EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

Master Thesis

Harnessing Protein Language Models to improve classic Local Pairwise Alignments for more sensitive and scalable Deep Homology Detection

Céline Behr

19th December 2022
Mentors: Dr. Hajk Georg Drost and Benjamin Buchfink
Supervisors: Jun.-Prof. Dr. Andreas Dräger and Prof. Dr. Detlef Weigel

1 Background and Motivation

Biological sequences are the foundation on which molecular biology research builds its comparative and analytical knowledge. As a result, high-throughput sequencing is becoming a *bona fide* technology that is employed across the life-science sector. This exponential growth of sequencing in the past decades dictates a new focus on organizing and classifying the immense amount of sequence data to capture the entire complexity of the biospheric sequence space. Classic global (e.g., Needleman-Wunsch) and local (Smith-Waterman) alignment approaches have been used to search and learn from this vast sequence space. However, newly emerging approaches incorporating machine learning are tested today with the ambition to build a fast classification procedure rather than a costly alignment process. Therefore, there are several models that capture the complexity of sequences and integrate the respective data into tractable probabilistic models [1, 2]. An important application of these efforts is exploring of the natural variation of the protein universe and the organization of species into a Tree of Life (ToL) based on protein sequence comparisons. In addition, recent milestones in structural biology (fold prediction based on AlphaFold2) are drawing new attention to biological sequences that are now proven predictive enough for structural explorations and functional annotation. This is relevant to various tasks, such as drug design and comparative genomics.

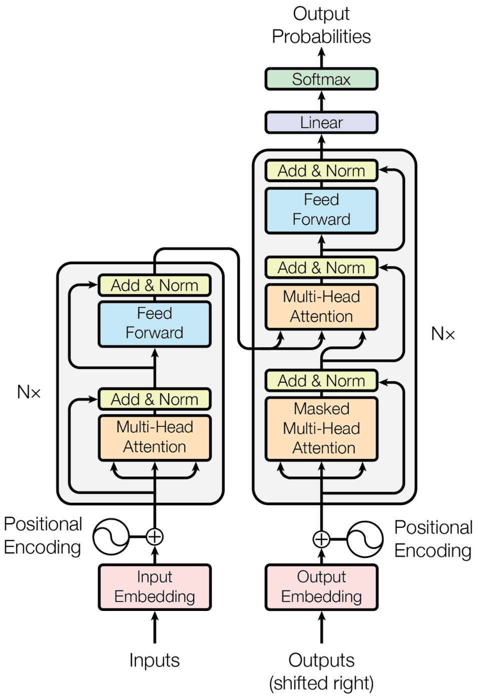

Figure 1 | Transformer model architecture of a Protein Language Model [1]

Since it is complex to determine protein structures experimentally, several new approaches are emerging that claim to be able to infer biological information from sequences alone. Information can be transferred by adapting new or unannotated sequences from annotated sequences to infer evolutionary, structural, and functional relations [2, 3]. This can be done by homology detection, where homology (common descent) is confirmed based on the similarity between protein sequences [3].

There are already many attempts in the literature in which machine learning-based approaches are applied to find homologs. These show good results and improvement and reveal the new potential in deep homology detection [4]. A specific approach is protein language models (pLM) derived from Natural Language Processing (NLP), which are currently widely used. These are deep learning models trained on the assessment of multiple sequences. [5]

They are based on a transformer architecture (Fig. 1) that captures structural and functional properties of amino acid sequences by learning context while tracing relationships in successive data. Therefore they apply an attention mechanism which is then interpreted in an amino acid alphabet context. [1]

The pLMs generalize input amino acid sequences using embeddings that capture the biophysical properties of amino acids [3]. Thus, e.g., nearest neighbor search and other approaches can then be applied to determine homology associations [3, 6].

There are already promising results in the literature, such as the finding that pLMs capture crucial constraints and information beyond sequence similarity and show comparatively good improvements in remote homology detection [3]. However, problems occur, especially with multidomain proteins, proteins with long amino acid chains, and disordered proteins where performance degrades [3, 6].

2 Aim and Approach

The goal of this master thesis is to implement a hybrid approach for the homology detection of protein sequences by combining classical sequence alignment approaches with approaches based on protein language transformers. This approach aims to take advantage of the strengths of both methods and combine them into a more accurate and robust model, as they may learn complementary information such as local alignment structure versus long-range amino acid interactions. DIAMOND [7] should be used as a classical sequence alignment method. The intention is to find an approach that addresses the abovementioned problems and determine novel solutions that improve accurate homology predictions.

3 Requirements

(a) Python programming, (b) fundamental understanding of bioinformatics and interest in developmental biology and evolutionary and comparative genomics, and (c) fundamental understanding of machine learning algorithms and techniques, especially language transformer.

References

[1] Ashish Vaswani et al. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762.

[2] Richard Durbin et al. “Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.” In: 1998.

[3] Konstantin Schütze et al. “Nearest neighbor search on embeddings rapidly identifies distant protein relations.” In: *bioRxiv* (2022). DOI: 10.1101/2022.09.04.506527.

[4] Maxwell L. Bileschi et al. “Using Deep Learning to Annotate the Protein Universe.” In: *bioRxiv* (2019). DOI: 10.1101/626507.

[5] Damiano Sgarbossa et al. “Generative power of a protein language model trained on multiple sequence alignments.” In: *bioRxiv* (2022). DOI: 10.1101/2022.04.14.488405.

[6] Kamil Kaminski et al. “pLM-BLAST – distant homology detection based on direct comparison of sequence representations from protein language models.” In: *bioRxiv* (2022). DOI: 10.1101/2022.11.24.517862.

[7] Benjamin Buchfink et al. “Fast and sensitive protein alignment using DIAMOND.” In: (2015). DOI: 10.1038/nmeth.3176.