AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 94:143–156 (1994)

Cladistic Analysis of Molecular and Morphological Data

BRENT D. MISHLER
Department of Integrative Biology, University and Jepson Herbaria, University of California, Berkeley, California 94720

KEY WORDS Systematics, Homology, Branch lengths, Hennig, Felsenstein, Phylogenetic reconstruction

ABSTRACT Considerable progress has been made recently in phylogenetic reconstruction in a number of groups of organisms. This progress coincides with two major advances in systematics: new sources have been found for potentially informative characters (i.e., molecular data) and (more importantly) new approaches have been developed for extracting historical information from old *or* new characters (i.e., Hennigian phylogenetic systematics or cladistics). The basic assumptions of cladistics (the existence and splitting of lineages marked by discrete, heritable, and independent characters, transformation of which occurs at a rate slower than divergence of lineages) are discussed and defended. Molecular characters are potentially greater in quantity than (and usually independent of) more traditional morphological characters, yet their great simplicity (i.e., fewer potential character states; problems with determining homology), and difficulty of sufficient sampling (particularly from fossils) can lead to special difficulties. Expectations of the phylogenetic behavior of different types of data are investigated from a theoretical standpoint, based primarily on variation in the central parameter λ (branch length in terms of expected number of character changes per segment of a tree), which also leads to possibilities for character and character state weighting. Also considered are prospects for representing diverse yet clearly monophyletic clades in larger-scale cladistic analyses, e.g., the exemplar method vs. "compartmentalization" (a new approach involving substituting an inferred "archetype" for a large clade accepted as monophyletic based on previous analyses). It is concluded that parsimony is to be preferred for synthetic, "total evidence" analyses because it appears to be a robust method, is applicable to all types of data, and has an explicit and interpretable evolutionary basis.

The central importance of phylogeny reconstruction in systematics, ecology, and evolutionary biology has become widely realized in recent years (Brooks and McLennan, 1991; Donoghue, 1989; Funk and Brooks, 1990; Harvey and Pagel, 1991; Huey, 1987; Wanntorp et al., 1990). Explicit cladistic phylogenies now provide a critical basis for classification as well as for studies of speciation, biogeography, ecology, and behavior (among many other areas).

Partly as cause (and partly as effect) of this realization, considerable progress has been made in phylogenetic reconstruction in a number of major groups of organisms. This progress coincides with two major advances in systematics: 1) new approaches to extracting historical information from characters (e.g., cladistics) and 2) new sources of potentially informative characters (i.e., molecular data). The question of which of these two advances is primarily responsible for progress in understanding phylogeny is often debated in diverse arenas ranging from purely academic to pragmatic concerns of

Received April 5, 1993; accepted November 12, 1993.

funding and hiring. I argue for the primacy of the first listed advance; the real breakthrough was the development of a logical framework (based on the Hennig Principle, more on which below) wherein the phylogenetic meaning of any data can be evaluated.

Co-incident with this development, and the theoretical advances it has entailed, the empirical data base of systematics has been augmented by various forms of molecular characters. After a rather naive, overly optimistic initial phase, wherein it was felt that such data would provide a panacea for systematics (Gould, 1985), a careful look is now being taken at exactly how historical information in molecules is to be discovered and used, and how it is expected to perform relative to more traditional morphological characters (Donoghue and Sanderson, 1992; Mishler, et al., 1988; Patterson, 1987). We have reached a stage where we can begin to see how the new molecular data should work (and in fact how they do work) in reconstructing phylogenetic relationships among organisms.

I begin by summarizing two different approaches to phylogeny reconstruction, then discuss features of molecular data bearing on their suitability for such approaches. I conclude by examining prospects for comparative analyses of molecular and morphological data, both in terms of theoretical expectations and empirical practice.

APPROACHES TO PHYLOGENY RECONSTRUCTION

Two different intellectual traditions in evolutionary biology have recently converged on the production of phylogenetic trees. Both traditions have figured in molecular systematics, yet until recently they have remained estranged. However, conditions are currently favorable for a "hybrid vigor" effect, since ideas from one area are being considered and used in the other.

The phylogenetic systematics tradition

This tradition is derived from comparative anatomy and morphology, focused on the implications of individual homologies (landmark works in this tradition include Hennig, 1966; Zimmermann, 1967). This tradition tends to conceive of the inference process as one of reconstructing history following deductive-analytic procedures (Farris, 1983; Gaffney, 1979; Nelson and Platnick, 1981; Wiley, 1981). The goal is seen as coming up with the best supported hypothesis to explain a unique past event.

Hennig's seminal contribution was to note that in a system evolving via descent with modification and splitting of lineages, characters that changed state along a particular lineage can serve to indicate the prior existence of that lineage, even after further splitting occurs. The "Hennig Principle" follows from this: homologous similarities among organisms come in two basic kinds, *synapomorphies* due to immediate shared ancestry (i.e., a common ancestor at a specific phylogenetic level), and *symplesiomorphies* due to more distant ancestry. "Special similarities" (synapomorphies; taxic homologies) are the key to reconstructing truly natural relationships of organisms, rather than overall similarity (which is an incoherent mixture of synapomorphy, symplesiomorphy, and nonhomology). In the Hennigian system then, individual hypotheses of putative homology are built up on a character-by-character basis, then a congruence test (using a parsimony principle, more on which below) is applied to identify *homoplasies* (i.e., apparent homologies that are not congruent with the plurality of characters), and finally, strictly monophyletic classifications are applied to the resulting branching diagram *(cladogram)*. This elegant correspondence between synapomorphy, homology, and monophyly is the basis of the cladistic revolution in systematics.

