

Improving progressive alignment for phylogeny reconstruction using parsimonious guide-trees

Usman Roshan^{*}, Dennis R. Livesay[†], Satish Chikkagoudar^{*}

^{*} Department of Computer Science, New Jersey Institute of Technology, GITC 4400, University Heights, Newark, NJ 07102, USA; usman@oak.njit.edu and sc243@oak.njit.edu; [†] Department of Chemistry and Center for Macromolecular Modeling & Materials Design, California State Polytechnic University-Pomona, 3801 W. Temple Ave., Pomona, CA 91768, USA; drivesay@csupomona.edu.

ABSTRACT

Motivation: Phylogenies are widely used for understanding the evolutionary histories of species and gene products. Moreover, phylogenies are increasingly utilized in phylogenomic and other comparative analyses. Maximum parsimony (MP) and maximum likelihood (ML) are commonly used optimization criteria for constructing phylogenies. However, phylogenetic descriptions depend not only on the employed reconstruction technique, but also on the underlying sequence alignment (Goldman 1998). Here, we establish a simple prescription to improve the underlying alignments used in phylogeny reconstruction.

Results: We adapt Gotoh's (1996) iterative strategy for improving a progressive alignment (by using better guide-trees) specifically for the purpose of constructing optimal MP phylogenies. We improve the progressive alignment heuristic as implemented in the MUSCLE alignment program by iterating with maximum parsimony guide-trees constructed using PAUP*, yielding both deterministic and randomized heuristics. We show that phylogenies on our alignments are better supported on real data than those on Dialign, Probcons, ClustalW, and the default MUSCLE alignments, none of which utilize parsimony guide-trees. We evaluate accuracy on simulated data under a wide range of model conditions and show that phylogenies produced using our technique are more accurate than phylogenies on other alignments.

1. INTRODUCTION

Phylogenies are a fundamental tool for understanding the evolutionary history of species (Gusfield 1997). In the post-genomic era, they play a key role in understanding protein family differentiation, and are commonly used in phylogenomic analyses (Eisen 1998). Phylogenies (protein and DNA) are generally constructed using the maximum parsimony (MP) or maximum likelihood (ML) optimization criteria (Swofford et. al. 1996), both of which are widely used. MP results in phylogenetic trees with the minimum number of mutations, whereas ML results in phylogenies with the maximum probability of observing the data given the tree. A widely used software package for constructing MP and ML trees is PAUP* (Swofford 1996). Bayesian approaches are increasingly popular; however, we exclude them from our analyses at present.

The most important input to a phylogeny reconstruction method is a multiple sequence alignment. The progressive alignment strategy of Feng and Dolittle (1987) is a fast and widely used heuristic for aligning multiple sequences to a guide-tree (i.e. phylogenetic tree sequence alignment). For example, the popular ClustalW (Thompson et. al. 1994) program uses a progressive alignment combined with improvements built around it such as sequence weighting and automatic gap penalties. Guide trees for progressive alignment are usually obtained by simple distance-based approaches such as neighbor joining or UPGMA (Swofford et. al. 1996), where distance matrices are constructed using pairwise alignments. More recent programs such

as MUSCLE (Edgar 2004) use a combination of progressive alignment and other approaches that optimize the sum-of-pairs criterion for alignment.

Most previous phylogenetic reconstruction studies have focused on constructing optimal trees with the alignment fixed. However, the input alignment is known to affect the reconstructed phylogeny (Morrison and Ellis 1997; Goldman 1998; Hall 2005). Consequently, improving the alignment input could lead to better phylogenies. In this report we demonstrate that a simple MP iterative refinement method, based on Gotoh's (1996) doubly nested randomized iterative technique, can result in significantly improved sequence alignments for phylogeny reconstruction. We compare our approach to Dialign, Probcons, the standard ClustalW, and different stages of MUSCLE on real and simulated data.

2. BACKGROUND

2.1 Phylogeny reconstruction. As stated above, MP and ML are two widely used optimization criteria for phylogeny reconstruction (Swofford et. al. 1996; Sanderson 1993). Both are known to be NP-hard. However, in practice, heuristic ML implementations are orders of magnitude slower than MP (Swofford et. al. 1996). Consequently, we only examine MP for constructing phylogenies in this preliminary investigation; future work will also investigate ML. We define the MP problem below and briefly review standard hill-climbing heuristics for solving it.

Given a phylogenetic tree and sequences (of the same length) assigned to internal nodes, we define the parsimony length as the sum of the residue mutations on each edge of the tree. The residue mutations on a given edge are measured by computing the mutations between the sequences assigned to each node of the edge. The MP problem is to find a tree and sequences assigned to the internal nodes such that the parsimony length is minimized. MP was proven NP-hard by Foulds and Graham (1982).

Hill-climbing heuristics for MP begin with a starting tree and then modify it to find better ones. A popular tree modification is Tree Bisection and Reconnection (Swofford 1996). In the TBR move, a tree T is bisected by removing an edge e , which yields two subtrees, and then connected by connecting any two edges of each subtree. After trying all possible TBR moves on a given tree, the one with the best score is selected and this process iterates until we reach a local minimum. TBR is efficiently implemented in PAUP* among other commonly used phylogeny reconstruction techniques.

