Cladistics **13**, 313–336 (1997)
WWW http://www.apnet.com

# Probabilism and Phylogenetic Inference

Mark E. Siddall and Arnold G. Kluge

*Museum of Zoology, University of Michigan, Ann Arbor, Michigan, 48109, U.S.A.*[1]

Accepted 8 July 1997

The maximum likelihood approach to phylogenetics rests on frequency probability theory. This stands in stark contrast to the logical probability of corroboration-based cladistic parsimony. History is particular and cannot be described in terms of universal statements about abstract generalities, the task of the historical sciences being one of explanation, not prediction. Thus, frequency probability methods of estimation are inappropriate for making historical inferences. Maximum likelihood estimation procedures are deconstructed from numerous perspectives in spite of their supposed impressive technicalities. Charges of parsimony's inconsistency are rendered mute, because its justification lies elsewhere, yet maximum likelihood is still subject to Wald's dilemma if realism is of any interest. Although all epistemologies make assumptions, the models employed by maximum likelihood are problematic and deterministic, as opposed to the unproblematic background knowledge characteristic of cladistics. Apart from issues of logical and sampling dependencies, the requirements of frequency probability theory are non-trivial and the maximum likelihood estimation of phylogeny can neither escape, nor satisfy the tenets of calculus independence (e.g. i.i.d.) inherent in the multiplicative relations of the method. If phylogeneticists are to maintain a rational foundation for their epistemology, neo-justificationist appeals to some metaphysical truth must be abandoned in favour of the realism of sophisticated falsification. 

[1]Correspondence to: M. E. Siddall.
Email: msiddall@umich.edu, fax:(313) 663 4080.

## INTRODUCTION

We are concerned with probabilism in phylogenetics. Much of the current disagreement relating to phylogenetic methods reduces to the markedly different concepts of logical probability and frequency probability. The former probability serves as the refutationists' justification for cladistic parsimony (Farris, 1983; Kluge, 1997), whereas the latter underlies several verificationist approaches. By "logical probability" we admit that in common parlance "probability" has varied interpretations and meanings. We, as others (Popper, 1959, 1983; Lakatos, 1970), distinguish between the calculus of frequency probability typified by Bayes' theorem

$$p(\mathrm{h,e}) = p(\mathrm{e,h})p(\mathrm{h}) / p(\mathrm{e})$$

in which the first term, $p$(e,h), formally is the likelihood of the evidence (e) in light of the hypothesis (h), and that logical probability exemplified in Popper's (1983) degree of corroboration

$$C(\mathrm{h,e,b}) = [p(\mathrm{e,hb})\text{-}p(\mathrm{e,b})] / [p(\mathrm{e,hb}) - p(\mathrm{eh,b}) + p(\mathrm{e,b})]$$

in which h is the hypothesis in question, e is the evidence and b is background knowledge.

No-one disputes what the alternative hypotheses are in phylogenetics. That is, for $N$ taxa there are exactly $(2N\text{-}3)! / 2^{N-2}(N\text{-}2)!$ possible bifurcating cladograms, all

0748-3007/97/040315+23/$25.00/0/cl970046

of which are capable of explaining observed character state distributions. These trees, then, comprise part of the premise for any phylogenetic analysis irrespective of method. Verification uses what Popper (1983) called the "mistaken solution of the problem of induction" by seeking the "induced hypothesis" with the highest probability and in which a probability of 1.00 would be certainty. In contrast, falsification seeks the hypothesis that best survives the severity of test offered by the data, that is, the most corroborated hypothesis. The problem with the verificationist program is that it denies nothing. For example, consider the premise "if bitten by a spider and given antivenom within three hours, the probability of surviving is 0.86". Suppose Pablo is bitten and receives the antivenom. If Pablo survives, this cannot be attributed to the antivenom, because there is the prior possibility that he could have died even with the antivenom. In fact, the 0.86 schema would equally explain why Pablo died, why he nearly died or even why he survived. Verificationist approaches to phylogenetics, like maximum likelihood, suffer from this failure as well, because all trees are assigned a non-zero probability, and yet no more than one tree actually can be correct — thus the probabilities are not *explanatory*. Other fields of science, including medicine, have already acknowledged that using "statistical estimates ... is unavoidably arbitrary, will often be contested and will have differential effects" upon our conclusions relating to singular cases (Lynn et al., 1997:56).

We focus on the verificationist methodology of maximum likelihood in this paper, because of its obvious reliance on frequency probability and in light of its increasing popularity in phylogenetics. Our conclusions, however, also apply to other neo-justificationist forms of induction in phylogenetics. For example, taxonomic congruence relies on the frequency of finding a clade in common among different data sets (Kluge and Wolf, 1993), the comparative method relies on the frequency of homoplasy in the characterization of adaptation (Harvey and Pagel, 1991; Foster et al., 1996; Larson and Losos, 1996), there are the resampling and permutation procedures which are intended to assess confidence in phylogenetic hypotheses statistically (Felsenstein, 1985; Faith, 1991), allelic frequencies are used as historically heritable transformations (Swofford and Berlocher, 1987; Wiens, 1995), and there are a variety of weighting criteria which are determined by the frequency of some observed character state (Farris, 1969; Carpenter, 1988; Williams and Fitch, 1990; Goloboff, 1993; Knight and Mindell, 1993). Our criticisms of frequency probability derive largely from the historical context in which that form of probabilism is employed. We begin with a philosophical consideration of universals and particulars, because their distinction makes it clear why frequency probability methods cannot apply in historical inference, neither to the instances of sister species common ancestry or the instances of descent with modification that characterize those historical patterns.

Devotees of maximum likelihood methods exhibit little concern for philosophy, and even denigrate cladistic parsimony for its appeal to such foundations. For example, Huelsenbeck (1996:7) has remarked that you "throw in a lot of philosophical mumbo-jumbo and you have the [cladistic] parsimony method". To the contrary, we believe that the coherence and generality of cladistics, and its reliance on parsimony, is due in large part to a solid grounding in philosophy, and it is with these same issues that we begin to build our case against the use of frequency probability in inferring history.

## THE NATURE OF INDIVIDUALS

There remains considerable confusion in comparative biology concerning universals and particulars. A simple question-answer exchange between a probabilist and a historian illustrates how easy it is to conflate the two.

*Probabilist*: "What is the chance of life evolving on earth?"

*Historian*: "Chance? It simply did."

*Probabilist*: "What is the chance that life has evolved, or could evolve, elsewhere in the universe?"

*Historian*: "None."

*Probabilist*: "Don't we have a good idea of the physical and chemical conditions necessary for life on earth, the number of appropriate stars and M-class planets, and, from that, would you not agree that we can predict the likelihood of there being life elsewhere?"

*Historian*: "Certainly not. Of course the answer might have been yes, if I had understood your question to

mean a *kind* of life. Obviously, your question is metaphysical, as opposed to scientific."

The difference in perspective between the probabilist and the historian is more than mere semantics. Biological life is earth-bound through a historically singular continuum of common ancestry. Even if DNA and RNA arose independently somewhere else in the universe, it would not be "life", because it would be ontologically independent. Metaphysically one might find something in common between *life* and some other independent thing in the cosmos that *looks-like-life*, but it could not be life in a biological sense if it arose independently. Consider "wings". Although it is clear that "wings" permit "flying" — that birds have wings and fly, that bats have wings and fly, and that flies have wings and fly — "flying" does not confer ontological identity on those things we call "wings". There is no spatio-temporally universal "wing" that encompases them, because the respective origins of these various wings constitute independent events in time. *Per contra*, wing-of-robin and wing-of-stork possess this identity. Such identity is by virtue of common ancestry alone. The following review of universals and particulars provides a more detailed and formal explanation for why the probabilist cannot make meaningful predictions about unique historical events, and why the phylogeneticist must embrace a different form of probabilism as the basis for *explanation*, i.e. historical inference with logical probability, not with frequency probability.

Universals and particulars play different roles in science (for additional examples see Frost and Kluge, 1994). Universals (classes, sets, generalities) are spatially unrestricted, and usually temporally unrestricted as well. In contrast, particulars (composite wholes, entities, things) are spatio-temporally restricted. Universals are abstractions, whereas particulars are the actual things that populate the universe. Prediction is achieved through abstract generalization, while particulars serve as the empirical basis, the explanans, for formulating and others for testing scientific generalities. Generalities cannot be tested with other generalities. For example, the universal abstract proposition "all swans are white" can only be tested with observational propositions of swans, not with other theoretical propositions like "all swans are black". Note that we refer to "observational propositions" instead of facts, because our so-called facts are themselves fallible in terms of our observations. Most fields of biology, like ecology and population genetics, employ frequency probability in their empirical search for generalities, and we are not denying the utility of this approach in those pursuits. Phylogenetic systematics, on the other hand, is concerned with the explanation of historical particulars, one of singularities, such as clades and evolutionary transformations.

Ordinarily, membership in a universal is determined intensionally (connotatively) — that is, there is a list of properties severally necessary and jointly sufficient for inclusion (or exclusion) in the set. These intensions cannot be "wrong"— it is impossible for a member of a set not to meet the set's definition. So, classes have "sharp edges", because of the exclusive middle, something being either in the set, or not. Members of a class have no spatio-temporal connections (Fig. 1A); historical origin and location in space are irrelevant to deciding membership. The reason for this is that the definition of a universal cannot change, without becoming a different universal. For example, consider "things round". In this class, there is, for instance, a coin, a plate, a discus, etc. Even two different pennies included in the class have no spatio-temporal connection, because their inclusion is determined by their roundness, not by being pennies, or having been reproduced by the same mint. The member / class relation is logically non-transitive. That some coins happen to be in the class "round" does not deny that there may be coins outside the class (Canadian nickels, for example). The names of universals are predicates.

A particular can be either a member of some class or a part of some other particular. Those particulars of the latter kind, which exhibit continuity, one to another, also are individuals (cf. bird wings above), and historical connections (like genealogy and descent with modification) count as continuity (Fig. 1B). The spatio-temporally restricted nature of particulars is necessary for continuity. In sharp contrast, as noted above, a universal is denied historicity, because it is unrestricted. The continuous existence of an individual has two consequences. First, there cannot be instances (pl.) of an individual. Identical twins are nonetheless two individuals, and there cannot be two Mammalia, anymore than there can be more than one Darrel Frost. There might be more than one definition of Mammalia, however, no two definitions can be held simultaneously to be correct. Likewise, each

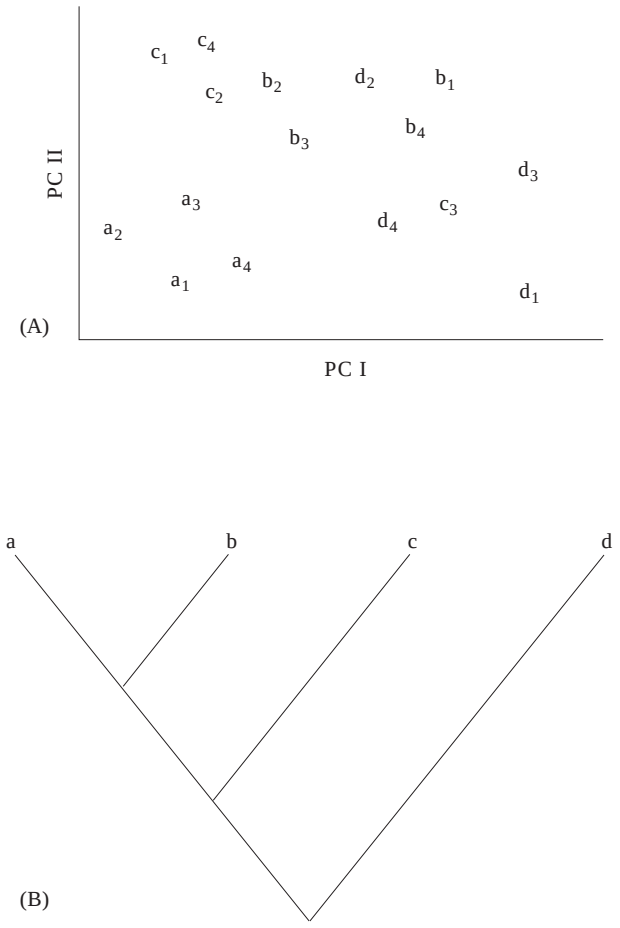


FIG. 1. Examples of universality (A) and individuality (B). In panel A, a principal components space (PC I by PC II) intensionally defines the similarity of four species' (a-d) representatives' (1-4) locations within that space. The phylogeny of four species (a-d), whose history is "defined" ostensively in terms of their relative recency of common ancestry, is illustrated in panel B.

evolutionary transformation is unique in a spatio-temporally restricted, historical, sense. Even physico-chemically identical nucleotide convergences in DNA cannot be the same historically, because each transformation is spatio-temporally exclusive of the other. Second, however, it is possible for an individual to change during its existence and still retain its individuality. For example, an organism, the paradigm individual, exhibits ontogeny and that change, no matter how great (e.g. metamorphosis in anurans), does not alter the wholeness of the individual, either spatially or temporally.

