A Bayesian Approach to Nested Clade Analyses

Ioanna Manolopoulou

Statistical Laboratory, University of Cambridge

February 20, 2008

Motivation:

Two types of data, **phenotypic** and **phylogeographic**.

(a) **Phenotypic data**. We have aligned DNA sequences and phenotypic measurements from a set of individuals.

Individual	DNA seq	trait
A	ATCGA	1.3
B	ATCGA	1.8
C	ATCCA	2.9
D	CTTGA	3.1
E	CTGAG	1.1

Want to associate a specific mutation on a specific nucleotide position with a significant change in the characteristic trait. In most cases no single SNP (Single Nucleotide Polymorphism) mutation can be identified as having a significant effect, but sometimes possible.

(b) **Phylogeographic data** Testing geographical history.
Have aligned DNA sequences from set of individuals and phenotypic measurements.

Individual	DNA seq	geographical location
A	ATCGA	(1.3, 2.5)
B	ATCGA	(1.7, 3.9)
C	ATCCA	(2.9, 0.1)
D	CTTGA	(3.1, 6.1)
E	CTGAG	(1.3, 2.5)

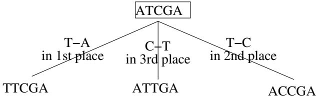
Want to split our data into significant clusters in order to draw conclusions about the geographical history, i.e. range expansion (inc. colonisation), past fragmentation and restricted gene flow (e.g. by isolation-by-distance).

The Evolution of DNA Sequences

- All individuals have a very long DNA sequence of A,T,C,G.
- Precise mutations and oldest sequence not clear from data.
- Not clear that the mutations which appear as Single Nucleotide Polymorphisms (SNP) indeed occurred in a parsimonious fashion.
- 3 sources of information:
  - the distinct sequences which appear,
  - the number of times each sequence is observed,
  - the phenotypic/geographical measurements.

Example:

Assume initially one sequence present, ATCGA, which then mutated to TTCGA, ATTGA and ACCGA. This process yields the graph, where each sequence may be observed more than once in the data.



The resulting data would be (ATCGA, ATTGA, ACCGA, TTCGA).

The Coalescent Process

Two processes evolving simultaneously:

- The splitting process, i.e. replication of sequences, which occurs at rate w and preserves the stationary distribution.
- The mutation process, typically with a much slower rate. We have l independent parallel Markov Processes (l length of sequences), 4 states: A, G, C, T, with Q-matrix:

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

where j represents the nucleotide position, to account for variable mutation rates across the sequence, and π is the stationary distribution. The mutation process also preserves stationarity and is time-reversible.

Parsimony & Missing Sequences

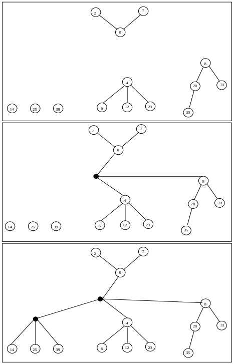
- Parsimony assumption: 1 SNP apart ⇒ 1 mutation apart
- Missing Sequences: sequences are not pairwise 1 SNP apart

Various possible algorithms for finding possible true trees:

(a) Insert minimal missing nodes so that they join most disconnected groups iteratively

(b) Insert minimal missing nodes so that they join the most disconnected groups of the ones closest together.

(c) Insert any number of nodes to connect groups together, taking the minimal path for each combination.



Homoplasy

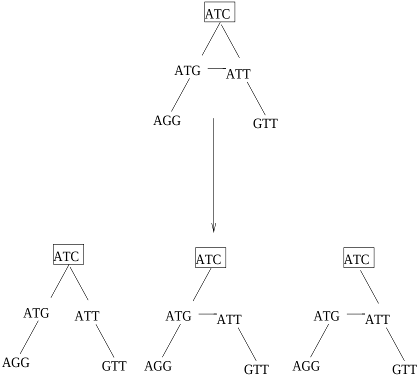
Homoplasy occurs when 2 or more mutations occur at the same nucleotide position. It can either be a mutation which is subsequently "reversed", or a nucleotide mutating to a different one and then to a 3rd one.

Consequences:

(a) Uncertainty in true underlying tree, even assuming parsimony.

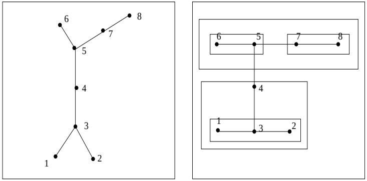
(b) Cannot classify individuals according to which nucleotide they have at a position.

