

Protein Multiple Alignments: Sequence-based vs Structure-based Programs

Mathilde Carpentier^{1*}, Jacques Chomilier²

¹ Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, Paris, France.

² Sorbonne Université, CNRS, MNHN, IRD, IMPMC, BiBiP, Paris, France

* To whom correspondence should be addressed. Tel: +33 1 40 79 48 15; Email: mathilde.carpentier@mnhn.fr

ABSTRACT

Facing the huge increase of information about proteins, classification has reached the level of a compulsory task, essential for assigning a function to a given sequence, by means of comparison to existing data. Multiple sequence alignment programs have been proven to be very useful and they have already been evaluated. In this paper we wished to evaluate the added value provided by taking into account structures. We compared the multiple alignments resulting from 24 programs, either based on sequence, structure, or both, to reference alignments deposited in five databases. Reference databases, on their side, can be split in two: more automatic ones, and more manually ones. Scores have been attributed to each program. As a global rule of thumb, five groups of methods emerge, with the lead to two of the structure-based programs. This advantage is increased at low levels of sequence identity among aligned proteins, or for residues in regular secondary structures or buried. Concerning gap management, sequence-based programs place less gaps than structure-based programs. Concerning the databases, the alignments from the manually built databases are the more challenging for the programs.

Abbreviations: ASA: Accessible Surface Area; CDD: Conserved Domain Database; MSA: Multiple Sequence Alignment; PDB: Protein Data Bank; RMS: Root Mean Square; SP: sum of pairs; CS: column score. DP: Dynamic Programming, AFP: Aligned Fragment Pairs, SSE: secondary structure element

INTRODUCTION

Multiple alignments of protein sequences constitute an essential tool to explore evolution, diversity, conservation and function of proteins (1–4). Despite the impressive increasing number of available structures, most of these alignments are still computed by software relying only on sequence information. Protein structures are mostly used in a second step to manually refine the alignment (5). Since it is generally admitted that structures are more conserved than sequences (6), structural information may guide a particularly difficult alignment of very divergent proteins (7). Nevertheless, multiple protein structure alignment methods, or methods combining sequence and structure, are not widespread.

A structural alignment may outline other types of information than homology (8). In protein sequence alignments we align amino acids which are considered as homologous, i.e. deriving from an ancestral

sequence by substitution at the same site. In structural alignments, we align positions similar from the point of view of local and/or global conformations. This structural similarity does not always imply homology (8). Indeed, sub domain fragments can be found in many different folds, with unrelated functions or various origins (9–11). 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 of the folding step of protein structures.

From all the previous arguments, it is difficult to claim that structure alignments provide the golden standard to evaluate the quality of sequence alignment. This is particularly the case when the proteins have a low level of similarity or if the homology of the whole genes is questionable. However, as structures are better conserved, alignments should be more reliable when information from sequences and structures are combined. We therefore compared 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 (12). The most widely used benchmarks are composed of a collection of reference alignments considered as the gold standard. The reference alignments are constructed mainly from the sequence and structural information, but also according to other information as the function. Some of them are manually curated. When a new alignment method or an improvement is published, its performance is usually assessed by comparison to other methods by aligning the proteins of these reference alignment databases. There are also many comparative studies of the performances of sequence alignment methods (13, 14). The second type of benchmarks relies on simulated sequences (15). A third type of benchmarks relies on a direct comparison of all computed alignments, without any reference alignment (16, 17). The fourth type of benchmarks is to calculate phylogenetic trees from the alignments and to check their validity (18). For structure-based alignment methods, less comparative studies have been conducted and most of them compare pairwise structural alignment programs (8, 19–24). Multiple structural alignment programs are compared in the study of Berbalk et. al. (25). The authors first remarked that the programs were generally very difficult to use and that there is room for improvements concerning usability and applicability. They concluded that combining different alignment approaches into a single program supported by 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 are not aware of a thorough comparative study of the performance of sequence-based and structure-based programs. We believe that such a study is important to address some questions: Are structure-based methods really superior to retrieve homologous residues? Or is it the sequence and structure ones? In what cases should we use structure methods, sequence and structure or sequence-based methods? These are the aims of this article.

MATERIAL AND METHODS

Databases

In this study, we used the most widely used type of benchmarks: the reference multiple alignments built from sequences, structures and function information, and considered as the gold standard. The usage of simulated sequence is not possible in our case because there is no structure associated. It is possible to compare all alignments without a reference but as programs may be consistently wrong; therefore we decided to avoid this approach in this article. The phylogeny-based approach would be very interesting but it requires a database of validated trees of genes with all known protein structures, which is beyond the scope of the article.

