УДК 577.112:577.29:004.9
https://doi.org/10.20538/1682-0363-2025-4-194-203

# Возможности анализа белков в биоинформационной системе NCBI

**Часовских Н.Ю.**

*Сибирский государственный медицинский университет (СибГМУ)*
*Россия, 634050, г. Томск, Московский тракт, 2*

**РЕЗЮМЕ**

**Цель** – рассмотреть и обобщить информацию об особенностях хранения данных о белках, а также о возможностях их анализа с помощью инструментов NCBI (National Center for Biotechnology Information, Национальный центр биотехнологической информации).

В лекции обобщены данные по существующим хранилищам белковых последовательностей и структур, проанализированы возможности биоинформационных инструментов для исследования белков на платформе NCBI. Первичные базы данных содержат информацию о белках (записи), полученную при проведении экспериментальных исследований. Помимо этого представлены базы с дополнительной информацией, добавленной кураторами после аналитики. Биоинформационный анализ белковых последовательностей и структур с помощью представленных в лекции инструментов позволяет выявить особенности филогенетического развития, спрогнозировать функции и структуры. Таким образом, извлечение обширной информации и возможность ее анализа с помощью специализированных сервисов помогает пролить свет при исследовании *in silico* на необнаруженные экспериментально характеристики белков, получить новые знания, служащие основой для дальнейших теоретических и экспериментальных исследований.

**Ключевые слова:** биоинформатика, последовательность белков, домен, выравнивание, трехмерная структура, NCBI

**Конфликт интересов.** Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

**Источник финансирования.** Автор заявляет об отсутствии финансирования при проведении исследования.

**Для цитирования:** Часовских Н.Ю. Возможности анализа белков в биоинформационной системе NCBI. *Бюллетень сибирской медицины.* 2025;24(4):194–203. https://doi.org/10.20538/1682-0363-2025-4-194-203.

# Protein analysis capabilities in the NCBI bioinformation system

**Chasovskikh N.Yu.**

*Siberian State Medical University (SibSMU)*
*2 Moskovsky trakt, 634050 Tomsk, Russia*

**ABSTRACT**

**Aim.** To review and summarize information about the features of protein data storage, as well as the possibilities for their analysis using NCBI tools.

The lecture summarizes data on existing repositories of protein sequences and structures, and analyzes the capabilities of bioinformatics tools for protein research on the NCBI platform (National Center for Biotechnology Information). The primary databases contain information about proteins (records) obtained through experimental

✉ *Часовских Наталия Юрьевна,* nch03@mail.ru

studies; in addition, databases with supplementary information added by curators after analysis are also presented. Furthermore, bioinformatic analysis of protein sequences and structures using the tools discussed in this lecture enables the identification of phylogenetic features, as well as the prediction of functions and structures. Thus, the extraction of extensive information and its analysis through specialized services facilitate insights into in silico research of experimentally undetected protein characteristics, providing new knowledge that forms the basis for further investigations.

**Keywords:** bioinformatics, protein sequence, domain, alignment, three-dimensional structure, NCBI

**Conflict of interest.** The author declares the absence of obvious and potential conflicts of interest associated with the publication of this article.

**Source of financing.** The author states that there was no funding for the study.

**For citation:** Chasovskikh N.Yu. Protein analysis capabilities in the ncbi bioinformation system. *Bulletin of Siberian Medicine.* 2025;24(4):194–203. https://doi.org/10.20538/1682-0363-2025-4-194-203.

## ВВЕДЕНИЕ

Анализ белковых последовательностей и протеомов является одной из важных областей биоинформационных исследований как способ реализации решений большого спектра медико-биологических задач. Ведущие ресурсы и платформы биоинформатики включают в себя инструменты для проведения данных исследований. В настоящей статье рассмотрены важнейшие из них, предлагаемые системой NCBI (National Center for Biotechnology Information, Национальный центр биотехнологической информации).

NCBI – информационно-поисковая интегрирующая система, содержащая большое количество баз данных и инструментов для биоинформационного анализа, в том числе и для белков. Они включают данные по последовательностям белков, структурам белковых молекул, инструменты для их сравнения и визуализации, а также инструменты и базы данных для анализа белковых доменов. В целом имеющиеся возможности системы реализуются через большой набор баз данных и биоинформационных сервисов, в настоящее время NCBI обеспечивает поиск и извлечение большей части данных из 31 отдельного репозитория и баз знаний [1, 2].

Одна из ключевых особенностей NCBI – наличие поисковой системы, которая позволяет обращаться к записям не только данной платформы, но и других хранилищ [3]. Доступ к записям – последовательностям белков важен для проведения таких основополагающих операций биоинформатики, как выравнивание последовательностей (парное и множественное) [4, 5]. Парное выравнивание – процесс сопоставления одной последовательности с другой (путем выстраивания их друг под другом) таким образом, чтобы достичь максимального сходства, множественное выравнивание – сопоставление трех и более последовательностей. Основная цель выравнивания – нахождение идентичных участков в разных последовательностях, подсчет оценки идентичности, что позволяет выявить гомологичные последовательности белков, отследить их эволюционные изменения, проанализировать функции на основе выявленного сходства [4, 6].

