Sequence analysis

A word-oriented approach to alignment validation

Robert G. Beiko, Cheong Xin Chan and Mark A. Ragan*

ARC Centre in Bioinformatics and Institute for Molecular Bioscience, The University of Queensland, Brisbane, Qld 4072, Australia

Received on July 14, 2004; revised on January 27, 2005; accepted on February 16, 2005
Advance Access publication February 22, 2005

ABSTRACT

Motivation: Multiple sequence alignment at the level of whole proteomes requires a high degree of automation, precluding the use of traditional validation methods such as manual curation. Since evolutionary models are too general to describe the history of each residue in a protein family, there is no single algorithm/model combination that can yield a biologically or evolutionarily optimal alignment. We propose a ‘shotgun’ strategy where many different algorithms are used to align the same family, and the best of these alignments is then chosen with a reliable objective function. We present WOOF, a novel ‘word-oriented’ objective function that relies on the identification and scoring of conserved amino acid patterns (words) between pairs of sequences.

Results: Tests on a subset of reference protein alignments from BAliBASE showed that WOOF tended to rank the (manually curated) reference alignment highest among 1060 alternative (automatically generated) alignments for a majority of protein families. Among the automated alignments, there was a strong positive relationship between the WOOF score and similarity to the reference alignment. The speed of WOOF and its independence from explicit considerations of three-dimensional structure make it an excellent tool for analyzing large numbers of protein families.

Availability: On request from the authors.
Contact: m.ragan@imb.uq.edu.au

INTRODUCTION

The goal of biological sequence alignment is to identify regions of similarity (often interpreted as homology) between two or more sequences, and associate these regions with one another to enable further comparisons. Existing algorithms for sequence alignment and validation are adequate for many problems, but two fundamental challenges persist: multiple sequence alignment is a computationally demanding, NP-complete problem (Bonizzoni and Della Vedova, 2001), and even if a complete solution for multiple sequence alignment were available, mathematical or statistical optimality (however defined) and biological optimality are not equivalent, due to the inevitable violations of implicit or explicit evolutionary models (Notredame, 2002; Sullivan and Swofford, 2001).

These challenges have led to the development of many heuristic algorithms and families on the one hand (reviewed in Notredame, 2002), and a plethora of different parameters and parameter settings on the other. The exact alignment method MSA (Gupta et al., 1995; Lipman et al., 1989) was followed by progressive algorithms such as CLUSTALW (Thompson et al., 1994), iterative methods such as IterAlign (Brocchieri and Karlin, 1998) and Prpp (Gotoh, 1996), and consensus-based methods including DiAlign (Morgenstern, 1999) and T-COFFEE (Notredame et al., 2000). The Poa alignment algorithm (Lee et al., 2002) represents multiple sequence alignments as a graph and is a novel approach to the idea of gaps in a sequence alignment. The growing body of protein structure data has led to databases of alignments based on three-dimensional structure such as SCOP (Lo Conte et al., 2002). However, the applicability and utility of structure-based alignment is presently limited by the small proportion of proteins with known three-dimensional structures, and by the different methods available for protein sequence comparison, each of which can have different optimal solutions (Feng and Sippl, 1996; Godzik, 1996; Koehl, 2001). The best sequence alignment algorithm for a given protein family may not be evident from observable family properties such as sequence length distribution or percent identity. If the best algorithm cannot be selected a priori, then it becomes a viable strategy to employ several alignment algorithms to construct alternative solutions, and subsequently to select the best among these. This principle is perhaps captured most effectively by the consensus-based approaches such as DiAlign and T-COFFEE, which generate multiple alternative alignments prior to selecting one that is ‘optimal’.

If multiple approaches are employed in parallel, a reliable, algorithm-independent validation method is required for choosing the winning alignment. While validation strategies are still subject to the limitations of mathematical optimality, an appropriate model can yield good selective power among alignments. Column-based strategies compute alignment scores that reflect the similarity among residues within the same column, with the fundamental assumption that residue similarity reflects common function or common evolutionary origin. Column scoring methods include the sum-of-pairs (SP; Carrillo and Lipman, 1988) and mean distance (MD: Thompson et al., 1997) scores, both of which attempt to quantify the similarity of residues within each alignment: in the case of SP, residues are scored directly against one another, while in MD the paired residues are compared in terms of all other possible pairings. AL2CO (Pei and Grishin, 2001) is a flexible scoring method that provides several ways of representing the residue composition of each alignment column, and then scoring the observed amino acid counts. The main purpose of this method is to identify regions within an alignment that have a dense arrangement of highly conserved columns. NorMD (Thompson et al., 2001) is a refinement of the MD score that corrects for sequence length and conservation, thus allowing a quality comparison of alignments of different sequences.