Cohesion is what keeps the parts of an individual together, and cohesiveness may be expected of an entity that has continuity. The cohesiveness of an individual is ascribed to the integrated actions of its parts (Kluge, 1990). For example, an organism's cohesiveness may be due to its parts being functionally integrated, as the result of genetic and epigenetic processes. It is in this functional sense that organisms are said to be self-delimiting. Some consider gene flow (recombination) to be responsible for the cohesion of biparental species. Developmental homeostasis might be another significant process, in so far as it may help to explain the apparent cohesiveness of uniparentals. Continuity and cohesion imply that individuals have the potential to be affected by, and participate in, natural processes. If an entity reacts and acts as a cohesive entity it is the focus of some natural process. Such participation gives those entities reality, whether or not humans are present to perceive or discover them. Universals come to be meaningful only to the extent that they are instantiated by particulars.

Three general kinds of individuals might be of interest to evolutionists and systematists. There is the interactor, the individual that acts in a unitary way in processes (e.g. an organism interacts in reproductive processes). Second, there is the replicator, an interactor that reproduces itself, such as mitochondria, cells, and DNA strands. And finally, there is the historical individual, some more inclusive part of history, whose cohesiveness is apt to be nothing more than a consequence of its history (Kluge, 1990; Ereshefsky, 1991; Frost and Kluge, 1994).

Parts of history can be referenced in different ways. For example, the clade Mammalia is specifiable by enumeration of all things mammal, mammal = {monotreme + marsupial + placental}, or by pointing to the common ancestor of same. Thus, individuals are "defined" ostensively. Notice, by either enumeration or pointing, the sister groups or the common ancestor are the parts, not the entity whose name may be intensionally defined (*contra* de Queiroz, 1992:305). Also, Mammalia can be diagnosed: mammal = {things with mammary glands, hair, etc.}. Here, the part/whole relation of the clade Mammalia is set forth extensionally (denotatively). "Definition" by extension occurs when a group of items or observations taken together are evaluated individually, and generalities about them are used descriptively in the form of an intension (Brady, 1983). A character diagnosis in cladistics is an extension. Primacy in extension and intension can be

easily confused, the things taken together being most important in the former, whereas it is the definition itself in the latter that is critical. Unlike intension, there is no basis for saying the individual has been "correctly" identified whatever form ostension takes, by enumeration, pointing, or extension. No completely accurate definition or essential characteristic exists for an individual. Relative to a precisely defined class, an individual has a relatively "fuzzy" boundary (physically speaking). For example, when an organism originates, and ceases to exist, cannot be specified precisely. Similarly, an evolutionary transformation cannot be defined with class-like precision.

Individuals are said to be parts, not members. The part / whole relation of individuals is functionally transitive. For example, Darrel Frost is simultaneously part of other individuals, *Homo sapiens*, Homonidae, Primates, Mammalia (and *cetera*), by virtue of common ancestry. Individuals are given proper names (e.g. Darrel Frost, *Mus musculus*, Vertebrata).

Thus, it follows that an intensionally defined set satisfies the requirements for frequency probability — precision and potentially indefinitely many independent members (Bartlett, 1962:10–36). However, it is equally clear that the application of frequency probability to individuals (clades, lineages, organisms, synapomorphies) is inappropriate, because of their uniqueness and *non*-independence. For frequency probability to apply to phylogeny there has to be a set of *simultaneously* possible trees, but if only one tree can be "true" then all others necessarily are false (Fig. 1B). So, historical particulars are singular (the same argument holds for other parts of history, like evolutionary transformations). Frequency probability can only assess the probability of particulars in a class of concurrently possible instances, such as in the principal components space illustrated in Fig. 1A. The class to which the particulars can legitimately belong has a size greater than one, and there is nothing that forbids two or more particulars from occupying exactly the same point in that space.

Those who apply frequency probability to the instances of sister-species common ancestry, as well as to the instances of descent with modification that serve as the tests of that history, must therefore seek special justification for their applications. All such justifications will be judged against other approaches used in reconstructing history, and especially those which do not involve frequency probability. Obviously, cladistic parsimony presents a particularly serious challenge to those who may argue for maximum likelihood. Cladistic parsimony is sufficient to explain the relation of individuals, and its justification lies in *logical* probability which does not violate any of the tenets of individuality.

## STATISTICAL CONSISTENCY

Homoplasy provides a useful basis with which to explore further the issue of probabilism. In phylogenetic systematics, homoplasy is considered just as deserving of explanation as is homology. On the other hand, the more statistically minded consider homoplasy to be problematic "noise", that is, something to be factored out in order to improve estimation. The importance of statistical consistency in evaluating methods, such as cladistic parsimony and maximum likelihood, depends on how homoplasy is to be understood.

Hillis (1995:5) argued that accuracy can be measured in terms of statistical consistency (see also Penny et al., 1992). A method is said to be statistically consistent ("convergent" in the parlance of philosophers) when it is certain to converge on the truth when applied to a data set of infinite size. Otherwise, the method is said to be statistically inconsistent. Statistical consistency follows from the law of large numbers, where the sample can be made conceptually large without limit. Hillis (1995) redefined "efficiency" in terms of how frequently the "correct" tree can be obtained. In fact, though, an unbiased (convergent or consistent) estimator is not an efficient one, whereas a biased one is bound to be more efficient.

The issue of statistical consistency in phylogenetic inference was raised by Felsenstein (1978), who attempted to show that cladistic parsimony, under his particular model of evolutionary change, not only would fail, but would do worse as more data are obtained. The statistical inconsistency space, in reference to cladistics, has been called the Felsenstein Zone, and the underlying general issue is referred to as the "long-branch attraction" problem. Consider the example in which two "long branches" exhibit evolution in proportion to the length of the branch, according to a

maximum likelihood model. Felsenstein's (1978:407–408) particular model in that example requires "that parallelism of changes be more probable than unique and unreversed changes in an informative part of the tree". Hendy and Penny (1989; see also Farris, 1973) claim that cladistic parsimony is statistically consistent under models other than the unequal rates of evolution example illustrated in Swofford et al. (1996: figure 8). The zone of statistical inconsistency is given the special name, the Felsenstein Zone, because "the only hope of getting the correct tree is by sampling few enough characters that we may be lucky enough to obtain more of the character patterns favoring the true tree than of the more probable character patterns favoring the wrong tree" (Swofford et al., 1996:427). This amounts to an assertion of truth from sampling error. Sober (1988: figure 16) generalized the issue of statistical consistency, which exemplifies how it has been used in evaluating different phylogenetic methods. But, is statistical inconsistency and the Felsenstein Zone, as it applies to cladistic parsimony, anything more than a red herring?

Concerns for consistency are inseparable from those regarding truth, but Hume's (1739) challenge (the problem of induction) has never been met, suggesting that "the search for truth" was a misguided venture in science from the start and one that has no basis in reality. The simple argument is that we never can know an objective truth, thus "accuracy" is rendered empty as an empirical aim. The frequency probabilist argument which claims convergence on the "true" parameter when increasing data consistently converge on one solution belies the more subtle problem of induction. Watkins (1984:163, see also Quine's (1975) Underdetermination Thesis, and Felsenstein's (1973:241) own admission) remarked that "whenever a highly exact theory T seems to be brilliantly confirmed by evidence E, there is a huge, indeed infinite, set of possible alternatives T', T''... to T, each having a relation to E similar to that which T enjoys". Even in Edwards' treatise on likelihood (1992:34) we are admonished that we should "also be influenced by the simplicity of the hypotheses, by their relevance ... and by a multitude of subtle considerations that defy explicit statement. The scientist must be the judge of his own hypotheses, not the statistician".

For example, why is it necessary for a rule of inference to be statistically consistent for it to be reasonable (Sober, 1988)? Consider that statistical inconsistency has no bearing on the reasonableness of cladistic parsimony as a refutationist research program in phylogenetic inference. Steel et al. (1993a; see also Farris, 1973) also point out that cladistic parsimony can be rendered a statistically consistent estimator of phylogeny if gene sequence data are corrected for unobserved substitutions. Such action, however, has the disadvantage that cladistic parsimony can then no longer be justified as a refutationist research program (Kluge, 1997). No practical consequences follow from demonstrating statistical inconsistency, because there is no reason to believe that the method necessarily will fail in the finite case (Farris, 1983). Moreover, statistical consistency does not suffice to justify a rule of inference. Two methods (even phenetics, under certain circumstances) might converge in the long run on the truth; however, in the finite they might not. As empirical scientists, we must operate in the finite. Consider a more familiar example: the mean and mode of a normally distributed parameter will converge on the same (true) answer in the infinite case; however, their estimates are different with finite data (Sober, 1988, 1993). Thus, statistical consistency is not a sufficient criterion with which to judge methods, either cladistic parsimony or maximum likelihood.

An estimator (rule of estimation) is consistent, or inconsistent, only in relation to some "model", and if the model is false then there is no guarantee that the method will converge on the truth when it is applied to the real world (Farris, 1983; Sober, 1988). It is, of course, easy to demonstrate the fallibility of any method, even maximum likelihood. Some have maintained that parsimony *generally* is inconsistent (Goldman, 1990; Yang, 1993, 1996; Zharkikh and Li, 1993; Takezaki and Nei, 1994; Tateno et al., 1994; Swofford et al., 1996; Huelsenbeck, 1997) as though it has been demonstrated that parsimony is always bound to converge on the "wrong" solution and is simply not to be trusted, whilst maximum likelihood is generally consistent. First, no method in and of itself can be said to be consistent or inconsistent. Second, consistency is not an emergent property of a method, it is a property of estimators for given situations and analyses. All methods potentially are consistent and all methods potentially are inconsistent, the difference being merely circumstantial; any method

> "that is consistent under one set of circumstances can be made inconsistent under others; it is only a matter of imagining the circumstances. If the model employed is not constrained by realism, 'consistency' is meaningless...Since the kind of argument [Felsenstein] employed shows the same 'fault' for every conceivable method, it in fact shows nothing" (Farris, 1986:25).

