Syst. Biol. 68(3):396–411, 2019

DOI:10.1093/sysbio/syy068
Advance Access publication October 17, 2018

Evaluating Statistical Multiple Sequence Alignment in Comparison to Other Alignment Methods on Protein Data Sets

Michael Nute[1], Ehsan Saleh[2], and Tandy Warnow[2,3,4,*]

[1]*Department of Statistics, University of Illinois at Urbana-Champaign, 725 S Wright St #101, Champaign, IL 61820, USA;* [2]*Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave, Urbana, IL 61801, USA;* [3]*Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1205 W. Clark St., Urbana, IL 61801, USA;* [4]*National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

[*]*Correspondence to be sent to: Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave, Urbana, IL 61801, USA;*
E-mail: warnow@illinois.edu

Received 18 April 2018; reviews returned 27 September 2018; accepted 11 October 2018
Associate Editor: Brant Faircloth

Abstract.—The estimation of multiple sequence alignments of protein sequences is a basic step in many bioinformatics pipelines, including protein structure prediction, protein family identification, and phylogeny estimation. Statistical coestimation of alignments and trees under stochastic models of sequence evolution has long been considered the most rigorous technique for estimating alignments and trees, but little is known about the accuracy of such methods on biological benchmarks. We report the results of an extensive study evaluating the most popular protein alignment methods as well as the statistical coestimation method BAli-Phy on 1192 protein data sets from established benchmarks as well as on 120 simulated data sets. Our study (which used more than 230 CPU years for the BAli-Phy analyses alone) shows that BAli-Phy has better precision and recall (with respect to the true alignments) than the other alignment methods on the simulated data sets but has consistently lower recall on the biological benchmarks (with respect to the reference alignments) than many of the other methods. In other words, we find that BAli-Phy systematically underaligns when operating on biological sequence data but shows no sign of this on simulated data. There are several potential causes for this change in performance, including model misspecification, errors in the reference alignments, and conflicts between structural alignment and evolutionary alignments, and future research is needed to determine the most likely explanation. We conclude with a discussion of the potential ramifications for each of these possibilities. [BAli-Phy; homology; multiple sequence alignment; protein sequences; structural alignment.]

Multiple sequence alignment is a basic step in many bioinformatics pipelines, including phylogenetic estimation, but also for analyses specifically aimed at understanding proteins. For example, protein alignment is used in protein structure and function prediction (Cuff and Barton 2000), protein family and domain identification (George and Heringa 2002; Mulder and Apweiler 2002), functional site identification (Sankararaman and Sjölander 2008; Alterovitz et al. 2009), domain identification (Bernardes et al. 2016), inference of ancestral proteins (Holmes 2017), detection of positive selection (Fletcher and Yang 2010), and protein–protein interactions (Xue et al. 2015). However, multiple sequence alignment is often difficult to perform with high accuracy, and errors in alignments can have a substantial impact on the downstream analyses (Lake 1991; Morrison and Ellis 1997; Ogden and Rosenberg 2006; Dessimoz and Gil 2010; Fletcher and Yang 2010; Simmons et al. 2010; Wang et al. 2011; Karin et al. 2014; Philippe et al. 2017). For this reason, the evaluation of multiple sequence alignment methods (and the development of new methods with improved accuracy), especially for protein sequences, has been a topic of substantial interest in the bioinformatics research community (e.g., Thompson et al. 2011; Wang et al. 2011; Iantorno et al. 2014; Pais et al. 2014; Le et al. 2017).

Protein alignment methods have mainly been evaluated using databases, such as BAliBase (Bahr et al. 2001), Homstrad (Mizuguchi et al. 1998), SABmark (Van Walle et al. 2005), Sisyphus (Andreeva et al. 2007), and Mattbench (Daniels et al. 2012), that provide reference alignments for different protein families and superfamilies based on structural features of the protein sequences (see discussions about these benchmarks in Aniba et al. 2010; Iantorno et al. 2014). Performance studies evaluating protein alignment methods using these benchmarks (e.g., Blackshields et al. 2006; Edgar and Batzoglou 2006; Kemena et al. 2011; Sievers et al. 2011; Thompson et al. 2011; Mirarab et al. 2015; Nguyen et al. 2015) have revealed conditions under which alignment methods degrade in accuracy (e.g., large data sets, or highly heterogeneous data sets with low average pairwise sequence identity) and have also revealed differences between alignment methods in terms of accuracy, computational efficiency, and scalability to large data sets. In turn, the databases have been used to provide training data for machine learning techniques to infer alignments on novel data sets (e.g., Do et al. 2006; Roshan and Livesay 2006). Method development for protein alignment is thus strongly influenced by these databases and has produced several protein alignment methods that are considered highly accurate and robust to many different challenging conditions.

An alternative approach to multiple sequence alignment has been developed within the statistical phylogenetics community in which an alignment is coestimated with a phylogenetic tree by considering stochastic models of evolution in which sequences evolve down a model tree under a process that includes substitutions, insertions, and deletions (jointly referred to as "indels"). Likelihood-based estimation of alignments and/or trees under these models provide a mathematically rigorous and therefore appealing approach, and was initially proposed in Bishop and Thompson (1986). Subsequent extensions of this basic approach were made in a sequence of papers (Thorne et al. 1991, 1992a, 1992b; Holmes and Bruno 2001; Miklós 2002, 2003; Hein et al. 2003; Lunter et al. 2003; Miklós et al. 2004; Fleissner et al. 2005; Lunter et al. 2005; Suchard and Redelings 2006; Redelings and Suchard 2007; Novák et al. 2008; Bradley et al. 2009; Redelings 2014). BAli-Phy (Suchard and Redelings 2006; Redelings and Suchard 2007; Redelings 2014), a Bayesian method that uses MCMC sampling to jointly estimate the multiple sequence alignment and phylogenetic tree under a stochastic sequence evolution model that allows for indels and substitutions, is the most well-known of these methods.

A related approach is PRANK (Löytynoja and Goldman 2008), which closely adheres to a phylogenetic model of sequence evolution but does not rely on a detailed stochastic model to the same degree. Because of its similarity in design objectives, Blackburne and Whelan (2013) refer to PRANK as a "heuristic to full statistical alignment." Blackburne and Whelan (2013) examined alignments computed using BAli-Phy and PRANK in comparison to other methods on biological protein data sets; BAli-Phy and PRANK were clearly outliers in their visualization using principle co-ordinate analysis (PCoA, a type of multidimensional scaling), while the remaining methods largely grouped together.

Only a few studies have evaluated BAli-Phy for accuracy on either biological or simulated data. Three studies (Liu et al. 2009; Redelings 2014; Nute and Warnow 2016) evaluated BAli-Phy on simulated nucleotide data sets and found it to have superior accuracy compared to the other alignment methods they examined; this question was examined directly in Liu et al. (2009), Redelings (2014), and indirectly in Nute and Warnow (2016) through the substitution of MAFFT (Katoh et al. 2002) by BAli-Phy within PASTA (Mirarab et al. 2015), a div ide-and-conquer meta-method that is designed to scale MSA methods to larger data sets.

Additionally, Katoh and Standley (2016) evaluated BAli-Phy on protein biological benchmarks as well as on simulated protein data sets (to the best of our knowledge, this is the only study that has evaluated BAli-Phy in terms of accuracy on biological data). In their study, BAli-Phy was much less accurate than some other MSA methods (PRANK, Muscle, Edgar 2004, and variants of MAFFT) on the biological data, but was very good (and for some criteria it was the best) on the simulated data. This study is intriguing but its evaluation of BAli-Phy was limited; the data analyzed were large for BAli-Phy (the simulated data sets had 100 sequences, and the biological data sets ranged up to 100 sequences), and each run was limited to 1000 MCMC iterations. As discussed by the authors (and in Redelings 2018), 1000 MCMC iterations may not have been sufficient to allow BAli-Phy to reach convergence on data sets of this size, and it is known that BAli-Phy can have reduced accuracy if stopped prematurely (Redelings 2018). The contrast in performance on biological and simulated data is notable but a more careful evaluation of BAli-Phy is necessary to determine whether these trends were spurious.

Here, we report on an extensive performance study in which we compared BAli-Phy version 2.3.8 to a collection of leading protein sequence alignment methods. We used 1192 data sets from four established benchmark databases of protein multiple sequence alignments (BAliBASE v3.0, Sisyphus v1.2, Mattbench, and Homstrad, all downloaded in March 2017) as well as 120 simulated data sets in order to characterize the relative and absolute accuracy of the alignment methods we explore. We limited our study to biological sequence data sets with at most 25 sequences and to simulated data sets (under 6 model conditions) with 27 sequences, so that we were able to run BAli-Phy for long enough to improve its chances of converging. In particular, we ran BAli-Phy on each data set using 32 independent runs, each for 48 h (i.e., BAli-Phy was run somewhat longer than 2 months on each data set). This analysis protocol enabled BAli-Phy to generate many hundreds of thousands (and in several cases more than 1,000,000) of MCMC samples for each data set that it analyzed and hence to achieve good ESS values that suggest that BAli-Phy may have converged on these data sets. Overall our study used more than 230 CPU years for the BAli-Phy analyses alone and provides a careful evaluation of how BAli-Phy performs on biological and simulated data sets.

In our simulation study, BAli-Phy produced alignments that had higher Modeler scores (a measure of precision) and SP-scores (a measure of recall) than other alignment methods; furthermore, BAli-Phy produced alignments that were very close in length to the true alignment. The results on biological data, however, were quite different. There, BAli-Phy generally had Modeler scores that were often better than most other alignment methods, but SP-scores that were lower than many other alignment methods; furthermore, BAli-Phy produced alignments that were generally longer than the reference alignment and also longer than all the other alignments. In other words, BAli-Phy tended to underalign on biological data but not on the simulated data and was visibly an outlier on the biological data in terms of alignment length. Our results are consistent with the prior studies discussed above in finding that BAli-Phy is an outlier among alignment methods and provides a more detailed examination of these differences. Finally, there are several possible explanations for why BAli-Phy has different performance on biological and simulated data (discussed below), and further research is needed to determine the causes of these differences.

