

Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

На правах рукописи

Матвеев Евгений Викторович

**Предсказание сайтов протеолиза на основе информации о пространственной
структуре потенциальных субстратов**

Специальность 1.5.8 —
«Математическая биология, биоинформатика»

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель:
кандидат технических наук
Казанов Марат Джамалудинович

Москва — 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.....	5
ГЛАВА I. ОБЗОР ЛИТЕРАТУРЫ	14
1.1. Протеолиз, ограниченный протеолиз	14
1.2. Разнообразие протеаз	15
1.3. Протеолитическая активация вирусных гликопротеинов	20
1.4. Влияние протеолиза на жизненный цикл SARS-CoV-2	27
1.4.1. Жизненный цикл коронавируса SARS-CoV-2.....	27
1.4.2. Структура S-белка	29
1.4.3. Рецепторы SARS-CoV-2.....	30
1.4.4. Протеолиз S-белка SARS-CoV-2.....	31
1.4.5. Влияние аминокислотных замен на протеолиз S-белка	33
1.5. Биоинформатическое прогнозирование протеолитических сайтов	35
1.5.1. Специфичность протеазы	35
1.5.2. Детерминанты ограниченного протеолиза	37
1.5.3. Базы данных и инструменты прогнозирования протеолитических событий.....	40
ГЛАВА II. РАЗРАБОТКА МЕТОДА ПРОГНОЗИРОВАНИЯ ПРОТЕОЛИТИЧЕСКИХ СОБЫТИЙ.....	47
2.1. Построение модели предсказания уязвимости участков белка к протеолизу на основе информации о структурных особенностях субстратов	47
2.1.1. Подготовка набора данных для построения предсказательной модели	47
2.1.2. Формирование обучающей выборки и выбор оптимального алгоритма машинного обучения для построения предсказательной модели	50
2.1.3. Визуализация результатов предсказаний на трёхмерных структурах потенциальных субстратов и определение порогового значения оценок восприимчивости к протеолизу для физиологических сайтов расщепления.....	53
2.2. Расширение обучающей выборки с использованием моделей трёхмерных структур субстратов из базы данных AlphaFoldDB	56

2.2.1. Анализ особенностей моделей трёхмерных структур белков AlphaFold с точки зрения построения предсказательной модели восприимчивости к протеолизу	56
2.2.2. Подготовка расширенного набора данных для построения модифицированной предсказательной модели.....	57
2.2.3. Построение модифицированной предсказательной модели на наборе данных, расширенном AlphaFold-моделями	58
2.2.4. Сравнение качества прогнозирования исходной и модифицированной предсказательных моделей	61
2.3. Интеграция структурной модели с моделями специфичности протеаз по аминокислотной последовательности и сравнение с существующими методами предсказания сайтов протеолиза.....	63
2.3.1. Построение интегративной модели на основе структурной модели и моделей специфичности протеаз по аминокислотной последовательности	63
2.3.2. Сравнение разработанного метода прогнозирования протеолитических событий с существующими методами	66
2.4. Обсуждение результатов	68
ГЛАВА III. БИОИНФОРМАТИЧЕСКИЙ АНАЛИЗ ПРОТЕОЛИТИЧЕСКОЙ АКТИВАЦИИ S-ГЛИКОПРОТЕИНА ОБОЛОЧКИ КОРОНАВИРУСА SARS-COV-2	71
3.1. Протеолитическая активация S-гликопротеина оболочки коронавируса SARS-CoV-2.....	71
3.2. Биоинформатическая идентификация протеаз человека, способных активировать S-белок оболочки коронавируса SARS-CoV-2	72
3.2.1. Разработка предсказательных моделей, описывающих субстратную специфичность протеаз человека по аминокислотной последовательности	72
3.2.2. Валидация моделей субстратной специфичности.....	74
3.2.3. Идентификация протеаз, потенциально способных активировать S-гликопротеин оболочки коронавируса SARS-CoV-2 в известных сайтах расщепления S1/S2 и S2`	76
3.3. Биоинформатический поиск сайта расщепления катепсином L в S-гликопротеине оболочки коронавируса SARS-CoV-2	77
3.4. Поиск альтернативных сайтов расщепления в S-белке коронавируса SARS-CoV-2.....	80
3.5. Влияние мутаций в S-гликопротеине оболочки коронавируса SARS-CoV-2 на эффективность расщепления сайтов протеолиза.....	87

3.6. Обсуждение результатов	90
ЗАКЛЮЧЕНИЕ	93
ВЫВОДЫ	97
БЛАГОДАРНОСТИ.....	99
СПИСОК ЛИТЕРАТУРЫ	100
СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ	116
СПИСОК ИЛЛЮСТРАЦИЙ	119

ВВЕДЕНИЕ

Актуальность 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 [1]. 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 [1, 2]. 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 [2], поскольку знание протеаз и их мишеней имеет значение для биомедицинских исследований и разработки новых терапевтических подходов [3].

О практической значимости протеолитических ферментов известно достаточно давно. С древних времен для приготовления пищи и напитков использовались экстракты животного и растительного происхождения, содержащие протеазы, например при изготовлении сыра или пива [4, 5]. В наши дни протеазы широко используются в промышленности как детергенты, средства для обработки тканей и переработки пищевых продуктов [6]. В медицине и фармацевтике протеазы применяются в качестве лекарственных средств для контроля коагуляции крови, ускорения пищеварительных процессов, лечения нарушений нервно-мышечной системы [7]. Протеолитические ферменты

являются лекарственными мишенями в терапии сердечно-сосудистых заболеваний, ВИЧ-инфекции, диабета [8].

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 [9]. Существует множество экспериментальных методов, направленных на идентификацию субстратов (например, гель-электрофорез), определение позиций сайтов расщеплений (методы детектирования N-концевых пептидов, масс-спектрометрия), а также способных оценивать динамику и масштабы протеолитических событий [10]. 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: одна и та же молекула может быть субстратом для нескольких протеолитических ферментов, а сами протеазы нередко являются субстратами других протеаз. Кроме того, запуск протеолитических каскадов часто инициируется внешними сигналами, что ещё больше усложняет интерпретацию экспериментальных данных [2, 11–15]. Во-вторых, на практике чрезвычайно трудно выделить продукт протеолиза из клеточной смеси белков и пептидов и точно идентифицировать его исходный субстрат [10]. 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

В настоящее время доступны несколько биоинформатических инструментов прогнозирования протеолитических событий [16]. 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.

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 [17], что открывает перспективу создания инструмента такого типа.

Цель и задачи исследования

Целью настоящего 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 (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:
 - а. 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 перекрытия 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 исследования

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. Впервые разработаны модели субстратной специфичности для 169 протеаз человека; функционал метода может быть расширен за счёт добавления моделей специфичности для других протеаз. Предсказана позиция сайта расщепления S-белка коронавируса SARS-CoV-2 цистеиновой протеазой катепсин L. Кроме этого, 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 разрабатывался 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 результатов

Обучение и валидация структурной модели выполнялись на основе данных о протеолитических событиях, полученных из базы данных CutDB [18]. Протеазо-специфические модели, описывающие субстратную специфичность протеаз человека, были построены с использованием информации о подтверждённых сайтах расщепления из базы данных MEROPS [19]. 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*.

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.

Апробация результатов

Результаты 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”.

Список публикаций по теме диссертации

По теме диссертации опубликовано три статьи, индексируемые в базах данных Scopus и РИНЦ:

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

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.

ГЛАВА I. ОБЗОР ЛИТЕРАТУРЫ

1.1. Протеолиз, ограниченный протеолиз

Протеолиз представляет собой ферментативный процесс гидролиза, при котором осуществляется расщепление (разрыв) пептидной связи в молекуле субстрата. Этот процесс катализируют специализированные ферменты — протеазы. В качестве субстратов протеаз могут выступать различные белки и пептиды [20]. Ключевой характеристикой протеаз является специфичность, то есть, способность избирательно распознавать и связываться с определённым участком молекулы субстрата — сайтом протеолиза — и осуществлять его расщепление [21].

Долгое время протеазы рассматривались исключительно в качестве ферментов, обеспечивающих распад белков и ответственных за белковый катаболизм [22]. Считалось, что протеазы не избирательны по отношению к субстратам, а лизосома — основной клеточный компартмент, в котором происходит протеолиз. Однако со временем стало понятно, что роль протеолитических ферментов не ограничивается участием в белковом обмене. Выяснилось, что протеазы могут осуществлять протеолитический процессинг белков [2] — катализировать специфическое расщепление ограниченного числа пептидных связей в молекуле субстрата. Этот процесс может происходить в различных компартментах клетки, начиная от лизосом и эндосом и заканчивая плазматической мембраной и митохондриями. Такой протеолиз называют ограниченным или регуляторным. Он играет ключевую роль в регуляции активности и локализации белков, модулирует их взаимодействие и участвует в передаче сигналов в сложных сетях молекулярных взаимодействий [11–15]. Протеолитический процессинг рассматривается как один из видов посттрансляционной модификации, поскольку приводит к появлению новых белковых форм и тем самым расширяет разнообразие протеома. Он является

повсеместно распространённым благодаря широкой локализации протеаз в клетке. Протеолиз является необратимым процессом, что отличает его от большинства других посттрансляционных модификаций, таких как гликозилирование или фосфорилирование, которые, как правило, представляют собой обратимое ковалентное присоединение определённых химических групп к аминокислотным остаткам белка [23].

1.2. Разнообразие протеаз

Согласно устоявшейся терминологии, ферменты, катализирующие гидролиз пептидных связей, называются протеазами или пептидазами — эти термины являются взаимозаменяемыми [20]. В зависимости от расположения расщепляемой пептидной связи в полипептидной цепи, все пептидазы подразделяют на эндопептидазы и экзопептидазы. Эндопептидазы, или протеиназы, расщепляют пептидные связи внутри полипептидной цепи белка. Экзопептидазы, наоборот, отщепляют аминокислотные остатки или небольшие пептиды от концов полипептидной цепи, причём аминопептидазы — со стороны N-конца, а карбоксипептидазы — со стороны C-конца полипептидной цепи. Кроме этого выделяют олигопептидазы [24] — подгруппу эндопептидаз, участвующих в расщеплении небольших субстратов длиной 20-30 аминокислотных остатков — олигопептидов [20], и дипептидазы — ферменты, расщепляющие дипептиды, то есть, пептиды, состоящие из двух аминокислотных остатков [25].

В зависимости от строения активного центра фермента и механизма катализа химической реакции выделяют следующие классы протеаз:

- сериновые протеазы (трипсин, тромбин, плазмин) [26];
- цистеиновые протеазы (папаин, катепсины В, L, К, каспазы) [27];
- треониновые протеазы (субъединицы протеасомы) [28];

- металлопротеазы (ангиотензин-конвертирующий фермент, коллагеназы, нейтрофильная эластаза) [29];
- аспартатные протеазы (пепсин, химозин, катепсин D, HIV-1 протеаза) [30];
- глутаматные протеазы (scytalidoglutamic peptidase, aspergilloglutamic peptidase) [31].

В названии классов отражены те аминокислотные остатки или химические элементы, которые составляют активный центр фермента и необходимы для осуществления протеолиза. Протеазы различных классов объединяются в семейства на основе сходства их аминокислотных последовательностей, а семейства протеаз, в свою очередь, объединяются в кланы на основе сходства трёхмерных структур их каталитических доменов [1].

В зависимости от структурной организации фермента, протеазы могут состоять как из одной полипептидной цепи, включающей каталитический домен, так и из множества субъединиц с различными функциональными доменами. Протеолитические ферменты первой группы представляют собой небольшие молекулы массой около 20 кДа, такие как папаин, тромбин, трипсин, эластаза. Вторая группа протеаз включает крупные молекулярные комплексы, обладающие протеолитической активностью, например, протеасому, апоптосому, меприны. Функциональные домены таких протеаз обеспечивают субстратную специфичность, определяют клеточную локализацию фермента, модифицируют его кинетические свойства и влияют на чувствительность к эндогенным ингибиторам [32].

Протеазы классифицируются не только по механизму действия, но и по типу расщепляемых пептидных связей. Большинство протеаз расщепляют наиболее распространённые альфа-пептидные связи. Однако существуют и специализированные протеазы, например, деубиквитинирующие ферменты, расщепляющие изопептидные связи между убиквитином и белками-мишенями [33]. Ферменты, такие как гамма-глутамилгидролаза и глутамат-карбоксипептидаза, способны расщеплять гамма-пептидные связи [1]. Отдельную

группу составляют протеазы, способные расщеплять пептидные связи внутри собственной молекулы — процесс, известный как аутопротеолиз. Примерами таких ферментов являются протеазы Nup98 (нуклеопорин) и полицистин-1, которые после саморасщепления теряют протеолитическую активность [34, 35].

Протеазы представляют собой одну из наиболее распространённых и функционально разнообразных групп ферментов. Они обнаруживаются во многих клеточных компартментах, а также во внеклеточном пространстве. В цитоплазме клетки протеазы участвуют как в полном расщеплении белков (тотальный протеолиз), так и в протеолитическом процессинге. Именно в цитоплазме клетки находится протеасома [36], ответственная за деградацию неправильно свернувшихся и повреждённых белков. Также в цитоплазме присутствуют кальпаины [37], осуществляющие ограниченный протеолиз рецепторов и мембранных белков, и другие протеазы, участвующие в метаболизме гормонов и нейропептидов, например, IDE (Insulin-Degrading Enzyme) [38] и пролил-олигопептидаза [39]. В лизосомах преобладают протеазы, катализирующие расщепление белков до пептидов и аминокислот. Хотя основная функция лизосом — тотальный протеолиз, некоторые лизосомные ферменты, такие как катепсин В, осуществляют ограниченный протеолиз, активируя коллагеназу, урокиназу и химозин. Катепсины — наиболее известные протеазы лизосом, активность которых зависит от уровня кислотности среды (pH) [40]. В эндоплазматическом ретикулуме и комплексе Гольджи протеолитический процессинг обеспечивает созревание белков. Здесь наиболее активны протеазы семейства PCSK и, в частности, фурин [41]. Специализированная протеаза митохондрий — MPP (Mitochondrial Processing Peptidase) — осуществляет ограниченный протеолиз белков, попадающих в митохондрии из цитоплазмы [42]. Некоторые протеазы связаны с плазматической мембраной [43, 44]. Одни из них являются интегральными белками, пронизывающими мембрану насквозь, тогда как другие проявляют активность опосредованно, например, после связывания с рецептором. Так неприлизин и энтеропептидаза представляют собой мембранные протеазы: первая обеспечивает распад нейропептидов, в то время как вторая активирует

трипсин путём протеолиза трипсиногена. Активаторы плазминогена урокиназного и тканевого типов, напротив, связываются с мембраной не напрямую, а через взаимодействие с соответствующими рецепторами. Многие клетки секретируют протеазы и их ингибиторы во внеклеточное пространство. Среди них — матриксные металлопротеазы, играющие ключевую роль в разрушении внеклеточного матрикса, особенно при росте опухолей [45], а также пищеварительные ферменты, такие как трипсин, пепсин, химотрипсин [20].

Каждой протеазе соответствует определённый набор субстратов, обозначаемый специальным термином — деградом [2]. Деградом отражает субстратную специфичность протеазы. Разные протеолитические ферменты отличаются по субстратной специфичности [46]. Одни из них характеризуются широкой субстратной специфичностью и способны расщеплять разнообразные белки. Обычно такие протеазы участвуют в полном разрушении белковых молекул. В качестве примера можно привести протеиназу К или протеазы, связанные с перевариванием пищи. Другие ферменты, напротив, расщепляют лишь ограниченное число субстратов. К таким протеазам относятся каспазы, фурин и многие другие протеазы, ответственные за регуляторный протеолитический процессинг белков. В редких случаях протеаза может проявлять моноспецифичность, то есть, расщеплять лишь один субстрат. Как правило, моноспецифичными являются аутопротеазы, расщепляющие собственную молекулу [34, 35].

Протеолитические ферменты активно изучаются в организмах различных царств живой природы. Протеазы обнаружены у животных, растений, грибов, архей, бактерий и вирусов. Согласно базе данных MEROPS [19], у человека известно около 700 протеолитических ферментов, включая предсказанные протеазы. Распределение протеаз по каталитическим классам неравномерно: наиболее многочисленны металлопротеазы и сериновые протеазы, тогда как треониновые протеазы составляют меньшинство, а глутаматные протеазы до настоящего времени не были обнаружены у млекопитающих. Протеазы в организме человека задействованы во многих физиологических и патологических

процессах, включая свёртывание крови, овуляцию и оплодотворение, ангиогенез, нейрогенез, заживление ран, воспаление, канцерогенез и старение [1]. У растений протеазы участвуют в регуляции роста, развития и ответных реакций на изменение условий окружающей среды. Протеазы также играют ключевую роль в иммунитете растений, в том числе в развитии реакции сверхчувствительности — специализированного механизма, ограничивающего распространение патогенов [47, 48]. У грибов протеолитические ферменты задействованы в процессах спорообразования и формирования колоний. Многие протеазы грибов секретируются, что делает их удобными объектами промышленной ферментации [49]. Примечательно, что глутаматные протеазы, за редким исключением, обнаруживаются только у грибов и бактерий. У грибов эти протеазы участвуют в расщеплении белков древесины погибших деревьев, а также в патогенезе растений, например, при инфицировании грибом белой плесени *Sclerotinia sclerotiorum* [50]. Бактериальные протеазы играют важную роль в регуляции деградации белков. Этот процесс обеспечивает контроль качества бактериального протеома, координацию клеточного цикла и адаптацию к стрессовым условиям окружающей среды [51, 52]. Внутримембранный протеолиз обеспечивает передачу сигналов в ответ на изменения условий окружающей среды у грамм-положительных и грамм-отрицательных бактерий [53]. У многих вирусов в геноме закодированы собственные протеазы, играющие критическую роль в жизненном цикле вируса [54]. Они участвуют в протеолитическом процессинге вирусных полипротеинов и созревании новых вирусных частиц. В связи с этим протеазы вирусов рассматриваются как перспективные мишени для антивирусной терапии. Примечательно, что некоторые протеазы вирусов схожи по аминокислотной последовательности с протеазами других организмов. К ним относятся химотрипсин-подобные сериновые и цистеиновые протеазы альфавирусов и пикорнавирусов. В то же время существуют уникальные вирусные протеазы, такие как протеазы вируса герпеса или аденовируса [54]. Кроме этого, при проникновении вируса в клетку гликопротеины оболочки

вирусов подвергаются протеолизу со стороны протеаз клеток-хозяев, а некоторые клеточные протеазы способствуют созреванию новых вирусных частиц [55].

1.3. Протеолитическая активация вирусных гликопротеинов

Протеолитическая активация вирусных гликопротеинов — широко распространённый механизм, обеспечивающий проникновение вирусов внутрь клеток [56]. Впервые этот механизм был изучен на примере расщепления гликопротеина F вируса Сендай, в результате которого высвобождается пептид слияния, инициирующий слияние вирусной оболочки с клеточной мембраной [55]. Подобное протеолитическое расщепление может также способствовать связыванию гликопротеинов с клеточными рецепторами. На примере вируса болезни Ньюкасла было показано, что способность вируса заражать различные типы клеток и степень патогенности коррелируют с тем, насколько эффективно его гликопротеин расщепляется клеточными протеазами: у вирулентных штаммов расщепление происходит в широком спектре тканей, в то время как у авирулентных расщепление ограничено определёнными типами клеток. Таким образом, протеолиз вирусных гликопротеинов может оказывать влияние как на инфекционность вируса — его способность заражать новые клетки, так и на патогенность, то есть, степень вызываемого им повреждения организма [55].

Важной группой клеточных протеаз, задействованных различными вирусами на этапах проникновения в клетку и созревания вирусных частиц, является семейство пропротеин-конвертаз — сериновых протеаз эукариот [57]. Ключевым представителем этого семейства является фурин. Помимо фурина, к семейству пропротеин-конвертаз относятся протеазы PC4, PC5, PACE4 и PC7, которые вместе с фурином образуют подсемейство кексиноподобных пропротеин-конвертаз. Эти протеазы локализуются в органеллах секреторных путей, таких как аппарат Гольджи, эндоплазматический ретикулум, эндосомы и другие [58].

Основная функция пропротеин-конвертаз заключается в посттрансляционной активации белков, что увеличивает функциональное разнообразие протеома клетки. Эти ферменты играют ключевую роль в регуляции процессов клеточного роста, пролиферации и дифференцировки. Большинство клеток экспрессируют все перечисленные кексиноподобные пропротеин-конвертазы, за исключением протеазы PC4, экспрессия которой ограничена клетками семенников, яичников и плаценты. Пропропротеин-конвертазы представляют собой крупные мультидоменные белки, состоящие из консервативных каталитических и переменных регуляторных Р-доменов. Фурин, PC5 и PC7 являются мембраносвязанными белками I типа, при этом фурин и PC5 могут высвобождаться из клетки во внеклеточное пространство. В отличие от них, PC4 является полностью секретируемым белком. Субстратная специфичность кексиноподобных пропротеин-конвертаз определяется консенсусной последовательностью: $P4Arg-P3X-P2X-P1Arg-P1'X$, где X — любой аминокислотный остаток, а между остатками P1Arg и P1'X происходит расщепление. Этот мотив часто встречается в гликопротеинах вирусов, активируемых на стадии созревания или проникновения в клетку. По характеру специфичности кексиноподобные пропротеин-конвертазы делятся на две группы: одну группу составляет фурин, обладающий наибольшей селективностью, тогда как PC4, PC5, PC6 и PC7 образуют вторую, менее специфичную группу [59].

Пропропротеин-конвертазы играют ключевую роль в протеолитическом процессинге белков, расположенных на поверхности вирусных частиц. Расщепление вирусных гликопротеинов в основном происходит внутри клеток-хозяев на стадии сборки вирионов, до их выхода во внеклеточную среду. Реже расщепление пропротеин-конвертазами происходит внеклеточно, то есть, непосредственно во время проникновения вируса в клетку. Несмотря на то, что другие протеазы также могут быть вовлечены в протеолитический процессинг вирусных гликопротеинов, было показано, что пропротеин-конвертазы чаще способствуют усилению инфекционности некоторых вирусов по сравнению с расщеплением другими протеолитическими ферментами [57-59].

Например, гликопротеин gB, входящий в состав оболочки вируса герпеса, играет ключевую роль в слиянии вирусной частицы с клеточной мембраной. На начальном этапе gB синтезируется в форме белка-предшественника, который затем подвергается протеолитическому расщеплению пропротеин-конвертазами. Расщепление происходит в участке с петлевой вторичной структурой, находящемся на определённом расстоянии от пептида слияния. Длина и аминокислотная последовательность этой петли существенно варьируют среди различных представителей вируса герпеса. Протеолитическое расщепление гликопротеина оболочки вируса герпеса gB активирует слияние вируса с клеточной мембраной [60].

Оболочка ортомиксовирусов, в том числе вируса гриппа, содержит два основных гликопротеина: гемагглютинин (HA) и нейраминидазу (NA). Оба белка играют важную роль в патогенности ортомиксовирусов [61]. Степень и разнообразие клеточной протеолитической активности также является существенным фактором, определяющим патогенность, распространение и тропизм вируса гриппа — основного представителя ортомиксовирусов. Гемагглютинин синтезируется в секреторных путях инфицируемых клеток в виде неактивного предшественника Н0, который ограничивает процесс слияния вирусной и клеточной мембран. Активация HA происходит в результате протеолитического расщепления, в ходе которого Н0 расщепляется на две субъединицы гемагглютинина — Н1 и Н2. Протеолитическое расщепление гемагглютинина происходит в петлевом участке, соединяющем две субъединицы, и который варьирует по длине и аминокислотной последовательности в зависимости от штамма вируса. Для гемагглютинина было установлено, что эффективность протеолиза зависит от длины линкерной последовательности. Линкерная последовательность, состоящая из одного аминокислотного остатка аргинина, подвергалась протеолитическому расщеплению гораздо реже, чем линкерная последовательность с множеством положительно заряженных аминокислотных остатков. Помимо длины петлевого участка, были выявлены и другие факторы, определяющие восприимчивость гемагглютинина к протеолизу,

включая точечные мутации, находящиеся на некотором расстоянии от сайта расщепления. Такие мутации могут повышать вероятность протеолиза гемагглютиниона со стороны большего числа протеаз, что увеличивает спектр инфицируемых клеток. Существенное значение также имеют посттрансляционные модификации аминокислотных остатков вблизи сайта расщепления, в частности, гликозилирование, которое может препятствовать доступу протеазы к сайту расщепления. Протеолитическое расщепление может происходить как во время синтеза новой вирусной частицы, так и после высвобождения вируса из клетки, а также перед проникновением вируса в новую клетку. Этот процесс может осуществляться различными клеточными протеазами. Например, сайт расщепления в линкерной последовательности гемагглютиниона высокопатогенного штамма вируса, ответственного за пандемию гриппа в Испании в 1918 году, включал только один аминокислотный остаток аргинина. Было показано, что две протеазы, которые активно экспрессируются в дыхательных путях человека, особенно в лёгких, TMPRSS2 и TMPRSS4, участвуют в расщеплении гемагглютиниона данного штамма вируса. Ещё одной протеазой, способной активировать гликопротеины вируса гриппа, является НАТ (Human airway trypsin-like protease) — трипсиноподобная протеаза дыхательных путей человека, которая экспрессируется, в основном, в гортани. Расщепление гликопротеинов некоторых штаммов вируса гриппа также было продемонстрировано для таких протеаз, как матриптаза, MSPL/MTPRSS13 [62]. Кроме того, частота расщепления пропротеин-конвертазами в определённых штаммах вируса гриппа увеличивалась при удалении сайтов гликозилирования вблизи позиции расщепления.

В активации гликопротеинов оболочки парамиксовирусов, включая ряд высокопатогенных представителей этой группы, протеазы также играют ключевую роль. У большинства парамиксовирусов гликопротеины подвергаются однократному протеолитическому расщеплению [63]. Так, пропротеин-конвертазы осуществляют расщепление гликопротеинов вирусов парагриппа, кори, а также вируса, вызывающего заболевание Ньюкасла. Респираторно-

синцитиальный вирус уникален среди парамиксовирусов тем, что его гликопротеин расщепляется пропротеин-конвертазами в двух местах: первое расщепление происходит до проникновения вируса в клетку, а второе — после эндоцитоза, в эндосомах. Вирусы Нипах и Хендра, которые вызывают инфекции, нередко заканчивающиеся смертельным исходом у людей, используют для активации своих гликопротеинов не пропротеин-конвертазы, а эндосомные катепсины [64].

Наиболее изученными вирусами семейства ретровирусов являются вирусы лейкоза крупного рогатого скота и мышей, которые относятся к Т-лимфотропным вирусам человека. Их гликопротеины оболочки расщепляются фурином и другими пропротеин-конвертазами [65]. Расщепление гликопротеинов вируса лейкоза происходит в два этапа. Пропропротеин-конвертазы осуществляют первое расщепление, которое запускает конформационные изменения, необходимые для дальнейшего протеолитического процессинга. Второе расщепление осуществляется протеазой самого вируса и приводит к полной активации гликопротеина. У вируса иммунодефицита человека HIV-1 предшественник гликопротеина gp160 расщепляется фурином в транс-сети Гольджи на две субъединицы — gp120 и gp41. После выхода из комплекса Гольджи gp120 далее расщепляется фурином на фрагменты gp77 и gp53. Было показано, что полиосновный участок, обогащённый положительно заряженными аминокислотными остатками и расположенный вблизи сайта расщепления пропротеин-конвертазами между gp120 и gp41, способен связывать гепарин и способствует расщеплению [66].

Гликопротеины оболочки филовирюсов, вызывающих смертельные геморрагические лихорадки, такие как вирусы Эбола и Марбург, подвергаются расщеплению фурином [67, 68]. В результате расщепления формируются две субъединицы, которые остаются связанными между собой дисульфидными связями. Гликопротеин большинства штаммов вируса Эбола имеет один сайт расщепления пропротеин-конвертазами, тогда как у гликопротеина вируса Марбург обнаружены два таких сайта, отличающиеся как по аминокислотной

последовательности, так и по положению относительно сайта расщепления в гликопротеине вируса Эбола. Кроме этого, вирус Эбола дополнительно использует эндосомные катепсины для активации своих белков оболочки [69]. Гликопротеин буньявируса Конго-Крымской геморрагической лихорадки также расщепляется фурином, однако его способна активировать и другая пропротеин-конвертаза — SKI-1 (S1P), относящаяся к пирролизиноподобному типу. Примечательно, что специфичность данной протеазы отличается от специфичности кексиноподобных пропротеин-конвертаз. Аналогичным образом, для активации гликопротеинов оболочки некоторых аренавирусов, включая вирус лимфоцитарного хориоменингита и вирус Ласса, также необходима активность SKI-1 [70].

Протеолиз также играет важную роль в жизненном цикле папилломавирусов, включая вирус папилломы человека (ВПЧ) [71]. Капсид ВПЧ состоит из двух белков — L1 и L2, которые претерпевают конформационные изменения при связывании вируса с протеогликанами на клеточной поверхности. Изменение конформации способствует расщеплению белка L2 внеклеточными и околочлеточными пропротеин-конвертазами. Протеолиз L2, в свою очередь, приводит к изменению конформации других вирусных белков, обеспечивая возможность взаимодействия вирусной частицы с дополнительными рецепторами. Это облегчает интернализацию вируса и способствует эффективному распространению инфекции.

Оболочка флавивирусов содержит два основных структурных белка — рgM и гликопротеин E — которые на ранних стадиях сборки вирусной частицы существуют в связанной форме. Созревание вируса сопровождается протеолитическим расщеплением белка рgM пропротеин-конвертазами, преимущественно при выходе вируса из клетки. Вирусы рода *Alphavirus*, такие как *Chikungunya*, *Semliki forest*, *Sindbis* и *Ross River* вирусы, являются арбовирусами и структурно родственны флавивирусам. Их основной гликопротеин оболочки также синтезируется в виде белка-предшественника, который затем расщепляется пропротеин-конвертазами. Это расщепление

регулирует взаимодействие гликопротеина с другими компонентами вирусной оболочки, что способствует слиянию вирусной частицы с мембраной клетки-хозяина [72].

