


Structural bioinformatics

Protein multiple alignments: sequence-based versus structure-based programs

Mathilde Carpentier ^{1,*} and Jacques Chomilier²

¹Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, 75005 Paris, France and ²Sorbonne Université, MNHN, CNRS, IRD, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), BiBiP, 75005 Paris, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 4, 2018; revised on March 5, 2019; editorial decision on March 21, 2019; accepted on April 2, 2019

Abstract

Motivation: Multiple sequence alignment programs have proved to be very useful and have already been evaluated in the literature yet not alignment programs based on structure or both sequence and structure. In the present article we wish to evaluate the added value provided through considering structures.

Results: We compared the multiple alignments resulting from 25 programs either based on sequence, structure or both, to reference alignments deposited in five databases (BALIBASE 2 and 3, HOMSTRAD, OXBENCH and SISYPHUS). On the whole, the structure-based methods compute more reliable alignments than the sequence-based ones, and even than the sequence+structure-based programs whatever the databases. Two programs lead, MAMMOTH and MATRAS, nevertheless the performances of MUSTANG, MATT, 3DCOMB, TCOFFEE+TM_ALIGN and TCOFFEE+SAP are better for some alignments. The advantage of structure-based methods increases at low levels of sequence identity, or for residues in regular secondary structures or buried ones. Concerning gap management, sequence-based programs set less gaps than structure-based programs. Concerning the databases, the alignments of the manually built databases are more challenging for the programs.

Availability and implementation: All data and results presented in this study are available at: <http://www.abi.snv.jussieu.fr/people/mathilde/download/AlimulComp/>.

Contact: mathilde.carpentier@mnhn.fr

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Multiple alignments of protein sequences are an essential tool for exploring the evolution, diversity, conservation and function of proteins (Feng and Doolittle, 1987; Lecompte *et al.*, 2001; Levasseur *et al.*, 2008; Wong *et al.*, 2008). Despite the impressive and increasing number of available structures, most of these alignments are still computed by softwares that rely only on sequence information. Protein structures are mostly used as a second step in order to manually refine the alignment (Lemey *et al.*, 2009) or to guide a particularly difficult alignment of very divergent proteins (Jean *et al.*, 1997). Since it is usually admitted that structures are more conserved than sequences

(Illergård *et al.*, 2009) it is somehow surprising that multiple protein structure alignment methods, or methods combining sequence and structure, are not more widespread.

The goal of protein sequence alignments is to align homologous amino acids that derive from an ancestral sequence by substitutions. In structural alignments, the aligned positions are similar from the point of view of local and/or global conformations, and this structural similarity does not always imply homology (Godzik, 1996). Indeed, similar sub-domain fragments can be found in many different folds, with unrelated functions or various origins (Alva *et al.*, 2015; Lamarine *et al.*, 2001; Nepomnyachiy *et al.*, 2017). The conceptual model behind sequence alignment explicitly considers three events for

evolution: insertion, deletion and mutation. The model behind structure alignment is not so clear, partly because the impact of those three events on the folding step of protein structures is not well understood. The design of such a model is one of the greatest challenges of our decade for structural biology (Liberles *et al.*, 2012).

Homology is difficult to assess, especially when the proteins show a low level of similarity or if the homology of the whole genes is questionable. Due to all the considerations above, it is difficult to claim that structure alignments provide a golden standard for evaluating the quality of sequence alignment. However, as structures are better conserved, alignments should be more reliable when information from sequences and structures are combined. We compared in this article the alignments computed from structure or both structure and sequence with those from sequence only.

Multiple sequence alignment methods have been compared in many articles and with several types of benchmarks reviewed in Iantorno *et al.* (2014). The most widely used benchmarks are composed of a collection of reference alignments considered as the gold standard. The reference alignments are constructed mainly by using the sequence and structural information, but also according to other information as the function (Thompson *et al.*, 2011). Other types of benchmarks rely on simulated sequences (Nuin *et al.*, 2006), on direct comparison of all computed alignments, without any reference alignment (Landan and Graur, 2007; Lassmann and Sonnhammer, 2005) or on the validity of phylogenetic trees computed from the alignments (Dessimoz and Gil, 2010).

For structure-based alignment methods, less comparative studies have been conducted and most of them compare pairwise structural alignment programs (Feng and Sippl, 1996; Gerstein and Levitt, 1998; Godzik, 1996; Kim and Lee, 2007; Mayr *et al.*, 2007; Sauder *et al.*, 2000; Slater *et al.*, 2013). Multiple structural alignment programs are compared in the study of Berbalk *et al.* (2009). The authors noticed that structure-based alignment programs were generally very difficult to use and that there is room for improvements concerning use and applicability. They concluded that combining different alignment approaches into a single program relying on an automated scoring could improve the alignment quality but that until such a method is implemented, it seems important for a user to apply different tools and to manually compare their results.

We have conducted here a thorough comparative study of the performances of sequence-based and structure-based programs in order to address the following questions: are structure-based methods really superior in order to retrieve homologous residues? Or is it the sequence and structure ones? In which cases should we use structure-based methods, sequence+structure-based methods or sequence-based methods?

2 Materials and methods

2.1 Databases

In this study, we used reference multiple alignments built from sequences, structures and function information, and considered them as the gold standard. We did not use the three other types of benchmarks mentioned above because: (i) the use of simulated sequences is not possible in our case because there is no associated structure; (ii) it is possible to compare all alignments without a reference but, as programs may be consistent with each other but all wrong, we decided to avoid this approach in this article; (iii) the phylogeny-based approach would be very interesting but it requires a database of validated trees which is beyond the scope of the article.

We have selected 847 alignments, containing at least three protein chains or domains, from five reference multiple alignment databases: BALIBASE 2 (Thompson *et al.*, 1999b), BALIBASE 3 (Thompson

et al., 2005), HOMSTRAD (Mizuguchi *et al.*, 1998a), OXBENCH (Raghava *et al.*, 2003) and SISYPHUS (Andreeva *et al.*, 2007). We restricted the databases to proteins present in the protein data bank that represent only the structured domains of protein sequences, thus discarding intrinsically disordered proteins. This restriction is necessary when using structure alignment methods. Some regions may be disordered in resolved protein structure but their proportion is low (1% of the residues in human protein-coding genes), whereas the proportion of these regions predicted in proteins of unknown structure is 20% (van der Lee *et al.*, 2014). Some other alignments have been discarded: those with two or more proteins with identical amino acid sequence, those with missing residues in structures or with various inconsistencies. We did not consider the alignments of other well-known databases listed in Blackshields *et al.* (2006) for various reasons: PREFAB (Edgar, 2004) because it is composed of pairwise alignments; IRMBASE (Subramanian *et al.*, 2005) because there is no structure associated to the simulated fragments and SABMARK (Van Walle *et al.*, 2005) because of some inconsistencies in the multiple alignments which are built from pairwise structural alignments, pointed by the author and in Edgar (2010). We also had difficulties accessing PALI (Balaji *et al.*, 2001) and could not download the database. For all the databases, we only consider the core of the alignments but its definition depends on the database.

