

БИОИНФОРМАТИЧЕСКИЙ МЕТОД ИДЕНТИФИКАЦИИ ПРОТЕАЗ ЧЕЛОВЕКА, АКТИВНЫХ ОТНОСИТЕЛЬНО ГЛИКОПРОТЕИНОВ ОБОЛОЧКИ ВИРУСОВ, НА ПРИМЕРЕ БЕЛКА ШИПА КОРОНАВИРУСА SARS-CoV-2

© 2024 г. Е. В. Матвеев^{a, b, c}, Г. В. Пономарев^{a, b}, М. Д. Казанов^{a, b, c, *}

^a Сколковский институт науки и технологий, Москва, 121205 Россия

^b Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, 127051 Россия

^c Национальный медицинский исследовательский центр детской гематологии, онкологии и иммунологии им. Дмитрия Рогачева, Москва, 117997 Россия

*e-mail: mkazanov@gmail.com

Поступила в редакцию 02.10.2023 г.

После доработки 02.10.2023 г.

Принята к публикации 02.10.2023 г.

Рекомендована к публикации М.С. Гельфандом

Множество вирусов, включая коронавирус-2 тяжелого острого респираторного синдрома (SARS-CoV-2), который стал причиной пандемии коронавирусной инфекции (COVID-19), проникает в клетку за счет активируемого протеолитическими ферментами процесса слияния клеточной и вирусной мембран. Как правило, роль протеолитических ферментов в этом случае выполняют протеазы хозяйской клетки. Идентификация активирующих протеаз — задача непростая, но важная для разработки новых противовирусных лекарственных средств. В рамках проведенного исследования мы разработали биоинформатический метод идентификации протеаз, воздействующих на белки оболочки вирусов. Предлагаемый подход включает в себя использование предсказательных моделей субстратной специфичности протеаз человека и применение метода предсказания уязвимости участков белка к протеолизу на основе его трехмерной структуры. Модели специфичности были построены для 169 протеаз человека на основании информации по их известным субстратам. Метод структурного анализа потенциальных сайтов протеолиза был разработан нами ранее и применен в представленной здесь работе совместно с моделями специфичности протеаз. Валидация предлагаемого подхода выполнялась применительно к белку шипа SARS-CoV-2, для которого хорошо изучены сайты протеолиза.

Ключевые слова: протеазы человека, вирусный гликопротеин, протеолитическая активация, белок шипа, SARS-CoV-2

DOI: 10.31857/S0026898424010176, EDN: NRZZBT

ВВЕДЕНИЕ

Использование вирусами ресурсов клетки для реализации жизненного цикла, в том числе для проникновения в клетку, репликации генома и сборки вирусных частиц, хорошо известный факт [1]. В геноме вирусов обычно отсутству-

ют гены, кодирующие используемые вирусом функции клеточного аппарата хозяина. По этой причине становится достаточно сложно выявить те факторы в клетке-хозяине, которые критически важны для успешной реализации жизненного цикла вируса. Идентификация партнеров вируса в хозяйской клетке имеет решающее значение для создания новых противовирусных препаратов и разработки различных стратегий лечения вирусных инфекций [2]. Эксперименты, направленные на выяснение ответа на данный вопрос, трудоемкие и ресурсозатратные, в то время как биоинформатическое предсказание таких кле-

Сокращения: ACE2 (Angiotensin-Converting Enzyme 2) — ангиотензинконвертирующий фермент-2; CS (Cleavage Score) — оценка протеолитической эффективности; PWM (Position Weight Matrix) — позиционно-весовая матрица; TMPRSS2 (Transmembrane Serine Protease 2) — трансмембранная сериновая протеаза-2.

точных факторов существенно упрощает проведение экспериментальных исследований.

Борьба с пандемией COVID-19 привела к массовому пониманию важности вирусологических исследований и необходимости всестороннего понимания жизненного цикла вызвавшего ее вируса, включая выявление всех компонентов, вовлеченных в процесс инфицирования клетки-хозяина [3]. Коронавирус-2 тяжелого острого респираторного синдрома (SARS-CoV-2) относится к семейству *Coronaviridae*, а его геном представлен одноцепочечной (+) РНК. Уже идентифицирован ряд факторов клетки-хозяина, взаимодействующих с компонентами SARS-CoV-2 [4], и среди них важная роль отводится клеточным протеазам [5]. Ключевая роль протеаз заключается в активации процессов конформационного перехода поверхностного гликопротеина вируса — белка шипа (S) — в компетентную для слияния форму и слияния мембраны вирусной частицы с клеточной [6, 7].

