MCMC methods on estimating the genetic and geographical history of individuals

Ioanna Manolopoulou

June 16, 2007

The problem: Data & Aims

Geographical data.

We have *aligned* DNA sequences from a set of individuals and their location.

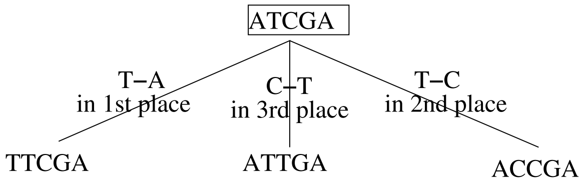
Individual	DNA seq	location
A	ATCGA	(1.3, 2.5)
B	ATCGA	(1.7, 3.9)
C	ATTGA	(2.9, 0.1)
D	ACCGA	(3.1, 6.1)
E	TTCGA	(1.3, 2.5)

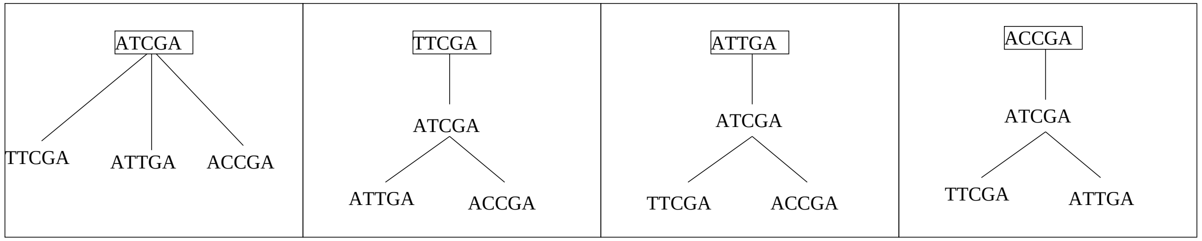
We want to split our data into significant clusters in order to draw conclusions about their geographical and genetic history.

The mutation process

All individuals have a very long DNA sequence of A,T,C,G.
e.g. Assume initially one sequence present, ATCGA, which then mutated to TTCGA, ATTGA and ACCGA. This process yields the graph:

The resulting data would be (ATCGA, ATTGA, ACCGA, TTCGA): Not clear from data which sequence is the oldest.





The mutation process

We have l independent and identical parallel Markov Processes (l length of seqs), 4 states: A, G, C, T, with Q-matrix:

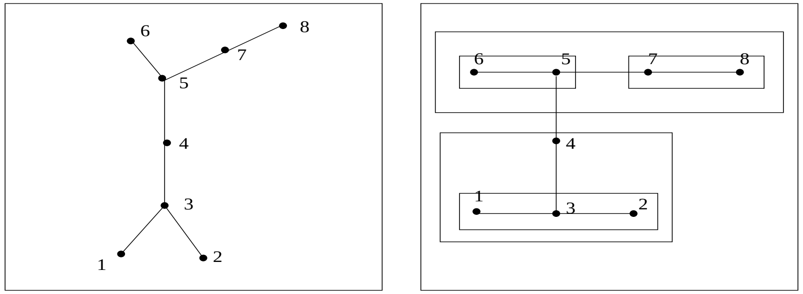
$$Q = \begin{pmatrix} \cdot & \pi_G\alpha & \pi_C\beta & \pi_T\gamma \\ \pi_A\alpha & \cdot & \pi_C\delta & \pi_T\epsilon \\ \pi_A\beta & \pi_G\delta & \cdot & \pi_T\zeta \\ \pi_A\gamma & \pi_G\epsilon & \pi_C\zeta & \cdot \end{pmatrix}$$

We assume we start at stationary distribution π, so any of the l nucleotide positions equally likely to mutate (independent of transition matrix) . After a mutation, the process remains stationary, and by Strong Markov property starts afresh.

Nested Clade Analyses (NCA)

Idea: Split data into nested groups and perform ANOVA so that significant branches are determined. Need unique tree.

Perform ANOVA for each level of nesting, taking branches as groups. Testing for significant groups of our geographical data.



