doi:10.1016/j.jmb.2004.04.058 *J. Mol. Biol.* (2004) **340**, 385–395

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments

Orla O'Sullivan[1], Karsten Suhre[2], Chantal Abergel[2]
Desmond G. Higgins[1] and Cédric Notredame[2,3]*

[1]*Conway Institute, University College Dublin, Belfield Dublin 4, Ireland*

[2]*Information Génomique et Structurale, CNRS UPR-2589 31, Chemin Joseph Aiguier 13402 Marseille, France*

[3]*Swiss Institute of Bioinformatics, Chemin des Boveresse, 155, 1066 Epalinges Switzerland*

**Corresponding author*

Most bioinformatics analyses require the assembly of a multiple sequence alignment. It has long been suspected that structural information can help to improve the quality of these alignments, yet the effect of combining sequences and structures has not been evaluated systematically. We developed 3DCoffee, a novel method for combining protein sequences and structures in order to generate high-quality multiple sequence alignments. 3DCoffee is based on TCoffee version 2.00, and uses a mixture of pairwise sequence alignments and pairwise structure comparison methods to generate multiple sequence alignments. We benchmarked 3DCoffee using a subset of HOMSTRAD, the collection of reference structural alignments. We found that combining TCoffee with the threading program Fugue makes it possible to improve the accuracy of our HOMSTRAD dataset by four percentage points when using one structure only per dataset. Using two structures yields an improvement of ten percentage points. The measures carried out on HOM39, a HOMSTRAD subset composed of distantly related sequences, show a linear correlation between multiple sequence alignment accuracy and the ratio of number of provided structure to total number of sequences. Our results suggest that in the case of distantly related sequences, a single structure may not be enough for computing an accurate multiple sequence alignment.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: multiple alignment; structural superposition; TCoffee; threading; sap

Introduction

It has long been assumed that using structural information can increase the accuracy of multiple protein sequence alignments (MSA).[1] Recent results[2,3] suggest that accurate MSAs obtained this way are useful for making functional assignments. These findings are quite exciting in a context where a structure may soon be available for each protein family (transmembrane proteins excepted).[4] However, making the best out of this wealth of data will require the development of new automatic methods, able to efficiently incorporate protein structure information within MSAs. The incentive for doing so is very strong, considering the critical role MSAs play in so many sequence analysis applications,[5] like phylogenetic reconstruction, structure prediction, functional characterization, database searches and non-synonymous single nucleotide polymorphism characterization.[6]

Despite their usefulness, accurate MSAs remain difficult to compute, owing to reasons that are both computational[7] and biological.[8] From a computational point of view, the assembly of an optimal MSA is a complex problem and an exact solution can be computed only for small sets of related sequences.[9] This is the reason why most packages use an approximate heuristic, the progressive alignment algorithm,[10] that gives no guarantee on delivering an optimal solution but can rapidly align large sets of sequences. On the biological side, one is limited by the lack of an objective and accurate criterion to assess MSA quality.[8] As a consequence, most methods use

Abbreviations used: MSA, multiple protein sequence alignment(s); S-MSA, structure-based MSA; DP, dynamic programming; NW, Needlman & Wunsch; CS, column score.

E-mail address of the corresponding author: cedric.notredame@europe.com

0022-2836/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.

sequence similarity (assessed with a substitution matrix) as a criterion for optimization. However, similarity is not informative enough to drive the correct alignment of distantly related sequences, a situation that typically requires using structure comparison methods so that a structure-based MSA (S-MSA) can be derived. S-MSAs constitute the *de facto* standard of truth for assessing sequence alignment accuracy and several established S-MSAs collections[11–13] are used routinely to evaluate MSA packages.[14–17] Although one may argue that these highly accurate MSAs (as judged from structural analysis) are not always optimal from an evolutionary point of view, they usually reflect well the structural and functional relationships between the considered proteins.

With 3DCoffee, we show that using a small amount of structural information when assembling an MSA makes it possible to improve alignment accuracy and emulate the computation of an S-MSA. Combining sequences and structures in this manner requires the integration of three types of methods: (i) sequence alignment methods; (ii) methods for comparing two or more structures and deduce a sequence alignment; (iii) methods for comparing sequences and structures, often referred to as threading.

Sequence–sequence comparison methods rely mostly on the dynamic programming (DP) algorithm to compute an alignment where gaps are disposed in such a manner that similarity is maximized between the two sequences.[18,19] Given a substitution matrix and a gap penalty scheme, DP can be used to compute global or local alignments[20,21] but accurate alignments can be obtained only with pairs of sequences that are at least 30% identical.[22] Structure–structure comparison has been approached using a wide variety of heuristics,[23,24] and to this day more than 30 algorithms have been reported. The simplest, like LSQman,[25] use rigid body superposition and let the algorithm look for an optimal superposition where intermolecular distances are minimized between superposed positions in the two structures. These methods perform well on similar structures where the 3D relationships of residues have been well preserved by evolution. These structures are usually encoded by closely related sequences. When dealing with more distantly related sequences, the residue equivalences can be worked out iteratively, as done in STAMP,[26] where the equivalences are used to drive a superposition that is used, in turn, to compute a distance matrix. The algorithm uses this updated matrix to refine the set of residue equivalences and make a new superposition. The process is carried out until it converges. SAP[27] uses a similar principal, although rather than being iterative, the algorithm computes the series of rigid superpositions associated with forcing the superposition of every possible pair of residues. The final alignment is computed by DP, using the summed distance matrices of all the superpositions considered. DALI produces alignments of comparable accuracy, computed by considering the local comparison of the distance maps associated with the considered structures.[28] Most of these methods make it possible to use structures for aligning sequences that are less than 30% identical. Although they diverge slightly in the alignment they produce, it is hard to establish which one (if any) performs better than the others.