Finally, there is "Wald's dilemma" (Wald, 1949) — if the number of nuisance parameters increases as a function of the addition of "data" then the method cannot be consistent. Likelihood justifications consider convergence in a method only in relation to the addition of ever-increasing character information. Although in some circumstances characters are of interest to phylogeneticists, the core of the cladistic research program concerns relationships among taxa. Adding taxa to a maximum likelihood analysis has received little attention, but adding taxa increases the nuisance parameters proportionately. This is no different for adding characters, of course. Adding one more nucleotide entails the estimation of ntax-1 new ancestral states. That is, the addition of any data monotonically increases the number of parameters needing estimation and renders the starting parameters of base frequencies and branch lengths in time as incidental, rather than structural (see Models and Assumptions below). Edwards (1992:109) saw "no reason to suppose that it is always possible to eliminate a nuisance parameter" and that "integrating them out of the model ... is a dubious virtue, for if a parameter is inextricably associated with another parameter ... I should prefer to face up to the fact that I may have inadequate information". All this indicates that, in principle, when applied in phylogenetic inference, maximum likelihood *cannot be saved* by an appeal to statistical consistency.

Goldman (1990:348) offered the observation that "statisticians have never doubted that consistency is a desirable property". This is simply false (e.g. Fisher, 1938; Hacking, 1965; Edwards, 1992). Leaving aside the notion that consistency is a "primitive postulate" (Fisher, 1938), it might yet deserve our attention. There is a fundamental distinction that can be drawn between the task of statistics and that of phylogenetics. Statistical measures are required for making estimations regarding collections of observational propositions by way of an abstract generality. That is, there is no spatio-temporally real thing that we can point to called a "mean", or a "variance". These are useful abstractions that exist only in our psychology as scientists. "Mean", "median" and "mode", for example are simply measures of central tendency for a population of individuals. But neither does "central tendency" exist in any physico-objective reality. Consider that the mean number of fin-rays in a group of gobiid fishes can be said to be 6.3. Whereas this may be a useful construct for estimating whether or not this group of gobies has more rays than some other group, we can simultaneously be certain that there is no goby anywhere that has 6.3 fin rays. We can also reasonably state that even though this group of gobies can be said to have "significantly more" fin rays than some other group, this does not mean that all gobies in this group necessarily have more fin rays than all gobies (or a particular goby) in the other group. How is "statistical error" to be interpreted in an evolutionary or other singular framework? Is there some super-psychological reason that any one goby should have 6.3 fin rays, for example?

The issue of statistical consistency has something to say about the reliability of our empirical measures as reasonable estimators of some meaningful abstraction, but it is silent on the reliability or specifics of the data themselves; the data merely are. In phylogenetics, however, we are not interested in some abstract generality regarding the group of taxa we are working with. We are concerned with uncovering the actual spatio-temporally real history of divergence, the species genealogy. Without wholly belabouring Sober's (1988) coin-flipping example, the irrelevance of consistency to objective reality amounts to common sense. If the best corroborated hypothesis is false, this merely means that the data are lying to us about the objective truth. If the data are lying to us, short of special knowledge or providential wisdom, surely our best corroborated hypothesis ought to be false! To assert otherwise is simply unempirical. The lack of recourse to objective truth (*sans* time-machine) renders these concerns immaterial anyway. We should be at least suspicious of a rubric that requires "convergence on the correct tree as the data available become infinite" (Hillis, 1995:4) when neither a correct tree nor infinite data ever *will* be available. In the practical case which seems to consume our colleagues with doubt and angst, should two long-branched taxa group together in a cladogram, it may well be that the two taxa actually *are* each other's closest relatives (see also, Farris, 1986:25). Carmean and Crespi (1995) and Huelsenbeck (1997)

were quick to judge the grouping of Diptera and Strepsiptera as indicative of statistical inconsistency, because it conflicted with their preconceived notions of relationships, and because they had long branches. Yet Whiting et al. (1997) demonstrated that independently two different genes and morphological data all point to these taxa actually being each other's closest relatives. The likelihoodist, in denying his own basic observations (e.g. that these two cytosines are the same) relinquishes any empirical foundation for specifying precisely why these two branches necessarily should be kept *apart*.

## MODELS AND ASSUMPTIONS

Arguably, it is not possible to explain something *de novo*. That is, all knowledge is inexorably intermingled in ways that often are less clear than we would like to believe. Thus, the charge that there can be no inference without making some assumptions of some kind is valid. Galileo's inferences regarding the moons revolving around Jupiter, though seemingly simple facts (observational propositions), cannot be separated from assumptions regarding optics and the bending of light through a series of lenses prior to arriving on his retina. Nonetheless it is a grave mistake to class all underlying propositions in one summary notion of "all assumptions are models". In the systematics community, these arguments often are made in a framework such as "you cannot have assumption-free science, we make our models explicit, parsimony does not, yet this does not free it from assumption". Comments such as these, and like Goldman's (1990:346) "no model, no inference", conflate all assumptions with model.

There are deterministic assumptions (which we call models) and there are non-deterministic assumptions (which we label background knowledge). Models, by their very nature in epistemology, determine (in part) the interpretation of observational propositions. A simple case is the evaluation of the difference between some parameter of two populations wherein an assumption of stochastic normality is made. We do not deny the value of these model-assumptions. In fact, the assessment of such an abstract population difference is not possible without making some sort of deterministic assumption. It is also clear, however, that whatever hypothesis is found to be corroborated, is necessarily a joint hypothesis. We may well reject the null hypothesis of sameness, but we do so with the understanding that we have rejected the null hypothesis either because it *is* in fact false, *or* because the model we have imposed on the question is not appropriate. In this sense the assumption is "problematic". "Problematic" should not be confused with "is a problem".

An assumption is problematic if it is deterministic to the outcome of the test. Background knowledge is, by definition (Popper, 1963), unproblematic. It is something we can assume as holding "true" while we conduct our test. We simply assume that it is on the one hand necessary but not deterministic to our conclusions. It need not be actually true in a metaphysical sense. For example, Galileo's observational propositions took as background knowledge assumptions about optics. This is unproblematic, because it hardly exerts an effect on the moons of Jupiter. Describing the orbit of the moon around the earth entails some background knowledge such as an assertion that the disk in the sky actually represents a sphere, not a disk, and that the moon is not created each dusk and destroyed each dawn. What seems frequently to get confused is the notion that any knowledge we might have necessarily is "background knowledge". We trust that the distinction between "problematic" and "unproblematic" clarifies the issue.

In phylogenetics, cladists do make assumptions. But are these assumptions models, or are they background knowledge? We take it as unproblematic that the species in our analysis are in fact related. This amounts to an assertion that there is some singular history of life and that somewhere there is a common ancestor for all species. Note that it is not a statement about how species are related. Is this a deterministic model or is it background knowledge? If species are not somehow related (life is polyphyletic or was providentially created), the background knowledge we have assumed is false and our hypothesis will be false. However, having tentatively held the assumption does not determine *how* the species will be judged to be related by character evidence. It is a necessary assumption but it is not a *problematic* assumption. More to the point (Ward Wheeler, pers. comm.), if we discover tomorrow that all life is the product only of special creation, we can still do cladistics, operationally, in terms of summarizing the observed character generalities.

Notably, we could not do likelihood, because there would be no meaning to the imposed models of histories of base-substitution, transition probabilities and branch-length nuisance parameters required by the method.

Cladists also make the assumption that the clades under consideration are not the product of hybridization between different species. Again, this is taken as unproblematic background knowledge. Having made this assumption will not determine the results. The results are determined by the competition among characters for groups. Nonetheless, it is an assumption, and a serious one (though *unproblematic*, it is a *problem*). Should we ever "know" that all taxa are the end products of interspecific hybridization, we would then have to consider abandoning cladistics, but we would only do so in favour of a yet-to-be-developed methodology which could deal non-arbitrarily with reticulations in terms of explanatory power.

Cladistic parsimony also takes as unproblematic that the patterns of character distribution are historically contingent (descent with modification). It may well be that this is not so. It may be that character patterns are due to something else. However, if they are due to something else, this does not change the calculus of the cladistic parsimony method. That is, the *making* of an inference does not *require* that characters be historically contingent, though this may be a "problem" for making a "true" inference, because we cannot know the truth, it is "unproblematic".

Consider the following situation: a man is found hanging from a noose in an empty room that is locked from the inside. We observe that he is dead. We observe that there is a puddle of water on the floor beneath him. We hypothesize that he committed suicide by standing on a block of ice, throwing a noose around his neck and waiting for the ice to melt. We have a hypothesis that is better corroborated than the hypothesis of foul-play. Note that there is no assertion of truth here. It may (in truth) be that he was murdered, but such a hypothesis would require additional data for it to be *corroborated*. For the moment, we have suicide as our best hypothesis. What assumptions accompany the suicide hypothesis, and are they background knowledge assumptions or are they deterministic assumptions? We have taken it as unproblematic that ice melts at temperatures greater than 0° C. But then if ice did not melt at 0° C the man is still hanging by his neck. More to the point, he is dead. We have taken it as unproblematic that this man knew how to lock a door. But if he did not, the door would still be locked from the inside. In fact, no deterministic models are assumed here; there is only background knowledge. Most importantly, there has been no statement about the *frequency* with which this kind of man kills himself from which we *calculate* the probability that he has done so. Nor is there any deterministic model regarding the frequency with which people *on the whole* commit suicide by standing on a block of ice from which we determine that it is likely he has done so. These models are not required for the inference of this singular case, and they have no bearing on it.

It has been argued (Felsenstein, 1978, 1982; Saether, 1986) that cladistic parsimony assumes that evolution proceeds parsimoniously. There appears to be an assumption that homoplasy is rare, because it minimizes homoplasy. This is a background-knowledge criticism, not a model-based criticism. There are many models that can be considered which would *cause* homoplasy to be rare, but this was not the criticism. The criticism is that homoplasy is *assumed* to be rare, and this clearly does not obtain. As many analyses with a global consistency index of less than 0.50 will evidence, finding the most parsimonious tree does not require history to be parsimonious. Likelihoodists have argued that because cladistic parsimony is unlikely to discover the correct tree unless rates of change *are* slow, it must then necessarily assume that rates of change are slow. This, however, makes a variety of presuppositions. The most serious of these is that there is an inference to be made from transformations that have gone unobserved. This, however, falls afoul of the fallacy that absence of evidence is necessarily evidence of absence. That is, maximum likelihood is supposed to be superior in its "allowing for" multiple substitutions. By definition, there are no tangible observations requiring this explanation. Thus, the concern for these unobserved changes amounts to an ad hoc declaration that there must be something wrong with our observations in the first place. This conclusion is necessarily based on a presumption of rates of change, predicated on a generality derived from all other nucleotides in the analysis. It is no accident that Popper (1983:133) saw ad hocness and circularity to be opposite to independence. In other words, one cannot

claim that nucleotide sites are treated independently and simultaneously "correct for" multiple substitutions. It also presupposes that a frequency probability framework is the most appropriate within which to assess history. The argument that a most parsimonious tree is *unlikely*, is no more damning than is the charge that a likelihood tree is *unparsimonious*. Though rates of change might be a *problem*, they are *unproblematic* to the calculus of cladistic parsimony.

Relating back to the issue of statistical consistency, wherein the *data* are lying, surely the best explanation of the data, the most corroborated hypothesis, ought to be a falsehood. Cladists contend that frequency probability is not the correct framework in which to judge historical hypotheses. We know (background knowledge) that the improbable is possible and that the possible can occur. To argue otherwise would deny the existence of life itself for surely it is improbable (see above). The probabilistic argument lies at the very heart of contemporary creationist arguments; after all, "there is no mathematical probability whatever for any known species to have been the product of a random occurrence of random mutations" (Cohen, 1984:205). The calculus of cladistic parsimony does not require anything regarding rates of change, amounts of homoplasy, base compositions or the like. Nor does it require that which we observe to be probable. It requires only that the preferred hypothesis be better corroborated by the data than the alternatives; that the explanation explains the explanans. This cannot be claimed (and indeed has not been) for maximum likelihood methods.