Ещё одним семейством вирусов, для которых протеолиз играет ключевую роль в регуляции проникновения в клетку и созревания вирусных частиц, являются коронавирусы. Характерной особенностью этих вирусов является наличие шиповидного белка, называемого spike-белком, или S-белком, который расщепляется протеазами как во время биосинтеза новых вирусных частиц, так и при проникновении вируса в клетки-мишени [73]. S-белок относится к гликопротеинам слияния I класса. Гликопротеины коронавирусов расщепляются различными протеазами, включая фурин и трансмембранные сериновые протеазы семейства TMPRSS; кроме того, было показано, что некоторые лизосомные катепсины, эластаза и фактор свёртывания крови Ха также способны расщеплять S-белок [74]. Расщепление S-белка происходит последовательно по двум позициям. Первое расщепление происходит в месте соединения двух субъединиц гликопротеина — S1 и S2 — на этапе формирования новых вирусных частиц. Субъединица S1 обеспечивает связывание вируса с рецепторами клетки, в то время как субъединица S2 опосредует слияние мембран и проникновение вируса внутрь клетки. Первое расщепление способствует второму расщеплению, которое происходит на участке субъединицы S2. У некоторых коронавирусов первое расщепление может не происходить [75]. Способность коронавирусов адаптироваться к протеолитической среде инфицированных клеток, зависящая от специфичности и профилей экспрессии клеточных протеаз, определяет тропизм и патогенность вируса. Гликопротеин коронавируса MERS-CoV, вызывающего ближневосточный респираторный синдром, содержит оба сайта расщепления пропротеин-конвертазами, что, по-видимому, связано с его расширенным тканевым тропизмом по сравнению с другими коронавирусами. У коронавируса SARS-CoV, вызывающего острый респираторный синдром, не наблюдали активности фурина в отношении его гликопротеина на стадии сборки новых вирусных частиц. При этом протеаза TMPRSS2 способствует проникновению

обоих коронавирусов [76]. Также было показано, что проникновение SARS-CoV внутрь клетки может быть опосредовано действием катепсина L [77]. Расщепление S-белка SARS-CoV катепсином L предположительно происходит после связывания гликопротеина с клеточным рецептором, что открывает доступ к ранее скрытым сайтам, служащих мишенями для этой протеазы. Также была установлена положительная корреляция между чувствительностью S-белка к протеолизу со стороны пропротеин-конвертаз и других протеаз и количеством положительно заряженных аминокислотных остатков в области сайта расщепления.

1.4. Влияние протеолиза на жизненный цикл SARS-CoV-2

1.4.1. Жизненный цикл коронавируса SARS-CoV-2

SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) — это оболочечный одноцепочечный РНК-вирус положительной полярности из рода бетакоронавирусов. Он является возбудителем заболевания COVID-19, вызвавшего пандемию в 2020 году. На сегодняшний день охарактеризовано семь коронавирусов человека (HCoV), четыре из которых вызывают лёгкие респираторные инфекции (HCoV-229E, HCoV-NL63, HCoV-OC43 и HCoV-NKU1), а три (SARS-CoV, MERS-CoV и SARS-CoV-2) ассоциированы с тяжёлым течением заболевания. Вирион SARS-CoV-2 имеет диаметр около 60-100 нм и содержит одноцепочечную РНК длиной около 30 тыс. нуклеотидов, упакованную в комплекс с нуклеокапсидными белками (N). Нуклеокапсид окружён оболочкой, которая содержит белки мембраны (M), белки оболочки (E) и шиповидные белки (S) [78].

Проникновение коронавируса SARS-CoV-2 в клетки опосредовано S-белками. Шиповидные белки обеспечивают прикрепление вирусной частицы к

клеточной поверхности через взаимодействие с рецептором — ангиотензинконвертирующим ферментом ACE2. Далее проникновение коронавируса может развиваться по двум путям: так называемым раннему и позднему путям проникновения [79, 80]. Ранний путь проникновения реализуется путём слияния вирусной оболочки и клеточной мембраны, при этом формируется пора, через которую вирусный геном попадает в цитоплазму клетки. Процесс слияния вирусной и клеточной мембран инициируется протеолитическим расщеплением S-белка трансмембранной сериновой протеазой TMPRSS2. Поздний путь проникновения происходит за счёт клатрин-опосредованного эндоцитоза вирусной частицы, с последующим её попаданием в эндосому [81, 82]. Высвобождение вируса в цитоплазме клетки происходит на этапе поздней эндосомы и инициируется протеолитическим расщеплением цистеиновыми катепсинами, главным образом, катепсином L. Активация позднего пути проникновения зависит от активности протеазы TMPRSS2. После высвобождения нуклеокапсида вируса в цитоплазму происходит диссоциация N белков и вирусной РНК. Репликация генома происходит в специальных репликационных органеллах, которые обеспечивают оптимальную среду для репликации. Эти органеллы являются производными от эндоплазматического ретикулума и представляют собой околядерные взаимосвязанные мембранные структуры, содержащие двумембранные везикулы, в которых происходит синтез новых вирусных РНК. Синтезированные РНК покидают двумембранные везикулы через трансмембранные поры и попадают в промежуточный компартмент, образованный комплексом Гольджи и эндоплазматическим ретикулумом. В просветной части этого компартмента происходит сборка новых частиц вируса. При образовании новых вирусных частиц также происходит созревание вновь синтезированных S-белков путём протеолитического расщепления фурином или другими пропротеин-конвертазами. Завершив формирование, вирусные частицы транспортируются к клеточной поверхности и выделяются из клетки при помощи экзоцитоза [83].

1.4.2. Структура S-белка

Шиповидный белок, или S-белок, представляет собой гликопротеин, относящийся к белкам слияния I класса, главным представителем которых является гемагглютинин вируса гриппа. S-белок образует гомотример, каждый мономер которого состоит из двух субъединиц — S1 и S2. В S-белке незрелой вирусной частицы субъединицы соединены линкерной последовательностью, тогда как в S-белке зрелой вирусной частицы S1 и S2 субъединицы удерживаются вместе благодаря нековалентным взаимодействиям. Функциональные роли обеих субъединиц известны: субъединица S1 обеспечивает связывание с клеточным рецептором, субъединица S2 опосредует процесс слияния мембран. На поверхности S-гликопротеина преобладают гликаны: известно о 66 сайтах N-гликозилирования [84], а также о нескольких O-гликозилированных аминокислотных остатках [83]. Гликозилирование влияет на стабильность конформации S-белка, его чувствительность к расщеплению протеазами клетки-хозяина и способность уклоняться от иммунного ответа [83].

Субъединица S1 включает четыре домена — N-концевой домен (NTD), рецептор-связывающий домен (RBD), а также два C-концевых домена (CTD1 и CTD2). Три рецептор-связывающих домена образуют вершину S-белка и способны находиться в двух состояниях: «открытом», при котором S-белок может взаимодействовать с рецептором ACE2, и «закрытом», при котором связывание с рецептором невозможно. Известно, что N-гликозилированные позиции N165 и N234 стабилизируют открытую конформацию рецептор-связывающего домена, обеспечивая эффективное связывание S-белка с рецептором ACE2. O-гликозилированные позиции T323 и S325 также играют важную роль во взаимодействии вируса с рецептором [83]. Основным элементом субъединицы S2 является пептид слияния, который обеспечивает слияние вирусной оболочки и клеточной мембраны. Также в этой субъединице присутствуют гептадные повторы HR1 и HR2, которые участвуют в процессе слияния [85]. После слияния

мембран, субъединицы S1 и S2 S-белка разделяются, при этом субъединица S2 принимает форму удлинённого тримера, который пронизывает клеточную мембрану инфицируемой клетки, тем самым формируя пору.

1.4.3. Рецепторы SARS-CoV-2

Вирусный тропизм во многом определяется репертуаром рецепторов, с которыми вирус может взаимодействовать. SARS-CoV-2 проникает в организм преимущественно через дыхательные пути, но также обнаруживается в почках, печени, мозге, крови и кишечнике [86]. Основным рецептором SARS-CoV-2, с которым взаимодействует S-гликопротеин, является ангиотензинконвертирующий фермент ACE2. Первоначально ACE2 был идентифицирован в 2003 году как рецептор вируса SARS-CoV [87], а позднее и как рецептор для альфакоронавируса HCoV-NL63 [88]. Помимо рецепторной функции, ангиотензинконвертирующий фермент ACE2 является цинковой металлопротеазой и карбоксипептидазой, катализирующей образование биологически активных пептидов. Кроме того, ACE2 участвует в регуляции транспорта незаряженных аминокислот в кишечнике [89]. Рецепторы некоторых коронавирусов также являются белками с протеолитической активностью, как, например, CD26 (DPP4) у MERS-CoV и APN у HCoV-229E. Однако каталитическая активность этих рецепторов не влияет на способность вируса связываться с рецептором. ACE2 существует как в мембраносвязанном состоянии, так и в растворимой форме, высвобождаемой после расщепления протеазой ADAM17. Было показано, что растворимый ACE2 также может способствовать вирусной инфекции, образуя комплекс с S-белком SARS-CoV-2 и вазопрессином. Помимо взаимодействия с ACE2, S-гликопротеин коронавируса SARS-CoV-2 способен взаимодействовать с гепарансульфатом на поверхности клеток, облегчая первоначальное прикрепление вирусной частицы к клеточной поверхности и увеличивая вероятность последующего взаимодействия с основным рецептором.

Альтернативными рецепторами для SARS-CoV-2 могут выступать лектины С-типа, включая DC-SIGN и L-SIG, которые могут способствовать транспортировке вируса между клетками-мишенями (транс-инфекция). Также в качестве альтернативных рецепторов рассматривают белки TIM1 и AXL [79].

1.4.4. Протеолиз S-белка SARS-CoV-2

S-гликопротеин коронавируса SARS-CoV-2 подвергается протеолитическому расщеплению только со стороны протеаз инфицируемых клеток. В S-белке выделяют два ключевых сайта расщепления. Первый из них — фуриновый сайт расщепления, способствующий созреванию S-гликопротеина, расположен в линкерной петле, соединяющей две субъединицы S-белка в незрелой вирусной частице. Этот участок обозначается как S1/S2 сайт и соответствует позиции R685 в аминокислотной последовательности S-гликопротеина. Расщепление S1/S2-сайта способствует переходу рецептор-связывающего домена S-белка в открытое состояние, необходимое для взаимодействия с ACE2-рецептором. Причём расщепление лишь одного протомера и его взаимодействие с ACE2 может способствовать последовательному переходу в открытое состояние рецептор-связывающих доменов остальных протомеров и их взаимодействие с ACE2-рецепторами [90, 91].

Отличительной особенностью S1/S2 сайта расщепления у коронавируса SARS-CoV-2 является наличие вставки четырёх аминокислотных остатков 681-PRRA-684. Этот аминокислотный мотив отличает SARS-CoV-2 от SARS-CoV и других сарбековиров (подрод бетакоронавирусов, связывающихся с ACE2 рецептором), S-белок которых не подвергается расщеплению со стороны фурина и других пропротеин-конвертаз во время созревания вируса в инфицированной клетке. Однако фуриновый мотив встречается у других вирусов. Например, его наличие коррелирует с повышенной вирулентностью вирусов птичьего гриппа А

и в целом является характерной чертой высокопатогенных вирусов птичьего гриппа. Однако наличие фуринового мотива не обязательно коррелирует с высокой вирулентностью у коронавирусов: сезонные коронавирусы человека (HCoV-NKU1 и HCoV-OC43), имеющие фуриновый мотив на участке S1/S2 сайта расщепления, менее вирулентны, чем SARS-CoV или SARS-CoV-2. Тем не менее, предполагается, что такая вставка аминокислотных остатков могла способствовать адаптации SARS-CoV-2 к человеку. Также на эффективность расщепления фуринового мотива влияет длина линкерной петли, определяющая доступность сайта расщепления для протеазы [90].

Основной протеазой, задействованной в расщеплении S1/S2 сайта, является фурин, относящийся к семейству пропротеин-конвертаз. Фурин экспрессируется почти во всех тканях и локализуется в компартментах, связанных с биосинтезом и созреванием новых белков, таких как аппарат Гольджи, а также на поверхности клеток. Фурин существует в двух формах: мембранной и секретируемой [79]. Расщепление S1/S2 сайта может происходить даже при дефиците фурина, предположительно с участием других пропротеин-конвертаз [92].

Второй сайт расщепления — S2' — располагается в субъединице S2 недалеко от пептида слияния и соответствует позиции R815 в аминокислотной последовательности S-белка. Расщепление сайта S2' запускает конформационные изменения в S-белке, высвобождение пептида слияния и последующее слияние вирусной оболочки и клеточной мембраны с образованием поры. Важно отметить, что сайт расщепления S2' становится доступным для протеолитического воздействия только после связывания S-белка с ACE2 рецептором, что, в свою очередь, зависит от расщепления сайта S1/S2. Основная протеаза, участвующая в расщеплении S2' сайта — трансмембранная сериновая протеаза TMPRSS2. Эта протеаза играет важную роль в респираторных вирусных инфекциях, включая вирус гриппа и коронавирусы SARS. Протеаза TMPRSS2 экспрессируется в тканях предстательной железы, желудочно-кишечного тракта, дыхательных путей, печени, почек, поджелудочной железы, слюнных желёз и тимусе [93]. Три основных типа клеток совместно экспрессируют TMPRSS2 и ACE2 —

пневмоциты II типа, абсорбционные энтероциты подвздошной кишки и секреторные клетки носовой полости. Предполагается, что в расщеплении S-гликопротеина коронавируса SARS-CoV-2 могут быть задействованы и другие протеазы, например, трипсиноподобная протеаза дыхательных путей человека (HAT), TMPRSS4, TMPRSS11A, TMPRSS11E, матриптаза и нейтрофильная эластаза [85]. Известно, что протеазы TMPRSS13 и TMPRSS11D способны активировать S-белок SARS-CoV-2, но с меньшей эффективностью по сравнению с TMPRSS2. В частности, протеаза TMPRSS4 может расщеплять S-белок коронавируса SARS-CoV-2 в эпителии кишечника [78].

При эндоцитарном (позднем) пути проникновения S-гликопротеин расщепляется цистеиновыми катепсинами, в частности, катепсином L. Катепсины относятся к семейству протеаз, ответственных за переработку клеточных белков внутри лизосом. Различают цистеиновые, аспартатные и сериновые катепсины. Катепсин L является цистеиновой протеазой, которая локализуется в эндосомах и лизосомах и экспрессируется в большинстве типов тканей. Существует также секретлируемая форма катепсина L [85]. Позиция расщепления катепсинами, включая катепсин L, до сих пор не установлена. Предполагается, что катепсин L расщепляет S-белок как вблизи сайта расщепления S1/S2, так и недалеко от пептида слияния, рядом с сайтом S2'. Предполагаемые сайты расщепления катепсиновыми протеазами были предложены в результате анализа *in silico* [94]. Протеолитическое расщепление S-белка катепсином L инициирует выход коронавируса SARS-CoV-2 из эндосомы в цитоплазму клетки.

1.4.5. Влияние аминокислотных замен на протеолиз S-белка

Наиболее известными штаммами коронавируса SARS-CoV-2 являются штаммы Альфа (B.1.1.7), Бета (B.1.351), Гамма (P.1), Дельта (B.1.617.2) и Омикрон (B.1.1.529). Установлено, что все пять штаммов обладают повышенной трансмиссивностью по сравнению с исходным, «диким типом» коронавируса

SARS-CoV-2, выявленным во время вспышки в Ухане в начале 2020 года. Каждый штамм характеризуется своим набором аминокислотных замен в S-белке. Аминокислотные замены затрагивают различные участки S-гликопротеина, наиболее значимыми из которых считают аминокислотные замены в рецептор-связывающем домене, N-концевом домене, вблизи фуринового сайта расщепления и в S2-субъединице [95]. Аминокислотные замены в S-белке могут влиять на тропизм, вирулентность или патогенность коронавируса SARS-CoV-2.

Например, аминокислотные замены G339D, N440K, L452R, S477N, T478K, E484K, N501Y, затрагивающие рецептор-связывающий домен, ассоциированы с повышенным сродством между S-белком и ACE2 рецептором. Большая часть замен в рецептор-связывающем домене выявлено в S-белке коронавируса штамма Омикрон [95]. Аминокислотная замена N501Y также встречается в S-белках штаммов Альфа, Бета и Гамма и является одной из ключевых в рецептор-связывающем домене S-белка для усиления взаимодействия S-гликопротеина с ACE2 рецептором, что увеличивает инфекционность и вирулентность данных штаммов коронавируса. S-белки штаммов Бета и Гамма SARS-CoV-2 имеют аминокислотные замены K417N/T и E484K, которые влияют на связывание с рецептором ACE2. В S-гликопротеине штамма Дельта наблюдаются замены L452R и T478K, которые также увеличивают сродство с ACE2. Большая часть аминокислотных замен в позициях, локализованных в N-концевом домене, ассоциирована со штаммом Омикрон SARS-CoV-2, который появился после широкого распространения вакцинации. Предполагается, что эти замены способствуют устойчивости вируса к нейтрализующим антителам, связывающимся с N-концевым доменом S-белка. Замена D614G является наиболее распространённой и встречается в S-белках всех вышеперечисленных штаммов. Эта мутация стабилизирует взаимодействие между субъединицами S1 и S2 и усиливает инфекционность вируса. Также показано, что рецептор-связывающий домен S-белка, который содержит аминокислотную замену D614G, с большей вероятностью принимает открытую конформацию благодаря повышению эффективности расщепления S1/S2 сайта. Аминокислотные замены в

681-й позиции влияют на эффективность расщепления S1/S2 сайта. Например, замена P681R, зафиксированная у штамма Дельта коронавируса SARS-CoV-2, увеличивает эффективность расщепления фуринового сайта и, как следствие, повышает вирулентность вируса. Замена P681H, характерная для штаммов Альфа и Омикрон, также усиливает расщепление S1/S2 сайта, однако её влияние на инфекционность вируса оказалось несущественным [95]. Кроме того, аминокислотные замены в 681 позиции могут устранять О-гликозилирование аминокислотного остатка, что, в свою очередь, дополнительно способствует более эффективному расщеплению сайта S1/S2 фурином [96]. Предполагается, что замена A570D, наблюдаемая в S-белке штамма Альфа, также способна оказывать влияние на эффективность расщепления фуринового сайта. Несмотря на удалённость 570-й позиции от сайта S1/S2, данная замена приводит к увеличению расстояния между цепями S-белка вследствие изменения заряда и гидрофобности, что потенциально повышает доступность сайта для фурина [97].

1.5. Биоинформатическое прогнозирование протеолитических сайтов

1.5.1. Специфичность протеазы

Специфичность протеазы — это способность протеолитического фермента избирательно распознавать и расщеплять определённые участки в полипептидной цепи субстрата. Специфичность протеазы определяется структурой её активного центра. Активный центр любого фермента выполняет две ключевые функции: связывает субстрат и катализирует специфическую реакцию. Важной особенностью активного центра является его пространственная организация и размер [21].

Активный центр протеазы рассматривают как совокупность субсайтов, каждый из которых, как правило, взаимодействует с одним аминокислотным

остатком субстрата. Субсайты активного центра протеазы, расположенные по разные стороны от расщепляемой пептидной связи, обозначают как S и S', а соответствующие им аминокислотные остатки субстрата — как P и P'. И субсайты, и аминокислотные остатки нумеруются. Нумерация ведётся от расщепляемой пептидной связи — она располагается между N-концевым и C-концевым аминокислотными остатками P1 и P1', соответственно. Остальные аминокислотные остатки и субсайты нумеруются в направлении к N- и C-концам полипептидной цепи: "...—P3—P2—P1—P1'—P2'—P3'—...". Пример записи для субсайтов: "...—S3—S2—S1—S1'—S2'—S3'—..." [21, 98] (Рис. 1).

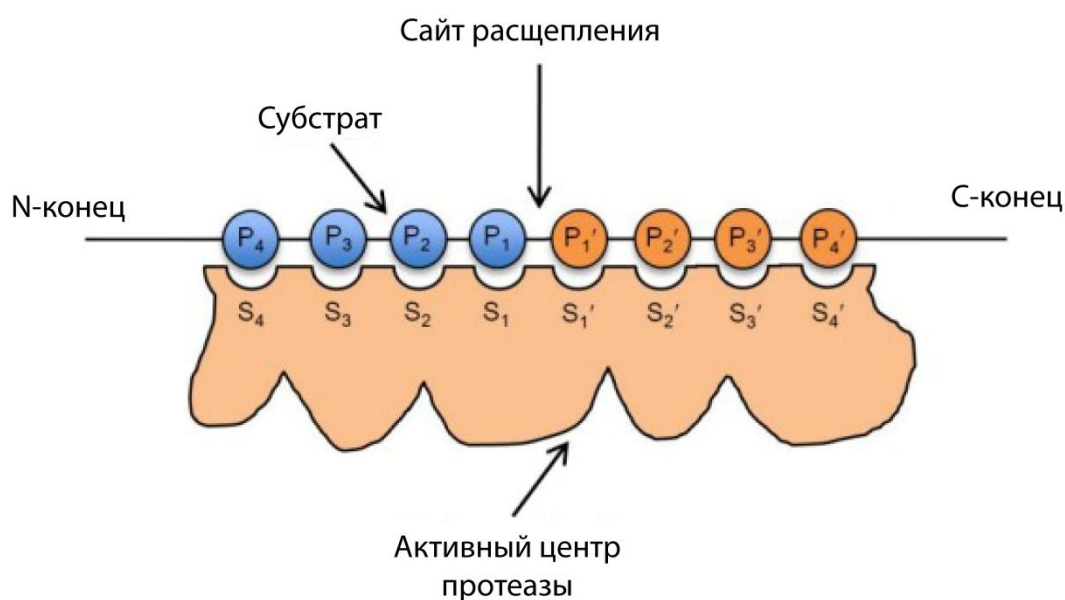


Рисунок 1. Представление сайта расщепления при взаимодействии активного центра протеазы с субстратом, согласно [21]. Аминокислотные остатки субстрата обозначаются "P", соответствующие субсайты активного центра протеазы — "S". Нумерация позиций задаётся относительно сайта расщепления. Рисунок адаптирован из [9].

Субсайты активного центра протеазы различаются по своей способности связывать аминокислотные остатки субстрата: одни субсайты проявляют строгое предпочтение к определённым аминокислотным остаткам, тогда как другие субсайты демонстрируют более широкий диапазон взаимодействий. Эти особенности и лежат в основе субстратной специфичности протеазы. Взаимодействие протеазы с субстратом должно быть достаточно прочным для обеспечения эффективного катализа, но при этом не чрезмерно сильным, чтобы не препятствовать прогрессу реакции [98]. Важно отметить, что специфичность

протеазы определяется не только субсайтами, взаимодействующими с остатками P1 и P1', но и отдалёнными субсайтами активного центра протеазы, которые тоже вносят вклад в общую конфигурацию комплекса фермент-субстрат [21, 99].

Специфичность протеаз не ограничивается механизмом катализа. Эволюционно родственные протеолитические ферменты, обладающие общими каталитическими механизмами, могут подразделяться на специфические и неспецифические подгруппы. Например, у цистеиновых протеаз выделяют два клана CD и CA. Каспазы являются высокоспецифичными протеазами и относятся к клану CD, в то время как кальпаины и катепсины, проявляющие широкую субстратную специфичность, относятся к клану CA. Ещё один пример эволюционно близких протеаз, отличающихся по диапазону специфичности — плазмин и трипсин [99].

1.5.2. Детерминанты ограниченного протеолиза

Ограниченный протеолиз представляет собой высокоспецифичный процесс, регулируемый рядом факторов, от которых зависит восприимчивость белков к расщеплению. Выделяют четыре основных детерминанты, влияющих на восприимчивость субстратов к ограниченному протеолизу [17, 100]. Во-первых, необходима пространственно-временная колокализация: протеаза и её субстрат должны одновременно присутствовать в одном и том же субклеточном компартменте. Во-вторых, протеолиз требует наличие специфического аминокислотного мотива, соответствующего субсайтам активного центра фермента. Только при наличии определённой последовательности аминокислот возможно стабильное связывание фермента с субстратом, необходимое для катализируемой реакции. В-третьих, структурный контекст сайта расщепления оказывает существенное влияние на вероятность протеолиза. Даже при наличии аминокислотного мотива, специфичного для той или иной протеазы, протеолиз не произойдёт, если этот мотив будет физически недоступен для контакта с

активным центром протеолитического фермента. Важно отметить, что структурная доступность является универсальной детерминантой протеолитических ферментов, не зависящей от их специфичности [17]. В-четвёртых, взаимодействие протеазы с аллостерическими сайтами, или экзосайтами, то есть сайтами, которые удалены от сайта расщепления, также влияет на вероятность протеолиза. Эта детерминанта характерна не для всех протеаз. Аллостерическое взаимодействие протеазы с субстратом наблюдалось, например, для тромбина и фибриногена [100].

Данные о специфичности протеазы, а также о структурных особенностях участков субстрата вблизи сайтов расщеплений, активно используются для предсказания потенциальных протеолитических сайтов в молекулах субстратов. Наиболее значимыми с точки зрения прогностической способности являются структурные характеристики, описывающие доступность участков молекулы для растворителя, подвижность участков субстрата, их вторичную структуру и другие параметры [17]. Было показано, что протеолитическое расщепление чаще происходит в доступных и подвижных неупорядоченных петлевых участках, и реже — в областях с устойчивой вторичной структурой, таких как альфа-спирали и бета-листы. Если протеолиз всё же затрагивает бета-структуры, то наиболее уязвимыми оказываются либо края бета-стрэндов, либо периферийные бета-стрэнды. Вероятно, это связано с тем, что аминокислотные остатки, локализованные по периметру бета-листа, более доступны физически и имеют более низкую энергию водородной связи по сравнению с аминокислотными остатками в центральных участках бета-листа. Тем не менее, известны протеолитические события, затрагивающие внутренние бета-стрэнды, однако они в таких случаях, как правило, примыкают к N- или C-концевым бета-стрэндам. Влияние вторичной структуры на уязвимость к протеолизу до сих пор остаётся предметом дискуссий. В целом, предполагается, что примерно одна треть всех пептидных связей в белке имеет уязвимость к протеолитическому расщеплению [17].

Кроме этого, остаётся предметом дискуссий вопрос о возможных локальных конформационных изменениях расщепляемого участка перед связыванием с протеазой. В частности, неизвестно, необходимо ли предварительное распрямление альфа-спиралей, или они могут быть расщеплены в исходной конформации. С одной стороны, некоторые протеазы, такие как стафилококковая протеаза GluC, обладают крупным активным центром, способным вместить альфа-спираль без её дестабилизации. С другой стороны, у ферментов с компактным активным центром, таких как каспаза-3, предполагается либо избирательное связывание линейных участков, либо частичное распрямление коротких спиралей, расположенных внутри петель [100]. Альтернативная гипотеза предполагает, что участки расщепления могут принимать конформации вытянутых бета-стрэндов, способствующие эффективному взаимодействию с активным центром. Это объясняется тем, что линейная форма участка расщепления обеспечивает более благоприятные условия для размещения в компактном активном центре протеазы и формирования межмолекулярных взаимодействий [98].

Особые механизмы протеолиза наблюдаются при шеддинге — процессе высвобождения внеклеточных (эктодоменных) фрагментов трансмембранных белков. Известно, что большинство шеддаз — протеаз, осуществляющих шеддинг эктодоменов — имеют нестрогую субстратную специфичность, и в этом случае ключевым фактором является фиксированное расстояние от поверхности мембраны в так называемом околосмембранном домене белка. Было показано, что делеции или вставки вблизи таких сайтов лишь смещают позицию расщепления, но не ингибируют шеддинг. В некоторых случаях для осуществления шеддинга необходимо, чтобы сайт расщепления находился в области с альфа-спиральной структурой. При этом установлено, что шеддинг может ингибироваться посттрансляционными модификациями, такими как О-гликозилирование [101], или присутствием отрицательно заряженных аминокислотных остатков, препятствующих связыванию с шеддазами и их кофакторами [102].

1.5.3. Базы данных и инструменты прогнозирования протеолитических событий

В настоящее время уже существуют несколько инструментов прогнозирования протеолитических событий. Данные методы решают задачи идентификации потенциальных субстратов протеаз и выявления ранее неизвестных сайтов расщеплений. При работе данные методы используют информацию о белке — потенциальном субстрате протеазы. Как правило, существующие инструменты прогнозирования событий протеолиза моделируют специфичность протеолитических ферментов по отношению к аминокислотной последовательности субстратов. Подавляющее большинство инструментов поиска новых субстратов и сайтов расщеплений являются протеазо-специфическими — они генерируют предсказания протеолитических событий, связанных с конкретными протеазами. Существуют как узкоспециализированные методы, предсказывающие сайты расщеплений одной конкретной протеазы [103, 104] или семейства протеаз [105], так и более универсальные инструменты, моделирующие специфичность разнообразных протеаз. В одном из последних исследований была предложена не отдельная модель, а целая платформа, которая позволяла пользователю самостоятельно создать собственную модель прогнозирования или воспользоваться уже готовыми моделями [106]. Некоторые инструменты имеют конкретное практическое применение: например, инструмент DeepDigest был разработан для того, чтобы способствовать более качественному разделению пептидов при выполнении масс-спектрометрического анализа протеомов [107]. Разработка протеазо-специфических инструментов прогнозирования обусловлена как разнообразием механизмов катализа реакции протеолиза, так и высокой вариабельностью субстратной специфичности протеаз [108]. В связи с этим универсальный инструмент прогнозирования протеолитических событий до сих пор не создан. В исследовании, посвящённом разработке инструмента прогнозирования протеолитических событий Pripper [109], предсказанные субстраты — белки AIF-1 и SMN1 — были валидированы экспериментально, что

подтверждает практическую ценность биоинформатических подходов и их значимость в планировании дальнейших исследований. Таким образом, биоинформатические методы не только дополняют, но и направляют экспериментальную работу.