Sequence–structure comparisons (or threading) can be achieved using two categories of methods.[29,30] One may use techniques inspired from molecular replacement to check whether a sequence is compatible with a 3D fold,[31] or sophisticated DP where the algorithm analyses the 3D-structure to determine local gap penalties and local substitution costs. Fugue is based on this principle and turns a structure into a position-specific substitution matrix, so that a sequence–structure alignment can be delivered using DP.[32]

Many of the structure-based alignment methods have been extended to generate S-MSAs. For instance, the double DP strategy of SAP has been coupled with a progressive algorithm to align more than two structures.[33] At least two other pairwise structural alignment methods have been incorporated in a progressive alignment strategy: STAMP and COMPARER. COMPARER[34] was used to assemble HOMSTRAD, the collection of multiple structural alignments used in this work for validation purposes. Other multiple structural alignment methods exist that use more specific procedures. For instance, DALI produces S-MSAs by aligning several structures to a master structure. One may use Fugue in a similar fashion by aligning several sequences to a single structural template. MNYFIT computes a consensus structure and uses it as a master to align all the others.[35] The lack of method-independent reference datasets makes it difficult to benchmark these packages accurately and establish their respective strength and weaknesses. Yet they all share a common drawback: they are all built around a specific pairwise alignment algorithm, making it difficult to combine the respective strengths of several algorithms into a single model. Furthermore, none of the available methods can seamlessly handle a mixture of sequences and structures, and when doing so, the most common strategy is to start aligning the structures into an S-MSA, before adding the sequences in a semi-manual fashion.[2]

We designed 3DCoffee to address this problem. 3DCoffee uses the TCoffee v2.00 MSA package. TCoffee computes MSAs using pre-compiled libraries of pairwise alignments. The libraries can be compiled using any method able to generate pairwise alignments, like threading and structure superposition. This makes the library a powerful means to incorporate structural information into the MSA assembly process. Using methods like SAP or Fugue, we studied the effect of compiling the library with a mixture of sequences and structures. Our methodology could easily be extended to incorporate methods that have not yet been considered so that biologists can integrate and combine their techniques of choice.

Principle of the 3DCoffee method

Computation of TCoffee multiple sequence alignments

We used TCoffee version 2.00 to compute non-structure-based MSAs (default mode), as well as S-MSAs. In its default mode, TCoffee does not use structures, it takes sequences as input and makes pairwise comparisons to compile a primary library. This primary library is a list of weighted pairs of residues.[36] A residue pair appears in the library when it has been observed in one of the pre-compiled pairwise alignments. The pairwise alignments compiled in the primary library can be computed using any method one finds suitable. By default, TCoffee computes for each pair of sequence a global pairwise alignment obtained with the Needlman & Wunsch (NW)[18] algorithm and the ten best-scoring local alignments as given by the SIM algorithm.[37] The weight associated with every residue pair obtained this way is set to the average percentage identity within the primary alignment (local or global). When two alignments contribute the same pair of aligned residues, the weights are added.

The weights within the primary library are then re-estimated according to the library self-consistency,[36] and the re-weighted library (named an extended library) is used as a position-specific substitution matrix to carry out a progressive multiple alignment.[38] Doing so involves computing a distance matrix by comparing every pair of sequences and using this matrix to compute a neighbor-joining guide tree.[39] The tree topology determines the order in which the sequences are incorporated within the MSA, using standard DP and the extended library as a position-specific substitution matrix.

Incorporation of structural information within the TCoffee library

Structural information is incorporated within the library by the means of structure-based pairwise sequence alignments. We used three methods, now fully integrated within TCoffee, providing the associated package is installed. Fugue is a threading method that aligns a protein sequence with a 3D-structure.[32] 3DCoffee directly submits sequence/structure pairs to the official Fugue server† and retrieves the resulting pairwise alignments that are integrated into the primary library using the standard TCoffee weighting scheme. SAP uses double DP to compute a pairwise alignment based on a non-rigid structure superposition.[27] When integrating these alignments within the primary library, we set to 100 the weight associated with each pair of aligned residues. This is the maximum weight an individual constraint can receive in a TCoffee primary library. Although this value is meant to reflect the high reliability of SAP, it only makes it more likely for these pairs to be aligned in the final MSA without explicitly forcing them to be so. Not forcing every pair of the structural alignments to find their way into the final alignment is important, as some portions of the SAP alignments correspond to non-superposable portions of the structures and are therefore unreliable. These segments usually have a low consistency within the primary library, and are therefore down-weighted at the extension stage. LSQman is a rigid body structure superposition package that makes structure-based sequence alignments.[40] When turning an LSQman structure superposition into a sequence alignment, we considered two residues to be aligned if they were less than 3 Å apart in the superposition. LSQman constraint weights are set to 100, like those of SAP and for similar reasons.

Producing multiple sequence structure alignments

We adapted TCoffee so that, given a collection of sequences and structures, one may specify which structures must be used and which methods should be applied on each possible pair. For instance, given a peptide file, 3DCoffee considers in turn every possible sequence pair within the dataset. For a given pair, the program computes a global alignment using NW and a series of local alignments using Lalign. If both sequences have an available structure, a pairwise alignment is computed using SAP and another one using LSQman. If one sequence only has a known structure, an alignment is made using the threading method Fugue. All these alignments are added to the TCoffee library using the standard procedure described above.

