Algorithms and equations utilized in poppr version 1.1.0.99

Zhian N. Kamvar[1] and Niklaus J. Grünwald[1,2]

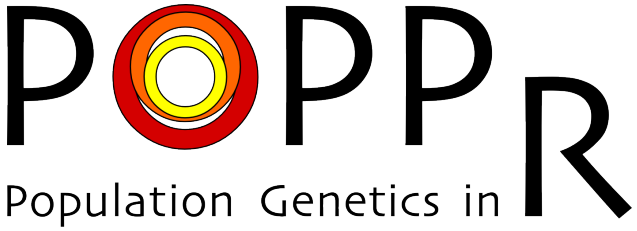
1) Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR

2) Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, OR

May 23, 2014

Abstract

This vignette is focused on simply explaining the different algorithms utilized in calculations such as the index of association and different distance measures. Many of these are previously described in other papers and it would be prudent to cite them properly if they are used.



Contents

1 The Index of Association

The index of association was originally developed by A.H.D. Brown analyzing population structure of wheat [2]. It has been widely used as a tool to detect clonal reproduction within populations [4]. Populations whose members are undergoing sexual reproduction, whether it be selfing or out-crossing, will produce gametes via meiosis, and thus have a chance to shuffle alleles in the next generation. Populations whose members are undergoing clonal reproduction, however, generally do so via mitosis.

This means that the most likely mechanism for a change in genotype is via mutation. The rate of mutation varies from species to species, but it is rarely sufficiently high to approximate a random shuffling of alleles. The index of association is a calculation based on the ratio of the variance of the raw number of differences between individuals and the sum of those variances over each locus [4]. It can also be thought of as the observed variance over the expected variance. If both variances are equal, then the index is zero after subtracting one (from Maynard-Smith, 1993 [4]):

$$I_A = \frac{V_O}{V_E} - 1 \tag{1}$$

Any sort of marker can be used for this analysis as it only counts differences between pairs of samples. This can be thought of as a distance whose maximum is equal to the number of loci multiplied by the ploidy of the sample. This can be calculated using Kronecker's delta, or, in the case of *poppr*, absolute genetic distance. Remember that in *poppr*, genetic data is stored in a table where the rows represent samples and the columns represent potential allelic states grouped by locus. Notice also that the sum of the rows all equal one. *Poppr* uses this to calculate distances by simply taking the sum of the absolute values of the differences between rows.

The calculation for the distance between two individuals at a single locus with a allelic states and a ploidy of k is as follows[1]:

$$d = \frac{k}{2} \sum_{i=1}^{a} \mid ind_{Ai} - ind_{Bi} \mid \tag{2}$$

As you can see, these values of d at locus one add up to 2, 2, and 1, respectively.

To find the total number of differences between two individuals over all loci, you just take d over m loci, a value we'll call D:

$$D = \sum_{i=1}^{m} d_i \tag{3}$$

These values are calculated over all possible combinations of individuals in the data set, $\binom{n}{2}$ after which you end up with $\binom{n}{2} \cdot m$ values of d and $\binom{n}{2}$ values of D. Calculating the observed variances is fairly straightforward (modified from Agapow and Burt, 2001) [1]:

$$V_O = \frac{\sum_{i=1}^{\binom{n}{2}} D_i^2 - \frac{(\sum_{i=1}^{\binom{n}{2}} D_i)^2}{\binom{n}{2}}}{\binom{n}{2}} \tag{4}$$

[1]Individuals with Presence / Absence data will have the $k/2$ term dropped.

Calculating the expected variance is the sum of each of the variances of the individual loci. The calculation at a single locus, j is the same as the previous equation, substituting values of D for d [1]:

$$var_j = \frac{\sum_{i=1}^{\binom{n}{2}} d_i^2 - \frac{(\sum_{i=1}^{\binom{n}{2}} d_i)^2}{\binom{n}{2}}}{\binom{n}{2}} \tag{5}$$

The expected variance is then the sum of all the variances over all m loci [1]:

$$V_E = \sum_{j=1}^{m} var_j \tag{6}$$

Now you can plug the sums of equations (4) and (6) into equation (1) to get the index of association. Of course, Agapow and Burt showed that this index increases steadily with the number of loci, so they came up with an approximation that is widely used, $\bar{r}_d$ [1]. For the derivation, see the manual for *multilocus*. The equation is as follows, utilizing equations (4), (5), and (6) [1]:

$$\bar{r_d} = \frac{V_O - V_E}{2\sum_{j=1}^{m}\sum_{k\neq j}^{m}\sqrt{var_j \cdot var_k}} \tag{7}$$

