Maximum Likelihood Estimation of Coalescence Times in Genealogical Trees

Loukia Meligkotsidou and Paul Fearnhead
Department of Mathematics and Statistics
Lancaster University

LMS Durham Symposium
on Mathematical Genetics,
14 July 2004

Outline

- Introduction. Population Genetics
- Part 1. Maximum Likelihood Method
  - Model and Assumptions
  - The Viterbi Algorithm
  - Implementation
- Part 2. Results
  - Simulation Experiments
  - Real Data Application

Population Genetics

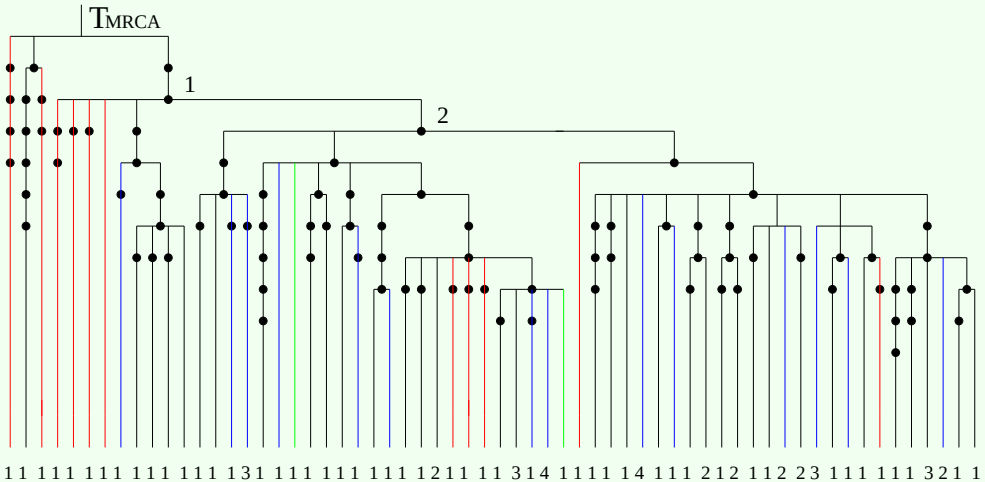
Population genetics is concerned with the study of the ancestry of a particular population in terms of its genetic evolution. For this purpose, a sample of DNA sequences from the population is analyzed.

It is assumed that all of the sequences have evolved from a common ancestral one and genetic variation in the sample is due to a number of mutations having occurred in its ancestry (no recombination).

The ancestral relationships in the sample are described by a genealogy.

The mutational history of the sample is model by a mutational process.

Genealogical tree for a worldwide sample of 82
human Y chromosomes



The Problem

The population genetics problem consists of inference about the genealogical tree and about the mutational process given the tree. It is of particular interest to infer the times, in the past, when coalescence events occurred and most importantly to estimate the TMRCA.

The Method

- Model-free Approach: only specifies a mutation model and makes no assumptions about the genealogy and the demography.
- Maximum likelihood estimates and likelihood-based confidence intervals for the coalescence times.
- Viterbi-type Algorithm which exploits the Markovian structure of the tree to maximize the joint likelihood and calculate the profile likelihoods of the coalescence times in a sequential manner.

Model and Assumptions

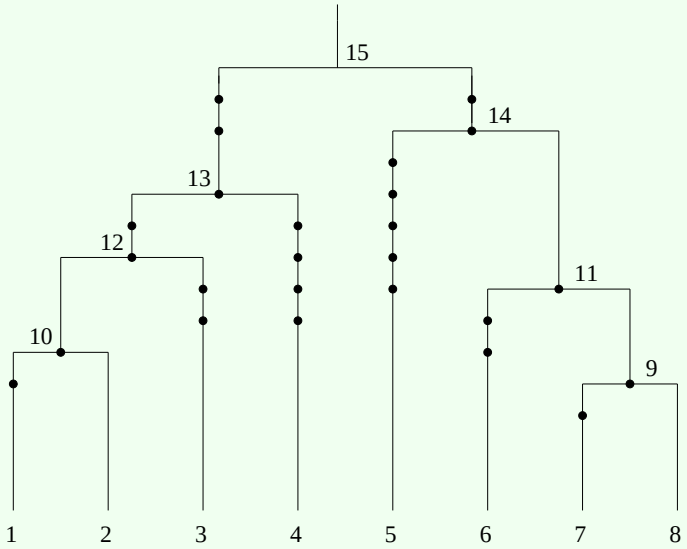
Mutation Model: The DNA sequence data are modeled by assuming that each site evolves independently of all others and mutations occur as events in a Poisson process with rate 1. The mutation rate defines the scale in which time is measured (1 mutation is expected in one unit of time).

Markov Property: Evolution is independent along different paths (lineages).