As indicated above, in model-based estimation, what may seem to be merely another hypothesis is in fact a conjunctive hypothesis which involves something other than that which the experimenter is interested in, and yet which the experimenter cannot ignore. Likelihood analyses must assess the probability of a given change in light of some model. For example, if transversions are less likely than transitions under the model, and one has the choice of a transversion here or a transition there, the transition will be considered more likely. What underlies the hypothesis of species relationships, in this case, is a requirement that transitions actually *are* more likely than transversions, and *always* have been in all *particular* cases. A mere supposition has now been taken as an absolute. If transitions are not more common in the history of these taxa, then the *method* has lied (even if the *data* have *not* lied). Moreover, because of issues like relative branch lengths, whether a transition is two versus five times more likely than a transversion is also deterministic to the result. That is so, because the experimenter's choice of transition / transversion ratio is taken as fact and that governs the end point of the test. Whether or not the choice of model has an *observable* effect is irrelevant. The calculus of the method is such that a choice will govern the probability of a kind of change, and thus has a bearing on the calculus of the method even if it does not effect the outcome of the analysis. The likelihoodist who asserts that because the shape of the tree "is robust to" changes in models misunderstands her own epistemology. Because frequency probability is the philosophical framework, if one model results in a tree with a probability of 0.004 and some other model results in the same tree with a probability of 0.002, the framework *requires* that the former tree+model estimate be explicitly preferred. Huelsenbeck (1997) admitted this in his preference for the HKY85+$\Gamma_5$ model's achieving a lower log-likelihood.

"Robustness", though over-used in systematics, is seen as a supposed virtue for a phylogenetic method with little regard for what it can be said to be robust to. Likelihoodists claim (Huelsenbeck and Hillis, 1993; Hillis et al., 1994) the robustness of their estimates to violations of the model, and thus appear to imply that this holds for any violation of model assumptions. An example: data are generated according to stochastic model M1 (e.g. Jukes–Cantor) and a tree is estimated employing stochastic model M2 (e.g. Kimura 2-parameter) and it is found that M2 makes the same (or similar) estimate as would have been made if M1 had been employed in the estimation. The *method* is then proclaimed as being "robust". In fact all that has been established is that if the *process* of all character change was in truth M1, M2 would also do a good job of recovering the true tree. So, is it that the *method* is robust in discovering? Or is the *modeled process* merely robust to being discovered? Even if a cornucopia of stochastic models are used in generating data, and then in estimating a tree from those data, the best that can be said is that the method is robust to the discovery of phylogenies when characters have changed due to purely stochastic processes. In this light it is not surprising that some are now declaring that "Evolution *is* a stochastic process" (Penny et al., 1991:160, emphasis

added). The neutral theory of evolution may well be an interesting abstract generalization about the universe of molecular sequence data (or perhaps not, Gillespie, 1991), but when this generalization is rejected by the data, surely we are compelled as rational scientists actually to reject it, and with it methods that rely upon it for their justification. To paraphrase Hillis (1995:6): "It makes no sense to simulate a tree using a [stochastic model]... and then conclude that a method that assumes a [stochastic model] is generally better".

The manner in which model-based assumptions are chosen in a likelihood analysis runs counter to several well-characterized logical fallacies. The most damning of these is "*post hoc ergo propter hoc*" (after this therefore because of this – confusing pattern with cause, or more simply, having the cart before the horse). An extreme of this is the syllogism: "All humans breathe. All humans eventually die. Therefore dying causes humans to breathe". The fallacy is merely arguing to a premise from a conclusion. Various authors (e.g. Thompson, 1975; Goldman, 1990) have recognized the need to avoid specious assumptions of equal rates of change and equal base composition in the model used by likelihood. This is also required, in part, to avoid a particular problem relating to independence of assigned probabilities (see below). The manner in which this is accomplished is to determine base composition from the extant taxa and relative branch lengths from those extant character distributions as well (on, for example, a preliminary parsimony tree). Because these values, the observed patterns, are employed in the model, and the model is taken to be causal of the observed patterns, the analysis explicitly has taken the *results* of an evolutionary process to be *causally* part of that process (the assigning of a base composition to some arbitrarily chosen root node entails the same problem). This is very much like concluding that the bending of a tree causes the wind to blow. Admittedly, this is not a new gambit in evolutionary studies. Evolutionary taxonomists, for example, might have believed the essence of tetrapods to be terrestrial locomotion, and thus assigned primacy to the pattern of limbs as especially causal: "Tetrapods have femurs so that they might walk". We trust that the incendiary debates about essentialism and teleology do not need repeating.

Likelihood requires assumptions, but these assumptions are models that are deterministic of the outcome, as opposed to assumptions which are unproblematic. Models are, by definition, assertions regarding generalities in the universe. Some may wonder if we cannot accept it as background knowledge that, on the whole, transitions do occur more frequently than transversions. To be taken thus would require that statement to be unproblematic, and non-deterministic to the results. Clearly this is not so. In a likelihood inference, the model assumptions that must be made relate to prior probabilities of base composition ($\pi$), branch lengths and the specifics of classes of transformation types (transitions and transversions). In order to achieve the probability of the tree in question one must find the product of the probabilities of characters, which are determined by the sums of the probabilities of each possible pattern of ancestral states, which in turn are determined by the product of the probabilities of each transformation, which finally, is determined by the probabilities assigned in the four-by-four Q matrix. Because all of these components go into the calculus of the method and determine the results, they cannot be said to be unproblematic. Our most serious concern is the static nature of Q. That is, topologically, the same Q matrix is required to apply across all branches, for all clades for the entire history of the tree. (That the Q matrix might be allowed to be iteratively adjusted does not save it from this criticism. The final Q matrix may well be different from the original but it is still applied as a universal across the tree.)

Whereas we might be able to demonstrate on the whole that the proportional representations of bases A, C, G and T are 0.26, 0.24, 0.27 and 0.23, respectively, it hardly seems reasonable to assert that this has been so in all aspects of the tree across all characters and throughout the entire history. However, this is precisely Felsenstein's (1988:529) assertion that "processes of base change probably do not differ much in related species". This concern is not trivial as it must be remembered again that the resulting hypothesis is conjunctive: this tree is a good hypothesis assuming the applicability of the model. Whereas some dismiss the concern for models as merely a matter of best-fit (e.g. Swofford et al., 1996; Huelsenbeck, 1997), others have acknowledged that the models used in maximum likelihood are contingencies which require testing outside of the context of the analysis itself (Thompson, 1975;

Farris, 1986; Sober, 1988; Goldman, 1990). That such testing is not regularly done does not reduce the force of this requirement. Consider a simple empirical case involving the protistan group Alveolata. Alveolata is comprised of three principal clades: ciliates, dinoflagellates and apicomplexans (the latter including malaria parasites). On the whole, for actin, nucleotides are about equally represented (between 0.23 and 0.26 each). Yet, when we break this down into the constituent phyla, we find that apicomplexans deviate considerably from the other two ($\pi_A$ = 0.29, $\pi_C$= 0.19, $\pi_G$ = 0.21, $\pi_T$ = 0.28). However, even these deviant values are not representative of the particular taxa within the Apicomplexa:

| | $\pi_A$ | $\pi_C$ | $\pi_G$ | $\pi_T$ |
|---|---|---|---|---|
| *Plasmodium falciparum* | 0.36 | 0.13 | 0.18 | 0.33 |
| *Toxoplasma gondii* | 0.22 | 0.30 | 0.26 | 0.23 |

Lest some static model be considered applicable to the ciliates, consider the marked variation *among* ciliates, and even *within* the genus *Oxytricha* :

| | $\pi_A$ | $\pi_C$ | $\pi_G$ | $\pi_T$ |
|---|---|---|---|---|
| *Oxytricha fallax* | 0.25 | 0.31 | 0.23 | 0.22 |
| *Oxytricha nova* | 0.29 | 0.23 | 0.23 | 0.25 |
| *Euplotes crassus* | 0.30 | 0.22 | 0.24 | 0.23 |
| *Tetrahymena thermophila* | 0.26 | 0.26 | 0.19 | 0.29 |

Even these values are not representative of patterns across all sites, as revealed by looking at 3rd positions of codons in isolation:

| | $\pi_A$ | $\pi_C$ | $\pi_G$ | $\pi_T$ |
|---|---|---|---|---|
| Alveolata | 0.17 | 0.35 | 0.18 | 0.27 |
| *T. gondii* | 0.05 | 0.48 | 0.24 | 0.21 |
| *P. falciparum* | 0.42 | 0.06 | 0.04 | 0.46 |
| *O. fallax* | 0.10 | 0.53 | 0.18 | 0.18 |
| *O. nova* | 0.25 | 0.32 | 0.17 | 0.25 |
| *E. crassus* | 0.26 | 0.30 | 0.20 | 0.22 |
| *T. thermophila* | 0.15 | 0.40 | 0.10 | 0.34 |

So, it is clear that a model of base substitution (any one model) cannot apply across *all* characters or across *all* taxa. We need not add to this the issue of rates being differential across clades. In short, we have no more reason to relegate substitution type and transition/transversion ratios to background knowledge than we have to proceed thus with codon usage biases (Eyre-Walker, 1991). Our knowledge of these phenomena is not unproblematic to our epistemology. For these reasons Swofford et al. (1996) asserted that maximum likelihood methods should not be used for amino-acid encoding nucleotides. Should we then infer that Swofford et al. meant we should only use maximum likelihood for structural RNA genes and for introns? These same concerns apply equally well to RNA genes wherein there are hypervariable and conserved regions relating to stems and loops in the secondary structure. We are left, then, with an epistemology that disallows use of morphological data, disallows use of amino-acid encoding data, disallows use of nucleotide data showing codon or dinucleotide biases, and yet claims to be more realistic than the use of cladistic parsimony which permits (indeed insists) on using all of the data available (Kluge, 1997). Is it any wonder that in the face of this methodological limitation, likelihoodists are now insisting that the data are limited, not the method, and that morphology never had anything to offer phylogeny and systematics (e.g. Hedges and Maxson, 1996). We see Miyamoto's epistemological abjuration, having once argued so strongly in favour of total evidence (Miyamoto, 1985) and now insisting that data never be combined (Miyamoto and Fitch, 1995), as a thinly veiled verification of his newly acquired acceptance of likelihood, notwithstanding the attempt to couch it in terms of vague "process-partitions".

In summary, a static Q matrix is *deterministic* to the outcome of a likelihood analysis and thus is not background knowledge. The Q matrix is applied across the entire history of the taxa under consideration, and across all characters, and yet the applicability of constant base compositions in separate clades is refuted easily, violating the requirement for spatio-temporal independence of the universal model. The deterministic imposition of constant frequencies of transformation types is similarly refuted. It may be suggested, then, that Q is only a weakness, not a flaw, because having recognized this problem we need only

apply multiple models to circumvent this failing (as is attempted with Gamma distribution corrections [Steel et al., 1993b; Yang, 1993;] and the like). Unfortunately, this will render likelihood untenable, not because of the computational difficulty (though this is serious enough), but because of how this impinges on the issue of *nuisance parameters*.

"Wald's dilemma", as we have characterized it, is precisely this: a likelihood analysis is impossible if the number of parameters increases proportionally with the addition of new data. Nuisance parameters are parameters which we are not necessarily interested in but which we need to estimate from the data in order to estimate those that we are interested in. In order to assess the difference between two populations we need only estimate one parameter for each population, for example, the mean. No matter how much new data we add, there is still only this one parameter per population. Here, likelihood is tenable. However, Wald's dilemma is encountered, for example, in estimating a line from a series of points. For each new data point added, the parameter of the squared distance of the point to a line must be calculated before we can get to the parameter of the line itself. Wald saw that the only way around this was to impose a *ceteris paribus* clause, his condition #1, that $F(x,\Theta)$ is discrete for all $\Theta$ or is absolutely continuous for all $\Theta$, and admitting some elementary probability law $f(x,\Theta)$. Or, simply, assume homogeneity. The regression line can be estimated in a likelihood framework if and only if there is an assumption of some constant probability density function like normality, but thus requiring independence and identical distributions. For the phylogenetic application of likelihood to be tenable, this assumption must hold as well, though we have shown above, and it is widely admitted (Cramer, 1946; Neyman and Scott, 1948; Lindgren, 1976; Goldman, 1990; Gaut and Lewis, 1995), that it does not.