Materials and Methods

Alignment Methods

We explored the following multiple sequence alignment methods: BAli-Phy v. 2.3.6, Clustal-Omega v. 1.2.4 (Sievers et al. 2011), CONTRAlign v. 1.04 (Do et al. 2006), DiAlign v. 2.2.2 (Morgenstern 1999; Golubchik et al. 2007), Kalign v. 2.04 (Lassmann and Sonnhammer 2005), MAFFT v. 7.305b (Katoh et al. 2002), Muscle v. 3.8.31 (Edgar 2004), PRANK v. 140603 (Löytynoja and Goldman 2005, 2008), PRIME v. 1.1 (Yamada et al. 2006), Probalign v. 1.4 (Roshan and Livesay 2006), ProbCons v. 1.12 (Do et al. 2005), PROMALS3D (retrieved January 31, 2018, installed and run with Python 2.7.8 and GCC v. 4.7.1) (Pei et al. 2008), and T-Coffee v. 11.00.8cbe486 (Notredame et al. 2000; O'Sullivan et al. 2004; Notredame 2007). We explore two ways of running MAFFT: MAFFT-G-INS-i and MAFFT-Homologs (using the SwissProt Database, Bairoch and Apweiler 2000).

All methods other than BAli-Phy and Promals3D were performed in default mode. Promals-3D enables structural alignment features, but we turned these off using the following sample command:

```
python promals < InputSequencesFile > -dali 0
-tmalign 0 -fast 0
```

BAli-Phy requires specific parameters (including the substitution model and the number of MCMC iterations) to be set by the user. We used RAxML (Stamatakis 2006) version 8.2.9 to select the protein sequence evolution model based on likelihood scores obtained on the alignment computed using MAFFT L-ins-i (see Supplementary Section 1.3 available on Dryad at http://dx.doi.org/10.5061/dryad.8k821ds, for details). We ran 32 independent runs of BAli-Phy, each for 48 h, discarding the first 25% of the alignments that were generated during the MCMC run, and then retaining every 10th alignment in the remaining sample. The point estimates of the alignments were computed using the posterior decoding (PD). According to the output from BAli-Phy, the vast majority of the BAli-Phy runs we performed had good ESS values, which suggests that BAli-Phy may have converged on those data; see Supplementary Section 2.1 available on Dryad, for these statistics.

Computational Resources

BAli-Phy and T-Coffee are the most computationally intensive methods we explored, and so these were run on the Blue Waters supercomputer at the National Center for Supercomputing Applications (NCSA); all other methods were run on the Campus Cluster at the University of Illinois at Urbana-Champaign.

Table 1. Empirical properties of the 1192 reference alignments from the four biological benchmark collections

Database	PID	# seqs.	Alignment length	% gapped	Gap length
BAliBase	0.30	12.4	772.0	37.7	8.1
Homstrad	0.37	6.9	257.3	16.6	2.7
Mattbench	0.20	7.3	416.4	44.6	2.8
Sisyphus	0.26	9.4	172.3	21.0	4.9

Note: We report the average pairwise sequence identity (PID), average number of sequences, average alignment length, average fraction of the reference alignment occupied by gaps, and median gap length.

Data Sets

Protein biological data sets.— We took all the alignments from the four databases we selected (BAliBASE, Mattbench, Homstrad, and Sisyphus) that had between 4 and 25 sequences. Each alignment with more than 25 sequences was then subsampled to produce a data set with between 5 and 25 sequences; see Supplementary Section 1.2 available on Dryad, for the protocol used for subsampling.

T-Coffee failed to align a number of data sets, returning empty folders; this was particularly pronounced on the BAliBase data, where 82 out of 742 alignments were not completed, although it also failed to align 2 data sets each from the other three benchmarks (see Supplementary Section 2.3 available on Dryad, for discussion). BAli-Phy was able to analyze all the data sets, but on two data sets the posterior decoding algorithm failed due to the high computational complexity of having a small number of very long sequences. After eliminating the data sets where T-Coffee and the Bali-Phy posterior decoding failed to complete, we still had a large number (1192) of reference alignments from the four benchmarks. Table 1 presents empirical properties for the reference alignments for these 1192 data sets, including average pairwise sequence identity (PID), average sequence length, average number of sequences, average percentage gapped, and mean gap length.

Simulated data sets.—We generated 120 simulated data sets (20 data sets from each of 6 different model conditions) to evaluate the alignment methods for this study. To obtain the basic model tree topology and branch lengths, we selected the 27-sequence serine protease data set from the Homstrad benchmark collection, computed a MAFFT L-ins-i alignment on the data set, and then used RAxML v8.2.9 to construct a phylogenetic tree with branch lengths (see Supplementary Section 1.1 available on Dryad, for exact command). RAxML selected the WAG model for this data set. We set the indel rate and the gap length distribution (a negative binomial) to match the empirical distribution for the serine protease data set. We then modified this basic model tree in two ways—by rescaling the branch lengths (by a factor of three) and reducing the indel rate—to produce six different model conditions (Table 2) that ranged in terms of the average percent gapped (from 18.3% to 46.4%) and average pairwise

TABLE 2. Empirical properties of the true alignments for the simulated data sets, each with 27 sequences

			Low subst. rate	High subst. rate
Indel rate	High	PID	0.24	0.11
		% gapped	46.4	42.6
	Medium	PID	0.24	0.11
		% gapped	29.8	31.5
	Low	PID	0.23	0.12
		% gapped	18.3	19.2

Note: Each submatrix represents one of the six model conditions, and the top row within each submatrix represents the mean pairwise sequence identity (PID) and the bottom row represents the percentage gapped.

sequence identity (from 0.11 to 0.24). Hence, this process produced six different model conditions with a range of average PID and percentage gapped that cover the characteristics of the biological benchmark data sets we explored. The sequence length at the root is 200, and then sequences evolve down the model tree with indels and substitutions using the Indelible (Fletcher and Yang 2009) simulator. We used WAG (Whelan and Goldman 2001) for the substitution model, and indels were generated with lengths drawn from a negative binomial distribution.

Evaluation Criteria

The accuracy of the estimated alignment was assessed in comparison to the reference alignment for the biological data sets, and to the true alignment for the simulated data sets. Each alignment on the same set of sequences can be represented by its set of "homology pairs," where a homology pair is a pair of residues, one from each of two different sequences, that are placed in the same column in the alignment (see Mirarab and Warnow 2011, 2017). Two different alignments can then be compared to each other by examining the shared or unique homology pairs. Furthermore, when one alignment is treated as a reference or true alignment, then the error in an estimated alignment can be evaluated by calculating the number of homology pairs in the true alignment that are missing from the estimated alignment (i.e., the number of false negatives) as well as the number of homology pairs in the estimated alignment that do not appear in the true alignment (i.e., the number of false positives). These error metrics are normalized to produce values between 0 and 1, which are then called the error rates. The first of these error rates is referred to as the sum-of-pairs false negatives (SPFN) score and the second is referred to as the sum-of-pairs false positives (SPFP) score. Finally, the error rates can also be expressed as accuracy measures in the obvious way: 1-SPFN is a measure of recall, and is referred to as the SP-Score, and 1-SPFP is a measure of precision, and is referred to as the Modeler Score.

We also report the expansion ratio, which is the ratio of the number of sites in the estimated alignment to the number of sites in the reference or true alignment; values below 1.0 represent overalignment (i.e., shorter alignments than the reference or true alignment) and values greater than 1.0 represent underalignment. We used FastSP v. 1.6.0 (Mirarab and Warnow 2011) to calculate these values. Note that the classical tradeoff between the false positive rate (FPR) and false negative rate (FNR) has an analog in multiple sequence alignment as well: just as a classifier can achieve zero FPR by classifying everything as negative, an MSA can have zero SPFP if the sequences are completely unaligned (i.e., all sites in the alignment have at most one nongap character). That is the extreme case, of course, but serves to indicate that although expansion ratios greater than 1.0 reflect underalignment, a pattern of low SPFP and high SPFN (equivalently high Modeler score and low SP-score) is also indicative of underalignment.

Finally, we examined the impact of alignment error on tree error. We computed maximum likelihood trees using RAxML v8.2.9 on estimated and true alignments for the simulated data sets, and then recorded the Robinson–Foulds (Robinson and Foulds 1981) error rate (i.e., the fraction of the number of branches in the true tree that are missing from the estimated tree), computed using Dendropy (Sukumaran and Holder 2010).

The impact of pairwise sequence identity (PID) on multiple sequence alignment accuracy is well established (e.g., Thompson et al. 1999; Blackshields et al. 2006; Liu et al. 2009; Sievers et al. 2011; Wang et al. 2011), with alignment accuracy generally decreasing as PID decreases, and expected to be very low when PID is below 0.20 (Aniba et al. 2010). Therefore, we evaluated the impact of PID on alignment accuracy in our experiments. We grouped the sequence data sets into four bins according to the PID within each data set, with the highest PID bin (where PID is at least 0.5) expected to contain the easiest data sets to align and the lowest PID bin (where PID is at most 0.15) expected to contain the most difficult data sets to align.

RESULTS

Results on Biological Data Sets

We began by exploring the overall accuracy of the different methods we examined with respect to Modeler score and SP-score (Fig. 1). The results shown are restricted to the 1192 data sets where all methods ran successfully.

