WILEY-BLACKWELL

Cladistics (2013) 1–9

Cladistics

10.1111/cla.12046

Maximum a posteriori probability assignment (MAP-A): an optimality criterion for phylogenetic trees via weighting and dynamic programming

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 06 June 2013

Abstract

One of the most time-consuming aspects of Bayesian posterior probability analysis in the analysis of phylogenetic trees is the use of Metropolis-coupled Markov chain Monte Carlo (MC^3) methods to determine relative posteriors and identify maximum a posteriori (MAP) trees. Here, analytical and numerical methods are presented to determine tree likelihoods that are integrated over edge-length (and other parameter) distributions. Given topological (tree) priors (flat or otherwise), this allows for identification of the maximum posterior probability assignment (MAP-A) of character states to non-leaf tree vertices via dynamic programming. Using this form of posterior probability as an optimality criterion, tree space can be searched using standard trajectory techniques and heuristically optimal MAP-A trees can be identified with considerable time savings over MC^3. Example cases are presented using aligned and unaligned molecular sequences as well as combined molecular and anatomical data.

Introduction

There is a variety of Bayesian (posterior probability) approaches employed in the analysis of phylogenetic trees (e.g. Edwards, 1970; Farris, 1973; Harper, 1979; Smouse and Li, 1987; Wheeler, 1991; Rannala and Yang, 1996; Larget and Simon, 1999; Ronquist et al., 2011). These may be divided into those that present clade or subtree posterior probabilities (Larget and Simon, 1999), termed clade-Bayes (Wheeler and Pickett, 2008); and those that present topologies with optimal tree posterior probability, termed maximum posterior probability (MAP; Rannala and Yang, 1996) or topology-Bayes (Wheeler and Pickett, 2008). The identification of optimal phylogenetic trees is well known to be a NP-hard problem (for parsimony by Foulds and Graham, 1982; for likelihood by Addario-Berry et al., 2004; Roch, 2006; Chor and Tuller, 2006), hence only heuristically optimal solutions can be found. Subtree-based measures are often employed as statements of support (Wheeler, 2010), while only the tree-optimality MAP can participate directly (by transitive pairwise competition) in hypothesis testing and act as an optimality criterion to identify heuristically best trees.

Commonly used implementations (e.g. Ronquist et al., 2011) offer tools to explore both these forms of posterior probability (clade and tree), usually via the Metropolis-coupled Markov chain Monte Carlo (MC^3; Geyer, 1991) procedure. This technique relies on random walks within a simulated annealing environment and can be quite time consuming when done properly, requiring exponential time in terms of number of characters to reach stationarity even in very small cases (five taxa with characters drawn from two trees under Jukes and Cantor, 1969 model; Mossel and Vigoda, 2005, 2006). For n taxa and k cliques of characters, $(2n–5)!!^k$ random starting points will be needed to sample the posterior distribution properly. The cause of this is the need to sample adequately the distributions of multiple parameters such as topology, branch length, and character transition models. These issues add time complex-

*Corresponding author:
E-mail address: wheeler@amnh.org

ity to maximum likelihood analysis as well, but in the form of the identification of point estimates of tree and model parameters.

When performed by MC^3, the sampling of parameter space takes place during the phylogenetic analysis. The examination of a candidate topology is based on a sampled value for each of the stochastic parameters, and is likely to be evaluated multiple times. This repeated tree calculation adds a large factor to the time complexity of the operation.

Here, I present analytical and numerical methods to permit the integration of various stochastic parameters before tree search, yielding character transformation cost matrices suitable for analysis with dynamic programming. This procedure identifies the maximum a posteriori probability assignment (MAP-A) of non-leaf vertex character states in an efficient manner. The resulting form of MAP-A tree optimality can be used to identify heuristically best solutions via standard tree-searching techniques. This tree search in all likelihood remains NP-hard, but is liberated from the stationarity requirements of MC^3-based analyses.

A simple case: the Neyman model and branch lengths

There are multiple models of nucleic acid sequence evolution that have been used in statistical phylogenetics. These range from completely homogeneous models of change with all nucleotides (A, C, G, and T) having the same stationary frequencies (0.25) and equal probabilities of mutation (Jukes and Cantor, 1969), to those with biases in favour of some transformations (transitions and transversions: HKY; Hasegawa et al., 1985), to highly parameterized models with potentially unique probabilities for all transitions and stationary frequencies (general time-reversible, GTR; Tavaré, 1986). The Neyman (1971) model is a generalization of the Jukes–Cantor for r states, where $r = 4$ for DNA.

Four taxa, one character, two states

Consider the simple case of four taxa with a single binary character evolving under the Neyman (1971) model (equation 1) yielding the probabilities of character change over an edge (p_{ii} and p_{ij} with μ as mutation rate, t time, and r the number of states):

$$p_{ii} = \frac{1}{r} + \frac{(r-1)}{r} e^{-\mu t}$$

$$p_{ij} = \frac{1}{r} - \frac{1}{r} e^{-\mu t} \tag{1}$$

with edge parameter μt and $r = 2$ in this case.

