Analysis of Molecular Variance

A.R. Rao
Indian Agricultural Statistical Research Institute, New Delhi - 12

1 Introduction

Analysis of Molecular Variance (AMOVA) is a method for studying molecular variation within a species. The analysis of molecular variance (AMOVA) was used to study the patterns and degree of relatedness revealed by Multidimensional scaling and the Clustering dendrogram. Further it is used to summarize the population structure with the marker data from different genotypes, while remaining flexible enough to accommodate different types of assumptions about the evolution of the genetic system. Electrophoresis, one of the most widely used methods for studying the structure of DNA, produces marker data in the form of 0s and 1s where 1 denotes the presence of a band and zero its absence. The vector of such 0’s and 1’s is called DNA haplotype of the individual/variety. Recently, Analysis of Molecular Variance (AMOVA) is used to calculate the ‘between groups’ and ‘within groups’ variance. This technique treats genetic distances as deviations from a group mean position, and uses the squared deviations as variances. The total sums of squares of genetic distances can then be partitioned into components that represent the ‘within group’ and the ‘between-group’ sum of squares. The resulting test statistic $\Phi_{ST}$ is analogous to Wright’s $F_{ST}$, and is the ratio of the between-group mean square to the total mean square (Wright, 1951; Cockerham, 1973). $\Phi_{ST}$ represents the correlation between random genetic accessions within a group relative to random accessions from the population at large. This statistic can take values between 0 and 1; higher values indicate greater partitioning of the population into sub-groups.

2 The Analysis of molecular variance procedure

When a population is divided into isolated subpopulations, there is less heterozygosity than there would be if the population was undivided. Founder effects acting on different demes generally lead to subpopulations with allele frequencies that are different from the larger population. Also, these demes are smaller in size than the larger population; since allele frequency in each generation represents a sample of the previous generation's allele frequency, there will be greater sampling error in these small groups than there would be in a larger undifferentiated population. Hence, genetic drift will push these smaller demes toward different allele frequencies and allele fixation more quickly than would take place in a larger undifferentiated population. For a given species, when several subpopulations are separated geographically, in absence of selection and with random mating, two trends are expected: (i) Gene frequencies for the total population remain constant over generations, and (ii) The variance of gene frequencies increase over time because of differentiation among subpopulations. Wright’s F statistic (Wright, 1965), quantify the differentiation among subpopulations and among individuals. However, molecular data reveals not only the frequency of molecular markers, but can also tell us something about the amount of mutational differences between different genes. Analysis of Molecular Variance (AMOVA) is a method of estimating population differentiation directly from
molecular data and testing hypotheses about such differentiation. AMOVA may be used to analyze STMS or AFLP molecular data.

AMOVA treats any kind of raw molecular data as a Boolean vector \( p_i \), that is, a \( 1 \times n \) matrix of 1’s and 0’s, 1 indicating the presence of a ‘i’ marker and 0 its absence. A marker could be a nucleotide base, a base sequence, a restriction fragment, or a mutational event. Euclidean distances between pairs of vectors are then calculated by subtracting the Boolean vector of one haplotype from another, according to the formula 

\[
\delta_{jk} = (p_j - p_k)^T W (p_j - p_k)
\]

where \( W \) is a weighting matrix; by default, it is an identity matrix and does not change the value of the final product; however, \( W \) can be a matrix with a number of values depending upon how one weights molecular change at different locations on a sequence or phylogenetic tree.

### 2.1 Partitioning a distance matrix into hierarchical components

Consider a haploid genetic system where inter-haplotype distances are identical to distances between individuals. One can arrange a set of \( N \) individuals from \( I \) populations into a distance matrix, \( D^2 \), partitioned into a series of submatrices corresponding to particular subdivisions as below:

\[
D^2 = \begin{bmatrix}
D_{11}^2 & D_{12}^2 & \cdots & D_{1I}^2 \\
D_{21}^2 & D_{22}^2 & \cdots & D_{2I}^2 \\
\vdots & \vdots & \ddots & \vdots \\
D_{II}^2 & \cdots & \cdots & D_{II}^2
\end{bmatrix}
\]

where the elements of the block-diagonal submatrices \( D_{ii}^2 \) contain pairwise squared-distances \( \delta_{jk}^2 \) between individuals of the same \( (i) \)th population, and those of the off-diagonal matrix blocks \( D_{ii}^2 \), contain pairwise squared-distances between individuals, one from the \( i \)th and other from the \( i \)'th population. Individuals may also be grouped at higher levels, according to such non-genetic criteria as geography, ecological environment, or language.