There was a large range in scores on these data, with the average Modeler score varying from 0.66 to 0.80 and average SP-score varying from 0.63 to 0.77. PROMALS, T-Coffee, CONTRAlign, and MAFFT-homologs each came in the top four places for both criteria, and Kalign, DiAlign, and PRANK had the lowest overall SP-scores and Modeler scores of all the methods we tested. BAli-Phy had the top average Modeler score (0.80) but among the lowest average SP-scores of all the methods (0.67, ranking 11 out of 14); thus, BAli-Phy's average Modeler score was substantially larger than its SP-score, a pattern that indicates underalignment. All other methods had

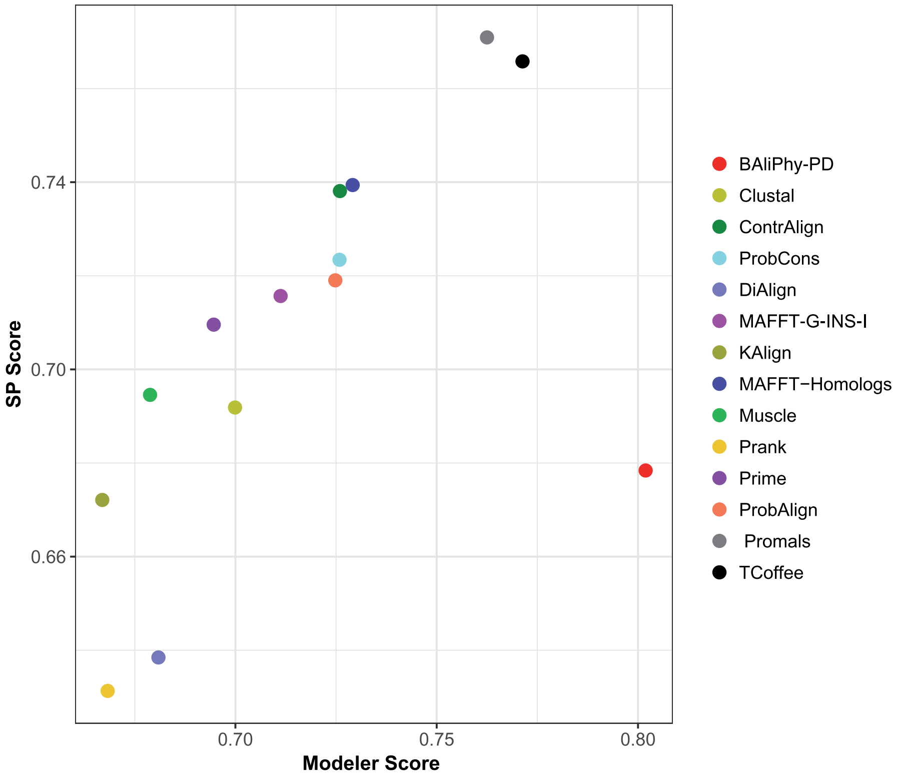

FIGURE 1. Average Modeler Score (i.e., precision) versus average SP-score (i.e., recall) of the full set of multiple sequence alignment methods on the biological benchmark data sets, each with at least 4 and at most 25 sequences; each data point represents analyses of 1192 data sets from the four benchmark collections (658 from BAliBase, 231 from Homstrad, 202 from Mattbench, and 101 from Sisyphus). See Supplementary Table S1 and Excel File available on Dryad for actual numeric values.

close average SP and Modeler scores (i.e., differences that were at most 0.04, and usually at most 0.01).

Results on the individual benchmarks for Modeler and SP-scores show similar trends (Fig. 2). For example, T-Coffee had the highest SP-scores on the Homstrad (0.89), Mattbench (0.78), and Sisyphus (0.80) benchmarks, and PROMALS had the highest SP-score (0.74) on the BAliBASE data (where T-Coffee had 0.71). Thus, T-Coffee was either best or close to best in terms of SP-score on these data sets. Similarly, although PROMALS was only top on the BAliBASE data sets, it came in second on the other benchmarks, where its average SP-scores were fairly close to the best score: 0.01 lower than best on the Homstrad data sets, 0.03 lower than best on the Sisyphus data sets, and 0.06 lower on the Mattbench data sets. MAFFT-Homologs had the second or third highest SP-score on all but the Sisyphus benchmark, and the third or fourth highest Modeler score on three of the benchmarks. Finally, BAli-Phy consistently ranked among the first three methods for Modeler Score (it was top on BAliBASE, in second place on Sisyphus and Mattbench, and in third place on Homstrad) and between eighth and twelfth for SP-score (Fig. 2). Thus, overall as well as on the individual benchmarks, BAli-Phy produced alignments with high Modeler scores and low SP-scores, a pattern that indicates underalignment.

To better understand this trend, we examined the expansion ratios of the different alignment methods. A few methods (notably Clustal, Probcons, and Promals) had excellent expansion ratios (in the range 0.95 to 1.05) across all PID values. However, all others either underaligned (i.e., expansion ratios greater than 1.05) or overaligned (i.e., expansion ratios less than 0.95) for some condition (Fig. 3). As expected, the lowest PID condition (PID $\geq$0.5) was the most challenging for the remaining methods. For this condition, BAli-Phy, DiAlign, and Prank underaligned the most, with expansion ratios 1.87, 1.31, and 1.17 respectively. The methods that over-aligned the most were Muscle (expansion ratio 0.81), Prime (expansion ratio 0.84), Kalign (expansion ratio 0.86), MAFFT-G-INS-i, MAFFT-Homologs, and CONTRAlign (all with expansion ratio 0.89). Most significantly, BAli-Phy's expansion ratio was the largest of all the methods for each bin, indicating that it underaligned the most of all the methods and produced substantially longer alignments than all other methods. Hence, BAli-Phy was an outlier among these methods in terms of alignment length.

The remaining experiments were restricted to the top-performing alignment methods. Therefore, we exclude Kalign, DiAlign, and PRANK, each of which had among the lowest overall accuracy in terms of SP-score and Modeler score on the biological data sets.

PID also impacted the SP-score and Modeler score of the top methods, as shown in Figures 4 and 5. As expected, all methods had their best SP-scores and

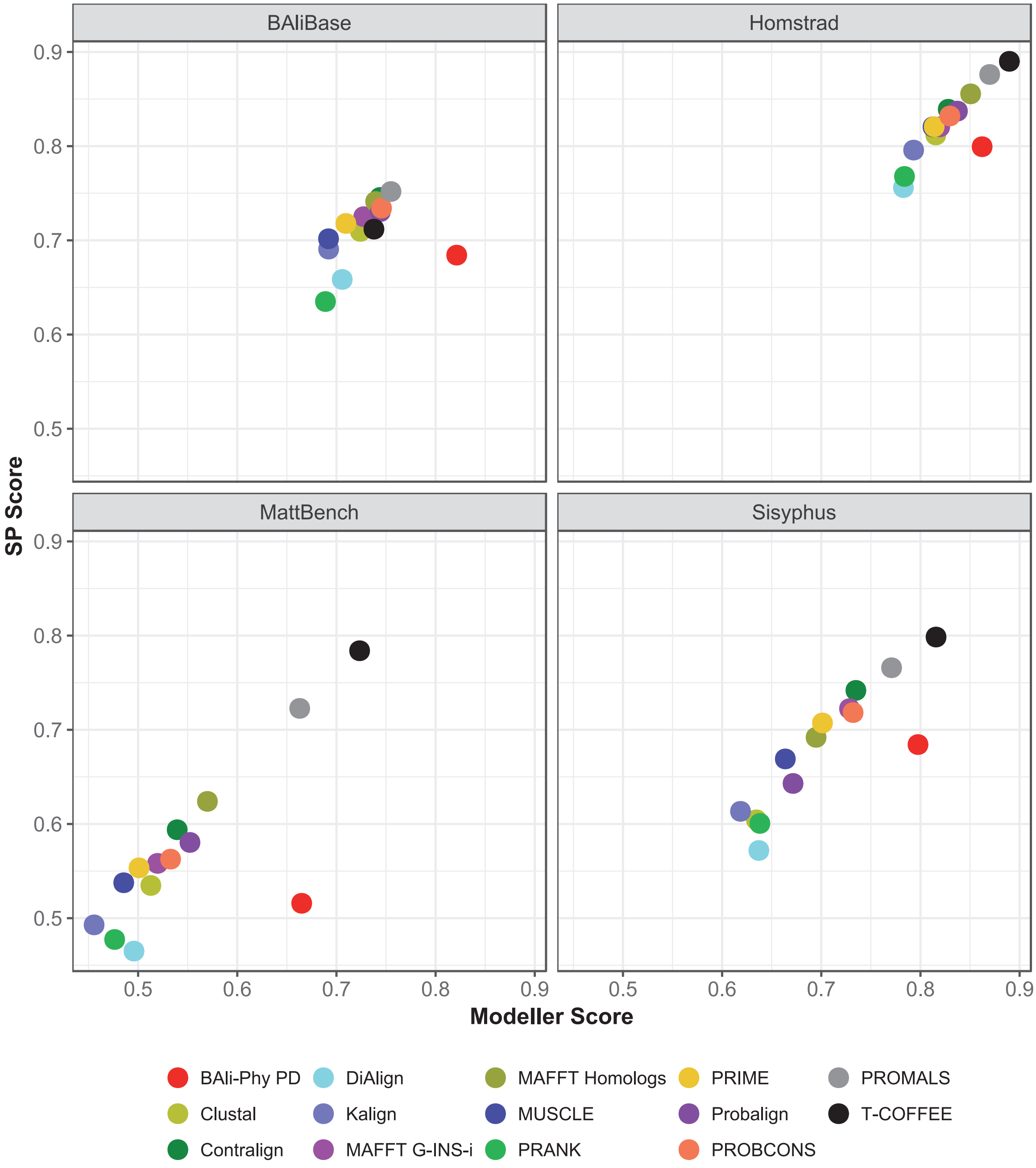

FIGURE 2. Average Modeler Score (i.e., precision) versus SP-Score (i.e., recall) of all alignment methods on the individual biological benchmarks. Results shown are for 1192 data sets from the four benchmark collections (658 from BAliBase, 231 from Homstrad, 202 from Mattbench, and 101 from Sisyphus) See Supplementary Table S2 and Excel File available on Dryad for actual numeric values.

Modeler scores under the highest PID bin (i.e., when PID ≥ 0.5) and their scores dropped as PID decreased. The range in SP-scores and Modeler scores was narrowest for the highest PID bin (at most 0.05 difference between the largest and smallest scores) and increased as PID dropped. For example, on the highest PID bin (Fig. 4), the Modeler scores ranged from 0.91 (attained by T-Coffee) to 0.95 (attained by BAli-Phy), while on the lowest PID bin the Modeler scores ranged from 0.26 (Clustal) to 0.53 (BAli-Phy). Furthermore, although the Modeler scores dropped for all methods as PID dropped, this effect was smaller for BAli-Phy, Promals, and T-Coffee than for the other methods (i.e., the change in average Modeler score between the top and bottom PID bins for these three methods was at most 0.47, and all other methods had changes of between 0.59 and 0.68).