Certain fundamental assumptions must be made in order to justify the use of cladistic parsimony in phylogenetic reconstruction. There are five basic assumptions necessary: 1) Reproduction (replication in the sense of Brandon, 1990; Hull, 1980) must occur to form *lineages* (the diachronic ancestor-descendent relationship). 2) Heritability (in the population genetic sense) must obtain, wherein particular features to be used as historical markers *(characters)* have discrete variants (*character states* empirically, *transformational homologs* ontologically) that show a strong correlation between parent and offspring. 3) Divergence (branching of lineages) must occur predominantly, as compared to reticulation, giving rise to patterns of *taxic homologs* (in the sense of Patterson, 1982) shared among *sister groups* (the synchronic monophyly relationship). 4) Independence must occur among different characters; i.e., no process (e.g., natural selection, gene conversion, developmental constraints) is operating to produce nonhomologous character associations that overwhelm taxic homologs indicating common history. 5) Transformation in particular characters must occur at a relatively low rate (see below for a discussion of just how low), as compared to divergence.

In sum, these five basic criteria amount to a joint assumption that an apparent taxic homology (N.B., this is a feature that has already passed strict observational and experimental tests of detailed similarity, heritability, and independence) is more likely to be due to true taxic homology than to homoplasy, unless evidence to the contrary exists, i.e., a plurality of apparent taxic homologies showing a different pattern (Funk and Brooks, 1990). Thus, cladistic analysis does not rely on an assumption that evolution proceeds parsimoniously (contra Cronquist, 1987), in fact, the application of this method of analysis provides the only rigorous basis we have for identifying homoplasy and thus demonstrating nonparsimonious evolution (Farris, 1983).

The population genetic tradition

This tradition is derived from studies of the fate of genes in populations. This tradition tends to see phylogenetic inference as a statistical estimation problem (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1988). The goal is seen to be choosing a set of trees out of a statistical universe of possible trees, while putting confidence limits on the choice. Molecular systematics was initially derived mainly from this tradition (Nei et al., 1983; Sogin et al., 1986; Wilson et al., 1977; Woese and Fox, 1977; Zurawski and Clegg, 1987). Probably because of the modern synthesis emphasis on genetics and the gradual divergence of lineages (Gould, 1982), overall genetic distance measures are often used to indicate "relationship." However, the recognition has been growing that such phenetic methods are at best able to mimic the results of a cladistic analysis, and that the disadvantages of the former approach are many, including: 1) such approaches usually assume a molecular clock, but empirical studies have demonstrated that such an assumption is often violated (Cracraft, 1987; Mindell and Honeycutt, 1990); 2) many distance measures used are nonmetric, therefore one can't interpret branch lengths evolutionarily (Farris, 1981); 3) distance measures hide homoplasy when character-by-character differences are boiled down into a single difference value (Farris, 1983). The bottom line is that distance methods throw away much information on individual characters that is so laboriously obtained (Penny et al., 1992). After all, evolution does mechanistically proceed by individual mutations, not by change in some measure of overall distance.

While character-based methods seem to be in the ascendancy, the jury is still out on the applicability of various statistical approaches (or even the desirability of such approaches). Issues under debate include the nature of the statistical universe being sampled and exactly what evolutionary assumptions are safe to use in hypothesis testing. Under standard views of hypothesis testing, one is interested in evaluating an estimate of some real but unknown parameter, based on samples taken from a relevant class of individual objects (the statistical universe). It might be argued that a particular phylogeny is one of many possible topologies, thus somehow one might talk about the probability of existence of that topology or of some particular branches. However, phylogenies are unique historical events ("individuals" in the sense of Hull, 1980); a particular phylogeny clearly is a member of a statistical universe of one. It is of course valid to try to set a frequency-based probability for such phylogenetic questions as: How often should we expect to find completely pectinate cladograms? or How often should we find a clade as well supported as the mammals? In such cases, there is a valid reference class ("natural kind" in the sense of Hull, 1980) about which one can attempt an inference.

It could be reasonably argued that characters in a particular group of organisms are sampled from a universe of possible characters (Felsenstein, 1985). The counterargument, however, is that characters are chosen based on a refined set of criteria of likely informativeness, e.g., presence of discrete states, invariance within operational toxonomic units (OTUs), and ability to determine potential homology (including alignability for molecular data). Therefore, the characters are at best a highly nonrandom sample of the possible descriptors of the organisms. It may perhaps be better not to view characters as a sample from a larger universe at all—a data matrix is (or at least should be) *all* the "good" characters available to the systematist.

My own view is that statistical considerations primarily enter systematics during the phase sometimes called "character analysis" (Neff, 1986), that is when the data matrix is being assembled. Based on expectations of "good" phylogenetic markers (characters), procedures have been developed that involve assessing the likely independence and evolutionary conservatism of potential characters using experimental and statistical manipulations (see discussion in Mishler and De Luna, 1991; Neff, 1986; Wiley, 1981). By the time a matrix is assembled, each column can be regarded as an independently justified hypothesis about phylogenetic grouping, an individual piece of evidence for the existence of a monophyletic group (a putative taxic homology). The parsimony method used to produce a cladogram from a matrix should then be viewed as a *solution* of that matrix, an analytic transformation of the information contained therein from one form to another, just as in the solution of a set of linear equations. No inductive, statistical inference has been made at that step, only a deductive, mathematical one. Now to assert that the resulting cladogram represents a model of a phylogenetic tree is another matter, an inductive inference requiring separate justification.

Maximum likelihood techniques remain the preferred statistical approach for such problems, at least for many workers from the population genetics tradition (e.g., Felsenstein, 1981, 1982; Penny et al., 1992). A maximum likelihood approach attempts to evaluate the probability of observing a particular set of data, given an underlying phylogenetic tree. Among competing phylogenetic trees, the most believable (likeliest) tree is one that makes the observed data most probable. To make such a connection between data and trees, it is necessary to have auxiliary assumptions about such parameters as the rate of character change, the length of branches, the number of possible character states, and relative probabilities of change from one state to another. The primary debate has involved these assumptions—how much is necessary or desirable or possible to assume about evolution before a phylogeny can be established? Sober (1988) has shown convincingly that *some* evolutionary assumptions are necessary to justify any method of inference, but he (and the field in general) remains unclear about exactly *what* the minimum assumptions are.

