Algorithms and equations utilized in poppr version 2.8.2

Zhian N. Kamvar\textsuperscript{1} and Niklaus J. Grünwald\textsuperscript{1,2}
\textsuperscript{1) Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR}
\textsuperscript{2) Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, OR}

March 11, 2019

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.
1 Mathematical representation of data in adegenet and poppr

The sections dealing with the index of association and genetic distances will be based on the same data structure, a matrix with samples in rows and alleles in columns. The number of columns is equal to the total number of alleles observed in the data set. Much of this description is derived from adegenet’s dist.genpop manual page.

1.1 Caveat

The information here describes the data as it would be passed into the distance functions. Data represented in the genind objects are now counts of alleles instead of frequencies of alleles. Otherwise, these equations still hold.

Let $A$ be a table containing allelic frequencies with $t$ samples\(^1\) (rows) and $m$ alleles (columns).

---

\(^1\)populations or individuals
Let \( \nu \) be the number of loci. The locus \( j \) gets \( m(j) \) alleles.

\[
m = \sum_{j=1}^{\nu} m(j)
\]

So, if you had a data set with 5 loci that had 2 alleles each, your table would have ten columns. Of course, codominant loci like microsatellites have varying numbers of alleles.

For the row \( i \) and the modality \( k \) of the variable \( j \), notice the value

\[
a_{ijk}(1 \leq i \leq t, 1 \leq j \leq \nu, 1 \leq k \leq m(j))
\]

\[
a_{ij} = \sum_{k=1}^{m(j)} a_{ijk}
\]

\[
p_{ijk} = \frac{a_{ijk}}{a_{ij}}
\]

The above couple of equations are basically defining the allele counts \( a_{ijk} \) and frequency \( p_{ijk} \). Remember that \( i \) is individual, \( j \) is locus, and \( k \) is allele. The following continues to describe properties of the frequency table used for analysis:

\[
p_{ij} = \sum_{k=1}^{m(j)} p_{ijk} = 1
\]

The sum of all allele frequencies for a single population (or individual) at a single locus is one.

\[
p_{i..} = \sum_{j=1}^{\nu} p_{ij} = \nu
\]

The sum of all allele frequencies over all loci is equal to the number of loci.

\[
p_{...} = \sum_{j=1}^{\nu} p_{ij} = t\nu
\]

The the sum of the entire table is the sum of all loci multiplied by the number of populations (or individuals).

2 The Index of Association

The index of association was originally developed by A.H.D. Brown analyzing population structure of wheat and has been widely used as a tool to detect clonal reproduction within populations (Brown et al., 1980; Smith et al., 1993). 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.

The most likely mechanism for a change in genotype, for a clonal organism, is 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 (Smith et al., 1993). 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 (Smith et al., 1993)):

$$I_A = \frac{V_O}{V_E} - 1$$ (8)

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 is calculated using an 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 $j$ with $m(j)$ allelic states and a ploidy of $l$ is as follows$^2$:

$$d(a, b) = \frac{l}{2} \sum_{k=1}^{m(j)} |p_{ajk} - p_{bjk}|$$ (9)

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

$$D(a, b) = \sum_{i=1}^{\nu} d_i$$ (10)

An interesting observation: $D(a, b)/(l\nu)$ is Prevosti’s distance.

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

$$V_O = \frac{\sum_{i=1}^{\binom{n}{2}} D_i^2 - \left( \frac{\sum_{i=1}^{\binom{n}{2}} D_i}{\binom{n}{2}} \right)^2}{\binom{n}{2}}$$ (11)

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 $d$ (Agapow & Burt, 2001):

$$\text{var}_j = \frac{\sum_{i=1}^{\binom{n}{2}} d_i^2 - \left( \frac{\sum_{i=1}^{\binom{n}{2}} d_i}{\binom{n}{2}} \right)^2}{\binom{n}{2}}$$ (12)

The expected variance is then the sum of all the variances over all $\nu$ loci (Agapow & Burt, 2001):

$$V_E = \sum_{j=1}^{\nu} \text{var}_j$$ (13)