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org
The information content (IC) score (Hertz and Stormo, 1995, 1999) differs in principle from the column scores described above in considering the significance of each alignment column in light of residue frequencies from the whole set of sequences. A high IC score corresponds to pattern arrangements that are unlikely to occur in a random alignment, such as a single column that contains a highly conserved, rare residue. While column scoring methods and information content can be ‘vertically sophisticated’ and are reasonable for the assessment of similarities, the consideration of each alignment column in isolation ignores important ‘horizontal’ information that defines regions of homology. Even if all of the residues in a single column are identical, a column score cannot confirm that these residues are in fact functionally or evolutionarily related.

Validating alignments against folded structure is becoming more feasible with increasing numbers of available protein structures, and more accurate structure prediction methods. Recent structure-based validation methods include APDB (O’Sullivan et al., 2003), which relies on Protein Data Bank (PDB) structures and the assumption that homologous members of a protein family will have structural consistency. However, PDB contains only a small subset of all orthologous families, so in the context of whole-genome analysis, structure-based validation methods cannot address the entire dataset.

As an alternative to structural analysis, it is possible to identify and weight the regions within a protein family that are likely to be homologous, and base the quality assessment on the degree to which these regions of putative homology are in fact aligned. These regions can be expressed as patterns, and in practice can be obtained through a comparison of protein sequences with biological databases such as ProSite (Falquet et al., 2002) or BLOCKS (Henikoff et al., 2000), or generated de novo using programs such as TEIRESIAS (Rigoutsos and Floratos, 1998) or Splash (Hart et al., 2000). An advantage of TEIRESIAS is that the patterns it identifies do not have to be contiguous, thus allowing the representation of interspersed conserved and non-conserved residues as a single pattern. While de novo patterns extracted from a set of proteins are not assumed to have functional relevance, TEIRESIAS has been shown to extract patterns that can be associated with conserved protein functions. This property has been exploited in the construction of the BioDictionary (Rigoutsos et al., 1999), and in the functional analysis of predicted coding sequences within prokaryotic and eukaryotic genomes (Dehal et al., 2002; Shibuya and Rigoutsos, 2002). We have developed a word-oriented objective function (WOOF) that uses conserved amino acid sequence patterns to score protein alignments. This approach is analogous to the informal ‘inspection’ phase of sequence alignment, where an individual verifies an alignment by visually identifying conserved regions and ensuring that they are correctly aligned with each other. WOOF applies this principle in a rigorous manner by performing a weighted analysis of a complete set of patterns.

**SYSTEM AND METHODS**

**Sequences and Alignment**

Seventy-eight protein families (1aab to 9rnt), each with a corresponding reference alignment, were extracted from Reference set 1 in BAliBASE 1.0 (Thompson et al., 1999). These families have up to six members each, and are subdivided by alignment length (short, 61–142 columns; medium, 209–326 columns; long, 398–1002 columns) and by mean percent identity (low, <25%; moderate, 25–35%; high, >35%). Since they are manually curated by biologists, the BAliBASE reference alignments were treated as the ‘gold standard’, biologically optimal alignments for these protein families.

In addition to the reference BAliBASE alignment for each protein family, alignments of the original ungapped sequences were performed using a wide range of settings for T-COFFEE, ClustalW, Poa, IterAlign and Prrp, producing a total of 1060 alternative alignments (Table 1). With the reference BAliBASE alignment, this yielded a total of 1061 multiple sequence alignments for each protein family.

**Pattern extraction and weighting**

The fundamental principle of WOOF is the extraction of patterns from each pair of sequences within a protein family, and a relative weighting of these patterns based on expectations of homology. While patterns could be identified in more than two proteins at a time, pairwise extraction was chosen for two reasons. First, since patterns may be conserved to varying degrees across all members of a protein family, an extraction of multiple patterns would require an arbitrary compromise between the required degree of conservation and the number of instances of the pattern within the family. Second, requiring patterns to match across many sequences would limit the influence of distantly related sequences, since they would be swamped by the patterns extracted from more similar proteins within the family. If we are judging the alignment of every protein from a given family, then pairwise pattern extraction can ensure that distantly related sequences contribute patterns to the scoring scheme.

WOOF can accept any set of patterns with associated weights, but we used the TEIRESIAS algorithm to generate sets of amino acid patterns for alignment scoring purposes. When generating patterns, TEIRESIAS can require exact matches, or equivalence classes can be specified to

<table>
<thead>
<tr>
<th>Program (version)</th>
<th>Reference</th>
<th>Parameters</th>
<th>Number of alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLUSTALW (1.83)</td>
<td>Thompson et al. (1994)</td>
<td>Substitution matrix (Gonnet_pam250, Blosum62, Jones110, Str)</td>
<td>4</td>
</tr>
<tr>
<td>Poa (1.0.0)</td>
<td>Lee et al. (2002)</td>
<td>—</td>
<td>1</td>
</tr>
<tr>
<td>IterAlign (sunos-4.1.3)</td>
<td>Brocchieri and Karlin (1998)</td>
<td>Substitution matrix (Blosum62, Pam120, Pam250, Str)</td>
<td>4</td>
</tr>
<tr>
<td>Prrp (3.1.0)</td>
<td>Gotoh (1996)</td>
<td>—</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>1060</td>
</tr>
</tbody>
</table>

