A Bayesian Framework for Analyses of Demographic DNA Sequence Data

Ioanna Manolopoulou[1]

Simon Tavaré[1], Steve Brooks[1], Lorenza Legarreta[2]

[1]University of Cambridge

[2]University of East Anglia

UNIVERSITY OF CAMBRIDGE

The Problem

We have aligned sections of the DNA sequences of n individuals along with their geographical locations. Example:

Individual	DNA seq	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)

One of the questions asked by biologists is how to split our data into significant clusters in terms of their geographical location, so that the results are consistent with the genetic history, and what demographic events occurred in history (e.g. colonisation).

The Mutation Process

Firstly we model the mutation process. Each sequence of l nucleotides A, G, C, T can be represented by l parallel independent Markov Processes where each position j (j =1…l) has transition matrix:

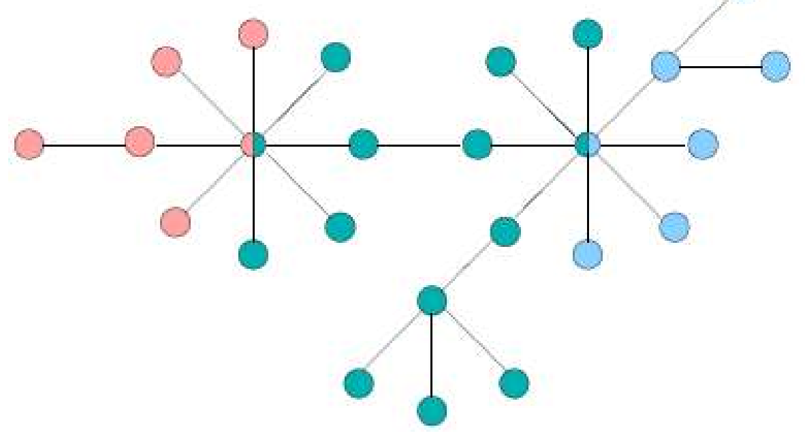
$$Q = \begin{pmatrix} \cdot & \phi_j\pi_G\alpha & \phi_j\pi_C\beta & \phi_j\pi_T\gamma \\ \phi_j\pi_A\alpha & \cdot & \phi_j\pi_C\delta & \phi_j\pi_T\epsilon \\ \phi_j\pi_A\beta & \phi_j\pi_G\delta & \cdot & \phi_j\pi_T\zeta \\ \phi_j\pi_A\gamma & \phi_j\pi_G\epsilon & \phi_j\pi_C\zeta & \cdot \end{pmatrix}$$

Where the π represent the stationary probabilities of the chain and the ϕ represent the relative mutation rates for each nucleotide position. It can easily be checked that this chain is time-reversible and that π is indeed the stationary distribution.

By considering a finite set of parsimonious and nearly parsimonious mutational histories, we draw inferences about the mutational tree by introducing it as one of the parameters.

The Colonisation Model

Now, for a colonisation event, we assume that one or more sequences left a geographical population to start a new one. Then an example of such a set of sequences would be:



Each node here represents a sequence and lines represent mutations. The colour represents the geographical cluster. So, in the figure above, the node which is green/red is a sequence which initially was present in one population (the green), but then colonised to a new one (the red), so that the individuals with that sequence have distinct descendant sequences in the 2 geographical clusters.

The geographical clusters are assumed to have a 2-dimensional Normal Distribution $\mathrm{MVN}_2(\boldsymbol{\mu}_i, \Sigma_i)$ and hence colonised sequences will have a mixture of Normals distribution

Materials & Methods

We use Reversible Jump Markov Chain Monte Carlo (RJMCMC) methods to sample from the posterior distribution of our parameter set $\boldsymbol{\theta}$. So, at each iteration $\boldsymbol{t}$ with parameters $\boldsymbol{\theta_t}$ we propose to move to $\boldsymbol{\theta}_{t+1}$ according to some proposal distribution q. We accept that move with probability $\min(1, A)$ where

