УЧЁНЫЕ ЗАПИСКИ ПСПбГМУ им. акад. И. П. ПАВЛОВА
The Scientific Notes of Pavlov University
journal homepage: www.sci-notes.ru

УДК 616-092 :575-004.4.019.941
https://doi.org/10.24884/1607-4181-2025-32-1-11-20

Д. С. Буг[1]*, А. Н. Наркевич[2], Н. В. Петухова[1]

[1] Первый Санкт-Петербургский государственный медицинский университет имени академика И. П. Павлова
197022, Россия, Санкт-Петербург, ул. Льва Толстого, д. 6-8
[2] Южно-Уральский государственный медицинский университет
454092, Россия, г. Челябинск, ул. Воровского, д. 64

ОБЗОР ПРОГРАММ ДЛЯ ОЦЕНКИ ПАТОГЕННОСТИ ГЕНЕТИЧЕСКИХ ВАРИАНТОВ

Поступила в редакцию 04.09.2024 г.; принята к печати 25.02.2025 г.

Резюме

В настоящее время молекулярно-генетические методы играют важную роль в диагностике ряда патологий. Внедрение массового параллельного секвенирования значительно увеличило объем данных, описывающих варианты ДНК пациентов с различными заболеваниями, но клиническое значение многих из этих результатов остается неизвестным. Для оценки эффекта генетических вариантов широко используется автоматическое определение клинической значимости вариантов при помощи программ-предикторов. Отечественные и международные руководства по интерпретации данных, полученных методами массового параллельного секвенирования, рекомендуют использовать программы-предикторы для определения клинического значения генетических вариантов. Однако принципы работы и характеристики этих программ в научной литературе описаны недостаточно. В данном обзоре на примере наиболее популярных программ-предикторов представлены основные принципы их работы, которые используются для оценки патогенности вариантов.

Ключевые слова: оценка патогенности, вариант, мутация, определение клинического эффекта

Для цитирования: Буг Д. С., Наркевич А. Н., Петухова Н. В. Обзор программ для оценки патогенности генетических вариантов. *Ученые записки ПСПбГМУ им. акад. И. П. Павлова*. 2025;32(1):11 – 20. https://doi.org/10.24884/1607-4181-2025-32-1-11-20.

* **Автор для связи:** Дмитрий Сергеевич Буг, ФГБОУ ВО ПСПбГМУ им. И. П. Павлова Минздрава России, 197022, Россия, Санкт-Петербург, ул. Льва Толстого, д. 6-8. E-mail: bug.dmitrii@yandex.ru.

Dmitrii S. Bug[1]*, Artem N. Narkevich[2], Natalia V. Petukhova[1]

[1] Pavlov University
6-8, L'va Tolstogo str., Saint Petersburg, Russia, 197022
[2] South-Ural State Medical University
64, Vorovskogo str., Chelyabinsk, Russia, 454092

REVIEW OF PROGRAMS FOR ASSESSING THE PATHOGENICITY OF GENETIC VARIANTS

Received 04.09.2024; accepted 25.02.2025

Summary

Currently, molecular genetic methods play an essential role in the diagnostic process for diverse pathologies. The introduction of mass parallel sequencing has significantly increased the amount of data on DNA variants in patients with various diseases, but the clinical significance of many of these findings remains unknown. Widely used methods of variant effect evaluation include the automatic determination of the pathogenicity of variants using specialized predictors. Domestic and international guidelines for the interpretation of data obtained through mass parallel sequencing recommend the use of predictive programs to determine the clinical significance of genetic variants. However, there is a lack of detailed information about the principles and characteristics of these programs in the scientific literature. In this review, we present the basic principles that are used to evaluate the pathogenicity of variations using the example of some of the most widely used predictive programs.

Keywords: pathogenicity assessment, variant, mutation, clinical effect evaluation

For citation: Bug D. S., Narkevich A. N., Petukhova N. V. Review of programs for assessing the pathogenicity of genetic variants. *The Scientific Notes of Pavlov University*. 2024;31(4):11 – 20. (In Russ.). https://doi.org/10.24884/1607-4181-2024-31-4-11-20. * **Corresponding author:** Dmitrii S. Bug, Pavlov University, 6-8, L'va Tolstogo str., Saint Petersburg, 197022, Russia. E-mail: bug.dmitrii@yandex.ru.

ВВЕДЕНИЕ

Развитие методов молекулярной биологии и интенсификация геномных исследований позволили раскрыть молекулярные механизмы, лежащие в основе нарушений, возникающих в клетке при возникновении генетических заболеваний. Благодаря этому удается определять новые биомаркеры, обнаружение которых у пациентов может помочь определить прогноз заболевания или назначить целевую терапию.

Однако большие объемы данных, полученные с помощью методов современной молекулярной биологии, требуют большого опыта и вычислительных мощностей для обработки, идентификации и классификации генетических вариантов, которые могут дать значимую информацию. Следует отметить, что под генетическим вариантом подразумевается любое изменение гена относительно референсной последовательности, которое может быть патогенным или доброкачественным. Таким образом, с внедрением современных технологий значительно увеличилось число генетических вариантов, обнаруживаемых у пациентов, значимость которых не может быть определена [1, 2].

В настоящее время проводятся различные исследования для оценки клинической значимости того или иного варианта. Широко используемый подход *in silico* предполагает автоматическое определение клинической значимости вариантов с помощью так называемых программ-предикторов. Эти программы, основанные на эволюционной консервативности и структурных характеристиках белков, рекомендованы для использования при диагностике заболеваний, вызванных герминальными и соматическими мутациями.

В частности, «Руководство по интерпретации данных последовательности ДНК человека, полученных методами массового параллельного секвенирования (MPS)» рекомендует использовать программы-предикторы в случае, если вариант нуклеотидной последовательности не был описан ранее и не представлен ни в одной из баз данных или сведения о нем недостаточны [3]. Также в работе «Интерпретация соматических генетических вариантов, выявленных методом высокопроизводительного секвенирования опухолевой ДНК, на примере онкологических заболеваний детского возраста» программы-предикторы предполагается использовать для оценки патогенности соматических мутаций [4].

В настоящее время программы для определения клинической значимости генетических вариантов основаны на выравнивании нуклеотидных последовательностей сходных генов [5, 6, 7, 8] или аминокислотных последовательностей их белковых продуктов [9, 10, 11, 12, 13, 14, 15]. Под выравниванием подразумевается сравнение всего множества генов, продуцирующих нормальный, полностью функционирующий белок, включая референсную последовательность гена человека и других организмов, и гена обследуемого пациента с вариантом. Родственные гены, которые способны производить функциональный продукт и выполняют одинаковую функцию, называют гомологами, или гомологичными генами.

