Ankh-score produces better sequence alignments than AlphaFold3

Julia Malec[a], Karina Rusen[b], G. Brian Golding[c], Lucian Ilie[a,*]

[a]*Department of Computer Science, University of Western Ontario, London, N6A 5B7, Ontario, Canada*
[b]*School of Rehabilitation Science, McMaster University, Hamilton, L8S 4L8, Ontario, Canada*
[c]*Department of Biology, McMaster University, Hamilton, L8S 4K1, Ontario, Canada*

Abstract

Protein sequence alignment is one of the most fundamental procedures in bioinformatics. Due to its many downstream applications, improvements to this procedure are of great importance. We consider two revolutionary concepts that emerged recently as candidates for improving the state-of-the-art alignment methods: AlphaFold and protein embeddings. Alignment improvements can come from structural alignment of AlphaFold-predicted structures or scoring based on similarity of protein embedding vectors, resp. Thorough comparison on many domains from BAliBASE and CDD demonstrates that the Ankh-score method produces much better sequence alignments than structural alignments using US-align of AlphaFold3-predicted structures. Both are better than the traditional method using BLOSUM matrices. This suggests that Ankh embedding vectors may possess certain information that is not available in the AlphaFold3-predicted structures.

Keywords:
Protein Sequence Alignment, AlphaFold, Protein Embedding, Ankh, US-align

*Corresponding author: `ilie@uwo.ca`

1. Introduction

Protein sequences contain information necessary for the many functions they perform, which makes protein sequence alignment one of the most fundamental procedures in bioinformatics. Improvements to the quality of alignments are essential for their many downstream applications. Two concepts emerged recently as candidates for improving the state-of-the-art alignment methods. The first is AlphaFold, which, through its high quality predicted structures, can produce sequence alignments via alignment of the structures. The second is protein embeddings, which can help produce improved amino acid scoring schemes. Our goal is to provide a thorough comparison of these methods as well as the traditional BLOSUM matrices.

Two revolutions have been undergoing in bioinformatics and biology in recent years. The first one is AlphaFold [14], which produced a much better solution to the fundamental problem of structure prediction, accelerating drug discovery, enzyme engineering, protein-protein interaction studies, etc. AlphaFold illuminated half of the dark human proteins [3]. The second one is protein embeddings, such as ProtT5 [7], ESM2 [26], or Ankh [6], which provide contextual representations that capture functional and evolutionary patterns, helping to create state-of-the-art solutions for many problems in proteomics, such as structure prediction [27, 29, 14], function prediction [16, 8, 17], interaction site prediction [20, 12], etc. AlphaFold itself uses a form of embedding (Evoformer) to make its own predictions. We shall use Ankh embeddings because of thier superior performance for alignment computation compared with the other embedding methods [19].

Because evolutionary pressure is higher on the structure than it is on the sequence, Rajapaksa et al [25] claimed that structure alignments are significantly more reliable than sequence-only alignments. Lesk and Konagurthu [18] suggest that applying AlphaFold to the sequences and aligning the predicted structures produces better sequence alignments than traditional methods based on the Needleman-Wunsch algorithm [23] with affine gap penalties [9]. They argue that "structure-based methods will for most cases supersede classical sequence-only-based methods for the central and fundamental problem of alignment."

On the other hand, the cosine similarity of embedding vectors associated with amino acid residues has been used recently as a successful scoring scheme for alignment computation by several studies: pLM-BLAST [15], vcMSA [22], PEbA [13] and E-score [2, 19]. In particular, PEbA and E-score show that

embedding-based scoring schemes produce much better pairwise alignments than traditional BLOSUM matrices.

Thus, we have two very promising candidates for replacing the traditional methods for computing sequence alignments. We present here a thorough comparison of the three methods: BLOSUM matrices [11], AlphaFold3 [1] (with US-align [31]), and Ankh-score [2, 19]. We use a wide variety of alignments selected from the BAliBASE database [28] and the Conserved Domain Database [21] and employ four different distance metrics to compare the results with the reference alignments.

The conclusion is that Ankh-score is clearly the best, followed by AlphaFold3 and then the traditional BLOSUM matrices. We provide also examples showing the superiority of the information provided by the contextual embedding vectors over the alignment of AlphaFold3-predicted structures.

In addition to identifying the best method for computing protein sequence alignments, we have two important observations. First, it appears that Ankh embedding vectors may possess certain information that is not available in the AlphaFold3-predicted structures. Second, our testing appears to indicate that the sequence alignments produced from experimentally determined structures are slightly inferior to those derived using AlphaFold-predicted structures, an intriguing hypothesis that remains to be investigated.

2. Material and Methods

2.1. Alignment methods

We compare three methods for computing protein sequence alignments. The first is dynamic programming [23] with affine gap penalties [9] using the score provided by the BLOSUM matrices [11]. The common matrices are considered: BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, and BLOSUM90. For most tests, BLOSUM45 gives the best results and, therefore, most of the analysis focuses on this one out of the five matrices.

The second method is the sequence alignment induced by aligning the structures predicted by AlphaFold3 [1] using US-align [31, 30]; we shall call this method *AF3US*. The idea is that the residues from the AlphaFold3-predicted structures that are aligned to be in close proximity by US-align become aligned as well in the sequence alignment.

US-align, the successor of TM-align [31] reports two TM-scores, one for each protein sequence, as a measure of structural similarity. The TM-score is length independent. A score above 0.5 suggests very good alignment, the two

structures sharing the same overall fold, while a score below 0.2 suggests that the proteins are likely unrelated. If two proteins of very different length are aligned, then the TM-score for the longer one will always be low, reflecting the unaligned remainder of the larger structure. In this case, it is possible that the TM-score for the shorter one is very high, meaning that the shorter protein aligns very well with a part of the longer one.

The third method employs also dynamic programming with affine penalties but uses for scoring the similarity of Ankh embedding vectors [6]. Precisely, the Ankh-score [2, 19] for two residues a_1 and a_2 is given by the cosine similarity between the vectors $v_i = Ankh(a_i)$, $i = 1, 2$, computed by Ankh for the two residues, that is:

$$Ankh\text{-}score(a_1, a_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \ .$$

2.2. Testing datasets

We have selected a wide variety of domains from the BAliBASE database [28] and the Conserved Domain Database (CDD) [21]. We attempted to cover many different identity levels, as a main criterion that influences the quality of the alignments. Figure 1 gives the list of 20 domains we use from each database, with the identity distribution and the number of sequences. More details are provided in the Supplementary Tables 1-2.

