BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 4 2012, pages 495–502
doi:10.1093/bioinformatics/btr701

Sequence analysis

Advance Access publication December 23, 2011

Measuring the distance between multiple sequence alignments

Benjamin P. Blackburne and Simon Whelan*

Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, Manchester M13 9PT

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Multiple sequence alignment (MSA) is a core method in bioinformatics. The accuracy of such alignments may influence the success of downstream analyses such as phylogenetic inference, protein structure prediction, and functional prediction. The importance of MSA has lead to the proliferation of MSA methods, with different objective functions and heuristics to search for the optimal MSA. Different methods of inferring MSAs produce different results in all but the most trivial cases. By measuring the differences between inferred alignments, we may be able to develop an understanding of how these differences (i) relate to the objective functions and heuristics used in MSA methods, and (ii) affect downstream analyses.

Results: We introduce four metrics to compare MSAs, which include the position in a sequence where a gap occurs or the location on a phylogenetic tree where an insertion or deletion (indel) event occurs. We use both real and synthetic data to explore the information given by these metrics and demonstrate how the different metrics in combination can yield more information about MSA methods and the differences between them.

Availability: *MetAl* is a free software implementation of these metrics in Haskell. Source and binaries for Windows, Linux and Mac OS X are available from http://kumiho.smith.man.ac.uk/whelan/software/metal/.

Contact: simon.whelan@manchester.ac.uk

Received on July 19, 2011; revised on December 06, 2011; accepted on December 19, 2011

1 INTRODUCTION

Many methods used in bioinformatics require one or more accurate multiple sequence alignments (MSAs) as input. Each MSA arranges a set of homologous amino acid or nucleotide sequences in a matrix, where each column of the matrix corresponds to a set of characters that are homologous, functionally related or superposable in a protein structure (Edgar and Batzoglou, 2006). These definitions may coincide when the sequences are closely related, but may conflict as the sequences diverge. Methods of phylogenetic tree reconstruction (Felsenstein, 2003), structure prediction (Arnold *et al.*, 2006), functional annotation (Eisen, 1998) and the creation of profile hidden Markov models for database searching (Eddy, 2009) all depend on an alignment, and the performance of these methods depends inevitably on the accuracy of these alignments. A variety of MSA methods have been developed over the past two decades that are still in active use (Notredame, 2007), including methods that co-estimate trees and alignment (Hagopian *et al.*, 2010; Liu *et al.*, 2009; Redelings and Suchard, 2005). These algorithms vary in their objective function and the heuristic used to find the best MSA.

The performance of MSA methods is frequently assessed by their ability to recover a reference alignment (e.g. Golubchik *et al.*, 2007), which may be produced from biological information, such as a conserved structure (Thompson *et al.*, 2005). The output of each method can then be scored against the reference alignment, either by the fraction of residue pairs in the reference alignment that are correctly identified by a given method, known as the Sum-of-Pairs (SP) score, or by the Total Column (TC) score, which describes the fraction of reference columns identified (Thompson *et al.*, 2005).

Several studies have noted the effect of sequence alignment algorithm on, for example, phylogenetic tree topology (Cantarel *et al.*, 2006; Hall, 2005; Morrison and Ellis, 1997; Ogden and Rosenberg, 2006; Wang *et al.*, 2011; Wong *et al.*, 2008). Many of these studies used simulated data and the parameter of interest was generally assessed against alignment quality, measured as TC or SP score against the true alignment. Neither the TC nor the SP score are true metrics because they violate the principles of symmetry and the triangle inequality (see below). Producing a valid metric would allow alignment to be projected as points in an alignment space, enabling the comparison of distances between different MSA methods or across different datasets. When these comparisons include a reference alignment, the metric would give some indication of the similarity of alignment methods as well as an accuracy score. Alternatively, such a metric could be used to characterize alignment uncertainty. Previous work on uncertainty has relied on annotating a 'best' alignment with uncertainty information based on the proportion of base pairs that are present when the sequences are reversed before alignment (Landan and Graur, 2007; Wise, 2010), or when the guide tree is varied (Penn *et al.*, 2010). A true metric could also provide direct comparison of equally or similarly high scoring paths through the dynamic programming matrix produced during MSA, allowing one to identify regions of high uncertainty in the final alignment.

In this study, we derive four true metrics for the comparison of MSAs: (i) a simple correction to the SP score; (ii) a metric that incorporates raw gap information; (iii) a metric that includes the position where gaps occur in a sequence; and (iv) a metric that includes the position where insertion/deletion (indel) events occur both in a sequence and on a phylogenetic tree. We proceed to demonstrate that the SP and TP score are not metrics and cannot be used to investigate many questions about how MSA methods perform. We then show the usefulness of our metrics on MSAs produced by a selection of methods on real data extracted from BAliBASE (Thompson *et al.*, 2005), and simulated data produced using INDELible (Fletcher and Yang, 2009). Through these analyses, we also demonstrate that the combined use of our metrics can identify similarities and differences between

*To whom correspondence should be addressed.

MSA methods and characterize the rate of decline in alignment performance over increasing evolutionary distances.

