Points of View

Syst. Biol. 44(3):421–429, 1995

Polymorphic Characters and Phylogenetic Analysis: A Statistical Perspective

BRUCE RANNALA

Department of Biology, Yale University, P.O. Box 208104, New Haven, Connecticut 06520-8104, USA[1]

Most data used in phylogenetics display variable degrees of intraspecific (or intrataxon) character polymorphism. The most important classes of intraspecific polymorphism include variable allele frequencies (for DNA sequences or allozyme data), variable phenotypic means (for morphometric data), and different ontogenetic states among individuals (for developmental data) (Mabee and Humphries, 1993). Polymorphism has a special interpretation in phylogenetics. In this paper, I adopt a commonly accepted phylogenetic meaning of the term *polymorphic* (as recently defined by Mabee and Humphries, 1993) as the presence of more than one character state within a terminal taxon.

A number of methods have been proposed for dealing with polymorphic characters in relation to character coding and phylogeny reconstruction (e.g., Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967; Felsenstein, 1973; Cavalli-Sforza and Piazza, 1975; Mickevich and Mitter, 1981, 1983; Patton and Avise, 1983; Buth, 1984; Swofford and Berlocher, 1987; Mabee and Humphries, 1993). Less attention has been focused on the possible sources of sampling error associated with polymorphic characters. Two broadly different approaches to phylogenetic analysis using polymorphic characters have developed: (1) the characters are first classified as either fixed or polymorphic, and characters in the two categories are treated differently in constructing the phylogeny (Mickevich and Mitter, 1981, 1983; Patton and Avise, 1983; Buth, 1984); or (2) the character-state frequencies are used directly in constructing the phylogeny (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967; Felsenstein, 1973; Cavalli-Sforza and Piazza, 1975; Swofford and Berlocher, 1987).

The accuracy of the first approach to phylogeny estimation may be reduced if characters that are actually polymorphic are coded as fixed for a taxon because of sampling error (i.e., the polymorphism is not represented in the sample from the taxon), and the accuracy of the second approach may be reduced by errors in estimates of the frequency of a polymorphism in a taxon. Other methods of phylogenetic analysis based on stochastic models of nucleotide substitution for DNA sequences (such as the Poisson process: Langley and Fitch, 1974; Felsenstein, 1981) may make implicit assumptions about particular nucleotides being fixed in different taxa.

The character sampling process may introduce two types of error for estimates of phylogeny: error in estimates of the topology and error in estimates of the branch lengths (see discussion by Nei, 1987). The influence of sampling variance on errors of the first type is poorly understood (but see Weir and Basten, 1990; Bulmer, 1991), but

[1] E-mail: rannala@minerva.cis.yale.edu.

its influence on estimates of branch lengths and genetic distances among taxa has received more study (Nei and Roychoudhury, 1974; Chakraborty, 1977; Nei et al., 1985; Takahata and Tajima, 1991; Tajima, 1992).

In estimating a species phylogeny using polymorphic characters, there are two distinct sources of sampling variance: that due to the sampling of individuals (or alleles) for each character state and that due to the sampling of characters (or loci). For molecular characters, these are the intralocus and interlocus sampling variances, respectively (Nei and Roychoudhury, 1974). In estimating phylogeny for a particular gene, intralocus sampling is the only source of sampling variance. If there is no intrataxon character polymorphism, interlocus sampling is the only source of sampling variance. Nei and Roychoudhury (1974) suggested that the interlocus sampling variance (which is a property of the evolutionary process as well as the gene sampling process) is of greatest importance when estimating genetic distance. However, their suggestion is based on an assumption that a simple random sample (with respect to genotype) is obtained from each population. Population genetic subdivision and the resulting nonrandom sampling of individuals may violate this assumption and greatly increase the importance of the intralocus sampling variance.

In this paper, I examine the influence of three primary factors on the sampling variance of a frequency estimate for a polymorphic character: finite population size, nonrandom mating within populations, and population subdivision and differentiation. The methods I present may be applied to a variety of different types of characters, although here I explicitly consider properties of an estimator of the allele frequency at a single genetic locus. Standard statistical methods of cluster sampling (Deming, 1950; Cochran, 1977) may be applied for estimating the average allele frequency (and the variance of this estimate) in a metapopulation (i.e., a collection of populations exchanging migrants; Levins, 1969) based on a stratified sample of individuals and populations.

