Supplementary Material: DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins

Hyrum Carroll[a]*, Wesley Beckstead[b], Timothy O'Connor[b], Mark Ebbert[a,b], Mark Clement[a], Quinn Snell[a] and David McClellan[b]

[a]Computer Science Department,[b]Biology Department, Brigham Young University
Provo, Utah 84602, USA
{hdc,clement,snell}@cs.byu.edu, {wesb,tdoconnor,david_mcclellan}@byu.edu, marktwe@byu.net

1 INTRODUCTION

Since Thompson *et al.* published BAliBASE (1999a), several other protein databases have also been published (Thompson *et al.*, 2005; Raghava *et al.*, 2003; Edgar, 2004b; Van Walle *et al.*, 2005; Letunic *et al.*, 2004; Subramanian *et al.*, 2005). Most of these databases leverage structural alignments to provide a suite of "gold standard" alignments. They have been well accepted by the community to provide evaluations of MPSAs (Thompson *et al.*, 1999b; Van Walle, 2004; Edgar, 2004b,a; Do *et al.*, 2005; Lassmann and Sonnhammer, 2002, 2005a; Karplus and Hu, 2001; Subramanian *et al.*, 2005). While these efforts help to evaluate protein data sets, an unequal effort has been manifest for DNA data sets.

We presented a method to convert multiple protein sequence alignments (MPSAs) into multiple DNA sequence alignments (MDSAs) (Carroll *et al.*, 2007), allowing MSAs to be evaluated on DNA data sets. This paper covers the details of the process of conversion and provides a case study analyzing the accuracy of multiple sequence alignment programs using MPSAs and these MDSAs.

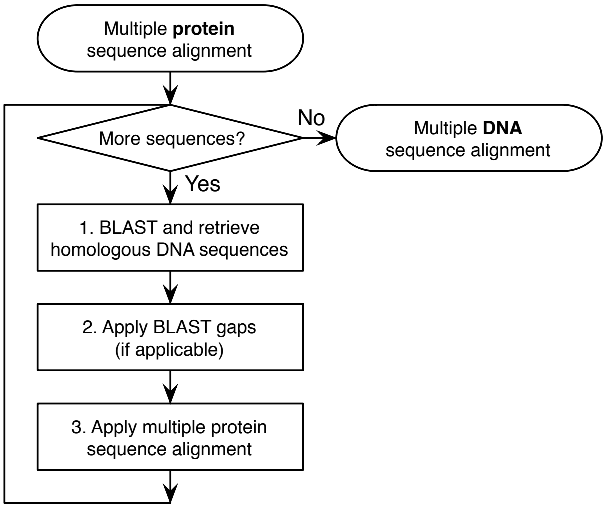
Multiple **protein** sequence alignment
More sequences?
No
Multiple **DNA** sequence alignment
Yes
1. BLAST and retrieve homologous DNA sequences
2. Apply BLAST gaps (if applicable)
3. Apply multiple protein sequence alignment

Fig. 1: *Flow chart for MPSA2MDSA.*

2 REFERENCE ALIGNMENTS

Estimating a DNA benchmark alignment from a protein alignment requires three steps: similarity searching, reconciling inconsistencies and applying the multiple protein sequence alignment (see Figure 1). First, TBLASTN is used to perform a similarity search of a protein sequence to get a corresponding DNA sequence. Second, any inconsistencies in the hit sequence (e.g., length or introduced gaps) are reconciled by inserting gaps or ambiguous characters. Finally, the gaps dictated by the MPSA are inserted into the MDSA to reflect biological accuracy. Each step is covered in detail in the remainder of this section.

2.1 Similarity Search

The first step in building a multiple DNA sequence alignment involves finding DNA sequences that are analogous to the protein sequences in the MPSA. Analogous sequences can be determined by similarity searches when an appropriate statistical test is used as the metric or scheme (Karlin and Altschul, 1990). The Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) is used for the similarity search algorithm. We chose BLAST for its long track record, speed, performance, ease of use, and statistical scoring metric (McGinnis and Madden, 2004). TBLASTN is a BLAST derivative, which translates nucleotide databases into protein sequences in all six reading frames, then identifies the most statistically probable sequence as hits (Altschul *et al.*, 1997). The input to TBLASTN is a protein sequence (the query), a database of nucleotide sequences and a cut-off threshold for the E-value. For this work, similarity searches were performed on the September 2006 version of the `nt` GenBank (Benson *et al.*, 2005) database, which has 16.9 billion base pairs in 3.8 million sequences. The second parameter to TBLASTN is a cut-off threshold value. In this study, matches with an E-value worse than 0.001 are ignored. The output of TBLASTN is the translated analogous sequences with the lowest E-value and corresponding identification information. Finally, the NCBI tool `fastacmd` retrieves the analogous DNA sequence from the `nt` database.

