Name:	Mr. Nayan Prabhakar Kasturi
Class:	M. Sc. Bioinformatics (Part I)
Roll Number:	110
Course:	M. Sc. Bioinformatics
Department:	Department of Bioinformatics
Paper:	Electives Paper I
Paper Name and Code:	Bioinformatics and Sequence Analysis (GNKPSBI3501)
Academic Year:	2023-24

SGCP's
Guru Nanak Khalsa College
of Arts, Science & Commerce (Autonomous)

DEPARTMENT OF BIOINFORMATICS

CERTIFICATE

This is to certify that Mr. Nayan Prabhakar Kasturi (Roll No: 110) of M.Sc. Bioinformatics (Part I) has satisfactorily completed the practical for Elective Paper 1: Bioinformatics and Sequence Analysis (GNKPSBI3501) for Semester I course prescribed by the University of Mumbai during the academic year 2023-2024.

TEACHER-IN-CHARGE	**HEAD OF THE DEPARTMENT**	**EXTERNAL EXAMINER**
(Mrs. Aparna Patil Kose)	(Dr. Gursimran Kaur Uppal)	

INDEX

	study of the query HBB Gene (ID: AY998983) in various file formats.			
3(D)	To explore the UniProt Database for further study of the query – thrombin protein (Accession ID – P25116).	08/09/23	87	
3(E)	To explore the PIR (Protein Information Resource) Database for the further study of the query casein (PRO ID – PR: 000028855) under various categories.	30/09/23	93	
4.	**Introduction to Domain Databases**	30/09/23	105	
4(A)	To study protein domain for query 'Lectin' (UniProt ID: Q9LW83) in PROSITE database.	30/09/23	108	
4(B)	To explore the InterPro database related to the protein family Amylase from organism Tetraodon nigroviridis (UniProt ID: CAD20312.1).	30/09/23	116	
5.	**Introduction to Structure Databases**	30/10/23	122	
5(A)	To study and explore the protein structure for the query 'Lysine' (PDB ID: 1OZV) using the Protein Data Bank (PDB) Database.	30/10/23	130	
5(B)	To explore the Nucleic Acid Knowledgebase (NAKB) / Nucleic Acid Database (NDB) for the study of the 3D structure of protein 'Helicase' (PDB ID: 8PJB).	30/10/23	140	
5(C)	To explore the CSDB for the query 'Antigen of Blood Group H2' (Compound ID: 13199).	30/10/23	146	
5(D)	To explore the Reactome pathway database with query 'Glycogenolysis pathway' (R-HSA-70221).	04/11/23	155	
5(E)	To explore and study the structure of the protein 'Tubulin' (PDB ID: 1TUB) under various categories using the structural database of PDBSum.	2/11/23	165	
5(F)	To explore the Protein Data Bank of Transmembrane Proteins (PDBTM) Database for studying transmembrane protein for the query of 'Aquaporin' (PDB ID: 1RC2).	2/11/23	172	
5(G)	To study the structural classification of proteins using CATH Database.	2/11/23	179	
5(H)	To study the structural classification of proteins using SCOPe Database.	2/11/23	188	

6.	**Introduction to Genome Browsers**	8/11/23	191	
6(A)	To explore the Genome Online Database (GOLD) in order to retrieve information about genome and metagenome subsequence.	8/11/23	199	
6(B)	To study various genomic databases like UCSC and ENSEMBL.	8/11/23	210	
6(C)	To explore the Microbial Genome Database for Comparative Analysis (MGBD) for query '*Escherichia coli*'.	8/11/23	230	
6(D)	To explore the International Committee on Taxonomy of Viruses Database (ICTVdb) using the query 'Measles Virus'.	8/11/23	235	
7.	**Introduction to Multiple Sequence Alignment**	6/11/23	239	
7(A)	To explore Multiple Sequence Alignment Tools, namely Clustal Omega, T-Coffee and MUSCLE for aligning 'cytochrome c oxidase subunit 1' protein sequences from five different species. The species used in this study and their UniProt IDs are as follows: *Homo sapiens* (UniProt ID: P00395), *Mus musculus* (UniProt ID: P00397), *Rattus norvegicus* (UniProt ID: P05503), Bos taurus (UniProt ID: P00396), *Ovis aries* (UniProt ID: 078749).	6/11/23	244	
8.	**Introduction to Restriction Enzyme Database (REBASE)**	10/11/23	258	
9.	**Introduction to Omics and Applications of Bioinformatics**	10/11/23	262	

DATE: 25/08/23

WEBLEM 1
INTRODUCTION TO NCBI AND USE OF FILTER OPTIONS
(URL: https://www.ncbi.nlm.nih.gov/)

INTRODUCTION:

Modern molecular biology is in a quest to unravel the elegant but silent language of living cells; the four alphabets representing the chemical subunits of DNA. The study of molecular biology is majorly focused on the syntax of life processes which emerges out of these four alphabets to form words and phrases. The staggering volume of molecular data and its cryptic and subtle patterns have led to an absolute requirement for computerized databases and analysis tools.

Late Senator Claude Pepper conducted a biomedical research and sponsored legislation that established the National Center for Biotechnology Information (NCBI) on November 4, 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), USA. NLM established an intramural research program in computational molecular biology. The collective research components of NIH make up the largest biomedical research facility in the world.

NCBI hosts approximately 40 online literature and molecular biology databases and is now a leading source for public biomedical databases, software tools for analyzing molecular and genomic data, and research in computational biology. There are over 3 million visitors daily to its website, approximately 27 terabytes of data downloaded per day, and the number of users as well as downloads increases dramatically each year.

The organizational structure in NCBI consists of three branches;

1. **Computational Biology Branch (CBB):**
   This branch conducts basic and applied research in computational, mathematical, and theoretical problems in molecular biology and genetics (genome analysis, sequence comparisons, etc.), establishes collaborative research projects in computational molecular biology with biologists, chemists, mathematicians, and computer scientists, consults and advises governmental agencies and research laboratories in the application of computer-based analytical tools, interacts with molecular biology groups to enhance laboratory-based research through the application of computational and theoretical approaches.
2. **Information Engineering Branch (IEB):**
   This branch performs applied research in data representation and analysis (develops systems for the storage, management, and retrieval of knowledge), designs database schema and specifications for representation of various forms of molecular biology information, coordinates public access to sequence, genetics, structural, and bibliographic information.
3. **Information Resources Branch (IRB):**
   This branch plans, directs, and manages the technical operations of NCBI, provides technical assistance to NCBI staff and provides support for external users of NCBI

network services, supervises network operations for the NCBI and coordinates with other government agencies.

To carry out its diverse responsibilities, NCBI:

1. Conducts research on fundamental biomedical problems at the molecular level using mathematical and computational methods.
2. Maintains collaborations with several NIH institutes, academia, industry, and other governmental agencies.
3. Fosters scientific communication by sponsoring meetings, workshops, and lecture series.
4. Supports training on basic and applied research in computational biology for postdoctoral fellows through the NIH Intramural Research Program.
5. Engages members of the international scientific community in informatics research and training through the Scientific Visitors Program.
6. Develops, distributes, supports, and coordinates access to a variety of databases and software for the scientific and medical communities.
7. Develops and promotes standards for databases, data deposition and exchange, and biological nomenclature.

Some of the major milestones from the many that have occurred over the past quarter of a century:

Year	Tool	About
1990	BLAST- Basic Local Alignment Search Tool	Optimized for speed, the sequence comparison algorithm quickly finds similar sequences to one's query.
1991	Entrez	The search and retrieval system for NCBI's linked databases, allowing users to easily find related information.
1992	GenBank	GenBank, a database of nucleotide sequences, and collaborates in its development with international partners at the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).
1994	NCBI Website	NCBI establishes its own website, mounting initially BLAST, Entrez, dbEST (Expressed Sequence Tags), and dbSTS (Sequence Tagged Sites).
1995	Genomes	This new resource organizes information on genomes, including sequences, maps, chromosomes, assemblies, and annotations.
1995	BankIt	The online tool is introduced to facilitate submissions to GenBank.
1996	OMIM- Online Mendelian Inheritance in Man	A directory of human genes and genetic disorders.
1997	PubMed	A freely accessible bibliographic retrieval system to the entire MEDLINE database.

1998	New NIH Disease-Based Services	Collaborations with NIH Institutes for Disease-Based Services are established such as CGAP (Cancer Genome Anatomy Project), to identify the human genes expressed in different cancerous states.
1999	Human Genome	First human chromosome sequence (chromosome 22) deposited by Human Genome Project researchers.
1999	Suite of Genomic Resources	Resources to support comprehensive analysis of the human genome, including: LocusLink—key descriptors of genetic loci; RefSeq—a non-redundant set of human reference sequences; and dbSNP—a collection of data on human genetic variation.
2000	PMC- PubMed Central	Fee full-text digital archive of biomedical and life sciences journal literature. Serves as an online counterpart to NLM's extensive print journal collection and is in keeping with the National Library's legislative mandate to collect and preserve the world's biomedical literature.
2000	GEO	The Gene Expression Omnibus database is launched in response to community interest in a public repository for data generated from high-throughput microarray experiments.
2001	Bookshelf	Entrez database to provide free access to books and documents in the life sciences and healthcare fields.
2002	WGS- Whole Genome Shotgun	GenBank begins including Whole Genome Shotgun sequences, which are generated by a semi-automatic technique.
2003	Entrez Gene	The Entrez Gene database (formerly known as LocusLink) is developed to supply key connections between maps, sequences, expression profiles, structure, function, homology data, and the scientific literature
2004	PubChem	Providing information on the chemical structure and biological activities of small molecules.
2006	dbGaP	Database of Genotypes and Phenotypes (dbGaP) to archive and distribute the results of studies that investigate the interaction of genotypes and phenotypes.
2007	Genome Reference Consortium	The Consortium of NCBI, EBI, Sanger Institute, and the Genome Institute is created to improve the sequence quality and accuracy of the human reference genome.
2010	dbVar	Archive of large-scale genomic variation data and associated defined variants with phenotypic information.
2013	ClinVar	Aggregates information about sequence variation and its relationship to human health.

REFERENCES:

1. Database resources of the National Center for Biotechnology Information. (2012). Nucleic Acids Research, 41(D1), D8–D20. https://doi.org/10.1093/nar/gks1189
2. Organizational Structure - NCBI. (n.d.). https://www.ncbi.nlm.nih.gov/home/about/structure/
3. Our mission - NCBI. (n.d.). https://www.ncbi.nlm.nih.gov/home/about/mission/
4. Smith, K. (2013). A brief history of NCBI's formation and growth. The NCBI Handbook - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK148949/

DATE: 25/08/2023

WEBLEM 1(A)
BASIC, LIMIT AND ADVANCED SEARCH USING NCBI DATABASE
(URL: https://www.ncbi.nlm.nih.gov/)

AIM:
To study query, 'Chitosan' in NCBI database and filter results using BASIC, LIMIT and ADVANCE search.

INTRODUCTION:
NCBI is a digital archive that stores and organizes material so that it can be easily found using a number of search criteria. A record, also known as an entry, has a number of fields that contain the actual data elements. A user can provide a specific piece of information, termed a value, to be found in a specific field and expect the computer to retrieve the entire data record to fetch a specific record from the database. This is known as making a query. Entrez, a biological database retrieval system, was created and is maintained by the NCBI. It is a text-based search engine for a wide range of data, including annotated genetic sequence information, structural information, citations and abstracts, entire articles, and taxonomy data. The ability of Entrez to integrate information, which results from cross-referencing different NCBI databases based on pre-existing and logical links between individual entries, is its defining feature. This is really convenient because users do not have to visit various databases located in different locations.
Firing a query on Entrez gives several results within several databases within NCBI; this is known as a 'Basic Search'. Although the effective use of Entrez requires an understanding of the main features of the search engine and narrowing down the search results; this is known as a 'Limit Search'. There are several options common to all NCBI databases that help to narrow the search which restrict the search to a subset of a particular database. One option being 'Limits' restricts the search to a particular database (e.g., the field for author or publication date) or a particular type of data (e.g., chloroplast DNA/RNA). Another option is 'Preview/Index,' which connects different searches with the Boolean operators and uses a string of logically connected keywords to perform a new search. The search can also be limited to a particular search field (e.g., gene name or accession number). Further, 'Advance Search Builder' can also be used which gives an even more structured search. There is a limit of five filters (including custom filters) that can be selected for all NCBI databases, except for PubMed, where the maximum number of filters allowed is 15.

Chitosan:
Chitosan is a β- 1-4-2-acetamido-2-deoxy-D-glucose molecule and is mainly known as partially deacetylated derivative of chitin and is reported to be bioactive. Chitosan is more soluble in water and organic acids than chitin and as such is easier to process. Due to this, it could be successfully used to increase the growth of crops and vegetables as well as to maintain the quality of harvested fruits and vegetables. It has applications in many industries like water and wastewater treatment, food and beverages, chemicals, feed, and cosmetics due to their

versatile nature. Information regarding the compound chitosan is present on many databases, like NCBI (National Center for Biotechnology Information).

Chitin is the most abundant polysaccharide in the marine ecosystem but the second most abundant polysaccharide in nature next to cellulose. It is found in crustaceans (shrimps, crabs), molluscs (shell oysters), insects(ladybug) and fungi (*Mucor rouxii*). Chitin is converted into its deacetylated form, i.e., Chitosan. Chitin and chitosan both have enormous economic value because of their biological properties. Crystallinity and insolubility of chitin demote its commercial applications.

In this weblem, the query 'Chitosan' will be investigated in the NCBI (the National Center for Biotechnology Information) database using basic search and limiting filters for appropriate results.

METHODOLOGY:

1. Go to the NCBI website.
2. NCBI Homepage appears with Entrez search engine and an 'All Databases' default basic search.
3. Enter the query, 'Chitosan' and click on search. An overview of the number of entries related to the query will be listed in 24 databases.
4. Limit the search using the 'Books' filter on the drop-down menu of the Entrez search engine. The search gives the results for the books present in the literature resources of NCBI.
5. Add more filters to the book search to get the latest publication on the topic by changing the 'Publication' filter to 1year. The results from this will give the books published in the past year on the topic queried.

OBSERVATIONS:

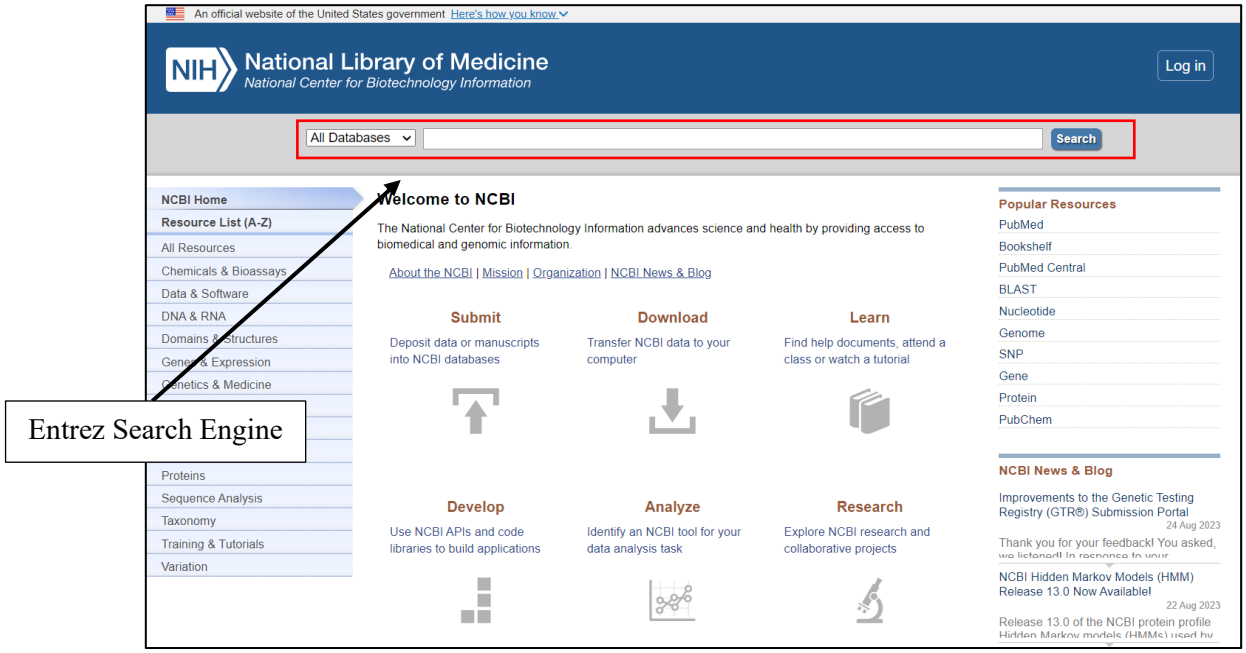


Figure 1: NCBI Homepage

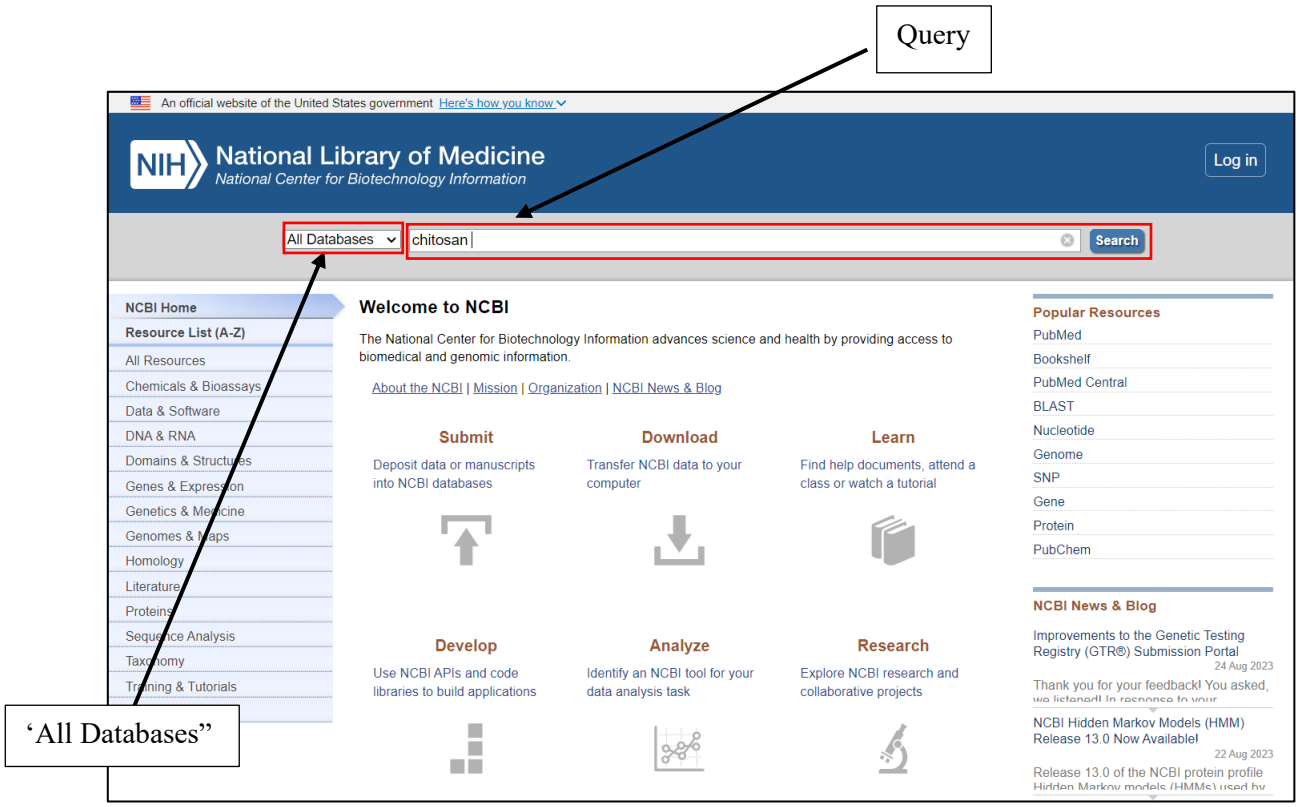


Figure 2: 'Chitosan' query with default 'All Databases'

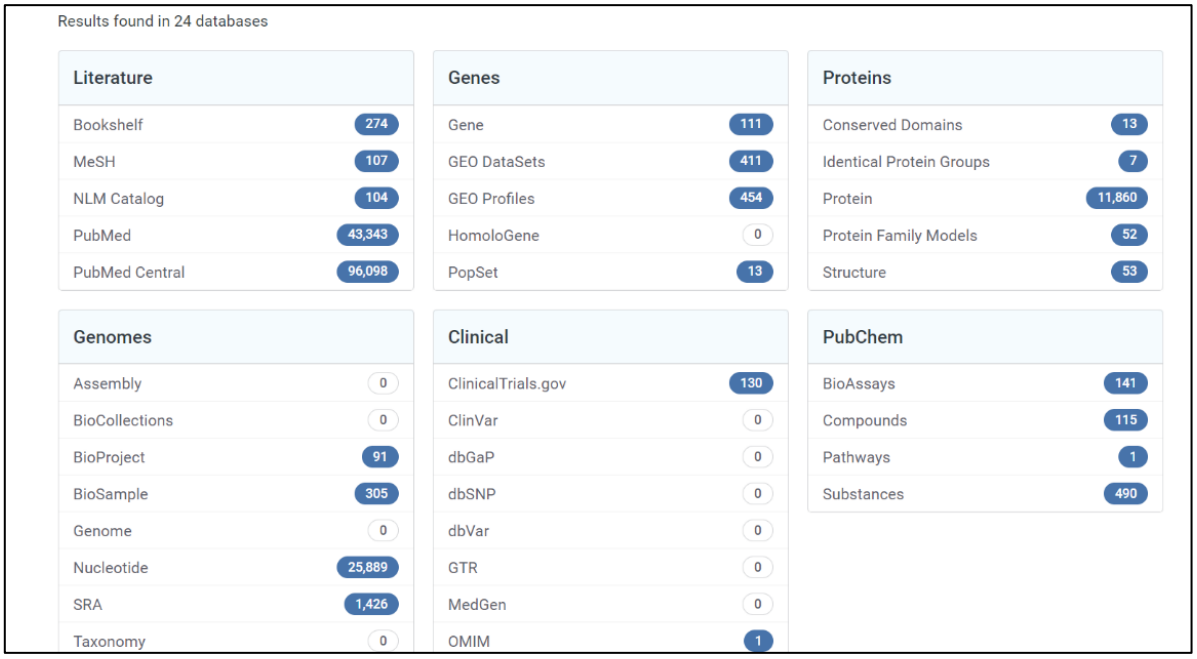


Figure 3: Entrez Search Result with entry mentioned in 24 databases.

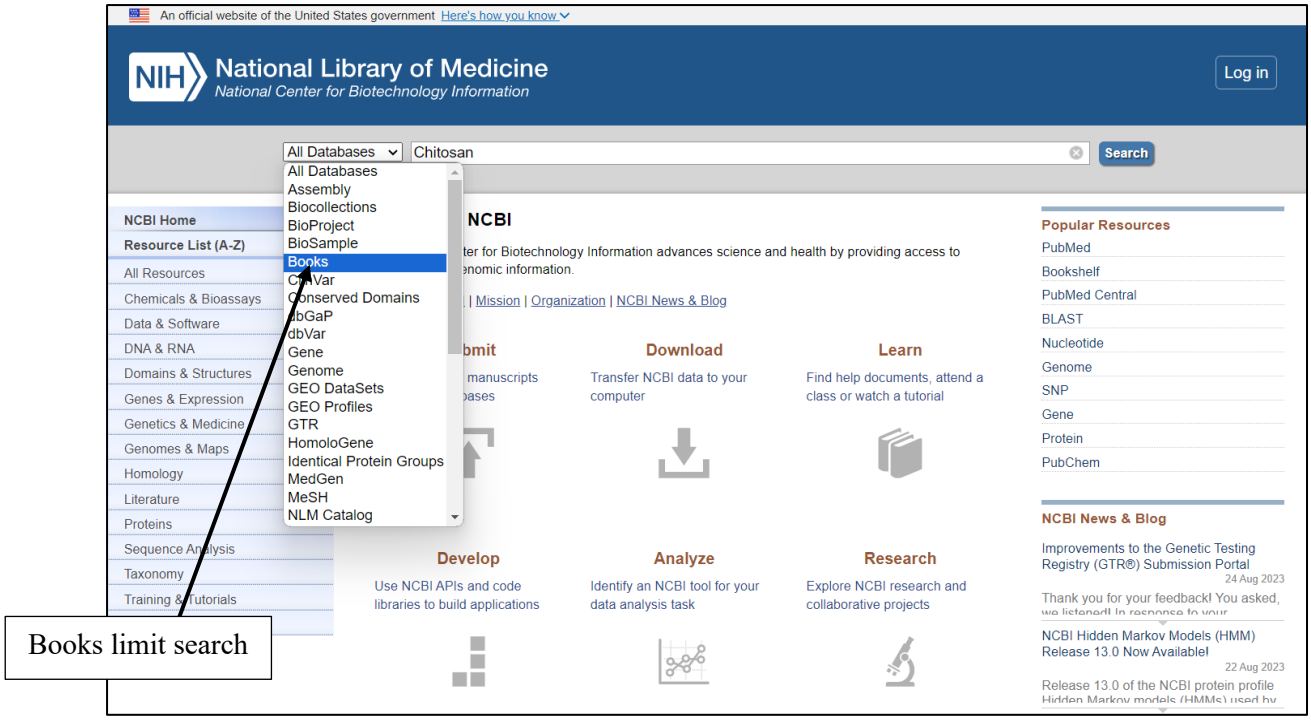


Figure 4: Adding limit search- 'Books' to the Entrez search engine.

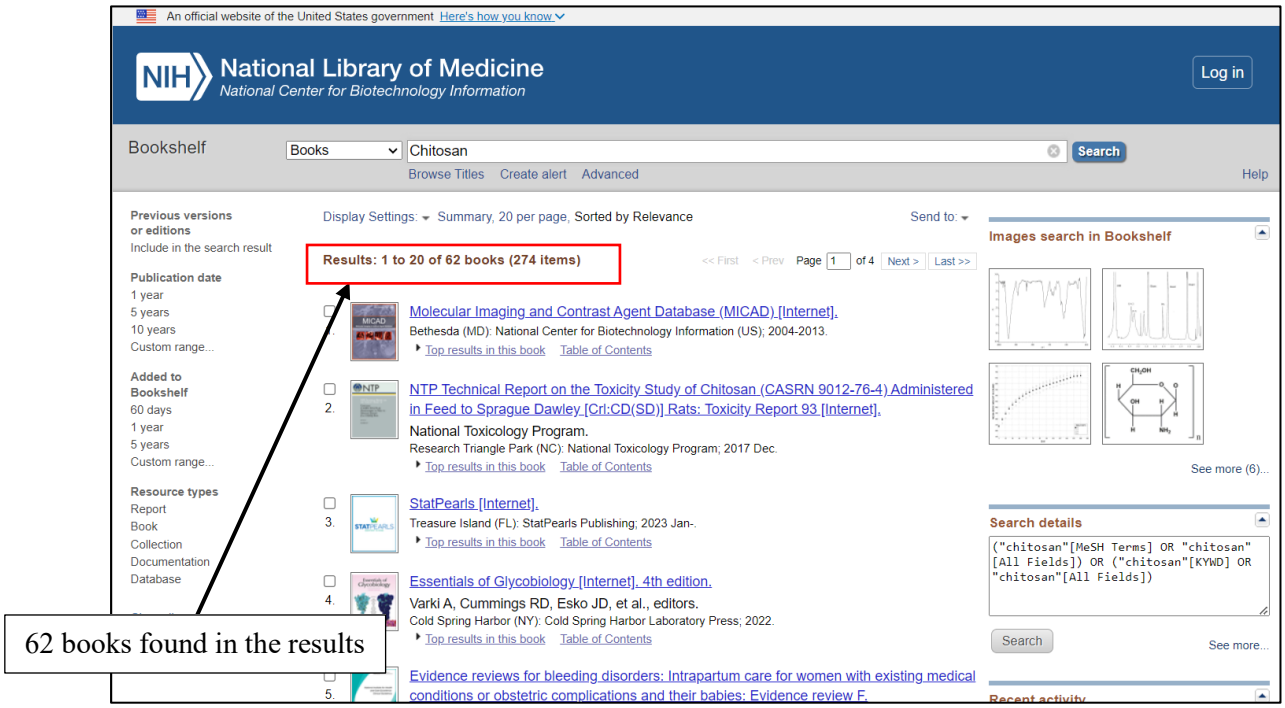


Figure 5: Basic Search result of the query 'Chitosan' in Books (limit filter).

Figure 6: Publication date limit filter results

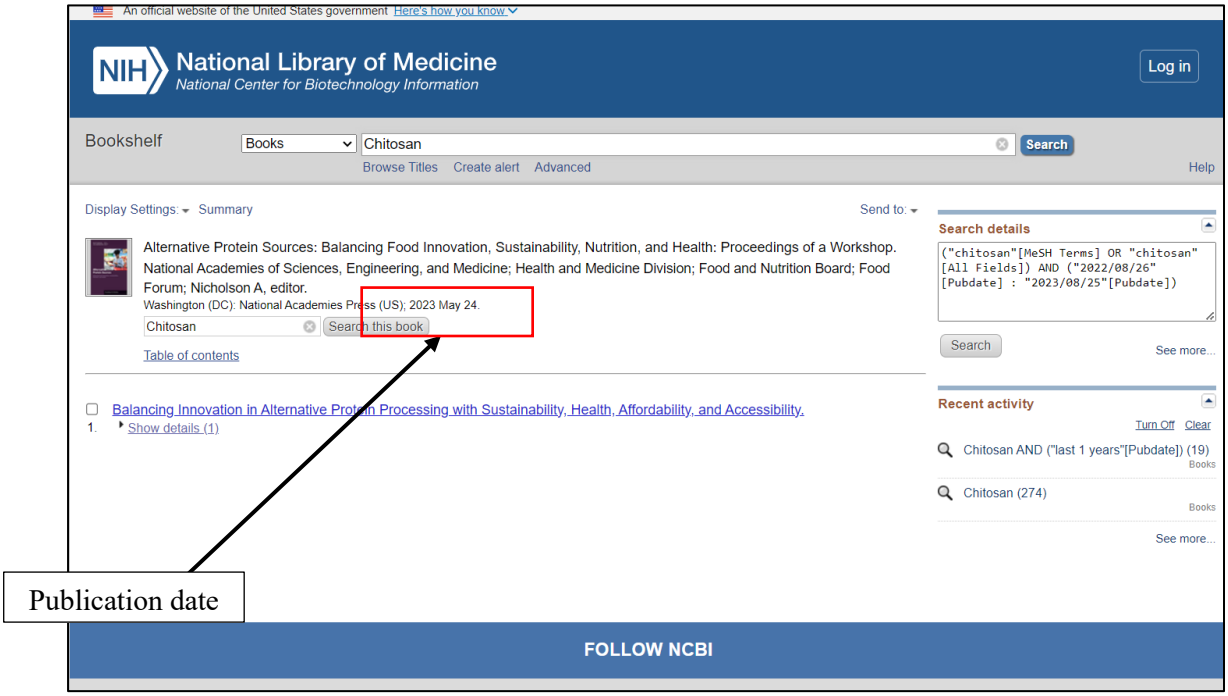


Figure 7: Most recent book published in the past year.

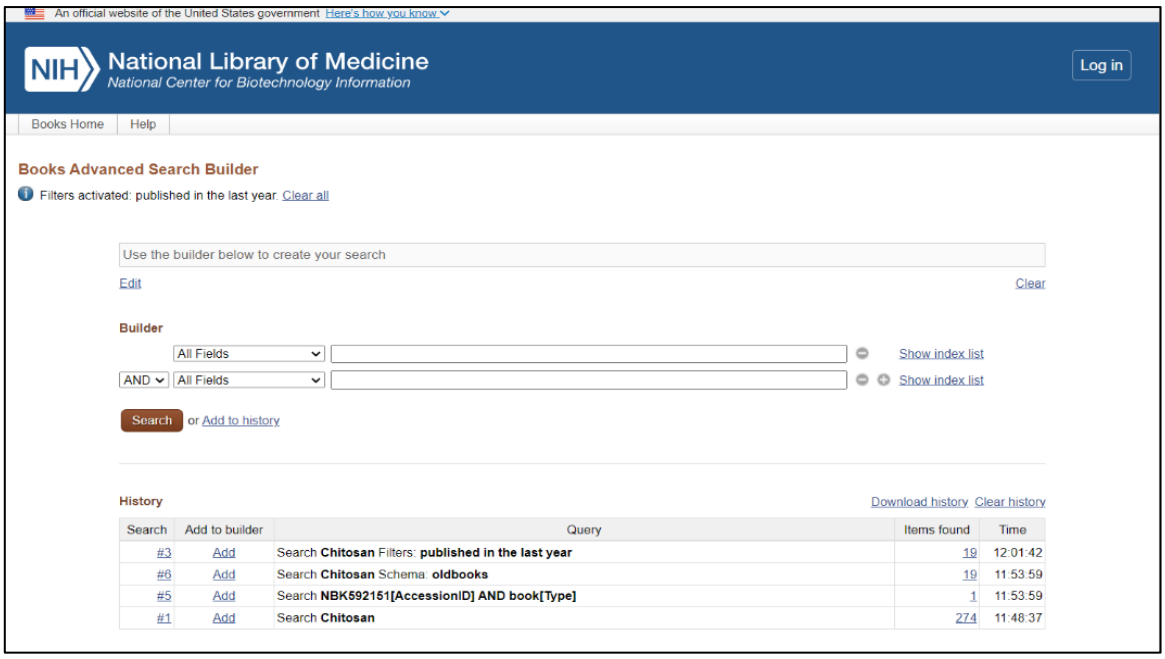


Figure 8: Advance Search filter

RESULTS:

Basic search for the query 'Chitosan' on the entrez search engine gave results as an overview of the number of entries related to the query in 24 databases. After using the limit filter, 'Books' on the search, a result of 62 books related to the query were observed. Results were further narrowed down to 4 books after adding filter to the publication date (1year). The most recent publication on the query was found.

CONCLUSION:

NCBI was searched for the query 'Chitosan' through basic and limit search to find the most recent publication as well as advanced search tool was explored.

REFERENCES:

1. Hoqani, H. a. S. A., Al-Shaqsi, N., Hossain, M. A., and Sibani, M. a. A. (2020). Isolation and optimization of the method for industrial production of chitin and chitosan from Omani shrimp shell. Carbohydrate Research, 492, 108001. https://doi.org/10.1016/j.carres.2020.108001
2. National Center for Biotechnology Information (US). (2021, June 28). Working with Filters. My NCBI Help - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK53591
3. Xiong, J. (2006). Essential bioinformatics. Cambridge University Press.
4. Yadav, M., Goswami, P., Paritosh, K., Kumar, M., Pareek, N., and Vivekanand, V. (2019). Seafood waste: a source for preparation of commercially employable chitin/chitosan materials. Bioresources and Bioprocessing, 6(1). https://doi.org/10.1186/s40643-019-0243-y

DATE: 25/08/23

WEBLEM 2
INTRODUCTION TO SPECIALIZED DATABASE

INTRODUCTION:

Specialized databases normally serve a specific research community or focus on a particular organism. The content of these databases may be sequences or other types of information. The sequences in these databases may overlap with a primary database, but may also have new data submitted directly by authors. Because they are often curated by experts in the field, they may have unique organizations and additional annotations associated with the sequences. Many genome databases that are taxonomic specific fall within this category. Examples include OMIM, KEGG, Flybase, WormBase, AceDB, and TAIR. In addition, there are also specialized databases that contain original data derived from functional analysis. For example, GenBank EST database and Microarray Gene Expression Database at the European Bioinformatics Institute (EBI) are some of the gene expression databases available.

A Specialized databases is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. Data derived from the results of analyzing Literature & specialized database like NCBI Literature database, PMC, PubMed, often draw upon information from numerous sources, including other databases-controlled vocabularies and the scientific literature.

They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science. The amount of computational processing work, however, varies greatly among the secondary databases, some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

NCBI Literature Database:

The NCBI database, operated by the National Center for Biotechnology Information, is a pivotal resource in the realm of biological and genetic information. Housing diverse databases like GenBank, PubMed, and others, NCBI serves as a comprehensive repository for genomic data, scientific literature, and bioinformatics tools. GenBank, within NCBI, catalogues genetic sequences, fostering global collaboration and knowledge exchange among researchers. PubMed, another integral part, offers an extensive collection of biomedical literature, empowering scientists and healthcare professionals to access a wealth of peer-reviewed articles. NCBI's commitment to organizing, maintaining, and disseminating biological information makes it an indispensable hub for advancing research and understanding in the fields of molecular biology and genomics.

1. **PubMed Database:**

PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics The United States National Library of Medicine (NLM) the National Institutes of Health maintain the database as part of the Entrez system of information retrieval

From 1971 to 1997, online access to the MEDLINE database had been primarily through institutional facilities, such as university libraries PubMed, first released in January 1996. ushered in the era of private, free, home- and office-based MEDLINE searching The PubMed system was offered free to the public starting in June 1997

In addition to MEDLINE, PubMed also provides access to

a. Older references from the print version of Index Medicus, back to 1951 and earlier
b. References to some journals before they were indexed in Index Medicus and MEDLINE, for instance Science, BMI, and Annals of Surgery
c. Very recent entries to records for an article before it is indexed with Medical Subject Headings (MeSH) and added to MEDLINE
d. A collection of books available full-text and other subsets of NLM records!"
e. PMC citations
f. NCBI Bookshelf

Many PubMed records contain links to full text articles, some of which are freely available, often in PubMed Central" and local mirrors, such as Europe PubMed Central Information about the journals indexed in MEDLINE, and available through PubMed, is found in the NLM Catalog As of 27 January 2020, PubMed has more than 30 million citations and abstracts dating back to

1966, selectively to the year 1865, and very selectively to 1809. As of the same date, 20 million of PubMed's records are listed with their abstracts, and 21.5 million records have links to full- text versions (of which 7.5 million articles are available, full-text for free). Over the last 10 years (ending 31 December 2019), an average of nearly 1 million new records were added each year. Approximately 12% of the records in PubMed correspond to cancer-related entries, which have grown from 6% in the 1950s to 16% in 2016. Other significant proportion of records correspond to ‘chemistry’ (8.69%), ‘therapy’ (8.39%), and ‘infection’ (5%). In 2016, NLM changed the indexing system so that publishers are able to directly correct typos and errors in PubMed indexed articles. PubMed has been reported to include some articles published in predatory journals MEDLINE and PubMed policies for the selection of journals for database inclusion are slightly different. Weaknesses in the criteria and procedures for indexing journals in PubMed Central may allow publications from predatory journals to leak into PubMed.

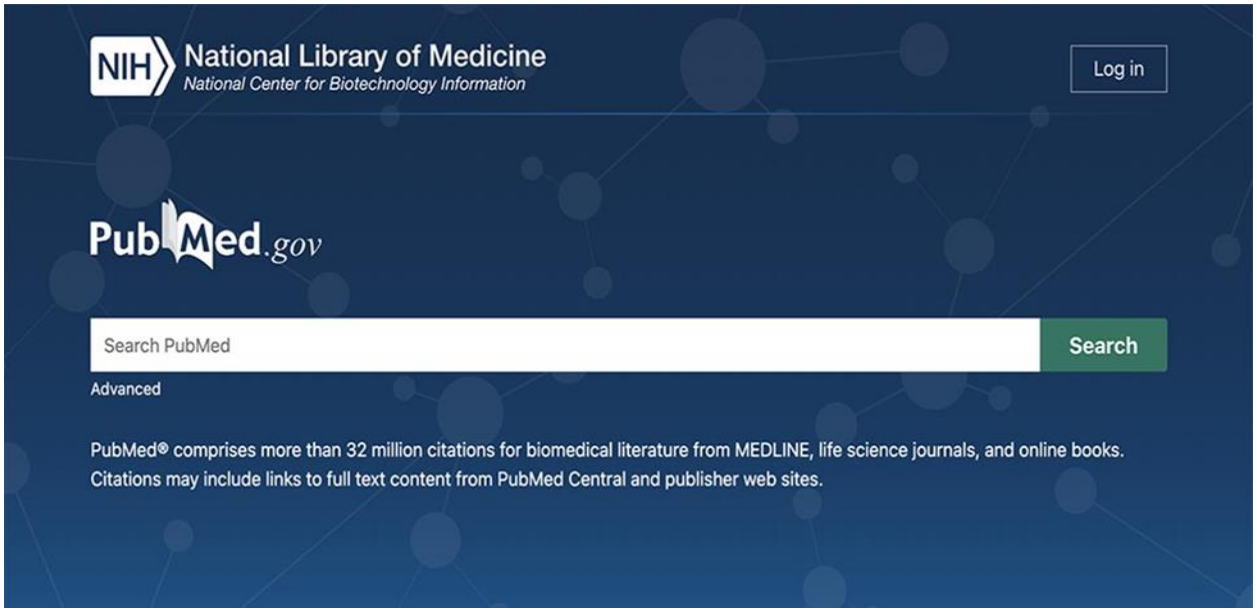


Figure 1: Homepage of PubMed Database

2. **<u>PubMed Central (PMC) Database:</u>**

PubMed Central (PMC) is a free digital repository that archives open access full-text scholarly articles that have been published in biomedical and life sciences journals. As one of the major research databases developed by the National Center for Biotechnology Information (NCBI). PubMed Central is more than a document repository. Submissions to PMC are indexed and formatted for enhanced metadata, medical ontology, and unique identifiers which enrich the XML structured data for each article. Content within PMC can be linked to other NCBI databases and accessed via Entrez search and retrieval systems, further enhancing the public's ability to discover, read and build upon its biomedical knowledge.

PubMed Central is distinct from PubMed. PubMed Central is a free digital archive of full articles, accessible to anyone from anywhere via a web browser (with varying provisions for reuse). Conversely, although PubMed is a searchable database of biomedical citations and abstracts, the full-text article resides elsewhere (in print or online, free or behind a subscriber paywall).

As of December 2018, the PMC archive contained over 5.2 million articles, with contributions coming from publishers or authors depositing their manuscripts into the repository per the NIH Public Access Policy. Earlier data shows that from January 2013 to January 2014 author-initiated deposits exceeded 103,000 papers during a 12-month period. PMC identifies about 4,000 journals which participate in some capacity to deposit their published content into the PMC repository. Some publishers delay the release of their articles on PubMed Central for a set time after publication, referred to as an ‘embargo period’, ranging from a few months to a few depending on the journal (Embargoes of six to twelve months are the most common). PubMed years Central is a key example of systematic external distribution by a third party which is still prohibited by the contributor agreements of many publishers.

The PMCID (PubMed Central identifier), also known as the PMC reference number, is a bibliographic identifier for the PubMed Central database, much like the PMID is the bibliographic identifier for the PubMed database. The two identifiers are distinct however. It consists of 'PMC' followed by a string of seven numbers. The format is: PMCID: PMC1852221. Authors applying for NIH awards must include the PMCID in their application.

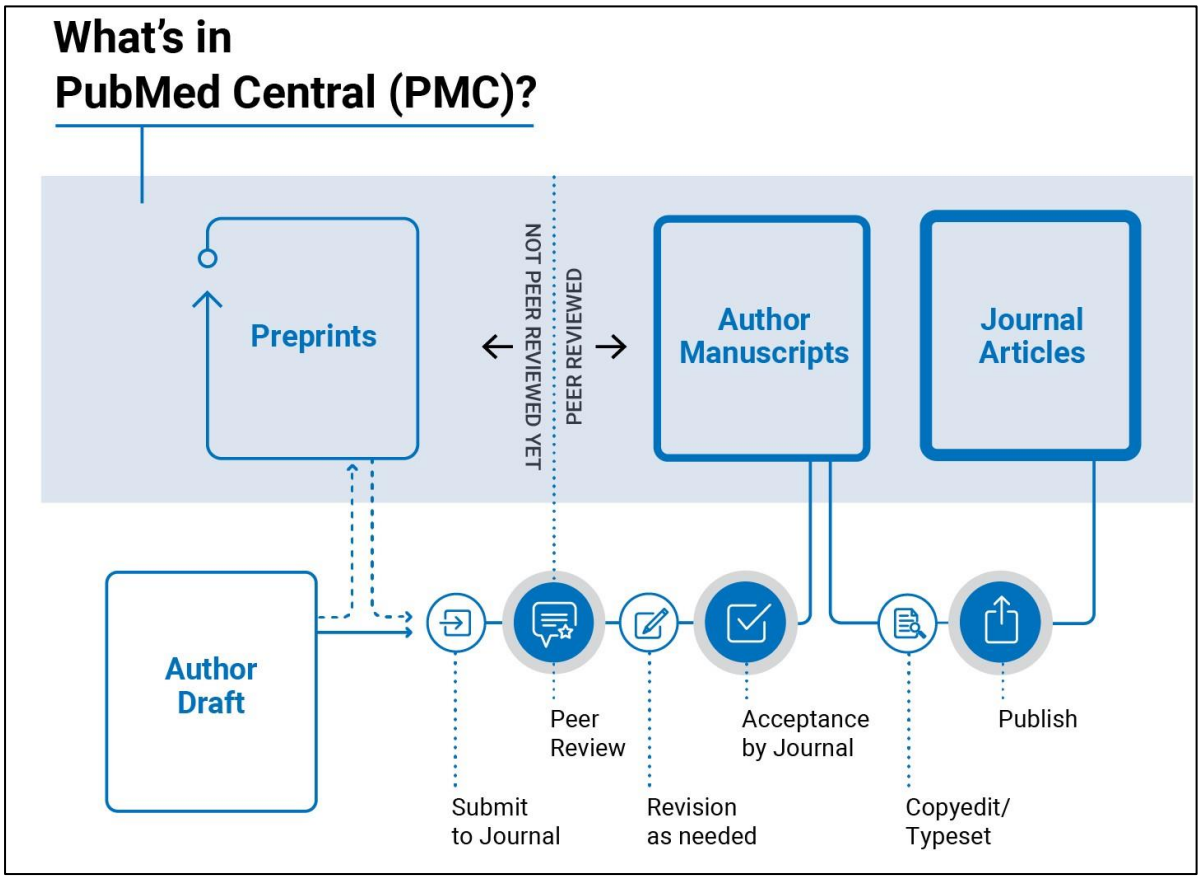


Figure 2: Overview of PubMed Central Database

3. **<u>Kyoto Encyclopedia of Genes and Genomes (KEGG) Database:</u>**

KEGG (Kyoto Encyclopedia of Genes and Genomes) is an effort to link genomic information with higher order functional information by computerizing current knowledge on cellular processes and by standardizing gene annotations. Generally speaking, the biological function of the living cell is a result of many interacting molecules; it cannot be attributed to just a single gene or a single molecule. The functional assignment in KEGG is a process of linking a set of genes in the genome with a network of interacting molecules in the cell, such as a pathway or a complex, representing a higher order biological function.

a. **Genomes to Biological System:**

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the biosphere, from genomic and molecular-level information. It is a computer model of the biological system, consisting of molecular building blocks of genes

and proteins (genomic information) and chemical substances (chemical information) that are integrated with molecular wiring diagrams of interaction and reaction networks (systems information). The KEGG model also contains disease and drug information (health information) in terms of perturbed molecular networks.

The concept behind developing KEGG is described in the webpage of Kanehisa Laboratories. KEGG is a reference knowledge base that links genomes to biological systems. It is widely used with the KEGG mapping procedure for integration and interpretation of large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies.

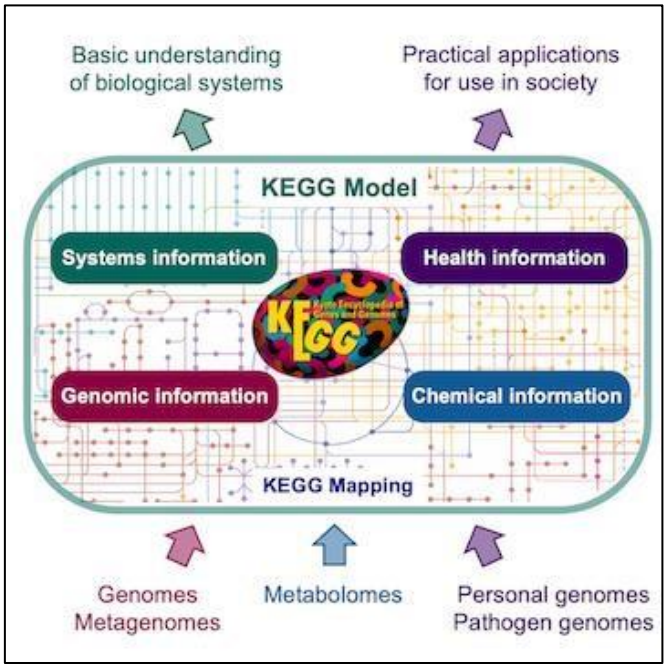


Figure 3: KEGG Model Mapping

b. Overview of KEGG database:

The KEGG model is implemented as an integrated database resource consisting of sixteen databases shown below. They are broadly categorized into systems information, genomic information, chemical information and health information, which are distinguished by color coding of web pages.

Category	Database	Content	Color
Systems Information	KEGG PATHWAY	KEGG pathway maps	
	KEGG BRITE	BRITE hierarchies and tables	
	KEGG MODULE	KEGG modules and reaction modules	

Genomic Information	KEGG ORTHOLOGY (KO)	Functional orthologs	
	KEGG GENES	Genes and proteins	
	KEGG GENOME	KEGG organisms and viruses	
Chemical Information (KEGG LIGAND)	KEGG COMPOUND	Metabolites and other chemical substances	
	KEGG GLYCAN	Glycans	
	KEGG REACTION KEGG RCLASS	Biochemical reactions Reaction class	
	KEGG ENZYME	Enzyme nomenclature	
Health Information (KEGG MEDICUS)	KEGG NETWORK	Disease-related network variations	
	KEGG DISEASE	Human diseases	
	KEGG DRUG KEGG DGROUP	Drugs Drug groups	
	KEGG VARIANT	Human gene variants	

c. **Network Variants:**

The KEGG database has been developed by focusing on conservation and variation of genes and genomes among different organisms. The reference datasets of KEGG pathway maps, BRITE hierarchies and KEGG modules have been developed with the concept of functional orthologs (KOs), so that KEGG pathway mapping and other procedures can be applied to any cellular organism.

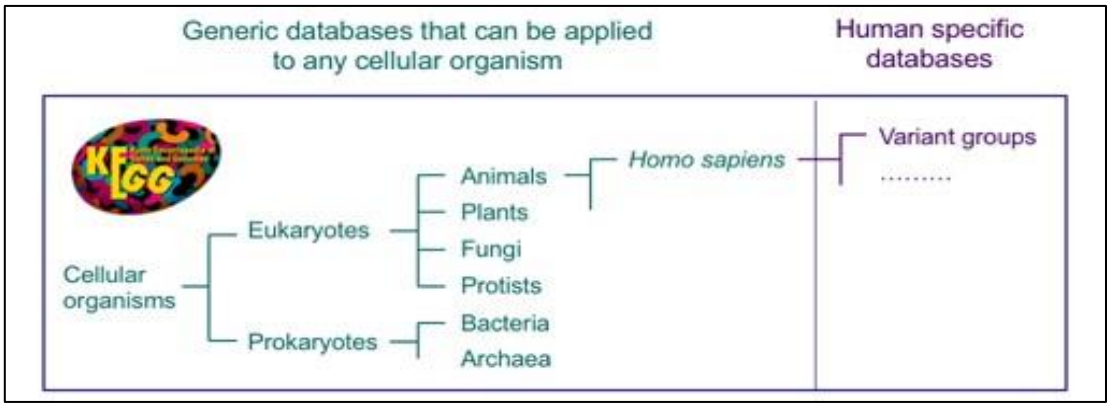


Figure 4: Network variants of KEGG database

However, this generic approach is inadequate for understanding more detailed features caused by variations of genes and genomes within a species, especially for understanding disease related variations of human genes and genomes. KEGG NETWORK represents a renewed attempt by KEGG to capture knowledge on diseases and drugs in terms of network variants caused by not only gene variants, but also viruses and other factors.

d. Pathway Identifier:

In the KEGG database, a regular map notation typically follows the format 'mapXXXXX,' where 'XXXXX' represents a five-digit numerical code assigned to a specific pathway or map in the KEGG pathway database. For example, the pathway for 'Citrate cycle (TCA cycle)' is represented by the notation 'map00020' in KEGG. You can replace 'XXXXX' with the specific numerical code corresponding to the pathway or map you are interested in.
Each pathway map is identified by the combination of 2-4 letter prefix code and 5-digit number.

The prefix has the following meaning:

Sr. No.	Prefix	Meaning
1.	Map	manually drawn reference pathway
2.	Ko	reference pathway highlighting KOs
3.	Ec	reference metabolic pathway highlighting EC numbers
4.	Rn	reference metabolic pathway highlighting reactions
5.	<org>	organism-specific pathway generated by converting KOs to gene identifier

And the numbers starting with the following:

Sr. No.	Numbers	Meaning
1.	010	global map (lines linked to KOs)
2.	012	overview map (lines linked to KOs)
3.	010	chemical structure map (no KO expansion)
4.	07	drug structure map (no KO expansion)

KEGG PATHWAY is integrated with MODULE and NETWORK databases as indicated below:

1. M - module
2. R - reaction module
3. N - network

e. Regular Map Notation:

In the KEGG pathway maps, various notation is used to represent different molecular components, reactions, and other entities.

Here are some common notations used in KEGG pathway maps:

Sr. No.	Notation	Meaning
1.	Rectangle	Represents proteins or protein complexes
2.	Rounded rectangle	Represents genes
3.	Ellipse	Represents small molecules, such as metabolites or ions
4.	Arrow	Indicate reactions or interactions between entities
5.	Dashed arrow	May indicate indirect interactions or regulatory effects
6.	T-bar (inhibitory symbol)	Represent inhibitory interactions
7.	Addition/Subtraction signs	Denote activation and inhibition, respectively
8.	Diamond	Represent other types of molecules or biochemical events

It's important to note that the specific symbols and their meanings can vary between different pathway maps. Each map in KEGG comes with a legend that explains the symbols used in that particular map. When interpreting a specific KEGG pathway map, always refer to the legend provided for accurate understanding.

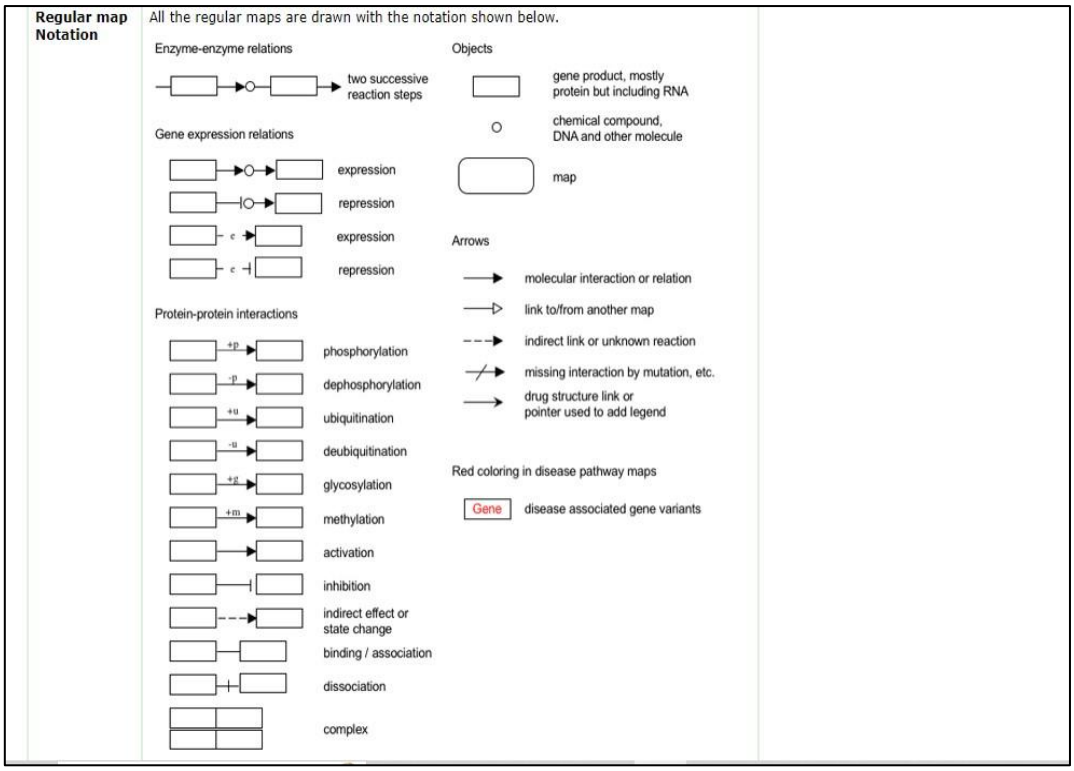


Figure 5: Regular Map Notation of KEGG Pathway Database

4. **OMIM Database:**

OMIM stands for Online Mendelian Inheritance in Man, which is a comprehensive and authoritative knowledgebase of human genes and genetic disorders. It was started by Dr. Victor A. McKusick as the definitive reference Mendelian Inheritance in Man and OMIM is now distributed electronically by the National Center for Biotechnology Information, where it is integrated with the Entrez Suite of databases.

OMIM is curated and edited at Johns Hopkins University with input from scientists and physicians around the world. Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.

The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 16,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources. OMIM focuses on the relationship between phenotype and genotype.

OMIM is based on the peer-reviewed biomedical literature, and criteria for inclusion of papers continue to evolve. In general, priority for inclusion is given to papers that provide significant insight into the gene-phenotype relationship, expand our understanding of human biology, or contribute to the characterization of a disorder. Information in each OMIM entry is cited, and the full reference is provided. OMIM is an easy and straightforward portal to the burgeoning Information in human genetics. PheneGene graphics (OMIM PheneGene graphics depict relationships between phenotypes, groups of related phenotypes (Phenotypic Series), and genes. They are graphical representations of the information in OMIM's Genemap and Phenotypic Series. These relationships are not hierarchical).

The numbering system used in OMIM describes:

Sr. No.	MIM Number	Type of Data
1.	100000- to 200000-	Autosomal loci or phenotypes (entries created before May 15, 1994)
2.	300000-	X-linked loci or phenotypes
3.	400000	Y-linked loci or phenotypes
4.	500000	Mitochondrial loci or phenotypes
5.	600000	Autosomal loci or phenotypes (entries created after May 15, 1994)
6.	MIM number of the entry, followed by a decimal point and a unique 4-digit variant number	Allelic variants

The symbols preceding a MIM number represents:

Sr. No.	Symbols	Meaning
1.	Asterisk (*)	Indicates a gene entry
2.	Number symbol (#)	Indicates that it is a descriptive entry, usually of a phenotype, and does not represent a unique locus. The reason for the use of the number symbol is given in the first paragraph of the entry. Discussion of any gene(s) related to the phenotype resides in other entries as described in the first paragraph
3.	Plus sign (+)	Indicates that the entry contains the description of a gene of known sequence and a phenotype
4.	Percent sign (%)	Indicates that the entry describes a confirmed mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known
5.	No symbol	Generally, indicates a description of a phenotype for which the mendelian basis, although suspected, has not been clearly established or that the separateness of this phenotype from that in another entry is unclear
6.	Caret (^)	Indicates the entry no longer exists because it was removed from the database or moved to another entry as indicated

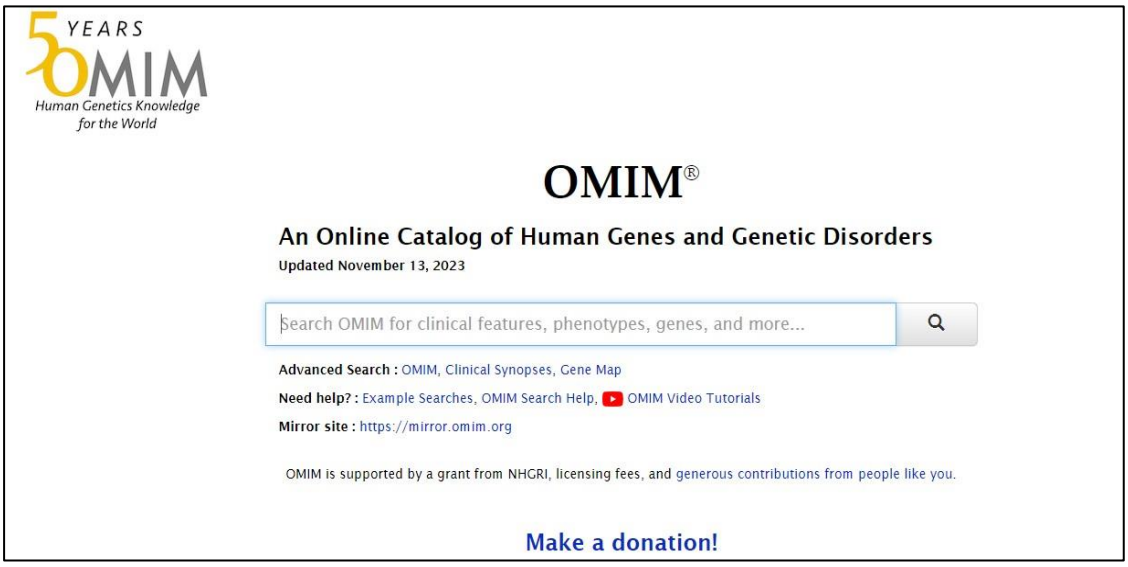


Figure 6: OMIM Database Homepage

REFERENCES:

1. Kanehisa, M. (2000, January 1). KEGG: Kyoto Encyclopedia of Genes and Genomes.
2. Nucleic Acids Research; Oxford University Press. https://doi.org/10.1093/nar/28.1.27
3. KEGG Pathway Map (Help). (n.d.). https://www.genome.jp/kegg/document/help_pathway.html
4. Xiong, J. (2006, March 13). Introduction. Cambridge University Press eBooks. https://doi.org/10.1017/cbo9780511806087.002

DATE: 25/08/2023

WEBLEM 2(A)
INTEGRATED DATABASE SYSTEMS VIZ. BASIC, ADVANCED LIMITS USING LITERATURE RESOURCES
(URL: https://www.ncbi.nlm.nih.gov/)

AIM:
To study literature database for query, "Melanin" in NCBI database and filter results using BASIC, LIMIT & ADVANCE search.

INTRODUCTION:
The National Centre for Biotechnology Information (NCBI) is a prominent and comprehensive resource in the field of bioinformatics and molecular biology. It is part of the United States National Library of Medicine (NLM), which is itself a branch of the National Institutes of Health (NIH). NCBI plays a central role in organizing, storing, and disseminating biological information, including genetic and genomic data, scientific literature, and various tools and resources for researchers, clinicians, and the general public.

Literature databases:

1. **Bookshelf:**
   Bookshelf, the books division of the NLM Literature Archive (LitArch) at the National Centre for Biotechnology Information (NCBI), is an online searchable collection of books, reports, databases, and other scholarly literature in biology, medicine, and the life sciences. The NCBI bookshelf is integrated with other NCBI resources allows seamless transition between the databases to access available information.
2. **MeSH (Medical Subject Headings):**
   The comprehensive resource that allows users to explore and search for MeSH terms, find related terms and locate descriptors that are relevant to research interests, also used to navigate through MeSH vocabulary.
3. **NLM Catalog:**
   The catalog contains bibliographic records and searches using keywords, titles, author names and other criteria to locate specific resources or explore topics of interest. It also provides advanced search options , allowing user to refine the searches based on specific fields, publication type and other criteria.
4. **PUBMED Central:**
   PMC is a free full text archive of biomedical and life sciences journal literature. Provides open access to vast collection of research articles, reviews, and other types of scientific literature making it valuable resource for researchers, healthcare professionals, and the general public.

5. **PUBMED:**
   Widely used for biomedical literature database that encompasses articles from various fields, including medicine, biology and healthcare. Indexed using medical subheadings, PubMed facilitates precise searching and categorization of articles. PubMed is a subset of PMC.

Melanin:

Melanin is a natural pigment found in many living organisms, including humans. It plays a crucial role in determining the color of various tissues and structures within the body, such as the skin, hair, eyes, and even the inner ear. It is produced by specialized cells called melanocytes, which are primarily found in the skin but are also present in other areas of the body.

There are two main types of melanin:

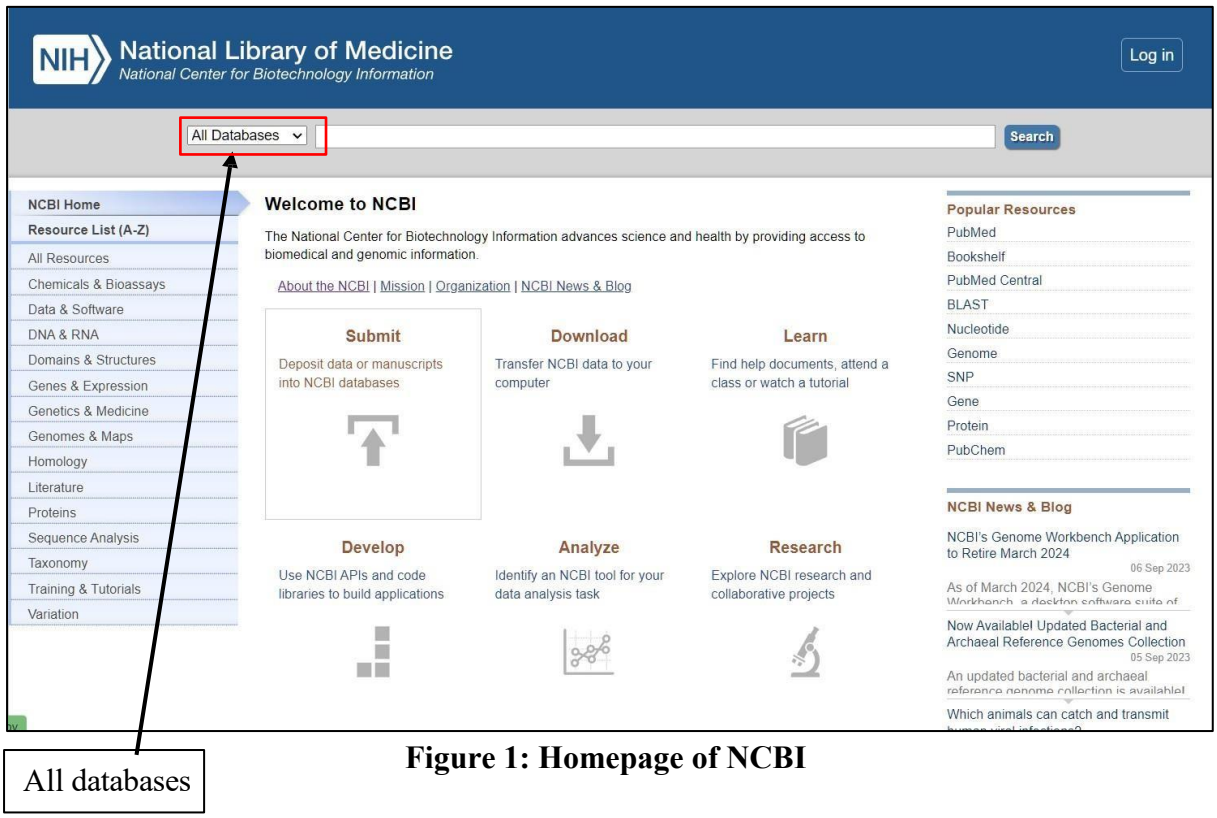
1. **Eumelanin:** This type of melanin is responsible for the brown and black pigmentation in hair, skin, and eyes. Eumelanin is associated with darker shades of color.
2. **Pheomelanin:** Pheomelanin is responsible for the red and yellow pigmentation found in hair and skin. It is generally associated with lighter shades of color

The production of melanin is influenced by genetics, hormones, and environmental factors, particularly exposure to ultraviolet (UV) radiation from the sun. When the skin is exposed to UV radiation, melanocytes produce more melanin as a protective mechanism. This increase in melanin production leads to tanning, which is the darkening of the skin's color. It acts as a natural defense against the harmful effects of UV radiation by absorbing and dissipating the UV energy, preventing it from causing DNA damage and mutations in skin cells.

METHODOLOGY:

1. Go to the NCBI website.
2. Enter the query ‘Melanin’ and click on search.
3. Perform basic search for Books, limit the result by applying 1 year filter and report resource type.
4. Perform basic search for MeSH, limit the search by clicking on subheadings.
5. Perform Basic search for PubMed for query ‘Melanin’ and apply filters, free full text and reduce the publication time from 2010 to 2023.
6. Perform basic search for PMC catalogue for the query.
7. Perform basic search for NLM for the query.

OBSERVATIONS:



All databases

Figure 1: Homepage of NCBI

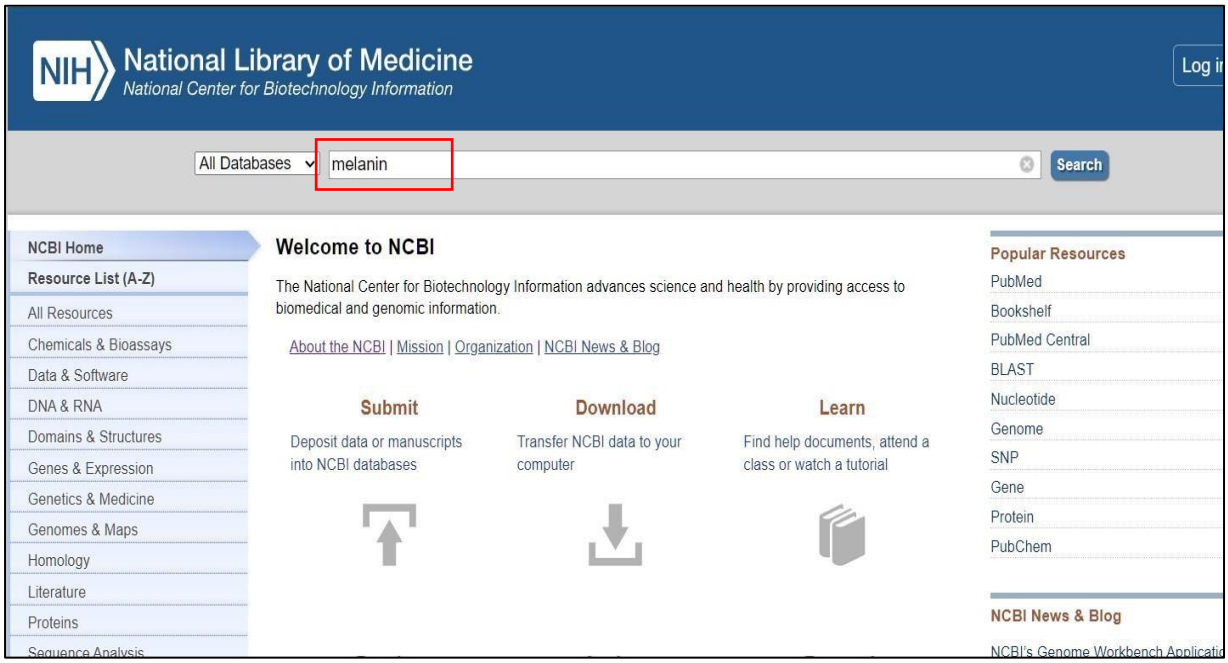


Figure 2: 'Melanin' Query Search

National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI melanin Search

Results found in 30 databases

Literature		Genes		Proteins	
Bookshelf	690	Gene	2,463	Conserved Domains	24
MeSH	28	GEO DataSets	194	Identical Protein Groups	106
NLM Catalog	64	GEO Profiles	117,829	Protein	49,719
PubMed	24,335	HomoloGene	5	Protein Family Models	53
PubMed Central	49,949	PopSet	41	Structure	66
Genomes		**Clinical**		**PubChem**	
Assembly	0	ClinicalTrials.gov	228	BioAssays	3,519
BioCollections	0	ClinVar	118	Compounds	15
BioProject	127	dbGaP	4	Pathways	12
BioSample	81	dbSNP	0	Substances	167

Figure 3: Entrez search result found in 30 databases

Database changed to Books

National Library of Medicine
National Center for Biotechnology Information

Log in

Bookshelf Books melanin Search

Browse Titles Create alert Advanced Help

Previous versions or editions
Include in the search result

Publication date
1 year
5 years
10 years
Custom range...

Added to Bookshelf
60 days
1 year
5 years
Custom range...

Resource types
Report
Book
Collection
Documentation
Database

Clear all

Show additional filters

Display Settings: Summary, 20 per page, Sorted by Relevance

Send to:

Results: 1 to 20 of 154 books (690 items)

<< First < Prev Page 1 of 8 Next > Last >>

1. StatPearls [Internet].
Treasure Island (FL): StatPearls Publishing; 2023 Jan-.
Top results in this book Table of Contents

2. Molecular Imaging and Contrast Agent Database (MICAD) [Internet].
Bethesda (MD): National Center for Biotechnology Information (US); 2004-2013.
Top results in this book Table of Contents

3. GeneReviews® [Internet].
Adam MP, Mirzaa GM, Pagon RA, et al., editors.
Seattle (WA): University of Washington, Seattle; 1993-2023.
Top results in this book Table of Contents

4. PDQ Cancer Information Summaries [Internet].
Bethesda (MD): National Cancer Institute (US); 2002-.
Top results in this book Table of Contents

5. Endotext [Internet].
Feingold KR, Anawalt B, Blackman MR, et al., editors.
South Dartmouth (MA): MDText.com, Inc.; 2000-.
Top results in this book Table of Contents

Textbook on Scar Management: State of the Art Management and Emerging Technologies

Images search in Bookshelf

See more (18)...

Search details

("melanins"[All Fields] OR "melanin"[All Fields]) OR ("melanins"[KYWD] OR "melanins"[All Fields] OR "melanin"[All Fields])

Search

See more...

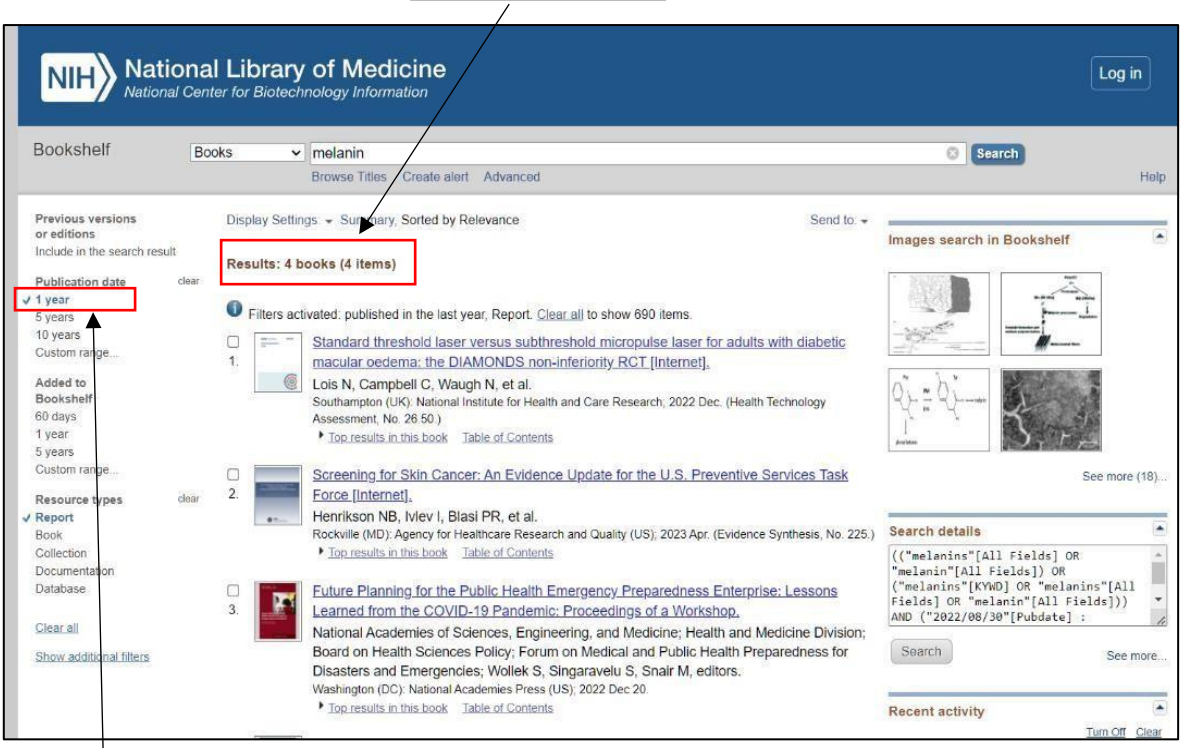
Recent activity

Turn Off Clear

melanin (690)

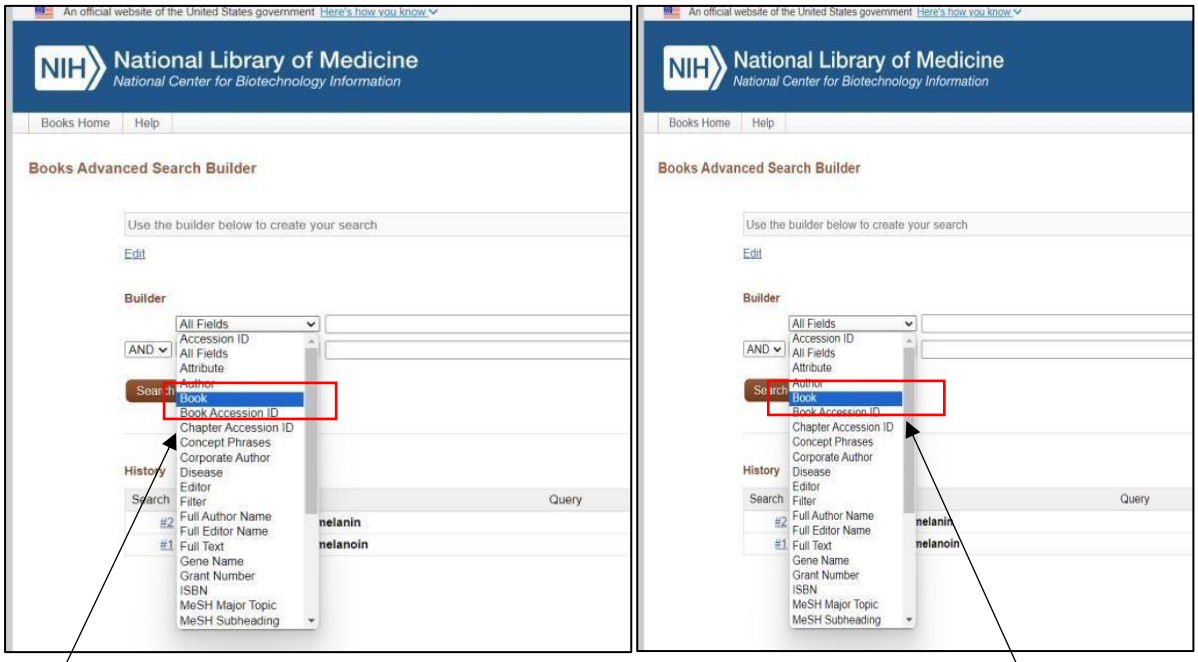
Figure 4: Results page for Basic search for books

4 books in result



Apply filter of 1 year

Figure 5: Result for books search with filter



Books filter

Figure 6: Advanced search for books

All fields

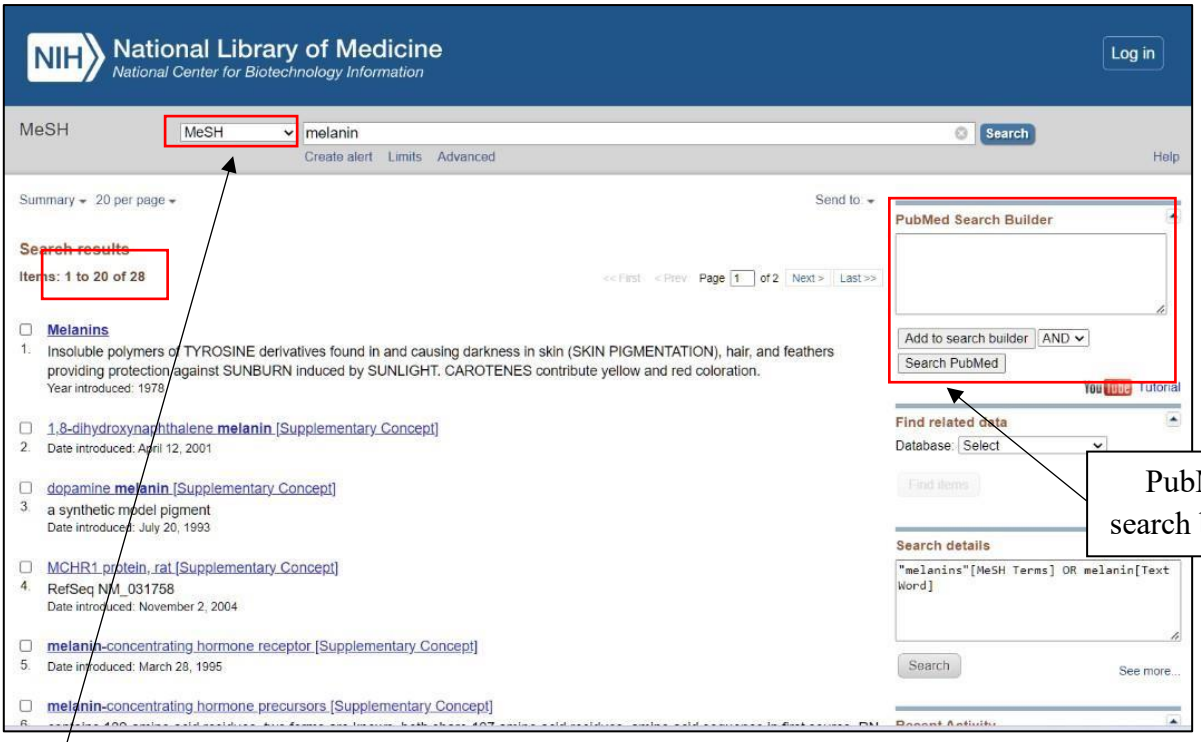


Figure 7: MeSH search for query 'Melanin' with PubMed search builder

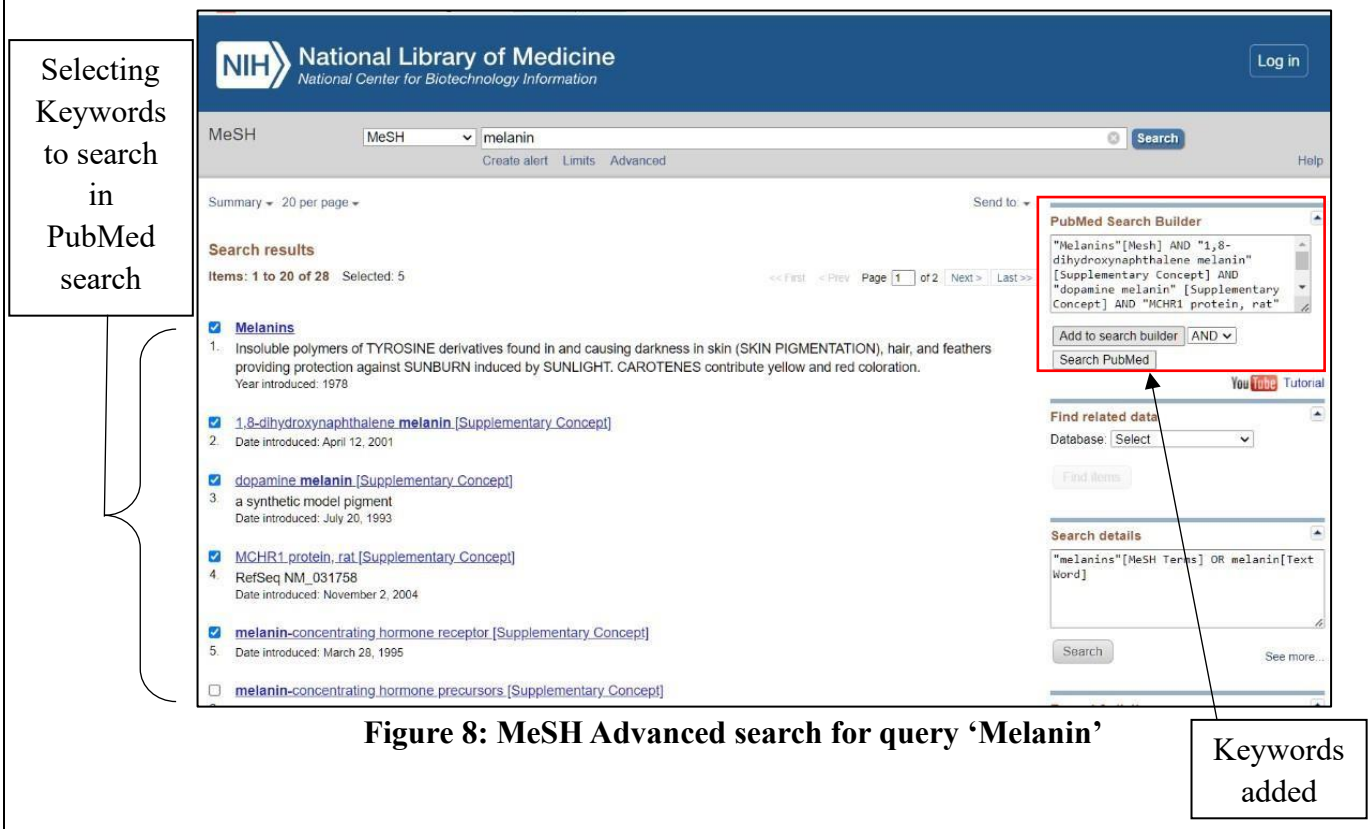
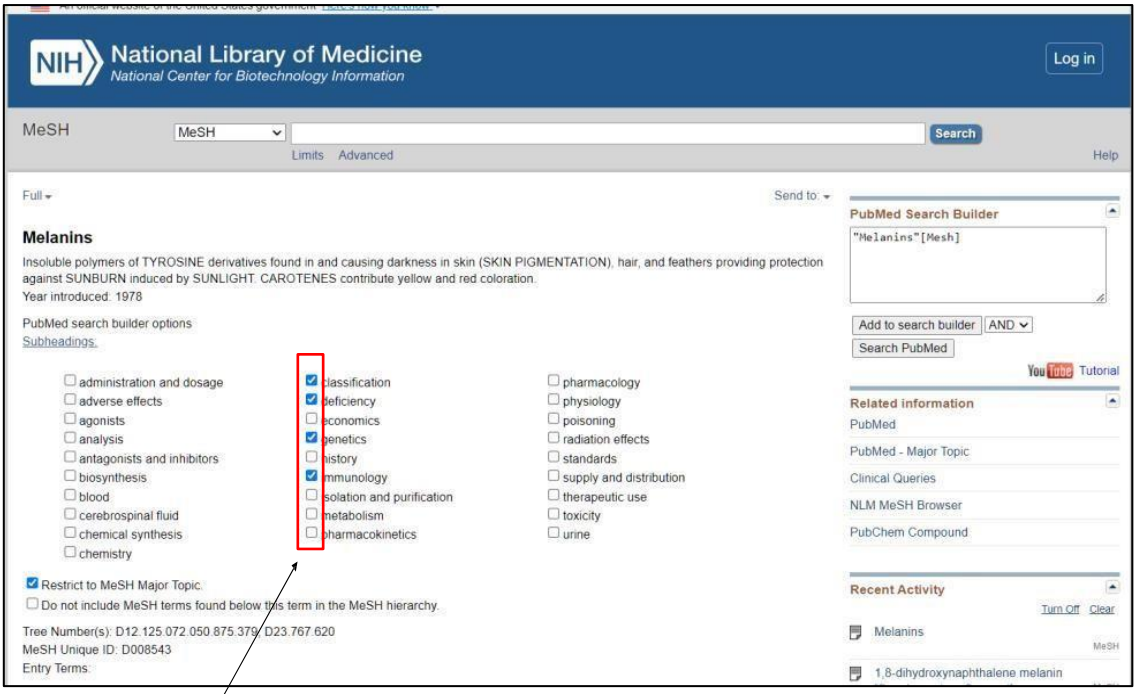


Figure 8: MeSH Advanced search for query 'Melanin'



Selection of subheadings

Figure 9: MeSH subheadings

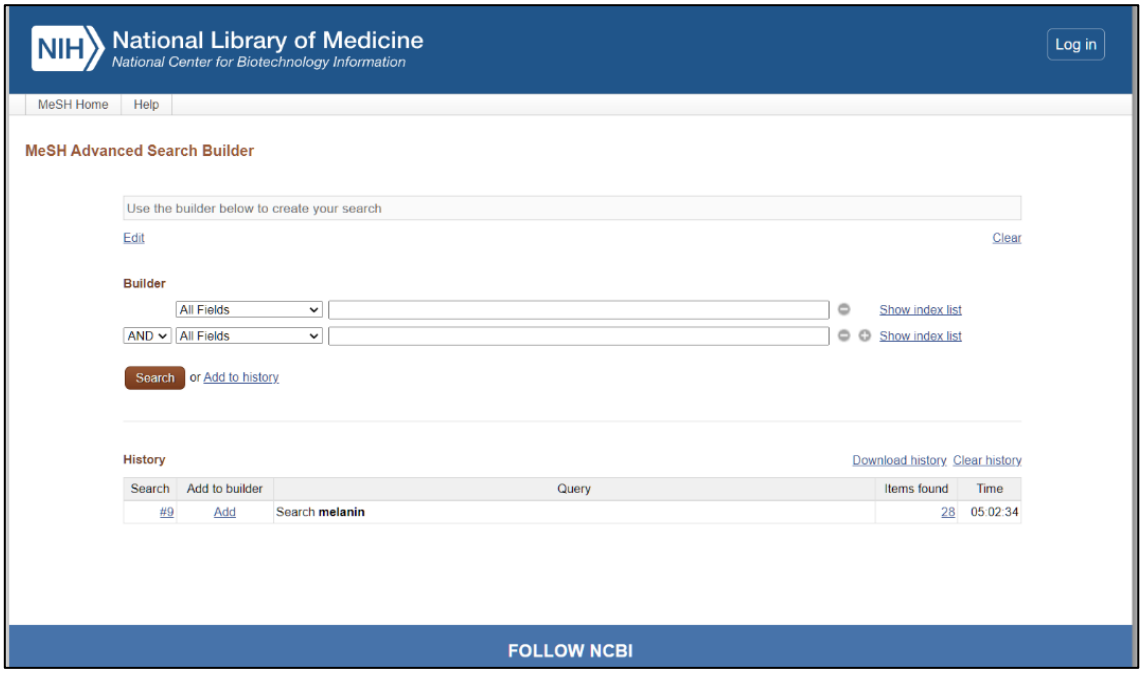


Figure 10: MeSH Advanced Search Builder

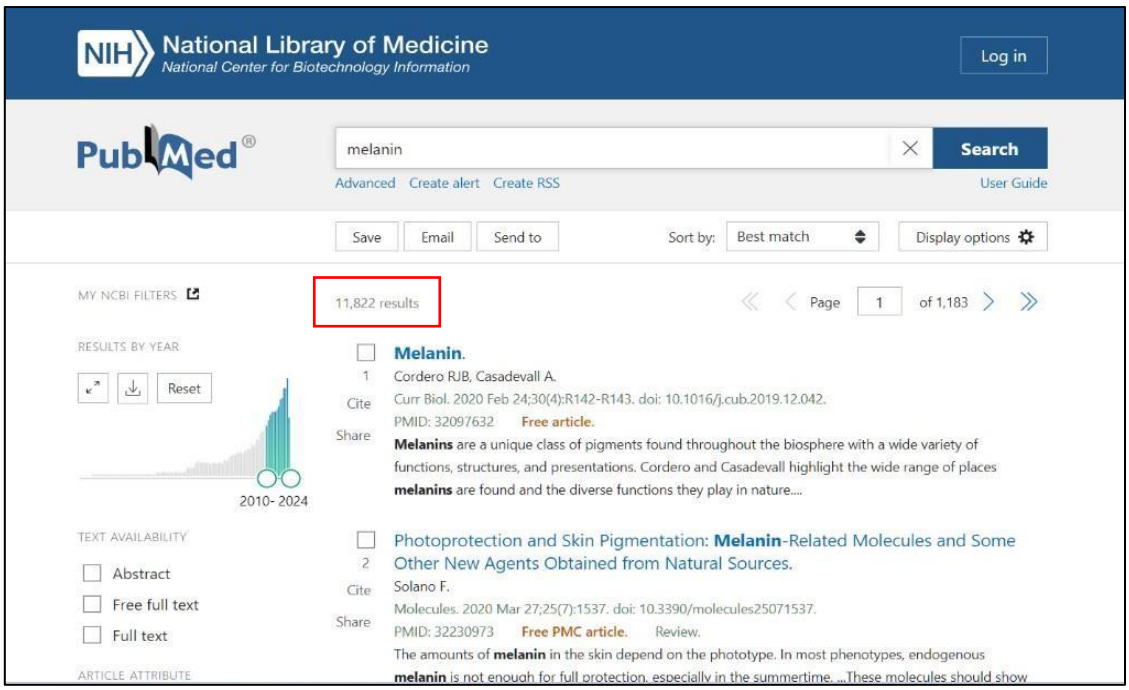


Figure 11: PubMed search results page for query 'Melanin'

Results reduced to 5957

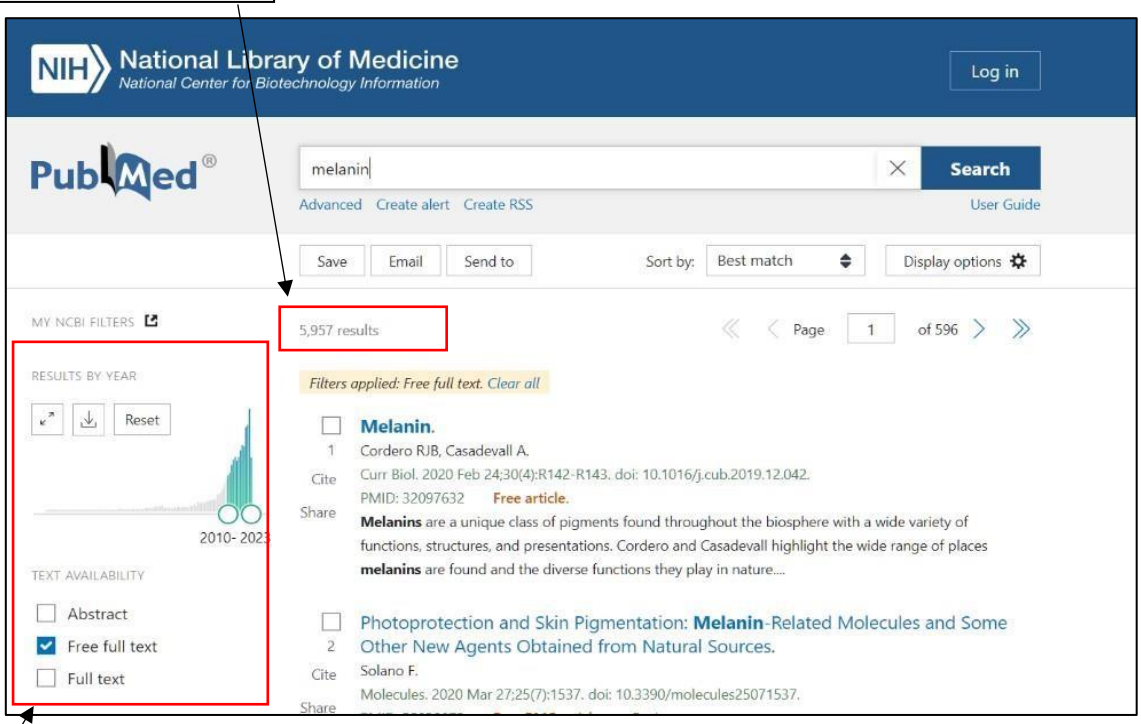


Figure 12: PubMed search for query 'Melanin' with year and availability filter

Year and Availability filters applied

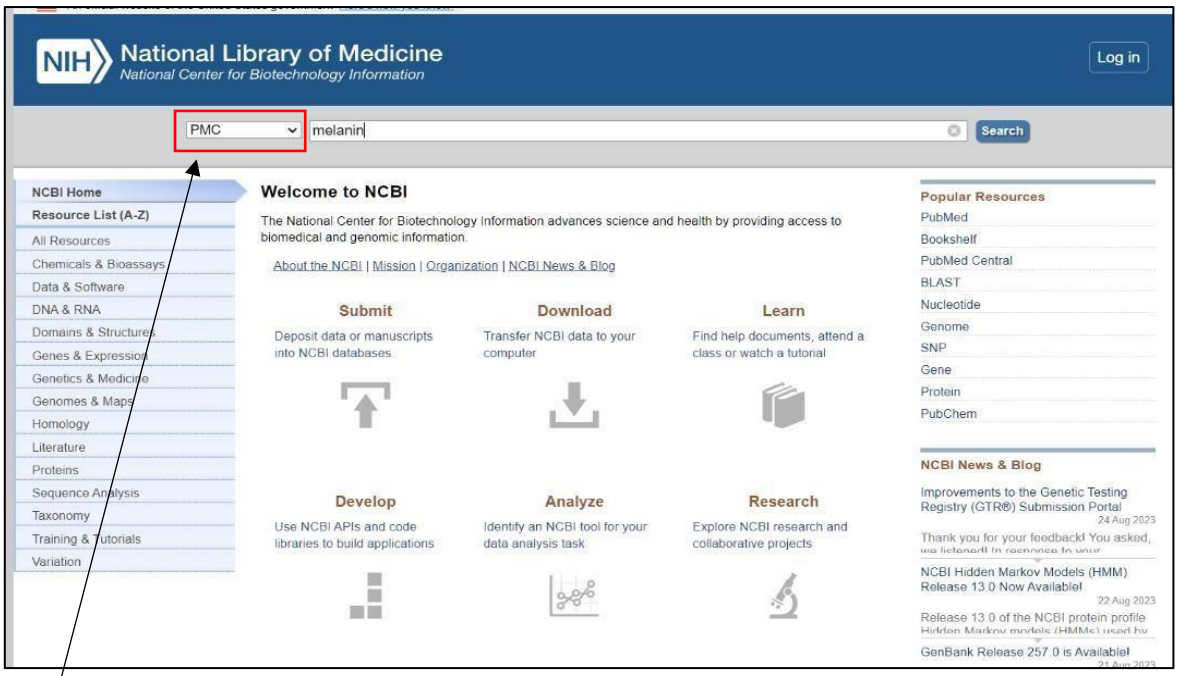


Figure 13: PMC search page for query 'Melanin'

Database changed to PMC

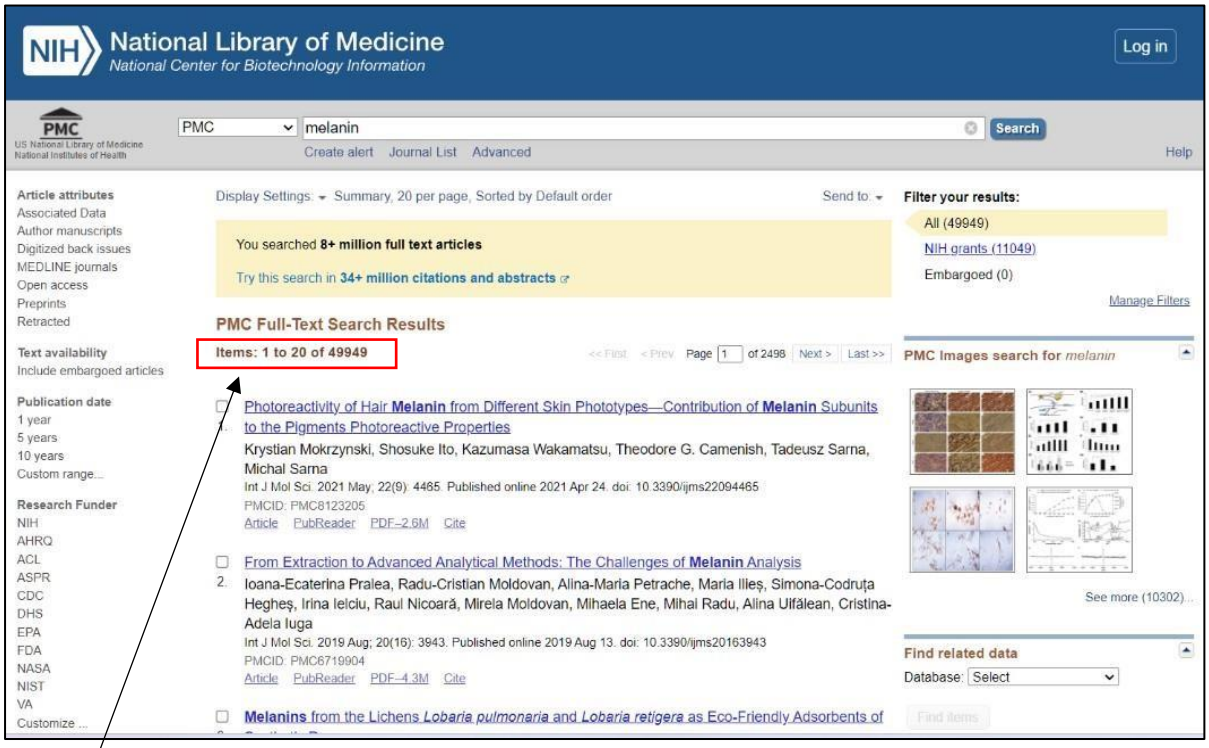


Figure 14: PMC Search result for query 'Melanin'

49949 results

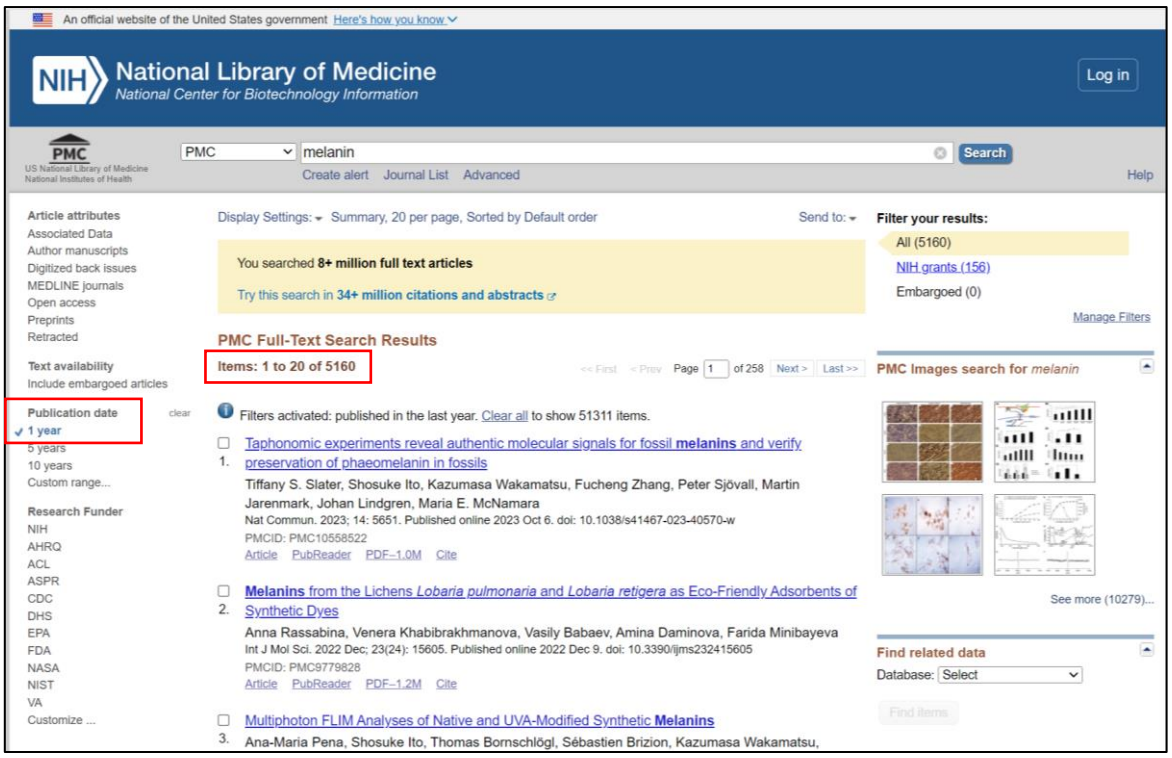


Figure 15: PMC Search result for query 'Melanin' with filter

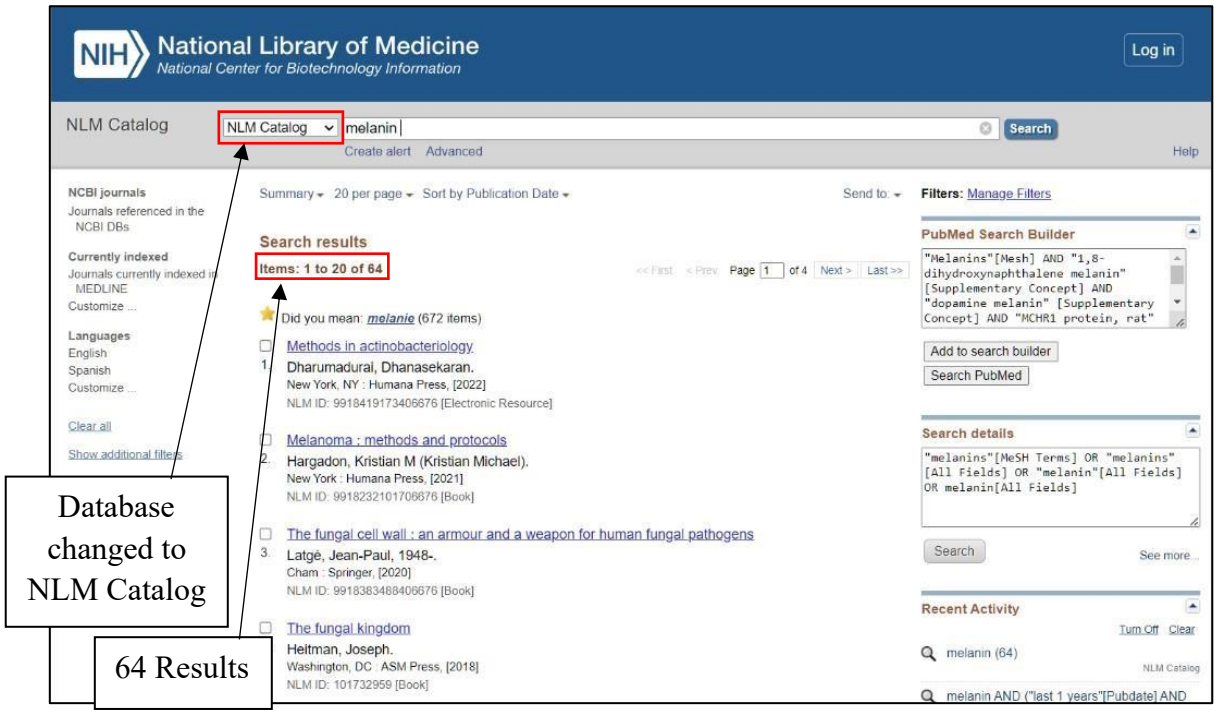


Figure 16: NLM Catalog search page for query 'Melanin'

National Library of Medicine
National Center for Biotechnology Information

Log in

NLM Catalog Home | Help

NLM Catalog Advanced Search Builder

Use the builder below to create your search

Edit | Clear

Builder

All Fields | Show index list

AND | All Fields | Show index list

Search or Add to history

History | Download history | Clear history

Search	Add to builder	Query	Items found	Time
#1	Add	Search **melanin** Sort by: **NLMID**	64	12:13:33
#10	Add	Search **melanin** Sort by: **PubDate** Filters: **English**	53	12:13:19
#9	Add	Search **melanie** Sort by: **PubDate** Filters: **Journals currently indexed in MEDLINE; English**	0	12:13:18
#8	Add	Search **melanin** Sort by: **PubDate** Filters: **Journals currently indexed in MEDLINE; English**	0	12:13:18
#7	Add	Search **melanie** Sort by: **PubDate** Filters: **Journals referenced in the NCBI DBs; Journals currently indexed in MEDLINE; English**	0	12:13:14
#6	Add	Search **melanin** Sort by: **PubDate** Filters: **Journals referenced in the NCBI DBs; Journals currently indexed in MEDLINE; English**	0	12:13:14

Figure 17: NLM Catalog advanced search builder page

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Log in

NLM Catalog | NLM Catalog | (melanin) AND nature[Publisher] | Search

Create alert | Advanced | Help

NCBI journals
Journals referenced in the NCBI DBs

Currently indexed
Journals currently indexed in MEDLINE
Customize ...

Languages
English
Spanish
Customize ...

Clear all

Show additional filters

Summary | Sort by Publication Date | Send to:

Search results

Items: 3

Showing results for ***(melanie) AND nature[Publisher]***. Your search for *(melanin) AND nature[Publisher]* retrieved no results.

1. The kappa opioid receptor
Liu-Chen, Lee-Yuan; Inan, Saadet.
Cham, Switzerland : Springer **Nature**, [2022]
NLM ID: 9918434484406676 [Book]

2. Pharmacology of the WNT signaling system
Schulte, Gunnar; Kozielewicz, Pawel.
Cham, Switzerland : Springer **Nature**, [2021]
NLM ID: 9918351184106676 [Book]

3. Bone toxicology
Smith, Susan Y; Varela, Aurore; Samadfam, Rana.
Cham, Switzerland : Springer **Nature**, [2017]
NLM ID: 101723146 [Book]

Summary | Sort by Publication Date | Send to:

Filters: Manage Filters

PubMed Search Builder

Add to search builder

Search PubMed

Search details

```
melanie[All Fields] AND
nature[Publisher]
```

Search | See more...

Recent Activity

Turn Off | Clear

(melanie) AND nature[Publisher] (3) NLM Catalog

(melanin) AND 1 year[Publication Year] (0) NLM Catalog

Figure 18: NLM Catalog Search result for query 'Melanin' with filter

RESULTS:

Different literature databases such as PubMed, PMC, Bookshelf, NLM catalog, MeSH of the National Centre for Biotechnology Information (NCBI) that centralizes all literature resources into individual search results and records. The query searched was 'Melanin' a pigment. Under the basic search the number obtained was 690 for bookshelf, 28 for MeSH, 64 for NLM catalog, 28,335 for PubMed, PubMed central 49,949 were obtained.

List of resources	Number of hits	Number of hits after filters
Bookshelf	690	4
MeSH	28	10
NLM catalog	64	3
PubMed	24335	612
PubMed Central	49949	5160

CONCLUSION:

The National Centre for Biotechnology Information's (NCBI) different literature resources integrates specific information from multiple literature resources. The query 'Melanin' was searched in all resources and basic, limit and advanced search was performed that helped to explore the NCBI's different literature resources.

