На правах рукописи

Матвеев Евгений Викторович

Предсказание сайтов протеолиза на основе информации о пространственной структуре потенциальных субстратов

Специальность 1.5.8 —

«Математическая биология, биоинформатика»

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата биологических наук

Москва — 2025

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научный руководитель:	**Казанов Марат Джамалудинович** кандидат технических наук, зав. сектором (Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук)
Официальные оппоненты:	**Головин Андрей Викторович** Доктор химических наук, профессор Факультета биоинженерии и биоинформатики Московского государственного университета имени М. В. Ломоносова **Попцова Мария Сергеевна** Кандидат физико-математических наук, доцент Факультета компьютерных наук / Департамента больших данных и информационного поиска Национального Исследовательского Университета «Высшая Школа Экономики».
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта».

Защита состоится 29 сентября 2025 г. в 17:00 на заседании диссертационного совета 24.1.101.01 при Институте проблем передачи информации имени А.А. Харкевича Российской академии наук (ИППИ РАН) по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН, а также на сайте ИППИ РАН по адресу: http://iitp.ru/upload/content/1731/Matveev%20EV%20dissertation.pdf

Автореферат разослан «_____» _____ 2025 г.

Учёный секретарь
диссертационного совета
24.1.101.01,
доктор биологических наук Казенников Олег Васильевич

Общая характеристика работы

Актуальность темы исследования. Протеолиз — это процесс расщепления белков, который регулирует активность множества молекул и играет ключевую роль в самых различных биологических процессах как внутри клетки, так и за её пределами. Протеазы активируют и инактивируют ферменты, цитокины, гормоны, факторы роста, влияют на статус рецепторов (агонист-антагонист) и определяют локализацию молекул. Протеолиз контролирует такие процессы, как репликация и транскрипция ДНК, прогрессия клеточного цикла, клеточная дифференциация, морфогенез и ремоделирование тканей, ангиогенез, свёртывание крови, гемостаз, некроз, апоптоз. Нарушение активности протеаз связано с развитием многих заболеваний, включая злокачественные новообразования, артрит, сердечно-сосудистые и нейродегенеративные заболевания. Настоящее исследование посвящено разработке и применению метода идентификации субстратов протеолитических ферментов и определения сайтов расщеплений в молекулах субстратов. Эта задача остаётся одной из ключевых проблем современной протеомики, поскольку знание протеаз и их мишеней имеет значение для биомедицинских исследований и разработки новых терапевтических подходов. Для понимания функциональной роли протеолитических ферментов в контексте биологических процессов принципиально важно знание их естественных субстратов, а также точное определение позиций расщепляемых пептидных связей в молекулах субстратов. Такая информация позволяет охарактеризовать субстратную специфичность протеаз и пролить свет на молекулярные механизмы их действия. Существует множество экспериментальных методов, направленных на идентификацию субстратов и определение позиций сайтов расщеплений, а также способных оценивать динамику и масштабы протеолитических событий. Однако идентификация субстратов в условиях эксперимента является крайне трудоёмким процессом. Во-первых, протеазы функционируют в клетке не изолированно, а в рамках сложных сетей протеолитических взаимодействий. Во-вторых, на практике чрезвычайно трудно выделить продукт протеолиза из клеточной смеси белков и пептидов и точно идентифицировать его исходный субстрат. Эти трудности, в сочетании с высокой стоимостью и сложностью проведения соответствующих экспериментов, подчёркивают актуальность разработки биоинформатических инструментов предсказания потенциальных субстратов и сайтов протеолитического расщепления как эффективной альтернативы трудоёмким лабораторным методам.

Степень разработанности темы исследования. На сегодняшний день доступен ряд биоинформатических инструментов, предназначенных для прогнозирования протеолитических событий. Большинство из них разработано на основе данных об известных событиях протеолиза и использует эту информацию для моделирования специфичности протеолитических ферментов. Как правило, такие модели являются протеазо-специфическими, то есть, они ориентированы на моделирование специфичности одной конкретной протеазы или ограниченного набора протеаз из одного семейства. Это существенно сужает область их применения. Кроме того, существующие подходы используют различные характеристики субстратов, прежде всего, аминокислотную последовательность и, реже, структурные особенности. Однако на сегодняшний день не разработано ни одного инструмента, способного абстрагироваться от специфичности конкретной протеазы и моделировать универсальные закономерности протеолиза, общие для всех или большинства протеаз. При этом известно, что именно структурные свойства субстратов являются универсальными детерминантами протеолиза, что открывает перспективу создания инструмента такого типа.

Целью исследования являлась разработка подхода к идентификации субстратов протеаз и определению сайтов расщепления, основанного на сочетании универсальной модели, оценивающей восприимчивость участков белков к протеолизу на основе структурных характеристик субстратов (далее — структурной модели), и протеазо-специфических моделей, отражающих предпочтения отдельных протеолитических ферментов к аминокислотному окружению расщепляемой пептидной связи.

Для достижения цели исследования были поставлены следующие **задачи**:

1. Разработать модель предсказания подверженности участков белка к протеолизу на основании информации о его трёхмерной структуре:
    a. Сформировать набор протеолитических событий из базы экспериментально верифицированных сайтов расщеплений CutDB.
    b. Идентифицировать трёхмерные структуры субстратов в базе PDB с фильтрацией по степени сходства и длине перекрывания с аминокислотной последовательностью субстрата.
    c. Картировать сайты расщепления на трёхмерные структуры субстратов с использованием выравнивания последовательностей субстратов и структур.

d. Выполнить фильтрацию протеолитических сайтов, исключив сайты, которые были получены в экспериментах с возможной денатурацией белка.
e. Определить набор ключевых структурных характеристик, потенциально влияющих на восприимчивость пептидных связей белка к протеолизу.
f. Апробировать различные алгоритмы машинного обучения и выбрать оптимальный метод предсказания восприимчивости участков белка к протеолизу.
g. Подобрать оптимальное соотношение размеров положительного и отрицательного классов в обучающей выборке.
h. Выполнить сравнительный анализ оценок восприимчивости к протеолизу для сайтов, зафиксированных в физиологических и лабораторных условиях.
i. Оценить влияние расширения обучающего множества за счёт моделей трёхмерных структур из базы AlphaFoldDB на качество предсказания.

2. Разработать модели специфичности по аминокислотной последовательности для максимально возможного количества протеаз человека.

3. Разработать подход к объединению предсказаний структурной модели и моделей специфичности протеаз по последовательности, а также сравнить качество его работы с существующими методами прогнозирования.

4. Использовать разработанный метод для изучения протеолитической активации S-гликопротеина оболочки коронавируса SARS-CoV-2:

a. Идентифицировать протеазы человека, потенциально способные участвовать в протеолитической активации S-гликопротеина.
b. Предсказать сайт расщепления в S-белке для катепсина L.
c. Оценить влияние мутаций в S-гликопротеине на эффективность протеолитического расщепления.

Научная новизна. В настоящем исследовании впервые разработана универсальная предсказательная модель, позволяющая оценивать подверженность различных участков белков протеолизу на основании их трёхмерной структуры. Модель применима ко всем типам протеолитических ферментов и не ограничена конкретными классами протеаз. Кроме этого, впервые разработаны модели субстратной специфичности для 169 протеаз человека; функционал метода может быть расширен за счёт добавления

моделей специфичности для других протеаз. Впервые предсказана позиция сайта расщепления S-белка коронавируса SARS-CoV-2 цистеиновой протеазой катепсин L, а также впервые получены оценки влияния как мутаций, обнаруженных в штаммах коронавируса SARS-CoV-2, так и теоретически возможных аминокислотных замен, на эффективность расщепления S-белка протеазами человека.

Теоретическая и практическая значимость работы. С теоретической точки зрения, разработанный метод идентификации протеолитических субстратов и сайтов расщепления позволяет выявлять ранее неизвестные элементы протеолитических сетей, что расширяет современные представления об участии протеолитических ферментов в биологических процессах в клетке.

С практической точки зрения, потенциальные субстраты и продукты протеолитического расщепления могут рассматриваться в качестве лекарственных мишеней, биомаркёров и биосенсоров, продуктов биотехнологической и пищевой отраслей промышленности, инструментов биохимического анализа. Наличие информации о субстратном профиле протеаз может способствовать разработке новых ингибиторов и созданию искусственных протеолитических систем с заданными свойствами. В целом, предложенный биоинформатический подход может служить эффективным дополнением к экспериментальным исследованиям, позволяя сузить область поиска и повысить их направленность, снижая затраты времени и ресурсов.

Методология и методы исследования. Для достижения цели исследования применялись следующие методы и инструменты: биоинформатические методы использовались для анализа и визуализации трёхмерных структур белков, а также для выравнивания аминокислотных последовательностей; для автоматизации поиска и обработки больших объёмов данных, визуализации статистических результатов был разработан программный код на языках Python и R; математические алгоритмы и статистические методы применялись для построения структурной модели и протеазо-специфических моделей, а также для разработки метода прогнозирования протеолитических событий.

Положения, выносимые на защиту:

1. Разработана модель предсказания подверженности участков белка к протеолизу на основании информации о его трёхмерной структуре.

2. Разработаны модели специфичности по аминокислотной последовательности для 169 протеаз человека, представленные в виде позиционно-весовых матриц (PWM).

3. Предложен подход интеграции структурной модели с моделями специфичности по последовательности и показано, что объединённая модель демонстрирует качество предсказания, сопоставимое с таковым для существующих методов, при этом охватывая больший спектр протеаз и обладая гибкостью и расширяемостью за счёт возможности добавления новых PWM-моделей.

4. Разработанный метод применён для анализа протеолитической активации S-гликопротеина коронавируса SARS-CoV-2. Качество работы метода подтверждает корректное предсказание двух наиболее изученных сайтов расщепления S-белка — S1/S2 и S2`. Установлено, что представители четырёх семейств сериновых протеаз — PCSK, TTSP, калликреины и факторы свёртывания крови — потенциально способны расщеплять указанные сайты в случае колокализации с вирусным белком.

5. В позиции K790 S-гликопротеина коронавируса SARS-CoV-2 предсказан ранее неизвестный протеолитический сайт катепсина L — цистеиновой протеазы, активной в эндосомах. Предполагается, что расщепление S-белка в данном сайте способствует проникновению коронавируса внутрь клетки из эндосомы. Пространственный анализ показал, что данный сайт располагается вблизи сайта S2` и пептида слияния — двух функционально значимых элементов S-белка, обеспечивающих слияние вирусной оболочки и клеточной мембраны.