Similar trends hold for SP-score (Fig. 5), with the following main difference: under the low PID bin, BAli-Phy's SP-score tied for the lowest of all methods, indicating it is more impacted by changes in PID than

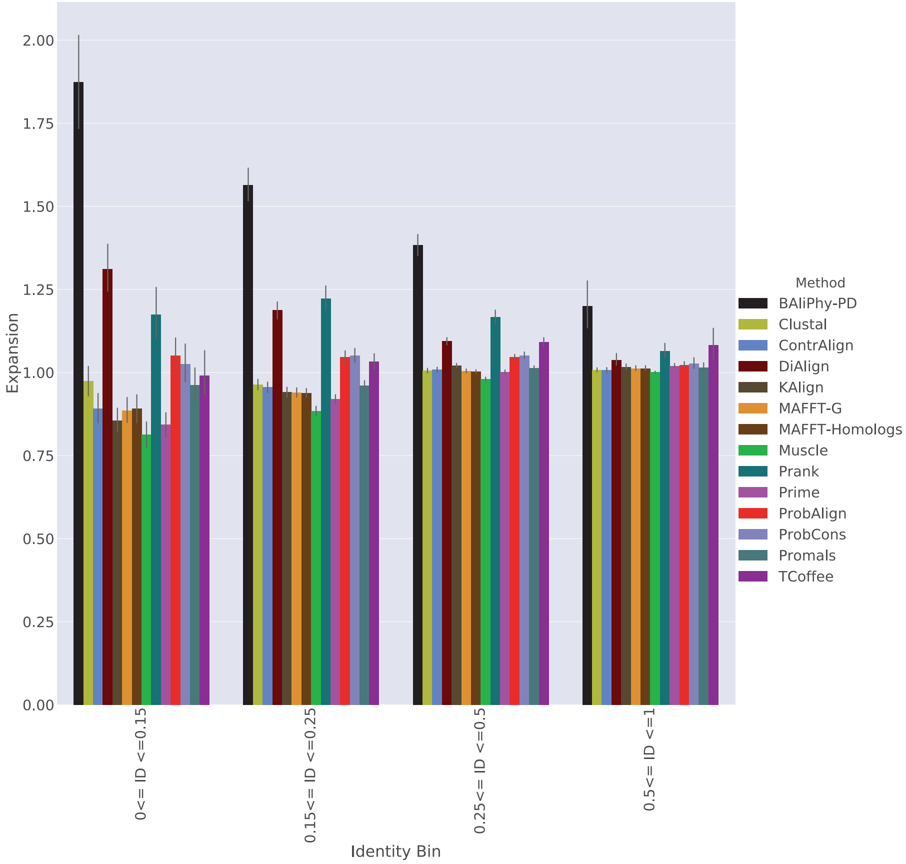

FIGURE 3. Average expansion ratios on the 1192 biological benchmark data sets, each with at most 25 sequences, by average percent ID (ID). Values more than 1.0 indicate underalignment (i.e., longer alignments than the reference alignment), while values less than 1.0 indicate overalignment (i.e., shorter alignments than the reference alignment). The four bins based on average sequence identity, ordered from smallest to largest, have 83, 417, 615, and 77 alignments, respectively. See Supplementary Table S3 and Excel File available on Dryad for actual numeric values.

we saw for its Modeler score (in particular, the change in BAli-Phy's SP-scores between the high and low PID bins was 0.64, which is approximately the same change as for the remaining methods other than Promals and T-Coffee). Finally, for the two bins where the differences between methods were large (i.e., the bottom two bins), T-Coffee and PROMALS had the top SP-scores.

Results on Simulated Data Sets

Methods that rely on gathering homologs from external databases (e.g., MAFFT-Homologs, T-Coffee, and Promals) are expected to have poor accuracy on these simulated data, a prediction we confirmed (see Supplementary Section 2.2 available on Dryad). We therefore omit these three methods for the rest of this section, but we include PRANK since, like BAli-Phy, it is a phylogeny-aware method.

We explored the relative and absolute accuracy of the multiple sequence alignment methods on simulated data sets with 27 sequences with 6 model conditions, each with 20 replicates. The accuracy of these methods varied across these six model conditions (Fig. 6). When both rates are low, all methods had excellent Modeler and SP-scores (i.e., at least 0.95) and the differences between them were small (e.g., the difference in score between any two methods under the easiest model condition was at most 0.02 for both criteria). However, with higher substitution rates or indel rates, the accuracy of all methods decreased and the range in scores increased.

The most striking observation on the simulated data sets is that BAli-Phy had the best accuracy of all methods with respect to both criteria. Furthermore, while the difference between BAli-Phy and the least accurate method was small (at most 0.02) for the easiest model condition, the difference in accuracy between BAli-Phy and the *second most accurate method* increased as the indel rate or the substitution rate increased. For example, under the most difficult model condition (where substitution and indel rates were the highest), BAli-Phy achieved an average SP-score of 0.93 and an average Modeler score also of 0.93; the second best SP-score was 0.87 (attained by Clustal) and the second best Modeler score was 0.84 (attained by MAFFT-G-ins-i). These are drops in accuracy in the 0.07 to 0.09 range (see Supplementary Table S8 and Excel File available on Dryad).

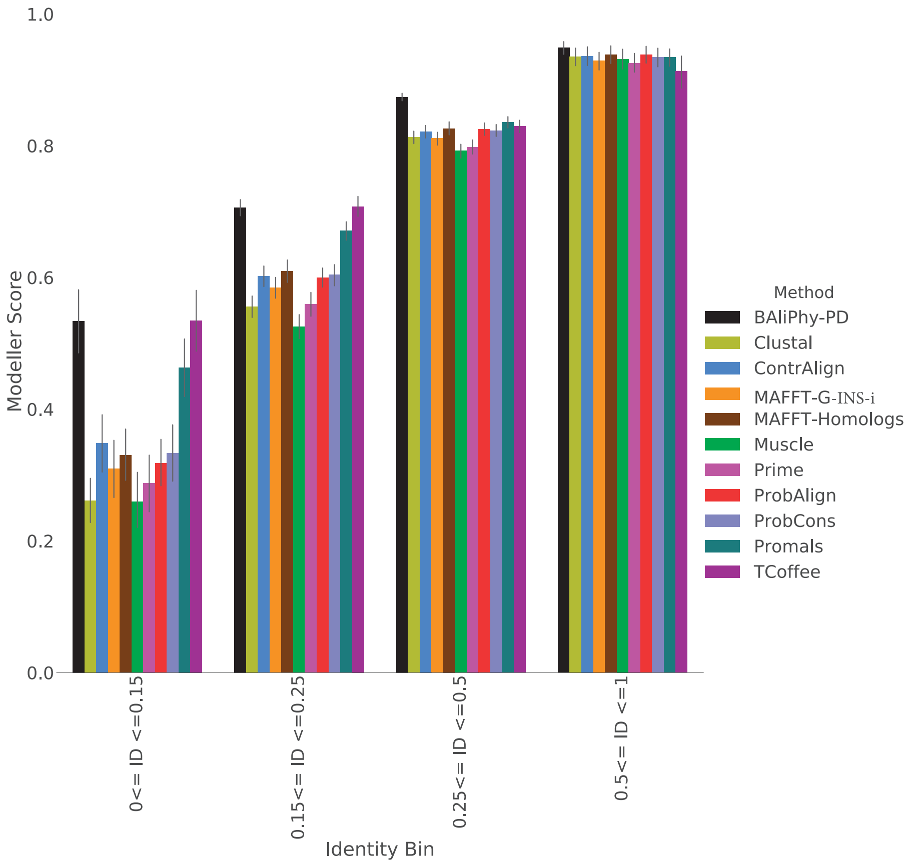

FIGURE 4. Average Modeler Scores (i.e., precision) for the top methods on the 1192 biological benchmark data sets, binned into different average pairwise sequence identity (ID) levels. The four bins based on average sequence identity, ordered from smallest to largest, have 83, 417, 615, and 77 alignments, respectively. See Supplementary Table S4 and Excel File available on Dryad for actual numeric values.

As shown in Figure 7, similar trends were seen with respect to expansion ratios. Results under the easiest model condition (with low mutation rates and indel rates that were at most moderate), all methods produced expansion ratios in the range 0.97–1.01 (i.e., nearly perfect). However, under the more difficult model conditions, the methods could be distinguished and we observed the following overall trends. BAli-Phy produced alignments with expansion ratios of 1.0 under all six model conditions (except when given the wrong substitution model, in which case it produced alignments with average expansion ratio 0.99). Most other methods overaligned under difficult conditions (e.g., Muscle, MAFFT G-INS-i, CONTRAlign, Clustal, and PRIME overaligned when mutation rates and indel rates were high, with expansion ratios less than 0.90). The three remaining methods (PRANK, Probalign, and ProbCons) showed somewhat different responses. PRANK tended to underalign (with an expansion ratio of 1.5 for the most difficult condition where both indel and mutation rates were high) and Probalign underaligned whenever substitution rates were high (expansion ratio of 1.06–1.08), overaligned for the condition with high indel rates and low substitution rates (expansion ratio of 0.92), and had nearly perfect expansion ratios (in the 0.98–0.99 range) for the remaining two conditions. ProbCons had excellent expansion ratios (in the range 0.97–1.0) under five of the six conditions, but overaligned (expansion ratio 0.91) when indel rates were high and substitution rates were low. Thus, of these methods, only BAli-Phy had consistently excellent expansion ratios under all six model conditions.

