Graphical Modeling of the Joint Distribution of Alleles at Associated Loci

Alun Thomas[1,2] and Nicola J. Camp[2]

[1]Department of Medical Informatics and [2]Center for High Performance Computing, University of Utah, Salt Lake City

Pairwise linkage disequilibrium, haplotype blocks, and recombination hotspots provide only a partial description of the patterns of dependences and independences between the allelic states at proximal loci. On the gross scale, where recombination and spatial relationships dominate, the associations can be reasonably described in these terms. However, on the fine scale of current high-density maps, the mutation process is also important and creates associations between loci that are independent of the physical ordering and that can not be summarized with pairwise measures of association. Graphical modeling provides a standard statistical framework for characterizing precisely these sorts of complex stochastic data. Although graphical models are often used in situations in which assumptions lead naturally to specific models, it is less well known that estimation of graphical models is also a developed field. We show how decomposable graphical models can be fitted to dense genetic data. The objective function is the maximized log likelihood for the model penalized by a multiple of the model's degrees of freedom. We also describe how this can be modified to incorporate prior information of locus position. Simulated annealing is used to find good solutions. Part of the appeal of this approach is that categorical phenotypes can be included in the same analysis and association with polymorphisms can be assessed jointly with the interlocus associations. We illustrate our method with genotypic data from 25 loci in the *ELAC2* gene. The results contain third- and fourth-order locus interactions and show that, at this density of markers, linkage disequilibrium is not a simple function of physical distance. Graphical models provide more flexibility to express these features of the joint distribution of alleles than do monotonic functions connecting physical and genetic maps.

Introduction

Graphical modeling is the statistical study of high-dimensional joint distributions that break down into products of simpler, lower-dimensional factors. A graphical model consists of a Markov graph that represents conditional independences between the variables and the set of parameters that define the component factors. Although the graphical models are not always made explicit, they enable computations to be made on complex models in a tractable way in many areas of genetics. Perhaps the best example is the *peeling* method of Elston and Stewart (1971) and Thompson et al. (1978), which enables linkage and segregation analyses in extended pedigrees. In this example, the variables are the genotypes of pedigree members, the Markov graph is constructed by the union of the triangles formed by connecting parent/offspring triplets, and the parameters are those that describe Mendelian inheritance, the penetrance probabilities, and the population genotype frequencies. This particular application displays almost all the features of the later general developments by Lauritzen and Spiegelhalter (1988). Most Markov chain Monte Carlo (MCMC) methods also have an underlying graphical model, although, again, this is not always made explicit (but see Thomas et al. [2000] for an example in which it is).

Although less commonly done, graphical models can be estimated empirically from observations from a joint distribution. The methods, developed by Højsgaard and Thiesson (1995) and implemented by them in the BIFROST program, are based on maximizing a penalized likelihood function. In this article, we describe the estimation of graphical models to describe linkage disequilibrium.

Modeling linkage disequilibrium, or the tendency of alleles observed at one locus not to be independent of alleles observed at nearby loci, in an appropriate way has become an important problem in statistical genetics. Dense maps of polymorphisms, particularly SNPs, are now available. More than 3,000,000 polymorphisms are currently reported in the human genome. Moreover, the cost of assaying genotypes for samples of individuals has decreased, and the speed with which this can be done increased. Since SNPs are usually diallelic, taken individually they are comparatively uninformative, so it is essential to consider the haplotypes formed by combinations of them. On the other hand, the number of possible haplotypes grows exponentially with the number of loci and, potentially, so does the number of pa-

Received September 30, 2003; accepted for publication March 10, 2004; electronically published April 26, 2004.

Address for correspondence and reprints: Dr. Alun Thomas, Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108. E-mail: alun@genepi.med.utah.edu

0002-9297/2004/7406-0003$15.00

rameters required to describe their frequencies. The approach we present here seeks to replace the full-dimensional contingency table with a product of tables of far lower dimension. We estimate not only the multinomial parameters of the contingency tables but also which sets of variables give the best factorization.

This approach contrasts with other techniques used in this field, such as the PHASE method for haplotype reconstruction (Stephens et al. 2001; PHASE Web site), which gains power by modeling the underlying processes of mutation and recombination. Although most of the work presented here concerns a straightforward application of graphical model estimation, we also describe how this can be adapted to incorporate prior information about locus position. In this way, we hope to claim some additional power without losing the advantages of robustness and computational efficiency of our empirical method. We illustrate our method with an analysis of 25 loci in and around the *ELAC2* gene in a sample of 688 unrelated individuals from a study of cancer in Utah.

Most approaches to modeling linkage disequilibrium impose strong constraints based on physical location. For instance, a monotone function is often used to transform the physical map into the genetic map, as in the analysis of loci on chromosome 22 by Dawson et al. (2002) and of haplotype blocks and recombination hotspots by Daly et al. (2001); Johnson et al. (2001) characterize linkage disequilibrium in terms of the behavior of contiguous genetic regions. Although such approaches are valuable for describing large genetic regions, at the scale of resolution we examine, as our results show, the tendency of linkage disequilibrium to decay with distance has less of an influence on the joint distribution than the mutation history. The greater flexibility of graphical models can better describe patterns of linkage disequilibrium under these circumstances and also allows for characterization of higher-order interactions between the loci considered.

Our approach is empirical. We do not use a model for population genetics; hence, we cannot directly make inferences about how the population evolved or how it will continue to do so in future. We can, however, provide a tractable, concise, and accurate representation of the current linkage disequilibrium structure that is directly relevant to mapping phenotypes by association and to choosing informative subsets of loci.

Methods

Haplotype Data

The method described below assumes the input of a list of haplotypes independently sampled from a population. Let $x_{i,j}$ be the allele of the ith haplotype at the jth locus, and let the number of possible alleles at the jth locus be a_j. Let $X = \{x_{i,j} \forall i,j\}$. Since most of the data available will be SNPs, a_j will usually be 2; however, the method is general enough to handle any number of alleles and, indeed, any categorical data.

Haplotypes may be determined experimentally, reconstructed from family data, or reconstructed from a sample of genotypes from a population through use of programs such as PHASE (Stephens et al. 2001), D. Clayton's SNPHAP (SNPHAP Web site), or GCHap (Thomas 2003). At this stage, we will also assume that the data are complete, with no unobserved variables.

For each locus j, we define L_j, the random variable that is the allele of a randomly chosen haplotype from the population at the jth locus. Let the full set of variables be $V = \{L_1, L_2, \ldots, L_n\}$. The full joint distribution of V can be considered as an n-dimensional contingency table. The high dimensionality and high number of multinomial parameters, however, mean that this is not a useful or tractable representation, and what follows will be aimed at estimating from the data, X, less complex but informative models.