Necessary to have measure of mutation probabilities in order to decide between possible trees.



Nested Clade Analyses (NCA)

Split data into nested groups. Perform ANOVA for each level of nesting, taking branches as groups.



- In the case of phenotypic data, the nested clades are cumulatively tested for significance by a standard ANOVA.
- In the case of phylogeographic data, the clade distance is compared with the geographical distance, i.e. the spread of each cluster compared to the next connected node of that cluster.

Problems

- Allows little uncertainty for parameters: local optima are not necessarily globally optimum.
- Nested ANOVA does not always give answer to our problem.

The significant mutation model

A specific mutation (or more than one) has a significant effect on the characteristic trait in question:

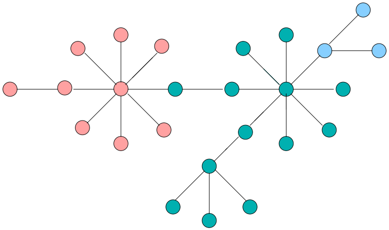


Figure: Example of a significant mutation model gene genealogy

This is equivalent to the ANOVA of the characteristic trait between clades followed in NCA.

The colonisation model

An individual with a specific sequence colonises to a different location, causing all of its descendants to have a significantly different location to its ancestors. The colonising sequence will exist in both locations, and hence its location will have a mixture distribution.

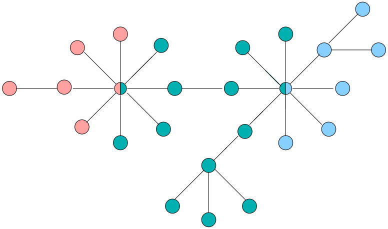


Figure: Example of a colonisation model gene genealogy

Equivalent to comparing the spread of a clade with the geographical position of the next sequence up, in this case the "colonising" sequence.

Probability of a Tree

- Knowing the parameters and the **exact order** of the mutation and splitting processes, can calculate the probability of a tree.
- Of course, millions of different orders result in the same tree.
- We need to either add the order of events as a parameter or sum over all possibilities (computationally infeasible).

Knowing the order, for a mutation and split respectively we check that:

$$\begin{aligned}\mathbb{P}(i\text{th site mutates first from state } i \text{ to } i') &= \frac{q_{i_0 i_t'}}{\sum n_j w + \sum q_j} \\ \mathbb{P}(\text{sequence } i \text{ splits}) &= \frac{n_i w}{\sum n_j w + \sum q_j}\end{aligned}$$

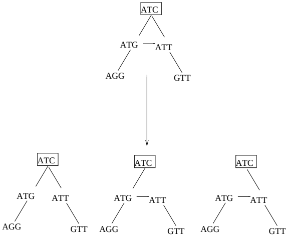
So the total probability of the tree given the root r will be given by

$$\mathbb{P}(T_i|r) = \frac{\mathbb{P}(\mathrm{r}) \prod \mathbb{P}(\text{each mutation/split})}{\sum \mathbb{P}(T_j|r)}$$

Missing Intermediate Sequences

There
are lots of approaches we can take to address the issue of missing intermediate sequences.

- Taking into
account both minimal and non-minimal trees, construct all "sensible" trees. These can subsequently be represented on a single network, where the possible trees are all subtrees of the graph.

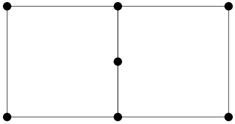


- Finding the tree(s) with a minimal number of edges is straightforward, and we relax the minimality assumption by allowing the tree to have at most, say, l extra edges (i.e. mutations).
- The above process yields a huge number of trees, often of the order 10^7 of higher.

Missing Intermediate Sequences

We need to find a way
to label and represent all possible tree topologies.

- A tree with n nodes has precisely $n-1$ edges
- In a network with n nodes and $n-1+k$ edges, we can always identify k loops (not necessarily uniquely) so that every subtree of the network can be obtained by deleting an edge from each of the k loops.



- Having fixed the k loops, the representation of each subtree in terms of deleting edges from loops is unique, so each tree is represented by a vector of length k.
- Each vector can be represented by a single number using a standard *hashing* algorithm.

