ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

УДК 577.322.4

Структурные особенности каротиноид-связывающих белков

М.М. Сурков[1], А.Ю. Литовец[1], А.А. Мамчур[1], Т.Б. Станишнева-Коновалова[2], И.А. Ярошевич[1, *]

[1]Кафедра биофизики, биологический факультет, Московский государственный университет имени М.В. Ломоносова, Россия, 119234, г. Москва, Ленинские горы, д. 1, стр. 24;
[2]Кафедра биоинженерии, биологический факультет, Московский государственный университет имени М.В. Ломоносова, Россия, 119234, г. Москва, Ленинские горы, д. 1, стр. 73
[*]e-mail: iyapromo@gmail.com

Каротиноид-белковые комплексы участвуют в процессах фотосинтеза, фоторецепции, защиты от окислительного стресса, обмена веществ и пигментации. В данной работе проведен детальный анализ доступных структурных данных с атомарным разрешением каротиноид-содержащих белков. В ходе исследования проанализированы молекулярные особенности каротиноид-связывающих областей белков и структурные особенности связанных каротиноидов. Полученные результаты указывают на общие принципы организации белок-каротиноидных взаимодействий, необходимые для разработки новых подходов к их направленной модификации. Методом машинного обучения создана модель для прогноза каротиноид-связывающей активности по первичной структуре белка.

Ключевые слова: *каротиноиды, каротинопротеины, белок-лигандные взаимодействия*

DOI: 10.55959/MSU0137-0952-16-80-3S-13

Введение

Каротиноиды объединяются в самый обширный класс биологических пигментов, который в настоящий момент насчитывает более 1200 представителей, выделенных из семи сотен различных организмов [1]. В природе каротиноиды выполняют разнообразные биологические роли: это эффективные антиоксиданты [1], стабилизаторы фосфолипидных мембран [2], светосборщики и светофильтры [3]. В биотехнологии и медицине они нашли применение в качестве провитаминов [4], предшественников пахучих веществ [5], аллелохимикатов [6], противораковых агентов [7], агентов борьбы с множественной лекарственной устойчивостью [8]. Потенциал использования каротиноидов значительно расширяется при рассмотрении каротиноид-белковых комплексов, особое внимание среди которых уделяется каротинопротеинам – стехиометрическим комплексам каротиноидов и белков, последние характеризуются специфичностью связывания каротиноидов и наличием строго определенного сайта посадки лиганда. На основе таких объектов уже созданы биосовместимые фотопереключатели [9] и температурные сенсоры [10].

Каротиноид-белковые комплексы способны существенно модифицировать физико-химические свойства связанных молекул каротиноидов за счет межмолекулярных взаимодействий [11].

Огромное разнообразие химических структур каротиноидов и широкие возможности направленной модификации белков открывают заманчивые перспективы для синтетической биологии. Однако на данный момент отсутствует общепринятая теория, объясняющая каротиноид-связывающую активность белков. Разработка такой теории могла бы значительно ускорить прогресс в целенаправленной разработке каротинопротеинов и открыть новые возможности для практического применения таких молекул в медицине, сельском хозяйстве, биотехнологии и других сферах.

Предсказать способность белков связывать каротиноиды – сложная задача, поскольку универсальные структурные мотивы, характерные для сайтов связывания, до сих пор не определены. Однако достижения в области вычислительных методов, структурного анализа и понимания белок-лигандных взаимодействий открывают путь к более точным прогнозам. В этой работе на основании структурных данных о каротиноид-связывающих белках, депонированных в международную базу данных RCSB PDB (http://www.rcsb.org/) [12], проведен анализ закономерностей их строения, и с помощью методов машинного обучения создана модель для прогнозирования каротиноид-связывающей активности белков по их первичной структуре.

Материалы и методы

Анализ структур PDB. Для формирования набора данных выполнен систематический поиск в базе данных RCSB PDB по списку кодов лигандов, соответствующих каротиноидам (табл. SI1). Поиск проводился с использованием API RCSB Search. Для каждой найденной структуры автоматически собрана метаинформация, включая экспериментальный метод, разрешение, таксономические данные и сведения о первичной публикации. Структурные данные для последующего анализа загружены в формате PDBx/mmCIF.

Анализ аминокислотного окружения каротиноидов проводился для двух типов атомов: 1) атомы иононовых колец и их односвязных заместителей (для оценки специфичности вблизи функциональных групп) и 2) все тяжелые атомы лиганда (для оценки общего характера сайта связывания). Атомы шестичленных колец в структуре лиганда идентифицировались автоматически с помощью алгоритмов поиска циклов из библиотеки NetworkX 3.2.1. Аминокислотные остатки белка считались соседними к каротиноиду, если расстояние от любого их атома до соответствующей группы атомов лиганда не превышало 4,5 Å. Поиск соседей осуществлялся с помощью модуля NeighborSearch библиотеки Biopython 1.85.

Для оценки специфичности белкового кармана рассчитана частота встречаемости 20 стандартных аминокислот в окружении лиганда. Далее было вычислено логарифмическое обогащение (Log2) как отношение наблюдаемой частоты встречаемости аминокислоты к фоновой, опубликованной для базы данных Swiss-Prot [13]. Статистическая значимость обогащения/обеднения для каждой аминокислоты оценивалась с помощью точного биномиального теста с последующей поправкой на множественное тестирование по методу Бенджамини-Хочберга (FDR, False Discovery Rate). Общее отклонение аминокислотного состава окружения от фонового распределения проверялось с помощью критерия хи-квадрат (χ^2).