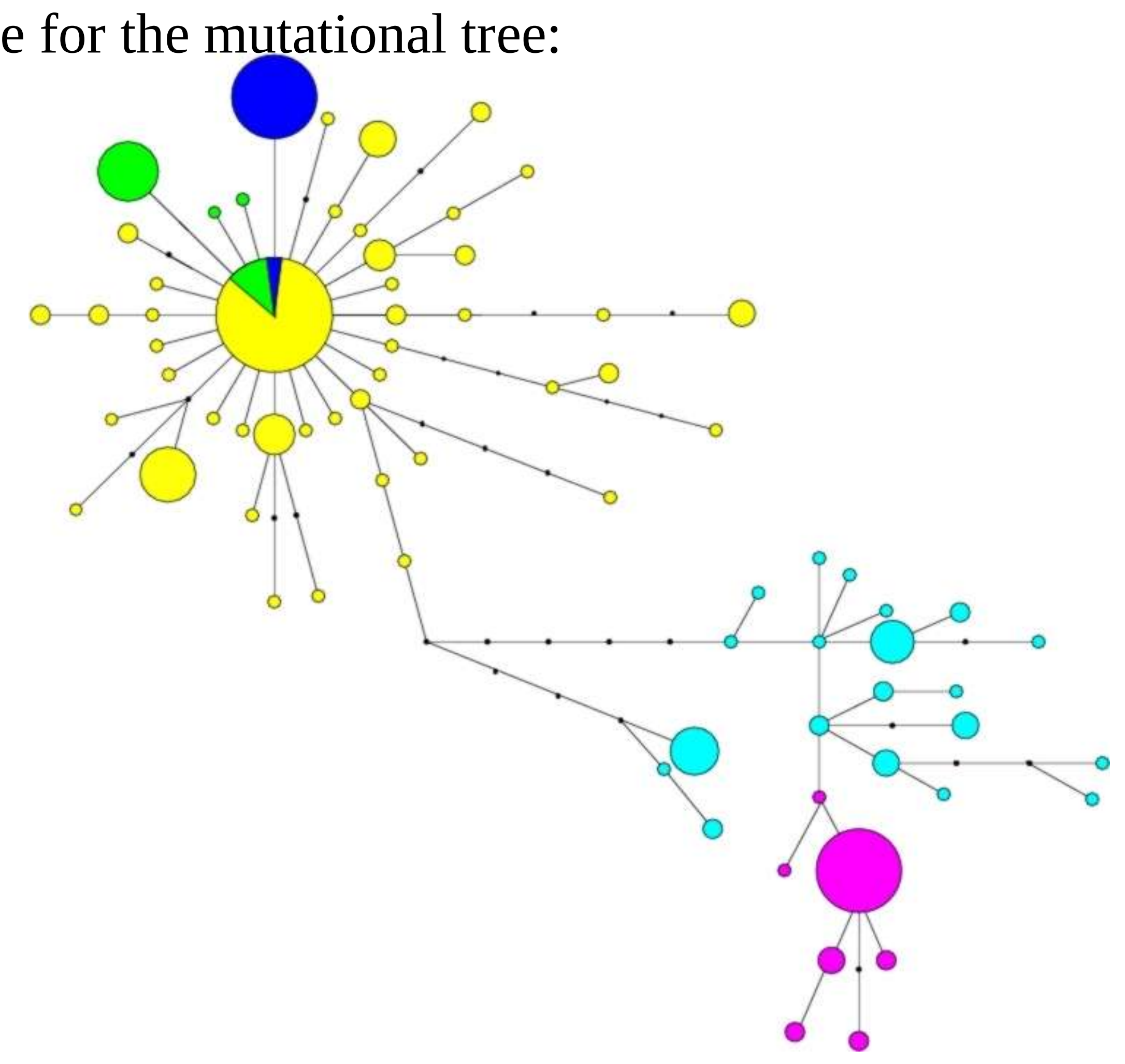
$$A = \frac{\mathbb{P}(data|\boldsymbol{\theta_{t+1}})p(\boldsymbol{\theta_{t+1}})q(\boldsymbol{\theta_{t+1}}, \boldsymbol{\theta_t})}{\mathbb{P}(data|\boldsymbol{\theta_t})p(\boldsymbol{\theta_t})q(\boldsymbol{\theta_t}, \boldsymbol{\theta_{t+1}})}|J|$$

Otherwise we set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$. Here p represents the prior distribution and $\mathbb{P}$ the likelihood of the data, and J is the Jacobian of the tranformation used if the size of the parameter space is allowed to vary.

We update the mutation process parameters, including the ancestral sequence (i..e. the Most Recent Common Ancestor) as well as the means and variances of the clusters and the number of clusters simultaneously. This holistic approach ensures that the uncertainty is carried throughout the analysis.