Missing Sequences of the form o-·-·-o

Often we have two sequences which are 2 or more mutations apart, with no other sequences being involved in those mutations. This results to a something like this:

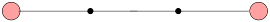


Figure: Example of missing sequences of degree 2

- We know which 3 mutations occurred to get from one sequence to the other
- We don't know the order in which they occurred.
- This does not affect the tree topology.
- It does affect inference about the mutation parameters.
- It does affect inference about the root of the tree.

Priors and Distributions

$$\begin{aligned}
\mathbf{n} &\sim \mathcal{M}ultinomial(N, \boldsymbol{\pi}) \\
(\pi_A, \pi_G, \pi_C, \pi_T) &\sim \mathcal{D}irichlet(B_1, B_2, B_3, B_4) \\
\alpha,\ \zeta &\sim \mathcal{N}(bias, \sigma_m^2) \quad \text{independent} \\
\beta,\ \gamma,\ \delta,\ \epsilon &\sim \mathcal{N}(1, \sigma_m^2) \quad \text{independent} \\
D|\boldsymbol{\Sigma}, \boldsymbol{\mu}, \mathbf{e} &\sim \mathcal{MVN}(\boldsymbol{\mu}_i, \Sigma_i) \ \text{independent} \\
\Sigma_i &\sim \mathcal{IW}(m, \Psi) \\
\boldsymbol{\mu}_i|\Sigma_i &\sim \mathcal{MVN}\left(\mathbf{0}, \frac{1}{\tau_{prior}}\Sigma_i\right) \\
e_i &\sim \mathcal{U}\{1, \ldots, n-1\} \ \text{w/o replacement}
\end{aligned}$$

Exact values of the parameters will depend on species and DNA loci.

Markov Chain Monte Carlo (MCMC) approach

Idea: inference about all the parameters in the model simultaneously. Estimate posterior distribution of all underlying parameters and find global optimum. We construct a Markov Chain s.t.:

1. Given that we are in a position $\boldsymbol{\theta}_t \in \Theta$ where $\boldsymbol{\theta}$ is the vector of parameters and Θ is the parameter space, we propose to jump to a value $\boldsymbol{\theta}_{t+1}$ according to a proposal distribution $q(\boldsymbol{\theta}, \boldsymbol{\theta'})$.
2. Accept this proposed move with probability $\min(1, A)$, otherwise set $\boldsymbol{\theta}_{\mathbf{t+1}} = \boldsymbol{\theta}_{\mathbf{t}}$
$$A = \frac{\mathbb{P}(data|\boldsymbol{\theta}_{\mathbf{t+1}})p(\boldsymbol{\theta}_{\mathbf{t+1}})q(\boldsymbol{\theta}_{\mathbf{t+1}}, \boldsymbol{\theta}_{\mathbf{t}})}{\mathbb{P}(data|\boldsymbol{\theta}_{\mathbf{t}})p(\boldsymbol{\theta}_{\mathbf{t}})q(\boldsymbol{\theta}_{\mathbf{t}}\ , \boldsymbol{\theta}_{\mathbf{t+1}})}$$

The chain described above (provided it is irreducible) is reversible and its equilibrium distribution is equal to the posterior distribution of our parameters given the data.
If the dimension of $\boldsymbol{\theta}_{\mathbf{t}}$ and $\boldsymbol{\theta}_{\mathbf{t+1}}$ is not the same, then the acceptance probability is multiplied by $||J||$, the Jacobian of the transformation from one space to another.

Updating the Mutation Parameters

The nucleotide probabilities and mutation parameters can be easily updated by using the probability of the tree, and the probability of accepting the proposed move becomes:

$$\frac{q((\phi',\pi',\alpha')\to(\phi,\pi,\alpha))}{q((\phi,\pi,\alpha)\to(\phi',\pi',\alpha'))}\times\frac{p(\phi',\pi',\alpha')}{p(\phi,\pi,\alpha)}\times\frac{\mathbb{P}(T|r,\phi',\pi',\alpha')}{\mathbb{P}(T|r,\phi,\pi,\alpha)}$$

Updating the Tree Topology

Similarly, the tree topology T can be updated easily knowing the root and mutation parameters. The acceptance probability becomes

$$\frac{q(T'\to T)}{q(T\to T')}\times\frac{p(T')}{p(T)}\times\frac{\mathbb{P}(T'|r,\phi,\pi,\alpha)}{\mathbb{P}(T|r,\phi,\pi,\alpha)}$$

What is the distribution of the root given the data?

