Blackwell
Publishing

Cladistics 22 (2006) 157–170

Cladistics

www.blackwell-synergy.com

Dynamic homology and the likelihood criterion

Ward C. Wheeler

Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024-5192, USA

Accepted 6 December 2005

Abstract

The use of likelihood as an optimality criterion is explored in the context of dynamic homology. Simple models and procedures are described to allow the analysis of large variable length sequence data sets, alone and in combination with qualitative information (such as morphology). Several approaches are discussed that have different likelihood interpretations in terms of maximum parsimony likelihood and maximum average likelihood. Implementation is discussed and an example in arthropod systematics presented. Topological congruence comparisons with parsimony are made.

Dynamic homology

Nucleic acid sequence data do not present themselves in neat packages. Nucleotide homologies are topology specific and identified by optimization processes that determine nucleotide correspondence and transformation via a specific criterion. This is the concept of dynamic homology (Wheeler, 2001; Wheeler et al., 2005b). Methods have been proposed to apply this manner of thinking using parsimony as an optimality criterion (Wheeler, 1996, 1999a, 2003a,b) which have their roots with Sankoff (1975), but less work has been done to explore the use of likelihood as an alternate interpretive framework.

Several methods for analyzing unaligned sequence data using likelihood as an optimality criterion in a dynamic homology framework are proposed here. These techniques are optimization methods in the sense of Wheeler (2005) in that they do not rely on *a priori* multiple alignments (although they may generate them *a posteriori*), but analyze variation directly, yielding optimal—in this case *likely*—cladograms. To accomplish this, models that can accommodate insertion-deletion information are required.

Likelihood alignment models

Although there exist a large diversity of nucleotide substitution models, there are far fewer which model the insertion-deletion process. The most prominent of these models is that of Thorne et al. (1991) (TKF) for the statistical alignment of two sequences. TKF allows for transitions among nucleotides, as well as their insertion and deletion through a birth/death process of gaps. The model yields the probability of one sequence evolving into another over a given time interval, and was expanded in Thorne et al. (1992) to include affine gaps and rate heterogeneity. Fleissner et al. (2005) used the latter in their multiple-alignment approach, and Redelings and Suchard (2005) added additional generalizations for their Bayesian approach.

An important advance of the TKF model over previous attempts (e.g., Bishop and Thompson, 1986) is that the total probability of transformation between sequences is the sum of all possible alignments between the sequences (and there may be many of them; Slowinski, 1998). The unique alignment that is usually

E-mail address: wheeler@amnh.org

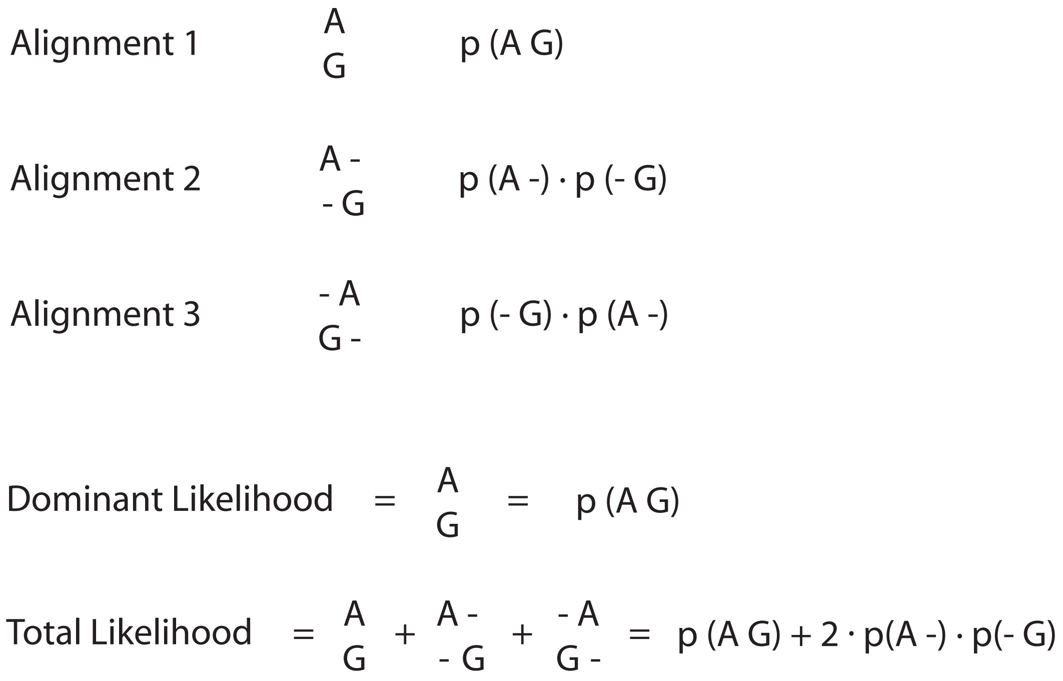

Alignment 1
A
G
p (A G)
Alignment 2
A -
- G
p (A -) · p (- G)
Alignment 3
- A
G -
p (- G) · p (A -)
Dominant Likelihood = A G = p (A G)
Total Likelihood = A G + A - - G + - A G - = p (A G) + 2 · p(A -) · p(- G)

Fig. 1. A simple alignment of two nucleotides demonstrates the difference between a dominant likelihood alignment and the *total* likelihood. Assuming probability of transformation between A and G [*p*(AG)] to have a higher probability than indels [*p*(–G); *p*(A–)], alignment 1 is the dominant (= highest likelihood) alignment. The total likelihood alignment would include contributions from all possible alignments, however, marginal. In this case, the total would include contributions from alignments 1, 2 and 3.

presented and employed is the so-called *dominant* alignment of the two sequences, that is, the one with highest individual probability, although it is most likely that this alignment may contain a very small fraction of the total transformation probability (Hein et al., 2000; Fig. 1). This behavior will reappear throughout the likelihood models and cladogram-based optimization implementations presented here.

The computational costs of the original TKF recursive method were vastly improved by Hein et al. (2000). Steel and Hein (2001) generalized this model to more than two sequences including a polynomial time approximation for three sequences. This is significant in that it allows for the optimization of a complete cladogram via methods akin to those of Sankoff (1975). The thread was completed by Hein et al. (2003) who presented a recursive method for optimizing complete (if small) binary trees. Hein et al. stated that their approach could be used for up to six or seven sequences, but that larger data sets would require a radically different method.

The limitations on the tractability and utility of the Hein et al. (2003) approach and the even more time-efficient Bayesian methods (Redelings and Suchard, 2005) required 1–8 days to analyze 12 taxa on a 1.7 GHz Athlon PC. Clearly, another approach is required to analyze the data sets of hundreds to thousands of sequences commonly available today. A simpler, less complex model that is able to accommodate such large data sets can have considerable utility.

