Genetic Distance

Rodney J. Dyer
Department of Biology
Virginia Commonwealth University
http://dyerlab.bio.vcu.edu

Synopsis

The analysis of genetic data is largely an analysis of distances; distances among frequencies, distances among centroids of populations, etc.

Genetic Distances Among Individuals

In these examples, the data from *Araptus attenuatus* will be used again but this time we'll use the subset of individuals from "CladeB" (mainland populations).

```
> require(gstudio)
> data(araptus_attenuatus)
> sonora <- araptus_attenuatus[araptus_attenuatus$Species == "CladeB",
+     ]
> summary(sonora)
```

```
   Species    Cluster    Pop       Individual      Lat             Long
 CladeB:36   SON-B:36   101: 9   101_10A: 1   Min.   :26.38   Min.   :-110.6
                        102: 8   101_1A : 1   1st Qu.:26.64   1st Qu.:-109.6
                        32 :19   101_2A : 1   Median :26.64   Median :-109.3
                                 101_3A : 1   Mean   :26.90   Mean   :-109.6
                                 101_4A : 1   3rd Qu.:26.95   3rd Qu.:-109.3
                                 101_5A : 1   Max.   :27.91   Max.   :-109.1
                                 (Other):30
    LTRS         WNT          EN           EF          ZMP          AML          ATPS
 01:01: 1   01:01:29   01:01: 7   01:01:23   01:01: 1   08:08: 1   02:02:28
 01:02:17   01:03: 1   01:03: 2   01:02:11   02:02:19   08:11: 1   02:03: 1
 02:02:18   NA   : 6   03:03:19   NA   : 2   NA   :16   08:12: 1   02:04: 2
                       03:04: 6                         10:11: 1   02:09: 3
                       04:04: 1                         11:11:12   04:04: 1
                       NA   : 1                         12:12: 5   09:09: 1
                                                        NA   :15
     MP20
 12:12  : 6
 03:13  : 4
 11:12  : 3
 13:13  : 3
 NA     : 3
 02:10  : 2
 (Other):15
```

Jaccard Distance

Jaccard distance is a set-theoretic distance quantifying dissimilarity. Assuming that loci are sets of alleles, the Jaccard dissimilarity between genotypes A and B is given by:

$$J_\delta(A,B) = \frac{|A \bigcup B| - |A \bigcap B|}{|A \bigcup B|} \tag{1}$$

Using the `LTRS` locus, we compute this distance as:

```
> d.jaccard <- genetic.distance(sonora, stratum = "Pop", loci = "EN",
+     mode = "Jaccard")
> dim(d.jaccard$LTRS)

NULL
```

YOu can look at the elements of the LTRS matrix (it is 36x36 so I am not printing it out here). With `mode="Jaccard"`, missing genotypes will result in `NA` rows and columns in the distance matrix. It is no entirely clear how this metric can easily handle missing genotypes.

Bray-Curtis Distance

Bray-Curtis Distance (Bray & Curtis 1957) has been primarily used to quantify differences in species composition. It is defined as the total number of species that are unique to either of the two sites standardized by the number of species in both sites.

$$BC_\delta = \frac{S_i + S_j - 2S_{ij}}{S_i + S_j} \tag{2}$$

where S_x is the species count and S_{ij} is the sum of minimum abundances. Lately, this has seen considerable use within individual-based landscape genetic studies. Missing genotypes are set to average allele frequencies, that is to say that every missing genotype is considered to have all the alleles present in the entire population, but with probability equal to their global frequencies. Essentially, this removes the `NA` problem like in the `mode="Jaccard"` situation and does so by taking the non-missing genotype's genetic distance from the global genetic centroid (it's cosmic man!). Here is the estimation using two loci.

```
> d.bray <- genetic.distance(sonora, stratum = "Pop", loci = c("LTRS",
+     "EN"), mode = "Bray")
> summary(d.bray)

     Length Class  Mode   
LTRS 1296   -none- numeric
EN   1296   -none- numeric
```

AMOVA Distance

The final individual-based approach is based upon the Analysis of Molecular Variance (AMOVA) analysis. A geometric interpretation of this genetic distance is given in Figure 1 indicating distances among diploid genotypes.