We have selected 29 families from BALIBASE 2 (BB2) and 38 from BALIBASE 3 (BB3), manually curated by checking the alignments of functional and other conserved residues. In each family, all proteins share the same structural fold, so the core can be reliably defined, excluding ambiguous or non-superimposable regions, unrelated secondary structure borders or some loop regions. BB2 and BB3 were kept even if they are from the same source because the protein families are different between BB2 and BB3. HOMSTRAD, from which we selected 357 families, is exclusively based on proteins with known structures, and each family is aligned with the programs MNYFIT (Sutcliffe *et al.*, 1987), STAMP (Russell and Barton, 1992) and COMPARE (Sali and Blundell, 1990). These structure-based alignments are annotated with JOY (Mizuguchi *et al.*, 1998b) and individually examined and modified if necessary. JOY produces core blocks annotations defined as the regular secondary structure elements (SSEs). We retrieved from OXBENCH 330 alignments with three or more proteins in each alignment (subset 'multi'), not split in domains (full-length sequences). These multiple alignments are computed by STAMP (Russell and Barton, 1992). All the aligned positions were taken as the core blocks. The last database, SISYPHUS, is based on the families of domains from the structural classification SCOP (Murzin *et al.*, 1995) with non-trivial structural relationships. Multiple alignments are manually constructed for structural regions that range from oligomeric biological units, or individual domains to fragments of different sizes and are manually curated. SISYPHUS annotates the structurally equivalent residues in the alignments and we consider them as the core blocks.

Many structure-based programs do not output all the residues of input protein structures (some residues are removed or ignored) or change the name of the sequences. We have developed two programs for solving this issue: the first matches the protein names in the reference alignments and the protein names in the program-calculated alignments and the second makes each sequence of a program-calculated alignment identical to the sequence in the reference alignment. The residues removed by some structure-based programs are inserted in the alignment and the rest of the column is filled with gaps.

2.2 Alignment quality evaluation

The alignments produced by each program are evaluated by comparison with the reference alignments through two scores, following

Thomson *et al.* (1999a): (i) the fraction of pairs of residues in the reference alignment correctly identified by a given method, known as the sum-of-pairs (SP) score; (ii) the column score (CS) that describes the fraction of reference columns identified. As usually done in alignment method comparisons (Do *et al.*, 2005; Golubchik *et al.*, 2007; Thompson *et al.*, 1999a), Friedman tests (Friedman, 1937) were performed. This test is more conservative than the Wilcoxon test that assumes a symmetrical difference, and this is not always the case. All tests, plots and heatmaps have been done with R (R Core Team, 2017). The average multiple root mean square (RMS) have been computed with THESEUS (Theobald and Wuttke, 2006) that has been applied to all alignments, reference ones or computed by the tested programs. We have counted the number of gaps in all columns between the first and last core elements. We present in the article only the proportion of columns containing one or more gap opening. Accessible surface area (ASA) is calculated with NACCESS for all the proteins, in order to split the amino acids in two classes: either buried (relative ASA <25%) or exposed (Petersen *et al.*, 2009). Secondary structure assignments have been performed with STRIDE (Frishman and Argos, 1995). The six classes given in the output of STRIDE are back coded in three classes: helices, strands and coils. All analyses have been led according to the following characteristics: the residues have been assigned either as buried or accessible, and either in helix, strand or other (loop).

2.3 Programs

We have three categories of multiple alignment programs: sequence-based, sequence+structure-based and structure-based. To be included in this study a program must: (i) be available for download, (ii) output a file containing the sequence alignment, (iii) run without error. Each multiple alignment had to be computed in <2 h, otherwise the job was canceled. The execution time has been measured for the alignments of the SISYPHUS database on a standard desk computer with an i7 processor (Table 2). Some programs failed to produce enough alignments to allow a significant analysis of their performance and they were excluded if they produced an alignment for <70% of the dataset. As we mainly aim at addressing the performance of structure-based or sequence+structure-based alignment methods, we tried to be as exhaustive as possible for them. We searched or tested more than 40 programs but many were unavailable or did not conform to our criteria. We were also surprised by the low number of sequence+structure-based alignment methods. We did not include methods improving alignments afterwards, like STACCATO (Shatsky *et al.*, 2005). There is a great number of sequence-based programs and we only tested the most popular ones according to the last studies (Le *et al.*, 2017; Thompson *et al.*, 2011). All the programs included in our study are listed with a short description in Table 1. We have selected 9 sequence-based programs, 5 sequence+structure-based programs, (TCOFFEE/3DCOFFEE is either run with SAP or TM-ALIGN) and 11 structure-based programs.

3 Results

3.1 Number of computed alignments

All programs have been run on the 847 alignments. All sequence-based programs calculated all the 847 alignments but some programs of the two other categories failed for some alignments (Table 2). Sequence-based programs, MATRAS and TCOFFEE_TM successfully computed all alignments but not the other programs. Sometimes failures were due to the time limit, but most of the time

they were due to errors returned by the programs. MAMMOTH encountered the most failures; it has obviously a limit of 25 proteins per alignment. In order to improve the robustness of our analysis, we decided to restrict our analysis to the alignments computed by all programs, resulting in 535 alignments: 24 from BB2, 24 from BB3, 287 from HOMSTRAD, 158 from OXBENCH and 42 from SISYPHUS. These 535 alignments involve more than 2000 different protein chains.

3.2 Databases

The distribution of mean pairwise sequence identity among the 535 core alignments is given in Figure 1. BB2, BB3 and SISYPHUS databases are more focused on low identity, while HOMSTRAD and OXBENCH present alignments of high level of identity. The proportion of amino acids included in regular secondary structures in the complete dataset is 60%; restricted to the core alignments, it increases to 79%. We checked the redundancy of the databases. The number and proportion of chains included in two databases are listed in Supplementary Table S1. There is some overlap between BB2 and BB3: 48 chains are present in both BB2 and BB3. However, the protein families are all different between BB2 and BB3 so we decided to keep them all. The overlaps are very weak for the other databases.