Constant Molecular Clock: The lengths of the branches connecting each internal node to its descendent tips are identical.

Assumption: The mutation rate is sufficiently low and repeat mutations are sufficiently rare that the number of mutations on the branches of the tree can be inferred from the data (similar to infinitely-many-sites assumption).

Known Tree
Topology



The log-Likelihood Function

We assume that, given observed DNA data for m individuals, the tree topology and the numbers of mutations on the branches of the tree will be known.

Let n_i denote the number of mutations on the ith branch and let b_i be the branch length. The log-likelihood of the "observed" data $\mathbf{n} = (n_1, \ldots, n_{2(m-1)})$ and the unobserved variables $\mathbf{b} = (b_1, \ldots, b_{2(m-1)})$, is given by

$$\ell(\mathbf{n}, \mathbf{b}) = \sum_{i=1}^{2(m-1)} g_i(b_i),$$

where $g_i(b_i) = n_i \log(b_i) - b_i$ is the contribution of the data on the ith branch, and each b_i is a function of at most two coalescence times.

For ML estimation of coalescence times, the log-likelihood has to be maximized over $\mathbf{b}$, subject to some constraints on the branch lengths.

The Viterbi Algorithm: HMM

Consider a HMM where the hidden underlying process $\{X_t\}_{t=1}^T$ is a homogeneous discrete-time Markov chain on a finite state-space $S = \{1, \ldots, m\}$, with transition probability matrix $\mathbf{P} = [p_{ij}]$ and stationary distribution $\pi = (\pi_1, \ldots, \pi_m)$. Let $g_j(Y_t \mid X_t = j)$, $j = 1, \ldots, m$ denote the probability distribution of the observed variable Y_t, $t = 1, \ldots, T$, given the unobserved state X_t.

The Viterbi algorithm exploits the Markovian structure of the HMM to built a reqursion for the calculation of the joint probability, δ_t, of $(X_1, \ldots, X_t)$ and $(Y_1, \ldots, Y_t)$, maximized over the states $(X_1, \ldots, X_{t-1})$:

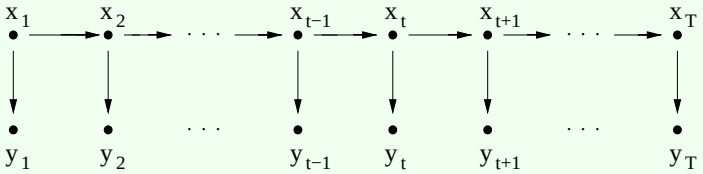
$$\delta_t(j) = \left[\max_i \delta_{t-1}(i) p_{ij}\right] g_j(y_{t-1} \mid x_{t-1} = j), \quad j = 1, \ldots, m.$$

Maximization of the joint likelihood is achieved for the value of j that maximizes $\delta_T(j)$.

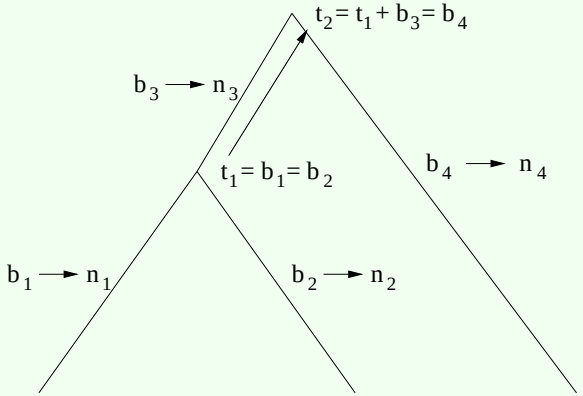
The Viterbi Algorithm: Estimation of Coalescence Times

We construct a Viterbi-type algorithm for the estimation of coalescence times in genealogical trees, by exploiting the Markov property in the tree: The time at the ith node only depends on the times at its immediate neighbouring nodes.

The numbers of mutations correspond to the observed data of the HMM, while the branch lengths correspond to the unobserved states that need to be restored. Given the branch lengths the coalescence times will be also known.



HMM



Genealogical Tree

The Algorithm

A Viterbi-type algorithm can be constructed to maximize the log-likelihood function ℓ.

This is given by:

- Start from pairs of tips which coalesce.
- Go up the tree, iteratively computing the contribution to the log-likelihood of the data on the subtree below the ith node, maximized over the coalescence times at the internal nodes of this subtree.
- At the final step of the sequential procedure, the profile log-likelihood of the time t at the root of the tree, $\ell(t)$, is obtained.

MLEs and CIs

The MLE for the TMRCA will be the value of t that maximizes $\ell(t)$. Since $\ell(t)$ is convex, a unique maximum exists.