Для детального изучения геометрии π-стэкинговых взаимодействий был проведен анализ взаимного расположения иононовых колец каротиноидов и ароматических колец аминокислот (PHE, TYR, TRP). Рассчитывались два ключевых параметра: 1) расстояние между центроидами (геометрическими центрами) колец и 2) угол между нормалями к их плоскостям. Плоскости колец аппроксимировались методом главных компонент (Principal Component Analysis, PCA), где нормаль к плоскости соответствовала вектору, описывающему наименьшую дисперсию координат атомов кольца. Взаимодействия учитывались, если расстояние между центроидами не превышало 7,0 Å.

Все расчеты, обработка данных и статистический анализ были выполнены с использованием языка программирования Python 3 и библиотек NumPy, pandas, SciPy, NetworkX и Biopython. Для генерации двумерных графиков использовались библиотеки Matplotlib и Seaborn. Трехмерные визуализации сайтов связывания и конформаций лигандов создавались программно с помощью API PyMOL (The PyMOL Molecular Graphics System, Version 2.5, Schrödinger, LLC.).

Машинное обучение. В данной работе использовалась дообученная на последовательностях белков модель ProtBERT [14]. В качестве входных данных использована аминокислотная последовательность, в качестве выходных – массив 1024-мерных векторов – по вектору на каждую аминокислоту последовательности. На выходных массивах рассчитаны два вектора для белка в целом: с определением среднего значения и с определением максимального значения. Таким образом, для каждой последовательности получено два 1024-мерных вектора: один отражающий усредненный контекст белка, а другой – отражающий наиболее значимые локальные характеристики. Эти векторы затем использовались для обучения нашей модели.

Для набора позитивных примеров (каротиноид-связывающие цепи) выбраны структуры из Protein Data Bank с разрешенным каротиноидом из целевого списка. Из каждой структуры выделены полипептидные цепи. Цепь обозначалась как каротиноид-связывающая, если хотя бы одна аминокислота в ее составе расположена на расстоянии не более 4,5 Å от каротиноида. Цепи из этих структур, обозначенные как не связывающие каротиноид, использовались нами как «сложные» негативные примеры: зачастую они связывают хлорофиллы и другие гидрофобные молекулы, что хорошо для обучения модели отличать связывание каротиноида от связывания гидрофобных молекул в целом. Из полученного набора данных были исключены последовательности-дубликаты.

В качестве «простых» негативных примеров, необходимых для того, чтобы отличать белки с иной функцией, взяты последовательности белков крысы (*Rattus norvegicus*) из базы SwissProt [15]. Последовательности были кластеризованы с помощью программы CD-HIT [16] с порогом идентичности в 0,9 для дедупликации – из каждого кластера бралась только одна последовательность. Таких последовательностей получилось 3865.

Разделение набора данных на тренировочную и тестовую выборки проведено с учетом артефакта наличия близкородственных последовательностей [17]. Все используемые последовательности кластеризованы программой MMseqs2 [18] с порогом покрытия 0,3 и минимальной идентичностью 0,6, режим покрытия 1 (двунаправленное покрытие). Затем случайным образом отделена валидационная выборка размером 12% от общего количества кластеров. Остальная часть кластеров использова-

лась для тренировки модели. Таким образом, близкородственные последовательности не попадают одновременно и в валидационную выборку, и в выборку для обучения.

В качестве алгоритма машинного обучения был выбран классификатор на основе градиентного бустинга, реализованный в библиотеке CatBoost [19]. После обучения классификатора для исследования качества модели были рассчитаны метрики Precision, Recall, F1, ROC-AUC, также была построена ROC-кривая, precision-recall-F1-кривая. На основании метрики F1 при разных порогах классификатора был выбран наилучший порог.

Учитывая, что негативные примеры в нашем наборе белков могут быть частично неверно размечены из-за неспецифичной, неосновной каротиноид-связывающей активности белков, была рассчитана метрика Precision@k, где k = 50, 100, 150, что позволяет оценить ранжирующую способность модели.

Результаты и обсуждение

Анализ структур PDB. В базе данных PDB идентифицировано 626 уникальных структур, содержащих каротиноиды из целевого списка. Аминокислотные цепи разбиты на два типа: контактирующие и не контактирующие с каротиноидами (рис. 1А). Общее число идентифицированных цепей составило более пяти с половиной тысяч, уникальных белковых последовательностей – более двух тысяч (рис. 1Б). Предварительный анализ каротиноид-связывающих белков указывает на то, что зачастую это короткие (рис. SI1) последовательности, представленные одной альфа-спиралью в составе больших молекулярных суперкомплексов фотосинтетического аппарата. Анализ метаданных полученного списка указывает на значительное увеличение за последние пять лет темпов определения новых структур, содержащих каротиноиды, причем это ускорение полностью связано с применением методов крио-ЭМ (рис 1В).

