УДК 519.68.02
https://doi.org/10.47533/2020.1606-146X.192

Е. С. ГОЛЕНКО, А. А. ИСМАИЛОВА*, Р. Н. МОЛДАШЕВА

Казахский агротехнический университет им. С. Сейфуллина, Астана, Казахстан

ПРИМЕНЕНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ СТРУКТУРЫ БЕЛКОВ

Быстрое развитие методов машинного и глубокого обучения привело к их повсеместному распространению во многих областях исследований, в том числе они нашли применение и для решения задач протеомики. Сегодня методы машинного обучения играют важную роль в прогнозировании трехмерных структурных компонентов белков, интерпретируя то, как белковые последовательности и их гомология управляют межостаточными контактами и структурной организацией. На последних CASP (Critical Assessment of Structure Prediction – критическая оценка предсказания структуры), где оценивалось положение дел в моделировании структуры белка на основе аминокислотной последовательности, был продемонстрирован значительный прогресс в моделировании конструкций без использования структурных шаблонов (исторически моделирование «ab initio»). Прогресс был обусловлен успешным применением методов глубокого обучения для прогнозирования расстояний между остатками. В свою очередь, эти результаты привели к значительному повышению точности трехмерной структуры при условии, что для семейства белков известно достаточное количество последовательностей. Кроме того, количество последовательностей, необходимых для выравнивания, существенно сократилось, а точность моделей на основе шаблонов также значительно улучшилась. В данной работе предлагается обзор последних успехов в применении методов глубокого обучения, использующихся для предсказания трехмерной структуры белка. Также рассмотрены возможности использования нейронных сетей для выявления неизвестных белковых структур и функций белков, что является одной из главнейших задач протеомики. Описываются проблемы, которые еще предстоит решить, однако ожидается, что в ближайшем будущем описанные методы будут играть решающую роль в структурной биоинформатике белков.

***Ключевые слова**: глубокое обучение, биоинформатика, структура белка, белковая последовательность, протеомика.*

Введение. Выявление структуры белков позволяет лучше понимать их функции, что, в свою очередь, значительно облегчает решение задач в сферах медицины и фармакологии. Недавние достижения в экспериментальных методах структурной биологии, таких как рентгеновская кристаллография, ядерный магнитный резонанс и криогенная электронная микроскопия, способствуют более точному определению структуры белков [1]. Однако из-за высокой стоимости и трудоемкости экспериментального определения разница между количеством белковых последовательностей и относительно небольшим количеством уже известных структур несоразмерно велика. Следовательно, существует необходимость развития теоретических методов, позволяющих быстро и точно выявлять структуру белка. После догмы Анфинсена о том, что нативная структура, по крайней мере, небольшого глобулярного белка определяется только последовательностью, были предприняты различные попытки идентифицировать структуру белка по его последовательности, начиная с предсказания состояний

* E-mail корреспондирующего автора: golenko.katerina@gmail.com

сворачивания белка в 1951г.[2]. Значительный прорыв в технологии секвенирования следующего поколения (Next-Generation Sequencing) привел к быстрому увеличению количества информации о последовательностях, и фундаментальная проблема структурной биоинформатики заключается в прогнозировании трехмерных структур с использованием данных о последовательностях.

Прогнозирование структуры белка стало более быстрым и точным благодаря развитию как традиционных статистических методов, так и методов машинного обучения (ML – Machine Learning) и глубокого обучения (DL – Deep Learning). Искусственная нейронная сеть, особенно глубокая нейронная сеть, хорошо подходит для прогнозирования структуры белка благодаря ее способности выражать широкий спектр функций и ее эффективности, в значительной степени зависящей от количества качественных данных. Первоначально для предсказания предназначаются некоторые элементы белковых структур, такие как спиральный статус или торсионные углы, а затем выводится вся структура с использованием предсказанных признаков, известных как аннотации структуры белка (Protein Structure Annotations – PSA). Использование искусственных нейронных сетей не только увеличивает скорость обработки данных, но и усиливает функциональный анализ крупномасштабных протеомных исследований. Технологии машинного и глубокого обучения, основанные на различных вычислительных методах, позволяют выявлять белок-белковые взаимодействия (PPI – Protein-Protein Interaction) в гетерогенных типах данных протеомики [3].

В данной работе рассмотрены методы прогнозирования структуры белка, основанные на глубоком обучении, наиболее часто используемые архитектуры глубокого обучения и разработанные методы, прогнозирующие аннотации, описывающие различные уровни детализации структуры белка. Также освещены текущие ограничения и проблемы, а также преимущества предсказания структуры белка на основе глубокого обучения.