**Table 1. Alignment programs and parameter sets used, with total number of generated alignments**
allow for some substitutions between sequences. In separate experiments, exact patterns and two different equivalence classes of ‘alphabets’ (ChenB: [AG], [DEF], [FY], [KR], [ILMV], [QN], [ST], StructB: [CSI], [DNL], [EQ], [FHFWY], [ITV], [KMR]) were used to define conserved ‘words’ between pairs of sequences. The initial extracted patterns were required to contain at least three conserved characters within a string of 15 consecutive amino acids (L = 3, W = 15 in TEIRESIAS). These minimal patterns were combined using the convolution algorithm of TEIRESIAS to yield maximal patterns. TEIRESIAS can assign a single residue to more than one maximal pattern between a pair of sequences.

In some cases, the same pattern was found more than once in either or both members of a pair of sequences. If the number of instances of a single pattern was equal in both sequences, then they were simply paired off in the order they occurred in both sequences. However, if unequal numbers of the same pattern were found in the two sequences, most-likely pairs were chosen based on their positions within the sequence. Pairing was performed by calculating the start point of each instance of a pattern divided by the total sequence length, yielding two lists of values between 0.0 and 1.0. Elements in the list were paired off in such a way as to minimize the total sequence distance between paired pattern instances, with ties resolved in favour of keeping the patterns closest to the N-terminus of the protein. Unpaired patterns were discarded.

Each pattern shared between two sequences was then assigned an overall ‘importance score’, calculated as the product of the pattern’s log-likelihood score and the positional weighting of the pattern. The log-likelihood was calculated using the ‘evaluate3plets’ program in TEIRESIAS, and represents the log-likelihood of the pattern occurring by chance. The background model used by TEIRESIAS to estimate the significance of a pattern is based on a second-order Markov model of protein sequences in the GenPept database (Rigoutsos et al., 1999). The positional weight \( P_w \) expressed the relative position of the pattern in both sequences, and was calculated using the formula:

\[
P_w = 1 - \left( \frac{O_1 - O_0}{b} \right)
\]

where \( O_1 \) and \( O_0 \) represent the positions of the pattern relative to all other paired patterns in the first and second sequences, and \( b \) is the total number of patterns. The importance score quantifies the expectation of homology for a given pair of patterns, characterized by two different modes of conservation: sequence similarity as expressed by the log-likelihood score from TEIRESIAS, and consistency in pattern order as expressed by the positional weight. These two measures ensure that the patterns that are least likely to occur by chance in two proteins contribute the most to the final WOOF score, while weak patterns, and patterns with an unconserved order, are downweighted. While sequence position could have again been used for positional weighting, an advantage of pattern order is that it is more likely to be robust when some members of a protein family have large terminal extensions: while the position of a pattern in one an extended sequence could change substantially, its position relative to other highly conserved patterns found in the family would not change.

A subset of all maximal patterns from each pair of sequences was used in the subsequent scoring by WOOF. Patterns were added in decreasing order of importance score until the residue density requirement was satisfied, with a pre-specified number of sites contained within the set of patterns. The number of required sites for a given pair of sequences was determined by multiplying the length of the shorter sequence in the pair by the pattern density (between 0.4 and 0.8 in this analysis). Since TEIRESIAS can potentially return large numbers of very short patterns, a low pattern density will restrict the WOOF scoring to only those patterns with high importance scores, while a high pattern density will allow many more short, dispersed patterns to contribute as well. The optimal pattern density depends on whether these short patterns contribute to the discrimination ability of WOOF, or confound it due to non-homology.

**Alignment scoring with WOOF**

For each multiple alignment, WOOF considers each pair of sequences in turn. For each shared pattern, a score is calculated that represents the proportion of residues in the pattern that are correctly aligned by a given algorithm. This score is multiplied by the importance score \( I \) to yield the pattern score \( \Phi \):

\[
\Phi = I \cdot r_i \cdot r_l
\]

where \( r_i \) is the number of correctly aligned residues and \( r_l \) is the total number of residues in the pattern. The ratio \( r_i/r_l \) scales from 0.0 to 1.0, so the pattern score can have values between 0.0 (complete misalignment of every residue of a pattern in a given alignment) and 1 (correct alignment of all residues in the pattern).

