Some background on the program TreeMix

Stephen (Alex) Townsend and Peter Beerli

The program TreeMix described by Pickrell and Pritchard [2012] uses allele frequencies of SNP (and microsatellite data: https://treemix.googlecode.com/files/microsat_model.pdf) to infer trees representing the evolutionary relationships between populations. Trees generally have a branch-and-leaf structure. In TreeMix the leaves are the populations and the branches are the inferred relationships among them. Additionally, TreeMix can account for situations where one more branches may lead to the same leaf suggesting population admixture and migration between populations.

I. Basic assumptions

TreeMix uses allele frequency data of single nucleotide polymorphisms (SNPs) assuming that only two alleles are recognized. Additionally, TreeMix has an option to run microsatellite data using a mean number of repeats per population per locus. For the method to work well, the allele frequency data should be accurate, thus it needs a decent number of sampled individuals in each population. The model also assumes a Wright-Fisher population model and that the populations are in Hardy-Weinberg Equilibrium.

II. Model and Implementation

TreeMix infers a bifurcating tree between the populations based on the maximum likelihood criterion. In a second step, populations or nodes that do not fit well into the tree will be found. The likelihood score is improved by adding additional branches between such nodes. Any additional branch can be interpreted as a migration event that led to admixture in the leaf population.

A. Brownian motion

Suppose we have a single SNP for which we have a *known* allele frequency data x_A. This allele frequency is the one in the ancestral population A. A second population B with unknown frequency X_B could be the 'offspring' of A. So that we can say

$$x_A \longrightarrow X_B \tag{1}$$

Assuming that changes in the frequency x_A are small per unit time we can assume that the allele frequencies follow a random walk. At any time there is chance to move the frequency up or down (Brownian motion). This process can be modeled assuming that X_B is actually x_A plus an error term:

$$X_B = x_A + \varepsilon_B, \tag{2}$$

the error term can be modeled as a Normal distribution with mean zero and standard deviation relative to x_A. Assuming that allele frequencies come from a Wright-Fisher population model we get

$$\varepsilon_B \sim N(0, c_B x_A(1 - x_A)) \tag{3}$$

If we wanted to move from population X_B to population X_C on our tree, we repeat the process and get

$$X_C = X_B + \varepsilon_C. \tag{4}$$

Since we already know that X_B can be modeled as x_A plus a normal-distributed error term, we have:

$$X_C = x_A + \varepsilon_B + \varepsilon_C. \tag{5}$$

The error term for X_C is modeled as

$$\varepsilon_C \sim N(0, c_C(1 - X_B)X_B). \tag{6}$$

Since we only know the allele frequency at the root, x_A, we may be more interested in the expected frequencies and variances than the actually (unknown) frequencies; we can calculate these

$$E(X_C) = E(x_A + \varepsilon_B + \varepsilon_C) = x_A \tag{7}$$

and

$$\begin{aligned} Var(X_C) &= Var(x_A + \varepsilon_B + \varepsilon_C) && (8)\\ &= Var(\varepsilon_B) + Var(\varepsilon_C) + Cov(\varepsilon_B, \varepsilon_C) && (9)\\ & && (10) \end{aligned}$$

Genetic drift is usually small so that we can assume that $X_B(1 - X_B \approx x_A(1 - x_A)$ and we further assume that there is no correlation between X_B and X_C so that we can ignore the covariance $Cov(\varepsilon_B, \varepsilon_C)$. We get for

$$\begin{aligned} Var(X_C) &\approx Var(\varepsilon_B) + Var(\varepsilon_C) && (11)\\ &\approx c_B x_A(1 - x_A) + c_C x_A(1 - x_A) && (12)\\ &\approx (c_B + c_C)x_A(1 - x_A) && (13) \end{aligned}$$

so we can see that X_C is

$$X_C \sim N(x_A, (c_B + c_C)x_A(1 - x_A)) \tag{14}$$

and X_B is

$$X_B \sim N(x_A, c_B x_A(1 - x_A)). \tag{15}$$

This suggests that the populations differentiation can be measured with the coefficient c. Using the Wright-Fisher model we can approximate

$$c = \frac{t}{2N} \tag{16}$$

where t is the time between ancestor and descendant and N is the population size. The parameter c can be thought of as the length of a branch on the tree.

III. The tree

If we are now interested in the tree (Figure 1) than we can calculate the covariance between X_1 and X_2 as the expected values of the differences to x_A:

$$\begin{aligned} Cov(X_1, X_2) &= E\left((X_1 - x_A)(X_2 - x_A)\right) && (17) \\ &= c2x_A(1 - x_A) && (18) \\ Cov(X_3, X_4) &= c1x_A(1 - x_A) && (19) \\ Cov(X_1, X_3) &= 0 && (20) \end{aligned}$$

If there is migration between different populations than the covariance matrix will fit the data not very well. To accommodate a better fit, one need to allow ancestry from multiple populations, this contributions is weighted by w. For example, Figure 2 and Table I shows the contribution to the covariance for three leaves (X_1, X_2, X_3), assuming that one (X_2) is admixed. We know now the structure of the covariance matrix V and approximate this with W which we estimate from the data