Анализ белковых последовательностей в контексте кластеров гомологов (ортологичных, родственных групп) важен для функционального и эволюционного анализа геномов. Чтобы извлечь максимальное количество информации из быстро накапливающихся записей последовательностей генома, все консервативные гены необходимо классифицировать по их гомологичным отношениям. Сравнение белков, закодированных в полных геномах определенных филогенетических линий, позволяет выявлять закономерности подобия последовательностей, выделять кластеры групп ортологов (Clusters of Orthologous Groups, COG). Каждый COG состоит из отдельных ортологичных белков (подобных белков в разных видах) или наборов паралогов (подобных белков внутри одного вида). Ортологи чаще всего выполняют одну и ту же функцию, что может обеспечить успешные функциональные прогнозы для недостаточно охарактеризованных геномов [7].

Поскольку идентификация функций белков использует данные о содержащихся доменах и мотивах, инструменты, позволяющие выявить их наличие, широко применяются в протеомике.

Работа с трехмерными структурами белков дает возможность изучать, моделировать молекулярные взаимодействия как для понимания механизмов происходящих в клетке процессов, так и для разработки лекарств. Инструменты для 3D-визуализации также описаны в данном сообщении.

Представленные в лекции примеры извлекаемых из биоинформационных репозиториев и инструментов данных рассмотрены для белков SARS-CoV-2 (Severe acute respiratory syndrome-related coronavirus

2, тяжелый острый респираторный синдром, связанный с коронавирусом 2).

В целом исследование и прогнозирование различных свойств белков требуют современных биоинформационных подходов и средств. Необходимо применять соответствующие конкретным исследовательским задачам базы данных и инструменты, ключевые характеристики которых описаны ниже.

## БАЗЫ ДАННЫХ NCBI, СОДЕРЖАЩИЕ ИНФОРМАЦИЮ О БЕЛКАХ

Основные базы данных NCBI, содержащие информацию о белках, включают BioProject, Conserved Domain Database (CDD), HIV-1, Human Protein Interaction Database, Identical Protein Groups, Protein Clusters, Protein Database, Protein Family Models, Reference Sequence (RefSeq) [1, 2]. Ниже будут рассмотрены примеры характерного для них извлечения информации на данных о белках SARS-CoV-2.

База данных BioProject (https://www.ncbi.nlm.nih.gov/bioproject) является организационным базисом для доступа к информации об исследовательских проектах со ссылками на сведения, которые депонируются в архивах, поддерживаемых членами Международного консорциума по базам данных нуклеотидных последовательностей [8]. Помимо данных по геномике и транскриптомике, в этой базе имеются записи белков, протеомов. Информация в хранилище представлена в виде набора связанных данных, т. е. «проекта». В BioProject выделяют два типа проектов: первичные и «зонтичные». Первичные – это впервые размещаемые (используя портал представления NCBI) данные, которые могут быть зарегистрированы лицами, представившими их. «Зонтичные» проекты с организационной структурой более высокого уровня для более крупных инициатив, которые обеспечивают дополнительный уровень отслеживания данных, такие проекты создаются по запросу [9]. В настоящее время проекты, связанные с исследованиями SARS-CoV-2, представлены в следующем количестве: «зонтичные» – 177 и впервые размещаемые – 3 639 (из них данные о белках содержат 185 проектов).

CDD (The Conserved Domain Database, база данных консервативных доменов, https://www.ncbi.nlm.nih.gov/cd) является ресурсом для аннотации функциональных модулей (т. е. доменов) в белках, его коллекция содержит набор курируемых NCBI данных (в том числе 3D-структур) [10]. CDD содержит хорошо аннотированные модели множественных выравниваний последовательностей для консервативных доменов и полноразмерных белков. Множественные выравнивания представляют собой профили расположенных друг под другом последовательностей. При этом гомологичные (подобные) участки выровнены в столбцах поперек длины последовательностей (пример приведен ниже). Предполагается, что сопоставляемые участки белков должны иметь общее происхождение, аналогичные функции и одинаково размещаться в пространстве. Эти модели используются для идентификации доменов в белковых последовательностях. Коллекции моделей выравнивания доменов очень важны для исследования эволюции белков, а также для аннотирования геномных последовательностей [11]. Так, при поиске в данной базе доменов по запросу SARS-CoV-2 пользователю выдается 81 результат, который включает семейства различных белков соответствующего вида. Один из этих результатов – ORF8-Ig_SARS-CoV-2-like (рис. 1) демонстрирует подсемейство, которое включает белок домена иммуноглобулина (Ig) ORF8 (SARS-CoV-2) и родственные белки ORF8 сарбековируса.

Результаты множественного выравнивания доменов данных белков демонстрируют высокую степень подобия и родство (рис. 2).

CDD также включает курируемые NCBI домены, которые используют информацию о 3D-структуре для определения границ доменов и предоставления информации о возможной связанности последовательности белка со структурой и его функцией [12]. Курирование данных о доменах позволяет пользователям получить представление о закономерностях сохранения консервативных остатков и дивергенции в ходе эволюции в семействах белков, их связи с функциональными свойствами. Чтобы обогатить традиционные множественные выравнивания последовательностей (а это основа моделей предметной области), в хранилище включают дополнительные типы информации:

– о трехмерных структурах и основных консервативных мотивах. Кураторы извлекают множественные выравнивания белков из внешних ресурсов, приводя их в соответствие с данными о трехмерных структурах и их суперпозиции (наложении в пространстве). В результате пользователям представляются выровненные блоки, включающие все строки множественного выравнивания без пробелов, и невыровненные области между ними. Указанные блоки демонстрируют консервативные основы соответствующего семейства доменов, а 3D-структуры могут быть интерактивно визуализированы с помощью инструмента Cn3D [11]. Так, для рассмотренного выше семейства SARS-CoV-2 ORF8 immunoglobulin (Ig) можно изучить трехмерную структуру белков в Cn3D, для чего необходимо выбрать данную опцию;

