«Сейфуллин оқулары – 18: « Жастар және ғылым – болашаққа көзқарас» халықаралық ғылыми -практикалық конференция материалдары = Материалы международной научно-практической конференции «Сейфуллинские чтения – 18: « Молодежь и наука – взгляд в будущее» - 2022.- Т.I, Ч.IV. - С. 42-45

МЕТОДЫ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ОДНОМЕРНЫХ И ДВУМЕРНЫХ АННОТАЦИЙ БЕЛКА

Голенко Е. С., докторант III курса
Казахский агротехнический университет им. С. Сейфуллина, Нур-Султан, Казахстан

Введение

Прогнозирование структуры белка стало более мощным и точным благодаря развитию методов от традиционных статистических методов до методов машинного обучения (Machine Learning - ML) и глубинного обучения (Deep Learning - DL) [1].

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. Он основан на фундаментальном наблюдении, что сходные последовательности из одного и того же эволюционного семейства обычно принимают сходные белковые структуры.

Белки и их функции отличаются своей структурой во многих аспектах, но скорость обнаружения белковых структур была намного ниже, чем скорость идентификации последовательностей из-за стоимости и сложности. Таким образом, предиктор структуры белка стал одним из наиболее эффективных и высокопроизводительных инструментов в биоинформатике для обработки данных об известных последовательностях с помощью развивающихся методологий, таких как статистические методы, методы машинного обучения и глубинного обучения. Функция, используемая в процессе прогнозирования, известна как PSA; он содержит упрощенную информацию для облегчения вычислительного процесса и используется в качестве промежуточного шага для оценки полной структуры белка. Большое внимание привлекли одномерные (1D-) и двумерные (2D-) PSA, где вторичная структура, доступность растворителя или внутренняя неупорядоченность в основном описываются как 1D-PSA, а карта контактов (Contact Map – CM) или подробная версия CM (мультиклассовый CM или карта расстояний) выражается с помощью 2D-PSA. Было разработано несколько приложений DL для прогнозов 1D- и 2D-

PSA, которые становятся более точными благодаря расширению доступности данных о последовательности и структуре.

Наиболее полезной особенностью 1D-PSA является вторичная структура, самый первый шаг для предсказания полной структуры белка по последовательности. Доступны две основные классификации: категоризация с тремя состояниями на α-спираль, β-цепь и спиральную область или восемь мелкозернистых категоризаций, которые дополнительно разделяют предыдущие три состояния. Точность может быть легко выражена процентной точностью с тремя состояниями (оценка $Q3$) или процентной точностью с восемью состояниями (оценка $Q8$), которая определяется как процент правильно предсказанных остатков вторичной структуры.

Одним из первых доступных серверов для предсказания вторичной структуры был JPred. Сервер использует шесть различных алгоритмов предсказания вторичной структуры: DSC, использующий линейную дискриминацию, PHD, использующий нейронные сети принятия решений жюри, NNSSP, использующий ближайших соседей, PREDATOR, использующий склонность к водородным связям, ZPRED, использующий взвешенное предсказание числа сохранения, и MULTIPRED, использующий согласованную комбинацию методов одной последовательности [2]. Стал доступен еще один сервер прогнозирования вторичной структуры, PSIPRED, где использовался метод сопряжения двух FFNN, обучающих нейронные сети на информации об эволюционном сохранении, полученной из PSIBLAST. Другая попытка под названием SSpro показала применение расширенного алгоритма с использованием BRNN-CNN [3]. В методе используется смесь оценщиков, которая использует эволюционную информацию, указанную в нескольких выравниваниях, как на входном, так и на выходном уровнях BRNN. Porter, Porter+ и PaleAle из серии Distill также основаны на ансамблях BRNN-CNN, каждый из которых используется для предсказания различных 1D-PSA. В следующих методах Distill последовательность обрабатывается на первом этапе BRNN-CNN, а затем вытягивается в набор средних значений, которые обрабатываются вторым этапом BRNN-CNN. Портер добился лучших результатов, используя как PSI-BLAST, так и HHBlits для использования эволюционной информации [4]. Точно так же Портер+ рассматривает локальные структурные мотивы для предсказания торсионных углов. PaleAle, связанный с относительной доступностью растворителя (RSA), структурирован с двойными стеками BRNN-CNN в самой последней версии 5.0, превосходя эталонные показатели других методов прогнозирования RSA [5]. NetSurfP-2.0, объединяющий CNN и BRNN, был разработан в 2019 году.

