Cladistics (1990)6: 269–275

COMBINATORIAL WEIGHTS IN PHYLOGENETIC ANALYSIS: A STATISTICAL PARSIMONY PROCEDURE

Ward C. Wheeler[1]

[1]*Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, 10024-5192 U.S.A.*

Abstract—A data dependent weighting procedure is developed to allow the comparison of phylogenetic trees based on nucleic acid sequence data. The sampling error of this cladogram "cost" is then examined, permitting statistical evaluation of the cost differential.

Introduction

Molecular sequence data are becoming an increasingly popular tool in systematic biology. They offer an abundant source of new information, rich with possibilities for analysis. Yet although these data can be quite illuminating in the construction of phylogenetic trees, several persistent problems undermine their use. One such problem is the occurrence of multiple, equally-parsimonious solutions to character distribution data. Because nucleic acid sequence data are multistate, the number of plausible solutions can be embarrassingly high.

At present three methods of divining pattern in these multistate characters have been proposed: (1) Fitch or non-additive optimization; (2) Mickevich's Transformation Series Analysis (TSA); and (3) Felsenstein's Maximum Likelihood procedure. Recently, Fitch and Williams (1989) have presented a fourth method to order DNA base transformations.

Fitch optimization (1971) is the most parsimonious, most conservative (even nihilistic) procedure. In this method, all transformations are equally credible. The most parsimonious cladograms have implicit models of character state transformation, but they derive from the tree itself; hence they cannot be used to construct trees *ab initio*. Moreover, all transformations are given equal weight, yet all transformations are clearly not equally probable. This cautious approach limits the resolving power of the data.

By contrast, Mickevich's TSA (1982) is an iterative process that converges on a character cladogram. This method involves the construction of a cladogram, and the determination of its most parsimonious character transformation series. The series is then used to construct yet another cladogram. The entire operation is repeated until a stable set of relationships is reached. A major drawback of this method is the seemingly arbitrary assumption that a character at an interior node will evolve through a singleton state. The construction of the cladograms and the derivation of character transformation hypotheses are inextricably linked, which may bias the result.

The third procedure is somewhat different in outlook. Felsenstein proposes Maximum Likelihood (1978, 1979) to avoid certain difficulties he perceives as compromising the effectiveness of strictly parsimonious approaches. At the crux of Felsenstein's method is the model of evolution employed. Maximum Likelihood requires an *a priori* probabilistic model of evolution. In my view, this assumption has problems all its own, since general models may not apply in specific cases, and specific models may not apply in general.

0748-3007/90/030269 + 7 $03.00/0

An extension of Farris' successive approximations weighting, Williams and Fitch's more recent method (1989) involves an iterative dynamic approach similar to TSA. However, it yields character state transformation weights, in addition to character weights, under a variety of models. Since it is an iterative process, this method also depends on topology to chart character transformations.

Ideally, a method for the construction of character transformations would extract more information from sequence data than Fitch optimization, without depending either on topology or on external hypotheses of character state evolution. Here I will propose such a method.

The Method

Aligned sequences are examined one position at a time. As in any study, these aligned positions constitute the character set of the phylogenetic analysis. If gaps are not included, each position may exhibit one to four nucleotides among the sequences. In other words, a position may be invariant, having the same base, or it may vary, having different combinations of bases.

The essence of this procedure lies in observing how frequently nucleotides co-occur in these variable positions. The greater the frequency that two nucleotides occur in the same position, the more likely they are to have a transformation relationship; that is, a high probability of interchange. An example is found in Fig. 1.

TAXON	POSITION
1	A A A G T A A T A C
2	A C G G T T T G A C
3	A C G G A C C T T G
4	A A G C A G C T T G

Fig. 1. Hypothetical sequence of 10 bases from four taxa.

The first variable position, number two, has two nucleotides, A and C. At a minimum there was one evolutionary transformation between A and C, either from A to C or from C to A. Of course, there may have been more than one transformation, if intermediate states have gone unobserved. I will define an association between two nucleotides as their co-occurrence in an aligned position, as with A and C in position two. Similar situations occur in positions three, four, eight, nine, and ten.

A more complex arrangement is found in the sixth and seventh positions. In position six, all four nucleotides are present, while in seven, three are observed. In these cases, the minimum number of evolutionary events is the number of nucleotides observed minus one, or $(n-1)$, where "n" is the number of nucleotide states in that position. Thus it requires one fewer evolutionary step than nucleotide to span the variation, regardless of the order in which the changes occurred.

Comparing position two to positions six and seven, we notice not only the increase in the number of steps, but also a greater complexity in the paths through which these changes may have taken place. By simple combinatorial logic, n nucleotides produce $\binom{n}{2}$ different pairs. In other words, the $n-1$ changes occur among $\binom{n}{2}$ pairs.