2.2 Multiple sequence alignment. Like MP and ML phylogenetic reconstruction, standard optimization criteria for multiple sequence alignment, i.e. sum-of-pairs and phylogenetic tree alignment (Gusfield 1997), are also NP-hard. Sum-of-pairs (SP) aims to maximize the sum of pairwise similarity between the input sequences, whereas phylogenetic tree alignment aims to minimize dissimilarity along the edges of a given tree. The progressive alignment strategy of Feng and Dolittle (1987) has been adapted into most software packages for alignment, the most popular being ClustalW (Thompson et. al. 1994), because of its speed and accuracy. It has polynomial running time as a function of the number and length of sequences and is considered to be biologically realistic as it aligns to guide-tree (Altschul and Lipman 1989).

This heuristic essentially performs a post-order walk on a given binary guide-tree aligning pairs of sequences or profiles at a time, and then a pre-order walk to insert gaps and compute the final alignment (Edgar 2004). The initial guide-tree is usually obtained by computing a UPGMA or neighbor joining tree on pairwise alignment distances. Profiles are a standard way to represent an alignment of sequences and can be used to align two alignments on separate sets of sequences.

Two profiles are aligned using the same Needleman-Wunsch (Gusfield 1997) algorithm for aligning two sequences as long as one defines how to score two profile positions.

Various programs have implemented improvements around the basic progressive alignment. ClustalW implements ideas such as sequence weighting and automatic gap penalties that are designed to improve the alignment based on biologically sound assumptions (Thompson et. al. 1994). ClustalW uses neighbor joining for a guide-tree. MUSCLE is a three-stage program each of which we study separately for this paper. Stage I is the basic progressive alignment on a UPGMA guide-tree. Stage II is Gotoh's (1996) iterative heuristic but without SP optimization, i.e. compute alignment on starting UPGMA tree, compute UPGMA tree on alignment, recompute alignment on UPGMA tree, and iterate until the UPGMA guide-tree in the current iteration is the same as the one from the previous one. Stage III is a SP optimization on the alignment from stage II. In this stage, the alignment is divided into sets of sub-alignments as determined by the UPGMA tree from stage II, columns containing only gaps in the sub-alignments are removed, and the profiles of the sub-alignments are realigned to determine if an alignment with a better SP score can be found. This continues until a specified number of iterations elapses or a better alignment is not found after iterating over all edges of the tree.

2.3 Simulation. Simulations are commonly used to evaluate phylogenetic accuracy since we have no way of knowing "true" evolutionary trees (Sjolander 2004). The ROSE software package (Stoye et. al. 1998) implements the HKY85 (Hasegawa et. al. 1985) model of DNA sequence evolution, but also allows for insertions and deletions. This allows simulation of biologically realistic sequences on which both phylogeny reconstruction and alignment programs can be tested. We focus here on accuracy of phylogeny reconstruction.

Given the true tree (which we know since we are simulating data) and an estimated tree, we can use the Robinson-Foulds distance (Robinson and Foulds 1981) to measure accuracy, defined as follows. Every edge e in a leaf-labeled tree T defines a bipartition (induced by the deletion of e); the tree T is uniquely encoded by the set of bipartitions of all internal edges of T . If T is the true tree and T' is the reconstructed tree then the RF distance is the number of bipartitions in the reconstructed tree that are not present in the true tree (false positives) and those in the true tree but not in the reconstructed tree (false negatives). We define the error rate to be the normalized RF distance, which is obtained by dividing the RF distance by $n-3$, the number of internal edges in a binary tree on n leaves. We present the error rate as percentages (between 0 and 100).

3. IMPROVED PROGRESSIVE ALIGNMENT

Gotoh (1996) introduced a doubly nested randomized iterative method which iterated between progressive alignments and distance-based UPGMA phylogenies. His approach was implemented in the PRRP software package (Gotoh 1996) and in stage II of the MUSCLE program (Edgar 2004). We modify this approach by alternating between MP trees and progressive alignments and output the pair of alignment and tree with the best MP score. This approach is expected to be more accurate as MP is widely considered to produce better phylogenies than UPGMA. We implemented this heuristic using the MUSCLE program (for computing the progressive alignment) and PAUP* (for computing MP trees) and call it MUSCLE-PARS (see Figure 1). Our approach is specifically designed to find alignments and phylogenies that optimize the MP score, and thus is likely to be more appropriate for phylogeny-centric applications, i.e. predicting functional sites with phylogenetic motifs (La et al. 2005; Roshan et al. 2005).

MUSCLE offers a simple and fast implementation of progressive alignment without any additional options, i.e. sequence weighting which affect the order in which the sequences are aligned. In MUSCLE-PARS we strictly follow the order of the tree in aligning sequences. PAUP* implements various hill-climbing heuristics for solving MP. The MP heuristic we use builds a starting tree by adding sequences in the order of their closeness (see Swofford et. al. 1996 for more details). Once the tree is constructed, a TBR-based hill-climbing search is applied to it (as explained in Section 2.1). The initial starting tree for the search can also be built by adding sequences in a random order instead of their closeness; this produces a randomized search heuristic since each time the search starts from a different tree. We use the former deterministic search for MP so that MUSCLE-PARS is also deterministic. We leave a thorough study of the randomized version of MUSCLE-PARS to a later study. Here we examine randomized MUSCLE-PARS only on real data and the deterministic version on both real and simulated data.

Figure 1: Description of MUSCLE-PARS.

Input: unaligned sequences S , initial guide-tree T , number of iterations n
Output: alignment A^* and guide-tree T^*
Algorithm:
 (1) Set best score bs to *infinity*.
 (2) Compute MUSCLE progressive alignment A on guide tree T .
 (3) Compute MP score $MP(T,A)$ of tree T on alignment A . If $MP(T,A) < bs$ then set bs to $MP(T,A)$, A^* to A , T^* to T .
 (4) Compute MP tree T on A using PAUP*. If number of iterations not done then go to 2. Else return A^* and T^* .