We have selected 846 alignments, containing at least three protein chains, from five reference multiple alignment databases: BALIBASE 2 (26), BALIBASE 3 (27), HOMSTRAD (28), OXBENCH (29) and SISYPHUS (30). Some alignments have been discarded: those with two or more proteins with identical amino acid sequence, NMR or theoretical model structures, structures with missing residues and those with various inconsistencies. We did not consider the alignments of other well-known databases listed in (31) for various reasons: PREFAB (32) because it is composed of pairwise alignments; IRMBASE (33) because there is no structure associated to the simulated fragments and SABMARK (34) because of some inconsistencies in the multiple alignments which are built from pairwise structural alignments, pointed by the author and in (35). We also had difficulties accessing PALI (36) and couldn't download the database. From 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. HOMSTRAD, from which we selected 357 families, is exclusively based on proteins with known structures, and each family is aligned with the programs MNYFIT (37), STAMP (38) and COMPARER (39). These produce structure-based alignments that are annotated with JOY (40) and individually examined and modified if necessary. JOY produces core blocks annotations defined as the regular secondary structures elements. We retrieved from OXBENCH 330 alignments from the subset multi with three or more proteins in each, not split in domains (full-length sequences). The multiple alignments are computed by STAMP (38). 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 (41) 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 size and are manually curated. Sisyphus annotates the structurally equivalent residues in the alignments and we consider them as the core blocks.

Many structure-based alignment programs don't 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 to solve this issue: the first one retrieves the correspondence of the protein between the

reference alignment and the output alignment and the second makes each sequence of an output alignment identical to the sequence in the reference alignment.

Alignment Quality Evaluation

The alignments produced by each program are evaluated by comparison with the reference alignments by two scores, following Thomson et al (13): 1) the fraction of pairs of residues in the reference alignment correctly identified by a given method, known as the Sum-of-Pairs (SP) score; 2) the Column Score (CS) score, which describes the fraction of reference columns identified. As usually done in alignment method comparisons (13, 42, 43), Friedman tests (44) 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 (45). The average multiple RMS have been computed with THESEUS (46) 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. Solvent accessibility (ASA) is calculated with NACCESS for all the proteins, in order to separate the amino acids in two classes: either buried (relative ASA < 25%) or exposed (47). Secondary structure assignments have been performed with STRIDE (48). The six classes given in the output of STRIDE are back coded in three classes: helices, strands and coils. All analyses have been made according to these characteristics: the residues of the core blocks have been attributed either as buried or accessible, and either in helix, strand or other (loop).

Programs

We have 3 categories of multiple alignment programs: sequence-based, sequence+structure-based and structure-based. We only included programs respecting the following conditions: (i) available for download, (ii) output a file containing the alignment in a standard format, (iii) run without error. Each multiple alignment had to be computed in less than two hours. Some programs failed to produce enough alignments to allow a significant analysis of their performance and were excluded if they produced an alignment for less than 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 didn't respect our criteria. We were also surprised by the few number of sequence+structure alignment methods. We didn't include methods improving alignments afterwards like STACCATO (49). There is a great number of sequence-based programs and we only tested the most popular according to the last studies (14, 50). 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 10 structure-based programs.

Results

Number of computed alignments

All programs have been run on the 846 alignments. All the sequence-based programs were able to calculate all the alignments but some programs of the two other categories failed for some alignments. The proportion of successful alignments is reported in Table 2. Only MATRAS, TCOFFEE_SAP, TCOFFEE_TM, and KPAX, successfully computed all alignments. The failure causes were sometimes the time limit, but most of the time, the programs returned some errors. MAMMOTH encountered the most failures; there is obviously a limit of 25 proteins per alignment for it. The greatest number of failures for all programs is with the SISYPHUS database which is not surprising because it is a benchmark built to be challenging. To improve the robustness of our analysis, we decided to restrict our analysis to the alignments computed by all programs, resulting in 531 alignments: 24 from BB2, 24 from BB3, 288 from HOMSTRAD, 155 from OXBENCH and 40 from SISYPHUS. These 531 alignments involve 2043 chains.

Databases

The distribution of mean pairwise sequence identity among the 531 multiple alignments of the databases 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%; but, restricted to the core alignments, the proportion increases to 79%.

We checked if some protein families were present in several datasets. We found some chains in several databases even if all the proteins of the family are not the same. The number and proportion of chains included in two databases are listed in Supplementary Table 1. There is some overlap between BB2 and BB3: 48 chains are present both in BB2 and BB3. However, the protein families are different between BB2 and BB3 so we decided to keep them all. The overlaps are very weak for the other datasets.

Global Analysis of Alignment Scores

The boxplot distribution of SP and CS scores of each program run on the 531 alignments are presented in Figure 2. The exact median values are reported in Table 2 of the Supplementary Data. Globally, for all programs, 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 methods. Similarly, in half of the alignments, more than 81% of the alignment columns are correct. The scores vary with the programs and it is strikingly clear that structure-based alignment programs have globally better results, except for MULTIPROT. The sorting is roughly the same for SP and CS scores except for FORMATT and MULTIPROT. STAMP has the greatest variability in its results and it is not the best despite the fact that it has been used to build the alignments of two databases (HOMSTRAD and OXBENCH). It is interesting to notice that FORMATT, a modified version of MATT to include sequence

