 doi: 10.1111/j.1471-8286.2007.01691.x

PROGRAM NOTE

TREES SIFTER 1.0: an approximate method to estimate the time to the most recent common ancestor of a sample of DNA sequences

PATRICK MARDULYN
Laboratory of Evolutionary Genetics, Institute of Molecular Biology and Medicine, Université Libre de Bruxelles, rue Jeener & Brachet 12, 6041 Gosselies, Belgium

Abstract

TREES SIFTER 1.0 implements an approximate method to estimate the time to the most recent common ancestor (TMRCA) of a set of DNA sequences, using population evolution modelling. In essence, the program simulates genealogies with a user-defined model of coalescence of lineages, and then compares each simulated genealogy to the genealogy inferred from the real data, through two summary statistics: (i) the number of mutations on the genealogy (M_{π}), and (ii) the number of different sequence types (alleles) observed (K_{π}). The simulated genealogies are then submitted to a rejection algorithm that keeps only those that are the most likely to have generated the observed sequence data. At the end of the process, the accepted genealogies can be used to estimate the posterior probability distribution of the TMRCA.

Keywords: approximate method, coalescence model, DNA sequences, genealogy, simulations, TMRCA

Received 15 October 2006; revision accepted 4 December 2006

Inferring the coalescence time of a sample of DNA sequences, i.e. the time to the most recent common ancestor (TMRCA) of these sequences, is important for many population genetics or phylogeography studies (e.g. Templeton 1993). For example, it may allow estimating the divergence time between two populations if it is large relative to populations sizes and if migration between populations is small (Rosenberg & Feldman 2002).

When inferring coalescence times from molecular genetic data, several authors have promoted the use of population evolution modelling (but see Tang *et al*. 2002; for another approach) to incorporate the uncertainty of the estimate associated to the stochastic nature of the evolutionary process (i.e. the random coalescence of lineages (going backward in time) and the random accumulation of mutations along the branches of the genealogy). In this approach, a coalescence model (e.g. Hudson 1990; Hein *et al*. 2005) is defined a priori and the probability distribution that the observed data were generated under this model (post-data probability distribution) is derived. Although full-likelihood methods have been developed to estimate the probability density function (pdf) of parameters like the TMRCA using Markov chain Monte Carlo simulations (Griffiths & Tavaré 1994, 1995; Nielsen & Wakeley 2001), these are however, computer-intensive, and thus time-consuming, and current implementations of methods estimating the TMRCA from the full data are limited to relatively simple coalescence models (GENETREE, available at http://www.stats.ox.ac.uk/~griff/software.html; IM, available at lifesci.rutgers.edu/~heylab/HeylabSoftware.htm).

A promising alternative approach consists in using approximate methods based on summary statistics (Beaumont *et al*. 2002; Rosenberg & Nordborg 2002). In this case, the full sequence data are replaced by a summary statistic. For example, Tavaré *et al*. (1997) have suggested the use of a simple rejection algorithm to estimate the TMRCA of a set of DNA sequences. The principle of this algorithm is to simulate many genealogies (of a number of sequences identical to our sample) following a coalescent model defined a priori, while possibly varying the values of some of the model parameters between each simulation. Each genealogy is then accepted or rejected with a probability proportional to the probability that it has generated