Сопоставление последовательностей исследуемого гена и его гомологов позволяет определить, существуют ли какие-либо различия в генетических вариантах между гомологичными генами. Наличие варианта среди гомологичных последовательностей указывает на то, что он может быть безвредным и не нарушать функцию гена. Таким образом, основой для оценки клинической значимости того или иного варианта является поиск гомологичных генов, который включает в себя изучение эволюции генов. Следует отметить, что изучение эволюции может производиться в ходе исследования или аминокислотных, или нуклеотидных последовательностей, однако результаты оказываются более информативными при сравнении белков [16].

ИСХОДНЫЕ БАЗЫ ДАННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Основой любого метода оценки патогенности вариантов с неопределенной значимостью является база данных нуклеотидных или аминокислотных последовательностей, которые будут сравниваться в процессе определения того, является ли тот или иной генетический вариант допустимым. Существует четыре основных источника данных о последовательностях белков: GenPept, RefSeq, TrEMBL и SWISS-PROT, каждый из которых связан с соответствующими базами данных о нуклеотидных последовательностях.

GenPept является частью GenBank, который представляет собой базу данных, содержащую генетические последовательности из Национального института здравоохранения США. GenBank также содержит информацию о длине каждой последовательности, типе молекулы, которую она представляет (белок, ген, транскрипт и др.), координатах транслируемых областей, источнике биоматериала и названии организма, к которому относится последовательность. База данных GenBank поддерживается самими исследователями и снабжается информацией из других баз данных, таких как Европейская лаборатория молекулярной биологии (EMBL) и Японский банк данных ДНК. Это взаимодействие координируется Международной организацией по сотрудничеству в области базы данных нуклеотидных последовательностей.

Структура базы данных нуклеотидных последовательностей EMBL, Европейского архива нуклеотидов (ENA), во многом схожа с GenBank. Однако она отличается тем, что не содержит аминокислотных последовательностей. Для систематизации белковых структур в Европейской лаборатории молекулярной биологии была создана специальная база данных UniProtKB/TrEMBL. Объекты в базах

данных ENA и UniProtKB/TrEMBL описывают длину последовательности, ее функцию, источник биоматериала и таксономическую классификацию исходного организма. Как и GenBank, база данных ENA обновляется самими исследователями. Кроме того, система EMBL обеспечивает проверку контаминации векторами в материалах, предоставляемых исследователями.

Кроме того, существует проект RefSeq, который поддерживается Национальным центром биотехнологической информации (NCBI). Объекты базы данных RefSeq создаются по-разному, в зависимости от типа изучаемого организма, при этом некоторые свойства последовательностей предсказываются автоматически при депонировании. Специалисты NCBI контролируют относительно небольшую часть бактериальных и эукариотических геномов, содержащихся в RefSeq.

Существует множество ресурсов, которые используют информацию из других баз данных. Примером этого является неизбыточная база данных белковых последовательностей NCBI NR, которая включает последовательности из GenPept, UniProtKB/Swiss-Prot, RefSeq и ряда других источников, таких как PIR, PDF и PDB.

В актуальных программах-предикторах в качестве источника последовательностей обычно используется UniProtKB или ее вариации, объединяющие в кластеры белки, идентичные на 90 % (UniRef90) или на 100 % (UniRef100). В частности, сравнительно более ранние программы, такие как MutationAssessor, SIFT, MutPred и MAPP, используют UniProtKB в качестве источника, в то время как более современные EVE и AlphaMissense основаны на UniRef100 и UniRef90 соответственно.

Другие программы используют Ensembl (MutationTaster, LRT), NCBI NR (PROVEAN) или собственные базы данных (PANTHER, Align-GVGD).

ФИЛОГЕНЕТИЧЕСКАЯ РЕКОНСТРУКЦИЯ

Филогенетическая реконструкция — это процесс установления эволюционных взаимосвязей, основанный на изучении структурных признаков. Это важный шаг в оценке патогенности генетических вариантов, который позволяет нам исключить из набора данных для исследования белки, функционально отличающиеся от продукта изучаемого гена [17].

В настоящее время программы, применяемые для оценки клинических эффектов генетических вариантов, используют автоматизированные инструменты для поиска гомологов. Эти инструменты направлены на поиск наиболее схожих белков или генов из баз данных последовательностей [9]. Этот поиск основан на исходном предположении о тесной взаимосвязи между структурой и функцией белка, из чего следует, что гомологи выполняют сходные функции. Таким образом, если в одной из гомологичных последовательностей обнаруживается некоторый вариант, то его наличие не свидетельствует о значимом изменении функции. И наоборот, если таких изменений не удается обнаружить ни в одном из большого числа гомологов, то можно предположить, что наличие данного варианта влияет на функцию белка и выживаемость организма в целом [9, 12, 17].

Поиск гомологичных последовательностей обычно проводится с помощью программы базового инструмента поиска локального выравнивания BLAST или ее модификаций. Этот первоначальный отбор последовательностей производится по принципу максимального сходства с целевым белком. В качестве меры сходства выступает E-значение, определяемое как количество последовательностей в базе данных, которые случайно оказываются более похожими на искомую последовательность, чем найденная.

Например, программы MutationAssessor, PolyPhen2, SNPs&GO, MutationTaster, PROVEAN и PANTHER используют метод BLAST. Однако в PROVEAN дополнительно применяется CD-HIT — алгоритм, осуществляющий кластеризацию последовательностей исходя из порогового уровня идентичности [18]. Например, при выборе порогового уровня идентичности в 80 % каждый кластер будет содержать последовательности, идентичные на 80 %. Также существуют усовершенствования BLAST, которые итеративно подстраивают систему поиска под уже обнаруженные последовательности. Они включают CS-BLAST и PSI-BLAST, применяемые в ConSurf, SIFT и Mutpred. Схожим итеративным механизмом поиска обладают программы JackHMMER и HHSearch, используемые в EVE и AlphaMissense, соответственно.

МНОЖЕСТВЕННОЕ ВЫРАВНИВАНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Следующим этапом после определения перечня последовательностей, структурно схожих с искомой, является определение консервативности позиций в последовательности при помощи множественного выравнивания. Следует отметить, что на этапе филогенетической реконструкции также может осуществляться множественное выравнивание, результат которого используется для вычисления расстояний между последовательностями, формирования матрицы и последующего построения филогенетического дерева.

