The Impact of Multiple Protein Sequence Alignment on Phylogenetic Estimation

Li-San Wang, Jim Leebens-Mack, P. Kerr Wall, Kevin Beckmann, Claude W. dePamphilis, and Tandy Warnow

Abstract—Multiple sequence alignment is typically the first step in estimating phylogenetic trees, with the assumption being that as alignments improve, so will phylogenetic reconstructions. Over the last decade or so, new multiple sequence alignment methods have been developed to improve comparative analyses of protein structure, but these new methods have not been typically used in phylogenetic analyses. In this paper, we report on a simulation study that we performed to evaluate the consequences of using these new multiple sequence alignment methods in terms of the resultant phylogenetic reconstruction. We find that while alignment accuracy is positively correlated with phylogenetic accuracy, the amount of improvement in phylogenetic estimation that results from an improved alignment can range from quite small to substantial. We observe that phylogenetic accuracy is most highly correlated with alignment accuracy when sequences are most difficult to align, and that variation in alignment accuracy can have little impact on phylogenetic accuracy when alignment error rates are generally low. We discuss these observations and implications for future work.

Index Terms—Simulation, biology and genetics, multiple protein sequence alignment, phylogeny reconstruction.

1 Introduction

Multiple sequence alignment (MSA) is an important computational problem that is fundamental to all sequence-based comparative analyses. In particular, applications of alignment estimation to problems in protein sequence analysis (structure, function, and subfamily identification) have led to many new protein alignment methods. Structural alignment databases, such as STAMP [1], SCOP [2], [3], HOMSTRAND [4], BaliBASE [5], [6], [7], and PREFAB [8], have been compiled to serve as benchmarks for developing and refining MSA methods. Reference alignments are also commonly derived through simulations of sequence evolution [9], [10], [11], [12]. Comparative studies based upon these benchmarks and simulations have concluded that many of the newer MSA methods, especially those designed for protein alignment (MAFFT [13], ProbCons [14], T-Coffee [15], Di-Align [16], Opal [17], Prank [18], AMAP [19], ProbAlign [20], and others), are substantially better than earlier ones including Clustal [21], in that they achieve higher alignment accuracy scores on these benchmarks. Among these newer methods, MAFFT and ProbCons are consistently among the best performing [22], [23], [24], [25].

These studies have implications for phylogenetic analyses, as the common wisdom is that alignment strategy can impact phylogeny estimation. For example, Morrison and Ellis [26] found that the choice of multiple sequence alignment method had a greater impact on the resultant phylogeny than the choice of phylogeny estimation method, and Wong et al. [27] showed that with high sequence divergence, different alignment strategies can produce alignments that yield conflicting gene trees.

Studies of the impact of alignment estimation on phylogeny estimation have also been performed using simulations of sequence evolution that include insertions, deletions (jointly referred to as "indels") as well as substitutions [20], [28], [29], [30]. These studies have generally (but not always) shown that alignment estimation can have an impact on phylogeny estimation, and have suggested that under some circumstances, alignment accuracy correlates with phylogenetic accuracy. Unfortunately, each of these studies evaluated a limited set of MSA methods, phylogeny reconstruction methods, and models of sequence evolution. For example, Roshan and Livesay [20], [31] aligned simulated DNA sequences using six MSA methods (ClustalW [21], MUSCLE [8], and four methods developed the authors that use MUSCLE in different ways) and performed maximum parsimony analsyes. Their main objective was to see if their new MSA methods provided an improvement in phylogenetic estimation when followed by a maximum parsimony analysis. Their simulation study (performed on large model trees containing from 100 to 400 taxa) showed that their new methods provided a very modest improvement (about one percent) in phylogenetic accuracy, as measured using the Robinson-Foulds score [32], over the best of the other alignment methods. They also observed that

- *L.-S. Wang is with the Department of Pathology and Laboratory Medicine and Penn Center for Bioinformatics, 1424 Blockley Hall, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA 19104. E-mail: lswang@mail.med.upenn.edu.*
- *J. Leebens-Mack is with the Department of Plant Biology, University of Georgia, Athens, GA 30602. E-mail: jleebensmack@plantbio.uga.edu.*
- *P.K. Wall is with BASF Plant Science, L.L.C., 26 Davis Drive, Research Triangle Park, NC 27709. E-mail: kerr.wall@basf.com.*
- *K. Beckmann and C.W. dePamphilis are with the Department of Biology, the Huck Institutes of the Life Sciences, and the Institute of Molecular Evolutionary Genetics, the Pennsylvania State University. E-mail: phage9@gmail.com, cwd3@psu.edu.*
- *T. Warnow is with the Department of Computer Science, One University Station, STOP C0500, the University of Texas at Austin, Austin, TX 78712. E-mail: tandy@cs.utexas.edu.*

Manuscript received 29 Oct. 2008; revised 3 Mar. 2009; accepted 30 July 2009; published online 1 Sept. 2009.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-10-0186. Digital Object Identifier no. 10.1109/TCBB.2009.68.

1545-5963/11/$26.00 © 2011 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

maximum parsimony trees computed on ClustalW alignments were generally as good as (and sometimes better than) maximum parsimony trees computed on Muscle alignments. Finally, a comparison of the entire set of parsimony trees (each based upon a different MSA method) showed that the choice of MSA method does impact the topological accuracy of phylogenetic trees. However, there are several limitations to this study: the focus on maximum parsimony for phylogeny estimation, the failure to include MAFFT or one of the other recently developed methods which has been shown to outperform ClustalW, and the lack of information on alignment error rates. Therefore, the results of this study do not provide a comprehensive understanding of the impact of alignment error on phylogeny estimation.

Ogden and Rosenberg [30] examined the impact of alignment error on phylogeny reconstruction. They performed a DNA sequence simulation study and examined the performance of phylogeny reconstruction methods applied to ClustalW alignments on small (16 taxon) trees. They explored all the most commonly used phylogeny reconstruction methods including neighbor joining (NJ), maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference (BI), thus enabling them to make inferences about the impact of alignment error for many phylogeny reconstruction methods. They concluded that statistical methods (ML and BI) provide the best estimations of phylogenies, but their main focus was on two other issues: the impact of model tree shape on alignment error, and the impact of alignment error on phylogeny reconstruction accuracy. Comparisons revealed that pectinate trees produced the largest alignment error compared to other shapes, and that NJ was impacted the most by alignment error, whereas MP was least sensitive to alignment error, and ML and BI were intermediate in this regard. These are interesting results, but general inferences are significantly limited by the restriction to ClustalW alignments and the small (16 taxon) model trees.

Hall [29] performed a simulation study to look at the impact of MSA choice on phylogeny reconstruction, but had a different objective than the Ogden and Rosenberg and the Roshan and Livesay studies. Rather than comparing standard MSA methods, Hall focused on how the choice of data (protein or DNA coding sequence) as well as alignment method impacted accuracy. Hall used his own software to simulate the evolution of a coding sequence, aligned the sequences (in some cases first translating the coding sequences into amino acid sequences), and then constructed phylogenies. This study used all the standard phylogeny reconstruction methods, so it provides a comparison between phylogeny reconstruction methods for various fixed MSA methods. However, only a few MSA methods were compared: Clustal-X with parameters set for codons or protein sequences, and routines for mapping nucleotides onto amino acid sequences. The study concluded that statistical methods for phylogeny estimation based upon DNA aligned using codon-models outperformed other two-phase approaches. A careful look at the data shows that the choice of MSA approach definitely impacts the phylogeny estimation in some cases, but not all. The cases where the MSA method had the most impact were those where the phylogenetic accuracy was quite poor for one or more alignments. However, Hall only analyzed small data sets (up to 16 taxa), and provided no analysis of overall alignment accuracy (although alignment "quality" scores, calculated by Clustal-X, were reported).