As an application of this approach, I consider the effect of genetic differentiation among populations on the variance of an estimate of the average allele frequency in a metapopulation. In addition, I examine the likelihood of mistaking a polymorphic allele for a fixed-state character when there is population subdivision and differentiation of allele frequencies among subpopulations. These sources of sampling variance (in particular population subdivision) have often been neglected in systematics studies, although they may have important effects on the accuracy of estimated phylogenies. Recent advances in DNA sequencing technology promise to reveal much more intrataxon polymorphism than is currently recognized, and previously hidden polymorphisms may be an important source of error for phylogenies based on sequences from only one (or a few) individual(s) per taxon.

In this study, I consider only the properties of an estimator of the frequency of a single allele at a single genetic locus (or equivalently a single character state for a single polymorphic morphological character) and its variance and simply highlight the potential importance of this source of sampling error when population subdivision exists. In addition to the interlocus variance, the variances for some distance measures employed in phylogenetic analysis using molecular data depend on higher order moments (and among-taxa covariances) of the intralocus allele frequency distribution as well as the mean and variance as considered here (see Nei and Roychoudhury, 1974).

Statistical Estimators of Allele Frequency

A maximum likelihood estimator of allele frequency for a diploid species (Li and Horvitz, 1953) is given by

$$\hat{p} = \frac{(2n_{11} + n_{12})}{2n}, \qquad (1)$$

where n individuals are sampled with n_{11} homozygous for the allele and n_{12} heterozygous. For a haploid species, an equivalent estimator is

$$\hat{p} = \frac{n_1}{n}. \qquad (2)$$

Estimation of the allele frequency for a single population is straightforward. Here, I am mostly concerned with estimating the average allele frequency in a collection of disjunct populations and the associated sampling variance. The results will be considered in light of their importance for phylogenetic analysis.

A number of authors (e.g., Nei, 1987; Swofford and Berlocher, 1987; Slatkin and Arter, 1991) have applied the following maximum-likelihood estimator for the sampling variance of the frequency of an allele in a population with an estimated frequency of $\hat{p}$ for a sample of n diploid individuals:

$$s^2_{\hat{p}} = \frac{\hat{p}(1 - \hat{p})}{2n}. \qquad (3)$$

This estimator is based on several assumptions: (1) the sample n is much smaller than the population size N, or sampling is with replacement; (2) the distribution of genotypes among individuals fits that expected under Hardy–Weinberg equilibrium; and (3) all individuals in the population are equally likely to be sampled. The effects of violations of each of these assumptions on an estimate of the average allele frequency (and its variance) will be considered here. A nonrandom sampling process associated with population subdivision generally has the most important influence on the variance of estimates of average allele frequency.

Influence of Finite Population Size

If the first assumption is violated, i.e., a large fraction of the individuals in a population are sampled (without replacement), then Equation 1 still provides an unbiased estimate of the average allele frequency, and the following unbiased estimator for the variance of this estimate (e.g., Rice, 1988:187) may be used (for a diploid species) in place of Equation 3:

$$s^2_{\hat{p}} = \frac{\hat{p}(1 - \hat{p})}{2n - 1}\left(1 - \frac{n}{N}\right), \qquad (4)$$

where $\hat{p}$ is an estimate of the allele frequency (i.e., obtained from Eq. 1), n is the sample size, and N is the total number of individuals in the taxon. Equation 4 will differ little from Equation 3 if n is small relative to N, and the estimated variance will always be less than that for Equation 3.

Influence of Nonrandom Mating

The effect of a violation of the second assumption, i.e., a deviation from the Hardy–Weinberg distribution of genotypes among individuals, can be examined using the following approach. Let F be the fixation index for a diploid population; F is a measure of the degree of departure of the distribution of genotypes among individuals from Hardy–Weinberg proportions due to nonrandom mating; F will assume positive values when local inbreeding occurs, negative values when assortative mating occurs, and a value of zero when mating is random (see Nei, 1987). Equation 1 still provides an unbiased estimate of the average allele frequency (for a diploid species) if there is nonrandom mating, and if $N \gg n$ then the standard result for the variance of this estimate (e.g., Li, 1976:242) is

$$s^2_{\hat{p}} = (1 + F)\frac{\hat{p}(1 - \hat{p})}{2n}. \qquad (5)$$

Thus, the variance of an estimate of p, when nonrandom mating occurs, is a simple linear function of the fixation index F. If $F = 0$, the population is in Hardy–Weinberg equilibrium, and the variance of an estimate of allele frequency based on a sample of n individuals (from Eq. 5) is identical to that given by Equation 3. If $F = 1$, so that the level of inbreeding is maximized, the variance of the estimated allele frequency is given by $p(1 - p)/n$ and is increased over that for a randomly mating population (e.g., Eq. 3) by a factor of 2.