2 METHODS

2.1 General definitions

Consider a set of n sequences $\mathbf{S}=\{S^1,S^2,...,S^n\}$ where the j-th character in sequence i is denoted S^i_j and the sum of the sequence lengths is given by $c=\sum_i|S^i|$. Also consider a phylogenetic tree, T, that describes the evolutionary relatedness of the sequences in $\mathbf{S}$ by a series of edges, $\mathbf{e}=\{e_1,e_2,...,e_{2n-3}\}$ where e_k describes an edge in the tree that splits $\mathbf{S}$ into two non-empty sets. Distances between MSAs should compare where in the MSAs each observable character, S^i_j, is placed in relation to the characters in the other sequences. We begin by assigning S^i_j to a homology set, which contains the characters in the other sequences that share a common ancestor with S^i_j. Where there is no such relationship due to an indel, the set may include a character representing the gap state. When assigning such homology sets, we may also wish to include additional information in the labelling of the gap states to alter the information given by the distance. First, we can include information about the placement of the gap characters in their respective sequences. Secondly, we can include information about the likely edge that corresponds to the location of the indel event on the tree that resulted in the gap.

2.2 Recoding the insertion and deletion history of an alignment

The next step in computing a distance between alignments is to recode the alignment so that all the characters and gaps can be correctly compared. The transition from $\mathbf{S}$ to an alignment requires the placement of gap characters, which represent all the indels that have occurred during the history of the sequences. We assume no direction to evolution and, therefore, cannot differentiate between insertions or deletions. To incorporate information from indels, we recode gap characters in our alignments in one of four ways (see Fig. 1 for examples):

(1) *SSP*: the 'Symmetrized SP' recoding ignores gaps and treats them as blanks in an alignment. The name reflects the similarity to the existing SP method for comparing alignments.

(2) *seq*: this recoding provides a simple record of gap information and treats all gaps in a sequence equally. Each gap is simply recoded as G^i, indicating it occurred in sequence i.

(3) *pos*: our third recoding incorporates the positional information about where a gap occurs in a sequence, but not the temporal (phylogenetic) location of the indel that produced the gap. Each gap is labelled as G^i_j, where j is the location of the real character to the left. In Figure 1, both gaps occurring after the first character in Sequence 3 are labelled as G^3_1. Note any gap occurring before the first character would be labelled G^i_0.

(4) *evol*: our final recoding includes all the information of *pos* recoding, and also incorporates where the indel event leading to that gap occurs in a phylogenetic tree. Each column in an alignment is considered independently and the indel history inferred as the most parsimonious set of events under Dollo parsimony (Huson and Steel, 2004). Dollo parsimony enforces that the history of each column is restricted to a maximum of one insertion and any number of deletion events, with the mapping of indel events to edges the same regardless of the tree rooting. Gap characters in the column are then recoded using the nomenclature $G^i_j(k)$, where k is an index used to label edges in the tree.

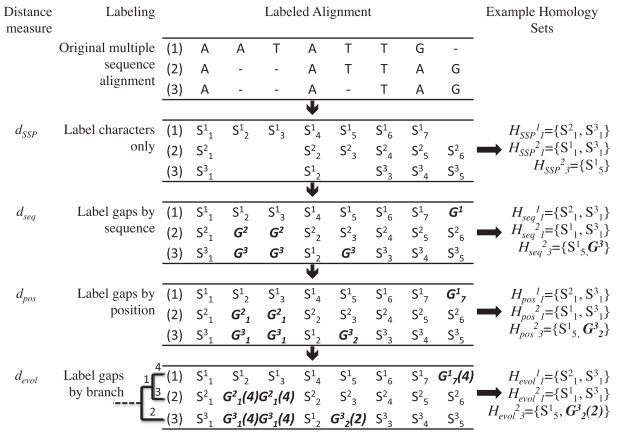

Fig. 1. Labelling of sequence alignments in order to determine distances by different metrics. Gap states are given in bold to highlight the difference between the four methods.

A

(1) A A A G T T - G
(2) A A A G - - A -
(3) A A A G T T A -

B

(1) A A A G T T G -
(2) A A A G - - - A
(3) A A A G T T - A

Fig. 2. Two alternative representations of the same alignment. Before our recoding procedure, the columns of representation B would be re-ordered to give representation A.

2.3 Enforcing a unique representation of an alignment

When comparing MSAs, the first step is to ensure that any given assignment of gaps has a unique representation in alignment space. When gaps are considered as independent events, the order of some columns in an alignment may be arbitrary. For example, the last two columns in Figure 2 could be placed in any order because they contain no overlapping bases. To enforce a unique alignment representation, we reorder sequences alphabetically by their names. We then sort the non-overlapping columns so that the leftmost column contains a gap in a higher row than the column to the right.

2.4 Including evolutionary information

The labelling step for the *evol* recoding requires a tree in order to label each gap by the inferred indel event that caused it, assuming that each position in a multi-site gap can be considered independently. Although this approach is not ideal from a biological perspective, it does allow efficient and simple computation of where indels occur on the tree. Allowing gaps to span multiple columns makes identifying their placement on the tree more difficult, because there may be many equally parsimonious solutions for gap placement (Simmons and Ochoterena, 2000), each of which may induce a different homology assignment to a site and result in a different distance between alignments. Our approach may be viewed as similar to using a linear gap penalty when identifying where an indel occurs.