Since μ and t always appear as a product, we can scale this factor and represent it by t, the expected number of changes per site, on the edge (as will be done subsequently). In this scenario, there are three possible trees, each of which has four possible assignments of states to its internal vertices (Fig. 1).

If we take this model to be fixed, and the prior probability of trees to be flat ($\forall i,j; \mathrm{pr}(T_i) = \mathrm{pr}(T_j)$, the posterior probability of a given tree ($T_i \in \tau$, the set of all trees) is then given by equation 2 with model θ (e.g. Neyman, GTR), data (D), and edge parameter distribution for the set of all edges in a tree ($\mathbf{t}_T$).

$$p(T_i \mid D) = \frac{\int p(\mathbf{t}_{T_i} \mid \theta, T_i) p(D \mid \theta, T_i, \mathbf{t}_{T_i}) d\mathbf{t}_{T_i}}{\sum_{T_j \epsilon \tau} \int p(\mathbf{t}_{T_j} \mid \theta, T_j) p(D \mid \theta, T_j, \mathbf{t}_{T_j}) d\mathbf{t}_{T_j}} \tag{2}$$

In this example, we can consider two popular continuous distributions for t, exponential and uniform with θ = Neyman. In these cases, the distribution of branch lengths is independent of the particular tree of which it is a component, so all edges in $\boldsymbol{t}$ have the same distribution. For parameter a, these distributions are given as:
exponential:

$$pr(t) = ae^{-at} \text{ for } t = [0, \infty)$$

uniform:

$$pr(t) = 1/a \text{ for } t = [0, a]$$

The integrated probabilities (over t) of change and no-change for a given character over an edge under the Neyman model can then be determined. For the exponential distribution with parameter a (and μt):

$$P_{ii} = \int p_{ii} dt = \int_0^\infty \left[\frac{1}{r} + \frac{r-1}{r} e^{-t}\right] ae^{-at} dt$$

$$P_{ij} = \int p_{ij} dt = \int_0^\infty \left[\frac{1}{r} - \frac{1}{r} e^{-t}\right] ae^{-at} dt$$

which evaluate to:

$$P_{ii} = (ar + 1)/[(a + 1)r] \tag{3}$$

$$P_{ij} = 1/[(a + 1)r]$$

For the uniform distribution with parameter a:

$$P_{ii} = \int p_{ii} dt = \int_0^a \left[\frac{1}{r} + \frac{r-1}{r} e^{-t}\right] \frac{1}{a} dt$$

$$P_{ij} = \int p_{ij} dt = \int_0^a \left[\frac{1}{r} - \frac{1}{r} e^{-t}\right] \frac{1}{a} dt$$

which evaluate to:

$$P_{ii} = [a - (r-1)(e^{-a} - 1)]/ar$$

$$P_{ij} = [a + e^{-a} - 1]/ar$$

With these integrated probabilities, we can calculate the probabilities of assignments and trees of Fig. 1 (Table 1) for a variety of values of a.

