BAYESIAN STATISTICS 8, pp. 1–26.
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2007

Assessing the Effect of Genetic Mutation: A Bayesian Framework for Determining Population History from DNA Sequence Data

S. P. Brooks, I. Manolopoulou
University of Cambridge, UK
steve@statslab.cam.ac.uk

B. C. Emerson
University of East Anglia, UK

Summary

In cases where genetic sequence data are collected together with associated physical traits it is natural to want to link patterns observed in the trait values to the underlying genealogy of the individuals. If the traits correspond to specific phenotypes, we may wish to associate specific mutations with changes observed in phenotype distributions, whereas if the traits concern spatial information, we may use the genealogy to look at population movement over time.

In this paper we discuss the standard approach to analyses of this sort and propose a new framework which overcomes a number of shortcomings in the standard approach. In particular, we allow for uncertainty associated with the underlying genealogy to fully propagate through the model to directly interact with the inferences of primary interest, namely the effects of genetic mutations on phenotype and/or the dispersal patterns of populations over time.

Keywords and Phrases: Reversible jump MCMC; posterior model probability; cladogram; cluster analysis; phenotypic analysis; phylogeographic analysis; nested clade analysis.

The authors gratefully acknowledge the financial support of Trinity College, Cambridge and the Leverhulme Trust. The authors would also like to thank Sara Goodacre for her invaluable input to this project.

1. INTRODUCTION

This paper is motivated by the increasing interest in phenotypic and phylogeographic analyses arising from recent developments in genetic research. These methods attempt to explain phenotypic or geographic variations in terms of the underlying genealogy in order to, for example, identify mutations associated with characteristic traits (phenotypic analysis) or infer population movement over time (phylogeographic analysis).

These methods are important for many reasons. Phenotypic analyses can be used to prevent or predict unwanted genetic traits which predispose the associated individual to disease, encourage the development of desired traits and even control specific genes to influence the associated phenotype. Phylogeographic analyses are used to help understand the long-term evolutionary history and behaviour of species such as colonisation and range expansion, as well as investigating the effect of geological events such as glaciation or seismic activity, for example.

Both forms of analysis require similar data and proceed in similar ways. We begin with data from a series of individuals comprising for each individual i a DNA sequence S_i together with an associated (perhaps vector) variable $\boldsymbol{T}_i$ recording the characteristic traits of interest. The latter might be indicators for the presence of disease or the geographic location of the associated individual, for example. The task then is to determine patterns in the traits observed across individuals that can be explained by the underlying genealogy e.g., do all those with a particular trait have commonalities between their corresponding sequences?

In answering such questions our first task is to determine the underlying genealogy, that is to construct a network (or so-called cladogram) of nodes which correspond to specific sequences and edges which correspond to specific genetic mutations (Cassens *et al.* 2005, and Posada and Crandall, 2001) . Once we have this network which shows how each individual is related to the rest through a sequence of one or more genetic mutations, we can look to see whether deleting any edge (or collection of edges) creates clusters of sequences whose associated traits are similar within a cluster but distinct between clusters. An alternative motivation for the analysis is to view it as a simple clustering problem in which we wish to allocate individuals to two or more groups in terms of the observed traits but in a way that is consistent with the underlying genealogy,

This approach is probably best explained via a simple example. Suppose we have data from 6 individuals with sequences

$$ATCGA, ATCGA, ATCGA, ATTGA, ACCGA \text{ and } TTCGA$$

and corresponding trait values $1.9, 3.5, 3.1, 4.4, 1.2$ and 2.0, say. We see that the first 3 individuals have the same sequence and so in our cladogram, they will all correspond to the same node. We thus have four distinct nodes and we connect any two nodes by an edge if those two nodes differ in value at only one point in the sequence. The corresponding cladogram is given in Figure 1.

Our aim is to draw some conclusion of the type "a mutation at the ith nucleotide position is associated with a significant change in the values of the observed trait". In order to do this, we first need to assign a distribution to the trait values (in order to determine "significance") and then consider each partition in turn that can be obtained by deleting a single edge from the associated cladogram. Here, we have only three edges and so only three potential arrangements of the observations into two clusters. We can thus cluster the traits as $(\{4.1\}, \{1.9, 3.5, 2.1, 1.2, 2.0\})$

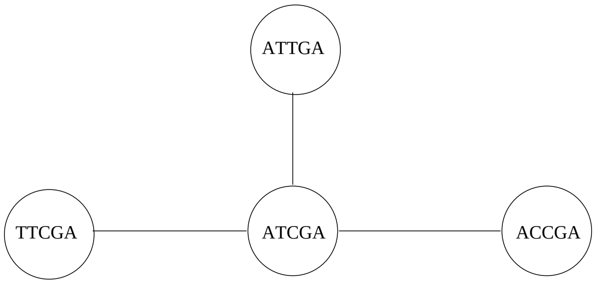

Figure 1: *Basic Cladogram showing the sequences observed and the mutations linking them.*

$(\{1.2\}, \{1.9, 3.5, 2.1, 4.1, 2.0\})$ or $(\{2.0\}, \{1.9, 3.5, 2.1, 4.1, 1.2\})$ by deleting the edge corresponding to a mutation at the 3rd, 2nd and 1st nucleotide position respectively. An ad-hoc interpretation might suggest that a mutation at the third nucleotide position causes a change in the associated trait and we will discuss and develop more formal methods for testing this in due course. However, we note here that without the sequence data, if we were to cluster the trait values directly, we might well wish to group 4.1 and 3.5 together, a combination that is not permitted by the underlying genealogy. Thus, by using the sequence data, we reduce the number of possible clusters to those which have a direct interpretation in terms of the mutations required to obtain the sequences observed in the data.

Of course, this is a very simple example. In practice the cladogram is not so easily determined. Problems associated with missing values, homoplasy (multiple mutations at the same nucleotide position) and recombinations (mutations not at a single nucleotide position but obtained by swapping a partial sequence from one node with that from another) mean that the cladogram is non-unique and so the clustering procedure should take account of the associated uncertainty in the underlying genealogy though, in practice, this uncertainty is often ignored.

Current best statistical practice relies on a series of analytic steps in which each stage is conditioned on the outcome of the previous one. In particular, the "best" network is usually determined using a combination of statistical and ad-hoc procedures and then a form of iterative ANOVA is used to determine the mutations linking significantly distinct clusters. This modular approach is not able to fully propagate uncertainty at each stage into the final analysis and so inferences tend to overestimate the support for any one hypothesis over another and may, potentially, result in substantial inferential bias.

In this paper we describe a Bayesian approach in which different networks are assigned appropriate weights and then network-averaged mixture models are used to determine appropriate clusters. It is worth noting here, the existence of the already large but growing literature on inferring network structures given sequence data. By far the largest literature is concerned with the derivation of phylogenetic trees which incorporate both network and temporal structure by allowing edge lengths to denote times between observed mutations. See, for example, Mau *et al.* (1999), Newton *et al.* (1999), Larget and Simon (1999), Huelsenbeck *et al.* (2002), Altekar *et al.* (2004) and Stamatakis *et al.* (2005). Similarly a large number of software packages have been developed which create the phylogenetic tree from sequence data e.g., MrBayes, BAMBE and RAxML. However, here we do not require temporal information but do wish instead to infer the state of missing nodes in the tree. Though there is a direct link between the cladogram and the phylogenetic tree, here we work directly with the network structure (or cladogram).

We begin, in Section 2 with a review of the traditional method of undertaking such analyses based upon an analysis of variance built around the so-called nested cladogram. Then, in Section 3, we introduce a Bayesian alternative based around the idea of bivariate normal mixture models to help determine appropriate geographic clusters using the underlying haplotype network to determine cluster membership. In Section 4 we extend the basic analysis to deal with problems associated with missing values and homoplasy both of which occur frequently in data of this sort. We conclude with the results of our analysis in Section 5 and finally a discussion in Section 6 which highlights a variety of potential future research areas.

2. NESTED CLADE ANALYSIS

The main method used for associating physical traits with underlying genetic mutations is the so-called nested clade analysis. The standard nested clade analysis proceeds in three steps: first, the cladogram is constructed; second, a nested cladogram is formed; and, finally, we test for associations between clades and the corresponding traits.

2.1. *The Cladogram*

As we discussed in the previous section, the cladogram is a graph comprising nodes which correspond to specific sequences and edges which correspond to mutations at specific nucleotide positions. For the basic cladogram, we join all nodes which differ at just one nucleotide position by an edge and, if the resulting graph is incomplete we introduce additional nodes which correspond to sequences that are believed either to be extant but were missed in the sampling, or to be extinct. In adding missing nodes we typically appeal to a parsimony argument and assume that cladogram with the smallest number of missing nodes is the correct one in that it most accurately reflects the evolutionary history of the species under study.

