Published in final edited form as:

Comput Stat Data Anal. 2009 March 15; 53(5): 1818–1828. doi:10.1016/j.csda.2008.02.003.

Estimation of graphical models whose conditional independence graphs are interval graphs and its application to modeling linkage disequilibrium

Alun Thomas¹
University of Utah

Abstract

Estimation of graphical models whose conditional independence graph comes from the general class of decomposable graphs is compared with estimation under the more restrictive assumption that the graphs are interval graphs. This restriction is shown to improve the mixing of the Markov chain Monte Carlo search to find an optimal model with little effect on the haplotype frequencies implied by the estimates. A further restriction requiring intervals to cover specified points is also considered and shown to be appropriate for modeling associations between alleles at genetic loci. As well as usefully describing the patterns of associations, these estimates can also be used to model population haplotype frequencies in statistical gene mapping methods such as linkage analysis and association studies.

Keywords

Decomposable graphs; Markov chain Monte Carlo; allelic association

1 Introduction

It is important in many type of statistical analyses in genetics to have accurate models for the frequency distributions of alleles or haplotypes at genetic loci of individuals in a population. This is true both for association testing involving unrelated cases and controls, and for linkage analysis where the founders of a pedigree are assumed to be randomly selected from the population at large. In each of these cases, using inappropriate distributions can lead to both the loss of power to detect the effects of genes on phenotypes, and to excessive false positive gene findings (Amos, Chen, Lee, Li, Kern, Lundsten, Batliwalla, Wener, Remmers, Kastner, Criswell, Seldina & Gregersen 2006). For an analysis using a single genetic locus, the requirement is simply a good estimate of the allele frequencies at that locus in the appropriate population. If a set of loci are used, and they are spaced sufficiently far apart, it is reasonable to assume that they are in *linkage equilibrium* in which case the probability of a haplotype is the product of the individual locus allele probabilities. However, if loci are densely spaced along the genome, strong correlations will exist between alleles at nearby loci. That is, the loci are in *linkage disequilibrium (LD)* (Ott 1985). Failure to incorporate LD in the modeled joint

¹Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA. alun@genepi.med.utah.edu, +1 801 587 9303 (voice), +1 801 581 6052 (fax).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

distribution of alleles will distort the frequencies of haplotypes and can again result in excessive false positives and loss of statistical power in linkage or association analyses. Other types of analysis, such as mapping based on segments of DNA shared between distant relatives (Thomas, Camp, Farnham, Allen-Brady & Cannon-Albright 2008) also need to model haplotype frequencies accurately.

Methods developed to estimate general multi locus LD models that can be used in genetic mapping include that of Stephens, Smith & Donnelly (2001), as implemented in the PHASE program and developed in the more recent FASTPHASE program (Scheet & Stephens 2006), and general graphical modeling implemented in the HapGraph program (Thomas & Camp 2004, Thomas 2005). Both these approaches, while computationally intensive, can be applied to hundreds of loci, and can be incorporated into linkage and association analyses (Thomas 2005, Thomas 2007). However, current genotyping assays involve over a million loci genome wide which makes the modeling problem intractable with current methods, even when the data is broken down to smaller components such as chromosome arms. In this paper, we consider the effects on computational tractability and model accuracy of modifications to graphical modeling methods which limit the class of graphs possible.

When the joint distribution of a set of random variables implies many independences or conditional independences between subsets of the variables, it can often be usefully considered as a graphical model. A graphical model has two elements: a *conditional independence*, or *Markov* graph, G , that represents the structure of the relationships between the variables, and a set of parameters, M . If the distribution of $X_1, \dots, X_n$ factorizes as

$$P(X_1, \dots, X_n) = \prod_i f(T_i) \text{ where } T_i \subset \{X_1, \dots, X_n\}, \quad (1)$$

