УДК 577.112:577.29:004.9
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 (National Center for Biotechnology Information) platform. The primary databases contain information about proteins (records) obtained through experimental studies; in addition, databases with supplementary information added by curators after analysis are also presented. Furthermore, bioinformatics 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.

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 филогенети-

✉ *Chasovskikh Natalia Yu.,* nch03@mail.ru

ческого 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.

INTRODUCTION

Analysis of protein sequences and proteomes is a key area of bioinformatics research, serving as a means to address a wide range of medical and biological problems. Leading bioinformatics resources and platforms include tools for conducting these studies. This article reviews the most important of these tools, provided by the National Center for Biotechnology Information (NCBI) System.

NCBI is an information retrieval and integration system that encompasses numerous databases and tools for bioinformatics analysis, including those for proteins. These include data on protein sequences, protein molecule structures, as well as tools for their comparison and visualization, along with tools and databases for protein domain analysis. Overall, the system's capabilities are realized through a comprehensive set of databases and bioinformatics services. Currently, NCBI provides search and data extraction from 31 separate repositories and knowledge bases [1, 2].

One of the key features of NCBI is its search system, which allows for access not only to records within this platform but also from other repositories [3]. Access to records – specifically protein sequences – is essential for performing fundamental bioinformatics operations, such as sequence alignment (pairwise and multiple) [4, 5]. Pairwise sequence alignment involves comparing one sequence with another (by lining them up one under the other) to achieve maximum similarity, while multiple alignment compares three or more sequences simultaneously. The primary goal of alignment is to identify identical regions across different sequences, calculate the identity score, and thereby facilitate the identification of homologous protein sequences, track their evolutionary changes, and analyze functions based on the observed similarities [4, 6].

Analysis of protein sequences within the context of homologous clusters (orthologous and related groups) is crucial for functional and evolutionary genome analysis. To maximize the information extracted from the rapidly accumulating number of genome sequence records, all conserved genes must be classified according to their homologous relationships. Comparing proteins encoded in complete genomes of certain phylogenetic lineages enables the identification of sequence similarity patterns and the delineation of Clusters of Orthologous Groups (COGs). Each COG comprises individual orthologous proteins (similar proteins across different species) or sets of paralogs (similar proteins within a single species). Orthologs most often perform equivalent functions, which can facilitate successful functional predictions for genomes that are poorly characterized [7].

Since the identification of protein functions uses the domains and motifs they contain, tools capable of detecting these features are widely used in proteomics.

Working with three-dimensional protein structures enables the study and modeling of molecular interactions, which is essential for understanding cellular processes and for drug development. Tools for 3D visualization are also described in this lecture.

The examples of data extracted from bioinformatics repositories and tools presented in the lecture focus on SARS-CoV-2 proteins (Severe Acute Respiratory Syndrome-related Coronavirus 2).

In general, the study and prediction of various protein properties require modern bioinformatics approaches and tools. It is necessary to utilize databases and tools tailored to specific research tasks, whose key characteristics are described below.

NCBI DATABASES CONTAINING PROTEIN INFORMATION

The main NCBI databases that include protein-related information are BioProject, Conserved Domain Database (CDD), HIV-1, Human Protein Interaction Database, Identical Protein Groups, Protein Clusters, Protein Database, Protein Family Models, and Reference Sequence (RefSeq) [1, 2]. Below, examples of their characteristic data extraction related to SARS-CoV-2 protein information will be discussed .

The BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) serves as an organizational platform for accessing information on research projects, with links to data deposited in archives maintained by members of the International Nucleotide Sequence Database Collaboration [8]. In addition to genomics and transcriptomics data, the database contains records of proteins and proteomes. Information within the repository is presented as a set of linked data, referred to as a "project". BioProject distinguishes between two types of projects: primary projects — data posted for the first time (using the NCBI submission portal) that can be registered by submitters; and umbrella projects with higher-level organizational structures for larger initiatives that provide an additional level of data tracking. These umbrella projects are created upon request [9]. Currently, regarding SARS-CoV-2 research, there are 177 umbrella projects and 3,639 primary projects posted for the first time (of which 185 contain data on proteins).