2.5 Comparing homology sets

Once a pair of alignments are recoded appropriately the process of calculating the distances between them can begin. The first step is to calculate the site-wise homology set for each alignment under each metric, so for alignment A and metric $X\in\{SSP,seq,pos,evol\}$ we have $\mathbf{H}_X(A)=\{H_X(A)^i_j\}$. To create the homology set $H_X(A)^i_j$, one identifies the column in alignment A with character S^i_j and store all the other characters in that column, including gaps if they are labelled. Note that gaps do not have a homology set of their

own because they do not represent observable data and their presence varies between alignments. For example, taking the *evol* recoding from Figure 1 the homology set for sequence character S_3^2 is $H_{\text{evol}}(A)_3^2 = \{S_5^1, G_2^3(2)\}$. We define the d_{SSP} metric as the Jaccard distance on the homology sets H_{SSP}. Previous research has shown the Jaccard distance to be a valid metric (Lipkus, 1999).

$$d_{\text{SSP}}(A,B) = 1 - \frac{\sum_i \sum_j |H(A)_j^i \cap H(B)_j^i|}{\sum_i \sum_j |H(B)_j^i \cup H(A)_j^i|} \quad (1)$$

We call this measure of distance between alignments our Symmetrized SP (SSP) metric, reflecting its inspiration from the original SP score. This metric has a mixture of desirable and undesirable properties. It has a direct link to the information used by most methods of sequence analysis (homology pairs), but the contribution of base-gap pairs to the metric is less clear. By explicitly encoding gap characters, we can extend the metric to include other information that may be considered important.

For the remaining three metrics, we can define them such that they take advantage of the fact that the homology set sizes remain constant for any given alignment. In Figure 1, for example, one can see that for *seq*, *pos* and *evol* recoding every homology set is of size $n-1$, in contrast to *SSP* where the size of homology sets varies between alignments. We take advantage of this consistency to use the following to compute the distance metric $d_X(A,B)_j^i$ for $X \in \{seq, pos, evol\}$:

$$d_X(A,B)_j^i = \frac{|H_X(A)_j^i \triangle H_X(B)_j^i|}{|H_X(A)_j^i| + |H_X(B)_j^i|} \quad (2)$$

The numerator of this equation is the symmetric difference, which in our case is equivalent to the Hamming distance, as each member of $H_X(A)_j^i$ has a corresponding member in $H_X(B)_j^i$. Previous research has shown the Hamming distance is a true metric (Deza and Deza, 2009). The inclusion of gaps in the homology sets means the normalizing denominator in equation 2 is constant for any given alignment of the same set of sequences, and therefore $|H_X(A)_j^i| = |H_X(B)_j^i|$. A function remains a metric after normalization by a constant. The formulation of a per character homology distance allows us to compute two useful quantities. First, we can compute the distance between a pair of alignments by taking the average across all characters in the sequences.

$$d(A,B) = \frac{1}{c} \sum_i \sum_j d(A,B)_j^i \quad (3)$$

The second quantity of interest is a measure of how differently the characters of a single sequence S^i are incorporated into two MSAs. This metric could, in principle, be used to identify sources of error in aligners, particularly those associated with a guide tree.

$$d(A,B)^i = \frac{1}{|S^i|} \sum_j d(A,B)_j^i \quad (4)$$

2.6 Interpreting alignment distances

The metrics described here provide a distance between two alignments that lies between 0 and 1. Note that this distance can equally be expressed as a percentage. For d_{seq}, d_{pos} and d_{evol}, these distances can be interpreted as the probability that a randomly selected base (x) will be aligned to a different location against a sequence randomly selected from the $n-1$ sequences that do not contain x. The remaining metric, d_{SSP}, represents the fraction of pairings between observable characters that overlap in the two alignments. The limitations of this metric are best demonstrated by example. Consider an arbitrary alignment of n sequences each of length l. If we modify this alignment by breaking the homology between two bases and aligning them both with gaps, the distance between the two alignments under d_{seq} will be $\frac{2}{ln(n-1)}$. Under d_{SSP}, the distance will depend on how many pairwise homologies are present in the starting alignment; if more pairwise homologies are inferred, then the denominator will be larger, and so the difference between the two alignments will be smaller. In other words, the structure of one part of the alignment affects the distance caused by a change in another part of the alignment.

2.7 Alignment programs examined

To demonstrate the performance and usefulness of our four metrics, we use them to compare the alignments produced from a range of popular alignment programs with their default options on two different datasets. We consider the progressive aligners ClustalW (Larkin *et al.*, 2007); Muscle (Edgar, 2004); MAFFT's FFT–NS–i algorithm (Katoh and Toh, 2008), which includes additional refinement steps; DIALIGN–TX (Subramanian *et al.*, 2008), which includes an additional greedy matching step; the consistency aligners T–Coffee (Notredame, 2000), ProbCons (Do *et al.*, 2005) and the L–INS–i algorithm of MAFFT; and the phylogenetically aware aligner Prank, with the recommended '+F' option (Löytynoja and Goldman, 2008). Although other non-default parameters have the potential to improve the quality of alignments, our aim is not to judge alignment quality but to show how the metrics presented here reveal differences in the algorithms.

We examine two datasets, one synthetic, and one real (described below). For each dataset examined, we align the sequences using each of the procedures detailed above. We also include the true or reference alignment for comparison. Given these alignments, we can calculate the all-against-all set of distances under each of the four metrics proposed here. We visualise mean distance matrices using the heatmap.2 package from the gplots R package. The mean distance is equivalent to a renormalized Manhattan distance; results using the euclidean distance are similar but a mean distance is more interpretable.

