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 2, 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 [9] and annotations as Gene Ontology [3, 1], UniProtKB Keywords and NCBI Taxonimy [4].

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 level: flat partitions of all possibles number of clusters are produced. We performed a comparison of the partitions with the same number of clusters using the adjusted rand score [5]. The mean of these score, starting from two clusters up to $\#elements - 1$ clusters is called mean adjusted rand score (MARS).

3.2.2 Enrichment analysis

We wanted to analyze the properties of the embeddings also at an higher level. The Gene Ontology (GO) describes our knowledge of the sequence with respect to: molecular function, cellular component and biological process; there ara also more specific controlled vocabulary as the UniProt Keywords and hierarchical classifications specific for sequences as the NCBI Taxonomy.

Whatever they are the sets of words to describe the sequences in our datasets, we want to build a distance between sequences among them. Given A and B the sets of annotations of two sequences we computed the distance in two possible ways:

$$d1 = \frac{2 * |A \cap B|}{|A| + |B|}$$

$$d2 = \max\{\frac{|A \cap B|}{|A|}, \frac{|A \cap B|}{|B|}\}$$

Both of them vary between 0 and 1, however $d1$ goes to 1 only when the two sets are equals while $d2$ goes to 1 also when one set is a subset of the other. After calculating one of these distance between all possible pair in the dataset we end up with a similarity matrix, that can be easly transformed in a distance matrix that is possible to comprare with the distance matrix derived from the distance between the embeddings using the MARS as described in subsection 3.2.1.

Name	description	number of sequences	type	avg length
hemoglobin	hemoglobin for various organisms	761	amminoacid	142
mouse	mouse proteome	974	amminoacid	516
bacterium	bacterium proteome	259	amminoacid	427
covid19	covid19 complete genome	77	nucleotides	29831
meningitis	meningitis complete genome	68	nucleotides	2240049

Table 1: Datasets used in the experiments.

Name	input length (chunk)	embedding dimension
embedding reproduction (rep)[10]	64	64 per chunk
dnabert [6]	512	768 per chunk
prose [2]	512	100 per amino acid
alphafold [7]	1024	384 per ammino acid
evolutionary scale modeling (esm2) [8]	1024	1280 per ammino acid

Table 2: Embedders used in the experiments, their maximum imput length and the dimension of the embedding produced.

4 Results

References

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

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

[3] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru,

Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 03 2023.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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