information, achieves worse than MATT. It highlights the difficulty to combine sequence and structure information. It is nevertheless possible: TCOFFEE_TM is the best sequence+structure program and it achieves clearly better than TCOFFEE_SEQ. It is however surprising that sequence+structure based methods do not achieve better than structure only methods: their alignments rely on the same structural information, enhanced with sequence 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 531 alignments (see method). In Figure 3, five groups of methods appear: the differences are mostly non-significant between the programs within each group, but they are significant with the programs outside the groups. The two first groups (blue symbols on the diagonal) contain all structure-based alignment methods but MULTIPROT and three of the sequence+structure-based methods. The next three groups (red symbols) contain all sequence-based alignment methods and two sequence+structure-based methods. The results of TCOFFEE_SAP are identical to TCOFFEE_SEQ. STAMP is particular: its results greatly vary and it is not significantly better or worse than the programs in the two middle groups. MULTIPROT is also particular: the differences are not significant for the SP scores with DIALIGN, PRANK, CLUSTALW and KALIGN2, but for the CS scores, the differences are not significant with all sequence-based methods but PRANK and DIALIGN.

From this analysis, we can conclude that there is a ranking of groups of programs according to their overall performance, and that structure-based programs achieve better scores.

We also proceeded to 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 Data Figure 1 for SP scores. The results are extremely similar with both scores. Considering alignment clustering, alignments for which all categories of methods succeed are more concentrated on the right side of the heatmap and the alignment clustering tree. In a thin central strip, the sequence-based methods have better scores: the score cells are in darker red for the structure-based. On the left side, the scores are better for structure-based and structure+sequence-based methods and in the extreme left side, all methods fail. Those difficult alignments are mostly from BB2, BB3 and SISYPHUS.

Considering program clustering, all structure-based programs but STAMP and sequence+structure-based programs but TCOFFEE_SAP are in the same sub-tree. All sequence-based are also pooled together. MUSTANG, MAMMOTH and MATRAS, which performances are undistinguishable according to the Friedman test, are very close in the tree. It means that their performances are similar, even if there are some discrepancies for some alignments. The next program in this branch is PROMALS3D, but its performances are significantly different from the three previous programs according to the Friedman tests. MATT and FORMATT are clustered together, but we know from the test that MATT results are better. The next closest method is TCOFFEE_TM. 3DCOMB, MTALIGN, GESAMT and SALIGN are clustered together but according to the previous tests, 3DCOMB achieves the best results in this group of programs.

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 2 for SP scores. As expected, the differences between structure-based and sequence-based methods are stronger for alignments of very divergent proteins. The difference is stronger in the case of CS scores but the effect is globally the same. It is more surprising to see that even at very high identity levels, structure-based programs still provide better scores than sequence-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 (see Supplementary Figure 3) but it is visible on the CS scores (Supplementary Figure 4).

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 core residues in helices, strands or loops; the same procedure was applied for exposed or buried residues. The results are presented Figure 6 for CS scores. The scores decrease for loop residues; this decrease is more importantly for structure-based and structure+sequence based methods than for sequence-based methods. Similarly, the scores decrease for exposed residues for all methods. In summary, buried or regular secondary structure regions are better aligned by all programs than exposed regions or loops.

Database effect

We wondered if the success rate of the programs was dependent on the databases. The composition of the various databases is different in terms of sequence identity and core definition. We tried to remove these biases by selecting alignments between 10% and 40% of sequence identity, because all databases are present in this range. Besides, only core positions in conserved regular secondary structures were selected. In Figure 7, it is clear that the CS scores fluctuate depending on the reference alignment origin: the median scores are globally higher and less variable for the two HOMSTRAD and OXBENCH which contain more alignments and whose generation procedure is more automatic than BB2, BB3 and SISYPHUS. However, the ranking of the programs is similar: the same structure-based or structure+sequence-based programs are the best, even if their order slightly varies. The most affected program is STAMP, which performances are poorer with the three last databases.

RMS and sequence identity

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 resemblances (see Figure 8). The RMS computed according to the reference alignments (Figure 8, black line) are in between the two categories except for alignments above 70% identity where these RMS are higher. These high RMS are mostly from OXBENCH alignments and are mainly due to an alignment in multidomain chains. The best programs, as resulting from the previous

sorting, are not those with the smallest RMS; on the contrary, the order is globally reversed. Figure 8 highlights the differences in the goal of the programs: either optimize structural similarities or consider some evolutionary aspects. As we are testing the capabilities of programs to retrieve homologous positions, the second category of programs is advantaged in our study: if they are better here, it is only in their capacity at retrieving homologous positions, but they may miss other structural similarities.

GAPS

The proportion of gap opening is clearly different in sequence-based and structure-based programs (see Figure 9). The structure-based programs but MAMMOTH tend to over-estimate the number of indels and the sequence-based tend to under estimate the number of gaps. MAMMOTH has a linear penalty gap function which seems to be quite efficient. PROMALSD3D has also a linear gap penalty function and tends to place less gaps than in the reference alignments. PRANK, which has been designed to correctly place the indels, is the closest method to the reference. As most of the structure-based methods work with small structural blocks, they don't have a gap penalty function, which explains this possible over-estimation. We believe that some improvement in the gap treatment for structure-based and sequence+structure based methods should improve their performance.