Степень достоверности и апробация результатов. Обучение и валидация структурной модели выполнялись на основе данных о протеолитических событиях, полученных из базы данных CutDB. Протеазо-специфические модели, описывающие субстратную специфичность протеаз человека, были построены с использованием информации о подтверждённых сайтах расщепления из базы данных MEROPS. Для валидации протеазо-специфических моделей использовались независимые данные из базы MEROPS, не включённые в обучающие выборки, а также известные сайты расщепления в S-белке коронавируса SARS-CoV-2. Сравнение с существующими методами предсказания протеолитических событий проводилось на независимом наборе данных, полученном в эксперименте по инкубированию матриксных металлопротеаз с белками *E.coli*. Результаты настоящего исследования были представлены на международных и российских конференциях в виде устных и стендовых докладов: в 2021 году

— “ASBMB Experimental Biology”, “ASBMB Serine proteases in pericellular proteolysis and signaling”, “Pacific Symposium on Biocomputing (PSB) ”, “XXVII Симпозиум Биоинформатика и компьютерное конструирование лекарств”; в 2022 году — “XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery”; в 2023 году — “11-я Московская конференция по вычислительной молекулярной биологии (MCCMB)”, “ASBMB Serine Proteases in Pericellular Proteolysis and Signaling Virtual Conference”.

Личный вклад автора. Все результаты были получены лично автором настоящего исследования либо при его непосредственном участии. В частности, автором была выполнена работа по поиску и анализу литературы по теме исследования, сбору и обработке данных, построению моделей, визуализации результатов. Обсуждение и интерпретация результатов, а также подготовка публикаций и тезисов конференций осуществлялись автором совместно с научным руководителем и соавторами.

Объём и структура работы. Диссертация изложена на 121 странице и включает введение, три главы, заключение и выводы. Работа содержит 32 рисунка и 2 формулы. Список литературы содержит 142 наименования.

Основное содержание работы

В **Главе I** представлен обзор литературы по теме диссертационного исследования. Она содержит 5 разделов.

В **разделе 1.1** изложены фундаментальные аспекты протеолиза.

В **разделе 1.2** обсуждается общепринятая классификация протеолитических ферментов, основанная на особенностях расположения расщепляемой пептидной связи, строении активного центра и структурной организации фермента, механизме катализа, локализации в клетке, специфичности, а также представленности среди живых организмов.

Раздел 1.3 посвящён известным механизмам протеолитической активации вирусных гликопротеинов и влиянию эффективности протеолиза на инфекционность и патогенность вирусов. В частности, рассматривается участие различных протеаз человека в активации гликопротеинов оболочек ортомиксовирусов, парамиксовирусов, ретровирусов, филовирусов, флавивирусов, папилломавирусов и коронавирусов.

В **разделе 1.4** рассматривается механизм протеолитической активации S-гликопротеина коронавируса SARS-CoV-2. Представлены общие сведения о жизненном цикле вируса, структуре S-гликопротеина его оболочки, клеточных рецепторах, обеспечивающих прикрепление вируса к мембране клетки. Описано, как происходит протеолиз S-гликопротеина SARS-CoV-2 и как мутации в S-белке влияют на его чувствительность к протеолитическому расщеплению.

В **разделе 1.5** описаны существующие методы прогнозирования протеолитических событий. Особое внимание уделено важности моделирования специфичности ферментов и учёта различных характеристик субстратов при построении прогностических моделей. Также приведен обзор баз данных и инструментов прогнозирования, которые уже доступны для использования.

Глава II посвящена разработанному методу прогнозирования протеолитических событий. Она состоит из трёх разделов.

В **разделе 2.1** описываются основные этапы построения модели предсказания уязвимости участков белка к протеолизу на основе информации о структурных особенностях субстратов.

Для построения структурной модели был подготовлен набор данных, включающий экспериментально подтверждённые события протеолиза из базы данных CutDB. Составление набора данных включало поиск известных трёхмерных структур субстратов в базе данных PDB, картирование сайтов расщеплений на структуры субстратов и их фильтрацию, формирование информативных структурных признаков субстратов, таких как, доступность для растворителя, тип вторичной структуры, температурный фактор, длина петлевых участков и принадлежность к терминальным участкам белка (Рис. 1). Значения структурных признаков были рассчитаны для каждого аминокислотного остатка субстрата, подразумевая, что расщепляемая пептидная связь располагается между рассматриваемым и последующим остатками (между субсайтами P1 и P1` на Рис. 2), поскольку ранее было показано, что уязвимость пептидной связи к протеолизу главным образом определяется структурными характеристиками субсайта P1.

Итоговый набор данных был представлен в виде таблицы, включающей набор признаков, характеризующих различные структурные особенности субстратов, и целевую бинарную переменную, отражающую позицию

расщепления в субстрате. Данная таблица содержала 69285 наблюдений, соответствующих пептидным связям в 190 трёхмерных структурах субстратов, из которых 445 представляли собой сайты расщепления 130-ти протеаз.

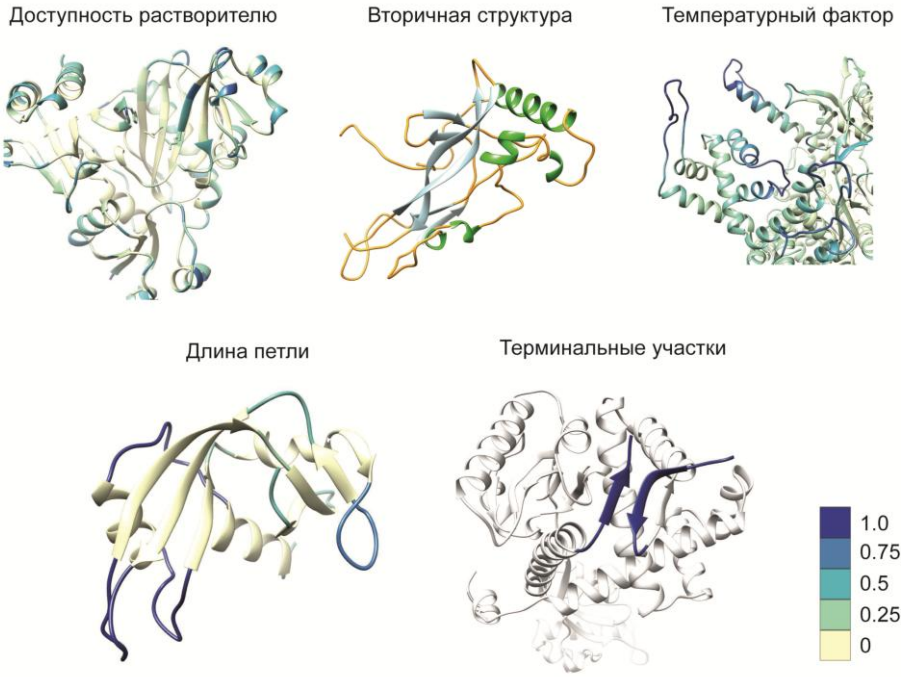

Рисунок 1. Визуализация признаков, использованных для обучения предсказательной модели, на трёхмерных структурах субстратов. Цветовая шкала из пяти градаций с диапазоном значений от 0 до 1 отображает нормализованные значения структурных признаков «Доступность растворителю», «Температурный фактор» и «Длина петли». Цветовая схема вторичной структуры: зелёный — альфа-спирали, оранжевый — петли и неупорядоченные области, светло-голубой — бета-стрэнды.

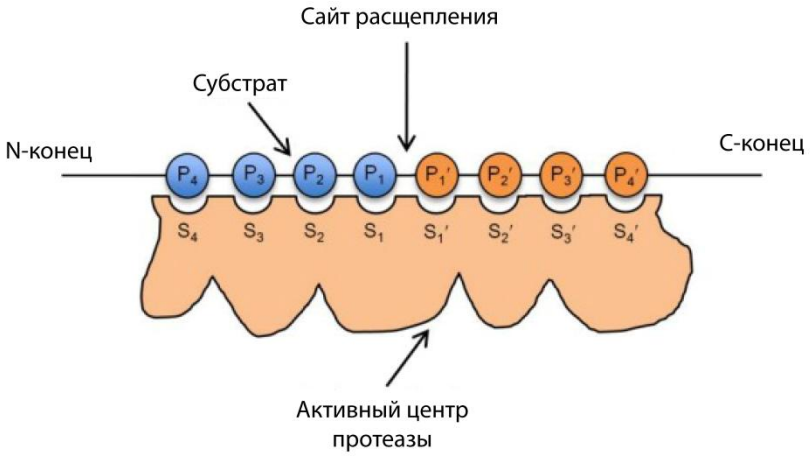

Рисунок 2. Представление сайта расщепления при взаимодействии активного центра протеазы с субстратом. Аминокислотные остатки субстрата обозначаются “P”, соответствующие субсайты активного центра протеазы — “S”. Нумерация позиций задаётся относительно сайта расщепления. Рисунок адаптирован из doi: 10.1142/s0219720011005288.

При построении структурной модели были опробованы различные алгоритмы машинного обучения, такие как случайный лес (Random Forest), дерево решений (Decision Tree), наивный байесовский классификатор (Naive Bayes), метод опорных векторов (SVM), логистическая регрессия (Logistic Regression), градиентный бустинг (XGBoost), линейный дискриминантный анализ (LDA), квадратический дискриминантный анализ (QDA), алгоритм поиска ближайшего соседа (KNN). Тестирование моделей осуществлялось методом 10-блочной стратифицированной кросс-валидации. Метрикой качества прогнозирования модели была выбрана площадь под ROC-кривой (ROC AUC).

В связи с выраженным дисбалансом классов в исходном наборе данных (445 положительных и 68840 отрицательных наблюдений), была проведена оценка влияния различных соотношений размеров классов на качество модели, что позволило определить оптимальное соотношение размеров классов для формирования финальной обучающей выборки. Согласно полученным результатам (Рис. 3), качество прогнозирования практически не зависело от соотношения классов в обучающей выборке, поэтому для оптимизации вычислительных затрат итоговая модель была обучена на сбалансированной выборке с соотношением положительных и отрицательных наблюдений 1:1. Наилучшие показатели качества прогнозирования (ROC AUC ≈ 0,7) продемонстрировала модель, построенная с использованием алгоритма линейного дискриминантного анализа.

Разработанная предсказательная модель позволяет оценивать восприимчивость участков потенциального субстрата к протеолизу на основании его структурных характеристик. Оценки восприимчивости присваиваются каждой пептидной связи белка и варьируются в интервале от 0 до 1 (Рис. 4). Анализ оценок восприимчивости к протеолитическому расщеплению на трёхмерных структурах субстратов показал, что модель присваивает более высокие оценки участкам, локализованным в выступающих петлях, в областях с высокими значениями температурного фактора, на подвижных N- и C-концах полипептидной цепи, а также в областях, доступных для растворителя, что согласуется с данными предыдущих исследований. Оценки умеренной восприимчивости к протеолизу наблюдались преимущественно на внешних поверхностях альфа-спиралей и в коротких петлях, тогда как участки, расположенные в гидрофобном ядре белка, демонстрировали низкие оценки восприимчивости к протеолитическому расщеплению.