Суть данного этапа состоит в выравнивании позиций различных последовательностей таким образом, чтобы максимальное количество аминокислот (или нуклеотидов) в соответствующих позициях совпадало. Влияние миссенс-варианта можно определить, посмотрев, какие аминокислоты (или нуклеотиды) присутствуют в положении, на которое влияет этот вариант, и в соответствующих положениях в гомологичных последовательностях [19].

Наиболее часто используемым алгоритмом является прогрессивное выравнивание. В ходе этого

процесса сначала соединяются две ближайшие последовательности, а затем постепенно добавляются новые. Порядок, в котором добавляются эти новые последовательности, определяется филогенетическим деревом, в котором используются эволюционные расстояния, рассчитанные на основе парных последовательностей. Примеры программ, использующих этот подход, включают ClustalW [20], Clustal Omega [21] и MAFFT [22]. Также существует итеративное прогрессивное выравнивание, которое отличается от предыдущего метода тем, что все последовательности корректируются на каждом этапе [23]. Этот подход используется в программе MUSCLE [24]. Другие менее распространенные методы включают сегментное выравнивание (DIALIGN [25]) и генетический алгоритм (SAGA [26]). В ряде исследований оценивалась надежность результатов, полученных с помощью различных программ, и было установлено, что MAFFT наиболее точно воспроизводит стандартные выравнивания [27].

Следует отметить, что BLAST и другие программы для отбора гомологичных последовательностей также продуцируют выравнивания, которые могут сразу использоваться для сравнения белков. Однако некоторые программы все же осуществляют множественное выравнивание последовательностей отдельным этапом после обнаружения гомологов. В частности, MutationAssessor и LRT используют MUSCLE, в ConSurf применяет MAFFT.

Выравнивание нуклеотидных последовательностей используется в таких программах, как phastCons [8] и phyloP [7] для оценки эволюционной консервативности фрагментов генома. Выравнивания геномной ДНК создаются путем сопоставления геномов разных видов с основным изучаемым геномом (чаще всего это геном человека) с помощью программы LASTZ. Затем эти парные выравнивания совмещаются для формирования множественного выравнивания.

ПРЕДИКТОРЫ НА ОСНОВЕ ВЫРАВНИВАНИЯ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

SIFT была первой программой, которая выявляла вредные мутации на основе информации об эволюции изучаемого гена [28]. Основной принцип SIFT заключается в выявлении законсервированных аминокислот (в данном случае — в значении наиболее распространенных аминокислот в определенном положении) и вычислении вероятности того, что замена на один из девятнадцати альтернативных остатков не приведет к значимому изменению функции белка. В частности, мутации консервативных аминокислот с большей вероятностью приведут к повреждению, в то время как изменения в вариабельных положениях с большей вероятностью не окажут значимое воздействие. Программы FATHMM [29] и PANTHER [30] используют схожий подход оценки консервативности для определения патогенности вариантов. В дополнение к этому ConSurf использует информацию о скорости эволюции аминокислот для оценки уровня консервативности, что подразумевает построение филогенетического дерева и оценку взаимоотношений между гомологами на его основе [31].

Программа MutationAssessor использует более продвинутый подход, который также основан на оценке сохранности аминокислот человеческого белка [14]. В этом случае рассчитываются два показателя консервативности: внутри семейств и внутри подсемейств. В данном случае под подсемействами подразумеваются кластеры схожих последовательностей внутри семейства. В результате консервативность рассчитывается по шкалам соответствующего семейства и подсемейства последовательностей, что позволяет уточнить возможные функциональные последствия замены.

Предиктор PROVEAN способен оценивать патогенность не только миссенс-вариантов, но и коротких вставок и делеций без сдвига рамки считывания. В данном случае осуществляется серия парных выравниваний исходного и мутантного белка с гомологичными последовательностями, и вычисляется разница между числовой оценкой схожести последовательностей в одном и другом случае [12].

Программа LRT [32] также учитывает происходящие в той или иной позиции синонимичные замены и выполняет тест отношения правдоподобия для оценки вероятности того, что та или иная позиция подвергается негативному отбору, то есть частота синонимичных замен в ней значимо выше, чем частота миссенс-вариантов.

Другие программы используют дополнительные свойства белков для оценки консервативности. Например, физико-химические свойства аминокислот используются предикторами MAPP и Align-GVGD [33, 34]. Для того, чтобы учитывать эти дополнительные данные, MAPP применяет метод главных компонент для снижения размерности данных, после чего вычисляет расстояние между двумя точками, соответствующими исходной и мутантной последовательностям. Align-GVGD, в свою очередь, осуществляет вычисление расстояний Грэнтема, которые учитывают физико-химические различия между аминокислотами.

Для определения патогенных и нейтральных изменений PolyPhen2 применяет наивный байесовский классификатор, основанный на эволюции и функции рассматриваемой позиции, а также структуре молекулы белка [9]. Данные о пространственных характеристиках белка используются также программой AlphaMissense, в основе которой лежит нейросеть Evoformer. Предиктор EVE также, хотя и опосредованно, основан на структурных данных, и использует особую нейросетевую архитектуру, вариационный автоэнкодер [13].

Одним из главных недостатков программ-предикторов является игнорирование медико-био-

логического контекста гена: какие заболевания ассоциированы с мутациями в исследуемом гене, насколько генетические варианты в нем (в частности, миссенс-замены) распространены в человеческой популяции, входит ли позиция мутации в какой-либо домен белка-продукта и прочие аспекты игнорируются для упрощения и последующей автоматизации процесса. Разработчики программы SNPs&GO попытались нивелировать этот фактор, используя информацию ресурса Gene Ontology, в котором основные характеристики генов, включая наличие или отсутствие белковых продуктов, функцию и другие, закодированы в виде категориальных переменных, удобных для использования в машинном обучении [11]. Патогенность определяется при помощи метода опорных векторов, на вход подаются данные ресурса Gene Ontology и сведения о консервативности позиции варианта.

Метод случайного леса применяется сразу в двух программах-предикторах: MutationTaster и MutPred. В них, помимо консервативности последовательности, оцениваются структурные и функциональные характеристики белка (например, связь аминокислоты в той или иной позиции с определенной функцией).

ПРЕДИКТОРЫ НА ОСНОВЕ ВЫРАВНИВАНИЯ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Существует проблема определения клинического эффекта вариантов, которые не относятся к миссенс-заменам, но также способны играть значительную роль в развитии различных заболеваний, включая злокачественные новообразования [35]. Некоторые из этих вариантов локализуются в местах сплайсинга. Для оценки их клинической значимости могут быть использованы специализированные программы с относительно высокой чувствительностью и специфичностью. Эти программы могут анализировать варианты в донорском месте сплайсинга с точностью до 99 % [36]. Однако их нельзя использовать для оценки патогенности вариантов других типов.