The performance of PRANK on the simulated data was generally not as strong as some of the other methods. Under the most difficult model condition, PRANK had the lowest average Modeler and SP-scores (0.65 and 0.52, respectively), and produced the longest alignments (with expansion ratio 1.5) of all the tested methods. However, under the two easiest conditions (low substitution rates with low or moderate indel rates), PRANK produced alignments that were very close to the correct length (expansion ratios between 0.96 and 1.0) and had SP-scores and Modeler scores of at least 0.96; it even achieved average SP-score and Modeler score of 0.92 and 0.93, respectively, for the simulated data under the low substitution rate with high indel rate. Thus, PRANK's accuracy was impacted by the substitution rate: it was competitive with the better methods under conditions with low substitution rates but not when substitution rates were high. Also, when

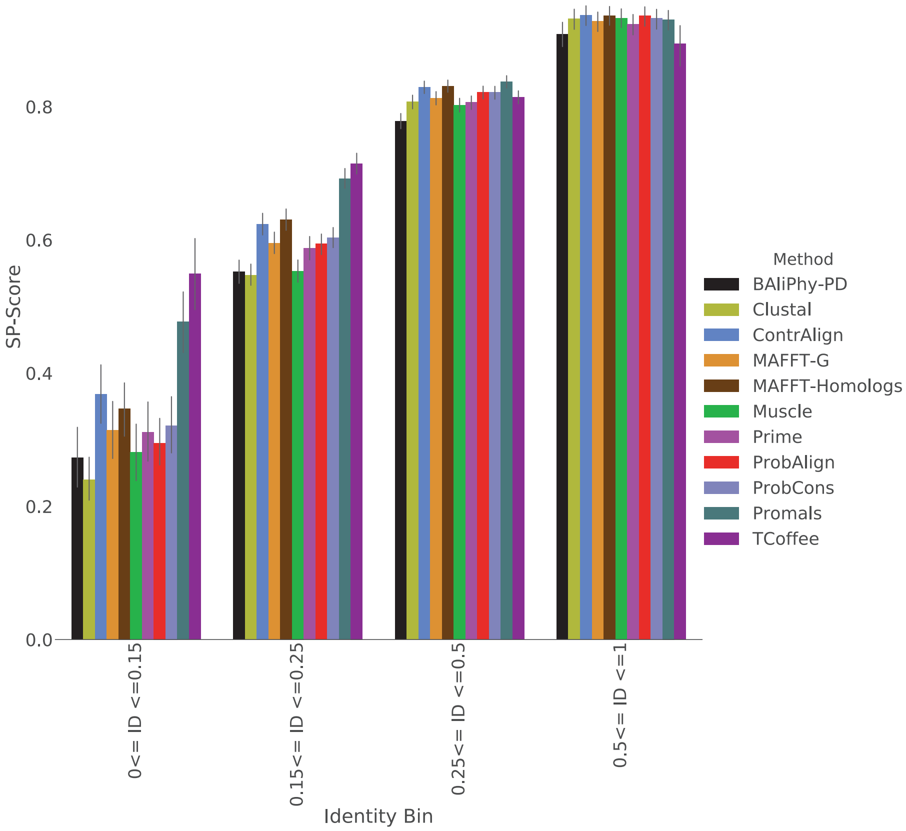

FIGURE 5. Average SP-scores (i.e., recall) for the top methods on the 1192 biological benchmark data sets, with data sets binned by average pairwise sequence identity (ID) levels. The four bins based on average PID, ordered from smallest to largest PID values, have 83, 417, 615, and 77 alignments, respectively. See Supplementary Table S5 and Excel File available on Dryad for actual numeric values.

PRANK was given the true (model) tree as a guide tree, these scores increased (in one case by 0.10, see Supplementary Table S8 and Excel File available on Dryad), but not enough to change its ranking within the experiment (e.g., PRANK used with the true tree was still in the bottom position for both SP-score and Modeler score under the most difficult model condition).

Running Time

We also did a small evaluation of the running time of a sample of the alignment methods. As noted, we always ran BAli-Phy for 48 h on 32 independent runs, in order to improve the chances of convergence. Hence, the total running time for BAli-Phy always exceeded 2 months on each data set. In other words, the way we ran BAli-Phy is by design computationally intensive.

We selected four data sets (one from each of the benchmark collections), each containing 17 sequences. This comparison is meant to be approximate, as we used different platforms for the methods and did not ensure that all methods were run using the same environments, and only examined four data sets; hence, the results are not necessarily indicative of running time on other data sets. T-Coffee and BAli-Phy were run on the National Center of Supercomputing Applications Blue Waters supercomputer and the rest of the methods were run on the Campus Cluster at the University of Illinois at Urbana-Champaign. Some of these methods were compiled from the source code, and we used the precompiled versions for other methods.

As shown in Table 3, BAli-Phy was the most computationally intensive of all the methods. T-Coffee was the next most computationally intensive, using from 7 to 59 min on these four data sets. PROMALS and PRANK were faster than T-Coffee, but each was slow on at least one data set: PROMALS used 24 min on one data set and PRANK used 4 min on another. All the others were much faster, never using even a full minute on any of the four data sets, and several of these (i.e., DiAlign, PRIME, Clustal, Muscle, and MAFFT-G-ins-i) never exceeded 2 s on any data set.

Although this was a limited study, the methods that were very fast on these data are likely to remain very fast for other data sets with similar characteristics (number of sequences and average sequence length), under other modern computational platforms. On the other hand, although we ran BAli-Phy 32 times, each for 48 h, similar accuracy might have been obtained from a reduced number of hours or number of independent runs; also, the new version of BAli-Phy (v.3.1.5) may converge more

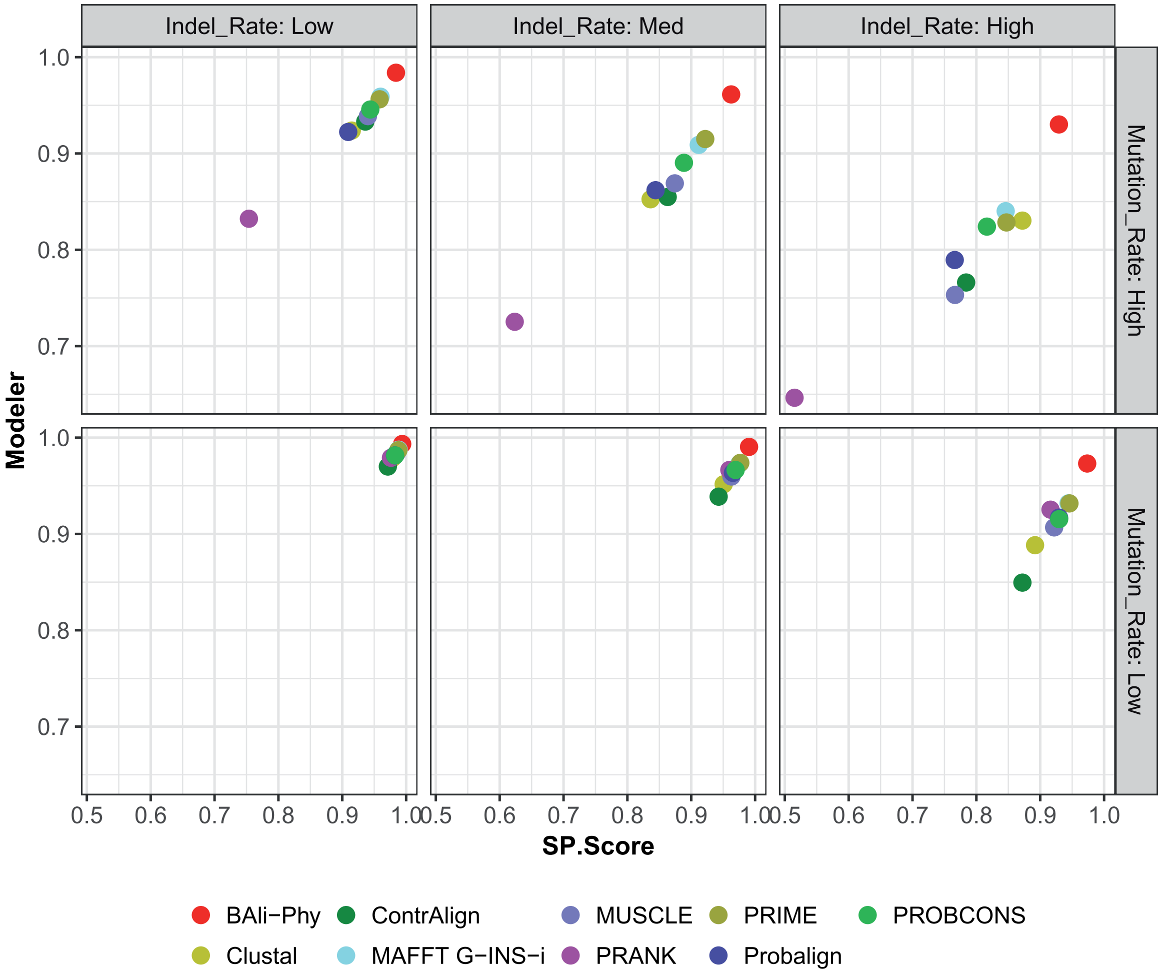

FIGURE 6. Modeler score (i.e., precision) versus SP-Score (i.e., recall) for MSA methods on simulated amino acid data sets with 27 sequences for 6 different model conditions that vary by the substitution rate and indel rate; averages over 20 replicates are shown. See Supplementary Excel File available on Dryad for actual numeric values.

quickly than the version we used (v.2.3.8). Thus, these running time values are not meant to be used to predict running time on other data sets or on other platforms, but mainly only to show that some of the better methods (e.g., MAFFT-G-ins-i) were very fast, and much faster than some of the other methods that also had very good accuracy.

Impact on Tree Estimation

Alignment estimation is known to have an impact on tree estimation (Dessimoz and Gil 2010; Wang et al. 2011; Mirarab et al. 2015), and so we explored this issue as well. We evaluated the topological error of maximum likelihood trees computed using RAxML v8.2.9 on the true alignment and on estimated alignments. We did not explore the impact on tree error on the biological data sets because true trees are unknown, and the true species tree can differ from the true gene tree as the result of multiple biological processes, including incomplete lineage sorting (Maddison 1997).

We let RAxML select the protein substitution model for each data set (see Supplementary Section 1.1 available on Dryad) and report the normalized Robinson–Foulds (RF) error for the single best ML tree found by RAxML. We report the normalized RF error rates in Supplementary Figure S1 available on Dryad and Delta-RF (the increase in error rate resulting from using an estimated alignment instead of the true alignment) in Supplementary Table S9 available on Dryad.