The pattern scores \( \Phi \) for a pair of sequences \( i \) are summed over all \( bi \) patterns to yield a pairwise score, and the pairwise scores are summed for all sequence pairs and divided by the total number of pairs \( \psi \) to yield the score for a multiple alignment. The generalized WOOF function is thus:

\[
\text{WOOF Score} = \sum_{i=0}^{\psi} \left( \sum_{q=0}^{b(i)} \Phi \right)^{0.5} \psi
\]

with an exponent of 0.5 applied to each pairwise score. The effect of this exponent is to reduce the weighting of highly conserved sequences relative to less conserved ones, and avoid the situation where the score is largely determined by very similar sequences in a protein family. A pairwise exponent of 0.5 yielded the best mean normalized scores across the 78 BAiBASE families as compared with exponents of 1.0 or 2.0 (data not shown). However, the choice of pairwise exponent does not affect the ranking of alignments within any single protein family, so the ranking of the BAiBASE reference alignment was the same for all of the tested exponent values.

The 1061 WOOF scores obtained from each trial combination of protein family, TEIRESIAS alphabet and pattern density were normalized by dividing each score by the maximum score obtained in the trial. Since the WOOF score is sensitive to the number and weighting of extracted patterns, normalization was necessary to allow the comparison of different scoring distributions across different trials. The range of normalized WOOF scores is therefore between 0.0 (every paired pattern is completely misaligned) and 1.0 (the best observed alignment of weighted TEIRESIAS patterns).

**Calculation of alignment similarity**

The similarity between a pair of alignments was calculated by comparing the column associations of protein residues. For each residue in each sequence within the test alignment, 1.0 was added to the cumulative score for each other residue in its alignment column that was consistent with the reference alignment. The sum of similarities obtained was then divided by the mean of the self-scores of each alignment to yield a similarity score between 0.0 and 1.0. Normalizing with the mean of the two alignment self-scores yielded symmetrical scores that could be compared across protein families.

Two alignment similarity scores were computed for every pair of alignments within every protein family: one based on the entire sequences and another that considered only the ‘core’ regions of the BAiBASE alignments. In the latter case, the cumulative similarity of two alignments was incremented by 1.0 only for similarly aligned residues that were both derived from core blocks of the BAiBASE alignment. Pairwise correlations between these two scores were computed for each protein family, and a very high correlation coefficient was observed for most families (mean \( r \) across BAiBASE families = 0.930). Only those alignment similarity scores that were based on the entire sequences are shown in the Results section, but there is a very strong, positive linear relationship between the two sets of scores, with the ‘core’ region similarity scores slightly higher than those reported in the following section.

**Other validation methods**

Two alternative methods were used to score the test and BAiBASE alignments. The NorMD program (Thompson et al., 2001) was downloaded from ftp://ftp-igbmc.u-strasbg.fr/pub/NORMD and executed with the gap opening (1.0) and gap extension (0.1) penalties used in the original paper. The NorMD scores thus obtained for each protein family were then divided by the maximum score obtained (as with WOOF above) to yield a set of scores between
Word-oriented objective function for alignment

Fig. 1. Mean rank of scores assigned to 78 BAliBASE alignments, relative to 1060 × 78 test alignments, for different variants of the WOOF objective function. Bars indicate the performance of different equivalence classes and pattern densities between 0.4 and 0.8. The mean rankings of the BAliBASE reference alignment using NorMD and IC as objective functions are also shown.

RESULTS

BAliBASE coverage of patterns

Only 58% of alignment columns in BAliBASE are ‘reliable’ in that they are believed to be homologous and correctly aligned based on structural criteria. The remainder were not subject to the same level of structural validation, and so are classified as ‘non-core’ alignment sites. Although we used WOOF to score the entire alignment of any protein family (as we would with any family whose ideal alignment was not known), it is worth considering the regions from which the patterns were extracted within the sequences. Since protein regions that correspond to ‘reliable’ regions within the alignment are the most highly conserved, it is reasonable to expect that the most reliable paired patterns (i.e. those with the highest importance scores) will come from these regions. However, the protein regions that yield ambiguous alignments may also contribute patterns owing to the homology of some but not all proteins from a family at a given site and to the detection of short patterns that occur by chance.

To assess the relative weight of conserved and ambiguous BAliBASE regions, we partitioned the importance score of each extracted pattern into ‘conserved’ and ‘ambiguous’ components, reflecting the proportion of the pattern that fell into these two categories. These two components were pooled across all patterns extracted from a given protein family, to estimate the relative contributions of the conserved and unconserved regions to the WOOF function. While only 58% of BAliBASE alignment columns correspond to ‘core’ regions, we found that, on average, 85.2% of the total importance score of patterns from a given protein family was associated with these conserved regions. The proportional weight of patterns that occurred in conserved regions was the highest in the ‘1dox’ family, which has a relatively high conservation (46% identity); on average, 98.7% of the WOOF score was determined by the conserved regions. In the ‘lhavA’ family, which has only 15% identity, only 23.1% of the total pattern weight was derived from ‘reliable’ regions. This observation is not necessarily surprising, since only 39 out of 245 alignment columns are considered reliable. Visual inspection of non-core regions in BAliBASE alignments reveals substantial pairwise sequence similarity, so the identification of some conserved patterns in these regions is to be expected. Patterns derived from such non-core regions are still in many cases likely to reflect homology of the underlying sequences.