Benchmarking procedure

We used the February 2002 release of HOMSTRAD[11] (e) to design a benchmark strategy for 3DCoffee. HOMSTRAD is a hand-curated database of high-quality S-MSAs built around the multiple structure alignment package COMPARER. We selected within HOMSTRAD the most demanding alignments using two criteria: at least four sequences and less than 25% average identity within the MSA. This yields a collection of 43 MSAs, four of which had to be discarded (FAD-Oxidase_C, FAD-Oxidase_NC, TPR and bv) because they are impossible to align with any of the available methods and are therefore uninformative for the analysis. The 39 remaining MSAs (245 sequences) constitute our HOM39

† http://www.cryst.bioc.cam.ac.uk/~fugue/

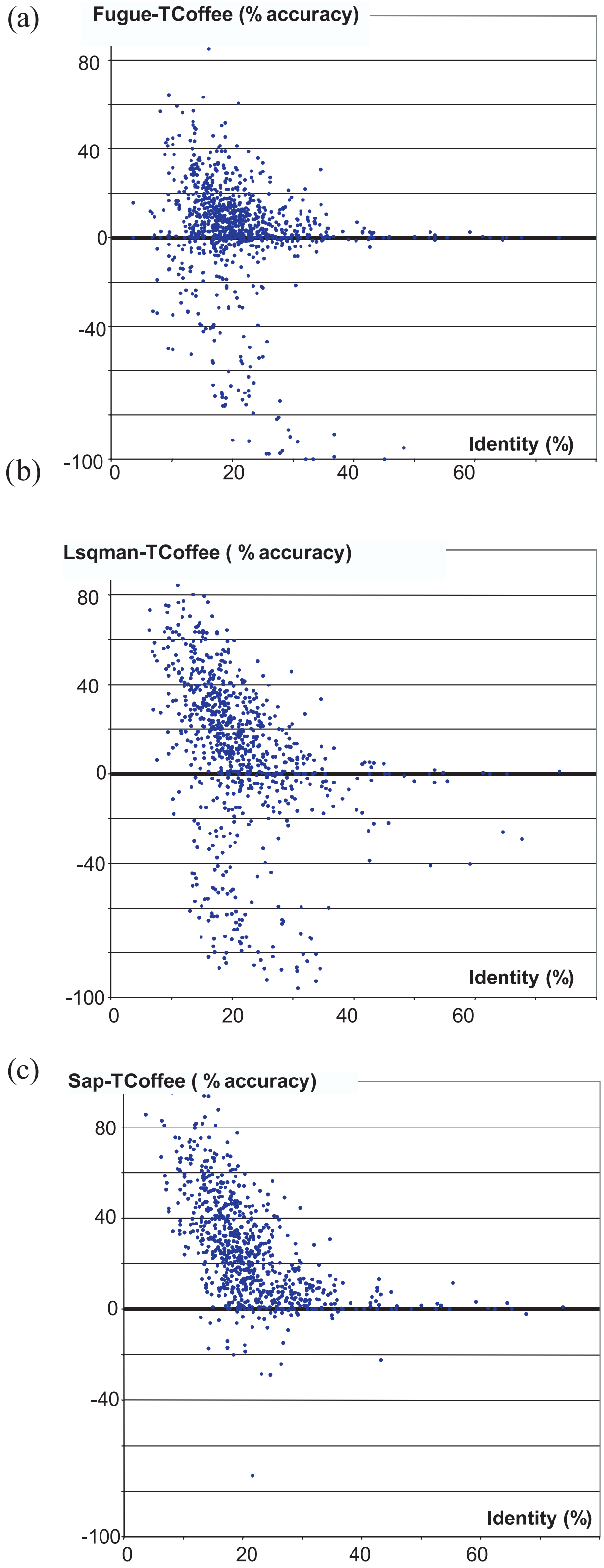

Figure 1. Performances of pairwise structure-based sequence alignment methods. Each dot corresponds to a parwise alignment taken from HOM39 (see method). The vertical axis represents the difference of alignment accuracy (Column Score) between TCoffee and (a) Fugue, (b) SAP and (c) LSQman. The horizontal axis shows the percent identity between the two sequences being considered, as measured on the reference HOM39 MSA.

dataset. It has the advantage of being both compact and discriminative.

We assessed the biological quality of our MSAs by comparing them with their HOM39 reference MSA, using the aln_compare package[36] that computes the column score (CS), which is a measure of the fraction of columns aligned identically between two MSAs.[41] We checked whether sequences without a known structure could benefit from being aligned with sequences whose structure is known. We named this measure the induced improvement, and measured it by removing the provided structure(s) from the reference and the target MSAs before comparing them.

System and packages

Academic licences (free of charge) to run TCoffee 2.00, SAP and LSQman were obtained for each package. These were installed on an SGI 02, running Irix 6.2. The protocols used here are now part of the TCoffee documentation.

Results

Improving MSA accuracy with a single structure

Single structures can be incorporated into an MSA only by using a threading method like Fugue. Before doing so, we evaluated the accuracy of Fugue as a pairwise method on the entire HOM39 dataset. Figure 1(a) shows a comparison between Fugue and TCoffee (TCoffee uses SIM and NW by default) where the relative performances of the two methods are assessed by comparison with the HOM39 reference. Fugue clearly outperforms TCoffee when making pairwise alignments. For instance, when comparing Fugue and TCoffee on all pairs of sequences from HOM39 (Figure 1(a)), we found Fugue to be three percentage points more accurate than TCoffee (61.8% accuracy for Fugue against 58.8% for TCoffee). The difference is significant with a *P*-value of 10^{-9} (Wilcoxon signed rank test).

