Preliminary statistical analysis of amino acid sequence embeddings of proteins

Krzysztof Fidelis[1,†], Mykhailo Luchkevych[2,†] and Yaroslav Teplyi[2,*,†]

[1] Genome Center, UC Davis, Davis, California, USA

[2] Lviv Polytechnic National University, Stepan Bandera Street 12 79013 Lviv, Ukraine

Abstract

Recent extensive research in the field of bioinformatics aimed at predicting the 3D structure of proteins from their amino acid sequence using LLMs has generated large datasets of numerical data on the relative positioning of amino acids in sequences, known as embeddings. These data banks are publicly accessible, enabling their analysis and utilization, particularly for tasks such as identifying sets of typical elements of protein structures. Recognizing typical substructures could significantly simplify the protein analysis process, which involves more than 240 millions of proteins.

This work explores the main statistical characteristics of amino acid sequence embeddings of protein pairs, both significantly similar and distinctly different in composition and structure, in order to identify patterns in their behavior parameters: linearity, stationarity, probability distribution laws, and others, ensuring the correctness of applying corresponding models and methods in the future.

Keywords

Protein structure analysis, ESM-2 model, sequence embeddings, statistical analysis, sequence alignment

1. Introduction

The use of Large Language Models (LLMs) has revolutionized various fields of study, extending their impact to the domain of protein amino acid sequence analysis [1]. Recent innovations have leveraged LLMs to decode protein sequences, significantly advancing our understanding and capabilities in constructing detailed spatial structure databases. Among these innovations, the ESM-2 database [2] stands out as a pivotal development. ESM-2, an open-access database, encapsulates an extensive array of protein spatial structures, thus providing a valuable resource for biochemists and bioinformatics researchers. This enables an in-depth exploration of the functional attributes of proteins in correlation with their three-dimensional conformations.

Utilizing LLMs for these purposes transforms the representation of proteins into a multidimensional vector space where each amino acid's embedding reflects its potential spatial relationships within the protein's folded structure. This approach not only enhances the precision of structural predictions but also introduces a quantitative method to assess the likelihood of proximal interactions among the amino acids in a given protein. The effectiveness and accuracy of these models are rigorously evaluated through the Critical Assessment of protein Structure Prediction (CASP) project [3], where computational predictions are juxtaposed with experimentally determined structures, validating the reliability of the models.

Central to the effectiveness of LLMs is the preliminary processing and statistical analysis of data. The architecture of the system and the specific algorithms employed, particularly the deep learning components of LLMs, critically influence the characteristics of the resulting embedding arrays. This initial data processing phase is crucial as it ensures that subsequent analyses and applications of the data are based on robust and reliable foundation.

MoMLeT-2024: 6th International Workshop on Modern Machine Learning Technologies, May, 31 - June, 1, 2024, Lviv-Shatsk, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

kfidelis@ucdavis.edu (K. Fidelis); luchkevychmm@gmail.com (M. Luchkevych); yaroslav.b.teplyi@lpnu.ua (Y. Teplyi)

0000-0002-8061-412X (K. Fidelis); 0000-0002-2196-252X (M. Luchkevych); 0000-0002-0877-7063 (Y. Teplyi)

2. Analysis of recent research

Protein language models (pLMs) have significantly advanced our understanding of the relationships within protein sequences, providing a numerical representation of their structural and evolutionary features. Recent developments, such as the Embedding-based Alignment (EBA), specifically an approach introduced in [4], highlights the potential of using high-dimensional sequence embeddings from pLMs in protein structure analysis. This approach was effectively used to detect distant homologies in the so-called 'twilight zone' [5] where sequence similarities are not readily apparent.

The authors demonstrate that EBA surpasses both traditional sequence alignment methods and other pLM-based approaches in detecting structural similarities, without the need for training or parameter optimization. The use of embeddings allows EBA to capture deeper evolutionary relationships, offering a significant improvement in identifying structural similarities in proteins with low sequence identity.

We utilize similar approach, where our research aims to expand on these results by exploring a variation of the EBA method, focusing specifically on the statistical characterization of sequence embeddings through autocorrelation and correlation analysis. Our methodology differs from the proposed EBA in [4] by analysing the spatial relationships within protein sequences, which are encoded in the embeddings generated by models like ESM-2.

Findings from [6], demonstrated that the statistical distribution of amino acid sequences supports Darwinian evolution. Their research showed that certain peptide combinations occur rarely, suggesting evolutionary constraints. These constraints may be reflected in the distribution of embeddings, which could indicate evolutionary pressure shaping protein structures. The confirmation of the statistical nature of amino acid distributions complements the statistical analysis of sequence embeddings.

3. Research purpose

This study conducts a preliminary analysis of protein sequence embeddings using autocorrelation functions. The goal is to detect repetitive patterns within these embeddings that may indicate underlying structural or functional elements in the proteins. By employing a sliding window across the sequence embedding dimensions, we assess the local repetitiveness of these patterns to identify motifs suggestive of structural features.

Additionally, this analysis evaluates the similarity between protein pairs by correlating their autocorrelation outcomes, providing a detailed view of how similarities are distributed across the entire sequence. This work sets the foundation for developing a variation of the Embedding-based Alignment (EBA) method.

4. Problem formulation

Given a language model, denoted as L, for predicting the three-dimensional structure of a protein from the sequence S. The input data for the model is the sequence of amino acids in the protein $S = \{s_1, s_2, \ldots, s_N\}$, where s_i represents an individual element of the sequence (the letter corresponding to the amino acid), and N indicates the number of amino acids in the sequence. The model L is defined by a set of pre-learned parameters θ.

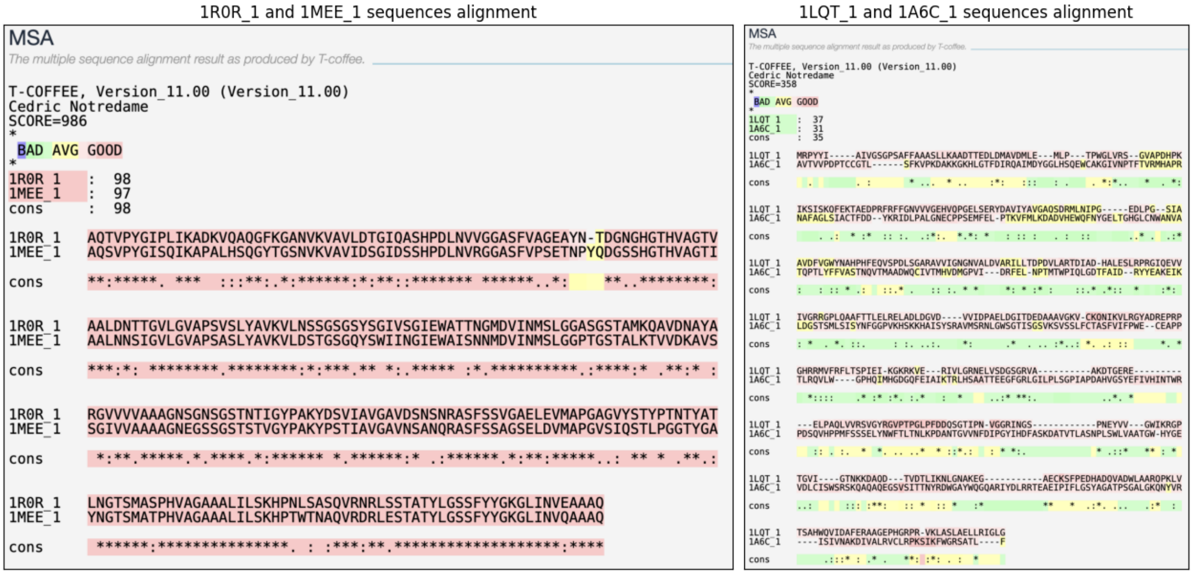
After processing the sequence S, the model outputs a set of parameters. The mapping of S to Y can be formally described as the function $f_\theta(S) = Y$, where f_θ summarizes the computational logic of model L with parameters θ.