REFERENCES:

1. Ferreira, J. G. P., Bittencourt, J. C., & Adamantidis, A. (2017, June). Melanin-concentrating hormone and sleep. *Current Opinion in Neurobiology*, *44*, 152–158. https://doi.org/10.1016/j.conb.2017.04.008
2. Schlessinger, D. I. (2023, May 1). *Biochemistry, Melanin*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK459156/

DATE: 30/10/2023

WEBLEM 2(B)
KYOTO ENCYCLOPEDIA OF GENES AND GENOMES (KEGG) DATABASE
(URL: https://www.genome.jp/kegg/)

AIM:
To explore the Kyoto Encyclopedia of Genes and Genomes (KEGG) Database with respect to the analysis of the functions of genes and enzymes, and the metabolic reactions involved in caffeine metabolism pathway.

INTRODUCTION:
The Kyoto Encyclopedia of Genes and Genomes, commonly known as KEGG database, stands as a pivotal resource in the realm of bioinformatics, genomics, and systems biology. Established in the late 20th century, KEGG database represents a dynamic and comprehensive repository that meticulously catalogs information pertaining to molecular pathways, biological systems, and genomic functions across diverse organisms.
At its core, KEGG database serves as an invaluable tool for researchers and scientists aiming to dissect and comprehend the intricate orchestration of genes and their products within cellular processes. By integrating a vast array of biological data, KEGG database not only offers a detailed understanding of metabolic pathways but also provides insights into cellular signaling, environmental information processing, and various other biological functions.
One of the standout features of KEGG database is its pathway mapping system, which visually represents intricate networks of interactions among genes, proteins, and small molecules. This visualization aids in the interpretation of complex biological phenomena, fostering a systems-level understanding of cellular activities. The KEGG database plays a crucial role in bridging the gap between genomic information and functional interpretations. The database incorporates genomic data, such as DNA sequences and protein information, and correlates them with functional annotations, enabling researchers to connect genetic information with physiological functions. As a resource that continually evolves with advancements in genomic research, KEGG database remains an indispensable tool for scientists delving into diverse fields, including molecular biology, pharmacology, and medicine.

Caffeine metabolism:
Caffeine metabolism, the intricate process by which the human body breaks down and processes caffeine, is a subject of significant interest due to the widespread consumption of caffeine-containing beverages and products. Caffeine, a natural stimulant found in coffee, tea, and certain energy drinks, exerts its effects by influencing neurotransmitters and adenosine receptors in the central nervous system. Understanding the metabolic pathways of caffeine is crucial for unraveling its physiological impact, potential health effects, and interactions with other substances. This metabolic journey involves enzymes, primarily in the liver, that transform caffeine into various metabolites, each with distinct properties. Delving into caffeine

metabolism provides insights into individual responses to this widely consumed compound and informs discussions on its role in health and well-being.

METHODOLOGY:

1. Open the homepage of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.
2. To explore the pathways, click on 'KEGG PATHWAY' under the data-oriented entry points section.
3. Select the required pathway or search for the desired pathway in the search bar given above. Enter the query 'caffeine metabolism' and initiate the search.
4. After the query retrieval, observe the pathway.
5. Explore the desired pathway map (here, map00232).
6. Click on the desired ID (here, CYP1A2) to get detailed information about the orthology, enzyme and reaction of that pathway map.

OBSERVATIONS:

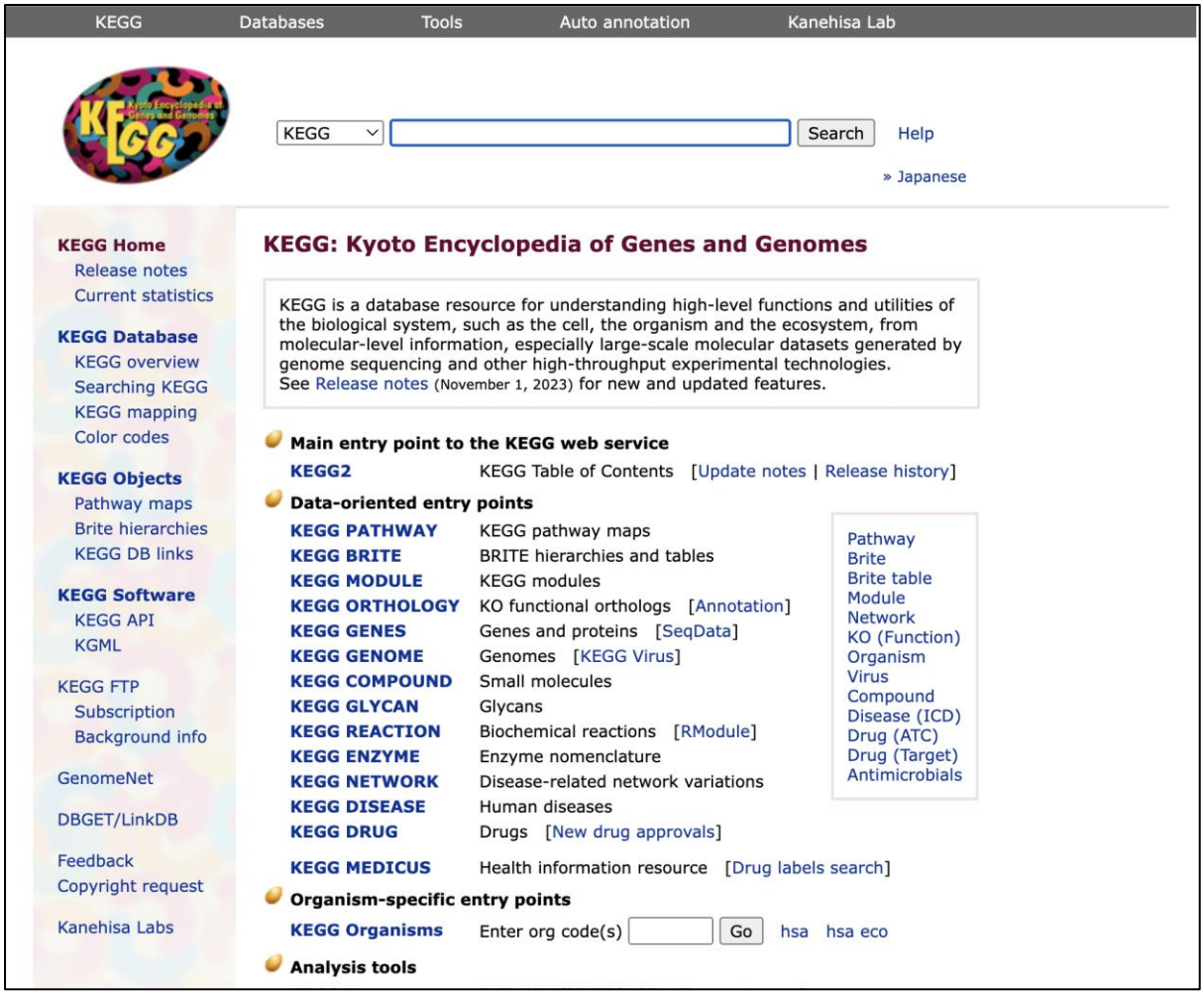


Figure 1: Homepage of the Kyoto Encyclopedia of Genes and Genomes (KEGG) Database

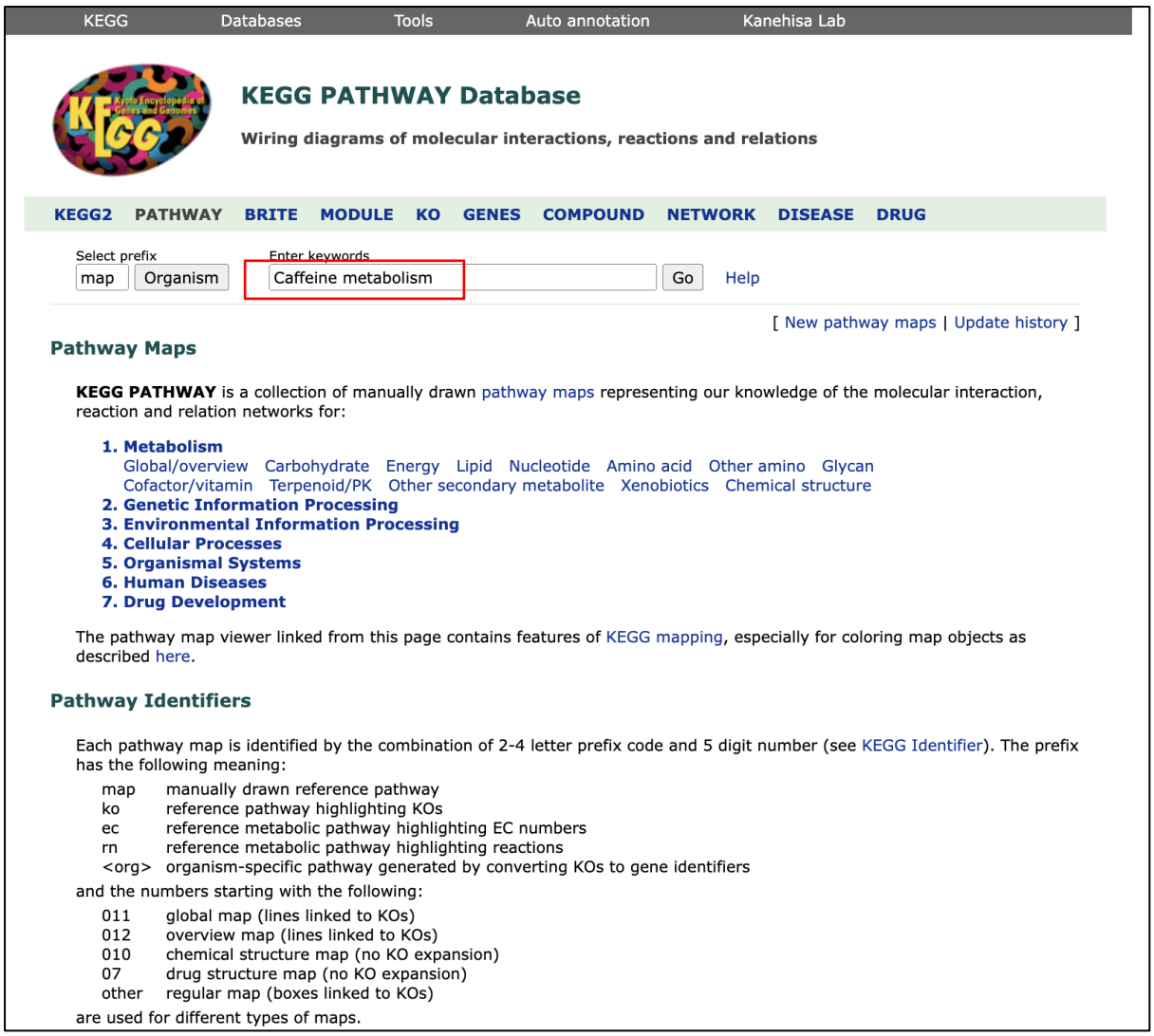


Figure 2: KEGG Pathway Maps

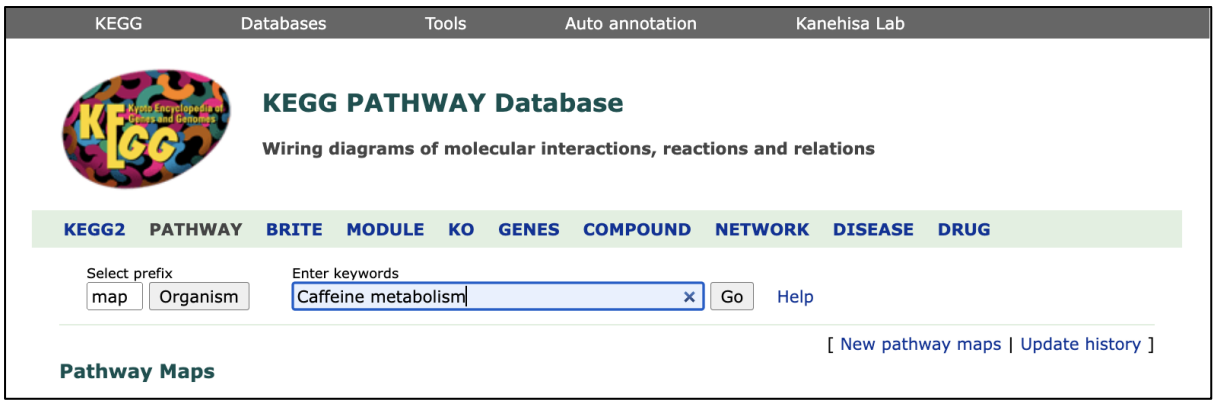
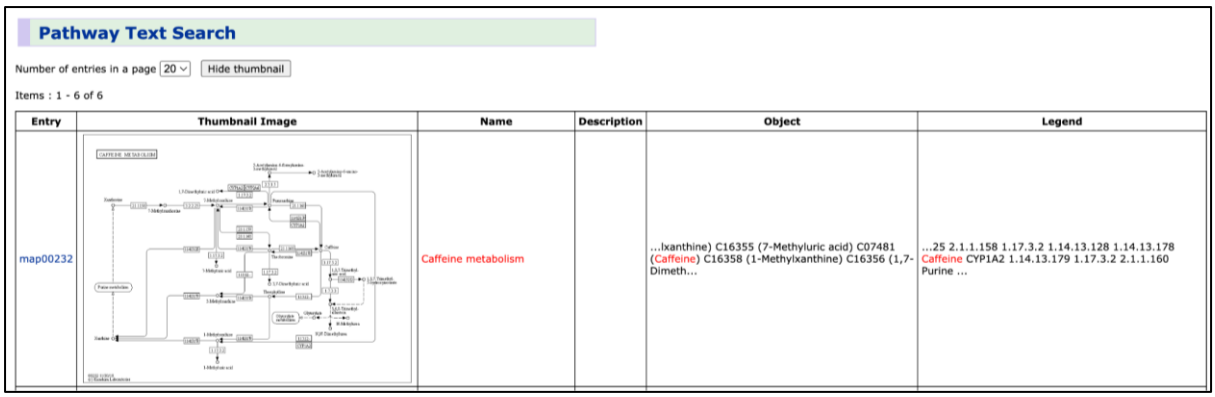


Figure 3: Searching for the query 'Caffeine Metabolism'



Pathway Text Search

Number of entries in a page 20 Hide thumbnail

Items : 1 - 6 of 6

Entry	Thumbnail Image	Name	Description	Object	Legend
map00232		Caffeine metabolism		...lxanthine) C16355 (7-Methyluric acid) C07481 (Caffeine) C16358 (1-Methylxanthine) C16356 (1,7-Dimeth...	...25 2.1.1.158 1.17.3.2 1.14.13.128 1.14.13.178 Caffeine CYP1A2 1.14.13.179 1.17.3.2 2.1.1.160 Purine ...

Figure 4: Results of the query searched

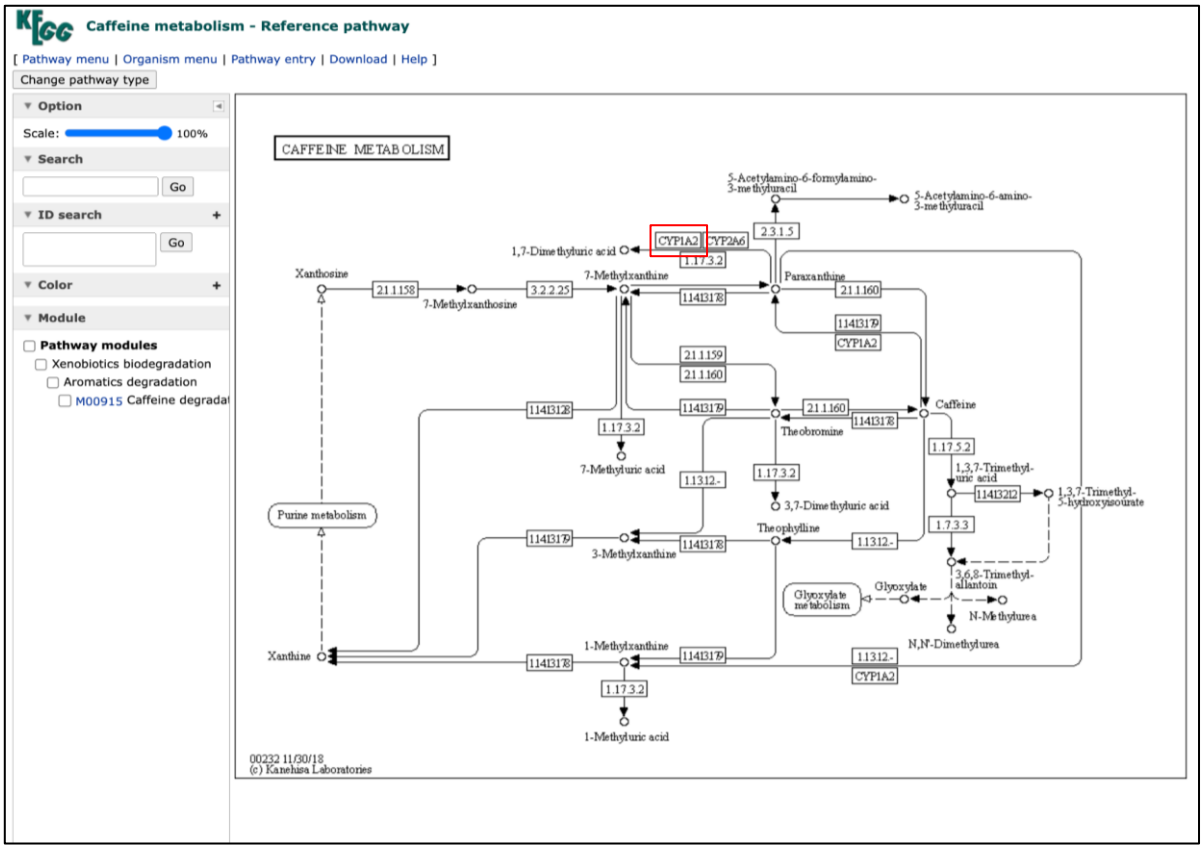


Figure 5: View of the Caffeine Metabolism Pathway Map.
Selected 'CYP1A2' for its further study

KEGG

ORTHOLOGY: K07409

Help

Entry	K07409 KO
Symbol	CYP1A2
Name	cytochrome P450 family 1 subfamily A2 [EC:1.14.14.1]
Pathway	map00140 Steroid hormone biosynthesis map00232 Caffeine metabolism map00380 Tryptophan metabolism map00591 Linoleic acid metabolism map00830 Retinol metabolism map00980 Metabolism of xenobiotics by cytochrome P450 map00982 Drug metabolism - cytochrome P450 map01100 Metabolic pathways map01110 Biosynthesis of secondary metabolites map05204 Chemical carcinogenesis - DNA adducts map05207 Chemical carcinogenesis - receptor activation map05208 Chemical carcinogenesis - reactive oxygen species
Reaction	R03408 R03629 melatonin,NADPH---hemoprotein reductase:oxygen oxidoreductase R07000 naphthalene,NADPH:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing) R07001 naphthalene,NADPH:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing) R07021 1-nitronaphthalene,NADPH:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing) R07022 1-nitronaphthalene,NADPH:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing) R07055 R07056 R07098 trichloroethene,NADPH:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing) R07099 trichloroethene,NADPH:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing) R07939 caffeine:oxygen oxidoreductase (N3-demethylating) R07943 R07945 R08293 R08294 R08392 R09405 R09407 R09408
Brite	KEGG Orthology (KO) [BR:ko00001] 09100 Metabolism 09103 Lipid metabolism 00140 Steroid hormone biosynthesis K07409 CYP1A2; cytochrome P450 family 1 subfamily A2 00591 Linoleic acid metabolism K07409 CYP1A2; cytochrome P450 family 1 subfamily A2 09105 Amino acid metabolism

All links

Ontology (3)
KEGG BRITE (3)
Pathway (24)
KEGG PATHWAY (24)
Chemical reaction (34)
KEGG ENZYME (1)
KEGG REACTION (19)
KEGG RCLASS (14)
Gene (167)
KEGG GENES (148)
KEGG MGENES (1)
RefGene (7)
OC (11)
Literature (1)
PubMed (1)
All databases (229)

Download RDF

Figure 6: Orthology information for 'CYP1A2'

ENZYME: 1.14.14.1

Help

Entry	EC 1.14.14.1 Enzyme
Name	unspecific monooxygenase; microsomal monooxygenase; xenobiotic monooxygenase; aryl-4-monooxygenase; aryl hydrocarbon hydroxylase; microsomal P-450; flavoprotein-linked monooxygenase; flavoprotein monooxygenase; substrate,reduced-flavoprotein:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing)
Class	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen; With reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen into the other donor BRITE hierarchy
Sysname	substrate,NADPH---hemoprotein reductase:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing)
Reaction(IUBMB)	RH + [reduced NADPH---hemoprotein reductase] + O2 = ROH + [oxidized NADPH---hemoprotein reductase] + H2O [RN:R04122]
Reaction(KEGG)	R04122 > R01842 R02354 R02355 R02356 R02503 R03088 R03089 R03090 R03408 R03629 R04121; (other) R03697 R05259 R07000 R07001 R07021 R07022 R07042 R07043 R07044 R07045 R07046 R07048 R07050 R07051 R07052 R07054 R07055 R07056 R07079 R07080 R07081 R07085 R07087 R07098 R07099 R07939 R07943 R07945 R08265 R08267 R08270 R08286 R08287 R08293 R08294 R08312 R08343 R08344 R08345 R08390 R08391 R08392 R09404 R09405 R09406 R09407 R09408 R09416 R09418 R09421 R09423 R09424 R09425 R09442 Reaction
Substrate	RH [CPD:C01371]; [reduced NADPH---hemoprotein reductase] [CPD:C03024]; O2 [CPD:C00007]
Product	ROH [CPD:C01335]; [oxidized NADPH---hemoprotein reductase] [CPD:C03161]; H2O [CPD:C00001]
Comment	A group of P-450 heme-thiolate proteins, acting on a wide range of substrates including many xenobiotics, steroids, fatty acids, vitamins and prostaglandins; reactions catalysed include hydroxylation, epoxidation, N-oxidation, sulfooxidation, N-, S- and O-dealkylations, desulfation, deamination, and reduction of azo, nitro and N-oxide groups. Together with EC 1.6.2.4, NADPH---hemoprotein reductase, it forms a system in which two reducing equivalents are supplied by NADPH. Some of the reactions

All links

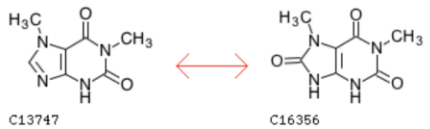
Pathway (26)
KEGG PATHWAY (26)
Chemical substance (108)
KEGG COMPOUND (108)
Chemical reaction (109)
KEGG REACTION (66)
KEGG RCLASS (43)
Gene (14365)
KEGG ORTHOLOGY (27)
KEGG GENES (8165)
KEGG MGENES (431)
RefGene (5742)
Protein sequence (10508)
UniProt (7907)
SWISS-PROT (220)
RefSeq(pep) (1801)
PDBSTR (483)
PMD (97)
DNA sequence (6777)
RefSeq(nuc) (5048)
GenBank (1089)
EMBL (640)
3D Structure (186)
PDB (186)
Protein domain (25)
InterPro (25)
Enzyme (1)
UMBBD-EC (1)
All databases (32105)

Download RDF

Figure 7: Enzyme related Information

REACTION: R07945

Help

Entry	R07945 Reaction
Definition	1,7-Dimethylxanthine <=> 1,7-Dimethyluric acid
Equation	C13747 <=> C16356
	
Comment	CYP2A6, CYP1A2
Reaction class	RC02017 C13747_C16356
Enzyme	1.14.14.1 1.14.14.-
Pathway	rn00232 Caffeine metabolism rn01100 Metabolic pathways
Brite	Enzymatic reactions [BR:br08201] 1. Oxidoreductase reactions 1.14 Acting on paired donors, with incorporation or reduction o 1.14.14 With reduced flavin or flavoprotein as one donor, and 1.14.14.1 R07945 1,7-Dimethylxanthine <=> 1,7-Dimethyluric acid 1.14.14.- R07945 1,7-Dimethylxanthine <=> 1,7-Dimethyluric acid BRITE hierarchy
Orthology	K07409 cytochrome P450 family 1 subfamily A2 [EC:1.14.14.1] K17683 cytochrome P450 family 2 subfamily A6 [EC:1.14.14.-]

All links

Ontology (1)
KEGG BRITE (1)
Pathway (4)
KEGG PATHWAY (4)
Chemical substance (2)
KEGG COMPOUND (2)
Chemical reaction (2)
KEGG ENZYME (1)
KEGG RCLASS (1)
Gene (159)
KEGG ORTHOLOGY (2)
KEGG GENES (157)
All databases (168)

Download RDF

DBGET integrated database retrieval system

Figure 8: Reaction Information

RESULTS:

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was explored to study the pathway of caffeine metabolism, [CYP1A2 (1.14.14.1)]. It provides a wealth of information on genes, proteins, biochemical pathways, and their interactions in various organisms. The KEGG database offers several key functionalities and features such as Genome and Gene Information, Enzyme information as well as orthology groups for query, Genome and Gene Information, Reaction associated to caffeine metabolism, etc.

CONCLUSION:

The Kyoto Encyclopedia of Genes and Genomes (KEGG) stands as a pivotal resource in the realm of biological information, offering a comprehensive and integrated database of biological systems, molecular interactions, and functional annotations KEGG's strength lies in its multi-faceted approach, combining genomic, chemical, and systemic information to elucidate the complex networks of molecular interactions within cells and organisms. It provides a wealth of data on pathways, genes, proteins, diseases, drugs, and their relationships, offering a holistic view of biological systems. KEGG remains an invaluable and comprehensive resource, providing a unified platform for understanding biological pathways, molecular interactions, and functional annotations. Its impact spans across diverse areas of research, from fundamental biology to drug discovery and clinical applications, serving as a catalyst for advancements in life sciences and contributing to the development of innovative therapeutics and treatments.

REFERENCES:

1. Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, Kanae Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D353–D361, https://doi.org/10.1093/nar/gkw1092
2. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, *34*(Database issue), D354–D357. https://doi.org/10.1093/nar/gkj102

DATE: 01/11/23

WEBLEM: 2(C)
ONLINE MENDELIAN INHERITANCE IN MAN (OMIM) DATABASE
(URL: https://www.omim.org/)

AIM:
To study the disease 'Hepatitis' (#114550) with a focus on chromosomal studies and investigate its genotypic and phenotypic relationships by exploring the Online Mendelian Inheritance in Man (OMIM) database.

INTRODUCTION:
OMIM database stands for Online Mendelian Inheritance in Man, which is a comprehensive and authoritative knowledgebase of human genes and genetic disorders. It was started by Dr. Victor A. McKusick as the definitive reference Mendelian Inheritance in Man and OMIM database is now distributed electronically by the National Center for Biotechnology Information, where it is integrated with the Entrez Suite of databases. OMIM database is curated and edited at Johns Hopkins University with input from scientists and physicians around the world. Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM database, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM database was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The full-text, referenced overviews in OMIM database contain information on all known mendelian disorders and over 16,000 genes. OMIM database focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources. OMIM database is based on the peer-reviewed biomedical literature, and criteria for inclusion of papers continue to evolve. In general, priority for inclusion is given to papers that provide significant insight into the gene-phenotype relationship, expand our understanding of human biology, or contribute to the characterization of a disorder. Information in each OMIM entry is cited, and the full reference is provided. OMIM database is an easy and straightforward portal to the burgeoning Information in human genetics. PheneGene graphics (OMIM PheneGene graphics depict relationships between phenotypes, groups of related phenotypes (Phenotypic Series), and genes. They are graphical representations of the information in OMIM's Genemap and Phenotypic Series. These relationships are not hierarchical).

Hepatitis:
Hepatitis is an inflammation of the liver that is caused by a variety of infectious viruses and noninfectious agents leading to a range of health problems, some of which can be fatal. There are five main strains of the hepatitis virus, referred to as types A, B, C, D and E. While they all cause liver disease, they differ in important ways including modes of transmission, severity of the illness, geographical distribution and prevention methods. Types B and C, in particular, lead

to chronic disease in hundreds of millions of people and together are the most common cause of liver cirrhosis, liver cancer and viral hepatitis-related deaths. An estimated 354 million people worldwide live with hepatitis B or C, and for most, testing and treatment remain beyond reach. Some types of hepatitis are preventable through vaccination. A WHO study found that an estimated 4.5 million premature deaths could be prevented in low- and middle-income countries by 2030 through vaccination, diagnostic tests, medicines and education campaigns. WHO's global hepatitis strategy, endorsed by all WHO Member States, aims to reduce new hepatitis infections by 90% and deaths by 65% between 2016 and 2030.

METHODOLOGY:

1. Open the homepage of OMIM database.
2. Enter search query of HEPATITIS in search box of OMIM database.
3. Interpret the PheneGene linear and radial graphs.
4. Interpret the Result of Description, Clinical features, Other features, Molecular genetics, Pathogenesis, Animal Mode.
5. Clinical synopsis from Advanced Search under the search bar.
6. Type 'hepatitis' in the search box.
7. In the 'Only Entries with' section, select Inheritance and Abdomen from the drop-down menu.

OBSERVATIONS:

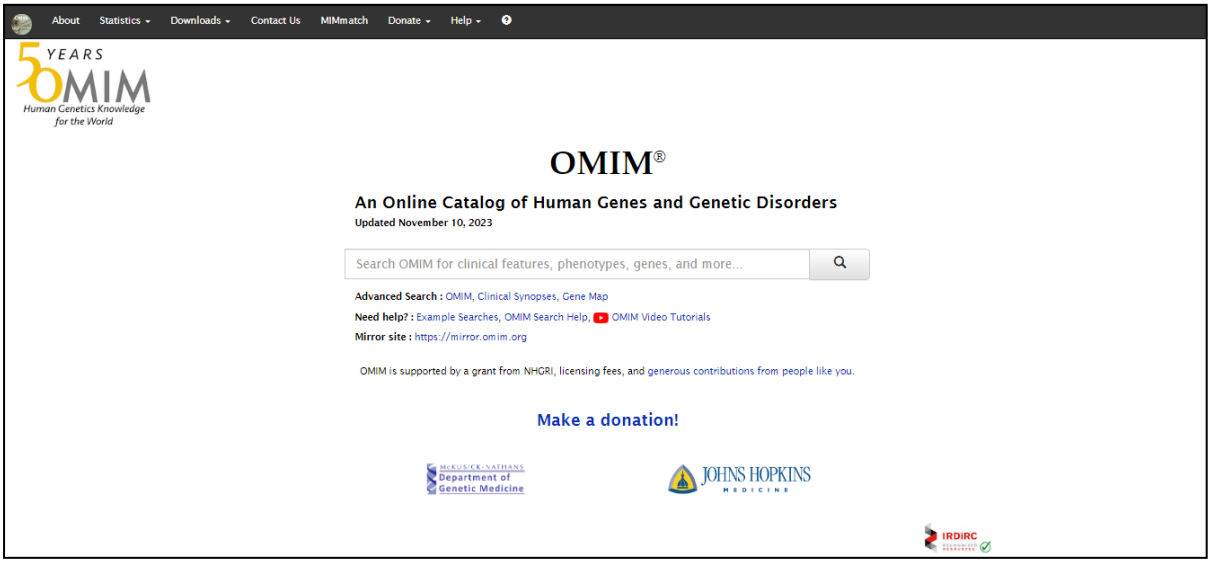


Figure 1: Homepage of the OMIM Database

A number symbol (#) before an entry number indicates that it is a descriptive entry, usually of a phenotype, and does not represent a unique locus.

No. of hits obtained

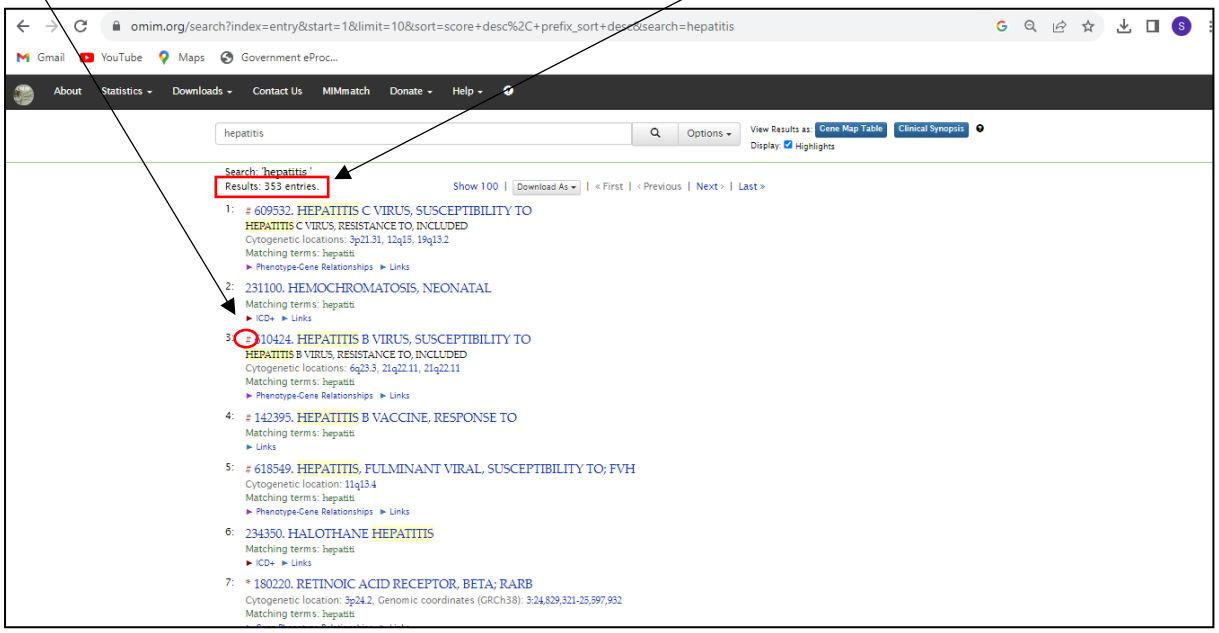


Figure 2: Hits for the query 'Hepatitis'

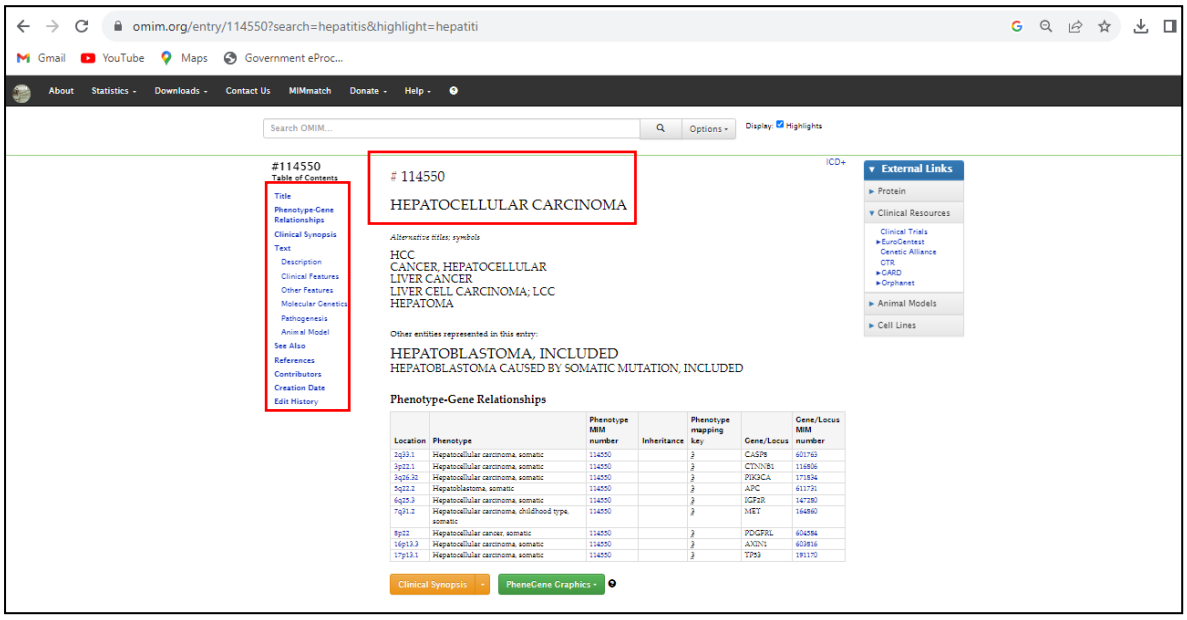


Figure 3: Result page for query 'Hepatitis'

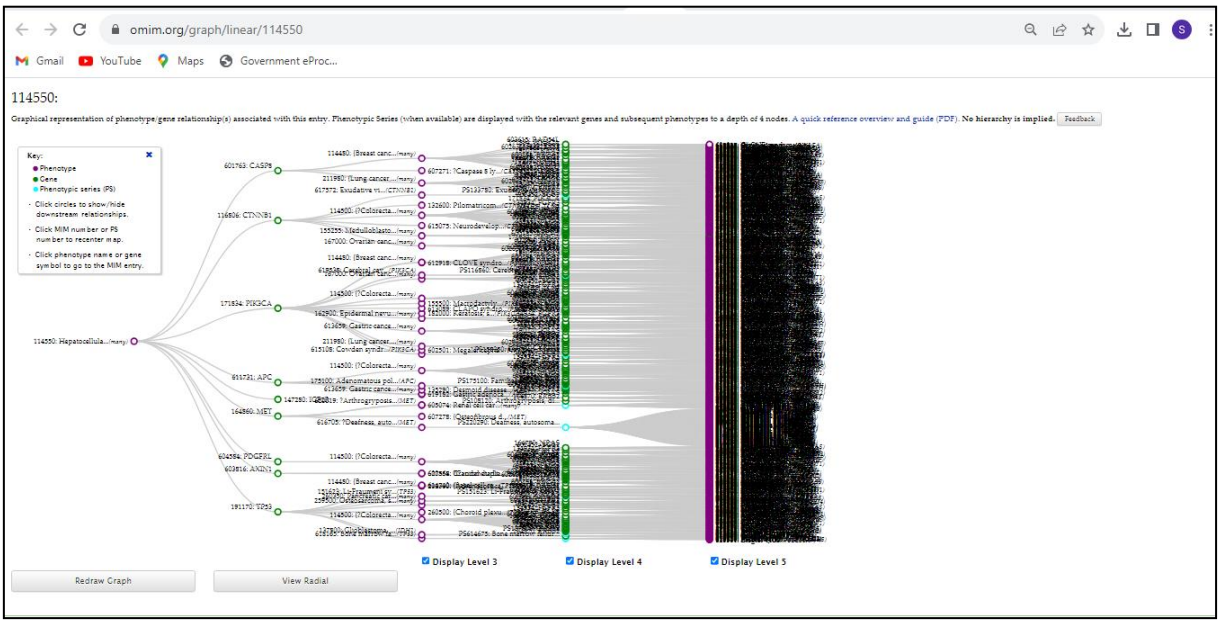


Figure 4: Result page of PheneGene Linear graph for query

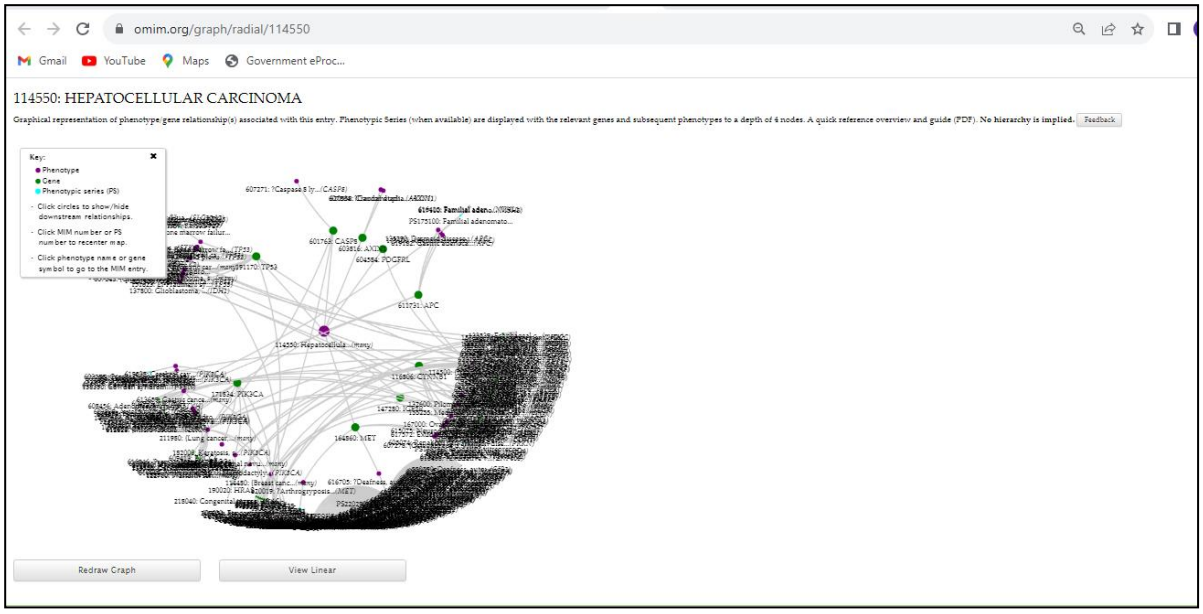


Figure 5: Result page of PheneGene Radial graph for query

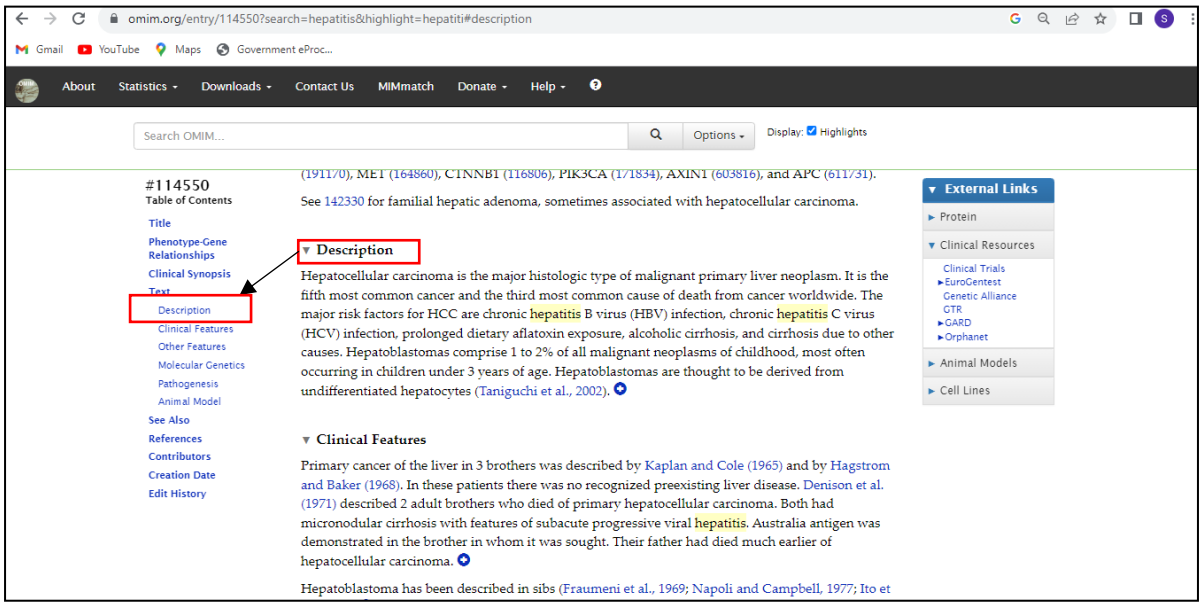


Figure 6: Result page of Description for the query

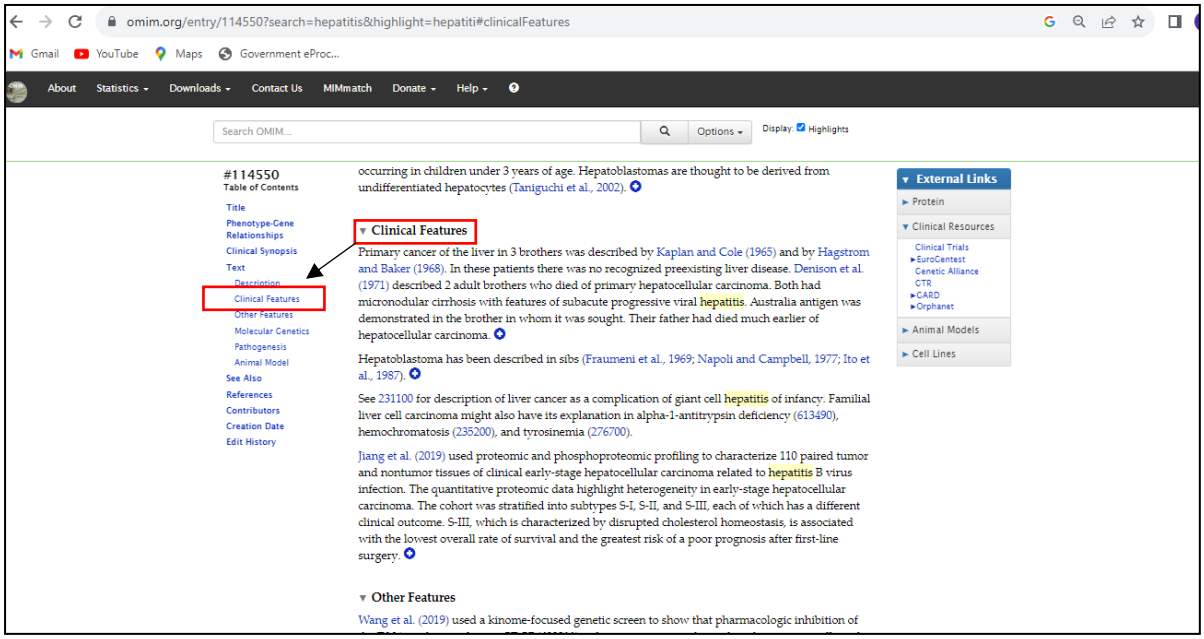


Figure 7: Result page of Clinical features for the query

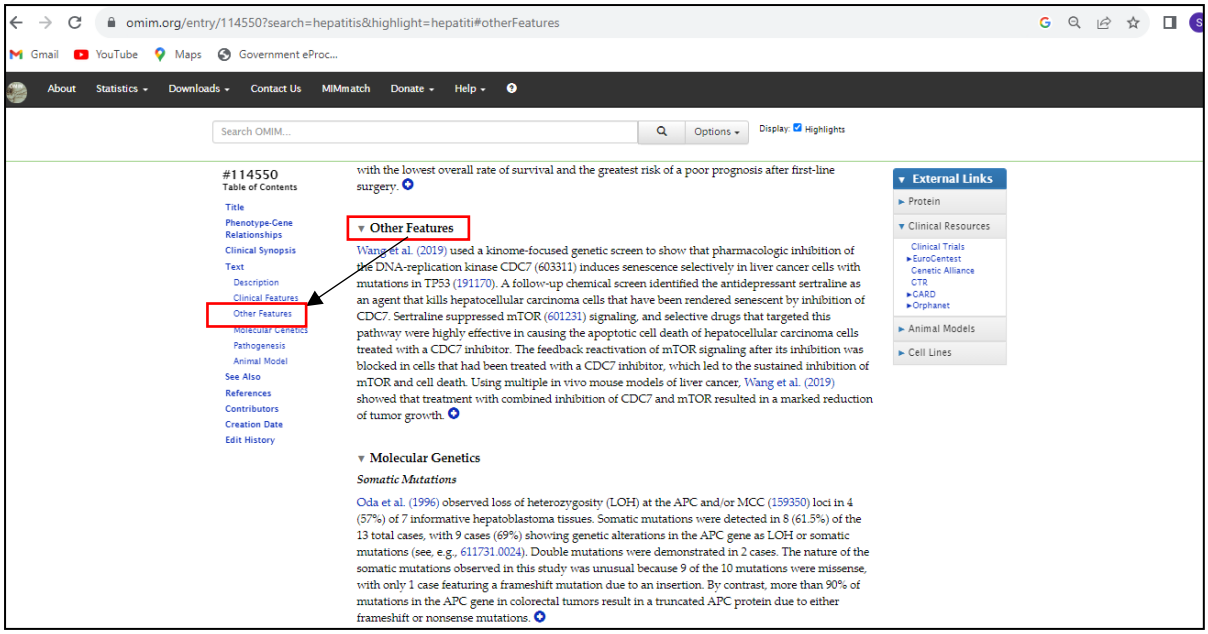


Figure 8: Result page of Other Features for the query

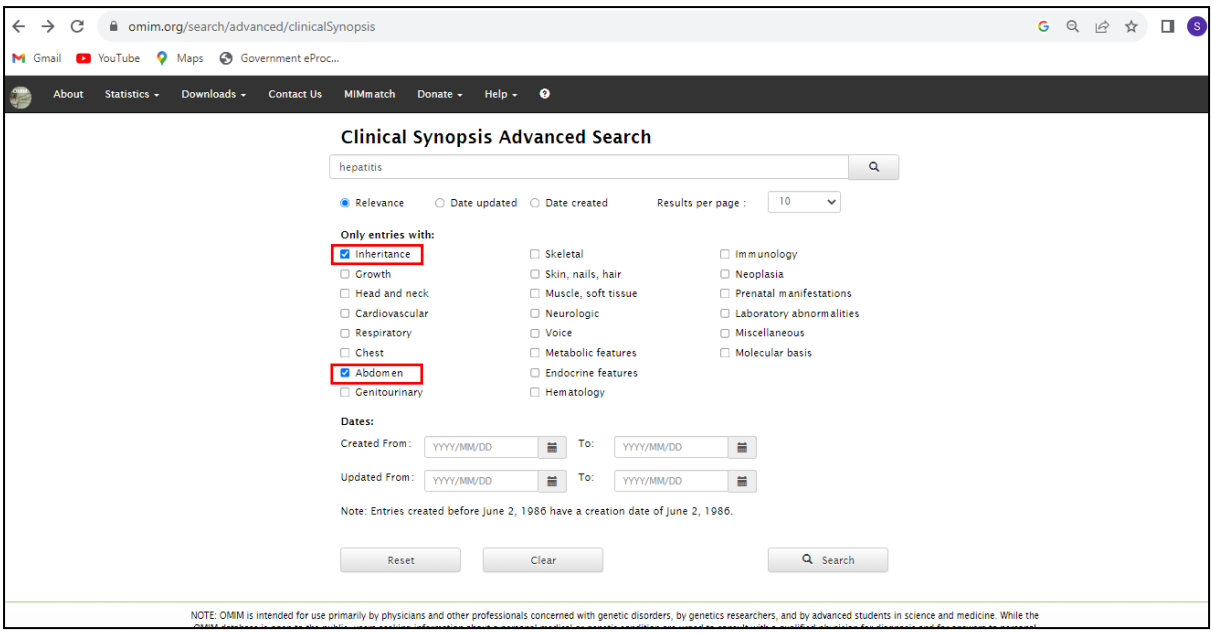


Figure 9: Clinical Synopsis Advanced Search

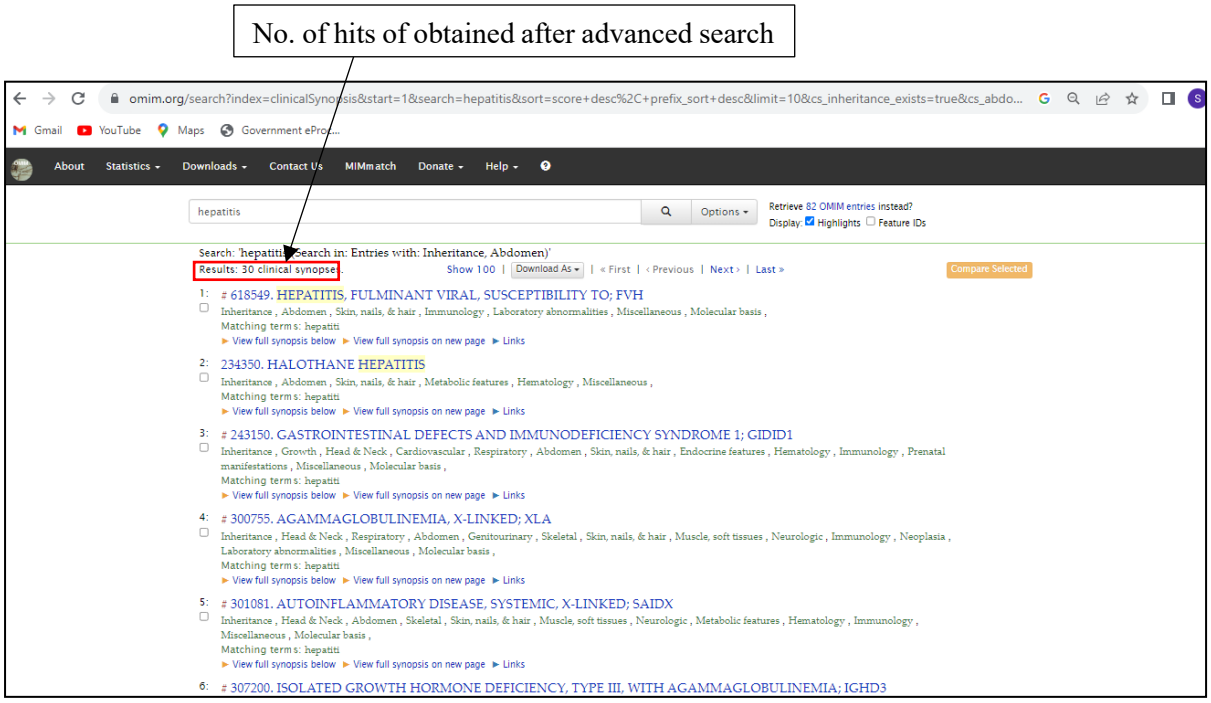


Figure 10: Result page of Clinical Synopsis Advanced Search for the query

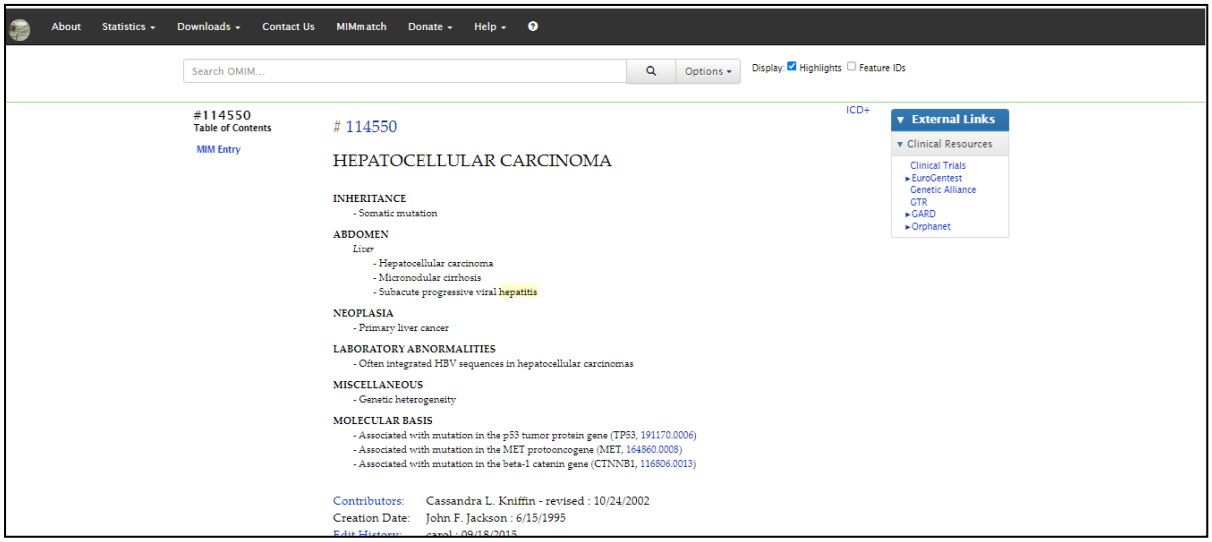


Figure 11: Result page of Clinical Synopsis for 'Hepatitis' in relevance of Inheritance and Abdomen

RESULTS:

The query 'Hepatitis' was searched 353 hits were obtained. In Text, 6 hits obtained which is Description, Clinical features, Other features, Molecular genetics, Pathogenesis, Animal model and other details about query. The clinical synopsis advanced search yielded 30 hits for only entries with inheritance and abdomen.

CONCLUSION:

The Online Mendelian Inheritance in Man (OMIM) database stands as a cornerstone in the field of medical genetics and genomics, providing a comprehensive and meticulously curated repository of information on human genes and genetic disorders. Since its inception, OMIM has been an invaluable resource for researchers, clinicians, and geneticists, offering a wealth of knowledge on the genetic basis of inherited diseases. OMIM remains an indispensable resource, playing a pivotal role in advancing our understanding of human genetic disorders. Its comprehensive, curated, and freely accessible data continue to empower researchers and healthcare professionals worldwide, driving progress in the diagnosis, treatment, and management of genetic diseases while laying the groundwork for personalized medicine.

REFERENCES:

1. Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, *33*(Database issue), D514–D517. https://doi.org/10.1093/nar/gki033
2. Amberger, J. S., Bocchini, C., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research, 43*(D1), D789–D798. https://doi.org/10.1093/nar/gku1205
3. World Health Organization: WHO. (2020, March 11). *Hepatitis*. https://www.who.int/healthtopics/hepatitis

DATE: 26/08/2023

WEBLEM 3

INTRODUCTION TO SEQUENCE DATABASES AND SUBMISSION TOOLS

INTRODUCTION:

In the field of bioinformatics, a sequence database refers to the type of biological database that is composed of a large collection of nucleotide nucleic acid sequences, protein sequences, or other polymer sequences stored on a computer. The utilization of sequence databases in the modern-day molecular biology, biotechnology, and bioinformatics has a profound influence on research by conserving time, energy and efforts.

These databases serve as structured archives of genetic and protein material, playing a critical role in facilitating research across several biological and medicinal disciplines. The databases encompass a diverse array of nucleotide and amino acid sequences, allowing researchers the chance to explore the genetic data of numerous entities and unravel fundamental biological mechanisms. Sequence databases encompass data in several different forms and patterns that are specifically tailored to meet specific requirements of the computational systems or programs.

In addition, these databases encompass genomic databases, which serve as repositories for entire genomes; nucleotide databases, which house individual DNA or RNA sequences; and protein databases, which contain amino acid sequences from an array of species from various different genera. The prominent examples of sequence databases involve GenBank, which is under the supervision of the National Center for Biotechnology Information (NCBI), EMBL-EBI (European Bioinformatics Institute), UniProt, a collaborative initiative involving EMBL-EBI, SIB (Swiss Institute of Bioinformatics), and PIR (Protein Information Resource), as well as DDBJ (DNA Data Bank of Japan). Several examples of sequence databases include the following:

1. **GenBank Database & Submission Tools**

   a. **GenBank Database:**
   GenBank Database plays a pivotal role in modern molecular biology, serving as an indispensable archive of annotated DNA sequences and providing essential resources for researchers worldwide. Additionally, GenBank actively contributes to the International Nucleotide Sequence Database Collaboration (INSDC), working in collaboration with the DNA Data Bank of Japan (DDBJ) and the European Nucleotide Archive (ENA). The National Centre for Biotechnology Information (NCBI) is responsible for the meticulously curation of the data stored in GenBank. The repository contains an extensive assortment of genetic data, encompassing a wide range of organisms and molecular mechanisms. The periodic releases of new data by GenBank serve to facilitate the ongoing expansion of this highly valuable resource, hence promoting scientific advancement and exploration in diverse areas of the biological sciences.

   b. **Submission Tools:**
   The data and results generated by numerous researches throughout the globe had to be made available for the study. With the intension of this, databases started to collect data from the researchers via their specified submission tools. The development of

submission tools has addressed researchers' demand for efficient data deposition, aiming to facilitate the seamless integration of new data. These tools enhance the process of submitting sequences, hence ensuring accurate curation and smooth integration into the databases. Various submission tools include:

1. **Submission Portal**, a comprehensive system accommodating many submission kinds, is focused on expanding its capabilities to encompass other forms of GenBank submissions.
2. **table2asn**, a command-line program, serves the purpose of automating the process of generating sequence records for submission to GenBank. It is specifically designed to be applicable for both annotated genomes and large sets of sequences. The software can be accessed over the File Transfer Protocol (FTP) on various operating systems, including MAC, PC, and Unix.
3. **Genome Workbench** software offers a wide range of tools that facilitate thorough genetic study. The Submission Wizard provided by the platform facilitates the process of submitting single eukaryotic and prokaryotic genomes. Additionally, this tool has the capability to modify and present ASN1 files generated by the table2asn software.
4. **BankIt** is an internet-based submission tool that incorporates interactive wizards to facilitate the process of submitting information.

   a. **BankIt:**
   The National Centre for Biotechnology Information (NCBI) has developed a well-known submission tool called BankIt, which serves as a crucial facilitator in accelerating the incorporation of genetic material into sequence databases. The web-based interface of BankIt is designed to accommodate the submission of genomic DNA, transcripts, and tiny genomes. This user-centric platform aims to streamline the submission process. The purposeful design of the system places a high priority on user-friendliness, ensuring that users are provided with clear and straightforward instructions for submitting data in a seamless manner. The accessibility of this platform is especially beneficial for researchers who wish to contribute individual or small groups of sequences, since it is supported by its strong annotation capabilities. The relevance of BankIt in expanding the frontiers of genetics and supporting collaborative research endeavors is underscored by its rise within the landscape of sequence submission methods. BankIt plays a vital role in facilitating the advancement of our comprehension of the intricate genetic aspects of life by virtue of its interconnection with sequence databases.

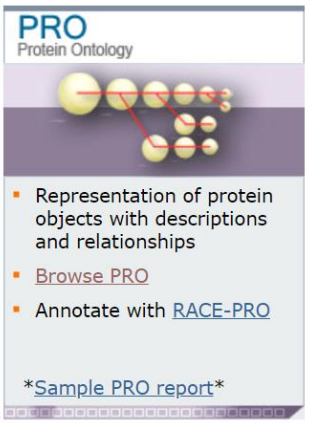
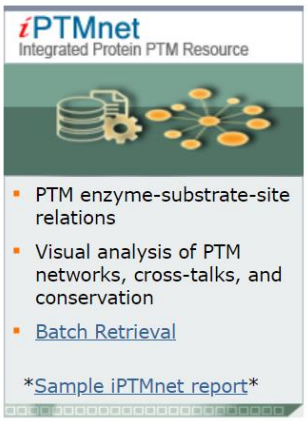
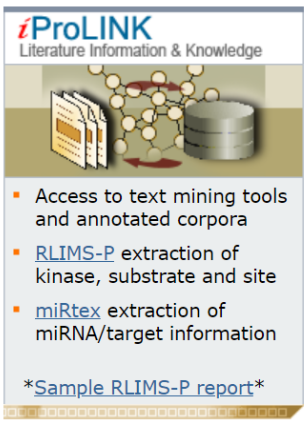
2. **European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL – EBI) Database:**
EMBL-EBI Database serves as a crucial center for bioinformatics research, providing a comprehensive range of integrated tools and resources that support investigations in molecular biology. Located on the Wellcome Genome Campus in close proximity to Cambridge, United Kingdom. By virtue of its comprehensive assortment of molecular databases, this platform enables researchers worldwide to investigate and unravel intricate biological data. The contributions of EMBL-EBI extend beyond the mere storing of data, as they also cover the supply of sophisticated analytical tools and comprehensive training workshops. These resources are designed to provide scientists with the necessary skills to effectively extract important insights from the vast amount of biological information available. EMBL-EBI serves as a dynamic hub that integrates the fields of

biology and informatics, facilitating scientific exploration and fostering advancements in several domains within the life sciences.

3. **UniProt Database: (SwissProt & Trembl):**
   UniProt Database, known as the Universal Protein Resource, plays a vital role in today's bioinformatics by providing a comprehensive and diligently maintained collection of protein sequence and functional information. UniProt is a helpful resource for researchers involved in proteomics and related investigations, which was established through a collaborative endeavor between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR). UniProt serves a wide range of research requirements through its three databases: the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The platform has undergone significant development and now serves as a reliable and authoritative resource for extensive protein-related data. This enables users to effectively investigate protein sequences, structures, functions, and annotations. UniProt's dedication to the principles of open access, precision, and cooperation enables the worldwide scientific community, facilitating advancements in diverse biological disciplines. UniProt encompassing Swiss-Prot - a curated database of annotated protein sequences, and TrEMBL - an automated resource for protein sequences awaiting curation. PIR's collaboration with global partners underscores its significance in advancing molecular research and providing comprehensive protein information.

4. **Protein Information Resource (PIR) Database:**
   The Protein Information Resource (PIR) Database is an essential bioinformatics resource that plays a major role in the advancement of genomic, proteomic, and systems biology research. The Protein Information Resource (PIR) was established in 1984 under the guidance of the National Biomedical Research Foundation (NBRF). Its primary objective has been to aid researchers in the analysis and interpretation of protein sequence data. The Protein Information Resource (PIR) has a notable historical background that can be traced back to the groundbreaking contributions made by Margaret O. Dayhoff. Over time, PIR has developed into a cohesive platform that offers a wide range of protein sequences and structured data, serving as a valuable resource for facilitating scientific research endeavors. PIR plays a vital role in the construction of UniProt, a definitive library of protein sequences and annotations, through collaboration with global partners such as the European Bioinformatics Institute (EMBL-EBI) and the SIB Swiss Institute of Bioinformatics. The organization's unwavering dedication to precise curation and its involvement as a contributor to the International Nucleotide Sequence Database Collaboration (INSDC) highlight its significant influence on contemporary bioinformatics and the wider scientific community.

Following are the 3 resources available on the PIR Database Homepage –

		
PRO (Protein Resource Ontology) PRO provides an ontological representation of protein-related entities by explicitly defining them and showing the relationships between them.	**iPTMnet (PTMs = Protein Post – Translational Modification)** iPTMnet connects multiple disparate bioinformatics tools and systems text mining, data mining, analysis and visualization tools, and databases and ontologies into an integrated cross – cutting research resource to address the knowledge gaps in exploring and discovering PTM networks.	**iproLINK (integrated Protein Literature Information and Knowledge)** iproLINK (integrated Protein Literature Information and Knowledge) is a resource with access to text mining tools and annotated corpora developed in house.

The text mining tools used to retrieve data from the PIR Database are:

1. **iTextMine:** An integrated text mining tools and relation extraction results from large–scale text processing.
2. **pGenN:** A gene normalization tool tailored for plants.
3. **miRTex:** A relation extraction tool that identifies miRNA – target relations as well as miRNA – gene and gene – miRNA regulation relations.
4. **eFTP:** A relation extraction tool that identifies information relevant to phosphorylated proteins and phosphorylation – dependent protein – protein interactions.
5. **emiRIT:** An integrative text mining system collecting miRNA information from the literature.

5. **DNA Data Bank of Japan (DDBJ) Database:**
   The DNA Data Bank of Japan (DDBJ) Database is a significant participant in the field of recent bioinformatics, effectively fulfilling its responsibilities as a member of the International Nucleotide Sequence Database Collaboration (INSDC) and making considerable contributions to the progress of genomics and life sciences research. DDBJ, which stands for DNA Data Bank of Japan, was established with the purpose of serving as a comprehensive repository for nucleotide sequence data. Its primary objective is to gather, organize, and distribute genetic information in order to support and enhance scientific investigation. The organization's dedication to facilitating worldwide research endeavors is apparent in its collaborative data sharing with esteemed partners, namely the European Nucleotide Archive (ENA) and GenBank. This collaboration serves to guarantee the widespread availability and ease of access to invaluable genetic resources. DDBJ, located in Japan, plays a crucial role in the advancement of our knowledge of the genetic underpinnings of life and the facilitation of international cooperation in the realm of bioinformatics through its ongoing contributions to the INSDC.

REFERENCES:

1. *GenBank Overview*. (n.d.). https://www.ncbi.nlm.nih.gov/genbank/
2. *About us*. (n.d.). EMBL-EBI. https://www.ebi.ac.uk/about
3. *UniProt*. (n.d.). https://www.uniprot.org/help/about
4. *History [PIR - Protein Information Resource]*. (n.d.). https://proteininformationresource.org/pirwww/about/
5. *About DDBJ Center*. (n.d.). https://www.ddbj.nig.ac.jp/about/index-e.html
6. *International Nucleotide Sequence Database Collaboration*. (n.d.). https://www.insdc.org/
7. *How to submit data to GenBank*. (n.d.). https://www.ncbi.nlm.nih.gov/genbank/submit/
8. *Submission Portal | NCBI | NLM | NIH*. (n.d.). https://submit.ncbi.nlm.nih.gov/

DATE: 26/08/2023

WEBLEM 3(A)(a)
GENBANK DATABASE
(**URL**: www.ncbi.nlm.nih/genbnk/)

AIM:
To explore GenBank Database for the query ABO gene (Accession ID – NC_008260.1).

INTRODUCTION:
GenBank Database is a database that contains publicly available nucleotide sequences for over 300,000 organisms. The database includes DNA sequences for more than 105,000 different organisms. GenBank Database contains data from:

1. Major DNA and protein sequence databases
2. Taxonomy
3. Genome
4. Mapping
5. Protein structure and domain information
6. Biomedical journal literature via PubMed

GenBank Database was created in 1979 at the Los Alamos National Laboratory. It was originally called the Los Alamos Sequence Database. In 1982, it was renamed GenBank and became a public database. GenBank Database is the most complete collection of annotated nucleic acid sequence data for almost every organism. The content includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms. There is also a GenPept database for protein sequences, the majority of which are conceptual translations from DNA sequences, although a small number of the amino acid sequences are derived using peptide sequencing techniques.

GenBank Database is a public database of all known nucleotide and protein sequences with supporting bibliographic and biological annotation, built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH). NCBI was created by Congress in 1988 to develop information systems, such as GenBank, to support the biomedical research community. NCBI was also mandated to conduct basic and applied research and, as part of the NIH Intramural Program, NCBI scientists work in areas of gene and genome analysis, computational structural biology, and mathematical methods for sequence analysis. NCBI builds GenBank Database primarily from the direct submission of sequence data from authors and secondarily from scanning the journal literature. A major source of data are bulk submissions of EST and other high-throughput data from sequencing centers. The data are supplemented by sequences from other public databases.

ABO Gene:
The ABO gene is located on chromosome 9 and has three alleles: A, B, and O. The ABO gene encodes proteins related to the ABO blood group system. The ABO gene indirectly encodes the ABO blood group antigens. The ABO gene determines the ABO blood group of an individual by modifying the oligosaccharides on cell surface glycoproteins.

The ABO gene has three main allelic forms:

1. A allele
2. B allele
3. O allele

The ABO system was discovered in 1900 by Landsteiner. It is one of the most important blood group systems in transfusion medicine.

The **A allele** produces α-1,3-N-acetylgalactosamine transferase (A-transferase), which catalyzes the transfer of GalNAc residues from the UDP-GalNAc donor nucleotide to the Gal residues of the acceptor H antigen, converting the H antigen into A antigen in A and AB individuals.

The **B allele** encodes α-1,3-galactosyl transferase (B-transferase), which catalyzes the transfer of Gal residues from the UDP-Gal donor nucleotide to the Gal residues of the acceptor H antigen, converting the H antigen into B antigen in B and AB individuals. Remarkably, the difference between the A and B glycosyltransferase enzymes is only four amino acids.

The **O allele** lacks both enzymatic activities because of the frameshift caused by a deletion of guanine-258 in the gene which corresponds to a region near the N-terminus of the protein. This results in a frameshift and thus of a truncated protein of only 117 amino acids. The truncated protein is unable to modify oligosaccharides which end in fucose linked to galactose. Thus, no A or B antigen is found in O individuals. This sugar combination is termed the H antigen. These antigens play an important role in the match of blood transfusion and organ transplantation Other minor alleles have been found for this gene.

METHODOLOGY:

1. Go to the GenBank database's homepage and choose the nucleotide option.
2. Use the Entrez search to look up the ABO gene.
3. Use limit filters to filter data, such as database sources, sequence type, molecule type, etc.
4. Use advanced search to get better outcomes.

OBSERVATIONS:

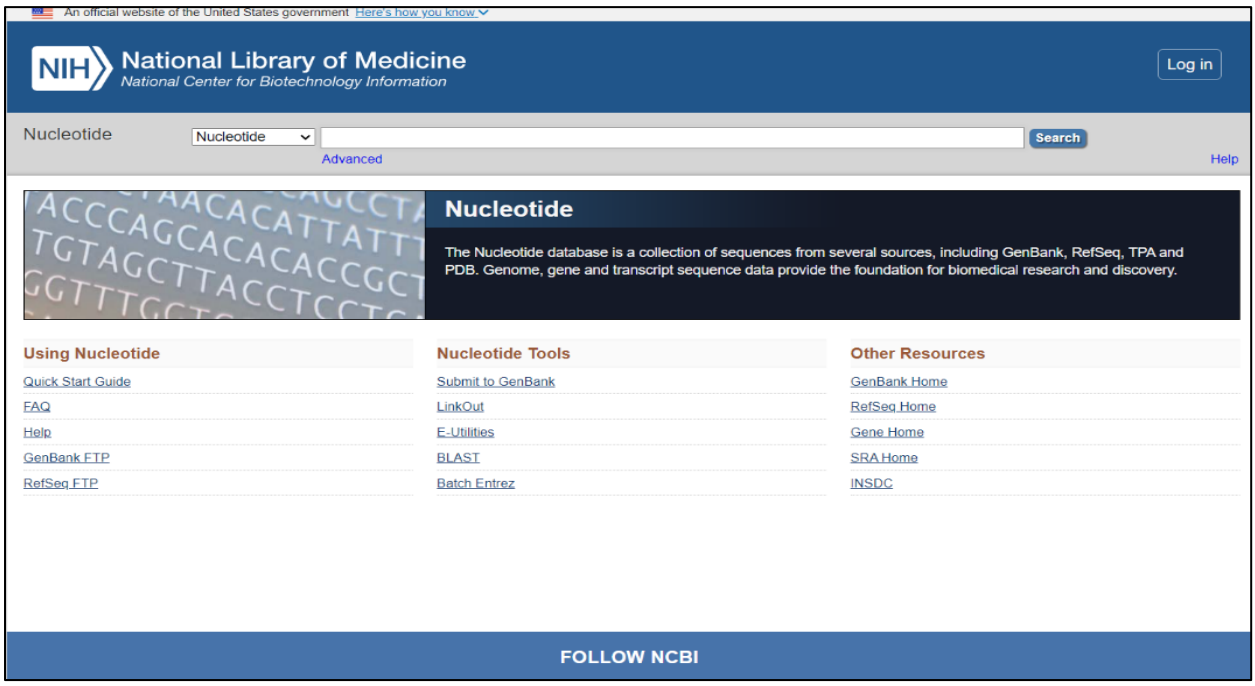


Figure 1: Homepage of GenBank Database

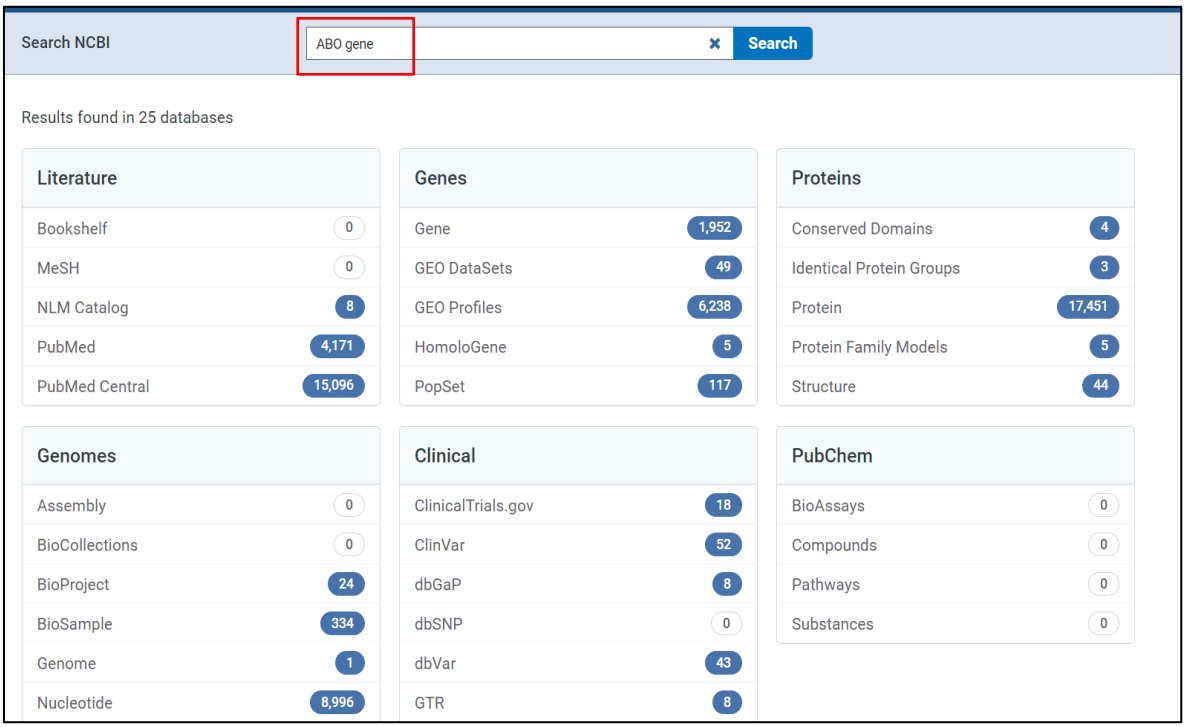


Figure 2: Searched ABO Gene using Entrez (All Databases)
In search of ABO gene show 25 databases including literature, gene, proteins, genomes, clinical, PubChem

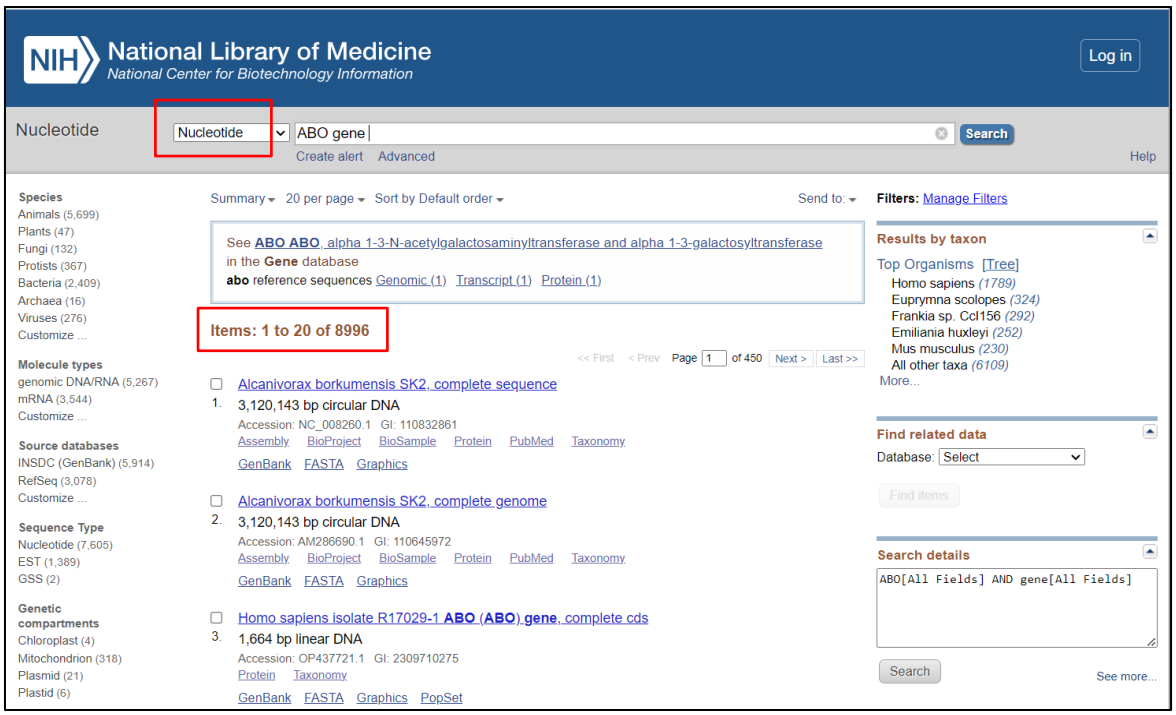


Figure 3: Searched for ABO Gene sequence identifiers and annotations with Entrez search (Category: Nucleotide).

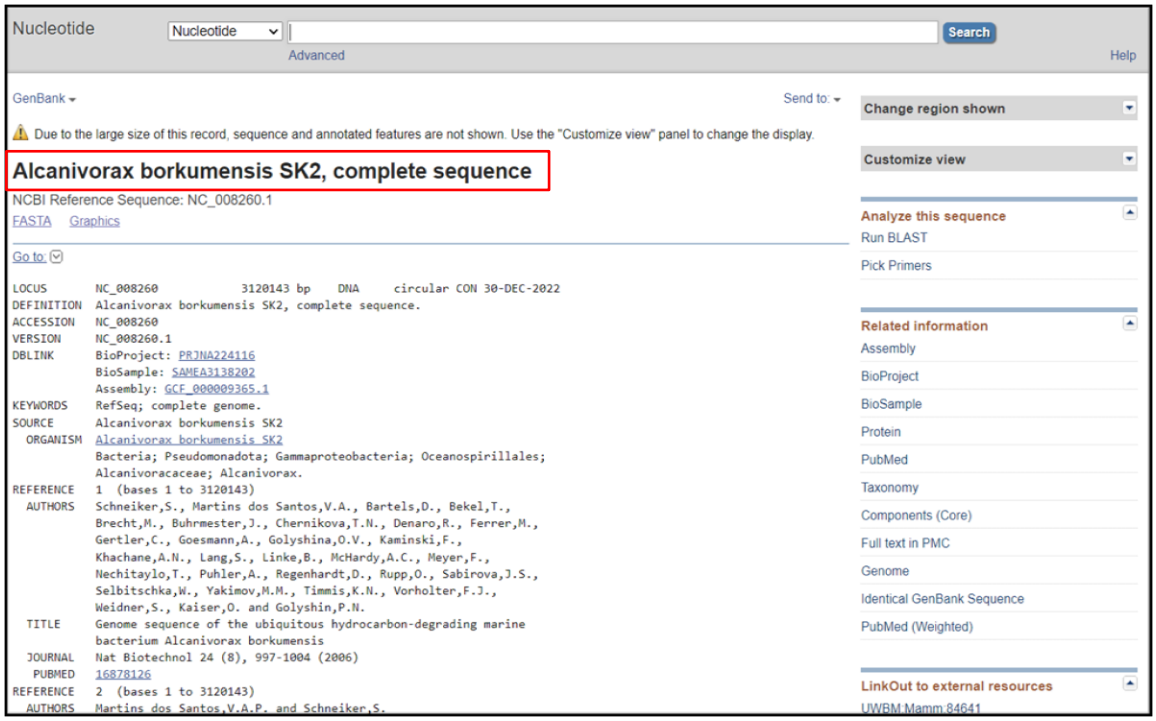
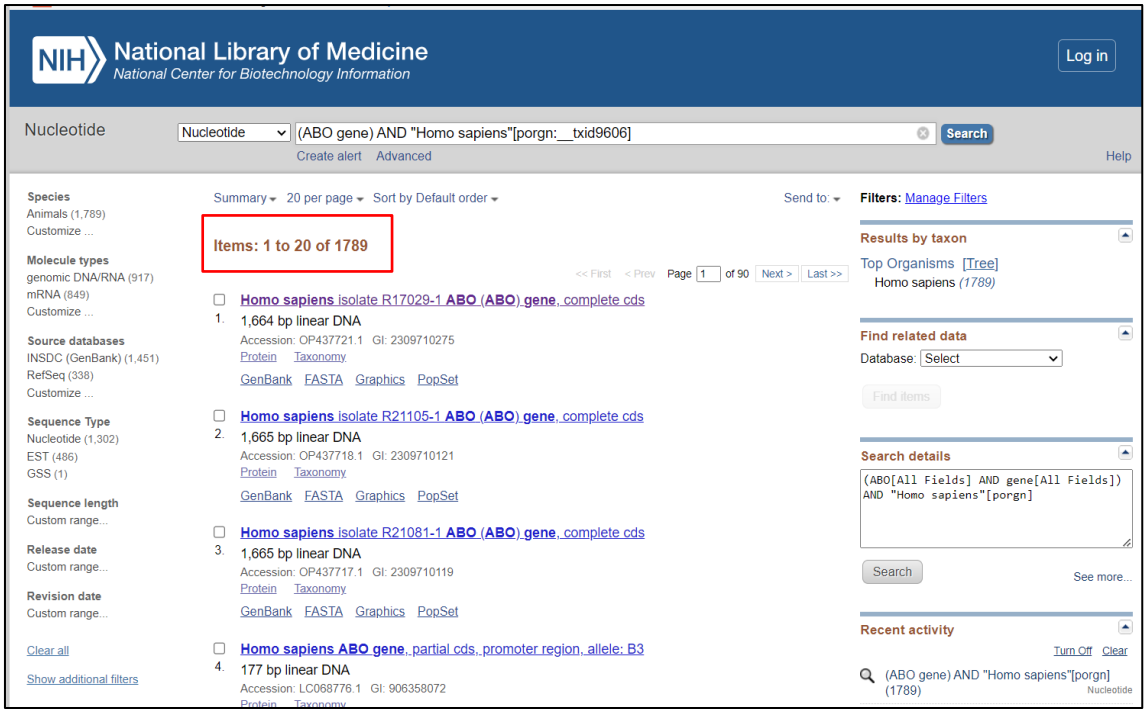


Figure 4: *Alcanivorax borkuumenis* (Organism) (Accession ID: NC_008260.1) Result using Nucleotide filter



(ABO gene) AND "Homo sapiens"[porgn:__txid9606]

Species
Animals (1,789)
Customize ...

Molecule types
genomic DNA/RNA (917)
mRNA (849)
Customize ...

Source databases
INSDC (GenBank) (1,451)
RefSeq (338)
Customize ...

Sequence Type
Nucleotide (1,302)
EST (486)
GSS (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

Items: 1 to 20 of 1789

<< First < Prev Page 1 of 90 Next > Last >>

☐ 1. **Homo sapiens** isolate R17029-1 **ABO** (**ABO**) **gene**, complete cds
1,664 bp linear DNA
Accession: OP437721.1 GI: 2309710275
Protein Taxonomy
GenBank FASTA Graphics PopSet

☐ 2. **Homo sapiens** isolate R21105-1 **ABO** (**ABO**) **gene**, complete cds
1,665 bp linear DNA
Accession: OP437718.1 GI: 2309710121
Protein Taxonomy
GenBank FASTA Graphics PopSet

☐ 3. **Homo sapiens** isolate R21081-1 **ABO** (**ABO**) **gene**, complete cds
1,665 bp linear DNA
Accession: OP437717.1 GI: 2309710119
Protein Taxonomy
GenBank FASTA Graphics PopSet

☐ 4. **Homo sapiens ABO gene**, partial cds, promoter region, allele: B3
177 bp linear DNA
Accession: LC068776.1 GI: 906358072
Protein Taxonomy

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Homo sapiens *(1789)*

Find related data
Database: Select
Find items

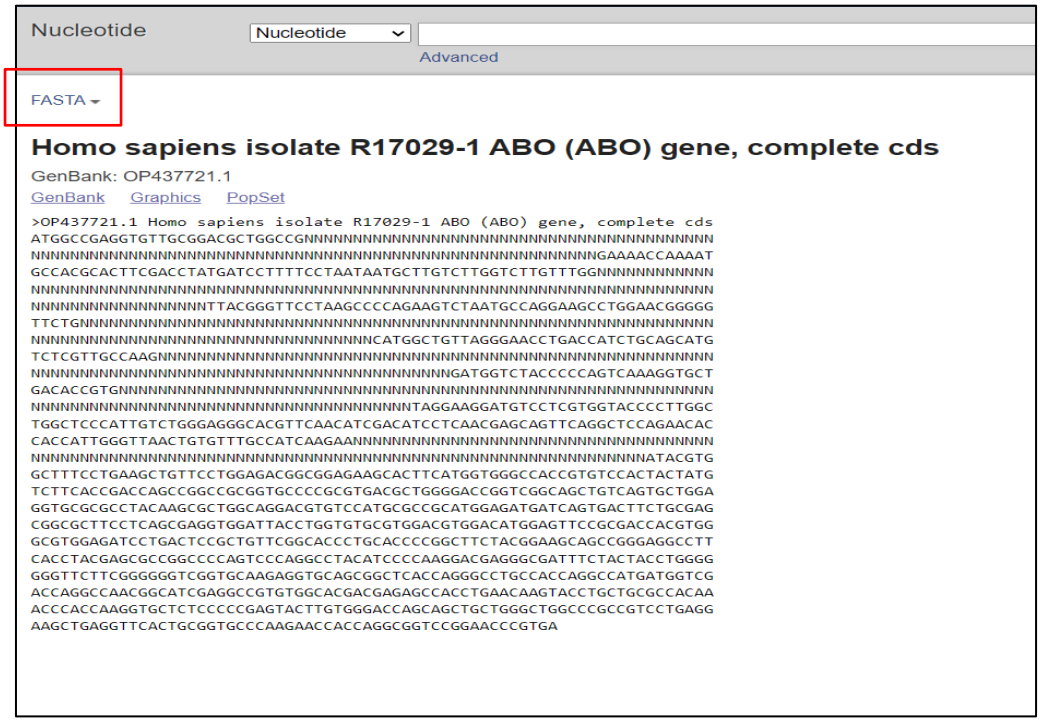
Search details

```
(ABO[All Fields] AND gene[All Fields])
AND "Homo sapiens"[porgn]
```

Search See more...

Recent activity
Turn Off Clear
(ABO gene) AND "Homo sapiens"[porgn]
(1789) Nucleotide

Figure 5: Result using Advance filter and *Homo sapiens* shows 1789 research articles



Homo sapiens isolate R17029-1 ABO (ABO) gene, complete cds

GenBank: OP437721.1

GenBank Graphics PopSet

```
>OP437721.1 Homo sapiens isolate R17029-1 ABO (ABO) gene, complete cds
ATGGCCGAGGTGTTGCGGACGCTGGCCGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGAAAACCAAAT
GCCACGCACTTCGACCTATGATCCTTTTCCTAATAATGCTTGTCTTGGTCTTGTTTGGNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNTTACGGGTTCCTAAGCCCCAGAAGTCTAATGCCAGGAAGCCTGGAACGGGGG
TTCTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCATGGCTGTTAGGGAACCTGACCATCTGCAGCATG
TCTCGTTGCCAAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGATGGTCTACCCCCAGTCAAAGGTGCT
GACACCGTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGGAAGGATGTCCTCGTGGTACCCCTTGGC
TGGCTCCCATTGTCTGGGAGGGCACGTTCAACATCGACATCCTCAACGAGCAGTTCAGGCTCCAGAACAC
CACCATTGGGTTAACTGTGTTTGCCATCAAGAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATACGTG
GCTTTCCTGAAGCTGTTCCTGGAGACGGCGGAGAAGCACTTCATGGTGGGCCACCGTGTCCACTACTATG
TCTTCACCGACCAGCCGGCCGCGGTGCCCCGCGTGACGCTGGGGACCGGTCGGCAGCTGTCAGTGCTGGA
GGTGCGCGCCTACAAGCGCTGGCAGGACGTGTCCATGCGCCGCATGGAGATGATCAGTGACTTCTGCGAG
CGGCGCTTCCTCAGCGAGGTGGATTACCTGGTGTGCGTGGACGTGGACATGGAGTTCCGCGACCACGTGG
GCGTGGAGATCCTGACTCCGCTGTTCGGCACCCTGCACCCCGGCTTCTACGGAAGCAGCCGGGAGGCCTT
CACCTACGAGCGCCGGCCCCAGTCCCAGGCCTACATCCCCAAGGACGAGGGCGATTTCTACTACCTGGGG
GGGTTCTTCGGGGGGTCGGTGCAAGAGGTGCAGCGGCTCACCAGGGCCTGCCACCAGGCCATGATGGTCG
ACCAGGCCAACGGCATCGAGGCCGTGTGGCACGACGAGAGCCACCTGAACAAGTACCTGCTGCGCCACAA
ACCCACCAAGGTGCTCTCCCCCGAGTACTTGTGGGACCAGCAGCTGCTGGGCTGGCCCGCCGTCCTGAGG
AAGCTGAGGTTCACTGCGGTGCCCAAGAACCACCAGGCGGTCCGGAACCCGTGA
```

Figure 6: ABO Gene Result sequence can be downloaded in FASTA format

RESULTS:

The searched result of ABO gene is indirectly encoding the ABO blood group antigens. ABO locus has three main allelic form A, B and O. On searching ABO gene by using nucleotide shows 8971 results from 25 databases. Top results using nucleotide shows *Alcanivorax borkumensis* sk2, complete sequence of 3,120,143 base pair circular DNA. The GI for this accession is 110838861. The organism is *Alcanivorax borkumeneis*, a gammaproteobacterial, pseudomonas, and oceanospirillales.

Results from the *Homo sapiens* filter return 1789 results. Humans extracted 1664 kb of linear DNA containing the ABO gene R17029-1. (Accession number: OP437721.1 and GI: -23209710275).

The output sequence is translated to the FASTA format, which includes the header, features, and origin. The header displays a summary, and the features display information on the order. With some gaps and variances, it contains genes, mRNA, and CDs. Origin displays the sequences' sources.

CONCLUSION:

GenBank Database is explored for query ABO gene and related information is searched.

REFERENCES:

1. *ABO ABO, alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase [Homo sapiens (human)] - Gene - NCBI*. (n.d.). https://www.ncbi.nlm.nih.gov/gene/28
2. Wikipedia contributors. (2022, October 18). ABO (gene). In *Wikipedia, The Free Encyclopedia*. Retrieved 00:17, October 6, 2023, from https://en.wikipedia.org/w/index.php?title=ABO_(gene)&oldid=1116756688
3. *National Center for Biotechnology Information*. (n.d.). https://www.ncbi.nlm.nih.gov/
4. *GenBank Overview*. (n.d.). https://www.ncbi.nlm.nih.gov/genbank/

DATE: 26/08/2023

WEBLEM 3(A)(b)
SUBMISSION TOOLS
(**URL**: https://www.ncbi.nlm.nih.gov/WebSub/)

AIM:
To submit eukaryotic and prokaryotic genome sequence in BankIt submission tool.

INTRODUCTION:
BankIt is an internet-based submission tool that incorporates interactive wizards to facilitate the process of submitting information. Submission Portal, a comprehensive system accommodating many submission kinds, is focused on expanding its capabilities to encompass other forms of GenBank submissions.

Eukaryotic cells form more complex and larger organisms. They have a nuclear membrane that comprises a nucleus. Eukaryotic cells can thrive in and maintain multiple environments as part of a single cell- a characteristic that helps them grow larger as compared to prokaryotic cells and also facilitates metabolic reactions. Some examples of eukaryotic cells are plants, animals, protists, and fungi. The Genetic material of Eukaryotic cells is structured in chromosomes. Golgi apparatus, Mitochondria, Ribosomes, and Nucleus are the parts of the eukaryotic cell. Animals, plants, fungi, and protozoa have eukaryotic cells and are classified under the Eukaryota kingdom.

Prokaryotic cells are single-celled microorganisms and include archaea and bacteria. These cells usually live freely by themselves or can be found in the gut of other organisms. The cells have a single membrane and consist of cytoplasm. Certain prokaryotic cells perform photosynthesis with the help of the cyanobacteria inside them.

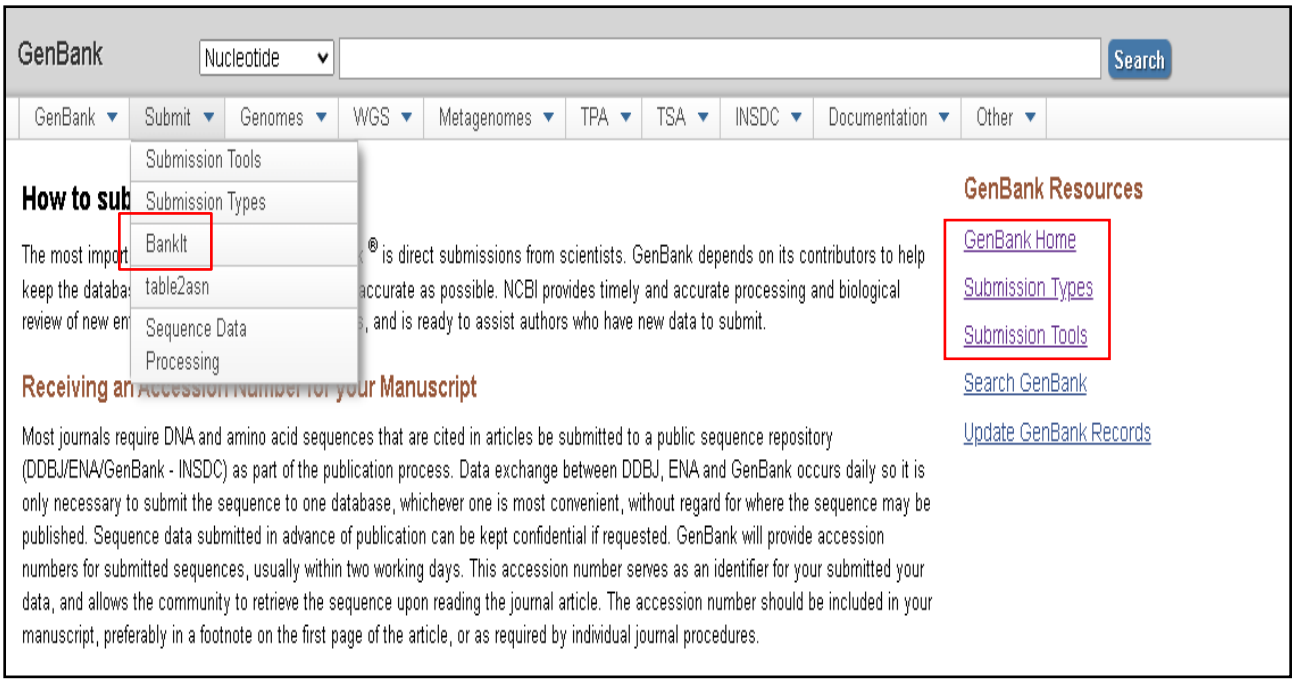


Figure 1: Homepage of GenBank Database

In the Genbank Database, 5 types of submit options such as submission tool, submission types, BankIt table2asn, sequence data processing are provided. Out of which, BankIt is used for submission of sequences.

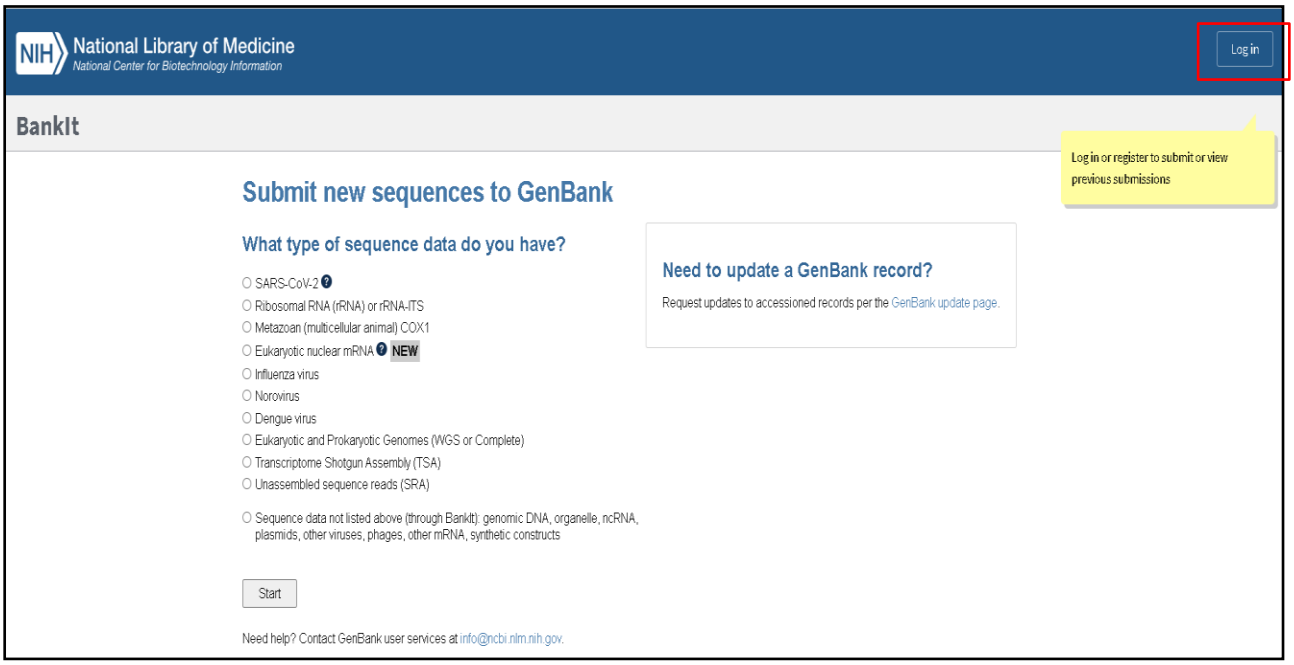


Figure 1a: Open BankIt option and create and NCBI Login for submission.

Figure 1b: NCBI Account Login Option

In order to do the submission, either create an NCBI account new login or use the existing credentials.

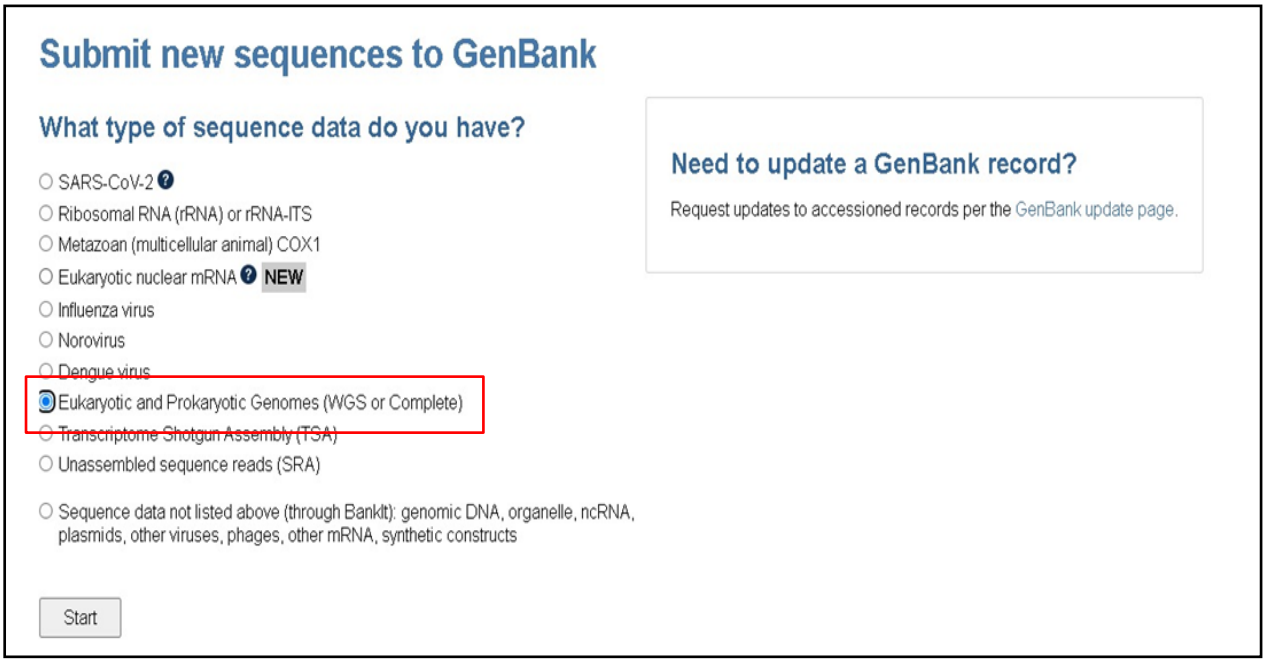


Figure 1c: Select the submission type.

Here selected the option as: Eukaryotic and prokaryotic genomes (WGS or Complete) submission.

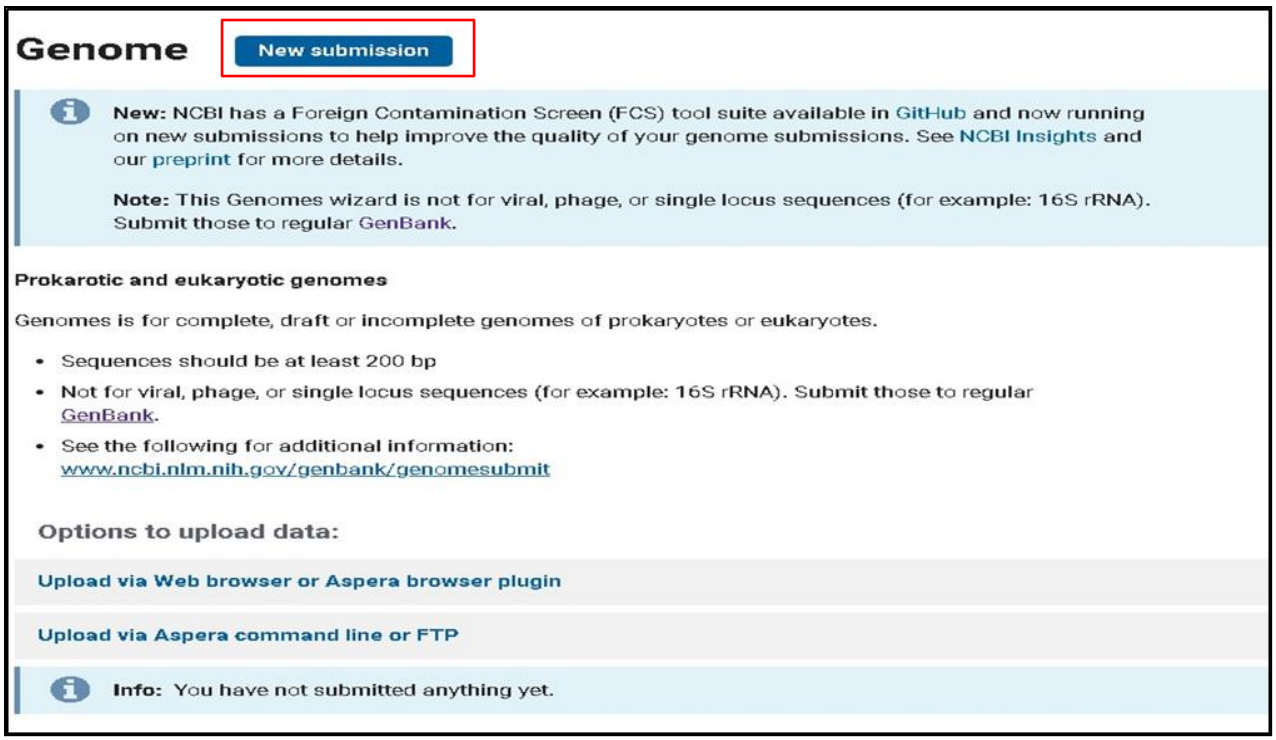


Figure 1d: Select on new submission option to continue the process

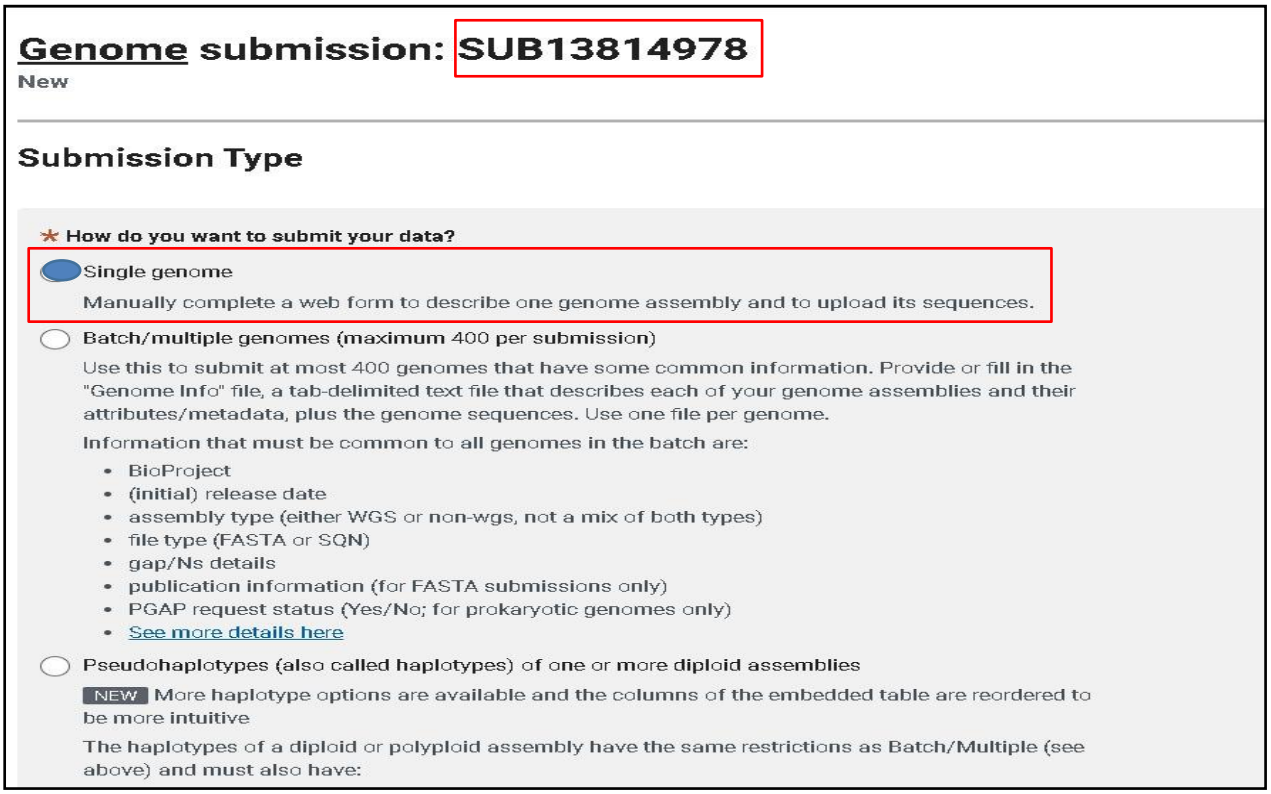


Figure 1e: Option for submission type

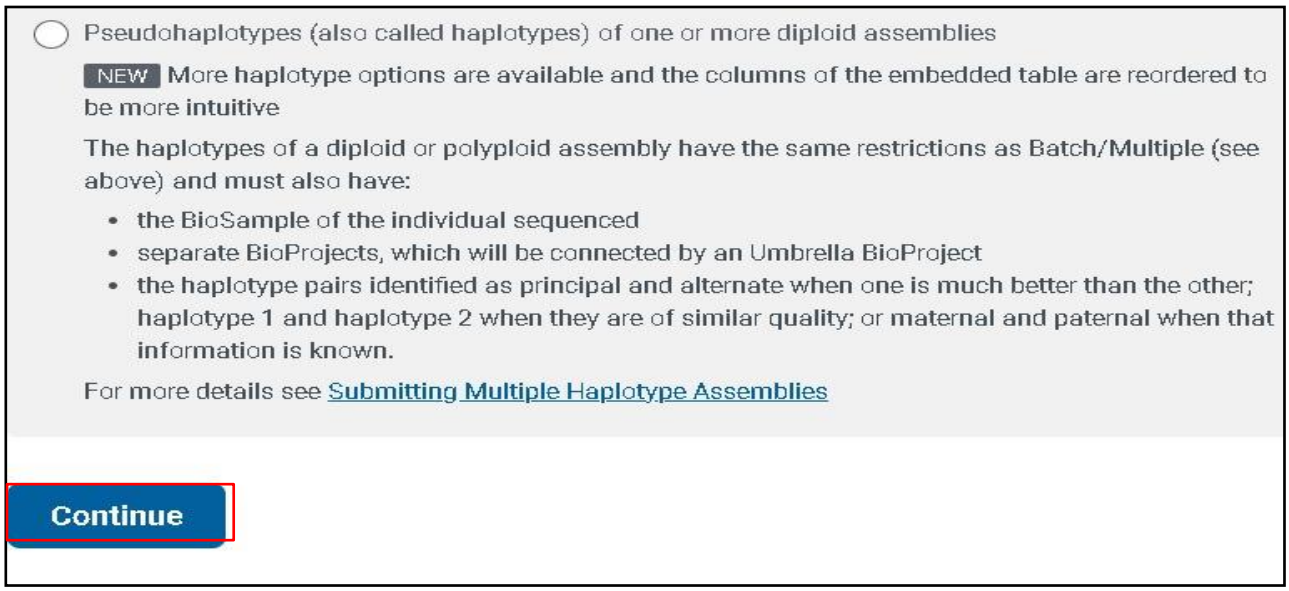


Figure 1f: Click on Continue option

After selecting a new submission option, genome data gets a new submission code and then 3 submission types panel is viewed which includes: Single genome, Batch/Multiple genomes, and Pseudo Haplotypes of one or more diploid assemblies. As per the requirement select the option. For the study Single genome submission option is being used. After the selection hit the continue button.