Under the model conditions with low mutation rates, all the methods had good accuracy, with Delta-RF error rates that were at most 1%. However, under the conditions with high substitution rates, the methods could be clearly distinguished (Supplementary Table S9 available on Dryad). For example, under the hardest model condition (where the indel and substitution rates were both high), the Delta-RF rates were 28% for Clustal, 20% for PRANK, 9% for Probalign, 7% for ProbCons, and 4% for Muscle; in contrast, BAli-Phy and PRIME had 1%, and MAFFT-G-ins-i had 0% Delta-RF rate. More generally, for all conditions, ML trees computed on the

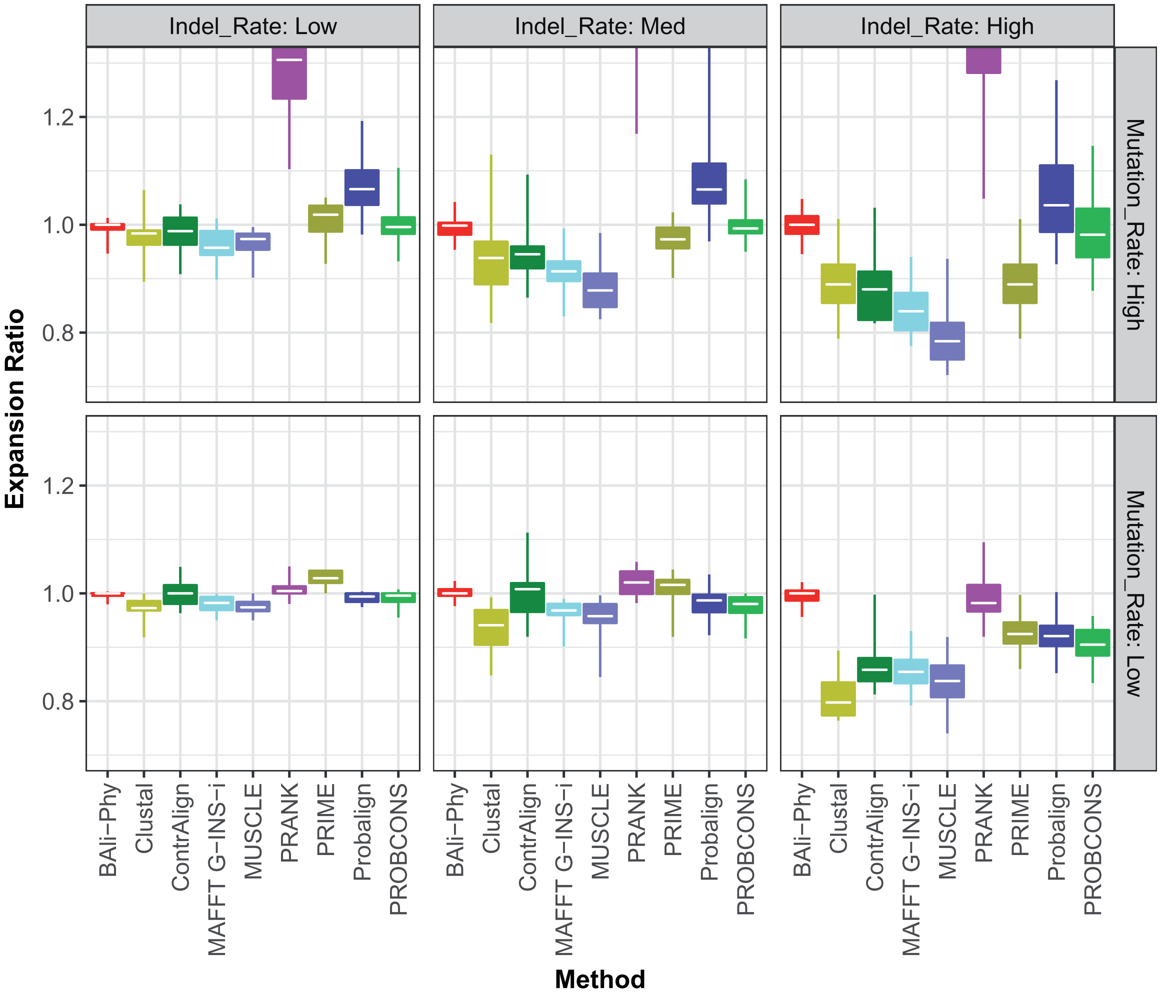

FIGURE 7. Box plot showing expansion ratios (1.0 is perfect, ratios below 1.0 indicate overalignment, and ratios above 1.0 indicate underalignment) for MSA methods on simulated amino acid data sets with 27 sequences for 6 different model conditions that vary by the substitution rate and indel rate; averages over 20 replicates are shown. Lines represent means and the lower and upper hinges of the box represents first and third quartiles; the upper whisker is the maximum value, and the lower whisker is the minimum value. See Supplementary Table S8 and Excel File available on Dryad for actual numeric values.

BAli-Phy, MAFFT-G-ins-i, and PRIME alignments had Delta-RF at most 1%, and so were very close in accuracy to ML trees computed on the true alignment. Thus, BAli-Phy came in among the top alignment methods with respect to topological accuracy of maximum likelihood trees computed on these simulated alignments.

DISCUSSION

Although our study was restricted to amino acid data sets with at most 27 sequences, the following trends were consistently observed. The best Modeler and SP-scores were obtained for the high PID conditions, and this held for both types of data (simulated and biological) and for all methods. In addition, SP-scores and Modeler scores decreased as PID decreased. We also saw that the expansion ratios were very close to 1.0 for high PID conditions, but when PID was low the expansion ratios could be far from 1.0. Similarly, our simulation study showed that under the low substitution rate conditions (where PID was moderate at 0.24) then alignment error did not have a noteworthy impact on tree estimation (i.e., maximum likelihood trees estimated on estimated alignments were on average within 1% Robinson–Foulds error of the maximum likelihood trees estimated on the true alignment); however, under the high substitution rate conditions (where PID was low at 0.11) then maximum likelihood trees for some estimated alignments (e.g., Clustal and PRANK) were very far from the maximum likelihood tree computed on the true alignment. Thus, decreases in PID resulted in decreases in the accuracy (for all three alignment criteria we evaluated) of alignment methods and also resulted in increases in the error of trees computed on estimated alignments. This reduction in accuracy under low PID conditions explains why some biological benchmarks were more difficult than others. For example, all alignment methods had lower average Modeler and SP-scores on Mattbench than on the other benchmarks, and

TABLE 3. Running time (in seconds) information of a single 17-sequence data set in each of the biological benchmarks for a sample of the alignment methods, with methods roughly sorted by running time from fastest to slowest

Benchmark	Mattbench	Homstrad	Sisyphus	BAliBASE
Data set	SF054	Proteasome	AL00048098	BALBS213
Max. Seq. Len.	*270*	*250*	*117*	*688*
DiAlign	0.0	0.0	0.0	0.0
PRIME	0.1	0.0	0.0	0.0
Clustal	0.4	0.3	0.1	1.5
Muscle	0.5	0.4	0.1	1.0
MAFFT-G-INS-i	0.7	0.7	0.3	2.0
Probalign	1.7	1.4	0.4	7.9
ProbCons	3.1	2.6	0.6	12.6
CONTRAlign	5.8	6.2	1.4	42.0
PRANK	48.5	1:16.1	9.4	4:14.7
PROMALS	14:11.5	12:22.1	5:06.2	24:03.2
T-Coffee	46:47.2	58:04.7	7:06.5	59:18.8
BAli-Phy	48:00:00.0	48:00:00.0	48:00:00.0	48:00:00.0

Note: The running times are rounded to the nearest hundredth of a second, and reflect wall clock time. The time reported for most methods is based on a single processor. However, BAli-Phy was run 32 independent times, each for 48 h (in order to improve the chances of convergence) but the running time reported is for a single run; MAFFT uses 4 threads, and Clustal uses 12 threads.

the average PID for the Mattbench data sets (0.20) is the lowest of the four biological collections we analyzed. Similarly, the Homstrad data sets have the highest average PID (0.37) of all these benchmarks, and the Modeler and SP-scores were highest on these data sets.

Another consistent trend throughout this study is that the differences between methods in terms of SP-score, Modeler score, and expansion ratio increased as PID decreased. Furthermore, under the high PID data sets, the differences between methods are very small, making distinctions between methods more difficult, but methods were easily distinguished on the low PID conditions. These trends suggest that the choice of alignment method may have little impact when PID is high but can be important when PID is low. The impact of PID on alignment accuracy and downstream analyses have been observed before (e.g., Blackshields et al. 2006; Liu et al. 2009; Sievers et al. 2011), so these observations confirm prior studies.

The best performing methods on the biological data sets were typically T-Coffee and PROMALS (although the relative performance depended on the PID level and the criterion). For example, T-Coffee had the highest average SP-scores for the low PID data sets but not for the high PID data sets where PROMALS and many other methods had higher SP-scores. MAFFT-homologs and CONTRAlign also had good Modeler and SP-scores on the biological data sets, coming in the first four positions for all benchmarks. The good overall performance of MAFFT-homologs, PROMALS, and T-Coffee is noteworthy since these methods share a common strategy of recruiting homologs from an external database to use in the alignment task. Finally, BAli-Phy produced the best Modeler scores but came in at position 11 (out of 14) for its SP-score.

Results on the simulated data sets showed different trends: as they are inherently unsuited for simulated data, T-Coffee and PROMALS were not among the better methods for SP-score or Modeler score, and BAli-Phy had better scores than all the other methods for both criteria. Hence, the relative performance of methods seems to depend on PID, the criterion (i.e., Modeler score or SP-score), and—to some extent—whether the data were biological or simulated. In particular, our study shows that BAli-Phy, a leading statistical method for coestimating alignments and trees, had the best Modeler scores and SP-scores of all the methods we examined on simulated data sets but lower SP-scores than many methods on the biological data sets.

