Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments

Tanya Golubchik, Michael J. Wise,† Simon Easteal,‡ and Lars S. Jermiin*§¶*

*School of Biological Sciences, University of Sydney, Sydney, Australia; †School of Biomolecular, Biomedical and Chemical Sciences, University of Western Australia, Perth, Australia; ‡John Curtin School of Medical Research, Australian National University, Canberra, Australia; §Sydney Bioinformatics, University of Sydney, Sydney, Australia; and ¶Centre for Mathematical Biology, University of Sydney, Sydney, Australia

Multiple sequence alignment (MSA) is a crucial first step in the analysis of genomic and proteomic data. Commonly occurring sequence features, such as deletions and insertions, are known to affect the accuracy of MSA programs, but the extent to which alignment accuracy is affected by the positions of insertions and deletions has not been examined independently of other sources of sequence variation. We assessed the performance of 6 popular MSA programs (ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS, and T-COFFEE) and one experimental program, PRANK, on amino acid sequences that differed only by short regions of deleted residues. The analysis showed that the absence of residues often led to an incorrect placement of gaps in the alignments, even though the sequences were otherwise identical. In data sets containing sequences with partially overlapping deletions, most MSA programs preferentially aligned the gaps vertically at the expense of incorrectly aligning residues in the flanking regions. Of the programs assessed, only DIALIGN-T was able to place overlapping gaps correctly relative to one another, but this was usually context dependent and was observed only in some of the data sets. In data sets containing sequences with non-overlapping deletions, both DIALIGN-T and MAFFT (G-INS-I) were able to align gaps with near-perfect accuracy, but only MAFFT produced the correct alignment consistently. The same was true for data sets that comprised isoforms of alternatively spliced gene products: both DIALIGN-T and MAFFT produced highly accurate alignments, with MAFFT being the more consistent of the 2 programs. Other programs, notably T-COFFEE and ClustalW, were less accurate. For all data sets, alignments produced by different MSA programs differed markedly, indicating that reliance on a single MSA program may give misleading results. It is therefore advisable to use more than one MSA program when dealing with sequences that may contain deletions or insertions, particularly for high-throughput and pipeline applications where manual refinement of each alignment is not practicable.

Introduction

Alignment of multiple amino acid sequences to identify regions of high conservation is fundamental to research in molecular biology and evolution. Multiple sequence alignment (MSA) is required by most phylogenetic prediction systems as the first step in the inference of evolutionary relationships between taxa represented by amino acid (or nucleotide) sequences. MSAs are also widely employed to assist the prediction of protein function. With the continuing rise in the number of amino acid sequences deposited in publicly available databases, MSA has emerged as a critical first step in automation of sequence analysis, and automatically generated alignments are commonly used as input for subsequent downstream analysis in program pipelines (Plewniak et al. 2003). Even where efforts are made to refine the alignment automatically, by using software such a RASCAL (Thompson et al. 2003), this practice still assumes a high degree of confidence in the quality of the initial alignment.

Traditionally, alignment of highly similar sequences has been seen as a relatively simple task, and assessments of MSA programs on sequences with low evolutionary divergence tend to indicate that most programs perform well when input sequences are very similar (Lassmann and Sonnhammer 2002). On the other hand, benchmark testing on large-scale data sets, such as BAliBASE (Thompson et al. 1999a), has demonstrated that most MSA programs have serious shortcomings when confronted with insertions and deletions (Edgar and Batzoglou 2006; Nuin et al. 2006). The question then arises: how do insertions and deletions affect the performance of alignment programs if the sequences are otherwise identical?

Given the prevalence of sequences with insertions and deletions in real data sets, limitations in the capacity to align such sequences are highly significant. Mechanisms such as alternative splicing of pre-mRNA in eukaryotes can produce a variety of protein products with insertions and deletions corresponding to the alternatively spliced exons (Caporale 2006). This situation is by no means uncommon: according to current estimates, as many as 45–60% of human and other animal genes and 10–20% of plant genes are alternatively spliced (Wang and Brendel 2006; Kim et al. 2007). Because sequences of alternatively spliced proteins will be highly similar outside the alternatively spliced regions, it may seem counterintuitive to anticipate a poor alignment—yet, the limitations of MSA programs need to be taken into account if realistic alignments are to be achieved. This is all the more important when a large volume of sequence data make manual fine-tuning of each individual alignments impracticable.

Previous studies of MSA programs have compared alignment methods for most general applications, with a focus on divergent sequences (Thompson et al. 1999b; Wallace et al. 2005; Edgar and Batzoglou 2006; Morrison 2006). Here we present a specific investigation of popular MSA programs on both simulated and real data sets containing overlapping and non-overlapping deletions. Our purpose was to identify the difficulties encountered by common alignment programs when confronted with sequences that contain deletions and to highlight the need for caution when dealing with MSAs, particularly where the output is to be fed through a pipeline to downstream applications.

Materials and Methods

Data sets used for the analysis of gap placement were constructed from sequences selected on the basis of their

Key words: multiple sequence alignment, ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS, T-COFFEE.

E-mail: lars.jermiin@usyd.edu.au.

Mol. Biol. Evol. 24(11):2433–2442. 2007
doi:10.1093/molbev/msm176
Advance Access publication August 20, 2007

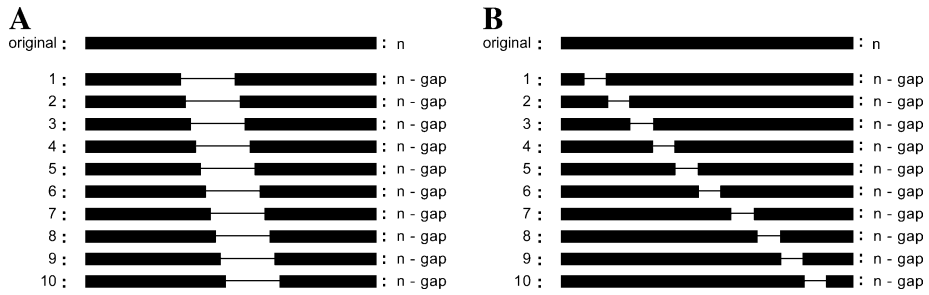

FIG. 1.—An amino acid sequence of size *n* was replicated 10 times, and a stretch of either 10 or 30 residues was deleted from each replicate, each time shifting the gap origin either (*A*) by one residue for overlapping deletions or (*B*) by the length of the deletion for non-overlapping deletions. A set of such gapped alignments was generated for each amino acid sequence, placing gaps along the length of the sequence.

different lengths and organisms of origin to control for sequence-specific effects in the alignments. The sequences used were the human erythrocyte membrane protein band 4.1 (EPB, 641 amino acids; GenBank accession AAA35797; Baklouti et al. 1997), the fibrinogen/fibronectin-binding protein (FBP) from *Staphylococcus epidermidis* RP62A (565 amino acids; YP_188358; Gill et al. 2005), and the necrosis and ethylene-inducing peptide 1 (NEP1) from *Fusarium oxysporum* (253 amino acids; AAC97382; Nelson 1998).