Информация о консервативности участков генома может быть использована для определения клинического воздействия вариантов, не относящихся к миссенс-заменам [37]. Преимущество этого подхода заключается в том, что он позволяет анализировать не только аминокислотные замены, но и другие изменения в нуклеотидных последовательностях, такие как синонимичные замены и варианты интронов. Эти изменения невозможно исследовать, используя только анализ последовательности белка. Существует несколько доступных программ, которые могут помочь идентифицировать консервативные участки в выравнивании нуклеотидных последовательностей: PhyloP [7], PhastCons [8], GERP [5], SiPhy [6].

GERP основан на вычислении частоты замен в каждой позиции выравнивания в нуклеотидных последовательностях и сравнении ее со средним числом замен во всех позициях выравнивания, исключая пробелы. Другими словами, эта программа сравнивает наблюдаемую и ожидаемую скорости эволюции.

GERP, наряду с несколькими другими статистическими филогенетическими методами, используется в phyloP. Этот подход основан на выявлении положений в геноме, в которых наблюдаются какие-либо отклонения от нейтральной эволюции, такие как консервация или ускорение эволюции. Эти явления определяются с помощью методов LRT [38], Score test [39], распределения числа замен [40] и GERP [5].

PhastCons использует филогенетические скрытые марковские модели для отнесения каждого сегмента генома к одной из двух категорий: консервативные и вариабельные. Этот подход аналогичен подходу, используемому в SiPhy, который учитывает не только сохранение в определенном положении, но и закономерности замен, наблюдаемые в последовательностях, подвергающихся эволюционному отбору.

Существует множество других программ, которые анализируют влияние вариаций в некодирующих областях. Все эти программы используют ансамблевый подход, что означает, что они объединяют результаты нескольких методов. Они основаны на показателях консервативности, которые рассчитываются с использованием таких методов, как phyloP, GERP, PhastCons и fitCons, среди прочих.

АНСАМБЛЕВЫЙ ПОДХОД

Суть ансамблевого подхода заключается в оценке влияния различных вариантов на основе информации из различных источников, таких как филогенетический анализ, функциональные и структурные данные, консервативности генома и прочих. Этот подход использует методы машинного обучения для анализа этой информации и составления прогнозов о влиянии мутаций на гены и их функции.

Большинство программ, использующих ансамблевый подход, сосредоточены на методах анализа сохранения генома, таких как GERP и phastCons. Они также используют результаты программ, изучающих эволюцию генов, таких как SIFT и PolyPhen, а также данные о регуляторных элементах и транскриптах. Методы контролируемого обучения также используют базы данных о мутациях и полиморфизмах для составления прогнозов. Хотя конкретные источники информации, используемые в этих программах, могут незначительно отличаться, основное различие заключается в способах обработки и анализа данных.

CADD — это первая реализация ансамблевого метода, основанная на методе опорных векторов (SVM) с учителем [41]. Основным недостатком CADD является его неспособность учитывать нелинейные взаимосвязи между объектами. Эта проблема была решена в программе DANN, которая использует глубокие нейронные сети [42].

Первой программой, в которой было реализовано неконтролируемое машинное обучение, является Eigen: в ней использовались мутации и полиморфизмы, клинические последствия которых известны, однако были скрыты от алгоритма [43]. В FATHMM-MKL реализован подход, основанный на множественном обучении ядра. Он обладает преимуществами с точки зрения большей чувствительности и специфичности при определении патогенности вариантов в некодирующих областях [44]. Кроме того, широкое распространение получили программы, основанные на тесте отношения правдоподобия (MetaLR) [45], градиентном бустинге (M-CAP) [46] и алгоритме Random forest (REVEL) [47].

ОБСУЖДЕНИЕ

Согласно рекомендациям по интерпретации вариантов, полученных с помощью массового параллельного секвенирования, процесс классификации вариантов требует проверки различных критериев патогенности (и доброкачественности) на нескольких уровнях, от вспомогательных до очень сильных (убедительных). Чтобы удалить варианты, связь которых с некоторым заболеванием сомнительно, можно использовать различные фильтры данных. Они основаны на оценке качества секвенирования, информации о частоте встречаемости в человеческой популяции, биологических и клинических данных, таких как сегрегация и информация о функции белка, положении вариантов в белке (активные сайты или «горячие точки»), типе варианта (например, синонимичный, миссенс, со сдвигом рамки считывания и прочие), а также на результатах работы предикторов. В данном обзоре были рассмотрены основные особенности некоторых программ-предикторов, которые изложены в таблице.

Однако современные программы для определения клинического значения вариантов обладают и рядом недостатков. В частности, в большинстве случаев они используют определенную базу данных последовательностей для определения консервативности, не позволяя исследователю воспользоваться более современным источником аминокислотных или нуклеотидных последовательностей. Актуальность данной проблемы подчеркивается тем, что многие программы-предикторы были созданы более 10 лет назад.

Одной из главных проблем программ-предикторов является игнорирование того факта, что гомологичные гены не являются однородной группой и делятся на ортологи и паралоги по отношению к искомой последовательности. Ортологи — это гены, которые возникли в результате дивергенции видов и существуют в разных организмах. Паралоги возникают в результате дупликации наследственного гена и существуют в пределах одного и того же организма. Если наследственный ген выполняет какую-то функцию, необходимую для выживания, то эту функцию должны выполнять все ортологи. Однако функция паралогов может меняться. Это явление называется «ортологической гипотезой»: ортологи обычно выполняют одни и те же функции, в то время как паралоги — разные [48].

Наличие последовательностей паралогов в наборе данных, используемом для определения клинической значимости варианта с неясным эффектом, может привести к ошибкам. Эти ошибки возникают из-за того, что структурные особенности паралогов, которые функционально отличаются от изучаемой последовательности, воспринимаются как нейтральные вариации, не влияющие на функцию. Это может привести к ошибке в определении значимости конкретного варианта у пациента, поскольку он может быть неверно истолкован как генетическая вариация в рамках нормы. Поэтому важно учитывать возможное наличие паралогов при интерпретации данных и определении клинической значимости генетических вариантов [17].

Следующая проблема автоматизированных программ заключается в использовании исключительно полноразмерных белковых последовательностей в качестве эволюционно-функциональной единицы. Однако эволюцию мультидоменных генов следует рассматривать с точки зрения эволюции отдельных доменов [49].