We do not claim that likelihoodists have been wholly unconcerned with this issue. The reality of rate differentials and the inapplicability of one class of transition probabilities has received some attention. It is variously argued that in order to compensate for these phenomena one could employ, for example, Gamma distribution rate corrections across characters, or a log-determinate transformation across branches. Use of these transformations of the instantaneous Q matrix amounts to an admission that one static matrix is unrealistic and merely substitutes multiple equivalents in its place. We submit that two static models in place of one is merely doubly unrealistic. Although we still object to the notion that anagenic change along an isolated branch is even static and stochastic, the logical end point to arguing for multiple transition matrices would be to have a different one for each internode. This would have us logically employ 2T-2 different matrices for T taxa in an analysis. But, as we have seen, different characters should also be expected to have different properties. If one then tries to modify the likelihood method by permitting *different* models for each character, and different models for each clade or branch, then one adds more parameters to the method for each new character or taxon added to the matrix. One new character would add one new set of base frequencies, and one new set of transformational probabilities to be estimated from the data; and one new taxon would add one new hypothetical ancestor to be estimated, and two new branch length relative rates, and so on. Felsenstein (1988:529) freely admitted that Barry and Hartigan's (1987) allowance for a realistic flexibility in transition probabilities was not tenable, concluding that "the truth must lie somewhere in between", and thus apparently asserting that truth does not require realism. As well, Kim (1996:363) was forced to admit that in maximum likelihood analyses fewer taxa are to be preferred over more taxa. So, the solution to the problem of a static matrix would itself violate the requirement that there be a limited set of parameters to be estimated, or as succinctly put by Goldman (1990:351) "this approach renders inferences virtually impossible".

In this section, we have attempted to clarify the distinction between models and background knowledge as two distinct kinds of assumptions that phylogeneticists might make. Most importantly we have underscored the distinction by asking whether or not the assumption is deterministic to the inference made in terms of the calculus of the method. In the body of its calculus the result provided by cladistic parsimony is determined only by the data. Hillis (1995:5) uncritically accepted the notion that "all methods are based on explicit or implicit assumptions about the evolutionary process". Regarding models of base substitution, this is plainly false. That cladistic parsimony might be said to deviate from some objective truth under conceivable historical contingencies may

be so, but it is not bound and determined to do so by its calculus. It might still render the objective truth or it might not. Because it is not determined to do so in its calculus, cladistic parsimony does not assume a process model. That is, there is no such thing as an implicit model. Models, by definition, are explicit. Likelihoodists cannot overcome the deficiencies of joint hypotheses by merely issuing a declaration that all methods employ them. We also see this concern for the objective truth as illusory. We admit that there is a truth, but this admission can only be metaphysical, it cannot be an empirical guide, because we have no way to assess accuracy (comparison to proven truth) for any singular part of history.

## INDEPENDENCE AND OPERATIONALISM

Kim (1996) argued that the "structure of the models [in maximum likelihood] is not unreasonable" and pointed both to symmetric transitions as well as the requirement for independent and identical distributions of character information. These models would only be reasonable if these criteria ever were met in the methodology. As indicated above, Wald (1949) also asserted the requirement for independence and identical distribution of the data ($X$), and Felsenstein (1973:246–247) admitted as much. We have already shown that these requirements do not obtain for nucleotide data any more than they can for any historically contingent patterns. The argument that cladistic parsimony too requires independence of characters such that likelihood estimation is no more flawed, misunderstands that non-independence exerts its effects at various levels. There are logical or causal dependencies in data (exemplified by interaction effects found in analyses of variance, or base pairing in stem regions of ribosomal sequences). Logical dependence applies equally to cladistic parsimony analyses as it does to likelihood analyses. For cladistic parsimony, there is the requirement that homoplasious transformations are logically independent, otherwise fewer ad hoc hypotheses are required than have been inferred from the data (Farris, 1983; Farris et al., 1995). There are sampling dependencies in data acquisition (such as the mesh size of a fish net used for collecting, or sequencing nucleotide characters in a row). Sampling dependencies, though of critical importance in likelihood analyses, are of limited concern for cladistic parsimony. With corroboration as the governing principle in cladistics, *any and all data* are worthy and capable of offering refutation of relationships. It matters not if those data were sampled in a particular order, in a row, or alphabetically for that matter. Sampling widely is still important if we are to claim severity of test, but sampling randomly is not. Finally, there are calculus dependencies which are not at all relevant to cladistic parsimony but can be shown to confound any application of likelihood. At the level of the calculus of a method (that is, the requirements of the mathematical procedures) cladistic parsimony makes no assertions. Suppose that we can assert that all character distributions are independent and that characters were sampled randomly (assertions, by the way, that never *can* be made insofar as "random sequencing" is an oxymoron), the phylogenetic application of maximum likelihood still fails regarding independence in the calculus of the method. The calculus of the likelihood of a tree involves two multiplicative levels and one additive level. For probabilities to be multiplicative in frequency probability calculus, they must be independent of each other. For probabilities to be additive, they must be disjunct. The summation of probabilities of the various ancestral patterns are disjunct. That is, one cannot be both A as well as G for a particular hypothetical ancestor. So likelihood satisfies this element of its calculus. However, the two multiplicative relations are (1) the probability of a certain transformation type having occurred (as determined from the instantaneous Q matrix) multiplied across branches in the tree for a character, and (2) the probabilities obtained for each individual character which are multiplied to obtain the final probability of the tree. For these to be multiplicative *in the calculus*, they must be independent quantities, or "since we assume independence of evolutions in different characters, we calculate the likelihood of the tree separately for each character and then multiply these" (Felsenstein, 1973:247). However, the Q matrix is derived from *all* characters. That is, the base compositional $\pi$ values (in the HKY85 or F81 models) applied to character $i$ are contingent on all of the compositions of all other characters. Thus, the probability obtained for character $i$, being dependent on Q, is transitively (in the logical sense, not the temporal sense) dependent on

character *j* (and all others), just as *j* is on *i*. There is a difference between assigning prior probabilities to characters as though they were independent and assigning conditional probabilities which admits that they are not. If one has the conditional probabilities based on *C* characters, the point at which the transition probabilities of *C*-1 characters is determined, that of the *C*th character is already determined. This is why, for example, the *t* distribution equals the *z* distribution only when degrees of freedom are infinite ("Student", 1908). Statisticians regularly acknowledge these biases in conditional probabilities and variously correct for their effects (e.g. Efron and Gong, 1983). Likelihoodists have yet to explain how their phylogenetic applications circumvent this problem of non-independence. The argument could be made that in the face of large amounts of data the effect would be small. In fact, this is demonstrably false. Huelsenbeck's (1997) results, in which likelihood consistently separated Strepsiptera from Diptera, rest on a single character (#203 in the alignment). Maximum likelihood analysis of the same data set used by Huelsenbeck (1997), with this one character removed, groups Strepsiptera and Diptera together irrespective of the model (Siddall and Whiting, 1998).

The LogDet transformation (Lockhart et al., 1994) is particularly revealing in this respect. Lockhart et al. (1994) acknowledged that it was unrealistic to assume that base compositions don't vary across a topology or at nodes. Like Barry and Hartigan (1987) they suggest adding a new parameter to the model ($d_{xy}$) to compensate each internode for this. In point of fact, to do so simultaneously admits historical non-independence of character state transformations on ancestral reconstructions while still requiring this independence in the calculus of the tree estimator. Goldman (1990:350) had already noted that if one estimates "these variables as though they were parameters of the model ... as the amount of data (i.e., the number of characters) increases, the number of parameters increases" and this too falls outside of the scope of consistent estimation (Wald, 1949; Felsenstein, 1973). This, and related approaches, aimed at correcting for the illegitimacy of models in relation to historical contingencies, violates the premise for a Poisson-directed instantaneous rate matrix requiring that we "have assumed that the probability of change during any time interval is independent of the times and numbers of previous changes" (Felsenstein, 1973:241). Similarly, the relative and mean rate parameters (*r* and $\mu$), which apply to character *i* through Q, are dependent upon all other characters, and likewise *all* other characters are dependent on *i*. A possible way around this, regarding rates, would be to assert a model of equal rates of change instead of character-derived relative rates. Leaving aside, for the moment, how untenable a molecular clock actually is, this would merely shift the problem of independence from one of characters to one of transformations. In a constant rate model, the probability of changing on branch *a* is not independent of branch *b* if the lengths of branches *a* and *b* are dependent, as they must be under a clock model. If branches *a* and *b* lie on the same trajectory, making branch *a* longer requires that branch *b* be shorter. The probabilities of transformations being dependent on these lengths cannot then be said to be independent. Felsenstein (1973, 1988) noted that in the absence of a molecular clock, consideration of the nodes in a tree estimator would entail too many nuisance parameters unless some prior distribution could be applied to the duration of the internodes (u dt). That Ranalla and Yang (1996) have now done so by applying a Yule branching process seems to simply ignore Felsenstein's (1981) and Goldman's (1990:356) admonishment that "the many factors involved in speciation ... do not readily permit such a simple model". Even in the case of real character independence, the calculus of maximum likelihood cannot be saved from violating the independence of calculated probabilities. Independence is not a philosophical triviality. It goes to the very core of frequency theory logic. Indeed, maximum likelihood has come a long way from having to "oversimplify the complex process of mutation, natural selection, and random genetic drift into sudden changes from one state to another" and requiring in the model "also the process of sampling by which we obtained the data" (Felsenstein, 1973:240–241). Its simplicity is now a virtue (Kim, 1996) and its models, once acknowledged to be restrictive and unrealistic, now constitute analytical power (Swofford et al., 1996; Huelsenbeck, 1997).

The notion that a likelihood analysis renders an assessment of how probable some *tree* is misunderstands the method. In fact, all that receives a posterior probability in a likelihood analysis, $p(e \mid M, T)$, are the character data (*e*) (that is, the probability of the evidence given the model and the tree), not the tree which

would require the expression to be $p(T \mid M, e)$. Assigning terms to likelihood analyses like "Maximum Posterior Probability" (Rannala and Yang, 1996) serves only to obfuscate this distinction further, as though the tree is somehow made more probable. The likelihood of a hypothesis, given the data, $L(h,e)$, is a probability, but it is a probability of the data [$L(h, e) = p(e, h)$], not of the hypothesis itself. Because only the data are probabilified, likelihood methods might yet be useful for asking questions about some property of the characters. Thus, a likelihood analysis could reasonably ask does a Jukes-Cantor or a Kimura model assign a higher probability to finding these data? Whatever answer is returned from this query could add insight into some *generality* concerning the characters in the group, but only a generality — it would say nothing about how well the data match some *other* stochastic or non-stochastic model's expectations, and nothing about expectations for some other group. However of what value is probabilifying data points that we already know? It says nothing about the probable relationships of taxa, because it is *not* designed to do this. It is for this reason that likelihood values cannot logically be compared on the different topologies as is explicitly attempted in the phylogenetic applications. What meaning is there for probabilifying these data given this tree if the trees are not the same and "why should we accept the probability model?" (Edwards, 1992:209). Relationships among taxa are not abstract generalities, they are real and they are singular. All that is needed, to ascertain the posterior probability of the character data (if this is our interest), is some external best-corroborated tree to add to the model. Cladistic parsimony offers this corroboration.