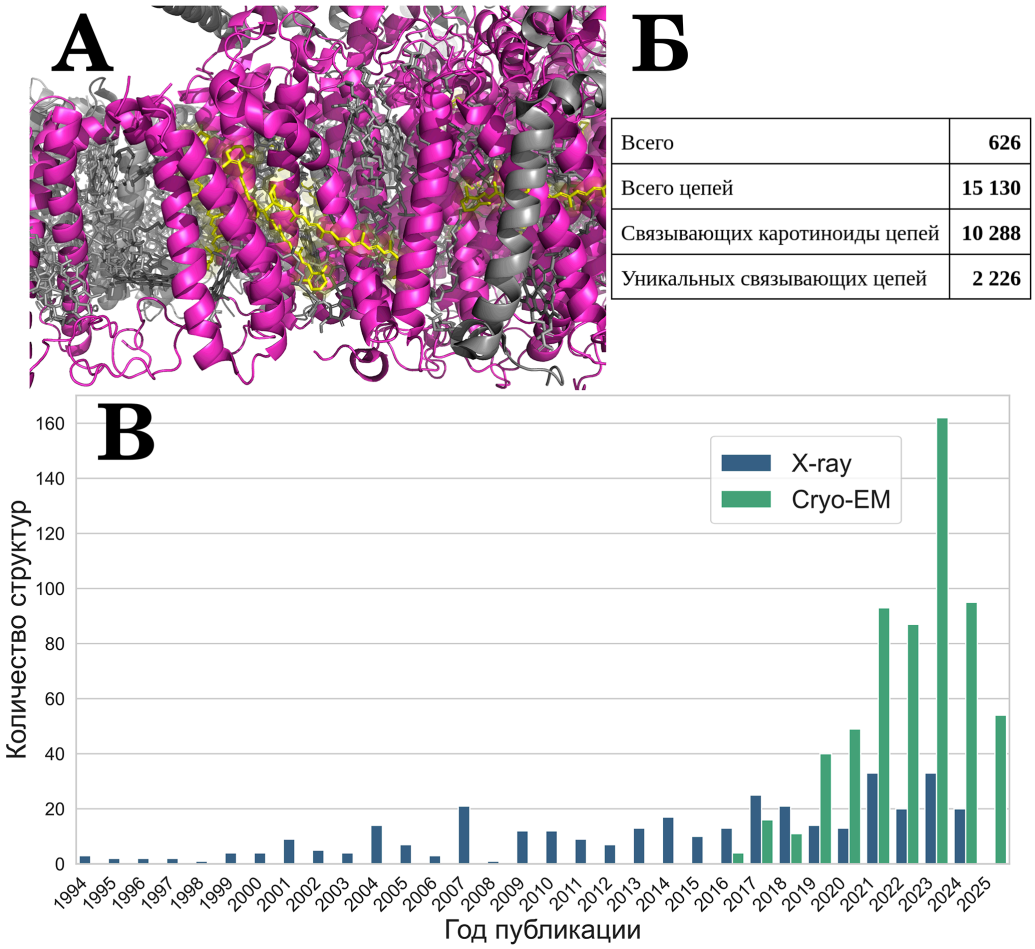

Рис. 1. Общий обзор анализируемых структур. **А** – Пример структуры, использованной в анализе. Изображен участок суперкомплекса фотосистемы II и светсобирающего комплекса II (PDB: 3JCU), желтым отмечены связанные молекулы каротиноида (BCR), фиолетовым – контактирующие с каротиноидами цепи, серым – цепи, не контактирующие с каротиноидами; **Б** – Число структур, «Всего» – число уникальных pdb id, «Всего цепей» – общее число аминокислотных цепей в списке определенных структур, «Связывающих каротиноиды цепей» – число связывающих каротиноиды цепей, «Уникальных связывающих цепей» – число уникальных связывающих каротиноиды цепей; **В** – Динамика публикаций структур каротиноид-связывающих белков по годам. Показано распределение по методам определения структуры: рентгеноструктурный анализ (X-ray, синий) и криоэлектронная микроскопия (Cryo-EM, зеленый).

Ввиду того, что подавляющее большинство учтенных структур содержат C40-циклические каротиноиды, которые аналогичны по своим размерам и конформационным свойствам, анализ объединяет структурные наблюдаемые всей выборки без разделения на отдельные классы в зависимости от лиганда. Для полученной выборки структур проведен детальный анализ аминокислотного окружения лигандов и геометрии их взаимодействия с ароматическими остатками. Для каротиноид-связывающих регионов характерен выраженный гидрофобный характер. Наблюдается сильное, статистически значимое обогащение окружения ароматическими (TRP, PHE) и алифатическими гидрофобными (LEU, ILE, VAL, ALA, MET) аминокислотами. Триптофан и фенилаланин стабильно занимают лидирующие позиции по степени обогащения (Log2(Набл./Ожид.) > 2), что указывает на их ключевую роль в формировании сайта связывания (рис. 2).