Algebraically, we can define an individual locus using a multivariate vector as an allele coding vector. The Locus class has a method, as.multivariate, that does the translation. The distance between the two alleles is defined as:

$$\delta_{ij}^2 = 2(p_i - p_j)^2 \tag{3}$$

as shown below.

The amova distance is simply the vector distance between these two vectors as demonstrated below

```
> locAA <- Locus(c("A", "A"))
> locBB <- Locus(c("B", "B"))
> locAB <- Locus(c("A", "B"))
> locBC <- Locus(c("B", "C"))
> vAA <- as.vector(locAA, c("A", "B", "C"))
> vBB <- as.vector(locBB, c("A", "B", "C"))
> vAB <- as.vector(locAB, c("A", "B", "C"))
> vBC <- as.vector(locBC, c("A", "B", "C"))
> dist.AA.BB <- 2 * ((vAA - vBB) %*% (vAA - vBB))
> dist.AA.BB
```

```
     [,1]
[1,]   16
```

```
> dist.AA.AB <- 2 * ((vAA - vAB) %*% (vAA - vAB))
> dist.AA.AB
```

```
     [,1]
[1,]    4
```

```
> dist.AA.BC <- 2 * ((vAA - vBC) %*% (vAA - vBC))
> dist.AA.BC
```

```
     [,1]
[1,]   12
```

While we will deal more with the AMOVA analysis in the section on Genetic Structure, the AMOVA genetic distance matrix can be estimated as follows, this time using *all* the loci. This metric is additive across loci, so only a single distance matrix is returned. The list key for the multilocus parameters is a list of the locus names, joined using a period.

```
> d.amova <- genetic.distance(sonora, stratum = "Pop", mode = "AMOVA",
+     loci = "EN")
> summary(d.amova)
```

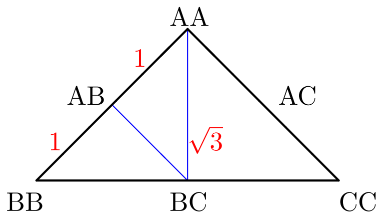

Figure 1: Geometry of AMOVA distances. The resulting squared distance is the square of the geometric distance.

```
   Length Class  Mode
EN 1296   -none- numeric
```

There are several other measures of individual-to-individual distance such as relatedness and coancestry. These are not currently implemented in `R` but may become available in the near future. That being said, it is probably something not too difficult for someone to extend these functions with their own code.

Differences Between Distances

These three distances are correlated, and here we can look at how close they are for this three allele locus in *Euphorbia lomelii*. They will be transformed from a `dist` matrix object into columns within a `data.frame` and then their relationship can be tested using `cor.test`.

```
> df <- data.frame(jaccard = d.jaccard$EN[lower.tri(d.jaccard$EN)],
+     bray = d.bray$EN[lower.tri(d.bray$EN)], amova = d.amova$EN[lower.tri(d.amova$EN)])
> summary(df)
```

```
    jaccard           bray            amova      
 Min.   :0.000   Min.   :0.0000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000  
 Median :0.500   Median :0.5000   Median :1.000  
 Mean   :0.527   Mean   :0.5238   Mean   :1.568  
 3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:4.000  
 Max.   :1.000   Max.   :1.0000   Max.   :4.000  
```

```
> cor(df)
```

```
          jaccard      bray     amova
jaccard 1.0000000 0.9985311 0.8883334
bray    0.9985311 1.0000000 0.8919370
amova   0.8883334 0.8919370 1.0000000
```

```
> pairs(df)
```