Forms of likelihood

Likelihood, at least in a phylogenetic context, is not a monolithic entity but presents itself in a variety of flavors. Before delving further into specifics, it is necessary to be clear as to the type of likelihood pursued here, and its relationship to other forms (Fig. 2). As pointed out by Steel and Penny (2000), Steel (2002) and Goloboff (2003), the likelihood methods used today are based on maximum relative likelihood (MRL), as opposed to maximum integrated likelihood (MIL). The distinction between these approaches is that for the MRL analysis, those branch lengths which maximize overall likelihood are used to establish cladogram likelihood and evaluate competing topologies. In contrast, MIL integrates the total value over all possible branch lengths for each cladogram. These two forms of likelihood may conflict as when an unlikely set of branch lengths (or other parameters) yields a high overall score. The methods proposed here are based on MRL, since a single branch length is used to determine the likelihood of a given hypothetical taxonomic unit (HTU) state. Within MRL, there are further subtypes. Barry and Hartigan (1987) distinguished between

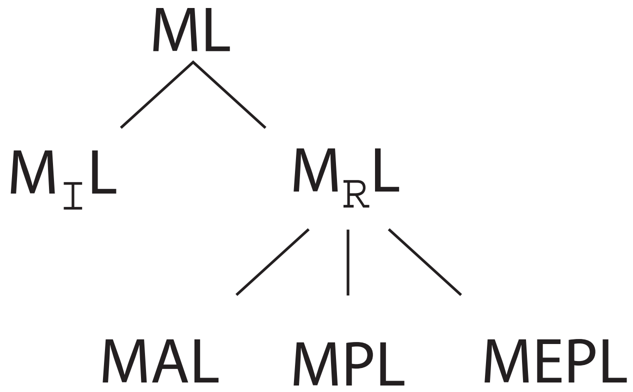


Fig. 2. Forms of likelihood after Barry and Hartigan (1987), Steel (2002) and Goloboff (2003). ML = maximum likelihood in the most general sense, MIL = maximum integrated likelihood, MRL = maximum relative likelihood, MAL = maximum average likelihood, MPL = maximum parsimony likelihood, and MEPL = maximum evolutionary path likelihood.

maximum average likelihood (MAL) and maximum parsimony likelihood (MPL). The MAL approach in essence averages over all possible internal nodal states (HTU assignments), whereas MPL assigns a single state or sequence that maximizes the likelihood. MPL methods have the benefit of being much simpler to compute, since only a single (but hopefully highly likely) scenario is required, as opposed to the potentially large universe of solutions. Both the MAL and MPL methods are presented here. A third approach is the evolutionary path method of Farris (1973). In this form of MRL, the actual sequence of intermediate states is chosen to maximize likelihood. This yields precisely the same result as parsimony, even under variations in mutation rates and branch lengths (Farris, 1973). The evolutionary path likelihood is not implemented by any of the methods proposed here.

A simple likelihood model

A simple five-state model of sequence change can be used (Durbin et al., 1998; McGuire et al., 2001) at the very least to approximate the results of more complex formalizations. In such a model, indels are accounted for by transformations between the state of the gap and nucleotides (Fig. 3). Such a model, coupled with a method for determining likelihood edits between sequence pairs, will yield a likelihood basis for the optimization of unequal length sequences on a cladogram. This type of model may be rough, but it does allow the calculation of an optimality value. This optimality value can then be used to choose hypothetical ancestral sequences and to assess the relative merits of phylogenetic topologies.

The fundamental model used in all the methods presented here is a 10-parameter symmetrical transition matrix Q, coupled with a five-parameter state frequency vector ($S_{10}F_5$) (Fig. 4). These 15 parameters (14 independent since the π-values must sum to 1) determine the overall probabilities of a transformation among each pair of nucleotides and gap (signifying an indel) for a given time t via elementary linear algebra (equation 1):

$$P(t) = e^{Rt}, \qquad (1)$$

where R (as in Tavaré, 1986) is derived from the symmetrical transition matrix Q and the state vector Π (Yang, 1994; equation 2).

$$\begin{array}{cccccc}
 & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} & - \\
\mathrm{A} & -(\pi_C\alpha+\pi_G\beta+\pi_T\gamma+\pi_-\delta) & \pi_C\alpha & \pi_G\beta & \pi_T\gamma & \pi_-\delta \\
\mathrm{C} & \pi_A\alpha & -(\pi_A\alpha+\pi_G\varepsilon+\pi_T\zeta+\pi_-\eta) & \pi_G\varepsilon & \pi_T\zeta & \pi_-\eta \\
\mathrm{G} & \pi_A\beta & \pi_C\varepsilon & -(\pi_A\beta+\pi_C\varepsilon+\pi_T\theta+\pi_-\kappa) & \pi_T\theta & \pi_-\kappa \\
\mathrm{T} & \pi_A\gamma & \pi_C\zeta & \pi_G\theta & -(\pi_A\gamma+\pi_C\zeta+\pi_G\theta+\pi_-\nu) & \pi_-\nu \\
- & \pi_A\delta & \pi_C\eta & \pi_G\kappa & \pi_T\nu & -(\pi_A\delta+\pi_C\eta+\pi_G\kappa+\pi_T\nu)
\end{array}$$

Fig. 3. General, though symmetrical, five state nucleotide and gap (‘–’) stationary Markov model. The π components refer to nucleotide and gap frequencies; α, β, γ, δ, ε, ζ, η, θ, κ and ν refer to the nucleotide and indel transition probabilities.

Frequency Model

F_1 $\pi_A = \pi_C = \pi_G = \pi_T = \pi_-$

F_2 $\pi_A = \pi_C = \pi_G = \pi_T \;;\; \pi_-$

F_5 $\pi_A \;;\; \pi_C \;;\; \pi_G \;;\; \pi_T \;;\; \pi_-$

Substitution Models

S_1 $\alpha = \beta = \gamma = \delta = \varepsilon = \zeta = \eta = \theta = \kappa = \nu$

S_{1G} $\delta = \eta = \kappa = \nu \;;\; \alpha = \beta = \gamma = \zeta = \varepsilon = \theta$

S_{2G} $\delta = \eta = \kappa = \nu \;;\; \beta = \zeta \;;\; \alpha = \gamma = \varepsilon = \theta$

S_{3G} $\delta = \eta = \kappa = \nu \;;\; \beta \;;\; \zeta \;;\; \alpha = \gamma = \varepsilon = \theta$

S_{6G} $\delta = \eta = \kappa = \nu \;;\; \alpha \;;\; \beta \;;\; \gamma \;;\; \zeta \;;\; \varepsilon \;;\; \nu$

S_{10} $\alpha \;;\; \beta \;;\; \gamma \;;\; \delta \;;\; \varepsilon \;;\; \zeta \;;\; \eta \;;\; \theta \;;\kappa \;;\; \nu$

Fig. 4. Alternate models of both state frequencies and transition class probabilities from the most complex ($S_{10}F_5$) to the most simple (S_1F_1). Abbreviations as in Fig. 3.

$$R_{i,j} = Q_{i,j} \cdot \pi_j,$$
$$\text{if } i \neq j;\ \text{else } R_{i,j} = -\sum Q_{i,s} \text{ for } s \in \{A, C, G, T, -\}. \tag{2}$$

The probability of change (P) between states i and j in time t is then:

$$P_{i,j}(t) = \sum e^{\lambda_s t} \cdot U_{s,i} \cdot U^{-1}_{j,k}, \tag{3}$$

with s as in eqn 1; λ_s are the eigenvalues of R and U the associated matrix of eigenvectors and U^{-1} its inverse.