3.3 Global analysis of alignment scores

The boxplot distribution of SP and CS scores of each program run on the 535 alignments are presented in Figure 2. The exact median values are reported in Supplementary Table S2. Globally, the results are impressively good: the SP score medians range from 0.86 to 0.97, meaning that in half of the alignments, more than 86% of the residue pairs are correctly aligned by any method. Similarly, in half of the alignments, more than 81% of the alignment columns are correct. Scores vary with the programs and structure-based programs give better results on the whole, except for MULTIPROT and MISTRAL. The sorting is the same for SP and CS scores except for FORMATT, MULTIPROT and MISTRAL that have a better CS score, and STAMP and KPAX that swap ranks. STAMP shows the greatest variability in its results, and it is not the best despite the fact that it has been used for building the alignments of two databases (HOMSTRAD and OXBENCH). FORMATT, a modified version of MATT that includes sequence information, is worse than MATT. It highlights the difficulty of combining sequence and structure information, which is nevertheless possible: TCOFFEE_TM is the best sequence+structure-based program, and achieves better than TCOFFEE_SEQ. However, sequence+structure-based methods do not perform better than structure only methods, despite the use of both sequence and structure information.

For each pair of programs, the significance of their differences has been evaluated by a Friedman rank test on their scores calculated for all 535 alignments (Section 2). In Figure 3, the programs are ranked according to their median CS score, and six groups of programs without significant differences within a group appear (black squares). The differences are significant between the programs outside the groups in most cases. The first group contains MAMMOTH and MATRAS that are the two best performing programs according to our study. The second group gathers MATT, TCOFFEE_TM, 3DCOMB, MUSTANG and TCOFFEE_SAP and their results are close to the two first programs. MUSTANG and TCOFFEE_TM are not significantly different from MAMMOTH and MATRAS despite their lower ranking. The three last groups contain all sequence-based programs and also FORMATT, MULTIPROT and MISTRAL. TCOFFEE_SEQ and PROBCONS

Table 1. Programs used in this study to align families of proteins from the reference datasets

Type	Name	Description	Rigid super-imposition	Version	References	Year
SEQ	CLUSTALO	Seeded guide trees and HMM profile–profile	NA	1.2.0	(Goujon <i>et al.</i> , 2010; Sievers <i>et al.</i> , 2011)	2010
SEQ	CLUSTALW	Classical progressive aligner	NA	2.1	(Larkin <i>et al.</i> , 2007; Thompson <i>et al.</i> , 1994)	1994
SEQ	DIALIGN	Greedy and progressive approaches for segment-based multiple alignment	NA	TX, 1.0.2	(Al Ait <i>et al.</i> , 2013; Morgenstern, 1999; Morgenstern <i>et al.</i> , 1998)	1998
SEQ	KALIGN2	Wu–Manber string-matching algorithm, improving both accuracy and speed	NA	2.04	(Lassmann <i>et al.</i> , 2009; Lassmann and Sonnhammer, 2005)	2005
SEQ	MAFFT_linsi	Fast progressive aligner with iteration and refinement using consistency score	NA	7.215	(Katoh <i>et al.</i> , 2002; Katoh and Standley, 2013)	2002
SEQ	MUSCLE	Fast progressive aligner with iteration and refinement	NA	3.8.31	(Edgar, 2004, 2004)	2004
SEQ	PRANK	Phylogeny-aware progressive aligner; correcting treatment of insertions	NA	v.100701	(Löytynoja and Goldman, 2005)	2005
SEQ	PROBCONS	Probabilistic variant of the consistency algorithm	NA	1.12	(Do <i>et al.</i> , 2005)	2005
SEQ	TCOFFEE_SEQ	Consistency-based progressive aligner	NA	11.00.8cbe486	(Notredame <i>et al.</i> , 2000)	2000
SEQ/STRUCT	PROMALS3D	Derives constraints through structure-based alignments; combines them with sequence constraints when constructing consistency-based multiple sequence alignments	No	NA	(Pei <i>et al.</i> , 2008; Pei and Grishin, 2007)	2008
SEQ/STRUCT	TCOFFEE_SAP	TCOFFEE + pairwise structure alignments by SAP	Yes	11.00.8cbe486	(O’Sullivan <i>et al.</i> , 2004; Orengo and Taylor, 1996)	2004
SEQ/STRUCT	TCOFFEE_TM	TCOFFEE + pairwise structure alignments by TM-ALIGN	Yes	11.00.8cbe486	(O’Sullivan <i>et al.</i> , 2004; Zhang and Skolnick, 2005)	2004
SEQ/STRUCT	SALIGN	DP with a score that is a sum of an affine gap penalty and of terms depending on various sequence and structure features	Yes	Modeler version: 9.18	(Madhusudhan <i>et al.</i> , 2009)	2007
SEQ/STRUCT	FORMAT	MATT with sequence information	No	1.02	(Daniels <i>et al.</i> , 2012)	2005
STRUCT	3DCOMB	Identifies structurally similar pairwise fragments and assemblies according to pivot structures	Yes	1.06	(Wang <i>et al.</i> , 2011)	2011
STRUCT	GESAMT	Score: TM-score (Zhang and Skolnick, 2004) Clustering of small structurally similar pairwise fragments	Yes	7.0	(Krissinel, 2012; Winn <i>et al.</i> , 2011)	2012
STRUCT	KPAX	Score: Q-score (Krissinel and Henrick, 2004) DP + alignment optimization	Yes	5.0.5	(Ritchie <i>et al.</i> , 2012)	2005
STRUCT	MAMMOTH	Score: Gaussian structural similarity score AFPs alignment by DP. Progressive multiple alignment with a guide tree	No	NA	(Lupyan <i>et al.</i> , 2005)	2005
STRUCT	MATRAS	Score: probability of residue random match of two different folds (Ortiz <i>et al.</i> , 2002) Progressive multiple alignment (guide tree) by DP	No	1.2	(Kawabata, 2003; Kawabata and Nishikawa, 2000)	2000
STRUCT	MATT	Score: PAM like matrices computed on SSE conservation or C α internal distances AFPs chaining by DP	Yes	1.0	(Menke <i>et al.</i> , 2008)	2008
STRUCT	MISTRAL	Score: based on RMS for AFP and on a geometrical transformations to allowing flexibility for chaining Superposition by minimizing interaction energy and residue one-to-one correspondence afterwards	Yes	3.6	(Micheletti and Orland, 2009)	2009
STRUCT	MTMALIGN	Score: interaction energy and RMS Progressive multiple alignment (guide tree) by DP	Yes	20171124	(Dong <i>et al.</i> , 2018)	2017
STRUCT	MULTIPROT	Score: TM-score With each structure as a pivot, detection of all AFPs, assembling to build the longest consistent alignment Score: alignment length, consistency and RMS	Yes	1.93	(Shatsky <i>et al.</i> , 2004)	2004

(continued)

Table 1. Continued

Type	Name	Description	Rigid super-imposition	Version	References	Year
STRUCT	MUSTANG	AFP and progressive multiple alignment with a tree. Score: C α internal distance [DALI like, (Holm and Sander, 1993)]	No	3.2.3	(Konagurthu et al., 2006)	2005
STRUCT	STAMP	Iterative superposition and alignment of C α by DP with a guide tree Score: C α distances and conformational similarity	Yes	4.4	(Russell and Barton, 1992)	1992

Note: Categories of programs: SEQ is a sequence-based alignment method; STRUCT is a structure-based alignment method; SEQ/STRUCT is a sequence+structure-based program. DP, dynamic programming; AFP, aligned fragment pairs.