MUSCLE-PARS differs from Gotoh's original implementation in several key ways. First, the original method of Gotoh (1996) used UPGMA trees instead of MP. Second, Gotoh's method performed SP optimization on the progressive alignment *before* recomputing a phylogeny on it. We do not perform this additional optimization step because it does not necessarily improve accuracy and extends running time (data not shown here). Third, the stopping criterion for Gotoh's method is when the UPGMA tree does not change; Gotoh's method usually reaches convergence in a few iterations. MUSCLE-PARS uses parsimony trees (that may be deterministic or randomized) which provides no guarantee of convergence; alignments and trees could get worse or improve with iterations. If the same alignment is obtained in two consecutive iterations, the MP trees (which are used for constructing the alignment of the following iteration) may not be the same if randomized heuristics are used. And fourth, the alignment outputted from Gotoh (1996) is the one from the most recent iteration. MUSCLE-PARS outputs the alignment and tree with the best MP score over all the iterations.

We note that the Poy (Wheeler 2002) software package is also designed to find highly optimal MP phylogenies. Poy starts from a guide-tree and performs a local search through tree space (using TBR and other tree modification operators) and selects better trees by recomputing the alignment and the MP score on each new tree it encounters. Poy does not perform a progressive alignment based on profiles (unlike MUSCLE-PARS); instead, it uses a different approach explained in Wheeler (1996). Profile progressive alignment, like that of ClustalW and MUSCLE, has acceptable performance on UPGMA trees which are also used as starting guide-trees for MUSCLE-PARS. However, Poy's alignment strategy performs better on optimal MP guide-trees than UPGMA trees (personal communication from Wheeler and unpublished studies). The approach we study here can greatly benefit local search heuristics like those in Poy and statistical alignments approaches (Redelings and Suchard 2005; Fleissner et. al. 2005) because it

can be used to obtain highly optimal starting trees very quickly. This will be studied in more detail in a forthcoming investigation.

We compare MUSCLE-PARS to the original Gotoh approach as implemented in MUSCLE stage II option. Stage II of MUSCLE is similar to Gotoh's (1996) original idea except that it does not perform a SP optimization between iterations. Thus we can directly compare the effect of using UPGMA and MP trees as iterating guide-trees for progressive alignment.

4. EXPERIMENTAL DESIGN

4.1 Software and methods studied. We compare ClustalW, Dialign (Morgenstern 1999), Probcons (Do et. al. 2005), MUSCLE and its three different stages to two variants of MUSCLE-PARS using default scoring matrices and gap penalties. The scoring matrices and gap penalties of the MUSCLE variants and MUSCLE-PARS are exactly the same; the only difference is in the guide-tree iterations. It is possible that better results could be obtained by selecting better gap penalties for all the methods studied here. Nevertheless, in this initial study, we use only default scoring matrices and gap penalties. We use the abbreviations MUSCLE-PROG to refer to stage I of MUSCLE (since it is just the progressive alignment), MUSCLE-UPGMA to refer to stage II, and MUSCLE to refer to the final stage III alignment. Additionally, we present two variants of MUSCLE-PARS. In the first, which we call MUSCLE-PARS1, the initial guide-tree is the UPGMA one constructed on pairwise alignment distances, and in the second one, which we call MUSCLE-PARS2, the initial guide-tree is the one used in the last iteration of MUSCLE-UPGMA. Thus, in a way, MUSCLE-PARS1 and MUSCLE-PARS2 attempt to further improve MUSCLE and MUSCLE-UPGMA by iterating with MP guide-trees. On both, real and simulated data, we perform 25 iterations of MUSCLE-PARS1 and MUSCLE-PARS2. We exclude Dialign and Probcons from the simulated data study due to limited time and space; both are much slower than MUSCLE and ClustalW. However, preliminary testing suggests that their performance is not better than MUSCLE-PARS on the simulated data (data not shown).

We construct MP phylogenies on all the alignments (on real and simulated data) using a more extensive TBR search heuristic than the basic strategy described in Section 2.1. In this version we repeat the basic hill-climbing search 100 times starting from a different randomized starting tree each time. This is known as the random restart method. The randomized starting tree is constructed by sequentially adding sequences (in a random order) on the edge that optimizes the MP score. Since PAUP* was used in MUSCLE-PARS, we use PAUP* for constructing MP phylogenies on all alignments (using the random restart heuristic). In line with standard phylogenetic studies (Swofford 1996; Felsenstein 1981), we treat gaps as missing data (default setting of PAUP*) when constructing MP trees on all the alignments and within MUSCLE-PARS. The effect of treating gaps as a fifth state increases the error rate for all alignment on the simulated data (data not shown here).

4.2 Real datasets. We study a real dataset of 208 Metazoa 18s sequences (average length of 1790 and std. dev. 96.6) provided by Ward Wheeler. We align each dataset using Dialign, Probcons, ClustalW, the three stages of MUSCLE, and 25 iterations each of MUSCLE-PARS1 and MUSCLE-PARS2. MP trees are constructed using the heuristic described above. We compute bootstrap supports (Felsenstein 1985) by performing 500 replicates and use the standard TBR hill-climbing search for MP (described in Section 2) to construct a tree for each replicate. We also study randomized MUSCLE-PARS2 which uses randomized MP trees within each iteration instead of deterministic ones (as explained in Section 3). Due to limited, we computed 55 randomized MUSCLE-PARS2 alignments with 10 iterations for each run.

