



Phylogenetic Reconstruction, a Critique

Richard H. Zander

Taxon, Vol. 47, No. 3 (Aug., 1998), 681-693.

Stable URL:

<http://links.jstor.org/sici?sici=0040-0262%28199808%2947%3A3%3C681%3APRAC%3E2.0.CO%3B2-9>

Taxon is currently published by International Association for Plant Taxonomy (IAPT).

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/iapt.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

POINTS OF VIEW

Edited by John McNeill

Phylogenetic reconstruction, a critiqueRichard H. Zander¹

The mathematical algorithms associated with phylogenetic analysis effectively eliminate grossly unreasonable trees, namely those that are too long or too improbable. This is done in part, for instance, with an initial minimum-spanning tree in parsimony analysis (Abbott & al., 1985). Problems occur, however, when the algorithms are executed maximally, and single trees (or small groups of trees) are presented as “reconstructions” of phylogeny and are offered as a basis for classification. Optimality criteria such as maximum parsimony, least falsifiability, maximum likelihood or maximum posterior probability are generally used in phylogenetic analysis as justifications for selecting a single tree (or a group of equally optimal trees). Although proffered as “approximations” or “converging on the true tree”, trees obtained with linear rank methods are almost always inappropriate as phylogenetic hypotheses in any but the most general sense.

A phylogenetic hypothesis resulting from a “discovery process” should, by implication, be based on more adequate evidence than simply being the one with the least evidence against it, but this is usually not the case. Harper (1979) suggested that the probability of two taxa sharing closer ancestry to one another than to others in the group, given that the two taxa share one or more uniquely derived advanced character states and the others do not, should at least exceed 0.5 for scientifically plausible models. This is a minimally acceptable definition of a “probabilistic hypothesis” in phylogenetic analysis.

The major problem shared by maximum likelihood and maximum parsimony techniques of phylogenetic analysis is that of “statistical relevance”. This is seldom referred to in the literature, since its application in phylogenetics is flawed. This concept, as introduced by Salmon (1971: 11), asserts that the most probable of several statistical hypotheses about a phenomenon has the higher explanatory power. Salmon demonstrated that requiring a probability to reflect more evidence for than against may not be the appropriate goal of predictive statistical analysis in some cases. His example is of a medical test showing an increase of the chance of a disease in an individual, rising from the chance of only one member of the general population having that disease to a higher probability (but less than 0.5). Even although there may not be more than a small chance of contracting the disease than not, this probabilistic increase is the accepted basis for, e.g., risk assessment in medical studies.

But Salmon (1971: 56) expanded the concept: “According to Hemple [1965], the basic requirement for an inductive explanation is that the posterior weight ... must be high, whereas I have been suggesting that the important characteristic is the increase

¹ Buffalo Museum of Science, 1020 Humboldt Parkway, Buffalo, NY 14211, U.S.A.

of the posterior weight over the prior weight as a result of incorporating the event into a homogeneous reference class... When the posterior weight of an event is low, it is tempting to think that we have not fully explained it ... [but] when the reference class is epistemically homogeneous in terms of our present knowledge, ... we have provided the most adequate explanation possible in view of the knowledge we possess." Likewise: "To explain an event is to provide the best possible grounds we could have had for making predictions concerning it. An explanation does not show that the event was to be expected; it shows what sorts of expectations would have been reasonable and under what circumstances it was to be expected... In some cases the explanation will show that the explanandum event was not to be expected, but that does not destroy the symmetry of explanation and prediction. The symmetry consists in the fact that the explanatory facts constitute the fullest possible basis for making a prediction of whether or not the event would occur..." (Salmon, 1971: 79).

As Salmon's criterion is used in phylogenetic analysis, second-best hypotheses (and third-best, etc.) are rejected because they involve somewhat less homogeneous reference classes. It is an argument for accepting, sometimes, an improbable tree as best hypothesis of phylogenetic relationships because that tree best explains the data (see also Farris, 1983; Hull, 1974). There is no apprehension for how badly it explains the data or for the relative quality of second-best, third-best, etc., explanations. There is similarly no concern for whether the best explanation is a good bet or not, even in situations where a good bet is necessary. Fischer (1970: 50), discussing the fallacy of the circular proof, pointed out that the best available proof of a historical event may not be truly convincing. A correct hypothesis may be an improbable one once it is known, but present optimality methods will select as a "reconstruction" one tree from among many reasonable but almost equally improbable hypotheses. The single most-adequate explanation can be scientifically inadequate as a hypothesis if it cannot support a confident bet, and when the posterior probability of that explanation is less than 0.5, only the features shared by all reasonable trees with probabilities summing to more than 0.5 have more evidence for than against.

Maximum likelihood analysis, which "chooses the value of a parameter that maximizes the probability of observing the data" (Huelsenbeck & Crandall, 1997), and other linear rank statistics are appropriate for situations in which loss through failure to identify increased risk is very great. Statistical relevance (as an optimality criterion) has been applied, however, to the results of evolutionary analyses as a justification for presenting the tree of maximum likelihood or maximum posterior probability, or the tree capable of least falsification, as the "best" phylogenetic hypothesis. This, however, simply substitutes a dubious attainable goal for one that is presently unattainable or rarely attainable, i.e., selecting the most adequate hypothesis from a number of nearly identical hypotheses instead of an entirely adequate hypothesis that can stand alone as a good bet in prediction. All trees with a significant increase in probability or decrease in falsifiability in light of information in a data set are possible reasonable phylogenetic hypotheses, and a more stringent criterion for a single hypothesis is necessary.

