DOI: 10.25205/978-5-4437-1843-9-15

РАЗРАБОТКА АЛГОРИТМА НА ОСНОВЕ РЕКУРРЕНТНЫХ И СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРЕДСКАЗАНИЯ СТРУКТУРЫ БЕЛКА

RECURRENT AND CONVOLUTIONAL NEURAL NETWORKS FOR PROTEIN STRUCTURE PREDICTION

А. А. Головкин[1,2], А. В. Беспалов[1]

[1] *Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, р. п. Кольцово*
[2] *Новосибирский государственный университет*

A. A. Golovkin[1,2], A. V. Bespalov[1]

[1] *State Research Center of Virology and Biotechnology "Vector", Koltsovo*
[2] *Novosibirsk State University*

✉ a.bespalov@alumni.nsu.ru

Аннотация

В данной работе описана разработка нейронной модели для предсказания структур белков на основе их первичной последовательности аминокислот. Разработанная модель показала высокую эффективность и статистическую достоверность с метриками Weighted PPV, Weighted SEN и F1-Score, равными 0,8901, 0,892 и 0,891 соответственно. Кроме того, значения AUC для каждого класса превысили 0,94, что свидетельствует о надежности и качестве модели.

Abstract

The proposed neural model has demonstrated effectiveness and statistical validity in predicting structures of single-chain polypeptides. This is supported by the metrics with weighted PPV, weighted sensitivity (SEN), and F1-score values of 0.8901, 0.892, and 0.891 respectively. Moreover, the AUC values for each class exceeded 0.94, further confirming the robustness of the model.

Белки представляют собой сложные биологически активные молекулы, состоящие из длинных цепочек аминокислотных остатков. Их структура организована на нескольких уровнях, каждый из которых имеет свои принципы организации, что позволяет воссоздать состояние белка на определенном этапе синтеза с помощью алгоритмов прогнозирования.

Исследование белков представляет значительный научный интерес благодаря их уникальным свойствам и широкому спектру применения. Особое внимание уделяется изучению фармакологических характеристик полипептидов и их роли в биологических процессах человека. Понимание структуры белков позволяет прогнозировать их физико-химические взаимодействия с другими молекулами, а также их функции в организме. Прогнозирование и моделирование белковых структур остаются актуальными задачами, и одним из наиболее эффективных методов их решения сегодня являются нейронные сети. Благодаря комбинации линейных и нелинейных вычислений успешно выявляются сложные взаимосвязи между архитектурой белков и их свойствами. Ключевое преимущество нейросетевых моделей — способность выявлять общие закономерности, применимые ко всей молекуле. Однако для решения различных задач требуется корректная подготовка входных данных. Поскольку белки обладают множеством свойств и конформаций, важным этапом становится выбор оптимального представления данных в зависимости от архитектуры модели и целевых результатов.

Современные алгоритмы предсказания вторичных структур белков, несмотря на их эффективность, имеют ряд ограничений: они могут содержать систематические погрешности и демонстрируют сниженную точность при отсутствии точных атомарных координат целевого белка.

В связи с этим разработка альтернативных подходов к прогнозированию вторичных структур сохраняет свою актуальность. В данной работе описывается создание статистически достоверной нейронной сети, способной с высокой точностью предсказывать вторичные структуры белков исключительно на основе их первичной аминокислотной последовательности.

Набор тестовых данных был создан из файлов, взятых из RCSB Protein Data Bank. При анализе использовались белки, состоящие из одной полипептидной цепочки с известным составом аминокислотных остатков на 70 % и более. В качестве выходных данных были представлены классы вторичных структур белков по системе

DSSP. Полученные последовательности аминокислотных остатков и вторичных структур переводились в численные значения в соответствии с заранее заготовленным словарем.

В описываемую модель данные поступали по 100 белков за одну итерацию. При поступлении входных значений, отражающих первичную структуру белка, первой стадией обработки служил слой Embedding, который создавал контекст и формировал векторное представление аминокислот. После формирования векторного представления матрица транспонировалась и проходила через следующий слой — одномерную свертку Conv1d, которая выделяла новые признаки. Затем матрица снова транспонировалась и подавалась на слой LSTM, выполняющий задачи долгосрочной краткосрочной памяти, что позволило учитывать зависимости в последовательности. На этом этапе каждый элемент данных проходил дальнейшую обработку с помощью специализированных функций. Результатом работы модели была матрица, содержащая вероятности принадлежности каждой аминокислоты к определенному классу вторичной структуры. Итоговый класс для каждого остатка определялся по максимальному значению вероятности. Обучение сети проводилось в течение 1000 эпох. В работе использовался алгоритм стохастического градиентного спуска (SGD) в качестве оптимизатора, а функцией потерь была выбранная NLLLoss. Для оценки качества модели строились ROC-кривые для каждого класса, а также рассчитывались статистические метрики: площадь под ROC-кривой (AUC), положительное прогностическое значение (PPV), чувствительность (SEN) и F1-мера (F1-Score).

В процессе обучения модель улучшила свою точность на тренировочной выборке с 0,8123 на первой эпохе до 0,8946 на тысячной. На тестовой выборке показатели точности были сопоставимы: 0,8110 при первой эпохе и 0,8925 при последней. Значение F1-Score составило 0,891, что высоко и близко к единице, указывая на хорошую способность модели точно предсказывать классы вторичных структур, снижая число ложных срабатываний и корректно выявляя истинные положительные случаи. Все рассчитанные метрики подтверждают, что нейронная сеть демонстрирует высокую и статистически значимую эффективность в предсказании вторичной структуры белков с одной полипептидной цепью на основе аминокислотных последовательностей.