Phylogenetic Inference from Conserved Sites Alignments

William Noble Grundy*
Department of Computer Science
University of California, Santa Cruz
Santa Cruz, CA 95064
(831) 459-2078 Fax: 459-4829
bgrundy@cse.ucsc.edu

Gavin J. P. Naylor
Department of Zoology and Genetics, Iowa State University and
Iowa Computational Biology Lab

Includes 8 figures and 3 tables.

Running head: Phylogenetic Inference from Conserved Alignments

Abstract

Molecular sequences provide a rich source of data for inferring the phylogenetic relationships among species. However, recent work indicates that even an accurate multiple alignment of a large sequence set may yield an incorrect phylogeny, and that the quality of the phylogenetic tree improves when the input consists only of the highly-conserved, motif regions of the alignment. This work introduces two methods of producing multiple alignments that include only the conserved regions of the initial alignment. The first method retains conserved motifs, whereas the second retains individual conserved sites in the initial alignment. Using parsimony analysis on a mitochondrial data set containing nineteen species among which the phylogenetic relationships are widely accepted, both conserved alignment methods produce better phylogenetic trees than the complete alignment. Unlike any of the nineteen inference methods used previously to analyze this data, both methods produce trees that are completely consistent with the known phylogeny. The motif-based method, on the other hand, employs far fewer alignment sites for comparable error rates. For a larger data set containing mitochondrial sequences from 39 species, the site-based method produces a phylogenetic tree that is largely consistent with known phylogenetic relationships and which suggests several novel placements.

1 Introduction

The Human Genome Project and similar work on other species are producing molecular sequence data at an accelerating rate. In addition to providing an increased understanding of the fundamental mechanisms of biology, this trove of sequence data offers a picture of the evolutionary past of genes and of the species that carry them. A phylogenetic tree outlining the evolutionary history of a set of species can be derived from a set of DNA or protein sequences taken from those species.

*Corresponding author

Because a phylogenetic tree represents an evolutionary history that is not directly observable, such trees must necessarily be constructed by inference. Many phylogenetic inference algorithms are available, most of which are based upon maximizing some measure of the goodness of candidate phylogenetic trees. For any of these inference algorithms, the reliability of the inferred tree can be evaluated via statistical means, using, for example, statistical bootstrapping (Felsenstein 1985) or decay indices (Bremer 1988; Donoghue *et al.* 1992). Such analyses can reveal the extent to which the historical signal is differentiable from non-hierarchical signals in the data (Swofford *et al.* 1996). However, if the data contain a signal that differs from the historical signal but which nonetheless orders the data in a hierarchical way, then this erroneous signal may be supported by tests of statistical significance. This is especially the case for distantly related organisms that have had sufficient time to accrue signals due to the influence of subtle non-historical forces acting on the genome.

The possibility that a multiple sequence alignment may contain a misleading signal in addition to the historical signal implies that even extremely well supported phylogenetic trees may be incorrect. Naylor and Brown (1998) demonstrate this phenomenon using a large set of widely divergent mitochondrial sequences derived from organisms whose phylogenetic relationships are uncontroversial. They infer phylogenetic trees using a large battery of phylogenetic inference techniques, including three equally weighted parsimony analyses (of nucleotides, transversions only, and amino acids) and sixteen distance analyses (using Jukes-Cantor (1969), Kimura two-parameter (1980), Hasegawa-Kishino-Yano (1985) and general time-reversible (Lanave *et al.* 1984; Tavaré 1986; Rodríguez *et al.* 1990) distances in conjunction with four different models of among-site rate variation). Each of these nineteen analyses yields statistically significant phylogenetic trees. None of the methods yields the true tree.

Analysis of the source of the misleading signal in the mitochondrial data indicates that the strongest historical signal resides in sites that are associated with conserved molecular motifs. Accordingly, we describe here a method for producing multiple sequence alignments that consist only of the conserved motif regions. The MEME (Bailey and Elkan 1994) and Meta-MEME (Grundy *et al.* 1997) motif-based modeling toolkit uses expectation-maximization to discover motif regions in sequence data, and hidden Markov models (HMMs) to produce a motif-only multiple alignment. Alignments produced in this way, when subjected to parsimony analysis, yield phylogenetic trees that are closer to the true phylogeny than a similarly produced tree from the complete multiple sequence alignment.

Each site included in a conserved motif alignment is characterized by high conservation among the input sequences and proximity to a cluster of other highly-conserved sites. To determine whether the clustering of sites is a significant property of the historical signal, we also investigate a non-motif-based method of producing conserved multiple alignments. This method involves discarding alignment sites for which the relative entropy falls below a given threshold. The resulting high relative entropy alignment yields better phylogenetic trees than the complete alignment and yields the widely accepted true tree for a range of relative entropy thresholds.

Both of these conserved alignment methods involve a free parameter that specifies the degree to which the initial alignment is constrained. These parameters are the number n of motifs to include in the alignment, and the relative entropy threshold t for sites included in the alignment. For the mitochondrial data set described above, the optimal settings for these parameters can be determined experimentally by comparing various inferred phylogenetic trees with the trusted tree. In general, however, the proper values of n and t will not be known *a priori*.

The correct values of n and t can be estimated if the phylogenetic relationships among a subset of the given species is known. We apply this method to a larger data set containing the mitochondrial coding sequences from 39 species, including 34 vertebrates and a collective outgroup comprising a cephalochordate and four echinoderms. A motif-only alignment is created by directly applying to the larger data set the hidden Markov model that provided the most accurate tree for the original, nineteen-species data set. The quality of the resulting tree is evaluated with respect to an incompletely resolved tree representing all known phylogenetic relationships among the 39 species. For the site-based conserved alignment, this incompletely resolved tree provides a benchmark against which to compare inferred trees for various settings of the constraint parameter t. The result is a phylogeny that is consistent with known phylogenetic relationships and which suggests several novel placements.

Species name	Common name	Amino acids
Mus musculus	Mouse	3785
Rattus norvegicus	Rat	3794
Bos taurus	Cow	3791
Balaenopterus physalus	Fin-back whale	3790
Balaenopterus musculus	Blue whale	3790
Didelphis virginiana	Opposum	3727
Gallus gallus	Chicken	3785
Xenopus laevis	Frog	3782
Cyprinus carpio	Carp	3793
Oncorhynchus mykiss	Trout	3802
Petromyzon marinus	Lamprey	3803
Branchiostoma floridae	Lancelet	3740
Paracentrotus lividus	Common urchin	3824
Strongylocentrotus purpuratus	Purple urchin	3825
Drosophila yakuba	Fruit fly	3829
Cepaea nemoralis	Snail	3533
Anopheles gambiae	Mosquito	3733
Ascaris suum	Nematode 1	3454
Caenorhabditis elegans	Nematode 2	3421

Table 1: **Species included in the metazoan data set.** The last five species in the table constitute a collective outgroup. The entire alignment contains 4078 sites and is available at http://www.cse.ucsc.edu/research/compbio/phylo.

