HIERFSTAT, a package for R to compute and test hierarchical F-statistics

Jérôme Goudet
Dept. of Ecology and Evolution
BB, UNIL, CH-1015 Lausanne, Switzerland
jerome.goudet@unil.ch

16th August 2004

Abstract

The package HIERFSTAT for the statistical software R (R Development Core Team [2003]) allows to estimate hierarchical F-statistics from a hierarchy with any numbers of levels. In addition, it allows testing the statistical significance of population differentiation for these different levels, using a generalized likelihood-ratio test. The package HIERFSTAT is available from `http://www.unil.ch/popgen/softwares/hierfstat.htm`.

Population biologists have a long standing interest in estimating population structure with the help of genetic markers. To this end, F-statistics (Wright [1951], Nei [1987],Weir [1996]) are very commonly used and many computer programs are available to estimate these quantities (see for instance an extensive list at `http://www.bio.ulaval.ca/louisbernatchez/links.htm#soft_pop_struct`). F-statistics are usually defined for a two level hierarchy, where individuals (usually diploids) are located in sub-populations. From such a hierarchy, the statistic F_{IS}, where I stands for individual and S for sub-population, estimates the departure from panmixia at the level of the sub-populations, while F_{ST} (where T stands for total) quantifies the differences in allele frequencies among populations.

This two level hierarchy has often been found to be too restrictive compared to the biology of the species under investigation. Patches are often nested in localities, themselves potentially nested in areas again nested in regions. For instance, Trouve et al. [2004] in an investigation of the genetic structure of the freshwater snail *Galba truncatula* identified patches as a first level of population structure, and patches were nested in localities. Many other empirical studies found it useful to distinguish between different geographical levels.

Weir [1990] gave formulae to obtain variance components from a three and four level hierarchy from which hierarchical F-statistics can be computed, and some

computer programs exist to analyse genetic data for such hierarchy (e.g. GDA (Lewis and Zaykin [2001]), ARLEQUIN (Schneider et al. [2000])). More recently, Yang [1998] developed an algorithm allowing to compute variance components from a fully random hierarchical design with any number of levels in the hierarchy.

The goal of this note is to announce HIERFSTAT, a package for the statistical software R (R Development Core Team [2003]) that implements Yang's algorithm and computes moment estimators of hierarchical F-statistics from any number of hierarchical level. In addition, this package allows testing the significance of the different levels of differentiation using randomization tests.

Variance component estimation

The algorithm and rational for estimating variance components are given in Yang [1998], and need not be further developed here. To obtain hierarchical F-statistics, we defined the following notations. σ_i^2 is the variance component for level i, $\sigma^2_{\sum i} = \sum_{k=1}^{i} \sigma_k^2$ is the sum of the variance components from the lowest hierarchical level to level i and $\sigma^2_{i(j)} = \sum_{k=(j+1)}^{i} \sigma_k^2$. With these definitions, the hierarchical F between level j and i is defined as $F_{ji} = \frac{\sigma^2 i(j)}{\sigma^2_{\sum i}}$.

Two functions in the package FSTATHIER carry out the estimation of variance components. The first is `varcomp`, which given the matrix of levels and the genotypic data at one locus, outputs:

- `df` the degrees of freedom from each level of the analysis.
- `k` the matrix of k coefficients necessary to extract each variance components.
- `res` the variance components for each allele at the locus.
- `overall` the variance components summed over alleles.
- `F` a matrix of hierarchical F-statistics type-coefficients with the first line corresponding to $F_{(n-1)/n}, F_{(n-2)/n}, \ldots, F_{1/n}$ and the diagonal corresponding to $F_{(n-1)/n}, F_{(n-2)/(n-1)}, \ldots, F_{1/2}$.

The second is `varcomp.glob`. This later function is simply a wrapper of the previous one as it produces estimates of variance components for each locus passed to the function (`loc`), an overall estimate (`overall`), and a matrix of hierarchical F-statistics estimated from all the loci as the ratio of sums of variance components in the numerator and denominator (see e.g. Excoffier [2001]).

Testing population differentiation

While the issue of testing population differentiation at one diploid locus has been addressed (Goudet et al. [1996]), little care has been given to differentiation

tests for data sets made of multiple loci (Petit et al. [2001]), or on appropriate permutation schemes for tests of the effect of different hierarchical levels (Excoffier [2001]).

Contrary to intuition, the best statistic to test for differentiation is not F_{ST} or its components, but Goudet et al. [1996] showed that the likelihood ratio G-statistic, defined as

$$G = 2\sum_{i=1}^{np}\sum_{j=1}^{na} O_{i,j}\log\frac{O_{i,j}}{E_{i,j}}$$

(where $O_{i,j}$ is the observed number of alleles in cell $[i,j]$, $E_{i,j}$ the expected number of alleles in the same cell, np is the number of populations and na is the number of alleles at the locus) is a powerful statistic to detect population differentiation. These authors also showed that for diploids, the appropriate unit to randomize is the diploid genotype rather than the allele (of course, the contingency tables are based on alleles, not genotypes). Petit et al. [2001] showed that an adequate and powerful multilocus test statistic is the sum of individual loci G-statistics. This global test statistic has the advantage over methods such as Fisher's procedure to combine P-values to account for the level of variation encountered at each locus and should therefore be more accurate.

