Hard and Easy Data Sets for Coalescent Likelihood Methods

Peter Beerli and Mary K. Kuhner
Genetics, Box 357360, University of Washington, Seattle WA 98195

Summary

The LAMARC samplers estimate population parameters using a coalescent likelihood method. When mutations are common estimation is accurate. When they are rare the sampler has difficulty estimating all of the branch lengths and produces too-narrow confidence intervals; heating and replication can help. When migrations are rare, the sampler may see asymmetry among migration rates where none really exists; replication and examination of confidence intervals can help here. Intermediate levels of migration work well. At high levels the confidence intervals may be more reliable than the estimate itself. LAMARC appears to do well on a different subset of cases than the Oxford samplers.

LAMARC: Likelihood Analysis with Metropolis algorithms using Random Coalescences – We use a procedure that samples different genealogies by a correlated walk (Markov chain Monte Carlo) through genealogy space to approximate

$$L(P) = \quad \text{Prob}(G|P)\ P(\text{Data}|G)$$

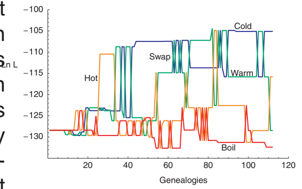
and then find the parameters P at the maximum likelihood. The method is described in detail in Kuhner et al 1985,1999, and Beerli and Felsenstein 1999.

Definitions

L	Likelihood
	4x effective population size x mutation rate per generation per site
4Nm	4x effective population size x migration rate
G	Genealogies
P	Parameters, such as and 4Nm
j,i,z	indicators

Parameter values between log(L) of 0 and -2 are in the 95% confidence interval with a df=1

Replication and Heating

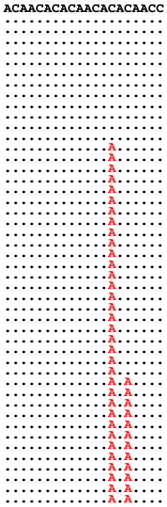


Heating: for difficult problems Markov chain Monte Carlo samplers sometimes get stuck in an isolated mode. This can be overcome by running several independent samplers that run at different "temperatures" and swap occasionally.

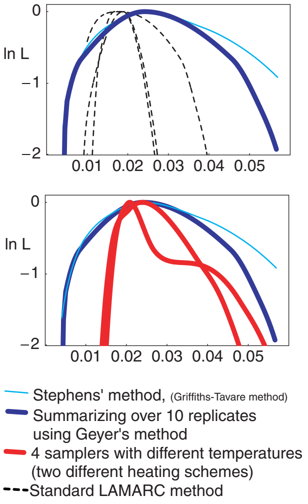
Replication: C. Geyer devised a method that uses a logistic regression scheme to weight different replicates of a Markov chain Monte Carlo run.

$$L(P) = \underset{j}{\overset{\text{Replicates}}{}}\ \underset{i}{\overset{\text{Genealogies}}{}}\ \frac{\text{Prob}(G|P)}{\underset{z}{\overset{\text{Replicates}}{}}\ \frac{\text{Prob}(G_i|P_z)}{L(P_z)}}$$

Low mutation rate



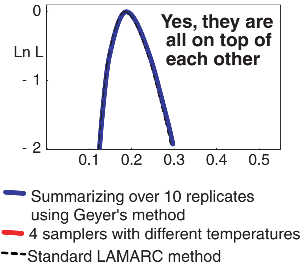
Data sets with few mutations are difficult for the LAMARC samplers because they need to estimate all branch lengths, in stark contrast to the Griffiths-Tavaré samplers (●), and will spend most of their time in rearranging identical sequences on the genealogy, and therefore do not sample effectively among the different possible genealogies. As a result of this under-exploration the confidence intervals of the parameters of interest are too narrow. Heated samplers widen the confidence interval. Replication may come close to the true distribution, but the true distribution is unknown.



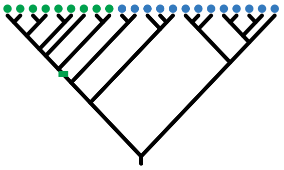
High mutation rate

```
ATGCATGGGAAAAAATTAGG
....................
..A.................
..AA................
..AA................
..A.................
..AA................
..AA................
..AA................
.CA................A
..A................A
..A................A
..A.................
..CA................
..CA........T.......
....................
....................
............G.......
............G.......
..........G.........
..AA................
..CA................
..AA.......G........
```

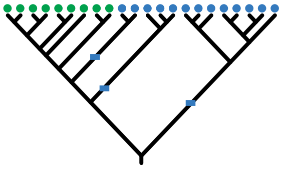
Data with lots of mutation, such as mtDNA D-loop or HIV data, which is difficult for the Griffiths-Tavaré sampler, is easy for the LAMARC sampler. We used 23 samples from the HIV database (1564 bp from the *gag* locus).
Neither heating nor replication seems to be necessary to recover "proper" confidence intervals. Of course, the truth is not known in this case.

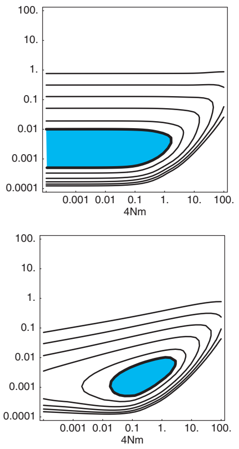


Low migration rate



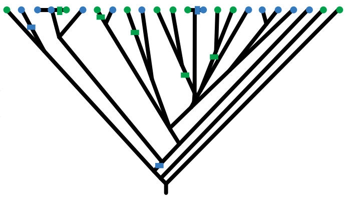
Some data sets do not contain enough information for some population parameters, such as migration between two populations. The data that produced these topologies can fit at least two alternative migration patterns which lead to very different maximum likelihood estimates for the immigration rates. The contour graphs show the likelihood surfaces of the two alternatives (● is the 95% confidence interval). A likelihood ratio test cannot reject equality of migration rate between these two populations. The estimate used 10 replicates.

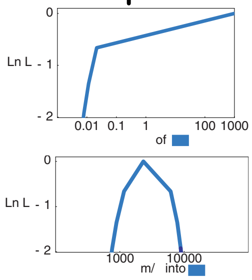




High migration rate

When the migration rate is high (4Nm > 10), and with very variable data and very few sampled individuals, the LAMARC samplers sometimes explore genealogies that are compatible with high values and low 4Nm and will often not return to better regions of the search space. This can produce strange results when we only look at the ML estimates and do not consider the confidence intervals. In these cases the confidence intervals are often very wide, containing both values near zero and very high values (see profile likelihood for (on the right).





Software

http://evolution.genetics.washington.edu/lamarc.html

(●) Griffiths-Tavaré-method is based on work of Griffiths and Tavaré (1994) and extended by Stephens and Donnelly (2000). It works with coalescence and mutation classes and does not need to infer the branch lengths of the genealogies. This works well with low numbers of mutations but gets rather slow with variable data sets. Additionally, current implementations require rather simple mutation models, that are inappropriate for e.g. HIV data or mtDNA. In contrast to the LAMARC samplers it samples the histories in a independent way.

Thanks and Support

We thank Jon Yamato and Joe Felsenstein for discussions, Cliff Cunningham and Stephen Irvin provided the migration data sets. We were supported through grants from NIH and NSF (all to Joe Felsenstein).

Evolution 2000, Bloomington