Each sequence was replicated 10 times, and regions of 10 or 30 residues were deleted from each copy to yield MSAs containing gaps. Deletions were staggered vertically with a displacement of one residue for each subsequent sequence (fig. 1*A*) or a displacement equal to the size of the deletion (so that the deletions do not overlap: fig. 1*B*). Several data sets were derived for each sequence, initiating the deletions at different positions along the sequence length. A reference alignment was created for each data set, with gaps inserted at sites where residues had been deleted. Reference alignments were used to assess the accuracy of alignments produced by MSA programs.

Each set of sequences, comprising the original sequence and 10 modified copies of the original sequence, was aligned using 6 popular MSA programs: ClustalW (Thompson et al. 1994), DIALIGN-T (Subramanian et al. 2005), MAFFT (Katoh et al. 2002), MUSCLE (Edgar 2004), PROBCONS (Do et al. 2005), and T-COFFEE (Notredame et al. 2000). We also tested a program, PRANK (Loytynoja and Goldman 2005), that still appears to be under development. Unless otherwise stated, the recommended default parameters for maximizing accuracy were retained for each program. In the case of MAFFT, which offers 3 modes of operation, the generic G-INS-I mode for globally alignable sequences was selected (Katoh et al. 2005). Test alignments were compared against the corresponding reference alignments, and accuracy was estimated by calculating the sum-of-pairs score (SPS). The SPS is a metric commonly used to assess MSA program performance and has been described in detail elsewhere (see, e.g., Thompson et al. 1999a; Nuin et al. 2006). Briefly, given a test alignment and a reference alignment of the same sequences, the SPS is defined as the number of residue pairs aligned identically in the test and the reference alignments, divided by the total number of aligned residue pairs in the reference alignment. "Residue pairs" refer to the residues located at the same position in any 2 sequences within a given alignment. Correctly aligned residue pairs scored 1, residues aligned with gaps in both the test and reference alignments scored 0.5, and misaligned residue pairs scored 0. The final score ranged from 0 to 1, where 1 indicated that the test alignment was identical to the reference alignment. This was converted to percentages for presentation in graphs and figures. Summary views of alignments were visualized in GeneDoc 2.6 (Nicholas et al. 1997).

For tests using alternatively spliced sequences, the same MSA programs were used to align amino acid sequences of alternatively spliced isoforms of human EPB and the rat tropomyosin-α (Tpm1) exons 1–8 (derived from GenBank GeneID 24851; Cooley and Bergtrom 2001) with the corresponding complete exon sequences. Results were compared with reference alignments based on published splice maps (Baklouti et al. 1997; Cooley and Bergtrom 2001; Kim et al. 2007).

To compare the performance of MSA programs, SPS values for alignments were rank transformed as a first step for the Friedman test (a nonparametric equivalent to blocked analysis of variance; Friedman 1937). For each set of sequences to be aligned, the MSA program that produced the highest scoring alignment was given a rank of 1, the MSA program that produced the lowest scoring alignment was given a rank of 6, and tied values were given the average of their ranks. Rank transformed SPS values were compared using the Friedman test followed by Fisher's protected lowest significant difference (LSD) for multiple pairwise comparisons (Fisher 1935). Results were visualized using the *R* statistics package, version 2.3.0 (R Development Core Team 2006).

Results and Discussion

Alignments Containing Overlapping Deletions

To assess whether MSA programs could correctly place overlapping gaps within sequences that contained overlapping deletions, alignments were initially generated based on the sequence of EPB. The alignments were expected to include a diagonal band of gapped regions (fig. 1*A*). Because alignment algorithms, such as ClustalW, include heuristics that dictate the likelihood of opening a gap at a particular site (Thompson et al. 1994), several data sets were created from the original amino acid sequence, each time with the deletion initiated at different