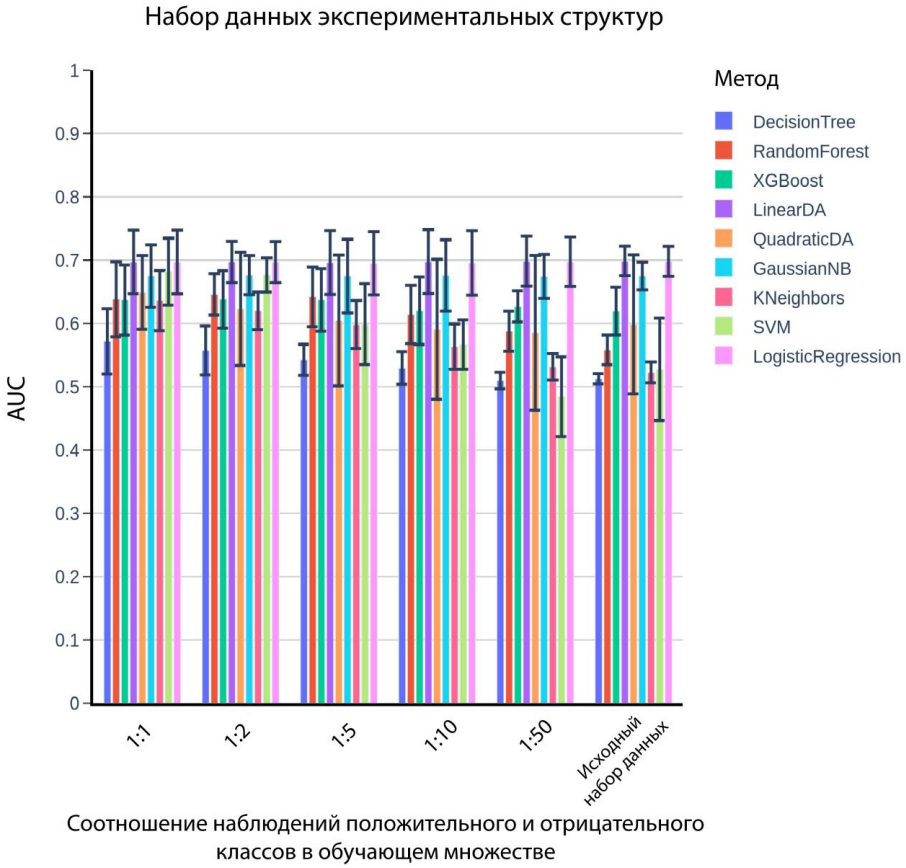

Рисунок 3. Зависимость качества предсказания модели от соотношения классов в обучающей выборке. В качестве метрики использовалась площадь под ROC-кривой (ROC AUC). Для построения моделей применялись различные алгоритмы машинного обучения. Лучшее качество прогнозирования на обучающей выборке с равным соотношением классов продемонстрировала модель, основанная на алгоритме линейного дискриминантного анализа.

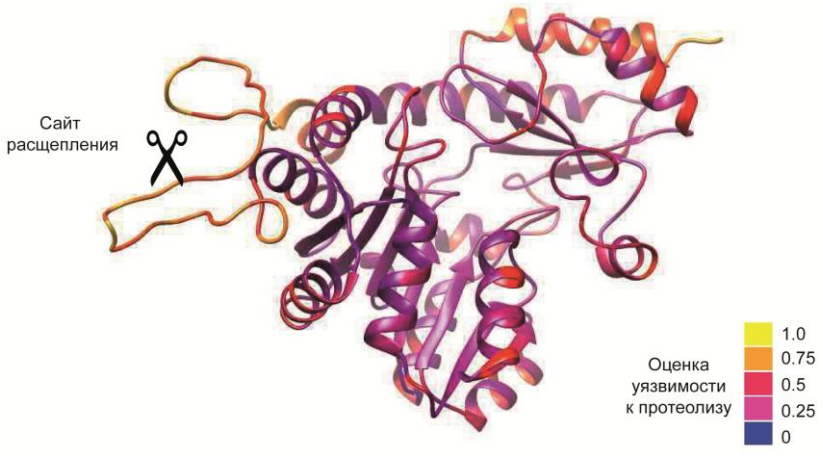

Рисунок 4. Визуализация предсказаний восприимчивости к протеолизу пептидных связей, полученных с помощью разработанной модели, на трёхмерной структуре субстрата.

Данные экспериментов свидетельствуют о том, что событие протеолитического расщепления зависит, в частности, от времени инкубации протеазы с субстратом. Таким образом, разработанная предсказательная

модель не позволяет однозначно прогнозировать, произойдёт ли протеолитическое расщепление в конкретной позиции, однако способна с высокой степенью достоверности оценить, какие пептидные связи в белке являются наиболее уязвимыми к расщеплению. Предполагая, что сайты расщепления в белках эволюционировали таким образом, чтобы обеспечивать достаточную восприимчивость к протеолизу, в данной работе была проанализирована возможность определения порогового значения для оценок уязвимости к протеолизу, генерируемых моделью. Для этого были проанализированы распределения оценок восприимчивости к протеолизу, соответствующие сайтам расщепления *in vivo* (физиологическим) и *in vitro* (определённым в лабораторном эксперименте). Согласно полученным результатам (Рис. 5), надёжно отличить сайты протеолиза, соответствующие событиям, происходящим в живой клетке, от регистрируемых в лабораторных условиях, крайне затруднительно: распределения оценок уязвимости участков белка, полученные структурной моделью, в значительной степени перекрываются (точность классификации = 0,622), несмотря на то, что тест Вилкоксона показал, что различия между наблюдаемыми распределениями являются статистически значимыми (p-value = $1{,}43 \times 10^{-5}$).

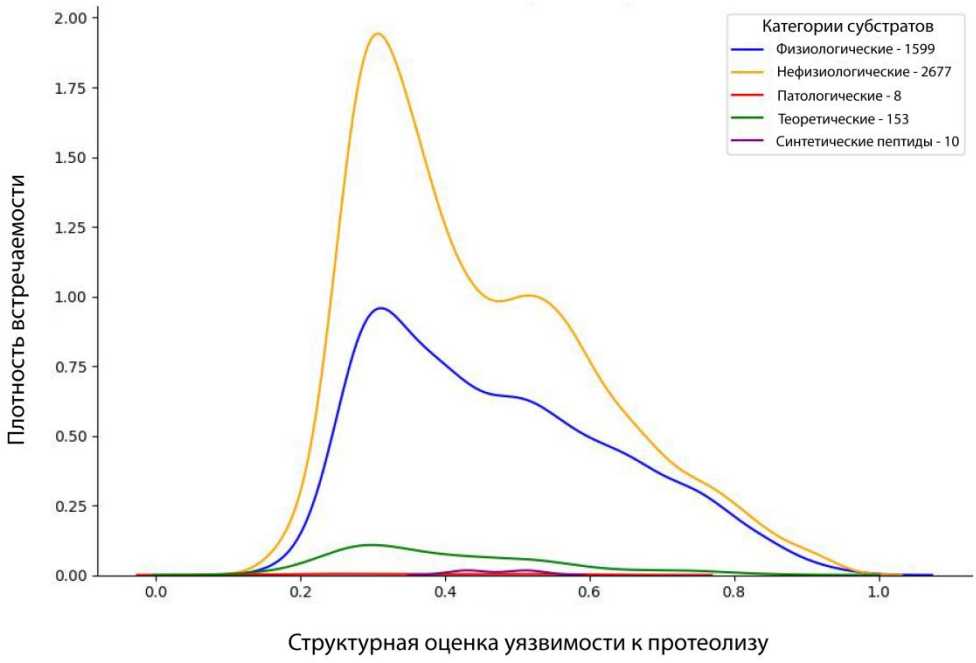

Рисунок 5. Распределение оценок восприимчивости к протеолизу, сгенерированных структурной моделью, для различных категорий протеолитических расщеплений.

В **Разделе 2.2** показано, что качество прогнозирования структурной модели можно повысить за счёт расширения обучающей выборки моделями структур AlphaFold.

Для построения модифицированной предсказательной модели были использованы те же протеолитические события и субстраты из базы данных CutDB, что и при обучении исходной предсказательной модели. Поиск моделей трёхмерных структур субстратов в базе данных AlphaFoldDB осуществлялся с помощью алгоритма BLAST. Остальные этапы подготовки набора данных были такими же, как и при построении исходной модели. Использование моделей трёхмерных структур белков из базы данных AlphaFoldDB позволило увеличить число положительных наблюдений в обучающей выборке более чем в шесть раз (2918 против 445).

Для построения модифицированной предсказательной модели на расширенном наборе данных использовались те же алгоритмы машинного обучения, что и при построении исходной модели предсказания уязвимости участков белка к протеолизу. Подход к тестированию модели и метрика оценки качества (ROC AUC) также оставались прежними. Единственным отличием построения модифицированной модели от построения её исходной версии стало отсутствие температурного фактора в расширенном наборе данных, поскольку этот параметр недоступен в структурах, предсказанных AlphaFold. Вместо него модельные структуры содержат уникальную характеристику — оценку достоверности предсказанной структуры — анализ предсказательной способности которой показал, что она обладает предсказательной силой. Таким образом, модифицированная модель была обучена на меньшем числе структурных признаков, чем исходная модель.

Согласно полученным результатам (Рис. 6), как и в случае с исходной моделью, качество прогнозирования модифицированной модели не зависело от соотношения классов в обучающей выборке, поэтому соотношение количества положительных и отрицательных наблюдений в итоговой обучающей выборке расширенного набора данных также составило 1:1. Наилучшие показатели точности прогнозирования среди протестированных алгоритмов продемонстрировала модель, обученная с использованием алгоритма XGBoost.

Для корректного сравнения качества прогнозирования исходной и модифицированной предсказательных моделей необходимо было обеспечить использование идентичных наборов структурных признаков субстратов, поскольку в противном случае было бы невозможно однозначно определить, связано ли изменение качества модели с расширением обучающего множества или с различием в используемых признаках. По этой причине исходная модель, построенная на основе характеристик экспериментально

определённых трёхмерных структур субстратов, была повторно обучена на том же наборе данных, но с исключением температурного фактора — признака, отсутствующего в моделях структур из базы AlphaFoldDB. Результаты сравнения показали, что модифицированная модель, обученная на расширенной выборке с включением структур из AlphaFoldDB, продемонстрировала более высокое качество прогнозирования по сравнению с моделью, основанной исключительно на экспериментальных структурах. Разница между медианными значениями метрики ROC AUC между двумя моделями составила 0,05 (Рис. 7).