The final simulation study relevant to this work was performed by Cantarel et al. [28]. Protein sequence evolution were simulated on 16 taxon model trees in order to understand the impact of model condition upon alignment and phylogeny accuracy. The main observation from this study is that the accuracy of sequence alignments (as estimated by T-Coffee) and trees (estimated in several ways) both degrade with evolutionary distance. However, true alignments produced better trees only than estimated alignments only for the hardest model conditions (many substitutions and indels on each edge), and estimated alignments sometimes yielded better trees for the easier model conditions.

In summary, these studies suggest that improvements in phylogenetic accuracy can be obtained through improving alignments, but sometimes the improvement is small, and sometimes there is no improvement at all. Even if we believe improving the alignment should help phylogeny estimation, we cannot predict how much of a difference it will make. Furthermore, these studies were relatively narrow in scope: with the exception of the study by Roshan and Livesay, all looked only at small model trees, and most did not consider the top performing MSA methods, such as MAFFT or ProbCons. Therefore, the impact of the newer MSA methods on phylogeny estimation is still unclear.

This paper will address several questions related to the impact of these newer MSA methods on phylogeny estimation, including:

1. How do the different MSA methods influence the accuracy of the phylogenies estimated on inferred alignments? Does an improvement in alignment accuracy yield an improvement in phylogenetic accuracy? If so, how strongly are these measures correlated, and does the impact depend upon specific model parameters?
2. Do different phylogeny reconstruction methods respond differently to errors in multiple sequence alignments? Also, which phylogeny reconstruction methods produce the most accurate phylogenies, given highly accurate alignments?

We will also seek to make informed recommendations about the best MSA and phylogeny estimation procedures when working with amino acid sequences and the objective is accuracy in the reconstructed phylogeny.

2 Methods

2.1 Overview

We used simulated data in this paper instead of real sequences for several reasons. First, we wanted to be able to quantify error in phylogeny estimations, and while some biological data sets have highly reliable curated alignments, there are no comparably reliable phylogenies for these data sets (in particular, species trees may differ from the gene trees, and estimated gene trees may differ from true gene

trees, even if based upon the true alignment). A second reason is that the use of simulated data allows us to explore a wide range of data set conditions, whereas biological data sets for which curated alignments are available are much more limited, and the alignments are often short.

Amino acid sequence evolution was simulated so we could include ProbCons (one of the two "best" MSA methods), which had only been developed for amino acid alignment. Our simulations included amino acid substitutions and indels evolving at various rates on model trees ranging from small (approximately 20-30 taxa) to large (100 taxa). We aligned sequences using eight alignment methods, and evaluated accuracy by comparing the alignments to the true alignment; trees were constructed on each alignment (including the true alignment) using six phylogeny reconstruction methods, and accuracy evaluated by comparing the estimated trees to the true tree. The rest of this section provides a brief summary of the simulation experiments. For the complete description of simulation model conditions, please see Online Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68.

2.2 Sequence Alignment Methods

The MSA methods we explored included new MSA methods that have been shown to produce the best alignments on the benchmark data sets, as well as the standard methods used by the phylogenetic research community. Thus, we included MAFFT (version 5.861 [13]), PROBCONS (version 1.1 [14]), MUSCLE (version 3.6 [8]), DIALIGN (version 2.2 [16]), POA (version 2 [33]), T-COFFEE (version 4.45 [15]), ClustalW (version 1.83 [21]), and DBClustal (version 1 [34]). The settings used for each MSA program are given in the Online Appendix B, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB. 2009.68.

2.3 ROSE Simulations of Protein Sequence Evolution

We simulated amino acid sequence evolution on model trees using a modified version of ROSE [11] (see Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB. 2009.68, for our modifications), which allows sequences to evolve under fairly general site substitution models, as well as with insertions and deletions of strings of amino acids. The input to ROSE is a model tree with branch lengths, and the simulation settings including sequence length, marginal distribution of amino acid frequencies at the root of the model tree, substitution model, the insertion/deletion (indel) length distribution, and the indel rates relative to the point substitution events. We chose ROSE as the sequence simulator for two reasons: it is open source (which allowed us to modify the code to output actual internal node sequences, and figure out details on the Markovian model for amino acid sequence evolution implemented in ROSE), and it is quite flexible with many adjustable parameters.

We modified ROSE to report the true multiple sequence alignment as well as the full indel history—that is, not only the sequence at every node (internal as well as leaf) of the model tree, but also the true pairwise alignment on every edge. This output also allows us to identify the edges of the model tree that lacked substitutions or indels, which are then collapsed in order to produce the "true tree." Note although the model tree is always bifurcating, that the realized true tree will not be bifurcating when there are zero edge lengths. We compute topological error rates for estimated trees by comparing them to the realized true trees.

We performed two experiments using the modified ROSE code. The first simulated sequence evolution on model trees based upon some BaliBASE2 (http://bips.ustrasbg.fr/fr/Products/Databases/BAliBASE2/) data sets with 19 to 28 protein sequences, with parameter settings based upon these data sets. For the second experiment, we examined larger model trees, from 25 to 100 proteins, and with higher rates of substitutions and indels. Some parameters of the simulation process were held constant (see Online Appendix C, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68) across all the experiments: we used the Dayhoff 250 PAM matrix [35] as the point substitution probability matrix, and we used one gap length distribution for all the experiments. This gap length distribution was based upon the average of the gap length distributions of the BaliBASE data sets we used for our first experiment, which had an average gap length of approximately 3.4 residues. Except for these two parameters (substitution matrix and gap length distribution), no other parameters were held constant in our experiments.

2.4 Phylogenetic Reconstruction Methods

We studied six phylogeny reconstruction methods: three variants of maximum parsimony implemented within PAUP* [36], two variants of neighbor joining also implemented within PAUP*, and RAxML [37], a fast method for estimating maximum likelihood trees. The commands we used for each of these methods are given in the online supplemental Appendix D, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68.

For NJ analyses, the substitution model was based on the PAM 250 matrix [35] and we used Protdist in PHYLIP [38] to compute the distance matrix. Distances were estimated for the NJ analyses using both pairwise and complete deletion of gapped columns [39]. NJ trees were then calculated using PAUP*. For RAxML analyses, we invoked the PAM 250 substitution model by specifying the PROTCATDAYHOFF option, and we ran RAxML in its default setting.

We performed several different maximum parsimony analyses, varying how we treated gaps, how we weighted the substitutions, and the heuristic we used to search for MP trees. For the gap treatment, we did an initial exploration on the BaliBASE model trees, and observed that treating the gaps as missing data produced slightly better results than treating the gaps as an extra state (data not shown). We explored three ways of weighting the substitutions: one in which all substitutions have the same cost (i.e., unweighted), one in which the amino acid substitution cost is based upon the minimum number of nucleotide substitutions needed (i.e., substitution-count

TABLE 1
Simulation Settings for Experiment 1

Parameter	Value
Insertion/deletion length distribution [a]	Between 1 and 10 residues, with the following probability distribution: [.265,.189,.145,.114,.089,.069,.052,.038,.025,.014]
Number of replicates per setting	80
Sequence type	Protein
Substitution model	Dayhoff (PAM) matrix; the distribution of the sequence at the root is the equilibrium distribution of the substitution matrix; also see Table M6 for the settings for the ROSE simulator
Model tree	Binary trees estimated from 12 BAliBASE datasets using maximum likelihood; see Appendix G.

[a] *The insertion and deletion lengths are based on the original alignments from BAliBASE. We used the lambda.pl script distributed with the DAWG sequence simulation software (Cartwright 2005).*

weighted), and one using the PAM250 substitution matrix. Finally, we used two different heuristics for the parsimony analyses, depending upon the data set size.