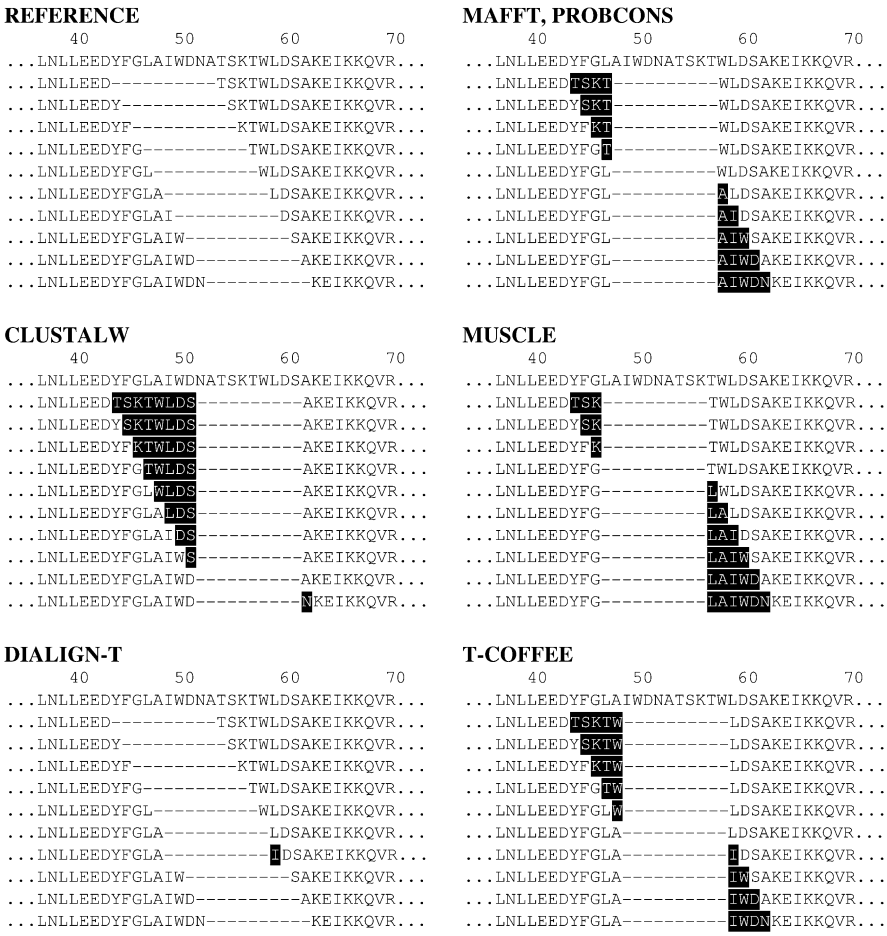

FIG. 2.—Summary view of alignments of highly similar sequences containing gaps (dashes) at overlapping positions. Misaligned residues are shaded. Alignments were generated by sequentially deleting a region of 10 residues from the EPB sequences and aligning the modified sequences with the original sequence using ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS, and T-COFFEE. Results were compared with the reference alignment (REFERENCE).

sites in the other sequences. This allowed detection of any contextual effect around the gap placement site.

Of the programs analyzed, ClustalW, MAFFT, MUSCLE, PROBCONS, and T-COFFEE preferentially aligned gaps in a single vertical column rather than the expected diagonally staggered band (fig. 2). Gaps were inserted at the same position in all sequences, flanked in each case by a region of more or less poorly aligned nonhomologous residues. This alignment of gaps was observed for all tested sequences, regardless of sequence length and gap length, and was unaffected by the input order of the sequences within a given alignment. However, different programs opened the gap at different positions (fig. 2).

The exception was DIALIGN-T, which was able to recreate the correct staggered arrangement in over 60% of the tested alignments (figs. 2 and 3). This was expected, as the DIALIGN-T algorithm has partial local alignment capability, and seeks to align large, highly similar regions at an early stage (Subramanian et al. 2005). None of the other programs produced the staggered alignments required; the high overall scores for these programs reflect the length of the deletions compared with the total length of the identical parts of the sequences.

Because compression of gapped regions, termed overalignment, is a common phenomenon in global alignment methods such as those evaluated here (Morrison 2006), we also examined the performance of PRANK (Loytynoja and Goldman 2005). This program attempts to generate alignments that are true to the evolutionary history of the sequences, even where this may entail the insertion of more gaps than would be expected for more conventional alignment methods. The result is that PRANK is not expected to align gaps where a better alignment can be achieved by lining up the residues around the gaps. However, aligning the same data set with PRANK produced a dramatically poorer alignment than did any of the other methods, including misalignment of large blocks of identical residues, placement of one or more sequences downstream of the rest of the alignment, and other artifacts (fig. 4). As with DIALIGN-T and the other MSA programs, gap placement was context dependent, but there was a much greater variability in overall alignment accuracy for PRANK, ranging from percent-converted SPS of 38.4–97.4%. By contrast, the variability in accuracy of alignments produced by DIALIGN-T, the most variable of the other MSA programs, ranged between 99.1% and 100%. PRANK's dramatically reduced accuracy

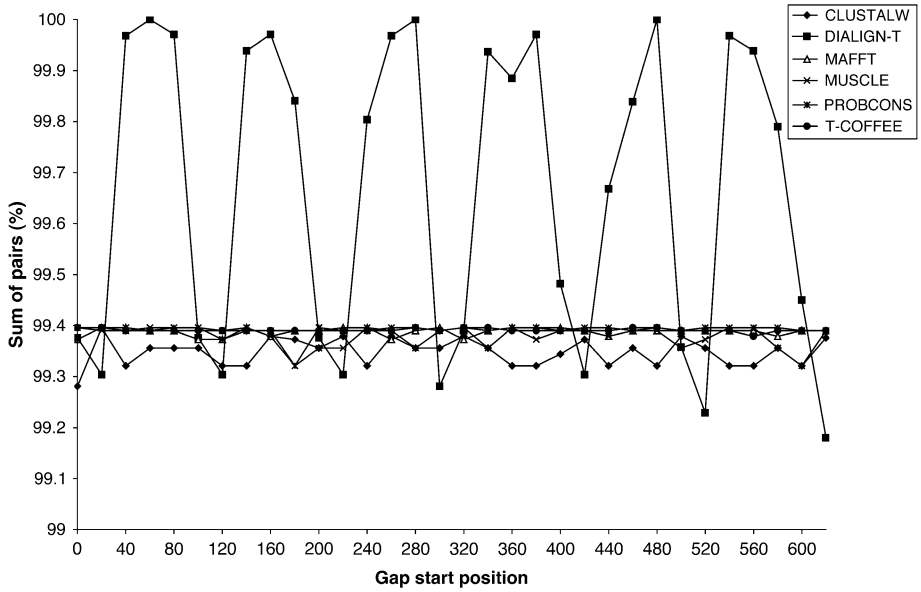

FIG. 3.—Alignment scores for ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS, and T-COFFEE on data sets containing alignments with overlapping deletions of length 10 (1.6% of sequence length). Alignments were generated from the EPB sequence. Each data point represents and individual alignment, with deletions originating at the position indicated.

and very slow computation speed did not make it a viable competitor to the other MSA programs tested, at least in its current implementation. For these reasons, PRANK was excluded from further tests.

It has been previously reported that the relative length of deletions has a small detrimental effect on alignment accuracy (Nuin et al. 2006). We therefore assessed the performance of MSA programs when the deletions were increased from 10 residues to 30 residues for the same sequence (EPB, 641 amino acids), as well as for 2 other sequences, FBP (565 amino acids) and the shorter sequence, NEP1 (253 amino acids). Sequences were duplicated for alignments, and deletions were positioned as before.