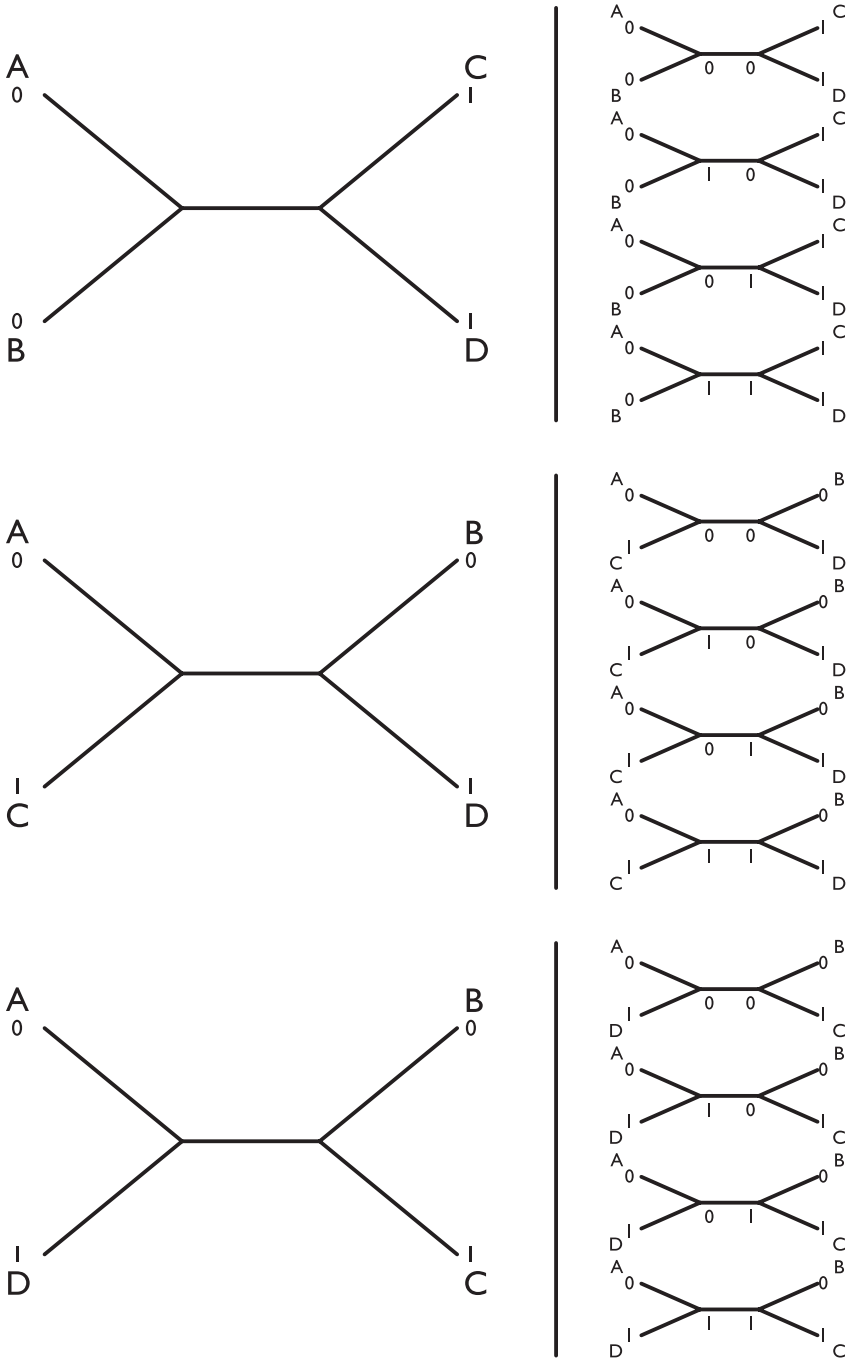

Fig. 1. The case of four taxa and one binary character showing the three possible trees (left) and the four possible assignments of internal node states for each tree (right).

As the expected branch lengths ($a/2$ and $1/a$ for uniform and exponential, respectively) grow longer, the probabilities of change and non-change become increasingly close, and the posterior probabilities of the three trees converge to $1/3[p_{ii} = p_{ij} = (1/r)]$ as $a \to \infty$ for uniform and $a \to 0$ for exponential. The assignments with maximum probability are always 0–1 for $AB|CD$ and both 0–0 and 1–1 for $AC|BD$ and $AD|BC$, given the initial assignments to the leaves in Fig. 1.

Although $AB|CD$ is the optimal MAP-A choice in all cases examined here (if marginally for very long branches), it is possible that summing the suboptimal solutions could yield a choice different from the maxi-

Table 1
Assignment probabilities for the three trees and four assignments of internal states for Fig. 1

		a parameter and distribution					
Tree	Assignment	10.0 Exp.	0.1 Uni.	5.0 Exp.	0.2 Uni.	0.1 Exp.	10.0 Uni.
AB|CD	0–0	0.001797	0.000544	0.005349	0.001899	0.033530	0.033691
	0–1	0.037737	0.021931	0.058839	0.038653	0.040236	0.041177
	1–0	0.000000	0.000000	0.000004	0.000000	0.019404	0.018453
	1–1	0.001797	0.000544	0.005349	0.001899	0.033530	0.033691
	Sum	0.041331	0.023018	0.069541	0.042451	0.126699	0.127012
	Posterior	0.845885	0.911739	0.748702	0.841951	0.340055	0.341396
AC|BD	0–0	0.001797	0.000544	0.005349	0.001899	0.033530	0.033691
	0–1	0.000086	0.000013	0.000486	0.000093	0.027941	0.027566
	1–0	0.000086	0.000013	0.000486	0.000093	0.027941	0.027566
	1–1	0.001797	0.000544	0.005349	0.001899	0.033530	0.033691
	Sum	0.003765	0.001114	0.011671	0.003984	0.122942	0.122513
	Posterior	0.077057	0.044130	0.125649	0.079025	0.329973	0.329302
AD|BC	0–0	0.001797	0.000544	0.005349	0.001899	0.033530	0.033691
	0–1	0.000086	0.000013	0.000486	0.000093	0.027941	0.027566
	1–0	0.000086	0.000013	0.000486	0.000093	0.027941	0.027566
	1–1	0.001797	0.000544	0.005349	0.001899	0.033530	0.033691
	Sum	0.003765	0.001114	0.011671	0.003984	0.122942	0.122513
	Posterior	0.077057	0.044130	0.125649	0.079025	0.329973	0.329302

"Sum" = total probability summed over the four assignments for that tree; "posterior" = posterior probability for that tree (sum divided by total of three trees). "Exp." and "Uni." denote exponential and uniform distributions with commonly used parameter values.

mal assignment (they are different optimality criteria, after all). This might occur with long (near-randomized, $\mu t = 10$ or so) expected branch lengths and large alphabets (as might be found in amino acid or developmental [Schulmeister and Wheeler, 2004;] sequence data). However, such cases are likely to have very low odds support for tree edges. This same suboptimal solution issue can occur with likelihood alignment ("dominant" versus "total" likelihood) (Thorne et al., 1991; Wheeler, 2006) and tree search (most parsimonious likelihood, MPL; maximum average likelihood, MAL) (Barry and Hartigan, 1987).

