11

Assessing Multiple Sequence Alignments Using Visual Tools

Catherine L. Anderson[1], Cory L. Strope[2] and Etsuko N. Moriyama[2,3]
[1]*Department of Computer Science and Engineering*
[2]*School of Biological Sciences and*
[3]*Center for Plant Science Innovation*
University of Nebraska-Lincoln,
U.S.A.

1. Introduction

Bioinformatics and molecular evolutionary analyses most often start with comparing DNA or amino acid sequences by aligning them. Pairwise alignment, for example, is used to measure the similarities between a query sequence and each of those in a database in BLAST similarity search, the most used bioinformatics tool (Altschul *et al.*, 1990; Camacho *et al.*, 2009). Evolutionary history among sequences can be reflected better when more than two sequences are aligned, in a multiple sequence alignment (MSA). When building an MSA, we assume that the sequences compared are derived from a common ancestral sequence. Then the process of MSA building is to infer homologous positions between the input sequences and place gaps in the sequences in order to align these homologous positions. These gaps represent evolutionary events of their own. Gaps (also called indels) are caused by either insertions or deletions of characters (nucleotides or amino acids) on a particular lineage of sequences during the evolution. Building an MSA is, therefore, to reconstruct the evolutionary history of the sequences involved. While it is easy to understand that the quality of MSAs affects the quality of phylogenetic tree reconstruction, the effect of MSA quality reaches far beyond this. Some examples of bioinformatics methods that utilize information extracted from MSAs include: profile building in similarity search (*e.g.*, PSI-BLAST: Altschul *et al.*, 1997), motif/profile recognition (*e.g.*, PROSITE: Hulo *et al.*, 2008), profile hidden Markov models for protein families/domains (*e.g.*, Pfam: Finn *et al.*, 2010), and protein secondary-structure prediction (for review, see Pirovano & Heringa, 2010). There are numerous bioinformatics and molecular evolutionary analyses that are affected by MSA quality and they can be benefited by having reliable MSAs.

Despite the significance of having good MSAs, assessing MSA quality is far from straightforward. Measuring the quality of MSAs requires two components: a benchmark dataset and a scoring method. A benchmark dataset includes *reference alignments*. These alignments are considered to represent the evolutionary history of the sequences truthfully. The same set of sequences included in a reference alignment is then aligned using the MSA methods to be tested. The *reconstructed* MSA can be compared with the reference MSA using a scoring method and the quality of the reconstructed MSA is assessed compared to the

reference MSA. Problems exist both in benchmark MSA datasets as well as in the methods used to measure the MSA quality.

The majority of benchmark MSA datasets are built on real sequences by aligning structural elements and in some cases with hand-curation (*e.g.*, PREFAB: Edgar, 2004b; OXBench: Raghava *et al.*, 2003; HOMSTRAD: Stebbings & Mizuguchi, 2004; BAliBASE: Thompson *et al.*, 2005; Thompson *et al.*, 2011; SABmark: Van Walle *et al.*, 2005). Since the true evolutionary history of the sequences included in these datasets is unknown, positional homologies among sequences are unknown and the accuracy of these reference MSAs is subjective (some issues on benchmark datasets, see Edgar, 2010). Some other benchmark datasets are generated by simulating sequence evolution based on specific molecular evolutionary models (*e.g.*, IRMBASE: Subramanian *et al.*, 2005). The advantage of these simulated datasets is that the evolutionary history of sequences (the guide tree) is known and the *true* alignment is given as an outcome of the simulation. Since the evolutionary history is known, these datasets can be used to assess the quality of both MSAs as well as phylogenetic reconstruction methods. The disadvantage is that the biological correctness of the simulation relies solely on the evolutionary models used.

Issues also exist in the methods used to measure the quality of MSAs. While a number of statistics has been proposed (*e.g.*, Position Shift Error score: Cline *et al.*, 2002; sum-of-pairs score and column score: Thompson *et al.*, 1999), there is no definite answer how to measure 'biological correctness' of MSAs. It remains for the end user to incorporate the statistics into their evaluation of this 'biological correctness'.

Due to its significant impact on many bioinformatics and molecular evolutionary studies, MSA is one of the most scrutinized bioinformatics fields (Kemena & Notredame, 2009; Thompson *et al.*, 2011). However, assessment of MSAs is usually reserved for power users. Often regular users simply run one MSA method and proceed to the next analysis without examining their alignment output (Morrison, 2009b). Considering how MSA quality affects the outcomes of further analysis, assessment of MSAs, however, should be included as regular part of sequence analysis. In order to facilitate comparative analysis of MSAs, we recently developed a software package called SuiteMSA (Anderson *et al.*, 2011). SuiteMSA provides several alignment-viewing tools that allow the user to compare MSAs both visually and quantitatively. SuiteMSA also includes a feature-rich biological sequence simulator, indel-Seq-Gen v2.1 (Strope *et al.*, 2009), with a user-friendly graphical interface, allowing the users to generate their own benchmark alignments for testing various MSAs.

In this chapter, we first review some of the statistics used to assess the quality of MSAs focusing on those used in SuiteMSA. We then describe how MSA comparison can be actually performed using various MSA viewers available in SuiteMSA. Five examples are chosen from diverse types of alignment problems: proteins with secondary structures, transmembrane proteins, proteins with length variation, simulated protein sequences, and ribosomal DNAs. These comparisons illustrate how various MSA methods perform differently based on their underlying assumptions. We also discuss how different alignment statistics should be used for assessing MSAs and their limitations.[1]

[1] All input files and alignments shown in this chapter are available from the following website: http://bioinfolab.unl.edu/~canderson/SuiteMSA/supplement.html

2. Statistics used to assess multiple sequence alignments

There are two types of alignment statistics. The first type of statistics is used to characterize a single alignment for the level of conservation in each alignment position and for various gap measures. These are descriptive measures for a specific alignment and should not be interpreted as a measure of the alignment quality. The second type of statistics can be used to compare any two alignments containing the same sequences.

2.1 Descriptive statistics on a single multiple sequence alignment

We describe the following two descriptive statistics: information content and average hydrophobicity. Both are calculated on a per column basis.

2.1.1 Information content

The Shannon entropy is a measure of the amount of uncertainty (Shannon, 1948). When it is applied to MSA analysis, it is interpreted as a measure of the diversity of characters within a given alignment column (Schneider & Stephens, 1990). The amount of information conveyed, or information content, is given by the decrease in this uncertainty and represents the level of sequence conservation within a column.

Formally defined, the entropy for the k^{th} column of an alignment is given as:

$$H(k) = -\sum_{s \in k} f(s,k) \log_2 f(s,k), \tag{1}$$

where s is any character contained in column k and $f(s,k)$ is the frequency of s as it appears in column k. If there are x_s of the character s in the column that has x of non-gap characters, $f(s,k)$ is calculated as x_s/x. The information content in the k^{th} column is given as:

$$I(k) = \log_2 S - H(k), \tag{2}$$

where S is the number of character types for an alignment (4 for a nucleotide alignment and 20 for an amino acid alignment). Both $H(k)$ and $I(k)$ have their units in bits.

It can be seen from these equations that the higher the number of distinct characters within a column, the higher the entropy value (H) and thus, the lower the information content (I) in the column. For a completely conserved column c, one which contains only one type of characters, the entropy $H(c)$ is 0; thus it contains the maximum amount of information. For a nucleotide alignment this maximum value is 2, while for an amino acid alignment it is 4.32.

Note that gaps are not considered in calculating $f(s,k)$ in equation (1). Excluding gaps from calculation could inflate the information content for a column that contains many gaps. A single character in a column of gaps, for example, can be erroneously attributed a maximum information content. In order to compensate for this situation, the column information calculation is normalized by multiplying each column's information content by the proportion of non-gap characters present in the column (Schneider & Stephens, 1990).

While the information content is a measure applicable to a single alignment, it can be useful to compare the information statistics among alternate alignments for trends.

2.1.2 Average hydrophobicity

Hydrophobicity is one of the most useful properties of amino acid residues, which is directly related to the function and structure of proteins. Many different types of

hydrophobicity indices are available (Kawashima *et al.*, 2008). By plotting hydrophobicity values along the sequence, the presence of functional/structural regions (*e.g.*, membrane-spanning regions in transmembrane proteins or core regions in globular proteins) can be predicted. For MSA analysis, comparing the distribution of hydrophobicity along the alignment among different MSAs can provide a visual aid for evaluating the consistency between alignments. Equation (3) below shows how the average hydrophobicity for column k, $h(k)$, is calculated for an alignment containing N sequences:

$$h(k) = \frac{\left(\sum_{i=1}^{N} h_i\right)}{N}, \tag{3}$$

where h_i is the hydrophobicity index value of i^{th} residue of column k. In SuiteMSA, the hydrophobicity index provided by Kyte and Doolittle (1982) is used and the value of 0 is assigned for a gap.

2.2 Measuring the similarity between two multiple sequence alignments

As mentioned earlier, many statistics have been proposed to compare two MSAs. The sum-of-pairs score (SPS) and the column score (CS) are the two used most often. Both scores were proposed by Thompson *et al.* (1999). The values of these two scores react differently to varying inconsistency between MSAs compared.

When comparing two alignments, one is referred to as the *reference alignment* and the other the *test alignment*. The test alignment is compared against the reference. If the reference alignment is known to be 'correct', these statistics can be used to measure the alignment quality. As mentioned before, however, the 'correctness' of an alignment can be highly subjective in the case of many available benchmark datasets. An alignment can be said to be truly 'correct' only if its exact evolutionary history is known and if the alignment reflects it correctly. Usually it is possible only if the alignment was generated by a sequence evolution simulator. Even if the 'true' alignment can be obtained by sequence simulation, however, 'biological realism' of the evolutionary model used with the simulation becomes an issue. In this chapter, SPS and CS are thus used more as general comparison measures.

2.2.1 Sum-of-pairs score (SPS)

To calculate the SPS for a test MSA against the reference MSA, each pair of characters within an alignment column is treated as an alignment unit. The per-column SPS is the number of alignment units within a specific column of the test alignment that are also aligned in the same column of the reference alignment. The total of all per-column scores from the entire alignment is obtained and normalized by dividing by the total number of character pairs. This is formally defined as follows:

i. Let an alignment of length M containing N sequences be an N by M array, **A**. Then the character in the i^{th} sequence and k^{th} column of the alignment is identified as A_{ik}.
ii. Let there be two alignments for comparison: alignment $\mathbf{A_r}$ (referred to as the reference alignment) of length M_r containing N sequences and alignment **A** (referred to as the test alignment) of length M containing N sequence, where M_r and M can be but are not required to be equal.
iii. To examine the k^{th} column of **A**, consisting of elements A_{1k}, A_{2k}, … A_{nk}, let p_{ijk} be defined as:

$$\begin{cases} p_{ijk} = 1 \text{ if } A_{ik} \text{ and } A_{jk} \text{ of alignment } \mathbf{A} \text{ are in the same column of } \mathbf{A_r}, \\ p_{ijk} = 0 \text{ otherwise.} \end{cases} \quad (4)$$

iv. Then the score for k^{th} column of **A** is defined as:

$$S_k = \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_{ijk} . \quad (5)$$

v. The score for the full alignment **A** is given as:

$$SPS = \left(\sum_{k=1}^{M} S_k \right) \Big/ \left(\sum_{k=1}^{M_r} S_{rk} \right) , \quad (6)$$

where S_{rk} is the score for the reference alignment, $\mathbf{A_r}$. This reference score is calculated as $S_{rk} = x(x\text{-}1)/2$ where x is the number of characters in column k excluding gaps.