Comparison of alignments

Within each BAliBASE family, the alignment scores specific to that family, TEIRESIAS alphabet (exact matches or chemical/structural equivalence classes) and pattern density were ranked from 1061 (the worst score) up to 1 (the best score). The mean of the 78 ranks thus obtained for each alignment method expressed its performance over the entire set of BAliBASE families. A similar ranking was applied to the scores obtained with NorMD and IC to permit comparisons with the ranked WOOF scores. If the BAliBASE alignments have been optimized through manual curation by biologists, then a good objective function should assign the highest score to the BAliBASE alignments and lesser scores to automatically generated, presumably suboptimal alignments.

Figure 1 shows the mean rank of the BAliBASE reference alignment in the list of 1061 alignments scored with WOOF, for each combination of alphabet and pattern density. Within the WOOF trials, the use of exact patterns for validation yielded higher ranks for the reference alignment than either the ChemB or StructB equivalence classes. The average rank of the reference alignment when exact patterns were used for scoring ranged from 67/1061 to 50/1061, with no automated alignment method yielding a better mean rank than the BAliBASE reference. When the ChemB equivalence classes were used to generate patterns, the mean rank dropped to between 268/1061 and 207/1061. Forty-one alignment algorithms
Fig. 2. Breakdown of WOOF: (a) Exact, (b) ChemB, (c) StructB and IC scores (d) by BAliBASE family. In each panel, the upper graph shows the normalized score of the BAliBASE reference, while the lower graph shows the rank of the BAliBASE alignment within the set of 1061 test and reference alignments. The BAliBASE families are organized by sequence length (first letter: S = short, M = medium, L = long) and by degree of conservation (second letter: L = low, M = medium, H = high). Within each BAliBASE category, protein families are sorted from left to right in order of increasing normalized score of the corresponding BAliBASE reference alignment.

yielded better ranks than the BAliBASE alignment for this set of equivalence classes, including Prrp, all 4 CLUSTALW alignments and 36 T-COFFEE alignments. The StructB patterns performed better than ChemB on average but had a larger range, with a minimum rank of 283/1061 and a maximum of 173/1061. As with the ChemB equivalence classes, the Prrp and CLUSTALW alignments achieved better ranks than the BAliBASE alignment on average, but only 2 of 1050 T-COFFEE alignments outperformed BAliBASE as well.

There was a tendency for the mean rank to improve with increasing pattern density for the ChemB and StructB equivalence classes, but no such effect was visible when exact matches were required. The high average BAliBASE alignment ranks seen with exact patterns are more susceptible to large fluctuations in the rank of poorly performing BAliBASE families (Fig. 2) and there is less room for improvement within this group, so the influence of pattern density is not as clear.

The performance of the other two scoring methods was mixed. The IC score tended to assign very high scores to the BAliBASE families, yielding a final average rank of 28/1061. In contrast, the NorMD function did not favour the BAliBASE alignments to the same
degree as did WOOF and IC, with an overall ranking of 317/1061. Interestingly, the best performance was observed with objective functions that considered only literal amino acid matches (WOOF with exact patterns and IC), while performance diminished when degenerate characters were considered (WOOF with equivalence classes and NorMD).

Figure 2 shows the normalized score and rank of the BAliBASE reference alignments for the subset of trials performed with a pattern density of 0.8 (Fig. 2a–c), as well as the IC score (Fig. 2d), over all 78 BAliBASE families considered. For each set of scores, there is a clear effect of sequence conservation on the normalized score, with the reference alignments of the least conserved (SL, ML and LL) BAliBASE families typically scoring below 0.9, while the reference alignments for the families with medium and high conservation tend to score above 0.9 and above 0.99 respectively. The observed difference in normalized BAliBASE scores between the different equivalence class scoring schemes is evident here as well: among the WOOF scores, the reference alignment score derived from exact patterns was more frequently in the top 1% of all scores (157/1061, 19%) than were the scores derived from the ChemB (7/1061, 19%) and StructB (37/1061, 4%) equivalence classes. However, when rankings alone are considered, the performance of the IC score is remarkable: the majority (72/1061, 92%) of reference alignments scored in the top 1% of all alignments. This observation is surprising, since the normalized reference alignment IC scores tend to equal or be less than the WOOF scores obtained with exact patterns.

The scores of individual alignment algorithms were also examined although they are not shown in Figure 2. The highest-scoring alignment for most families and scoring schemes was typically a T-COFFEE alignment, which is not surprising since T-COFFEE alignments constituted 98.5% of the entire alignment set. The Prp and CLUSTALW alignments shown also attained high WOOF scores across many families, with average ranks from exact patterns of 53/1061 and 97/1061, respectively. The average performances of Poa and IterAlign were considerably worse, with an average rank of 154/1061 for Poa and 302/1061 for IterAlign. In contrast, the clear winner among normalized IC scores was the set of alignments generated with Poa, which attained a mean rank of 20/1061 over all families, higher than even the BAliBASE reference alignment mean rank of 28/1061 (Fig. 1). This observation explains the inconsistency between normalized IC scores and their ranks, since Poa alignments were so strongly favoured over the reference alignments of many protein families. IterAlign (102/1061), Prp (107/1061) and CLUSTALW (152/1061) all had worse mean ranks than Poa, but still outperformed the majority of the T-COFFEE alignments.

