On Reduced Amino Acid Alphabets for Phylogenetic Inference

Edward Susko and Andrew J. Roger†*

*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada; and †Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

We investigate the use of Markov models of evolution for reduced amino acid alphabets or bins of amino acids. The use of reduced amino acid alphabets can ameliorate effects of model misspecification and saturation. We present algorithms for 2 different ways of automating the construction of bins: minimizing criteria based on properties of rate matrices and minimizing criteria based on properties of alignments. By simulation, we show that in the absence of model misspecification, the loss of information due to binning is found to be insubstantial, and the use of Markov models at the binned level is found to be almost as effective as the more appropriate missing data approach. By applying these approaches to real data sets where compositional heterogeneity and/or saturation appear to be causing biased tree estimation, we find that binning can improve topological estimation in practice.

Introduction

Most probabilistic models of amino acid evolution implicitly assume that it is the amino acids themselves that evolve according to a time-reversible Markov model. Gross departures from these assumptions can, and frequently do, give rise to incorrect topological estimation. Common difficulties include compositional heterogeneity (Jermiin et al. 2004) and saturation: the number of amino acid changes along lineages is so large that it effectively erases the historical signal (Ho and Jermiin 2004). More generally, the substitution process may not be homogeneous and could change in any number of ways throughout the tree. To deal with compositional heterogeneity and/or saturation, nucleotide data is often recoded into pyrimidine (Y) and purine (R) bins. This procedure has been shown, in some cases, to improve phylogenetic estimates (Phillips and Penny 2003). Recoding amino acids is also possible, although the choice of bins is not as obvious. There is thus some value in exploring formal ways of binning character states into groups that are more appropriate for evolutionary analysis in the presence of some of these difficulties for phylogenetic reconstruction. Good choices of alphabets or bins of amino acids can serve as diagnostic tools: large bootstrap support for splits under an amino acid model that are no longer supported under a binned model provide evidence that something is awry.

A set of bins that has received some attention in the literature are the "Dayhoff classes": *AGPST*, *DENQ*, *HKR*, *ILMV*, *FWY*, and *C*. Users of this choice of reduced alphabet in phylogenetic reconstructions include Hrdy et al. (2004), Martin et al. (2005), and Embley et al. (2003). The classes were obtained from a log odds matrix of probabilities of pairs of amino acids appearing together. Pairs of amino acids that are more likely to appear together in short evolutionary distances than random are expected to have log odds greater than zero. The classes are chosen so that they (almost) satisfy that log odds for pairs of amino acids within groups are greater than 0. The single exception to this rule is the pair *G* and *P* in the group *AGPST* (fig. 84 in Dayhoff et al. 1978). Other approaches to the construction of bins include those of Wang J and Wang W (1999), who obtained reductions in a more formal way based on the Miyazawa and Jernigan matrix (Miyazawa and Jernigan 1996). Their goals, however, were not evolutionary; the initial question they posed is how many residue types are required to form a structured protein? Cannata et al. (2002) gave a globally convergent Branch-and-Bound routine to derive bins for certain classes of criterion functions. The criteria used, however, are more directly related to scoring an alignment than evolutionary processes. Kosiol et al. (2003) presented methods most similar to those that will be used here in the construction of what we will refer to as the saturation bins, and their approach is similar in motivation to the construction of the Dayhoff classes. The approach taken by these authors resembles a rearrangement of the rows and columns of a substitution matrix to obtain an approximate block diagonal matrix with relatively rapid rates of exchange within blocks and slow rates of exchange between blocks.

An outline of the article is as follows. First, we consider methods for dealing with reduced alphabets in the absence of model misspecification. Next, we propose criteria for choices of bins. These come in 2 flavors: once-and-for-all choices of bins based on the properties of empirical amino acid rate matrices or bin choices based on the sequence data being analyzed. A brief simulation study is conducted to investigate how much information is lost through binning. The major sections of the paper conclude with several real data examples.

Methods for Dealing with Reduced Alphabets

For the bins considered in this paper, once the various amino acids are placed into a given bin, they are then treated as the same single character state. For instance, for the Dayhoff groups: *AGPST*, *C*, *FWY*, *HRK*, *MILV*, and *NDEQ*, we assign the bin 1 to any amino acid that is an *A*, *G*, *P*, *S*, or *T*. For any amino acid *C*, we assign the character state 2, and so forth. The simple 3-site alignment

Sequence 1 ARI
Sequence 2 DFV
Sequence 3 NWI

becomes

Sequence 1 145
Sequence 2 635
Sequence 3 635

Key words: protein evolution, amino acid alphabets, Markov models, compositional heterogeneity.

E-mail: susko@mathstat.dal.ca.

Mol. Biol. Evol. 24(9):2139–2150. 2007
doi:10.1093/molbev/msm144
Advance Access publication July 25, 2007

We first consider methods for using binned data in the absence of model misspecification at the amino acid level. There are 2 approaches to modeling binned data that we will consider: the missing data (MD) method and the binned Markov chain (BMC) method. The MD method is the more appropriate one in the absence of model misspecification but is not the natural choice otherwise. The methods differ primarily in the manner in which they do probability calculations. Maximum likelihood (ML) trees and/or ML distances, that are the primary methods used here, can be calculated given any one of the choices for probability calculations.

The MD method treats the amino acids within bins as MD and averages over what they may have been in much the same way that the MD is treated in standard phylogenetic analysis. For instance, for the Dayhoff groups: *AGPST*, *C*, *FWY*, *HRK*, *MILV*, and *NDEQ*, if bin 1 were the observed data for one taxon separated by evolutionary distance t from another taxon having bin 3, the probability of the observed bins for those 2 taxa would be the probability that the amino acid for the first taxon was *A*,*G*, *P*, *S*, or *T* and that the amino acid for the second was *F*, *W*, or *Y*. Under a Markov model for amino acids, this can be calculated as

$$\sum_{i\in AGPST}\sum_{j\in FWY}\pi_i P_{ij}(t)$$

where the π_i and $P_{ij}(t)$ are the stationary frequencies and substitution probabilities under the amino acid model. This is analogous to the summing over unobserved amino acids at internal nodes and corresponds to the widely agreed upon correct method for dealing with MD.

The BMC method treats the process of substitution of bins along an edge as a continuous time Markov chain and uses the rate matrices at the amino acid level to obtain rate matrices for the substitutions of bins. This method was used in Yang et al. (1998) to derive a rate matrix for a Markov model of amino acids based on a Markov model for codons.

For $q_{IJ}^{(b)}$ to be a rate from bin I to bin J, the probability that bin I is substituted with bin J should be $q_{IJ}^{(b)}h + o(h)$, for a small evolutionary distance h. Here $o(h)$ is the standard notation for a function of h that converges to 0 faster than h. Similarly as in Yang et al. (1998), if the original data were generated under a Markov model with rates q_{uv}, we obtain

$$q_{IJ}^{(b)} = \sum_{u\in I}\sum_{v\in J}\pi_u q_{uv} \Big/ \sum_{u\in I}\pi_u \qquad (1)$$

To see this, note that the Markov model for amino acids gives the probability, at the start of a small time interval $[t, t + h)$, that the amino acid is in bin I as $\sum_{u\in I}\pi_u$. The probability that the amino acid is u at time t and v at $t + h$ is $\pi_u q_{uv}h + o(h)$ and so the probability that it is in bin I at t and bin J at time $t + h$ is obtained by summing over u in I and v in J; the numerator of equation (1). The conditional probability of going to J from I is obtained as the ratio of the probability of bin I at time t and bin J at time $t + h$ divided by the probability that bin I is observed initially. This is exactly the ratio given in equation (1).

If a Markov chain model also applied for bins, using the rates in equation (1) for Markov chain–based calculations, it would give exactly the same probabilities as the MD method. However, although the rates in equation (1) may be appropriate, it is rarely the case that the Markov property holds; given the last 2 bins in the evolutionary process, the probability of the next bin will not be dependent only on the last.