We define S to be any subset of the elements in V. For each S, there are $\prod_{i:L_i \in S} a_i$ possible allelic combinations across the loci in S. If S contains s loci, then classification of the data in X by the loci in S can be done with an s-dimensional contingency table. If we let y_i be the number of observations in the ith cell of the table and let p_i be the model probability of an observation in that cell, the usual maximum likelihood estimate of p_i is then given by

$$\hat{p}_i = \frac{y_i}{\sum_i y_i} \ ,$$

and the maximized log likelihood, up to the addition of a constant, is given by

$$\log[\hat{L}(S)] = \sum y_i \log(\hat{p}_i) \ . \qquad (1)$$

The degrees of freedom for the contingency table are

$$\mathrm{df}(S) = \prod_{i:L_i \in S} a_i - 1 \ . \qquad (2)$$

Note that the sum in equation (1) requires only the nonzero values of y_i and, thus, requires computational time that is a linear function of the number of observed haplotypes, or rows, of X. It can, therefore, be calculated even for contingency tables of very high dimension.

Graphical Models

Graphical models can specify stochastic systems of lower complexity to describe data in high-dimension contingency tables. For a full discussion, see Lauritzen and Spiegelhalter (1988) and Højsgaard and Thiesson (1995). In what follows, we will consider graphs in which the vertices represent the variables of our system, $V = \{L_1, L_2, \ldots, L_n\}$, and edges represent the associations between them. To avoid another layer of notation, we will have L_i denote both the ith variable and the vertex of the graph that represents it.

A graphical model for a set of random variables $V = \{L_1, L_2, \ldots, L_n\}$ is defined by a graph G and a model M. The vertices of G are V, and G has edges E such that the joint distribution of V can be factorized into a product of k nonnegative functions, $\phi(C_l)$, whose arguments $\{C_l\}$ are the complete subgraphs, or cliques, of G. That is,

$$P(V) = \prod_l \phi(C_l) \ , \ C_l \subseteq V \ ,$$

for $l = 1, \ldots, k$, where, if $L_i \in C_l$ and $L_j \in C_l$ for any C_l, then the edge, (L_i, L_j), is contained in E. The model M specifies the probabilities of the distributions that $P(V)$ factorizes into.

The conditional independences of the model can then be determined from G. For example, if S_1 and S_2 are subsets of V such that all paths from any vertex in S_1 to any vertex in S_2 must pass through a third subset S_3, then S_1 and S_2 are conditionally independent, given S_3, or

$$P(S_1, S_2 | S_3) = P(S_1 | S_3) P(S_2 | S_3) \ ,$$

and S_3 is said to *separate* S_1 and S_2. Also, the conditional distribution of any variable L_i in V, given all the other variables, depends only on the values of the variables that neighbor L_i in G. That is,

$$P(L_i | V - L_i) = P[L_i | \Delta(L_i)] \ ,$$

where $\Delta(L_i)$ contains all vertices L_j such that $(L_i, L_j) \in E$.

The simplest graphical model would be a graph with vertices V but no edges, which represents complete independence between all the variables in V: this is the *trivial* graphical model. Another simple example is a first-order Markov chain, which can be represented by $G(V,E)$, where E contains the $n - 1$ edges (L_i, L_{i+1}). The most complex graphical model has all $n(n-1)/2$ possible edges and represents the n-dimensional contingency table. This is called the *saturated* model.

Decomposable graphical models are a tractable subclass of models for which the graph G has the running intersection condition for its cliques—that is, the cliques C_i can be ordered, and subsets S_i defined, such that

$$S_i = C_i \cap \bigcup_{j:j>i} C_j \subseteq C_l \ ,$$

for some $l > i$. S_i are called the "clique separators." An equivalent condition is that G is a *triangulated* graph—that is, there are no cycles of length >3 that are unchorded. For instance, four vertices cannot form a rectangle unless one of the diagonal edges is also present. If the four vertices are loci A, B, C, and D, say, such that (A,B), (B,C), (C,D), and (D,A) all form pairs in linkage disequilibrium, triangulation requires that at least one of the pair (A,C) or (B,D) must also be in linkage disequilibrium. For further information about triangulated graphs, see Golumbic (1980). A decomposable graphical model can be specified by such an ordering of cliques, which can be found in linear time in a decomposable graph using the *maximum cardinality search* (Tarjan and Yannakakis 1984). This decomposition of G allows the joint distribution of V to be factorized as

$$P(V) = \prod_{i=1}^{k} P(C_i | S_i) = \prod_{i=1}^{k} \frac{P(C_i)}{P(S_i)} \ ,$$

and, thus, the n-dimensional contingency table for V is decomposed as a function of lower-dimensional tables. Note that S_k is always the empty set.

The Y Chromosome and Mitochondrion

As an illustrative aside, we describe how to derive the graphical model for regions of no recombination, such as the Y chromosome, mitochondrion, and small autosomal haplotype blocks. We apply the simplifying assumption of the infinite sites model that no second mutation occurs at a polymorphic site. Hence, all polymorphisms are diallelic, and a mutated allele never reverts back to the original allele. We also assume that all haplotypes ancestral to those in the sample are also present in the sample. Although this is not a plausible assumption, it simplifies the illustration, and, in fact, a similar relationship between the ancestral tree and graphical model exists for estimated trees containing unobserved but inferred ancestral haplotypes. Under our assumptions, we can reconstruct the ancestry completely by drawing a tree in which haplotypes are nodes and haplotypes that differ at one base only are connected. If we know from an outgroup, or related species, which haplotype is the wild type, we can also root the tree; however, this is not necessary to derive the graphical model. A small example is shown in figure 1. The ancestral haplotype is 00000000, and mutations are the places in which 1s replace 0s.

Under these assumptions, a SNP is represented by a

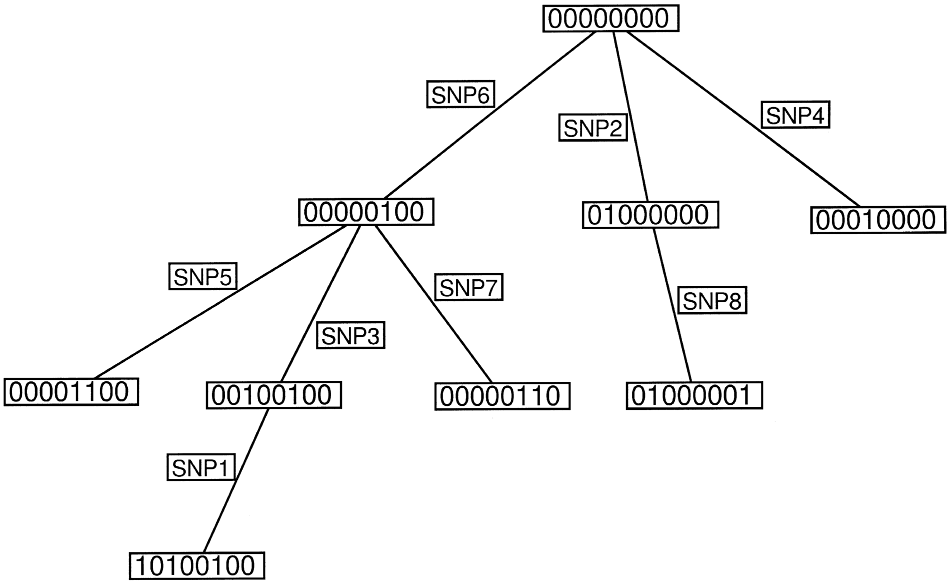

Figure 1 An example of the ancestry of haplotypes in a region with no recombination

unique branch in the tree, the one that connects haplotypes that differ only at the SNP in question. From this tree, we can construct a new graph in which vertices represent the SNPs and edges join SNPs whose branches in the original tree shared an end point. This derived graph is the Markov graph for the SNP locus variables and is shown in figure 2. The mode of derivation ensures that the graph is triangulated, and, therefore, the graphical model is decomposable.

