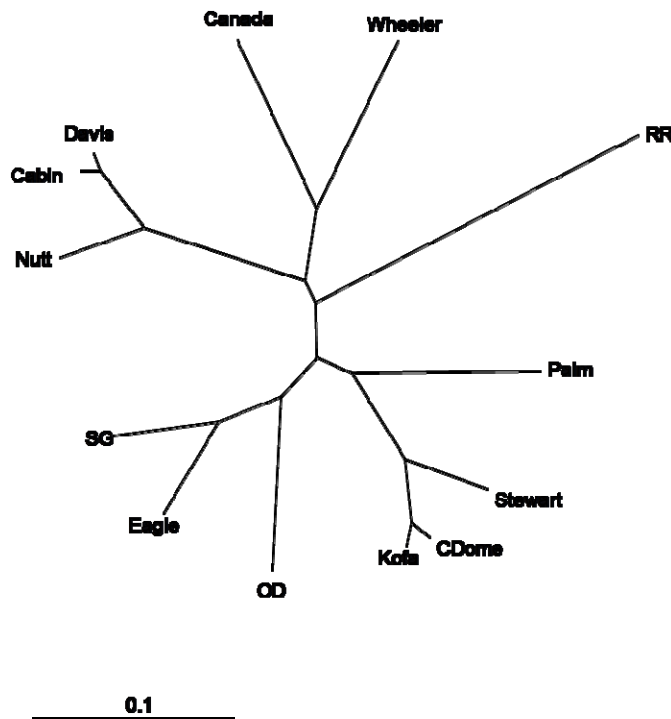


TREEFIT

A COMPUTER PROGRAM FOR EVALUATING HOW WELL EVOLUTIONARY
TREES FIT GENETIC DISTANCE DATA

Steven T Kalinowski
310 Lewis Hall
Department of Ecology
Montana State University
Bozeman, MT 59717
skalinowski@montana.edu
(406) 994-3232

Introduction



Unrooted evolutionary trees are frequently used to describe genetic relationships among populations. They are interpreted by measuring the length of the branches separating each population. For example, in the tree shown at left, populations “Cabin” and “Davis” are more similar to each other than they are to population “Nutt.” Hierarchical, bifurcating trees are a reasonable model for the evolution of DNA sequences and species, but may or may not be appropriate for describing patterns of genetic similarity among in populations connected by gene flow. For example, if populations are arranged in a stepping stone pattern (either one or two dimensional), the genetic relationships between populations may not follow a hierarchical pattern,

and traditional neighbor-joining or UPGMA trees may not be appropriate tools for describing the structure of such populations.

The most commonly used approach for constructing evolutionary trees for populations is to compare allele frequencies among the populations. Genetic differences between populations are summarized with a matrix of pair wise genetic distance (e.g. pair wise F_{ST}), and a tree is then constructed from these distances so that the relationships in tree are similar to the relationships in the matrix of genetic distances. The two most commonly used distance based methods for building trees are the unweighted pair group method with arithmetic mean (UPGMA), and neighbor-joining (NJ). Both methods build trees by searching the genetic distance matrix for the most similar populations (i.e. the ones with the smallest distance between them), and then connecting these populations at a node. Once populations are connected, they are removed from the distance matrix and replaced with the node connecting them. The tree building algorithm continues until all populations are connected in a tree.

The purpose of TreeFit is to analyze how well a tree fits the genetic data the tree was calculated from. TreeFit creates NJ and UPGMA trees from a genetic distance matrix, and then compares the observed genetic distance between populations with the genetic distance in the tree. The main output from TreeFit, is an R^2 value for a tree—the proportion of variation in the genetic distance matrix that is explained by the tree.

Input files

TreeFit reads two types input files: distance matrix files (example shown below), and GENEPOP genotype files. If a GENEPOP file is used as input, Weir and Cockerham's (1984) theta is used a genetic distance. An example of a distance matrix file for the microsatellite data of Gutiérrez-Espeleta (2000) is show below.

```
Pairwise FST for BHS microsatellite data of Gutierrez et al. (2000)
Cabin
CDome      0.24
Davis      0.02 0.25
Eagle      0.20 0.13 0.20
Canada     0.23 0.26 0.24 0.21
Kofa       0.22 0.02 0.24 0.10 0.24
Nutt       0.07 0.32 0.10 0.27 0.28 0.34
OD         0.29 0.29 0.31 0.13 0.31 0.29 0.37
Palm       0.31 0.20 0.30 0.14 0.32 0.18 0.39 0.26
RR         0.36 0.25 0.38 0.23 0.31 0.24 0.41 0.35 0.30
SG         0.34 0.28 0.33 0.11 0.31 0.27 0.39 0.22 0.25 0.39
Stewart    0.27 0.07 0.26 0.15 0.27 0.10 0.34 0.27 0.19 0.24 0.30
```


The main function of TreeFit is to calculate the proportion of variation, R^2 , in the genetic distance matrix that is explained by the tree. This is done in the usual manner,

$$R^2 = 1 - \sum \frac{(D_{ij} - d_{ij})^2}{(D_{ij} - \bar{D})^2}$$

where summation is taken over all pairs of populations (i.e. all elements in the genetic distance matrix). If R^2 is near 1.0, the tree represents a good summary of the genetic relationships shown in the distance matrix.