$^2$Individuals with Presence / Absence data will have the $l/2$ term dropped.
Now you can plug the sums of equations (11) and (13) into equation (8) 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, $\bar{r}_d$ (Agapow & Burt, 2001). For the derivation, see the manual for multilocus. The equation is as follows, utilizing equations (11), (12), and (13) (Agapow & Burt, 2001):

$$\bar{r}_d = \frac{V_O - V_E}{2 \sum_{j=1}^{\nu} \sum_{k \neq j}^{\nu} \sqrt{\text{var}_j \cdot \text{var}_k}}$$

(14)

3 Genetic distances

Genetic distances are great tools for analyzing diversity in populations as they are the basis for creating dendrograms with bootstrap support and also for AMOVA. This section will simply present different genetic distances along with a few notes about them. Most of these distances are derived from the ade4 and adegenet packages, where they were implemented as distances between populations. Poppr extends the implementation to individuals as well (with the exception of Bruvo’s distance).

<table>
<thead>
<tr>
<th>Method</th>
<th>Function</th>
<th>Assumption</th>
<th>Euclidean</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevosti</td>
<td>prevosti.dist</td>
<td>-</td>
<td>No</td>
<td>(Prevosti et al., 1975)</td>
</tr>
<tr>
<td></td>
<td>diss.dist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nei</td>
<td>nei.dist</td>
<td>Infinite Alleles</td>
<td>No</td>
<td>(Nei, 1972, 1978)</td>
</tr>
<tr>
<td>Edwards</td>
<td>edwards.dist</td>
<td>Genetic Drift</td>
<td>Yes</td>
<td>(Edwards, 1971)</td>
</tr>
<tr>
<td>Reynolds</td>
<td>reynolds.dist</td>
<td>Genetic Drift</td>
<td>Yes</td>
<td>(Reynolds et al., 1983)</td>
</tr>
<tr>
<td>Rogers</td>
<td>rogers.dist</td>
<td>-</td>
<td>Yes</td>
<td>(Rogers, 1972)</td>
</tr>
<tr>
<td>Bruvo</td>
<td>bruvo.dist</td>
<td>Stepwise Mutation</td>
<td>No</td>
<td>(Bruvo et al., 2004)</td>
</tr>
</tbody>
</table>

3.1 Distances that assume genetic drift

3.1.1 Nei’s 1978 Distance

$$D_{Nei}(a,b) = -\ln \left( \frac{\sum_{k=1}^{\nu} \sum_{j=1}^{m(k)} p_{ajk} p_{bjk}}{\sqrt{\sum_{k=1}^{\nu} \sum_{j=1}^{m(k)} (p_{ajk})^2 \sqrt{\sum_{k=1}^{\nu} \sum_{j=1}^{m(k)} (p_{bjk})^2}}} \right)$$

(15)

Note: if comparing individuals in poppr, those that do not share any alleles normally receive a distance of $\infty$. As you cannot draw a dendrogram with infinite branch lengths, infinite values are converted to a value that is equal to an order of magnitude greater than the largest finite value.

3.1.2 Edwards’ angular distance

$$D_{Edwards}(a,b) = \sqrt{1 - \frac{1}{\nu} \sum_{k=1}^{\nu} \sum_{j=1}^{m(k)} \sqrt{p_{ajk} p_{bjk}}}$$

(16)

3.1.3 Reynolds’ coancestry distance

$$D_{Reynolds}(a,b) = \sqrt{\frac{\sum_{k=1}^{\nu} \sum_{j=1}^{m(k)} (p_{ajk} - p_{bjk})^2}{2 \sum_{k=1}^{\nu} \left(1 - \sum_{j=1}^{m(k)} p_{ajk} p_{bjk}\right)}}$$

(17)
3.2 Distances without assumptions

3.2.1 Rogers’ distance

\[ D_{\text{Rogers}}(a, b) = \frac{1}{\nu} \sum_{k=1}^{\nu} \left( \frac{1}{2} \sum_{j=1}^{m(k)} (p_{a,j,k} - p_{b,j,k})^2 \right) \]  

(18)

3.2.2 Prevosti’s absolute genetic distance

\[ D_{\text{Prevosti}}(a, b) = \frac{1}{2\nu} \sum_{k=1}^{\nu} \sum_{j=1}^{m(k)} |p_{a,j,k} - p_{b,j,k}| \]  

(19)

Note: for AFLP data, the 2 is dropped.