To see that the graph does indeed represent the appropriate conditional independences, consider, for example, the derivation of the haplotype 01000001 from 01000000. First, note that, since allele 1 occurs at SNP2 because of a mutation on the original ancestral haplotype, if we observe allele 1 at SNP2, all SNPs other than SNP8 must have allele 0, regardless of the allele at SNP8. That is, for $a = 0$ or $a = 1$,

$$\begin{aligned}
&P\big(\{\text{SNP1,SNP3,}\dots\text{,SNP7}\} = \{0,0,\dots,0\}\\
&\qquad |\text{SNP2} = 1,\text{SNP8} = a\big)\\
&= P\big(\{\text{SNP1,SNP3,}\dots\text{,SNP7}\} = \{0,0,\dots,0\}\\
&\qquad |\text{SNP2} = 1\big)\\
&= 1\ .
\end{aligned}$$

On the other hand, since allele 1 at SNP8 occurs as a mutation on a haplotype that has allele 1 at SNP2, if we observe allele 0 at SNP2, then the allele at SNP8 must also be 0, regardless of the alleles at the remaining loci. That is,

$$\begin{aligned}
&P\big(\text{SNP8} = 0|\text{SNP2} = 0,\\
&\qquad \{\text{SNP1,SNP3,}\dots\text{SNP7}\} = \{a_1,a_3,\dots a_7\}\big)\\
&\quad = P(\text{SNP8} = 0|\text{SNP2} = 0)\\
&\quad = 1\ ,
\end{aligned}$$

where each $a_i = 0$ or 1.

Hence, given either allele at SNP2, {SNP8} and {SNP1,SNP3,…,SNP7} are independent sets. The other conditional independences can be similarly derived.

Fitting Graphical Models

For any of the cliques C_i or separators S_i, we can find the maximized log likelihood and degrees of freedom from equations (1) and (2). Moreover, the maximized log likelihood for the whole graphical model G is given by

$$\begin{aligned}
\log\big[\hat{L}(G)\big] &= \max_M \log[L(G,M)]\\
&= \sum_{i=1}^{k} \log\big[\hat{L}(C_i)\big] - \sum_{i=1}^{k} \log\big[\hat{L}(S_i)\big]\ , \qquad (3)
\end{aligned}$$

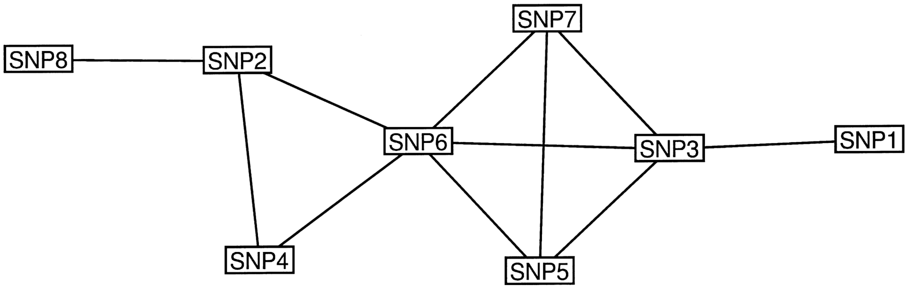

Figure 2 The graphical model for the SNPs derived from figure 1. The cliques are as follows, from left to right: {SNP8, SNP2}, {SNP2, SNP4, SNP6}, {SNP6, SNP7, SNP5, SNP3}, and {SNP3, SNP1}. The clique intersections are {SNP2}, {SNP6}, and {SNP3}.

and the degrees of freedom are given by

$$\mathrm{df}(G) = \sum_{i=1}^{k} \mathrm{df}(C_i) - \sum_{i=1}^{k} \mathrm{df}(S_i).$$

Thus, both $\log[\hat{L}(G)]$ and $\mathrm{df}(G)$ can be computed efficiently.

Straightforward maximization of the log likelihood over Ω, the set of decomposable graphical models, will choose saturated models; hence, we follow the strategy of the BIFROST implementation (Højsgaard and Thiesson 1995) by using the heuristic information criterion,

$$\mathrm{IC}(G) = \log\left[\hat{L}(G)\right] - \alpha \mathrm{df}(G) \ ,$$

where α is a complexity penalty to be specified. In the current implementation, we use $\alpha = [\log(m)]/2$, where m is the number of haplotypes observed, which is the Bayesian information criterion of Schwarz (1978).

We can specify a partial ordering on Ω by describing G as a submodel of G' if and only if $E \subset E'$, and the BIFROST program exploits this to choose a model by a stepwise backward elimination strategy. However, the result of this is usually a local optimum. We attempt to improve on this by using simulated annealing (Kirkpatrick and Gellatt 1982). The strategy is as follows:

1. Find an initial incumbent graph, G, which represents a decomposable graphical model and evaluate IC(G).
2. Propose a new model, G'.
3. If G' is decomposable and

$$\mathrm{IC}(G') \geqslant \mathrm{IC}(G) - \gamma \times Z \ ,$$

then G' becomes the new incumbent. Otherwise, G remains the incumbent. In either case, iterate from step 2.

For each iteration, Z is a new independent random realization from the Exponential(1) distribution, and γ is the annealing parameter, which is set at a large value and gradually reduced to 0, at which time the process becomes a random downhill search.

The way in which G is perturbed to form G', the proposed new incumbent, is the most important feature of the search. At each iteration we choose, at random, one of the following perturbations:

1. Randomly select two vertices of G. If they are connected, form G' by disconnecting them. If they are disconnected, form G' by connecting them.
2. Choose a random vertex, L_i, of G, which has at least one neighbor and at least one nonneighbor in the graph. Choose a random neighbor L_j of L_i and a random nonneighbor L_l. Form G' by disconnecting (L_i,L_j) and connecting (L_i,L_l).
3. Choose, at random, r nonoverlapping pairs of vertices. Apply perturbation 1 above to each pair simultaneously. Implement for $r = 2,\dots,5$.
4. Choose, at random, a set of r distinct vertices. Apply perturbation 1 above to each of the $r(r-1)/2$ pairs simultaneously. Implement for $r = 3,\dots,5$.
5. Choose, at random, two distinct vertices of G. Form G' by exchanging their neighbor sets.
6. Choose a random vertex of G. Form G' by disconnecting the vertex from its neighbors and connecting it to a new random set of neighbors of the same size.

The underlying Markov chain induced by this process is irreducible, since any decomposable graph G can be disassembled to the trivial graph with no edges by removing one edge at a time in such a way that all intermediate graphs are also decomposable. Although only perturbation 1 is necessary to ensure irreducibility, test runs have shown that the others significantly improve the mixing properties.