Among the set of output parameters Y, we focus on one specific parameter S^s, which is the subjects of this study. This parameter provides an internal protein's sequence representation in a form of vectors, also called embeddings. The parameter S^s is a matrix that represents the mapping of the input sequence into a higher-dimensional space $(N, 1024)$, where N is a length of the sequence. Hence, $S^s \in R^{N \times 1024}$, where each row in S^s corresponds to an element from S transformed into a 1024-dimensional vector (embedding), which encodes contextual information about the given element, its properties, and its interaction with other elements in the sequence.

5. Statistical Analysis of Sequence Embeddings

5.0.1. Sequence Alignment and Similarity Metrics

The proteins selected for this analysis are $1R0R_1$ and $1MEE_1$, which have been identified as similar, as well as $1LQT_1$ and $1A6C_1$, which are considered dissimilar. The criteria for this categorization are based on structural features and evolutionary relationships inferred from sequence homology.



1R0R_1 and 1MEE_1 sequences alignment

MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=986
*
 BAD AVG GOOD
*
1R0R_1    :  98
1MEE_1    :  97
cons      :  98

1R0R_1    AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDLNVVGGASFVAGEAYN-TDGNGHGTHVAGTV
1MEE_1    AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETNPYQDGSSHGTHVAGTI

cons      **:*****. ***  :::**:.*:******:*:**::******* ******..*:    **..********:


1R0R_1    AALDNTTGVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIEWATTNGMDVINMSLGGASGSTAMKQAVDNAYA
1MEE_1    AALNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWIINGIEWAISNNMDVINMSLGGPTGSTALKTVVDKAVS

cons      ***:*: ********.********:*:***.** *:.***** :*.**********.:****:* .**:* :


1R0R_1    RGVVVVAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVDSNSNRASFSSVGAELEVMAPGAGVYSTYPTNTYAT
1MEE_1    SGIVVAAAAGNEGSSGSTSTVGYPAKYPSTIAVGAVNSANQRASFSSAGSELDVMAPGVSIQSTLPGGTYGA

cons       *:**.*****.*.****.*:****** *.******:* .:******.*:**:*****..: ** * .**.:


1R0R_1    LNGTSMASPHVAGAAALILSKHPNLSASQVRNRLSSTATYLGSSFYYGKGLINVEAAAQ
1MEE_1    YNGTSMATPHVAGAAALILSKHPTWTNAQVRDRLESTATYLGSSFYYGKGLINVQAAAQ

cons       ******:**************. : :***:**.******************:****
```

1LQT_1 and 1A6C_1 sequences alignment

MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=358
*
 BAD AVG GOOD
*
1LQT_1    :  37
1A6C_1    :  31
cons      :  35

1LQT_1    MRPYYI-----AIVGSGPSAFFAAASLLKAADTTEDLDMAVDMLE---MLP---TPWGLVRS--GVAPDHPK
1A6C_1    AVTVVPDPTCCGTL------SFKVPKDAKKGKHLGTFDIRQAIMDYGGLHSQEWCAKGIVNPTFTVRMHAPR

cons        .         . :       * ...  * ..     :*:   :::    : .     . *:*..   *  . *:


1LQT_1    IKSISKQFEKTAEDPRFRFFGNVVVGEHVQPGELSERYDAVIYAVGAQSDRMLNIPG-----EDLPG--SIA
1A6C_1    NAFAGLSIACTFDD--YKRIDLPALGNECPPSEMFEL-PTKVFMLKDADVHEWQFNYGELTGHGLCNWANVA

cons          . .:   * :*  :: :.  .:*:.  *.*: *   : :: :    . :  ::         ..* .  .:*


1LQT_1    AVDFVGWYNAHPHFEQVSPDLSGARAVVIGNGNVALDVARILLTDPDVLARTDIAD-HALESLRPRGIQEVV
1A6C_1    TQPTLYFFVASTNQVTMAADWQCIVTMHVDMGPVI---DRFEL-NPTMTWPIQLGDTFAID--RYYEAKEIK

cons      :   : :: * .:   ::.* .   :: :. * *     *: * :* :     ::.* .*::  *    :*:


1LQT_1    IVGRRGPLQAAFTTLELRELADLDGVD----VVIDPAELDGITDEDAAAVGKV-CKQNIKVLRGYADREPRP
1A6C_1    LDGSTSMLSISYNFGGPVKHSKKHAISYSRAVMSRNLGWSGTISGSVKSVSSLFCTASFVIFPWE--CEAPP

cons      : *  . *. ::.     : :. ..:.     *:        .*  . .. :*..: *. .: ::       *. *


1LQT_1    GHRRMVFRFLTSPIEI-KGKRKVE---RIVLGRNELVSDGSGRVA----------AKDTGERE---------
1A6C_1    TLRQVLW----GPHQIMHGDGQFEIAIKTRLHSAATTEEGFGRLGILPLSGPIAPDAHVGSYEFIVHINTWR

cons        *::::     .* :* :*. :.*    :   *      ..:* **:.               ..*. *


1LQT_1    ---ELPAQLVVRSVGYRGVPTPGLPFDDQSGTIPN-VGGRINGS-----------PNEYVV---GWIKRGP
1A6C_1    PDSQVHPPMFSSSELYNWFTLTNLKPDANTGVVNFDIPGYIHDFASKDATVTLASNPLSWLVAATGW-HYGE

cons         :: . :.   *   *. .. ..*   * ::*.:    : * *:.               * .::*   ** : *


1LQT_1    TGVI----GTNKKDAQD---TVDTLIKNLGNAKEG----------AECKSFPEDHADQVADWLAARQPKLV
1A6C_1    VDLCISWSRSKQAQAQEGSVSITTNYRDWGAYWQGGQARIYDLRRTEAEIPIFLGSYAGATPSGALGKQNYVR

cons      ..:       :::  :**:    :: *   :: *    :*              **    *    .:*. ...     .:*   :


1LQT_1    TSAHWQVIDAFERAAGEPHGRPR-VKLASLAELLRIGLG
1A6C_1    ----ISIVNAKDIVALRVCLRPKSIKFWGRSATL----F

cons            .:::* :  .*  .     **:  :*:  .  :    *
```

Figure 1: MSA alignment of a pair of similar proteins $1R0R_1$ and $1MEE_1$ and a pair of dissimilar proteins $1LQT_1$ and $1A6C_1$

Multiple sequence alignment (MSA), is a crucial tool in bioinformatics and has been employed to align the amino acid sequences of the chosen proteins. The quality of alignment is quantified by an MSA score, which assesses the degree of conservation and similarity between sequences. A higher score denotes a greater level of similarity. Alignments were generated using the T-Coffee program [7].