3.3 Bruvo’s distance (stepwise mutation for microsatellites)

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 (Bruvo et al., 2004). The distance between each pair of alleles is calculated as \(^3\) (Bruvo et al., 2004):

\[ m_x = 2^{-|x|} \]  

(20)

\[ d_a = 1 - m_x \]  

(21)

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 \(|57 - 61| = 4\). 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 (Bruvo et al., 2004).

\[ s_i = \sum_{a=1}^{k} d_a \]  

(22)

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 (Bruvo et al., 2004):

\[ d_i = \frac{1}{k} \left( \min_{i...k!} s_i \right) \]  

(23)

Finally, after all of this, we can get the average distance over all loci (Bruvo et al., 2004).

\[ D = \frac{1}{l} \sum_{i=1}^{l} d_i \]  

(24)

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

\(^3\)Notation presented unmodified from Bruvo et al, 2004
### 3.3.1 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 (Bruvo et al., 2004).

To help deal with these situations, Bruvo has suggested three methods for dealing with these differences in ploidy levels (Bruvo et al., 2004):

- **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: $[69, 70, 69, 69]$, $[69, 70, 69, 70]$, $[69, 70, 70, 69]$, 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 $n^k$ where $n$ is the number of potential replacements and $k$ is the number of alleles to be replaced.

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

### 3.3.2 Choosing a model

By default, the implementation of Bruvo’s distance in `poppr` will utilize the combination model. This is implemented by setting both the `add` and `loss` arguments to `TRUE`. For other models use the following table for reference:

<table>
<thead>
<tr>
<th>Model</th>
<th>Arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infinite</td>
<td><code>add = FALSE, loss = FALSE</code></td>
</tr>
<tr>
<td>Genome Addition</td>
<td><code>add = TRUE, loss = FALSE</code></td>
</tr>
<tr>
<td>Genome Loss</td>
<td><code>add = FALSE, loss = TRUE</code></td>
</tr>
<tr>
<td>Combination (default)</td>
<td><code>add = TRUE, loss = TRUE</code></td>
</tr>
</tbody>
</table>

### 3.4 Tree topology

All of these distances were designed for analysis of populations. When applying them to individuals, we must change our interpretations. For example, with Nei’s distance, branch lengths increase linearly with mutation (Nei, 1972, 1978). When two populations share no alleles, then the distance becomes infinite. However, we expect two individuals to segregate for different alleles more often than entire populations, thus
we would expect exaggerated internal branch lengths separating clades. To demonstrate the effect of the different distances on tree topology, we will use 5 diploid samples at a single locus demonstrating a range of possibilities:

<table>
<thead>
<tr>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
</tr>
<tr>
<td>1/2</td>
</tr>
<tr>
<td>2/3</td>
</tr>
<tr>
<td>3/4</td>
</tr>
<tr>
<td>4/4</td>
</tr>
</tbody>
</table>

Table 2: Table of genotypes to be used for analysis

```r
library("poppr")
dat.df <- data.frame(Genotype = c("1/1", "1/2", "2/3", "3/4", "4/4"))
dat <- as.genclone(df2genind(dat.df, sep = "/", ind.names = dat.df[[1]]))

We will now compute the distances and construct neighbor-joining dendrograms using the package ape. This allows us to see the effect of the different distance measures on the tree topology.

distances <- c("Nei", "Rogers", "Edwards", "Reynolds", "Prevosti")
dists <- lapply(distances, function(x) {
  DISTFUN <- match.fun(paste(tolower(x), "dist", sep = "))
  DISTFUN(dat)
})
names(dists) <- distances

# Adding Bruvo's distance at the end because we need to specify repeat length.
dists$Bruvo <- bruvo.dist(dat, replen = 1)
library("ape")
par(mfrow = c(2, 3))
x <- lapply(names(dists), function(x){
  plot(nj(dists[[x]]), main = x, type = "unrooted")
  add.scale.bar(lcol = "red", length = 0.1)
})
```
4 AMOVA

AMOVA in *poppr* acts as a wrapper for the *ade4* implementation, which is an implementation of Excoffier’s original formulation (Excoffier *et al.*, 1992). As the calculation relies on a genetic distance matrix, *poppr* calculates the distance matrix as the number of differing sites between two genotypes using equation 10. This is equivalent to Prevosti’s distance multiplied by the ploidy and number of loci. This is also equivalent to Kronecker’s delta as presented in (Excoffier *et al.*, 1992). As is the practice from *Arlequin*, within sample variance is calculated by default. Within sample variance can only be calculated on diploid data.
5 Genotypic Diversity

Many of the calculations of genotypic diversity exist within the package vegan in the diversity function. Descriptions of most calculations can be found in the paper by Grünwald et al. (2003).