Гомология белковой последовательности, трехмерная структура и глубокое обучение. Центральная догма молекулярной биологии гласит, что последовательности ДНК транскрибируются в информационную РНК (мРНК), а затем эти последовательности мРНК транслируются в белковые последовательности. Поиск похожих последовательностей можно использовать для выявления «гомологичных» генов или белков путем выявления статистически значимого сходства, указывающего на общее происхождение. Предполагается, что эта белковая последовательность в структурной биологии определяет трехмерную структуру и функцию белка. Это основано на фундаментальном наблюдении, что сходные последовательности из одного и того же эволюционного семейства обычно принимают сходные белковые структуры. Более того, структуры белков очень консервативны в эволюции по сравнению с их последовательностями, и обычно считается, что количество уникальных структурных складок ограничено в природе.

Структура белка может быть определена как один из четырех уровней: первичная, вторичная, третичная или четвертичная структуры. Первичная структура представляет собой линейную последовательность аминокислот. Для образования белка доступно 20 стандартных аминокислот, и каждая аминокислота связана со следующей через пептидные связи. Первичная структура часто представляется в виде цепочки букв, на-

пример, «AESVL…», так как каждая стандартная аминокислота имеет соответствующий однобуквенный код. Это уже дает много полезной информации о структуре белка в трехмерном пространстве благодаря отличительным характеристикам каждой аминокислоты. Многие предсказания структуры белка ab-initio начинаются с этой последовательности аминокислот, первичной структуры. Вторичная структура определяет форму локальных сегментов белков. Обычно это определяется структурой водородных связей полипептидного остова или двугранными углами остова (ϕ, ψ). Двумя распространенными вторичными структурами являются -спирали и β-цепи. α-спираль представляет собой сегмент аминокислот, в котором основная цепь образует спираль, направленную наружу боковыми цепями. Две водородные связи на остаток стабилизируют это образование спирали. -цепь, скорее, связана латерально, где боковые цепи направлены перпендикулярно плоскости, а каждый последующий остаток обращен в противоположную сторону. Эта форма обычно требует наличия партнерской -цепи для своей стабильности. При использовании в предсказании структуры белка вторичная структура обычно попадает в категорию с тремя состояниями или с мелкозернистой классификацией с восемью состояниями. Категоризация с тремя состояниями состоит из двух правильных типов -спирали и β-цепи и одного неправильного типа спиральной области.

Третичная и четвертичная структуры объясняют трехмерное расположение одиночных и множественных белков соответственно. Их можно представить с помощью декартовых координат каждого атома в трехмерном пространстве. Из-за водной природы белков основной движущей силой, определяющей третичную и четвертичную структуру, является гидрофобное взаимодействие между аминокислотами и молекулами воды. Таким образом, белки, как правило, обладают гидрофобным ядром, где боковые цепи скрыты, избегая полярных молекул воды. Такая трехмерная информация может быть выведена, когда уже есть первичная структура, вторичная структура и промежуточный остаток карты контактов (Contact Map – CM).

В отличие от последовательностей, число которых практически бесконечно, белки могут принимать конечное число различных форм, чтобы выполнять свои функции в клетке. Можно наблюдать более сильное сохранение структуры, чем сохранение последовательности; например, сильная взаимозависимость для полярных остатков существует в ядре белка с плохой доступностью для растворителя, но при рассмотрении только последовательностей не обнаруживается значительной корреляции [4]. Это позволяет предсказать структуру белка, более консервативный домен, на основании большого количества данных о последовательностях. Следовательно, были предприняты различные попытки разгадать взаимосвязь между структурой и последовательностью, включая методологии глубокого обучения.

Материалы и методы исследования. Обзор методов глубокого обучения. Глубокое обучение – это ветвь машинного обучения, в которой используется искусственная нейронная сеть со множеством встроенных слоев, напоминающая нервную систему человека. Работая как универсальные аппроксиматоры функций, глубокие нейронные сети используются для решения различных задач.

Искусственные нейронные сети состоят из узлов входного, выходного и скрытого слоев, где каждый узел связан с узлами соседних слоев. Эти соединения имеют

разные веса, и входные данные обрабатываются (т. е. умножаются и суммируются) в каждом узле. Затем он подвергается преобразованию на основе функции активации, такой как сигмовидная или выпрямительная, а выходные функции используются в качестве входных данных для следующего уровня. Обучение – это процесс поиска оптимальных весов, которые заставляют нейронную сеть вести себя так, как нужно. Существует два типа обучения: контролируемое обучение обрабатывает помеченные наборы данных для целей классификации или прогнозирования, а неконтролируемое обучение обрабатывает немаркированные наборы данных для анализа или кластеризации данного набора данных. Объем необходимых обучающих данных для построения эффективных моделей глубокого обучения зависит от сложности и количества функций в обучающих данных. Для обновления и оптимизации весов используется обратное распространение для вычисления градиента функции потерь, которая вычисляет ошибку для каждой итерации обучения. Однако при использовании слишком большого количества слоев градиенты либо исчезают, либо взрываются, что делает процесс обучения неэффективным. Для преодоления этой проблемы существуют определенные приемы, такие как модификации функций активации и использование пропущенных соединений (остаточная нейронная сеть) [5].