In addition, the perturbations have all been constructed so that the probability of proposing a graph G' from G is the same as that of proposing G from G'. Thus, the scheme is reversible, and, therefore, for a fixed value of γ, the whole process is equivalent to Metropolis sampling

(Metropolis et al. 1953) from the set of all graphs with n vertices with ergodic distribution:

$$P(G) \propto \begin{cases} \left[\hat{L}(G)e^{-\alpha \mathrm{df}(G)}\right]^{1/\gamma} & \text{if } G \text{ is decomposable} \\ 0 & \text{otherwise} \end{cases} .$$

Thus, setting $\gamma = 1$ will allow posterior sampling of graphical models in an approximate or pseudo-Bayesian manner. Hence, we can approximate the posterior probability that particular edges exist or that particular conditional independences hold. For arbitrary values of γ, we can use importance sampling with weights

$$\left[\hat{L}(G)e^{-\alpha \mathrm{df}(G)}\right]^{[1-(1/\gamma)]} = e^{\mathrm{IC}(G)\times[1-(1/\gamma)]} .$$

Using $\gamma > 1$ will typically give a chain with far better mixing properties. To avoid the numerical problems of dealing with very small numbers, it is better to scale the weights. If G_0 is some fixed high scoring graph, we can use the weights

$$e^{\{[\mathrm{IC}(G)-\mathrm{IC}(G_0)]\times[1-(1/\gamma)]\}} .$$

If G_0 can be chosen to be the graph that globally maximizes the information function, or if we can find an upper bound for $\mathrm{IC}(G_0)$, then the same weights can also be used for rejection sampling.

Incorporating Information on Physical Location

The method we have presented so far is a standard application of graphical modeling that could be applied to any categorical data. We have attempted to describe the data as presented, without attempting to model the processes that gave rise to it. As Stephens et al. (2001) have shown with their development of the PHASE haplotype estimation procedure, statistical power can be improved by incorporating knowledge of the population genetic processes. The cost is typically more involved computation and possibly a lack of robustness to deviation from the assumed model.

In our illustration below, in which we have a large number of haplotypes from which to estimate relationships between a tightly linked set of markers, the effects of more involved modeling is likely small. However, for data spanning larger genomic regions, the effect of our knowledge that linkage disequilibrium tends to decay with distance will be greater. Such knowledge can be incorporated very naturally within our estimation framework, by altering the prior distribution to reflect the fact that, before looking at the data, our prior belief is that proximal loci are more likely to be dependent than distant ones. We can do this by adding another penalty term to the score of a graphical model that penalizes edges in the graph by how far apart the joined loci are. If we let z_i be the position of the ith locus and use the square root of distance between loci as our penalty, the altered score becomes

$$\mathrm{IC}'(G) = \log\left[\hat{L}(G)\right] - \alpha \mathrm{df}(G) - \beta \sum_{(i,j)\in G} \sqrt{|z_i - z_j|} . \tag{4}$$

The square-root function has been used to mimic the slow decay rate in linkage disequilibrium seen in the graphs of Dawson et al. (2002); however, this and the appropriate value of β need further analysis.

A far stronger possible use of prior information could be made by restricting the Markov graph to the subset of interval graphs. Interval graphs are triangulated graphs that are defined by associating a vertex with an interval on the real line. Vertices are joined if and only if the associated intervals have a nonempty intersection. Each locus can be represented by a region on the genome that contains the locus position, with the bounds of the region to each side of the locus representing, roughly speaking, the limits to which linkage disequilibrium with the locus extends. Such a restriction has some intuitive appeal, would be straightforward to program, should greatly increase the power to estimated models within the subclass, and so should be considered when mapping a large genetic region. However, as can be seen from the results below, for small regions in tight linkage disequilibrium, there will be relationships in the data that cannot be represented in this way.

Testing for Disease Association

We do not have to specifically assume that the variables considered represent allelic states at loci; the method applies to any categorical data. It is straightforward, therefore, to include categorical variables associated with the haplotype variables. In particular, we can add indicators as to whether the haplotype came from an individual affected by a certain disease. We can also include covariates; for example, in familial analyses of prostate cancer, we have included whether the diagnosis was made before or after prostate-specific-antigen testing was routine and whether onset was early or late. We would also, of course, include the sex of the individual, so that the linkage disequilibrium structure could be estimated using haplotypes from males and females while simultaneously accounting for a male-specific disease. In fitting the graphical model, the phenotypes have no special status, and the fitted model may or may not indicate that there are dependences between them and the loci. We can, therefore, construct a standard χ^2 likelihood-ratio test for association comparing the estimated model with the submodel derived by removing any links between the phenotype variable and the loci.

Choosing Subsets of Loci

Shannon's information function (Shannon 1948) is an obvious criterion on which to select subsets of informative loci (Hampe et al. 2003), and, as might be suggested from the form of equation (1), its calculation is tractable for graphical models. For a contingency table with estimated probabilities $\hat{p}_i$, it is given, up to an additive constant, by the negative of the maximized log likelihood; for an estimated graphical model, it is $-\log[\hat{L}(G)]$, as given in equation (3).

To find the information in a subset of variables according to a given graphical model, we have to sum over the variables omitted. In this case, we can achieve this without explicitly making the intensive summation calculations. We do this by first finding the Markov graph obtained when the omitted variables are summed out and then calculating the maximized log likelihood of the corresponding model from the data, using equation (3). The first step is done by taking each omitted variable in turn, joining all its neighbors in the graph, and then removing that variable from the graph. The result is invariant to the order in which the variables are removed. The cliques and intersections of this resulting graph, which is always triangulated, are found, and the log likelihood—and, hence, the Shannon information—is calculated. Note again that, although this process results in graphs with large cliques, the computation in equation (3) depends only on the nonzero entries in the corresponding contingency table and, hence, again takes computational time that is a linear function of the number of observed haplotypes.

Results

We have implemented the above method in a Java program that can be downloaded from A.T.'s Web site. The data used below are also available at this site.

The program has a graphical interface that shows the current estimate of the graph and allows the user to control the annealing parameter γ. The maximized IC with log likelihood and degrees of freedom of the current best graph are output to the screen.

As an illustration, we present an analysis of association between loci in the *ELAC2* gene. From several large extended families involved in prostate, breast, and ovarian cancer studies performed by the Genetic Epidemiology group at the University of Utah, we selected all individuals whose parents were not in the pedigrees—that is, the founders. Of these, we found 688 individuals who had been genotyped at at least 12 of our panel of 25 loci, which are all within 93 kb in and around the *ELAC2* gene. The locations and brief descriptions of the loci are given in table 1. The five most downstream loci, *s21–s25,* are in the neighboring gene KIAA0672, which is an uncharacterized GTPase-activator protein for Rho-like GTPases. Two of the loci are reported to associate with prostate cancer (Tavtigian et al. 2001; Camp and Tavtigian 2002): $SL = s10$, which changes a serine to a leucine at amino acid position 217, and $AT = s15$, which changes an alanine to a threonine at position 541. All samples were obtained under appropriate institutional-review-board approval at the University of Utah.