All MSA programs tested were affected by the length of the deletions, with the exception of DIALIGN-T, for which the same alignment accuracy was observed for deletions of 30 residues as for deletions of 10 residues ($P < 0.05$, Fisher's LSD; table 1). However, the reduction in SPS values due to longer deletions was small compared with the reduction in SPS values due to shorter alignment length (fig. 5). When the gapped region formed a similar proportion of the total alignment length, alignments based on the longer EPB sequence scored higher than those based on the shorter NEP1 sequence (fig. 5*B* and *C*). This can be explained by considering that a deletion of given size causes a constant number of misaligned residue pairs, regardless of alignment length. Because the SPS metric involves dividing the number of misaligned residue pairs by the total length of the alignment, a longer alignment scores higher than a shorter alignment with the same number of misaligned residue pairs. It should be noted that the different locations of the data do not affect the nonparametric

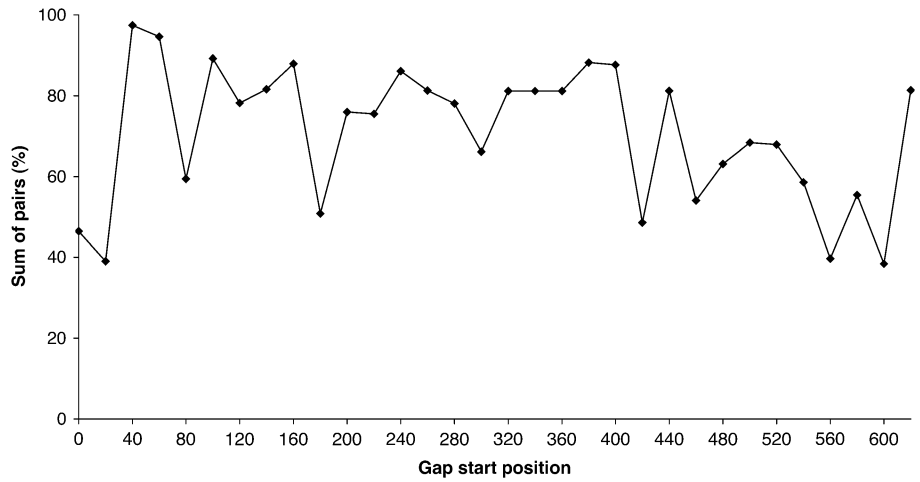

FIG. 4.—Alignment scores for PRANK on data sets containing alignments with overlapping deletions of length 10 (1.6% of sequence length). Alignments were generated from the EPB sequence. Each data point represents and individual alignment, with deletions originating at the position indicated.

Table 1
Accuracy (Mean SPS %) of 6 MSA Programs for Alignments Containing Deletions of 10 or 30 Residues

Mean SPS (%)

	Overlapping Deletions		Non-overlapping Deletions	
Gap Size (residues)	10	30	10	30
ClustalW	99.14	99.08	96.88	87.86
DIALIGN-T	99.47	99.67	99.26	99.25
MAFFT	99.19	99.15	99.94	99.93
MUSCLE	99.17	99.13	98.51	94.72
PROBCONS	99.19	99.14	97.03	96.62
T-COFFEE	99.19	99.15	95.02	90.60

Friedman test, which examines only the relative ranks at each data point.

Alignments Containing Non-overlapping Deletions

The second test set of artificially generated sequences comprised alignments of identical amino acid sequences with non-overlapping deletions (fig. 1*B*). This pattern of deletions resembles the arrangement expected for amino acid sequences encoded by isoforms of alternatively spliced mRNAs, where one or more exons may be omitted.

Once again, most MSA programs demonstrated a tendency to align gapped regions vertically, resulting in over-alignment of the nonhomologous flanking regions (fig. 6). Furthermore, the MSA programs frequently inserted additional gaps into the original sequence and into modified copies of the original sequence. Due largely to these extra gap insertions, the alignments often scored significantly lower than the comparable alignments with overlapping gaps, although most programs did manage to insert at least some gapped regions at the correct positions.

Alignment accuracy for this test set varied greatly for the different MSA programs (fig. 7). As was the case for alignments containing overlapping deletions, DIALIGN-T performed well but inconsistently, and its placement of gapped regions was once again context dependent. Remarkably, however, MAFFT produced alignments as good as or better than those produced by DIALIGN-T, and unlike DIALIGN-T, it did not exhibit context-dependent gap placement on any of the alignments generated from sequences of EPB, NEP1, or FBP. Consequently, alignments by MAFFT had near-perfect accuracy (>98.9% SPS compared with the reference alignments for all data sets tested).

When the length of deletions was increased from 10 residues to 30 residues, MAFFT and DIALIGN-T were the only programs unaffected by the length of deletions. Alignment scores for other MSA programs, particularly