Результаты. Самым простым и ранним примером глубокой нейронной сети является нейронная сеть с прямой связью (FeedForward Neural Network – FFNN), иногда называемая многослойным персептроном (MultiLayer Perceptron – MLP). Персептрон, однослойная нейронная сеть, может обрабатывать информацию только первого порядка для получения результатов, сравнимых с результатами, полученными с помощью множественной линейной регрессии. Когда используется несколько слоев, нейронные сети могут извлекать признаки более высокого порядка. В FFNN информация течет в одном направлении – от входного слоя к скрытым слоям, если таковые имеются, пока не достигнет выходного слоя. Сеть имеет соединения между каждым узлом и каждым другим узлом на следующем уровне.

Рекуррентная нейронная сеть (Recurrent Neural Network – RNN) содержит циклы, в которых выход слоя становится входом. Этот цикл генерирует нейроны состояния, которые позволяют сети сохранять память о предыдущем состоянии. Получение будущей памяти благоприятно для предсказания и осуществимо с RNN за счет введения задержки, но скорость предсказания падает, если задержка слишком велика. Чтобы решить эту проблему, была разработана двунаправленная рекуррентная нейронная сеть (Bidirectional Recurrent Neural Network – BRNN), разделяющая нейроны состояния на положительное и отрицательное направления времени. 2D-BRNN – двумерное приложение BRNN – широко используется для правильного прогнозирования карты контакта остатков (CM), обычно с использованием векторов с четырьмя состояниями, обрабатывающих четыре основных угла карты [6]. Долгосрочная память (Long-Short Term Memory – LSTM) – это вариант элементарной ячейки, используемый в RNN, предназначенный для решения проблемы исчезающего градиента путем введения в элементарную ячейку вентильных функций. Этот контроль ошибок позволяет LSTM изучать долгосрочные зависимости между точками данных. Благодаря своей способности разрешать последовательность в качестве входных и выходных данных, RNN известны своей отличной производительностью при решении любых задач, основан-

ных на последовательностях, подходящих для предсказания структуры белка с белковой последовательностью в качестве входных данных.

Сверточная нейронная сеть (Convolutional neural network – CNN) часто включает в себя три типа слоев: сверточный, объединяющий и полностью связанный слои. CNN обычно принимает входные данные, такие как 2D-изображение, и сверточные слои применяют различные ядра для его свертки, где каждое ядро действует как персептрон, генерируя карты признаков. Затем следует слой объединения, чтобы выполнить уменьшение размерности параметров сети и карт объектов. Результаты передаются в полносвязные слои, отображая карты 2D-объектов в 1D-вектор для дальнейшего представления объектов. Основным преимуществом применения схемы свертки является массовый параллелизм, обеспечивающий большую вычислительную эффективность.

Сверточная сеть графов (Graph Convolutional Network – GCN), обобщение оператора свертки на неевклидово структурированных данных, содержит несколько спектральных или пространственных сверточных слоев [7]. Его уникальные стратегии определения характеристик на уровне ввода с продуманной архитектурой подходят для сложных задач, таких как белок-белковые взаимодействия.

Основываясь на коэволюционном анализе и методах глубокого обучения, методы прогнозирования структуры белка в последние годы достигли значительного прогресса за счет использования множественного выравнивания последовательностей целевого белка и его гомолога. Комбинация упомянутых выше архитектур широко используется в этом типе методов предсказания структуры белка. Одним из известных примеров может быть комбинация двунаправленных RNN и CNN (BRNN-CNN) [8]. В этой схеме сверточное ядро отображает окно памяти BRNN в локальное состояние. Существуют варианты, такие как двунаправленный LSTM, за которым следует CNN (BLSTM–CNN) [9]. Доступно неограниченное количество гибридных топологий, но необходимо тщательно спроектировать архитектуру, учитывая сложность обучения, вычислительную сложность и требования к памяти, чтобы получить максимальную точность.

Прогнозирование белковых 3D-структур. Одной из основных целей структурной биоинформатики является раскрытие взаимосвязи между отдельными аминокислотами, входящими в состав белка, и соответствующими трехмерными структурами, т.е. выявление взаимосвязи между генотипом и фенотипом. Существует несколько экспериментальных методов определения структуры, которые постоянно развиваются. Однако именно глубокое обучение стало доминирующей технологией для предсказания белковых структур на основе контактных или эволюционных карт.

Критическая оценка предсказания структуры белка (Critical Assessment of Protein Structure Prediction – CASP). CASP (https://predictioncenter.org/index.cgi) – это соревнование, проводимое два раза в год глобальными совместными усилиями, предназначенное для оценки современных методов предсказания структуры белка. Алгоритм предсказания третичной структуры можно разделить на следующие части: моделирование гомологии, при котором в качестве шаблона используется известная структура с аналогичной последовательностью (моделирование на основе шаблонов); распознавание складок, также называемое нитями белков (требуются шаблоны); и предсказа-

ние структуры de novo, то есть моделирование без шаблонов. Недавние достижения в методах, связанных с глубоким обучением, повысили точность предсказания контактного расстояния и коэволюционного анализа остаток-остаток, и, наконец, за последние несколько лет был достигнут значительный прогресс в предсказании структуры белка без шаблонов, а также моделирование на основе шаблонов.

