An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network. J Anim Breed Genet 129(6):474–487

Neapolitan RE (2003) Learning Bayesian Networks. Prentice Hall, New Jersey

Newton MA (1999) Thoughts on gibbs distributions and markov random fields. Course notes http://wwwstatwiscedu/∼newton/st775/materials/notes/gibbspdf. Accessed 15 August 2012

Park T, Casella G (2008) The bayesian LASSO. J Am Stat Assoc 157:1819–1829

Peng J, Wang P, Zhou N, Zhu J (2009) Partial correlation estimation by joint sparse regression model. J Am Stat Assoc 104:735–746

Pérez P, de∼los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3(2):106–116

Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional Ising model selection using L1 regularized logistic regression. Ann Stat 38(3):1287–1319

Scutari M (2010) Learning Bayesian networks with the bnlearn R package. J Stat Softw 35(3):1–22

Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. Nat Biotechnol 24:427–433

Thomas A, Camp NJ (2004) Graphical modeling of the joint distribution of alleles at associated loci. Am J Hum Genet 74(6):1088–1101

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc 58:267–288

Tsamardinos I, Aliferis CF, Statnikov A (2003) Algorithms for large scale Markov blanket discovery. In: Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference

VanRaden PM (2008) Efficient methods to compute genomic prediction. J Dairy Sci 91:4414–4423

Wang P, Chao DL, Hsu L (2011) Learning oncogenic pathways from binary genomic instability data. Biometrics 67:164–173

Wright S (1921a) Correlation and causation. J Agric Res 20:557–585

Wright S (1921b) Systems of mating. I. The biometric relations between parents and offspring. Genetics 6:111–123