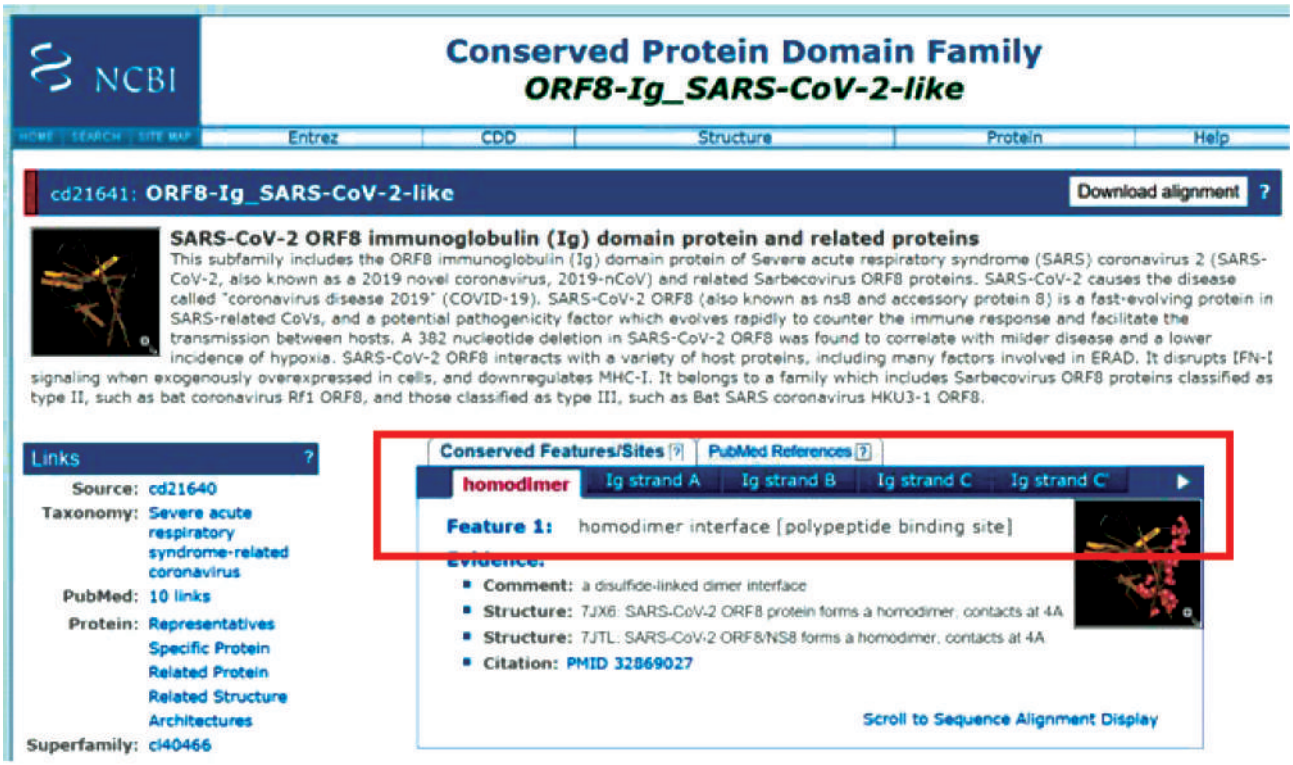


Рис. 1. Подсемейство, которое включает белок домена иммуноглобулина (Ig) ORF8 (SARS-CoV-2) и белки ORF8 сарбековируса: красным выделен раздел с данными (вкладками) о консервативных сайтах, найденных с помощью хранилища курируемых NCBI доменов (NCBI-curated domains)

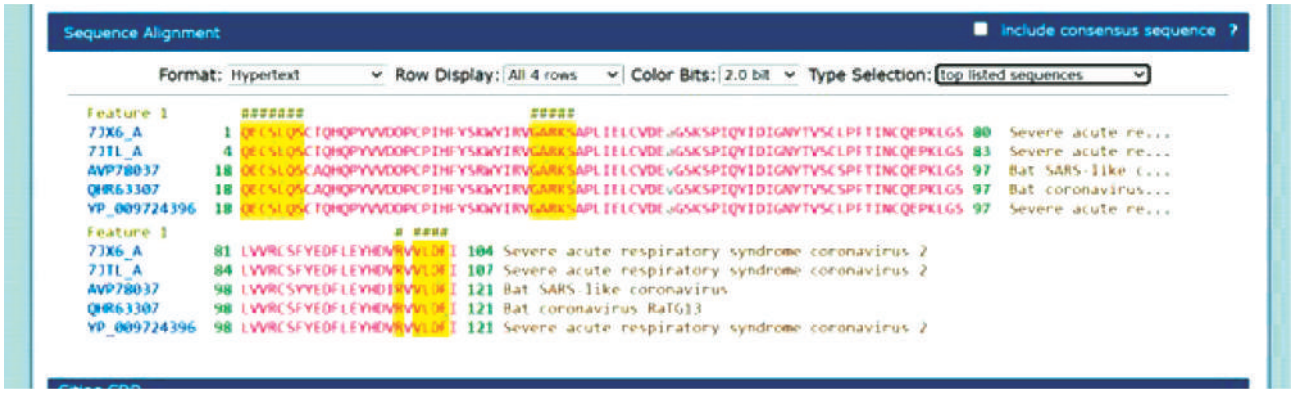


Рис. 2. Множественное выравнивание белков: представлены все остатки в каждой строке последовательности, при этом выровненные остатки отображаются прописными буквами, невыровненные – строчными. Горизонтальная шкала показывает количество остатков в общей последовательности. Цифры в начале и конце каждой строки последовательности указывают на диапазон элементов последовательности, которые были импортированы из полной записи последовательности белка