2.3. Comparison procedure

Each domain consists of a number of protein sequences, say n, and a reference multiple sequence alignment (MSA) of all n sequences. A test is any pair of sequences, $\frac{1}{2}n(n-1)$ in total for a domain with n sequences. The two sequences have a reference alignment induced by the original MSA. This is the reference alignment with which all computed alignments between the two sequences are compared.

For each method, an alignment is computed and its distance to the reference alignment is measured according to one of several distance metrics. We have used four different distances. The first one is the inter-alignment distance [25], $\mathrm{d_{ia}}$, and measures the area between the alignments paths in the dynamic programming matrix. The second one is the relative displacement distance [2], $\mathrm{d_d}$, and consists of all position differences of all residue pairs. The third one, $\mathrm{d_{cc}}$, measures the distance to the closest position with the same context [2]. The last one, $\mathrm{d_{pos}}$, comes from three distances, $\mathrm{d_{ssp}}, \mathrm{d_{seq}}, \mathrm{d_{pos}}$,

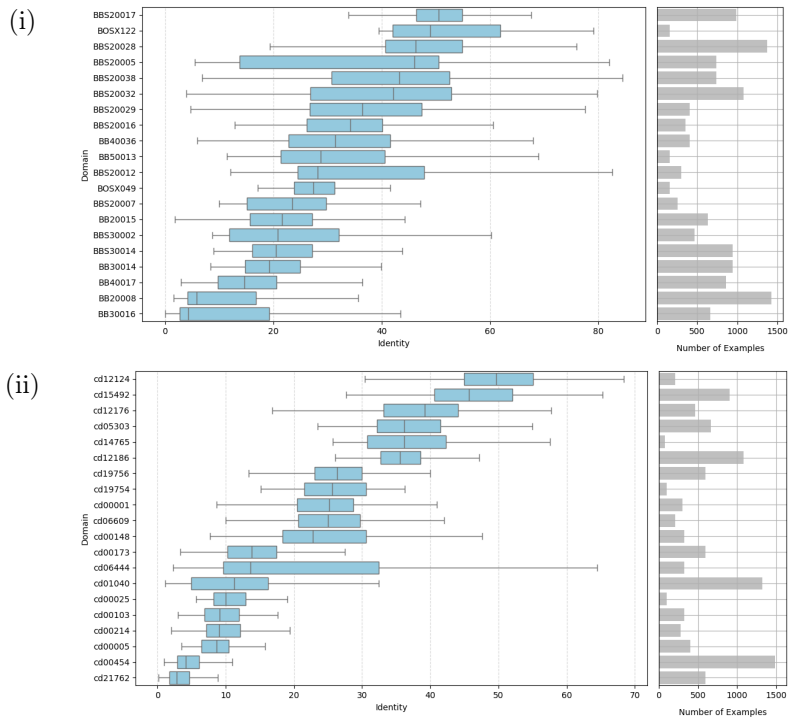


Figure 1: Testing domains from (i) BAliBASE and (ii) CDD. For each domain we plot the identity distribution as box-and-whiskers on the left and the number of sequences on the right. Further details are given in the Supplementary Tables 1-2.

which are true metrics inspired from the sum-of-pairs score, differing from each other in the way they consider gaps [4]. As d_{pos} considers the most information about gaps – the sequence and the position of the gap in the sequence – it is the most relevant of the three. In order to reduce the complexity of the comparison, we use only d_{pos} out of the three.

For each method tested on a domain and evaluated according to a distance, we report (i) the average distance from the alignments computed by the method to the reference alignments and (ii) the number of tests won by the method. The methods are compared in pairs, enabling the use of Wilcoxon signed-rank test; we consider p-values smaller than 0.01 significant.

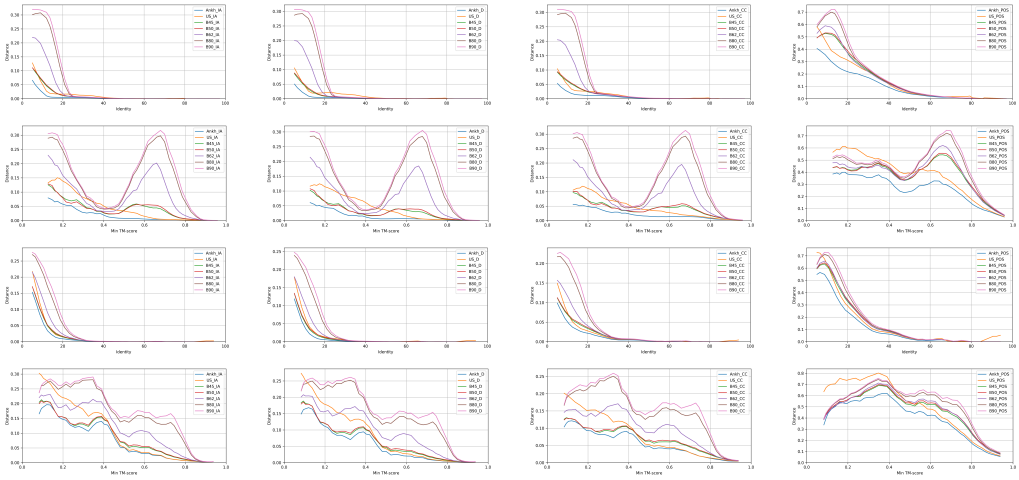


Figure 2: All methods compared on all tests. In each row, the distances used from right to left are $d_{ia}, d_d, d_{cc}, d_{pos}$, resp. The first two plots are for BAliBASE datasets, and the last two for CDD. In the first and third rows, the scores are sorted by the identity level of the sequences, while in the second and fourth, they are sorted by the minimum TM-score.

3. Results

3.1. All-against-all comparison

We first compare all methods together, separately on BAliBASE and CDD datasets. The comparison is shown in Figure 2. All methods are plotted: BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, BLOSUM90, AF3US, and Ankh-score. Each row contains four plots, corresponding, from left to right, to the distances $d_{ia}, d_d, d_{cc}, d_{pos}$, resp. The first two rows correspond to the BAliBASE datasets and the last two to the CDD datasets. In each group, there are two rows of plots. In the first of those, the tests are sorted by identity level, whereas in the second one they are sorted by minimum TM-score. Identity level and minimum TM-score are the most important parameters, as sequences with higher identity level are easier to align, and higher TM-score indicates better performance for the structural alignment.