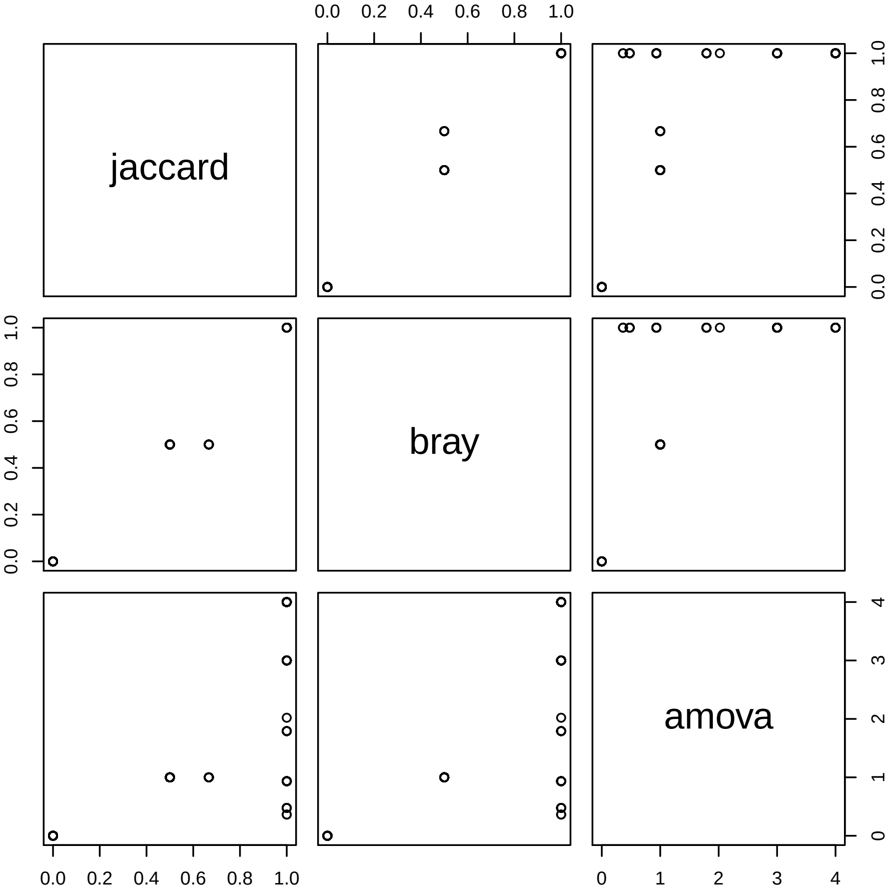

Figure 2: Relationship among three individual genetic distance metrics estimated for individual *Araptus attenuatus* individuals in Sonora & Sinoloa, Mexico.

Genetic Distance Among Strata

Genetic distances can also be estimated among groups of individuals. The same data will be used here but since there are only three populations, we'll be able to see the whole distance matrix.

Euclidean Distance

Euclidean distance is the most straight-forward distance metric available as it is essentially straight-line distance based upon the allele frequencies in each population. It is given by:

$$d_{eucl} = \sqrt{\sum_{j=1}^{L} (p_{ij} - p_{kj})^2}$$

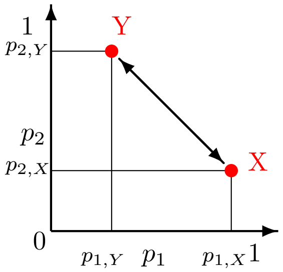

Figure 3: Geometry of euclidean distance based upon a two-allele locus denoted as frequencies p_1 & p_2.

where p_{ij} and p_{kj} are the frequencies of the j^{th} allele in both the i^{th} and j^{th} population. In this and the following distance examples, I am going to take the resulting distance matrix among all pairs of populations and put them into a Neighbor joining tree (via the `nj` function from the `ape` package) as it may be easier to see differences in topologies rather than matrices.

It is perhaps easiest to think of Euclidean distance in x,y coordinate space (Figure 3). This distance can be estimated by `stratum.distance` using the optional parameter `method='eucl'` and it will return a `dist` matrix.