– о консервативных свойствах и (или) сайтах. Помимо множественного выравнивания белковых последовательностей, кураторы NCBI фиксируют, когда это возможно, расположение и свойства объектов в семействе доменов. При этом обычно описывают упоминаемые в литературе каталитические остатки, сайты связывания или мотивы. Также в базу добавляются функции, предположительно применимые к анализируемому семейству доменов, если найдены доказательства связи этих функций с набором элементов выравнивания белков семейства [13]. В рассматриваемом примере консервативные функции (или) сайты аннотированы в домене, курируемом NCBI, поэтому они отмечаются в поле сводки вверху страницы, с отдельной вкладкой для каждой функции (см. рис. 1);

– о филогенетической организации. Основываясь на данных сравнения последовательностей, кураторы объединяют модели связанных доменов в филогенетическую иерархию семейств. Последняя представляет собой набор родственных доменов, которые имеют общего предка, единый набор консервативных остатков и общую функцию. Однако при этом они имеют некоторые отличия в особенностях филогении, специфических функциях и дополнительных наборах консервативных остатков. Иерархии помогают получить представление о том, как закономерности сохранения остатков и дивергенции в семействе белков связаны с их функциональными свойствами [12];

– о ссылках на ресурсы электронной литературы. NCBI-curated domains также предоставляет активные ссылки на цитаты с информацией (если она существует) о домене (доказательства биологической функции, данные по эволюции и классификации) в PubMed (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#Link_cdd_pubmed) и NCBI Bookshelf (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#Link_cdd_books).

CDD также содержит данные, импортированные из ряда внешних баз данных (Pfam, SMART, COG, PRK, TIGRFAM) [10]. Так, текущая версия CDD, v3.20, содержит 59 693 белковые и белковые доменные модели, полученные из Pfam (19 178 количество моделей), SMART (1009), коллекции COGs (4871), $TIGRFAM_S$ (4488), коллекции белковых кластеров NCBI, NCBIfam (1125) и собственных результатов CDD по курированию данных (18 882) [10]. Хотя перечисленные внешние базы созданы для различных целей, рассматривают определенные подмножества пространства белков и различаются по размеру, в совокупности они позволяют осуществлять масштабный анализ доменов.

Наибольшую коллекцию множественных выравниваний для CDD представляет Pfam (http://pfam.sanger.ac.uk/), охватывая данные о распространенных белковых доменах и семействах, каждое из которых представлено множественным выравниванием последовательностей и скрытыми марковскими (вероятностными) моделями. Различные комбинации доменов обеспечивают разнообразие существующих белков, а идентификация доменов в белках позволяет получить представление об их функциях [14]. Так, для исследований возбудителя и механизма заболевания COVID-19, а также для определения вариантов лечения Pfam предоставляет полезную аннотацию для SARS-CoV-2, периодически обновляя модели, названия семейств и аннотации для этого вируса. К настоящему времени рассмотрены почти все генные продукты, кодируемые SARS-CoV-2. Orf10, небольшой белок, кодируемый на 3›-конце генома SARS-CoV-2, является единственным белком, который остается необозначенным Pfam [14].

Другой инструмент – SMART (http://smart.embl-heidelberg.de/). Помимо идентификации и аннотирования белковых доменов в исследуемых последовательностях, реализует функции сравнительного изучения сложных доменных архитектур в белках. Он содержит вручную курируемые модели для более чем 1300 доменов белков, семейства которых подробно аннотированы с точки зрения филогении, функциональных классов, 3D-структур и функционально важных остатков молекул [15].

COG (https://www.ncbi.nlm.nih.gov/research/cog-project/) – это ресурс классификации белков, также курируемый NCBI. Изначально проект был создан для обеспечения функциональной аннотации распространенных генов бактерий и архей, кластеризации их белковых продуктов по подобию последовательностей, отражающему их общее эволюционное происхождение. Включение в COG только генов из полностью секвенированных геномов позволяет точно идентифицировать ортологичные гены (или группы генов). COG предоставляет точную и актуальную аннотацию к наиболее распространенным семействам бактериальных и архейных белков, а также к малоизученным и неохарактеризованным белкам [16].

TIGRFAMs (https://www.ncbi.nlm.nih.gov/Structure/cdd/docs/tigrfams.html) представляет собой коллекцию вручную курируемых семейств белков с фокусом на последовательности прокариот, рассчитанную в первую очередь на исследователей, работающих в данной области [17].

Protein Clusters (гластеры белков, https://www.ncbi.nlm.nih.gov/proteinclusters) – это коллекция родственных белковых кластеров NCBI, состоящих из белков референсных последовательностей базы RefSeq (https://www.ncbi.nlm.nih.gov/RefSeq/), извлеченных из полных геномов, органелл и плазмид. Каждый кластер белков представлен списком идентификаторов белков и кодирующих их геномов. Protein Clusters на текущий момент ограничен данными архей, бактерий, растений, грибов, простейших и вирусов; включает в себя как курируемые, так и некурируемые (автоматически генерируемые) кластеры [18].