Ранее в S-белке SARS-CoV-2 идентифицировано два критически важных сайта фузионной активности вируса под действием протеаз клетки-хозяина [8]. Первый сайт расщепления, известный как S2', расположен в домене S2 и исходно недоступен для протеаз. Только после связывания S-белка с клеточным рецептором — ангиотензинконвертирующим ферментом-2 (ACE2) — на инфицируемой клетке этот сайт "открывается" для протеаз [9]. Известно, что позиция S2' подвергается расщеплению трансмембранной сериновой протеазой-2 (TMPRSS2), тем самым инициируя процесс слияния мембраны вириона с мембраной клетки [10]. Второй протеолитический сайт, известный как S1/S2, расположен в области соединения субъединицы S1, содержащей рецепторсвязывающий домен (RBD), и субъединицы S2, содержащей домен слияния. Согласно современным представлениям, этот сайт расщепляется фурином в комплексе Гольджи перед упаковкой вируса [6, 11]. Это протеолитическое расщепление вызывает конформационные изменения S-белка, переводя его из так называемой "закрытой" конформации в "открытую", что облегчает связывание S-белка с ACE2 [12]. Наличие в S-белке SARS-CoV-2 короткой аминокислотной вставки в области S1/S2, отсутствующей у близкородственных бетакоронавирусов, вызвало дискуссии о природе ее происхождения [13–16].

В ряде исследований показано, что даже в отсутствие TMPRSS2 SARS-CoV-2 может проникать в клетку через альтернативный эндоцитарный путь, называемый также "поздним" [17]. Так, Z. Liu и др. [18] обнаружили, что катепсин L активирует слияние мембраны вирусной частицы с эндолизосомальной посредством рас-

щепления S-белка вблизи S1/S2, в то время как J. Park и др. [19] предполагают, что расщепление происходит рядом с сайтом S2'. Данный путь проникновения вируса в клетку все еще остается малоизученным по сравнению с классическим, "ранним", ACE2-опосредованным. Альтернативный эндоцитарный путь инфицирования остается предметом активных исследований [20], что особенно актуально в свете сообщений о повышенной зависимости штамма SARS-CoV-2 Omicron BA.1 именно от этого пути [21].

Целью проведенного нами исследования была разработка биоинформатического метода идентификации протеаз человека, потенциально вовлеченных в процессинг рассматриваемого вирусного белка. Для этого мы разработали модели специфичности протеаз человека относительно аминокислотной последовательности, а также использовали разработанный нами ранее метод предсказания подверженности участков белка протеолиту на основе информации о его трехмерной структуре. Валидация предложенного подхода проведена на хорошо изученном S-гликопротеине SARS-CoV-2, для которого известны как вовлеченные в его процессинг протеазы клеток человека, так и сайты расщепления.

МЕТОДЫ

Модели субстратной специфичности протеаз человека. Информация о субстратах и протеолитических сайтах протеаз человека была извлечена из базы данных MEROPS [22]. Модели специфичности протеаз по аминокислотной последовательности были созданы в виде позиционно-весовых матриц (PWM), как описано в Wasserman & Sandelin [23]. Далее изложено краткое описание алгоритма построения модели. Для каждого протеолитического фермента человека из базы данных MEROPS, имеющего не менее 8 протеолитических сайтов, были рассчитаны частоты встречаемости аминокислот относительно сайта разрезания в позициях P3–P3' (обозначение согласно Schechter & Berger [24]) — f_{ij} , где i — аминокислота, а j — относительная позиция в аминокислотной последовательности. Фоновые частоты встречаемости аминокислот b_{ij} были рассчитаны для всего протеома человека. Каждой ячейке в PWM, столбцы которой представляют собой позиции P3–P3', а строки — 20 стандартных аминокислот, было присвоено значение r_{ij} , рассчитанное по формуле: $r_{ij} = \log_2 (f_{ij}/b_{ij})$. При допущении, что каждая позиция в PWM является независимой, оценка протеолитической эффективности (CS) для конкретной пептидной связи белка была рассчитана как сумма соответствующих значений из каждого столбца P3–P3' для соответствующих аминокислот.