DISCUSSION

In this article, we have compared the ability of sequence-based, structure-based and sequence+structure based multiple alignment programs to retrieve homologous positions of defined in reference alignments from five well known datasets. The structure-based programs have globally better performances than sequence-based, but also than most of the structure+sequence-based programs. A group of five structure-based plus one sequence+structure-based programs are significantly better than the others: MAMMOTH, MATRAS, TCOFFEE_TM (sequence+structure), 3DCOMB, MATT and MUSTANG. All those six programs build the alignments from pairwise aligned fragments of few residues. No obvious superior methodology has been identified. However, it is interesting to notice that those program performances are different, according to a hierarchical clustering of their scores: they do not all cluster together, meaning that their success or failure depends upon the alignment. It is therefore possible that a consensus method achieves better results if it is possible to identify the cases where each method succeeds, as it have been suggested also in the article of Berbalk et al. (25). The performance differences between sequence and structure-based programs are stronger for low identity alignments as it has been highlighted in Kim and Lee (23), but they are still present at high identity. They weakly depend on the 3D localization of the residues.

TCOFFEE_TM is the only sequence+structure-based program in the group of best programs. The adjunction of structure information clearly improves the alignment done by TCOFFEE_SEQ. In sequence-based programs, in our study, the consistency-based programs (TCOFFEE, PROBCONS) are the best ranked as in (14, 51) but without MAFFT. These consistency-based methods are quite efficient and it would be interesting to see their results with only structure information. We can conclude that while aligning proteins for the identification of homologous positions and if all structures are known, it is better to align the protein with their structures. In the case where not all structures are

known, it is probable that it is better to use a sequence+structure-based method as TCOFFEE_TM, but this exact case has not been addressed in this study. Knowing the difficulty of combining structure and sequence information, this case has to be tested before counseling it.

The second performance group of structure-based programs contains four structure-based programs (MTALIGN, GESAMT, KPAX and STAMP) and two sequence+structure-based programs (SALIGN and PROMALS3D). STAMP is apart, but all three other structure-based programs tend to produce low RMS alignments; they are also clustered together. These programs seem to be more dedicated to the identification of structurally similar regions, which are not always homologous. It would be very interesting to compare the regions identified as structurally similar by those programs and not by the six previous ones: these regions may sequence similar-structure dissimilar which would help to understand the complex evolution of protein structure. STAMP has a very different behavior which renders comparison difficult: globally, it succeeds for databases that are more automatically constructed (HOMSTRAD and OXBENCH) and mostly fails for the others (BB2, BB3 and SISYPHUS). However, it is surprising that its performances are not the best for the two first databases as it is used to compute their alignments. Maybe the program version or parameters are different, or the afterwards refinement of the reference alignments may explain it. MULTIPROT is also very different from the other structure-based programs: it is dedicated to the identification of reliable aligned columns which are locally structurally similar. It does not align the other regions, which explains its poor scores. However, it is very efficient for column identification: its CS score are clearly better than its SP scores, and it is the only structure-based program with this behavior.

The scores of all programs and their dispersions are similar for the two databases HOMSTRAD and OXBENCH which are the most automatically generated. The scores are different for the three other databases which are more manually built: they are globally lower and more variable among the programs, meaning that these alignments seem to be more difficult to retrieve. 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. However, the ranking may vary: in the subset of alignments from the three manual databases, with 10 to 40% of identity and only the positions perfectly conserved in terms of SSE, MATT, MATRAS, FORMATT and MUSTANG achieve the best results. Another bias in this study is that these 5 benchmarks are built from protein structural information which may advantage structure-based methods. Other broad of benchmarks exist (12). It is not possible to use simulated sequences in the case of structure-based programs but it would be interesting to compare the program alignments altogether without a reference to check their consistency. The last type of benchmark is to compute phylogenetic trees from the program alignments to compute a score from the correctness of the trees. This analysis could also be done for all type of programs.

All structure-based programs except MAMMOTH and all sequence+structure-based programs except PROMALS3D and TCOFFEE_SAP have a greater proportion of columns with a gap opening than reference alignments and all other methods have a lesser proportion. Most of structural-based methods don't have a penalization function of gaps which explains this behavior. It is possible that a penalization of gaps would improve the alignment quality.

Finally, some improvement improvements concerning usability and applicability of structure-based programs would generally been valuable.

CONCLUSION

We can conclude from this study that it is indeed better to use structure information than sequence information only to identify homology in proteins, but the difficulty of combining sequence and structure information is obvious: the sequence+structure-based methods are not better than the structure-based method. Several programs are globally equivalent in performance but their behavior vary for each alignment and maybe, a consensus method could achieve better results. However, a real model of sequence and structure protein evolution would surely greatly improve the methods but such a model is quite difficult to design notably because of the folding process which 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.

AVAILABILITY

All results presented in this study are available upon request to the corresponding author.

ACKNOWLEDGEMENT

The authors greatly appreciated the informal discussion with the authors of most of the programs used in this study.

