Computer Note

FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics

J. Goudet

Computation of Wright's (1943, 1951) fixation indices ($F_{st}$, $F_{st}$, and $F_{st}$) is widespread among population biologists to assess genetic differentiation of populations. $F_{st}$ is a measure of the within-population heterozygote deficit, $F_{st}$ is a measure of the among-population heterozygote deficit (the Wahlund effect), and $F_{st}$ is a measure of the global heterozygote deficit. These indices could be defined as the correlation between uniting gametes (Wright 1943, 1951), as a function of gene diversity in the total population (Nei 1973, 1977; Nei and Chesser 1983), or as a function of variance components from a nested analysis of variance (Cockerham 1969, 1973; Cockerham and Weir 1986; Weir and Cockerham 1984). Deriving unbiased estimators of these quantities has been at the center of an argument between Nei on the one hand (Nei 1986; Nei and Chesser 1983) and Weir and Cockerham (Cockerham 1973; Cockerham and Weir 1986, 1987) on the other. An in-depth review of the properties of both families of estimators can be found in Chakraborty and Dander-Hopple (1991). Cockerham and Weir (1993) presented an interpretation of the differences between Nei's estimator of $F_{st}$, $G_{st}$, and their own, $F_{st}$, in terms of the probabilities of identity by descent.

The most often used package for analysis of allele and genotype frequencies is Swafford and Selander's (1981) BIOSYS-1. Although very useful for calculation of genetic distances and building of phylogenetic trees, its module on $F$ statistics estimation is outdated, because it is based on Nei's (1977) article, which does not take sampling affects into account. I propose a PASCAL program that performs the calculation of Weir and Cockerham's (1984) estimators of $F$ statistics, based on the Fortran listing published in Weir's (1990) Genetic Data Analysis, with several new features.

The Program FSTAT

The program FSTAT performs the following:

- Estimated frequency of alleles per sample and overall (from which indices such as Nei's gene diversity can be directly calculated).
- Observed and expected heterozygosity per allele and sample. Expected heterozygosity is calculated using Levene's (1949) correction for small samples.
- $F_{st}$ (in Weir and Cockerham's notation) estimated per sample over loci.
- $F_{st}$ ($F$), $F_{st}$ ($0$), and $F_{st}$ ($t$) estimated per allele, locus, and globally over all samples. A fourth index has been added, a measure of Hamilton's (1971) relatedness, $r = 2F_{st}(1 + F_{st})$, using the estimator given in Queller and Goodnight (1988). This measure is the average relatedness of individuals within samples when compared to the whole. It is often used in studies of social insects.
- Confidence intervals based on resampling schemes are provided: (1) Mean and variance of $F$ statistics per locus, estimated from jackknifing over samples. (2) Mean and variance of $F$ statistics over loci, estimated from jackknifing over loci. (3) Bootstrap confidence intervals of $F$ statistics performed on the loci.
- Calculation of $F_{st}$ per pair of samples. The output is a matrix of $F_{st}$ values that can be used to carry out Slatkin's (1993) method to test for isolation by distance because $F_{st}$ is closely related to a genetic distance (Reynolds et al. 1983). This matrix could also be used in Mantel tests (e.g., Manly 1985).
- Test of the significance of $F_{st}$, $F_{st}$, and $F_{st}$ per locus and over all loci using permutations (Excoffier et al. 1992; Hudson et al. 1992; Manly 1991). The aim is to obtain the distribution of the null hypothesis, namely $F_{st} > 0$, and to compare this null distribution with the observed $F_{st}$. The probability of obtaining by chance a value as large or larger than the observed is given: (1) For $F_{st}$, alleles are permuted among individuals within samples. (2) For $F_{st}$, alleles are permuted among samples. (3) For $F_{st}$, two types of tests can be carried out, depending on the results of the test on $F_{st}$. If $F_{st}$ is not significantly different from zero, it is valid to permute alleles among samples to test $F_{st}$ because alleles can be considered as independent. If $F_{st}$ is different from zero, however, alleles within individuals are not independent anymore, and the appropriate permutation units are the genotypes, to be permuted among samples.

These tests were developed to avoid the caveats of existing tests, such as those of Workman and Niswander (1970), based on $X^2$ and therefore relying on large samples (expected classes larger than 5). Raymond and Rousset (in press) generalized Fisher's exact test for Hardy-Weinberg equilibrium to among samples differentiation. Some problems remain however with this test, because combining information from different loci is carried out using Fisher's procedure (Fisher 1954; Sokal and Rohlf 1981), which does not weight loci. Furthermore, their test for between sample differentiation is based on the assumption that there is Hardy-Weinberg equilibrium within samples. If there is departure from it, then alleles within individuals are not independent, and the exact test for differentiation would lead to erroneous results. Permutations eliminate those caveats. Slatkin (1994) pointed out that, when studying population differentiation, it may...
be more appropriate to use $F_{st}$, a statistic arising naturally, as a test statistic.

The Random Number Generator proposed by L'Ecuyer (1988) was chosen for the Bootstrap and permutation procedures. It combines two of the best Multiplicative Linear Congruential Generators known and has passed all the tests for random number generators.

The format of the output file (tab separated) allows direct reading into many commercially available spreadsheets, facilitating printing and graphical representation of the data.

A real mode version of the program runs on 80286 (and above) PC compatibles. No coprocessor is required, but will speed up calculations. A protected mode version will run on 80386 (and above) PC compatibles. Again, no coprocessor is required. This version uses all the available extended memory, therefore allowing the processing of larger data sets. The actual limits are:

- Number of samples: 200
- Number of locus: 50
- Number of alleles at the most polymorphic locus: 99
- Maximum number of individuals: 5,000
- Maximum number of permutations: 15,000

The program is also suited for haploid data and appropriately handles missing data, such as a locus missing completely from one sample. The program is distributed with no charges. It can be sent electronically in a Binhexed or unencoded format (requests should be sent to Jerome.goudet@izea.unil.ch). Alternatively, it can be retrieved from the ftp server oracle.bangor.ac.uk after anonymous login, in the directory pub/fstat.

From the School of Biological Sciences, University of Wales, Bangor, U.K., and Institut de Zoologie and d’Ecologie Animale, Bat Biologie, Universite de Lausanne, Lausanne CH-1015, Switzerland. This work forms part of a research program into the evolutionary effect of gene flow and has been partly funded by the University of Wales, Bangor, the Department of Environment, U.K., and the Swiss National Science Foundation. It is a contribution to the Biodiversity Module of the Swiss Priority Program on the environment. I am indebted to Thierry DeMeeus for thorough checking of the program capabilities. Many thanks are due to Chris Gliddon, Michel Raymond, and Francois Rousseau.

The Journal of Heredity 1995:86(5)

**References**


Received August 30, 1994

Accepted March 14, 1995

Corresponding Editor: Stephen O'Brien