Where d_{evol} was used a reference phylogenetic tree topology was required. For the BAliBASE alignments, a maximum-likelihood tree was generated using RAxML (Stamatakis, 2006) on the BAliBASE reference alignment. For the synthetic data, the reference tree was the tree used to generate the data.

2.7.1 Test data 1: BAliBASE We examine alignment performance on three sections of the BAliBASE 3 database, which has been widely used to assess the performance of alignment algorithms at recovering reference alignments derived from protein structure information. We use sections RV12 (medium-to-divergent sequences with 20–40% identity), RV30 (subfamilies) and RV50 (internal insertions), with 20 alignments randomly chosen from each section for subsequent analysis. Results are also presented for the truncated versions of the same sequences, containing only the homologous regions. Alignments contain between 4 and 142 sequences (mean of 34.9).

2.7.2 Test data 2: synthetic data Simulated data provides a useful test of alignment where all properties of the data are known, at the cost of not reflecting the complexity of real data. In this study, we use synthetic data to assess the effect of evolutionary distance on alignment. We generate data under a balanced tree topology with eight leaves, where all 14 branches are of equal length and the total tree branch length is one of $\{0.14, 0.42, 0.84, 1.4, 2.8, 7\}$. We also generate a second dataset with an unbalanced tree where each internal branch has equal length and the overall tree length is scaled the same as the corresponding balanced tree. All trees used are clock-like. INDELible v1.03 (Fletcher and Yang, 2009) was used to simulate data under the WAG (Whelan and Goldman, 2001) substitution matrix, with an indel rate parameter of 0.1 and a power-law indel length distribution with parameter $a = 1.7$. The root sequence has length 1000, and 20 replicates were made of each run.

3 RESULTS

3.1 The problem with existing scores

A non-negative function, $d(x,y)$, is a metric if it obeys the following conditions:

(1) $d(x,y) = 0$ if and only if $x = y$ (identity of indiscernibles).

(2) $d(x,y) = d(y,x)$ (symmetry).

(3) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality).

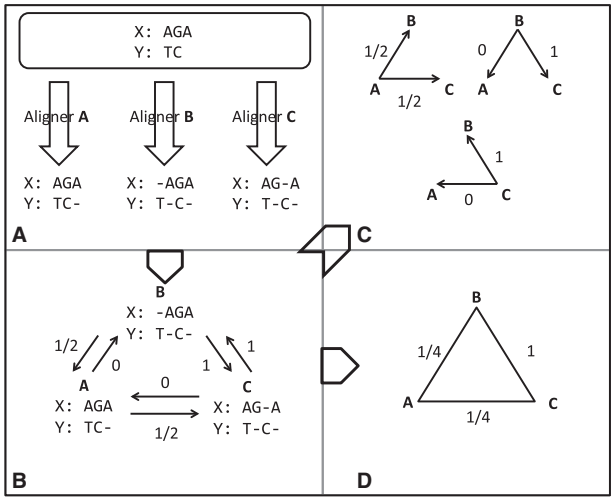

Fig. 3. The process of analysing the alignment of two toy sequences with three sequence aligners, using the SP score. See Section 3.1 for discussion. (**A**) Perform sequence alignments; (**B**) calculate SP differences; (**C**) choose one reference and compare; (**D**) use mean difference as pseudo-distances.

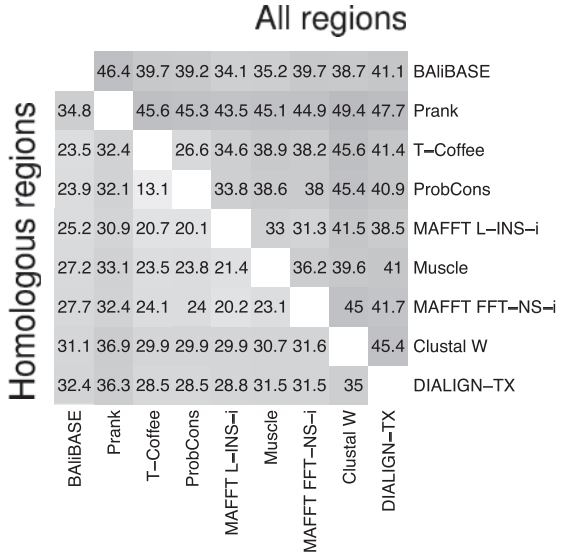

Fig. 4. Distance matrix computed using d_{evol}. The upper part of the matrix gives the (mean) distance between two alignments over all of BAliBASE RV12, RV30 and RV50, scaled as percentages. The lower part gives the same information but for the homologous regions only. Darker shades of grey indicate increasing distance.

We demonstrate that existing methods for comparing alignments are not metrics, starting with the widely used SP similarity (Thompson *et al.*, 2005), recoded as a dissimilarity, d_{SP}. Using our labelling definitions (Fig. 1), the dissimilarity $d_{\text{SP}}(A,B)$ between a test alignment A and a reference alignment B is:

$$d_{\text{SP}}(A,B) = 1 - \frac{\sum_i \sum_j |H_{\text{SSP}}(A)^i_j \cap H_{\text{SSP}}(B)^i_j|}{\sum_i \sum_j |H_{\text{SSP}}(B)^i_j|} \tag{5}$$

This dissimilarity score is not a true metric because it fails to satisfy the conditions of symmetry and the identity of indiscernibles. The asymmetry occurs because the denominator may be different for each alignment. Note this asymmetry is also true for the similarly defined TC score. The overlap score (Lassmann and Sonnhammer, 2002), another related measure of alignment similarity, introduces symmetry and is defined as:

$$Q_{ab} = \frac{\sum_i \sum_j |H_{\text{SSP}}(A)^i_j \cap H_{\text{SSP}}(B)^i_j|}{(\sum_i \sum_j |H_{\text{SSP}}(B)^i_j| + |H_{\text{SSP}}(A)^i_j|)/2} \tag{6}$$

This formulation is also known as Dice's Coefficient and its dissimilarity measure, defined as $1-Q_{ab}$, does not satisfy the triangle inequality and so is not a valid metric (Leach and Gillet, 2003). Failure to satisfy the triangle inequality means that the alignment distances A↔B, A↔C, and B↔C cannot be projected into Euclidean space. In other words, the path A→B→C may be shorter than the direct path A→C.