4.3 Simulation study parameters. Simulation parameters are selected such that the MP tree on the true alignment has, at most, 15% error. Details of the simulation parameters are described below.

Model trees: We use birth-death model trees produced using the r8s software package (Sanderson 2004). This package has been used extensively in previous simulation studies and produces model trees that reflect our understanding of evolutionary processes and trees on real data. Birth-death trees produced by r8s are scaled to be ultrametric by default, which means that the evolutionary distance from the root to each leaf is the same. Biological trees on real data are not necessarily ultrametric; therefore, to deviate the tree from ultrametricity we randomly multiply each edge length by a deviation factor as described in Nakhleh et. al. (2002). A deviation of 1 means no deviation, 2 means small, and 4 is moderate deviation. We also multiply the edge lengths of each tree by scaling factors of 16, 32, and 64 to produce different levels of evolutionary rates. We generated 20 model trees of sizes 100, 200, and 400 taxa for each setting of deviation and scale to produce a total of 360 different model trees.

Sequence evolution: For each model tree we generate DNA sequences using ROSE under the HKY85 (Hasegawa 1985) model with transition/transversion ratio set to 2. We study two sequence lengths used at the root, 500 and 1000, and examine two different indel probabilities of 0.00005 and 0.0005. On each of the 360 model trees we evolved DNA sequences for each setting of sequence length and indel probability; thus, producing a total of 1,440 simulated datasets.

5. EXPERIMENTAL RESULTS

5.1 Real data. Bootstrap results: Bootstrapping (Felsenstein 1985) is frequently used to estimate confidence levels for phylogenies on real data. Although this technique has been criticized for a systematic bias towards lower values (Zharkikh and Li 1995; Newton 1996), it has been shown (Efron et. al. 1996) that it can be seen as a first order approximation of the accuracy of the tree's topology. Furthermore bootstrap supports can be used to evaluate phylogenetic signal of different alignments (personal communication with Bernard Moret and Joe Felsenstein). In Table 1, we list the percentage of highly supported edges in the MP trees for each method using the original bootstrapping technique of Felsenstein (1985). MUSCLE-PARS phylogenies yield the highest number of highly supported edges for different levels of cutoff. Table 1 also shows the bootstrap supports of guide-trees (when appropriate). Most of the edges with 95% support are also present in the guide-trees, but as we move to lower thresholds, fewer guide-tree edges are supported except for MUSCLE-PARS1 and MUSCLE-PARS2. In fact, the 50%, 75%, and 95% edges which are supported in the MP tree on MUSCLE-PARS1 and MUSCLE-PARS2 alignments are the same as those on the best MUSCLE-PARS1 and MUSCLE-PARS2 guide-trees. Consequently, we use the best MUSCLE-PARS1 and MUSCLE-PARS2 guide-trees as estimates of the MUSCLE-PARS1 and MUSCLE-PARS2 phylogenies henceforth.

To further understand the relationship between bootstrap edges on different phylogenies, Table 2 lists the number of edges with over 95% support that are present in one alignment and missing from the other. The last column of Table 2 shows that the MUSCLE-PARS2 phylogeny contains most of the highly supported edges from the end MP phylogenies on the other alignments. Furthermore, as seen from the last row of Table 2, MUSCLE-PARS2 phylogenies contain several highly supported edges that are not present in phylogenies on any of the other alignments. Both of these observations suggest that the MUSCLE-PARS phylogenies are more "informative" than trees on other alignments, and so can be used in a data-mining context. For example, programs such as Orthotrapp (Sonnhammer and Strom 2002), which use bootstrapped phylogenies for detecting orthologous proteins, should yield more useful

relationships if used with informative phylogenies like those of MUSCLE-PARS. Investigations along these lines are currently underway.

Randomized MUSCLE-PARS: Each run of randomized MUSCLE-PARS produces a different (but parsimonious) alignment and phylogeny pair. As discussed earlier, we use the MUSCLE-PARS guide-tree as the phylogeny estimate on its alignment. We computed the set of edges common to at least 95% of the best randomized MUSCLE-PARS2 guide-trees. This tree, which we call MPARS2-R95, contains all of the 95% supported edges on phylogenies from Probcons, Dialign, and MUSCLE-UPGMA alignments. For the other alignments, MPARS2-R95 contains 97% of ClustalW, 98% of MUSCLE, 98% of MUSCLE-PROG, 94% of MUSCLE-PARS1, and 89% of MUSCLE-PARS2 highly supported edges (i.e. over 95% support) in their respective phylogenies. Furthermore, MPARS-R95 contains several additional edges that are not among highly supported ones on any of the other alignments. This suggests that randomized MUSCLE-PARS is more informative than the deterministic counterpart. Furthermore, since randomized MUSCLE-PARS produces many parsimonious phylogeny-alignment pairs, they can be mined for various statistical measures.

Running time: Finally, in Table 3 we list the running time of computing each alignment, MP tree, and the total time of alignment plus phylogeny. Clearly Dialign and Probcons are much slower than the other methods shown here. The total running time of MUSCLE-PARS1 and MUSCLE-PARS2 is larger than MUSCLE-PROG and MUSCLE-UGMA. However, since we are taking the MUSCLE-PARS1 and MUSCLE-PARS2 guide-tree as the estimate of the phylogeny we can compare their alignment time to the total alignment plus phylogeny time of MUSCLE-PROG and MUSCLE-UPGMA---in which case they are comparable.