Проблема идентификации новых субстратов и сайтов расщеплений важна с различных точек зрения — начиная от углубленного понимания биологии протеаз и механизмов их взаимодействия с субстратами до разработки новых ингибиторов протеаз, а также идентификации потенциальных терапевтических мишеней. Существенный вклад в решение этой задачи вносят экспериментальные методы. Более того, именно накопление данных, полученных в ходе экспериментальных исследований, сделало возможным разработку вычислительных подходов для поиска потенциальных субстратов протеаз и ранее неизвестных сайтов протеолиза. Все существующие на сегодняшний день биоинформатические инструменты поиска новых субстратов и сайтов расщеплений основаны на доступной информации о протеолитических событиях, значительная часть которой аккумулирована в специализированных базах данных. Наиболее известной и объемной является база данных MEROPS, в которой собраны сведения о протеазах, их субстратах, ингибиторах и сайтах протеолиза, охватывающие широкий спектр организмов [19]. CutDB [18] и Degradome [110] — базы данных, в которых сосредоточена информация о протеазах млекопитающих, причём в базе данных Degradome сделан акцент на медицинскую значимость некоторых протеаз. Такие базы данных, как CamPDB [111] и CASBAN [112] содержат информацию о протеолитических событиях, ассоциированных с отдельными семействами протеаз, такими как кальпаины и каспазы, соответственно.

Инструменты прогнозирования протеолитических событий можно условно разделить на две основные группы. Первую группу составляют методы, основанные на оценке схожести аминокислотных последовательностей. Ко второй группе относятся инструменты, основанные на алгоритмах машинного обучения. Методы обеих групп основываются на доступной информации о

протеолитических событиях с экспериментально подтверждёнными позициями расщеплений [16].

Инструменты первой группы исходят из предположения, что схожие аминокислотные последовательности обладают аналогичными биологическими свойствами, в том числе в контексте узнавания и расщепления протеазами. Предсказание осуществляется путём сравнения исследуемой последовательности с известными субстратами, в которых были зафиксированы позиции расщепления конкретных протеаз. Результаты такого сравнения выражаются в виде оценок, отражающих степень сходства: чем выше оценка, тем выше вероятность того, что соответствующая позиция является потенциальным сайтом расщепления. Различие между инструментами первой группы заключается в подходах к подсчёту данных оценок. В качестве математической основы инструментов первой группы используются такие методы, как позиционно-весовые матрицы, индексы, частоты встречаемости, оценки консервативности и их комбинации, которые позволяют моделировать субстратную специфичность протеолитического фермента. Главным преимуществом таких инструментов является скорость получения результатов [16].

Инструменты второй группы основаны на алгоритмах машинного обучения и, как правило, обеспечивают более высокую точность предсказания. Подобно первой группе инструментов, они моделируют специфичность протеаз, но делают это на основе принципов машинного обучения. Построение моделей машинного обучения включает несколько этапов, основными из которых являются сбор данных, выбор алгоритма, обучение модели и её валидация. Задача поиска новых сайтов расщеплений с точки зрения машинного обучения относится к задаче бинарной классификации. Модель обучается на размеченных данных, в которых сайты расщепления обозначены как положительные примеры, а остальные участки субстрата — как отрицательные. Задачи подобного рода, как правило, решаются при помощи алгоритмов обучения с учителем, поэтому набор данных разбивается на обучающую и тестовую выборки, предпочтительно без пересечения между ними. Обучение модели подразумевает поиск при помощи

выбранного алгоритма машинного обучения определённых закономерностей в данных, которые позволят отделить сайты расщеплений от остальных участков субстрата. Для улучшения качества предсказания в обучающие данные могут быть добавлены различные признаки, характеризующие субстраты. Большинство разработанных моделей опирается на характеристики, полученные на основе аминокислотной последовательности субстрата: типы аминокислотных остатков, их физико-химические свойства, PEST-последовательности и другие. Некоторые современные модели прогнозирования протеолитических событий дополнительно используют пространственные характеристики, если структура белка известна. Известно, что специфичность некоторых протеаз лучше моделируется с использованием определённых типов признаков, например, PEST-последовательностей [113]. PEST-последовательности — это участки полипептидной цепи, обогащённые пролином, глутаминовой кислотой, серином и треонином. Известно, что PEST-последовательности вблизи сайтов расщеплений могут быть ответственны за распознавание последних кальпаинами и протеасомой. Увеличение числа и разнообразия информативных признаков, а также их комбинирование, значительно повысило качество предсказаний.

Для решения задачи идентификации новых субстратов и сайтов расщеплений широко используются такие алгоритмы, как метод опорных векторов, деревья решений, метод ближайших соседей, наивный Байесовский алгоритм, скрытые марковские модели. Качество прогнозирования моделей оценивается преимущественно с использованием независимой тестовой выборки. Однако при ограниченном наборе данных применяют методы кросс-валидации, в рамках которой исходный набор данных делят на определённое количество частей: для тестирования выбирается одна из частей данных, а сама модель обучается на оставшихся частях. Этот процесс повторяется циклически, пока качество модели не будет проверено на каждой из частей исходного набора данных. Существуют разновидности методов кросс-валидации, наиболее надёжной из которых считается метод Leave-One-Out, при котором каждый пример из исходного набора данных используется в качестве тестового, а все

остальные примеры — для обучения. Для проверки качества моделей используют различные метрики, основанные на учёте ошибок первого и второго рода. Наиболее распространёнными метриками являются чувствительность, специфичность, точность, F-мера, площадь под ROC-кривой (AUC) [114], коэффициент корреляции Мэтьюса. Каждая из них обладает своими достоинствами и недостатками. Например, коэффициент корреляции Мэтьюса устойчив к дисбалансу классов, тогда как F-мера более чувствительна к различиям в размерах положительного и отрицательного классов. Площадь под ROC-кривой не зависит от порогового значения, поэтому часто используется как универсальный показатель качества бинарной классификации.

При построении моделей машинного обучения для прогнозирования протеолитических событий необходимо учитывать ряд методологических и технических сложностей, способных негативно повлиять на качество генерируемых прогнозов. Например, с увеличением пространства признаков могут возникнуть проблемы переобучения моделей и усиления шума в данных. Для снижения этих эффектов применяют методы отбора информативных признаков при помощи различных алгоритмов, проектирования комплексных или интегральных признаков, а также тщательно настраивают параметры модели машинного обучения. Альтернативный подход к решению данной проблемы заключается в использовании алгоритмов глубокого обучения. В отличие от традиционных методов, требующих явного задания признаков, глубокие нейросетевые модели способны самостоятельно извлекать информативные представления из исходных данных, например, непосредственно из аминокислотных последовательностей [115].

Ещё одной серьёзной проблемой является дисбаланс между количеством наблюдений положительного и отрицательного классов в обучающей выборке: число стабильных пептидных связей в белке значительно превышает число реально расщепляемых. Для решения этой проблемы применяют различные стратегии: балансировку выборки путём подбора сравнимого числа положительных и отрицательных примеров, использование весов,

компенсирующих недопредставленность одного из классов, или, наоборот, штрафов для избыточно представленного класса. В некоторых случаях источником ошибок служит избыточность данных, которая проявляется в виде утечки информации между обучающей и тестовой выборками. Такая ситуация возникает, когда в выборках присутствуют сильно похожие последовательности субстратов. Для предотвращения этого эффекта при формировании обучающей и тестовой выборок применяют пороговые значения сходства и специальные фильтры.

Отдельно выделяют проблему точности прогнозирования сайтов протеолиза конкретных протеаз. Например, известно, что сайты расщепления тромбина или некоторых матриксных металлопротеаз довольно сложно точно предсказать. В первую очередь, это связывают с ограниченностью обучающих наборов данных. Предполагается, что с увеличением количества накопленных данных точность таких предсказаний будет возрастать. В некоторых случаях выявляется предвзятость генерируемых прогнозов, связанная с тем, что обучающая выборка основана на данных, полученных лишь с использованием одного экспериментального подхода. В связи с этим рекомендуется формировать обучающие наборы из данных, полученных с применением различных методов, чтобы обеспечить более полную представленность закономерностей протеолитического процесса. Наконец, подчёркивается важность адекватного выбора алгоритма: во многих случаях более простые алгоритмы обеспечивают достаточную предсказательную мощность, не уступая по эффективности более сложным нелинейным алгоритмам [116].

Существуют различные перспективные направления развития инструментов биоинформатического прогнозирования протеолитических событий. Некоторые авторы полагают, что улучшения качества генерируемых прогнозов можно добиться за счёт использования более сложных алгоритмов, таких как свёрточные и рекуррентные нейронные сети, большие языковые модели, или за счёт применения ансамблевых методов, объединяющих прогнозы, генерируемые различными алгоритмами и моделями — в совокупности с облачными и

распараллеленными вычислениями. Другие исследователи возлагают надежды на возможности использования специализированных подходов, таких как обучение с позитивными и немаркированными примерами (Positive Unlabeled Learning) [16]. Отдельное направление связано с разработкой моделей, способных учитывать кинетику протеолитических реакций, особенности реакционной среды (например, pH и клеточную локализацию), а также механизмы аллостерической регуляции [9].

ГЛАВА II. РАЗРАБОТКА МЕТОДА ПРОГНОЗИРОВАНИЯ ПРОТЕОЛИТИЧЕСКИХ СОБЫТИЙ

2.1. Построение модели предсказания уязвимости участков белка к протеолизу на основе информации о структурных особенностях субстратов

2.1.1. Подготовка набора данных для построения предсказательной модели

Для построения модели, предсказывающей восприимчивость участков белка к протеолизу, на первом этапе работы были подготовлены данные, содержащие информацию об экспериментально подтверждённых протеолитических событиях и структурных особенностях субстратов. Информация об экспериментально подтверждённых событиях протеолиза была получена из базы данных CutDB [18]. Каждому протеолитическому событию соответствовали три атрибута: идентификатор субстрата, позиция сайта расщепления в последовательности субстрата и идентификатор протеазы. В общей сложности были собраны данные о 4576 протеолитических событиях, относящиеся к 2062 уникальным субстратам и 457 уникальным протеазам.

Далее был выполнен поиск известных трёхмерных структур субстратов в базе данных PDB [117] на основе их аминокислотных последовательностей с использованием инструмента BLAST [118]. К результатам поиска применялись два критерия фильтрации: 1) процент схожести аминокислотных последовательностей структуры и субстрата не менее 90%; 2) длина выравнивания аминокислотных последовательностей структуры и субстрата более $\frac{2}{3}$ от длины каждой из последовательностей. Для каждого рассматриваемого субстрата выбиралась одна трёхмерная структура, удовлетворяющая критериям фильтрации. В случаях, когда не удавалось найти структуру, удовлетворяющую указанным критериям, рассматриваемый субстрат

исключали из анализа. В результате было отобрано 585 трёхмерных структур субстратов, на долю которых приходилось 1499 протеолитических событий, индуцированных 256 уникальными протеазами.

На следующем этапе сайты расщеплений были отображены на трёхмерных структурах соответствующих субстратов. Для этого выполнялось дополнительное выравнивание исходной аминокислотной последовательности субстрата и последовательности, соответствующей трёхмерной структуре, при помощи инструмента Clustal Omega [119]. Визуализация сайтов расщеплений на трёхмерных структурах была выполнена с помощью программы UCSF Chimera [120]. В результате на 323 структурах удалось отобразить 777 протеолитических сайтов, соответствующих активности 183 различных протеаз. Общее число сайтов протеолиза после этапа картирования сократилось, поскольку позиции многих сайтов соответствовали незакристаллизованным участкам структуры субстрата.

После этапа картирования было проведено визуальное курирование позиций сайтов расщеплений, отображённых на трёхмерных структурах субстратов. Этот этап был необходим, поскольку в данных ряда экспериментов наблюдалось значительное количество сайтов, локализованных во внутренних, гидрофобных областях глобулы белка, которые, с высокой степенью вероятности, недоступны для взаимодействия с протеазой. Подобное может наблюдаться в экспериментах, в которых детектирование сайтов расщеплений происходило после утраты субстратом нативной конформации. Если при курировании данных конкретного эксперимента значительная часть сайтов расщепления локализовалась в гидрофобном ядре белка, все протеолитические события, относящиеся к данному эксперименту, исключались из рассмотрения. В результате курирования было отобрано 445 протеолитических событий, соответствующих 130 протеазам и 190 трёхмерным структурам субстратов. Среднее количество сайтов протеолиза на один субстрат составило 2.34, а медианное количество — 1.

Далее, для формирования обучающего множества были сгенерированы информативные структурные признаки субстратов (Рис. 2). В данном исследовании были выбраны структурные характеристики, для которых в ранних исследованиях была найдена позиционная корреляция с уязвимостью к протеолизу [17, 121-126], а именно: доступность для растворителя (solvent accessibility, ACC), тип вторичной структуры (secondary structure, SS) и температурный фактор (B-factor). Кроме этого, основываясь на наблюдениях, представленных в литературе [17], свидетельствующих о повышенной плотности сайтов протеолитических расщеплений в длинных выступающих петлях, а также на N- и C-концах полипептидной цепи субстратов, были введены два дополнительных структурных признака. Первый из них — длина петли (loop length), отражающая протяжённость неупорядоченных участков. Вторым признаком характеризует принадлежность остатка к N- или C-концам полипептида и имеет бинарное представление. Значения структурных признаков доступность для растворителя и типа вторичной структуры были получены при помощи метода DSSP [127]. Значения температурного фактора были извлечены непосредственно из PDB-файлов трёхмерных структур субстратов. Далее численные значения структурных признаков — доступности для растворителя, температурного фактора и длины петли — были нормированы на интервал [0, 1]. Значения структурного признака «тип вторичной структуры» были преобразованы в бинарный формат при помощи преобразования одноразрядного кодирования (one-hot encoding).

Поскольку ранее было показано, что уязвимость пептидной связи к протеолизу главным образом определяется структурными характеристиками субсайта P1 [17, 98, 100], все структурные признаки для каждой анализируемой позиции пептидной связи вычислялись именно для данного субсайта. Таким образом, итоговый набор данных был представлен в виде таблицы, включающей набор признаков, характеризующих различные структурные особенности субстратов, а также целевой бинарной переменной, отражающей позиции расщепления в субстрате. Данная таблица содержала 69285 наблюдений,

соответствующих пептидным связям в трёхмерных структурах субстратов, из которых 445 представляли собой сайты расщепления.

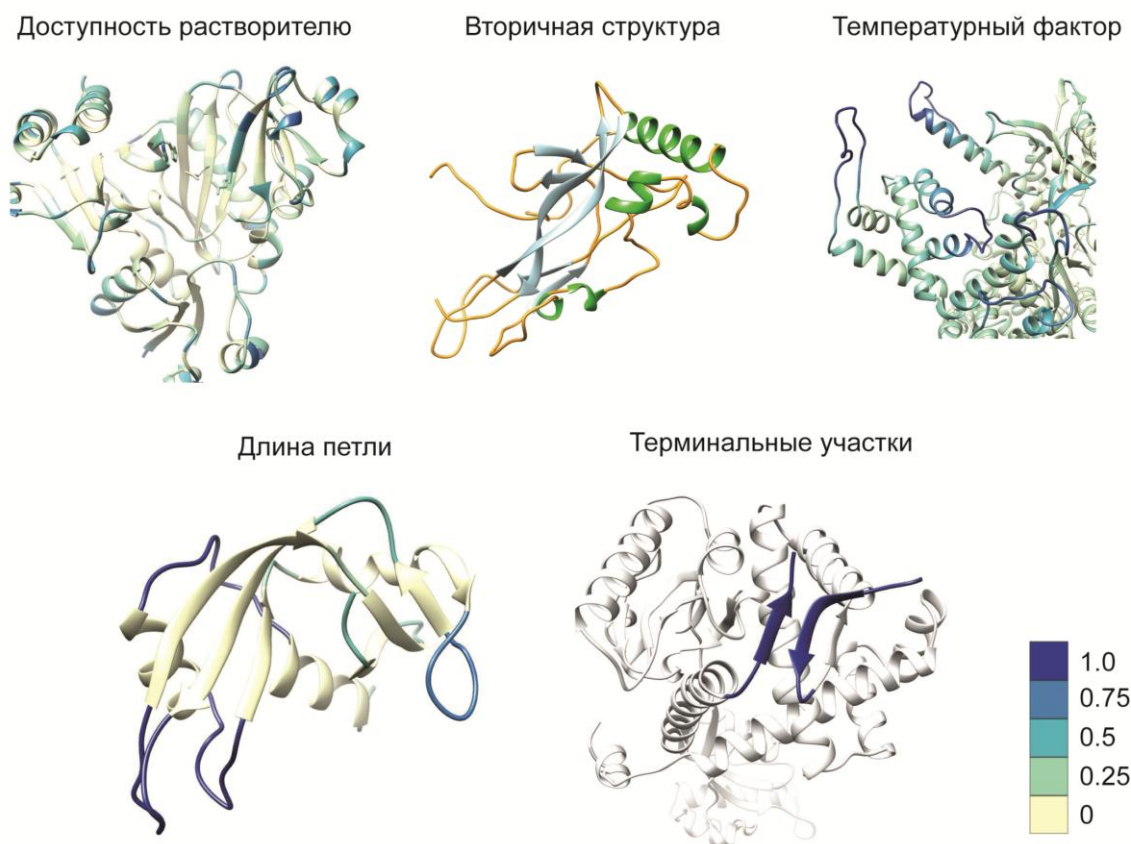


Рисунок 2. Визуализация признаков, использованных для обучения предсказательной модели, на трёхмерных структурах субстратов. Цветовая шкала из пяти градаций с диапазоном значений от 0 до 1 отображает нормализованные значения структурных признаков «Доступность растворителю», «Температурный фактор» и «Длина петли». Цветовая схема вторичной структуры: зелёный — альфа-спирали, оранжевый — петли и неупорядоченные области, светло-голубой — бета-стрэнды.

2.1.2. Формирование обучающей выборки и выбор оптимального алгоритма машинного обучения для построения предсказательной модели

Для построения предсказательной модели были опробованы различные методы машинного обучения из библиотеки Scikit-learn [128], с целью выбора наилучшего подхода. В рамках сравнительного анализа тестировались следующие алгоритмы: 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). Поскольку в исходном наборе данных наблюдался дисбаланс классов (445 положительных наблюдений против 68840 отрицательных наблюдений), было также проанализировано влияние соотношения размеров положительного и отрицательного классов на качество прогнозирования с целью определения оптимального соотношения размеров классов в финальной обучающей выборке. Для этого все данные были предварительно разделены на две части: в первую часть входили наблюдения с заданным соотношением положительных и отрицательных примеров, вторую часть составляли оставшиеся данные. Далее первую часть разбивали на 10 равных подмножеств с равномерным распределением положительных примеров между ними. Обучение модели выполнялось на 9 из 10 подмножеств, применяя вышеописанные алгоритмы машинного обучения, а тестирование проводилось на объединённой выборке, включающей оставшееся десятое подмножество первой части и все данные из второй части. Такая процедура повторялась таким образом, чтобы каждое из десяти подмножеств первой части данных использовалось для тестирования. В результате было получено 10 значений ROC AUC, которые затем усреднялись. 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.

Дополнительно была проведена проверка на возможную утечку данных из обучающего множества в тестовое при оценке качества прогнозирования. Для

этого был применён алгоритм выравнивания трёхмерных структур DALI [129] с использованием порога $3\text{\AA}$ для выявления структур субстратов, близких по пространственной организации. Выявленные схожие структуры были сгруппированы методом односвязной кластеризации. После исключения таких структур из обучающей и текстовой выборок, качество прогнозирования разработанной модели изменилось незначительно (<0.02).

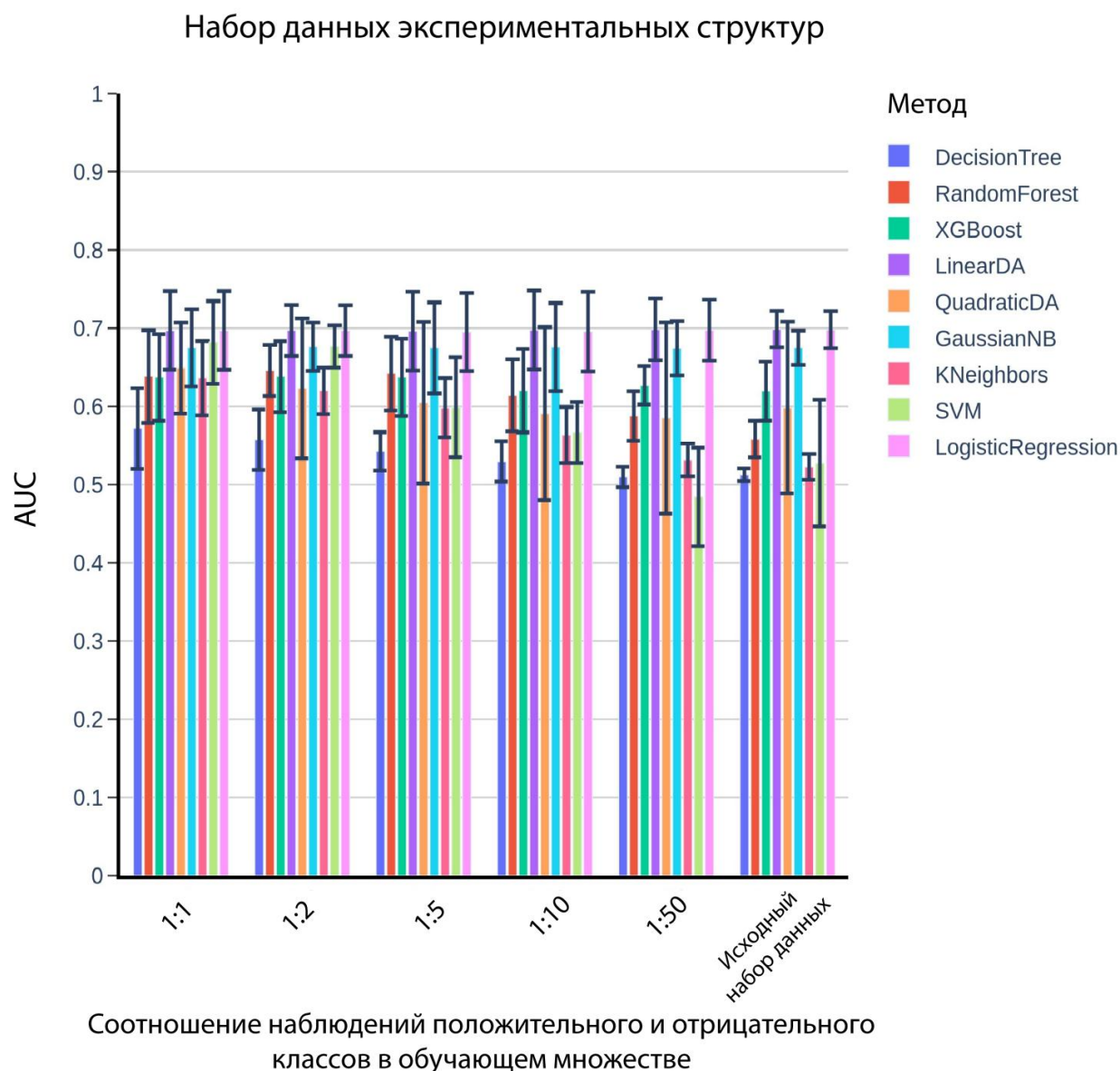
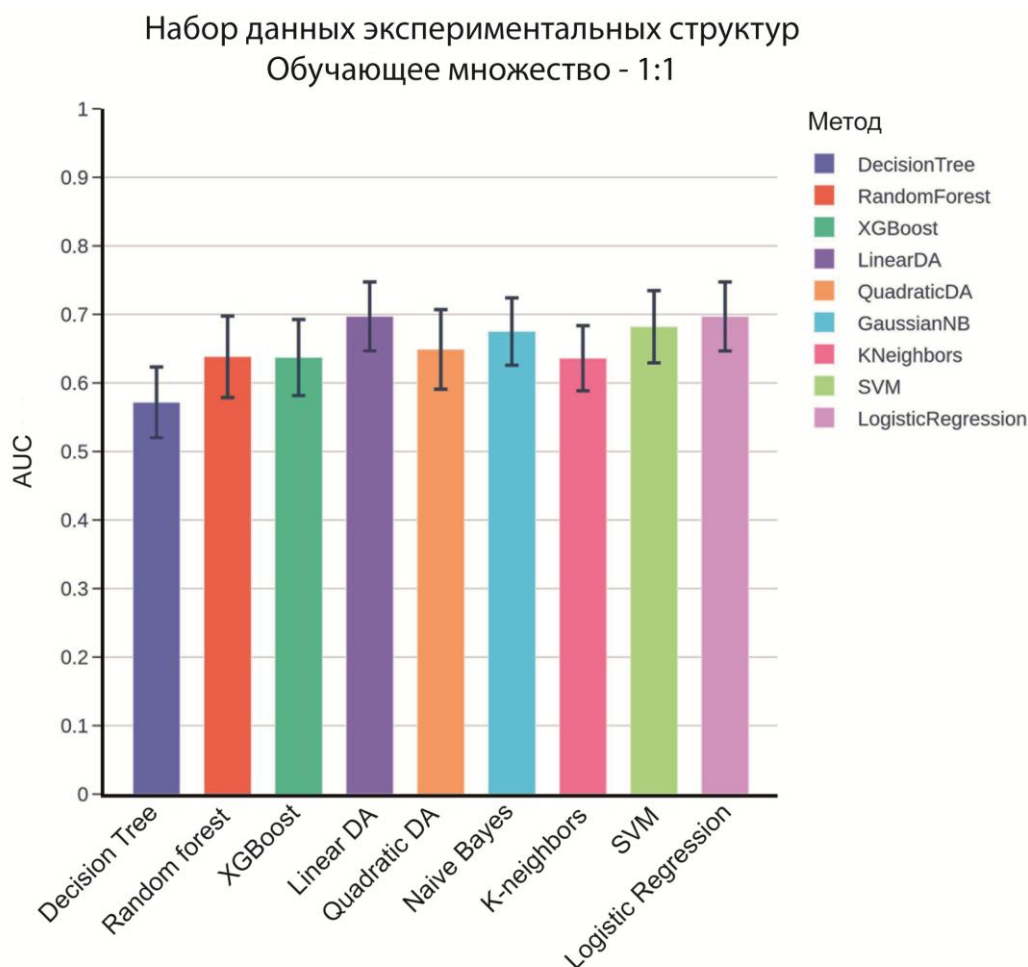


Рисунок 3. Зависимость качества предсказания модели от соотношения классов в обучающей выборке. В качестве метрики использовалась площадь под ROC-кривой (ROC AUC). Для построения моделей применялись различные алгоритмы машинного обучения.

Также согласно полученным результатам (Рис. 4), наилучшие показатели качества прогнозирования продемонстрировала модель, основанная на алгоритме линейного дискриминантного анализа.



Методы машинного обучения

Рисунок 4. Оценка качества предсказания модели при соотношении наблюдений положительного и отрицательного классов 1:1 в обучающей выборке. В качестве метрики использовалась площадь под ROC-кривой (ROC AUC). Наилучшее качество предсказания продемонстрировала модель, построенная с использованием алгоритма линейного дискриминантного анализа.

2.1.3. Визуализация результатов предсказаний на трёхмерных структурах потенциальных субстратов и определение порогового значения оценок восприимчивости к протеолизу для физиологических сайтов расщепления

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. Для визуализации этих оценок на трёхмерных структурах потенциальных субстратов использовался

инструмент UCSF Chimera. 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. Эти наблюдения согласуются с данными предыдущих исследований [17, 121] (Рис. 5).

Тем не менее, в ряде случаев модель занижает оценки восприимчивости к протеолизу для участков, которые, с высокой вероятностью, являются неструктурированными, доступными и подвижными в условиях водного окружения, но в зафиксированной конформации оказываются экранированными другими подвижными участками, что затрудняет их корректную оценку. Таким образом, использование закристаллизованных трёхмерных структур белков вносит определённые ограничения в предсказательные способности модели, поскольку не отражает динамических свойств белковых молекул в растворе.

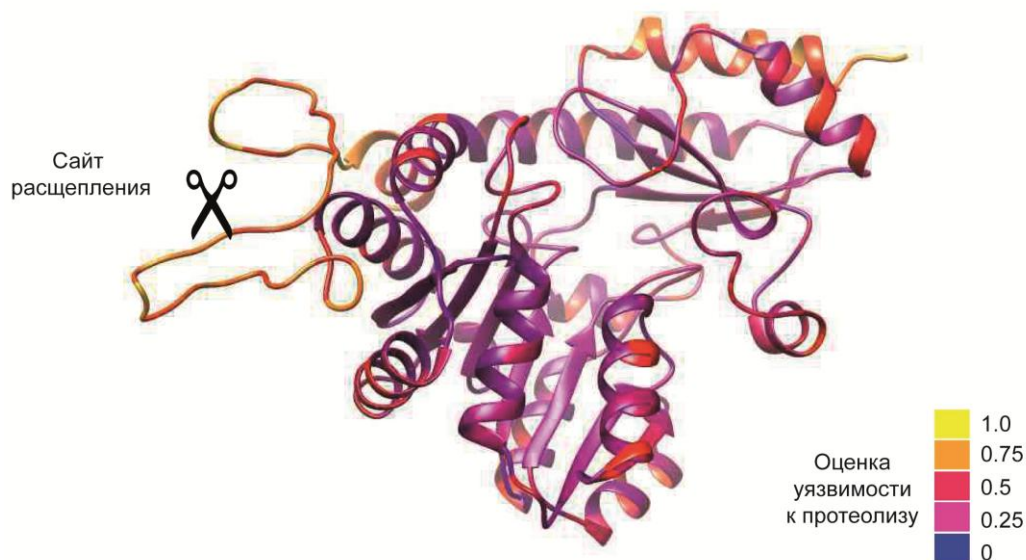


Рисунок 5. Визуализация предсказаний восприимчивости к протеолизу пептидных связей, полученных с помощью разработанной модели, на трёхмерной структуре субстрата. Наиболее высокие значения уязвимости к протеолизу наблюдаются в протяжённой выступающей петле, а также на N- и C-концах белка. Средние оценки характерны для внешних поверхностей альфа-спиралей и коротких петель. Гидрофобное ядро белка демонстрирует низкие оценки восприимчивости к протеолитическому расщеплению.

