SHOW ME WHAT YOU GOT

Making sense of protein language models

SMTB

ZIMIN FOUNDATION

ABSTRACT

WHY:
Any machine learning algorithm has its limitations and disadvantages. And if in the case of natural language processing or image recognition, these problems are visible to the naked eye, with amino acid sequences it can be not so obvious. Therefore, we decided to check if ESM, the most popular large protein language model at the moment.

HOW:
We examined remote homology using randomly generated sequences and compared ESM data with existing amino acid substitution tables.

INTRODUCTION

Natural Language Processing (NLP) is a dynamically developing field of machine learning. One of the main breakthroughs in this area is the invention of transformer models. From the idea to apply transformers to protein sequences, the ESM (evolutionary scale model) was born.

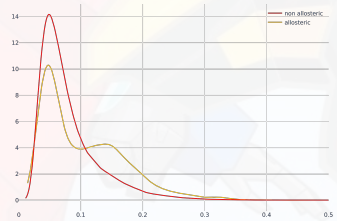

Fig. 1. Mutations in allosteric sites

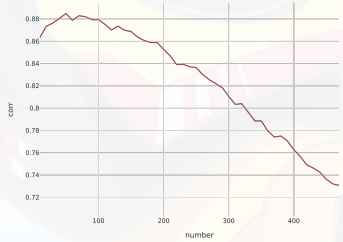

Fig. 2. Correlation between PAM and ESM

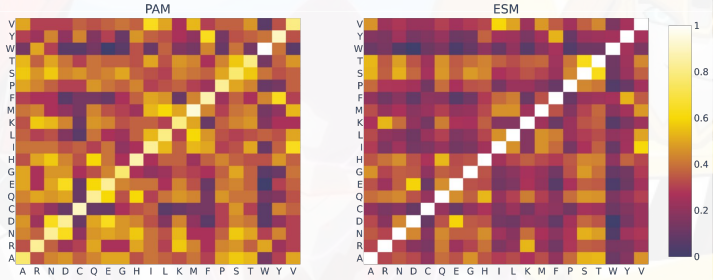

Fig. 3. PAM and ESM matrices

Fig. 4. PCA-mut-prot plot

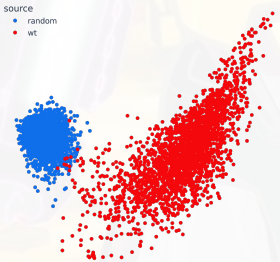

Fig. 5. t-SNE embedding

WHO are we:

Anna Toidze

Roman Joeres

Alper Yurtseven

Ilya Senatorov

Daria Guseva

Lidia Rebriy

Aleksandra Seravkina

Prof. Olga Kalinina

ATTENTION

Figures

GitHub