It seems generally agreed that only the fewest and least controversial assumptions should be used. Given its assumptions as discussed above, the Wagner parsimony method appears to give a robust connection between data and preferred tree(s). In other words, assuming that characters are heritable and independent, and that changes in state are relatively slow as compared to branching events in a lineage, reconstructions for a character showing one change on a branch will be more likely than reconstructions showing more than one change elsewhere. For example, this method has recently been shown to work well in reconstructing a known bacteriophage phylogeny (Hillis et al., 1992).

In my opinion, the basic framework for future work will remain Hennigian, but with significant input and justification from the population genetics tradition. A growing group of molecular systematists (e.g., Albert et al., 1992; Donoghue and Sanderson, 1992; Doyle, 1992; Hillis et al., 1992; Mindell, 1991; Mishler et al., 1988) are deeply concerned with adapting standard methods of cladistic analysis to molecular data. In applying methods of phylogenetic analysis to such data most of the issues discussed for years with regard to morphological and anatomical data are involved, such as: the nature of homology, including alignment problems; defining independent characters and character states (the problem of functional, mutational, and developmental constraints); and weighting issues (e.g., gains vs. losses in restriction site data, transitions vs. transversions in nucleotide data, compensatory substitutions in RNA due to secondary structure).

SPECIAL FEATURES OF MOLECULAR DATA

Various techniques have been developed for molecular systematic studies. Some of these, such as DNA hybridization and restriction fragment length polymorphisms (RFLPs), give only distance information. Other techniques, including mapping of restriction enzyme sites, direct RNA and DNA sequencing methods (initially via cloning but more recently via the polymerase chain reaction [PCR]) yield information about specific characters, suitable for cladistic analysis. It is the latter techniques I will focus on here, because of the arguments given above in favor of character-based methods for phylogeny reconstruction.

Potential advantages to the molecular data include: 1) a huge number of characters, especially useful in organisms with simple morphology and chemistry; and 2) the ability to homologize across very broad groups, at least in highly conserved genes. Other purported advantages to molecular data need further examination, because they are probably unfounded.

Frequently cited (but controversial) advantages to molecular data include the idea that it is somehow closer to (or even equivalent to) *the* genetic information (Sibley and Ahlquist, 1986). This appears to be a reductionistic hold-over from the modern synthesis point of view in population genetics that "genetic relationship" is synonymous with "phylogenetic relationship." Some authors have even equated "phenetic" with phenotypic (morphological) and "genetic" with phylogenetic (e.g., Manhart and Palmer, 1990). This intuitively attractive idea turns out to be nonsensical upon further consideration. Based on recent considerations of developmental processes, it has become clear that "information" in biological lineages is not passed down solely (or even mainly) by raw DNA sequences (Nijhout, 1990). Whole cytological systems are inherited, often considerably influenced by a maternal environment. Historical information may be preferentially preserved in epigenetic developmental systems that are buffered against considerable variation at the individual nucleotide level.

Another questionable argument for the superiority of molecular data is their independence from morphological characters that are supposedly more subject to adaptive convergence (Bobrova et al., 1987; Olmstead, 1989). It is indeed true that the highly canalized epigenetic systems referred to above may in some cases be quite subject to parallelism. However, constraints that affect the independence of characters may operate at any level. Selective constraints can be very strong at the molecular level; in fact, sometimes stronger than anything that would be seen at the morphological level (e.g., consider the likely selection coefficient associated with having functional ribosomes as compared to having a particular petal shape). This is particularly true for those highly conserved genes that give us "the ability to homologize across very broad groups." There are also nonadaptive constraints at the molecular level that are at least somewhat analogous with developmental constraints for morphological data; these include error-correcting mechanisms, secondary and tertiary structural constraints, etc. Again, the resolution of these possibilities in any particular case will require investigation of congruence in many different characters and character systems. The hope is that if one looks at enough characters, all perhaps subject to varying degrees of natural selection or other constraints, the various noise-producing factors will "cancel-out" and a common historical signal can be detected.

Clear weaknesses of molecular level characters include the simplicity of characters; they may have no ontogeny (in the case of DNA) and few possible character states, leading to special problems with homoplasy. The alignment phase in a molecular analysis is equivalent to the morphologist's initial hypotheses of homology (Mindell, 1991; Mishler et al., 1988), yet there is not as much information to go on. An adenine residue at the "same" position in different taxa must be scored as the same, while superficially identical morphological states can sometimes be separated by detailed ultrastructural or developmental study. Molecules come basically as a one-dimensional string (albeit sometimes with additional information available from codon position or secondary structure models). Much further work remains to be done on the logic behind the alignment phase, taking advantage of feed-back from preliminary phylogenetic knowledge, much in the way morphologists have always worked.

There are problems with sampling at the molecular level—it is time consuming and expensive to sample within OTUs (to check for polymorphisms) at the level that is possible for many morphological characters, and of course fossil taxa generally cannot be included. The need for sampling the true phylogeny evenly has been discussed by Swofford and Olsen (1990). Fossils may be of critical importance for this (Donoghue et al., 1989), because of the need to reduce asymmetries in branch lengths and detect more character state changes (more on which below; see also Albert et al., 1992).

MOLECULES AND MORPHOLOGY: PROSPECTS

The necessity of using all data

It is clear that all appropriate data should be included in an analysis, whether one's own focus has been exclusively on morphology or on molecules. It makes no sense to ignore older data just because newer data have been generated (or vice versa). This calls into question the basis for the field of "molecular systematics" (wherein whole journals exist whose apparent purpose is to present new molecular phylogenies to the exclusion of all previous data, even earlier molecular data!). Clearly the field should be "systematics" in an inclusive sense. However, it is far from clear how different data sets are to be evaluated and compared (Swofford, 1991). Some have argued that data sets derived from fundamentally different sources should be analyzed separately, and only common results taken as well supported (i.e., consensus tree approaches). Others have argued that all putative homologies should be combined into one matrix (perhaps with unequal weighting if this can be independently justified). Theoretical arguments at present favor the latter approach (i.e., "total evidence"; Barrett et al., 1991; Donoghue et al., 1989; Doyle, 1992; Kluge, 1989; Miyamoto, 1985). If characters have been independently judged to be candidates as taxic homologies (i.e., with detailed similarity across OTUs, discrete states, heritability, and independence; Mishler and De Luna, 1991; Wiley, 1981), then they are equivalent and should be analyzed together. However, the availability of fast personal computers and efficient software means that the prudent course of action is always to explore both approaches, even if the latter is favored. The importance of this more ecumenical procedure is shown by the existence of cases where one particular data set exhibits a clear and explainable discordance with other data sets. The best examples of such discordance are organellar genomes that may have different phylogenies (because of hybridization and lineage sorting) than those of associated nuclear genomes and morphologies (Doyle, 1992; Rieseberg and Soltis, 1991; Smith and Sytsma, 1990).