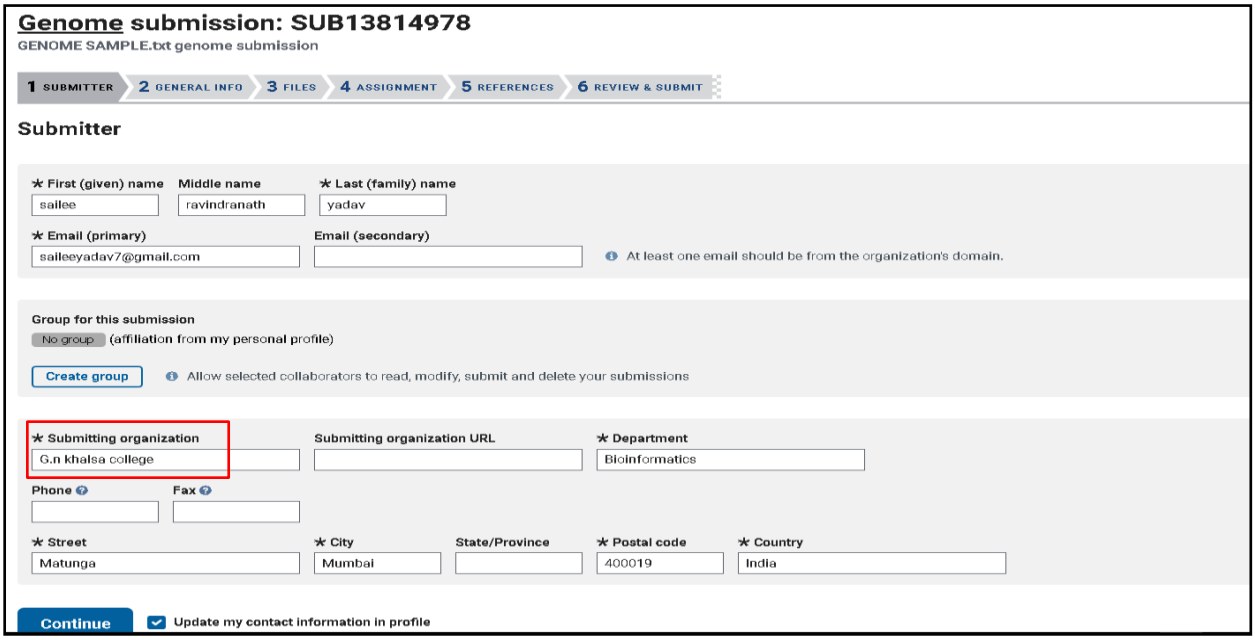


Figure 2: Different option for submission: STEP 1: **SUBMITTER Information**

STEP 1: SUBMITTER Information: The panel is provided to fill the details of the researchers such as first name, middle name, last name, and email ids (primary & secondary). Further the details of the researcher organization has to be provided and then click on the continue button.

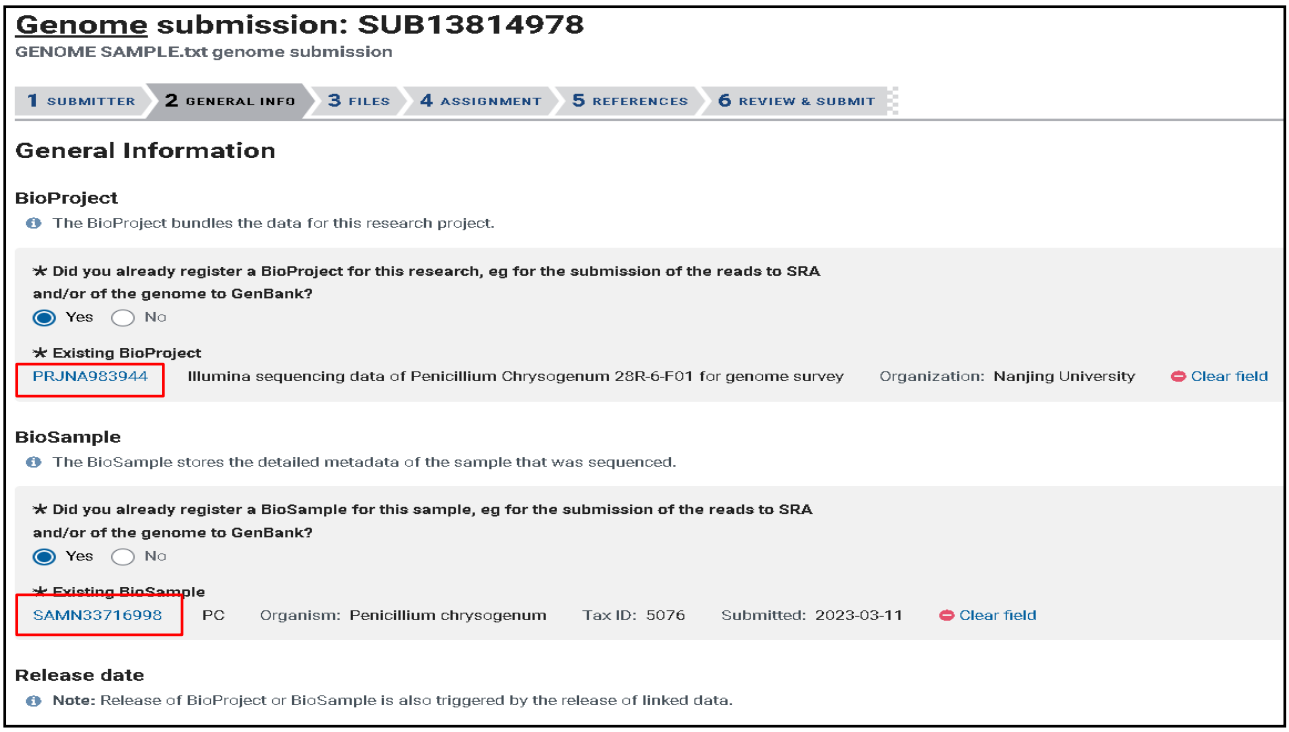


Figure 3: Different option for submission: STEP 2: **GENERAL Information**

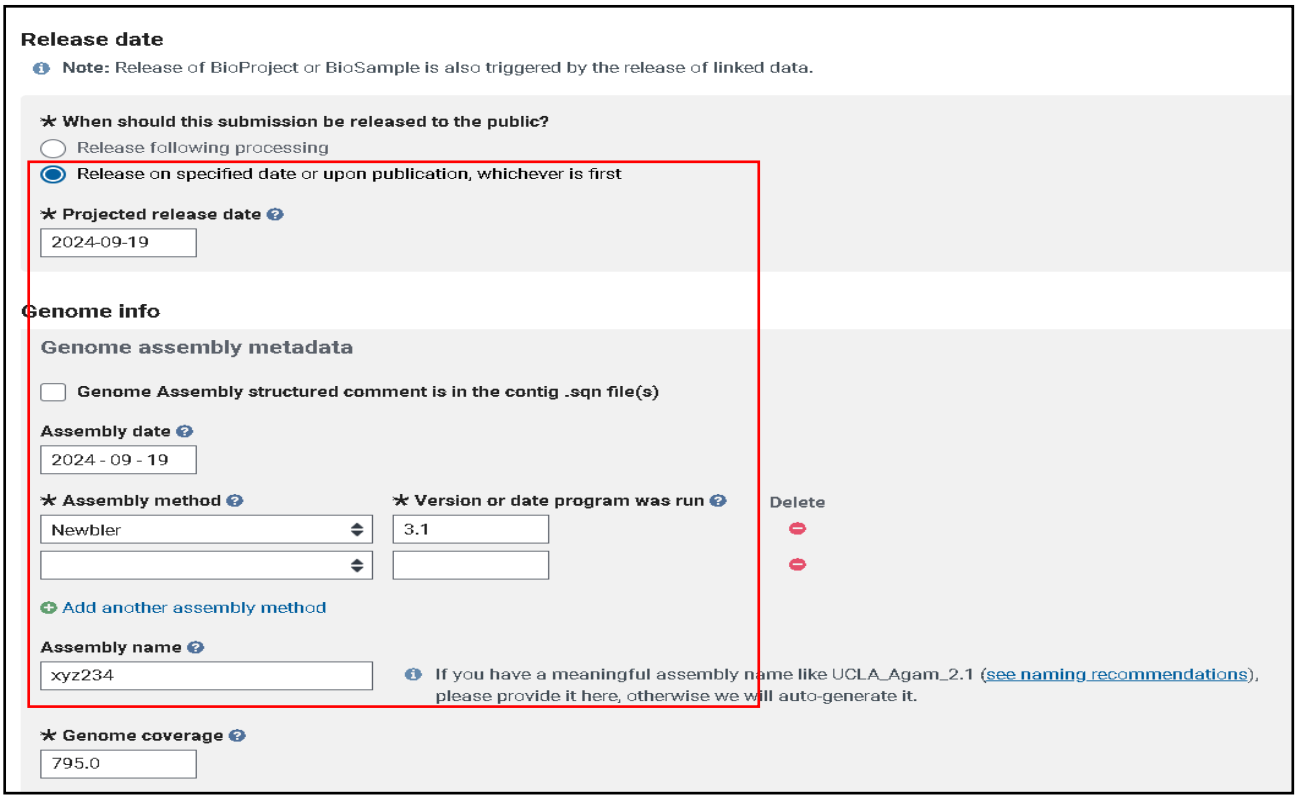


Figure 3a: Different option for submission: STEP 2: **GENERAL Information**

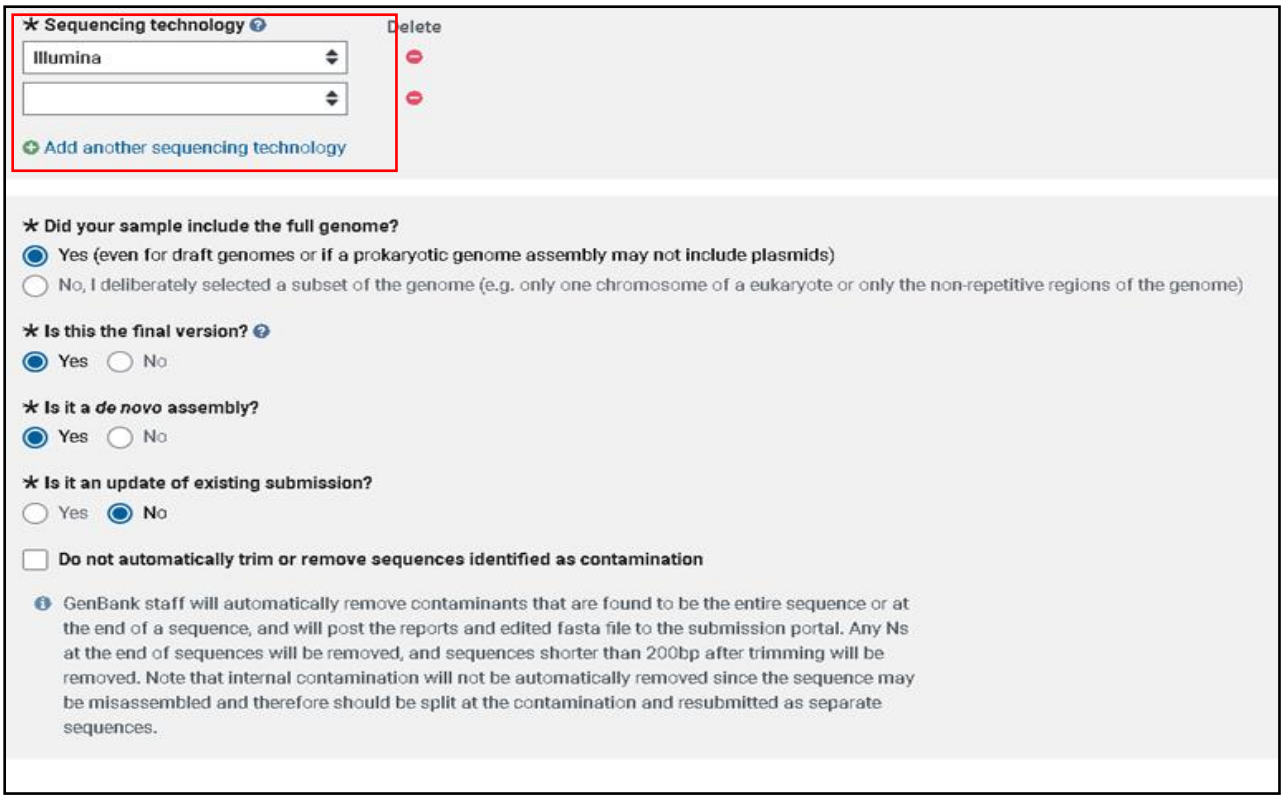


Figure 3b: Different option for submission: STEP 2: **GENERAL Information**

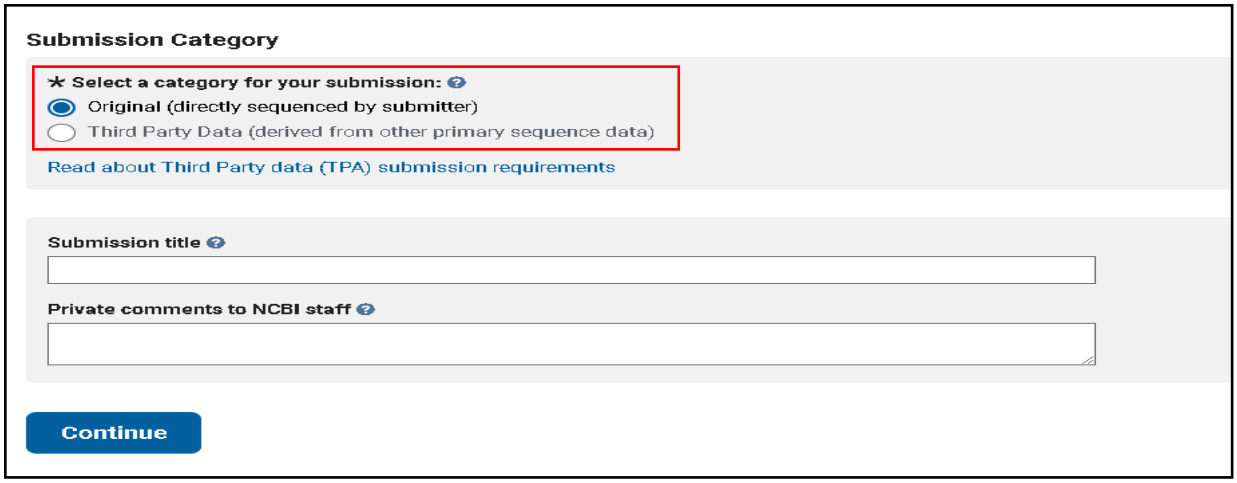


Figure 3c: Different option for submission: STEP 2: **GENERAL Information**

STEP 2: **General Information:** The researcher has to provide the Bio Project as well as BioSample Accession id for further submission process. Also requires to select whether the project submission should be released to the public or to be released on a specific date then mention the date. Further provide the Genome information such as assembly metadata, assembly date, assembly method along with the version of program used and later the name of the assembly with query coverage information. The tool also needs the sequencing method information as well as the other details related to the same. Later on, Select the submission category as original or third party and continue the step.

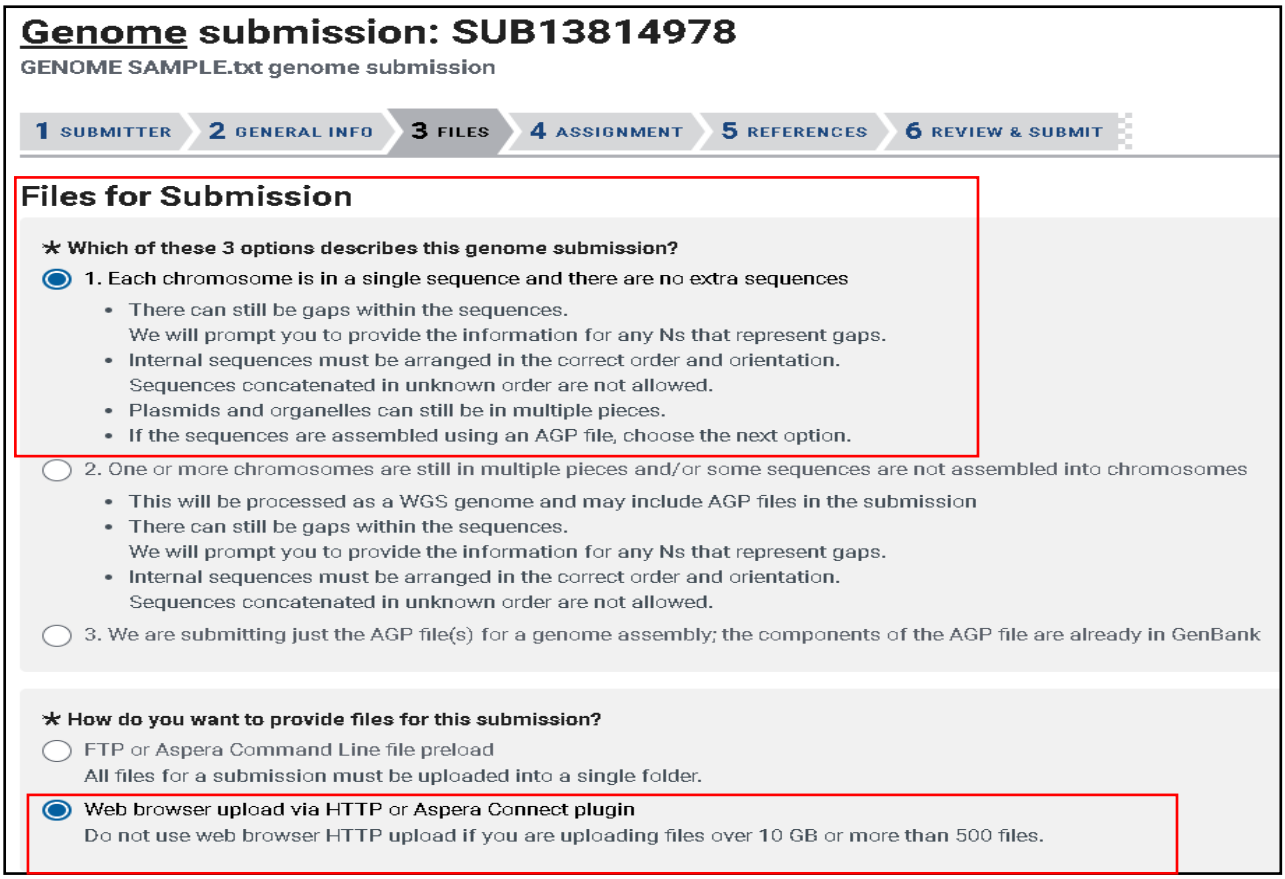


Figure 4: Different option for submission: STEP 3: **FILES**

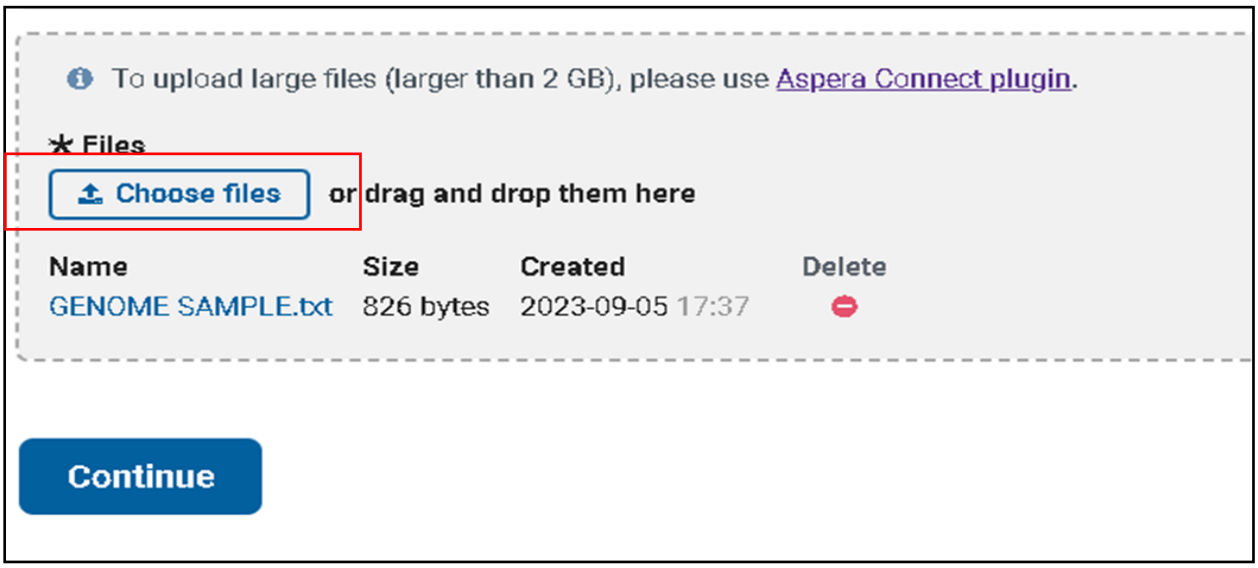


Figure 4a: Upload the genome sequence file of eukaryotic and continue:
STEP 3: **FILES**

FASTA

```
>seq1
ATGGCCGCTGTCCAGGGTGCTATCTCCAAGCGCAGGAAGTTCGTCGCCGACGGTGTCTTCTACGCCGAGC
TGAACGAGTTCTTCCAGCGCGAGCTCGCTGAGGAGGGCTACTCCGGTGTCGAAGTCCGTGTCACTCCCAC
CGTCACCGACATCATCATCCGTGCCACCCACACCCAGGAGGTTCTCGGCGAGCAGGGCCGCCGCATCCGT
GAGCTCACCTCGCTCATCCAGAAGCGTTTCAAGTTCCCCGAGAACTCGGTCTCCCTCTATGCCGCCAAGG
TCCAGAACCGCGGTCTGTCCGCCGTCGCTCAGTGCGAGTCCCTCCGCTACAAGCTCCTGAACGGTCTCGC
CGTCCGCCGTGCCTGCTACGGTGTCCTCCGCTTCATCATGGAGTCCGGTGCCAAGGGTTGCGAGGTTGTT
GTTTCCGGCAAGCTCCGTGCCGCCCGTGCTAAGTCCATGAAGTTCACTGACGGCTTCATGATCCACTCCG
GTCAGCCCGCCAAGGAGTTCATTGACTCCGCCACCCGCCACGTTCTCCTCCGCCAGGGTGTCCTTGGTAT
CAAGGTCAAGATCATGCGCGGCTCCGACCCCGAGGGCAAGTCCGGCCCCCAGAAGACCCTCCCCGACTCG
GTCACCATCATCGAGCCCAAGGAGGAGCAGCCCGTTCTCCAGCCCATGAGCCAGGACTACGGTGCCAAGG
CCATTGCCGCCCAGCAGCTCGCTGAGCAGCAGCGTCTGGCTGAGCAGCAGGCCGGTGAGGCTGAGGGTGG
TGCCGAGGGTTACGCTCAGGAGTAA
```

Figure 4b: The eukaryotic genome sequence file: STEP 3: **FILES**

STEP 3: FILES: The files for submission of genome has to be provided in the given step. Select the option as per the requirement. and upload the files using FTP or HTTP option. Sequence file submission will be done via uploading txt file and click on Continue option.

Figure 5: Different option for submission: STEP 4: **ASSIGNMENT**

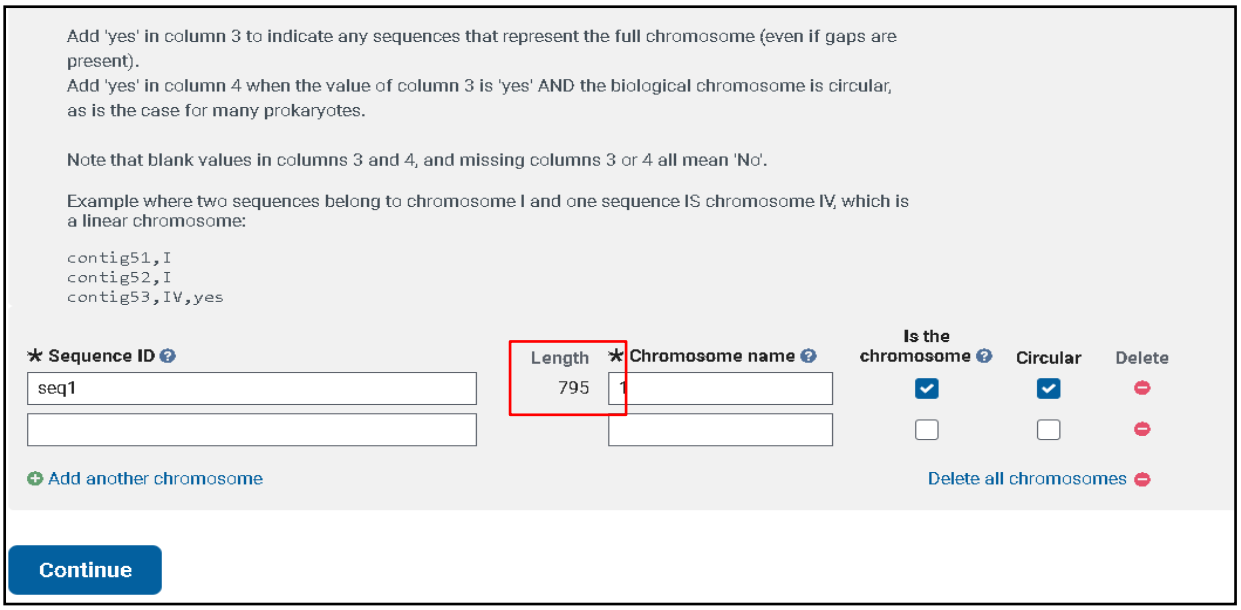


Figure 5a: Different option for submission: STEP 4: **ASSIGNMENT**

STEP 4: ASSIGNMENT: In this select if the genome sequence belongs to any organelle or plasmid select accordingly and further provide the sequence information, its length as well as chromosome name and select the further option as per requirement. Then click on continue.

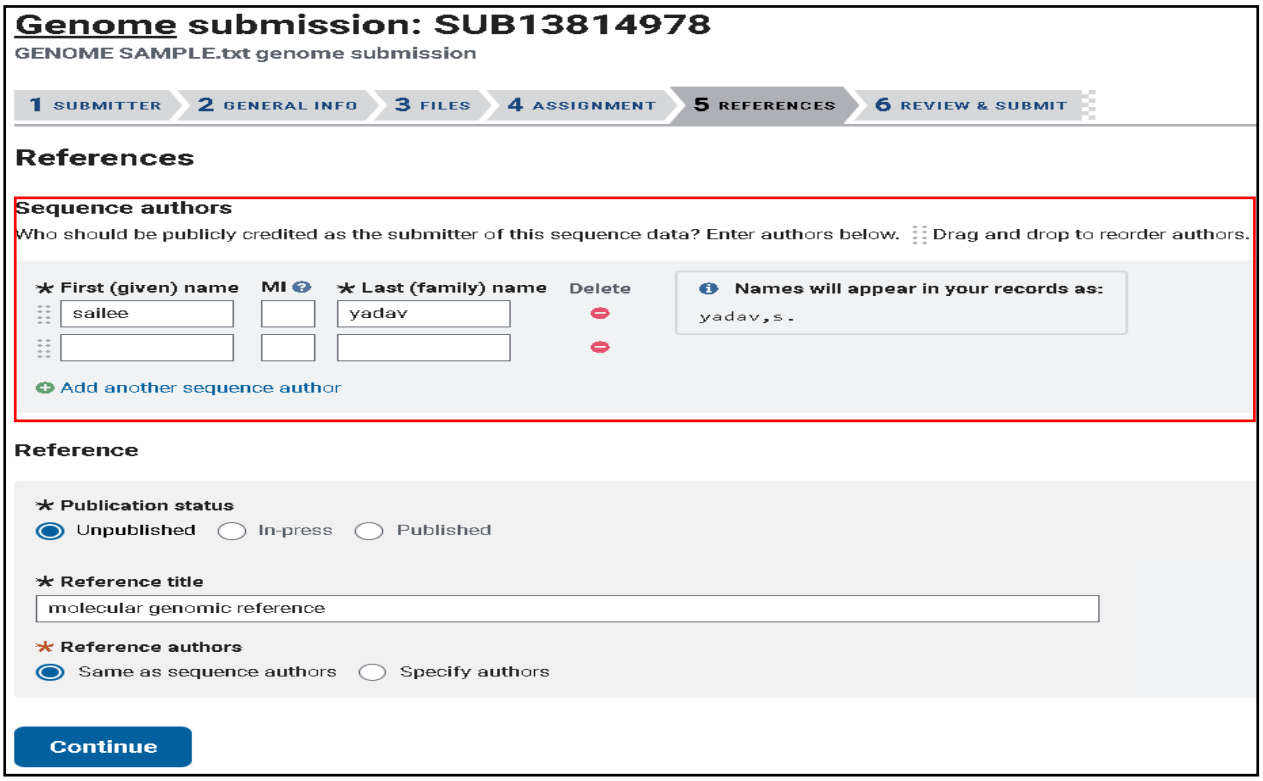


Figure 6: Different option for submission: STEP 5: **REFERENCES**

STEP 5: REFERENCES: The panel provides options for submitting the author information and decides the sequence in which it has to be displayed in the submission. Also, the publication status, title and author sequence confirmation have to be provided. Then click on Continue.

1 SUBMITTER | 2 GENERAL INFO | 3 FILES | 4 ASSIGNMENT | 5 REFERENCES | 6 REVIEW & SUBMIT

Review & Submit

This WGS submission will be released on **2024-09-19** or upon publication, whichever is first.
Note: Release of BioProject or BioSample is also triggered by the release of linked data.

Submitter

Submitter	sailee yadav saileeyadav7@gmail.com
Submitting organization	G.n khalsa college
Department	Bioinformatics
Street	Matunga
City	Mumbai
State/Province	None
Postal code	400019
Country	India

General Information

BioProject ID	PRJNA983944
BioSample ID	SAMN33716998
Genome assembly structured comment is in the contig .sqn file	No
Assembly date	2024 - 09 - 19

Figure 7: Different option for submission: STEP 6: **REVIEW AND SUBMIT**

Assembly date	2024 - 09 - 19
Assembly methods	Newbler 3.1
Assembly name	xyz234
Genome coverage	795.0
Sequencing technologies	Illumina
Did your sample include the full genome?	yes
Is this the final version?	yes
Is it *ade novo*assembly?	yes
Is it an update of existing submission?	no
GenBank will remove detected contamination, if possible	Yes

Files

Complete genome	yes
FASTA contigs	GENOME SAMPLE.txt

Assignment

Chromosomes

Sequence ID	Chromosome name	Complete	Circular
seq1	1	yes	yes

Figure 7a: Different option for submission: STEP 6: **REVIEW AND SUBMIT**

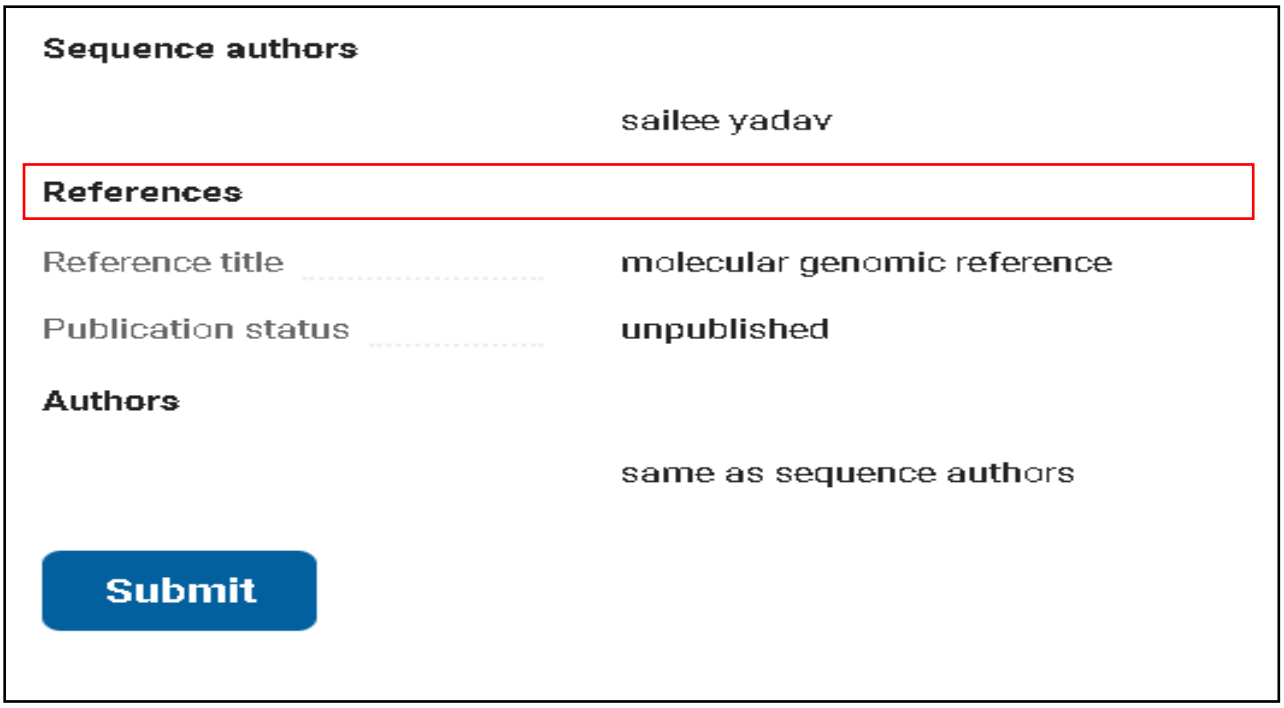


Figure 7b: Different option for submission: STEP 6: **REVIEW AND SUBMIT**

STEP 6: REVIEW AND SUBMIT: This is the final submission step where data needs to be reviewed and finalized. If no more corrections, then provide for submission.

RESULTS:

BankIt tool was explored and learned various steps for eukaryotic and prokaryotic genome submission.

CONCLUSION:

The genome sequence submission tool BankIt plays a pivotal role in advancing our understanding of genetics and genomics. It simplifies the process of sharing and archiving genetic data, making it accessible to researchers and scientists worldwide. This tool streamlines data management, enhances collaboration, and promotes transparency within the scientific community. As we continue to explore the intricacies of the genome, the genome sequence submission tool remains an indispensable resource for accelerating scientific discoveries and driving progress in fields such as medicine, agriculture, and evolutionary biology. Its continued development and widespread adoption are crucial for unlocking the full potential of genomics in addressing complex biological questions and improving human and environmental health.

REFERENCES:

1. National Institutes of Health, Office of Extramural Research (December 3, 1997). NIH Guide Using The TOC Notification LISTSERV, National Institutes of Health.
2. Enguita, F. J., Pereira, S., & Leitão, A. L. (2023, March 27). Transcriptomic Analysis of Acetaminophen Biodegradation by Penicillium chrysogenum var. halophenolicum and Insights into Energy and Stress Response Pathways. *Journal of Fungi*, *9*(4), 408. https://doi.org/10.3390/jof9040408
3. Petersen, C., Sørensen, T., Nielsen, M. R., Sondergaard, T. E., Sørensen, J. L., Fitzpatrick, D. A., Frisvad, J. C., & Nielsen, K. L. (2023, February 1). Comparative genomic study of the Penicillium genus elucidates a diverse pangenome and 15 lateral gene transfer events. *IMA Fungus*, *14*(1). https://doi.org/10.1186/s43008-023-00108-7

DATE: 26/08/2023

WEBLEM 3(B)
EUROPEAN MOLECULAR BIOLOGY LABORATORY – EUROPEAN BIOINFORMATICS INSTITUTE (EMBL – EBI) DATABASE (URL - https://www.ebi.ac.uk/)

AIM:
To explore the EMBL – EBI (European Molecular Biology Laboratory – European Bioinformatics Institute) database in terms of basic search and further study of the query angiotensinogen (Accession ID: P01019) under various categories.

INTRODUCTION:
EMBL – EBI (European Molecular Biology Laboratory – European Bioinformatics Institute) Database is a primary nucleotide sequence database in Europe. It is a biological database that houses data spanning from genomics, proteins, expression, small molecules, protein structures, systems, ontologies and scientific literature.
Single cell read data can be submitted in the form of BAM, CRAM or Multi – FASTQ formats. Genome Assembly data files can be submitted in the form of flat files that adhere to the ENA's set of documented guidelines.
Data from EMBL can be retrieved by –

1. Using accession numbers (unique identifiers)
2. Sequence identifiers, for instance, SV X99911.3 (sequence version line type used for nucleotide sequence identifier) and protein_id = 'CAA45406.1' (protein sequence identifier for valid CDS features)
3. By directly searching for the required nucleotide or protein sequence.

Sequence annotation is an essential part of EMBL sequence records. EMBL records must have either Expressed Sequence Tag sites (ESTs) or Unfinished High Throughput Genome Sequences (HTGs), that are necessary for locating coding regions, to allow the inclusion of the corresponding translated protein sequence in the protein databases – TrEMBL and SWISS – PROT.

Angiotensinogen:
Angiotensinogen is a peptide prohormone that is an alpha – globulin precursor of angiotensin, a hormone involved in regulating the blood pressure and fluid balance of the body. It is primarily synthesized in the liver, kidney, adrenal glands, brain and other tissues. With a composition of 485 amino acids including a 33 – amino acid signal peptide, it is a member of the serpin family of proteins. Angiotensinogen is generally considered as a passive substrate of the renin – angiotensinogen system. It has a key physiological function as the carrier of the angiotensin peptides that control blood pressure.

METHODOLOGY:

1. Open the browser and search for EMBL – EBI database.
2. Enter the EMBL – EBI database and search for the query ‘angiotensinogen’.
3. Note down the total number of hits obtained for the basic search of the EMBL Portal.
4. Explore any 4 categories for limiting and obtaining information about the query “angiotensinogen”. Various categories explored were –
    a. Category 1: Genomes & Metagenomes
    b. Category 2: Protein Sequences
    c. Category 3: Macromolecular Structures
    d. Category 4: Gene Expressions
5. Note down the total number of hits obtained for each of the 4 categories.
6. After applying limits (for instance, Organism: *Homo sapiens*), note down the total number of hits obtained for each of the 4 categories.

OBSERVATIONS:

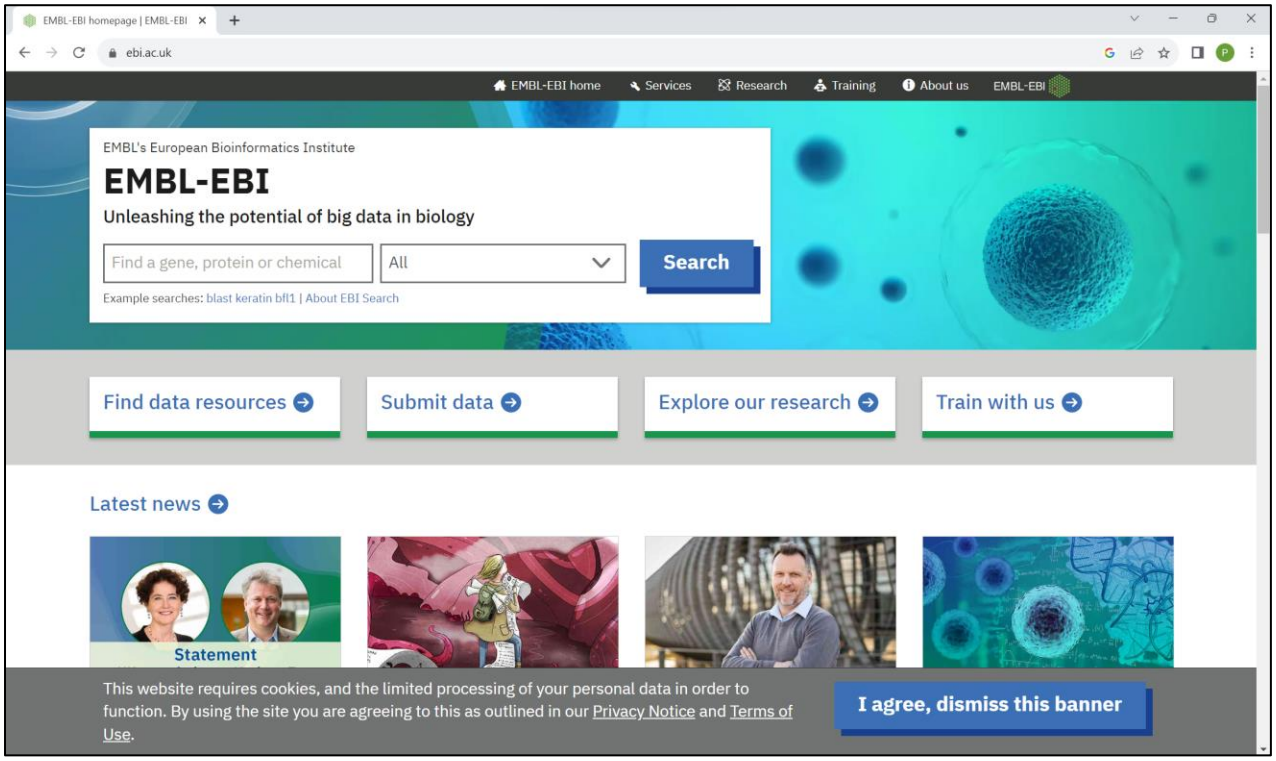


Figure 1: Homepage of EMBL – EBI Database

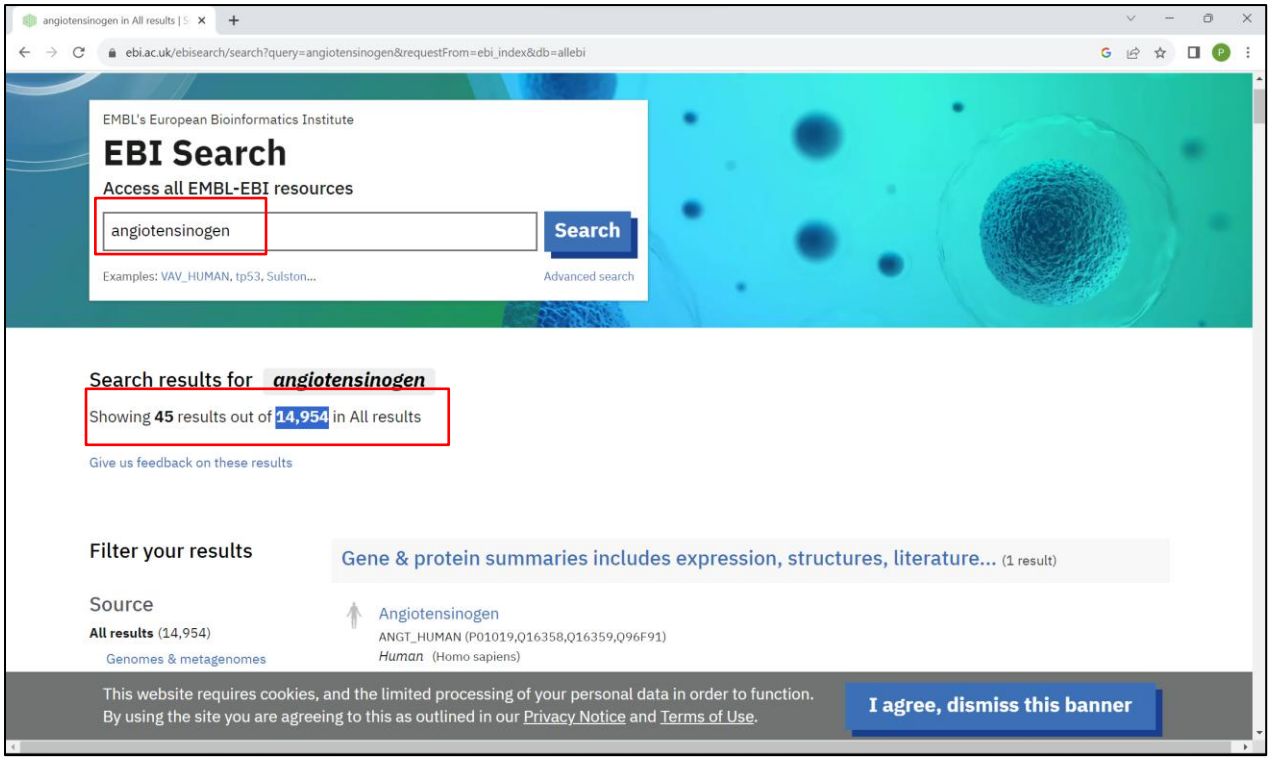


Figure 2: Number of hits obtained for Basic Search for the query Angiotensinogen (Accession ID: P01019)

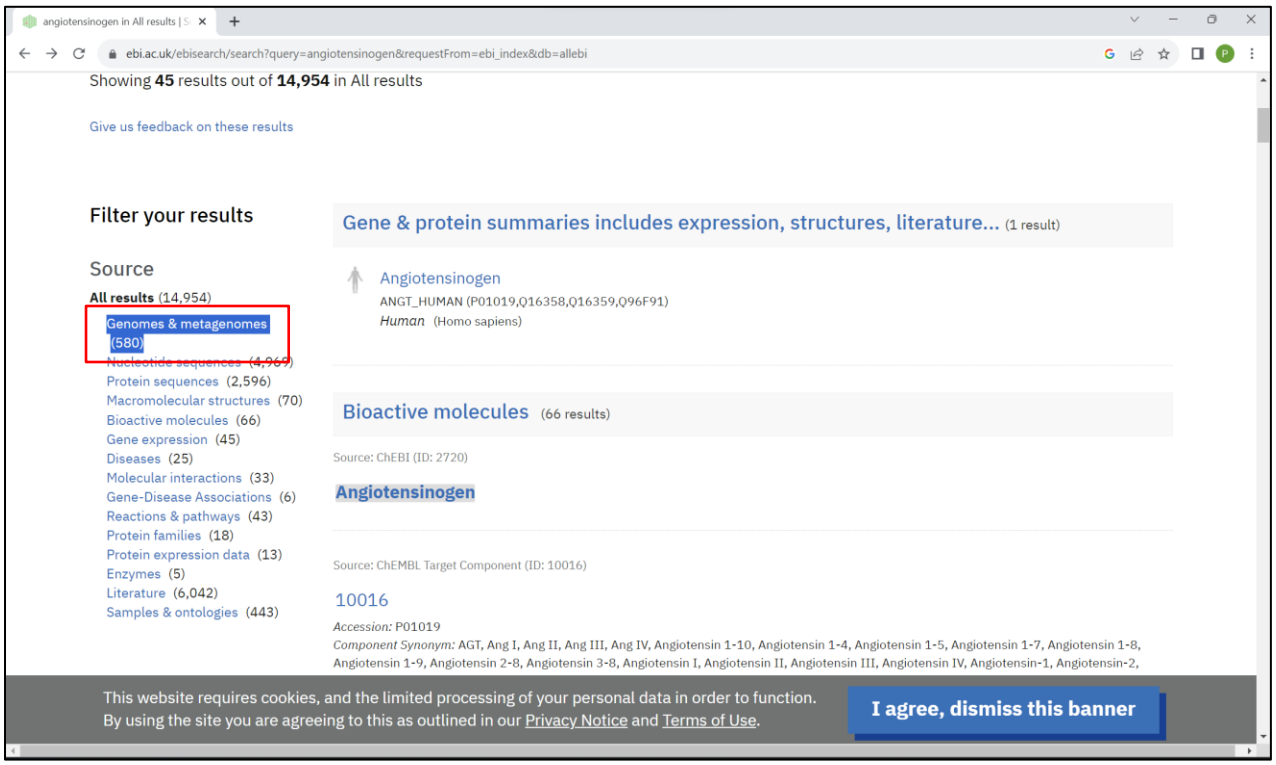


Figure 3: Total Number of Hits obtained for Category 1: Genomes & Metagenomes

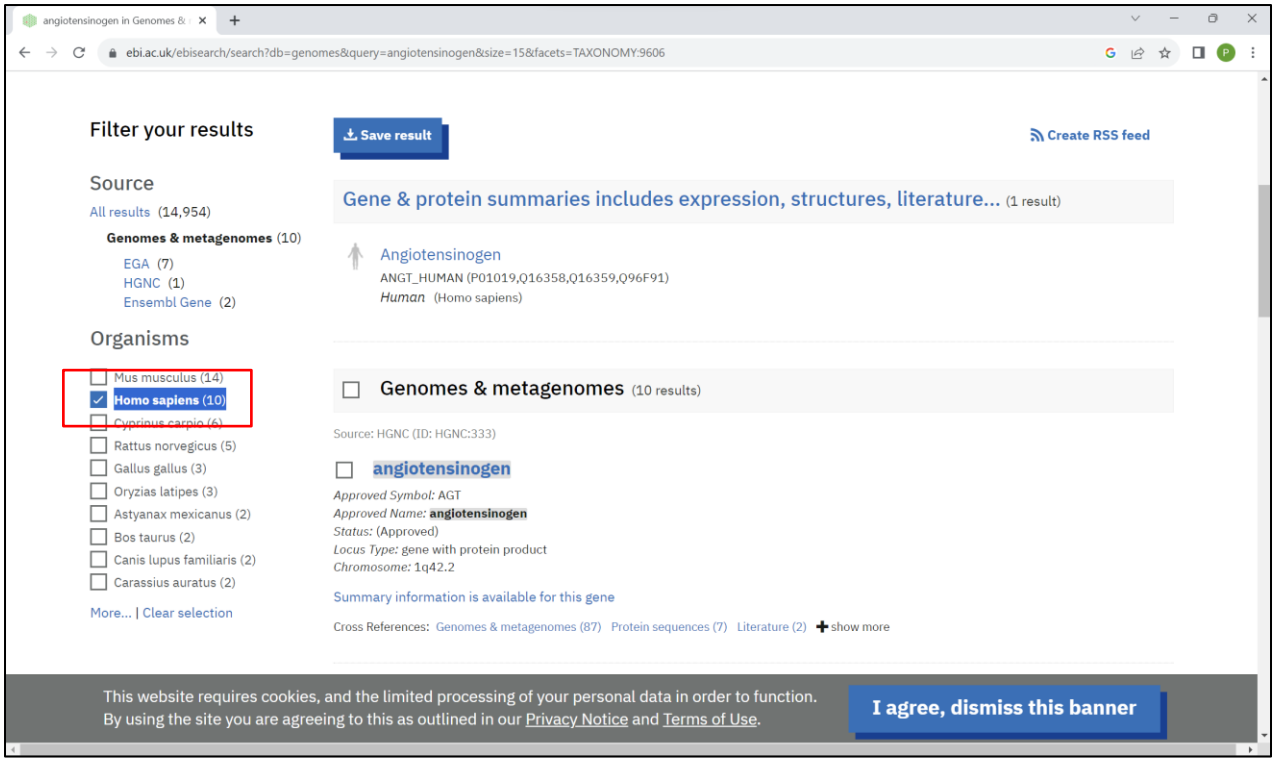


Figure 4: Total Number of Hits obtained for Category 1: Genomes & Metagenomes, Limits – Organism: *Homo sapiens*

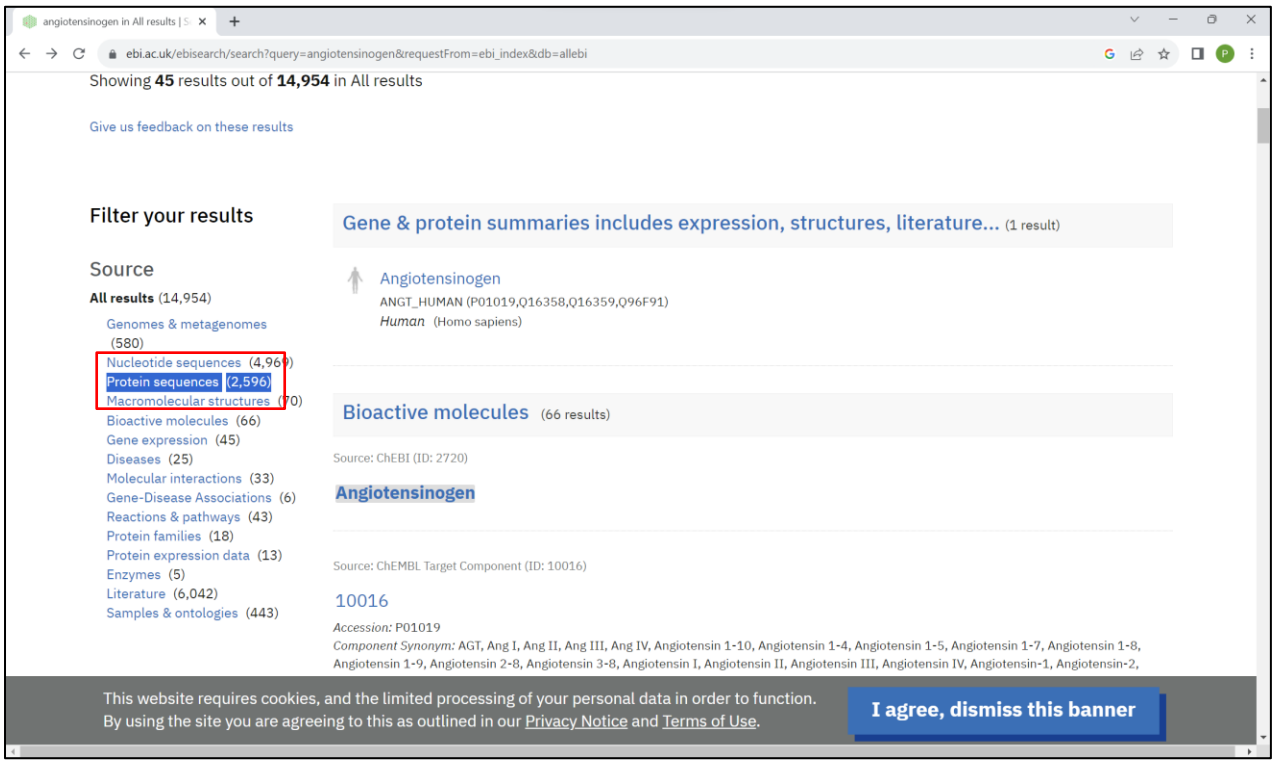


Figure 5: Total Number of Hits obtained for Category 2: Protein Sequences

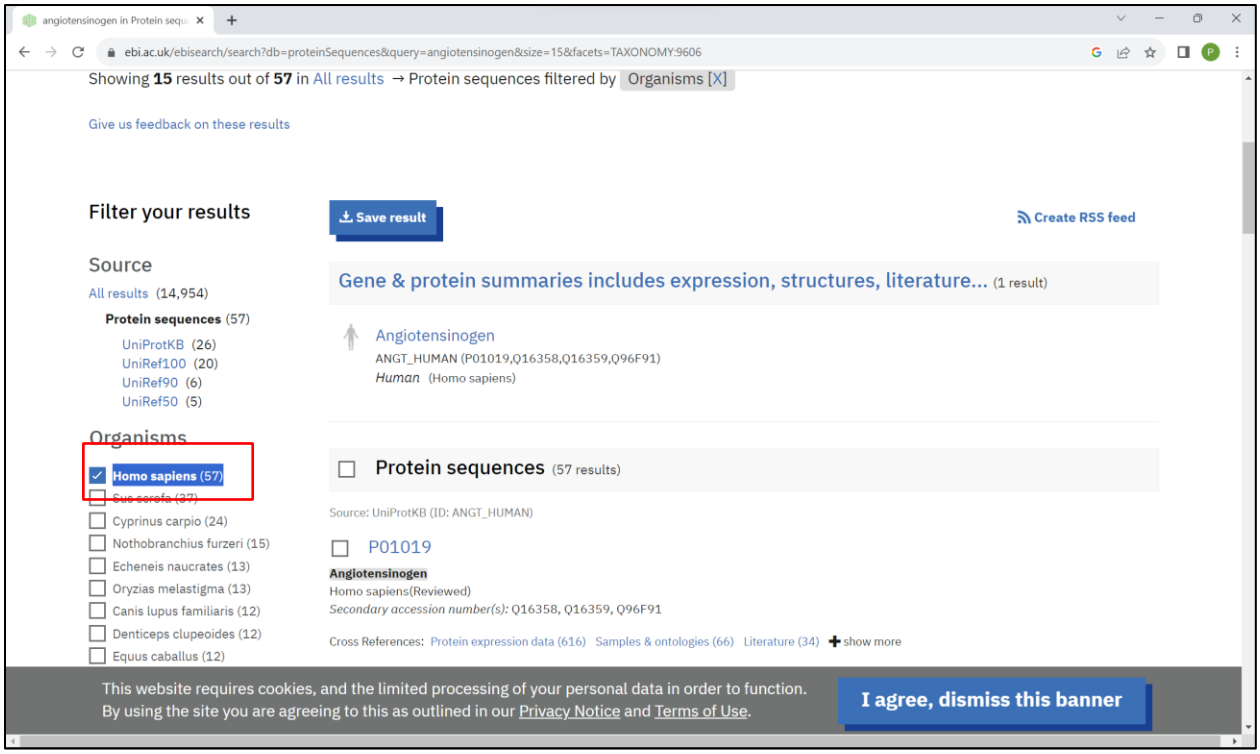


Figure 6: Total Number of Hits obtained for Category 2: Protein Sequences, Limits – Organism: *Homo sapiens*

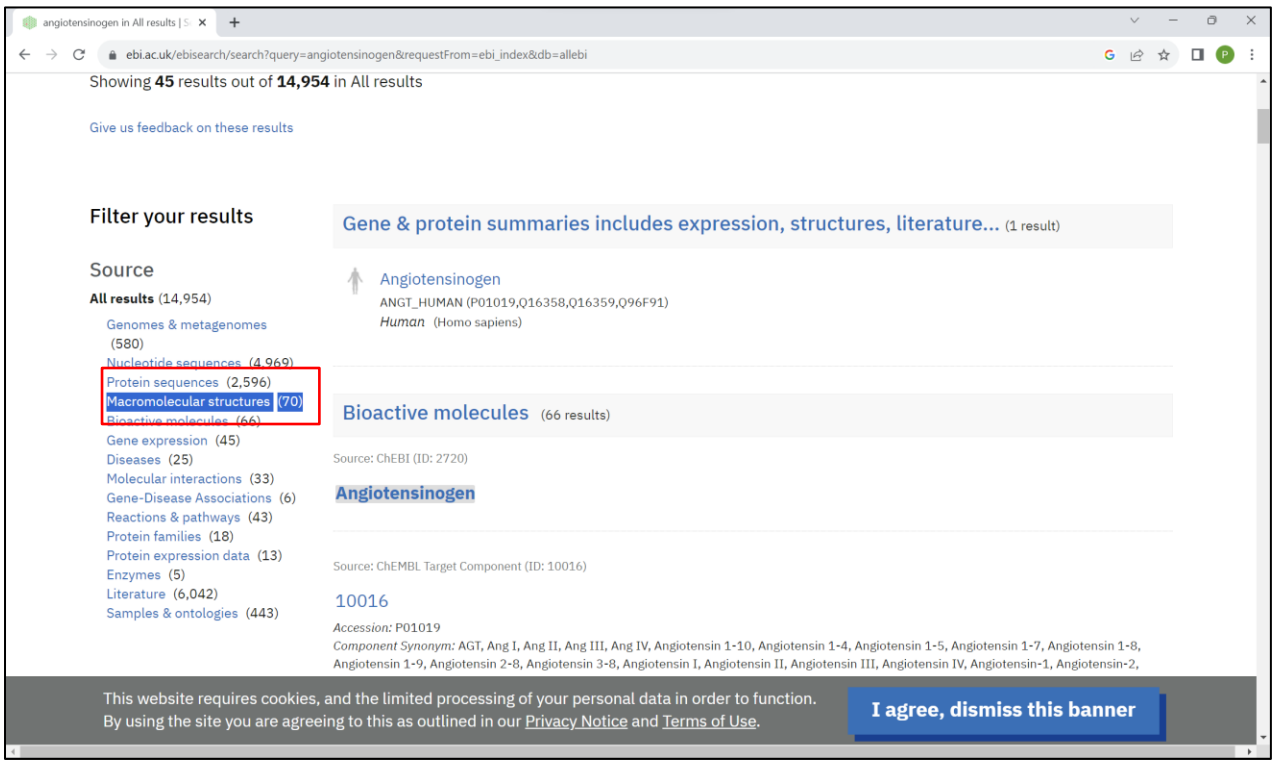


Figure 7: Total Number of Hits obtained for Category 3: Macromolecular Structures

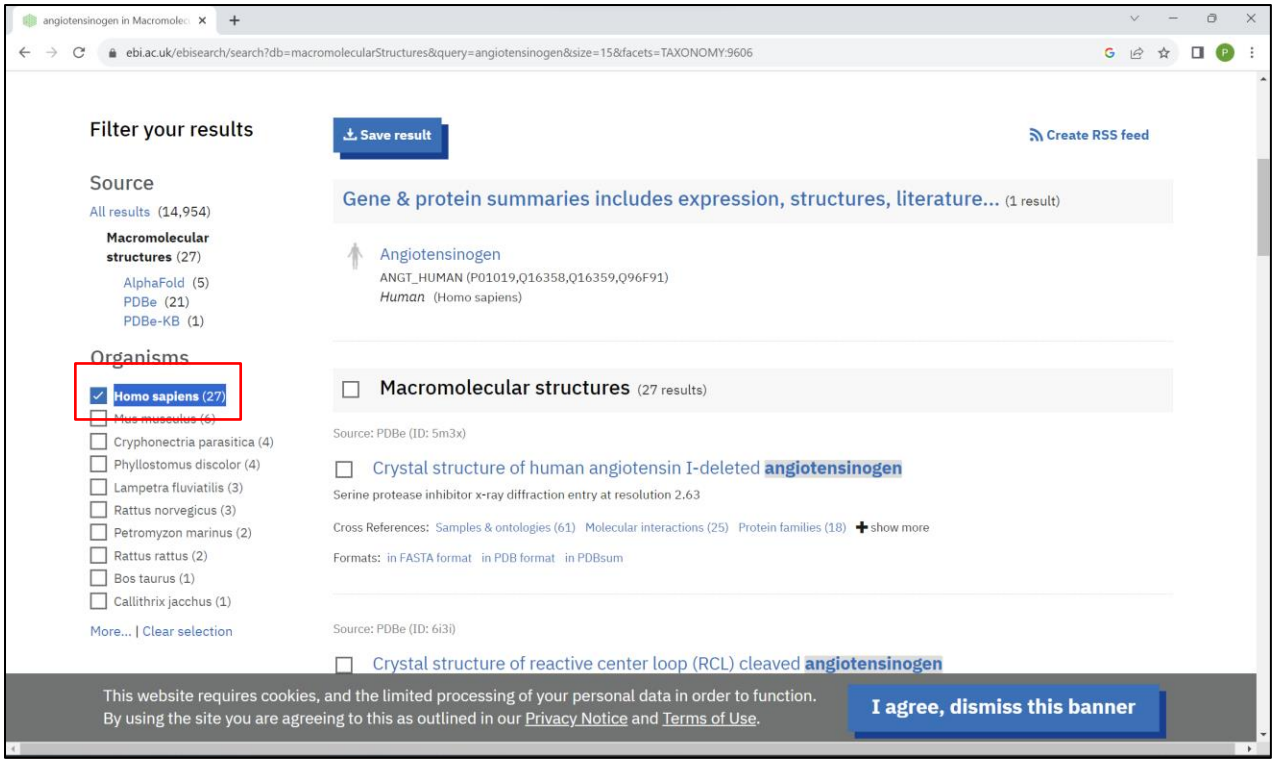


Figure 8: Total Number of Hits obtained for Category 3: Macromolecular Structures, Limits – Organism: *Homo sapiens*

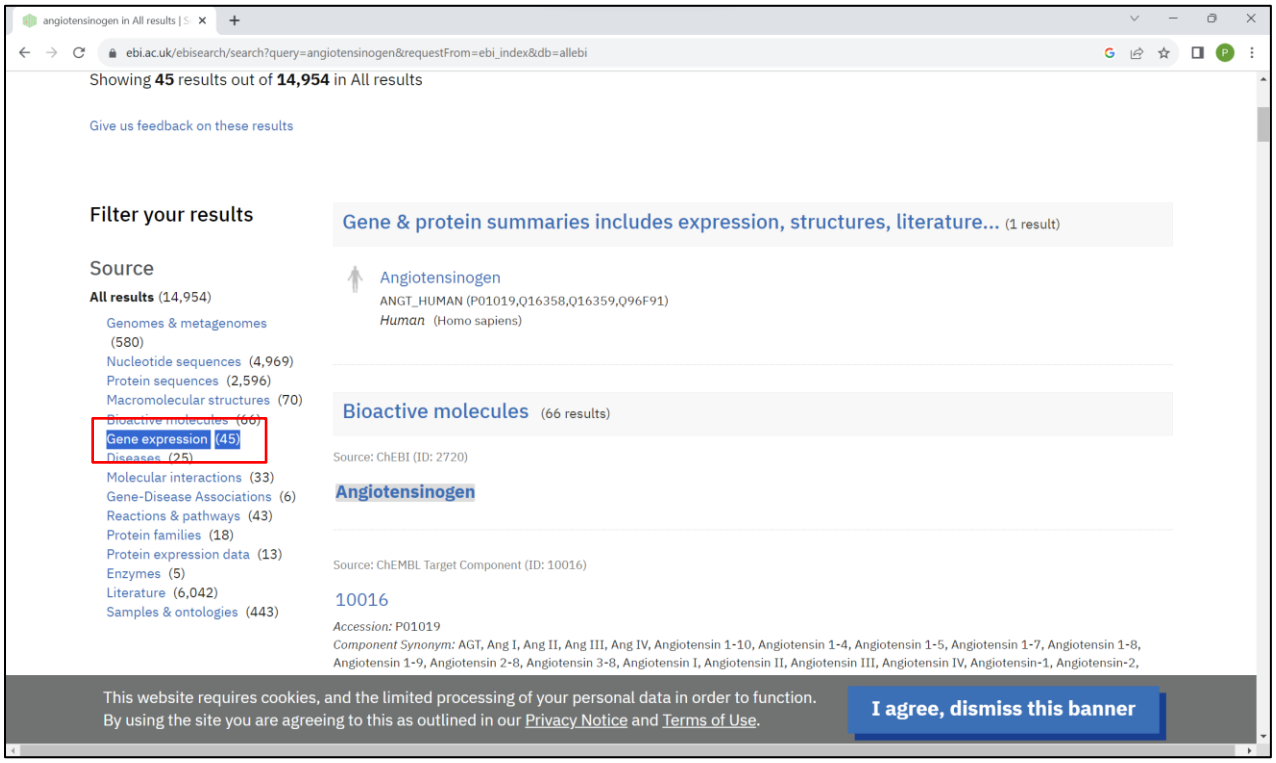


Figure 9: Total Number of Hits obtained for Category 4: Gene Expression

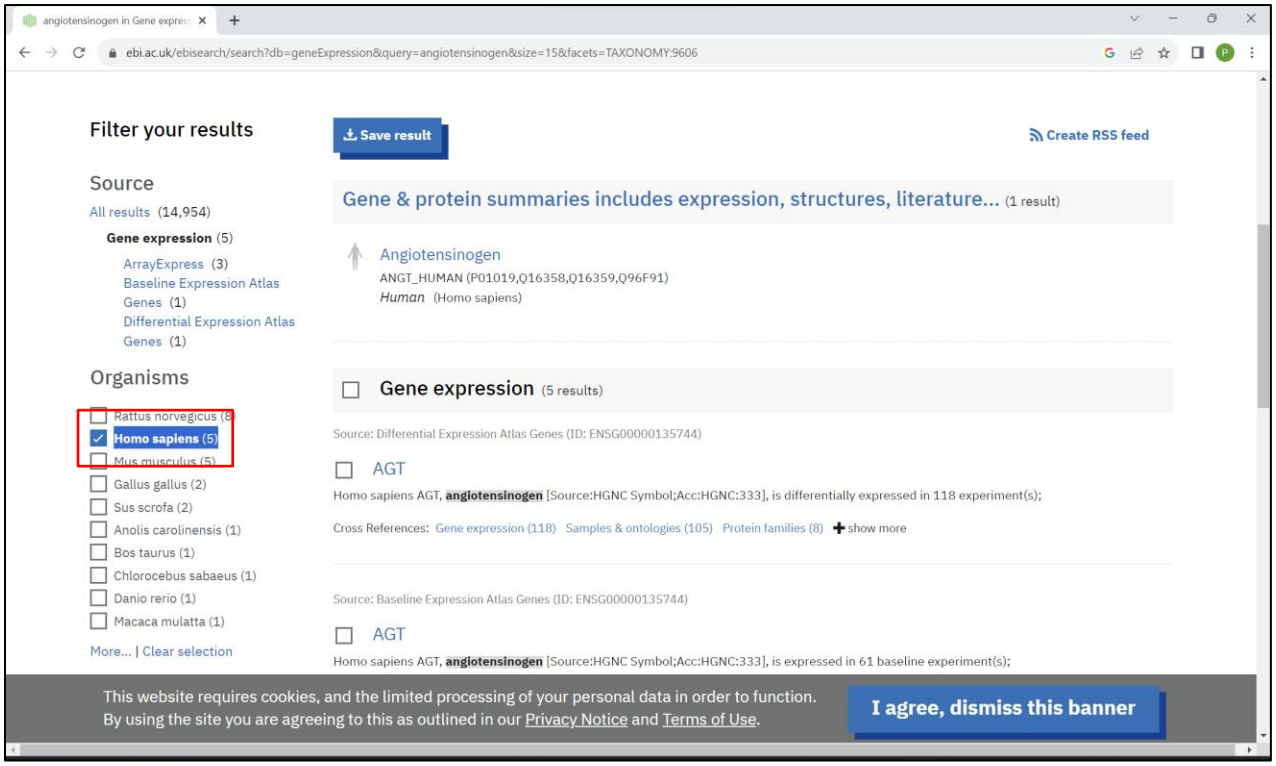


Figure 10: Total Number of Hits obtained for Category 4: Gene Expression, Limits – Organism: *Homo sapiens*

RESULTS:

The query 'angiotensinogen' (Accession ID: P01019) was searched and explored in the EMBL – EBI database. 14954 hits were obtained for basic search on the EMBL Portal. The following 4 categories were selected for further study – Genomes & Metagenomes, Protein Sequences, Macromolecular Structures and Gene Expression. Following hits were obtained for each of the mentioned categories –

Sr. No.	Category	No. of hits obtained	No. of hits obtained for Organism: *Homo sapiens*
1	Genomes & Metagenomes	580	10
2	Protein Sequences	2596	57
3	Macromolecular Structures	70	27
4	Gene Expression	45	5

CONCLUSION:

EMBL – EBI (European Molecular Biology Laboratory – European Bioinformatics Institute) Database was explored and the query 'angiotensinogen' was searched for and studied under the following 4 categories namely – Genomes & Metagenomes, Protein Sequences, Macromolecular Structures and Gene Expression.

REFERENCES:

1. Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., & Tuli, M. A. (2000). *The EMBL nucleotide sequence database. Nucleic acids research, 28(1), 19–23.* https://doi.org/10.1093/nar/28.1.19
2. Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F. G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Apweiler, R. (2005). *The EMBL Nucleotide Sequence Database. Nucleic acids research, 33(Database issue), D29–D33.* https://doi.org/10.1093/nar/gki098
3. Pevsner, J. (2009). Access to Sequence Data and Literature Information. Bioinformatics and Functional Genomics. pp. 12 – 45. https://doi.org/10.1002/9780470451496.ch2

DATE: 31/08/2023

WEBLEM 3(C)
DNA DATA BANK OF JAPAN (DDBJ) DATABASE
(URL - https://www.ddbj.nig.ac.jp)

AIM:

To explore the DDBJ (DNA Data Bank of Japan) Database with respect to ARSA search and further study of the query HBB Gene (ID: AY998983) in various file formats.

INTRODUCTION:

The DNA Data Bank of Japan (DDBJ) Database plays a crucial role in the field of life science research, functioning as a fundamental biological resource database for the scientific community all over the globe. DDBJ was founded in 1986 at the National Institute of Genetics in Shizuoka, Japan, and acts as an essential member in the International Nucleotide Sequence Database Collaboration (INSDC). The nucleotide sequence information in DDBJ Database is periodically synchronized with the European Molecular Biology Laboratory, GenBank, and other archives which facilitates regular data interchange and updating.

The DDBJ database is a comprehensive repository that predominantly collects DNA sequences generated by Japanese researchers, while also encouraging scientists from throughout the globe to make valuable additions. The DDBJ Center is responsible for the management of archival databases that save nucleotide sequences, study particulars, and sample data. Its primary objective is to facilitate unrestricted access to these resources, hence facilitating progress in the field of life science research. DDBJ adheres to stringent data management and sharing protocols, overseen by its advisory groups such as the DNA Database Advisory Committee and the International Advisory Committee to INSDC.

In the year 2020, the DNA Data Bank of Japan (DDBJ) Database experienced a notable influx of nucleotide sequence submissions, with a majority of these submissions being attributed to research groups based in Japan. The database also provides training courses in the field of bioinformatics via various programs, such as DDBJing, which aids in the process of data submission and analysis. DDBJ plays a crucial role as an invaluable resource, facilitating scientific inquiry and advancing life science research on a worldwide level.

Sickle Cell Anemia:

Sickle cell anemia is a hereditary blood disorder characterized by abnormal hemoglobin molecules, specifically hemoglobin S (HbS), which can deform red blood cells into a crescent shape. The manifestation of this particular medical issue arises as a consequence of genetic mutations occurring in the HBB gene, which is situated on chromosome 11p15.5. The HBB gene is primarily responsible for encoding beta-globin, a vital constituent of hemoglobin. The HBB gene has significant variability, encompassing a wide range of variants that have the potential to result in diverse manifestations of sickle cell disease. Within these multiple variations, a single beta-globin subunit is changed with HbS, while the other is replaced with several defective hemoglobin variations such as hemoglobin C or hemoglobin E.

Sickle cell anemia is characterized by an autosomal recessive mode of inheritance and predominantly impacts populations residing in areas with a notable historical incidence of malaria. This is due to the fact that carriers of HBB mutations experience some kind of immunity against malaria. The average frequency of these mutations in the African American population is estimated to be around 8%. When individuals who have a homozygous genotype for HbS, they are exposed to circumstances such as reduced oxygen levels or increased hemoglobin concentrations, the HbS molecules have the ability to undergo polymerization, resulting in the distinctive creation of sickle-shaped red blood cells. The comprehensive knowledge of the genetic causes of this condition is essential for the advancement of the study and development of remedies. DNA Databases, like DDBJ Database, helps providing the comprehensive genomic data to the researcher to study on Sickle Cell Anemia and HBB Gene and find the appropriate treatment.

METHODOLOGY:

1. Visit the DNA Data Bank of Japan's (DDBJ) Database homepage.
2. Use the 'ARSA' search option to retrieve annotated/assembled data from the DDBJ Database using accession numbers and/or keywords.
   i. To retrieve data about BioProject/BioSample/SRA & JGA Data, go to 'DDBJ Search'.
   ii. To retrieve data about Taxonomy, do to 'TXSearch'.
   iii. To retrieve data about DDBJ annotated/assembled data using only accession numbers, go to 'getentry'.
   iv. To use web API for searching DDBJ data without navigating to the web front-end, go to 'WABI'.
3. Search for the query "HBB" in the Quick Search section.
4. After retrieving a list of relevant entries, select the desired entries and click the 'View Selected' option to view further details regarding the selected entries. A compressed version of the selected entries may also be downloaded in a particular file format such as FlatFile, XML or FASTA file format for further study.
5. The data can then be imported into the desired software after decompressing and extracting the downloaded data and further analyzed.

OBSERVATIONS:

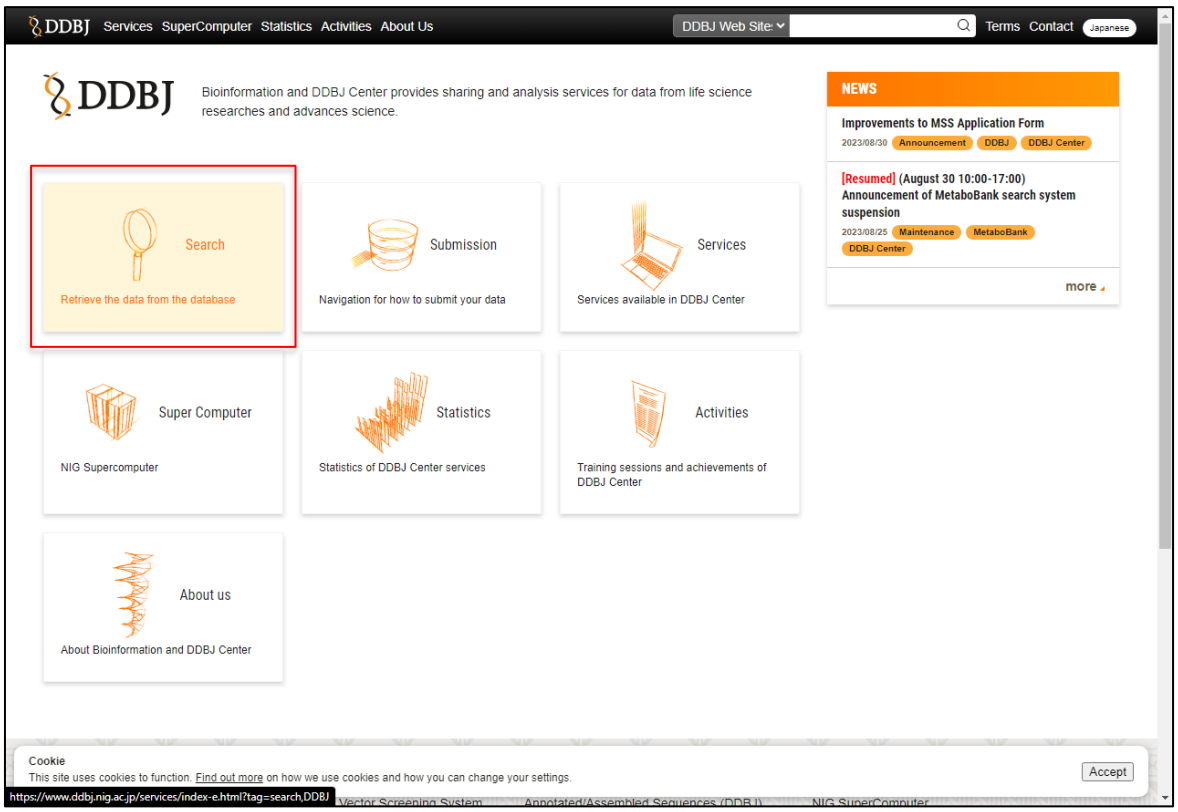


Figure 1: Homepage of DDBJ Database

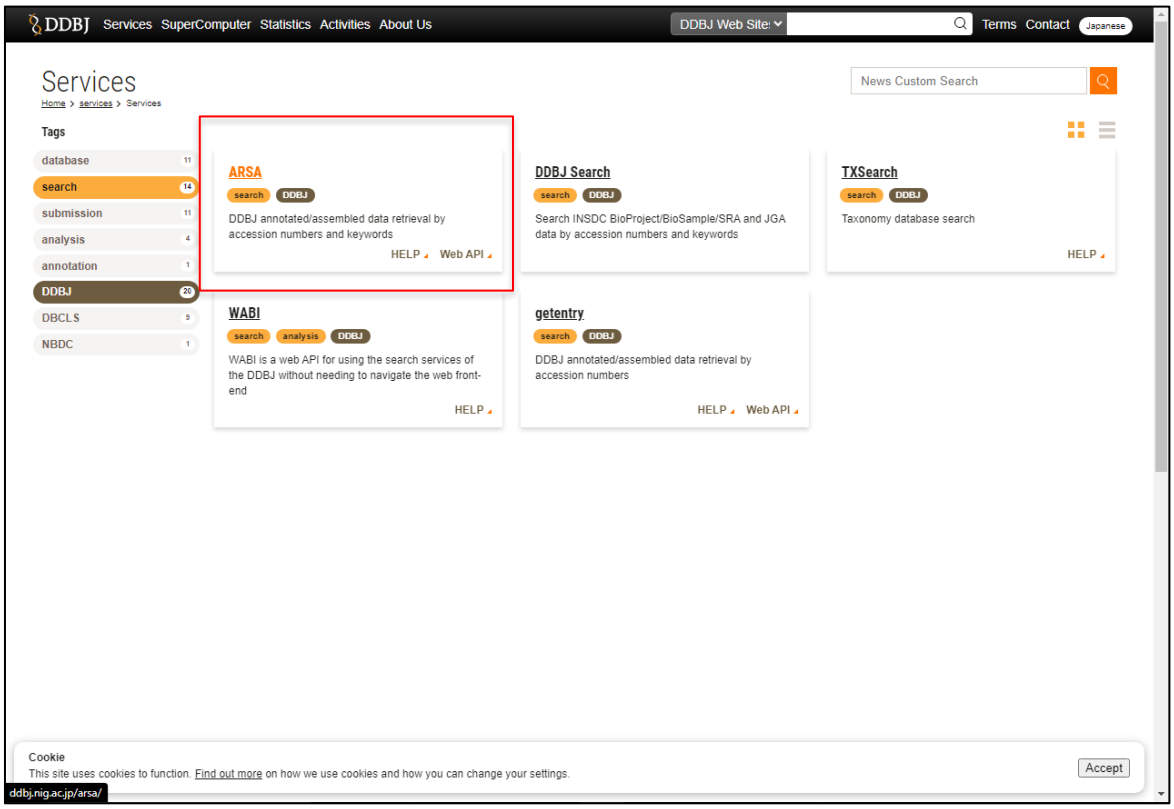


Figure 2: Service page of DDBJ Database with ARSA search option

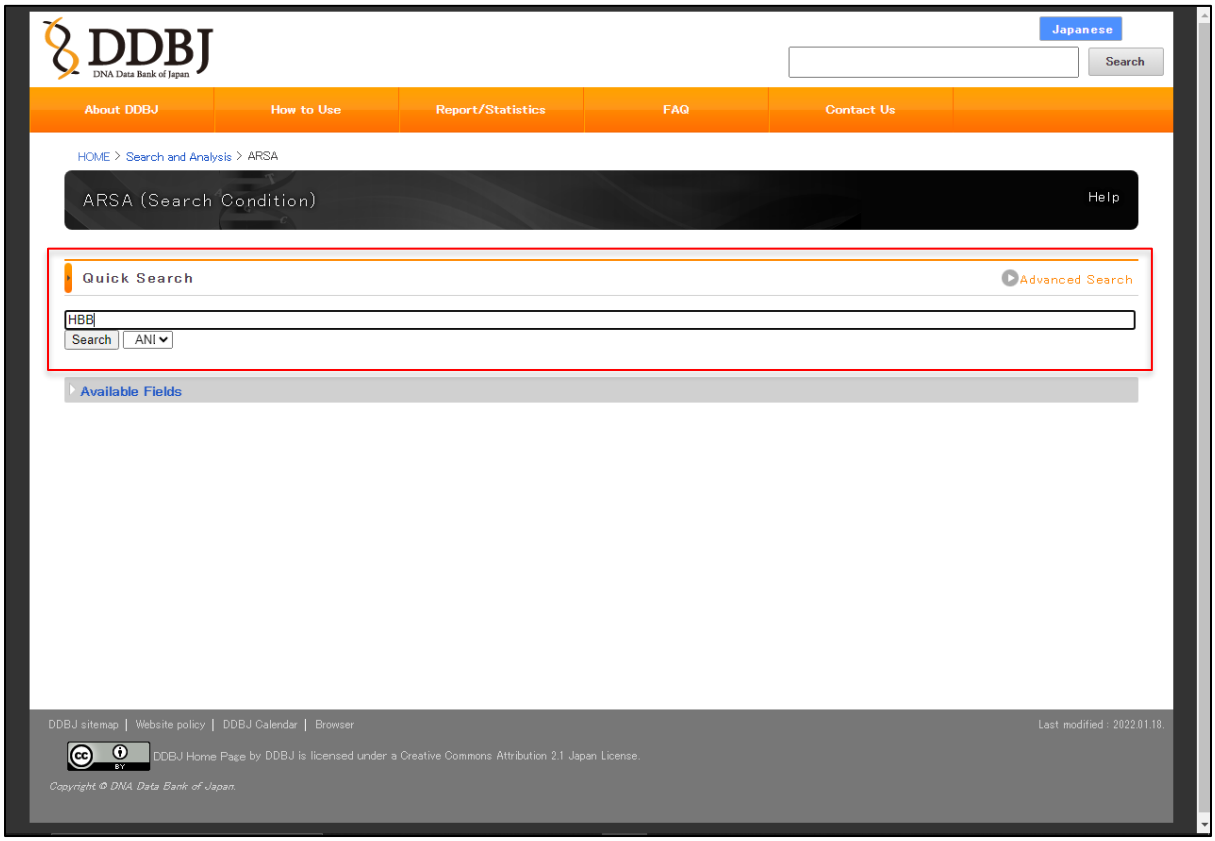


Figure 3: ARSA Search Page with the query - HBB

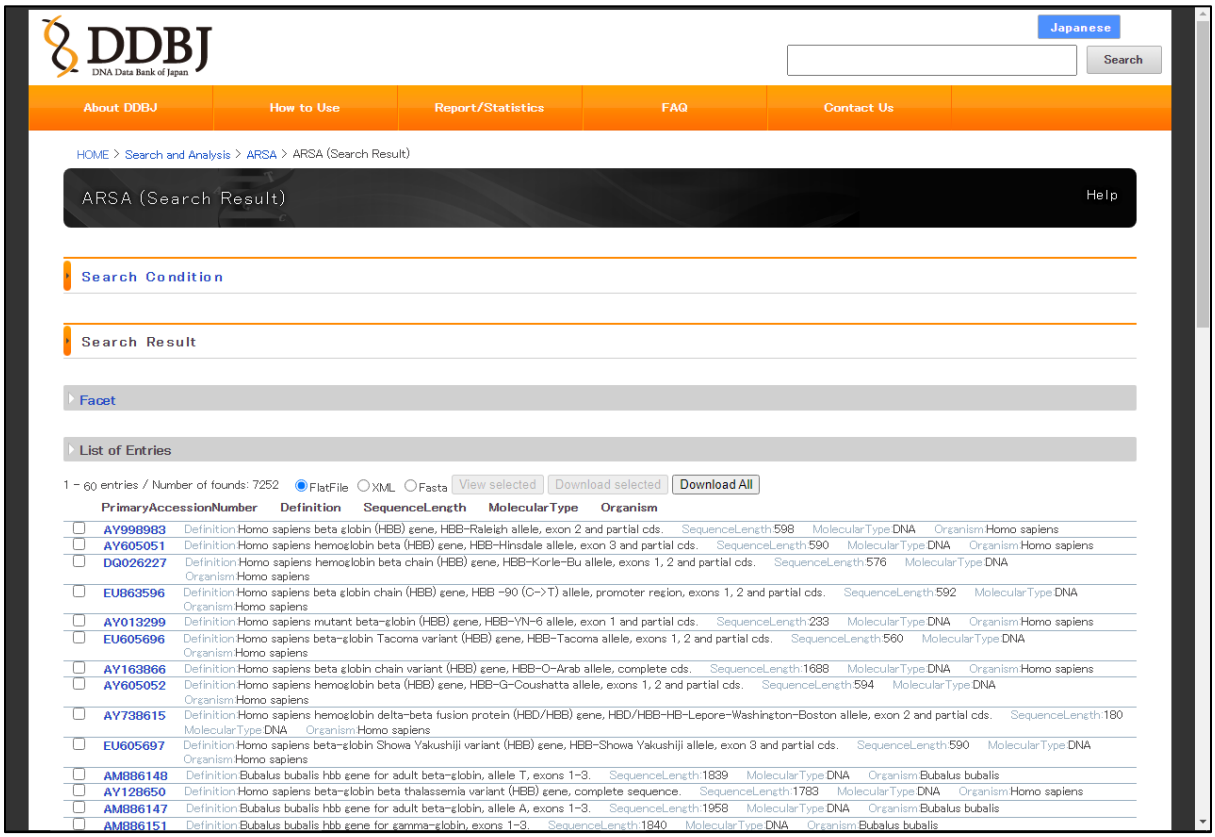


Figure 4: ARSA Results Page for the query

```
LOCUS       AY998983                 598 bp    DNA     linear   HUM 26-JUL-2016
DEFINITION  Homo sapiens beta globin (HBB) gene, HBB-Raleigh allele, exon 2 and
            partial cds.
ACCESSION   AY998983
VERSION     AY998983.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 598)
  AUTHORS   Knovich,M., Davis,D.H., Nechtman,J., Elam,D., Kutlar,A. and
            Kutlar,F.
  TITLE     Hemoglobin A-Raleigh: a low oxygen affinity beta chain variant
            (GTG->GCG/Val-1-Ala) has been detected on a 'Cambodian' individual
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 598)
  AUTHORS   Knovich,M., Davis,D.H., Nechtman,J., Elam,D., Kutlar,A. and
            Kutlar,F.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-APR-2005) Medicine, Medical College of Georgia, Laney
            Walker Bulv. AC-1000, Augusta, GA 30912, USA
FEATURES             Location/Qualifiers
     source          1..598
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="11"
                     /map="11p15.5"
                     /sex="female"
                     /cell_type="whole blood"
                     /country="Cambodia"
     gene            <1..>598
                     /gene="HBB"
                     /allele="HBB-Raleigh"
     mRNA            join(<1..102,233..>455)
                     /gene="HBB"
                     /allele="HBB-Raleigh"
     exon            <1..102
                     /gene="HBB"
                     /allele="HBB-Raleigh"
                     /number=1
     CDS             join(11..102,233..>455)
                     /gene="HBB"
                     /allele="HBB-Raleigh"
                     /note="low oxygen affinity hemoglobin variant"
                     /codon_start=1
                     /product="beta globin"
                     /protein_id="AAY15222.1"
                     /translation="MAHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFE
                     SFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPE
                     NFR"
     variation       15
                     /gene="HBB"
                     /note="heterozygous; results in Val to Ala; rare mutation
                     occurring in Caucasian and Swedish families; Hb. Raleigh"
                     /replace="t"
```

General information section

Detailed features section containing gene segment descriptions

Figure 5.a.1: Data visualization using 'View Selected' option, in FlatFile format with General information & Detailed features section

```
                     /replace="c"
     variation       529
                     /gene="HBB"
                     /note="heterozygous"
                     /replace="g"
BASE COUNT          137 a          127 c          175 g          159 t
ORIGIN
        1 aacagacacc atggcgcatc tgactcctga ggagaagtct gccgttactg ccctgtgggg
       61 caaggtgaac gtggatgaag ttggtggtga ggccctgggc aggttggtat caaggttaca
      121 agacaggttt aaggagacca atagaaactg ggcatgtgga gacagagaag actcttgggt
      181 ttctgatagg cactgactct ctctgcctat tggtctattt tcccaccctt aggctgctgg
      241 tggtctaccc ttggacccag aggttctttg agtcctttgg ggatctgtcc actcctgatg
      301 ctgttatggg caaccctaag gtgaaggctc atggcaagaa agtgctcggt gcctttagtg
      361 atggcctggc tcacctggac aacctcaagg gcacctttgc cacactgagt gagctgcact
      421 gtgacaagct gcatgtggat cctgagaact tcagggtgag tctatgggac gcttgatgtt
      481 ttctttcccc ttcttttcta tggttaagtt catgtcatag gaaggggata agtaacaggg
      541 tacagtttag aatgggaaac agacgaatga ttgcatcagt gtggaagtct caggatcg
//
```

Origin section containing gene sequence

Figure 5.a.2: Data visualization using 'View Selected' option, in FlatFile format with Origin Section

```
>AY998983|AY998983.1 Homo sapiens beta globin (HBB) gene, HBB-Raleigh allele, exon 2 and partial cds.
aacagacaccatggcgcatctgactcctgaggagaagtctgccgttactgccctgtgggg
caaggtgaacgtggatgaagttggtggtgaggccctgggcaggttggtatcaaggttaca
agacaggtttaaggagaccaatagaaactgggcatgtggagacagagaagactcttgggt
ttctgataggcactgactctctctgcctattggtctattttcccacccttaggctgctgg
tggtctacccttggacccagaggttctttgagtcctttggggatctgtccactcctgatg
ctgttatgggcaaccctaaggtgaaggctcatggcaagaaagtgctcggtgcctttagtg
atggcctggctcacctggacaacctcaagggcacctttgccacactgagtgagctgcact
gtgacaagctgcatgtggatcctgagaacttcagggtgagtctatgggacgcttgatgtt
ttctttccccttcttttctatggttaagttcatgtcataggaaggggataagtaacaggg
tacagtttagaatgggaaacagacgaatgattgcatcagtgtggaagtctcaggatcg
>AY605051|AY605051.1 Homo sapiens hemoglobin beta (HBB) gene, HBB-Hinsdale allele, exon 3 and partial cds.
ctaaagaataacagtgataatttctgggttaaggcaatagcaatatctctgcatataaat
atttctgcatataaattgtaactgatgtaagaggtttcatattgctaatagcagctacaa
tccagctaccattctgcttttattttatggttgggataaggctggattattctgagtcca
agctaggccctttttgctaatcatgttcatacctcttatcttcctcccacagctcctgggc
aacgtgctggtctgtgtgctggcccatcactttggcaaagaattcaccccaccagtgcag
gctgcctatcagaaagtggtggctggtgtggctaaagccctggcccacaagtatcactaa
gctcgctttcttgctgtccaatttctattaaaggttcctttgttccctaagtccaactac
taaactgggggatattatgaagggccttgagcatctggattctgcctaataaaaaacatt
tattttcattgcaatgatgtatttaaattatttctgaatattttactaaaaagggaatgt
gggaggtcagtgcatttaaaacataaagaaatgaagagctagttcaaacc
```

Figure 5.b: Data visualization using 'View Selected' option, in Fasta file format


```
This XML file does not appear to have any style information associated with it. The document tree is shown below.

▼<INSDSet>
  ▼<INSDSeq>
     <INSDSeq_locus>AY998983</INSDSeq_locus>
     <INSDSeq_length>598</INSDSeq_length>
     <INSDSeq_moltype>DNA</INSDSeq_moltype>
     <INSDSeq_topology>linear</INSDSeq_topology>
     <INSDSeq_division>HUM</INSDSeq_division>
     <INSDSeq_update-date>26-JUL-2016</INSDSeq_update-date>
     <INSDSeq_definition>Homo sapiens beta globin (HBB) gene, HBB-Raleigh allele, exon 2 and partial cds</INSDSeq_definition>
     <INSDSeq_primary-accession>AY998983</INSDSeq_primary-accession>
     <INSDSeq_accession-version>AY998983.1</INSDSeq_accession-version>
     <INSDSeq_source>Homo sapiens (human)</INSDSeq_source>
     <INSDSeq_organism>Homo sapiens</INSDSeq_organism>
     <INSDSeq_taxonomy>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
     Homo.</INSDSeq_taxonomy>
    ▼<INSDSeq_references>
     ▼<INSDReference>
        <INSDReference_reference>1</INSDReference_reference>
        <INSDReference_position>1..598</INSDReference_position>
       ▼<INSDReference_authors>
          <INSDAuthor>Knovich,M.</INSDAuthor>
          <INSDAuthor>Davis,D.H.</INSDAuthor>
          <INSDAuthor>Nechtman,J.</INSDAuthor>
          <INSDAuthor>Elam,D.</INSDAuthor>
          <INSDAuthor>Kutlar,A.</INSDAuthor>
          <INSDAuthor>Kutlar,F.</INSDAuthor>
        </INSDReference_authors>
        <INSDReference_title>Hemoglobin A-Raleigh: a low oxygen affinity beta chain variant (GTG->GCG/Val-1-Ala) has been detected on a 'Cambodian' individual
        </INSDReference_title>
        <INSDReference_journal>Unpublished </INSDReference_journal>
      </INSDReference>
     ▼<INSDReference>
        <INSDReference_reference>2</INSDReference_reference>
        <INSDReference_position>1..598</INSDReference_position>
       ▼<INSDReference_authors>
          <INSDAuthor>Knovich,M.</INSDAuthor>
          <INSDAuthor>Davis,D.H.</INSDAuthor>
          <INSDAuthor>Nechtman,J.</INSDAuthor>
          <INSDAuthor>Elam,D.</INSDAuthor>
          <INSDAuthor>Kutlar,A.</INSDAuthor>
          <INSDAuthor>Kutlar,F.</INSDAuthor>
        </INSDReference_authors>
        <INSDReference_title>Direct Submission </INSDReference_title>
        <INSDReference_journal>Submitted (04-APR-2005) Medicine, Medical College of Georgia, Laney Walker Bulv. AC-1000, Augusta, GA 30912, USA </INSDReference_journal>
      </INSDReference>
    </INSDSeq_references>
    ▼<INSDSeq_feature-table>
     ▼<INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..598</INSDFeature_location>
       ▼<INSDFeature_intervals>
        ▼<INSDInterval>
           <INSDInterval_from>1</INSDInterval_from>
           <INSDInterval_to>598</INSDInterval_to>
```


Figure 5.c: Data visualization using 'View Selected' option, in XML format.

RESULTS:

Through the DNA Databank of Japan Database –

1. The entry with Accession ID – AY998983 was studied and analyzed.
2. The total of 7275 entries were found for the query of HBB. However, only 180 entries loaded initially and later as scrolled down subsequent entries loaded.

CONCLUSION:

The DDBJ Database was explored using the ARSA search to retrieve data for the query HBB (AY998983). The query was further studied and analyzed using both the 'View Selected' option on the web browser itself and downloading the sequence in the FASTA file format.

REFERENCES:

1. Inusa, B., Hsu, L., Kohli, N., Patel, A., Ominu-Evbota, K., Anie, K., & Atoyebi, W. (2019, May 7). Sickle Cell Disease—Genetics, Pathophysiology, Clinical Presentation and Treatment. *International Journal of Neonatal Screening*, *5*(2), 20. https://doi.org/10.3390/ijns5020020
2. Bethesda & National Center for Biotechnology Information (US). (1998). *Genes and Disease* [Internet]. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK22238/
3. Thomson, A. M., McHugh, T. A., Oron, A. P., Teply, C., Lonberg, N., Vilchis Tella, V., Wilner, L. B., Fuller, K., Hagins, H., Aboagye, R. G., Aboye, M. B., Abu-Gharbieh, E., Abu-Zaid, A., Addo, I. Y., Ahinkorah, B. O., Ahmad, A., AlRyalat, S. A. S., Amu, H., Aravkin, A. Y., . . . Kassebaum, N. J. (2023, August). Global, regional, and national prevalence and mortality burden of sickle cell disease, 2000–2021: a systematic analysis from the Global Burden of Disease Study 2021. *The Lancet Haematology*, *10*(8), e585–e599. https://doi.org/10.1016/s2352-3026(23)00118-7
4. Mashima, J., Kodama, Y., Fujisawa, T., Katayama, T., Okuda, Y., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., & Takagi, T. (2016, October 24). DNA Data Bank of Japan. *Nucleic Acids Research*, *45*(D1), D25–D31. https://doi.org/10.1093/nar/gkw1001

DATE: 08/09/2023

WEBLEM 3(D)
UNIPROT DATABASE
(URL: https://www.uniprot.org/)

AIM:

To explore the UniProt Database for further study of the query – thrombin protein (Accession ID – P25116.).

INTRODUCTION:

The UniProt database is a free resource for protein sequence and functional information. It contains over 60 million sequences, including over half a million that have been curated by experts. The database was originally created as a primary database for protein sequences and functional annotation based on experimental evidence. It now combines a network of sister databases that centralize all levels of annotation for protein sequences.

The UniProt databases are:

1. UniProt Knowledgebase (UniProtKB)
2. UniProt Reference Clusters (UniRef)
3. UniProt Archive (UniParc)

UniProt Database was created by combining Swiss-Prot, TrEMBL, and PIR. Many entries in the database are derived from genome sequencing projects.

The Protein Data Bank (PDB) is the central archive of all experimentally determined protein structure data. The PDB was established in 1971 and is maintained by an international consortium known as the Worldwide Protein Data Bank (wwPDB).

Thrombin:

Thrombin is a protein in the bloodstream that helps blood clot. It is the final enzyme in the blood coagulation cascade and is a member of the trypsin family of serine proteases.

Thrombin's two main actions are:

1. Cleaving fibrinogen to release fibrin
2. Activating platelets through a specific receptor
3. Thrombin also catalyzes other coagulation-related reactions.

Thrombin is produced when prothrombin is activated by tissue thromboplastin in the presence of calcium chloride. The first step of the cleavage is at residue R320 and produces meizothrombin.

Thrombin is a multifunctional enzyme that has been implicated in brain development. It also has a mitogenic effect, which stimulates the growth of mammalian cells, fibroblasts, and macrophage-like tumor cell lines.

Protease-activated receptor-1 (PAR-1) is a G protein-coupled receptor that regulates the endothelium. It blocks cytokine signaling, adhesion molecule expression, vascular permeability, apoptosis, and leukocyte migration and adhesion.

PAR-1 was the first member of the PARs (protease-activated receptors) family. The other members of the family are PAR2, PAR3, and PAR4.
PAR-1 inhibitors are a new class of antiplatelet agents. They are used to reduce the risk of a heart attack in people with coronary artery disease. They work by inhibiting thrombin-related platelet aggregation.

METHODOLOGY:

1. Go to the UniProt database homepage and type "thrombin protein" into the search box.
2. Decide whether you choose to view your results as a table or cards.
3. Use several filters to look for thrombin, such as organism popularity, taxonomy, proteins having 3D structures, sequence length, etc.
4. Save data in the FASTA format.
5. Results can be sorted by functions, name, taxonomy, subcellular location, disease and variations, structure, family & domains, sequence, and related proteins when you click on a result.

OBSERVATIONS:

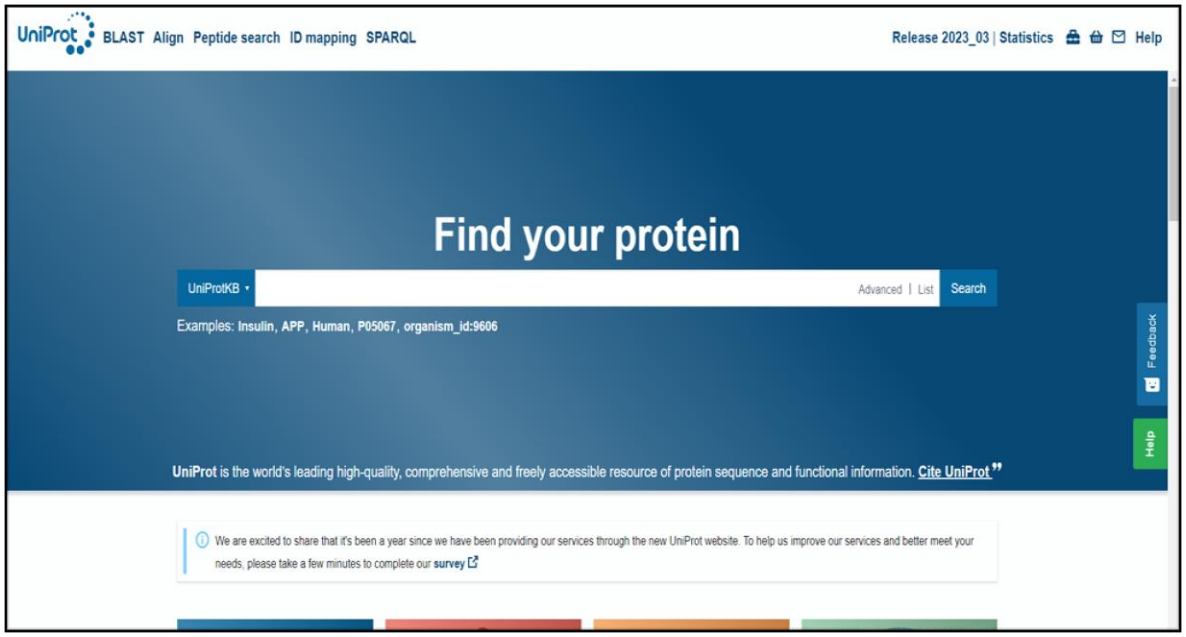


Figure 1: Homepage of UniProt Database

(A drop-down list next to the search box allows you to specify the protein you want to look up, and the search box itself can be used to look up many proteins.)

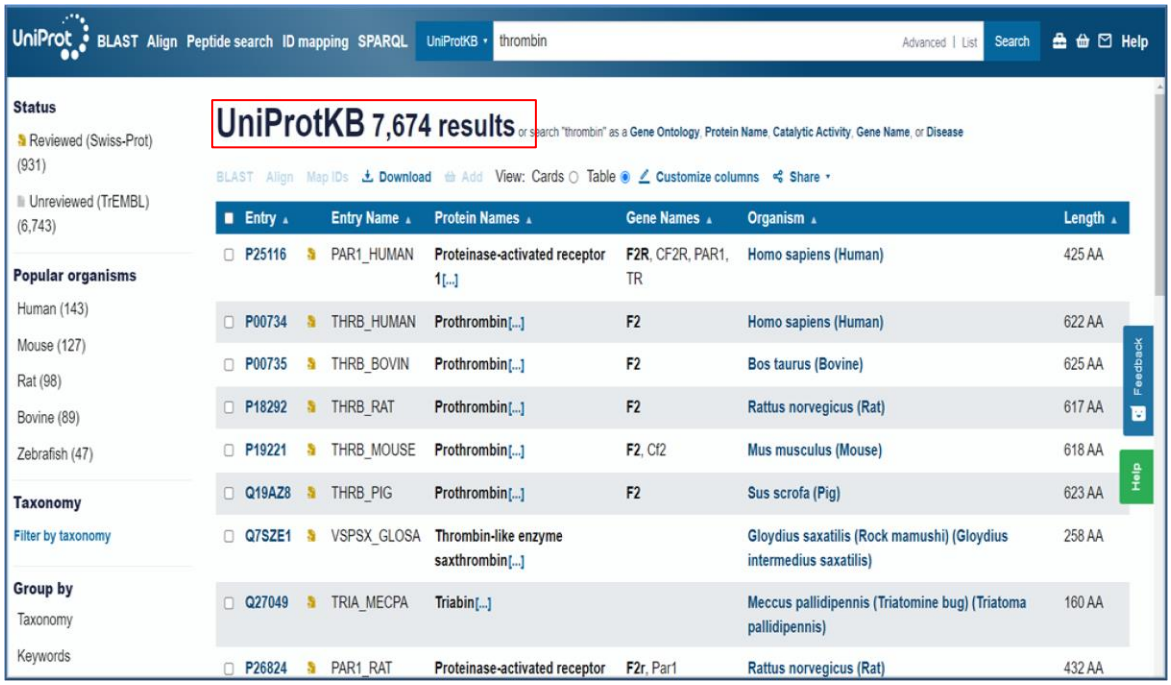


Figure 2: Thrombin protein reviewed (SwissProt) search (931 results) 7,674 hits are displayed in the search results.

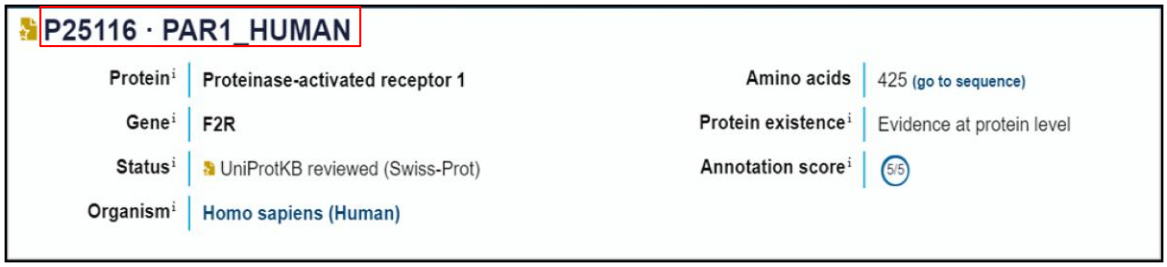


Figure 3: The first result on a search for "thrombin protein" is protein activated receptor 1 with 425 amino acids.