For each plot, in order to have good control over which data is being used to draw the line, the average distance of all tests in a sliding window of fixed size is computed and the window is shifted by a fixed step. For the identity

plots, the window size is 10 and the step is 2. For the min TM-score plots, the window size is 0.1 and the step is 0.02.

The first observation, common to all plots, is that the Ankh-score alignments are always the best. The BLOSUM matrices with lower index perform better, BLOSUM45 being the best with BLOSUM50 a very close second. For that reason, from now on, when discussing the behaviour of BLOSUM matrices, we will restrict our attention exclusively to BLOSUM45.

AF3US has an interesting behaviour. It starts the worst (compared to Ankh-score and BLOSUM45) and then it gradually becomes second best, behind Ankh-score. This happens in the identity plots around identity level 10-15% and in the min TM-score plots around TM-score 0.5. After that, in the min TM-score plots it remains second best until the end, inching closer and closer to the performance of the Ankh-score algorithm. In the identity plots, the behaviour is somewhat different. The distances for all methods eventually become very close to zero, except for the AF3US, which is sometimes hovering above the abscissa, even for very high identity values. This is more pronounced in the BAliBASE plots, but visible also in the CDD plots. This behaviour may be partly explained by the tests with TM-score below 0.5, which may cause poor alignments, thus increasing the average distance. However, as we discuss below, this is not the only factor.

As structural alignments are considered reliable when the TM-score is above 0.5, we have plotted separately the results including only tests with minimum TM-score above 0.5 in Figure 3. The performance of AF3US is improved compared with the case of all tests. It now starts as second best, behind Ankh-score, from the beginning. However, the "wavering" behaviour is still present. We have provided zoom inserts to the identity plots in Figure 3 to make this more clearly visible. In the BAliBASE plots, AF3US is the only one displaying this behaviour, whereas in the CDD plots all methods appear to falter to some degree. Finally, we notice that Ankh-score alignments are still the best, even when the tests with minimum TM-score below 0.5 are removed, which greatly favours AF3US.

3.2. Head-to-head comparison

A more precise comparison is done by considering two methods at the time and computing the statistical significance, using Wilcoxon signed-rank test, of the comparison on each individual domain. We compare Ankh-score, AF3US and BLOSUM45 in pairs. Figure 4 gives a visual presentation for all comparisons in the form of dot plots. The Ankh-score dominates AF3US

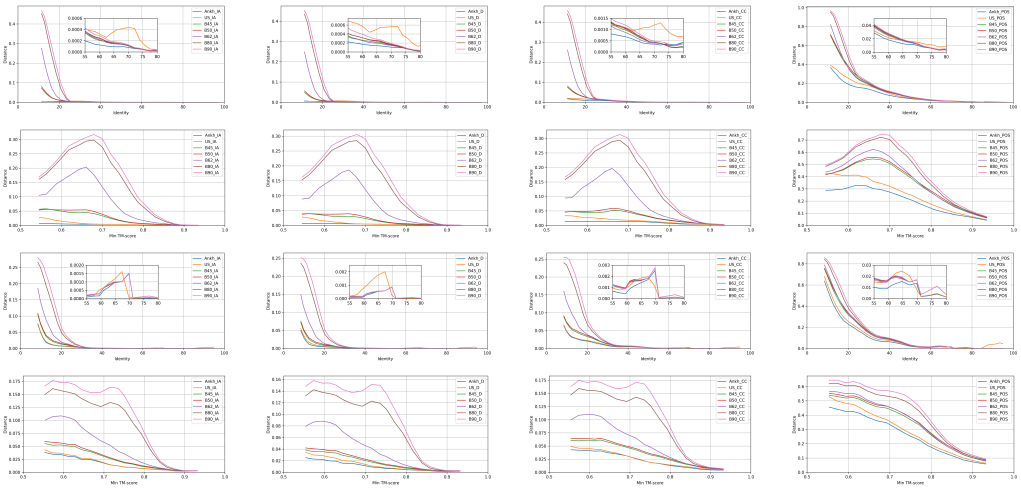


Figure 3: All methods compared on tests with minimum TM-score > 0.5. All methods are compared: BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, BLOSUM90, AF3US, and Ankh-score. In each row, the distances used from right to left are $d_{ia}, d_d, d_{cc}, d_{pos}$, resp. The first two plots are for BAliBASE datasets, and the last teo for CDD. In the first and third rows, the scores are sorted by the identity level of the sequences, while in the second and fourth, they are sorted by the minimum TM-align score.

clearly, and even more so BLOSUM45. The performance of AF3US and BLOSUM45 appear very close in the all-tests plots, but AF3US dominates BLOSUM45 in the high TM-score plots.

A summary of results when all tests are considered, irrespective of the TM-scores, is given in Supplementary Table 3. The table gives separate results for each of the four distances considered: $d_{ia}, d_d, d_{cc}, d_{pos}$, resp. The table has three sections, one for each pair of methods considered. In each case, the number of domains won and percentage of tests won is given, separately for each database, BAliBASE and CDD, that is, by summing/averaging the individual results for the domains from each database, as well as combined. When the number of domains won by each method is given, "null" represents the number of domains for which the result is not statistically significant. When the percentage of tests won is given, "equal" represents the fraction of tests for which the performance of the two methods was identical.

While different distances present slightly different results, they are consistent, revealing similar trends. Also, BAliBASE and CDD domains behave differently, but present similar pictures. Ankh-score is always winning with

