BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 24 2011, pages 3385–3391
doi:10.1093/bioinformatics/btr587

Structural bioinformatics

Advance Access publication October 28, 2011

STRIKE: evaluation of protein MSAs using a single 3D structure

Carsten Kemena[1], Jean-Francois Taly[1], Jens Kleinjung[2] and Cedric Notredame[1,*]

[1]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG) and UPF, Aiguader, 88, 08003 Barcelona, Spain and [2]Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Evaluating alternative multiple protein sequence alignments is an important unsolved problem in Biology. The most accurate way of doing this is to use structural information. Unfortunately, most methods require at least two structures to be embedded in the alignment, a condition rarely met when dealing with standard datasets.

Result: We developed STRIKE, a method that determines the relative accuracy of two alternative alignments of the same sequences using a single structure. We validated our methodology on three commonly used reference datasets (BAliBASE, Homestrad and Prefab). Given two alignments, STRIKE manages to identify the most accurate one in 70% of the cases on average. This figure increases to 79% when considering very challenging datasets like the RV11 category of BAliBASE. This discrimination capacity is significantly higher than that reported for other metrics such as Contact Accepted mutation or Blosum. We show that this increased performance results both from a refined definition of the contacts and from the use of an improved contact substitution score.

Contact: cedric.notredame@crg.eu

Availability: STRIKE is an open source freeware available from www.tcoffee.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on 27 July 2011; revised on 22 September 2011; accepted on 16 October 2011

1 INTRODUCTION

The assembly of accurate multiple sequence alignments (MSAs) is one of the most important tasks in Biology. Judging from citation index, MSA packages have become the most widely used tools for biological sequence modeling. This should not come as a surprise given their wide range of applications that include phylogenetic tree reconstruction, profile building and structural modeling. These diverse modeling schemes are becoming increasingly important in defining the way experimental information is being transferred across uncharacterized sequences. Yet, trees, profiles and structural models all have one thing in common: their strong dependency on the MSA framework they are based upon. For instance, it is well established that homology modeling is very dependent on MSA accuracy (Claude *et al.*, 2004; Zhang and Skolnick, 2004). In a similar fashion, Wong *et al*. (2008) have recently described the influence of MSAs onto phylogenetic tree topologies, showing how minor variations across different aligners can lead to significantly different tree topologies. Albeit these authors have not formally established a direct dependency between MSA accuracy and proper phylogenetic tree reconstruction, they have nonetheless shown the existence of a strong methodological bias induced by the aligner's choice. More recently, Markova-Raina *et al.* (2011) have used a different set up to show how the use of different aligners can result in significantly different estimates for positive selection rates.

These observations confront biologists with a complex situation. On the one hand, it is clear that the choice of an alignment strategy will have an impact on any result drawn upon the MSAs; on the other hand, there is no simple way to quantify this impact and determine its positive or negative nature. Using sequence information alone, it is very hard to decide which one is the best among two or more alternative MSAs of the same sequences. This difficulty stems from the impossibility of accurately aligning proteins with <25% sequence identity using sequence information only. Sequence similarity is only a useful indicator of alignment accuracy for closely related sequences. The most biologically relevant alignment (defined with respect to structural information in standard datasets) is rarely the one having the highest possible score (Sierk *et al.*, 2010). The practical consequences of these observations are important as they imply the need for external evaluation criteria in order to evaluate MSAs in a biologically meaningful way.

The problem of estimating the biological relevance of an alignment is particularly acute considering alternative MSA methods. As shown in M-Coffee (Wallace *et al.*, 2006), different packages tend to deliver significantly different MSAs. Most scientists address this problem by ignoring it and using a single aligner [most frequently ClustalW (Thompson *et al.*, 2002)]. However, for those ready to consider the aligner as yet another parameter in a complex modeling pipeline, at least two sequence-based solutions exist. The first one involves combining all the alternative alignments into one unique consensus MSA. This approach has been shown to improve slightly on all the combined methods (Wallace *et al.*, 2006). Along the same lines, it has also been shown that regions showing a high agreement across methods are also more likely to be correctly aligned, as estimated by comparison with a reference MSA (Lassmann and Sonnhammer, 2005). The second approach is to evaluate the alignment for specific features difficult to take into account while building the MSA, but measurable on a complete model (e.g. completely conserved positions) (Thompson *et al.*, 2003). The AlexSys (Aniba *et al.*, 2010) strategy is built on this approach and combines features measured on the unaligned sequences (length, divergence, etc.) to determine

*To whom correspondence should be addressed.

the aligner that should give the best accuracy/efficiency trade off. In practice, meta-methods tend to be slightly superior, as they benefit from the combination of MSAs, but in any case, none of these two approaches manages to fully recapitulate the accuracy estimations one obtains when using structural information.