We then computed each HOM39 MSA while providing TCoffee with one structure *via* the *-struc_to_use* flag. In each test case, we chose the most distantly related sequence (as judged with the average percentage identity in the HOM39 reference). The extent of identity between the selected structures and the rest of their MSA ranged between 12% and 24%. A new flavor of TCoffee (TC-Fugue) was designed, that uses three pairwise alignment methods: SIM, NW, and Fugue (Table 1A). We also used TCoffee associated with the Fugue method only (Fugue) as a control. This last procedure amounts to aligning the sequences one after the other onto the sequence with known structure, using the Fugue algorithm. Two other controls were set up using TCoffee in default mode and CLUSTAL W version 1.83 (CW183).

Table 1. Direct (A) and induced (B) improvement when providing one structure of the HOM39 datasets

Method	N str.	Avg. acc.	Difference with TCoffee	*P*-value (Wilcoxon signed-rank test)
A. *Direct improvement*				
TCoffee	0	42.24	–	–
CW-183	0	38.43	−3.8	2×10^{-2}
Fugue	1	31.26	−10.9	2×10^{-4}
TC-Fugue	**1**	**46.33**	**+4.1**	$\mathbf{1 \times 10^{-3}}$
B. *Induced improvement*				
TCoffee	0	52.83	–	–
CW-183	0	45.75	−7.1	1×10^{-3}
Fugue	1	35.53	−17.3	3×10^{-4}
TC-Fugue	**1**	**54.73**	**+1.9**	$\mathbf{4 \times 10^{-1}}$

Method indicates the method being used: TCoffee (TCoffee with NW and SIM), CW-183 (CLUSTAL W, 1.83), TC-Fugue (TCoffee with NW, SIM and Fugue), Fugue (TCoffee + Fugue, without NW or SIM). N str. indicates the number of structures provided. Avg. acc. indicates the average accuracy as measured with the CS score by comparison with the HOM39 reference alignments. *P*-value estimates the statistical significance of the observed difference between the considered method and the default TCoffee. The best performing method is in bold.

Our results (Table 1A) show that providing a structure to TC-Fugue improves MSAs by four percentage points over TCoffee (or by a litle less than eight percentage points over CLUSTAL W). The difference is significant with a *P*-value of 10^{-3}, and an observed improvement on 23 of the 31 alignments that are not tied between the two methods. We found (Figure 2(a)) that the amount of reported improvement depends loosely on the structure/sequence ratio, with high ratios yielding greater improvements. The low performances of the Fugue control are probably explained by the stringency of the CS measure that requires every sequence to be aligned correctly and is not well adapted to the pairwise-based alignment method used here.

We measured the induced improvement in the TC-Fugue alignments by removing the provided structure and found the average TC-Fugue accuracy to remain higher than that of TCoffee (Table 1B and Figure 2(b)), although in this case the difference is not statistically significant, as the observed difference is associated with a *P*-value of only 0.4. Note that the values in Table 1B are higher than the corresponding values in Table 1A because in Table 1B the evaluation is carried out while ignoring the provided structure (usually the hardest sequence to align).

Improving MSA accuracy with two structures

Using two structures offers the possibility of making structure–structure (SAP, LSQman) as well as structure–sequence comparisons. Before using these methods to compute an MSA, we evaluated their pairwise accuracy (Figure 1(b) and (c)). As expected, we found SAP and LSQman to outperform TCoffee significantly. A measure made on the SAP alignments of every HOM39 pair of sequence (Figure 1(b)) indicates an average accuracy of 86.3%. The difference with TCoffee is highly significant with a *P*-value of 10^{-11} (Wilcoxon signed rank test). Under the same conditions, LSQman outperforms TCoffee by 12 points with an average accuracy of 70.3%, and a difference also highly significant.

We computed every HOM39 MSA while providing TCoffee with two structures: the one used previously with TC-Fugue and its most distantly related homologue (lowest percentage identity) within the considered HOM39 MSA. An attempt to use the most informative pairs guided this choice. In order to judge the individual contribution of each of the three structure-based methods (Fugue, SAP and LSQman) to the overall accuracy of 3DCoffee, we used them separately, each time in conjunction with SIM and NW (Table 2A). These three new flavors of TCoffee are named TC-Fugue, TC-SAP and TC-LSQ, and the combination of all the available pairwise methods (Fugue, SAP, LSQman, SIM and NW) constitutes the new 3DCoffee method (TC-3D in the Tables).

As expected, TC-Fugue, TC-SAP and TC-LSQ all outperform TCoffee (Table 2A). Furthermore, TC-3D outperforms every alternative flavor and, given two structures, it produces MSAs on average ten percentage points better than TCoffee and 4.5 percentage points better than TC-Fugue (Table 2A). As indicated in Table 2A, all the differences reported between the new methods and TCoffee are statistically significant. Here as well, the extent of the improvement depends on the structure/sequence ratio (Figure 3(a)). Similar trends were observed when measuring the induced improvement (Figure 3(b)), which amounts to slightly less than 3.5 percentage points when comparing TC-3D with TCoffee (Table 2B). Although limited in amplitude, this improvement is also statistically significant.

Improving MSAs accuracy with many structures

We examined the effect of varying the structure/sequence ratio for every HOM39 MSA. We did so by applying TC-3D on each HOM39 dataset, using structural sets that contained between one and *N* structures (*N* being the total number of sequences).