Problems:

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

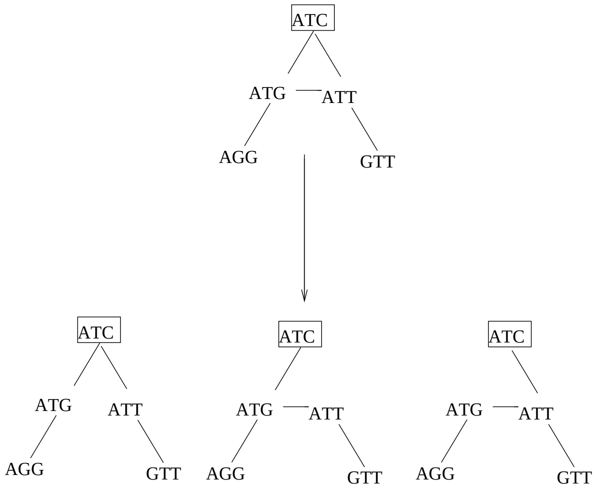
Priors and Distributions

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

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

Missing Sequences

Assume that the underlying true tree contains the minimum number of unobserved sequences (nodes). Adding more nodes means adding mutations which have very small probability, hence a reasonable assumption.



Add nodes so that each node we add connects (or brings closer together) the maximum number of closest disconnected groups of nodes. This procedure only has to be done once, outside the Monte Carlo chain.

Minimum tree assumption is useful because it means that we have fixed size of tree and so, conditional on that, we can assume multinomial distributions for the nucleotides and mutation probabilities.

Nucleotide frequencies and mutation probs

Using the observed nucleotide frequencies, we generate estimates of the true frequencies from the exact posterior Dirichlet distribution (this is called a Gibb's sampler, it's a MCMC in which the acceptance probability simplifies to 1):

$$(\pi_A, \pi_G, \pi_C, \pi_T)|D, \boldsymbol{B} \quad \sim \quad \mathrm{Dir}(B_1 + n_1, B_2 + n_2, B_3 + n_3, B_4 + n_4)$$

We then propose estimates for the mutation probabilities assuming the present tree is true. Typically very few loops, so this assumption does not have a crucial impact.

Cladogram: Root and Tree

We propose a node to be the root uniformly according to nodes' degrees. Assuming a simulated prior so that $p(x) = f(\text{degree}(x))$, we calculate the acceptance probability using current nucleotide frequencies and transition probabilities.

Once we have established a root, propose to remove loops by deleting edges uniformly (from the loops) so that every set of deleted edges has the same probability. We then calculate the probability of the data assuming current nucleotide frequencies and mutation probabilities.

Clusters and their means and variances

For a fixed number of significant mutations K, we want to find $K+1$ means. We propose a new covariance matrix Σ using the posterior distribution of Σ if the data only is known:

$$\Sigma_i | D, \boldsymbol{e} \sim \text{InvWishart}\Big(n+m, \Psi + \sum_{c(j)=i} \boldsymbol{x}_j \boldsymbol{x}_j^T + n_i \overline{\boldsymbol{x}}_i \overline{\boldsymbol{x}}_i^T \Big)$$

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

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

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

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

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

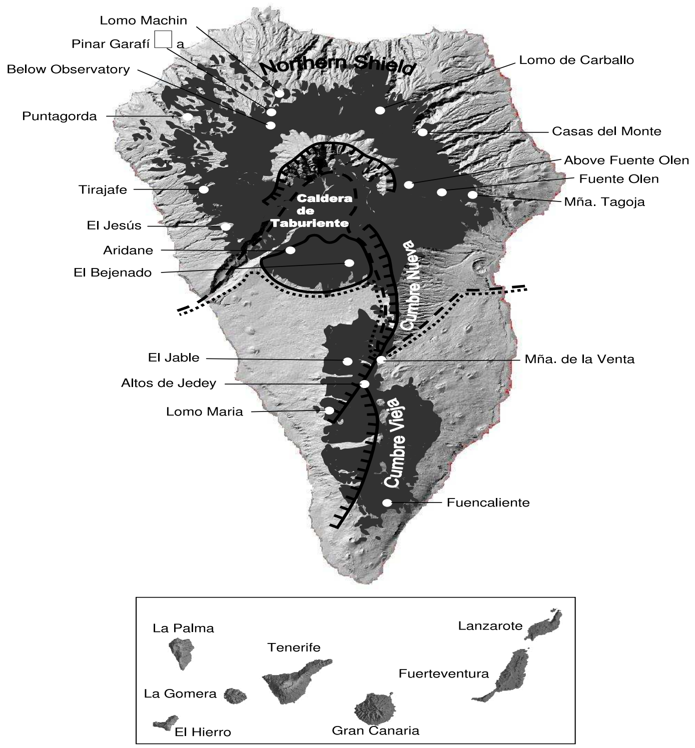
$$\sum_{j=0}^{n} \log \frac{1}{t} \left(\sum_{i=1}^{t} \mathbb{I}_{c_j^{(i)} = c_j^{(t)}} \right)$$

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”.