Table 2. Number of computed alignments from structure-based or sequence+structure-based methods

	Alignments		Average time
MATRAS	847	100.0%	<10 s
TCOFFEE_TM	847	100.0%	<1 min
KPAX	846	99.9%	<10 s
PROMALS3D	846	99.9%	<10 min
TCOFFEE_SAP	845	99.8%	<10 s
MTMALIGN	845	99.8%	<10 s
FORMATT	844	99.6%	<1 min
GESAMT	841	99.3%	<1 s
MUSTANG	840	99.2%	<10 min
MISTRAL	828	97.8%	<10 min
STAMP	826	97.5%	<1 s
MATT	824	97.3%	<10 min
3DCOMB	822	97.0%	<10 s
SALIGN	796	94.0%	<10 min
MULTIPROT	766	90.4%	<1 min
MAMMOTH	622	73.4%	<10 s
#Alignments	847		

Note: The average computation time has been measured for the 42 SISYPHUS families that all programs successfully aligned. All sequence-based methods compute the alignments in less than a second on average except PRANK (time < 1 min). KALIGN2, CLUSTALO, CLUSTALW and MUSCLE are the fastest (<0.1 s).

are the two best sequence-based programs. STAMP, FORMATT, MULTIPROT and MISTRAL performances are not significantly different from the performances of programs with a lower ranking.

We also performed a hierarchical clustering on the basis of the scores of the various programs and the various alignments. A heatmap of this clustering is presented in Figure 4 for CS scores and in Supplementary Figure S1 for SP scores. The results are extremely similar regardless of the score (CS or TC). Considering program clustering (left tree Fig. 4), all sequence+structure-based programs and structure-based programs except STAMP are in the same subtree. All sequence-based ones are also pooled together. We have three groups of programs in the upper sub-tree (see the pink dashed line). TCOFFEE_SAP is alone on its branch; its score profile is different from the others: it sometimes fails when others succeed (see the red scores at the right extremity of its profile). The second group is composed of MUSTANG, MAMMOTH, MAT, FORMATT and MATRAS, whose performances are almost undistinguishable according to the Friedman tests. In the third group, 3DCOMB, MTMALIGN, KPAX and GESAMT have very similar profiles; they are also pooled with MULTIPROT and MISTRAL that are more

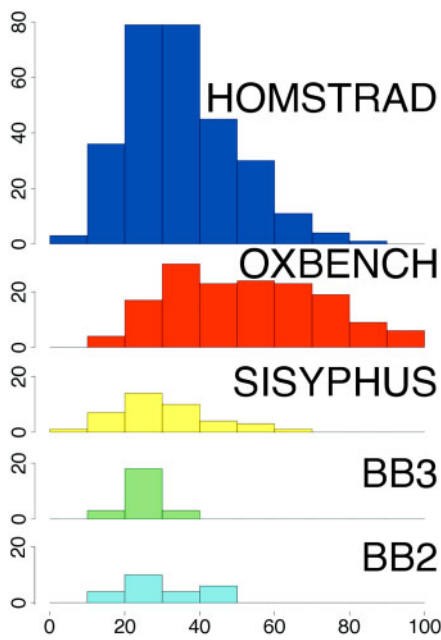


Fig. 1. Distribution of core block sequence identity percentages among the five databases. X-axis: identity percentage, y-axis: number of reference alignments

designed to find conserved structural blocks than to align whole proteins. This splitting of the structure-based programs into two clusters is consistent with the performance of rigid superimposition at some step in the methods, except for MATT but it explicitly compensates for the rigidity by introducing flexibility in its score. The profiles of sequence-based programs are very similar to each other.

We analyzed the most difficult alignments. Nine alignments show CS scores below 0.5 for all the programs (Supplementary Table S3 and Figure S7). The sequence identities of the core reference alignments are low (21% on average). The structural challenges of these alignments are: large insertions or deletions for some proteins of the families (five alignments), structural repetitions (one alignment) or large alignments with strong structural variations (three alignments) examined the difficult alignments for structure-based programs. There are 78 alignments where structure-based programs do not have the highest CS score. For 66 of them, the difference between the maximum CS score of all programs and structure-based programs is <0.1. The remaining 12 alignments are listed in Supplementary Table S4 and Figure S8. The sequence identity is globally higher (31% on average) and the RMS calculated from the reference alignment is high

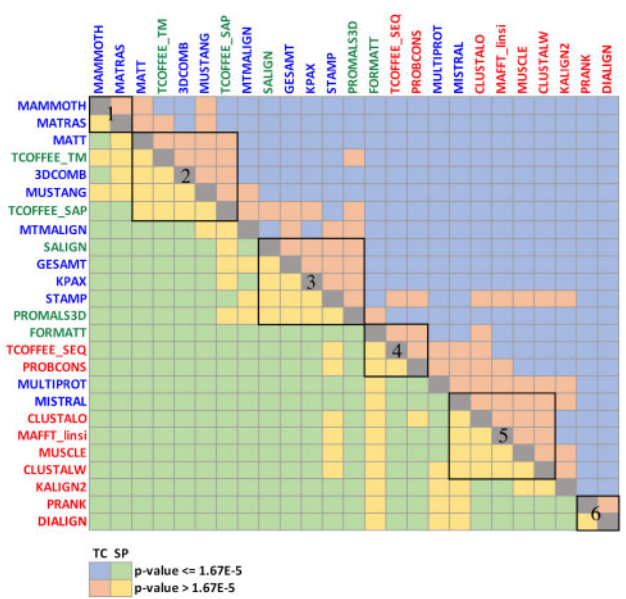
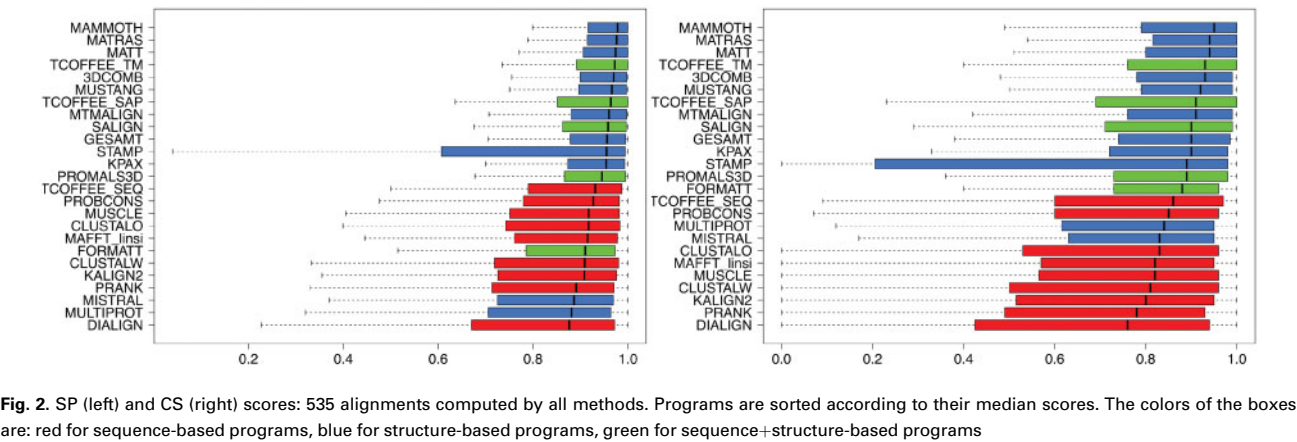