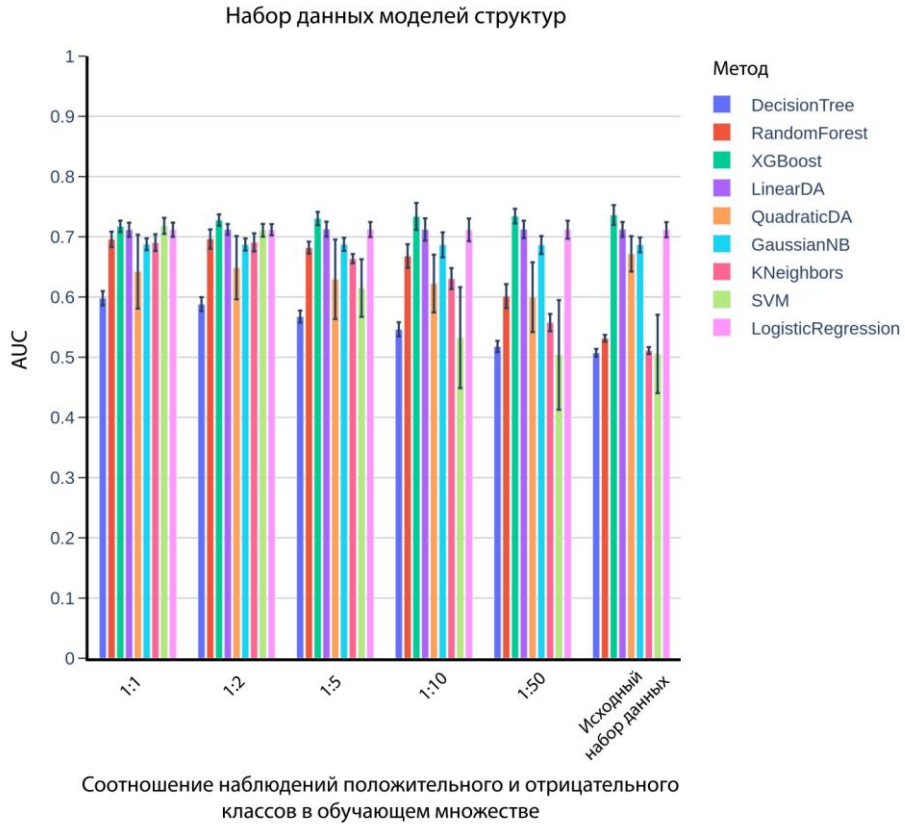

Рисунок 6. Оценка качества прогнозирования модифицированной модели в зависимости от соотношения классов в обучающей выборке. В качестве метрики использовалась площадь под ROC-кривой (ROC AUC). Для построения моделей применялись различные алгоритмы машинного обучения. Лучшее качество прогнозирования на обучающей выборке с равным соотношением классов продемонстрировала модель, основанная на алгоритме XGBoost.

Таким образом, были разработаны две версии моделей предсказания уязвимости участков белка к протеолизу на основе известной трёхмерной структуры: одна использует в качестве входных данных экспериментально определённую структуру, тогда как другая — модель трёхмерной структуры, сгенерированную методом AlphaFold.

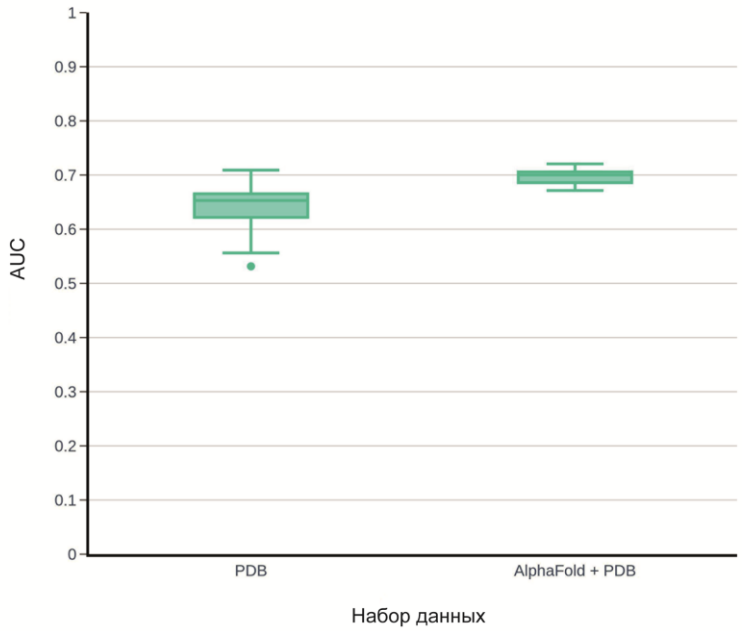

Рисунок 7. Улучшение качества прогнозирования модифицированной модели при расширении обучающего набора данных за счёт включения моделей трёхмерных структур из базы данных AlphaFoldDB.

В **разделе 2.3** описаны интеграция структурной модели предсказания уязвимости участков белков к протеолизу с моделями специфичности протеаз по последовательности и результаты сравнения разработанного метода прогнозирования протеолитических событий с существующими методами.

В настоящем исследовании следующим этапом развития модели стало включение информации о специфичностях протеаз по аминокислотной последовательности с целью повышения точности предсказаний, основанных на пространственных характеристиках субстратов. Наиболее распространённым подходом к моделированию специфичности протеаз является использование позиционно-весовых матриц (PWM). Интеграция структурной модели с PWM-моделями специфичности протеаз позволяет создать консолидированный подход, учитывающий как пространственные особенности субстрата, так и предпочтения протеазы к определённым аминокислотным окружениям. Кроме того, такой подход позволяет напрямую сравнивать результирующую модель с существующими методами (Рис. 8).

Для построения интегративной модели был использован независимый набор данных, предоставленный лабораторией J. Smith (Burnham Institute), включающий информацию о 249 уникальных протеолитических событиях,

относящихся к двум матриксным металлопротеазам человека — MMP9 и MMP25. В качестве субстратов использовались 25 белков *E.coli*, для которых доступны трёхмерные структуры в базе PDB. Из набора данных были исключены структуры, схожие со структурами субстратов из обучающей выборки, использованной при построении исходной структурной модели. В конечном наборе данных оказалось 174 протеолитических расщепления, соответствующих 24 уникальным структурам субстратов.

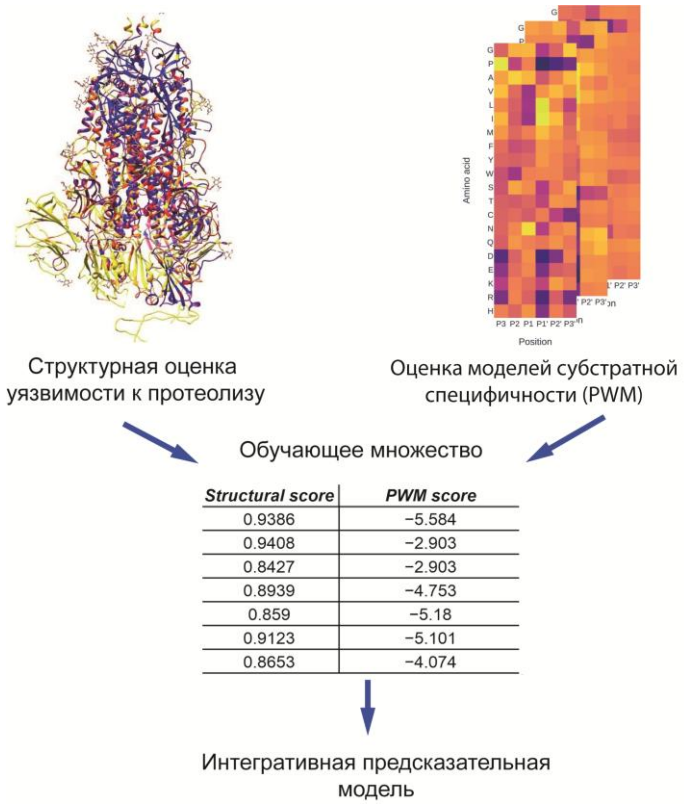

Рисунок 8. Схематическое представление объединения результатов предсказаний структурной моделью и протеазо-специфическими моделями, описывающими субстратную специфичность конкретных протеаз.

К полученному набору данных были применены как структурная модель, так и модели специфичности протеаз MMP9 и MMP25, полученные в раннем исследовании[1]. Результирующий набор предсказаний включал оценки восприимчивости к протеолизу, сгенерированные для каждой пептидной связи рассматриваемых белков двумя независимыми моделями — структурной моделью и моделью специфичности по последовательности.

[1] *Fedonin, G. G.* Predictive Models of Protease Specificity based on Quantitative Protease-Activity Profiling Data / G. G. Fedonin, A. Eroshkin, P. Cieplak, et al. // Biochim Biophys Acta Proteins Proteom. — 2019. — Vol. 1867(11). — P. 140253. — doi: 10.1016/j.bbapap.2019.07.006

Для объединения этих двух наборов предсказаний в итоговую модель были апробированы различные алгоритмы машинного обучения: метод k-ближайших соседей (KNN), метод опорных векторов с линейной и радиальной базисными функциями (Linear SVM и RBF SVM), алгоритм гауссовского процесса (Gaussian Process), дерево решений (Decision Tree), случайный лес (Random Forest), наивный байесовский классификатор (Naïve Bayes), квадратический дискриминантный анализ (QDA), нейронная сеть (Neural Net) и градиентный бустинг (AdaBoost). Тестирование модели проводилось на сбалансированной выборке с равным соотношением наблюдений положительного и отрицательного классов. В качестве метрики качества прогнозирования использовалась площадь под ROC-кривой (ROC AUC).

Согласно полученным результатам, наилучшую интеграцию предсказаний структурной моделью и моделями специфичности по последовательности обеспечили алгоритмы наивного байесовского классификатора и квадратического дискриминантного анализа. Относительно высокие показатели значения ROC AUC, продемонстрированные моделями, основанными на алгоритмах случайного леса, ближайшего соседа и градиентного бустинга, вероятнее всего обусловлены эффектом переобучения, что особенно заметно по двумерным визуализациям результатов (Рис. 9). Интегративная модель, объединяющая оценки, выдаваемые структурной моделью и моделями специфичности по последовательности, была построена на основе алгоритма наивного байесовского классификатора.

Для разработанного подхода к прогнозированию протеолитических событий было выполнено сравнение с одним из передовых инструментов — Procleave, который использует комплексную информацию о субстратах, включая информацию о структуре и аминокислотной последовательности. Однако, поскольку Procleave является протеазо-специфическим методом, его применимость ограничена 27 протеазами, в числе которых матриксные металлопротеазы и катепсины. Для корректного сравнения эффективности разработанного метода с Procleave был сформирован независимый набор данных для тестирования из базы данных MEROPS, путем исключения протеолитических событий, ранее включенных в обучающие и тестовые выборки как разработанной в ходе настоящего исследования модели, так и метода Procleave. Итоговый набор для тестирования включал 43 позиции расщепления, соответствующие 28 субстратам и 3 протеазам — S01.269,

C14.003 и M10.003. Для полученного набора данных были сгенерированы предсказания восприимчивости к протеолизу двумя компонентами разработанного метода: а) структурной моделью и б) PWM-моделями, отражающими специфичность протеаз по аминокислотной последовательности. PWM-модели для протеаз были разработаны с помощью информации о сайтах расщепления данных протеаз, полученной из базы данных MEROPS.