The maximum possible SPS is a value of 1.0 when $\mathbf{A} = \mathbf{A_r}$. The SPS is not symmetric in that the score will be different if the reference and test alignments are switched.

2.2.2 Column score (CS)

To calculate the CS, the test and reference alignments are compared column-wise. The column score is the number of 'matched' columns between the test alignment and the reference alignment divided by the total number of 'considered' columns in the test alignment. This is formally defined as follows:

i. For the k^{th} column of **A**:

$$\begin{cases} C_k = 1 \text{ if all the characters in the column } k \text{ of alignment } \mathbf{A} \text{ are matched in alignment } \mathbf{A_r}, \\ C_k = 0 \text{ otherwise.} \end{cases} \quad (7)$$

ii. The column score for the full alignment **A** is given as:

$$CS = \left(\sum_{k=1}^{M} C_k \right) \Big/ M . \quad (8)$$

In SuiteMSA, two types of CS are calculated: un-gapped and gapped.

Un-gapped CS: This score considers only un-gapped columns (columns that have no gaps), where M of equation (8) equals the number of un-gapped columns in the alignment (shown in red in Fig. 1). For example, if an alignment has 500 columns and only 200 contain no gaps and of these 200, 150 columns are exactly as they appear in the reference alignment, then the un-gapped CS is given as 150/200 = 0.75. The disadvantage to these criteria is that very gappy alignments with very few un-gapped columns can still produce a high column score if those un-gapped columns are all 'matched'. For instance, a test alignment of any length, even if only one column is un-gapped and matches a column in the reference alignment, will yield a column score of 1.0.

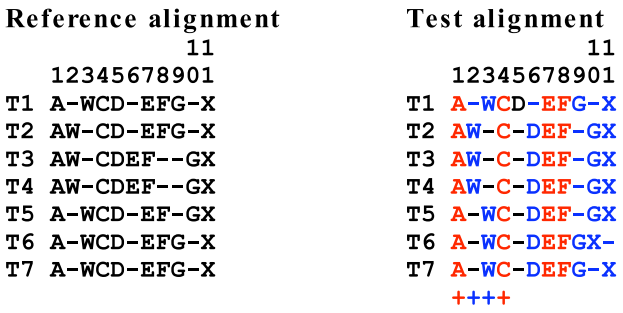

Fig. 1. Illustration of the column score calculation. In the Test alignment, 'un-gapped' columns are shown in red. 'Un-gapped matched' columns are indicated with red '+' under the alignment. For 'gapped' CS, all but 5th column of the Test alignment are considered and these columns are shown in blue as well as red. However, only those columns indicated with '+' (both red and blue) are counted as 'matched' against the Reference alignment. In this example, 'un-gapped' CS is 0.5 (2 out of 4 columns are matched) and 'gapped' CS is 0.4 (4 out of 10 columns are considered to be matched).

Gapped CS: This score considers columns that contain more than 20% non-gap characters. To be 'matched' the characters that appear in a column of the test alignment must appear in a column of the reference alignment with no additional characters. For example, in Fig. 1, all but 5th column of the Test alignment are considered. The columns 6-11 are not counted as 'matched'. This is because, for example, while in the Test alignment, 'G' of T1 position 9 is aligned only with 'G' of T6 and T7, in the Reference alignment, 'G' of T1 position 9 is aligned with 'G' of T2 as well as T6 and T7. The advantage to 'gapped' CS is that it allows more columns to be considered; columns with gaps can be matched if the same non-gap characters (but no other characters) are aligned in the reference alignment. This does offset the disadvantage of the potentially inflated un-gapped CS mentioned before.
Exclusion of any alignment columns that include gaps can be justified since gaps represent evolutionary events that are often not traceable. They are either the insertion of new characters, the deletion of existing characters, or a combination of the two. Therefore, while they are represented by the same gap symbol in the alignment, they are not equivalent. It is often not possible to infer if a gap in one alignment was generated by the same event as a gap in the second alignment. On the other hand, excluding all alignment positions with gaps even for those containing only a small number of gaps may not be desirable. In SuiteMSA, as described above, a column is considered as long as it contains a number of non-gap characters above the 20% threshold. A third column score is also provided in SuiteMSA as '% consistency', which considers all columns regardless of the number of gaps. Comparing these values can help assessing the difference between two alignments.

2.2.3 Implementation of SPS and CS

In addition to SuiteMSA, several implementations of SPS and CS are available as listed in Table 1. Note that not all of these programs generate the same value for the same alignment. The difference is caused by different criteria used to define, for example, 'matched' columns and which columns should be 'considered' for counting. When comparing scores, due to this inconsistency among programs, it is necessary to use the same implementation of scoring methods.

Program	Reference	Note
bali_score	(Thompson *et al.*, 1999)	standalone; C program; MSF format.
qscore	(Edgar, 2004b)	standalone; C++ program; calculates Q score (SPS), TC (CS), Modeler score, and Shift scores; fasta format.
VerAlign		available from http://www.ibi.vu.nl/programs/veralignwww MSF format.
SuiteMSA	(Anderson *et al.*, 2011)	part of the GUI software; fasta format.

Table 1. Programs available to calculate SPS and CS. The actual SPS and CS values for alignments discussed in this chapter given by different programs are available from our website (see footnote 1).

3. Visual inspection of MSAs

In the following sections, using various examples, we will show how MSAs can be compared using SuiteMSA's visual tools and statistics. See Anderson *et al.* (2011) and SuiteMSA User's Manual for detailed description of various tools available in SuiteMSA.
Among the numerous MSA methods currently available, we chose seven MSA methods listed in Table 2 for comparative analysis. We chose these methods based on their general popularity in various bioinformatics analyses, their availability, and some of their features useful for aligning particular types of proteins (*e.g.*, transmembrane proteins).

Method (version)	Reference	Description
ClustalW2 (2.1)	(Larkin *et al.*, 2007)	Progressive alignment; weights sequences based on branch lengths and adjusts gap penalties; one of the earliest methods implemented. http://www.clustal.org/
MUSCLE (3.8.31)	(Edgar, 2004a, 2004b)	Progressive alignment; fast distance estimation using kmer counting; iterative refinement using tree-dependent restricted partitioning. http://www.drive5.com/muscle/
MAFFT (6.843)	(Katoh & Toh, 2008)	Progressive alignment; L-INS-i method is used for iterative refinement incorporating local pairwise alignment information in this study. http://mafft.cbrc.jp/alignment/software/
Probalign (1.4)	(Roshan & Livesay, 2006)	Uses partition function posterior probability estimates to compute maximum expected accuracy alignments. [eProbalign] http://probalign.njit.edu/probalign/login
PRANK (web version)	(Löytynoja & Goldman, 2005, 2008)	Phylogeny-aware gap handling; not meant for divergent sequences; recognizes insertions and deletions as distinct evolutionary events. [webPRANK] http://www.ebi.ac.uk/goldman-srv/webprank/

Method (version)	Reference	Description
PRALINE (web version)	(Pirovano *et al.*, 2008)	Progressive alignment with profile pre-processing; incorporates secondary structure and transmembrane information; PSIPRED and Phobious (for GPCR alignment) chosen for this study. http://www.ibi.vu.nl/programs/pralinewww/
PROMALS (web version)	(Pei & Grishin, 2007)	Progressive alignment enhanced with profiles and secondary structure information; a hidden Markov model using a combined scoring of amino acids and secondary structures. http://prodata.swmed.edu/promals/

Table 2. The seven MSA methods compared in this study. All methods are used with the default options unless noted otherwise.

3.1 Examining a protein MSA with secondary structure prediction

When protein sequences are aligned, it is useful to identify the location of their functional or structural landmarks to determine if such landmarks are aligned properly. Useful landmarks include secondary structures, transmembrane regions, and conserved domains or motifs. Color-coding MSAs based on properties of amino acids also helps determine if the distribution of different types of amino acids is consistent or varied among sequences.

3.1.1 Inspecting a single MSA

In Fig. 2, eight protein sequences of the lipocalin family (Pfam PF00061; Finn *et al.*, 2010) are aligned. The lipocalin family proteins are highly divergent at the sequence level yet highly conserved at the structure level (Flower *et al.*, 2000). The common structural feature among these proteins is a single eight-stranded antiparallel beta-barrel. The MSA shown in Fig. 2 was originally produced using PROMALS3D (Pei *et al.*, 2008) with manual adjustment (Strope *et al.*, 2009). Using SuiteMSA's secondary structure viewer, we aligned the lipocalin MSA with the secondary structures predicted from the eight sequences using PSIPRED (Jones, 1999). It can be seen in Fig. 2 that eight beta-strand regions (shown as brown-colored clusters of 'E' letters) are clearly well aligned with very few gaps.

Fig. 2 also shows the per-column information content displayed as a blue bar chart below the MSA. The information content reflects the level of conservation for each column. This display is especially useful when dealing with alignments containing a large number of sequences and/or long sequences. When comparing such large alignments, the information content display can be used to quickly scan along the alignment to search for, *e.g.*, high conservation areas (indicated as high information content regions). In Fig. 2, fully conserved columns (positions 51, 53, 148, 150, and 179 are readily identifiable by the full-height bars. In fact, these positions are part of the three conserved motifs shared among lipocalin proteins. These motifs (indicated as M1, M2, and M3 in Fig. 2) are described as "structurally conserved regions" (SCR1, 2, and 3, respectively) by Flower *et al*. (2000). SCR1 corresponds to PROSITE lipocalin motif (PS00213; Hulo *et al.*, 2008).

Several summary statistics are given at the top of MSA Viewer window (Fig. 2). The following statistics are available:

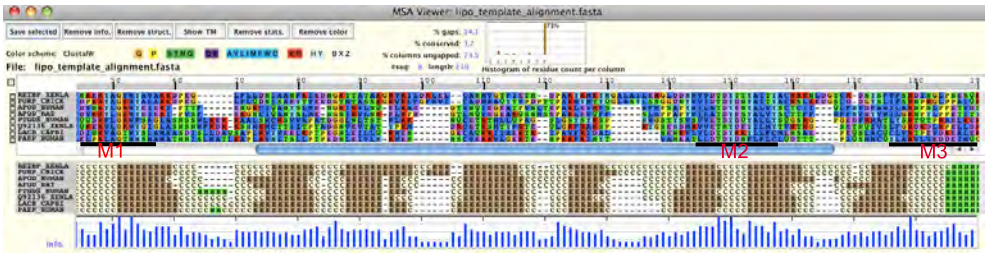

Fig. 2. The alignment of eight protein sequences from the lipocalin family. The MSA Viewer is used to display the MSA aligned with the predicted secondary structures. Black thick lines marked with M1, M2, and M3 indicate the locations of the three conserved motifs. Symbols used for the secondary structure prediction are: H (green) for helix, C (cream) for coil, and E (brown) for beta-strand. The alignment statistics are shown above the MSA. The column information content is displayed as a blue bar chart at the bottom indicating the level of conservation for each column.

- % gaps. The number of gap symbols within the alignment divided by the total number of characters within the alignment (alignment length times number of sequences). This should not be confused with the number of insertion/deletion events in the alignment since an individual event can span multiple positions.
- % conserved. The number of completely conserved columns divided by the total number of columns. A conserved column is defined as an un-gapped column containing a single type of characters.
- % columns un-gapped. The number of un-gapped columns divided by the total number of columns.
- The histogram of character count per column. This histogram represents the gappiness of the MSA using a non-gap character frequency distribution (the inverse of gap frequency distribution). For the lipocalin MSA, 73% of the columns have no gap (this is also shown as % columns un-gapped).