CDD (The Conserved Domain Database, CDD, https://www.ncbi.nlm.nih.gov/cd) is a resource for annotating functional modules (i.e., domains) in proteins. Its collection contains a set of NCBI-curated data, including 3D structures [10]. CDD provides well-annotated models of multiple sequence alignments for conserved domains and full-length proteins. These multiple alignments generate profiles of aligned sequences, with homologous (similar) regions arranged in columns across the length of the sequences (an example is given below). The aligned protein regions are expected to share a common origin, perform similar functions, and exhibit spatial similarity. These models are used to identify domains within protein sequences. Collections of domain alignment models are crucial for studying protein evolution and for annotating genomic sequences [11]. When searching this domain database with a query related to SARS-CoV-2, the user receives 81 results, which include families of different proteins from the corresponding species. One of these results – ORF8-Ig_SARS-CoV-2-like (Fig. 1) – represents a subfamily that includes the immunoglobulin (Ig) domain protein ORF8 (SARS-CoV-2) and related proteins from ORF8 sarbecoviruses.

The results of multiple domain alignment of these proteins demonstrate a high degree of similarity and relatedness (Fig. 2):

The CDD also includes NCBI-curated domains that utilize information about 3D structures to define domain boundaries and provide insights into the possible relationship between a protein sequence, its structure, and its function [12]. Domain curation enables users to gain insights into patterns of conservation of residues and divergence during evolution within protein families, as well as their relationship to functional properties. To enhance traditional multiple sequence alignments – the foundation of domain models – the repository incorporates additional types of information.

– Regarding 3D structures and major conserved motifs: curators extract multiple protein alignments from external resources, aligning them with 3D structural data and their superposition (spatial overlap). As a result, users are presented with aligned blocks that include all lines of the multiple alignment without gaps, along with unaligned regions between them. These blocks illustrate the conserved core regions of the corresponding domain family, and the 3D structures can be interactively visualized using the Cn3D tool [11]. Thus, for the SARS-CoV-2 ORF8 Ig family discussed above, the 3D structure of proteins can be studied in Cn3D by selecting this option.

– Regarding conserved features/sites. In addition to multiple sequence alignments of proteins, NCBI curators record locations and properties of entities within a domain family, when it is possible. This typically includes catalytic residues, binding sites, or motifs that are referenced in the literature.