Dynamic programming

As with other forms of character-transformation weighting where the total cost is the sum of the weighted character changes over the tree, the MAP-A can be identified by dynamic programming. By using a logarithmic transform of the integrated character-change probabilities, the overall probability can be determined by summing the logarithmically weighted transformations.

In the case of the exponential distribution with parameter $a = 10$ (corresponding to a commonly used expected branch length parameter of 0.1), the negative logarithm of the probabilities of change (for $r = 2$; $P_{ii} = 0.9545$, $P_{ij} = 0.04546$) is used as a weighting scheme ($W = \{w_{ii} = 0.04652,\ w_{ij} = 3.091\}$) for determining the minimum assignment cost for a tree. To be precise, as with MAP determination, we are identifying the mode of the posterior distribution with a 0/1 loss function. In this way, we do not need to determine the value of the partition function (denominator of Bayes theorem) to identify the optimal value.

The dynamic programming method of Sankoff and Rousseau (1975) assumes that W is metric, which it is not, given the non-zero diagonals ($w_{ii} > 0$). The cost matrix, however, is symmetric ($w_{ij} = w_{ji}$) with non-zero transformation costs ($\forall_{i,j};\ w_{ij} > 0$), and conforms to the triangle inequality (provably for uniform and exponential, equation 4, and verifiable numerically for more complex models on a case-by-case basis).

$$\forall_{i,j,k} :\ w_{ij} + w_{jk} \geq w_{ik} \tag{4}$$

since $P_{ij} = P_{jk} = P_{ik}$ for exponential and uniform distributions;

$$2w_{ij} \geq w_{ij}$$

which will always be true since $w_{ij} > 0$.

Due to the presence of the non-zero identities, the dynamic programming procedure of Sankoff and Rousseau (1975) requires slight modification. The post-order pass from leaves down is unmodified until the root vertex is reached. At the root, the identity transforms (w_{ii}) are doubly counted (the root vertex adds an additional edge to the unrooted tree) and hence must be subtracted from the calculated tree cost. This factor is simply the sum of the identity costs over the number of characters (for m characters $c_{0,\ldots,m-1} \sum_{i=0}^{i=m-1} w_{c_i,c_i}$).

This tree cost calculation procedure can be extended over tree space to find the tree with the maximum value for MAP-A, which would be $1/e$ raised to the tree cost (since the weights are negative logs of probabilities).

Sequence characters

The case of a single binary character can be extended to sequence data, again using the Neyman model, this time with $r = 5$ for the nucleotides A, C, G, T, and the indel or gap '–'. A modified direct optimization algorithm (DO; Wheeler, 1996; Varón and Wheeler, 2012) can then be used to identify heuristic (this optimization is also NP-hard; Wang and Jiang, 1994) sequence medians (non-leaf assignments) and calculate an upper-bound MAP-A for a tree with unaligned sequences as leaves (terminal taxa) in a Bayesian dynamic homology analysis (Wheeler, 2001).

There are three steps to this process. First, the set of sequence transformation costs must be determined. Second, these costs are applied via the DO algorithm to identify sequence medians; and third, a post-order traversal of the tree is performed to determine the MAP-A for the DO tree medians.

Sequence transformation costs. In order to identify the cost of sequence medians in the following steps, the cost of individual sequence state assignments is required. This is the cost of assigning a given state or combination of states to a vertex with two descendent states (or combination of states). Hence, there are three elements to the cost: the vertex assignment state (s) and the two descent states. The overall cost of the assignment is the sum of the minimum transformation costs between that assignment and each of the child states. To make this determination efficient, the cost of all possible assignment states (and their combinations —the power set) given all possible descendent states are precalculated (equation 5; Varón and Wheeler, 2012).

$$\forall_{i,j,k \in \wp \geq 1} \{A, C, G, T, -\} : \quad c_{i,j,k} = w_{ij} + w_{ik} \quad (5)$$

For an exponential distribution with $a = 10.0$ and $r = 5$, the costs would be $w_{ii} = 0.07551$ and $w_{ij} = 4.007$.

Direct optimization and median construction. The DO algorithm assigns median sequences during paired post-order (down) and pre-order (up) traversals of the tree. During the post-order pass, a sequence median is constructed such that the median assignment minimizes the cost to the two descendent sequences. This is accomplished via dynamic programming using a modified string-matching algorithm (Wheeler, 1996). The original algorithm assumed that all match events (e.g. A → A) occurred with zero cost (Wheeler, 1993). The MAP-A case here does not share this property but, since the cost of assignment (equation 5) takes into account match cost, this does not cause a problem. It is worth noting that there may be multiple elements assigned to a specific median position. For instance, under the Neyman model the median for two sequences A and G would be either A or G, represented as R in IUPAC code.