Once the matrix has been estimated, you can visualize it in many ways. One of the most straight-forward approaches it to visualizing the relationships among rows and columns is to put it into a bifurcating tree.

```
> d.eucl <- genetic.distance(sonora, stratum = "Pop", loci = "EN",
+     mode = "Euclidean")
> d.eucl
```

```
$EN
          [,1]      [,2]      [,3]
[1,] 0.0000000 0.5611959 0.6908633
[2,] 0.5611959 0.0000000 0.2698923
[3,] 0.6908633 0.2698923 0.0000000
```

Cavalli-Sforza Distance

Another distance approach that is commonly used for microsatellite loci is Cavalli-Sforza distance, D_C (Cavalli-Sforza and Edwards, 1967). Here population allele frequencies are plot on the surface of a sphere (radius=1) using the square root of the allele frequencies.

$$D_C = \frac{2}{\pi}\sqrt{(2-2cos\theta)}$$

The genetic distance, D_C is measured as the chord distance as indicated in Figure **??**. The resulting Neighbor joining tree from this distance is shown in Figure 4

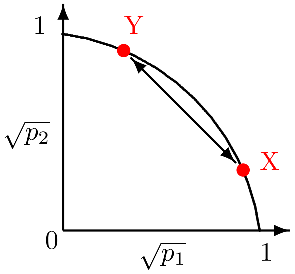

Figure 4: Geometry of Cavalli-Sforza distance. Population allele frequencies at two loci are plot at $\sqrt{p_1}$ and $\sqrt{p_2}$ and D_C is the chord between the populations.

```
> d.cavalli <- genetic.distance(sonora, "Pop", "EN", "Cavalli")
> d.cavalli

$EN
          [,1]      [,2]      [,3]
[1,] 0.0000000 0.4131725 0.7554523
[2,] 0.4131725 0.0000000 0.5155875
[3,] 0.7554523 0.5155875 0.0000000
```

Nei's Genetic Distance

Nei's genetic distance is based upon mutation drift equilibrium therefore you should be reasonably comfortable with the notion that your populations have been separated a sufficient period of time such that drift and mutation may have played a significant role in their structure.

The formula for Nei's distance that is used here is:

$$D_{Nei} = -ln\left(\frac{(2N-1)\sum_{i=1}^{L}\sum_{j=1}^{\ell} p_{ij,x}p_{ij,y}}{\sqrt{\sum_{i=1}^{L}(2N\sum_{j=1}^{\ell} p_{ij,x}-1)(2N\sum_{j=1}^{\ell} p_{ij,y}-1)}}\right)$$

where the summation L is across loci and ℓ is across alleles at each locus in population x and y.

```
> d.nei <- genetic.distance(sonora, "Pop", "EN", "Nei")
> d.nei

$EN
          [,1]     [,2]      [,3]
[1,] 0.0000000 1.200027 0.5848357
[2,] 1.2000270 0.000000 2.8444285
[3,] 0.5848357 2.844428       Inf
```

Conditional Genetic Distance

Conditional genetic distance (*cGD*, Dyer et al. 2010) is a graph-theoretic genetic distance derived from Population Graphs (Dyer and Nason 2004)

Isolation-By-Distance

Under models with restrictions in gene flow, there is an expectation that genetic distance should increase with physical separation. Using populations found along the Baja Peninsula, it is pretty easy to see which one of these among-strata distance approaches provides a better fit to the data.

```
> baja <- araptus_attenuatus[araptus_attenuatus$Species != "CladeB",
+     ]
> euc <- genetic.distance(baja, "Pop", "EN", "Euclidean")$EN
> cav <- genetic.distance(baja, "Pop", "EN", "Cavalli")$EN
> nei <- genetic.distance(baja, "Pop", "EN", "Nei")$EN
> phys <- stratum.distance(baja, "Pop", lat = "Lat", lon = "Long")
> df <- data.frame(Euclidean = euc[lower.tri(euc)], Cavalli = cav[lower.tri(cav)],
+     Nei = nei[lower.tri(nei)], Physical.Dist = phys[lower.tri(phys)])
> pairs(df)
> cor(df)
```

```
              Euclidean   Cavalli Nei Physical.Dist
Euclidean     1.0000000 0.9414622 NaN     0.2935694
Cavalli       0.9414622 1.0000000 NaN     0.2704443
Nei                 NaN       NaN   1           NaN
Physical.Dist 0.2935694 0.2704443 NaN     1.0000000
```