NCBI Virus (вирусы NCBI, https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) является интегрированным ресурсом, обеспечивающим оптимизированный поиск и анализ курируемых коллекций последовательностей вирусов и больших наборов данных из

GenBank и других хранилищ NCBI. В базе присутствует информация о различных вирусах: гепатита B и C, Денге, энтеровируса A, гриппа; для SARS-CoV-2 выделен отдельный репозиторий (SARS-CoV-2 Data Hub). В настоящее время NCBI Virus включает данные о последовательностях записей GenBank, RefSeq. Информация в репозитории включает аннотированные библиографии опубликованных отчетов о белковых взаимодействиях со ссылками на соответствующие записи последовательностей [19].

Identical Protein Groups (идентичные группы белков, https://www.ncbi.nlm.nih.gov/ipg) – коллекция консолидированных записей, описывающих белки, идентифицированные по аннотированным кодирующим регионам в базах данных GenBank и RefSeq, а также SwissProt и PDB. Этот ресурс позволяет исследователям получить более целенаправленные результаты поиска и быстро выявить интересующий белок. Обычно при поиске интересующего белка по базе Protein возникает сложность, связанная с большим количеством возвращаемых записей-результатов. Identical Protein Groups упрощает данный процесс, осуществляя поиск по группам записей, каждая из которых связана с одной уникальной последовательностью. Отдельная группа – это один результат, что позволяет формировать меньший набор найденных записей и быстрее идентифицировать интересующий белок [20].

Protein Database (база данных белков, https://www.ncbi.nlm.nih.gov/protein) включает записи белковой последовательности из различных источников, в том числе GenPept, RefSeq, Swiss-Prot, PIR, PRF и PDB [20]. Обширное хранилище дает возможность обращаться к коллекциям последовательностей из разных источников (в том числе к результатам трансляции из аннотируемых кодирующих регионов баз данных GenBank, RefSeq и TPA, к записям из других белковых баз данных SwissProt, PIR, PRF, PDB). Protein Database является важным ресурсом для работы с белками, поскольку последовательности белков – основа, определяющая их структуру и функции. Анализ последовательностей позволяет установить гомологию, определить филогенетические отношения, охарактеризовать функции и осуществить моделирование структур, а обширная база данных облегчает эти задачи. Так, большое количество белков можно извлечь из Protein Database для различных исследований SARS-CoV-2, например, в настоящий момент в базе содержится 1461 последовательность только белка Mpro SARS-CoV-2.

Protein Family Models (модели семейства белков, https://www.ncbi.nlm.nih.gov/protfam) содержит набор моделей, представляющих гомологичные белки с общей функцией. База включает в себя консервативные доменные архитектуры (CDD), скрытые модели Маркова и BlastRules [21]. Семейства, основанные на скрытых моделях Маркова, создаются преобразованием множественного выравнивания последовательностей с известной функцией и являются вероятностными моделями для определения принадлежности конкретного белка семейству. BlastRules представляет собой тип доказательства функциональной классификации белков на основе инструмента BLAST (рассмотрен в следующем разделе). BlastRules коллекционирует «модели» белков с известной биологической функцией, а BLAST помогает найти подобные определенной модели белки [21]. В частности, для белков SARS-CoV-2 репозиторий идентифицирует 47 моделей, из них скрытых моделей Маркова – 17, консервативных доменных архитектур – 10.

Reference Sequence (RefSeq, https://www.ncbi.nlm.nih.gov/RefSeq/) содержит полный, хорошо аннотированный набор последовательностей геномных ДНК, транскриптов (РНК) и белковых последовательностей NCBI, в связи с чем пользуется особой популярностью у исследователей. RefSeqs обеспечивает качественный информационный базис для различных исследований: аннотации генома, идентификации и характеристики генов, анализа мутаций и полиморфизма, исследований экспрессии и сравнительного анализа. К коллекции RefSeq можно получить доступ через базы данных нуклеотидов и протеинов [22, 23].

## БИОИНФОРМАЦИОННЫЕ ИНСТРУМЕНТЫ NCBI ДЛЯ АНАЛИЗА БЕЛКОВ

Basic Local Alignment Search Tool (BLAST, базовый инструмент поиска по локальному выравниванию, https://blast.ncbi.nlm.nih.gov/) находит области локального сходства у биологических последовательностей. Программа сравнивает нуклеотидную или белковую последовательность запроса с последовательностями баз данных (базы для поиска, как и его параметры, может задавать пользователь) и вычисляет статистическую значимость совпадений [24]. Инструмент BLAST позволяет находить регионы локального подобия последовательностей, что используется для анализа функциональных и эволюционных отношений между ними.

BLAST был создан в 1990 г. на основе метода k-кортежей и с тех пор был внедрен в GenBank, претерпевая многочисленные обновления для повышения эффективности и точности. Метод k-кортежей [5, 25, 26] – это быстрый эвристический метод попарного выравнивания, который обычно используется в каче-

стве начального шага при большом объеме выборки. Показатель сходства $S_{ij}$ между последовательностями *i* и *j* определяется как количество совпадений k-кортежей при наилучшем попарном выравнивании за вычетом фиксированного критерия штрафа за разрыв в последовательности. Для ДНК и РНК k обычно находится в диапазоне от 2 до 4, а для аминокислот k равно 1 или 2. Этот метод не гарантирует оптимального выравнивания, но является быстрым эвристическим методом и может быть использован для инициализации BLAST и выравнивания нескольких последовательностей. Программа BLAST сначала создает список слов из k-букв. Затем она выполняет поиск возможных совпадающих слов из k-букв в банке данных и оценивает их. Все слова, набравшие больше порогового значения, сохраняются, набравшие большее количество баллов, находятся в дереве поиска. Затем этот процесс распространяется на пары с высоким баллом (high-scoring pairs, HSP), которые также ищут похожие (а не только точно совпадающие) слова [27, 28]. Как базовый инструмент данный подход используется для обнаружения, идентификации или поиска похожих последовательностей в базе данных. Например, таким образом исследователями были обнаружены подобные коронавирусу последовательности у других организмов, таких как ящеры [29] и летучие мыши [30].