Structural information is the best conserved signal in proteins across long evolutionary distances. This is the reason why structural similarity comparisons have long been established as the acid test for estimating or evaluating protein sequence alignments. For more than a decade, multiple sequence aligners have been optimized for their capacity to restitute structure-based MSAs. Unfortunately, structural data are scarce and beyond well-established benchmarks, the only realistic way to use that information is to embed within a dataset at least two sequences with a known structure (O'Sullivan *et al.*, 2004). One can then either build the MSA (by combining sequences and structures) or estimate its accuracy *a posteriori*, by evaluating the structural superposition implied by the sequence alignment. This approach is very powerful, but remains hampered by the two structures requirement, a prerequisite only met within a small minority of protein families.

A very desirable solution would be to adapt this approach and make it work with a single structure, so as to deal with the rapidly growing number of globular protein families having at least one structurally resolved member. This can be achieved by aligning any sequence with another homologous sequence having a known 3D structure. One can then evaluate the structural correctness of the potential contacts projected from the structure via the alignment on the first sequence. This strategy is applied in the development of fold recognition methods (also called threading) (Bowie *et al.*, 1991; Jones *et al.*, 1992; Marin *et al.*, 2002; Wu and Zhang, 2008), where the compatibility between a query sequence and a template structure is assessed by knowledge-based potentials extracted from highly resolved protein structures. Over time many other potentials have been developed either based on physical principals or taking into account more empirical data like residue spatial environment. Such methods include Verify3D (Lüthy *et al.*, 1992), ProsaII (Sippl, 1993), MIGeval (Taly *et al.*, 2008) or Fugue (Shi *et al.*, 2001). While threading had originally been designed mostly for the recognition of remote homology relationship, it is only recently that its potential for the improvement of MSAs has also been considered. In 2004, O'Sullivan *et al.* used the T-Coffee package to combine threading and structure-alignment methods. Their results, however, were inconclusive and suggested that the contribution of threading to the MSA of sequences and structures to be relatively modest. More recently, Lin *et al.* (2003) addressed the same problem from a different perspective. Rather than using threading in order to improve the alignments, they asked if the equivalent of threading potentials could be used to measure the relative accuracy of alternative MSAs. Their approach relied on the computation of a contact substitution matrix [Contact Accepted mutation (CAO)]. The rationale behind CAO is that contacts should be preserved by evolution as they constitute the main network of interactions within a protein fold, a highly conserved feature in proteins. Unfortunately, the estimation of the CAO matrix was hampered by the limited amount of available data with respect to the high dimensionality of the matrix. Indeed, the matrix contains one entry for each pair of possible contacts which makes a total of 400 × 400 entries. To estimate this matrix, one needs pairwise alignments of sequences with known structures, a relatively rare commodity. As a consequence, the CAO matrix is largely underdetermined and lacks the statistical power it would need to achieve its initial goal of discriminating between alignments of different accuracy.

In this work, we have approached the same problem as CAO, but from a different angle. Rather than a contact matrix, we have estimated a potential matrix, conceptually similar to that described by Sippl (1993). Yet, in contrast to Sippl's approach, our metric [Single sTRucture Induced Evaluation (STRIKE)] relies on a purely empirical matrix, whose estimation does not involve any Boltzmann modeling. Since the STRIKE matrix does not require pairs of homologous structures, far fewer parameters can be estimated on a much larger dataset compared with CAO, thus avoiding both underdetermination and overfitting problems. The rest of our approach is conceptually similar to that used in CAO and we show here that STRIKE can be used to compare alternative MSAs in terms of their relative accuracy while using one structure only.

2 METHODS

2.1 Contact estimation

Intramolecular contacts were estimated using the Connolly framework (Connolly, 1983) where two atoms are considered to interact if a solvent molecule cannot be inserted between their molecular surfaces. Following common practice, water molecules were approximated with a single oxygen atom. An all-atom representation of protein structures was used here. This approach departs significantly from that of Lin *et al.* (2003), who only considered interactions between spheres representing the residue side chains (coarse-grained C^{β} atoms). Previous results (Taly *et al.*, 2008), based on molecular dynamics simulations suggest that such an approximation is not precise enough to reflect the complexity of existing interactions. In order to avoid any bias introduced by near-neighbor interactions (1–2, 1–3, 1–4, 1–5), that dominate secondary structure contacts, we have only considered long-range residue-to-residue contacts, involving two amino acids separated by at least five amino acids within the primary structure. Three types of contacts are considered here: main chain to main chain (MC–MC) contacts involving only atoms forming part of the protein backbone, side chain to side chain (SC–SC) made of side chain atoms and ALL–ALL in which atoms from either subset may form the contact.