The alignment results (Figure 1) for the similar proteins, $1R0R_1$ and $1MEE_1$, show a high degree of conservation, as indicated by a score of 986. The corresponding MSA visual (left side of the attached figure) shows a significant number of residues are identical (marked by an asterisk '*') or have strong similarities (marked by a colon ':' or a period '.'). This suggests these proteins may share functional and structural properties.

In contrast, the MSA for dissimilar proteins, $1LQT_1$ and $1A6C_1$, yields a score of 358, reflecting a low level of similarity. The alignment (right side of the attached figure) has fewer conserved residues and indicates considerable variation between these sequences, which means they have different functions or structures.

These MSA results provide a baseline for the subsequent statistical analysis. By establishing the degree of similarity through MSA scores and visual inspection, we can set expectations for how these similarities or differences might manifest in various statistical measures such as embeddings' magnitude distributions, distance, angles etc.

5.1. Outlier Normalization

Prior to the application of statistical methods to analyze protein sequence embeddings, initial view of the embeddings showed the presence of extreme outlier values. These outlier values are significantly higher/lower than the general dataset, observed consistently at identical indices across all sequence embeddings. Due to their magnitude, these outliers have the potential to influence subsequent statistical computations, thereby skewing the analysis results.

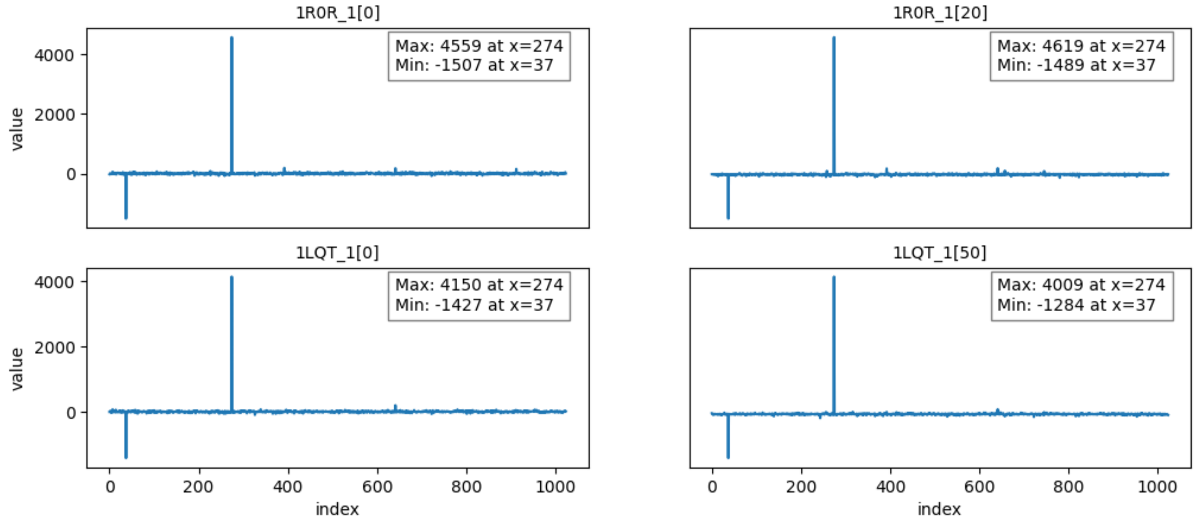


Figure 2: Embedding values before normalization showing outliers

The visualization of the embeddings is depicted in the first set of plots, illustrating spikes (Figure 2). These peaks are consistent across all embeddings of different proteins, suggesting a systematic anomaly of the ESM-2 model rather than random or natural variation within the protein structure representation.

To address this anomaly, a normalization method was applied, where the top five maximum and the bottom five minimum values were adjusted by taking the average value of two adjacent values. This threshold of first five values was chosen based on empirical observations and analysis to effectively remove the outliers without affecting the informational content of the embeddings.

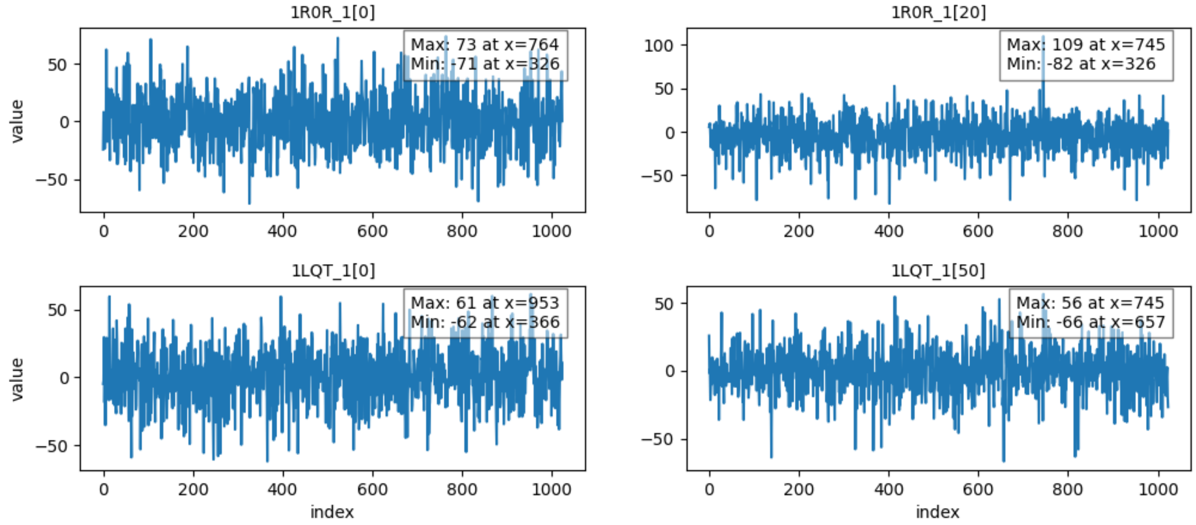


Figure 3: Embedding values after normalization with outliers adjusted

The second set of plots (Figure 3) displays the embeddings after normalization, where the previously visible spikes have been truncated. The consistency across different embeddings indicates that the removing top and bottom five values is addressing the issue. This normalization step is critical as it ensures that the subsequent analytical methods are reflective of structural properties rather than artifacts introduced by outlier data points.