2 Methods

Evaluations of the two conserved alignment methods are performed using amino acid sequences from the mitochondria of nineteen metazoan taxa, including several vertebrate classes, two echinoderms and a collective outgroup of five species (Naylor and Brown 1998). The data set contains thirteen proteins from each species. The sequences range in length from 3421 to 3829 amino acids, with a total of 71 001 amino acids in the entire data set. A list of the species included in the data set is provided in Table 1, and the complete data set is available on the web at http://www.cse.ucsc.edu/research/compbio/phylo.

Motifs are discovered in the unaligned sequences using MEME (Multiple Elicitation of Motifs by Expectation-maximization) (Bailey and Elkan 1994). To reduce potential bias in the data set, sequences are weighted using a binary weighting scheme. The `purge` program (Lawrence *et al.* 1993) uses the BLAST algorithm (Altschul *et al.* 1990) to remove highly similar sequences from a given set of sequences. For this analysis, a bit score threshold of 1500 is used. For the purposes of evaluating the consistency of the overall method, this process is repeated ten times with different random seeds, yielding ten divergent training sets containing three or four sequences each. MEME analyzes each training set using the default parameter settings from the web interface (Grundy *et al.* 1996). The defaults include empirical Dirichlet mixture priors (Brown *et al.* 1995; Sjolander *et al.* 1996) weighted according to the megaprior heuristic (Bailey and Gribskov 1996), a minimum motif width of 12 and a maximum of 55, and a motif model biased toward zero or one motif occurrence per sequence. No attempt was made to optimize these parameters for the specific data sets under consideration.

MEME employs a modified likelihood ratio test to compute the relative significance of the motif models it discovers. However, a computationally feasible means of determining the absolute statistical significance of a MEME motif model is not known. Therefore, in the absence of a theoretical significance threshold, sensitivity analysis is performed on n, the total number of motifs from which the multiple sequence alignment is inferred. MEME discovers a total of 100 motifs in the weighted training set. Motifs that appear in less than 75% of the training set sequences are discarded. This procedure results in an average of 73 motifs from each of the ten training sets.

For each value of n, up to the total number of motifs discovered, the n most significant motifs are

```
TLFLIPMNLFSIVFALSWIAFIYPTNWAPSRFQSIWASFR
TFFLIPMNVFSMAFCLSWLVFIYPVNWAPSRFQSIWLGFR
TFGMLPLAWLAMLAPSLMLVSQTPVKFIKSRYHTLLTPIL
```

$$\begin{aligned}
I_1 &= f_{T1} * log_2(f_{T1}/b_T) \\
&= 1.00 * log_2(1.00/0.06) \\
&= 4.06
\end{aligned}$$

$$\begin{aligned}
I_2 &= [f_{F2} * log_2(f_{F2}/b_F)] + [f_{L2} * log_2(f_{L2}/b_L)] \\
&= [0.67 * log_2(0.67/0.07)] + [0.33 * log_2(0.33/0.15)] \\
&= 2.56
\end{aligned}$$

$$\begin{aligned}
I_3 &= [f_{F3} * log_2(f_{F3}/b_F)] + [f_{G3} * log_2(f_{G3}/b_G)] \\
&= [0.67 * log_2(0.67/0.07)] + [0.33 * log_2(0.33/0.05)] \\
&= 3.08
\end{aligned}$$

Figure 1: **Calculating relative entropy.** The figure illustrates how to calculate the relative entropy I for positions within a multiple alignment containing three sequences. The calculations shown correspond to the three sites within the box. f_{ij} is the frequency of amino acid i at position j in the alignment. The alignment shown is a fragment of a larger multiple alignment from which the background frequencies b_i are drawn. Position 1 is completely conserved and therefore has the highest relative entropy of the three. Position 2 has a slightly higher relative entropy than position 3 because glycine (G) in position 3 is less common than leucine (L) in column 2.

combined into a linear hidden Markov model, which is then used to align the motif regions of the entire data set. HMMs have been introduced relatively recently to the field of computational biology but have gained widespread acceptance as an effective means of modeling proteins (Krogh *et al.* 1994; Eddy 1995; Baldi *et al.* 1994; Eddy 1998). A hidden Markov model may be used either to detect homologs of the modeled sequences or to build a multiple sequence alignment of a set of known homologs.

In this work, HMMs are built using MAST (Bailey and Gribskov 1998) and the Meta-MEME toolkit (Grundy *et al.* 1997). MAST finds the canonical order and spacing in the training set sequences of the given MEME motif models. Meta-MEME then uses this order and spacing information to combine the motif models into a single HMM of the entire sequence. In a Meta-MEME HMM, the spacer regions between motifs are modeled only imprecisely, using a single parameter to approximate the observed spacer length. Therefore, Meta-MEME produces multiple alignments only of the motif regions; the non-motif regions are discarded from the alignment. The resulting motif-only alignment serves as the conserved alignment from which a phylogenetic tree may be inferred.

In addition to motif-only alignments, conserved alignments are constructed that include only sites with high relative entropy. The relative entropy, or information content, of an alignment site containing amino acids with frequencies $f_1, f_2, \ldots, f_{20}$ is $\sum_{i=1}^{20} f_i * log_2(f_i/b_i)$, where b_i is the global background frequency of amino acid i. Figure 1 illustrates how this calculation is carried out. To create a high relative entropy multiple alignment, a complete alignment of the nineteen sequences described above is created using Clustal W (Thompson *et al.* 1994). This alignment is then checked for higher order structural concordance using the codon-coloring feature of Aligner (Eernisse 1995). The relative entropy of each position in the alignment is calculated, and sites containing less relative entropy than the given threshold are discarded from the

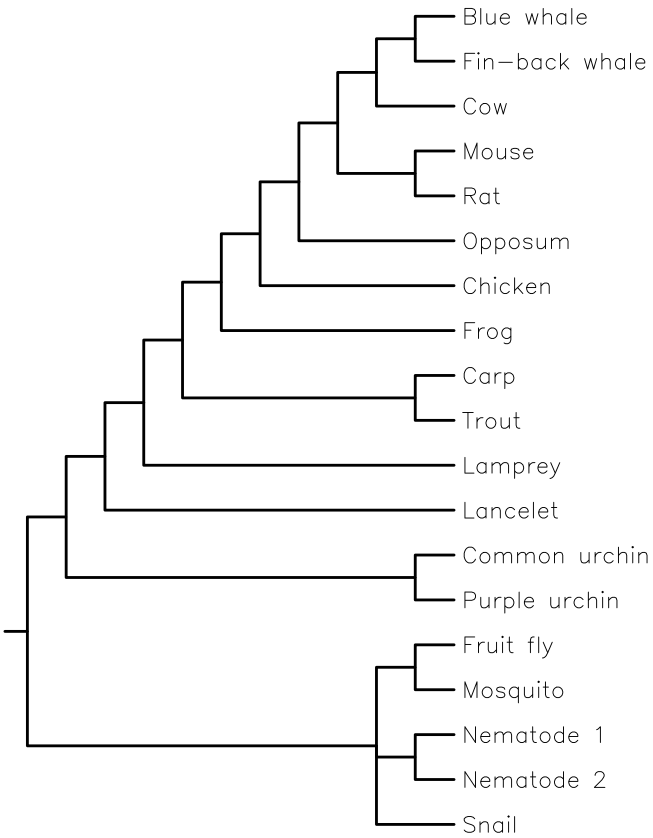

Figure 2: **The accepted phylogenetic tree for the metazoan data set.** One branch of the tree is incompletely resolved because the relationships among the molluscs, nematodes and arthopods are not universally accepted.

alignment.