To see that the binned process will rarely satisfy the Markov property, assume a JTT substitution model and suppose that one of the bins is the single amino acid V and that another is $\{I, P\}$. For the JTT substitution model, when the last amino acid is an I, the next will be a V 49% of the time, but when the last amino acid is a P, there is only a 2% chance the next will be a V. Given only that the last bin is $\{I, P\}$, calculations give a probability of 0.33 that the next amino acid will be V. However, given that the last amino acid was a V, the next will be an I 41% of the time and a P only 1% of the time. Thus, if we know that the last 2 bins are V and $\{I, P\}$, it is much more likely that the last amino acid was I and, consequently, that the next will be V; calculation gives the probability as 0.48.

Reduced Alphabets to Adjust for Saturation

A site in an alignment is considered saturated if many repeated substitutions have occurred over the period of time under consideration. Saturation usually refers to situations in which many sites are saturated with changes, leading to some edges being long. Sequences corresponding to such terminal edges can be considered to have been (almost) random or independent of the other sequences in the analysis, which makes their placement highly variable. Saturation is also known to create biases whereby saturated sequences are more likely to be placed, correctly or incorrectly, near other long branches in the tree (Susko et al. 2005; Wenzel and Siddall 1999). In cases where rates of substitution within a group of similar amino acids, say V, I, and L, is high, an ancestral character state of V will likely be multiply substituted along a long edge, but the group, V, I, and L are more likely to remain the same. Thus, with a reduced amino acid alphabet, ancestral nodes of long terminal edges will appear more similar to their child terminal nodes; the edges will be less susceptible to multiple substitutions of bins.

We illustrate with a simple, contrived example. Consider the following pair of sequences of length 1000:

Sequence 1 ETMIYDNKFC. . .MA
Sequence 2 EQLEWRDCTA. . .AT

Although not independent, one can clearly see a large number of differences in these sequences. The estimated evolutionary Jukes–Cantor (Jukes and Cantor 1969) distances were 2.48. (Although the Jukes–Cantor model was originally described for nucleotides, it can be applied with amino acids or, more generally, bins as well. It is the model in which the rate of substitution from any one bin to any other is constant.) The edge length for this simplest of trees, with only one edge, is quite large and the 2 sequences can be thought to be saturated with changes. If, however, we create bins for the amino acids based on the alphabetical order of the 3-letter codes for amino acids, with 5 amino acids in each bin (the first 5 alphabetically ordered amino

acids are placed in the first bin, the second 5 in the second bin, and so forth), the sequence data become

Sequence 1 2432411331...31
Sequence 2 2232411141...14

One can see that there is much more conservation between the sequences. The estimated Jukes–Cantor distances were 0.53. The example was created by simulating bins according to a Jukes–Cantor model, with evolutionary distance 0.5 between the sequences, and then, for a given bin, randomly assigning an amino acid from that bin.

The goal in adjusting for saturation is to create groups so that when multiple substitutions occur along edges, they will usually be substitutions within the same group. The empirically derived rates from the JTT substitution model (Jones et al. 1992) should give an indication of which groups of amino acids tend to see large numbers of substitutions within groups but smaller numbers of substitutions between groups. Indeed, the use of an empirically derived model like the JTT substitution model should ameliorate the effects of saturation because larger numbers of substitutions within groups are expected. Still, empirical models obtain rates by averaging over a large number of sites and a large number of proteins. Site-specific increases in the rates of exchange within groups may still lead to problems of saturation. In any case, we will choose bins so that the rate of substitution, under the JTT model, within bins is large and the rate of substitution between bins is small.

However, simply considering only the rates of substitution under the JTT model cannot be expected to be an effective way of choosing bins. We illustrate why here, using the Jukes–Cantor model as a baseline for comparison. In the Jukes–Cantor model, no bin choice should be preferred since, for an edge of length t, the expected number of substitutions is $t/380$ for a pair of amino acids, regardless of what those amino acids are. It is important to use the Jukes–Cantor model as a baseline for comparison because of size effects. For instance, for 2 bins, one with a single amino acid, there are $19 \cdot 18 = 342$ substitutions of amino acids within bins so that the expected number of substitutions within bins under the Jukes–Cantor model is $342t/380$. In contrast, with 2 bins each containing 10 amino acids, there are $2 \cdot 10 \cdot 9 = 180$ substitutions of amino acids within bins so that the expected number of substitutions within bins is $180t/380$.

The criterion that we use to select bins for saturation is to choose the bins that maximize the ratio of the expected number of substitutions within bins under the JTT model, W, to those expected number under the Jukes–Cantor model, W^{JC}. Alternatively, because W^{JC} is proportional to the number of pairs of amino acids within bins, the average rate of substitutions, within bins, is being maximized. The resulting bins are given in table 1.

It is interesting to note that the bins are related to the chemical properties of amino acids. For instance, the hydroxylated polar amino acids *S* and *T* are frequently together, the positively charged amino acids *K* and *R* always group together, as do the negatively charged *D* and *E*. For comparison, we include a set of bins based on hierarchical clustering of the Grantham distance matrix given in table 2 of Grantham (1974). This distance matrix is based on the chemical properties of the amino acids alone and not rates of exchange between them. We do see some similarities between the Grantham bins and the saturation bins (such as the groupings *ILV* and *AGPT*), but there are substantial differences as well.

Algorithms for Criterion Minimization

In principle, for a fixed number of bins, determination of the set of bins that maximizes the ratio W/W^{JC} is straightforward. More generally, the problem is to minimize a function of the bins, $F(b)$ (note that maximization is the same as minimization of—$F(b)$), for a fixed number of bins. Because there are a finite number of sets of bins, one can obtain $F(b)$ for each of these sets and choose the one that gave the smallest value. The difficulty is that the number of possible bins is very large. The numbers of partitions of n things into k groups are known as the Stirling numbers of the second kind and can be recursively calculated from

$$S(n,k)=kS(n-1,k)+S(n-1,k-1),\ 2\le \mathrm{k}\le \mathrm{n}-1,$$

where $S(n, 1) = S(n, n) = 1$ give starting conditions (cf. Abramowitz and Stegun 1972). Considering the number of partitions of 20 amino acids into 2 groups, the number is a feasible 524,287, but with 8 bins, there are roughly 1.5×10^{13} choices of bins.

We experimented with a number of different heuristic algorithms to minimize criterion functions, but all results reported here are for a bin rearrangement algorithm with 1000 random sets of starting bins for each choice of the number of bins. Given an initial set of bins, all possible exchanges of one amino acid from a bin to another are checked until an improved criterion function is found. The procedure was repeated with the resulting bins that improved the criterion function and continued until, for a current set of bins, all possible exchanges had been exhausted without an improvement.

