15

Applications of Graphical Clustering Algorithms in Genome Wide Association Mapping

K.J. Abraham[1*] and Rohan Fernando[2]
[1]*Programa de Pós Graduação em Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo*
[2]*Dept. of Animal Science, Iowa State Univ*
[1]*Ribeirão Preto SP, Brazil*
[2]*Ames IA USA*

1. Introduction

The field of statistical genetics has been the area of a great deal of active research in recent years; due in part to dramatic advances in sequencing technology which has led to vast amounts of genomic data becoming available at lower and lower costs. The data from these sequencing efforts is not only copious, but is also characterized by significant levels of experimental noise. Processing data of this nature to draw statistically significant inferences requires dealing with a number of challenges, both statistical and algorithmic. In this chapter we will not discuss the very substantial statistical issues which arise in extracting genome sequences, but rather will focus on computational and algorithmic issues which arise in drawing biological inferences once the sequence is known. Event though our discussion will be oriented more towards applications of graph theory, it is worth keeping in mind that statistical considerations will still play a role due the intrinsically probabilistic nature of Mendelian Genetics as well as finite sample sizes. We will first begin by reviewing earlier well known work which will serve to illustrate the utility of graph theoretical concepts in dealing with genomic data. The data at our disposal are assumed to consist of observations of at a large number of locations (tens of thousands or possibly much more) on multiple chromosomes for a collections of individuals, or plants or animals; for now we assume that these individuals are related with known parent offpsring information. We restrict our attention to species which have just two chromosomes, but much of what we will discuss can be generalized to species with more than two chromosomes although the computational implementation could be challenging. At any locus (precise location on a specified chromosome) we assume that there are two or more possible alleles in the population, the precise number is assumed to differ from locus to locus. The number of possible observable genotypes at each locus will thus also vary from locus to locus. In a population of related individuals with known parent offspring relations between individuals (*i.e.* pedigree) it is possible to predict the probability for an offspring to receive a given allele from a parent based on Mendelian Genetics. If we represent a given locus for a given individual by a vertex, we can assign multiple possible states to each vertex depending on how many genotypes are possible. Since genetic information flows from parents to offspring

*Bolsista CAPES/Brasil

we can construct a directed graph where the arrows flow from parental vertices to offspring vertices. The indegree of each vertex is maximally two, while the outdegree will depend on how many offspring an individual has. We thus have a directed graph where each vertex has multiple states (genotypes or alleles) associated with it. Since individuals cannot be their own ancestors, the graph is acyclic as well as being directed. With each edge we can associate a transmission probability which is determined by Mendelian Genetics; in addition there is a Markov Field Property involved as conditional on parental genotypes offspring genotypes are independent of all other ancestral genotypes and sib genotypes. We thus have all the ingredients of a Bayesian Network. Many important problems relating to genetic inference from data on pedigrees have been formulated in the language of Bayesian Networks (Fishelson & Geiger, 2002); (Fishelson & Geiger, 2004); finding exact and approximate solutions to these problems particulary on large data sets has led to the development of very sophisticated algorithms which we will not discuss here. The key feature underlying the data is that it consists of a potentially large but discrete number of observations. These observations have a very complex correlational structure, some of the observations are heavily correlated (*eg.* the genotypes of parents and offspring at the same locus) while others may be very loosely correlated (*eg.* the genotypes of individuals and ancestors going back several generations). The discrete nature of the data points permits us to assign a vertex in a graph to each data point while the edge structure (which arises from the pedigree structure) is a reflection of the association between data points; seen in this light the use of a graph theoretical formulation is quite natural. In what follows, the data under consideration will have the same features, suggesting the use of graphical models but our discussion will focus more on the use of undirected graphical models.

In order to motivate the application of undirected graphical models, we consider two fixed loci on the same chromosome in a population of individuals. It is frequently observed that the joint distribution of alleles at loci which are in close physical proximity on the same chromosome, is not uniform. More specifically, if there are two alleles at one locus A or a and B or b at another locus, then the probability of finding allele B in some arbitrary individual in a population may depend on the which allele (A or a) is present at the other locus. In the language of probabilities $\mathcal{P}(A,B) \neq \mathcal{P}(A)\mathcal{P}(B)$ for certain pairs of loci; this is the phenomenon of linkage disequilibrium (LD) (Weir, 1996). The extent and statistical significance of the non-randomness of the alleleic association can be quantified by analyzing a (3×3) contingency table whose entries are genotype counts. If there are just two of the three possible genotypes present, or if the population from which the unrelated individuals are sampled is subject to certain other constraints, it becomes possible to estimate the linkage disequilibrium between each pair of distinct markers using just a (2×2) contingency table, alternatively the non-randomness in the asociation between the alleles depends one just one function of the allele counts. What is actually computed is just the sample LD, the standard errors on the LD will depend inversely on the size of the population. The magnitude of the observed LD can vary quite dramatically depending on which loci are being compared, large LD is very much more common among loci close together than loci on different chromosomes. Furthermore, while LD tends to decrease as the distance between loci increases, the decrease is often neither uniform nor monotonic. This discussion can be extended to multiple loci by considering larger contingency tables, more sophisticated multivariate discrete probability distributions and also multiple coefficients of association. Our data once again consists of a large number of discretized observations with a possibly complex correlation structure between the observations; suggesting the use of a graph theoretical formulation. If we

represent each locus by a vertex, the LD structure can be captured graphically by introducing an undirected edge between a pair of vertices whenever the LD between the vertices in the pair is statistically significantly different from zero. It is assumed that there is a user defined significance threshold. The edge is undirected because the statment of statistical association between loci relates only to correlation, and does not carry any implications of causality. This defines an undirected graph whose edge structure is indicative of the LD patterns between the loci under consideration. Unlike in (Fishelson & Geiger, 2002); (Fishelson & Geiger, 2004); we do not associate states with a vertex in a graph, all that information has been averaged over all indviduals in determining the LD between markers. The use of graphical models to elucidate the LD structure between loci is well established (Thomas & Camp, 2004) as is the connection between graphical models and discrete multivariate probability distributions (Lauritzen, 1996). To summarize, datasets in statistical genetics consist of a vast collection of discretized observations with potentially very complex correlations between the observations; graph theoretical methods can be adopted for describing and analyzing data of this nature. In the rest of this chapter, we will discuss some applications of this nature, some open problems and possible solutions.