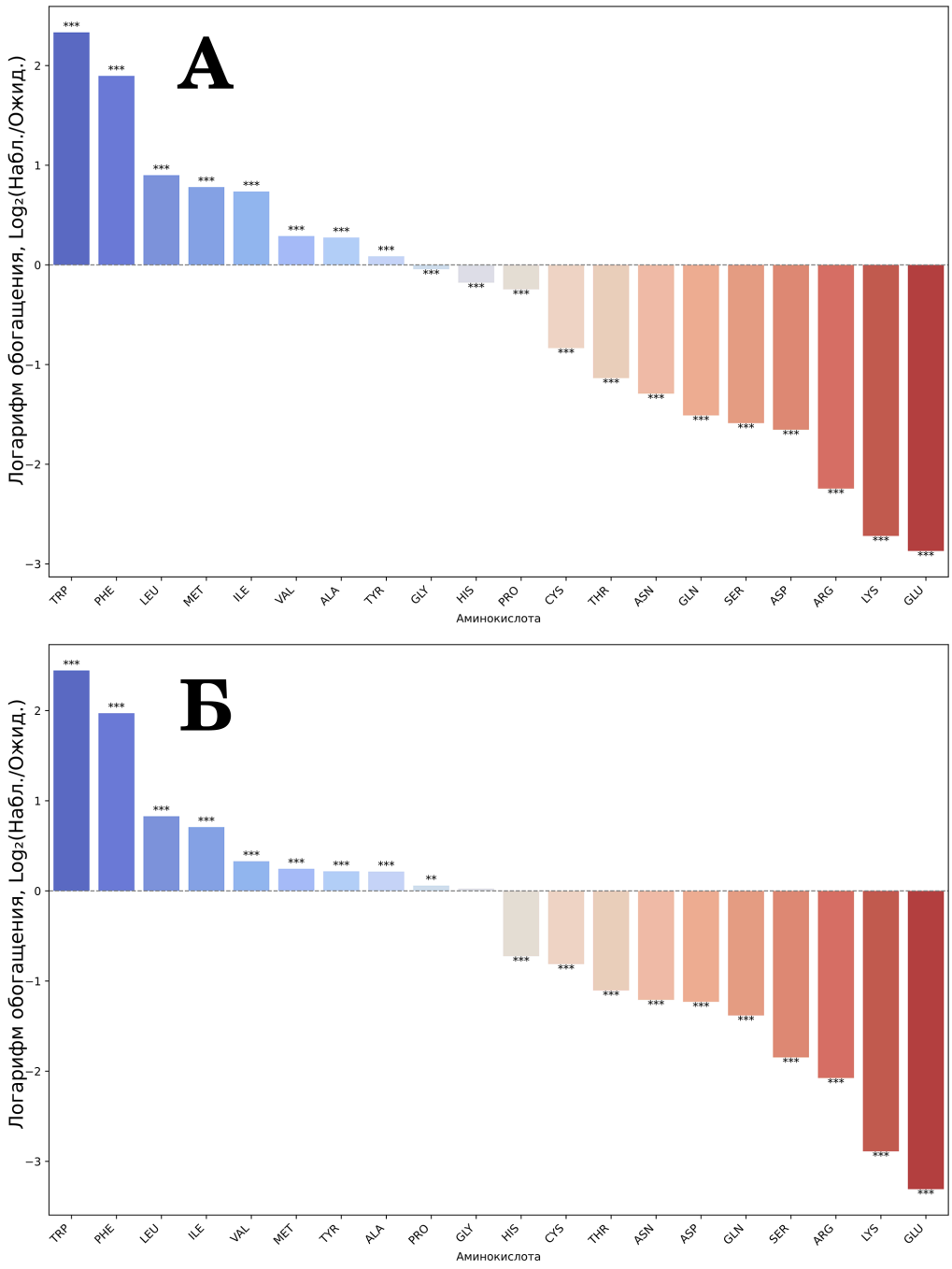

Рис. 2. Аминокислотное обогащение в сайтах связывания каротиноидов. Представлен логарифм отношения наблюдаемой частоты аминокислот в окружении (в пределах 4,5 Å) лиганда к ожидаемой (фоновой). Положительные значения (синие столбцы) указывают на обогащение, отрицательные (красные) – на обеднение. Символ «***» обозначает статистически значимое отклонение, при p-value < 0,001 (скорректированное по методу Бенджамини-Хохберга), символ «**» обозначает статистически значимое отклонение, при 0,01 ⩽ p-value < 0,05 (скорректированное по методу Бенджамини-Хохберга). **А** – обогащение в сайте связывания; **Б** – Аминокислотное обогащение в непосредственной близости (до 4,5 Å) от иононовых колец каротиноидов.

Одновременно с этим наблюдается драматическое обеднение окружения полярными и, в особенности, заряженными аминокислотами. Остатки лизина (LYS), глутамата (GLU) и аргинина (ARG) встречаются на порядки реже, чем ожидалось бы при случайном распределении (Log2(O/E) от -5 до −7), что соответствует в 32–128 раз более низкой частоте (рис. 2А).

Полученные данные свидетельствуют об универсальной стратегии связывания каротиноидов. Белок формирует гидрофобный участок, с которым контактирует значительная часть молекулы лиганда. Высокая частота ароматических остатков, вероятно, обусловлена их способностью формировать эффективные Ван-дер-Ваальсовы и π-стэкинговые взаимодействия с сопряженной системой лиганда, обеспечивая как аффинность, так и специфичность связывания.