Table 1: Percentage of internal edges with bootstrap support above indicated threshold; bootstraps on guide-trees, when appropriate, are shown in parentheses.

Bootstrap support cutoff	Dialign	Prob cons	ClustalW	MUSCLE	MUSCLE-PROG	MUSCLE-UPGMA	MUSCLE-PARS1	MUSCLE-PARS2
50%	61	61	73 (63)	62	60 (35)	62 (39)	83 (83)	81 (81)
75%	46	45	54 (50)	44	46 (30)	49 (34)	66 (66)	68 (68)
95%	31	33	32 (32)	29	30 (24)	33 (27)	41 (41)	43 (43)

Table 2: Each entry in the table contains percentage of edges with at least 95% support in the MP phylogeny on the left alignment that are not present in the MP phylogeny listed at across the top. For MUSCLE-PAR1 and MUSCLE-PARS2 we use the best guide-tree since it is as well supported as the MP tree (see Table 1).

	Dialign	Prob Cons	ClustalW	MUSCLE	MUSCLE-PROG	MUSCLE-UPGMA	MUSCLE-PARS1	MUSCLE-PARS2
Dialign	---	9.4	15.6	21.9	15.6	15.6	10.9	7.8
Probcons	13.4	---	14.9	20.9	11.9	11.9	10.4	9.0
ClustalW	16.9	12.3	---	26.2	15.4	12.3	12.3	6.2
MUSCLE	15.3	10.2	18.6	---	13.6	10.2	6.8	6.8
MUSCLE-PROG	12.9	4.8	11.3	17.7	---	4.8	9.7	4.8
MUSCLE-UPGMA	20.6	13.2	16.2	22.1	13.2	---	8.8	4.4
MUSCLE-PARS1	32.9	29.4	32.9	35.3	34.1	27.1	---	10.6
MUSCLE-PARS2	33.7	31.5	31.5	38.2	33.7	27.0	14.6	---

Table 3: Running time for computing each alignment, tree, and total. For MUSCLE-PARS1 and MUSCLE-PARS2, the time at which the best phylogeny and alignment pair was found is shown in parentheses.

Running time (hours)	Dialign	Prob cons	ClustalW	MUSCLE	MUSCLE-PROG	MUSCLE-UPGMA	MUSCLE-PARS1	MUSCLE-PARS2
Alignment	43.30	26.30	1.80	0.30	0.01	0.03	0.31(0.18)	0.30(0.20)
Tree	0.34	0.36	0.21	0.37	0.34	0.30	0.21	0.17
Total	43.64	26.66	2.01	0.67	0.35	0.33	0.52(0.39)	0.47(0.37)

5.2 Simulated data. For each set of simulated unaligned sequences, we compute ClustalW, MUSCLE (all three stages), and MUSCLE-PARS (both variants) alignments. Subsequently, we construct MP trees, using the random restart TBR heuristic described earlier, on the ClustalW and MUSCLE alignments. However, in line with our observations on real data, we use the best guide-trees for MUSCLE-PARS1 and MUSCLE-PARS2 as estimates of the phylogeny. The accuracy of each phylogeny, computed using the RF distance, is compared against the true tree. In Tables 4a and 4b, we report the average error rate for each parametric setting. The improvement, in terms of percentage differences, is also provided for the best scoring alignment. We also report the improvement in MUSCLE-PARS1 and MUSCLE-PARS2 error rates over the best error rate of the other methods. While the average gain is modest, the overall results clearly indicate that improvement when using the two MUSCLE-PARS methods is a robust result.

General trends: Our results follow some of the general trends one would expect to see in simulation studies. For example, the error rates decrease as the sequence length increases. Conversely, error rates tend to increase as the evolutionary rates, number of taxa, deviations, or indel probabilities increases, all of which are known to make the phylogeny estimation problem harder. However, trees at evolutionary rates of 32 fare better than 16. Overall MUSCLE-PARS1 and MUSCLE-PARS2 have the lowest error rates. At sequence lengths of 1000 and low indel probability of 0.00005 the improvement using MUSCLE-PARS is the smallest (especially at 100 taxa), if any at all. We take a closer look at the part of the parameter space where improvement is over 1% in topological accuracy.

Parameters where improvement with MUSCLE-PARS is over 1%: There are 9 parametric settings at which MUSCLE-PARS has error rate lower than 1% than the other methods. Out of those 7 are for sequence lengths of 500. Thus MUSCLE-PARS can be most effective when sequence lengths are short relative to the number of sequences. On 6 of these settings the indel probability is 0.0005 (the higher value) thus showing that MUSCLE-PARS can be useful for data that has undergone a modest number of insertions and deletions. The largest improvement is of 2.2% for 200 sequences, 500 sequence length, 64 scaling, 4 deviation, and 0.0005 indel probability, which can be considered a hard setting.

SP vs. progressive alignment: A curious observation is that MUSCLE has high error rates, especially when considering high evolutionary rates and indel probabilities. In fact, the error rates sometimes go above 25%, which is much higher than that of the other methods. Even on real datasets, the bootstrap supports on MUSCLE alignments are generally lower than the other methods. Recall that MUSCLE computes a SP optimization in stage III after the progressive alignments are done. We conjecture that this significantly decreases the quality of the alignment for phylogeny reconstruction. However, for other tasks, such as aligning structurally conserved regions, it may be more appropriate as seen from performance on BALiBASE (Thompson et al. 1999) structural alignment benchmarks. When considering protein data, we have also noticed this anti-correlation between phylogeny reliability (using bootstraps) and BALiBASE accuracy.