2. Graphical methods in association mapping

2.1 Population stratification

Understanding the LD structure is of considerable interest not only from the viewpoint of population genetics, but is vital for deducing the location of genes influencing traits in populations by Genetic Association Mapping. The Case Control design is a popular design for Genetic Association Mapping (Thomas, 2004). Here the data are assumed to consist of a large number of genotypes at fixed loci observed on two distinct groups of individuals, healthy controls and diseased cases, and we assume that the individuals are unrelated to each other. We assume that the genotypes are observed at marker loci, *i.e.* locations on the chromosome where there are no genes directly influencing the disease. Nonethless, if genotypes are observed at a large number of sufficiently closely spaced markers , there may be some markers physically close to the gene influencing the disease and which are potentially in strong LD with the diesease gene leading to a statisticaly significant association between certain genotypes and disease status. The goal of Case Control studies is to discover which markers (if any) show statistically significant associations with disease status and then draw conclusions about the physical location of a gene causing the disease with relation to these markers. As the association is statistical in nature it is important to understand potential causes of false positives in order to minimize Type I error. Two very important causes of Type I error in Case Control studies are population stratification and multiple testing artifacts. As we will see, undirected graphical models can be used to acquire new insights on both these problems. We begin with an analysis of population stratification; population stratification can be understood in terms of a difference in genetic content between cases and controls over and above any differences at loci in high LD with the disease gene. For example, if all the diseased cases are from one ethnic group, and all the healthy controls are from another ethnic group, then there will be statistically significant associations between case/control status and genotypes at loci which reflect differences in ethnicity *i.e.* population structure , in addition to loci which are possibly liked to the disease. This is an example of population stratification, and will lead to false positive associations at loci reflecting the ethnic differences between cases and controls but unrelated to the disease under study. In this very simple

instance we just considered two ethnic groups, with all the cases drawn from one group and all the controls from the other, in more typical instances, both the cases and controls will themselves be mixtures of two or more ethnic groups. In these more realistic scenarios population stratification will be a problem when the proportions of various distinct groups are different in cases and controls and when genotypes are observed at loci where the frequencies of the various genotypes are different in the various ethnic groups represented in the case and control samples. The effects of population stratification can be ameliorated by carefully matching ethnicities between cases and controls but this is not always possible or feasible. In real life situations, it is safer to assume that population stratification exists, which then must be taken into account before testing any markers for association with disease status. There are two broad approaches to correcting for population stratification in genetic association studies, non-parametric and parametric. The most popular parametric method for dealing with population structure is using the program *Structure* (Pritchard, 2000) ;(Falush, 2003) in which specific scenarios for population admixture are assumed. *Structure* attempts to assign the individuals to specific clusters based on a specific model, the genotypes are the feature vectors used to decide how to assign individuals to clusters. In addition to the genotype data the number of populations present in the data must be supplied by the user, this is analogous to specifying the number of clusters in k-means or other clustering methods. In the most sophisticated scenario (the Linkage Model) the genotypes for any individual reflect a mixture of different populations with different chromosomal segments possibly arising from different populations. The number of populations however is not specified and must be supplied by the user. The precise assignment of individuals to different populations frequently arises only after a long MCMC simulation which uses genotypes for all individuals at all loci as input. Any LD between the loci is corrected for in the process of assigning individuals to constituent sub-populations. Since it is not uncommon to have genotypes at tens of thousands (frequently more) of loci implementing the methodology of *Structure* can be time consuming partly due to the sheer size of the data set and partly due the overhead of correcting for LD between the loci which is typically present when there are a large number of loci under consideration. The presence of LD between the loci also suggests that even though the number of loci may be large, the various loci do not necessarily contribute additional independent information on population stratification. This suggests that with a judicious choice of mutually independent loci, population stratification can be analyzed with a smaller and more manageable subset of the data in less CPU time. We will next explain how exactly this can be done using graph theoretical ideas and mention an application to real data.

An optimal set of loci for discerning population structure should be sufficiently large so that loci characteristic of populations whose frequency in the sample is relatively small, are nonetheless included, while ensuring a high degree of statistical independence between the loci. The requirement of statistical independence between the loci can be made more precise by ensuring that the loci are in low LD with one another. What we are then looking for is the largest possible subset of loci such that the LD between any arbitrary pair of loci is low. We will recast the problem in the language of graph theory using the correspondence between vertices in a graph and loci on a chromosome we discussed earlier and show a correspondence between a well known combinatorial optimization problem, that of finding the maximum independent set on an undirected graph. As there is no known polynomial time solution for this problem, a randomized heuristic algorithm will be described along with its performance on a real data set.