Bias and Variance in the Absence of Misspecification

We first consider performance of methods using binned data in the absence of model misspecification and without much consideration for the choice of bins. Although it seems clear that some information will be lost constructing these bins, it is far from clear how much information will be lost. Figure 1 gives the mean and variances of the expected estimated distances between 2 taxa as functions of the true distance between them. These are based on 1000 simulated amino acid data sets of sequence length 250 simulated using Seq-Gen (Rambaut and Grassly 1997) under a JTT Markov model at each of the indicated true distances. The bins considered are the Dayhoff groups as well as the 10 saturation bins given in table 1. Plots using 6 saturation bins were very similar to those of the Dayhoff groups; although this choice of bins is similar to the Dayhoff groups, there are some differences. Considering the bias and variance plots, the differences in the performance for both of these bin choices and performance with 10 saturation bins suggests that the number of bins was the more

Table 1
The Estimated Saturation Bins and the Bins Obtained by Hierarchical Clustering Applied to the Grantham Distance Matrix

Saturation	Grantham
ADEGKNPQRST CFHILMVWY	ACDEFGHILMNPQRSTVWY K
ADEGNPST CHKQRW FILMVY	ACDFGMPQRSTW EHILNVY K
AGNPST CHWY DEKQR FILMV	AGPT CDFMQRSW EHILNVY K
AGPST CFWY DEN HKQR ILMV	AGPT CDQ EHILNVY FMRSW K
APST CW DEGN FHY ILMV KQR	AG CDQ EHILNVY FMRSW K PT
AGST CW DEN FY HP ILMV KQR	AG CDQ EHNY FMRSW ILV K PT
AST CG DEN FY HP ILV KQR MW	AG C DQ EHNY FMRSW ILV K PT
AST CW DE FY GN HQ ILV KR MP	AG C DQ EHNY FMW ILV K PT RS
AST CW DE FY GN HQ IV KR LM P	A C DQ EHNY FMW G ILV K PT RS
AST C DE FY GN HQ IV KR LM P W	A C DQ EHNY FM G ILV K PT RS W
AST C DE FY G HQ IV KR LM N P W	A C DQ EHNY FM G IL K PT RS V W
AST C DE FY G H IV KR LM N P Q W	A C DQ E FM G HNY IL K PT RS V W
AST C DE FL G H IV KR M N P Q W Y	A C D E FM G HNY IL K PT Q RS V W
AST C DE F G H IV KR L M N P Q W Y	A C D E FM G HNY IL K PT Q R S V W
AT C DE F G H IV KR L M N P Q S W Y	A C D E F G HNY IL K M PT Q R S V W
AT C DE F G H IV K L M N P Q R S W Y	A C D E F G HNY IL K M P Q R S T V W
A C DE F G H IV K L M N P Q R S T W Y	A C D E F G HNY I K L M P Q R S T V W
A C D E F G H IV K L M N P Q R S T W Y	A C D E F G HN I K L M P Q R S T V W Y

NOTE.—The saturation bins were constructed from the JTT rate matrix. The criterion used to choose bins is to maximize the ratio of the expected number of substitutions within bins to the expected number under the Jukes-Cantor model.

important factor in degradation of performance rather than the content of those bins.

The biases of the MD method are comparable to the bias in the unbinned JTT distances (fig. 1). All distances are ML distances. By contrast, the BMC bins give substantial apparent biases. This is expected as the interpretation of distances under the BMC method is the expected number of substitutions between bins where it remains unchanged under the MD method. What is important to note is that the mean BMC distances vary in a roughly linear fashion with the true distances, suggesting that such forms of binning will not give rise to topological inconsistency with larger sequence lengths. All variances increase as a function of true distance with a more rapid increase for the MD methods. To make the variances more comparable, we rescaled the distances so that the means were the same for each of the methods. Interestingly, the BMC variances are smaller than corresponding MD method variances, suggesting that the BMC distances might perform better in topological estimation.

In figure 2, each cell of a heat map indicates the proportion of times that the correct tree was estimated for a particular edge length setting on a 4-taxon tree. As expected, we see a degradation of performance as we move from estimation without binning to 10 to 6 bins. Nevertheless, performance is reasonable even with 6 bins. Interestingly, both the BMC and, in the case of no model misspecification, the more appropriate MD method give comparable performance for the Dayhoff groups. This suggests that the BMC method can be used with potential gains in the case of model misspecification and without too much loss in the case that models are not misspecified. The setting for the simulation results given in figure 2 is the widely used long-branch attraction simulation setting used in, for instance, Huelsenbeck (1995). The simulating tree for a cell in the heatmap is ((1:*b*, 2:*a*):*a*, (3:*b*, 4:*a*)). We follow the convention of Huelsenbeck (1995) by transforming *a* and *b* to corresponding probabilities of different amino acids for taxa separated by those evolutionary distances. The *x* value for a cell indicates the probabilities corresponding to *a*, and the *y* values indicate the probabilities corresponding to *b*. The general variation in proportions of missestimations in the heat maps is as expected being greatest when the middle edge length, *a*, is small, but terminal edges are long. Each cell was based on 1000 simulations, with sequences of length 250, from the JTT substitution model. For each simulated data set, ML distances were calculated for the binned model, and a neighbor-joining tree was constructed from this.

Criterion Minimization for Particular Alignments

The saturation bins in table 1 are once-and-for-all bins that are determined from a rate matrix. Because most amino acid models are empirically derived from large databases of alignments, the resulting bins reflect average properties over many alignments. In some cases, it may be preferable to obtain minimizers of criteria that are specific to a particular alignment of interest. For instance, violation of the

Table 2
The *P* Values for a Chi-Squared Compositional Heterogeneity Test for the Metazoan Data Sets Using Maximum Chi-Square Statistic Bins

Number of Bins	2	3	4	5	6	7	8	9	10	11	≥12
Mitochondrial	0.53	0.61	0.56	0.45	0.27	0.21	0.10	0.06	0.02	0.01	<0.001
Nuclear	0.74	0.70	0.66	0.61	0.53	0.37	0.24	0.17	0.09	0.02	0.000

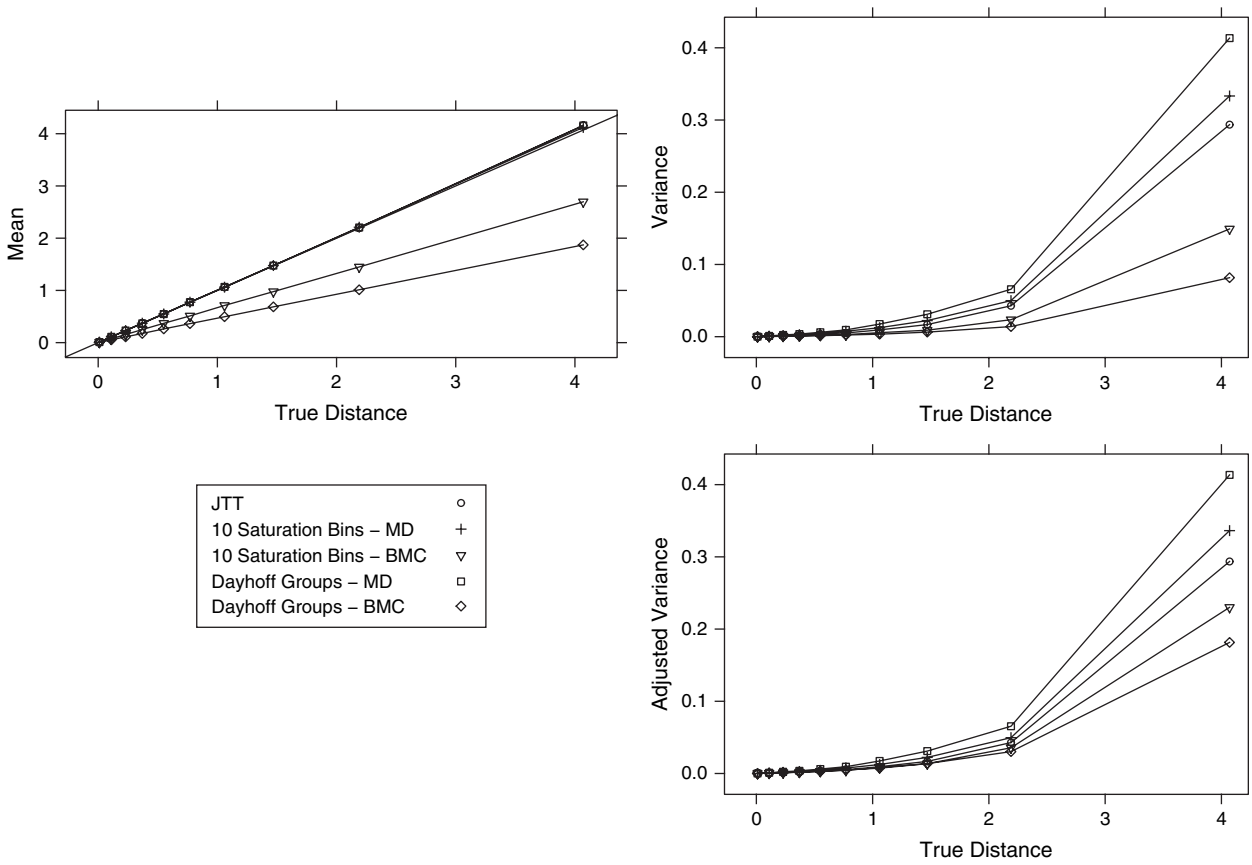

FIG. 1.—Plots of the mean and variance of the estimated distances under various binned models as functions of the true distances. The unbinned estimated distances are JTT, the Dayhoff groups are a set of 6 bins, and the saturation bins are given in table 1. Means and variances for binned distances were obtained under both a MD (no label) and BMC method. Included as well are variances after rescaling distances so that each method has the same mean as the JTT distances. All values are based on 1000 simulated amino acid data sets of sequence length 250 simulated under a JTT Markov model at each of the indicated true distances. All figures presented here were produced using the R statistical package (R Development Core Team 2007).

