Title

Author

Date

1 Introduction

In the latest few years, machine learning and deep learning have revolutionized the natural language processing field (NLP)[11]. At the same time some fundamental ideas developed in NLP have been successfully applied to another type of language, the biological one: DNA, RNA and amminoacid sequences bringing to excellent results even in the complex task of protein structure prediction[10, 13]. One of these fundamental ideas is word embedding [15] because it transforms words into points in space, therefore easy to process. It is well known that sequence similarity do not always correspond to functional similarity [12]. Application of protein language model embeddings to downstream tasks was first demonstrated by Bepler and Berger (2019) [2]

2 Methods

In this section we describe the pipeline used to analyze the embeddings, i.e. vectors in a high-dimensional real space that represent a chunk of a genetic sequence (of a hole one if it is short enough). As shown in Table 1, the input length accepted is different between different models as well as the dimensionality of the output produced. The maximum length of the input sequence is usually determined by the length of the examples the model was training on or by limitations due to the structure of the model such as the size of the attention module in the models that use it. These numerical representation of sequences are then used to various downstream applications.

Because of their structure, these model output a vector for each single item of the input sequence (amminoacid or nucleotides) so we end up with a matrix representation of a sequence with different number of rows for each different sequence. Futhermore, if the sequence that we want to analyze is longer then the model's limit we have to split it in subsequences and thus the representation became a tensor.

It is not clear which is the best way to aggregate this tensor in a fixed size representation.

We want to address the following problems: 1) compare different methods to join together the amminoacid-specific contextual represen-

tations in order to have a representation for the whole chunk and subsequently join together the representations of the chunks in order to have a representation for the whole protein; 2) find out if these representations reflect known properties of the proteins.

It is important to note that the vectorial representation of an item of the sequence (nucleotides or amminoacid) is not fixed, it is instead contextual, i.e. it depends, theoretically, on each other item in the sequence. This is intuitively good in natural language because a word can assume different meaning in different context [17], and this is also good for nucleotides or amminoacids in biological sequences.

2.1 Represent protein sequences as continuous vectors

2.1.1 DNABert

2.1.2 Prose [3]

The model structure is a multi-layer bidirectional Long Short Term Memory (bi-LSTM). Following the intuition that some aspect of proteins structure and semantic can never be discoverable by statistical sequence models alone, they came up with the idea of multi-task learning with structural supervision. The learning tasks are: 1) classical masked language modeling task, 2) residue-residue contact prediction, 3) structural similarity prediction. This model outperformed existing approaches in both transmembrane position labeling and phenotypes prediction of sequence variants.

2.1.3 Alphafold [10]

This is the only approach that do not relay on a language model but is a combination of a bioinformatics and physical approaches. It rely on large datasets of protein sequences that are similar enough to be aligned with high confidence but contain enough divergence to confidently infer statistical couplings between positions. It consists in two modules, Evoformer module and structure module. The Evoformer builds separate MSA and residue-pairwise embedding spaces. As it is structured, this system is not able to learn patterns across large-scale databases of possibly unrelated proteins [3]. In the experiments we used the "single" representation of the sequence, without the MSA and the templates.

2.1.4 SeqVec [7]

To build their model Heinzinger et al.[7] adapted the standard ELMo configuration [16] to work with protein sequences modifying the number of tokens and the unroll steps. It is composed by 1 CharCNN and 2 LSTM-Layers. Given a protein sequence of arbitrary length it returns 3072 features for each residue derived by concatenating the outputs of the three layers of ELMo, each describing a token with a vector of length 1024. In order to obtain a smaller representation for each amminoacid we computed the mean of the three layers (as also suggested in the official repository). Given the architecture of the ELMo, these representation are contextual-

dependent.

2.1.5 Evolutionary scale modeling [14]

The underlying architecture is a BERT [6] style encoder transformer with modifications in the number of layers, number of attention heads, hidden size and feed forward hidden size. A further improvement is the use of Rotary Position Embedding [20] that allows to generalize beyond the context window it is trained on. The main advantages compared to AlphaFold are the removal of the the need of multiple sequence alignment and an increasing up to two order of magnitude of the speed of the prediction pipeline. We used the version with 33 layers and 650B parameters (compare with alphafold)

2.2 Combining the (contextual) representations

We tried four methods to join together the amminoacid embeddings in order to produce a fixed size embedding for the chunk: average, maximum, sum and principal component analysis (PCA).

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

2.3 Comparison with known information

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 [18] and higher level annotations as Gene Ontology [4, 1], UniProtKB Keywords and NCBI Taxonomy [5].

The alignment distance matrix provide an evolutionary related distance between sequences [19], we also wanted to analyze the properties of the embeddings 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 easily transformed in a distance matrix that is possible to compare with the distance matrix derived from the distance between the embeddings.

2.3.1 Similarity between distance matrices

In order to compare two distance matrices we performed an agglomerative clustering on both, the resulting tree is then cut at each level obtaining flat partitions of all possibles number of clusters. We performed a pairwise comparison of the partitions having the same number of clusters using the adjusted rand score [8]. The mean of these scores, starting from two clusters up to $\#elements - 1$ clusters is called mean adjusted rand index (MARI). We compared different distance metrics as well as different methods to perform the hierarchical clustering.