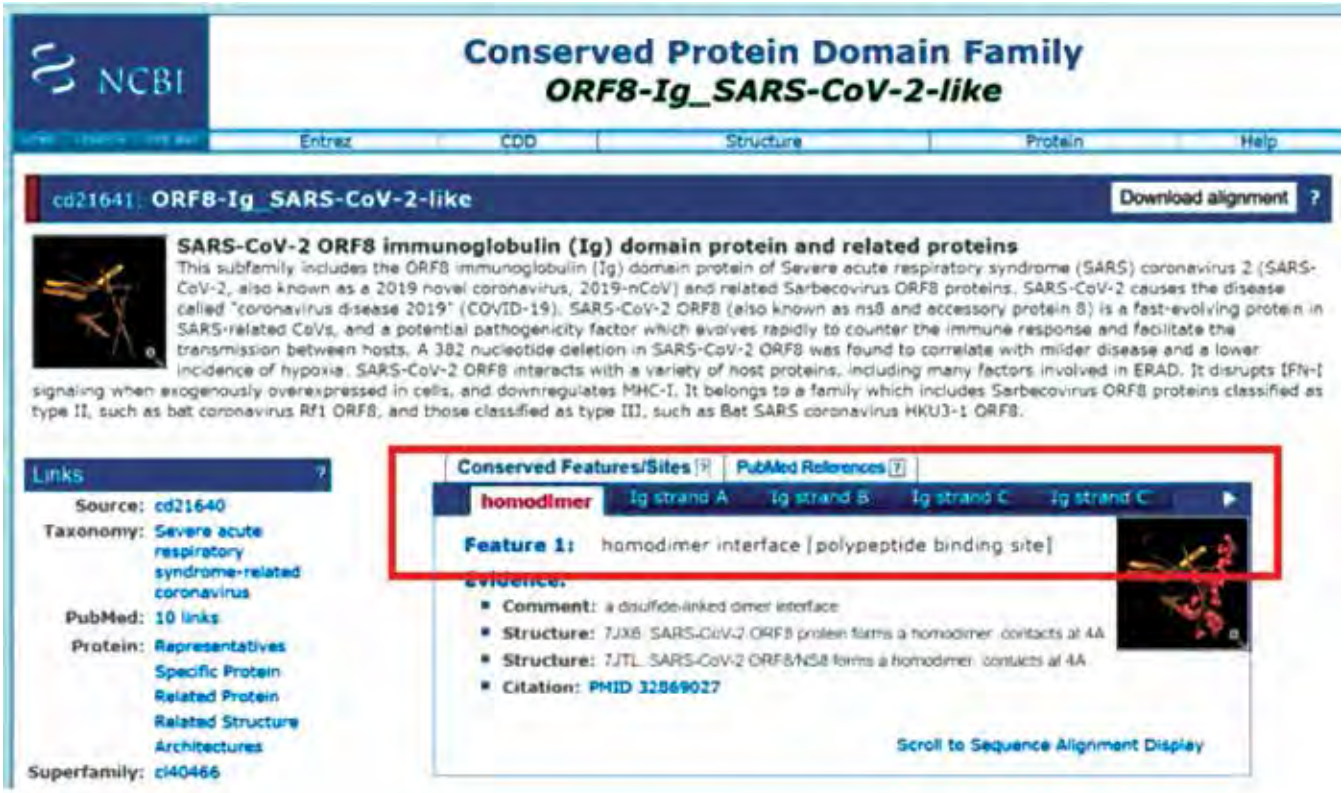

Fig. 1. The subfamily that includes the immunoglobulin (Ig) domain protein ORF8 (SARS-CoV-2) and the ORF8 proteins of sarbecoviruses. The section with data (tabs) on conserved sites identified using the NCBI-curated repository is highlighted in red.

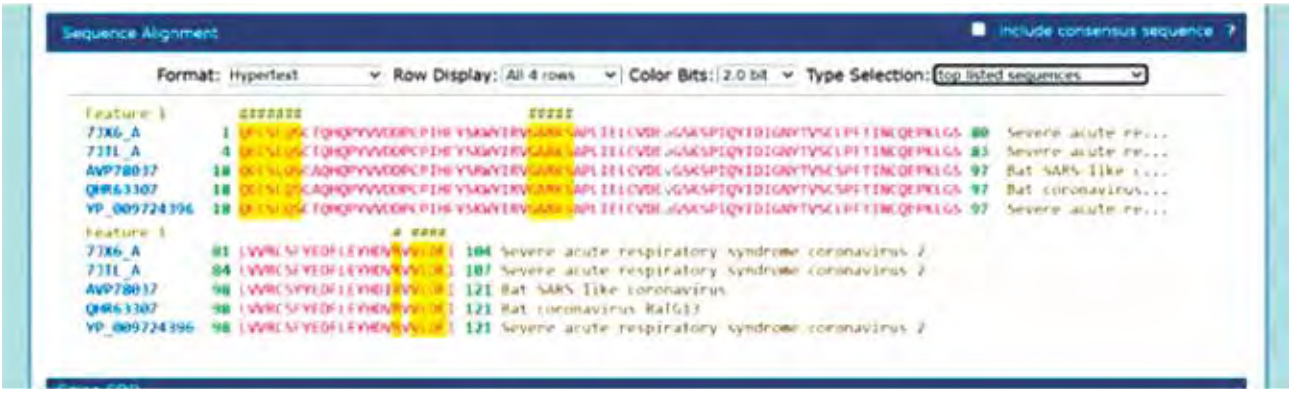

Fig. 2. Multiple protein alignment. All residues in each row of the sequence are shown, with aligned residues in uppercase and unaligned residues in lowercase. The horizontal scale indicates the number of residues in the overall sequence. The numbers at the beginning and end of each sequence row specify the range of sequence elements imported from the complete protein record

Functions that are potentially applicable to the domain family under analysis are also incorporated into the database if there is evidence linking these functions to a set of alignment elements within the family [13]. In this example, conserved features and sites are annotated within the NCBI-curated domain and are noted in the summary field at the top of the page, with a separate tab for each function (Fig. 1);

– Regarding phylogenetic organization: Based on sequence comparison data, curators organize models of related domains into a phylogenetic hierarchy of families. A domain family hierarchy consists of related domains sharing a common ancestor, a set of conserved residues, and a common function. However, these families may exhibit differences in phylogenetic features, specific functions, and additional conserved residues. Such hierarchies facilitate understanding how patterns of residue conservation and divergence within a protein family relate to their functional properties [12].

