Analysis of molecular variance (AMOVA)

Introduction

We’ve already encountered $\pi$, the nucleotide diversity in a population, namely

$$\pi = \sum_{ij} x_i x_j \delta_{ij},$$

where $x_i$ is the frequency of the $i$th haplotype and $\delta_{ij}$ is the fraction of nucleotides at which haplotypes $i$ and $j$ differ. It shouldn’t come to any surprise to you that just as there is interest in partitioning diversity within and among populations when we’re dealing with simple allelic variation, i.e., Wright’s $F$-statistics, there is interest in partitioning diversity within and among populations when we’re dealing with nucleotide sequence or other molecular data. Let’s stick with nucleotide sequence data for the moment.

Analysis of molecular variation (AMOVA)

The notation now becomes just a little bit more complicated. We will now use $x_{ik}$ to refer to the frequency of the $i$th haplotype in the $k$th population. Then

$$x_i = \frac{1}{K} \sum_{k=1}^{K} x_{ik}$$

is the mean frequency of haplotype $i$ across all populations, where $K$ is the number of populations. We can now define

$$\pi_t = \sum_{ij} x_i x_j \delta_{ij}$$

and

$$\pi_s = \frac{1}{K} \sum_{k=1}^{K} \sum_{ij} x_{ik} x_{jk} \delta_{ij},$$

where $\pi_t$ is the nucleotide sequence diversity across the entire set of populations and $\pi_0$ is the average nucleotide sequence diversity within populations. Then we can define

$$\Phi_{st} = \frac{\pi_t - \pi_s}{\pi_t},$$

(1)

© 2001-2006 Kent E. Holsinger
which is the direct analog of Wright’s $F_{st}$ for nucleotide sequence diversity. Why? Well, that requires you to remember stuff we covered eight or ten weeks ago.

To be a bit more specific, refer back to http://darwin.eeb.uconn.edu/eeb348/lecture-notes/wahlund/node4.html. If you do, you’ll see that we defined

$$F_{it} = 1 - \frac{H_i}{H_t},$$

where $H_i$ is the average heterozygosity in individuals and $H_t$ is the expected panmictic heterozygosity. Defining $H_s$ as the average panmictic heterozygosity within populations, we then observed that

$$1 - F_{it} = \frac{H_i}{H_t} = \frac{H_i H_s}{H_s H_t} = (1 - F_{is})(1 - F_{st}).$$

In short, another way to think about $F_{st}$ is

$$F_{st} = \frac{H_t - H_s}{H_t}. \quad (2)$$

Now if you compare equation (1) and equation (2), you’ll see the analogy.

Excoffier et al. [1] pointed out that other types of molecular data can easily be fit into this framework. We simply need an appropriate measure of the “distance” between different haplotypes or alleles. Even with nucleotide sequences the appropriate $\delta_{ij}$ may reflect something about the mutational pathway likely to connect sequences rather than the raw number of differences between them. The idea is illustrated in Figure 1. This procedure for partitioning diversity in molecular markers is referred to as an analysis of molecular variance or AMOVA (by analogy with the ubiquitous statistical procedure analysis of variance, ANOVA). Like Wright’s $F$-statistics, the analysis can include several levels in the hierarchy.

**An AMOVA example**

Excoffier et al. [1] illustrate the approach by presenting an analysis of restriction haplotypes in human mtDNA. They analyze a sample of 672 mitochondrial genomes representing two populations in each of five regional groups (Figure 2). They identified 56 haplotypes in that sample. A minimum spanning tree illustrating the relationships and the relative frequency of each haplotype is presented in Figure 3.
Figure 1: Converting raw differences in sequence (or presence and absence of restriction sites) into a minimum spanning tree and a mutational measure of distance for an analysis of molecular variance (from [1]).

Figure 2: Locations of human mtDNA samples used in the example analysis (from [1]).
It’s apparent from the figure that haplotype 1 is very common. In fact, it is present in substantial frequency in every sampled population. An AMOVA using the minimum spanning network in Figure 3 to measure distance produces the results shown in Table 1. Notice that there is relatively little differentiation among populations within the same geographical region ($\Phi_{SC} = 0.044$). There is, however, substantial differentiation among regions ($\Phi_{CT} = 0.220$). In fact, differences among populations in different regions is responsible for nearly all of the differences among populations ($\Phi_{ST} = 0.246$). Notice also that $\Phi$-statistics follow the same rules as Wright’s $F$-statistics, namely

$$1 - \Phi_{ST} = (1 - \Phi_{SC})(1 - \Phi_{CT})$$
$$0.754 = (0.956)(0.78),$$

within the bounds of rounding error.$^1$

$^1$There wouldn’t be any rounding error if we had access to the raw data.
<table>
<thead>
<tr>
<th>Component of differentiation</th>
<th>Φ-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among regions</td>
<td>Φ_{CT} = 0.220</td>
</tr>
<tr>
<td>Among populations within regions</td>
<td>Φ_{SC} = 0.044</td>
</tr>
<tr>
<td>Among all populations</td>
<td>Φ_{ST} = 0.246</td>
</tr>
</tbody>
</table>

Table 1: AMOVA results for the human mtDNA sample (from [1]).

An extension

As you may recall, Slatkin [3] pointed out that there is a relationship between coalescence time and $F_{st}$. Namely, if mutation is rare then

$$F_{ST} = \frac{\bar{t} - \bar{t}_0}{\bar{t}}$$

where $\bar{t}$ is the average time to coalescence for two genes drawn at random without respect to population and $\bar{t}_0$ is the average time to coalescence for two genes drawn at random from the same populations. Results in [2] show that when $\delta_{ij}$ is linearly proportional to the time since two sequences have diverged, $\Phi_{ST}$ is a good estimator of $F_{ST}$ when $F_{ST}$ is thought of as a measure of the relative excess of coalescence time resulting from dividing a species into several population. This observation suggests that the combination of haplotype frequency differences and evolutionary distances among haplotypes may provide insight into the evolutionary relationships among populations of the same species.

References


Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/2.5/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.