The input to the algorithm is $\mathcal{N}$, a set of N markers, an $N \times N$ symmetric matrix $\mathcal{M}$ with positive off diagonal values, and a positive constant c. The precise value of the diagonal matrix elements of $\mathcal{M}$ are not relevant as long as they are smaller than c . If we denote the elements of $\mathcal{N}$ by N_j $(1 \leq j \leq N)$ then each row of $\mathcal{M}$ corresponds to a unique marker; with this assignment $\mathcal{M}_{ij}(i \neq j)$ is simply the magnitude of the association between markers N_i and N_j. All the different M_{ij} can be easily computed given a rectangular data matrix of genotypes in which individuals are indexed by rows and each column contains the genotypes for one particular marker. As mentioned earlier we assign to each marker a vertex in an undirected graph $\mathcal{G}$. Thus $\mathcal{G}$ has a set of vertices $\mathcal{V}$ with N elements denoted by $V_i \in \mathcal{V}$, where $1 \leq i \leq N$. Since each marker is assigned to a unique row of $\mathcal{M}$, we can now uniquely associate to each row of $\mathcal{M}$ a vertex $V_i \in \mathcal{V}$, where $1 \leq i \leq N$. Let the set of edges of $\mathcal{G}$ be denoted by $\mathcal{E}$. An undirected edge E_{ij} exists between any two elements V_i and V_j of $\mathcal{V}$ if $\mathcal{M}_{ij} > c$. This condition is adequate to define all the elements of $\mathcal{E}$. There can clearly be no edges from any vertex to itself due to the choice of the diagonal matrix elements of $\mathcal{M}$. By a suitable choice of c, any two unlinked vertices in $\mathcal{G}$ can be made to correspond to two unassociated markers in $\mathcal{N}$. The precise value of c needed to achieve this correspondence will depend on some user specified threshold for defining significant association. Thus given a subset of vertices $\{V_i, V_j, V_k, V_m\}$ with no edges connecting any of the six possible pairs of vertices that can be formed from this subset, we can find a corresponding subset of markers $\{N_i, N_j, N_k, N_m\}$ which are mutually unassociated. This argument can be extended to any $\mathcal{V}_s \subset \mathcal{V}$ which gives rise to a corresponding $\mathcal{N}_s \subset \mathcal{N}$ of mutually unassociated markers. Furthermore, each unique $\mathcal{N}_s \subset \mathcal{N}$ corresponds to a unique $\mathcal{V}_s \subset \mathcal{V}$. However, any $\mathcal{V}_s$ corresponds to a clique on $\mathcal{G}^c$, the complement graph of $\mathcal{G}$. If we want the largest possible $\mathcal{N}_s \subset \mathcal{N}$ of mutually unassociated markers we must find the maximum independent set of vertices in $\mathcal{G}$.

We have thus transformed the problem of finding the largest possible set of mutually unassociated markers to a well known problem in graph theory that of finding the maximum independent subset of vertices in an undirected graph, (or equivalently the clique of largest size on the complement graph) a problem for which there is no known efficient solution. Thus we are forced to resort to heuristics which yield only approximate solutions, more precisely a subset of vertices which may be smaller in size than the true maximum independent subset. As a cross-check on any given solution it would be useful to have a different solution for the sake of comparision. This motivates the use of a stochastic greedy heuristic which can generate multiple solutions, rather than use of one of the many well known published heuristic algorithms for this problem. The graph that we have is unweighted, although we could have considered a weighted graph with the LD between markers playing the role of weights. Although the precise LD information has been ignored in the construction of the graph and in our heuristic algorithm, this does not matter. LD represents a statistical correlation and all that matters for our purposes is whether the correlation is significant or not. One complication we have ignored here is that the presence or absence of edges is determined by comparing sample LD values with some threshold; in a more sophisticated scheme the standard errors on the sample LD values could also be used to assign probabilities for the presence of edges in the graph where the corresponding sample LD values are close to the significance threshold.

We will next describe a stochastic greedy heuristic for finding the clique of maximum size on an undirected graph, which due to the exact corespondence between maximum cliques and maximum independent sets can readily be applied to our maximum independent set problem.

• Description of Algorithm

As before we assume we have an undirected graph $\mathcal{G}$ in which $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges. $\mathcal{G}$ is not assumed related to any genetic marker map so the algorithm is at this stage perfectly general. The algorithm also requires as input a positive parameter γ which is a measure of how far the algorithm deviates from a deterministic greedy heuristic. The larger the value of γ the closer the algorithm resembles a deterministic heuristic. We define L_i the set of neighbors of V_i and n_i size of L_i, and assume $n_i > 0 \; \forall \, V_i \in \mathcal{V}$. We define sets of vertices *CandSet*, *TempSet* as well as *ReturnSet* which is the output from the program.

Informally, the algorithms starts by picking a seed vertex which has a relatively large number of neighbors, relatively large being defined with respect to the number of neighbors of all the vertices in the graph. This seed vertex V_s is inserted into *ReturnSet*. *CandSet* is initialized by the set of neighbors V_s, while *TempSet* is initialized by the empty set. An element V_n of *CandSet* is chosen on the basis of having a relatively large number of neighbors and *TempSet* is the set of neighbors of V_n not already included in *ReturnSet*. Next $CandSet \leftarrow (CandSet \cap TempSet)$, which has the effect of ensuring that all all surviving elements of *CandSet* are elements of both V_s and V_n, *i.e.* all elements of *CandSet* are connected to all elements of *ReturnSet*. Once this step is carried out, it is safe to pick another element from *CandSet* and repeat the cycle untill *CandSet* is the null set. At this stage *ReturnSet* cannot be further augmented and the algorithm halts. A more precise description of the algorithm is given below.

• Initialization

1. Compute $norm = \sum_{i=1}^{N} n_i^{\gamma}$.
2. Evaluate $p_i = (n_i^{\gamma}/norm) \;\; \forall \, i \; 1 \leq i \leq N$.
3. Pick some j with probability p_j.
4. Insert V_j in *ReturnSet*.
5. $CandSet \leftarrow L_j$.
6. $TempSet \leftarrow \emptyset$.

• Main Loop

while $CandSet \neq \emptyset$ do

1. Evaluate $n_i^{\gamma} \; \forall \; V_i \in CandSet$
2. Compute $norm = \sum n_i^{\gamma}$
   with the summation restricted to elements of *CandSet*
3. Compute $p_k = (n_k^{\gamma}/norm) \; \forall \, V_k \in CandSet$
4. Select $V_n \in CandSet$ with probability p_n.
5. Insert V_n in *ReturnSet*.
6. $TempSet \leftarrow \{V_m \in L_n : V_m \notin ReturnSet \text{ where } 1 \leq m \leq N\}$
7. $CandSet \leftarrow (CandSet \cap TempSet)$
8. $TempSet \leftarrow \emptyset$

If $CandSet \equiv \emptyset$ return *ReturnSet*.