- Regarding links to electronic literature resources: NCBI-curated domains also provide active links to citations containing information — if available — about the domain's biological function, evolutionary context and classification data in PubMed (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#Link_cdd_pubmed) and NCBI Bookshelf (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#Link_cdd_books).

CDD also includes data imported from several external databases (Pfam, SMART, COG, PRK, TIGRFAM) [10]. As a result, the current version of CDD, v 3.20, contains 59,693 protein and protein domain models derived from Pfam (19,178 models), SMART (1,009), COGs collections (4,871), TIGFAMS (4,488), NCBI protein cluster collections, NCBIfam (1,125) and CDD's own data curation results (18,882) [10]. Although these external databases are designed for different purposes, they address specific subsets of the protein space and vary in size, together they facilitate large-scale domain analysis.

The largest collection of multiple alignments within CDD is Pfam (http://pfam.sanger.ac.uk/), which covers data on common protein domains and families. Each family is represented by multiple sequence alignments and hidden Markov models (HMMs). The diversity of existing proteins arises from various combinations of domains, and identifying these domains in proteins provides valuable insights into their functions [14]. For example, in research related to the pathogen and disease process of COVID-19, as well as in the search for treatment options, Pfam offers useful annotations for SARS-CoV-2. The models, family names, and annotations for this virus are periodically updated. Nearly all gene products encoded by SARS-CoV-2 have now been reviewed; however, Orf10 – a small protein encoded at the 3' end of the SARS-CoV-2 genome – remains unannotated by Pfam [14].

Another tool, SMART (http://smart.embl-heidelberg.de/), in addition to identifying and annotating protein domains in the studied sequences, offers functions for the comparative analysis of complex domain architectures in proteins. It contains manually curated models for more than 1,300 protein domains, with families that are thoroughly annotated in terms of phylogeny, functional classes, 3D structures, and functionally important residues of the molecules [15].

COG (https://www.ncbi.nlm.nih.gov/research/cog-project/) is a protein classification resource also curated by NCBI. The project was originally developed to provide functional annotation of common bacterial and archaeal genes, clustering their protein products based on sequence similarity, reflecting their shared evolutionary origin. By including only genes from fully sequenced genomes, COG enables accurate identification of orthologous genes (or gene groups). COG offers precise and up-to-date annotations of the most prevalent bacterial and archaeal protein families, as well as those that are poorly studied or uncharacterized [16].

TIGRFAMs (https://www.ncbi.nlm.nih.gov/Structure/cdd/docs/tigrfams.html) is a collection of manually curated protein families with a focus on prokaryotic sequences, primarily intended for researchers working in this field [17].

Protein Clusters (https://www.ncbi.nlm.nih.gov/proteinclusters) is a collection of related NCBI protein clusters composed of proteins from the RefSeq reference sequence database (https://www.ncbi.nlm.nih.gov/RefSeq/), extracted from complete genomes, organelles, and plasmids. Each protein cluster is represented by a list of protein identifiers and the genomes that encode them. Currently, Protein Clusters includes data from archaea, bacteria, plants, fungi, protozoa, and viruses; it encompasses both

curated and uncurated (automatically generated) clusters [18].

NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) is an integrated resource that offers optimized search and analysis of curated collections of virus sequences and large datasets from GenBank and other NCBI repositories. The database contains information on various viruses, including hepatitis B and C, dengue virus (DENV), enterovirus A, and influenza; a dedicated repository is allocated for SARS-CoV-2 (SARS-CoV-2 Data Hub). Currently, NCBI Virus includes sequence data from GenBank and RefSeq. The repository also features annotated bibliographies of published protein interaction reports, with links to relevant sequence records [19].

Identical Protein Groups (https://www.ncbi.nlm.nih.gov/ipg) is a collection of consolidated records describing proteins identified by annotated coding regions in GenBank and RefSeq databases, as well as SwissProt and PDB. This resource enables researchers to obtain more targeted search results and quickly identify a protein of interest. Typically, searching for a specific protein in the Protein database can be challenging due to the large number of records returned. Protein Groups simplifies this process by searching through groups of records, each associated with a single unique sequence. Each group is considered as a single hit, resulting in a smaller set of results and faster identification of the protein of interest [20].