the vertices of the Markov graph are the variables $X_1, \dots, X_n$ with edges connecting pairs of variables if they appear together in one or more of the T_i . While the structure of a graphical model is often apparent from modeling assumptions, it is also possible to estimate it from a set of multivariate observations. This was originally developed by Højsgaard & Thiesson (1995) with more recent work by Giudici & Green (1999) and Thomas & Camp (2004) on continuous and discrete variables respectively. In all this work models are restricted to the class of *decomposable* graphical models that are well behaved, tractable and flexible. This class is defined and the main features of estimation methods are described below. The Markov graph of a decomposable model is a *decomposable graph*. Both Giudici & Green (1999) and Thomas & Camp (2004) use stochastic search methods for finding an optimal model, or Markov chain Monte Carlo (*MCMC*) methods for sampling from the posterior distribution of models. In each of these cases it is necessary, given a decomposable graph G to propose a new graph G' and accept or reject it as the new incumbent according to appropriate probabilities. If G' is decomposable, then Giudici & Green (1999) have shown that the value of the target function for the proposed model can be found very quickly, in time independent of the size of the graph. However, it is not straightforward to ensure the decomposability of G' in advance so that it is necessary to check for this condition and reject graphs that are not decomposable. As we show below, this rejection step makes this general approach extremely inefficient when the number of variables is large. However, it can be avoided if we restrict the graphs considered to those in a more manageable subclass: the class of *interval graphs*.

A graph is an interval graph if its vertices can be made to correspond to sub intervals of the real line with pairs of vertices joined by an edge if and only if their corresponding intervals overlap. This is illustrated in figure 1 and described more fully below. As Golumbic (1980)

shows, all interval graphs are decomposable. If we now work with a set of intervals, one for each of the random variables in our model, it is easy to perturb these by moving and resizing them and yet be sure to stay within the class of interval graphs. If, furthermore, we find that the restriction to the set of interval graphs does not seriously affect our ability to accurately model the data, then we have a simpler and more computationally efficient estimation method.

Although this idea is developed generally to model associations between variables in any context, heuristically, it seems particularly appropriate for LD modeling. On average, according to Malecot's model, pairwise LD decreases as the distance between loci increases (Morton 2002), however, on a fine scale, more complex patterns appear. Thomas (2007) showed that, at least when dealing with small genomic regions, the haplotype frequencies estimated by both the PHASE program and graphical modeling are quite different to those obtained by fitting low order Markov models. Because of the linear arrangement of genetic loci along a chromosome, and the expectation that LD decreases with distance, modeling with interval graphs has clear intuitive appeal: most statistical geneticists have some informal notion of the *extent* of LD around a locus. In what follows, therefore, we consider not only the complete class of interval graphs which may have general applications, but also a more constrained subclass which will be appropriate when there is some reason to expect that a linear arrangement of the variables affects correlation.

2 Methods

2.1 Estimating graphical models

Consider a graph $G = (V, E)$ with vertices V and edges E . A subset of vertices $U \subseteq V$ defines an *induced subgraph* of G which contains all the vertices U and any edges in E that connect vertices in U . A subgraph induced by $U \subseteq V$ is *complete* if all pairs of vertices in U are connected in G . A *clique* is a complete subgraph that is maximal, that is, it is not a subgraph of any other complete subgraph.

A graph G is *decomposable* if and only if the set of cliques of G can be ordered as $(C_1, C_2, \dots, C_c)$ so that

$$\text{if } S_i = C_i \cap \bigcup_{j=i+1}^c C_j \text{ then } S_i \subseteq C_k \text{ for some } k > i. \quad (2)$$

This is called the *running intersection property*. This condition is equivalent to requiring that the graph is *triangulated*, or *chorded*, (Golumbic 1980), that is, it contains no un-chorded cycles of 4 or more vertices. The sets S_i are called the *separators* of the graph, and although several orderings typically give the running intersection property the cliques and separators are uniquely determined by the graph structure.

A graphical model with a decomposable Markov graph is a *decomposable model* and joint distribution of the variables in the model can be decomposed in terms of the marginal distributions of the cliques and separators:

$$P(X_1, \dots, X_n) = \prod_i \frac{P(C_i)}{P(S_i)}. \quad (3)$$

For discrete variables these marginals are simple multinomials, and so, given a set of observations, it is straightforward to calculate maximum likelihood estimators of the

parameters, the maximized likelihood, and the degrees of freedom. Multivariate Gaussians are similarly tractable in the continuous case. The decomposability then allows us to combine these to obtain the overall maximized log likelihood and degrees of freedom:

$$\log \widehat{L}(G) = \sum_i \log \widehat{L}(C_i) - \sum_i \log \widehat{L}(S_i), \quad (4)$$