## LOGIC AND PROBABILITY

Scientific philosophy has undergone considerable change since Aristotle's metaphysical essentialism, and probably more so in this century than in any other previous period (Reichenbach, 1951). Evolutionary biologists, and indeed biologists in general, are rarely concerned with the logical formulations of their epistemology, certainly markedly less so in comparison, for example, to mathematicians and physicists (e.g. Boltzman, Einstein, Galileo, Hanson, Hertz, Mach, Newton, Poincaré, Schrödinger). By this we do not mean to imply that biologists ignore "logic" (in the small "l", Vulcan sense) as it is used colloquially (e.g. the sentence, 2 + 2 = 5, is illogical). In contrast, we are concerned with the formalisms of "Logic" (written large with a capital "L") in the deeper sense. Biologists are no more immune to the requirements for a sound philosophical foundation than are these other sciences if our occupation ever is to be more than a simple cataloguing of the experiences of our senses. Evolutionary biology, and phylogenetics in particular, demands this even more because, like the quantum physicist, we are not able to observe that which we seek to explain. Perhaps it is the dual nature of the phylogeneticists' occupation that has led to this abdication of the requirement for a formal philosophy. With the exception of the end point of our endeavours (i.e. the tree), we are consumed as naturalists with the gathering of data. This snake has 168 ventral scales. Bats have wings. There is a cytosine in the 1236th position of this gene in this leech. And so, insofar as all which systematists do tends to be straightforward and seemingly philosophy-free, there is a tendency to consider the last step as but a continuation of that which precedes it: mere operationalism. Experimenters A and B wonder whether or not this population of fish is part of some larger stock. Experimenter A conducts genetic analyses of fishes from a larger region and finds homogeneity. Experimenter B conducts telemetric tagging surveys availing himself of GPS technology and finds that the fish migrate over some larger distance coincident with the region found by experimenter A. The two lines of inquiry have converged on the same answer and it is legitimately considered well supported. The choice of technique does not amount to a deep philosophical issue in this case. However, it is a grave error for biologists, so used to this regime, to extend this to the uncritical acceptance of a variety of phylogenetic methods (neighbor-joining, maximum likelihood, cladistic parsimony, genetic distance, etc.), and then to seek the consensus of solutions found by these techniques. Carl Woese, for example, asserted at the 1996 meeting of the Society of Protozoologists in Tucson, that the groups jointly found by maximum likelihood, neighbor-joining, and cladistic parsimony (and ostensibly UPGMA too) are explicitly to be preferred. There has even been a kind of revisionist history regarding the demise of phenetic methods, as

though they reached their ungainly scientific death only because of some mathematical sensitivity to unequal rates (Saitou and Nei, 1987; Hillis et al., 1994). Phenetics perished in the face of a failure to achieve greater explanatory power over its competitor, cladistic parsimony, a failure to realize its factor-asymtote claims, and because to assert that similarity is antecedant to phylogeny is to argue that carts push horses. The differences in these techniques run much deeper than mere methodological calculus. The competing methods for phylogenetic explanation are exclusive of each other. Cladistic parsimony denies frequency probabilism, likelihood denies corroboration, and neighbor-joining fails to be explicit at all (Farris et al., 1996). Mutually exclusive methodologies cannot be simultaneously invoked by an experimenter. Contemporary phylogeneticists tend to look to others for their epistemology, perhaps considering the frequency of publications invoking this or that method to be a suitable guide for the acceptability of their science, when in truth one needs to look elsewhere to decide what method is defensible. To act otherwise is to allow arbitrariness and to invite ridicule from the rest of science. Whether it is that cladists "hide behind" philosophy or that likelihoodists "hide behind" naked operationalism depends on whether or not one thinks one needs a rational basis for one's science.

Philosophies of science are not so nearly well defined as are the cursory expositions which follow here, and yet there is need of a clearer sense of a "philosophy-space" so that systematists can better decide what is defensible, and so that the phylogeny debate can continue more intelligibly. Some general philosophies can be quickly discarded as non-scientific and irrelevant to phylogenetic inference. *Academic scepticism*, for example, simply denies the existence of knowledge. *Intellectualism* considers as knowledge only those things which we can fathom independent of our senses (such as geometry, mathematical calculus, etc.). *Relativism* would have us consider knowledge to be unique to the individual's psychology, and thereby deny any objective reality. This is not to say that these philosophies are wrong, only that they are concerned with metaphysical issues, like truth, and cannot guide us as scientists. Scientific philosophies are *empiricist* as opposed to intellectualist, and fall, generally, into *verificationism/justificationism*, or into *falsificationism*. A simplistic view of these generalities is that justificationists are concerned that their theories *be* true, whereas falsificationists are concerned that their theories *not be* false. There are distinctions within these broad groups though. *Classical justificationism* holds that knowledge consists of proven propositions — proven by wisdom, experience or the power of intellect — and has been the hallmark of science for centuries. The claim that truth can be deduced from observation and the induction of logical proof by factual evidence are both justificationist paradigms. For the justificationists (Kant, 1781; Mill, 1878), Euclidean and Newtonian proofs were exemplary. The work of one individual, Einstein (1950), unintentionally destroyed the classical justificationist framework through non-Euclidean geometry and through Relativity's superiority over Newtonian mechanics. All theories thereafter must be viewed as equally unprovable — truth through scientific inquiry vaporized. *Neo-justification*, now espousing probability as opposed to certainty, was required to save the classical justificationist framework from sophistry. For the neo-justificationist, although all theories may be unprovable, some theories are more probably true than others (Carnap, 1956) — a theory is neither true nor false, it is merely more or less probable given the evidence. This offered a comfort zone to scientists, who need not be concerned with being right or being wrong, but need only assign a *P*-value to be able to get on to the next question.

Falsificationists differ markedly in their interpretation of the nature of evidence, but not all falsification is the same. Justificationists (whether neo- or classical) object to the notion that a single disconfirming datum should overthrow a theory in the face of so much confirming evidence. This conflates falsificationism *sensu lato* with the more restricted naive or *dogmatic falsification*. Felsenstein (1988:530) has taken an identically unintelligible stance regarding cladistic parsimony, suggesting that because falsification "is not absolute... statistical concepts must be admitted through the back door". Arguably there are no, and never were any, dogmatic falsificationists like the lurking absolutists imagined by Felsenstein. Falsificationism merely states, in contrast to classical justification, that no amount of positive evidence can prove something true, but this hardly admits that any amount of negative evidence can prove something false. Felsenstein cannot be held wholly to blame for his naive perspective insofar

as Weyl (1949) and Braithwaite (1953) also handed to nature the task of saying "No" to scientists' theories. However, scientists themselves, generally acknowledging the fallibility of their own observations, have not. The *methodological falsificationist* differs somewhat in acknowledging the fallibility of observation, such that a theory is held tentatively, and, in the face of disconfirming observation, three things are possible: either the negative *data* are false (as can, one hopes, be determined by repeating experiments with appropriate controls), or the *theory* is false and must be discarded, or the theory must be expanded in order to account for the observation. The expansion of a theory weakens it by the addition, ad hoc, of auxiliary components. In methodological falsification, a theory teeters on the edge of the explanatory power engendered on the one hand by how falsifiable it is, and on the other, of the tendency to explain everything (and so nothing) as it becomes encumbered with ad hocisms (Lakatos, 1970).

*Sophisticated falsificationism* differs still, and it is to this which we contend cladistic parsimony belongs. We do not pretend to speak for all cladists. As we suggested above, each must decide what they are doing and why they are doing it. Sophisticated falsification does not judge theories by their closeness to some objective "truth" or distance from some objective "falsehood", as neither can be known either for certain or with any degree of certainty. Truth, though not irrelevant to science, is nonetheless irrelevant to the choice among scientific theories, because it is unknowable, and yet we still need a rational reason to choose among competing theories. It is this notion of competing theories which is the hallmark of sophisticated falsification. Suppose we have a Keplerian theory of the movement of planets and a Newtonian theory, and suppose that both of these theories are in some objective sense false, but of course, short of divine intervention, we do not and cannot know that they are false. We are in need of a theory to explain the movement of planets. Kepler's Laws fully explain the elliptical movement of the planets. So do Newton's. Which theory is to be preferred? Kepler's theory has as many confirming instances as Newton's, in fact more because Newton's theory fails to explain Mercury's perihelion (see also Berkeley, 1710). The justificationist cannot choose between the two. The methodological falsificationist might prefer the mechanistic theories of Newton but is troubled by the disconfirming evidence in Mercury ("maybe there is an unseen moon"). The probabilist (neo-justificationist) must assert on the basis of frequency theory that Kepler's theory is *probably* true and that Newton's theory also is *probably* true, or she might even suggest that Newton's theory actually has a low probability, because of the many things it tries to explain. That is, if a theory like Kepler's tries to explain one thing it has a higher probability than a theory which tries to explain many more things, because there is a greater chance, in the latter, that one of those things will not be true (like Mercury). That is, "high probability is the dubious reward for saying very little, or nothing" (Popper, 1983:223). The sophisticated falsificationist sees this all as beside the point. A theory is to be preferred if it has a higher *empirical content* than some other theory. Thus, Newton's theory is to be preferred over Kepler's, because the theory not only says something about planetary motion but also asserts that a 200 lb stone and a 4 oz apple both hit the ground at the same time and at a rate of acceleration of 9.8 meters per second squared. By saying more about more things the theory has a larger suite of potential falsifiers. Insofar as it says reasonable other-things that are borne out, it is more corroborated than Kepler's theory, even though Kepler's theory may have no disconfirming evidence and *in spite of* Newton's theory being born false (because of Mercury). This too is why General Relativity is to be preferred over Newtonian gravitational theory. Even if General Relativity *also* failed to explain Mercury's orbit, it would be preferred for its empirical content. General Relativity embodies Keplerian and Newtonian phenomena and adds to this the contraction of space and time under motion, the speed of light as an absolute, and so on, all in one theory. There is a larger class of potential falsifiers, and at face value Einstein's theory has a very low probability (by virtue of the sheer number of phenomena it claims it can explain).

Having considered philosophy of science thus, it is clear that maximum likelihood is predicated on probabilism (neo-justificationism) where there is a premium on the frequency of unknowable truths, and cladistic parsimony is predicated on sophisticated falsification where there is a premium on corroboration and empirical content, even coming at the expense of probability. We hold that in addition to whether or not a likelihood analysis can withstand the requirements

of its calculus, which in terms of calculus independence we have shown it cannot, it is perhaps more important whether or not its logical framework holds up under rigorous scrutiny. We must beware, as succinctly put by Popper (1983:60), "the cult of impressive technicalities or the cult of precision [which] may get the better of us, and interfere with our search for clarity [and] simplicity".

The neo-justificationist framework has come under serious criticism by philosophers of science for its focus on probability as a guide to scientific endeavor (Russell, 1946) and for placing a premium on auxiliary hypotheses (Popper, 1959, 1963, 1983; Lakatos, 1970). Long ago, Hume (1739:139) remarked that "even after the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience". This is not to suggest that statistical estimation is beyond the pale of serious science, but it is limited to abstractions. Statistical estimation procedures are designed to make generalities about the universe. Whether or not humans are more closely related to chimps or to gorillas is not an abstract generality. When this logical relation is taken to be a symmetrical one, that is, when *specifics* are inferred from *generalities*, probability theory steps beyond its logical bounds. Bernoullian (Bernoulli, 1713) probability theory, for example, states that in the case of large numbers, the objective (parametric) probability will lie very close to the observed frequency and thus the observed frequency will probably be a good estimate of the objective probability. What likelihoodists confuse in this theoretical framework is that this relies on *repetitive* cases and large numbers (truth being obtained in the infinite). In phylogenetics we are not faced with large numbers or repetitive cases. So long as time is taken to be linear, history has occurred but once. Large numbers and their generalities cannot be relevant to finite singular cases. Consider some event's occurrence (1) or non-occurrence (0) and the series of observations:

0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1.