Трехмерное прогнозирование структуры на основе контактных карт. Геномные последовательности могут быть эффективно использованы для обнаружения корреляций или ковариаций между остатками в белках (так называемые «эволюционные связи»). Анализ этой ковариации может помочь идентифицировать непосредственно контактирующие остатки в 3D-конформациях, функциональные остатки в связывании субстрата или остатки, участвующие в белок-белковых взаимодействиях. Как обсуждалось в предыдущем разделе, карта контактов представляет собой двумерную матрицу, кодирующую отсутствие/присутствие или вероятность контакта между парами остатков в данном белке.

Значения вблизи главной диагонали в CM тривиальны, потому что это значения из соседних пар аминокислот. Наиболее актуальная информация в карте контактов расположена далеко от главной диагонали. Анализ элементов, удаленных от главной диагонали, может дать полезную информацию о структурных свойствах и пространственных деталях белкового остова. Следовательно, эти контакты или карты мультиклассовых контактов могут предоставить информацию о пространственной организации белка и могут быть использованы для улучшения качества предсказанной третичной структуры. Ожидается, что в случае типичного глобулярного белка почти 90% всех пар остатков будут неконтактирующими, так что только небольшая часть расстояний между аминокислотами должна точно использоваться в качестве ограничений для прямого определения структуры. Усовершенствованные методы глубокого обучения показали себя многообещающе в прогнозировании точных контактов между остатками. Чтобы повысить точность CM, может потребоваться рассмотреть несколько ключевых факторов, таких как методы глубокого обучения, надежное выравнивание нескольких последовательностей (Multiple Sequence Alignments – MSA), прогнозирование распределения расстояния и интеграция контактов на основе домена.

AlphaFold (Google DeepMind), последняя тенденция в этой области, была впервые выпущена на CASP13 (2018 г.) и превратилась в AlphaFold2 на CASP14 (2020 г.) [10]. DL с алгоритмом внимания обучал нейронные сети примерно 170 000 известных белковых структур. Во-первых, коэволюционный анализ используется для сопоставления ковариаций аминокислотной последовательности с физическим контактом в трехмерной структуре белка, а затем исследуется с использованием нейронных сетей для изучения моделей коэволюционных взаимодействий и преобразования их в CM. Основываясь на эволюционно связанных белковых последовательностях и парах аминокислотных остатков, модель итеративно генерирует структуру, передавая информацию туда и обратно между обоими представлениями. В AlphaFold1 карта расстояний создается на основе информации о множественном выравнивании последовательностей и используется для создания направляющего потенциала. Используется простой метод градиентного спуска, чтобы непосредственно свернуть белок в структуру, со-

вместимую с предсказанными расстояниями. Затем энергетическая функция Розетты используется для уточнения окончательной складчатой структуры. Подробный процесс AlphaFold2 еще не опубликован, но известно, что процесс направляющего потенциала заменен системой, полностью основанной на распознавании образов, а уточнение энергии на основе силового поля AMBER применяется в качестве последнего шага уточнения. Модель показала выдающиеся результаты со средним баллом теста глобального расстояния (Global Distance Test – GDT) около или выше 90 по всем целям. Программа смогла надежно предсказать структуры мембранных белков, которые до сих пор было чрезвычайно трудно распознавать.

Хотя во время двух экспериментов CASP серия AlphaFold была более заметной, чем другие конкуренты, множество других программ прогнозирования были разработаны на основе глубокого обучения и продемонстрировали значительный прогресс. Например, RaptorX [11] – это сервер для прогнозирования структуры и функций белков, основанный на DL. Он предсказывает вторичную и третичную структуры белка, доступность растворителя, неупорядоченные области, функциональную аннотацию и возможные сайты связывания. Он также обеспечивает распределение межостаточного/межатомного расстояния и вероятности ориентации, которые могут использоваться другими алгоритмами укладки для восстановления трехмерных моделей белков. В программе качество профилей MSA оценивается методом профильно-энтропийной оценки с учетом имеющихся неизбыточных гомологов. Затем условные случайные поля используются для интеграции множества биологических сигналов в нелинейную функцию оценки потоков. Пакет Rosetta использует алгоритм для предсказания структуры de novo, который также используется при работе со складками белков в расходящихся доменах моделей гомологии. Исходный фолдинг коротких сегментов выбирается из базы данных структуры белков, тогда как более длинные сегменты строятся из фрагментов из трех и девяти остатков, выбранных из базы данных, а затем собираются с использованием алгоритма Rosetta [12]. SPOT-fold – это инструмент предсказания структуры белка без фрагментов ab initio, основанный на предсказанной структуре скелета и CM из SPOT-Contact, а также на предсказанных двугранных углах из SPIDER3 [13]. MULTICOM – это сервер моделирования белковой структуры с поддержкой глубокого обучения и прогнозирования контактного расстояния [14]. EVfold разделяет прямые и непрямые корреляции остаток-остаток в больших множественных выравниваниях последовательностей и выводит прямые эволюционные связи остаток-остаток [15].

