# Statistical phylogeography

Peter Beerli

November 13, 2005

## 1 Haplotype networks

Phylogeneticists who turned to population analyses seem to like haplotype networks, these are relationships among haplotypes (not individuals) often measure using a simple genetic distance, such as how many different mutation apart two haplotypes are. In contrast to bifurcating trees, haplotype networks can have multifurcations and, in principle, cycles (biologists might call them loops). Typically haplotype networks presented in papers show only some minimal reprensentation of the network: for example a minimum spanning tree. In figure 1 are the differences among a genealogy of sampled individuals and the haplotype network shown. Haplotype networks are then often colored by population. Inferences that try to establish historical relationships among different populations because of the haplotype network connections are very popular (for example open a random edition of the journal Molecular Ecology).

**Algorithm 1** Prim's algorithm to find a minimum spanning tree. The time required by Prim's algorithm is $O(|V|^2)$ where $V$ is the number of nodes (vertices).

Start with a fully resolved graph (every node would connect to every other (in the extreme).
Pick a random node (vertex) $v$ out of all vertices $V$ and color it red
Remove $v_1$ from the vertices list $V$.
**while** $V$ has more than 1 element **do**
  Find nearest node $v_i$ close to any red colored vertex or edge and color connection red.
  Remove $v_i$ from the vertices list $V$.
**end while**

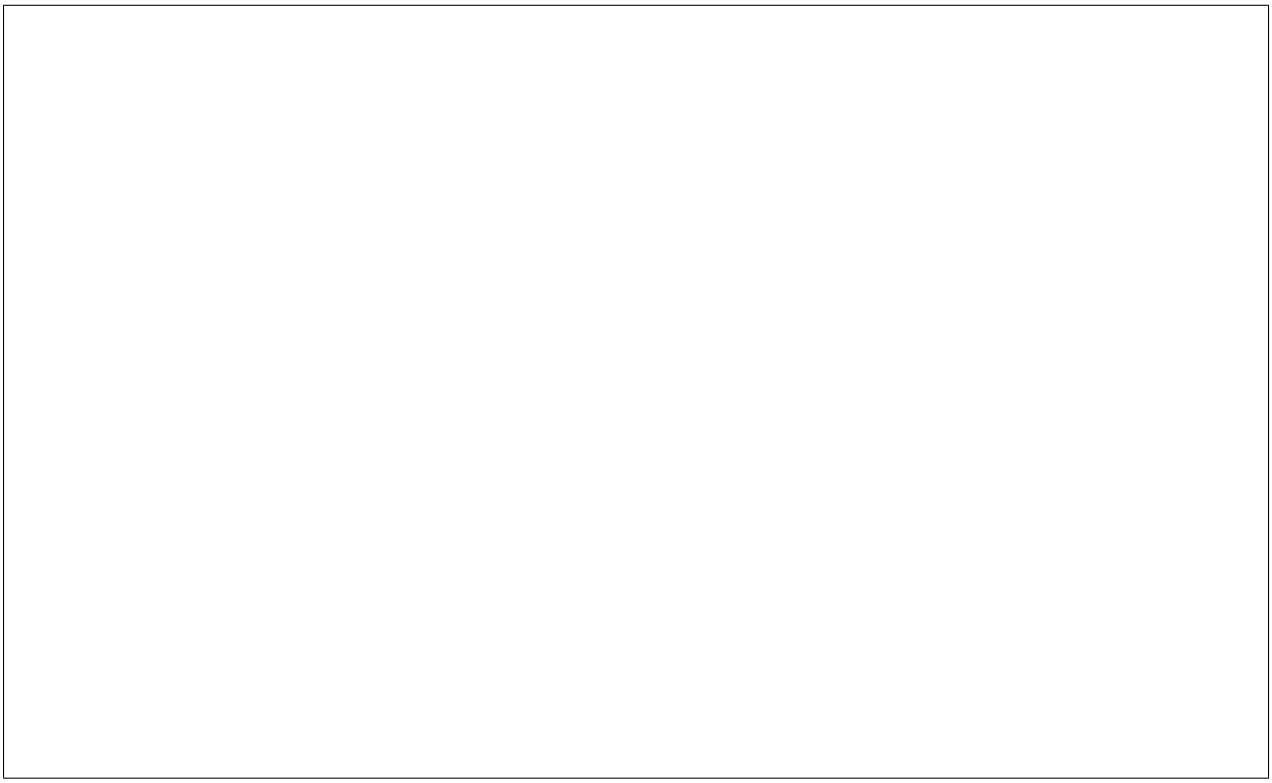

Figure 1: Example of haplotype network and its correlation to a genealogy

## 2 Nested clade population analysis [NCPA]

In 1995 Templeton formalized the correlation between number of haplotypes in a sample and their connection to infer population history

Alan Templeton showed a method that appealed to many biologists but seems to fail a rigorous test with simulated data (fro example Knowles and Maddison 2003). The approach consists of two major steps

**Algorithm 2** Kruskal's algorithm to find a minimum spanning tree. The time required by Kruskal's algorithm is $O(|E|log(|V|))$ where $V$ is the number of nodes (vertices) and $E$ is the number edges (branches).

