От моделирования языка к моделированию генома: опубликована первая в мире нейросетевая модель, обученная на самом полном геноме человека

Исследователи из Института искусственного интеллекта AIRI обучили нейросетевую модель на самой полной на сегодняшний день сборке генома человека. Модель, названная GENA_LM, выложена в open source и доступна биологам по всему миру для использования в научных исследованиях.

Геном – это совокупность наследственного материала, заключенного в клетку организма. ДНК же содержит в себе генетическую информацию, которая определяет характеристики человека – от цвета глаз до предрасположенности к определенным заболеваниям. Последовательность ДНК представляет из себя «текст», закодированный чередованием 4 «букв» – нуклеотидов. Размер генома человека составляет более 3 млрд. таких символов. Однако менее 2% нашего генома кодируют гены, с которых впоследствии образуются молекулы РНК, участвующие в синтезе белков. Остальные 98% генома – последовательность ДНК, которая не кодирует белки и до сих пор мало изучена.

В последние годы в биоинформатике набирают популярность подходы, заимствующие методы обработки естественного языка. Эти методы позволяют выучить закономерности или, другими словами, построить модель последовательности элементов. Особенно важно, что знание, аккумулированное в модели ДНК в процессе обучения, можно использовать повторно для решения широкого класса исследовательских задач: поиск участков генома, выполняющих регуляторные функции в процессах считывания РНК, синтеза белков; определение влияния отдельных мутаций на интенсивность работы генов; предположение патогенного или доброкачественного эффекта от мутаций в ДНК, меняющих одну аминокислоту в белке, классификации живых организмов на основе данных секвенирования и многих других.

В данный момент в мире уже представлен набор достаточно хороших моделей для последовательностей белков *(например, ESM),* но для последовательностей ДНК публично доступна только разработанная коллективом ученых из США модель *DNABERT*. По сравнению с белковыми последовательностями, ДНК намного длиннее, поэтому строить модель на последовательностях ДНК достаточно сложно.

«Наша модель - первая языковая модель для ДНК, обученная на самой полной версии генома человека – T2T-CHM13, которая была опубликована в конце

марта 2022 года[1]. Она может обрабатывать последовательности в 6 раз длиннее, чем DNABERT. Тестирование полученной ДНК модели на одной из задач генетики – предсказании последовательностей, способных «включать» гены (промотеров) – уже показало результаты превосходящие аналогичные с использованием DNABERT»

Ольга Кардымон, руководитель научной группы «Биоинформатика» Института искусственного интеллекта AIRI.

В ближайшем будущем планируется улучшение самой модели и расширение ее возможностей. А для решения прикладных задач скоро будут выложены версии модели для предсказания сайтов сплайсинга, поиска функционально-важных малых рамок считывания белка (uORF), предсказания изменения интенсивности работы генов. Решение этих задач поможет понять больше о механизмах возникновения заболеваний и процессах образования злокачественных клеток. Список задач может быть расширен, исходя из научно-практических интересов биологов, биоинженеров и врачей-генетиков.

«Опубликованная модель – лишь первый шаг нашего исследования. Впереди эксперименты по применению трансформерных архитектур с памятью, которые позволят увеличить размер входной последовательности ещё в несколько раз. Это позволит повысить точность модели и в итоге увеличит качество решения прикладных задач»

Михаил Бурцев, директор по фундаментальным исследованиям Института искусственного интеллекта AIRI, руководитель научной группы «Новые нейронные архитектуры»

Ссылка на репозитории с моделью:

Репозиторий Hugging Face - https://huggingface.co/AIRI-Institute/gena-lm-bert-base
Репозиторий на GitHub - https://github.com/AIRI-Institute/dna-lm

Научно-исследовательский Институт искусственного интеллекта AIRI — автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 90 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами

[1] https://www.science.org/doi/10.1126/science.abj6987#:~:text=3A