The relative entropy threshold t used in constructing the conserved alignment is analogous to n, the number of motifs used for the motif-only alignments. Both parameters indirectly determine the number of sites in the conserved alignment, and a theoretically correct means of determining either parameter with respect to a given data set is not currently known. Therefore, as for n, sensitivity analysis is performed on t, varying its value from 0 bits to the maximum relative entropy in the alignment (6.8 bits) in increments of 0.05.

From each conserved multiple alignment, a maximum parsimony phylogenetic tree is inferred using Phylip (Felsenstein 1989). For the alignments with high relative entropy, this tree inference procedure is repeated ten times. Inferred trees are compared to the known, true tree for this data set (see Figure 2) by counting the number of branches by which the two trees differ (Bourque 1978; Robinson and Foulds 1981). The number of errors assigned to an inferred tree is the total number of incorrect or missing branches that it contains, relative to the true tree. Branches that appear only in the inferred tree and resolve previously unresolved portions of the true tree are not counted as errors.

Species name	Common name	Amino acids
Struthio camelus	Ostrich	3739
Dasypus novemcinctus	Armadillo	3786
Cavia porcellus	Guinea pig	3790
Crossostoma lacustre	Loach	3800
Xenopus laevis	Frog	3782
Halichoerus grypus	Grey seal	3795
Gallus gallus	Chicken	3785
Didelphis virginiana	Opossum	3829
Gadus morhua	Cod	3799
Rhinoceros unicornis	Rhinoceros	3792
Macropus robustus	Kangaroo	3785
Ceratotherium simum	White rhino	3793
Mus musculus	Mouse	3785
Equus caballus	Horse	3789
Rattus norvegicus	Rat	3794
Erinaceus europaeus	Hedgehog	3790
Gorilla gorilla	Gorilla	3789
Bos taurus	Cow	3791
Homo sapiens	Human	3789
Cyprinus carpio	Carp	3793
Phoca vitulina	Harbor seal	3795
Oncorhynchus mykiss	Trout	3802
Hylobates lar	Gibbon	3789
Latimeria chalumnae	Coelocanth	3790
Petromyzon marinus	Lamprey	3803
Pan paniscus	Chimpanzee	3789
Pongo pygmaeus	Orangutan	3789
Felis catus	Cat	3792
Balaenopterus physalus	Fin-back whale	3790
Balaenopterus musculus	Blue whale	3790
Ornithorhynchus anatinus	Platypus	3786
Polypterus ornatipinnis	Birchir	3787
Equus asinus	Donkey	3791
Protopterus dolloi	Lungfish	3788
Paracentrotus lividus	Common urchin	3824
Strongylocentrotus purpuratus	Purple urchin	3825
Arbacia lixula	Black urchin	3667
Asterina pectinifera	Starfish	3823
Branchiostoma floridae	Lancelet	3740

Table 2: **Species included in the vertebrate data set.** The last five species in the table constitute a collective outgroup. The entire alignment contains 3972 sites and is available at http://www.cse.ucsc.edu/research/compbio/phylo.

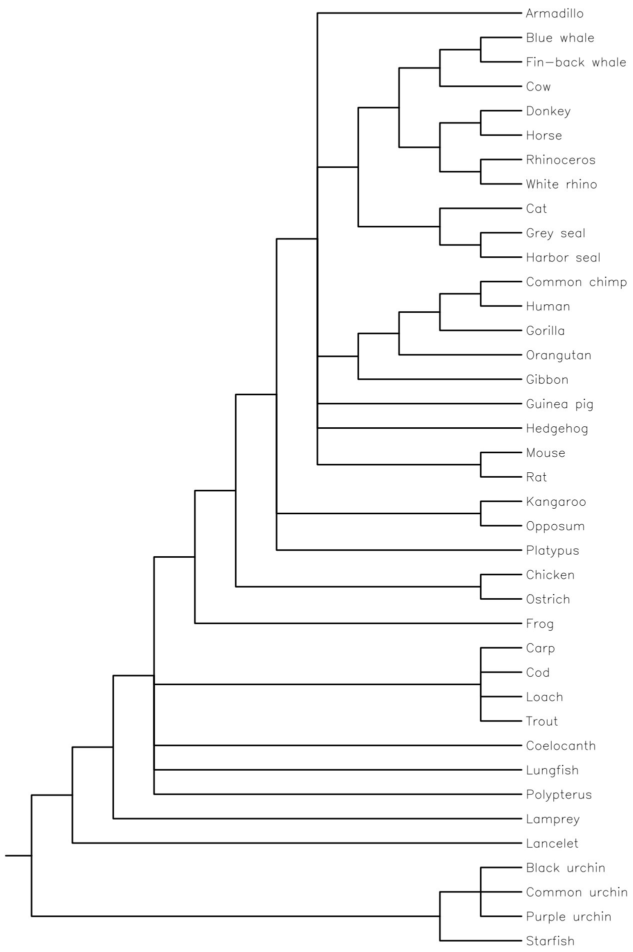

Figure 3: **Known phylogenetic relationships among the vertebrate data set.**

As a further test, both conserved alignment inference methods are applied to a larger data set containing the mitochondrial coding sequences from 34 vertebrate species plus a five-member collective outgroup (see Table 2). This data set contains a total of 147 755 amino acids. Unlike the previous data set, the phylogenetic relationships among many of these 39 species are not known. Figure 3 shows the incompletely resolved tree representing the phylogenetic relationships that are known with relative certainty, based upon morphological and fossil evidence (Benton 1993; Gauthier *et al.* 1988; Maisey 1986; 1988).

3 Results

Phylogenetic trees inferred from a Meta-MEME motif alignment are more accurate than trees inferred from the complete alignment. We believe this approach may be especially useful in cases involving deep divergences where there is a lot of sequence data, and will become increasingly important as comparative genomic data bases become established. Figure 4(b) shows the number of errors, relative to the true tree, for trees inferred from alignments containing varying numbers of MEME motifs. Ten trees inferred from the complete alignment uniformly contain six errors relative to the true tree (see Figure 6(a)). However, alignments based upon 50 or more motifs provide more accurate trees, and sixteen of the inferred trees are completely consistent with the known phylogeny. When the number of motifs exceeds 80, most trees contain two or fewer errors.