Many special cases can be generated from this general expression (10 substitution parameters and five state frequencies—$S_{10}F_5$ in POY Wheeler et al. (1996–2003) see below) by reducing degrees of freedom and pooling various events into classes. There are six general substitution models and three state frequency scenarios

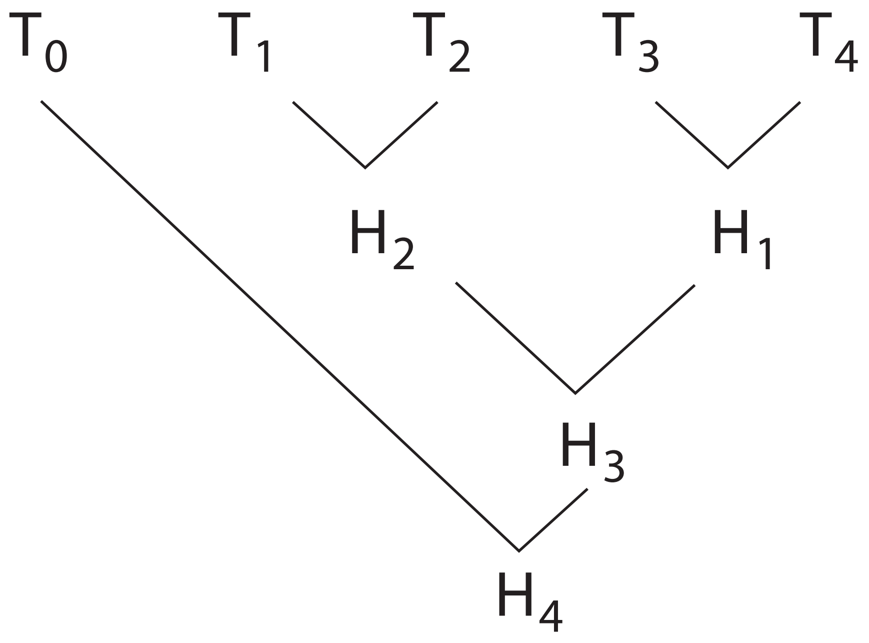


Fig. 5. Down-pass used to calculate likelihood values under Direct Optimization (DO; Wheeler 1996). The immediate ancestors of terminal nodes (T_i) are optimized first then proceeding down the cladogram to the root (H_4) until all internal nodes (H_i)are optimized.

for 18 model combinations from a *InDel-Jukes-Cantor* (Jukes-Cantor + Gaps; S_1F_1) where all transitions and all state frequencies are equal, to the 14 parameter *InDel-GTR* (the 11 parameter would also be a GTR model) (GTR + Gaps; Fig. 5). Furthermore, these model parameters can be determined individually for different fragments of DNA and even for the individual branches of the topologies examined during a search. In general, investigators seem to prefer homogeneous indel probabilities (or costs) such that indels of A, C, G or T do not differ. For this reason, the S_{1G}, S_{2G}, S_{3G} and S_{6G} models restrict indels to a single probability. The S_{10} model removes this restriction. Comparison of the behavior of these models would allow a test of the assumption of indel homogeneity.

Following parsimony-based techniques, there are two classes of methods presented here to employ likelihood as a criterion for dynamic homology-based analysis: estimation and search methods (*sensu* Wheeler, 2005). The estimation methods are heuristics that seek to create likely HTU sequences such that the overall cladogram likelihood is maximized. Likelihood versions of Direct Optimization (Wheeler, 1996) and Iterative-Pass optimization (Wheeler, 2003a) fall into this camp. In their focus on likely HTUs, these methods are MPL methods in the sense of Barry and Hartigan (1987), though there are elements that can be described as MAL. The search-based methods, Fixed-State (Wheeler, 1999a) and Search-Based Optimization (Wheeler, 2003b), can be implemented as either MPL or MAL methods and both are discussed here. All of these methods are heuristic cladogram optimization procedures of the same basic model. The problems of cladogram search and tree-alignment (Wang and Jiang, 1994) are known to be NP-complete and none of these methods (other than exhaustive Search-Based Optimization) guarantees an exact solution for more than two sequences.

Direct optimization (DO)

As with parsimony, Direct Optimization using likelihood (DO-lik) begins by determining the HTU sequence of a node with two terminals (OTUs) as descendants (Two Sequences below). This is repeated for each node in a down-pass until each HTU has been determined from its descendants (Fig. 5). The likelihood of the cladogram is the product of the likelihoods of the HTUs. The differences from DO-parsimony come in the determination of the optimization cell costs (sum versus minimum value), the incorporation of branch-length information, and the creation of the HTU sequence. All these events occur within the modified Needleman and Wunsch (1970) algorithm used to determine the HTU.

Two sequences

The core of DO-lik is the determination of an HTU sequence from two descendant sequences. In order to begin, three sorts of values are required. Given the most general sequence change model ($S_{10}F_5$), 10 transition probabilities (Q), 5 state probabilities (Π), and a time (t) or expected branch length factor are needed (as well as any others such as the gamma α

and invariant sites θ). *Q* and Π can be asserted or estimated by various means (see Implementation section below), but *t* must be determined for the HTU problem at hand. One could simply begin at an arbitrarily small *t* and determine HTU likelihood for ever-larger values keeping the maximum. This approach could be improved by using an initial estimate based on the parsimony-based DO. This step would provide an estimate of the number of changes on the branch (and potentially *Q* and Π as well) which could then be refined through iteration.

With the necessary parameters in place, dynamic programming can be used to determine the HTU likelihood and sequence. The procedure would follow the parsimony DO procedure (Wheeler, 1996) with two modifications. First, the transformation costs between nucleotides and indels would follow the usual absolute value of the logarithm of the transition probabilities determined from the likelihood model and *t* (eqn 3). This makes the costs additive and the problem one of minimization; hence we can follow the modified Needleman–Wunsch algorithm. Second, the costs ($c_{i,j}$) of individual cells in the Needleman–Wunsch matrix would be calculated from the sum of the three paths to that cell as opposed to the minimum value (eqn 4).

$$c_{i,j} = (P_{i,j} \cdot c_{i-1,j-1}) + (P_{i,gap} \cdot c_{i-1,j}) + (P_{gap,j} \cdot c_{i,j-1}). \qquad (4)$$

Following Thorne et al. (1991), this yields the total likelihood value of the transformation between the two input sequences. If the minimum value alone were used, the dominant likelihood value would be produced (this may yield different homologies). This is a noteworthy distinction, since a unique HTU (although it might have ambiguities) is produced. Operationally, this HTU is created during the traceback step of the procedure based on the minimum cost/maximum likelihood match/indel patterns in the Needleman–Wunsch matrix. Since the method produces a single HTU, this is a MPL procedure and the use of dominant likelihood values most consistent. The total likelihood cost would be a MAL method for two sequences.

After the HTU likelihood (total or dominant) is determined for a given *t*, transition probabilities would be recalculated for a new *t*, and the process repeated iteratively until a stable solution was found.