3.1.2 Comparing two MSAs

In Fig. 3A, we compared the previously shown lipocalin MSA (listed as 'Reference') with the MSA generated by ClustalW2 using the MSA Comparator. Under the blue selection bar and the green range bar, alignment positions are color-coded for the consistency with respect to the reference MSA. Blue characters illustrate where completely consistent columns are, and red characters depict those inconsistently aligned. Compared against the reference, ClustalW2 MSA is more compacted with very few gaps, making the alignment shorter (201 positions compared to 219 in the reference). We further examined the ClustalW2 MSA using the secondary structure display function of the MSA Viewer. As illustrated in Fig. 3B, the ClustalW2 MSA does not have the beta-strand regions (shown as brown-colored clusters of 'E' letters) aligned as well as the reference MSA does.

As mentioned earlier, the information content is the indicator of sequence divergence within a single MSA, and not a direct comparison between two alignments. However, as shown in Fig. 3A, the information content distributions (blue and green bar charts) can be compared between the alignments. It is especially useful when dealing with large alignments containing

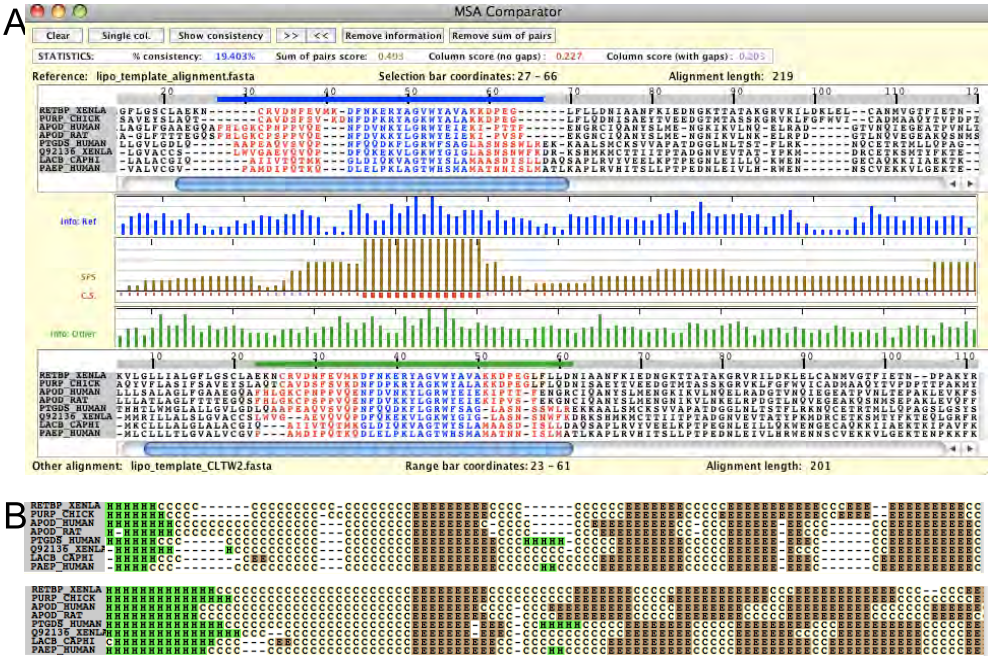

Fig. 3. Comparison of the ClustalW2 MSA with the reference alignment of the lipocalin family. A. The two MSAs are compared using the MSA Comparator (the reference and ClustalW2 alignments shown at the top and bottom, respectively). The column SPS display (brown bar chart) is positioned between the two MSAs and is aligned to the ClustalW2 alignment. At the bottom of the column SPS display is the column score (CS) indicator. The un-gapped CS uses those columns marked with purple squares, and the gapped CS uses columns marked with both purple and red squares (small and large squares indicate 'considered' and 'matched' columns, respectively). Summary statistics shown above the reference alignment include: % consistency, SPS, and two types of CS. B. The MSA Viewer is used to generate the secondary structure representation for the reference and ClustalW2 MSAs (shown at the top and bottom, respectively). Symbols used for the secondary structure prediction are: H (green) for helix, C (cream) for coil, and E (brown) for beta-strand.

many/long sequences. On the other hand, SPS is the result of a direct comparison between two MSAs. The per-column SPS (brown bar chart) displayed in Fig. 3A clearly shows where the test alignment (ClustalW2 in this case) is consistent (and to what degree) with the reference.

3.1.3 Comparing multiple MSAs

In Fig. 4, we compared MSAs produced by four methods against the reference lipocalin family MSA (MSA 1). Using the Pixel Plot, we can clearly see different patterns among the MSAs. The magenta-highlighted areas illustrate how the corresponding characters are aligned (or not) in each MSA. The PRALINE MSA (MSA 2) is fairly consistent compared to the reference MSA. This is expected since PRALINE uses secondary structure information when optimizing the alignments. On the other hand, MAFFT, MUSCLE, and ClustalW2 MSAs show a similar displacement of the same sequences, apparent from the ragged edges

Fig. 4. Comparison of the lipocalin family reference MSA (MSA 1) with four reconstructed MSAs (PRALINE, MAFFT, MUSCLE, and ClustalW2). The Pixel Plot is used to show the alignment patterns with each non-gap character represented with a solid colored pixel and a gap with a blank pixel. Characters corresponding to those under the blue selection bar for the reference MSA are highlighted in magenta in all MSAs. The green range bars for MSAs 2-5 show the column ranges where corresponding characters are located.

of the magenta areas. Alignments generated by MAFFT, MUSCLE, and ClustalW2 (MSAs 3-5) are roughly consistent to each other, but not consistent with the reference and PRALINE alignments. All four MSA methods tested produced shorter alignments (201-214 positions) compared to the reference alignment (219 positions). The shortest alignment was obtained from ClustalW2 (201 positions).

4. Aligning transmembrane protein sequences

In the previous section, we showed that comparing MSAs and secondary structure predictions help us assess the quality of MSAs. In this section, we will examine alignments of another type of proteins, transmembrane proteins.

4.1 G-protein coupled receptors

G protein-coupled receptor (GPCR) proteins contain seven transmembrane (TM) regions. They constitute a large protein superfamily grouped into three major and several minor classes (Horn *et al.*, 2003; Vroling *et al.*, 2011). Although the TM regions are relatively constant in length (22~24 amino acids or aa), the lengths of the N-/C-terminal and loop

regions are highly varied especially among different classes (Inoue *et al.*, 2004; Wistrand *et al.*, 2006). GPCR sequences are also highly divergent. These features make aligning GPCR sequences a challenge. We sampled 25 protein sequences from three major classes of GPCRs (Classes A, B, and C). The lengths of these GPCR sequences vary from 201 to 972 aa.

4.2 Alignment of GPCR sequences

Fig. 5 shows the alignment of the 25 GPCRs generated by PRALINE (showing only the first three TM area). Since PRALINE incorporates information from secondary structure, TM structure, as well as profiles based on PSI-BLAST similarity search (Table 2), it is expected to perform well in aligning TM regions. In order to confirm this, TM regions were predicted for each of the 25 GPCR sequences using MEMSAT3 (http://bioinf.cs.ucl.ac.uk/psipred/; Nugent & Jones, 2009). The predicted TM structural information was then aligned with the PRALINE MSA. Fig. 5 shows that the predicted TM regions (depicted with 'X' in green color) are clearly well aligned and visualized as green-colored clusters. The 'hydrophobicity' color scheme used for the MSA display as well as the average column hydrophobicity plot also confirm that more hydrophobic amino acids are found in predicted TM regions.

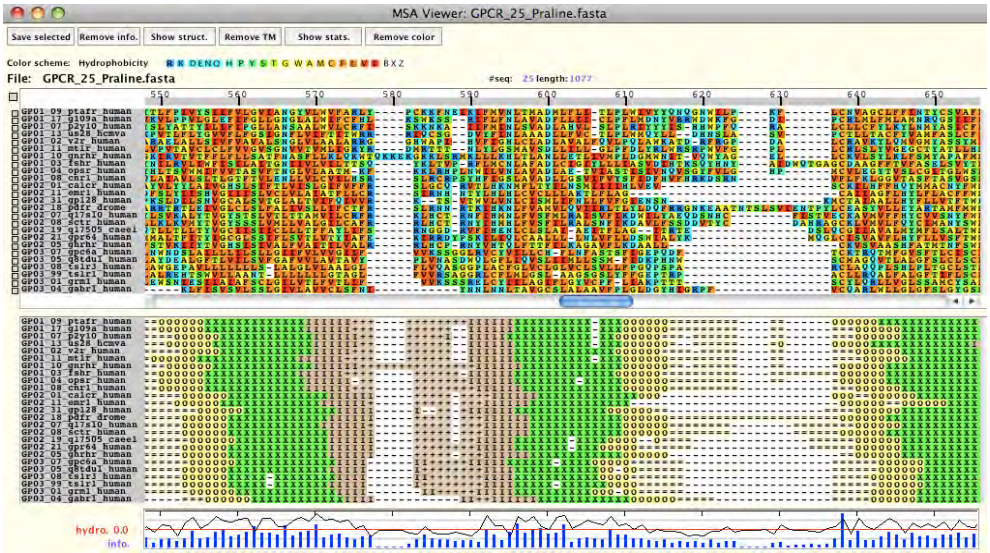

Fig. 5. Alignment of 25 GPCR proteins generated by PRALINE compared with TM structural predictions. Using the MSA Viewer, the PRALINE MSA is displayed using the 'hydrophobicity' color scheme showing hydrophobic amino acids more toward red and hydrophilic amino acids more toward blue. The predicted TM structure corresponding to each sequence of the MSA is aligned below. The symbols (based on MEMSAT3 prediction) used to show different TM structural components are as follows: 'X' (green) for the TM region, '+' (light brown) for the inside loop, 'I' (brown) for the inside helix cap, '=' (cream) for the outside loop, and 'O' (yellow) for the outside helix cap. The first three TM regions are depicted as three clusters of green letter X's. At the bottom of the display is the information content for each column (blue bar) and the average hydrophobicity for each column (black line plot). The average hydrophobicity (see equation (3)) is based on the index given by Kyte and Doolittle (1982).

4.3 Comparison of GPCR MSAs reconstructed by seven methods

We aligned the 25 GPCR protein sequences using seven methods. The seven MSAs produced were compared using Pixel Plot in Fig. 6. Compared to the terminal or loop regions, seven TM regions are expected to have fewer gaps. Using the Pixel Plot we can confirm such patterns. Approximate areas predicted to have TM regions can be located as clusters of solid colored pixels. In Fig. 7, the seven MSAs are represented in the predicted TM structures. The area includes the first five TM regions shown as the green-colored clusters. Both PRALINE and PROMALS utilize information from secondary structure prediction (also TM prediction for PRALINE) as well as profiles based on PSI-BLAST similarity search (Table 2). As expected, in the MSAs reconstructed by these two methods, predicted TM structures are aligned better than other methods. Other methods with the exception of PRANK also generated MSAs that aligned the area containing the first three TM regions relatively well. The rest of the sequences were more difficult for alignment. Probalign had a difficulty in reconstructing also the third TM region. With all MSA methods, all positions after the third TM region were not well reconstructed in terms of conservation of TM regions. The difficulty in aligning the second half of the protein sequences is likely caused by the large length variation found among GPCR classes, especially in the fourth and fifth loops (between TM4 and TM5, and TM5 and TM6, respectively) (Wistrand *et al.*, 2006).

