PBGG Notes 2011: The use of assignment methods in population genetics.

Graham Coop[1]

[1] Department of Evolution and Ecology & Center for Population Biology,

Here I describe a simple probabilistic assignment to find the probability that an individual of unknown population comes from one of K predefined populations. I then briefly explain how to extend this to cluster individuals into K initially unknown populations. This method is a simplified version of what Bayesian population genetics clustering algorithms such as STRUCTURE (Pritchard et al. Genetics 2000).

A simple assignment method We have genotype data from unlinked S bi-allelic loci for K populations. The allele frequency of allele 1 at locus l in population k is denoted by $p_{k,l}$, so that the allele frequencies in population 1 are $p_{1,1}, \cdots p_{1,L}$ and population 2 are $p_{2,1}, \cdots p_{2,L}$ and so on.

You type a new individual from an unknown population at these L loci. This individual's genotype at locus l is g_l, where g_l denotes the number of copies of allele 1 this individual carries at this locus $g_l = 0, 1, 2)$.

The probability of this individual's genotype at locus l conditional on coming from population k (i.e. their alleles being a random HW draw from population k) is

$$P(g_l|\text{pop k}) = I(g_l = 0)(1-p_{k,l})^2 + I(g_l = 1)2p_{k,l}(1-p_{k,l}) + I(g_l = 2)p_{k,l}^2 \quad (1)$$

where $I(g_l = 0)$ is an indicator function which is 1 if $g_l = 0$ and zero otherwise, and likewise for the other indicator functions.

Therefore, assuming that the loci are independent, the probability of individual's genotypes conditional on them coming from population k is

$$P(\text{new ind.}|\text{pop k}) = \prod_{l=1}^{S} P(g_l|\text{pop k}) \quad (2)$$

We wish to know the probability that this new individual comes from population k, i.e. $P(\text{pop k}|\text{new ind.})$. We can obtain this through Bayes rule

$$P(\text{pop k}|\text{new ind.}) = \frac{P(\text{new ind.}|\text{pop k})P(\text{pop k})}{P(\text{new ind.})} \tag{3}$$

where

$$P(\text{new ind.}) = \sum_{k=1}^{K} \frac{P(\text{new ind.}|\text{pop k})P(\text{pop k})}{P(\text{new ind.})} \tag{4}$$

is the normalizing constant. We interpret $P(\text{pop k})$ as the prior probability of the individual coming from population k, unless we have some prior knowledge we will assume that the new individual has a equal probability of coming from each population $P(\text{pop k}) = 1/K$.

We intepret

$$P(\text{pop k}|\text{new ind.}) \tag{5}$$

as the posterior probability that our new individual comes from each of our $1, \cdots, K$ populations.

More sophisticated versions of this are now used to allow for hybrids, e.g, we can have a proportion q_k of our individual's genome come from population k and estimate the set of q_k's.

clustering based on assignment methods We wish to cluster our individuals into K unknown populations. We begin by assigning our individuals at random to these K populations.

- Given these assignments we estimate the allele frequencies at all of our loci in each population.
- Given these allele frequencies we chose to reassign each individual to a population k with a probability given by eqn. (2).

We iterate steps 1 and 2 for many iterations. If the data is sufficiently informative the assignments and allele frequencies will quickly converge.

Technically the iterations of this procedure represents (correlated) draws from the joint posterior of our allele frequencies and assignments. This is a Monte Carlo Markov Chain algorithm to explore this joint posterior.