and

$$\text{df}(G) = \sum_i \text{df}(C_i) - \sum_i \text{df}(S_i). \quad (5)$$

Model estimation can then be based on optimizing a penalized likelihood *information criterion*

$$IC(G) = \log \widehat{L}(G) - \alpha \text{df}(G), \quad (6)$$

where α is some arbitrary constant. Højsgaard & Thiesson (1995) use a deterministic optimization while Giudici & Green (1999) and Thomas & Camp (2004) use stochastic search or sampling methods. The stochastic methods require that an *incumbent* decomposable graph G is perturbed, for example by adding or deleting an edge, to give a proposed new graph G' . If G' is not decomposable it is immediately discarded, otherwise it is accepted or rejected with the appropriate probabilities for Metropolis (Metropolis, Rosenbluth, Rosenbluth & Teller 1953) or Hastings (Hastings 1970) sampling, or simulated annealing optimization (Kirkpatrick, Gellatt & Vecchi 1982). Giudici & Green (1999) give very fast methods for evaluating the rejection probability whose computational requirements do not increase with the number of variables being considered. Their algorithm for determining whether G' is decomposable can take order n time in the worst case, but in practice is very quick. However, for large graphs the probability that a random perturbation to G will result in decomposable G' is small. For instance if we consider adding or subtracting an edge there are $n(n-1)/2$ pairs of vertices to choose from, whereas, intuitively we would expect $O(n)$ of these flips to result in a decomposable proposal.

2.2 Interval graphs

A graph is an interval graph if its vertices can be made to correspond to intervals of the real line and its edges connect pairs of vertices if and only if the corresponding intervals overlap. This is illustrated in Figure 1. Intuitively, an interval graph would be expected to be long and thin, and this is indeed the case: this notion can be formalized in terms of the longest path in the graph and how far a vertex can be from this path (Golumbic 1980). Moreover, an interval graph is always decomposable. Thus, if we restrict our search for decomposable models to those with interval Markov graphs, we can work with the more tractable interval representations of the graphs instead of the graphs themselves. Whatever perturbations to the solution then involve, for example, moving an interval or changing its length or more complex manipulations involving multiple intervals, the result will always give an interval graph and a decomposable model. The benefits of this can be twofold. First, the perturbations can be more radical than simply adding or deleting an edge and so can potentially give better mixing properties for the sampler or optimizer. Second, we do not need to waste time proposing non-decomposable solutions.

It should be recognized, however, that we are sampling interval sets, not graphs directly. Since interval graphs can be represented as interval sets in different numbers of ways, those graphs with more interval set representations will be over sampled, and those with fewer will be under sampled. While this might be accounted for in the Metropolis or Hastings rejection probability, we will assume that this effect is small when we are sampling graphs of similar probability, and justify this empirically below.

2.3 Efficient implementation

In order to take advantage of this idea, we need two things. One is to have a data structure that allows interval sets to be managed and queried efficiently. The other is to be able to evaluate the maximized log likelihood and degrees of freedom of a proposed model quickly, and preferably in time that does not depend on the size of the problem.

The first issue is resolved by using a standard data structure called an *interval tree* (de Berg, van Kreveld, Overmars & Schwarzkopf 2000). The root of the tree is associated with a fixed point, typically the mid point of a finite region that contains all the intervals. This root node stores a list of the intervals that cover the fixed point. All intervals that lie completely to the left of the point are delegated to a daughter node whose fixed point is the mid point of the left region, and similarly for intervals who lie completely to the right of the fixed point. The structure is built up recursively in this way until all intervals are stored in a list at one of the nodes in the tree. This structure allows addition of new intervals, deletion of existing intervals, querying for intervals that cover a particular point, and querying for intervals that overlap with a given interval to be done in $O(\log n)$ time.