In order to gain more insights on the difference among GPCR protein MSAs quantitatively, we gathered SPS values from all pairwise comparisons among the seven MSAs. Each of the seven MSAs was used as the reference and other six MSAs were tested against. Fig. 8 clearly shows that SPS is not symmetric. As expected, PRALINE and PROMALS, both of which utilize secondary structure and TM prediction information, had very high SPS' when they are compared to each other (0.546 and 0.543). Interestingly, using PRALINE or PROMALS as the reference, MAFFT was found to perform very well although MAFFT does not incorporate secondary structure nor profile information. It should be also noted that SPS' are among the highest when Probalign was compared to MAFFT (either as the reference or the test MSA).

The most drastic difference between the row and column averages of SPS' is found in PRANK. The SPS' obtained when the PRANK MSA was used as the reference (shown in the PRANK column) are all higher than those obtained when the PRANK MSA was tested against others (shown in the PRANK row). This can be explained by the gappy nature of the PRANK MSA (see Figs. 6 and 7). The PRANK MSA tends to have more gaps because of the underlying design of the method. It attempts to identify distinct insertions and deletions and tries not to collapse such independent events into the same column. For the same set of sequences, the reference alignment that has more gaps has a fewer number of character-pairs available (denominator in equation (6)) when averaging the total SPS, which tends to generate a higher SPS. Note also that the phylogeny-aware algorithm used with PRANK cannot perform well when sequences are too diverged (Löytynoja & Goldman, 2008). With extremely diverged GPCR sequences, PRANK was not expected to perform very well, which was indicated by constantly low SPS' obtained with PRANK. Although in the absence of 'true' reference alignment, low SPS values do not necessarily indicate incorrect alignment but rather inconsistency between the alignments, virtually no TM region was conserved in the PRANK MSA (Figs. 6 and 7). We will examine more on PRANK in the next section.

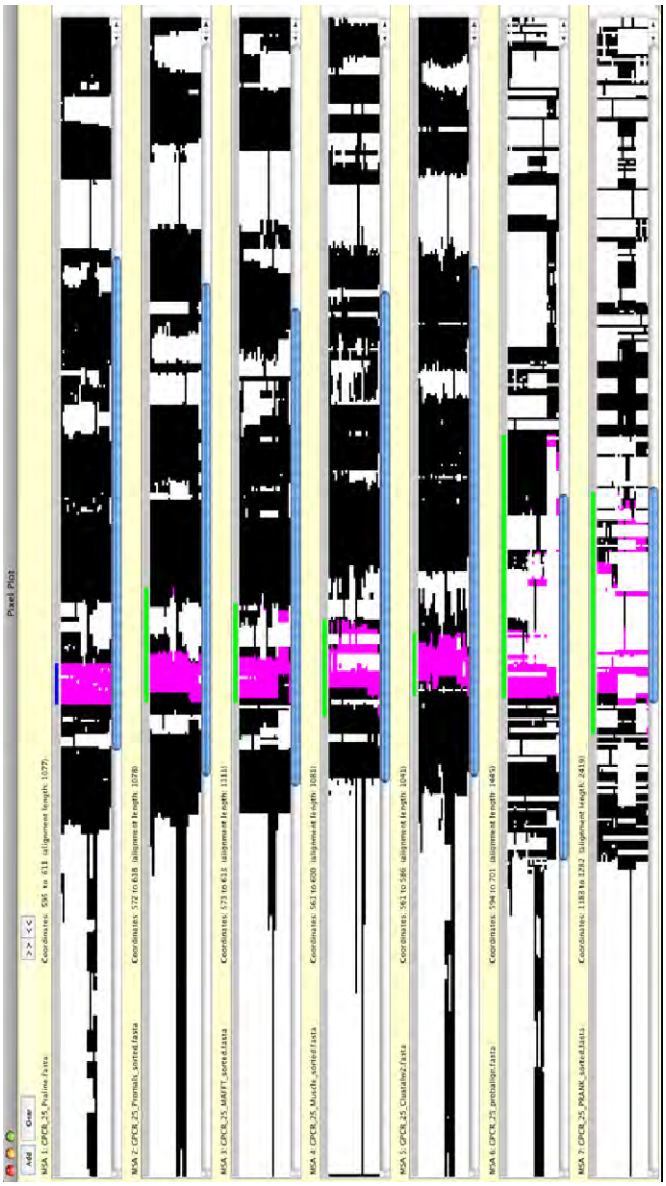

Fig. 6. Comparison of seven GPCR protein MSAs. The characters corresponding to the second TM (TM2) regions are highlighted with magenta-colored pixels. It shows that the TM2 region is reconstructed well in the MSAs by PRALINE (MSA 1) and PROMALS (MSA 2), relatively well by MAFFT (MSA 3), MUSCLE (MSA 4) and ClustalW2 (MSA 5), and not very well by Probalign (MSA 6) and PRANK (MSA 7). For the PRALINE MSA, the positions for the seven TM regions are as follows: 555-572, 595-611, 641-663, 683-703, 738-756, 813-830, and 854-875, which roughly correspond to solid-colored regions.

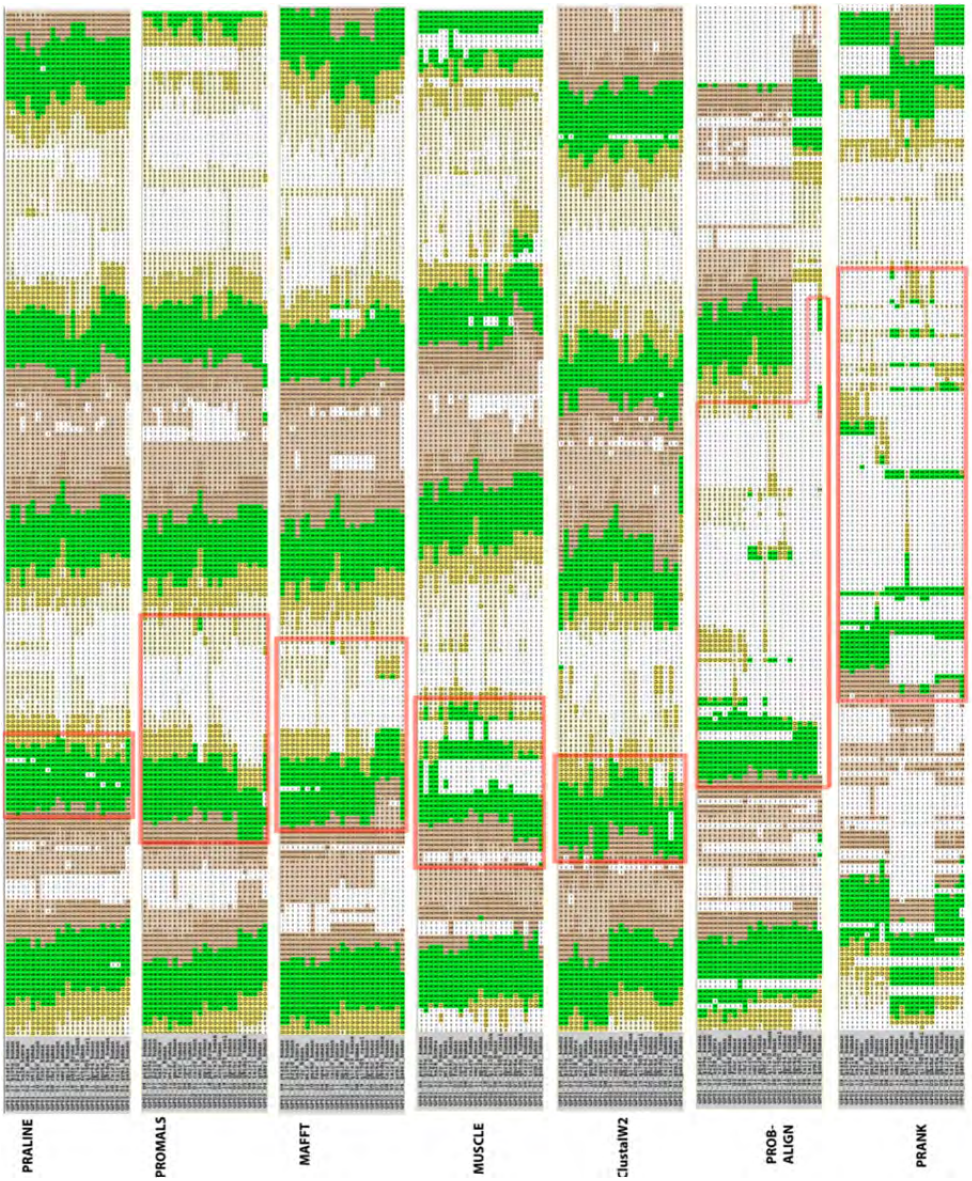

Fig. 7. Seven GPCR protein MSAs represented in TM structures predicted by MEMSAT3. The areas covering the first five predicted TM regions are shown. The red boxes indicate the areas containing amino acids predicted for the second TM regions. Wider red boxes in reconstructed MSAs indicate TM regions with a higher number of gaps (*e.g.*, Probalign and PRANK). The symbols representing amino acids predicted for different TM structural components are as follows: 'X' (green) for the TM region, '+' (light brown) for the inside loop, 'I' (brown) for the inside helix cap, '=' (cream) for the outside loop, and 'O' (yellow) for the outside helix cap.

Test \ Reference	PRALINE	PROMALS	Probalign	MAFFT	MUSCLE	ClustalW2	PRANK	[Average]
PRALINE	(1,077)	0.543	0.563	**0.592**	0.474	**0.381**	0.351	**0.484**
PROMALS	0.546	(1,078)	0.538	0.568	**0.500**	0.345	0.344	0.474
Probalign	0.477	0.455	(1,111)	0.520	0.457	0.324	**0.362**	0.433
MAFFT	**0.569**	**0.546**	**0.590**	(1,081)	0.498	0.346	0.357	**0.484**
MUSCLE	0.462	0.487	0.526	0.505	(1,445)	0.325	0.322	0.438
ClustalW2	0.383	0.346	0.384	0.362	0.335	(1,041)	0.283	0.349
PRANK	0.231	0.227	0.227	0.245	0.218	0.186	(2,419)	0.232
[Average]	0.445	0.434	0.471	**0.465**	0.414	0.318	0.337	

Fig. 8. Pairwise comparison of the sum-of-pairs scores (SPS) between GPCR protein MSAs reconstructed by the seven methods. The numbers in parentheses are the alignment lengths (the number of columns in each alignment). The highest score in each comparison is shown in boldface.

5. A different perspective on gaps

In this section we highlight the alignment method PRANK, which is unique in emphasizing a different perspective on the evolutionary process producing insertions and deletions. As shown in the previous section, it tends to produce more gaps in alignments compared to other methods. We compare the alignment generated by PRANK with four other methods.

5.1 Viral envelope glycoprotein, gp120

Löytynoja and Goldman (2008) used the viral exterior envelope glycoprotein, gp120, from human and simian immunodeficiency viruses (HIVs and SIVs, respectively) as an example to demonstrate how PRANK works. In this section, we used the same set of sequences they used (the seed alignment of Pfam Family GP120, PF00516, excluding SIVGB, SIVV1, and SIVG1). The entry of HIVs and SIVs into the host requires the interaction of the viral gp120 with the cell-surface proteins of the host. In order to avoid the host's immune system, several regions of the gp120 proteins evolve fast. Fig. 9 shows the MSA of gp120 proteins compared with the predicted secondary structures.