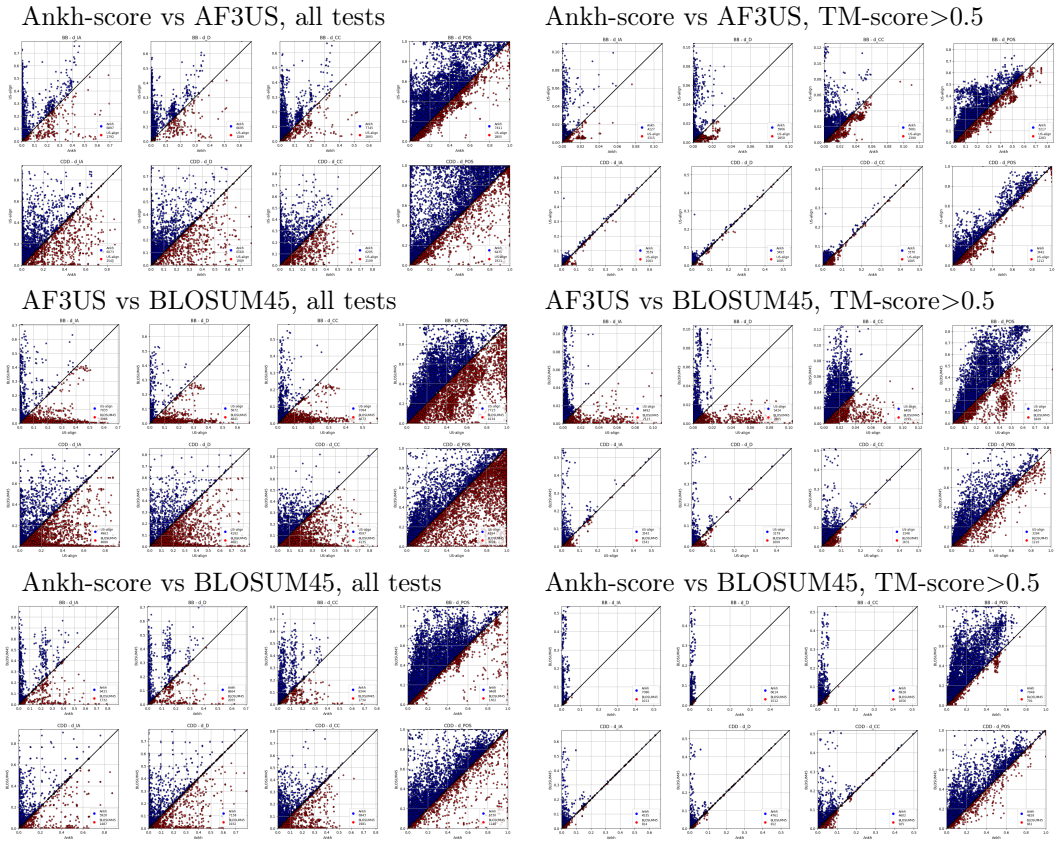


Figure 4: Dot plots comparing the three methods pairwise, as indicated in the table. In each group the first row is for BAliBASE and the second for CDD database, while the four columns correspond to the distances $d_{ia}, d_d, d_{cc}, d_{pos}$, resp.

a very wide margin. The winning margin is larger over BLOSUM45 than over AF3US. AF3US wins always over BLOSUM45 – with a single exception provided by the case of CDD domains, percentage of tests won for the d_d distance – but the winning margin is not very large. The competition is closer for the CDD domains than for BAliBASE ones.

An identically organized table but using only the tests with minimum TM-score above 0.5 is shown in Supplementary Table 4. The general trends are preserved, with AF3US performing slightly better than before, given the advantage of high TM-scores.

A "summary of summaries" is presented in Table 1. The number of "null"

tests is larger in the right half of the table, because the elimination of many tests causes the p-values to increase. The general trends mentioned above are clearly visible. Ankh-score wins 78.75% of domains against AF3US, whereas AF3US wins only 10.63%. AF3US wins 59.38% of domains against BLOSUM45, with BLOSUM45 winning 30.00%. In the case of tests with minimum TM-score above 0.5, the performance of AF3US is slightly improved but the trends are preserved.

Full details are given for interested readers in Supplementary Tables 5-16. Wilcoxon test p-values below 0.01 are considered statistically significant and only these tests are considered when summary results are calculated. Those with p-value above 0.01 are ignored; they are shown in red in the tables.

Table 1: Summary of summaries. Sum of totals and average of averages are presented for all comparisons performed, combining the results for all distances and the two databases (Supplementary Tables 3-4).

All tests			High TM-score tests		
Domains won					
Ankh-score	AF3US	null	Ankh-score	AF3US	null
126	17	17	**114**	16	30
AF3US	BLOSUM45	null	AF3US	BLOSUM45	null
95	48	17	**92**	38	30
Ankh-score	BLOSUM45	null	Ankh-score	BLOSUM45	null
140	6	14	**136**	2	22
Percentage won					
Ankh-score	AF3US	equal	Ankh-score	AF3US	equal
67.42	22.69	9.88	**66.47**	22.48	11.05
AF3US	BLOSUM45	equal	AF3US	BLOSUM45	equal
53.24	40.47	6.29	**56.21**	37.15	6.65
Ankh-score	BLOSUM45	equal	Ankh-score	BLOSUM45	equal
73.76	15.11	11.12	**73.74**	12.92	13.35

4. Case studies

We present in this section three examples of proteins pairs with their three alignments: the reference, Ankh-score and AF3US. The proteins' organism

and length are given in Table 2. Their domains have been identified using SUPFAM (`supfam.org`) [10, 24].

Table 2: Three case studies. The protein pairs studied are given their ID, name, organism and length.

Protein ID	Protein name	Organism	Length
2D1K_C	WH2 domain of human Metastasis Suppressor Protein 1 (MTSS1)	*Homo sapiens*	32
3UE5_B	Domain D of Spire protein	*Drosophila melanogaster*	66
HT16_HYDAT	Tyrosine-protein kinase HTK16	*Hydra vulgaris*	757
1d4t_A	SH2 domain of human SAP protein 1A	*Homo sapiens*	104
YXJL_BACSU	Uncharacterized transcriptional regulatory protein YxjL	*Bacillus subtilis* (strain 168)	218
DEGU_BACSU	Transcriptional regulatory protein DegU	*Bacillus subtilis* (strain 168)	229

4.1. MTSS1 vs Spire

The MTSS1 protein has one domain: WH2 superfamily, Wiskott Aldrich syndrome homology region 2 (WH2 motif) found in Metastasis suppressor protein [1..31]. The Spire protein has two domains: WH2 superfamily, Wiskott-Aldrich Syndrome Homology (WASP) region 2 (WH2 motif) [5..30] and WH2 superfamily, fourth tandem Wiskott-Aldrich Syndrome Homology (WASP) region 2 (WH2 motif) [37..63]. The TM-scores are 0.51 and 0.31, which means one is borderline good, the other not very good. Yet, this is a simple test, as the protein sequences are very short. The alignments are shown in Figure 5. The reference aligns the WH2 domain in MTSS1 with the second WH2 of the Spire protein. Ankh-score does the same, producing an identical alignment with the reference. AF3US aligns the WH2 domain in MTSS1 with the first WH2 of the Spire protein, producing a completely different alignment.

