Received 25 July 1994

How should we bootstrap allozyme data?

S. VAN DONGEN

Department of Biology, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium

The use of the bootstrap to estimate the distribution of statistics from allozyme data is examined. The different loci are often used as the unit of resampling. Since the interpretation and validity of the bootstrap is affected by the unit of resampling, and since resampling over loci, in most practical cases, does not conform to the three basic assumptions of the bootstrap, this method should be avoided. Resampling over individual genotypes may provide a valid alternative approach.

Keywords: bootstrap, *F*-statistics, genetic distance, jackknife, phylogenies, repeatability.

Introduction

The bootstrap (Efron, 1979) is a general resampling technique for estimating the distribution of statistics, which has been recently applied in several fields of population genetics and systematics. Felsenstein (1985) introduced a resampling scheme for bootstrapping phylogenetic trees to estimate confidence limits of internal branches. This method proceeds by resampling the species (or subpopulation) × character matrix across characters, with replacement, such that the resamples consist of a set of characters with some of them duplicated, triplicated,... and others absent. The program SEQBOOT in the PHYLIP 3.5 computer package (Felsenstein, 1993), for example, resamples across loci for allozyme data. Felsenstein (1985) suggests that the results from the bootstrap replicates can then be combined in a consensus tree to asses confidence in the branches. This approach of bootstrapping was first used to assess the repeatability of a given result (Hillis & Bull, 1993). Whether it can be given a statistical interpretation in the sense of a *P*-value has been investigated and criticized by several studies (Zharkikh & Li, 1992; Hillis & Bull, 1993; but see Felsenstein & Kishino, 1993).

A similar way of resampling is used by the programs HAPLOID and DIPLOID by Weir (1990a,b). Here also the loci are the units across which the resampling is performed to estimate the distribution and consequently confidence intervals of *F*-statistics. In the program by Lessios (1990), the standard errors of the means of Nei's and Hillis's genetic distances are estimated by jackknifing over loci. Jackknifing is a technique which is very similar to bootstrapping and consequently requires the same assumptions (Miller, 1974; Efron, 1979) (see below). Although this way of resampling has been applied in several studies (Zharkikh & Li, 1992; Prout & Baker, 1993), there are three basic statistical problems involved with it if the data consist of allele frequencies from allozyme data or RFLPs. In this note I will argue that resampling across loci, for such datasets, is in most cases inappropriate and that the development and investigation of new resampling schemes should rather focus on resampling over individual genotypes.

Assumptions of the bootstrap

Bootstrapping consists of approximating the sampling distribution of $R(X_n,F)$, where $X_n=(X_1, X_2, ..., X_n)$ is a sample of independent, identically distributed (i.i.d.) random variables with common distribution function F, by the bootstrap distribution of $R^*=R(X^*,F_n)$, where $X_n^*=(X_1^*, X_2^*, ..., X_n^*)$ denotes a random sample of size n from F_n, the empirical distribution function. Efron (1979) suggests a Monte Carlo approximation to obtain the bootstrap distribution of R^*. By repeatedly generating random samples of size n (say B times) from F_n and calculating R_i^*, the sampling distribution of R_1^*, R_2^*, ..., R_B^* can be taken as an approximation of the actual bootstrap distribution of R^*. This approximation can be made arbitrarily accurate by taking B sufficiently large. The bootstrap has been successfully applied to estimate bias, variances and confidence intervals of a broad range of statistics (see for example Efron, 1979, 1981; Freedman, 1981; Stine, 1985) and for hypothesis testing (Hall & Wilson, 1991). More elaborate reviews are given by Efron (1979), Efron & Tibshirani (1986), and Hall (1988). The bootstrap distribution has been proven to be asymptotically accurate (Bickel & Freedman, 1981; Singh, 1981), but that is no guarantee of a good small sample behaviour (Efron & Tibshirani, 1986; Van Dongen & Backeljau, 1995), and there has been insufficient basic research to characterize when the bootstrap can be expected to be reliable (Noreen, 1989; Manly, 1991).