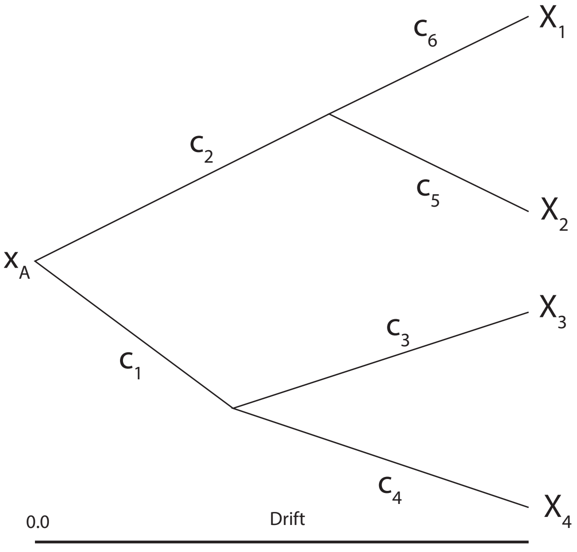

Figure 1. Figure of a population tree with 4 populations and labeled branch lengths.

Table I
Relationships among the three leaves in Figure 2. Each entry needs to be multiplied with $x_A(1 - x_A)$

	X_1	X_2	X_3
X_1	$V_{11} = c_4 + c_6$	$V_{12} = (1 - w)c_4$	$V_{13} = 0$
X_2		$V_{22} = c_1 + w^2(c_2 + c_5)$ $+(c_3 + c_4)(1 - w)^2$	$V_{23} = wc_5$
X_3			$V_{33} = c_5 + c_7$

$$W_{ij} = E[(X_i - \hat{\mu})(X_j - \hat{\mu})] \tag{21}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{m} X_i}{m} \tag{22}$$

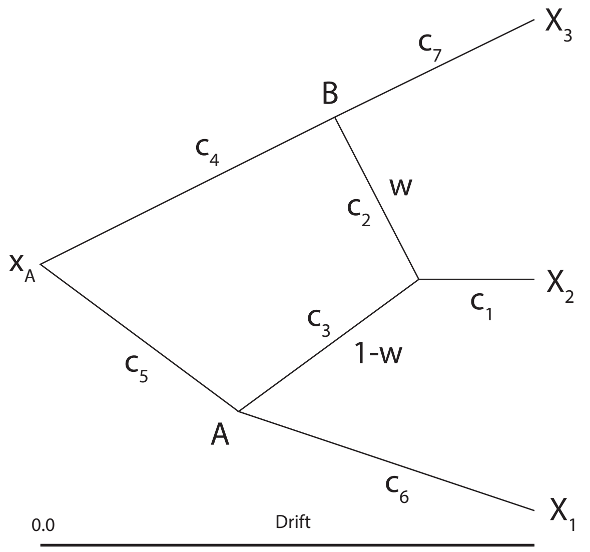

Figure 2. Tree showing admixture calculation

This will be biased because of finite sampling. The authors supply a correction for finite sampling

$$W_{ij} = V_{ij} - \frac{1}{m}\left(\sum_{k}^{m}(V_{ik} + V_{kj})\right) + \frac{1}{m^2}\sum_{k}^{m}\sum_{k'}^{m} V_{kk'} \tag{23}$$

where we assume that X_i and X_j are the allele frequencies from the sample data and that m is the number of populations in the sample. This covariance matrix assumes 1 SNP for m populations. For n SNPs we simply take the averages of all individual allele frequencies. We know now how to calculate the covariance matrix from the data W, but what we want is the covariance structure $\hat{W}$ given the tree G If we know everything then we can calculate a likelihood as

$$P(W|\hat{W}) = \prod_{i=1}^{m}\prod_{j=1}^{m} f_{ij} \tag{24}$$

$$f_{ij} \sim N(W_{ij}, G\hat{\sigma}^2_{ij}) \tag{25}$$

where G is a weighted, directed graph and the term $N(W_{ij}, G\hat{\sigma}^2_{ij}$ is a Normal distribution whose mean is the covariance of the current tree and where $\hat{\sigma}^2_{ij}$ is the current sample variance. Using a maximization routine we can find the optimal $\hat{W}$ for the data. The methods starts out by assuming there is no migration and calculates covariances, then it uses the residuals

$$R = \hat{W} - W \tag{26}$$

to find the components that do not fit well and adjust those components by adding branches to improve the fit. If there is rampant migration, this model may struggle to be able to account for that due to the fact that it only adds migration to the populations with the largest residuals rather than considering all possible trees. Taking all possible trees is computationally infeasible for data with many populations.

IV. Significance testing

The likelihood is a composite likelihood that assumes that blocks of SNPs are independent of each other, this may or may not be true. Standard confidence intervals based on the Fisher information will be too conservative. Instead, the methods uses a resampling technique, jackknifing, to test of significance and confidence intervals.

Bibliography

Joseph K Pickrell and Jonathan K Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, 8(11):e1002967, 11 2012.

V. Disclaimer

This text was written by Stephen (Alex) Townsend and Peter Beerli, Florida State University for a course on practical population genetics inference, Fall 2015.