To address the second issue of efficient likelihood recalculation, we first note that the set of intervals that cover any point on the line correspond to a *complete cutset* of the graph (Golumbic 1980). A set of vertices K is a cutset if it partitions the vertices of G into L , M and K itself such that all paths in G from a vertex in L to a vertex in M must pass through a vertex in K . The separators S_i of G are all complete cutsets, in fact, all the minimal complete cutsets. The complete cutsets defined by points on the line will include these separators and also complete cutsets that are not minimal. For any graphical model, if K is a complete cutset then the variables L are conditionally independent of M given the value of K . That is

$$P(KLM) = P(L|K)P(M|K)P(K) = \frac{P(LK)P(MK)}{P(K)}. \quad (7)$$

If we now consider a sub region (x, y) of the line we can define three induced subgraphs of G : A , B and D the subgraphs induced by the intervals that overlap with $(-\infty, x)$, (y, ∞) and (x, y) respectively, so that $A \cap D$ and $B \cap D$ will be the complete cutsets defined by the intervals that cover the points x and y . This is illustrated in Figure 2. The sub region (x, y) thus defines conditional independences that can be expressed as

$$P(ABD) = \frac{P(A)P(B)P(D)}{P(A \cap D)P(B \cap D)}. \quad (8)$$

If we now alter the graph G to make G' in such a way that only intervals that lie completely in (x, y) are changed, D may change to D' but A and B will not be affected. Moreover, $A \cap D' = A \cap D$ and $B \cap D' = B \cap D$. Hence,

$$\frac{P(G')}{P(G)} = \frac{P(A)P(B)P(D')}{P(A \cap D')P(B \cap D')} \times \frac{P(A \cap D)P(B \cap D)}{P(A)P(B)P(D)} = \frac{P(D')}{P(D)}. \quad (9)$$

In this way, the change in the global joint probability can be evaluated very quickly from local changes.

As with equation 3, this extends to allow us to quickly evaluate changes in the maximized log likelihood and degrees of freedom, and hence the information criterion $IC(G')$. So, for perturbations of G that involve changing just one interval, we need only consider the graph corresponding to the portions of the line that lie under the interval before it is changed and after it is changed. Hence, we can very efficiently evaluate the target function for the proposed graph G' .

In our implementation of this scheme, intervals are initially allocated with midpoints evenly distributed between 0 and 1, with small lengths so that no intervals overlap. Perturbations consist of randomly choosing an interval and either giving it a new midpoint chosen uniformly at random in $(0,1)$, or giving it a new length chosen from an exponential distribution, or both.

2.4 Constrained interval graphs

When the variables being modeled can be ordered linearly it may be appropriate to reflect this in the structure of the interval graph. For example, genetic loci have physical positions along a chromosome and we strongly expect the greatest correlations to be between alleles at loci that are nearest each other. In this case we can require the interval that represents a particular locus to cover its physical location. We also alter the definition of the graph to require intervals to overlap by some minimal amount in order to add an edge between the corresponding vertices. Any vertex corresponding to an interval of length less than this minimal amount will therefore not be connected to any other vertices. This is illustrated in Figure 3. This extra condition gives some flexibility to the model. For example, with reference to Figure 3, suppose that locus 2 appears from the data to be independent of all other loci, but that loci 1 and 3, and 3 and 4 are very strongly correlated. Without this final requirement, the interval structure would force an edge between 2 and 3 making the model more complex than necessary. Such a situation may often arise with genetic loci where the frequency of the less frequent allele is very low. It is trivial to show that requiring a minimal overlap still gives an interval graph.

In this case the intervals are initially set as for the general interval graph. Perturbations involve randomly extending or reducing the spans to each side of the required fixed point by amounts generated from an exponential distribution.

This approach can also be used if an ordering of the variables is known but distances are not. In this case we can assign the variables to arbitrary evenly spaced points along the line.

2.5 Programs

General and constrained interval graph searches have been incorporated into the author's HapGraph program (Thomas & Camp 2004) which can be used both as a generic graphical model estimator, or for the specific case of modeling allelic association. This latter case requires an extra step to account for observing unordered genotypes as opposed to complete phase known haplotypes. Both versions allow for missing data by random imputation. Full details of the methods are given by Thomas (2005). The program is written completely in Java thus is platform independent, and can be obtained from <http://bioinformatics.med.utah.edu/~alun>.

3 Results

We illustrate the effects of the model restrictions described here using data for subsets of the single nucleotide polymorphisms on chromosome 1 genotyped in the sample of Yoruba people from Ibadan, Nigeria by the HapMap project (The International HapMap Consortium 2005). This sample is conventionally abbreviated as YRI, and the data was from build 36 dated 2 May 2007. The loci that were monomorphic in this sample were not considered in these analyses. We used subsets of the first 20,000 remaining loci in what follows.