assumption of compositional homogeneity (Ho and Jermiin 2004) can create substantial problems for phylogenetic reconstruction. Alternatively one might be concerned with violation of time-reversibility assumptions.

The form of model misspecification that we focus upon here is compositional heterogeneity. The criterion function that we minimize is the maximum chi-squared statistic:

$$t_s = n \sum_{i,s} (\pi_{is} - \pi_i)^2 / \pi_i$$

where n is the sequence length, π_{is} is the frequency of the ith bin for the sth species, and π_i is the overall frequency of the ith bin. For a fixed number of bins, n_b, we will refer to the resulting set of bins as the minimax chi-squared bins. The P value for any one of these observed t_s test statistics would be calculated as $P(T > t_s)$ where T has a $\chi^2_{n_b-1}$ distribution. The same sequence that will give the maximum t_s will give the minimum P value, and so minimizing the maximum chi-squared statistic is equivalent to maximizing the corresponding minimum P value. Given that the concern is with compositional heterogeneity, one of the reasons for considering this criterion is that it is readily interpretable: If for a given choice of bins we end up with a P value that is larger than 0.1, compositional homogeneity cannot be rejected for this set of bins even if it could be for the original amino acid data.

One other unexpected advantage with this choice of criterion function is the stability of its calculation across the large number of different choices of bins encountered during the progression of the algorithm. There is some possibility that difficulties will arise when the frequency of one or more amino acids is 0, but this tends not to occur too frequently in practice. By contrast, an alternative choice that was considered was Bowker's test statistic as described in Ababneh et al. (2006). The null hypothesis for the latter test is compositional symmetry: a property implied by most time-reversible Markov models. Because compositional heterogeneity implies compositional asymmetry, the test can detect compositional heterogeneity as well. However, for a given pair of taxa, the test statistic requires that the sum of pairwise frequencies for each pair of bins be nonzero. With large numbers of bins, this is frequently not the case.

Another kind of data set–specific binning is what we will refer to as general time reversible (GTR) binning. The setting is similar to Waddell and Steel (1997) and Weiss and

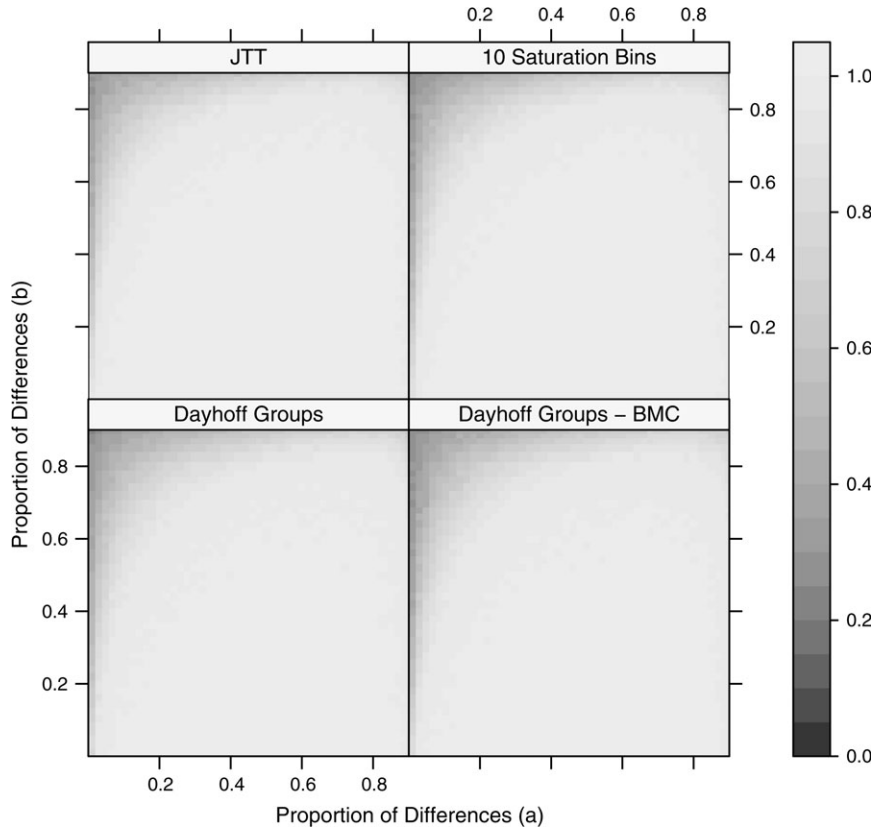

FIG. 2.—Each cell of a heat map indicates the proportion of times that the correct tree was estimated for the 4-taxon tree ((1:*b*, 2:*a*):*a*, (3:*b*, 4:*a*)). The *x* value for a cell indicates the probability of different amino acids for along an edge of length *a*, and the *y* values indicate the corresponding probability for an edge of length *b*. Each cell was based on 1000 simulations, with sequences of length 250, from the JTT substitution model. For each simulated data set, ML distances were calculated for the binned model and a neighbor-joining tree was constructed from this. Results for unbinned estimation are indicated as JTT. For the Dayhoff groups, results are reported for both the MD (no label) and BMC method. The saturation bins are given in table 1.

von Haeseler (2003). In brief, fixing a pair of taxa, an eigenvector decomposition of the form $\Pi^{-1}F = U\Omega U^{-1}$ is obtained, where Π is a diagonal matrix with entries equal to the frequencies of the amino acids, or bins in our case, and the *ij*th entry of F gives the frequency with which the pair of amino acids *i* and *j* arose. In our implementation, the F frequency matrix is symmetrized so that F_{ij} is the frequency with which *i* occurred in the first sequence and *j* in the second or *i* in the second and *j* in the first sequence. It follows from the symmetry of F that $\Pi^{-1}F$ has a real-valued eigenvector decomposition (cf. Keilson 1979; Waddell and Steel 1997). Under a gamma rates-across-sites model, an estimate of the *ii*th entry of the Λ in the eigenvalue decomposition of the matrix $Q = U\Lambda U^{-1}$ can be obtained through the transformation $\alpha - \alpha(\Omega_{ii})^{-1/\alpha}$; this is the inverse moment generating function for the gamma rate distribution. Because the decomposition of $\Pi^{-1}F$ may give negative values of Ω_{ii} it is possible that Λ cannot be computed. We ignored all choices of bins for which this occurred. A value of α is required, and we obtain this through a 2-step process. First a neighbor-joining tree is obtained using uncorrected ML distances (BMC method using bins). Then ML estimates of the edge lengths and α are obtained given the tree. Estimation of α in this way can be affected by poor topological estimation due to the use of uncorrected distances, but our anecdotal experience, using simulated amino acid data, is that reasonable estimates are obtained.

