

Отзыв

**официального оппонента на диссертацию Матвеева Евгения Викторовича
«Предсказание сайтов протеолиза на основе информации о пространственной
структуре потенциальных субстратов», представленную на соискание учёной
степени кандидата биологических наук по специальности 1.5.8 –«Математическая
биология, биоинформатика»**

Актуальность темы диссертации

Протеолиз лежит в основе механизма регуляции большого числа клеточных процессов. Так, посредством протеолиза происходит регуляция активации ферментов, гормонов, цитокинов, рецепторов, запуск апоптоза, клеточная дифференцировка, иммунный ответ, а также передача сигналов. Нарушение регуляции протеолиза связано со многими заболеваниями, включая онкологические нейродегенеративные, сердечно-сосудистые, инфекционные и аутоиммунные. Таким образом, разработка моделей предсказания сайтов расщепления важна для понимания молекулярных механизмов протеолитических процессов и разработки таргетной терапии.

Большинство существующих биоинформатических инструментов для предсказания сайтов протеолиза являются специфичными для конкретного семейства протеаз. Перечисленные модели предсказывают сайты расщепления для конкретной протеазы или узкого семейства протеаз, что существенно сужает область их применения. Таким образом, на момент начала данного исследования отсутствовал универсальный инструмент, который мог бы оценивать общие структурные характеристики белка и определять сайты, более подверженные протеолизу без учета специфичности конкретной протеазы.

Структура и содержание

Диссертация Е.В. Матвеева изложена на 121 странице и содержит введение, три главы, заключение и выводы. Работа содержит 32 рисунка и 2 формулы. Список литературы содержит 142 наименования.

Название работы в полной мере отражает ее содержание, поскольку в работе представлена модель по прогнозированию сайтов протеолитического расщепления белков на основе их трёхмерной структуры, а также представлена апробация модели для реального S-белка SARS-CoV-2.

Во Введении обосновывается актуальность исследования протеолиза – ферментативного расщепления белков, которое играет ключевую роль в регуляции биологических процессов и связано с многими заболеваниями. Подчёркивается, что экспериментальные методы идентификации субстратов и сайтов расщепления трудоёмки и дороги, что делает биоинформатические подходы важной альтернативой. Автор отмечает, что существующие методы прогнозирования в основном ориентированы на конкретные протеазы и используют данные о последовательностях, но не учитывают универсальные структурные особенности субстратов. Цель работы – разработать интегративный метод, сочетающий структурную модель (оценку уязвимости участков белка на основе 3D-структуры) и модели специфичности протеаз по аминокислотной последовательности.

В первой главе приведен обзор литературы. В ней автор суммирует современные научные знания по теме диссертации. Дается определение протеолиза, объясняется его ключевая роль в регуляции биологических процессов. Обозначено разнообразие протеаз, описываются классы протеаз, такие как сериновые, цистеиновые, металлопротеазы и др., обсуждаются механизмы действия, специфичность и распространение по дереву жизни. Рассказывается о роли протеолиза в вирусологии, на примере гриппа, герпеса, ВИЧ, эболы подробно разбирается, как протеолитическая активация вирусных белков критически важна для проникновения вируса в клетку и его патогенности. Особое внимание уделяется жизненному циклу коронавируса, структуре его S-белка, роли протеаз (фурина, TMPRSS2, катепсина L) в активации этого белка и проникновении вируса в клетку. Анализируются существующие биоинформатические подходы к предсказанию сайтов протеолиза, их сильные и слабые стороны, базы данных (MEROPS, CutDB) и подчёркивается отсутствие универсальных структурно-ориентированных методов.