Many examples exist of congruence between morphological and molecular data (i.e., a clade that is strongly supported by both kinds of data), while few examples exist of incongruence between *strong* results of molecular vs. morphological data sets. The few examples of the latter type that do exist seem to be explainable by the processes mentioned at the end of the previous paragraph that can result in discordance between organellar and species phylogenies (basically, both phylogenies may be correct, but the organeller phylogeny is not tightly coevolving with the phylogeny of the rest of the organism). Most examples of incongruence between morphology and molecules instead result either from weakness in both data sets or from one data set strongly resolving a clade about which the other data set is weak (i.e., not decisive in the sense of Goloboff, 1991). Based on their comparative

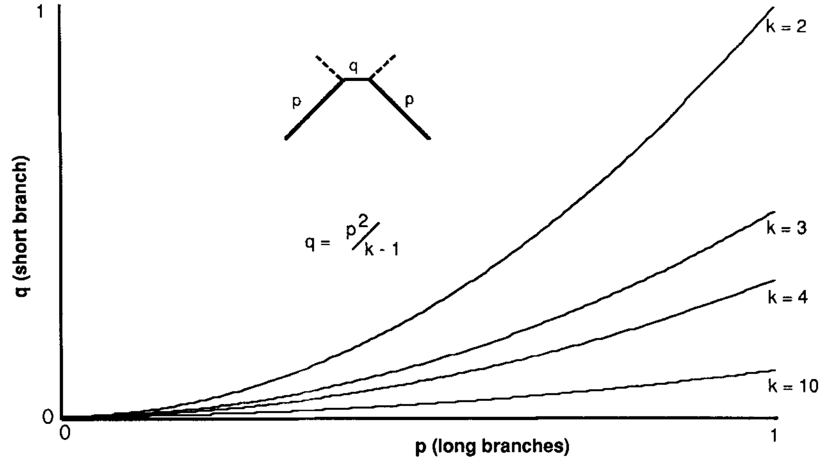

Fig. 1. The relationship between number of character states and size of the Felsenstein Zone. The probability of change in a short branch is indicated by q; the probability of change in an adjacent long branch is indicated by p (k = number of character states). The inset network shows the position of these branches relative to each other. The Felsenstein Zone (i.e., the region where parsimony will give inconsistent results) is below and to the right of the line given by the equation $q = p^2/(k - 1)$. Note that with increasing numbers of character states, a larger asymmetry in these probabilities can be tolerated (based on discussion in Albert et al., 1992).

surveys of the literature, Sanderson and Donoghue (1989) detected no indication of any overall trend in the relative phylogenetic utility of morphology vs. molecules.

The central parameter λ and its properties

The best way to predict phylogenetic behavior of characters (that otherwise meet the criteria of detailed similarity, heritability, and independence) is by examining variation in the central parameter λ, the quantitative measure of assumption 5 above, defined as *branch length* in terms of expected number of character changes per branch (segment) of a tree (following Albert et al., 1992; DeBry and Slade, 1985). The advantage of using this parameter rather than the more commonly used "rate of character change per unit time" is that the former measure incorporates both rate of change per unit time and the length of time over which the branch existed. Thus, a high λ can be due to either a high rate of change or a historically long branch (both have an equivalent effect on parsimony reconstruction). This parameter, either for a single character, or averaged over a number of characters ($\bar{\lambda}$) defines a "window of informativeness" (Mishler et al., 1988) for those data. In other words, a very low value of $\bar{\lambda}$ indicates data with too few changes on each segment to allow all branches to be discovered; this would result in polytomies in reconstructions because of too little evidence. Too high a value of $\bar{\lambda}$ indicates data that are changing so frequently that problems arise with homoplasy through multiple changes in the same character. At best a high λ causes erasure of historical evidence for the existence of a branch, at worse it creates "evidence" for false branches through parallel origins of the same state.

The effects of differential λ values have been investigated by several workers. In an important early paper, Felsenstein (1978) showed that branch length asymmetries within a tree can cause parsimony reconstructions to be inconsistent. That is, if the probability of a parallel change to the same state in each of two long branches is greater than the probability of a single change in a short connecting branch (see inset network in Fig. 1), then the two long branches will tend to falsely "attract" each other in parsimony reconstructions using a large number of characters (see also Lake, 1987; Sober, 1988). The region where branch length asymmetries will tend to cause such problems has been called the "Felsenstein Zone" (e.g., Albert et al., 1992). The seriousness of this problem (i.e., the size of the Felsenstein Zone) is affected by several factors, the most important of which are: i) the number of possible character states per character; and ii) the overall rate of change of characters.

The relationship between number of character states (k) and size of the Felsenstein Zone is shown in Figure 1. Felsenstein investigated the most extreme case, where $k = 2$. However, it stands to reason that the problem with parallelism will be most acute when the number of possible changes is highly restricted. It can be seen in Figure 1 that the Felsenstein Zone rapidly shrinks as more character states are added. This difference is of course relevant to the comparison of molecular and morphological data; with respect to this factor alone, one would be most suspicious of two-state characters such as restriction site maps, less so of nucleic acid sequence data, and even less so of most morphological data (how many potential distinct states are there for the shape of a bone?).