### Relationship between WOOF/IC scores and reference alignment similarity

Since experimental sequence alignment is typically conducted without a known reference alignment, a desirable property of an objective function is a strong positive relationship between the objective function score and the biological optimality of the generated alignment. We assessed this property of WOOF and IC by comparing the similarity between each test alignment and the BAliBASE reference against the WOOF and IC scores obtained for that alignment. The linear regression coefficients obtained for each protein family and grouped by class of protein family are shown for WOOF (Fig. 3a) and IC (Fig. 3b). The weakest relationships between score and similarity for WOOF are seen in families with low sequence conservation, most prominently when the sequence length is low as well. Most of the regression coefficients from this group are <0.3, while the mean $R^2$ values for the other two categories with low sequence conservation are ~0.6. Mean regression coefficients for the six categories of families with moderate and high sequence conservation were all between 0.75 and 0.85. Two outliers with very weak score–similarity relationships were found in the family with short, highly conserved sequences: the ‘lmnb’ family with an $R^2$ of 0.013 and ‘1km’ with an $R^2$ value of 0.64. In both these cases, each of the 1060 generated test alignments was at least 97% similar to the BAliBASE reference alignment, and the worst WOOF scores were only 20% lower than the best. This narrow range of data obscured the relationship that was observed in all other protein families with high sequence conservation.

Notwithstanding these unusual cases, there was a strong positive relationship between the WOOF score and biological relevance.
(as defined by similarity to the reference alignment) for most BAliBASE families. While the relationship between similarity to the reference alignment and IC score was also statistically significant for most protein families, the regression coefficients obtained were typically much smaller than those seen with WOOF. The relationship was particularly weak for families with low sequence conservation, with the majority of these families yielding an $R^2$ value < 0.1. The weakest relationship was observed in the ‘1ubi’ protein family ($R^2 = 1.0 \times 10^{-7}$) and the strongest was seen with the ‘1ied’ family ($R^2 = 0.81$). The largest regression coefficient for alignment IC score and reference alignment similarity is less than 35 such coefficients of the relationship between WOOF score and reference alignment similarity.

**Relationship with sequence length and conservation**

The raw WOOF score reflects the extent to which presumed identical patterns between sequences are correctly aligned, with weighted pattern scores to reflect both the likelihood of paired patterns occurring by chance and the positional consistency of these patterns. Since longer and more highly conserved protein families should contain more patterns of greater statistical significance, we expect to see positive relationships between these quantities and the WOOF score. Figure 4 shows the relationship between the maximum WOOF score from exact patterns for each set of the 1061 alignments and two separate parameters, one representing the length of the shortest sequence in a given protein family (Fig. 4a) and the other representing the mean identity between all pairs of sequences within a family (Fig. 4b). In both cases there is evidence for a positive relationship, with a particularly strong linear relationship between the WOOF score and the sequence length.

The strength of these relationships was confirmed with regression analysis. A linear regression of the maximum WOOF score versus both of the protein family parameters shown in Figure 4 yielded a regression coefficient of 0.936, thus explaining the majority of the variations in the WOOF score. The $p$-value associated with this regression was less than $1.0 \times 10^{-10}$.

**Similarity of inferred and optimal alignments**

In only a few instances did both WOOF and IC scoring methods favour the same alignment. However, it is possible to assess the similarity between the winning alignments for each protein family as defined in the System and Methods section, and compare them with the average similarity of the entire set of alignments generated for that protein family. Figure 5 shows the distribution of similarity scores between the best WOOF alignment and the best IC alignment for all 78 protein families. With the exception of 0.0–0.1, every similarity subdivision of size 0.1 is represented at least once, and although the median (0.76) and the mean (0.64) of the set of similarity scores are both > 0.5, there are a substantial number of very low similarity scores. These low scores correspond to protein families with very low sequence similarity such as ‘1ubi’ (IC/WOOF similarity = 0.10), while more conserved protein families such as ‘1ezm’ (similarity = 0.96) had winning alignments that were much more similar. In 7 of 78 cases the favoured alignments were the same: 5 of these instances were protein families of high sequence conservation and 2 were families with moderate conservation. This correlation between sequence similarity and similarity between the winning WOOF and IC alignments is not surprising, since alignments of weakly conserved protein families may be more

---

**Fig. 4.** Relationship between the maximum exact pattern WOOF score obtained from 1061 alignments of each protein family, and two properties of the family: the length of the shortest sequence (a) and the mean percent identity between all pairs of proteins in the family, as calculated from the BAliBASE reference alignment (b). Regression lines with corresponding $R^2$ values are shown in the two graphs.