A conventional sum of squares [SS(Total)] may be written, barring a constant \((2N)\), as the sum of squared differences between all pairs of \( N \) items. In the multidimensional case, using vectors instead of scalars, the conventional sum of squares becomes a sum of squared deviations (SSD) from the centroid of a multidimensional space. Thus,

\[
SSD_{(Total)} = \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{N} (p_j - p_k)^T W (p_j - p_k)
\]
= \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{N} \delta_{jk}^{2}

because \( \delta_{jj}^{2} = 0 \) for all haplotype \( h_j \). This transformation applies equally to the total array of individuals in the data set, to those within each population separately (within the diagonal blocks, \( D_{ii}^{2} \)), and to those belonging to a particular subdivision (within the diagonal blocks, \( D_{11}^{2}, D_{12}^{2}, D_{21}^{2} \) and \( D_{22}^{2} \)).

### 2.2 Model for AMOVA

Where individuals are arranged into populations and populations nested within groups defined \textit{a priori} on nongenetic criteria, a linear model can be defined on the pattern first described by Cockerham (1969, 1973) and refined upon by others (Weir and Cockerham 1984; Long 1986)

\[ p_{jig} = p + a_g + b_{ig} + c_{jig} \]  

(1)  

where \( p_{jig} \) indexes the \( j \)th chromosome, here equivalent to the \( j \)th individual (\( j = 1, \ldots, N_{ig} \)) in the \( i \)th population (\( i = 1, \ldots, I_g \)) in the \( g \)th group (\( g = 1, \ldots, G \)) and \( p \) is the unknown expectation of \( p_{jig} \) averaged over the whole study. The effects are \( a \) for group, \( b \) for populations and \( c \) for individuals within populations. The effects will be assumed to be additive, random, uncorrelated, and to have the associated variance components \( \sigma_a^2 \), \( \sigma_b^2 \) and \( \sigma_c^2 \) respectively.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>MSD</th>
<th>Expected MSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among regions</td>
<td>( G-1 )</td>
<td>MSD/(AG)</td>
<td>( \sigma_c^2 + n'\sigma_b^2 + n''\sigma_a^2 )</td>
</tr>
<tr>
<td>Among populations within regions</td>
<td>( \sum_{g=1}^{G} I_g - G )</td>
<td>MSD/(AP/WG)</td>
<td>( \sigma_c^2 + n\sigma_b^2 )</td>
</tr>
<tr>
<td>Among individuals within populations</td>
<td>( N - \sum_{g=1}^{G} I_g )</td>
<td>MSD(WP)</td>
<td>( \sigma_c^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( N-1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relying on the standard decomposition, one can write note that for any choice of hierarchical partition of the \( N \) individuals into strata,

\[ \text{SSD(Total)} = \text{SSD(Among Strata)} + \text{SSD(Within Strata)}, \]

placing in traditional analysis of variance framework, designated here as Analysis of Molecular variance, AMOVA (Table 1). The total sum of squared deviations, SSD(Total), can be partitioned into components for variation within populations, SSD(WP), variation among populations within regional groups, SSD(AP/WG), and variation among regional groups, SSD(AG). The corresponding sums of squares are
SSD(WP) = \sum_{g=1}^{G} \sum_{i=1}^{I_g} \sum_{k=1}^{N_{g}} \sum_{j=1}^{I_g} \frac{\delta_{jk}^2}{2N_{ig}}

SSD(AP/WG) = \sum_{g=1}^{G} \left( \sum_{i=1}^{I_g} \sum_{j=1}^{I_g} \sum_{k=1}^{N_{g}} \sum_{l=1}^{I_g} \frac{\delta_{jk}^2}{2N_{ig}} \right) - \sum_{i=1}^{I_g} \sum_{j=1}^{I_g} \sum_{k=1}^{N_{g}} \sum_{l=1}^{I_g} \frac{\delta_{jk}^2}{2N_{ig}}

and

SSD(AG) = \left( \sum_{j=1}^{I_g} \sum_{k=1}^{N_{g}} \delta_{jk}^2 \right) \frac{1}{2N_{ig}} - \sum_{g=1}^{G} \left( \sum_{g=1}^{G} \sum_{i=1}^{I_g} \sum_{j=1}^{I_g} \sum_{k=1}^{N_{g}} \sum_{l=1}^{I_g} \frac{\delta_{jk}^2}{2N_{ig}} \right)