В сочетании с моделированием на основе шаблонов. Одним из популярных и успешных подходов к предсказанию структуры белка является моделирование гомологии, основанное на двух принципах: (i) аминокислотная последовательность определяет типичную укладку или трехмерную структуру белка, и (ii) трехмерная структура сохраняется по отношению к первичным последовательностям. Использование известных структур гомологичных белков, обладающих определенной степенью сходства последовательностей, является достаточно удобным и эффективным способом построения исходной модели. Однако проблемы, связанные со слабым сходством последовательности и структуры, выравниванием последовательностей со структурами, моделированием сдвигов твердого тела и точными конформациями

петель и боковых цепей, а также обнаружением ошибок в модели, все еще актуальны на сегодняшний день. Комбинация с подходами, основанными на глубоком обучении, в последнее время превосходит традиционные методы, достигая значительного повышения точности модели. Результаты CASP13 и 14 демонстрируют, что сложное сопоставление аминокислотной последовательности и трехмерной структуры белка может быть эффективно изучено с помощью нейронной сети и обобщено на ранее недоступные случаи.

Моделирование гомологии обычно выполняется с помощью следующих шагов: (i) идентификация и выбор подходящих матриц, т.е. других гомологичных белков с известными трехмерными структурами (связанные программы: BLAST, PSI-BLAST, HHsuite, JackHMMer); (ii) множественное выравнивание последовательностей (Omega, MUSCLE и т. д.) [16]; (iii) построение 3D-моделей (SWISS-MODEL, MODELLER, I-TASSER); (iv) моделирование петель, представляющих собой вариабельные и неконсервативные области; (v) моделирование боковой цепи на основе библиотеки ротамеров, функции оценки и метода сканирования (OPUS-Rota2, FASPR, SCWRL); (vi) оптимизация модели, повышающая качество конечной модели (как правило, минимизация энергии, молекулярная динамика или моделирование методом Монте-Карло); и (vii) оценка и проверка модели.

Методы на основе глубокого обучения могут использоваться для повышения точности на каждом этапе. Например, DLPAlign является примером подхода глубокого обучения в сочетании с выравниванием последовательностей [17]. Это новый и простой подход к повышению точности прогрессивного метода MSA путем обучения модели принятия решений на основе сверточных нейронных сетей. DESTINI (глубокий структурный вывод для белков) — это новый вычислительный подход, который сочетает в себе алгоритм глубокого обучения для предсказания контактов белковых остатков и остатков со структурным моделированием на основе шаблонов [18]. ThreaderAI сначала применяет глубокое обучение для прогнозирования матрицы вероятности выравнивания остатков путем интеграции профиля последовательности, предсказанных последовательных структурных особенностей и предсказанных контактов остаток-остаток, а затем использует динамическое программирование для выравнивания шаблона и запроса в соответствии с матрицей вероятностей [19].

Обсуждение. Методологии DL сегодня являются ведущим трендом в пограничных областях, а именно в оценке качества модели (Quality Assessment - QA), которая является последующим шагом для предсказания структуры белка. Как для методов, основанных на шаблонах, так и для методов без шаблонов, за QA следуют прогнозы структуры для измерения расхождения со структурами белка с нативной укладкой. Начиная с CASP7 (2006 г.), QA была включена в категорию конкурентов для разработки методов оценки качества и правильности моделей структуры белка. Более ранние статистические методы, включая PROCHECK и WhatCheck, были сосредоточены на стереохимии структуры белка, такой как двугранные углы основной цепи или несвязанные расстояния между остатками. Предсказанные модели также могут быть оценены с использованием энергий взаимодействия остаток-остаток, где пики в энергетическом профиле будут означать ошибочное предсказание области. Позже были разработаны и выделены методы обеспечения качества на основе DL.

AngularQA использовала свойства последовательности, такие как вторичные структуры в дополнение к углам, при задаче обеспечения качества, став первой попыткой использования ячеек LSTM для задач обеспечения качества. GraphQA решал проблемы обеспечения качества с GCN для желаемых свойств, таких как обучение представлению, геометрическая инвариантность, явное моделирование трехмерной структуры и т.д. [20].