A second factor has been investigated using λ-based probabilities (Albert et al., 1992, 1993; Albert and Mishler, 1992). The allowable asymmetry between different branches itself varies as a function of overall expectation of character change ($\bar{\lambda}$); when $\bar{\lambda}$ is very small (<0.01), the λs for individual branches can vary by several orders of magnitude, but when $\bar{\lambda}$ is large (approaching 1.0) the individual λs must be nearly equal. Based on this consideration, these authors suggested that data should have an expected $\bar{\lambda}$ of less than about 0.1 to be considered reliable phylogenetic markers. They point out that this is an intuitively reasonable rule-of-thumb; it is hard to imagine a phylogenetically informative data set where the characters were varying at a rate approaching one change in every character on every branch.

What is the theoretical relationship between λ and homoplasy (see also discussion in Albert et al., 1993)? One common measure of homoplasy over a tree is the ensemble consistency index (CI), i.e., the ratio between the theoretical minimum number of changes for a set of characters and the actual number of changes (M/S using the symbolism of Kluge and Farris, 1969). The expected CI for a set of data is constant for any given $\bar{\lambda}$ (i.e., it is not affected by either number of characters or number of taxa). Given the way in which λ is defined (expected number of changes per character per branch), there is an expectation of a certain number of changes, regardless of how many branches or characters there are. The expected homoplasy (number of characters that will change more than once on a tree) is likewise constant for any given $\bar{\lambda}$. This can be shown in the following way:

If I = number of characters, and N = number of branches (which for an unrooted tree is itself $2n\text{-}3\text{-}p$, where n = the number of taxa in the analysis and p = the number of branches in excess of three per node), then for any given $\bar{\lambda}$ the number of state changes expected in a data set (S) is: $IN\lambda$.

The expected number of characters changing on a tree no times, one time, two times, etc. can be derived from the Poisson distribution (as is appropriate for low probability of change, assuming a low variance in λs for different characters):

IN (total chances for character change) $= INe^{-\lambda}$ (i.e., the number of characters expected not to change) $+ IN\lambda e^{-\lambda}$ (i.e., the number of characters expected to change once) $+ (1/2)IN\lambda^2e^{-\lambda}$ (i.e., the number of characters expected to change twice) $+ \ldots$

The number of characters that would be observed to be variable (M) is thus:

$$M = IN - INe^{-\lambda}$$

CI for binary data can thus be approximated as the ratio between M and the actual number of state changes (S):

$$CI = \frac{M}{S} = \frac{IN - INe^{-\lambda}}{IN\lambda} = \frac{1 - e^{-\lambda}}{\lambda}$$

CI for good data ($\bar{\lambda} < 0.1$) should thus always be greater than 95%, calculated over all characters (i.e., including autapomorphies). This is assuming, of course, that the expectation of changes follows the Poisson distribution. In real data sets this often appears not to hold (Langley and Fitch, 1974); if certain characters have a higher intrinsic probability of change, then there will be an excess of characters that show multiple changes. This probably accounts for the frequent observation of lower CIs, even with "good" data. This effect is exaggerated when CIs are calculated "excluding autapomorphies," as is commonly done. In terms of judging the general suitability of a data set for a particular problem, it appears better to calculate CI across *all* characters. Furthermore, if differential probabilities of change can be identified and taken into account via weighting (see below), the CI of a weighted analysis would be expected to be higher than an unweighted one.

As $\bar{\lambda}$ goes up, the expected CI will go down. There is a common observation that increasing the number of OTUs in an analysis tends to decrease CI (Goloboff, 1991; Klassen et al., 1991; Sanderson and Donoghue, 1989). This effect may be largely explainable as a tendency for an increased $\bar{\lambda}$ as OTUs are added. However, such an effect would be expected only if the branches added are "outside" the true tree shared by the original, smaller set of OTUs (i.e., if the added OTUs make the analysis more inclusive phylogenetically). If new segments are added "inside" the original tree, thus bisecting original segments, then $\bar{\lambda}$ should actually be reduced.

Character and character state weighting

Considerations of λ also lead to possibilities for character and character-state weighting (Albert et al., 1992, 1993; Albert and Mishler, 1992). If differential λs for different characters (or types of characters) can be discovered a priori, then maximum likelihood-based weights can be specified (e.g., weights taking into account differential probabilities of change at different codon positions in a protein-coding gene). Differential probabilities of transformation that can be specified among states *within* characters can be modeled similarly (e.g., weights taking into account gains vs. losses in restriction site data, or transition/transversion bias in sequence data).

It obviously is difficult to specify expectations for λ before an analysis; currently such approaches can only be attempted for molecular data (one advantage of its relative simplicity), therefore we are far from being able to use this sort of approach for combining molecular and morphological data. Fortunately, one important conclusion of our attempts (Albert et al., 1992, 1993; Albert and Mishler, 1992) at modeling the major known transformational asymmetries is that the differential weights thus produced have little effect on parsimony reconstructions. With data having a reasonable $\bar{\lambda}$ (≤ 0.1, as discussed above), optimal weighted parsimony topologies are usually a subset of the unweighted (or more properly, equally weighted) ones. Thus, paradoxically, our pursuit of well-supported weighting schemes has ended up convincing us of the broad applicability and robustness of equally weighted parsimony.

Choice of characters: independence

Establishing the phylogenetic independence of different characters (assumption 4, above) remains as difficult a problem as specifying expectations for λ for different characters, and for the same reason (ignorance). We know too little about molecular, developmental, and ecological processes that can impart dependence among characters. On the other hand, of course, we need a phylogenetic framework in place before an understanding of evolutionary constraints can be achieved (Hennig's reciprocal illumination in action).