Fig. 3. *P*-value heatmap of the Friedman tests. Entries show the *P*-values computed using a Friedman rank test. Values above the diagonal were calculated with CS scores and values under the diagonal were calculated with SP scores. The programs are ordered according to their median CS scores. The colors of the program names are the same as in Figure 2. The yellow or orange cells denote a non-significant *P*-value according to the 0.05 alpha risk, with Bonferroni correction for multiple tests. The green or blue cells mean a significant *P*-value

(>4Å) for all but two families that include structural repetition (Leucine rich repeats and many beta strands).

3.4 The effect of sequence identity

We have investigated the effect of sequence conservation on the quality of the alignments computed by the different programs. The results are presented in Figure 5 for CS scores and in Supplementary Figure S2 for SP scores. As expected, the differences between structure-based and sequence-based methods are stronger for alignments of very divergent proteins. For alignments above 50% of sequence identity, sequence-based programs have similar or even better performances than structure-based programs. We also checked the effect of the number of proteins to align. The effect is very weak in the case of SP scores for all programs except MULTIPROT (Supplementary Figure S3) but it is noticeable on the CS scores (Supplementary Figure S4).

3.5 The effect of structural variations

We have measured the structural divergence by computing the RMS from a superimposition built according to the reference alignments. The performances of the programs as a function of these RMS are presented in Figure 6 for the CS scores and in Supplementary Figure S5 for SP scores. We have split our dataset in alignments below 30% of sequence identity (left, 199 alignments) and above (right, 335 alignments). There is no alignment below 30% with an average RMS below 1 Å. For the alignments below 30% of sequence identity, the scores of all programs globally decrease while RMS increases. This is understandable for structure and sequence+structure-based programs, but it is less obvious for sequence-based programs. The average sequence identity of these alignments is almost constant whatever the RMS (between 21 and 25%). The decrease of CS scores for sequence-based programs may be associated with the increase of the number of gaps—eight indels on average for alignments below 1 Å of RMS to 22 indels for all alignments above 3 Å—and to the increase of the number of proteins to align—from 3.5 proteins on average to 5.7. For the alignments above 30% of sequence identity, the CS scores decrease for alignments below 3 Å; this decrease is associated with a decrease of sequence identity (68% of sequence identity for alignments in the interval [0 Å, 1 Å], 52% for [1 Å, 2 Å] and 40% for [2 Å, 3 Å]). The variations are non-significant afterward (43% for [4 Å, 5 Å] and 41% above 5 Å) which is coherent with the stability of the sequence-based CS scores. When the RMS is high (>6Å) and the sequence identity not too low (>30% sequence identity), several sequence-based programs perform better than the structure-based programs, and the best program is a sequence+structure-based program. The structural variability may be due to unstructured regions of the proteins which may be seen in some structures of the difficult cases presented in Supplementary Figures S7 and S8.

We have also computed the RMS on the basis of the alignments resulting from the programs. The results are presented in Supplementary Figure S6. The multiple RMS among proteins of the families are smaller for structure-based methods than for sequence-based methods as expected because structure-based methods align proteins while optimizing the structural similarities. The RMS computed according to the reference alignments (black line) are in between the two.

3.6 SSE and burying effect

We also investigated whether structure-based methods are strongly dependent on secondary structures and solvent exposure. We computed the SP and CS scores independently for core residues

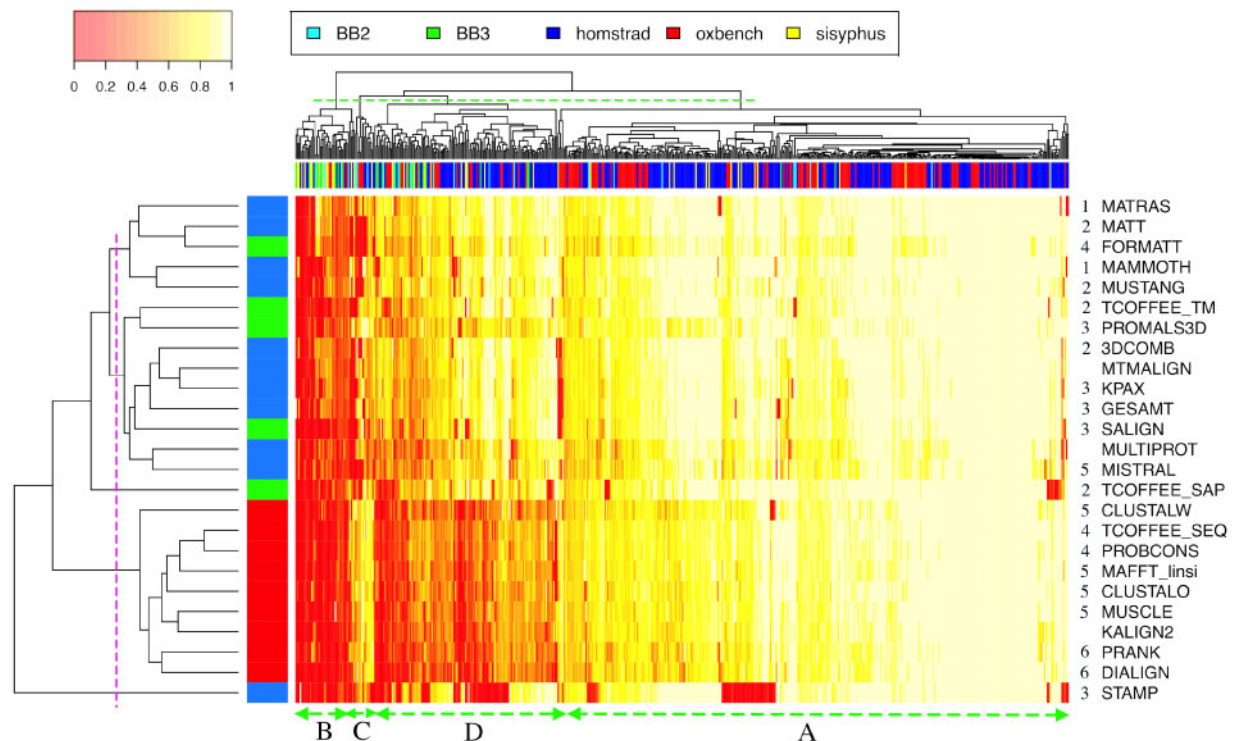


Fig. 4. CS scores heatmap and hierarchical classification of the programs and of the alignments (complete method, Euclidian distance). The program colors are the same as in Figure 2. The numbers before each program correspond to those in Figure 3

