SHOW ME WHAT YOU GOT

Making sense of protein language models

SMTB

ZIMIN FOUNDATION

ABSTRACT

ЗАЧЕМ:
У любого алгоритма машинного обучения есть свои ограничения и недостатки. И если при работе с текстом или изображениями эти проблемы видны невооружённым глазом, то с аминокислотным последовательностями все не так очевидно. Поэтому мы решили проверить, так ли хороша ESM - самая популярная большая языковая белковая модель на данный момент.
КАК:
Мы исследовали remote homology, используя случайно сгенерированные последовательности, также сравнили данные полученные с помощью ESM с существующими таблицами частотности аминокислотных замен.

INTRODUCTION

Обработка естественного языка (NLP) - это динамично развивающаяся область машинного обучения. Один из главных прорывов в этой области - изобретение моделей-трансформеров. Из идеи применить трансформеры к белковым последовательностям и родилась ESM (evolutionary scale model).

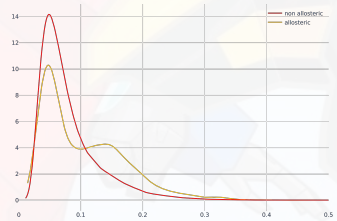

Fig. 1. Mutations in allosteric sites

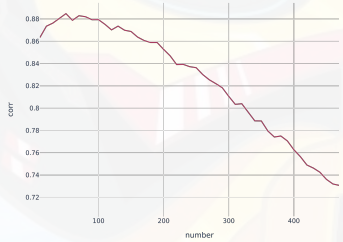

Fig. 2. Correlation between PAM and ESM

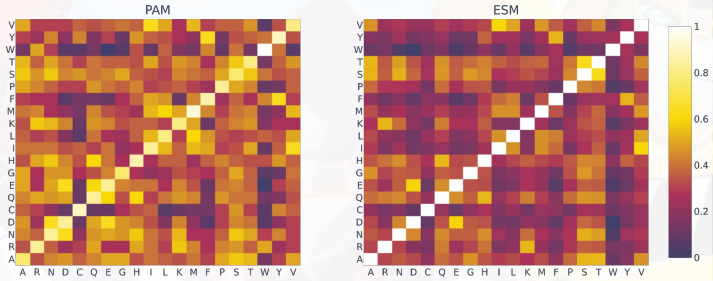

Fig. 3. PAM and ESM matrices

Fig. 4. PCA-mut-prot plot

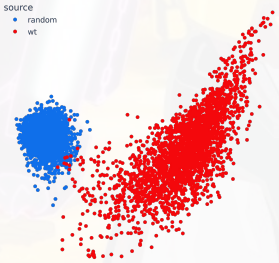

Fig. 5. t-SNE embedding

WHO ARE WE:

Anna Toidze

Roman Joeres

Alper Yurtseven

Ilya Senatorov

Daria Guseva

Lidia Rebriy

Aleksandra Seravkina

Prof. Olga Kalinina

ATTENTION

Figures

GitHub