2.2 Reconcile Inconsistencies

The second step to building a MDSA is to account for the occasional gaps introduced by the similarity search. BLAST, among other similarity searches uses a pairwise alignment criteria for

*to whom correspondence should be addressed

Table 1. Reference Protein Alignment Benchmark Suites

Name	Version	# of Alignments
BAliBASE (Thompson *et al.*, 2005)	3.0	498
OXBench (Raghava *et al.*, 2003)	1.3	672
PREFAB (Edgar, 2004b)	4.0	1682
SMART (Ponting *et al.*, 1999)	June 7, 2006	701

matches. Adding gaps into the hit sequence (which account for insertions/deletions) sometimes improves the calculated likelihood that the query and the modified hit sequence are analogous. This produces two sources of gaps in the hit sequence: terminal gaps and interior gaps. Terminal gaps occur when the section of the hit sequence that corresponds to the query sequences does not run the entire length of the query sequence (it either does not start early enough and / or it is not long enough). The user can choose to account for interior gaps by either ignoring them or adding additional gaps into the MDSA. Finally, if a hit sequence does not provide the DNA for a section of the query (due to gaps), the least ambiguous characters possible are inserted to account for the missing data. For example, if the amino acid in the query sequence is Tyrosine, then the first two nucleotides are known to be T and A and the most resolution that the third character can have is a Y (a T or C).

2.3 Apply Multiple Protein Sequence Alignment

The last step to producing a MDSA is to apply the alignment from the MPSA. This step is important to preserve the alignment features obtained by higher order methods (e.g., secondary and tertiary structure and chemical properties) or in other words, to preserve the higher order benchmark alignment. Since the amino acid sequence is known, gaps are inserted such that they do not cause frame shifts. To do this, for each gap in the MPSA, three gaps are introduced into the MDSA at the respective locations.

2.4 Reference Multiple Protein Sequence Alignment Databases

Recently, a number of protein sequence data sets have been presented to provide a benchmark for alignment algorithms (see Table 1). These benchmarks leverage structural alignments to provide a suite of “gold standard” alignments. These alignments are assumed to be the “true” alignments, and calculated alignments are generally evaluated by comparison against them. They have been well accepted by the scientific community and used in numerous studies to compare the quality of protein alignments generated by multiple sequence alignment programs (Thompson *et al.*, 1999b; Van Walle, 2004; Edgar, 2004a,b; Do *et al.*, 2005; Lassmann and Sonnhammer, 2002, 2005a; Karplus and Hu, 2001).

BAliBASE (Benchmark Alignment dataBASE) contains reference alignments that have been manually refined and validated by superposition of known tertiary structures (Thompson *et al.*, 2005). OXBench (from the University of Oxford), contains automated protein alignments that were benchmarked using tertiary structure associations (Raghava *et al.*, 2003). PREFAB (Protein REFerence Alignment Benchmark) contains protein alignments based on pairs of protein sequences that have been structurally aligned and supplemented with as many as 50 homologs found by PSI-BLAST (Edgar, 2004b). SMART (Simple Modular Architecture Research Tool) alignments, were also manually refined with structure comparisons, but where no structure was available, automated alignment techniques were used (Ponting *et al.*, 1999).

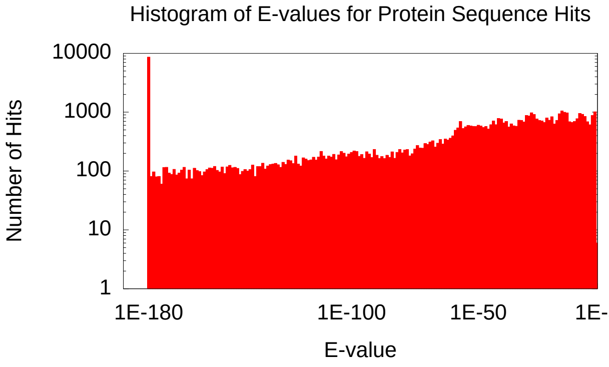

Fig. 2: *Histogram of the E-values from the hits of the protein sequences. (Note: To conservatively correct for scores reported by BLAST to have an E-value of 0.0, scores less than or equal to 1E-180 are reported as 1E-180.)*

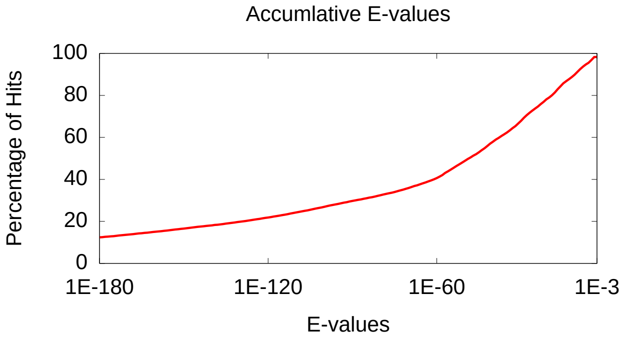

Fig. 3: *Aggregates of the number of hits plotted against E-values. (Note: To conservatively correct for scores reported by BLAST to have an E-value of 0.0, scores less than or equal to 1E-180 are reported as 1E-180.)*

2.5 Results

In general, MDSA2MPSA finds good matches in the `nt` GenBank database (Benson *et al.*, 2005) in terms of sequence identity and E-value. The majority, 69.0%, of protein sequences have matches in the database that have 100% sequence identity with the translated DNA sequences. Furthermore, another 3.9% of the hits have only one mismatched amino acid with the protein query. In terms of E-values, 98.3% of the protein sequences found a DNA sequence in the database with a score of 0.001 or better. Figures 2 and 3 illustrate all of the E-values for the protein sequence hits. The lower