2.5 Performance Assessment

For each alignment and phylogenetic tree constructed, we recorded the alignment error and phylogenetic tree error as follows: We used the Lobster package [8] to score each alignment we computed using three accuracy measures: column accuracy (TC, i.e., the percentage of correctly aligned columns of residues in the inferred alignment), sum-of-pairs (SP accuracy, i.e., the percentage of correctly aligned pairs of residues in the inferred alignment), and Cline Shift (CS, which reflects the total of the shift lengths between the sequences [40]). The alignment error rates are calculated by subtracting the accuracy measure from 100 percent. We scored each phylogenetic reconstruction using the False Negative (FN) rate, also known as the "Missing Edge Rate," which is the percentage of the edges in the true tree that are missing in the reconstructed tree. We use this measure instead of the normalized Robinson-Foulds (RF) distance [32], since the Robinson-Foulds distance is inappropriate for evaluating estimated trees when trees may not be binary [41]. However, when the estimated trees are binary, then the FN rate and the normalized RF distance are the same. (A summary of the measurements is given in Online Appendix E, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68, which also shows the False Positive (FP) rate, which is the percentage of the estimated tree's edges that do not appear in the true tree.)

2.5.1 Experiment 1: Simulation Study Using Model Trees Based on a BAliBASE Subset

In the first experiment, we explored performance on model trees defined so as to reproduce the properties of selected BaliBASE data sets. For this experiment, we simulated alignments based on 12 BAliBASE2, Reference 3 data sets, which included 19 to 28 sequences per alignment (http://bips.ustrasbg.fr/en/Products/Databases/BAliBASE2/align_index.html#ref3). For each data set, we first estimated an evolutionary tree using a maximum likelihood phylogeny estimator, PhyML [42], on the reference alignment for the data set. This produced a tree topology and branch lengths specific to the model that were used in subsequent simulations. See Table 1 for a summary of simulation settings. The estimated trees are included in Online Appendix F, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68.

For each model tree, we set the simulation parameters to produce sequences having the same empirical properties as we estimated for the complete structural alignments for the reference data set for that model tree. We computed the indel rate and gap length distribution for each model using the lambda.pl script distributed with the DAWG sequence simulation software [9]. These parameters were used as run parameters for the modified ROSE software, and the model tree diameters were scaled so the sequences having the same average p-distance as the empirical data sets (Appendix H, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68). These model trees produced data sets with varying properties, so that the maximum p-distances varied from 0.58 to 0.80, and the percentage of the true alignment matrix that was gapped varied from 11 to 35 percent. For each of the 12 model trees, we generated 80 data sets, and compared performance of the eight MSA methods and six phylogeny reconstruction methods described above. The parameter estimation procedure is described in detail in the Online Appendix G, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68.

2.5.2 Experiment 2: Varying Taxon Number and Alignment Lengths

In the second experiment, we examined the performance of two-phase methods on larger model trees having from 25 to 100 sequences, and with a wider range of evolutionary rates, thus producing a wider range of maximum p-distances and gappiness. For this experiment, we examined three representative MSA methods (ClustalW, MAFFT, and POA) and three phylogeny reconstruction methods (weighted MP using the PAM250 transition matrix, NJ with pairwise deletion, and RAxML). MAFFT was included since it is one of the top two MSA methods (the other being ProbCons). ClustalW was selected since it is the most commonly used MSA method in phylogenetic studies, and POA was selected since it was the worst performing in Experiment 1.

We generated pure-birth model trees using r8s [43] with 25, 50, and 100 taxa. The branch lengths of each model tree were modified so as to deviate (slightly) from the strong molecular clock as follows: for a given edge in the tree, we

TABLE 2
Simulation Settings for Experiment 2

Parameter	Value
Insertion/deletion length distribution	Between 1 and 10 residues, distributed as follows: [.265,.189,.145,.114,.089,.069,.052,.038,.025,.014] (see footnote in Table 1 (a))
Relative rates for insertion/deletion events	0.0025 or 0.0100 per point mutation
No. repeats per setting	30
Sequence type	Protein, sequence length at root = 150, 300
Substitution model	Dayhoff (PAM) matrix; the distribution of the sequence at the root is the equilibrium distribution of the substitution matrix; also see Table M6 for the settings for the ROSE simulator
Model tree	**Tree topology and branch lengths:** 1. We first generate an ultrametric random-birth tree (Yule-Harding) using r8s (Sanderson 2006). 2. We then deviate the tree from ultrametricity; this is achieved through multiplying each branch by a random number X where ln X is distributed uniformly between [-ln 1.5, ln 1.5]. Then all branchs of the tree are scaled uniformly by the same scaling constant so the tree diameter equals the given diameter setting (see below). The diameter is the maximum expected number of changes per site between any pair of leaves in the model tree. **Number of taxa:** 25, 50, 100. **Diameter of the model tree:** 1, 2, 3, 4, 6 (number of mutations per site, as input to ROSE).

selected a random number x from $[-\ln 1.5, \ln 1.5]$, and multiplied the branch length by exp(x). We then rescaled the tree to produce the desired maximum evolutionary distance (also called the "tree diameter") of 1, 2, 3, 4, or 6 expected substitutions per site. We then used ROSE to simulate evolution with two different initial sequence lengths (150 or 300), two different indel rates (0.0025 and 0.01), and with the same gap length distribution as we used in the first experiment; please see Table 2. In total, we had 60 model conditions (five evolutionary diameters, two indel rates, two sequence lengths, and three numbers of taxa), and we generated 50 data sets for each model condition. The data sets we generated had maximum p-distances that ranged from 59 to 97 percent and gappiness, measured as the proportion of cells in the alignment matrix, ranged from 4 to 54 percent (Appendix I, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68). Each data set was aligned using the three MSA methods listed above (MAFFT, ClustalW, and POA) and the true alignment. Trees were constructed using three phylogeny reconstruction methods (PAMwMP, NJ-pw, and RAxML) on each of the alignments.

We have made the simulated trees and alignments in Experiments 1 and 2 available online at http://people.pcbi.upenn.edu/~lswang/alignexp/.

3 Results

These two experiments explored different parts of model tree space, with Experiment 1 focusing on small model trees based upon BaliBASE data sets, and Experiment 2 exploring larger trees (25-100 taxa), with a greater range of substitution rates and indel rates. In the first experiment, we examined the full set of multiple sequence alignment tools, but in the second experiment, we limited our attention to a subset of these tools.

3.1 Experiment 1: Analysis of BALIBASE Model Trees

This experiment involved simulations on 12 different model trees, each based upon one of the 12 BaliBASE data sets (see the Methods section). While absolute performance varied between the different model conditions, the relative performance was quite consistent. These trends are evident in the results for three of the 12 model trees (Fig. 1). The results on all 12 model trees can be seen in the companion Website (http://people.pcbi.upenn.edu/~lswang/alignexp/).

The results of these analyses revealed the best performing approach for each phylogeny reconstruction method, in terms of treatment of gaps, pruning data sets, etc. We observed that pruning data sets to eliminate sites with more than 50 percent gapped positions did not improve the accuracy of the phylogenies constructed using any of the phylogeny reconstruction methods (data not shown). For neighbor joining, we observed that computing the distance matrix using pairwise deletion yielded slightly better results than computing the distance matrix using complete deletion of all gapped columns.

The study also revealed differences in performance between different phylogeny reconstruction methods. PAM-weighted maximum parsimony performed better in our experiments than unweighted MP and weighting by the number of nucleotide substitutions required to transition from one amino acid to another. RAxML produced the most accurate phylogenies on the true alignments and the greatest variance in phylogenetic accuracy among alignments (Figs. 1c, 1d, and 1e).

We observed that while MSA methods varied with respect to alignment accuracy whether measured using SP, TC, or Cline Shift scores on these data sets, the relative performance was largely consistent among treatments (Fig. 1), and indicated three clearly defined levels of performance among the MSA programs. The top group consisted of MAFFT, ProbCons, T-Coffee, and Muscle, in that order, though MAFFT and ProbCons were very close in performance. The next group consisted of DBClustal, ClustalW, and then Di-Align, also roughly in that order. Finally, POA performed the worst.