5.2 Gap treatment

The 'gap' within an alignment is a general expression for two very different types of evolutionary events. It represents either an insertion of one or more characters or the deletion of one or more. Both types of events are unobservable, and as such it is difficult to distinguish which event creates a gap in an alignment. For example, the 'gappy' section of an alignment, such as the V1 section of HIV/SIV gp120 (Fig. 9), can be interpreted either as the result of a high substitution rate along with frequent independent deletions or as the result of frequent independent short insertions and deletions. Optimization functions used in most MSA methods over-infer the former scenario, stacking independent insertions in the same column and potentially erroneously inflating substitution rates in such regions. Using phylogenetic information, PRANK, on the other hand, allows for the inference of both deletions and insertions as separate events.

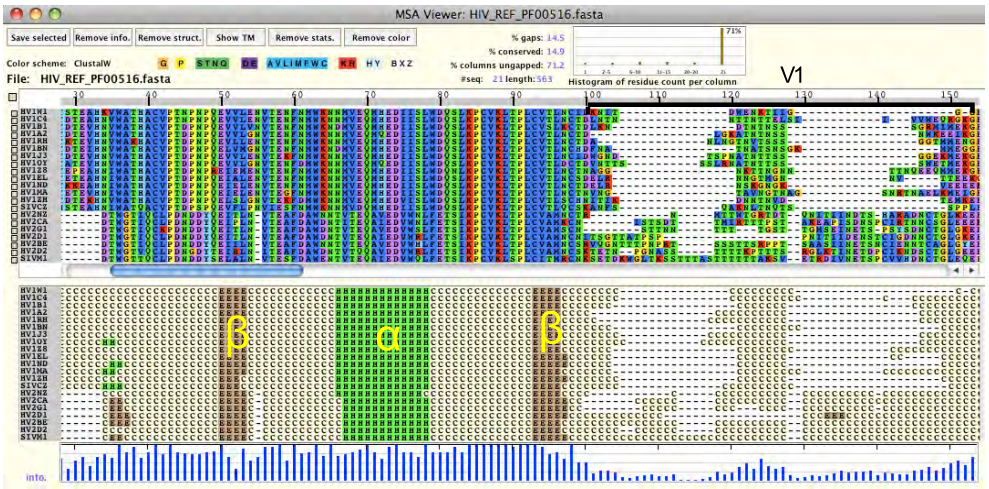

Fig. 9. Reference MSA of HIV/SIV gp120 proteins. The reference alignment was based on the seed alignment of Pfam Family GP120 (PF00516). The secondary structure is predicted for each sequence by PSIPRED (Jones, 1999) and is displayed using the MSA Viewer. The symbols depicting different secondary structures are as follows: H (green): alpha-helix, E (brown): beta-strand, and C (cream): coil. The MSA region shown above contains the N-terminal conserved area and the V1 variable region (positions 100-150). The predicted regions of the beta-strands (positions 50-53 and 93-96) and the alpha-helix (positions 66-78) correspond to the known HIV gp120 protein structure.

Progressive alignment methods such as ClustalW2, MAFFT, and MUSCLE build an alignment based on aligning the profiles of previously aligned sequences. The presence of a gap in the profiles is not checked to determine if the addition of another gap is parsimonious with the guide tree. The decision of whether adding a gap or not is based on the optimization function score. Since the inference of additional gaps penalizes the optimization function score, it often results in incorrectly matching potentially independent insertions, creating incorrect homologies.

PRANK attempts to avoid the above-mentioned pit-falls in progressive alignments by utilizing "phylogeny-aware" handling of gaps and treating insertions and deletions differently. The overall effect of the PRANK method compared to other progressive alignment methods is that the alignment is extended due to the separation of the independent insertions. As Löytynoja and Goldman (2008) stated, "the resulting alignments may be fragmented by many gaps and may not be as visually beautiful as the traditional alignments, but if they represent correct homology, we have to get used to them."

5.3 Comparison of the PRANK MSA with others

We aligned the set of 21 gp120 sequences using PRANK and other five alignment methods. For PRANK, we reconstructed the phylogeny using PhyML 3.0 (Guindon et al., 2010) and used it as the input phylogeny (with rooting between HIV1 and HIV2 clusters, the topology was identical with the one given in Löytynoja & Goldman, 2008). As shown in Fig. 10, the

Fig. 10. Comparison of gp120 MSAs. The Pixel Plot is used to compare five reconstructed MSAs (MSA 2: PRANK, MSA 3: PROMALS, MSA 4: Probalign, MSA 5: MAFFT, and MSA 6: MUSCLE) with the reference alignment (MSA 1, based on the seed alignment of Pfam Family GP120, PF00516). The area highlighted in magenta color is part of the V1 variable region, where the patterns show that the PRANK MSA is highly inconsistent with other MSAs.

major differences among MSAs are found starting at the first highly variable area, V1. Within this area, PRANK infers far more insertions than the other methods. The number of sites covered by the blue selection bar in the reference alignment (MSA 1) is 53. The corresponding sites in the other alignments are spread over from 51 columns with MAFFT (MSA 5) to 84 columns in PRANK (MSA 2).

Table 3 summarizes alignment statistics. As expected, PRANK generated the longest alignment. This is indicated in the PRANK MSA having a higher % gaps, lower % consistency, and lower % no-gap columns. Note also that the reference alignment used was the Pfam seed alignment, which in principle was generated using an alignment strategy similar to methods other than PRANK. These comparisons clearly illustrate the point made by Löytynoja and Goldman (2008). Depending on the MSA method used, a very different evolutionary mechanism would be emphasized to explain fast evolving gp120 sequences: either accelerated substitution rates or extremely high rate of short insertions or deletions. Another important point is that scores devised for MSA comparison (*e.g.*, SPS) should be used with the knowledge of the assumption underlying the design of the method used as well as the nature of the reference alignment.

Method	SPS	CS with no-gap	CS with gaps	% consistency	MSA length (aa)	% conserved	% gaps	% no-gap columns
Reference	-	-	-	-	563	14.90	14.50	71.20
PRANK	0.872	0.845	0.702	53.08	633	13.60	**23.90**	61.30
PROMALS	0.919	0.855	**0.775**	**68.24**	548	15.30	12.10	76.60
Probalign	**0.920**	0.838	0.761	62.50	579	15.00	16.80	72.70
MAFFT	0.907	0.827	0.727	63.90	557	15.40	13.60	74.70
MUSCLE	0.910	**0.926**	0.750	66.20	548	**15.50**	12.10	**77.60**

Table 3. Alignment statistics for gp120 MSAs. SPS, CS, and % consistency are obtained against the reference alignment. The highest value in each comparison is shown in boldface.

Fig. 11 illustrates a comparison between the MSAs generated by MAFFT (top) and PRANK (bottom). In Fig. 11A, the region under the blue selection bar for the MAFFT MSA (positions 104 to 128; 25 aa long) is more compact than the region covered by the corresponding amino acids in the PRANK MSA as indicated by the long green range bar (ranging from positions 106 to 159; 54 aa long). In this region, PRANK shows, for example, two independent insertions marked with light blue boxes ('GL' and 'MIR') both happening in SIV/HIV2 sequences. In the MAFFT MSA, these two sequences are part of a much longer insertion region unique to SIV/HIV2, implying that frequent deletion events shortened this region in various HIV2. Another insertion found in HIV1 by PRANK, 'SSSLR' (in a light green box), is shown to be almost independent. However, in the MAFFT MSA, the corresponding region appears to have experienced many deletion events instead. This shows the "gap magnet" phenomenon found in many progressive-alignment methods. Fig. 11B from the same MSA area highlights another possible artifact often found in MSAs generated by progressive alignment methods. In the red area in the MAFFT MSA, all sequences are aligned (matched) generating the "collapsed insertions", implying homologous relationships among these sequences. However, in the PRANK MSA, the corresponding sequences are spread out in a wide range of columns. These examples show that the inferred evolutionary scenarios can be completely different depending on the alignment methods used to analyze sequences.

6. Using simulated sequences for testing MSA methods

In this section we will discuss the use of simulation data in the comparison of alignment methods. The advantage of using simulated sequences is the availability of the 'true' alignment. In the simulation example discussed in this section, we simulate two sets of eight lipocalin sequences described in Section 3. The lipocalin protein family has a common structural feature, a single eight-stranded antiparallel beta-barrel. They also share three conserved motifs. We will use the simulation program indel-Seq-Gen version 2.1 (iSGv2.1; Strope *et al.*, 2009) to simulate this lipocalin family proteins. iSGv2.1 is included in the SuiteMSA package and the simulation can be done using its graphical user interface.

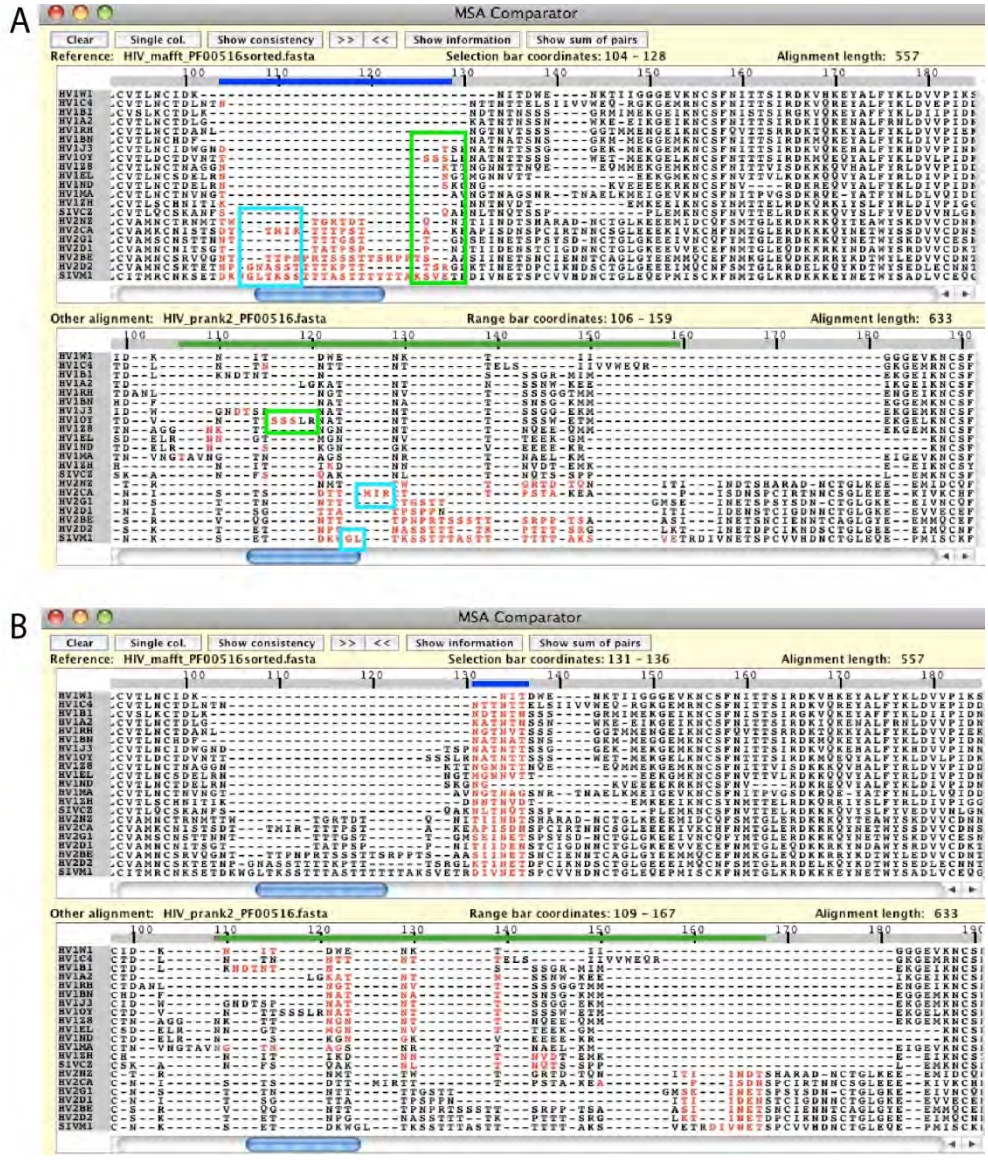

Fig. 11. Comparison of gp120 alignment regions generated by MAFFT and PRANK. For both panels A and B, the MAFFT MSA is used as the reference (top) and the PRANK MSA (bottom) is compared against. The blue selection bar for the MAFFT MSA shows the alignment area selected, and the green range bar for the PRANK MSA shows the column range where corresponding amino acids are found in the MSA. The corresponding amino acids in the two MSAs are shown with red color ('red' indicates that the alignment of these characters is inconsistent between the MSAs). See the main text for the description on the sequences marked with light blue and light green boxes.