Correspondence: Patrick Mardulyn, Fax: 32-2-378950; E-mail: pmarduly@ulb.ac.be

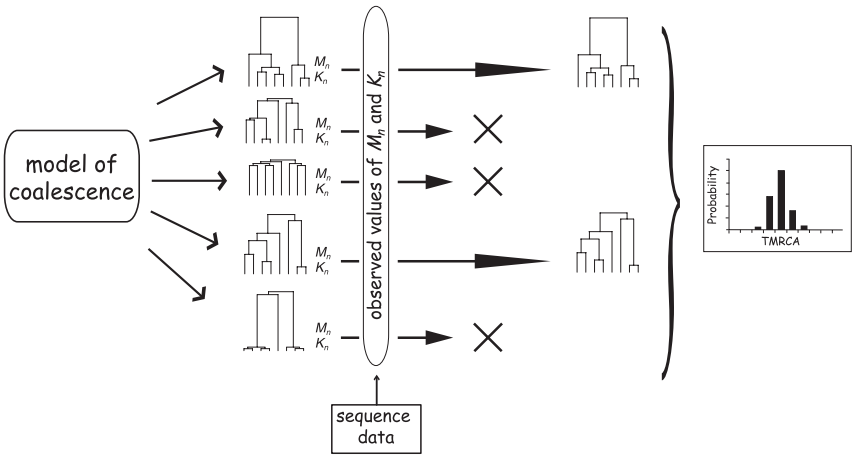

Fig. 1 Principle of TREES SIFTER 1.0: genealogies are simulated using a user-defined model of coalescence of lineages, and are accepted (or rejected) only if the simulated and real values of two summary statistics (M_n and K_n) are sufficiently similar. The accepted genealogies can then be used to estimate the posterior probability distribution of the TMRCA.

the same number of segregating sites (S_n) that is observed in the sampled sequences (i.e. depending on how similar the simulated value of S_n is to the observed value of S_n). At the end of the process, the TMRCA of all accepted genealogies are used to generate the estimate of the pdf of this parameter.

Replacing the full data by a summary statistic like S_n greatly reduces the calculations, and thus allows one to investigate a much wider range of model parameters (Tavaré *et al*. 1997) and to implement more complex coalescence models. On the other hand, one obvious pitfall for this approach is that the chosen summary statistic will be less informative than the full sequence data. The use of more than one summary statistic can therefore help to improve this approach. For example, Pritchard *et al*. (1999) have implemented a method derived from Tavaré *et al*. (1997) for microsatellite data that makes the rejection/acceptance decision based on three summary statistics. Instead of calculating the probability that the tested genealogy has generated the observed values of each summary statistic, it simply calculates the difference between the values of the observed and simulated genealogies. The genealogies are then accepted only if, for each summary statistic, this difference is smaller than a threshold value.

TREES SIFTER version 1.0 implements a similar rejection algorithm that bases its acceptance/rejection decision on two summary statistics: the number of mutations on the genealogy (which is equivalent to S_n under the infinite sites model) and the number of alleles in the sample (K_n). The hope is that these two statistics capture most of the information enclosed in the genealogy. The first statistic, the number of mutations on the genealogy, can easily be measured from the observed data by counting the number of mutations on the inferred most parsimonious tree, as already suggested by Fu (1997). Along with each accepted genealogy, TREES SIFTER provides a weight that is proportional to the difference between the simulated and observed values of M_n and K_n. This weight can be used to estimate the posterior distribution of the TMRCA (or any other estimated parameter) as suggested by Beaumont *et al*. (2002). The principle of TREES SIFTER is depicted in Fig. 1.

It is important to note that this estimation procedure assumes that the number of mutations on the genealogy of the sampled sequences is relatively low. Indeed, if the number of mutations is too high, the number of allelic types will simply be equal to the sample size, and will not anymore depend upon the shape of the genealogy. Also, it is assumed that the DNA fragment of interest was not recombined during the time that separates the generation of the TMRCA from the present generation.

Implementation

The program allows the simulation of genealogies under various models with or without population structure and migration. The model can be defined with any number of populations, each with a specific size. A different migration rate can be specified for each pair of populations, as defined in a migration matrix by the user. Change of population size can be modelled either by exponential growth, or by an instantaneous size increase/decrease defined at a specific time point (time measured in number of generations). Modification of parameters (population sizes, population growth rates, and migration rates) can be scheduled at specific time points in the past, as well as the merging of two or more populations into one or more ancestral populations (going backward in time). Although the classical coalescent model makes the approximation that the population size is always much larger than the sample size, and thus that only one coalescence event per generation can occur, this approximation does not hold in many situations like in a severely bottlenecked population. TREES SIFTER implements a simulation process occurring backward in time, one generation at a time, which allows the occurrence of more than one coalescence event per generation.