In the absence of homoplasy (and, indeed recombination, which we do not deal with here as we focus only on mitochondrial DNA) the cladogram is structurally unique in terms of the observed sequences. Thus, we know exactly which sequences are connected to which and, where there are missing values, we know the number of missing values between any two observed nodes and the locations at which the mutations for those missing sequences occur (though we may not know their order). For example, the cladogram corresponding to the sequences

$$ATCGA, ACCGA, TTCGA, ATTGA, ATTGC$$

is unique and has no missing values (see Figure 2). However, if the sequence $ATCGA$ were not observed then we know, in the absence of homoplasy, exactly what the missing sequence is and where it lies because of its two neighbours. If ATTGA were missing, we would know it, but we not know what the missing sequence actually was (it could be ATTGA or ATCGG). Note, however, that if $TTCGA$ were missing, we would not know it and this will be true for any leaf node.

The presence of homoplasy may create cycles or loops in the cladogram which means that one or more sequences can join the graph in two or more distinct ways. The presence of homoplasy can also lead to non-uniqueness of missing sequences

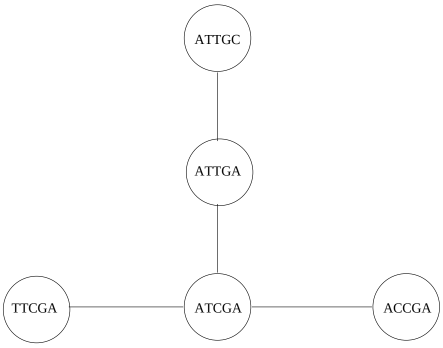

Figure 2: *A simple illustration of a cladogram with 5 nodes.*

and, indeed, can lead to uncertainty as to the number of missing sequences required to complete the graph.

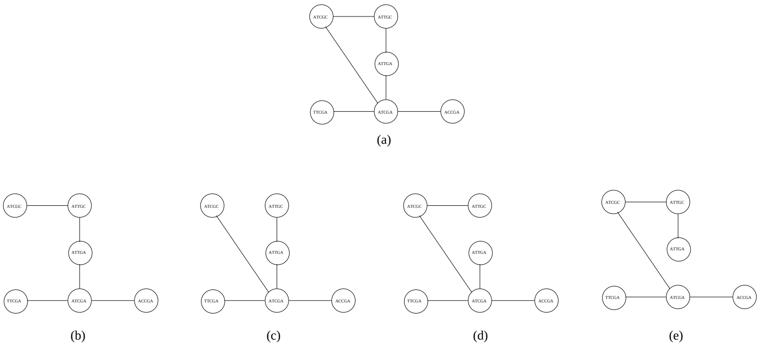

Figure 3: *(a) A 6-node cladogram with homoplasy. Figures (b)-(e) provide the corresponding cladograms with the homoplasy removed.*

For example, suppose we observe a 6th sequence: *ATCGC*. This sequence is just one nucleotide position apart from the sequences *ATCGA* and *ATTGC* and so could be joined to either one or both of these nodes (see Figure 3a). Again, the principle of parsimony is used to justify the assumption that any sequence will not occur via mutation from any other sequence in more than one way. This means that the true cladogram will be a sub-graph of the cyclical graph in Figure 3a formed by deleting one or more edges so that the graph is both complete and contains no loops. In this case, there are four possible graphs, see Figure 3b-e. Note that if the sequence *ATTGA* were not observed (as above) then no cycle would exist, there would be no observed homoplasy and the cladogram would be unique.

Again, it is common to use the parsimony principle to assume that both homoplasy and missing values should only be considered if they are required to complete the graph. Templeton *et al.* (1992) develops a sequential method for determining the number of missing values and for minimising homoplasy (and, indeed, recombination events). See also Lloyd and Calder (1991), Clement *et al.* (2000,2002), Crandall and Templeton (1993), Templeton *et al.* (1988) and Felsenstein (1992). We shall see in the following section how similar methods may be incorporated into the Bayesian analysis to account for the uncertainty associated with the underlying cladogram.

2.2. *The Nested Cladogram*

Once we have the cladogram, we next form the so-called nested cladogram. This is done following the guidelines set out in Templeton *et al.* (1987) as follows.

We start with our 0-step clades, which are just the sequences represented as leaves in our cladogram. Given the n-step clades, the $n+1$-step clades are formed by taking the union of all n-step clades which can be joined up by moving one mutational step back from the terminal node of each n-step clade. We continue the process recursively until all the nodes in our cladogram have been nested. As before, the nesting process is probably best explained via an example.

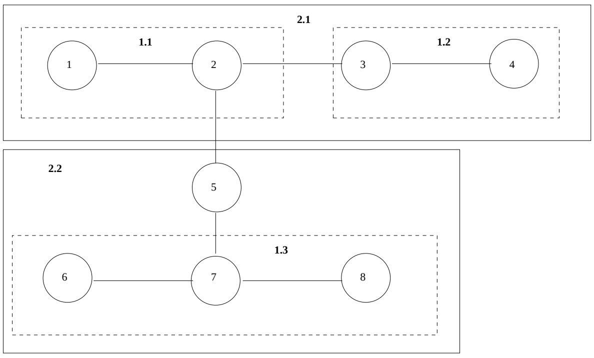

Figure 4: *A simple illustration of a nested cladogram comprising 3 1-step clades and 2 2-step clades.*

Figure 4 provides a simple illustration of a cladogram with 8 nodes. Following Templeton's algorithm, we first join the leaves to their neighbours to form the 1-step clades. In the case where two leaves are joined to the same neighbour, as for example occurs with nodes 6, 7 and 8, the corresponding clades are joined together. Thus, we obtain 1-2, 3-4 and 6-7-8 as our 1-step clades which we arbitrarily label as clades 1.1, 1.2 and 1.3 respectively. At the next stage these three groups are essentially treated as a single node and form the two step clades by combining 1-2 with 3-4 and 6-7-8 with 4. We label the corresponding 2-step clades 2.1 and 2.2. At the third stage we would combine the two 2-step clades to form a single group containing all nodes in the cladogram.

We can see here the link between the clusters obtained following Templeton's algorithm and those obtained by deleting edges in the graph as discussed in the previous section. However, not all edge-based clusters are obtained using Templeton's method and there are many cases where Templeton's approach leads to ambiguities or inappropriate clusters. Templeton's method might be compared with that of stepwise regression in which we attempt to identify the best model by successively adding and deleting parameters from an initial starting model. It is well recognised that this process does not necessarily lead to the best model and there are many examples where more exhaustive comparison have identified significantly better solutions than these sequential methods. Ad-hoc comparisons are used within Templeton's algorithm (based upon merging pairs of neighbouring clusters with smallest sample size, for example - Templeton and Sing, 1993) which introduces a somewhat arbitrary element to the clustering process in the final stage of the analysis. We will show how the Bayesian approach based upon creating clusters by deleting sets of edges from the graph can be used to more efficiently and reliably identify potentially significant clusters of sequences as well as allowing for the uncertainty associated with different clustering combinations to be properly propagated to the final stage of the analysis.

2.3. *Nested Clade Analysis*

The final stage of the standard phenotypic and phylogeographic analysis concerns the identification of significant clades i.e., those clusters of nodes that appear to well explain the variation in the observed traits. We distinguish here between the two forms of analysis as slightly different procedures are used for each.

The Phenotypic Analysis

The phenotypic analysis is based upon a standard (nested) analysis of variance (Templeton *et al.*, 1987). Suppose we have a cladogram with 0-step, 1-step, up to k-step clades (where the $(k+1)$-step clade contains all nodes in the graph) then we perform a $(k+1)$-way (M)ANOVA using the trait values to determine which clade levels exhibit significant deviation from the null hypothesis of equality of means. Sums of squares that are deemed significant are further decomposed to gain finer resolution on the mutations associated with differences in the means. For example, if the 1-step clades are deemed significant, then the associated sum of squares is sub-divided into independent sums of squares formed within each 2-step clade. For example, with our nested cladogram in Figure 4, we would decompose the Sum of Squares for the 1-step clades into two: the sum within clade 2.1 and that within 2.2. If the former were found to be significant, this suggests that the mutation between nodes 2 and 3 is associated with a significant change in the mean trait level.

This approach suffers from a number drawbacks associated with the risk of the effect of significant mutations carrying through to different clades masking the true patterns in the data. In practice, multiple comparison statistics are used together with Bonferoni-type corrections to ensure the correct overall significant level. We shall see in the next section, how the traditional ANOVA can be replaced by Bayesian mixture modelling which, once again, provides a more comprehensive and robust procedure for identifying mutations associated with changes in phenotype values.