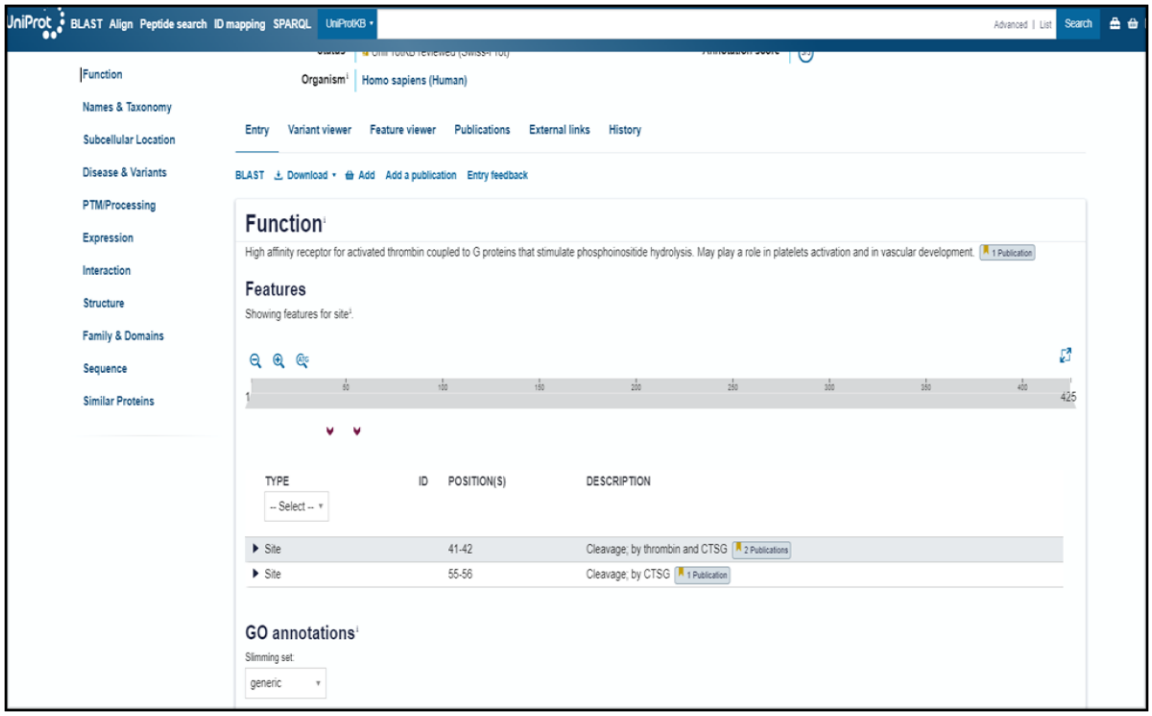


Figure 4: P25116 protein present in Human searched shows functions and features G protein-coupled high-affinity receptor for active thrombin that promotes phosphoinositide hydrolysis. may be involved in vascular growth and platelet activation.

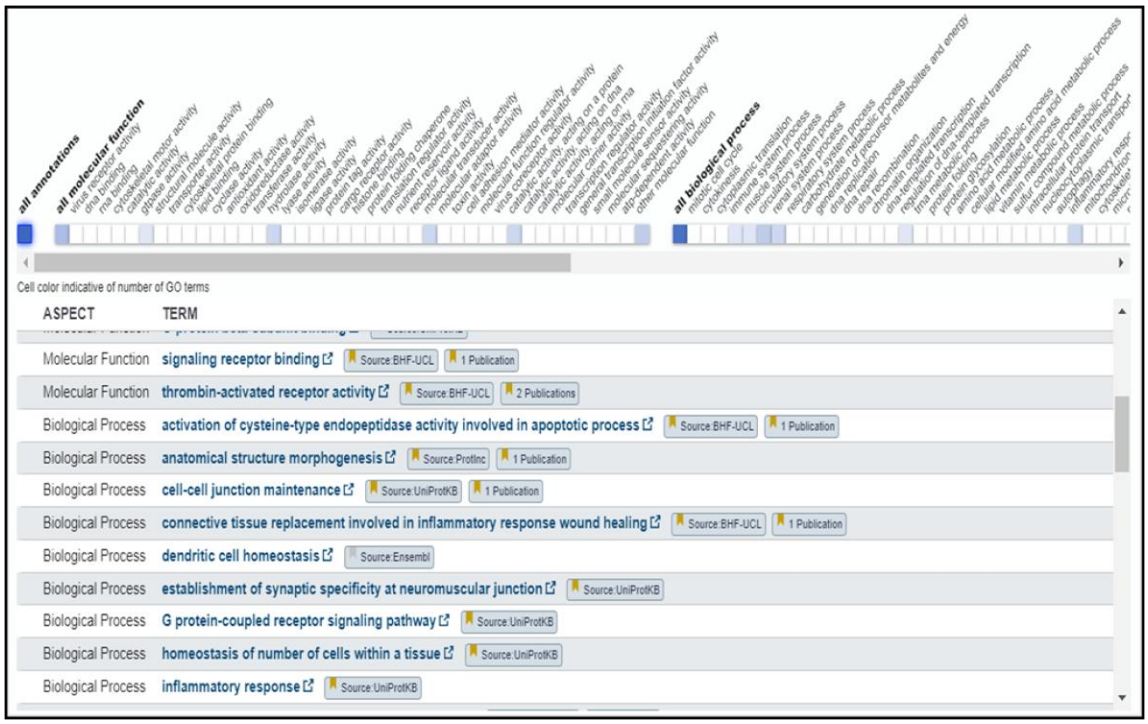


Figure 4.a: Number of Annotations and all molecular functions of site

Names & Taxonomy

Protein names

Recommended name	Proteinase-activated receptor 1
Short names	PAR-1
Alternative names	Coagulation factor II receptor Thrombin receptor

Gene names

Name	F2R
Synonyms	CF2R, PAR1, TR

Organism names

Organism	Homo sapiens (Human)
Taxonomic identifier	9606 NCBI
Taxonomic lineage	cellular organisms > Eukaryota (eucaryotes) > Opisthokonta > Metazoa (metazoans) > Eumetazoa > Bilateria > Deuterostomia > Chordata (chordates) > Craniata > Vertebrata (vertebrates) > Gnathostomata (jawed vertebrates) > Teleostomi > Euteleostomi (bony vertebrates) > Sarcopterygii > Dipnotetrapodomorpha > Tetrapoda (tetrapods) > Amniota (amniotes) > Mammalia (mammals) > Theria > Eutheria (placentals) > Boreoeutheria > Euarchontoglires > Primates > Haplorrhini > Simiiformes > Catarrhini > Hominoidea (apes) > Hominidae (great apes) > Homininae > Homo

Accessions

Primary accession	P25116
Secondary accessions	Q53XV0 Q96RF7 Q9BUN4

Proteomes

Identifier	UP000005640
Component	Chromosome 5

Organism-specific databases

HGNC	HGNC:3537 F2R	VEuPathDB	HostDB:ENSG00000181104
MIM	187930 gene	neXtProt	NX_P25116

Figure 5: Name and Taxonomy of PAR – 1

Structure

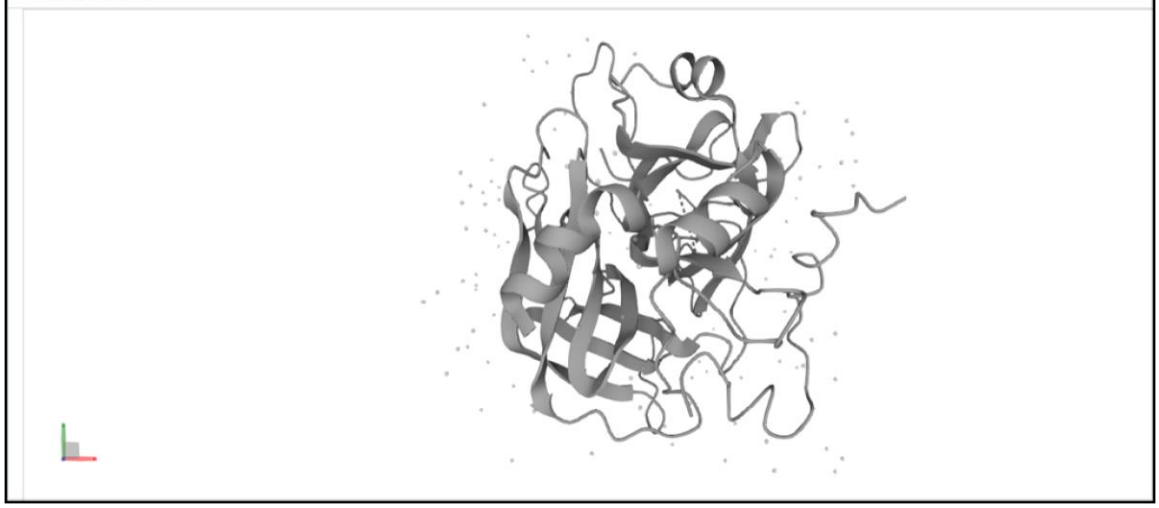
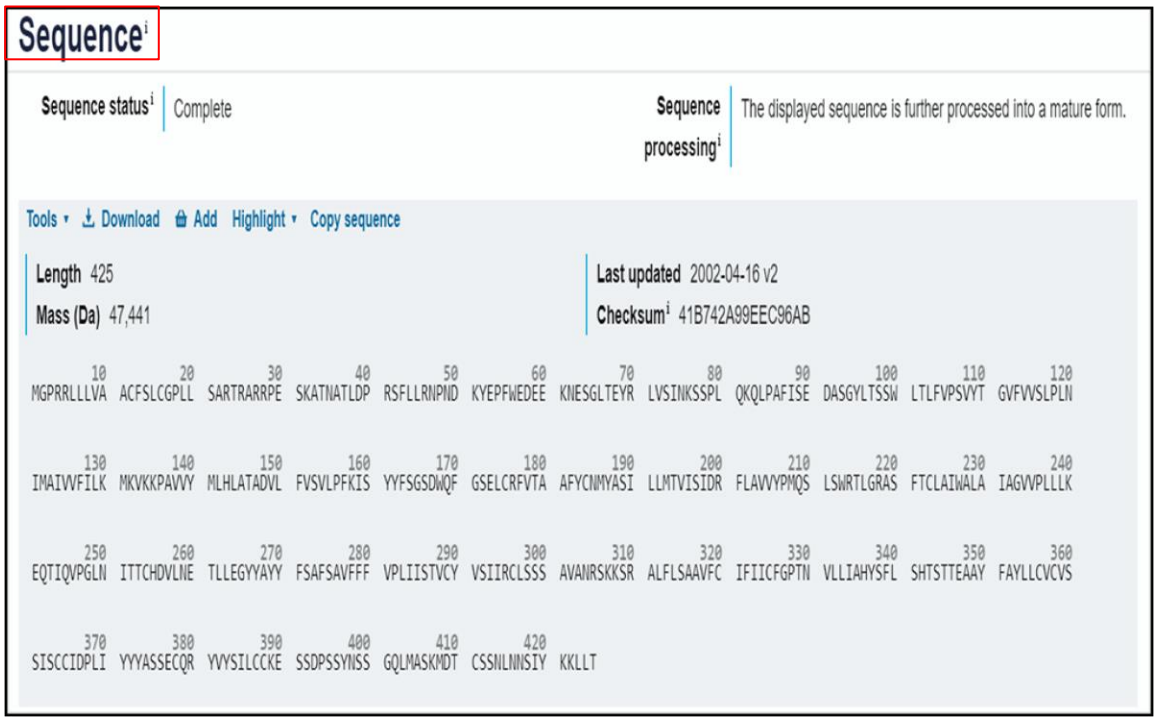


Figure 6: Structure of PAR – 1



Sequence

Sequence status: Complete

Sequence processing: The displayed sequence is further processed into a mature form.

Tools ▾ Download Add Highlight ▾ Copy sequence

Length 425

Mass (Da) 47,441

Last updated 2002-04-16 v2

Checksum: 41B742A99EEC96AB

```
        10         20         30         40         50         60         70         80         90        100        110        120
MGPRRLLLVA ACFSLCGPLL SARTRARRPE SKATNATLDP RSFLLRNPND KYEPFWEDEE KNESGLTEYR LVSINKSSPL QKQLPAFISE DASGYLTSSW LTLFVPSVYT GVFVVSLPLN

       130        140        150        160        170        180        190        200        210        220        230        240
IMAIVVFILK MKVKKPAVVY MLHLATADVL FVSVLPFKIS YYFSGSDWQF GSELCRFVTA AFYCNMYASI LLMTVISIDR FLAVVYPMQS LSWRTLGRAS FTCLAIWALA IAGVVPLLLK

       250        260        270        280        290        300        310        320        330        340        350        360
EQTIQVPGLN ITTCHDVLNE TLLEGYYAYY FSAFSAVFFF VPLIISTVCY VSIIRCLSSS AVANRSKKSR ALFLSAAVFC IFIICFGPTN VLLIAHYSFL SHTSTTEAAY FAYLLCVCVS

       370        380        390        400        410        420
SISCCIDPLI YYYASSECQR YVYSILCCKE SSDPSSYNSS GQLMASKMDT CSSNLNNSIY KKLLT
```

Figure 7: Sequence of PAR – 1

RESULTS:

The first entry in the 7,674 findings for thrombin protein hits is a *Homo sapiens* (human) creature with 425 amino acids. Activated thrombin-coupled G protein receptor affinity is shown using a function filter; the positions 41 and 42 indicate thrombin and CTSC cleavage, and the positions 55 and 56 exhibit CTSC cleavage. Name and taxonomy indicate proteinase activated receptor 2 (PAR-1) and altermatic name is coagulation factor 2 receptor gene name (F2R accession main ID is P25116).

CONCLUSION:

The UniProt, Swiss-Prot and TrEMBL databases were explored for the query thrombin protein (Accession ID: P25116) and related information was searched.

REFERENCES:

1. Esmon, C. T. (1995, July). Thrombomodulin as a model of molecular mechanisms that modulate protease specificity and function at the vessel surface. *The FASEB Journal*, *9*(10), 946–955. https://doi.org/10.1096/fasebj.9.10.7615164
2. Narayanan S. (1999). Multifunctional roles of thrombin. *Annals of clinical and laboratory science*, *29*(4), 275–280.https://pubmed.ncbi.nlm.nih.gov/10528826
3. *UniProt*. (n.d.). https://www.uniprot.org/

DATE: 30/09/2023

WEBLEM 3(E)
PROTEIN INFORMATION RESOURCE (PIR) DATABASE
(URL: https://proteininformationresource.org/)

AIM:
To explore the PIR (Protein Information Resource) Database for the further study of the query casein (PRO ID – PR: 000028855) under various categories.

INTRODUCTION:
The Protein Information Resource (PIR) Database is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies.
PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information. PIR is the most extensively classified protein sequence database. Classification of protein sequences into superfamilies and families aids scientists in searching against gene families and in determining the functional and evolutionary relationships among family members.
Dr. Dayhoff and her research group pioneered in the development of computer methods for the comparison of protein sequences, for the detection of distantly related sequences and duplications within sequences, and for the inference of evolutionary histories from alignments of protein sequences. PIR has provided protein databases and analysis tools freely accessible to the scientific community including the Protein Sequence Database (PSD).
PIR has the following 3 resources:

1. PRO
2. iPTMnet
3. iproLINK

1. PRO (Protein Resource Ontology)
PRO provides an ontological representation of protein-related entities by explicitly defining them and showing the relationships between them. Each PRO term represents a distinct class of entities (including specific modified forms, orthologous isoforms, and protein complexes) ranging from the taxon-neutral to the taxon-specific (e.g. the entity representing all protein products of the human SMAD2 gene is described in PR: Q15796; one particular human SMAD2 protein form, phosphorylated on the last two serines of a conserved C-terminal SSxS motif is defined by PR:000025934).
PRO encompasses three sub-ontologies: proteins based on evolutionary relatedness (ProEvo); protein forms produced from a given gene locus (ProForm); and protein – containing complexes (ProComp).

2. iPTMnet (PTMs = Protein Post–Translational Modification)
iPTMnet is a bioinformatics resource for integrated understanding of protein post–translational modifications (PTMs) in systems biology context.

It connects multiple disparate bioinformatics tools and systems text mining, data mining, analysis and visualization tools, and databases and ontologies into an integrated cross–cutting research resource to address the knowledge gaps in exploring and discovering PTM networks.

3. **iproLINK (integrated Protein Literature Information and Knowledge)**
   iproLINK (integrated Protein Literature Information and Knowledge) is a resource with access to text mining tools and annotated corpora developed in house. The collection of data sources can be utilized by computational and biological researchers to explore literature information on proteins and their features or properties.

Text Mining Tools:

1. **iTextMine –** integrated text mining tools and relation extraction results from large–scale text processing.
2. **pGenN –** a gene normalization tool tailored for plants.
3. **miRTex –** a relation extraction tool that identifies miRNA–target relations as well as miRNA–gene and gene–miRNA regulation relations.
4. **eFTP –** a relation extraction tool that identifies information relevant to phosphorylated proteins and phosphorylation–dependent protein–protein interactions.
5. **emiRIT –** an integrative text mining system collecting miRNA information from the literature

Casein:

Casein, also known as calcium caseinate and casein protein isolate, is a protein found in milk that gives milk its white color. Cow's milk consists of around 80% casein protein. In addition to milk, casein protein is found in yoghurt, cheese and infant formulas, as well as in a variety of dietary supplements. Unlike casein, casein peptides are made by breaking casein protein down into smaller pieces. Casein protein can be consumed to improve athletic performance, nutrition and to treat diabetes, liver disease due to alcohol consumption, and many other conditions.

METHODOLOGY:

1. Enter the PIR Homepage. Following 3 different types of databases can be viewed within the PIR Homepage – PRO, iPTMnet and iproLINK.

A. **Methodology for PRO (Protein Resource Oncology):**

1. Select the PRO database on the homepage of PIR.
2. Search for the query casein and note down the number of hits that appear.
3. Apply limits on the quick browse and search either a string / ID in the view entry in DAG to retrieve information for the protein query. Note down the total number of hits after applying limits.
4. Select any one node from the category, for instance – family, gene, etc. to further view information regarding the query.

B. Methodology for iPTMnet:

1. Select the iPTMnet database on the homepage of PIR.
2. Search for the query casein in iPTMnet database and note down the number of hits that appear.
3. Apply the limits for instance, PTM Type and Restrict by Organism – and note down the number of hits obtained for the query searched.
4. Select any one entry to study iPTMnet report.

C. Methodology for iproLINK:

1. Select the iproLINK database on the homepage of PIR.
2. Use the iTextMine as the Text Mining Tool.
3. Enter the keyword or PMID to retrieve information regarding the query casein and note down the number of hits (documents) obtained.
4. Apply limits, for instance, Query Type and Collection Type and note down the number of hits obtained.
5. Select any one entry to view information for the respective document.

OBSERVATIONS:

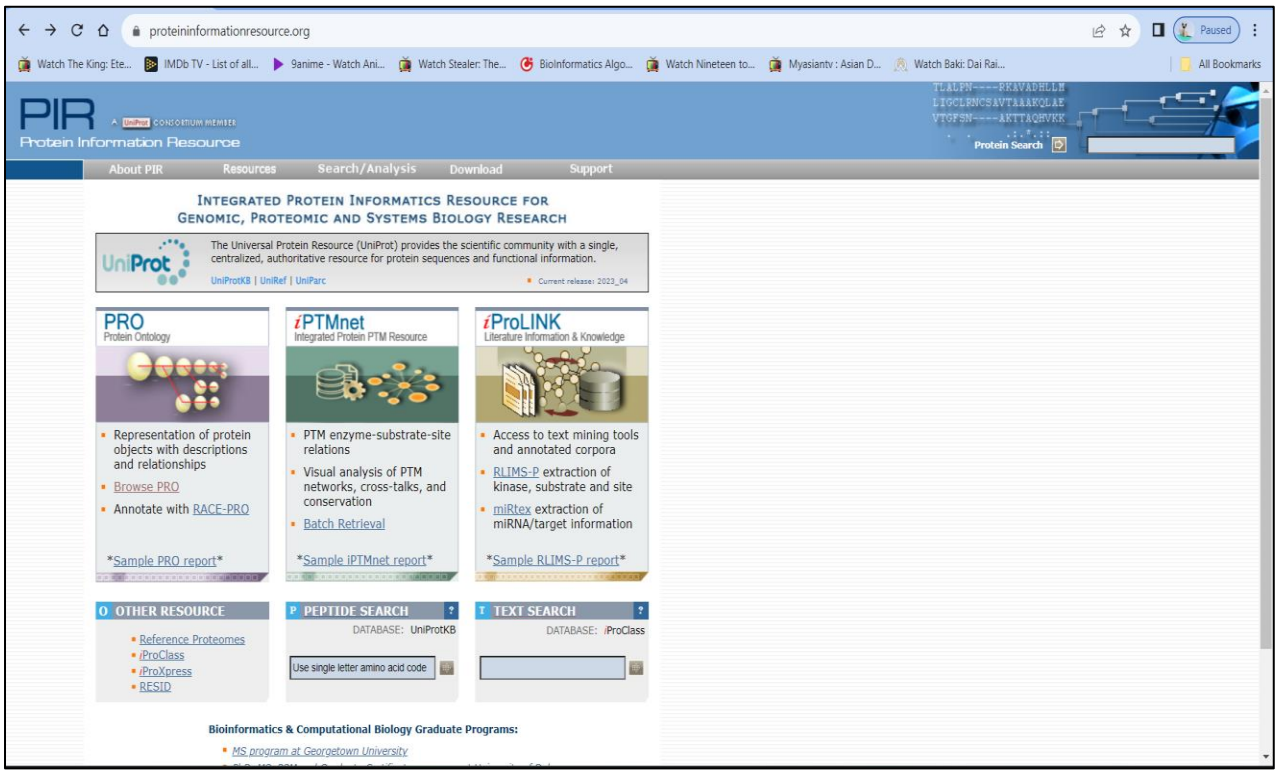


Figure 1: Homepage of PIR (Protein Information Resource) Database

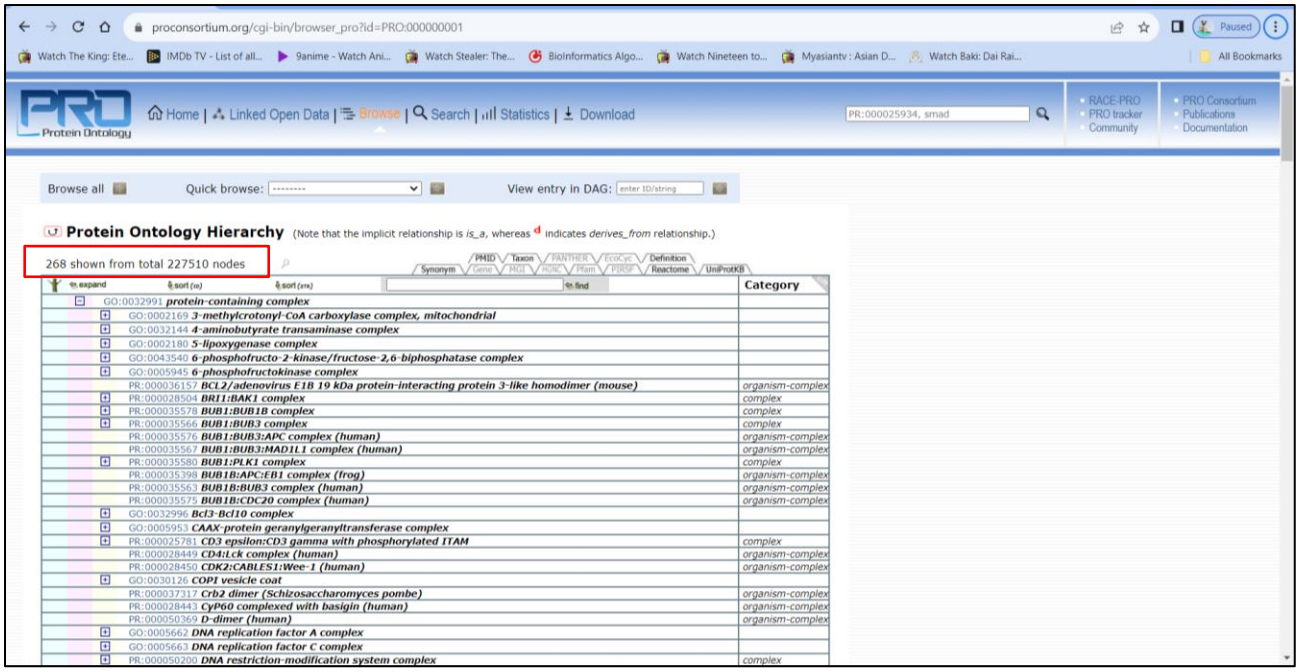


Figure 2: Query – Casein searched on the PRO (Protein Resource Oncology) Database.

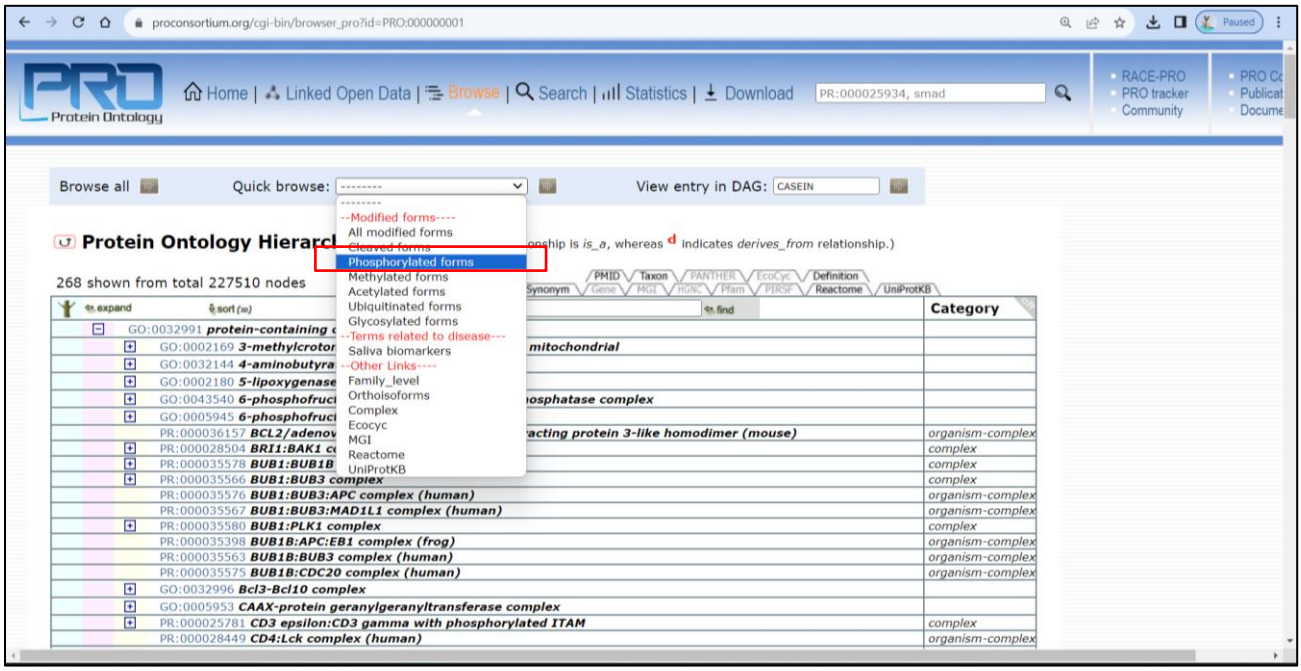


Figure 2a: Applying limits to obtain information regarding the phosphorylated forms of the query from the Quick Browse option in the PRO (Protein Resource Oncology) Database.

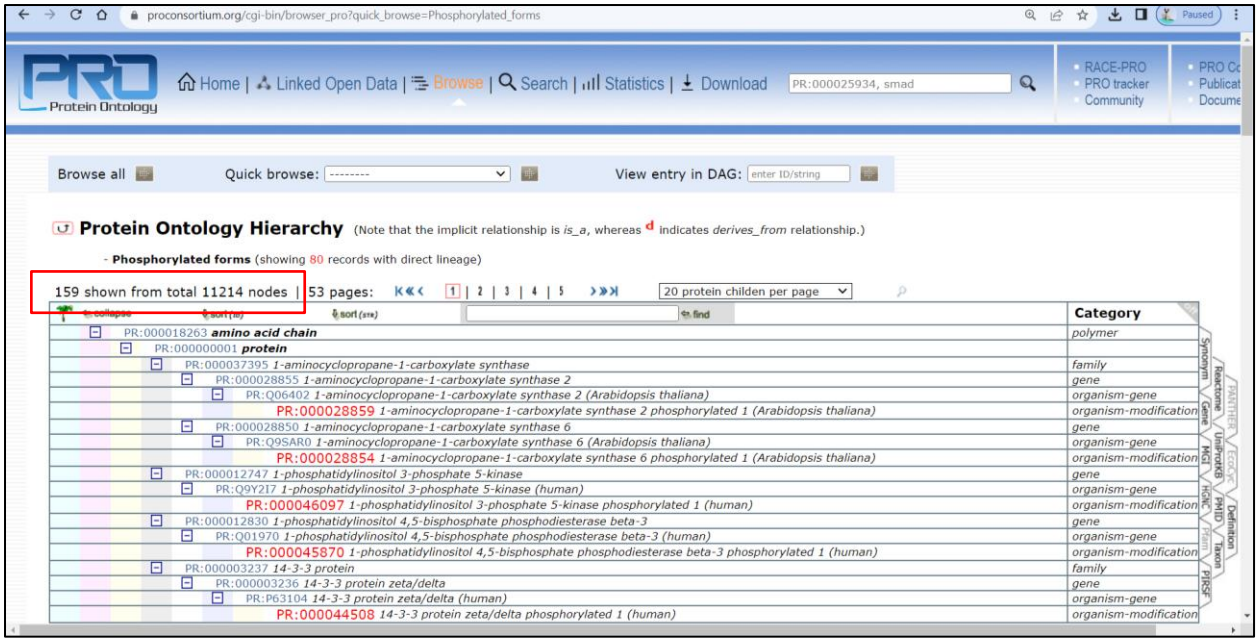


Figure 2b: Number of hits obtained after applying limits (phosphorylated forms)

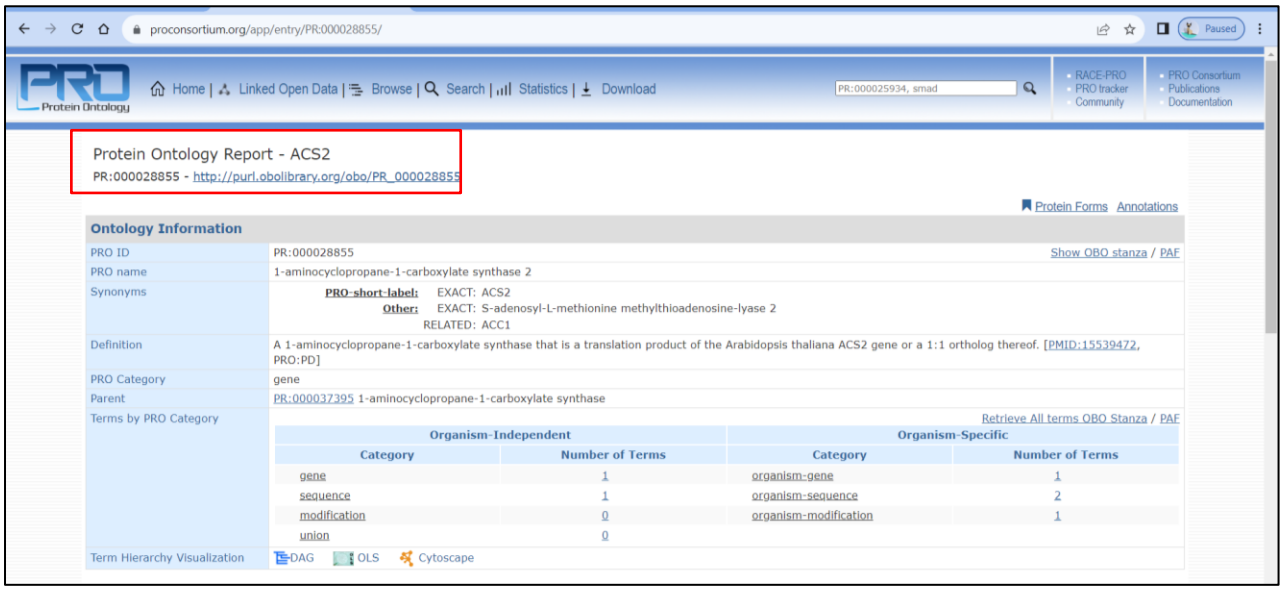


Figure 2c: To retrieve the Protein Ontology Report for ACS2 (PR: 000028855) under the category gene (node)

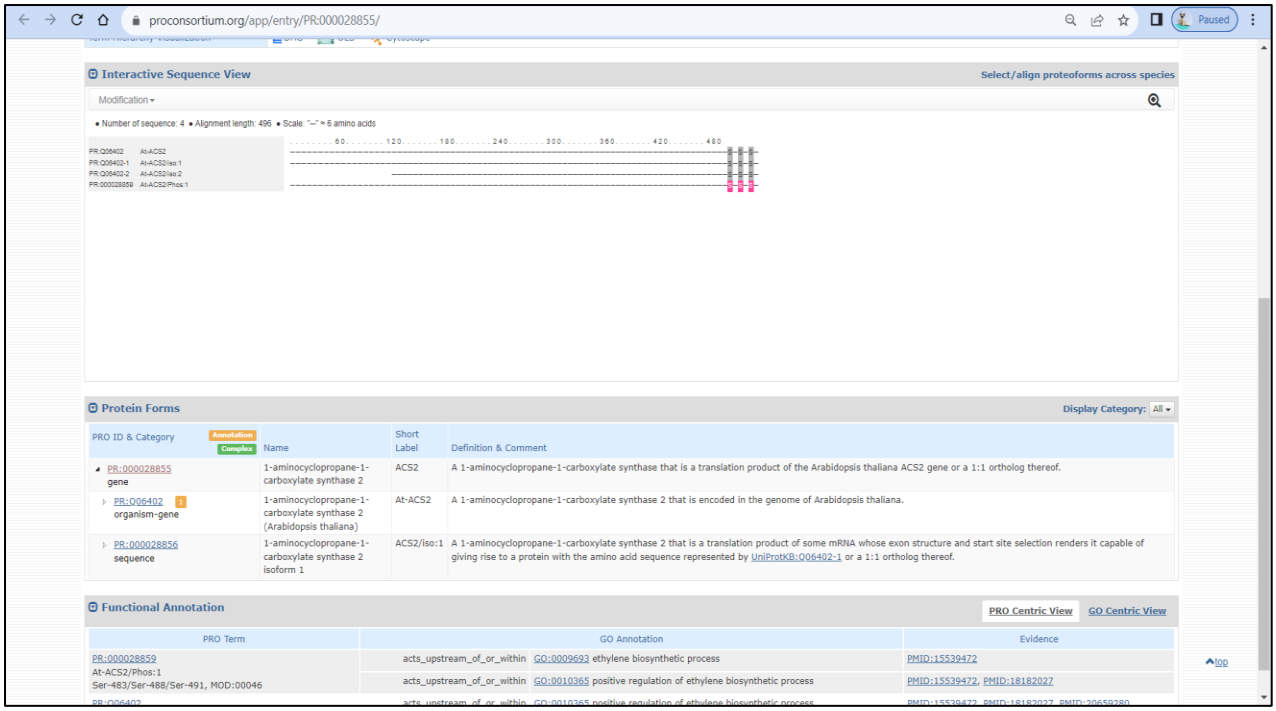


Figure 2d: Information regarding interactive sequence view, protein forms, functional annotation

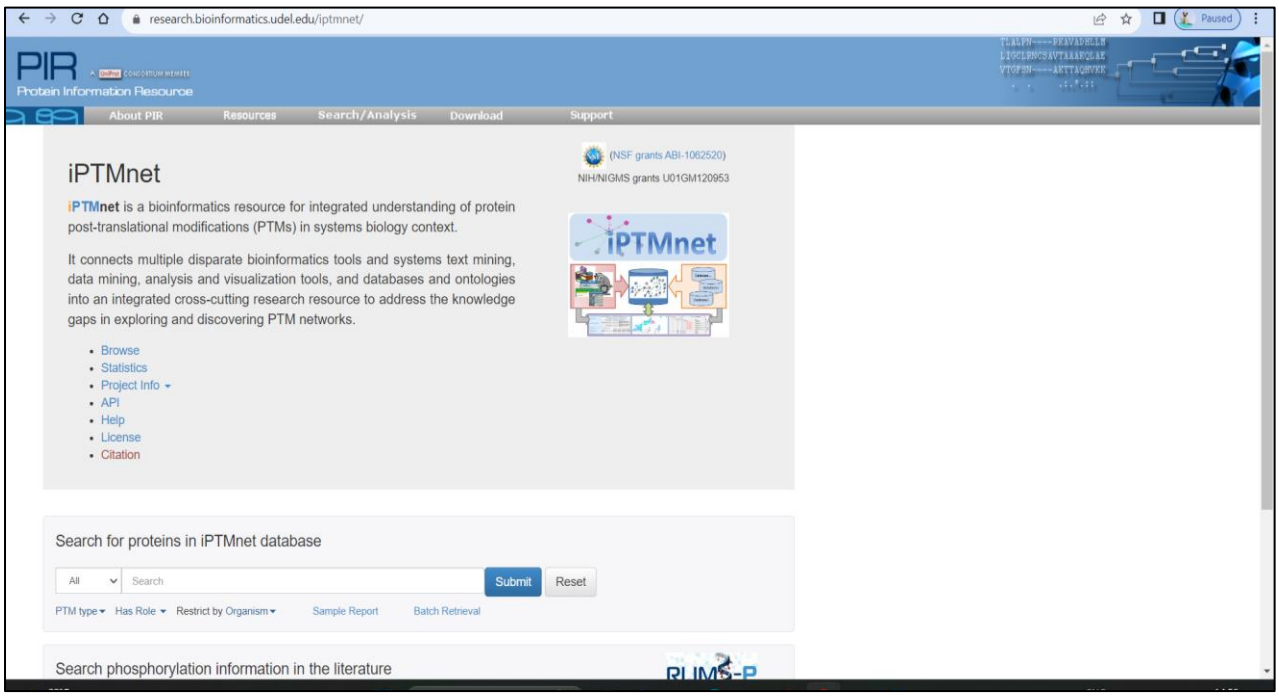


Figure 3: Homepage of iPTMnet (PTMs = Protein Post–Translational Modification) Resource

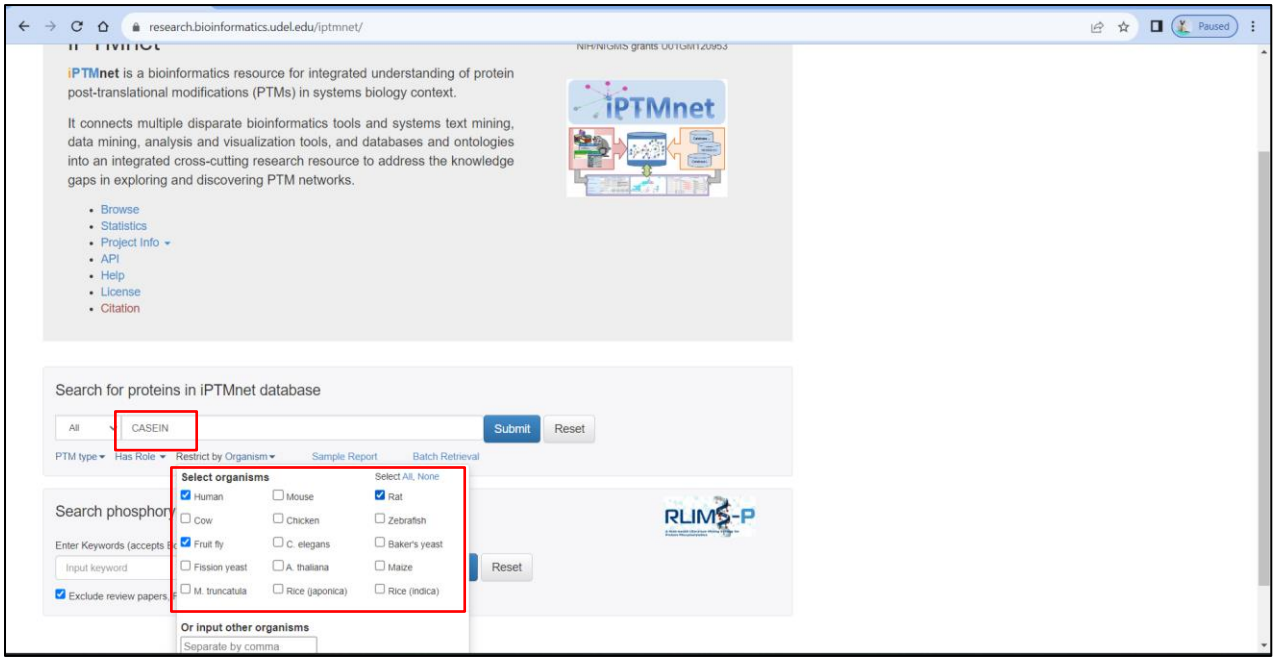


Figure 3a: Query – Casein searched on the iPTMnet (PTMs = Protein Post–Translational Modification). Further limits applied (Restrict by Organism: Human, Fruit fly, Rat)

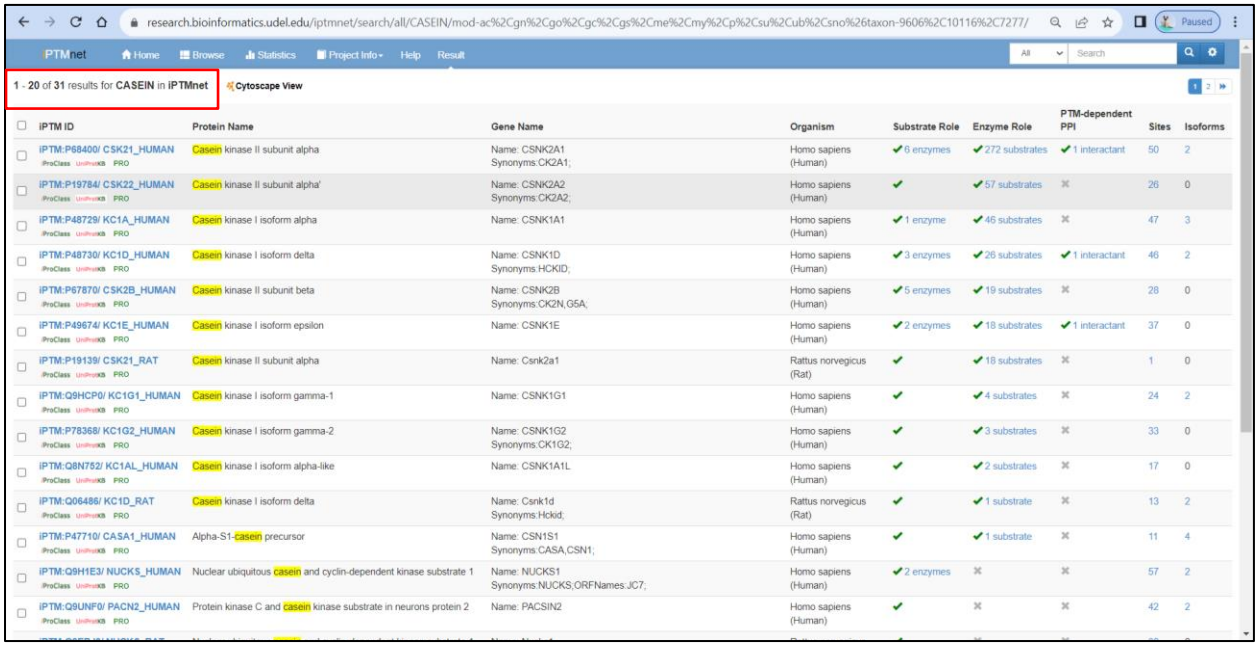


Figure 3b: View of Number of hits obtained for the query casein. For further study, iPTM ID: P684000/CSK21 – HUMAN was selected.

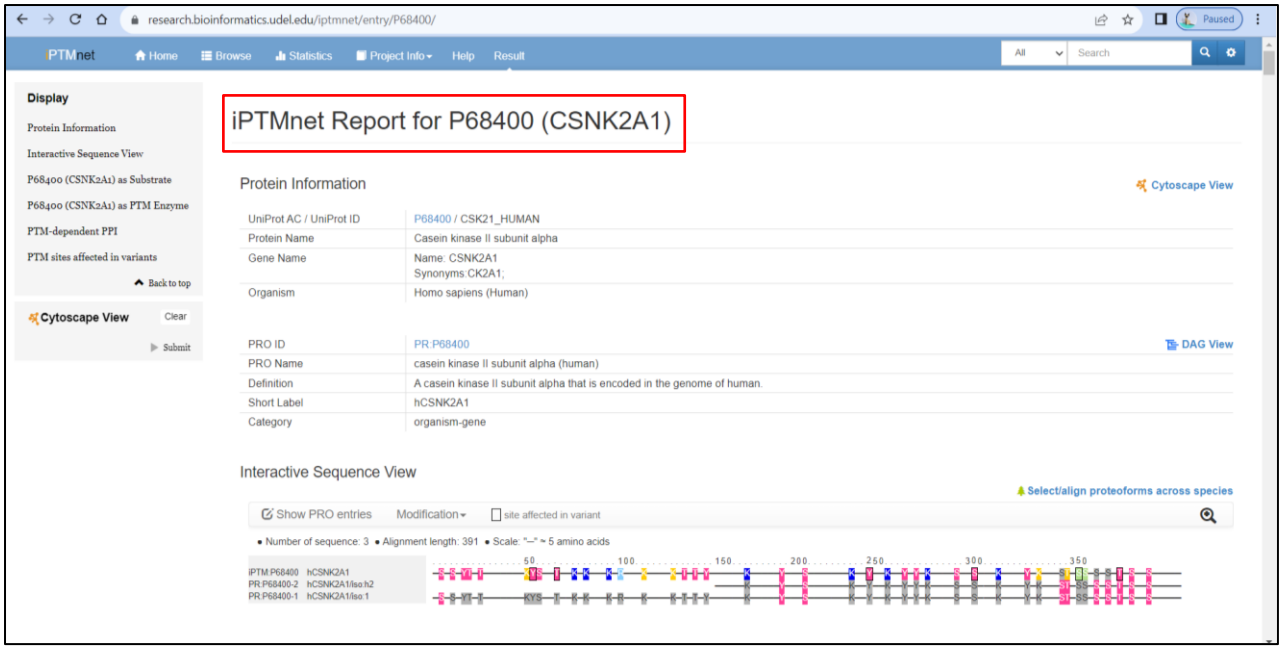


Figure 3c: Study of iPTMnet report displayed for the UniProt ID / UniProt AC: (P68400/CSK21 – HUMAN)

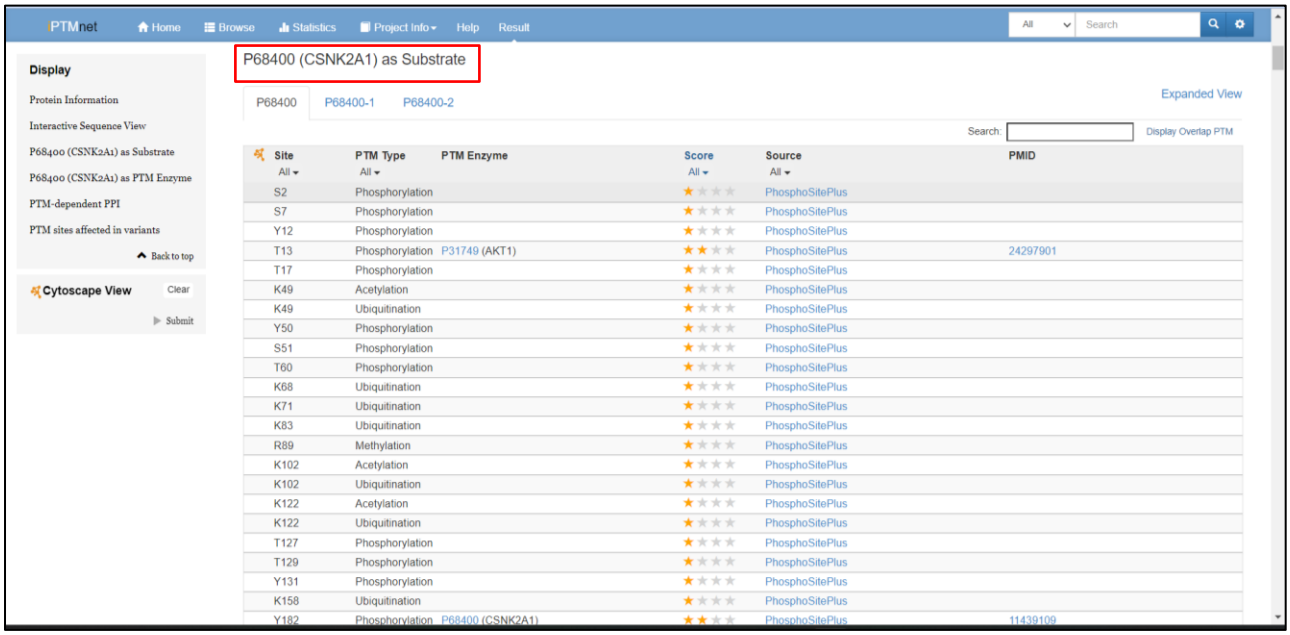


Figure 3d: Display of information regarding P68400 (CSNK2A1) as Substrate list, PTM Type, PTM Enzyme along with the rating score. Best result is represented with 4 stars.

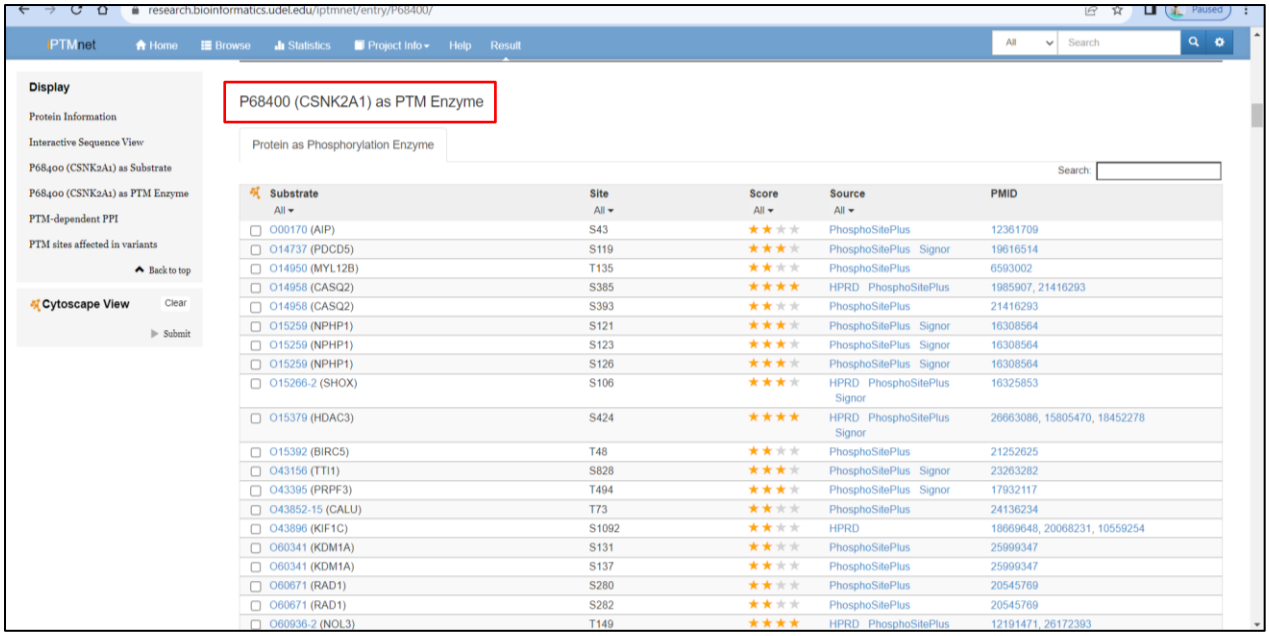


Figure 3e: Casein present in P68400 (CSNK2A1) as PTM Enzyme

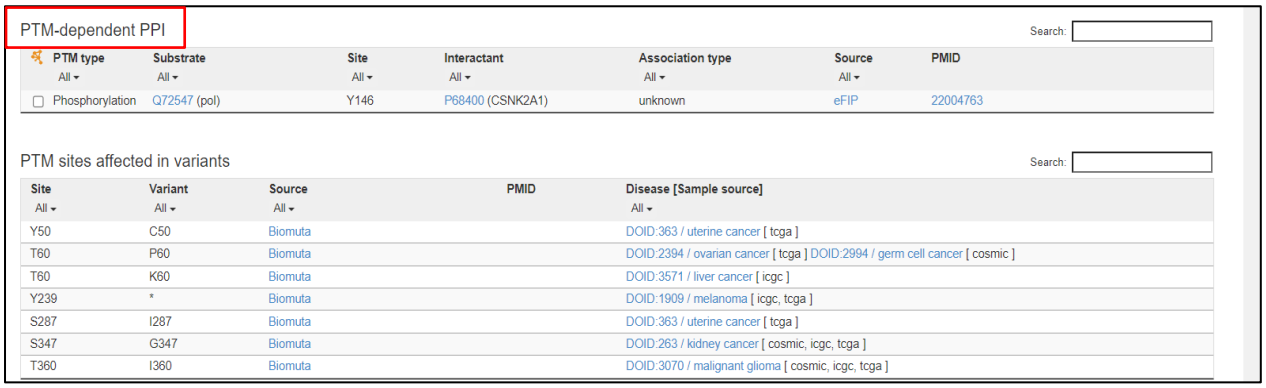


Figure 3f: View of PTM Dependent PPI. It displays PTM sites that are affected in variants and subsequently shows various functions and features.

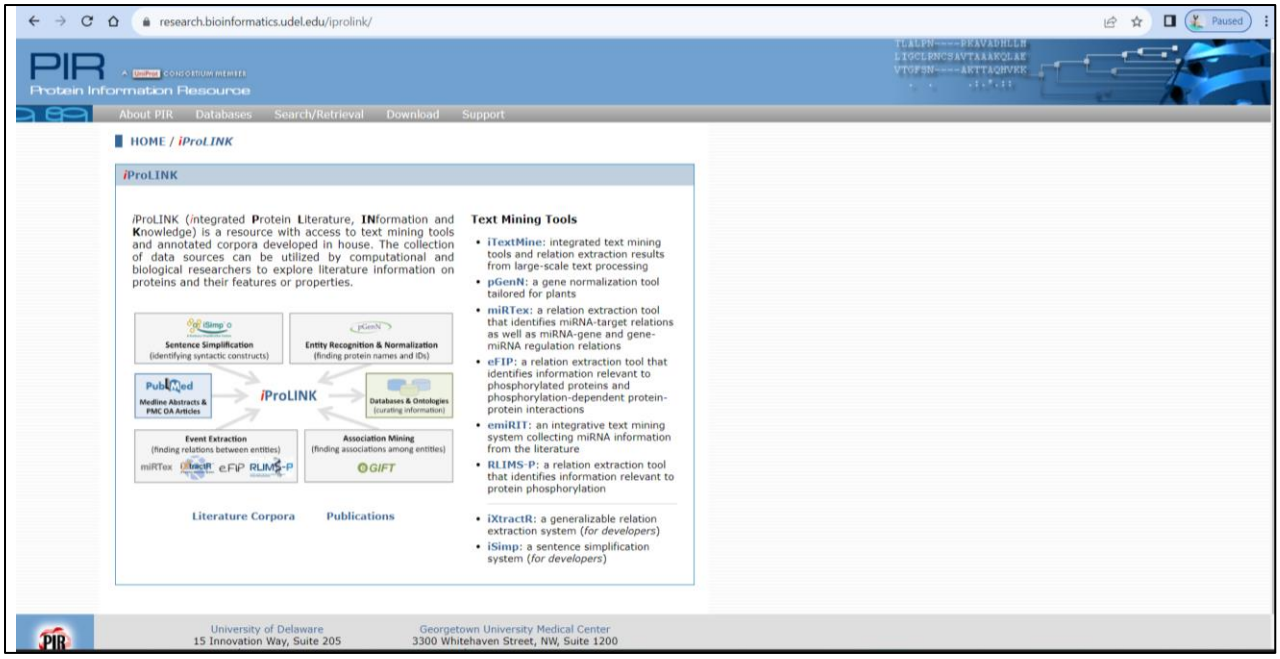


Figure 4: Homepage of iproLINK (integrated Protein Literature Information and Knowledge) Resource

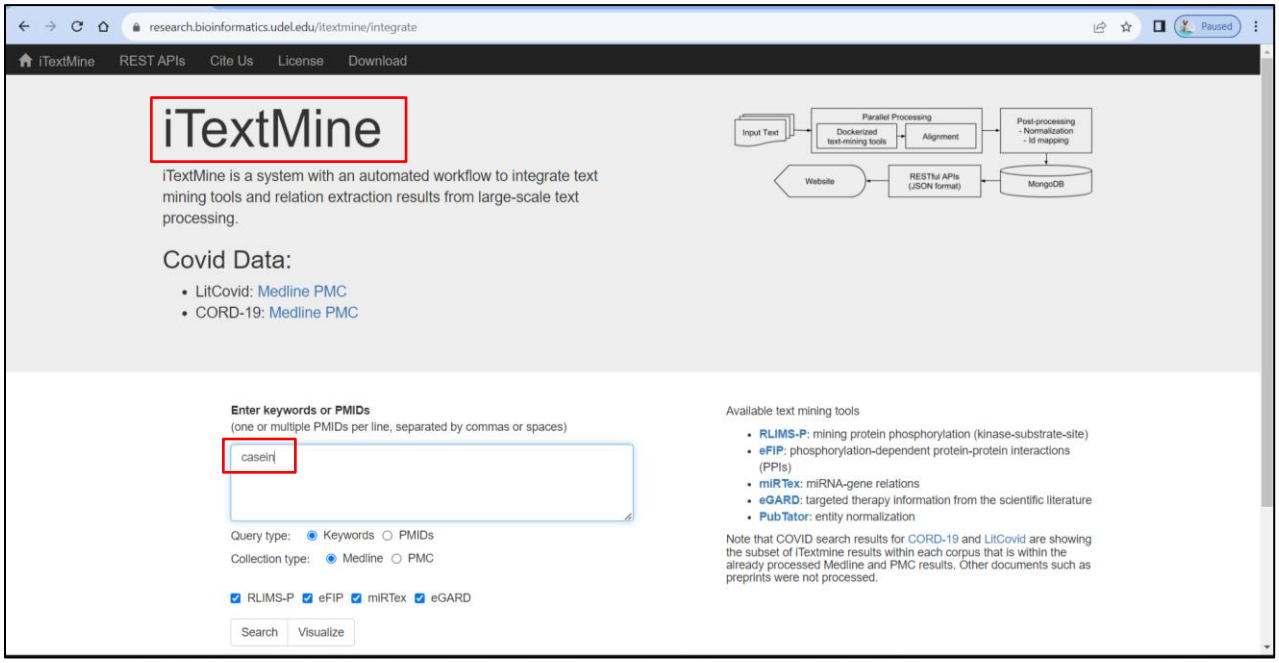
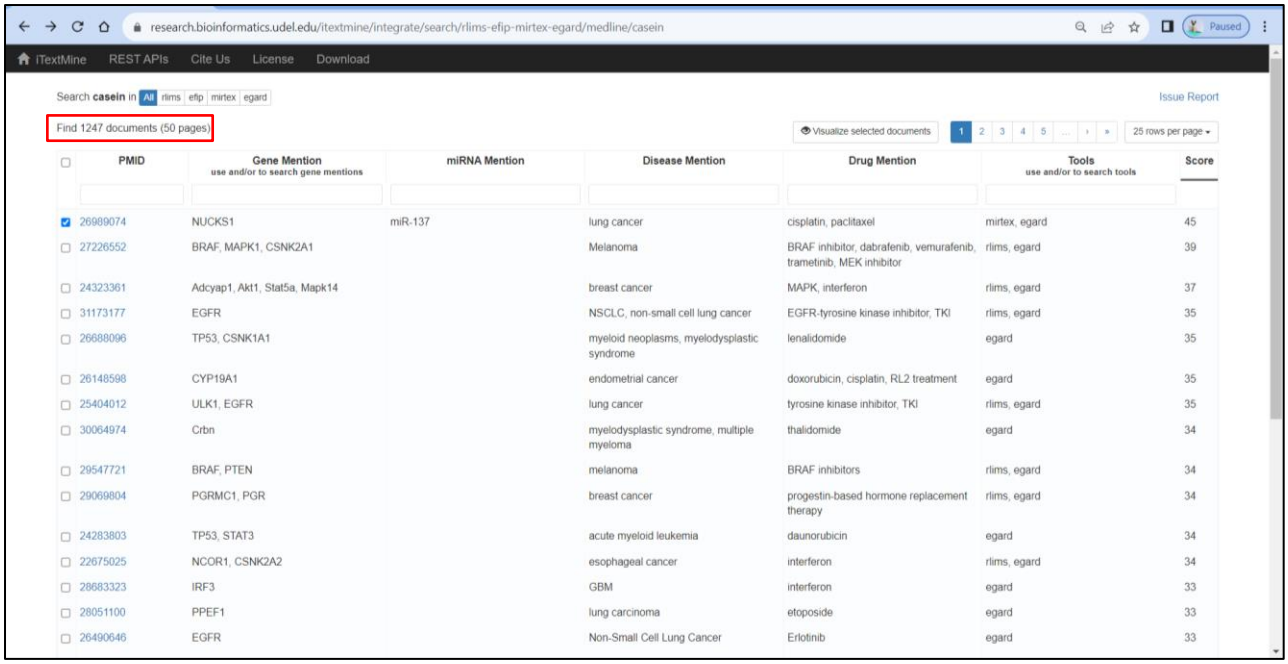


Figure 4a: After selecting the iTextMine as the text mining tool, the query casein is searched in the dialog box titled "Enter keywords or PMDs). Further limits applied: Query Type – keywords and Collection Type – Medline



PMID	Gene Mention (use and/or to search gene mentions)	miRNA Mention	Disease Mention	Drug Mention	Tools (use and/or to search tools)	Score
26989074	NUCKS1	miR-137	lung cancer	cisplatin, paclitaxel	mirtex, egard	45
27226552	BRAF, MAPK1, CSNK2A1		Melanoma	BRAF inhibitor, dabrafenib, vemurafenib, trametinib, MEK inhibitor	rlims, egard	39
24323361	Adcyap1, Akt1, Stat5a, Mapk14		breast cancer	MAPK, interferon	rlims, egard	37
31173177	EGFR		NSCLC, non-small cell lung cancer	EGFR-tyrosine kinase inhibitor, TKI	rlims, egard	35
26688096	TP53, CSNK1A1		myeloid neoplasms, myelodysplastic syndrome	lenalidomide	egard	35
26148598	CYP19A1		endometrial cancer	doxorubicin, cisplatin, RL2 treatment	egard	35
25404012	ULK1, EGFR		lung cancer	tyrosine kinase inhibitor, TKI	rlims, egard	35
30064974	Crbn		myelodysplastic syndrome, multiple myeloma	thalidomide	egard	34
29547721	BRAF, PTEN		melanoma	BRAF inhibitors	rlims, egard	34
29069804	PGRMC1, PGR		breast cancer	progestin-based hormone replacement therapy	rlims, egard	34
24283803	TP53, STAT3		acute myeloid leukemia	daunorubicin	egard	34
22675025	NCOR1, CSNK2A2		esophageal cancer	interferon	rlims, egard	34
28683323	IRF3		GBM	interferon	egard	33
28051100	PPEF1		lung carcinoma	etoposide	egard	33
26490646	EGFR		Non-Small Cell Lung Cancer	Erlotinib	egard	33

Figure 4b: Number of hits obtained for searching the query casein.

Results display information regarding – PMID, Gene Mention, miRNA Mention, Drug Mention, Tools, Score. For further study, PMID: 2698074 was selected.

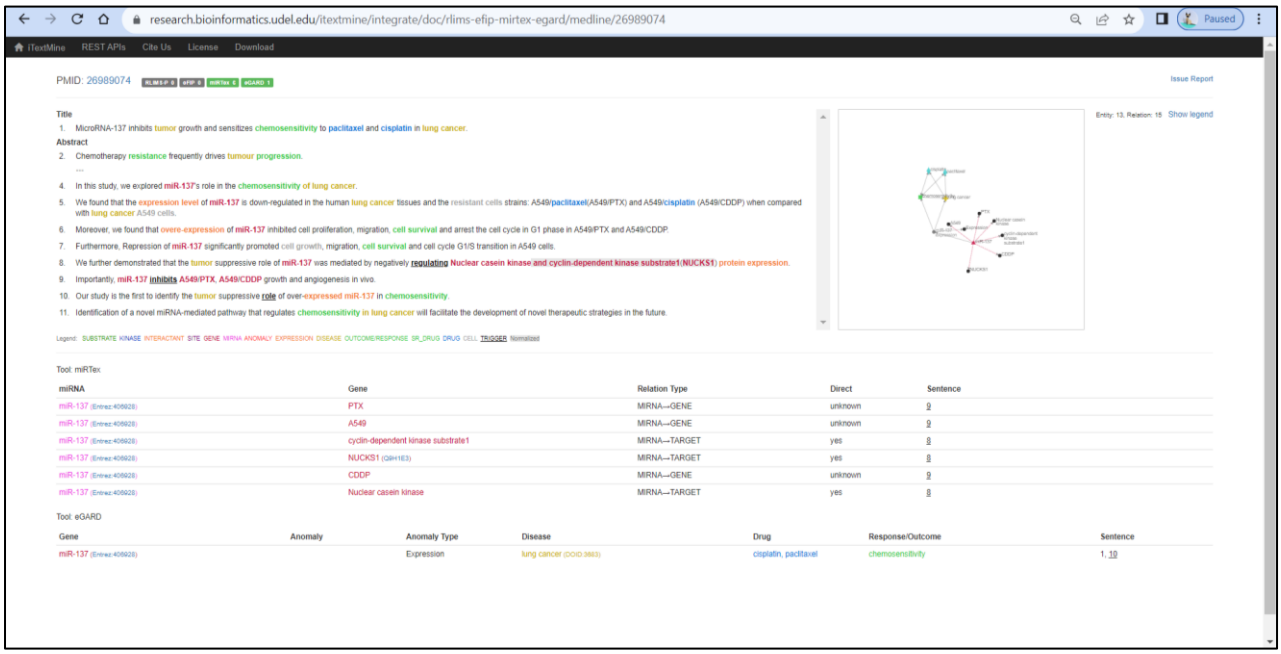


Figure 4c: Information regarding PMID: 26989074 is displayed. The title and the abstract of the research are mentioned using different colors for different legends. A diagram of the entity relation is also displayed.

RESULTS:

The query "casein" (PRO ID – PR: 000028855) was searched and explored in the PIR (Protein Information Resource) Database. Following hits were obtained –

Sr. No.	Resource	No. of hits obtained after applying limits
1	PRO (Protein Resource Oncology)	159
2	iPTMnet (PTMs = Protein Post – Translational Modification)	31
3	iproLINK (integrated Protein Literature Information and Knowledge)	1247

CONCLUSION:

PIR (Protein Information Resource) Database was viewed and explored for the query casein (PRO ID – PR: 000028855) and all the three resources - PRO (Protein Resource Ontology), iPTMnet (PTMs = Protein Post–Translational Modification and iproLINK (integrated Protein Literature Information and Knowledge) were studied for the related query.

REFERENCES:

1. *Protein Ontology*. (n.d.). https://proconsortium.org/pro.shtml
2. *iProLINK [PIR - Protein Information Resource]*. (n.d.). https://research.bioinformatics.udel.edu/iprolink/
3. Huang H, Arighi CN, Ross KE, Ren J, Li G, Chen SC, Wang Q, Cowart J, Vijay-Shanker K, Wu CH. iPTMnet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Res. 2018 Jan 4;46(D1): D542-D550. doi: https://doi.org/10.1093/nar/gkx1104.
   PMID: https://www.ncbi.nlm.nih.gov/pubmed/29145615;
   PMCID: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753337.
4. *Welcome to PIR [Protein Information Resource]*. (n.d.). https://proteininformationresource.org/
5. Wikipedia contributors. (2023, September 28). Casein. In *Wikipedia, The Free Encyclopedia*. Retrieved 00:25, October 6, 2023, from https://en.m.wikipedia.org/wiki/Casein

DATE: 30/09/2023

WEBLEM 4
DOMAIN DATABASES

AIM:

To study protein domains database for functional characterization and annotation.

INTRODUCTION:

Secondary databases refer to databases that are derived from primary databases, which include manually curated or computationally processed information. Secondary databases provide an added layer of information by curating, processing, and analyzing the raw data from primary databases. Protein databases have become a crucial part of modern biology, huge amount of data for protein structure, functions and particularly sequences are being generated. Comparison between protein and protein classification provide information about the relationships between protein within a genome or across different species. A protein domain is an independently folded, structurally compact unit that forms a steady 3D structure and shows a certain level of evolutionary conservation. A conserved domain contains one or more motifs. Protein sequence motif is a set of conserved amino acid residues that are important for protein function and are located within a certain distance from one another. These motifs usually provide clues to the functions of otherwise uncharacterized proteins.

PROSITE Database:

The PROSITE database consists of documentation entries describing protein domains, families, and functional sites as well as associated patterns and profiles to identify them. PROSITE database is a database of protein families, domains, and functional sites that contains manually curated information on amino acid patterns and profiles of proteins. It is a secondary protein database that provides tools for the analysis of protein sequences and the identification of motifs. The database contains a large collection of signature patterns or profiles that hold biological importance. Each signature is associated with important biological information such as protein family, domain, or functional site. PROSITE database uses two types of signatures, patterns, and generalized profiles, to identify conserved regions. These signatures can be used to predict the function and structure of proteins and help in the annotation of new protein sequences. PRINT is a database for protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family. PRINTS uses a fingerprinting method that detects distant relatives of large and highly divergent protein superfamilies by exploiting conserved regions within sequence alignments. SWISS-PROT database is a protein sequence database that provides high levels of annotations, including information on the protein's function, domain structure, post-translational modifications. Pfam database is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. Blocks are ungapped multiple alignments of related protein sequence segments that correspond to the most conserved regions of the proteins. The Blocks database is a collection of blocks representing known protein families that can be used to compare a protein or DNA sequence with documented families of proteins. SMART is a highly reliable

and sensitive tool for domain identification. COG is a database and a convenient tool for motif and domain identification.

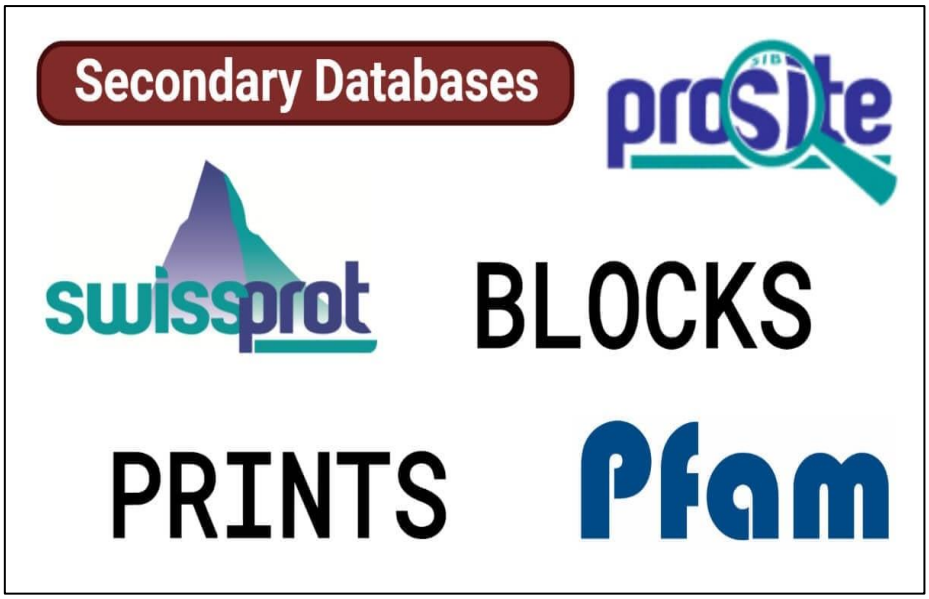


Figure 1: Different sites to study Protein domain databases.

<u>InterPro Database:</u>

Databases with signatures diagnostic for protein families, domains or functional sites are important tools for the computational functional classification of newly determined sequences that lack biochemical characterization. InterPro database is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. The InterPro database provides an integrative classification of protein sequences into families, and identifies functionally important domains and conserved sites. InterPro Scan is the underlying software that allows protein and nucleic acid sequences to be searched against InterPro's signatures. Signatures are predictive models which describe protein families, domains, or sites, and are provided by multiple databases. InterPro database combines signatures representing equivalent families, domains, or sites, and provides additional information such as descriptions, literature references and Gene Ontology (GO) terms, to produce a comprehensive resource for protein classification. InterPro database integrates 13 protein signature databases into one central resource: CATH-Gene3D, the Conserved Domains Database (CDD), HAMAP, PANTHER, Pfam, PIRSF, PRINTS, PROSITE Patterns, PROSITE Profiles, SMART, the Structure–Function Linkage Database (SFLD), SUPERFAMILY, TIGRFAMs and MobiDB. Pfam, focuses on divergent domains, PROSITE on functional sites and PRINTS focuses on families, specializing in hierarchical definitions from super-family down to sub-family levels to describe specific functions. Several sequence cluster databases, for example ProDom, are also commonly used in sequence analysis to facilitate domain identification. Unlike signature databases, the clustered resources are derived automatically from sequence databases, using different clustering algorithms. Databases like Blocks provide ungapped multiple alignments for protein families.

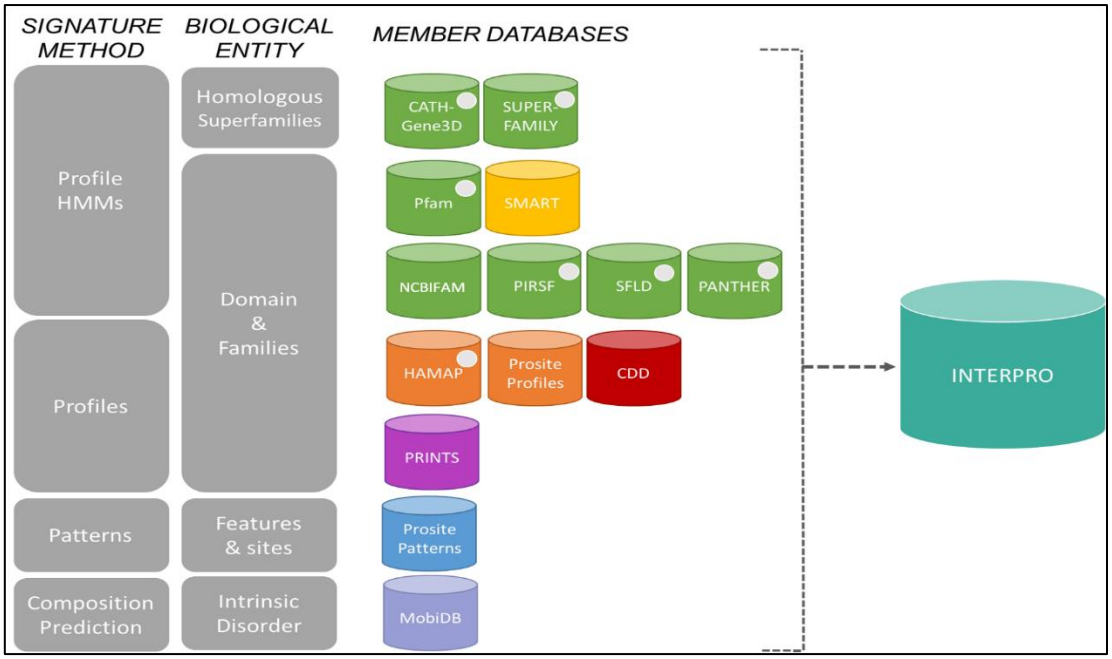


Figure 2: InterPro database integrates 13 protein signature databases into one central resource

REFERENCES:

1. Alok, K., & Shrivastava. (n.d.). Introduction to bioinformatics (databases) Course Code -BOTY 4204 Course Title-Techniques in plant sciences , biostatistics and bioinformatics. https://mgcub.ac.in/pdf/material/20200406015739416c3962e5.pdf
2. Secondary Databases - Bioinformatics. (2019, April 6). Microbe Notes. https://microbenotes.com/secondary-databases/
3. Magadh Mahila College – Patna University – Patna – Bihar. (n.d.). https://magadhmahilacollege.org/
4. Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2009). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Research, 38(suppl_1), D161–D166. https://doi.org/10.1093/nar/gkp885
5. Apweiler, R. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research, 29(1), 37–40. https://doi.org/10.1093/nar/29.1.37

DATE: 30/09/2023

WEBLEM 4(A)
PROSITE DATABASE
(URL: https://prosite.expasy.org/)

AIM:

To study protein domain for query 'Lectin' (UniProt ID: Q9LW83) in PROSITE database.

INTRODUCTION:

PROSITE is a database of protein families and domains. Database of protein families, protein domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterize them. It is based on the observation that, while there is a huge number of different proteins, most of them can be grouped, on the basis of similarities in their sequences, into a limited number of families. Properties from well-studied genes can be propagated to biologically related organisms, and for different or poorly known genes biochemical functions can be predicted from similarities. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. PROSITE database currently contains patterns and profiles specific for more than a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins. The database ProRule builds on the domain descriptions of PROSITE database. It provides additional information about functionally or structurally critical amino acids. The rules contain information about biologically meaningful residues, like active sites, substrate- or co-factor-binding sites, post-translational modification sites or disulfide bonds, to help function determination. These can automatically generate annotations based on PROSITE motifs. PROSITE database are used to identify and annotate specific protein features. It's a valuable resource for studying protein structure and function, aiding in tasks such as predicting protein function and detecting potential functional sites. Expasy is operated by the SIB Swiss Institute of Bioinformatics. PROSITE database is a widely used database of protein families and domains. The sequence of a protein is usually notated as a string of letters, to the order of the amino acids from the amino-terminal to the carboxyl-terminal of the protein. A protein domain is a region of a protein's polypeptide chain that is self-stabilizing and that folds independently from the rest. Each domain forms a compact folded three-dimensional structure's independent folding unit where as a motif is a chain-like biological structure made up of connectivity between secondary structural pieces.

Lectins:

Lectins are carbohydrate-binding proteins that are highly specific for sugar groups. Lectins have a role in recognition at the cellular and molecular level and play numerous roles in biological recognition phenomena involving cells, carbohydrates, and proteins. Lectins also mediate attachment and binding of bacteria, viruses, and fungi to their intended targets. Lectins can act as an antioxidant, which protects cells from damage caused by free radicals. They also slow down digestion and the absorption of carbohydrates, which may prevent sharp rises in blood sugar and high insulin levels. The use of non-toxic low amounts of certain

lectins to help stimulate gut cell growth in patients who are unable to eat for long periods, and in anticancer treatments due to the ability of lectins to cause cancer cell death.

Pattern syntax:

1. The standard IUPAC one letter code for the amino acids (http://www.bioinformatics.org/sms/iupac.html) is used in PROSITE.
2. The symbol 'x' is used for a position where any amino acid is accepted.
3. Ambiguities are indicated by listing the acceptable amino acids for a given position, between square brackets '[ ]'. For example: [ALT] stands for Ala or Leu or Th.
4. Ambiguities are also indicated by listing between a pair of curly brackets '{}' the amino acids that are not accepted at a given position. For example: {A}.
5. Each element in a pattern is separated from its neighbor by a '-'.
6. Repetition of an element of the pattern can be indicated by following that element with a numerical value or, if it is a gap ('x'), by a numerical range be
7. Examples:
    a. x(3) corresponds to x-x-x
    b. x(2,4) corresponds to x-x or x-x-x or x-x-x-x
    c. A(3) corresponds to A-A-A
8. When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern respectively starts with a '<' symbol or ends with a '>' symbol.
9. In some rare cases (e.g., PS00267 (/PS00267) or PS00539 (/PS00539), '>' can also occur inside square brackets for the C-terminal element. 'F-[GST].

Note:

1. Ranges can only be used with 'x', for instance 'A(2,4)' is not a valid pattern element.
2. Ranges of 'x' are not accepted at the beginning or at the end of a pattern unless restricted/anchored to respectively the N- or C-terminal of a sequence.

METHODOLOGY:

1. Go to the PROSITE database website.
2. Query can be searched by text, search by sequence, search by PRO ID engine.
3. Go to UniProt database and search query 'Lectin' in search Entrez.
4. Copy the query 'Lectin' sequence Q9LW83 ID of G-type lectin document from the list of documents.
5. Enter the query, 'Lectin' sequence in Quick Scan mode of ScanProsite and click on scan.
6. After searching query 'Lectin' we get a list of relevant information.
7. Analyze the sequence of "Lectin" for study of different protein domain, families and functional sites.

OBSERVATIONS:

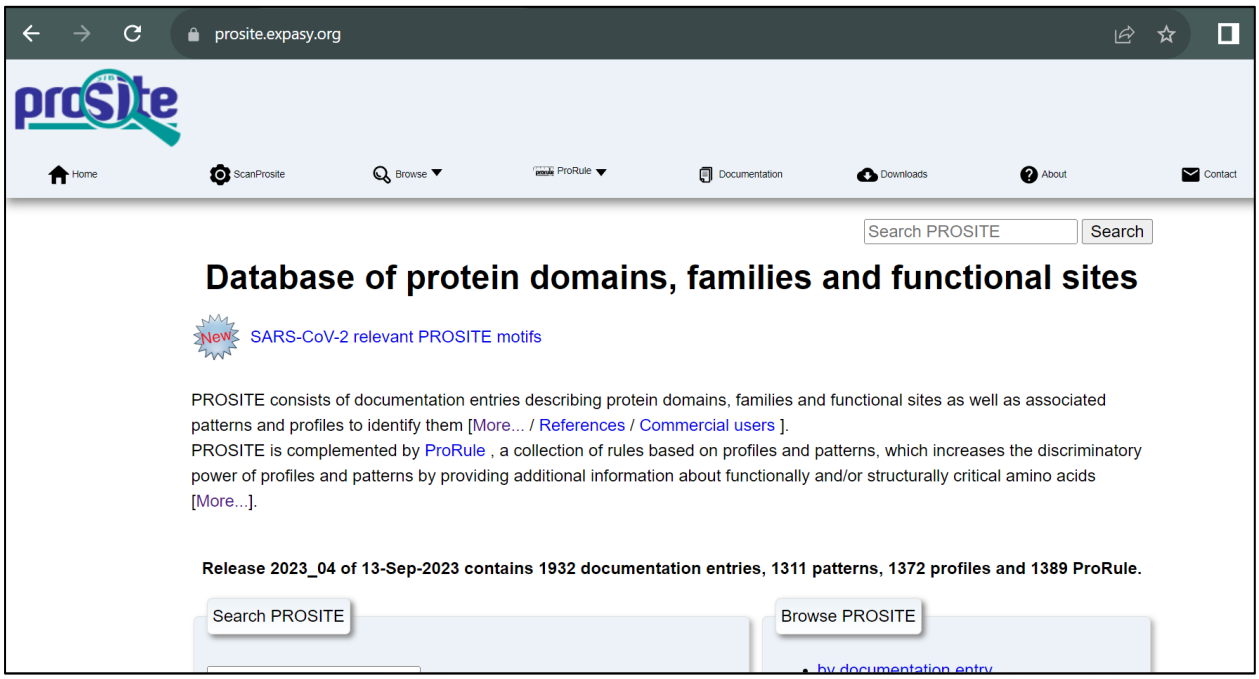


Figure 1: Homepage of PROSITE database

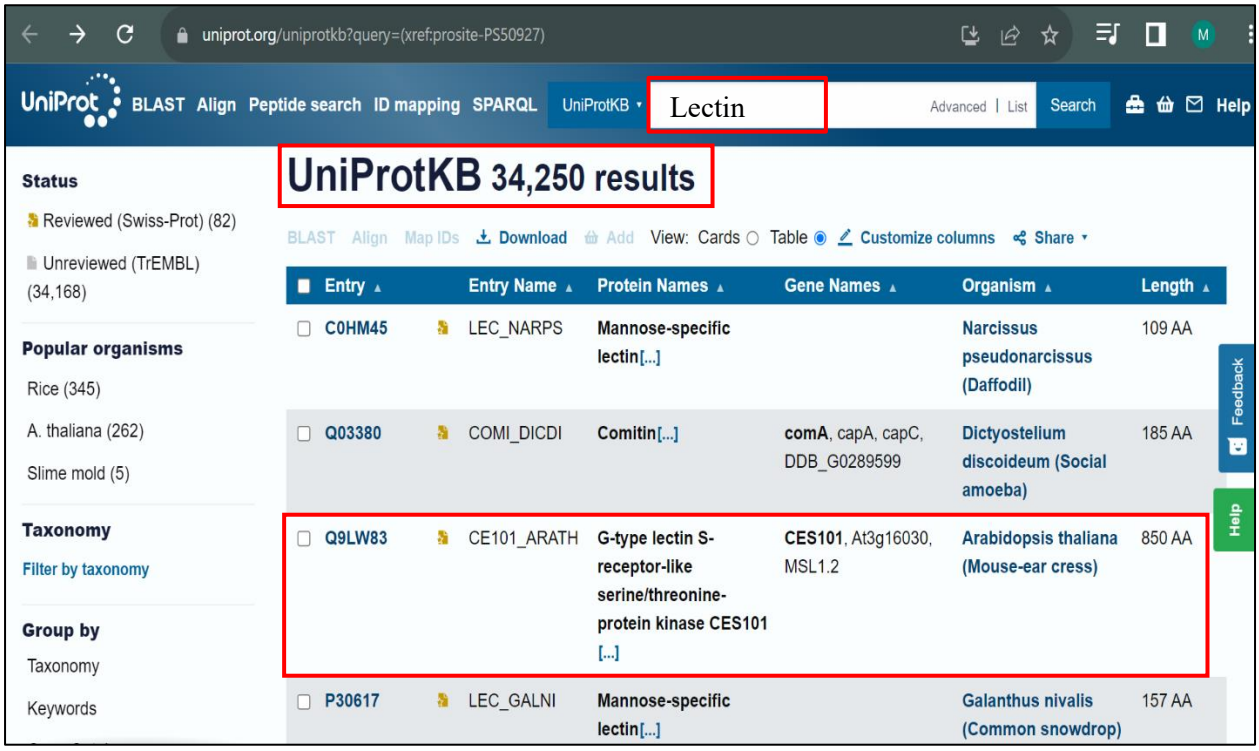


Figure 2: Search query 'Lectin' using UniProt database

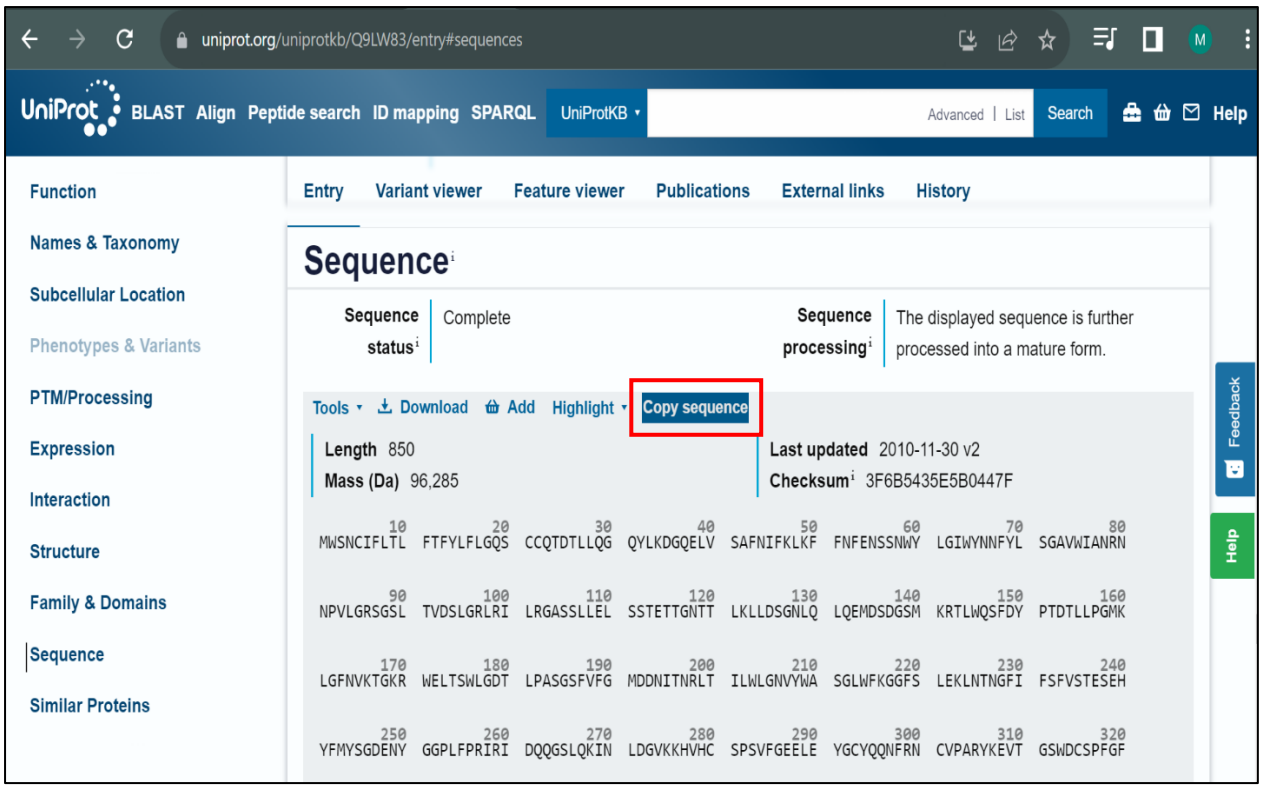


Figure 3: Query 'Lectin' sequence is copied from UniProt database

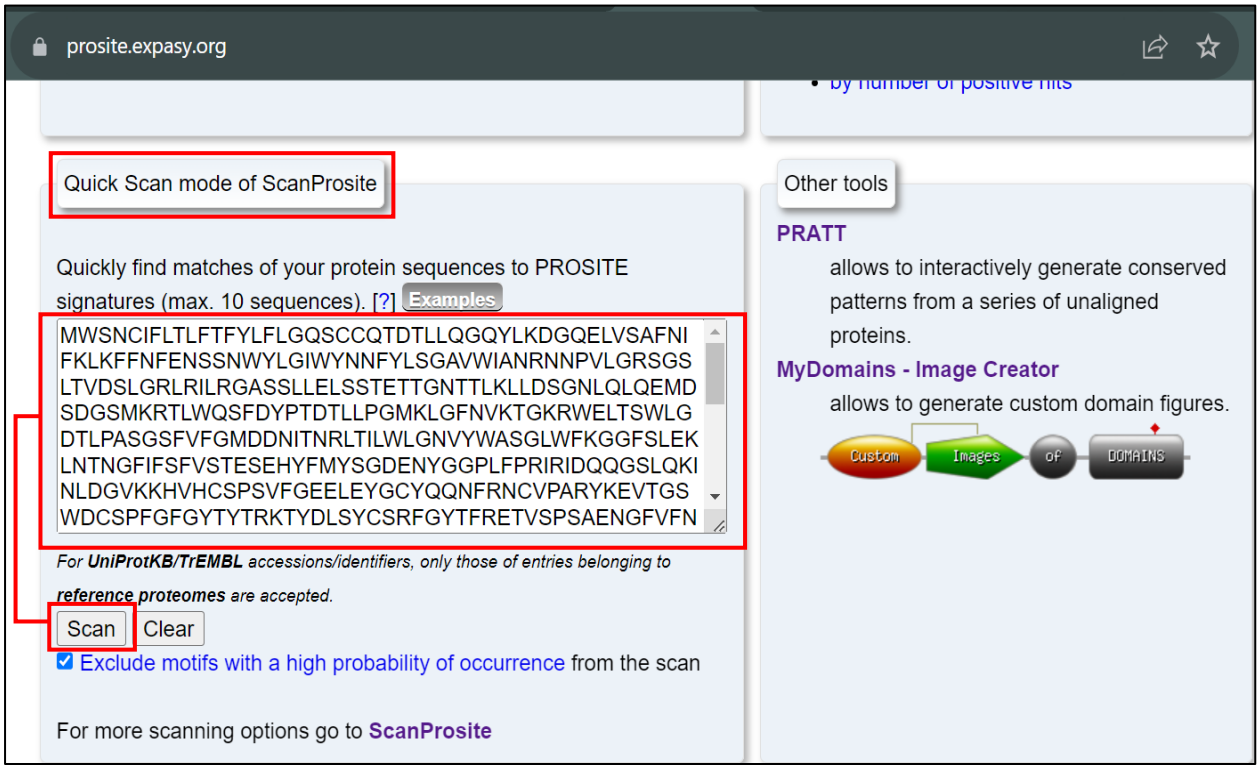


Figure 4: Query 'Lectin' sequence is searched in Quick Scan Mode of ScanProsite in PROSITE database

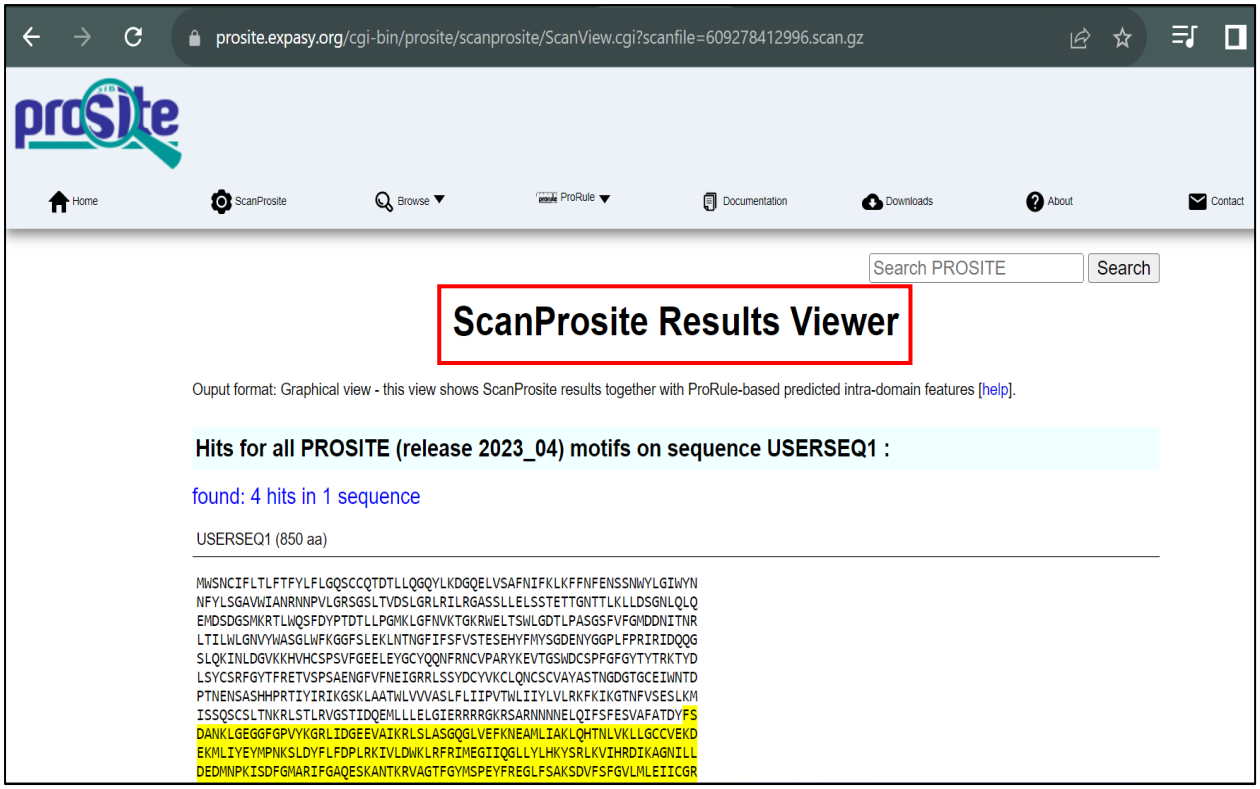


Figure 5: Result of 'Lectin' sequence after scanning in PROSITE database.

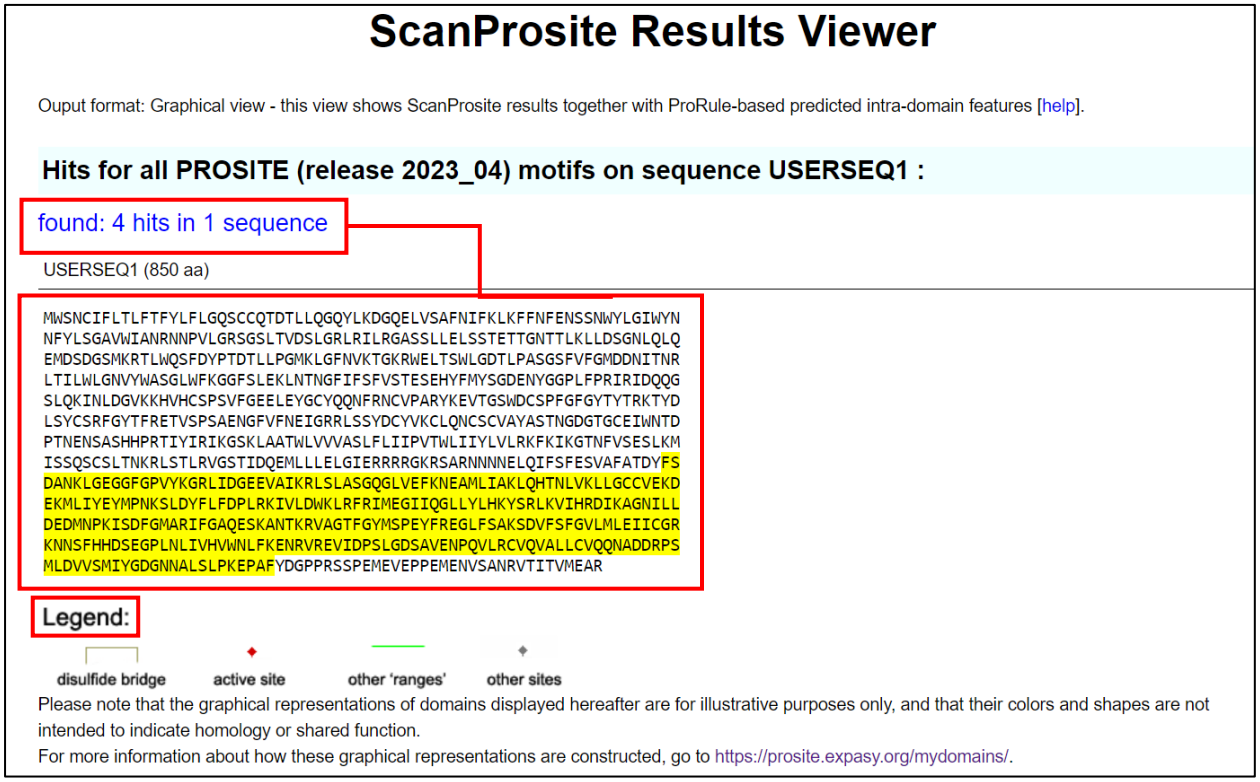


Figure 6: 4 Hits found for query sequence 'Lectin' in ScanProsite

PS50948 **PAN** *PAN/Apple domain profile* :

334 - 416: **score =** 8.226

```
CSRFgyTFRETVSPSAENgFVFNEIgRRLSSYDCYVKCLQNCSCVAYASTNGdGTGCEIW
NTDPTNENSASHHPRTIYIRIKG
```

Predicted features:

DOMAIN	334	416	Apple	[condition: not(<Any1={C}> or <Any78={C}>)]
DISULFID	367	390		[condition: C-x*-C]
DISULFID	371	377		[condition: C-x*-C]

Absent feature:

DISULFID	334	416		[condition not true: C-x*-C]

PS50011 **PROTEIN_KINASE_DOM** *Protein kinase domain profile* :

527 - 816: **score =** 37.149

```
FSDANKLGEGGFGPVYKGRLI-DGEEVAIKRLSL---ASGqGLVEFKNEAMLIAKLQHTN
LVKLLGCCVEKDEKMLIYEYMPNKSLDYFLFDplRKIVLDWKLRFRIMEGIIQGLLYLHK
YsrlKVIHRDIKAGNILLDEDMNPKISDFGMARIFGAQESkaNTKRVAGTFGYMSPEYFR
-EGLFSAKSDVFSFGVLMLEIICGRKnnSFHHDsEGPLNLIVHVWNLfkenrvreVIDPS
LgdsavenP----QVLRCVQVALLCVQQNADDRPSMLDVVSmiygdgnnalslpkEPAF
```

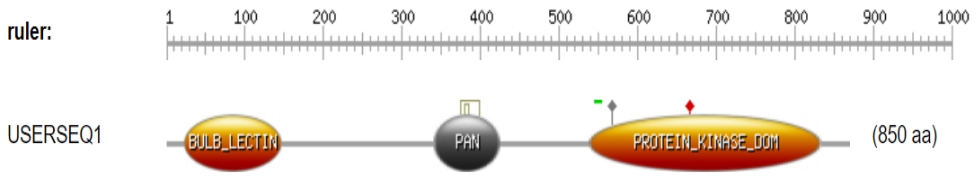
Predicted features:

DOMAIN	527	816	Protein kinase	[condition: none]
BINDING	533	541	/ligand="ATP" /ligand_id="ChEBI:CHEBI:30616	[condition: <30=K> or <Feature:PS00107>]
BINDING	555		/ligand="ATP" /ligand_id="ChEBI:CHEBI:30616	[condition: <30=K>]
ACT_SITE	652		Proton acceptor	[condition: D and <FTTag:ATP>]

Figure 7: PAN and PROTEIN KINASE Domain Profile of 'Lectin' sequence

hits by profiles: [3 hits (by 3 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



PS50927 **BULB_LECTIN** *Bulb-type lectin domain profile* :

24 - 144: **score =** 13.950

```
TDTLLQGQYLKDGQELVSAFNifKLKFFNFENSSNWYlgIWYNNFYlsGAVWIANRNNPV
lGRSGSLTVDSLGRLRIL-RGASSLLELSSTETTGNTTLKLLDSGNLQLQEMDSDGSMKR
TL
```

Predicted feature:

DOMAIN	24	144	Bulb-type lectin	[condition: none]

Absent feature:

DISULFID	56	88		[condition not true: C-x*-C]

Figure 8: Hit by Profile Diagrammatically Representation of 'Lectin' sequence

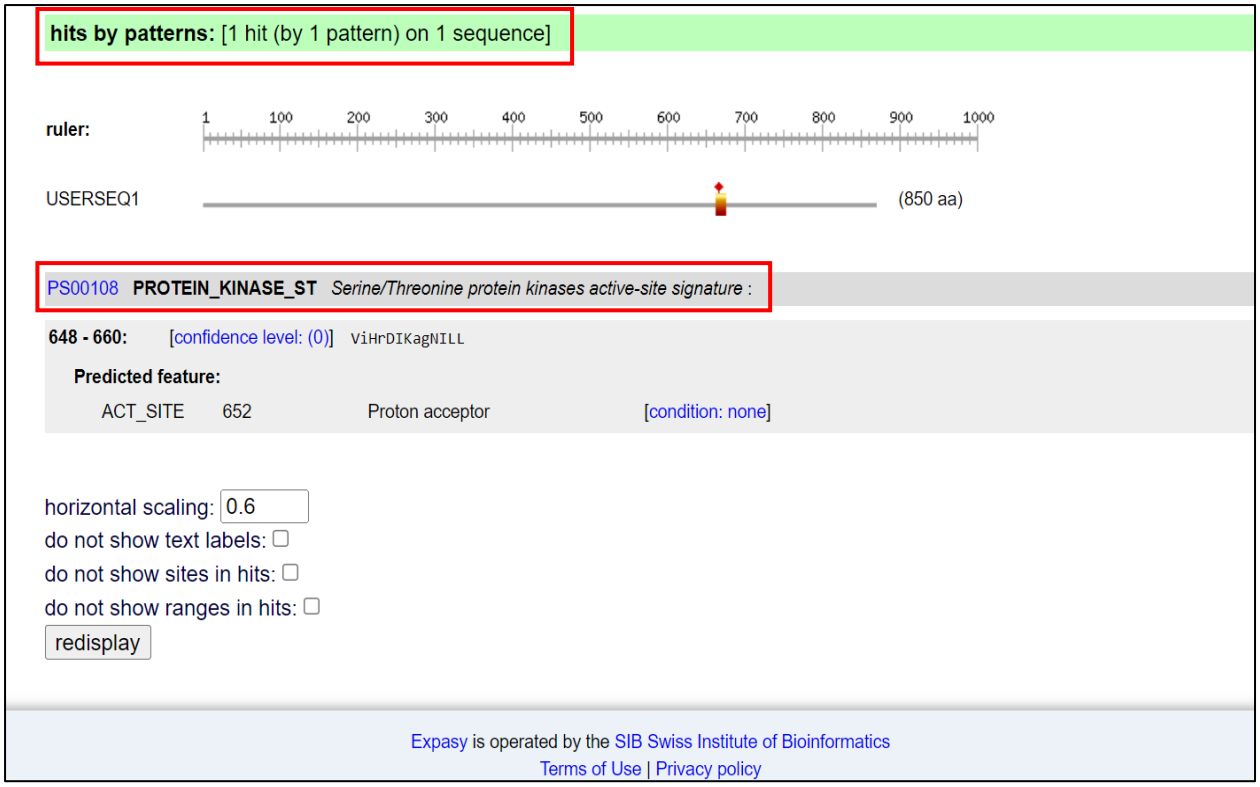


Figure 9: Hit by Pattern and Active Site Signature of the 'Lectin' sequence.

RESULTS:

Using the PROSITE database, Lectin (UniProt ID: Q9LW83) query was studied where FASTA sequence was considered for study. Overall, 4 hits were observed out of which 3 were hit by profile where the profile helps to characterize protein domains over their entire length, and they are more sensitive than patterns whereas 1 was hit by pattern with the unique signature for the functional sites & are biologically significant information. The result further helps to understand the query elaborately with respect to various section such as description, technical, references, copyright and miscellaneous, domain architecture. Thus, the database helps to understand the conserved regions from the query.

CONCLUSION:

The PROSITE database has played a crucial role in the field of bioinformatics and molecular biology for several decades. This comprehensive resource, maintained by the Swiss Institute of Bioinformatics, is a valuable tool for the identification and analysis of protein sequences and their functional domains. In conclusion, the PROSITE database offers several key advantages such as Domain and Motif Identification, Annotated and Curated Data, Compatibility, User-Friendly Interface, Support for Biomedical Research, Wide Range of Applications.

In summary, the PROSITE database remains a fundamental resource for bioinformatics, offering a wealth of information for researchers and scientists working in various life sciences disciplines. Its continued updates and the dedication of the Swiss Institute of Bioinformatics to maintain its quality ensure that it will remain an asset in the years to come.

REFERENCES:

1. Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2009). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Research, 38(suppl_1), D161–D166. https://doi.org/10.1093/nar/gkp885
2. Worst Foods High in Lectins. (2020, November 3). WebMD. https://www.webmd.com/diet/foods-high-in-lectins
3. Lectins. (2023, February 2). The Nutrition Source. https://www.hsph.harvard.edu/nutritionsource/anti-nutrients/lectins/
4. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D523–D531, https://doi.org/10.1093/nar/gkac1052
5. Christian J. A. Sigrist, Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, Ioannis Xenarios, New and continuing developments at PROSITE, *Nucleic Acids Research*, Volume 41, Issue D1, 1 January 2013, Pages D344–D347, https://doi.org/10.1093/nar/gks1067

DATE: 30/09/2023

WEBLEM 4(B)
INTERPRO DATABASE
(URL: https://www.ebi.ac.uk/interpro/)

AIM:

To explore the InterPro database related to the protein family Amylase from organism *Tetraodon nigroviridis* (UniProt ID: CAD20312.1)

INTRODUCTION:

InterPro database is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. To classify proteins in this way, InterPro database uses predictive models, known as signatures, provided by several collaborating databases (referred to as member databases) that collectively make up the InterPro consortium. A key value of InterPro database is that it combines protein signatures from these member databases into a single searchable resource, capitalizing on their individual strengths to produce a powerful integrated database and diagnostic tool. We add further value to InterPro database entries by providing detailed functional annotation as well as adding relevant GO terms that enable automatic annotation of millions of GO terms across the protein sequence databases. InterPro database integrates signatures from the following 13 member databases: CATH, CDD, HAMAP, MobiDB Lite, Panther, Pfam, PIRSF, PRINTS, Prosite, SFLD, SMART, SUPERFAMILY AND NCBIFAMs. The member databases use a variety of different methods to classify proteins. Each of the databases has a particular focus (e.g. protein domains defined from structure or full length protein families with shared function). We strive to integrate the signatures from the member databases into InterPro database entries and to identify where different member database entries are the same entity. InterPro database is updated approximately every 8 weeks. The release note pages contain information about what has changed in each release. All information in InterPro database is freely available. You can download InterPro data for local analyses from the Download page, or use the InterPro API. Find out more about the project by exploring the latest papers.

Amylase:

Amylase is an enzyme that occurs naturally in the saliva of some mammals and humans that aids in the process of digestion. It accelerates the breakdown or hydrolysis of starch into simple sugars. The pancreas and the salivary glands mainly synthesize amylase to hydrolyze dietary starch into disaccharides and disaccharide's that are converted into glucose and used as energy. Amylase was one of the first enzymes to be discovered in the 1800s. It was initially named distaste but later renamed amylase in the late 20th century.

Amylase, any member of class enzymes that catalyze the hydrolysis of starch into smaller carbohydrate molecules such as maltose (a molecule composed of two glucose molecules). Three categories of amylases, denoted alpha, beta, and gamma, differ in the way they attack the bonds of the starch molecules.

METHODOLOGY:

1. Go to the InterPro database website.
2. Search query Amylase using either of the search option like search by sequence, search by text, and search by domain architecture.
3. Hits are obtained, and filter options are enabled.
4. Click on entries for detailed information and external links for results.
5. Interpret the results.