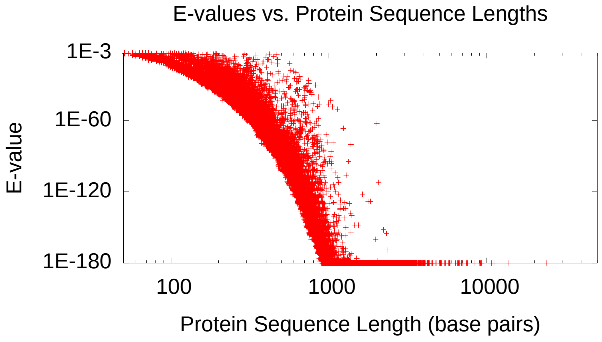

Fig. 4: *E-values of all the hits plotted against the length of the protein sequence query. (Note: To conservatively correct for scores reported by BLAST to have an E-value of 0.0, scores less than or equal to 1E-180 are reported as 1E-180.)*

Table 2. Categorization of Multiple Sequence Alignment Programs

Program	Progressive	Iterative	Local
CLUSTALW	X		
DIALIGN		X	X
Kalign	X		
MAFFT-GINSI	X	X	
MAFFT-LINSI	X	X	X
MAFFT-NS1	X		
MAFFT-NSI	X	X	
MUSCLE-Default	X	X	
MUSCLE-Fast	X	X	
POY	X	X[1]	
ProbCons/ProbConsRNA	X[2]	X	
T-Coffee	X		X

The progressive column indicates programs that use progressive alignment algorithm (Feng and Doolittle, 1987). Iterative refers to programs to refine the multiple sequence alignment. Programs that incorporate local alignment (in addition to global alignment) have a mark in the local column.

[1] Optimization Alignment, [2] Markov model

Table 3. Arguments Used For Multiple Sequence Alignment Programs

Program	Version	Arguments
CLUSTALW	1.83	defaults
DIALIGN	2.2.1	defaults
Kalign	2.0	defaults
MAFFT-GINSI	5.861	–maxiterate 1000 –globalpair
MAFFT-LINSI	5.861	–maxiterate 1000 –localpair
MAFFT-NS1	5.861	–maxiterate 0 –retree 1
MAFFT-NSI	5.861	–maxiterate 1000
MUSCLE-Default	3.6	-stable
MUSCLE-Fast	3.6	-stable -maxiters 1 -diags
POY	3.0.11	-replicates 10 -repintermediate
ProbCons/ProbConsRNA	1.10	-ir 1000
T-Coffee	4.58	defaults

the E-value, the greater the similarity between the query and the hit sequences. In the graphics, the x-axis in both graphics is logarithmic and in Figure 2 the y-axis is as well. Also, the values are bucketed logarithmically. E-values with a score better than or equal to 1E-180, are display at the 1E-180 location. This adjustment accounts for the E-values that BLAST reports as having an E-value of 0.0. (8,567 of them, or 12.5% of all hits). While the tool finds a high percentage of quality matches, databases are growing at an exponential rate, thereby increasing the number and quality analogous hits of protein queries.

E-value scores are calculated from the length and similarity of the query and hit sequences. Figure 4 shows the correlation between the length and the E-value of the hits found in the nt database. Note both scales in the graphic are logarithmic. The data here suggests that as a database increases, in terms of the length and quality of the sequences, BLAST will find hits with greater similarity to the query.

In total we converted 3,545 reference alignments, comprising of 68,581 sequences and 35,600,958 bases. These reference alignments are publicly available at http://csl.cs.byu.edu/mdsas/.

3 MULTIPLE SEQUENCE ALIGNMENT CASE STUDY

We performed an alignment case study by testing the leading alignment programs on our newly created MDSAs (the mdsa_all version). The purpose of this case study is two-fold: 1) to show the usefulness of our DNA reference sets, and 2) to test the empirical performance of these alignment programs on DNA. The alignment programs that we chose have been tested on the protein benchmarks discussed in section 2.4 and have proven to be effective at aligning protein sequences. Even so, their performance on DNA sequences is virtually unknown, hence benchmarking these programs on our new DNA reference data sets is important. In addition to running each alignment program on our reference MDSAs, we also chose to run them on the reference MPSAs. We decided that this would give us a uniform method of assessing each alignment algorithm on protein sequences and comparing these results to the accuracy of each alignment algorithm on the corresponding DNA sequences that are found in our DNA alignments. All test alignments and accuracy measures were performed with the supercomputers in the Ira and Mary Lou Fulton Supercomputing Laboratory at Brigham Young University, using Dual-core Intel Xeon EM64T processors (2.6GHz) and 8 GB of memory.

3.1 Alignment Programs

We chose eight different alignment programs to benchmark on the DNA reference alignments: CLUSTALW (Thompson *et al.*, 1994), DIALIGN (Morgenstern *et al.*, 1998), Kalign (Lassmann and Sonnhammer, 2005b), MAFFT (Katoh *et al.*, 2005), MUSCLE (Edgar, 2004a), POY (Gladstein and Wheeler, 2000), ProbCons (Do *et al.*, 2005), and T-Coffee (Notredame *et al.*, 2000). These programs use a variety of strategies to construct a multiple sequence alignment, such as progressive alignment, iterative refinement, probabilistic alignment and so forth (see Table 2). They are widely used in biology and bioinformatics today. For each alignment program, we used default parameters, unless noted otherwise (see Table 3).