It is also possible to define a deterministic greedy heuristic in which the vertices selected in step 3 of the initialization and step 4 of the main loop are just those with the largest number of neighbors. If the parameter γ is made larger and larger, then the output from the program will increasingly resemble that from a deterministic greedy heuristic. It is worth pointing out that our algorithm does not make use of the relative location of markers with respect to each other along the chromosome, thus the methodology we outline is applicable whether the LD decays rapidly or slowly as a function of the distance between markers on the chromosome. The output from the algorithm returns a subset of markers which is not necessarily the largest subset of independent markers, nontheless the number of markers returned could still be large enough to get an accurate handle on population stratification. We now briefly discuss the application of of this to real data, more details are available in (Hamblin, 2010).

In conjunction with the Barley Coordinated Agricultural Project (www.BarleyCAP.org), 1816 Barley lines (treated as individuals for our purposes) were genotyped at 1415 markers. Five initial attempts to run *Structure* with 500,000 iterations and 100,000 burn in steps taking into account the association between markers and allowing for admixture between populations were unsuccessfull due to non-convergence of the MCMC iterations. At this stage, our algorithm was used to identify a subset of markers with linkage disequilibrium r^2 between any two markers in the subset to be less than 0.25, a criteria used to decide when to consider markers unlinked. A subset of 486 markers was identified and used as input for *Structure* allowing for admixture between individuals but no association between markers. Eight runs of *Structure* with 100,000 burn in and 200,000 analysis iterations all converged with consistent likelihood estimates, illustrating the utility of selecting unassociated markers as opposed to using the entire set and allowing for association between markers. It is worth pointing out that constituent populations identified by *Structure* have differing linkage disequilibrium structure at both short and large distances, with some SNPs in a few but not all of the subpopulations in high LD even when 50cM apart. As mentioned earlier, our algorithm does not use map distances in selecting markers, neither the presence of significant LD between unlinked markers nor the very different patterns of LD in the subpopulations is an issue. This feature gives rise to a complex edge structure on the graph, similar to the examples considered in (Thomas & Camp, 2004).

We next turn our attention to the relevance of the maximum independent set problem to non-parametric methods for analyzing population stratification. The most widely used non-parametric approach for analyzing population stratification is Principal Components Analysis, a number of popular implementations such as EIGENSTRAT (Price, 2006) and EIGENSOFT (Patterson, 2006) are available. We will discuss the relevance of the apparoaches just described to EIGENSOFT and then show how some of the statistical methodology in EIGENSOFT might have applications to machine learning problems outside of statistical genetics. The input data for EIGENSOFT is a rectangular data matrix where the rows correspond to individuals and there is one column for each marker, the entries of the data-matrix correspond to the genotypes suitably parametrized and standardized. The key idea behind the implementation in (Patterson, 2006) is the realization that in the absence of population stratification, the largest Singular Value of the data matrix is distributed according to the Tracy Widom distribution (Tracy & Widom, 1994). However, even in the absence of stratification, deviations from the Tracy Widom distribution are possible if there is LD between the markers. One way to avoid false signals of population stratification is to choose a set of markers which are mutually uncorrelated, preferrably as large an unrelated set of markers

as possible. As we have seen in the discussion of model dependent population stratification, choosing this set is tantamount to solving an instance of a maximum independent set on an undirected graph defined by the LD matrix. In practical instances, alternative methods have been used to find a set of unrelated markers, for example by exploiting special features of the LD structure or by other approximations(Heerwarden, 2010). It is not clear that these approaches will find the largest possible set of uncorrelated markers which is required for putting the most stringent bounds on the extent of population stratification. Very few (if any) attempts have been made to use the methodology previously described to identify as large a subset of unrelated markers as possible, such an analysis could be fruitful. This concludes our discussion on the application of graph theoretical methods for analyzing population stratification. The key point of our discussion is the relevance of the problem of finding the maximum independent set to understanding population stratification. This connection has not been established before (to the best of our knowledge) and can be exploited to speed up the the analysis of population stratification in real data sets. Before going on to discuss the application of graph theoretical methods to multiple testing, it is worth mentioning how the methodology developed in EIGENSOFT could possibly be used to address a long standing problem in cluster analysis, *i.e.* how to identify the number of distinct groupings in a dataset. As mentioned in our discussion of *Structure* the number of constituent populations to fit in *Structure* is user defined, in (Patterson, 2006) the authors point out how this number might be reliably estimated using the sample singular values of the data matrix and the details of the Tracy Widom Distribution. What this amounts to is computing a non-parametric statistic of the dataset which is then used to estimate the number of distinct groupings in the dataset. Since the methodology is very general and model independent it could conceivably be applied to a whole range of problems far removed from statistical genetics.

2.2 Multiple testing

Another major source of Type-I error in Genome Wide Association studies (GWAS) is false positives arising from multiple testing, and as mentioned earlier these can arise even if population stratification between cases and controls is fortuitously negligible or has been controlled for in some manner. Before we discuss the relevance of graph theoretical methods for understanding multiple testing artefacts, it is worth outlining the root of the problem and some common remedies. A large number of markers (N) are tested one after another for association with the trait or disease of interest. Under H_0 none of the markers are associated with the trait, and in addition the p-values for the test statistic are distributed like $\sim U(0,1)$. For a significance level α the expected number of significant tests under H_0 will be $\sim N\alpha$, since N in modern GWAS can be $\mathcal{O}(10^6)$ this leads to a sizeable number of false positives even if α is small. One way to ensure that there are no false positives with N independent tests is by choosing α so that $N\alpha \ll 1$ (the Bonferroni correction). However this leads to such stringent significance thresholds that only markers with very strong effects are picked up and many markers associated with the trait are ignored because their effects are not large enough to survive the stringent significance threshold, *i.e* there are is a large Type II error rate. Furthermore, if all the tests are not independent due to correlations between the markers the correction is excessively conservative. This problem can be avoided by permutation testing which leads to a non-parametric estimate of the number of significant test under H_0 given the correlation structure between the markers. While this approach certainly works it can become computationally very intensive when there are hundreds of thousands of markers to be tested. A less computationally intensive method to lower the number of Type II errors at

