Idea for species tree estimation using unlinked biallelic markers

Graham Jones, www.indriid.com

2016-02-19, March 14, 2016

1 Introduction

This is aimed at the same sort of analysis as SNAPP (Bryant et al., 2012). The data consists of unlinked biallelic markers such as SNPs. The two alleles are labelled 'red' and 'green'. SNAPP uses 'backwards' equations for the coalescent process. During the likelihood calculation it keeps track of quantities (nearly but not quite partial likelihoods) for each locus and each species tree branch, and for every pair (n, r) where n is a number of lineages and r the number which are red. Here, $1 \leq r \leq n \leq m$, where m is the maximum possible number of lineages for this locus and species tree branch. Thus m varies from the number of markers assigned to a species in external branches up to the total number at the root. The calculation is sophisticated.

The idea here is cruder, and uses 'forward' equations for whole population. Some smallish integer M (perhaps 100 or 1000) is fixed for the analysis. The aim is to approximate the behaviour of a population of any size using one of size M, with suitable scaling. The likelihood calculation is straightforward, using Felsenstein's pruning algorithm with length $M+1$ vectors and $(M+1) \times (M+1)$ rate matrices on the species tree. Suppose X is a species tree branch, and y_X is the data at the tips which are descendants of X. Then the $M+1$ quantities represent partial likelihoods

$$\Pr(y_X | \text{ a fraction } (i/M) \text{ of the population at } X \text{ is red}) \qquad (0 \leq i \leq M).$$

The basic intuition is that since continuous diffusion methods (which use a limit as the population size tends to infinity) can approximate the process for quite small population sizes, the reverse should be true. A modest value for M may serve as a reasonable approximation to any size of population (except a very small one). The conditional probabilities along branches can be calculated using a rate matrix R.

$$\Pr(\text{a fraction } (i/M) \text{ is red} \mid (j/M) \text{ were red time } t \text{ ago}) = \exp(Rt)_{ij}$$

There are other ways the likelihood calculation might be implemented. One could introduce parameters representing fractions of the populations that are red at the ends of branches and sample them during the MCMC. There would be one at each tip, one at each rootward end of a branch, for each locus. That is a lot of parameters, similar to the number of node heights in gene trees for each locus. There are no gene tree topologies to worry about, and no incompatibilities between species tree and the per-locus parameters, which should make the MCMC more efficient, and easier to design operators for. It also seems the most flexible and extendible approach. More, similar, parameters could be added for fractions of nucleotides, and so on. The rest of this note is about integrating things out analytically and approximately.

The time complexity of SNAPP is $O(sn^2)$ per locus for the calculations along the branches and $O(sn^2 \log n)$ per locus for the calculation to merge two branches, where s is the number of species and n the total number of individuals. I think the time complexity for the algorithm here is $O(sM)$ per locus for the calculation to merge two branches, with a small constant, and so not the main problem. For the calculations along the branches, it could be done using eigendecompostion in $O(sM^2)$ per locus, and should be BEAGLE-friendly. Alternatively, the $\exp(Rt)v$ calculations (where v a vector of partial likelihoods) could be done using a Caratheodory-Fejer approximation, like SNAPP. This is $O(sM)$ per locus, with quite a large constant.

2 Population genetics

The Wright-Fisher model from 1931 seems most commonly used, but the Moran model (Moran, 1958) seems to be the most suitable for the current method. This is because the matrix of transition probabilities is tridiagonal, and there are analytical expressions for the eigenvalues. (I don't know if the latter are useful computationally, but the tridiagonal property is good.)

In this model, processes of birth, death and mutation are intertwined. I plan to use a variant of the Moran model introduced in Maruyama (1977), where the birth-death process and the mutation process are regarded as independent. I'll call this the Maruyama-Moran model. Aalto (1989) says 'All practical results from these two models seem to be essentially identical.' Blythe and McKane (2007) make the connection precise.

The model here can be described as follows. Suppose the total number of alleles in the population is M and there are j red alleles.

- A red and a green are converted to two reds at rate $(M-j)j$
- A red and a green are converted to two greens, also at rate $(M-j)j$
- greens mutate to reds at rate $j\alpha$
- reds mutate to greens at rate $j\beta$

The first two processes represent birth and death. One can imagine a population of bacteria, in which two are chosen at random, the first to divide, and the other to die, maintaining the population at M. If the two bacteria have the same allele, there is no change in the number of reds and greens. If they are different, the number changes by one. Since there are $(M-j)j$ such pairs, the rate is proportional to this. The second two processes represent mutation.