**Fig. 5.** Distribution of similarity between the alignments with the best WOOF and IC scores. Alignment similarities were calculated as described in the System and Methods section, and were rounded to two decimal places.
susceptible to differences in algorithms and parameter settings. The influence of non-core alignment regions will be strongest in protein families with low similarity; since there is less signal from conserved regions to ‘anchor’ the alignment, correct alignment of residues in non-core blocks will be even more difficult and more highly variable. We defined two alignment similarity sets against which the winning alignment similarity scores were compared: the first was a ‘complete’ all-versus-all comparison of the 1061 reference and test alignments, and the second was restricted to one example of each of the five main algorithms considered. Across all protein families, there was a strong Spearman rank correlation (Spearman, 1904) between the winning alignment similarity and the average similarity of the complete set of alignments \( r = 0.84, P < 1 \times 10^{-15}, 76 \text{ df} \) and the average similarity of the restricted set \( r = 0.88, P < 1 \times 10^{-15}, 76 \text{ df} \). Though neither set of alignments constitutes a proper ‘background’ distribution, we calculated a pair of \( Z \)-scores for each protein family, in which the average similarity of either the complete or the restricted comparison set was subtracted from the similarity between the two winning alignments and the result divided by the standard deviation (SD) of the appropriate comparison set. When the complete alignment set was used, the average \( Z \)-score was only 0.27, with a lower median (0.18) and a large SD (1.03). When the restricted alignment set was used, the average \( Z \)-score across all 78 protein families was 0.25, with an even lower median (0.08) and an SD of 1.38. Thus, while both WOOF and IC tend to assign a high rank to the reference alignment, and both scores show some correlation with alignment quality, the alignments they favour are only marginally more similar to each other than are an average pair of alignments sampled from the set of 1061. The large SDs show that this weak trend is not even consistent across all protein families, with nearly half of all families exhibiting a similarity \( Z \)-score < 0.

**Similarity to the BAliBASE alignment**

Another way to assess the performance of different alignment algorithms is to compare their results directly with the BAliBASE reference using the alignment similarity score. For each of the 78 protein families, we compared the restricted set of alignments (one alignment from each type of algorithm) to the BAliBASE reference and ranked them according to their similarity. CLUSTALW (33 winners) and Prrp (35 winners) produced the majority of most-similar alignments, with the remainder due to T-COFFEE (9 winners) and, in a single case, Poo. The algorithms that tended to yield more winning alignments also tended to have higher margins of victory: winning CLUSTALW alignments were, on average, 4% more similar to the BAliBASE reference than the next most similar alignment, with smaller margins for Prrp (2%) and T-COFFEE (1%). The single winning Poo alignment was only 0.3% more similar to the corresponding reference alignment (1dox) than the CLUSTALW alignment of that protein family. The good performance of CLUSTALW and Prrp in terms of alignment similarity mirrors their tendency to score well when WOOF is used to assess alignment quality. In contrast, the Poo algorithm performed extremely well when IC was used to score alignments, but tended to yield alignments that were not as similar to the BAliBASE reference as those produced by other algorithms.

Finally, we compared the similarity of the BAliBASE reference alignment with the alignments yielding the best WOOF and IC scores, and with the entire set of inferred alignments. In seven of the 78 BAliBASE families considered, the alignments yielding the highest WOOF and IC scores were the same. The BAliBASE reference alignment was more similar to the winning WOOF alignment than to the winning IC alignment in 37 of the 71 remaining cases, and more similar to the winning IC alignment in the remaining 34 cases. Though the winning WOOF and IC alignments were usually more similar to the BAliBASE reference than the average similarity of all 1060 test alignments to the reference, in a minority of cases (11/78 for WOOF and 21/78 for IC) the winning alignment was less similar. These cases were usually observed in BAliBASE families where the majority of alignments were nearly identical, and the winning WOOF alignment was only slightly less similar (0–5%) to the BAliBASE reference than the average test alignment.

**DISCUSSION**

**Performance of WOOF**

Our assessment of WOOF, as well as the IC and NorMD scores, was based on the assumption that a good objective function should favour a manually curated, ‘biologically optimal’ (or ‘optimal to biologists’) reference alignment over a set of alignments generated automatically for the same protein family without reference to biological considerations. The experiments described here show that WOOF scores based on conserved exact patterns assign very high scores to these reference alignments. For cases where a reference alignment is not available, our trials with BAliBASE suggest that WOOF is an effective tool for identifying good (biologically or evolutionarily reliable) alignments from a set of automated alignments, even if none of these is, individually, biologically optimal. This is of particular concern in the ‘twilight zone’ of alignment problems, since alignment programs do not perform well on sequences of very low identity (Elófsson, 2002). This assertion is supported by our automated alignments of BAliBASE families with low sequence conservation: in many cases, even the best (highest-scoring and most similar to the reference) automated alignments were <50% similar (as defined in the System and Methods section) to the reference alignment. Poorly conserved sequences tended to yield few patterns with high statistical significance and positional weight relative to the overall sequence length. However, if the extracted patterns are indeed homologous, then they can still provide a ‘lattice’ onto which the rest of the alignment can be overlaid. Extracted patterns can also serve as a guide to the regions of a protein family for which alignment should even be attempted.