the cost of allowing a certain number of false positives is by controlling the False Discovery Rate (FDR)(Benjamini & Hochberg, 1995); variants on this idea have also been considered. It has been suggested by (Nyholt, 2004) that the Bonferroni corrections be modified by replacing N with N^*, the number of independent tests, where hopefully N^* is very much smaller than N. With this replacement, the significance threshold can be made less stringent lowering the Type II error rate. We will next examine this sugestion in the language of graph theory, more specifically we will consider the problem of finding a suitable subset of independent markers, the size of this subset is the number of independent markers. One key observation is that restricting ourselves to a subset of markers is most meaningfull if it is possible to choose that subset of markers not only to be statistically independent also to be serve as surrogates for all the markers under consideration. If the latter condition is fulfilled, then testing only the markers in the independent subset can be regarded as testing each and every marker. If this condition is not fulfilled, we run the risk of skipping association tests on some markers which are part of the panel. It is worth pointing out that the idea that a limited subset of markers can be used as surrogates for an entire panel is well established; this is the notion underlying the use of tag SNPs in GWAS (Carlson, 2004). Identifying an optimal tag SNPs in a marker panel given the LD matrix between the markers can be reformulated as a variety of different well known graph theoretical problem, including the search for a dominating set of smallest size (Li & Wang, 2011). In order not to underestimate the number of independent tests it is necessary that the subset of markers be as large as possible; from our earlier discussion of population stratification it is clear than once again we are dealing with finding a maximum independent set on a undirected graph defined by the LD structure between the markers and a user defined specification of statistical independence between markers. The condition that the markers we select be proxies for all the markers in the panel can be fulfilled by requiring that each vertex be connected by an edge to at least one of the markers in the maximum independent set. In other words, the maximum independent set should be a dominating set for the graph. Since any maximal independent set is a dominating set (Foulds, 1992) the maximum independent set satisfies this condition. If the heuristic used to find the maximum independent set only returns a maximal independent set, this attractive feature will be preserved. Thus estimating the number of independent tests via maximum independent set heuristics seems to have some advantages. One can also approach the problem of estimating N^* in terms of the the size of the smallest dominating set on the graph. If the smallest dominating set turns out not to be an independent set, then the resulting estimate of N^* would be smaller than what we would obtain from analyzing independent sets, but not easy to estimate precisely, given the dependence of the markers. This point is illustrated in Fig. 1. where$\{A, B, E, F\}$ is the maximum independent set, but all markers can be tested by considering just two (not independent) markers C and D. In situations such as these, it is not clear what to value use for N^*.

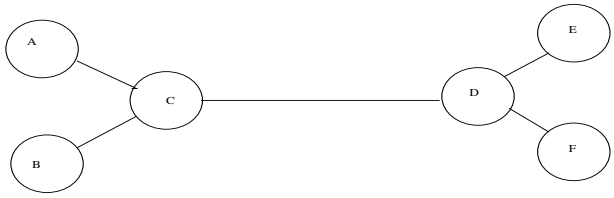


Fig. 1. Ambuguity in N^*

In practise implementing the prescription of (Nyholt, 2004) has been shown to be problematic in real and simulated datasets (Dudbridge & Koeleman, 2004);(Salyakina, 2005);(Coneely & Boehnke, 2007), but there has been no general model independent analysis as to why these difficulties arise. Our graph theoretical analysis sheds light on possible ambiguities in the prescription of (Nyholt, 2004) which may be at part of the reason for its observed limitations. Before concluding our discussion on the applications of the maximum independent set it worth recalling that all that heuristics can deliver is a lower bound on the size of the maximum complete set. What is missing is some means using the data to obtain an estimate of the upper bound, such an estimate could potentially improve the performance and applicability of the heuristics. One feature of the data which has not been exploited is that the $N \times N$ matrix of correlation values is obtained from a data matrix with dimensionality $N_{ind} \times N$, where N_{ind} is the number of individuals and typically $N_{ind} < N$. Thus both the data matrix and the correlation matrix can be expected to have rank less than N. The redundancy in the rows of the correlation matrix due to the reduced rank could possibly be exploited in order to obtain an upper bound on the size of the maximum independent set. As has been shown in (Li & Li, 2005) this redundancy can be used to obtain an alternative estimate for the number of independent tests; combining the approach of (Li & Li, 2005) with the maximum independent set heuristic we describe here could be a fruitfull line of future research.

3. Blocking Gibbs

In the remainder of this chapter we will focus on possible applications of graph theoretical methodology to analysing pedigree data; more precisely we will consider individuals with known parent offspring relations and genotypic information at a possibly large number of loci. As was mentioned earlier, the pedigree structure and the genotypes can be combined to form a Bayesian Network where the conditional probabilities along the edges are defined by Mendelian Genetics. Since there are known genotypes there are vertices in the Bayesian networks where evidence is available. From the standpoint of genetic linkage analysis one of the most important quantities to be computed from a pedigree and associated genotypic data is the Likelihood (Ott, 1999). Computing the Likelihood involves evaluating a very complex series of nested sums and products of conditional probabilities over expressions such as the one shown below:

$$\cdots P(g^G \mid g^F, g^E)\, P(g^H \mid g^F, g^E)\, P(g^E \mid g^A, g^B) P(g^I \mid g^C, g^D) P(g^C \mid g^A, g^B)\, P(g^D \mid g^A, g^B) \cdots \tag{1}$$

g^A, g^B, *etc.* are discrete random variables representing either genotypes or alleles, and the $\cdots$ indicate the presence of many more such conditional probabilities. The conditional probabilities shown above are typical of the factors that would appear in the Joint Probability Distribution defined by the Bayesian Network, however realistic pedigrees often contain far more factors than can be written down. Computing the Likelihood involves summing over all allowed values of all the random variables, (*i.e.* all consistent genotypes), and in realistic situations where there are huge numbers of conditional probabilities, this is analytically intractable. Numerical solutions are hypothetically possible due to the local structure of the computations (Lauritzen & Spiegelhalter, 1988). The computational effort involved depends critically on the order in which the summations are performed (Jordan, 2004) and determining the lowest cost summation order is $\mathcal{NP}$ Hard (Arnborg, 1987). If no good heuristic algorithm for determining the most efficient summation order can be found the multiple sum cannot be performed exactly, and must be approximated by sampling the most significant terms.