We then turned to considering the relative sensitivity to alignment error for each of the phylogeny estimation methods. We observed that (Figs. 1c, 1d, and 1e) of all the phylogeny reconstruction methods, NJ seems to vary the least between different alignments, and ML the most. The most interesting observation, however, was that although the improvement in phylogenetic accuracy is

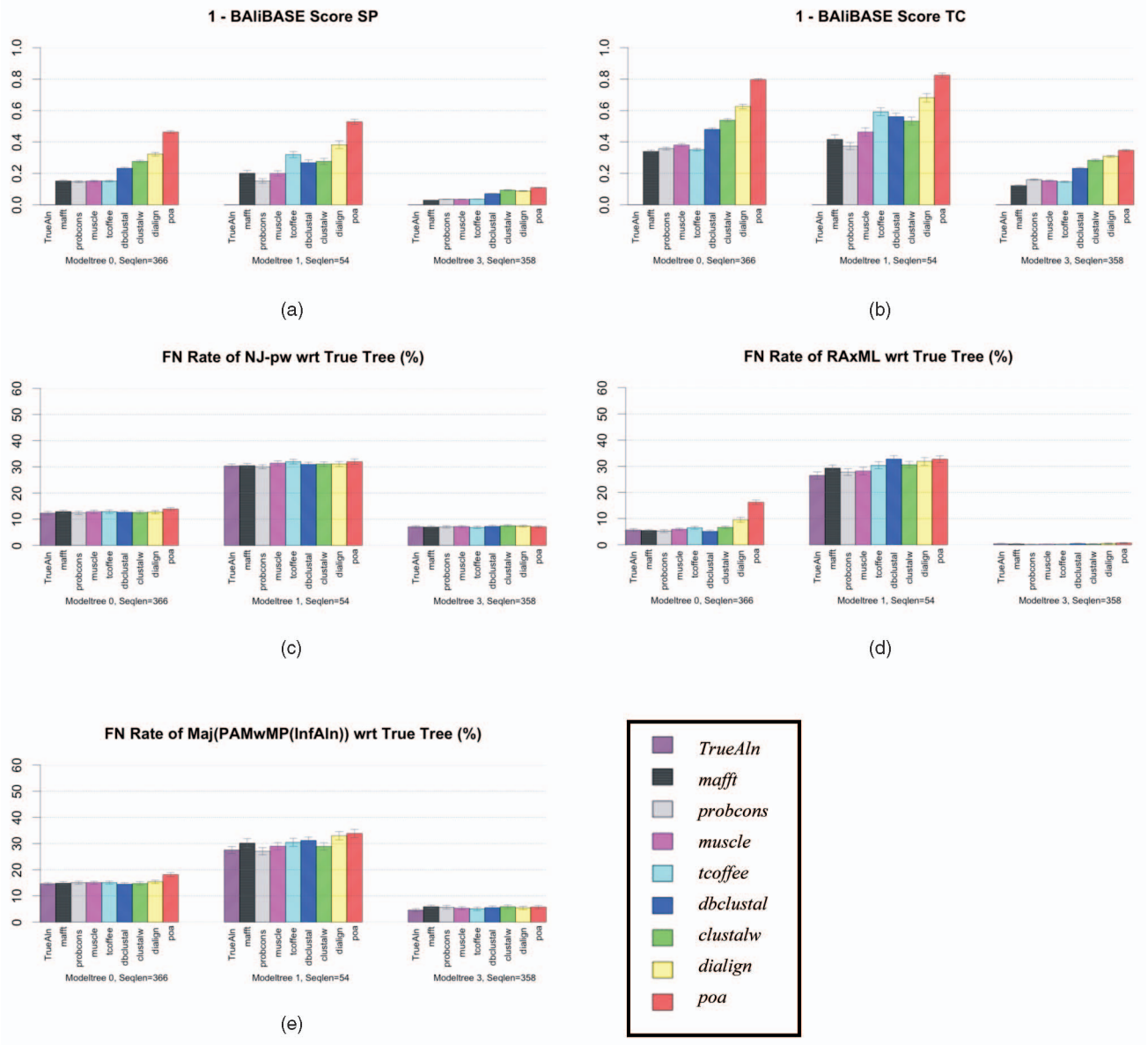

Fig. 1. Performance of alignment methods under the BAliBASE simulation study, using model trees 0, 1, and 3 (see Methods for the settings). (a) 1-BAliBASE SP score; (b) 1-BAliBASE TC score; (c) FN rate (missing edge rate) of neighbor joining with respect to the true tree (gap removal based on pairwise sequence comparison); (d) FN rate (missing edge rate) of RAxML with respect to the true tree; (e) FN rate (missing edge rate) of the majority consensus of maximum parsimony with respect to the true tree (using PAM-weighted parsimony scoring matrix).

statistically significant, the amount of improvement depends upon the model condition. For example, in model tree 3, alignments vary in their 1-SP error rates (from 3 to 11 percent), but in that model all the ML trees have less than one percent missing edge error rates. In contrast, model trees 0 and 1 produced more variable alignments (from 15 to 46 percent for the 1-SP error in model tree 0), and RAxML trees also vary substantially in their missing edge rates (from 5 to 17 percent). On the other hand, if we consider only MAFFT, ProbCons, Muscle, T-Coffee, DBClustal, and ClustalW alignments, ignoring the significantly worse DIALIGN and POA alignments, the picture becomes a bit less clear. Even, here, the alignments vary in their 1-SP error rates for these three model conditions (from 15 to 38 percent for model trees 0 and 1, and from 3 to 9 percent for Model Tree 3), but trees based upon these "good" alignments are much closer in accuracy. Indeed, on two of the three model conditions (model trees 0 and 3), there seems to be no real difference between the RAxML trees based on these good alignments; only on Model Tree 1 do we see any appreciable difference. And, interestingly, it is only on Model Tree 1 that tree errors are high, ranging from 28 (for trees estimated on the true alignment) to 32 percent (for trees estimated on T-Coffee and DBClustal). Otherwise, although there are detectable differences in tree error rates, they are somewhat small.

We analyzed the 12 real BAliBASE alignments using the same experimental procedure (Appendix N, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68). We find that the ranking of alignment methods using the various measures of accuracy (including alignment and phylogenetic accuracy) is roughly the same here as in the simulated data.