Figure 6(b) shows the tree most commonly inferred from motif alignments based upon 80 or more motifs. This tree incorrectly groups the frog with the fishes. Occasionally, trees inferrred from motif aligments instead incorrectly group the lancelet with the echinoderms. The misplacement of the frog also occurs in trees inferred from the complete alignment. However, that tree contains two additional misplacements: grouping the chicken with the frog and fishes, and placing the lancelet outside of a clade comprising vertebrates and echinoderms. The elimination of these two misplacements when using motif alignments does not result from a reduction of errors in the alignment itself: due to the high conservation of this data set, aligning the motif regions is relatively easy. Inspection of one Meta-MEME motif alignment showed it to be identical to the corresponding motif regions of the complete alignment. Thus, the improvement of phylogenies derived from motif-based alignments must derive from the elimination of non-motif sites that harbor a misleading, hierarchical signal.

The bootstrap percentage values and decay indices shown in Figures 6(a) and (b) show that the tree based on the entire multiple alignment, while inaccurate, is better supported than the more accurate tree based upon only the motif regions. This difference is not surprising: tree (a) is derived from all 4078 alignment sites, whereas tree (b) is derived from only 1024 sites.

The MEME motif analysis identifies clusters of highly conserved sites. Removing this clustering constraint further increases the accuracy of the inferred phylogenies. Figure 5(b) shows the error rate of inferred phylogenies as a function of the relative entropy threshold t. For values of t between 1.7 and 2.2 bits, most of the inferred trees are completely consistent with the true tree (see Figure 6(c)). Thus, for this data set, the degree of conservation of a site, rather than its appearance within a conserved motif, is the most useful selection criterion for creating a conserved alignment for phylogenetic analysis.

The data shown in Figure 5(b) is U-shaped, with high error rates occurring at high and low values of t. For low values of t, the conserved alignment is large, approaching the size of the entire alignment. Thus, errors from alignments with $t < 1.7$ bits arise from the same, misleading signal that causes errors in trees derived from the entire alignment. On the other hand, when $t > 2.2$ bits very few sites are included in the conserved alignment. Consequently, the tree inferred from the alignment is unresolved, as evidenced by the large number of missing branches relative to the true tree. Thus, for low values of t, errors arise from the misleading, non-historical signal, whereas for high values of t, errors arise from the under-determination of the inferred tree.

For a fixed error rate, motif alignments contain far fewer sites than high relative entropy alignments. Figure 4 shows that motif-based alignments containing more than 80 motifs consistently yield an average of two or fewer errors. On average, these alignments contain no more than 955 sites. By contrast, the smallest high relative entropy alignment yielding two errors contains 2119 sites. This difference suggests that the motif constraint will be useful for smaller or less conserved data sets, when fewer sites with high relative entropy are available.

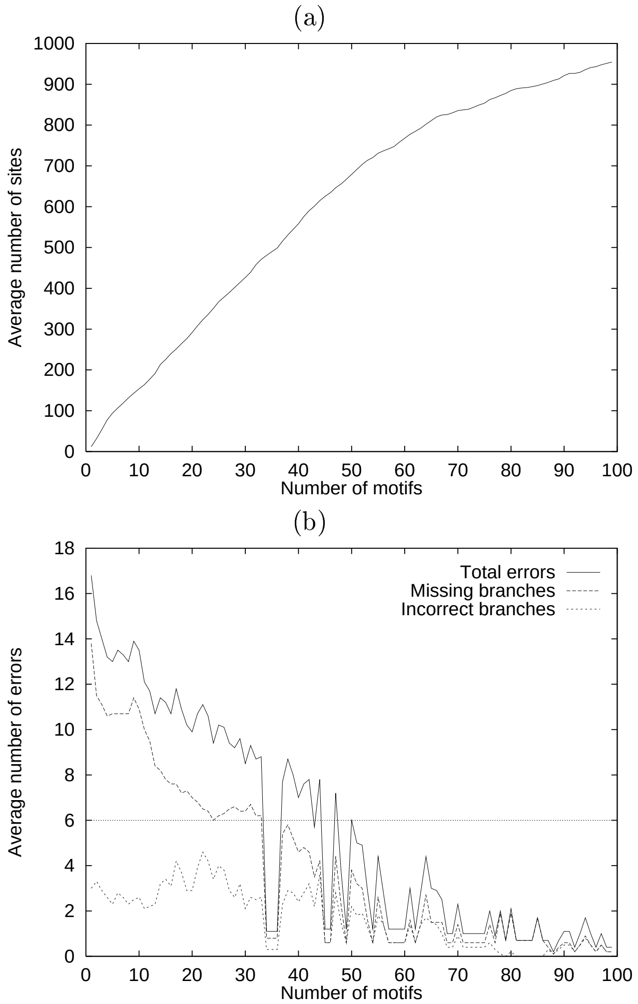

Figure 4: **Improved phylogenetic trees derived from alignments of motif regions.** Figure (a) shows the total number of sites in the conserved alignment as a function of n, the number of motifs discovered by MEME. Figure (b) plots the total number of phylogenetic inference errors as a function of n, averaged over ten runs. The total number of errors is the sum of the number of missing branches (i.e., branches that appear in the true tree but not the inferred tree) and the number of incorrect branches (i.e., branches that appear in the inferred tree but not the true tree). Branches that appear only in the inferred tree and resolve previously unresolved portions of the true tree are not counted as errors. The horizontal line in Figure (b) represents the total number of errors in a tree inferred from the entire alignment.