BLAST также использовался для обнаружения вируса SARS-CoV-2 в окружающей среде [31, 32], в том числе в сточных водах [33, 34]. В работе M. Parmar и соавт. [35] с помощью BLAST последовательности белка Mpro SARS-CoV-2 попарно сопоставлялись с другими последовательностями Mpro для определения возможной идентичности. Полученные результаты позволили оценить степень подобия Mpro SARS-CoV-2 с его ближайшими известными гомологами (SARS-CoV, MERS-CoV, Bat-CoV-RaTG13, HCoV-HKU, HCoV-OC43, HCoV-NL63 и HCoV-229E) для выявления консервативных сегментов. При данном анализе последовательностей удалось установить, что большинство (8/12) изменений остатков были обнаружены в доменах I и II Mpro β-цепей, где расположен ингибитор и (ИЛИ) каталитический участок; остальные (4/12) остатки были обнаружены в домене III [35]. В исследовании R. Naderi Beni и соавт. [36] для биоинформационного анализа структуры белка Mpro SARS-CoV-2 и его лигандов и ингибиторов также применялся BLAST, однако анализ осуществлялся по базе данных, содержащей 3D-структуры белков [36].

Помимо этого существуют специализированные варианты BLAST, предназначенные для решения более узких задач:

– SmartBLAST (https://blast.ncbi.nlm.nih.gov/smartblast/?LINK_LOC=BlastHomeLink) для поиска белков с высокой степенью подобия;

– IgBLAST (https://www.ncbi.nlm.nih.gov/igblast/) для поиска последовательностей иммуноглобулинов и рецепторов Т-клеток;

– CDART (https://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps) для поиска последовательностей с подобными архитектурами консервативных доменов.

Инструмент Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez) позволяет получать записи из многих баз данных NCBI, для чего необходимо загрузить файл с идентификаторами (индивидуальными номерами) последовательностей из соответствующих хранилищ. Результаты поиска – записи последовательностей, которые можно сохранить в файле для дальнейшей работы.

COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi) вычисляет множественные выравнивания последовательностей белков, используя информацию о консервативных доменах и локальном подобии последовательностей (на базе инструментов вышеупомянутого семейства BLAST) [37].

Cn3D (https://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) представляет собой автономное приложение – визуализатор для просмотра трехмерных структур NCBI, которое необходимо устанавливать на свой компьютер. Cn3D одновременно отображает разные данные: структуру, последовательность, множественное выравнивание последовательностей, и предлагает возможности для редактирования выравнивания. Помимо этого NCBI обеспечивает интерактивную трехмерную визуализацию макромолекул с помощью iCn3D (https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html), без инсталляции приложения. iCn3D позволяет визуализировать поверхности взаимодействий, сайты связывания, экспортировать модели для 3D-печати, выравнивать две структуры или две цепи, а также выравнивать (сопоставлять) последовательность белка со структурой [38].

Conserved Domain Search Service (CD Search, https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) используется для поиска белков, идентифицируя консервативные домены, присутствующие в последовательности. Если CD Search находит специфичное совпадение, существует высокая степень наличия ассоциации между последовательностью запроса и соответствующим консервативным доменом. В свою очередь, это может служить основой для понимания принадлежности выявленной функции белку запроса [10].

E-Utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/) – инструменты, которые обеспечивают доступ к данным в системе NCBI за пределами обычного интерфейса веб-запросов. Они предоставляют способ автоматизации задач в программных приложениях. Каждая утилита выполняет специализированную задачу поиска и может быть использована при написании специально отформатированного URL. E-utilities используют фиксированный синтаксис URL, который преобразует стандартный набор входных параметров в значения, необходимые для программных компонентов NCBI для поиска и извлечения запрошенных данных (последовательности нуклеотидов и белков, записи генов, трехмерные молекулярные структуры и литература) [39].

Инструмент ProSplign (https://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html) служит для выравнивания отдаленно родственных белков, обладающих небольшим сходством (используя данные о последовательностях геномных ДНК). Он основан на вариации алгоритма глобального выравнивания и, в частности, учитывает наличие интронов [2].

Рассмотренные инструменты, использующие методы и алгоритмы выравнивания последовательностей, широко применяются при анализе SARS-CoV-2, например для выявления мутаций и сравнения вирусных последовательностей у разных видов и организмов [40–42], для расшифровки механизмов передачи бессимптомной инфекции COVID-19 [43], для изучения влияния мутаций на диагностику и лечение данного заболевания [44], для идентификации и сопоставления вирусных последовательностей SARS-CoV-2 с другими коронавирусами животных, человека и родственными искусственными конструкциями [45, 46]. Приведенные данные демонстрируют, что выравнивание последовательности является незаменимым подходом для анализа и моделирования свойств белков [40–46].

## ЗАКЛЮЧЕНИЕ

Анализ последовательностей, эволюции, структуры и функций белков может быть более исчерпывающим и полным при использовании репозиториев и биоинформационных инструментов платформы NCBI. Благодаря предоставляемой информации, в том числе как результатов аналитики кураторов, можно прояснить строение белков интереса, их доменный состав, консервативность и дивергенцию в процессе видообразования, исследовать структуру и функции для большого спектра задач протеомики и не только.