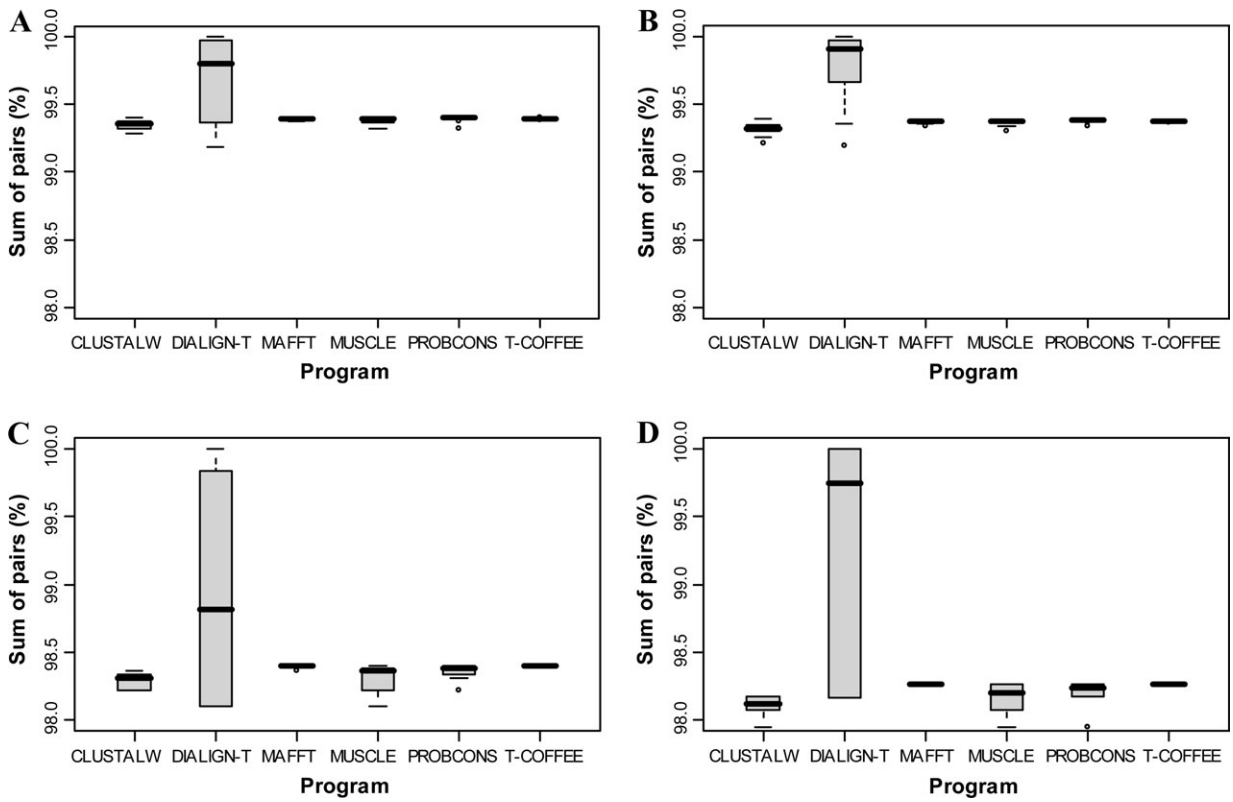

FIG. 5.—Effect of sequence length and gap of deletion on accuracy of alignments containing overlapping deletions. Boxplot indicates spread of SPS (%) for alignments based on EPB sequence with (*A*) deletions of 10 residues (1.6% of alignment length) and (*B*) deletions of 30 residues (4.7% of alignment length) compared with alignments based on the shorter NEP1 sequence with (*C*) deletions of 10 residues (4% of alignment length) and (*D*) deletions of 30 residues (11.9% of alignment length). Boxes indicate the interquartile range (middle 50% of the data); thick horizontal lines denote the medians. Plot whiskers (dashed vertical lines capped by horizontal lines) extend to the most extreme data point up to 1.5 times the interquartile range away from the box. Dots denote outliers.

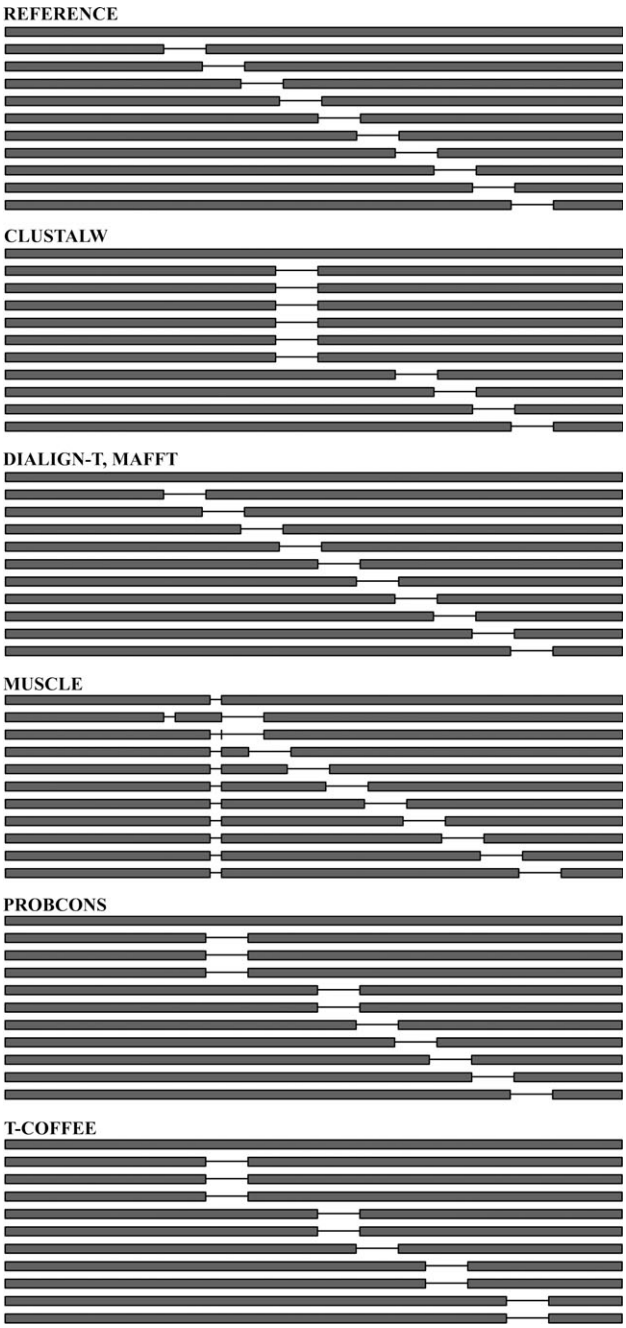

FIG. 6.—Summary view of alignments of highly similar sequences (boxes) containing deletions (horizontal lines) at non-overlapping positions. Alignments were generated by sequentially deleting a region of 10 residues from the EPB sequences and aligning the modified sequences with the original sequence using ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS, and T-COFFEE. Results were compared with the reference alignment (REFERENCE).

ClustalW and T-COFFEE, dropped markedly for the longer deletion (table 1).

Overall scores for each MSA program were ranked from 1 (most accurate) to 6 (least accurate) for each of the data points as a first step for the nonparametric Friedman test. As is evident from the ranked data, MAFFT produced the best scoring alignments most often followed by DIALIGN-T (fig. 8). By contrast, ClustalW and T-COFFEE performed quite poorly, particularly in alignments with longer deletions. Of these, T-COFFEE generally performed worse than ClustalW. This is notable, given that T-COFFEE is one of the newer MSA programs and has been previously shown to be highly accurate when aligning closely related sequences, including those with high frequency of insertions and deletions (Nuin et al. 2006).

Context-dependent gap placement was examined further by reversing the amino acid sequences from the C to the N terminal and realigning these using the different MSA programs. In the absence of structural information, all MSA programs were expected to place the gap at the same position in both the forward and the reverse orientations. Instead, the most common observation on the reversed alignments was that the gap position was shifted up- or downstream by one or more columns compared with the forward orientation. T-COFFEE and ClustalW were the most likely to align gapped regions differently in the forward and reverse orientations for the data sets with overlapping and non-overlapping gaps, respectively (table 2). The shift in gap placement was not due to the original position of the deletion within each sequence.

Alignments of Alternatively Spliced Sequences

To determine whether the results observed for alignments of artificially generated sequences were comparable with those from biological data, alignments were manually generated from published sequences of proteins derived from differentially spliced isoforms of EPB (used in the previous 2 test sets) and rat tropomyosin-α (Tpm1). The alignments generated by all MSA programs were compared with these references (Baklouti et al. 1997; Cooley and Bergtrom 2001).