3 Results

3.1 Phylogenetic

3.2 Enrichment

3.3 Projections?

3.4 Classification?

3.5 Pointwise representation similarity?

Name	input length (chunk)	embedding dimension
embedding reproduction (rep)[21]	64	64 per chunk
seqvec [7]	1024	1024 per amminoacid
dnabert [9]	512	768 per chunk
prose [3]	512	100 per amino acid
alphafold [10]	1024	384 per ammino acid
evolutionary scale modeling (esm2) [14]	1024	1280 per ammino acid

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

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

Table 2: Datasets used in the experiments.

combiner	dimensional PCA	seqvec	prose	alphafold	esm
pca	10	0.449140	0.351432	0.417911	0.495245
	20	0.494195	0.395091	0.468677	0.545581
	30	0.509105	0.389776	0.478603	0.584235
	40	0.533046	0.388992	0.488803	0.589202
	50	0.544053	0.391228	0.489178	0.590380
	all	0.564945	0.394987	0.519607	0.605519
average	10	0.440643	0.376631	0.427137	0.511643
	20	0.493682	0.407182	0.467401	0.554667
	30	0.519014	0.407446	0.481296	0.578104
	40	0.534462	0.408016	0.488328	0.588670
	50	0.544056	0.407953	0.486588	0.601595
	all	0.577440	0.406862	0.351767	0.610861
sum	10	0.437402	0.382771	0.437270	0.493417
	20	0.480627	0.411288	0.474695	0.536754
	30	0.499964	0.407558	0.480758	0.568252
	40	0.525452	0.407691	0.490450	0.572879
	50	0.524109	0.409393	0.497554	0.583100
	all	0.558963	0.405949	0.359855	0.605848
max	10	0.429831	0.415769	0.374368	0.514790
	20	0.464959	0.498250	0.397569	0.579752
	30	0.477753	0.506698	0.415304	0.587261
	40	0.487472	0.514747	0.428992	0.595889
	50	0.499499	0.522068	0.435171	0.611635
	all	0.542751	0.533704	0.378067	0.680641

Table 3: Emoglobin Phylogenetic results.

combiner	dimensional PCA	seqvec	prose	alphafold	esm
pca	10	0.190254	0.212593	0.123817	0.154119
	20	0.223326	0.215994	0.158653	0.211017
	30	0.233909	0.216776	0.169543	0.236152
	40	0.248802	0.216817	0.165458	0.259220
	50	0.255239	0.216956	0.170329	0.268735
	all	0.284410	0.212964	0.166672	0.284101
average	10	0.203596	0.234889	0.133038	0.194540
	20	0.252383	0.238390	0.160997	0.267515
	30	0.265740	0.238646	0.168218	0.294172
	40	0.282326	0.239766	0.170593	0.308133
	50	0.293596	0.239742	0.170439	0.318216
	all	0.337058	0.239850	0.126641	0.330147
sum	10	0.158431	0.215487	0.151187	0.141865
	20	0.168831	0.220564	0.154411	0.169305
	30	0.181248	0.221456	0.165493	0.186215
	40	0.182201	0.221607	0.162561	0.187705
	50	0.183775	0.222080	0.160401	0.190762
	all	0.195369	0.214632	0.112511	0.207289
max	10	0.213904	0.285951	0.088717	0.280855
	20	0.276499	0.307956	0.122844	0.342551
	30	0.294216	0.316039	0.135043	0.380750
	40	0.310357	0.316190	0.143899	0.397330
	50	0.319428	0.328722	0.154333	0.413625
	all	0.271817	0.334124	0.099899	0.438918

Table 4: Mouse Phylogenetic results.

combiner	dimensional PCA	seqvec	prose	alphafold	esm
pca	10	0.064957	0.072183	0.012768	0.034683
	20	0.085704	0.076620	0.005497	0.050977
	30	0.091166	0.076632	0.005371	0.052120
	40	0.090972	0.076629	0.008708	0.052657
	50	0.096276	0.076608	0.022336	0.058888
	all	0.092587	0.072544	0.007800	0.086005
average	10	0.063419	0.089908	0.017157	0.035883
	20	0.078590	0.089790	0.011811	0.053720
	30	0.087748	0.089808	0.007368	0.053789
	40	0.092050	0.089814	0.008073	0.054903
	50	0.094008	0.089814	0.006989	0.066082
	all	0.099638	0.092655	0.010589	0.084124
sum	10	0.017759	0.060247	-0.000334	0.005171
	20	0.020039	0.059339	0.004637	0.007319
	30	0.022856	0.059442	0.005062	0.007984
	40	0.024805	0.059451	0.005057	0.008785
	50	0.024293	0.059436	0.004787	0.008381
	all	0.025664	0.060248	0.000751	0.023070
max	10	0.035823	0.074083	0.000905	0.034337
	20	0.042759	0.088803	0.003758	0.064972
	30	0.039637	0.091899	0.000352	0.081378
	40	0.038814	0.093861	-0.000994	0.091238
	50	0.036684	0.098092	-0.000269	0.093263
	all	0.031241	0.097953	0.010153	0.112645

Table 5: Bacterium Phylogenetic 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 protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

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

[4] 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.

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

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17, 2019.

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

[9] 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.

[10] 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.

[11] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[12] Mickey Kosloff and Rachel Kolodny. Sequence-similar, structure-dissimilar protein pairs in the pdb. *Proteins: Structure, Function, and Bioinformatics*, 71(2):891–902, 2008.

[13] 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.

[14] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[16] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[17] Shimi Salant and Jonathan Berant. Contextualized word representations for reading comprehension. *arXiv preprint arXiv:1712.03609*, 2017.

[18] 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.

[19] Mohammad Yaseen Sofi, Afshana Shafi, and Khalid Z. Masoodi. Chapter 6 - multiple sequence alignment. In Mohammad Yaseen Sofi, Afshana Shafi, and Khalid Z. Masoodi, editors, *Bioinformatics for Everyone*, pages 47–53. Academic Press, 2022.

[20] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

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