Tree traversal and MAP-A determination. Each internal (non-leaf) vertex is assigned a median sequence based on its two descendant and one ancestor sequences, and the total cost of the tree is the sum of the edge (costs) between each pair of vertices. As with the pre-aligned or non-sequence case, care must be taken at the root not to overcount matching events. This can be overcome by simply treating one of the leaf sequences as the root sequence. This optimization problem-tree alignment (Sankoff, 1975) is known to be NP-hard (Wang and Jiang, 1994), hence the DO procedure is a heuristic upper bound on tree cost.

Consider the case of four sequences A, A, AC, and AC under a Neyman model with $r = 5$ and a uniform branch length distribution ($a = 10$). The MAP-A tree and assignment (Fig. 2) results in a cost of 4.536 (six A → A, two C → C, and one C-insertion event; posterior probability = 0.01072). If the assignment were based on multiple sequence alignment, the cost increases (integrated likelihood decreases) due to the extra events required (2 × – → –).

It is important to emphasize that this procedure identifies a single assignment of maximum posterior probability and is not the total posterior for the tree. The value is akin to the "dominant" likelihood in sequence comparison (Thorne et al., 1991). Determination of the total posterior probability would require (as in Fig. 1) the sum over all possible assignments. For DO analysis of sequence characters, this presents a problem. While it is possible to determine the total posterior probability for a given pair of sequences (by summing probabilities during median calculation; Wheeler, 2006), the DO method does not represent relative probabilities of alternate state assignments or length distribution when median states are assigned. The algorithm determines subsequent assignments based on these medians. The cost calculation will be optimal for a given specific assignment, but this assignment is not guaranteed to be to be optimal for the entire tree, and there is a potentially exponential number of assignments that contribute to the total posterior probability.

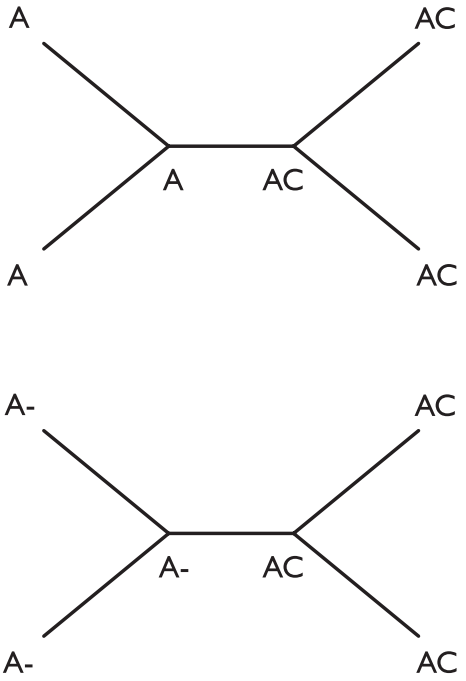

Fig. 2. The case of four simple sequences under a Neyman model with $r = 5$ and an exponential branch length distribution ($a = 10$). The upper case shows DO assignments to internal nodes (cost = 4.536, MAP-A = 0.01072) with a single indel on the central branch (insert or delete C); the lower case is based on multiple alignment (cost = 4.687, MAP-A = 0.009218) including the central indel (C ↔ –) and two gap matches (– ↔ –).

General case

The cases above employ restricted models, but the method of analysis they suggest can be extended to more complex scenarios. In the general case, the analytical integrations of the Neyman model can be replaced by numerical integration over a broad variety of distributions and parameters. These typically might include edge parameters (branch length), rate parameters including the discrete gamma distribution (with rate classes), invariant sites, and state transformation models (e.g. GTR). As with the sequence example above, indel parameters can be included as an additional state in sequence models to allow for analysis of unaligned sequences.

Since the result of this process is the transformation cost matrix, the numerical integration can be performed a single time for each set of parameters before a search is used to identity the MAP-A solution, as in the simple cases above.

To demonstrate the use of this approach for nucleotides and indels (five elements), analytical parameters were drawn with symmetric Dirichlet distribution (all parameters 1.0) for the five element priors and 10 instantaneous rate change parameters, uniform distribution [0.0–1.0] for invariant sites, uniform distribution [0.0–50.0] for the α-shape parameter, uniform discrete distribution [1–7] for rate classes, and exponential distribution ($a = 10.0$) for edge (branch) lengths. This process was repeated 2×10^8 times, in each case calculating the transition probabilities between each pair of elements (using the program MAPA, available with POY5). Due to the symmetry in the Dirichlet parameters for element priors and transition rates, the only probabilities that are asymptotically unique are those of an element remaining untransformed along an edge ($p_{ii} = 0.089422$) and that of transformation having occurred ($p_{ij} = 0.64231$). This results in transformed weights of $w_{ii} = 0.44268$ and $w_{ij} = 2.41438$.