The trends observed for the alternatively spliced sequences were the same as those observed for the artificially generated sequences. Thus, gapped regions were preferentially aligned at the expense of misalignment of flanking residues or entire exons. This was the case for all MSA programs with the exception of DIALIGN-T and MAFFT and was particularly poorly done by ClustalW (figs. 9 and 10). Overall, the results were similar to those observed for alignment of non-overlapping deletions (fig. 11). It was noted that the alignments by MAFFT could occasionally be improved further by selecting one of the local alignment modes (L-INS-I or E-INS-I), but as this study focused on the recommended generic high-accuracy options for each program, we did not include results obtained with these other parameter optimizations.

Conclusions

Alignments of similar sequences, although not as difficult a problem as alignments of highly divergent sequences, present a considerable challenge to commonly used MSA programs. Our results indicate that despite improvements in the accuracy of modern MSA programs, even the reportedly highly accurate programs, such as T-COFFEE (Notredame et al. 2000), encounter difficulties with the placement of gaps, with resulting reductions in accuracy

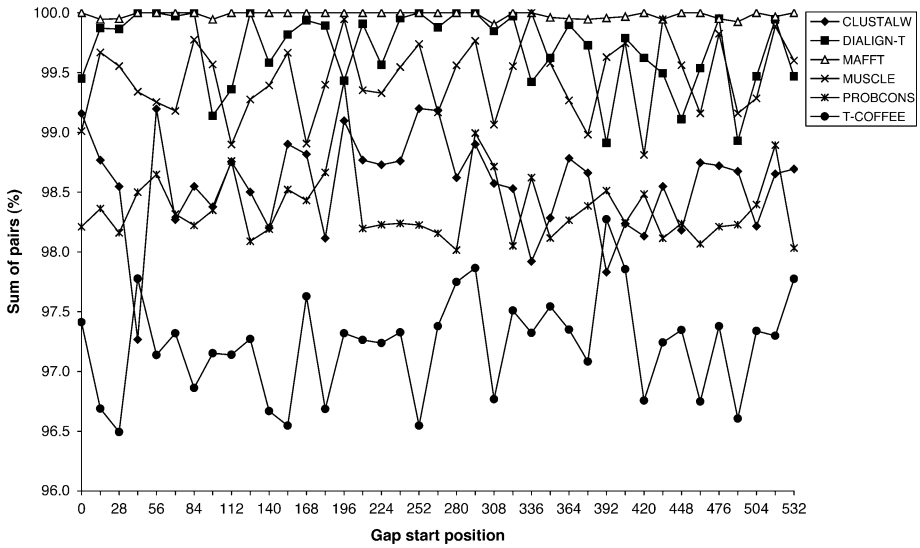

FIG. 7.—Alignment scores for ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS, and T-COFFEE on data sets containing alignments with non-overlapping deletions of length 10 (1.6% of sequence length). Alignments were generated from the EPB sequence. Each data point represents and individual alignment, with deletions originating at the position indicated.

that, in some cases, drop below the level of the older program ClustalW. On the other hand, the very accurate and fast MAFFT program performs extremely well on alignments containing non-overlapping deletions, even when its local alignment options (L-INS-I or E-INS-I) are not selected. Alignments by MAFFT using the global G-INS-I option were generally as good as or better than those by DIALIGN-T and T-COFFEE, both of which have local alignment capabilities. This is particularly important, as MAFFT has also been shown to be more accurate than DIALIGN-T on other types of data (Edgar 2004; Nuin et al. 2006), and may thus be a good general option when selecting an MSA program both for alignments of highly similar sequences and for alignments with a greater level of sequence divergence.

It is easy to understand why progressive alignment methods, such as ClustalW, perform poorly in the context of the present data sets. Progressive alignment methods first produce a guide tree (based on alignment scores obtained from the pairwise sequence alignments) and then use the guide tree to direct the progressive production of the MSA. This means that if a gap is inserted incorrectly in one of the first sequences to be included in the MSA, then that gap will be present also in the final MSA. T-COFFEE addresses this problem by using the guide tree and the pairwise alignments concomitantly during the progressive production of the MSA. Notredame et al. (2000) reported "a significant increase in alignment accuracy," but our study shows that this is not always the case. A possible explanation for T-COFFEE's inability to infer the correct MSA here might be that the guide tree's contribution to the final MSA is too large. It is difficult to explain why MAFFT performs better than MUSCLE and PROBCONS, given that they all incorporate iterative refinement as a default option. PROBCONS is trained on alignments from BAliBASE (Thompson et al. 1999a), so it is possible that values of alignment parameters may not be suitable for the present data. Finally, it is likely that the implementations of the iterative refinement differ among these programs. A

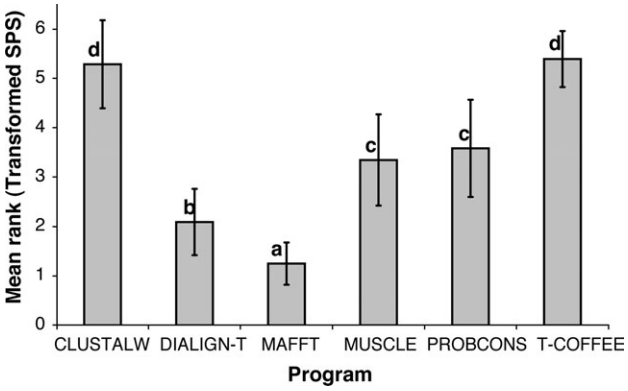

FIG. 8.—Rank transformed alignment scores (SPS) for alignments containing non-overlapping deletions. SPS values were transformed into ranks as a first step for Friedman analysis. *Y* error bars indicate standard deviation from the mean. Values marked by the same letter were derived from data sets not significantly different ($P < 0.05$) using Fisher's protected LSD for ranked data.