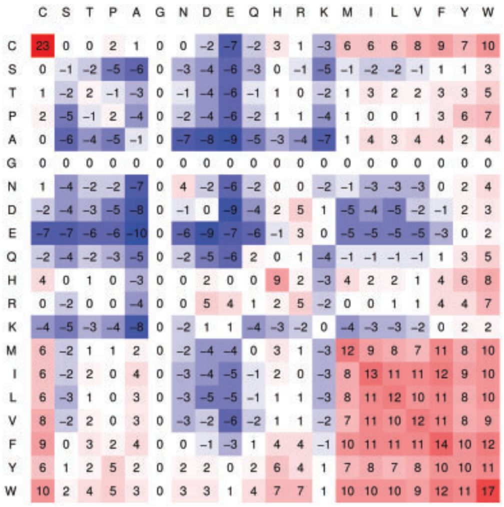
2.2 Structural dataset

We assembled a dataset of non-redundant protein structures from the ASTRAL database (Chandonia *et al.*, 2004) (version 1.75), a high-quality protein domain collection derived from the SCOP database (Murzin *et al.*, 1995) and the Protein Data Bank (Berman *et al.*, 2000). We followed the CAO strategy by first filtering out all low-resolution structures having an AEROSPACI score (Chandonia *et al.*, 2004) <0.5. The resulting subset was further trimmed by keeping only structures with <40% sequence identity at the pairwise level. The result was a dataset of 5242 structures, comprising a total of 979 934 residues and referred to as the ASTRAL subset in the rest of the text.

2.3 Contact matrix calculation

Using the reference structural dataset as a source for amino acid frequencies and contacts, we estimated a log-odds contact matrix by measuring the ratio between the frequency of each possible contact and its expectation given the background frequency of each single amino acid. Albeit it captures a different kind of signal, this matrix is conceptually comparable to a Blosum (Henikoff and Henikoff, 1992) or a PAM (Dayhoff *et al.*, 1979) matrix. Given any pair of amino acids i and j, the score for their contact can be estimated as:

$$M_{ij} = 10 \times \ln\left(\frac{f_{ij}}{f_i f_j}\right),$$



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	23	0	0	2	1	0	0	-2	-7	-2	3	1	-3	6	6	6	8	9	7	10
S	0	-1	-2	-5	-6	0	-3	-4	-6	-3	0	-1	-5	-1	-2	-2	-1	1	1	3
T	1	-2	2	-1	-3	0	-1	-4	-6	-1	1	0	-2	1	3	2	2	3	3	5
P	2	-5	-1	2	-4	0	-2	-4	-6	-2	1	1	-4	1	0	0	1	3	6	7
A	0	-6	-4	-5	-1	0	-7	-8	-9	-5	-3	-4	-7	1	4	3	4	4	2	4
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	1	-4	-2	-2	-7	0	4	-2	-6	-2	0	0	-2	-1	-3	-3	-3	0	2	4
D	-2	-4	-3	-5	-8	0	-1	0	-9	-4	2	5	1	-5	-4	-5	-2	-1	2	3
E	-7	-7	-6	-6	-10	0	-6	-9	-7	-6	-1	3	0	-5	-5	-5	-5	-3	0	2
Q	-2	-4	-2	-3	-5	0	-2	-5	-6	2	0	1	-4	-1	-1	-1	-1	1	3	5
H	4	0	1	0	-3	0	0	2	0	0	9	2	-3	4	2	2	1	4	6	8
R	0	-2	0	0	-4	0	0	5	4	1	2	5	-2	0	0	1	1	4	4	7
K	-4	-5	-3	-4	-8	0	-2	1	1	-4	-3	-2	0	-4	-3	-3	-2	0	2	2
M	6	-2	1	1	2	0	-2	-4	-4	0	3	1	-3	12	9	8	7	11	8	10
I	6	-2	2	0	4	0	-3	-4	-5	-1	2	0	-3	8	13	11	11	12	9	10
L	6	-3	1	0	3	0	-3	-5	-5	-1	1	1	-3	8	11	12	10	11	8	10
V	8	-2	2	0	3	0	-3	-2	-6	-2	1	1	-2	7	11	10	12	11	8	9
F	9	0	3	2	4	0	0	-1	-3	1	4	4	-1	10	11	11	11	14	10	12
Y	6	1	2	5	2	0	2	2	0	2	6	4	1	7	8	7	8	10	10	11
W	10	2	4	5	3	0	3	3	1	4	7	7	1	10	10	10	9	12	11	17

Fig. 1. STRIKE contact matrix. Each entry corresponds to the score of a contact between two amino acids. The color code merely reflects the numeric value of the entry (blue: negative, red: positive). Amino acids in the vertical column correspond to the more N-terminal amino acid of the considered interaction while the top row corresponds to the more C-terminal amino acids.

where f_{ij} is the frequency of contacts involving i and j across all observed residue–residue contacts, f_i and f_j are the single residue frequencies in the considered dataset. This formula therefore amounts to estimate whether the two considered residues i and j are more (or less) often in contact than one would expect by chance. In this framework, positive values imply an evolutionary favored contact, whereas negative values imply a disfavored contact. This approach defines the STRIKE matrix, displayed in Figure 1. Note that the analysis presented was carried out with full precision values (Supplementary Material).