5.2. Distribution Analysis of Embedding Dimensions

The visualization presented in the Figure 4 demonstrates the distribution of normalized embedding values across selected dimensions of protein sequences. Histograms are utilized to compare the distributions between protein pairs that are considered similar and dissimilar, respectively.

For similar proteins ($1R0R_1$ and $1MEE_1$), the first two histograms in the top row represent the distribution of values at specific dimensions (index 10 and index 100). The distributions are noticeably overlapping, meaning a high degree of similarity in these embedding dimensions. This suggests that the embeddings capture similar structural or functional features within this dimension.

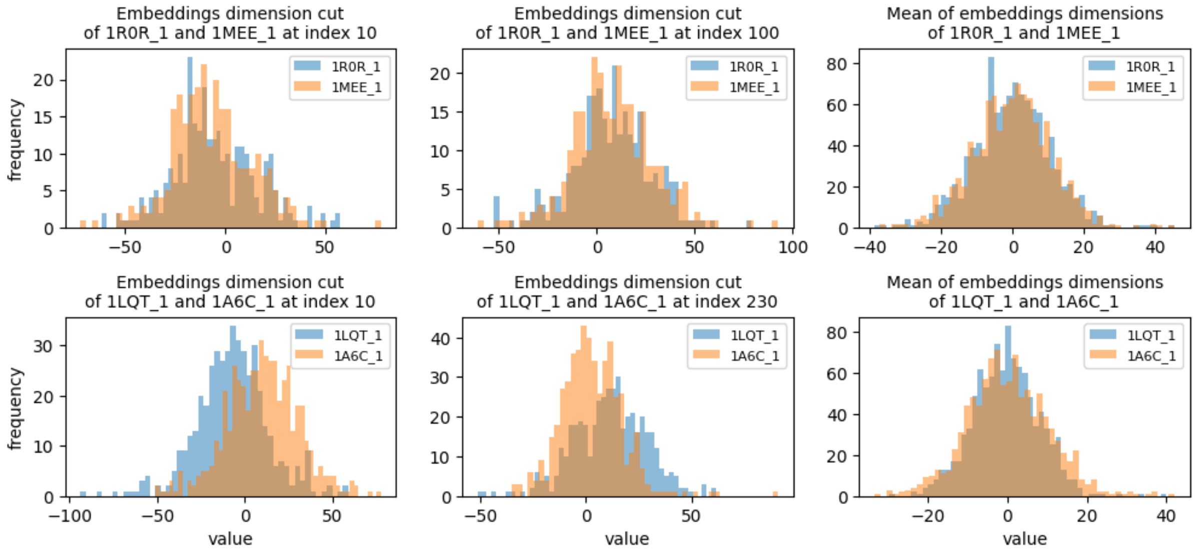


Figure 4: Comparative histograms of embedding dimensions for two protein sequence pairs.

The third histogram in the top row depicts the mean value distribution across all dimensions. The concentration of values around the center and the bell-shaped distribution is indicative of the embeddings capturing a consistent pattern across dimensions.

Conversely, the bottom row compares the dimension value distributions for proteins $1LQT_1$ and $1A6C_1$, which are dissimilar. Here, the first two histograms (index 10 and index 230) show a shift in the frequency of values, suggesting a difference in the structural or functional properties encoded by these dimensions, though it's not always the case, as some

The mean value histogram for these dissimilar proteins shows a distribution is overlapping with the one observed in similar proteins. This indicates that while there is a commonality in the overall embedding pattern (as shown by the shape of the distribution), the distribution across specific dimensions may differ.

The central, bell-shaped distributions seen across the mean histograms shows the consistency and validity of using distribution analyses in protein structure comparison studies. This consistency also suggests that the embeddings may follow an underlying statistical distribution.

5.3. Embedding Magnitude Analysis and Statistical Measures

The computation of embedding magnitudes serves to quantify the strength or intensity of the protein sequence embeddings. We assess the stationarity of this magnitude distribution, as stationary processes allow for the reliable application of statistical measures such as mean, median, and variance over time, yielding consistent and interpretable results across different segments of the sequence.

The histograms in the Figure 5 provide the frequency distribution of embedding magnitudes for both similar and dissimilar proteins. Contrary to initial expectations, the magnitude distributions between similar ($1R0R_1$ and $1MEE_1$) and dissimilar ($1LQT_1$ and $1A6C_1$) protein pairs are analogous, which means that the magnitude alone does not distinguish between the similarities or differences in protein structures.

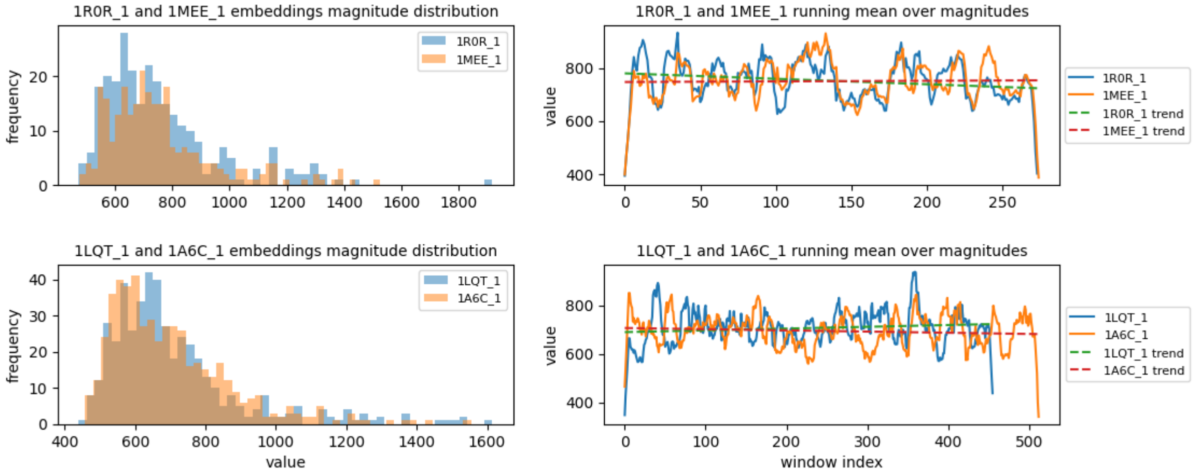


Figure 5: Distribution of embedding magnitudes and running mean.

For the similar proteins, the running mean over magnitudes show a considerable overlap and the local fluctuations are largely aligned, indicating that the embedding magnitudes change similarly over the course of the sequences. This alignment in local fluctuations suggests that the similar proteins have analogous dynamic behaviors in their structure over time.

In contrast, the running mean plots for dissimilar proteins do not show the same degree of overlap or alignment. While the overall trend lines appear to be stationary for both similar and dissimilar proteins, the patterns of local fluctuation provide evidence that the embeddings are sensitive to differences in protein structures.

