Scoring alignments by embedding vector similarity

Sepehr Ashrafzadeh[1], G. Brian Golding[2], Lucian Ilie[1,*]

[1]Department of Computer Science, University of Western Ontario, London, N6A 5B7, Ontario, Canada
[2]Department of Biology, McMaster University, Hamilton, L8S 4K1, Ontario, Canada

August 30, 2023

Abstract

Sequence similarity is of paramount importance in biology, as similar sequences tend to have similar function and share common ancestry. Scoring matrices, such as PAM or BLOSUM, play a crucial role in all bioinformatics algorithms for identifying similarities, but have the drawback that they are fixed, independent of context. We propose a new scoring method for amino acid similarity that remedies this weakness, being contextually dependent. It relies on recent advances in deep learning architectures that leverage the power of enormous amounts of unlabelled data in order to generate contextual embeddings, which are vector representations for words. These ideas have been applied to protein sequences, producing embedding vectors for protein residues. We propose the *E-score* between two residues as the cosine similarity between their embedding vector representations. Thorough testing on a wide variety of reference multiple sequence alignments indicate that the alignments produced using the best such method, ProtT5-score, are significantly better than those obtained using BLOSUM matrices. The new method proposes to change the way alignments are computed, with far reaching implications in all areas of textual data that use sequence similarity. The program to compute alignments based on various E-scores is available as a web server at `e-score.csd.uwo.ca`. The source code is freely available for download from `github.com/lucian-ilie/E-score`.

1 Introduction

Sequence similarity is fundamental in sequence analysis. Significant similarities do not appear by chance but are evolutionarily motivated, with corresponding implications on their associated functions. Given a new sequence, its functionality is much faster investigated by performing a database search to identify known similar sequences and inferring the function of the new sequence from the existing information, resulting often in accurate identification. This procedure is sometimes the only way to infer function. Sequence similarity search is the most widely used procedure in bioinformatics, the BLAST program [2, 3] being one of the most cited contributions in the history of science.[1]

The most important component of similarity search algorithms, such as BLAST, is the function that describes the similarity scores of various components, such as nucleotides for DNA and amino acids for proteins. We consider here the much more complicated case of proteins, where scoring of alignments is done using PAM [6] or BLOSUM [11] matrices, which give the substitution scores for pairs of amino acids. These matrices are fixed, which means that the substitution score for a pair of amino acids does not depend on their context, that is, the surrounding sequence of the protein containing them. We propose a new scoring function for residue pairs that is dependent of the context in which the residues appear. The new score makes use of protein embeddings.

Many of the best current techniques in Natural Language Processing (NLP) involve transfer learning based on pre-trained models using self-supervised learning on unlabelled data. Such data is freely available in enormous amounts, such as the Common Crawl corpus (commoncrawl.org). Words that share similar meanings or are related often appear in similar contexts. This enables scalable models to map words to numerical vectors (embeddings) in a high-dimensional space. Words with similar co-occurrence patterns are represented by vectors that are closer in this space. Such text embeddings are able to capture the semantic meaning of words. One of the best known earlier models is word2vec [21]. Denoting by w_v the vector associated with a word w, an example ([21]) of meaning capture is shown by the fact that $\texttt{king}_v - \texttt{man}_v + \texttt{woman}_v$ produces a vector that is closest to $\texttt{queen}_v$.

Embeddings such as word2vec [21] or GloVe [24] are contextual independent as the embedding vector for a given word is the same regardless of context. The next generation of embeddings, called contextual embeddings, overcome this limitation, e.g., ELMo [25], BERT [7], RoBERTa [19], XLNet [34], T5 [26]. The most successful contextual embeddings rely on the transformer architecture [32], achieving state-of-the-art results in many NLP benchmarks [7, 19, 34, 18].

Protein sequences present a similar situation, with entire proteins seen as sentences and residues as words, enabling the use of the embedding techniques developed for NLP to take advantage of the vast amount of protein sequences from

*Corresponding author: `ilie@uwo.ca`

[1]The two BLAST papers [2, 3] combined exceed 190,000 citations on Google Scholar as of Aug. 2023.

UniProt [1]. Many protein embeddings have been proposed, including ProtVec [4], SeqVec [10], PRoBERTa [22], MSA-transformer [27], ESM-1b and ESM2 [28], and the models in the ProtTrans project: ProtTXL, ProtBert, ProtXLNet, ProtAlbert, ProtElectra, and ProtT5 [8]. These protein embeddings have many important applications to various areas, such as: structure prediction [29, 33, 15], function prediction [16, 9, 17], interaction site prediction [13, 35, 12], etc. In many applications they help achieving state-of-the-art performance.

Our new scoring function uses any embedding method that associates a vector to each residue. For any such method, E, we define the *E-score* between two residues as the cosine similarity between their associated vectors. We investigate this new scoring function for contextual embeddings from the point of view of correct alignment generation. We have selected a wide variety of reference multiple sequence alignments (MSAs) from NCBI's CDD (Conserved Domain Database) [20], that are manually curated according to the function. For each, we randomly selected many pairs of protein sequences in order to check how close the pairwise alignment is to the original reference MSA. Several alignment distances [5], including one that we introduce here, were used to compare the best new contextual ProtT5-score, with the best old one, the BLOSUM62 matrix. ProtT5-score performs overwhelmingly better than BLOSUM62, producing an average alignment distance that is better in 35 MSAs out of the 37 tested.

The new scoring method proposes completely changing the way alignments are computed, with far reaching implications in all areas that use sequence similarity.