These observations underscore the reality that no single assessment strategy can be considered perfect when evaluating alignments and phylogenies.

6. CONCLUSIONS AND FUTURE WORK

Our experiments on both simulated and real data show that MUSCLE-PARS1 and MUSCLE-PARS2 produce phylogenies of better accuracy and better support than phylogenies on Probcons and Dialign (real data only), ClustalW, MUSCLE-PROG, MUSCLE-UPGMA, and MUSCLE. Furthermore, MUSCLE-PARS is efficient in the running time required to produce an alignment and phylogeny, which means it can be used to analyze datasets containing even hundreds to thousands of sequences. We expect MUSCLE-PARS to quickly produce very good starting trees for expensive simultaneous alignment and phylogeny reconstruction local search strategies, such as those conducted in Poy and statistical alignment packages. MUSCLE-PARS can easily be implemented using existing easily available software packages with a simple Perl script. We are currently developing an open source MP solver so that MUSCLE-PARS can be distributed as an open source freely available package.

Many improvements are possible over this approach. For example, we could expect higher accuracy by iterating with ML trees; however, this would be at a cost of running time since constructing ML trees in between iterations is costlier than MP trees. For this study we used a deterministic version of MUSCLE-PARS implemented using PAUP*. Further gains in speed could be obtained with a faster implementation of TBR-based search heuristics for MP.

We have presented only preliminary but promising results on randomized MUSCLE-PARS. Randomized MUSCLE-PARS presents an alternative view of computing support for edges. Instead of bootstrapping, where the same alignment is resampled, one can sample phylogenies from different (but parsimonious) alignments. We plan to explore this in a future study.

We note that the gap penalties selected here may not be optimal for phylogeny reconstruction. Therefore, a study which examines this in detail for the methods used in this paper may yield worthwhile results. One simple and fast heuristic for “optimal” gap selection, which we plan to investigate, is to select the penalties (from a given set) which produce the most parsimonious alignment on the UPGMA guide-tree. We plan to examine this in a forthcoming study.

We also plan to conduct a rigorous examination of highly supported edges on the Metazoa 18s dataset across phylogenies from different alignments. Of particular interest are the additional highly supported edges in MUSCLE-PARS2 and randomized MUSCLE-PARS that are not present in other phylogenies.

Finally, a critical assessment of MUSCLE-PARS performance on protein data is currently underway. Several real protein datasets have been identified and aligned with all of the above methods (and more). In general, the trends outlined here are consistent with the protein results. Investigations that critically compare the observed anti-correlation between BALiBASE accuracy and bootstrap supports are planned.

7. REFERENCES

- Alschul,S.F., and Lipman D.J., (1989) Tree, stars, and multiple biological sequence alignment. Lecture notes in Computer Science, 49, 197-209.
- Do C.B., Mahabhashyam M.S.P., Brudno M., and Batzoglou S., (2005) PROBCONS: Probabilistic Consistency-base Multiple Sequence Alignment, *Genome Research*, 15, 330-340.
- Edgar,R.C., (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, 5(113)
- Effron,B., Halloran,E., and Holmes,S., (1996) Bootstrap confidence level for phylogenetic trees, *Proc. Natl. Acad. Sci. USA.*, 93, 13429-13434

- Eisen, J.A., (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis, *Genome Research*, 8(3), 163-167
- Felsenstein, J., (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368-376.
- Felsenstein, J., (1985) Confident limits on phylogenies: an approach using the bootstrap, *Evolution*, 39, 783-791
- Feng D-F. and Dolittle R.F., (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees, *Journal of Molecular Evolution*, 25, 351-360.
- Fleissner, R., Metzler D., and Haeseler A. (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction, *Systematic Biology*, 54(4), 548-561
- Foulds, L.R., and Graham, R.L., (1982) The Steiner problem in phylogeny is NP-complete, *Advances in Applied Mathematics*, 3, 43-49
- Goldman N., (1998) Effects of sequence alignment procedures on estimates of phylogeny. *Bioassays*, 20, 287-290
- Gotoh O., (1996) Significant improvement in accuracy of multiple protein sequence alignment by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.*, 264, 823-838.
- Gusfield D., (1997) Algorithms in strings, trees, and sequences, Cambridge University Press.
- Hall B.G., (2005) Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences, *Molecular Biology and Evolution*, 22, 792-802 .
- Hasegawa, M., Kishino, H., and Yano, T., (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, 22, 160-174
- La D., Sutch B., and Livesay D.R., (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins: structure, function and bioinformatics*, 58(2), 309-320.
- Morgenstern B., (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3), 211-218.
- Morrison D.A. and Ellis, J.T. (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. *Mol. Biol. Evol.* 14:428-441
- Newton, M.A., (1996) Bootstrapping phylogenies: large deviations and dispersion effects, *Biometrika*, 82, 315-328
- Robinson, D.F., and Foulds, L.R., (1981) Comparison of phylogenetic trees, *Mathematical Biosciences*, 53, 131-147
- Roshan U, Livesay D.R., and La D., (2005) Improved phylogenetic motif identification using parsimony. *BIBE05*, 19-26
- Sanderson M.J., (2004) r8s software package available from <http://ginger.ucdavis.edu/r8s/>.
- Sjolander K., (2004) Phylogenomic inference in protein molecular function: advances and challenges, *Bioinformatics*, 20(2), 170-179.
- Stoye J., Evers D., and Meyer F., (1998) Rose: generating sequence families, *Bioinformatics*, 14(2), 157-163.
- Redelings, B.D and Suchard, M.A. (2005) Joint Bayesian estimation of alignment and phylogeny, *Systematic Biology*
- Swofford, D.L., (1996) PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods), Sinauer Associates, Sunderland, Massachusetts, Ver 4.0
- Swofford, D.L., and Olsen, G.J., Phylogeny Reconstruction, In Hillis, D., Moritz, C., and Marble, B.K., editors, *Molecular Systematics*, chapter 11, pages 407-514, Sinauer Ass. Inc., Sunderland, Massachusetts, USA, 1996, 2nd edition
- Thompson J.D., Higgins D.G., and Gibson T.J., (1994) CLUSTALW: improving the sensitivity of multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673-4680.
- Thompson J.D., Plewniak, F., and Poch O., (1999) BALiBASE: a benchmark alignment database for evaluation of multiple alignment programs, *Bioinformatics*, 15, 87-88.
- Wheeler W., (1996) Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics*, 12, 1-9
- Wheeler W., (2002) POY: the optimization of alignment characters. Version 3.0.4. Program and Documentation, New York, NY, Available from <ftp://amnh.org/pub/molecular>.
- Zharikikh, A., and Li, W.H., (1995) Estimate of confidence in phylogeny: the complete and partial bootstrap technique, *Molecular Phylogenetics and Evolution*, 4, 44-63.