Другая проблема, связанная с доступными в настоящее время программами для анализа эффектов вариантов, основанными на выравнивании нуклеотидной последовательности, заключается в том, что они базируются на количественном показателе консервативности в каждой позиции, не учитывая качества изменений в ортологичных генах.

Наконец, существенной проблемой является использование методов неинтерпретируемого машинного обучения для интерпретации результатов, которые используются в некоторых программах. Хотя использование этих методов может привести к очевидному улучшению результатов, их внутренняя работа может игнорировать важные аспекты биомедицинского контекста [17, 50].

Следует помнить, что анализ *in silico* является лишь одним из параметров для классификации вариантов в соответствии с отечественными и международными рекомендациями. Кроме него используются, например, анализ популяционных данных и функциональных характеристик белка, сведения о сегрегации генетического варианта и о возможном возникновении его *de novo* [3]. Применение программ-предикторов является необходимым, но не единственным этапом для проведения качественного анализа патогенности вариантов, который, в свою очередь, может выявить ряд значимых диагностических, прогностических и предиктивных биомаркеров, позволяющих улучшить качество медицинской помощи за счет индивидуализации тактики ведения пациентов.

Основные особенности рассмотренных предикторов

The main features of the reviewed predictors

Наименование	База данных	Методы отбора и выравнивания последовательностей	Метод оценки патогенности
MutationAssessor	UniProtKB	BLAST, MUSCLE	Оценка изменения энтропии в результате мутации
ConSurf	UniRef90	CS-BLAST, MAFFT	Оценка консервативности аминокислот на основе вычисления скорости эволюции позиции
SIFT	UniProtKB	PSI-BLAST	Оценка вероятности возникновения альтернативной аминокислоты в позиции
PolyPhen2	UniRef100	BLAST	Наивный байесовский классификатор
EVE	UniRef100	JackHMMER, EVcouplings	Вариационный автоэнкодер
PROVEAN	NCBI NR	BLAST, CD-HIT	Оценка изменения веса выравнивания с гомологами (alignment score) в результате мутации
SNPs&GO	UniRef90	BLAST	Метод опорных векторов
MutationTaster	Ensembl	BLAST	Random forest
MutPred	UniProtKB	PSI-BLAST	Random forest
FATHMM	UniRef90	JackHMMER	Оценка вероятности возникновения альтернативной аминокислоты в позиции
PANTHER	PANTHER database	BLAST	Оценка эволюционной сохранности аминокислот у предковых последовательностей
LRT	Ensembl	EnsemblCompara Genetrees, MUSCLE	Тест отношения правдоподобия
MAPP	UniProtKB	SEMPHY, ClustalW, ProbCons	Метод главных компонент
Align-GVGD	Собственная база данных	Собственные наборы последовательностей, 3DCoffee	Вычисление расстояний Грэнтема
AlphaMissense	UniRef90	HHSearch	Evoformer

Конфликт интересов

Авторы заявили об отсутствии конфликта интересов.

Conflict of interest

Authors declare no conflict of interest.

Соответствие нормам этики

Авторы подтверждают, что соблюдены права людей, принимавших участие в исследовании, включая получение информированного согласия в тех случаях, когда оно необходимо, и правила обращения с животными в случаях их использования в работе. Подробная информация содержится в Правилах для авторов.

Compliance with ethical principles

The authors confirm that they respect the rights of the people participated in the study, including obtaining informed consent when it is necessary, and the rules of treatment of animals when they are used in the study. Author Guidelines contains the detailed information.

ЛИТЕРАТУРА

1. *Cook C. E., Bergman M. T., Finn R. D. et al.* The European Bioinformatics Institute in 2016: Data growth and integration // Nucleic Acids Res. – 2015. – Vol. 44, № D1. – P. D20–D26. https://doi.org/10.1093%2Fnar%2Fgkv1352.

2. *Gagan J., Van Allen E. M.* Next-generation sequencing to guide cancer therapy // Genome Med. – 2015. – Vol. 7, № 1. – P. 80. https://doi.org/10.1186/s13073-015-0203-x.

3. *Рыжкова О., Кардымон О., Прохорчук Е. и др.* Руководство по интерпретации данных последовательности ДНК человека, полученных методами массового параллельного секвенирования (MPS) (редакция 2018, версия 2) // Медицинская генетика. – 2020. – Т. 18, № 2. – С. 3–23. https://doi.org/10.25557/2073-7998.2019.02.3-23.

4. *Спектор М., Ясько Л., Друй А.* Интерпретация соматических генетических вариантов, выявленных методом высокопроизводительного секвенирования опухолевой ДНК, на примере онкологических заболеваний детского возраста // Медицинская генетика – 2021. – Т. 20, № 3. – С. 3–25. https://doi.org/10.25557/2073-7998.2021.03.3-25.

5. *Cooper G. M., Stone E. A., Asimenos G. et al.* Distribution and intensity of constraint in mammalian genomic sequence // Genome Res. – 2005. – Vol. 15, № 7. – P. 901–913. https://doi.org/10.1101/gr.3577405.

6. *Garber M., Guttman M., Clamp M. et al.* Identifying novel constrained elements by exploiting biased substitution patterns // Bioinformatics. – 2009. – Vol. 25, № 12. – P. i54–i62. https://doi.org/10.1093%2Fbioinformatics%2Fbtp190.

7. *Pollard K. S., Hubisz M. J., Rosenbloom K. R. et al.* Detection of nonneutral substitution rates on mammalian phylogenies // Genome Res. – 2009. – Vol. 20, № 1. – P. 110–121. https://doi.org/10.1101/gr.097857.109.

8. *Siepel A., Bejerano G., Pedersen J. S. et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes // Genome Res. – 2005. – Vol. 15, № 8. – P. 1034–1050. https://doi.org/10.1101/gr.3715005.

9. *Adzhubei I. A., Schmidt S., Peshkin L. et al.* A method and server for predicting damaging missense mutations // Nat

Methods. – 2010. – Vol. 7, № 4. – P. 248–249. https://doi.org/10.1038/nmeth0410-248.

10. *Ashkenazy H., Abadi S., Martz E. et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules // Nucleic Acids Res. – 2016. – Vol. 44, № W1. – P. W344–W350. https://doi.org/10.1093/nar/gkw408.

11. *Calabrese R., Capriotti E., Fariselli P. et al.* Functional annotations improve the predictive score of human disease-related mutations in proteins // Hum. Mutat. – 2009. – Vol. 30, № 8. – P. 1237–1244. https://doi.org/10.1002/humu.21047.