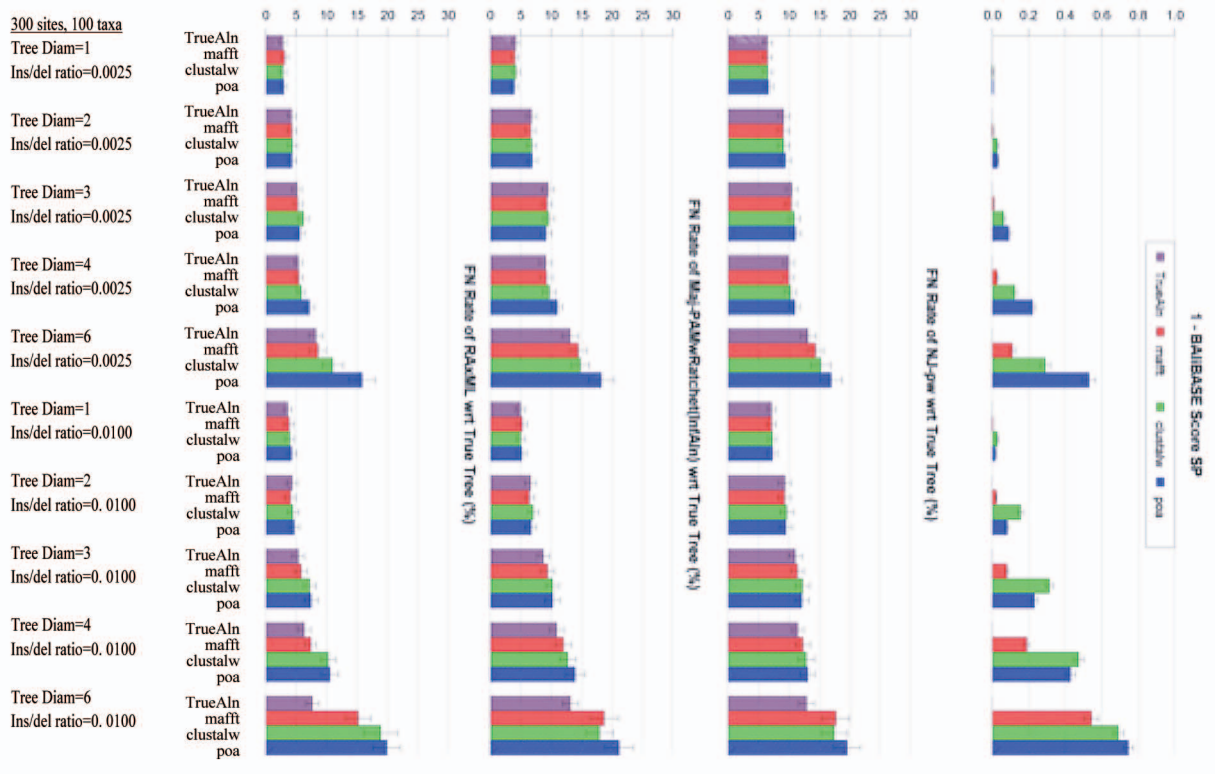

Fig. 2. Performance of alignment methods in the random model tree simulation study using 100 taxa and 300 amino acids. See the table of coefficients at http://people.pcbi.upenn.edu/~lswang/alignexp/correlation.xls for the full set of correlations, and Appendix J, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68, for the complete results using 25 taxa and/or 150 amino acids.

The conclusions we can draw from this experiment are limited, but intriguing. We see that while substantial differences in phylogenetic accuracy can result from changing from a very poor alignment (e.g., some POA alignments) to a very good (e.g., MAFFT or ProbCons) alignment, trees based upon two "good" alignments (one of the top four) can often be of comparable accuracy, even if the alignments have quite different levels of accuracy.

3.2 Experiment 2: Analysis of Birth-Death Trees

In this simulation study, we focused on parametric model tree generation to explore the interaction between various model parameters (including model tree diameter, insertion-deletion rate, sequence length, and the number of taxa), and their effects on the accuracies of alignment and phylogeny reconstruction (see the Methods section for details). The data sets generated for this experiment had a wider range of number of taxa, p-distances, and gappiness, so that some of the data sets were very close to saturated and many were very gappy (some even had almost 60 percent of the true matrix gapped, see Appendix I, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68). Thus, this experiment produced a range of data sets, some of which were harder to align than the data sets produced in the first experiment. This experiment thus showed, unsurprisingly, that the hardest model conditions for alignment estimation are those with many indels and substitutions, and that trees estimated on these alignments are also the poorest.

The results of this simulation study on random model trees exhibited several clear trends regarding the relative performance of multiple sequence alignment methods and phylogeny estimation methods. The relative performance between MSA methods in this experiment was also similar to what we saw in the first experiment (Figs. 1 and 2, Appendix J, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68): MAFFT alignments were more accurate than both ClustalW and POA alignments, and POA generally the least accurate. Interestingly, however, the comparison between ClustalW and POA occasionally varied. For most of the model conditions, ClustalW is better than POA, but given the hardest model conditions, ClustalW is worse than POA. Further, ClustalW's accuracy seems to degrade with increasing numbers of taxa, but this trend does not hold for the other MSA methods. Also, ML consistently produces better trees than MP or NJ. However, there is no clear relative advantage between MP and NJ, as the relative performance seems to depend upon the particular model conditions.

An examination of the correlation between alignment error and tree error, for each of the three phylogeny estimation methods, shows that alignment error and tree error are positively correlated, with the strongest correlation obtained by ML-estimated trees, then by MP-estimated trees, and finally by NJ-estimated trees. These positive correlations are, however, relatively weak: when computing alignment error using 1-SP, the Spearman correlation for ML is 0.443, for MP it is 0.386, and for NJ it is 0.339 (Fig. 3 for RAxML; see Appendix K, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68, for NJ and MP trees). Note also the

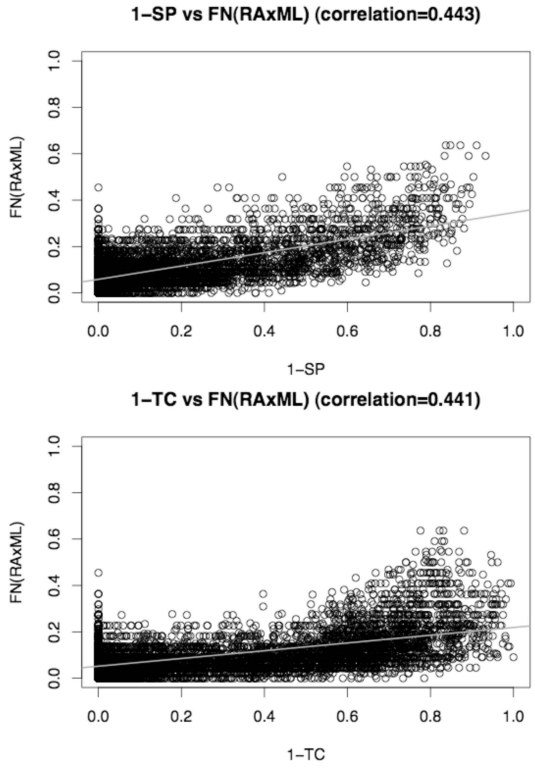

Fig. 3. Relationship of alignment error (1-SP and 1-TC) and RAxML reconstruction error (FN(RAxML)) for Experiment 2. In each scatter plot, each of the 1,800 points corresponds to one of the 30 replicate runs from the 60 model settings; its x coordinate is the average alignment error, and its y coordinate is the topological error of the reconstructed tree with respect to the true tree. The dark gray line corresponds to the linear regression over the 1,800 points.

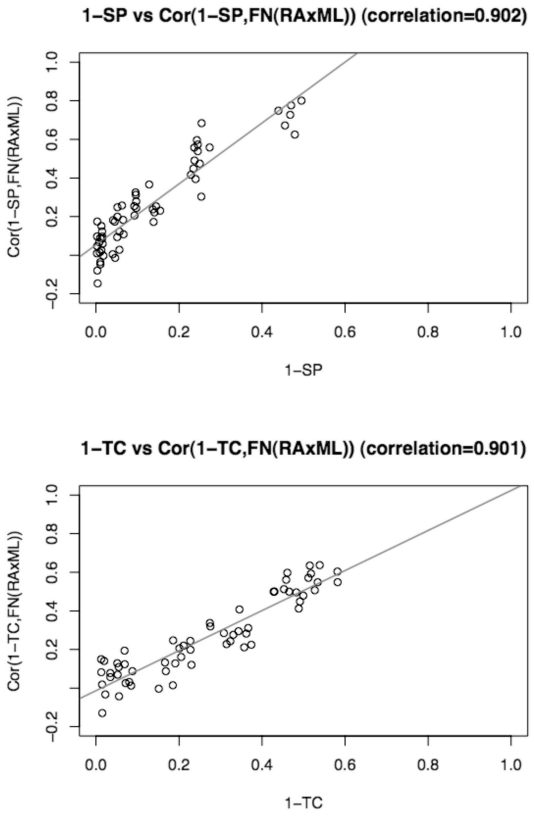

Fig. 4. Relationship between alignment error (1-SP and 1-TC) and the correlation between alignment error versus RAxML reconstruction error (FN(RAxML)) for Experiment 2. In each scatter plot, each point corresponds to one of the 60 model settings; its x coordinate is the average alignment error, and its y coordinate is the Spearman correlation coefficient between the alignment error and the topological error of the reconstructed tree with respect to the true tree. The dark gray line corresponds to the linear regression over the 60 points.

greater variability in tree error for ML-estimated trees than for MP and NJ-estimated trees (Fig. 2), which reflects the greater correlation between alignment error and tree error for ML.