An example: metazoan ribosomal DNA

A dataset of 208 complete metazoan 18S rDNA sequences (Giribet and Wheeler, 2001; Wheeler, 2007) was analysed using the MAP-A weight function and compared with other forms of analysis. An aligned version of the data was created using CLUSTAL ver. 2.0.012 (Larkin et al., 2007) under default parameters. The data were treated as both aligned and unaligned using parsimony with substitutions = indels = 1, and using MAP-A weights as determined above ($w_{ii} = 0.44268$, $w_{ij} = 2.41438$). In each case, a simple search with 10 random-addition Wagner build sequences, tree bisection and reconnection (TBR) branch swapping, and tree fusing were performed.

The analyses based on equally weighted parsimony (counting indels) show a typical pattern of multiple sequence alignment (MSA)-based analysis with higher cost than for DO-based analysis. In this case, the MSA-based tree is 16% more costly than the DO (30 990 versus 26 744). The same pattern follows for the MAP-A analysis with the MSA-based MAP-A of 584 316.21396 and the DO-based 380 663.93472. This difference is, in large part, due to the necessity of accounting for "gap-to-gap" events in the MSA, a situation that will never occur in a median optimization approach such as DO. If an adjustment is made to remove this contribution (by arbitrarily setting $w_{\text{gap} \rightarrow \text{gap}} = 0$), the difference in values still favours the DO-based tree, but by a smaller margin (380 663.93472 versus 385 973.23538)[1] . The MAP tree generated by MrBayes is included for comparative purposes, but its MAP value of 0.002660 is not directly comparable with the MAP-A numbers. The MAP-A score of this tree is 382 437.61768 (DO via diagnosis), which is very close to the cost of the DO-based MAP-A tree found with a search.

[1]This kludge allows for verification and comparison, but does not result from, or imply, any possible stochastic model of change.

Combined analysis: arthropod systematics

As a demonstration of how a combined ("total evidence" or simultaneous) analysis can be performed under MAP-A, the arthropod data set of Giribet et al. (2001) was used. In this data set of 54 taxa with eight molecular loci and 303 morphological characters, MAP-A weight matrices were created as above using an exponential distribution (with parameter 10.0) for edge lengths. Molecular sequence character models were as above in the initial sequence example ($w_{ii} = 0.07551$, $w_{ij} = 4.007$) and a Neyman-based model was created for the characters with alphabet sizes varying from two to 10 (equation 4). The results are shown in Fig. 3 for equally weighted parsimony (morphological changes = nucleotide substitutions = indels), MAP-A, and MAP (via MrBayes ver. 3.2.1).

Several groups are common to the three optimality criteria, including Pycnogonida, Chelicerata, Remipedia, and paraphyletic Crustacea and Myriapoda. Others are specific to parsimony (Entognatha) or parsimony and MAP (Tetraconata). While the myriapods are never (in these analyses) monophyletic, several lineages (e.g. Pauropodinae) group with the entognathan Protura. The MAP-A tree is more similar to that of parsimony (Robinson–Foulds distance, RF = 40) than either is to MAP (RF = 56) (via TREEDIST in Phylip ver. 3.69, Felsenstein, 2004 and POY5β, Varón et al., 2013).

Group support

Similarly to likelihood ratios (Wheeler, 2006), the commonly used Bremer support (Goodman et al., 1982; Bremer, 1994), where tree costs are expressed as the –log of posterior probabilities, yields the log of the Bayes factor for each subtree (difference in log of MAP-A cost with and without subtree). That is, the