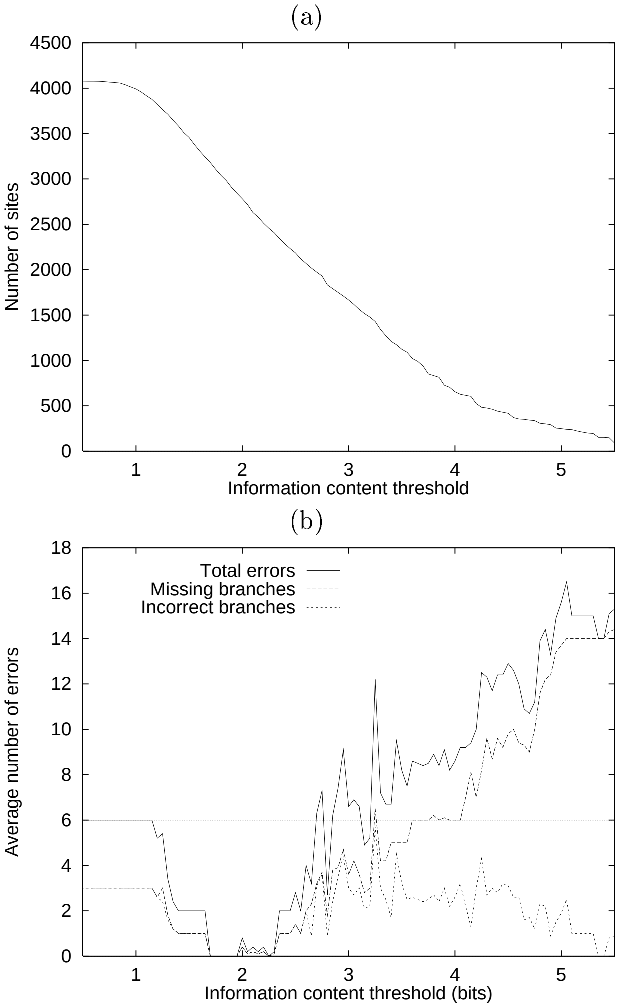

Figure 5: **Improved phylogenetic trees derived from alignments of high relative entropy regions.** Figure (a) shows the total number of sites in the conserved alignment as a function of t, the relative entropy threshold. Figure (b) plots the total number of phylogenetic inference errors as a function of t, averaged over ten runs. Errors are computed as described in Figure 4. The horizontal line in Figure (b) represents the total number of errors in a tree inferred from the entire alignment.

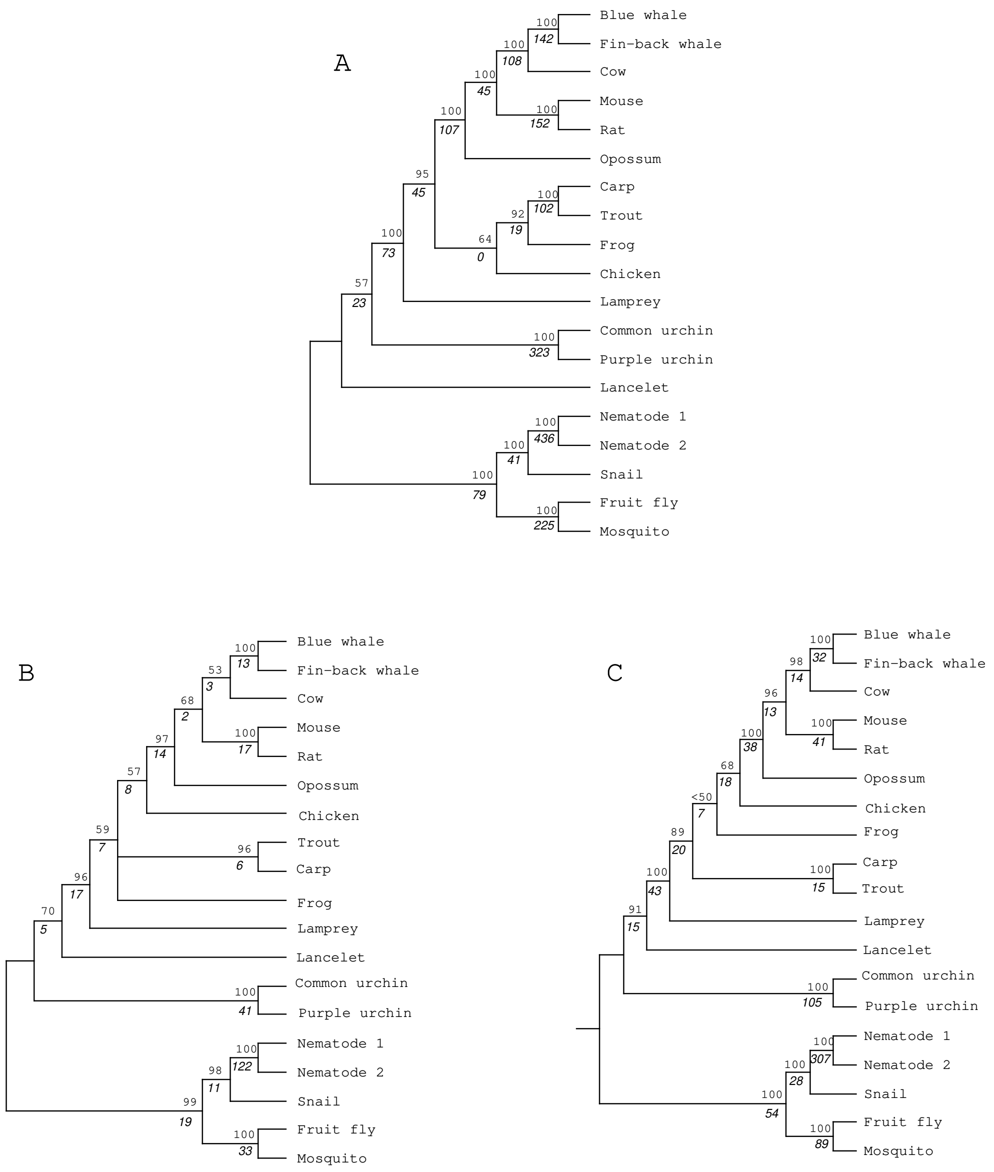

Figure 6: **Comparison of phylogenetic trees inferred from three different alignments.** Figure (a) shows the phylogenetic tree that results from a parsimony analysis of the complete multiple alignment. Figure (b) shows the most common tree given by motif-only alignments containing more than 80 motifs, and (c) shows the best tree from a high relative entropy alignment. Bootstrap percentage values based upon 1000 repetitions are given above each branch, and decay indices (the number of steps under parsimony before the node collapses) are in italics below the branch. None of the trees include branch length information.

Topology	Complete	Motifs	Relative entropy
Most parsimonious tree	[0.0995]	0.0158	0.0055
Switch *Amphioxus* and echinoderms	0.2649	0.1336	0.0548
Cluster fish and frog together	0.6037	1.0000	0.1260
Cluster fish, frog and chicken together	0.8442	0.0124	0.0109

Table 3: **Templeton test analyses.** Each entry in the table is the p-value for the true tree fitting an alignment better than the given tree, except for the single value in brackets, which is the p-value for the most parsimonious tree fitting the complete alignment better than the true tree. The most parsimonious tree is defined relative to the complete alignment and is shown in Figure 6(a). The other trees are variations on the most parsimonious tree, as described in the table. The three alignments are those used to generate the three trees in Figure 6.

The difference in the number of sites included in the two types of conserved alignments provides improved statistical support for high relative entropy alignments. The tree in Figure 6(c) is based upon 2407 sites. Consequently, the bootstrap percent values and decay indices shown in Figure 6(c) are much stronger than the corresponding values for the tree inferred from a motif alignment. All of the decay values for the relative entropy alignment are more than double those seen in the motif alignment, with a corresponding increase in bootstrap support values. Bootstrap support values for the relative entropy alignment are above 68 percent in all cases except for tetrapods. The latter is due to the weak molecular support for the inclusion of the frog in the tetrapod clade.