The Protein Database (https://www.ncbi.nlm.nih.gov/protein) includes protein sequences from several sources, such as GenPept, RefSeq, SwissProt, PIR, PRF, and PDB [20]. This extensive repository provides access to sequence collections from different databases – including translation results from annotated coding regions of GenBank, RefSeq, and TPA – as well as records from other protein databases like SwissProt, PIR, PRF, and PDB. The Protein Database is an essential resource for working with proteins, since protein sequences form the basis for determining their structure and function. Sequence analysis allows for establishing homology, determining phylogenetic relationships, characterizing functions, and modeling structures. The large size of the database facilitates these tasks. For example, numerous proteins can be retrieved from the Protein Database for various studies of SARS-CoV-2, since currently, the database contains 1,461 SARS-CoV-2 Mpro protein sequences.

Protein Family Models (https://www.ncbi.nlm.nih.gov/protfam) contain a set of models representing homologous proteins with a common function. The database includes conserved domain databases (CDD), hidden Markov models, and BlastRules [21]. Families based on hidden Markov models are created by transforming multiple sequence alignments with known functions and serve as probabilistic frameworks for determining whether a particular protein belongs to a specific family. BlastRules are a type of evidence used for functional classification of proteins based on BLAST tool (Basic Local Alignment Search Tool, which is discussed in the next section). A BlastRule consists of one or more "model" proteins with known biological function, and BLAST helps identify proteins similar to a given "model" [21]. For SARS-CoV-2 proteins, the repository currently identifies 47 models, including 17 hidden Markov models and 10 conserved domain architectures.

The Reference Sequence database (RefSeq, https://www.ncbi.nlm.nih.gov/RefSeq/) contains a comprehensive, well-annotated collection of NCBI genomic DNA, transcript (RNA), and protein sequences, making it particularly popular among researchers. RefSeq provide a high-quality information base for various studies, including genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analysis. The RefSeq collection can be accessed through nucleotide and protein databases [22, 23].

NCBI BIOINFORMATICS TOOLS FOR PROTEIN ANALYSIS

Basic Local Alignment Search Tool (BLAST, https://blast.ncbi.nlm.nih.gov/) identifies regions of local similarity in biological sequences. The program compares a query nucleotide or protein sequence against sequences in databases (the search databases and parameters can be specified by the user) and calculates the statistical significance of the matches [24]. BLAST enables the identification of local sequence similarities, which are used to analyze functional and evolutionary relationships between sequences.

BLAST was developed in 1990 based on the k-tuple method, since then it has been integrated into

GenBank, undergoing numerous updates to enhance efficiency and accuracy. The word or k-tuple method [5, 25, 26] is a rapid heuristic pairwise alignment approach typically used as an initial step when dealing with large datasets. The similarity score S_{ij} between sequences *i* and *j* is defined as the number of matching k-tuples in the best pairwise alignment minus a fixed gap penalty. For DNA and RNA, *k* generally ranges from 2 to 4, while for amino acids, *k* is usually 1 or 2. Although this method does not guarantee an optimal alignment, it is a fast heuristic technique that can be used to initialize BLAST and facilitate multiple sequence alignments. The BLAST algorithm begins by creating a list of k-letter words from the query sequence. It then searches the database for potential matching k-letter words and scores them; all words scoring above a certain threshold are retained. Higher-scoring words are stored in a search tree. This process is extended to identify high-scoring pairs (HSPs), which involve searching for similar words (not necessarily exact matches) [27, 28]. As a fundamental tool, BLAST is used to detect, identify, or find similar sequences within a database. For example, researchers have identified coronavirus-like sequences in other organisms, such as pangolins [29] and bats [30]. BLAST has also been employed to detect SARS-CoV-2 in environmental samples [31, 32] , including wastewater [33, 34].

In work by M. Parmar et al. [35], pairwise BLAST comparisons of SARS-CoV-2 Mpro protein sequences with other Mpro sequences were performed to assess potential identity. The results allowed for evaluation of the similarity between SARS-CoV-2 Mpro and its closest known homologs (SARS- CoV , MERS- CoV , Bat-CoV-RaTG13, HCoV -HKU, HCoV-OC43, HCoV-NL63, and HCoV-229E) facilitating the identification of conserved regions. Sequence analysis revealed that most residue changes (8 out of 12) occurred in domains I and II of the Mpro β-chains – regions containing the catalytic/inhibitor region – while the remaining four residues were located in domain III [35]. In another study by R. Naderi Beni et al. [36], BLAST was also used for bioinformatic analysis of the SARS-CoV-2 Mpro structure and its ligands and inhibitors; however, this analysis utilized a database containing three-dimensional protein structures [36].