2.4 Matrix entropy calculation

The formula below estimates the entropy of a matrix. This measure reflects the average information gain per aligned residue (as defined by Shannon) and may be described as an estimate of how much the considered matrix departs from the null model (zero average information gain). In the context of this work, we used the formula of Henikoff and Henikoff (1992) and Yu *et al.* (2003).

$$H=\sum_{ij}\log_2\left(\frac{p_{ij}}{p_i p_j}\right)p_{ij},$$

where p_{ij} is the probability for i and j to be in contact while p_i and p_j are the probability of their occurrence in the sequences. Probabilities were estimated using frequencies measured on the structural dataset.

2.5 Random model estimation

For the sake of validation, 1000 random matrices were generated and compared with the STRIKE matrix. These were obtained by randomizing the ASTRAL subset sequences using the original amino acid frequencies. We concatenated all sequences, shuffled them and split them up again. The random sequences were then projected onto the cognate structure and used to derive random contact counts. Note that no 3D modeling has been performed and the contacts were those estimated on the *bona fide* structure.

2.6 Alignment evaluation

Given an alignment between a sequence with a known structure (Template) and a sequence of unknown structure (Target), the STRIKE score is estimated by projecting all the contacts measured on the template onto the target. The scores of these induced contacts $c_i c_j$ (as given by the STRIKE matrix) are then summed up and normalized by the total number of contacts |C| within the template.

$$\text{Score}(\text{Target})=\frac{\sum_{ij}\text{STRIKE}(c_i c_j)\times\text{IsContact}(c_i c_j)}{|c|},$$

with

$$\text{IsContact}\,(c_i c_j)=\begin{cases}1, & \text{if } c_i\ c_j \text{ are in contact}\\ 0, & \text{else}\end{cases}.$$

Given a MSA A, the scores for each of the N target sequences are then added up to yield to the final alignment score $\text{STRIKE}_{\text{ali}}(\text{A})$:

$$\text{STRIKE}_{\text{ali}}(A)=\sum_{\substack{i\\ i\neq\text{Template.}}}^{N}\text{Score}(\text{Target}_i)$$

Whereas this formula has no explicit gap penalty, in practice the normalization by the number of contacts in the template plays a similar role and target sequences missing many contacts because of unaligned residues end up with a global score lowered accordingly.

The contact score can be further normalized by dividing its value with that of the Template sequence. This measure gives an indication as to whether the overall score of the Target sequences is significantly lower (<1), comparable (=1) or higher (>1) than that of the only known structure.

2.7 Validation databases

The various scoring schemes described above were tested on three different reference collections: BAliBASE3 (Thompson *et al.*, 2005), Homstrad (Mizuguchi *et al.*, 1998) and Prefab (Edgar 2004). These collections are made of structure-based sequence alignments. It is a common practice to consider these alignments as gold standards when evaluating multiple aligners. BAliBASE3 consists of six subsets of structural reference alignments, with a total of 379 usable alignments with empirically defined core regions. Homstrad is also made of multiple structure-based sequence alignments. In the context of this work, we only kept the 233 alignments containing more than four sequences. Prefab4 is one of the most extensive databases. It contains 1682 structure-based pairwise sequence alignments. Each structure comes along with ∼25 homologous sequences, thus making up datasets of ∼50 sequences with two known structures. The datasets come along with core regions defined by the agreement between two structural aligners.

2.8 Evaluation of datasets

Reference databases are normally used to determine the accuracy of an aligner. Sequences are aligned with the chosen aligner whose merits are then quantified by a comparison of its output with the reference alignment. In the context of this work, we were not so much concerned with establishing the accuracy of an aligner, but rather with a comparison of the STRIKE score with established evaluation metrics on the output of different aligners. Therefore, a number of different MSA evaluation metrics was used here. First, the companion scores of some databases (BaliScore for BAliBASE, qscore for Prefab) were used in order to establish the absolute accuracy of each individual MSA tested in this study. As reference score (RS) we used the sum-of-pair scores, which denote the proportion of amino acid pairs that are aligned in the test alignment the same way as in the reference. The same MSAs were also evaluated in terms of their sum-of-pair scores (using a Blosum62 matrix), their CAO score and STRIKE score using the metrics has been defined earlier in this section.