12. *Choi Y., Sims G. E., Murphy S. et al.* Predicting the Functional Effect of Amino Acid Substitutions and Indels // PLoS ONE. – 2012. – Vol. 7, № 10. – P. e46688. https://doi.org/10.1371/journal.pone.0046688.

13. *Frazer J., Notin P., Dias M. et al.* Disease variant prediction with deep generative models of evolutionary data // Nature. – 2021. – Vol. 599, № 7883. – P. 91–95. https://doi.org/10.1038/s41586-021-04043-8.

14. *Reva B., Antipin Y., Sander C.* Predicting the functional impact of protein mutations: application to cancer genomics // Nucleic Acids Research. – 2011. – Vol. 39, № 17. – P. e118–e118. https://doi.org/10.1093/nar/gkr407.

15. *Sim N., Kumar P., Hu J. et al.* SIFT web server: predicting effects of amino acid substitutions on proteins // Nucleic Acids Research. – 2012. – Vol. 40, № W1. – P. W452–W457. https://doi.org/10.1093/nar/gks539.

16. *Capriotti E., Fariselli P.* Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants // Hum Genet. – 2022. – Vol. 141, № 10. – P. 1649–1658. https://doi.org/10.1007/s00439-021-02419-4.

17. *Adebali O., Reznik A. O., Ory D. S. et al.* Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations // Genetics in Medicine. – 2016. – Vol. 18, № 10. – P. 1029–1036. https://doi.org/10.1038/gim.2015.208.

18. *Li W., Godzik A.* Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences // Bioinformatics. – 2006. – Vol. 22, № 13. – P. 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

19. *Ng P. C.* SIFT: predicting amino acid changes that affect protein function // Nucleic Acids Research. – 2003. – Vol. 31, № 13. – P. 3812–3814. https://doi.org/10.1093/nar/gkg509.

20. *Thompson J. D., Higgins D. G., Gibson T. J.* CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucl Acids Res. – 2007. – Vol. 22, № 22. – P. 4673–4680. https://doi.org/10.1093/nar/22.22.4673.

21. *Sievers F., Wilm A., Dineen D. et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega // Molecular Systems Biology. – 2011. – Vol. 7, № 1. – P. 539. https://doi.org/10.1038/msb.2011.75.

22. *Katoh K.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform // Nucleic Acids Research. – 2002. – Vol. 30, № 14. – P. 3059–3066. https://doi.org/10.1093/nar/gkf436.

23. *Mount D.W.* Using Iterative Methods for Global Multiple Sequence Alignment // Cold Spring Harb Protoc. – 2009. – Vol. 2009, № 7. – P. pdb.top44. https://doi.org/10.1101/pdb.top44.

24. *Edgar R.C.* MUSCLE: a multiple sequence alignment method with reduced time and space complexity // BMC Bioinformatics. – 2004. – Vol. 5, № 1. – P. 113. https://doi.org/10.1186/1471-2105-5-113.

25. *Morgenstern B.* DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment // Bioinformatics. – 2002. – Vol. 15, № 3. – P. 211-218. https://doi.org/10.1093/bioinformatics/15.3.211.

26. *Notredame C.* SAGA: sequence alignment by genetic algorithm // Nucleic Acids Research. – 2002. – Vol. 24, № 8. – P. 1515–1524. https://doi.org/10.1093%2Fnar%2F24.8.1515.

27. *Sievers F., Higgins D.G.* QuanTest2: benchmarking multiple sequence alignments using secondary structure prediction // Bioinformatics. – 2019. – Vol. 36, № 1. – P. 90–95. https://doi.org/10.1093/bioinformatics/btz552.

28. *Ng P.C., Henikoff S.* Predicting Deleterious Amino Acid Substitutions // Genome Res. – 2002. – Vol. 11, № 5. – P. 863–874. https://doi.org/10.1101/gr.176601.

29. *Shihab H.A., Gough J., Cooper D.N. et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models // Human Mutation. – 2012. – Vol. 34, № 1. – P. 57–65. https://doi.org/10.1002/humu.22225.

30. *Tang H., Thomas P. D.* PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation // Bioinformatics. – 2016. – Vol. 32, № 14. – P. 2230–2232. https://doi.org/10.1093/bioinformatics/btw222.

31. *Glaser F., Pupko T., Paz I. et al.* ConSurf: Identification of Functional Regions in Proteinsby Surface-Mapping of Phylogenetic Information // Bioinformatics. – 2002. – Vol. 19, № 1. – P. 163–164. https://doi.org/10.1093/bioinformatics/19.1.163.

32. *Chun S., Fay J. C.* Identification of deleterious mutations within three human genomes // Genome Res. – 2009. – Vol. 19, № 9. – P. 1553–1561. https://doi.org/10.1101/gr.092619.109.

33. *Stone E. A., Sidow A.* Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity // Genome Res. – 2005. – Vol. 15, № 7. – P. 978–986. https://doi.org/10.1101/gr.3804205.

34. *Tavtigian S. V.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral // Journal of Medical Genetics. – 2005. – Vol. 43, № 4. – P. 295–305. https://doi.org/10.1136/jmg.2005.033878.

35. *Frazer J., Notin P., Dias M. et al.* Disease variant prediction with deep generative models of evolutionary data // Nature. – 2021. – Vol. 599, № 7883. – P. 91–95. https://doi.org/10.1038/s41586-021-04043-8.

36. *Moles-Fernández A., Duran-Lozano L., Montalban G. et al.* Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? // Front. Genet. – 2018. – Vol. 9. – P. 366. https://doi.org/10.3389/fgene.2018.00366.

37. *Mahmood K., Jung C., Philip G. et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics // Hum Genomics. – 2017. – Vol. 11, № 1. – P. 10. https://doi.org/10.1186/s40246-017-0104-8.

38. *Pollard K. S., Salama S. R., Lambert N. et al.* An RNA gene expressed during cortical development evolved rapidly in humans // Nature. – 2006. – Vol. 443, № 7108. – P. 167–172. https://doi.org/10.1038/nature05113.

39. *Rao C.R.* Score Test: Historical Review and Recent Developments / Eds. N. Balakrishnan, H. N. Nagaraja, N. Kannan. – Boston, MA: Birkhäuser, 2005. – P. 3–20. https://doi.org/http://dx.doi.org/10.1007/0-8176-4422-9_1.

40. *Siepel A., Pollard K. S., Haussler D.* New Methods for Detecting Lineage-Specific Selection / Eds. A. Apostolico et al. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. – P. 190–205. https://doi.org/10.1007/11732990_17.