For more clarity, we represent the alignments as structural alignments in Figure 6 (first row). Here we used PyMol [5] to construct structural alignments from the sequence alignments, based on the AlphaFold-predicted structures. We used the 'pair-fit' command of PyMol which forces a residue-by-residue alignment between the two AlphaFold-predicted structure's backbone alpha-carbon atoms according to the sequence alignment, then finds the

The reference and Ankh-score alignment:

```
Seq 1 : 1     -------------------------------------DTPQGEDMLNAIRRGVKLKKTTTNDRSAPRFS    32
                                                      E     IR G  LK  T     A
Seq 2 : 1     GPLGSKKDAHAMILEFIRSRPPLKKASDRQLGPPRMCEPSPREQLMESIRKGKELKQITPPEAAAS---    66
```

The AF3US alignment:

```
Seq 1 : 1     --DTPQGEDMLNAIRRGVKLKKTTT--NDRSAPRFS--------------------------------------    32

Seq 2 : 1     GPLGSKKDAHAMILEFIRSRPPLKKASD--------RQLGPPRMCEPSPREQLMESIRKGKELKQITPPEAAAS    66
```

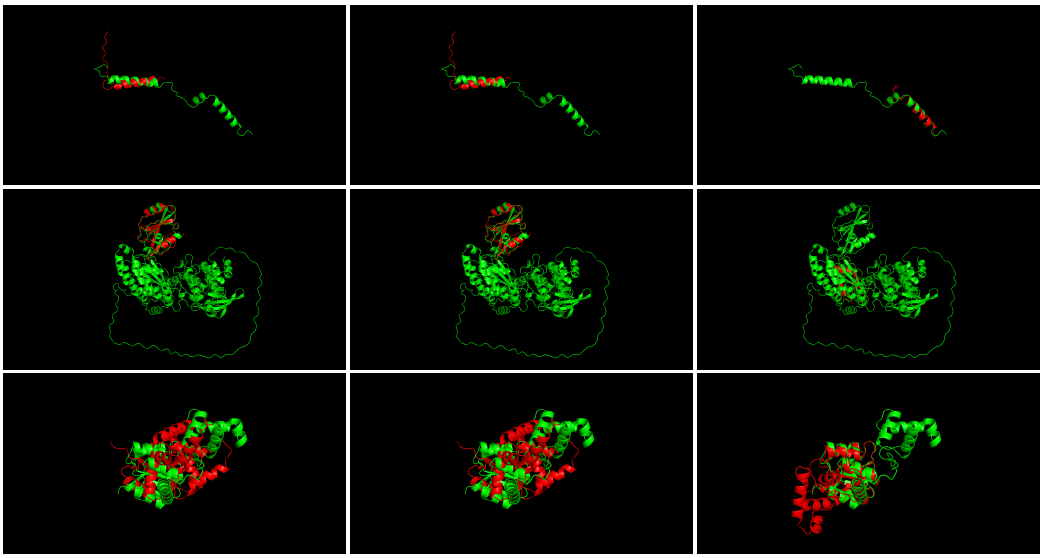
Figure 5: The sequence alignments for MTSS1 (`Seq 1`) and Spire (`Seq 2`) proteins. (Only two alignments are shown because the Ankh-score alignment is identical with the reference.)

best rigid-body transformation, that minimizes RMSD (root mean square deviation). The result is a structural alignment that can be used to better visualize the differences between the sequence alignments. It is very clearly visible in the first row of Figure 6 that the Ankh-score alignment is identical with the reference, whereas the one by AF3US is completely different.

4.2. HT16 vs SH2 SAP

The HT16 protein has four domains: SH2 (Src Homology 2 domain) [5..105], Ankryin repeat domain [99..274], another SH2 [284..503], and protein kinase, catalytic subunit [477..741]. The SH2 SAP domain has one SH2 domain [7..102]. The TM-scores are 0.84 and 0.13. These are two proteins of very different length, which means one of the TM-scores has to be low. The higher score is very high, which means good alignment of the shorter protein with a subregion of the longer. The two SH2 domains have similar shape but potentially different functions based on possible subtle differences in their surface loops, which confuses the structural alignment.

The alignments are shown in Figure 7. The SH2 domains differ in length – 101 and 220 for the first protein and 96 for the second – and the ones of similar length are aligned by the reference, that is, it aligns the SH2 domain of the second protein with the first SH2 domain of the first protein. The Ankh-score alignment does the same, being nearly identical with the reference. The AF3US alignment, on the other hand, aligns the SH2 domain of the second protein with the second, much longer, SH2 domain of the longer protein. The alignments are also visualized in second row of Figure 6.



(i) (ii) (iii)

Figure 6: Visualizations of the three alignments: (i) reference, (ii) Ankh score, (iii) AF3US. Row 1: MTSS1 (red) and Spire (green) proteins; row 2: HT16 (green) and SH2 SAP (red) proteins; row 3: YxjL (red) and DegU (green) proteins.

4.3. YxjL vs DegU

The Yxjl protein has two domains: CheY related [5..132] and GerE-like (LuxR/UhpA family of transcriptional regulators) [158..304]. The DegU protein has two domains: CheY related [3..136] and GerE-like (LuxR/UhpA family of transcriptional regulators) [149..217]. The TM-scores are 0.62 and 0.59. These are two proteins of similar lengths for which both TM-scores are over 0.5, indicating good structural alignment. They have two common domains, in the same order.