When several levels of population structure are present, it would be convenient to be able to test for the effect of each level independently of the effect of the lower levels in the hierarchy. This can be achieved by acknowledging that the units to randomize are the units defined by the level just below that of interest in the hierarchy. For instance, if we have individuals nested in subpopulations nested in populations, the association of individuals in subpopulations would be tested by permuting individuals among subpopulations but keeping them within their populations of origin. The function `test.within` of package HIERFSTAT carries out such a test. And the association of subpopulations in populations should be tested by permuting whole subpopulations among populations. The function `test.between` of the package carries out this test. If there are more levels in the hierarchy, for instance `level1` nested in `level2` nested in `level3`, testing for the effect of `level2` implies permuting whole units of `level1` among units defined by `level2`, but keeping them within units defined by `level3`. The function `test.between.within` of the package allows carrying out such a test.

It is beyond the scope of this note to quantify the power of these tests. It is however necessary to check that when the null hypothesis of no differentiation at a given level is true, the test gives a proportion of significant results equal to the type I error α usually set at 5% . To this end, I simulated for the four level hierarchy given in Yang [1998] (see the dataset `yangex`) 1000 datasets of three loci each by sampling for each locus two random integers between 1 and 9 according to a uniform distribution. For each of these datasets, I tested for the effect of `sspop` by permuting individuals among `sspop` within `spop` using the

command

```
test.within(datataset,test=sspop,within=spop)
```

The effect of `spop` was tested using the command

```
test.between.within(dataset,within=pop,rand.unit=sspop,test=spop)
```

Last, the effect of `pop` was tested using the command

```
test.between(dataset,rand.unit=spop,test=pop)
```

The number of datasets for which I obtain significant results at $\alpha = 5\%$ were respectively of 56, 39 and 40, giving in each case a proportion of significant results close to the nominal α, as it should (I also checked that the distribution of p-values did not differ from uniform).

A quick tutorial

I assume that you have downloaded and installed the HIERFSTAT package. After having opened an R session and loaded the package (`library(hierfstat)`), simplified procedures for loading data set have been defined. The function `read.fstat.data` allows you to read in `R` a data file formatted for the program FSTAT (Goudet [1995]). Alternatively, you can read in `R` (using the `R` command `read.table(filename)`) any text file with columns separated by space or tabs. Two such files are provided in the folder data: `yangex.txt` and `exhier.txt`.

As an example, you can load the data set `gtrunchier`. It is directly accessible by using the command `data(gtrunchier)`. The data set is made of 370 diploid multi-locus genotypes of *Galba truncatula*, a freshwater snail sampled from 29 different patches (column `Patch`) belonging to 6 different localities (column `Locality`) of western Switzerland (see Trouve et al. [2004]). The other columns of the data set consist in the genotypes of the different individuals at each of 6 microsatellite loci. By issuing the command `attach(gtrunchier)`, you obtain access to each column of the dataset separately.

From the above data, the commands

```
loci<-data.frame(L21.V,L37.J,L20.B,L29.V,L36.B,L16.J)
varcomp.glob(data.frame(Locality,Patch),loci)
```

will produce the estimation of the variance components for each locus and overall, as well as the matrix of hierarchical F-statistics, which reads as follows: on the first line from left to right $F_{Locality/Total} = 0.516, F_{Patch/Total} = 0.643, F_{Ind/Total} = 0.915$, on the second line $F_{Patch/Locality} = 0.262, F_{Ind/Locality} = 0.825$ and on the third line $F_{Ind/Patch} = 0.763$.

To test for the effect of localities, issue the command

```
test.between(loci,rand.unit=Patch,test=Locality,nperm=1000)
```

I obtain a p-value of 0.001 meaning that none of the randomized data set gave a global $G_{Locality}$ statistic larger than that obtain from the observed data set. To test for the effect of patches, issue the command

```
test.within(loci,test=Patch,within=Locality,nperm=1000)
```

Again, I obtain a p-value of 0.001 meaning that none of the randomized data set gave a global G_{Patch} statistic larger than that obtain from the observed data set. Hence I conclude that there is both a strong effect of patches within localities and of localities on the genetic structure of this snail.

The package HIERFSTAT should complement nicely the collection of softwares available to analyse population genetics data. It is available from `http://www.unil.ch/popgen/softwares/hierfstat.htm`.

References

L. Excoffier. *Handbook of Statistical Genetics*, chapter Analysis of population subdivisions, pages 271–307. John Wiley and Sons, 2001.

J. Goudet. Fstat (version 1.2): A computer program to calculate f- statistics. *Journal of Heredity*, 86(6):485–486, 1995. URL `http://www.unil.ch/popgen/softwares/fstat.htm`.

J. Goudet, M. Raymond, T. deMeeus, and F. Rousset. Testing differentiation in diploid populations. *Genetics*, 144(4):1933–1940, 1996.

P.O. Lewis and D. Zaykin. *Genetic data analysis: computer program for the analysis of allelic data*, 2001. URL `http://lewis.eeb.uconn.edu/lewishome/software.html`.

M. Nei. *Molecular Evolutionary Genetics*. Columbia University press, first edition, 1987.

E. Petit, F. Balloux, and J. Goudet. Sex-biased dispersal in a migratory bat: A characterization using sex-specific demographic parameters. *Evolution*, 55(3): 635–640, 2001.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

S. Schneider, D. Roessli, and Excoffier L. *Arlequin version 2.000: a software for population genetics data analysis*, 2000. URL `http://lgb.unige.ch/arlequin`.

S. Trouve, L. Degen, and J. Goudet. Ecological components and evolution of selfing in the freshwater snail galba truncatula. *Journal of Evolutionary Biology*, 17:In Press, 2004.

B.S. Weir. *Genetic Data Analysis*. Sinauer, first edition, 1990.

B.S. Weir. *Genetic Data Analysis II*. Sinauer, second edition, 1996.

S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15:323–354, 1951.

R.C. Yang. Estimating hierarchical f-statistics. *Evolution*, 52:950–956, 1998.