In many cases, no prior information is available to help setting the value of some of the parameters of the model

used to simulate genealogies, e.g. the population size or the mutation rate. In these cases, the program can be set to generate genealogies under various models, changing the value of some parameters between each simulation.

The simulated genealogies can be submitted to a rejection algorithm that keeps each of them only if the difference between the simulated and observed value of both summary statistics is below a threshold (i.e. the genealogy is accepted when it is likely to have generated the observed values of both summary statistics). This threshold is specified by the user. The differences δ_1 and δ_2 between the simulated and observed values of M_n and K_n are calculated as follows:

$$\delta_1 = \frac{|E(M_n) - k|}{k} \quad \text{and} \quad \delta_2 = \frac{|E(K_n) - h|}{h}$$

where k and h are the observed values of M_n and K_n, respectively, and $E(M_n)$ and E(K_n) are the mean M_n and K_n values for the simulated genealogy. $E(M_n)$ is calculated, assuming a Poisson distribution, as $E(M_n) = \mu l$, where μ is the mutation rate and l the length of the genealogy. $E(K_n)$ is estimated by simulating the mutation process 10 times along the branches of the genealogy, each time recording the resulting K_n. It is then defined as the geometric mean of these simulated values. A weight provided by the program with each accepted genealogy is simply calculated as.

$$w = \frac{(\delta - \delta_1)(\delta - \delta_2)}{\delta_2}$$

All accepted genealogies are output in Newick format (evolution.genetics.washington.edu/phylip/newicktree.html), as well as the corresponding TMRCA and the values of other parameters of the model that were varied during the simulations. These data can then be used to estimate the posterior probability distribution of the TMRCA.

The program is written in Java, and the source code is available as well as executables for Mac OSX and Windows, at http://ueg.ulb.ac.be/treesSifter. The program can be used for the sole purpose of simulating genealogies or to estimate the TMRCA of a sample of DNA sequences.

Testing the program

In order to test the program, we have simulated 100 genealogies using TREES SIFTER under the following model of population evolution: six completely isolated populations (no migration) of 1000 sequences are all merged, going backward in time, into a single ancestral population ($N = 1000$), exactly 3000 generations ago. The genealogies were simulated from a sample of 30 sequences, five sequences taken from each of the six contemporary populations. The evolution of DNA sequences along each generated genealogy was simulated using SEQ-GEN (Rambaut & Grassly 1997; sequence length of 2000 nucleotides, model HKY85, Ti/Tv = 2, equal base frequencies, equal substitution rate per site, tree scaled at 0.003). The resulting sequence data sets were used to estimate the posterior probability distribution of (i) the TMRCA, and (ii) the time of divergence among populations, using TREES SIFTER 1.0. For these estimations, a δ threshold of 0.1 was chosen, and it was assumed that the population sizes and the time of divergence were unknown, so that these two parameters were varied during the estimation procedure ($N = 100$–10 000 and Time = 100–10 000 generations). Figure 2