The Phylogeographic Analysis

Phylogeographic analyses are treated slightly differently since the aims of the analysis are typically somewhat different from those of phenotypic analysis. Here we are interested in determining associations between observed genetic mutations and spatial patterns observed in the data. See Ibrahim *et al.* 1996, Templeton 1995, Gomez-Zurita *et al.* 2000, and Emerson and Hewitt 2005, for example. There are 3 major biological factors that can cause a significant spatial association with haplotype variation: i) restricted gene flow; ii) past fragmentation; and iii) range expansion (including colonisation). The standard phylogeographic analysis proceeds in two steps: a permutational contingency test is carried out to determine whether or not haplotypes are distributed randomly at all clade levels; then, if the hypothesis of randomness is rejected then further tests under assumptions of the three different scenarios above are conducted and the results interpreted via an inference key provided originally by Templeton (1998) and based upon a series of simulation studies.

The so-called nested clade phylogeographic analysis (NCPA) works by estimating two distinct geographic measures in the context of the hierarchical design inherent in the nested cladogram: the clade distance $D_c(X)$ for clade X, which represents the geographical range of that clade; and D_n (X), which measures how that particular clade is geographically distributed relative to its closest evolutionary clades (i.e. clades in the same higher-level nesting category as X). Specifically, $D_c(X)$ is the mean distance of clades within clade X from the geographical centre of that clade and $D_n(X)$ is the mean distance of the clades within clade X to the geographical centre of all same-level clades that lie within the higher-level clade that contains X. See Templeton *et al.* (1995), for example. The distribution of these distance measures under the null hypothesis of no geographical association within clade X is obtained by using a Monte Carlo approach, resampling the measures under permutations of the low-level nesting clades. See Templeton and Sing (1993) and Templeton *et al.* (1995), for example.

If the null hypothesis of no association is rejected then the permutation test is repeated under different assumptions as to the underlying nature of the geographic association. In practice, however, the D_c and D_n values are calculated and a prescriptive key provided by Templeton (1998) is followed to determine whether any of the three factors described above can be identified as adequately describing the spatial patterns of observed haplotypes.

In essence, the first stage of the NCPA is equivalent to an ANOVA based upon distance and suffers from similar problems associated with multiple testing and poor resolution in terms of identifying the level of clade at which significant mutations occur. As we shall see in the next section, this inferential process can be easily incorporated into the Bayesian framework that we develop here.

The three-stage approach described here, which represents current best practice for problems of this sort, suffers from problems at each stage that can be resolved by using a Bayesian approach. A further advantage of the Bayesian approach is that all three stages can be combined into a single analysis so that it is no longer necessary to conduct any stage conditioning on the results from the previous stage. In this way, we fully propagate all sources of uncertainty into our final inference and avoid the need to conduct conditional analyses based upon what are often ad-hoc procedures for picking between different solutions at each stage.

3. THE BAYESIAN ALTERNATIVE

The problem with Nested Clade Analyses is that by conditioning on the best option at every stage of the process, we lose the information concerning the associated uncertainties. In this section we describe a Bayesian approach which combines all three stages into one so that uncertainty in the underlying tree structure is fully propagated throughout the model and is properly accounted for when assigning individuals to clusters.

We will focus on the phylogeographic analysis here though essentially identical methods may be applied to the phenotypic analysis. In line with the classical NCA we will use Euclidean distance between individuals as basis for clustering which is essentially equivalent to adopting a bivariate normal distribution to describe the spatial distribution of individuals within a cluster. Thus we will use a bivariate normal mixture model in which the clustering of individuals arises from the separation of the underlying tree into disjoint parts. For phenotypic analyses alternative distributions to the normal may be more appropriate.

For the moment, we will assume the simplest case in which the data contains no missing values and there is no homoplasy. In this case the underlying tree is unique and can be determined analytically. The only uncertainty, then, arises from the parameters of the bivariate normal parameters and the partition of the tree into disjoint sets to determine component membership. We will assume here that we do not know how many clusters exist in the data and so use a combination of standard and reversible jump MCMC methods to undertake the analysis. We begin with a brief description of the mixture modelling framework.

3.1. *The Mixture Model*

Suppose we observe n individuals so that our data $\mathcal{D}$ comprises the set $\{(S_i, \boldsymbol{T}_i) : i = 1, ..., n\} = \{\mathcal{S}, \mathcal{T}\}$, say, of sequences and associated traits (in this case geographic locations so that $\boldsymbol{T}_i = (x_i, y_i)$, say). Suppose also that within $\mathcal{D}$ we observe $m \leq n$ distinct haplotypes so that our associated cladogram comprises m nodes $v_1, ..., v_m$ and $m-1$ edges $e_1, ..., e_{m-1}$. Under the assumption of no homoplasy or missing values, the cladogram is a complete connected graph with no loops. Thus, any individual i will be associated with exactly one node $v_{r(i)}$, say, which corresponds to the associated sequence S_i for individual i. Clearly, any node v_j may be associated with more than one individual if and only if $m < n$.

To model the locations, we adopt a mixture of k bivariate normals for the variables $\boldsymbol{T}_i$ in which allowable groupings of the n individuals correspond only to partitions of the graph obtained by deleting $k-1$ edges. That is, deleting any set of $k-1$ edges creates a partition of the graph into k disjoint sub-graphs. If we label the subgraphs $j = 1, ..., k$ then the individuals associated with component j of the mixture are those individuals associated with nodes within subgraph j. We index the partitions by $\boldsymbol{e}(k)$ which denotes a set of k edges chosen from the complete set $\mathcal{E} = \{e_1, ..., e_{m-1}\}$ for deletion. Under partition $\boldsymbol{e}(k)$, we allocate individual i to component $z_i(\boldsymbol{e}(k))$ for all $i = 1, ..., n$. Note that any two individuals with the same haplotype will therefore be assigned to the same component. Finally, for any individual i in component j we assume that $\boldsymbol{T}_i \sim \mathcal{N}_2(\boldsymbol{\mu}_j, \Sigma_j)$ with associated density $f(\boldsymbol{T}_i|\boldsymbol{\mu}_j, \Sigma_j)$, say.

Thus, we obtain a joint probability distribution for the data given the number of components k, the partition $\boldsymbol{e}(k)$ and the set $\mathcal{M}_k$ of component means $\mathcal{V}_k$ of component covariance matrices:

$$f(\mathcal{T}|k, \boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k) = \prod_{i=1}^{n} f(\boldsymbol{T}_i|\boldsymbol{\mu}_{z_i(\boldsymbol{e}(k))}, \Sigma_{z_i(\boldsymbol{e}(k))}).$$

Note that since at this stage we are modelling only the geographic locations, we have here only a probability distribution for the traits $\mathcal{T}$. We shall derive a similar distribution for the sequences $\mathcal{S}$ in the next section.

We shall assume, for the moment, that the value of k is fixed and adopt a uniform prior over all valid partitions, together with independent bivariate normal and inverse Wishart priors for the component mean vectors and covariance matrices respectively. Thus

$$\pi(\boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k|\mathcal{T}, k) \propto f(\mathcal{T}|k, \boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k)p(\boldsymbol{e}(k)) \prod_{i=1}^{k} p(\boldsymbol{\mu}_i, \Sigma_i). \tag{1}$$

3.2. *The MCMC Algorithm*

In order to explore the posterior distribution given in Equation (1) above we use a standard MCMC algorithm comprising a series of Gibbs and Metropolis Hastings updates for the different parameters. Here, we take the following priors for the $\boldsymbol{\mu}_i$ and Σ_i:

$$\begin{aligned} \Sigma_j &\sim \text{InvWishart}(t, \Psi); \text{ and} \\ \boldsymbol{\mu}_j|\Sigma &\sim \mathcal{N}_2\left(\boldsymbol{0}, \frac{1}{\tau}\Sigma_j\right) \end{aligned}$$

Thus, we obtain the following posterior conditional distributions for $j = 1, ..., k$:

$$\begin{aligned} \boldsymbol{\mu}_j|\Sigma_j, \boldsymbol{e}(k), \mathcal{T}, k &\sim \mathcal{N}_2\left(\frac{n_j\overline{\boldsymbol{T}}_j}{n_j+\tau}, \frac{1}{n_j+\tau}\Sigma_j\right); \text{ and} \\ \Sigma_j|\boldsymbol{\mu}_j, \boldsymbol{e}(k), \mathcal{T}, k &\sim \text{InvWishart}\Big(n+t, \Psi + \sum_{i:z_i(\boldsymbol{e}(k))=j} \boldsymbol{T}_i\boldsymbol{T}_i' \\ &\qquad - \big(n_j\overline{\boldsymbol{T}}_j\overline{\boldsymbol{T}}_j' + \frac{n_j\tau}{n_j+\tau}(\overline{\boldsymbol{T}}_j - \boldsymbol{\mu}_j)(\overline{\boldsymbol{T}}_j - \boldsymbol{\mu}_j')\big)\Big), \end{aligned}$$

where n_j denotes the number of individuals assigned to component k (dropping the dependence on $\boldsymbol{e}(k)$ for notational convenience) and $\overline{\boldsymbol{T}}_j$ denotes the vector of all $\boldsymbol{T}$ vectors associated with individuals in cluster j i.e.,

$$\overline{\boldsymbol{T}}_j = \frac{1}{n_j} \sum_{i:z_i(\boldsymbol{e}(k))=j} \boldsymbol{T}_i.$$

Thus, the MCMC algorithm for drawing from the posterior distribution given in Equation (1) proceeds by updating the component means and variances using Gibbs updates, in which new values are drawn from the posterior conditional distributions given above, followed by a move which updates the edge set $\boldsymbol{e}(k)$. Obviously, $\boldsymbol{e}(k)$ is a discrete random variable and so can be updated directly using a Gibbs update. However, for datasets with large numbers of distinct sequences (i.e., large m), a Gibbs update for the edge set can prove extremely computationally expensive due to the fact that the posterior probabilities associated with each of the "$m-1$ choose k" potential edge sets must be calculated. A far more efficient update is obtained by using a Metropolis Hastings procedure in which only the posterior probabilities associated with the existing and proposed new edge set need be calculated.