In order to first consider the computational effects of model restrictions we ran three versions of the HapGraph program. The first fitted a general decomposable graph using the rejection method of Giudici & Green (1999), which is the standard form of the program. The other two implemented a general interval graph and a constrained interval graph search as described above. HapGraph's graphical user interface that shows the graph as it is being updated was not used so as to avoid incorporating the processor time needed for graphical rendering in the comparisons. Figure 4 shows the times taken by each of the 3 methods to perform one million Metropolis updates of the graph for data on sets of between 20 and 20,000 loci. Figure 5 plots the largest penalized log likelihood score seen in each of the runs. All the programs were run on the author's laptop computer which has a 2.33 GHz dual core central processing unit running Red Hat Linux and Java version 1.5.

For the decomposable graph search we recorded both the number of random proposals that resulted in a decomposable graph, and the number of these proposals that were accepted based on the usual Metropolis probabilities. For both types of interval graph searches we recorded the number of proposed new interval configurations that were accepted and also the number of these that resulted in a different implied graph. These counts are shown in Figure 6. For all the versions of the program, the starting configuration used was the trivial graph, that is the graph with a vertex for each locus but no edges. Thus, in the early stages of each search the graph was very sparse and almost all randomly chosen pairs of vertices could legitimately be connected to give a decomposable model. Also, in the early stages almost any change would tend to be accepted. Therefore, in order to check the performance of the methods closer to the equilibrium state we also recorded these counts in the last 100,000 (10%) of iterations. These are also shown in Figure 6.

We then compared the haplotype frequencies implied by models optimized for each of the three classes of graph using simulated annealing. To avoid comparing very small frequencies we considered only haplotypes for the first 20 polymorphic loci on chromosome 1. Figure 7 gives pairwise scatter plots of the frequencies estimated under the general decomposable models against those seen for general and constrained interval graphs. As an external reference we also show the haplotype frequencies estimated using the FASTPHASE program, those estimated with no accommodation for LD, that is, under the assumption of linkage equilibrium, and those estimated under the assumption that dependence is limited to a first order Markov chain and a fifth order Markov chain.

4 Discussion

In absolute terms, as shown in Figure 4, the computational performances for the 3 methods are similar. In the long run, the time required is quadratic although for up to about 10,000 variables performance is very close to linear. This difference is probably due to the increasing amounts of work done by the Java garbage collector to reclaim heap space. Even for the substantial numbers of loci used here, none of the methods takes prohibitive time or storage.

Although for below around 15,000 loci each of the interval graph methods takes more absolute time than the decomposable graph method, the amount of work done is substantially more as

shown by Figure 6. Figures 6(b) and (c) show that around 25% of updates for the interval graph methods are accepted, of which 5% to 10% give rise to new graphs. For the decomposable graph method the percentage of proposals accepted decreases rapidly, see Figure 6(a). The difference is far more marked in the last 100,000 iterations when the effect of initial conditions is minimized. Of the last 100,000 times that a random pair of 20,000 loci were selected, in only 70 cases could the pair be either disconnected, if they were previously connected, or connected, if they were previously disconnected, so that the resulting graph was decomposable: clearly the rejection method becomes very inefficient, see Figure 6(d). On the other hand for constrained interval graphs, the acceptance rate settles down very quickly at about 25%, and the accepted interval configurations that give a new graph settles at about 5%, Figure 6(f).

The acceptance rate for general interval graphs actually increases with the number of loci, even for the last 100,000 iterations. However, this is likely to be due to long term residual effects of initial conditions: in effect, for general interval graphs on large numbers of vertices the Markov chain is not mixing well. This poor mixing is also reflected in Figure 5. Since constrained interval graphs are a subset of general interval graphs which are a subset of decomposable graphs, the true optimal values of the penalized log likelihood scores must increase through that sequence of inclusion. However, the maxima actually found reverse that order showing that the smaller space of constrained interval graphs is far more efficiently searched than its supersets.