Given estimates $Q^{(kj)}$, of the rate matrices for all pairs of sequences (k, j), we use the test statistic of Weiss and von Haeseler (2003) as a criterion for minimization. This test statistic is the sum of 1 plus the eigenvalues of the sample covariance matrix, treating all pairs as the sample, of the off-diagonals of the $Q^{(kj)}$. The matrix provides a measure of the differences between the individual pairwise $Q^{(kj)}$ matrix estimates from the overall mean, and the test statistic will be large when there is evidence of heterogeneity in the Q matrices throughout the tree.

We refer to bins obtained through the maximization of the Weiss and von Haeseler (2003) test statistic as GTR binning because the underlying Q matrix estimates are estimates under the general time-reversible model. One of the reasons that this type of estimation is not frequently pursued with amino acid data is that sparseness issues can lead to difficulties with the eigenvalue decompositions. With smaller numbers of bins, these problems are not as likely and make a full GTR method feasible. In cases where GTR bins are used, this is indeed what we do. Given a set of bins, we obtain GTR distances for these bins: $d = -\text{trace}[\Pi Q]$. The estimated Q matrix for the pair is constructed as discussed above.

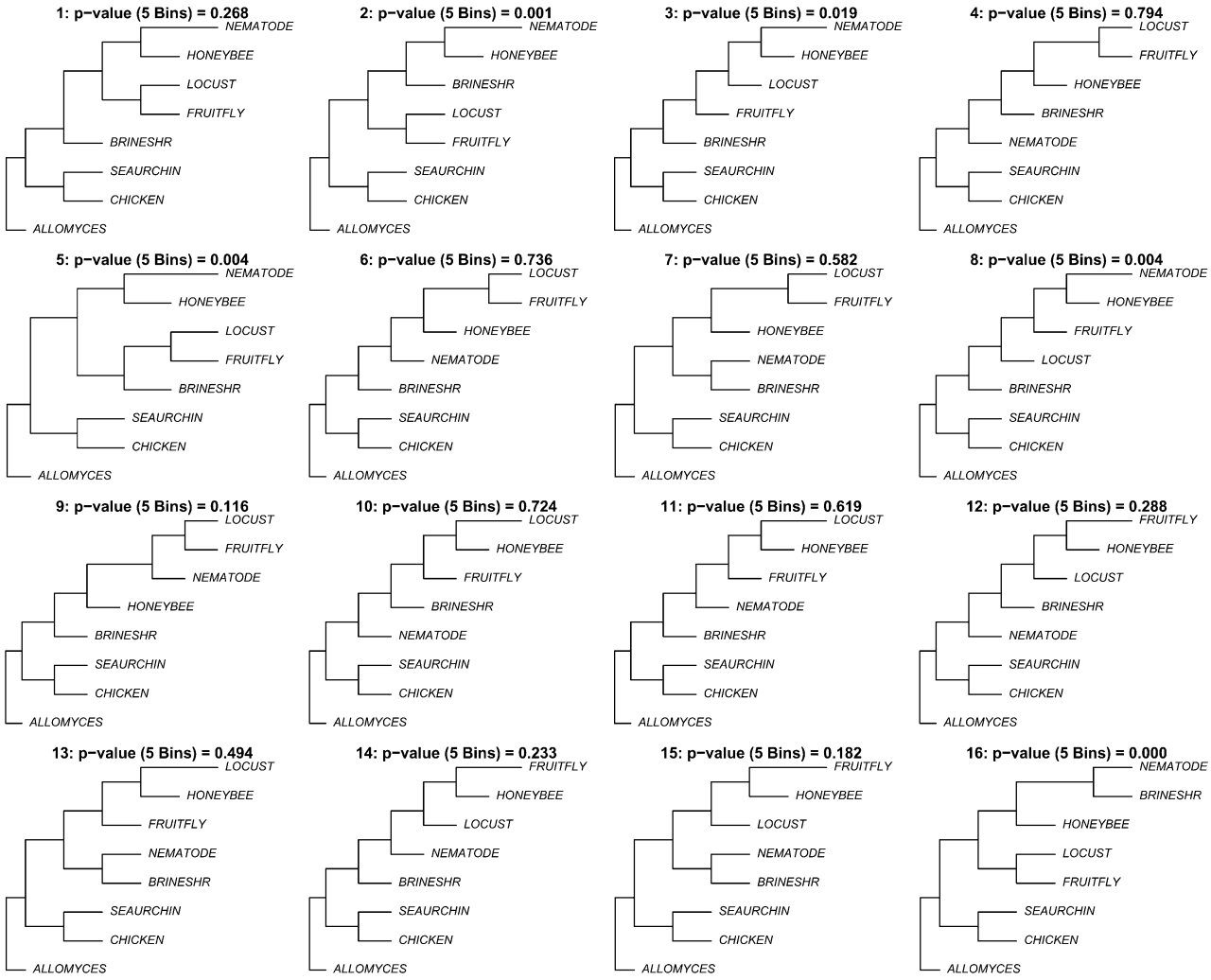

FIG. 3.—The top ranked ML topologies for the metazoan mitochondrial data using the original amino acid data are listed across rows. The JTT model with 8 gamma rate categories was used in fitting. *P* values for the SDNB test using amino acids were 0.29 and 0.03 for the top 2 ranked trees, with the next 5 trees giving *P* values of 0.01. *P* values for the SDNB test when 5 maximum chi-squared bins are also given; any tree with a *P* value larger than 0.05 would be included in a 95% confidence region.

Metazoan Mitochondrial Data

The first example data set that we consider is the metazoan mitochondrial data considered previously in Foster et al. (1998). This amino acid data set is fairly large with 8 taxa and 3713 sites. The topologies for the top 16 ML trees are given in figure 3; the fitted model was the JTT model with 8 gamma rate categories. The top 3 trees and 5 out of the top 10 trees have the honeybee and nematode sequences grouped together, a clearly incorrect split. This is consistent with the observations of Foster et al. (1998) that these 2 sequences share significantly elevated ratio of the amino acids *FYMINK* to *GARP* due to their heightened $A + T$ nucleotide content relative to other sequences in this data set. Indeed, the Ecdysozoa hypothesis tree, ranked 12th by ML, is widely believed to be the correct tree for this data (cf. Turbeville et al. 1997; Philippe et al. 2005).

To assess whether the unusual groupings observed may be due to uncertainty in the data, we used the single distribution nonparametric bootstrap (SDNB) test of Shi et al. (2005). This test gives a *P* value as the proportion of bootstrap samples where the log likelihood ratio $l(\hat{\tau}^*) - l(\hat{\tau})$ was greater than the observed log likelihood ratio, $l(\hat{\tau}) - l(\tau_0)$, for a hypothesized topology τ_0; here $\hat{\tau}$ is the ML topology and $\hat{\tau}^*$ represents the ML topology for a bootstrap sample. Using 100 resampling of estimated log-likelihoods resampled data sets (Kishino et al. 1990), only a single tree was contained in a 95% confidence region for trees. The *P* values from the SDNB test ranked from highest to lowest were 0.29, 0.03, 5 trees had *P* values of 0.01, and the rest were less than 0.01. The observed groupings of honeybee and nematode are not a consequence excess noise in the data set.

The results of chi-square tests confirm that compositional heterogeneity is a serious problem in this data set. Only the fruit fly passed the compositional chi-squared test with a *P* value of 0.29. Seven of the 8 taxa gave *P* values that were 0 up to round-off error. In light of these difficulties, the suggested compositional homogeneity of the fruit fly is not meaningful as the overall frequencies are a mix of very heterogeneous frequencies.