2 Materials and Methods

2.1 Cosine similarity

For two vectors, $A = (A_i)_{i=1..n}$ and $B = (B_i)_{i=1..n}$, the *cosine similarity* between A and B is defined as the cosine of the angle, θ, between the two vectors:

$$\mathrm{CosSim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \ .$$

The cosine similarity takes values between -1 and 1, with -1 for vectors of opposite direction, 1 for the same direction, and 0 for orthogonal vectors; Fig. 1 illustrates these three cases.

2.2 E-score

For a protein sequence P, of length m, denote its ith residue by $P[i]$. Consider an embedding method, E, that produces embedding vectors of size n. Assume that E, when applied to protein P, generates an $m \times n$ matrix, $E(P)$, that has as its rows all m embedding vectors of the residues of P; denote the embedding of the ith residue by $E(P)_i = E(P)(i, :)$.

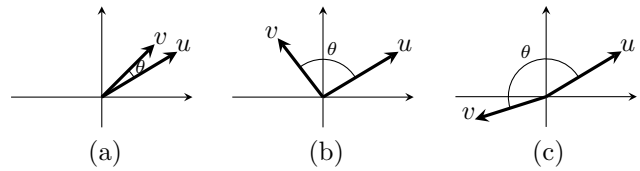


Figure 1: Cosine similarity of vectors u and v is the cosine of the angle, θ, between them: (a) similar vectors: $\theta \approx 0$, $\cos(\theta) \approx 1$; (b) orthogonal (independent) vectors: $\theta \approx \frac{\pi}{2}$, $\cos(\theta) \approx 0$; (c) opposite vectors: $\theta \approx \pi$, $\cos(\theta) \approx -1$.

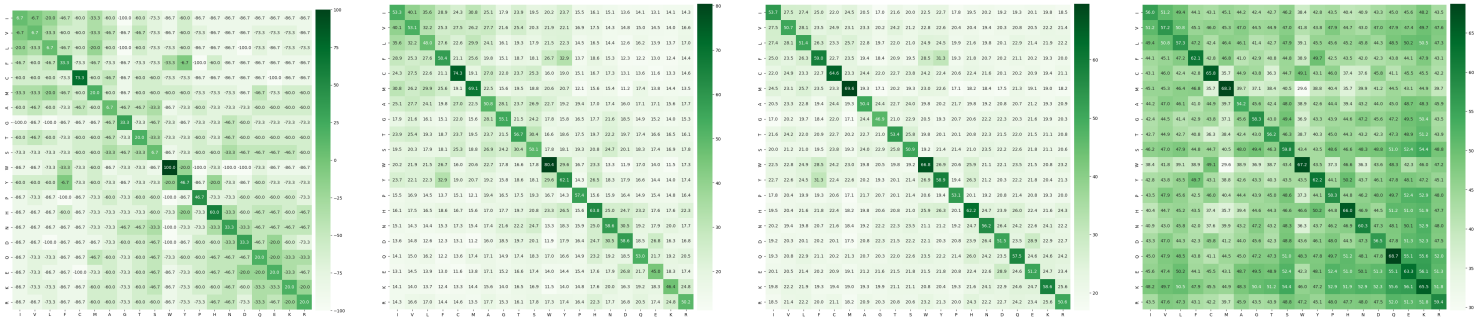
For two proteins P and Q, and two positions, i within P and j within Q, the *E-score* between the residues $P[i]$ and $Q[j]$ is defined as the cosine similarity between the corresponding embedding vectors:

$$E\text{-score}(i, j) = \mathrm{CosSim}(E(P)_i, E(Q)_j) \ .$$

Note that, while not reflected by our notation, E-score(i, j) depends also on the protein sequences P and Q. Thus, when the embedding method is ProtT5, ProtBert, ESM2, etc., we obtain ProtT5-score, ProtBert-score, ESM2-score, and so on. Clearly, for any contextual embedding method E, the E-score changes with context as well.

2.3 Visualizing E-scores

To have an idea what these new E-scores look like, we constructed some scoring matrices using those, and compared them with the BLOSUM matrices. While a BLOSUM matrix is fixed, the E-scores change with context, so we computed some average scores for each pair of amino acids. We picked the longest multiple sequence alignment (MSA), *NBD_sugar-kinase_HSP70_actin*, from the selected six representative MSAs; see the first column of Supplementary Tables 4-7. Then, we distinguished between "aligned" residues, which appear in the same column of the MSA, and "unaligned" residues, which occur in different columns. Since the E scores use the context, it is expected that aligned residues have significantly higher scores compared to unaligned ones. For each pair of amino acids, we randomly selected occurrences of their residues in the MSA and computed the average score; 100,000 pairs were selected in each of the categories aligned and unaligned. We do this separately for the aligned and the unaligned case. This way, for a given embedding type and an MSA, we obtain two 20×20 matrices. We show the aligned matrices for three embeddings: ProtT5-score, ProtAlbert-score, and ProtXLNet-score, alongside the BLOSUM62 matrix, in Fig. 2. It is apparent from Fig. 2 that the matrix for the ProtT5-score is the most similar with the BLOSUM62 matrix. The heatmaps of all five BLOSUM matrices and all twelve E-score matrices (two matrices for each embedding method, for the aligned and the unaligned case) are shown in the Supplementary Table 11. It is interesting to notice the wide variety of ranges among the E-score matrices. The unaligned matrices have lower, and narrower, ranges. It appears that a wider range of values is an indication of better performance for alignments.



(a) (b) (c) (d)

Figure 2: Heatmaps of (a) BLOSUM62 matrix (scaled to −1..1) and three aligned matrices of average E-scores for the *NBD_sugar-kinase_HSP70_actin* MSA: (b) ProtT5-score, (c) ProtAlbert-score, and (d) ProtXLNet-score.