5.1 Rarefaction, Shannon-Wiener, Stoddart and Taylor indices

Stoddart and Taylor’s index is also known as Inverse Simpson’s index. A detailed description of these can be found in the “Diversity” vignette in vegan. You can access it by typing vignette("diversity-vegan")

5.2 Evenness ($E_5$)

Evenness ($E_5$) is essentially the ratio of the number of abundant genotypes to the number of rarer genotypes. For the function locus_table, these indices represent the abundance of alleles.

This is simply calculated as

$$E_5 = \frac{1}{\lambda} - 1$$

(25)

Where $1/\lambda$ is Stoddart and Taylor’s index and $H$ is Shannon diversity (Stoddart & Taylor, 1988; Shannon, 1948). The reason for this formulation is because $1/\lambda$ is weighted for more abundant genotypes and $H$ is weighted for rarer genotypes.

5.3 Hexp

$H_{exp}$ stands for Nei’s 1978 expected heterozygosity, also known as gene diversity (Nei, 1978). Previously, in poppr, it was calculated over multilocus genotypes, which was more analogous to a correction for Simpson’s index. This has been corrected in version 2.0 and is now correctly calculated over alleles. The following is adapted from equations 2 and 3 in (Nei, 1978):

$$h = \left( \frac{n}{n-1} \right) - \sum_{i=1}^{k} p_i^2$$

(26)

$$H_{exp} = \sum_{l=1}^{m} h_l/m$$

(27)

where $p_i$ is the population frequency of the $i^{th}$ allele in a locus, $k$ is the number of alleles in a particular locus, $n$ represents the number of observed alleles in the population, and $h_l$ is the value of $h$ at the $l^{th}$ locus.

Nei’s definition originally corrected the sum of squared allele frequencies with $2N/(2N - 1)$, where $N$ is the number of diploid samples (Nei, 1978). This definition is still congruent with that since we would observe $2N$ alleles in a diploid locus. The only caveat is that populations where allelic dosage is unknown will have an inflated estimate of allelic diversity due to the inflation of rare alleles.

6 Probability of Genotypes: $p_{gen}$

The metric $p_{gen}$ represents the probability that a given multilocus genotype will occur in a population (Parks & Werth, 1993; Arnaud-Haond et al., 2007). For $\nu$ loci, it is calculated as:

$$h = \begin{cases} 
1 & \text{if heterozygous,} \\
0 & \text{if homozygous or haploid}
\end{cases}$$

(28)
\[ p_{\text{gen}} = \prod_{j=1}^{\nu} \left( 2^h \prod_{i=1}^{m(j)} p_{ij} \right) \]  

(29)

where \( p_{ij} \) represents the allele frequency of the \( i^{th} \) allele present in the sample at the \( j^{th} \) locus derived using the round-robin method (Parks & Werth, 1993; Arnaud-Haond et al., 2007). The round-robin allele frequencies for a given locus are derived by first clone-correcting with the genotypes found at all other loci (Parks & Werth, 1993; Arnaud-Haond et al., 2007). Any allele that ends up with a zero frequency are assigned a value of \( 1/N \), where \( N \) is the number of samples.

7 Probability a genotype is derived from sexual reproduction: \( p_{\text{sex}} \)

The probability of a genotype can provide information as to how many times we expect to see that genotype repeated in our population. Parks & Werth (1993) defined the probability of encountering a genotype a second time from an independent reproductive event as:

\[ p_{\text{sex}} = 1 - (1 - p_{\text{gen}})^G \]  

(30)

Where \( G \) is the number of multilocus genotypes in the data. Finding a genotype \( n \) times out of \( N \) samples is derived from the binomial distribution (Parks & Werth, 1993; Arnaud-Haond et al., 2007):

\[ p_{\text{sex}} = \sum_{i=n}^{N} \binom{N}{i} (p_{\text{gen}})^i (1 - p_{\text{gen}})^{N-i} \]  

(31)

The `poppr` function `psex()` can calculate both of these quantities.

References