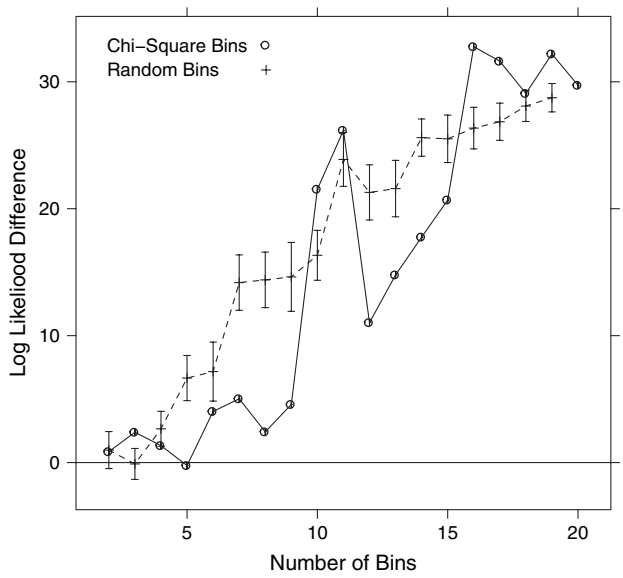

FIG. 4.—The difference in log likelihoods between the ML topology (topology 1 in Figure 3) and the Ecdysozoa hypothesis tree (topology 12) for the bins giving the maximum chi-squared statistic. For comparison, the plot includes the same log likelihood differences averaged over 100 choices of random bins of the same size. Approximately 95% confidence bounds are indicated for the mean log likelihood difference.

The minimum *P* value for any taxa and for each choice of bins is given in table 2. Binning does help in dealing with the compositional heterogeneity difficulties, giving large *P* values for 5 or less bins and insignificant test results for 9 or less bins.

A plot of the log likelihood difference between the ML topology (topology 1 in fig. 3) and the Ecdysozoa hypothesis tree (topology 12) is given in figure 4. For comparison, the plot includes the same log likelihood differences averaged over 100 choices of random bins of the same size. The increasing trend in log likelihoods for the random bins is due in part to the loss of information due to binning but may be due as well to an amelioration of compositional heterogeneity effects for some of the random choices. Nevertheless, because the random bins were not selected with composition corrections in mind, numbers of bins where the maximal chi-squared bins give smaller differences in likelihood are of particular interest. We see that the gap is particularly pronounced for less than or equal to 9 bins—the numbers of bins that gave insignificant chi-squared homogeneity test results—and that with 5 bins, the likelihood for the Ecdysozoa hypothesis tree is actually higher than for the ML topology.

Considering the choice of 5 bins further, we obtained the likelihood values (optimizing edge lengths and an α parameter for 8 gamma rate categories) for the top 100 amino acid ML topologies. The top 8 topologies were topologies 4, 6, 10, 11, 7, 13, 12, and 1 in figure 3. With the exception of topology 1, all these topologies have broken up the honey bee, nematode split, and correctly have honey bee, locust, and fruit fly split from the rest. The 95% confidence region of trees from the SDNB test included these 8 trees plus the trees with amino acid ML rankings: 14, 15, 9, 54, 38, 41, and 58, with the *P* values for the last 5 topologies being between 0.10 and 0.05. Among the topologies in the confidence region, only topology 1 had honey bee and nematode split from the rest. The 5 bins for this data were *AFMP*, *CKQR*, *DE*, *GHN,* and *ILSTVWY*. There is not a lot of overlap with the Dayhoff groups or saturation bins. Nevertheless, some of the groupings, *DE* and *VIL*, for instance, are consistent with what one would expect based on chemical properties of amino acids.

Metazoan Nuclear Data

We consider next a phylogenomic data set with 50,462 sites and 10 taxa considered by Dopazo H and Dopazo J (2005). Of primary interest for this data is whether one of 2 trees, the Coelomata tree (fig. 5*A*) or the Ecdysozoa tree (fig. 5*B*), is the correct tree; these are figure 3*a* and *b* (Dopazo H and Dopazo J 2005). The 2 trees differ in their placement of *Caenorhabditis elegans*. Different data sets and methods have yielded either the Coelomata or Ecdysozoa trees, although currently most evidence suggests the latter is correct (Philippe et al. 2005; Dopazo H and Dopazo J 2005). In Dopazo J and Dopazo J (2005), for instance, analysis of the full data set gives the Coelomata tree, whereas removal of fast evolving sites gives the Ecdysozoa tree. This is a natural data set for the saturation bins which are intended to deal with difficulties due to fast evolving sites without throwing away all the information coming from these sites.

For the 2 hypothesized trees, likelihoods were obtained with ML edge lengths for both the saturation bins and the maximum chi-squared bins. In each case, 8 gamma rate categories were used. For the saturation bins, the Ecdysozoa tree gave a larger likelihood only when the number of bins were 3. For the maximum chi-squared bins, the Ecdysozoa tree gave a larger likelihood only when there were 9 bins. Notably, the minimum chi-squared statistics gave rejections of compositional homogeneity with more than 9 bins for this data (table 2). The Coelomata tree was the estimated tree more frequently than not for both binning methods. In addition, the neighbor-joining tree obtained using GTR distances and GTR bins was the Coelomata tree in each case. Still the support was frequently not strong. The *P* values from the SDNB tests comparing the 2 trees are given in table 3 and are often larger than 0.05.

Chloroplast Data

Years of controversy have surrounded the identity of the basal-most node in the angiosperm phylogeny. For instance, the placement of *Amborella* within the radiation of angiosperms has evoked a debate about the basal node in angiosperm phylogeny (Goremykin et al. 2003, 2004; Soltis et al. 2004; Lockhart and Penny 2005; Martin et al. 2005). Using the chloroplast genome data, Leebens-Mack et al. (2005) found weak support for *Amborella* and water lilies (*Nymphaea* and *Nuphar* here) at the base of the angiosperms.

A binned analysis of 61 concatenated chloroplast-encoded proteins, yielding 15,688 sites for 24 taxa, gave interesting results in this regard. The neighbor-joining trees for the ML distances using bins that gave the minimum chi-squared statistic among 12 and 13 bins are given in figure 6.

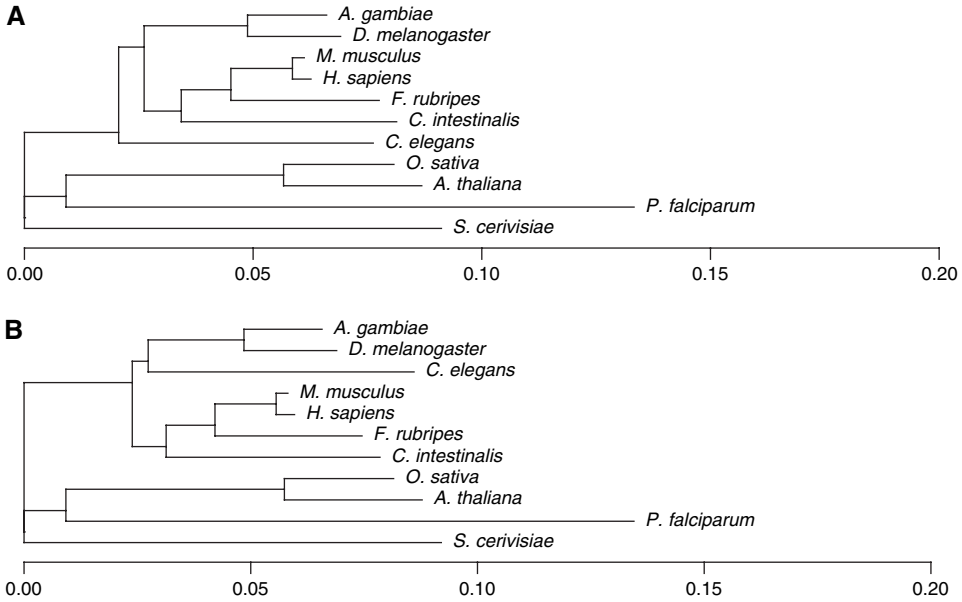

FIG. 5.—The Coelomata (*A*) and Ecdysozoa (*B*) hypothesis trees for the metazoan nuclear data. Edge lengths were obtained from the full amino acid data fitted with a JTT+Γ model having 8 Γ rate categories. For the GTR bins, neighbor joining with GTR distances was used to reconstruct the tree.