This provides the motivation for introducing Markov Chain Monte Carlo (MCMC) methods which have been used extensively in linkage analysis (Thompson, 2005). The simplest form of MCMC sampler to implement is the Gibbs sampler, which however can be very tricky to implement on a large pedigree with many known genotypes. One of the key conceptual difficulties in implementing the Gibbs Sampler can be illustrated on a very simple situation involving just four individuals as shown in the following figure.

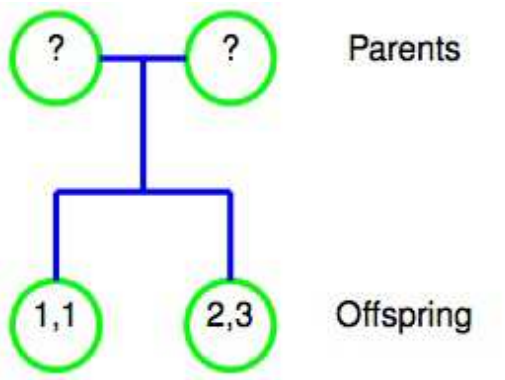


Fig. 2. Trouble with Gibbs

The two offspring have observed genotypes, and the laws of Mendelian Inheritance dictate that the two parental genotypic combinations are either $\{\{1,2\},\{1,3\}\}$ or $\{\{1,3\},\{1,2\}\}$ in obvious notation. In order for the sampler to be irreducible transitions between configurations should be possible, and in keeping with textbook Gibbs sampling one parental genotype would be updated at a time keeping the other fixed. Let us begin from the configuration$\{\{1,2\},\{1,3\}\}$ with a view to reaching $\{\{1,3\},\{1,2\}\}$. If we sample conditional on any one parent and the known genotypes, there is no way in which we can update the genotype of the other parent; *i.e.* the sampler gets stuck in the starting configuration. Thus a single site update is problematic and it is easy to see that the root of the problem lies in the stringent constraints arising from Mendelian Genetics which lead to strong correlations between the variables to be updated. A possible solution within the framework of Gibbs sampling is to update both parental genotypes simultaneously, *i.e.* use a blocking Gibbs sampler where a block consists of multiple stochastic variables which are strongly correlated and must be updated simultaneously. The idea of updating multiple strongly correlated variables during a single MCMC update in order to improve convergence is well established and outperforms standard Gibbs sampling in statistical genetics (Totir, 2003) and other applications. (Swendsen & Wang, 1987); (Roberts & Sahu, 1997). Furthermore, this approach has been applied in Bayesian Networks arising not only in Statistical Genetics (Jensen & Kong, 1999);(Thomas, 2000), but also in expert systems (Jensen, 1995). For our purposes the optimal choice of blocks is not only crucial for ensuring the irreducibility of the sampler but also for improving the mixing and convergence properties of the sampler. In the rest of this chapter we will study the issue of block definition and relate this problem to a well known problem in machine learning, that of partitioning data sets into semi-autonomous clusters. Before doing so we will briefly mention another aspect of likelihood computations on large pedigrees which has attracted recent attention, *i.e.* the relation to constraint satisfaction. The problem of finding assignments of unknown genotypes consistent with known genotypes, the pedigree

structure and the laws of Mendelian Inheritance can be viewed as finding the solution of a constraint satisfaction problem, and is known to be computationally hard (Aceto, 2001). There is however one additional complication which arises in dealing with pedigrees, each solution can be assigned a posterior probability and what is required are the solutions with higher posterior probabilities. What an irreducible ergodic MCMC sampler should do is not only find solutions to a very complex contraint satisfaction problem, but also assign the correct posterior probability to the various solutions. Seen in this light it is easy to see why the MCMC sampling on pedigrees can be so challenging.

The key difficulty in constructing blocks is correctly grouping strongly correlated variables together, followed by updating them simultaneously in a manner consistent with the known genotypes. In simple instances like Fig. 2, grouping variables is easy, but in more complicated cases such as the pedigree in Fig. 3 it can be highly non-trivial.

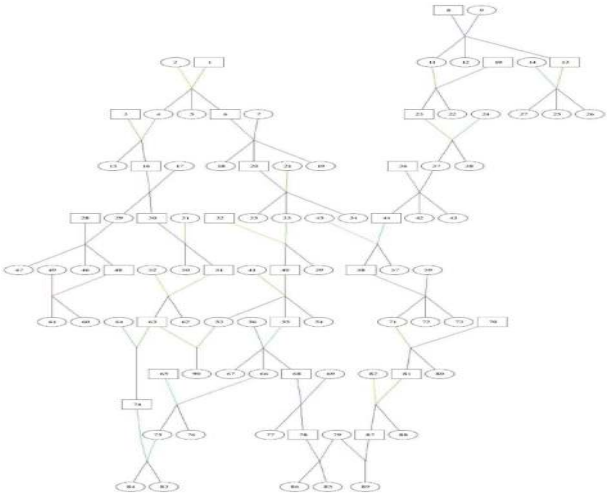


Fig. 3. Large pedigree

A moments reflection will suggest that the difficulty in partitioning the pedigree of Fig. 3 into blocks arises because of the large number of cycles in the graph. A more general analysis of the difficulties has been undertaken in (Jensen & Sheehan, 1998), the presence of cycles is indeed a problem and no general solution for dividing an arbitrary pedigree into blocks is known. However, a few features of a well motivated Blocking Scheme can be identified

- The assignment of blocks should be such that all the variables are assigned to at least one block. If this condition is not satisfied, some variables may not be updated leading to biased MCMC estimates.
- All strongly correlated variables must be contained in the same block, if not the irreducubility issues mentioned earlier will arise
- Within a given block variables should be more strongly correlated with each other than with variables outside the block.