It was early "generally agreed that the reconstruction of evolutionary trees should ideally be regarded as a problem in statistical inference..." (Farris, 1973). According to Sober (1986: 33), in a sense "the parsimonious hypothesis is the hypothesis of maximum likelihood". All statistical methods deal with expectations. Analyses of relative frequency describe (within stated error margins) the long-run outcomes of

series of instances, but probabilities involved in a single event (as discussed by Bernardo & Smith, 1994: 478; Pfaffenberger & Patterson, 1987; Mises, 1957) are largely identified with Bayesian analysis techniques in which probability models frequency (Frank & Althoen, 1994). Bayesian analysis may yield the same probabilistic expectation of a phenomenon as classical statistics, but many extra assumptions are necessarily involved in modelling. For instance, in classical statistics, the frequency data for particular thrown dice may be used to form expectations about the results of future throws, but in Bayesian analysis expectations are based on assumptions that the dice are not loaded and are fairly cast (as prior probabilities). Phylogenetic analysis is, at least implicitly, Bayesian. A phylogenetic data set is a view of the phylogeny taken at one instant in time, and Bayesian generalisations and inferences are required for thorough analysis (Harper, 1979). Probabilistic methods are used to deal with single events because, according to Salmon (1971: 56), such events can be usually be referred to reference classes (see also Pap, 1962: 175, 216) of known initial probabilities.

Bayesian analysis estimates probabilities for single events through regularity assumptions. Such assumptions may eventually be proved correct or frequencies may be shown to even out in the long run. But before actual frequency data are known, Bayesian analysis remains the best method of dealing with incomplete information. On the other hand, a phylogenetic “reconstruction” should not be based on belief-oriented Bayesian analysis when the bet is at long odds, especially when one would build upon the results, which compounds the effects of being wrong.

A Bayesian bet selecting hypotheses that best fit data is shown here in a simple example with two dice: a four-sided regular die (tetrahedron) and an ordinary six-sided die. These, hidden from you and selected randomly, are thrown one at a time until a “1” turns up (one pip up – or down in the case of the tetrahedron). You then must guess which hidden die was used to generate the data set “1”. The prior probability of getting a 1 with a four-sided die is $1/4$, but from a six-sided die it is $1/6$: this reflects regularity assumptions that the dice are not loaded and are fairly cast. The likelihood is proportional (in this case 1:1) to the initial probabilities and thus you conclude that the four-sided die has the maximum likelihood of being the die that was used to generate the data set (in this case of a single datum). This is the best explanation and, if you must bet, the best bet. By Bayes’s Theorem (Harper, 1979; Winkler, 1972), the posterior probability that one’s tetrahedric die hypothesis is correct is 0.6 (while that of a cubic die is 0.4, both adding to “probability 1”). This gives you a better chance than the 0.5 (random) expectation you had before knowledge of the additional information (data set “1”).

In another example, consider a series of 10 bean bags sewn of increasing numbers of segments, the first bag of 1 segment, the second of 2, through 20, the segments equal in size and numbered sequentially on each bag. The bags are like different regular-solid dice but with unit incremental sides. One bag, hidden from you, is selected randomly and is cast. You are informed that “4” is the number of the side that turned up. Eliminating the first three bags because they had fewer than four segments, you identify the bag with only four segments as having maximum likelihood of being the one that was cast, since all six other bags have more segments. The posterior probabilities, however, by Bayes’s Theorem, in order of bags with increasing numbers of segments are 0.228, 0.183, 0.152, 0.130, 0.114, 0.101, and 0.091. You must be able to select all three most likely bags (the ones with the least

segments) at once to have a better than even bet of picking the bag that generated the datum “4” (the posterior probabilities of these four bags sum to 0.563). A bet with confidence (greater than 0.95 chance) requires considering all bags. With 20 unit-incremental bags and a conditional datum “4”, a better than even bet requires selecting all four most likely bags (summing the posterior probabilities 0.146, 0.117, 0.098, 0.084), and a bet with confidence (0.95 or greater in summed probabilities) requires selecting all but two bags. These are simple examples with few alternative hypotheses, but the problem with correctly using optimality in maximum likelihood analysis of phylogeny is similar.