Output

TreeFit performs two types of analysis. First, it constructs UPGMA or NJ trees and outputs the tree as a TreeView file. (TreeView is a free computer program available on the internet for displaying and printing trees). Second, TreeFit compares observed genetic distances between populations to the fitted genetic distance between the populations within UPGMA and NJ trees.

The microsatellite data of Gutiérrez-Espeleta (2000) is used as an example throughout this manual. In this study, ten microsatellite loci were genotyped for at 13 populations of bighorn sheep, most of which were in the deserts of Southern California and Arizona.

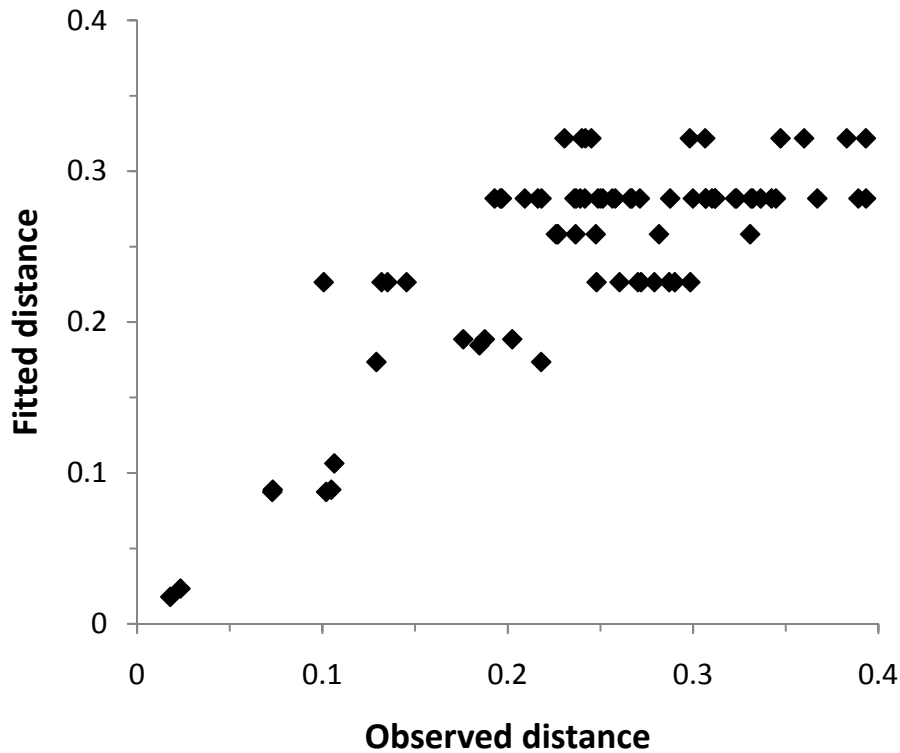
TreeFit output from the data of Gutiérrez-Espeleta (2000) is shown below for a UPGMA tree (the tree is shown as an example in the introduction).

UPGMA Tree

		Obs D	Fitted D
Cabin	CDome	0.2391	0.2820
Cabin	Davis	0.0179	0.0179
Cabin	Eagle	0.1968	0.2820
Cabin	Forbes	0.2270	0.2583
Cabin	Kofa	0.2182	0.2820
Cabin	Nutt	0.0732	0.0890
Cabin	OD	0.2876	0.2820
Cabin	Palm	0.3069	0.2820
Cabin	RR	0.3599	0.3219
Cabin	SG	0.3447	0.2820
Cabin	Stewart	0.2670	0.2820
Cabin	Wheeler	0.2260	0.2583
CDome	Davis	0.2487	0.2820
CDome	Eagle	0.1319	0.2265
CDome	Forbes	0.2564	0.2820
CDome	Kofa	0.0234	0.0234
CDome	Nutt	0.3235	0.2820
CDome	OD	0.2901	0.2265
CDome	Palm	0.2024	0.1887

CDome	RR	0.2451	0.3219
CDome	SG	0.2791	0.2265

The first column of numbers contains the observed genetic distance between pairs of populations (shown at left), and the second column of numbers contains the genetic distance between each pair of populations in the tree. A graph of these data (made in Microsoft Excel) is shown below.



The graph shows that there is a rough correspondence between the observed and fitted genetic distance. The R^2 for this UPGMA tree is 0.61, which is relatively low. The R^2 for a NJ tree for the same data is 0.91, which is much better, but still relatively low for a NJ tree.

Literature Cited

- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates Inc., Sunderland, Massachusetts.
- Gutiérrez-Espeleta GA, ST Kalinowski, WM Boyce, PW Hedrick (2000) Genetic variation and population structure in desert bighorn sheep: implications for conservation. *Conservation Genetics* 1:3-15.
- Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.