The mean squared deviations (MSD) are then obtained by dividing such SSD by the appropriate degrees of freedom as reported in Table 1. The \( n \) coefficients in Table 1 represent the average sample sizes of particular hierarchical levels, allowing for unequal sample sizes,

\[
n = \frac{\sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig} - \left( \sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig}^2 \right)}{\sum_{g=1}^{G} I_g}
\]

\[
n' = \frac{\sum_{g=1}^{G} \left( \sum_{i=1}^{I_g} N_{ig}^2 \right) - \left( \sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig} \right)^2}{(G - 1) \sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig}}
\]

\[
n'' = \frac{\sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig} - \left( \sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig} \right)^2}{(G - 1) \sum_{g=1}^{G} \sum_{i=1}^{I_g} N_{ig}}
\]

The variance components (\( \sigma^2 \)'s) of each hierarchical level are extracted by equating the mean squares (MSDs) to their expectations. It may also be useful to employ haplotypic correlation measures, which are termed as \( \Phi \)-statistics. The different variance components can be expressed in terms of \( \Phi \)-statistics as

\[
\sigma_c^2 = (1 - \Phi_{sc}) \sigma^2
\]
\( \sigma_b^2 = (\Phi_{ST} - \Phi_{CT})\sigma^2 \)
\( \sigma_a^2 = \Phi_{CT}\sigma^2 \)

where \( \sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 \); \( \Phi_{ST} \) is viewed as the correlation of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the whole species; \( \Phi_{CT} \) as the correlation of random haplotypes within a group of populations, relative to that of random pairs of haplotypes drawn from the whole species, and \( \Phi_{SC} \) as the correlation of the molecular diversity of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the region. One can rewrite the above equations in terms of the \( \Phi \)-statistics as

\( \Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2} \), \( \Phi_{CT} = \frac{\sigma_a^2}{\sigma^2} \), \( \Phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2} \)

### 2.3 Testing significance of the variance components and \( \Phi \)-statistics

For testing variance components, the traditional analysis of molecular variance procedure cannot be adopted because the molecular data consist of Euclidean distances derived from vectors of 1’s and 0’s, and the data are unlikely to follow a normal distribution. A null distribution is therefore computed by resampling of the data or by a permutation procedure. Excoffier, et al., (1992) has discussed the methods for testing the significance of the variance components obtained from analysis of molecular variance. Under this procedure each individual is allocated to a randomly chosen population, while holding sample sizes constant at the realized values so as to obtain null distribution. This amounts to random permutation of the rows (and corresponding columns) of the squared distance matrix. The variance-components are estimated from each of a large number (say 500) of permuted matrices.

Further, they suggested two other permutation schemes that are useful for testing \( \Phi_{SC} \), \( \sigma_b^2 \) and \( \Phi_{CT} \), \( \sigma_c^2 \). The first assumes that the regions are real but that the populations within them are not, permuting individuals within regional groups without regard to population. The second assumes that while the populations are real, the regional groupings are artificial, permuting whole populations across groups. In this case, the sizes of the groups (but not those of the populations) vary with each permutational run.

### 2.4 Restriction site sampling

The sampling of nucleotides shows a major source of variability for the estimation of molecular diversity (Lynch and Crease 1990). One can ask whether the results depend on a particular array of marker sites employed. Excoffier, et al., (1992) examined the influence of site sampling on the genetic structure of the populations, using a site resampling plain similar to the bootstrap used by Efron (1982). Under the assumption the observed n sites are representative of all molecular markers. They obtain the distribution of the variance components and associated \( \Phi \)-statistics by Monte Carlo simulation, using 500 random collection sites. For each collection, the procedure is as follows: (a) Draw a given number of sites from the observed array of m sites, at random and with replacement. Given the choice of sites, the haplotype of each individual is then taken as
the combination of the original states of those randomly chosen sites; (b) compute interhaplotypic distances on the basis of the newly defined haplotypes and perform an AMOVA analysis. The distances are simply computed from euclidean distance; and (c) permute the matrix 500 times, and test the significance of the different statistics with the previously described procedures.

References