It is easy to understand the relevance of these criteria, the last two criteria are not only relevant for pedigree analysis but are also similar to what might expected from an optimal partitioning

of a data set into clusters. Note that we have not specified the number of partitions in advance; if this scheme were to be implemented on an arbitrary dataset it would not only assign all the elements of the dataset to clusters but could also return the number of partitions. Thus any heuristic solution for finding optimal blocks could be very useful in a number of machine learning applications. One important distinction between pedigree analysis and other applications is the requirement that the sampler mix rapidly, this was the motivation for the use of overlapping blocks. The relevance to data partitioning problems of a more general nature is greatest when the clusters are expected to overlap. The other important distinction is that many genotypes in pedigrees may be unknown, corresponding to vertices with no information. Missing or ambiguous data are not as widely considered in other data sets, so the analogy works better when there are not too many unknown genotypes.

One outline of a Blocking Gibbs scheme was made in (Abraham, 2007) where the problem of Block Identification and consistent assignment of genotypes were addressed simultaneously. The algorithmic insight exploited in (Abraham, 2007) was that the genotypes of any given individual are strongly dependent by just a handfull of close relatives; in the language of the pedigree graph the state of a vertex is influenced by just a handfull of neighbouring vertices. This is because the edge structure in the graph reflects a combination of either relatedness between individuals or physical distance between loci. The notion of neighbouring vertices can be made more precise by defining distances between vertices in terms of a breadth first search. Due to the underlying Markov Field Property, vertices which are far apart as defined by the breadth first search are expected to be roughly independent. Once there is a guideline for deciding which vertices can be expected to be independent of other vertices, it becomes possible to partition the graph into overlapping blocks in which consistent genotype assignments in a block can be made with little input from the evidence from other blocks. The dataset used in (Abraham, 2007) is very complex and has many of the features discussed in (Jensen & Sheehan, 1998) which are known to lead to difficulties, nonetheless it was possible to generate a consistent set of genotypes using the scheme just outlined. Furthermore, it was checked that the posterior probabilities of the genotypes found in this manner were consistent with those that would have been obtained in the absence of any approximations. This suggests that separation of vertices on the graph is a useful guideline for assessing the approximate independence of the corresponding random variables. Criteria similar to these have been successfully used to construct blocks and mcmc samplers in other complex examples, (Habier, 2009);(Habier, 2010) indicating that the basic idea may have a broad general applicability.

If we consider the problem in a more general light, what we have done is to use the known correlations in a data set containing many discrete observations to identify subsets of variables which have strong correlations with each other but weaker correlations with the other variables. If this methodology were to be applied to cluster a general data set with a a known matrix of correlation values it would be first necesary to define graph and identify edges between the vertices (datapoints). Identifying edges could be acchieved through a user defined threshold which could be defined independent of the data values as described in our discussion of population stratification , or could be defined in terms of some suitable number of sample standard deviations above the sample mean of all the correlation values. Once the edges are specified in this manner, the procedure used in (Abraham, 2007) could be used to define blocks which in a more general case would correspond to a cluster in the data set. One advantage of this procedure is that it has been shown to work in the context of pedigree graphs where inaccurate assignment of vertices to clusters will often be penalized by poor

MCMC convergence or in extreme cases by a lack of irreducibility of the sampler. Adapting the blocking methodologies in (Abraham, 2007);(Habier, 2009) and (Habier, 2010) to other cluster identification in general ohmics data sets could prove to be fruitful.

We next consider a long standing issue which is relevant in both block assignment in blocking gibbs and cluster assignment in general, *i.e.* the problem of determining the number of independent subgroups in the data set making as few model dependent assuptions as possible. As was mentioned in our discussion of non-parametric population stratification , the authors of (Patterson, 2006) suggest that from the elements of a suitably constructed correlation matrix a test statistic can be obtained which can be used to decide on the appropriate number of populations to use as input for parametric population stratification analysis. The treatment of this issue in (Patterson, 2006) is so general that it would appear to be the basis for a model-free approach that could be used to estimate the number of subgroups in an arbitrary ohmics data set given a matrix of correlation values. As applied to blocking gibbs, the matrix of correlation values could be substituted by the distance matrix used in (Abraham, 2007) or some more sophisticated variant thereof. In this regard it is worth recalling that number zero eigenvalues of the Laplacian of an undirected graph is the number of connected components, which supplies a lower bound on the number of clusters. Thus the connection between the entries of a suitable correlation matrix and the number of clusters is well established, by applying the results of (Patterson, 2006) it might be possible to extract more detailed information on the number of clusters present in a dataset.

4. Conclusions

In this chapter we have discussed the relevance and applications of graph theoretical methods to a number of problems in statistical genetics. In particular, some novel applications of the maximum independent set on an undirected graph to population stratification were presented. Some key issues in the construction of Blocking Gibbs Samplers on complex pedigrees were discussed along with their relevance outside of statistical genetics.

5. Acknowledgments

RLF and KJA both received supprot from the United States Department of Agriculture, National Research Iniative Grant USDA NRI-2009-03924. KJA also acknowledges financial support of the program Professor Visitante do Exterior of Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) , Brasil. KJA thanks Prof. Jean-Eudes Dazard and members of the Department of Computation and Mathematics, University of São Paulo (Ribeirão Preto) for valuable discussions.

6. References

Abraham, K.J., *et.al.*. (2007). Improved techniques for sampling complex pedigrees with the gibbs sampler, *Genetics, Selection and Evolution* 39: 27–38.

Aceto, L., *et. al.*. (2001). The complexity of checking consistency of pedigree information and related problems, *The Journal of Computer Science and Technology* 19(1): 42–59.

Arnborg, S., *et.al.*. (1987). Complexity of finding embeddings in a k-tree, *SIAM Journal on Algebraic and Discrete Methods* 8: 277–284.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B* 57: 289–300.

Carlson, C.S., *et.al.*. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium, *American Journal of Human Genetics* 74: 106–120.

Coneely, K. & Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustments of p values for multiple correlated tests, *American Journal of Human Genetics* 81(6): 2074–2093.

Dudbridge, F. & Koeleman, B. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies, *American Journal of Human Genetics* 75(3): 424–435.

Falush, D., *et.al.*. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies, *Genetics* 164: 1567–1587.