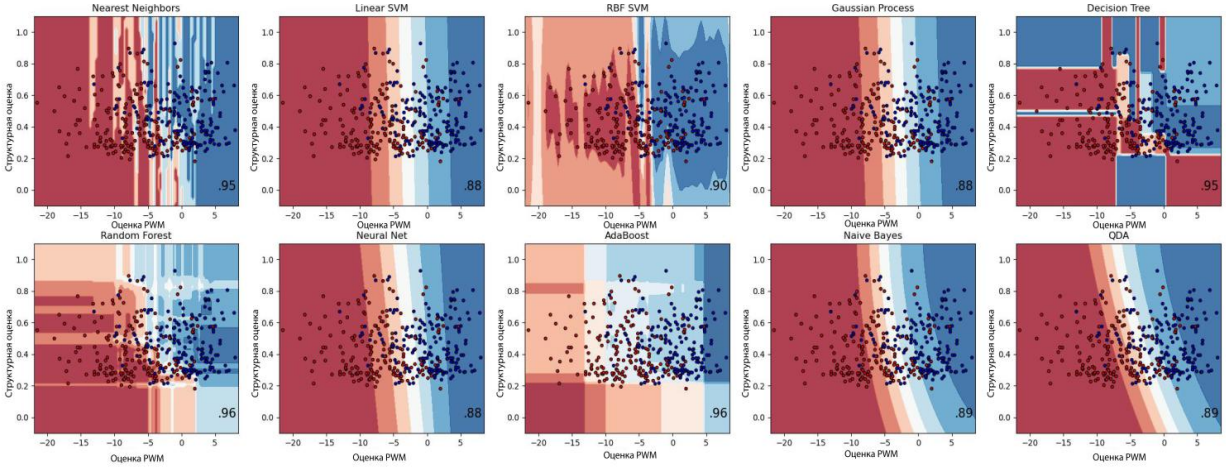

Рисунок 9. Двумерные визуализации предсказаний интегративной модели для наблюдений различных классов при использовании различных алгоритмов машинного обучения. Значения метрики ROC AUC указаны в правом нижнем углу каждой визуализации. Цветовой градиент от красного к синему отражает оценки восприимчивости к протеолизу, сгенерированные моделью: от меньших значений к большим.

Для всех пептидных связей рассматриваемых субстратов были получены оценки подверженности протеолизу, сгенерированные инструментом Procleave. Далее предсказания, полученные интегративным методом, объединяющим предсказания структурной модели и моделей специфичности протеаз по последовательности, были сопоставлены с результатами Procleave. Сравнение качества прогнозирования проводилось с использованием метрики площадь под ROC-кривой (ROC AUC). Согласно полученным результатам, интегративный метод продемонстрировал более высокое качество прогнозирования по сравнению с Procleave: средние значения ROC AUC составили 0,962 и 0,937, соответственно, а медианные значения — 0,97 и 0,966. Однако статистический анализ с использованием критерия Вилкоксона показал, что наблюдаемая разница не является статистически значимой (Рис. 10).

Раздел 2.4 обобщает основные результаты, представленные в главе II.

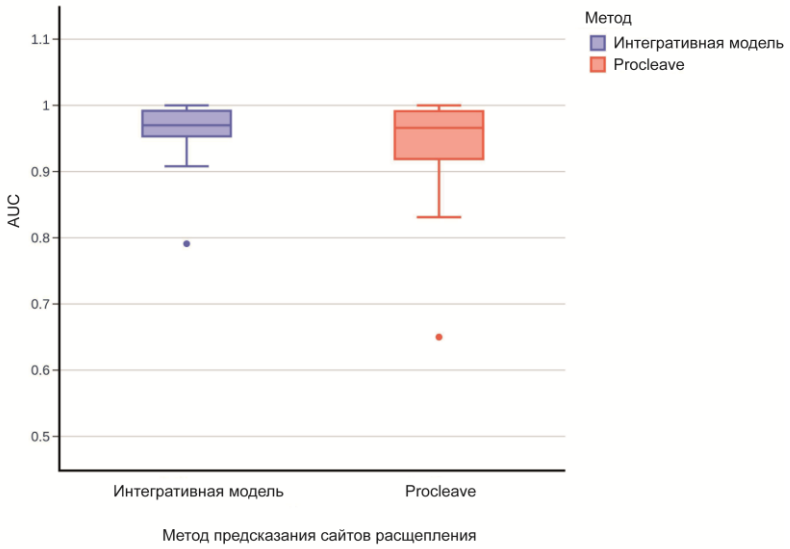

Рисунок 10. Сравнение качества прогнозирования между разработанным интегративным методом, объединяющим предсказания структурной модели и модели специфичности по последовательности, и методом Procleave.

Глава III посвящена применению разработанного метода для анализа протеолитической активацией S-гликопротеина оболочки коронавируса SARS-CoV-2. Она состоит из 6 разделов.

В **разделе 3.1** рассматриваются ключевые нерешённые вопросы, связанные с протеолитической активацией S-гликопротеина оболочки коронавируса SARS-CoV-2:

1. Какие протеазы, помимо известных фурина, TMPRSS2 и катепсина L, способны активировать S-белок коронавируса SARS-CoV-2?
2. Какую позицию в полипептидной цепи S-гликопротеина расщепляет катепсин L при эндоцитарном механизме проникновения коронавируса в клетку?

Раздел 3.2 посвящён разработке и валидации моделей субстратной специфичности протеаз человека, а также применению этих моделей для идентификации протеаз, способных активировать S-гликопротеин расщеплением известных сайтов S1/S2 (сайт расщепления фурином) и S2` (сайт расщепления TMPRSS2).

Модели специфичности были разработаны в виде позиционно-весовых матриц (PWM) на основе информации об аминокислотных последовательностях известных субстратов протеаз человека, полученной из

базы данных MEROPS. В результате были построены модели специфичности по аминокислотной последовательности для 169 протеаз человека.

Для того чтобы использовать модели специфичности с целью идентификации протеаз человека, способных активировать S-гликопротеин коронавируса SARS-CoV-2, была проведена двухэтапная валидация разработанных моделей. На первом этапе было выполнено сравнение распределений оценок расщепления, сгенерированных моделями, для двух выборок: известных позиций протеолитических расщеплений, полученных из базы данных MEROPS, и такого же количества случайных позиций. Каждая выборка содержала по 38841 позиции. Согласно полученным результатам (Рис. 11), первый и третий квартили распределения оценок расщепления для сайтов протеолиза составили 1.64 и 5.02, соответственно, среднее — 3.36, медиана — 3.16. Для случайных позиций первый и третий квартили составили –5.89 и –1.03, соответственно, среднее — –3.57, медиана — –3.42. Эти данные позволяют интерпретировать шкалу оценок, получаемых моделями субстратной специфичности.

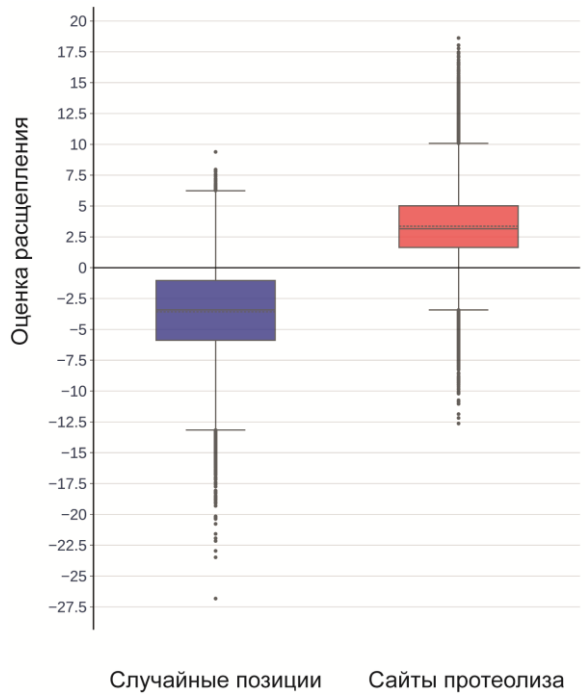

Рисунок 11. Распределение оценок расщепления, рассчитанных с помощью моделей субстратной специфичности протеаз, для известных сайтов расщеплений и случайных позиций в белках.

На втором этапе валидации разработанные модели были применены к аминокислотной последовательности S-белка исходного уханьского штамма коронавируса SARS-CoV-2 (UniProtKB ID: P0DTC2). Целью было определить, какие протеазы демонстрируют наивысшие оценки в позициях

известных сайтов расщепления — S1/S2 (R685) и S2` (R815). Для повышения биологической обоснованности анализа были дополнительно учтены данные о клеточной локализации (база данных COMPARTMENTS) и о профилях экспрессии протеаз в тканях (база данных TISSUES).

Согласно полученным результатам, наивысшую оценку расщепления в позиции R685 показала модель субстратной специфичности фурина (Рис. 12), а в позиции R815 — модель TMPRSS2 (Рис. 13). Эти результаты полностью согласуются с текущими представлениями о ключевой роли фурина и TMPRSS2 в активации S1/S2 и S2` сайтов, соответственно. Таким образом, валидация моделей субстратной специфичности подтвердила их точность и применимость в дальнейших исследованиях.