In a paper on Dollo cladistic analysis, Farris (1977) found that “the more parsimonious of two rooted trees differing by only one in total steps would be at least 4 times as probable as the other”. The probability rises to 16, 64 and 256 times for 2, 3 and 4 steps longer, respectively. One can assume that Farris is correct that “preferring a tree with 4 fewer total steps than an alternative tree for the same data is statistically better justified than preferring an alternative to a null hypothesis when the latter can be rejected at $\alpha = 0.001$ ”. There are, however, commonly many more additional possible longer trees at each of 1, 2, 3, 4 or more steps than the shortest tree. The sum of the probabilities of these many trees (if Farris’s probability assignments are acceptable) is generally far greater than that of the shortest tree. Earlier, Rogers & al. (1967) pointed out this same problem, to which Kluge & Farris (1969) responded, inadequately, that convergence is shown in cladistic homoplasy thus “demonstrating that evolution is not parsimonious”. According to Fischer (1970: 53) “valid empirical proof requires not merely the establishment of possibility, but an estimate of probability. Moreover, it demands a balanced estimate of probabilities pro and con.” A consistency index even as high as 0.85 (sometimes attained in molecular cladistic analyses) implies that there is still considerable cladistic homoplasy, which implies a proportional amount of evolutionary convergence “hidden” among the synapomorphies of the shortest tree (least state changes). Patristically distant convergence is identifiable as cladistic homoplasy, but patristically close convergence, in morphological or gene data, is lost among trees up to a few steps longer than the shortest tree.

Schemes for estimating confidence sets for cladistics studies have been suggested by, e.g., Sanderson (1989) using bootstrap replicates, and Faith (1991) using random character correlation, but these refer to the shortest tree as methodologically central. According to Swofford & Maddison (1992): “One way to minimize the impact of incorrect assumptions regarding the phylogeny when examining hypotheses of character evolution is to reconstruct the character(s) on a variety of reasonable trees, ideally a large enough set of trees that the probability of including the true tree is relatively high.” Felsenstein (1985) cautioned, in the case of multiple equally shortest trees, against the assumption of good support for those subclades appearing identically in all trees: “the confidence interval on phylogenies appears to be much larger than the set of all most parsimonious trees”. Bremer (1988) decided, as to protein sequence data, that “Not only the shortest cladograms, but also those with an increasing number of steps should be combined into strict-consensus trees... Only those groups present in the consensus trees may be hypothesized to be monophyletic with any confidence. There is no easy way to determine how many extra steps should be allowed.”

Maximum likelihood estimation (reviewed recently by Huelsenbeck & Crandall, 1997) is considered by some to be superior to parsimony methods. Yang (1997), a

statistical phylogeneticist, referred to maximum parsimony as using “intuitive clustering algorithms for phylogeny reconstruction, which lack a rigorous statistical basis... Under quite general regularity conditions, maximum-likelihood estimators have desirable large-sample properties: they are consistent, asymptotically unbiased, and most efficient.” H. E. Ballard, Jr. (pers. comm.) found that maximum likelihood analysis of ITS molecular data may generate more parsimonious explanations of evolution than do parsimony methods in that for the oceanic island groups that he studied fewer biogeographic dispersal events or ecological shifts are required. Kluge (1997), on the other hand, argued that likelihood techniques, as verificationist methodologies, are opposed to Popperian falsificationism, and cladists are “not preoccupied with knowing the absolute truth, unlike verificationists”. Siddall & Wenzel (1997) admonished phylogeneticists “to abandon neojustificationist statistical interpretations”. In the present paper, I argue that both methods can produce unacceptable results.

A phylogenetic tree may be viewed stochastically as a branching Markov chain (Sanderson, 1993) of conditional probabilities. Each evolutionary event depends only on the event immediately preceding. The abstruse mathematics of maximum likelihood analysis of molecular data in the literature is a result of having to deal with many parameters. Variables are treated as random and continuous. Maximum likelihood is the point on the curve of probability density where the slope of a tangent is zero, i.e., the top of the curve, using the infinitesimal calculus. For reasons of computational speed and memory, maximum likelihood analysis was limited to small data sets until recently. Log likelihoods are used as measures, in part, because they are easily distinguished to the left of the decimal point, while actual likelihood values are often very small decimal fractions.

Because gene mutations are readily dealt with as stochastic events, maximum likelihood is presently a much used method for molecular phylogenetic analysis. The rate of evolutionary change of morphological characters is difficult to estimate (Martins, 1994). Likelihood analysis simulates DNA sequences by calculating probabilities of oligonucleotides by correlation between base frequencies in various positions of the sequence (Bralley, 1996). Markov chains tend toward a steady state or an equilibrium (Rolf & Williams, 1991), which allows analysis of long-term trends. A maximum likelihood Markov chain Monte Carlo method (parametric bootstrapping of a martingale, Bernardo & Smith, 1994: 353; Goldman, 1990; Ross, 1997: 211; Rolf & Williams, 1991) was used by Mau & al. (1997) for nine species of *Clarkia* Pursh (*Onagraceae*) with cpDNA restriction-site data. The tree of maximum likelihood was chosen as the best phylogenetic hypothesis. Marginalized posterior probabilities of the five most likely phylogenies were reported (the tree of maximum likelihood as usual has highest posterior probability) as 0.649, 0.179, 0.168, 0.002 and 0.001. The probabilities of the three most likely trees added to a posterior probability of 0.996, resulting in a pool of three trees in this “credible region”, which is the Bayesian equivalent of a confidence interval (Pfaffenberger & Patterson, 1987: 1116). The posterior probability of 0.649 means in classical frequentist terms that if this exact, same data set were to occur many times, the tree of maximum likelihood will also be the true tree in about 13 out of every 20 duplications. Rannala & Yang (1996) used a similar method with primate pseudo-genes to obtain a (((human, chimpanzee), gorilla), orang-utan) tree with posterior probability of 0.84. This same paper reported a study of 11 mitochondrial tRNA genes of primates giving the (((common chimpanzee, pygmy chimpanzee), human) gorilla) orang-utan) tree a

