УДК [61+615.1] (043.2)
ББК 5+52.81
А 43
ISBN 978-985-21-1864-4

Наумов С.А., Хонов В.Р.

ПРЕДСКАЗАНИЕ ЛОКАЛИЗАЦИИ ФУНКЦИОНАЛЬНЫХ ДОМЕНОВ БЕЛКОВ С ПОМОЩЬЮ КОНТЕКСТНЫХ МОДЕЛЕЙ

Научные руководители: канд. хим. наук Страшнов П.В.
Кафедра прикладной математики и информатики
Московский Государственный Технический Университет им. Баумана. г. Москва

Актуальность. Традиционные методы выделения функциональных доменов белков сложные и трудозатратные, а классические нейросетевые подходы требуют увеличения модели по мере увеличения количества данных, поэтому разработка новой архитектуры нейросетевых моделей для автоматического предсказания этих функций представляет собой важную задачу,

Цель: цель данной работы проанализировать данные о доменах в белках, и исследовать использование похожих последовательностей с уже заранее известными ответами - контекстом в нейросетевых моделях с целью улучшения производительности и скорости.

Материалы и методы. В рамках работы использовался язык Python вместе с набором библиотек - для анализа данных, визуализации, конструирования и обучения нейросетевых моделей. В качестве данных были взяты белковые последовательности с заранее выделенными функциональными доменами с ресурса InterPro. Для получения численного представления белковых последовательностей использовались два метода - Word2Vec и ESM. Далее были построены, обучены и сравнивались три модели: модель на основе линейной регрессии - пытается отнести каждую аминокислоту к домену не учитывая окружающие элементы последовательности, модель на основе BiLSTM, которая уже учитывая окружающие аминокислоты выделяет подпоследовательности в домены, модель на основе BiLSTM с контекстом - для входной белковой последовательности из фиксированной базы данных белков с правильными ответами на задачу предсказания доменов выбирается самый похожий помощью выравнивания, и с помощью дополнительного модуля информацию о ‘похоже решенной задаче’ передается в BiLSTM модель, которая используя эту информацию пытается предсказать домены.

Модели обучались на задачу сегментации последовательностей - где каждый элемент последовательности нужно отнести к одному из заданных классов (существующие домены)

Результаты и их обсуждение. Линейная модель справляется с задачей плохо - она не может уловить паттерны, по которым последовательности образуют домены, модель BiLSTM справляется лучше, но имеет проблему - чем больше доменов мы хотим предсказывать, тем больше растет модель, и тем больше данных ей требуется для того чтобы корректно обучиться. Контекстная модель не столько старается запомнить паттерны доменов, сколько учится решать задачу, на основе похожих, поэтому она почти не требует масштабирования по мере роста кол-ва рассматриваемых доменов, а также имеет намного более стабильное обучение и требование к количеству данных, качество же не ниже, чем у модели BiLSTM.

Выводы. Добавления контекста в модель это принцип, который позволяет модели стабильно обучиться на большом количестве данных с сколько угодно большим количеством рассматриваемых классов для предсказания, при этом не требуя увеличения размера самой модели, так как модель начинает обучаться не задаче запоминания паттернов, а задаче сопоставления задачи и похожих на задачу примеров.