3.2 Alignment Benchmarks

Not only were each of the alignment programs evaluated on the following MPSAs: BAliBASE, OXBench, PREFAB, and SMART, but also on their respective MDSAs. The one exception is POY, which restricts its analysis to DNA sequences. For BAliBASE, OXBench and SMART, we excluded alignments that have over 100 sequences in order to make the test manageable for T-Coffee and POY, which require excessive CPU time for larger sets. In addition, we discarded alignments that did not complete within 2 weeks for 1 or more MSA programs.

We used reference sets 1-5 of BAliBASE for assessing each alignment algorithm on DNA sequences. Reference sets 6-8 contain repeats, inversions and transmembrane helices. We excluded these reference sets because none of the chosen alignment programs were designed to handle these cases. Our tool, MPSA2MDSA, converted all of the protein alignments in reference sets 1-5 to DNA alignments. We excluded 8 of these alignments because they contain more than 100 sequences, allowing 378 DNA alignments to be included in the case study for BAliBASE. To test each alignment algorithm on protein sequences, we used the 378 corresponding protein alignments in BAliBASE.

For OXBench, MPSA2MDSA was able to convert all 672 MPSAs to MDSAs. We discarded 4 alignments that were over 100 sequences and 4 alignments that aborted or did not finish after 2 weeks while being analyzed by POY. In total, 664 DNA alignments were included in the case study for OXBench. For analyzing each alignment algorithm on protein sequences, we used 668 corresponding protein alignments in OXBench.

We used the June 7, 2006 version of the SMART database to convert to MDSAs and to use for the case study. Our tool converted 698 of the 701 MPSAs in SMART to MDSAs. We excluded 108 alignments that either contained over 100 sequences or did not complete on POY or DIALIGN after 2 weeks. This gives a total of 590 MDSAs that were used for the case study. 592 corresponding MPSAs were used from SMART to assess each algorithm on protein sequences.

MPSA2MDSA converted 1676 of the 1682 protein alignments in PREFAB to DNA alignments. All 1676 DNA alignments were used in the case study and all 1682 protein alignments were used in the evaluation of the alignment programs on protein sequences.

3.3 Accuracy Measurement and Statistical Analysis

To ascertain the accuracy of the alignments generated by each alignment program we used a variety of scoring metrics that compare a calculated multiple sequence alignment to a reference alignment. In general we used the scoring metrics that were provided by or suggested for each respective database. These scoring metrics are all forms of the Q (Quality) and TC (Total Column) scores. The Q score, previously termed as the developer score (Sauder *et al.*, 2000) or SPS (Sum of Pairs Score) (Thompson *et al.*, 1999b), is defined as the number of correctly aligned residue pairs in the generated alignment divided by the number of residue pairs in the reference alignment. The TC score, also known as the CS score, is the number of correctly aligned columns in the generated alignment divided by the number of columns in the reference alignment. The TC score is the same as the Q score in the case of pairwise alignment. For BAliBASE, OXBench and SMART we used the Q and TC scores. We only used the Q score for PREFAB since the alignments in these databases are pairwise.

For an individual database we averaged each score across all of the alignments in the database. To measure statistical significance in the accuracy differences between alignment programs, we performed a Friedman rank test with the accuracy scores (Friedman, 1937). This test is more conservative than the Wilcoxon test, which has also been used to determine statistical significance in past alignment studies (Edgar, 2004b).

Table 4. DNA BAliBASE scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
CLUSTALW	.445	5.78	.120	5.45	52.0	6.53
DIALIGN	.389	4.34	.099	5.24	169.7	7.80
Kalign	.408	5.27	.105	5.40	1.7	**1.44**
MAFFT-GINSI	.617	**11.21**	.277	**9.91**	58.1	5.87
MAFFT-LINSI	.607	10.75	.275	9.78	47.9	6.28
MAFFT-NS1	.459	6.68	.141	6.44	2.3	2.25
MAFFT-NSI	.559	9.57	.207	8.30	21.7	4.02
MUSCLE-Default	.516	8.02	.198	7.90	188.0	8.02
MUSCLE-Fast	.291	4.42	.099	5.05	6.3	3.11
POY	.305	2.59	.045	3.79	26364.4	11.14
ProbCons	.452	6.78	.124	6.39	3228.9	10.10
T-Coffee	.308	2.56	.071	4.35	10453.7	11.45

The average Q scores, TC scores, and times (in seconds) for BAliBASE MDSAs are shown. For each category the ranks according to the Friedman test are given. For the Q and TC scores, the higher the rank indicates higher accuracy. The alignment programs that ranked the highest for the Q score and TC score are highlighted. For the times, a lower rank indicates better performance in comparison to other programs. The best program in terms of CPU times is also highlighted.

3.4 DNA Alignment Results

All of the Friedman rank tests performed on scores of the DNA alignment benchmarks are statistically significant (p-value of 2e-16). Table 4 shows the average scores and times for the BAliBASE DNA alignments as well as the ranks according to the Friedman test for each category. Three MAFFT strategies (GINSI, LINSI, and NSI) rank first, second, and third for both the Q and the TC scores. POY and T-Coffee are ranked last for both scores. Kalign and MAFFT-NS1 do very well in terms of time, averaging 1.7 seconds and 2.3 seconds per alignment and rank first and second respectively. T-Coffee, POY, and ProbCons come in last in terms of time. It is interesting to note the wide spread in execution time among the different alignment methods, ranging from 1 second to 7 hours on average.

