Bioinformatics, 38(22), 2022, 5007–5011
https://doi.org/10.1093/bioinformatics/btac625
Advance Access Publication Date: 21 September 2022
Original Paper

Sequence analysis

Highly significant improvement of protein sequence alignments with AlphaFold2

Athanasios Baltzis [1,†], Leila Mansouri[1,†], Suzanne Jin[1], Björn E. Langer[1], Ionas Erb[1], Cedric Notredame[1,2,*]

[1]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Spain and [2]Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Pier Luigi Martelli

Received on July 7, 2022; revised on August 29, 2022; editorial decision on September 6, 2022

Abstract

Motivation: Protein sequence alignments are essential to structural, evolutionary and functional analysis, but their accuracy is often limited by sequence similarity unless molecular structures are available. Protein structures predicted at experimental grade accuracy, as achieved by AlphaFold2, could therefore have a major impact on sequence analysis.

Results: Here, we find that multiple sequence alignments estimated on AlphaFold2 predictions are almost as accurate as alignments estimated on experimental structures and significantly closer to the structural reference than sequence-based alignments. We also show that AlphaFold2 structural models of relatively low quality can be used to obtain highly accurate alignments. These results suggest that, besides structure modeling, AlphaFold2 encodes higher-order dependencies that can be exploited for sequence analysis.

Availability and implementation: All data, analyses and results are available on Zenodo (https://doi.org/10.5281/zenodo.7031286). The code and scripts have been deposited in GitHub (https://github.com/cbcrg/msa-af2-nf) and the various containers in (https://cloud.sylabs.io/library/athbaltzis/af2/alphafold, https://hub.docker.com/r/athbaltzis/pred).

Contact: cedric.notredame@crg.eu

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

1 Introduction

Protein multiple sequence alignment (MSA) is the most widely used modeling technique in biology (Van Noorden *et al.*, 2014). Its many applications include structural, functional and evolutionary analyses (Mistry *et al.*, 2021; Spence *et al.*, 2021). Their computation typically relies on amino-acid substitution matrices and only achieves sufficient levels of accuracy when comparing sequences that are more than 20% identical (Rost, 1999). Alignments based on structural comparisons are, by contrast, much less sensitive to low sequence identity levels and have routinely been employed as standards of truth when evaluating sequence alignment algorithms (Thompson *et al.*, 1999). There remains, however, some discussion as to whether structure-based sequence alignments lend themselves to support accurate phylogenetic reconstruction. Indeed, compelling evidence suggests that non-structural signals such as indels (Dessimoz and Gil, 2010) or non-structurally supported sequence alignments (Nute *et al.*, 2019) provide stronger evolutionary support than structure-based sequence alignments. In this context, having access to unlimited amounts of high-quality structural data would obviously expand the usability of sequence alignments for any application involving homology modeling, but it will also make it possible to power any analysis aimed at quantifying the capacity of structural comparisons to support phylogenetic analysis.

While the scarcity of experimental structural data currently limits the use of structure-based sequence alignments, recent results indicate that inferring protein structure using deep learning techniques is becoming increasingly effective, with claims that AlphaFold2 (AF2) (Jumper *et al.*, 2021) can predict structures at angstrom-level accuracy for most proteins (Porta-Pardo *et al.*, 2022; Tunyasuvunakool *et al.*, 2021; Varadi *et al.*, 2022). Furthermore, other work indicates that predicted protein folding code elements can be used to improve sequence homology detection (Gao and Skolnick, 2021). Given these advances, we asked whether AF2 models can be used to compute highly accurate protein MSAs.