The rate matrix (infintesimal stochastic matrix) R is $(M+1)\times(M+1)$ and tridiagonal. The nonzero entries are

$$\begin{aligned} R_{j-1,j} &= j(M-j)+j\alpha \quad (0<j\le M) \text{ (upper)} \\ R_{j,j} &= -\big(2j(M-j)+j\alpha+(M-j)\beta\big) \quad (0\le j\le M) \text{ (diagonal)} \\ R_{j+1,j} &= j(M-j)+(M-j)\beta \quad (0\le j<M) \text{ (lower)} \end{aligned}$$

where we adopt the conventions that rows and columns are numbered from zero, and that columns sum to zero and column vectors of probabilities appear on the right. An example with $M=5$ is below, where the diagonal entries are omitted for clarity.

$$R=\begin{pmatrix} * & 4+\alpha & 0 & 0 & 0 & 0 \\ 5\beta & * & 6+2\alpha & 0 & 0 & 0 \\ 0 & 4+4\beta & * & 6+3\alpha & 0 & 0 \\ 0 & 0 & 6+3\beta & * & 4+4\alpha & 0 \\ 0 & 0 & 0 & 6+2\beta & * & 5\alpha \\ 0 & 0 & 0 & 0 & 4+\beta & * \end{pmatrix}$$

The equilibrium state is a beta-binomial distribution with parameters M,α,β. This is the distribution you get if you sample a beta distribution with parameters α,β to get a value $p\in[0,1]$, then toss a coin which has a probability p of landing heads for M times, and count the heads. The probability of x reds (or heads) is

$$\Pr(x|M,\alpha,\beta)=\frac{\Gamma(M)}{\Gamma(x)\Gamma(M+1-x)}\frac{\Gamma(x-1+\beta)\Gamma(M-x+\alpha)}{\Gamma(M-1+\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

As M tends to infinity, it increasingly resembles a beta distribution which has been stretched to fill the interval $[0,M]$. This equilibrium state is the eigenvector belonging to the eigenvalue 0.

Figure 1 compares the asymptotic distribution with the distribution various M. To me it suggests that a small M is a better approximation than infinite M for large M, at least for the U-shaped distributions which mostly occur in practice