The average scores and times for the MDSAs of OXBench are shown in Table 5. The program rankings according to the Friedman test are also given. MAFFT-LINSI comes in first for both Q and TC score. All four strategies (GINSI, LINSI, NS1, and NSI) rank higher than any other method. CLUSTALW and MUSCLE come in behind MAFFT. Kalign is the fastest alignment method on OXBench. Again POY and T-Coffee rank the worst in accuracy and time.

The scores, times and Friedman ranks for the DNA alignments of PREFAB are displayed in Table 6. Kalign comes on top in CPU time again followed by MUSCLE-Fast and CLUSTALW. ProbCons and T-Coffee are ranked last in comparison to the times of other alignment methods. MAFFT-LINSI is ranked number

Table 5. DNA OXBench scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
CLUSTALW	.766	7.46	.671	7.61	1.6	4.89
DIALIGN	.696	4.57	.604	5.50	1.5	5.23
Kalign	.756	6.30	.645	5.88	.2	**1.57**
MAFFT-GINSI	.789	8.50	.687	8.07	1.2	7.67
MAFFT-LINSI	.795	**9.11**	.699	**8.72**	1.4	7.29
MAFFT-NS1	.782	7.97	.677	7.70	.8	5.16
MAFFT-NSI	.789	8.33	.687	8.03	.5	6.12
MUSCLE-Default	.755	6.75	.660	7.02	1.4	6.47
MUSCLE-Fast	.743	6.02	.643	6.40	.4	3.51
POY	.694	3.85	.574	3.91	79.4	8.95
ProbCons	.741	5.57	.626	5.34	8.2	9.58
T-Coffee	.692	3.56	.577	3.80	49.0	11.55

The average Q scores, TC scores, and times (in seconds) are shown for the DNA alignments of OXBench. The ranks according to the Friedman test are given for each category. For the Q and TC scores, the higher the rank indicates better accuracy in comparison with other programs. For the times, a lower rank indicates better performance in comparison to other programs. The best program for each category is highlighted.

Table 6. DNA PREFAB scores, times, and ranks

	Q Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank
CLUSTALW	.351	6.88	.34	3.94
DIALIGN	.248	4.70	.77	4.90
Kalign	.344	7.19	.33	**1.88**
MAFFT-GINSI	.376	8.15	1.0	7.66
MAFFT-LINSI	.380	**8.39**	1.1	7.41
MAFFT-NS1	.376	8.15	.7	5.44
MAFFT-NSI	.375	8.03	.8	6.70
MUSCLE-Default	.297	5.73	1.5	6.44
MUSCLE-Fast	.297	5.73	.4	3.80
POY	.254	4.74	2.4	8.59
ProbCons	.298	5.82	4.5	10.11
T-Coffee	.254	4.50	7.2	11.14

The average Q score and times (in seconds) are shown for PREFAB. The ranks according to the Friedman test are also given for each category. For the Q score, the higher the rank indicates better accuracy in comparison with other programs. For the times, a lower rank indicates better performance in comparison to other programs. The best program for each category is highlighted.

Table 7. DNA SMART scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
CLUSTALW	.577	4.41	.224	4.68	10.7	5.61
DIALIGN	.515	3.29	.183	3.58	37.3	8.23
Kalign	.687	7.27	.294	6.68	.5	**1.52**
MAFFT-GINSI	.833	**11.08**	.468	**10.81**	9.6	6.19
MAFFT-LINSI	.812	10.67	.460	10.65	.5	2.55
MAFFT-NS1	.673	7.07	.288	6.62	4.2	4.60
MAFFT-NSI	.790	9.93	.415	9.70	8.2	6.00
MUSCLE-Default	.700	7.71	.331	7.48	1.3	3.15
MUSCLE-Fast	.550	3.79	.194	3.78	19.3	7.48
POY	.555	3.51	.200	4.07	4163.6	11.15
ProbCons	.701	7.49	.301	7.38	544.4	9.86
T-Coffee	.444	1.77	.146	2.67	2507.1	11.66

The average Q score, TC score, and time (in seconds) are shown for the MDSAs of SMART. The ranks according to the Friedman test are also given for each category. For the Q and TC scores, the higher the rank indicates better accuracy in comparison with other programs. For the times, a lower rank indicates better performance in comparison to other programs. The best program for each category is highlighted.

one for this benchmark in terms of the accuracy of the generated alignments compared to the reference DNA alignments produced by MPSA2MDSA. All four MAFFT strategies (GINSI, LINSI, NS1, and NSI) rank higher than any other alignment method. They are followed by Kalign and CLUSTALW. As has been shown with other benchmarks, T-Coffee, POY, and DIALIGN are in the bottom end of the rankings on accuracy.

Table 7 shows the average scores and times for the DNA alignments of SMART. MAFFT-GINSI comes in first again for both scores with MAFFT-LINSI and MAFFT-NSI ranking second and third respectively. MUSCLE with default parameters follows in accuracy, along with Kalign and ProbCons. DIALIGN and T-Coffee rank last in accuracy for both scores. Again Kalign ranks number one in CPU time with POY and T-Coffee coming in last.