The overall weakness of the positive correlation would tend to suggest that alignment error does not have a large impact on phylogeny reconstruction, but this turns out to be a premature conclusion. A careful analysis of the data reveals that the correlation varies tremendously between model conditions. For example, the correlations between alignment error measured using 1-SP and the missing edge error rates for ML-estimated trees vary between 0.03 (very weakly positive) and 0.8 (very strongly positive), depending upon the model condition. Similar trends, but with smaller values, occur for MP (-0.1 to 0.56) and NJ (0.3 to 0.54). Thus, the model condition is an important consideration in terms of predicting whether alignment error will have a large or small impact on phylogenetic accuracy.

An investigation of the differences between the model conditions reveals the striking observation that the correlation between alignment error and tree error is strongly based upon the average alignment error (averaged over all alignment methods) for the model condition: when the average alignment error is low, the correlation will be weak, and when the average alignment error is high, the correlation will be large. Fig. 4 shows scatter plots comparing alignment error (here 1-SP and 1-TC against RAxML are shown; see Appendix L, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68, for NJ and MP trees) and the correlation between alignment error and tree error, for each of the phylogeny estimation methods. Note the strong correlation for each method when alignment error is calculated using 1-SP: ML exhibits the strongest correlation of 0.902, MP exhibits the correlation 0.844, and NJ exhibits 0.846. Also notice that when the alignment error is low (below 20 percent, say), the correlation is quite weak—not above 0.4 for ML, and not above 0.3 for MP and NJ. Indeed, it is only for cases where the average alignment error is quite high (around 50 percent) that alignment error is strongly correlated to tree error. This striking observation helps explain why substantial differences in alignment error can result in modest differences in phylogenetic tree error (i.e., ClustalW and MAFFT have large differences in alignment error but result in trees with approximately the same error for Model Tree 0 from the BaliBASE model trees, as well as for several of the random birth-death model trees).

We computed the standard deviation of the alignment error and phylogeny FN error scores, and the correlations between three alignment error scores and FN scores of the three reconstructed phylogenies, with the three alignments analyzed separately (Appendix O, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68). We found that the correlations range from 0.49 to 0.78; the correlation is highest in POA, followed by ClustalW, and lowest in

Mafft. We hypothesize that this is due to the higher variability of FN rates and alignment errors in POA (the least accurate method) relative to Mafft (the most accurate of the three).

We performed analysis using regression with additive effects of log-transformed number of taxa, sequence length, insertion/deletion rate, and diameter; regressions are performed separately for the four alignment methods (Appendix M, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68) to find the effects of simulation parameters. Overall, increasing the number of taxa, indel rate, or diameter all increase the width and gappiness of alignments. Increases in the diameter almost always increase alignment and phylogeny error. Longer sequences in simulation decrease alignment width and gappiness for POA and ClustalW, but has no effect for TrueAln and Mafft. Increases in the indel rate increase alignment and phylogeny error in general for the three alignment methods, though this has no effect on phylogeny reconstruction using TrueAln. Thus, increases in the indel rate seem to make alignment more difficult, which then results in phylogenetic errors increasing.

In summary, there is a generally positive correlation between alignment and tree error when model conditions produce data sets with relatively high average alignment errors. The positive correlation between alignment error and ML tree accuracy is much weaker when alignment errors are low and evaporates for MP and NJ analyses of data sets with low average alignment error. This suggests that except for model conditions that produce data sets that are quite difficult to align, there will only be small consequences to choosing between a very good alignment and a somewhat poorer alignment—a prediction that is supported by Fig. 2, which shows that except for the most difficult model conditions (with the highest indel rates and most gaps), RAxML trees based upon MAFFT, ClustalW, and the true alignment all tend to have about the same error, even though the estimated alignments can still have substantial error. (For the full set of correlation coefficients between various measures in our study across the replicate runs using all 48 settings, see the table of coefficients at http://people.pcbi.upenn.edu/~lswang/alignexp/correlation.xls.)

4 Discussion

Our study supports earlier studies that showed that the new MSA methods produced by the protein alignment research community do provide better estimates of alignments than ClustalW, and that MAFFT and ProbCons have the highest accuracy of the methods we examined. We also confirm earlier studies showing that maximum likelihood produces better trees than maximum parsimony or neighbor joining, and that improving the alignment will tend to improve the tree. However, here, we find that the impact of the alignment method on phylogenetic accuracy depends upon the model condition, so that under some conditions the impact is relatively small. More precisely, when the model condition produces data sets that are generally difficult to align (i.e., when indel rates and substitution rates are both high, see Appendix P, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68), then alignment accuracy and phylogenetic accuracy will be strongly correlated, but when the model condition produces data sets that are easier to align, then alignment and phylogenetic accuracy will only be weakly correlated. As a result, if the choice is between a very good method (such as ProbCons or MAFFT) and a very poor method (such as POA), then it is likely that choosing the better method will result in a better tree, but if the choice is between two very good methods, or even between a very good method and a moderately good one (like ClustalW), there maybe little impact on the resultant phylogenetic accuracy. Indeed, it seems that only when the estimated alignments are sufficiently poor (perhaps with alignment SP-error rates above 20 or 30 percent) will differences in alignment error reliably produce an appreciable impact on the resultant phylogeny. Furthermore, data sets with higher evolutionary rates (larger diameters) and more indels tend to show bigger differences in the accuracy of phylogenies estimated on different alignments. For now, we hypothesize that when alignments are relatively easy, there is enough phylogenetic signal in any "reasonable" alignment (even one with perhaps 20 percent of the homologous pairs missing) to reproduce much of the tree one would get if one had the true alignment. These observations may help resolve the seeming contradictory findings of earlier studies, in which alignments have sometimes been shown to have a big impact on phylogenetic estimation, but not always. From a practical point of view, it would be desirable to develop a metric that predicts whether a particular data set falls into the particular model conditions for which alignment is likely to have a significant impact on phylogenetic accuracy. One obvious approach would be to compare the phylogenies based upon different MSAs to see if the phylogenies change, but other approaches merit inquiry.

While sequence alignment quality can influence phylogentic accuracy, using the historical signal in insertions and deletions for phylogeny reconstruction is also a challenge. Our research has focused on how alignment quality influences phylogenetic analysis of amino acid substitution, but improvements in phylogentic estimation might be expected if the information content of indels were utilized in a more sensitive and appropriate manner.

Several attempts have been made to produce phylogenetic reconstruction methods that handle gaps appropriately. Our study showed that removing "gappy" sites did not improve phylogenetic estimations; however, other methods for preprocessing alignments prior to phylogenetic estimation might be able to improve the final phylogeny. For example, gap-coding techniques that add indel characters to the data set have only been developed for parsimony analyses (e.g., [44]). However, gaps can also be considered in the calculation of the distance matrix given to distance-based methods like neighbor joining. It is possible that the incorporation of such techniques may enable better estimations of phylogenies with the consequence that alignment accuracy could have a bigger impact on phylogenetic accuracy. However, to date, these techniques have not been widely used by the phylogenetics community.