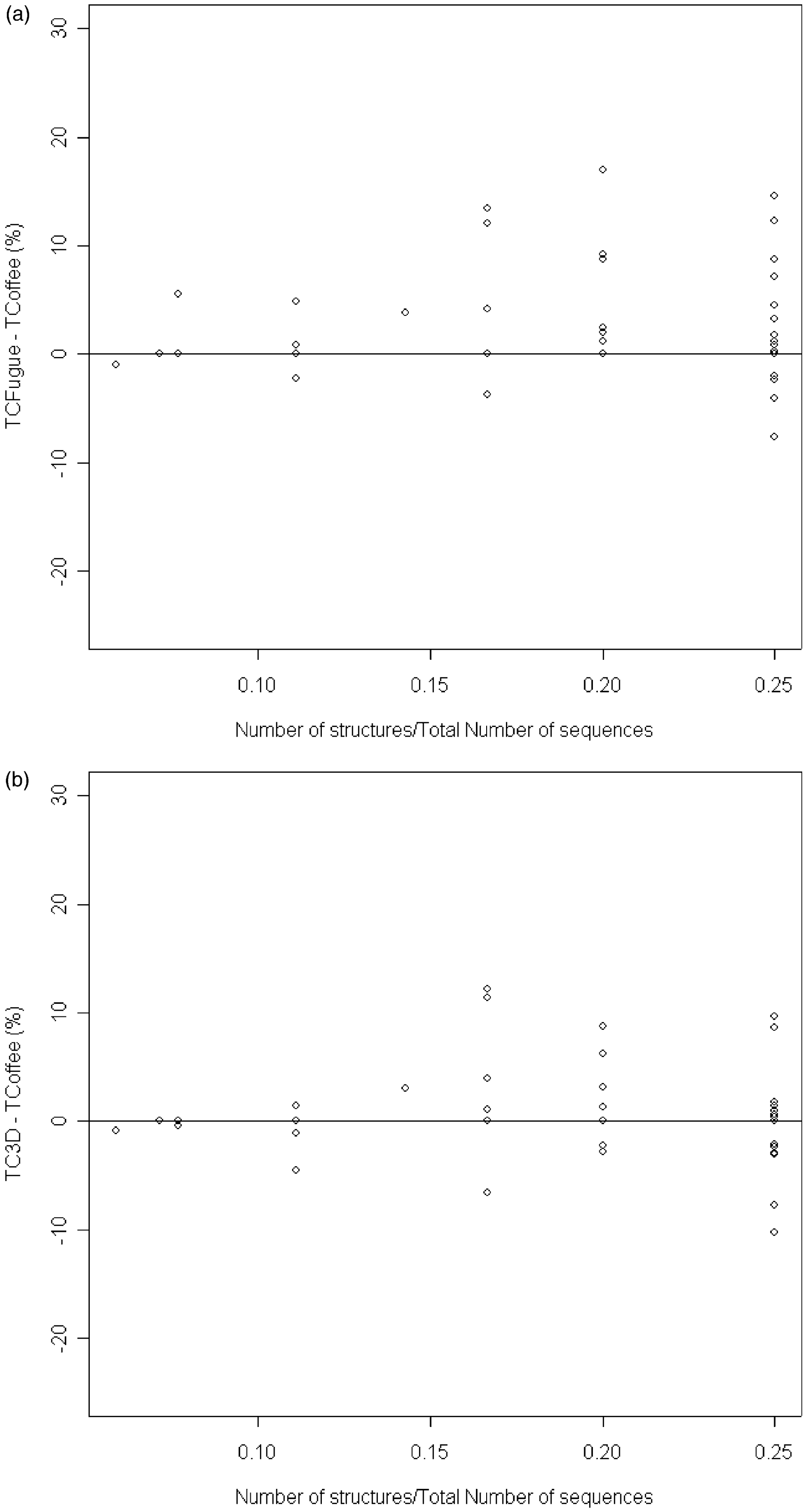

Figure 2. Comparative performances of TC-Fugue and TCoffee when using one structure. (a) Direct improvement. Each dot corresponds to an MSA taken from HOM39. The vertical axis indicates the difference of accuracy between a TC-Fugue and a TCoffee MSA. The horizontal axis indicates the ratio between the number of provided structures (1 structure) and the total number of sequences contained in the MSA. (b) Induced improvement. Similar to Figure 2(a), the MSA Accuracy is measured while ignoring the contribution of the provided structure.

The structural sets were assembled in an incremental manner. Given an MSA, one starts with the most distantly related structure (as shown above) before adding the structure of the less similar remaining sequences one by one, until *N* structural sets are defined for each HOM39 MSA. We then realigned every HOM39 MSA with each of its associated structural sets and compared the resulting alignments with the HOM39 reference. This makes a total of 200 MSA (between four and 15 for each HOM39 protein family) that were used to compute the data presented in Figure 4(a) and 161 for Figure 4(b).

The results are presented in the form of a

Table 2. Direct (A) and Induced (B) improvement when providing two structure of the HOM39 datasets

Method	N Str.	Avg. acc.	Difference with TCoffee	P-value Wilcoxon signed-rank test
A. *Direct improvement*				
TCoffee	0	42.24	0.0	1.0
CW-183	0	38.43	−3.8	2×10^{-2}
TC-Fugue	2	46.39	+4.0	5×10^{-3}
TC-SAP	2	50.81	+8.5	6×10^{-6}
TC-LSQ	2	47.26	+5.0	2×10^{-3}
TC-3D	**2**	**52.52**	**+ 10.3**	$\mathbf{1 \times 10^{-5}}$
B. *Induced improvement*				
TCoffee	0	56.12	0.0	1.0
CW-183	0	50.22	−5.9	1×10^{-1}
TC-Fugue	2	58.07	+1.9	2×10^{-1}
TC-SAP	2	58.49	+2.4	2×10^{-1}
TC-LSQ	2	57.52	+1.4	4×10^{-1}
TC-3D	**2**	**59.55**	**+ 3.4**	$\mathbf{2 \times 10^{-2}}$