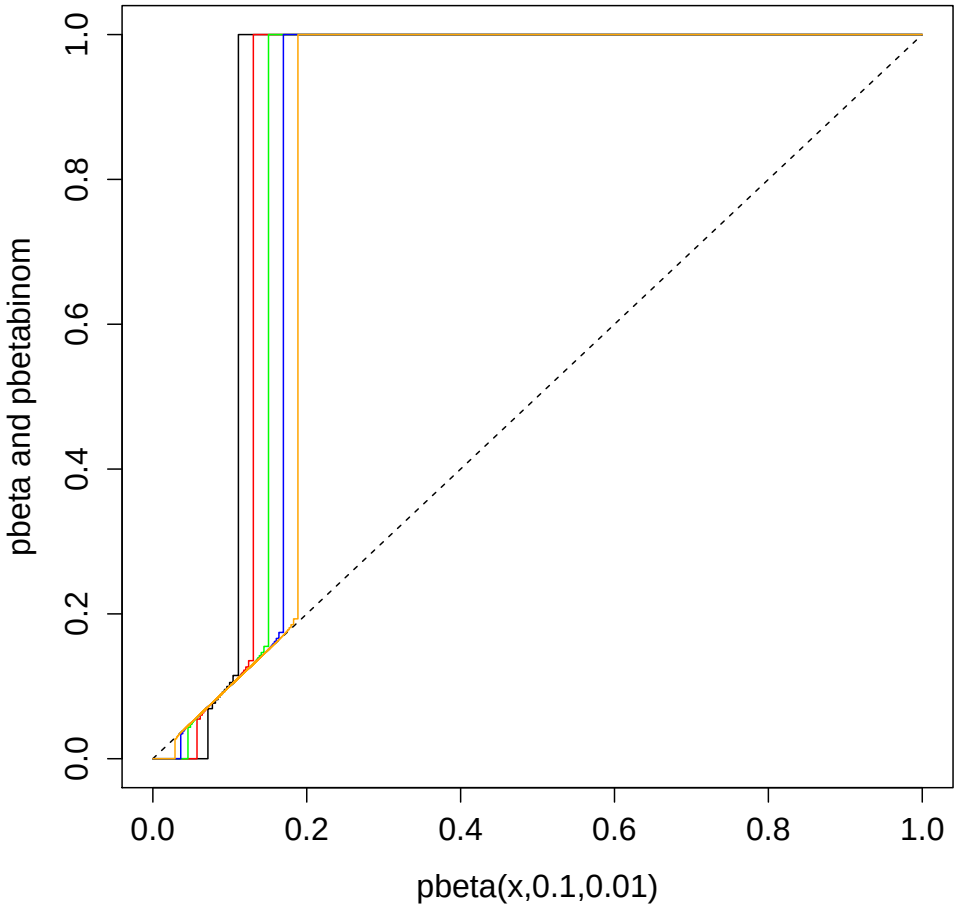

Figure 1: Comparing a Beta(0.1, 0.01) distribution with Betabinomial(M,0.1,0.01). They are sort-of quantile-quantile plots. Dotted is Beta, black is $M = 10$, red is $M = 100$, green is $M = 1000$, blue is $M = 10000$, orange is $M = 100000$.

3 Main calculation

3.1 Scaling for arbitrary population sizes

The approximation goes like this. [TODO check.]

Suppose the number of gene copies in the population is η. (I think it will be convenient for this to be continuous.) First replace α and β in the Maruyama-Moran model with $\eta\mu$ and $\eta\nu$, where μ and ν are the red $\to$ green and green $\to$ red mutation rates, in subtitutions per site per generation. (Roughly speaking, $\eta = 2N_e$, and if θ is the usual population parameter $4N_e u$, then $\eta\mu \approx \eta\nu \approx \theta/2$. However, the Maruyama-Moran model has a different effective population size.) This gives approximately the right shape for the equilibrium distribution. It should be BetaBinomial(η, $\eta\mu$, $\eta\nu$) and the approximation looks similar to this discretized to $\{0, \ldots, M\}$. Then divide all the rates by η, so that the rate at which equilibrium is approached, and transition times between 0 and M are approximately correct for η. (I have probably missed at least one factor of 2 here.) In units of generations,

$$\begin{aligned} R_{j-1,j} &= j(M-j)/\eta + j\mu \quad (0 < j \le M) \text{ (upper)} \\ R_{j,j} &= -\big(2j(M-j)/\eta + j\mu + (M-j)\nu\big) \quad (0 \le j \le M) \text{ (diagonal)} \\ R_{j+1,j} &= j(M-j)/\eta + (M-j)\nu \quad (0 \le j < M) \text{ (lower)} \end{aligned}$$

The equilibrium fractions of red and green are $\nu/(\mu+\nu)$ and $\mu/(\mu+\nu)$. The expected rate of mutation at equilibrium is $2\nu\mu/(\mu+\nu)$. So in units of expected substitutions,

$$\begin{aligned} R_{j-1,j} &= d_j + j(\mu+\nu)/2\nu \quad (0 < j \le M) \text{ (upper)} \\ R_{j,j} &= -\big(2d_j + j(\mu+\nu)/\nu + (M-j)(\mu+\nu)/\mu\big) \quad (0 \le j \le M) \text{ (diagonal)} \\ R_{j+1,j} &= d_j + (M-j)(\mu+\nu)/2\mu \quad (0 \le j < M) \text{ (lower)} \end{aligned}$$

where

$$d_j = \frac{j(M-j)(\mu+\nu)}{2\nu\mu\eta}.$$

So R changes whenever μ, ν or η changes. η varies over branches, but not loci, since μ and ν are assumed the same for all loci. I think there is no absolute information about population sizes, only relative ones, like SNAPP. [TODO non-polymorphic sites?]

TODO. It is not obvious that a population of fixed size M can successfully mimic one of any size, even if the equilibriums are similar, and the overall rate is similar. I have done a few experiments in R. My first impression is that things work well in the middle of the distribution (away from 0 and M) but it cannot match the ends well. It is not clear how this would affect a statistical inference of a species tree. If you get the same results using $M = 100$ and $M = 200$, its probably working OK, otherwise you'll have to try $M = 400$... My intuition is that if M is as large as the number of individuals in the analysis it will be OK, since there is enough resolution to match the available information.

3.2 Tips

At the tips, conditional probabilities for the data are found for each fraction $p_i = (i/M)$ of the population for $i \in \{0, 1, \ldots, M\}$. It seems best to regard the individuals as labelled, so the observation is that there are n specific individuals of which r specific ones are red. So

$$\Pr(\text{data at tip}|i) = p_i^r \left(1 - p_i\right)^{n-r}.$$

This differs from the unlabelled case by a normalisation constant (a binomial coefficient).