Table 1

Positions of the 25 Polymorphisms in and around *ELAC2*

Locus	Position[a] (bp)	Alleles	Description
s1	−12682	T, C	Promoter region, evolutionarily conserved
s2	−12479	G, A	Promoter region, evolutionarily conserved
s3	−6634	C, G	Promoter region
s4	−6280	G, A	Promoter region
s5	−3831	C, T	Promoter region
s6	−689	T, G	Promoter region
s7	−381	G, A	Promoter region
s8	1005	T, C	Intron
s9	4659	T, C	Intron
SL (s10)	6256	C, T	Exon, coding
s11	7241	A, G	Intron
s12	12100	ins GAT	Intron
s13	15691	A, T	Intron
s14	19577	C, T	Intron
AT (s15)	21363	G, A	Exon, coding
s16[b]	21383	ins G	Exon, coding
s17	22970	A, G	Exon, coding
s18	24991	G, A	Exon, coding
s19	25281	C, G	Exon, 3′ UTR
s20	25584	del 7	3′ flanking region
s21	26853	T, C	3′ flanking region, in KIAA0672
s22	27065	A, C	3′ flanking region, in KIAA0672
s23	48681	G, A and ins/del	3 alleles, 3′ flanking region, in KIAA0672
s24	69894	T, A	3′ flanking region, in KIAA0672
s25	80582	C, T	3′ flanking region, in KIAA0672

[a] Position is given in bp away from the first base of the first exon of *ELAC2*.

[b] The minor allele for *s16,* ins G, is not seen in this data set, but the locus is included for compatibility with an analysis, by N.J.C. (unpublished data), of the full data set from which the data here are drawn.

From the original set of genotypes, we reconstructed the haplotypes through use of the ApproxGCHap program, which is an implementation of an approximate gene counting or expectation-maximization algorithm method for this problem (Thomas 2003). Thus, although our original data were not complete, we circumvented the problem by estimating the missing observations. The complexity penalty α was set at $[\log(1{,}376)]/2 = 3.613$.

Several runs of simulated annealing with various cooling schedules resulted in the graph shown in figure 3 as our best guess at the global optimum. Some runs converged to other local maxima, indicating that improvements on mixing properties and the annealing schedule are needed. The estimated graph has a log likelihood of −5,364.3 with 115 df.

Although the graph shows some spatial effects, the structure of dependence is not dominated by chromosomal location. The loci *s5* and *s24* show association although they are physically distant, as do *s4* and *s20.* The loci *s16* and *s18* are independent of the others; on inspection, this lack of association turned out to be due to a lack of information: *s16* is totally uninformative, and the rarer allele for *s18* appears only eight times in the data set. These loci could be removed from this analysis, although the rarer alleles for each appear more often in the extended data set from which ours is drawn. There are also high-order interactions, such as the association between *s5, s12, s17,* and *s24,* which form one of many 4-cliques. There are no 5-cliques in the graph.

To assess the strengths of the estimated associations, we compared the optimal model with submodels obtained by removing single edges. Some care is needed here, since omitting some edges results in nontriangulated subgraphs. For example, disconnecting *s24* and *s23* leaves the unchorded 4-cycle {*s*11,*s*24,*s*25,*s*23}. We checked each edge of the graph in turn and, if its removal gave a triangulated graph, we compared the nested models with the usual χ^2 likelihood-ratio test. The results are given in table 2. Since the test statistics are so extreme, we give the statistic and degrees of freedom rather than the *P* values. The least significant link is that between *s13* and *s20,* which has a *P* value of .00016.

We also performed the converse operation of adding each edge, checking whether the resulting graph was

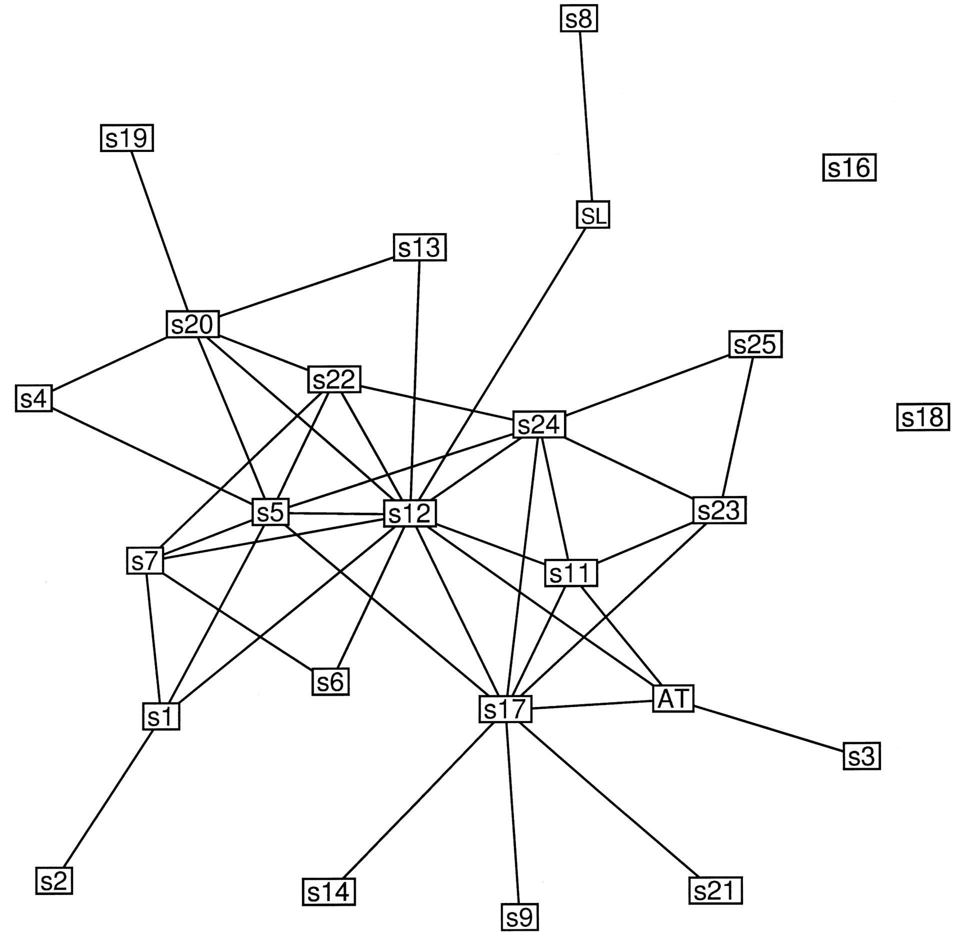

Figure 3 Estimated Markov graph from a study of 688 individuals genotyped at 25 loci in and around the *ELAC2* gene

Table 2

Test Statistics and Degrees of Freedom Submodels in which One Edge at a Time is Deleted from the Estimated Model

Edge	χ^2	df
s1, s2	1,110.5	1
s1, s5	320.3	4
s1, s7	132.0	4
s1, s12	396.4	4
s3, AT	379.3	1
s4, s5	905.0	2
s4, s20	755.6	2
s5, s17	482.0	4
s6, s7	1,196.2	2
s6, s12	20.8	2
s7, s22	173.2	4
s8, SL	1,630.5	1
s9, s17	1,085.0	1
SL, s12	1,628.2	1
s11, AT	216.4	4
s11, s23	490.6	8
s12, s13	546.5	2
s12, AT	382.1	4
s13, s20	17.4	2
s14, s17	1,059.3	1
AT, s17	220.9	4
s17, s21	1,072.5	1
s17, s23	689.3	8
s19, s20	1,772.0	1
s20, s22	452.6	4
s22, s24	130.8	4
s23, s25	283.7	4
s24, s25	603.0	3

NOTE.—Edges whose omission results in a nontriangulated subgraph are not tested.

triangulated, and assessing its significance. Table 3 gives the results for the most significant supermodels, as assessed using the χ^2 test once more. On the basis of this table, several edges could reasonably be added to the graph, so our information criterion appears to be somewhat parsimonious in allocating edges. However, note that the likelihood-ratio statistics for most of those edges included in the graph are orders of magnitude larger than those omitted.