The differences in the 2 trees involve the placement of the *Nymphaea*, *Nuphar*, *Amborella* clade, hereafter referred to as the *NNA* clade, as well as the placement of *Calycanthus*. With 12 bins, *Calycanthus* groups with the monocots, and the NNA clade is basal. This would usually be considered the correct relationship, although there has been considerable debate. With 13 bins, *Calycanthus* groups with *NNA*, and these groups split from the monocots after an initial split giving rise to the eudicots.

The transition from 12 to 13 bins represents a sort of phase transition. With 3–12 bins, the neighbor-joining trees obtained are exactly the same and with 13–20 bins the neighbor-joining trees are the same. What is striking is the bootstrap support for the features that differ from 12 to 13 bins. With 12 bins, the bootstrap support for a basal *NNA* clade is 84% and similarly high for less than 12 bins. With 13 bins, the shallower *NNA* + *Calycanthus* split has 59% bootstrap support, and this increases to 82% with the original amino acid data (20 bins).

Surprisingly, there is overlap in the bins. The bins giving the minimum chi-squared statistic with 12 bins were *AIV*, *CW*, *EF*, *GN*, *KR*, and *LQS* with the rest of the bins consisting of single amino acids. The 13 minimax chi-squared bins that differ from the 12 minimax bins are *AL*, *CQV*, and *IS*. In addition, *W* has split from *C* to give its own new bin.

Part of the explanation for the differences between the trees may have to do with a joint composition and saturation correction when there are 12 bins. Considering the JTT rate matrix, the rate of substitution from *I* to any other amino acid, for instance, is highest when that amino acid is *V* and the highest rate from *V* is to *I*. Generally, when the number of bins is 12, the bins that differ from those when the number of bins is 13 involve amino acids that have relatively high substitution rates to and from each other. Homoplastic substitutions of amino acids are less likely to be substitutions at the binned level for 12 bins.

Some confirmation that the differing rates of substitution between bins were important is obtained by considering the sites in the alignment where there were 2 and only 2 amino acids. Among these sites, there were 577 that involved amino acids that were within bins when the number of bins was 12 but in different bins when the number of bins was 13; for instance, a site with all *A* and *V*. Out of these sites, 205 require as an evolutionary explanation multiple substitutions between the 2 amino acids. In contrast, there were only 16 such sites with amino acids in the same bin when the number of bins was 13 but different bins when the

Table 3
The *P* Values for the Metazoan Nuclear Data, Using the SDNB Test, when the Null Hypothesis Is that the Ecdysozoa Tree Is the True Tree and the Coelomata Tree Is the Alternative Tree

					Saturation Bins					
Number	2	3	4	5	6	7	8	9	10	11
P value	0.01	0.51	0.34	0.08	0.14	0.08	0.09	0.21	0.05	0.06
Number	12	13	14	15	16	17	18–20			
P value	0.04	0.03	0.20	0.10	0.02	0.01	<0.01			
					Maximum Chi-Square Bins					
Number	2	3	4	5	6	7	8	9	≥10	
P value	0.41	0.33	0.00	0.00	0.22	0.04	0.33	0.42	0.01	

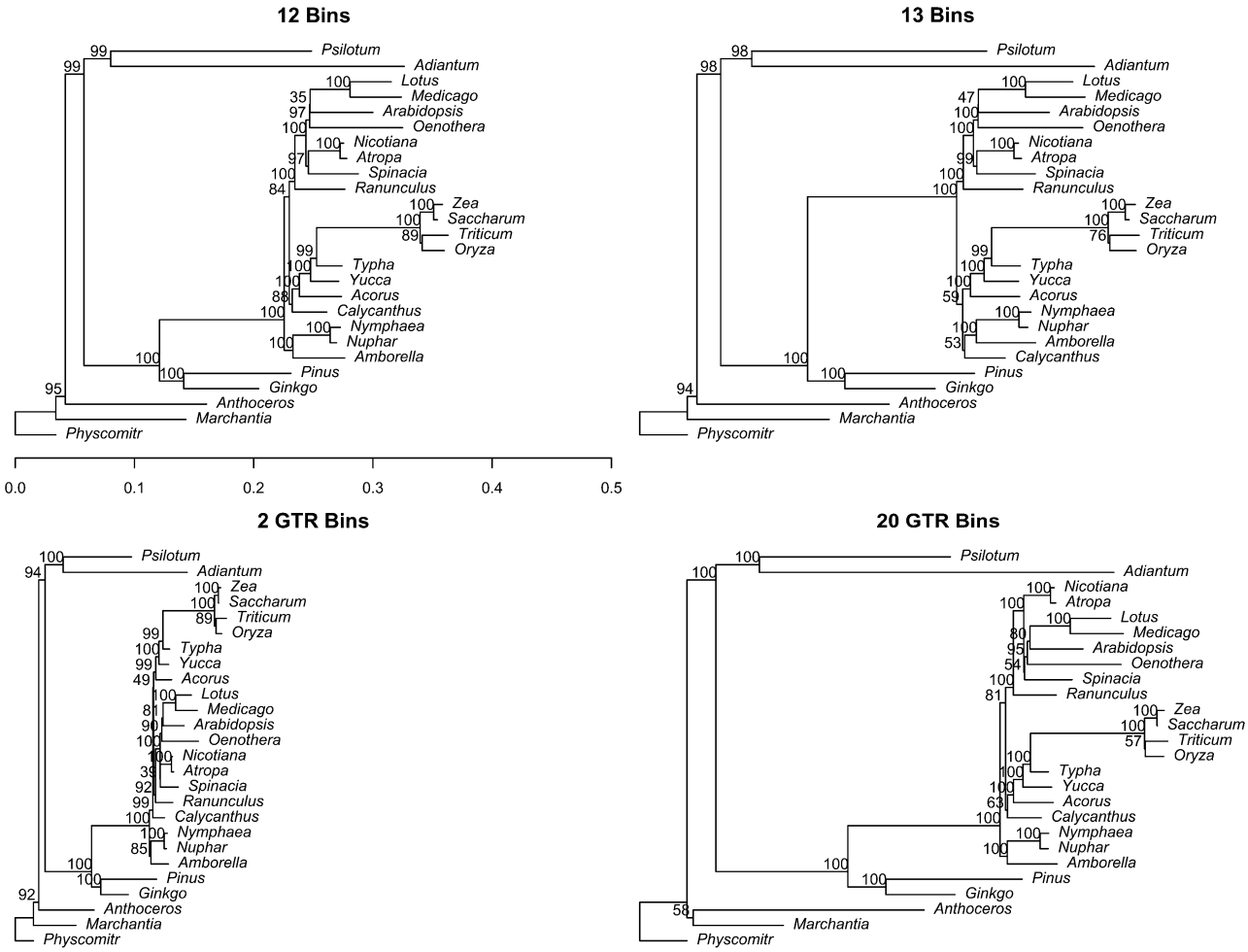

FIG. 6.—Trees with bootstrap support for the chloroplast data estimated by applying the neighbor-joining algorithm. The top panels give trees from ML distances from 12 and 13 minimax chi-squared bins. The bottom panels give trees from GTR distances with 2 GTR bins and with the original amino acid data.

number of bins was 12. Only one of these sites requires multiple substitutions between the amino acids.

It does appear, however, that it was important that both saturation and composition are adjusted for with 12 minimax chi-squared bins. The same topology as with 13 bins was obtained for every choice of the number of bins when the saturation bins were used.

The neighbor-joining trees obtained using GTR distances and corresponding GTR bins, with gamma rates-across-sites adjustment, consistently placed the *NNA* clade at the base of the tree. The only difference in the estimated tree across choices of bins was the placement of *Calycanthus* which was more basal with 2 GTR bins. With 3 or more bins, it grouped with the eudicots with bootstrap support varying between 40 and 70%.