The alignments are shown in Figure 8. The reference alignment indicates, as expected, that the common domains in the two protein sequences are aligned against each other. The Ankh-score alignment is, again, identical with the reference. AF3US aligns the first domain perfectly, but then completely misses the second one. It aligns each occurrence of the second domain entirely with gaps in the other sequence. The alignments are also visualized in third row of Figure 6

```
Seq 1 : 1     ----MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSESVPGVYCLCVLYHGYIYTYR   55
                  DA  YHGKI RE    LL  G DG  L RD   P  Y L      I
Seq 2 : 1     MSKNSDALLWYHGKITREVAVQVLLRKGGRDGFFLIRDCGNAPEDYVLSMMFRSQILHFQ   60

Seq 1 : 56    VSQTETGSWSAETAPGVHKRYFRKIKNLISAFQKPDQGIVIPLQYFVEK-----------  104
                                 F    LIS      G   L
Seq 2 : 61    INCLG-DNKFSIDNGP----IFQGLDMLISYYKVISDGLPCKLVDFCVGKIAPLYALKYG  115

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 116   LDTRLHLACEEKNPNTVKELLQDSVIKEHVNARSISGLTALHISCSNGDNDIVAMLLNAG  175

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 176   ADASAIDANGRTPVQVVCFYNHASTLHLLISKGSADFLKRSPNNGWVPLHEAAMRGSLEC  235

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 236   VKVLLSFNASMYPRSLDGDTPRDLALQYENYNVVEFFDNYPVNQPKTSITQWLHQNLDRN  295

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 296   GALIILQNASMADGSFLIRSSIKCHGYYVLTLVYEKKTYHFQIKSRADRWFYIDDGPLFE  355

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 356   TLPHLVDHYMQYADGLPTLLQFPVPSAENRKRPLPPTPTKNQLKLPVPPSRPIKNNNGLP  415

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 416   QPLPYPEFTNESDSDIFTRLECEKEKPLPKLPRPVVNHTEVPHSVNVGQKGDQTMKNNAQ  475

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 476   QNIILKESISFGKELGVGEFGSVIKGIWLSPGGKEINVAMKTLNKDKMVQGEKEFLREAL  535

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 536   VMSQLNHPCIVSLLGVCLGPPMILVQELVEMGALLDYLMDYQPEIQEVDLKLWASQIAFG  595

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 596   MMYLELKRFVHRDLAARNILLANKKQVKISDFGLSRAVGTGSDYYQAKQGGRWPVRWYAP  655

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 656   ESINYGTFSTKSDVWSYGITLWEMFTFGDLPYGEMTGNEVVSFLERCGRLEKPDECPINT  715

Seq 1 : 104   ---------------------------------------------  104

Seq 2 : 716   YSIMLSCWHIDPNKRPTFNELHSTFSTDPEYEDVRIYRDRIK   757
```

(i)

```
Seq 1 : 1     ----MDAVAVYHGKISRETGEKLLLAT-GLDGSYLLRDSESVPGVYCLCVLYHGYIYTYR   55
                  DA  YHGKI RE    LL  G DG  L RD   P  Y L      I
Seq 2 : 1     MSKNSDALLWYHGKITREVAVQVLLRKGGRDGFFLIRDCGNAPEDYVLSMMFRSQILHFQ   60

Seq 1 : 56    VSQTETGSWSAETAPGVHKRYFRKIKNLISAFQKPDQGIVIPLQYFVEK-----------  104
                         S          F    LIS      G   L
Seq 2 : 61    INCLGDNKFSIDNG-----PIFQGLDMLISYYKVISDGLPCKLVDFCVGKIAPLYALKYG  115

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 116   LDTRLHLACEEKNPNTVKELLQDSVIKEHVNARSISGLTALHISCSNGDNDIVAMLLNAG  175

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 176   ADASAIDANGRTPVQVVCFYNHASTLHLLISKGSADFLKRSPNNGWVPLHEAAMRGSLEC  235

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 236   VKVLLSFNASMYPRSLDGDTPRDLALQYENYNVVEFFDNYPVNQPKTSITQWLHQNLDRN  295

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 296   GALIILQNASMADGSFLIRSSIKCHGYYVLTLVYEKKTYHFQIKSRADRWFYIDDGPLFE  355

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 356   TLPHLVDHYMQYADGLPTLLQFPVPSAENRKRPLPPTPTKNQLKLPVPPSRPIKNNNGLP  415

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 416   QPLPYPEFTNESDSDIFTRLECEKEKPLPKLPRPVVNHTEVPHSVNVGQKGDQTMKNNAQ  475

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 476   QNIILKESISFGKELGVGEFGSVIKGIWLSPGGKEINVAMKTLNKDKMVQGEKEFLREAL  535

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 536   VMSQLNHPCIVSLLGVCLGPPMILVQELVEMGALLDYLMDYQPEIQEVDLKLWASQIAFG  595

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 596   MMYLELKRFVHRDLAARNILLANKKQVKISDFGLSRAVGTGSDYYQAKQGGRWPVRWYAP  655

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 656   ESINYGTFSTKSDVWSYGITLWEMFTFGDLPYGEMTGNEVVSFLERCGRLEKPDECPINT  715

Seq 1 : 104   ---------------------------------------------  104

Seq 2 : 716   YSIMLSCWHIDPNKRPTFNELHSTFSTDPEYEDVRIYRDRIK   757
```

(ii)

```
Seq 1 : 1     MD--------------------------------------------------------   2

Seq 2 : 1     --MSKNSDALLWYHGKITREVAVQVLLRKGGRDGFFLIRDCGNAPEDYVLSMMFRSQILH   58

Seq 1 : 2     ------------------------------------------------------------   2

Seq 2 : 59    FQINCLGDNKFSIDNGPIFQGLDMLISYYKVISDGLPCKLVDFCVGKIAPLYALKYGLDT  118

Seq 1 : 2     ------------------------------------------------------------   2

Seq 2 : 119   RLHLACEEKNPNTVKELLQDSVIKEHVNARSISGLTALHISCSNGDNDIVAMLLNAGADA  178

Seq 1 : 2     ------------------------------------------------------------   2

Seq 2 : 179   SAIDANGRTPVQVVCFYNHASTLHLLISKGSADFLKRSPNNGWVPLHEAAMRGSLECVKV  238

Seq 1 : 3     ----------------------------------------AVA-YVHGKISRETGE   17
                                                                   H   R
Seq 2 : 239   LLSFNASMYPRSLDGDTPRDLALQYENYNVVEFFDNYPVNQPKTSITQWLHQNLDRNGAL  298

Seq 1 : 18    KLLLATG-LDGSYLLRDSESVPGVYCLCVLYHGYIYTYRVSQTETGSWSAETAPGVHKRY   76
                 L     DGS L R S   G Y L   Y   Y
Seq 2 : 299   ITLQNASMADGSFLIRSSIKCHGYYVLTLVYEKKTYHFQIKSRADRWFYID-----DGPL  353

Seq 1 : 77    FRKIKNLISAFQKPDQGIVIPLQYPVEK------------------------------  104
                F    L                G   LQ PV
Seq 2 : 354   FETLPHLVDHYMQYADGLPTLLQFPVPSAENRKRPLPPTPTKNQLKLPVPPSRPIKNNNG  413

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 414   LPQPLPYPEFTNESDSDIFTRLECEKEKPLPKLPRPVVNHTEVPHSVNVGQKGDQTMKNN  473

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 474   AQQNIILKESISFGKELGVGEFGSVIKGIWLSPGGKEINVAMKTLNKDKMVQGEKEFLRE  533

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 534   ALVMSQLNHPCIVSLLGVCLGPPMILVQELVEMGALLDYLMDYQPEIQEVDLKLWASQIA  593

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 594   FQMMYLELKRFVHRDLAARNILLANKKQVKISDFGLSRAVGTGSDYYQAKQGGRWPVRWY  653

Seq 1 : 104   ------------------------------------------------------------  104

Seq 2 : 654   APESINYGTFSTKSDVWSYGITLWEMFTFGDLPYGEMTGNEVVSFLERCGRLEKPDECPI  713

Seq 1 : 104   ---------------------------------------------  104

Seq 2 : 714   NTYSIMLSCWHIDPNKRPTFNELHSTFSTDPEYEDVRIYRDRIK   757
```