Table 2
Proportion of Alinments Where deletions Were Inserted in the Same Position When the Amino Acids Were Reversed from the C to the N Terminal

Proportion of Alignments Identical in Forward and Reverse Orientations (%)		
Program	Overlapping Deletions	Non-overlapping Deletions
ClustalW	40.6	0.0
DIALIGN-T	30.0	92.3
MAFFT	29.0	30.8
MUSCLE	25.0	5.1
PROBCONS	30.0	71.8
T-COFFEE	13.0	2.6

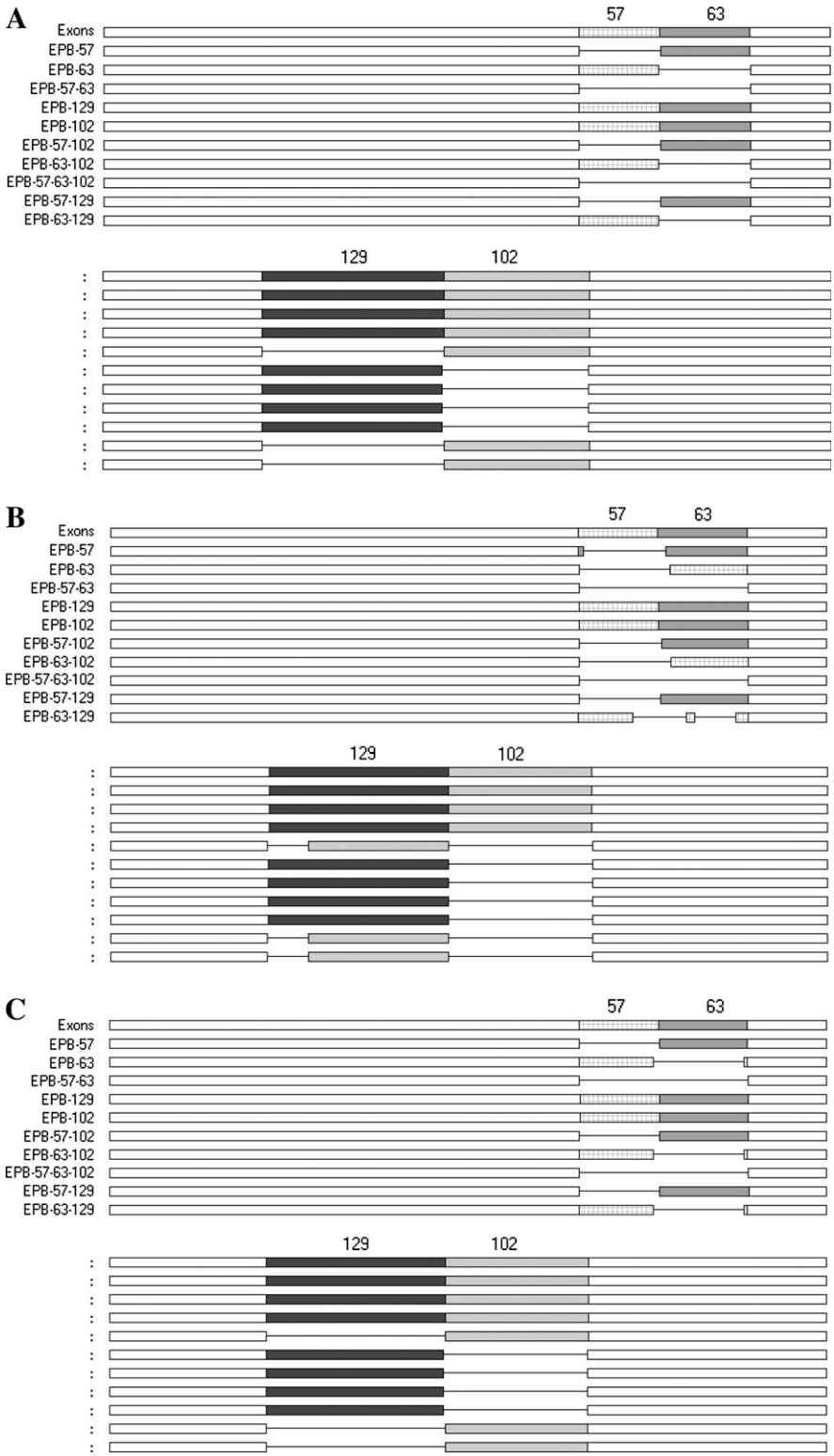

FIG. 9.—(*A*) Schematic representation of translated exons in the 80 kDa variant of the alternatively spliced EPB gene product. Complete exon sequence (Exons) shown aligned with 10 alternative spliced isoforms. The same sequences aligned by (*B*) ClustalW and (*C*) DIALIGN-T.

plausible reason for the good performance of DIALIGN-T is that it uses a segment-based approach and an objective function that 1) assumes the sequences are unrelated and 2) ignores the gaps in the alignment (Subramanian et al. 2005).

In view of the variability in the performance of all MSA programs on the different alignment types, reliance on a single MSA program for all alignment needs is undesirable. It has been previously recommended that 1) 2 or more alignment programs be used and 2) the results be

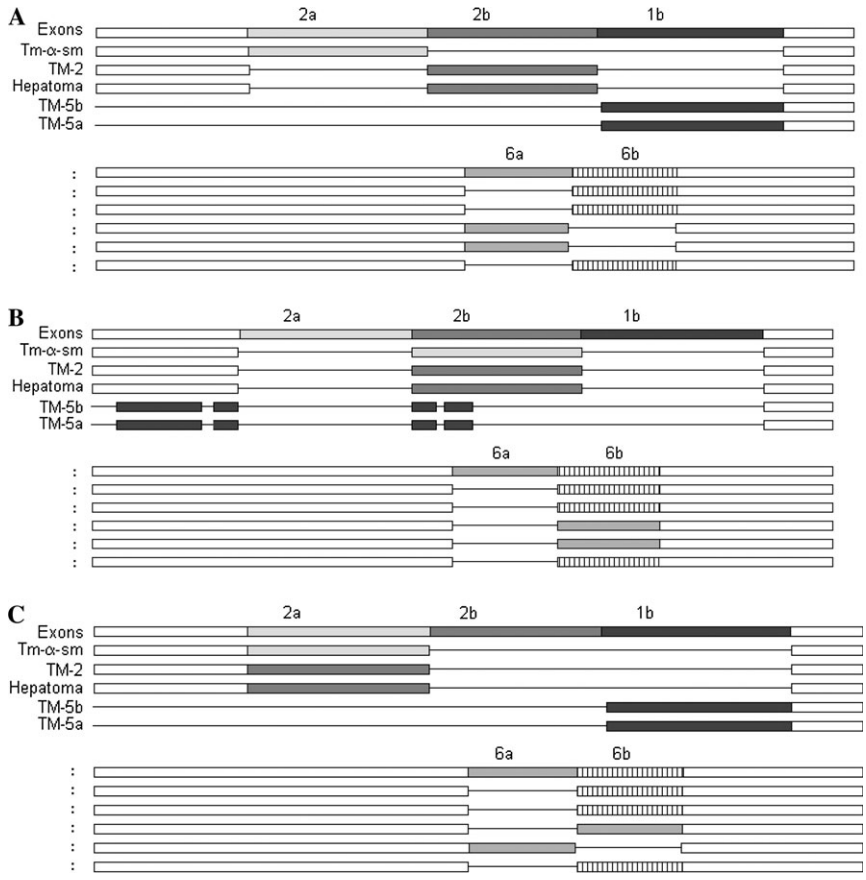

FIG. 10.—(*A*) Schematic representation of translated exons in the alternatively spliced Tpm1 gene product. Complete exon sequence (Exons) shown aligned with 5 alternative spliced isoforms. The same sequences aligned by (*B*) ClustalW and (*C*) DIALIGN-T.