кислот, наблюдаемых в данных позициях. Для обозначения координат расщепляемых пептидных связей использовалась позиция P1.

Модель предсказания подверженности областей белка протеолитическому расщеплению. Эта модель основана на анализе трехмерной структуры субстрата и описана нами ранее [25]. Далее представлено краткое описание метода. Информация о протеолитических событиях была получена из базы данных CutDB [26]. Сайты расщепления протеазами были картированы на трехмерные структуры соответствующих белков-субстратов. Набор предикторов был составлен из следующих структурных характеристик: доступность для растворителя, температурный фактор, тип вторичной структуры, а также дополнительных особенностей структуры белка, таких как длина петель и принадлежность к концевым участкам. Обучающий набор был прокурирован вручную с целью исключения данных, полученных из экспериментов с потенциально денатурированными субстратами. Такие эксперименты мы выявляли на основе наличия большого числа сайтов расщепления в гидрофобном ядре субстратов. Итоговая модель была обучена с использованием алгоритма линейного дискриминантного анализа, который продемонстрировал лучшие результаты по сравнению с другими алгоритмами машинного обучения из библиотеки scikit-learn [27].

Данные по аминокислотной последовательности и трехмерной структуре S-белка SARS-CoV-2. Аминокислотная последовательность S-белка SARS-CoV-2 исходного штамма Wuhan была получена из базы данных UniProt (ID: P0DTC2) [28]. Трехмерные структуры белка-шипа в “открытой” конформации (PDB ID: 6VYB), “закрытой” конформации (PDB ID: 6VXX) и в комплексе с рецептором ACE2 (PDB ID: 7DF4) были скачаны из базы данных PDB [29]. Неструктурированные участки трехмерных структур были смоделированы с помощью инструмента SWISS-MODEL с параметрами, заданными по умолчанию [30].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Модели субстратной специфичности протеаз человека

Предсказательные модели специфичности протеаз по аминокислотной последовательности были построены для 169 протеаз человека (см. табл. S1 Дополнительных материалов). Размер данного набора протеолитических ферментов человека был определен на основе анализа информации по идентифицированным субстратам протеаз в базе данных MEROPS. Для построения предсказательных моделей специфичности фер-

ментов по аминокислотной последовательности в виде PWM мы отобрали протеазы с широким спектром субстратов (см. раздел “Методы”). Для корректной интерпретации результатов по протеолитической эффективности, полученных построенными моделями, мы сравнили распределения оценок для известных сайтов протеолитического расщепления из базы данных MEROPS и для случайных позиций, которые с большой вероятностью не относятся к сайтам протеаз человека (рис. 1а). Результаты анализа распределений показали, что Q1–Q3 значения CS для протеолитических сайтов лежат в интервале [1.64, 5.02] со средним и медианным значениями 3.36 и 3.16 соответственно, тогда как для случайных позиций границы интервала Q1–Q3 для CS равны [–5.89, –1.03] со средним и медианным значениями –3.57 и –3.42 соответственно.

Валидация моделей субстратной специфичности протеаз на примере известных сайтов расщепления S-белка SARS-CoV-2

Полученные модели специфичности протеаз человека апробированы на аминокислотной последовательности S-гликопротеина SARS-CoV-2 исходного штамма Wuhan. Вычисленные значения CS для сайтов S2' и S1/S2 отображены соответственно на рис. 1а и 1б (полные версии доступны на рис. S1 и S2 Дополнительных материалов, см. на сайте http://www.molecbio.ru/downloads/2024/1/supp_Matveev_rus.zip). Для позиции S2' наибольшее значение CS относится к протеазе TMPRSS2, что полностью согласуется с текущим представлением об активации процесса слияния мембран коронавируса и инфицируемой клетки. Аналогично для позиции S1/S2 максимальному показателю CS соответствовал фурин, что также согласуется с текущими представлениями по инфекционному циклу SARS-CoV-2 [31].