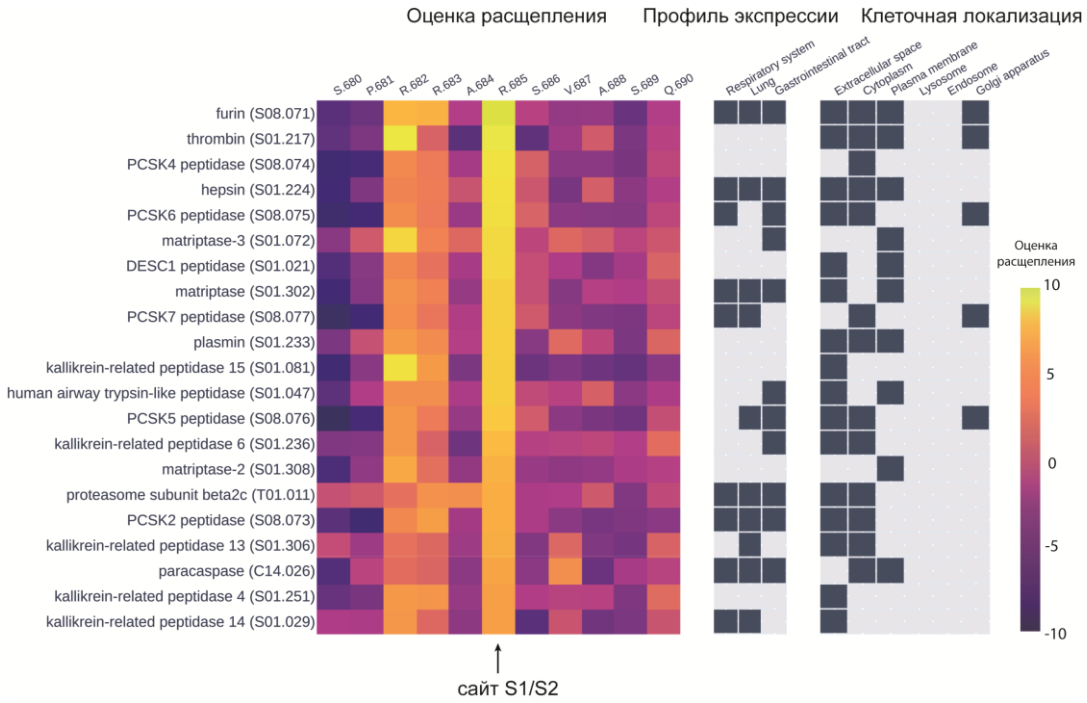

Рисунок 12. Двадцать наивысших оценок расщепления в позиции R685 (сайт S1/S2), полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель фуриновой протеазы. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

Согласно полученным результатам (Рис. 12), помимо фурина в активации S1/S2 сайта могут участвовать и другие протеазы. Высокие оценки расщепления демонстрируют модели, описывающие субстратную специфичность других представителей семейства пропротеин-конвертаз кексинового типа (PCSK), к которому принадлежит и сам фурин. Это позволяет предположить возможность частичной функциональной взаимозаменяемости данных ферментов. Некоторые из них, такие как PCSK5, PCSK6 или PCSK7, обладают схожими профилями экспрессии в тканях и

клеточной локализации с фурином. Кроме того, высокие оценки в этой позиции показали модели таких протеаз, как тромбин, гепсин, матриптаза и представители семейства калликреинов, обладающих частично пересекающимися профилями экспрессии и локализации. Учитывая, что процесс созревания вирусных частиц коронавируса SARS-CoV-2 происходит в аппарате Гольджи, особое внимание уделялось протеазам, локализованным в данном компартменте. К таким протеазам относятся, в первую очередь, представители семейства PCSK, а также тромбин.

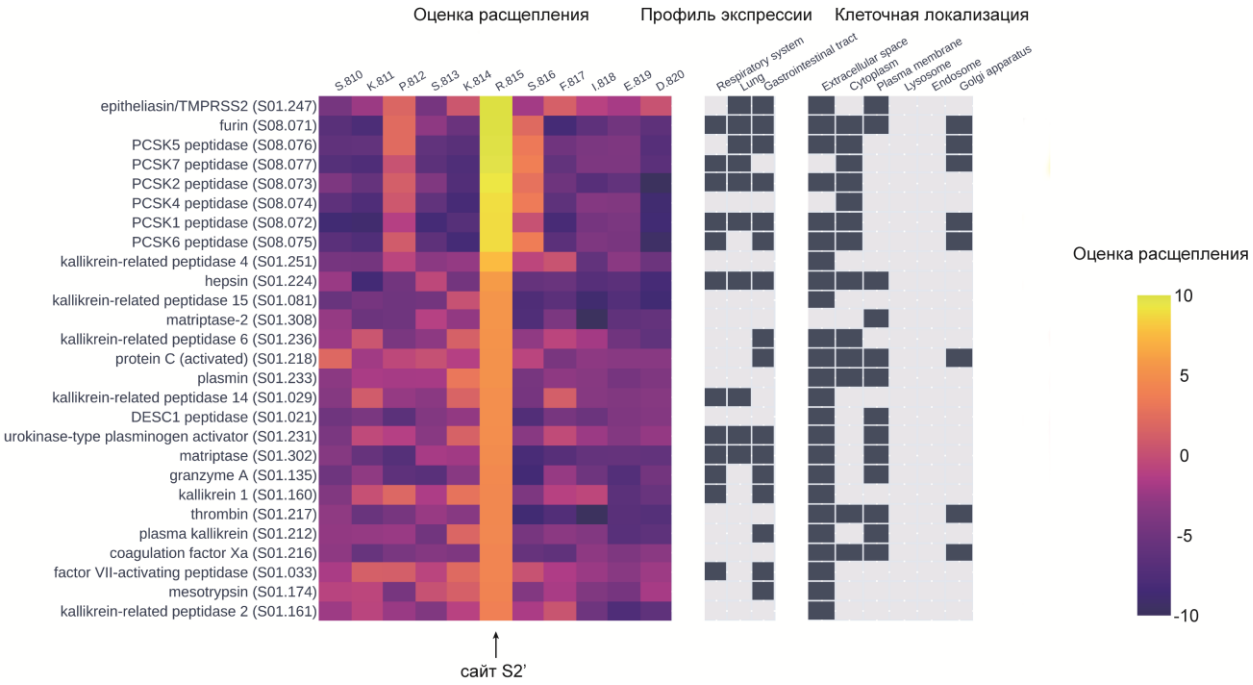

Рисунок 13. Тридцать наивысших оценок расщепления в позиции R815 (сайт S2`), полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель протеазы TMPRSS2. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

Для сайта S2` (позиция R815), по результатам моделирования (Рис. 13), потенциальными участниками протеолитической активации, помимо TMPRSS2, могут быть представители тех же семейств — PCSK, калликреины, а также некоторые факторы свёртывания крови. Однако, поскольку расщепление сайта S2` происходит на поверхности клетки в момент взаимодействия вирусной частицы с клеточным рецептором, ключевую роль, вероятно, играют протеазы, локализованные на клеточной мембране. К таким протеазам относятся фурин, гепсин, матриптазы, активатор плазминогена урокиназного типа и некоторые другие протеазы, причём гепсин и активатор плазминогена урокиназного типа имеют наиболее близкие к TMPRSS2 профили экспрессии в тканях. Важно также отметить, что гепсин и матриптаза относятся к тому же семейству трансмембранных

сериновых протеаз, что и TMPRSS2. При этом нельзя исключать участие и внеклеточных протеаз, к которым относятся калликреиновые протеазы и некоторые протеазы семейства PCSK. Наиболее сходными с TMPRSS2 профилями экспрессии среди этих протеаз обладают PCSK5, PCSK2, PCSK1, калликреин 1.

В **разделе 3.3** представлено применение разработанного метода для идентификации сайта расщепления катепсином L в S-гликопротеине оболочки коронавируса SARS-CoV-2.

Для идентификации потенциального сайта расщепления S-белка катепсином L модель субстратной специфичности данной протеазы была применена к аминокислотной последовательности S-белка. В результате анализа были получены оценки расщепления для всех возможных пептидных связей. Согласно полученным результатам (Рис. 14), максимальная оценка расщепления соответствует позиции K790.

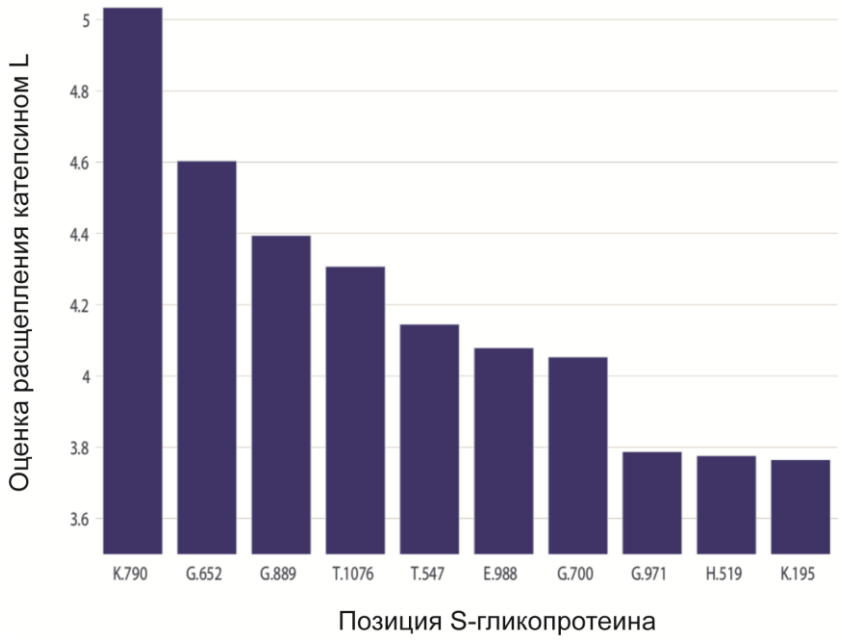

Рисунок 14. Десять позиций последовательности S-гликопротеина, получивших наивысшие оценки расщепления по модели субстратной специфичности катепсина L.

Далее было проанализировано расположение исследуемой позиции на трёхмерной структуре S-гликопротеина. Из базы данных PDB были получены экспериментальные структуры S-белка в "открытой" (PDB ID: 6VYB) и "закрытой" (PDB ID: 6VXX) конформациях, а также в комплексе с рецептором ACE2 (PDB ID: 7DF4). Незакристаллизованные участки, охватывающие сайты расщепления S1/S2 и S2`, были смоделированы с помощью инструмента SWISS-MODEL с параметрами по умолчанию. К полученным структурам была применена универсальная структурная модель

с целью получить оценку восприимчивости к протеолизу, с точки зрения структурных особенностей S-гликопротеина, участка белка в позиции K790. Полученные оценки восприимчивости к протеолизу были отображены на структурах S-белка при помощи инструмента UCSF Chimera (Рис. 15). Оценки, полученные универсальной структурной моделью, показали, что участок белка в позиции K790 обладает высоким потенциалом к расщеплению. Примечательно, что этот участок расположен в непосредственной близости от сайта S2` и примыкает к пептиду слияния, отвечающего за формирование поры при слиянии вирусной и клеточной мембран. Более того, оценка восприимчивости к протеолизу для позиции K790 оказалось даже выше, чем для позиции R815 — 0.61 против 0.58, соответственно. На основании этих данных можно предположить, что позиция K790 является сайтом протеолиза катепсином L, расщепление которого приводит к активации процесса слияния вирусной частицы с мембраной эндосомы при реализации эндоцитарного пути проникновения SARS-CoV-2.