Intuitively, can see that the root is more likely to be in the “middle” of the tree. We know the distribution of the tree given the root, but what is the actual distribution of the root?

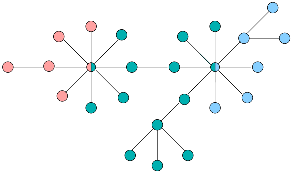


Figure: Example of a colonisation model gene genealogy

One measure of the likelihood of a node being the root would be to count the number of different orders of events in which that root would have given the assumed tree topology. Although this generally works quite well, it is computationally infeasible to calculate all possible combinations, and an approximation needs to be taken.

Clusters and their means and variances

- For a fixed number of significant mutations (or colonising sequences) K, we want to find $K+1$ means
- For significant mutation model, all observations from the same sequence belong to the same cluster
- For the colonisation model, this need not be true, and we need an indicator allocation variable z_i for sequences which are assumed to have colonised.
- This greatly increases the size of the parameter space, and so efficient proposals are essential.

We propose a new covariance matrix Σ using the posterior distribution of Σ if the data only is known:

$$\Sigma_i | D, \mathbf{e} \sim \mathcal{IW}\left(n+m, \Psi + \sum_{c(j)=i} \mathbf{x}_j \mathbf{x}_j^T + n_i \overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^T\right)$$

Using that we generate new means from the posterior for $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}_i | \Sigma_i, D \sim \mathcal{MVN}\left(\frac{n_i \overline{\mathbf{x}}_i}{n_i + \tau_{prior}}, \frac{1}{n_i + \tau_{prior}} \Sigma_i\right)$$

We accept/reject this proposal and then generate new values for the covariance and means from the posterior:

$$\begin{aligned}\Sigma_i | D, \mathbf{e}, \boldsymbol{\mu} \quad \sim \quad & \mathcal{IW}\Big(n + m, \Psi + \sum_{c(j)=i} \mathbf{x}_j \mathbf{x}_j^T \\ & - \left(n_i \overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^T + \frac{n_i \tau_{prior}}{n_i + \tau_{prior}} (\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i)(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i)^T\right)\Big)\end{aligned}$$

If the number of significant mutations is not fixed, have to use Reversible-Jump MCMC, where the size of the parameter space is allowed to vary from iteration to iteration. Propose to increase or decrease the number of clusters by 1, and either combine 2 adjacent clusters or split an existing one, proposing new means as before.

The label-switching problem

In order to draw conclusions about the means and variances of the clusters, need to have a consistent way of labelling them. Although the label-switching problem also exists in 1-dimensional data, it is especially important in higher dimensions, hence indeed in geographical data. To reduce the effect of label-switching, we use the following algorithm:
At each iteration, pick a labelling to maximise

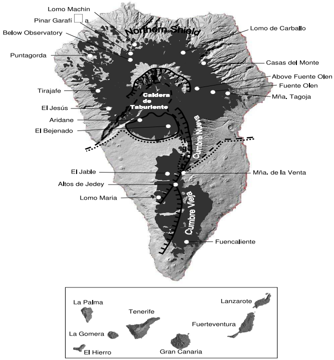
$$\sum_{j=0}^{n} \log \frac{1}{t} \left(\sum_{i=1}^{t} \mathbb{I}_{c_j^{(i)} = c_j^{(t)}} \right)$$

where c_j denotes the cluster of node j.
This algorithm picks labels for each cluster so that "as many nodes as possible belong to their favourite cluster so far".

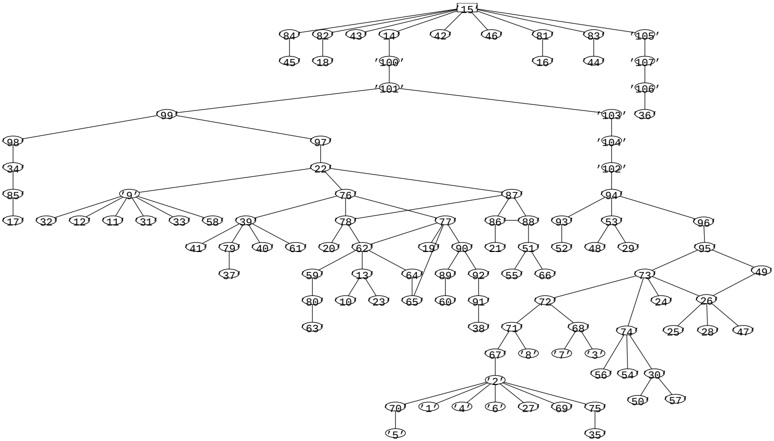
The Beetle Dataset

We have mitochondrial sequence data of 138 individuals from 18 locations of La Palma island (in the Canary islands). The sequences are 570 letters long, but in fact only 66 of them are actually variable. Of the 138 sequences, we obtain 69 distinct ones.