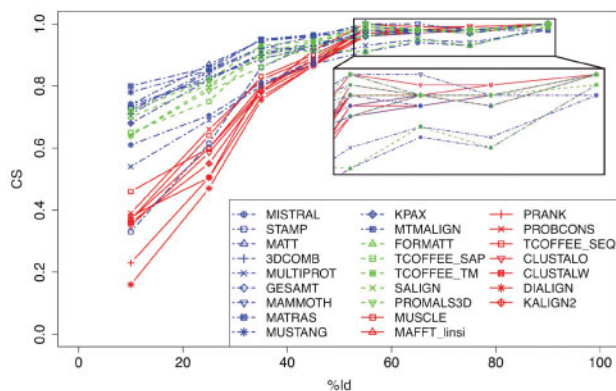


Fig. 5. CS scores as a function of percentage identity of the core reference alignments. Color code as in Figure 2

respectively in helices, strands or loops; the same procedure was applied for exposed or buried residues. The results are presented in Figure 7 (CS scores only). CS scores fall sharply for loops and a little for helices for structure-based and structure+sequence-based methods compared with sequence-based methods. The structure-based methods are sensitive to regular secondary structures. The scores decrease for exposed residues for all type of methods. The structural variability of exposed regions explains the difficulties of structure-based and sequence+structure-based programs. For the sequence-based programs, the decrease is probably due to the decrease of sequence identity (46% versus 33%).

3.7 Database effect

We wondered whether the success rate of the programs was dependent on the database. The composition of the various databases is

different in terms of sequence identity and core definition. We tried to eliminate these biases by selecting alignments between 10 and 40% of sequence identity since all databases are present in this range. Besides only core positions in conserved regular secondary structures were selected. In Figure 8, it is clear that the CS scores fluctuate depending on the reference alignment origin. The median scores are globally higher and less variable for HOMSTRAD and OXBENCH that contain more alignments and whose generation procedure is automatic. For BB2, BB3 and SISYPHUS, the discrepancy of the scores is larger. One may consider that a bias in favor of the structure-based methods is present in HOMSTRAD and OXBENCH. Yet, the ranking of the programs is similar: the same structure-based or structure+sequence-based programs are the best, even though their order varies slightly. The most affected program is STAMP, whose performances are poorer with the last three databases. It is used in the building procedure of HOMSTRAD and OXBENCH nevertheless its performances for those two databases are not the best. The best program in this 10–40% sequence identity subset is MATT, followed by MATRAS and FORMATT. Therefore, these programs have good results with divergent proteins.

3.8 Gaps

The proportion of gap opening is clearly different in sequence-based and structure-based programs (Fig. 9). The structure-based programs except MAMMOTH tend to over-estimate the number of indels and the sequence-based ones tend to under-estimate the number of gaps. MAMMOTH has a linear penalty gap function that seems to be quite efficient. PROMALS3D has also a linear gap penalty function and tends to put fewer gaps than in the reference alignments. PRANK which has been designed for placing correctly indels is the closest to the reference. As most of the structure-based methods work with small

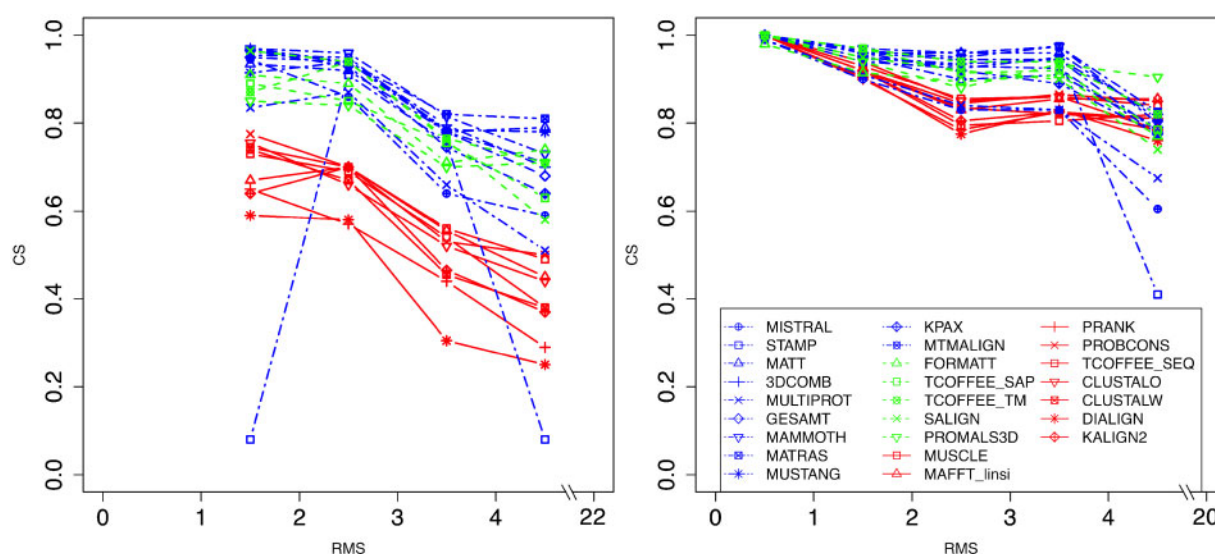


Fig. 6. CS median score as a function of the RMS computed from the reference alignments. Left: alignments below 30% of sequence identity. Right: alignments above 30% of sequence identity. Color code as in Figure 2

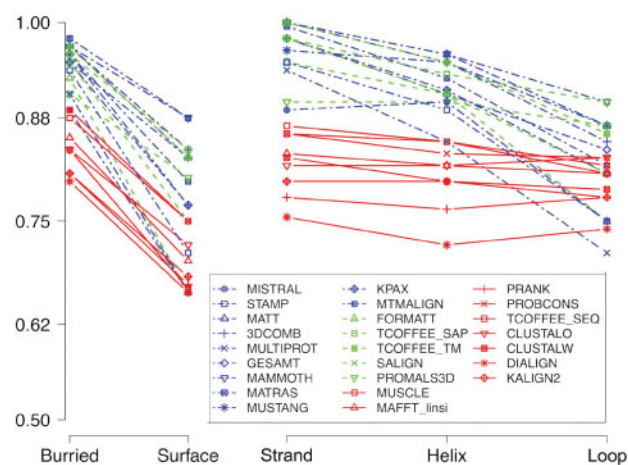


Fig. 7. Median CS scores for each program. Left: residues either buried or exposed. Right: core columns where the residues are either in helix or strand or loop. Color code as in Figure 2

structural blocks, they do not have a gap penalty function, which may explain this possible over-estimation of gaps. We believe that some improvement in the gap treatment for structure-based and sequence+structure-based methods should improve their performance.