Figures 4 and 5 show the Markov graphs estimated with the additional penalty for the square root of distances, in base pairs, between connected loci, as in equation (4) above. The parameter β was arbitrarily set to 1 and 2, following some experimentation to obtain values illustrating the range of models estimated. The models had log likelihoods and degrees of freedom, respectively, of $-6{,}765.3$ and 58 for $\beta = 1$ and of $-7{,}416.1$ and 47 for $\beta = 2$. For the case of $\beta = 2$, the optimal graph was simple enough that the perturbation scheme was able to reach this solution even with γ fixed at 0. That is, a straightforward random downhill search sufficed, and annealing was not necessary.

As might be expected, the results of using prior location information are simpler graphs. However, they are not subgraphs of the $\beta = 0$ solution in figure 3 but have some additional edges to replace those forced out by the distance penalty. Note, however, that there are still strong associations between distant loci. Even with $\beta = 2$, *s4* and *s22* are seen to be in disequilibrium (see fig. 5).

To compare our findings with those from a nongraphical method, we also analyzed the same data using the principal-components analysis method introduced by Horne and Camp (2004), to determine the groups of variants in strong linkage disequilibrium and the loci that effectively tag these groups. This method allows for noncontiguous and overlapping linkage disequilibrium groups, which is important when analyzing small genomic regions in which both mutation and recombination rates must be considered. Many other methods require contiguous and mutually exclusive groups, such as SNPtagger (Ke and Cardon 2003) or HaploBlockFinder (Zhang and Jin 2003), or allow for mutation but do not provide the linkage disequilibrium groups—just the tagging loci—such as tagSNPs (Stram et al. 2003) or Clayton's STATA utility used by Johnson et al. (2001). Principal-components analysis was performed on all haplotypes, with variants considered as constituent components, and those factors with eigenvalues >0.7 were extracted. The number of extracted factors represents the number of linkage disequilibrium groups. These underwent oblique rotation to remove

Table 3

Test Statistics, Degrees of Freedom, and *P* Values of Supermodels Created by Adding an Edge to the Estimated Model

Edge	χ^2	df	*P* Value
SL, s20	13.4	2	.0012
SL, s17	12.0	2	.0024
s2, s12	9.4	2	.0091
s4, s22	13.4	4	.0096
s18, s25	6.2	1	.0127
s5, s9	8.0	2	.0185
SL, s24	7.9	2	.0189
s2, s7	7.8	2	.0204
s21, s23	11.4	4	.0222
s8, s12	7.4	2	.0243
SL, s13	7.4	2	.0245
s18, s24	4.7	1	.0309
s18, s20	4.5	1	.0333
s18, s19	4.4	1	.0351
s17, s25	12.9	6	.0450

NOTE.—Edges whose addition results in a nontriangulated supergraph are not tested. Results are given for the supermodels whose *P* value is <.05.

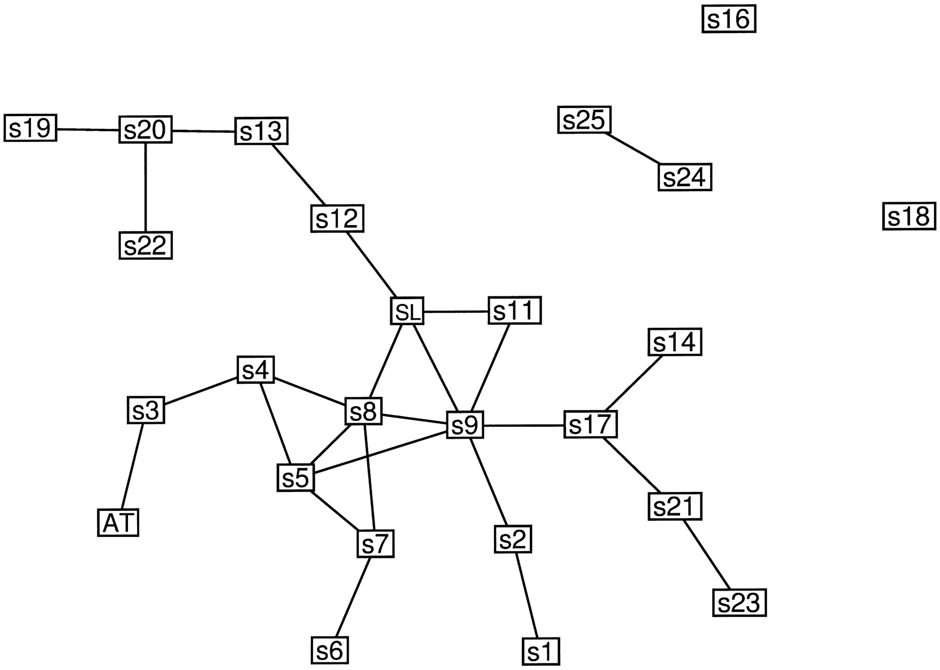

Figure 4 The graphical model estimated with an additional penalty for physical distance between connected loci. $\beta = 1.0$.

interfactor correlation, and variants with coefficients >0.4 in each factor were considered the members of the linkage disequilibrium groups (see Jolliffe [1986] and Stevens [1992] for a justification of these thresholding values). To determine tagging loci for each group, principal-components analysis was performed again for each linkage disequilibrium group by considering only the variants contained within each group separately. The highest-loading variant in each within group factor extracted was considered a tagging locus. For complete details of this method, see Horne and Camp (2004).

Seven factors with eigenvalues of at least 0.7 were extracted, representing seven linkage disequilibrium groups, which together explained 91.6% of the variance of all haplotypes observed. These linkage disequilibrium groups, ordered by eigenvalue, are given in table 4 and have also been indicated in figure 6. This shows a strong correspondence between the linkage disequilibrium groups identified by principal-components analysis and the links between loci in the estimated graphical model. The correspondence is the best for the graph generated without imposing priors for physical location. However, it remains strong even when the physical location priors are used (see figs. 4 and 5), although some links disappear. For figure 4, in which $\beta = 1$, *s22* is not connected to other variants in linkage disequilibrium group 3, and for figure 5, in which $\beta = 2$, two links have been lost, *AT* and *s3* are no longer connected, and *s5* is no longer connected to the other variants in linkage disequilibrium group 2.