To understand this difference in performance, it is helpful to consider the tendency of methods to either underalign (i.e., produce alignments that are longer than the reference alignment) or overalign (i.e., produce alignments than are shorter than the reference alignment). Our study shows that that many methods tended to overalign (producing expansion ratios substantially less than 1.0) under challenging conditions; the major exceptions to this were BAli-Phy (which underaligned the most of all methods), DiAlign, and PRANK (some other methods also underaligned but to lesser degrees). Interestingly, in contrast to the other alignment methods, BAli-Phy never underaligned on the simulated data, even under the most challenging conditions. Underalignment is also demonstrated by higher Modeler scores than SP-scores, a trend consistently produced by BAli-Phy on the biological data (where the overall gap was 0.13), but never on the simulated data (where BAli-Phy had average Modeler and SP-scores that were within 0.01 for every model condition). In other words, our data show that BAli-Phy underaligned on the biological data with respect to the reference alignment, but did not underalign on the simulated data with respect to the true alignment. The fact that BAli-Phy underaligned on biological data but not on simulated data explains the change in performance for BAli-Phy between biological and simulated data.

The performance of PRANK in our study is interesting to consider, since PRANK is designed to be "phylogeny-aware," and so has some similarities to BAli-Phy in terms of approach. On biological data Prank produced slightly higher Modeler scores than SP-scores (but on average within 0.04 of each other); on the simulated data Prank also produced larger Modeler scores, but the gap was larger (0.15), at least for the most difficult model condition. Prank underaligned on both simulated and biological data, but the degree to which it underaligned was larger on the simulated data. Thus, like BAli-Phy, PRANK tended to underalign on the biological data and responded differently to the biological and simulated data. However, PRANK was not competitive with the better methods in our study on either the biological or simulated data for any criterion, while BAli-Phy generally had the best (or close to the best) Modeler scores under all conditions, and only had reduced SP-scores on the biological data. As we have seen, PRANK had very good accuracy (even if not the best accuracy) under conditions with high PID, but relatively poor accuracy (compared to the better methods) under the low PID conditions, such as occur under high rates of evolution. PRANK's reduced accuracy on the simulated data sets with lower PID is perhaps surprising, given that PRANK had superior alignment accuracy in prior simulation studies (Löytynoja and Goldman 2008). However, a careful examination of Löytynoja and Goldman (2008) reveals that the simulation conditions in which PRANK provided outstanding accuracy had substitutions operating under the simplest model (Jukes–Cantor with a strict molecular clock), which may have favored PRANK in some way.

Conclusions

Statistical sequence alignment, and in particular statistical coestimation of multiple sequence alignments and phylogenetic trees under phylogenetic models of sequence evolution, has been considered by many to be the most rigorous approach to alignment estimation. This study examined the accuracy of BAli-Phy, a leading method for statistical coestimation of alignments and trees, on both biological and simulated data under a range of model conditions, and explored the impact of alignment accuracy on tree accuracy.

Our study shows that BAli-Phy has the best (or close to best) accuracy of all methods for all criteria we examined (SP-score, Modeler score, expansion ratio, and accuracy of maximum likelihood trees estimated on the alignment) for the simulated data; however, BAli-Phy underaligns on biological data to a sufficient extent that its overall SP-score drops to 11th place (out of 14) even though its Modeler score remains in top place. In other words, BAli-Phy has superior accuracy on simulated data but mixed accuracy on biological data caused by underalignment. We do not know why BAli-Phy exhibits this difference in performance between biological and simulated data.

Understanding this distinction in performance requires some care as there are multiple possible explanations, including the distinctions between evolutionary and structural alignments, the potential for model misspecification between the model assumed in BAli-Phy and how proteins evolve, and the possibility that reference alignments could have errors (Aniba et al. 2010; Iantorno et al. 2014). However, each explanation is potentially valid, and each may contribute to a greater or lesser degree to this distinction in performance. Furthermore, interactions between these and other factors could also be contributing to the differences we observed.

The first potential explanation is that the reference alignments in the biological benchmarks are accurate as structural alignments but not as evolutionary alignments. This explanation is consistent with the argument made by some authors (notably Iantorno et al. 2014; Chatzou et al. 2015) that the two types of alignments have different objectives and that an alignment may be correct in terms of structural features and yet be incorrect in terms of evolutionary descent (or vice-versa); in addition, the argument has been made that similarities between sequences based on shared structures, even if substantial, may not be due to descent from a common ancestor (Reeck et al. 1987). For example, as noted by (Chatzou et al., 2015), convergent evolution could lead to proteins having very similar or even identical structural features, as well as potentially the same functions, and alignments based on these structures will not always reflect descent from a common ancestor. If this is the major cause for this discordance, then this study would suggest that even if BAli-Phy is suitable for alignments used in phylogenetic inference, it may not be suitable for alignments used to predict protein structures. However, others (e.g., Dover 1987) have argued that a high degree of structural similarity should nearly always indicate true homology, suggesting that the true structural alignment ought to be very close to (and perhaps the same as) the true phylogenetic alignment; if Dover (1987) is correct in this assertion, then the distinction between structural and phylogenetic alignments is unlikely to be the main cause for the discordance we observe.

The second potential explanation is that the reference alignments are accurate evolutionary alignments, but the model assumed by BAli-Phy is a poor match to the true model under which the proteins evolve. There are many critiques of sequence evolution models used in phylogeny estimation (Liberles et al. 2012; Wilke 2012) and in simulation studies (Boyce et al. 2014; Iantorno et al. 2014), with two of the major concerns being the assumption that the sites evolve identically and independently (the *i.i.d.* assumption) and without any selection occurring. Although the model underlying BAli-Phy is more complex than the standard models discussed in these papers in that it addresses insertions and deletions (i.e., indels) rather than only substitutions, the BAli-Phy model nevertheless also has those two problematic features (*i.i.d.* site evolution and no selection operating) that are clearly violated by protein sequence evolution. If the degree of misspecification between the model in BAli-Phy and how proteins actually evolve is sufficient to explain much of the distinction in performance between BAli-Phy on biological and simulated data sets, then phylogeny estimation under standard models may also be impacted since many genomic regions (e.g., protein-coding sequences) are acted on by selection and evolve under processes that are not *i.i.d.*

Finally, since the structural alignments in the four benchmark collections are estimated rather than known, the accuracy of the biological benchmark alignments can be questioned (Aniba et al. 2010; Iantorno et al. 2014). It is therefore interesting to consider the possibility that the reference alignments, since they are estimated using a combination of manual and automated techniques, may themselves be overaligned; in this case, the true alignment would have a high Modeler score and a low SP-score with respect to the reference alignment, which is what we see with BAli-Phy on the biological data sets. Since we observed the same general trends across all four benchmark collections, whatever the issues are, they are likely to be impacting each collection rather than just one. If some of the reference alignments are incorrect, then more accurate structural alignments would need to be developed in order to provide reliable benchmarks for evaluating protein alignment methods.

Investigating these different possible explanations will require additional study. For example, the impact of model misspecification could be explored using simulations in which the various assumptions of the stochastic model assumed in BAli-Phy could be violated. Supplementary Section 2.4 and Table S8 available on Dryad include an initial evaluation of the impact of model misspecification of the substitution model (which shows that using JTT instead of WAG can reduce SP-score or Modeler score by 0.01), but other types of model misspecification are likely to be more impactful. For example, selection is clearly relevant to protein sequence evolution, and so simulating under sequence evolution models with varying degrees and types of selection could potentially reveal the degree to which selection complicates alignment estimation. Similarly, heterotachy (Lopez et al. 2002; Taylor et al. 2006; Zhou et al. 2007), where sites evolve independently not under identical models, is also expected to be present in many data sets and may complicate the inference of alignment using the stochastic sequence evolution model within BAli-Phy. Fortunately, some simulation tools have been developed that could be used for such studies, as described in Arenas et al. (2013) and Goldstein and Pollock (2016). Determining whether the reference alignments in these biological benchmarks have errors will depend on experimental data that provide structural features of folded proteins as well as on alignments of multiple protein structures and so may require new computational methods. Thus, determining the relative contribution of each of these possible explanations will require substantial effort and should be the focus of future research.

Other directions for future work include examining these questions on larger data sets. Our study examined BAli-Phy version 2.3.8, but a new version has been developed (version 3.1.5) that is faster and uses reduced memory compared to the version we studied, and is designed to handle larger data sets; therefore, any subsequent evaluation of these issues on larger sequence data sets should use this new version. The questions we raise here are also relevant to RNA and DNA sequence evolution (see, e.g., the thoughtful discussion in Morrison 2018 about multiple sequence alignment for nucleotides sequences), and so future work should examine how statistical alignment methods perform compared to other methods on nucleotide data sets with structural alignments and also on simulated nucleotide data sets.

Data Availability

All data sets studied in this article, along with the BAli-Phy posterior decoding alignments of these data sets, are available in Dryad.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org//10.5061/dryad.8k821ds.

Funding

This research was part of the Blue Waters sustained petascale computing project, which is supported by the U.S. National Science Foundation [Grant Numbers OCI- 0725070 and ACI-1238993] and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This research made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program in conjunction with the National Center for Supercomputing Applications and which is supported by funds from the University of Illinois at Urbana-Champaign. This work was supported in part by the U.S. National Science Foundation (NSF) [ABI-1458652 to T.W.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank the Editor, Associate Editor, and three anonymous reviewers whose helpful comments led to improvements in the manuscript. We also thank the NSF Center for Pure and Applied Mathematics at UCLA, whose support of a workshop on multiple sequence alignment stimulated this research project.

References

Alterovitz R., Arvey A., Sankararaman S., Dallett C., Freund Y., Sjölander K. 2009. ResBoost: characterizing and predicting catalytic residues in enzymes. BMC Bioinformatics 10:197.

Andreeva A., Prlic A., Hubbard T., Murzin A. 2007. SISYPHUS—structural alignments for proteins with non-trivial relationships. Nucleic Acids Res. 35:D253–D259.

Aniba M.R., Poch O., Thompson J.D. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res. 38:7353–7363.

Arenas M., Dos Santos H., Posada D., Bastolla U. 2013. Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics 29:3020–3028.

