Learning co-evolution information with language models.

Зверев Егор

Московский физико-технический институт

Курс: Автоматизация научных исследований (практика, В. В. Стрижов)/Группа 821

Эксперт: Сергей Грудинин

Консультант: Илья Игашов

2021

Исследование применимости языковых моделей к fold classification problem

Цель

Исследовать возможности применения языковых моделей для получения информации о ко-эволюции белков.

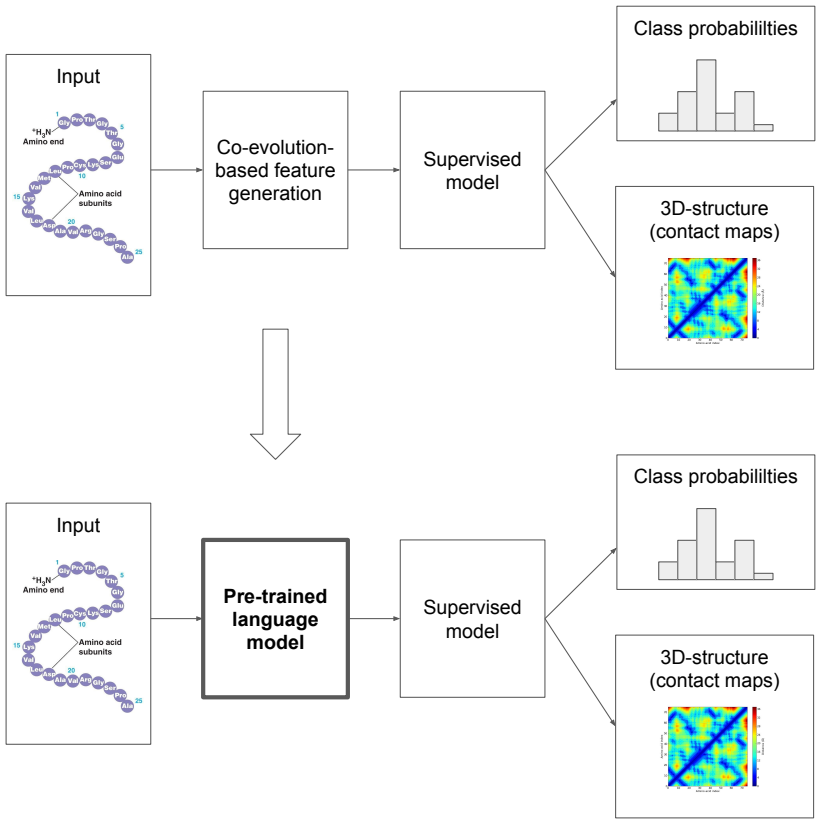
Задача

Решается задача классификаций белков (fold classification problem).

Предложение

Применить предобученный BERT к последовательностям аминокислот, вычленить attention heads, применить CNNs.

Суть подхода



Столбец 1

Идея: заменить MSA генерацию признаков на языковую модель.

Публикации по теме


S. de Oliveira and C. Deane.
Co-evolution techniques are reshaping the way we do structural bioinformatics.
F1000Res, 6:1224, 2017.


Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

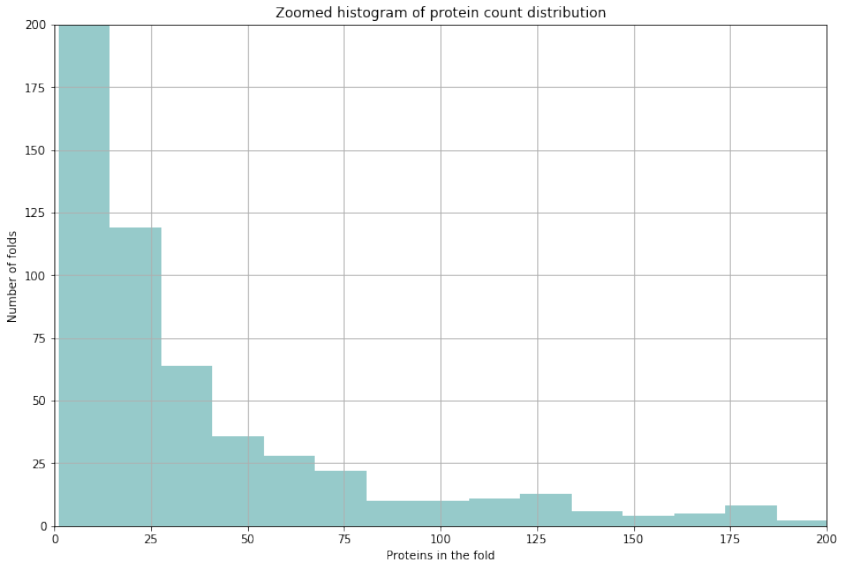

W. Elhefnawy, M. Li, J. Wang, and Y. Li.
DeepFrag-k: a fragment-based deep learning approach for protein fold recognition.
BMC Bioinformatics, 21(Suppl 6):203, Nov 2020.