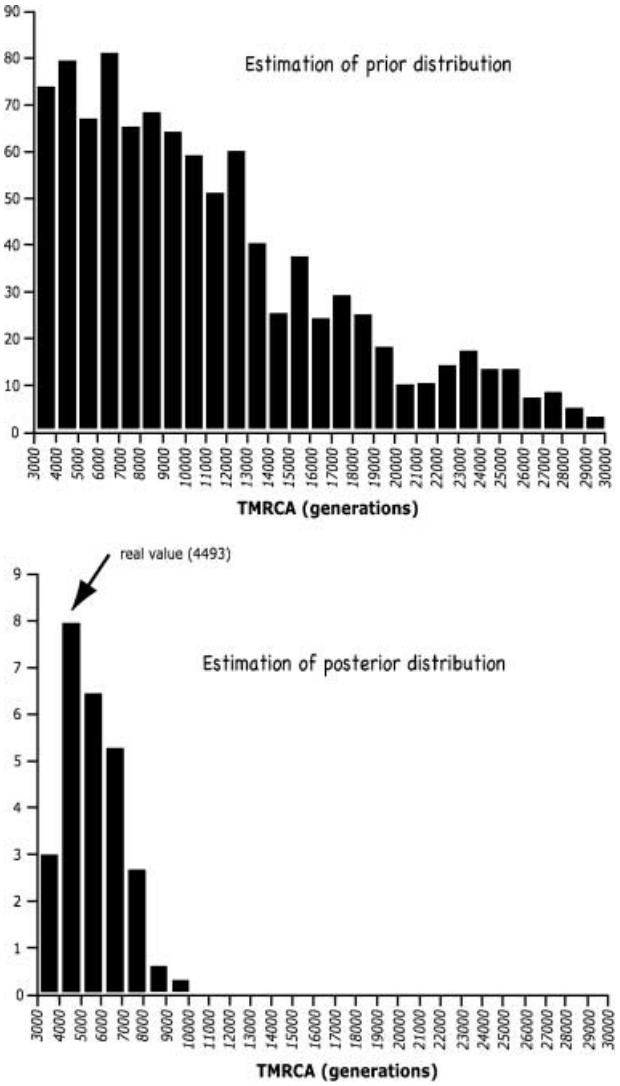

Fig. 2 Estimation of prior and posterior probability distributions of the TMRCA with TREES SIFTER 1.0 using the sequence data simulated along the first simulated genealogy (see text). The prior distribution is based on 1000 simulations, the *y*-axis values being the number of genealogies falling in each category. The posterior distribution estimate is based on 20 000 simulated genealogies (from which 122 only were accepted), the *y*-axis values in this case represent the sum of the weights associated to the genealogies included in each category. An arrow indicates the category of the real TMRCA value.

shows, as an example, the estimated prior and posterior probability distributions for the first simulated genealogy.

In that case, the real value of the TMRCA (4493 generations) is clearly included in the estimated 95% confidence interval. This was the case for 94 simulated genealogies out of 100. The real divergence time (3000 generations for all simulated genealogies), on the other hand, was included in the estimated 95% confidence interval only for 80 genealogies out of 100. It is worth noting that estimating the TMRCA from the simulated sequence data, but this time assuming that we know the time of divergence among the six populations (set to 3000 generations in the model used for the estimation procedure) resulted in the same proportion of correct estimations (94/100), but with an estimated confidence interval in average half the size of the one obtained in the previous estimation test. As expected, increasing the number of unknown variables can therefore considerably increase the uncertainty around the estimate.

Acknowledgements

Raphaël Helaers is gratefully acknowledged for his helpful advice at different stages of writing the program. Bettina Harr and one anonymous reviewer provided useful comments on a previous version of the manuscript and program. P.M. is Research Associate at the Belgian National Fund for Scientific Research (FNRS).

References

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915–925.

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. *Statistical Sciences*, **9**, 307–319.

Griffiths RC, Tavaré S (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical Biosciences*, **127**, 77–98.

Hein J, Schierup MH, Wiuf C (2005) *Gene Genealogies, Variation and Evolution*. Oxford University Press, Oxford.

Hudson RR (1990) Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (eds Futuyma DG, Antonovics J), pp. 1–44. Oxford University Press, Oxford.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

Rambaut A, Grassly NC (1997) SEQ-GEN: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, **13**, 235–238.

Rosenberg NA, Feldman MW (2002) The relationship between coalescence times and population divergence times. In: *Modern Developments in Theoretical Population Genetics* (eds Slatkin M, Veuille M), pp. 130–164. Oxford University Press, New York.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphism. *Nature Review in Genetics*, **3**, 380–390.

Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW (2002) Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics*, **161**, 447–459.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequences. *Genetics*, **145**, 505–518.

Templeton AR (1993) The 'Eve' hypotheses: a genetic critique and reanalysis. *American Anthropologist*, **95**, 51–72.