Example: Beetle data from La Palma

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.

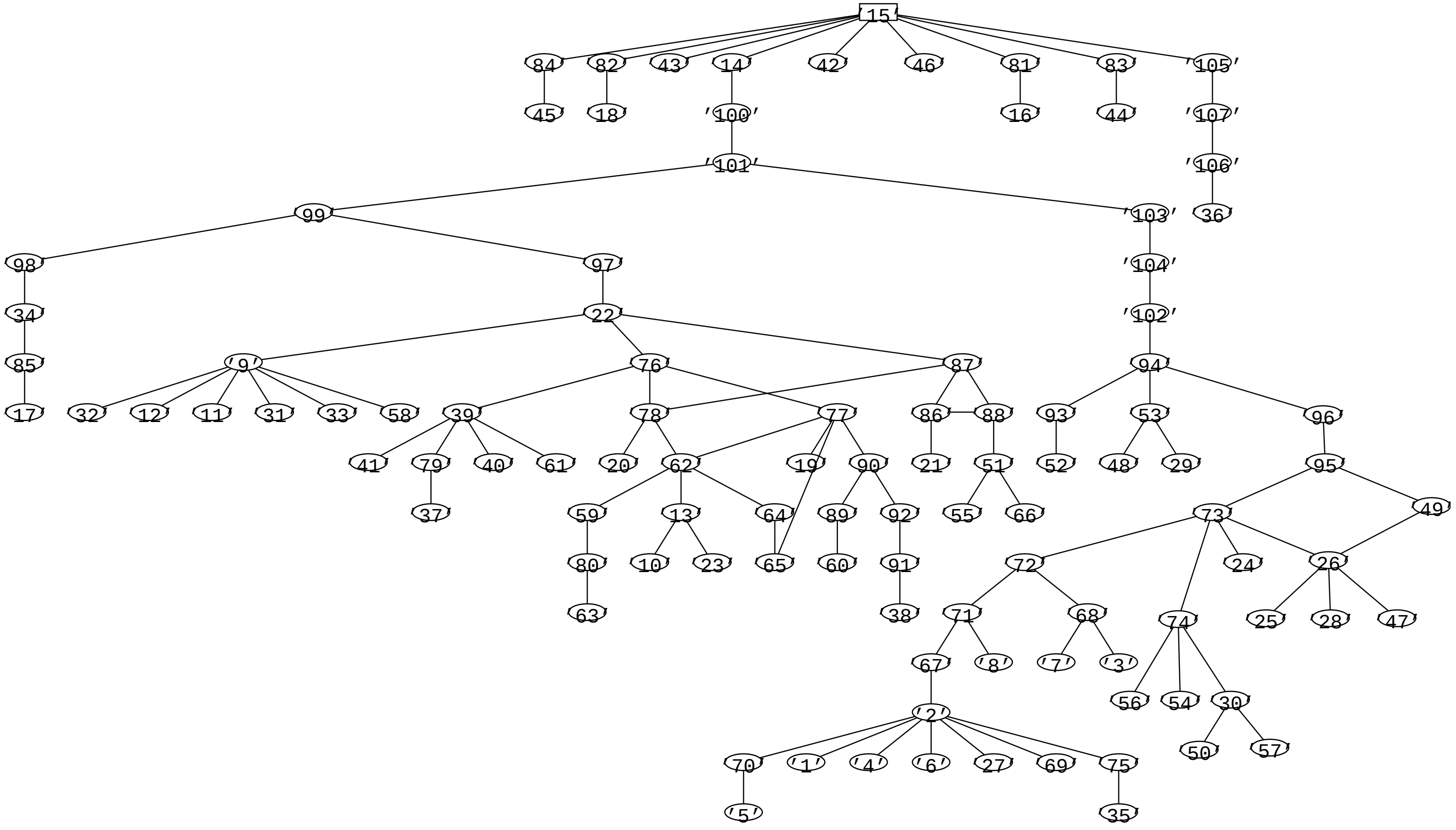


The volcano present on the west side of the island suggests that the clusters on either side will belong to different clusters as beetles cannot fly.