2 Bruvo's distance

Bruvo's distance between two individuals calculates the minimum distance across all combinations of possible pairs of alleles at a single locus and then averaging that distance across all loci [3]. The distance between each pair of alleles is calculated as [3]:

$$m_x = 2^{-|x|} \tag{8}$$

$$d_a = 1 - m_x \tag{9}$$

Where x is the number of steps between each allele. So, let's say we were comparing two haploid ($k = 1$) individuals with alleles 228 and 244 at a locus that had a tetranucleotide repeat pattern $(CATG)^n$. The number of steps for each of these alleles would be $228/4 = 57$ and $244/4 = 61$, respectively. The number of steps between them is then $\mid 57 - 61 \mid = 4$. Bruvo's distance at this locus between these two individuals is then $1 - 2^{-4} = 0.9375$. For samples with higher ploidy (k), there would be k such distances of which we would need to take the sum [3].

$$s_i = \sum_{a=1}^{k} d_a \tag{10}$$

Unfortunately, it's not as simple as that since we do not assume to know phase. Because of this, we need to take all possible combinations of alleles into account. This means that we will have k^2 values of d_a, when we only want k. How do we know which k distances we want? We will have to invoke parsimony for this and attempt to take the minimum sum of the alleles, of which there are $k!$ possibilities [3]:

$$d_l = \frac{\left(\min_{i\ldots k!} s_i\right)}{k} \tag{11}$$

Finally, after all of this, we can get the average distance over all loci [3].

$$D = \frac{\sum_{i=1}^{l} d_i}{l} \tag{12}$$

This is calculated over all possible combinations of individuals and results in a lower triangle distance matrix over all individuals.

2.1 Special Cases of Bruvo's distance

As shown in the above section, ploidy is irrelevant with respect to calculation of Bruvo's distance. However, since it makes a comparison between all alleles at a locus, it only makes sense that the two loci need to have the same ploidy level. Unfortunately for polyploids, it's often difficult to fully separate distinct alleles at each locus, so you end up with genotypes that appear to have a lower ploidy level than the organism [3].

To help deal with these situatons, Bruvo has suggested three methods for dealing with these differences in ploidy levels [3]:

- Infinite Model - The simplest way to deal with it is to count all missing alleles as infinitely large so that the distance between it and anything else is 1. Aside from this being computationally simple, it will tend to inflate distances between individuals.
- Genome Addition Model - If it is suspected that the organism has gone through a recent genome expansion, the missing alleles will be replace with all possible combinations of the observed alleles in the shorter genotype. For example, if there is a genotype of [69, 70, 0, 0] where 0 is a missing allele, the possible combinations are: [69, 70, 69, 69], [69, 70, 69, 70], and [69, 70, 70, 70]. The resulting distances are then averaged over the number of comparisons.
- Genome Loss Model - This is similar to the genome addition model, except that it assumes that there was a recent genome reduction event and uses the observed values in the full genotype to fill the missing values in the short genotype. As with the Genome Addition Model, the resulting distances are averaged over the number of comparisons.
- Combination Model - Combine and average the genome addition and loss models.

As mentioned above, the infinite model is biased, but it is not nearly as computationally intensive as either of the other models. The reason for this is that both of the addition and loss models requires replacement of alleles and recalculation of Bruvo's distance. The number of replacements required is equal to the multiset coefficient: $\left(\binom{n}{k}\right) == \binom{(n-k+1)}{k}$ where n is the number of potential replacements and k is the number of alleles to be replaced. So, for the example given above, The genome addition model would require $\left(\binom{2}{2}\right) = 3$ calculations of Bruvo's distance, whereas the genome loss model would require $\left(\binom{4}{2}\right) = 10$ calculations.

To reduce the number of calculations and assumptions otherwise, Bruvo's distance will be calculated using the largest observed ploidy. This means that when comparing [69,70,71,0] and [59,60,0,0], they will be treated as triploids.

References

[1] Paul-Michael Agapow and Austin Burt. Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, 1(1-2):101–102, 2001.

[2] A.H.D. Brown, M.W. Feldman, and E. Nevo. Multilocus structure of natural populations of hordeum spontaneum. *Genetics*, 96(2):523–536, 1980.

[3] Ruzica Bruvo, Nicolaas K. Michiels, Thomas G. D'Souza, and Hinrich Schulenburg. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology*, 13(7):2101–2106, 2004.

[4] J M Smith, N H Smith, M O'Rourke, and B G Spratt. How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90(10):4384–4388, 1993.