posterior probability close to one (0.9999). A study of mitochondrial genome segments, being “parts of two protein-coding genes and three tRNA genes”, by Yang & Rannala (1997) found a high posterior probability for nine primates: ((((((human, chimpanzee) gorilla) orang-utan) gibbon) crab-eating macaque) squirrel monkey)(tarsier, lemur) of between 0.95 to 0.96 depending on the evolutionary model used, all models obtaining the same tree.

The curves of probability density in the above studies are high-peaked. These particular probabilistic bets are, however, for small data sets, and assume that no other data (sharing significant numbers of advanced characters) applies. Yang & Rannala (1997) asserted that posterior probabilities did not change much among different analytical variations, and their method “is robust to variations in the prior” because “most information concerning phylogeny derives from the data”. In any case, the analyses appear to have successfully established well-supported gene tree hypotheses for the primate data sets, given the neutralist expectation toward gene selection (Avice, 1994), certain non-informative (ignorance) priors (non-objective according to Bernardo & Smith, 1994: 357), the possible problems with treating character states as independent (see example of Huelsenbeck & Crandall, 1997: 447) and uniformly distributed random variables in these particular genes, the amount of sampling for intraspecific variation in traits (Doyle, 1992), optimality criteria for sequence alignment, and a host of additional regularity assumptions for the Bayesian-style analysis (Zander, 1998).

In another study cited in the Mau & al. (1997) paper above, mitochondrial DNA for 31 species of African cichlid fish (plus an outgroup) were analysed at 1044 aligned sites. The posterior probabilities of the five most likely phylogenies were 0.11, 0.07, 0.06, 0.04 and 0.03. Here the curve of probability density is much flattened. With this larger data set, the chance is about 1 in 10 that the most likely tree is the true tree. This is about the same probability of guessing correctly that the tetrahedric die generated the data set “1” when one tetrahedric die and an additional 14 cubic dice are cast randomly until a “1” appears. The very strong Bayesian bet is that the most likely tree is not the true tree, and one of the cubic dice (statistics does not tell us which one) most probably generated the “1” because the number of dice involved outweighed the effect of their likelihoods. Though the best theory available, this is not the kind of probabilistic estimation from which, say, outgroups should be selected for analyses of taxa higher in the tree of life, unless the pertinent subclades in the several most likely clades are identical and their trees add to a strong posterior probability (not the case for most subclades in this study).

Although Mau & al. (1997) eliminated in their cichlid study most of c. 10^{40} trees to get the 250 trees that comprised their “95 percent credible region” (the most likely trees with posterior probabilities adding to 95 %), one would have to somehow eliminate all but the two most likely trees to make the one most likely tree a good bet to be the true tree. It is this “somehow” that is the proper focus of new research in phylogenetic methodology. Taking a mathematical elimination process to the limit may be wrong if the justification for eliminating the majority of possibilities (phylogenetically grossly long trees or improbable trees) is different from the justification for eliminating all but one from the pool of credible trees. Thus, a corollary to Occam’s Razor is that explanations must remain multiple when no one of them is probabilistically adequate. In maximum likelihood studies a single improbable tree is often presented, possibly to match the apparent success of parsimony studies. A

strict or majority-rule consensus tree of how trees in the 95 % credible region actually agree would seem to be the better interpretive result in likelihood analyses.

A comparison of a consensus gene tree from the 0.5 and 0.95 credibility regions obtained with statistical methods with the shortest gene tree (from the same data set) might throw light on what the word “approaches” means when cladists assert that a particular result approaches, approximates or converges on the true tree, and on what Bremer support signifies statistically for molecular data. Hendry’s (1996) account of “convergence realism” includes a caution that the expectation of science converging on the truth is only applicable to mature sciences; this is hardly a description of phylogenetic systematics. According to Hendry: “Approximate truth is a difficult concept. On any reasonable construal, approximate truth does not explain predictive success.” As for likelihood analysis, the consistency argument, that the method ensures that the tree of maximum likelihood must converge probabilistically to 1 as data increase to infinity (Shenton & Bowman, 1977), is commonly offered as a reason to accept low posterior probabilities, but this is merely intuitive and liable to sampling errors according to Yang (1994), and was flatly denied by Sober (1983, 1986) as impractical. Also, the numbers of species that may supply additional data are limited (Sanderson, 1993). According to Yang & Rannala (1997), “A method assuming a wrong model may still be consistent and may have smaller sampling errors than one using the right model.” Bernardo & Smith (1994: 456, 480) discuss relevant problems with likelihood approaches and approximate methods based on asymptotic theory. Huelsenbeck & Crandall (1997) stated that “the maximum likelihood method can be robust to a variety of model violations”, but this is a judgment based largely on linear rank studies and comparison of alternate methodologies. Their positive examination of the “power of the likelihood-ratio test” (of rejecting the null hypothesis, described by Bernardo & Smith, 1994: 487) rests on simple data simulation and null rejection at 0.05, and is dubious because actual molecular data from different genes commonly conflict (see below). In their description of the likelihood-ratio test statistic of whether a parameter provides a significant improvement (read statistical relevance) in the likelihood, only the null hypothesis is rejected, a carryover from classical statistical methods; in Bayesian statistics, the probabilities of both null and alternate hypotheses must be examined (Pfaffenberger & Patterson, 1987) and more than one alternative hypothesis should be considered (J. Neyman & E. Pearson *fide* Bower, 1997). It is not clear from Huelsenbeck & Crandall’s treatment exactly what conditions of wrong assumptions and poor data (if any) will cause maximum likelihood analysis of phylogenetic relationships to actually fail (but see Schöniger & Haeseler, 1995 on inconsistency). Improper use of null hypothesis testing has recently been flagged in the psychological sciences where “... more theoretical courage” is called for (G. Gigerenzer in Bower, 1997).