In total, nine tagging loci were identified using principal-components analysis, which accounted for 81.3% of the total genetic variance: *s1, s6, s11, AT, s17, s18, s20, s22,* and *s24*. These also are indicated in figure 6. It should be noted that *s2* could be substituted for *s1, s25* for *s24,* and *s3* for *AT,* since these loci loaded equally on their respective factors. We assessed this nine-locus set for information content through use of the graphical model. The log likelihood of the model obtained by summing out the other variables was −3,716.51; hence, these loci contain 3,716.51/5,364.34 = 69.3% of the information in the whole set. This was fractionally above the 80th percentile as estimated from a sample of 100,000 randomly chosen nine-locus sets. The best set sampled was {*s2, s4, s5, s12, s19, s22, s23, s24, s25*}, which holds 80.7% of the information available from using all loci. Although at first glance it seems strange that all of *s22*–*s25* are chosen, table 1 shows that, between them, these four loci cover more than half of the genomic region covered by the whole set, so the result is reasonable.

Discussion

The primary challenges for this method are ones of scaling and efficiency. Although the data sets we have analyzed to date have been relatively small—the example above is

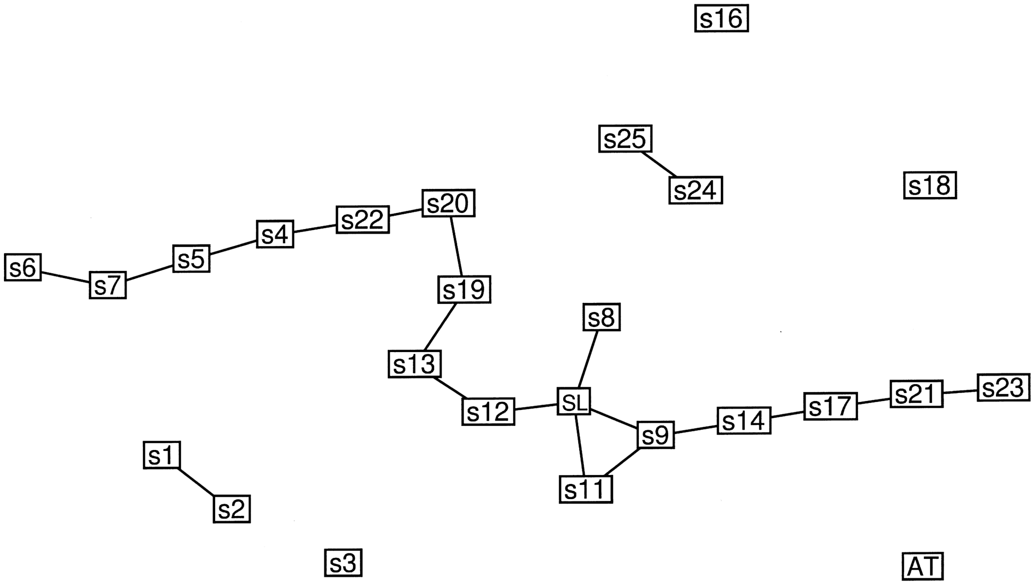

Figure 5 The graphical model estimated with an additional penalty for physical distance between connected loci. $\beta = 2.0$.

typical—the ultimate aim is to apply the method to substantial genomic regions with hundreds or thousands of loci and substantial numbers of genotyped individuals. The current program is quite efficient, and computation times scale linearly with the number of loci. The speed is approximately constant in the number of haplotypes, since the implementation avoids recalculating log likelihoods by storing the results. The cost of this is, of course, some increased storage space. The issue of efficiency to be addressed is not so much one of algorithmic efficiency as one of how the stochastic behavior of the simulated annealing search or Metropolis sampling changes as the state space grows.

The current perturbation scheme involves a limited set of changes, ranging from simply adding or deleting an edge to swapping the adjacency sets for a pair of vertices. The graphs produced by larger changes have a smaller probability of being accepted in the annealing scheme, but the rare acceptances are important. Other ways of improving mixing properties include Metropolis random restarts (George and Thompson, in press), using restarts proposed by sequential imputation (Kong et al. 1993), and simulated tempering (Geyer and Thompson 1995), under which the annealing parameter is randomly changed. With simulated tempering, all realizations produced can be used with the appropriate importance sampling weight, or, more simply, only those produced with annealing parameter 1 are used.

Several MCMC methods are readily amenable to parallelization. The precursor to simulated tempering was Metropolis-coupled MCMC, or MCMCMC, in which several chains are run simultaneously with different annealing parameters (Geyer 1991). Instances are then swapped between chains with Metropolis probabilities, and realizations are selected from the chain with annealing parameter 1 or from all chains with appropriate importance weights. This was originally conceived of as being run on a parallel processing machine; however, using it effectively requires many processors, each with temperature differing only slightly from its neighbors. In practice, therefore, it was usually implemented in a sequential fashion and, under such circumstances, simulated tempering is a better option. However, MCMCMC is perfectly suited to current massively parallel metaclusters. Each processor can be given a Markov chain to run and can do so almost independently of the other processors. The interprocessor communication needed is minimal: just the random swapping of states when the Metropolis probabilities require it. In

Table 4

Linkage Disequilibrium Blocks Estimated by Principal-Components Analysis

Linkage Disequilibrium Group	% of Variation	Cumulative % of Variation
s8, SL, s11, s12, s13, s19, s20, s22	38.7	38.7
s5, s9, s11, s14, s17, s21, s23	17.5	56.2
s4, s5, s6, s7, s22	12.9	69.3
s3, AT	7.9	77.2
s24, s25	5.6	82.8
s1, s2	4.8	87.6
s18	4.0	91.6