2 Materials and methods

2.1 Reference datasets

We assembled a structure-based dataset suitable for AF2 (Jumper *et al.*, 2021) predictions, highly discriminative and independent of any AF2 training set. This was achieved by collecting all the 23 300 Protein Data Bank (PDB) (Berman *et al.*, 2000) entries (07/2020 release) with a release date posterior to AF2 (04/2018). These entries were assigned to PFAM (Finn *et al.*, 2016) families (release 28) using HMMER3/f (Version 3.1b1 | May 2013) with default parameters (E-value threshold set at 10.0). This procedure gave rise to 39 476 domain segments longer than 80 AA that were turned into a non-redundant dataset featuring 3419 segments [CD-HIT (Fu *et al.*, 2012; Li and Godzik, 2006) version 4.8.1, threshold 70% sequence identity], each unambiguously assigned to a PFAM family. We grouped all the segments labeled with the same PFAM family and kept the 31 datasets featuring 10 or more segments (461 segments in total). Some of these datasets were poorly informative because of the high similarity level between the sequences. In order to build the most compact and discriminative reference dataset, we therefore selected the families for which the agreement between the sequence and structure-based MSAs was lower than 75%. This procedure generated a total of 15 datasets. We also removed two datasets featuring sequences longer than 400 amino acids and one dataset containing a multi-domain protein. On top of this, we discarded two segments from the PF00520 dataset whose very short length suggests truncation. Overall this procedure led to a test set featuring 153 PDB chains belonging to 12 distinct PFAM families and comprising between 8 and 16 sequences each (Supplementary Table S1). The complete list of all the selected segments, including the discarded ones, is available from Zenodo (https://doi.org/10.5281/zenodo.7031286).

2.2 Sequence and structure-based MSAs

On the basis of recent benchmarks (Carpentier and Chomilier, 2019; Deorowicz *et al.*, 2016), five different sequence-based MSA algorithms [FAMSA v.1.6.218 (Deorowicz *et al.*, 2016), MAFFT G-INS-i v.7.45319 (Katoh *et al.*, 2005), MSAProbs v.0.9.720 (Liu *et al.*, 2010), T-Coffee v.13.45.57.844f401 (Notredame *et al.*, 2000), PSI-Coffee v.13.45.57.844f401 (Kemena and Notredame, 2009)] and three structure-based MSA algorithms [3D-Coffee_SAP+TMalign v.13.45.57.844f401 (O'Sullivan *et al.*, 2004), 3D-Coffee_TMalign v.13.45.57.844f401 (O'Sullivan *et al.*, 2004) and mTM-align v.20180725 (Dong *et al.*, 2018)] were tested.

The sequences of the PDB-derived structures were multiply aligned using either structure-based or sequence-based alignment algorithms. The best algorithms for sequences or structures respectively were selected on the basis of their average structural correctness using normalized intra-molecular root mean squared deviation (NiRMSD, Supplementary Table S4). NiRMSD is a normalized distance-based-RMSD (Armougom *et al.*, 2006) that compares intramolecular distance variations to assess the structural correctness of sequence alignments. 3D-Coffee_SAP+TMalign, the best-performing structure-based MSA algorithm, was used as a reference method to perform all the structural MSAs using either PDB or AF2 structures (MSA-PDB and MSA-AF2, respectively). The best sequence-based MSA alignment algorithm was MAFFT G-INS-i which was used for MSA-Seq. The PSI-Coffee alignments (MSA-PSI) were assembled using the psicoffee mode of T-Coffee (v.13.45.57.844f401, -mode=psicoffee) where we fetched for each input sequence the 100 most informative homologs by conducting a BLAST search against UniRef50. The collected BLAST alignments were stacked onto the corresponding query sequence, such that the result could be used as profile templates when generating the T-Coffee pairwise libraries (Chang *et al.*, 2012) used to produce a regular T-Coffee MSA.

2.3 *De novo* structure prediction and evaluation

Structure predictions were carried out using AlphaFold2 (AF2) v.2.0.0 (Jumper *et al.*, 2021) with the default parameters:

-full_dbs preset -max_template_date 2020-05-14

AF2 provides for each predicted structure a reliability index named predicted local distance difference test (pLDDT) that reports the predicted fraction of preserved local distances between the C_α atoms of all the residues. Furthermore, the correctness of the predicted structures was estimated via the Global Distance Test Total Score [GDT-TS, TM-Score (Zhang and Skolnick, 2004) package], one of the standard CASP measures. Given an experimental structure (PDB-derived) and its corresponding prediction, GDT-TS reports the fraction of C_β (C_α in case of glycines) superimposed below different distance thresholds after a rigid superposition.

2.4 Alignment accuracy measures

The sequence-based alignments (MSA-Seq, MSA-PSI) and the structure-based alignments estimated on AF2 predicted models (MSA-AF2) were compared with the structure-based MSAs computed on experimental structures (MSA-PDB) and evaluated using the sums of pairs (SoP) measure as implemented in the T-Coffee package (Version_13.45.57.844f401). The SoP reports the fraction of aligned pairs occurring in the reference MSA that are recovered in the evaluated one. The SoP can be estimated on complete MSAs as well as pairwise projections (i.e. pairs of sequences extracted from the MSA they are part of while retaining the gaps). Here, the SoP scores used in the main analysis were computed only on those residues that were assigned to the same structural state by DSSP (Kabsch and Sander, 1983) while excluding loops. For the sake of completeness, the equivalent SoP scores computed on all the pairs of residues are also provided and yield a similar trend to the one measured on residues of identical structural categories (Supplementary Table S1a and b).