OBSERVATIONS:

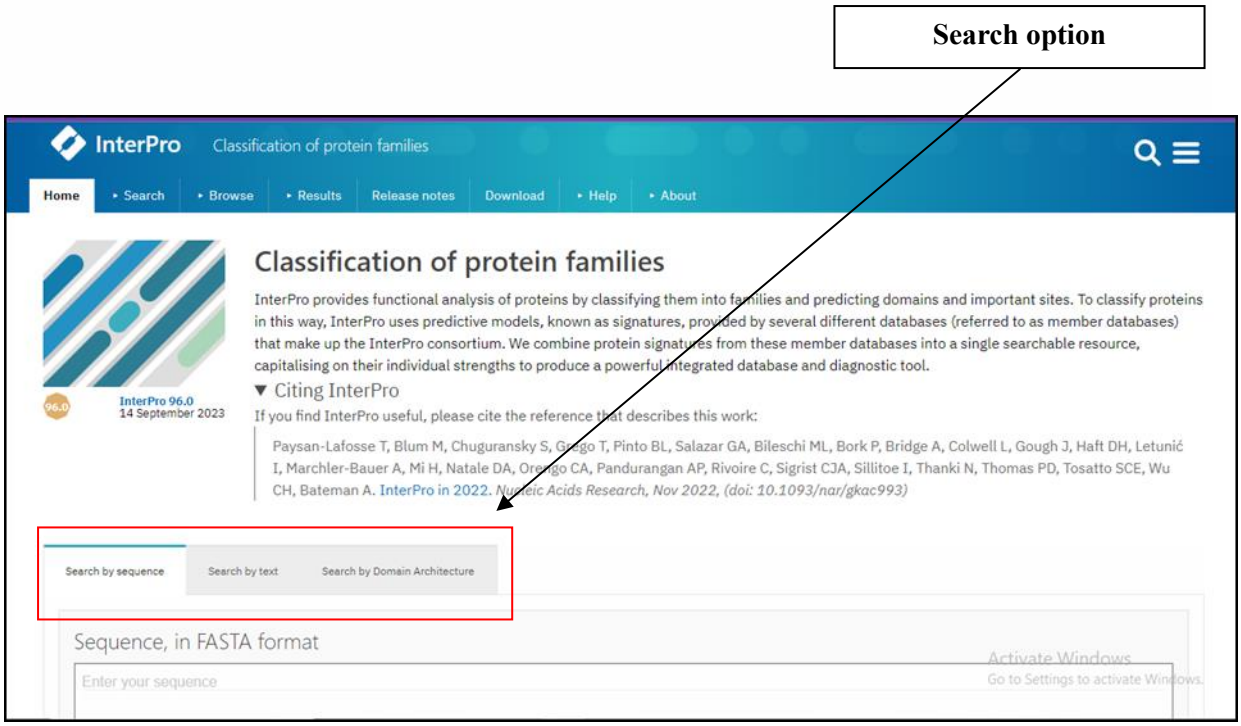


Figure 1: Homepage of InterPro Database

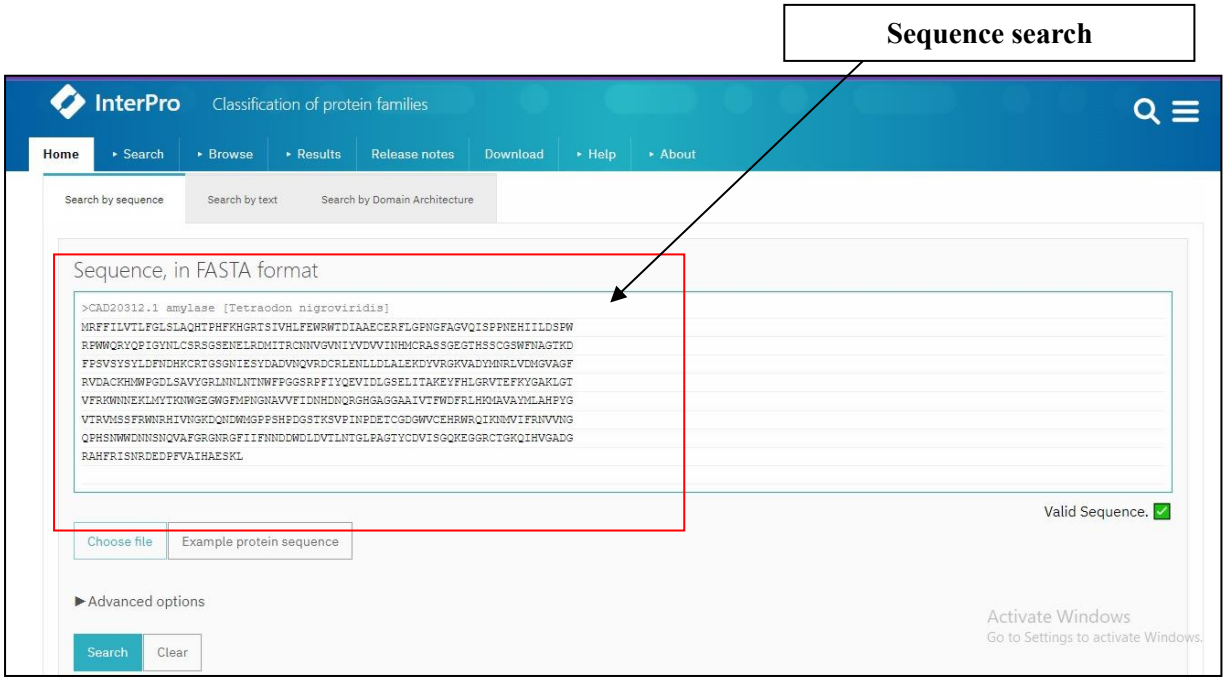


Figure 2: Query searched by sequence, in FASTA format

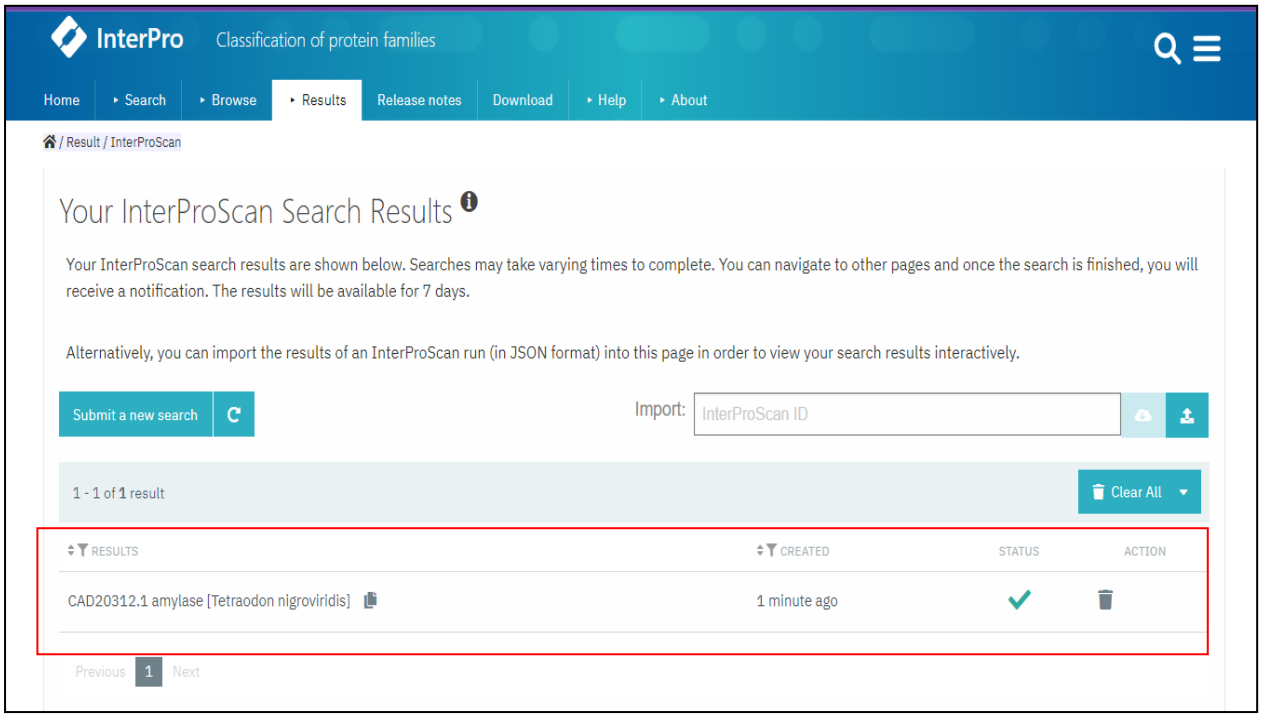


Figure 3: Results found by sequence search

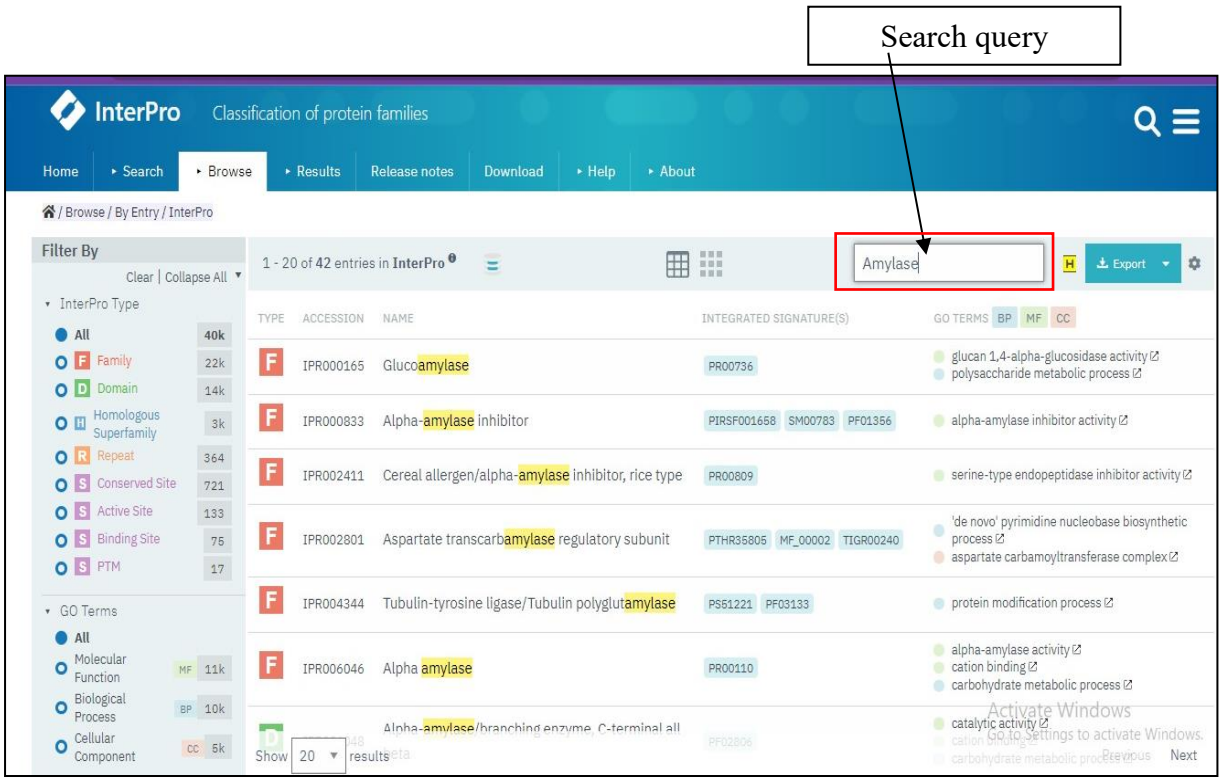


Figure 4: Search query in Browse option of InterPro database

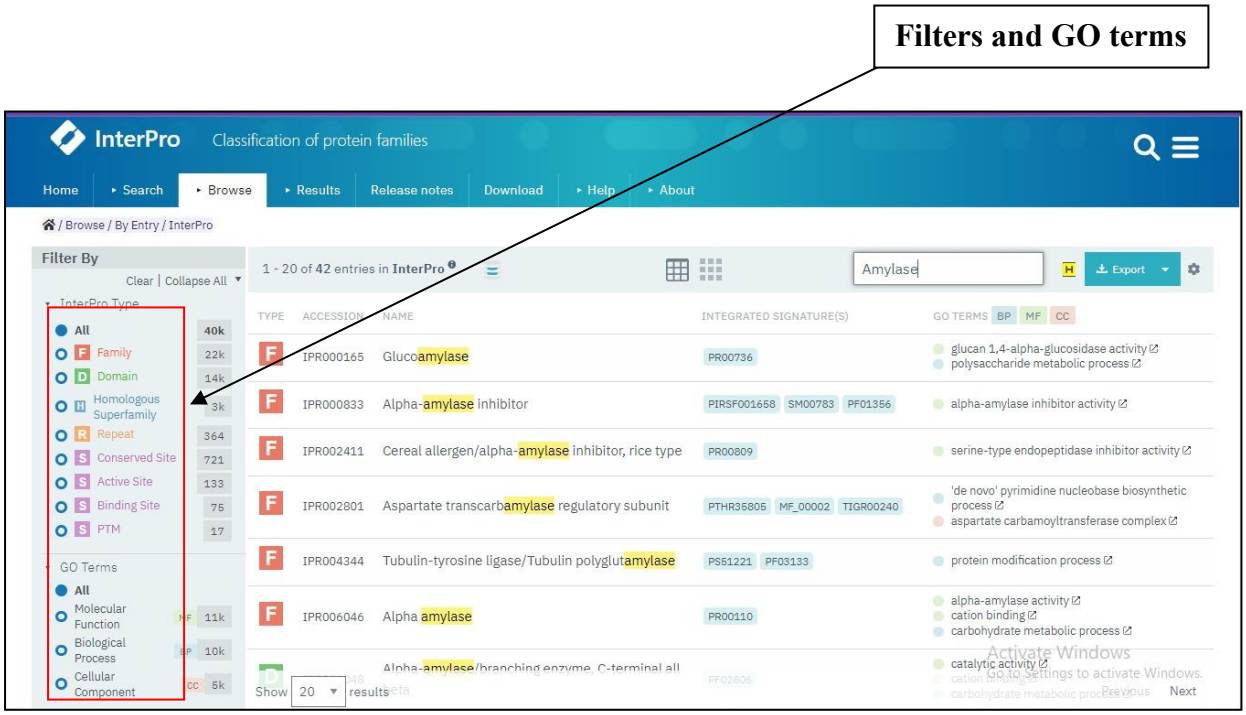


Figure 4a: Filters and GO (gene ontology) terms available in Browse search option

InterPro Classification of protein families

Home | Search | Browse | Results | Release notes | Download | Help | About

/ Browse / By Entry / Pfam / PF00128 / Overview

Pfam PF00128 Alpha amylase, catalytic domain

Pfam entry

Overview
Proteins 249k
Domain Architectures 2k
Taxonomy 41k
Proteomes 10k
Structures 580
Signature
AlphaFold 203k
Alignment
Curation

Member database	Pfam
Pfam type	domain
Short name	*Alpha-amylase*
Set	TIM_barrel

Add your annotation

Integrated to

IPR006047

Description

Alpha amylase is classified as family 13 of the glycosyl hydrolases. The structure is an 8 stranded alpha/beta barrel containing the active site, interrupted by a ~70 a.a. calcium-binding domain protruding between beta strand 3 and alpha helix 3, and a carboxyl-terminal Greek key beta-barrel domain.

References

1. Crystal structure of yellow meal worm alpha-amylase at 1.64 A resolution. Strobl S, Maskos K, Betz M, Wiegand

2. Refined molecular structure of pig pancreatic alpha-amylase at 2.1 A resolution. Larson SB, Greenwood A,

Activate Windows
Go to Settings to activate Windows.

Figure 4b: Entries pattern of InterPro database

Protein structure

InterPro Classification of protein families

Home | Search | Browse | Results | Release notes | Download | Help | About

Glycoside hydrolase family 13 Wikipedia

In molecular biology, **glycoside hydrolase family 13** is a family of glycoside hydrolases.

Alpha amylase, N-terminal ig-like domain

crystal structure of thermoactinomyces vulgaris r-47 alpha-amylase 1 (tvai) mutant d356n/e396q complexed with p2, a pullulan model oligosaccharide

Identifiers	
Symbol	Alpha-amylase_N
Pfam	PF02903
InterPro	IPR004185

Activate Windows
Go to Settings to activate Windows.

Figure 4c: Protein structure in InterPro database entries

RESULTS:

The query Amylase was studied using the InterPro database, where the data from various sources are integrated and classification as well as analysis of protein sequences and domains are performed. The query was fired using search by FASTA option and relevant information is studied.

CONCLUSION:

The InterPro database is a resource used in bioinformatics and genomics for the classification and analysis of protein sequences and domains. By combining information of different sources of data, InterPro database provides a comprehensive resource for researchers to classify, annotate, and analyze proteins. It helps in functional annotation, prediction of protein properties, and understanding the relationships between proteins based on shared domains and features.

REFERENCES:

1. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G., Bileschi, M., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D., Orengo, C., Pandurangan, A., Rivoire, C., . . . Bateman, A. (2022, November 9). InterPro in 2022. *Nucleic Acids Research*, 51(D1), D418–D427. https://doi.org/10.1093/nar/gkac993
2. InterPro. (n.d.). *About InterPro — InterPro Documentation*. https://interpro-documentation.readthedocs.io/en/latest/interpro.html

DATE: 30/10/2023

WEBLEM 5
INTRODUCTION TO STRUCTURE DATABASES

INTRODUCTION:

Structural bioinformatics, a branch of bioinformatics, is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA. The main objective of structural bioinformatics is to create new methods for analyzing and manipulating biological macromolecular data to solve problems in biology and generate new insights.

Structural databases in bioinformatics are crucial resources that are modelled around experimentally determined protein structures, providing the biological community with access to valuable experimental data in a useful way. These databases aim to organize and annotate protein structures, and they often include three-dimensional coordinates, experimental information (such as unit cell dimensions and angles for x-ray crystallography determined structures), and sequence information. The primary attribute of a structure database is structural information, whereas sequence databases focus on sequence information and contain no structural information for the majority of entries. Protein structure databases are critical for many efforts in computational biology, such as structure-based drug design, and they are used to provide insights about the function of proteins.

Prominent examples of structural databases include the Protein Data Bank (PDB), which contains experimentally determined three-dimensional structures of biomolecules, the Nucleic Acid Data Base (NDB), which contains experimentally determined information about nucleic acids, the carbohydrate structure databases (CSDB) ,which providing a curated repository of structural, taxonomical, bibliographic, and NMR-spectroscopic data on natural carbohydrates and carbohydrate-related molecules from bacterial, fungal, and plant origins, the Reactome databases which provides information about metabolic pathways, the PDBSum databases provides a pictorial summary and detailed analyses of 3D macromolecular structures deposited in the Protein Data Bank, the PDBTM databases provides information about transmembrane proteins from the PDB, the CATH classifies protein domains based on their architecture, topology, and homology and the Structural Classification of Proteins (SCOP), which provides a comprehensive description of the structural and evolutionary relationships between structurally known proteins. These examples are introduced in detail below.

1. **Protein Data Bank (PDB) Database:**

   Protein data bank is an online structural library of biological macromolecules, which is the only worldwide repository of macromolecular structure. The PDB was organized in 1971 at Brookhaven National Laboratories (BNL) as a platform of crystal structures of biomolecules. Over the years, the data submitted to the PDB was modified and approaches to access the PDB have changed, as a result of advancements in technology.

In October 1998, Research Collaborator for Structural Bioinformatics (RCSB) has started to manage and maintain the activities of PDB. The major task of the RCSB is to generate such measures that allow the use and analysis of structural data. PDB stores 3D structural information of biological molecules mainly nucleic acid and proteins. The structural information of biomolecules is commonly acquired experimentally by NMR spectroscopy, X-ray crystallography, electron microscopy etc. Structural information of some chemical ligands and nucleotides are also available on PDB. PDB ID is a four- character identifier that is actually entitled as PDB entry. A Searching through PDB is done by a vast range of search engines ranges from PDB ID and keywords to structural features of proteins and other biomolecules.

There are two formats that PDB uses to keep structural data: The PDB file format and macromolecular crystallographic information file format (mmCIF). PDB file design is more commonly used in protein community ascompared to mmCIF. PDB offers various molecular structural visualization soft wares including RasMol, Jmol, PDB simple viewer, PDB protein workshop and RCSB-Kiosk. Structural confirmation of secondary structure is also provided by PDB. The PDB depository is run by an association, named the Worldwide Protein Data Bank (wwPDB) which guarantees that the information is freely accessible to the public. Structures for huge numbers of the proteins and nucleic acids required in the central procedures of life are available on PDB.

PDB file format:

1. **ATOM:** atomic coordinate record containing the X,Y,Z orthogonal Å coordinates for atoms in standard residues (amino acids and nucleic acids).
2. **HETATM:** atomic coordinate record containing the X, Y, Z orthogonal Å coordinates for atoms in non-standard residues (ligands, cofactors, etc.).
3. **TER:** record indicating the end of a chain of residues.
4. **HEADER:** record containing general details about the molecules in the file, as well as the experiment(s) used to elucidate their structures.
5. **COMPND:** record containing information about the compound, including its name, synonyms, and other identifiers.

Bank (PDB) file format is a standard for files containing atomic coordinates of biological macromolecules. The PDB file format consists of lines of information in a text file, with each line of information in the file called a Record. A PDB file generally contains several different types of records, arranged in a specific order to describe a structure.

The most common record types include:

1. **ATOM:** atomic coordinate record containing the X, Y, Z orthogonal Å coordinates for atoms in standard
2. **REMARK:** record containing additional information about the structure, such as refinement details, experimental conditions, and other annotations.

The formats of these record types are given in the PDB file specification. The PDB file format is limited to 80 columns per line, with each line terminated by an end-of-line indicator. The columns in the PDB file format for the ATOM record type include the atom

serial number, atom name, residue name, chain identifier, residue sequence number, and atomic coordinates. The HETATM record type is similar to the ATOM record type, but is used for non-standard residues. The TER record type indicates the end of a chain of residues. The HEADER, COMPND, and REMARK record types contain general information about the structure, such as the name of the molecule, the authors of the structure, and the method of structure determination.

2. **Nucleic Acid Knowledgebase (NAKB) Database:**

The Nucleic Acid Database (NDB) played a pivotal role as the first comprehensive resource for three-dimensional (3D) structures of nucleic acids. Established in the 1990s at Rutgers University, NDB facilitated collaborative studies through a SQL-relational database, offering curated information from X-ray and nuclear magnetic resonance (NMR) experiments. Over its three-decade tenure, NDB evolved to become a valuable repository, collecting data from the Protein Data Bank (PDB) and the Cambridge Structural Database (CSD).

In response to the growing landscape of nucleic acid structures and emerging technologies like cryoelectron microscopy (EM), the Nucleic Acid Knowledgebase (NAKB) emerged as the modern successor to NDB. Initiated in 2019 and officially launched in May 2023, NAKB aimed to preserve and enhance NDB's functionality while incorporating structures from diverse methods, providing comprehensive functional and structural annotations, and establishing links to broader nucleic acid-focused resources.

NAKB provides search, report, statistics, atlas and visualization pages for all nucleic-acid containing experimentally determined 3D structures held by NDB and by the Protein Data Bank (PDB), including all major methods: X-ray, NMR, and Electron Microscopy. For each structure, links are provided to external resources that annotate and analyze nucleic acid structures and their complexes.

The NAKB website (nakb.org), introduced in July 2022, offers efficient search tools, tabular reports, 2D and 3D structure visualizations, educational content, standards information, and a curated nucleic acid community web and software resource list. With a user-friendly interface and modern web architecture, NAKB ensures an enhanced experience for users, supporting accessibility on both large and small devices. The website undergoes weekly updates, maintaining its commitment to providing timely and relevant nucleic acid structural information. Notably, NDB was officially retired in July 2023, marking the seamless transition to the advanced capabilities of NAKB in serving the scientific community.

NOTE: NAKB replaces the Nucleic Acid Database (NDB) resource that will be retired in July 2023.

3. **Carbohydrate Structure Database (CSDB)/ CCSD /Gly-Tou-Can Database:**

The Carbohydrate Structure Database (CSDB) is a free curated database and service platform in glycoinformatics, launched in 2005 by a group of Russian scientists from N.D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences. The database aims to provide structural, bibliographic, taxonomic, NMR spectroscopic, and other information

on glycan and glycoconjugate structures of prokaryotic, plant, and fungal origin. It serves as a platform for multiple glycoinformatic studies and web tools.

CSDB covers nearly all structures published up to the previous year in the scope of bacterial carbohydrates. Prokaryotic, plant, and fungal mean that a glycan was found in the organisms belonging to these taxonomic domains or was obtained by modification of those found in these organisms. Carbohydrate means a structure composed of any residues linked by glycosidic, ester, amidic, ketal, phospho- or sulpho-diester bonds in which at least one residue is a sugar or its derivative, except DNA/RNA.

The main source of data is retrospective literature analysis. About 20% of data were imported from CCSD (Carbbank, University of Georgia, Athens; structures published before 1996) with subsequent manual curation and approval. CSDB contains manually curated natural carbohydrate structures, taxonomy, bibliography, NMR, and other data from literature. Coverage is close to complete up to the year 2020 for bacterial and fungal carbohydrates. Users can search the database by IDs, bibliographic data and keywords, biological source, structural fragments, and NMR data. The substructure search supports graphic input, structure wizard, selection from the library, and query language (expert form).

4. **REACTOME Database:**

Reactome stands as a cornerstone in the landscape of pathway databases, offering an open-source, open-access, and meticulously curated resource dedicated to human pathways and biological processes. Developed through the collaborative efforts of expert biologists and Reactome editorial staff, pathway annotations within this database undergo a rigorous peer-review process. Notably, Reactome's annotations are intricately cross-referenced with various authoritative sources, including protein and gene information from UniProt, NCBI EntrezGene, Ensembl, UCSC, and HapMap, as well as small molecule data from KEGG Compound and ChEBI. Primary research literature from PubMed and GO controlled vocabularies further enriches the annotations, ensuring a comprehensive and well-rounded knowledgebase.

The unique data model employed by Reactome broadens the traditional concept of a reaction, encompassing diverse biological events such as entity transformations, compartmental transport, interactions leading to complex formation, and classical biochemical reactions. This inclusive approach allows Reactome to capture a wide spectrum of biological processes spanning signaling, metabolism, transcriptional regulation, apoptosis, and synaptic transmission. The resulting dataset is presented in a single, internally consistent, and computationally navigable format, making Reactome an indispensable resource for basic research, genome analysis, pathway modeling, systems biology, and education.

In response to the rapid growth of knowledge in the field, Reactome has not only doubled in size over the past two years but has also introduced new tools for data aggregation and analysis. To support this continuous evolution, Reactome has undergone a redesign, encompassing both its web interface and data analysis software. This redesign reflects Reactome's commitment to staying at the forefront of pathway databases, providing an up-to-date and user-friendly platform for researchers.

5. **PDBSum Databases:**

In the early years of the Protein Data Bank (PDB), researchers faced challenges navigating experimentally determined protein structures due to text file storage, lack of a user-friendly interface, and laborious methods for identifying entries of interest. The growing repository necessitated innovative solutions to efficiently access and analyze structural information.

In response to these challenges, the advent of the World Wide Web (WWW) in the early 1990s ushered in a transformative era for protein structure analysis. Among the pioneering platforms that leveraged the emerging web technology was PDBsum, developed at University College London (UCL) in 1995. Designed to harness the capabilities of the WWW, PDBsum sought to streamline the exploration of structural information in the PDB by creating a visually-oriented catalog. This compendium aimed to provide a rich array of pictorial representations, including unique structural analyses not readily available elsewhere. Alongside PDBsum, other early servers such as PDBBrowse, the Swiss-3Dimage collection, and the IMB Jena Image Library of Biological Macromolecules emerged, each contributing distinct approaches to presenting and visualizing protein structures.

PDBsum's development persisted at UCL until its transfer to the European Bioinformatics Institute (EBI) in 2001, marking a pivotal moment in its evolution. Subsequent enhancements and additions have further refined the database, while concurrent advancements in other servers, particularly those operated by members of the worldwide Protein Data Bank (wwPDB) consortium, have collectively propelled the field of protein structure analysis into a new era of accessibility and functionality. This narrative encapsulates the dynamic evolution of databases like PDBsum, which, through strategic adaptation to technological advancements, continue to play pivotal roles in facilitating the exploration and understanding of protein structures on a global scale.

6. **PDBTM Databases:**

The Protein Data Bank (PDB) is a critical repository of biological macromolecular structures, yet the representation of transmembrane proteins within this vast resource is notably scarce, constituting less than 2% of entries, as highlighted by the PDBTM database. Established in 2004, the PDBTM database emerged to address the challenges associated with identifying and characterizing transmembrane protein structures within the PDB.

Transmembrane proteins, pivotal for cellular functions such as energy production, regulation, and metabolism, are also frequent targets for drug development, with approximately half of contemporary drugs impacting these proteins. Recognizing the importance of these proteins, the PDBTM database pioneered a methodology reliant solely on 3D coordinates to identify transmembrane segments, circumventing the limitations of existing annotations in PDB entries.

Given the experimental intricacies in determining the orientation of transmembrane proteins relative to the lipid bilayer, the PDBTM database introduced the TMDET method to tackle this challenge. In the absence of solved atomic structures for the double lipid layer, theoretical methods, such as those employed by the PDBTM database, become indispensable for determining protein orientations.

Several other databases, each utilizing diverse theoretical algorithms, contribute to the understanding of transmembrane proteins. The OPM database offers a well-structured classification, emphasizing the protein-membrane relationship. The CGDB database employs sophisticated physics-based models derived from coarse-grained simulations, while Mpstruct stands out as a reliable resource for regularly updated membrane protein classifications.

In the landscape of transmembrane protein databases, PDBTM plays a distinctive role by systematically collecting and verifying the structures of transmembrane proteins from the PDB. This meticulous curation includes the correction of biologically active oligomer forms, definition of membrane orientation, and identification of transmembrane segments, re-entrant loops, and interfacial helices. Through these efforts, PDBTM significantly contributes to unraveling the complexities of transmembrane protein structures and their roles in cellular processes.

7. **Class, Architecture, Topology, And Homologous Superfamily (CATH) Databases:**

CLASS, ARCHITECTURE, TOPOLOGY, AND HOMOLOGOUS SUPERFAMILY (CATH) CATH, a database for hierarchical classification of protein domains was developed at University of London.The CATH database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains. It was created in the mid-1990s by Professor Christine Orengo and colleagues, and continues to be developed by the Orengo group at University College London.

At its core, CATH utilizes experimentally-determined protein three-dimensional structures sourced from the Protein Data Bank (PDB). These structures are meticulously dissected into their constituent polypeptide chains, and the identification of protein domains within these chains is a nuanced process involving a combination of automated methodologies and manual curation. The ensuing classification within the CATH structural hierarchy follows a multi-tiered approach.

The Class (C) level classification categorizes domains based on their secondary structure content, distinguishing between all-alpha, all-beta, a combination of alpha and beta, or domains with minimal secondary structure. Moving up the hierarchy, the Architecture (A) level considers the spatial arrangement of secondary structures in three-dimensional space. The Topology/fold (T) level focuses on the connectivity and arrangement of secondary structure elements. Finally, domains are assigned to the Homologous Superfamily (H) level when there is compelling evidence of evolutionary relatedness, indicating homology.

To supplement experimentally determined structures, CATH incorporates additional sequence data from Gene3D, a related resource. Gene3D provides information on domains lacking experimentally determined structures, aiding in the population of homologous super families, UniProtKB and Ensembl contribute to this process by having their protein sequences scanned against CATH Hidden Markov Models (HMMs), facilitating the prediction of domain sequence boundaries and the assignment to homologous super families.

This intricate classification process, combining automated tools and manual curation, results in a wealth of information that is freely accessible to the scientific community and beyond. Furthermore, the CATH database remains dynamic, receiving periodic updates to

ensure that the latest advancements in protein domain classification are reflected, demonstrating its commitment to serving as a valuable resource for researchers and bioinformaticians alike.

8. **SCOPe Databases (Structural Classification of Proteins – Extended):**

The Structural Classification of Proteins (SCOPe) database, established 27 years ago as the successor to the classic SCOP, continues to be a cornerstone in the field of protein structure and evolution. Designed as a manually curated hierarchy of domains from known protein structures, SCOPe's primary objective is to unravel the structural and evolutionary relationships among proteins.

SCOPe maintains a dynamic knowledgebase that evolves with the influx of new protein structures from the Protein Data Bank (PDB). Its hierarchical organization encompasses Families, Superfamilies, Folds, and Classes, providing a comprehensive framework for understanding the relationships between related proteins at various structural and functional levels. Expert curation, particularly at the Superfamily level, integrates diverse information to discern common ancestry.

The database excels in uncovering ancient homologous relationships, utilizing structural evidence when sequence similarity is absent. SCOPe annotates these relationships, grouping homologous domains into Superfamilies or, when evidence is inconclusive, categorizing them under common Folds.

Beyond classification, SCOPe offers valuable resources for computational analyses. It provides sequences and PDB-style coordinate files for all domains, ensuring accessibility for researchers. Post-translationally modified amino acids are meticulously translated, and sequences are curated to eliminate errors.

In alignment with FAIR principles (Findable, Accessible, Interoperable, Reusable), SCOPe ensures data availability through versioned releases, enabling findability and traceability over time. Major stable releases, accompanied by periodic updates, reflect the commitment to maintaining a stable and accurate database. The monthly updates, synchronized with the PDB, reflect the dedication to staying current in the rapidly evolving field.

Since 2001, SCOPe has adhered to stable identifiers, ensuring consistency across releases. The database is designed for both machines and humans, supporting download in various formats and archived on Zenodo, an open-access data repository. The current SCOPe release, 2.08, stands as a testament to its growth, classifying 344,851 domains from 106,976 PDB entries. With each release, SCOPe continues to be a vital resource for researchers exploring the intricate world of protein structure and evolution.

REFERENCES:

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic acids research*, 28(1), 235–242. https://doi.org/10.1093/nar/28.1.235
2. About RCSB PDB: *Enabling Breakthroughs in Scientific and Biomedical Research and Education. RCSB PDB*;[cited 2018 March 19]. Available from: http://www.rcsb.org/pages/about-us/index.
3. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, *The Protein Data Bank (2000) Nucleic Acids Research* 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
4. Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A. I., Sweeney, B., Zirbel, C. L., Leontis, N. B., & Berman, H. M. (2014). *The Nucleic Acid Database: new features and capabilities*. *Nucleic acids research*, 42(Database issue), D114–D122. https://doi.org/10.1093/nar/gkt980.
5. *Introduction to Protein Data Bank Format.* (n.d.). https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html.
6. Catherine L Lawson, Helen M Berman, Li Chen, Brinda Vallat, Craig L Zirbel, *The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids, Nucleic Acids Research*, 2023;, gkad957, https://doi.org/10.1093/nar/gkad957.
7. S.I. Shcherbinina, Ph.V. Toukach "Three-dimensional structures of carbohydrates and where to find them", Int J Mol Sci, 2020, 21(20): ID 7702. (PMID 33081008, DOI 10.3390/ijms21207702).
8. David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Christina Yung, Ewan Birney, Peter D'Eustachio, Lincoln Stein, *Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Research*, Volume 39, Issue suppl_1, 1 January 2011, Pages D691–D697, https://doi.org/10.1093/nar/gkq1018.
9. Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S., & Thornton, J. M. (2018). *PDBsum: Structural summaries of PDB entries. Protein science: a publication of the Protein Society*, 27(1), 129–134. https://doi.org/10.1002/pro.3289.
10. Kozma, D., Simon, I., & Tusnády, G. (2012). *PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Research*, 41(D1), D524–D529. https://doi.org/10.1093/nar/gks1169.
11. Knudsen, M., & Wiuf, C. (2010). *The CATH database. Human genomics*, 4(3), 207–212. https://doi.org/10.1186/1479-7364-4-3-207.
12. Chandonia, J., Guan, L., Lin, S., Yu, C., Fox, N., & Brenner, S. E. (2021). *SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. Nucleic Acids Research*, 50(D1), D553–D559. https://doi.org/10.1093/nar/gkab1054.

DATE: 30/10/2023

WEBLEM 5 (A)
PROTEIN DATA BANK (PDB) DATABASE
(URL: https://www.rcsb.org/pdb/)

AIM:

To study and explore the protein structure for the query "Lysine" (PDB ID: 1OZV) using the Protein Data Bank (PDB) Database.

INTRODUCTION:

The Protein Data Bank (PDB) is a database that contains three-dimensional structural data of biological macromolecules, such as proteins and nucleic acids. The PDB was established in 1971 and is managed by the Worldwide Protein Data Bank (wwPDB), an international consortium that collaboratively oversees deposition, validation, Bio-curation, and open access dissemination of 3D macromolecular structure data. The PDB is a key resource in areas of structural biology, such as structural genomics, and is used by structural biologists to study the 3D structure of biological macromolecules. The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. The primary information stored in the PDB archive consists of coordinate files for biological molecules, which list the atoms in each protein and their 3D location in space. Features of the PDB include its historical significance as the first open-access digital resource in biology for sharing three-dimensional protein structures, its role as a critical resource for computational biology, such as structure-based drug design, and its constant growth as a reflection of the research happening in laboratories across the world. The PDB file format is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank. The PDB format provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity. The PDB format is the legacy file format for the Protein Data Bank which now keeps data on biological macromolecules in the newer mmCIF file format.

Lysine:

Lysine, denoted as Lys or K, is a critical α-amino acid serving as a precursor to numerous proteins, essential for fundamental biological processes. Its significance lies in being an essential amino acid for humans, necessitating its intake through the diet due to the body's inability to synthesize it. Lysine contributes to proteinogenesis, collagen cross-linking, nutrient uptake, and the production of carnitine, crucial in fatty acid metabolism. Additionally, lysine plays a role in histone modification, impacting the epigenome.

The PDB is crucial for understanding lysine's structural aspects, exemplified by the crystal structure (PDB ID 1OZV) showcasing LSMT's SET domain bound to lysine and AdoHcy. LSMT, a lysine-specific methyltransferase, catalyzes lysine methylation in RuBisCO, contributing insights into multiple lysine methylation mechanisms. These structures unravel lysine-specific methyltransferase intricacies, influencing gene expression, chromatin structure, and cellular processes. Studying lysine in the PDB is essential for advancing structural comprehension and functional roles, providing avenues for further research.

METHODOLOGY:

1. Open the homepage of the Protein Data Bank (PDB) Database.
2. Enter the query 'Lysine' and initiate the search.
3. After the retrieval of the query, observe the results. Apply specific refinements (filters) to narrow down the results based on the query.
4. Select a particular entry of interest ['1OZV: Crystal structure of the SET domain of LSMT bound to Lysine and AdoHcy'] for further study in terms of its Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome, Ligands, and Versions.
5. To display and download the 3D structure of the protein, click on the 'Display and Download' option and select the desired format.

OBSERVATIONS:

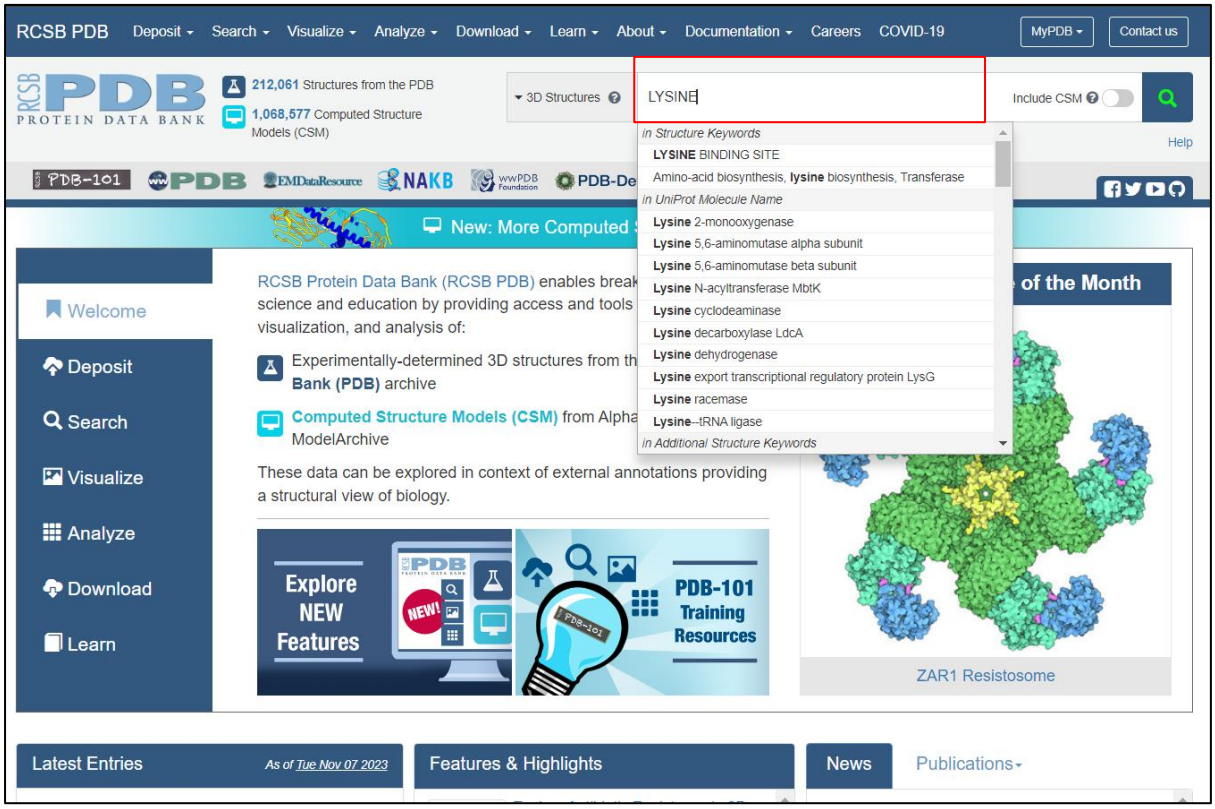


Figure 1: Homepage of the Protein Data Bank (PDB) Database

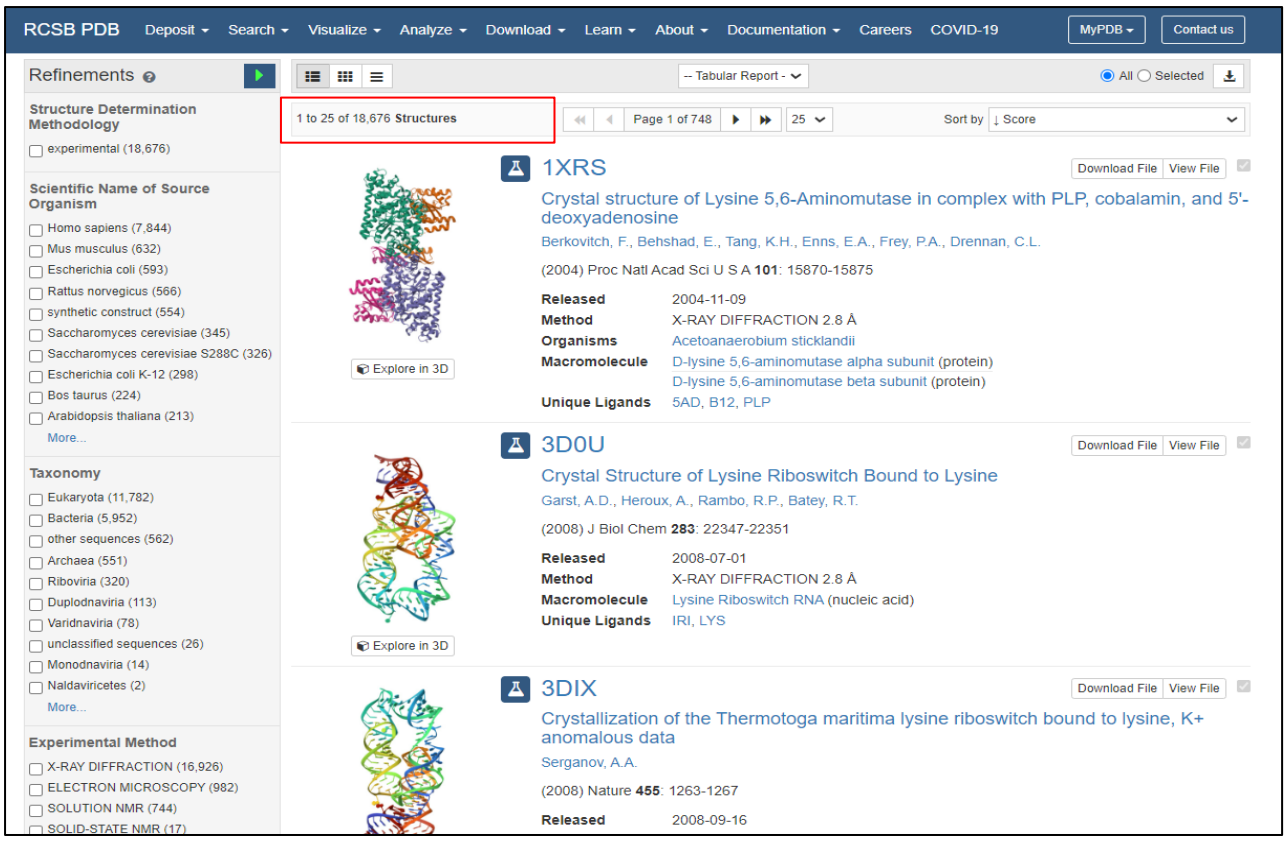


Figure 2: Number of hits obtained for Basic Search for the query

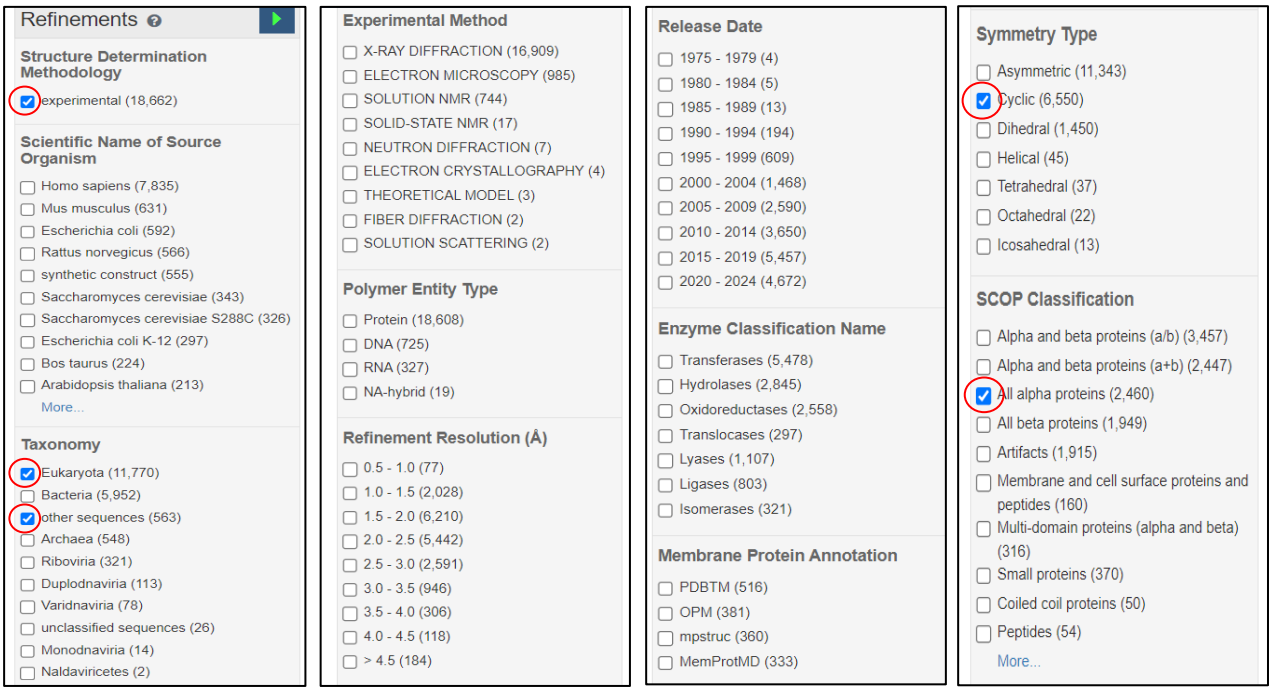


Figure 3: List of Refinements (Filters) applied

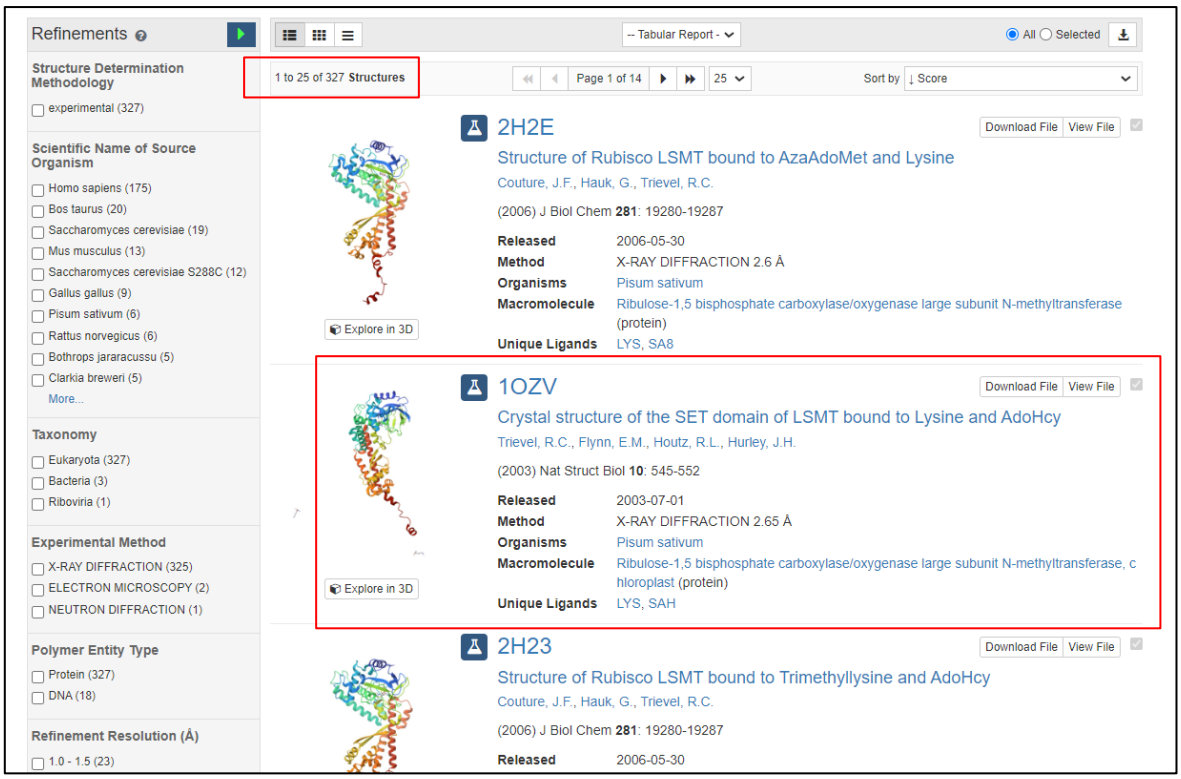


Figure 4: Results obtained after applying refinements (filters) and select the query

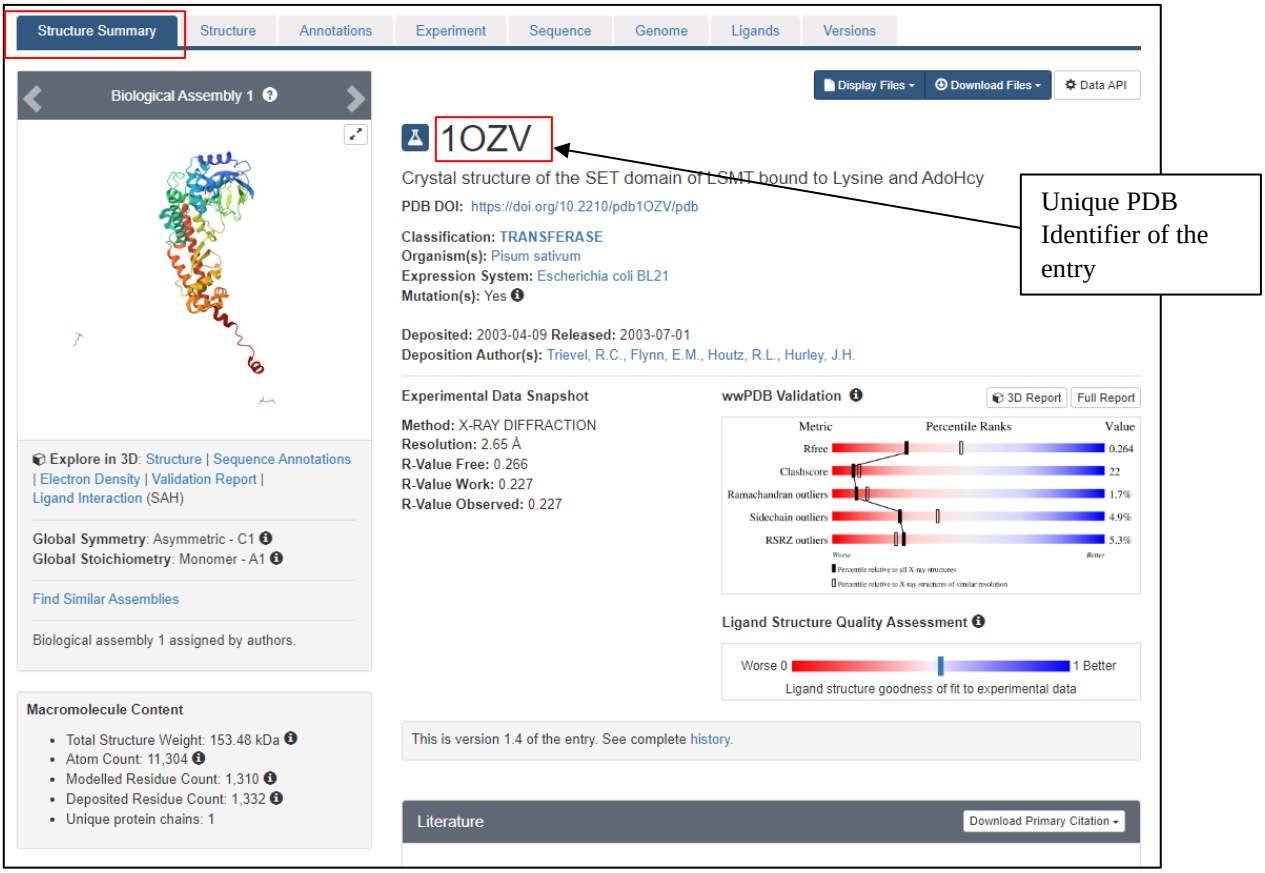


Figure 5: Entry opened that displays the Structure Summary

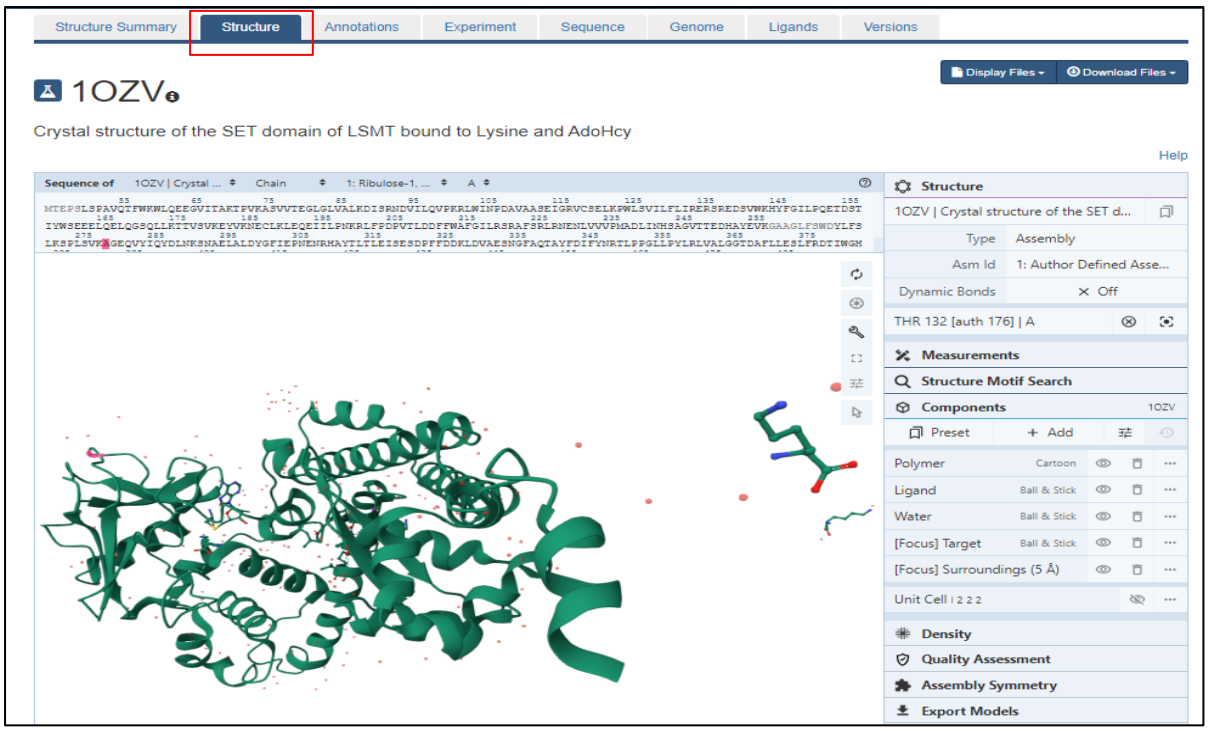
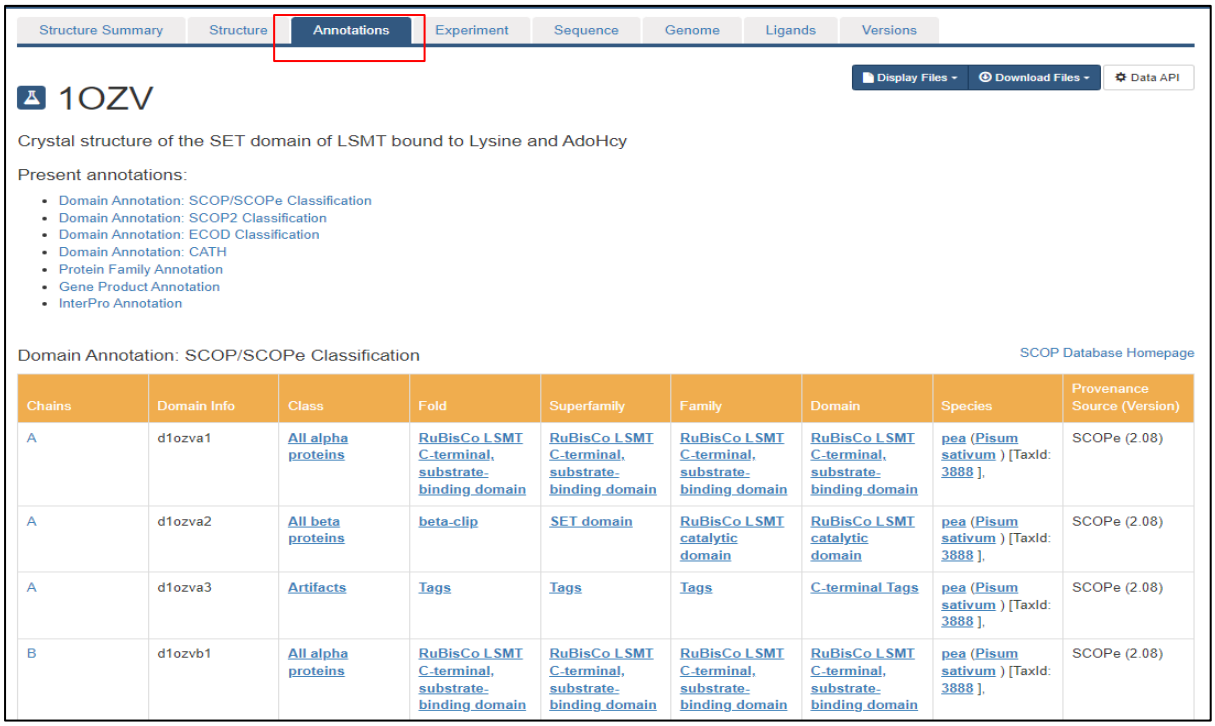


Figure 6: 3D View of the structure



Chains	Domain Info	Class	Fold	Superfamily	Family	Domain	Species	Provenance Source (Version)
A	d1ozva1	All alpha proteins	RuBisCo LSMT C-terminal, substrate-binding domain	RuBisCo LSMT C-terminal, substrate-binding domain	RuBisCo LSMT C-terminal, substrate-binding domain	RuBisCo LSMT C-terminal, substrate-binding domain	pea (Pisum sativum) [TaxId: 3888],	SCOPe (2.08)
A	d1ozva2	All beta proteins	beta-clip	SET domain	RuBisCo LSMT catalytic domain	RuBisCo LSMT catalytic domain	pea (Pisum sativum) [TaxId: 3888],	SCOPe (2.08)
A	d1ozva3	Artifacts	Tags	Tags	Tags	C-terminal Tags	pea (Pisum sativum) [TaxId: 3888],	SCOPe (2.08)
B	d1ozvb1	All alpha proteins	RuBisCo LSMT C-terminal, substrate-binding domain	RuBisCo LSMT C-terminal, substrate-binding domain	RuBisCo LSMT C-terminal, substrate-binding domain	RuBisCo LSMT C-terminal, substrate-binding domain	pea (Pisum sativum) [TaxId: 3888],	SCOPe (2.08)

Figure 7: View of the Annotations Section

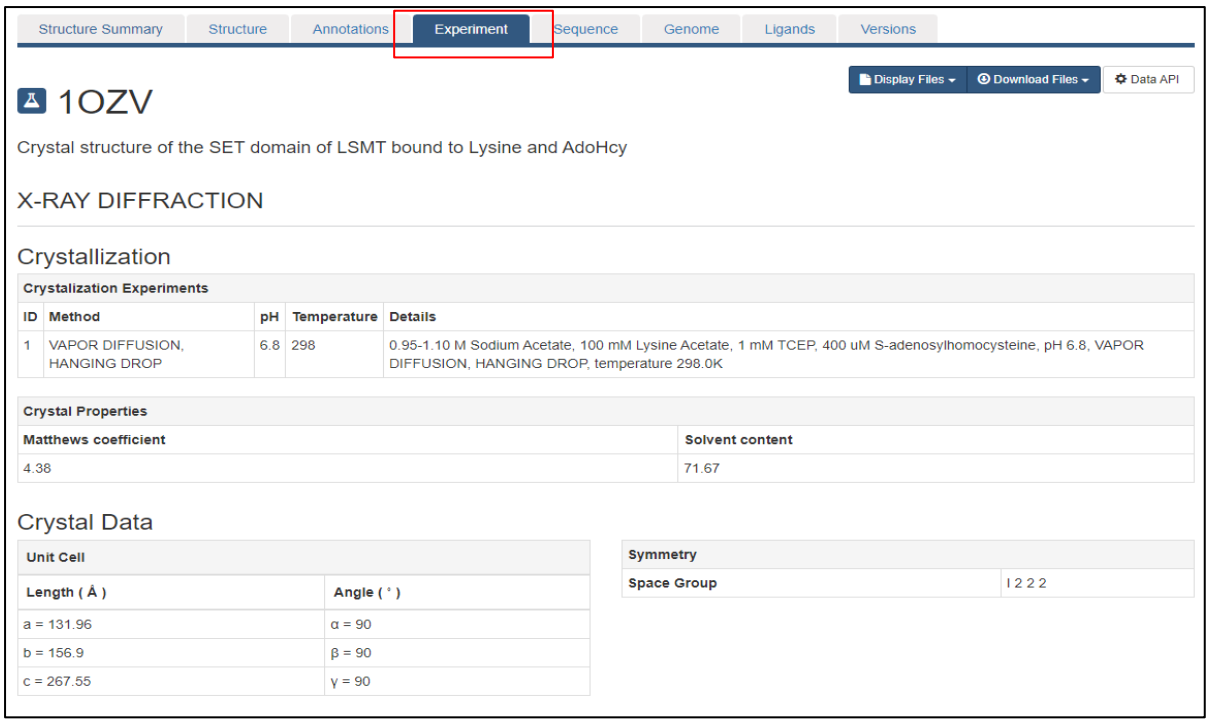


Figure 8: View of the Experiment Section

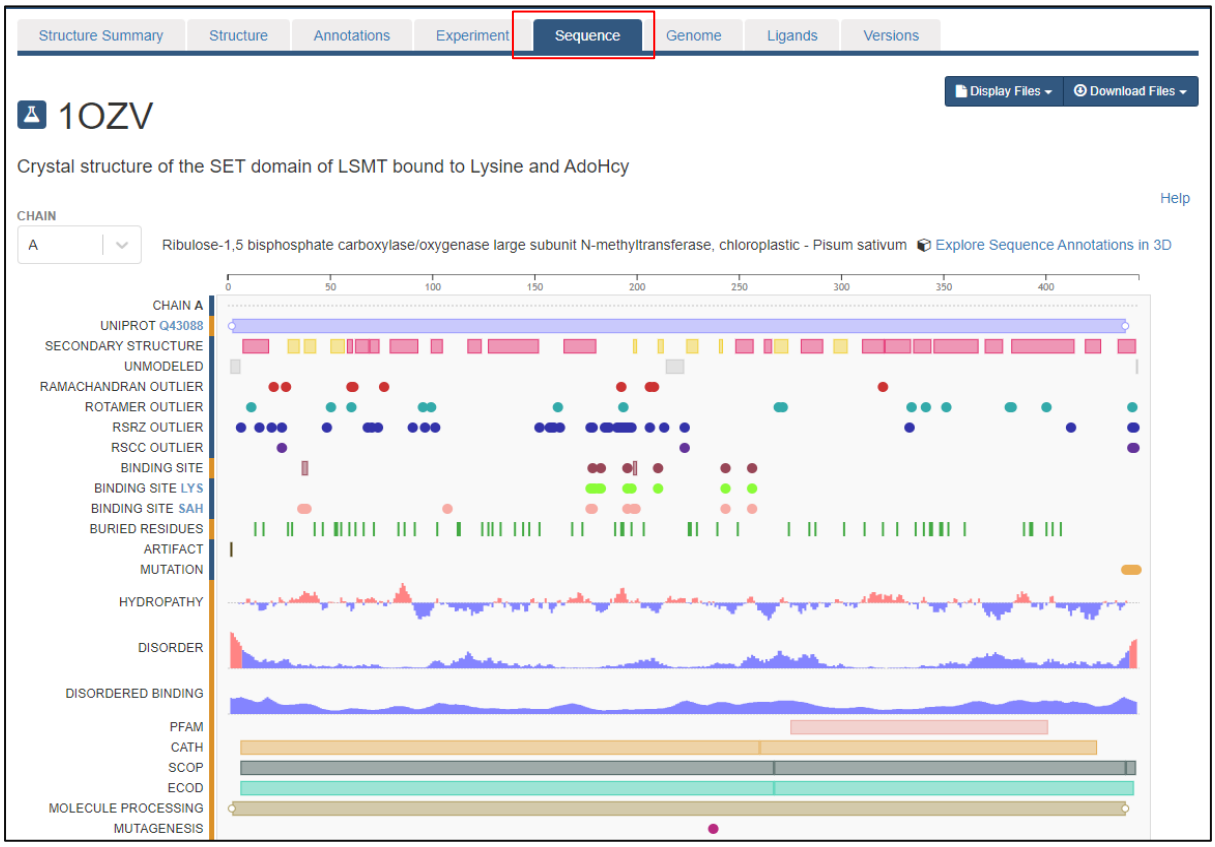


Figure 9: View of the Sequence Section

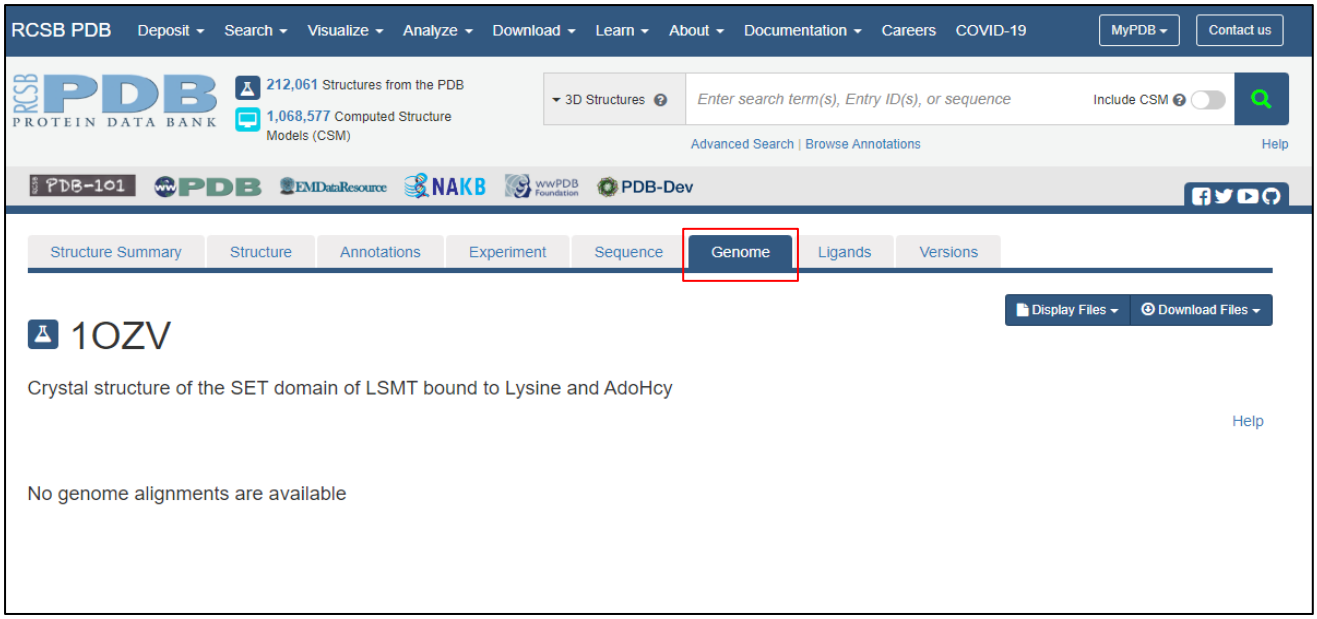


Figure 10: View of the Genome Section

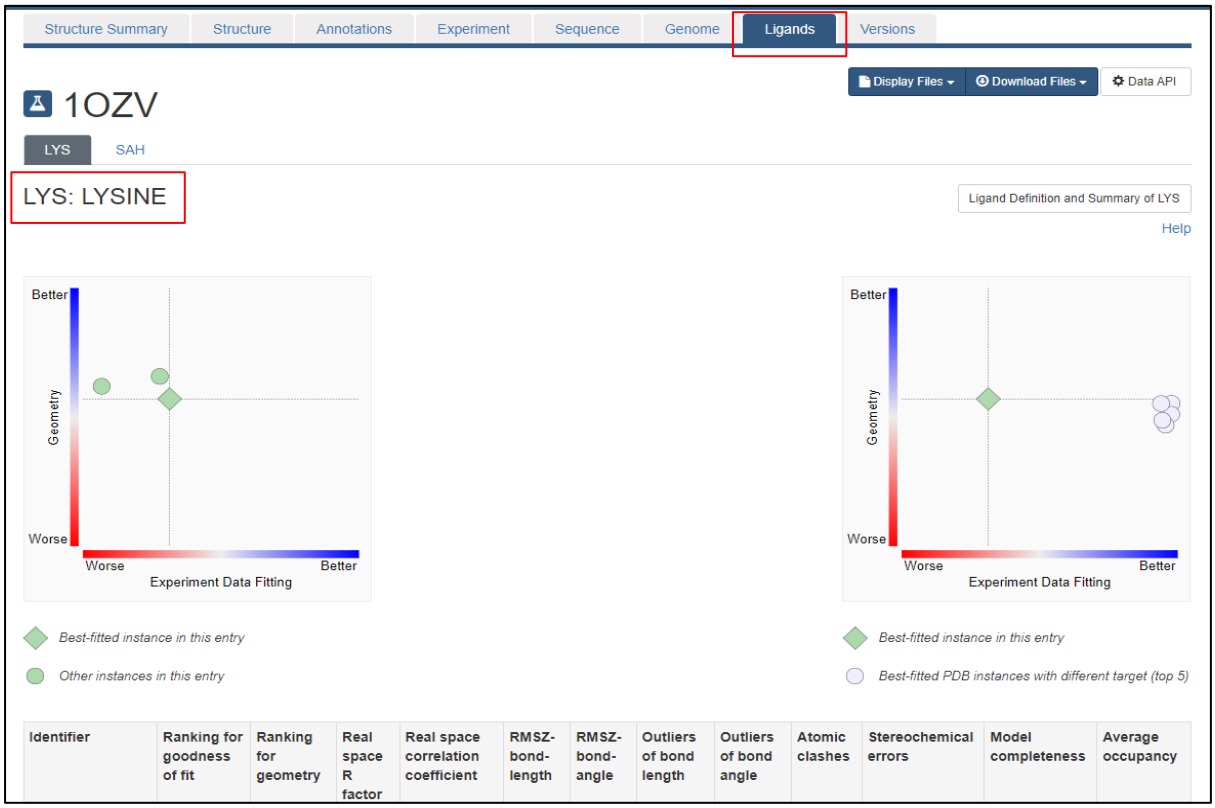


Figure 11: View of the Ligand Section showing LYS: Lysine

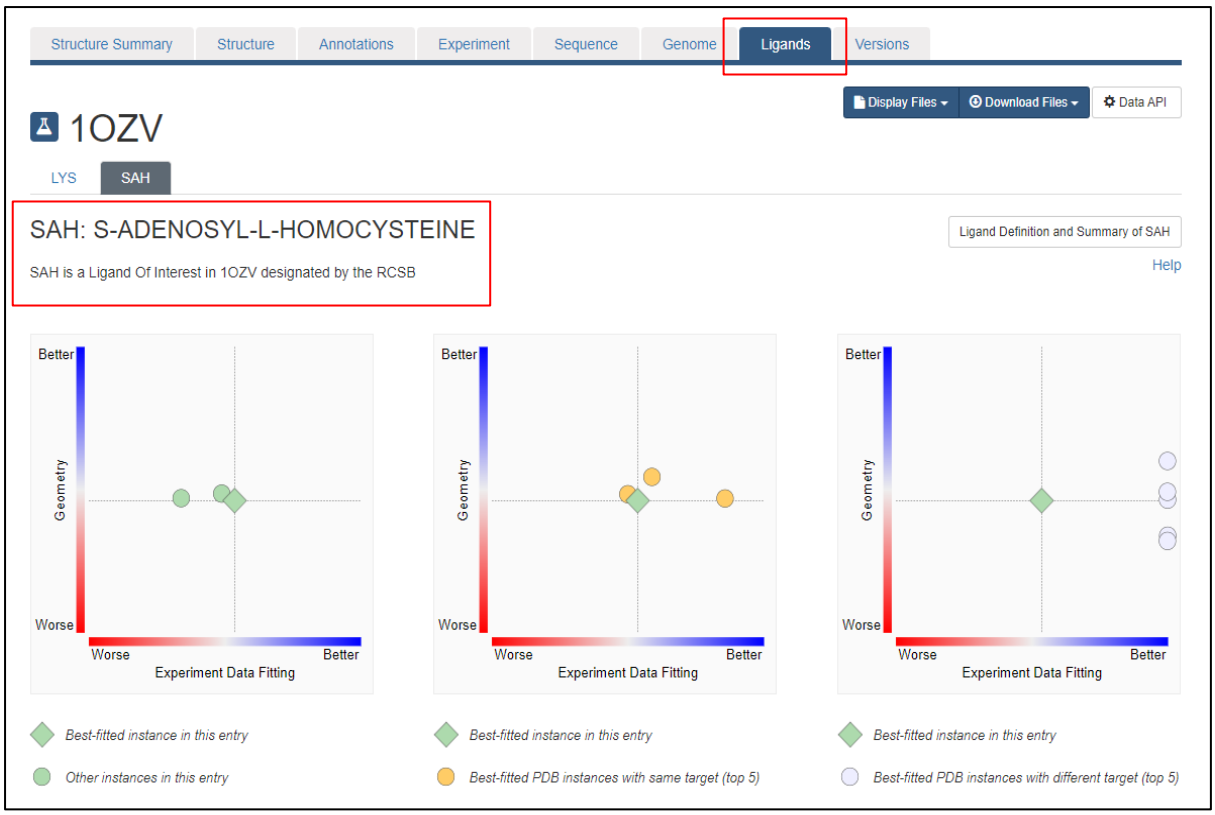


Figure 11a: View of the Ligand Section showing SAH: S-ADENOSYL-L-HOMOCYSTEINE

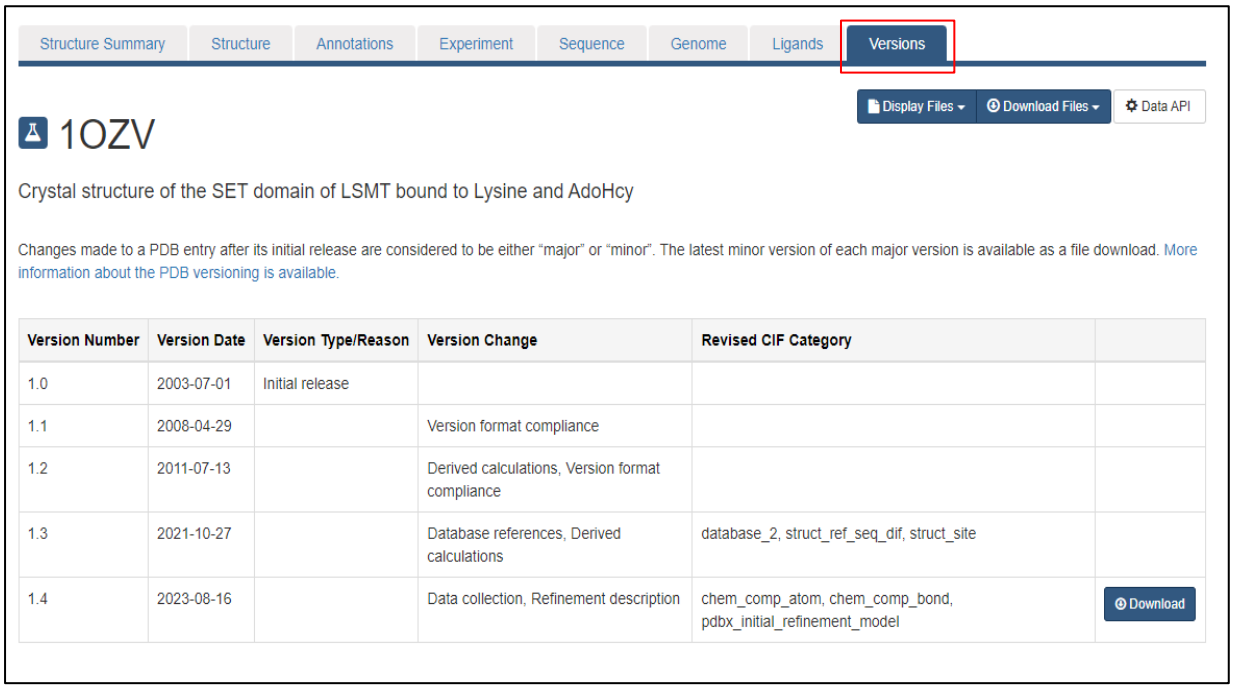


Figure 12: View of the Version Section

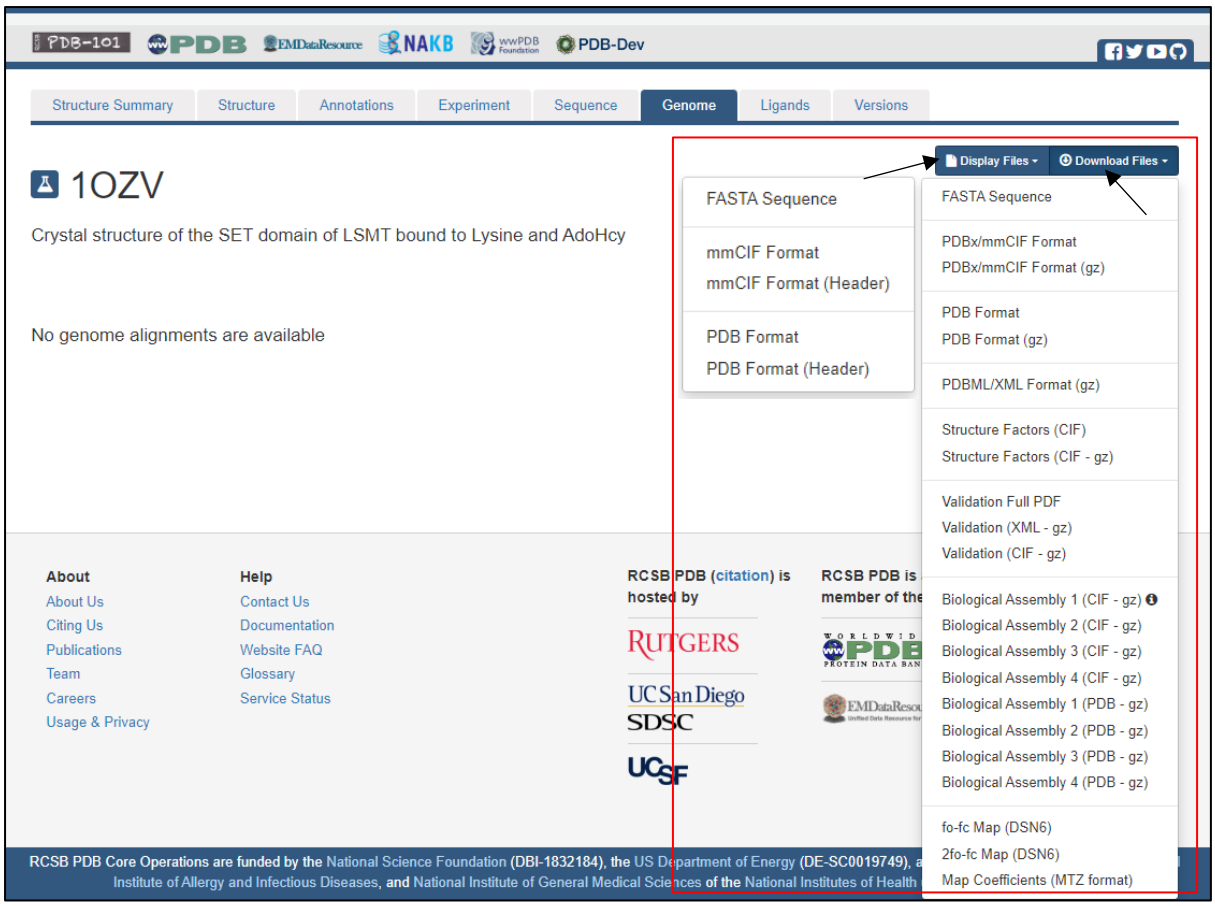
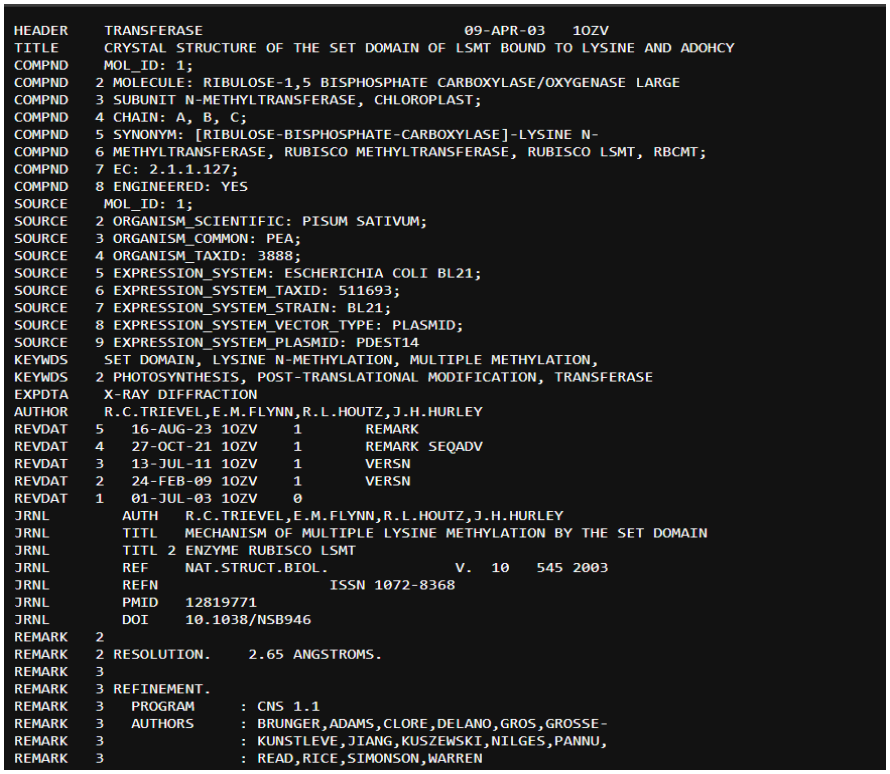


Figure 13: Display And Download Options



```
HEADER    TRANSFERASE                             09-APR-03   1OZV
TITLE     CRYSTAL STRUCTURE OF THE SET DOMAIN OF LSMT BOUND TO LYSINE AND ADOHCY
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: RIBULOSE-1,5 BISPHOSPHATE CARBOXYLASE/OXYGENASE LARGE
COMPND   3 SUBUNIT N-METHYLTRANSFERASE, CHLOROPLAST;
COMPND   4 CHAIN: A, B, C;
COMPND   5 SYNONYM: [RIBULOSE-BISPHOSPHATE-CARBOXYLASE]-LYSINE N-
COMPND   6 METHYLTRANSFERASE, RUBISCO METHYLTRANSFERASE, RUBISCO LSMT, RBCMT;
COMPND   7 EC: 2.1.1.127;
COMPND   8 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PISUM SATIVUM;
SOURCE   3 ORGANISM_COMMON: PEA;
SOURCE   4 ORGANISM_TAXID: 3888;
SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI BL21;
SOURCE   6 EXPRESSION_SYSTEM_TAXID: 511693;
SOURCE   7 EXPRESSION_SYSTEM_STRAIN: BL21;
SOURCE   8 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PDEST14
KEYWDS    SET DOMAIN, LYSINE N-METHYLATION, MULTIPLE METHYLATION,
KEYWDS   2 PHOTOSYNTHESIS, POST-TRANSLATIONAL MODIFICATION, TRANSFERASE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.C.TRIEVEL,E.M.FLYNN,R.L.HOUTZ,J.H.HURLEY
REVDAT   5   16-AUG-23 1OZV    1       REMARK
REVDAT   4   27-OCT-21 1OZV    1       REMARK SEQADV
REVDAT   3   13-JUL-11 1OZV    1       VERSN
REVDAT   2   24-FEB-09 1OZV    1       VERSN
REVDAT   1   01-JUL-03 1OZV    0
JRNL        AUTH   R.C.TRIEVEL,E.M.FLYNN,R.L.HOUTZ,J.H.HURLEY
JRNL        TITL   MECHANISM OF MULTIPLE LYSINE METHYLATION BY THE SET DOMAIN
JRNL        TITL 2 ENZYME RUBISCO LSMT
JRNL        REF    NAT.STRUCT.BIOL.                V.  10   545 2003
JRNL        REFN                   ISSN 1072-8368
JRNL        PMID   12819771
JRNL        DOI    10.1038/NSB946
REMARK   2
REMARK   2 RESOLUTION.    2.65 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : CNS 1.1
REMARK   3   AUTHORS     : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK   3               : KUNSTLEVE,JIANG,KUSZEWSKI,NILGES,PANNU,
REMARK   3               : READ,RICE,SIMONSON,WARREN
```

Figure 14: View of the sequence in PDB file format (Header)

RESULTS:

The Protein Data Bank (PDB) database was examined to investigate protein structures using the query 'lysine' with the PDB ID: 1OZV. A total of 18,676 protein structure entries were initially obtained through a basic search, and further refinement led to the identification of 327 structures. The results have been categorized into different sections, including Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome, Ligands, and Versions. The entry can be displayed and downloaded in the desired format for further analysis.

CONCLUSION:

The Protein Data Bank (PDB) stands as an essential and foundational resource in structural biology and bioinformatics. It serves as a repository for experimentally determined three-dimensional structures of biological macromolecules, including proteins, nucleic acids, and complex assemblies. Key features and contributions of the PDB include Comprehensive Repository, Global Collaboration, Structural Insights, etc. Thus, the Protein Data Bank remains an indispensable resource for structural biologists, researchers, educators, and clinicians worldwide. Its wealth of structural information plays a pivotal role in advancing scientific knowledge, aiding in various research endeavors, and paving the way for innovations in biomedicine and biotechnology.

REFERENCES:

1. Berman, H. M. (2000, January 1). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235
2. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977, May). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*(3), 535–542. https://doi.org/10.1016/s0022-2836(77)80200-3
3. Trievel, R. C., Flynn, E. M., Houtz, R. L., & Hurley, J. H. (2003, June 22). Mechanism of multiple lysine methylation by the SET domain enzyme Rubisco LSMT. *Nature Structural & Molecular Biology*, *10*(7), 545–552. https://doi.org/10.1038/nsb946

DATE: 30/10/23

WEBLEM 5(B)
NUCLEIC ACID KNOWLEDGEBASE (NAKB)/ NUCLEIC ACID DATABASE (NDB)
(URL: https://www.nakb.org)

AIM:
To explore the Nucleic Acid Knowledgebase (NAKB) / Nucleic Acid Database (NDB) for the study of the 3D structure of protein 'Helicase' (PDB ID: 8PJB).

INTRODUCTION:
One of the largest online databases devoted to experimentally discovered structures of DNA and RNA is the Nucleic Acid Knowledgebase (NAKB). The Nucleic Acid Database (NDB), which was created in 1992 with the main goal of storing and sharing structural data pertaining to nucleic acids, is expected to be replaced by these. By providing a multitude of features, such as search, report, statistics, atlas, and visualization sites, the NAKB Database outperforms the previous version. These include all experimentally discovered 3D structures containing nucleic acids that are stored in both the Protein Data Bank (PDB) Database and the NDB Database. NAKB Database includes data obtained from various methods like X-rays, NMR, and Electron Microscopy.

Each of the entry in the NAKB Database is well annotated and cross-liked to various databases like the PDB Database, the UniProt Database and the PMC Database, which enables the users to study the structure more detailed. The primary objective of the NAKB Database is to enable easy access for users to find and download relevant structures and metadata for their research, regardless of how general or specific it may be. Consistent weekly updates on Thursdays, guarantees the most recent data, which is prominently featured in the banner at the top of every page.

Prominent annotations contained within the NAKB Database include 2D fold diagrams, RNA view v.1.0, NA parameter tables, annotations, X3DNA-DSSR v.2.4, NA and Protein Sequence Clusters, CD-HIT, CD-HIT-ESTv.4.8.1, and more. The National Institutes of Health (NIH) R01 GM085328 provides funding for this essential resource, which is run out of the Institutes for Quantitative Biomedicine at Rutgers University, 174 Frelinghuysen Rd, Piscataway, NJ USA, 08854-8076. Citations to the NAKB Database are encouraged for researchers who use it in their research.

Helicase:
Helicases are essential enzymes involved in all aspects of nucleic acid metabolism, including DNA replication, repair, recombination, transcription, ribosome biogenesis, and RNA processing, translation, and decay. They are part of molecular complexes that include components required for each specific step of nucleic acid metabolism, utilizing the energy derived from nucleoside triphosphate hydrolysis to translocate along nucleic acid strands, unwind/separate the helical structure of double-stranded nucleic acid, and, in some cases,

disrupt protein-nucleic acid interactions. Helicases are ubiquitous and evolutionary conserved proteins, making them crucial for the dynamic behavior and structural integrity of cells.
The Nucleic Acid Knowledgebase (NAKB) can be used to study and analyze 'Helicase' (PDB ID: 8PJB) to understand essential function of helicases in various cellular processes and their role in maintaining the proper structure and function of nucleic acids. By analyzing the structure and function of helicases, researchers can gain insights into mechanisms and interactions of helicase, which can help develop a deeper understanding of their role in various cellular processes and potentially lead to new therapeutic strategies for diseases related to nucleic acid metabolism.

METHODOLOGY:

1. Open the Protein data Bank (PDB) database and search for the query of 'Helicase'.
2. From the results page, open the protein of interest and retrieve its PDB ID (Here, 8PJB).
3. Open the homepage of the NAKB database.
4. Enter the PDB ID retrieved for the query of 'Helicase' (PDB ID: 8PJB) and click on basic search.
5. Information regarding query is displayed with respect to components, assemblies and images.
6. To view and download the structures, click on the 'View and Download' option. Then, select the desired format for view and download.
7. Interpret the results obtained.

OBSERVATIONS:

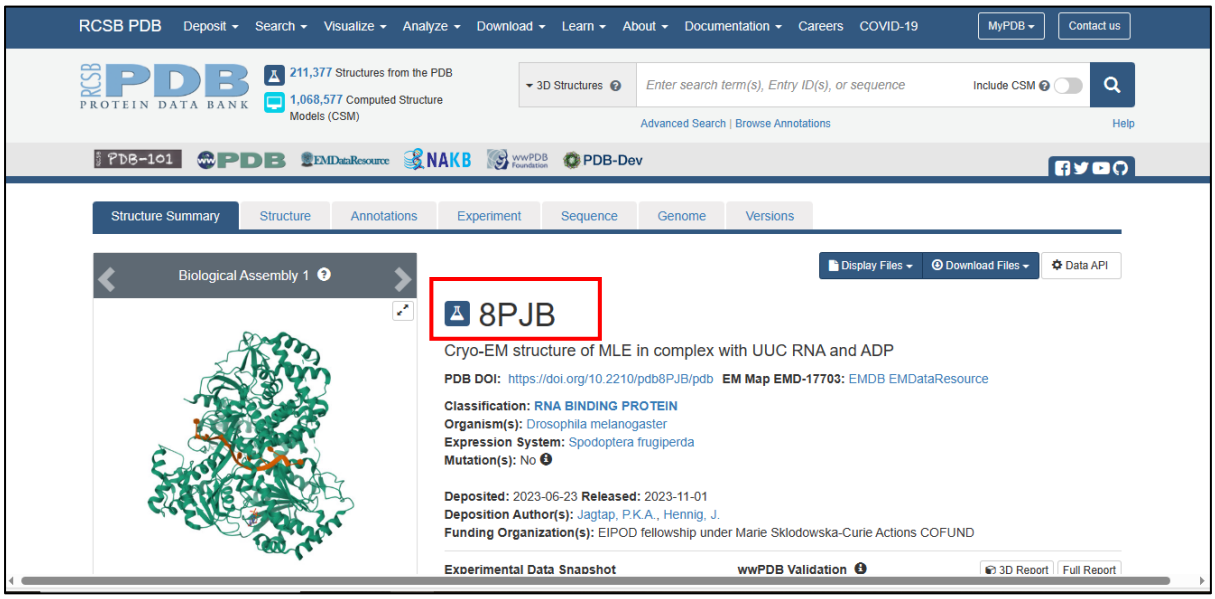


Figure 1: Protein Helicase (PDB ID: 8PJB) on the Protein Data Bank (PDB) Database

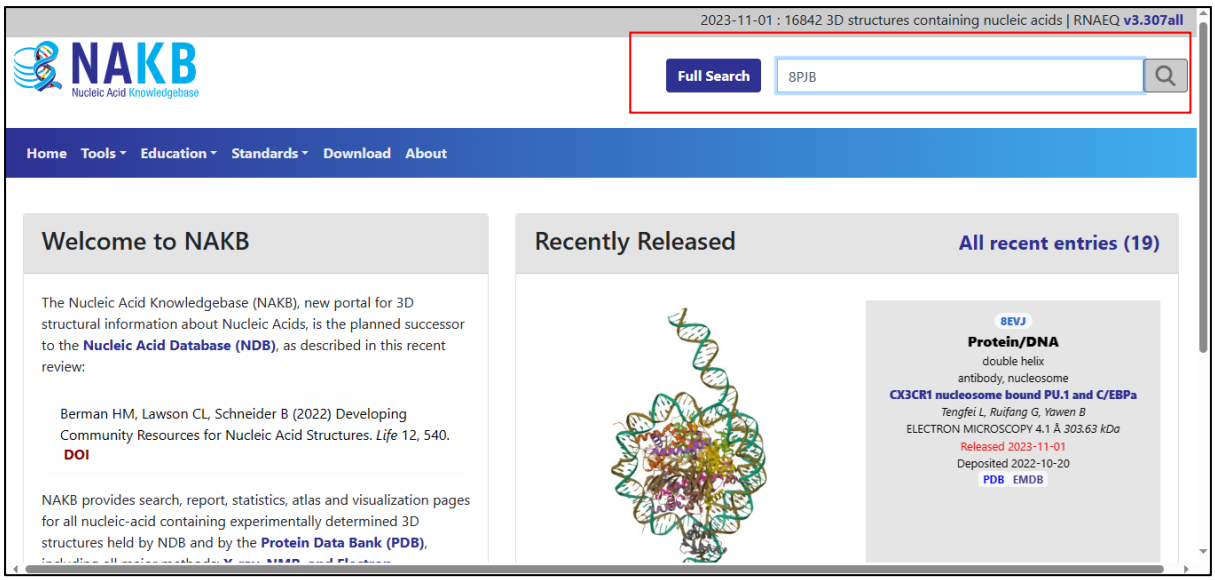


Figure 2: Homepage of the Nucleic Acid Knowledgebase (NAKB) Database

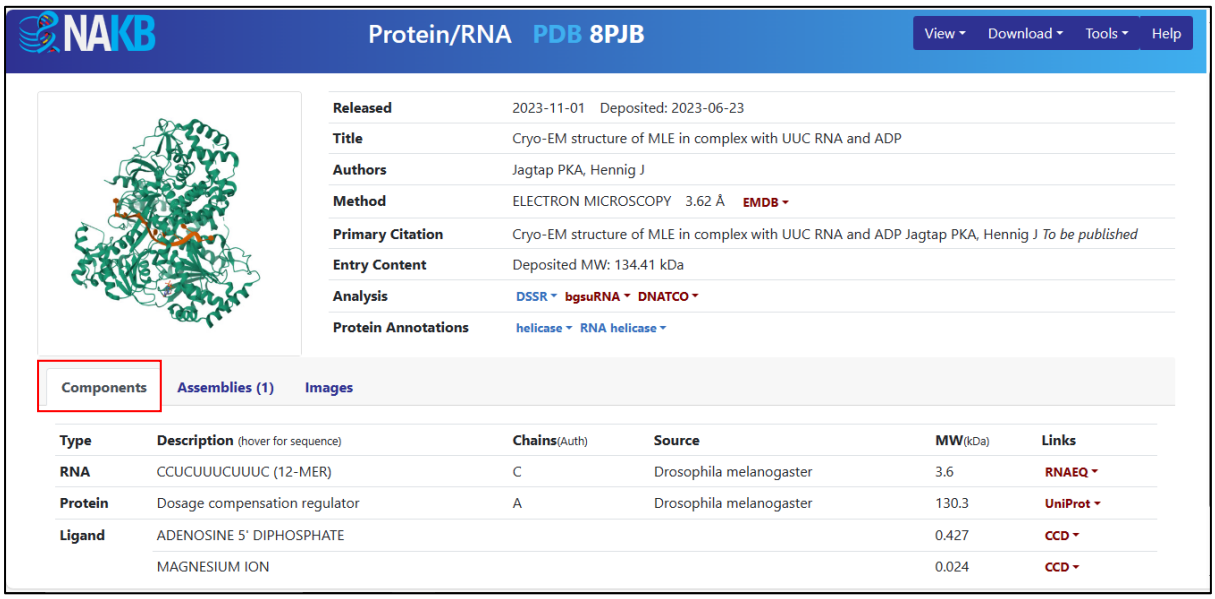


Figure 3: Entry of protein Helicase (PDB ID: 8PJB) with Components result section

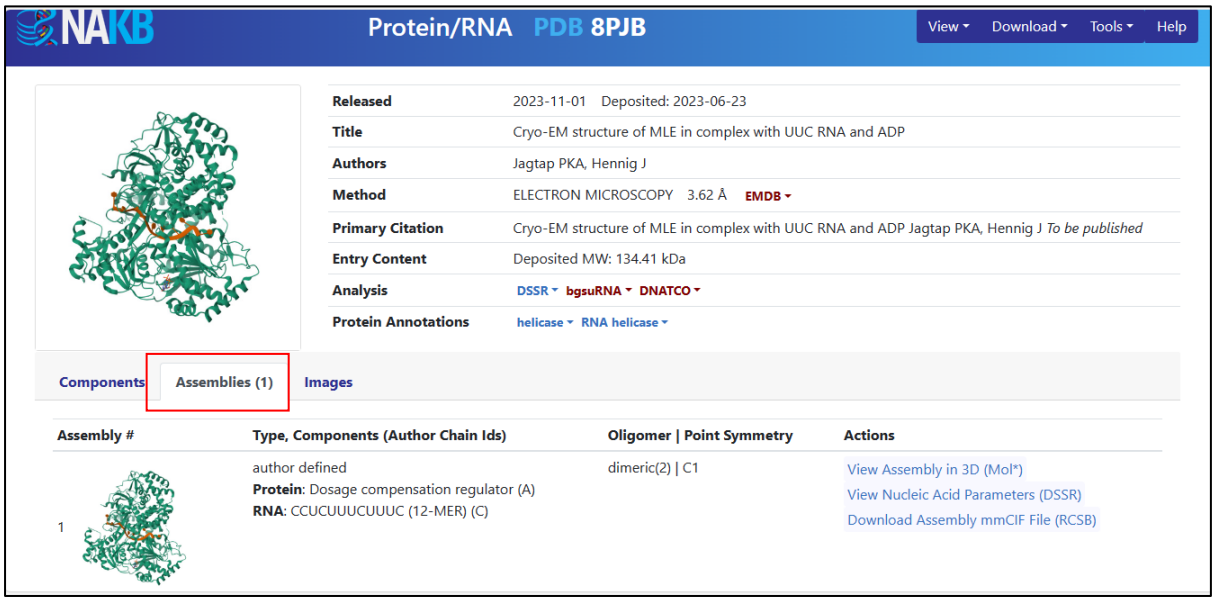


Figure 4: Entry of protein Helicase (PDB ID: 8PJB) with Assemblies result section

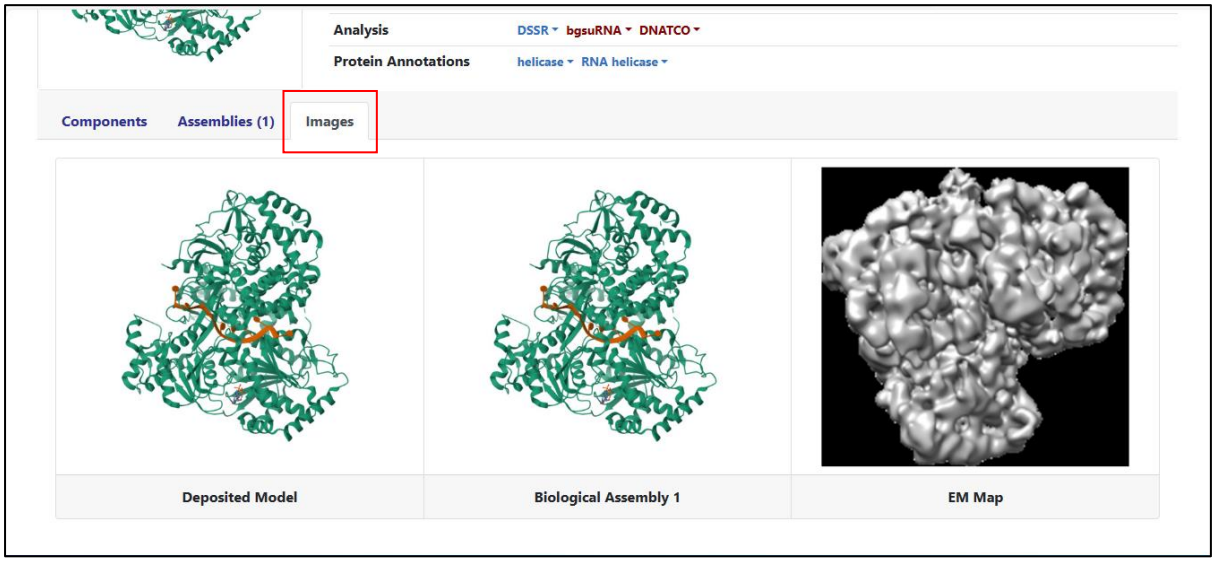


Figure 5: Entry of protein Helicase (PDB ID: 8PJB) with Images result section

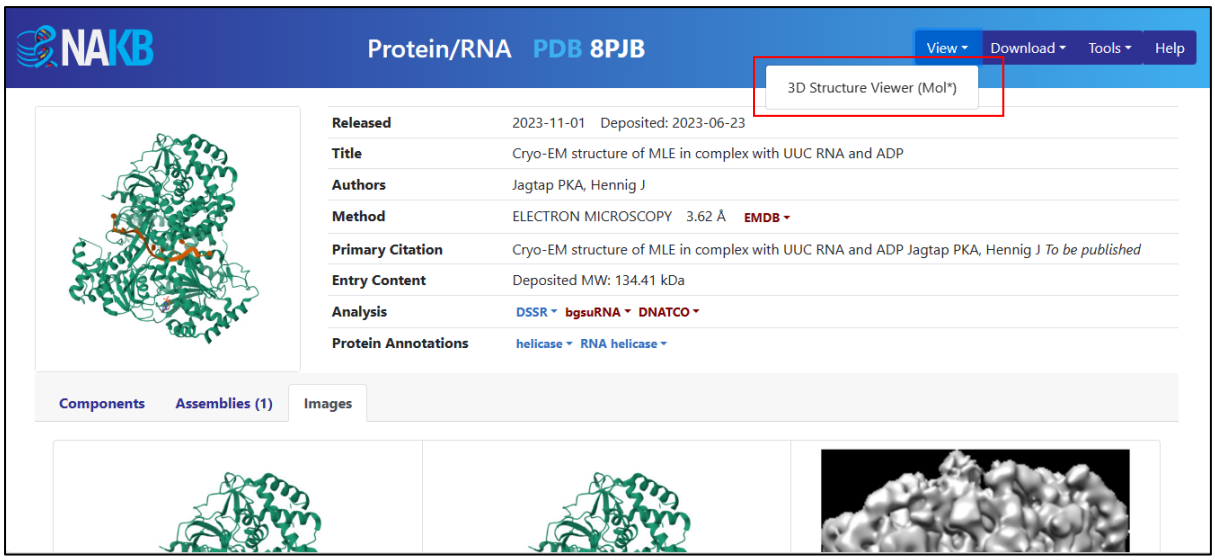


Figure 6: View option to view protein 3D structure viewer (Mol*)

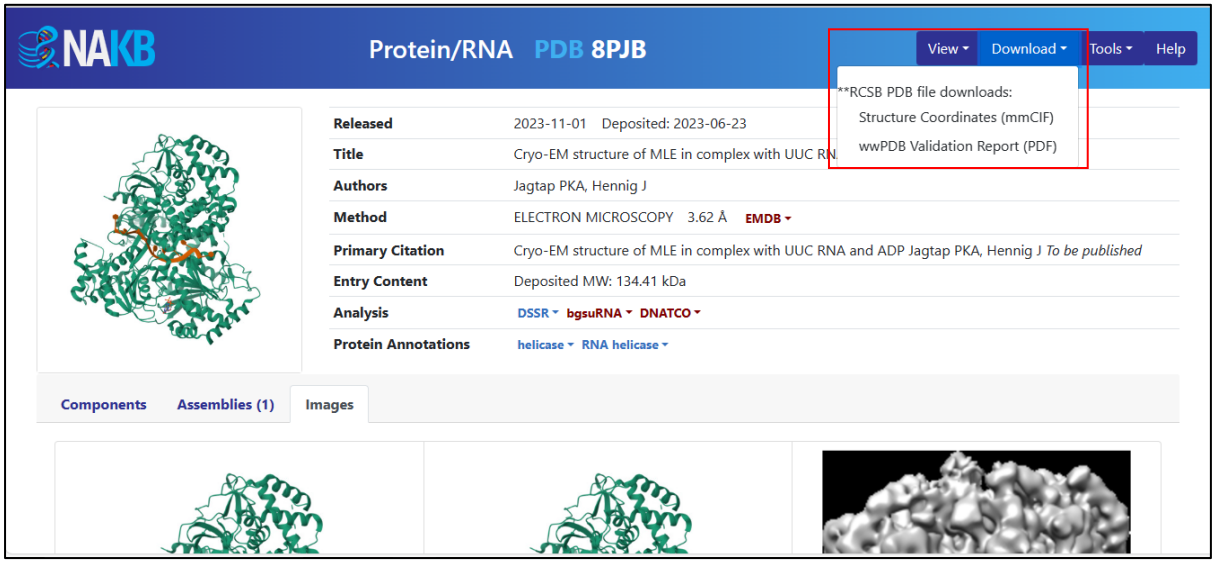


Figure 7: Download option to download the structure coordinates (mmCIF) and wwPDB validation report (PDF) files

RESULTS:

The NAKB database was explored for the query 'Helicase' (PDB ID: 8PJB). The results were observed in three sections: Components, Assemblies (with only one result obtained), and Images. Structural images were viewed and downloaded in the preferred file format.

CONCLUSION:

The Nucleic Acid Knowledgebase (NAKB) was explored to study the 3D structure of protein 'Helicase' (PDB ID: 8PJB) along with its associations with the nucleic acids.

REFERENCES:

1. Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A. I., Sweeney, B., Zirbel, C. L., Leontis, N. B., & Berman, H. M. (2013, October 31). The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Research*, *42*(D1), D114–D122. https://doi.org/10.1093/nar/gkt980
2. Berman, H. M., Lawson, C. L., & Schneider, B. (2022, April 6). Developing Community Resources for Nucleic Acid Structures. *Life*, *12*(4), 540. https://doi.org/10.3390/life12040540
3. Berman, H. M., Zardecki, C., & Westbrook, J. (1998, November 1). The Nucleic Acid Database: A Resource for Nucleic Acid Science. *Acta Crystallographica Section D Biological Crystallography*, *54*(6), 1095–1104. https://doi.org/10.1107/s0907444998007926

DATE: 30/09/2023

WEBLEM 5(C)
CARBOHYDRATE STRUCTURE DATABASE (CSDB)/ CCSD / GLY-TOU-CAN DATABASE
(URL: http://csdb.glycoscience.ru/)

AIM:

To explore the CSDB for the query Antigen of Blood group H2. (Compound ID: 13199)

INTRODUCTION:

Carbohydrate Structure Database (CSDB) is a regularly updated database containing structural, taxonomic, bibliographic, NMR spectroscopic, and other information on carbohydrates and their derivatives obtained from prokaryotes, plants, and fungi. CSDB claims for full coverage and high data quality. It serves as a platform for various search strategies, tools for NMR spectrum and structure prediction, and instruments for statistical analysis. It was launched in 2005 by a group of Russian scientists from N.D.Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences. The time lag between publishing and deposition is one year. CSDB serves as a platform for multiple services of glycoinformatics, such as NMR spectra simulation, NMR-based structure elucidation, molecular geometry prediction, taxon clustering, glycobiological statistical tools, etc. The project aims at coverage close to complete in selected taxonomic domains and at high data quality achieved by manual literature analysis, annotation, verification, and data approval. The data in bacterial part of CSDB are regularly updated. The time lag between publishing and deposition is one year. CSDB is cross-linked to other glycoinformatics projects and NCBI databases.