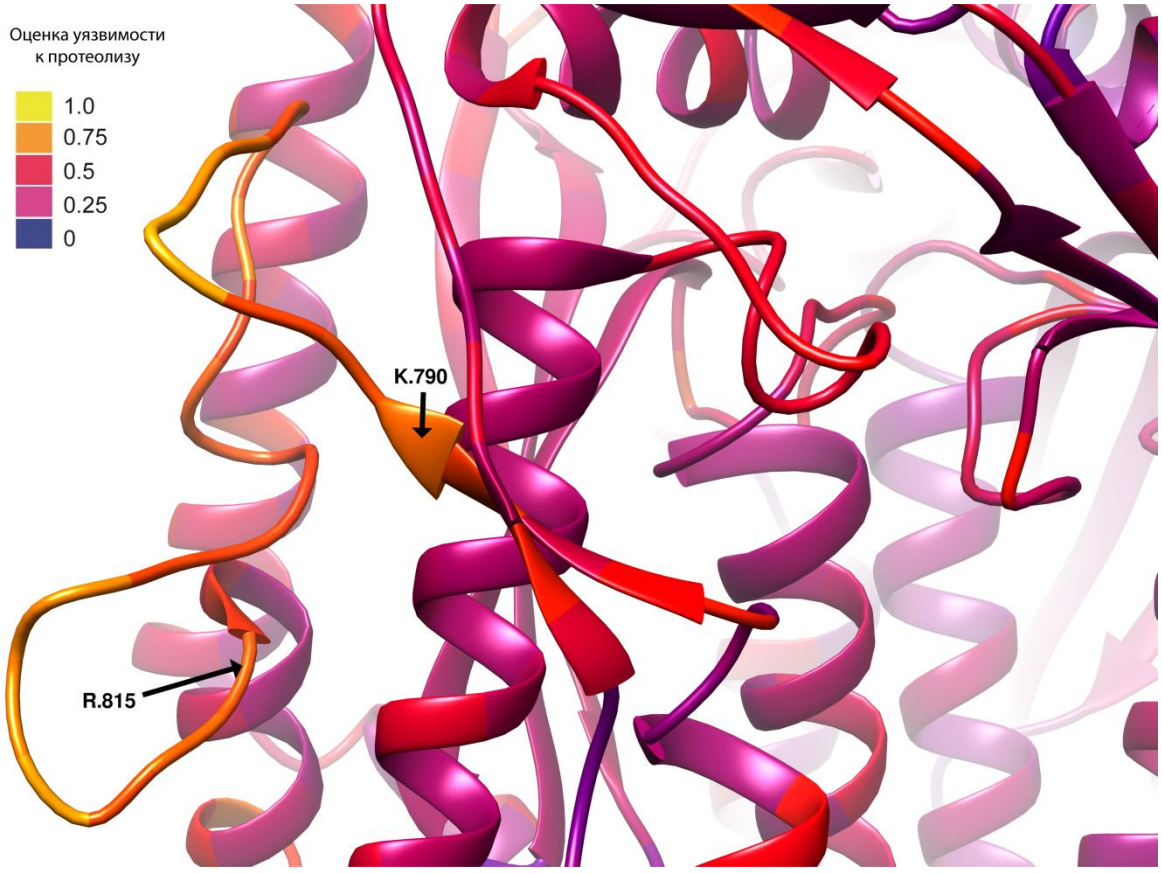

Рисунок 15. Отображение оценок восприимчивости к протеолизу, сгенерированных универсальной структурной моделью, на трёхмерной структуре S-гликопротеина. Стрелками указаны предполагаемый сайт расщепления катепсином L (K790) и известный сайт расщепления TMPRSS2 — S2` (R815).

Примечательно, что модель специфичности катепсина L продемонстрировала вторую по величине оценку в позиции K790: первая оценка соответствовала протеазе мезотрипсин, а третья — катепсину V. Мезотрипсин не локализуется в эндосомах или лизосомах, что делает его участие в активации S-белка при эндоцитарном пути проникновения вируса маловероятным, тогда как катепсин V относится к тому же семейству протеаз, что и катепсин L, и при этом обладает схожим профилем экспрессии и клеточной локализации.

Раздел 3.4 посвящён поиску альтернативных сайтов расщеплений в S-белке оболочки коронавируса SARS-CoV-2. По результатам анализа были предложены пять потенциальных позиций (R78, R346, R466, R634 и R847) для экспериментальной проверки, получивших высокие оценки подверженности протеолизу как по структурной модели, так и по моделям субстратной специфичности.

В **разделе 3.5** продемонстрировано применение разработанного метода для оценки влияния мутаций в S-белке оболочки коронавируса SARS-CoV-2 на эффективность расщепления сайтов протеолиза.

Модели субстратной специфичности протеаз, разработанные в настоящем исследовании, были далее использованы для анализа влияния аминокислотных замен в S-гликопротеине оболочки коронавируса SARS-CoV-2 на эффективность расщепления сайтов протеолиза. Информация о мутациях в S-белке различных вариантов коронавируса SARS-CoV-2 была получена из базы данных CoVariants. Особое внимание было уделено мутациям P681H и P681R, первая из которых характерна для вариантов коронавируса SARS-CoV-2 Альфа и Омикрон, а вторая — для вариантов Каппа и Дельта, соответственно. Эти мутации ассоциируются с повышенной инфекционностью данных штаммов коронавируса, что может быть обусловлено повышением эффективности расщепления сайта S1/S2. Применение модели специфичности фурина к аминокислотной последовательности S-гликопротеина с вышеописанными заменами в позиции P681 показало, что оценки расщепления в сайте S1/S2, в случае мутаций P681H и P681R превосходят значение оценки, вычисленной для исходного (уханьского) штамма (Рис. 18). Таким образом, с точки зрения биоинформатического моделирования, мутации вблизи сайта расщепления S1/S2 в S-белке штаммов Альфа, Омикрон, Каппа и Дельта повышают восприимчивость S-гликопротеина к протеолитическому расщеплению фурином.

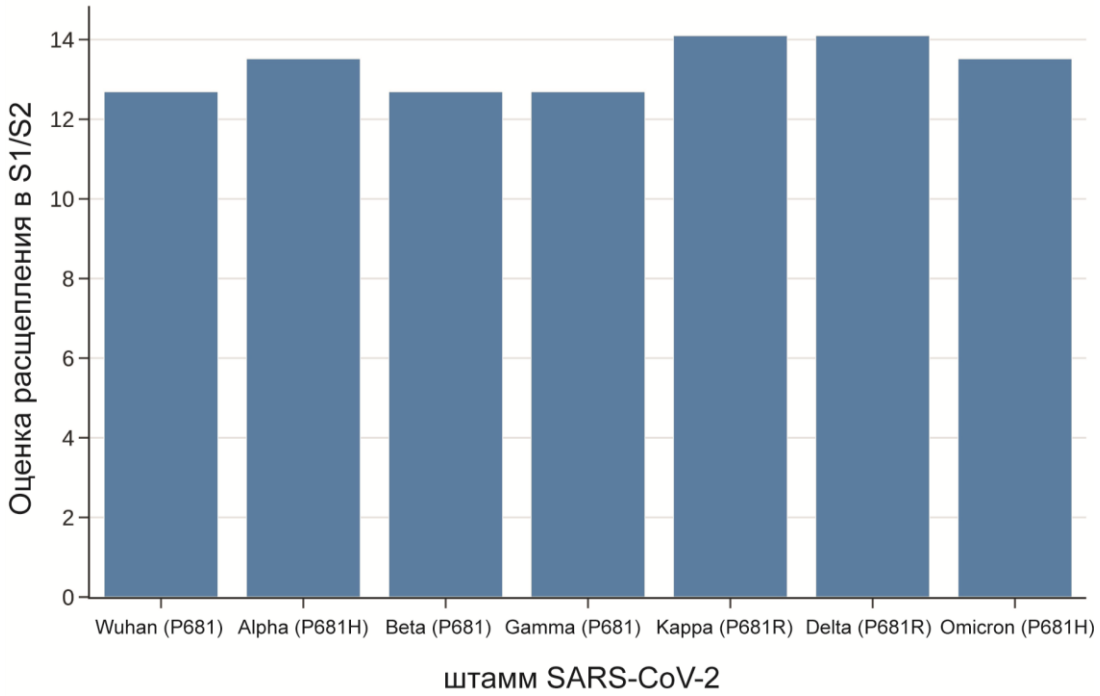

Рисунок 18. Оценки расщепления сайта S1/S2 S-гликопротеина, рассчитанные моделью субстратной специфичности фуриновой протеазы, для различных мутаций в позиции 681, наблюдаемых у отдельных штаммов коронавируса SARS-CoV-2.

Далее были рассчитаны оценки расщепления сайта S1/S2 моделью субстратной специфичности фурина для всех возможных аминокислотных замен в позиции P681. Согласно полученным результатам (Рис. 19), две наивысшие оценки расщепления фурином сайта S1/S2 соответствуют заменам P681R и P681H. Третья по величине оценка наблюдается для замены P681S, но такая замена не была обнаружена в S-белках какого-либо штамма коронавируса SARS-CoV-2. Четвёртая по величине оценка расщепления фурином сайта S1/S2 соответствует исходному варианту P681, который встречается в S-белках уханьского штамма коронавируса SARS-CoV-2.

Кроме того, в ходе анализа были выявлены варианты коронавируса SARS-CoV-2 с мутациями в S-белке в позиции R346, ранее отнесённой к числу потенциальных сайтов протеолитического расщепления. В различных штаммах коронавируса SARS-CoV-2 зафиксированы две замены в данной позиции — R346K и R346T. Анализ их влияния на эффективность расщепления сайта R346 показал значительное снижение оценок расщепления для данной позиции. Это может указывать на действие отбора, направленное на деградацию этого потенциального участка расщепления.

В **разделе 3.6** подведены итоги, обобщающие результаты, представленные в главе III.

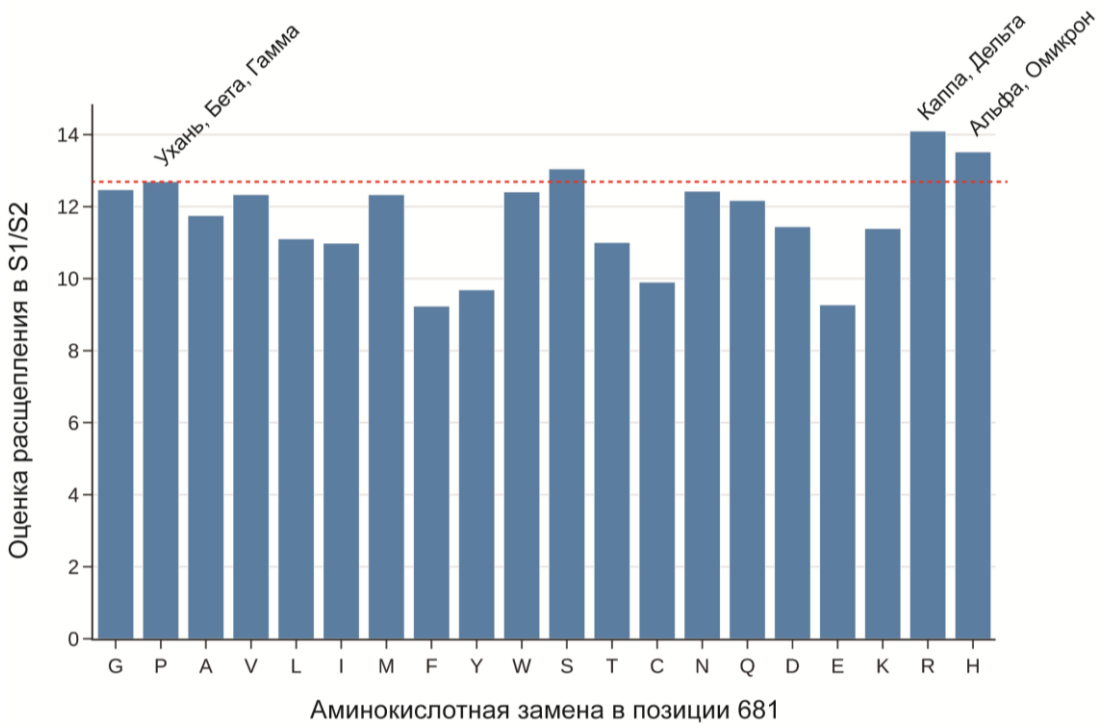

Рисунок 19. Оценки расщепления сайта S1/S2, рассчитанные моделью субстратной специфичности фуриновой протеазы для всех возможных замен в позиции 681.