FUNDING

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

CONFLICT OF INTEREST

None declared

REFERENCES

1. Levasseur,A., Pontarotti,P., Poch,O. and Thompson,J.D. (2008) Strategies for Reliable Exploitation of Evolutionary Concepts in High Throughput Biology. *Evol Bioinform Online*, **4**, 121–137.
2. Wong,K.M., Suchard,M.A. and Huelsenbeck,J.P. (2008) Alignment Uncertainty and Genomic Analysis. *Science*, **319**, 473–476.
3. Lecompte,O., Thompson,J.D., Plewniak,F., Thierry,J.-C. and Poch,O. (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*, **270**, 17–30.
4. Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, **25**, 351–360.
5. Lemey,P., Salemi,M. and Vandamme,A.-M. (2009) The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing Cambridge University Press.

6. Illergård, K., Ardell, D.H. and Elofsson, A. (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins*, **77**, 499–508.
7. Jean, P., Pothier, J., Dansette, P.M., Mansuy, D. and Viari, A. (1997) Automated multiple analysis of protein structures: application to homology modeling of cytochromes P450. *Proteins*, **28**, 388–404.
8. Godzik, A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci*, **5**, 1325–1338.
9. Lamarine, M., Mornon, J.P., Berezovsky, I.N. and Chomilier, J. (2001) Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *CMLS, Cell. Mol. Life Sci.*, **58**, 492–498.
10. Alva, V., Söding, J. and Lupas, A.N. (2015) A vocabulary of ancient peptides at the origin of folded proteins. *eLife Sciences*, **4**, e09410.
11. Nepomnyachiy, S., Ben-Tal, N. and Kolodny, R. (2017) Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *P Natl Acad Sci Usa*, **114**, 11703–11708.
12. Iantorno, S., Gori, K., Goldman, N., Gil, M. and Dessimoz, C. (2014) Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol. Biol.*, **1079**, 59–73.
13. Thompson, J.D., Plewniak, F. and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, **27**, 2682–2690.
14. Thompson, J.D., Linard, B., Lecompte, O. and Poch, O. (2011) A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLoS ONE*, **6**, e18093.
15. Nuin, P.A., Wang, Z. and Tillier, E.R. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.
16. Lassmann, T. and Sonnhammer, E.L.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Res*, **33**, 7120–7128.
17. Landan, G. and Graur, D. (2007) Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. *Mol Biol Evol*, **24**, 1380–1383.
18. Dessimoz, C. and Gil, M. (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology*, **11**, R37.
19. Feng, Z.K. and M J Sippl (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold Des*, **1**, 123–132.
20. 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.
21. Sauder, J.M., Arthur, J.W. and Dunbrack, R.L. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.
22. Mayr, G., Domingues, F.S. and Lackner, P. (2007) Comparative analysis of protein structure alignments. *BMC Struct Biol*, **7**, 50.
23. Kim, C. and Lee, B. (2007) Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinformatics*, **8**, 355.
24. Slater, A.W., Castellanos, J.I., Sippl, M.J. and Melo, F. (2013) Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics*,

29, 47–53.

25. Berbalk,C., Schwaiger,C.S. and Lackner,P. (2009) Accuracy analysis of multiple structure alignments. *Protein Science*, **18**, 2027–2035.
26. Thompson,J.D., Plewniak,F. and Poch,O. (1999) BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
27. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
28. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*, **7**, 2469–2471.
29. Raghava,G., Searle,S.M., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 1–23.
30. Andreeva,A., Prlić,A., Hubbard,T.J.P. and Murzin,A.G. (2007) SISYPHUS--structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res*, **35**, D253-9.
31. Blackshields,G., Wallace,I.M., Larkin,M. and Higgins,D.G. (2006) Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol (Gedruckt)*, **6**, 321–339.
32. Edgar,R. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797.
33. Subramanian,A.R., Weyer-Menkhoff,J., Kaufmann,M. and Morgenstern,B. (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
34. Van Walle,I., Lasters,I. and Wyns,L. (2005) SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.
35. Edgar,R.C. (2010) Quality measures for protein alignment benchmarks. *Nucleic Acids Res*, **38**, 2145–2153.
36. Balaji,S., Sujatha,S., Kumar,S.S. and Srinivasan,N. (2001) PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res*, **29**, 61–65.
37. Sutcliffe,M.J., Haneef,I., Carney,D. and 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.
38. 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.
39. 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.
40. JOY: protein sequence-structure representation and analysis. - PubMed - NCBI.
41. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536–540.
42. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, **15**, 330–340.
43. Golubchik,T., Wise,M.J., Easteal,S. and Jermin,L.S. (2007) Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments. *Molecular Biology and Evolution*, **24**, 2433–2442.