Во второй главе автор подробно описывает разработку собственного метода предсказания протеолитических сайтов. Для этого используются данные об известных сайтах расщепления из базы CutDB, а потом находятся и отбираются 3D-структуры белков-субстратов из PDB. Для каждого участка белка рассчитываются структурные признаки (доступность растворителю, вторичная структура, подвижность и др.). Далее тестируются разные алгоритмы машинного обучения, и лучший результат показал алгоритм на основе линейного дискриминантного анализа. Дальнейшее улучшение модели достигается с помощью расширения обучающей выборки за счёт высококачественных предсказанных структур из базы AlphaFoldDB. Далее предложена

интеграционная модель, где структурная модель объединена с позиционно-весовыми матрицами (PWM), которые описывают предпочтения конкретных протеаз к аминокислотным последовательностям. Представлено сравнение с существующим методом Procleave, и показано, что разработанный метод показал точность, сопоставимую с лучшим на сегодняшний день существующим инструментом (Procleave), но при этом является более универсальным и охватывает больше протеаз.

В третьей главе разработанный метод прогнозирования протеолитических сайтов применяется для анализа S-белка коронавируса SARS-CoV-2. Были идентифицированы протеазы человека, способные расщеплять S-белок в известных сайтах S1/S2 и S2'. Помимо известных протеаз фурина и TMPRSS2, высокий потенциал показали протеазы семейств PCSK, TTSP, калликреины и факторы свёртывания крови. Был предсказан новый сайта расщепления для катепсина L — позиция K790, расположенный рядом с пептидом слияния и сайтом S2', что может объяснять активацию вируса при эндоцитозе. С помощью разработанной автором модели был произведен поиск альтернативных сайтов расщепления в S-белке. Выявлены пять перспективных позиций (R78, R346, R466, R634, R847), которые могут быть мишенями для протеаз. Так же был произведен анализ влияния мутаций в S-белке (например, P681H/R в штаммах Альфа, Омикрон, Дельта) на эффективность расщепления. Показано, что эти мутации повышают восприимчивость к расщеплению фурином. Таким образом, глава подтверждает практическую ценность разработанного метода для изучения механизмов инфекции и эволюции вируса.

Основные выводы по результатам исследования изложены в разделе Выводы. Работа дополнена списком таблиц, рисунков и условных обозначений.

Научная новизна

Научная новизна работы обусловлена созданием первой универсальной структурной модели предсказания сайтов протеолиза. Разработана первая в своём роде прогностическая модель, которая предсказывает степень предрасположенности участков белка к протеолизу на основе исключительно структурных характеристик (доступность растворителю, вторичная структура, температурный фактор, длина петли). Модель является универсальной, так как не привязана к специфичности какой-либо конкретной протеазы и применима ко всем их классам. В отличие от большинства существующих методов, которые опираются на аминокислотную последовательность, данная модель использует 3D-структуру белка, что позволяет учитывать пространственную

доступность сайта. Таким образом, создан интегративный подход, который учитывает и структуру белка, и специфичность протеазы.

Практическая и теоретическая значимость

Теоретическая значимость исследования заключается в создании принципиально новой универсальной структурной модели предсказания протеолиза, не зависящей от специфичности протеазы. Работа вносит вклад в фундаментальное понимание механизмов инфекции SARS-CoV-2, предсказывая новый сайт расщепления катепсина L (K790) и предлагая модель для объяснения влияния мутаций в S-белке на повышенную инфекционность штаммов. Практическая ценность состоит в создании общедоступного инструментария (169 моделей специфичности протеаз и интегративного подхода) для исследователей. Метод позволяет ускорить поиск потенциальных мишеней для терапии, оценивать патогенность новых вариантов вируса и сокращает время на планирование экспериментов в вирусологии и протеомике.

Обоснованность научных положений и достоверность результатов

Обоснованность научных положений и достоверность результатов в диссертации обеспечивается использованием проверенных данных и методов. Работа построена на данных из авторитетных международных баз: CutDB (протеолитические события), MEROPS (специфичность протеаз), PDB (экспериментальные структуры белков), AlphaFoldDB (предсказанные структуры). Все разработанные модели прошли тщательную валидацию, а модели специфичности протеаз валидировались на независимых выборках из MEROPS. Достоверность подтверждается тем, что модель корректно предсказала известные сайты расщепления (S1/S2 и S2') в S-белке SARS-CoV-2, присвоив наивысшие оценки именно тем протеазам (фурину и TMPRSS2), чья роль экспериментально доказана. Основные положения и выводы работы опубликованы в рецензируемых научных журналах, индексируемых в Scopus и Web of Science, а также представлены и опубликованы в виде тезисов на трех конференциях.