2.5 Transitive consistency score

The transitive consistency score (TCS) (Chang *et al.*, 2014) estimates the level of agreement between the alignment of any pair of sequences in an MSA and the realignment of this same pair across all possible sequence triplets within the same dataset. Triplets are estimated by combining all pairs of pairwise alignments featuring a common sequence. The method has been extensively described in Chang *et al.* (2014) and has shown to be a robust predictor of sequence alignment accuracy. Given an MSA, the pairwise alignments required for the triplet analysis are generated by separately applying to every pair of sequences the same method used to generate the MSA being evaluated. This procedure is supported by the T-Coffee package in version v.13.45.57.844f401 and can be invoked by the following commands:

t_coffee -infile <MSA-Seq> -evaluate -method mafftginsi_pair

t_coffee -infile <MSA-AF2> -evaluate -method sap_pair, TMalign_pair -pdb_dir <AF2>

2.6 Kendall correlation analysis

Kendall correlation analysis quantifies the concordance between two different measures evaluated on the same samples. It estimates the frequency with which any increase of one measure is associated with an increase or decrease of the other measure. We applied this analysis on each of the structural accuracy measures (GDT-TS and pLDDT) and on the sequence accuracy predictor (TCS) in order to explore the agreement between these three measures, and the true sequence alignment accuracy as defined by the SoP. Since our datasets are independent from one another, we only included in the analysis the differences measured across pairs within families while ignoring differences across families. Kendall's correlation (τ) coefficients were computed using:

$$\tau = \frac{C - D}{\sqrt{(C + D + Xo)(C + D + Yo)}}, \quad (1)$$

where C is the number of the concordant pairs within families, D is the number of the discordant pairs within families, Xo is the number

of pairs tied only in the first variable, and *Yo* is the number of pairs tied only in the second variable. Statistical significance was estimated by shuffling 1000 times the vector of the first variable, estimating the resulting τ, and comparing them to the true τ value. The *p*-value is then defined as the fraction of cases for which the true value was exceeded or equaled.

2.7 Computation

All computation was carried out on a cluster running Scientific Linux release 7.2. The structure predictions as well as the alignments, comparisons and evaluations were run within separate containers based on Ubuntu and Debian operating systems. The whole computational pipeline was implemented in the Nextflow language (Di Tommaso *et al.*, 2017) and was deployed in a containerized form using Singularity. We collected the time needed for the computation of all the alignments and the AF2 prediction in Supplementary Table S5.

3 Results

3.1 Alignments computation and comparisons

To validate the potential of AF2 models for the generation of accurate structure-based sequence alignments, we filtered the PDB (Berman *et al.*, 2000) (https://www.rcsb.org/) to select protein sequences that were not part of the AF2 training set (see Section 2). This highly stringent process left us with a total of 153 proteins that can be associated with 12 PFAM (Finn *et al.*, 2016) families. Sequences within each family were multiply aligned using: (i) experimental structures (MSA-PDB, 3D-Coffee_SAP+TMalign); (ii) predicted structures (MSA-AF2, 3D-Coffee_SAP+TMalign); or (iii) sequence information only (MSA-Seq, MAFFT G-INS-i). We used MSA-PDB as a reference to assess the alignment accuracy of every pair of sequences within MSA-Seq and MSA-AF2. Accuracy was estimated using the SoP measure, which reports the fraction of pairs of residues identically aligned across two alternative alignments of the same sequences. Our results indicate that MSA-AF2 is clearly more similar to the structural reference than sequence-based alignments, with 78% of the sequence pairs being more accurately aligned in MSA-AF2 than in MSA-Seq (Fig. 1). Across all MSAs, the difference is highly significant, with an average improvement by 23.62 percent points of MSA-AF2 over MSA-Seq (93.95% versus 70.33% and Wilcoxon test *p*-value = 0.0002, Table 1, Supplementary Table S1a). This difference is particularly clear in datasets containing repeated elements, a feature known to compromise most sequence alignment procedures (e.g. PF13306, Supplementary Fig. S1). The high levels of accuracy measured on MSA-AF2 alignments confirm that the quality of these alignments is comparable to their reference (Carpentier and Chomilier, 2019). These results support the notion that AF2 predicted structures could be systematically used to replace sequence information.