Iterative-Pass (IP)

The likelihood version of Iterative Pass (IP-lik) follows the steps of its parsimony incarnation (Wheeler, 2003a) with initialization, three-way HTU determination, and cladogram cost calculation (Fig. 6). The modifications are quite straightforward. The HTUs are initialized with DO-lik and then updated using a three sequence version of the DO-lik described above. In this case, all three branches leading to the HTU are iterated and the optimization itself is more time consuming. The HTUs are revisited in turn (and minimally) until they are stable in sequence. The cladogram cost determination uses the two-sequence DO-lik above, to determine the branch-likelihood for each ancestor–descendant pair on the cladogram and this sum is the cladogram likelihood (Fig. 7).

As with DO-lik, the method has total and dominant likelihood calculations based on summing (over seven paths in this 3-D case) or taking the maximum likelihood path for HTU determination. A single HTU is produced either way however, hence the method is MPL.

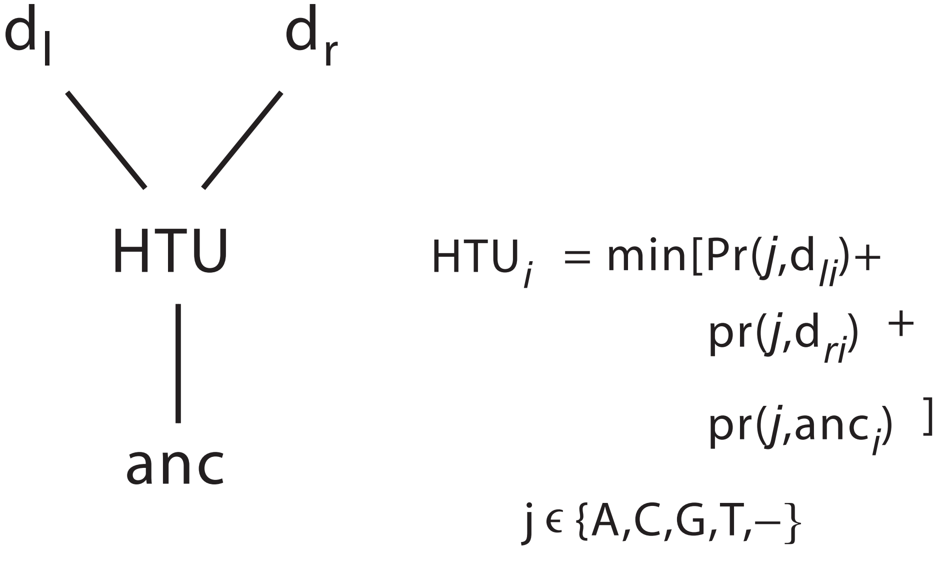
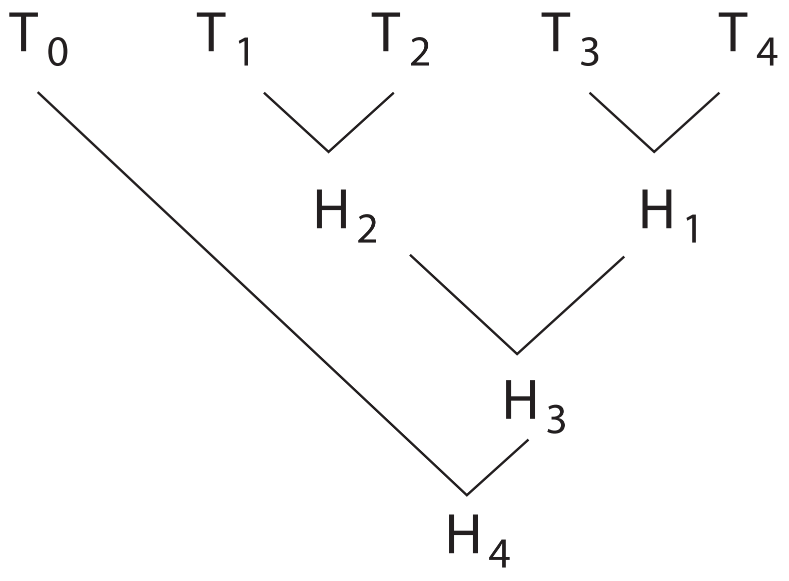


Fig. 6. Calculation of HTU sequence under likelihood using Iterative-Pass optimization (IP; Wheeler 2003a). HTU$_i$ is chosen such that it minimizes the sum of the likelihoods of transformation to each of the three adjacent nodes (one ancestor and two descendants).



$$\text{log likelihood}_{\text{cladogram}} = \text{logPr}_{(H_4)} + \text{loglik}_{(H_4, T_0)} + \text{loglik}_{(H_4, H_3)} + \text{loglik}_{(H_3, H_2)} + \text{loglik}_{(H_3, H_1)} + \text{loglik}_{(H_2, T_1)} + \text{loglik}_{(H_2, T_2)} + \text{loglik}_{(H_1, T_3)} + \text{loglik}_{(H_1, T_4)}$$

Fig. 7. Determination of cladogram likelihood in IP-lik. The total log-likelihood is the sum of the log-likelihood values of each branch.

Fixed-State/Search-Based Optimization

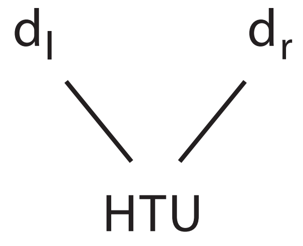
These methods are referred to as search methods as opposed to estimation methods (Wheeler, 2005) due to their reliance on the examination of a heuristic subset of the known range of possible sequences. Sequences, like cladograms are (in principle) enumerable before any analysis. Fixed-States (FS; Wheeler, 1999a) and Search-Based Optimization (SBO; Wheeler, 2003b) examine very small sets of possible sequences as HTU states. They do not attempt to construct HTU sequences (as do the estimation methods above) in creating optimal cladograms, but choose among pre-specified sequences. For FS, the search set is defined by the observed sequences, yielding a minimum-size search space. SBO expands the sequence state set, hence producing better, i.e. more optimal, cladograms at a cost of increased execution time (scaling quadratically with the number of states). In principle, SBO could examine all possible sequences yielding an exact solution for that cladogram. This is unlikely to be tractable for anything but the shortest sequences.

As with the parsimony version of these methods, dynamic programming (Sankoff and Rousseau, 1975) is used to determine both the optimal cladogram cost and the HTU state set. The fundamental requirement for this is a sequence edit cost matrix. As the parsimony cost matrix is determined by the pair-wise parsimony DO, the likelihood edit costs are determined by the DO-lik described above.

Once more, this raises the issue of total/dominant likelihood. When intermediate costs/likelihoods are calculated, either the minimum cost/maximum likelihood can be assigned to HTU sequence states, or the sum of all paths to that state (Fig. 8).

This corresponds to the dynamic programming operations for DO-lik above. If the total likelihood cost is calculated for the edit cost matrix, and HTU state likelihoods are summed as well, FS/SBO will be a MAL method, if dominant (minimum cost/maximum likelihood) are used on both, a MPL version will result.