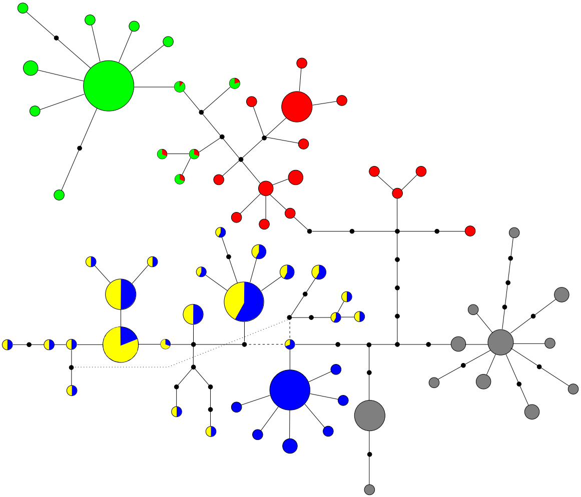
The volcano present on the west side of the island suggests that the clusters on either side will belong to different clusters as beetles cannot fly.



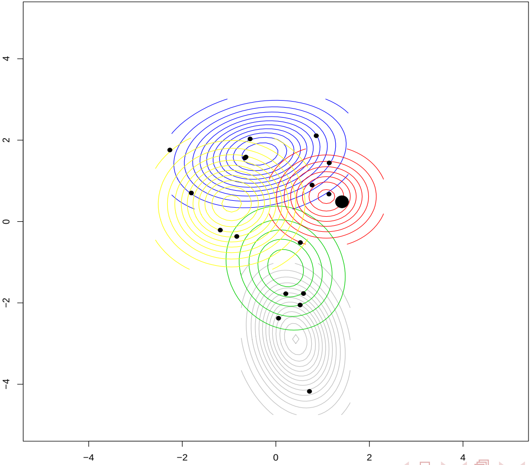
Total Sequence Network



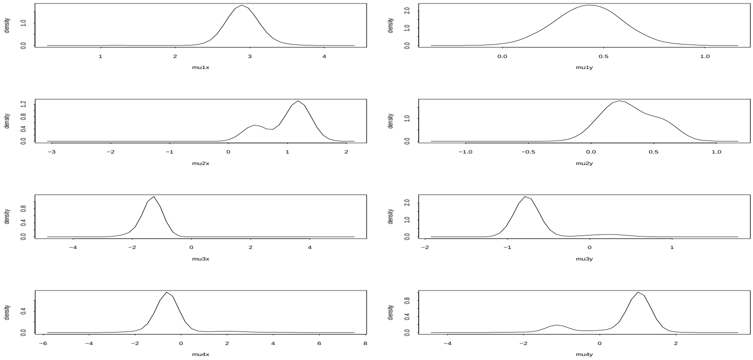
MAP sequence network showing clustering



Contour plot of the clusters



Estimated density of the means of the x and y coordinates in the 4 clusters



Future Work/Problems

- Unknown nucleotides
- Improve overall speed
- Use particle filters
- Improve efficiency of clustering proposal for the phylogeographic case
- Improve convergence to tree
- Improve model for inference about the root
- Normalisation constant
- Use standard phylogenetic trees for the phylogeographic analysis

References

- J. C. Avise. *Molecular Markers, Natural History and Evolution*. Chapman and Hill, 1994.
- A. R. Templeton. Nested clade analysis of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, 7(3):381–397, 1998.
- S. P. Brooks, I. Manolopoulou and B. C. Emerson. Assessing the Effect of Genetic Mutation: A Bayesian Framework for Determining Population History from DNA Sequence Data. *Bayesian Statistics 8,* 2007.
- I. Manolopoulou, S. P. Brooks and L. Legarreta. A Bayesian Framework for Analyses of Demographic DNA Sequence Data. *Proceedings of the 20th Panhellenic Statistics Conference 2007 "Statistics and Society"*, to appear.