44. Friedman, M. (1937) The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, **32**, 675–701.
45. R Core Team (2018) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
46. Theobald, D.L. and Wuttke, D.S. (2006) THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, **22**, 2171–2172.
47. Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. and Lundegaard, C. (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol*, **9**, 51.
48. Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–79.
49. Shatsky, M., Nussinov, R. and Wolfson, H.J. (2005) Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, **62**, 209–217.
50. Le, Q., Sievers, F. and Higgins, D.G. (2017) Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*, **33**, 1331–1337.
51. Pais, F.S.-M., Ruy, P. de C., Oliveira, G. and Coimbra, R.S. (2014) Assessing the efficiency of multiple sequence alignment programs. *Algorithms for molecular biology: AMB*, **9**, 4.
52. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
53. Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*, **38**, W695–W699.
54. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (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.
55. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
56. Al Ait, L., Yamak, Z. and Morgenstern, B. (2013) DIALIGN at GOBICS—multiple sequence alignment using various sources of external information. *Nucleic Acids Res*, **41**, W3–W7.
57. Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
58. Morgenstern, B., Frech, K., Dress, A. and Werner, T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
59. Lassmann, T., Frings, O. and Sonnhammer, E.L. (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res*, **37**, 858–865.
60. Lassmann, T. and Sonnhammer, E.L. (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298–298.
61. 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.

62. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, **30**, 3059–3066.
63. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
64. Löytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *P Natl Acad Sci Usa*, **102**, 10557–10562.
65. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205–217.
66. Pei,J., Kim,B.H. and Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res*, **36**, 2295–2300.
67. Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.
68. O'Sullivan,O., Suhre,K., Abergel,C., Higgins,D.G. and Notredame,C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, **340**, 385–395.
69. Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol*, **266**, 617–635.
70. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, **33**, 2302–2309.
71. Madhusudhan,M.S., Webb,B.M., Marti-Renom,M.A., Eswar,N. and Sali,A. (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Engineering Design and Selection*, **22**, 569–574.
72. Daniels,N.M., Nadimpalli,S. and Cowen,L.J. (2012) Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC Bioinformatics*, **13**, 259.
73. Wang,S., Peng,J. and Xu,J. (2011) Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. **27**, 2537–2545.
74. 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.
75. Krissinel,E. (2012) Enhanced fold recognition using efficient short fragment clustering. *J Mol Biochem*, **1**, 76–85.
76. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G.W., McCoy,A., *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.
77. Ritchie,D.W., Ghoorah,A.W., Mavridis,L. and Venkatraman,V. (2012) Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, **28**, 3274–3281.
78. Lupyan,D., Leo-Macias,A. and Ortiz,A.R. (2005) A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**, 3255–3263.
79. Kawabata,T. (2003) MATRAS: A program for protein 3D structure comparison. **31**, 3367–3369.
80. Kawabata,T. and Nishikawa,K. (2000) Protein structure comparison using the markov transition model of evolution. *Proteins*, **41**, 108–122.

81. Menke,M., Berger,B. and Cowen,L. (2008) Matt: Local Flexibility Aids Protein Multiple Structure Alignment. *PLoS Comput Biol*, **4**, e10.
82. Dong,R., Peng,Z., Zhang,Y. and Yang,J. (2017) mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, 10.1093/bioinformatics/btx828.
83. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.
84. Konagurthu,A., Whisstock,J., Stuckey,P. and Lesk,A. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.

TABLE AND FIGURES LEGENDS

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

Figure 2: SP (left) and CS (right) Scores: 531 alignments computed by all methods. Programs are sorted according to the median. The colors of the boxes are: red for sequence-based alignment programs, blue for structure-based alignment programs, green for sequence+structure based alignment programs.

Figure 3. P-value heatmap of the Friedman tests Entries show the p-value computed using a Friedman rank test. Values above the diagonal where calculated with CS scores and values under the diagonal where calculated with SP scores. The programs are ordered according to their median SP scores. The colors of the boxes are: red for sequence-based alignment programs, blue for structure-based alignment programs, green for sequence+structure based alignment programs. The yellow or orange cells denote a non-significant p-value according to the 0.05 alpha risk, with a Bonferroni correction for multiple tests. The green or blue cells denote a significant p-value.

Figure 4: CS scores Heatmap and hierarchical classification of the programs and of the alignments (Complete method, Euclidian distance). The program colors are: red for sequence-based programs, blue for structure-based programs and green for sequence+ structure based programs. The symbols characterizing each program are the same as in Figure 3.

Figure 5: CS scores as a function of percentage identity of the core reference alignments

Figure 6: Median CS scores for each program. Left: core columns where the residues are either in helix or strand or loop. Right: residues either buried or exposed.

Figure 7: Median CS scores for each program and each database, restricted to alignments in the range 10-40% sequence identity. Besides, only core positions in regular secondary structures perfectly conserved in all the proteins of the family are considered. Color code is the same as in Figure 6.

Figure 8: RMS as a function of core alignments identity percentage

Figure 9: Distribution of number of alignment columns containing one or more gap opening. The colors of the boxes are: red for sequence-based alignment programs, blue for structure-based alignment programs, green for sequence+structure based alignment programs, grey for reference alignments.

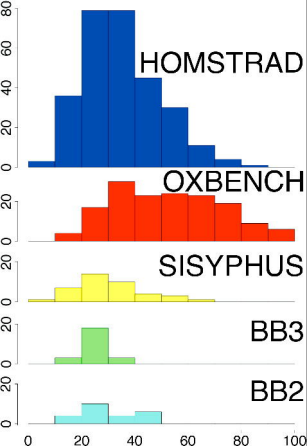
Table 1: Programs used in this study to align families of proteins from the reference datasets. Categories of programs: SEQ is a sequence-based alignment method; STRUCT is a structure-based alignment method; SEQ/STRUCT is a sequence+structure based program. DP=Dynamic Programming, AFP=Aligned Fragment Pairs, SSE=secondary structure element.