Results

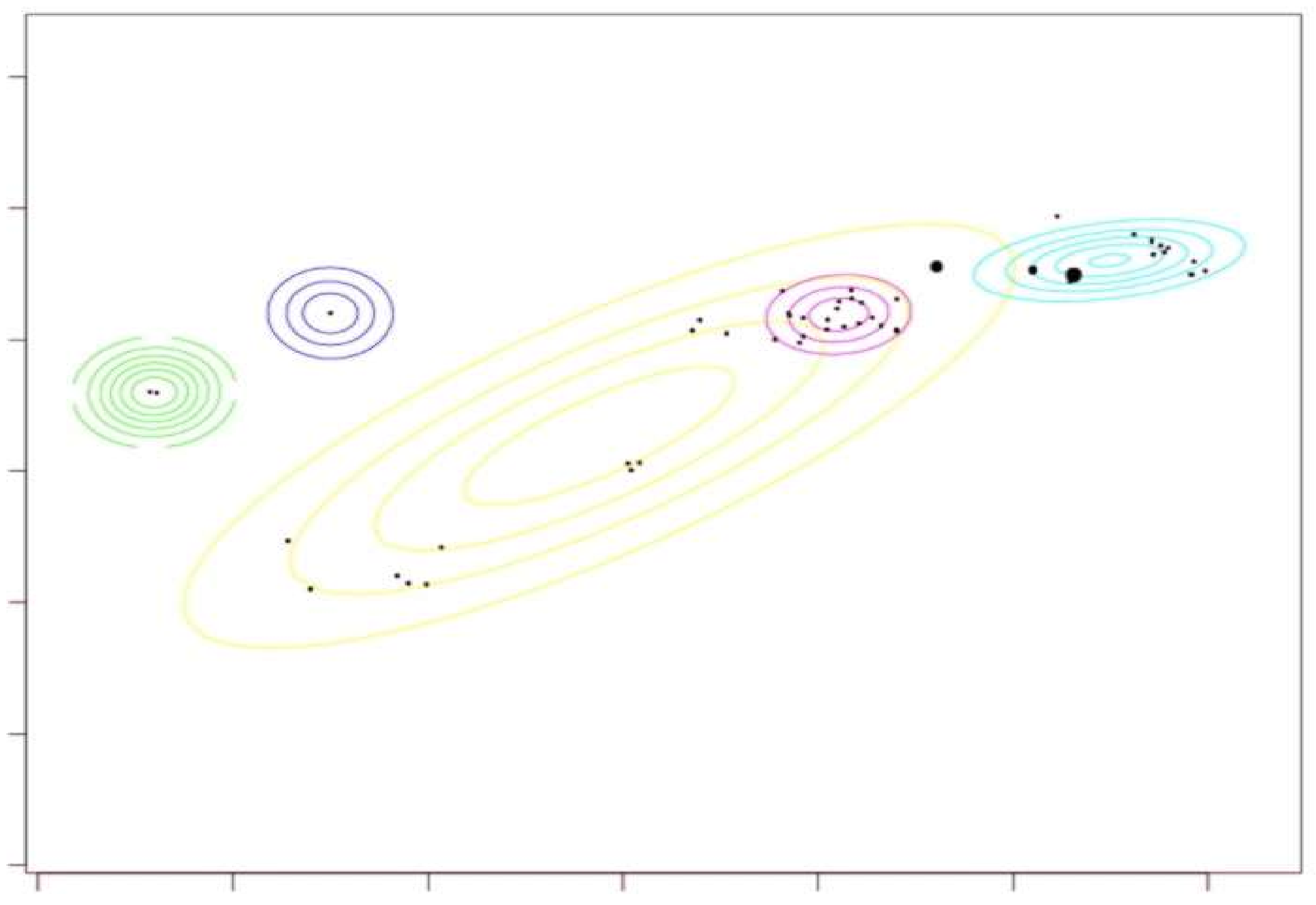
We use a dataset of DNA sequences from weevils in South-West Europe and obtain the following Maximum A Posteriori estimate for the mutational tree:



We see that 2 clusters are distinctly separated, whereas the blue, yellow and green share a node, representing the sequence which colonised the populations.

Conclusions

Looking at the contour plot of the posterior estimates for the 5 clusters below, and taking into account the ancestral locations (represented by larger dots) we see that the area was colonised from the Rhône area easternly into the alpine region and westernly into the iberian region. The North-West locations are clearly totally isolated from all others.



Our approach gives results accurate with biologists' prior beliefs and provides a quantitative comparison to possible scenarios.

References & Acknowledgements

[1] Brooks, S.P., Manolopoulou, I. and Emerson, B.C. (2006) "Assessing the Effect of Genetic Mutation – A Bayesian Framework for Determining Population History from DNA Sequence Data". *Bayesian Statistics 7*.

Many thanks to Trinity College for funding this research and B. C. Emerson for his invaluable contribution.