The approach to take in the absence of detailed knowledge of biological causes and interactions behind taxonomic characters is twofold. First, avoid known or suspected cases of character interdependence, such as compensating substitutions in stem regions of ribosomal RNA (due to secondary structure), or introgressive hybridization (a problem with clonally inherited organellar genomes), or adaptive syndromes (e.g,. cave organisms, or hummingbird-pollinated flowers), or developmental correlations (often approachable experimentally through controlled growth studies). Second, hedge your bets by using a number of presumably functionally unrelated character systems. It is much better to sequence portions of several different genes, scattered around the nuclear and organelle genomes, than it is to concentrate on extensive sequencing of a single gene, because of the problem of tight selective constraints on any one highly conserved region. It is imperative to use a spectrum of different morphological characters as well, for the same reason. Any one character system (or maybe all) is influenced by constraints that tend to bias phylogeny reconstruction one way or another, yet a combination of very different character data can allow the "noise" to cancel out, and the historical signal to come through. There is only one known process that can impose a common pattern across all these widely different character systems: phylogeny.

Choice of taxa: compartmentalization

One might hope that extensive nucleotide sequence data from a number of related organisms would somehow obviously reveal the true phylogeny. This has not been borne out in practice in those groups for which considerable sequence data have accumulated—analysis is still required and different analyses often give different results. If one had the complete genome sequence from all species in a given clade, one would simply have a mess of data, at best weakly supporting a phylogeny. Some parts of the genome would be invariant, other parts very "noisy" through multiple mutations. Not only might the true phylogenetic signal be obscured by the noise, but the sheer amount of data would make proper analysis computationally impossible. Fortunately, given the wealth of potential characters made accessible recently through technological advances in ultrastructure, development, and molecular biology (and our improved understanding of the expected properties of reliable characters), it is both feasible and necessary to be highly selective about characters and taxa to be used for any specific phylogenetic problem.

It appears better from a theoretical standpoint, based on the discussions of λ above, to break large surveys down into local analyses (e.g., instead of putting all eukaryotes into one huge matrix, work on relationships within smaller groups accepted as monophyletic based on previous analyses, and later link those groups together), to avoid spurious homoplasy due to incorporation of characters with an inappropriate λ for a given level. This is also clearly necessary from a practical, computational standpoint; the rapidly accumulating data from well-studied groups of organisms have outstripped improvements in software and (especially) hardware, e.g., the number of *rbc*L sequences available for green plants is approaching 1,000.

The theoretical underpinnings of the process of representing diverse yet clearly monophyletic clades in larger-scale cladistic analyses are as yet little explored. The most common approach is to select a few representatives of such a group to be entered into the data matrix (the *exemplar* method). Sometimes care is taken to select representatives that are "basal" clades within the group to be represented; however, this still does not avoid two important problems: i) within-group variation is not being fully represented in the analysis, and ii) an increase in both $\bar{\lambda}$ and asymmetry between λs for different branches is being introduced.

A new approach, called *compartmentalization* (Mishler, 1993; by analogy to a water-tight compartment on a ship—homoplasy is not allowed in or out), that avoids these problems, involves substituting an inferred "archetype" or hypothetical ancestor for a clade accepted as monophyletic a priori into an inclusive analysis. It differs from the exemplar approach in that the representative character states coded for the archetype are based on all the taxa in the compartment (but the archetype is likely to be different from all the real taxa). In brief, the procedure is to: 1) perform global analyses, determine the best supported clades (these are the compartments; more on justification of these below); 2) perform local analyses *within* compartments, often with augmented data sets (since more characters can usually be used within compartments due to the improved homology assessments, as discussed below); 3) return to a global analysis, in one of two ways, either a) with compartments represented by single OTUs (the archetypes), or b) with compartments constrained to the topology found in local analyses (for smaller data sets—this approach is better because it allows the character states of the archetypes to "float" with character optimizations based on the overall tree topology).

Only the most strongly supported clades in the initial global analysis should be used as compartments in the sort of analyses discussed above. It is often necessary to compare clades *within* a cladogram for their relative support for other purposes as well (e.g., when making decisions on classification or when carrying out adaptive or biogeographic studies, only the best-supported nodes in a cladogram should be used). Popular methods for evaluating support for clades include a simple count of the number of characters supporting a node or the use of bootstrap percentages. Both approaches have problems (the former because it is based on a strictly local parsimony measure; the latter because of problems with statistical interpretation); an alternative method shows promise (Bremer, 1988; Donoghue et al., 1992; Graham et al., 1991; Källersjö et al., 1992; Mishler et al., 1991). This method (called "decay analysis," Donoghue et al., 1992; Graham et al., 1991; or "Bremer support," Källersjö et al., 1992) works by obtaining the strict consensus of trees that are one step longer than the most parsimonious tree(s), two steps longer, and so on until all resolution is lost. A "decay index" can be defined as the number of steps maximum parsimony must be relaxed to cause a particular clade to lose its support (Mishler et al., 1991). Based on analyses of real and hypothetical data sets (Mishler et al., in preparation), this index appears to be a sensitive measure of relative support, and is recommended for defining compartments.

The goals of compartmentalization are to cut large data sets down to manageable size (the most obvious effect, but not the most important theoretically), suppress the effect of "spurious" homoplasy, and allow use of more information in analyses. The last is the most subtle point (but probably the most important)—improved homology assessments can be made within compartments. This has been instinctively done by morphologists; when characters are being defined, only the "relevant" organisms (i.e., previously accepted as related) are compared (e.g., leaf-cell size is an important cladistic character within the moss genus *Tortula,* yet obviously this character would have to be eliminated if character state divisions had to be justified across all the mosses together). There are also analogous advantages in molecular data. Alignments can be done more easily, and most accurately, when closely related organisms are compared first (Mindell, 1991). Regions that are too variable to be used globally (and thus must be excluded from a global analysis) can often be aligned and included in a local analysis within a compartment. These goals are self-reinforcing; as better understanding of phylogeny is gained, the support for compartments will be improved, leading in turn to refined understanding of appropriate characters and OTUs.