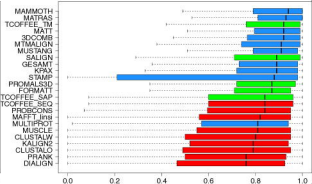
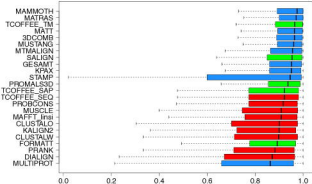
Type	Name	Description	Version	Ref.	Year
SEQ	CLUSTALO	Seeded guide trees and HMM profile-profile	1.2.0	(52, 53)	2010
SEQ	CLUSTALW	Classical progressive aligner	2.1	(54, 55)	1994
SEQ	DIALIGN	Greedy and progressive approaches for segment-based multiple alignment	TX, 1.0.2	(56–58)	1998
SEQ	KALIGN2	Wu-Manber string-matching algorithm, to improve both accuracy and speed	2.04	(59, 60)	2005
SEQ	MAFFT_linsi	Fast progressive aligner with iteration and refinement using consistency score	7.215	(61, 62)	2002
SEQ	MUSCLE	Fast progressive aligner with iteration and refinement	3.8.31	(32, 63)	2004
SEQ	PRANK	Phylogeny-aware progressive aligner; correct treatment of insertions	v.100701	(64)	2005
SEQ	PROBCONS	Probabilistic variant of the consistency algorithm	1.12	(42)	2005
SEQ	TCOFFEE_SEQ	Consistency-based progressive aligner	11.00.8c be486	(65)	2000
SEQ/ STRUCT	PROMALS3D	Derives constraints through structure-based alignments; combines them with sequence constraints to construct MSAs	NA	(66, 67)	2008
SEQ/ STRUCT	TCOFFEE_SAP	TCOFFEE + pairwise structure alignments by SAP	11.00.8c be486	(68, 69)	2004
SEQ/ STRUCT	TCOFFEE_TM	TCOFFEE + pairwise structure alignments by TMALIGN	11.00.8c be486	(68, 70)	2004
SEQ/ STRUCT	SALIGN	DP with a score that is a sum of an affine gap penalty and terms dependent on various sequence and structure features	Modeler version: 9.18	(71)	2007
SEQ/ STRUCT	FORMATT	MATT with sequence information	1.02	(72)	2005
STRUCT	3DCOMB	Identify structurally similar pairwise fragments (TM-Score (70)) + assembly according to pivot structures	1.06	(73)	2011
STRUCT	GESAMT	Clustering of small structurally similar pairwise fragments (Q-Score (74)) + Alignment refinement	7.0	(75, 76)	2012
STRUCT	KPAX	DP with a Gaussian structural similarity score + alignment optimization	5.0.5	(77)	2005
STRUCT	MAMMOTH	Alignment of small structurally similar pairwise fragments by dynamic programming.	NA	(78)	2005

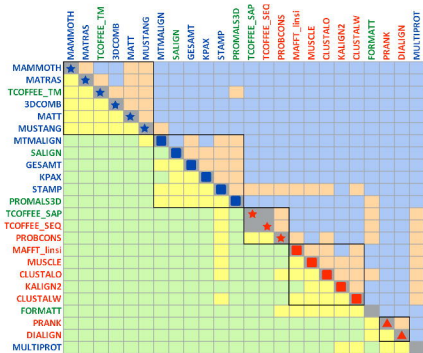
		<i>Progressive multiple alignment with a guide tree.</i>			
STRUCT	MATRAS	<i>Progressive multiple alignment (guide tree) by dynamic programming with scores depending on PAM like matrices computed on SSE conservation or Ca internal distances.</i>	1.2	(79, 80)	2000
STRUCT	MATT	<i>AFPs chaining by DP with a score allowing flexibility (translations and twists).</i>	1.0	(81)	2008
STRUCT	MTMALIGN	<i>Progressive multiple alignment (guide tree) by DP with TM-Score (70)</i>	20171124	(82)	2017
STRUCT	MULTIPROT	<i>With each structure as a pivot, detection of all AFPs, assembly to build the longest consistent alignment.</i>	1.93	(83)	2004
STRUCT	MUSTANG	<i>AFP identification and pairwise alignment to build a guide tree. Progressive multiple alignment. Score=Ca internal distance (DALI like)</i>	3.2.3	(84)	2005
STRUCT	STAMP	<i>Iterative superposition and alignment of Ca by DP with a guide tree.</i>	4.4	(38)	1992