2.4 Correlations

It is of interest to see how well the new E-scores are correlated with the BLOSUM matrices. We consider here five BLOSUM matrices: BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, and BLOSUM90, and six embedding methods: ProtT5, ProtBert, ProtAlbert, ProtXLNet, ESM-1b and ESM2. For each embedding method, we use for computing the correlations the same matrices computed above for the *NBD_sugar-kinase_HSP70_actin* MSA. Figure 3 gives the Pearson correlations between the five BLOSUM matrices and the twelve E-score matrices. It can be seen that the ProtT5 matrices have the highest correlation with all BLOSUM matrices, followed, in order, by ESM2 and ProtBert, and then ProtAlbert, ESM-1b and, by far the last, ProtXLNet. For each method, the aligned matrix has higher correlation with BLOSUM matrices than the unaligned one. Interestingly, the unaligned ProtT5 matrix comes in second place, after the aligned ProtT5 and before all the other aligned matrices. With the exception of ProtXLNet, the two matrices, aligned and unaligned, for the same embedding method are highly correlated.

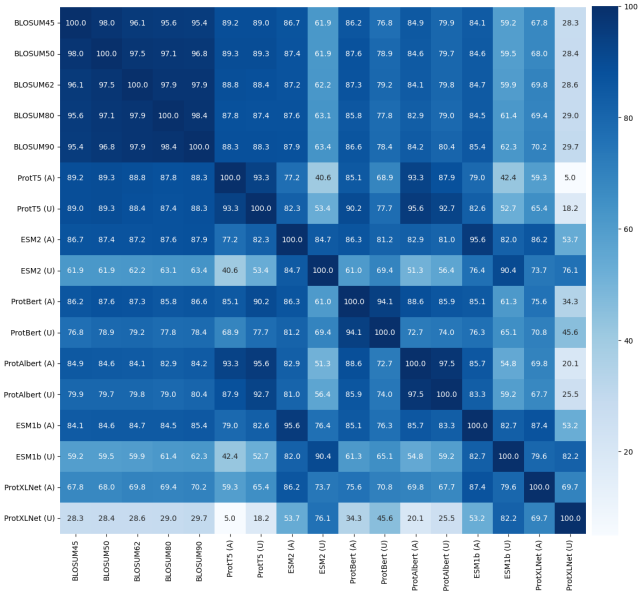


Figure 3: Pearson correlations between five BLOSUM matrices and twelve E-score matrices. For each embedding method we considered, the aligned and unaligned matrices for the *NBD_sugar-kinase_HSP70_actin* MSA have been included.

2.5 Alignment distances

Our goal is to show that the use of the new E-scores provides better alignments compared to the BLOSUM scores. We do this by comparing their ability to produce true alignments, as given by the reference alignments from CDD; see Reference alignments subsection below. We selected a number of reference MSAs, discussed in the Results section, and investigated how close pairwise alignments produced using E-score and BLOSUM matrices, resp., are from the alignments between the same sequences, as induced by the reference MSAs. To compare how close two alignments are, we need to measure distances between alignments. We use four distance metrics, three existing ones, $\mathrm{d_{ssp}}$, $\mathrm{d_{seq}}$, and $\mathrm{d_{pos}}$ from [5], and one we introduce here, $\mathrm{d_N}$.

We give first a formal definition of an alignment. Consider an alphabet Σ. For a string P over Σ of length $|P| = n$, let $P[i], 1 \leq i \leq n$, denote the ith character of P. For two strings P, Q over Σ, an *alignment* between P and Q is a pair of strings $A = (P_A, Q_A)$, of the same length $|P_A| = |Q_A|$, over $\Sigma \cup \{\texttt{-}\}$ (the '-' character stands for a *gap*), with P_A (Q_A) being obtained by inserting gaps at any positions in P (Q, resp.), with the restriction that, for any position i, at least one of the characters $P_A[i]$ and $Q_A[i]$ is not a gap. Here are two examples of alignments between peptides $P = \texttt{QAVG}$ and $Q = \texttt{VQG}$:

$$A_1 : \begin{array}{l}\texttt{-QAVG}\\ \texttt{VQG--}\end{array} \qquad A_2 : \begin{array}{l}\texttt{QAV-G}\\ \texttt{--VQG}\end{array}$$

We have, $P_{A_1} = \texttt{-QAVG}$ and $Q_{A_1} = \texttt{VQG--}$, both of length 5. Note that there is no column with two gaps.

Four distances have been introduced by [5], of which we use three that are suitable for our purpose: $\mathrm{d_{ssp}}$, symmetrized sum-of-pairs score, that ignores gaps; $\mathrm{d_{seq}}$, that treats all gaps in the same sequence equally; and $\mathrm{d_{pos}}$, that incorporates positional information about the gaps. The original alignment is

modified to reflect each gap treatment; for our example above, alignment A_1 is converted into the following three alignments, to be used for each of the three distances $\mathrm{d_{ssp}}, \mathrm{d_{seq}}, \mathrm{d_{pos}}$, respectively:

$$\begin{array}{cccccc} & \mathtt{Q} & \mathtt{A} & \mathtt{V} & \mathtt{G} \\ \mathtt{V} & \mathtt{Q} & \mathtt{G} & & \end{array} \qquad \begin{array}{ccccccc} G^1 & \mathtt{Q} & \mathtt{A} & \mathtt{V} & & \mathtt{G} & \\ & \mathtt{V} & \mathtt{Q} & \mathtt{G} & G^2 & G^2 & \end{array} \qquad \begin{array}{ccccccc} G_1^1 & \mathtt{Q} & \mathtt{A} & \mathtt{V} & & \mathtt{G} & \\ & \mathtt{V} & \mathtt{Q} & \mathtt{G} & G_3^2 & G_3^2 & \end{array}$$