Direct improvement is measured on the complete alignment, including the used structures. The induced improvement is measured only on the sequences whose structures were not used. Method indicates the method being used: TCoffee (TCoffee with SIM and NW), TCW-183 (CLUSTAL W version 1.83) TC-Fugue (TCoffee + NW + SIM + Fugue), TC-SAP (TCoffee + SIM + NW + SAP), TC-LSQ (TCoffee + SIM + NW + LSQman), TC-3D (TCoffee + SIM + NW + Fugue + SAP + LSQman). N str. indicates the number of structures provided. Avg. acc. indicates the average accuracy as measured with the CS score by comparison with the HOM39 reference alignments. P-value estimates the statistical significance of the difference between the considered method and TCoffee default using the Wilcoxon signed-rank test. The best performing method is in bold.

boxplot in Figure 4(a) (direct improvement) and Figure 4(b) (induced improvement). Figure 4(a) indicates the existence of a reasonable correlation between the structure/sequence ratio and the MSA accuracy, although the data are not distributed evenly. One gains roughly ten percentage points in accuracy with every 20 percentage points increase of the structure/sequence ratio. An individual analysis of each protein family suggests that this trend is consistent across most of the HOM39 dataset, although the phenomenon varies in amplitude. When using 3DCoffee and all the available structures in a procedure that amounts to assembling a multiple structural alignment, we obtained a score of 71.9% accuracy, a value short of the theoretical maximum of 100 that might have been expected if the unreliable regions of HOM39 had been removed from the evaluation. This value is an estimate of the correlation between the two-structure superposition method SAP and COMPARER rather than an estimate of accuracy. The induced improvement follows a similar trend, albeit more modestly (Figure 4(b)), and yields a gain of roughly two percentage points for every 20 percentage points of ratio increase. The distribution of this induced improvement is even less regular than that of the direct improvement. It indicates that in the HOM39 dataset, sequences benefit only modestly from the incorporation of the 3D information associated with one of their remote homologue.

Conclusion

3DCoffee is a novel method that takes advantage of structural information for aligning sequences. We benchmarked 3DCoffee using HOM39, a collection of high-quality reference S-MSAs. We used the TCoffee package to mix sequences, structures and structure/sequence alignment methods, and found this new protocol to improve MSA accuracy in a manner that depends on the structure/sequence ratio within the considered dataset. Our results suggest that using structures can improve the alignment accuracy of sequences without a known structure.

The 3DCoffee protocol bears several advantages. It is relatively fast: given all the pairwise alignments, it takes a few seconds to align ten sequences 200 residues long on a standard workstation. It is also very flexible and could easily be adapted to include any structure analysis method able to deliver a sequence alignment. We show here that one can effectively use this protocol to combine the output of methods based on different principles, like a rigid structure superposition method (LSQman) and a non-rigid one (SAP). This makes 3DCoffee a versatile tool that could easily be used in MSAs computation the way meta-methods are used in structure prediction.[42]

Yet, this study lends itself to a more paradoxical conclusion. Although structural information clearly helps improve MSA accuracy, it is surprising to find that its usage lacks the dramatic effect one may have expected. For instance, using one structure on a dataset of distantly related sequences increases the average accuracy by only an average four percentage points (and a maximum of ten). One may have hoped that the first or the first two structures would have delivered a larger share of the potential improvement. Yet this does not happen and every extra structure has about the same effect as the others on the overall accuracy, thus yielding a quasi-linear correlation between the structure/sequence ratio and the overall MSA accuracy.