Bibliography

Bray, JR and JT Curtis (1957) An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* **27**:325-349.

Cavalli-Sforza LL and AWF Edwards 1967. Phylogenetic analysis: models and estimation procedures. American Journal Human Genetics **19**: 233-257.

Dyer RJ and JD Nason 2004 Population Graphs: The graph-theoretic shape of genetic structure. Molecular Ecology **13**: 1713-1728.

Dyer RJ, JD Nason, and RC Garrick 2010 Landscape modeling of gene flow: Improved power using conditional genetic distance derived from the topology of population networks. Molecular Ecology **19**: 3746-3759.

Smouse, PE and R Peakall (1999) Genetics

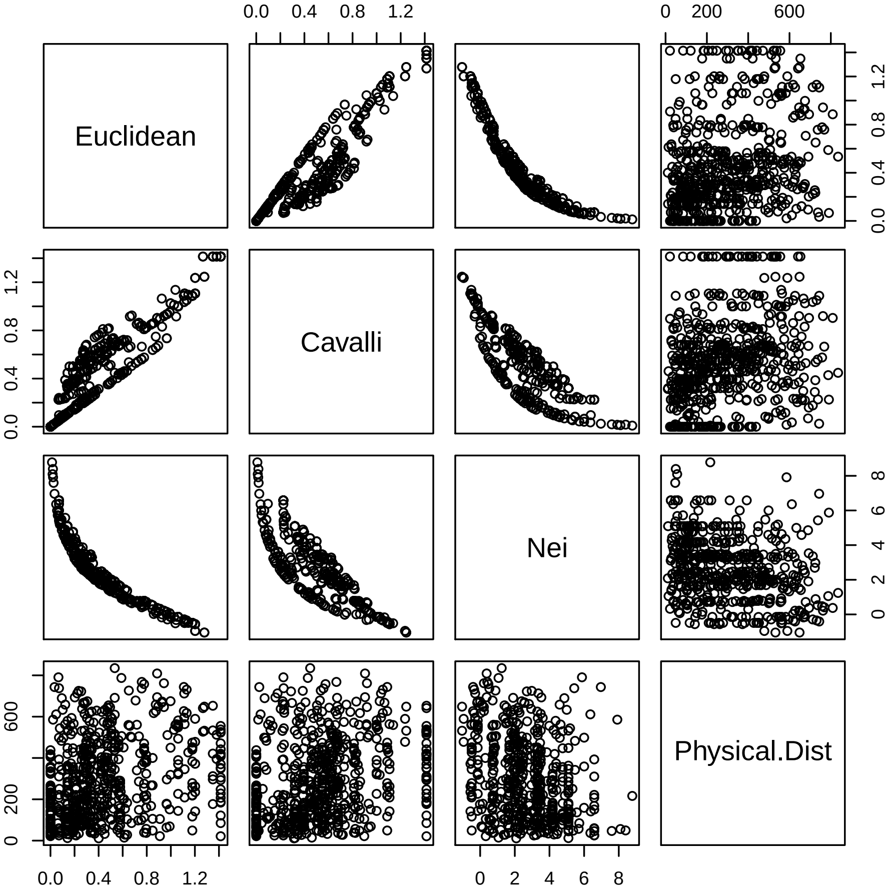

Figure 5: Relationship among strata genetic distance metrics estimated for *Araptus attenuatus* sites in Baja California along with physical distance.