These theoretical failings expose a major limitation of existing scores: although they can calculate a measure of similarity to a reference alignment, they cannot be used for all-against-all comparisons, with or without a reference. This limitation is illustrated in Figure 3, where we show that existing metrics prevent any meaningful comparison between a set of toy MSAs. In Figure 3A, we align two short sequences using three alignment programs. In Figure 3B, we calculate SP differences between the three alignments, but because we need to normalize the SP (and TC) score using one of the alignments as a reference alignment, the distance (e.g.) A→B is different to the distance B→A, demonstrating that the SP (and TC) violates the symmetry requirement for a distance.

To demonstrate how the choice of reference alignment affects these scores, Figure 3C demonstrates the different conclusions drawn when each MSA is taken to be the reference. If we take A to be the reference, we infer that B and C are equally similar to A, perhaps concluding that B and C are equally good aligners. However, if we take B to be the reference alignment, we find A and B produce the same alignment, whereas C is maximally different. In this case, we would conclude B is an excellent method and C is a poor method.

One may try to solve these problems by taking the mean of the score A↔B and B↔A. This approach enforces symmetry, but the resultant distances do not obey the triangle inequality. In Figure 3D, we demonstrate that averaged distances result in a triangle between A, B and C that cannot exist in Euclidean space.

These observations make evident the need for a true alignment metric that allows accurate comparison between alignment methods. We have implemented the metrics described in Section 2 in the command-line program *MetAl*. Below, we demonstrate the utility of these metrics by examining the performance of MSA methods using real and simulated sequence data.

3.2 Comparing alignment methods using BAliBASE

Mean distances across the full BAliBASE alignments from RV12, RV30 and RV50 are plotted for d_{evol} as the upper part of Figure 4. The trends for other metrics are similar and generally highly correlated (d_{evol} and d_{seq} is the least correlated pair, with $r^2 = 0.68$). The different subgroups tested (RV12, RV30 and RV50) are also correlated, with $r^2 > 0.85$ for d_{evol}. Our results show that all methods produce alignments in which bases share the majority of their base–base or base–gap pairings with the reference alignment (distance:

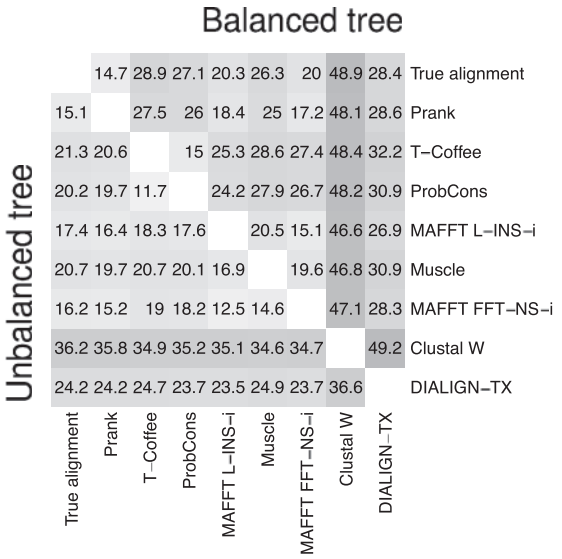

Fig. 5. Distance matrix computed using d_{evol}, as Figure 4 but for synthetic data on balanced (upper) and unbalanced (lower) trees. The tree length of the trees used to simulate the data was 1.4.

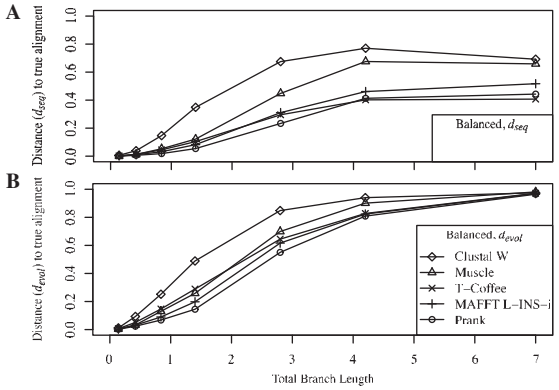

Fig. 6. Mean distances [(**A**) d_{seq} and (**B**) d_{evol}] between inferred and true alignments calculated for data simulated on increasing total tree length on balanced trees.

34.1–46.4%). Each method also tends to produce alignments that are substantially different from one another. Prank and ClustalW, for example, show the greatest dissimilarity (49.4%), whereas T–Coffee and ProbCons give the most similar alignments (26.6%). Prank has a noticeable tendency to have high distances to other aligners (43.5–49.4%).