The relative frequency of occurrence is $f(0) = 0.5 = f(1)$. In the infinite this holds as well. In the more limited case, however, it does not. Suppose we only observed five of these in a row. There are 26 ways in which we could have done this. We are most likely to have concluded that $f(1) = 0.4$ [$f = 0.27$], equally likely to have concluded that $f(1) = 0.0, 0.2, 0.6$ or $1.0$ [$f = 0.15$, respectively] and least likely to have concluded that $f(1) = 0.8$ [$f = 0.12$]. Even if we observed all possible series of five observations in a row we would come to the conclusion that $f(1) = 0.48$. The reason for this anomaly is that in the above series of events, even though under the law of large numbers, given the series in its infinite iteration, we can be certain that $f(1) = 0.5$, in the finite case there is more structure to the data than mere stochastic process. Bernoullian laws cannot distinguish the above series from the following:

0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0.

The observation that some phenomenon matches the expectations of a stochastic process "in the long run" does not allow us to infer that the process is, in fact, stochastic or that it will have been in any particular case. An example of this distinction, in a phylogenetic framework, is given in Fig. 2. Across the taxa in this illustration, there is an equal representation of A, C, G and T ($\pi = 0.25$), but the process leading to that representation differs in the two trees. In one we may well have the Poisson-directed Markov rule, but surely we do not on the second representation of their relationships. In fact, as suggested by the analogy above, although the complete representation of taxa in the second tree indicates equal base compositions, if we had been interested only in the relationships of a sub-clade of this group, we would have come to a very different conclusion. The probabilist would be forced to conclude that the underlying processes vary according to our sampling routine as opposed to history, that our observations exert a direct causal effect on history. This is a clear example of the logical Fallacy of Division in which a property of the whole is also taken to be a property of its parts (see also the distinction between specification and scalar hierarchies in Frost and Kluge, 1994).

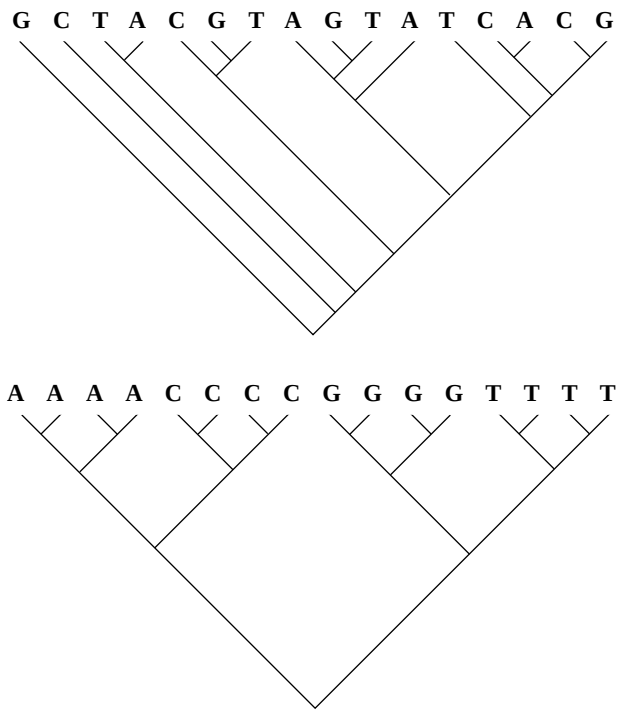


FIG. 2. Two sets of nucleotides with identical frequencies (0.25), but with different histories.

Now consider flipping a coin 10 times. In one series we obtain:

H, H, H, H, H, H, H, H, H, H.

In another series we get:

H, T, H, H, H, T, H, T, H, H.

Assuming the coin is equally weighted, which observation is more probable? The likelihoodist would hasten to judge the second to be more probable than obtaining 10 heads in a row. However, if the events are conjoined somehow, this need not be so. That is, if getting "heads" depends on what happened before it (is historically contingent), probability theory falls apart in the face of its requirement for sampling independence. What, then, if the events are wholly independent? If the coin is equally weighted, the probability of getting a heads on a toss is 0.5, and so is tails.There are 256 ways in which we could have got heads or tails in 10 successive throws, thus $p$(H, H, H, H, H, H, H, H, H, H) = 0.004 and $p$(H, T, H, H, H, T, H, T, H, H) = 0.004. They are equally probable. What if we take the likelihoodist's perspective of saying that the probability of heads, as ascertained from the second series, is 0.7? In this case, the first series of throws ($p$= 0.028) is actually *more* likely than the second ($p$ = 0.0005) from which the rule was derived. Again, this is only damning of frequency probability calculus applied to the particularness of the singular series, not of its relevance to abstract generalities like "a throw on average". Regarding phylogenetic inference, Hillis (1995:4) argued for "efficiency... measured in terms of [finding] the correct solution at a given frequency". This might be germane if, in fact, there were frequencies of history. Insofar as there is but one history for a given set of taxa, Hillis' rubric is left empty.

There is also great confusion about prediction and explanation (Scheffler, 1957), the most important being that "an explanation is not fully adequate unless it could have served as a basis for predicting the phenomenon under consideration" (Hempel and Oppenheim, 1948:137–138), which invokes a structural identity for explanation and prediction. However, they are not the same. Predictive success requires the possibility of predictive failure. That is, the dual truth of some predictor (like a model) and some phenomenon (character distributions) does not make the predictor explanatory any more than a clairvoyant's correctly predicting that something will occur and it then occurring (see also Wenzel and Carpenter, 1994). Take, for example, the gambler's fallacy: Roberto Alomar is batting 0.300. He comes to bat three times in a game and fails to get a hit. The naive gambler bets heavily on Alomar's getting a hit on the fourth at-bat, because he is "due". Our objective probabilist, like the likelihoodist, sees this differently and asserts that, because he is batting 0.300, he still has only a 30% chance of getting a hit, but this too fails to take into account the full scope of knowledge. In the first place, because Alomar failed to get a hit in his last three times at bat, he is actually batting 0.297; the probabilities have changed, because they are historically contingent phenomena[1]. More to the point, Alomar either will or he will not get a hit and there is no probability that can be assigned to that one event: betting on one event alone is foolish. Historical sciences are explanatory (or postdictive) not predictive. The explanation that best explains the explanans must necessarily be preferred.

Whether or not it is predictive of some other phenomenon is irrelevant.

Popper's (1983) critique of probabilism carried two main points demonstrating the illogic of this inductive framework. If a highly probable theory is our objective, we must certainly fail. In light of an infinite possible number of theories, each with non-zero probability, all theories must actually have a probability of zero. In the likelihood framework, any hypothesis is joint (tree + model) and there *is* an infinite number of possible models, never mind that only a certain number of models has been offered so far. It was, at one time, admitted that "there seems to be an infinite variety of possible models [and] No pretense is made that Table 1 is exhaustive" (Felsenstein, 1973:241). The likelihoodist must accept that any theory proffered must then have a probability of zero. However, the probabilist might counter in a Bernoullian appeal that this theory utters more truths than some other theory. Leaving aside the fact that this could never be demonstrated, if the frequency of truthhoods is taken to be of primary concern, the probabilist enters a quagmire. If the frequency of falsehoods stemming from a theory is non-zero, then in the infinite case there will be an infinite number of falsehoods, and in the face of an infinite number of falsehoods, surely the theory is falsified (see also Nagel, 1939). Neo-justificationists like Bernoulli, Carnap and Keynes each struggled with these logical flaws and admitted that there was no avoiding them. Felsenstein, Goldman, Hillis, Huelsenbeck, Rannala, Swofford, and Yang have neither admitted these, nor struggled with them, but they *cannot* avoid them any more than could those who came before them.

[1]Actually (Jim Carpenter, pers. comm.) whether three at-bats drop Alomar's average to 0.297 depends on where we are in the season, and basically would be a very special circumstance. Early in the season, three 0-fer's on a 0.300 average would drop the average a lot more, so his average would be well below 0.297, while late in the season the same sequence would drop the average less, so that he would still be above 0.297. For illustrative numbers, consider that a 0.300 average the first week of the season would be based on something like three hits in 10 at-bats, so the collar drops him to 0.230, while the last day in the season, where a key starter like Alomar might have 600 at-bats, that means 180 hits gets 0.300, and the 0-for-three leaves him at 0.298 (with a chance to end at 0.300 — because it's rounded up — if he gets a hit his last at-bat). Which, contract incentives being what they are these days, he might really press to do, so if he hadn't spat on any umpires recently you might bet on him to get that hit, because he'll press a little, not too much, and late in the game he's probably going to be seeing mostly fastballs anyway, even if it's a blowout, because everybody will be looking to get out of there. All of which goes to prove the point about historical contingency.

Evidence is merely that. Hume (1739) was correct in asserting that there is nothing ampliative in probability that will garner us a metaphysical truth from finite data. The probability of the evidence (*e*)

> "does not make [the hypothesis] *h* 'probable': for to say that *h* is probable is to say that it is more probable than not that *h* is true. This would mean that it is more probable than not that *h* agrees with all of the facts in the world: that there exists no counter example, no fact that contradicts it. But no finite evidence *e* can ever tell us that" (Popper, 1983:346).

There is, however, an alternative to a frequency probability of historical estimation, and this is a logical probability regarding corroboration in historical inference. When it is recognized that explanation and prediction are separate precepts in scientific understanding, then evidence and probability are seen to stand in different relations: evidence to explanation and frequency probability to prediction. Our task in the historical sciences is one of explanation, not of prediction. It matters little whether or not Caesar *probably* crossed the Rubicon. It matters only if Caesar crossing the Rubicon is the better corroborated explanation of the various explanans surrounding the historical question. Cladistic parsimony certainly can be thought of in an inductive probabilist framework (Felsenstein, 1973; Sober, 1988; Goldman, 1990), but this is a mistake. Casting cladistic parsimony in this light deviates from its ontological basis as *explanation* of character distributions *evidenced* in the taxa we use (Hennig, 1966; Farris, 1979, 1983; Frost and Kluge, 1994). Swofford et al.'s (1996:426) dismissal of cladistic parsimony in a citationless footnote is shameful.

The search for truth is certainly what drives the humanistic component of biologists' desire to discover and uncover. Our assertions regarding the terminal elusiveness of this truth may be seen by some as troubling or even nihilistic. We counter that it is the impossibility of achieving truth that ensures the continuation of scientific endeavour, and that guarantees our perpetual realization of that which is more valuable than truth itself — understanding.

## ACKNOWLEDGMENTS

The Systematics Discussion Group, Museum of Zoology, University of Michigan, provided many useful comments on the ideas contained in this paper. We especially thank Jennifer Ast, Andy Brower, Jim Carpenter, Tom DiBenedetto, Darrel Frost, Sharon Jansa, Richard Lewontin, Barb Lundrigan, Brad Moon, Olivier Rieppel, Mike Sorensen, Chris Thacker,

John Wenzel and Ward Wheeler for their comments and criticisms of earlier drafts. We additionally thank Darrel Frost for the phrase "naked operationalism" and for being singular himself, and we thank John Wenzel for his musings over what "evolutionary error" might be. Engaging disagreements with David Swofford contributed to some of our final arguments. We also greatly appreciate critical input and encouragement from the Michigan Society of Fellows. This manuscript was completed at the Cladistics Institute, Harbor Springs, Michigan.

## REFERENCES

Barry, D. and Hartigan, J. A. (1987). Statistical analysis of hominoid molecular evolution. *Stat. Sci.* **2**, 191-210.

Bartlett, M. S. (1962). "Essays on Probability and Statistics". John Wiley, New York.

Berkeley, G. (1710). "The Principles of Human Knowledge". (1937 reprint, A. Brown and Sons, London).

Bernoulli, J. (1713). "Ars Conjectandi". Basiliae.