3 RESULTS

Our first task was to determine which type of contact yields the most informative log-odds substitution matrix: for this we checked the MC–MC, SC–SC and ALL–ALL contacts. We determined the corresponding matrix values on the ASTRAL subset and measured the entropy of the resulting matrix. As one would expect, the side chain contacts are those yielding the most informative matrix. Indeed, while the MC–MC matrix has an entropy of 0.06 and the ALL–ALL has an entropy of 0.12, the SC–SC reaches 0.46. This value is in the same range as that reported by Altschul (1991) for the Blosum45 (0.4) and Blosum62 (0.7) matrix. In order to assess the statistical significance of this value, we generated 1000 random matrices and found their entropies to be distributed normally, with mean entropy of 2.7×10^{-4} and a SD of 3.8×10^{-5}. Altogether these results suggest that the SC–SC matrix (named STRIKE matrix in the rest of this text) has all the desirable properties for evaluating the conservation of contacts in MSAs. These finding is in agreement with previous results (Taly *et al.*, 2008) where it is shown that main chain contacts are less informative for the task of discriminating correct 3D predicted models from incorrect ones if the decoys are based on templates sharing structural motifs with the native structure. MC atoms mainly contribute to the stabilization of secondary structure elements and therefore constitute energetically favorable components of the fold. In the context of STRIKE, MC–MC connections disturb the signal because they increase the frequency of contacts between residues not sharing the adequate physicochemical properties for an interaction. As one can see in Figure 1, the STRIKE matrix recapitulates quite well the best-known properties of protein structures. For instance, the entry with the highest score is the cystein–cystein interaction, an observation that is in good agreement with the well-established importance of disulfide bridges. Likewise, contacts between hydrophobic residues interactions tend to have positive values, whereas contacts between charged residues with the same sign (+1 or −1) have a clear negative trend. The interactions between amino acids with opposite charges are only slightly positive, as one would expect given the tendency of these residues to interact with the solvent rather than forming salt bridges within the hydrophobic core.

Only glycine contacts remained undetermined, owing to the lack of a side chain. The corresponding entries were set to 0. It is interesting to note that the matrix is slightly asymmetric. This reflects the underlying asymmetry of residue–residue interactions along the peptide chain. Indeed, the contacts are determined in Nter → Cter direction and can therefore not be assumed to be reversible. The most asymmetric entry is the proline–tryptophan interaction, with $P\rightarrow W=5$ and $W\rightarrow P=7$. Aside from this extreme observation, most of the other values have limited variation when comparing the Nter → Cter with the Cter → Nter entry.

Our next task was to estimate whether the STRIKE method is suited to discriminate between accurate and less accurate sequence alignments (accuracy being defined with respect to the reference alignment). For this purpose, we estimated the correlation between the reference score (RS, as produced by the BAliBASE3 program BaliScore) and the normalized STRIKE score (see Section 2 for details). For each dataset, a STRIKE score was computed independently for each structure and its related sequences (i.e. one sequence with known structure was used as a template while others were considered as targets). The result is displayed in Figure 2. The resulting correlation with a Pearson's coefficient of $r=0.54$ (Spearman's rho = 0.51) is very weak. Nonetheless, the graph suggests the existence of a trend in the relation between structure-based accuracy estimates and STRIKE scores.

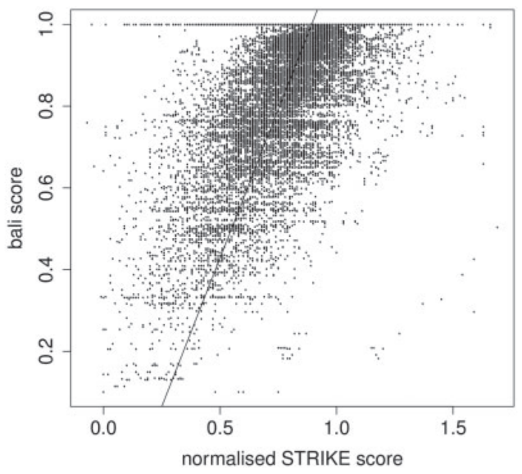


Fig. 2. The correlation factor on Homstrad of the normalized score with the BaliScore. For better display some outliers (38 points) were removed from this picture.