Предсказание подверженности протеолитическому расщеплению областей белка на основе 3D-структуры для известных сайтов расщепления S-белка SARS-CoV-2

Для предсказания протеолитической активности протеаз человека по отношению к вирусным белкам, помимо использования моделей специфичности протеаз, имеет смысл дополнительно оценивать структурную уязвимость участков исследуемого белка к протеолизу. Соответствующий метод был разработан нами ранее [25] и теперь применен к известным сайтам расщепления S-белка SARS-CoV-2. Используя этот алгоритм, мы дали оценку подверженности протеолизу участков трехмерных структур S-белка в “открытой” и “закрытой” конфор-

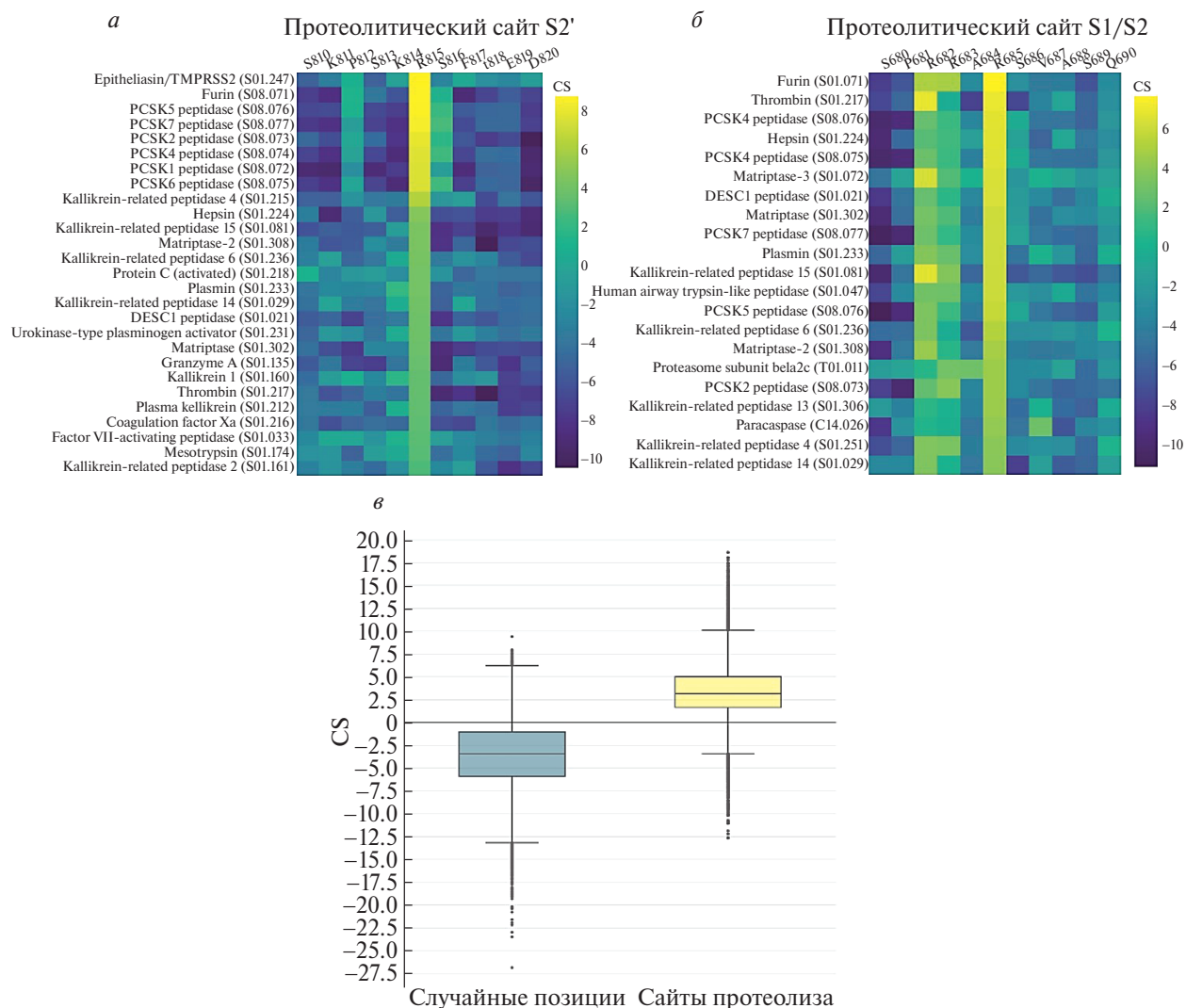


Рис. 1. Оценка протеолитической эффективности протеаз человека относительно сайтов S-гликопротеина SARS-CoV-2. Протеолитические ферменты человека с наивысшими значениями CS в позиции R815 сайта S2' (а) и в позиции R685 сайта S1/S2 (б). Распределение значений CS для известных протеолитических сайтов и случайно выбранных позиций (в).