3.2 Alphafold2 performance on alignment accuracy

To determine whether AF2 performance is the main driver of sequence alignment accuracy (Supplementary Fig. S2), we compared sequence alignment and structural prediction accuracies. We did so by measuring the alignment score of every pair of sequences as they appear in their respective MSAs (SoP) versus the average correctness of the corresponding AF2 structures computed as the percentage of Cα distances under definite distance thresholds after superposing every experimental structure with its corresponding prediction (GDT-TS geometric mean). The results paint a more complex picture than anticipated. Many aligned pairs achieve high alignment accuracy despite being based on AF2 models scoring lower than the 75% GDT-TS threshold required for atomic resolution precision (Kryshtafovych *et al.*, 2019). Specifically, 80.7% of such pairs give rise to AF2-based alignments that are more similar to the reference than their sequence-based counterparts. Altogether, these pairs of lower correctness structures occur in 18.3% of all the 938 possible pairs of sequences (Supplementary Table S2). Given that the

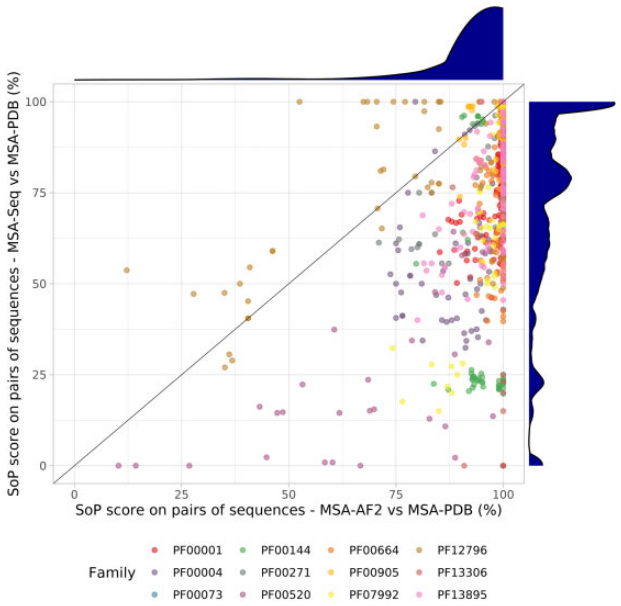

Fig. 1. Comparison of alignment accuracy of MSAs based on sequence information (MSA-Seq) and on predicted structural information (MSA-AF2). Accuracy is measured against the reference MSAs based on experimental structural data (MSA-PDB) using the sums-of-pairs (SoP) measure estimated of every pair of sequences within the respective MSA. Points are colored by dataset and the marginal density plots represent their distribution across the considered axis (A color version of this figure appears in the online version of this article.)

structure-based alignments of sequence pairs rely only on Cα superpositions and ignore sequence information (see Section 2), it is unclear why so many incorrect AF2 models contribute to highly accurate sequence alignments.

One possible explanation is that some of the AF2 predictions, albeit different from the experimental reference (i.e. low GDT-TS), may preserve amino-acid homologous relationships across protein family members. This preservation could make the AF2 structures easier to superpose and would therefore lead to more consistent alignments (see Section 2). We quantified this effect using the TCS, an alignment reliability index based on the comparison between the alignment of a pair of sequences in the MSA and the realignment of this same pair across all possible sequence triplets within the same dataset (Chang *et al.*, 2014). Our results indicate that MSA-AF2 alignments featuring low GDT-TS sequences can display a wide range of TCS scores (gradient overlay Fig. 2). Overall the 22.8% pairs with a GDT-TS lower than 75% have TCS scores ranging between 54.5% and 95.5%. The very existence of low GDT-TS/high TCS alignments confirms that some of the incorrectly predicted structures are sufficiently superposable to yield consistent alignments. Their relatively high accuracy shows the benefits of such predictions for sequence analysis.