In practice we found that updating only the edge set led to extremely poor mixing because of the strong posterior association between the edge set and the corresponding component means and covariance. Thus, we propose updating the edge set as follows. Select an edge within the edge set with equal probability and propose to replace it uniformly at random with any other edge within the graph that is not already in the edge set i.e., from the set $\mathcal{E} \setminus \boldsymbol{e}(k)$. We then propose new values for all means in $\mathcal{M}_k$ and covariance matrices in $\mathcal{V}_k$ from the corresponding posterior conditionals under the proposed new edge set. This new edge set, the associated cluster means and covariances are then accepted with the usual Metropolis Hastings acceptance ratio.

This, then, completes the updates required to generate a series of realisations from the posterior distribution given in Equation (1) from which we can obtain empirical estimates for any posterior statistics of interest.

3.3. *Label Switching*

As is common with mixture modelling, our model suffers from the so-called label-switching problem caused by symmetry in the joint probability distribution of the data given the model parameters. This has the practical consequence that the collection of nodes labelled in one iteration of the MCMC algorithm as group j (with associated mean and covariance μ_j and Σ_j) may may be labelled as group $l \neq j$ in the next iteration.

The standard approach to overcoming such difficulties is to introduce an essentially arbitrary identifiability constraint such as ordering the components in terms of the associated component parameters e.g., the mean. This sort of approach was unsatisfactory in the range of examples we considered due, mainly, to the multi-dimensional nature of the mixture distributions and the tendency for there to be considerable overlap between the marginal distributions associated with the bivariate normal components. Ordering based upon distance of the mean vector from a fixed point as well as ordering based upon the properties of the tree (e.g., start at the left of the tree and label the components as you visit them) all failed to produce sensible results.

Thus, we adopt the relabelling algorithm of Stephens (2000) to draw inference about component parameters from the sample of observations obtained via the MCMC algorithm described earlier. Suppose we have a sample of T observations. For any observation t we require a permutation ν_t of the the associated component labels that provides us with a consistent labelling for all $t = 1, ..., T$. Stephens suggests the following iterative algorithm. Given suitable permutations for iterations $1, ..., t$ choose a permutation ν_{t+1} for iteration $t + 1$ that maximises

$$\sum_{i=1}^{n} \log \frac{1}{t} \left(\sum_{j=1}^{t} \mathbb{I}_{c_i^{(j)} = c_i^{(t)}} \right).$$

This, then, provides a series of permutations that assign labels to the different components that leads to “as many nodes as possible belonging to their favourite group so far”.

From a practical perspective, this approach proved extremely reliable with only minimal additional computational complexity. Indeed the labelling process can even be introduced as an additional (deterministic) step within the MCMC algorithm itself.

3.4. *How Many Clusters?*

Finally, we consider the case where the number of clusters in the data is unknown *a priori*. In this case, we must first extend the posterior distribution in Equation (1) to allow for uncertainty in the value of k, which is easily achieved by specifying an appropriate prior $p(k)$ so that

$$\pi(\boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k, k|\mathcal{T}) \propto \pi(\boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k, k|\mathcal{T}, k)p(k). \tag{2}$$

Here, we assume k follows a uniform distribution on the discrete values $0, ..., k_{\max}$ though other priors may be also be appropriate. For example, we may wish to specify a prior which weights each value of k by the reciprocal of the number of possible edge sets available when k edges can be deleted. In any case, appropriate re-weighting of the posterior model probabilities under the uniform prior for k can provide the required the probabilities associated with any other and so we stick with the uniform here. Note here that the priors for $\boldsymbol{e}(k)$, $\mathcal{M}_k$ and $\mathcal{V}_k$ must be normalised since they will be functions of k. For the fixed-k case this was not necessary. The normalisation constants are all easily obtained.

As we have extended our posterior distribution, we must now extend our MCMC algorithm to allow for the additional move required to update the parameter k. As this necessitates changing the number of parameters in the model (i.e., adding or deleting component means and covariances) this requires a reversible jump MCMC (Richardson and Green, 1997) update.

We might begin each model move by first deciding to increase k by splitting an existing cluster or to decrease k by merging two adjacent clusters (where adjacency is determined in terms of the tree). In practice this means that we add or delete an edge to the edge set with probabilities P_{add} and P_{delete} respectively. If we propose to move beyond the $[0, k_{\max}]$ range we automatically reject the proposal. A simpler way is to simply select an edge and if it is already in the current edge set we propose to delete it and if it is not, then we propose to add it. This simplifies the acceptance ratio, but leads to fairly high probabilities for a split move. In practice, we found that this caused us no computational difficulties, but for other problems it may be better to specify fixed split/merge probabilities to gain better control over the algorithm's performance.

Suppose we propose to add a new edge to the edge set, moving from k to $k+1$, then this automatically splits an existing cluster. We therefore need to propose values for the normal parameters associated with the two new clusters obtained by splitting the original. We do this by drawing values from the associated posterior conditionals, conditioning on the proposed new edge set and associated assignments of nodes (individuals) to each cluster. Similarly, when merging two clusters, we draw new normal parameter values from the posterior conditional with individuals from the two original clusters are assigned to the proposed new one. In this way, the Jacobian term required for the reversible jump acceptance ratio reduces to 1. In fact, the entire acceptance ratio therefore reduces to a ratio of appropriate normalisation constants and is easily derived.

Combining this step with the Gibbs and Metropolis Hastings steps described in Section 3.2 we obtain an MCMC algorithm capable of obtaining draws from the posterior distribution described in Equation (2).

Recall that at the beginning of this Section we made the assumption that there was (i) no missing values and (ii) no homoplasy. In practice these two assumptions will rarely be valid. Thus, in practice we will require a number of extensions to our method in order to deal with these two common situations.

4. EXTENDING THE BASIC FRAMEWORK

Here we extend our analysis to allow both for missing values and homoplasy and therefore need to deal with uncertainties in the underlying tree. Note the distinction here with the previous section which essentially assumes that the underlying tree is fixed and focuses only on modelling the traits. In order to assess tree probabilities, we must derive a form of model for the observed sequences $\mathcal{S}$ from which the associated model parameters can be inferred. This requires estimation of the mutation probabilities from the mutations observed to occur in the data which, in turn requires the derivation of a root node which determines the direction of the mutations in the tree.

4.1. *Finding the Root*

Here, we discuss methods for determining the root of the tree, that is the “oldest” haplotype from which the other observed haplotypes can be assumed to have been derived. We can make various assumptions about the properties of the root. For example, we would normally expect the root to be towards the “centre” of the cladogram and that it might often be surrounded by missing nodes representing haplotypes that might now be extinct. It might also be assumed that the root node would tend to be of higher degree than other nodes, though this can be masked by population dynamics such as changing productivity rates. Such information can be used to determine an appropriate prior for the root node. In the absence of suitable data on what constitutes a “central” node, we will assume a flat prior for the root over all nodes (both observed and missing) in the graph.

4.2. *Estimating Mutation Rates*

Having determined the root, the graph now becomes directed and it is possible to estimate the mutation rates. We begin by parameterising the mutation probabilities as follows.