мациях, а также в комплексе с ACE2 (рис. 2а,б; табл. S1 Дополнительных материалов, см. на сайте http://www.molecbio.ru/downloads/2024/1/supp_Matveev_rus.zip). Предварительно для каждой субъединицы тримера S-гликопротеина нами были смоделированы все неструктурированные участки белка, в том числе отсутствующая на структурах позиция R685. После получения оценок для каждой позиции каждой субъединицы полученные значения усредняли по всем трем субъединицам тримера. Для удобства анализа позиции в белке были отсортированы по убыванию усредненных оценок. В результате применения метода к 3D-структуре S-гликопротеина в “закрытой” конформации (PDB ID: 6VXX) для позиции S1/S2 была получена оценка 0.77, что указывает на ее высокую доступность для протеолиза: занимает пятую строчку в таблице значений для 1121 исследованной позиции. Для по-

зиции сайта S2' была получена оценка 0.58, что соответствует средней доступности сайта. Аналогично применение метода к 3D-структуре S-белка в “открытой” конформации (PDB ID: 6VYB) дало оценку 0.72 для сайта S1/S2, что также соответствует хорошей доступности для протеолиза — 15 строчка в таблице с 1121 позицией. Для позиции сайта S2' была получена оценка 0.58. Наконец, при применении метода к комплексу S-белка с рецептором ACE2 (PDB ID: 7DF4) для позиции сайта S1/S2 была получена оценка 0.73, а для позиции сайта S2' — 0.67, что уже соответствует хорошей доступности для протеолиза. Полученные результаты соответствуют текущим представлениям о том, что сайт S2' становится доступным для протеазы TMPRSS2 после связывания S-гликопротеина с клеточным рецептором ACE2, а сайт S1/S2 доступен при любой конформации S-белка [5].