Let ℓ^* denote the maximum of the profile log-likelihood of t and define the deviance of $\ell(t)$ from ℓ^* by

$$D^*(t) = 2(\ell^* - \ell(t)).$$

An $100(1-\alpha)\%$ likelihood-based CI for t, is given by

$$\{t : D^*(t) \leq c_\alpha\},$$

where c_α is the uper $100(1-\alpha)\%$ point of the χ^2_1 distribution.

The MLEs and profile log-likelihoods for the other coalescence times can then be obtained, going down the tree, in an annalogous manner.

Generalizations

1 Partially known tree topologies can also be analyzed with our method. These include ambiguous parts where the exact order of some coalescence events is not known. When maximizing sequentially the log-likelihood, more complicated steps are needed for the ambiguous parts in the tree.

2 The previous discussion was based on the assumption that the available chromosomal segment was the same for all individuals and, therefore, all the DNA sequences in the sample had the same length. Our method can be easily adopted to analyze samples of sequences of different lengths. In this case, the branch lengths are scaled by the mutation rates of the respective sequences.

Simulation Study

We have conducted several simulation experiments in order to assess the performance of our method.

We have considered simulated data sets of different sample sizes from the coalescent, under various assumptions for the population model (constant population size, exponentially growing population, structured population).

We used the infinitely-many-sites model as the mutation model and considered different mutation rates.

The results have shown that the ML method is more accurate than other model-free methods, and more robust to demographic forces generating the data than model-based methods.

As the sequence lengths increase, the MLEs are asymptotically efficient (regular problem).

Simulation Results

We compared our method with two other model-free approaches under different demographic models (a Tang et al., 2002 and Thomson et al., 2000). The comparison is with respect to the accuracy of the estimation of the TMRCA. (Model-based approaches suffer from non-robustness.)

The results given below correspond to a simulation study where 500 genealogies for samples of $n = 20$ sequences were generated from the coalescent, assuming (1) constant population size, (2). exponential growth and (3) structure: 2 migrating subpopulations.

		ML Method	Tang et al.	Thomson et al.
(1)	Relative Error	0.1106	0.1208	0.1293
	Coverage probability	0.9600	0.9600	0.9980
(2)	Relative Error	0.1464	0.1547	0.1530
	Coverage probability	0.9440	0.9400	1.000
(3)	Relative Error	0.1007	0.1064	0.1101
	Coverage probability	0.9480	0.9480	0.9960

Application to Human Y-chromosome DNA Data

We have analazed a worldwide sample of 82 DNA sequences from the human Y chromosome (tree shown before).

The available sequences were of different lenghts. Previous studies only analyzed a subsample of these data (43 sequences) for which the available chromosomal segment was the same for all of the individuals.

Apart from estimating the TMRCA, it is of interest to estimate the ages of coalescences 1 (separation from the deep African clade) and 2 (human population expansion).

Small dataset: The subset of these data was analyzed by Thomson et al. (2000) using `genetree` under the assumption of constant population and of exponential growth, as well as using their simple model-free estimator for the TMRCA. The results are not robust to different demographic models. The method of Tang et al. (2002) is not able to estimate the coalescence times at internal nodes.

Results: Small Dataset

MLEs (in thousands of years) and 95% CIs, based on the profile likelihood, for the times of important events are given below. Results from previous analysis are also given for comparison purposes.

Method	T_{MRCA}	CI
Maximum Likelihood	63	$(40 - 100)$
Tang et al.	63	$(29 - 98)$
Thomson et al.	70	$(10 - 130)$
`genetree` (const. pop.)	84	$(55 - 149)$
`genetree` (exp. growth)	59	$(40 - 140)$

Method	T_1	CI	T_2	CI
Maximum Likelihood	45	$(32 - 74)$	37	$(27 - 67)$
`genetree`	47	$(35 - 89)$	40	$(31 - 79)$
Segregating Sites	43	$(37 - 111)$	42	$(36 - 109)$

Results: Full Dataset

MLEs (in thousands of years) and 95% CIs, based on the profile likelihood, for the times of important events are given below.

	MLE	CI
T_{MRCA}	55	$(37 - 82)$
T_1	35	$(25 - 49)$
T_2	30	$(21 - 41)$

Note: The estimates obtained by analyzing the full data are smaller.

Conclusions

- We have developed a Viterbi-type algorithm for ML estimation of coalescence times in genealogical trees.
- CIs for the coalescence times, based on the profile likelihood, are also calculated.
- The ML method is more accurate than existing model-free methods and more robust to demographic forces than model-based methods.
- Generalizations of the method to deal with partially known tree topologies and sequences of different lengths are very important for real data applications.

References

Tang, H., Siegmund, D. O., Shen, P., Oefner, P. J. and Feldman, M. W. (2002). Frequentist estimation of coalescence times from nucleotide sequence data using a tree based partition. *Genetcs* **161**, 447-459.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. and Feldman, M. W. (2000). Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl. Acad. Sci. US* **97**, 7360-7365.