(iii)

Figure 7: The sequence alignments for HT16 (`Seq 1`) and SH2 SAP (`Seq 2`): (i) reference, (ii) Ankh score, (iii) AF3US.

```
Seq 1 : 1     MNKDKIRVALADDQPLVREGFRYVINAQTDMTVSGEAGDGHDIIALAKQTKPDVILMDVQ   60
                  K      DD  L REG           V  E  DG          PDV  MD
Seq 2 : 1     --MTKVNIVIIDDHQLFREGVKRILDFEPTFEVVAEGDDGDEAARIVEHYHPDVVIMDIN   58

Seq 1 : 61    MPRCSGIEAAKDIMSALPNTKIVILTTFDTEEYVFEGIRAGAVGYLLKDTLPEELIDAIR  120
              MP   G EA K      P  K  IL   D E YV      GA GYLLK      LI A
Seq 2 : 59    MPNVNGVEATKQLVELYPESKVIILSIHDDENYVTHALKTGARGYLLKEMDADTLIEAVK  118

Seq 1 : 121   AAARGEAIFRTVTAAKIISETFRAKQQ----------THAEELAEPFTKRELEVLQQMA  169
                A G              E  R                      T RE EVLQ  A
Seq 2 : 119   VVAEGGSYLHPKVTHNLVNEFRRLATSGVSAHPQHEVYPEIRRPLHILTRRECEVLQMLA  178

Seq 1 : 170   YGLRNEDIAEKLFVSESTVKTHVHRILQKCNAQDRTQAVVFAIRNGIVQ--   218
               G  N  I E LF SE TVK HV  ILQK N  DRTQAVV AI NG V
Seq 2 : 179   DGKSNRGIGESLFISEKTVKNHVSNILQKMNVNDRTQAVVVAIKNGWVEMR   229
```

(i)

```
Seq 1 : 1     MNKDKIRVALADDQPLVREGFRYVINAQTDMTVSGEAGDGHDIIALAKQTKPDVILMDVQ   60
                  K      DD  L REG           V  E  DG          PDV  MD
Seq 2 : 1     --MTKVNIVIIDDHQLFREGVKRILDFEPTFEVVAEGDDGDEAARIVEHYHPDVVIMDIN   58

Seq 1 : 61    MPRCSGIEAAKDIMSALPNTKIVILTTFDTEEYVFEGIRAGAVGYLLKDTLPEELIDAIR  120
              MP   G EA K      P  K  IL   D E YV      GA GYLLK      LI A
Seq 2 : 59    MPNVNGVEATKQLVELYPESKVIILSIHDDENYVTHALKTGARGYLLKEMDADTLIEAVK  118

Seq 1 : 121   AAARGEAIFRTVTAAKIISETFRAK-QQTHAEELAEPFTKRELEVLQQMAYGLRNEDIAE  179
                A G              E  R
Seq 2 : 119   VVAEGGSYLHPKVTHNLVNEFRRLATSG-------------------------------  146

Seq 1 : 180   KLFVSESTVKTHVHRILQKCNAQDRTQAVVFAIRNGIVQ----------------------  218

Seq 2 : 147   --------------------------------------VSAHPQHEVYPEIRRPLHILT  167

Seq 1 : 218   ------------------------------------------------------------  218

Seq 2 : 168   RRECEVLQMLADGKSNRGIGESLFISEKTVKNHVSNILQKMNVNDRTQAVVVAIKNGWVE  227

Seq 1 : 218   --   218

Seq 2 : 228   MR   229
```

(ii)

Figure 8: The sequence alignments for YxjL (`Seq 1`) and DegU (`Seq 2`): (i) reference and Ankh score and (ii) AF3US. (Only two alignments are shown because the Ankh-score alignment is identical with the reference.)

5. Experimentally determined structures

One aspect that needs to be investigated is that of experimentally determined structures. They are expected to be superior to the AlphaFold-predicted structures, implying that the sequence alignments produced by

US-aligning experimentally determined structures are expected to be better as well. We haver attempted to clarify this aspect, with limited success. The problem is the lack of experimentally determined structures for the exact sequences in our MSAs. The sequences must be identical, otherwise the distances cannot be computed. Through considerable manual effort, we identified and tested a few protein pairs with suitable experimentally determined structures. We identified 11 suitable sequences from the cd14765 domain, giving us, with the four distances, 220 tests. The results are summarized in Table 3. The Ankh-score method is still clearly superior to both AF3US and experimental structures. The unexpected result is AF3US against experimental (aligned with US-align). AF3US wins 41.82% of tests, whereas the experimental structures, US-aligned, win only 35.45% of tests, with the remaining 22.73% being ties. Any conclusion drawn from this comparison is not reliable, due to the small number of samples, the fact that they all came from the same domain, and the fact that not even reference MSAs are perfect. Yet, the results presented are intriguing and sufficient to justify further investigation.

Table 3: Comparing US-aligned experimentally determined structures with Ankh-score and AF3US. A total of 220 tests from the cd14765 domain were used.