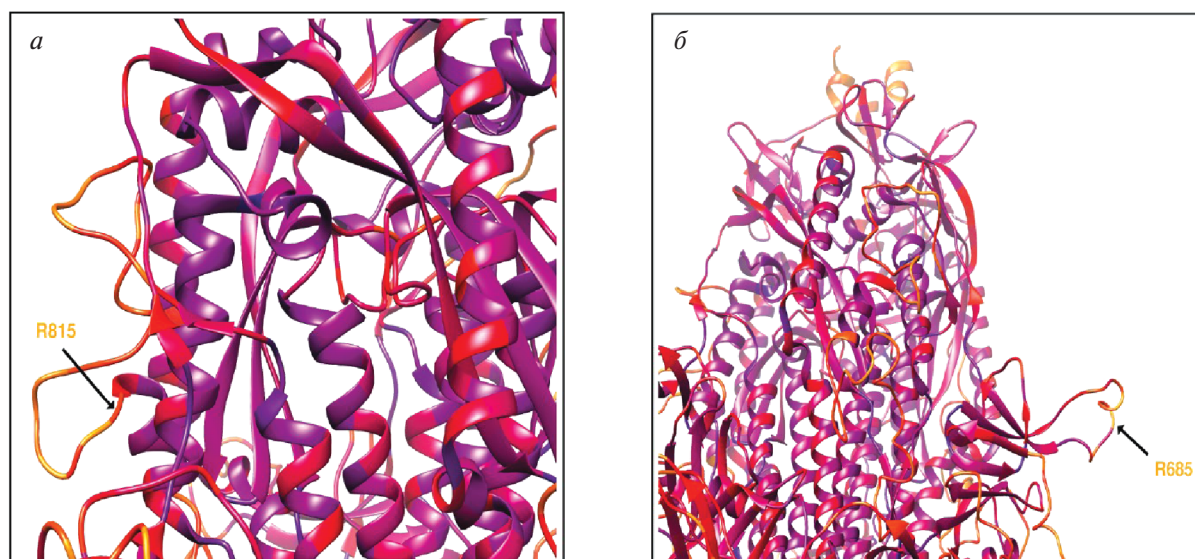


Рис. 2. Оценка протеолитической доступности сайтов S-гликопротеина SARS-CoV-2. *а* – Отображение на структуре S-гликопротеина (PDB ID: 6VXX) оценок протеолитической доступности в области сайта S2' (позиция R815). *б* – Отображение на структуре S-гликопротеина оценок протеолитической доступности в области сайта S1/S2 (позиция R685). Цветовой градиент от синего к красному и далее к желтому соответствует изменениям от минимальной оценки протеолитической доступности, равной 0, к максимальной, равной 1.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Мы представили биоинформатический метод идентификации протеаз человека, способных к протеолитическому расщеплению гликопротеинов оболочки вирусов. Идея предлагаемого подхода заключается в вычислении оценок протеолитической эффективности (CS) для пептидных связей белка с использованием модели специфичности протеаз человека и применении разработанного нами ранее метода оценки структурной уязвимости участков белка к протеолиту. Теперь мы разработали модели специфичности протеолитических ферментов к аминокислотной последовательности на основе известных субстратов для 169 протеаз человека. Предлагаемый подход применен к S-белку SARS-CoV-2 штамма Wuhan. В результате разработанный метод прошел успешную валидацию. Так, применив модели специфичности протеаз к аминокислотной последовательности, мы идентифицировали протеазы фурина и TMPRSS2 как характеризующиеся наивысшими значениями CS для сайтов S1/S2 и S2' соответственно. Применив метод предсказания уязвимости участков белка к протеолиту на основе анализа трехмерной структуры S-белка, мы убедились в доступности известных сайтов протеолиза этого вирусного гликопротеина для протеаз. Полученные нами результаты согласуются с результатами исследований сайтов протеолиза S-гликопротеина SARS-CoV-2 [5]. Представленный метод, разработанный для решения задачи поиска протеаз, активных относительно бел-

ков оболочки вирусов, может применяться для идентификации субстратов и активных в их отношении протеаз, задействованных в любых клеточных процессах, связанных с протеолитической регуляцией, — например, таких как передача сигнала в клетке, коагуляция и апоптоз.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-04-60066.

В данной работе отсутствуют исследования человека или животных.

Авторы заявляют об отсутствии конфликта интересов.

Разработанный метод находится в свободном доступе по адресу: <https://github.com/KazanovLab/ProteaseSpecificityModels>

СПИСОК ЛИТЕРАТУРЫ

1. Ramage H., Cherry S. (2015) Virus-host interactions: from unbiased genetic screens to function. *Annu. Rev. Virol.* **2**, 497–524. doi: 10.1146/annurev-virology-100114-055238
2. Li G., Hilgenfeld R., Whitley R., De Clercq E. (2023) Therapeutic strategies for COVID-19: progress and lessons learned. *Nat. Rev. Drug Discov.* **22**, 449–475. doi: 10.1038/s41573-023-00672-y
3. V'kovski P., Kratzel A., Steiner S., Stalder H., Thiel V. (2021) Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **19**, 155–170. doi: 10.1038/s41579-020-00468-6