Выводы. Были достигнуты значительные успехи в прогнозировании карт контактов белков на основе множественного выравнивания последовательностей гомологичных белков путем анализа сигналов, связанных с коэволюцией. Сочетание подходящих методов глубокого обучения стало мощной основой для понимания лежащих в основе взаимосвязей между последовательностями и структурными элементами. В этом обзоре были рассмотрены текущие тенденции в области прогнозирования структуры белков, в частности современные методы в сочетании с архитектурой глубокого обучения для прогнозирования карт контактов. Глубокое обучение только начало применяться к биомолекулярной структуре, но показало успешную стратегию в области прогнозирования. Современные методы, основанные на глубоком обучении, обеспечивают значительный прогресс, однако это не означает, что они, в конечном счете, полностью решили проблемы протеомики. Например, некоторые белки имеют гибкие, «внутренне неупорядоченные» части в своих структурах, а не четко определенные формы. Эти неупорядоченные части могут действовать как функциональная единица. Подходы, основанные на DL, также показали высокую эффективность при предсказании этих областей, но они не интерпретируют функциональный механизм этих гибких областей. Таким образом, необходимо разработать подходы к глубокому обучению, которые также смогут учесть некоторые из этих предостережений.

В некоторых случаях из-за ограничений, присущих исследованиям, основанным на данных, может быть сложно построить надежные модели из-за отсутствия высококачественных наборов данных. Тем не менее, это ограничение может быть преодолено за счет включения экспертных знаний в предметной области и постоянного увеличения высококачественных наборов данных. Ожидается, что прогнозирование карты контактов и моделирование на основе гомологии станут ведущими подходами для повышения точности в крупномасштабной задаче прогнозирования.

ЛИТЕРАТУРА

1 Либшнер Д., Афонин П.В., Бейкер М.Л., Бункочи Г., Чен В.Б., Кролл Т.И., Хинце Б., Хунг Л.-В., Джейн С., Маккой А.Дж. и др. Определение макромолекулярной структуры с использованием рентгеновских лучей, нейтронов и электронов: Последние разработки в Phenix. Акт кристаллографии. Раздел D. 2019. № 75 – стр.861–877.

2 Анфинсен К.Б. Принципы, управляющие сворачиванием белковых цепей. Наука. 1973. №181– с.223–230.

3 Сун Т., Чжоу Б., Ли Л., Пей Дж. Предсказание взаимодействия белок-белок на основе последовательности с использованием алгоритма глубокого обучения. BMC Bioinform. 2017. №18 – с.1-8.

4 Родионов М.А., Бланделл Т.Л. Сохранение последовательности и структуры в белковом ядре. Структура белков. Функция. Биоинформ. 1998. №33 – с.358–366.

5 Ху Ю., Хубер А., Анумула Дж., Лю С.- С. Преодоление проблемы исчезающего градиента в простых рекуррентных сетях. arXiv. 2018. arXiv: 1801.06105.

6 Ди Лена П., Нагата К., Балди П. Глубокие архитектуры для прогнозирования контактной карты белков. Биоинформатика. 2012. №28 – с.2449-2457.

7 Глигориевич В., Ренфрю П.Д., Костюлек Т., Леман Дж.К., Беренберг Д., Ватанен Т., Чандлер К., Тейлор Б.К., Фиск И.М., Вламакис Х. Структурное предсказание функций с использованием сверточных сетей графов. bioRxiv. 2020.

8 Торриси М., Кейл М., Полластри Г. Более глубокие профили и каскадные рекуррентные и сверточные нейронные сети для современного прогнозирования вторичной структуры белка. Sci. Rep. 2019. №9 – с.1–12.

9 Чжан Ю., Цяо С., Цзи С., Ли Ю. DeepSite: Двунаправленные модели LSTM и CNN для прогнозирования связывания ДНК с белком. Инт. Дж. Мах. Учить. Кибернетика. 2020. №11. – С.841–851.

10 Сеньор А.У., Эванс Р., Джампер Дж., Киркпатрик Дж., Сифре Л., Грин Т., Цинь К., Зидек А., Нельсон А.У.Р. Бриджленд А. и др. Улучшенное предсказание структуры белка с использованием возможностей глубокого обучения. Природа. 2020. №577 – с.706–710.

11 Ван С., Сун С., Ли З., Чжан Р., Сюй Дж. Точное предсказание De Novo карты контактов белков с помощью модели сверхглубокого обучения. PLoS Вычисляет. Биол. 2017. №13 – e1005324.

12 Леман Дж.К., Вайснер Б.Д., Льюис С.М., Адольф-Брайфогл Дж., Алам Н., Элфорд Р.Ф., Апрахамян М., Бейкер Д., Барлоу К.А., Барт П. и др. Макромолекулярное моделирование и проектирование в Розетте: новейшие методы и основы. Натуральный. Методы. 2020. №17. – С.665–680.

13 Cai Y., Li X., Sun Z., Lu Y., Zhao H., Hanson J., Paliwal K., Litfin T., Zhou Y., Yang Y. ТОЧЕЧНАЯ складка: Прогнозирование структуры белка без фрагментов На основе прогнозируемой структуры остова и карты контактов. J. Вычисл. Химия. 2020. №41 – с.745-750.