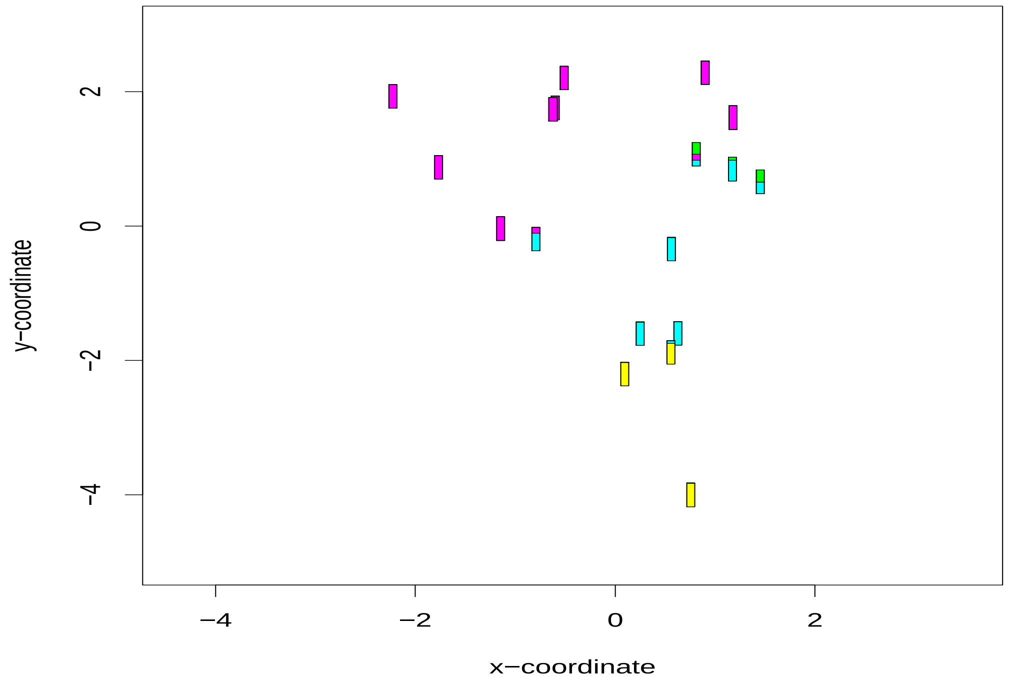
Parameters

$$\begin{aligned}
(\pi_A, \pi_G, \pi_C, \pi_T) &\sim \text{Dirichlet}(1,1,1,1) \\
\mathbf{n} &\sim \text{Multinomial}(138 \times 570, \boldsymbol{\pi}) \\
\alpha,\ \zeta &\sim \text{N}(2, 10) \quad \text{independent} \\
\beta,\ \gamma,\ \delta,\ \epsilon &\sim \text{N}(1, 10) \quad \text{independent} \\
\Sigma &\sim \text{InvWishart}(40, 100 \times \mathbb{I}_2) \\
\boldsymbol{\mu}_i | \Sigma &\sim \text{MVN}\left(\mathbf{0}, 10 \times \Sigma\right) \\
K &\sim U\{2, \ldots, 5\} \\
p_{split} &= 0.5
\end{aligned}$$

Graph formed using the Beetle data from La Palma



Graph showing the mode clustering in each location for 3 mutations



Output

group of edges	$\mathbb{P}(sig\|D)$
22-97 97-99	0.1723
94-102 101-103 102-104 103-104	0.0898
14-15	0.0439
100-101 14-100	0.0409
9-22	0.0409
15-105 105-107 106-107 36-106	0.0216
94-96 95-96	0.0218

<table>
<tr><th>node</th><th>deg</th><th>$\mathbb{P}(\text{root}|D)$</th></tr>
<tr><td>15</td><td>9</td><td>0.2909</td></tr>
<tr><td>2</td><td>8</td><td>0.3704</td></tr>
<tr><td>9</td><td>7</td><td>0.1210</td></tr>
<tr><td>26</td><td>5</td><td>0.0442</td></tr>
<tr><td>39</td><td>5</td><td>0.0244</td></tr>
<tr><td>73</td><td>5</td><td>0.0391</td></tr>
<tr><th># mutns</th><th colspan="2">post prob</th></tr>
<tr><td>2</td><td colspan="2">0.1545</td></tr>
<tr><td>3</td><td colspan="2">0.3300</td></tr>
<tr><td>4</td><td colspan="2">0.3400</td></tr>
<tr><td>5</td><td colspan="2">0.1754</td></tr>
<tr><td colspan="3">acceptance ratio 0.2588</td></tr>
<tr><td colspan="3">27 mins approximately</td></tr>
</table>

Future work

- Variable Q-matrix for different nucleotide posititions
- Distribution of root
- Consider non-minimal networks
- Introduce time T of mutations
- Introduce migration process