A more precise measure of the statistical support for a tree is provided by the Templeton test (Templeton 1983). This test can provide an estimate of the p-value of one tree matching a given alignment better than a second tree. As indicated in Table 3, the complete alignment matches the most parsimonious tree in Figure 6(a) better than the true tree with a p-value of 0.0995. The same is not true for the two conserved alignments. For both the motif alignment and the high relative entropy alignment, the true tree matches the data better than the original, most parsimonious tree, with p-values 0.0158 and 0.0055, respectively. Thus, both methods of constructing conserved alignments yield statistically significant support for the correct phylogeny.

Templeton tests were also conducted to determine the degree of support for individual branch arrangements when they differed from those of the expected tree. In the relative entropy alignment, the placement of amphioxus in its expected position was favored over its placement outside echinoderms with a p-value of 0.0548. Similarly, the expected arrangement of frog, chicken and the two fishes is favoured over the arrangement depicted in fig 6a with a p-value 0.0109

The relationships among the molluscs, nematodes and arthropods are not universally accepted (Aguinaldo *et al.* 1997). The results from the conserved alignment analyses suggest that the grouping shown in Figure 6 is correct: the molluscs belong with the nematodes. This grouping appears in every inferred tree that is consistent with the true tree, whether inferred from a motif alignment, a high relative entropy alignment, or the original Clustal W alignment.

Applying the conserved alignment methods to a larger data set of mitochondrial sequences yields a tree that is nearly consistent with known phylogenetic relationships. The tree derived from a parsimony analysis of the complete alignment of 39 species yields eight errors relative to the trusted phylogeny shown in Figure 3. Applying to this data set one of the best HMMs from the previous analysis yields a tree with only six errors. The error rate improves even further using high relative entropy alignments. Figure 7 shows the results of varying the relative entropy threshold, yielding a tree containing one error when $t = 2.12$ bits. The error is a lack of resolution among the perisodactyls, whale-cow and carnivore.

The inferred tree depicted in Figure 8 is nearly consistent with the known relationships among these taxa (Figure 3). Furthermore, the tree contains a number of noteworthy inferred relationships. Most striking, perhaps, is the fact that neither the lungfish (*Protopterus*) nor the coelocanth (*Latimeria*) fall as the sister group to the tetrapods, as generally contended (Helfman *et al.* 1997; Cloutier and Ahlberg 1996). Instead, the sister group to the tetrapods is a clade containing the birchir (*Polypterus*), the coelocanth and the neopterigian fishes. The lungfish is the inferred sister taxon to the clade containing both tetrapods and

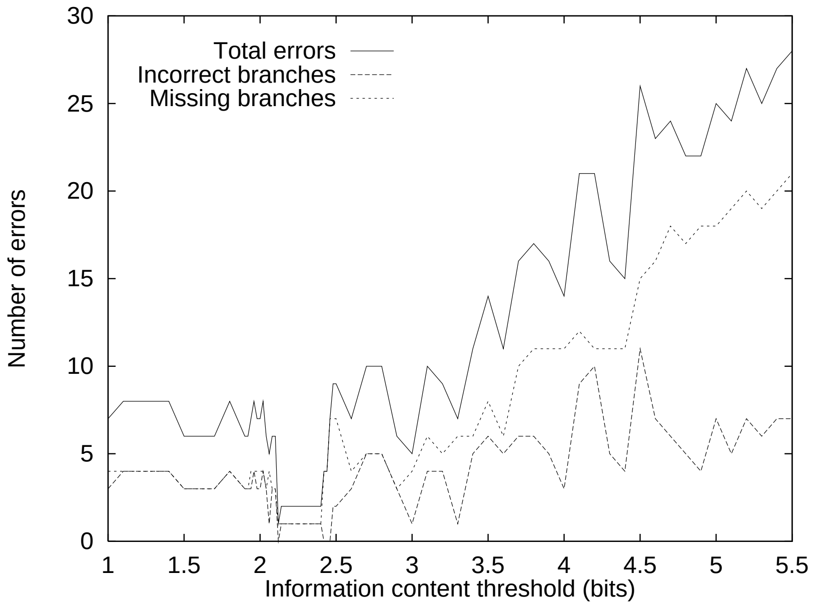

Figure 7: **Selecting the optimal relative entropy threshold for the vertebrate data set.** The figure plots the number of errors in an inferred phylogenetic tree as a function of relative entropy threshold t. Errors are computed relative to the set of known phylogenetic relationships depicted in Figure 3 using the method described in Figure 4.

the aforementioned novel clade. Other unexpected placements are the guinea pig as the sister-group to the primates, and the hedgehog and armadillo as sequential outgroups to the ferrungulata (carnivores and ungulates).

4 Discussion

This work suggests that inferring a phylogenetic tree from a conserved multiple sequence alignment can provide a significantly more accurate phylogeny than would be inferred from the complete alignment. Focusing on the conserved regions of the alignment, either within motifs or at individual alignment sites, strengthens the historical signal relative to any misleading signal in the data.

For the two data sets examined here, the relative-entropy constraint provides more accurate phylogenies than does the motif constraint. This result suggests that the level of conservation of a site is a more accurate guide in selecting historically informative sites than is the site's proximity to other conserved sites.

On the other hand, the site-clustering constraint used by the motif analysis leads to improved phylogenies when the total number of sites in the conserved alignment is small. For a fixed error rate, a motif alignment contains far fewer sites than a high relative entropy alignment. This indicates that motif analysis may be particularly appropriate for smaller data sets in which fewer sites are available. Indeed, in the MEME motif analyses described above, the motif models for both data sets are learned from only three or four mitochondrial sequences, rather than the entire data set. By using two constraints (relative entropy and site clustering), the motif analysis can do a better job of separating signal from noise in small data sets than using one constraint (relative entropy). Thus, for smaller or more divergent data sets, which contain fewer or more difficult to recognize highly conserved sites, the clustering constraint aids in identifying a small but accurate set of historically informative sites.

In addition to focusing on conserved regions, both of these conserved alignment methods eliminate noisy portions of the input alignment. Any phylogenetic inference algorithm that takes as input a multiple sequence