14 Хоу Дж., Ву Т., Цао Р., Ченг Дж. Моделирование третичной структуры белка на основе глубокого обучения и прогнозирования расстояния контакта в CASP13. Белки. 2019. №87 – с.1165–1178.

15 Хопф Т.А., Шарфе К.П., Родригес Дж.П., Грин А.Г., Кольбахер О., Сандер С., Бонвин А.М., Маркс Д.С. Коэволюция последовательностей дает трехмерные контакты и структуры белковых комплексов. eLife. 2014. №3 – e03430.

16 Сиверс Ф., Хиггинс Д.Г. Кластерная Омега для точного выравнивания многих белковых последовательностей. Белковая наука. 2018. №27 – с.135–145.

17 Куанг М., Лю Ю., Гао Л. DLPAlign: Основанный на глубоком обучении метод прогрессивного выравнивания для нескольких белковых последовательностей. В трудах CSBio'20: Материалы Одиннадцатой Международной конференции по вычислительным системам - Биологии и биоинформатике, Бангкок, Таиланд.19-21 ноября 2020 г. Стр.83–92.

18 Гао М., Чжоу Х., Скольник Дж. СУДЬБА: подход с глубоким обучением к прогнозированию структуры белка на основе контактов. Sci. Rep. 2019. №9 – с.3514.

19 Чжан Х., Шен Ю. Предсказание структуры белка на основе шаблонов с помощью глубокого обучения. BMC Genom. 2020. №21 – с. 878.

20 Бальдассарре Ф., Менендес Уртадо Д., Элофссон А., Азизпур Х. GraphQA: оценка качества белковой модели с использованием сверточных сетей графов. Биоинформатика. 2021, №37 – с.360–366.

REFERENCES

1 Liebschner D., Afonine P.V., Baker M.L., Bunkoczi G., Chen V.B., Croll T.I., Hintze B., Hung L.-W., Jain S., McCoy A.J., et al. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. Acta Crystallogr. Sect. D. 2019. No.75 – pp.861–877.

2 Anfinsen C.B. Principles that Govern the Folding of Protein Chains. Science. 1973. No.181- pp.223–230.

3 Sun,T., Zhou B., Lai L., Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinform. 2017. No.18 – pp.1–8.

4 Rodionov M.A., Blundell T.L. Sequence and structure conservation in a protein core. Proteins Struct. Funct. Bioinform. 1998. No.33 – pp.358–366.

5 Hu Y., Huber A., Anumula, J., Liu S.-C. Overcoming the vanishing gradient problem in plain recurrent networks. arXiv. 2018. arXiv:1801.06105.

6 Di Lena P., Nagata K., Baldi P. Deep architectures for protein contact map prediction. Bioinformatics. 2012. No.28 – pp.2449–2457.

7 Gligorijevic V., Renfrew P.D., Kosciolek T., Leman J.K., Berenberg D., Vatanen T., Chandler C., Taylor B.C., Fisk I.M., Vlamakis H. Structure-based function prediction using graph convolutional networks. bioRxiv. 2020.

8 Torrisi M., Kaleel M., Pollastri G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Sci. Rep. 2019. No.9 – pp.1–12.

9 Zhang Y., Qiao S., Ji S., Li Y. DeepSite: Bidirectional LSTM and CNN models for predicting DNA–protein binding. Int. J. Mach. Learn. Cybern. 2020. No.11 – pp.841–851.

10 Senior A.W., Evans R., Jumper J., Kirkpatrick J., Sifre L., Green T., Qin C., Zidek A., Nelson A.W.R. Bridgland A., et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020. No.577- pp.706–710.

11 Wang S., Sun S., Li Z., Zhang R., Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput. Biol. 2017. No.13 – e1005324.

12 Leman J.K., Weitzner B.D., Lewis S.M., Adolf-Bryfogle J., Alam N., Alford R.F., Aprahamian M., Baker D., Barlow K.A., Barth P., et al. Macromolecular modeling and design in Rosetta: Recent methods and frameworks. Nat. Methods. 2020. No.17 – pp.665–680.

13 Cai Y., Li X., Sun Z., Lu Y., Zhao H., Hanson J., Paliwal K., Litfin T., Zhou Y., Yang Y. SPOT-Fold: Fragment-Free Protein Structure Prediction Guided by Predicted Backbone Structure and Contact Map. J. Comput. Chem. 2020. No.41 – pp.745–750.

14 Hou J., Wu T., Cao R., Cheng J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins. 2019. No.87 – pp.1165–1178.

15 Hopf T.A., Scharfe C.P., Rodrigues J.P., Green A.G., Kohlbacher O., Sander C, Bonvin A.M., Marks D.S. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife. 2014. No.3 – e03430.