Beyond these modifications to distance-based and parsimony-based analyses, other methods have also been

attempted that incorporate gap events into the phylogeny reconstruction process. Indeed, more than 30 years ago, Sankoff asserted that the best alignment would be based upon the true phylogeny [45], and more recent work [18] has supported this assertion. This assertion would suggest that alignments and trees should be sought simultaneously, as some methods of phylogenetic inference attempt [17], [46], [47], [48]. However, there are surprisingly few formal investigations of how these methods perform, and the few that have been done have concluded that the current techniques for simultaneous estimation of alignments and trees are not as accurate as the best two-phase techniques (Fleissner et al. [46] showed that their statistical coestimation technique was not as accurate as a phylogenetic analysis based upon ClustalW alignments, while Kjer et al. [49] and Ogden and Rosenberg [50] showed problems with using POY as compared to phylogenies based upon ClustalW alignments). While some researchers [49] remain optimistic about the potential for statistical methods to provide high-quality alignments and trees, these are computationally too intensive [51] to be used except on small data sets. We have not included these programs in our simulation study due to their long running times, which make comprehensive simulation studies difficult; nonetheless, this is an area of active research, and advancements in implementation of simultaneous estimation of alignments and phylogenies [24] and model-based alignment [52] will promote improvements in both simultaneous and two-phase approaches to sequence alignment and phylogeny reconstruction.

How the alignment algorithms affect other aspects of phylogeny reconstruction such as branch lengths and internal node sequence estimation is an interesting research direction that has important biological significance for other topics including molecular dating and detection of directed evolution. Finally, incorporating uncertainty into a phylogenetic analysis, through the exploration of the uncertainty in the alignment itself, is also desirable [27], [52].

Another possible direction is to develop more informative alignment error metrics that can better predict the accuracy of inferred phylogeny. We have shown that the three alignment errors used in our simulation studies provide only limited information; we showed (see Appendix Q, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.68) that even by varying the alignment method, choosing the most accurate alignment using TC or SP does not lead to the most accurate phylogeny in Experiment 1; the difference in FN rate can be as high as seven percent. We tested the newly proposed AMA alignment error metric [19] in the Yule experiment and found that AMA is highly correlated with SP and TC scores (correlation>0.98), and all observations regarding SP and TC apply to AMA. It is likely that phylogentically relevant error metrics still need to be developed, since the current accuracy measures are not sufficiently informative.

In summary, the following questions are of major concern: First, the development of alignment methods that can produce more accurate alignments than current methods will only improve phylogeny estimation for the more difficult model conditions investigated in our study (high indel rates and tree diameters). Second, the development of phylogeny estimation methods that can utilize the historical signal in indel events would be expected to improve phylogenetic accuracy.

Finally, although our study focused on amino acid alignment and we believe these results will extend to nucleotide alignments. However, it is possible that there is sufficiently greater phylogenetic signal in nucleotides as compared to protein sequences, that there could be a higher correlation between DNA sequence alignment and tree error rates over a wider set of model conditions. The implications of our results may also be limited to protein-coding genes (versus rDNA or noncoding sequence alignments). For example, previous work has shown that the quality of ribosomal DNA alignments can have a significant impact on phylogenetic accuracy [26]. Future investigation of these issues and the utility of simultaneous alignment and phylogeny estimation are merited.

Acknowledgments

This research was supported in part by the US National Science Foundation (NSF) grants DBI 0638595 and DBI 0115684 to Claude W. dePamphilis and Jim Leebens-Mack, DEB 0733029 to Tandy Warnow and Jim Leebens-Mack, and ITR 0331453, ITR/AP 0121680, and DEB 0120709 to Tandy Warnow. Tandy Warnow was also supported by the Program for Evolutionary Dynamics at Harvard, by the Radcliffe Institute for Fundamental Research, and by the University of Texas Faculty Research program.

References

[1] R.B. Russell and G.J. Barton, "Multiple Protein Sequence Alignment from Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels," *Proteins,* vol. 14, pp. 309-323, Oct. 1992.

[2] A. Andreeva et al., "SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data," *Nucleic Acids Research,* vol. 32, pp. D226-D229, Jan. 2004.

[3] A.G. Murzin et al., "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," *J. Molecular Biology,* vol. 247, pp. 536-540, Apr. 1995.

[4] K. Mizuguchi et al., "HOMSTRAD: A Database of Protein Structure Alignments for Homologous Families," *Protein Science,* vol. 7, pp. 2469-2471, Nov. 1998.

[5] A. Bahr et al., "BAliBASE (Benchmark Alignment dataBASE): Enhancements for Repeats, Transmembrane Sequences and Circular Permutations," *Nucleic Acids Research,* vol. 29, pp. 323-326, Jan. 2001.

[6] J.D. Thompson et al., "BAliBASE 3.0: Latest Developments of the Multiple Sequence Alignment Benchmark," *Proteins,* vol. 61, pp. 127-136, Oct. 2005.

[7] J.D. Thompson et al., "BAliBASE: A Benchmark Alignment Database for the Evaluation of Multiple Alignment Programs," *Bioinformatics,* vol. 15, pp. 87-88, Jan. 1999.

[8] R.C. Edgar, "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput," *Nucleic Acids Research,* vol. 32, pp. 1792-1797, 2004.

[9] R.A. Cartwright, "DNA Assembly with Gaps (Dawg): Simulating Sequence Evolution," *Bioinformatics,* vol. 21, no. 3, pp. iii31-iii38, Nov. 2005.

[10] A. Pang et al., "SIMPROT: Using an Empirically Determined Indel Distribution in Simulations of Protein Evolution," *BMC Bioinformatics,* vol. 6, p. 236, 2005, doi:10.1186/1471-2105-6-236.

[11] J. Stoye et al., "Rose: Generating Sequence Families," *Bioinformatics,* vol. 14, pp. 157-163, 1998.

[12] C.L. Strope et al., "indel-Seq-Gen: A New Protein Family Simulator Incorporating Domains, Motifs, and Indels," *Molecular Biology and Evolution,* vol. 24, pp. 640-649, Mar. 2007.