Similarly, A_2 is converted into the following three versions:

$$\begin{array}{cccccc} \mathtt{Q} & \mathtt{A} & \mathtt{V} & & \mathtt{G} \\ & & \mathtt{V} & \mathtt{Q} & \mathtt{G} \end{array} \qquad \begin{array}{ccccccc} \mathtt{Q} & & \mathtt{A} & \mathtt{V} & G^1 & \mathtt{G} \\ G^2 & G^2 & \mathtt{V} & \mathtt{Q} & & \mathtt{G} \end{array} \qquad \begin{array}{ccccccc} \mathtt{Q} & & \mathtt{A} & \mathtt{V} & G_3^1 & \mathtt{G} \\ G_0^2 & G_0^2 & \mathtt{V} & \mathtt{Q} & & \mathtt{G} \end{array}$$

The three distances, $\mathrm{d_{ssp}}, \mathrm{d_{seq}}, \mathrm{d_{pos}}$, are then defined using the three versions above, in the same order, using the homology set for each character, $H_X(A)^i_j$, for alignment A, $X \in \{\text{ssp, seq, pos}\}$, sequence i, position j, that contain the characters from the other sequences in the same column as j. For example, $H_{\mathrm{pos}}(A_1)^1_1 = \{\mathtt{Q}\}$, $H_{\mathrm{pos}}(A_1)^1_3 = \{G^2_3\}$. $\mathrm{d_{ssp}}$ is defined as the Jaccard distance [14] on the homology sets H_{ssp}:

$$\mathrm{d_{ssp}}(A_1, A_2) = 1 - \frac{\sum_i \sum_j |H_{\mathrm{ssp}}(A_1)^i_j \cap H_{\mathrm{ssp}}(A_2)^i_j|}{\sum_i \sum_j |H_{\mathrm{ssp}}(A_1)^i_j \cup H_{\mathrm{ssp}}(A_2)^i_j|} \ .$$

The other two distances use symmetric difference on the same sets; for $X \in \{\text{seq, pos}\}$:

$$\mathrm{d}_X(A_1, A_2)^i_j = \frac{|H_X(A_1)^i_j \triangle H_X(A_2)^i_j|}{|H_X(A_1)^i_j \cup H_X(A_2)^i_j|} \ .$$

We refer the reader to [5] for details and examples.

2.6 Normalized relative displacement distance

The fourth distance, $\mathrm{d_N}$, is new. It measures the relative displacement between the positions of the letters in the two strings. Put $|P| = n$ and $|Q| = m$ and let $1 \le p_1 < p_2 < \cdots < p_n \le |P_A|$ be such that $P_A[p_i] = P[i]$, for any $1 \le i \le n$. Similarly, let $q_1 < q_2 < \cdots < q_m$ be such that $Q_A[q_j] = Q[j]$, for any $1 \le j \le m$. For any $i, 1 \le i \le n$, and $j, 1 \le j \le m$, denote the distance between the positions of $P[i]$ and $Q[j]$ in A by $d_A(i, j) = p_i - q_j$.

Given two alignments, A_1 and A_2, between the same strings, we define the *relative displacement distance* between them as:

$$d(A_1, A_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} |d_{A_1}(i, j) - d_{A_2}(i, j)| \ .$$

For example, consider the alignment A_1 in our example above. We have $n = 4, m = 3$, and $(p_1, p_2, p_3, p_4) = (2, 3, 4, 5)$, $(q_1, q_2, q_3) = (1, 2, 3)$. For A_2, we have $(p_1, p_2, p_3, p_4) = (1, 2, 3, 5)$ and $(q_1, q_2, q_3) = (3, 4, 5)$. Then, $d_{A_1}(1, 1) = 2 - 1 = 1$, $d_{A_2}(1, 1) = 1 - 3 = -2$. The distance between the alignments is $d(A_1, A_2) = |1-(-2)|+|0-(-3)|+|-1-(-4)|+|2-(-1)|+|1-(-2)|+|0-(-3)|+|3-0|+|2-(-1)|+|1-(-2)|+|4-2|+|3-1|+|2-0| = 33$.

It can be shown that d is a distance metric whose maximum value it can take is $nm(n+m)$. Thus, we define the *normalized* relative displacement distance as:

$$\mathrm{d_N}(A_1, A_2) = \frac{d(A_1, A_2)}{nm(n+m)} \ .$$

$\mathrm{d_N}$ is still a distance metric and it takes values between 0 and 1. For our example above, we have $\mathrm{d_N}(A_1, A_2) = \frac{33}{84} \approx 0.39286$.

3 Results

3.1 Reference alignments

There are multiple benchmark databases available to validate the accuracy of alignments. Most of these databases include comparatively similar sequences [31], albeit with the addition of features that are often difficult to align such as repeats, and many of the benchmarks were created using simulation techniques [30]. Most alignment algorithms can properly align similar sequences and the purpose of the method reported here is to align very distantly related protein sequences. For distantly related proteins we do not agree that simulation methods have completely captured the vagaries of protein sequence evolution. We therefore wanted to use very distantly related protein sequence alignments that had been done by hand and based on the function of the proteins. To this end we choose to make use of the Conserved Domain Database (CDD), [20], from NCBI. This database has hand curated alignments and furthermore, it lists these alignments in a hierarchical fashion. Thus alignments can be chosen at different degrees of divergence from the same types of proteins.