Another point worth noting concerns the interpretation of branch-lengths in the FS/SBO context. Whether under MAL or MPL, the likelihood sequence edit cost



$$\text{Total Likelihood HTU}_i = \sum_j \text{lik } d_{lj} \cdot \Pr(d_{lj}, i) + \sum_j \text{lik } d_{rj} \cdot \Pr(d_{rj}, i)$$

$$\text{Dominant Likelihood HTU}_i = \max;\{\text{lik } d_{lj} \cdot \Pr(d_{lj}, i)\} + \max;\{\text{lik } d_{rj} \cdot \Pr(d_{rj}, i)\}$$

Fig. 8. Calculation of total and dominant likelihood values under Search-Based optimization procedures. The total likelihood sums the likelihood contributions of all possible sequence state assignments at each HTU, whereas the dominant likelihood is found by determining the highest likelihood sequence state for each HTU.

matrix is the likelihood of transforming one sequence into another. This requires an estimate of *t*, the branch length. The likelihood values for all the sequence pairs are unlikely to be the same. Hence, each sequence state change implies a different branch length and this is embedded in the sequence edit cost. These differing branch lengths imply heterogeneity in models of change among the sequences and a mixed model among sequence states and fragment characters. The lack of a necessary homogeneous model may alleviate some of the problems in likelihood analyses of real data, discussed theoretically by Chang (1996) and through simulation by Kolaczkowski and Thornton (2004). These results show that likelihood methods will be inconsistent (Chang, 1996) and under-perform parsimony (Kolaczkowski and Thornton, 2004) in an environment of heterogeneous evolutionary processes (a point also made by Steel and Penny, 2000). The FS-SBO procedures do not truly create a mixed model among nucleotides, but do among sequence fragments. The dynamic homology framework cannot accommodate (at least at present) such a mixing of models. This is due to the necessary interrelationship among the nucleotides in homology determination. A truly heterogeneous model approach would have to assign parameters to each nucleotide in some manner akin to that of Tuffley and Steel (1998).

Simultaneous analysis

One of the great advances of recent systematics is the integration of disparate sources and forms of data to simultaneously test cladistic hypotheses. There is no reason why likelihood methods must be left out of the total evidence (Kluge, 1989) approach (the Bayesian approach implemented in MrBayes (Huelsenbeck and Ronquist, 2003) allows this). The missing component has been likelihood models for morphological features. The most general of these is that of Tuffley and Steel (1998). This followed that of Goldman (1990) and preceded others (e.g., Lewis, 2001). As the most general model, Tuffley and Steel (NCM; no common mechanism) likelihood values can be used in combination with those determined from molecular data. The *r*-states Jukes and Cantor (1969); Neyman (1971) model used, conforms to a nonadditive, or unordered character interpretation. Furthermore, since additive characters can be transformed into a series of binary nonadditive characters, the vast majority of qualitative morphological data can be analyzed this way. An additional attractiveness of the NCM model is its identity with parsimony (the same can be said for Goldman, 1990).

In combination with the methods proposed here, a likelihood-based total evidence analysis of morphological and unaligned molecular sequence data can be performed.

Support

It is worthy of mention that given a common interpretive framework for analysis, several support measures have convergent meaning. Such measures as Bremer support (Bremer, 1994) and Jackknife values (Farris et al., 1996) can be calculated under likelihood. Quite clearly, the log-likelihood ratio of cladograms with and without a clade is identical to the Bremer support, which is based on log-likelihood cladogram costs. Jackknife values may prove to be especially interesting if the summary cladogram is determined from weighted (likelihood weighted) cladogram costs. The clade support values can be quite similar to those proposed as posterior probabilities by Bayesian phylogenetic software (Huelsenbeck and Ronquist, 2003). This is an area for future investigation.

Implementation

An implementation was created in the computer program POY (Wheeler et al., 1996–2003, 2005a) to explore these models and procedures. Several general topics merit discussion here, and more option specific information can be found in the POY documentation (ftp.amnh.org/pub/molecular/poy).

Estimation of model parameters

Estimates of model parameters are required for HTU likelihood calculations. In the most complex of cases, the 10 values of the R matrix, five of Π, as well as the gamma shape parameter α and invariant sites proportion θ will need to be determined (and the number of discrete classes for the gamma shape distribution specified). In general, these parameters can be calculated by two distinct methods: (1) before any search begins via pair-wise comparisons, or (2) during the search on a branch-by-branch basis.

The most simple and straightforward manner of estimating R and Π parameters is to perform a series of pair-wise parsimony alignments, enumerate the number of transitions of each type (including indels), and count the relative numbers of each nucleotide state and gap. Such an empirical process would not be an explicit attempt at maximizing any likelihood, but provides a useful estimate. This could be refined by iterating the values of the parameters until a summed likelihood value was optimized, but this would be extremely time consuming (PAUP (Swofford, 2003) can use a cladogram to estimate many parameters). Such a refinement for α and θ would be quite reasonable and is the default for POY.

A second, more specific, estimation occurs during the determination of each HTU. A preliminary parsimony alignment of the two descendant sequences (in the case of DO-lik) or three adjacent sequences (in the case of IP-lik) could be used to estimate transition and state frequency parameters specific to a branch of a cladogram. The strength of such an estimate is that it would be tailored to that area of a cladogram. A drawback would be the large multiplication of effective parameters.

Estimation of branch lengths

An initial t (expected number of changes) is calculated from parsimony DO and iteration proceeds from there. The branch length is incremented and decremented by a factor (step interval) and likelihoods recalculated. For Iterative Pass Optimizations, each of the three branches connected to the HTU are iterated in turn and repeatedly until all are stable.

Dominant and total likelihood

As mentioned above, dominant and total likelihood costs can be calculated for DO-lik and IP-lik as well as the FS/SBO procedures. For reasons of efficiency, dominant likelihoods are usually calculated; hence, the methods are MPL by default. The only real difference in the calculation for DO/IP is found in Needleman and Wunsch's (1970) cell costs; summed for total likelihood and maximum value for dominant.

For FS-lik/SBO-lik, the total likelihood option is exact (given the state set) summing the likelihoods of all the paths to a given state (i.e., from all others).

Combination of data

The combination of data for total evidence analysis proceeds by adding the log likelihoods of the component characters. Qualitative morphological likelihoods are determined under No Common Mechanism (Tuffley and Steel, 1998). Molecular likelihood values are determined by the analytical procedure (DO-lik, IP-lik, FS-lik or SBO-lik).

Cladogram likelihood

Cladogram log-likelihood is determined by summing the log-likelihoods of the HTUs for DO-lik, the branch log-likelihoods for IP, and the final root-state log-likelihoods for FS/SBO with the log probability of the root node.

An example

To illustrate the behavior of these methods, the arthropod data set of Giribet et al. (2001) was used. This data set contains 54 taxa, 303 morphological characters and eight molecular loci (16S mt rRNA, 18S rRNA, 28S rRNA, mt cytochrome *c* oxidase subunit I, elongation factor-1α, RNA Polymerase II, histone H3 and U2 snRNA). Not all taxa were complete for all loci. Other examples can be found in Okusu et al., (2003) and Edgecombe and Giribet (2004).