For extreme inbreeding with $F = 1$, the correlation between allele types from the same individual is 1, so that the type of the second allele is immediately known upon examining the first allele. Thus, a sample of $2n$ alleles from a highly inbred population provides no more information about allele frequency than a sample of n alleles from a randomly mating population (in which the correlation between allele types in individuals is zero). Inbreeding tends to increase the variance of an estimate of allele frequency, whereas assortative mating tends to decrease it. Nei and Roychoudhury (1974) suggested a similar effect of inbreeding on an estimate of population heterozygosity, although they provided no explicit mathematical formulation. The increase in variance of an allele frequency estimate due to nonrandom mating will never exceed twice that expected for a population in Hardy–Weinberg equilibrium and may often be neglected in practice.

Influence of Nonrandom Sampling

Violations of assumptions 1 and 2 will have a limited effect on the variance of an estimate of allele frequency. A more important source of variance is that due to a violation of assumption 3, that all individuals in a taxon have an equal likelihood of being sampled. There are a number of ways in which this assumption may be violated; only one of these possibilities, the effect of population subdivision, is considered in detail here. In general, population subdivision acts to increase the variance of an estimate of average allele frequency based on a sample of individuals and populations.

Estimating Allele Frequency in a Subdivided Population

Consider a taxon composed of M populations, where p_i is the frequency of a particular allele in the the ith population and N_i is the total number of haploid individuals in the ith population, with $i = 1, \ldots, M$. A haploid species is used for illustration to avoid assumptions concerning genotype distributions, but the same methods may be applied to a diploid species in Hardy–Weinberg equilibrium (within populations) by replacing N with $2N$ in all cases (i.e., for both n and N). A sample of n_i individuals is taken from each of $i = 1, \ldots, m$ populations ($m \leq M$). Populations are chosen with equal probability, and individuals are sampled within each chosen population with equal probability. A ratio estimator of the allele frequency in the taxon as a whole is the weighted average of the allele frequency estimates for the sampled populations,

$$\hat{p}_{\mathrm{T}} = \frac{\sum_{i=1}^{m} N_i \hat{p}_i}{\sum_{j=1}^{m} N_j}, \qquad (6)$$

where $\hat{p}_i$ is the allele frequency in the ith population, estimated using Equation 2. This is a consistent estimator of p_{T}, although it is biased. In practice, the population sizes (N_i) will often not be known. An equivalent estimator that may then be more easily applied is

$$\hat{p}_{\mathrm{T}} = \sum_{i=1}^{m} \varphi_i \hat{p}_i, \qquad (7)$$

where

$$\varphi_i = \frac{N_i}{\sum_{j=1}^{m} N_i}. \qquad (8)$$

It follows that φ_i is the proportion of the total individuals in the sampled populations that are contained in the ith population. This relative proportion may often prove easier to estimate than the actual population sizes.

For a two-stage sampling process in which populations are chosen with equal probability and individuals within each sampled population are chosen by simple random sampling, the following formula for the variance of the ratio estimator of allele frequency given by Equation 7 may be obtained by a modification of the two-stage cluster sampling theory derived by Cochran (1977; see Appendix):

$$\mathrm{Var}(\hat{p}_\mathrm{T}) \approx \left(1 - \frac{m}{M}\right) \sum_{i=1}^{m} \varphi_i^2(\hat{p}_i - \hat{p}_\mathrm{T})^2 + \frac{m}{M} \sum_{i=1}^{m} \varphi_i^2(1 - \gamma_i)\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}, \tag{9}$$

where $\gamma_i = n_i/N_i$ is the proportion of individuals in the ith population that are sampled. Because N_i is generally unknown, γ_i will often be difficult to estimate; in most cases it will be small enough to be neglected, and in any case neglecting γ_i will necessarily result in an overestimate, rather than an underestimate, of the sampling variance. Equation 9 gives the variance of an allele frequency estimate for a metapopulation subdivided into any number of populations, with any degree of genetic differentiation among them, provided that each population in the collection of populations as a whole has an equal probability of being sampled.