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). В качестве исходных данных использовалась информация из базы MEROPS, содержащая аннотированные события протеолиза, классифицированные по следующим категориям: 1) *physiological* — события протеолиза в условиях *in vivo*; 2) *non-physiological* — события протеолиза в условиях *in vitro*; 3) *pathological* — события протеолиза, происходящие *in vivo* и связанные с патологическими процессами; 4) *theoretical* — предполагаемые события расщепления; 5) *synthetic* — события протеолиза, подтверждённые на искусственно синтезируемых субстратах. Из этой выборки были отобраны только те белки, для которых известна трёхмерная структура. Для каждой позиции была получена оценка восприимчивости к протеолизу при помощи разработанной предсказательной модели, на основании чего были построены распределения оценок для различных категорий сайтов (Рис. 6).

Согласно графику, распределения оценок уязвимости участков белка, полученные структурной моделью, в значительной степени перекрываются. Тем не менее, тест Вилкоксона показал, что различия между этими распределениями являются статистически значимыми ($p\text{-value} = 1,43 \times 10^{-5}$). Однако, несмотря на это, на практике довольно затруднительно надёжно отличить сайты протеолиза,

соответствующие событиям, происходящим в живой клетке, от регистрируемых в лабораторных условиях (точность классификации = 0,622).

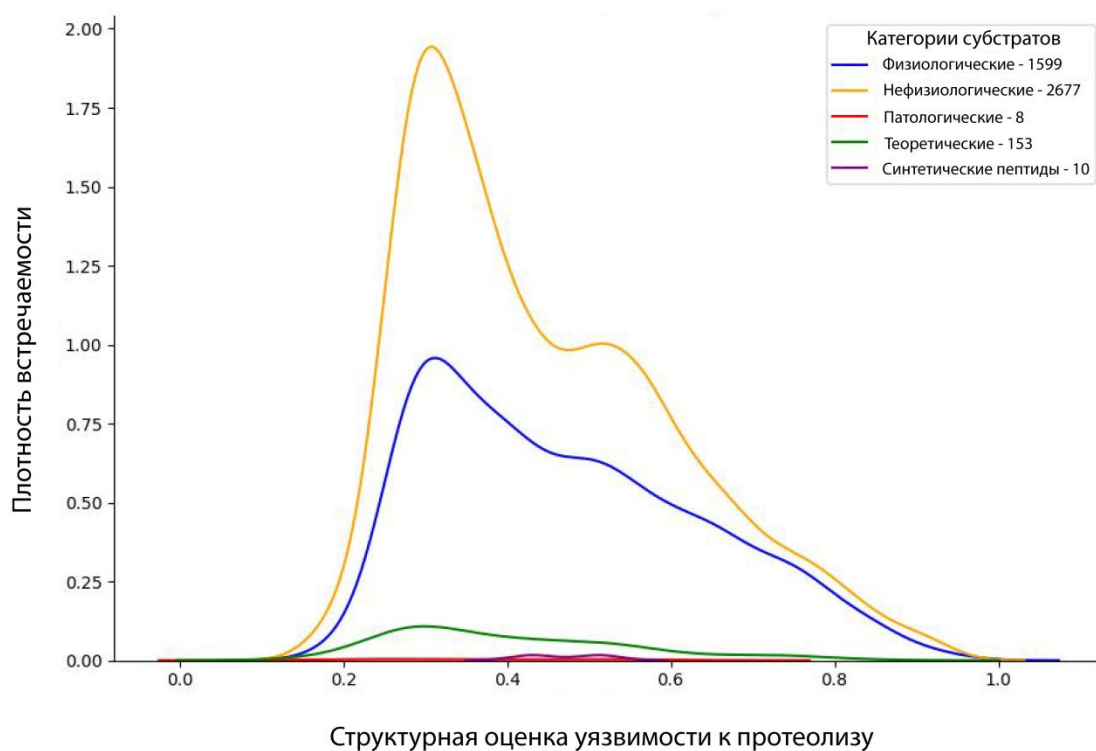


Рисунок 6. Распределение оценок восприимчивости к протеолизу, сгенерированных разработанной моделью, для различных категорий протеолитических расщеплений.

2.2. Расширение обучающей выборки с использованием моделей трёхмерных структур субстратов из базы данных AlphaFoldDB

2.2.1. Анализ особенностей моделей трёхмерных структур белков AlphaFold с точки зрения построения предсказательной модели восприимчивости к протеолизу

Прогресс в области искусственного интеллекта привёл к прорывам в различных научных дисциплинах, включая вычислительную биологию. Недавно разработанный метод предсказания трёхмерной структуры белка по

аминокислотной последовательности AlphaFold, основанный на алгоритмах глубокого обучения, значительно превзошёл другие методы в этой области и продемонстрировал качество моделей, сопоставимое с результатами, полученными с использованием экспериментальных методов [130]. Впоследствии AlphaFold был применён ко всем аминокислотным последовательностям белков протеома человека, а полученные высококачественные модели трёхмерных структур белков были опубликованы в открытом доступе в базе данных AlphaFoldDB [131]. Использование моделей трёхмерных структур белков из базы данных AlphaFoldDB позволит увеличить число положительных наблюдений в обучающем наборе данных, что, в свою очередь, может положительно сказаться на качестве прогнозирования сайтов восприимчивости к протеолизу.

В то же время важной особенностью моделей трёхмерных структур из AlphaFoldDB является отсутствие температурного фактора — параметра, измеряемого в ходе экспериментального определения трёхмерных структур белков методом рентгеноструктурного анализа или криоэлектронной микроскопии. Хотя температурный фактор отсутствует в моделях AlphaFold, они содержат оценку достоверности предсказания для каждого участка структуры, которая теоретически может обладать предсказательной значимостью. Поскольку температурный фактор недоступен при использовании моделей из AlphaFoldDB, прямое расширение обучающего множества невозможно. Однако исключение температурного фактора из предсказательной модели и её повторное обучение на объединённой выборке, включающей модели трёхмерных структур из AlphaFoldDB, позволяет оценить, улучшает ли увеличение объёма данных качество предсказания.

2.2.2. Подготовка расширенного набора данных для построения модифицированной предсказательной модели

Для построения модифицированной предсказательной модели были использованы те же протеолитические события и субстраты из базы данных CutDB, что и при обучении исходной предсказательной модели: всего 4576 протеолитических событий, соответствующих 2062 уникальным субстратам и 457 уникальным протеазам. Поиск моделей трёхмерных структур субстратов в базе данных AlphaFoldDB осуществлялся с помощью алгоритма BLAST. При этом применялись те же критерии отбора, что и ранее, подробно описанные в параграфе 2.1.1 данной работы. По результатам данного этапа было найдено 1209 моделей субстратов, соответствующих 3168 протеолитическим событиям и 317 уникальным протеазам.

Отображение сайтов протеолиза на AlphaFold-моделях, а также курирование позиций расщепления проводилось аналогично процедуре, применённой для экспериментально определённых структур. По итогам картирования на AlphaFold-модели осталось 2925 позиций расщепления, 1209 субстратов и 317 протеаз. После этапа курирования протеолитических событий финальный набор составил 2918 позиций расщепления, соответствующих 1205 субстратам и 314 протеазам.

Таким образом, 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). Среднее количество сайтов протеолиза на один субстрат в расширенном наборе данных составило 2.42, а медианное количество — 1.

2.2.3. Построение модифицированной предсказательной модели на наборе данных, расширенном AlphaFold-моделями

Для построения модифицированной предсказательной модели на расширенном наборе данных использовались те же алгоритмы машинного обучения, что и ранее, как описано в параграфе 2.1.2. данной работы. Подход к

тестированию модели и метрика оценки качества (ROC AUC) также оставались неизменными. Единственным отличием при построении модифицированной модели от её исходной версии стало отсутствие температурного фактора в расширенном наборе данных, поскольку этот параметр недоступен в структурах, предсказанных AlphaFold. Таким образом, модифицированная модель была обучена на меньшем числе структурных признаков, чем исходная модель.

Следует также отметить, что анализ предсказательной способности уникальной характеристики AlphaFold-моделей — оценки достоверности предсказанной структуры — показал, что данная характеристика имеет предсказательную силу (Рис. 7).

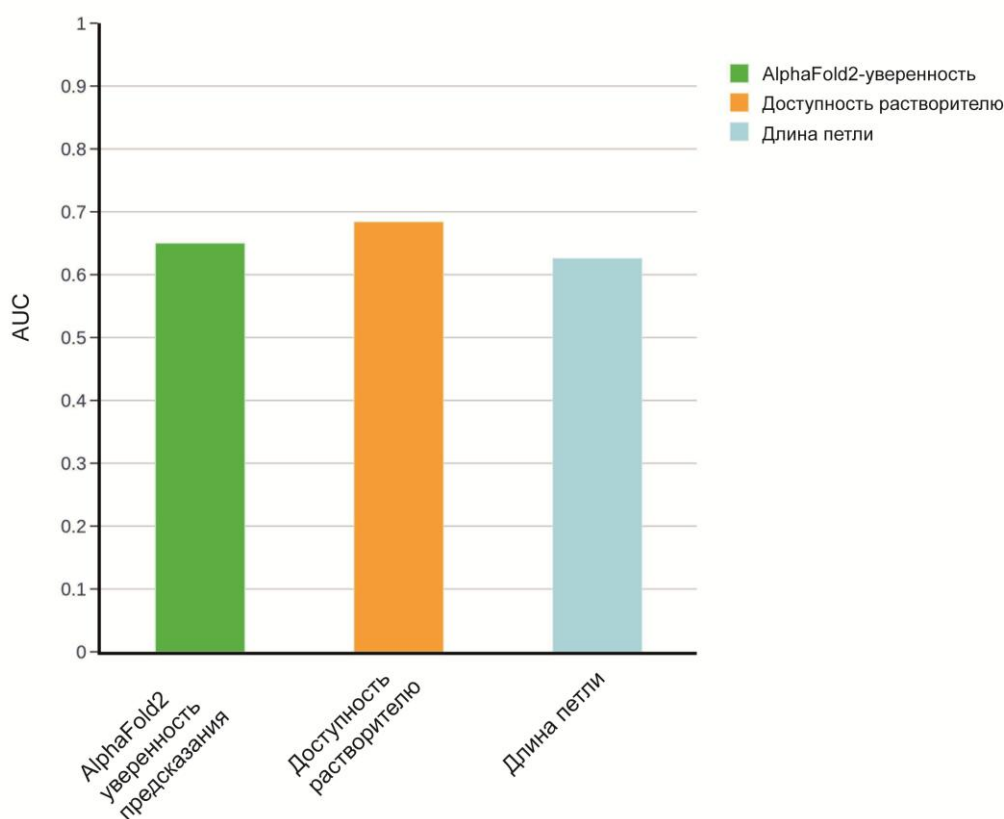


Рисунок 7. Сравнение предсказательной способности характеристики, описывающей достоверность структурного предсказания AlphaFold-модели, с предсказательной способностью других структурных признаков.

Согласно полученным результатам (Рис. 8), как и в случае с исходной моделью, качество прогнозирования модифицированной модели не зависит от соотношения классов в обучающей выборке. Поэтому соотношение количества

положительных и отрицательных наблюдений в итоговой обучающей выборке расширенного набора данных также составило 1:1.

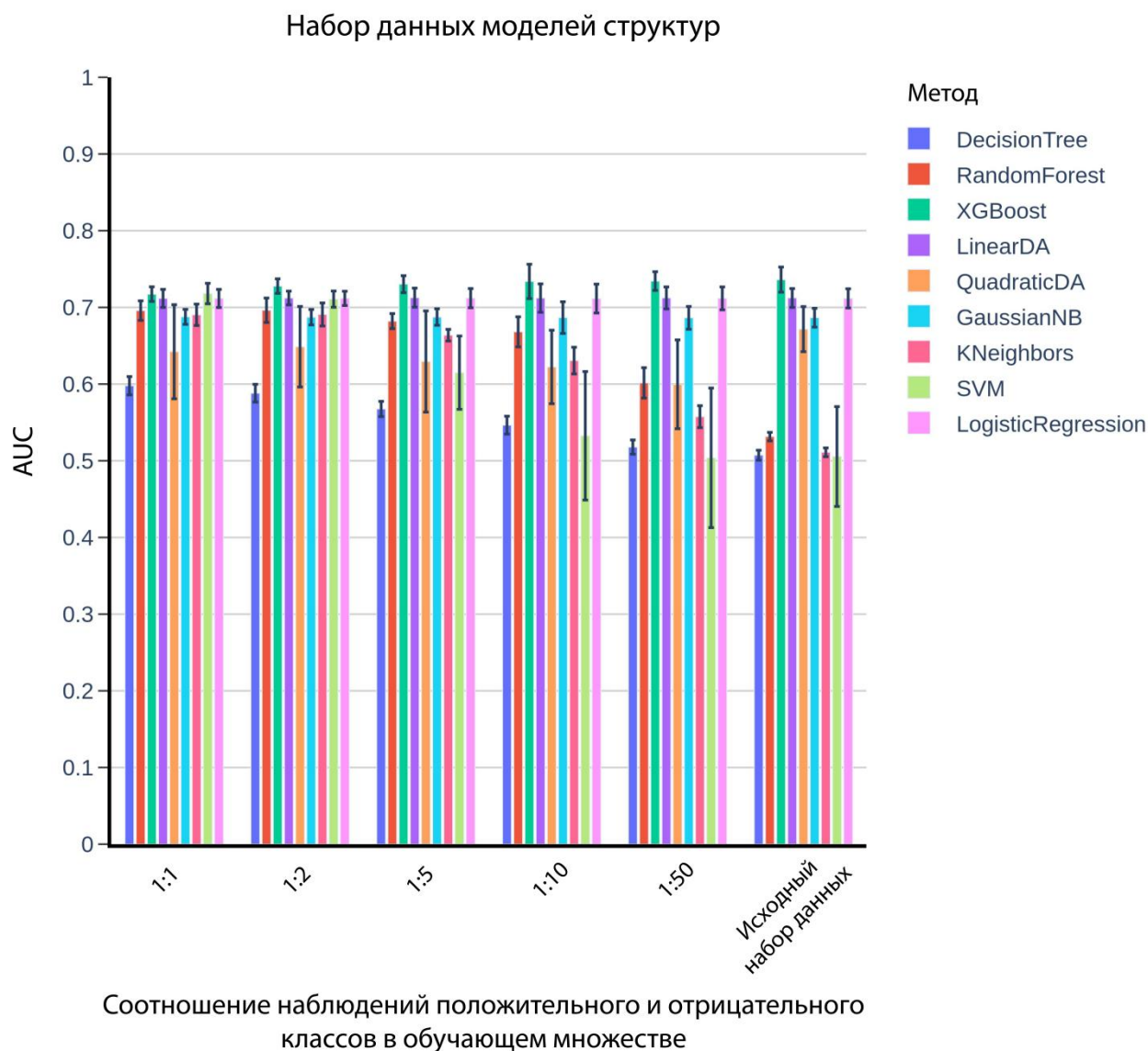


Рисунок 8. Оценка качества прогнозирования модифицированной модели в зависимости от соотношения классов в обучающей выборке. В качестве метрики использовалась площадь под ROC-кривой (ROC AUC). Для построения моделей применялись различные алгоритмы машинного обучения.

Результаты тестирования показали (Рис. 9), что наилучшие показатели точности прогнозирования среди протестированных алгоритмов продемонстрировала модель, обученная с использованием алгоритма XGBoost.

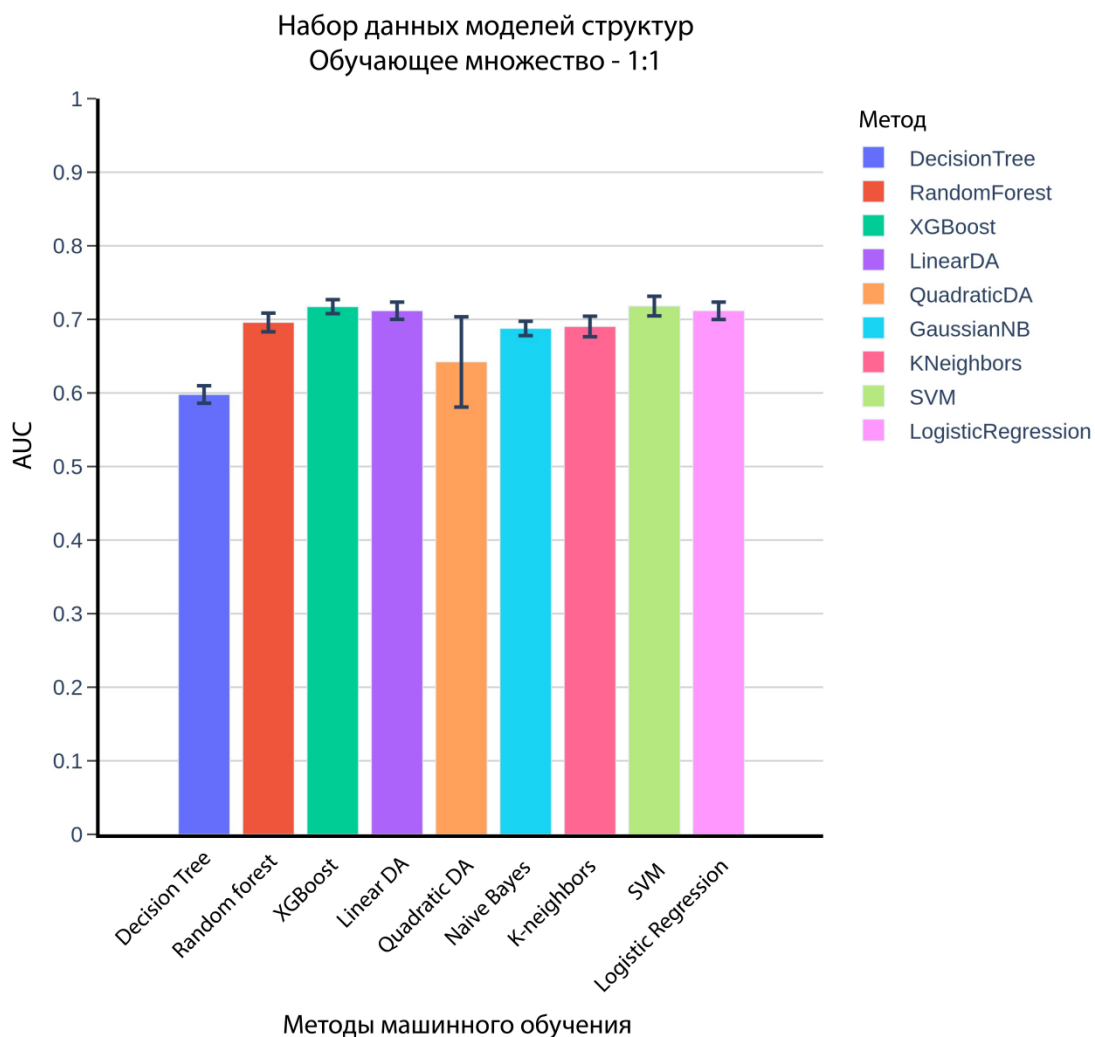


Рисунок 9. Оценка качества прогнозирования модифицированной модели при соотношении положительных и отрицательных классов 1:1 в обучающей выборке. В качестве метрики использовалась площадь под ROC-кривой (ROC AUC). Наилучшие результаты показала модель, обученная с использованием алгоритма XGBoost.

2.2.4. Сравнение качества прогнозирования исходной и модифицированной предсказательных моделей

Для корректного сравнения качества прогнозирования исходной и модифицированной предсказательных моделей было необходимо 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 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. 10).

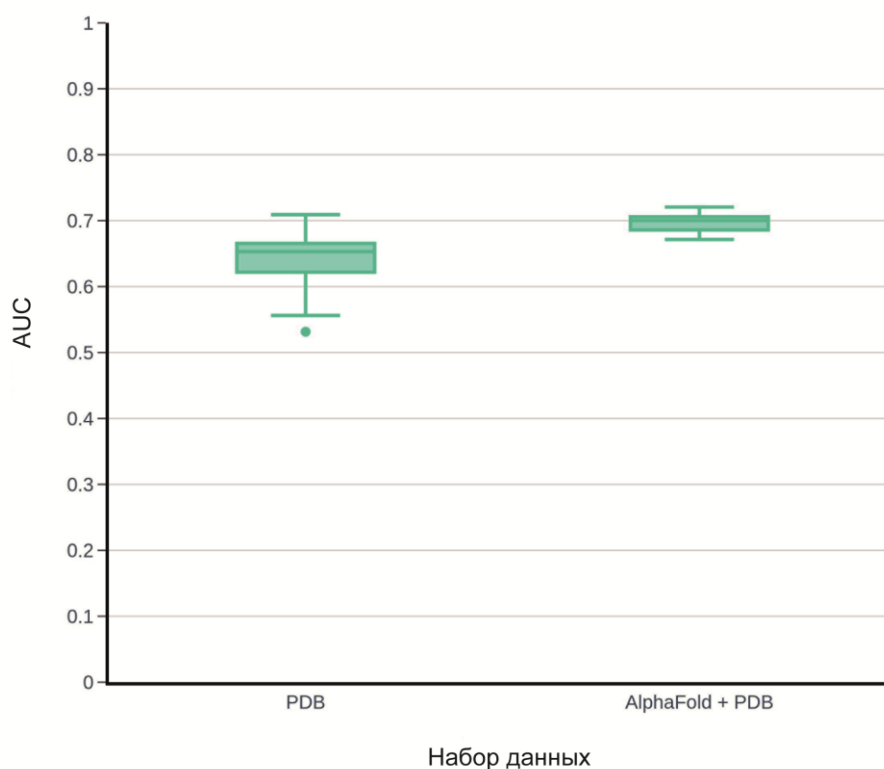


Рисунок 10. Улучшение качества прогнозирования модифицированной модели при расширении обучающего набора данных за счёт включения моделей трёхмерных структур из базы данных AlphaFoldDB.

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.

2.3. Интеграция структурной модели с моделями специфичности протеаз по аминокислотной последовательности и сравнение с существующими методами предсказания сайтов протеолиза

2.3.1. Построение интегративной модели на основе структурной модели и моделей специфичности протеаз по аминокислотной последовательности

Как было показано в разделе 2.1.2 данной работы, разработанная структурная модель, предсказывающая восприимчивость участков белка к протеолизу, продемонстрировала уверенные результаты (ROC AUC ~ 0.7). Однако очевидно, что включение дополнительной информации о специфичности протеазы по аминокислотной последовательности может существенно повысить точность предсказаний. Как правило, специфичность по последовательности моделируется с использованием позиционно-весовых матриц (Position-Weight Matrices, PWM), особенно в случаях, когда зависимости между позициями можно считать несущественными. 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. 11).

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. Картирование сайтов расщепления на структуры выполнялось по методике, описанной в параграфе 2.1.1. Из набора данных были исключены структуры, схожие с уже использованными для обучающей выборки

при построении структурной модели. В конечном наборе данных оказалось 174 протеолитических расщепления, соответствующих 24 уникальным структурам субстратов.

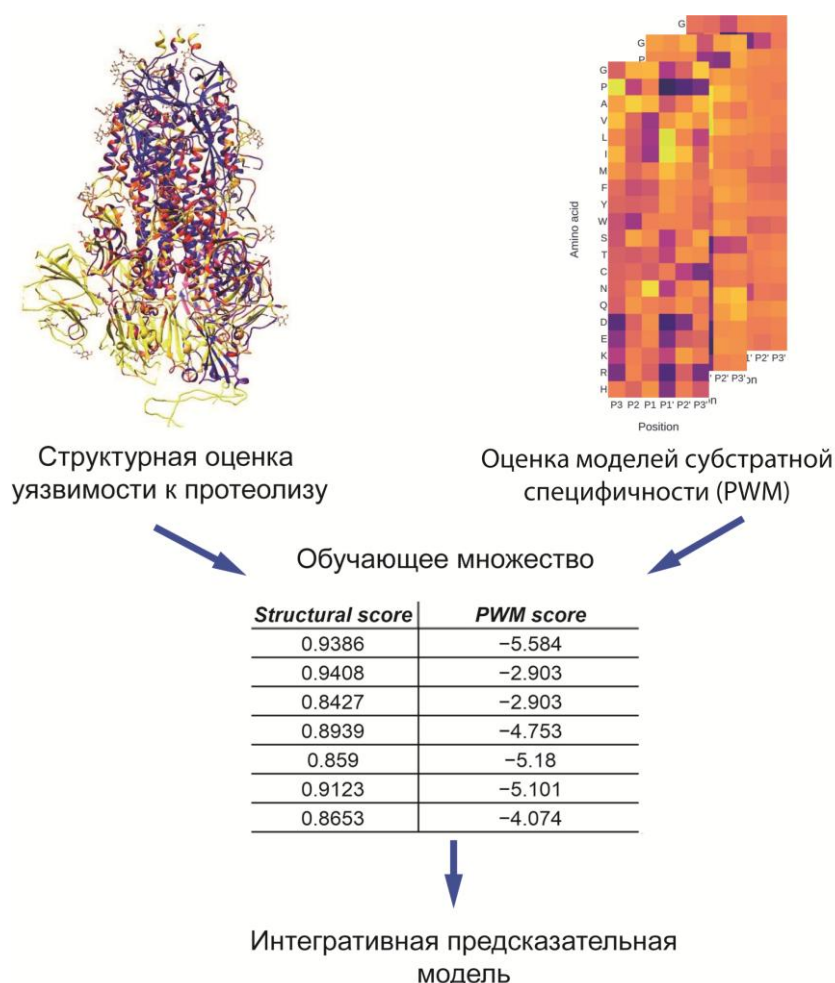


Рисунок 11. Схематическое представление объединения результатов предсказаний структурной моделью и протеазо-специфическими моделями, описывающими субстратную специфичность конкретных протеаз.

К полученному набору данных были применены как структурная модель, так и модели специфичности протеаз MMP9 и MMP25, полученные в раннем исследовании [132]. 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.

Для объединения этих двух наборов предсказаний в итоговую модель были апробированы различные алгоритмы из библиотеки Scikit-learn, включая: метод k-ближайших соседей (KNN), метод опорных векторов с линейной и радиальной

базисными функциями (Linear SVM и RBF SVM), алгоритмы гауссовского процесса (Gaussian Process), дерево решений (Decision Tree), случайный лес (Random Forest), наивный байесовский классификатор (Naïve Bayes), квадратический дискриминантный анализ (QDA), нейронную 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. 12). Таким образом, интегративная модель, объединяющая оценки, выдаваемые структурной моделью и моделями специфичности по последовательности, была построена на основе алгоритма наивного байесовского классификатора.

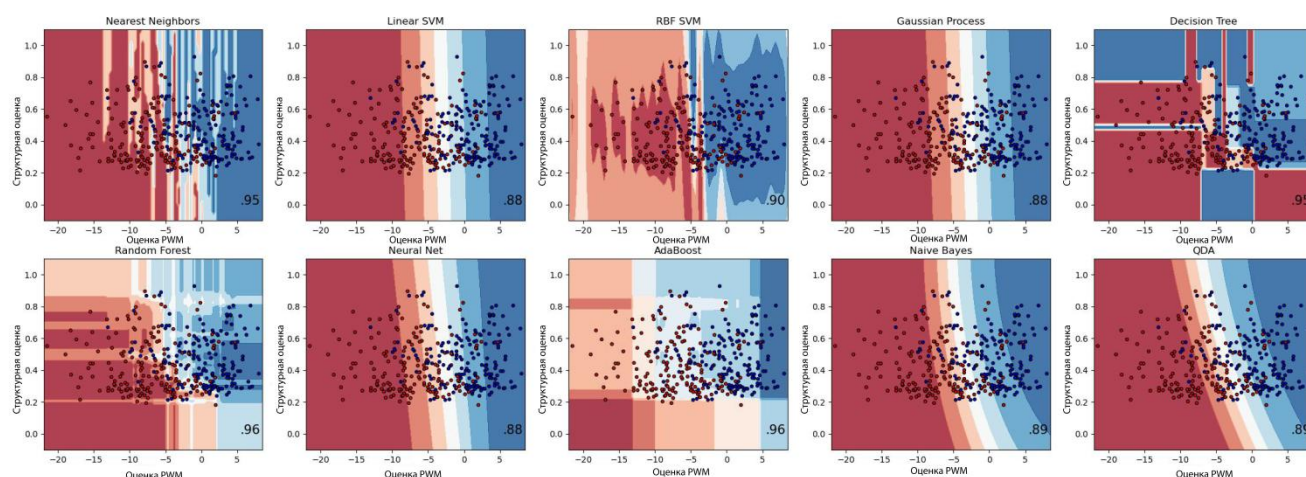


Рисунок 12. Двумерные визуализации предсказаний интегративной модели для наблюдений различных классов при использовании различных алгоритмов машинного обучения. Значения метрики ROC AUC указаны в правом нижнем углу каждой визуализации. Цветовой градиент от красного к синему отражает оценки восприимчивости к протеолизу, сгенерированные моделью: от меньших значений к большим.