All searches were performed using POY ver. 3.0.11 (Wheeler et al., 1996–2003) on the AMNH DEMETER

Table 1
Optimality values for partition analyses under parsimony and likelihood (-log). Data of Giribet et al. (2001)

		Cladogram costs									
Optimality	Method	Total	Morph	16S	18S	28S	COI	EF1α	H3	POL1	U2
Parsimony	DO	35262	596	2865	13176	552	3311	5679	1189	4746	304
	IP	35006	596	2832	13027	546	3307	5689	1188	4751	305
	FS	42412	596	3477	14528	630	4541	7976	1507	6332	385
	SBO	37026	596	3204	14116	569	3383	5829	1209	4915	312
Likelihood	DO	113090.23	766.90	6942.82	33611.75	2125.98	13033.37	23795.95	5427.81	17727.47	1477.40
	IP	113459.95	766.90	6892.47	33621.80	2147.69	13147.63	23619.13	5415.99	17835.97	1485.03
	FS	116570.82	766.90	7364.70	32242.45	1907.76	15789.13	28795.68	6037.71	21311.82	1497.24
	SBO	109441.22	766.90	6998.29	30431.90	1762.34	14456.72	25995.42	5730.46	20227.55	1358.63

cluster computer (2.8 GHz PIV Xeon CPU Linux) using 50 processors. Searches were based on a single addition sequence, followed by TBR branch swapping and tree fusing (Goloboff, 1999) if there were multiple cladograms. Up to 25 equally costly cladograms were stored, randomly adding new cladograms if the buffer were full (option `fitchtrees`). A final round of TBR swapping was performed examining cladograms within 1% of the minimum cost to control errors in the cladogram length heuristic calculations. The following command line options were used for all runs: `-norandomizeoutgroup -treefuse -checkslop 10 -maxtrees 25 -fitchtrees`. Additionally, indel cost were set to two and nucleotide substitutions to one. These would be constant for parsimony, but revised under likelihood as model parameters (including indels) were estimated. Morphological changes were accorded a weight of two to equal the indel cost, and this weighting was carried through the likelihood analysis. Likelihood runs added the options `-likelihood -totallikelihood -gammaclasses 2 -invariantsitesadjust` to use the sum of dynamic homology alternatives (as opposed to the dominant likelihood), Γ rate distribution with two classes (default), and adjustment for invariant sites. Likelihood calculations were based on the $S_{6G}F_5$ model (default) where nucleotide and indel frequencies were initially estimated via pair-wise comparison and held constant over the search. The nucleotide–indel substitution model ($S_{6G}F_5$) is equivalent to GTR + GAPS where there is a single indel transition probability for all nucleotides. This matrix was estimated, eigenvalues calculated, and transition probabilities derived uniquely for each branch of each cladogram examined. Morphological likelihood values were calculated using the methods of Tuffley and Steel (1998).

These are not the analytical parameters used by Giribet et al. (2001) in their analysis. The searches here were much less exhaustive and only a single (and different) parameter set used. This was largely due to the time requirements of the likelihood analyses.

Analyses were performed on each of the nine data partitions (morphology and eight molecular loci) separately and in combination. Each data set was analyzed under likelihood and parsimony optimality criteria using the four heuristic methods (except for the morphological data) described above, resulting in a total of 78 analyses.

Results

The cladograms produced by the data set-criterion-heuristic combination are shown in Fig. 10. These are summarized in Tables 1 and 2. As an example of comparative execution times, the 5S data set of Redelings and Suchard (2005) required 8 s on a 2.4 GHz PIII running Linux to complete a simple search with TBR branch swapping using Direct Optimization and Fixed States, and 86 s for Iterative-Pass. Their 12 taxon EF-1α data set (as DNA not the amino acids originally analyzed) required 1072 s for Direct Optimization and 7 s for Fixed States, both supporting their “eocyte” result.

Morphological analysis

Parsimony analysis of the 303 morphological characters resulted in 12 cladograms of length 596, while

Table 2
Topological incongruence among data partitions under parsimony and likelihood

		Topological congruence	
Optimality	Method	TILD	RILD
Parsimony	DO	0.495	0.691
	IP	0.422	0.699
	FS	0.450	0.751
	SBO	0.427	0.709
Likelihood	DO	0.380	0.637
	IP	0.357	0.601
	FS	0.415	0.687
	SBO	0.409	0.703

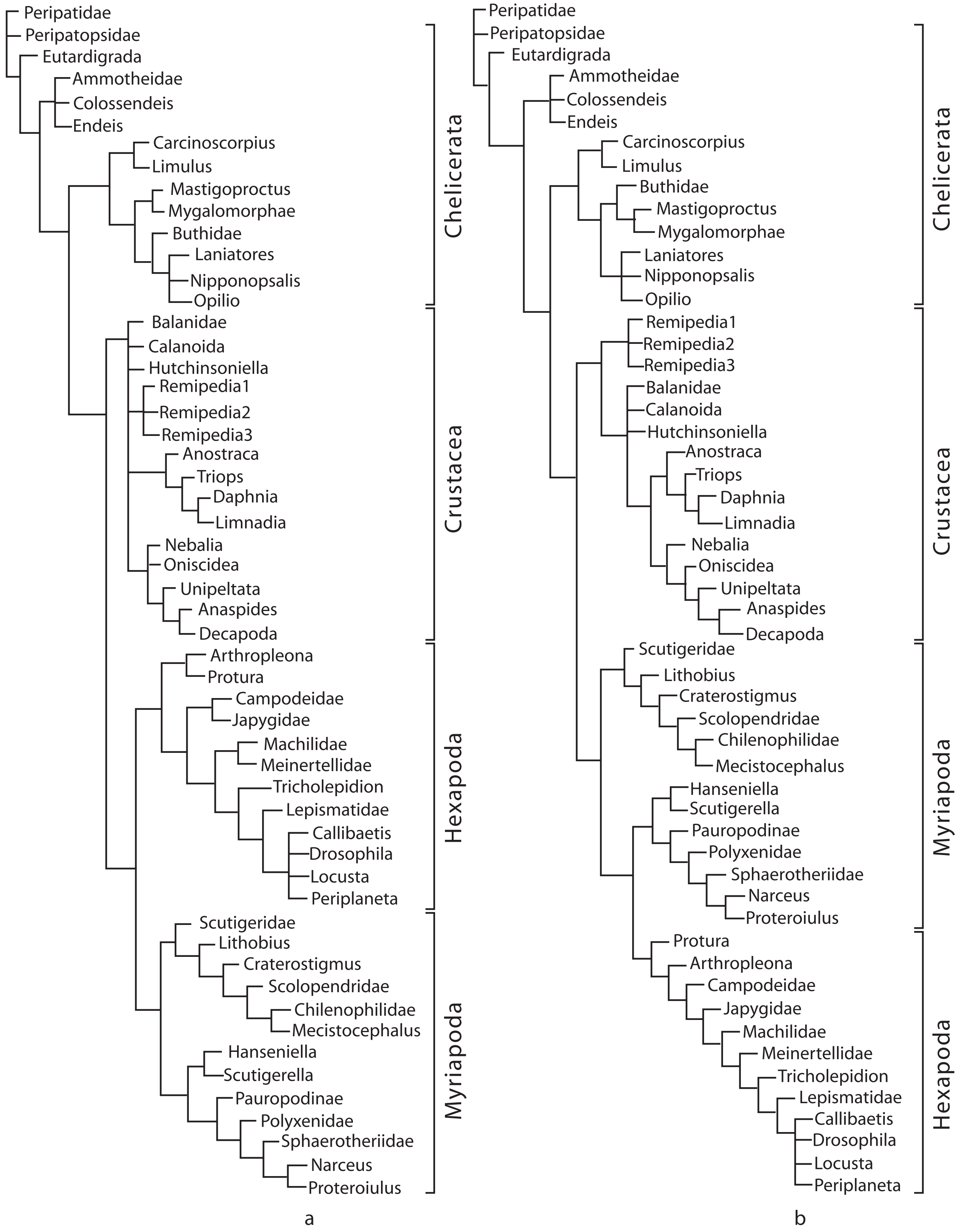


Fig. 9. Parsimony (a) and likelihood (b) analysis of arthropod morphological data Giribet et al. (2001). The likelihood analysis was performed using the model of Tuffley and Steel (1998). The differences between the two cladograms are due to the differential weight factor assigned to transitions by likelihood based on the number of states (*r*) in each character.