Принятие во внимание других 1D-PSA наряду с вторичной структурой и учетом физико-химических свойств, а также информации об эволюции помогло повысить общую точность определения структуры. DESTRUCT многократно использовал каскадно-корреляционные нейронные сети как для вторичной структуры, так и для торсионных углов. Итерация состоит из пер-

вой FFNN, обученной прогнозировать вторичную структуру и двугранник ϕ, и фильтрации FFNN, последовательно вмешивающейся для преобразования прогнозов в новые значения. Группа Херста обновила DESTRUCT до DISSPred, который опирался на машину опорных векторов (SVM) и получил лучшую производительность. Этот метод также можно использовать для предсказания доступности растворителя и торсионного угла. SPIDER2 запустил ожидаемые множественные 1D-PSA — вторичную структуру, доступную для растворителя площадь поверхности (SASA) и торсионные углы — все сразу с тремя итерациями глубинных нейронных сетей. Его преемник, SPIDER3, улучшил производительность в целом, и теперь метод предсказывает четыре PSA одновременно, включая контактный номер с четырьмя итерациями для предсказания. ProteinUnet, опубликованный в 2020 году, обеспечивает такую же точность предсказания вторичной структуры, как и SPIDER3-single, но использует половинные параметры с 11-кратно более быстрым временем обучения [6]. Большинство обсуждаемых серверов и методов в последних версиях с более глубинными нейронными сетями и лучшими алгоритмами имеют показатель $Q3$ более 84%. Учитывая взрывной рост надежности оценки $Q3$ с помощью методов DL, может пройти не так много времени, прежде чем будет достигнут теоретический предел в 88–90%.

Один особый вид 1D-PSA нацелен на неупорядоченные участки белков. Многие белки содержат внутренне неупорядоченные области (Intrinsically Disordered Regions - IDR), которые обладают высокой гибкостью. Имея несколько доступных структур, IDR участвуют в сборке, передаче сигналов и многих генетических заболеваниях. Таким образом, этот PSA представляет особый интерес в дополнение к тому, что он является компонентом предсказания полной структуры белка. IDR были предсказаны с использованием статистических потенциалов, SVM или искусственных нейронных сетей. IUPred использует статистический парный потенциал, выраженный в виде матрицы 20×20, которая выражает общие предпочтения каждой контактирующей пары аминокислот.

Рассчитывается парный энергетический профиль и соответственно оценивается вероятность беспорядка. Метод DISOPRED3 сформулирован на SVM, модели машинного обучения с учителем, для различения упорядоченных и неупорядоченных областей. DISOPRED3 обучен по профилю PSI-BLAST, потому что он превосходит модели, обученные на отдельных последовательностях, демонстрируя улучшения, основанные на эволюционной информации. SPOT-Disorder2 предлагает прогнозирование нарушений по остаткам на основе глубинной нейронной сети, использующей клетки LSTM. Более высокая точность была достигнута за счет обновления его архитектуры с одной топологии LSTM, используемой в предыдущей версии, SPOT-Disorder, до ансамблевого набора гибридных моделей, состоящих из остаточных CNN с начальными путями, за которыми следуют слои LSTM.

Методы глубокого обучения для прогнозирования

Имея на руках информацию, полученную из 1D-PSA, может потребоваться 2D-PSA для полного построения трехмерной структуры белка. Недавние усилия по 2D-PSA сосредоточены на CM и мультиклассовых CM, которые выражают близость между парами остатков в белке. CM принимает бинарную двумерную матричную структуру $N \times N$, где N — длина белковой последовательности, оценивая каждую пару остатков как 1 (присутствие) или 0 (отсутствие) для матричных элементов на основе заданного пользователем порогового евклидова расстояния (а типичное значение составляет ~8 Å между атомами $C\alpha$). Мультиклассовый CM выражается в двумерной матрице, но элементы матрицы подробно квантуются, классифицируются более чем по двум состояниям. Важность этого CM для предсказания структуры белка прямо показана в оценках; раннее исследование показало, что можно собрать структурную модель в пределах 5 Å RMSD от нативной структуры, если известно $N/4$ белковых контактов дальнего действия, а другое исследование показало, что один контакт на двенадцать остатков обеспечивает надежное и точное моделирование белковой укладки.

Сама CM, безусловно, дает полезную информацию о пространственной организации данного белка, но следует отметить, что CM часто содержат транзитивный шум, возникающий из-за «косвенных» корреляций между остатками. Для устранения этого шума используются методы прямого корреляционного анализа, такие как взаимная информация (Mutual Information - MI), анализ прямого связывания (Direct Coupling Analysis - DCA) и оценка обратной ковариации с разреженным белком (Protein Sparse Inverse Covariance Estimation - PSICOV) [7]. DCA предполагает прямые ко-эволюционные связи между парами остатков в таблице MSA для выявления нативных внутридоменных и междоменных контактов между остатками в семействах белков.