4 Discussion

In this article, we have compared the ability of sequence-based, structure-based and sequence+structure-based alignment programs to retrieve supposed homologous positions defined in reference alignments from five well-known databases. The structure-based programs have globally better performances than the sequence-based ones, but also better than most of the structure+sequence-based programs. A first group of two structure-based programs—MAMMOTH and MATRAS—scores significantly better than the others. A second group is close: MATT, MUSTANG and 3DCOMB (structure-based), TCOFFEE_TM and TCOFFEE_SAP (sequence+structure). All these

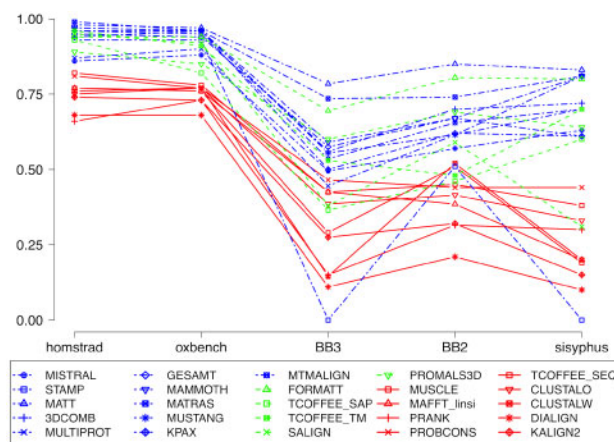


Fig. 8. Median CS scores for each program and each database, restricted to alignments in the range 10–40% sequence identity. Besides only are considered core positions in regular secondary structures perfectly conserved in all the proteins of the family. Color code is the same as in Figure 2

seven programs build the alignments from pairwise aligned fragments of few residues. The program performances are different according to the hierarchical clustering of their results: they do not all cluster together, meaning that their success or failure varies with the alignments. A consensus method may achieve better results if it can identify the cases where each method succeeds, as it has been also suggested in the *Berbalk et al. (2009)*. In TCOFFEE_TM and TCOFFEE_SAP, adding structure information clearly improves the alignment achieved by TCOFFEE_SEQ, but it is not the case for MATT and FORMATT. The consistency-based programs (TCOFFEE, PROBCONS) are the best ranked as far as the sequence-based programs are concerned in *Pais et al. (2014; Thompson et al., 2011)* except for MAFFT. The performance differences between sequence and structure-based programs are stronger for low identity alignments as it has been highlighted by *Kim and Lee (2007)*. The sequence-based program performances fall sharply for low sequence identity alignments but their performances are similar to structure-based programs above 50% of sequence identity. When the structural variations are large, structure-based program

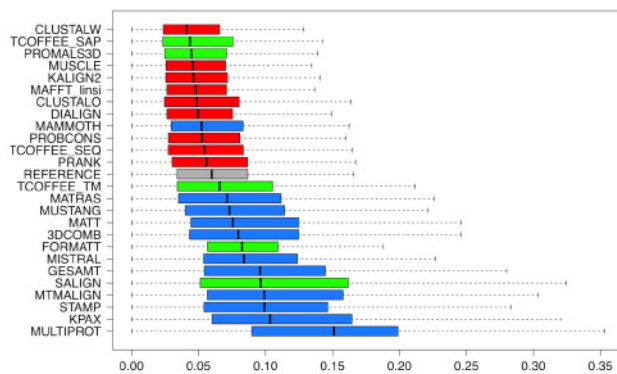


Fig. 9. Distribution of the number of columns containing one or more gap opening. The colors are as in Figure 2, and gray is for reference alignments

results may be worse than sequence-based programs. Other difficult cases for structure-based programs are the loops, the proteins with structural repetitions as Leucine rich repeats and proteins with large insertions or deletions. We can conclude that while aligning proteins for the identification of homologous positions and if all its structures are known, it is better to align the proteins only according to their structures. Today we can consider that the proportion of protein families with available structure amounts to 8700 over a total of 17425 families or domains in PFAM database (El-Gebali *et al.*, 2019). In the cases where not all structures are known, it should be better to use a sequence+structure-based method such as TCOFFEE_TM, but this particular case has not been addressed in this study.

As the five reference alignment databases are built from protein structural information we wondered whether it would advantage structure-based methods. For the two automatically generated databases, the alignments are computed by structure-based programs and it should favor homologous or even non-homologous but structurally similar positions that are more easily retrieved with structure than with sequence only. The case is different with manually curated alignments because no structure-based method has been used to build them and all kinds of information have also been used: sequence, function and structure. The scores of all programs and their dispersions are similar if the two automatic databases, HOMSTRAD and OXBENCH are used, and they are globally lower and more variable using the three other databases. Moreover, whatever the database used, the first ranked program is always a structure-based program. Although, structure-based and sequence+structure-based programs have better scores than sequence-based programs. It would be interesting to compare program alignments altogether without a reference in order to check their consistency, or to compute phylogenetic trees from the program alignments and derive a score from the accuracy of the trees. Finally, some improvements concerning usability and applicability of structure-based programs would be worthwhile and structure-based programs could improve the gap placement in the alignments.

5 Conclusions

For identifying homology in proteins, we can conclude that it is better to use structure information than sequence information only, yet the difficulty of combining sequence and structure information is obvious: the sequence+structure-based methods are not better than the structure-based methods. Several programs are globally equivalent in performance but their behavior varies for each alignment, and a consensus method might achieve better results. A real model of sequence and structure protein evolution would greatly improve the methods

but such a model is quite difficult to design mainly because of the folding process that may drastically change the structure even if the sequence difference is not that strong. There is also still room for improvement in term of software ergonomics and gap treatments. This study showed that, if several structures of a family are known, the most reliable alignment is the structural one. However, usually far more sequences than structures of a family are available so the use of sequence+structure-based methods with all sequences and known structures would gather all available information and may produce the best alignment. Computing several kinds of alignments using tools like STRAP (Gille and Frömmel, 2001) that allow combining alignments would be the most advisable approach.