examined for differences in gap placement (Edgar and Batzoglou 2006; Morrison 2006). Where appropriate, such as for phylogenetic analysis, columns that align differently could then be eliminated. We reiterate these recommendations and suggest MAFFT as a good candidate program for initial alignments, particularly where little is known about the data. Likewise, we recommend the Heads or Tails (HoT) approach to scoring MSAs (Landan and Graur 2007). For a given MSA program, HoT involves comparing the forward alignment (i.e., an alignment where the residues occur in their normal order) with the reverse alignment (i.e., an alignment produced after reversing the order of the residues). Where a second MSA program is not available, the HoT approach could be used to reduce context-dependent variability in gap placement by demarcating columns that differ between the forward and reverse alignments. These can be eliminated using alignment editors such as JalView (Clamp et al. 2004), SEAVIEW (Galtier et al. 1996), or GDE (Smith et al. 1994). Neither of these methods should be considered foolproof, however, and where structural information or other experimentally determined data are available, these should be taken into consideration.

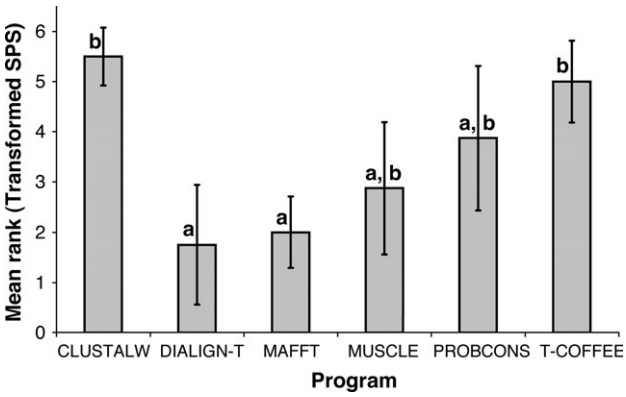

FIG. 11.—Rank transformed alignment scores (SPS) for alignments of isoforms of alternatively spliced genes. SPS values were transformed into ranks as a first step for Friedman analysis. *Y* error bars indicate standard deviation from the mean. Values marked by the same letter were derived from data sets not significantly different ($P < 0.05$) using Fisher's protected LSD for ranked data.

The arrangement of gaps presented here is extreme compared with most biological situations, but our analysis clearly shows that many MSA programs bias the placement of gapped regions. Our use of sequences that are identical except for the presence of deletions demonstrates the specific contribution of deletions to the alignment problem, which is expected to become progressively more difficult to solve as sequences diverge. It is not known to what extent this will affect downstream analysis, such as phylogenetic tree reconstruction and identification of conserved motifs. However, incorrect placement of gaps represents a source of error in any analysis that can be minimized by careful

application of an alignment strategy based on judicious choice of the available algorithms.

Supplementary Data

Files with unaligned and correctly aligned isoforms of the human erythrocyte membrane protein band 4.1 (EPB) and the rat tropomyosin-α (Tmp1) are available from http://www.bio.usyd.edu.au/jermiin/publications.html and *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was partly funded by Discovery Grants to S.E. (DP0450066) and L.S.J. (DP0453173 and DP0556820) from the Australian Research Council. The authors gratefully acknowledge the constructive comments and suggestions from T. M. Embley, C. L.-C. Ip, K. Katoh, A. W. D. Larkum, and 2 anonymous reviewers.

Literature Cited

Baklouti F, Huang S-C, Vulliamy TJ, Delaunay J, Benz EJ. 1997. Organization of the human protein 4.1 genomic locus: new insights into the tissue-specific alternative splicing of the pre-mRNA. Genomics. 39:289–302.

Caporale LH. 2006. The implicit genome. Oxford: Oxford University Press.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. Bioinformatics. 20:426–427.

Cooley BC, Bergtrom G. 2001. Multiple combinations of alternatively spliced exons in rat tropomyosin-[alpha] gene mRNA: evidence for 20 new isoforms in adult tissues and cultured cells. Arch Biochem Biophys. 390:71–77.

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

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

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

Fisher RA. 1935. The logic of inductive inference. J R Stat Soc [Ser A]. 98:39–54.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 32:675–701.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci. 12:543–548.

Gill SR, Fouts DE, Archer GL, et al. (29 co-authors). 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J Bacteriol. 187:2426–2438.

Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

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.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35:125–131.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Lassmann T, Sonnhammer ELL. 2002. Quality assessment of multiple alignment programs. FEBS Lett. 529:126–130.

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

Morrison DA. 2006. L.A.S. Johnson Review No. 8. Multiple sequence alignment for phylogenetic purposes. Aust Syst Bot. 19:479–539.

Nelsoni AJ. 1998. Sequence announcements. Plant Mol Biol. V38:911–912.

Nicholas KB, Nicholas HB Jr., Deerfield DW II. 1997. GeneDoc: analysis and visualization of genetic variation. EMBNEW News. 4:14.

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

Nuin P, Wang Z, Tillier E. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 7:471.

Plewniak F, Bianchetti L, Brelivet Y, et al. (16 co-authors). 2003. PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res. 31:3829–3832.

R Development Core Team. 2006. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM. 1994. The genetic data environment an expandable GUI for multiple sequence analysis. Bioinformatics. 10:671–675.

Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B. 2005. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics. 6:66.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thompson JD, Plewniak F, Poch O. 1999a. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics. 15:87–88.

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

Thompson JD, Thierry JC, Poch O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics. 19:1155–1161.

Wallace IM, Blackshields G, Higgins DG. 2005. Multiple sequence alignments. Curr Opin Struct Biol. 15:261–266.

Wang B-B, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA. 103:7175–7180.

Martin Embley, Associate Editor

Accepted August 14, 2007