The trend lines in the running mean plots remain relatively flat and parallel to the x-axis for all protein pairs, supporting the stationarity of the process. This confirms that the embedding magnitudes do not display long-term trends or drifts, ensuring that subsequent statistical analyses like mean and variance calculations are meaningful.

Table 1
Statistical descriptors of embedding profiles for analyzed proteins.

Protein	Mean	Median	Variance	STD
$1R0R_1$	-0.064	0.119	603.573	24.567
$1MEE_1$	-0.098	0.082	595.101	24.395
$1LQT_1$	-0.404	-0.283	528.686	22.993
$1A6C_1$	0.094	0.182	502.625	22.419

Table 1 encapsulates key statistical descriptors derived from the embeddings. The mean value varies near zero for both similar and dissimilar proteins, the median also indicate a central tendency that aligns with the means. Variance and standard deviation, as measures of data spread, reinforce these findings, with both similar and dissimilar proteins exhibiting a comparable dispersion.

5.4. Probability Distribution

As the last step of an embeddings analysis, we attempt to evaluate the probability distributions of embedding features. We focus on two aspects: the distribution of dimension cuts and the magnitudes of embeddings.

Our findings (Figure 6) reveal that the distribution of both embedding dimension cuts at a fixed index and mean value across all the dimensions, adheres to a normal distribution. For biological data, where

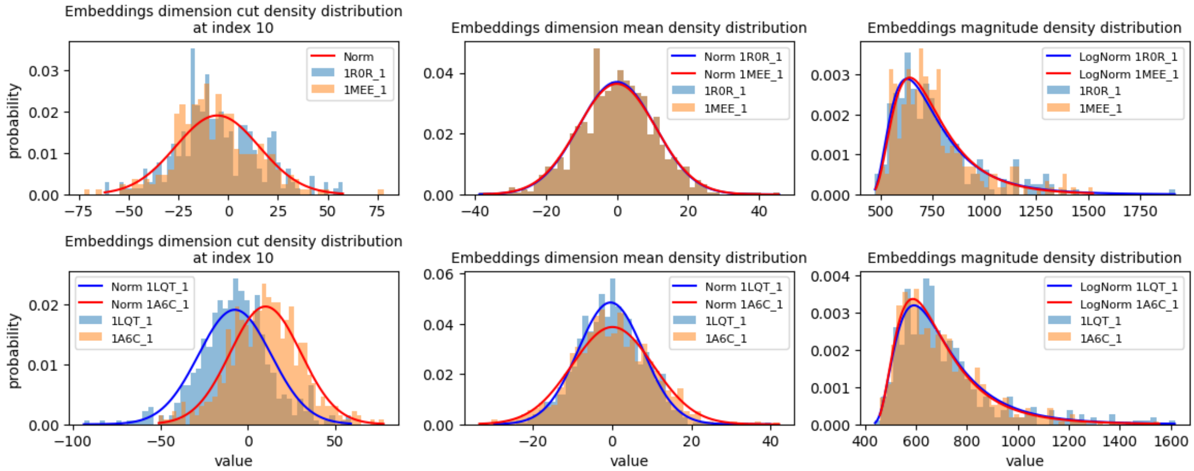


Figure 6: Density distributions for key aspects of protein sequence embeddings': dimension cuts at a specific index, mean values across all dimensions, and magnitudes.

a multitude of factors contribute to the final observation, such distribution is indicative of a robust underlying model that produces a stable, predictable pattern.

In contrast, the magnitudes of the embeddings follow a lognormal distribution, characteristic of processes governed by multiplicative factors. The lognormal nature of the magnitudes could reflect the exponential growth processes that underlie protein folding and development, where factors multiply, leading to the right-skewed distribution observed in our results.

6. Methodology

6.1. Autocorrelation Function

We define the autocorrelation function for a vector $V \in R^N$, where N is the length of the vector, using a sliding window of size w, and denote it as $ACF_e(V, w)$. By applying the autocorrelation function to each window w sliding over the input vector V, we obtain a matrix of autocorrelation functions $A \in R^{N-w-1 \times w}$. We then normalize the matrix A using the normalization function $N(A)$.

$$ACF_e(V, w) = R_{XX}(V_{i:j+w}) \tag{1}$$

where R_{XX} is the autocorrelation function, $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, N-w-1\}$.

6.2. Normalization Function

We define the normalization function $N(X)$ for the autocorrelation function that takes as input a matrix X and normalizes it by the maximum value of the corresponding row, resulting in a matrix X', where each row is divided by its maximum value:

$$N(X) = \frac{X}{max(X_i)} \tag{2}$$

where $i \in \{1, \ldots, N\}$.

6.3. Self-Similarity of Embeddings

Given the matrix $S^s \in R^{N \times 1024}$ that represents the embeddings of the protein sequence, we transpose this matrix $(S^s)^T \in R^{1024 \in N}$, to compute autocorrelation. Thus, computing self-similarity between corresponding dimensions of embeddings across the entire sequence. We define the self-similarity function for the sequence $ACF_p(P, w)$:

$$ACF_p(P, w) = ACF_e((S_i^s)^T, w) = A^{1024 \times N - w - 1 \times w} \tag{3}$$

where $i \in \{1, \ldots, 1024\}$.

6.4. Similarity of Two Sequences

For the first sequence, we compute $ACF_p(S_1^s, w) = A_1^{1024 \times N - w - 1 \times w}$, and for the second sequence, accordingly $ACF_p(S_2^s, w) = A_2^{1024 \times M - w - 1 \times w}$. We then calculate the Pearson correlation coefficient between each fragment of ACF of length w from the first sequence relative to all fragments of ACF from the second sequence in the given dimension. Let v_n be the set of ACF fragments from A_1 of length w where $n = \{1, \ldots, N - w - 1\}$, and v_m be the set of fragments of length w from A_2, where $m = \{1, \ldots, M - w - 1\}$. For each fragment of v_n, we compute the Pearson correlation coefficient with each fragment of v_m. The result is a correlation matrix $Corr \in R^{N - w - 1 \times M - w - 1}$, where each element $Corr_n m$ represents the correlation coefficient between fragments of v_n and fragments of v_m.

$$Corr = \frac{cov(v_n, v_m)}{\sigma(v_n) \times \sigma(v_m)} \tag{4}$$

By applying this function to all dimensions, we calculate a correlation matrix of two sequences $Corr \in R^{1024 \times N - w - 1 \times M - w - 1}$. The resulting matrix will contain information about the mutual similarity between the ACF of the sequences, describing the local similarity of the two protein sequences.

