Title

Author

Date

1 Introduction

2 Related Work

3 Methods

In this section we describe the pipeline used to analyze the embeddings. As shown in Table 1, the input length is different between the models as well as the output produced. We want to address the following problems: 1) compare different methods to join togheter the amminoacid-specific contextual representations in order to have a representation for the whole chunk; 2) compare different methods to join togheter the representations of the chunks in order to have a representation for the whole protein; 3) find out if these representations reflect known properties of the proteins.

3.1 Combining the contextual representations

We tried four methods to join togheter the amminoacid embeddings in order to produce a fixed size embedding for the chunk: average, maximum, sum and principal component analysis (PCA). Note that even if these operators are commutative, the overall process do takes into account the order of the amminoacids precisely because the embeddings are contextual.

The same operator used to combine the amminoacid embeddings is also used to combine the embeddings of the chunks of the sequence.

3.2 Comparison with known informations

Given a set of embeddings of sequences we want to analyze their distribution in the embedding space comparing it with both the distance matrix produced during the multiple sequence alignment with Clustal Omega [7] and protein annotations as gene ontology, UniProt keywords and taxonimy [2].

3.2.1 Allignment distance

In order to compare the distances between the sequences in the embedding space with the alignment distance we performed an agglomerative clustering on both the matrices,

the resulting tree is then cut at each levels: flat partitions of all possibles number of clusters are produced. We perform a comparison of the partitions with the same number of clusters using the adjusted rand score [3]. The mean of these score, starting from two clusters up to $\#elements - 1$ clusters is called mean adjusted rand score (MARS).

Name	input length (chunk)	embedding dimension
embedding reproduction (rep)[8]	64	64 per chunk
dnabert [4]	512	768 per chunk
prose [1]	512	100 per amino acid
alphafold [5]	1024	384 per ammino acid
evolutionary scale modeling (esm2) [6]	1024	1280 per ammino acid

Table 1: Embedders used in the experiments, their maximum imput length and the dimension of the embedding produced.

4 Results

References

[1] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

[2] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022.

[3] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[4] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[5] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[6] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[7] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

[8] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 2018.