Jie Hou, Badri Adhikari, and Jianlin Cheng.
DeepSF: deep convolutional neural network for mapping protein sequences to folds.
Bioinformatics, 34(8):1295–1303, 12 2017.

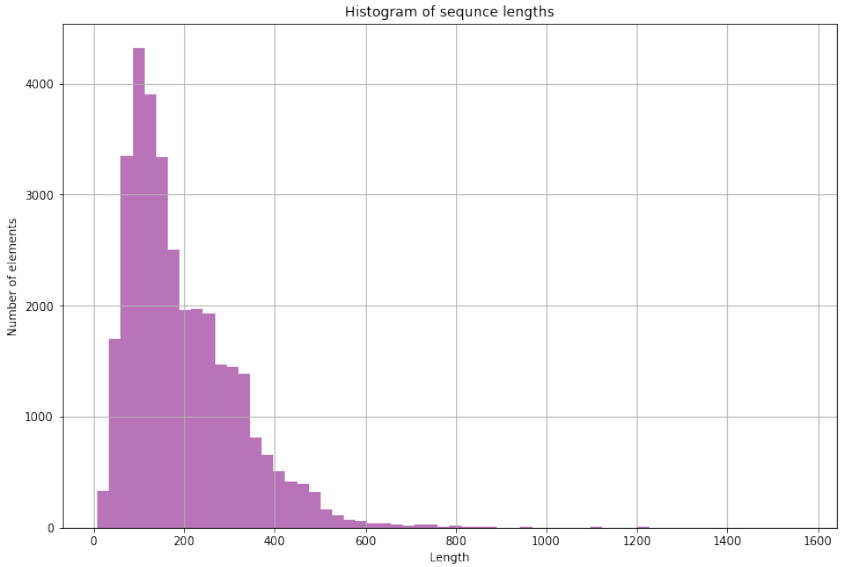
Данные - последовательности аминокислот

Используем SCOP2 dataset.

Более 30000 последовательностей, 1517 классов.



Количество протеинов по фолдам.



Распределение длин последовательностей

Задача классификации белков

Дано: $A = \{a_i\}_{i=1}^n$ - последовательности аминокислот.

$$A = A_{train} \cup A_{val}$$

$\mathbb{Y} = \{y_i\}_{i=1}^n$ - истинные классы.

Требуется: используя A_{train}, построить модель, предсказывающую вероятности принадлежности белка классам.

Внешняя метрика качества: Accuracy Score на A_{val}

Решение в два шага

Шаг1

Сгенерировать дескрипторы для белков. На выходе имеем $S = \{(x_i, Y_i)\}_{i=1}^{n}$, где $x_i \in \mathbb{R}^D$ - дескрипторы.

Шаг 2:

Обучить многослойную нейронную сеть решать стандартную задачу классификации на S.

Данная работа предлагает новый способ генерировать дескрипторы. На шаге 2 берётся модификация существующей архитектуры DeepSF.

Применение BERT

Интуиция

Последовательности аминокислот не случайны. В некотором смысле это язык. Имеет смысл применять BERT.

Новизна

Нет ресурсов обучать BERT, но можно взять предобученный. Предполагаем, что co-evolution скрыта в attention heads. Будем извлекать attention и подавать на вход нейронке.

Первый подход

Применить AlBERT ко всем данным. Обучить нейросеть, аналогичную DeepSF, на полувшихся данных.

Результаты первого эксперимента

```
model, history = train(
    model, criterion, optimizer, train_batch_gen, val_batch_gen,
    "first_exp", num_epochs=50)
```

```
Epoch 1 of 50 took 11465.916s
  training loss (in-iteration):         6.003195
  validation loss (in-iteration):       5.941781
  training accuracy:                    3.69 %
  validation accuracy:                  3.82 %
```

Первая эпоха DeepSF

Обучалась 3 часа, accuracy score = 4

Вывод: имеет смысл продолжать эксперимент, поменяв конфигурацию.

Проблемы первого эксперимента

В начале было 8 мегабайт данных. После применения AlBERT стало 108 гигабайт. В результате:

Сложные инженерные решения

Надо подгружать данные на сервер их облачного хранилища.

Невозможно долгое обучение

3 часа на одной эпохе. Нет ресурсов проверять, что будет на 50 эпохах.

Предлагаемое решение

Переформулировка задачи

Предлагается выбрать 2 класса из 1517 и решить на них задачу бинарной классификации. Успешное решение этой задачи позволит сделать выводы о применимости языков моделей, при этом вычисления будут легче на несколько порядков.