Like the WOOF score, the IC score also strongly favoured the reference alignments over the majority of test alignments. While the IC score considers each alignment column in turn, it takes into account the background probabilities of each amino acid (Hertz and Stormo, 1999), and thus incorporates a degree of the horizontal sophistication we sought in designing WOOF. Like WOOF, IC had the greatest difficulty in identifying good alignments when the aligned sequences were of low conservation.

It is somewhat surprising that the objective functions that consider only exact pattern matches (WOOF with exact patterns, and IC) so dramatically outperform those that consider equivalence classes (WOOF with degenerate patterns, and NorMD). Degenerate patterns extracted with TEIRESIAS were more extensive and had better log-likelihood scores than did exact patterns similarly extracted, but the additional information in these patterns seems not to convey a clear biological or evolutionary signal. The inclusion of degenerate characters may permit the inclusion of too many non-homologous residues, thus confounding the homologous signal within patterns.
There may be instances where the use of degenerate characters provides the better approach for a WOOF analysis. The exact patterns generated for some of the families with low sequence identity were few and poorly supported, thus providing minimal information to score the sequence alignments. In such cases, it may be preferable to use degenerate characters and risk including misleading signals to obtain a larger set of patterns. While WOOF relies on patterns detected between pairs of sequences, an approach that requires patterns to be present in a larger sequence set (such as all members of a small family, or 50% of all sequences in a larger family) might filter out some of the residues that are similar but not homologous, since the coincidental patterns would have to occur many times.

‘Shotgun’ multiple sequence alignment

While some sequence alignment algorithms performed better than others in the overall WOOF analysis, there was no single algorithm and set of parameters that performed well over all BALIBASE families. This outcome supports the idea that different alignment algorithms and parameters are suited to different classes of alignment problems. No single alignment algorithm was favoured within different subclasses of protein families, suggesting that the properties of a given protein family do not indicate which alignment strategy will be best for that family. A ‘shotgun’ approach in which several algorithms and parameter sets are used to align a single protein family may therefore be the best approach, if a reliable objective function can then be used to select the best alignment from this set.

There is a great deal of literature dealing with the identification of reliable, highly conserved regions within an alignment. These approaches are typically column based (Pei and Grishin, 2001; Valdar, 2002), though there is sometimes a component that favours stretches of conserved residues over islands of conservation (Castresana, 2000). The patterns used to assign a WOOF score to an alignment could also be used as the decision criterion for keeping or discarding alignment regions. Such a method could require a region within an alignment to have a certain amount of coverage by patterns, with the additional constraint that the patterns are correctly aligned.

Flexibility of WOOF and future directions

An important component of WOOF is the positional weighting of patterns between a pair of proteins. While this score yields the desired effect of diminishing the impact of patterns with inconsistent positions relative to other conserved sequences, two refinements to its implementation are worthy of consideration in the future. The first improvement would take into account the log-likelihood (or some other measure of support) of different patterns when the positional weighting is considered, such that weak patterns would have little or no impact on the weight of strong ones. Another improvement would involve the analysis of sets of correlated positions between a pair of proteins or within a whole protein family, to identify domains that may have been subject to translocation in their evolutionary history.

While the present analysis was based on patterns derived from TEIRESIAS, WOOF can use any set of patterns for scoring alignments of a protein family, as long as a set of pattern weights (TEIRESIAS log-likelihood scores in this analysis) is provided. Our incorporation of pattern log-likelihoods into the importance score is unusual, but these statistical units provide a convenient representation of similarity within a practical range (a factor of 10 for most non-trivial patterns) that permits a meaningful consideration of strong and weak patterns. The patterns extracted with TEIRESIAS are inferred to be similar due to a common evolutionary origin, with no assumptions about inherent functional or structural meaning. The preservation of these patterns through evolutionary time may be due to selective constraints on structure and function, or simply due to insufficient time for evolutionary divergence and fixation. Patterns from biological databases such as ProSite (Falquet et al., 2002) or BLOCKS (Henikoff et al., 2000) could be substituted, though it is unlikely that such patterns would cover the proteins to a sufficient degree to yield adequate discriminatory power between test alignments. A final option would be to define regions of conserved secondary structure within an alignment, and then use WOOF to assign scores to the residues that ‘should’ be aligned under a structural model of sequence divergence.

ACKNOWLEDGEMENTS

We wish to thank the two anonymous referees for helpful comments and Gerald Hertz for providing an updated version of the IC scoring package. This work was supported by ARC Grants DP0342987 and CE0348221.

REFERENCES

Hands-on objective function for alignment