The two terms in Equation 9 represent two distinct sources of variance for an allele frequency estimate. The first term represents the variance due to population sampling, and the second term represents the variance due to sampling of individuals within populations. If $M = m$, the first term of Equation 9 vanishes, and if $N_i = n_i$ for all $i = 1, \ldots, m$, the second term vanishes, as expected. Most considerations of the error in estimates of allele frequency due to sampling in a phylogenetic context (i.e., Swofford and Berlocher, 1987; Mabee and Humphries, 1993) have considered only the second component of variance, that due to the random sampling of individuals from a single population.

Influence of Genetic Differentiation

Differences in allele frequency among populations due to restricted gene flow and genetic drift or due to selection can greatly increase the sampling error associated with estimates of the average allele frequency in the collection of populations as a whole. Alternatively, if there is a great deal of gene flow among populations, so that they are genetically homogeneous, the variance in allele frequency estimates due to differences among populations will be negligible. Consider the relationship between the first term of Equation 9 and Wright's (1969) measure of the genetic differentiation among populations F_ST. If n_i is moderately large for all $i = 1, \ldots, m$ or $m \ll M$, the second term of Equation 9 will have little effect on the variance and may be neglected. I use the following definition of F_ST, taken from Nei (1987):

$$F_\mathrm{ST} = \frac{\sigma^2}{p_\mathrm{T}(1 - p_\mathrm{T})}, \tag{10}$$

where p_T is the average allele frequency (weighted by population size) in the taxon as a whole and is given by

$$p_\mathrm{T} = \sum_{i=1}^{M} \frac{N_i}{N_0} p_i, \tag{11}$$

which is the value achieved by the estimator (Eq. 6) above if all the individuals are sampled. The weighted variance in allele frequency among populations (σ^2) is given by

$$\sigma^2 = \sum_{i=1}^{M} \frac{N_i}{N_0}(p_i - p_\mathrm{T})^2. \tag{12}$$

If all the populations are of equal size, Equation 12 reduces to

$$\sigma^2 = \sum_{i=1}^{M} \frac{1}{M}(p_i - p_\mathrm{T})^2, \tag{13}$$

which may be estimated from a sample of m populations using

$$\hat{\sigma}^2 = \sum_{i=1}^{m} \frac{1}{m}(\hat{p}_i - \hat{p}_\mathrm{T})^2. \tag{14}$$

If all the populations are of equal size, the first term of Equation 9 reduces to

$$\frac{1}{m}\left(1 - \frac{m}{M}\right) \sum_{i=1}^{m} \frac{1}{m}(\hat{p}_i - \hat{p}_\mathrm{T})^2, \tag{15}$$

and substituting Equation 14 into Equation 15 produces

$$\frac{1}{m}\left(1 - \frac{m}{M}\right)\hat{\sigma}^2. \tag{16}$$

Because $\hat{\sigma}^2 = F_{ST}\hat{p}_T(1 - \hat{p}_T)$, the first term of Equation 9 may be written

$$\frac{1}{m}\left(1 - \frac{m}{M}\right)\hat{p}_T(1 - \hat{p}_T)F_{ST}. \qquad (17)$$

Thus, when populations are of about equal size, the first term for the variance of average allele frequency is a linearly increasing function of F_{ST}. Therefore, greater genetic differentiation among populations results in a linear increase in the sampling variance of an estimate of average allele frequency.

An Example

As an illustration of the effect of genetic differentiation on estimates of average allele frequency for a subdivided population when a limited number of populations are sampled, consider the following situation. A taxon is composed of 20 populations in total, with 100 individuals in each population and an allele with a frequency of 0.5 in two populations and 0.0 in the remaining populations. In this case, the value of F_{ST}, obtained by an application of Equation 10, is 0.474. The average frequency of the allele is 0.05. Now suppose we choose 10 populations at random and sample 10 individuals from each. The probability that zero populations containing the allele are in the sample (obtained using the hypergeometric distribution) is about 0.24.