Violations of the assumptions

In the context of bootstrapping phylogenies according to the method of Felsenstein (1985) with allozyme data, the statistic of interest ($R(\boldsymbol{X}_n,F)$) is the presence or absence of a node in the dendrograms based on the resamples, and the observations ($\boldsymbol{X}_i$) are the vectors of allele frequencies in the different subpopulations (or species) for the ith locus. These allele frequencies for the different loci are often estimated from the same (or partly the same) set of individuals. Therefore, the vectors of allele frequencies at the different loci (the $\boldsymbol{X}_i$s), are not independent from each other. Furthermore, the allele frequencies of the different loci must have the same distribution. This means that the same evolutionary processes must act upon the different loci. In many practical cases this will certainly be violated. Selection, either diversifying or balancing, may play a significant part in the evolution of allele frequencies (see, e.g. Nevo & Beiles, 1988; Goulson, 1993; Skibinski *et al.*, 1993), while the rate of divergence of selectively neutral loci is higher for more variable systems (Skibinski *et al.*, 1993). We can thus conclude that, for allozyme and single locus RFLP data, the units of resampling (i.e. the loci) are in many practical cases not i.i.d. and that irrespective of other possible critisisms, based on these problems, bootstrap results of phylogenies based on resampling over characters should not be interpreted as a P-value but only as a measure of repeatability.

Thirdly, there is the problem of small sample sizes. Although the number of individuals sampled in most studies is large enough to expect the bootstrap to be reliable, the number of loci scored is usually relatively limited and rarely exeeds 10–15. This causes the bootstrap distribution to be discrete and to have some peculiar properties, since the number of possible different resamples is limited. Van Dongen and Backeljau (1995), for example, showed that the bootstrap fails if the sample size is < 20 for the estimation of the distribution of single locus inbreeding coefficients, probably because the small samples did not contain all the 'important' information on heterozygosity.

The same three arguments hold for the resampling over loci to estimate confidence intervals of F-statistics and Nei's and Hillis's genetic distance. Although the loci can be expected to provide nearly independent replicates of the genetic sampling process (Weir, 1990b), the data (i.e. the allele frequencies) for the different loci are often not independent of each other, and in most practical cases different loci are influenced by at least partly different evolutionary forces such that the distributions of the data may be unequal.

Alternative approach

The independent units of observations are the individual genotypes. The obvious way of resampling is thus across individuals (see Van Dongen & Backeljau, 1995 for such an approach), such that all three previous problems can be avoided. If more than one subpopulation (or species) is involved in the analysis, one should rather keep the resampling separate (Crowley, 1992). For estimating confidence of phylogenetic dendrograms, this can be achieved by making for each resample a new data set for each subpopulation (of size n_i) by resampling across individuals within subpopulations. By combining these results in a consensus tree, one can assess the repeatability and statistical confidence of the originally obtained result, and how robust the estimated phylogenetic relationships are against resamples within each subpopulation. The difference in interpretation of this repeatability and possibly confidence measure from the one estimated by resampling across loci is that for the latter, variation in the resamples is of a different nature. Adopting the terminology of Slatkin & Arter (1991), there are three sources of variation in population genetic data. The first is *sampling variation*, arising from the sampling process when the data are collected. The second source of variation, *stochastic variation*, is the result of the stochastic processes governing allele frequencies at that locus. While the third source of variation, the so-called *parametric variation*, results from differences in mutation rate among loci. By resampling over loci, the variation in the resamples is mainly the result of parametric variation and differences in stochastic variation among loci and thus, the variation in the resamples reflects the variation arising from the selection of the loci, while resamping across individuals provokes variation in the resamples reflecting the sampling variation. For the latter approach, the investigator has to select (and this selection must not be random to ensure that the bootstrap can be applied correctly, contrary to the previous approach) a set of loci from which he wants to estimate the phylogenetic relationships. By resampling over idividuals within the subpopulation (or species), he can investigate how robust these phylogenetic relationships are against resamples of the original data. Whether the obtained measure of repeatability can be given a statistical interpretation in the sense of a P-value remains to be tested by simulation studies because the bootstrap may fail in some situations (Bickel & Freedman, 1981; Singh, 1981). At least, this new resampling approach conforms to the basic requirments of the bootstrap.

For the estimation of confidence intervals for F-statistics and the genetic distances, one can either resample over the complete genotype array and estimate the distribution of the average fixation index, or do the analysis separately for the different loci and combine the results by the method of Rubin & Schenker (1991). Under the 'Random model' (sensu Weir, 1990a,b), however, resampling over individual genotypes is not valid since the individuals cannot be regarded as independent (Weir, 1990b). Obviously, most of these alternative methods have to be examined for their reliability by simulation studies (see Van Dongen & Backeljau, 1995 for such an approach).

Although P-values obtained by resampling across loci (or populations) should in most practical cases not be given a statistical interpretation, this approach may yield informative results for the determination of the presence of disturbing forces such as selection (see Weir, 1990b, p.151). Generally, jackknifing can be applied to detect outliers, like for example loci that are not selectively neutral, by investigating the magnitude of the different pseudo-values (Miller, 1974).