4. Baggen J., Vanstreels E., Jansen S., Daelemans D. (2021) Cellular host factors for SARS-CoV-2 infection. *Nat. Microbiol.* **6**, 1219–1232. doi: 10.1038/s41564-021-00958-0
5. Takeda M. (2022) Proteolytic activation of SARS-CoV-2 spike protein. *Microbiol. Immunol.* **66**, 15–23. doi: 10.1111/1348-0421.12945
6. Jackson C.B., Farzan M., Chen B., Choe H. (2022) Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23**, 3–20. doi: 10.1038/s41580-021-00418-x
7. Walls A.C., Park Y.J., Tortorici M.A., Wall A., McGuire A.T., Veesler D. (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell.* **181**, 281–292.e6. doi: 10.1016/j.cell.2020.02.058
8. Zabiegala A., Kim Y., Chang K.O. (2023) Roles of host proteases in the entry of SARS-CoV-2. *Anim. Dis.* **3**(1), 12. doi: 10.1186/s44149-023-00075-x
9. Benton D.J., Wrobel A.G., Xu P., Roustan C., Martin S.R., Rosenthal P.B., Skehel J.J., Gamblin S.J. (2020) Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature.* **588**, 327–330. doi: 10.1038/s41586-020-2772-0
10. Matsuyama S., Nao N., Shirato K., Kawase M., Saito S., Takayama I., Nagata N., Sekizuka T., Kato H., Kato F., Sakata M., Tahara M., Kutsuna S., Ohmagari N., Kuroda M., Suzuki T., Kageyama T., Takeda M. (2020) Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. *Proc. Natl. Acad. Sci. USA.* **117**, 7001–7003. doi: 10.1073/pnas.2002589117
11. Shang J., Wan Y., Luo C., Ye G., Geng Q., Auerbach A., Li F. (2020) Cell entry mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA.* **117**, 11727–11734. doi: 10.1073/pnas.2003138117
12. Callaway E. (2020) The coronavirus is mutating — does it matter? *Nature.* **585**, 174–177. doi: 10.1038/d41586-020-02544-6
13. Lubinski B., Whittaker G.R. (2023) The SARS-CoV-2 furin cleavage site: natural selection or smoking gun? *Lancet Microbe.* **4**(8), e570. doi: 10.1016/S2666-5247(23)00144-1
14. Whittaker G.R. (2021) SARS-CoV-2 spike and its adaptable furin cleavage site. *Lancet Microbe.* **2**(10), e488–e489. doi: 10.1016/S2666-5247(21)00174-9
15. Wu Y., Zhao S. (2021) Furin cleavage sites naturally occur in coronaviruses. *Stem Cell Res.* **50**, 102–115. doi: 10.1016/j.scr.2020.102115
16. Chan Y.A., Zhan S.H. (2021) The emergence of the spike furin cleavage site in SARS-CoV-2. *Mol. Biol. Evol.* **39**(1), msab327. doi: 10.1093/molbev/msab327
17. Whittaker G.R., Daniel S., Millet J.K. (2021) Coronavirus entry: how we arrived at SARS-CoV-2. *Curr. Opin. Virol.* **47**, 113–120. doi: 10.1016/j.coviro.2021.02.006
18. Liu Z., Zheng H., Yuan R., Li M., Lin H., Peng J., Xiong Q., Sun J., Li B., Wu J., Ke C., Hulswit R.J.G., Bowden T.A., Rambaut A., Pybus O.G., Loman N., Lu J. (2020) Identification of common deletions in the spike protein of SARS-CoV-2. *J. Virol.* **94**, e00790-20. doi: 10.1128/JVI.00790-20
19. Park J.E., Li K., Barlan A., Fehr A.R., Perlman S., McCray P.B., Gallagher T. (2016) Proteolytic processing of middle east respiratory syndrome coronavirus spikes expands virus tropism. *Proc. Natl. Acad. Sci. USA.* **113**, 12262–12267. doi: 10.1073/pnas.1608147113
20. Baggen J., Jacquemyn M., Persoons L., Vanstreels E., Pye V.E., Wrobel A.G., Calvaresi V., Martin S.R., Roustan C., Cronin N.B., Reading E., Thibault H.J., Vercruysse T., Maes P., De Smet F., Yee A., Nivitchanyong T., Roell M., Franco-Hernandez N., Rhinn H., Mamchak A.A., Young-Chapon M.A., Brown E., Cherepanov P., Daelemans D. (2023) TMEM106B is a receptor mediating ACE2-independent SARS-CoV-2 cell entry. *Cell.* **186**, 3427–3442. doi: 10.1016/j.cell.2023.06.005
21. Meng B., Abdullahi A., Ferreira I.A.T.M., Goonawardane N., Saito A., Kimura I., Yamasoba D., Gerber P.P., Fatihi S., Rathore S., Zepeida S.K., Papa G., Kemp S.A., Ikeda T., Toyoda M., Tan T.S., Kuramochi J., Mitsunaga S., Ueno T., Shirakawa K., Takaori-Kondo A., Brevini T., Mallery D.L., Charles O.J., CITIID-NIHR BioResource COVID-19 Collaboration, Genotype to Phenotype Japan (G2P-Japan) Consortium, Ecuador-COVID19 Consortium, Bowen, J. E., Joshi A., Walls A.C., Jackson L., Martin D., Smith K.G.C., Bradley J., Briggs J.A.G., Choi J., Madissoon E., Meyer K.B., Mlcochova P., Ceron-Gutierrez L., Doffinger R., Teichmann S.A., Fisher A.J., Pizzuto M.S., de Marco A., Corti D., Hosmillo M., Lee J.H., James L.C., Thukral L., Veesler D., Sigal A., Sampaziotis F., Goodfellow I.G., Matheson N.J., Sato K., Gupta R.K. (2022) Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. *Nature.* **603**, 706–714. doi: 10.1038/s41586-022-04474-x
22. Rawlings N.D., Barrett A.J., Thomas P.D., Huang X., Bateman A., Finn R.D. (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **46**, D624–D632. doi: 10.1093/nar/gkx1134
23. Wasserman W.W., Sandelin A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287. doi: 10.1038/nrg1315
24. Schechter I., Berger A. (1968) On the active site of proteases. 3. Mapping the active site of papain, specific peptide inhibitors of papain. *Biochem. Biophys. Res. Commun.* **32**, 898–902. doi: 10.1016/0006-291x(68)90326-4
25. Matveev E.V., Safronov V.V., Ponomarev G.V., Kazanov M.D. (2023) Predicting structural susceptibility of proteins to proteolytic processing. *Int. J. Mol. Sci.* **24**, 10761. doi: 10.3390/ijms241310761
26. Igarashi Y., Eroshkin A., Gramatikova S., Gramatikoff K., Zhang Y., Smith J.W., Osterman A.L., Godzik A. (2007) CutDB: a proteolytic event data-