The strength of association between nucleotides i and j in a position k (a_{ijk}), is defined as the minimum number of substitutions required by the nucleotides in that position,

divided by the number of pairs of nucleotides through which these changes may have occurred:

$$a_{ijk} = (n_k - 1)/\binom{n}{2}$$

where

$$a_{ijk} = a_{jik}$$

This number will vary from 1 when there are only two nucleotides, to 2/3 when there are three, to 1/2 when there are four. As with protein sequences, the pattern can be extended to any number of states.

	A	C	G	T
A	—	2.2	1.5	3.2
C	2.2	—	2.5	1.2
G	1.5	2.5	—	1.5
T	3.2	1.2	1.5	—

Fig. 2. The association matrix {A} calculated from the sequences of Fig. 1.

A matrix of associations {A} is then created from the associations at each position (Fig. 2). The elements of {A}, A_{ij}, are calculated by summing the position associations, a_{ijk}, over all k positions:

$$A_{ij} = \sum_k a_{ijk}$$

The matrix is symmetrical, since the pairings make no directional distinction.

The next step is the creation of a transformation matrix {T}. This matrix is constructed by normalizing the columns of {A}. The columns are normalized in order to represent the transformation frequencies from a specific starting point to each of the three alternatives. Consider the case where A is present in 10 positions, C is present in 100, and A and C co-occur in 5. It seems reasonable that transformations from A to C (5/10) should be more frequent than from C to A (5/100). The column normalization reproduces this asymmetry (Fig. 3). Hence the matrix {T} represents the relative transformation frequencies from column to row. These frequencies are the basis for the weights used in phylogenetic analysis.

	A	C	G	T
A	—	0.37	0.27	0.54
C	0.32	—	0.45	0.20
G	0.22	0.43	—	0.26
T	0.46	0.20	0.27	—

Fig. 3. The transformation matrix {T} calculated from the association matrix {A} of Fig. 2.

There are several characteristics that enhance the appeal of a phylogenetic weighting scheme. First of all, less frequent events should be more costly; second, if two events occur with probability p_i and p_j, they should be as costly as a single event with probability $p_i \times p_j$. If we use the absolute value of the result, a logarithmic transformation of {T} conforms to both these criteria. The phylogenetic weight matrix {W} with elements W_{ij} is constructed from {T}, with elements T_{ij}, such that:

$$W_{ij} = |\ln (T_{ij})|$$

where "ln" is the natural logarithm (Fig. 4).

	A	C	G	T
A	—	0.99	1.3	0.62
C	1.1	—	0.80	1.6
G	1.5	0.84	—	1.3
T	0.78	1.6	1.3	—

Fig. 4. The phylogenetic weight matrix {W} calculated from the transformation matrix {T} of Fig. 3.

This new matrix {W} is then used to determine the weighted length or cost of a cladogram, C. The total cost of the cladogram is the product of the number of transformations between nucleotides, m_{ij}, and the cost of that type of transformation, W_{ij}, summed over all types of transformations:

$$C = \sum_i \sum_j m_{ij} W_{ij}$$

Since this cost statistic is based on weights calculated from nucleotide co-occurrences, it is subject to sampling error. The longer the sequence, or indeed, the greater the number of associations, the more precise the estimate of the transformation frequencies. To examine the nature of this sampling error, computer simulations were performed in which the observed associations A_{ij} were the basis for a series of multinomial trials. As an example, if 100 associations were observed in the sequences, the A_{ij} values were calculated and 100 trials were performed in which the probability of observing association *ij* was $A_{ij}/100$. In this way, a new series of association values was determined and used to calculate the matrices {A}, {T}, {W}, and the cladogram cost C. By repeating this procedure, the variance in the estimate of C was observed.

An Example

The entire procedure was performed on a set of insect 18SrDNA sequences (Wheeler, 1989) to determine the phylogenetic arrangement of the Holometabola, Paraneoptera, and Polyneoptera. There were three possibilities (Fig. 5). With all characters unordered (PAUP version 2.4.1; Swofford, 1985), cladogram I, at 132 steps, was the most parsimonious; but cladograms II, at 133, and III, at 137, were not far behind. The corresponding weighted lengths (to three significant figures) were 314, 317, and 327. These values, especially I and II, were very close. Are they "significantly" different in a statistical sense?

The simulation performed on these data was based on the T_{ij} values in Fig. 6, which were drawn from 216 associations. The results are summarized in Table 1. The cost of

Table 1

The cladogram length, weighted length (based on {T} of Fig. 6), and confidence bounds for the sampling error in weighted cladogram length. The upper and lower 2.5% regions were determined by the simulated probability distributions of the sampling process, not through assumptions of normality. In a similar manner, the empirical "p" values were determined through the overlap of these simulated distributions.