Brady, R. H. (1983). Parsimony, hierarchy, and biological implications. In "Advances in Cladistics", (N. I. Platnick, and V. A. Funk, eds), vol. 2, pp. 49-60. Columbia University Press, New York.

Braithwaite, R. B. (1953). "Scientific Explanation: A study of the Function of Theory, Probability and Law in Science". Cambridge University Press, Cambridge.

Carmean, D., and Crespi B. J. (1995). Do long branches attract flies? *Nature* **373**, 666.

Carnap, R. (1956). The methodological characterization of theoretical concepts. In "Minnesota Studies in the Philosophy of Science". (H. Feigl, and M. Scriven, eds.), pp. 38-76. University of Minnesota Press, Minneapolis.

Carpenter, J. M. (1988). Choosing among multiple equally parsimonious cladograms. *Cladistics* **4**, 291-296.

Cohen, L. L. (1984). "Darwin was Wrong — A Study in Probabilities". New Research Publications, Greenvale, New York.

Cramer, H. (1946). "Mathematical Methods of Statistics". Princeton University Press, Princeton.

De Queiroz, K. (1992). Phylogenetic definitions and taxonomic philosophy. *Biol. Philos.* **7**, 295-313.

Edwards, A. W. F. (1992). "Likelihood". Expanded edition. Johns Hopkins University Press, Baltimore.

Efron, B., and Gong. G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.*, **37**, 36-48.

Ereshefsky, M. (1991). Species, higher taxa, and the units of evolution. *Philosophy of Sci.* **58**, 84-101.

Einstein, A. (1950). Physics and Reality. In "Out of my Later Years", pp. 59-65. Philosophical Library, New York.

Eyre-Walker, A. C. (1991). An analysis of codon usage in mammals: Selection or mutation bias? *J. Mol. Evol.* **33**, 442-453.

Faith, D. P. (1991). Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* **40**, 366-375.

Farris, J. S. (1969). A successive approximations approach to character weighting. *Syst. Zool.* **18**, 374-385.

Farris, J. S. (1973). A probability model for inferring evolutionary trees. *Syst. Zool.* **22**, 250-256.

Farris, J. S. (1979). The information content of the phylogenetic system. *Syst. Zool.* **28**, 483-519.

Farris, J. S. (1983). The logical basis of phylogenetic analysis. In "Advances in Cladistics", (N. I. Platnick, and V. A. Funk, eds), vol. 2., pp. 7-36. Columbia University Press, New York.

Farris, J. S. (1986). On the boundaries of phylogenetic systematics. *Cladistics* **2**, 14-27.

Farris, J. S., Kallersjo, M., Albert, V. A., Allard, M., Anderberg, A., Bowditch, B., Bult, C., Carpenter, J. M., Crowe, T. M., Delaet, J., Fitzhugh, K., Frost, D., Goloboff, P., Humphries, C. J., Jondelius, U., Judd, D., Karis, P. O., Lipscomb, D., Luckow, M., Mindell, D., Muona, J., Nixon, K., Presch, W., Seberg, O., Siddall, M. E., Struwe, L., Tehler, A., Wenzel, J., Wheeler, Q., and Wheeler, W. (1995). Explanation. *Cladistics* **11**, 211-218.

Farris, J. S., Albert, V. A., Kallersjo, M., Lipscomb, D., and Kluge, A. G. (1996). Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12**, 99-124.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**, 240-249.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quart. Rev. Biol.* **57**, 379-404.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783-791.

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Ann. Rev. Gen.* **22**, 521-565.

Fisher, R. (1938). Comments on H. Jeffrey's "Maximum Likelihood, Inverse Probability, and the Method of Moments". *Ann. Eugenics* **8**, 146-151.

Foster, S. A., Cresko, W. A., Johnson, K. P., Tlusty, M. U., and Willmott, H. E. (1996). Patterns of homoplasy in behavioral evolution. In "Homoplasy: The Recurrence of Similarity in Evolution". (M. J. Sanderson, and L. Hufford, eds), pp. 245-269. Academic Press, San Diego.

Frost, D. R., and Kluge, A. G. (1994). A consideration of epistemology in systematic biology, with special reference to species. *Cladistics* **10**, 259-294.

Gaut, B. S., and Lewis, P. O. (1995). Success of maximum likelihood in the four-taxon case. *Mol. Biol. Evol.* **12**, 152-162.

Gillespie, J. H. (1991). "The Causes of Molecular Evolution". Oxford University Press, New York.

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.

Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics* **9**, 83-91.

Hacking, I. (1965). "The Logic of Statistical Inference". Cambridge University Press, Cambridge.

Harvey, P. H., and Pagel, M. D. (1991). "The Comparative Method in Evolutionary Biology". Oxford University Press, New York.

Hedges, S. B., and Maxson, L. R. (1996). Re: Molecules and morphology in amniote phylogeny. *Mol. Phylogen. Evol.* **6**, 312-314.

Hempel, C. G., and Oppenheim, P. (1948). Studies in the logic of explanation. *Philos. Sci.* **15**, 135-175.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38**, 297-309.

Hennig, W. (1966). "Phylogenetic Systematics". University of Illinois Press, Chicago.

Hillis, D. M. (1995). Approaches for assessing phylogenetic accuracy. *Syst. Biol.* **44**, 3-16.

Hillis, D. M., Huelsenbeck, J. P., and Swofford, D. L. (1994). Consistency: Hobgoblin of phylogenetics? *Nature* **369**, 363-364.

Huelsenbeck, J. P. (1996). Phylogenetic methods. pp. [1-8]. http://mw511.biol.berkeley.edu/john/lecture.html

Huelsenbeck, J. P. (1997). Is the Felsenstein Zone a fly trap? *Syst. Biol.* **46**, 69-74.

Huelsenbeck, J. P., and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42**, 247-264.

Hume, D. (1739). "Treatise of Human Nature". (L. A. Selby-Bigge, ed.) Clarendon Press, Oxford.

Kant, I. (1781). "Critique of Pure Reason". Konisberg. (1881 reprint, Macmillan Company, London).

Kim, J. (1996). General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* **45**, 363-374.

Kluge, A. G. (1990). Species as historical individuals. *Biol. and Philos.* **5**, 417-431.

Kluge, A. G. (1997). Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics* **13**, 81–96.

Kluge, A. G., and Wolf, A. J. (1993). Cladistics: What's in a word? *Cladistics* **9**, 183-199.

Knight, A., and Mindell, D. P. (1993). Substitution bias, weighting of DNA sequence evolution, and the phylogenetic position of Fea's viper. *Syst. Biol.* **42**, 18-31.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. *In* "Criticism and the Growth of Knowledge". (I. Lakatos and A. Musgrave, eds.), pp. 91-196. Cambridge University Press, Cambridge.

Larson, A., and Losos, J. B. (1996). Phylogenetic systematics of adaptation. *In* "Adaptation". (M. R. Rose, and G. V. Lauder eds.), pp. 187-220. Academic Press, New York.

Lindgren, B. W. (1976). "Statistical Theory". 3rd edition. Macmillan, New York.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11**, 605-612.

Lynn, J., Harrell, F., Cohn, F., Wagner, D., and Connors, A. F., (1997). Prognosis of seriously ill hospitalized patients on the days before death: Implications for patient care and public policy. *New Horizons* **5**, 56-61.

Mill, J. B. (1878). "Analysis of the Phenomena of the Human Mind". New edition, with notes illustrative and critical by A. Bain, A. Findlater, and G. Grote. Edited with additional notes by J. S. Mill. Longmans, Green, Reader, and Dyer, London.

Miyamoto, M. M. (1985). Consensus cladograms and general classifications. *Cladistics* **1**, 186-189.

Miyamoto, M. M., and Fitch, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **44**, 64-76.

Nagel, E. (1939). Principles of the theory of probability. *In* "The International Encyclopedia of Unified Science", Vol. 1, no. 6., pp. 62-75. University of Chicago Press, Chicago.

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

Penny, D., Hendy, M. D., and Steel, M. A. (1991). Testing the theory of descent. In "Phylogenetic analysis of DNA sequences" (M. M. Miyamoto, and J. Cracraft, eds) pp. 155-183. Oxford University Press, Oxford.

Penny, D., Hendy, M. D., and Steel, M. A. (1992). Progress with methods for constructing evolutionary trees. *Trends Ecol. Evol.* **7**, 73-78.

Popper, K. (1959). "The Logic of Scientific Discovery". Harper and Row, New York.

Popper, K. (1963). "Conjectures and Refutations: The Growth of Scientific Knowledge". Routledge and Kegan Paul, London.

Popper, K. (1983). "Realism and the Aim of Science". Routledge, London.

Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis* **9**, 313-328.

Ranalla, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304-311.

Reichenbach, H. (1951). "The Rise of Scientific Philosophy". University of California Press, Berkeley.

Russell, B. (1946). "History of Western Philosophy, and its Connection with Political and Social Circumstances from the Earliest Times to the Present Day". Simon and Schuster, New York.

Saether, O. A. (1986). The myth of objectivity — post-Hennigian deviations. *Cladistics* **2**, 1-13.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for constructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-422.

Scheffler I. (1957). Explanation, prediction and abstraction. *In* "Philosophy of Science" Danto, A., and Morgentaler, S., eds.), pp. 274-287. The World Publishing Company, New York.

Siddall, M. E., and Whiting, M. F. (1998). Long branch abstractions. *Cladistics* (submitted).

Sober, E. (1988). "Reconstructing The Past: Parsimony, Evolution, and Inference". MIT Press, Cambridge, Massachusetts.

Sober, E. (1993). "Philosophy of Biology". Westview Press, San Francisco.

Steel, M. A., Hendy, M. D., and Penny, D. (1993 a). Parsimony can be consistent! *Syst. Biol.* **42**, 581-587.

Steel, M. A., Szekely, L., Erdos, P. L., and Waddell, P. J. (1993 b). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *N. Z. J. Bot.* **31**, 289-296.

"Student". (1908). The probable error of a mean. *Biometrika* **6**, 1-25.

Swofford, D. L., and Berlocher., S. H. (1987). Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Syst. Zool.* **36**, 293-325.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics" ( D. M. Hillis, C. Moritz, and B. K. Mable, eds), pp. 407-514. Sinauer Associates, Sunderland, Massachusetts.

Tateno, Y., Takezaki, N., and Nei, M. (1994). Relative efficiencies of the maximum- likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11**, 261-277.

Takezaki, N., and Nei, M. (1994). Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *J. Mol. Evol.* **39**, 210-218.

Thompson, E. A. (1975). "Human Evolutionary Trees". Cambridge University Press, Cambridge.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* **20**, 595-602.

Watkins, J. (1984). "Science and Scepticism". Princeton University Press, Princeton.

Wenzel, J. W., and Carpenter, J. M. (1994). Comparing methods: Adaptive traits and tests of adaptation. *In* "Phylogenetics and Ecology" ( P. Eggleton, and R. I. Vane-Wright, eds), pp. 79-101. Linnean Society Symposium Series, no. 17, Academic Press, London.

Weyl, H. (1949). "Philosophy of Mathematics and Natural Science", (translation by Olav Helmer). Princeton University Press, Princeton.

Whiting, M. F., Carpenter, J. C., Wheeler, Q. D., and Wheeler, W. C. (1997). The Strepsiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Syst. Biol.* **46**, 1-68.

Wiens, J. J. (1995). Polymorphic characters in phylogenetic systematics. *Syst. Biol.* **44**, 482-500.

Williams, P. L., and Fitch, W. M. (1990). Phylogeny determination using dynamically weighted parsimony method. *Methods Enzymol.* **183**, 615-626.

Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**, 1396-1401.

Yang, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**, 294-307.

Zharkikh, A, and Li., W. H. (1993). Inconsistency of the maximum parsimony method: The case of five taxa with a molecular clock. *Syst. Biol.* **42**, 113-125.