Table 4a: Summary of simulation results (continued on next page)¹.

Scale / Deviation	ClustalW	MUSCLE	MUSCLE-PROG	MUSCLE-UPGMA	MUSCLE-PARS1	MUSCLE-PARS2	Best Diff. ²
Percent error rates for 100 taxa, 500 sequence length, indel probability 5x10 ⁻⁵							
16 / 2	9.4	9.1	9.5	9.4	8.7 (0.1)	8.8	0.4
32 / 2	8.7	8.7	8.6	8.7	8.2 (0.3)	8.5	0.4
64 / 2	10.2	10.1	10.9	10.3	11.0	9.9 (0.2)	0.2
16 / 4	13.8	13.8	13.9	13.7	13.5 (tie)	13.5 (tie)	0.2
32 / 4	13.3	13.0 (tie)	13.1	13.2	13.0 (tie)	13.0 (tie)	---
64 / 4	14.1	14.3	15.0	14.0 (0.1)	15.5	14.7	-0.7
Percent error rates for 100 taxa, 1000 sequence length, indel probability 5x10 ⁻⁵							
16 / 2	6.1	6.0	5.9	5.8	5.4 (tie)	5.4 (tie)	0.4
32 / 2	4.8	5.0	4.9	4.8	4.6 (0.1)	4.7	0.2
64 / 2	7.0	6.9	7.1	6.5	6.4 (0.1)	6.9	0.1
16 / 4	9.1	9.1	8.9	9.0	8.9	8.8 (0.1)	0.1
32 / 4	8.9	8.5	8.7	8.4	8.0 (0.1)	8.1	0.4
64 / 4	13.4	11.5 (0.7)	13.8	12.2	14.0	12.5	-1.0
Percent error rates for 100 taxa, 500 sequence length, indel probability 5x10 ⁻⁴							
16 / 2	10.9	10.9	10.8	10.7	10.3 (tie)	10.3 (tie)	0.4
32 / 2	11.5	10.2	9.0 (0.2)	9.2	9.3	9.2	-0.2
64 / 2	16.6	25.3	19.7	17.6	17.3	16.4 (0.2)	0.2
16 / 4	13.9	13.9	13.7	13.5	12.7 (0.3)	13.0	0.8
32 / 4	17.5	16.4	14.9	14.6	14.4	13.7 (0.7)	0.9
64 / 4	24.4	30.6	24.2	23.3	22.9	22.6 (0.3)	0.7
Percent error rates for 100 taxa, 1000 sequence length, indel probability 5x10 ⁻⁴							
16 / 2	5.7	5.6	5.6	5.8	5.2 (0.3)	5.5	0.4
32 / 2	6.9	6.7	6.1	6.4	6.0	5.9 (0.1)	0.2
64 / 2	13.9 (0.4)	18.3	18.5	15.9	15.9	14.3	-0.4
16 / 4	8.8	9.0	8.5	8.5	8.3 (tie)	8.3 (tie)	0.2
32 / 4	13.4	12.3	10.6	10.5 (tie)	11.0	10.5 (tie)	---
64 / 4	23.4	26.7	23.7	21.7	23.1	20.6 (1.1)	1.1
Percent error rates for 200 taxa, 500 sequence length, indel probability 5x10 ⁻⁵							
16 / 2	11.1	11.2	11.2	11.3	10.5 (0.7)	10.5 (0.7)	0.7
32 / 2	8.3	8.2	7.9 (tie)	7.9 (tie)	7.9 (tie)	8.0	---
64 / 2	10.2	11.2	11.4	9.6 (0.4)	11.3	10.0	-0.4
16 / 4	15.3	15.5	15.4	15.5	13.8 (0.2)	14.0	1.5
32 / 4	11.5	11.3	11.4	11.4	11.3	11.2 (0.1)	0.1
64 / 4	17.0	16.5	17.3	15.4 (0.3)	17.0	15.7	-0.3
Percent error rates for 200 taxa, 1000 sequence length, indel probability 5x10 ⁻⁵							
16 / 2	6.2	6.3	6.3	6.3	5.7 (tie)	5.7 (tie)	0.6
32 / 2	5.6	5.6	5.5	5.5	5.4 (tie)	5.4 (tie)	0.1
64 / 2	7.2	7.7	8.2	6.9 (tie)	8.2	6.9 (tie)	---
16 / 4	9.4 (tie)	9.5	9.4 (tie)	9.4 (tie)	9.5	9.5	-0.1
32 / 4	9.0	8.9	8.8 (tie)	8.8 (tie)	8.9	8.8 (tie)	---
64 / 4	14.4	13.6	14.4	12.8	14.2	12.7 (0.1)	0.1
Percent error rates for 200 taxa, 500 sequence length, indel probability 5x10 ⁻⁴							
16 / 2	11.9	11.7	11.2	11.2	10.2	9.7 (0.5)	1.5
32 / 2	12.5	14.7	10.3	10.0	10.0	9.5 (0.5)	0.5
64 / 2	19.0 (0.4)	37.4	22.2	20.7	19.9	19.4	-0.4
16 / 4	16.1	16.4	15.3	15.3	14.4	14.2 (0.2)	1.1
32 / 4	17.0	19.6	15.6	15.4	14.6	14.5 (0.1)	0.9
64 / 4	26.6	44.0	26.6	26.1	25.6	23.9 (1.7)	2.2