[13] K. Katoh et al., "MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment," *Nucleic Acids Research,* vol. 33, pp. 511-518, 2005.
[14] C.B. Do et al., "ProbCons: Probabilistic Consistency-Based Multiple Sequence Alignment," *Genome Research,* vol. 15, pp. 330-340, Feb. 2005.
[15] C. Notredame et al., "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment," *J. Molecular Biology,* vol. 302, pp. 205-217, Sept. 2000.
[16] A.R. Subramanian et al., "DIALIGN-T: An Improved Algorithm for Segment-Based Multiple Sequence Alignment," *BMC Bioinformatics,* vol. 6, p. 66, 2005, doi:10.1186/1471-2105-6-66.
[17] W. Wheeler et al., *Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY*. Am. Museum of Natural History, 2006.
[18] A. Loytynoja and N. Goldman, "Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis," *Science,* vol. 320, pp. 1632-1635, June 2008.
[19] A.S. Schwartz and L. Pachter, "Multiple Alignment by Sequence Annealing," *Bioinformatics,* vol. 23, pp. e24-e29, Jan. 2007.
[20] U. Roshan and D.R. Livesay, "Probalign: Multiple Sequence Alignment Using Partition Function Posterior Probabilities," *Bioinformatics,* vol. 22, pp. 2715-2721, Nov. 2006.
[21] J.D. Thompson et al., "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice," *Nucleic Acids Research,* vol. 22, pp. 4673-4680, Nov. 1994.
[22] G. Blackshields et al., "Analysis and Comparison of Benchmarks for Multiple Sequence Alignment," *In Silico Biology,* vol. 6, pp. 321-339, 2006.
[23] R.C. Edgar and S. Batzoglou, "Multiple Sequence Alignment," *Current Opinion in Structural Biology,* vol. 16, pp. 368-373, June 2006.
[24] S. Nelesen et al., "The Effect of the Guide Tree on Multiple Sequence Alignments and Subsequent Phylogenetic Analyses," *Proc. Pacific Symp. Biocomputing,* pp. 15-24, 2008.
[25] G.P. Raghava et al., "OXBench: A Benchmark for Evaluation of Protein Multiple Sequence Alignment Accuracy," *BMC Bioinformatics,* vol. 4, p. 47, Oct. 2003, doi:10.1186/1471-2105-4-47.
[26] D.A. Morrison and J.T. Ellis, "Effects of Nucleotide Sequence Alignment on Phylogeny Estimation: A Case Study of 18S rDNAs of Apicomplexa," *Molecular Biology and Evolution,* vol. 14, pp. 428-441, Apr. 1997.
[27] K.M. Wong et al., "Alignment Uncertainty and Genomic Analysis," *Science,* vol. 319, pp. 473-476, Jan. 2008.
[28] B.L. Cantarel et al., "Exploring the Relationship Between Sequence Similarity and Accurate Phylogenetic Trees," *Molecular Biology and Evolution,* vol. 23, pp. 2090-2100, Nov. 2006.
[29] B.G. Hall, "Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences," *Molecular Biology and Evolution,* vol. 22, pp. 792-802, Mar. 2005.
[30] T.H. Ogden and M.S. Rosenberg, "Multiple Sequence Alignment Accuracy and Phylogenetic Inference," *Systematic Biology,* vol. 55, pp. 314-328, Apr. 2006.
[31] U. Roshan and D.R. Livesay, "Improving Progressive Alignment for Phylogeny Reconstruction Using Parsimonious Guide-Trees," *Proc. Sixth IEEE Symp. Bioinformatics and Bioeng.,* 2006.
[32] D.F. Robinson and L.R. Foulds, "Comparison of Phylogenetic Trees," *Math. Biosciences,* vol. 53, pp. 131-147, 1981.
[33] C. Grasso and C. Lee, "Combining Partial Order Alignment and Progressive Multiple Sequence Alignment Increases Alignment Speed and Scalability to Very Large Alignment Problems," *Bioinformatics,* vol. 20, pp. 1546-1556, July 2004.
[34] J.D. Thompson et al., "DbClustal: Rapid and Reliable Global Multiple Alignments of Protein Sequences Detected by Database Searches," *Nucleic Acids Research,* vol. 28, pp. 2919-2926, Aug. 2000.
[35] M.O. Dayhoff, "Observed Frequencies of Amino Acid Replacements between Closely Related Proteins," *Atlas of Protein Sequence and Structure,* M.O. Dayhoff, ed., vol. 5, Nat'l Biomedical Research Foundation, 1978.
[36] D.L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.*, Sinauer Assoc., 2003.
[37] A. Stamatakis, "RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models," *Bioinformatics,* vol. 22, pp. 2688-2690, Nov. 2006.
[38] J. Felsenstein, "PHYLIP—Phylogeny Inference Package (Version 3.2)," *Cladistics,* vol. 5, pp. 164-166, 1989.
[39] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics.* Oxford Univ. Press, 2000.
[40] M. Cline et al., "Predicting Reliable Regions in Protein Sequence Alignments," *Bioinformatics,* vol. 18, pp. 306-314, Feb. 2002.
[41] B. Rannala et al., "Taxon Sampling and the Accuracy of Large Phylogenies," *Systematic Biology,* vol. 47, pp. 702-710, Dec. 1998.
[42] S. Guindon and O. Gascuel, "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood," *Systematic Biology,* vol. 52, pp. 696-704, Oct. 2003.
[43] M.J. Sanderson, "r8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock," *Bioinformatics,* vol. 19, pp. 301-302, Jan. 2003.
[44] M.P. Simmons et al., "The Relative Performance of Indel-Coding Methods in Simulations," *Molecular Phylogenetics and Evolution,* vol. 44, pp. 724-740, Aug. 2007.
[45] D. Sankoff, "Minimal Mutation Trees of Sequences," *SIAM J. Applied Math.,* vol. 28, pp. 35-42, 1975.
[46] R. Fleissner et al., "Simultaneous Statistical Multiple Alignment and Phylogeny Reconstruction," *Systematic Biology,* vol. 54, pp. 548-561, Aug. 2005.
[47] G. Lunter, A.J. Drummond, I. Miklos, and J. Hein, "Statistical Alignment: Recent Progress, New Applications, and Challenges," *Statistical Methods in Molecular Evolution (Statistics for Biology and Health),* R. Nielsen, ed., pp. 375-406, Springer, 2005.
[48] B.D. Redelings and M.A. Suchard, "Joint Bayesian Estimation of Alignment and Phylogeny," *Systematic Biology,* vol. 54, pp. 401-418, June 2005.
[49] K. Kjer et al., "Opinions on Multiple Sequence Alignment, and An Empirical Comparison of Repeatability and Accuracy between POY and Structural Alignment," *Systematic Biology,* vol. 56, pp. 133-146, 2007.
[50] T.H. Ogden and M.S. Rosenberg, "Alignment and Topological Accuracy of the Direct Optimization Approach via POY and Traditional Phylogenetics via ClustalW + PAUP*," *Systematic Biology,* vol. 56, pp. 182-193, Apr. 2007.
[51] G. Lunter et al., "Bayesian Coestimation of Phylogeny and Sequence Alignment," *BMC Bioinformatics,* vol. 6, p. 83, 2005, doi:10.1186/1471-2105-6-83.
[52] G. Lunter et al., "Uncertainty in Homology Inferences: Assessing and Improving Genomic Sequence Alignment," *Genome Research,* vol. 18, pp. 298-309, Feb. 2008.

Li-San Wang received the BS and MS degrees in electrical engineering from the National Taiwan University, in 1994 and 1996, respectively, and the MS and PhD degrees in computer science from the University of Texas at Austin, in 2000 and 2003, respectively. He was a postdoctoral fellow at the University of Pennsylvania between 2003 and 2006. Currently, he is an assistant professor of Pathology and Laboratory Medicine, a faculty member of Penn Center for Bioinformatics, and a fellow of the Institute on Aging, the University of Pennsylvania. His research interests include phylogenetics, comparative genomics, and microarray and other high-throughput biotechnologies.

Jim Leebens-Mack received the PhD degree in botany from the University of Texas at Austin in 1995, and was a postdoctoral researcher at Vanderbilt University from 1995 to 1998. Currently, he is a plant evolutionary biologist and assistant professor of plant biology at the University of Georgia. He is the University of Georgia PI for two ATOL projects (the Monocot ATOL project and the Multiple Sequence Alignment Project), and co-PI for the Ancestral Angiosperm Genome Project.

P. Kerr Wall received the PhD degree in biology from Pennsylvania State University in 2008. From 2001 to 2008, he was the lead bioinformatics programmer for the Floral Genome and Ancestral Angiosperm Genome projects. Following postdoctoral research in plant genomics at Penn State, he has taken a position as bioinformatics programmer with BASF Plant Genomics.

Kevin Beckmann received the dual bachelor's degrees in biology and minored in computer science from Colgate University in 2002, and was a graduate student in biology at Pennsylvania State University from 2003 to 2006.

Claude W. dePamphilis received the PhD degree from the University of Georgia. He was a US National Science Foundation (NSF) postdoctoral fellow in plant biology with Jeff Palmer at the University of Michigan and Indiana University, and is now professor of biology at Pennsylvania State University. He is a plant evolutionary biologist with broad interests in processes and patterns of evolution at both the molecular and organismal levels. His research focuses on three project areas—the evolution of the flower and the floral developmental program, the study of parasitic plants, and chloroplast genome and phylogenomic evolution. He is the PI of the Floral Genome and Ancestral Angiosperm Genome projects, and co-PI of the Parasitic Plant Genome Project and the Monocot ATOL.

Tandy Warnow received the PhD degree in mathematics at UC Berkeley in 1991 under the direction of Gene Lawler, and did postdoctoral training with Simon Tavare and Michael Waterman at USC. She is professor of computer science at the University of Texas at Austin. Her research combines mathematics, computer science, and statistics to develop improved models and algorithms for reconstructing complex and large-scale evolutionary histories in both biology and historical linguistics. She received the US National Science Foundation (NSF) Young Investigator Award in 1994, and the David and Lucile Packard Foundation Award in science and engineering in 1996.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.