Извлекаемые данные о белках NCBI включают записи о последовательностях, коллекции консервативных доменов и семейств белков, трехмерные структуры, информацию об исследовательских проектах и предназначены для решения разных задач. Объединяет их наличие перекрестных ссылок, обращение к внешним репозиториям, доступность для пользователей.

Инструменты NCBI призваны использовать алгоритмы биоинформатики для анализа данных о белках в представленных выше базах NCBI: для поиска подобных последовательностей белков, идентификации консервативных доменов, определения функций, структуры и визуализации информации.

Разнообразие сервисов позволяет применять исследователям достаточно широкий спектр подходов для анализа данных о белках. При этом постоянно совершенствующиеся версии инструментов, оптимизация алгоритмов, добавление новых разделов в хранилища данных делают их незаменимыми элементами процесса современных исследований.

## СПИСОК ИСТОЧНИКОВ


1. Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Chetvernin V. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2007;35:D5–D12. DOI: 10.1093/nar/gkl1031.
2. Sayers E.W., Beck J., Bolton E.E., Brister J.R., Chan J., Connor R. et al. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res*. 2025;53(D1):D20–D29. DOI: 10.1093/nar/gkae979.
3. Schuler G.D., Epstein J.A., Ohkawa H., Kans J.A. Entrez: molecular biology database and retrieval system. *Methods Enzymol*. 1996;266:141–162. DOI: 10.1016/s0076-6879(96)66012-1.
4. Часовских Н.Ю. Биоинформатика. М.: ГЭОТАР-Медиа, 2020:352. DOI: 10.33029/9704-5542-5-DIL-2020-1-352.
5. Mount D. Bioinformatics: sequence and genome analysis/ Cold Spring Harbor Laboratory Press: New York, 2004:692.
6. Polyanovsky V.O., Roytberg M.A., Tumanyan V.G. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms Mol. Biol*. 2011;6(1):25. DOI: 10.1186/1748-7188-6-25.
7. Tatusov R.L., Koonin E.V., Lipman D.J. A genomic perspective on protein families. *Science*. 1997;278(5338):631–637. DOI: 10.1126/science.278.5338.631. PMID: 9381173.
8. Karsch-Mizrachi I., Arita M., Burdett T., Cochrane G., Nakamura Y., Pruitt K.D. et al. The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res*. 2025;53(D1):D62–D66. DOI: 10.1093/nar/gkae1058.
9. Barrett T., Clark K., Gevorgyan R., Gorelenkov V., Gribov E., Karsch-Mizrachi I. et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res*. 2012;40:D57–D63. DOI: 10.1093/nar/gkr1163.
10. Wang J., Chitsaz F., Derbyshire M.K., Gonzales N.R., Gwadz M., Lu S. et al. The conserved domain database in 2023. *Nucleic Acids Res*. 2022;51(D1):D384–D388. DOI: 10.1093/nar/gkac1096.