Discussion

With small numbers of bins, the saturation bins ended up being similar to other choices of bins like the Dayhoff classes; the 5 saturation bins differ from the Dayhoff classes only in that *C* is not split from *W* and *Q* is with *H K R* rather than *D E N*. This is not very surprising as the reasons for the choices given in Hrdy et al. (2004) are similar to the motivations for the saturation bins. Kosiol et al. (2003) similarly desire choices of bins where rates of substitution are high within bins but low between bins. However, their bins differ substantially from the saturation bins here. Differences may be due to their derivation from the Helan and Goldman (WAG) and point accepted mutation (PAM) rate matrices rather than JTT.

To some degree, the saturation bins also match up with groupings based on the chemical similarities between the groups; that is, bins based on hierarchical clustering of the Grantham chemical property distance matrix which is constructed entirely from chemical properties. The bins match up to some degree but show a fair number of differences. Furthermore, the saturation bins also reflect groupings based on chemical properties described elsewhere (cf. fig. 4.3 in Orengo et al. [2003] and Sections 2.4–2.6 of Higgs et al. [2005]). For instance, the small hydroxylated amino acids *S* and *T* are frequently together, and *A*, *P* and *G* are often part of this group. The positively charged amino

acids *K* and *R* always group together; *H* on the other hand tends not to group with these. The aromatic amino acids *F* and *Y* tend to appear together but not with *W* and *H*. Another notable grouping is the negatively charged amino acids *D* and *E*. Because the saturation bins are directly a consequence of empirical observations of the frequencies of exchange of amino acids, via the empirically derived JTT rate matrix, it is clear, as anticipated, that these rates of exchange are interlinked with chemical properties of the amino acids, but are not solely determined by them.

Data set–based criteria for bins also showed some similarities with saturation bins. For example, for the chloroplast data, groups of amino acids like *IV*, *KR*, and *CW* that occur in some saturation bins were in the 12-bin maximum chi-squared solution that, excluding single amino acids, consisted of the bins *AIV*, *CW*, *EF*, *GN*, *KR*, and *LQS*. There are differences as well, however. For instance, the amino acid *A* is never grouped with *IV* in the saturation bins. As the chloroplast example illustrated there may be an interaction between saturation and compositional heterogeneity that the alignment-based criteria are dealing with. It seems plausible that, more generally, the root causes of saturation and other problematic issues for phylogenetic inference will be related to the causes of compositional heterogeneity, so that bins constructed with compositional homogeneity as a target may correct other problems as well.

The *P* values for the compositional heterogeneity tests in table 2 are fairly large. When the number of bins is small, one can overcorrect for compositional heterogeneity; during the course of optimizations, a number of different bin choices gave maximum chi-squared statistics consistent with the chi-squared variation expected in the absence of compositional difficulties. For smaller numbers of bins, it thus seems clear that more model misspecifications could potentially be fixed. The GTR bins considered here provide an example where heterogeneity of entire rate matrices rather than heterogeneity of composition is the target of correction. Because of the same types of sparseness issues that arise with GTR distances in amino acids, this type of method is primarily applicable when the number of bins is small enough that matrix logarithms can be expected to exist.

A somewhat surprising finding was that, in the absence of model misspecification, there was not too much loss of information in binning amino acids and, moreover, that using a (misspecified) binned Markov model as opposed to a MD method did not lead to substantial difficulties. It appears that the model misspecification induced by binning leads to increased variation but not to substantial biases. The situation is likely more complex with real data where model misspecification will usually be present both when the data is represented in terms of bins and when it is represented as amino acids.

The use of reduced amino acid alphabets will provide a useful diagnostic tool in this era of large concatenated data sets where information content is so high that concern rests more with biases due to model misspecification rather than with excess variance due to overly elaborate modeling. Differing, well-supported estimated trees for different choices of bins provide evidence of difficulties and food for additional thought and investigation.

Funding

This research was supported by Discovery grants awarded to E.S. and A.J.R. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. and E.S. are fellows of the Canadian Institute for Advanced Research Program in Evolutionary Biology. A.J.R. is supported by a fellowship from the Peter Lougheed New Investigator Award from the Canadian Institutes of Health Research and the E.W.R. Steacie fellowship from NSERC.

Acknowledgments

We would like to thank Peter Foster for supplying the metazoan mitochondrial data set and James Leebens-Mack for the chloroplast data.

Literature Cited

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics. 22:1225–1231.

Abramowitz M, Stegun IA. 1972. Handbook of mathematical functions with formulas, graphs, and mathematical tables. United States Department of Commerce. Washington (DC): U.S. Government Printing Office.

Cannata N, Toppo S, Romualdi C, Valle G. 2002. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. Bioinformatics. 18: 1102–1108.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6:R41.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Vol. 5, supplement 3. Silver Spring (MD): National Biomedical Research Foundation. p. 345–352.

Embley TM, van der Giezen M, Horner DS, Dyal PL, Foster P. 2003. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. Philos Trans R Soc Lond B Biol Sci. 358:191–203.

Foster P, Hickey DA. 1998. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol. 48:284–290.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that Amborella is not a basal angiosperm. Mol Biol Evol. 20:1499–1505.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 21:1445–1454.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185:862–864.

Higgs PG, Attwood TK. 2005. Bioinformatics and molecular evolution. Oxford: Blackwell Publishing.

Ho SYW, Jermiin LS. 2004. Tracing the decay of the historical signal in biological sequence data. Syst Biol. 53:623–637.

Hrdy I, Hirt RP, Dolezal P, Bardonova L, Foster PG, Tachezy J, Embley TM. 2004. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature. 432:618–622.

Huelsenbeck J. 1995. Performance of phylogenetic methods in simulation. Syst Biol. 44:17–48.

Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol. 53:638–643.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS. 8:275–282.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.
Keilson J. 1979. Markov chain models—rarity and exponentiality. New York: Springer-Verlag.
Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol. 30:151–160.
Kosiol C, Goldman N, Buttimore NH. 2003. A new criterion and method for amino acid classification. J Theor Biol. 228:97–106.
Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphlis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol. 22:1948–1963.
Lockhart P, Penny D. 2005. The place of Amborella within the radiation of angiosperms. Trends Plant Sci. 10:201–202.
Martin W, Deusch O, Stawski N, Grünheit N, Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:1360–1385.
Miyazawa S, Jernigan RL. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term. J Mol Biol. 256:623–644.
Orengo C, Jones D, Thornton J. 2003. Bioinformatics: genes proteins and computers. Oxford: Bios Scientific Publishers.
Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa. Protostomia. Mol Biol Evol. 22:1246–1253.
Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol. 28:171–185.
R Development Core Team. 2007. R:a language and environment for statistical computing [Internet].. Vienna (Austria): R Foundation for Statistical Computing;.[cited 2005 May 30]. Available from: http://www.R-project.org.
Rambaut A, Grassly NC. 1997. Seq-gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 13:235–38.
Shi X, Gu H, Susko E, Field C. 2005. The comparison of confidence regions in phylogeny. Mol Biol Evol. 22: 2285–2296.
Soltis D, Albert V, Savolainen V, et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and ending incongruence: a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.
Susko E, Spencer M, Roger AJ. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. J Mol Evol. 61:351–359.
Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387:489–493.
Waddell PJ, Steel MA. 1997. General time-reversible distances with unequal rates across sites: mixing Γ and inverse gaussian distributions with invariant sites. Mol Phylogenet Evol. 8: 398–414.
Wang J, Wang W. 1999. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol. 6: 1033–1038.
Weiss G, von Haeseler A. 2003. Testing substitution models within a phylogenetic tree. Mol Biol. Evol. 20:572–578.
Wenzel JW, Siddall ME. 1999. Noise. Cladistics. 15:51–64.
Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 15:1600–1611.

Martin Embley, Associate Editor

Accepted July 12, 2007