3.3 Predicting alignment accuracy

These results appear to be at odds with the expectation that structural correctness measures such as GDT-TS and the AF2 prediction reliability index, pLDDT, should be the best predictors of accuracy for alignments evaluated against structure-based references. To analyze this effect further, we compared the capacity of GDT-TS, pLDDT and TCS to discriminate between high and low accuracy alignments. We did so by quantifying the average SoP of each sequence within its MSA and by estimating its concordance with each of the three measures. When running a Kendall correlation analysis within families only (see Section 2), we observed considerably less discordance between TCS and SoP (21.43% of discordant pairs, Kendall $\tau = 0.3$, Fig. 3a) than for GDT-TS (33.48%, $\tau = 0.13$, Fig. 3b) or pLDDT (32.73%, $\tau = 0.15$, Fig. 3c) (Supplementary Table S3). This indicates that when aligning AF2 structures,

Table 1. Average alignment length, percent identity (PID) with its standard deviation, SoP, TCS and NiRMSD scores per MSA algorithm over the 12 datasets

MSA algorithm	Average alignment length	Average PID (%)	SD PID	SoP (%)	TCS (%)	NiRMSD (A)
MSA-Seq	292.67	23.61	4.98	70.33	53.08	2.00
MSA-PSI	299.50	21.23	5.08	77.30	70.97	1.77
MSA-AF2	292.33	21.31	5.23	93.95	83.43	1.52
MSA-PDB	295.42	21.26	5.05	100.00	83.17	1.46

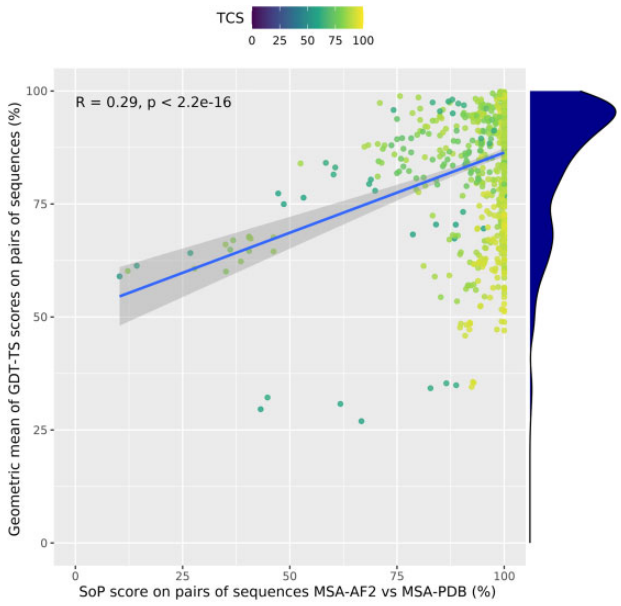

Fig. 2. Comparison between the geometric mean of the model structural correctness (GDT-TS) and alignment accuracy (SoP) for every pair of aligned sequences within the MSA-AF2 MSAs. Points are colored by the geometric mean of the predicted sequence accuracy (TCS) for each pair. The marginal density plot represents the distribution of the number of pairs along the considered axis. The blue line represents the Pearson correlation ($R = 0.29$ and p-value $= 2.2e{-}16$) as provided by the ggpubr package

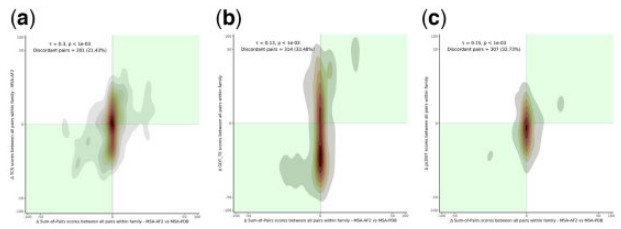

Fig. 3. Log/log representation of discordance between predicted sequence alignment accuracy (TCS) (**a**), model structural accuracy (GDT-TS) (**b**) or predicted model structural accuracy (pLDDT) (**c**) and the equivalent differences in sequence accuracy (SoP) scores when comparing MSA-AF2 against MSA-PDB. The green-shaded areas correspond to the first and third quadrants. The relationship between the two variables is indicated by a Kendall correlation coefficient (τ) and its assigned p-value. The text label also includes the number and percentage of discordant pairs over all the given pairs

consistency, as estimated by the TCS, is a much better predictor of alignment accuracy than the estimators of structural correctness.