We assume that the nucleotide frequencies are at equilibrium and that the mutation process is time-reversible. We use a Generalised Time-Homogeneous Time-Reversible Markov Process model (Tavaré, 1986) for the mutation rates (which are assumed to be independent and identical across all sites) generated by a Q-matrix:

$$Q = \begin{pmatrix} \cdot & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & \cdot & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & \cdot & \zeta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \zeta\pi_C & \cdot \end{pmatrix}$$

where the π_i's ($i = A,\ G,\ C,\ T$) represent the equilibrium probabilities of the nucleotides (we assume that the gene frequencies are stationary), and $\alpha, \ldots, \zeta$ the relative transition and transversion probabilities, so that we have 9 degrees of freedom. For simplicity we refer to ($\pi_A,\ \pi_G,\ \pi_C,\ \pi_T$) as ($\pi_1,\ \pi_2,\ \pi_3,\ \pi_4$) respectively. This chain is time-reversible since $\pi_i q_{ij} = \pi_j q_{ji},\quad i, j = 1, \ldots, 4$. Here we assume that the generator matrix is identical and independent across all sites, though it would be possible to extend the methods discussed her to relax this assumption.

The Q-matrix above is normalised to obtain the following matrix of mutation probabilities:

$$P = \begin{pmatrix} 0 & c_1\alpha\pi_2 & c_1\beta\pi_3 & c_1\gamma\pi_4 \\ c_2\alpha\pi_1 & 0 & c_2\delta\pi_3 & c_2\epsilon\pi_4 \\ c_3\beta\pi_1 & c_3\delta\pi_2 & 0 & c_3\zeta\pi_4 \\ c_4\gamma\pi_1 & c_4\epsilon\pi_2 & c_4\zeta\pi_3 & 0 \end{pmatrix}$$

where the c_i are simple normalisation constants chosen so that the entries in each row sum to 1.

Since the π_i sum to 1, we adopt a Dirichlet prior for the vector of nucleotide frequencies, $\boldsymbol{\pi}$. Here, we assume that $\boldsymbol{\pi} \sim Dir(B_1, B_2, B_3, B_4)$ where the $B_i \sim \mathcal{N}(1, \sigma_B^2)$. Alternatively, we might adopt the Jeffreys prior in which the B_i are fixed to be equal for $i = 1, ..., 4$. Again, we observed no clear sensitivity to the choice of prior in the primary posterior summary statistics of interest. Under the hierarchical Dirichlet prior, it is simple to show that the posterior conditional for $\boldsymbol{\pi}$ given the other parameters is itself a Dirichlet with parameters $B_i + b_i$, where b_i denotes the number of observed genes of type i within the observed sequences.

For the transition/transversion rates, we assume independent normal priors chosen on the basis of expert information as to the relative probabilities of transitions and transversions so that

$$\alpha, \zeta \sim \mathcal{N}(\mu_s, \sigma_s^2) \text{ and } \beta, \gamma, \delta, \epsilon \sim \mathcal{N}(c\mu_s, \sigma_s^2)$$

for some appropriate values of c, μ_s and σ_s^2. Note also, that the value of c might itself have a prior distribution if the likely transition/transversion ratio is unknown. For the moment, we assume that the prior parameters are fixed and, if $\boldsymbol{\theta}$ denotes the set of transition/transversion rates, we have prior $p(\boldsymbol{\theta}|c, \mu_s, \sigma_s^2)$.

4.3. *Assessing Tree Probabilities*

Now that we have the mutation probabilities, we can assess the probabilities associated with any given tree. In particular, we can use these to determine the state of missing nodes and, more importantly, break loops by deleting appropriate edges.

The basic tree probability is simply the product over all edges of the mutation probability associated with that edge. These probabilities are normalised by the sum over all trees consistent with the data and here we will make the simplifying assumption that the minimum number of missing nodes will be added to ensure completeness of the tree. Suppose that our tree contains η edges, $\mathcal{E} = \{e_1, ..., e_\eta\}$ with probability of mutation $p(e_j)$ for edge e_j derived from the mutation probability matrix given in Section 4.2 and the directions indicated by the root node identified in Section 4.1. Then the probability of the tree with edges $\mathcal{E}$ is given by

$$p(\mathcal{E}|r, \phi, \boldsymbol{\theta}, \eta) \propto \prod_{i=1}^{\eta} p(e_i).$$

Often the state of any missing node (i.e., the sequence with which it is associated) will be uniquely determined but in the presence of homoplasy or where strings of more than one missing nodes occur in the cladogram, the true state of the missing node will be unknown *a priori* and can be estimated. Of course, once the set of edges $\mathcal{E}$ is known, the states of the missing nodes are uniquely determined and so they are automatically estimated as a by-product of the estimation of the non-deterministic edges that we discuss below.

Homoplasy has the effect in our network of creating two or more different edges representing a mutation on the same restriction site. In some cases, loops may be formed and these need to be eradicated in order to form a permissible cladogram. In practice, loops are removed by deleting one or more edges (as few as possible) in much the same way as edges are removed in order to form clusters for the trait variables. The mutation probabilities can be used to weight edges according to the their relative probabilities and so the clustering algorithm described in the previous section can be averaged over the trees obtained by breaking loops in the full tree using the associated probabilities to weight the inference.

An additional complication is that when we wish to undertake a phenotypic analysis, if any edge in the edge set corresponds to a mutation which happens elsewhere in the graph (this is only possible if homoplasy occurs) then all edges corresponding to that mutation must be added to the edge set. Since we are not attempting to make causal associations in the same way for the phylogeographic analyses, this problem can be ignored in this case.

4.4. *The MCMC Algorithm*

The posterior distribution from the previous section is amended so that the distribution of the edge set depends upon the set of edges $\mathcal{E}$ within the tree. We decompose this set of edges into two: the set $\mathcal{E}_S$ of edges in the full graph that can be derived with certainty from the observed sequence data $\mathcal{S}$ i.e., edges only between observed nodes that are not part of a loop within the cladogram; and $\mathcal{E}_0$ which denotes the set of uncertain edges i.e., edges connected to missing nodes and subsets of the edges in the original tree that form loops. Thus, $\mathcal{E}_0 \cup \mathcal{E}_S$ forms a connected tree without loops. Then we amend our posterior in Equation (2) to obtain

$$\begin{aligned}
&\pi(\boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k, \mathcal{E}_0, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{B}, r, k | \mathcal{D}, \mathcal{E}_S) \\
&\quad \propto \quad f(\mathcal{T}|k, \boldsymbol{e}(k), \mathcal{M}_k, \mathcal{V}_k) p(\boldsymbol{e}(k)|\mathcal{E}) p(\mathcal{E}|r, \phi, \boldsymbol{\theta}) \\
&\qquad\quad \times p(\boldsymbol{\theta}) p(\boldsymbol{\pi}|\boldsymbol{B}) p(\boldsymbol{B}) p(r) \prod\nolimits_{i=1}^{k} p(\boldsymbol{\mu}_i, \Sigma_i) p(k)
\end{aligned} \tag{3}$$

The associated MCMC algorithm therefore involves updating the original parameters in much the same way as described in Section 3.2. The additional parameters introduced in this section are updated as follows:

Updating r: Here, we propose a new root randomly from the observed nodes in the graph and accept the move with the standard Metropolis Hastings acceptance probability. An alternative is to propose only neighbouring nodes to the current root. The latter tends to perform better in general but it is possible to get stuck in local modes of the distribution from which it is hard for the chain to escape.

Updating $\boldsymbol{\theta}$: The transition/transversion probabilities are updated using a random walk Metropolis Hastings update.

Updating $\boldsymbol{\pi}$: The nucleotide frequencies are updated via a Gibbs step, since their posterior conditional is Dirichlet.

Updating $\boldsymbol{B}$: The hyperprior parameters for the Dirichlet prior on the nucleotide frequencies are updated using a random walk Metropolis Hastings proposal on the log scale.

Updating $\mathcal{G}_0$: Here, we retain a list of edges involved in loops within the graph and at each iteration we draw uniformly at random a collection of edges that both removes all loops within the graph and which provides a complete tree. These edges are then accepted using the usual Metropolis Hastings acceptance ratio. If we wish to update the state of the missing nodes and, indeed, their associated trait value, these can be sampled directly from their corresponding posterior conditionals.

Now that we have the MCMC methodology required to sample from the posterior of interest, we may begin to analyse data and interpret the results.

5. EXAMPLE: A PHYLOGEOGRAPHIC ANALYSIS

Here, we consider a specific data set and conduct a phylogeographic analysis using the methodology described above. We begin with a description of the data before providing the results and associated interpretation and discussion.

5.1. *The Data*

Here, we focus on data from the geologically young and well-characterised island of La Palma from within the Canary Islands archipelago to generate phylogeographic predictions for *Brachyderes rugatus rugatus*, a flightless curculionid beetle species occurring throughout the island in the forests of *Pinus canariensis.*

The data comprises a sample of 135 beetles from 18 locations across the island and for each we record 570 base pairs of sequence data for the mitochondrial DNA cytochrome oxidae II gene. The data are summarised in Figure 5 which superimposes the sampling locations on a map of La Palma together with forest density. At each location the number of distinct haplotypes observed is recorded, with 69 distinct haplotypes observed in all. See Emerson *et al.* (2000,2006), for example.