Для более детального анализа специфичности взаимодействия был отдельно проанализирован состав окружения иононовых колец каротиноида (рис. 2Б), являющихся его наиболее функционально вариабельной частью. Сравнение с общим окружением (рис. 2А) показывает, что общие закономерности сохраняются, однако эффект для большинства аминокислот более выражен. Например, обогащение триптофаном и фенилаланином здесь еще сильнее. Интересное исключение представляют метионин (MET) и тирозин (TYR). TYR показывает умеренное, но значимое обогащение в окружении колец, в то время как его частота в окружении всей цепи ближе к фоновой. Вероятно, это связано со способностью тирозина формировать водородные связи с полярными группами ксантофиллов, которые обычно располагаются на кольцах, что невозможно для PHE и TRP. MET, в свою очередь, показывает высокое обогащение вдоль всей цепи, однако обогащение в области только колец снижено. Это может быть связано с наличием сульфидной группы, которая оказывается менее предпочтительной вблизи иононовых колец и их заместителей.

Анализ различных типов лигандов (рис. SI2) выявляет дополнительные закономерности. Например, для неполярного β-каротина (BCR) характерно наиболее сильное обеднение полярными остатками. В то же время для ксантофиллов, содержащих кето-группы (кантаксантин CRT, астаксантин A86), наблюдается менее выраженное обеднение полярными аминокислотами и даже некоторое обогащение тирозином и гистидином, способными образовывать водородные связи. Это указывает на то, что белки эволюционно адаптируют сайты связывания под конкретные химические свойства каротиноидов.

Для детального изучения геометрии взаимодействий между каротиноидом и ароматическими остатками был проведен анализ взаимного расположения иононового кольца каротиноида и ароматических колец аминокислот (PHE, TYR, TRP). Анализ распределения углов между плоскостями этих колец (рис. 3А) показал сильное смещение в сторону перпендикулярных ориентаций. Пик распределения находится в диапазоне 70–80°, что характерно для классических Т-образных взаимодействий. Параллельные (stacking) ориентации с углами < 30° встречаются крайне редко.

Распределение расстояний между центроидами колец (рис. 3Б) также оказалось нетривиальным. Основной пик наблюдается на относительно большом расстоянии, около 6,3–6,7 Å, со вторым плечом в районе 5,5–6,0 Å. Взаимодействия на близких расстояниях (< 4,5 Å), характерных для плотного параллельного стэкинга, практически отсутствуют.

Совокупность этих данных указывает на то, что классический стэкинг не является доминирующим мотивом взаимодействия. Вместо этого ароматические остатки чаще формируют стенки гидрофобного кармана, располагаясь перпендикулярно плоскости иононового кольца лиганда. Такое расположение позволяет максимизировать благоприятные контакты между C-H группами одного кольца и π-системой другого, что является ключевой особенностью Т-образных взаимодействий. Относительно большие расстояния между центроидами могут объясняться наличием объемных метильных групп на иононовом кольце, создающих стерические препятствия для плотного сближения.

При разделении данных по типам ароматических аминокислот (рис. 3В) выявляются интересные различия. Фенилаланин почти исключительно участвует в Т-образных взаимодействиях (> 60°). Тирозин демонстрирует удивительную неспецифичность: его кольцо может быть ориентировано практически под любым углом, с незначительным смещением в сторону перпендикулярных ориентаций. Триптофан показывает бимодальное поведение: он предпочитает Т-образные взаимодействия, но также образует небольшую, но заметную популяцию с параллельной ориентацией (~ 15–30°), которая практически отсутствует у PHE.

Это позволяет дифференцировать ароматические остатки по их более специфичным взаимодействиям. Фенилаланин выступает в роли «классического» строительного блока гидрофобного кармана. Уникальная гибкость тирозина, возможно, связана с его способностью формировать H-связи, что делает точную ориентацию π-системы менее критичной. Двойственный характер триптофана, самого большого ароматического остатка, позволяет ему эффективно участвовать как в формировании стенок кармана, так и, в отдельных случаях, образовывать классические стэкинг-взаимодействия, стабилизируя определенные конформации лиганда.