Because the probability is very small that one (or both) of the populations that contain the allele is sampled and that none of the 10 alleles sampled from either (or both) are the allele of interest, the probability that the allele is missed in the sample depends almost entirely on the probability that the populations containing the allele are missed in the sample; the value is about 0.24. By contrast, if we consider only the average allele frequency of 0.05 and view the taxon as a single population, the probability that a total of 100 individuals are sampled without observing the allele is 0.95^{100}, or about 0.006 (using the method of calculation of Swofford and Berlocher, 1987). Thus, if we ignore the structure of the sample, missing the allele seems quite improbable, whereas by taking the structure into account it appears quite likely (we expect to miss the allele in about 25% of our samples on average).

Now consider the variance of an estimate of the frequency of this allele. First, applying Equation 3 and ignoring the population structure we obtain $\text{Var}[\hat{p}_T] = 0.0005$. If one of the populations containing the allele is sampled (as would be the case, on average), applying Equation 9 and taking population structure into account results in $\text{Var}[\hat{p}_T] = 0.0113$. Thus, Equation 3 may seriously underestimate the variance of an estimate of allele frequency when there is population subdivision and genetic differentiation.

Implications for Phylogenetic Analysis

The sampling theory considered in this paper relates to only one aspect of the overall sampling process associated with phylogenetic estimation: the intralocus or intracharacter error, which arises from the sampling of individuals for a particular polymorphic character. Population subdivision and differentiation may greatly increase this component of the sampling variance. Two approaches may prove useful in reducing the variance due to character-state sampling for phylogenetic estimation: (1) increase the number of populations sampled or (2) choose conservative characters that are less likely to display polymorphism at the level of the taxon. For molecular data, such conservative characters might include regions of the genome with unusually low rates of mutation or strong selective constraints. The difficulty with this second approach to sampling error reduction is that choosing regions of the genome for an analysis based on their relative rates of substitution may itself lead to systematic biases in phylogenetic estimates by affecting the interlocus component of the sampling error.

I have considered here only the sampling properties of an estimator of the average frequency for a polymorphic character state. Because this ratio estimator is consistent, the estimated allele frequency will converge to its true value for the metapopulation as the number of populations and individuals sampled becomes large. For small samples, the estimator is biased, and it is at present unclear precisely what effect this bias may have on estimates of phylogeny.

There are a number of additional sources of phylogenetic bias that may arise due to the intracharacter sampling process. The general procedure of phylogenetic inference involves attempts to predict the branching relationships among taxa based on covariances in their character states. These covariances may also arise as a result of nongenealogical processes, however, such as natural selection acting on polymorphic populations. Thus, phylogenetic bias may be introduced by the process of sampling individuals for a particular character if individuals in the different taxa are sampled from similar regions. In such cases, shared alleles may be the result of selection in a common environment rather than a shared ancestry among taxa. Collection of an independent stratified random sample for each taxon over its range should help to reduce this source of bias; as should choosing characters that appear likely to be selectively neutral.

An essential requirement for the variance of an allele frequency estimate given by Equation 14 to be meaningful is that all populations of the taxon under investigation have some probability of being included in the sample. The technique presented here has an even stricter requirement: each population must also have an equal probability of being sampled. Fortunately, other statistical methods may be used in cases where the populations under consideration have unequal probabilities of being chosen (e.g., Cochran, 1977). In some cases, if certain populations are very difficult or expensive to sample, one can assign a much lower sampling probability to these populations in collecting the samples and still achieve error estimates using statistical techniques that are an extension of those presented here.

If phylogenies are to be constructed using allele frequencies or by invoking assumptions about alleles being either fixed or polymorphic in a taxon (i.e., Patton and Avise, 1983; Buth, 1984), then an effort should be made to collect a stratified sample from a known collection of existing populations of the taxon to allow confidence limits to be placed on these assumptions. A survey of the degree of genetic differentiation among populations of a taxon, using a measure such as Wright's F_{ST}, may also be helpful in establishing the scale of the sample needed for robust estimates of allele frequencies. If F_{ST} is very near zero for populations of the taxon under consideration, then a sample from a small number of populations is likely to suffice. If F_{ST} is near 1, then a much larger sample of populations will be required. Of course, the statistical error associated with an estimate of F_{ST} must also be taken into account in such cases (see Lynch and Crease, 1990).