These results suggest that although the different MSA methods are producing alignments that are roughly of equal similarity to the BAliBASE alignment, the way they achieve these alignments can be quite different. BAliBASE also provides alignments where only regions of identified homology are included. Results for these datasets are plotted below the diagonal of Figure 4. Trends are similar for both the full sequences and the homologous regions, although absolute distances are larger for the full sequences.

3.3 Comparing alignment methods using data simulated by INDELible

To investigate the performance of our metrics on alignments with a known origin and history, we apply the metrics to data simulated using INDELible on a balanced and an unbalanced tree, each with tree length of 1.4. The mean values of d_{evol} are plotted in a matrix in Figure 5 across the balanced (upper matrix) and unbalanced (lower matrix) datasets. These figures show that the differences between methods are more pronounced in simulated data relative to real data, with ClustalW performing noticeably worse than other methods. In both examples, the phylogenetically aware aligner, Prank, has the best performance, although MAFFT L–INS–i also performs well. The performance of Prank may be expected because its alignment model closely resembles that used by INDELible to generate the data.

To explore the effect of increasing evolutionary distance on accuracy, we examine the average distance from the true alignment for a subset of methods for d_{seq} and d_{evol} under a range of evolutionary divergences on balanced trees. The trends for unbalanced trees are similar. For d_{seq}, we find that as the evolutionary divergence increases, the distance between alignments inferred by the MSA methods also begins to increase, but it reaches a maximum distance when the tree length is ∼4, and then plateaus or declines (Fig. 6A). In contrast, for d_{evol} the distance between inferred MSAs and the true MSA steadily increases as evolutionary divergence increases (Fig. 6B). Curves for d_{pos} and d_{SSP} are similar to d_{evol}, reaching 1.0 at or after a tree length of 7. The differences between d_{seq} and d_{evol} in Figure 6 are due to the way they treat gaps in the divergent sequences. For d_{seq}, the decrease in alignment distance for divergent sequences occurs because the majority of amino acids are aligning with gap characters, which cannot be distinguished from one another. On the other hand, d_{evol} includes evolutionary and positional information about gaps, which allows the metric to correctly identify falsely inferred indel events.

Equation (4) makes it possible to calculate how differently a pair of MSAs treat a single sequence, potentially highlighting sequences that are aligned particularly inconsistently between two methods. We investigate this approach by examining the performance of the different MSA methods on simple balanced and unbalanced tree topologies. We expect that each sequence on the balanced tree will have a similar distance under the same method, but for the unbalanced tree the most distantly related sequences will be harder to align than the more closely related sequences. In Figure 7, we use simulation to investigate the distance from the true alignment of individual sequences for equal length balanced and unbalanced trees using the d_{evol} metric. The results match our expectations. For balanced trees, all leaf nodes are topologically identical, and the observed distances between sequence are broadly similar. For the unbalanced tree, we observe it is harder to align sequences from long terminal branches than those from short terminal branches.

3.4 Relationship between metrics

The structure of our metrics enforces the relationship $d_{evol} \geq d_{pos} \geq d_{seq}$, with the lack of gap information in d_{SSP} making it somewhat different from the other three metrics. Our results demonstrate that incorporating positional and evolutionary information can provide useful information to a bioinformatician. BAliBASE, example BB12038, contains one sequence that is much longer than the others,

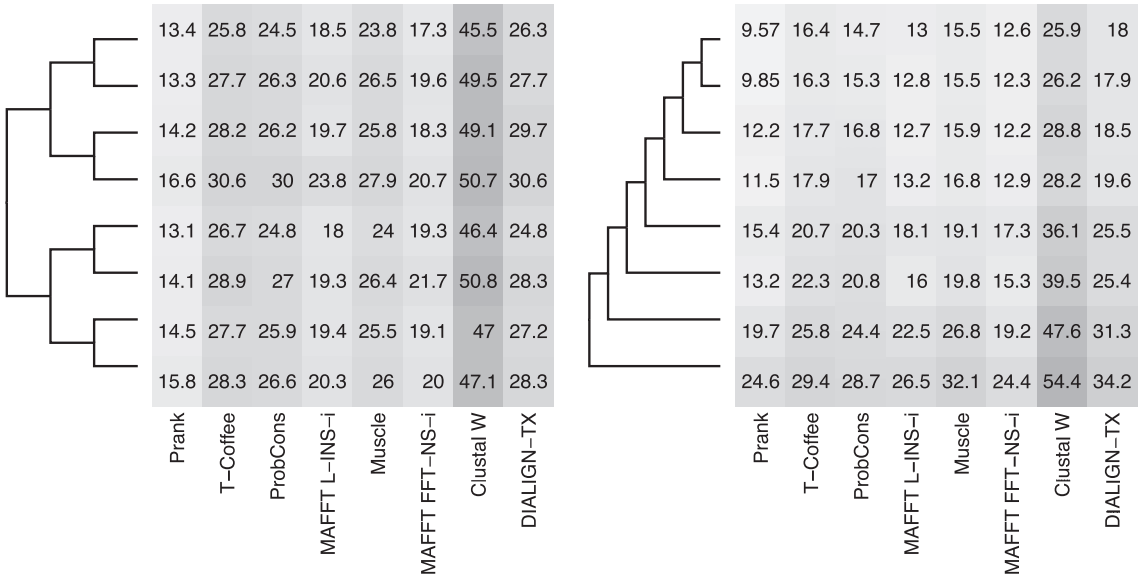

Fig. 7. The effect of tree topology on sequence-wise distances from the true alignment. Sequence-wise distances (d_{evol}) are similar (within a margin of error) on the balanced tree, but increase with terminal branch length for the unbalanced trees.