AF3US vs Experimental	AF3US wins	Experimental wins	equal
	92	78	50
Ankh-score vs Experimental	Ankh-score wins	Experimental wins	equal
	150	30	40
Ankh-score vs AF3US	Ankh-score wins	AF3US wins	equal
	150	27	43

6. Conclusion

Protein sequence alignments are fundamental in bioinformatics and biology because they provide explicit residue-to-residue correspondence, essential for tracing evolutionary relationships, detecting conserved motifs, and building profile databases. Alignments uniquely reveal positional conservation across homologs, grounding molecular biology in evolutionary context. Their

function is essential and improvements to the quality of computational alignments are very important. We investigated here two powerful candidates for improving protein sequence alignments: AlphaFold and protein embeddings.

AlphaFold marked a breakthrough in biology, providing a much better solution to the decades-long protein structure prediction challenge. By providing near-experimental accuracy at scale, it transforms structural biology, accelerating drug discovery, enzyme design, and functional annotation. Alignment of its predicted structures is considered as a challenger for traditional alignments.

The appearance of protein embeddings introduced a transformative shift in bioinformatics, enabling rich vector representations of sequences learned from massive unlabelled datasets. Embeddings capture subtle biochemical and structural patterns, allowing functional prediction, remote homology detection, and clustering beyond sequence identity.

We compared the two proposed improvements thoroughly and the result is the clear domination of the Ankh-score as the current best method for computing protein sequence alignments. This result has the important implication that Ankh embedding vectors may possess certain information that is not available in the AlphaFold3-predicted structures, which requires further investigation.

We considered very briefly experimental structures, which are expected to surpass AF3US. Very limited testing indicates, unexpectedly, the opposite. This issue clearly deserves more investigation.

Acknowledgements

All our computations were performed on Digital Research Alliance of Canada servers.

Competing interests

No competing interest is declared.

Author contributions

J.M. selected the data, performed all tests, including installing and running the necessary libraries, and contributed to the methodology and analysis. K.R. performed the detailed biological analysis of case studies. G.B.G. advised on choosing the data and contributed to the biological analysis. L.I. proposed the study, designed the methodology, analyzed the results, supervised the work and wrote the manuscript.

Funding

This work was supported by NSERC Discovery RGPIN-2021-03978 to L.I. and RGPIN-2020-05733 to G.B.G.

References

[1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 630(8016):493–500, 2024.

[2] S. Ashrafzadeh, G. B. Golding, S. Ilie, and L. Ilie. Scoring alignments by embedding vector similarity. Briefings in Bioinformatics, 25(3):bbae178, 2024.

[3] J. L. Binder, J. Berendzen, A. O. Stevens, Y. He, J. Wang, N. V. Dokholyan, and T. I. Oprea. Alphafold illuminates half of the dark human proteins. Current opinion in structural biology, 74:102372, 2022.

[4] B. P. Blackburne and S. Whelan. Measuring the distance between multiple sequence alignments. Bioinformatics, 28(4):495–502, 2012.

[5] W. L. DeLano et al. Pymol: An open-source molecular graphics tool. CCP4 Newsl. protein crystallogr, 40(1):82–92, 2002.

[6] A. Elnaggar, H. Essam, W. Salah-Eldin, W. Moustafa, M. Elkerdawy, C. Rochereau, and B. Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. arXiv preprint arXiv:2301.06568, 2023.

[7] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 44(10):7112–7127, 2021.

[8] V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. Nature communications, 12(1):3168, 2021.

[9] O. Gotoh. An improved algorithm for matching biological sequences. Journal of molecular biology, 162(3):705–708, 1982.

[10] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. Journal of molecular biology, 313(4):903–919, 2001.

[11] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 89(22):10915–10919, 1992.

[12] S. Hosseini, G. B. Golding, and L. Ilie. Seq-InSite: sequence supersedes structure for protein interaction site prediction. Bioinformatics, 40(1):btad738, 01 2024.

[13] B. G. Iovino and Y. Ye. Protein embedding based alignment. BMC bioinformatics, 25(1):85, 2024.

[14] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583–589, 2021.

[15] K. Kaminski, J. Ludwiczak, K. Pawlicki, V. Alva, and S. Dunin-Horkawicz. plm-blast: distant homology detection based on direct comparison of sequence representations from protein language models. Bioinformatics, 39(10):btad579, 2023.

[16] M. Kulmanov and R. Hoehndorf. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics, 36(2):422–429, 2020.

[17] B. Lai and J. Xu. Accurate protein function prediction via graph attention networks with predicted structure information. Briefings in Bioinformatics, 23(1):bbab502, 2022.

[18] A. M. Lesk and A. S. Konagurthu. Protein structure prediction improves the quality of amino-acid sequence alignment. Proteins: Structure, Function, and Bioinformatics, 90(12):2144–2147, 2022.

[19] J. Malec, G. B. Golding, and L. Ilie. Protein embeddings and local alignments. bioRxiv, pages 2025–07, 2025.

[20] M. Manfredi, C. Savojardo, P. L. Martelli, and R. Casadio. Ispred-seq: Deep neural networks and embeddings for predicting interaction sites in protein sequences. Journal of Molecular Biology, 435(14):167963, 2023.

[21] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, et al. CDD: NCBI's conserved domain database. Nucleic acids research, 43(D1):D222–D226, 2015.

[22] C. D. McWhite, I. Armour-Garb, and M. Singh. Leveraging protein language models for accurate multiple sequence alignments. Genome Research, 33(7):1145–1153, 2023.

[23] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3):443–453, 1970.

[24] A. P. Pandurangan, J. Stahlhacke, M. E. Oates, B. Smithers, and J. Gough. The superfamily 2.0 database: a significant proteome update and a new webserver. Nucleic acids research, 47(D1):D490–D494, 2019.

[25] S. Rajapaksa, A. S. Konagurthu, and A. M. Lesk. Sequence and structure alignments in post-alphafold era. Current Opinion in Structural Biology, 79:102539, 2023.

[26] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15):e2016239118, 2021.

[27] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, et al. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792):706–710, 2020.

[28] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch. Balibase 3.0: latest developments of the multiple sequence alignment benchmark. Proteins: Structure, Function, and Bioinformatics, 61(1):127–136, 2005.

[29] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker. Improved protein structure prediction using predicted inter-residue orientations. Proceedings of the National Academy of Sciences, 117(3):1496–1503, 2020.

[30] C. Zhang, M. Shine, A. M. Pyle, and Y. Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nature methods, 19(9):1109–1115, 2022.

[31] Y. Zhang and J. Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. Nucleic acids research, 33(7):2302–2309, 2005.