Cladogram	Length	Weighted-Length	Lower-2.5%	Upper-2.5%	Empirical "p"
I	132	314	308	320	—
II	133	317	311	323	0.16
III	137	327	321	333	<0·001

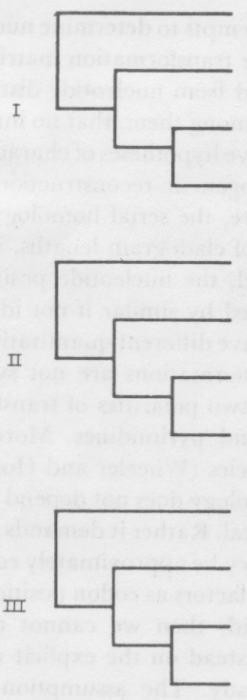

Fig. 5. Three possible relationships among the Paraneoptera, Polyneoptera, and Holometabola with respect to an outgroup.

	A	C	G	T
A	—	0.25	0.44	0.19
C	0.31	—	0.31	0.56
G	0.48	0.27	—	0.25
T	0.21	0.48	0.25	—

Fig. 6. The transformation matrix {T} for insect 18SrDNA sequences.

cladogram II is well within the 95% confidence interval of cladogram I, while the cost of III is not. Hence I is not significantly less costly than II but is significantly less costly than III.

Conclusions

Character data clearly have an order, but how can we tell what it is? Even if we can divine the pattern in a particular case, can we make any generalizations? Are there general laws of character evolution, such as a bias in transition-transversion ratio, or are these "laws" specific to molecules, regions of molecules, or even particular nucleotide positions? These questions are of great importance not only to our understanding of the evolutionary process, but to the methods by which we reconstruct historical relationships. If we knew, for instance, that transitions were always exactly three times as frequent as transversions, it would be foolish not to use this information in phylogenetic reconstruction. Such added knowledge might finally resolve some of the central arguments in molecular systematics. Unfortunately, as yet, we lack this kind of insight.

Here I present a method which attempts to determine nucleotide transformation and use it in a phylogenetic context. The transformation matrix {T} is a statement about character evolution. It is constructed from nucleotide distributions among taxa, and relies on several assumptions, chief among them, that no intermediate states have gone unobserved. The resulting quantitative hypotheses of character evolution can be used to assess the merits of alternative phylogenetic reconstructions.

One of the assumptions made here, the serial homology of nucleotide positions, is crucial in the statistical evaluation of cladogram lengths. In order for the length and variance estimates to be meaningful, the nucleotide positions used to construct the association matrix must be controlled by similar if not identical evolutionary forces. However, different molecules may have different quantitative character transformation hypotheses. Furthermore, the transformations are not symmetrical. There may be different levels of bias between the two polarities of transformation, or there may be different biases between purines and pyrimidines. Moreover, even regions within molecules can show these discrepancies (Wheeler and Honeycutt, 1988).

Yet the requirement of serial homology does not depend on different nucleotide base positions being evolutionarily identical. Rather it demands that they be robust to minor violations of this assumption, that they be approximately equal. If the evolution of each position is truly unique, due to such factors as codon position, or the hydrophobicity or hydrophilicity of coded amino acid, then we cannot create statistical models in systematics at all, and must rely instead on the explicit minimization of parsimony, where all changes are equally costly. The assumption of serial homology seems reasonable for certain areas within the molecule, especially for codon positions. These are the circumstances in which the method functions best.

The matrices and cost functions described here are specific to the taxa and sequences under study. In fact, a unique model is calculated in each case, making it unnecessary to assume a general model of evolution for all molecules and taxa. Such specificity is one of the strengths of this procedure. Not only can we use it to decide between equally or near-equally parsimonious cladograms. By gathering transformation matrices from different molecules, we may even be able to make general statements about character evolution, and its importance in phylogenetic reconstruction.

Acknowledgments

I would like to acknowledge the help of Elise Broach and Paul Vrana in the preparation of this manuscript. This research was funded by NSF grant BSR-870096.

REFERENCES

Felsenstein, J. 1978. Cases under which parsimony or compatibility will be positively misleading. Syst. Zool. 27: 401–410.

Felsenstein, J. 1979. Alternate methods of phylogenetic inference and their interrelationship. Syst. Zool. 28: 49–62.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20: 406–416.

Mickevich, M. F. 1982. Transformation series analysis. Syst. Zool. 31: 461–478.

Swofford, D. L. 1985. PAUP. Computer program and manual. Illinois Natural History Survey, Urbana.

WHEELER, W. C. 1989. Evolution and systematics of insect rDNA. In B. Fernholm et al., eds, *The Hierarchy of Life*. Elsevier Press, Amsterdam, pp. 307–321.

WHEELER, W. C. AND R. L. HONEYCUTT. 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. Mol. Biol. and Evol. 5: 90–96.

WILLIAMS, P. L. AND W. M. FITCH. 1989. Finding the minimal change in a given tree. In B. Fernholm et al., eds, *The Hierarchy of Life*. Elsevier Press, Amsterdam, pp. 453–470.

(Received for publication 18 November 1988; accepted 19 February 1990)