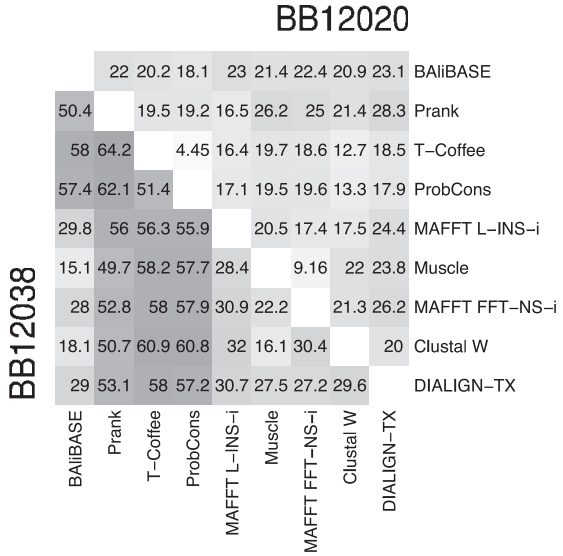

Fig. 8. Distance matrix computed using d_{evol}, as Figure 4 but for BAliBASE alignments BB12038 (lower) and BB12020 (upper).

with a length of 2314 aa relative to the next longest of 437 aa. The extreme nature of this alignment makes it a useful case study. Results comparing alignment methods on BB12038 are plotted in Figure 8 below the diagonal. For comparison, scores for a more typical alignment are plotted above the diagonal. By considering each of the four metrics in turn, using the distance from the reference sequence to ClustalW and Prank, we can evaluate the information encoded by each metric. In the case of d_{SSP}, the distance between ClustalW and the reference is 0.37, so of all the base–base pairs across the two alignments, 63% of them are present in both. For Prank, the distance is larger (0.44). Substantially lower distances are seen under d_{seq} (Prank = 0.19 and ClustalW = 0.14). The d_{SSP} distance is not nested in the other distances, so care should be taken when comparing the other distances with d_{SSP}. In this case, we interpret the large decrease as evidence that the alignments are quite gappy and those locations aligned to gaps in the reference alignment also tend to be aligned with gaps by the MSA methods. Under d_{pos}, the distance increases only slightly for ClustalW (0.17) but Prank increases to 0.47, indicating that many of the gap locations differ between the Prank and reference alignments. A small further increase under d_{evol} is observed (ClustalW = 0.18, Prank = 0.50). Prank frequently aligns bases to gaps located differently from the reference alignment, but only occasionally corresponding to different events in time.

To explore further the similarities and differences between our metrics, we investigate the distances they produce on BAliBASE RV12. We plot the distances between all pairs of aligners across all data from RV12 as scatterplots. Figure 9 shows comparisons between d_{SSP} and our gap-aware metrics d_{seq} (Fig. 9A) and d_{pos} (Fig. 9B). The inclusion of non-specific gap information in d_{seq} means that it tends to produce lower distances than d_{SSP} as characters aligned with gaps in both alignments reduce the distance. This observation reinforces the result presented in Figure 6A, where d_{seq} tends to decrease after a certain degree of divergence. The comparison between d_{SSP} and d_{pos} shows no such clean relationship, with d_{SSP} sometimes being greater than or less than d_{pos}. This variation is because for d_{pos} the location of the gap is important, so if two methods align gappy regions similarly d_{pos} is lower than d_{SSP}, whereas for inconsistent gappy regions d_{pos} is greater than d_{SSP}.

The substantial differences between d_{pos} and d_{seq} are also illustrated in Figure 9C. In common with the d_{SSP} and d_{seq} comparison, these differences arise due to the treatment of gaps. The definitions enforce the relationship $d_{\text{pos}} \geq d_{\text{seq}}$, but sometimes d_{pos} is substantially higher than d_{seq} and we infer that such alignments are gappy with much variation in the base-gap homology assignments.

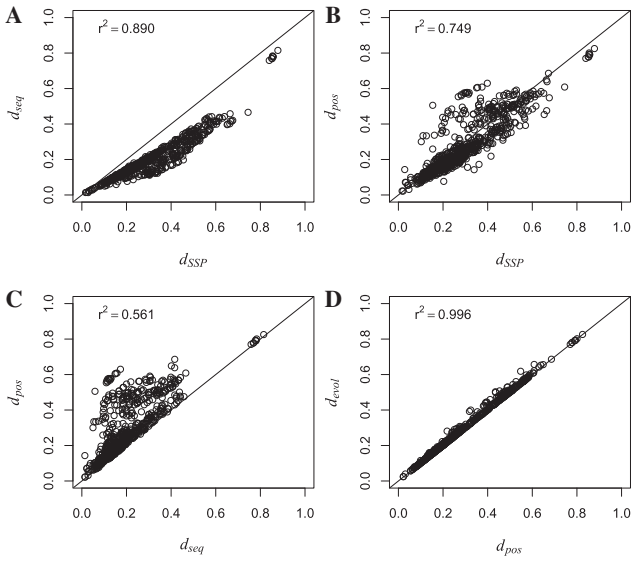

Fig. 9. Scatterplots of distances calculated according to different metrics on the RV12 BAliBASE dataset. Each circle corresponds to a single distance computed on one set of homologous sequences from RV12 between alignments performed by two of: ClustalW, Muscle, ProbCons, T–Coffee, Prank, MAFFT L–INS–i and MAFFT FFT–NS–i. A solid line is plotted to indicate $x = y$.