- base. *Nucleic Acids Res.* **35**(Database issue), D546–9. doi: 10.1093/nar/gkl813
27. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., Louppe G. (2011) Scikit-Learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830. doi:10.48550/arXiv.1201.0490
 28. The UniProt Consortium (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699. doi: 10.1093/nar/gky092
 29. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528. doi: 10.1093/nar/gky949
 30. Waterhouse A., Bertoni M., Bienert S., Studer G., Tauriello G., Gumienny R., Heer F.T., De Beer T.A.P., Rempfer C., Bordoli L., Lepore R., Schwelde T. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303. doi: 10.1093/nar/gky427
 31. Hoffmann M., Kleine-Weber H., Pöhlmann S.A. (2020) Multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell.* **78**, 779–784.e5. doi: 10.1016/j.molcel.2020.04.022

A Bioinformatics Method for Identification of Human Proteases Active against Viral Envelope Glycoproteins: A Case Study on the SARS-CoV-2 Spike Protein

E. V. Matveev^{1, 2, 3}, G. V. Ponomarev^{1, 2}, and M. D. Kazanov^{1, 2, 3, *}

¹ Skolkovo Institute of Science and Technology, Moscow, 121205 Russia

² Kharkevich Institute for Information Transmission Problem, Moscow, 127051 Russia

³ Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, 117997 Russia

*e-mail: mkazanov@gmail.com

Many viruses, including SARS-CoV-2, the coronavirus responsible for the COVID-19 pandemic, enter host cells through a process cell-viral membrane fusion that is activated by proteolytic enzymes. Typically, these enzymes are host cell proteases. Identifying the proteases that activate the virus is not a simple task but is important for the development of new antiviral drugs. In this study, we developed a bioinformatics method for identifying proteases that can cleave viral envelope glycoproteins. The proposed approach involves the use of predictive models for the substrate specificity of human proteases and the application of structural analysis method for predicting the vulnerability of protein regions to proteolysis based on their 3D structures. Specificity models were constructed for 169 human proteases using information on their known substrates. A previously developed method for structural analysis of potential proteolysis sites was applied in parallel with specificity models. Validation of the proposed approach was performed on the SARS-CoV-2 spike protein, the proteolysis sites of which had been well studied.

Keywords: human proteases, virus glycoprotein, proteolytic activation, spike protein, SARS-CoV-2