6.5. Algorithm

The algorithm described below provides a methodology for computing the correlation between two sets of protein sequence embeddings. It is designed to be invariant to outliers described in the previous section, since it operates on dimensions of embeddings.

```
function Autocorrelate(vector V, integer window_size)
    length = size of V
    Initialize array results with size (length - window_size + 1)
    for i from 0 to (length - window_size) do
        segment = slice of V from i to i + window_size
        autocorrelation = correlate segment with itself
        autocorrelation = normalize(autocorrelation)
        results[i] = autocorrelation
    end for
    return results
end function
```

Listing 1: Autocorrelation Computation

The $Autocorrelate$ function computes the autocorrelation of a given vector, segment by segment, within a defined window size. Normalization can be applied to each autocorrelation result to further ensure that the analysis is not skewed by extreme values.

```
function EmbeddingsCorrelation(matrix S1, matrix S2, integer window_size)
    smaller, larger = order matrices S1 and S2 by size
    autocorr_smaller = Autocorrelate(smaller, window_size)
    autocorr_larger = Autocorrelate(larger, window_size)
    Initialize correlation matrix
    for i from 0 to size of autocorr_smaller do
        for j from 0 to size of autocorr_larger do
            correlation[i, j] = Pearson correlation of autocorr_smaller[i] and autocorr_larger[j]
        end for
    end for
    return correlation
```

```
end function
```

Listing 2: Embedding Correlation Computation

$EmbeddingsCorrelation$ uses the autocorrelated data to compute the Pearson correlation coefficients across all pairs of autocorrelated segments between the two input matrices. The process accounts for the relative nature of the data, which is why the presence of outliers in specific indices does not distort the analysis.

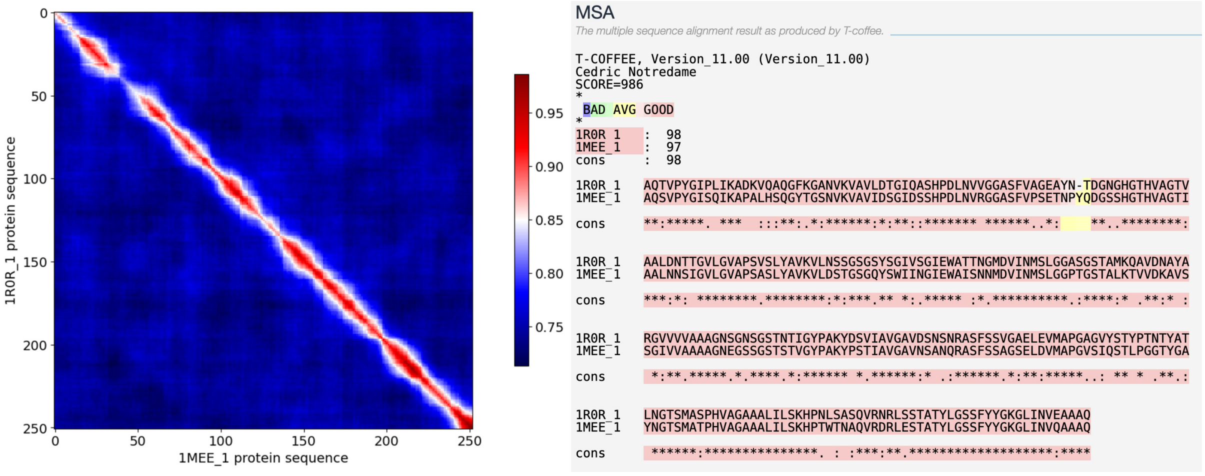
```
function ProteinCorrelation(matrix A, matrix B, integer window_size)
    A = transpose A
    B = transpose B
    Initialize corr matrix
    for each dimension d in A and B do
        corr[d] = EmbeddingsCorrelation(A[d], B[d], window_size)
    end for
    return corr
end function
```

Listing 3: Protein Correlation Analysis

Finally, $ProteinCorrelation$ iterates over each embedding dimension, applying $EmbeddingsCorrelation$ to build a correlation matrix for the entire set of embeddings. This matrix contains a detailed view of the similarities between the two protein sequences across all embedding dimensions, reflecting both local and global patterns in the data.

7. Experimental Results

We analyze several pairs of protein sequences by evaluating the correlation between their embeddings using the established methodology. We experimentally verify the ability of the ESM-2 model to learn the dependencies and evolutionary context of sequences and encode this informantion in seuquence embeddings. The resulting correlation matrix was visualized as a heatmap and compared with an MSA alignment. We present three cases of protein sequence comparison:

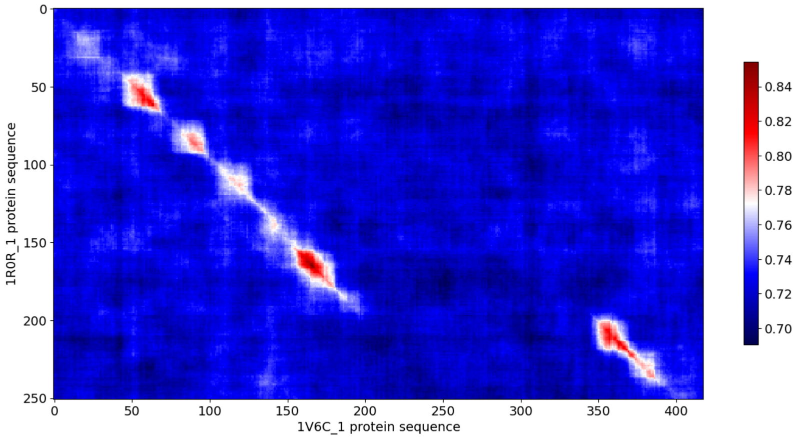


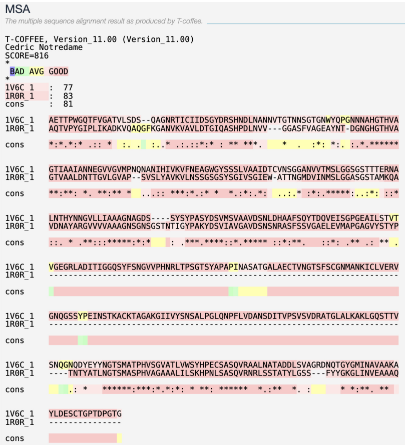
```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=986
*
 BAD AVG GOOD
*
1R0R_1    :  98
1MEE_1    :  97
cons      :  98

1R0R_1    AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDLNVVGGASFVAGEAYN-TDGNGHGTHVAGTV
1MEE_1    AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETNPYQDGSSHGTHVAGTI

cons      **:*****. ***  :::**:.*:******:*:**::******* ******..*:    **..********:


1R0R_1    AALDNTTGVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIEWATTNGMDVINMSLGGASGSTAMKQAVDNAYA
1MEE_1    AALNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWIINGIEWAISNNMDVINMSLGGPTGSTALKTVVDKAVS

cons      ***:*: ********.********:*:***.** *:.***** :*.**********.:****:* .**:* :


1R0R_1    RGVVVVAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVDSNSNRASFSSVGAELEVMAPGAGVYSTYPTNTYAT
1MEE_1    SGIVVAAAAGNEGSSGSTSTVGYPAKYPSTIAVGAVNSANQRASFSSAGSELDVMAPGVSIQSTLPGGTYGA

cons       *:**.*****.*.****.*:****** *.******:*  .:******.*:**:*****..: ** *  .**.:


1R0R_1    LNGTSMASPHVAGAAALILSKHPNLSASQVRNRLSSTATYLGSSFYYGKGLINVEAAAQ
1MEE_1    YNGTSMATPHVAGAAALILSKHPTWTNAQVRDRLESTATYLGSSFYYGKGLINVQAAAQ

cons       ******:**************. : :***:**.******************:****
```

Figure 7: Comparison of the correlation heatmap of the 1R0R_1 and 1MEE_1 protein sequence embeddings with the MSA alignment of these sequences

The left section of Figure 7 presents a heatmap generated by applying formula 4, which computes the correlation between two protein sequences. The heatmap's axes correspond to the sequences of two proteins $1R0R_1$ and $1MEE_1$, where diagonally aligned signal indicates similarity or identity, suggesting functional and structural parallels between the proteins, which corresponds to these proteins' MSA alignment.