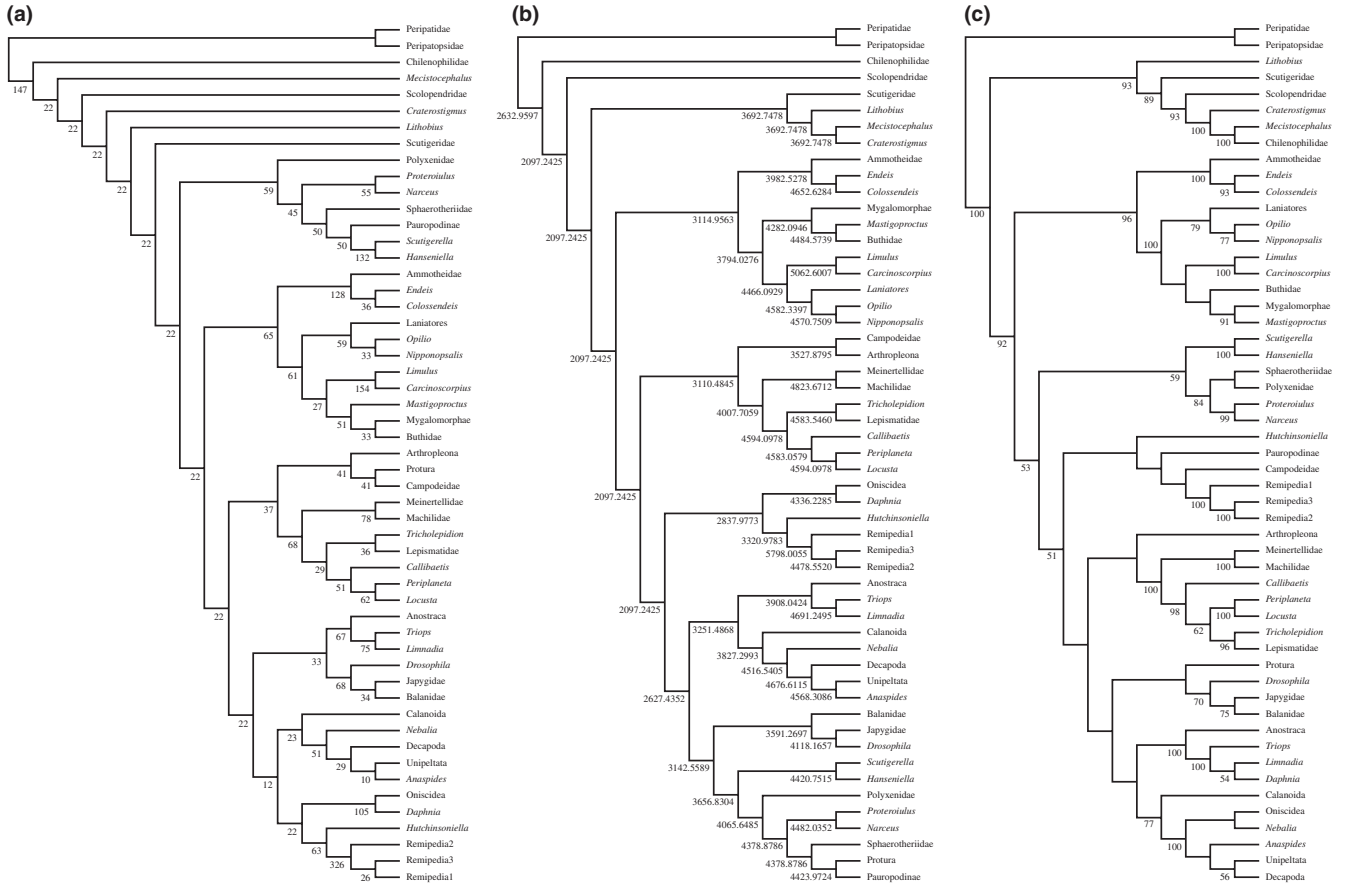

Fig. 3. Combined arthropod analysis under equally weighted parsimony (a), MAP-A (b), and MAP (c). MP and MAP-A using direct optimization and POY5β, MAP based on MSA (CLUSTAL) and MrBayes ver. 3.2.1. In the MP and MAP-A cases, simple searches of 100 RAS + TBR with tree fusing were performed; in the MAP case, 200M generations with 2 × 4 chains. The parsimony tree (left) has a score of 29 945 (MAP-A = 160 476.65905, –ln units); the MAP-A tree (centre) 159 533.61236 (parsimony score = 30 108); and the MAP tree (right) has a posterior probability of 0.000197 (MAP-A = 162 772.33991; parsimony score = 30 553). Bremer support values (a), log MAP-A ratios (b), and clade posterior probabilities (c, when > 50%) are displayed beneath branches. Values are determined by comparison with trees in the TBR neighbourhood of the best tree.

log ratio of odds for the best tree found and the best tree without that subtree. This is shown in Fig. 3, where groups such as Pycnogonida, Chelicerata, and especially Remipedia have relatively high log odds compared with others such as Tetraconata ("Crustacea" + Hexapoda), which are more equivocal. Overall, group support (whether by Bremer, odds ratio, or clade-posterior probability) is fairly consistent in terms of well supported groups (e.g. Pycnogonida: 128 parsimony Bremer, 3982.7478 log odds ratio, 100% clade posterior probability) versus weakly supported ones (e.g. Tetraconata: 122 parsimony Bremer, < 1.0 log odds ratio, 51% clade posterior probability).

Heuristic utility

MAP-A searches can be achieved in much less time than MC^3-based MAP searches: CPU hours versus CPU days. A 100 Wagner build and TBR branch swap for the unaligned (using DO) combined arthropod data set (above) took 26 CPU hours on a 2.93 Ghz Intel HexaCore i7 desktop running OSX 10.7 (a single TBR swap takes 128 s). The 200M 2×4 chain runs of the aligned (CLUSTAL) data set with MrBayes 3.2.1 required 1084 CPU hours. These values are not meant to represent completely adequate analytical effort, but example timings.

Discussion

MAP-A is a novel form of posterior probability-based (Bayesian) optimality criteria. As such, we can use this value in hypothesis testing, hence the identification of heuristically useful phylogenetic scenarios. Unlike MAP, MAP-A is based on a single assignment of maximal posterior probability, not the integration of all possible vertex state assignments on a tree. Like MAP, MAP-A does not depend on the potentially exponential time complexity (Mossel and Vigoda, 2005, 2006) of numerical stationarity that clade-based posterior probabilities require. Hence MAP-A is a relatively efficient, optimality based method of identifying heuristically best phylogenetic trees and associated odds-based support values for subcomponents.