Several alignments were chosen from CDD to explore different degrees of divergence and to explore different lengths of alignments. Pairs of protein groups were chosen to be very divergent (fold or superfamily level) and then a superfamily or family from within that protein group. Here are several examples: *Globin-like*, cd01067 (consensus alignment length 161 sites with 119 aa) and *Hb-alpha-like*, cd08927 (consensus alignment length 142 sites with 140 aa); *7tm_GPCRs*, cd14964, (consensus alignment length 420 sites with 267 aa) and *7tmA_photoreceptors_insect*, cd15079 (consensus alignment length 301 sites with 292 aa); *NBD_sugar-kinase_HSP70_actin*, cd00012 (consensus alignment length 1154 sites with 185 aa) and *FGGY_YpCarbK_like*, cd07782 (consensus alignment length 660 sites with 509 aa). The difference between the alignment length and the consensus number of aa gives an indication of the divergence within the grouping.

We have chosen 49 multiple sequence alignments from CDD. They are presented in the first four columns of Supplementary Table 1, each with the conserved domain name, source, number of protein sequences, and the length of the alignment. The list is sorted by alignment length.

3.2 Case studies

Before presenting a comparison between the best E-score and BLOSUM matrix, we show in this section two examples of the advantage presented by the alignments produced by the ProtT5-score compared with the BLOSUM62 matrix. The first example is for two sequences from the *KAZAL_FS* (cd00104) MSA. The sequences and the two alignments are in Figure 4. The BLOSUM62 alignment is correct in the first half, whereas the second half is significantly different, with several long gaps added both within and at the end of the sequence.

– Sequence `S1`:

```
>gi|46560124|gb|AAT00502.1|
CPEVHDPVCGTDKVTYPNECDLGLAQCAHPERNITVFARSTC
```

– Sequence `S2`:

```
>gi|24583741|ref|NP_723690.1|
CPRNYEPVCGSNLVTYPNRCEFDCVRRNVERQGRSMGLLRDGTC
```

– ProtT5-score end-gap-free alignment, which is identical with the reference MSA induced alignment:

```
S1: 1   CPEVHDPVCGTDKVTYPNECDLGLAQCAH--PERNITVFARSTC   42
        CP    PVCG   VTYPN C             R        TC
S2: 1   CPRNYEPVCGSNLVTYPNRCEFDCVRRNVERQGRSMGLLRDGTC   44
```

– BLOSUM62 matrix end-gap-free alignment:

```
S1: 1   CPEVHDPVCGTDKVTYPNECDLGLAQCAHPERNITVFARSTC-------   42
        CP  ++PVCG++ VTYPN C+     C    RN+    RS
S2: 1   CPRNYEPVCGSNLVTYPNRCEF---DCVR--RNVERQGRSMGLLRDGTC   44
```

Figure 4: Two sequences from the *KAZAL_FS* (cd00104) MSA and their pairwise end-gap-free alignments produced using ProtT5-score and BLOSUM62 matrix. The ProtT5-score alignment is identical with the one induced by the reference MSA.

The second example involves two sequences from the *globin_sensor* (cd01068) MSA. The sequences and the two alignments are shown in Figure 5. In this example, the BLOSUM62 alignment diverges drastically from the true alignment; the algorithm is deceived by a number of identities between the prefix of the first sequence and the middle of the second one.

It appears that the BLOSUM62 matrix often introduces false gaps to achieve a higher number of matches. That is, it blindly works to achieve a maximum score while the ProtT5-score avoids this trap, due to its contextual nature. For example, note in the first example that BLOSUM62 removes three matches to instead introduce five matches, one of which is a high scoring C match. The ProtT5-score apears to be taking more information from the properties of the amino acids than are encoded in the matrix or capable of being encoded in the matrix without context.

– Sequence `S1`:

```
>gi|385735151|gb|EIG55349.1|
RFRELLGLSELTENPLARWAGTLGADPEAFADRIQAFVLEMPQLRIVQENLPDPAKLRRNWGEWFSGLF
ARGLGDELLTRLWRSGQSHVAYNVDHRLISTAYALARAYLHEKAAESLPALELAGALRSIDAMVDFSLL
VETDAF
```

– Sequence `S2`:

```
>gi|445776913|gb|EMA27889.1|
WRKELVGFDQQDAERVEALEGLFADNADEIASRFYGNLTEHERTQSVIDRSPKGIDELERTQQAYVRTL
TEGTYGTDYFRNRARIGKLHELLEMPVAQYVGQYGFYFELLFEVMRERRADRIAATLEESGVEDETIAE
VRSQVETTSEETLSVLKLLMLDMQVAMETY
```

– ProtT5-score end-gap-free alignment, which is identical with the reference MSA induced alignment:

```
S1: 1   RFRELLGLSELTENPLARWAGTLGADPEAFADRIQAFVLEMPQLRIVQENLP-DPAKLRR   59
           EL G              G         A R      E      V    P     L R
S2: 1   WRKELVGFDQQDAERVEALEGLFADNADEIASRFYGNLTEHERTQSVIDRSPKGIDELER   60

S1: 60  NWGEWFSGLFARGLGDELLTRLWRSGQSHVAYNVDHRLISTAYALARAYLHEKAAESLP-  118
                L     G        R G  H             Y      L E   E
S2: 61  TQQAYVRTLTEGTYGTDYFRNRARIGKLHELLEMPVAQYVGQYGFYFELLFEVMRERRAD  120

S1: 119 -----------------------ALELAGALRSIDAMVDFSLLVETDAF  144
                                             S          V
S2: 121 RIAATLEESGVEDETIAEVRSQVETTSEETLSVLKLLMLDMQVAMETY  168
```