6.1 Setting up the iSGv2.1 simulation

iSGv2.1 simulation requires a guide tree and a root sequence or MSA. By providing a root MSA, instead of generating a random root sequence, the site-specific amino acid (or nucleotide) frequency distribution derived from each MSA column can be used to generate a simulation root sequence (for details, refer to iSG user manual). For the root MSA, we used the 8-protein alignment of the lipocalin family we described in Section 3. The evolutionary parameters chosen to simulate the lipocalin protein family are listed below. We performed two simulations: the second more divergent than the first. For any parameters not mentioned, default values were used. Three input files are prepared: a guide tree file, a lineage file, and a root MSA file. For details on preparing the guide tree, the three motifs used, and how to set up the length-limitation template, refer to Strope *et al.* (2009) as well as Anderson *et al.* (2011). All input files used for this simulation are available from: http://bioinfolab.unl.edu/~canderson/SuiteMSA/supplement.html.

i. Basic parameters
   - Guide tree file: lipo8_3.tre (provides the guide tree and option parameters listed below)
   - Substitution model: PAM

ii. Advanced parameters
   - Lineage file: lipo8_3.spec (provides the motif and template information)
   - Branch scale: 0.5 (first simulation), 2.0 (second simulation)
   - Random number seed: 6262

iii. Guide tree options (information included in the guide tree file)
   - Use root msa file: lipo8_3template.root_in
   - Maximum indel length: 10
   - Insertion probability = deletion probability = 0.02 (first simulation), 0.025 (second simulation)
   - Indel length distribution = deletion length distribution: file name = inDL (provides indel length distribution)

After running each simulation, we obtained a set of eight simulated sequences, the true alignment of the eight sequences, and a record of all insertion and deletion events. As shown in Fig. 12, the 'true' alignments from both simulations (the first more conserved and the second more diverged) maintained the three conserved lipocalin motifs (M1, M2, and M3) specified in the simulations. As expected, the second MSA derived from the simulated sequences with a higher rate of substitutions (longer branch lengths) and a higher rate of indel probability is about 100 aa longer (Fig. 12B, 303 aa) than the first MSA (Fig. 12A, 215 aa). We used these 'true' alignments as the references for the next analysis.

6.2 Comparison of MSA reconstruction using simulated sequences

We used four MSA methods (MAFFT, MUSCLE, Probalign, and PRANK) to align both sets of the eight simulated sequences. For PRANK, the simulation guide trees (with branch lengths scaled for the 'more conserved' and 'more divergent' simulations) were used as the input phylogenies. In Fig. 13, the Pixel Plot is used to compare the reconstructed MSAs against the reference MSAs (the true alignments obtained from the two simulations). Tables 4 and 5 summarize the alignment statistics for the two sets of simulated data.

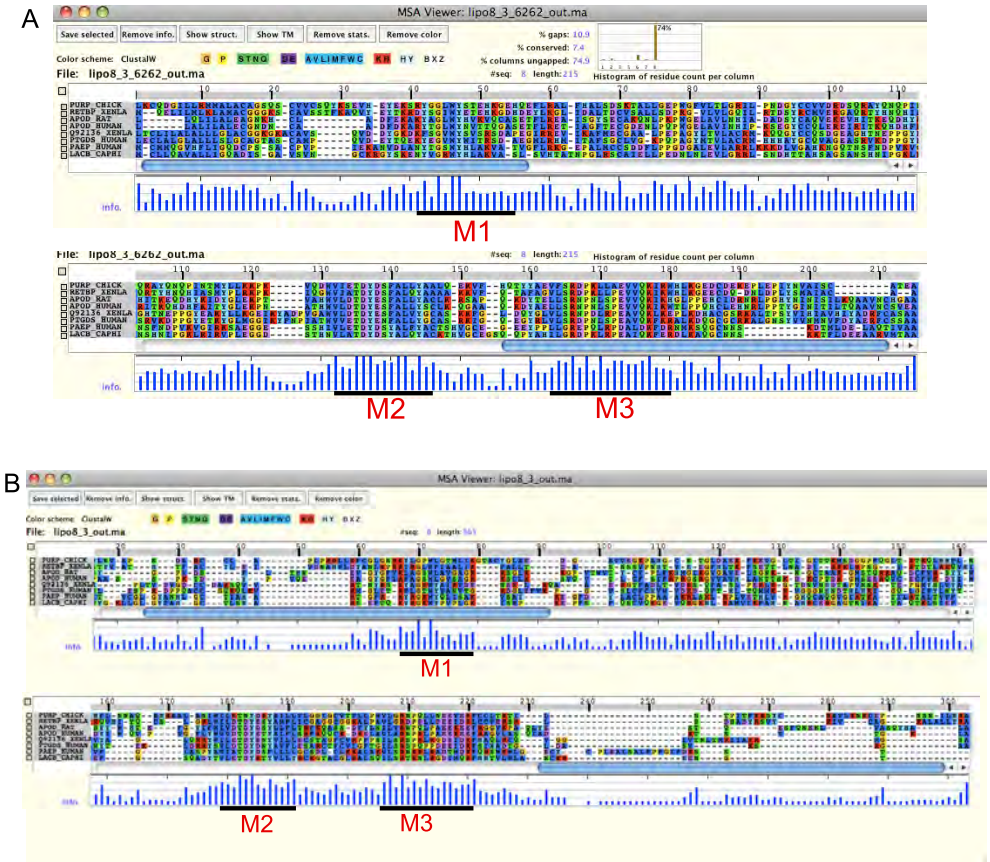

Fig. 12. 'True' alignments of the two sets of simulated lipocalin protein sequences (A: more conserved and B: more divergent simulations). Both alignments clearly show that the three motifs (M1, M2, and M3) are conserved among these two sets of simulated protein sequences.

Fig. 13. Comparison of simulated lipocalin protein MSAs. The Pixel Plot is used to compare four reconstructed MSAs with the reference alignments (A: more conserved and B: more divergent simulations). The 'true' alignments obtained from the simulations are used as the reference alignment (MSA 1). M1 (red box), M2 (blue box), and M3 (green box) show the location of three conserved regions. The regions highlighted in magenta show an example of inconsistent alignments found in reconstructed alignments relative to the true reference alignments. MSA methods used are MAFFT (MSA 2), MUSCLE (MSA 3), Probalign (MSA 4), and PRANK (MSA 5).

Fig. 13A shows that all four methods produced highly consistent MSAs for the sequences obtained from the more conserved simulation. While two of the three conserved motifs were identified correctly in all MSAs, in the region of the first motif (M1), all reconstructed MSAs contained gaps. Consistently very high SPS' (0.91~0.93, Table 4) indicate that all methods performed very well. The proportion of gaps is also consistent between all reconstructed MSAs and the reference (~10%, Table 4).

Method	SPS	CS with no-gap	CS with gaps	% consistency	MSA length (aa)	% conserved	% gaps	% no-gap columns
Reference	-	-	-	-	215	7.4	10.9	74.9
MAFFT	**0.933**	**0.850**	**0.782**	**72.77**	213	7.0	10.0	75.1
MUSCLE	0.921	0.817	0.751	70.28	212	7.1	9.6	**77.4**
Probalign	0.912	0.787	0.704	63.76	218	**7.8**	**12.1**	73.9
PRANK	0.932	0.826	0.776	71.5	215	7.5	10.5	75.2

Table 4. Alignment statistics for the simulated 'more conserved' lipocalin family MSAs. SPS, CS, and % consistency are obtained against the reference alignment. The highest value in each comparison is shown in boldface.

Method	SPS	CS with no-gap	CS with gaps	% consistency	MSA length (aa)	% conserved	% gaps	% no-gap columns
Reference	-	-	-	-	303	2.00	**39.70**	39.60
MAFFT	0.533	**0.481**	0.263	20.64	252	2.40	27.50	42.90
MUSCLE	0.532	0.421	0.314	**28.11**	219	2.70	16.60	**63.00**
Probalign	**0.585**	0.465	**0.321**	23.56	258	2.30	29.20	49.20
PRANK	0.504	0.395	0.264	24.88	213	**2.80**	14.30	60.60

Table 5. Alignment statistics for the simulated 'more divergent' lipocalin family MSAs. SPS, CS, and % consistency are obtained against the reference alignment. The highest value in each comparison is shown in boldface.

In Fig. 14, we compared PRANK and MAFFT alignments more in detail using the MSA Comparator. This is the same region highlighted in magenta color in Fig. 13A. The alignment columns that are fully consistent between the reference and PRANK (Fig. 14A) or MAFFT (Fig. 14B) are shown with blue color. Black characters, on the other hand, indicate inconsistently aligned columns. For example, the characters contained in the red square in the reference MSA are aligned exactly the same in the PRANK MSA. However, as shown in Fig. 14B for the MAFFT MSA, the gap in column 35 (in the reference) is filled with the characters shifted from the left. The Pixel Plot in Fig. 13 shows that the same shifting and filling of the gap happened in all but the PRANK MSA. This demonstrates the "gap magnet" phenomenon described in the previous section. Using the simulated data, we know the origins of gaps. In Fig. 14, '-' in yellow cells are derived from deletion events. Characters in green cells, on the other hand, are derived by insertion events. Therefore, stacking up the 'QVD' sequences and avoiding inserting gaps as done by MAFFT (Fig. 14B) is evolutionary incorrect. With commonly used affine-gap penalty systems, opening new gaps is highly penalized as opposed to extending an existing gap. This is reinforced with progressive alignment methods. This situation is clearly illustrated in the example shown in Fig. 14B. Using its "phylogeny-aware" gap handling, PRANK was able to correctly align these gaps.

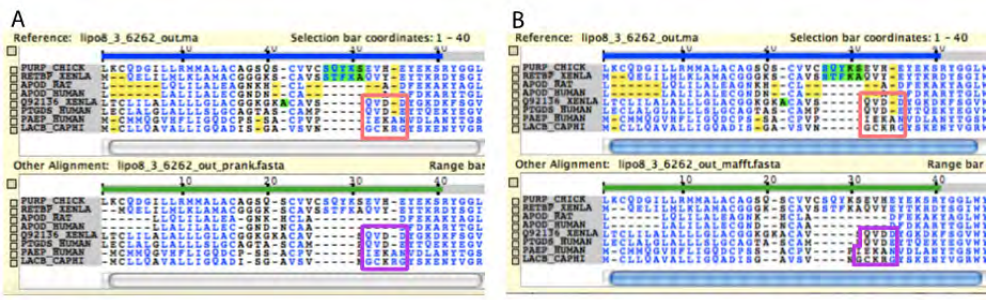

Fig. 14. Comparisons of PRANK (A) and MAFFT (B) alignments against the reference alignment (the simulated 'true' alignment). The region is taken from the area highlighted in magenta in Fig. 13A. The MSA Comparator is used to show the actual insertion (marked with green) and deletion (marked with yellow) events in the reference alignment. These events are traced during the iSGv2.1 simulation.