41. *Kircher M., Witten D. M., Jain P. et al.* A general framework for estimating the relative pathogenicity of hu-

man genetic variants // Nat Genet. – 2014. – Vol. 46, № 3. – P. 310–315. https://doi.org/10.1038/ng.2892.

42. *Quang D., Chen Y., Xie X.* DANN: a deep learning approach for annotating the pathogenicity of genetic variants // Bioinformatics. – 2014. – Vol. 31, № 5. – P. 761–763. https://doi.org/10.1093/bioinformatics/btu703.

43. *Ionita-Laza I., McCallum K., Xu B. et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants // Nat Genet. – 2016. – Vol. 48, № 2. – P. 214–220. https://doi.org/10.1038/ng.3477.

44. *Shihab H.A., Rogers M.F., Gough J. et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation // Bioinformatics. – 2015. – Vol. 31, № 10. – P. 1536–1543. https://doi.org/10.1093/bioinformatics/btv009.

45. *Dong C., Wei P., Jian X. et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies // Human Molecular Genetics. – 2015. – Vol. 24, № 8. – P. 2125–2137. https://doi.org/10.1093/hmg/ddu733.

46. *Jagadeesh K. A., Wenger A. M., Berger M. J. et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity // Nat Genet. – 2016. – Vol. 48, № 12. – P. 1581–1586. https://doi.org/10.1038/ng.3703.

47. *Ioannidis N. M., Rothstein J. H., Pejaver V. et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants // The American Journal of Human Genetics. – 2016. – Vol. 99, № 4. – P. 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016.

48. *Koonin E. V.* Orthologs, Paralogs, and Evolutionary Genomics // Annu. Rev. Genet. – 2005. – Vol. 39, № 1. – P. 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725.

49. *Han J., Batey S., Nickson A. A. et al.* The folding and evolution of multidomain proteins // Nat Rev Mol Cell Biol. – 2007. – Vol. 8, № 4. – P. 319–330. https://doi.org/10.1038/nrm2144.

50. *Lipton Z. C.* The mythos of model interpretability // Queue. – 2020. – Vol. 16, № 3. – P. 31–57. https://doi.org/10.48550/arXiv.1606.03490.

REFERENCES

1. Cook C. E., Bergman M. T., Finn R. D. et al. The European Bioinformatics Institute in 2016: Data growth and integration // Nucleic Acids Res. 2015;44(D1):D20–D26. https://doi.org/10.1093%2Fnar%2Fgkv1352.

2. Gagan J., Van Allen E. M. Next-generation sequencing to guide cancer therapy // Genome Med. 2015;7(1):80. https://doi.org/10.1186/s13073-015-0203-x.

3. Ryzhkova O. P., Kardymon O. L., Prohorchuk E. B. et al. Guidelines for the interpretation of massive parallel sequencing variants (update 2018, v2) // Medical genetics. 2019;18(2):3–23. (In Russ.).

4. Spektor M. A., Yasko L. A., Druy A. E. The interpretation of somatic genetic variants identified with high-throughput sequencing of DNA from paediatric solid tumors // Medical Genetics. 2021;20(3):3–25. (In Russ.).

5. Cooper G. M., Stone E. A., Asimenos G. et al. Distribution and intensity of constraint in mammalian genomic sequence // Genome Res. 2005;15(7):901–913. https://doi.org/10.1101/gr.3577405.

6. Garber M., Guttman M., Clamp M. et al. Identifying novel constrained elements by exploiting biased substitution patterns // Bioinformatics. 2009;25(12):i54–i62. https://doi.org/10.1093%2Fbioinformatics%2Fbtp190.

7. Pollard K. S., Hubisz M. J., Rosenbloom K. R. et al. Detection of nonneutral substitution rates on mammalian phylogenies // Genome Res. 2009;20(1):110–121. https://doi.org/10.1101/gr.097857.109.

8. Siepel A., Bejerano G., Pedersen J. S. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes // Genome Res. 2005;15(8):1034–1050. https://doi.org/10.1101/gr.3715005.

9. Adzhubei I. A., Schmidt S., Peshkin L. et al. A method and server for predicting damaging missense mutations // Nat Methods. 2010;7(4):248–249. https://doi.org/10.1038/nmeth0410-248.

10. Ashkenazy H., Abadi S., Martz E. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules // Nucleic Acids Res. 2016;44(W1):W344–W350. https://doi.org/10.1093/nar/gkw408.

11. Calabrese R., Capriotti E., Fariselli P. et al. Functional annotations improve the predictive score of human disease-related mutations in proteins // Hum. Mutat. 2009;30(8):1237–1244. https://doi.org/10.1002/humu.21047.

12. Choi Y., Sims G. E., Murphy S. et al. Predicting the Functional Effect of Amino Acid Substitutions and Indels // PLoS ONE. 2012;7(10):e46688. https://doi.org/10.1371/journal.pone.0046688.

13. Frazer J., Notin P., Dias M. et al. Disease variant prediction with deep generative models of evolutionary data // Nature. 2021;599(7883):91–95. https://doi.org/10.1038/s41586-021-04043-8.

14. Reva B., Antipin Y., Sander C. Predicting the functional impact of protein mutations: application to cancer genomics // Nucleic Acids Research. 2011;39(17):e118–e118. https://doi.org/10.1093/nar/gkr407.

15. Sim N., Kumar P., Hu J. et al. SIFT web server: predicting effects of amino acid substitutions on proteins // Nucleic Acids Research. 2012;40(W1):W452–W457. https://doi.org/10.1093/nar/gks539.

16. Capriotti E., Fariselli P. Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants // Hum Genet. 2022;141(10):1649–1658. https://doi.org/10.1007/s00439-021-02419-4.

17. Adebali O., Reznik A. O., Ory D. S. et al. Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations // Genetics in Medicine. 2016;18(10):1029–1036. https://doi.org/10.1038/gim.2015.208.

18. Li W., Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences // Bioinformatics. 2006;22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

19. Ng P. C. SIFT: predicting amino acid changes that affect protein function // Nucleic Acids Research. 2003;31(13):3812–3814. https://doi.org/10.1093/nar/gkg509.

20. Thompson J. D., Higgins D. G., Gibson T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucl Acids Res. 2007;22(22):4673–4680. https://doi.org/10.1093/nar/22.22.4673.

21. Sievers F., Wilm A., Dineen D. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega // Molecular Systems Biology. 2011;7(1):539. https://doi.org/10.1038/msb.2011.75.

22. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform // Nucleic Acids Research. 2002;30(14):3059–3066. https://doi.org/10.1093/nar/gkf436.