Многие группы разработали предикторы CM, используя многоступенчатые глубинные нейронные сети. Ранее представленный сервер Distill также предоставляет предиктор CM под названием XX-Stout. Разработчики включили профиль контактной плотности в качестве промежуточного шага, используя другой модуль Distill, названный BrownAle. При расчете этого профиля контактной плотности главный собственный вектор значительно повысил общую производительность. DNCON воспользовались разработками графических процессоров для обучения значительно усиленных ансамблей предикторов контакта между остатками. MetaPSICOV — еще один предиктор CM, известный первым методом, использующим сигналы коэволюции из 1D-SA, извлеченные с помощью трех разных алгоритмов. Затем для вывода CM использовалась двухслойная нейронная сеть. Его последовательные версии, названные MetaPSICOV2 и DeepMetaPSICOV, существуют в которых ис-

пользуется более глубокая сетевая архитектура и модули ReLU. RaptorX-Contact из серии RaptorX использовал коэволюционные сигналы для повышения точности. RaptorX-Contact предсказывает свойства локальной структуры, матрицу контактов и расстояний, межостаточную ориентацию и третичную структуру белка с использованием сверхглубинной сверточной остаточной нейронной сети из первичной последовательности или множественного выравнивания последовательностей. DNCON2 реализован с шестью CNN и примененным коэволюционным сигналом от 1D PSA. Этот метод предсказывает CM с различными порогами расстояния 6, 7,5, 8, 8,5 и 10 Å, а затем уточняет их, чтобы оставить только 8 Å CM с улучшенной скоростью предсказания. TripletRes начинает со сбора MSA через базы данных последовательностей всего генома и метагенома, а затем строит три взаимодополняющие коэволюционные матрицы признаков (матрица ковариации, матрица точности и максимизация псевдовероятности) для создания моделей контактных карт посредством глубинного обучения остаточной сверточной нейронной сети. DeepContact также представляет собой подход на основе CNN, который обнаруживает коэволюционные мотивы и использует эти паттерны, чтобы обеспечить точный вывод вероятностей контактов. DeepCov использует полностью сверточные нейронные сети, работающие с данными о частоте или ковариации пар аминокислот, полученными непосредственно из выравнивания последовательностей, без использования глобальных статистических методов, таких как разреженная обратная ковариация или оценка псевдоправдоподобия. В отличие от других программ, для которых требуются сторонние программы, Pconsc4 представляет собой простой инструмент прогнозирования контактов, который не использует никаких внешних программ [8].

Заключение

В 2019 году, был разработан DeepCDPred, который включает в себя многоклассовый предиктор CM, использующий условия ограничения расстояния. Были использованы четыре модели на основе FFNN, чтобы выделить четыре класса диапазонов контактов: 0–8, 8–13, 13–18 и 18–23 Å. AlphaFold того же года генерирует наиболее детализированную мультиклассовую CM, 64 равных бина дистограммы (гистограмму расстояний) вдоль 2–22 Å, став передовым для этой области [9]. Архитектура глубинной двумерной расширенной сверточной остаточной сети с 220 остаточными блоками использовалась для предсказания карты расстояний в AlphaFold.

Список использованной литературы:

[1] Torrisi, M.; Pollastri, G.; Le, Q. (2020) Deep learning methods in protein structure prediction. Comput. Struct. Biotechnol. J., 18, 1301–1310.

[2] Cuff, J.A.; Barton, G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins Struct. Funct. Bioinform, 40, 502–511.

[3] Magnan, C.N.; Baldi, P. (2014) SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics, 30, 2592–2597.

[4] Torrisi, M.; Kaleel, M.; Pollastri, G. (2018) Porter 5: Fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. bioRxiv.

[5] Kaleel, M.; Torrisi, M.; Mooney, C.; Pollastri, G. (2019) PaleAle 5.0: Prediction of protein relative solvent accessibility by deep learning. Amino Acids, 51, 1289–1296.

[6] Kotowski, K.; Smolarczyk, T.; Roterman-Konieczna, I.; Stapor, K. (2021) ProteinUnet—An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. J. Comput. Chem., 42, 50–59.

[7] Bitbol, A.-F. (2018) Inferring interaction partners from protein sequences using mutual information. PLoS Comput. Biol., 14, e1006401.

[8] Michel, M.; Menendez Hurtado, D.; Elofsson, A. (2019) PconsC4: Fast, accurate and hassle-free contact predictions. Bioinformatics, 35, 2677–2679.

[9] 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. (2020) Improved protein structure prediction using potentials from deep learning. Nature, 577, 706–710.