At least with larger phylogenetic data sets the most probable tree is not “probably the true tree”. Few papers to date using maximum likelihood discuss relevant posterior probabilities: Bohs & Olmstead (1997) do not mention them, Huelsenbeck & Rannala (1997) promoted the use of likelihood-ratio tests (e.g. as used by Yang, 1996) instead and did not mention posterior probabilities. I agree with Rannala & Yang (1996) that “the posterior probability provides a natural measure of the reliability of the estimated phylogeny” given the various assumptions required for it to be calculated, but it must be used relative to the sum of the reasonable alternative probabilities.

A second problem is that in parsimony studies, convergence between very closely related ancestors or terminal taxa is necessarily interpreted as synapomorphy in the shortest tree. Such forced algorithmic interpretation may result in a shorter tree than the true tree. Because the consistency index is commonly at best 0.85 in phylogenetic analyses and in most cases much lower, the shortest tree is seldom probabilistically the tree of maximum parsimony.

With artificially generated phylogenies, Heijerman (1997) found certain clustering methods better at retrieving “true” trees than parsimony methods when homoplasy is relatively great. Neither clustering nor parsimony, however, achieved better than 74 % similarity with the true tree. In another study (Heijerman, 1990), parsimony methods were found to be more accurate than clustering methods when the consistency index is above 0.8. Significantly, however, parsimony methods found shorter trees than true trees, misinterpreting some “true” convergence as ancestrally shared states.

Maximum parsimony analyses are parsimonious in eliminating myriads of unreasonable trees of overly complex hypotheses. This leaves a pool of tens or hundreds of trees (similar to the Bayesian credible interval above) that are reasonable under Darwinian theory, thus the phrase “maximum parsimony” as used in the literature is a semantic distortion. The explanation of descent from common stock applies to all plausible trees, whether common descent is maximised or not. Maximum parsimony methods interpret convergence when possible as due to shared ancestors, something not required by Darwinian theory. This problem, termed “apparent synapomorphy”, has been previously pointed out by Lyons-Weiler & al. (1996). Though the shortest tree may not necessarily be the tree of maximum congruence or minimum homoplasy (Scotland, 1997), clearly the idea of optimality is used to discount other interpretations.

Contra Farris (1983) and Kluge (1997), eliminating all trees longer than the shortest from the pool of credible or reasonable trees includes unjustifiable ad hoc assumptions about the degree of joint ancestry of terminal taxa. When the number of assumptions against convergence is needlessly increased when searching for a shortest tree, the method becomes antiparsimonious and overly interpretive. A shortest tree with all covariance in the data set interpreted as ancestrally based when possible is not a phylogenetic reconstruction through parsimony. The shortest tree may be used for classification if it is understood that at least the fine structure (close relationships easily reinterpreted through minor convergence) is artificial and reflects a too-simple theory of evolution. Those who insist on the original ad hoc argument should consider the second, third, fourth, etc., least falsifiable trees, which compete as reasonable hypotheses in quantity if not in quality. It would be surprising if cladistic analysis of molecular data resulted in a smaller pool of justifiably reasonable trees than likelihood analysis could produce.

Molecular data may conflict as between sources, or there may be conflict between molecular data and morphological data (Avise, 1994: 314; Philippe & al., 1996; Seberg & al., 1997). According to Sites & al., (1996), “...diverse data sets do not always yield the same estimates of phylogeny for the same organisms”. Milinkovitch & al. (1996) found that “different phylogenetic analyses of the same genetic data set can yield conflicting results, depending on the choice of parameter settings and included taxa”, and resorted to a sensitivity analysis to identify “portions of the multidimensional parameter space where phylogenetic signal is most reliably recov-

ered". Naylor & Brown (1997) found a poor match between a bootstrap consensus parsimony tree based on "the entire protein-coding portion (12,234 base pairs) of the mitochondrial genome of 19 taxa whose interrelationships are widely accepted..." with the taxonomically accepted species tree (based on morphology). They used the retention index to find phylogenetically reliable functional classes of sites, though they recognised that using the expected tree to discover resilient sites was not an independent test. They pointed out that molecular data may have a similar co-variation due to both shared history and functional requirements as do morphological characters, resulting in a need to ascertain the "relative importance of particular co-varying combinations of residues for protein structure, function and folding".