23. Mount D. W. Using Iterative Methods for Global Multiple Sequence Alignment // Cold Spring Harb Protoc. 2009;2009(7):pdb.top44. https://doi.org/10.1101/pdb.top44.

24. Edgar R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity // BMC Bioinformatics. 2004;5(1):113. https://doi.org/10.1186/1471-2105-5-113.

25. Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment // Bioinformatics. 2002;15(3):211–218. https://doi.org/10.1093/bioinformatics/15.3.211.

26. Notredame C. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Research. 2002;24(8):1515–1524. https://doi.org/10.1093%2Fnar%2F24.8.1515.

27. Sievers F., Higgins D. G. QuanTest2: benchmarking multiple sequence alignments using secondary structure prediction // Bioinformatics. 2019;36(1):90–95. https://doi.org/10.1093/bioinformatics/btz552.

28. Ng P. C., Henikoff S. Predicting Deleterious Amino Acid Substitutions // Genome Res. 2002;11(5):863–874. https://doi.org/10.1101/gr.176601.

29. Shihab H. A., Gough J., Cooper D. N. et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models // Human Mutation. 2012;34(1):57–65. https://doi.org/10.1002/humu.22225.

30. Tang H., Thomas P. D. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation // Bioinformatics. 2016;32(14):2230–2232. https://doi.org/10.1093/bioinformatics/btw222.

31. Glaser F., Pupko T., Paz I. et al. ConSurf: Identification of Functional Regions in Proteinsby Surface-Mapping of Phylogenetic Information // Bioinformatics. 2002;19(1):163–164. https://doi.org/10.1093/bioinformatics/19.1.163.

32. Chun S., Fay J. C. Identification of deleterious mutations within three human genomes // Genome Res. 2009;19(9):1553–1561. https://doi.org/10.1101/gr.092619.109.

33. Stone E. A., Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity // Genome Res. 2005;15(7):978–986. https://doi.org/10.1101/gr.3804205.

34. Tavtigian S. V. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral // Journal of Medical Genetics. 2005;43(4):295–305. https://doi.org/10.1136/jmg.2005.033878.

35. Frazer J., Notin P., Dias M. et al. Disease variant prediction with deep generative models of evolutionary data // Nature. 2021;599(7883):91–95. https://doi.org/10.1038/s41586-021-04043-8.

36. Moles-Fernández A., Duran-Lozano L., Montalban G. et al. Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? // Front. Genet. 2018;9:366. https://doi.org/10.3389/fgene.2018.00366.

37. Mahmood K., Jung C., Philip G. et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics // Hum Genomics. 2017;11(1):10. https://doi.org/10.1186/s40246-017-0104-8.

38. Pollard K. S., Salama S. R., Lambert N. et al. An RNA gene expressed during cortical development evolved rapidly in humans // Nature. 2006;443(7108):167–172. https://doi.org/10.1038/nature05113.

39. Rao C. R. Score Test: Historical Review and Recent Developments / Eds. Balakrishnan N., Nagaraja H. N., Kannan N. Advances in Ranking and Selection, Multiple Comparisons, and Reliability. – Boston, MA: Birkhäuser; 2005. – P. 3–20. https://doi.org/http://dx.doi.org/10.1007/0-8176-4422-9_1.

40. Siepel A., Pollard K. S., Haussler D. New Methods for Detecting Lineage-Specific Selection / Eds. Apostolico A. et al. Research in Computational Molecular Biology. –Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. – P. 190–205. https://doi.org/10.1007/11732990_17.

41. Kircher M., Witten D. M., Jain P. et al. A general framework for estimating the relative pathogenicity of human genetic variants // Nat Genet. 2014;46(3):310–315. https://doi.org/10.1038/ng.2892.

42. Quang D., Chen Y., Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants // Bioinformatics. 2014;31(5):761–763. https://doi.org/10.1093/bioinformatics/btu703.

43. Ionita-Laza I., McCallum K., Xu B. et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants // Nat Genet. 2016;48(2):214–220. https://doi.org/10.1038/ng.3477.

44. Shihab H. A., Rogers M. F., Gough J. et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation // Bioinformatics. 2015;31(10):1536–1543. https://doi.org/10.1093/bioinformatics/btv009.

45. Dong C., Wei P., Jian X. et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies // Human Molecular Genetics. 2015;24(8):2125–2137. https://doi.org/10.1093/hmg/ddu733.

46. Jagadeesh K. A., Wenger A. M., Berger M. J. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity // Nat Genet. 2016;48(12):1581–1586. https://doi.org/10.1038/ng.3703.

47. Ioannidis N. M., Rothstein J. H., Pejaver V. et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants // The American Journal of Human Genetics. 2016;99(4):877–885. https://doi.org/10.1016/j.ajhg.2016.08.016.

48. Koonin E. V. Orthologs, Paralogs, and Evolutionary Genomics // Annu. Rev. Genet. 2005;39(1):309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725.

49. Han J., Batey S., Nickson A. A. et al. The folding and evolution of multidomain proteins // Nat Rev Mol Cell Biol. 2007;8(4):319–330. https://doi.org/10.1038/nrm2144.

50. Lipton Z. C. The mythos of model interpretability // Queue.2020;16(3):31–57. https://doi.org/10.48550/arXiv.1606.03490.

Информация об авторах

Буг Дмитрий Сергеевич, младший научный сотрудник, НИЦ биоинформатики НОИ биомедицины, Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова (Санкт-Петербург, Россия), ORCID: 0000-0002-5849-1311; **Наркевич Артем Николаевич**, доктор медицинских наук, доцент, профессор кафедры общественного здоровья и здравоохранения, Южно-Уральский государственный медицинский университет (г. Челябинск, Россия), ORCID: 0000-0002-1489-5058; **Петухова Наталья Витальевна**, кандидат медицинских наук, руководитель центра НИЦ биоинформатики НОИ биомедицины, Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова (Санкт-Петербург, Россия), ORCID: 0000-0001-6397-824X.

Information about authors

Bug Dmitrii S., Junior Research Fellow, Bioinformatics Research Center of the Research Institute of Biomedicine, Pavlov University (Saint Petersburg, Russia), ORCID: 0000-0002-5849-1311; **Narkevich Artem N.,** Dr. of Sci. (Med.), Associate Professor, Professor of the Department of Public Health, South-Ural State Medical University (Chelyabinsk, Russia), ORCID: 0000-0002-1489-5058; **Petukhova Natalia V.,** Cand. of Sci. (Med.), Head of the Bioinformatics Research Center of the Research Institute of Biomedicine, Pavlov University (Saint Petersburg, Russia), ORCID: 0000-0001-6397-824X.