The database can be searched as:

1. **Database search by structures**: This form lets you search the database by fragments of chemical structure. As you enter the structure by one of these methods, pressing Return the structure to the search page, returns you back to the structure search form with the pre-filled search term field.
2. **Database search by Composition**: This form allows searching structures by their residue composition, e.g., MS data. The default composition is a single hexose residue. Drop-down list lets you select a residue base type (e.g., HEX, Glc, GlcN etc.) without configurations and ring size. Only most widespread residues are included.
3. **Database search by Organism**: This form allows retrieving organisms and associated data by their taxonomic names. Alphabetical lists of genera, species and strains/serogroups provide taxonomic specification (position of an organism in the tree of life).
4. **Database search by Publications**: This form is proposed for search using bibliographic data and keywords. If search criteria are provided in several sections of this form, the intersection of queries will be returned. The queries are case-insensitive and accent-independent
5. **Database search for NMR signals:** This form allows searching for compounds with NMR spectra containing the specified signals. Selector allows selection of a particular

nucleus. The sub spectrum to search for should be typed in window. You can separate signals with spaces or new line characters, the sorting is not required; the allowed characters are numerals and decimal dot.

Useful tools:

1. **Predict NMR**: This feature is available from the NMR simulation link in Extras section of the main menu, and from the (Sub)structure search form. To simulate the NMR data, you should first enter the structure of interest. This structure is previewed in SNFG format in area and copied to the structure field as a term in CSDB Linear encoding. The structure can be refined by manual editing of this field.

2. **Elucidate**: This tool aims at helping in structural elucidation studies and NMR spectrum assignment. It uses GRASS algorithm, which stands for Generation, Ranking and Assignment of Saccharide Structures. It iterates through all possible carbohydrates and their derivatives limited by specified constraints. For each generated structure, an empirical 13C NMR spectrum is simulated and is compared to the provided experimental data. Not more than 500 best fitting structures are further refined to give a few top-matching structures. These structures are displayed as best matches together with the simulated NMR data.

3. **Fragments**: This feature provides the distribution of monomeric or dimeric fragments in structures from specified taxonomic group(s) and data on their uniqueness and location in the structures. Rank selector determines the rank of taxa to analyze: Domain, Phylum, Class, Genus, Species or Strain. Depending on the selected rank, the taxa available in the database are displayed as list. For domains, no lists are displayed, and domains can be selected using the Display groups checkboxes. For other ranks, Display groups filter the taxa to those included in the checked domains only. For species or strains, an additional list of genera is provided for easier selection.

4. **Cluster Taxa**: This feature generates a distance matrix between mono- or dimeric fragment pools from taxa populated in both the (bacterial and plant and fungal) databases. Based on this matrix, the taxa are clustered into groups, and the corresponding dendrograms are displayed. The exported matrix can be used for clustering of taxa according to the glycans they biosynthesize, can be visualized as phenetic trees or processed externally. The exported matrix can be used for clustering of taxa according to the glycans they biosynthesize, can be visualized as phenetic trees or processed externally.

5. **GT activities**: The CSDB GT sub database provides close-to-full coverage data on a few species (including the most studied plant, bacterium, and fungus): Arabidopsis thaliana, Escherichia coli, Saccharomyces cerevisiae, Acinetobacter baumannii.

a. **Names / IDs:** enzyme name or enzyme group name, CAZY family, enzyme Uniprot ID, gene GenBank ID, internal CSDB GT identifier.
b. **Organism:** you can select the origin species and optionally type subspecies/strain. Currently, only E. coli and A. thaliana are available, however more data will be returned if species is set to ANY.
c. **Molecule role:** allows filtering enzymes to certain cellular roles of a product they synthesize, or its analog. This selector contains roles like O-antigen, lipid A, CPS, etc.
d. **Synthesized bond:** the main result of glycosyltransferase activity presented as a dimeric fragment with residues linked by a specific bond.
e. **Donor and/or acceptor:** the structure of the carrier of the transferred monosaccharide, and of the substrate it is carried to. The input and preview options are the same as for the synthesized dimeric fragment.
f. **Confirmation status:** the type of evidence for the glycosyltransferase activity. With direct or indirect evidence - GTs with activity supported by direct, semidirect, or indirect evidences; Confirmed strictly in vivo - GTs with activity supported by direct evidences.

Blood Group Antigen: Blood Group H2

The H antigen is a carbohydrate sequence with carbohydrates linked mainly to protein (with a minor fraction attached to ceramide moiety). The H antigen is a precursor to the ABO blood group antigens. Generally, BG-H2 antigens are expressed on red blood cells and vascular endothelium. They are present on the surface of the surface of the RBCs. The H-2 system of antigens is a highly complex one, comprising at least 30 different alloantigenic specificities. More than 20 alleles or haplotypes are well-defined, and each determines a different combination of these specificities. The H-2 system constitutes the major set of histocompatibility and blood group antigens in the mouse. These antigens are widely distributed on the tissues, and when incompatibilities occur, they play a major role in graft rejection.

METHODOLOGY:

1. Go to the CSDB website.
2. Click structures under database search and select library option.
3. Click on blood group H2, from carbohydrate library blood group antigen section.
4. Click on, "Return the above structure".
5. Click "Go!"

OBSERVATIONS:

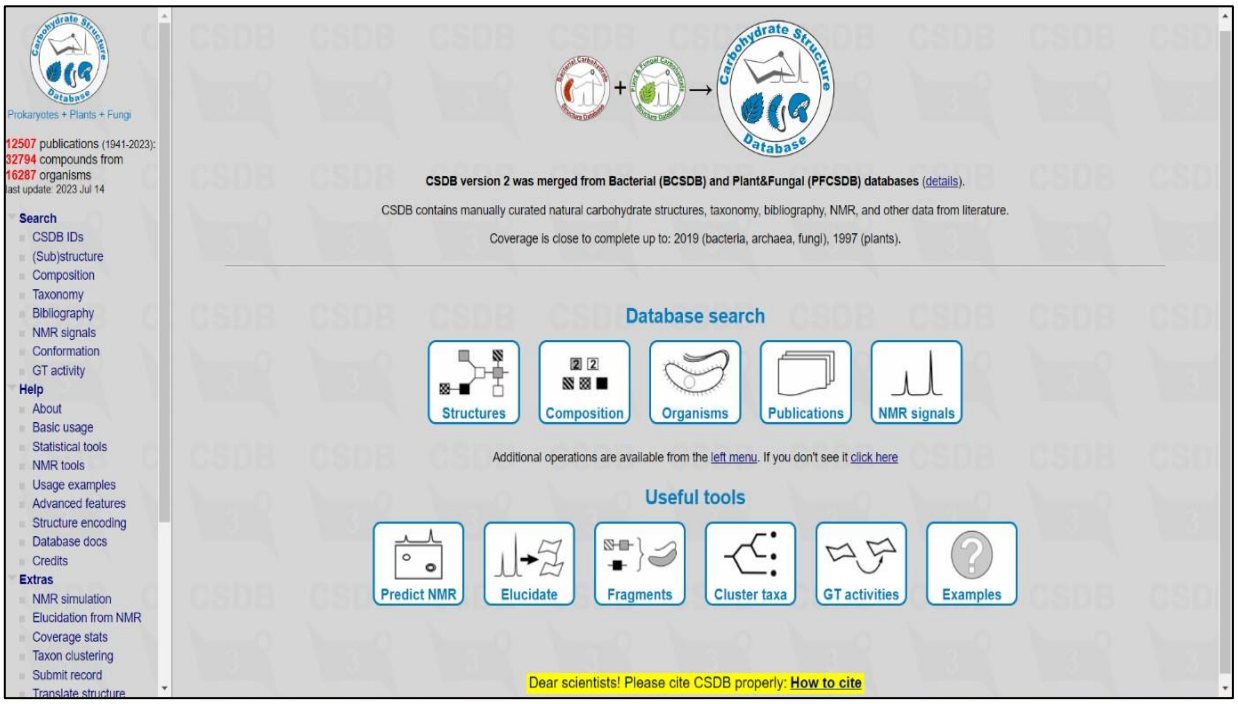


Figure 1: Homepage of CSDB Database

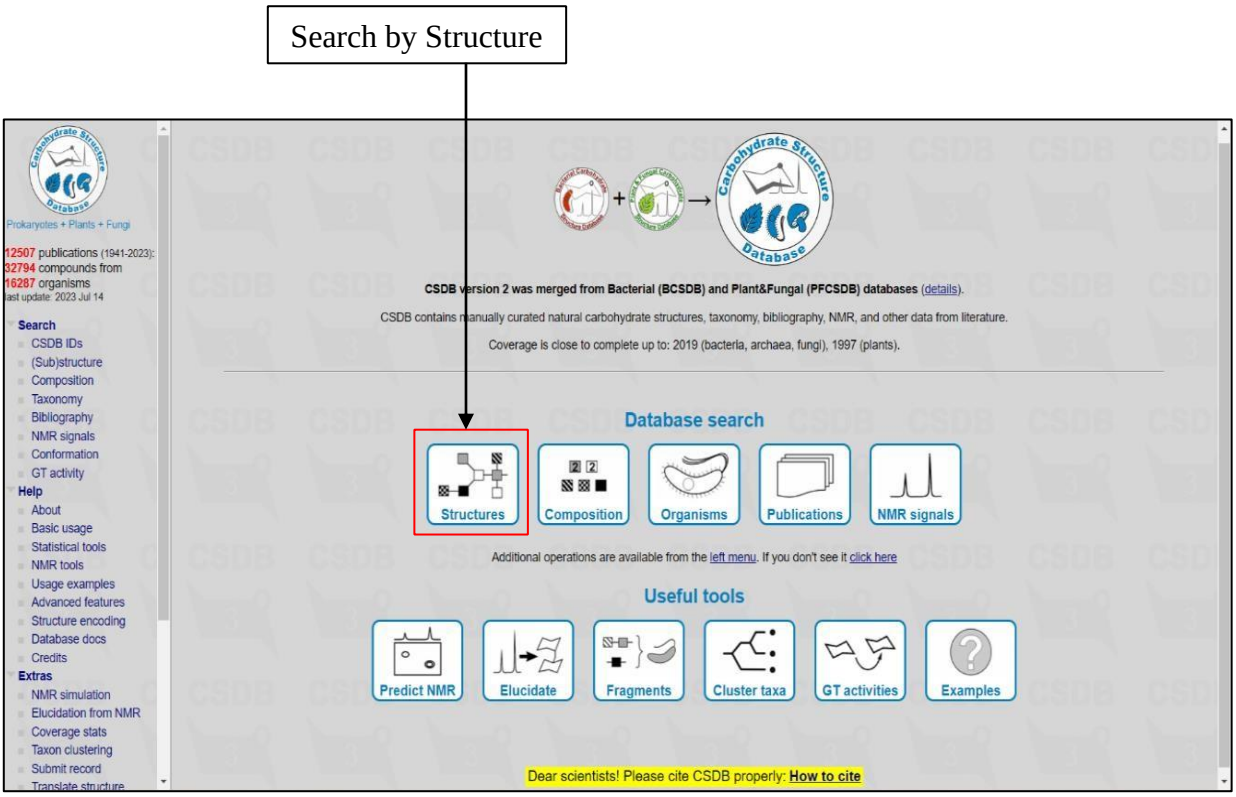


Figure 2: Structure search for the query

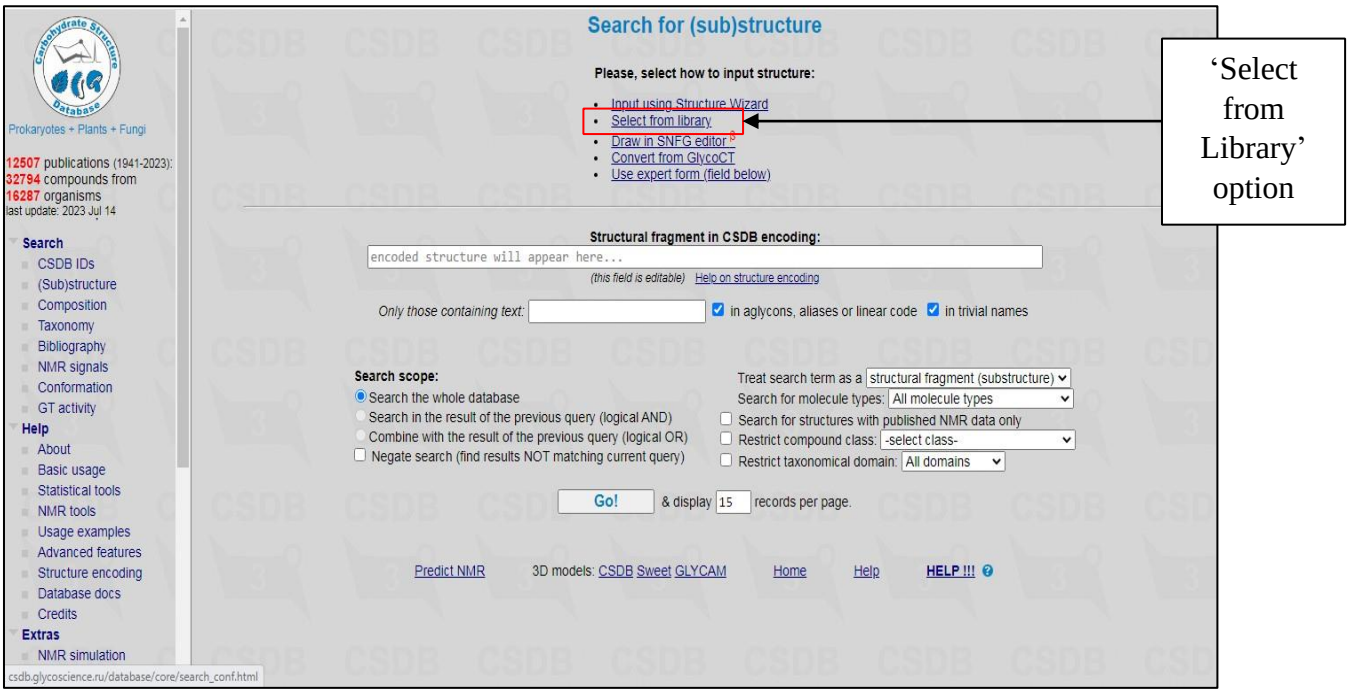


Figure 3: Selecting from Library

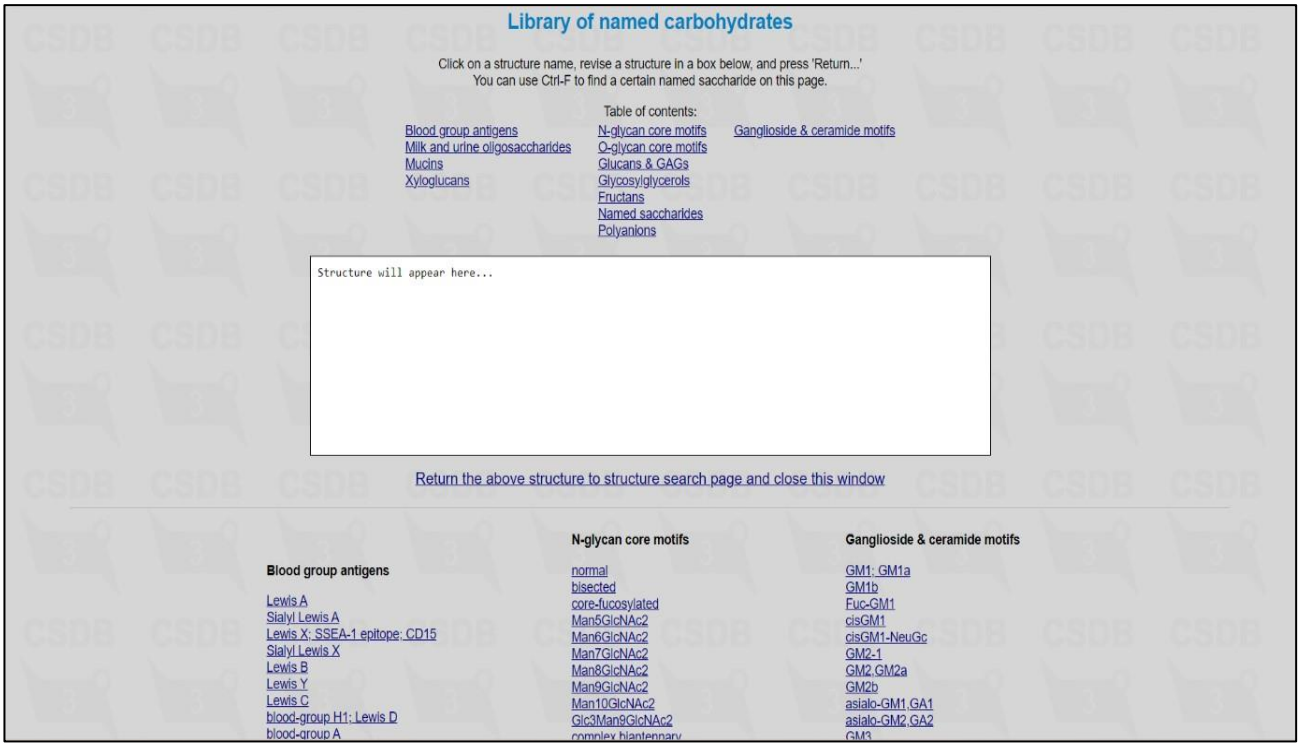


Figure 4: Library page after clicking on structure search

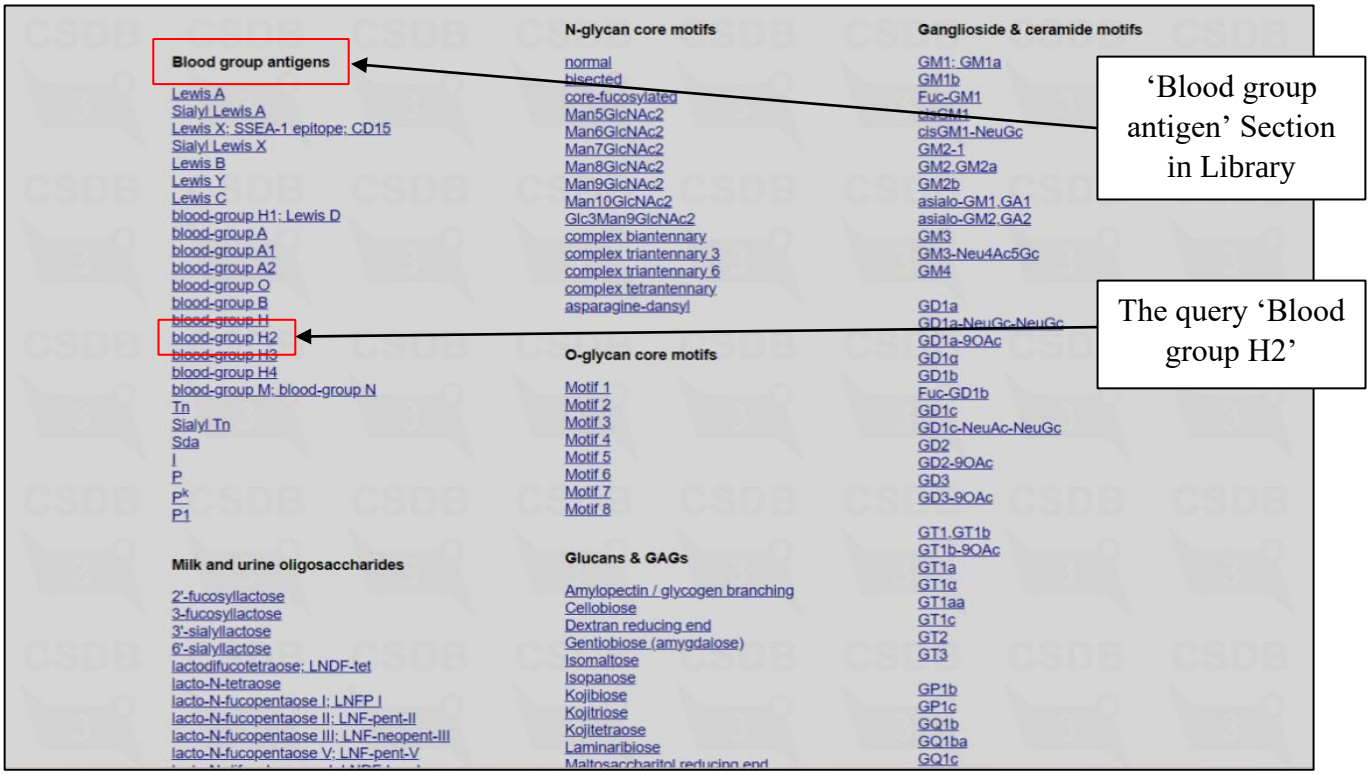


Figure 5: Blood Group Antigen section in the Library and the query Blood Group H2

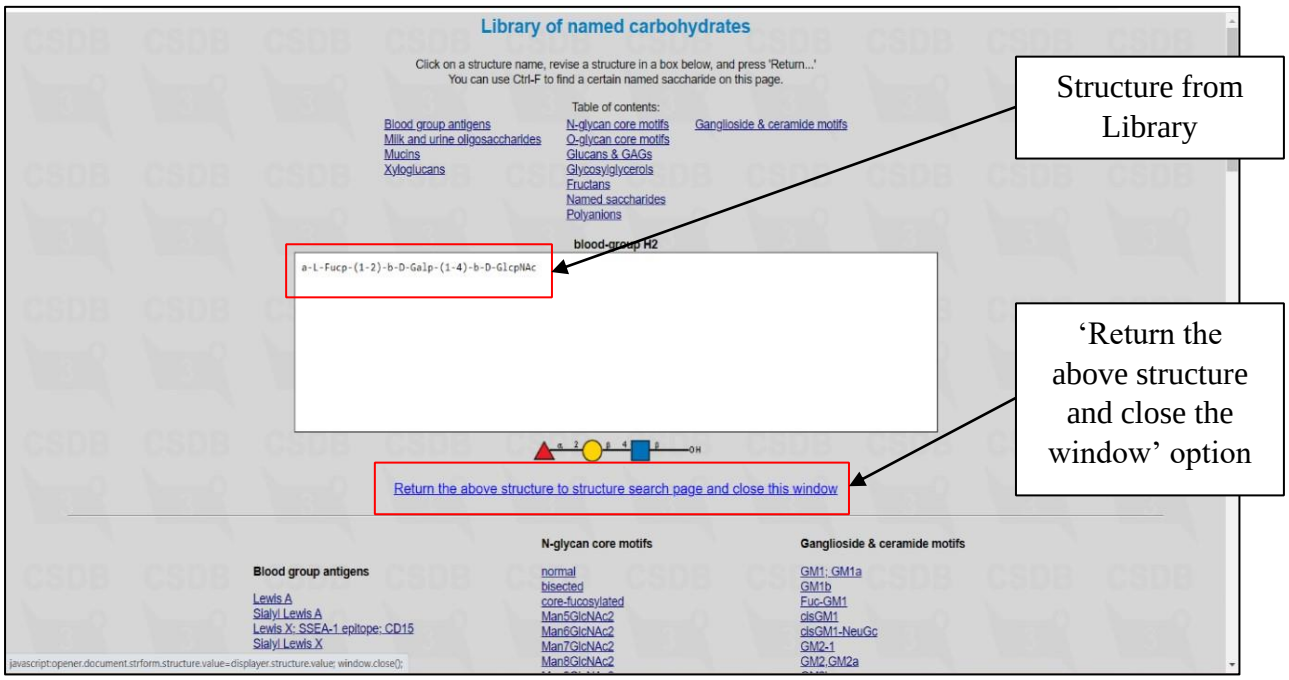


Figure 6: Structure for the Blood Group H2 antigen in the box and return above structure to search page

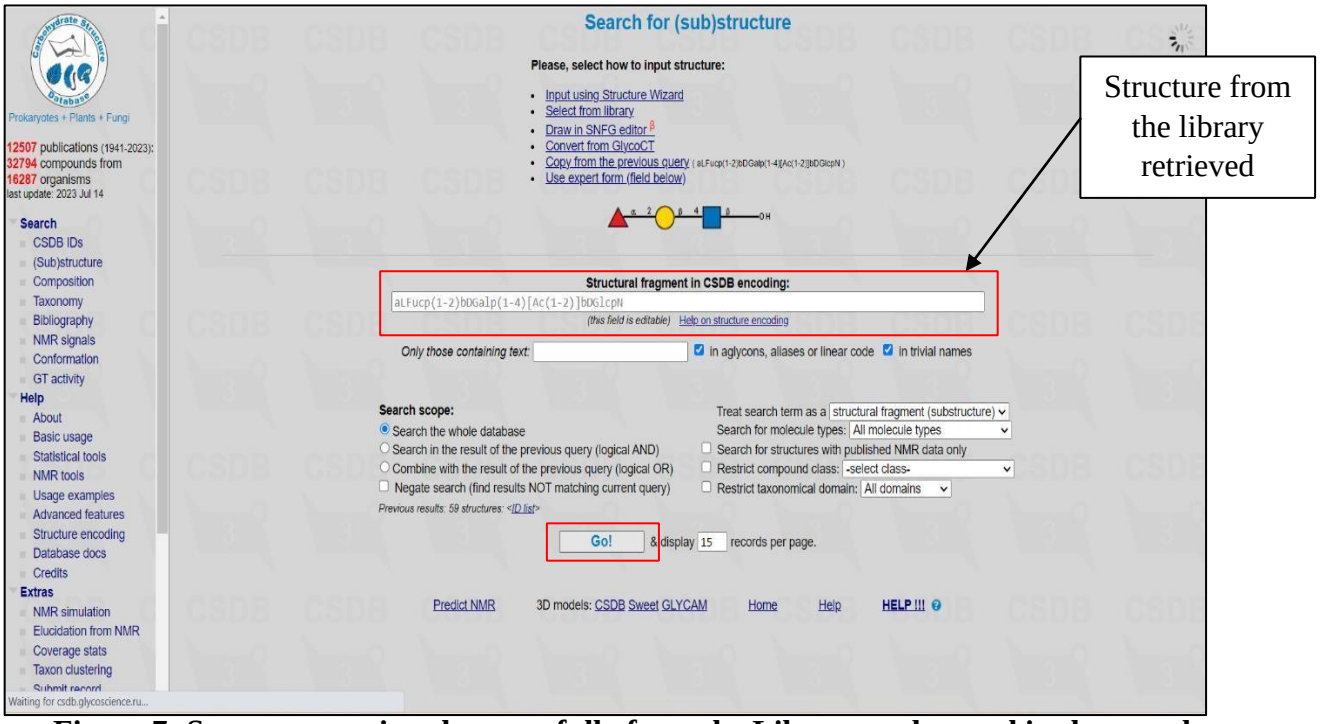


Figure 7: Structure retrieved successfully from the Library and pasted in the search box

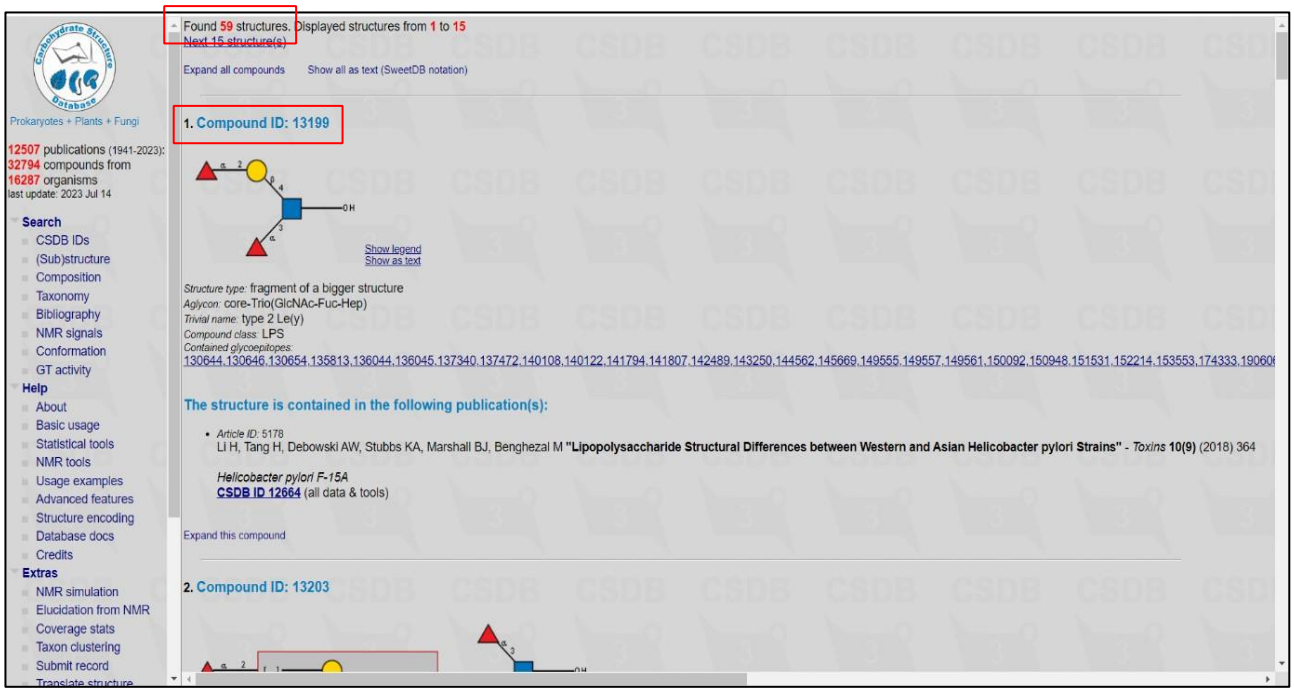


Figure 8: Hits for the query Blood Group Antigen: Blood Group H2 (59 entries were obtained)

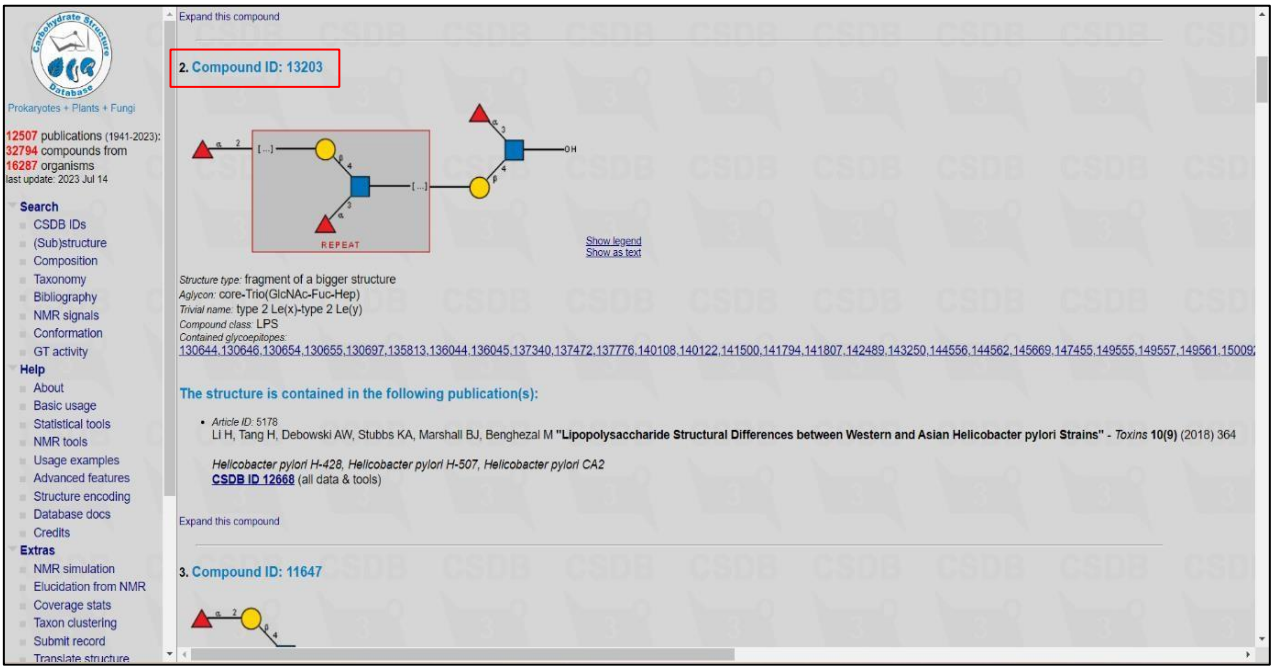


Figure 9: Second Entry for the Hit (Compound ID: 13203)

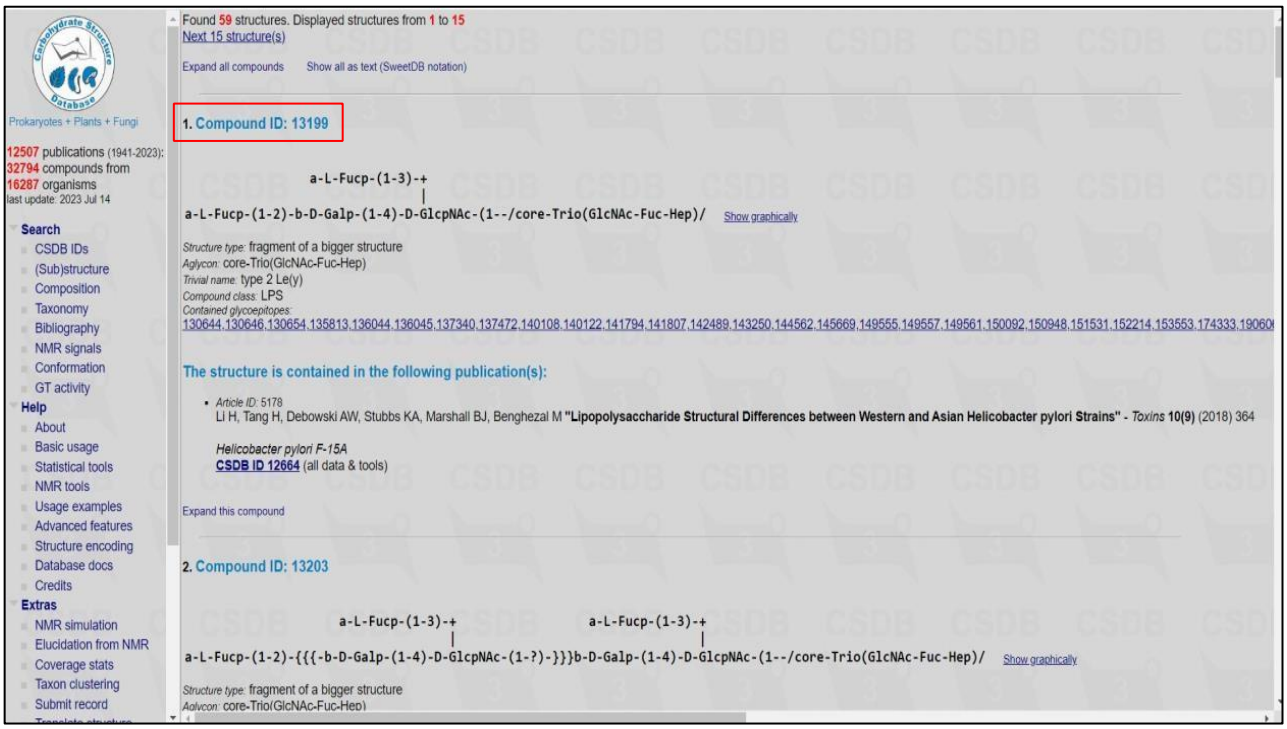


Figure 10: Result of structure in Text form (Compound ID: 13199)

RESULTS:

The query 'Blood Group H2' under Blood group antigen was studied using the Carbohydrate Structure Database (CSDB) where the data for specific query are integrated from the various sources and the data about the Structure type, Aglycon, Trivial name, Compound class, the glycoepitopes, the publications describing the compound, Compound ID are displayed below the Structure of the compound. The query was fired using the 'Search by structure' option and the relevant information is retrieved from the database.

CONCLUSION:

The Carbohydrate Structure Database is a resource used to retrieve the structure of the published carbohydrates and their derivatives and it also provides access to several carbohydrate related search tools. The CSDB integrates information from various sources and provides information to the researchers. Here, the CSDB was explored for the query 'Blood group H2' and the required results were obtained and used for understanding the properties and the interaction of the antigen in the body.

REFERENCES:

1. Toukach, F. V., & Egorova, K. S. (2015). Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. Nucleic Acids Research, 44(D1), D1229–D1236. https://doi.org/10.1093/nar/gkv840
2. Egorova, K. S., Kalinchuk, N. A., Knirel, Y. A., & Toukach, F. V. (2015). Carbohydrate Structure Database (CSDB): new features. Russian Chemical Bulletin. https://doi.org/10.1007/s11172-015-1003-6
3. Cox, J., & Pavic, A. (2014). SALMONELLA | Introduction. In Elsevier eBooks (pp. 322–331). https://doi.org/10.1016/b978-0-12-384730-0.00294-9

DATE: 04/11/23

WEBLEM 5(D)
REACTOME PATHWAY DATABASE
(URL: https://reactome.org/)

AIM:

To explore the Reactome pathway database with query Glycogenolysis pathway (R-HSA-70221).

INTRODUCTION:

Reactome Pathway Database is a curated database of pathways and reactions in human biology. Reactions can be considered as pathway 'steps'. Reactome defines a 'reaction' as any event in biology that changes the state of a biological molecule. Binding, activation, translocation, degradation and classical biochemical events involving a catalyst are all reactions. Information in the database is authored by expert biologists, entered and maintained by Reactome Pathway Database's team of Curators and Editorial staff. Reactome Pathway Database content frequently cross-references other resources e.g., Ensembl, UniProt, KEGG (Gene and Compound), ChEBI, PubMed and GO. Inferred orthologous reactions. NCBI, Ensembl, UniProt, KEGG (Gene and Compound), ChEBI, PubMed and GO. Inferred orthologous reactions Inferred orthologous reactions are available for 15 non-human species including mouse, rat, chicken, puffer fish, worm, fly, yeast, rice, and Arabidopsis.

Here are some key points about Reactome Pathway Database:

1. **Pathway Curation:** Reactome Pathway Database extensively curates pathways, capturing the sequence of molecular events involved in various biological processes. This includes pathways related to metabolism, signaling, cell cycle, and immune response.
2. **Molecular Entities:** The database includes detailed information about molecular entities such as proteins, small molecules, and complexes. This allows users to explore the relationships and interactions between these entities in the context of specific pathways.
3. **Data Integration:** Reactome Pathway Database integrates data from diverse sources, including literature, experimental data, and other databases. This ensures a comprehensive and reliable representation of biological processes.
4. **Accessibility:** Reactome Pathway Database is freely accessible to the scientific community and the public. The web interface provides user-friendly tools for searching, browsing, and visualizing pathway information.
5. **Analysis Tools:** In addition to pathway information, Reactome Pathway Database offers analysis tools that enable users to perform data interpretation and visualize data in the context of pathways. This is particularly valuable for researchers studying complex biological systems.
6. **Updates and Collaboration:** Reactome Pathway Database is regularly updated to incorporate new findings and maintain the accuracy of pathway information. It also

encourages collaboration, allowing researchers to contribute their expertise and insights to improve pathway annotations.

Glycogenolysis:

Glycogenolysis is the enzymatic process of breakdown of glycogen in liver and muscles. Insulin hormone inhibits glycogenolysis in liver cells and muscles and serves as anabolic hormone. The glycogen phosphorylase or also termed as phosphorylase mainly regulates the rate of glycogenolysis in liver and muscles. Phosphorylase exists in active state as (phosphorylase a) and inactive state as (phosphorylase b).

There are two separate pools of phosphorylases in muscles as well as in liver cells. Muscle phosphorylase and liver phosphorylase are encoded by separate genes but have common regulatory mechanisms. The phosphorylase b (inactive form) in liver and muscle is activated through its phosphorylation by addition of high energy phosphate group from ATP. The reaction is catalyzed by phosphorylase kinase enzyme resulting in the formation of phosphorylase a (active form). The activation of phosphorylase b is regulated by glucagon in liver cells and adrenaline in muscles. The hormones attach to the cell membrane receptors on-target cell and activate the adenylate cyclase enzyme which in turn converts ATP into cAMP leading to rise in cytosolic levels of cAMP (2nd messenger in cell). Further, protein kinase A is phosphorylated and activated due to rise in cAMP and activated PKA in turn activates phosphorylase kinase leading to phosphorylation and activation of phosphorylase b into phosphorylase a. it catalyzes the rate-limiting step in glycogenolysis.

METHODOLOGY:

1. Go to the website of Reactome pathway database.
2. Search Glycogenolysis pathway in the search option bar.
3. Hits are obtained. Filter the results based on specific criteria such as species, datatype, disease, cellular compartment, etc.
4. Click on pathways for detailed information and external links for results.
5. Interpret the results.

OBSERVATIONS:

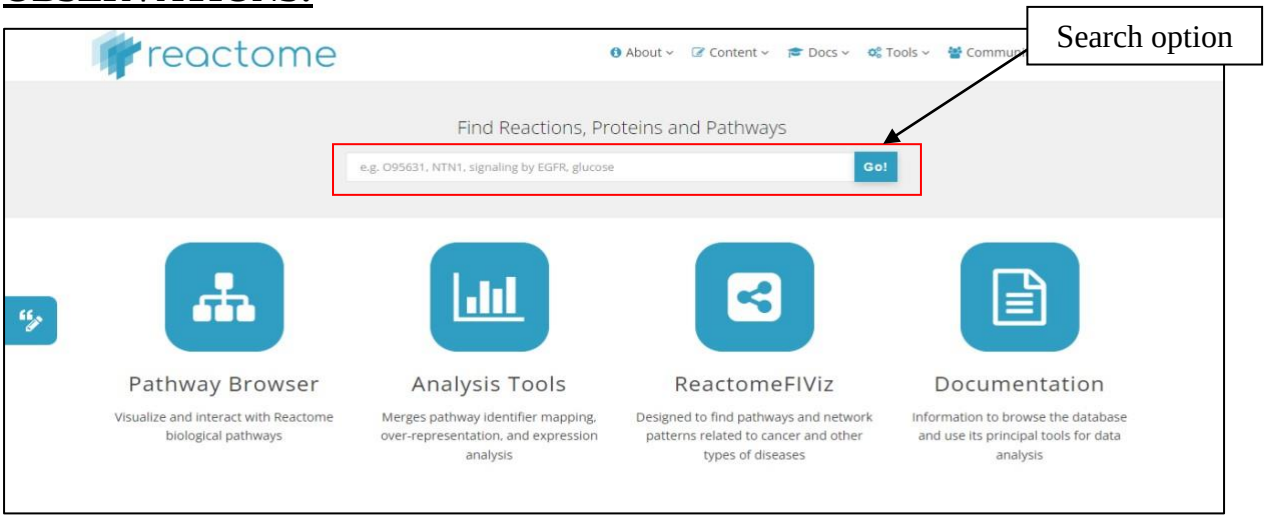


Figure 1: Homepage of Reactome pathway

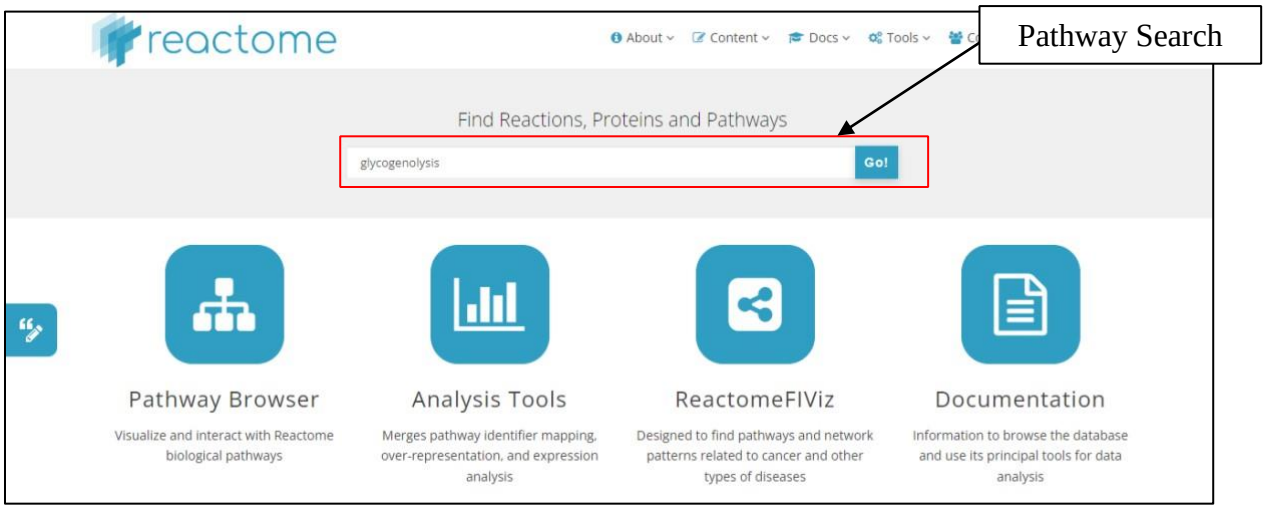


Figure 2: Pathway search in the search bar of Reactome database

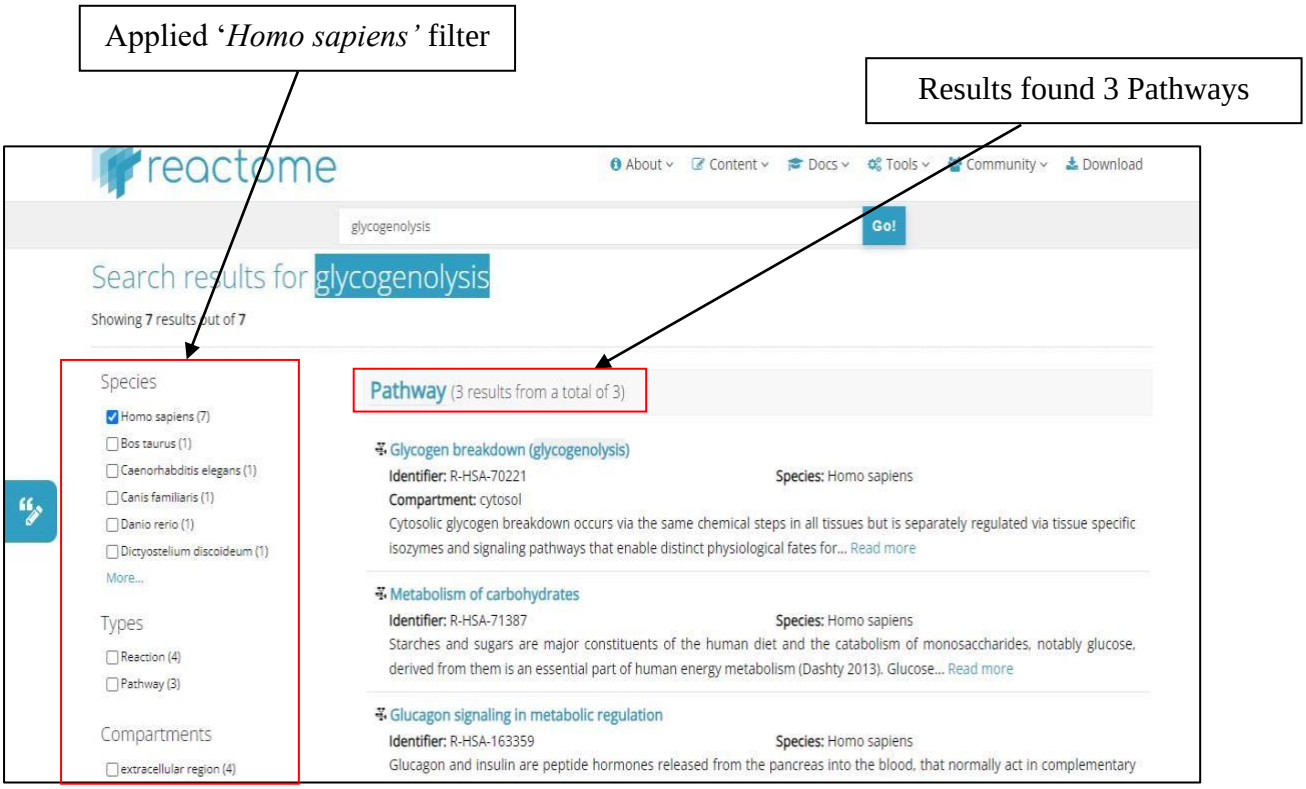


Figure 3: Pathways and filters available in Reactome Database

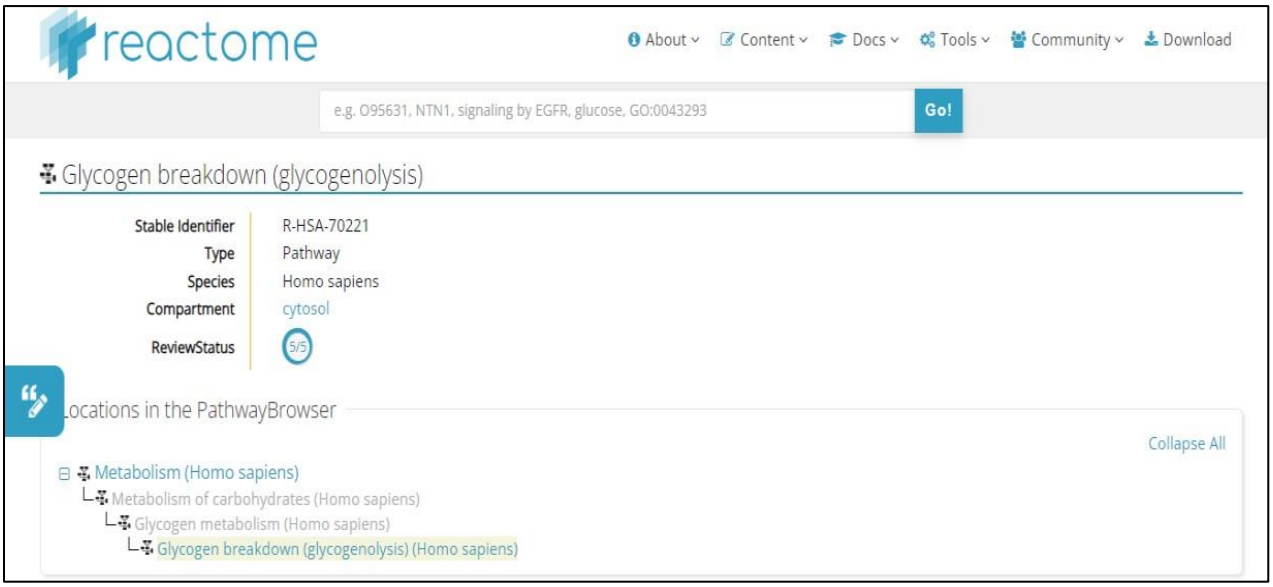


Figure 4: Entries pattern of Reactome database

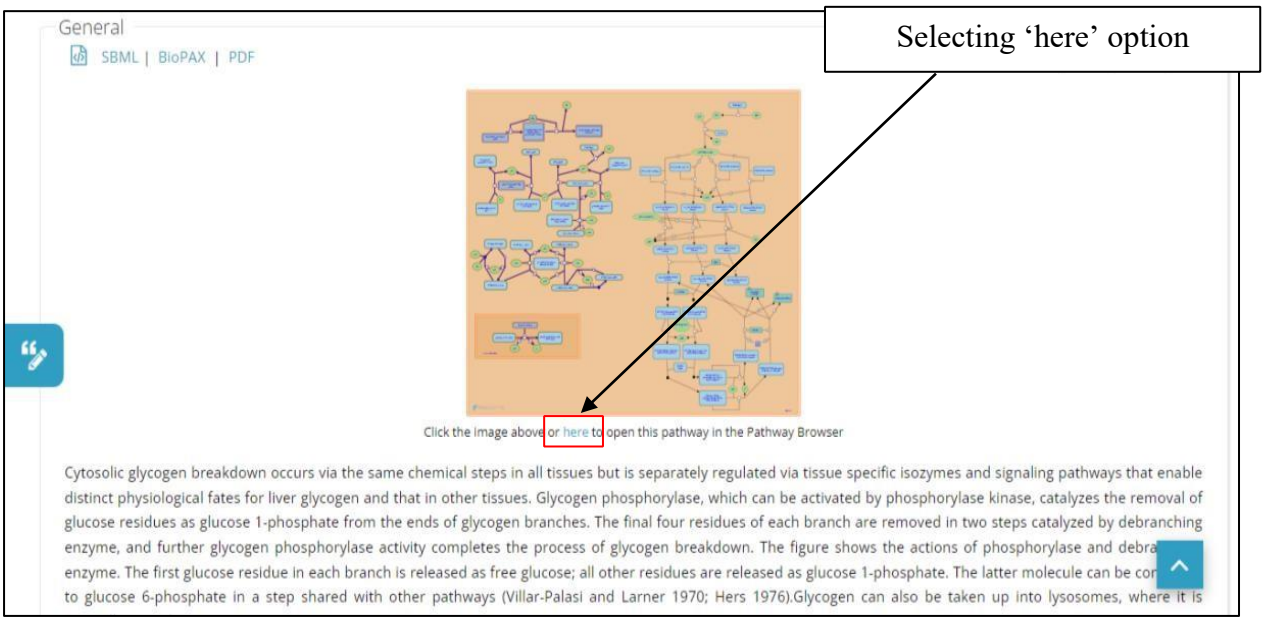


Figure 4a: Pathway image which is open in pathway browser

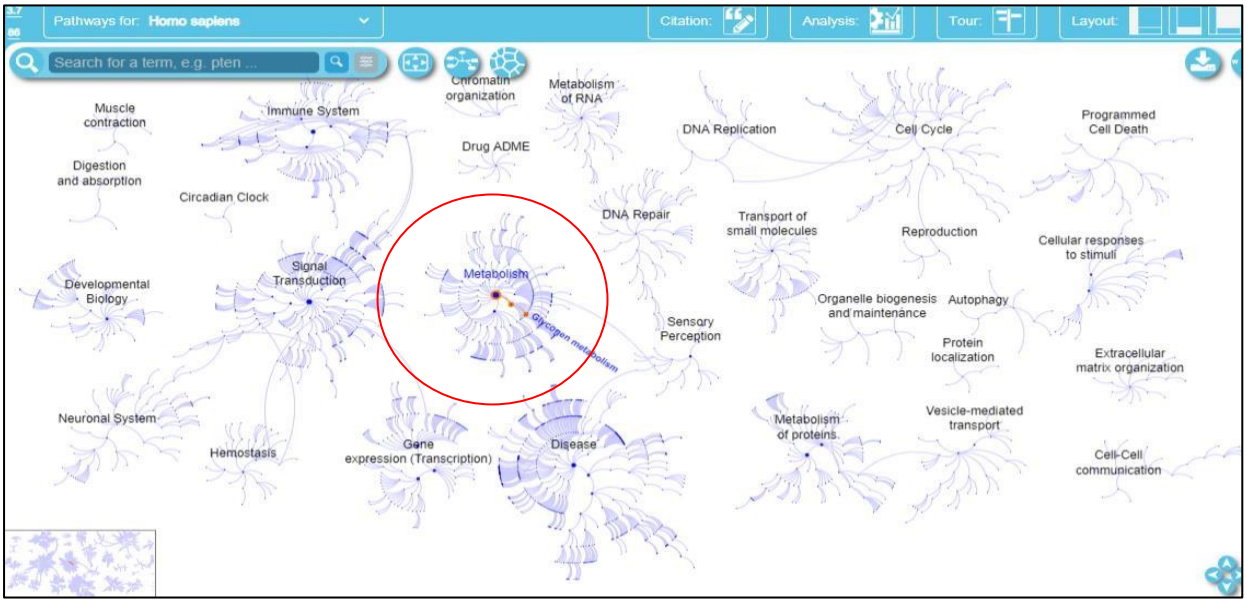
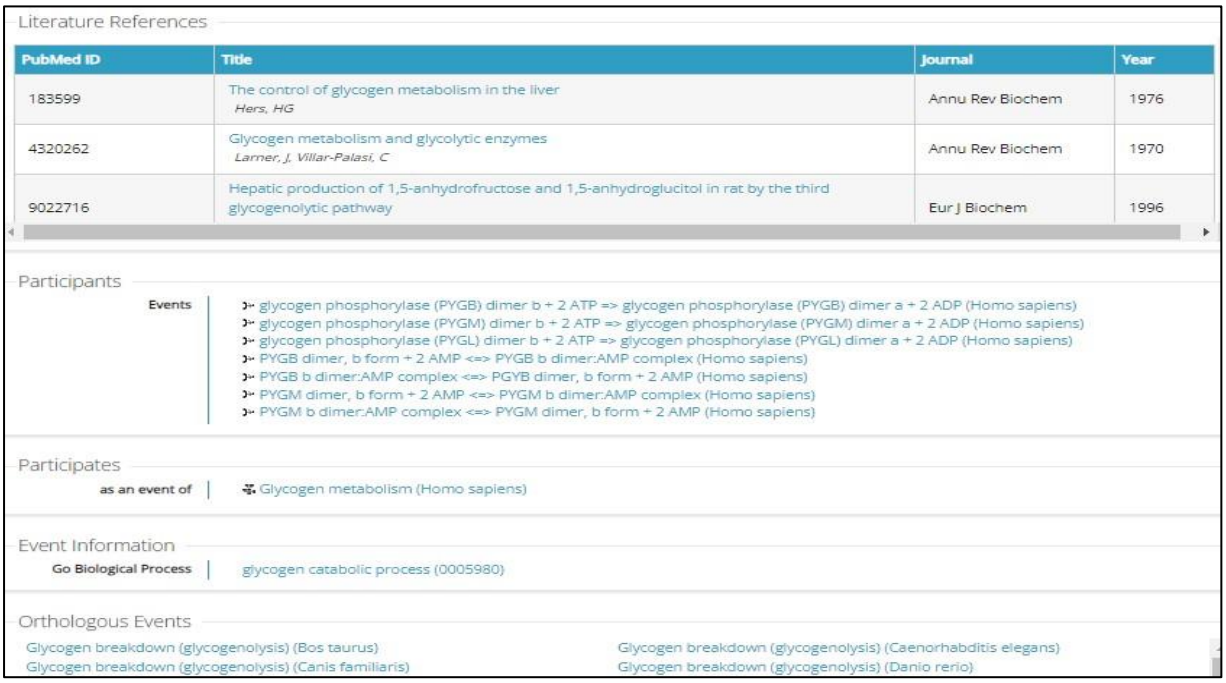


Figure 4b: Genome-wide, hierarchical visualization of glycogenolysis pathway



Literature References

PubMed ID	Title	Journal	Year
183599	The control of glycogen metabolism in the liver *Hers, HG*	Annu Rev Biochem	1976
4320262	Glycogen metabolism and glycolytic enzymes *Larner, J, Villar-Palasi, C*	Annu Rev Biochem	1970
9022716	Hepatic production of 1,5-anhydrofructose and 1,5-anhydroglucitol in rat by the third glycogenolytic pathway	Eur J Biochem	1996

Participants

Events
- glycogen phosphorylase (PYGB) dimer b + 2 ATP => glycogen phosphorylase (PYGB) dimer a + 2 ADP (Homo sapiens)
- glycogen phosphorylase (PYGM) dimer b + 2 ATP => glycogen phosphorylase (PYGM) dimer a + 2 ADP (Homo sapiens)
- glycogen phosphorylase (PYGL) dimer b + 2 ATP => glycogen phosphorylase (PYGL) dimer a + 2 ADP (Homo sapiens)
- PYGB dimer, b form + 2 AMP <=> PYGB b dimer:AMP complex (Homo sapiens)
- PYGB b dimer:AMP complex <=> PGYB dimer, b form + 2 AMP (Homo sapiens)
- PYGM dimer, b form + 2 AMP <=> PYGM b dimer:AMP complex (Homo sapiens)
- PYGM b dimer:AMP complex <=> PYGM dimer, b form + 2 AMP (Homo sapiens)

Participates

as an event of: Glycogen metabolism (Homo sapiens)

Event Information

Go Biological Process: glycogen catabolic process (0005980)

Orthologous Events

Glycogen breakdown (glycogenolysis) (Bos taurus)
Glycogen breakdown (glycogenolysis) (Canis familiaris)
Glycogen breakdown (glycogenolysis) (Caenorhabditis elegans)
Glycogen breakdown (glycogenolysis) (Danio rerio)

Figure 5: Link for literature and cross references

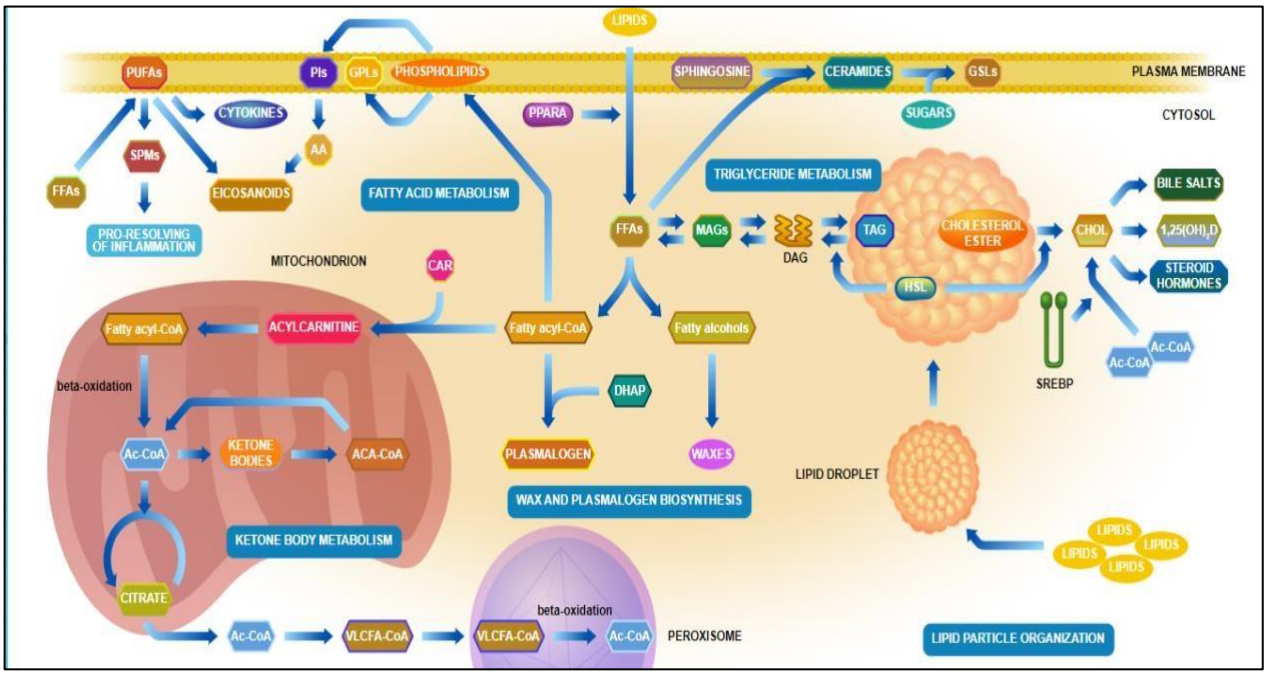


Figure 6: Illustration of Glycogenolysis pathway

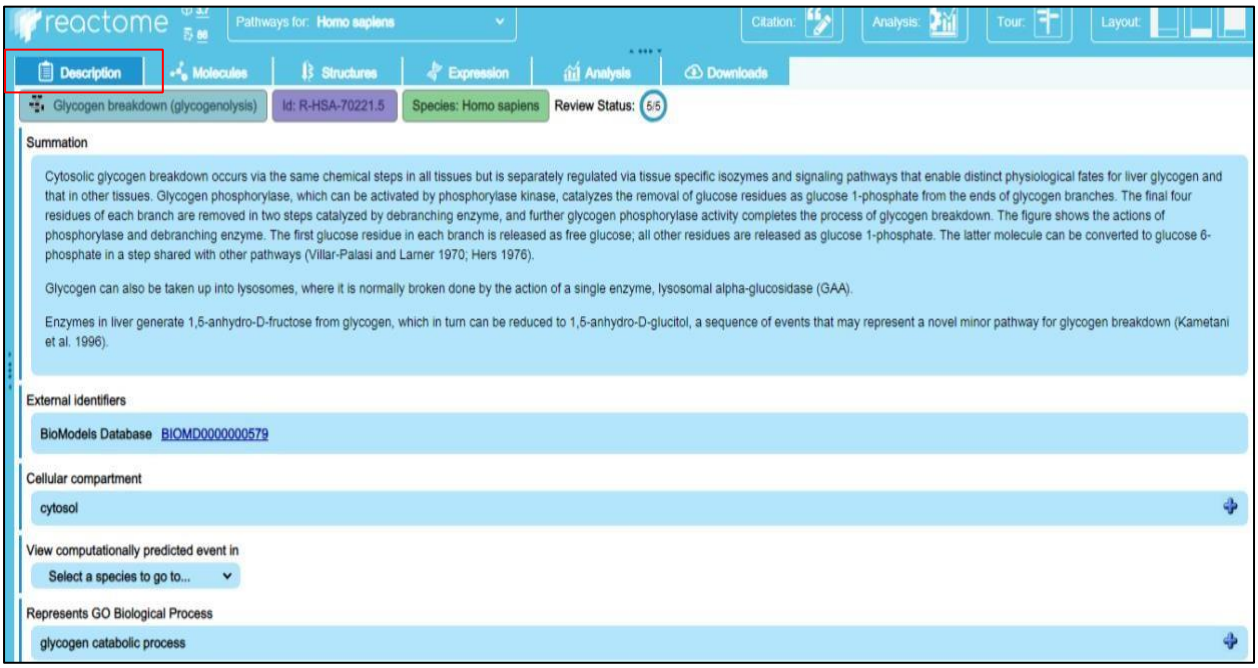


Figure 7: Description of the Glycogenolysis pathway

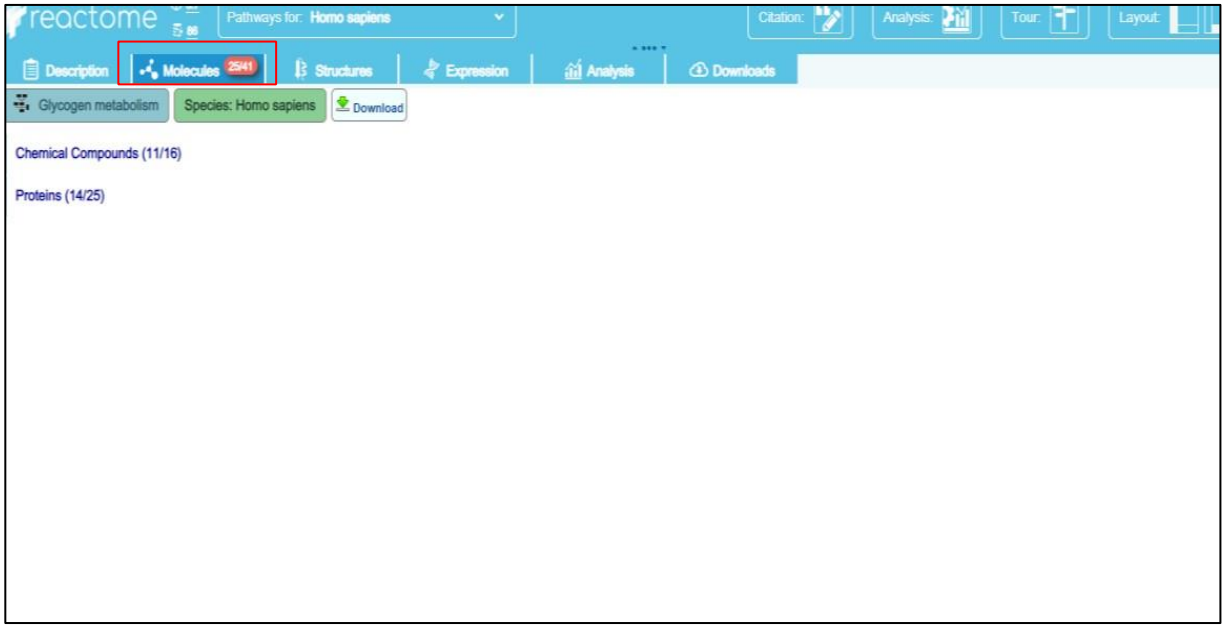


Figure 8: Detail of molecules involved in Glycogenolysis pathway

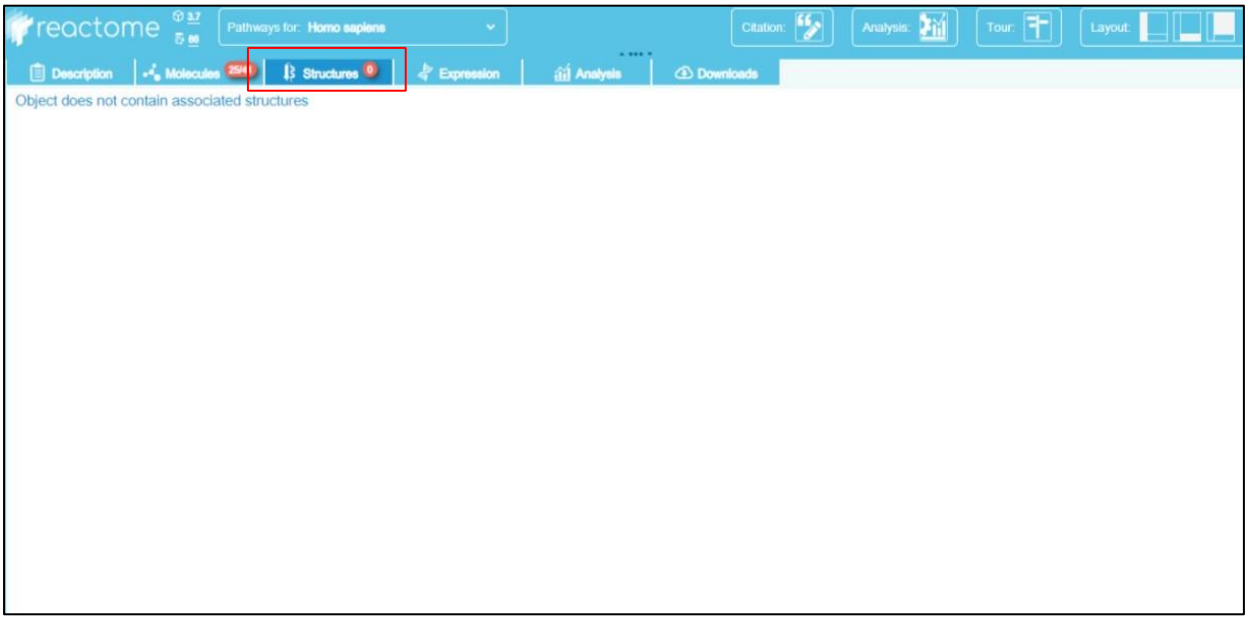


Figure 9: Structures related to the pathway

Figure 10: Expression of the glycogenolysis pathway in *Homo sapiens*

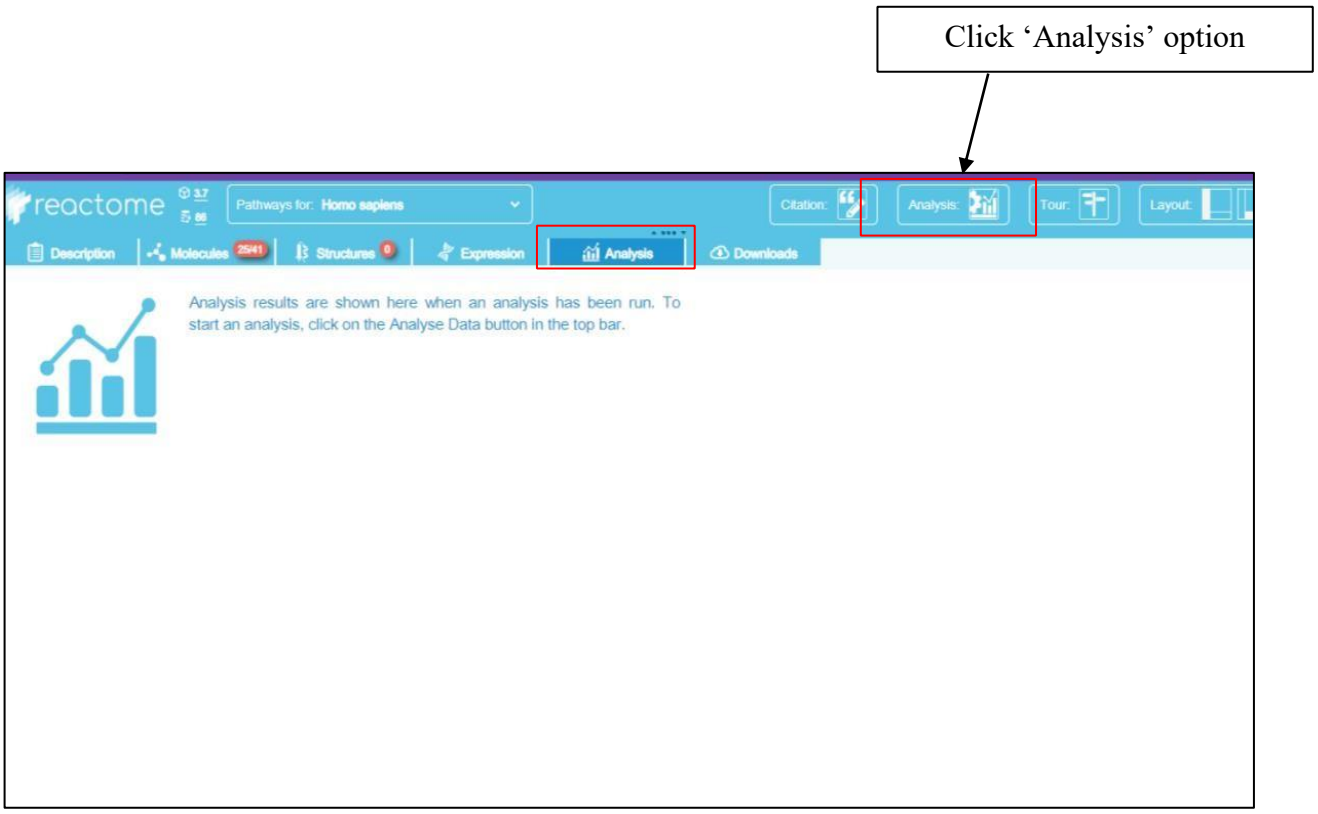


Figure 11: Analysis of the pathway

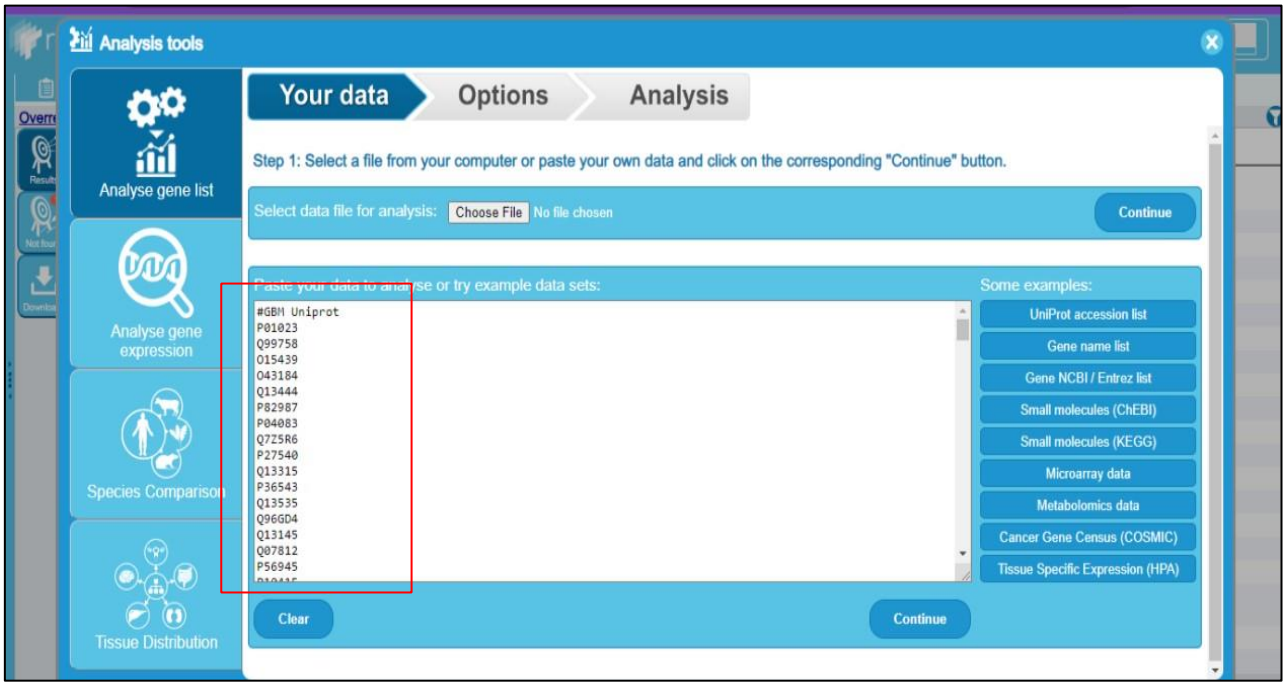
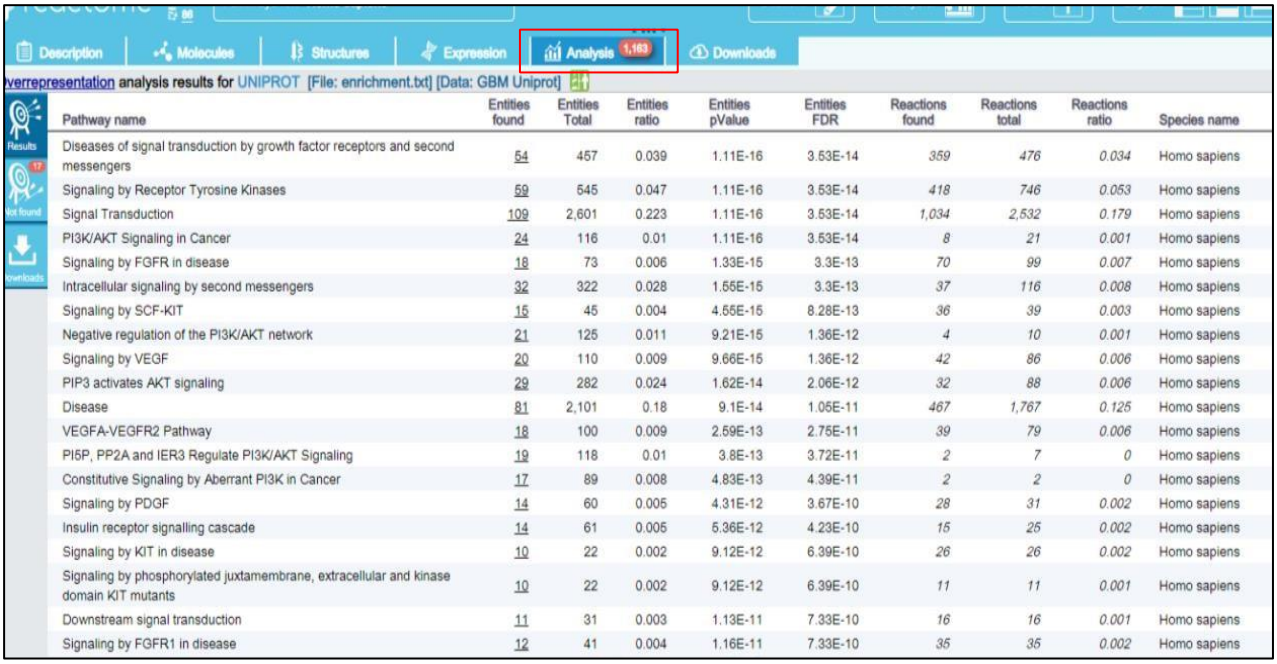


Figure 11a: Pasting the data for analysis



Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Species name
Diseases of signal transduction by growth factor receptors and second messengers	54	457	0.039	1.11E-16	3.53E-14	359	476	0.034	Homo sapiens
Signaling by Receptor Tyrosine Kinases	59	545	0.047	1.11E-16	3.53E-14	418	746	0.053	Homo sapiens
Signal Transduction	109	2,601	0.223	1.11E-16	3.53E-14	1,034	2,532	0.179	Homo sapiens
PI3K/AKT Signaling in Cancer	24	116	0.01	1.11E-16	3.53E-14	8	21	0.001	Homo sapiens
Signaling by FGFR in disease	18	73	0.006	1.33E-15	3.3E-13	70	99	0.007	Homo sapiens
Intracellular signaling by second messengers	32	322	0.028	1.55E-15	3.3E-13	37	116	0.008	Homo sapiens
Signaling by SCF-KIT	15	45	0.004	4.55E-15	8.28E-13	36	39	0.003	Homo sapiens
Negative regulation of the PI3K/AKT network	21	125	0.011	9.21E-15	1.36E-12	4	10	0.001	Homo sapiens
Signaling by VEGF	20	110	0.009	9.66E-15	1.36E-12	42	86	0.006	Homo sapiens
PIP3 activates AKT signaling	29	282	0.024	1.62E-14	2.06E-12	32	88	0.006	Homo sapiens
Disease	81	2,101	0.18	9.1E-14	1.05E-11	467	1,767	0.125	Homo sapiens
VEGFA-VEGFR2 Pathway	18	100	0.009	2.59E-13	2.75E-11	39	79	0.006	Homo sapiens
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	19	118	0.01	3.8E-13	3.72E-11	2	7	0	Homo sapiens
Constitutive Signaling by Aberrant PI3K in Cancer	17	89	0.008	4.83E-13	4.39E-11	2	2	0	Homo sapiens
Signaling by PDGF	14	60	0.005	4.31E-12	3.67E-10	28	31	0.002	Homo sapiens
Insulin receptor signalling cascade	14	61	0.005	5.36E-12	4.23E-10	15	25	0.002	Homo sapiens
Signaling by KIT in disease	10	22	0.002	9.12E-12	6.39E-10	26	26	0.002	Homo sapiens
Signaling by phosphorylated juxtamembrane, extracellular and kinase domain KIT mutants	10	22	0.002	9.12E-12	6.39E-10	11	11	0.001	Homo sapiens
Downstream signal transduction	11	31	0.003	1.13E-11	7.33E-10	16	16	0.001	Homo sapiens
Signaling by FGFR1 in disease	12	41	0.004	1.16E-11	7.33E-10	35	35	0.002	Homo sapiens

Figure 11b: Display of various pathway names after providing data for analysis

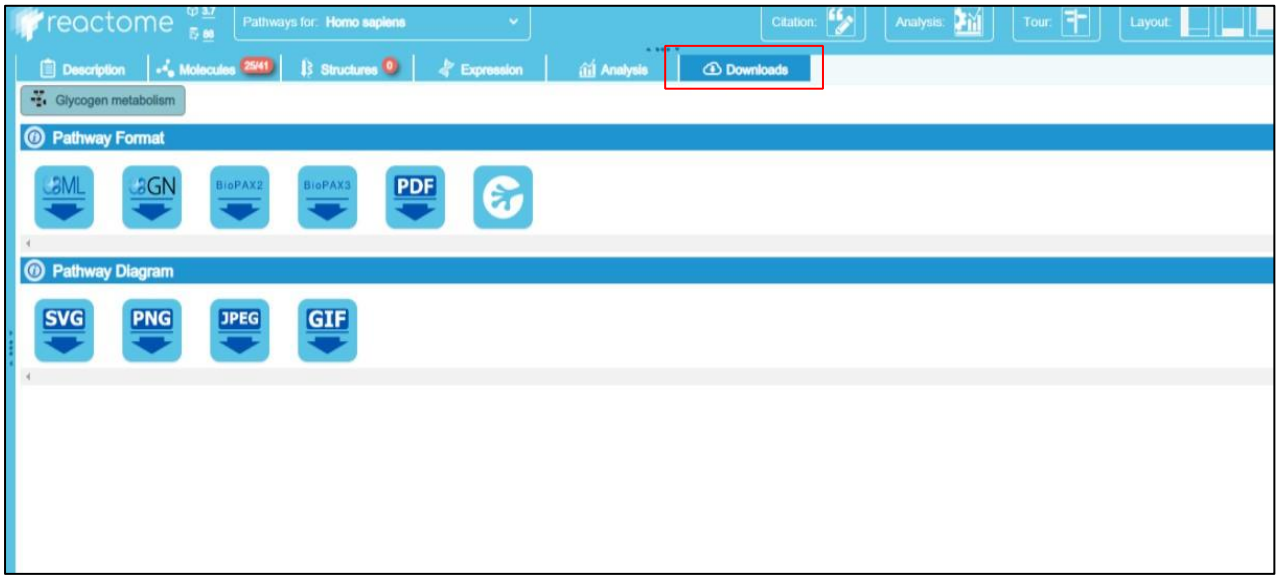


Figure 12: Options for downloading the pathway in various file formats

RESULTS:

Glycogenolysis pathway searched on Reactome pathway database, 3 pathways were found.

CONCLUSION:

The Reactome pathway database serves as a valuable resource for understanding biological pathways and processes. By leveraging Reactome, researchers can gain insights into complex molecular interactions, aiding in the interpretation of experimental data and facilitating a deeper understanding of cellular functions. However, the effectiveness of utilizing Reactome depends on the specific research goals and the comprehensiveness of the pathway data for the organism or system of interest. Regular updates and community contributions enhance its utility, making Reactome a powerful tool for exploring and analysing biological pathways in various contexts.

REFERENCES:

1. Home - Reactome Pathway Database. (n.d.). https://reactome.org/
2. Hayes, J. M. (2017, January 1). *Computer-Aided Discovery of Glycogen Phosphorylase Inhibitors Exploiting Natural Products*. Elsevier eBooks. https://doi.org/10.1016/b978-012-809450-1.00002-8
3. User Guide - Reactome Pathway Database. (n.d.). https://reactome.org/mzbxf/39#:~:text=Reactome%20is%20a%20free%2C%20open,mod eling%2C%20systems%20biology%20and%20education

DATE: 02/11/2023

WEBLEM 5(E)
PDBSUM DATABASE
(URL: https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/)

AIM:
To explore and study the structure of the protein 'Tubulin' (PDB ID: 1TUB) under various categories using the structural database of PDBSum.

INTRODUCTION:
PDBsum is an online database that provides a largely graphical overview of all the important data for every macromolecular structure that has been deposited at the Protein Data Bank (PDB) database. It was created in 1995 by Roman Laskowski and collaborators at University College London. The unique feature of PDBsum is that it is a pictorial database that provides a brief synopsis of all the 3D structures stored in the PDB database. PDBSum database comprises structural pictures, annotated secondary structure plots for each protein chain, PROMOTIF structural analyses, PROCHECK results summarized, and schematic diagrams exhibiting protein-ligand and protein-DNA interactions.
The principal aim of PDBsum database is to provide the researchers from all around the world with a concise and summarized data of the compounds present in a specific protein entry of interest from the PDB Database, comprising small-molecule ligands, metal ions, proteins, and DNA/RNA strands. It also provides analyses and annotations of important structural aspects. The molecule's thumbnail image is displayed first, followed by the structural data in a uniform format. PDBsum database contains comprehensive data about the protein, including linkages, clefts, pores, and protein-protein interactions. Diagrams that show how these parts interact are called schematic diagrams.

Tubulin:
Living cells' primary microtubule component is tubulin, a class of proteins essential to the eukaryotic cytoskeleton. These proteins combine to form lengthy chains and filaments that form hollow fibers that serve as the cell's skeleton. Six proteins are known to belong to the tubulin superfamily; of these, five subgroups are found in humans: α-Tubulin, β-Tubulin, γ-Tubulin, δ and ε-Tubulin, and ζ-Tubulin. β-Tubulin, which forms microtubules solely in neurons, is of particular importance. The importance of tubulin in creating medicinal medications is still being studied, particularly with regard to anticancer therapies. A number of currently available anticancer medications target tubulins, including vinblastine, vincristine, and paclitaxel. PDBsum is one of the databases that can be used to investigate the 3D structure of Tubulin to conduct additional research on this protein.

METHODOLOGY:

1. Open PDB database and search for the query of 'Tubulin'.
2. From the results page, open the protein of interest and copy its PDB ID (1TUB).
3. Open PDBsum database and paste the copied PDB ID of Tubulin in the 'Search with ID' box and then click on 'Find'.
4. Interpret the results displayed for the PDB ID: 1TUB on PDBSum database.

OBSERVATIONS:

Figure 1: Homepage of PDB Database (Protein Data bank)

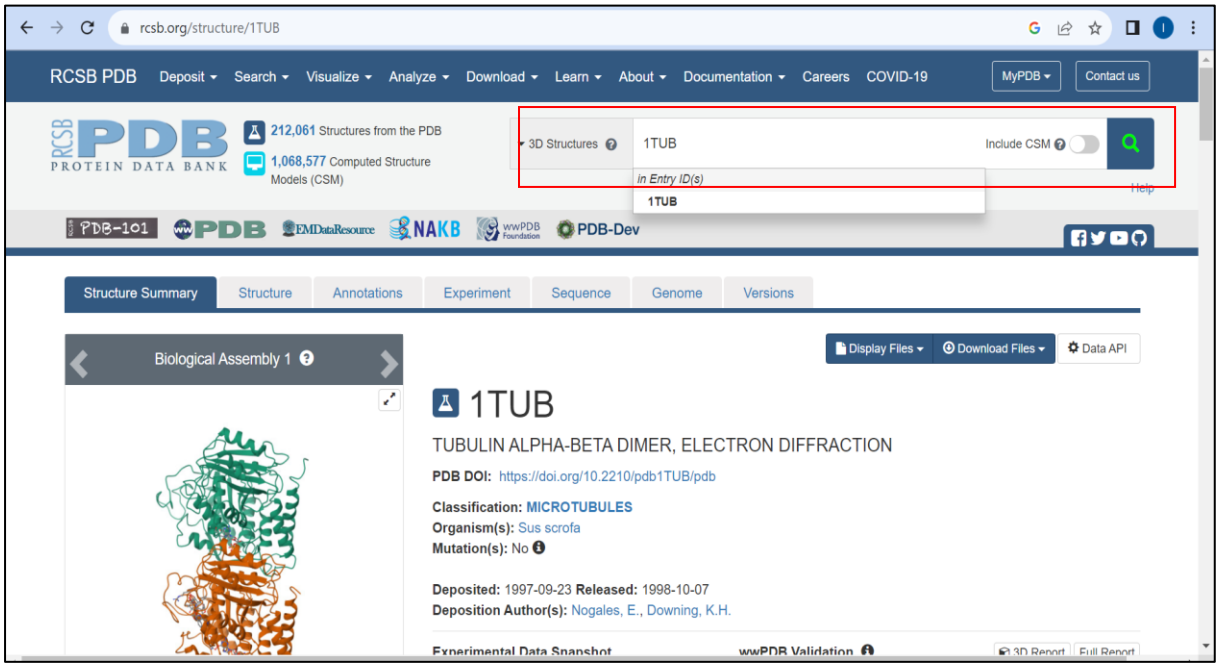


Figure 2: PDB Database search for the query 'Tubulin' (PDB ID: 1TUB)

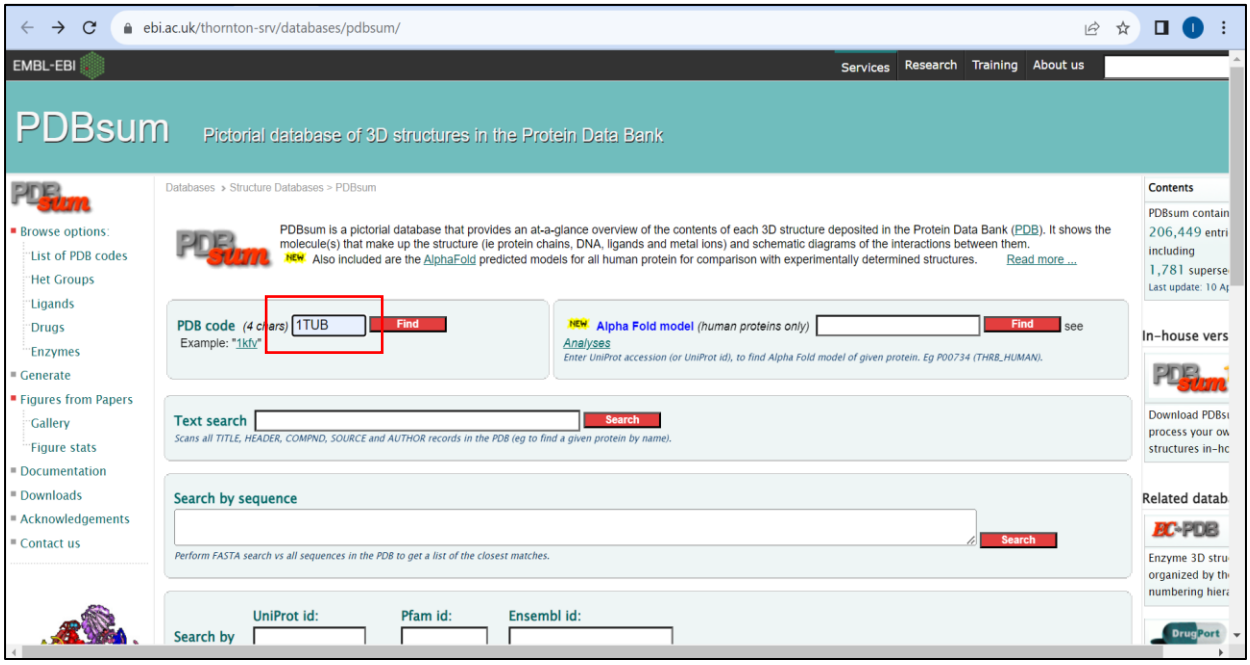


Figure 3: Homepage of PDBsum database query searched, PDB ID: 1TUB

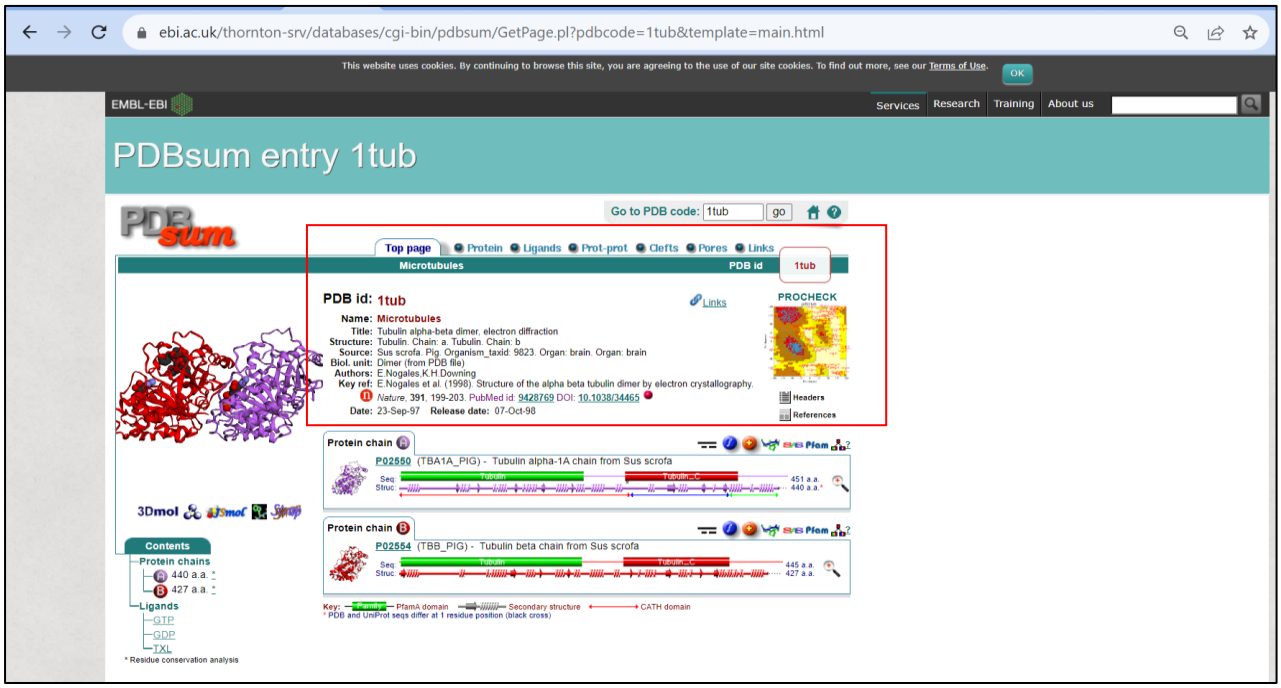


Figure 4: Description about the query, 1TUB (Title, Structure, scores and chains present in the protein structure)

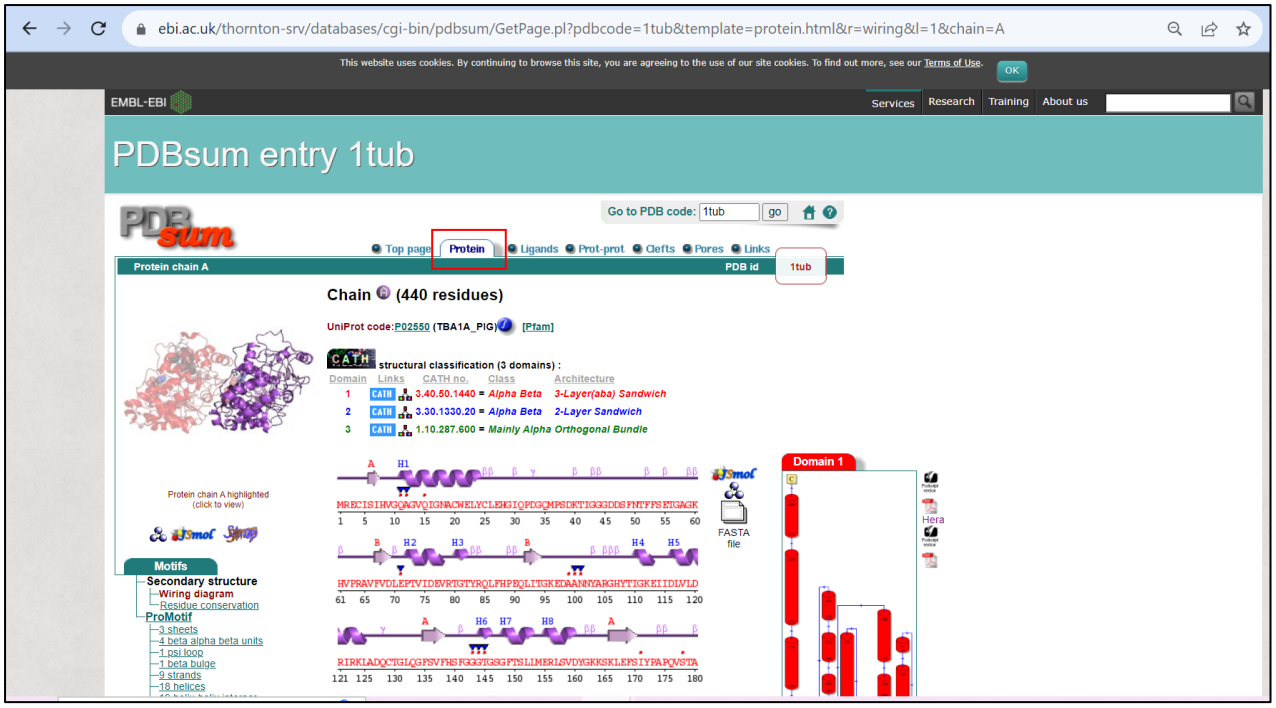


Figure 5: Secondary structure, Domains, Chains related information of the query '1TUB' in the section 'Protein'

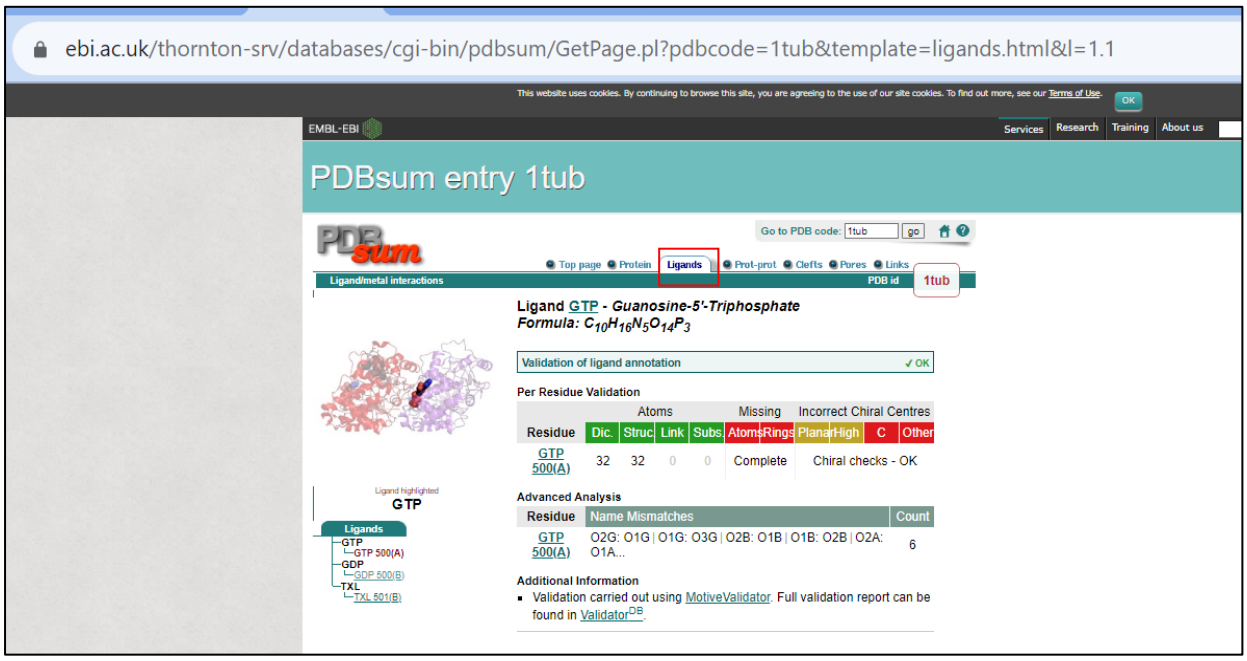


Figure 6: Information about the ligands (GTP, GDP, TXL) sites where the drugs can bind present in the protein structure '1TUB'

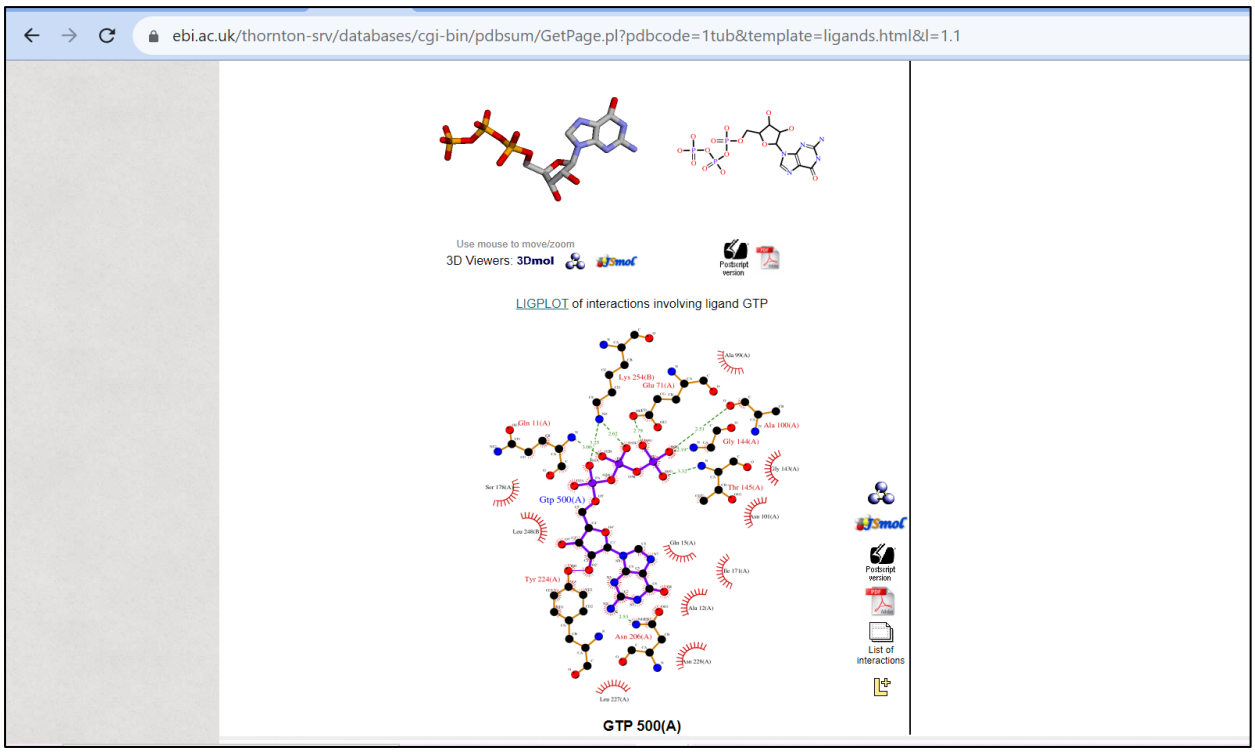


Figure 6a: The LIGPLOT structure of the ligand GTP

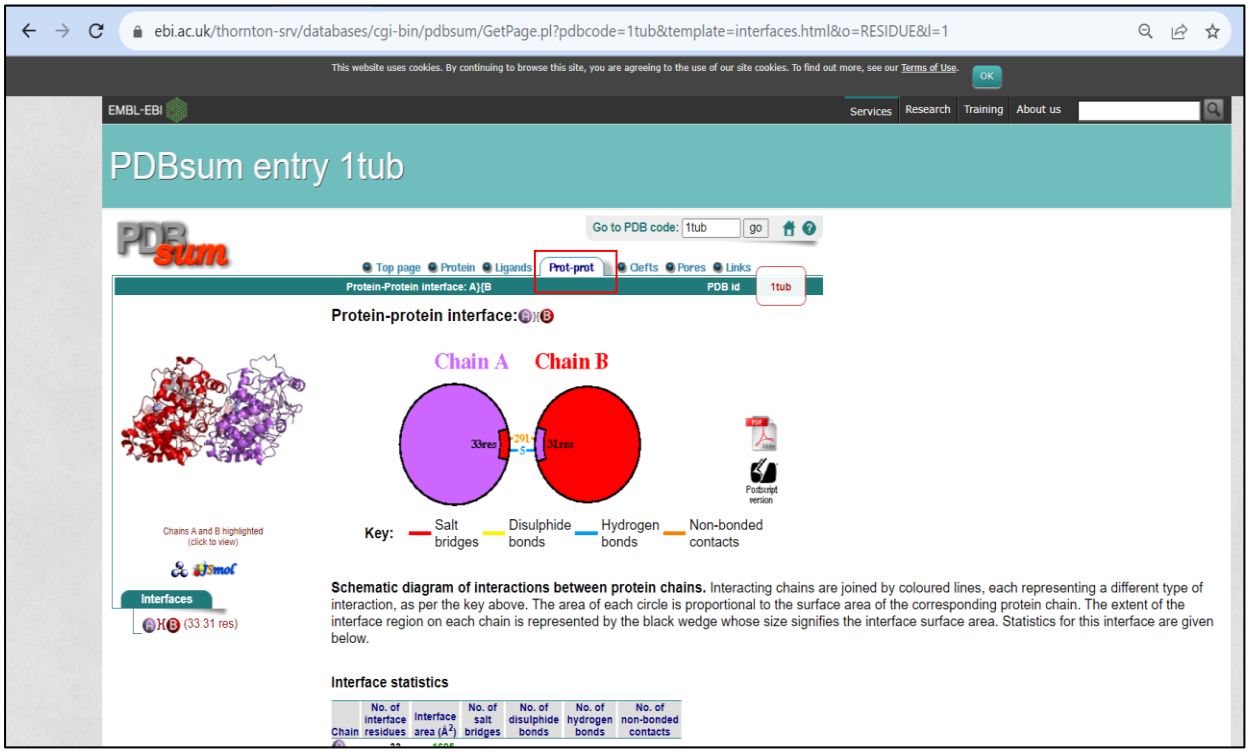


Figure 7: Diagrammatic information about the Protein-Protein Interactions and the bonds present in between them

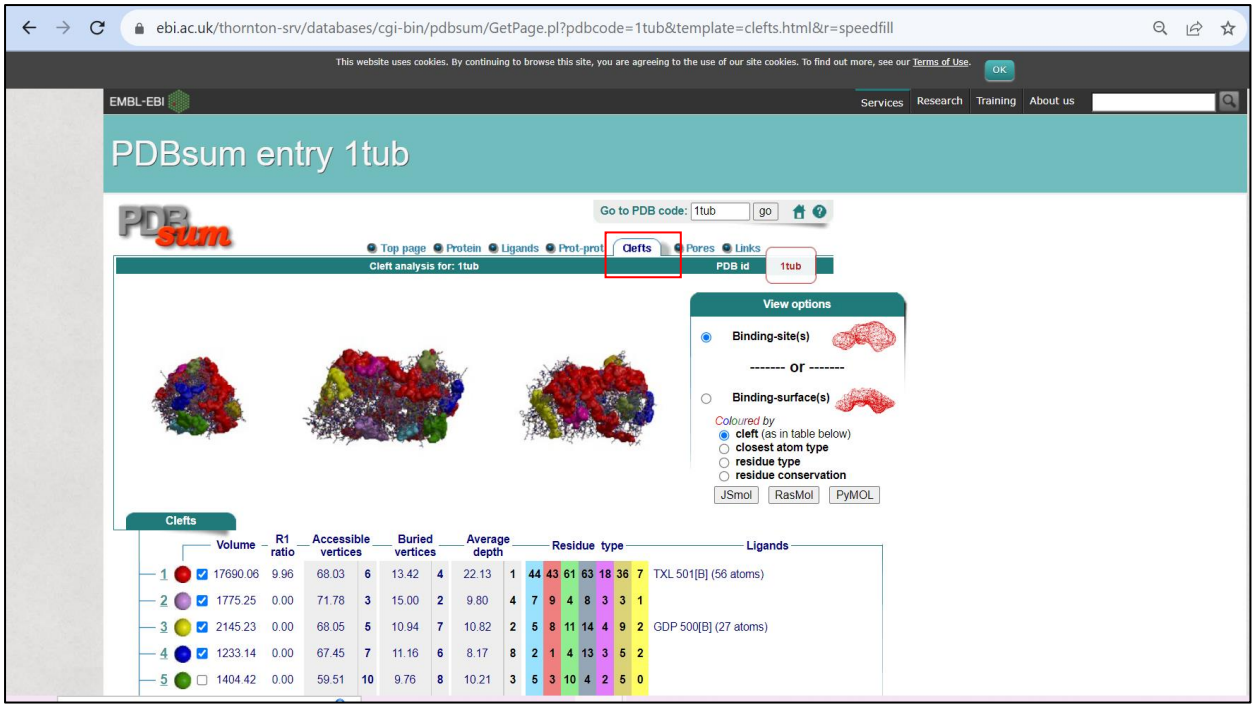


Figure 8: Diagrammatical information about the Clefts present in the Protein

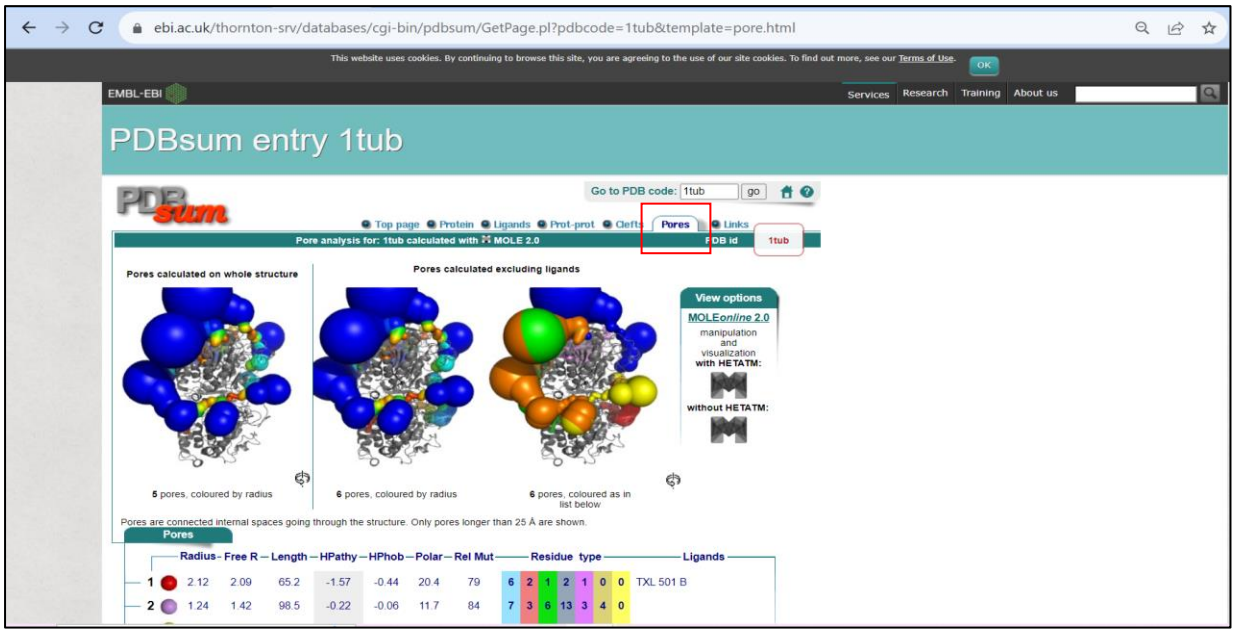


Figure 9: Structural representation and information of Pores present in the protein

RESULTS:

PDBsum database was explored and studied for the query 'Tubulin', (PDB ID: 1TUB) the 3D structure was studied and the results were observed in section like Ligands where in the drug binding sites present in the protein were studied, in protein- protein interaction the chains present in the protein were studied. The Clefts present in the protein were studied in Clefts. And then the Pores were studied which were represented in blue.