Geological studies of the island provide us with a fairly complete understanding of the island's geological history. The northern part of the island is mainly older volcanic terrain with the southern part comprising a ridge of more recent volcanic origin. It is a reasonable assumption that the *Brachyderes* beetle population, with their limited mobility, would have been strongly influenced by La Palma's volcanic and erosional history and there is also strong evidence that the population was seeded by immigration from the nearby island of Tenerife to the East. Thus we might expect the oldest haplotypes to be concentrated in the northern part of the island and to observe evidence of a more recent range expansion to the southern tip. If we were to cluster the haplotypes geographically we should therefore find that those to the North should be more central to the haplotype network, and those to the south should be placed towards the tips of the network.

5.2. *The Analysis*

Here, we apply the methodology described in the previous sections to the beetle data.

Table 1: *Posterior probabilities for different numbers of clusters*

No. Clusters	2	3	4	5	6
Post. Model Prob.	0.02	0.13	0.24	0.36	0.25

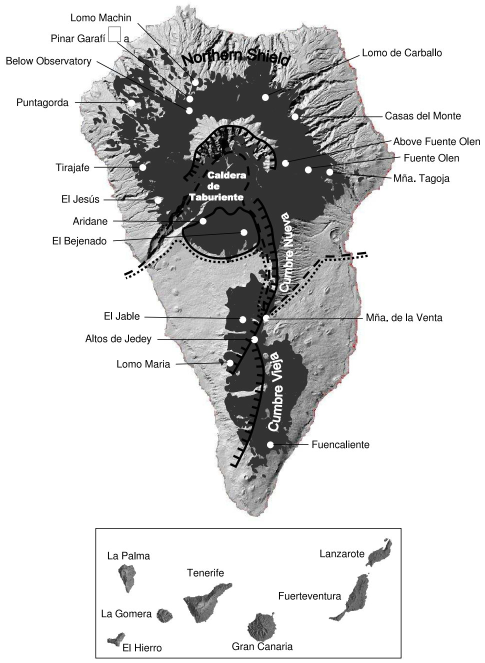

Figure 5: *Map of La Palma with sampling locations and tree density super-imposed.*

Table 1 provides the posterior model probabilities associated with models containing between 2 and 6 clusters. Clearly, the 5-cluster model is *a posteriori* most probable, but there is little to choose between many of the models under consideration.

Figure 6 provides the maximum *a posteriori* (MAP) estimate of the underlying cladogram, together with an indication of the clusters to which each node belongs. We can see that the assignment of nodes to specific clusters is certain *a posteriori* for many nodes, except a few towards the top of the cladogram which are most likely associated with the green cluster but also have non-zero probability of being allocated to the red one. Perhaps most intriguingly there are no nodes that are assigned to the yellow cluster with absolute certainty with most nodes being assigned with considerable probability to the blue cluster if they have any probability of being assigned to the yellow. At first sight, this may seem strange especially when you consider that in general we would expect posterior certainty for the yellow cluster to increase as we move closer to the leaf node towards the bottom left-hand corner of the cladogram. The reason for this is that there is considerable uncertainty as the underlying cladogram structure in this region with relatively substantial probability that the dotted edge given in Figure 6 might replace one of the two dashed lines in the graph. The uncertainty here arises from two distinct ways to delete a loop in the original graph and this clearly highlights the value of incorporating the uncertainty of the underlying graph into the analysis rather than simply conditioning on the most likely graph determined from a separate analysis.

The *a posteriori* most probable root node under the flat prior is the single red node of degree 5, though there is considerable uncertainty as to the exact location with the second most likely root node attracting 80% of the posterior probability associated with the first. The red cluster is the *a posteriori* most probable to contain the root node.

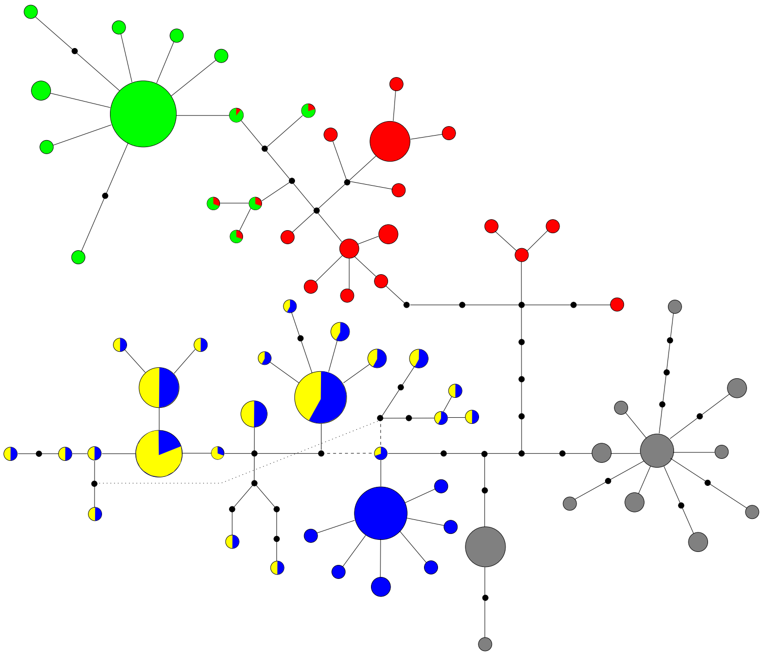

Figure 6: *The MAP estimate of the underlying cladogram. Each node is represented by a pie chart indicating the proportion of times it is assigned to each of the 5 clusters in the MAP model. The dotted line indicates the next most probable edge that could be added and the two dashed lines indicate the two edges either of which would be deleted in order to remove the corresponding loop in the original graph.*

To get a feel for the spatial structure of the haplotype dispersal, Figure 7 provides a plot of the contours of the bivariate normal distributions (evaluated at the posterior means of the mean and covariance parameters) associated with each component of the mixture. Here, we use the same colour scheme as for the cladogram in Figure 7, for comparison. It is clear the the identification of the three clusters at the North of the island could not have been possible without the additional information provided by the underlying cladogram. The individual locations are marked on the figure as black circles, with the largest circle corresponding to the location most likely *a posteriori* to contain the root node under a flat prior. This is located to the

East of the island in the so-called northern shield at Mna Tagoja, though the neighbouring locations around Fuente Olen also have substantial posterior probability of containing the root node.

We can see a clear distinction between clusters in terms of the locations, spread and associated covariance structures. With some clusters, such as the red cluster, being roughly circular and with only local spread, whereas others, such as the grey cluster, has much greater spread in one direction (North-South) than the other. These latter shapes are a result of clusters containing haplotypes spread over a wide area and typically in a specific direction. This type of pattern is consistent with range expansion in the population, whereas the more symmetric and focused clusters would be more indicative of a geographically stable population.

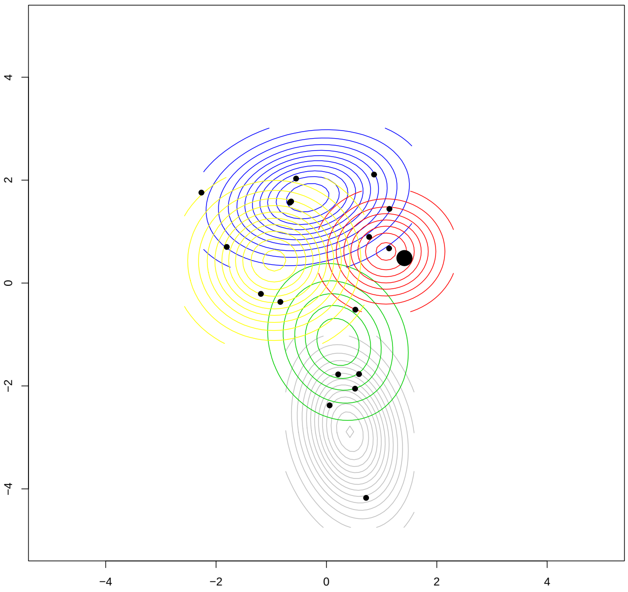

Figure 7: *Bivariate normal contour plots for the five components under the MAP model and evaluated at the posterior means. Each component is assigned the same colour as in Figure 6. The circles indicate sampling locations, with the larger circle indicating the location most likely to include the root node in the cladogram under a flat prior.*

6. DISCUSSION

The results provided in the previous section consistently point to the red cluster as being the population source as it both most likely to contain the root node and exhibits a suitably stable spatial structure. The MAP cladogram suggests a comparatively recent range expansion South East to the Cumbre Nueva terrain running vertically down the southern half of the island from the centre. We infer the recency of the expansion from the lack of missing nodes (indicating both missed samples and haplotype extinctions) between the two clusters in the cladogram. The largest green node is clearly a source for local expansion due to its high degree.