The right section of Figure 1 displays the sequence alignment for $1R0R_1$ and $1MEE_1$. Each block's alignment includes a conservation score, where an asterisk '*' denotes identical amino acids at that position, suggesting a perfect match. A colon ':' marks positions with chemically similar, yet different, amino acids—indicative of conservative substitutions. A space ' ' represents positions where the amino acids significantly differ, termed non-conservative substitutions. A period '.' denotes semi-conservative substitutions where the amino acids are moderately similar. The color gradient from green to red across the panel reflects the alignment's varying quality, from low to high.



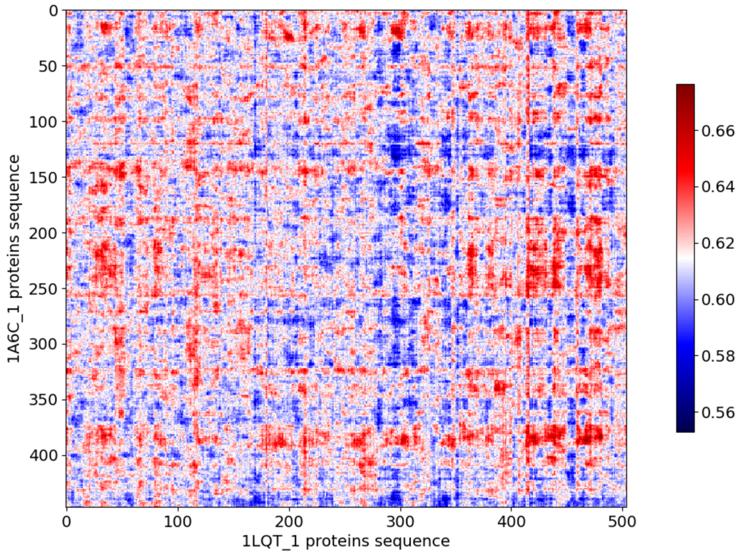

MSA

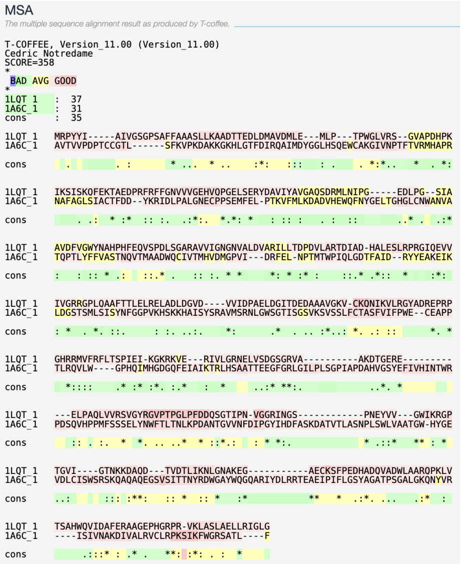
The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=816
*
 BAD AVG GOOD
*
1V6C_1     :  77
1R0R_1     :  83
cons       :  81

1V6C_1     AETTPWGQTFVGATVLSDS--QAGNRTICIIDSGYDRSHNDLNANNVTGTNNSGTGNWYQPGNNNAHGTHVA
1R0R_1     AQTVPYGIPLIKADKVQAQGFKGANVKAVLDTGIQASHPDLNVV---GGASFVAGEAYNT-DGNGHGTHVA

cons       *:*.*:* .:: *  :. .  :..* .:.::*:* : ** ***.    *  .  :*: *:.  :.*.******

1V6C_1     GTIAAIANNEGVVGVMPNQNANIHIVKVFNEAGWGYSSSLVAAIDTCVNSGGANVVTMSLGGSGSTTTERNA
1R0R_1     GTVAALDNTTGVLGVAP--SVSLYAVKVLNSSGSGSYSGIVSGIEW-ATTNGMDVINMSLGGASGSTAMKQA

cons       **:**: *. **:** *  ...:: ***:*.:* *  *.:*:.*:  ..:.* :*:.*****:..:*:  ::*

1V6C_1     LNTHYNNGVLLIAAAGNAGDS----SYSYPASYDSVMSVAAVDSNLDHAAFSQYTDQVEISGPGEAILSTVT
1R0R_1     VDNAYARGVVVVAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVDSNSNRASFSSVGAELEVMAPGAGVYSTYP

cons       ::. * .**:::*****:*:*    : .***.****::*.***** ::*:**.   ::*: .** .: ** .

1V6C_1     VGEGRLADITIGGQSYFSNGVVPHNRLTPSGTSYAPAPINASATGALAECTVNGTSFSCGNMANKICLVERV
1R0R_1     ------------------------------------------------------------------------

cons

1V6C_1     GNQGSSYPEINSTKACKTAGAKGIIVYSNSALPGLQNPFLVDANSDITVPSVSVDRATGLALKAKLGQSTTV
1R0R_1     ------------------------------------------------------------------------

cons

1V6C_1     SNQGNQDYEYYNGTSMATPHVSGVATLVWSYHPECSASQVRAALNATADDLSVAGRDNQTGYGMINAVAAKA
1R0R_1     ----TNTYATLNGTSMASPHVAGAAALILSKHPNLSASQVRNRLSSTATYLGSS---FYYGKGLINVEAAAQ