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

	<u>ALL</u>	<u>BB2</u>	<u>BB3</u>	<u>OXBENCH</u>	<u>HOMSTRAD</u>	<u>SISYPHUS</u>
MATRAS	846 100,0%	29 100,0%	38 100,0%	330 100,0%	357 100,0%	93 100,0%
TCOFFEE(SAP/TM)	846 100,0%	29 100,0%	38 100,0%	330 100,0%	357 100,0%	93 100,0%
KPAX	845 100,0%	29 100,0%	38 100,0%	330 100,0%	357 100,0%	93 100,0%
PROMALS3D	845 99,9%	29 100,0%	38 100,0%	330 100,0%	357 100,0%	92 98,9%
GESAMT	840 99,3%	28 96,6%	38 100,0%	330 100,0%	352 98,6%	93 100,0%
3DCOMB	838 99,1%	29 100,0%	38 100,0%	326 98,8%	356 99,7%	90 96,8%
MTMALIGN	836 98,8%	29 100,0%	38 100,0%	330 100,0%	356 99,7%	84 90,3%
FORMATT	831 98,2%	29 100,0%	38 100,0%	328 99,4%	354 99,2%	83 89,2%
STAMP	825 97,5%	29 100,0%	38 100,0%	330 100,0%	357 100,0%	72 77,4%
MATT	823 97,3%	29 100,0%	38 100,0%	322 97,6%	354 99,2%	81 87,1%
MUSTANG	813 96,1%	29 100,0%	38 100,0%	326 98,8%	350 98,0%	71 76,3%
SALIGN	802 94,8%	29 100,0%	37 97,4%	318 96,4%	345 96,6%	74 79,6%
MULTIPROT	761 90,0%	28 96,6%	38 100,0%	255 77,3%	357 100,0%	84 90,3%
MAMMOTH	<u>621 73,4%</u>	<u>24 82,8%</u>	<u>24 63,2%</u>	<u>212 64,2%</u>	<u>306 85,7%</u>	<u>56 60,2%</u>
#Alignments	846	29	38	330	357	93



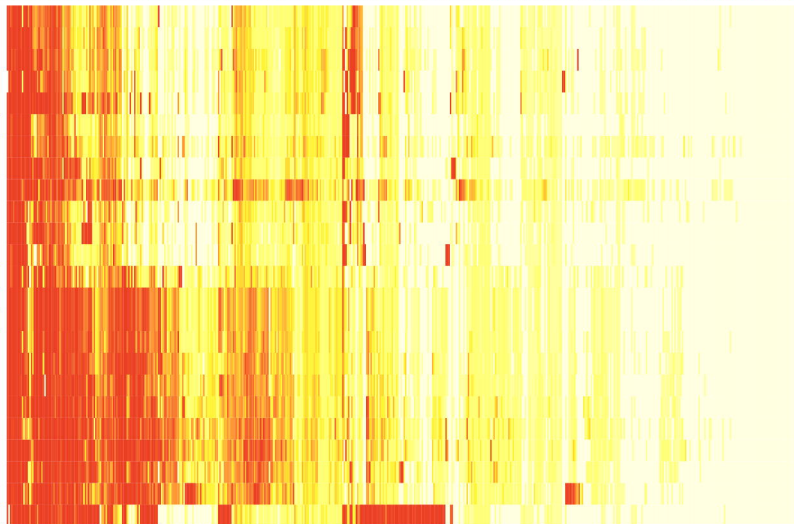
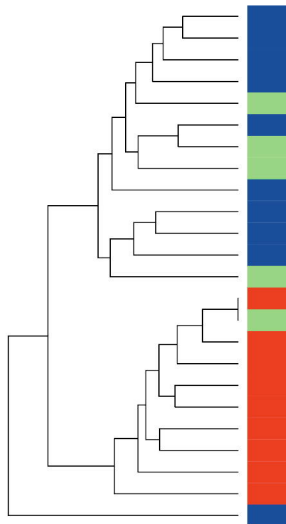
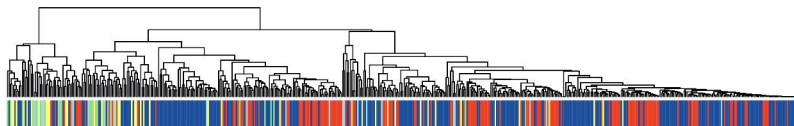
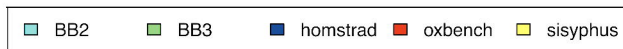
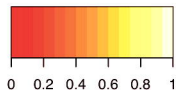




SP TC

$p\text{-value} \leq 1.89E-5$

$p\text{-value} > 1.89E-5$



- ★ 3DCOMB
- ★ MTALIGN
- ★ KPAX
- ★ GESAMT
- ★ SALIGN
- ★ MATT
- ★ FORMATT
- ★ TCOFFEE_TM
- ★ MULTIPROT
- ★ MUSTANG
- ★ MAMMOTH
- ★ MATRAS
- ★ PROMALS3D
- ★ TCOFFEE_SEQ
- ★ TCOFFEE_SAP
- ★ PROBCONS
- ★ MAFFT_linsi
- ★ MUSCLE
- ★ KALIGN2
- ★ PRANK
- ★ DIALIGN
- ★ CLUSTALW
- ★ CLUSTALO
- ★ STAMP