Table 4b: Summary of simulation results (continued from previous page)¹.

Scale / Deviation	ClustalW	MUSCLE	MUSCLE-PROG	MUSCLE-UPGMA	MUSCLE-PARS1	MUSCLE-PARS2	Best Diff. ²
Percent error rates for 200 taxa, 1000 sequence length, indel probability 5×10^{-4}							
16 / 2	7.2	7.6	7.1	6.9	6.6	6.4 (0.2)	0.5
32 / 2	9.5	10.4	6.8	6.8	6.6 (tie)	6.6 (tie)	0.2
64 / 2	15.8 (0.9)	28.4	19.8	18.4	17.6	16.7	-0.9
16 / 4	11.1	11.4	10.1	10.1	9.5 (0.2)	9.7	0.6
32 / 4	14.4	16.0	11.9	11.8	11.2 (0.1)	11.3	0.6
64 / 4	23.7	36.0	24.6	22.9	22.6	21.5 (1.1)	1.4
Percent error rates for 400 taxa, 500 sequence length, indel probability 5×10^{-5}							
16/2	12.6	12.6	12.6	12.6	11.5 (0.1)	11.6	1.1
32/2	8.7	8.6	8.6	8.6	8.3	8.1 (0.2)	0.5
64/2	9.0	10.1	9.6	8.6	9.0	8.3 (0.3)	0.3
16/4	17.8	17.9	17.9	17.9	16.2 (0.2)	16.4	1.6
32/4	13.3	13.3	13.2	13.2	12.8 (0.1)	12.9	0.4
64/4	15.1	15.7	14.7	13.9	14.5	13.5 (0.4)	0.4
Percent error rates for 400 taxa, 1000 sequence length, indel probability 5×10^{-5}							
16 / 2	7.4	7.3	7.4	7.3	7.0 (tie)	7.0 (tie)	0.3
32 / 2	5.5 (tie)	5.6	5.5 (tie)	5.5 (tie)	5.5 (tie)	5.5 (tie)	---
64 / 2	6.5	7.1	6.8	6.0 (0.1)	6.4	6.1	-0.1
16 / 4	10.3	10.3	10.3	10.3	9.6 (tie)	9.6 (tie)	0.7
32 / 4	8.8	8.9	8.5 (tie)	8.7	8.5 (tie)	8.5 (tie)	---
64 / 4	12.2	11.9	11.5	10.9 (0.1)	11.9	11.0	-0.1
Percent error rates for 400 taxa, 500 sequence length, indel probability 5×10^{-4}							
16 / 2	13.1	14.5	12.8	12.7	12.0 (tie)	12.0 (tie)	0.7
32 / 2	11.8	16.3	10.0	9.8	9.4	9.3 (0.1)	0.5
64 / 2	15.9	40.1	17.9	16.6	15.5	15.3 (0.2)	0.6
16 / 4	18.2	19.7	17.6	17.6	15.9 (tie)	15.9 (tie)	1.7
32 / 4	17.0	21.2	15.4	15.6	14.5 (0.1)	14.6	0.9
64 / 4	22.8	44.5	22.9	21.8	22.4	21.5 (0.3)	0.3
Percent error rates for 400 taxa, 1000 sequence length, indel probability 5×10^{-4}							
16 / 2	8.0	9.3	7.6	7.6	7.2 (tie)	7.2 (tie)	0.4
32 / 2	8.2	10.0	6.6	6.4	6.2 (0.1)	6.3	0.2
64 / 2	12.3 (0.6)	33.3	15.0	14.5	13.5	12.9	-0.6
16 / 4	11.6	13.4	11.0	11.0	10.3 (0.1)	10.4	0.7
32 / 4	12.9	15.1	10.6	10.6	10.2 (0.1)	10.3	0.4
64 / 4	20.4	39.5	19.8	19.0	18.6	18.2 (0.4)	0.8
Overall results: number of times each method was best (ties are counted in each occurrence)							
Dev. = 2	5	0	3	5	16	20	---
Dev. = 4	1	2	3	6	16	21	---
Total	6	2	6	11	32	41	---

¹ Best scoring alignments (across all six possibilities) also included the percent difference between it and the next best scoring alignment (again, across all six possibilities) in parentheses.

² In the final column, the difference between the best scoring MUSCLE-PARS alignment and the best of the remaining four alignments is presented.