– BLOSUM62 matrix end-gap-free alignment:

```
S1: 1   ------------------------------------------------------------    0

S2: 1   WRKELVGFDQQDAERVEALEGLFADNADEIASRFYGNLTEHERTQSVIDRSPKGIDELER   60

S1: 1   --------------------RFRELLG-LSELTENPLARWAGTLGADPEAF--------AD   32
                             R R  +G L EL E P+A++ G  G   E          AD
S2: 61  TQQAYVRTLTEGTYGTDYFRNRARIGKLHELLEMPVAQYVGQYGFYFELLFEVMRERRAD  120

S1: 33  RIQAFVLEMPQLRIVQENLPDPAKLRRNWGEWFSGLFARGLGDELLTRLWRSGQSHVAYN   92
        RI A + E          +    +           L    +   + T
S2: 121 RIAATLEESGVEDETIAEVRSQVETTSEETLSVLKLLMLDMQVAMETY------------  168

S1: 93  VDHRLISTAYALARAYLHEKAAESLPALELAGALRSIDAMVDFSLLVETDAF  144

S2: 169 ---------------------------------------------------  168
```

Figure 5: Two sequences from the *globin_sensor* (cd01068) MSA and their pairwise end-gap-free alignments produced using ProtT5-score and BLOSUM62 matrix. The ProtT5-score alignment is identical with the one induced by the reference MSA.

3.3 Comparison procedure

We compare the E-scores against the BLOSUM matrices on the 49 MSAs introduced above as follows. For each MSA, we pick 300 randomly selected protein pairs (if the total number of pairs is less than 300, then we pick all pairs), compute pairwise alignment of each pair using either the E-score or BLOSUM matrix, according to the Needleman-Wunsch dynamic programming global alignment algorithm [23], but without any penalties for gaps at the ends; we call this *end-gap-free* alignment. This is the most suitable alignment type, as we expect these sequences to align globally, however, using end-gap penalties could artificially alter the alignments. We used experimentally set gap penalties: -1 for gap opening and -0.2 for gap extension.

Each pairwise alignment is then evaluated against the reference alignment from CDD by computing the distance to the pairwise alignment induced by the reference MSA for the two sequences. The distance is computed using all four distances

introduce above: d_N, d_{ssp}, d_{seq}, and d_{pos}.

The comparison for all E-scores and all BLOSUM matrices on all data is very time consuming, therefore, we need to choose the best E-score and the best BLOSUM matrix. For that purpose, we have selected six representative MSAs of the above 49; see the first column of Supplementary Tables 4-7. We have first compared all six E-scores on these six MSAs; the results are shown in Supplementary Table 4. The two top performers are the ProtT5-score and the ESM2-score. They are compared head-to-head in Supplementary Table 5, using the same six MSAs, but considering also the Wilcoxon signed-rank test P-values, with a threshold of 0.01, to check the validity of the findings. The result is that ProtT5 is the best E-score.

Similarly, to find the best BLOSUM matrix for our purposes, we performed the same procedure. Here we had three candidates, BLOSUM50, BLOSUM62 and BLOSUM80, which we compared head-to-head. Two of these comparisons ended in ties (with one P-value in favour of BLOSUM50 over BLOSUM80), while BLOSUM62 outperformed BLOSUM50. Therefore, BLOSUM62 is the best matrix for our purpose. Note that Wilcoxon P-values are now much larger, almost all above the .01 threshold; see Supplementary Tables 6-7. This appears to indicate that it makes little difference which of the three BLOSUM matrices we use for comparison.

3.4 ProtT5-score vs BLOSUM62 matrix

In this section we compare head-to-head the ProtT5-score against the BLOSUM62 matrix, the top performers, as decided by the previous section. We have compared the two, as described above, on all data; see Supplementary Table 1, where the Wilcoxon test P-values larger than .01 are indicated in red. We have removed all tests with P-values higher than .01 and present the comparison for the d_N and d_{pos} distances in Table 1; the results for the d_{ssp} and d_{seq} distances are similar with those for the d_{pos} distance; see Supplementary Table 1. ProtT5-score is better in all but two tests on the shortest MSAs. The standard deviation values and the number of better cases for each scoring method are given in the Supplementary Tables 2-3. Regarding the overall number of better cases, ProtT5 has 66.47% of the cases, BLOSUM62 14.77%, while for the remaining 18.76% they were equal. However, the distribution of these better cases, together with the performance on each individual case, are so that the ProtT5-score performs better in all but the first two cases on the average.

For a better visualization of the comparison, we plotted in Figure 6, for each of the four distances, the ratio between the distance to the reference alignment from the BLOSUM62 alignment and from the ProtT5 alignment, respectively. That means higher than 1 is better in favour of ProtT5. It is very clear from the figure that ProtT5 has a much better similarity with the reference alignment, the advantage being most clear for the d_N distance.

The comparison for the purely global alignment case (where end-gaps are considered) is presented in the Supplementary Tables 8-10. The conclusions are similar with those presented here for the end-gap-free case.