This finding suggests that the standard methods

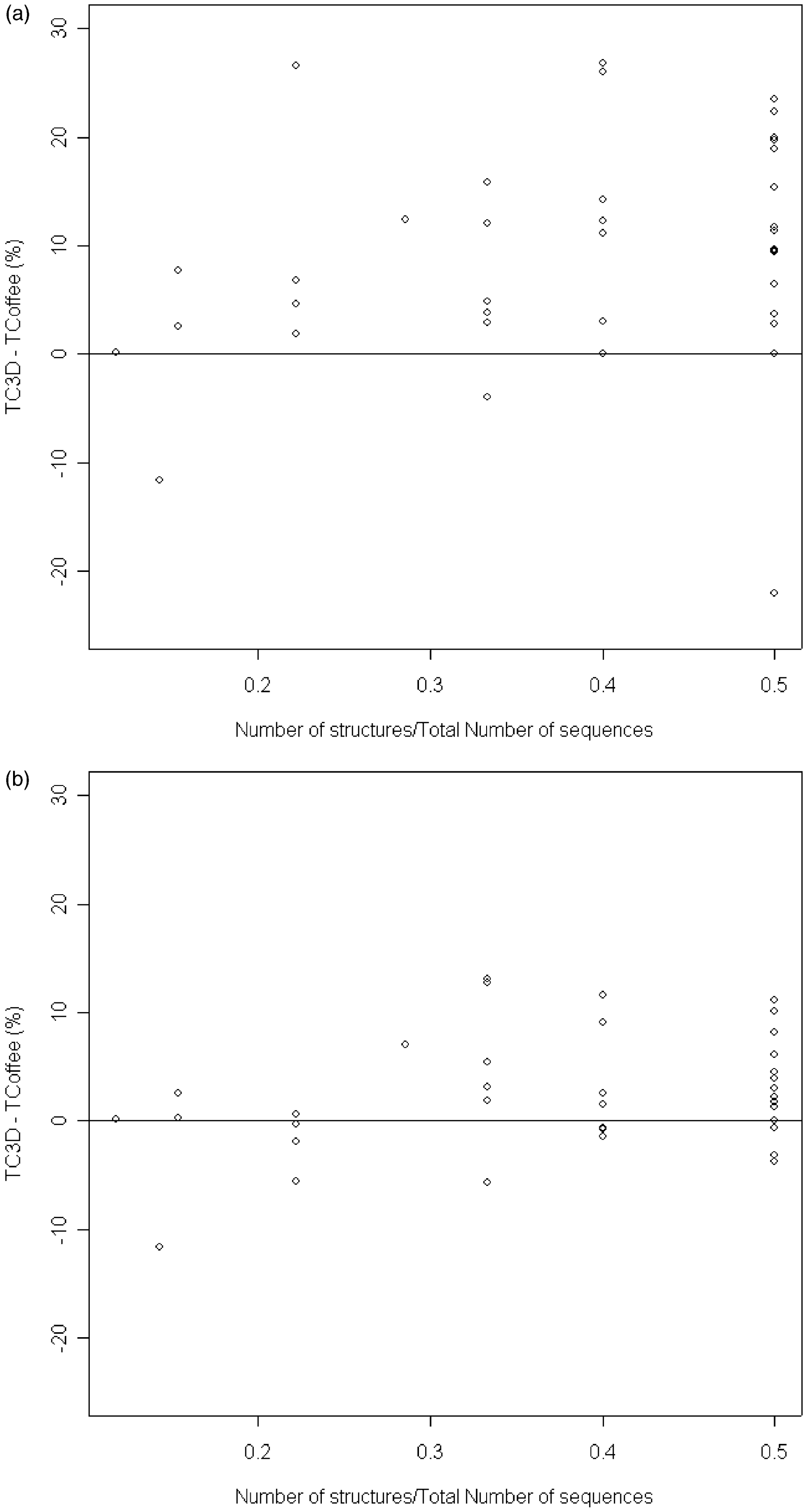

Figure 3. Comparative performances of TC-3D and TCoffee when using two structures. (a) Direct improvement. Each dot corresponds to an MSA taken from HOM39 (see method). The vertical axis indicates the difference in accuracy between a TC-3D and a TCoffee MSA. The horizontal axis indicates the ratio between the number of provided structures (2 structures) and the total number of sequences contained in the MSA. (b) Induced improvement. Similar to (a) with the MSA accuracy computed on the sequences without known structure.

we used here are not yet able to let the structural information diffuse optimally onto distantly related sequences. As a consequence, the best way to obtain a highly accurate MSA of remote homologues is to use more than one structure and, if possible, one structure for each sequence (or group of closely related sequences). On the basis of these results one may argue that given current methods, the "one structure for every protein family" strategy[43] may prove short of solving all the alignments problems faced by homology modeling. Achieving this purpose will require either better sequence comparison methods or more structures.