Figure 9D shows that $d_{\text{pos}} \approx d_{\text{evol}}$, which suggests that the different placement of gaps by different aligners rarely alters the inferred evolutionary history of indel events in the RV12 dataset.

4 DISCUSSION

In this study, we have defined four metrics to compare MSAs. Our metrics are superior to the most commonly used scores (i.e. the SP and TC scores) in two ways. First, our scores are valid metrics, whereas existing methods, such as the SP and TC scores are not. Secondly, our methods can incorporate indel information, including their location and when they occur during evolution. In contrast to existing scores, our metrics also allow the simple comparison of sets of sequence alignments, allowing us to investigate how similar methods of sequence alignment are to one another rather than how well individual alignments compare to a reference alignment. This is an important difference; we find that for real data, different alignment methods tend to produce alignments with similar magnitude in distance to the BAliBASE reference alignment, but that these inferred alignments are frequently at least as different from one another as they are from the reference alignment.

A multitude of studies have examined the effect of sequence misalignment on, for example, phylogenetic inference (Cantarel *et al.*, 2006; Hall, 2005; Morrison and Ellis, 1997; Ogden and Rosenberg, 2006; Wang *et al.*, 2011; Wong *et al.*, 2008), detection of positive selection (Fletcher and Yang, 2010; Markova-Raina and Petrov, 2011; Schneider *et al.*, 2009; Wong *et al.*, 2008), detection of co-varying sites in proteins (Dickson *et al.*, 2010), studies of non-coding DNA (Pollard *et al.*, 2006) and protein structure prediction (Miklos *et al.*, 2008). The metrics introduced here will permit several important analyses that may benefit such studies. First, our metrics allow the direct comparison of alternative alignments and their outcomes, which may be especially important as different algorithms may be subject to similar kinds of bias (Golubchik *et al.*, 2007). Second, comparing the differences between our metrics allows one to examine how variation in an alignment correlates with differences observed in downstream analyses. Finally, comparing sets of optimal or near-optimal alignments may allow one to discriminate between the effects brought about by uncertainty in an alignment and bias introduced by the alignment algorithms. By providing valid metrics for the direct comparison of alignments, we hope future studies will be able to more thoroughly understand the similarities and differences between alignment methods, and make progress in the task of disentangling alignment heuristics from their effect on downstream inference.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for comments that improved the manuscript.

Funding: Biotechnology and Biological Sciences Research Council (UK) (grant number BB/H000445/1).

Conflict of Interest: none declared.

REFERENCES

Arnold,K. *et al.* (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.

Cantarel,B. *et al.* (2006) Exploring the relationship between sequence similarity and accurate phylogenetic trees. *Mol. Biol. Evol.*, **23**, 2090–2100.

Deza,M.M. and Deza,E. (2009) *Encyclopedia of Distances*. Springer, Berlin, Heidelberg.

Dickson,R.J.*et al.* (2010) Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. *PLoS One*, **5**, e11082.

Do,C.B. *et al.* (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.

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

Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Felsenstein,J. (2003) *Inferring Phylogenies*. Sinauer Associates Inc., Sunderland, Massachusetts.

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

Fletcher,W. and 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.

Golubchik,T. *et al.* (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 2433–2442.

Hagopian,R. *et al.* (2010) SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction. *Nucleic Acids Res.*, **38**, W29–W34.

Hall,B.G. (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol. Biol. Evol.*, **22**, 792–802.

Huson,D.H. and Steel,M. (2004) Phylogenetic trees based on gene content. *Bioinformatics*, **20**, 2044–2049.

Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics*, **9**, 286–298.

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

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

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

Leach,A.R. and Gillet,V.J. (2003) *An Introduction to Chemoinformatics*. Dordrecht Kluwer Academic Publishers, Dordrecht, Netherlands.

Lipkus,A. (1999) A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.*, **26**, 263–265.

Liu,K. *et al.* (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**, 1561–1564.

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

Markova-Raina,P. and Petrov,D. (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. *Genome Res.*, **21**, 863–874.

Miklos,I. *et al.* (2008) How reliably can we predict the reliability of protein structure predictions? *BMC Bioinformatics*, **9**, 137.

Morrison,D.A. and Ellis,J.T. (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol. Biol. Evol.*, **14**, 428–441.

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

Notredame,C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.*, **3**, e123.

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

Penn,O. *et al.* (2010) An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.*, **27**, 1759–1767.

Pollard,D.A. *et al.* (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, **7**, 376.

Redelings,B.D. and Suchard,M.A. (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, **54**, 401–418.

Schneider,A. *et al.* (2009) Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.*, **1**, 114–118.

Simmons,M.P. and Ochoterena,H. (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.*, **49**, 369–381.

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

Subramanian,A.R. *et al.* (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.

Thompson,J.D. *et al.* (2005) Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Wang,L.-S. *et al.* (2011) The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 1108–1119.

Whelan,S. and 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.

Wise,M. (2010) No so hot - heads or tails is not able to reliably compare multiple sequence alignments. *Cladistics*, **26**, 438–443.

Wong,K.M. *et al.* (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.