Less recently, the red cluster appears to have been the source for two major expansions one much further to the South resulting in the grey cluster and one to the North West resulting in first the blue and then yellow clusters. This, again, is inferred from the structure of the cladogram in Figure 6 and the spatial structure apparent in Figure 7.

These conclusions are consistent with complementary evidence and previous analyses of similar data. For example, La Palma is at the western end of the Canary Islands and so any colonisation of the island is most likely to have occurred along the Eastern coast. Previous studies of these data by Emerson *et al.* (2006) drew very similar conclusions in that the island was colonised in the mid-east and saw three range expansions, two to the South and one to the West. Indeed the MAP cladogram in Figure 6 is fairly similar to their cladogram C, though quite distinct structurally from the other two potential cladograms that they discuss. Emerson *et al.* (2006) hypothesised that the observed range expansions followed similar expansions in the host species *Pinus canariensis* and it would be interesting to conduct a complementary phylogeographic study to determine the extent of any such association. The only major discrepancy between our results and those of Emerson *et al* (2006) concern the colonisation of the Northern Shield. Our analysis suggests that the Northern part of the island was colonised first by a movement to the North West and then South West down the coast, essentially moving in an anti-clockwise direction. Emerson *et al.* (2006) on the other hand suggest that the colonisation occurred first to the West and then to the North East, moving in a clockwise direction.

Our analysis has several advantages over the classical nested clade analysis. Perhaps the most important is that when it comes to making inference about the underlying dynamics of the population over time, our approach is able to provide quantitative measures of statistics of interest. For example, we are able to associate probabilities with any one location as being the source population for later colonisation of the island. The second advantage is that by combining all three stages of the analysis we are able to make sure that, for example, information from the geographical location of the individuals is allowed to feedback and inform the associated genealogy represented by the cladogram. It is clear that certain cladograms may lead only to nonsensical clusterings, but these can be avoided in the integrated analysis. This did not cause any problems for these particular data but, for others, the ability to incorporate all information into any one component of the analysis is bound to be a distinct advantage.

It is worth pointing out here that there are several ways in which this work can and should be further extended. In constructing our original graph we allowed only the minimum number of missing nodes necessary to complete the graph. It would be interesting to relax this assumption so that an even greater range of cladograms is permissible. In addition, it would be useful to allow for recombinations with the population. These won't have occurred in these particular data, because we are using mitochondrial DNA but other datasets may well exhibit both recombinations and homoplasy. Adding the possibility of recombinations significantly adds to the complexity of the graph construction, and various assumptions (based primarily on parsimony) will be required to make the problem tractable. Finally, it would be very useful to determine a quantitative measure along the lines of Templeton's inference key so that we could actually associate (posterior) probabilities with different clusters being associated with different forms of dispersal over evolutionary timescales. Each of these tasks are the focus of current work by the authors.

REFERENCES

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407–415.

Cassens, I., Mardulyn, P. and Milinkovitch, M. C. (2005). Evaluating intraspecific "network" construction methods using simulated sequence data: Do existing algorithms outperform the global maximum parsimony approach? *Systematic Biology* **54**, 363–372.

Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: A computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657–1659.

Clement, M., Snell, Q., Walker, P., Posada, D., and Crandall, K. (2002). TCS: estimating gene genealogies. *International Workshop on High Performance Computational Biology.*

Crandall, K. A. and Templeton, A. R. (1993). Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* **134**, 959–969.

Emerson, B. C., Forgie, S., Goodacre, S. and Oromi, P. (2006). Testing phylogeographic predictions on an active volcano island: *Brachyderes rugatus* (Coleoptera: Curculionidae) on La Palma (Canary Islands). *Molecular Ecology* **15**, 449–458.

Emerson, B. C. and Hewitt, G. M. (2005). Phylogeography. *Current Biology* **15**, 369–371.

Emerson, B. C., Oromi, P. and Hewitt, G. M. (2000). Colonization and diversification of the species *Brachyderes Rugatus* (Coleoptera) on the Canary Islands: Evidence from mitochondrial DNA COII gene sequences. *Evolution* **54**, 911–923.

Felsenstein, J. (1992). Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution* **46**, 159–173.

Gomez-Zurita, J., Petitpierre, E. and Juan, C. (2000). Nested cladistic analysis, phylogeography and speciation in the *Timarcha goettingensis* complex (Coleoptera, Chrysomelidae). *Molecular Ecology* **9**, 557–570.

Huelsenbeck, J. P., Larget, B., Miller, R. E. and Ronquist, F. (2002). Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic Biology* **51**, 673–688.

Ibrahim, K. M., Nichols, R. A. and Hewitt, G. M. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity* **77**, 282–291.

Larget, B., and Simon, D. (1999). Markov Chain Monte Carlo Algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* **16**, 750–759.

Lloyd, D. G., and Calder, V. L. (1991). Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses *J. Evolutionary Biology.* **4**, 9–21.

Mau, B., Newton, M., and Larget, B. (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55**, 1-12.

Newton, M. A., Mau, B., and Larget, B. (1999). Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. *Statistics in Molecular Biology and Genetics.* (F. Seillier-Moseiwitch, ed.) IMS Lecture Notes-Monograph Series, **33**, 143–162.

Posada, D. and Crandall, K. A. (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution*, **16**, 37-45.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–792.

Stamatakis, A., Ludwig, T. and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463.

Stephens, M. (2000). Dealing with label-switching in mixture models. *J. Roy. Statist. Soc. B* **62**, 795–809.

Tavaré, S. (1986). *Some Probabilistic and Statistical Problems in the Analysis of DNA sequences.* Lectures on Mathematics in the Life Sciences 17. American Mathematical Society.

Templeton, A. R. (1995). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and apoprotein E locus. *Genetics* **140**, 403–409.

Templeton, A. R. (1998) Nested clade analysis of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology* **7**, 381-397.

Templeton, A. R., Boerwinkle, E., and Sing, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila. Genetics* **117**, 343–351.

Templeton, A. R., Crandall, K. A., and Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. III. Cladogram estimation.*Genetics* **132**, 619-633.

Templeton, A. R., Routman, E. and Phillips, C (1995). Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger Salamander, *Ambystoma tigrinum. Genetics* **140**, 767–782.

Templeton, A. R., and Sing, C. F. (1993). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134**, 659–669.

Templeton, A. R., Sing, C. F., Kessling, A. and Humphries, S. (1988). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics* **120**, 1145–1154.

DISCUSSION

BANI K. MALLICK (*Texas A&M University, USA*)

I congratulate the authors on their very fine paper. It contains novel methodology emerged from a very interesting application. Process of grouping a collection of objects into subsets or clusters such that elements within each cluster are more closely related to one another than objects assigned to different clusters. This clustering has been done using some type of similarity measure (usually measuring the Euclidean distance between the data). In this particular problem there are additional information like DNA sequence data which makes it a constrained clustering problem. Cladograms are trees or graphs that are helpful in relating individuals from a population(s) given their sequence data. Partitioning such trees may allow us to associate specific genetic mutations with differences within the studied population. The authors have proposed novel Bayesian framework for creating disjoint partitions of cladograms.

There are two components of the data: S: Sequences, T: Traits. A Bayesian model will specify a joint distribution for $p(S,T)$ It has been done conveniently by individually modeling $p(T|S)$ and $P(S)$. Mixture model has been used for $p(T|S)$ as well as extra uncertainties have been added by modeling S. I have some comments about the general methodology.

Hierarchical Clustering?: Hierarchical clustering develops a tree (dendrogram) whose leaves are the data points and whose internal nodes represent nested clusters of various sizes. The nested cladogram method by Templeton et al. (1987) looks like very similar to hierarchical clustering method. Due to a nice structure in the data through the cladogram is hierarchal clustering is more scientifically interpretable than the random clustering method proposed here?

Bayesian hierarchical clustering is of recent interest (Heard et al., 2006; Heller and Ghahramani). In this problemt to calculate the normalizing constant a determinant of the matrix

$$\phi^{-1} + S_j + \frac{\tau . n_j}{(\tau + n_j)} \bar{T}_j \bar{T}_j{}'$$

has to be calculated where S_j is the sample variance-covariance matrix of the jth cluster. This could be time consuming specially when the dimension of the trait is very high. In a hierarchal clustering framework it is simple to find an iterative relationship between S (or $\bar{T}\bar{T}'$) and their children so you do not need to recalculate these quantities at each iteration. This way it could lead to considerable savings in computation time specially when the dimension is high.

Specification of Prior distributions: As it is an applied problem, I was expecting more subjective priors in this analysis. Also prior sensitivity needs to be verified. How would the clustering structures change if you change the prior for k from uniform to other distributions which will penalize too many clusters? If there are no strong subjective priors available then you can exploit model-based clustering methods using maximum likelihood estimation via EM algorithm (McLachlan and Peel, 2000). That way you can avoid the problem of specifying hyper priors and related prior sensitivities.