According to Doyle (1992), because a gene tree may be uncoupled from a species tree by introgression, lineage sorting, or mistaken orthology, molecular systematics has many of the faults of one-character taxonomy (unless many genes are tested in a cladogram) and molecular analysis alone is not a better alternative to morphological analysis (this is contested by Olmstead & Palmer, 1997 for relatively distant relationships they studied in *Solanaceae*). Doyle goes on "...additional data for any particular gene, while it may produce a better gene tree, cannot increase confidence in that gene as representative of the species phylogeny". For this reason, recent studies have begun to analyse several genes (e.g. Nei & Takezaki, 1996). Also, Avise (1994: 314 ff.) reviewed cases of successful application of molecular techniques to the clarification of difficult systematic problems, often with congruent results from more than one genic element. Avise (1994: 354) recommended multiple lines of evidence as important in addressing such problems as "shared retention of ancestral states by the taxa in question, extreme molecular rate heterogeneities across lineages, convergent evolution to a shared molecular condition, introgressive hybridization, and a mistaken assumption of orthology when the loci in question might truly be paralogous...", and he reviewed evidence for at least occasional horizontal transmission of particular genes, which may be mediated by parasites.

Because of the nature of the strongly belief-oriented probabilistic analysis used in phylogenetics and the poor or misleading results in practice, the degree of assurance that frequency-based long-run statistical analyses give to other scientific studies may never be attained. The shortest or most likely tree has a possible pragmatic (Pap, 1962: 228) value in that, if one must choose from among many hypotheses, even through a less-than-probabilistic reductionism, the perceived risk is then lowest. In many sciences, one can immediately test the correctness of the most-likely hypothesis, or each of the several most likely. The prospect of successful *post hoc* testing for correctness may be the psychological justification for the otherwise illogical (Wittgenstein, 1961: 70) idea of simplicity (Sober, 1975) or Occam's Razor (Jefferys & Berger, 1992). This is not so in systematics, where immediate, clear-cut tests of correctness are unavailable, especially if "total evidence" (Carnap, 1962; Kluge, 1989) is used. The best hypothesis as "least wrong" in parsimony or maximum likelihood analysis is not a scientifically acceptable result. This reminds one of the casino gambler, who, when asked how luck was holding out, replied: "Fine! I have not won in two hours, but my friend here has not won in four hours."

It is possible that molecular systematics can provide data for evolutionary trees of high probability (as at least good Bayesian bets), but there is as yet successful demonstration for only a few small data sets and the problems involved with prior assumptions are immense. Further advancement in evolutionary analysis must include

addressing prior assumptions (e.g. Felsenstein & Churchill, 1996; Philippe & al., 1996) and also finding ways of distinguishing synapomorphies in the shortest tree that are due to shared ancestry from those due to evolutionary convergence. An ingenious method of distinguishing at least some apparent synapomorphy from evolutionary synapomorphy was proposed by Lyons-Weiler & al. (1996), based on identifying fidelity of phylogenetic signal by “how much unique similarity exists between two taxa with no redundant information added”.

In my opinion, a probabilistic estimate of species phylogeny (including relationships of very similar taxa) should (at least): (1) use a model incorporating variable evolutionary rates if possible; (2) with a necessarily Bayesian statistical analysis of data sets of several selectively neutral, independent genes (total evidence is better) resulting in a multiple-gene tree; (3) which should have a posterior probability greater than 0.5 (greater than 0.95 is better if the results are to be used as a basis for analysing concatenated trees higher in the tree of life); and (4) which is congruent to a species tree from a pool of short reasonable trees derived from a morphological data set that passes at least a fidelity of phylogenetic signal test.

The resultant molecularly supported species tree may even turn out to be the same as the so-called tree of maximum parsimony. A pool of candidate species trees would be developed through a parsimony analysis of non-gene characters with “accepted” relationships constrained (as per Milinkovitch & al., 1996) and then the shortest tree and all trees one step longer (at least) retained. This produces a set of trees with grossly unreasonable trees eliminated. It remains a problem that the only test of phylogenetic hypotheses is congruence of the model with information about the past obtained from other sources, since the essentially Bayesian bet on a single chained past evolutionary phenomenon can have no direct corroboration. Probabilistic estimation of the branching pattern relationships of similar, closely related taxa may not be possible for many or most groups in which character states are fairly simple and appear many times in different groups. The degree to which phylogenetic analysis allows scientific predictions may be testable to some extent in biogeographic study, but this may depend more on the elimination of grossly unreasonable trees than identification of one optimum tree.

That much is expected from computerised phylogenetic analysis is readily demonstrable by the many university positions presently being advertised for molecular phylogenetic systematists, and from the amount of U.S. National Science Foundation (NSF) grant support. As to the latter, for the year 1997 (Anonymous, 1997), more than \$14 million was awarded in 96 grants for systematic research. Of these, 73 awards had the words “phylogeny”, “cladistic”, “molecular systematics”, or “evolution” (evolution in systematics studies being almost certainly used in the sense of phylogenetics) or some variant of these in their title. These grants totalled about \$9 million. The 23 remaining systematics research grants (23 %) totalled about \$5 million; for these, any emphasis on phylogenetics could not be told from their titles. Thus, support for modern computerised evolutionary analysis by NSF in just the one year 1997 is conservatively estimated at about \$9 million (of a \$14 million pot), taking about 75 % of the awards.