For fixed species delimitations, these can be found once for the whole analysis, and the normalisation constant does not matter. If the delimitation is estimated, the observations at the tips can be merged as labelled individuals. If a

second tip has r' red out of n', we get

$$\begin{aligned}\Pr(\text{data for both tips}|i) &= p_i^r (1-p_i)^{n-r} p_i^{r'} (1-p_i)^{n'-r'} \\ &= p_i^{r+r'} (1-p_i)^{n+n'-r-r'} .\end{aligned}$$

This means that the likelihood for the merged tips is the same as that for two tips with a zero height for their MRCA. (See section 3.4 below.) If the unlabelled version was used, there would be normalisation constants to worry about, similar to Leaché et al. (2014).

3.2.1 Dominant markers

If I understand this correctly, there is an observation that r (labelled) individuals out of n are red, where red is dominant. The formula is then

$$\Pr(\text{data at tip}|i) = (p_i (2-p_i))^r (1-p_i)^{2(n-r)} .$$

3.3 Along branches

3.3.1 Eigendecomposition

For this, the eigendecomposition

$$R = V^{-1} \Lambda V$$

where V contains the eigenvectors, and the diagonal matrix Λ contains the eigenvalues can be used. The calculation is right to left for each partial likelihood v:

$$R = V^{-1}(\Lambda(Vv))$$

I assume BEAGLE does this.

The eigenvalues and the left and right eigenvectors for the Moran model were expressed analytically in Karlin and McGregor (1962). The eigenvalues for the Maruyama-Moran model are

$$k(k-1+\alpha+\beta) \quad (0 \leq k \leq M).$$

[TODO: I think this is just a matter of translating the Moran model to the Maruyama-Moran model using Blythe and McKane (2007). In any case, the numerical evidence is strong!]

The expressions for eigenvectors for the Moran model are in terms of Hahn polynomials. [TODO: I think the eigenvectors are the same for the Maruyama-Moran model too.] From a computational point of view, it seems better to exploit the tridiagonal form of R to find the eigenvectors. [TODO understand Hahn polynomials.] This allows the eigenvectors to be found in $O(M^2)$ time. A naive algorithm may have numerical issues (Fernando, 2006). This article may solve the issue too [TODO]. Even if $O(M^3)$ is needed, the result will be used for all loci for at least one branch.

3.3.2 Sidje and Caratheodory-Fejer methods

From SNAPP source code comments, file `likelihood/MatrixExponentiator.java`, for methods `expmv()` and `cf_expmv()`.

```
expmv(): computes an approximation of w = exp(t*A)*v for a
general matrix A using Krylov subspace  projection techniques.
It does not compute the matrix exponential in isolation but instead,
it computes directly the action of the exponential operator on the
operand vector. This way of doing so allows for addressing large
sparse problems. The matrix under consideration interacts only
via matrix-vector products (matrix-free method).
```

```
cf_expmv()+: Uses the Caratheodory-Fejer approximation for the
exponential (on the negative real line) to evaluate the exponential
of A (assumed to be negative semi-definite) times a matrix.
Currently uses 12 degree.
```

These are probably more suitable than eigendecomposition. They are approximations, but the method is an approximation already, and it seems unlikely the exponentiation will be the main source of trouble.

3.4 Merging branches

When a species splits, it assumed that two populations, both of size M are formed instantaneously, with the same fraction of red alleles. The calculation is then just element-wise multiplication. I assume BEAGLE does this.

3.5 At root

The beta-binomial distribution for the equilibrium can be used.

References

Erkki Aalto. The Moran model and validity of the diffusion approximation in population genetics. *Journal of theoretical biology*, 140(3):317–326, 1989.

R A Blythe and A J McKane. Stochastic models of evolution in genetics, ecology and linguistics. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(07):P07018, 2007. URL `http://stacks.iop.org/1742-5468/2007/i=07/a=P07018`.

D Bryant, R Bouckaert, J Felsenstein, N A Rosenberg, and A RoyChoudhury. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol*, 29(8):1917–32, 2012. doi: 10.1093/molbev/mss086.

K. V. Fernando. On computing an eigenvector of a tridiagonal matrix. Part I: Basic results. *SIAM. J. Matrix Anal. & Appl.*, 18(4):1013–1034, 2006.

Samuel Karlin and James McGregor. On a genetics model of Moran. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 58, pages 299–311. Cambridge Univ Press, 1962.

A D Leaché, M K Fujita, V N Minin, and R R Bouckaert. Species delimitation using genome-wide snp data. *Systematic Biology*, 63(4):534–542, 2014.

Takeo Maruyama. *Stochastic problems in population genetics*, volume 17. Springer, Berlin, 1977.

P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54:60–71, 1 1958. ISSN 1469-8064. doi: 10.1017/S0305004100033193.