cons         .: *   ******:***:*.*:*: * **: ******  *.:**  *. :     * *:**. **

1V6C_1     YLDESCTGPTDPGTG
1R0R_1     ---------------

cons
```

Figure 8: Comparison of the correlation heatmap of the 1R0R_1 and 1V6C_1 protein sequence embeddings with the MSA alignment of these sequences

In Figure 8, we observe a correlation heatmap for a pair of protein sequences with more complicated alignment patterns. The heatmap's primary diagonal shows areas where the sequences align, indicating similarity. Notably, in the center, the alignment shifts and later realigns, suggesting a gap followed by a return to similarity. This observation is reflected in the MSA on the right, where asterisks and colons mark similar regions, and dashes '-' indicate sequence gaps, mirroring the heatmap's diagonal shifts.




MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=358
*
 BAD AVG GOOD
*
1LQT_1     :  37
1A6C_1     :  31
cons       :  35

1LQT_1     MRPYYI-----AIVGSGPSAFFAAASLLKAADTTEDLDMAVDMLE---MLP---TPWGLVRS--GVAPDHPK
1A6C_1     AVTVVPDPTCCGTL------SFKVPKDAKKGHKHLGTFDIRQAIMDYGGLHSQEWCAKGIVNPTFTVRMHAPR

cons         .          . :        * ...  * ..     :*:    :::    : .      . *:*..   *  . *:

1LQT_1     IKSISKQFEKTAEDPRFRFFGNVVVGEHVQPGELSERYDAVIYAVGAQSDRMLNIPG-----EDLPG--SIA
1A6C_1     NAFAGLSIACTFDD--YKRIDLPALGNECPPSEMFEL-PTKVFMLKDADVHEWQFNYGELTGHGLCNWANVA

cons           . .:  * :*  :: :.  .:*:.  *.*: *   : :: :   . :  ::        ..* .  .:*

1LQT_1     AVDFVGWYNAHPHFEQVSPDLSGARAVVIGNGNVALDVARILLTDPDVLARTDIAD-HALESLRPRGIQEVV
1A6C_1     TQPTLYFFVASTNQVTMAADWQCIVTMHVDMGPVI---DRFEL-NPTMTWPIQLGDTFAID--RYYEAKEIK

cons       :   : :: * .:   ::.* .   :: :. * *     *: * :* :     ::.* .*::  *     :*:

1LQT_1     IVGRRGPLQAAFTTLELRELADLDGVD----VVIDPAELDGITDEDAAAVGKV-CKQNIKVLRGYADREPRP
1A6C_1     LDGSTSMLSISYNFGGPVKHSKKHAISYSRAVMSRNLGWSGTISGSVKSVSSLFCTASFVIFPWE--CEAPP

cons       : *  . *. ::.     : :. ..:.    *:      .*  . .. :*..: *. .: ::       *. *

1LQT_1     GHRRMVFRFLTSPIEI-KGKRKVE---RIVLGRNELVSDGSGRVA----------AKDTGERE---------
1A6C_1     TLRQVLW----GPHQIMHGDGQFEIAIKTRLHSAATTEEGFGRLGILPLSGPIAPDAHVGSYEFIVHINTWR

cons        *::::    .* :* :*. :.*   :  *     ..:* **:.             ..*. *

1LQT_1     ---ELPAQLVVRSVGYRGVPTPGLPFDDQSGTIPN-VGGRINGS-----------PNEYVV---GWIKRGP
1A6C_1     PDSQVHPPMFSSSELYNWFTLTNLKPDANTGVVNFDIPGYIHDFASKDATVTLASNPLSWLVAATGW-HYGE

cons         :: . :.   *  *. .. ..*  * ::*.:   : * *:.           * .::*   ** : *

1LQT_1     TGVI----GTNKKDAQD---TVDTLIKNLGNAKEG-----------AECKSFPEDHADQVADWLAARQPKLV
1A6C_1     VDLCISWSRSKQAQAQEGSVSITTNYRDWGAYWQGQARIYDLRRTEAEIPIFLGSYAGATPSGALGKQNYVR

cons       ..:       :::  :**:   :: *   :: *   :*           **   *   .:*. ...   .:*   :

1LQT_1     TSAHWQVIDAFERAAGEPHGRPR-VKLASLAELLRIGLG
1A6C_1     ----ISIVNAKDIVALRVCLRPKSIKFWGRSATL----F

cons            .:::* : .* .   **: :*: . :   *
```

Figure 9: Comparison of the correlation heatmap of the $1A6C_1$ and $1LQT_1$ protein sequence embeddings with the MSA alignment of these sequences

Figure 9 showcases a scenario where protein sequences $1A6C_1$ and $1LQT_1$ exhibit minimal similarity. The heatmap lacks distinct patterns, aligning with the infrequent and scattered matches in the MSA, suggesting that the sequences share only isolated regions of structural or functional commonality.

8. Conclusions

Preliminary statistical analysis of the embedding arrays from selected protein amino acid sequences has been conducted. It was found that there are individual characteristic outliers in the numerical values of certain embeddings projections, namely the 247 and 37 dimensions always represent extreme maximum and minimum outlier values respectively. Normalizing these by replacing them with the average value of two adjacent readings allows for further processing and analysis of the data.

In the statistical analysis of protein sequence embeddings, histograms were employed to examine the distribution across various dimensions for both similar and dissimilar protein pairs. The results revealed that similar proteins exhibited overlapping distribution patterns in specific dimensions, suggesting shared structural or functional features through spatial proximity, while dissimilar proteins showed shifted distribution indicating varying structural characteristics. Further, the magnitude of these embeddings was analyzed and confirmed the stationarity of the process using running mean method, which allowed to compute statistical measures such as mean, median, and variance. Both similar and dissimilar protein pairs displayed analogous statistical and magnitude characteristics. Additionally, the probability distribution analysis showed that embedding dimensions generally follow a normal distribution, whereas embedding magnitudes adhered to a log-normal distribution, that may reflect the multiplicative biological processes inherent in protein folding. These findings enhance our understanding of protein folding process and support the initiative to use correlation and autocorrelation analysis to develop a Embedding-based Alignment (EBA) method.

The application of covariance and autocorrelation analysis to ESM-2 sequence embeddings showed that the model can learn the evolutionary character of sequence development and sequence interrelationships. By evaluating the correlation between embeddings of different protein pairs, we observed clear patterns, and experimentally verified the results with MSA alignment of these sequences, which further confirmed the proposed analysis method.

References

[1] C. Wang, H. Fan, R. Quan, Y. Yang, Protchatgpt: Towards understanding proteins with large language models, arXiv preprint arXiv:2402.09649 (2024). doi:10.48550/arXiv.2402.09649.

[2] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130. doi:10.1126/science.ade2574.

[3] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (casp)—round xv, Proteins: Structure, Function, and Bioinformatics 91 (2023) 1539–1549. doi:10.1002/prot.26617.

[4] L. Pantolini, G. Studer, J. Pereira, J. Durairaj, G. Tauriello, T. Schwede, Embedding-based alignment: combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone, Bioinformatics 40 (2024) btad786. doi:10.1093/bioinformatics/btad786.

[5] B. Rost, Twilight zone of protein sequence alignments, Protein Engineering, Design and Selection 12 (1999) 85–94. doi:10.1093/protein/12.2.85.

[6] K. Eitner, U. Koch, T. Gawęda, J. Marciniak, Statistical distribution of amino acid sequences: a proof of Darwinian evolution, Bioinformatics 26 (2010) 2933–2935. doi:10.1093/bioinformatics/btq571.

[7] P. Di Tommaso, S. Moretti, I. Xenarios, M. Orobitg, A. Montanyola, J.-M. Chang, J.-F. Taly, C. Notredame, T-coffee: a web server for the multiple sequence alignment of protein and rna sequences using structural information and homology extension, Nucleic acids research 39 (2011) W13–W17. doi:10.1093/nar/gkr245.