Our goal was not so much to estimate the MSA accuracy in absolute terms, but rather to tell apart two or more alternative alignments of the same structure/sequence pair. For that purpose, we were therefore more concerned with the existence of a non-parametric correlation allowing such comparisons. A non-parametric correlation exists whenever the relation of order defined between two observations is similar across the two considered metrics. For instance, if the reference score indicates that a T-Coffee MSA is more accurate than its ClustalW counterpart, the relation of order is conserved if the STRIKE score of the T-Coffee MSA is superior to the ClustalW MSA STRIKE score. In order to estimate the existence of such a non-parametric correlation, we computed alternative alignments of each available dataset using seven methods: ClustalW (Thompson *et al.*, 1994), Mafft (Katoh *et al.*, 2005), Muscle (Edgar, 2004), PCMA (Pei *et al.*, 2003), POA (Grasso and Lee, 2004), Probcons (Do *et al.*, 2005) and T-Coffee (Notredame *et al.*, 2000). We also included the reference alignments themselves, and treated them as an eighth additional method. We then computed the RS for each MSA, using the evaluation package suitable for the considered reference dataset resulting in a single accuracy score for each MSA. Using the same alignments, we evaluated the scores we wanted to test, including a Blosum62 sums-of-pairs, a Pam250 sums-of-pairs, the CAO score (one for each structure) and the STRIKE score. Given any of these datasets and two alternative alignments generated with two different methods, we plotted the difference in RS versus the difference in the test score.

The results obtained on the RV11 component of the BAliBASE 3 data when comparing the difference in RS with the difference in STRIKE scores are shown in Figure 3. In this graph, each point corresponds to one dataset, aligned with two different aligners.

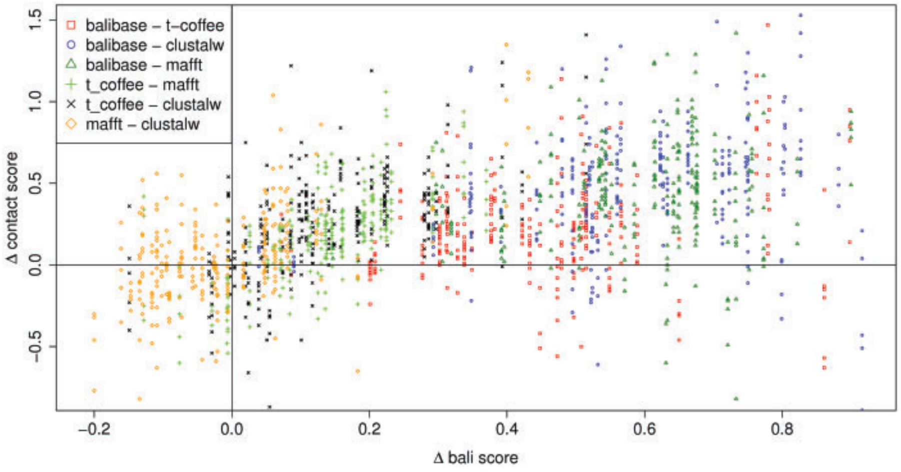


Fig. 3. Comparison of Δ BaliScore and Δ STRIKE score on BAliBASE3 RV11 using alignments produced by T-Coffee, Mafft and ClustalW as well as the reference alignment. All points which have the same algebraic sign are correctly classified.

The horizontal axis indicates the difference of RS between the two MSAs while the vertical axis represents the difference in STRIKE score (considering the same template structure). The total number of points is therefore a product of the number of datasets, the number of structures they contain and the number of possible pairwise method comparisons. As suggested by Figure 2, the absolute correlation between these differences is relatively weak. Nonetheless, observations made in Figure 3 suggest a very strong non-parametric correlation. Indeed, the two quadrants corresponding to RS differences and STRIKE differences having the same sign contain the most data points (top right and bottom left quadrant). In total, a significant proportion of 79% of the points fall into these quadrants. Note that the striped patterns that can be seen on this graph result from the same MSA being evaluated several times for its STRIKE score (once for each structure it contains). By definition, the reference alignments (red squares) always have the highest accuracy and it is interesting to note that a vast majority of these reference alignments are in the proper quadrants of the graph.

A graph like the one shown on Figure 3 can be summarized with an estimate of the fraction of measures falling within the two 'correct' quadrants. We therefore repeated the same analysis on the whole BAliBASE3, Prefab3 and Homstrad databases in order to compare the effect of using different metrics like sequence-based measures (PAM and Blosum) or CAO. The results are summarized in Table 1. It is interesting to note that these results are in broad agreement across the three datasets. As one would expect, the sequence-based measures (PAM and Blosum) have a limited capacity of discriminating MSAs for their structural accuracy. Overall, using them to rank two alignments is only slightly more accurate than flipping a coin (50–55%). Surprisingly, the CAO-based metrics is not significantly more informative than a direct sequence analysis. This surprising result is most likely a consequence of the underdetermination of the CAO substitution matrices. The last column summarizes the tests carried out with STRIKE. It shows that in this context, our approach yields the best results. On the most challenging dataset (RV11, made of distantly related sequences with a known 3D structure), the discrimination capacity of STRIKE is >79%. On the rest of the datasets, this capacity ranges from 65% to 70%, a significant improvement over the alternative methods considered here. Interestingly, this discriminative capacity seems to be independent of the nature of the considered proteins and the performances are relatively even when considering most structural subclasses (Table 2). Categories with <50 members were excluded.