In addition, there are specialized versions of BLAST designed to address more specific issues:

SmartBLAST (https://blast.ncbi.nlm.nih.gov/smartblast/?LINK_LOC=BlastHomeLink) for searching proteins with a high degree of similarity;

IgBLAST (https://www.ncbi.nlm.nih.gov/igblast/) for searching immunoglobulin and T-cell receptor sequences,

CDART (https://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps) for identifying sequences with similar architectures of conserved domains.

Another tool is Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez), which allows users to retrieve records from multiple NCBI databases by uploading a file containing sequence identifiers (individual numbers) from the relevant repositories. The search results are sequence records that can be saved to a file for further analysis.

COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi) performs multiple alignments of protein sequences using information about conserved domains and local sequence similarity, based on tools from the BLAST family [37].

Cn3D (https://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) is a standalone visualization application for viewing NCBI 3D structures that must be installed on your computer. Cn3D simultaneously displays various data types: structure, sequence, multiple sequence alignment, and offers alignment editing capabilities. Additionally, NCBI provides an interactive 3D visualization tool for macromolecules called iCn3D (https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html), which does not require installation. It allows for visualization of interaction surfaces, binding sites, export of models for 3D printing, alignment of two structures or chains, and comparison of protein sequences with structures [38].

The Conserved Domain Search Service (CD Search, https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) is used to identify conserved domains within protein sequences. If CD Search detects a specific match, it indicates a high degree of association between the query sequence and the corresponding conserved domain. This information can serve as a basis for understanding the functional classification of the query protein [10].

E-Utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/) are tools that provide access to data within the NCBI system beyond the standard web query interface. They offer a means to automate

tasks within software applications. Each utility performs a specific search task and can be used by constructing a specially formatted URL. E-utilities utilize a fixed URL syntax that converts a standard set of input parameters into the values required by NCBI software components to search and retrieve the requested data (including nucleotide and protein sequences, gene transcripts, three-dimensional molecular structures, and literature) [39].

The ProSplign tool (https://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html) is used to align distantly related proteins with low sequence similarity, based on genomic DNA sequence data. It is built on a variation of the global alignment algorithm and specifically accounts for the presence of introns [2].

The tools described above, which employ sequence alignment methods and algorithms, are widely used in SARS-CoV-2 research. They are applied to identify mutations and compare viral sequences across different species and organisms [40, 41, 42], to elucidate mechanisms of transmission of asymptomatic COVID-19 infection [43], to study the impact of mutations on its diagnosis and treatment [44], and to compare SARS-CoV-2 sequences with other animal and human coronaviruses as well as related artificial constructs [45, 46]. The data presented demonstrate that sequence alignment is an essential approach for analyzing and modeling protein properties [40–46].

CONCLUSION

Analysis of protein sequences, evolution, structure, and functions can be more comprehensive and complete when utilizing the NCBI platform repositories and bioinformatics tools. The information provided, including the results of curatorial analysis, can help clarify the structure of proteins of interest, their domain composition, conservation, and divergence during speciation and facilitate the exploration of structure and function for a wide range of proteomics issues and beyond.

NCBI protein data mining encompasses sequence records, collections of conserved domains and protein families, 3D structures, and research project information, all designed to address diverse needs. However, all data share cross-referencing, access to external repositories, and user accessibility.

NCBI tools are designed to employ bioinformatics algorithms for analyzing protein data within the NCBI databases: to find similar protein sequences, identify conserved domains, determine function and structure, and visualize the information.

The variety of services enables researchers to apply a broad spectrum of approaches for protein data analysis. Continuous improvements in tool versions, algorithm optimization, and the addition of new sections to data repositories make these resources indispensable elements of modern research.

REFERENCES

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. Chasovskikh N.Yu. Bioinformatics. Moscow: GEOTAR-Media. 2020:352. (In Russ.). 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.

Author information

Chasovskikh Natalia Yu. – Dr. Sci. (Med.), Associate Professor, Head of the Medical and Biological Cybernetics Division, Siberian State Medical University, Tomsk, chasovskih.ny@ssmu.ru, https://orcid.org/0000-0001-6077-0347

(✉) **Chasovskikh Natalia Yu.,** nch03@mail.ru

Received in 07.05.2025;
approved after peer review in 22.05.2025;
accepted in 29.05.2025