Acknowledgements

Our thanks to the authors of the programs used in this study for the informal discussion, Joël Pothier for discussion and critical reading of the manuscript and to Theresé Pothier for the English proof reading. We also thank the referees for their pertinent remarks that permitted us to make noticeable improvements in the manuscript.

Funding

This study has been supported from regular supplies provided both involved laboratories.

Conflict of Interest: none declared.

References

- Ait,L. *et al.* (2013) DIALIGN at GOBICS—multiple sequence alignment using various sources of external information. *Nucleic Acids Res.*, **41**, W3–W7.
- Alva,V. *et al.* (2015) A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, **4**, e09410.
- Andreeva,A. *et al.* (2007) SISYPHUS—structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res.*, **35**, D253–D259.
- Balaji,S. *et al.* (2001) PALI-a database of Phylogeny and ALIGNment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.
- Berbalk,C. *et al.* (2009) Accuracy analysis of multiple structure alignments. *Protein Sci.*, **18**, 2027–2035.
- Blackshields,G. *et al.* (2006) Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol. (Gedruckt)*, **6**, 321–339.
- Daniels,N.M. *et al.* (2012) Formatt: correcting protein multiple structural alignments by incorporating sequence alignment. *BMC Bioinformatics*, **13**, 259.
- Dessimoz,C. and Gil,M. (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.*, **11**, R37.
- Do,C.B. *et al.* (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
- Dong,R. *et al.* (2018) mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics.*, **34**, 1719–1725.
- Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
- Edgar,R.C. (2010) Quality measures for protein alignment benchmarks. *Nucleic Acids Res.*, **38**, 2145–2153.
- El-Gebali,S. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
- Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.
- Feng,Z.K. and Sippl,M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold. Des.*, **1**, 123–132.
- Friedman,M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, **32**, 675–701.
- Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

- Gerstein, M. and Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.*, **7**, 445–456.
- Gille, C. and Frömmel, C. (2001) STRAP: editor for STRuctural alignments of proteins. *Bioinformatics*, **17**, 377–378.
- Godzik, A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
- Golubchik, T. *et al.* (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 2433–2442.
- Goujon, M. *et al.* (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.
- Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
- Iantorno, S. *et al.* (2014) Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol. Biol.*, **1079**, 59–73.
- Illergård, K. *et al.* (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins*, **77**, 499–508.
- Jean, P. *et al.* (1997) Automated multiple analysis of protein structures: application to homology modeling of cytochromes P450. *Proteins*, **28**, 388–404.
- Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
- Katoh, K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
- Kawabata, T. (2003) MATRAS: a program for protein 3D structure comparison. **31**, 3367–3369.
- Kawabata, T. and Nishikawa, K. (2000) Protein structure comparison using the Markov transition model of evolution. *Proteins*, **41**, 108–122.
- Kim, C. and Lee, B. (2007) Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinformatics*, **8**, 355.
- Konagurthu, A. *et al.* (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
- Krissinel, E. (2012) Enhanced fold recognition using efficient short fragment clustering. *J. Mol. Biochem.*, **1**, 76–85.
- Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.
- Lamarine, M. *et al.* (2001) Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cell. Mol. Life Sci.*, **58**, 492–498.
- Landan, G. and Graur, D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 1380–1383.
- Larkin, M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
- Lassmann, T. and Sonnhammer, E.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Res.*, **33**, 7120–7128.
- Lassmann, T. and Sonnhammer, E.L. (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.
- Lassmann, T. *et al.* (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res.*, **37**, 858–865.
- Le, Q. *et al.* (2017) Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*, **33**, 1331–1337.
- Lecompte, O. *et al.* (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*, **270**, 17–30.
- Lemey, P. *et al.* (2009) *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge.
- Levasseur, A. *et al.* (2008) Strategies for reliable exploitation of evolutionary concepts in high throughput biology. *Evol. Bioinform. Online*, **4**, 121–137.
- Liberles, D.A. *et al.* (2012) The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.*, **21**, 769–785.
- Löytynoja, A. and Goldman, N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA*, **102**, 10557–10562.
- Lupyan, D. *et al.* (2005) A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**, 3255–3263.
- Madhusudhan, M.S. *et al.* (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, **22**, 569–574.
- Mayr, G. *et al.* (2007) Comparative analysis of protein structure alignments. *BMC Struct. Biol.*, **7**, 50.
- Menke, M. *et al.* (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol.*, **4**, e10.
- Micheletti, C. and Orland, H. (2009) MISTRAL: a tool for energy-based multiple structural alignment of proteins. *Bioinformatics*, **25**, 2663–2669.
- Mizuguchi, K. *et al.* (1998a) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
- Mizuguchi, K. *et al.* (1998b) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.
- Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
- Morgenstern, B. *et al.* (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
- 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.
- Nepomnyachiy, S. *et al.* (2017) Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc. Natl. Acad. Sci. USA*, **114**, 11703–11708.
- Notredame, C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
- Nuin, P.A. *et al.* (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.
- O’Sullivan, O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.
- Orengo, C.A. and Taylor, W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.
- Ortiz, A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
- Pais, F.S.-M. *et al.* (2014) Assessing the efficiency of multiple sequence alignment programs. *Algorithms Mol. Biol.*, **9**, 4.
- Pei, J. and Grishin, N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.
- Pei, J. *et al.* (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.
- Petersen, B. *et al.* (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.
- R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Raghava, G. *et al.* (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
- Ritchie, D.W. *et al.* (2012) Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, **28**, 3274–3281.
- Russell, R.B. and Barton, G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.
- Sali, A. and Blundell, T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.
- Sauder, J.M. *et al.* (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.
- Shatsky, M. *et al.* (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.
- Shatsky, M. *et al.* (2005) Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, **62**, 209–217.
- Sievers, F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

- Slater, A.W. *et al.* (2013) Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics*, **29**, 47–53.
- Subramanian, A.R. *et al.* (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
- Sutcliffe, M.J. *et al.* (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.
- Theobald, D.L. and Wuttke, D.S. (2006) THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, **22**, 2171–2172.
- Thompson, J.D. *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**, 4673–4680.
- Thompson, J.D. *et al.* (1999a) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
- Thompson, J.D. *et al.* (1999b) BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
- Thompson, J.D. *et al.* (2005) BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
- Thompson, J.D. *et al.* (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*, **6**, e18093.
- van der Lee, R. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
- Van Walle, I. *et al.* (2005) SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.
- Wang, S. *et al.* (2011) Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. **27**, 2537–2545.
- Winn, M.D. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.
- Wong, K.M. *et al.* (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.
- Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
- Zhang, Y., and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.