2.3.2. Сравнение разработанного метода прогнозирования протеолитических событий с существующими методами

Одним из передовых инструментов прогнозирования протеолитических событий является Procleave [133], использующий 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 открытой базы данных MEROPS. Известно, что инструмент Procleave обучали на данных из MEROPS версии 9.0. и валидировали на данных из версии 12.0. На момент проведения данного сравнения двух методов была доступна версия MEROPS 12.4. Поэтому в тестовый набор включались только те субстраты и протеолитические события, которые были добавлены в базу, начиная с версии 12.0. Итоговый набор для тестирования включал 213 сайтов протеолиза, 129 субстратов и 3 протеазы — S01.269, C14.003 и M10.003.

Далее были осуществлены этапы поиска структур по аминокислотным последовательностям, картирования сайтов расщепления на структуры и курирование протеолитических событий, в соответствии с процедурой, описанной в параграфе 2.1.1 данной работы. Из набора также были исключены те структуры, которые входили в обучающую выборку при построении структурной модели. После фильтрации конечный набор данных для тестирования составили 43 позиции расщепления, соответствующие 28 субстратам и 3 протеазам.

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

S01.269, C14.003 и M10.003 были разработаны с помощью информации о сайтах расщепления данных протеаз, полученной из базы данных MEROPS. Метод разработки PWM-моделей подробно рассматривается в третьей главе данной работы.

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. 13).

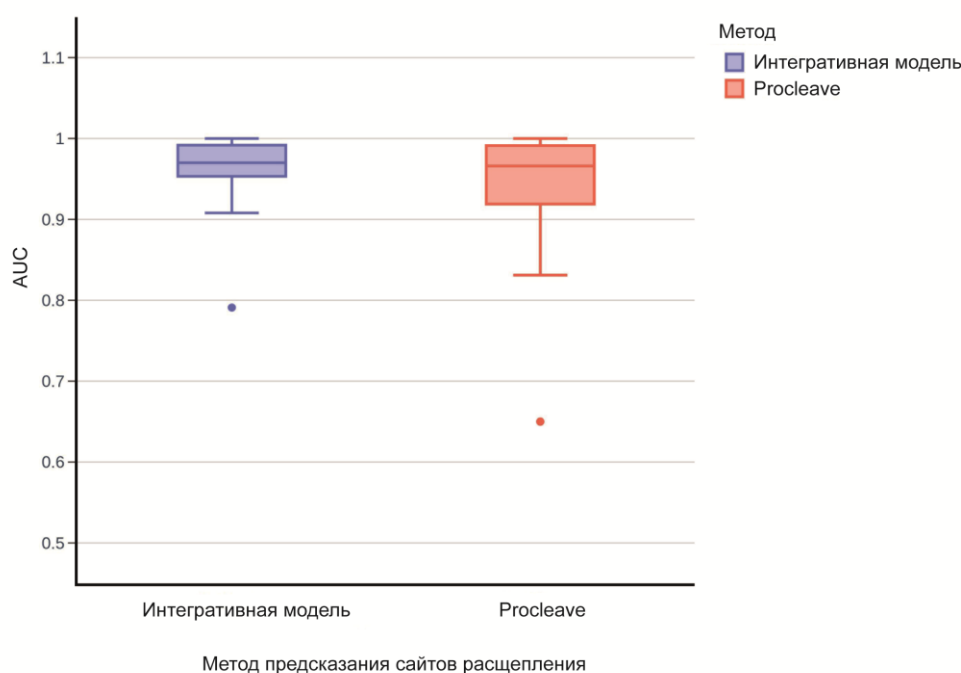


Рисунок 13. Сравнение качества прогнозирования между разработанным интегративным методом, объединяющим предсказания структурной модели и модели специфичности по последовательности, и методом Procleave.

2.4. Обсуждение результатов

1) Был сформирован набор данных для последующего построения модели предсказания протеолитических событий на основе структурных характеристик субстратов. В качестве исходной информации использовались аннотированные сайты расщепления из базы данных CutDB. Для соответствующих субстратов были идентифицированы трёхмерные структуры в базе PDB с последующей фильтрацией на схожесть по последовательности и по длине выравнивания. Картирование сайтов расщепления на структуры осуществлялось с учётом выравнивания аминокислотных последовательностей субстратов и их трёхмерных структур, а затем было дополнено этапом экспертного курирования, позволившем исключить случаи, потенциально связанные с денатурацией белков в условиях эксперимента. На основе анализа литературы были определены ключевые структурные характеристики, потенциально влияющие на восприимчивость пептидной связи к протеолизу. Кроме того, были разработаны и внедрены дополнительные признаки, описывающие пространственные особенности субстратов. Все количественные признаки были нормированы, а категориальные — перекодированы для последующего применения в алгоритмах машинного обучения. Результатом данного этапа работы стал набор данных, включающий пять характеристик структурного окружения для 69285 пептидных связей, соответствующих 190 субстратам с известной трёхмерной структурой, из которых для 445 пептидных связей были зафиксированы подтверждённые протеолитические события.

2) Была построена и оптимизирована модель, предсказывающая уязвимость участков белка к протеолизу на основе анализа его трёхмерной структуры. В ходе анализа были апробированы девять различных алгоритмов машинного обучения, среди которых наилучшие результаты продемонстрировал метод линейного дискриминантного анализа, выбранный в качестве основы для финальной модели. В процессе оптимизации также было установлено, что

наилучшее качество прогнозирования достигается при равном соотношении положительного и отрицательного классов (1:1) в обучающей выборке. Были разработаны скрипты для визуализации результатов на трёхмерных структурах субстратов, обеспечившие их интерпретируемость. Кроме того, был проведён сравнительный анализ оценок восприимчивости к протеолизу для сайтов, выявленных в физиологических и нефизиологических условиях.

3) Был подготовлен обучающий набор данных на основе моделей трёхмерных структур белков из базы данных AlphaFoldDB с использованием процедуры, аналогичной применённой к экспериментально определённым структурам. Включение AlphaFold-моделей позволило существенно — более чем в шесть раз — увеличить объём обучающей выборки (2918 событий протеолиза против 445). Для обеспечения корректного сравнения с исходной моделью, обученной на экспериментальных структурах, из набора признаков был исключён температурный фактор, отсутствующий в моделях AlphaFold, а сама модель была повторно переобучена. Предсказательная модель, построенная на расширенной выборке, продемонстрировала более высокое качество прогнозирования по сравнению с моделью, основанной исключительно на экспериментальных структурах. Дополнительно была проанализирована уникальная характеристика AlphaFold-моделей — оценка достоверности предсказанной структуры. Было показано, что эта характеристика обладает предсказательной значимостью и может быть использована в качестве дополнительного признака при обучении модели на основе предсказанных трёхмерных структур AlphaFold.

4) Для сравнения с существующими методами разработанная структурная модель была интегрирована с подходами, моделирующими специфичность протеаз по аминокислотной последовательности на основе позиционно-весовых матриц (PWM). Построение модели осуществлялось на независимом наборе данных, полученном в эксперименте по расщеплению белков *E. coli* матричными металлопротеазами. Для объединения предсказаний двух типов моделей — структурной и PWM — были апробированы десять алгоритмов машинного обучения, из которых наилучшие результаты продемонстрировал

наивный байесовский классификатор. Тестирование проводилось на новых, независимых данных из MEROPS, не использовавшихся при обучении моделей. Сравнение разработанного метода с современным инструментом Procleave показало сопоставимое качество прогнозирования: средние значения ROC AUC разработанного метода и инструмента Procleave составили 0,962 и 0,937, соответственно, а медианные — 0,97 и 0,966.

ГЛАВА III. БИОИНФОРМАТИЧЕСКИЙ АНАЛИЗ ПРОТЕОЛИТИЧЕСКОЙ АКТИВАЦИИ S-ГЛИКОПРОТЕИНА ОБОЛОЧКИ КОРОНАВИРУСА SARS-COV-2

3.1. Протеолитическая активация S-гликопротеина оболочки коронавируса SARS-CoV-2

Основные аспекты протеолитической активации S-гликопротеина оболочки коронавируса SARS-CoV-2 были изложены в главе 1. В данном подразделе сделан акцент на нерешённые вопросы протеолитической активации S-белка коронавируса SARS-CoV-2. Одной из таких нерешённых задач остаётся идентификация протеаз, потенциально способных активировать S-белок. Известно, что протеолиз занимает важное место в жизненном цикле вируса SARS-CoV-2, участвуя в двух критически важных процессах — проникновении вируса в клетку-хозяина и созревании вирионов. Созревание вируса сопровождается протеолитической активацией исходной формы S-белка, локализованного на поверхности вирусной частицы. На данном этапе активную роль играет фурин, расщепляющий полипептидную цепь S-белка в позиции R685, известной как S1/S2. Протеолитическая активация S1/S2 сайта приводит к изменению конформации S-гликопротеина, благодаря чему, вирус с большей эффективностью связывается с клеточными рецепторами на поверхности инфицируемой клетки. Проникновение вируса в клетку также сопровождается протеолитическим процессом. Так называемый «классический путь» проникновения коронавируса SARS-CoV-2 включает расщепление сайта S2' в позиции R815, осуществляемое трансмембранной сериновой протеазой II типа (TMPRSS2). Расщепление в этой позиции приводит к значительным конформационным изменениям в S-белке коронавируса, в результате чего происходит слияние клеточной мембраны и оболочки вируса с образованием поры, через которую вирус попадает внутрь клетки. Однако, в ряде исследований

показано, что, по-видимому, в протеолитической активации могут быть задействованы и другие протеазы [78, 85, 92, 94], поэтому в рамках настоящей работы была проведена биоинформатическая идентификация дополнительных протеаз человека, способных активировать S-белок.

Особый интерес представляет альтернативный путь проникновения коронавируса SARS-CoV-2 в клетку, связанный с эндоцитозом. Известно, что проникновение коронавируса SARS-CoV-2 не ограничивается классическим механизмом, поскольку вирус может попадать в клетку посредством эндоцитоза. При этом механизме вирусная частица захватывается клеткой, попадает в эндосому, и далее высвобождается в цитоплазму. Известно, что в этом процессе ключевую роль играет цистеиновая протеаза катепсин L, осуществляющая активацию S-белка внутри эндосом. Однако, до настоящего времени позиция сайта расщепления катепсином L в S-белке остаётся не установленной. В данной работе разработанный метод предсказания сайтов протеолиза был применён для идентификации позиции расщепления катепсином L. Также в рамках данной работы было проанализировано влияние мутаций на эффективность протеолитической активации S-гликопротеина, что может иметь значение для оценки патогенности новых вариантов вируса.

3.2. Биоинформатическая идентификация протеаз человека, способных активировать S-белок оболочки коронавируса SARS-CoV-2

3.2.1. Разработка предсказательных моделей, описывающих субстратную специфичность протеаз человека по аминокислотной последовательности

Для решения проблемы идентификации протеаз человека, потенциально способных активировать S-гликопротеин оболочки коронавируса SARS-CoV-2, было необходимо построить предсказательные модели, отражающие их

субстратную специфичность по аминокислотной последовательности. Модели специфичности были разработаны в виде позиционно-весовых матриц (PWM), как это описано в [134], основываясь на информации об аминокислотных последовательностях известных субстратов протеаз человека, полученной из базы данных MEROPS. Для каждой протеазы человека, имеющей в базе данных MEROPS не менее восьми известных последовательностей субстратов, были рассчитаны частоты встречаемости аминокислотных остатков в позициях P3-P3' — f_{ij} , где i — аминокислотный остаток, а j — позиция аминокислотного остатка. Также были рассчитаны фоновые частоты встречаемости аминокислотных остатков по всему протеому человека b_j . Данные по протеому были получены из базы данных UniProtKB [135]. Каждой ячейке позиционно-весовой матрицы, со столбцами, представляющими позиции P3-P3', и строками, представляющими 20 стандартных аминокислотных остатков, было присвоено значение r_{ij} , рассчитанное как логарифм отношения частоты встречаемости аминокислотных остатков в сайтах расщеплений к фоновой частоте встречаемости аминокислотных остатков:

$$r_{ij} = \log_2 \frac{f_{ij}}{b_j} \quad (1.1)$$

Предполагая независимость аминокислотных остатков в рамке расщепления P3-P3', оценка расщепления (cleavage score, CS) для конкретной пептидной связи в белке вычислялась как сумма соответствующих значений матрицы для аминокислотных остатков в позициях P3-P3':

$$CS = \sum_{P3}^{P3'} r_{ij} \quad (1.2)$$

Таким образом, для 169 человеческих протеаз были построены предсказательные модели, описывающие субстратную специфичность протеаз по аминокислотной последовательности.

3.2.2. Валидация моделей субстратной специфичности

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. 14), 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.

На втором этапе валидации модели были применены к аминокислотной последовательности S-белка исходного уханьского штамма коронавируса SARS-CoV-2 (UniProtKB ID: P0DTC2) [136]. 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 [137]) и о профилях экспрессии протеаз в тканях (база данных TISSUES [138]).

Согласно полученным результатам, наивысшую оценку расщепления в позиции R685 показала модель субстратной специфичности фурина (Рис. 15), а в позиции R815 — модель TMPRSS2 (Рис. 16). 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.

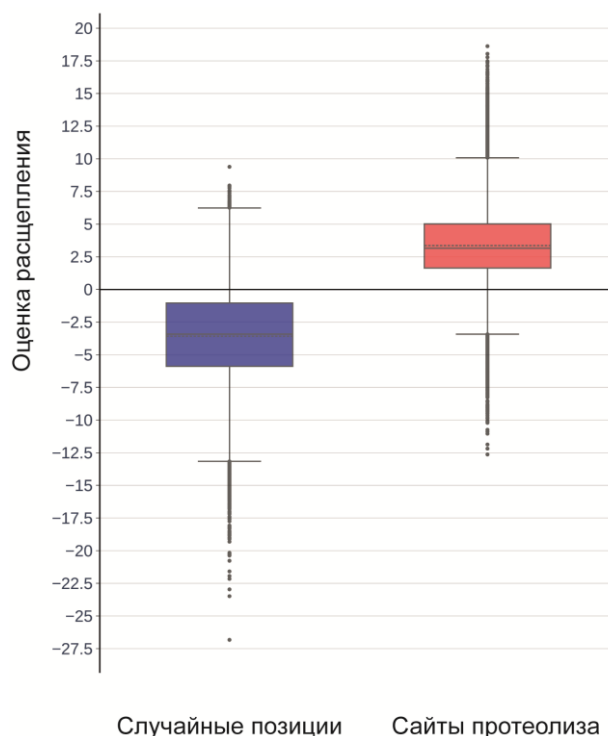


Рисунок 14. Распределение оценок расщепления, рассчитанных с помощью моделей субстратной специфичности протеаз, для известных сайтов расщеплений и случайных позиций в белках.

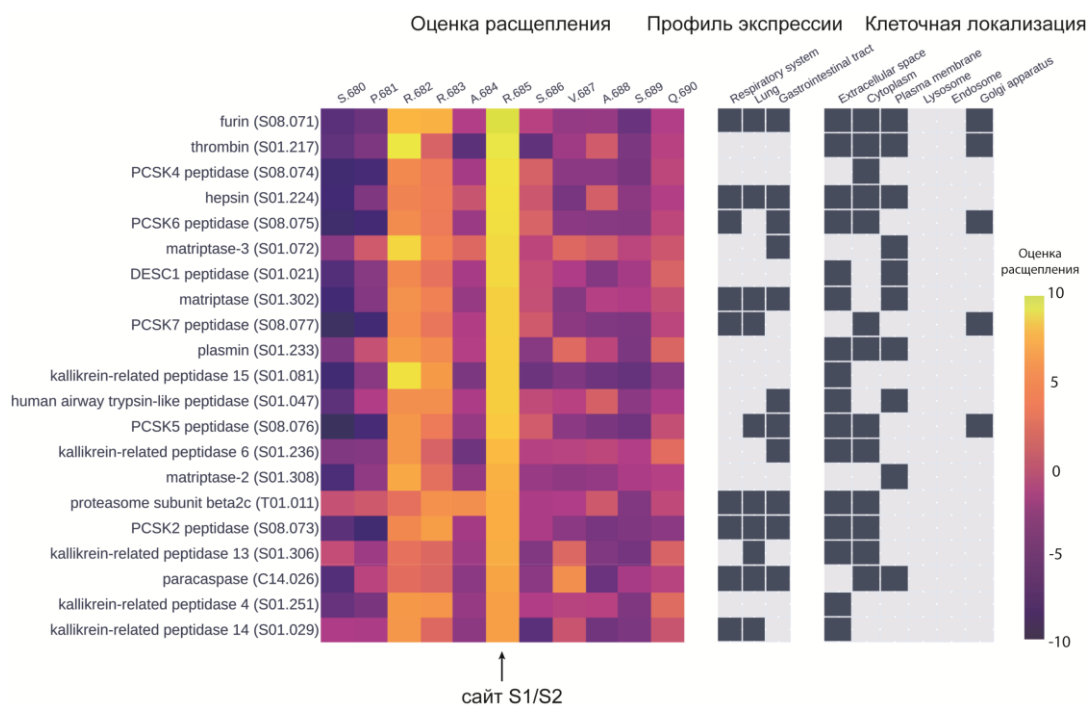


Рисунок 15. Двадцать наивысших оценок расщепления в позиции R685 (сайт S1/S2), полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель фуриновой протеазы. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

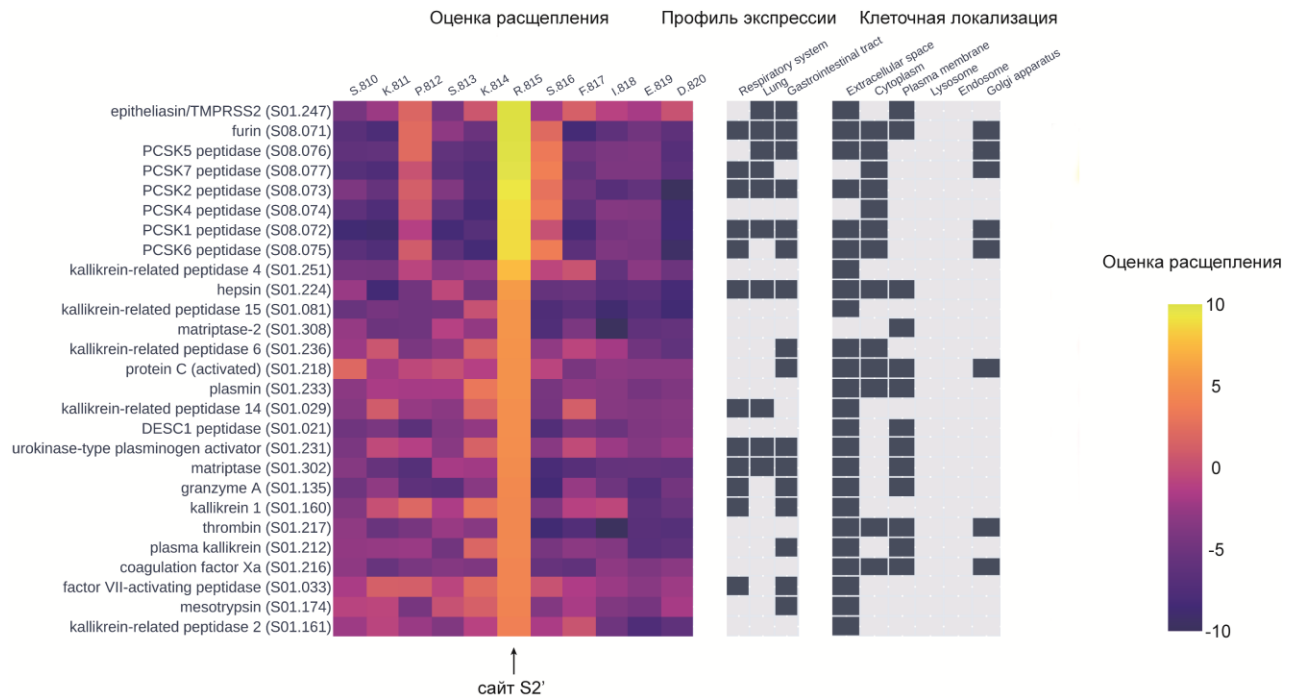


Рисунок 16. Тридцать наивысших оценок расщепления в позиции R815 (сайт S2'), полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель протеазы TMPRSS2. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

3.2.3. Идентификация протеаз, потенциально способных активировать S-гликопротеин оболочки коронавируса SARS-CoV-2 в известных сайтах расщепления S1/S2 и S2'

Согласно полученным результатам (Рис. 15), 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.

Для сайта S2` (позиция R815), по результатам моделирования (Рис. 16), 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.

3.3. Биоинформатический поиск 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 вероятности расщепления для всех возможных пептидных

связей. Согласно полученным результатам (Рис. 17), максимальная оценка расщепления соответствует позиции K790.

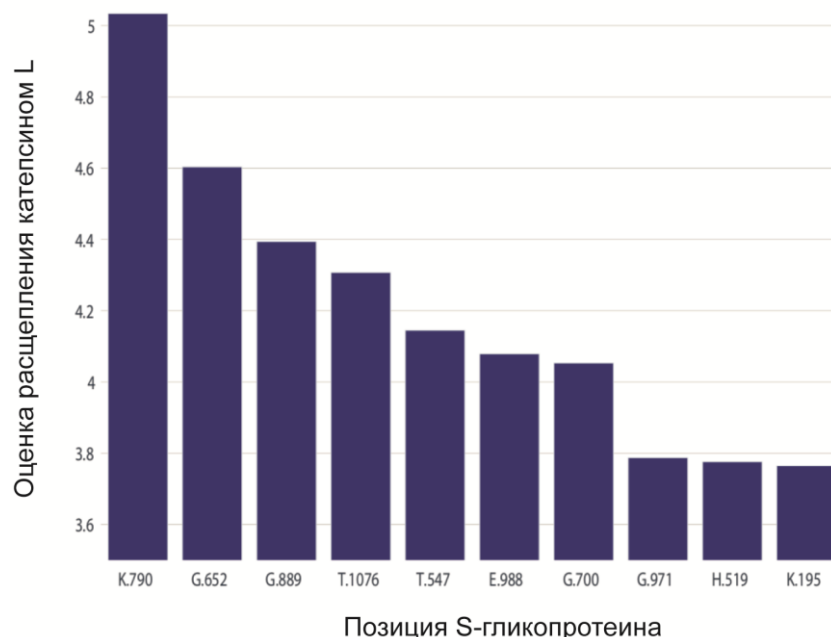


Рисунок 17. Десять позиций последовательности S-гликопротеина, получивших наивысшие оценки расщепления по модели субстратной специфичности катепсина L.

Далее было проанализировано расположение этой позиции на трёхмерной структуре S-гликопротеина. Из базы данных PDB были получены экспериментальные структуры S-белка в "открытой" (PDB ID: 6VYB) [139], "закрытой" (PDB ID: 6VXX) [139] конформациях, а также в комплексе с рецептором ACE2 (PDB ID: 7DF4) [140]. Незакристаллизованные участки, охватывающие сайты расщепления S1/S2 и S2', были смоделированы с помощью инструмента SWISS-MODEL с параметрами по умолчанию [141]. К полученным структурам была применена универсальная структурная модель, описанная в главе 2, с целью оценки 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. 18). 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.

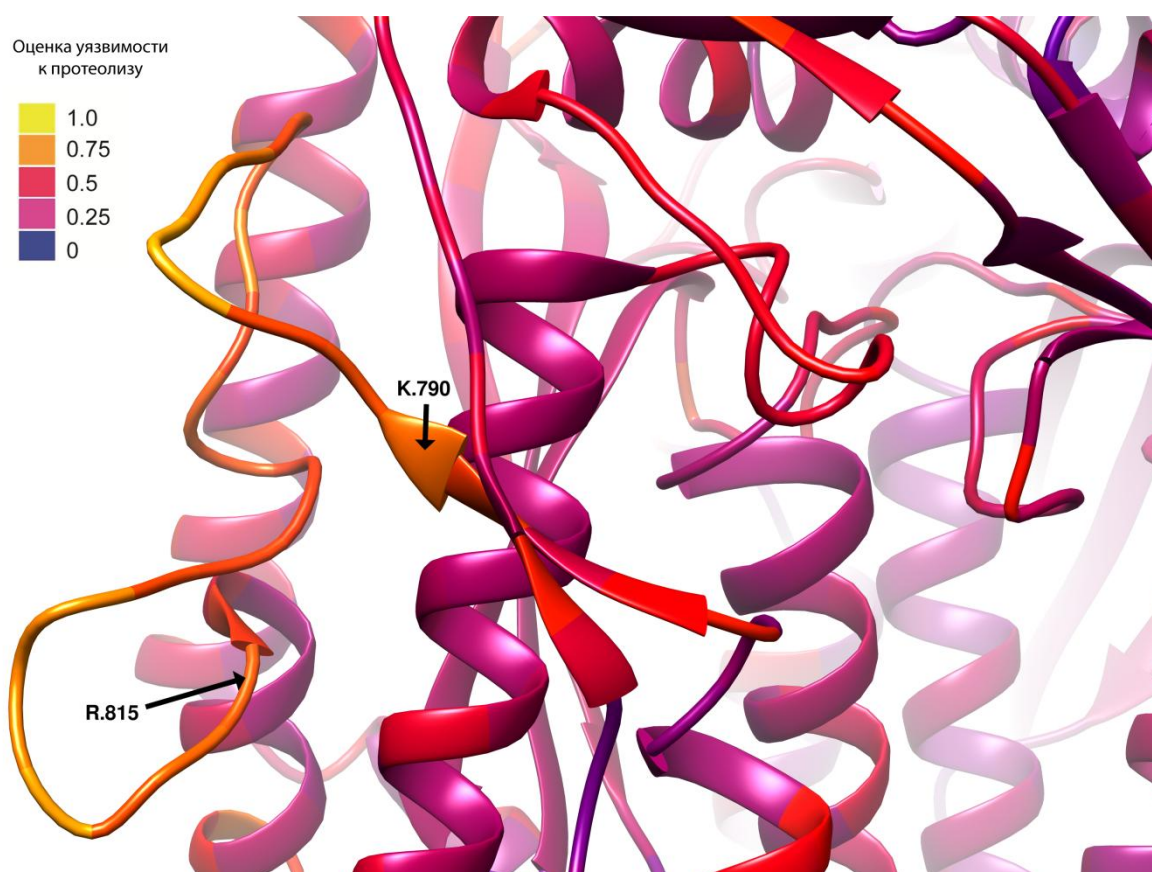


Рисунок 18. Отображение оценок восприимчивости к протеолизу, сгенерированных универсальной структурной моделью, на трёхмерной структуре S-гликопротеина. Стрелками указаны предполагаемый сайт расщепления катепсином L (K790) и известный сайт расщепления TMPRSS2 — S2' (R815).

Дополнительно было проанализировано потенциальное участие других протеаз в протеолитической активации S-белка в позиции K790 (Рис. 19). Хотя модель специфичности катепсина L показала лишь вторую по величине оценку для позиции K790, наивысшую оценку продемонстрировала модель мезотрипсина. Однако, мезотрипсин, в отличие от катепсина L, не локализуется в эндосомах или лизосомах, что делает его участие в активации S-белка

маловероятным. Примечательно, что третье по величине значение было получено для катепсина V — протеазы того же семейства, что и катепсин L, обладающей схожим профилем экспрессии и клеточной локализацией.

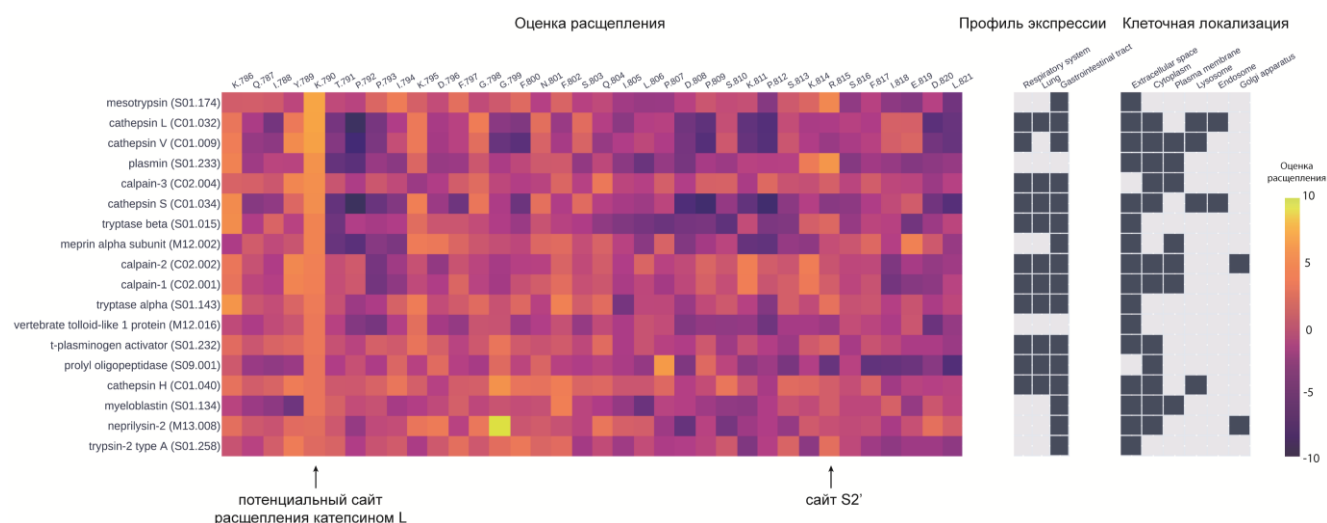


Рисунок 19. Двадцать наивысших оценок расщепления в позиции K790, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель мезотрипсина, модель катепсина L — вторую по величине. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

3.4. Поиск альтернативных сайтов расщепления в S-белке коронавируса SARS-CoV-2

Для выявления потенциальных сайтов протеолиза в S-гликопротеине SARS-CoV-2, за исключением известных сайтов S1/S2 и S2', был проведён анализ всех оценок расщепления, сгенерированных моделями субстратной специфичности 169 протеаз человека. Анализ ограничивался четырьмя ранее упомянутыми семействами протеаз — PCSK, TTSP (трансмембранные сериновые протеазы), калликреинами и факторами свёртывания крови. Для каждой позиции в последовательности S-гликопротеина были рассчитаны средние значения оценок расщепления для четырёх семейств протеаз. Далее были отобраны 20 позиций с наибольшими значениями оценок расщепления, и для них дополнительно получены оценки восприимчивости к протеолизу при помощи универсальной структурной модели (Рис. 20).