Table 1. The sum-of-pairs score, CAO score and STRIKE applied to BAliBASE 3, Homstrad and Prefab

Dataset	#comp.	PAM	Blosum	#comp.	CAO	STRIKE
RV11	1036	56.3	55.8	7000	42.5	79.2
RV12	1148	59.2	58.4	3556	50.9	70.4
RV20	1148	56.8	56.3	5544	48.7	64.9
RV30	840	57.4	57.5	4480	49.4	66.1
RV40	1316	58.4	58.3	6328	51.6	66.8
RV50	420	55.0	55.5	2520	55.2	66.8
BAliBASE total	5908	57.5	57.2	29428	48.8	69.7
Homstrad	6496	54.5	52.7	46200	43.7	67.0
Prefab	47012	57.5	57.9	91644	47.4	67.4

The number of comparison (# comp.) is much higher for the structural measurements because a score can be computed for each structure included.

Table 2. Performance measurement of STRIKE on all three databases dependent on their classification according to SCOP

Class	#chains	#comp.	STRIKE
All α	312	15568	64.3
All β	437	21560	69.4
α and β (α/β)	724	41300	67.7
α and β ($\alpha+\beta$)	636	32424	67.3
Multidomain proteins (α and β)	59	3080	69.6
Small proteins	103	3752	62.9

#chains represents the number of different PDB chains found in this class.

When using a method like STRIKE to estimate the relative accuracy of two MSAs, it is important to have a notion of how

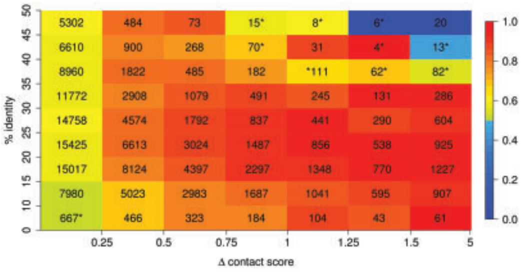


Fig. 4. STRIKE classification values as function of the differences in sequence identity and delta contact score. Numbers marked with at '*' have a $P > 0.001$. The numbers in the cells give the overall number of alignments. The color denotes the percentage of correctly classified alignment pairs.

meaningful the score may be and whether it might be influenced by the level of sequence identity in the considered dataset. We therefore binned the results of the analysis presented in Table 1 according to the average level of sequence identity in the considered MSAs (as estimated on the reference). Results are displayed in Figure 4, where the color code indicates the fraction of residues appearing in the 'correct' quadrants. The P-values were computed using a two-sided binomial test with $P_0 = 0.5$. Given two alignments and a difference in STRIKE score, this figure gives an estimate of how much trust one can have that the STRIKE difference reflects the difference in structural accuracy. The strongest trend on this graph is the poor level of information provided by STRIKE score differences <0.25. It seems one can rarely use this level of difference to effectively distinguish between two alternative alignments. Above this value, if we ignore the situations with <100 counts, the level of information provided by a difference in STRIKE score is roughly constant across all ranges of identity. With sequences within the twilight zone (20–30%), differences of STRIKE score >0.5 make it possible to identify the most accurate MSA in nearly 90% of the cases. On our datasets, such discrimination can be carried out on a significant fraction of the data.

The reason why our approach is much more informative than CAO is not entirely clear. There are two main differences between our method and CAO. First, our metric evaluates contacts rather than the conservation of contacts. Second, the STRIKE matrix was determined using all-atom contacts instead of coarse-grained C^{β} spheres as in CAO. We therefore determined an additional STRIKE matrix using the coarse-grained side chain contacts of CAO instead of the all-atom STRIKE contacts (Lin et al., 2003). The resulting matrix has an entropy <0.3 and when tested on Homestrad, it was only able to correctly classify 52% of the alignments. While this figure is higher than that of CAO itself (46%), it is also significantly lower than the original STRIKE matrix whose improved performances are therefore quite likely to be a combination of better contact estimation complemented with a larger amount of usable information.

