УДК 004.89

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

Латыпова А. А. (ИТМО), **Нам Е. В.** (ИТМО)
Научный руководитель – кандидат химических наук, Серов Н. С. (ИТМО)

Введение

Проникающие в клетку пептиды (Cell-Penetrating Peptides, CPP) — это короткие аминокислотные последовательности длиной 5–30 остатков, способные проникать сквозь клеточную мембрану. CPP имеют потенциал применения в качестве векторов доставки лекарственных молекул. Этому способствуют такие их преимущества, как низкая цитотоксичность, возможность нацеливания на определенные органеллы и способность переносить широкий спектр грузов. Традиционно поиск новых CPP осуществляется методами высокопроизводительного скрининга и является время- и ресурсозатратным. Применение методов *in silico* позволяет сократить эти затраты.

Большинство существующих работ предлагает в качестве решения применение бинарной классификации [1], а немногочисленные генеративные подходы [2, 3] обладают такими недостатками, как ограниченная способность к обобщению или высокие вычислительные требования. Также ни одна из этих работ по генерации CPP de novo не учитывает влияние клеточных линий на проникающую способность пептидов.

Целью данной работы является разработка условного генератора проникающих в клетки пептидов, а также создание независимых прогностических моделей для первичного отбора сгенерированных последовательностей.

Основная часть

В связи с ограниченным количеством данных и решением использовать прогностические модели для дополнительной фильтрации, особые усилия были приложены к разделению обучающих наборов данных, чтобы снизить утечку информации и обеспечить объективную оценку генератора.

Для обучения были собраны данные об экспериментально подтвержденных CPP и не-CPP последовательностях. После сравнения различных моделей и представлений среди регрессионных алгоритмов лучшую производительность показал метод опорных векторов (R^2 = 0,52 и MSE = 2,66 на тестовой выборке). Из классификаторов оптимальный результат показала модель Random Forest, которая далее была настроена на приоритезацию точности (precision). Для минимизации размера обучающего набора классификатора применялась стратегия активного обучения (Active Sampling). Это позволило сократить выборку до 500 последовательностей при сохранении precision 0,946 и recall 0,496. Анализ физико-химических свойств подтвердил репрезентативность набора данных, оставшегося для обучения генератора.

В качестве базовой модели использовалась предварительно обученная модель белкового языка ProtGPT2 [4]. Далее модель была дообучена на 1053 CPP с учетом информации о 15 клеточных линиях. Поскольку тонкая настройка модели в условиях столь ограниченного количества данных малоэффективна, были применены методы Parameter-Efficient Fine-Tuning (PEFT): Soft-Prompt tuning и Prefix tuning.

Обе модели генерировали последовательности, в которых доля лизина и аргинина была выше, чем в реальных CPP, что указывает на смещение генерации в сторону катионных CPP. Prefix модель генерировала более длинные последовательности с аминокислотным составом, более близким к естественному, тогда как Soft-Prompt модель воспроизводила диапазон длин, соответствующий реальным последовательностям, а также демонстрировала более высокие уникальность и сходство распределений. В качестве финальной модели была выбрана Soft-Prompt.

Внешняя оценка с помощью PMIPred [5] классифицировала 85,86% сгенерированных последовательностей как связывающиеся с мембраной (кандидаты в CPP), 11,66% — как промежуточный класс. Статистический анализ (χ^2-тест) показал, что доля связывающихся пептидов значимо отличается между различными клеточными линиями, что свидетельствует о влиянии задаваемых условий генерации на свойства последовательностей. Анализ значений среднего суммарного заряда не выявил простой зависимости между суммарным зарядом и связыванием пептида с мембраной, что указывает на влияние свойств более высокого порядка.

Выводы

Подводя итог, применение подхода PEFT для донастройки предобученной модели белкового языка в сочетании с активным обучением позволило не только разработать условный генератор CPP в условиях ограниченного объема данных, но и сохранить разнообразие и ключевые физико-химические свойства CPP. Кроме того, применение независимых прогностических моделей обеспечило точность отбора и интерпретируемую оценку сгенерированных последовательностей.

Таким образом, полученные результаты подтверждают способность модели генерировать разнообразные CPP с учетом особенностей клеточных линий и демонстрируют потенциал предложенного подхода для 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.