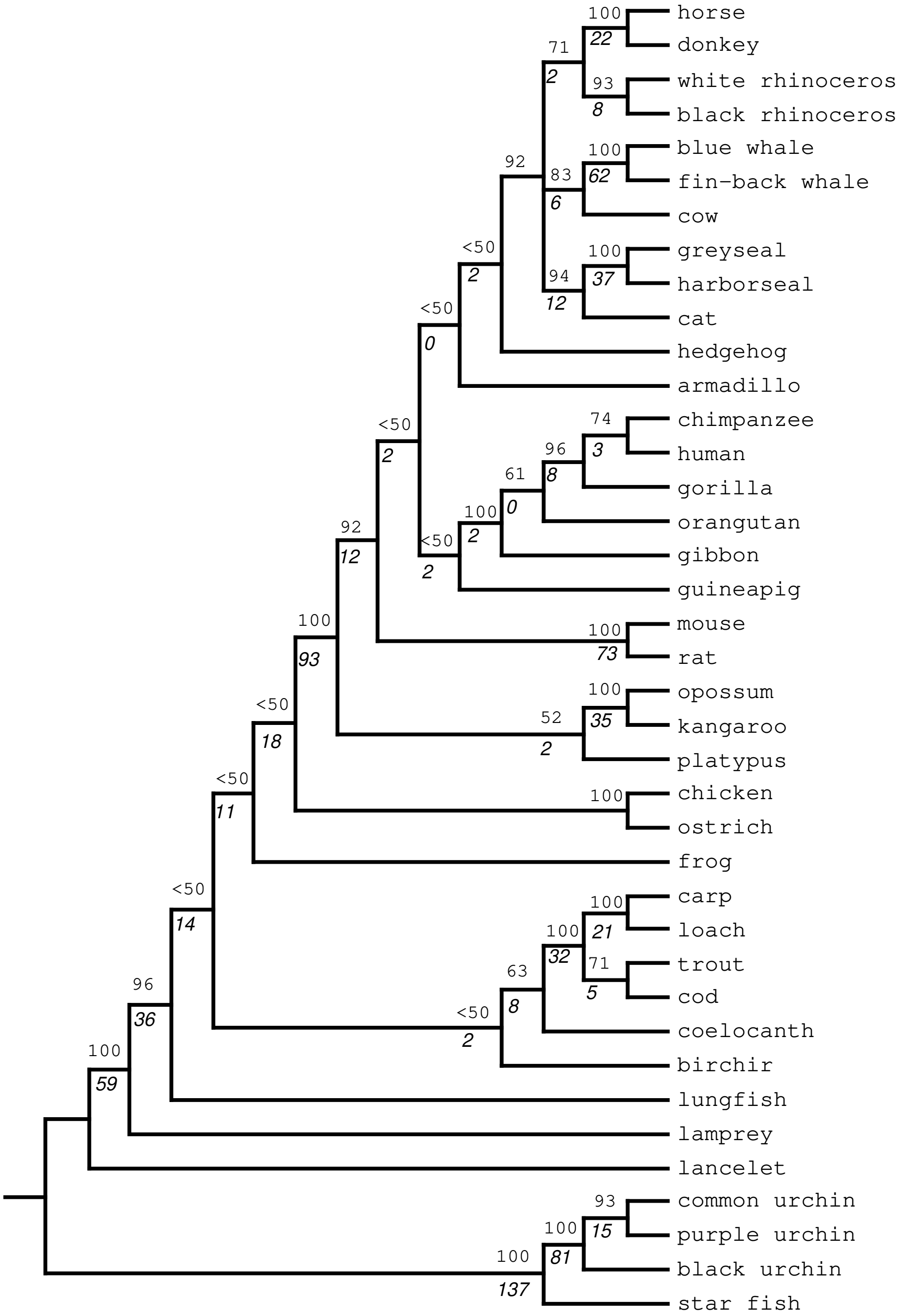

Figure 8: **Predicted phylogeny for the vertebrate data set.** This tree is inferred via parsimony analysis from a multiple alignment containing only sites with relative entropy greater than 2.12 bits. Bootstrap percentage values based upon 1000 repetitions are given above each branch, and decay indices are in italics below the branch. The tree is consistent with the known relationships depicted in Figure 3, except for the lack of resolution at the node above the horse-donkey-rhinos, whales-cow and seals-cat.

alignment will necessarily perform poorly if the input contains alignment errors. Therefore, researchers commonly restrict the alignment input to regions for which they are relatively certain of the alignment (Baldauf *et al.* 1996; Laudet *et al.* 1992; Farrell 1998). The two conserved alignment methods described here provide a principled means of selecting regions of the alignment to discard.

The idea of constraining a sequence alignment to produce a more accurate phylogeny is not itself novel. Clustal W (Thompson *et al.* 1994), for example, provides an *ad hoc* means of constraining alignments prior to phylogenetic inference. In this method, any site in the initial alignment that contains an insertion or deletion is discarded. However, for the nineteen-species data set described above, this constraint eliminates only 874 positions and yields exactly the same phylogenetic errors as the complete alignment. The conserved alignment methods described here are more flexible and, for these data, more effective.

The methods developed here analyze the level of conservation at each site within a multiple alignment. As such, the methods are sensitive to the degree of evolutionary divergence among the aligned sequences. This sensitivity is reflected in the parameters n and t, as described above. This paper presents a method for estimating values for n or t that are specific to a given data set using a subset of the data for which the true phylogenetic relationships are well established. Future work will investigate methods for estimating these parameters in the absence of known phylogenetic relationships.

In addition, it will be important to determine the degree to which these site selection methods depend upon the number of sites, number of taxa, degree of divergence and types of genes included in a particular study. The MEME software has been shown to accurately discover motifs in very small sequence sets (Bailey and Elkan 1995); however, this ability is strongly dependent upon the degree of conservation of the motifs. The mitochondrial data sets analyzed here are among the most conserved for which motif analysis is useful. The information content criterion, on the other hand, is capable of differentiating more- and less-conserved sites even in extremely conserved alignments. MEME can extract conserved regions from highly divergent sequences (Bailey and Grundy 1999), and recent evidence (Hudak and McClure 1999) suggests that, for highly divergent data, alignments based upon such motifs are more accurate than whole-sequence alignments. Thus, for difficult-to-align data, a MEME motif-based alignment may provide a more accurate phylogeny due to the reduced number of alignment errors. This hypothesis, as well as other dependencies upon data set features of both site selection methods presented here, will be the subject of future work.

In this paper, all phylogenetic inferences are carried out using the Phylip implementation of maximum parsimony. There is no reason to suppose, however, that the results are specific to that inference program. Similar improvements would likely occur if conserved alignments were provided to other implementations of maximum parsimony or to other algorithms, such as maximum likelihood or neighbor joining.

One drawback to the MEME/Meta-MEME method is its computational expense. The problem of discovering motifs in unaligned sequences is much more difficult than the corresponding problem for aligned sequences. Thus, a motif-finding method that takes as input a multiple alignment would be computationally cheaper and, for this highly conserved data, would likely yield the same results.

The majority of the phylogenetic placements in Figure 8 corroborate traditional placements based on morphology. This lends credence to the conserved alignment approach proposed and raises the question, "If conserved alignments are reliable enough to endorse placements of which we are confident, should we not also give serious consideration to those placements that have not been hitherto proposed?" There are two such placements implied by the tree in Figure 8. The first is the unusual clustering of the fishes: the four teleost representatives form a monophyletic group with coelocanth and polypterus, while the lungfish falls as the sister-taxon to a clade containing the fishes and the tetrapods. The second is the placement of the guinea pig basally among the eutherian mammals (between the rodents, the armadillo and the stem branch leading to mammals). The inferred phylogenetic arrangement for the fishes is particularly interesting. Traditionally, fishes have been regarded as a grade that comprises multiple paraphyletic lineages. This traditional perspective is almost universally endorsed. The notion that the Actinopterigii (teleosts and polypterus) should have a sarcopterigian (the coelocanth) buried within the clade has not, to our knowledge, been proposed previously. This said, we caution that the inference may be the consequence of sparse sampling at the base of the vertebrate clade. In any event, the unorthodox phylogenetic arrangement warrants further investigation.