Fishelson, M. & Geiger, D. (2002). Exact genetic linkage computations for general pedigrees, *Bioinformatics* 18 Suppl. 1: S189–S198.

Fishelson, M. & Geiger, D. (2004). Optimizing exact genetic linkage computations, *Journal of Computational Biology* 11(2-3): 263–275.

Foulds, L. (1992). *Graph Theory Applications*, Springer Verlag.

Habier, D., *et.al.*. (2009). Genomic selection using low density marker panels, *Genetics* 182: 343–353.

Habier, D., *et.al.*. (2010). A two-stage approximation for analysis of mixture genetic models in large pedigrees, *Genetics* 185: 655–670.

Hamblin, M.T., *et.al.*. (2010). Population structure and linkage disequilibrium in us barley germplasm: Implications for association mapping, *Crop Science* 50(2): 556–566.

Heerwarden, J. V., *et.al.*. (2010). Fine scale genetic structure in the wild ancestor of maize zea mays ssp parviglumis, *Molecular Ecology* 19: 1162–1163.

Jensen, C. & Kong, A. (1999). Blocking gibbs sampling for linkage analysis in large pedigrees, *American Journal of Human Genetics* 65(3): 885–901.

Jensen, C. & Sheehan, N. (1998). Problem with determination of noncommunicating classes for markov chain monte carlo applications in pedigree analysis, *Biometrics* 54: 416–425.

Jensen, C.S., *et.al.*. (1995). Blocking gibbs sampling in very large probabilistic expert systems, *International journal of human computer studies* 42(6): 573–704.

Jordan, M. (2004). Graphical models, *Statistical Science* 19(1): 140–155.

Lauritzen, S. L. (1996). *Graphical Models*, Oxford University Press.

Lauritzen, S. & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems, *Journal of the Royal Statistical Society Series B* 50: 157–224.

Li, J. & Li, K. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix, *Human Heredity* 95: 221–227.

Li, J. & Wang, W.-B. (2011). Tag snp selection, *in* E. Zeggini & A. Morris (eds), *Analysis of Complex Disease Association Studies, A Practical Guide*, Academic Press, pp. 49–65.

Nyholt, D. (2004). A simple correction for multiple testing for single nucleotide polymorphisms in linkage disequilibrium with each other, *American Journal of Human Genetics* 74(2): 765–769.

Ott, J. (1999). *Statistical Methods in Genetic Epidemiology*, The Johns Hopkins University Press.

Patterson, N. *et.al.*. (2006). Population structure and eigenanalysis, *PLoS Genetics* 2(12): 2074–2093.

Price, A.L., *et.al.*. (2006). Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics* 38: 904–909.

Pritchard, J.K., *at.al.*. (2000). Inference of population structure using multilocus genotype data, *Genetics* 155(2-3): 945–959.

Roberts, G. & Sahu, S. (1997). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler, *Journal of the Royal Statistical Society Series B* 59(6): 573–704.

Salyakina, D., *et.al.*. (2005). Evaluation of nyholt's procedure for multiple testing correction, *Human Heredity* 60: 19–25.

Swendsen, R. & Wang, J.-S. (1987). Nonuniversal critical dynamics in monte carlo simulations, *Physical Review Letters* 58: 86–88.

Thomas, A. & Camp, N. (2004). Graphical modelling of the joint distributions of alleles at associated loci, *American Journal of Human Genetics* 74: 1088–1101.

Thomas, A., *et.al.*. (2000). Multilocus linkage analysis by blocked gibbs sampling, *Statistics and Computing* 10: 259–269.

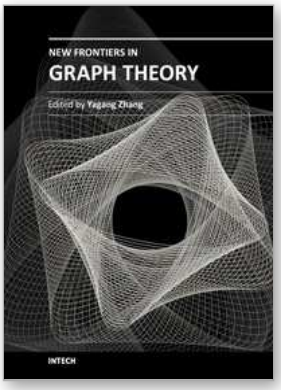
Thomas, D. T. (2004). *Statistical Methods in Genetic Epidemiology*, Oxford University Press.

Thompson, E. A. (2005). Mcmc in the analysis of genetic data on pedigrees, *in* W. S. Kendall, F. Liang & J.-S. Wang (eds), *Markov Chain Monte Carlo Innovations and Applications*, World Scientific Publishing, pp. 183–217.

Totir, L.R., *et.al.*. (2003). A comparison of alternative methods to compute conditional genotype probabilities for genetic evaluation with finite locus models, *Genetics,Selection and Evolution* 35: 585–604.

Tracy, C. & Widom, H. (1994). Level spacing distribution and the airy kernel, *Communications in Mathematical Physics* 159: 151–174.

Weir, B. S. (1996). *Genetic Data Analysis II*, Sinauer Associates, Sunderland MA 01375 USA.



New Frontiers in Graph Theory

Edited by Dr. Yagang Zhang

ISBN 978-953-51-0115-4
Hard cover, 526 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

Nowadays, graph theory is an important analysis tool in mathematics and computer science. Because of the inherent simplicity of graph theory, it can be used to model many different physical and abstract systems such as transportation and communication networks, models for business administration, political science, and psychology and so on. The purpose of this book is not only to present the latest state and development tendencies of graph theory, but to bring the reader far enough along the way to enable him to embark on the research problems of his own. Taking into account the large amount of knowledge about graph theory and practice presented in the book, it has two major parts: theoretical researches and applications. The book is also intended for both graduate and postgraduate students in fields such as mathematics, computer science, system sciences, biology, engineering, cybernetics, and social sciences, and as a reference for software professionals and practitioners.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

K.J. Abraham and Rohan Fernando (2012). Applications of Graphical Clustering Algorithms in Genome Wide Association Mapping, New Frontiers in Graph Theory, Dr. Yagang Zhang (Ed.), ISBN: 978-953-51-0115-4, InTech, Available from: http://www.intechopen.com/books/new-frontiers-in-graph-theory/applications-of-graphical-clustering-algorithms-in-genome-wide-association-mapping

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.