Perhaps the greatest difficulty in applying Equations 6 and 9 to natural populations is the necessity of estimating φ_i, the proportion of the total individuals in the sampled populations contained in the ith population, and m/M, the proportion of the total populations that are sampled. Certain species may be more amenable to study than others, given these sampling considerations. A conspicuous, highly vagile species confined to a small region of habitat should show little genetic differentiation, and its relative population numbers should be easy to estimate accurately. A sessile, cryptic species that is distributed over large and possibly disjunct areas, however, may show a greater degree of genetic differentiation, and its relative population numbers may be more difficult to quantify.

The results of this analysis suggest that a cautious approach should be taken in estimating phylogenies using polymorphic characters because the sources of sampling error are poorly understood and have not been adequately formulated. In many cases, the sampling errors for intrataxon allele frequencies may be much higher than previously assumed or may be difficult to estimate because of limited knowledge of the distributions and the relative sizes of the taxon populations. Many different types of phylogenetic characters may turn out to be polymorphic at the taxon level, including amino acid sequences for particular proteins and DNA nucleotide sequences for particular genes. Population-level studies of the DNA nucleotide sequence variation at individual loci are increasing in number, and intrataxon character-state variability may soon be shown to be much more widespread than is currently recognized.

Much work remains to be done to relate the distributions of characters and character states and the statistical procedures for sampling characters and individuals to the accuracy of phylogenetic estimation procedures. This paper has been restricted to a consideration of the process of sampling individuals to estimate character-state frequencies for a polymorphic character; it remains to be shown precisely how this sampling process will affect the overall accuracy of estimated phylogenies.

ACKNOWLEDGMENTS

Research support was provided by grants from the NSF (DEB-9224949) and the National Geographic Society (5189-94). Personal support was provided by a postgraduate scholarship from the Natural Sciences and Engineering Research Council (NSERC) of Canada and a Fellowship from Yale University. I thank James Archie and two anonymous reviewers for their comments on an earlier version of this paper.

REFERENCES

BULMER, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8:868–883.

BUTH, D. G. 1984. The application of electrophoretic data in systematic studies. Annu. Rev. Ecol. Syst. 15: 501–522.

CAVALLI-SFORZA, L. L., AND A.W. F. EDWARDS. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 21:550–570.

CAVALLI-SFORZA, L. L., AND A. PIAZZA. 1975. Analysis of evolution: Evolutionary rates, independence, and treeness. Theor. Popul. Biol. 8:127–165.

CHAKRABORTY, R. 1977. Estimation of time of divergence from phylogenetic studies. Can. J. Genet. Cytol. 19:217–223.

COCHRAN, W. G. 1977. Sampling techniques. Wiley, New York.

DEMING, W. E. 1950. Some theory of sampling. Wiley, New York.

FELSENSTEIN, J. 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 155:279–284.

LANGLEY, C. H., AND W. M. FITCH. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

LEVINS, R. 1969. Some demographic consequences of environmental heterogeneity for biological control. Bull. Entomol. Soc. Am. 15:237–240.

LI, C. C. 1976. First course in population genetics. Boxwood Press, Pacific Grove, California.

LI, C. C., AND D. G. HORVITZ. 1953. Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. 5:107–117.

LYNCH, M., AND T. J. CREASE. 1990. The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. 7:377–394.

MABEE, P. M., AND J. HUMPHRIES. 1993. Coding polymorphic data: Examples from allozymes and ontogeny. Syst. Biol. 42:166–181.

MICKEVICH, M. F., AND C. MITTER. 1981. Treating polymorphic characters in systematics: A phylogenetic treatment of electrophoretic data. Pages 45–58 *in* Advances in cladistics, Volume 1 (V. A. Funk and D. R. Brooks, eds.). New York Botanical Garden, New York.

MICKEVICH, M. F., AND C. MITTER. 1983. Evolutionary patterns in allozyme data: A systematic approach. Pages 169–176 *in* Advances in cladistics, Volume 2 (N. I. Platnick and V. A. Funk, eds.). Columbia Univ. Press, New York.

NEI, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

NEI, M., AND A. K. ROYCHOUDHURY. 1974. Sampling variances of heterozygosity and genetic distance. Genetics 76:379–390.

NEI, M., J. C. STEPHENS, AND N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. 2:66–85.

PATTON, J. C., AND J. C. AVISE. 1983. An empirical evaluation of qualitative Hennigian analysis of protein electrophoretic data. J. Mol. Evol. 19:244–254.