4 Discussion

The high accuracy of the MSA-AF2 alignments, regardless of the correctness of their predicted structures, may also indicate the AF2 algorithm's capacity to integrate structural and evolutionary data. Indeed, the training of AF2 involves combining the information content of experimentally determined structures along with MSAs consisting of large compilations of naturally occurring mutations selected for maintaining a functional fold. Evolutionary information is well established as a means to improve sequence alignments (Kemena and Notredame, 2009; Pei and Grishin, 2007). For instance, PSI-Coffee encompasses evolutionary information by replacing every sequence with a position-specific scoring scheme, often referred to as a profile. This provides a powerful way to average information content across protein family members, but it discards any information associated with covariation between sites. When we applied the PSI-Coffee profile-based approach to our dataset, it systematically produced MSAs of intermediate accuracy between sequence and AF2-based alignments (Table 1, Supplementary Table S1). This observation suggests that the way AF2 models incorporate evolutionary information possibly accounts for some aspects of covariation between sites.

We showed here that the structural models generated using AF2 lead to alignments closer to the structural reference than the ones achieved using the best profile-based solution (Nute *et al.*, 2019). We hypothesize that this improvement results from deep networks identifying higher-order relationships among sites thus allowing for more accurate alignments than simpler methods like PSI-Coffee. It is worth pointing out that our method was only tested on small datasets of single domain families due to the limitation imposed by excluding the AlphaFold2 training set. There is, however, no reason to suspect the results would not scale with a larger number of sequences. Limitations, if any, would most likely arise from the computational cost of consistency-based MSAs that are currently limited to a few hundred sequences. Furthermore, even though our benchmark suggests a slightly improved accuracy for consistency-based MSAs that rely on structures (i.e. 3D-Coffee), one could also deploy non-consistency-based aligners that are expected to scale better on large datasets, a prospect very likely to be made possible thanks to the expansion of the AlphaFold Protein Structure Database (Varadi *et al.*, 2022). The question remains, however, of how well these methods would perform on multi-domain proteins. The scarcity of PDB multi-domain chains makes it challenging to propose a quantified benchmark, and in the current state of the art, one should probably restrict the application of these methods to proteins processed into single-domain peptides. This situation may soon change as an increasing amount of protein complex data is acquired using next-generation structure determination methods such as cryogenic electron microscopy.

Acknowledgements

We thank Guy Riddihough for revisions and comments on the manuscript.

Author contributions

C.N., A.B. and L.M. designed the analysis, A.B. and L.M. carried out the validation and all the authors designed the validation procedure and wrote the manuscript.

Funding

This project was supported by the Centre for Genomic Regulation and the Spanish Plan Nacional and the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa'.

Conflict of Interest: none declared.

References

Armougom,F. *et al.* (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–e39.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Carpentier,M. and Chomilier,J. (2019) Protein multiple alignments: sequence-based versus structure-based programs. *Bioinformatics (Oxf. Engl.)*, **35**, 3970–3980.

Chang,J.-M. *et al.* (2012) Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *BMC Bioinformatics*, **13 Suppl 4**, S1.

Chang,J.-M. *et al.* (2014) TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.*, **31**, 1625–1637.

Deorowicz,S. *et al.* (2016) FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.*, **6**, 33964.

Dessimoz,C. and Gil,M. (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.*, **11**, R37.

Di Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.

Dong,R. *et al.* (2018) mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, **34**, 1719–1725.

Finn,R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–285.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxf. Engl.)*, **28**, 3150–3152.

Gao,M. and Skolnick,J. (2021) A novel sequence alignment algorithm based on deep learning of the protein folding code. *Bioinformatics (Oxf. Engl.)*, **37**, 490–496.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

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

Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics (Oxf. Engl.)*, **25**, 2455–2465.

Kryshtafovych,A. *et al.* (2019) Critical assessment of methods of protein structure prediction (CASP)-round XIII. *Proteins*, **87**, 1011–1020.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Liu,Y. *et al.* (2010) MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, **26**, 1958–1964.

Mistry,J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.

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

Nute,M. *et al.* (2019) Evaluating statistical multiple sequence alignment in comparison to other alignment methods on protein data sets. *Syst. Biol.*, **68**, 396–411.

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

Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics (Oxf. Engl.)*, **23**, 802–808.

Porta-Pardo,E. *et al.* (2022) The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput. Biol.*, **18**, e1009818.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Spence,M.A. *et al.* (2021) A comprehensive phylogenetic analysis of the serpin superfamily. *Mol. Biol. Evol.*, **38**, 2915–2929.

Thompson,J.D. *et al.* (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics (Oxf. Engl.)*, **15**, 87–88.

Tunyasuvunakool,K. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.

Van Noorden,R. *et al.* (2014) The top 100 papers. *Nature*, **514**, 550–553.

Varadi,M. *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.