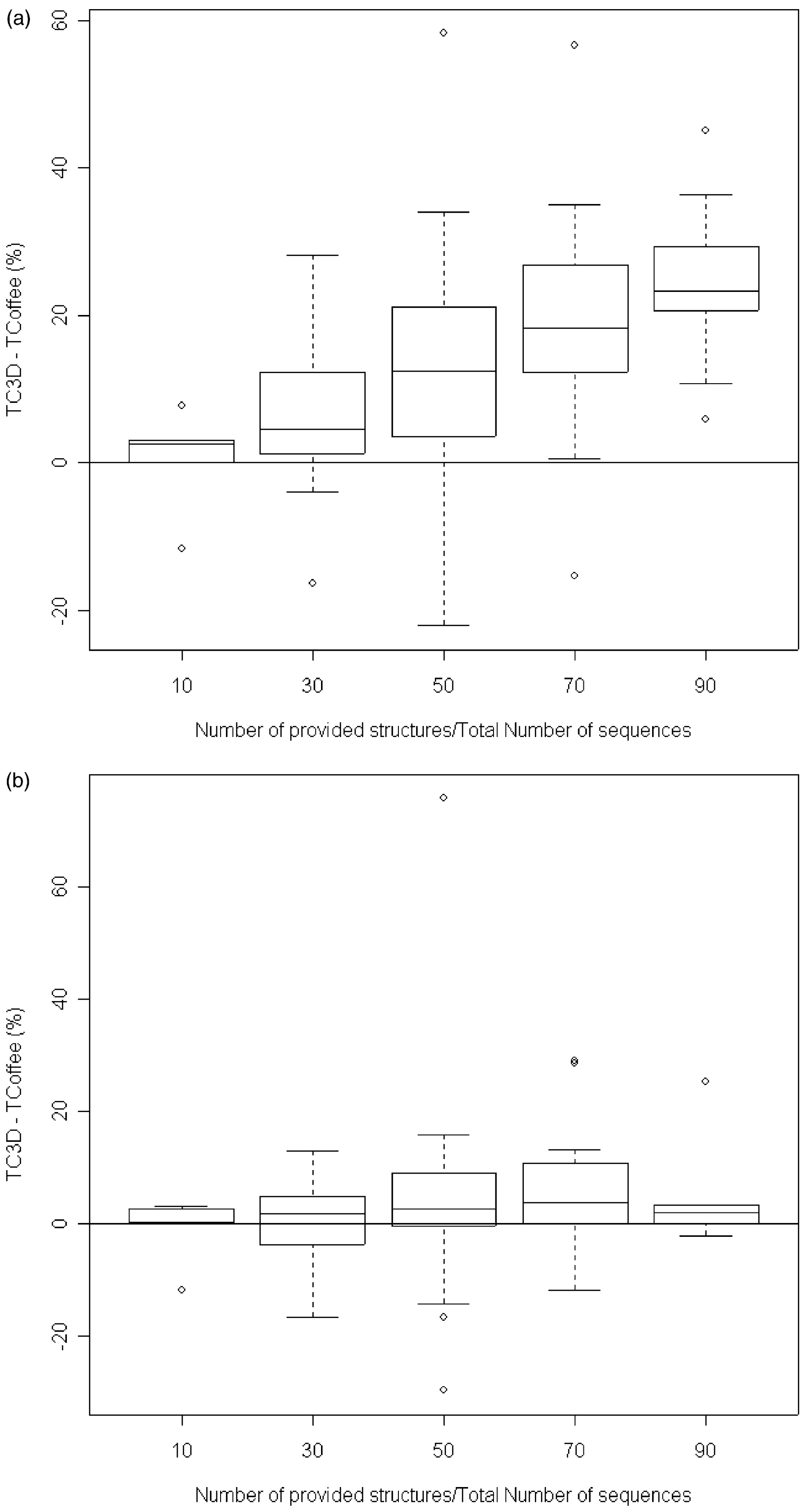

Figure 4. Alignment accuracy and structure/sequence ratio. (a) Each box indicates the average accuracy difference between TC-3D and TCoffee when computing HOM39 MSAs with various structure/sequence ratios: [0–20] (6 values), [21–40] (27 values), [41–60] (44 values), [61–80] (44 values), [81–100] (20 values). The vertical axis shows the average difference of accuracy and the horizontal axis the average structure/sequence ratio. The boxplot was generated with the R package using standard settings. Each box stretches from its lower hinge (defined as the 25th percentile) to its upper hinge (the 75th percentile). The median is shown as a line across the box. The top and the bottom whisker indicate the smallest data value larger then lower inner fence. The lower inner fence (not drawn) is equal to 1.5* spread to the 25th percentile. Values below the lower inner fence are plotted as a dot. The upper whisker is plotted in a similar fashion while using the 50th percentile as reference. (b) Induced improvement. Identical to 3b, with the measure of accuracy made on the sequences without known structure only.

Acknowledgements

Orla O'Sullivan was paid from Enterprise Ireland and Hewlett Packard provided some support. We thank Willie Taylor for helping us with setting up SAP, and Kenji Miziguchi for helping with Fugue. The comments of both referees were very helpful in improving the manuscript. We thank Jean-Michel Claverie for his many suggestions in improving and clarifying this manuscript.

References

1. Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270.
2. Al-Lazikani, B., Sheinerman, F. B. & Honig, B. (2001). Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl Acad. Sci. USA*, **98**, 14796–14801.
3. Marchler-Bauer, A., Panchenko, A. R., Ariel, N. & Bryant, S. H. (2002). Comparison of sequence and structure alignments for protein domains. *Proteins: Struct. Funct. Genet.* **48**, 439–446.
4. Brenner, S. E. (2001). A tour of structural genomics. *Nature Rev. Genet.* **2**, 801–809.
5. Duret, L. & Abdeddaim, S. (2000). Multiple alignment for structural, functional, or phylogenetic analyses of homologous sequences. In *Bioinformatics, Sequence, Structure and Databanks* (Higgins, D. & Taylor, W., eds), pp. 135–147 Oxford University Press, Oxford.
6. Ng, P. C. & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12**, 436–446.
7. Wang, L. & Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.* **1**, 337–348.
8. Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C. & Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.* **314**, 937–951.
9. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA*, **86**, 4412–4415.
10. Hogeweg, P. & Hesper, B. (1984). The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. *J. Mol. Evol.* **20**, 175–186.
11. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471.
12. 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*, **4**, 47.
13. Thompson, J. D., Plewniak, F. & Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
14. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.* **30**, 3059–3066.
15. Lassmann, T. & Sonnhammer, E. L. (2002). Quality assessment of multiple alignment programs. *FEBS Letters*, **529**, 126–130.
16. Lee, C., Grasso, C. & Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
17. Thompson, J. D., Plewniak, F. & Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucl. Acids Res.* **27**, 2682–2690.
18. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
19. Pearson, W. R. & Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol.* **210**, 575–601.
20. Huang, X., Hardison, R. C. & Miller, W. (1990). A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.* **6**, 373–381.
21. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
22. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
23. Eidhammer, I., Jonassen, I. & Taylor, W. R. (2000). Structure comparison and structure patterns. *J. Comput. Biol.* **7**, 685–716.
24. Sillitoe, I. & Orengo, C. (2002). Protein structure comparison. In *Bioinformatics: genes, proteins and computers* (Orengo, C., Jones, D. & Thornton, J., eds), pp. 250–265, BIOS Scientific Publisher, Oxford.
25. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A*, **34**, 827–828.
26. Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309–323.
27. Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.
28. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
29. Jones, D. T., Orengo, C. A. & Thornton, J. M. (1996). *Protein Folds and their Recognition From Sequence* Protein Structure Prediction (Sternberg, M. J. E., ed.), 1st edit., vol. 170, pp. 173–206, Oxford University Press, Oxford.
30. Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. & Elofsson, A. (2001). A study of quality measures for protein threading models. *BMC Bioinform.* **2**, 5.
31. Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.
32. Shi, J., Blundell, T. L. & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257.
33. Taylor, W. R., Flores, T. P. & Orengo, C. A. (1994). Multiple protein structure alignment. *Protein Sci.* **3**, 1858–1870.
34. Sali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. *J. Mol. Biol.* **212**, 403–428.
35. Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). Knowledge based modelling of homologous proteins. Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.
36. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
37. Huang, X. & Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. *Advan. Appl. Math.* **12**, 337–357.
38. Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W.: improving the sensitivity of progressive

multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4690.

39. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
40. Jones, T. A. & Kleywegt, G. J. (1999). CASP3 comparative modeling evaluation. *Proteins: Struct. Funct. Genet. Suppl.* **3**, 30–46.
41. Karplus, K. & Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**, 713–720.
42. Bourne, P. E. (2003). CASP and CAFASP experiments and their findings. *Methods Biochem. Anal.* **44**, 501–507.
43. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. *Nature Struct. Biol.* **8**, 559–566.

Edited by J. Thornton

(Received 14 November 2003; received in revised form 20 April 2004; accepted 22 April 2004)