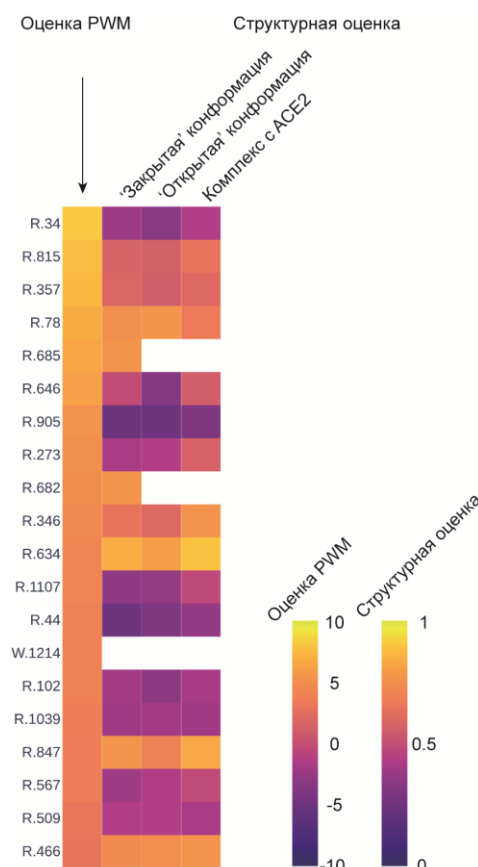


Рисунок 20. Двадцать позиций S-гликопротеина с наивысшими оценками расщепления, полученных с помощью моделей субстратной специфичности протеаз семейств PCSK, TTSP, калликреинов и факторов свёртывания крови. Для каждой из этих позиций с помощью универсальной структурной модели были рассчитаны оценки восприимчивости к протеолизу, используя трёхмерные структуры S-белка в «открытой» и «закрытой» конформациях, а также в комплексе с рецептором ACE2.

Согласно полученным результатам, помимо известных сайтов расщепления в позициях R685 и R815, только пять позиций из двадцати показали высокие значения как по оценкам расщепления, сгенерированным моделями субстратной специфичности, так и по оценкам восприимчивости к протеолизу, полученным универсальной структурной моделью. К этим позициям относятся: R78, R346, R466, R634 и R847. Анализ пространственной структуры S-гликопротеина показал, что позиция R78 локализуется в N-концевом домене белка (Рис. 21а), а позиции R346 и R466 — в рецептор-связывающем домене (Рис. 21б). Сайт протеолиза в позиции R634 находится внутри петли, соединяющей домены S1 и S2, известной как петля 630 [79] (Рис. 22а). Позиция R847 расположена вблизи пептида слияния (Рис. 22б).

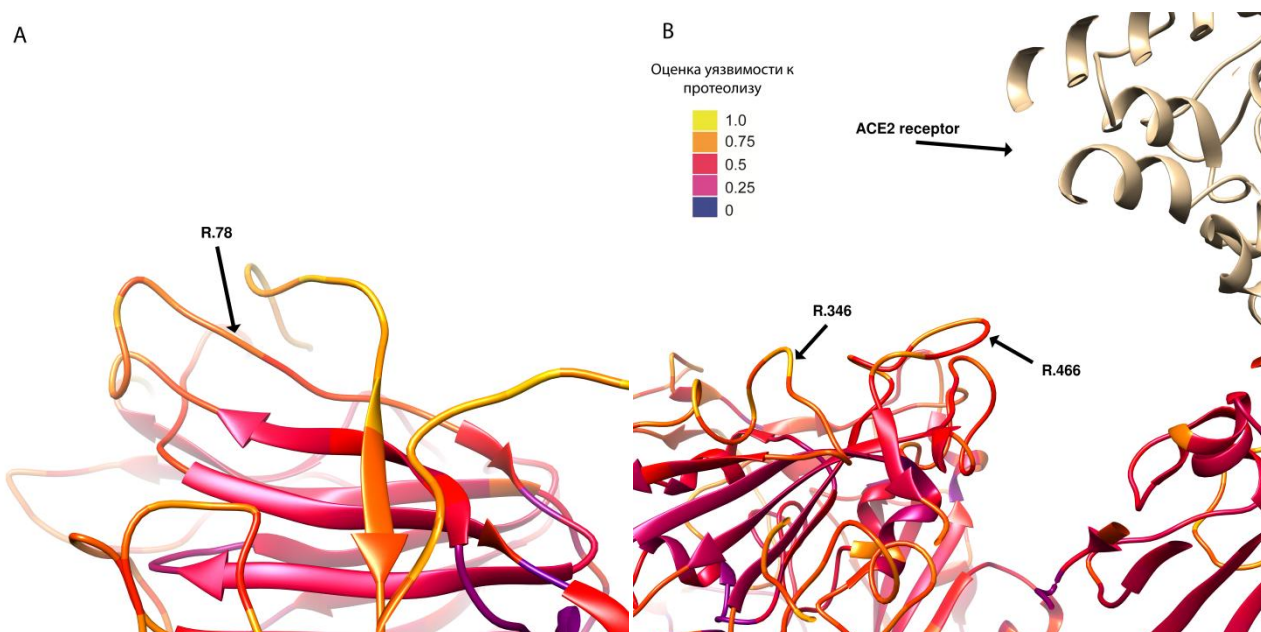


Рисунок 21. (а) Потенциальный сайт расщепления в позиции R78 расположен в N-концевом домене S-гликопротеина. (б) Потенциальные сайты расщепления в позициях R346 и R466 расположены в пределах рецептор-связывающего домена S-гликопротеина.

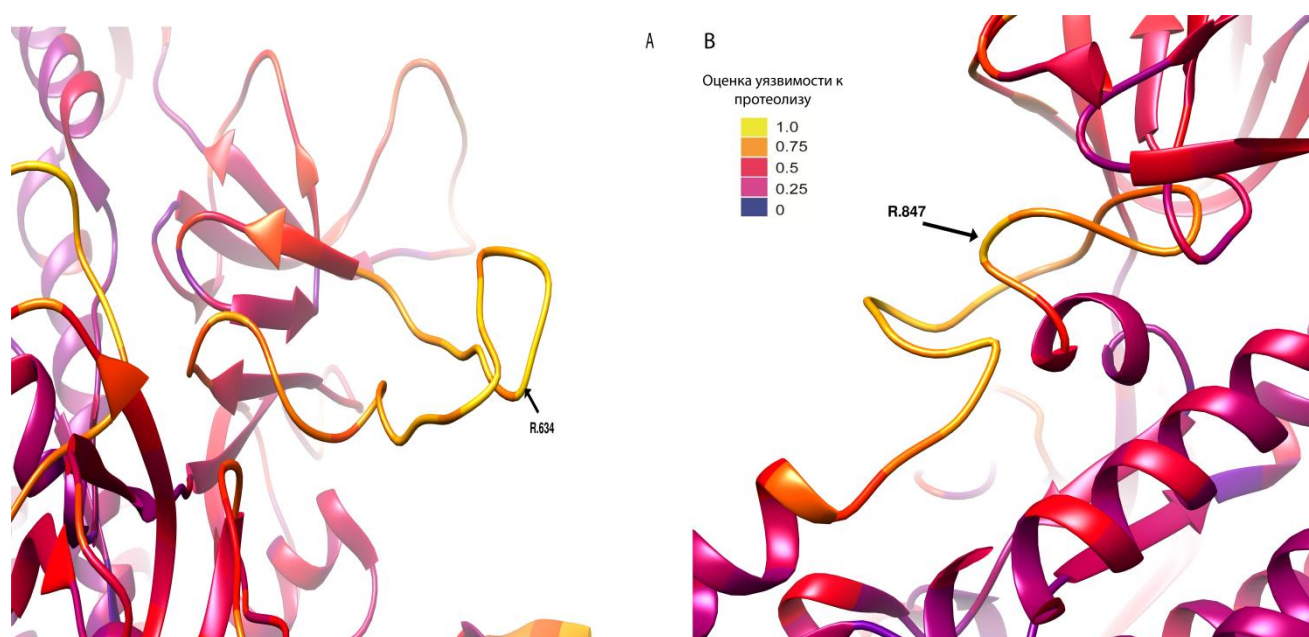


Рисунок 22. (а) Потенциальный сайт расщепления в позиции R634 расположен в, так называемой, «петле 630» S-гликопротеина. (б) Потенциальный сайт расщепления в позиции R847 примыкает к пептиду слияния.

Дополнительно были проанализированы оценки расщепления в указанных позициях, полученные всеми 169 моделями субстратной специфичности протеаз человека. Согласно результатам (Рис. 23), сайт протеолиза в позиции R78 потенциально могут расщеплять протеазы семейства PCSK, причём оценки

расщепления для данных протеаз были сопоставимы между собой и значительно превышали соответствующие оценки, полученные для других протеаз.

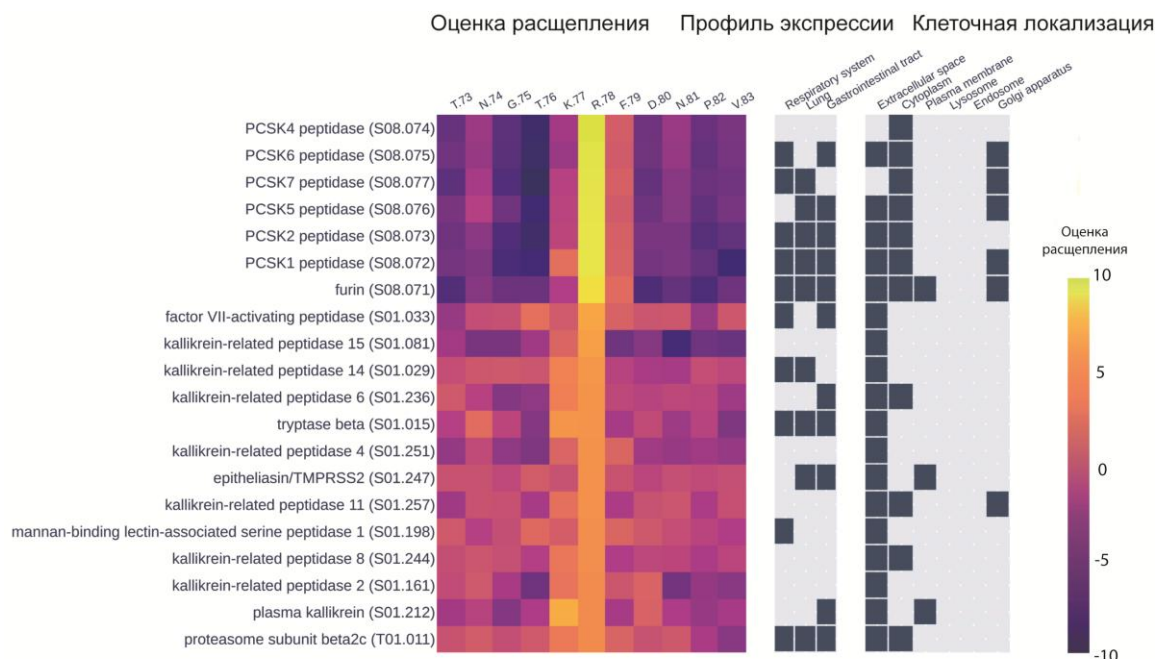


Рисунок 23. Двадцать наивысших оценок расщепления в позиции R78, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель протеазы PCSK4. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

Высокие оценки расщепления для позиции R346 также продемонстрировали модели семейства пропротеин-конвертаз, причем, наивысшие оценки были получены моделями протеаз PCSK5 и фурина (Рис. 24). Кроме того, потенциальную способность к расщеплению в данной позиции показали модели протеазы DESC1 и трипсиноподобной протеазы дыхательных путей человека, которые являются трансмембранными сериновыми протеазами, а также матриксной металлопротеазы MMP11 и некоторых других протеаз, не относящихся к четырём выбранным семействам.

Наивысшую оценку расщепления в позиции R466 продемонстрировала модель субстратной специфичности катепсина С, также известного как дипептидил-пептидаза 1. Примечательно, что многие протеазы, которые ранее не фигурировали в анализе, обладают потенциалом к расщеплению в данной позиции: PHEX протеаза, гранзим К и некоторые другие протеазы (Рис. 25).

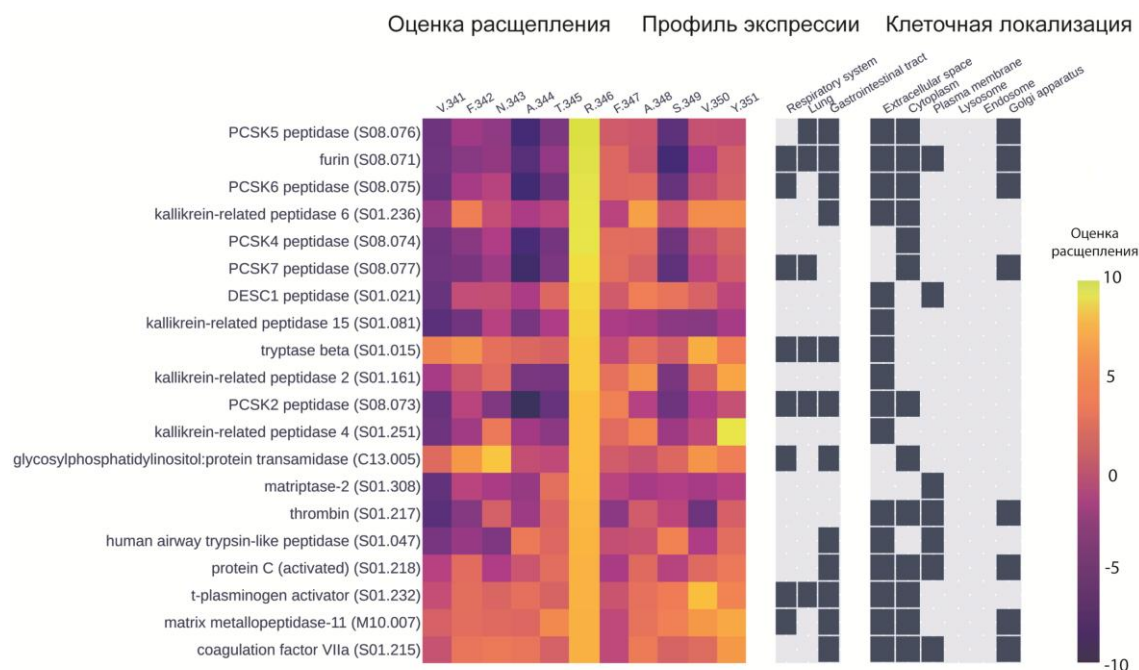


Рисунок 24. Двадцать наивысших оценок расщепления в позиции R346, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель протеазы PCSK5. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

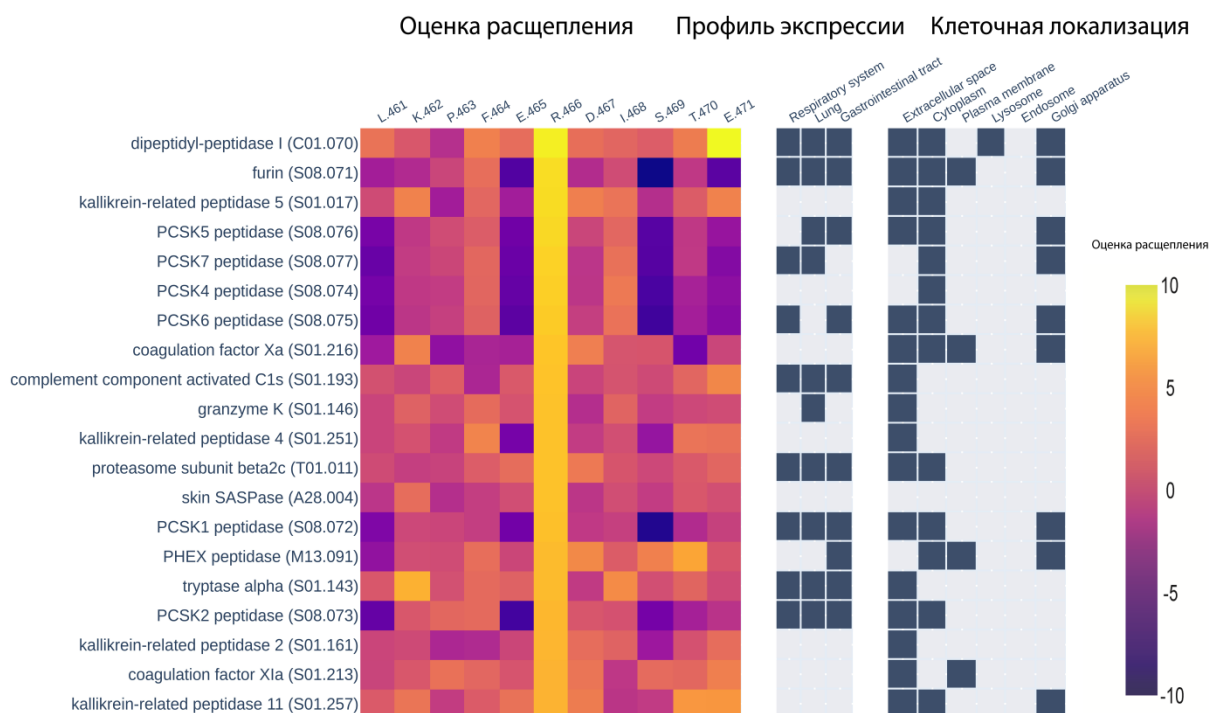


Рисунок 25. Двадцать наивысших оценок расщепления в позиции R466, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель протеазы дипептидил-пептидаза 1. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

В позиции R634 наивысшую оценку продемонстрировала модель активатора плазминогена урокиназного типа. Многие факторы свёртывания крови, такие как

плазмин, факторы свёртывания крови IXa, Xa и XIa, также показали существенные значения. Среди цистеиновых протеаз, лишь протеазы SENP5 и SENP2 показали высокие оценки протеолитической активации в позиции R634 (Рис. 26).

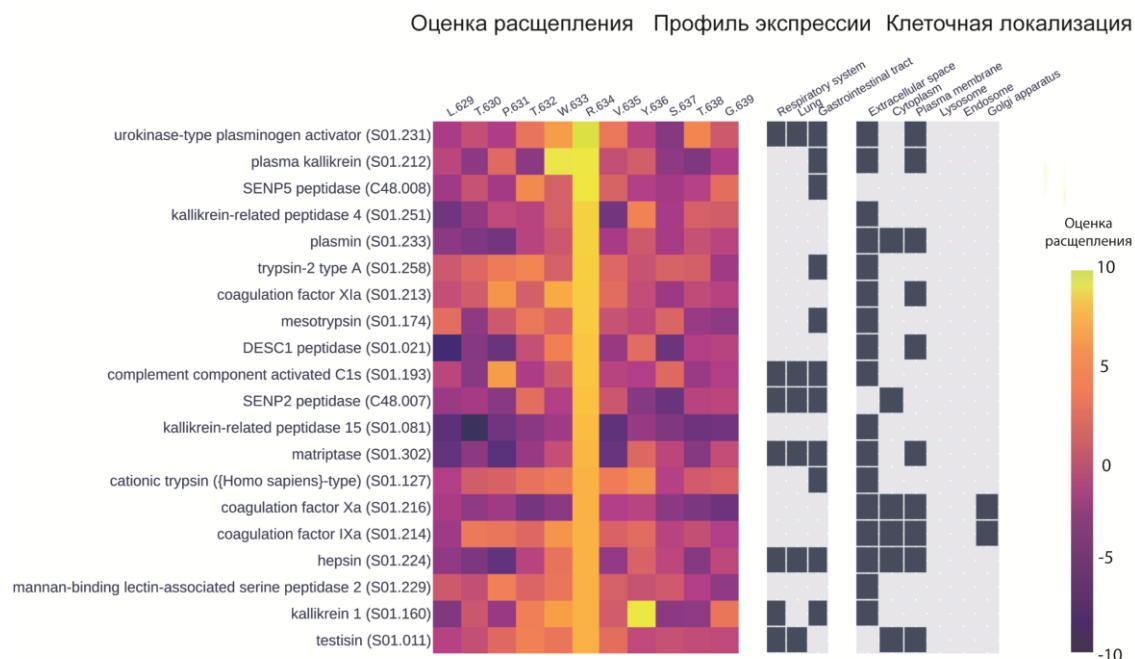


Рисунок 26. Двадцать наивысших оценок расщепления в позиции R634, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель активатора плазминогена урокиназного типа. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

В позиции R847 наивысшую оценку расщепления показала модель, описывающая субстратную специфичность протеазы DESC1. В целом, можно отметить, что многие протеазы, связанные с мембраной, такие как матриптазы, а также ряд металлопротеаз, включая ADAM17 и меприн, имеют высокие оценки расщепления в данной позиции (Рис. 27).

Для наглядного представления о различиях между оценками расщепления для потенциальных и известных сайтов расщепления, была сформирована тепловая карта, отображающая оценки расщепления всех протеолитических сайтов, упомянутых в этом исследовании. Оценки были рассчитаны только для протеаз из четырёх семейств — PCSK, TTSP, калликреинов и факторов свёртывания крови, а также для катепсина L (Рис. 28).

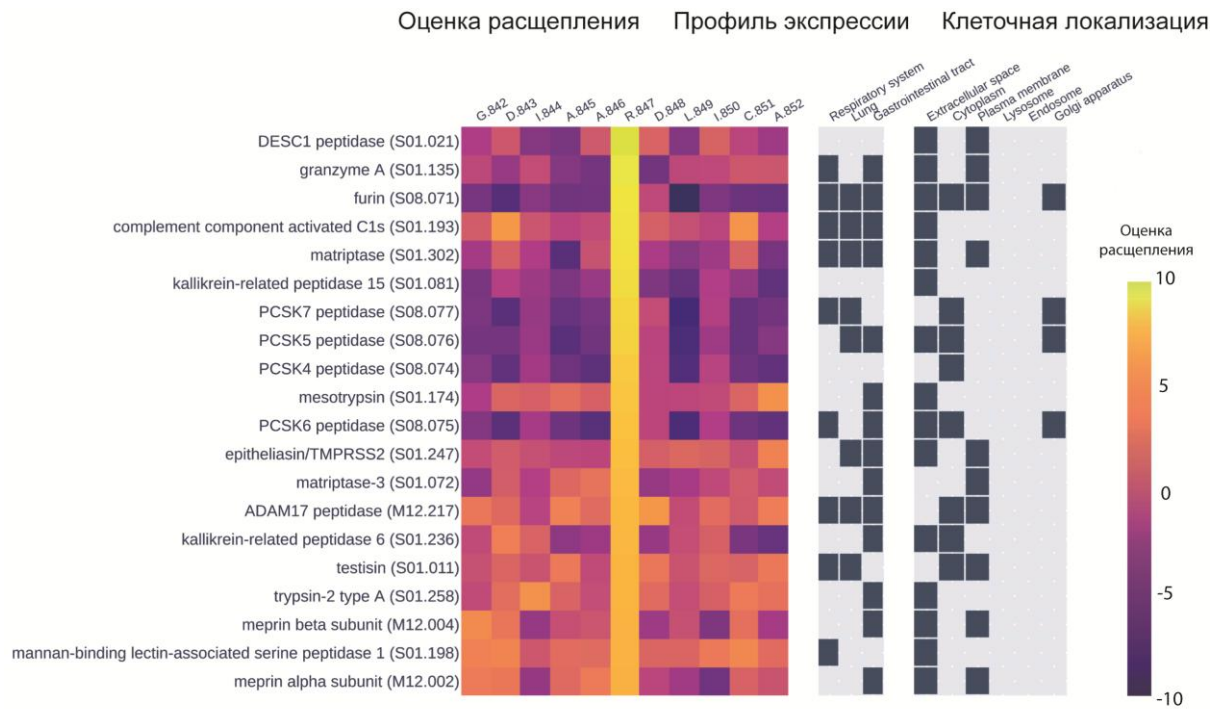


Рисунок 27. Двадцать наивысших оценок расщепления в позиции R847, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Наивысшую оценку демонстрирует модель протеазы DESC1. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

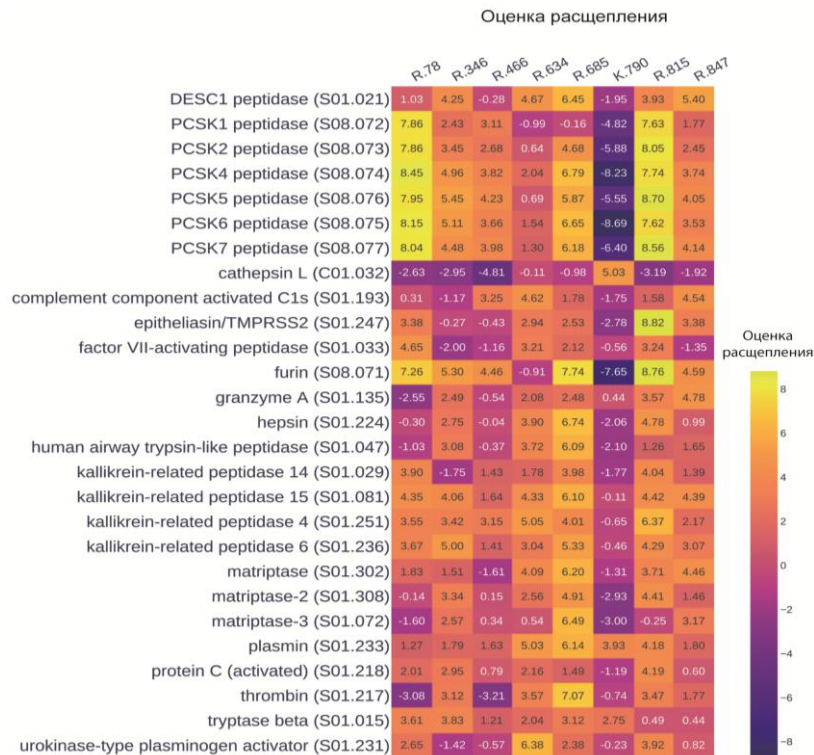


Рисунок 28. Оценки расщепления, рассчитанные для известных и потенциальных сайтов в S-гликопротеине, с помощью моделей субстратной специфичности протеаз семейств PCSK, TTSP, калликреинов и факторов свёртывания крови.

3.5. Влияние мутаций в S-гликопротеине оболочки коронавируса SARS-CoV-2 на эффективность расщепления сайтов протеолиза

Модели субстратной специфичности протеаз, разработанные в данном 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 [142].

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. Для оценки влияния указанных аминокислотных замен была построена расширенная модель субстратной специфичности фуриновой протеазы, охватывающая рамку P5-P5', что позволило учесть влияние удалённых от точки расщепления аминокислот. Применение данной модели 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. 29). 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.

Далее были рассчитаны оценки расщепления сайта S1/S2 моделью субстратной специфичности фурина для всех возможных аминокислотных замен в позиции P681. Согласно полученным результатам (Рис. 30), 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 (Рис. 31, Рис. 32).

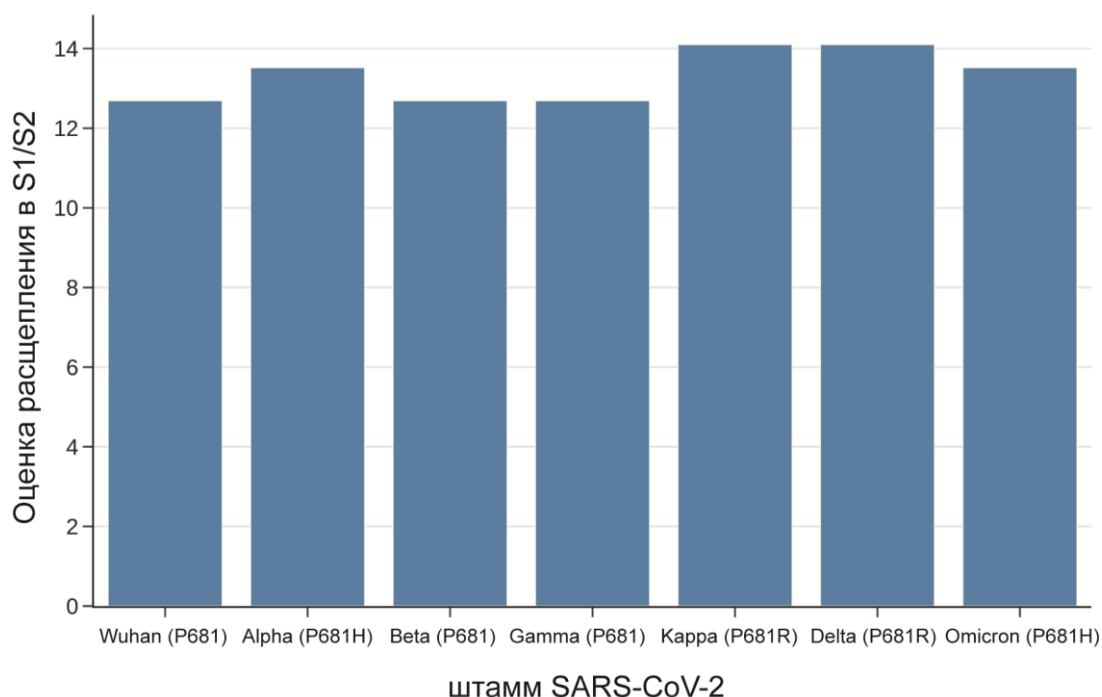


Рисунок 29. Оценки расщепления сайта S1/S2 S-гликопротеина, рассчитанные моделью субстратной специфичности фуриновой протеазы, для различных мутаций в позиции 681, наблюдаемых у отдельных штаммов коронавируса SARS-CoV-2.