When the divergence level was much higher, as shown in Fig. 13B, all methods could still identify all of the three conserved motif sites. However, all MSAs were highly inconsistent within the unconstrained areas. SPS' are significantly lower (0.50-0.59, Table 5). For this dataset, PRANK produced the most inconsistent MSA, which was expected as PRANK is recommended for aligning closely related sequences. It should also be noted that there is little agreement among the alignments. This indicates that regardless of the statistics used, no one method can be concluded as ideal. Using multiple methods is recommended so that a selection of alignment hypotheses can be used to generate a more robust hypothesis.

7. Aligning ribosomal DNA sequences

We have so far concentrated our discussion on protein sequence alignments. In order to obtain the full picture of alignment issues, in this section, we will examine the alignment of ribosomal DNA (rDNA) sequences.

7.1 Small-subunit ribosomal DNA sequences and secondary structure

The ribosomal RNA genes contain large stretches of highly conserved sites (stem or knot binding sites) interspersed with regions of varying sites (loop regions). These two types of regions within the gene have different information content due to strong selective constraints on the secondary structures and function within the stem and knot areas *versus* very weak constraints on the loop area. Fig. 15 shows a predicted secondary structure of the small-subunit ribosomal RNA (or 18S rRNA) from a parasitic protozoa *Toxoplasma gondii*, a member of the family Sarcocystidae (Phylum Apicomplexa; Class Conoidasida; Subclass Coccidia).

Figs. 16 and 17 show part of the 18S rDNA MSA of 60 Coccidia species (D. A. Morrison personal communication; Morrison, 2009a). As shown in Fig. 16, stem regions are highly conserved. This alignment illustrates the high level of conservation found in approximately 45% of the 18S rDNA alignment. On the other hand, large loop regions as the one shown in Fig. 15 have much lower functional constraints. As shown in Fig. 17, sequences of such regions are highly variable and alignment reconstruction of such regions often requires

laborious manual adjustment, iteratively incorporating information from the predicted rRNA secondary structures (Morrison, 2009a).

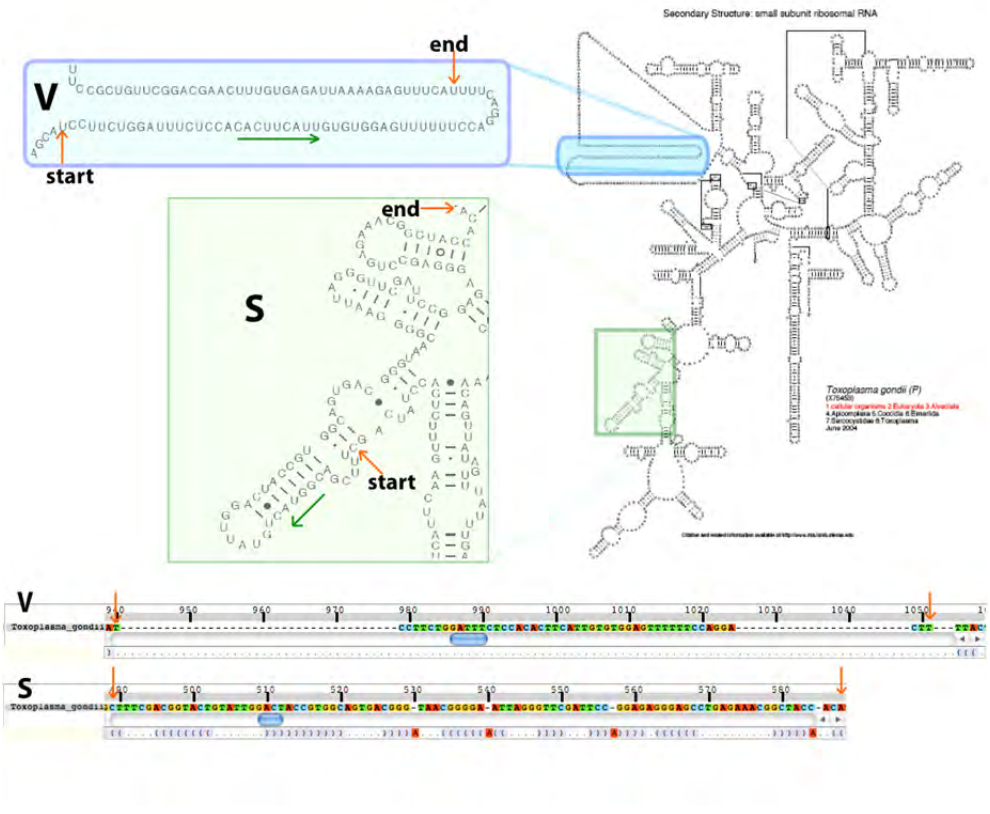

Fig. 15. Predicted secondary structure of the *Toxoplasma gondii* 18S rRNA. The secondary structure was obtained from Comparative RNA Web Site (http://www.rna.ccbb.utexas.edu/; Cannone *et al.*, 2002). The callouts 'S' and 'V' show the 'stem' and the large 'loop' regions, respectively. Their sequence-structure alignments are shown at the bottom (the orange arrows pointing to the beginning and ending of the regions).

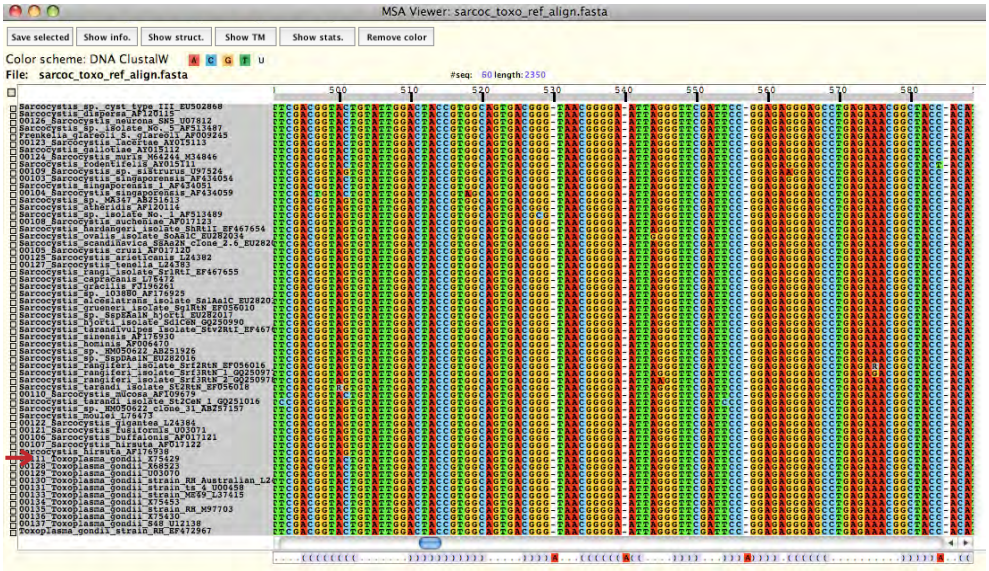

Fig. 16. Alignment of a highly conserved stem region of 18S rDNA from 60 Coccidia species. Using the MSA Viewer, the rRNA secondary structure information from *T. gondii* is displayed below the alignment. This alignment corresponds to the region in the callout 'S' shown in Fig. 15. The sites that are considered to be ambiguously aligned for this family are indicated by a red 'A' in the structural representation. These positions do not appear in the *T. gondii* structure. The alignment was provided by D. A. Morrison.

7.2 Comparison of 18S rDNA MSA reconstruction

We generated the alignments of full 18S rDNA sequences using four MSA methods. Using the above-mentioned alignment provided by D. A. Morrison as the reference, we compared the performance of the MSA methods. The alignment statistics are summarized in Table 6.

Method	SPS	CS with no gap	CS with gaps	% consistency	MSA length (nuc)	% conserved	% gaps	% no-gap columns
Reference	-	-	-	-	2095	50.60	15.80	62.40
Probalign	**0.953**	0.919	0.863	65.96	2389	44.30	**26.10**	55.60
MAFFT	0.950	0.898	0.844	73.08	2088	50.10	15.50	**64.10**
MUSCLE	0.950	0.917	**0.867**	73.77	2116	50.20	16.60	63.00
ClustalW2	0.948	**0.946**	0.855	**75.23**	2055	**51.40**	14.10	62.50

Table 6. Alignment statistics for the 18S rDNA MSAs. SPS, CS, and % consistency are obtained against the reference alignment. The highest value in each comparison is shown in boldface.

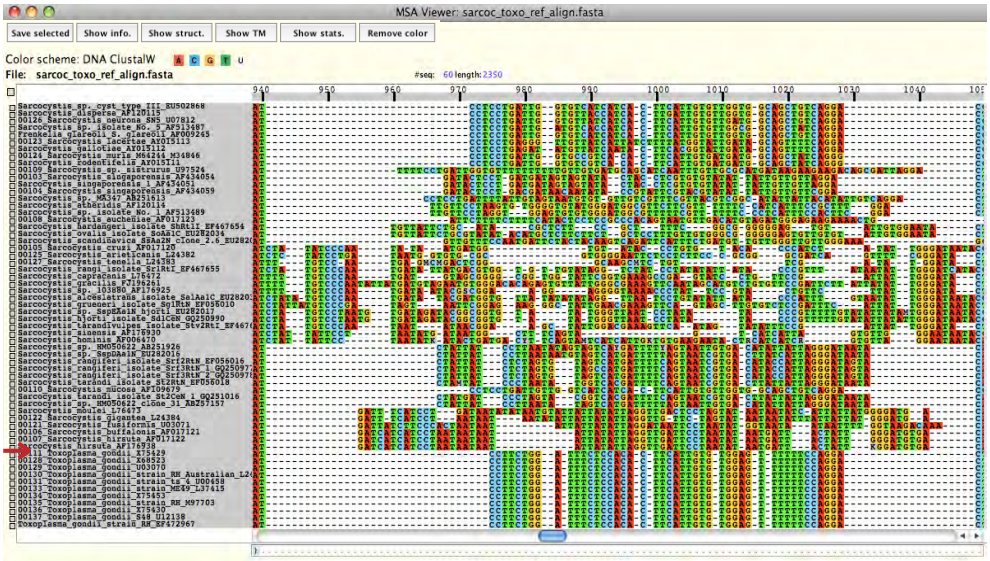

Fig. 17. Alignment of a highly variable loop region of 18S rDNA from 60 Coccidia species. One of the secondary structures used to refine the alignment was from *T. gondii*. This structure is displayed below the alignment. This alignment corresponds to the region in the callout 'V' shown in Fig. 15. The alignment was provided by D. A. Morrison.

All four methods appear to have produced alignments highly consistent with the reference. This must be owing to highly conserved stem or functional regions that cover almost 50% of the sequence regions. Such consistency is reflected by the high CS values particularly when gapped columns are excluded (CS with no gap) and also the small differences in CS values among MSAs. ClustalW2 has the highest un-gapped CS, indicating that ClustalW2 has the highest number of columns that match the reference alignment, and likewise, the highest % consistency. While the ClustalW2 MSA is the shortest (2,055 nucleotides), the longest and most 'gappy' MSA was obtained by Probalign. Similar trends are found in the other examples described in this chapter. Note, however, that the Probalign MSA had the highest SPS (0.953) and second highest 'CS with gaps' (0.863; MUSCLE had a slightly better score, 0.867).