The statistical effects of model subclassing are shown in Figure 7. For this example the differences between haplotype frequencies estimated from models in each of the three classes of graphs are very similar, see Figure 7(a) and (b). The results from FASTPHASE are also similar, Figure 7(c). However, frequencies under linkage equilibrium or simple Markov dependence, even up to fifth order, show marked differences with far more points along or close to the axes of Figures 7(d), (e) and (f). The distribution of distances between the 20 loci used here is quite skewed, with a mean of 37.97 kilo bases but median of only 2.33 kilo bases. Thus, haplotype frequencies derived from general decomposable, general interval, and constrained interval models are similar to each other and also similar to those derived from FASTPHASE. In contrast, ignoring LD completely or modeling it with small lag Markov models gives misleading results even though they may require more parameters than interval graphs.

Overall, therefore, there are clear computational benefits and little costs in terms of model flexibility to using interval graphs. In particular, in the context of LD modeling, constrained interval graphs have considerable practical advantages. As a final comment, note that the localization of the interactions implied in the constrained interval graph method means that loci sufficiently far apart can be considered separately. Thus, although not exploited by the programs described here, this would allow a moving window implementation that scales linearly with the number of loci and be feasible on a genome wide level.

Acknowledgments

This work was supported by grants number R01 GM070710 and R01 GM081417 to Alun Thomas from the National Institutes of General Medical Sciences. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of General Medical Sciences or the National Institutes of Health.

References

- Amos CI, Chen WV, Lee A, Li W, Kern M, Lundsten R, Batliwalla F, Wener M, Remmers E, Kastner DA, Criswell LA, Seldina MF, Gregersen PK. High density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Genetic and Immunity* 2006;7:277–286.

- de Berg, M.; van Kreveld, M.; Overmars, M.; Schwarzkopf, O. Computational Geometry. Algorithms and Applications. 2. Springer-Verlag; 2000.
- Giudici P, Green PJ. Decomposable graphical Gaussian model determination. *Biometrika* 1999;86:785–801.
- Golumbic, MC. Algorithmic Graph Theory and Perfect Graphs. Academic Press; 1980.
- Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;57(1):97–109.
- Højsgaard S, Thiesson B. BIFROST — Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques. *Computational Statistics and Data Analysis* 1995;19:155–175.
- Kirkpatrick, S.; Gellatt, CD., Jr; Vecchi, MP. Optimization by simulated annealing, Technical Report RC 9353. IBM; Yorktown Heights: 1982.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. Equations of state calculations by fast computing machines. *Journal of Chemistry and Physics* 1953;21:1087–1091.
- Morton, NE. Applications and extensions of Malecot's work in human genetics. In: Slatkin, M.; Veuille, M., editors. *Modern developments in theoretical population genetics*. Oxford University Press; Oxford: 2002. p. 20-36.
- Ott, J. Analysis of Human Genetic Linkage. The Johns Hopkins University Press; Baltimore: 1985.
- Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 2006;78:629–644. [PubMed: 16532393]
- Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 2001;68:978–989. [PubMed: 11254454]
- The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299–1320. [PubMed: 16255080]
- Thomas A. Characterizing allelic associations from unphased diploid data by graphical modeling. *Genetic Epidemiology* 2005;29:23–35. [PubMed: 15838847]
- Thomas A. Towards linkage analysis with markers in linkage disequilibrium. *Human Heredity* 2007;64:16–26. [PubMed: 17483593]
- Thomas A, Camp NJ. Graphical modeling of the joint distribution of alleles at associated loci. *American Journal of Human Genetics* 2004;74:1088–1101. [PubMed: 15114533]
- Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Annals of Human Genetics*. 2008 In press.

Intervals

Graph

Fig. 1.
An interval set and its corresponding interval graph.

Fig. 2.
A sub region partitions the interval graph.

Fig. 3.
An interval graph constrained by the physical location of genetic loci.

Fig. 4. The computer times required for one million MCMC iterations by number of genetic loci when the search is over general decomposable graphs, general interval graphs and constrained interval graphs.

Largest penalized log likelihood sampled

Fig. 5. The largest penalized log likelihood score seen in a sample of 1,000,000 MCMC simulations by number of loci.

Fig. 6. The numbers of accepted proposals in all 1,000,000 MCMC simulations and in the final 100,000 simulations under the three classes of graphs considered by number of loci, shown as percentages.

Fig. 7.

Haplotype frequencies for the first 20 loci estimated for the YRI data for an optimized model with general decomposable graph compared with models with general interval and constrained interval graphs. Also compared are haplotype frequencies estimated using FASTPHASE, and those estimated under linkage equilibrium, and under first and fifth order Markov dependence.