A discussion on these results are given in the remainder of this section.

3.5 Protein Alignment Results

As in the case of the DNA alignments, all of the Friedman rank tests performed on scores of the protein alignment benchmarks are statistically significant (p-value of 2e-16). Table 8 shows the average Q and TC scores as well as the CPU times for the protein alignments of BAliBASE. The scores were only measured across the core blocks defined by BAliBASE developers. Kalign follows the pattern shown in the DNA benchmarks and is the fastest in execution time. T-Coffee is ranked the worst in terms of time. Though ProbCons is ranked second to last in CPU time, it comes in first in terms of accuracy, achieving the highest ranking for both the Quality score and the Total Column score. ProbCons is followed by MAFFT-LINSI and then MAFFT-GINSI for both scores. T-Coffee and the MUSCLE strategies ranks follow in accuracy. DIALIGN and CLUSTALW are ranked the lowest in both the Q score and TC score by the Friedman rank test. We see that the average times are dramatically lower on the protein alignments of BAliBASE than on the DNA alignments of BAliBASE (Table 4). This is expected given that the protein alignments are one third of the length of the DNA alignments created by MPSA2MDSA.

Table 9 shows the average Q and TC scores for the MPSAs of OXBench. CLUSTALW and MUSCLE-Default are tied for the highest accuracy in terms of Q score and MUSCLE-Default is the most accurate in terms of TC score. But they are closely followed by MAFFT-LINSI, ProbCons, and T-Coffee. As in the case of BAliBASE, DIALIGN is ranked last in accuracy for both scores. Again Kalign ranks first for execution time, averaging only 0.05 seconds on our supercomputers per protein alignment in OXBench.

Table 10 shows the average Q scores and times for PREFAB. The program rankings according to the Friedman test are also shown. ProbCons, CLUSTALW, and T-Coffee rank in the top three in terms of accuracy. They are followed by the MUSCLE strategies and

Table 8. Protein BAliBASE scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
CLUSTALW	.755	3.59	.447	4.22	5.1	4.94
DIALIGN	.743	2.97	.435	3.42	20.3	7.75
Kalign	.811	5.37	.526	5.16	.3	**1.21**
MAFFT-GINSI	.847	7.88	.586	7.51	8.8	6.83
MAFFT-LINSI	.860	8.59	.616	8.20	6.8	6.56
MAFFT-NS1	.784	4.20	.484	4.33	.6	3.15
MAFFT-NSI	.832	6.79	.571	6.87	3.8	5.32
MUSCLE-Default	.828	6.43	.550	6.33	7.0	6.65
MUSCLE-Fast	.776	4.37	.473	4.52	1.4	3.27
ProbCons	.866	**9.06**	.620	**8.76**	250.0	9.86
T-Coffee	.815	6.74	.557	6.68	150.6	10.46

The average Q and TC scores(measures only on core blocks) and times (in seconds) for BAliBASE MPSAs are shown. For each category the ranks according to the Friedman test are given. For the Q and TC scores, the higher the rank indicates higher accuracy. The alignment programs that ranked the highest for the Q score and TC score are highlighted. For the times, a lower rank indicates better performance in comparison to other programs. The best program in terms of CPU times is also highlighted.

Table 9. Protein OXBench scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
CLUSTALW	.861	**6.78**	.772	6.78	.93	3.68
DIALIGN	.823	4.04	.733	4.31	.58	4.48
Kalign	.854	6.18	.766	6.25	.05	**1.19**
MAFFT-GINSI	.853	5.67	.760	5.52	.51	7.53
MAFFT-LINSI	.852	6.42	.766	6.36	.57	7.25
MAFFT-NS1	.847	5.06	.752	5.07	.50	6.67
MAFFT-NSI	.852	5.64	.760	5.56	.83	7.78
MUSCLE-Default	.861	**6.78**	.775	**6.88**	.81	6.33
MUSCLE-Fast	.859	6.66	.772	6.74	.79	4.90
ProbCons	.859	6.47	.768	6.21	.91	5.95
T-Coffee	.856	6.29	.767	6.31	4.45	10.24

The average Q scores, TC scores, and times (in seconds) for the protein alignments of OXBench. The ranks according to the Friedman test are also shown. For the Q and TC scores, the higher the rank indicates higher accuracy. The alignment programs that ranked the highest for the Q score and TC score are highlighted. For the times, a lower rank indicates better performance in comparison to other programs. The best program in terms of CPU times is also highlighted.

Kalign. Again DIALIGN is ranked last in terms of accuracy. Kalign also ranks the fastest for CPU time, followed by CLUSTALW. T-Coffee comes in last in terms of time.

The average scores and times for the MPSAs of SMART are shown in Table 11. The rankings for each score and time, given by the Friedman test, are also shown. ProbCons ranks the highest for both the Q and TC scores. MAFFT-GINSI, MAFFT-LINSI, and MAFFT-NSI come in second, third, and fourth place. Again DIALIGN does the worst in accuracy as compared to other alignment methods. As in other benchmarks, DNA and Protein versions alike, Kalign comes in first in CPU time while maintaining a decent ranking in accuracy. ProbCons and T-Coffee rank last in terms of execution time.