There are two relevant kinds of phylogenetic studies: (1) those that offer “reconstructions” by methodologically wrongly falling back on optimality criteria of, e.g., maximum parsimony, minimum falsifiability, or maximum likelihood and maximum posterior probabilities when posterior probability greater than 0.5 cannot be demon-

strated, and (2) those that treat their results not as optimisations but as contributions towards building a modern method for probabilistically estimating phylogenetic relationships for at least some groups and otherwise using “best” explanations simply to guide classification. No new probabilistic scientific conclusions can be derived from projects in the first category, other than the generation of data sets, and so the funding support is questionable. In view of the well-known world-wide critical status of biological diversity, one can only hope that alpha taxonomic studies (keys, descriptions, nomenclature, typification, discussion of range and variation, illustration, etc.) will be increasingly part of morphologically based phylogenetic projects, as is sometimes now the case, and that these are supported by NSF.

If we must base classifications on explanations of single past events that are merely the best of a number of competing explanations requiring in addition a host of regularity assumptions, then this is the sorry burden of systematics that has not been alleviated to any significant extent by modern computerised evolutionary analysis.

Acknowledgements

I thank T. DiBenedetto, P. M. Eckel, W. K. Gall, J. Lyons-Weiler, C. A. Maynard, J. McNeill, G. S. Mogensen, K. P. Smith, T. Schultz, and T. Stuessy for suggestions regarding particular portions of this work or for making helpful observations about some of the ideas in this paper.

*Literature cited (including *electronic publications)*

- Abbott, L. A., Bisby, F. A. & Rogers, D. J. 1985. *Taxonomic analysis in biology: computers, models, and databases*. New York.
- *Anonymous [National Science Foundation] 1997. *Award list for systematics [1997]*. [<http://www.nsf.gov>]. Washington, D.C.
- Avise, J. C. 1994. *Molecular markers, natural history and evolution*. New York.
- Bernardo, J. M. & Smith, A. F. M. 1994. *Bayesian theory*. New York.
- Bohs, L. & Olmstead, R. G. 1997. Phylogenetic relationships in *Solanum* (*Solanaceae*) based on *ndhF* sequences. *Syst. Bot.* 22: 5-12.
- Bower, B. 1997. Null science: psychology's statistical status quo draws fire. *Science News* 151: 356-357.
- Bralley, P. 1996. An introduction to molecular linguistics. *BioScience* 46: 146-153.
- Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42: 796-803.
- Carnap, R. 1962. *Logical foundations of probability*, ed. 2. Chicago.
- Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* 17: 144-163.
- Faith, D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* 40: 366-375.
- Farris, J. S. 1973. A probability model for inferring evolutionary trees. *Syst. Zool.* 22: 250-256.
- 1977. Phylogenetic analysis under Dollo's Law. *Syst. Zool.* 26: 77-88.
- 1983. The logical basis of phylogenetic analysis. Pp. 7-36 in: Platnick, N. I. & Funk, V. A. (ed.), *Advances in cladistics*, 2. New York.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.
- & Churchill, G. A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Molec. Biol. Evol.* 13: 93-104.
- Fischer, D. H. 1970. *Historian's fallacies: toward a logic of historical thought*. New York.