4 DISCUSSION

In this work, we present the STRIKE score, a new metric using a single structure to estimate the structural correctness of MSAs. We show that using STRIKE, one can distinguish between alternative MSAs of the same sequences more reliably than when using similarity based scoring schemes, like Blosum or PAM. STRIKE is not the first attempt to compare alternative MSAs using the information from a single sequence, and in this work we have extensively compared our approach with that developed in CAO, which is conceptually similar albeit more complex in its implementation. Our results suggest that comparisons based on STRIKE result are more likely to reveal the most trustworthy MSA than those using CAO. Two reasons explain this difference. First, the STRIKE matrix is not estimated on structural alignments but on single structures. Consequently, the STRIKE matrix could be estimated on a much larger reference dataset. In comparison, the CAO matrix that considers the substitution cost for every contact pair of amino acids was probably underdetermined. The second reason for the reported improvement is the all-atom definition of contacts. While CAO was determined using coarse-grained C^{β} side chain contacts, STRIKE uses a more sophisticated definition of contacts discriminating between side chains and backbone. We showed that this choice accounts for the largest part of the difference between CAO and STRIKE.

We also explored the factors influencing STRIKE's capacity to discriminate between accurate and inaccurate MSAs. Against our expectation, we found the protein class to have only a weak influence on the trustworthiness of STRIKE. On the datasets analyzed here, the variations between all α, all β and $\alpha-\beta$ structures are <5 percentage points. We found that the differences in STRIKE scores are most useful when considering datasets with low average identity, and unsurprisingly when the difference in the scores is high. For instance, a STRIKE difference of 1 nat on a dataset with 30% average sequence identity makes it possible to identify with a reliability close to 90% the most correct of the two considered.

This makes STRIKE a potentially very useful tool to evaluate the alternative MSAs. Indeed, following the work of Wong *et al.* (2008) on the uncertainty of phylogenetic trees, biologists are now confronted with a complex situation. On the one hand, it has now been established that alternative alignments can result in significantly different phylogenetic trees. On the other hand, no solution has yet been provided as to how biologists should proceed to select the most useful aligners or alignments. The method we introduce here could conveniently address this issue by providing users with an objective criterion to select the MSAs that are most likely to be structurally correct. Such a criterion would also be useful in the context of homology modeling.

Funding: Plan Nacional BFU2008-00419 from the Spanish Ministry of Science; LEISHDRUG (no 223414) and Quantomics (KBBE-2A-222664) projects both financed by the 7th Framework Programme of the European Commission. Computational resources are provided by the Centre for Genomic Regulation (CRG) of Barcelona. Medical Research Council project U117581331 (to J.K.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Aniba,M.R. *et al.* (2010) . AlexSys: a knowledge-based expert system for multiple sequence alignment construction and analysis. *Nucleic Acids Res.*, **38**, 6338–6349.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bowie,J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Chandonia,J.M. *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

Claude,J.B. *et al.* (2004) CaspR: a web server for automated molecular replacement using homology modelling. *Nucleic Acids Res.*, **32**, W606–W609.

Connolly,M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.

Dayhoff,M.O. *et al.* (1979) A model of evolutionary change in proteins. Detecting distant relationships: computer methods and results. In Dayhoff,M.O. (ed) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 353–358.

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

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

Grasso,C. and Lee,C. (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, **20**, 1546–1556.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Jones,D.T. *et al.* (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

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

Lin,K. *et al.* (2003) Testing homology with Contact Accepted mutatiOn (CAO): a contact-based Markov model of protein evolution. *Comput. Biol. Chem.*, **27**, 93–102.

Lüthy,R. *et al.* (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Marin,A. *et al.* (2002) FROST: a filter-based fold recognition method. *Proteins*, **49**, 493–509.

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

Mizuguchi,K. *et al.* (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

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

O'Sullivan,O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.

Pei,J. *et al.* (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.

Shi,J. *et al.* (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.

Sierk,M.L. *et al.* (2010) Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinformatics*, **11**, 146.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins Struct. Funct. Genet.*, **17**, 355–362.

Taly,J.F. *et al.* (2008) Can molecular dynamics simulations help in discriminating correct from erroneous protein 3D models? *BMC Bioinformatics*, **9**, 6.

Thompson,J. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4690.

Thompson,J.D. *et al.* (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics*, **Chapter 2**, Unit 2 3.

Thompson,J.D. *et al.* (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.

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

Wallace,I.M. *et al.* (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.

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

Wu,S. and Zhang,Y. (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.

Yu,Y.K. *et al.* (2003) The compositional adjustment of amino acid substitution matrices. *Proc. Natl Acad. Sci. USA*, **100**, 15688–15693.

Zhang,Y. and Skolnick,J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl Acad. Sci. USA*, **101**, 7594–7599.