Table 10. Protein PREFAB Q scores, and ranks

	Q Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank
CLUSTALW	.585	6.99	.59	3.64
DIALIGN	.513	4.07	.73	4.38
Kalign	.588	6.53	.19	**1.58**
MAFFT-GINSI	.558	5.14	.65	8.27
MAFFT-LINSI	.571	5.89	.64	7.80
MAFFT-NS1	.558	5.14	.49	7.15
MAFFT-NSI	.558	5.14	.41	7.58
MUSCLE-Default	.584	6.62	.32	4.53
MUSCLE-Fast	.584	6.62	.38	4.32
ProbCons	.590	**7.18**	.43	6.54
T-Coffee	.583	6.69	1.76	10.21

The average Q score and times (in seconds) are shown for PREFAB. The ranks according to the Friedman test are also given for each category. For the Q score, the higher the rank indicates better accuracy in comparison with other programs. For the times, a lower rank indicates better performance in comparison to other programs. The best program for each category is highlighted.

Table 11. Protein SMART scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
CLUSTALW	.819	4.55	.481	5.43	1.31	3.85
DIALIGN	.766	2.15	.395	2.84	6.18	7.87
Kalign	.830	5.23	.478	5.00	.27	**1.45**
MAFFT-GINSI	.871	8.57	.549	7.95	2.25	6.95
MAFFT-LINSI	.858	7.78	.533	7.46	2.04	6.57
MAFFT-NS1	.818	4.38	.460	4.49	.53	3.90
MAFFT-NSI	.853	7.04	.534	7.29	1.34	6.10
MUSCLE-Default	.851	6.80	.520	6.72	1.86	5.89
MUSCLE-Fast	.823	4.73	.461	4.59	.52	3.18
ProbCons	.873	**8.87**	.550	**8.32**	39.49	9.32
T-Coffee	.836	5.90	.490	5.91	78.03	10.91

The average Q scores, TC scores, and times for SMART. The ranks according to the Friedman test are also shown. For the Q and TC scores, the higher the rank indicates higher accuracy. The alignment programs that ranked the highest for the Q score and TC score are highlighted. For the times, a lower rank indicates better performance in comparison to other programs.

3.6 Discussion

The assessment of the chosen alignment programs on protein-coding DNA benchmarks gives valuable insight into current alignment techniques. Two general points are worth noting. First, we see in the case of the protein benchmarks there is generally high accuracy scores among all the programs and that they do not differ dramatically from one another. For example, in the protein assessment, SMART average accuracy scores range from 0.76 to 0.87 (Table 11) and OXBench scores only vary between 0.82 and 0.86 (Table 9). ProbCons, MAFFT, and MUSCLE tend to have the highest scores but they are always closely followed by T-Coffee, Kalign and other methods. When these same alignment programs are tested on DNA, we see dramatic differences and variability in the accuracy scores of each program in comparison with one another. In general we see lower accuracy scores, ranging from 0.4 to 0.8 on the DNA alignments of SMART, and from 0.3 to 0.6 on the DNA alignments of BAliBASE.

Second, the results show certain programs that do well on protein sequences but tend to rank low in accuracy for DNA sequences. T-Coffee and ProbCons, for example, rank very high on protein benchmarks but they are the least accurate of all the alignments methods for many of the DNA benchmarks. We also see some programs that do well on protein sequences and rank just as high on DNA sequences. The MAFFT strategies that come in behind ProbCons and MUSCLE on the protein benchmarks retain their high accuracy scores on DNA and rank number one, two, and three on every benchmark.

These two points indicate that many of these alignment programs have been trained and optimized on protein sequences and are not ideal for DNA alignment. Their performance on DNA has been altogether unknown. All programs produce relatively high accuracy scores on protein sequences and rank relatively close to one another. But when tested on DNA many programs drop in accuracy and other programs surpass them in the rankings.

DIALIGN is consistently the least accurate on protein sequences. On DNA, DIALIGN does better in the rankings merely because T-Coffee drops below it in accuracy. DIALIGN is not particularly fast either. On protein and DNA sequences alike, DIALIGN ranges in rank from the third to the ninth fastest alignment program.

Kalign is extremely fast and consistently ranks number one in execution time on all databases. The highest average time Kalign produces on a database is 1.7 seconds on our supercomputers for the MDSAs of BAliBASE. This is inordinately fast, considering T-Coffee averages 10,000 seconds and ProbCons averages 3,000 seconds on the same database. This would seem to indicate that Kalign takes a great reduction in accuracy in order to achieve this type of speed, but the results suggest otherwise. Kalign consistently takes first place in CPU time while maintaining moderately high accuracy scores on DNA and protein sequences and a decent ”middle ground” ranking according to Q and TC scores. This is important to many biologists who are interested in aligning large data sets quickly without taking a radical reduction in accuracy.

As mentioned before, MAFFT does very well on proteins sequences but is surpassed in many instances by Probcons and MUSCLE. On DNA benchmarks, MAFFT maintains its high accuracy scores and MAFFT-GINSI and MAFFT-LINSI take first and second on all DNA benchmarks. MAFFT-NS1 ranks third in accuracy on every single DNA database. In the case of the DNA alignments from PREFAB, all four MAFFT strategies do better than any other alignment method. MAFFT does this without a significant loss in execution time. MAFFT strategies generally rank around fifth or sixth in terms of time. For these reasons, MAFFT is a good choice for any biologist interested in aligning either DNA or protein sequences in a decent amount of time.