CONCLUSION:

PDBsum is a valuable resource in the realm of structural biology, offering detailed and comprehensive information about macromolecular structures deposited in the Protein Data Bank (PDB). It serves as a user-friendly platform that summarizes and presents crucial structural information, annotations, and analyses of protein structures available in the PDB. PDBsum serves as an indispensable resource for researchers in structural biology, providing comprehensive and user-friendly summaries of macromolecular structures. Its contribution to understanding protein structures and their functions is substantial, fostering advancements in various scientific disciplines and facilitating the development of novel therapies and treatments.

REFERENCES:

1. Katsetos, C. D., & Dráber, P. (2012). Tubulins as therapeutic targets in cancer: from bench to bedside. *Current pharmaceutical design*, *18*(19), 2778–2792. https://doi.org/10.2174/138161212800626193
2. Parker, A. L., Kavallaris, M., & McCarroll, J. A. (2014). Microtubules and their role in cellular stress in cancer. *Frontiers in oncology*, *4*, 153. https://doi.org/10.3389/fonc.2014.00153
3. Roman A. Laskowski, PDBsum new things, *Nucleic Acids Research*, Volume 37, Issue suppl_1, 1 January 2009, Pages D355–D359, https://doi.org/10.1093/nar/gkn860
4. Binarová, P., & Tuszynski, J. (2019). Tubulin: Structure, Functions and Roles in Disease. *Cells*, *8*(10), 1294. https://doi.org/10.3390/cells8101294

DATE: 02/11/2023

WEBLEM 5(F)
PROTEIN DATABANK OF TRANSMEMBRANE PROTEINS (PDBTM) DATABASE
(URL: http://pdbtm.enzim.hu)

AIM:
To explore the Protein Data Bank of Transmembrane Proteins (PDBTM) Database for studying transmembrane protein for the query of Aquaporin (PDB ID: 1RC2).

INTRODUCTION:
The Protein Databank of Transmembrane Proteins (PDBTM) Database is the first comprehensive and up-to-date transmembrane protein database based on the Protein Data Bank (PDB) Database and was launched in 2004. The database was created and has been continuously updated by the TMDET algorithm that is able to distinguish between the transmembrane and the non-transmembrane proteins using their 3D atomic coordinates, along with locating the spatial positions of transmembrane proteins in lipid bilayer as well. During the last 8 years not only the size of the PDBTM database has been steadily growing from ~400 to 1700 entries but also new structural elements have been identified, in addition to the well-known α-helical bundle and β-barrel structures.
Numerous 'exotic' transmembrane protein structures have been curated and annotated since the first release. This has necessitated the definition of new structural elements, such as membrane loops or interfacial helices, in the PDBTM database. The PDBTM database collects all transmembrane proteins for which structures have been curated, checks them, and, if necessary, corrects their biologically active oligomer form as given in PDB files. Additionally, it defines their membrane orientation and sets their transmembrane segments, membrane re-entrant loops, and interfacial helices (IFHs). The PDBTM database is updated every week. The update includes a combined process of automatic and manual steps.

TMDET Algorithm:
The TMDET algorithm detects candidate transmembrane proteins from PDB and defines its most likely position in the lipid bilayer membrane. Although the entire process is automatic, human intervention is necessary for the validation of the protein and its relative position in the lipid bilayer. The manual intervention is also necessary to rule out the false positives from the raw output, or to better reconstruct the membrane bilayer.

Aquaporin:
Aquaporins (AQPs) are a family of membrane water channel proteins that osmotically modulate water fluid homeostasis in several tissues; some of them also transport small solutes such as glycerol. At the cellular level, the AQPs regulate not only cell migration and transepithelial fluid transport across membranes, but also common events that are crucial for the inflammatory response. Emerging data reveal a new function of AQPs in the inflammatory

process, as demonstrated by their dysregulation in a wide range of inflammatory diseases including edematous states, cancer, obesity, wound healing, and several autoimmune diseases.

METHODOLOGY:

1. The PDB ID for the desired protein for the query of 'Aquaporin (1RC2)' was retrieved from Protein Data Bank (PDB) Database.
2. Open the Homepage of Protein Data Bank of Transmembrane Proteins (PDBTM) Database and paste the retrieved PDB ID in the query box of the PDBTM database.
3. On the results page, membrane localization, structure, and topology of the chains were observed.
4. The structure can be downloaded and observed in the various file formats, such as, XML, JSON, Transformed structure.
5. Interpret the results displayed for the PDB ID: 1RC2 on PDBTM Database.

OBSERVATIONS:

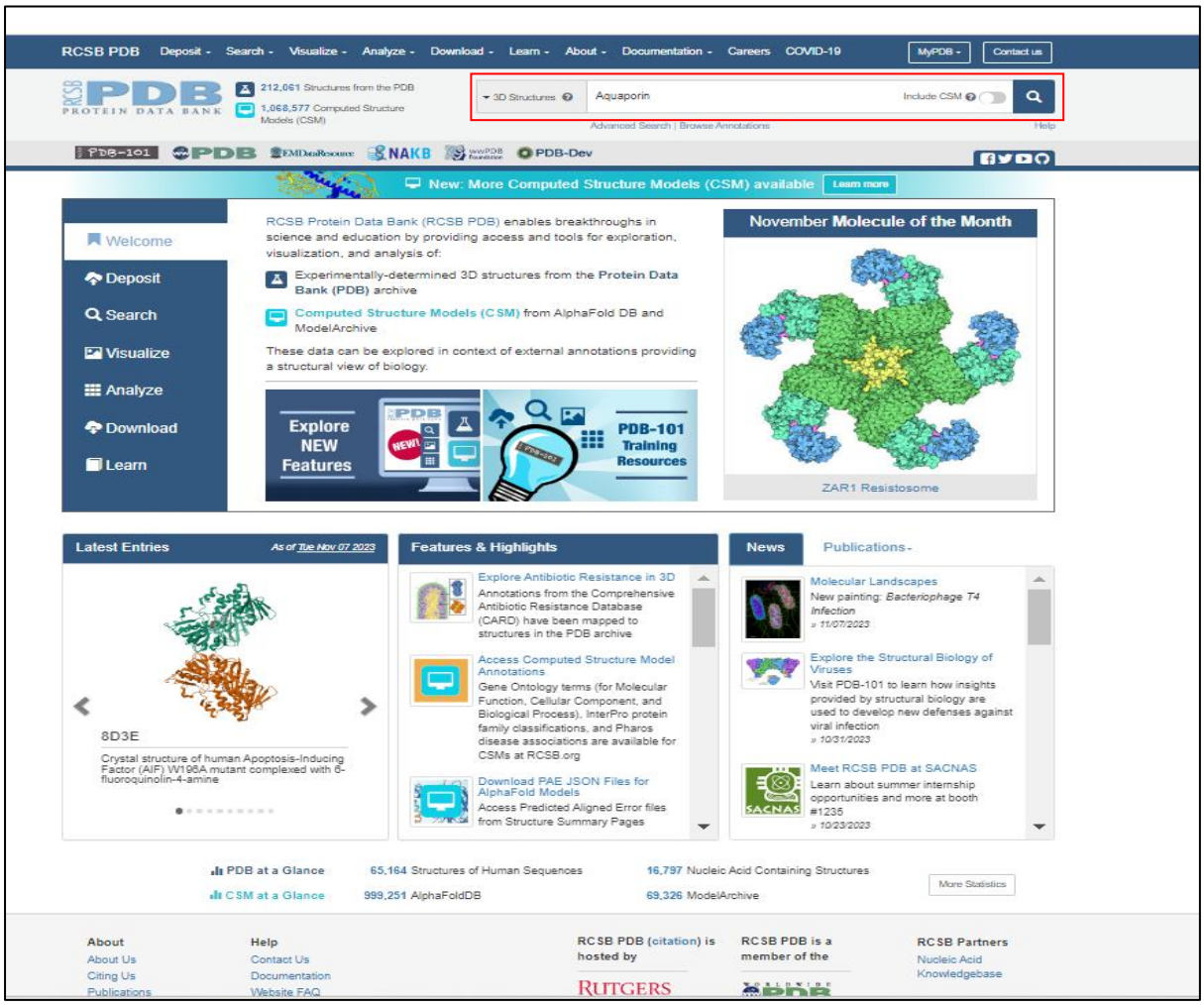


Figure 1: Homepage of the Protein Data Bank (PDB) Database with query of Aquaporin

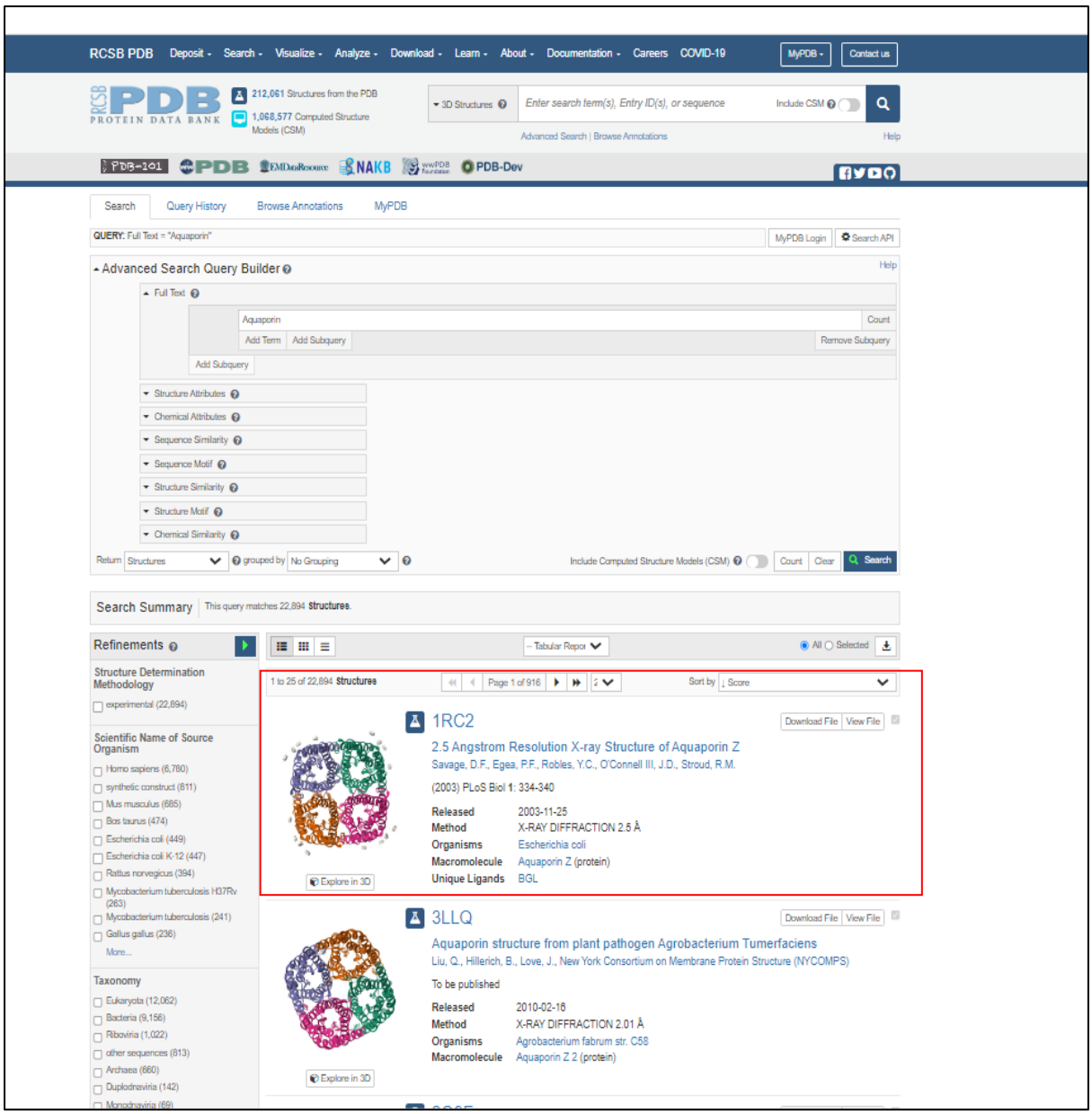


Figure 2: Results page of PDB Database with the desired query of Aquaporin (PDB ID: 1RC2)

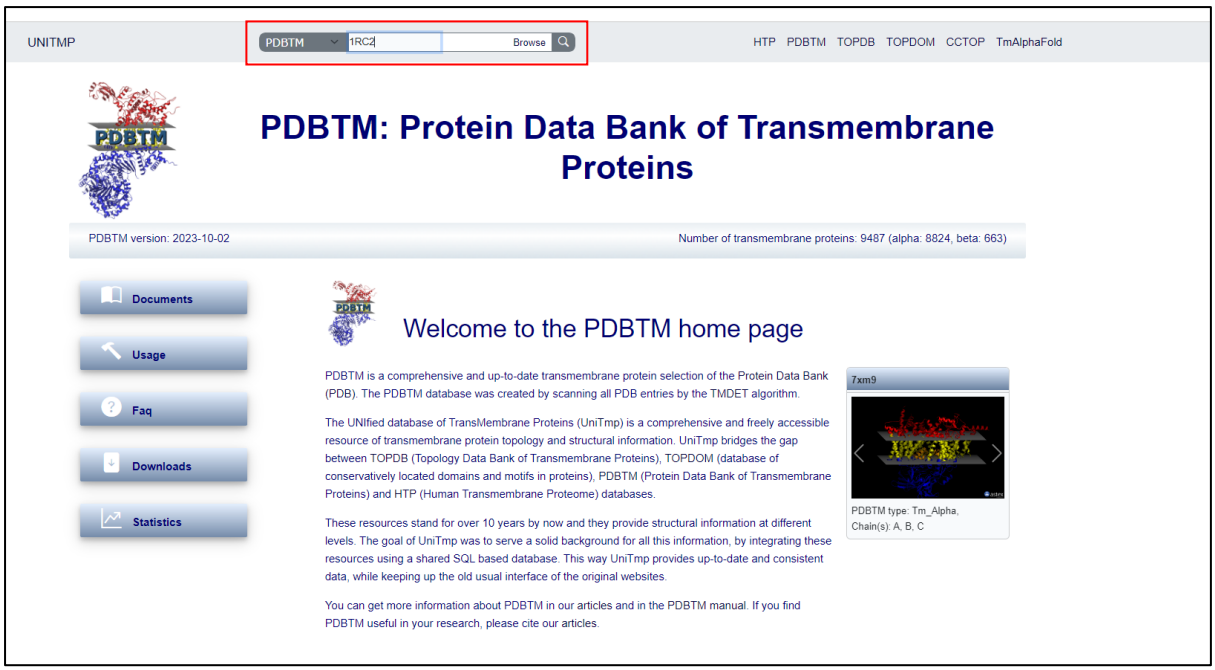


Figure 3: Homepage of PDBTM Database with the query of PDB ID: 1RC2

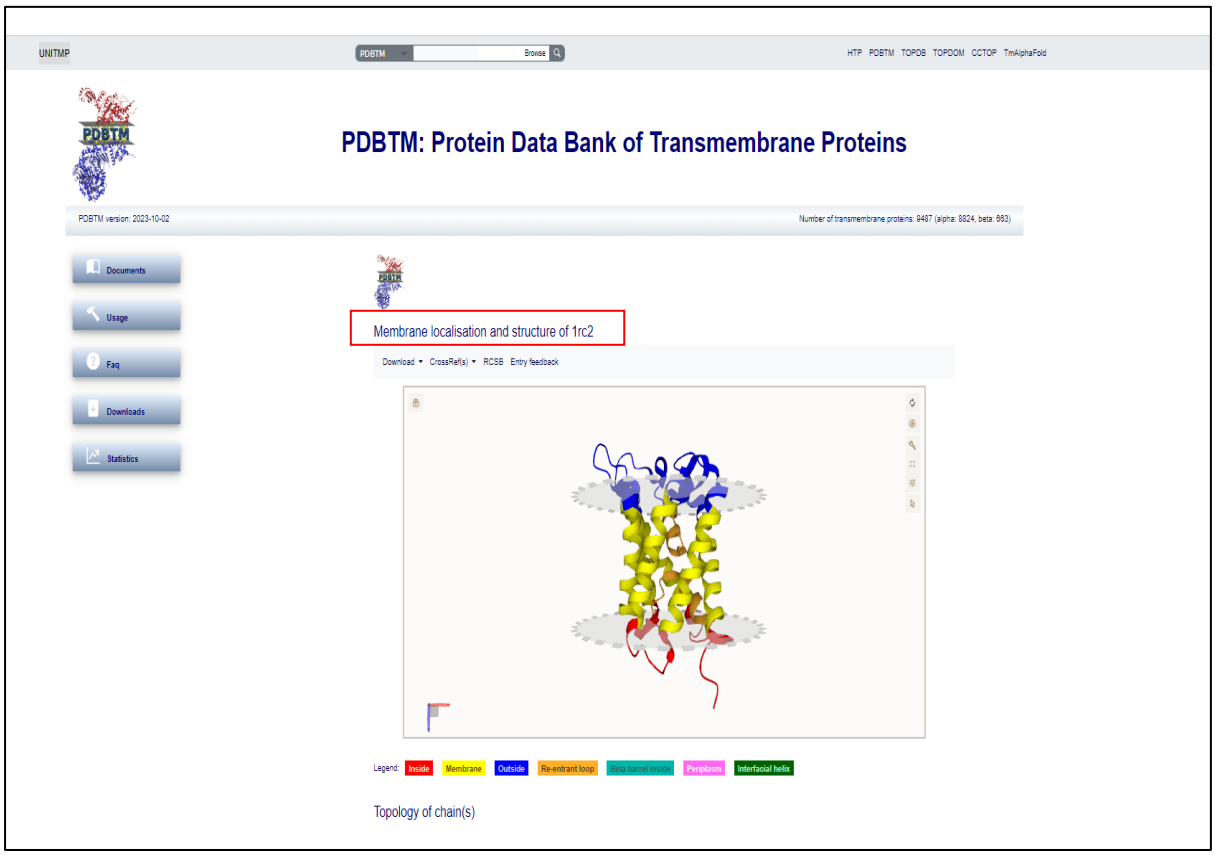


Figure 4: Membrane localization and structure of query aquaporin (PDB ID: 1RC2)

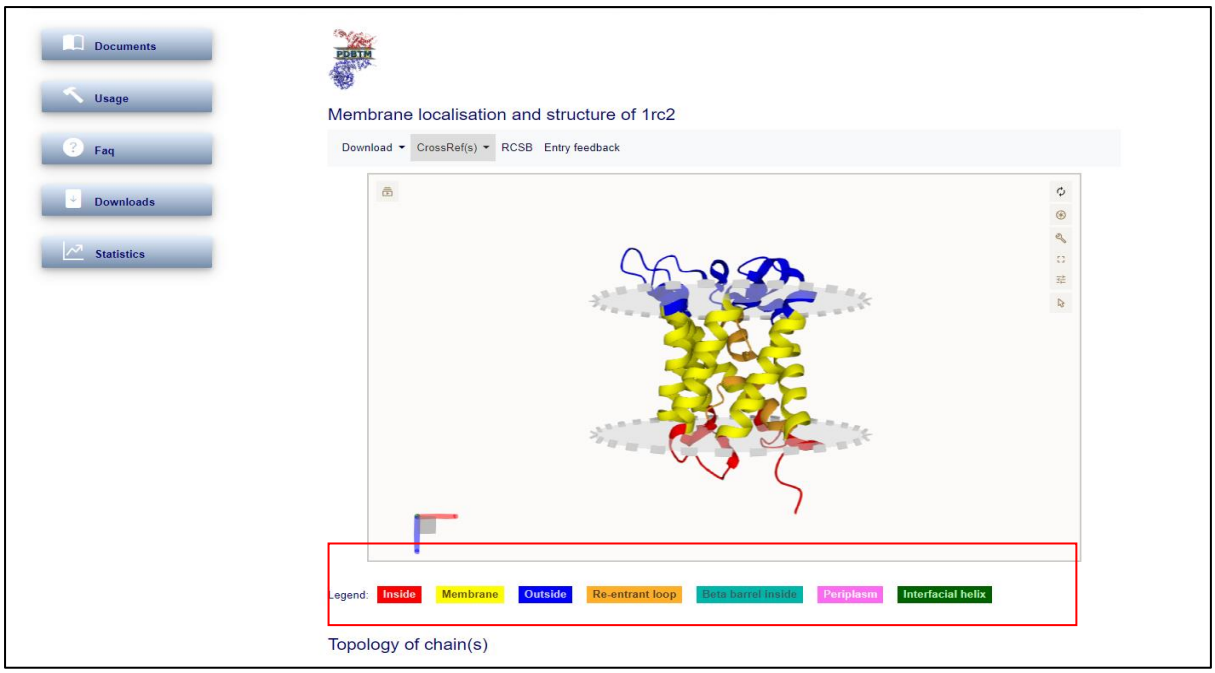


Figure 4a: Legend of query of 1RC2

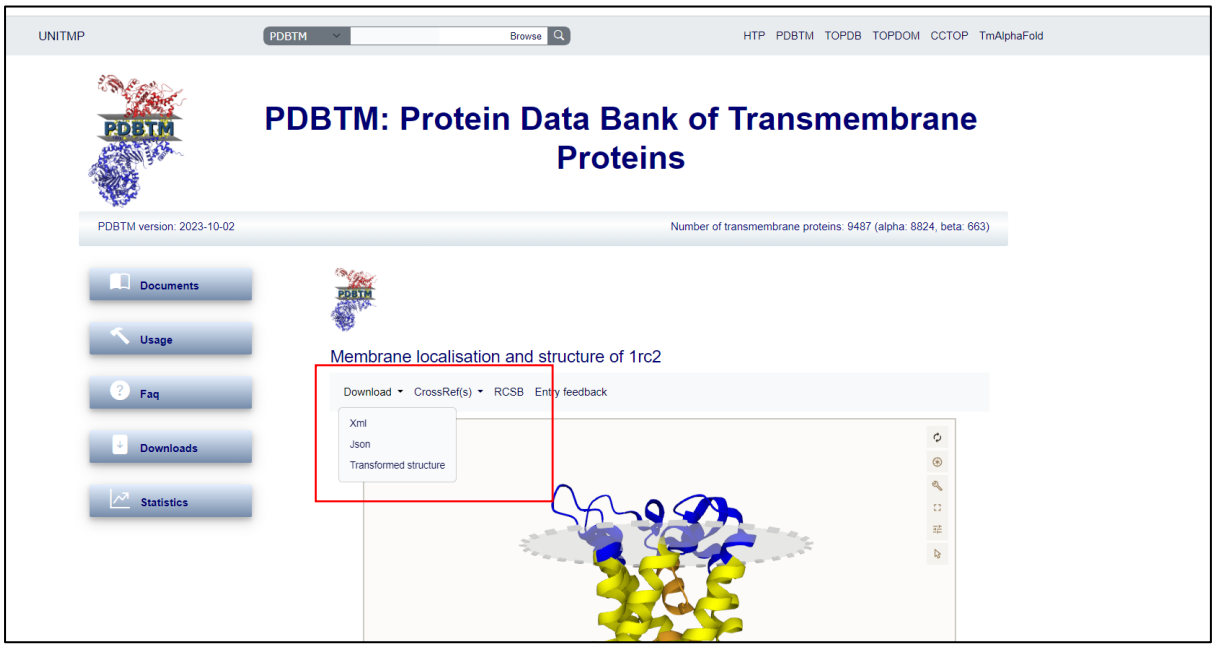


Figure 5: File formats for structure download

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
▼<pdbtm xmlns="http://pdbtm.enzim.hu" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://pdbtm.enzim.hu/data/pdbtm.xsd pdbtm.xsd" ID="1rc2" TMP="yes">
  <COPYRIGHT> All information, data and files are copyright. PDBTM database is produced in the Institute of Enzymology, Budapest, Hungary. There are no restrictions on its use by non-profit
  institutions as long as its content is in no way modified and this statement is not removed from entries. Usage by and for commercial entities requires a license agreement (send an email
  to pdbtm at enzim dot hu). </COPYRIGHT>
  <CREATE_DATE>2004-03-10</CREATE_DATE>
 ▼<MODIFICATION>
    <DATE>2005-04-06</DATE>
    <DESCR>Format has been changed to pdbtm format v2.0</DESCR>
  </MODIFICATION>
 ▼<MODIFICATION>
    <DATE>2005-07-22</DATE>
    <DESCR>Model has been recalculated by TMDET version 2.0</DESCR>
  </MODIFICATION>
 ▼<MODIFICATION>
    <DATE>2012-03-08</DATE>
    <DESCR>PDB entry has been changed</DESCR>
  </MODIFICATION>
 ▼<MODIFICATION>
    <DATE>2013-05-10</DATE>
    <DESCR>Sidedefinition added</DESCR>
  </MODIFICATION>
 ▼<RAWRES>
    <TMRES>71.74</TMRES>
    <TMTYPE>Tm_Alpha</TMTYPE>
    <SPRES>Unknown</SPRES>
    <PDBKWRES>yes</PDBKWRES>
  </RAWRES>
 ▼<BIOMATRIX>
  ▼<MATRIX ID="1">
      <APPLY_TO_CHAIN CHAINID="B" NEW_CHAINID="C"/>
    ▼<TMATRIX>
        <ROWX X="-1.00000000" Y="0.00000000" Z="0.00000000" T="-93.55000305"/>
        <ROWY X="0.00000000" Y="-1.00000000" Z="0.00000000" T="93.55000305"/>
        <ROWZ X="0.00000000" Y="0.00000000" Z="1.00000000" T="0.00000000"/>
      </TMATRIX>
    </MATRIX>
  ▼<MATRIX ID="2">
      <APPLY_TO_CHAIN CHAINID="B" NEW_CHAINID="D"/>
    ▼<TMATRIX>
        <ROWX X="0.00000000" Y="-1.00000000" Z="0.00000000" T="0.00000000"/>
        <ROWY X="1.00000000" Y="0.00000000" Z="0.00000000" T="93.55000305"/>
```


Figure 5a: XML file format


```json
{
    "data_resource": "PDBTM",
    "resource_version": " 1017",
    "software_version": "3.2.134",
    "resource_entry_url": "https://pdbtm.unitmp.org/entry/1rc2",
    "model_coordinates_url": "https://pdbtm.unitmp.org/api/v1/entry/1rc2.trpdb",
    "release_date": "10/03/2004",
    "pdb_id": "1rc2",
    "includes_het_groups": false,
    "chains": [
        {
            "chain_label": "A",
            "additional_chain_annotations": {
                "type": "alpha",
                "num_tm": "6"
            },
            "residues": [
                {
                    "site_data": [
                        {
                            "site_id_ref": 2,
                            "confidence_classification": "curated"
                        }
                    ],
                    "pdb_res_label": "1",
                    "aa_type": "MET"
                },
                {
                    "site_data": [
                        {
                            "site_id_ref": 2,
                            "confidence_classification": "curated"
                        }
                    ],
                    "pdb_res_label": "2",
                    "aa_type": "PHE"
                },
                {
                    "site_data": [
                        {
                            "site_id_ref": 2,
                            "confidence_classification": "curated"
                        }
```


Figure 5b: JSON file format