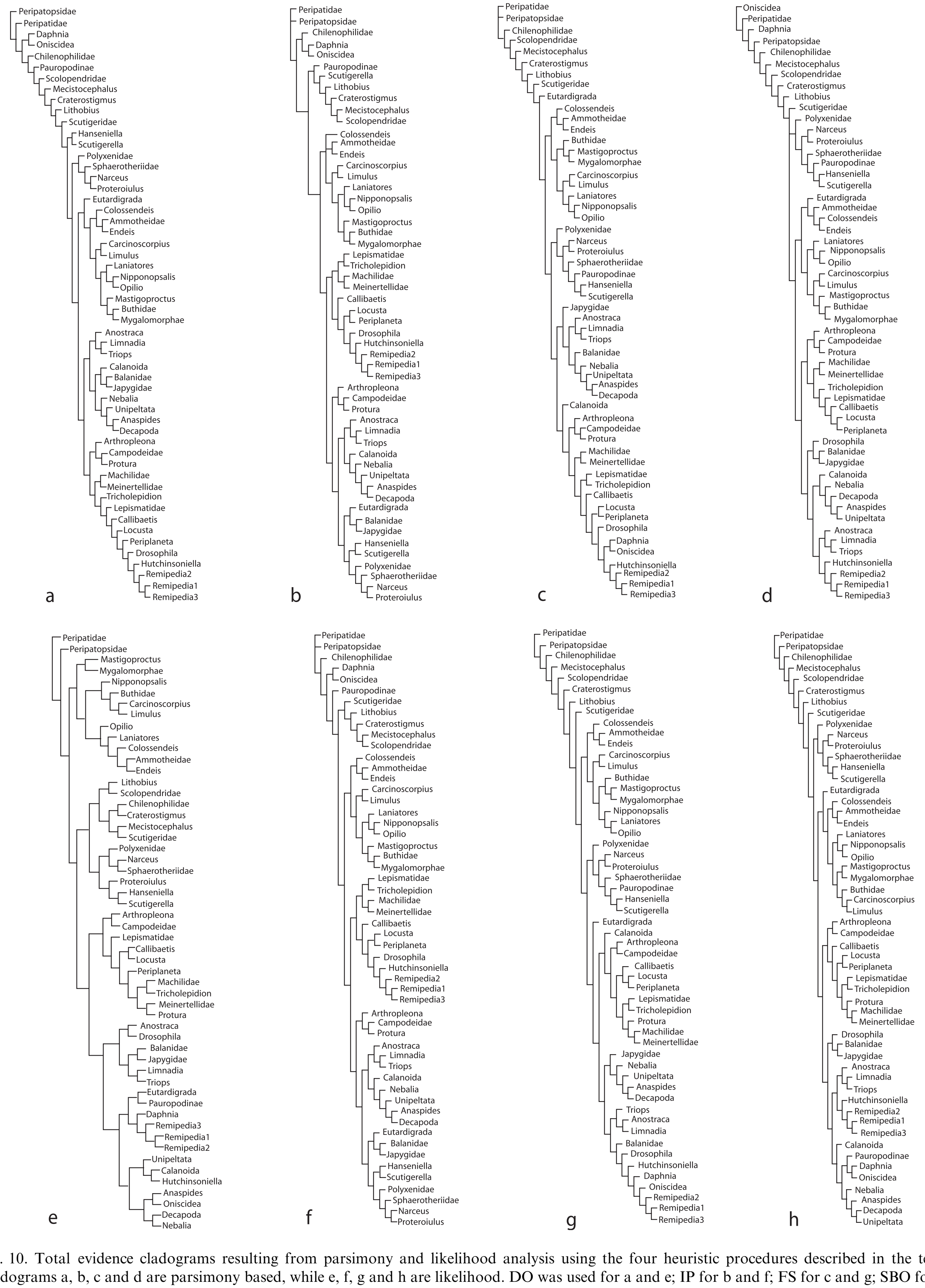


Fig. 10. Total evidence cladograms resulting from parsimony and likelihood analysis using the four heuristic procedures described in the text. Cladograms a, b, c and d are parsimony based, while e, f, g and h are likelihood. DO was used for a and e; IP for b and f; FS for c and g; SBO for d and h.

likelihood analysis resulted in 6 cladograms of cost 766.9 (-log likelihood; Fig. 9). The consensus cladograms of these two analyses are largely similar, but with a few differences (such as myriapod monophyly). It is often said that the Tuffley and Steel (1998) likelihood model shows the identity of parsimony and likelihood however, this is not strictly the case. In Tuffley and Steel (1998), the weight function of the contribution of an individual change is inversely proportional to the number of states. Although a given character or a suite of characters with identical numbers of states will yield identical results, data sets without such homogeneity will not, hence the differences between Figs 9a and b. The differences are subtle but real.

Molecular loci

The individual cladograms for the molecular partitions-optimality-heuristic procedure are contained in the accessory materials. In general, the results of IP analyses were superior (lower cost) to those of DO for parsimony but not under likelihood. SBO outperformed FS for both parsimony and likelihood.

Combined analysis

Strict cladograms for these eight analyses are shown in Fig. 10. The parsimony and likelihood results for SBO were identical. As expected, IP outperformed DO for parsimony runs. FS were more costly than IP, DO and SBO. For the likelihood analysis, SBO was by far the lowest cost (minimum -log lik) at 109 441.22 versus 113 090.23 for DO (Table 1). This is consistent with the MAL version of likelihood employed here. SBO summed the likelihoods (however marginal) of a larger set of potential HTU sequences than FS.

It is somewhat surprising that SBO outperformed all the other heuristics, which was not the case for parsimony. While the parsimony scores for the heuristics can be compared (since the tree lengths all represent the same weighted sum of events), it is not clear if the likelihoods can be. Certainly FS and SBO can be compared, as they are attempting to do the same thing (using the same form of likelihood and HTU determination). The estimation heuristics (DO and IP) are approaching the likelihood problem in a very different way.

Comparison

The numerical values (character congruence) produced by likelihood and parsimony analyses are largely incomparable. In order to assess their behavior, a comparison can be made, however, through an analysis of topological consistency among data partitions. Such an analysis was performed here using the topological incongruence metrics of Wheeler (1999b; Table 1).