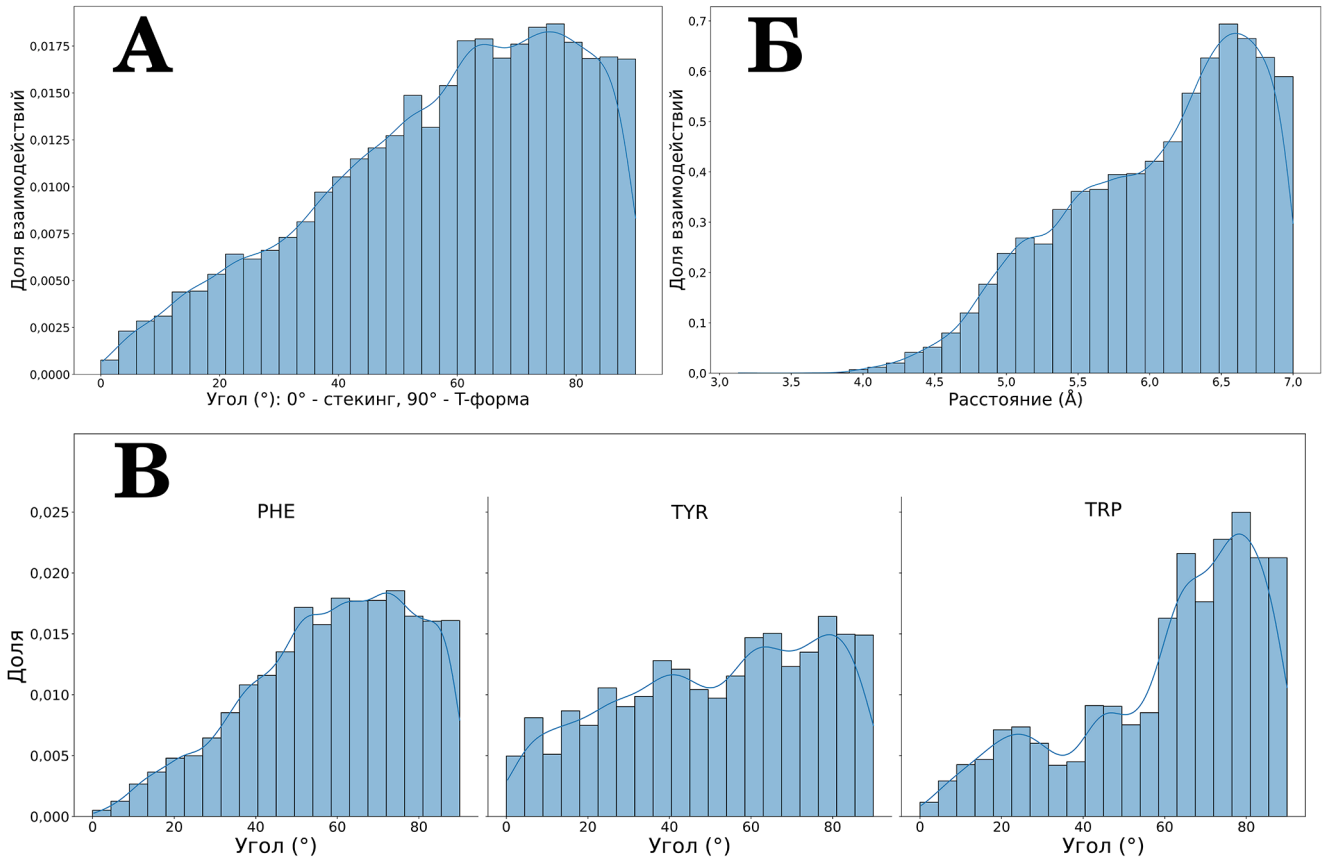

Рис. 3. Геометрические наблюдаемые относительного расположения каротиноида и аминокислотных остатков локального окружения проанализированных структур. **А** – Распределение углов между нормалями к плоскостям иононового кольца каротиноида и ароматического кольца аминокислоты. Угол 0° соответствует параллельной (stacking) ориентации, 90° – перпендикулярной (Т-образной); **Б** – Распределение расстояний между центроидами взаимодействующих иононового и ароматического колец; **В** – Распределение углов взаимодействия (аналогично А) в разрезе по типам ароматических аминокислот: фенилаланин (PHE), тирозин (TYR) и триптофан (TRP).

Проведенный анализ позволил сформулировать несколько общих принципов организации белок-каротиноидных взаимодействий. Во-первых, фундаментальным требованием является наличие гидрофобного кармана, исключающего полярные и заряженные аминокислотные остатки. Во-вторых, стенки этого кармана преимущественно сформированы ароматическими остатками (TRP, PHE), которые создают протяженный каркас для эффективных взаимодействий с сопряженной системой лиганда. В-третьих, геометрия этих взаимодействий в области иононовых колец преимущественно перпендикулярна (Т-образна), а не параллельна, что диктует строгие стерические требования к архитектуре сайта связывания. Наконец, в-четвертых, на фоне этих общих правил наблюдается принцип «тонкой настройки»: сайт связывания адаптируется под химические особенности конкретного лиганда, например, за счет введения способных к водородным связям остатков (таких как тирозин) для стабилизации полярных групп ксантофиллов. Совокупность этих принципов формирует структурную основу как для высокой аффинности, так и для специфичности каротиноид-белковых комплексов.

Машинное обучение. Подготовленный набор данных включает 6956 вхождений, из которых 1952 являются положительными – каротиноид-связывающими последовательностями. 906 вхождений были определены как тестовые (из них положительных – 271), остальные использовались для обучения.

На основе полученных из модели ProtBERT 1024-мерным векторам по каждой из аминокислот в последовательности были сформированы два 1024-мерных векторных признака для каждого белка (с вычислением среднего и с вычислением максимального значения по последовательности), что позволило получить для белков числовую репрезентацию, на которой обучена классификационная модель градиентного бустинга. Созданная модель оценивает первичную последовательность белка, выдавая для нее выходное значение от 0 до 1. Порог классификации (при достижении которого белок считается связывающим каротиноид) установлен на уровне 0,32, что соответствует максимальной метрике F1. Модель демонстрирует высокую производительность: F1 = 0,77, Precision = 0,72, Recall = 0,83, ROC-AUC = 0,92. Построенные ROC- и Precision-Recall-F1-кривые изображены на рис. SI3A и рис. SI3Б соответственно.

Таблица

Результаты работы созданной модели на ряде белков (с эмпирически известной активностью в отношении связывания каротиноидов): «–» – не связывает каротиноиды, «+» – неспецифическое связывание, «++» – специфическое связывание (основная функция)

Белок	Выходное значение модели	Эксперимент
альфа-казеин	0,13	+ [20]
бета-казеин	0,19	+ [20]
каппа-казеин	0,32	+ [20]
DBXN, цепь A	0,34	++ [22]
DBXN, цепь B	0,54	++ [22]
DBXN, цепь C	0,26	++ [22] *мало контактов с каротиноидом
Aster A	0,02	+ [21]
Изоцитратдегидрогеназа	0,05	–
Протромбин	0,01	–
Бета-2 адренорецептор	0,03	–
Гексокиназа-1	0,02	–

По рассчитанным метрикам Precision@k (Таблица SI2) видно, что наиболее вероятные по мнению модели каротиноид-связывающие белки действительно в большинстве своем проявляют такую активность. Это говорит о хорошей способности модели ранжировать белки по аффинности к каротиноидам.