Now let us visually examine these alignments. Keep in mind that the sequences are highly conserved, and that phylogenetic information will be derived mainly from the regions that are sufficiently variable. In Fig. 18, the Pixel Plot was used to compare the four reconstructed MSAs against the reference MSA. The selected area of the reference MSA under the blue selection bar includes the subsequence shown in the callout 'V' of Fig. 15 (Fig. 16 also shows the same area of the reference MSA). The magenta-colored pixels show the distribution of characters included in this selected area. In the reference MSA, the magenta-colored area has relatively small amount of gaps, providing the largest aligned overlap, putatively the most phylogenetically informative region, within this large loop region. However, in the alternative MSAs (MSAs 2-5), magenta-colored corresponding characters are spread over much wider regions (green bars show the ranges covered by corresponding characters). Note that each MSA method found a few conserved subsequences (matched columns) within this region.

However, each method also introduced a large number of gaps, affecting the consistency in the alignments of the surrounding areas immediately before and after the selected region. In spite of the high SPS' observed with these MSAs and the degree of conservation within the alignments, there is little consensus among the MSAs of this phylogentically critical area. Through visual comparisons among alternative MSAs it becomes possible to recognize that very different hypotheses could emerge depending on the MSA chosen. Such significant differences among the MSAs are, and should be, alarming to researchers, since such inconsistency in MSAs could affect phylogenetic hypotheses.

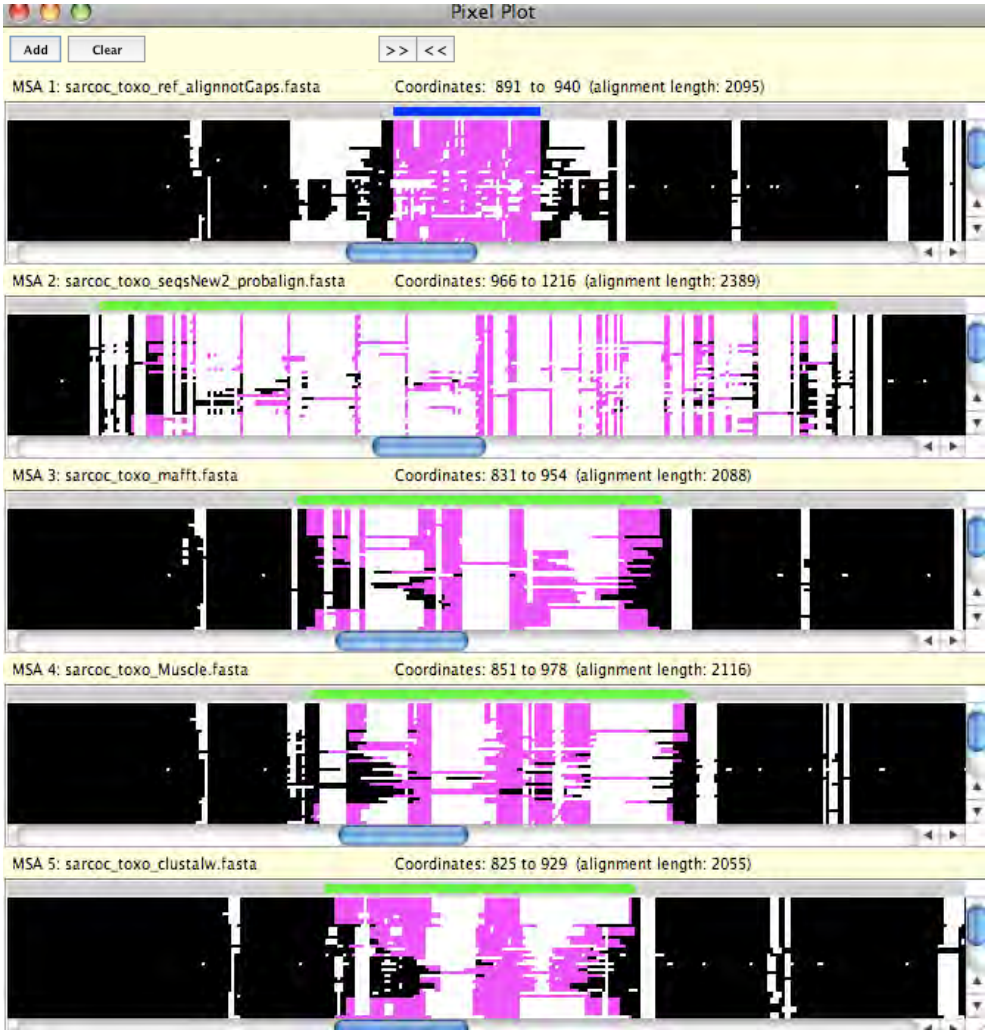

Fig. 18. Comparison of 18S rDNA MSAs. Pixel Plot is used to compare four reconstructed MSAs with the reference. The alignment provided by D. A. Morrison was used as the reference and compared with MSAs generated by Probalign, MAFFT, MUSCLE, and ClustalW2.

8. Conclusion

Advancements in the field of bioinformatics and molecular evolution have resulted in many different methods for reconstructing MSAs. While each MSA method has a different objective function and different heuristics to maximize the objective function for building the alignment, if they were in fact meant to reconstruct alignments that reflect the evolutionary history of sequences, we would expect some level of consensus between them. Such is not the case in reality. We used five types of alignment problems in this chapter. Using seven different MSA methods, we discussed the similarity and difference among MSAs built by these methods. We have shown that assessment of MSAs can be performed using a combination of descriptive statistics both for individual alignments and the comparison of two alternate alignments. We have also shown that using visual tools provided by SuiteMSA, we can examine MSAs based on the alignment of structural features such as secondary structure and transmembrane predictions. We further demonstrated how the sequence simulator included in SuiteMSA can be used to produce benchmark alignments.

We should keep in mind that alignments reconstructed by any MSA methods are only hypotheses on the evolutionary relation of the sequences. Furthermore, while these alignments can be assessed as consistent (or not) with the accepted model for the given sequences (the reference alignment), this reference is itself a hypothesis unless generated by a simulation program and may not be 'correct'. It is important for the researcher to understand the underlying assumptions of the alignment methods as well as the characteristics of the biological sequences to be aligned and to assess the resulting alignments. User friendly graphical tools such as SuiteMSA can assist in the critical assessment of MSAs prior to their use in further studies.

9. Acknowledgements

We would like to thank Dr. David A. Morrison (Swedish University of Agricultural Sciences) for providing us 18S rDNA alignments and discussion. This work has been partially supported by NSF AToL grant 0732863 to ENM.

10. References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology,* Vol. 215, No. 3, pp. 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research,* Vol. 25, No. 17, pp. 3389-3402.

Anderson, C. L., Strope, C. L., & Moriyama, E. N. (2011). SuiteMSA: Visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC bioinformatics,* Vol. 12, pp. 184.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics,* Vol. 10, pp. 421.

Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M., Pande, N., Shang, Z., Yu, N., & Gutell, R. R. (2002). The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics,* Vol. 3, pp. 2.

Cline, M., Hughey, R., & Karplus, K. (2002). Predicting reliable regions in protein sequence alignments. *Bioinformatics,* Vol. 18, No. 2, pp. 306-314.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics,* Vol. 5, No. 1, pp. 113.

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* Vol. 32, No. 5, pp. 1792-1797.

Edgar, R. C. (2010). Quality measures for protein alignment benchmarks. *Nucleic acids research,* Vol. 38, No. 7, pp. 2145-2153.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., & Bateman, A. (2010). The Pfam protein families database. *Nucleic acids research,* Vol. 38, No. Database issue, pp. D211-222.

Flower, D. R., North, A. C., & Sansom, C. E. (2000). The lipocalin protein family: structural and sequence overview. *Biochimica et biophysica acta,* Vol. 1482, No. 1-2, pp. 9-24.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, Vol. 59, No. 3, pp. 307-321.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F. E., & Vriend, G. (2003). GPCRDB information system for G protein-coupled receptors. *Nucleic acids research,* Vol. 31, No. 1, pp. 294-297.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., & Sigrist, C. J. (2008). The 20 years of PROSITE. *Nucleic acids research,* Vol. 36, No. Database issue, pp. D245-249.

Inoue, Y., Ikeda, M., & Shimizu, T. (2004). Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Computational biology and chemistry,* Vol. 28, No. 1, pp. 39-49.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology,* Vol. 292, No. 2, pp. 195-202.

Katoh, K. & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics,* Vol. 9, No. 4, pp. 286-298.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic acids research,* Vol. 36, No. Database issue, pp. D202-205.

Kemena, C. & Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics,* Vol. 25, No. 19, pp. 2455-2465.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology,* Vol. 157, No. 1, pp. 105-132.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics,* Vol. 23, No. 21, pp. 2947-2948.

Löytynoja, A. & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National academy of sciences of the United States of America,* Vol. 102, No. 30, pp. 10557-10562.

Löytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* Vol. 320, No. 5883, pp. 1632-1635.

Morrison, D. A. (2009a). Evolution of the Apicomplexa: where are we now? *Trends in parasitology,* Vol. 25, No. 8, pp. 375-382.

Morrison, D. A. (2009b). Why would phylogeneticists ignore computerized sequence alignment? *Systematic biology,* Vol. 58, No. 1, pp. 150-158.

Nugent, T. & Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC bioinformatics,* Vol. 10, pp. 159.

Pei, J. & Grishin, N. V. (2007). PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics,* Vol. 23, No. 7, pp. 802-808.

Pei, J., Kim, B. H., & Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research,* Vol. 36, No. 7, pp. 2295-2300.

Pirovano, W., Feenstra, K. A., & Heringa, J. (2008). PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics,* Vol. 24, No. 4, pp. 492-497.

Pirovano, W. & Heringa, J. (2010). Protein secondary structure prediction. *Methods in molecular biology,* Vol. 609, pp. 327-348.

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

Roshan, U. & Livesay, D. R. (2006). Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics,* Vol. 22, No. 22, pp. 2715-2721.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research,* Vol. 18, No. 20, pp. 6097-6100.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal,* Vol. 27, pp. 379-423.

Stebbings, L. A. & Mizuguchi, K. (2004). HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic acids research,* Vol. 32, No. Database issue, pp. D203-207.

Strope, C. L., Abel, K., Scott, S. D., & Moriyama, E. N. (2009). Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Molecular biology and evolution,* Vol. 26, No. 11, pp. 2581-2593.

Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M., & Morgenstern, B. (2005). DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC bioinformatics,* Vol. 6, pp. 66.

Thompson, J. D., Koehl, P., Ripp, R., & Poch, O. (2005). BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins,* Vol. 61, No. 1, pp. 127-136.

Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one,* Vol. 6, No. 3, pp. e18093.

Thompson, J. D., Plewniak, F., & Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics,* Vol. 15, No. 1, pp. 87-88.

Van Walle, I., Lasters, I., & Wyns, L. (2005). SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics,* Vol. 21, No. 7, pp. 1267-1268.

Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., Oliveira, L., de Vlieg, J., & Vriend, G. (2011). GPCRDB: information system for G protein-coupled receptors. *Nucleic acids research,* Vol. 39, No. Database issue, pp. D309-319.

Wistrand, M., Kall, L., & Sonnhammer, E. L. (2006). A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein science,* Vol. 15, No. 3, pp. 509-521.