11. Marchler-Bauer A., Panchenko A.R., Shoemaker B.A., Thiessen P.A., Geer L.Y., Bryant S.H. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*. 2002;30(1):281–283. DOI: 10.1093/nar/30.1.281.
12. Marchler-Bauer A., Anderson J.B., Derbyshire M.K., DeWeese-Scott C., Gonzales N.R., Gwadz M. et al. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res*. 2007;35:D237–240. DOI: 10.1093/nar/gkl951.
13. Marchler-Bauer A., Anderson J.B., Chitsaz F., Derbyshire M.K., DeWeese-Scott C., Fong J.H. et al. CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res*. 2009;37:D205–D210. DOI: 10.1093/nar/gkn845.
14. Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412–D419. DOI: 10.1093/nar/gkaa913.
15. Letunic I., Khedkar S., Bork P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res*. 2021;49(D1):D458–D460. DOI: 10.1093/nar/gkaa937.
16. Galperin M.Y., Vera Alvarez R., Karamycheva S., Makarova K.S., Wolf Y.I., Landsman D. COG database update 2024. *Nucleic Acids Res*. 2025;53(D1):D356–D363. DOI: 10.1093/nar/gkae983.
17. Haft D.H., Selengut J.D., Richter R.A., Harkins D., Basu M.K., Beck E. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res*. 2013;41:D387–D395. DOI: 10.1093/nar/gks1234.
18. Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Chetvernin V. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2008;36:D13–D 21. DOI: 10.1093/nar/gkm1000.
19. Brister J.R., Ako-Adjei D., Bao Y., Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res*. 2015;43:D571–D 577. DOI: 10.1093/nar/gku1207.
20. Entrez Sequences Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US) 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK44864/
21. Lu S., Wang J., Chitsaz F., Derbyshire M.K., Geer R.C., Gonzales N.R. et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48(D1):D265–D268. DOI: 10.1093/nar/gkz991.
22. Pruitt K., Brown G., Tatusova T., Maglott D. The Reference Sequence (RefSeq) Database. 2002 [Updated 2012]. In: McEntyre J., Ostell J., ed. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); Chapter 18. URL: https://www.ncbi.nlm.nih.gov/books/NBK21091/
23. Haft D.H., DiCuccio M., Badretdin A., Brover V., Chetvernin V., O'Neill K. et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res*. 2018;46(D1):D851–D860. DOI: 10.1093/nar/gkx1068.
24. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–3402. DOI: 10.1093/nar/25.17.3389.
25. Wilbur W.J., Lipman D.J. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*. 1983;80(3):726–730. DOI: 10.1073/pnas.80.3.726.
26. Rich D.H. Evaluation of enzyme inhibitors in drug discovery: a guide for medicinal chemists and pharmacologists. *Clin. Chem*. 2005;51:2219–2219. DOI: 10.1373/clinchem.2005.051946.
27. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol*. 1990;215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.
28. Ye J., McGinnis S., Madden T.L. BLAST: improvements for better sequence analysis. *Nucleic Acids Res*. 2006;34:W6–W9. DOI: 10.1093/nar/gkl164.
29. Xiao K., Zhai J., Feng Y., Zhou N., Zhang X., Zou J.-J. et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*. 2020;583:286. DOI: 10.1038/s41586-020-2313-x.
30. Wang H., Pipes L., Nielsen R. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. *Virus Evol*. 2021;7(1):veaa098. DOI: 10.1093/ve/veaa098.
31. La Rosa G., Mancini P., Bonanno F.G., Veneri C., Iaconelli M., Bonadonna L. et al. SARS-CoV-2 has been circulating in northern Italy since December 2019: Evidence from environmental monitoring. *Sci. Total Environ*. 2021;750:141711. DOI: 10.1016/J.SCITOTENV.2020.141711.
32. Sah R., Rodriguez-Morales A. J., Jha R., Chu D.K., Gu H., Peiris M. et al. Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. *Microbiol. Resour. Announc*. 2020;9:e00169–20. DOI: 10.1128/MRA.00169-20.
33. La Rosa G., Iaconelli M., Mancini P., Bonanno F.G., Veneri C., Bonadonna L. et al. First detection of SARS-CoV-2 in untreated wastewaters in Italy. *Sci. Total Environ.* 2020;736:139652. DOI: 10.1016/J.SCITOTENV.2020.139652.
34. Westhaus S., Weber F.-A., Schiwy S., Linnemann V., Brinkmann M., Widera M. et al. Detection of SARS-CoV-2 in raw and treated wastewater in Germany - Suitability for COVID-19 surveillance and potential transmission risks. *Sci. Total Environ.* 2021;751:141750. DOI: 10.1016/J.SCITOTENV.2020.141750.
35. Parmar M., Thumar R., Patel B., Athar M., Jha P.C., Patel D. Structural differences in 3C-like protease (Mpro) from SARS-CoV and SARS-CoV-2: molecular insights revealed by Molecular Dynamics Simulations. *Struct. Chem*. 2022:1–18. DOI: 10.1007/s11224-022-02089-6.
36. Naderi Beni R., Elyasi-Ebli P., Gharaghani S., Seyedarabi A. In silico studies of anti-oxidative and hot temperament-based phytochemicals as natural inhibitors of SARS-CoV-2 Mpro. *PLoS One*. 2023;18(11):e0295014. DOI: 10.1371/journal.pone.0295014.
37. Papadopoulos J.S., Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics.* 2007;23(9):1073–1079. DOI: 10.1093/bioinformatics/btm076.
38. Wang J., Youkharibache P., Zhang D., Lanczycki C.J., Geer R.C., Madej T. et al. iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *Bioinformatics*. 2020;36(1):131–135. DOI: 10.1093/bioinformatics/btz502.
39. Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US), 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK25501/

40. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*. 2020;112:3588–3596. DOI: 10.1016/j.ygeno.2020.04.016.
41. Li T., Liu D., Yang Y., Guo J., Feng Y., Zhang X. et al. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. *Sci. Rep*. 2020;10:1–9. DOI: 10.1038/s41598-020-79484-8.
42. Bianchi M., Borsetti A., Ciccozzi M., Pascarella S. SARS-Cov-2 ORF3a: Mutability and function. *Int. J. Biol. Macromol*. 2021;170:820–826. DOI: 10.1016/j.ijbiomac.2020.12.142.
43. Wang R., Chen J., Hozumi Y., Yin C., Wei G.-W. Decoding asymptomatic COVID-19 infection and transmission. *J. Phys. Chem. Lett*. 2020;11:10007–10015. DOI: 10.1021/acs.jpclett.0c02765.
44. Wang R., Hozumi Y., Yin C., Wei G.-W. Decoding SARS-CoV-2 Transmission and Evolution and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine. *J. Chem. Inf. Model.* 2020;60:5853. DOI: 10.1021/acs.jcim.0c00501.
45. Dallavilla T., Bertelli M., Morresi A., Bushati V., Stuppia L., Beccari T. et al. Bioinformatic analysis indicates that SARS-CoV-2 is unrelated to known artificial coronaviruses. *Eur. Rev. Med. Pharmacol Sci.* 2020;24:4558–4564. DOI: 10.26355/eurrev_202004_21041.
46. Trigueiro-Louro J., Correia V., Figueiredo-Nunes I., Gíria M., Rebelo-de-Andrade H. Unlocking COVID therapeutic targets: A structure-based rationale against SARS-CoV-2, SARS-CoV and MERS-CoV Spike. *Comput Struct Biotechnol J.* 2020;18:2117–2131. DOI: 10.1016/j.csbj.2020.07.017.

## Информация об авторе

**Часовских Наталия Юрьевна** – д-р мед. наук, доцент, зав. кафедрой медицинской и биологической кибернетики, СибГМУ, г. Томск, chasovskih.ny@ssmu.ru, https://orcid.org/0000-0001-6077-0347

(✉) **Часовских Наталия Юрьевна,** chasovskih.ny@ssmu.ru

Поступила в редакцию 07.05.2025;
одобрена после рецензирования 22.05.2025;
принята к публикации 29.05.2025