RICE, J. A. 1988. Mathematical statistics and data analysis. Wadsworth and Brooks/Cole, Pacific Grove, California.

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

SWOFFORD, D. L., AND S. H. BERLOCHER. 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. Syst. Zool. 36:293–325.

TAJIMA, F. 1992. Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. Mol. Biol. Evol. 9:168–181.

Takahata, N., and F. Tajima. 1991. Sampling errors in phylogeny. Mol. Biol. Evol. 8:494–502.

Weir, B. S., and C. J. Basten. 1990. Sampling strategies for distances between DNA sequences. Biometrics 46:551–582.

Wright, S. 1969. Evolution and the genetics of populations, Volume II. The theory of gene frequencies. Univ. Chicago Press, Chicago.

Received 28 February 1994; accepted 23 January 1995

Appendix

The variance of the ratio estimator of average allele frequency given by Equation 7 may be obtained from equation 11.30 of Cochran (1977:305) as follows. With an appropriate change of notation, the variance of an estimate of the total number of alleles of a type with average frequency p_T in the collection of populations as a whole (where $Y = N_o p_T$) may be written as

$$\mathrm{Var}(\hat{Y}) = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)\sum_{i=1}^{m}\frac{N_i^2(\hat{p}_i - \hat{p}_T)^2}{m - 1} + \frac{M}{m}\sum_{i=1}^{m}N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}. \qquad \text{(A1)}$$

To obtain $\mathrm{Var}(\hat{p}_T)$ from Equation A1, we simply divide both sides by N_o^2. An estimate of N_o is given by

$$\hat{N}_o = \frac{M}{m}\sum_{i=1}^{m}N_i. \qquad \text{(A2)}$$

Dividing Equation A1 by the squared inverse of Equation A2 and simplifying gives

$$\mathrm{Var}(\hat{p}_T) = \left(\frac{m}{m - 1}\right)\left(1 - \frac{m}{M}\right)\sum_{i=1}^{m}\varphi_i^2(\hat{p}_i - \hat{p}_T)^2 + \frac{m}{M}\sum_{i=1}^{m}\varphi_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}, \qquad \text{(A3)}$$

where $\varphi_i = N_i/\Sigma_{i=1}^{m} N_i$. If $m \gg 1$, Equation A3 simplifies to the approximate result:

$$\mathrm{Var}(\hat{p}_T) \cong \left(1 - \frac{m}{M}\right)\sum_{i=1}^{m}\varphi_i^2(\hat{p}_i - \hat{p}_T)^2 + \frac{m}{M}\sum_{i=1}^{m}\varphi_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}. \qquad \text{(A4)}$$

Syst. Biol. 44(3):429–434, 1995

Sponges, Plants, and T-PTP

John W. H. Trueman

CSIRO Division of Entomology, P.O. Box 1700, Canberra, ACT 2601, Australia[1]

In the 19th century, sponges sometimes were classified as plants rather than animals, but today there no longer is any doubt that their true relationships lie with the animal kingdom. The question of whether green plants are the sister group to animals or whether one or more fungal and/or protist lineages occupy intervening branches in the eukaryote phylogenetic tree is less clear cut. Wainwright et al. (1993) analyzed sequence data from the small subunit (18S) ribosomal RNA (rRNA) of a range of higher plants, metazoan animals, fungi, alveolates (ciliates, dinoflagellates, apicomplexans), and stramenopiles (chromophyte algae, heterokont protists) and reported a sister-group relationship between animals and fungi. More recently, Rodrigo et al. (1994) analyzed a 300-base sequence from the V4 region of 18S rRNA from two plants (*Arabidopsis* sp., *Glycine max*), three sponges (*Chondrosia* sp., *Dictyodendrilla* sp., *Fasciospongia* sp.), and four nonsponge metazoan animals (mollusc [*Placopecten magellanicus*], brine shrimp [*Artemia salina*], human [*Homo sapiens*], and sea anemone [*Anemonia sulcata*]), obtaining most-parsimonious trees in which the sponges and sea anemone apparently are more closely related to plants than to other animals. A fungus, the yeast *Saccharomyces cerevisiae*, was used to root the trees. Equal weighting of transitions and transversions gave the tree shown here as Figure 1. Weighting transversions

[1] E-mail: johnt@ento.csiro.au.