Предсказательная сила модели была проверена на отдельном наборе белков: туда вошли альфа-, бета-, каппа-казеины – для них известна способность связывать каротиноиды [20]; недавно открытый новый каротиноид-связывающий белок дибилиноксантин (DBXN, PDB: 9KUE), белок млекопитающих Aster A, для которого также известна каротиноид-связывающая активность [21], а также некоторые белки домашнего хозяйства (таблица). Модель классифицирует как каротиноид-связывающий каппа-казеин и цепи A, B дибилиноксантина при пороге в 0,32. Каппа-казеин действительно описан как казеин с самой сильной аффинностью к бета-каротину среди белков того же семейства; цепи A и B дибилиноксантина имеют больше всего контактов с каротиноидом в структуре. Белки Aster A и альфа, бета-казеины не определены моделью как каротиноид-связывающие. Альфа-, бета- казеины связывают бета-каротин хуже, чем каппа- – это согласуется с выходными значениями из модели. Возможно, механизм связывания или структурные особенности Aster A сильно отличаются от большинства белков, на которых обучалась модель. Белки домашнего хозяйства правильно определены как не связывающие каротиноиды.

Для улучшения модели необходимо сфокусироваться на нескольких ключевых направлениях: включение большего числа каротиноид-связывающих белков с различными механизмами связывания и из разных семейств (таких, как Aster A и другие); снижение «шумности» негативных примеров; а также включение структурных признаков помимо первичной последовательности белка. Более того, примененные методы агрегации векторных представлений могут быть недостаточно информативными – для преодоления данного недостатка возможно применение механизмов внимания.

В результате нами разработан универсальный подход для быстрой оценки каротиноид-связывающей активности белка по его аминокислотной последовательности. В основе подхода лежит способ числовой репрезентации первичной структуры белка с использованием модели ProtBERT. Числовая репрезентация позволяет использовать классические методы машинного обучения, в нашем случае градиентный бустинг, для классификации последовательности на предмет связывания каротиноидов. Хотя текущая версия модели сталкивается с проблемами при распознавании трудных негативных примеров и некоторых истинных положительных примеров, модель обладает предсказательной силой (верно предсказано наличие аффинности к каротиноидам у каппа-казеина и каротиноид-связывающего белка дибилиноксантина). К тому же, модель способна ранжировать белки по степени каротиноид-связывающей активности. Полученный результат закладывает основу для создания инструментов для разметки белков по аффинности к каротиноидам.

Исследование выполнено за счет гранта Российского научного фонда № 25-74-20001 (https://rscf.ru/project/25-74-20001/). Работа проведена без использования животных и без привлечения людей в качестве испытуемых. Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Yabuzaki J. Carotenoids Database: structures, chemical fingerprints and distribution among organisms. *Database*. 2017;2017:bax004.

2. Havaux M. Carotenoids as membrane stabilizers in chloroplasts. *Trends Plant Sci*. 1998;3(4):147–151.

3. Yaroshevich I.A., Krasilnikov P.M., Rubin A.B. Functional interpretation of the role of cyclic carotenoids in photosynthetic antennas via quantum chemical calculations. *Comput. Theor. Chem*. 2015;1070:27–32.

4. Tanumihardjo S.A., Palacios N., Pixley K.V. Provitamin A carotenoid bioavailability: What really matters? *Int. J. Vitam. Nutr. Res*. 2010;80(45):336–350.

5. Winterhalter P., Rouseff R.L., Eds. *Carotenoid-derived aroma compounds*. American Chemical Society; Distributed by Oxford University Press; 2002. 323 pp.

6. Sasamoto H., Suzuki S., Mardani-Korrani H., Sasamoto Y., Fujii Y. Allelopathic activities of three carotenoids, neoxanthin, crocin and β-carotene, assayed using protoplast co-culture method with digital image analysis. *Plant Biotechnol*. 2021;38(1):101–107.

7. Sharoni Y., Danilenko M., Walfisch S., Amir H., Nahum A., Ben-Dor A., Hirsch K., Khanin M., Steiner M., Agemy L., Zango G., Levy J. Role of gene regulation in the anticancer activity of carotenoids. *Pure Appl. Chem*. 2002;74(8):1469–1477.

8. Eid S.Y., El-Readi M.Z., Wink M. Carotenoids reverse multidrug resistance in cancer cells by interfering with ABC-transporters. *Phytomedicine*. 2012;19(11):977–987.

9. Piccinini L., Iacopino S., Cazzaniga S., Ballottari M., Giuntoli B., Licausi F. A synthetic switch based on orange carotenoid protein to control blue–green light responses in chloroplasts. *Plant Physiol*. 2022;189(2):1153–1168.