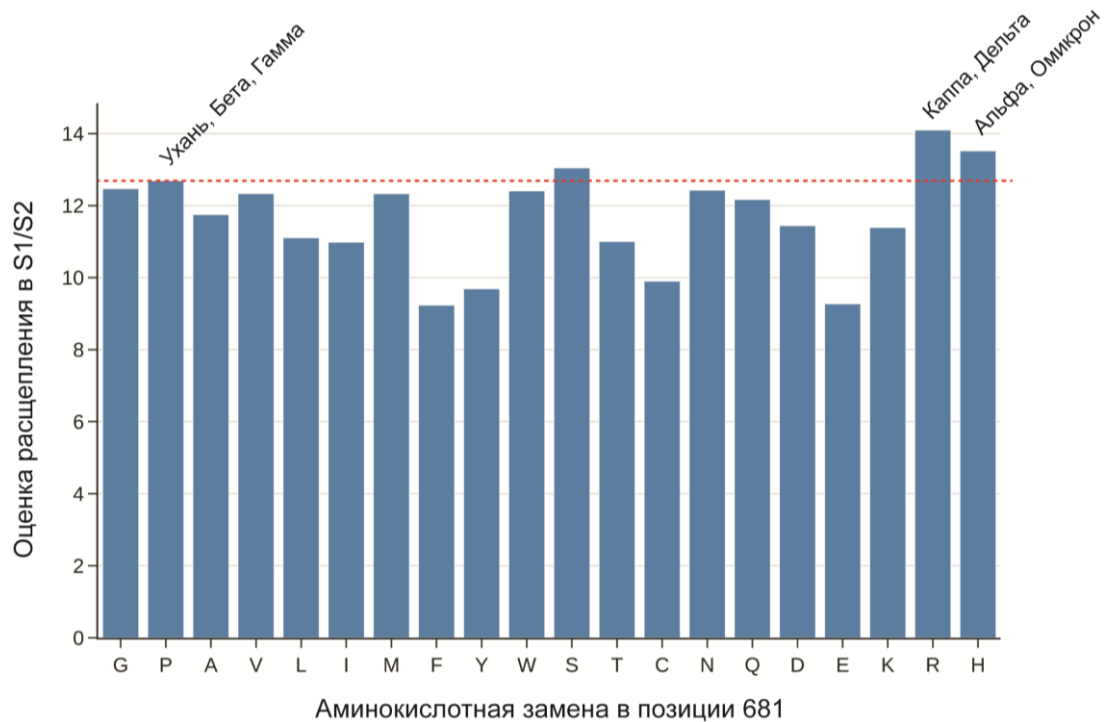


Рисунок 30. Оценки расщепления сайта S1/S2, рассчитанные моделью субстратной специфичности фуриновой протеазы для всех возможных замен в позиции 681.

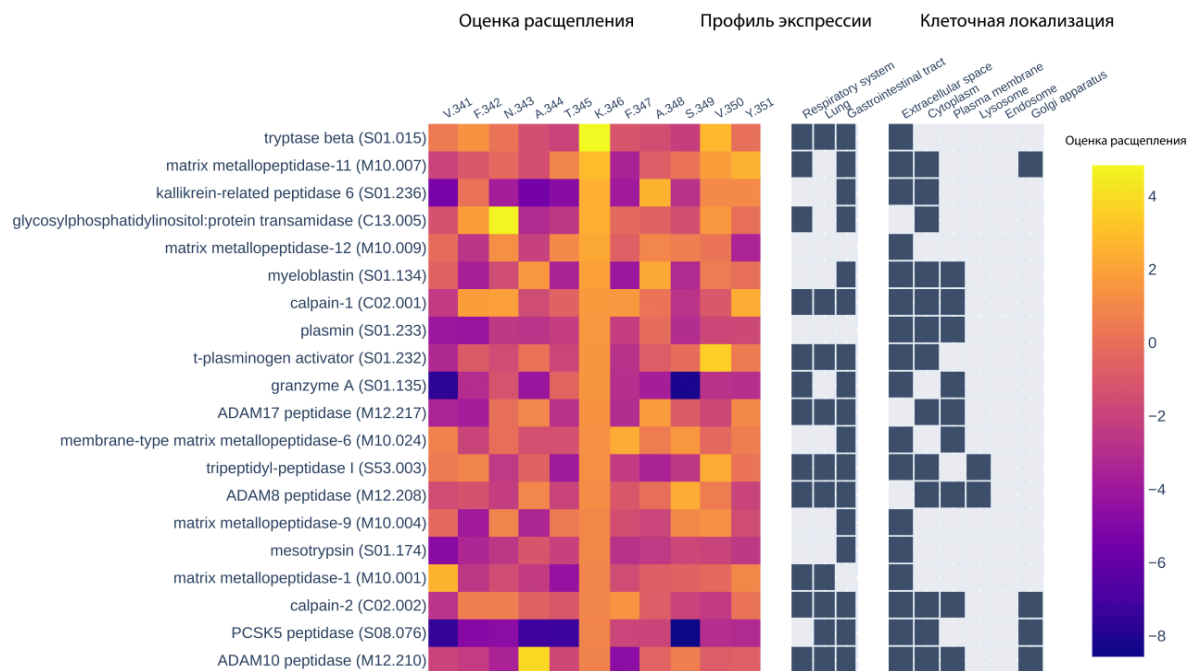


Рисунок 31. Двадцать наивысших оценок расщепления в позиции 346 при замене R346K, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

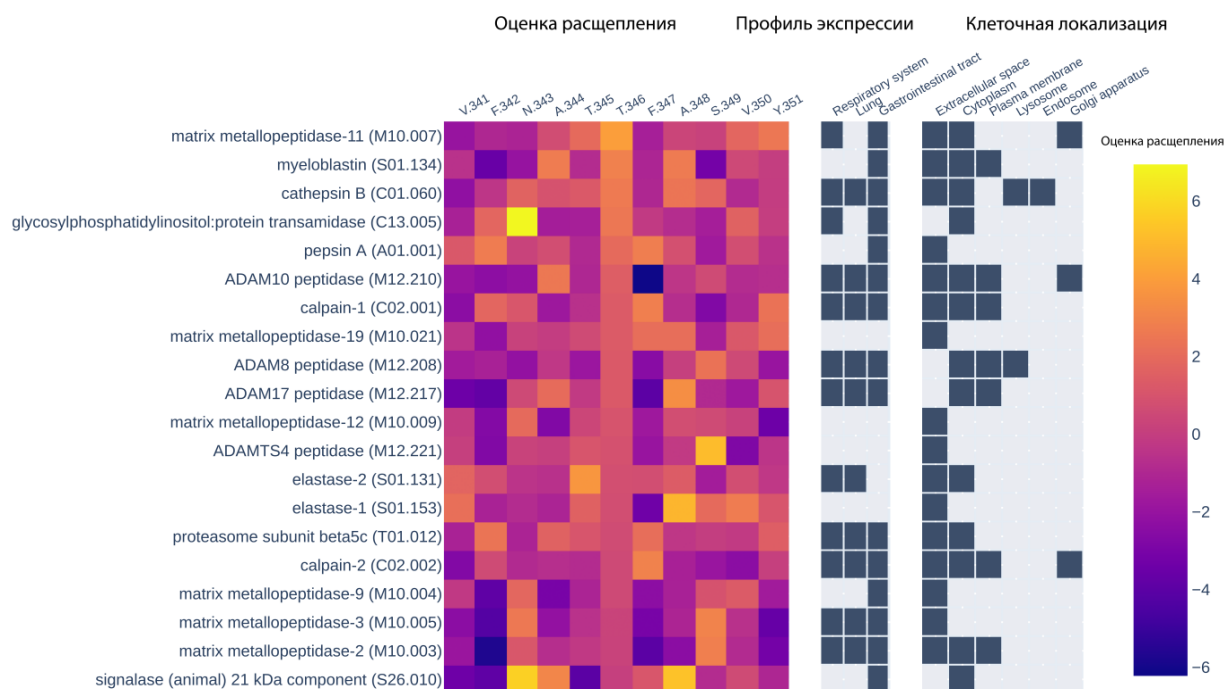


Рисунок 32. Двадцать наивысших оценок расщепления в позиции 346 при замене R346T, полученных с помощью моделей субстратной специфичности 169 протеаз человека. Также представлены данные о клеточной локализации и профилях экспрессии соответствующих протеаз.

3.6. Обсуждение результатов

1) Были разработаны 169 моделей субстратной специфичности протеаз человека на основе информации из базы данных MEROPS. Проведённый анализ распределения оценок расщепления показал, что модели позволяют уверенно отличать известные сайты протеолиза от случайных позиций. Для валидации подхода модели были применены к S-белку коронавируса SARS-CoV-2, где наивысшие оценки расщепления в позициях S1/S2 и S2' были получены для фурина и TMPRSS2, соответственно, что согласуется с современными представлениями о протеолитической активации S-белка.

2) Разработанные модели были использованы для биоинформатической идентификации протеаз человека, которые потенциально могут участвовать в протеолитической активации S-гликопротеина. Анализ показал, что ряд пропротеин-конвертаз кексинового типа, такие как PCSK5, PCSK6 и PCSK7, обладают высоким потенциалом расщепления в позиции S1/S2. В отношении

сайта S2` наивысшие оценки расщепления продемонстрировали протеазы, аналогично TMPRSS2, ассоциированные с клеточной мембраной, а именно — гепсин, матриптазы, активатор плазминогена урокиназного типа. В целом, к числу семейств протеаз, потенциально способных активировать S-белок, были отнесены следующие четыре семейства: пропротеин-конвертазы кексинового типа (PCSK), трансмембранные сериновые протеазы (TTSP), калликреины и факторы свёртывания крови.

3) На основе оценок расщепления, полученных с помощью модели субстратной специфичности катепсина L и универсальной структурной модели, был предсказан потенциальный сайт расщепления S-белка в позиции K790. Данный сайт расщепления предложен впервые и не совпадает с ранее предполагаемыми участками расщепления катепсином L. Пространственный анализ подтвердил функциональную значимость данной позиции: K790 располагается в непосредственной близости к сайту S2` и примыкает к пептиду слияния. Это позволяет предположить, что расщепление в данной позиции может вызывать конформационные изменения S-белка, аналогичные тем, которые происходят при активации сайта S2`, приводя к высвобождению пептида слияния, образованию поры и проникновению вирусной частицы в цитоплазму клетки.

4) Было проанализировано влияние мутаций в S-гликопротеине известных штаммов коронавируса SARS-CoV-2 на эффективность расщепления сайтов протеолиза. Варианты вируса были проверены на наличие мутаций в окрестностях сайтов расщепления S1/S2 и S2`. Было обнаружено, что в четырёх штаммах коронавируса (Альфа, Омикрон, Каппа и Дельта) присутствуют мутации в позиции P681, входящей в рамку P5-P5' фуринового сайта S1/S2: замена P681H — в штаммах Альфа и Омикрон, P681R - в штаммах Каппа и Дельта. Оценки расщепления, рассчитанные моделью субстратной специфичности фурина, показали, что данные замены повышают эффективность расщепления сайта S1/S2 фуриновой протеазой, по сравнению с исходным (уханьским) штаммом, что согласуется с предположением о связи этих мутаций с повышенной инфекционностью соответствующих штаммов. Анализ остальных возможных

замен в позиции Р681 показал, что все они, за исключением замены Р681S, ассоциированы со снижением оценок расщепления S1/S2 сайта. Таким образом, данный анализ показывает возможность предсказания эволюции штаммов в участках генома, кодирующих сайты протеолиза, используя биоинформатические методы.

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.

БЛАГОДАРНОСТИ

Автор выражает глубочайшую признательность своему научному руководителю Марату Джамалудиновичу Казанову за неоценимую и всестороннюю поддержку и наставничество при выполнении данной работы, а также своей семье и друзьям за моральную поддержку и понимание на протяжении всего времени исследования и написания текста диссертации. Работа посвящается Владимиру Николаевичу Шелудянкину (1938-2025).

СПИСОК ЛИТЕРАТУРЫ

1. *López-Otín, C.* Proteases: Multifunctional Enzymes in Life and Disease. / C. López-Otín, J. S. Bond // J. Biol. Chem. — 2008. — Vol. 283 (45). — P. 30433–30437. — doi: 10.1074/jbc.R800035200.
2. *López-Otín, C.* Protease degradomics: a new challenge for proteomics / C. López-Otín, C. M. Overall // Nat Rev Mol Cell Biol. — 2002. — Vol. 3(7). — P. 509–19. — doi: 10.1038/nrm858.
3. *Turk, B.* Protease signaling: the cutting edge / B. Turk, D. Turk, V. Turk // EMBO J. — 2012. — Vol. 31(7). — P. 1630–43. — doi: 10.1038/emboj.2012.42.
4. *Gutierrez-Mendez, N.* Proteolysis and Rheological Properties of Cream Cheese Made with a Plant-Derived Coagulant from *Solanum elaeagnifolium* / N. Gutierrez-Mendez, A. Balderrama-Carmona, S. E. Garcia-Sandoval, et al. // Foods. — 2019. — Vol. 8(2). — P. 44. — doi: 10.3390/foods8020044
5. *Mamo, J.* The Role of Microbial Aspartic Protease Enzyme in Food and Beverage Industries / J. Mamo, F. Assefa // Journal of Food Quality. — 2018. — Vol. 3. — P. 1–15. — DOI:10.1155/2018/7957269
6. *Dyer, R. P.* Making the Cut with Protease Engineering / R. P. Dyer, G. A. Weiss // Cell Chem Biol. — 2021. — Vol. 29(2). — P. 177–190. — doi: 10.1016/j.chembiol.2021.12.001
7. *Jamal, G. A.* Proteases, a powerful biochemical tool in the service of medicine, clinical and pharmaceutical / G. A. Jamal, E. Jahangirian, M. R. Hamblin, et al. // Prep Biochem Biotechnol. — 2025. — Vol. 55(1). — P. 1–25. — doi: 10.1080/10826068.2024.2364234.
8. *Drag, M.* Emerging principles in protease-based drug discovery / M. Drag, G. S. Salvesen // Nat Rev Drug Discov. — 2010. — Vol. 9(9). — P. 690–701. — doi: 10.1038/nrd3053

9. *Song, J.* Bioinformatic approaches for predicting substrates of proteases / J. Song, H. Tan, S. E. Boyd, et al. / J. Bioinform Comput Biol. — 2011. — Vol. 9(1). — P. 149–78. — doi: 10.1142/s0219720011005288.
10. *Bhagwat, S. R.* Proteolysis to Identify Protease Substrates: Cleave to Decipher / S. R. Bhagwat, K. Hajela, A. Kumar // Proteomics. — 2018. — Vol. 18(13). — P. e1800011. — doi: 10.1002/pmic.201800011.
11. *Overall, C. M.* Degradomics: systems biology of the protease web. Pleiotropic roles of MMPs in cancer / C. M. Overall, R. A. Dean // Cancer Metastasis Rev. — 2006. — Vol. 25(1). — P. 69–75. — doi: 10.1007/s10555-006-7890-0.
12. *auf dem Keller, U.* Protease research in the era of systems biology / U. auf dem Keller, A. Doucet, C. M. Overall // Biol Chem. — 2007. — Vol. 388(11) — P. 1159–62. — doi: 10.1515/BC.2007.146.
13. *Doucet, A.* Metadegradomics: toward in vivo quantitative degradomics of proteolytic post-translational modifications of the cancer proteome / A. Doucet, G. S. Butler, D. Rodriguez, et al. // Mol Cell Proteomics. — 2008. — Vol. 7(10). — P. 1925–51. — doi: 10.1074/mcp.R800012-MCP200.
14. *Kruger, A.* Functional genetic mouse models: promising tools for investigation of the proteolytic internet / A. Kruger // Biol Chem. — 2009. — Vol. 390(2). — P. 91–7. — doi: 10.1515/BC.2009.015.
15. *Mason, S. D.* Proteolytic networks in cancer / S. D. Mason, J. A. Joyce // Trends Cell Biol. — 2011. — Vol. 21(4). — P. 228–237. — doi: 10.1016/j.tcb.2010.12.002.
16. *Li, F.* Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods / F. Li, Y. Wang, C. Li, et al. // Brief Bioinform. — 2019. — Vol. 20(6). — P. 2150–2166. — doi: 10.1093/bib/bby077.
17. *Kazanov, M. D.* Structural determinants of limited proteolysis / M. D. Kazanov, Y. Igarashi, A. M. Eroshkin, et al. // J Proteome Res. — 2011. — Vol. 10(8). — P. 3642–51. — doi: 10.1021/pr200271w.

18. *Igarashi, Y.* CutDB: a proteolytic event database / Y. Igarashi, A. Eroshkin, S. Gramatikova, et al. // *Nucleic Acids Res.* — 2007. — Vol. 35. — P. D546–9. — doi: 10.1093/nar/gkl813.
19. *Rawlings, N. D.* MEROPS: the peptidase database / N. D. Rawlings, F. R. Morton, C. Y. Kok, et al. // *Nucleic Acids Res.* — 2008. — Vol. 36. — P. D320–5. — doi: 10.1093/nar/gkm954.
20. *Barrett, A. J.* Cellular proteolysis. An overview / A. J. Barrett // *Ann N Y Acad Sci.* — 1992. — Vol. 674. — P. 1–15. — doi: 10.1111/j.1749-6632.1992.tb27472.x.
21. *Schechter, I.* On the size of the active site in proteases. I. Papain / I. Schechter, A. Berger // *Biochem Biophys Res Commun.* — 1967. — Vol. 27(2). — P. 157–62. — doi: 10.1016/s0006-291x(67)80055-x.
22. *Bond, J. S.* Proteases: History, discovery, and roles in health and disease / J. S. Bond // *J Biol Chem.* — 2019. — Vol. 294(5). — P. 1643–1651. — doi: 10.1074/jbc.TM118.004156.6
23. *Ramazi, S.* Posttranslational modifications in proteins: resources, tools and prediction methods / S. Ramazi, J. Zahiri // *Database (Oxford).* — 2021. — Vol. 2021. — P. baab012. — doi: 10.1093/database/baab012
24. *Barrett, A. J.* Oligopeptidases, and the emergence of the prolyl oligopeptidase family / A. J. Barrett, N. D. Rawlings // *Biol Chem Hoppe Seyler.* — 1992. — Vol. 373(7). — P. 353–60. — doi: 10.1515/bchm3.1992.373.2.353.
25. *Mentlein, R.* Cell-Surface Peptidases / R. Mentlein // *Int Rev Cytol.* — 2004. — Vol. 235. — P. 165–213. — doi: 10.1016/S0074-7696(04)35004-7
26. *Di Cera, E.* Serine Proteases / E. Di Cera // *IUBMB Life.* — 2009. — Vol. 61(5). — P. 510–515. — <https://doi.org/10.1002/iub.186>
27. *Grzonka, Z.* Cysteine Proteases / Z. Grzonka, F. Kasprzykowski, W. Wiczak // *Industrial Enzymes.* — 2007. — P. 181–195. — https://doi.org/10.1007/1-4020-5377-0_11

28. *Baird Jr, T. T.* Conversion of trypsin to a functional threonine protease / T. T. Baird Jr, W. D. Wright, C. S. Craik // *Protein Sci.* — 2006. — Vol. 15(6). — P. 1229–38. — doi: 10.1110/ps.062179006.
29. *Nagase, H.* Metalloproteases / H. Nagase // *Curr Protoc Protein Sci.* — 2001. — P. 21.4.1–21.4.13. — doi: 10.1002/0471140864.ps2104s24.
30. *Dunn, B. M.* Retroviral proteases / B. M. Dunn, M. M. Goodenow, A. Gustchina, et al. // *Genome Biol.* — 2002. — Vol. 3(4). — P. reviews3006.1–reviews3006.7. — doi: 10.1186/gb-2002-3-4-reviews3006
31. *Oda, K.* New families of carboxyl peptidases: serine-carboxyl peptidases and glutamic peptidases / K. Oda // *The Journal of Biochemistry.* — 2012. — Vol. 151(1). — P. 13–25. — <https://doi.org/10.1093/jb/mvr129>
32. *Bertenshaw, G. P.* Structure of homo- and hetero-oligomeric meprin metalloproteases. Dimers, tetramers, and high molecular mass multimers / G. P. Bertenshaw, M. T. Norcum, J. S. Bond // *J Biol Chem.* — 2003. — Vol. 278(4). — P. 2522–32. — doi: 10.1074/jbc.M208808200.
33. *Fischer, J. A.* Deubiquitinating enzymes: their roles in development, differentiation, and disease / A. J. Fischer // *Int Rev Cytol.* — 2003. — Vol. 229. — P. 43–72. — doi: 10.1016/s0074-7696(03)29002-1.
34. *Rosenblum, J. S.* Autoproteolysis in nucleoporin biogenesis / J. S. Rosenblum, G. Blobel // *Proc Natl Acad Sci U S A.* — 1999. — Vol. 96(20). — P. 11370–5. — doi: 10.1073/pnas.96.20.11370.
35. *Wei, W.* Characterization of cis-autoproteolysis of polycystin-1, the product of human polycystic kidney disease 1 gene / W. Wei, K. Hackmann, H. Xu, et al. // *J Biol Chem.* — 2007. — Vol. 282(30). — P. 21729–37. — doi: 10.1074/jbc.M703218200.
36. *Ciechanover, A.* Tracing the history of the ubiquitin proteolytic system: the pioneering article / A. Ciechanover // *Biochem Biophys Res Commun.* — 2009. — Vol. 387(1). — P. 1–10. — doi: 10.1016/j.bbrc.2009.06.065.

37. *Ono, Y.* Calpains: an elaborate proteolytic system / Y. Ono, H. Sorimachi // *Biochim Biophys Acta*. — 2012. — Vol. 1824(1). — P. 224–36. — doi: 10.1016/j.bbapap.2011.08.005.
38. *Tundo, G. R.* Multiple functions of insulin-degrading enzyme: a metabolic crosslight? / G. R. Tundo, D. Sbardella, C. Ciaccio, et al. // *Crit Rev Biochem Mol Biol*. — 2017. — Vol. 52(5). — P. 554–582. — doi: 10.1080/10409238.2017.1337707.
39. *Babkova, K.* Prolyl oligopeptidase and its role in the organism: attention to the most promising and clinically relevant inhibitors / K. Babkova, J. Korabecny, O. Soukup, et al. // *Future Med Chem*. — 2017. — Vol. 9(10). — P. 1015–1038. — doi: 10.4155/fmc-2017-0030.
40. *Turk, B.* Lysosomes as “Suicide Bags” in Cell Death: Myth or Reality? / B. Turk, V. Turk // *J Biol Chem*. — 2009. — Vol. 284(33). — P. 21783–21787. — doi: 10.1074/jbc.R109.023820
41. *Seidah, N. G.* The activation and physiological functions of the proprotein convertases / N. G. Seidah, G. Mayer, A. Zaid, et al. // *Int J Biochem Cell Biol*. — 2008. — Vol. 40(6–7). — P. 1111–25. — doi: 10.1016/j.biocel.2008.01.030.
42. *Braun, H. P.* The mitochondrial processing peptidase / H. P. Braun, U. K. Schmitz // *Int J Biochem Cell Biol*. — 1997. — Vol. 29(8–9). — P. 1043–5. — doi: 10.1016/s1357-2725(97)00032-0.
43. *Sedo, A.* Cell membrane-bound peptidases: not “only” proteolysis / A. Sedo, V. Mandys, E. Krepela // *Physiol Res*. — 1996. — Vol. 45(3). — P. 169–76.
44. *Wolfe, M. S.* Intramembrane Proteolysis / M. S. Wolfe // *Chem Rev*. — 2009. — Vol. 109. — P. 1599–1612. — doi: 10.1021/cr8004197
45. *Sevenich, L.* Pericellular proteolysis in cancer / L. Sevenich, J. A. Joyce // *Genes Dev*. — 2014. — Vol. 28. — P. 2331–2347. — doi:10.1101/gad.250647.114
46. *Schauperl, M.* Characterizing Protease Specificity: How Many Substrates Do We Need? / M. Schauperl, J. E. Fuchs, B. J. Waldner, et al. // *PLoS One*. — 2015. — Vol. 10(11). — P. e0142658. — doi: 10.1371/journal.pone.0142658.

47. *Fernandez-Fernandez, A. D.* Mechanisms controlling plant proteases and their substrates / A. D. Fernandez-Fernandez, S. Stael, F. Van Breusegem // *Cell Death Differ.* — 2023. — Vol. 30(4). — P. 1047–1058. — doi: 10.1038/s41418-023-01120-5
48. *Balint-Kurti, P.* The plant hypersensitive response: concepts, control and consequences / P. Balint-Kurti // *Mol Plant Pathol.* — 2019. — Vol. 20(8). — P. 1163–1178. — doi: 10.1111/mpp.12821
49. *Monod, M.* Secreted proteases from pathogenic fungi / M. Monod, S. Capoccia, B. Lechenne, et al. // *Int J Med Microbiol.* — 2002. — Vol. 292(5–6). — P. 405–19. — doi: 10.1078/1438-4221-00223.
50. *Hegedus, D. D.* *Sclerotinia sclerotiorum*: when "to be or not to be" a pathogen? / D. D. Hegedus, S. R. Rimmer // *FEMS Microbiol Lett.* — 2005. — Vol. 251(2). — P. 177–84. — doi: 10.1016/j.femsle.2005.07.040.
51. *Mahmoud, S. A.* Regulated Proteolysis in Bacteria / S. A. Mahmoud, P. Chien // *Annu Rev Biochem.* — 2018. — Vol. 87. — P. 677–696. — doi: 10.1146/annurev-biochem-062917-012848
52. *Joshi, K. K.* Regulated Proteolysis in Bacteria: *Caulobacter* / K. K. Joshi, P. Chien // *Annu Rev Genet.* — 2017. — Vol. 50. — P. 423–445. — doi: 10.1146/annurev-genet-120215-035235
53. *Kristensen, S. S.* The role of site-2-proteases in bacteria: a review on physiology, virulence, and therapeutic potential / S. S. Kristensen, D. B. Diep, M. Kjos, et al. // *MicroLife.* — 2023. — Vol. 4. — P. uqad025. — doi: 10.1093/femsml/uqad025
54. *Tong, L.* Viral proteases / L. Tong // *Chem Rev.* — 2002. — Vol. 102(12). — P. 4609–26. — doi: 10.1021/cr010184f.
55. *Klenk, H. D.* Host cell proteases controlling virus pathogenicity / H. D. Klenk, W. Garten // *Trends Microbiol.* — 1994. — Vol. 2(2). — P. 39–43. — doi: 10.1016/0966-842x(94)90123-6.
56. *Yamauchi, Y.* Virus entry at a glance / Y. Yamauchi, A. Helenius // *J Cell Sci.* — 2013. — Vol. 126(Pt 6). — P. 1289–95. — doi: 10.1242/jcs.119685.