CONCLUSIONS

It is clear that parsimony works best with "good" data, i.e., with copious, independent, historically informative characters (homologies), evenly distributed across all the branches of the true phylogeny. Indeed, many competing methods tend to converge in their results with such data. It is in more problematic data (e.g., with limited information, a high rate of change, or strong functional constraints) that results of different methods begin to diverge. Data that are marginal or poor will be problematic for any approach, but different approaches account for (or are affected by) "noise" differently. Maximum likelihood approaches (e.g., Felsenstein, 1981; Lake, 1987), including asymmetrical weighting algorithms (e.g., Albert et al., 1992, 1993; Albert and Mishler, 1992) may be able to extend the "window of informativeness" for problematic data, but only if the evolutionary parameters that are biasing rates of change are known. As biases become greater, precise knowledge of them becomes ever more important for avoiding spurious reconstructions. Therefore, given the large number of potential characters made available by modern technology, it is both desirable and possible to be highly selective about the characters that are used to address any particular phylogenetic question, and to avoid the use of data known to be strongly biased.

Furthermore, as discussed above, in cases where evolutionary biases are known or suspected, it is better to use many different character systems (each with their own bias) to filter out noise and discover underlying phylogenetic signal. Specific maximum likelihood models can be developed now for "simple" data types such a DNA sequences or restriction site maps, but not in the foreseeable future for complex data types such as morphology. The constraints operating at epigenetic levels are too indirect, and too taxon specific, for our current biological knowledge. Maximum likelihood approaches are therefore not feasible at present for synthetic "total evidence" analyses, and appear to be unnecessary anyway with "good" data as defined above.

The future of phylogenetic analysis thus appears to be in careful selection of appropriate characters (discrete, heritable, independent, and with a low λ) for use at a carefully defined phylogenetic level. The broadest possible array of different types of characters should be used. Straight, evenly weighted parsimony is to be preferred for such data, not because it is infallible, but because it appears to be a robust method (insensitive to variation over a broad range of possible biasing factors), it is applicable to all types of data ranging from DNA to cell ultrastructure, from anatomy to behavior, and it has an explicit, interpretable, and testable evolutionary basis.

ACKNOWLEDGMENTS

I thank V.A. Albert and M.J. Donoghue for helpful discussion of these issues, and V.L. Roth for comments on the manuscript. Funding was provided in part by NSF Grant BSR-9107484.

LITERATURE CITED

Albert VA, and Mishler BD (1992) On the rationale and utility of weighting nucleotide sequence data. Cladistics *8:*73–83.

Albert VA, Mishler BD, and Chase MW (1992) Character-state weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. In PS Soltis, DE Soltis, and JJ Doyle (eds.): Molecular Systematics of Plants. New York: Chapman & Hall, pp. 369–403.

Albert VA, Chase MW, and Mishler BD (1993) Character-state weighting for cladistic analysis of protein-coding DNA sequences. Ann. Missouri Bot. Gard. *80:* 752–766.

Barrett M, Donoghue MJ, and Sober E (1991) Against consensus. Syst. Zool. *40:*486–493.

Bobrova VK, Troitsky AV, Ponomarev AG, and Antonov AS (1987) Low-molecular-weight rRNAs sequences and plant phylogeny reconstruction: Nucleotide sequences of chloroplast 4.5 S rRNAs from *Acorus calamus* (Araceae) and *Ligularia calthifolia* (Asteraceae). Plant Syst. Evol. *156:*13–27.

Brandon RN (1990) Adaptation and Environment. Princeton, NJ: Princeton University Press.

Bremer K (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution *42:*795–803.

Brooks DR, and McLennan DA (1991) Phylogeny, Ecology, and Behavior. Chicago: University of Chicago Press.

Cavalli-Sforza LL, and Edwards AWF (1967) Phylogenetic analysis: Models and estimation procedures. Evolution *21:*550–570.

Cracraft J (1987) DNA hybridization and avian phylogenetics. Evol. Biol. *21:*47–96.

Cronquist A (1987) A botanical critique of cladism. Bot. Rev. *53:*1–52.

DeBry RW, and Slade NA (1985) Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Zool. *34:*21–34.

Donoghue MJ (1989) Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. Evolution *43:*1137–1156.

Donoghue MJ, and Sanderson MJ (1992) The suitability of molecular and morphological evidence in reconstructing plant phylogeny. In PS Soltis, DE Soltis, and JJ Doyle (eds.): Molecular Systematics of Plants. New York: Chapman & Hall, pp. 340–368.

Donoghue MJ, Doyle JA, Gauthier J, Kluge A, and Rowe T (1989) The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. *20:*431–460.

Donoghue MJ, Olmstead RG, Smith JF, and Palmer JD (1992) Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Missouri Bot. Gard. *79:*333–345.

Doyle JJ (1992) Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. *17:* 144–163.

Farris JS (1981) Distance data in phylogenetic analysis. In VA Funk and DR Brooks (eds.): Advances in Cladistics: Proceedings of the First Meeting of the Willi Hennig Society. Bronx, New York: New York Botanical Garden, pp. 3–23.

Farris JS (1983) The logical basis of phylogenetic analysis. In N Platnick and V Funk (eds.): Advances in Cladistics, Vol. 2. New York: Columbia University Press, pp. 7–36.

Felsenstein J (1978) Cases in which parsimony and compatibility 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. Q. Rev. Biol. *57:*127–141.

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. Annu. Rev. Genet. *22:*521–565.

Funk VA, and Brooks DR (1990) Phylogenetic Systematics as the Basis of Comparative Biology. Washington, D.C.: Smithsonian Institution Press.

Gaffney ES (1979) An introduction to the logic of phylogeny reconstruction. In J Cracraft and N Eldredge (eds.): Phylogenetic Analysis and Paleontology. New York: Columbia University Press, pp. 79–111.

Goloboff P (1991) Homoplasy and the choice among cladograms. Cladistics *7:*215–232.

Gould SJ (1982) Introduction. In T Dobzhansky (ed.): Genetics and the Origin of Species. New York: Columbia University Press, pp. xvii–xli.

Gould SJ (1985) A clock for evolution. Nat. Hist. *94:* 12–25.

Graham LE, Delwiche CF, and Mishler BD (1991) Phylogenetic connections between the "green algae" and the "bryophytes". Adv. Bryol. *4:*213–244.

Harvey PH, and Pagel MD (1991) The Comparative Method in Evolutionary Biology. London: Oxford University Press.