10. Maksimov E.G., Yaroshevich I.A., Tsoraev G.V., Sluchanko N.N., Slutskaya E.A., Shamborant O.G., Bobik T.V., Friedrich T., Stepanov A.V. A genetically encoded fluorescent temperature sensor derived from the photoactive Orange Carotenoid Protein. *Sci. Rep*. 2019;9(1):8937.

11. Britton G., Helliwell J.R. Carotenoid-Protein Interactions. *Carotenoids*. Eds. G. Britton, S. Liaaen-Jensen, and H. Pfander. Basel: Birkhäuser Basel; 2008:99–118.

12. Berman H.M. The Protein data bank. *Nucleic Acids Res*. 2000;28(1):235–242.

13. Shen S., Kai B., Ruan J., Torin Huzil J., Carpenter E., Tuszynski J.A. Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences. *Phys. A: Stat. Mech. Appl*. 2006;370(2):651–662.

14. Elnaggar A., Heinzinger M., Dallago C., Rehawi G., Wang Y., Jones L., Gibbs T., Feher T., Angerer C., Steinegger M., Bhowmik D., Rost B. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022;44(10):7112–7127.

15. The UniProt Consortium, Bateman A., Martin M.J., et al. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res*. 2025;53(D1):D609–D617.

16. Li W., Jaroszewski L., Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 2001;17(3):282–283.

17. Teufel F., Gíslason M.H., Almagro Armenteros J.J., Johansen A.R., Winther O., Nielsen H. GraphPart: homology partitioning for biological sequence analysis. *NAR Genom. Bioinform*. 2023;5(4):lqad088.

18. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol*. 2017;35(11):1026–1028.

19. Dorogush A.V., Ershov V., Gulin A. CatBoost: gradient boosting with categorical features support [Электронный ресурс]. arXiv. 2018. URL: https://arxiv.org/abs/1810.11363 (дата обращения: 06.07.2025).

20. Shahidi F., Dissanayaka C.S. Binding of carotenoids to proteins: a review. *J. Food Bioact*. 2023;13–28.

21. Bandara S., Ramkumar S., Imanishi S., Thomas L.D., Sawant O.B., Imanishi Y., Von Lintig J. Aster proteins mediate carotenoid transport in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A*. 2022;119(15):e2200068119.

22. Egorkin N.A., Aleksin A.M., Sedlov I.A., Zhiganov N.I., Bodunova D.V., Varfolomeeva L.A., Slonimskiy Y.B., Ziganshin R.H., Popov V.O., Boyko K.M., Vassilevski A.A., Maksimov E.G., Sluchanko N.N. A green dichromophoric protein enabling foliage mimicry in arthropods. *Proc. Natl. Acad. Sci. U.S.A*. 2025;122(23):e2502567122.

Поступила в редакцию 30.06.2025
После доработки 05.09.2025
Принята в печать 08.09.2025

RESEARCH ARTICLE

Structural features of carotenoid-binding proteins

M.M. Surkov[1], A.Yu. Litovets[1], A.A. Mamchur[1], T.B. Stanishneva-Konovalova[2], I.A. Yaroshevich[1,*]

[1]Department of Biophysics, Faculty of Biology, Lomonosov Moscow State University, 1–24 Leninskie gory, 119234, Moscow, Russia;
[2]Department of Bioengineering, Faculty of Biology, Lomonosov Moscow State University, 1–73 Leninskie gory, 119234, Moscow, Russia
[*]e-mail: iyapromo@gmail.com

Carotenoid-protein complexes play a crucial role in photosynthesis, photoreception, protection against oxidative stress, metabolism, and pigmentation. This study conducts a detailed analysis of structural data with atomic resolution for carotenoid-containing proteins. The research examines molecular features of carotenoid-binding regions and structural characteristics of bound carotenoids. The findings reveal general principles of the protein-carotenoid interface organization, essential for developing new approaches to targeted modification. Additionally, a machine learning model is created to predict carotenoid-binding activity based on the primary protein structure.

Keywords: *carotenoids, carotenoproteins, protein-ligand interactions*

Funding: This work was supported by the Russian Science Foundation (project no. 25-74-20001).

Сведения об авторах

Сурков Макар Максимович – студент магистратуры кафедры биофизики биологического факультета МГУ. Тел.: 8-495-939-11-16; e-mail: macsurmak.m02@mail.ru; ORCID: https://orcid.org/0009-0001-5688-8753

Литовец Андрей Юрьевич – студент кафедры биофизики биологического факультета МГУ. Тел.: 8-495-939-11-16; e-mail: litovetsay@my.msu.ru; ORCID: https://orcid.org/0009-0000-5329-9796

Мамчур Александра Александровна – аспирант кафедры биофизики биологического факультета МГУ. Тел.: 8-495-939-11-16; e-mail: al.mam4ur@yandex.ru; ORCID: https://orcid.org/0000-0002-6025-7663

Станишнева-Коновалова Татьяна Борисовна – канд. биол. наук, ст. науч. сотр. кафедры биоинженерии биологического факультета МГУ. Тел.: 8-495-939-57-38; e-mail: stanishneva-konovalova@mail.bio.msu.ru; ORCID: https://orcid.org/0000-0002-8427-8178

Ярошевич Игорь Александрович – канд. биол. наук., ст. науч. сотр. кафедры биофизики биологического факультета МГУ. Тел.: 8-495-939-11-16; e-mail: iyapromo@gmail.com; ORCID: https://orcid.org/0000-0002-8525-5568