16 Sievers F., Higgins D.G. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018. No.27 – pp.135–145.

17 Kuang M., Liu Y., Gao L. DLPAlign: A Deep Learning based Progressive Alignment Method for Multiple Protein Sequences. In Proceedings of the CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics, Bangkok, Thailand.19–21 November 2020. pp.83–92.

18 Gao M., Zhou H., Skolnick J. DESTINI: A deep-learning approach to contact-driven protein structure prediction. Sci. Rep. 2019. No.9 – pp.3514.

19 Zhang H., Shen Y. Template-based prediction of protein structure with deep learning. BMC Genom. 2020. No.21 – pp. 878.

20 Baldassarre F., Menéndez Hurtado D., Elofsson A., Azizpour H. GraphQA: Protein model quality assessment using graph convolutional networks. Bioinformatics. 2021, No.37 – pp.360–366.

Е. С. ГОЛЕНКО, А. А. ИСМАИЛОВА, Р. Н. МОЛДАШЕВА

С. Сейфуллин атындағы Қазақ агротехникалық университеті, Астана, Қазақстан

АҚУЫЗДЫҢ ҚҰРЫЛЫМЫН БОЛЖАУ ҮШІН ТЕРЕҢ ОҚУ ӘДІСТЕРІН ҚОЛДАНУ

Машиналық оқытудың және тереңдету әдістерінің қарқынды дамуы олардың көптеген зерттеу салаларында кеңінен қолданылуына, соның ішінде протеомика мәселелерін шешуде қолданылуына әкелді. Бүгінгі таңда машиналық оқыту әдістері белок тізбегі мен олардың гомологиясы қалдық байланыстар мен құрылымдық ұйымды қалай басқаратынын түсіндіру арқылы белоктардың үш өлшемді құрылымдық компоненттерін болжауда маңызды рөл атқарады. Аминқышқылдарының тізбегі негізінде ақуыз құрылымын модельдеу техникасының жай-күйін бағалайтын соңғы CASP (Құрылымды болжауды сыни бағалау) құрылымдық үлгілерді қолданбай (тарихи түрде «ab initio» модельдеу) құрылымдарды модельдеуде айтарлықтай прогресті көрсетті. Прогресс қалдықтар арасындағы қашықтықты болжау үшін терең оқыту әдістерін сәтті қолдану арқылы жүзеге асты. Өз кезегінде бұл нәтижелер белоктар отбасы үшін белгілі реттіліктердің жеткілікті саны болған жағдайда үш өлшемді құрылымның дәлдігінің айтарлықтай артуына әкелді. Сонымен қатар, туралау үшін қажетті реттіліктердің саны айтарлықтай қысқарды және үлгіге негізделген үлгілердің дәлдігі де айтарлықтай жақсарды. Бұл құжат ақуыздың үш өлшемді құрылымын болжау үшін қолданылатын терең оқыту әдістерін қолданудағы соңғы жетістіктерге шолу жасайды. Протеомиканың маңызды міндеттерінің бірі болып табылатын белоктардың белгісіз белок құрылымдары мен функцияларын анықтау үшін нейрондық желілерді қолдану мүмкіндіктері де қарастырылады. Әлі шешілмеген мәселелер сипатталған, бірақ таяу болашақта сипатталған әдістер белоктардың құрылымдық биоинформатикасында шешуші рөл атқарады деп күтілуде.

Түйін сөздер*: тереңдетіп оқыту, биоинформатика, ақуыз құрылымы, ақуыз тізбегі, протеомика.*

Y. S. GOLENKO, A. A. ISMAILOVA, R. N. MOLDASHEVA

S. Seifullin Kazakh Agrotechnical University, Astana, Kazakhstan

APPLICATION OF DEEP LEARNING METHODS FOR PROTEIN STRUCTURE PREDICTION

the rapid development of machine learning and deep learning methods has led to their widespread use in many areas of research, including their use in solving problems of proteomics. Today, machine learning methods play an important role in predicting the three-dimensional structural components of proteins by interpreting how protein sequences and their homology govern interresidual contacts and

structural organization. Recent CASPs (Critical Assessment of Structure Prediction) assessing the state of the art in modeling protein structure based on amino acid sequence have demonstrated significant progress in modeling structures without the use of structural templates (historically "ab initio" modeling). The progress was driven by the successful application of deep learning techniques to predict the distances between residuals. In turn, these results led to a significant increase in the accuracy of the three-dimensional structure, provided that a sufficient number of sequences are known for a family of proteins. In addition, the number of sequences required for alignment has been significantly reduced, and the accuracy of template-based models has also improved significantly. This paper provides an overview of recent advances in the application of deep learning methods used to predict the three-dimensional structure of a protein. The possibilities of using neural networks to identify unknown protein structures and functions of proteins, which is one of the most important tasks of proteomics, are also considered. Problems that have yet to be solved are described, but it is expected that in the near future the described methods will play a decisive role in the structural bioinformatics of proteins.

***Key words**: deep learning, bioinformatics, protein structure, protein sequence, proteomics.*