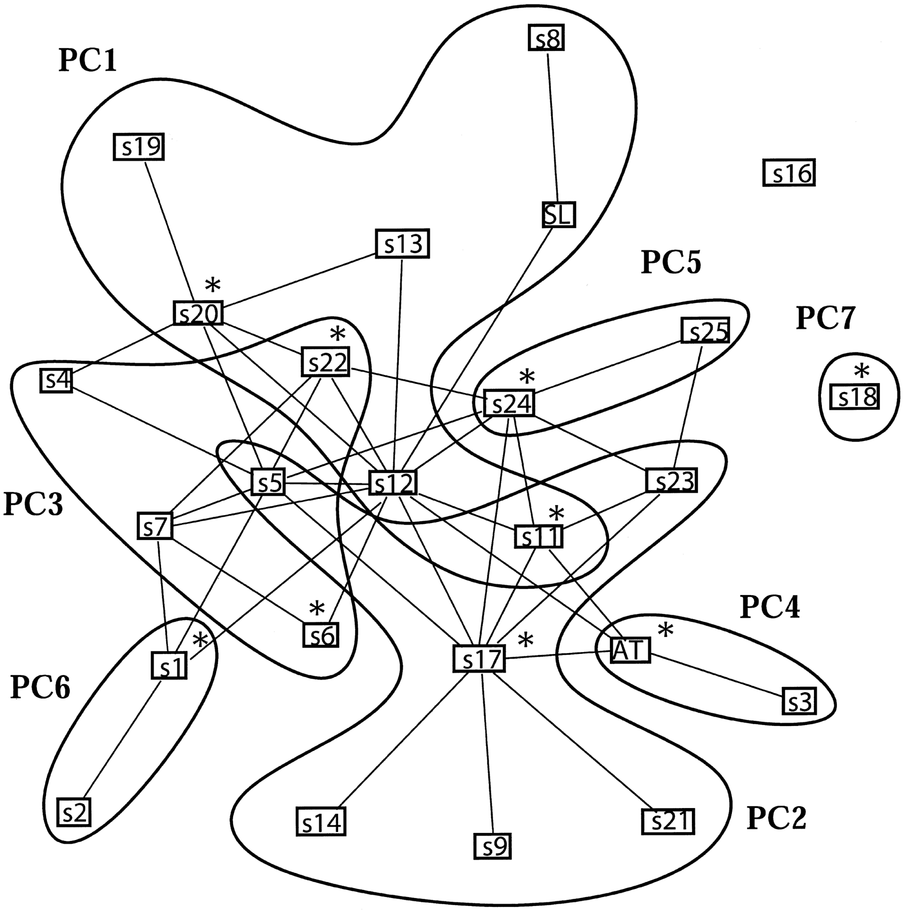

Figure 6 The estimated graphical model from figure 3, showing linkage disequilibrium groups and haplotype tagging loci estimated by principal-components analysis. Note that including *AT, s1,* and *s24* as tagging loci was somewhat arbitrary, since the analysis showed that *s3, s2,* and *s25,* respectively, could replace each of them equally well.

fact, even this can made faster simply by swapping the annealing parameters rather than the whole graphical state.

So far, we have essentially assumed that we can consider humans as a haploid species. Given the reported performance of haplotype reconstruction programs (Stephens et al. 2001; Niu et al. 2002), this is probably a reasonable assumption to apply when estimating associations between polymorphisms only and when the amount of missing data is low. However, when also applied to phenotypes and covariates, it is a very rough and ready approach. Simply classifying each haplotype by whether it belongs to an affected or an unaffected individual fails to account for the mode of inheritance of the disease and may give misleading results. We are currently extending the graphical model to include a layer corresponding to individuals and are also making extensions to deal with pedigree data to enable linkage analysis with linkage disequilibrium. This, in effect, would incorporate the sampling methods presented here with the MCMC method developed for linkage analysis by Thomas et al. (2000).

The reasons for choosing graphical modeling for analyzing associations between polymorphisms are numerous. The approach is theoretically rigorous. It requires no prior restrictions on the order of associations between loci: whether they are pairwise, three-way, or more complex is freely estimated in the procedure. It estimates empirically the independences and dependences between loci caused by the recombination process, the mutation process, and population history without modeling any of these. It can be used to select informative subsets of SNPs. Testing for association between phenotypes and loci fits naturally into the framework, and the tests developed will be statistically efficient. Finally, output and results can be easily visualized graphically and understood intuitively. All in all, we believe that these “Hap-Graphs” have considerably more potential to describe associated loci and facilitate analysis than do Hap-Maps.

Acknowledgments

This work was supported by National Institutes of Health (NIH) National Cancer Institute grants RO3 CA99844 (to N.J.C.) and R01 CA90752 (to L. A. Cannon Albright), a subcontract from Johns Hopkins University with funds provided by NIH National Cancer Institute grant R01 CA89600 (to L. A. Cannon Albright), and the Utah Cancer Registry supported by NIH Contract NO1-PC-67000, with additional support from the Utah Department of Health and the University of Utah. Genotyping was provided by Myriad Genetics. We would also like to thank two anonymous reviewers, whose comments led to the development of methods for incorporating prior information on locus positions and to a more detailed analysis of subgraphs of the estimated model.

Electronic-Database Information

The URLs for data presented herein are as follows:

A.T.'s Web site, http://bioinformatics.med.utah.edu/˜alun/ (for the Java program implementing the simulated annealing search procedure for fitting a graphical model to haplotype data, as well as an example data set, version information, documentation, and instructions for use; and for GCHap, which uses gene counting to estimate haplotype frequencies)

PHASE Web site, http://www.stat.washington.edu/stephens/phase.html (for M. Stephens's PHASE program)

SNPHAP Web site, http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt (for D. Clayton's SNPHAP program)

References

Camp NJ, Tavtigian SV (2002) Meta-analysis of associations of the Ser217Leu and Ala541Thr variants in ELAC2 (HPC2) and prostate cancer. Am J Hum Genet 71:1475–1478

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High resolution haplotype structure in the human genome. Nat Genet 29:229–232

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaar E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Krug A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I (2002) A first-generation linkage disequilibrium map of the human chromosome. Nature 418: 544–548

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

George AW, Thompson EA. Multipoint linkage analysis for disease mapping in extended pedigrees: a Markov chain Monte Carlo approach. Stat Sci (in press)

Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. In: Keramigas EM (ed) Computer science and statistics; 23rd Symposium on the Interface. Interface Foundation, Fairfax Station, VA, pp 156–163

Geyer CJ, Thompson EA (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. J Am Stat Assoc 90:909–920

Golumbic MC (1980) Algorithmic graph theory and perfect graphs. Academic Press, New York

Hampe J, Schreiber S, Krawszak M (2003) Entropy-based SNP selection for genetic association studies. Hum Genet 114: 36–43

Højsgaard S, Thiesson B (1995) BIFROST—block recursive models induced from relevant knowledge, observations, and statistical techniques. Comp Stat Data Anal 19:155–175

Horne BD, Camp NJ (2004) Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genet Epidemiol 26:11–21

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Genova GD, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Jolliffe IT (1986) Principal component analysis. Springer-Verlag, New York

Ke X, Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287–288

Kirkpatrick S, Gellatt CD Jr, Vecchi MP (1982) Optimization by simulated annealing. Technical report RC 9353, IBM, Yorktown Heights

Kong A, Cox N, Frigge M, Irwin M (1993) Sequential imputation and multipoint linkage analysis. Genet Epidemiol 10:483–488

Lauritzen SL, Spiegelhalter DJ (1988) Local computations with probabilities on graphical structures and their applications to expert systems. J R Stat Soc Ser B 50:157–224

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1091

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Shannon CE (1948) The mathematical theory of communication. Technical Report 27, Bell Systems Technical Journal

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stevens JP (1992) Applied multivariate statistics for the social sciences, 2nd ed. Erlbaum, Hillsdale, NJ

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC (2003) Choosing haplotype-tagging SNPs base on unphased data from a preliminary sample of unrelated subjects with and example from the Multi Cohort Study. Hum Hered 55:27–36

Tarjan RE, Yannakakis M (1984) Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM J Comput 13:566–579

Tavtigian SV, Simard J, Teng DHF, Abtin V, Baumgard M, Beck A, Camp NJ, et al (2001) A candidate prostate cancer susceptibility gene at chromosome 17p. Nat Genet 27:172–180

Thomas A (2003) GCHap: fast MLEs for haplotype frequencies by gene counting. Bioinformatics 19:2002–2003

Thomas A, Gutin A, Abkevich V, Bansal A (2000) Multipoint linkage analysis by blocked Gibbs sampling. Stat Comput 10:259–269

Thompson EA, Cannings C, Skolnick MH (1978) Ancestral inference I: the problem and the method. Ann Hum Genet 32:445–452

Zhang K, Jin L (2003) HaploBlockFinder: haplotype blocks analyses. Bioinformatics 19:1300–1301