Заключение

В данном исследовании был разработан подход к идентификации потенциальных протеолитических субстратов и поиску новых сайтов расщеплений, основанный на моделировании субстратной специфичности протеаз и структурной уязвимости участков белка к протеолизу. Была построена универсальная модель, независимая от конкретного типа протеазы, которая позволяет оценивать восприимчивость к расщеплению на основании структурных характеристик белка. Получаемые с её помощью оценки отражают потенциальную уязвимость участков полипептидной цепи к протеолитическому расщеплению. В настоящее время универсальные модели, аналогичные разработанной, отсутствуют среди доступных инструментов предсказания протеолитических событий.

Несмотря на высокое качество предсказаний, разработанная модель обладает рядом ограничений, которые необходимо учитывать при интерпретации результатов. Одним из таких ограничений является невозможность учитывать динамику конформационных изменений структуры белка. Так, оценки восприимчивости к протеолизу потенциально расщепляемых участков субстрата могут быть занижены моделью, если они экранируются подвижными структурными элементами белка. Частичным решением этой проблемы является использование нескольких доступных

трёхмерных структур одного и того же белка, полученных в разных экспериментах, и последующее усреднение полученных оценок по структурам для каждой пептидной связи. Также важно отметить, что универсальная структурная модель оценивает восприимчивость к протеолизу, исходя из текущей конформации белка, зафиксированной в его трёхмерной структуре, однако первое расщепление часто вызывает конформационные изменения субстрата, что делает прогнозирование последующих расщеплений ненадёжным. Таким образом, корректное предсказание вторичных и последующих расщеплений требует знания новой конформации, принимаемой белком после первичного расщепления, и наличия соответствующих трёхмерных структур. Другое ограничение универсальной структурной модели заключается в том, что она генерирует оценки вероятности расщепления, но не позволяет однозначно определить, какие пептидные связи субстрата действительно подвергнутся протеолизу. Для однозначного ответа на данный вопрос требуется установка порогового значения, классифицирующего оценки восприимчивости к протеолизу пептидных связей белка на расщепляемые и нерасщепляемые. С целью определения такого порога был проведён анализ распределения оценок восприимчивости в зависимости от условий проведения эксперимента по выявлению протеолитического события — *in vivo* или *in vitro*. Распределения оценок оказались схожими и не позволили установить чёткое значение порога. Более того, задание универсального порога, по-видимому, невозможно без учёта времени протекания протеолитической реакции. Предполагается, что использование экспериментальных данных, включающих информацию о времени протекания протеолитической реакции, может способствовать преодолению данного ограничения.

Несмотря на указанные ограничения, универсальная структурная модель в большинстве случаев демонстрирует высокое качество предсказаний, а визуализация результатов на трёхмерной структуре белка показывает их согласованность с современными представлениями о влиянии пространственной организации субстрата на эффективность протеолиза. Так, высокие оценки восприимчивости к протеолизу, как правило, соответствуют доступным и подвижным неструктурированным участкам белка, в то время как низкие значения наблюдаются в гидрофобном ядре и устойчивых элементах вторичной структуры. Согласно литературным данным, посттрансляционные модификации также могут существенно влиять на чувствительность белка к протеолитическому расщеплению, поэтому использование структур с наличием посттрансляционных модификаций может повысить точность прогнозов универсальной структурной модели.

Объединение универсальной структурной модели и моделей субстратной специфичности по аминокислотной последовательности позволило достичь качества прогнозирования на уровне передовых инструментов предсказания субстратов протеаз и протеолитических сайтов.

Разработанные модели были применены для прогнозирования протеолитических событий с целью решения двух важных задач, связанных с изучением механизмов протеолитической активации S-гликопротеина оболочки коронавируса SARS-CoV-2. Были построены 169 моделей субстратной специфичности протеаз человека, и на основе этих моделей, а также универсальной структурной модели, были идентифицированы протеазы, потенциально вовлечённые в протеолитическое расщепление S-белка в известных сайтах. Полученные данные позволили идентифицировать четыре семейства протеаз, потенциально способных активировать S-гликопротеин: пропротеин-конвертазы кексинового типа, трансмембранные сериновые протеазы, факторы свёртывания крови и калликреины.

Также в ходе работы был предсказан потенциальный сайт расщепления катепсином L в S-белке коронавируса SARS-CoV-2 — позиция K790, прилегающая к известному сайту расщепления S2` и пептиду слияния. На сегодняшний день в литературе отсутствует единое мнение относительно точной позиции расщепления катепсином L, и предсказанный в данной работе вариант не совпадает с ранее описанными. Однако анализ трёхмерной структуры показывает, что позиция K790 является надёжным кандидатом для экспериментальной проверки, поскольку расщепление в данной позиции, вероятнее всего, приводит к конформационным изменениям в S-гликопротеине, аналогичным тем, которые происходят при расщеплении сайта S2` (позиция R815).

Кроме того, разработанный подход позволил оценить влияние мутаций в S-белке на эффективность протеолитического расщепления. В частности, было показано, что замены в фуриновом сайте S1/S2, наблюдаемые в высокоинфекционных штаммах коронавируса SARS-CoV-2 (например, P681H и P681R), ассоциированы с повышенными оценками расщепления. Эти результаты демонстрируют потенциал разработанных биоинформатических моделей для предсказания эволюционных изменений в участках вирусного генома, кодирующих сайты протеолиза.

Все разработанные модели были выложены в открытый доступ на платформе GitHub (https://github.com/KazanovLab/ProteolysisStructuralPrediction и https://github.com/KazanovLab/ProteaseSpecificityModels) и могут быть использованы исследователями для решения прикладных задач, связанных с

идентификацией потенциальных субстратов и сайтов протеолитического расщепления, а также с определением протеаз, способных активировать интересующий белок.

В заключении следует отметить, что поставленная цель исследования была достигнута, все запланированные задачи выполнены, а разработанный подход к прогнозированию протеолитических событий продемонстрировал свою актуальность и эффективность. Предложенная методология может быть успешно использована для идентификации потенциальных субстратов протеолитических ферментов и определения новых сайтов расщепления, что делает её перспективным инструментом для дальнейших исследований в области протеолиза.

Выводы:

1. Разработана модель предсказания восприимчивости участков белка к протеолизу на основании информации о его пространственной структуре.
2. Разработаны модели специфичности по аминокислотной последовательности для 169 протеаз человека, представленные в виде позиционно-весовых матриц (PWM).
3. Разработан подход интеграции структурной модели с моделями специфичности по последовательности и показано, что объединённая модель демонстрирует качество предсказания, сопоставимое с таковым для существующих методов, при этом охватывая больший спектр протеаз и обладая гибкостью и расширяемостью за счёт возможности добавления новых PWM-моделей.
4. Разработанный метод применён для анализа протеолитической активации S-гликопротеина коронавируса SARS-CoV-2. Качество работы метода подтверждает корректное предсказание двух наиболее изученных сайтов расщепления S-белка — S1/S2 и S2`. Установлено, что представители четырёх семейств сериновых протеаз — PCSK, TTSP, калликреины и факторы свёртывания крови — потенциально способны расщеплять указанные сайты в случае колокализации с вирусным белком.
5. В позиции K790 S-гликопротеина коронавируса SARS-CoV-2 предсказан ранее неизвестный протеолитический сайт катепсина L — цистеиновой протеазы, задействованной при проникновении коронавируса внутрь клетки эндоцитарным способом. Пространственный анализ показал, что данный сайт располагается вблизи сайта S2` и пептида слияния — двух

функционально значимых элементов S-белка, обеспечивающих слияние вирусной оболочки и клеточной мембраны.

6. Разработанный метод был применен для анализа влияния мутаций, выявленных в S-гликопротеине известных штаммов коронавируса SARS-CoV-2, на эффективность его протеолитического расщепления.

Список работ, опубликованных автором по теме диссертации

I. Matveev, E. V. Predicting Structural Susceptibility of Proteins to Proteolytic Processing / E. V. Matveev, V. V. Safronov, G. V. Ponomarev, M. D. Kazanov // International Journal of Molecular Sciences. — 2023. — Vol. 24 (13). — P. 10761. — doi: 10.3390/ijms241310761

II. Матвеев, Е. В. Биоинформатический метод идентификации протеаз человека, активных относительно гликопротеинов оболочки вирусов, на примере белка шипа коронавируса SARS-CoV-2 / Е. В. Матвеев, Г. В. Пономарёв, М. Д. Казанов // Молекулярная биология. — 2024. — Т. 58(1). — С. 171–177. — DOI: 10.31857/S0026898424010176

III. Matveev, E. V. Genome-wide bioinformatics analysis of human protease capacity for proteolytic cleavage of the SARS-CoV-2 spike glycoprotein / E. V. Matveev, G. V. Ponomarev, M. D. Kazanov // Microbiology Spectrum. — 2024. — Vol. 12 (2). — P. e03530-23. — doi: 10.1128/spectrum.03530-23

Результаты работы, опубликованные в сборниках тезисов конференций:

I. Матвеев Е.В., Сафронов В.В., Казанов М.Д. Биоинформатическая идентификация субстратов протеаз на основе известной трёхмерной структуры белков / XXVII Симпозиум «Биоинформатика и компьютерное конструирование лекарств» // онлайн, 5–7 апреля 2021 г.

II. Matveev E.V., Kazanov M.D. Computational prediction of susceptibility to limited proteolysis for proteins with known 3D structure / XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery // Москва, Россия, 24–26 мая 2022 г.

III. Matveev E.V., Kazanov M.D. Bioinformatics screening of human proteases for ability of activating SARS-CoV-2 spike protein / Proceedings of 11th Moscow Conference on Computational Molecular Biology MCCMB’23 // Москва, Россия, 3–6 августа 2023 г.