- Frank, H. & Althoen, S. C. 1994. *Statistics: concepts and applications*. Cambridge.
- Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson Process Model of DNA substitution and to parsimony analyses. *Syst. Zool.* 39: 345-361.
- Harper, C. W. 1979. A Bayesian probability view of phylogenetic systematics. *Syst. Zool.* 28: 547-553.
- Heijerman, T. 1990. GENESIS: a simulation model of phylogeny. 2. A comparative study based on simulation experiments. *Z. Zool. Syst. Evolutionsforsch.* 28: 81-93.
- 1997. Adequacy of numerical taxonomic methods: why not be a pheneticist? *Syst. Biol.* 21: 309-319.
- Hempel, C. G. 1965. *Aspects of scientific explanation*. New York.
- *Hendry, R. 1996. *Realism*. Nov. 27, 1996. [<http://www.dur.ac.uk/~df10www/modules/philsoci/PHILSCI.htm>]. Durham, U.K.
- Huelsenbeck, J. P. & Crandall, K. A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Ann. Rev. Ecol. Syst.* 28: 437-466.
- & Rannala, B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276: 227-232.
- Hull, D. L. 1974. *Philosophy of biological science*. Englewood Cliffs, NJ.
- Jefferys, W. H. & Berger, J. O. 1992. Ockham's razor and Bayesian analysis. *Amer. Sci.* 80: 64-72.
- Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (*Boidae*, *Serpentes*). *Syst. Zool.* 38: 7-25.
- 1997. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics* 13: 81-96.
- & Farris, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18: 1-32.
- Lyons-Weiler, J., Hoelzer, G. A. & Tausch, R. J. 1996. Relative apparent synapomorphy analysis (RSA) I: The statistical measurement of phylogenetic signal. *Molec. Biol. Evol.* 13: 749-757.
- Martins, E. P. 1994. Estimating rates of character change from comparative data. *Amer. Naturalist* 144: 193-209.
- Mau, B., Newton, M. A. & Larget, B. 1997. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Molec. Biol. Evol.* 14: 717-724.
- Milinkovitch, M. C., LeDuc, R. G., Adachi, J., Farnir, F., Georges, M. & Hasegawa, M. 1996. Effects of character weighting and species sampling on phylogeny reconstruction: a case study based on DNA sequence data in cetaceans. *Genetics* 144: 1817-1833.
- Mises, R. von, 1957. *Probability, statistics and truth*, ed. 2 [English ed. transl. H. Geiringer. Dover Edition.] New York.
- Naylor, G. J. P. & Brown, W. M. 1997. Structural biology and phylogenetic estimation. *Nature* 388: 527-528.
- Nei, M. & Takezaki, N. 1996. The root of the phylogenetic tree of human populations. *Molec. Biol. Evol.* 13: 170-177.
- Olmstead, R. G. & Palmer, J. D. 1997. Implications for the phylogeny, classification, and biogeography of *Solanum* from cpDNA restriction site variation. *Syst. Bot.* 22: 19-29.
- Pap, A. 1962. *An introduction to the philosophy of science*. New York
- Pfaffenberger, R. C. & Patterson, J. H. 1987. *Statistical methods for business and economics*, ed. 3. Homewood, IL.
- Philippe, H., Lecointre, G., Hoc Lanh Vân Lê & Le Guyader, H. 1996. A critical study of homoplasy in molecular data with the use of a morphologically based cladogram, and its consequences for character weighting. *Molec. Biol. Evol.* 13: 1174-1186.
- Rannala, B. & Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Molec. Evol.* 43: 304-311.

- Rogers, D. J., Fleming, H. S. & Estabrook, G. 1967. Use of computers in studies of taxonomy and evolution. Pp. 169-196 in: Dobzhansky, T., Hecht, M. K. & Steere, W. C. (ed.), *Evolutionary biology*, 1. New York.
- Rolf, H. L. & Williams, G. 1991. *Finite mathematics*, ed. 2. Dubuque, IA.
- Ross, S. M. 1997. *Introduction to probability models*, ed. 6. San Diego.
- Salmon, W. C. 1971. *Statistical explanation and statistical relevance*. Pittsburgh, PA.
- Sanderson, M. J. 1989. Confidence limits on phylogenies: the bootstrap revisited. *Cladistics* 5: 113-129.
- 1993. Reversibility in evolution: a maximum likelihood approach to character gain/loss bias in phylogenies. *Evolution* 47: 236-252.
- Schöniger, M. & Haeseler, A. von, 1995. Performance of the maximum likelihood, neighbor joining and maximum parsimony methods when sequence sites are not independent. *Syst. Biol.* 44: 533-547.
- Scotland, R. W. 1997. Parsimony neither maximizes congruence nor minimizes incongruence or homoplasy. *Taxon* 46: 743-746.
- Seberg, O., Petersen, G. & Baden, C. 1997. Taxonomic incongruence – a case in point from plants. *Cladistics* 13: 180.
- Shenton, L. R. & Bowman, K. O. 1977. *Maximum likelihood estimation in small samples*. New York.
- Siddall, M. & Wenzel, J. 1997. Random cladistics: new frontiers in molecular phylogenetics or old hat? *Cladistics* 13: 180.
- Sites, J. W., Davis, S. K., Guerra, T., Iverson, J. B. & Snell, H. L. 1996. Character congruence and phylogenetic signal in molecular and morphological data sets: a case study in the living iguanas (*Squamata, Iguanidae*). *Molec. Biol. Evol.* 13: 1087-1105.
- Sober, E. 1975. *Simplicity*. Oxford.
- 1983. Parsimony methods in systematics. Pp. 37-47 in: Platnick, N. I. & V. A Funk (ed.), *Advances in cladistics*, 2: 37-47.
- 1986. Parsimony and character weighting. *Cladistics* 2: 28-42.
- Swofford, D. L. & Maddison, W. P. 1992. Parsimony, character-state reconstructions, and evolutionary inferences. Pp. 186-223 in: R. Mayden (ed.), *Systematics, historical ecology and North American freshwater fishes*. Stanford.
- Winkler, R. L. 1972. *An introduction to Bayesian inference and decision*. New York.
- Wittgenstein, L. 1961. *Tractatus logico-philosophicus*. London.
- Yang, Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.* 43: 329-342.
- 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Molec. Evol.* 42: 587-596.
- 1997. How often do wrong models produce better phylogenies? *Molec. Biol. Evol.* 14: 105-108.
- & Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molec. Biol. Evol.* 14: 717-724.
- Zander, R. H. 1998. A phylogrammatic evolutionary analysis of the moss genus *Didymodon* in North America North Of Mexico. *Bull. Buffalo Soc. Nat. Sci.* 36: 81-115.