Mixture model vs DPP based clustering: An alternative way to produce clusters consist of using Dirichlet process prior (DPP) based models. An advantage of this method is the number of clusters has been determined automatically. Is it possible to use DPP based method in this setup?

Clustering of graphs: Graph cutting, partitioning or clustering as it is variously known is the decomposition of a graph into roughly equal sized pieces while minimizing the number of edges between those pieces (Chung, 1997). Looks like the method developed in this paper is closely related to clustering of graphs. The main difference is in this problem the cladogram (the dependence structure) has been provided or when there is some uncertainty in the tree structure still biologically interpretable mutation models are helpful to rebuild the structure. In graph clustering problem the dependence structure is completely unknown, so is it possible to model the dependence structure as well as the clustering simultaneously? This will be very useful to develop gene regulatory networks. To understand the nature of the cellular function, it is necessary to study the behavior of the genes in a holistic rather than individual way. Gene networks can be developed based on gene expression data where the dependence structure among the genes are completely unknown. With very large number of genes, it is believed that the whole gene network can be partitioned into small sub-networks. This is same as clustering the network model in small pieces. Graph cutting is a popular tool (Chung, 1997) where you adaptively create the graph as well as cluster it in sub-graphs. Recently it has been shown that the graph cutting algorithm is equivalent to kernel k-means clustering algorithm (Dhillon et al., 2003). Existing Graph cutting techniques fail to provide uncertainty measures. In a recent paper Ray, Mallick, Dougherty (2006) developed a Bayesian graphical model with DPP priors to obtain the posterior distribution of clusters. The method developed by Brooks et al. is a novel way to produce clusters using the the set of edges and it will be interesting to see how this can be extended in the complex situations to identify the clusters within an adaptive graph model.

PAUL FEARNHEAD (*Lancaster University, UK*)

I would like to congratulate the authors for demonstrating how Bayesian methods can be used to coherently address challenging problems in statistical genetics; and the advantages they have over more classical approaches which split the analysis into a number of stages and that essentially assume that the inferences within each preceeding stage are exact.

The work implicitly assumes no recombination - so that there is a unique genealogical tree that describes the historical relationship of the sample; and this is appropriate for the phylogeographic analysis considered as mitochondrial DNA does not recombine. However, I wonder if you have considered how to extend your method to situations where there is recombination? As most DNA data comes from recombining regions of the genome, extending the method to this case is of particular importance, especially for possible applications relating to mapping disease genes.

In these cases there is no single genealogical tree, but (potentially) different genealogical trees for each site in the DNA sequence - this information can be described by something called the Ancestral Recombination Graph (ARG). Applying phylogenetic methods to this case is inappropriate - while a tree could be inferred, it no longer has the natural interpretation that there is in the no recombination case, and ignoring recombination has shown to lead to biases in various situations (Sheirup and Hein 2000). In theory the Bayesian approach can be adapted to this case, by replacing mixing over possible trees with mixing over ARGs, but in practice this is likely to be computationally infeasible due to the dimension of the space of ARGs.

On a related point, there is a substantial literature within the field of population genetics that is relevant to this work. In particular there is a natural population genetic model for the underlying tree (or ARG) for the data based around a stochastic process called the coalescent (Kingman 1982). There are numerous methods for inference under coalescent models (see for example Stephens and Donnelly 2000, Drummond *et al.* 2002 and references therein), some of which can include geographic information (Griffiths and Bahlo 1998, Beerli and Felsenstein 1999). While these are primarily only for non-recombining DNA, there are a few methods for the recombining case (e.g. Fearnhead and Donnelly 2001). There are a number of advantages that these coalescent-based methods have over the more phylogenetic approach described in the paper; for example they would enable trees to be inferred without assuming parsimony, better estimates of the root of the tree, and inferences about times on the tree (for example to date possible migration events). Though these are likely to come at a higher computational cost.

Also of relevance are various accurate approximate methods that have been devised in population genetics. The most generic of which is the Product of Approximate Conditional Likelihood method of Li and Stephens (2003) - see also Wilson and McVean (2006) for an application. This framework may give computationally feasible way of extending the work in this paper to the recombining DNA case. There is also related work by Zöllner and Pritchard (2005) on methods for mapping disease genes; and by Pritchard *et al.* (2000) and Falush *et al.* (2003) on detecting population structure from genetic data.

REPLY TO THE DISCUSSION

We are very grateful to Bani Mallick for his thoughtful discussion of our paper. He makes a number of interesting observations to which we respond below.

Hierarchical Clustering. Calculation of the normalisation constant with our bivariate normal clusters is a trivial exercise in the context of phylogeographic analyses. However, for phenotypic analyses higher dimensional distributions may well be appropriate in which case techniques such as those described may well be very useful. That being said, it is not clear that the normal assumption will necessarily be valid. In many cases, phenotypic traits may well be binary, noting the presence/absence of disease for example, and so entirely different distributions would need to be used.

Priors. We use vague priors in this analysis to demonstrate the ability of the method to extract information from the data that is consistent with what is already known. This paper represents a very early foray into this area and we are keen to demonstrate the utility of our method at this early stage. To a large degree, the results of our analysis are consistent with expert opinion. As a demonstration of the method's value the message would be less clear if we adopted more informative priors only to gain a posterior which then agreed with them. A more statistically rigorous defense of our choice of vague priors is based on the observation that the knowledge we already have about the model parameters (and, indeed, model probabilities) is predominantly based on previous analyses of these same data and so would be inappropriate for use as a prior. Of course, there is some independent information from studies on related species or on nearby islands, but it's very difficult to disentangle these from knowledge based upon earlier analyses of these data. Current research undertaking analyses of new data adopts a more rigorous approach to this issue using reference priors on occasion and carefully sourcing information for more informative priors where both possible and appropriate.

We were intrigued by the suggestion that in the absence of informative prior information, we should resort to a classical analysis as a means to "avoid the problem of specifying hyper-priors and related prior sensitivities". Ignoring any philosophical objections, it is worth noting that the EM algorithm would necessarily have to ignore the uncertainty in the underlying tree, converging to that with the highest likelihood. Therefore, such an algorithm would offer little advance over traditional methods. The Bayesian method, on the other hand, can incorporate this uncertainty and allow it to propogate through to the clustering component of the analysis.

Mixtures vs DPP. It may well be possible to produce an alternative scheme based upon DPP models. This is not something that we have explored but may well be worth investigating in due course. Our proposed method determines the number of clusters automatically too, so it is unclear to us the extent to which DPP-based models would provide an improvement as opposed to a simple alternative at this stage. It may, nonetheless, be worth exploring in more detail.

Clustering Graphs. We had not thought to link our work with the literature on partitioning graphs. The similarity measures in the latter case are functions of the graph structure itself, whereas in our case, they are functions of trait values associated with the relevant nodes. It is certainly possible that some cross-fertilisation of ideas might benefit one or other of the two areas and we will explore this as we continue our research.

Paul Fearnhead also kindly contributes to the discussion by providing a tutorial overview of a few related ideas together with additional references that augment our own. In particular, he discusses the problem in which recombinations may occur. This is not appropriate for the data we discuss here but may be appropriate for other datasets and readers may well find these additional references useful.

ADDITIONAL REFERENCES IN THE DISCUSSION

Bahlo, M. and Griffiths, R. C. (1998). Inference from gene trees in a subdivided population. *Theor. Pop. Biol.* **57**, 79–95.

Beerli, P. and Felsenstein, J. (1999). Maximum-Likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152** 763–773.

Chung, F. R. K. (1997). *Spectral Graph Theory*, CBMS Lecture Notes, AMS Publication.

Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means, spectral clustering and normalized cuts. *Proc. 10th ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining (KDD)*, 551–556.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. and Solomon W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320.

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

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without Dirichlet process. *Scandinavian J. Statist.* **28**, 355–375.

Heard, N., Holmes, C. and Stephens, D. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.* **473**, 18–29.

Heller, K. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. *Proceedings of the 22nd International conference on machine learning.*

Kingman, J. F. C. (1982). The coalescent. *Stoch. Proc. and App.* **13**, 235–248.

Li, N. and Stephens, M. (2003). Modelling LD, and identifying recombination hotspots from SNP data. *Genetics* **165**, 2213–2233.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*, New York: Wiley

Pritchard, J.K̃., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. , *J. Roy. Statist. Soc. B* **68**, 305–320.

Schierup, M. H. and Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics (with discussion). *J. Roy. Statist. Soc. B* **62**, 605–655.

Ray, S., Mallick, B. and Dougherty, E. (2006). Bayesian graph cutting. *Tech. Rep.*, Texas A&M University..

Wilson, D. J. and McVean, G. A. T. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**, 1411–1425.

Zöllner, S. and Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092.