A few computer programs are already available which resample over individual genotypes. The program FIXTEST performs one- and two-sample bootstrap tests on single-locus inbreeding coefficients (Van Dongen & Backeljau, 1995). FSTAT (Goudet, 1994) estimates the distribution of the null hypothesis of F-statistics by permutation over alleles and/or genotypes. The advantage of the bootstrap over this permutation approach is that with the first, the distribution of the F-statistic itself is estimated such that it can be statistically compared with any expected value and not only with zero, and that two observed values can be compared (see also Van Dongen & Backeljau, 1995). A Turbo Pascal program to estimate Weir & Cockerham's (1984) estimator of F_{st} (theta), and to perform one- and two-sample bootstrap tests by resampling over individual genotypes can be obtained from me. Send me a formatted 3.5 inch floppy disk and an envelope with your name and address on it.

Acknowledgements

S.V.D. is Research Assistant at the National Fund for Scientific Research (Belgium).

References

BICKEL, P. J. AND FREEDMAN, D. A. 1981. Some asymptotic theory for the bootstrap. *Ann. Stat.*, **9**, 1196–1217.

CROWLEY, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Ann. Rev. Ecol. Syst.*, **23**, 405–447.

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

EFRON, B. 1981. Censored data and the bootstrap. *J. Am. Stat. Ass.*, **76**, 312–319.

EFRON, B. AND TIBSHIRANI, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.*, **1**, 54–77.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

FELSENSTEIN, J. 1993. PHYLIP (Phylogenetic Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington.

FELSENSTEIN, J. AND KISHINO, H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.*, **42**, 193–200.

FREEDMAN, D. A. 1981. Bootstrapping regression models. *Ann. Stat.*, **9**, 1218–1228.

GOULSON, D. 1993. Allozyme variation in the butterfly, *Maniola jurtina* (Lepidoptera: Satyrinae) (L.): evidence for selection. *Heredity*, **71**, 386–393.

GOUDET, J. 1994. *FSTAT, a program for IBM PC compatibles to calculate Weir and Cockerham's estimators of F-statistics.* Biology Department, Lausanne University, Switzerland.

HALL, P. 1988. Theoretical comparison of bootstrap confidence intervals. *Ann. Stat.*, **16**, 927–953.

HILLIS, D. M. AND BULL, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, **42**, 182–192.

LESSIOS, H. A. 1990. Program for calculating genetic distances and jackknifed confidence intervals. *J. Hered.*, **81**, 490.

MANLY, B. F. J. 1991. *Randomization and Monte Carlo Methods in Biology.* Chapman and Hall, London.

MILLER, R. G. 1974. The jackknife – a review. *Biometrika*, **61**, 1–15.

NEVO, E. AND BEILES, A. 1988. Genetic parallelism of protein polymorphism in nature: ecological test of the neutral theory of molecular evolution. *Biol. J. Linn. Soc.*, **35**, 229–245.

NOREEN, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses.* John Wiley, New York.

PROUT, T. AND BARKER, J. S. F. 1993. *F* statistics in *Drosophila buzzatii*: selection, population size and inbreeding. *Genetics*, **134**, 369–375.

RUBIN, D. B. AND SCHENKER, N. 1991. Multiple imputation in health-care databases: an overview and some applications. *Stat. Med.*, **10**, 585–598.

SINGH, K. 1981. On asymptotic accuracy of efron's bootstrap. *Ann. Stat.*, **9**, 1187–1195.

SKIBINSKI, D. O. F., WOODWARK, M. AND WARD, R. D. 1993. A quantitative test of the neutral theory using pooled allozyme data. *Genetics*, **135**, 233–248.

SLATKIN, M. AND ARTER, H. E. 1991. Spatial autocorrelation methods in population genetics. *Am. Nat.*, **138**, 499–517.

STINE, R. A. 1985. Bootstrap prediction intervals for regression. *J. Am. Stat. Ass.*, **80**, 1026–1031.

VAN DONGEN, S. AND BACKELJAU, T. 1995. One- and two-sample tests on single locus inbreeding coefficients using the bootstrap. *Heredity*, **74**, 127–133.

WEIR, B. S. 1990a. Intraspecific differentiation. In: Hillis, D. M. and Moritz, C. (eds) *Molecular Systematics*, pp. 373–410. Sinauer, Sunderland, MA.

WEIR, B. S. 1990b. *Genetic Data Analysis.* Sinauer Ass. Inc.

ZHARKIKH, A. AND LI, W. H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Mol. Biol. Evol.*, **9**, 1119–1147.