We believe that concentrating on sites that are conserved across taxa enhances the hierarchical signal-to-noise ratio among deeply divergent taxa. We base this claim on the empirically supported notion that protein function is predominantly attributable to a core of important residues and their interactions (Golding and

Dean 1998; Chothia and Lesk 1986). Those residues not vital to function are free to vary and covary with one another in ways that are not critical. Variation in non-critical residues causes noise; covariation among non-critical residues generates misleading hierarchical signal. Removing such sites from analysis is likely to amplify any historical signal that might otherwise have become obscured by their inclusion. While such signal amplification approaches can be highly effective (as shown by the examples herein), we caution that they are not guaranteed to yield correct phylogenetic inferences all of the time. Shifts in function or constraint release events in unrelated taxa could lead to convergent patterns that appear as anciently conserved motifs. Indeed, in the 39 taxon data set examined for this paper, it is possible that the unorthodox placements of the lungfish, guinea pig, hedgehog and armadillo may reflect such convergent forces. Ultimately, of course, the key to accurately eliminating the misleading signal in a multiple sequence alignment is to understand the forces that have shaped its variation across taxa. This will come from a better appreciation of the mapping between genotypes and phenotypes for particular genes across multiple taxa.

Acknowledgments

William N. Grundy is supported by a Sloan/DOE Fellowship in Computational Molecular Biology. Gavin J. P. Naylor is supported by NSF grant number DEB-9707145.

References

A. M. A. Aguinaldo, J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387:489–493, 1997.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

T. L. Bailey and C. P. Elkan. Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, 1994.

T. L. Bailey and C. P. Elkan. The value of prior knowledge in discovering motifs with MEME. In C. Rawlings, D. Clark, R. Altman, L. C. Hunter, and L. C. Rawlings, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 21–29. AAAI Press, 1995.

T. L. Bailey and M. Gribskov. The megaprior heuristic for discovering protein sequence patterns. In D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 15–24. AAAI Press, 1996.

T. L. Bailey and M. Gribskov. Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics*, 14(1):48–54, 1998.

T. L. Bailey and W. N. Grundy. Classifying proteins by family using the product of correlated p-values. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, April 1999. To appear.

T. L. Bailey. MEME – Multiple EM for Motif Elicitation. http://www.sdsc.edu/MEME, 1999.

S. L. Baldauf, J. D. Palmer, and W. F. Doolittle. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences of the United States of America*, 93:7749–7754, 1996.

P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, 91(3):1059–1063, 1994.

M. J. Benton. *The fossil record 2*. Chapman and Hall, 1993.

M. Bourque. *Arbres de Steiner et Reseaux dont Varie l'Emplagement de Certains Sommets*. PhD thesis, Univ. Montréal, Montréal, Canada, 1978.

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

M. Brown, R. Hughey, A. Krogh, I. Mian, K. Sjolander, and D. Haussler. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In C. Rawlings, editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 47–55. AAAI Press, 1995.

C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5:823–826, 1986.

R. Cloutier and P.E. Ahlberg. *Morphology characters and the inter-relationships of basal sarcopterigians*, pages 445–479. Academic Press, 1996.

M. J. Donoghue, R. G. Olmstead, J. F. Smith, and J. D. Palmer. Phylogenetic relationships of dipscales based on *rbcL* sequences. *Announcements of the Missouri botanical gardens*, 79:333–345, 1992.

S. R. Eddy. Multiple alignment using hidden Markov models. In C. Rawlings, editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120. AAAI Press, 1995.

S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

D. J. Eernisse. Stacks: HyperCard software utilities for molecular systematists, version 1.1. ftp://ftp.biology.-indiana.edu, 1995.

B. Farrell. "Inordinate fondness" explained: Why are there so many beetles? *Science*, 5376:555–559, 1998.

J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–791, 1985.

J. Felsenstein. PHYLIP — phylogeny inference pacakge (version 3.2). *Cladistics*, 5:164–166, 1989.

J. Gauthier, A. G. Kluge, and T. Rowe. Amniote phylogeny and the importance of fossils. *Cladistics*, 4:105–209, 1988.

G. B. Golding and A. M. Dean. The structural basis of molecular adaptation. *Molecular Biology Evolution*, 15(4):355–369, 1998.

W. N. Grundy and C. P. Elkan. Meta-MEME version 2.0.1. http://metameme.sdsc.edu/, 1999.

W. N. Grundy, T. L. Bailey, and C. P. Elkan. ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *Computer Applications in the Biosciences*, 12(4):303–310, 1996.

W. N. Grundy, T. L. Bailey, C. P. Elkan, and M. E. Baker. Meta-MEME: Motif-based hidden Markov models of protein families. *Computer Applications in the Biosciences*, 13(4):397–406, 1997.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of the mitochondrial DNA. *Journal of Molecular Evolution*, 21:160–174, 1985.

G. S. Helfman, B.B. Collette, and D.E. Facey. *The Diversity of Fishes.* Blackwell Science, 1997.

J. Hudak and M. A. McClure. A comparative analysis of computational motif-detection methods. In *Pacific Symposium on Biocomputing*, volume 4, pages 138–149, 1999.

T. H. Jukes and C. R. Cantor. *Evolution of protein molecules*, pages 21–132. Academic Press, 1969.

M. Kimura. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

V. Laudet, C. Hanni, J. Coll, F. Catzelflis, and D. Stehelin. Evolution of the nuclear receptor gene superfamily. *EMBO Journal*, 11:1003–1013, 1992.

C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

J. G. Maisey. Heads and tails: A chordate phylogeny. *Cladistics*, 2:201–256, 1986.

J. G. Maisey. *Phylogeny of Early vertebrate Skeletal Induction and Ossification Patterns*, volume 22. Plenum, 1988.

G. J. P. Naylor and W. M. Brown. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Systematic Biology*, 47(1):61–76, 1998.

D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.

F. Rodríguez, J. L. Oliver, A. Marín, and J. R. Medina. The general stochastic model of nucleotide substitions. *Journal of Theoretical Biology*, 142:485–501, 1990.

K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences*, 1996.

D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. *Phylogenetic Inference*, pages 407–514. Sinauer Associates, 1996.

S. Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on the Mathematical Life Sciences*, 17:57–86, 1986.

A. R. Templeton. *Statistical Analysis of DNA Sequence Data*, chapter Convergent evolution and non-parametric inferences from restriction fragment and DNA sequence data, pages 151–179. Marcel Dekker, New York, 1983.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.