Текст диссертации и представленные в нем данные полностью обосновывают сделанные в работе выводы и выносимые на защиту положения. Автореферат диссертации полностью отражает ее содержание.

Представленная работа вносит важный вклад в развитие методов определения протеолитических сайтов расщепления белков за счёт создания первой универсальной структурной модели, не ограниченной специфичностью конкретных протеаз, и

демонстрирует её практическую значимость на примере анализа ключевого вирусного белка SARS-CoV-2.

В то же время к работе можно выдвинуть ряд замечаний, пожеланий и дискуссионных вопросов:

1). При сравнении качества моделей машинного обучения, используются метрики, полученные моделями, который были использованы с дефолтными гиперпараметрами. Вероятно, LDA не остался бы лучшим алгоритмом при тонкой настройке гиперпараметров. Известно, что LDA работает лучше всего на нормально распределенных данных.

2). Пробовал ли автор строить модели на большем числе параметров, чем на пяти отобранных? Преимущество алгоритмов машинного обучения как раз и состоит в том, что они могут интегрировать информацию от не сильно коррелирующих признаков? Могло ли бы это улучшить качество предсказаний модели?

3). Для сравнения с Procleave был использован очень небольшой тестовый набор (43 события, 3 протеазы). Достаточно ли этого для статистически надежного вывода о сопоставимости качества? Планировали ли вы валидацию на более масштабных наборах данных?

4). Как именно технически вы решали проблему сильного дисбаланса классов (445 позитивных против 68 840 негативных примеров) при обучении моделей, кроме подбора соотношения 1:1? Использовались ли методы аугментации данных?

5). В диссертации обсуждается, что посттрансляционные модификации (например, гликозилирование) сильно влияют на доступность сайта. Какие Вы видите способы интеграции такой информации в будущем?

6). Вопрос дискуссионный: возможно ли построить универсальную модель определения сайтов расщепления с помощью представлений больших языковых белковых моделей (например, ESM-2, ProtBERT)? Были ли опубликованы подобные работы?

7). Предсказание сайта для катепсина L в позиции K790 является ключевым результатом. Планируется ли его экспериментальная проверка (например, с помощью масс-спектрометрии) для подтверждения?

Несмотря на имеющиеся замечания и вопросы, считаю, что диссертационная работа Матвеева Евгения Викторовича «Предсказание сайтов протеолиза на основе информации о пространственной структуре потенциальных субстратов» является

законченной научно-квалификационной работой и отвечает требованиям п. 9 «Положения о порядке присуждения ученых степеней», утвержденным постановлением Правительства Российской Федерации No 842 от 24 сентября 2013 г., а ее автор Е.В. Матвеев заслуживает присуждения искомой степени кандидата биологических наук по специальности 1.5.8 – математическая биология, биоинформатика.

Официальный оппонент:

кандидат физико-математических наук,
заведующий международной лабораторией биоинформатики Института
искусственного интеллекта и цифровых наук Факультета компьютерных наук
Федерального государственного автономного образовательного учреждения высшего
образования «Национальный исследовательский университет
«Высшая школа экономики»

ПОПЦОВА Мария Сергеевна



подпись

Дата подписания

Контактные данные:

тел.: 7(962) 909 5153, e-mail: mpoptsova@hse.ru

Специальность, по которой официальным оппонентом
защищена диссертация:

03.01.02 – Биофизика

Адрес места работы:

109028, г. Москва, Покровский бульвар, д. 11,

Национальный исследовательский университет «Высшая школа
экономики», Факультет компьютерных наук, Институт искусственного интеллекта и
цифровых наук, Международная лаборатория биоинформатики

Тел.: +7 (495) 531-00-00 27335; e-mail: mpoptsova@hse.ru

