УДК 004.89

PARAMETER-EFFICIENT FINE-TUNING ПРЕДОБУЧЕННОЙ ЯЗЫКОВОЙ МОДЕЛИ ДЛЯ УСЛОВНОЙ ГЕНЕРАЦИИ ПРОНИКАЮЩИХ В КЛЕТКУ ПЕПТИДОВ С УЧЕТОМ КЛЕТОЧНОЙ ЛИНИИ

Латыпова А. А.[1], **Нам Е. В.**[1]
Научный руководитель – канд. хим. наук, Серов Н. С.[1]
[1]Университет ИТМО
latypova_adelina@scamt-itmo.ru

Работа выполнена в рамках программы Приоритет 2030.

Введение

Проникающие в клетку пептиды (Cell-Penetrating Peptides, CPP) — это короткие аминокислотные последовательности длиной 5–30 остатков, способные проникать сквозь клеточную мембрану. CPP рассматриваются в качестве перспективных векторов доставки лекарственных молекул благодаря таким преимуществам, как низкая цитотоксичность, возможность нацеливаться на определенные органеллы и способность переносить широкий спектр соединений. Поиск новых CPP традиционно осуществляется методами высокопроизводительного скрининга, что сопряжено со значительными затратами времени и ресурсов. Применение методов *in silico* позволит существенно сократить затраты.

Большинство работ в данной области сводят задачу к бинарной классификации и определяют принадлежность пептидных последовательностей к классу CPP [1]. Существующие генеративные подходы [2, 3], в свою очередь, имеют ряд ограничений, таких как высокая специфичность к конкретной задаче или значительные требования к вычислительным ресурсам. Кроме того, ни одна из предложенных генеративных моделей не учитывает влияние клеточной линии на проникающую способность CPP.

Целью данной работы является разработка условного генератора CPP, учитывающего клеточную линию, а также создание независимых прогнозных моделей для первичного отбора сгенерированных последовательностей.

Основная часть

В условиях ограниченного объема данных и использования вспомогательных моделей особое внимание было уделено разделению обучающих выборок для предотвращения утечки информации и обеспечения объективной оценки генератора.

Для обучения были собраны данные об экспериментально подтвержденных CPP и не-CPP последовательностях. После скрининга моделей и представлений среди регрессионных моделей наилучший результат показал метод опорных векторов (R^2 = 0,52 и MSE = 2,66 на тестовой выборке). Среди классификаторов оптимальной оказалась модель Random Forest, которая далее была настроена на приоритезацию точности (precision). Для минимизации объема обучающей выборки классификатора использовалась стратегия активного обучения (Active Sampling). В результате размер набора был уменьшен до 500 последовательностей при сохранении precision 0,946 и recall 0,496. Анализ показал, что датасет, предназначенный для обучения генератора, остался репрезентативным.

Генеративная модель была реализована на основе предварительно обученной языковой модели ProtGPT2 [4] и дообучена на 1053 CPP с учетом информации о 15 клеточных линиях. В связи с ограниченной эффективностью полного дообучения применялись методы Parameter-Efficient Fine-Tuning (PEFT): Soft-Prompt и Prefix tuning.

Обе модели продемонстрировали генерацию с большей долей лизина и аргинина

по сравнению с реальными CPP, что указывает на смещение в сторону катионных CPP. Prefix модель генерировала более длинные последовательности с более естественным аминокислотным составом, тогда как Soft-Prompt обеспечивала длину, сопоставимую с реальными данными, более высокую уникальность и меньшее отклонение распределений. В качестве финальной модели была выбрана Soft-Prompt.

При внешней оценке с помощью PMIPred [5] 85,86 % сгенерированных последовательностей были классифицированы как связывающиеся с мембраной (кандидаты в CPP), 11,66 % — как промежуточный класс. Доля связывающихся пептидов статистически значимо различалась между клеточными линиями (χ^2-тест), что свидетельствует о влиянии условия генерации на свойства последовательностей. Анализ среднего суммарного заряда последовательностей не выявил простой зависимости между суммарным зарядом и связыванием, что указывает на влияние характеристик более высокого порядка

Выводы

Комбинация PEFT подхода для донастройки предобученной языковой модели с использованием активного обучения позволила реализовать условный генератор CPP в условиях ограниченного объема данных с сохранением разнообразия и ключевых физико-химических свойств последовательностей. Интеграция независимых прогнозных моделей обеспечила высокую точность отбора и интерпретируемую оценку сгенерированных последовательностей.

Полученные результаты подтверждают способность модели генерировать разнообразные CPP с учетом особенностей клеточных линий и демонстрируют потенциал предложенного подхода для рационального дизайна пептидных терапевтических средств.

Литература

1. AI-Driven Design of Cell-Penetrating Peptides for Therapeutic Biotechnology / H. Ma [et al.] // International Journal of Peptide Research and Therapeutics. – 2024. – Vol. 30. – № 6. – P. 69.
2. Automatic generation of functional peptides with desired bioactivity and membrane permeability using Bayesian optimization / I. Fukunaga [et al.] // Molecular Informatics. – 2024. – Vol. 43. – № 4. – P. e202300148.
3. Using molecular dynamics simulations to prioritize and understand AI-generated cell penetrating peptides / D.P. Tran [et al.] // Scientific Reports. – 2021. – Vol. 11. – № 1. – P. 10630.
4. Ferruz N. ProtGPT2 is a deep unsupervised language model for protein design / N. Ferruz, S. Schmidt, B. Höcker // Nature Communications. – 2022. – Vol. 13. – № 1. – P. 4348.
5. PMIpred: a physics-informed web server for quantitative protein-membrane interaction prediction / N. van Hilten [et al.] // Bioinformatics. – 2024. – Vol. 40. – PMIpred. – № 2. – P. btae069.