The weights used to identify MAP-A solutions can also be viewed as a posterior probability-based weighting function that can be used in typical phylogenetic analysis. Such a weighting scenario is employable in a dynamic or static homology analysis. The main difference between MAP-A and standard likelihood analysis is in the use of integrated branch lengths as opposed to point estimates. In this sense, MAP-A is more akin to integrated likelihood than to maximum average likelihood. The most important distinction between MAP-A weights and those used in standard parsimony analysis is the non-zero identity weights. Unlike familiar scenarios, there is a non-zero cost for a transformation between identical states. This is, of course, reasonable given stochastic models of change where the probability of non-change is non-zero.

Acknowledgements

I would like to thank Ronald Clouse, John Denton, Gonzalo Giribet, Prashant Sharma, Katherine St John, and two anonymous reviewers for many useful comments on the manuscript. This material is based on work supported by, or in part by, the US Army Research Laboratory and the US Army Research Office under grant number W911NF- 05-1-0271.

References

Addario-Berry, L., Chor, B., Hallett, M., Lagergren, J., Panconesi, A., Wareham, T., 2004. Ancestral maximum likelihood of evolutionary trees is hard. J. Bioinform. Comput. Biol. 2, 257–271.

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

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

Chor, B., Tuller, T., 2006. Finding a maximum likelihood tree is hard. J. ACM 53, 722–744.

Edwards, A.W.F., 1970. Estimation of the bracing points of a branching diffusion process. JR. Stat. Soc. B 32, 155–174.

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

Felsenstein, J., 2004. Phylip (phylogeny inference package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Foulds, L.R., Graham, R.L., 1982. The Steiner problem in phylogeny is NP-complete. Adv. Appl. Math. 3, 43–49.

Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood. In: Keramidas, E.M. (Ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium. Interface Foundation, Fairfax Station, VA, pp. 156–163.

Giribet, G., Wheeler, W., 2001. Some unusual small-subunit ribosomal RNA sequences of metazoans. Am. Mus. Novit. 3337, 1–14.

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

Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J., 1982. New perspectives in the molecular biological analysis of mammalian phylogeny. Acta Zool. Fennica 169, 19–35.

Harper, C.W., 1979. A Bayesian probability view of phylogenetic systematics. Syst. Zool. 28, 547–553.

Hasegawa, M., Kashina, H., Yano, T., 1985. Dating the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

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.

Larget, B., Simon, D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16, 750–759.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins,

D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.
Mossel, E., Vigoda, E., 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309, 2207–2209.
Mossel, E., Vigoda, E., 2006. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. Ann. Appl. Prob. 16, 2215–2234.
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. New York, Academic Press. pp. 1–27.
Rannala, B., Yang, Z., 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43, 304–311.
Roch, S., 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Trans. Comput. Biol. Bioinform. 3, 92–94.
Ronquist, F., Huelsenbeck, J.P., Teslenko, M., 2011. MrBayes: Bayesian inference of phylogeny 3.2.1. program and documentation, available at http://morphbank.uuse/mrbayes/.
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.
Schulmeister, S., Wheeler, W.C., 2004. Comparative and phylogenetic analysis of developmental sequences. Evol. Dev. 6, 50–57.
Smouse, P.E., Li, W.-H., 1987. Likelihood analysis of mitochondrial restriction-cleavage patterns for the human–chimparnzee–gorilla trichotomy. Evolution 41, 1162–1176.
Tavaré, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. 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.
Varón, A., Wheeler, W.C., 2012. The tree-alignment problem. BMC Bioinformatics 13, 293.
Varón, A., Lucaroni, N., Hong, L., Wheeler, W.C., 2013. POY 5.0 beta. American Museum of Natural History, http://research.amnh.org/scicomp/projects/poy.php.
Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348.
Wheeler, W.C., 1991. Congruence among data sets: a Bayesian approach. In: Miyamoto, M.M., Cracraft, J. (Eds.), Phylogenetic Analysis of DNA Sequences. Oxford University Press, Oxford, pp. 334–346.
Wheeler, W.C., 1993. The triangle inequality and character analysis. Mol. Biol. Evol. 10, 707–712.
Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.
Wheeler, W.C., 2001. Homology and the optimization of DNA sequence data. Cladistics 17, S3–S11.
Wheeler, W.C., 2006. Dynamic homology and the likelihood criterion. Cladistics 22, 157–170.
Wheeler, W.C., 2007. The analysis of molecular sequences in large data sets: where should we put our effort? In: Hodkinson, T.R., Parnell, J.A.N. (Eds.), Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa. Systematics Association/Oxford University Press, Oxford, pp. 113–128.
Wheeler, W.C., 2010. Distinctions between optimal and expected support. Cladistics 26, 657–663.
Wheeler, W.C., Pickett, K.M., 2008. Topology-Bayes versus clade-Bayes in phylogenetic analysis. Mol. Biol. Evol. 25, 447–453.