Overall, parsimony outperformed likelihood, with each heuristic procedure having higher topological congruence values (TILD or RILD) for the parsimony version. These differences are not great (about 9% for the TILD values and 7.5% for RILD) and it is hard to say whether such distinctions are significant in a statistical sense. Furthermore, given the abbreviated nature of the searches and parameter space explored, this example is more of a demonstration than a general analysis of a congruence-based comparison technique.

The heuristics that yielded the highest congruence are the Search-Based FS and SBO. This holds for both parsimony and likelihood. While SBO also yielded the best likelihood score, this was not true of parsimony. The Search-Based techniques have the virtue of converging on an exact solution as the potential state set enlarges, for both optimality criteria. This convergence behavior may be keeping the partitions more consistent with each other as they add additional HTU options.

Discussion

The simple models presented here show that likelihood methods can be applied to scenarios of dynamic homology and combined analysis for real-sized data sets. This enlarges the world of problems amenable to likelihood analysis and allows us to bring to bear the greatest diversity of evidence to systematic problems within this probabilistic framework. More parameterized models (including affine gaps, lineage heterogeneity, etc.) might well improve the quality of the likelihood results, but the approximate techniques employed here are certainly usable now and provide a productive heuristic for more elaborate and time-consuming procedures.

Acknowledgments

I would like to thank Illya Bomash, Louise Crowley, Torsten Erikson, Gonzalo Giribet, Pablo Goloboff, Taran Grant, Cyrille D'Haese, Daniel Janies, Jerôme Murienne, Vamsikrishna Kalapala, Kurt Pickett, Usman Roshan, Wm. Leo Smith, Steven Thurston, Andrés Varón and two anonymous reviewers for many helpful critisms of the manuscript and the support of the Fundamental Space Biology Program of NASA.

References

Barry, D., Hartigan, J., 1987. Statistical analysis of hominid molecular evolution. Stat. Sci. 2, 191–210.

Bishop, M.J., Thompson, E.A., 1986. Maximum likelihood alignment of DNA sequences. J. Mol. Biol. 190, 159–165.

Bremer, K., 1994. Branch support and tree stability. Cladistics, 10, 295–304.

Chang, J.T., 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Bio. 134, 189–215.

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.

Edgecombe, G.D., Giribet, G., 2004. Molecular phylogeny of australasian anopsobiine centipedes (Chilopoda: Lithobiomorpha). Inv. Syst. 18, 235–249.

Farris, J.S., 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22, 250–256.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics, 12, 99–124.

Fleissner, R., Metzler, D., von Haeseler, R., 2005. Simultaneous statistical mulitple alignment and phylogeny reconstruction. Syst. Biol. 54, 548–561.

Giribet, G., Edgecombe, G.D., Wheeler, W.C., 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature, 413, 157–161.

Goldman, N., 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of dna substitution and to parsimony analysis. Syst. Zool. 39, 345–361.

Goloboff, P., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics, 15, 415–428.

Goloboff, P.A., 2003. Parsimony, likelihood, and simplicity. Cladistics, 19, 91–103.

Hein, J.C., Jensen, J.L., Pedersen, C.N.S., 2003. Recursions for statistical multiple alignment. Proc. Natl Acad. Sci. USA, 100, 14960–14965.

Hein, J., Wiuf, C., Knudsen, B., Moller, M.B., Wibling, G., 2000. Statistical alignment: computational properties, homology testing, and goodness-of-fit. J. Mol. Biol. 302, 265–279.

Huelsenbeck, J.P., Ronquist, F., 2003. MrBayes: Bayesian inference of phylogeny, 3.0 edition. Program and documentation available at http://morphbank.uuse/mrbayes/.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, N.H. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

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

Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature, 431, 980–984.

Lewis, P.O., 2001. A likelihoods approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913–925.

McGuire, G., Denham, M.C., Balding, D.J., 2001. Balding: Models of sequence evolution for dna sequences containing gaps. Mol. Biol. Evol. 18, 481–490.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48, 443–453.

Neyman, J., 1971. Molecular studies in evolution: a source of novel statistical problems. In: Gupta, S.S., Yackel, J. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, pp. 1–27.

Okusu, A., Schwabe, E., Eernisse, D.J., Giribet, G., 2003. Towards a phylogeny of chitons (mollusca, polyplacophora) based on combined analysis of five molecular loci. Org. Divers. Evol. 3, 281–302.

Redelings, B.D., Suchard, M.A., 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54, 401–418.

Sankoff, D.M., 1975. Minimal mutation trees of sequences. SIAM. J. Appl. Math, 28, 35–42.

Sankoff, D.M., Rousseau, P., 1975. Locating the vertices of a Steiner tree in arbitrary space. Math. Program. 9, 240–246.

Slowinski, J.B., 1998. The number of multiple alignments. Mol. Phylogen. Evol. 10, 264–266.

Steel, M., 2002. Some statistical aspects of the maximum parsimony method. In: Desalle, R., Giribet, G., Wheeler, W.C. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhauser-Verlag, Basel, Switzerland, pp. 124–140.

Steel, M., Hein, J., 2001. Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. Appl. Math. Let. 14, 679–684.

Steel, M., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.

Swofford, D.L., 2003. PAUP*: Phylogenetic analysis using parsimony (*and other methods), Version 4.0b.10. Sinauer Associates, Sunderland, MA.

Tavaré, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lec. Math. Life Sci. 17, 57–86.

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of dna sequences. J. Mol. Evol. 33, 114–124.

Thorne, J.L., Kishino, H., Felsenstein, J., 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34, 3–16.

Tuffley, C., Steel, M., 1998. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.

Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Computational Biol. 1, 337–348.

Wheeler, W.C., 1996. Optimization alignment: The end of multiple sequence alignment in phylogenetics? Cladistics, 12, 1–9.

Wheeler, W.C., 1999a. Fixed character states and the optimization of molecular sequence data. Cladistics, 15, 379–385.

Wheeler, W.C., 1999b. Measuring topological congruence by extending character techniques. Cladistics, 15, 131–135.

Wheeler, W.C., 2001. Homology and the optimization of DNA sequence data. Cladistics, 17, S3–S11.

Wheeler, W.C., 2003a. Iterative pass optimization. Cladistics, 19, 254–260.

Wheeler, W.C., 2003b. Search-based character optimization. Cladistics, 19, 348–355.

Wheeler, W.C., 2005. Alignment, dynamic homology, and optimization. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, pp. 73–80.

Wheeler, W.C., Aagesen, L., Arango, C.P., Faivoich, J., Grant, T., D'Haese, C., Janies, D., Smith, W.L., Varón, A., Giribet, G., 2005a. Dynamic Homology and Systematics: A Unified Approach. American Museum of Natural History.

Wheeler, W.C., Aagesen, L., Arango, C.P., Faivovich, J., Grant, T., D'Haese, C.A., Janies, D., Smith, W.L., Varón, A., Giribet, G., 2005b. Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY. American Museum of Natural History.

Wheeler, W.C., Gladstein, D.S., Laet, J.D., 1996–2003. POY, 3.0.11 edition, 1996–2003. ftp.amnh.org/pub/molecular/poy (current version 3.0.11). Documentation by D. Janies and W. Wheeler. Commandline Documentation by J. De Laet and. W. C. Wheeler.

Yang, Z., 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39, 105–111.