Hennig W (1966) Phylogenetic Systematics. Urbana: University of Illinois Press.

Hillis DM, Bull JJ, White ME, Badgett MR, and Molineux IJ (1992) Experimental phylogenetics: Generation of a known phylogeny. Science *255:*589–592.

Huey RB (1987) Phylogeny, history, and the comparative method. In ME Feder, AF Bennett, WW Burggren, and RB Huey (eds.): New Direction in Ecological Physiology. Cambridge: Cambridge University Press, pp. 76–101.

Hull DL (1980) Individuality and selection. Annu. Rev. Ecol. Syst. *11:*311–332.

Källersjö M, Farris JS, Kluge AG, and Bult C (1992) Skewness and permutation. Cladistics *8:*275–287.

Klassen GJ, Mooi RD, and Locke A (1991) Consistency indices and random data. Syst. Zool. *40:*446–457.

Kluge AJ (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidea, Serpentes). Syst. Zool. *38:*7–25.

Kluge A, and Farris JS (1969) Quantitative phyletics and the evolution of Anurans. Syst. Zool. *18:*1–32.

Lake JA (1987) Determining evolutionary distances from highly diverged nucleic acid sequences: Operator metrics. J. Mol. Evol. *26:*59–73.

Langley CH, and Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. *3:*161–177.

Manhart JR, and Palmer JD (1990) The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. Nature *345:*268–270.

Mindell DP (1991) Aligning DNA sequences: Homology and phylogenetic weighting. In MM Miyamoto and J Cracraft (eds.): Phylogenetic Analysis of DNA Sequences. New York: Oxford University Press, pp. 73–89.

Mindell DP, and Honeycutt RL (1990) Ribosomal RNA in vertebrates: Evolution and phylogenetic applications. Annu. Rev. Ecol. Syst. *21:*541–566.

Mishler BD (1993) Compartmentalization: Local versus global parsimony [abstract]. Hennig XII [Annual meeting of the Willi Hennig Society], Fullerton, California.

Mishler BD, and De Luna E (1991) The use of ontogenetic data in phylogenetic analyses of mosses. Adv. Bryol. *4:*121–167.

Mishler BD, Bremer K, Humphries CJ, and Churchill SP (1988) The use of nucleic acid sequence data in phylogenetic reconstruction. Taxon *37:391–395.*

Mishler BD, Donoghue MJ, and Albert VA (1991) The decay index as a measure of relative robustness within a cladogram [abstract]. Hennig X [Annual meeting of the Willi Hennig Society], Toronto, Ontario.

Miyamoto MM (1985) Consensus cladograms and general classifications. Cladistics *1:*186–189.

Neff NA (1986) A rational basis for a priori character weighting. Syst. Zool. *35:*110–123.

Nei M, Tajima F, and Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J. Mol. Evol. *19:*153–170.

Nelson G, and Platnick N (1981) Systematics and Biogeography. Cladistics and Vicariance. New York: Columbia University Press.

Nijhout HF (1990) Metaphors and the role of genes in development. Bioessays *12:*441–446.

Olmstead RG (1989) Phylogeny, phenotypic evolution, and biogeography of the *Scutellaria angustifolia* complex (Lamiaceae): Inference from morphological and molecular data. Syst. Bot. *14:*320–338.

Patterson C (1982) Morphological characters and homology. In KA Joysey and AE Friday (eds.): Problems of Phylogenetic Reconstruction. London: Academic Press, pp. 21–74.

Patterson C (ed.) (1987) Molecules and Morphology in Evolution: Conflict or Compromise. Cambridge: Cambridge University Press.

Penny D, Hendy MD, and Steel MA (1992) Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. *7:*73–79.

Rieseberg LH, and Soltis DE (1991) Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Plants *5:*65–84.

Sanderson MJ, and Donoghue MJ (1989) Patterns of variation in levels of homoplasy. Evolution *43:*1781–1795.

Sibley CG, and Ahlquist JE (1986) Reconstructing bird phylogeny by comparing DNA's. Sci. Am. *254:*82–92.

Smith RL, and Sytsma KJ (1990) Evolution of *Populus nigra* (sect. *Aigeiros*): Introgressive hybridization and the chloroplast contribution of *Populus alba* (sect. *Populus*). Am. J. Bot. *77:*1176–1187.

Sober E (1988) Reconstructing the Past. Cambridge, MA: MIT Press.

Sogin ML, Elwood HJ, and Gunderson JH (1986) Evolutionary diversity of eukaryotic small-subunit rRNA genes. Proc. Natl. Acad. Sci. U.S.A. *83:*1383–1387.

Swofford DL (1991) When are phylogeny estimates from molecular and morphological data incongruent? In MM Miyamoto and J Cracraft (eds.): Phylogenetic Analysis of DNA Sequences. New York: Oxford University Press, pp. 295–333.

Swofford DL, and Olsen GJ (1990) Phylogeny reconstruction. In DM Hillis nd C Moritz (eds.): Molecular Systematics. Sunderland, MA: Sinauer Associates, pp. 411–501.

Wanntorp H-E, Brooks DR, Nilsson T, Nylin S, Ronquist F, Stearns SC, and Wedell N (1990) Phylogenetic approaches in ecology. Oikos *57:*119–132.

Wiley EO (1981) Phylogenetics: The Theory and Practice of Phylogenetic Systematics. New York: John Wiley and Sons.

Wilson AC, Carlson SS, and White TJ (1977) Biochemical evolution. Annu. Rev. Biochem.*46:*573–639.

Woese CR, and Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. *74:*5088–5090.

Zimmerman W (1967) Methoden der Evolutionswissenschaft [= Phylogenetik]. In G Heberer (ed.): Die Evolution der Organismen, 3. Stuttgart: Aufl. G. Fischer, pp. 61–160.

Zurawski G, and Clegg MT (1987) Evolution of higher-plant chloroplast DNA-encoded genes: Implications for structure-function and phylogenetic studies. Annu. Rev. Plant Physiol. *38:*391–418.