Start with a fully resolved graph (every node would connect to every other (in the extreme).
Pick the shortest edge (branch) and color it red
**while** all vertices are not connected **do**
  Pick the shortest non-colored edge (branch) and color it red
**end while**

1. construct a network of haplotypes, his own procedure called statistical parsimony. A "minimal" spanning tree is formed using parsimony criteria, each haplotype can connect to other haplotypes using single mutation steps and missing haplotypes are filled in. Posada and Crandall (2001) describe the method this way *The statistical parsimony algorithm [...] begins by estimating the maximum number of differences among haplotypes as a result of single substitutions (i.e. those that are not the result of multiple substitutions at a single site) with a 95% statistical confidence. This number is called the parsimony limit (or parsimony connection limit). After this, haplotypes differing by one change are connected, then those differing by two, by three and so on, until all the haplotypes are included in a single network or the parsimony connection limit is reached. The statistical parsimony method emphasizes what is shared among haplotypes that differ minimally rather than the differences among the haplotypes and provides an empirical assessment of deviations from parsimony.*

2. Use this network and and construct groups that are 1 mutation, 2 mutation etc apart using Templeton's key (2004).

Templeton's key uses the coalescent and common sense reasoning. Under the assumption of the netural coalescent we can make some suggestions what to expect if we run an experiment many times. Remember that the coalescent is a noisy distribution and the variability is large, up to the mean squared.

- Older alleles have higher frequency.
- Older alleles are distributed over a larger geographic range.
- Haplotypes with greater frequency have more connections.
- Singletons are more often connected to nonsingletons.
- Singletons are more likley connected to the haplotypes in the same population.

The list above was mentioned by Posada and Crandall (2001), see figure 2 for position of events.

Knowles and Maddison (2002) studied the success of the method and found that out of ten replicates the answers are not different than random guesses.

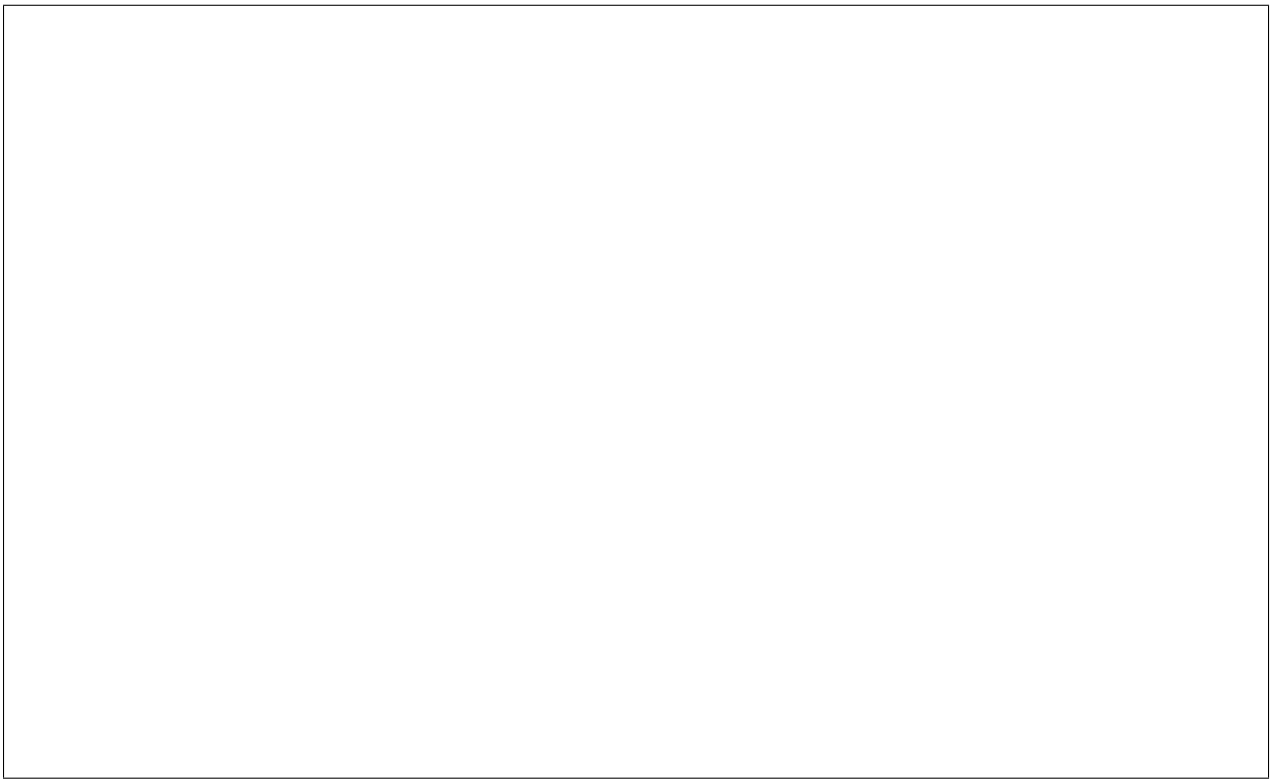

Figure 2: Coalescent and hapltoype network

## 3 Statistical phylogeography

Phylogeography was first used by Avise who mapped phylogenetic relationships of mtDNA onto maps. We came a long way from haphazard interpretation of such maps to statistical phylogepography Statistical phylogeography aims to infer the history and processes of population interaction, and speciation, and to provide objective, rather than ad hoc explanations. Such inferences are based on model-based approaches (new news here we nost talk about these).