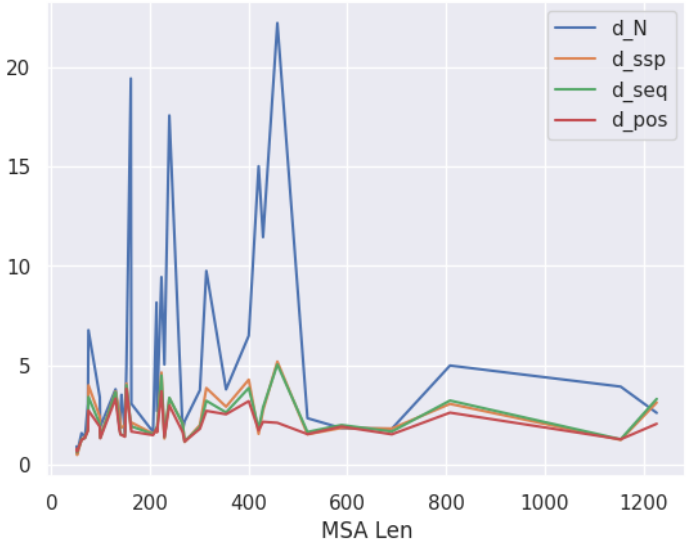


Figure 6: The ratio between the performance of BLOSUM62 and ProtT5-score in terms of distances to the reference from their alignments. Higher than 1 indicates ProtT5-score is better.

4 Discussion and conclusion

We have proposed a new method for scoring sequence alignments, E-score, using the cosine similarity between the vectors associated using any word embedding algorithm, E. When using contextual embeddings, the E-score has the advantage of incorporating into the score associated with each position contextual information from the neighbouring positions in the sequence.

We have thoroughly tested the new E-score in the context of protein alignments, by comparing the top scoring function, ProtT5-score, against the BLOSUM62 matrix on a large set including a wide variety of reference multiple sequence alignments from NCBI's Conserved Domain Database. The pairwise alignments produced by using the ProtT5-score are much closer to the true alignments, as induced by the reference multiple sequence alignments, compared to BLOSUM62.

While we have tested the new E-score in the context of biological sequences, particularly proteins, the new method can be applied to any sequences where computation of embedding vectors is possible, which include practically all existing textual data, the largest type of data available. All Large Language Models (LLM) have the possibility to produce word embedding vectors for any input text. Considering the importance and implications of text similarity in general, the potential for applications of our new method appears to be virtually boundless.

The language models are expected to improve at a rapid pace in the future, with expected corresponding improvements for those involving biological sequences. The ideas

Table 1: ProtT5-score against BLOSUM matrix: comparison using 37 MSAs, for which the Wilcoxon test P-values are smaller than .01. The best values are shown in boldface. We show only the distances d_N and d_{pos}; the comparisons using d_{ssp} and d_{seq} are similar with the one for d_{pos}.

MSA				d_N			d_{pos}		
Conserved domain	Source	Seq.	Len.	ProtT5	BLOSUM62	P-values	ProtT5	BLOSUM62	P-values
DEFL_defensin-like	cd21806	78	51	0.00440	**0.00402**	1.12E-03	0.13315	**0.10519**	3.88E-08
Bbox2_TRIM37_C-VIII	cd19779	25	52	0.00166	**0.00101**	2.05E-08	0.05781	**0.03770**	1.92E-06
CBD_like	cd12204	41	61	**0.00558**	0.00881	4.80E-15	**0.17165**	0.21346	4.92E-11
Bbox1_CYLD	cd19816	27	68	**0.00500**	0.00691	1.10E-05	**0.15611**	0.20869	1.77E-19
KAZAL_FS	cd00104	273	74	**0.00383**	0.01167	3.17E-24	**0.15631**	0.27250	9.83E-23
bHLH_SF	cd00083	79	75	**0.00265**	0.01790	8.88E-27	**0.09477**	0.25772	7.49E-27
CD_CSD	cd00024	522	98	**0.00197**	0.00668	1.95E-32	**0.10842**	0.20584	4.41E-32
C1	cd00029	281	99	**0.00497**	0.00947	2.56E-12	**0.19219**	0.25412	5.55E-12
TrHb	cd14756	8	130	**0.00137**	0.00518	3.27E-04	**0.10662**	0.35440	1.96E-04
Hb	cd14765	15	138	**0.00031**	0.00055	3.29E-03	**0.04222**	0.07250	1.65E-08
Hb-beta-like	cd08925	27	140	**0.00017**	0.00045	1.50E-12	**0.02242**	0.03413	1.21E-10
Hb-alpha-like	cd08927	39	142	**0.00004**	0.00015	1.88E-03	**0.01522**	0.02261	9.53E-03
Mb	cd08926	9	149	**0.00005**	0.00008	9.13E-03	**0.01415**	0.01984	5.09E-03
GS_GGDEF_2	cd14759	26	152	**0.00036**	0.00176	9.01E-36	**0.03284**	0.12525	1.69E-37
Globin-like	cd01067	17	161	**0.00843**	0.16366	9.05E-21	**0.42919**	0.83581	6.99E-21
FHb-globin	cd08922	47	162	**0.00012**	0.00037	2.72E-23	**0.02151**	0.03560	2.51E-21
SH2_STAT_family	cd09919	67	206	**0.00758**	0.01242	9.29E-21	**0.21899**	0.32420	1.64E-39
PBP-like	cd08919	31	213	**0.00507**	0.04135	6.66E-33	**0.22913**	0.41271	1.15E-42
SH2	cd00173	352	214	**0.01128**	0.03033	1.53E-38	**0.31303**	0.51107	2.58E-41
globin_sensor	cd01068	193	223	**0.00099**	0.00935	3.11E-46	**0.07105**	0.26224	1.76E-47
PFM_monalysin-like	cd17904	31	229	**0.00150**	0.00755	3.32E-34	**0.22581**	0.31391	1.71E-32
Mb-like	cd01040	384	239	**0.00243**	0.04274	2.56E-45	**0.17458**	0.51810	3.20E-48
FYVE_like_SF	cd00065	392	266	**0.01506**	0.02576	3.86E-34	**0.20270**	0.33013	3.06E-36
PFM_aerolysin_family	cd10140	65	270	**0.01442**	0.03145	6.72E-30	**0.64491**	0.74921	2.82E-26
7tmA_photoreceptors_insect	cd15079	35	301	**0.00020**	0.00076	6.89E-38	**0.04523**	0.08140	9.52E-38
7tmA_Melanopsin-like	cd15083	12	314	**0.00013**	0.00123	2.54E-10	**0.04100**	0.11071	2.61E-10
ClyA_AhlB-like	cd22652	39	354	**0.00017**	0.00064	3.55E-31	**0.02743**	0.06931	2.45E-33
7tmA_Opsins_type2_animals	cd14969	71	400	**0.00047**	0.00303	5.33E-47	**0.06303**	0.20091	4.08E-50
7tm_GPCRs	cd14964	18	420	**0.00889**	0.13344	6.27E-24	**0.53411**	0.92090	2.09E-23
7tmA_Anaphylatoxin_R-like	cd14974	18	429	**0.00112**	0.01285	7.92E-23	**0.10699**	0.23012	2.09E-23
7tmA_Opioid_R-like	cd14970	16	458	**0.00093**	0.02061	2.17E-18	**0.11708**	0.24558	5.84E-19
ClyA-like	cd21116	118	519	**0.02048**	0.04763	4.64E-40	**0.44748**	0.68647	7.03E-46
FGGY	cd00366	31	588	**0.00257**	0.00472	2.28E-08	**0.18464**	0.35333	2.98E-50
FGGY_YpCarbK_like	cd07782	35	690	**0.00074**	0.00133	1.63E-44	**0.06808**	0.10353	7.27E-48
7tm_classA_rhodopsin-like	cd00637	405	808	**0.00247**	0.01231	7.20E-48	**0.17469**	0.45518	6.33E-51
NBD_sugar-kinase_HSP70_actin	cd00012	125	1154	**0.03795**	0.14892	2.10E-48	**0.71817**	0.91310	1.07E-49
7tmA_amine_R-like	cd14967	78	1227	**0.00997**	0.02596	5.39E-49	**0.18722**	0.38419	7.51E-51