Bahr A., Thompson J.D., Thierry J.C., Poch O. 2001. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res. 29:323–326.

Bairoch A., Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28:45–48.

Bernardes J., Zaverucha G., Vaquero C., Carbone A. 2016. Improvement in protein domain identification is reached by breaking consensus, with the agreement of many profiles and domain co-occurrence. PLOS Comput. Biol. 12:e1005038.

Bishop M.J., Thompson E.A. 1986. Maximum likelihood alignment of DNA sequences. J. Mol. Evol. 190:159–165.

Blackburne B., Whelan S. 2013. Class of multiple sequence alignment algorithm affects genomic analysis. Mol. Biol. Evol. 30:642–653.

Blackshields G., Wallace I.M., Larkin M., Higgins D.G. 2006. Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol. 6:321–339.

Boyce K., Sievers F., Higgins D. 2014. Simple chained guide trees give high-quality protein multiple sequence alignments. Proc. Natl. Acad. Sci. USA 111:10556–10561.

Bradley R., Roberts A., Smoot M., Juvekar S., Do J., Dewey C., Holmes I., Pachter L. 2009. Fast statistical alignment. PLoS Comput. Biol. 5:e1000392.

Chatzou M., Magis C., Chang J.-M., Kemena C., Bussotti G., Erb I., Notredame C. 2015. Multiple sequence alignment modeling: methods and applications. Brief. Bioinformatics 17:1009–1023.

Cuff, J.A., Barton G.J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins Struct. Funct. Genetics 40:502–511.

Daniels N., Kumar A., Cowen L., Menke M. 2012. Touring protein space with Matt. IEEE/ACM Trans. Comput. Biol. Bioinformatics 9:286–293.

Dessimoz C., Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11:R37.

Do C., Gross S., Batzoglou S. 2006. CONTRAlign: discriminative training for protein sequence alignment. Research in Computational Molecular Biology: 10th Annual International Conference (RECOMB 2006), Venice, Italy, April 2–5, 2006. Berlin: Springer. p. 160–174.

Do C.B., Mahabhashyam M.S.P., Brudno M., Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

Dover G. 1987. Nonhomologous views of a terminology muddle. Cell 51:515.

Edgar R., Batzoglou S. 2006. Multiple sequence alignment. Curr. Opin. Struct. Biol. 16:368–373.

Edgar R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fleissner R., Metzler D., von Haeseler A., Lewis P. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst. Biol. 54:548–561.

Fletcher W., Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–1888.

Fletcher W., Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. 27:2257–2267.

George R.A., Heringa J. 2002. Protein domain identification and improved sequence similarity searching using PSI-BLAST. Proteins Struct. Funct. Genetics 48:672–681.

Goldstein R.A., Pollock D.D. 2016. The tangled bank of amino acids. Protein Sci. 25:1354–1362.

Golubchik T., Wise M.J., Easteal S., Jermiin L.S. 2007. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol. Biol. Evol. 24:2433–2442.

Hein J., Jensen J.L., Pedersen C. 2003. Recursions for statistical multiple alignment. Proc. Natl. Acad. Sci. USA 100:14960–14965.

Holmes I.H. 2017. Historian: accurate reconstruction of ancestral sequences and evolutionary rates. Bioinformatics 33:1227–1229.

Holmes I., Bruno W.J. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics 17:803–820.

Iantorno S., Gori K., Goldman N., Gil M., Dessimoz C. 2014. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. In: Multiple sequence alignment methods. Totowa, NJ: Humana Press. p. 59–73.

Karin E.L., Susko E., Pupko T. 2014. Alignment errors strongly impact likelihood-based tests for comparing topologies. Mol. Biol. Evol. 31:3057–3067.

Katoh K., Standley D. 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics 32:1933–1942.

Katoh K., Misawa K., Kuma K.-I., Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kemena C., Taly J.-F., Kleinjung J., Notredame C. 2011. STRIKE: evaluation of protein MSAs using a single 3D structure. Bioinformatics 27:3385–3391.

Lake J.A. 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8:378–385.

Lassmann T., Sonnhammer E.L.L. 2005. Automatic assessment of alignment quality. Nucleic Acids Res. 33:7120–7128.

Le Q., Sievers F., Higgins D.G. 2017. Protein multiple sequence alignment benchmarking through secondary structure prediction. Bioinformatics 33:1331–1337.

Liberles D.A., Teichmann S.A., Bahar I., Bastolla U., Bloom J., Bornberg-Bauer E., Colwell L.J., de Koning A.P.J., Dokholyan N.V., Echave J., Elofsson A., Gerloff D.L., Goldstein R.A., Grahnen J.A., Holder M.T., Lakner C., Lartillot N., Lovell S.C., Naylor G., Perica T., Pollock D.D., Pupko T., Regan L., Roger A., Rubinstein N., Shakhnovich E., Sjölander, K., Sunyaev S., Teufel A.I., Thorne J.L., Thornton J.W., Weinreich D.M., Whelan S. 2012. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21:769–785.

Liu K., Raghavan S., Nelesen S., Linder C.R., Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324:1561–1564.

Lopez P., Casane D., Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

Löytynoja A., Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. USA 102:10557–10562.

Löytynoja A., Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635.

Lunter G., Miklós I., Drummond A., Jensen J.L., Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6:83.

Lunter G.A., Miklós I., Song Y.S., Hein J. 2003. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. J. Comput. Biol. 10:869–889.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.

Miklós I. 2002. An improved algorithm for statistical alignment of sequences related by a star tree. Bull. Math. Biol. 64:771–779.

Miklós I. 2003. Algorithm for statistical alignment of sequences derived from a Poisson sequence length distribution. Discret. Appl. Math. 127:79–84.

Miklós I., Lunter G.A., Holmes I. 2004. A "long indel model" for evolutionary sequence alignment. Mol. Biol. Evol. 21:529–540.

Mirarab S., Warnow T. 2011. FASTSP: Linear time calculation of alignment accuracy. Bioinformatics 27:3250–3258.

Mirarab S., Nguyen N., Guo S., Wang L.-S., Kim J., Warnow T. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol. 22:377–386.

Mizuguchi K., Deane C.M., Blundell T.L., Overington J.P. 1998. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci. 7:2469–2471.

Morgenstern B. 1999. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211–218.

Morrison D.A. 2018. Multiple sequence alignment is not a solved problem. arXiv preprint arXiv:1808.07717.

Morrison D.A., 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.

Mulder N.J., Apweiler R. 2002. Tools and resources for identifying protein families, domains and motifs. Genome Biol. 3(1):reviews2001.1.

Nguyen N., Mirarab S., Kumar K., Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. Genome Biol. 16:124.

Notredame C. 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput. Biol. 3:1405–1408.

Notredame C., Higgins D., Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Novák Á., Miklós I., Lyngsø R., Hein J. 2008. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics 24:2403–2404.

Nute M., Warnow T. 2016. Scaling statistical multiple sequence alignment to large datasets. BMC Genomics 17:135–144.

Ogden T.H., Rosenberg M.S. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. 55:314–328.

O'Sullivan O., Suhre K., Abergel C., Higgins D.G., Notredame C. 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol. 340:385–395.

Pais F.S., Ruy P.C., Oliveira G., Coimbra R.S. 2014. Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol. Biol. 9:4.

Pei J., Kim B.-H., Grishin N.V. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36:2295–2300.

Philippe H., Vienne D., Ranwez V., Roure B., Baurain D., Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. Eur. J. Taxon. 283:1–25.

Redelings B. 2014. Erasing errors due to alignment ambiguity when estimating positive selection. Mol. Biol. Evol. 31:1979–1993.

Redelings B. 2018. BAli-Phy's User's Guide v3.0. Available from: http://www.bali-phy.org/README.html#mixing_and_convergence (accessed February 27, 2018).

Redelings B.D., Suchard M.A. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC Evol. Biol. 7:40.

Reeck G., de Haen C., Teller D., Doolitte R., Fitch W., Dickerson R., Chambon P., McLachlan A., Margoliash E., Jukes T., Zuckerkandl E. 1987. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell 50:667.

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

Roshan U., Livesay D.R. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics 22:2715–2721.

Sankararaman S., Sjölander K. 2008. INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 24:2445–2452.

Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D., Higgins D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.

Simmons M.P., Müller K.F., Norton A.P. 2010. Alignment of, and phylogenetic inference from, random sequences: the susceptibility of alternative alignment methods to creating artifactual resolution and support. Mol. Phylogenet. Evol. 57:1004–1016.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Suchard M.A., Redelings B.D. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22:2047–2048.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Taylor M., Kai C., Kawai J., Carninci P., Hayashizaki Y., Semple C. 2006. Heterotachy in mammalian promoter evolution. PLoS Genet. 2:e30.

Thompson J.D., Linard B., Lecompte O., Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6(3):318093.

Thompson J.D., Plewniak F., Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27:2682–2690.

Thorne J.L., Kishino H., Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.

Thorne J.L., Kishino H., Felsenstein J. 1992a. Erratum—an evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 34:91–91.

Thorne J.L., Kishino H., Felsenstein J. 1992b. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34:3–16.

Van Walle I., Lasters I., Wyns L. 2005. SABmark—a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 21:1267–1268.

Wang L.-S., Leebens-Mack, J., Wall P.K., Beckmann K., de Pamphilis C.W., Warnow T. 2011. The impact of multiple protein sequence alignment on phylogenetic estimation. IEEE/ACM Trans. Comput. Biol. Bioinform. 8:1108–1119.

Warnow T. 2017. Computational phylogenetics: an introduction to designing methods for phylogeny estimation. Cambridge University Press.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Wilke C. 2012. Bringing molecules back into molecular evolution. PLoS Comput. Biol. 8(6):e1002572.

Xue L.C., Dobbs D., Bonvin A. M., Honavar V. 2015. Computational prediction of protein interfaces: a review of data driven methods. FEBS Lett. 589:3516–3526.

Yamada S., Gotoh O., Yamana H. 2006. Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost. BMC Bioinformatics 7:524.

Zhou Y., Rodrigue N., Lartillot N., Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol. Biol. 7:206.