The MUSCLE strategies consistently rank well on the protein benchmarks. Even MUSCLE-Fast, which does not do iterative refinement, does better than many alignment programs. MUSCLE retains its accuracy on DNA but is surpassed by the MAFFT strategies.

As mentioned before, POY was chosen in order to assess the quality of DNA alignments that are produced as POY performs its optimization alignment and creates a tree without the use of a MSA as input. There have been many that have criticized POY for attempting to join alignment and phylogeny into one step, citing that alignment and phylogeny are logically independent of one another (e.g., (Simmons, 2000)). Perhaps the results of this case study shed light on the subject. The accuracy of the DNA alignments used by POY to build a tree has been virtually unknown due to the lack of DNA benchmarks in the past. The results of the case study show that POY has low accuracy scores compared to other alignment methods. POY consistently ranks second or third to last in accuracy. The goal of the POY analysis is to eliminate errors produced by preliminary alignment programs. It does this by attempting to produce the alignment in conjunction with the phylogenetic tree. Though this is a worthy goal, the results of this case study suggest that the alignments of POY are not as accurate as standard alignment programs and this may affect the resulting tree that is produced.

ProbCons does very well in the alignment of protein sequences. It ranks first in accuracy on three of the four protein databases. ProbCons does this with a great cost in time, being one of the slowest methods tested in this case study. Tested on DNA, ProbCons drops in the rankings and in general ranks around seventh place in terms of accuracy. This suggests that ProbCons has been optimized for protein sequences but it may not be the best choice for aligning DNA sequences.

T-Coffee also does well on protein benchmarks but at a great cost in time. T-Coffee comes in last place on protein benchmarks in the rankings according to CPU time. When tested on DNA, T-Coffee, like ProbCons, drops in accuracy and consistently takes last place. Again it comes in last place in terms of time on the DNA benchmarks as well. This would tend to make T-Coffee an undesirable choice for the alignment of DNA even though has historically done well in aligning protein sequences.

4 CONCLUSION

MPSA2MDSA finds high to extremely high similarity matches in the public `nt` database. Over two-thirds of the protein sequences (69.0%) found a perfect match with TBLASTN. While these results are encouraging, as databases continue to grow, in terms of length and number of sequences, the quality of the matches will increase too.

The results of the study show that many alignment programs are optimized and trained on protein sequences but vary greatly in accuracy when applied to DNA sequences. MAFFT-LINSI, MAFFT-GINSI, and MAFFT-NSI strategies are the most accurate on DNA sequences while T-Coffee and POY are the least accurate. This case study also shows that our newly created protein-coding DNA benchmarks are extremely useful at determining the accuracy of calculated alignments generated by currently used alignment programs. We feel that these benchmarks will become an integral part in the assessment of forthcoming alignment methods.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Meyers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research*, **33**, D34–38.

Carroll, H., Beckstead, W., O'Connor, T., Ebbert, M., Clement, M., Snell, Q., and McClellan, D. (2007). DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins. *Bioinformatics*, **23**(19), 2648–2649.

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

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(113-131).

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Feng, D.-F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, **25**(4), 351–360.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, **32**(200), 675–701.

Gladstein, D. S. and Wheeler, W. C. (1997-2000). *POY: The Optimization of Alignment Characters*. New York, NY.

Karlin, S. and Altschul, S. F. (1990). Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *Proceedings of the National Academy of Sciences*, **87**(6), 2264–2268.

Karplus, K. and Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**(8), 713–720.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**(2), 511–518.

Lassmann, T. and Sonnhammer, E. L. L. (2002). Quality assessment of multiple alignment programs. *FEBS Letters*, **529**, 126–130.

Lassmann, T. and Sonnhammer, E. L. L. (2005a). Automatic assessment of alignment quality. *Nucleic Acids Research*, **33**(22), 7120–7128.

Lassmann, T. and Sonnhammer, E. L. L. (2005b). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**(298).

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**, D142–D144.

McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, **32**, 20–25.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998). DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**(3), 290–294.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.

Ponting, C., Schultz, J., Milpetz, F., and Bork, P. (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Research*, **27**(1), 229–232.

Raghava, G. P. S., Searle, S. M. J., Audley, P. C., Barber, J. D., and Barton, G. J. (2003). OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**(47).

Sauder, J., Arthur, J., and Dunbrack Jr, R. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins Structure Function and Genetics*, **40**(1), 6–22.

Simmons, M. (2000). Gaps as Characters in Sequence-Based Phylogenetic Analyses. *Systematic Biology*, **49**(2), 369–381.

Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M., and Morgenstern, B. (2005). DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**(66).

Thompson, J., Plewniak, F., and Poch, O. (1999a). BAliBASE: A benchmark alignments database for the evaluation of multiple sequence alignment programs. *Bioinformatics*, **15**(1), 87–88.

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

Thompson, J. D., Plewniak, F., and Poch, O. (1999b). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, **27**(13), 2682–2690.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). BAliBASE 3.0 latest developments of the multiple sequence alignmnet benchmark. *Proteins: Structure, Function, and Bioinformatics*, **61**(1), 127–136.

Van Walle, I. (2004). Align-m–a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, **20**(9), 1428–1435.

Van Walle, I., Lasters, I., and Wyns, L. (2005). SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**(7), 1267–1268.