57. *Izaguirre, G.* The Proteolytic Regulation of Virus Cell Entry by Furin and Other Proprotein Convertases / G. Izaguirre // *Viruses*. — 2019. — Vol. 11(9). — P. 837. — doi: 10.3390/v11090837
58. *Seidah, N. G.* The biology and therapeutic targeting of the proprotein convertases / N. G. Seidah, A. Prat // *Nat Rev Drug Discov*. — 2012. — Vol. 11(5). — P. 367–83. — doi: 10.1038/nrd3699.
59. *Izaguirre, G.* Specific and Selective Inhibitors of Proprotein Convertases Engineered by Transferring Serpin B8 Reactive-site and Exosite Determinants of Reactivity to the Serpin α 1PDX / G. Izaguirre, M. Arciniega, A. G. Quezada // *Biochemistry*. — 2019. — Vol. 58(12). — P. 1679–1688. — doi: 10.1021/acs.biochem.8b01295
60. *Glauser, D. L.* Glycoprotein B Cleavage Is Important for Murid Herpesvirus 4 To Infect Myeloid Cells / D. L. Glauser, R. Milho, B. Frederico, et al. // *J Virol*. — 2013. — Vol. 87(19). — 10828–10842. — doi: 10.1128/JVI.00709-13
61. *Garten, W.* Influenza virus activating host proteases: Identification, localization and inhibitors as potential therapeutics / W. Garten, C. Braden, A. Arendt, et al. // *Eur J Cell Biol*. — 2015. — Vol. 94(7–9). — P. 375–83. — doi: 10.1016/j.ejcb.2015.05.013.
62. *Chaipan, C.* Proteolytic Activation of the 1918 Influenza Virus Hemagglutinin / C. Chaipan, D. Kobasa, S. Bertram // *J Virol*. — 2009. — Vol. 83(7). — P. 3200–3211. — doi: 10.1128/JVI.02205-08
63. *Ortmann, D.* Proteolytic cleavage of wild type and mutants of the F protein of human parainfluenza virus type 3 by two subtilisin-like endoproteases, furin and Kex2. / D. Ortmann, M. Ohuchi, H. Angliker, et al. // *J Virol*. — 1994. — Vol. 68(4). — P. 2772–2776. — doi: 10.1128/jvi.68.4.2772-2776.1994
64. *Pager, C. T.* Cathepsin L Is Involved in Proteolytic Processing of the Hendra Virus Fusion Protein / C. T. Pager, R. E. Dutch // *J Virol*. — 2005. — Vol. 79(20). — P. 12714–12720. — doi: 10.1128/JVI.79.20.12714-12720.2005
65. *Zarkik, S.* Comparative processing of bovine leukemia virus envelope glycoprotein gp72 by subtilisin/kexin-like mammalian convertases / S. Zarkik,

- E. Decroly, R. Wattiez, et al. // FEBS Lett. — 1997. — Vol. 406(1–2). — P. 205–10. — doi: 10.1016/s0014-5793(97)00275-5.
66. *Pasquato, A.* Heparin enhances the furin cleavage of HIV-1 gp160 peptides / A. Pasquato, M. Dettin, A. Basak, et al. // FEBS Lett. — 2007. — Vol. 581(30). — P. 5807–5813. — doi: 10.1016/j.febslet.2007.11.050
 67. *Volchkov, V.* Processing of the Ebola virus glycoprotein by the proprotein convertase furin / V. Volchkov, H. Feidmann, V. A. Volchkova, et al. // Proc Natl Acad Sci U S A. — 1998. — Vol. 95(10). — P. 5762–5767. — doi: 10.1073/pnas.95.10.5762
 68. *Volchkov, V.* Proteolytic processing of Marburg virus glycoprotein / V. Volchkov, V. A. Volchkova, U. Stroner, et al. // Virology. — 2000. — Vol. 268(1). — P. 1–6. — doi: 10.1006/viro.1999.0110.
 69. *Chandran, K.* Endosomal Proteolysis of the Ebola Virus Glycoprotein Is Necessary for Infection / K. Chandran, N. J. Sullivan, U. Felbor, et al. // Science. — 2005. — Vol. 308(5728). — P. 1643–1645. — doi: 10.1126/science.1110656
 70. *Rojek, J. M.* Targeting the Proteolytic Processing of the Viral Glycoprotein Precursor Is a Promising Novel Antiviral Strategy against Arenaviruses / J. M. Rojek, G. Pasqual, A. B. Sanchez, et al. // J Virol. — 2009. — Vol. 84(1). — P. 573–584. — doi: 10.1128/JVI.01697-09
 71. *Schiller, J. T.* Current understanding of the mechanism of HPV infection / J. T. Schiller, P. M. Day, R. C. Kines // Gynecol Oncol. — 2010. — Vol. 118. — P. S12–S17. — doi: 10.1016/j.ygyno.2010.04.004
 72. *Mukhopadhyay, S.* A structural perspective of the flavivirus life cycle / S. Mukhopadhyay, R. J. Kuhn, M. G. Rossmann // Nat Rev Microbiol. — 2005. — Vol. 3(1). — P. 13–22. — doi: 10.1038/nrmicro1067.
 73. *Miller, J. K.* Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis / J. K. Millet, G. R. Whittaker // Virus Res. — 2014. — Vol. 202. — P. 120–134. — doi: 10.1016/j.virusres.2014.11.021
 74. *Simmons, G.* Different host cell proteases activate the SARS-coronavirus spike-protein for cell-cell and virus-cell fusion / G. Simmons, S. Bertram,

- I. Glowacka, et al. // *Virology*. — 2011. — Vol. 413(2). — P. 265–274. — doi: 10.1016/j.virol.2011.02.020
75. *Belouzard, S.* Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites / S. Belouzard, V. C. Chu, G. R. Whittaker // *Proc Natl Acad Sci U S A*. — 2009. — Vol. 106(14). — P. 5871–5876. — doi: 10.1073/pnas.0809524106
 76. *Iwata-Yoshikawa, N.* TMPRSS2 Contributes to Virus Spread and Immunopathology in the Airways of Murine Models after Coronavirus Infection / N. Iwata-Yoshikawa, T. Okamura, Y. Shimizu, et al. // *J Virol*. — 2019. — Vol. 93(6). — P. e01815–18. — doi: 10.1128/JVI.01815-18
 77. *Bosch, B. J.* Cathepsin L Functionally Cleaves the Severe Acute Respiratory Syndrome Coronavirus Class I Fusion Protein Upstream of Rather than Adjacent to the Fusion Peptide / B. J. Bosch, W. Bartelink, P. J. M. Rottier // *J Virol*. — 2008. — Vol. 82(17). — P. 8887–8890. — doi: 10.1128/JVI.00415-08
 78. *Baggen, J.* Cellular host factors for SARS-CoV-2 infection / J. Baggen, E. Vanstreels, S. Jansen et al. // *Nat Microbiol*. — 2021. — Vol. 6(10). — P. 1219–1232. — doi: 10.1038/s41564-021-00958-0.
 79. *Jackson, C. B.* Mechanisms of SARS-CoV-2 entry into cells / C. B. Jackson, M. Farzan, B. Chen et al. // *Nat Rev Mol Cell Biol*. — 2022. — Vol. 23(1). — P. 3–20. — doi: 10.1038/s41580-021-00418-x.
 80. *Whittaker, G. R.* Coronavirus entry: how we arrived at SARS-CoV-2 / G. R. Whittaker, S. Daniel, J. K. Millet // *Curr Opin Virol*. — 2021. — Vol. 47. — P. 113–120. — doi: 10.1016/j.coviro.2021.02.006.
 81. *Dimitrov, D. S.* Virus entry: molecular mechanisms and biomedical applications / D. S. Dimitrov // *Nat. Rev. Microbiol*. — 2004. — Vol. 2(2). — P. 109–22. — doi: 10.1038/nrmicro817.
 82. *Mercer, J.* Virus entry by endocytosis / J. Mercer, M. Schelhaas, A. Helenius // *Annu Rev Biochem*. — 2010. — Vol. 79. — P. 803–33. — doi: 10.1146/annurev-biochem-060208-104626.

83. *Shajahan, A.* Glycosylation of SARS-CoV-2: structural and functional insights / A. Shajahan, L. E. Pepi, D. S. Rouhani et al. // *Anal Bioanal Chem.* — 2021. — Vol. 413(29). — P. 7179–7193. — doi: 10.1007/s00216-021-03499-x.
84. *Watanabe, Y.* Site-specific glycan analysis of the SARS-CoV-2 spike / Y. Watanabe, J. D. Allen, D. Wrapp et al. // *Science.* — 2020. — Vol. 369(6501). — P. 330–333. — doi: 10.1126/science.abb9983.
85. *Gomes, C. P.* Cathepsin L in COVID-19: From Pharmacological Evidences to Genetics / C. P. Gomes, D. E. Fernandes, F. Casimiro, et al. // *Front Cell Infect Microbiol.* — 2020. — Vol. 10. — P. 589505. — doi: 10.3389/fcimb.2020.589505
86. *Shankar, P.* Organ Involvement in COVID-19: A Molecular Investigation of Autopsied Patients / P. Shankar, J. Singh, A. Joshi, et al. // *Microorganisms.* — 2022. — Vol. 10(7). — P. 1333. — doi: 10.3390/microorganisms10071333
87. *Li, W.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus / W. Li, M. J. Moore, N. Vasilieva, et al. // *Nature.* — 2003. — Vol. 426(6965). — P. 450–454. — doi: 10.1038/nature02145
88. *Hofmann, H.* Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry / H. Hofmann, K. Pyrc, L. van der Hoek et al. // *Proc Natl Acad Sci U S A.* — 2005. — Vol. 102(22). — P. 7988–93. — doi: 10.1073/pnas.0409465102.
89. *Turner, A. J.* ACE2 Cell Biology, Regulation, and Physiological Functions / A. J. Turner // *The Protective Arm of the Renin Angiotensin System (RAS).* — 2015. — P. 185–189. — doi: 10.1016/B978-0-12-801364-9.00025-0
90. *Takeda, M.* Proteolytic activation of SARS-CoV-2 spike protein / M. Takeda // *Microbiol Immunol.* — 2022. — Vol. 66(1). — P. 15–23. — doi: 10.1111/1348-0421.12945.
91. *Zabiegala, A.* Roles of host proteases in the entry of SARS-CoV-2 / A. Zabiegala, Y. Kim, K-O. Chang // *Anim Dis.* — 2023. — Vol. 3(1). — P. 12. — doi: 10.1186/s44149-023-00075-x.

92. *Fitzgerald, K.* Furin Protease: From SARS-CoV-2 to Anthrax, Diabetes, and Hypertension / K. Fitzgerald // Perm J. — 2020. — Vol. 24. — P. 20.187. — doi: 10.7812/TPP/20.187
93. *Wettstein, L.* The Transmembrane Protease TMPRSS2 as a Therapeutic Target for COVID-19 Treatment / L. Wettstein, F. Kirchhoff, J. Munch // Int J Mol Sci. — 2022. — Vol. 23(3). — P. 1351. — doi: 10.3390/ijms23031351
94. *Bollavaram, K.* Multiple sites on SARS-CoV-2 spike protein are susceptible to proteolysis by cathepsins B, K, L, S, and V / K. Bollavaram, T. H. Leeman, M. W. Lee et al // Protein Sci. — 2021. — Vol. 30(6). — P. 1131–1143. — doi: 10.1002/pro.4073.
95. *Magazine, N.* Mutations and Evolution of the SARS-CoV-2 Spike Protein / N. Magazine, T. Zhang, Y. Wu, et al. // Viruses. — 2022. — Vol. 14(3). — P. 640. — doi: 10.3390/v14030640
96. *Zhang, L.* Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation / L. Zhang, M. Mann, Z. A. Syed et al. // Proc Natl Acad Sci U S A. — 2021. — Vol. 118(47). — P. e2109905118. — doi: 10.1073/pnas.2109905118.
97. *Ostrov, D. A.* Structural Consequences of Variation in SARS-CoV-2 B.1.1.7 / D. A. Ostrov // J Cell Immunol. — 2021. — Vol. 3(2). — P. 103–108. — doi: 10.33696/immunology.3.085
98. *Ng, N. M.* Subsite cooperativity in protease specificity / N. M. Ng, R. N. Pike, S. E. Boyd // Biol Chem. — 2009. — Vol. 390(5–6). — P. 401–7. — doi: 10.1515/BC.2009.065.
99. *Fuchs, J. E.* Cleavage entropy as quantitative measure of protease specificity / J. E. Fuchs, S. von Grafenstein, R. G. Huber, et al. // PLoS Comput Biol. — 2013. — Vol. 9(4). — P. e1003007. — doi: 10.1371/journal.pcbi.1003007.
100. *Timmer, J. C.* Structural and kinetic determinants of protease substrates / J. C. Timmer, W. Zhu, C. Pop, et al. // Nat Struct Mol Biol. — 2009. — Vol. 16(10). — P. 1101–8. — doi: 10.1038/nsmb.1668.
101. *Shirakabe, K.* Mechanistic insights into ectodomain shedding: susceptibility of CADM1 adhesion molecule is determined by alternative splicing and O-

- glycosylation / K. Shirakabe, T. Omura, Y. Shibagaki, et al. // *Sci Rep.* — 2017. — Vol. 7. — P. 46174. — doi: 10.1038/srep46174
102. *Lichtenthaler, S. F.* To cut or not to cut: New rules for proteolytic shedding of membrane proteins / S. F. Lichtenthaler, E. Meinl // *J Biol Chem.* — 2020. — Vol. 295(35). — 12353–12354. — doi: 10.1074/jbc.H120.015304.
 103. *Chou, K. C.* Prediction of human immunodeficiency virus protease cleavage sites in proteins / K. C. Chou // *Anal Biochem.* — 1996. — Vol. 233(1). — P. 1–14. — doi: 10.1006/abio.1996.0001.
 104. *Fan, Y-X.* LabCaS: labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields / Y-X. Fan, Y. Zhang, H-B Shen // *Proteins.* — 2013. — Vol. 81(4). — P. 622–34. — doi: 10.1002/prot.24217.
 105. *Song, J.* Cascleave: towards more accurate prediction of caspase substrate cleavage sites / J. Song, H. Tan, H. Shen, et al. // *Bioinformatics.* — 2010. — Vol. 26(6). — P. 752–60. — doi: 10.1093/bioinformatics/btq043.
 106. *Li, F.* ProsperousPlus: a one-stop and comprehensive platform for accurate protease-specific substrate cleavage prediction and machine-learning model construction / F. Li, C. Wang, X. Guo // *Brief Bioinform.* — 2023. — Vol. 24(6). — P. bbad372. — doi: 10.1093/bib/bbad372.
 107. *Yang, J.* DeepDigest: Prediction of Protein Proteolytic Digestion with Deep Learning / J. Yang, Z. Gao, X. Ren, et al. // *Anal Chem.* — 2021. — Vol. 93(15). — P. 6094–6103. — doi: 10.1021/acs.analchem.0c04704.
 108. *Barkan, D. T.* Prediction of protease substrates using sequence and structure features / D. T. Barkan, D. R. Hostetter, S. Mahrus, et. al. // *Bioinformatics.* — 2010. — Vol. 26(14). — P. 1714–22. — doi: 10.1093/bioinformatics/btq267.
 109. *Piippo, M.* Pripper: prediction of caspase cleavage sites from whole proteomes / M. Piippo, N. Lietzen, O. S. Nevalainen, et al. // *BMC Bioinformatics.* — 2010. — Vol. 11. — P. 320. — doi: 10.1186/1471-2105-11-320.
 110. *Perez-Silva, J. G.* The Degradome database: expanding roles of mammalian proteases in life and disease / J. G. Perez-Silva, Y. Espanol, G. Velasco, et al. //

- Nucleic Acids Res. — 2016. — Vol. 44(D1). — P. D351–5. — doi: 10.1093/nar/gkv1201.
111. *duVerle, D.* CaMPDB: a resource for calpain and modulatory proteolysis / D. duVerle, I. Takigawa, Y. Ono, et al. // Genome Inform. — 2010. — Vol. 22. — P. 202–13. — DOI:10.1142/9781848165786_0017
 112. *Luthi, A. U.* The CASBAH: a searchable database of caspase substrates / A. U. Luthi, S. J. Martin // Cell Death Differ. — 2007. — Vol. 14(4). — P. 641–50. — DOI: 10.1038/sj.cdd.4402103
 113. *Rechsteiner, M.* PEST sequences and regulation by proteolysis / M. Rechsteiner, S. W. Rogers // Trends Biochem Sci. — 1996. — Vol. 21(7). — P. 267–71. — DOI: 10.1016/S0968-0004(96)10031-1
 114. *Fawcett, T.* An Introduction to ROC Analysis / T. Fawcett // Pattern Recogn. Lett. — 2006. — Vol. 27. — P. 861–874. — doi:10.1016/j.patrec.2005.10.010
 115. *Li, F.* DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites / F. Li, J. Chen, A. Leier, et al. // Bioinformatics. — 2020. — Vol. 36(4). — P. 1057–1065. — doi: 10.1093/bioinformatics/btz721.
 116. *Rognvaldsson, T.* Why Neural Networks Should Not be Used for HIV-1 Protease Cleavage Site Prediction / T. Rognvaldsson, L. You // Bioinformatics. — 2004. — Vol. 20(11). — P. 1702–9. — DOI:10.1093/bioinformatics/bth144
 117. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data // Nucleic Acids Res. — 2019. — Vol. 47. — P. D520–D528. — DOI: 10.1093/nar/gky949
 118. *Altschul, S. F.* Basic Local Alignment Search Tool / S. F. Altschul, W. Gish, W. Miller, et al. // J. Mol. Biol. — 1990. — Vol. 215(3). — P. 403–410. — [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
 119. *Sievers, F.* Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega / F. Sievers, A. Wilm, D. Dineen, et al. // Mol. Syst. Biol. — 2011. — Vol. 7. — P. 539. — <https://doi.org/10.1038/msb.2011.75>

120. *Pettersen, E. F.* UCSF Chimera — A Visualization System for Exploratory Research and Analysis / E. F. Pettersen, T. D. Goddard, C. C. Huang, et al. // *J. Comput. Chem.* — 2004. — Vol. 25(13). — P. 1605–1612. — doi: 10.1002/jcc.20084.
121. *Belushkin, A. A.* Sequence-Derived Structural Features Driving Proteolytic Processing / A. A. Belushkin, D. V. Vinogradov, M. S. Gelfand, et al. // *Proteomics.* — 2014. — Vol. 14. — P. 42–50. — DOI: 10.1002/pmic.201300416
122. *Fontana, A.* Correlation between Sites of Limited Proteolysis and Segmental Mobility in Thermolysin / A. Fontana, G. Fassina, C. Vita, et al. // *Biochemistry.* — 1986. — Vol. 25(8). — P. 1847–1851. — DOI: 10.1021/bi00356a001
123. *Hubbard, S. J.* Assessment of Conformational Parameters as Predictors of Limited Proteolytic Sites in Native Protein Structures / S. J. Hubbard, R. J. Beynon, J. M. Thornton // *Protein Eng.* — 1998. — Vol. 11(5). — P. 349–359. — <https://doi.org/10.1093/protein/11.5.349>
124. *Hubbard, S. J.* Molecular Recognition: Conformational Analysis of Limited Proteolytic Sites and Serine Proteinase Protein Inhibitors / S. J. Hubbard, S. F. Campbell J. M. Thornton // *J. Mol. Biol.* — 1991. — Vol. 220(2). — P. 507–530. — [https://doi.org/10.1016/0022-2836\(91\)90027-4](https://doi.org/10.1016/0022-2836(91)90027-4)
125. *Hubbard, S. J.* Modeling Studies of the Change in Conformation Required for Cleavage of Limited Proteolytic Sites / S. J. Hubbard, F. Eisenmenger, J. M. Thornton // *Protein Sci.* — 1994. — Vol. 3. — P. 757–768. — doi: 10.1002/pro.5560030505.
126. *Novotný, J.* Correlation among Sites of Limited Proteolysis, Enzyme Accessibility and Segmental Mobility / J. Novotný, R. E. Bruccoleri // *FEBS Lett.* — 1987. — Vol. 211(2). — P. 185–189. — DOI: 10.1016/0014-5793(87)81433-3
127. *Kabsch, W.* Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features / W. Kabsch, C. Sander // *Biopolymers.* — 1983. — Vol. 22. — P. 2577–2637. — DOI: 10.1002/bip.360221211

128. *Pedregosa, F.* Scikit-Learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort, et al. // J. Mach. Learn. Res. — 2011. — Vol. 12. — P. 2825–2830.
129. *Holm, L.* DaliLite workbench for protein structure comparison / L. Holm, J. Park // Bioinformatics. — 2000. — Vol. 16(6). — P. 566-7. — doi: 10.1093/bioinformatics/16.6.566.
130. *Jumper, J.* Highly Accurate Protein Structure Prediction with AlphaFold / J. Jumper, R. Evans, A. Pritzel, et al. // Nature. — 2021. — Vol. 596(7873). — P. 583–589. — doi: 10.1038/s41586-021-03819-2.
131. *Varadi, M.* AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models / M. Varadi, S. Anyango, M. Deshpande, et al. // Nucleic Acids Res. — 2022. — Vol. 50. — P. D439–D444. — DOI:10.1093/nar/gkab1061
132. *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
133. *Li, F.* Procleave: Predicting Protease-specific Substrate Cleavage Sites by Combining Sequence and Structural Information / F. Li, A. Leier, Q. Liu, et al. // Genomics Proteomics Bioinformatics. — 2020. — Vol. 18(1). — P. 52–64. — doi: 10.1016/j.gpb.2019.08.002.
134. *Wasserman, W. W.* Applied Bioinformatics for the Identification of Regulatory Elements. / W. W. Wasserman, A. Sandelin // Nat. Rev. Genet. — 2004. — Vol. 5(4). — P. 276–287. — doi: 10.1038/nrg1315.
135. *UniProt Consortium, T.* UniProt: The Universal Protein Knowledgebase. // Nucleic Acids Res. — 2018. — Vol. 46. — P. 2699. — doi:10.1093/nar/gky092.
136. *Wu, F.* A new coronavirus associated with human respiratory disease in China / F. Wu, S. Zhao, B. Yu, et al. // Nature. — 2020. — Vol. 579(7798). — P. 265–269. — doi: 10.1038/s41586-020-2202-3.

137. *Binder, J. X.* COMPARTMENTS: Unification and Visualization of Protein Subcellular Localization Evidence. / J. X. Binder, S. Pletscher-Frankild, K. Tsafou, et al. // Database. — 2014. — Vol. 2014. — P. 1–9. — doi:10.1093/database/bau012.
138. *Palasca, O.* TISSUES 2.0: An Integrative Web Resource on Mammalian Tissue Expression. / O. Palasca, A. Santos, C. Stolte, et al. // Database. — 2018. — Vol. 2018. — P. 1–12. — doi:10.1093/database/bay003.
139. *Walls, A. C.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein / A. C. Walls, Y-J. Park, M. A. Tortorici // Cell. — 2020. — Vol. 181(2). — P. 281–292. — <https://doi.org/10.1016/j.cell.2020.02.058>
140. *Xu, C.* Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM / C. Xu, Y. Wang, C. Zhang, et al. // Sci Adv. — 2021. — Vol. 7(1). — P. eabe5575. — doi: 10.1126/sciadv.abe5575.
141. *Waterhouse, A.* SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. / A. Waterhouse, M. Bertoni, S. Bienert, et al. // Nucleic Acids Res. — 2018. — Vol. 46. — P. W296–W303. — doi:10.1093/nar/gky427.
142. *Hodcroft, E. B.* CoVariants: SARS-CoV-2 Mutations and Variants of Interest.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ДНК — дезоксирибонуклеиновая кислота

ВИЧ (HIV, human immunodeficiency virus) — вирус иммунодефицита человека

SARS-CoV-2 (severe acute respiratory syndrome-related coronavirus 2) —
коронавирус, вызывающий острую респираторную инфекцию COVID-19

PWM (position-weight matrix) — позиционно-весовая матрица

PCSK (proprotein convertase subtilisin/kexin) — семейство пропротеин-конвертаз
субтилизин/кексинового типа

TTSP (type 2 transmembrane serine protease) — семейство трансмембранных
сериновых протеаз второго типа

S1/S2 — сайт расщепления по позиции R685 в полипептидной цепи
гликопротеина оболочки коронавируса SARS-CoV-2

S2` — сайт расщепления по позиции R815 в полипептидной цепи гликопротеина
оболочки коронавируса SARS-CoV-2

E.coli — организм кишечной палочки (*Escherichia coli*)

CutDB / MEROPS — базы данных протеолитических событий

PDB — база данных экспериментально подтверждённых структур белков

AlphaFold — инструмент предсказания трёхмерных структур белков по их
аминокислотной последовательности

AlphaFoldDB — база данных структур белков, смоделированных при помощи
инструмента AlphaFold

HIV-1 протеаза — протеаза вируса иммунодефицита человека 1

Nup98 — нуклеопорин

IDE (insulin-degrading enzyme) — металлопротеаза, участвующая в расщеплении
инсулина

pH — уровень кислотности среды

MPP (mitochondrial processing peptidase) — митохондриальная протеаза

Sclerotinia sclerotiorum — фитопатогенный гриб

PC4 / PC5 / PC7 — протеазы семейства пропротеин-конвертаз

PACE4 (paired basic amino acid-cleaving enzyme 4) — протеаза семейства пропротеин-конвертаз

gB — гликопротеин оболочки вируса герпеса

HA — гемагглютинин

NA — нейраминидаза

H0 — исходная форма гемагглютиниона

H1 — субъединица 1 гемагглютиниона

H2 — субъединица 2 гемагглютиниона

TMPRSS2 / TMPRSS4 — протеазы семейства трансмембранных сериновых протеаз

HAT (human airway trypsin-like protease) — трипсиноподобная протеаза дыхательных путей человека

MSPL / MTPRSS13 — протеазы семейства трансмембранных сериновых протеаз

gp160 — молекула-предшественник гликопротеина оболочки вируса HIV-1

gp120 / gp41 — субъединицы гликопротеина оболочки вируса HIV-1

SKI-1/S1P — протеаза семейства пропротеин-конвертаз

ВПЧ — вирус папилломы человека

L1 / L2 — белки капсида ВПЧ

prM — структурный белок оболочки флавивирусов

Alphavirus — род РНК-вирусов

Chikungunya / Semliki forest / Sindbis / Ross River — вирусы из рода Альфа вирусов

Spike-белок / S-белок — белок-шип, гликопротеин оболочки коронавируса

S1 / S2 — субъединицы гликопротеина оболочки коронавируса

MERS-CoV (Middle East respiratory syndrome) — коронавирус, вызывающий ближневосточный респираторный синдром

SARS-CoV (severe acute respiratory syndrome) — коронавирус, вызывающий острый респираторный синдром

HCoV — коронавирусы человека, к ним относятся HCoV-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV и SARS-CoV-2

ACE2 (angiotensin-converting enzyme 2) — ангиотензин-конвертирующий фермент 2

РНК — рибонуклеиновая кислота

NTD (N-terminal domain) — N-концевой домен гликопротеина оболочки коронавируса SARS-CoV-2

RBD (receptor-binding domain) — рецептор-связывающий домен гликопротеина оболочки коронавируса SARS-CoV-2

CTD1 / CTD2 (C-terminal domain) — C-концевые домены гликопротеина оболочки коронавируса SARS-CoV-2

HR1 / HR2 (heptad repeat) — гептадный повтор, структурный мотив гликопротеина оболочки коронавируса SARS-CoV-2

Degradome / CamPDB / CASBAH — базы данных протеолитических событий

PEST — аминокислотные последовательности, обогащённые пролином, серином, треонином и глутаминовой кислотой

ROC (receiver operating characteristic) — кривая, позволяющая оценить качество бинарной классификации

AUC (area under curve) — площадь под кривой

Leave-One-Out — метод кросс-валидации, при котором каждый пример из исходного набора данных используется в качестве тестового, а все остальные примеры используются для обучения модели.

СПИСОК ИЛЛЮСТРАЦИЙ

Рисунок 1. Представление сайта расщепления при взаимодействии активного центра протеазы с субстратом, согласно [21].	36
Рисунок 2. Визуализация признаков, использованных для обучения предсказательной модели, на трёхмерных структурах субстратов	50
Рисунок 3. Зависимость качества предсказания модели от соотношения классов в обучающей выборке	52
Рисунок 4. Оценка качества предсказания модели при соотношении наблюдений положительного и отрицательного классов 1:1 в обучающей выборке	53
Рисунок 5. Визуализация предсказаний восприимчивости к протеолизу пептидных связей, полученных с помощью разработанной модели, на трёхмерной структуре субстрата	54
Рисунок 6. Распределение оценок восприимчивости к протеолизу, сгенерированных разработанной моделью, для различных категорий протеолитических расщеплений	56
Рисунок 7. Сравнение предсказательной способности характеристики, описывающей достоверность структурного предсказания AlphaFold-модели, с предсказательной способностью других структурных признаков	59
Рисунок 8. Оценка качества прогнозирования модифицированной модели в зависимости от соотношения классов в обучающей выборке	60
Рисунок 9. Оценка качества прогнозирования модифицированной модели при соотношении положительных и отрицательных классов 1:1 в обучающей выборке	61
Рисунок 10. Улучшение качества прогнозирования модифицированной модели при расширении обучающего набора данных за счёт включения моделей трёхмерных структур из базы данных AlphaFoldDB	62
Рисунок 11. Схематическое представление объединения результатов предсказаний моделью универсальной специфичности и протеазо-специфическими моделями, описывающими субстратную специфичность конкретных протеаз	64
Рисунок 12. Двумерные визуализации предсказаний интегративной модели для наблюдений различных классов при использовании различных алгоритмов машинного обучения	65
Рисунок 13. Сравнение качества прогнозирования между разработанным интегративным методом, объединяющим предсказания структурной модели и модели специфичности по последовательности, и методом Procleave	67

Рисунок 14. Распределение оценок расщепления, рассчитанных с помощью моделей субстратной специфичности протеаз, для известных сайтов расщеплений и случайных позиций в белках.	75
Рисунок 15. Двадцать наивысших оценок расщепления в позиции R685 (сайт S1/S2), полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	75
Рисунок 16. Тридцать наивысших оценок расщепления в позиции R815 (сайт S2`), полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	76
Рисунок 17. Десять позиций последовательности S-гликопротеина, получивших наивысшие оценки расщепления по модели субстратной специфичности катепсина L	78
Рисунок 18. Отображение оценок восприимчивости к протеолизу, сгенерированных универсальной структурной моделью, на трёхмерной структуре S-гликопротеина.....	79
Рисунок 19. Двадцать наивысших оценок расщепления в позиции K790, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	80
Рисунок 20. Двадцать позиций S-гликопротеина с наивысшими оценками расщепления, полученных с помощью моделей субстратной специфичности протеаз семейств PCSK, TTSP, калликреинов и факторов свёртывания крови	81
Рисунок 21. (а) Потенциальный сайт расщепления в позиции R78 расположен в N-концевом домене S-гликопротеина. (б) Потенциальные сайты расщепления в позициях R346 и R466 расположены в пределах рецептор-связывающего домена S-гликопротеина.....	82
Рисунок 22. (а) Потенциальный сайт расщепления в позиции R634 расположен в, так называемой, «петле 630» S-гликопротеина. (б) Потенциальный сайт расщепления в позиции R847 примыкает к пептиду слияния.....	82
Рисунок 23. Двадцать наивысших оценок расщепления в позиции R78, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	83
Рисунок 24. Двадцать наивысших оценок расщепления в позиции R346, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	84
Рисунок 25. Двадцать наивысших оценок расщепления в позиции R466, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	84

Рисунок 26. Двадцать наивысших оценок расщепления в позиции R634, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	85
Рисунок 27. Двадцать наивысших оценок расщепления в позиции R847, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	86
Рисунок 28. Оценки расщепления, рассчитанные для известных и потенциальных сайтов в S-гликопротеине, с помощью моделей субстратной специфичности протеаз семейств PCSK, TTSP, калликреинов и факторов свёртывания крови.	86
Рисунок 29. 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.	88
Рисунок 30. Оценки расщепления сайта S1/S2, рассчитанные моделью субстратной специфичности фуриновой протеазы для всех возможных замен в позиции 681.....	89
Рисунок 31. Двадцать наивысших оценок расщепления в позиции 346 при замене R346K, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	89
Рисунок 32. Двадцать наивысших оценок расщепления в позиции 346 при замене R346T, полученных с помощью моделей субстратной специфичности 169 протеаз человека.....	90