presented apply directly to any new model, meaning that the E-score will improve as well.

Specialized models, obtained usually by fine tuning the general ones, or adapting them to specialized tasks, are expected to provide improved performance. Regarding the case of biological sequences, where alignments play a particularly vital role, the possibility exists for training, or fine tuning, models for highly specialized tasks, such as specific types of proteins.

Finally, aligning nucleic acid sequences – DNA, RNA – though simpler than the case of proteins, is a whole area where E-scores can be applied as well.

5 Availability

The program to compute alignments based on various E-scores is available as a web server at `e-score.csd.uwo.ca`. The source code is freely available for download from `github.com/lucian-ilie/E-score`.

6 Author contributions statement

L.I. proposed the problem, defined the E-score, proposed the evaluation using pairwise alignments induced by MSAs, introduced the new normalized relative displacement distance, supervised the implementation and testing, and wrote

the manuscript. S.A. implemented the alignment programs, downloaded and installed the embedding models, downloaded and processed the data, performed all testing, and built the GitHub page and web server. G.B.G. proposed the Conserved Domain Database, selected all reference MSAs used for testing, identified the six representative ones used for pre-testing, and wrote the description of the data. All authors read and approved the final version of the manuscript.

7 Acknowledgments

SeyedMohsen Hosseini provided assistance with building the web server and installing some embedding models. All computations were performed on Compute Canada servers.

8 Funding

This research was funded by NSERC Discovery Grants R3143A01 to L.I. and RGPIN-2020-05733 to G.B.G.

9 Supplementary data

A supplementary file containing Supplementary Tables 1-11 is available online.

References

[1] Uniprot: the universal protein knowledgebase. Nucleic acids research, 45(D1):D158–D169, 2017.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. Journal of molecular biology, 215(3):403–410, 1990.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research, 25(17):3389–3402, 1997.

[4] E. Asgari and M. R. Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. PloS one, 10(11):e0141287, 2015.

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

[6] M. Dayhoff, R. Schwartz, and B. Orcutt. 22 a model of evolutionary change in proteins. Atlas of protein sequence and structure, 5:345–352, 1978.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] 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.

[9] 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.

[10] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost. Modeling aspects of the language of life through transfer-learning protein sequences. BMC bioinformatics, 20(1):1–17, 2019.

[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. bioRxiv, pages 2023–06, 2023.

[13] S. Hosseini and L. Ilie. PITHIA: Protein Interaction Site Prediction Using Multiple Sequence Alignments and Attention. International Journal of Molecular Sciences, 23(21):12814, 2022.

[14] P. Jaccard. The distribution of the flora in the alpine zone. 1. New phytologist, 11(2):37–50, 1912.

[15] 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.

[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] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[20] 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.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[22] A. Nambiar, M. Heflin, S. Liu, S. Maslov, M. Hopkins, and A. Ritz. Transforming the language of life: transformer neural networks for protein prediction tasks. In Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics, pages 1–8, 2020.

[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] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[25] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. arXiv:1802.05365, 2018.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683, 2020.

[27] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. MSA transformer. In International Conference on Machine Learning, pages 8844–8856. PMLR, 2021.

[28] 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.

[29] 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.

[30] J. Stoye, D. Evers, and F. Meyer. Rose: generating sequence families. Bioinformatics (Oxford, England), 14(2):157–163, 1998.

[31] 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.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

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

[34] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32, 2019.

[35] Q. Yuan, J. Chen, H. Zhao, Y. Zhou, and Y. Yang. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics, 38(1):125–132, 2022.