```
HEADER    MEMBRANE PROTEIN                        03-NOV-03   1RC2
TITLE     2.5 ANGSTROM RESOLUTION X-RAY STRUCTURE OF AQUAPORIN Z
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: AQUAPORIN Z;
COMPND   3 CHAIN: B, A;
COMPND   4 SYNONYM: BACTERIAL NODULIN-LIKE INTRINSIC PROTEIN;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI;
SOURCE   3 ORGANISM_TAXID: 562;
SOURCE   4 GENE: AQPZ, BNIP, B0875, C1009, SF0832, S0873;
SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   6 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE   7 EXPRESSION_SYSTEM_STRAIN: C43;
SOURCE   8 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PET28B
REMARK   2
REMARK   2 RESOLUTION.    2.50 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : CNS
REMARK   3   AUTHORS     : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK   3               : KUNSTLEVE,JIANG,KUSZEWSKI,NILGES,PANNU,
REMARK   3               : READ,RICE,SIMONSON,WARREN
REMARK   3
REMARK   3  REFINEMENT TARGET : NULL
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.50
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 50.00
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   DATA CUTOFF HIGH         (ABS(F)) : NULL
REMARK   3   DATA CUTOFF LOW          (ABS(F)) : NULL
REMARK   3   COMPLETENESS (WORKING+TEST)   (%) : 89.9
REMARK   3   NUMBER OF REFLECTIONS             : 21720
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : NULL
REMARK   3   R VALUE            (WORKING SET) : 0.227
REMARK   3   FREE R VALUE                     : 0.268
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : NULL
REMARK   3   FREE R VALUE TEST SET COUNT      : 1504
```

Figure 5c: Transformed structure file format

RESULTS:

By exploring Protein Databank of Transmembrane Proteins (PDBTM) Database, the membrane localization and structure were observed and the structure file formats were studied for the query Aquaporin (PDB ID: 1RC2). For this query of Aquaporin (PDB ID: 1RC2), the topology of chains was not found.

CONCLUSION:

The Protein Data Bank of Transmembrane Proteins (PDBTM) stands as a critical resource in the field of structural biology, specifically focusing on the three-dimensional structures of transmembrane proteins. Through its comprehensive collection of experimentally determined transmembrane protein structures, PDBTM has significantly contributed to our understanding of these vital biological components, which play pivotal roles in cell functions, signalling, and disease mechanisms. Thus, PDBTM remains an invaluable asset in the scientific community, serving as a cornerstone for structural and functional studies of transmembrane proteins.

REFERENCES:

1. Dániel Kozma, István Simon, Gábor E. Tusnády *Nucleic Acids Research*, Volume 41, Issue D1, 1 January 2013, Pages D524–D529, https://doi.org/10.1093/nar/gks1169
2. Margherita Sisto, ... Sabrina Lisi, in *Advances in Protein Chemistry and Structural Biology*, 2019. https://doi.org/10.1016/bs.apcsb.2018.11.010

DATE: 02/11/2023

WEBLEM 5(G)
CLASS ARCHITECTURE, TOPOLOGY AND HOMOLOGOUS SUPERFAMILY (CATH) DATABASE
(URL: http://www.cathdb.info/)

AIM:
To study the structural classification of proteins using CATH Database.

INTRODUCTION:
The CATH [Class (C)), Architecture (A), Topology (T), and Homology (H)] database is a hierarchical domain classification of protein structures maintained at UCL. The resource is largely derived using automatic methods, but manual inspections are necessary where automatic methods fail. There are four main levels: Class, Architecture, Topology, and Homology.

Level	Description
Class	The overall secondary-structure composition of each domain.
Architecture	Summarizes the shape revealed by the orientation of the secondary structure units, such as barrels and sandwich.
Topology	All the topology level, sequential connectivity is considered, such that members of the same architecture might have quite different topologies.
Homologous Superfamily	Indicative of a demonstrable evolutionary relationship. Equivalent to the superfamily level of SCOP.

Class, C- level:
Class is determined according to the secondary structure composition and packing within the structure. Three major classes are recognized; mainly- α, β and α-β. The last class α-β includes both alternating α/β structure and α+β structure.

Architecture, A-level:
This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually using a simple description of the secondary structure arrangement e.g., barrel or 3-layer sandwich.

Topology (Fold family), T-level:
Structures are grouped according to whether they share the same topology or fold in the core of the domain, that is, they share the same overall shape and connectivity of the secondary structures in the domain core. Within a given topology level the structures are similar but may have diverse functions. Where possible the name chosen for a given T-level is either the name

of the first structure in the family to be solved or the common name for the family [e.g., the globin fold or the immunoglobulin fold].

Homologous Superfamily, H-level:

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified either by high sequence identity or structure comparison using SSAP. Domains within each H-level are subclustered into sequence families using multi linkage.

METHODOLOGY:

A. Searching domains in the CATH database:

1. Go to the CATH database homepage.
2. The CATH database can be browsed searched and downloaded using the links present in the "Links for Researchers section".
3. To search the CATH database use "Search CATH by ID/sequence/text" link.
4. Search by ID can be done using "CATH DOMAIN ID", "CATH CHAIN ID" or "PDB ID" whereas text search can be performed using structural or functional term like —"chaperone" or —"helix". Search by sequence can be performed by pasting proteins sequences in FASTA format.
5. Search by ID displays; Domain ID, structure, CATH code, the different chains present, PDB code and functional annotation if known. To identify homologous domains, click in the —"CATH code". The classification architecture, topology, homologous superfamily, sequence, and structural information related to the domain can be obtained by clicking on the "Domain ID": search using keyword displays the PDB files and CATH classification code and name that contains the search term. More information on the domain can be obtained by clicking the "CATH code" of the classification entries.
6. Search by — "sequence" blasts against sequences derived from structural domains present in CATH database and displays domains with similar sequences present in CATH with their "Domain ID" and "CATH" code. More information on the domains can be obtained using the "Domain ID" and "CATH code" link.

B. Performing structural and functional analysis of proteins in CATH:

1. Retrieve the PDB ID for the protein of interest.
2. Paste the "PDB ID" or upload a — "PDB file", check the "PDB" versus — "CATH CATHEDRAL Scan" and click on —"Continue".
3. CATHEDRAL server identifies the domains present in the protein.
4. Structural and Topological information on the domains present in the protein of interest can be obtained using "Domain ID" and — "Topology ID". Functional information, taxonomic distributions, multi-domain architectures and protein-protein interaction data can be obtained using the Geno3D Server link.

OBSERVATIONS:

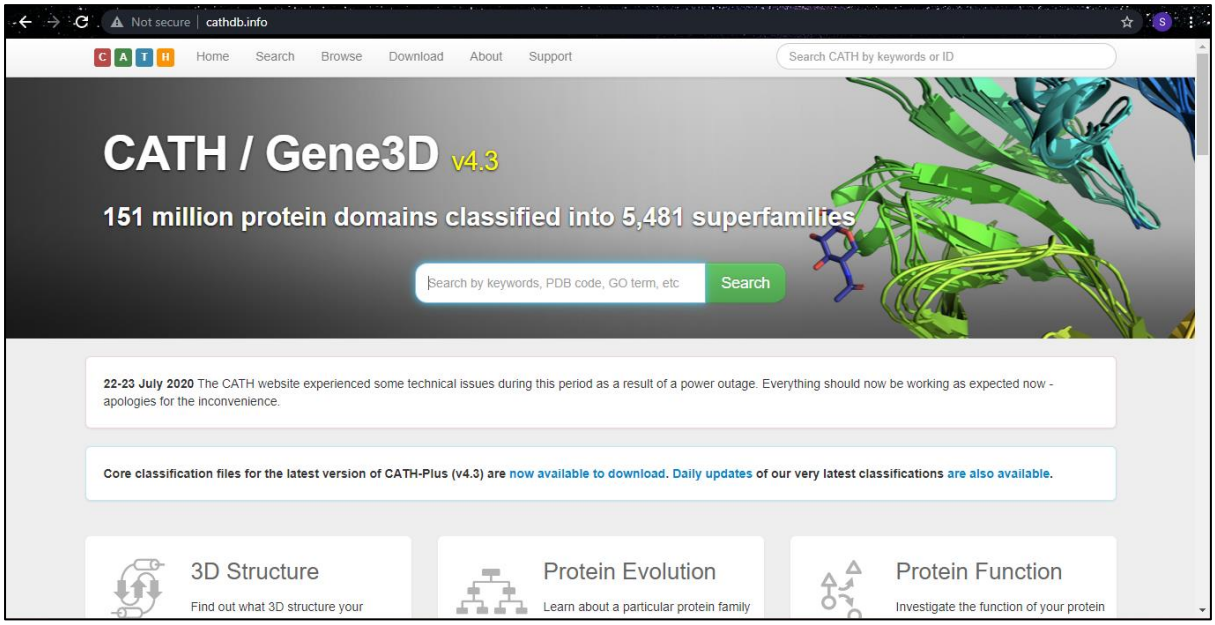


Figure 1: Homepage of CATH Database

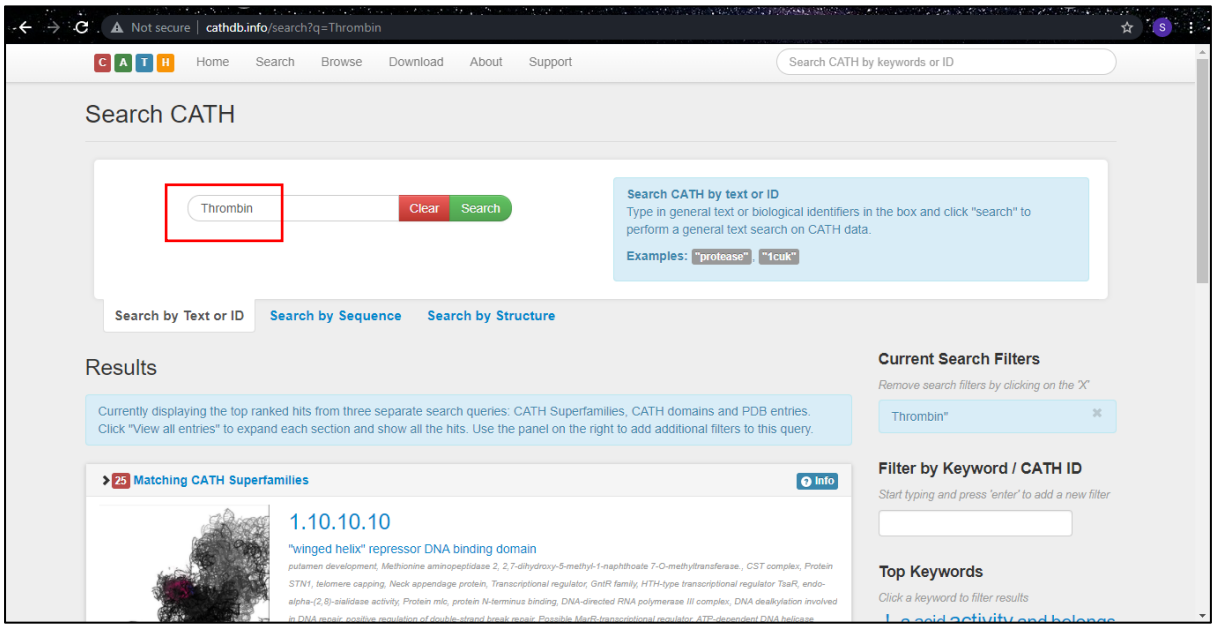


Figure 2: Result page for Query- THROMBIN in CATH Database

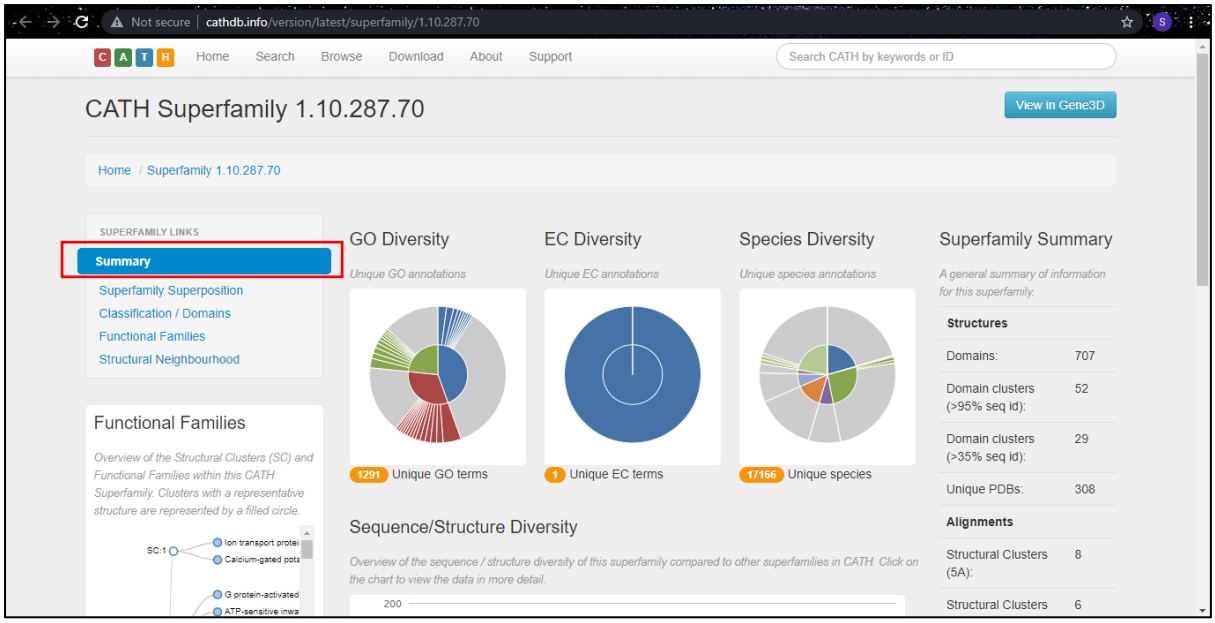


Figure 3: Summary page for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

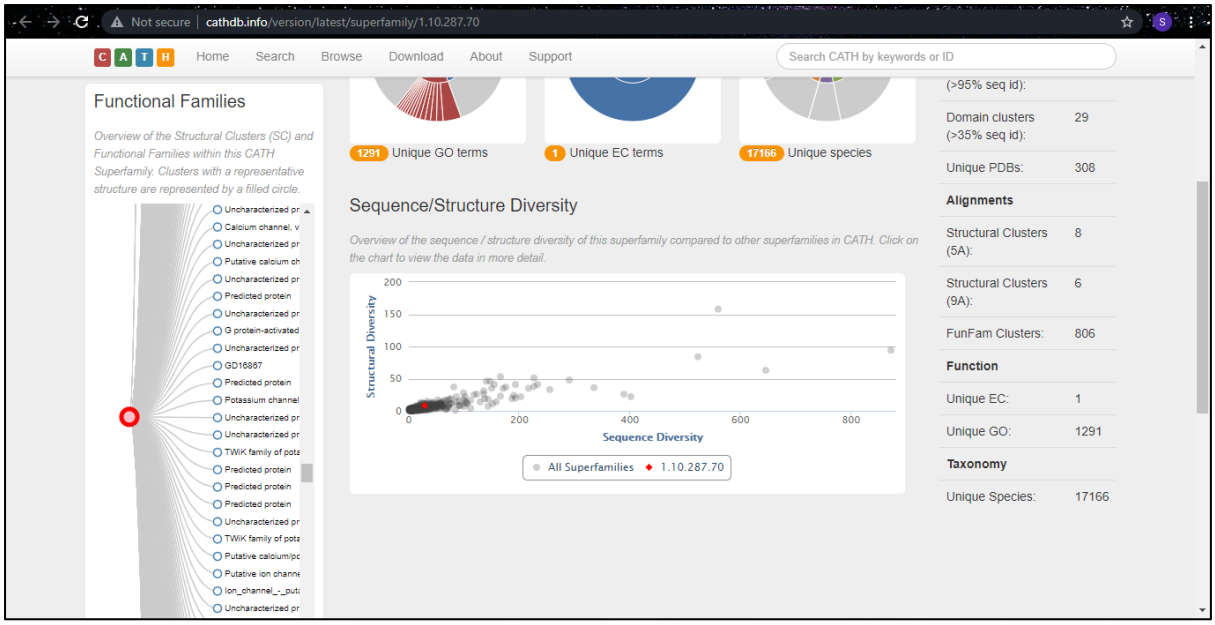


Figure 3a: Summary page for CATH Superfamily 1.10.287.70 displaying Sequence/ Structure diversity- Thrombin query in CATH Database

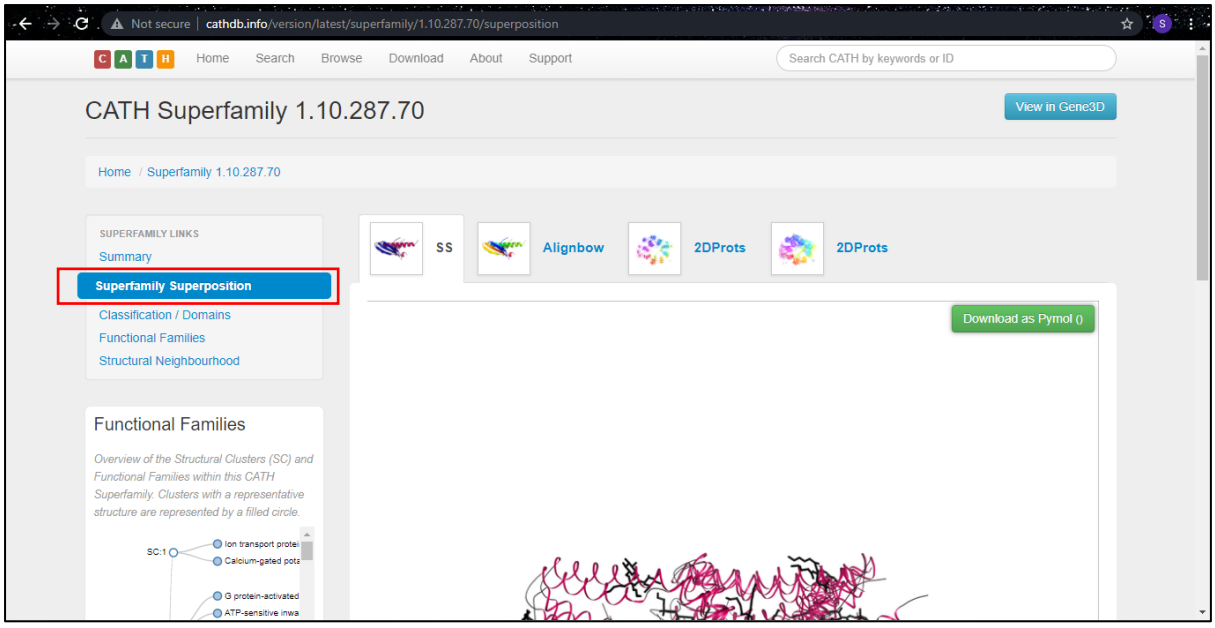


Figure 4: Information for Superfamily Superposition section for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

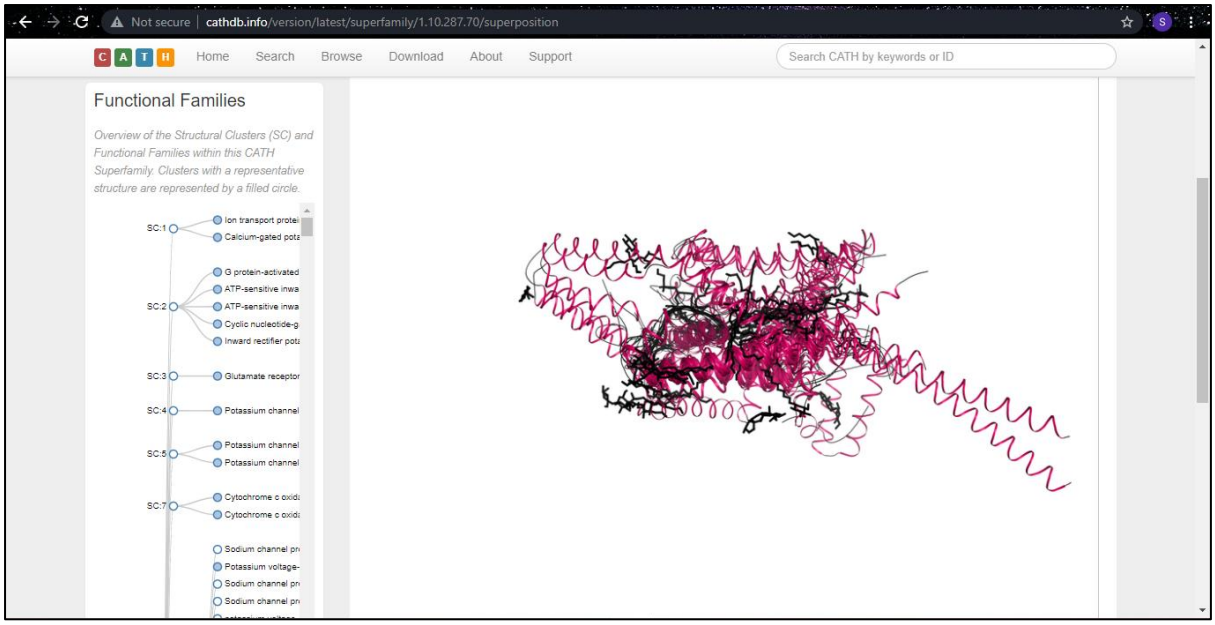


Figure 4a: Super Secondary structure in Superfamily Superposition section for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

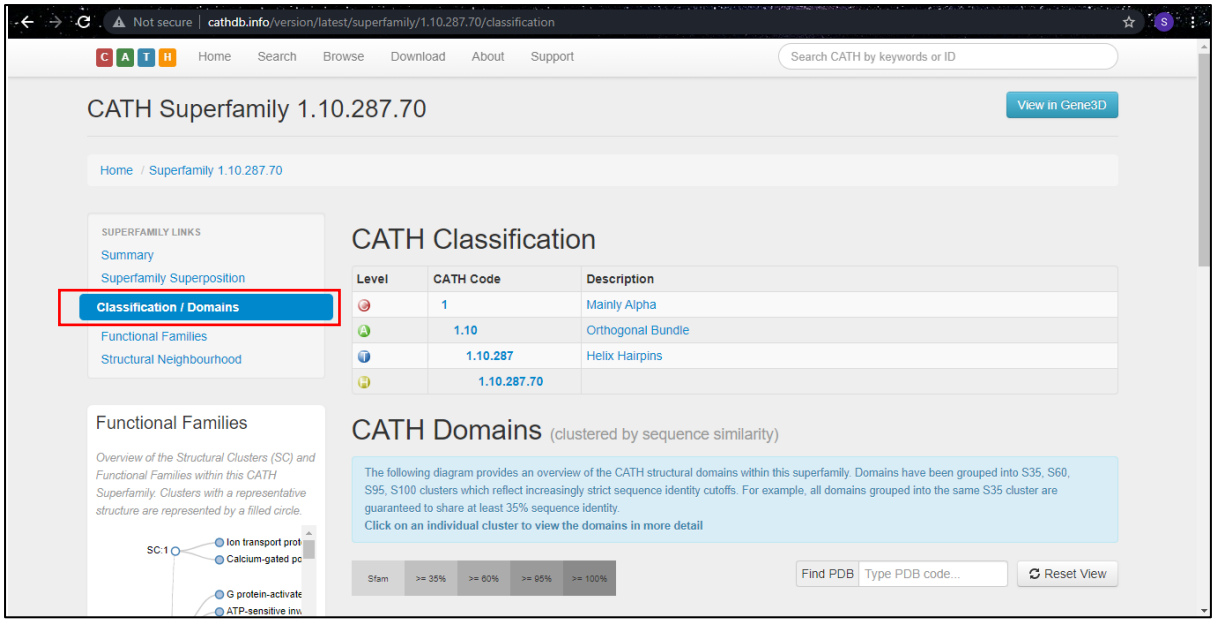


Figure 5: Information for Classification/ Domains section displaying CATH classification & CATH domains for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

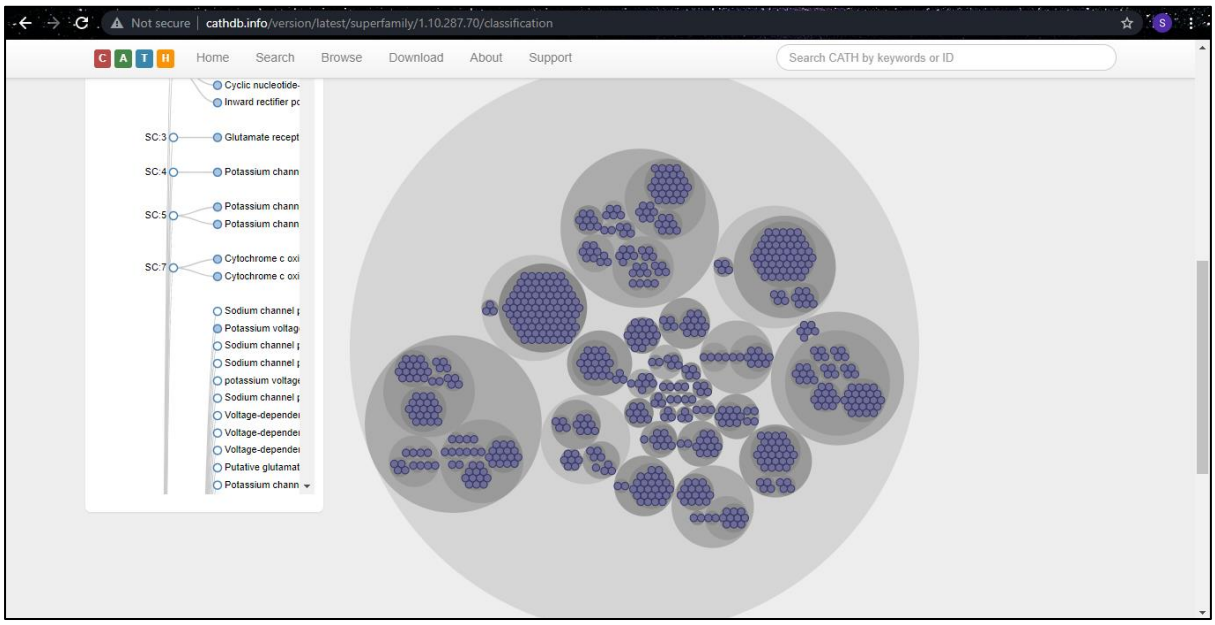


Figure 5a: CATH domains for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

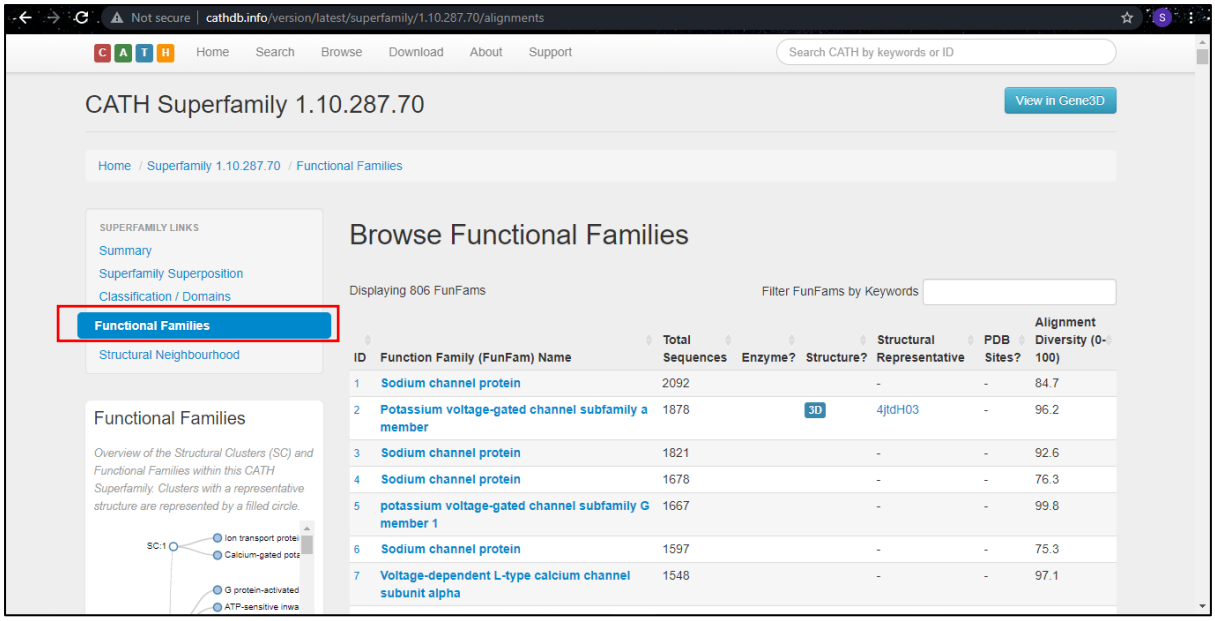


Figure 6: Information for Functional Families (Overview) for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

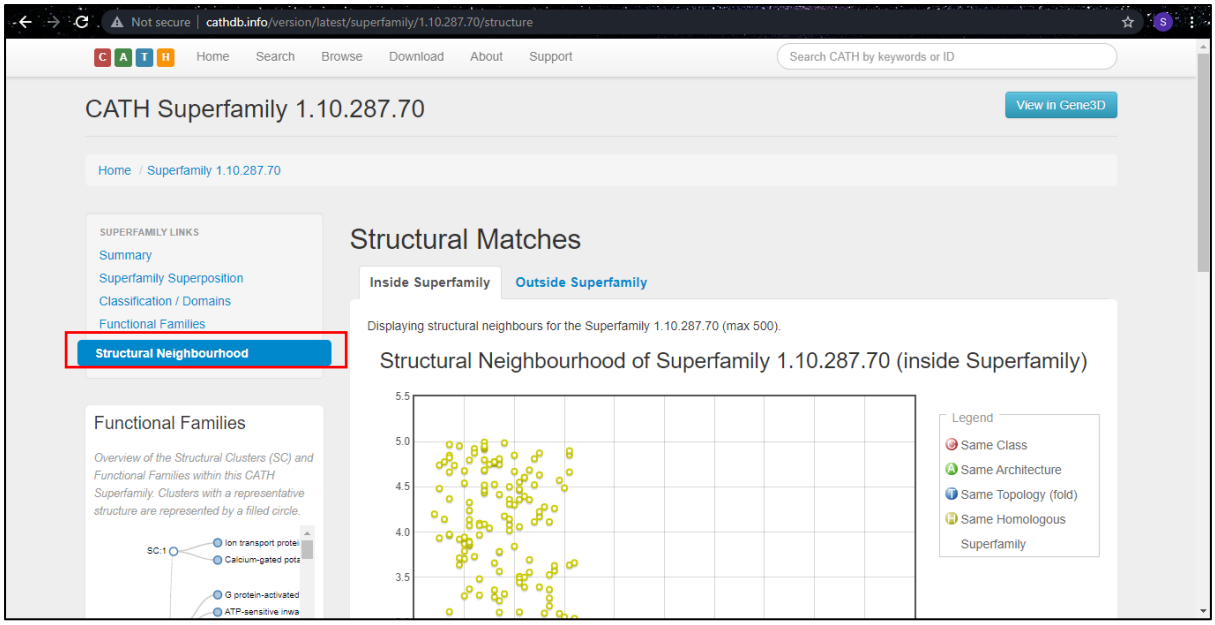


Figure 7: Information for Structural Neighborhood for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

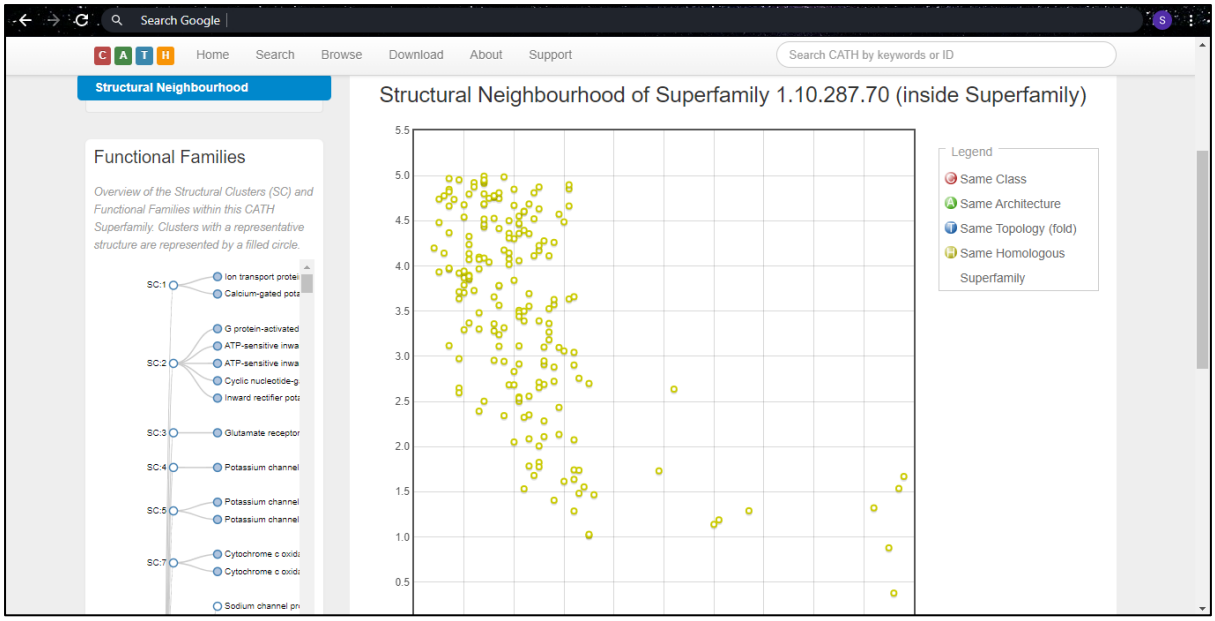


Figure 7a: Structural Neighborhood (inside Superfamily) for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

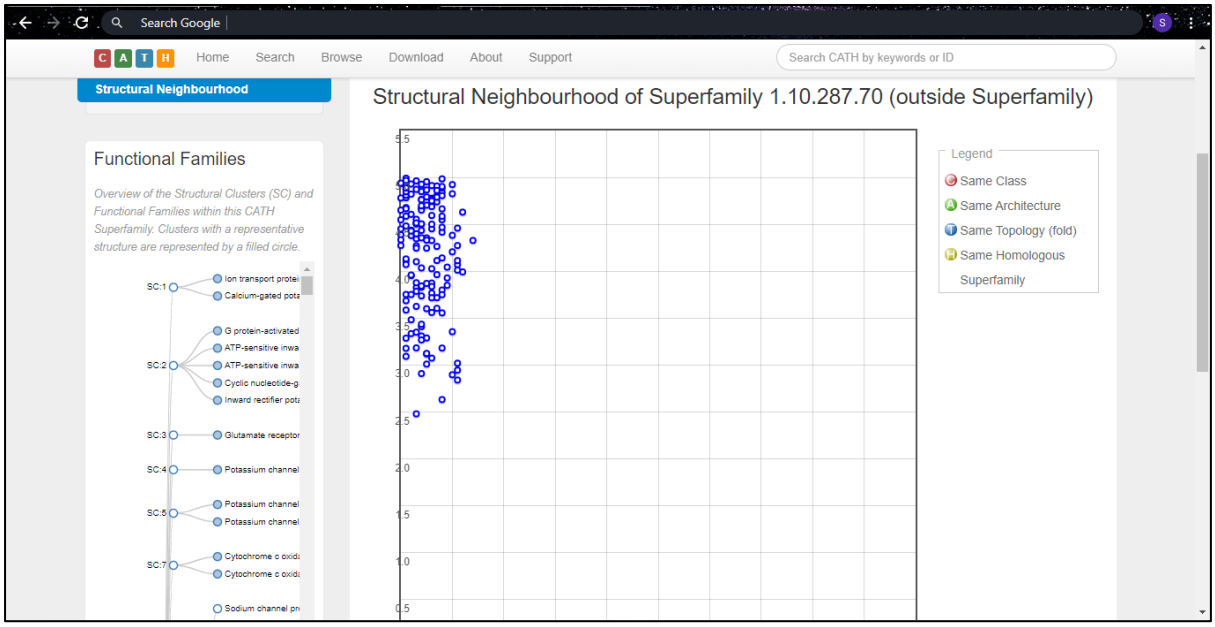


Figure 7b: Structural Neighborhood (outside Superfamily) for CATH Superfamily 1.10.287.70- Thrombin query in CATH Database

RESULTS:

Using the CATH database, Thrombin protein was considered as a query. After firing the query, a match from CATH superfamilies namely, CATH Superfamily 1.10.287.70 was explored for Thrombin. Various sections such as Summary, Superfamily Superposition, Classification/ Domains, Functional families, and Structural neighborhood were studied. The CATH database provides classification of protein domains and the information on the evolutionary relationships of protein domains.

CONCLUSION:

The CATH database was explored and information was retrieved for Thrombin protein.

REFERENCES:

1. Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: increased structural coverage of functional space. Nucleic acids research, 49(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

DATE: 02/11/2023

WEBLEM 5(H)
STRUCTURAL CLASSIFICATION OF PROTEINS (SCOPE) DATABASE
(URL: https://scop.berkeley.edu/)

AIM:
To study the structural classification of proteins using SCOPe Database.

INTRODUCTION:
The SCOP (structural classification of proteins) database maintained at the MRC Laboratory of Molecular Biology and Center for Protein Engineering describes structural relationships between proteins of known structure. Proteins are classified in a hierarchical fashion to reflect their structural and evolutionary relatedness. Within the hierarchy there are many the levels, but principally these describe the family, superfamily, and the fold. Some basic terminologies:

1. **Domain:** structure and evolution independent stretch of amino acids in polypeptides.
2. **Family:** clear evolution relationship. Members show structure homology and final similarity, pairwise sequence similarity.
3. **Superfamily:** usually low sequence similarity between members.
4. **Fold:** major secondary structure elements are in same arrangement and with same topological connection. Turns and loops can be different. Proteins in the same fold may not have common evolutionary origin.

METHODOLOGY:
1. Open the website of SCOPe..
2. Data in SCOPe can be browsed using ―ENTER SCOP at the top of the hierarchy or searched using the ‘Keyword search of SCOP entries.
3. Search in SCOP can be performed using the ― sunid: a unique identifier for all entries in the SCOP hierarchy‖ for e.g., search using ‘57942’ displays results for the coiled coil proteins, ―sccs: an identifier for class (alphabetical), fold, superfamily (all numerical), for e.g., to search for truncated hemoglobin family the sccs is ―a.1.1.1 where ―a is the class code and the remaining numericals refer to the fold, superfamily, family code respectively and ―keyword search.
4. Multiple keywords can be given by combining the keywords with '+' or '-' operator, for e.g., to search for ‘hemoglobin’ in pig only, the search would be represented as ― ‘hemoglobin + pig’ whereas to search for ―all hemoglobin entries except those present in human" the search keyword would be ― hemoglobin – human’. Search for new families, superfamilies, folds or classes is possible using keywords like ‘newfa’, ‘newsf’, ‘newcf’ and ‘newcl’ respectively.
5. Search using 'MSDlite" searches text fields in PDB and returns links to the corresponding CATH entry.

6. Identify the class and describe the fold of the protein. Using the superfamily link find the functional annotation given to the protein.
7. The 3D structures of the domain scan are studied by clicking on the rasmol or chime link. Identify the secondary structures elements present in the structures. Other domain related information can be obtained with link option.

OBSERVATIONS:

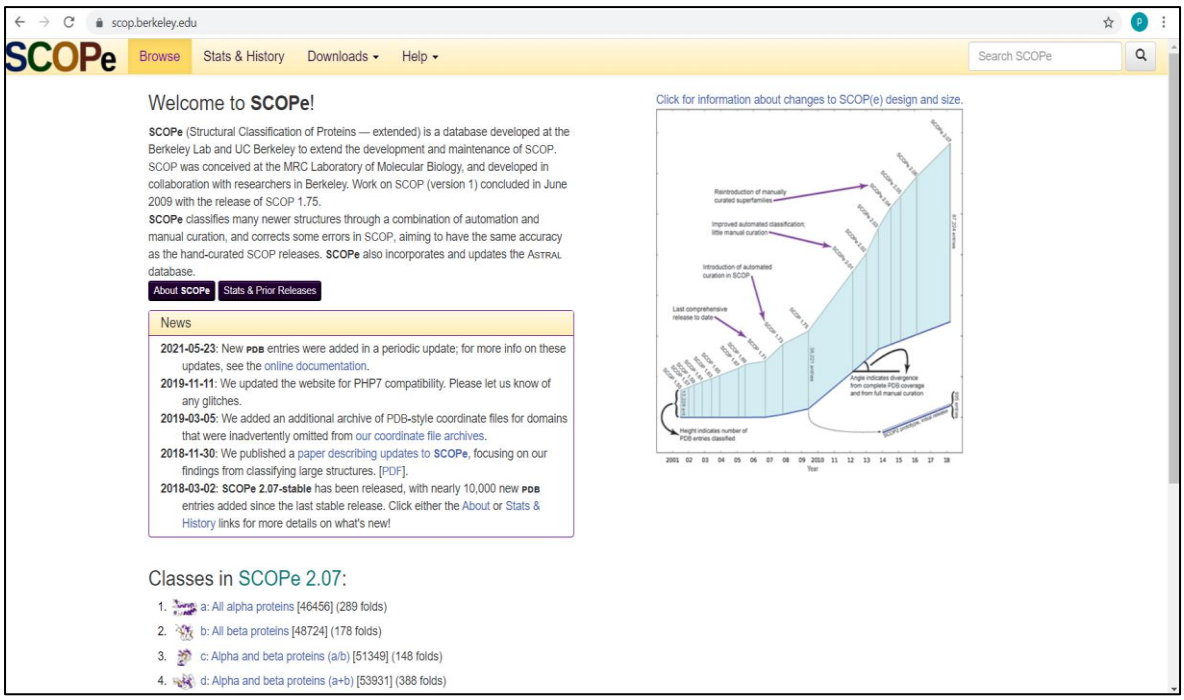


Figure 1: Homepage of SCOPe database

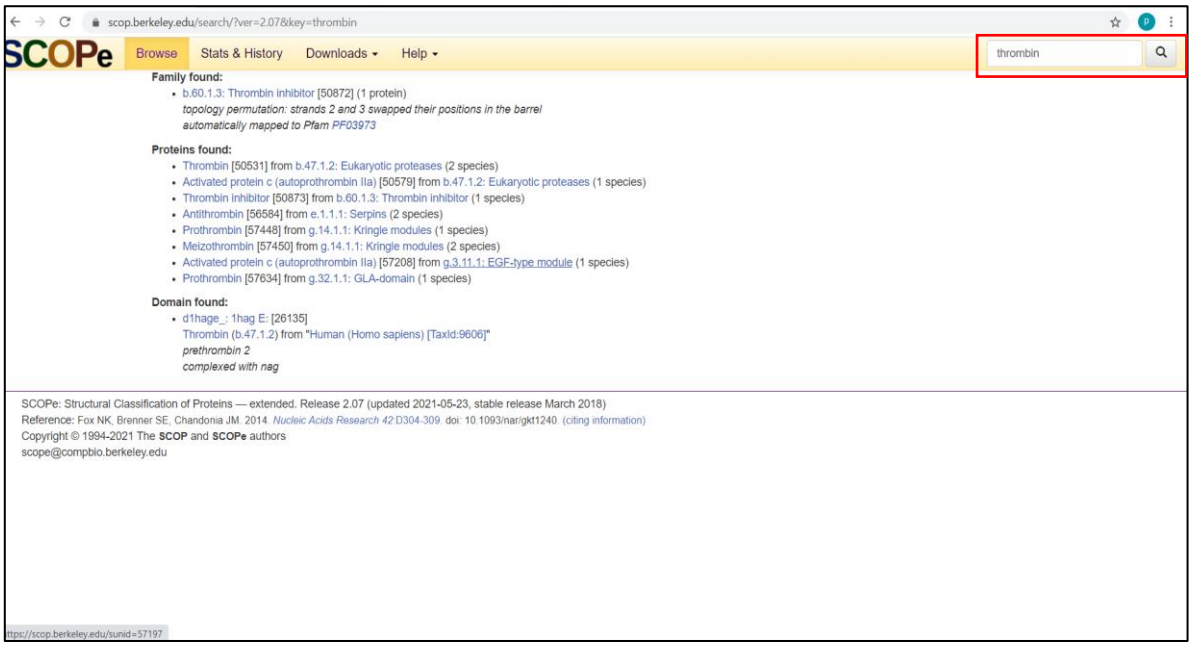


Figure 2: Hit page for query 'thrombin' in SCOPe database

Figure 3: Lineage of the query protein 'thrombin' shown in the SCOPe database

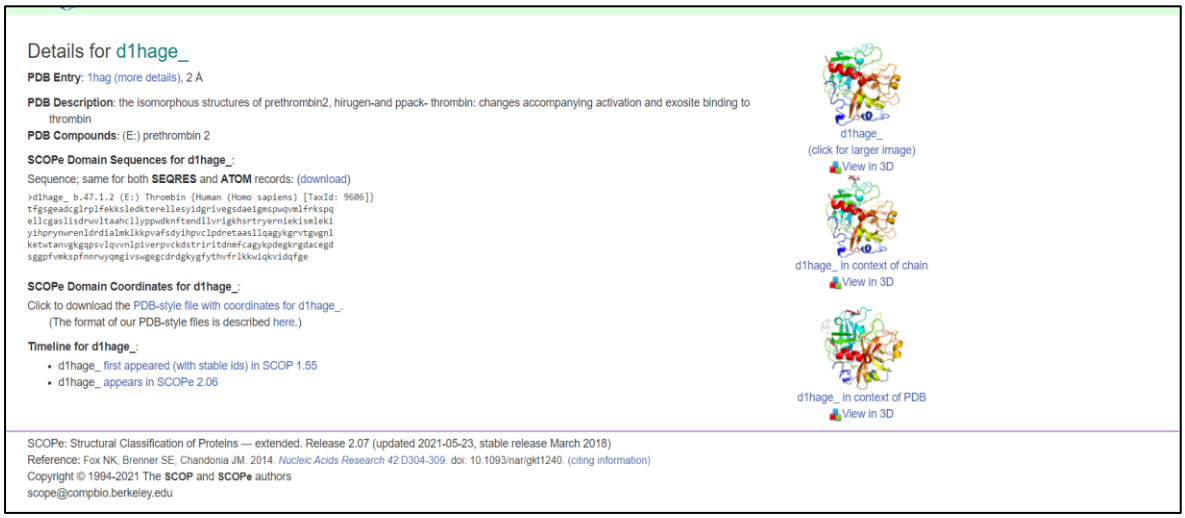


Figure 4: Structural details shown in the SCOPe database

RESULTS:

Using the SCOPe database, thrombin protein was considered as a query. After firing a query, information from different sections such as domain relationship, lineage of the protein, detailed structure, and proteins related to thrombin was obtained. It is easy and convenient to use this database to get all types of information about the protein of our interest.

CONCLUSION:

The SCOPe database was explored and information was retrieved for the thrombin protein.

REFERENCES:

1. Naomi K. Fox, Steven E. Brenner, John-Marc Chandonia, SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures, Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D304–D309, https://doi.org/10.1093/nar/gkt1240

DATE: 08/11/2023

WEBLEM 6
INTRODUCTION TO GENOMICS AND ITS VARIOUS BROWSERS

Computer databases are an increasingly necessary tool for organizing the vast amounts of biological data currently available and for making it easier for researchers to locate relevant information. In 1979, the Los Alamos Sequence Database was established as a repository for biological sequences. In 1982, this database was renamed GenBank and, later the same year, moved to the newly instituted National Center for Biotechnology Information (NCBI), where it lives today. By the end of 1983, more than 2,000 sequences were stored in GenBank, with a total of just under 1 million base pairs (Cooper & Patterson, 2008). At about the same time, a joint effort between NCBI, the European Molecular Biology Laboratory (EMBL), and the DNA Databank of Japan (DDBJ) created the International Nucleotide Sequence Database Collaboration (INSDC) to collect and disseminate the burgeoning amount of nucleotide and amino acid sequence data that was becoming available. Since then, the INSDC databases have grown to contain over 95 billion base pairs, reflecting an exponential growth rate in which the amount of stored data has doubled every 18 months. The advent of next-generation sequencing technologies, metagenomics, genome-wide association studies (GWAS), and endeavors such as the 1000 Genomes Project will only increase the tremendous volume and complexity of this and other sequence data collections. The sheer volume of the raw sequence data in these repositories has led to attempts to reorganize this information into various kinds of smaller, specialized databases. Such databases include various genome browsers, model organism databases, molecule- or process-specific databases, and others. To get an understanding of the growth of these resources, one need only look at the annual database issue of the journal Nucleic Acids Research. In one of the first database issues—the one in which GenBank is described—only a few dozen databases are listed. In contrast, the latest database issue describes over 1,000 genomics databases and tools. However, even this list of resources is only part of the overall picture. Today, it appears that there are upwards of 3,000 distinct genomic resources, tools and databases ate publicly available on the internet.
Various genome browser is as follow:

1. **NCBI Genome Browser:**
   NCBI Taxonomy, maintained by the National Center for Biotechnology Information (NCBI), is a comprehensive database system that organizes and categorizes biological species into a hierarchical structure. It serves as a standardized classification system for all organisms, providing a framework to understand evolutionary relationships among different species. Some Keys features are as follow:
   a. **Hierarchical Organization:** NCBI Taxonomy arranges organisms into a hierarchical tree-like structure known as a taxonomic hierarchy. This hierarchy begins with the root node, representing all life forms, and branches into kingdoms, phyla, classes, orders, families, genera, and species.
   b. **Scientific Nomenclature:** It assigns unique scientific names to each recognized species, ensuring standardized naming conventions based on binomial nomenclature (genus and species name). For example, humans are identified as Homo sapiens.
   c. **Taxonomic IDs:** Each organism in the NCBI Taxonomy is assigned a unique numerical identifier, known as a Taxonomic Identifier (taxid), facilitating easy reference and retrieval of information.
   d. **Evolutionary Relationships:** NCBI Taxonomy reflects the evolutionary relationships among organisms. It is based on scientific evidence derived from

various fields such as genetics, morphology, and phylogenetics, aiming to show the evolutionary connections between species.

e. **Integration with NCBI Databases:** Taxonomic information is linked across various NCBI databases, allowing users to access taxonomic classifications, sequences, literature, and other data associated with specific organisms.
f. **Taxonomic Browser and Search Tools:** The NCBI Taxonomy website provides a user-friendly interface for browsing the taxonomic tree, searching for specific organisms, retrieving taxonomic information, and accessing related resources.
g. **Standard Reference:** It serves as a reference for researchers, biologists, and database curators to maintain consistency in naming and classification of organisms across scientific studies and databases.

NCBI Taxonomy plays a crucial role in biological research, genomics, biodiversity studies, evolutionary biology, and many other disciplines. Researchers use this resource to classify newly discovered species, study evolutionary patterns, understand genetic relationships, and navigate the diversity of life on Earth. Overall, NCBI Taxonomy serves as a fundamental tool for organizing and understanding the vast array of life forms, providing a standardized framework that aids in biological research and knowledge dissemination.

2. **<u>UCSC Genome Browser:</u>**

The UCSC Genome Browser is a powerful tool used by researchers, scientists, and students worldwide to visualize and analyze genomic data. Developed by the University of California, Santa Cruz (UCSC), this web-based browser allows users to explore genomes, compare genetic sequences, and study various genomic annotations and data tracks. Here's an overview of the UCSC Genome Browser's key features and functionalities:

a. **Genome Visualization:** The browser provides a graphical representation of genomes, allowing users to view DNA sequences, genes, exons, introns, regulatory elements, and more.
b. **Multiple Genome Assembly Support:** It supports various genome assemblies from different species, including human, mouse, fruit fly, yeast, and many others. Users can switch between assemblies and compare genomic features across different species.
c. **Data Tracks:** Users can overlay diverse genomic data tracks onto the reference genome, such as gene expression data, chromatin accessibility, epigenetic modifications, evolutionary conservation, and more. These tracks offer valuable insights into gene regulation, evolutionary conservation, and other genomic features.
d. **Custom Tracks:** Users can upload and display their own experimental or computational data on the browser, enabling personalized analysis and visualization.
e. **Tools and Utilities:** The UCSC Genome Browser provides tools for searching genes, sequences, or specific genomic regions. It also includes features for data retrieval, analysis, and exporting data for further research
f. **Continuous Updates:** The browser regularly incorporates new genome assemblies, annotations, and data tracks to keep pace with the evolving field of genomics.
g. **User-Friendly Interface:** The browser interface is intuitive and user-friendly, allowing easy navigation and customization of displayed tracks and data.

h. **Educational Resources:** UCSC provides tutorials, guides, and documentation to help users understand how to use the browser effectively for research or educational purposes. Researchers use the UCSC Genome Browser for various purposes, including studying gene function, identifying regulatory elements, understanding genetic variation, exploring evolutionary relationships, and more. Its accessibility and wealth of genomic data make it an invaluable resource in genomic research and education.

It's important to note that while the UCSC Genome Browser is a robust tool, it may require some familiarity with genomics and bioinformatics to fully utilize its capabilities. However, its user-friendly interface also caters to beginners, providing a platform for learning and exploration in the field of genomics.

3. **ENSEMBL Browser:**

The Ensembl genome browser is an extensively used web-based platform that offers a comprehensive and user-friendly interface for exploring and analyzing genomic data from a wide array of species. Developed by the Ensembl project, a collaborative effort between the European Bioinformatics Institute (EBI) and the Wellcome Sanger Institute, the browser provides access to various genome assemblies, gene annotations, comparative genomics data, and more. The Ensembl genome browser include:

a. **Genome Annotations:** Ensembl provides detailed annotations of genomes, including genes, transcripts, exons, introns, regulatory elements, and other functional genomic elements. These annotations are available for numerous species, ranging from humans and model organisms to less-studied species.
b. **Multiple Genome Alignments:** It allows users to visualize and compare genomic sequences across different species through multiple sequence alignments. This feature assists in identifying conserved regions, evolutionary relationships, and understanding genome evolution.
c. **Variation Data:** Ensembl incorporates data on genetic variations, including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants. Users can explore genetic variations within populations and their potential impact on genes and phenotypes.
d. **Customizable Data Views:** Users can customize their views by selecting and overlaying diverse data tracks, such as gene expression, epigenetic modifications, evolutionary conservation, and more. This flexibility aids in tailored analysis and visualization.
e. **Gene Trees and Homology:** Ensembl provides gene trees and information on homologous genes across species, enabling the study of gene families, orthologs, and paralogs, which can offer insights into gene function and evolutionary relationships.
f. **Functional Analysis Tools:** The browser integrates various tools for functional analysis, allowing users to predict gene function, explore pathways, analyze protein domains, and perform other bioinformatics analyses.
g. **User-Friendly Interface:** Ensembl offers an intuitive and user-friendly interface that facilitates navigation, data retrieval, and analysis, making it accessible to researchers with varying levels of expertise.
h. **Regular Updates:** The Ensembl project continuously updates its database with new genome assemblies, annotations, and data tracks, ensuring that users have access to the latest genomic information.

Ensembl's rich collection of genomic data and its user-friendly interface make it an invaluable resource for researchers studying genomics, genetics, evolutionary biology, and related fields. It serves as a crucial tool for understanding genomes, gene function, genetic variations, and evolutionary processes across different species.

4. **GOLD Database:**

The Genomes Online Database (GOLD) serves as a comprehensive resource cataloging metadata associated with genome and metagenome sequencing projects worldwide. Developed by the Joint Genome Institute (JGI), which is part of the U.S. Department of Energy, GOLD offers a platform for researchers to access, organize, and compare information related to sequencing projects across various organisms and environments. Key features and aspects of the GOLD database include:

a. **Metadata Repository:** GOLD contains a wealth of metadata related to genome and metagenome sequencing projects. This includes information on project details, sequencing strategies, environmental data, organism classifications, sample origins, and other pertinent information.
b. **Cataloging Diversity:** It aims to capture and represent the diversity of life by encompassing data from a broad spectrum of organisms, ranging from microbes and viruses to eukaryotes, both in isolation (genomes) and in their natural environments (metagenomes).
c. **Standardized Information:** GOLD implements standardized data curation and annotation practices, ensuring that metadata across projects adhere to consistent formatting and quality standards. This facilitates data sharing, comparison, and analysis among researchers.
d. **Search and Access Tools:** Users can search and access project metadata using various parameters such as organism name, habitat type, sequencing technology, geographical location, and more. This allows for targeted retrieval of relevant information.
e. **Integration with Other Databases:** GOLD is linked to other genomic and metagenomic databases and resources, enabling cross-referencing and integration of data with platforms such as NCBI, EMBL-EBI, and others.
f. **Research and Analysis Support:** The database supports research and analysis by providing access to metadata essential for understanding the context and characteristics of sequencing projects. Researchers can use this information for comparative genomics, evolutionary studies, biodiversity assessments, and other analyses.
g. **Updates and Community Involvement:** GOLD is regularly updated to include new sequencing projects and additional metadata. It also encourages community involvement and contributions to enhance the database's comprehensiveness.

GOLD serves as a valuable resource for researchers interested in exploring genomic and metagenomic diversity, understanding environmental microbial communities, conducting comparative genomics, and tracking the progress of sequencing projects globally. It plays a pivotal role in advancing our knowledge of the world's genetic diversity and its applications across various scientific disciplines.

5. **MBGD Database:**

The Microbial Genome Database (MBGD) is a valuable resource that focuses on comparative analysis of microbial genomes. MBGD is designed to facilitate the comparative study of prokaryotic genomes, providing a platform for researchers to analyze and explore the genetic relationships among various microorganisms and characteristics of MBGD include:

a. **Microbial Genome Comparison:** MBGD offers tools and resources for the comparative analysis of microbial genomes. It enables users to compare multiple genomes of bacteria and archaea, identifying similarities and differences in gene content, orthologous genes, conserved domains, and evolutionary relationships.
b. **Orthologous Groups:** The database classifies genes into orthologous groups, which consist of genes derived from different species that are evolutionarily related by descent from a common ancestral gene. Orthologous groups aid in understanding gene function and evolution across microbial species.
c. **Functional Annotation:** MBGD provides functional annotations for genes within microbial genomes. These annotations include predicted functions, protein domains, pathways, and other relevant information, aiding researchers in understanding the biological roles of genes.
d. **Phylogenetic Relationships:** The database offers phylogenetic trees and diagrams that illustrate the evolutionary relationships between different microbial species. This information helps researchers trace the evolutionary history and relatedness of microorganisms.
e. **User-Friendly Interface:** MBGD features an intuitive interface that allows users to perform searches, access comparative genomics data, and visualize results. The user-friendly design facilitates navigation and analysis for researchers interested in microbial genomics.
f. **Integration of Genomic Data:** MBGD integrates genomic data from various sources, compiling information from publicly available databases and research studies. This integration enables comprehensive comparative analysis across multiple microbial genomes.
g. **Regular Updates:** The database is regularly updated to include new genomic data, annotations, and improvements in analysis tools, ensuring that users have access to the latest information.

MBGD serves as a valuable resource for researchers studying microbial genomics, evolution, ecology, and functional genomics. By providing tools for comparative analysis and detailed annotations of microbial genomes, MBGD contributes to our understanding of microbial diversity, adaptation, and evolutionary processes in the microbial world.

6. **ICTVdb Database:**

The International Committee on Taxonomy of Viruses (ICTV) maintains the ICTV Virus Metadata Database (ICTVdb), which serves as a comprehensive repository of information related to the taxonomy and classification of viruses. The ICTV is a global authority responsible for naming, classifying, and categorizing viruses, and ICTVdb plays a crucial role in organizing and disseminating this information. Some of the key aspects and features of ICTV Virus Metadata Database (ICTVdb) include:

a. **Virus Taxonomy:** ICTVdb contains standardized information on virus taxonomy and classification. It provides a systematic framework for organizing

viruses into hierarchical categories based on genetic and phenotypic characteristics.

b. **Virus Nomenclature:** The database maintains a standardized nomenclature system for naming and identifying viruses, ensuring consistency and clarity in virus classification across the scientific community.
c. **Comprehensive Virus Information:** ICTVdb offers detailed metadata on various aspects of viruses, including their genome structure, viral families, genera, species demarcation, host range, geographical distribution, and other relevant information.
d. **Reference Database:** It serves as a reference resource for researchers, virologists, and public health professionals seeking authoritative information on virus taxonomy and classification.
e. **Regular Updates:** The database is regularly updated to incorporate newly identified viruses, taxonomic revisions, and advancements in virus classification based on scientific research and discoveries.
f. **Access and Search Tools:** ICTVdb provides user-friendly tools for accessing and searching virus taxonomy and associated metadata. Users can query the database to retrieve specific virus-related information using various search parameters.
g. **Integration with Other Databases:** ICTVdb integrates with other virus-related databases and resources, facilitating cross-referencing and linking of virus taxonomy and classification information.

ICTVdb is an essential resource for researchers and professionals involved in virology, infectious disease research, epidemiology, and public health. It plays a vital role in standardizing virus taxonomy and classification, aiding in the identification, characterization, and understanding of viruses, ultimately contributing to global efforts in virus surveillance, diagnosis, and control.

REFERENCES:

1. **NCBI Genome Browser/ NCBI Taxonomy:**
   a. Federhen, S. (2011, December 1). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178
   b. Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., . . . Ye, J. (2009, January 1). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *37*(Database), D5–D15. https://doi.org/10.1093/nar/gkn741
   c. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2007, December 23). GenBank. *Nucleic Acids Research*, *36*(Database), D25–D30. https://doi.org/10.1093/nar/gkm929
2. **UCSC Browser:**
   a. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, A. D. (2002, May 16). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

b. Karolchik, D. (2004, January 1). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(90001), 493D – 496. https://doi.org/10.1093/nar/gkh103
c. Raney, B. J., Dreszer, T. R., Barber, G. P., Clawson, H., Fujita, P. A., Wang, T., Nguyen, N., Paten, B., Zweig, A. S., Karolchik, D., & Kent, W. J. (2013, November 13). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, *30*(7), 1003–1005. https://doi.org/10.1093/bioinformatics/btt637

3. ENSEMBL Browser:

a. Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., Ruffier, M., Taylor, K., Vullo, A., & Flicek, P. (2014, September 17). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*, *31*(1), 143–145. https://doi.org/10.1093/bioinformatics/btu613
b. Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Aranganathan, N. K., Langridge, N., Lowy, E., McDowall, M. D., Maheswari, U., Nuhn, M., . . . Staines, D. M. (2015, November 17). Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, *44*(D1), D574–D580. https://doi.org/10.1093/nar/gkv1209
c. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016, June 6). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0974-4

4. GOLD Database:

a. Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., Coates, R. C., Hadjithomas, M., Pavlopoulos, G. A., Paez-Espino, D., Yoshikuni, Y., Visel, A., Whitman, W. B., Garrity, G. M., Eisen, J. A., Hugenholtz, P., Pati, A., Ivanova, N. N., Woyke, T., . . . Kyrpides, N. C. (2017, June 12). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*, *35*(7), 676–683. https://doi.org/10.1038/nbt.3886
b. Reddy, T., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. A., & Kyrpides, N. C. (2014, October 27). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Research*, *43*(D1), D1099–D1106. https://doi.org/10.1093/nar/gku950
c. Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., Woyke, T., Göker, M., Parker, C. T., Amann, R., Beck, B. J., Chain, P. S. G., Chun, J., Colwell, R. R., Danchin, A., Dawyndt, P., Dedeurwaerdere, T., DeLong, E. F., Detter, J. C., De Vos, P., Donohue, T. J., Dong, X. Z., . . . Klenk, H. P. (2014, August 5). Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. *PLoS Biology*, *12*(8), e1001920. https://doi.org/10.1371/journal.pbio.1001920

5. **MBGD Database:**
    a. Uchiyama, I. (2003, January 1). MBGD: microbial genome database for comparative analysis. *Nucleic Acids Research*, *31*(1), 58–62. https://doi.org/10.1093/nar/gkg109
    b. Uchiyama, I., Mihara, M., Nishide, H., & Chiba, H. (2012, October 30). MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Research*, *41*(D1), D631–D635. https://doi.org/10.1093/nar/gks1006
    c. Uchiyama, I. (2017). Ortholog Identification and Comparative Analysis of Microbial Genomes Using MBGD and RECOG. *Methods in Molecular Biology*, 147–168. https://doi.org/10.1007/978-1-4939-7015-5_12
6. **ICTVdb Database:**
    a. King, A. M., Lefkowitz, E., Adams, M. J., & Carstens, E. B. (2012). *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. Academic Press.
    b. Lauber, C., & Gorbalenya, A. E. (2012, August 31). Genetics-Based Classification of Filoviruses Calls for Expanded Sampling of Genomic Sequences. *Viruses*, *4*(9), 1425–1437. https://doi.org/10.3390/v4091425
    c. Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., . . . Zerbini, F. M. (2017, January 3). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168. https://doi.org/10.1038/nrmicro.2016.177

DATE: 08/11/23

WEBLEM 6(A)
GENOMES ONLINE DATABASE (GOLD)
(URL: https://gold.jgi.doe.gov/)

AIM:

To explore the Genome Online Database in order to retrieve information about genome and metagenome subsequence.

INTRODUCTION:

Genomes Online Database (GOLD) is a freely available information rich resource of sequencing projects and their associated metadata. GOLD serves as a major re-source that catalogues and monitors genome and metagenome projects from around the world. Since its inception in 1997, GOLD has grown exponentially, keeping pace with the growth in number of sequencing projects. DNA sequencing recently celebrated its 40th anniversary. There have been several technological revolutions in this relatively short period of time. A reduction in sequencing cost as well as advancements in sequencing technologies and bioinformatics analyses have led to an increase in both the number and diversity of genomes that were sequenced. Large-scale, multi-institute projects such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA), Genome 10k Project, Earth Bio Genome Project are few of the several recent initiatives to sequence thousands or hundreds of thousands of isolate organisms. This exponential growth in genomics, compared to other data science fields, has promoted calls for substituting the term 'Astronomical' with 'Genetical' to describe the vast quantities of genomic data being generated Sequencing of cultured isolate microorganisms is important as they serve as references for related, less known organisms. At the same time, a large fraction of the prokaryotic diversity on Earth remains uncultured. Our knowledge about these organisms stems from the analyses of single cells, environmental DNA and metagenome-assembled genomes (MAGs). An example of a large-scale project targeting the uncultured diaspora is the Earth Microbiome Project (EMP), which aims to create a global catalogue of the Earth's uncultured microorganisms. In fact, cultivation-independent sequencing is predicted to outrank the rate at which cultured isolates are being sequenced. Sequence data facilitates comparative analysis and leads to discoveries only when it is accompanied with accurate metadata. This is precisely where GOLD, with its collection of rich and carefully curated metadata, comes in. The data in GOLD is freely available through an easy-to-use web interface. A user can browse through the entire collection of public genome and metagenome projects, study the metadata and download statistics and figures for use in publications and presentations. GOLD Metadata is used in research projects carried out by individual researchers as well as other resources. GOLD geolocation metadata powered Bio Atlas that provides geographic profiles of 16S rRNA sequences from metagenomes, whether they came from geographical and/or host-oriented locations.

GOLD Data Structure:

GOLD is based on a four-level classification system namely Study, Bio sample/Organism, Sequencing Project and Analysis Project. A GOLD Study lies at the top of this hierarchical classification system and broadly describes the overall objective of the sequencing projects that it contains. The physical material collected from the environment is called a Bio sample, while living biological material such as bacteria, fungus, plant or animal is termed as an Organism in GOLD. The sequencing output of a GOLD Bio sample or Organism makes up a Sequencing Project (SP) and the subsequent analysis and data processing methods are described in an Analysis Project (AP). This organization structure ensures that the different aspects of sequencing projects and their related metadata are connected to each other in a coherent manner.

GOLD Study:

The overall research objectives and goals are captured in a Study, which lies at the helm of the four-level organization structure. A Study is similar to the concept of NCBI's umbrella Bio Project and comprises one or many Organism(s) or Bio sample(s) as well as their respective Sequencing and Analysis Projects. Studies can vary in the type of samples collected. For instance, they may include a group of cultured bacteria or soil sample(s) from a rainforest or a mixture of both, provided they answer a common research question. Subsequently, a single Study may have several Sequencing and Analysis Projects that differ in their methodology and application such as Whole Genome Sequencing (WGS) and analysis, metagenome analysis or a combination of both.

Bio sample:

Bio sample is the description of the environment from where the DNA/RNA sample as collected. A Bio sample is described with metadata such as habitat, ecosystem, geographical location, latitude and longitude. Currently GOLD Bio sample entities are defined for all metagenome and meta transcriptome samples, i.e., for non-isolate genomes. DNA or RNA extracted from the same physical sample can be used for metagenome and meta transcriptome projects respectively.

Organism:

Organism is used to describe an individual entity such as a bacterium, fungus, plant, animal or a virus. It can be a cultured isolate of a pure strain of bacterium or an uncultured single-cell isolated using cell sorting. Metagenome-Assembled Genomes (MAGs) associate with a new type of uncultured, non-living Organism in GOLD. All Organisms are required to have basic taxonomic information such as genus, species, strain, NCBI taxonomy ID, phylum etc. Defining an Organism entity is essential to create GOLD Sequencing Projects with sequencing strategies such as Whole genome sequencing and Transcriptome.

Sequencing Project:

The process of generating sequencing data from a Bio sample or Organism is described in a Sequencing Project (SP). GOLD currently has 15 different types of SPs, from which whole-genome sequencing (WGS) and metagenome are most commonly used. The input material for an SP can either be DNA or RNA corresponding to a genome or transcriptome project, respectively. This material can come from either an organism, in the case of WGS and transcriptomes, or from a Bio sample, in the case of metagenomes and meta transcriptomes. A cultured organism can sometimes be sequenced by more than one institution at different times, resulting in multiple SPs for the same organism. The same Organism entity will be the basis for these SPs. In the case of environmental samples, the same Bio sample may be used for both metagenome (DNA) and meta transcriptome (RNA) SPs. Some of the critical metadata present in Sequencing Projects include the type of nucleic acid, sequencing instrument, library method, sequencing institution and funding agency as well as NCBI identifiers such as Bio Project/Bio Sample Accession and SRA Experiment IDs.

GOLD Analysis:

A GOLD Analysis Project describes the assembly and annotation processes that are performed on a Sequencing Project. A user must create a GOLD Analysis Projecting order to submit sequence data to IMG for analysis. The different types of Analysis Projects in GOLD are Genome, Metagenome, Meta transcriptome, Single Cell (Unscreened), Single Cell (Screened), Genome from Metagenome, Transcriptome and Combined Assembly Analysis Projects. A single Sequencing Project may have multiple Analysis Projects. A user has the option to choose which of these Analysis Projects the primary one for analysis is. The remaining Analysis Projects from the same Sequencing Project become designated as reanalysis. Assembly method, gene calling method, sequencing depth, estimated genome size are some of the key metadata fields in an Analysis Project. Submitted data sets may take between 2 and 4 weeks for processing and loading into IMG database, depending upon the number of submissions in queue.

Exploring Working:

1. Users have unrestricted access to projects and associated metadata for research and comparative analysis.
2. The GOLD homepage provides an option to download the complete list of public projects as an Excel file.
3. Users can get a summary of the different types of GOLD Studies, Organisms, Bio samples, Sequencing Projects and Analysis Projects along with their respective counts directly on the homepage.
4. These counts are updated daily and are presented as clickable links. For example, a user can click on the number next to ‘Sequencing Projects’ on the top left corner of the homepage and go to a page with complete list of available Sequencing Projects.
5. This list is sortable and searchable.
6. Using ‘Select Columns for Table’ option, one can configure the list to display more columns of interest such as library method, GC percent, NCBI Bio Project Accession, NCBI Bio Sample Accession etc.

7. Using the magnifying lens on top of each column one can further filter the list based on project status, sequencing strategy, project name etc.

Application:

1. GOLD database has brought about reduction in sequencing cost as well as an advancement in sequencing technologies and bioinformatics analysis which has led to an increase in both number and diversity of a genome that were sequenced.
2. Sequencing of a cultured microorganisms is important as they serve as references for related, less known organism were many of the uncultured organism diversity can know because of sequencing.
3. Sequenced data facilitates comparative analysis and leads to discoveries only when it is accomplished with accurate metadata.
4. The data in GOLD is freely available through an easy-to use web interface where a user can browse through the entire collection of public genome and metagenome projects, study the metadata and download statics and figures for use in publications and presentations.
5. The growth number of and variety of sequencing projects and analysis strategies are promising in term of providing better results into research questions and hypothesis testing.

METHODOLOGY:

1. Open GOLD database.
2. As per the objectives of study various options are selected such as sample, Bio project, sequencing projects, Analysis projects etc. to check the information on the desire query.
3. Filters are applied to refine the results via advanced search or metadata search.
4. New projects or new submission can also be done via LOGIN in option on the homepage, user ID and password is created and then submission are done by using all the relevant information asked in the section.

OBSERVATIONS:

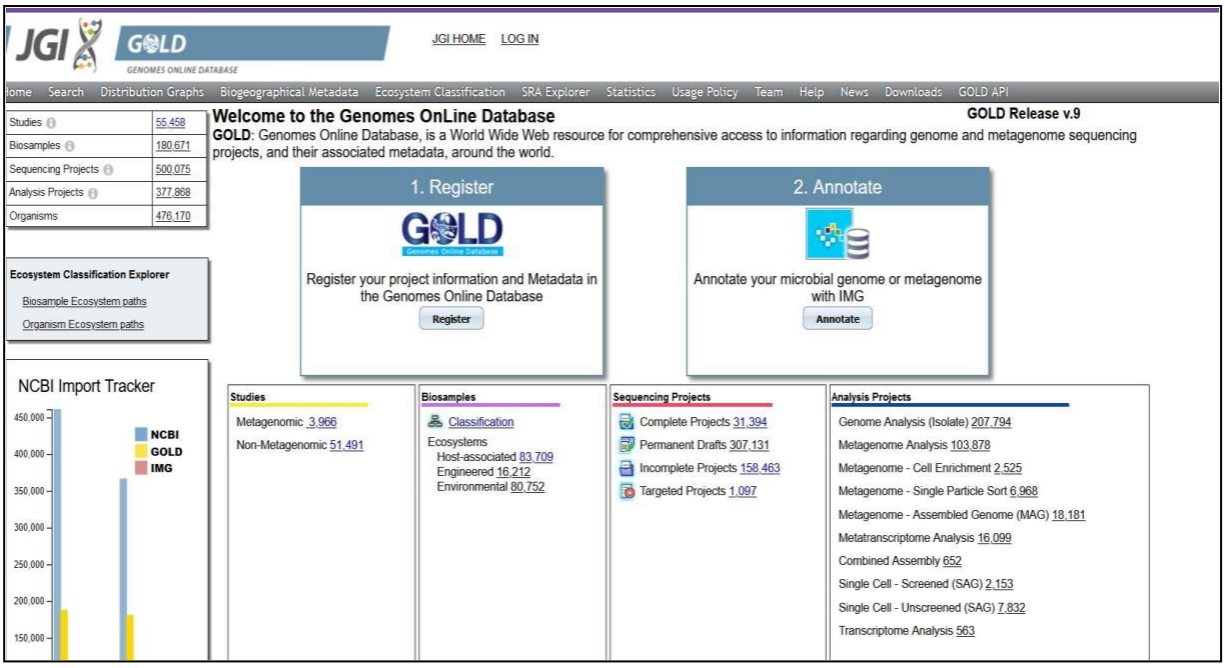


Figure 1: Homepage of the GOLD database

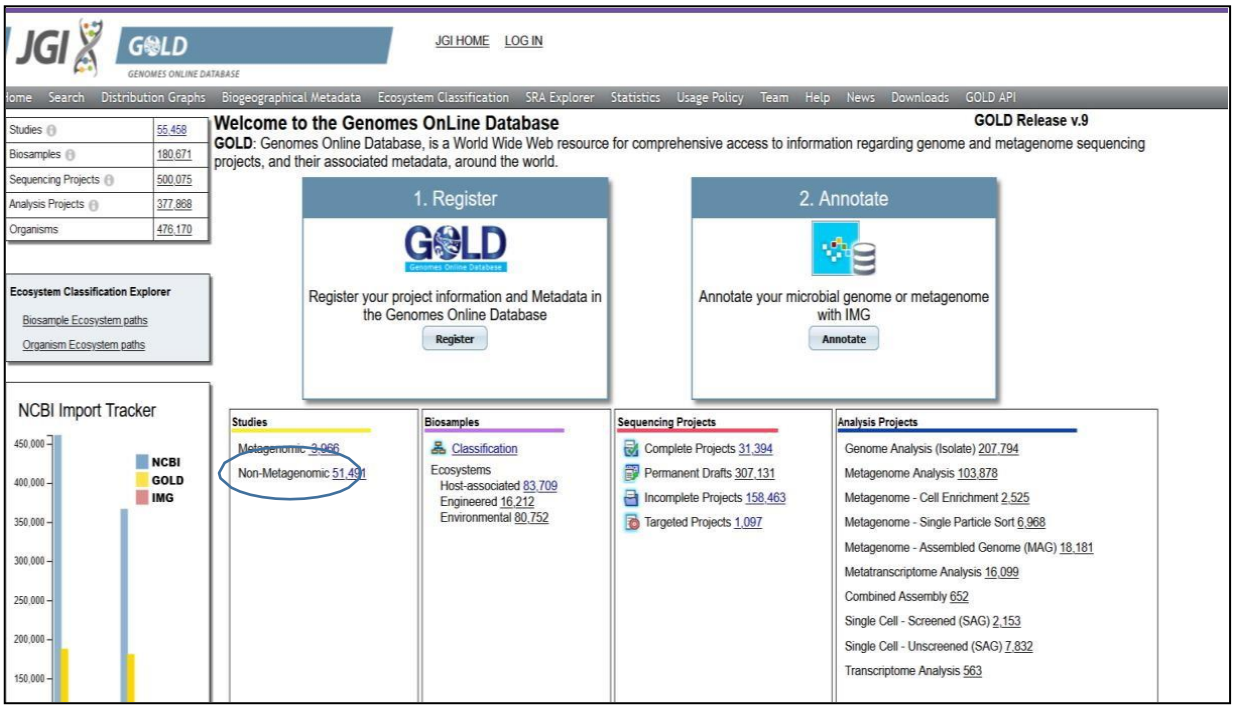


Figure 2: Selecting study section from Homepage of GOLD database

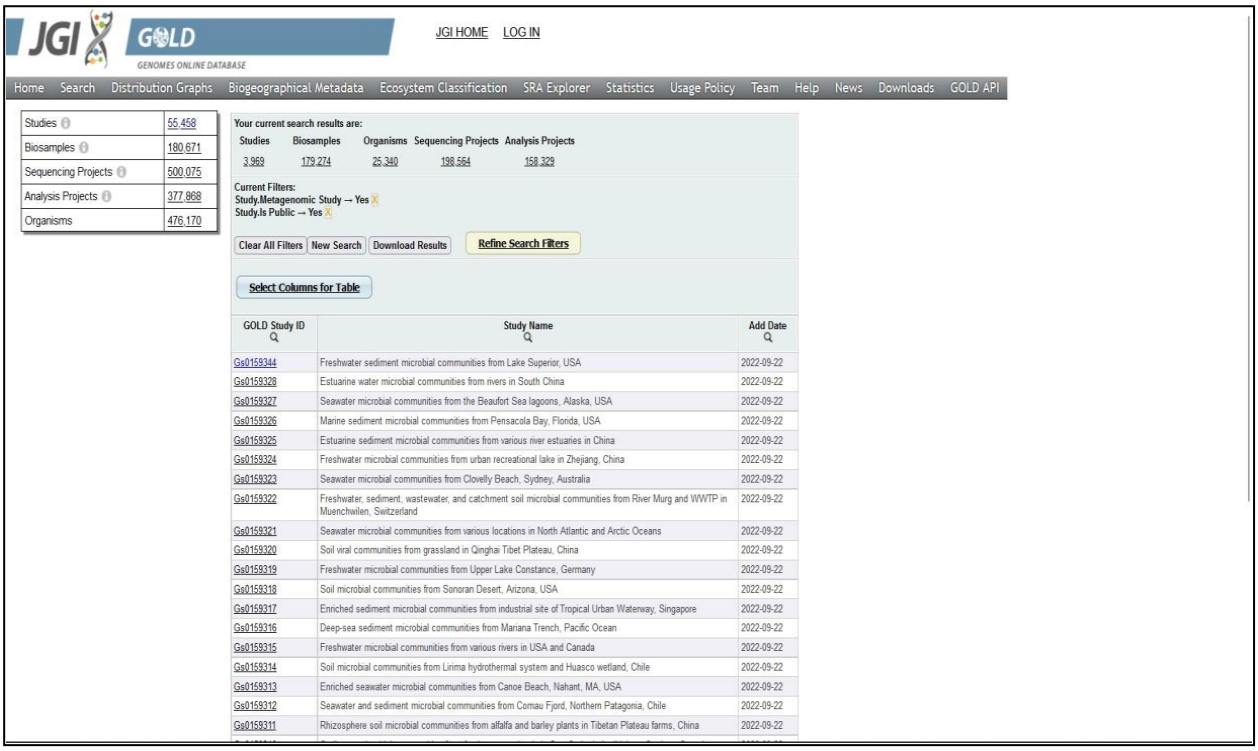


Figure 2a: Result of study section from GOLD database

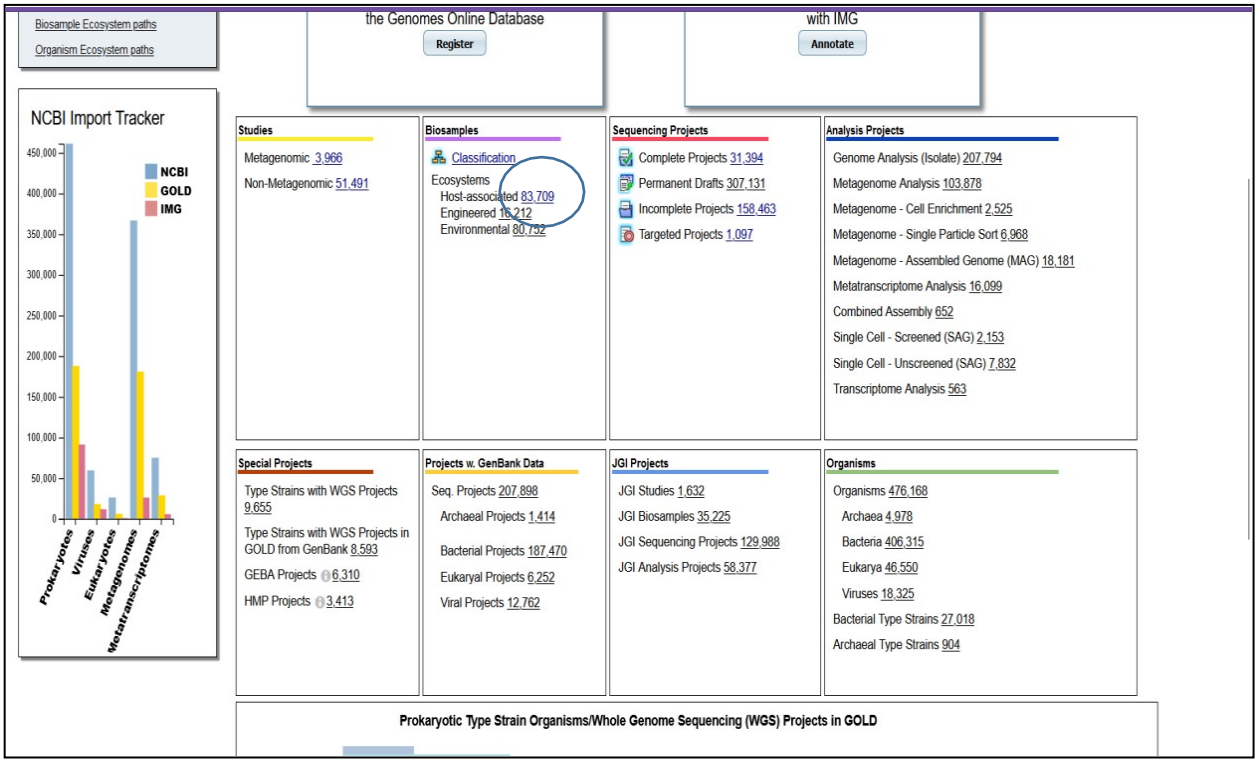


Figure 3: Selecting Bio sample section from Homepage of GOLD database

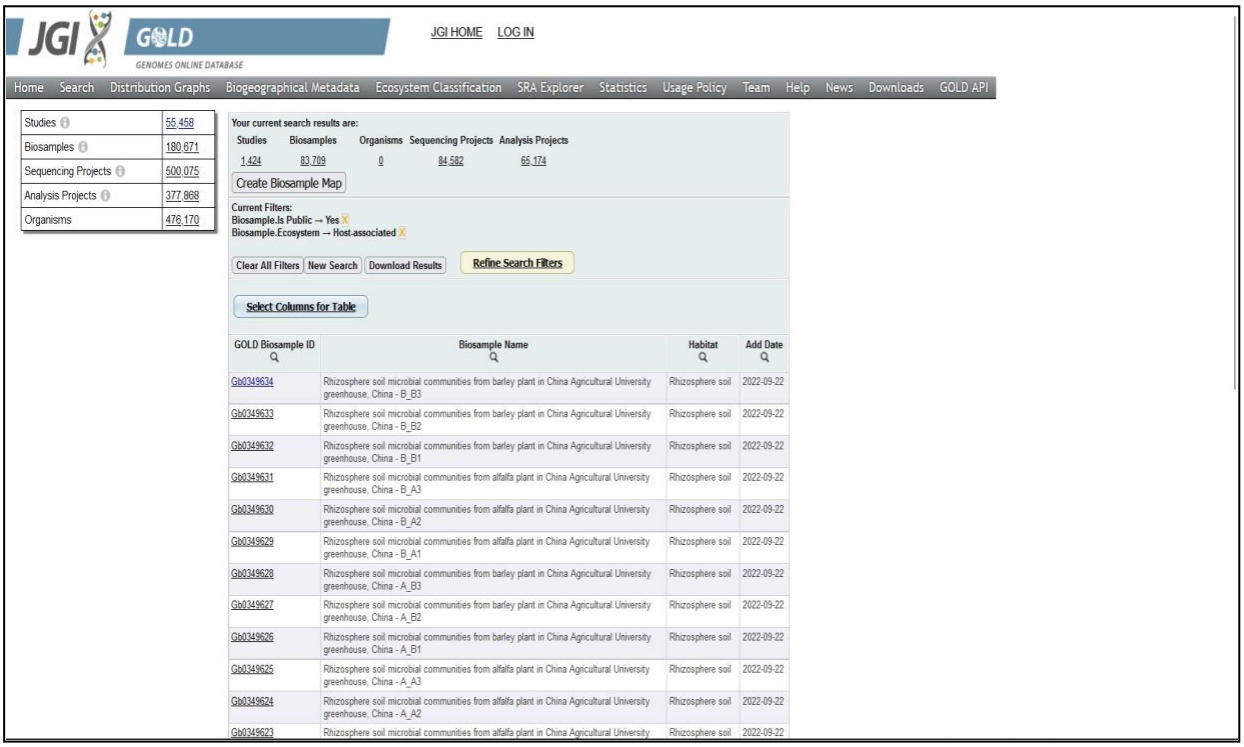


Figure 3a: Result of Bio sample section from GOLD database

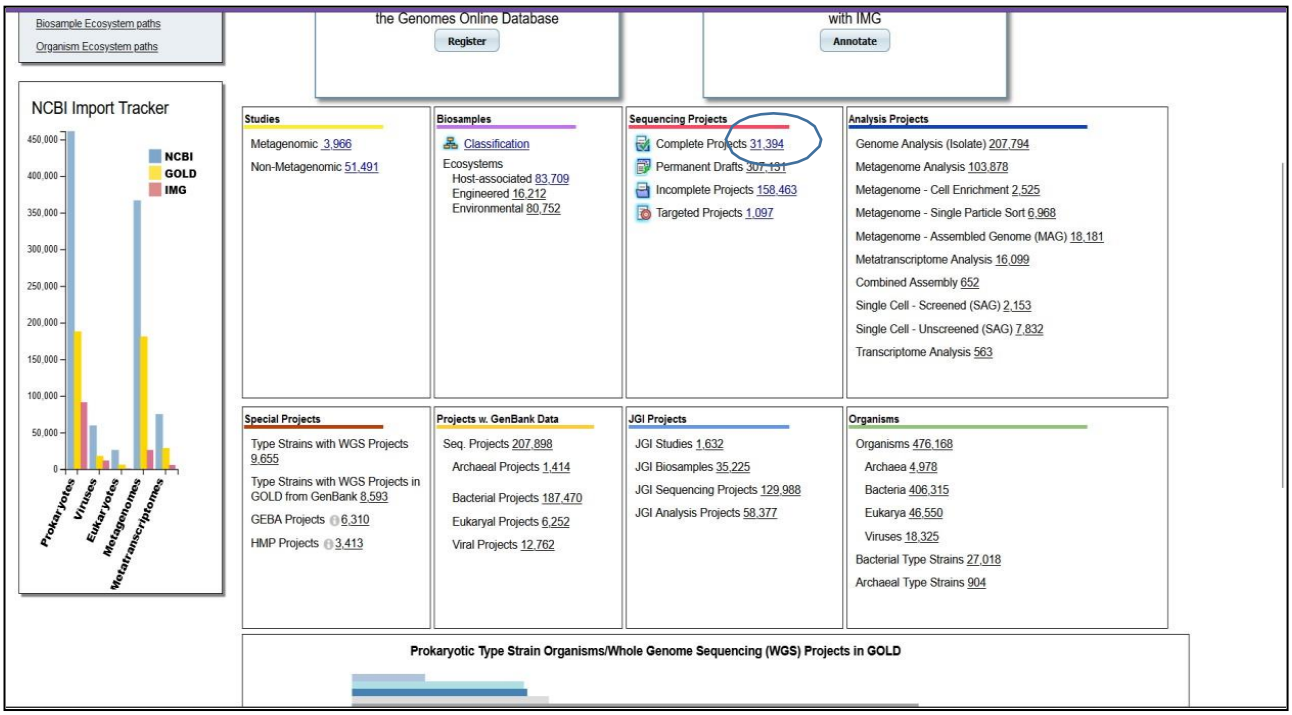


Figure 4: Selecting 'Complete Sequencing Project' section from Home page of GOLD database

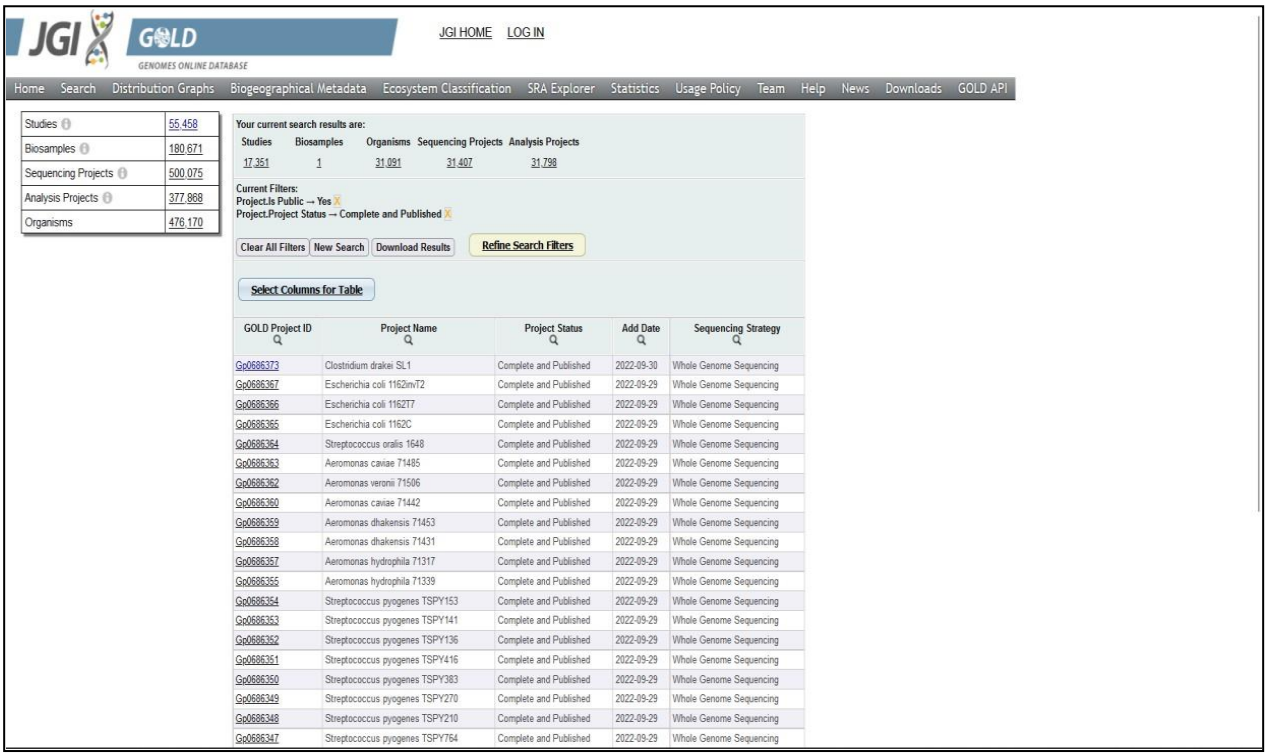


Figure 4a: Result of 'Complete Sequencing Project' section from GOLD database

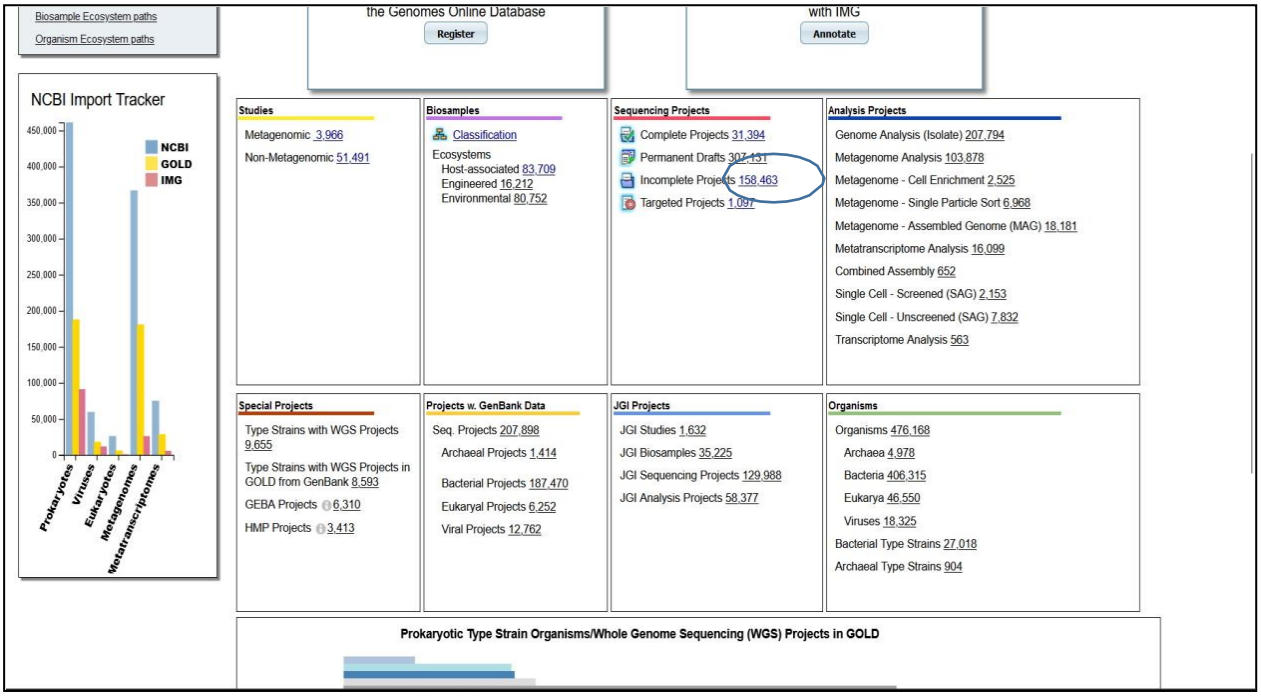


Figure 5: Selecting 'Incomplete Sequencing Project' section from Homepage of GOLD database

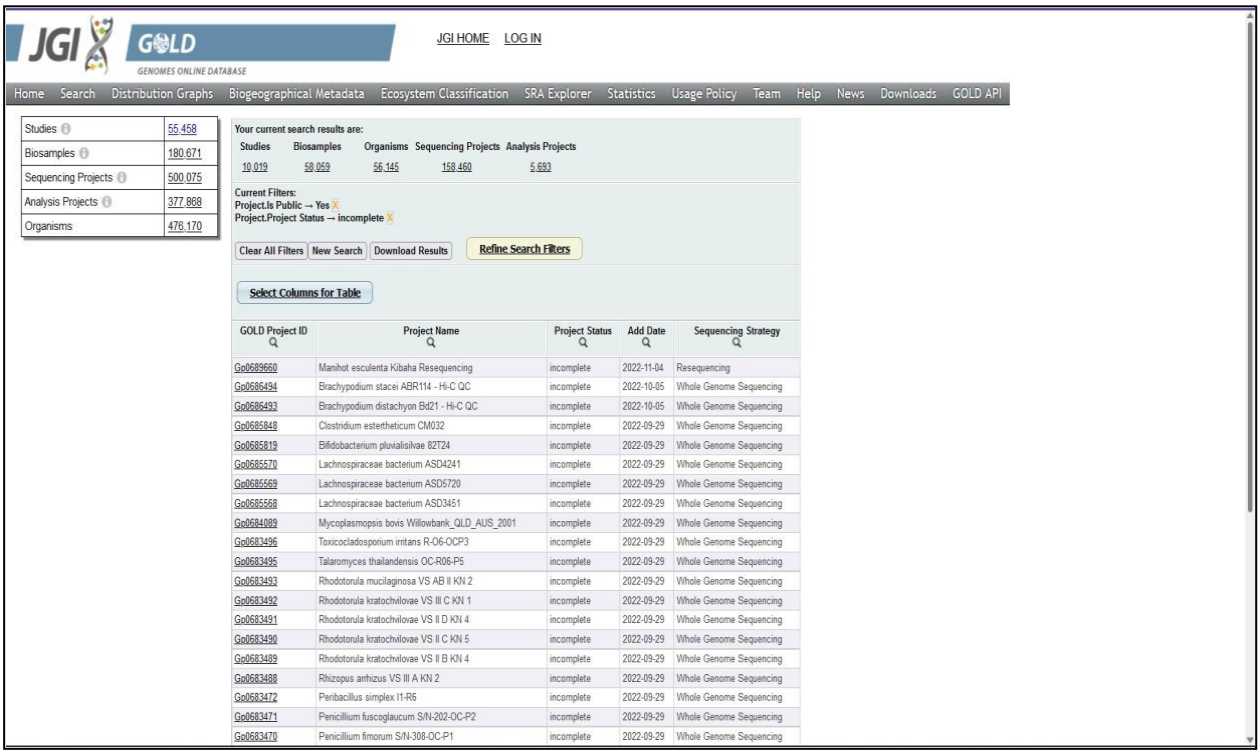


Figure 5a: Result of 'Incomplete Sequencing Project' section from GOLD database

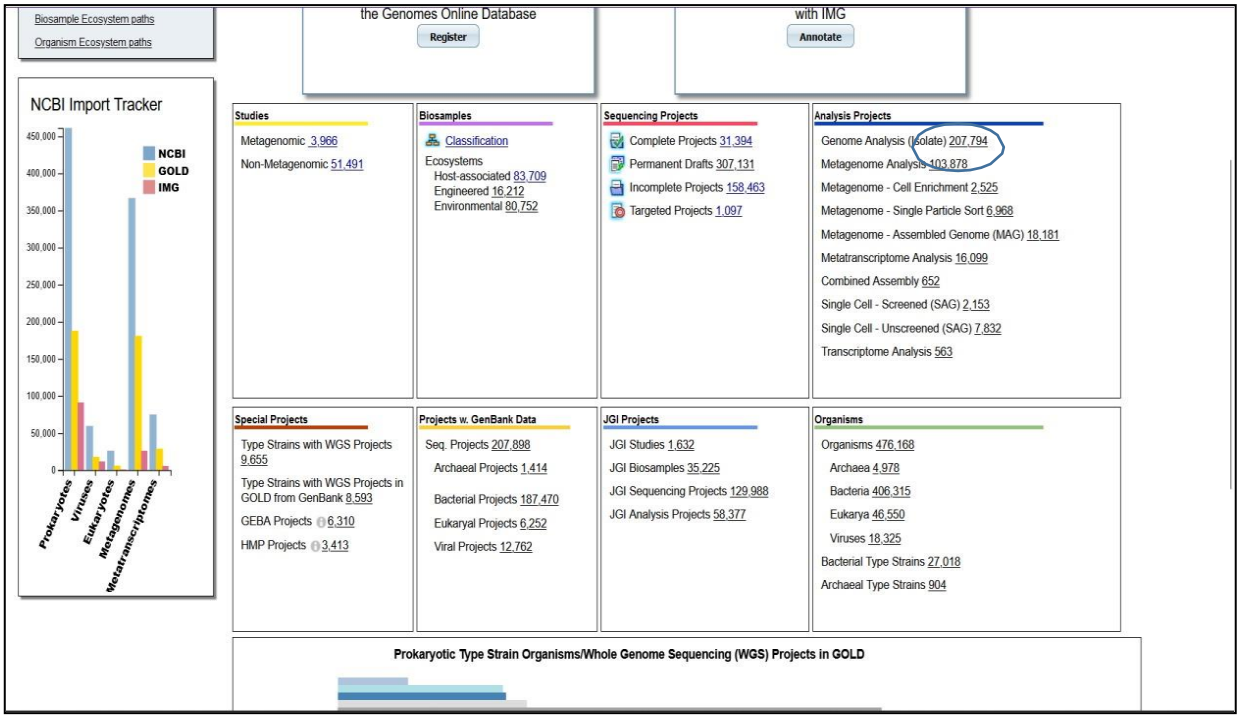


Figure 6: Selecting 'Analysis Project' section from the Homepage of GOLD database

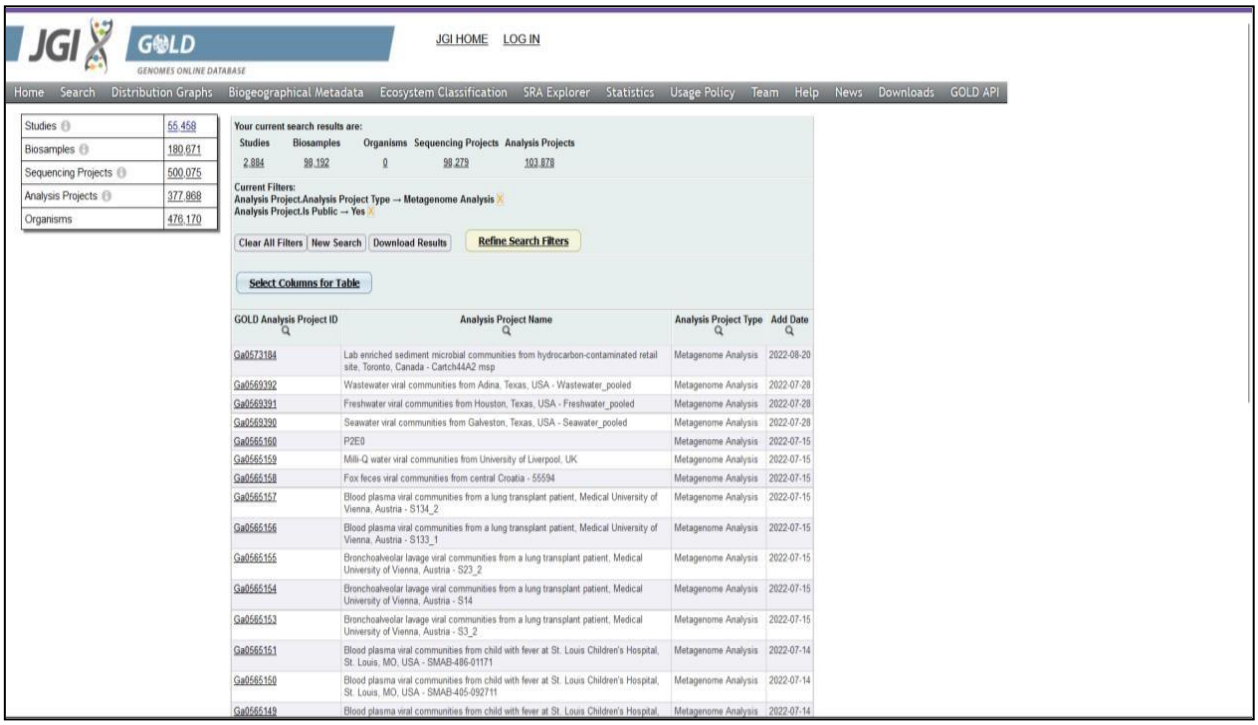


Figure 6a: Result of 'Analysis Project" section from GOLD database

RESULTS:

The Genome Online Database (GOLD) which is an online open resource maintains an up-to-date catalo of genome and metagenome project in the context of comprehensive list of associated metadata. It serves as a major resource that catalogues and monitors genome and metagenome projects around the world in various sections. As data in GOLD is freely available through an easy-to-use web interface user can browse through the entire collection of public genome and metagenome projects, study the metadata and download statistics and figures for use in publications. Also new submission can be submitted in the database where data is carefully curated and available for public use.

CONCLUSION:

Thus, GOLD provides a user-friendly web interface to browse sequencing projects and launch advanced search tools to retrieve the desired results. Therefore, it serves as a doorkeeper for sequencing projects for analysis and guides to gather comprehensive information available in database.

REFERENCES:

1. Bernal, A. (2001, January 1). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research*, *29*(1), 126–127. https://doi.org/10.1093/nar/29.1.126
2. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., Thomas, A. D., Ali, R., Sharma, K., Kyrpides, N. C., & Reddy, T. (2016, October 27). Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, *45*(D1), D446–D456. https://doi.org/10.1093/nar/gkw992

3. Gold, D. B., & Wegner, D. M. (1995, July). Origins of Ruminative Thought: Trauma, Incompleteness, Nondisclosure, and Suppression. *Journal of Applied Social Psychology*, *25*(14), 1245–1261. https://doi.org/10.1111/j.1559-1816.1995.tb02617.x
4. Anderson, D. R., Serxner, S. A., & Gold, D. B. (2001, May). Conceptual Framework, Critical Questions, and Practical Challenges in Conducting Research on the Financial Impact of Worksite Health Promotion. *American Journal of Health Promotion*, *15*(5), 281–288. https://doi.org/10.4278/0890-1171-15.5.281
5. Liolios, K., Mavromatis, K., Tavernarakis, N., & Kyrpides, N. C. (2007, December 23). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, *36*(Database), D475–D479. https://doi.org/10.1093/nar/gkm884

DATE: 08/11/23

WEBLEM 6(B)
INTRODUCTION TO GENOMICS AND ITS VARIOUS BROWSERS
(URLs: UCSC: https://genome.ucsc.edu
ENSEMBL: http://www.ensembl.org/index.html?redirect=no**)**

AIM:
To study various genomic databases like UCSC and ENSEMBL.

INTRODUCTION:

A. UCSC:
Since the debut of the UCSC Genome Browser in 2001, the web-based data visualization tool has served as a digital microscope to cross-reference, interpret and analyze genome assemblies.

From base pairs to contigs to chromosomes, the visualization tool allows for genome annotations to be positioned alongside the genomic DNA itself for a large number of vertebrate species and other clades of life. In this era of big data, the UCSC Genome Browser team aspires to quickly incorporate and contextualize vast amounts of genomic information. Apart from incorporating data from researchers and consortia, the Browser also provides tools available for users to view and compare their data with ease. Custom tracks allow users to quickly view a dataset, and track hubs allow users to extensively organize their data and share it privately using a URL. Saving a session and sharing the session URL with a colleague allows easy access to the pre-configured views of an interactive Browser image. Public data access also enables creators to submit their hub to our list of available 'public hubs 'or 'public sessions'.

Accessing the underlying track data can be achieved in a variety of ways. The Table Browser and RESTful API are useful to extract data from a region in many file formats such as BED or wiggle. The public MySQL server allows users to query data tables directly, and table dumps are available on the download server (https://hgdownload.soe.ucsc.edu/downloads.html) to enable bulk download and local processing of information in our database tables. Binary indexed files, liftOver files, and other large files can be found in the /gbdb/ directory hierarchy on the download server (https://hgdownload.soe.ucsc.edu/downloads.html#gbdb).

Currently, 211 genome assemblies are available on the UCSC Genome Browser, representing 107 different species. In early 2020, as a response to the urgency of supporting biomedical research for COVID-19, the SARS-CoV-2 genome assembly was released along with relevant biomedical datasets. With the growing number of datasets related to the RNA genome causing the pandemic, a COVID-19 landing page (https://genome.ucsc.edu/covid19.html) was created to consolidate and serve as a directory for certain information and research resources. Given the constant production of new datasets from researchers around the world, the UCSC Genome Browser team has added

support for new data types and several new display features, some of which have been suggested by the user community. New features including Hi-C, vcfPhasedTrio, and bigDbSnp data visualizations are designed to assist in the interpretation of genetic variants in clinical and research settings. As always, all data and software are freely available for personal, non- profit, and academic research use. Updating existing data tracks and displaying new annotations is a key goal for the UCSC Genome Browser team as a means to better serve the genomics community. The addition of new vertebrate genome assemblies ensures that new sequences are incorporated into the Browser as consortia work to resolve gaps, repetitive regions, and update chromosome assemblies. A total of four genome assemblies have been added to the Genome Browser within the last year; two of these are new to the Browser. In collaboration with the Monterey Bay Aquarium, the genome assembly for Gidget, a southern sea otter (enhLutNer1), was created and released. The other new genome assembly was the coronavirus, SARS-CoV-2 (wuhCor1), released as part of the effort to consolidate sequence and annotation information in one place for the virus and vaccine research communities. The assemblies for a horse (equCab3), rhesus macaque (rheMac10), and gorilla (gorGor6) were updated.

B. <u>ENSEMBL:</u>

Ensembl provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species. Ensembl provides genes and other annotation such as regulatory regions, conserved base pairs across species, and mRNA protein mappings to the genome. These data are accessible via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interface (Perl API). Ensembl comparative analyses, variation mappings and gene determinations are freely available to the scientific community within the context of genomic assemblies. The Ensembl gene set reflects a comprehensive transcript set based on protein and mRNA evidence in UniProt and NCBI RefSeq databases.

Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute that annotates chordate genomes (i.e., vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and fly are also imported for comparative analysis by the Ensembl ‘compara’ team. Most annotation is updated every two months, leading to increasing Ensembl versions, however the gene sets are determined on the order of once a year. A new browser at www.ensemblgenomes.org is now being set up to access non-chordates such as bacteria, plants, fungi and more.

Ensembl strives to provide the most accurate and up-to-date gene set possible. If available, manually curated datasets are imported, such as the SGD (Saccharomyces Genome Database) gene set for Saccharomyces cerevisiae, the WormBase gene set for C. elegans, and the VEGA/Havana set for Homo sapiens. The VEGA (vertebrate genome annotation) consortium provides manual annotation of vertebrate genomes, focusing on regions in human, mouse, zebrafish, pig and dog. For species where manually curated evidence is not available, Ensembl annotates the gene set using a gene prediction pathway (or annotation pipeline). This is termed as the genebuild, which determines the Ensembl gene set using biological evidence, namely mRNA and protein information in databases such as

UniProt/Swiss-Prot and annotated entries in RefSeq. Every resulting gene is based on at least one mRNA or protein, and in most cases, one Ensembl gene has been determined using multiple pieces of evidence from comprehensive biological databases. The Ensembl annotation pipeline is carefully followed by the genebuild team. A typical genebuild is performed over weeks, resulting in the Ensembl gene set of 'known' and 'novel' genes for a species.

A genome sequence provides a natural framework about which to organise biological data. In the short time in which genome sequences have been available, genome databases have proved invaluable resources to researchers. In the case of human, the range of existing biological data and the types of researchers is even wider than for other organisms, stretching from clinical genetics to molecular biology. The availability of the draft human genome sequence enables these huge amounts of data, ranging from records of disease in our species to the sequences of related organisms, to be brought together systematically for the first time. The Ensembl project is actively addressing this by providing a database of human genome annotation (http://www.ensembl.org/). This is being continuously expanded to include an increasing range of data types (vertical integration) as well as to build comparative genome sequence views as sequences of vertebrate genomes, such as mouse, rat and zebrafish, become available (horizontal integration). The database is being built on a very general and carefully engineered software framework that is being developed in parallel with the data integration. By making all software freely available and designing the system to be completely portable, Ensembl aims to provide a bioinformatics framework that is easy to apply to different organisms and types of data.

METHODOLOGY:

1. **UCSC:**
   a. Access the UCSC Genome Browser.
   b. On the homepage in the 'Our tools' section, click on the 'Genome Browser'. UCSC Browser Gateway page will open.
   c. Select/Set the assembly to → Dec.2013 (GRCh38/hg38). Hit "GO" without entering a search term to go to a default location.
   d. It will open the genome browser view. Displaying assembly title, ideogram, genome view, track categories, etc.
   e. Options are provided to 'move', 'zoom in/out', and configures the assembly as per user requirements.
   f. By dragging and selecting the tracks user can explore through other various options.
   g. In the search box by entering specific search parameters (such as gene name, amino acid position, RefSeq accession, SNPid, cytological band, chromosome coordinate, etc.) user can navigate through the assembly per the study requirement.
   h. The current working assembly images can also be downloaded and exported in PDF/EPS format by clicking on the 'View' option and selecting 'PDF/PS'.

2. **ENSEMBL:**
   a. Open the Ensembl database server.
   b. Enter the keywords (BRCA) in the search tab and click on go option.
   c. From the hit page, select the entry of interest.
   d. Explore the options available in the result page.

OBSERVATIONS:

A. UCSC:

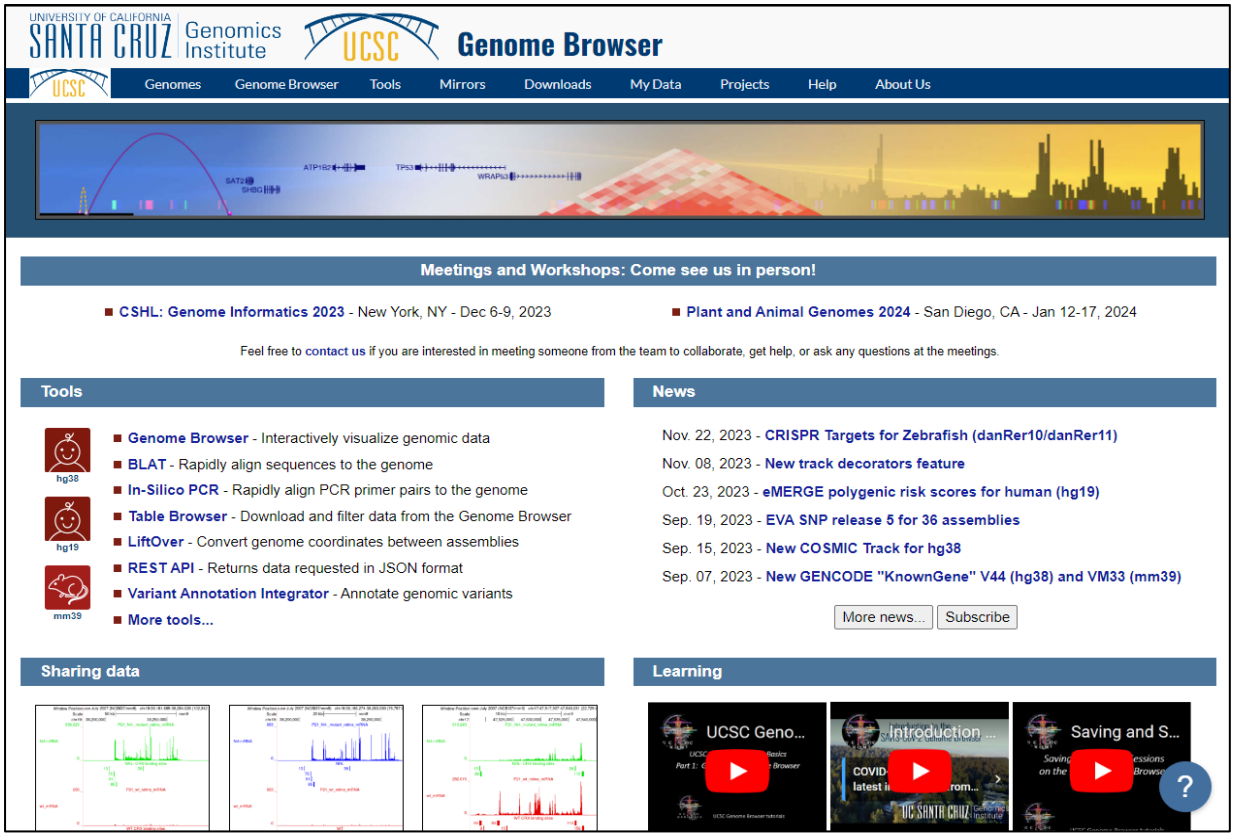


Figure 1: Homepage of UCSC genome browser

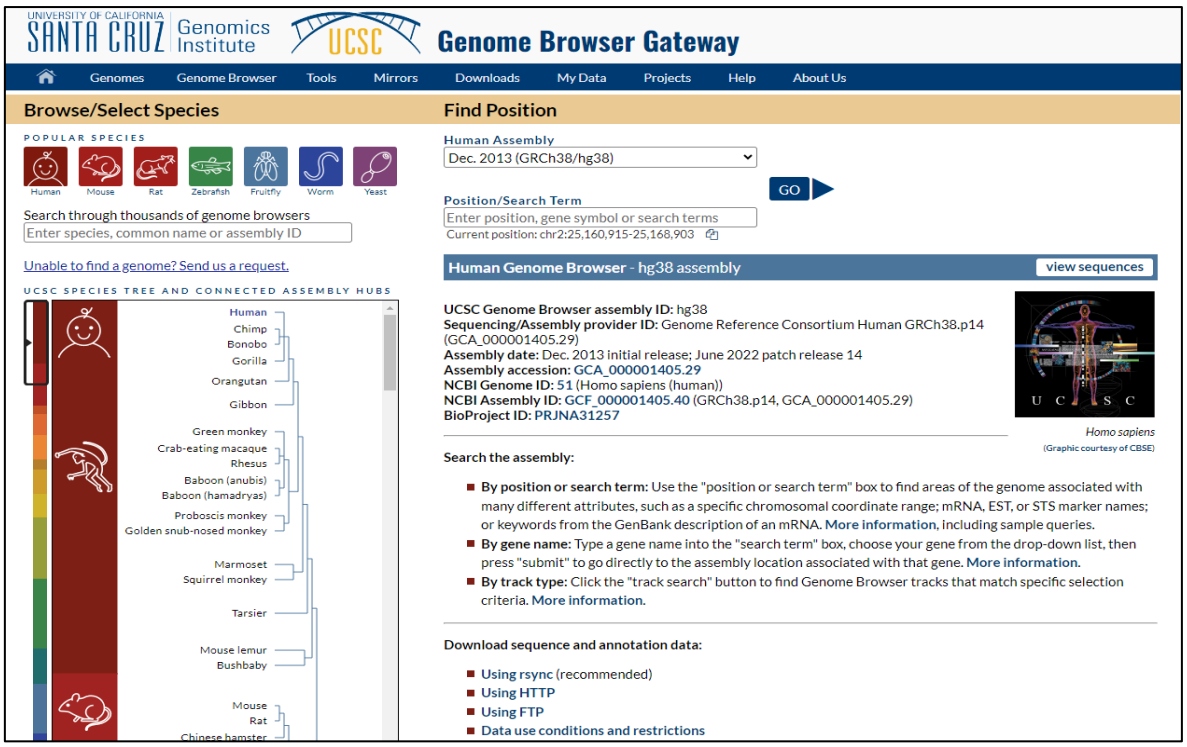


Figure 2: Gateway: Start page [Selecting the genome browser assembly 'Dec. 2013 (GRCh38/hg38)']

Figure 3: View of Genome browser (Navigating through set assembly 'hg38')

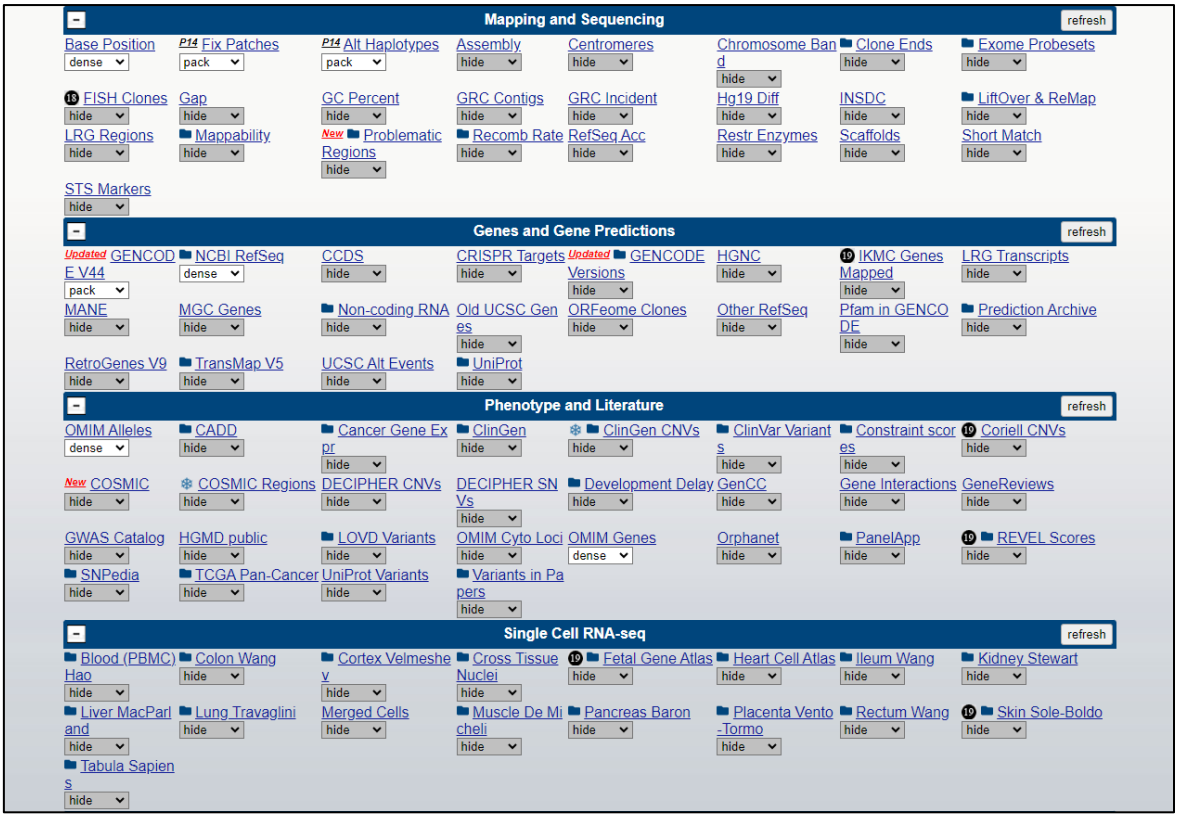


Figure 4: Track Categories

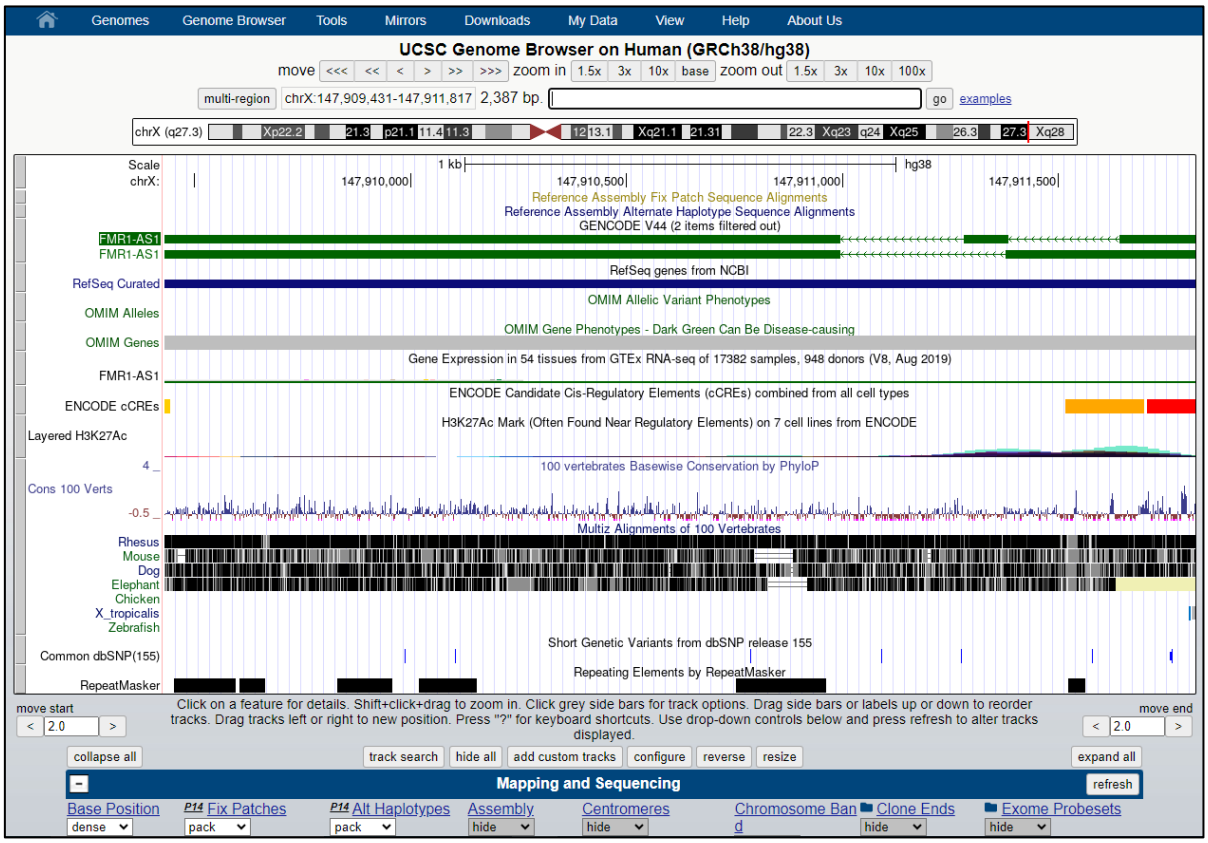


Figure 5: Searching assembly via gene name 'FMR1-AS1'

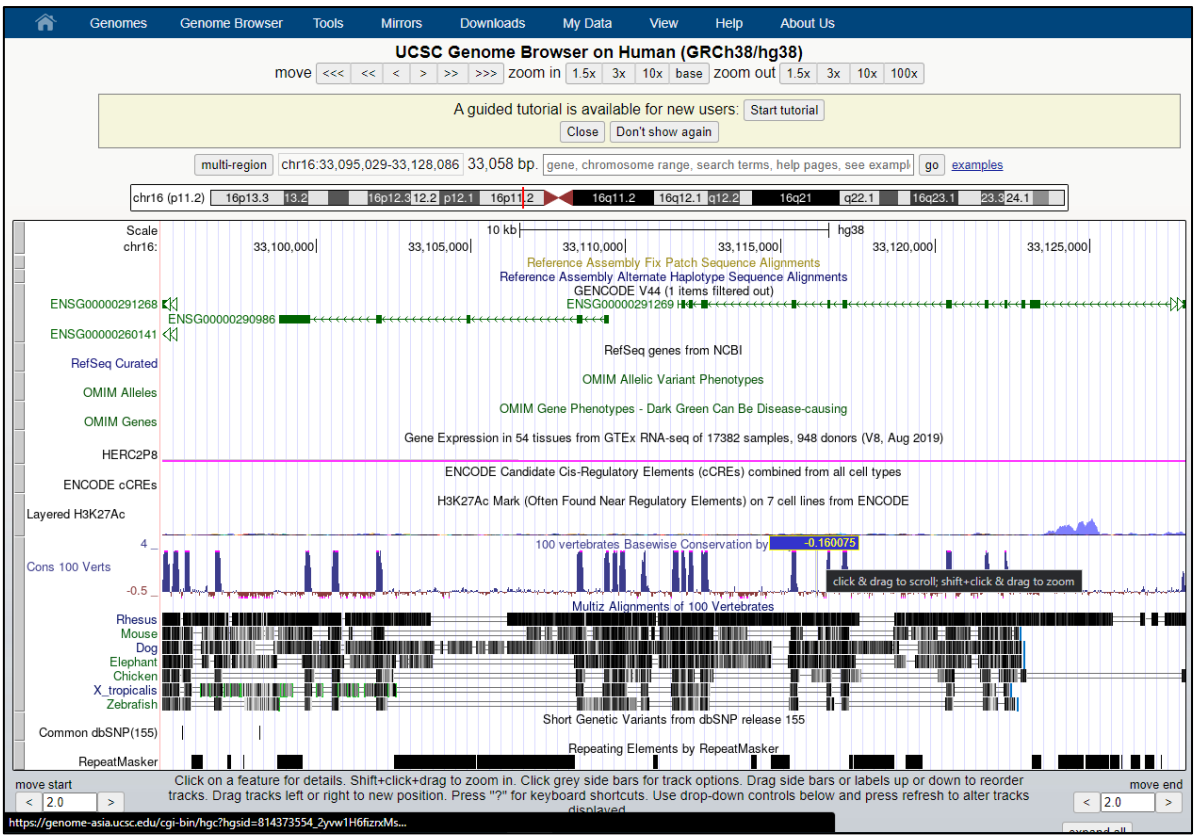


Figure 6: Searching assembly via amino acid position 'HERC2P8'

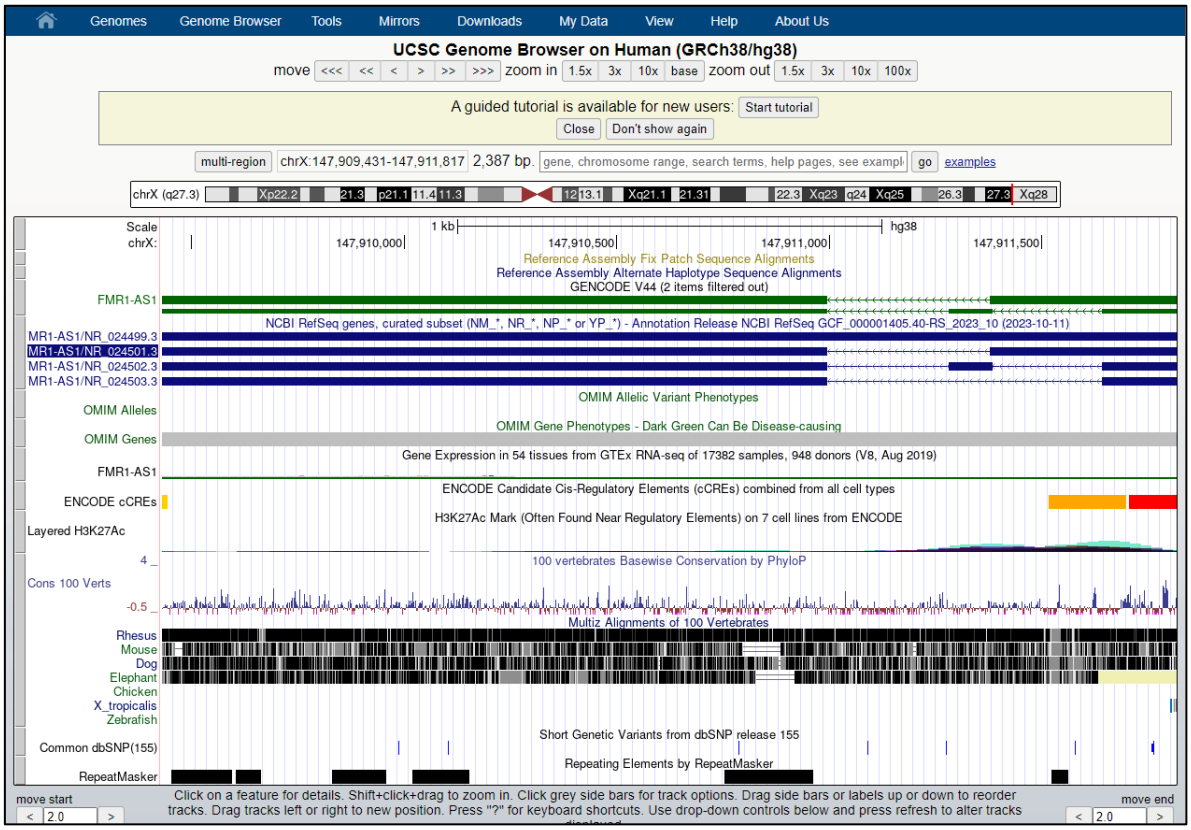


Figure 7: Searching assembly via RefSeq accession 'NR_024501'

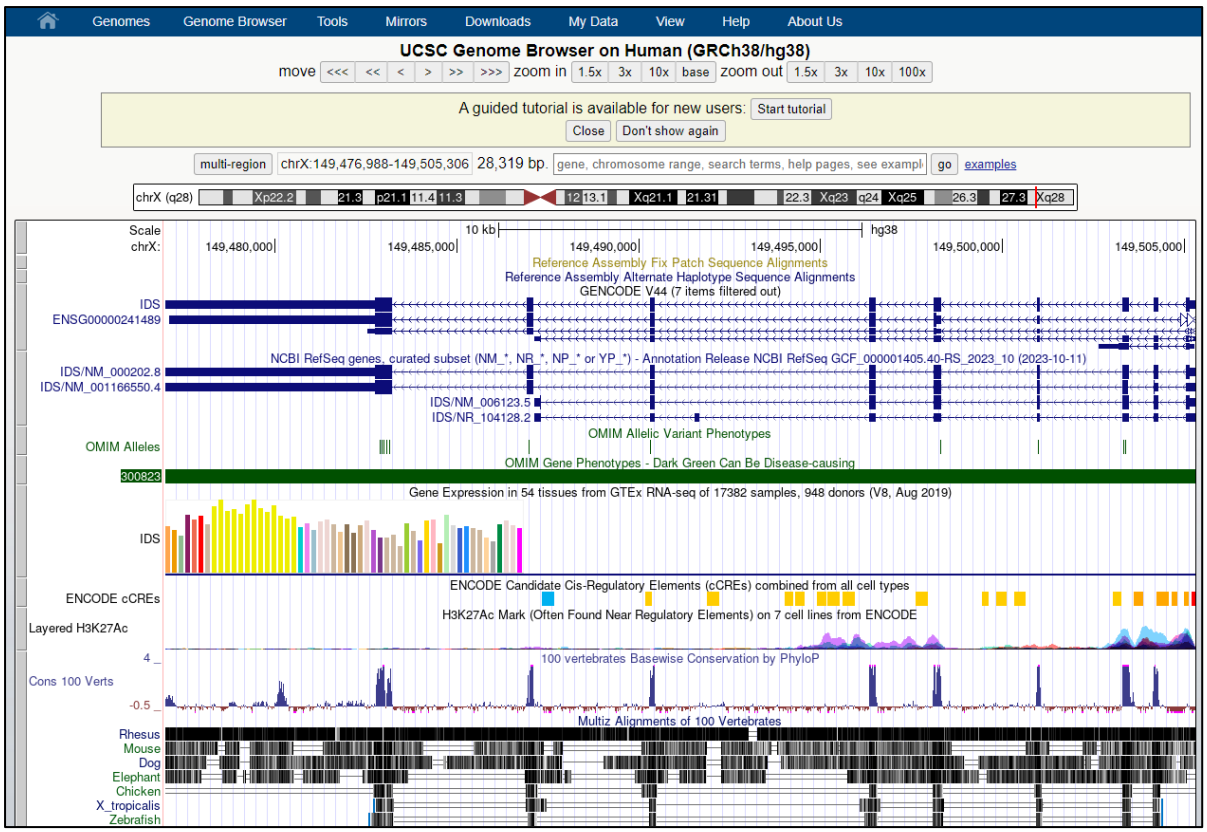


Figure 8: Searching assembly via OMIM identifier '#300823'

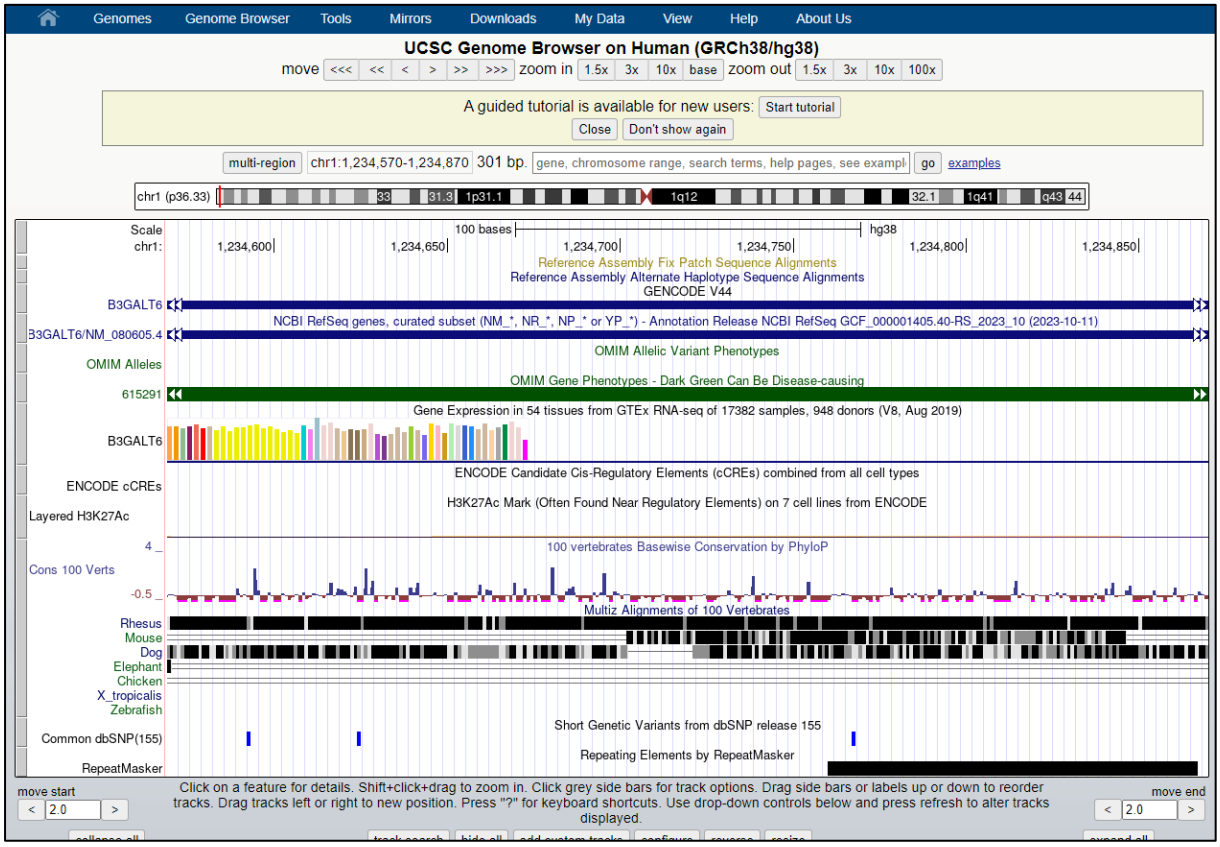


Figure 9: Searching assembly via genomic co-ordinates 'chr1:1234570-1234870'

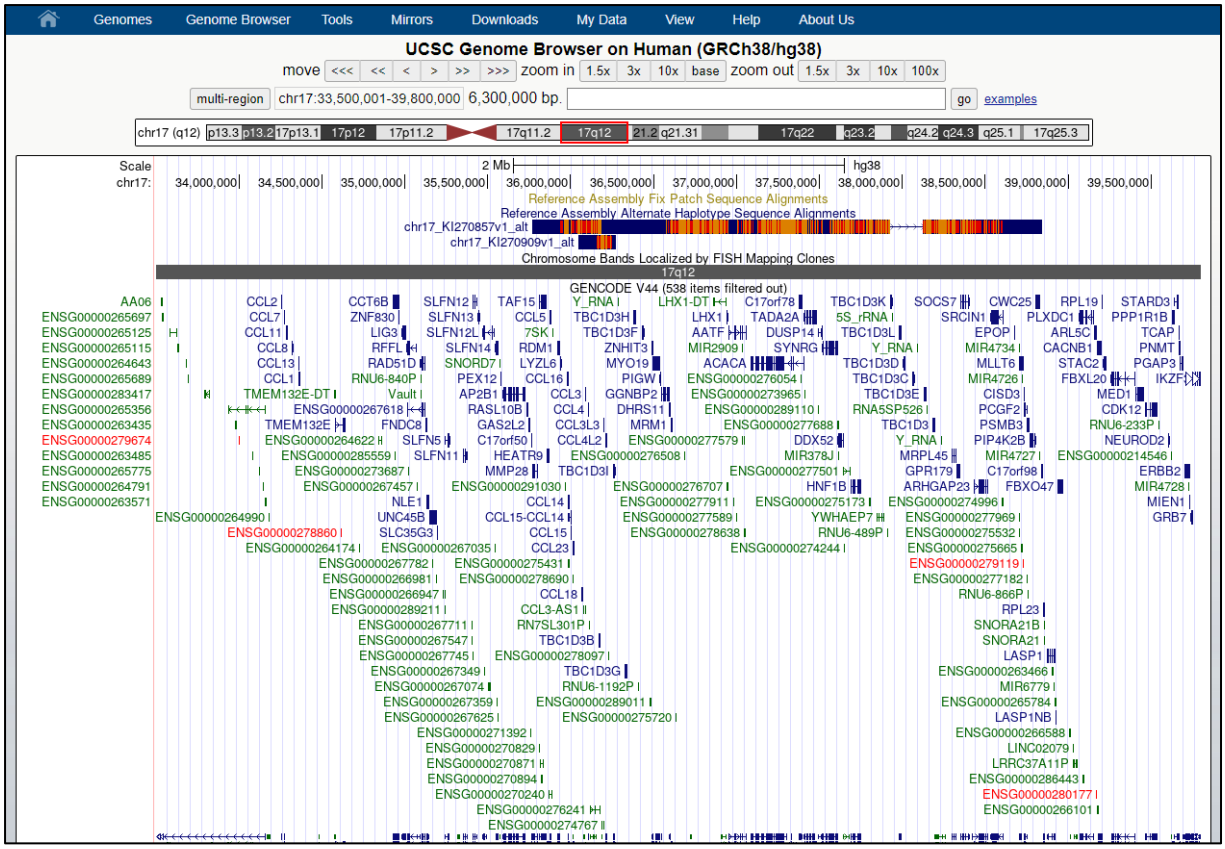


Figure 10: Searching assembly via cytological band '17q12' (Gene information, Coding/Non-coding regions)

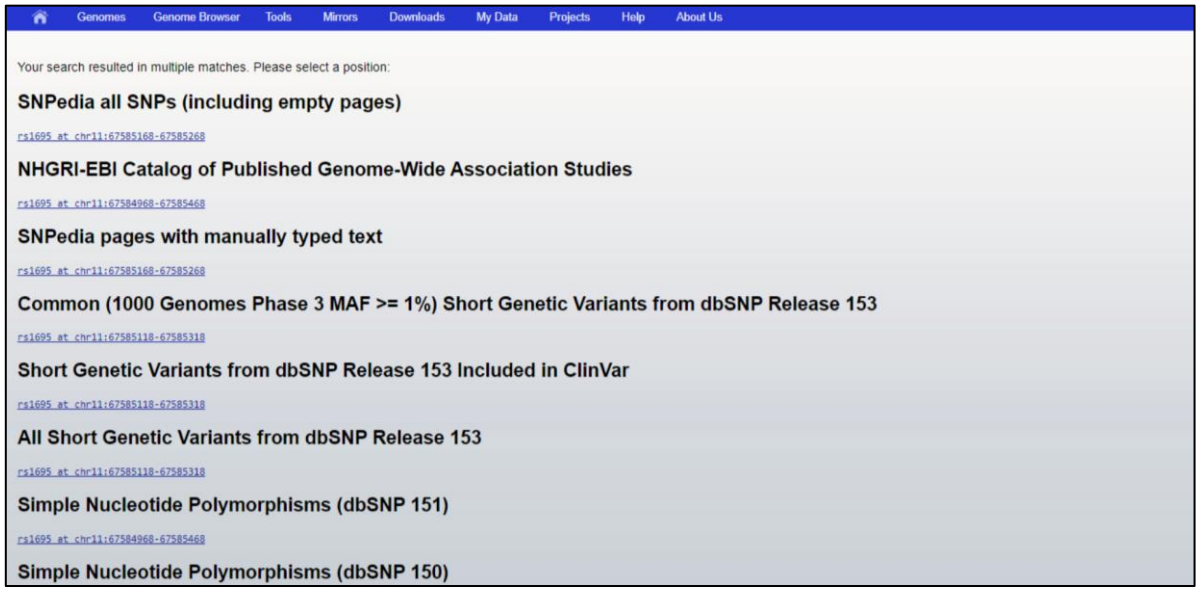


Figure 11a: Searching assembly via SNP ID 'rs1695' (Results- multiple matches)

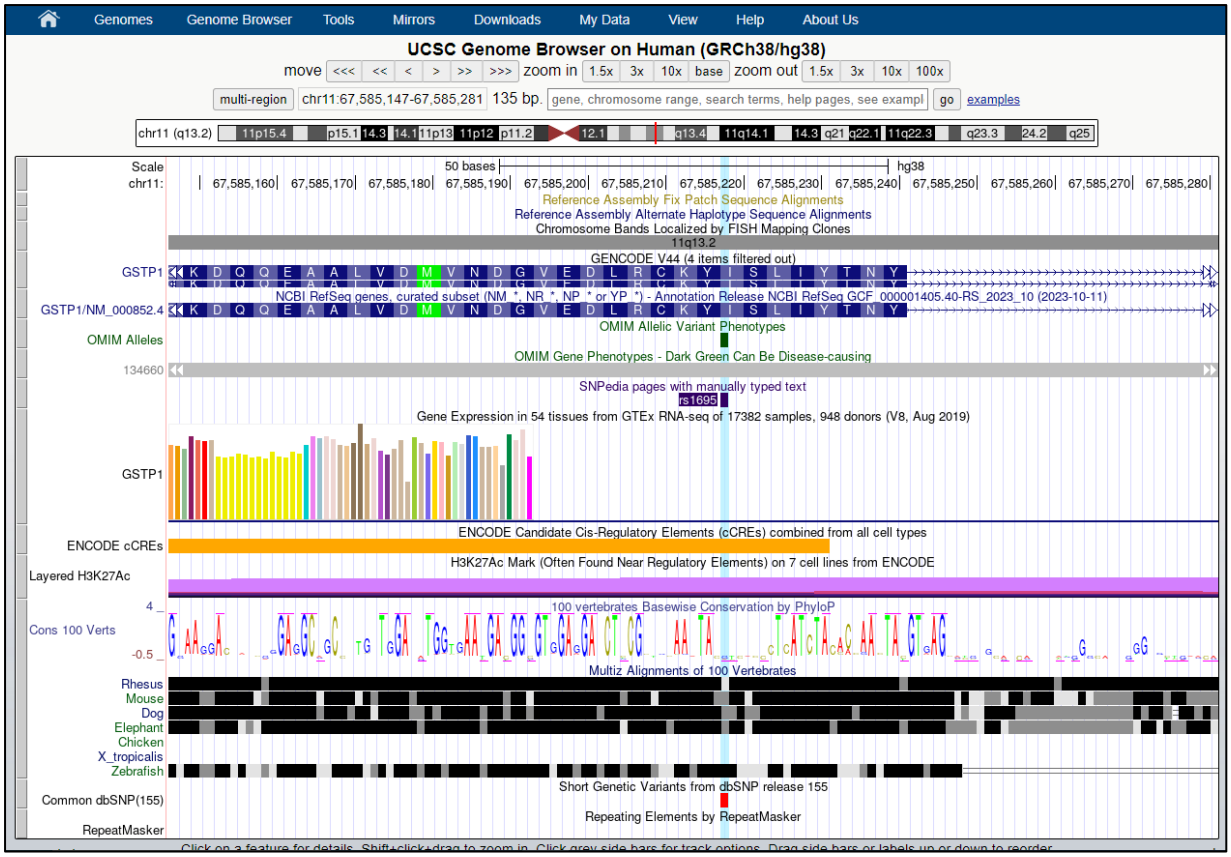


Figure 11b: Searching assembly via SNP ID 'rs1695' (Results for – 'rs1695' at chr11:67585168-67585268)

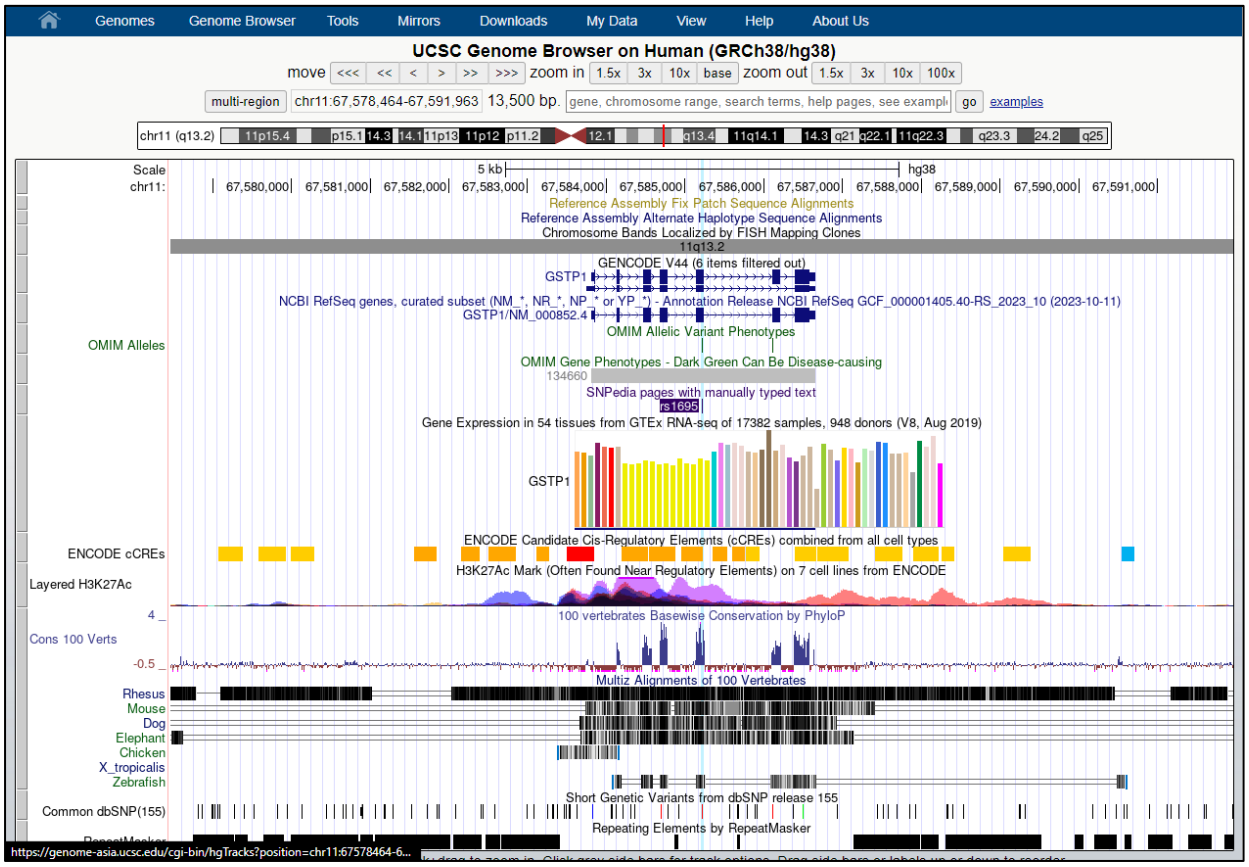


Figure 12: Zooming option applied (zoom out – 100X)

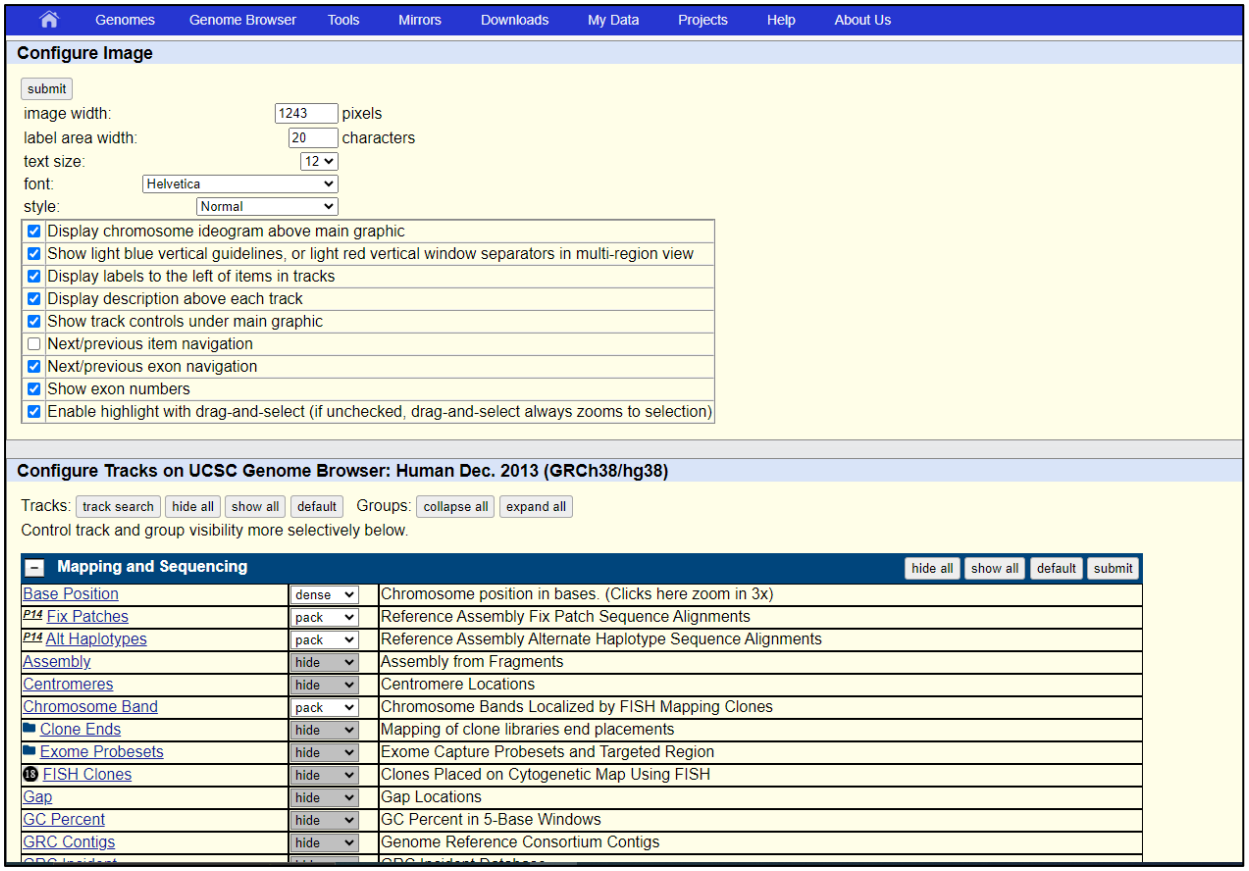


Figure 13: Configuring Tracks

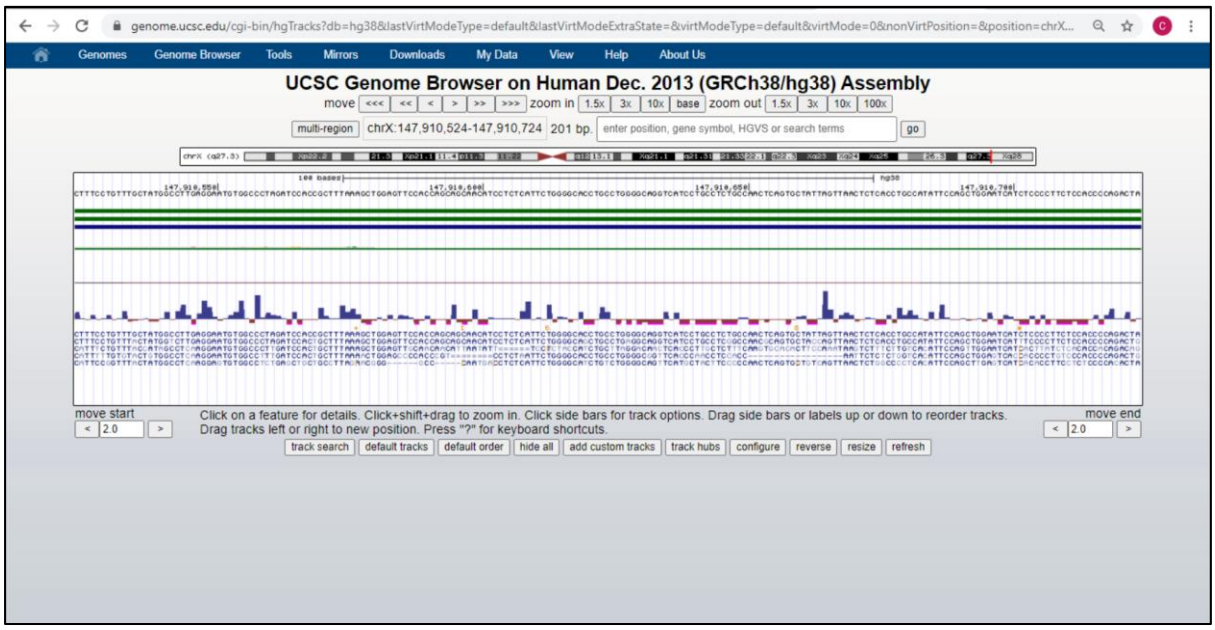


Figure 13a: Tracks after configuration

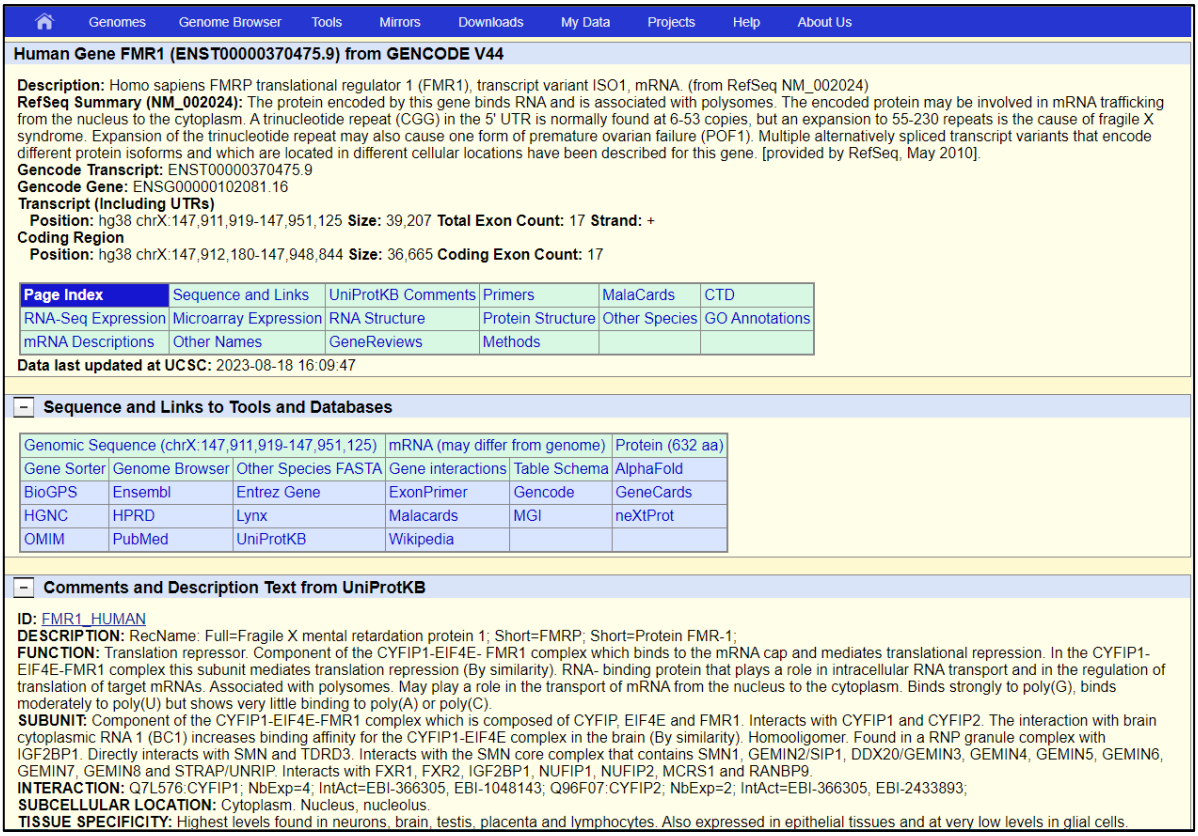


Figure 14: Track Details (Description Page)

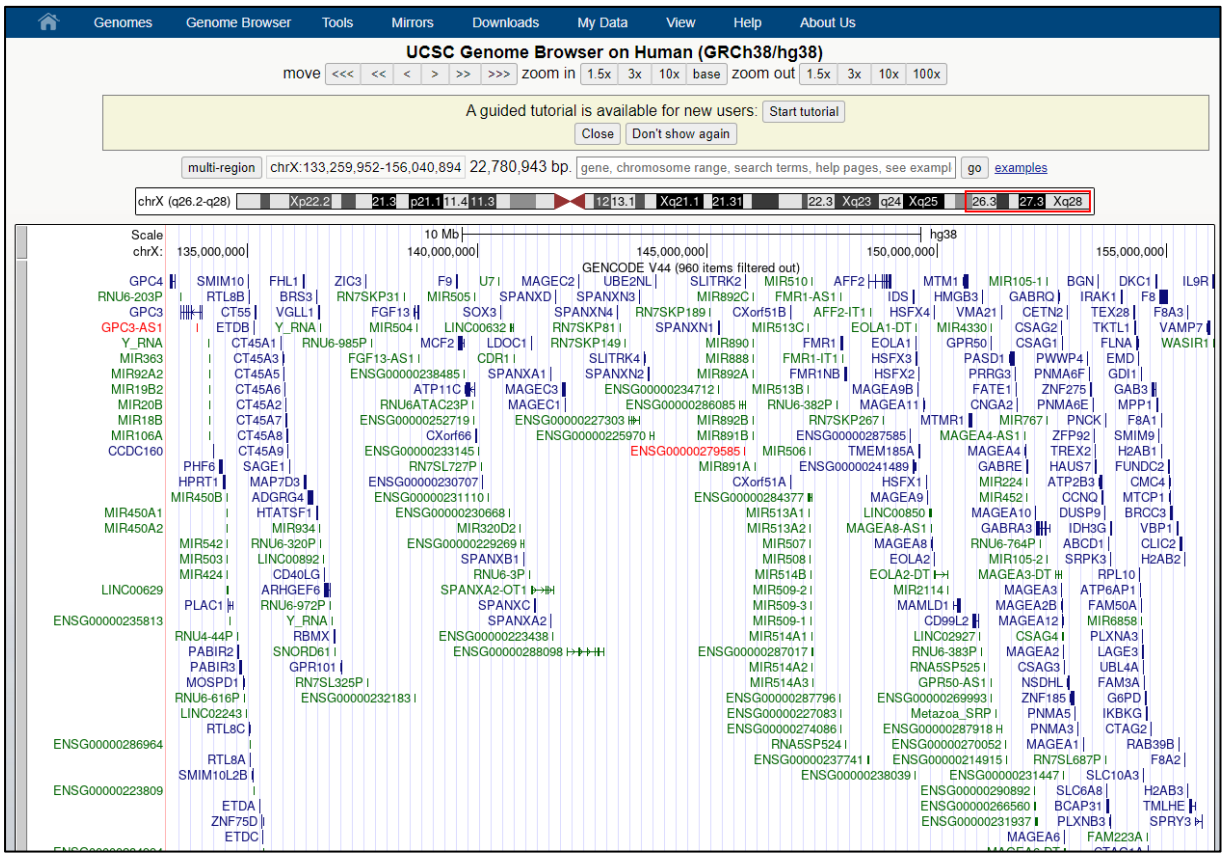


Figure 15: Moving through the assembly

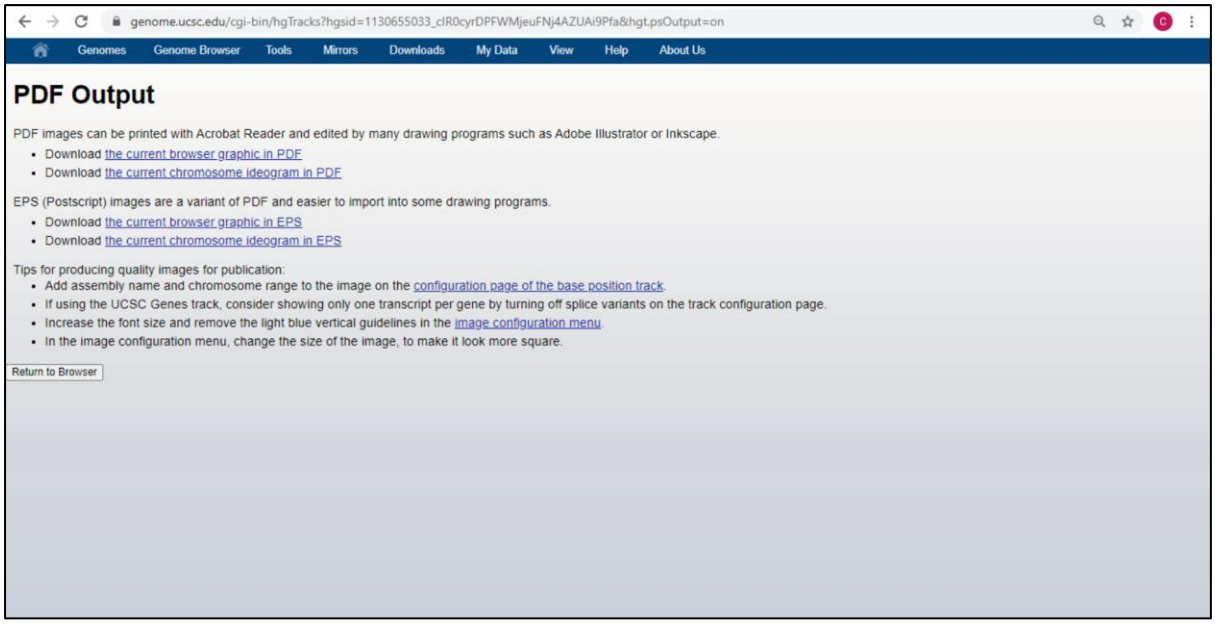


Figure 16: PDF output for 'View' option

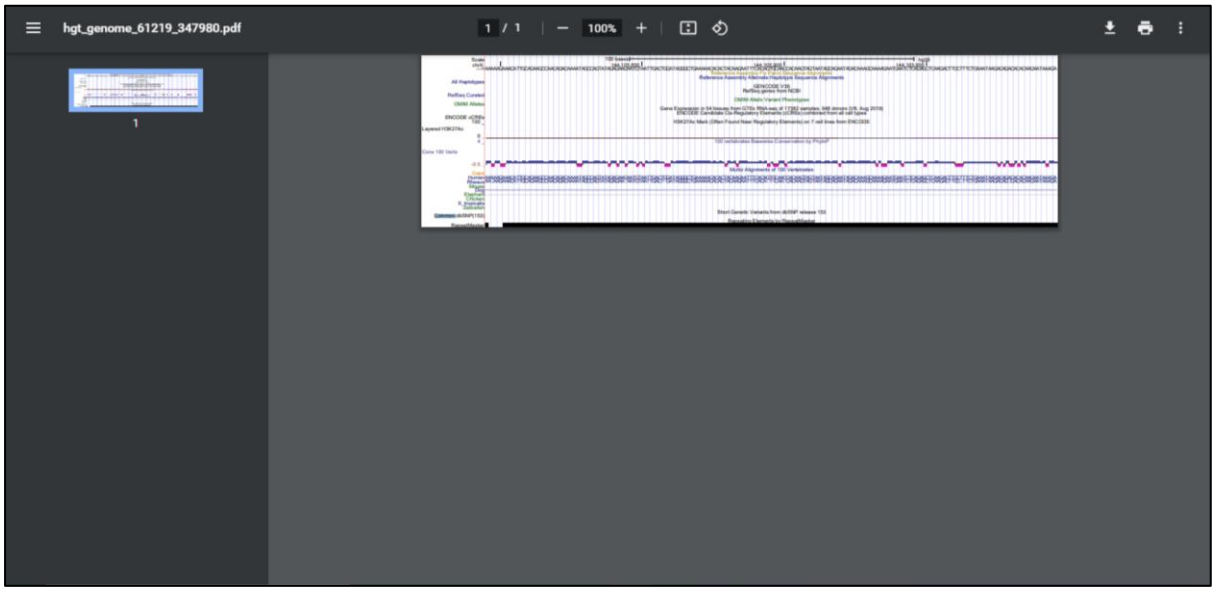


Figure 16a: Browser graphic in PDF format

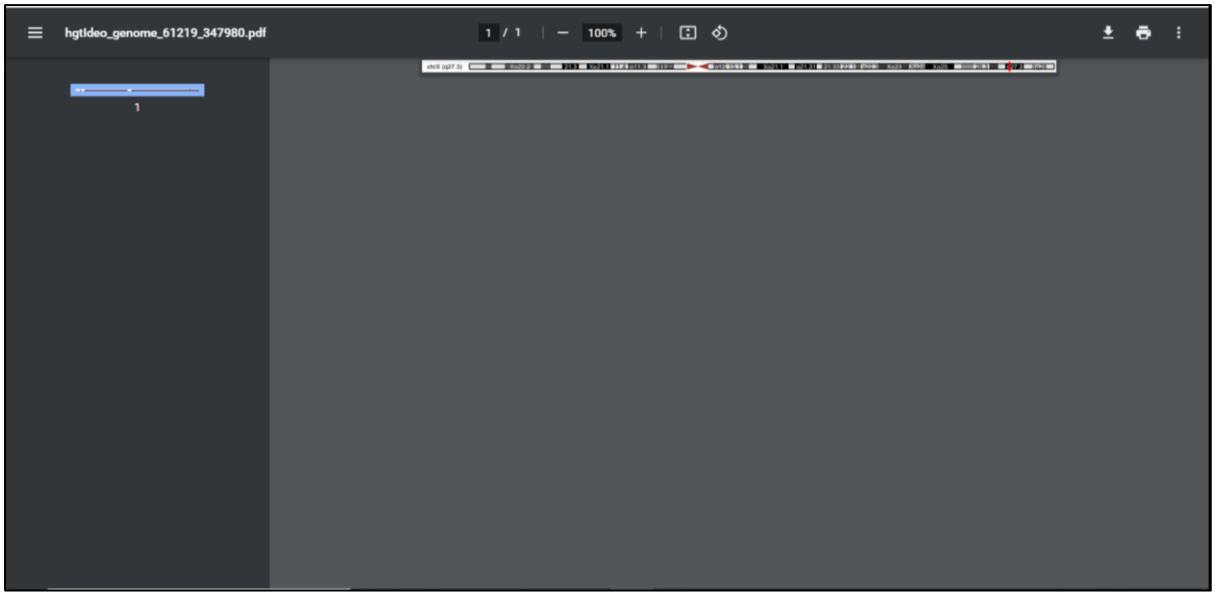


Figure 16b: Chromosome Ideogram in PDF format

B. ENSEMBL:

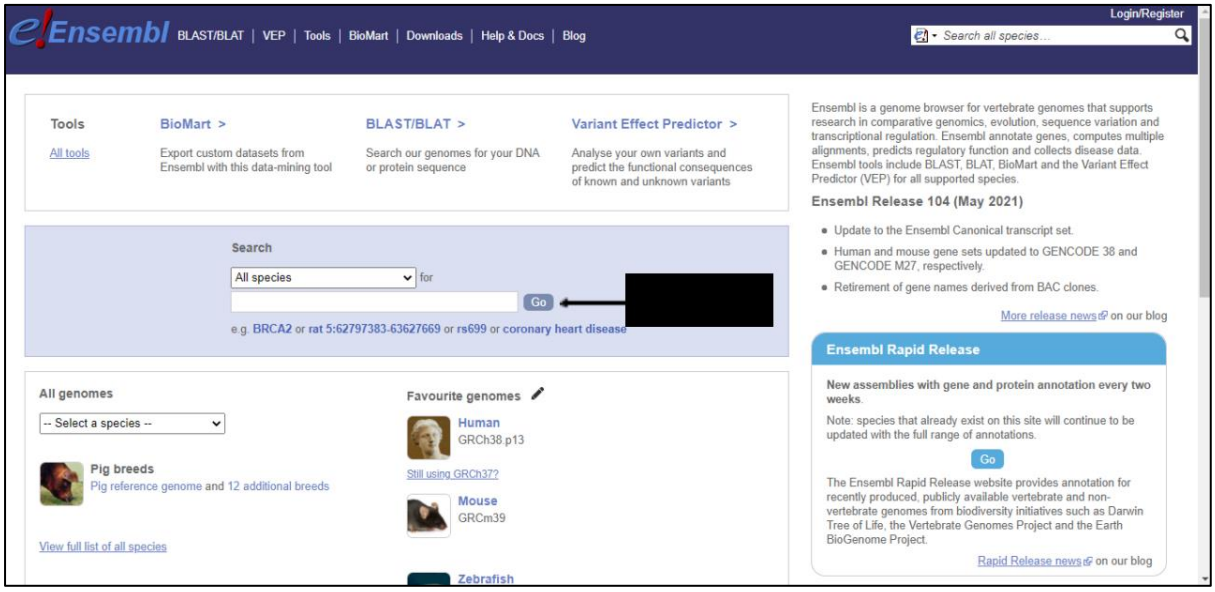


Figure 1: Homepage of Ensemble database web server

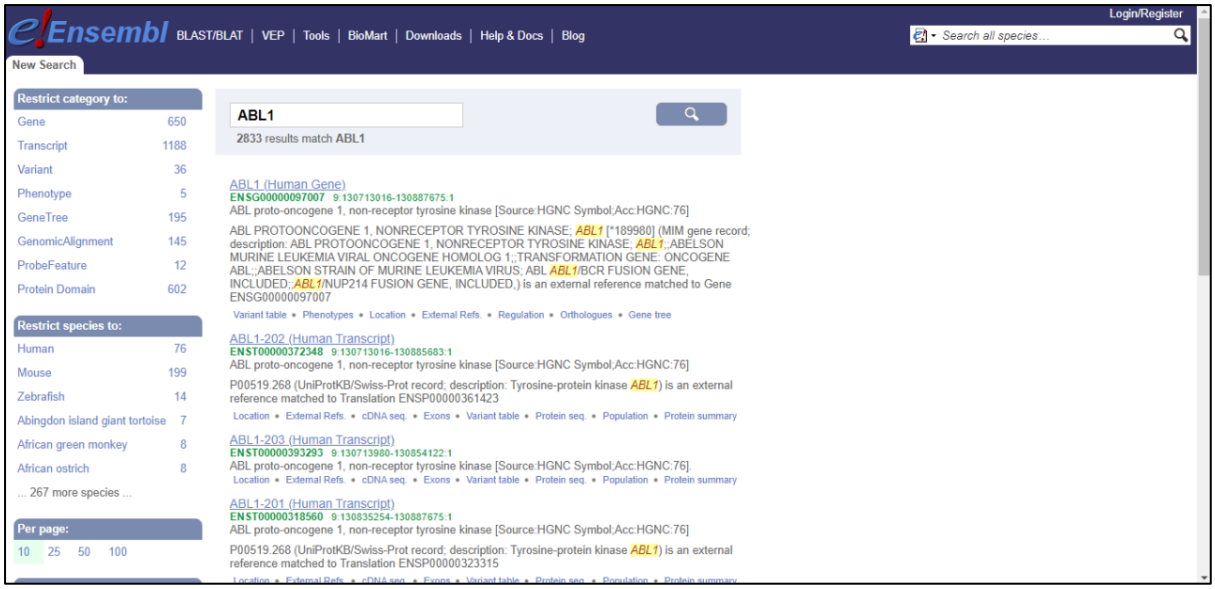


Figure 2: Hit page of Ensemble database when query (ABL1) was fired

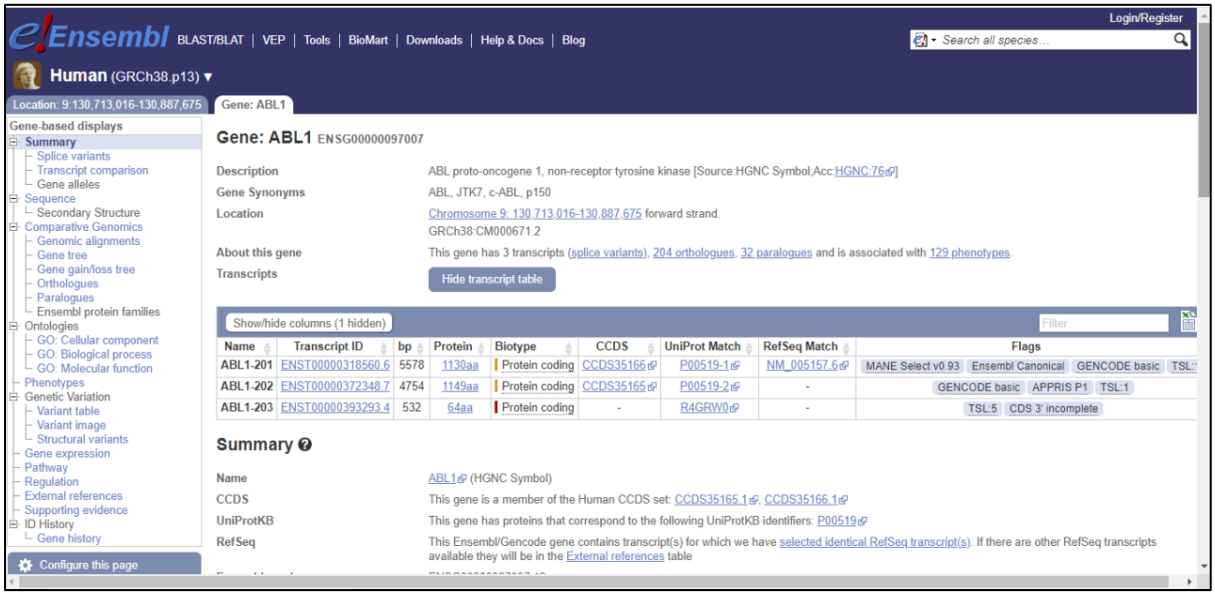


Figure 3: Result page in gene section showing summary

Description: Summary shows the description about the ABL1 gene and its location in the chromosome. The gene shows three transcripts named ABL1-201, ABL1-202, ABL1-203. Apart from this, the ensemble version, gene type, annotation method is also depicted. The annotation method used is automatic annotation from Ensembl and Havana Manual curation

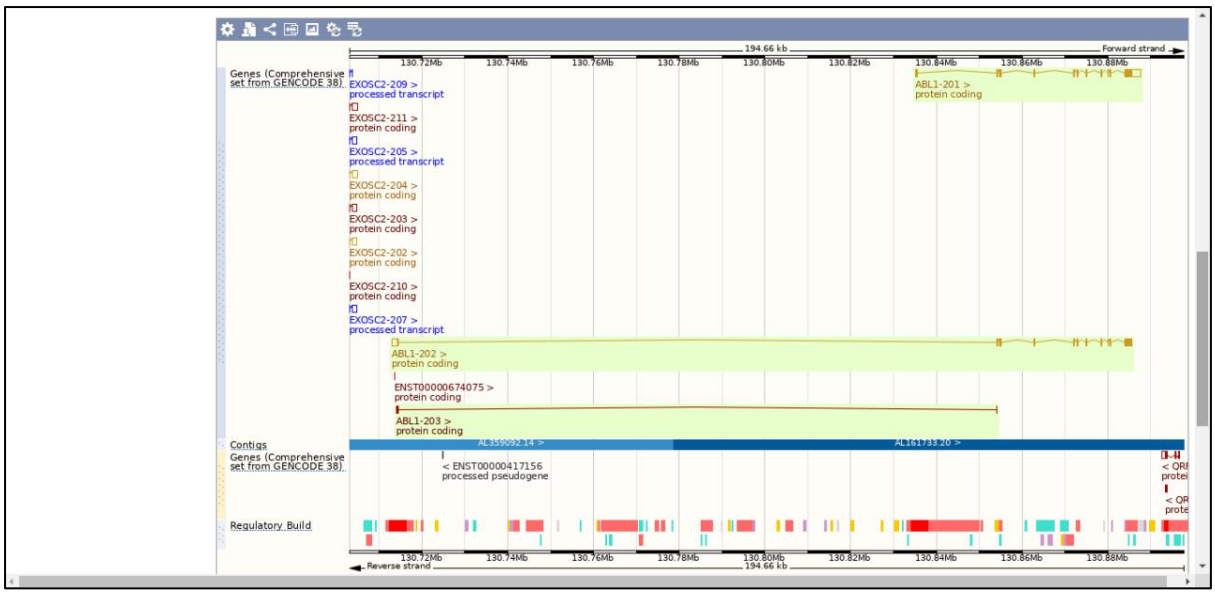


Figure 3a: Result page in gene section showing summary.

Description: Further in summary section, tracks were observed. Golden tracks indicate Ensembl annotation which is merged to Havana server. The blue tracks depict the non-coding transcripts and red tracks depicts Ensembl annotated pipeline. The lines indicate introns and boxes indicate exons. In boxes, filled boxes shows the coding sequences and unfilled boxes shows non-coding sequences

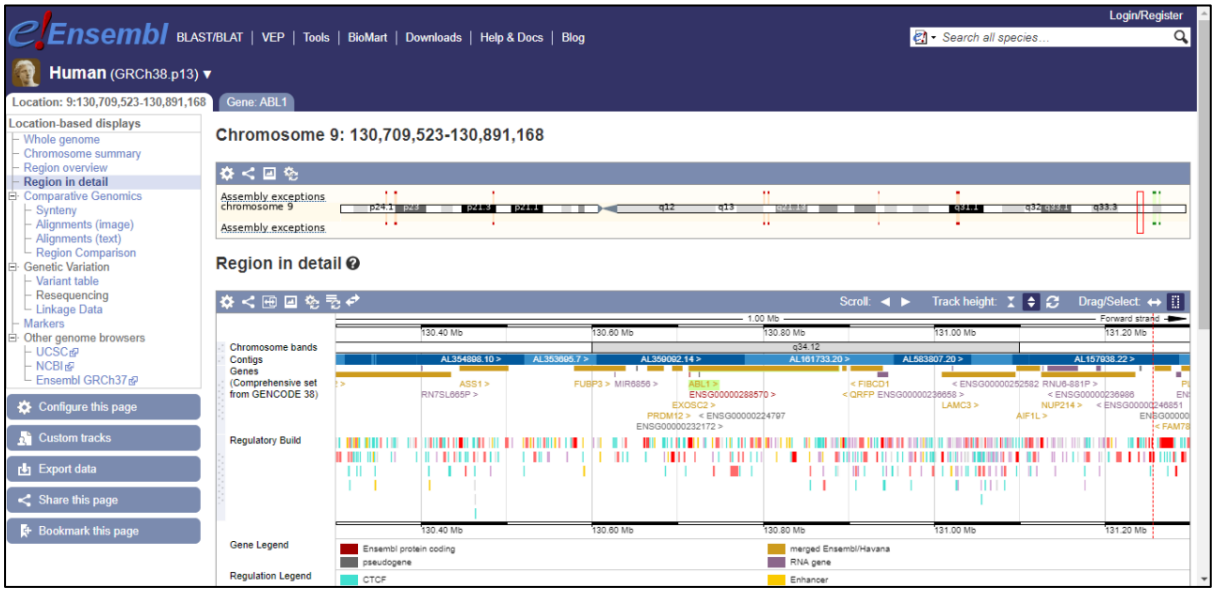


Figure 4: Location section showing the position of ABL1 gene.

Description: The chromosome image at the top depicts the ideogram. Region in details segment depicts the tracks, showing the band range in 43,167,274-43,171,963 in chromosome no. 9

Figure 4a: Location section showing the position of ABL1 gene Description: Further in section, in-depth location of the ABL1 gene was observed

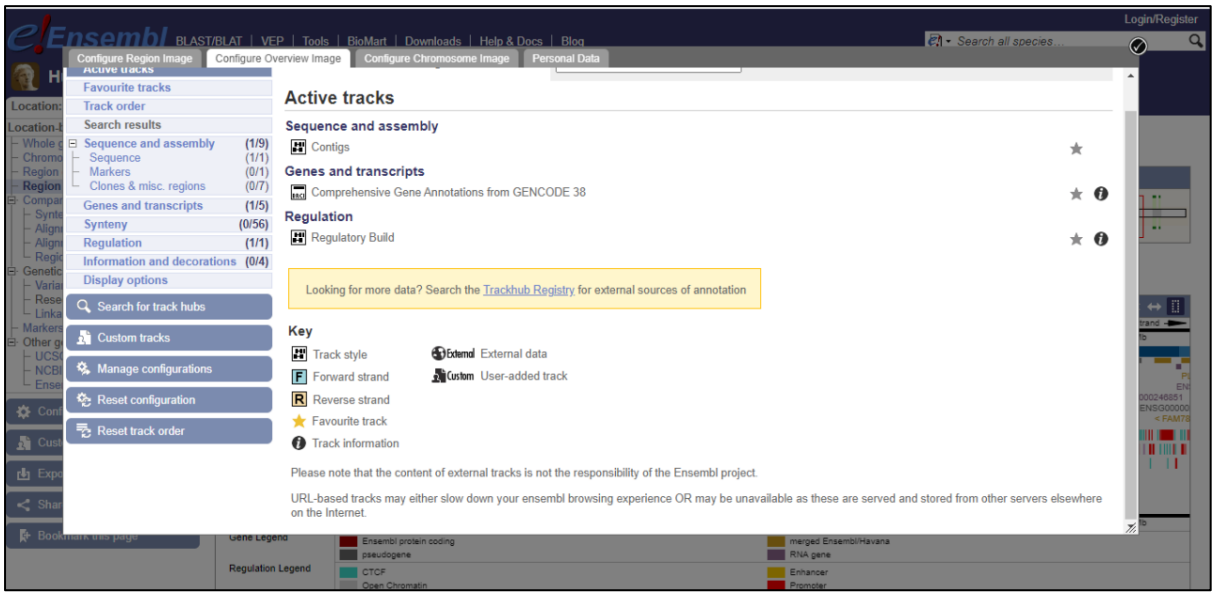


Figure 4b: Track configuration page. Description: In tracks configuration page, tracks can be modified according to the application

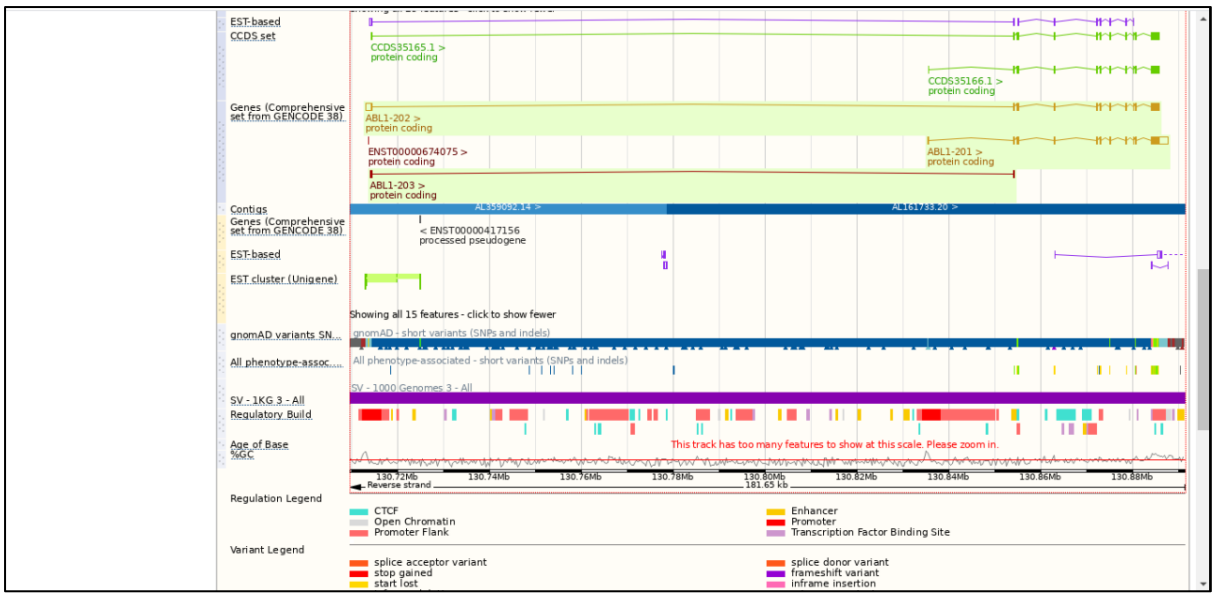


Figure 4c: Configured tracks Description: After the configuration was done, changes were observed in the track. Apart from the basic information, CCDS (Consensus coding DNA sequence) dataset and EST based sequences were also marked

RESULTS:

A. UCSC:

In the UCSC genome browser, information for GRCh38/hg38 is searched then it is navigated to hg38.

Searching options in UCSC genome browser:

1. search assembly by genes e.g., FMR1-AS1
2. search by amino acid position e.g., HERC2P8
3. search by Refseq accession e.g., NR_024501
4. search by OMIM identifier e.g., #300823
5. search by genomic co-ordinates e.g., chr1:1234570-1234870
6. search by cytological band e.g., 17q12 (gene information coding/Non coding regions)
7. search by SNP id e.g., rs1695 (results multiple matches)

The results show zoom options (zoom in 1.5x, etc. zoom out 1.5x), jump options. tracking of details shows description. configuration of tracks can be change by the option configure tracks after applying selections we can control and group them. We can drag and select option to go on specific track and get their information. Output can be viewed in PDF format. PDF format shows browser graphic information chromosome ideogram.

B. ENSEMBL:

To gather information for the ABL1 gene, the Ensembl genome browser has been used. Following are the results as per the observations:

1. The gene has shown the transcript, ABL1-201, ABL1-202 & ABL1-203. The tracks had been observed. The golden tracks indicate Ensembl annotation, which is merged to the Havana server. The blue tracks depict the non-coding transcripts, and the red tracks represent Ensembl annotated pipeline. The lines indicate introns, and boxes indicate exons. In the boxes, filled boxes show the coding sequences, and unfilled boxes show non-coding sequences.
2. The gene shows band range in 43,167,274-43,171,963 in chromosome no. 9.
3. Subsequently, the conformation was done, changes had been delineated in the track. And the CCDS (Consensus coding DNA sequence) dataset, as well as the EST-based sequences, have been marked.

CONCLUSION:

The UCSC genomic browser used to gather information about GRCh38/hg38.Secondary links from individual entries within annotation tracks lead to sequence details and supplementary off-site databases. To control information overload, tracks need not be displayed in full. Tracks can be hidden, collapsed into a condensed or single-line display, or filtered according to the user's criteria. Zooming and scrolling controls help to narrow or broaden the displayed chromosomal range to focus on the exact region of interest. Clicking on an individual item within a track opens a details page containing a summary of properties and links to off-site repositories such as PubMed, GenBank, Entrez, and OMIM. The page provides item-specific information on position, cytoband, strand, data source, and encoded protein, mRNA, genomic sequence and alignment, as appropriate to the nature of the track. The UCSC genome browser does not draw conclusion rather it collates all relevant information in one location leaving the exploration in one location, leaving the exploration and interpretation to the user. The UCSC genome browser supports text and sequence-based searches that provide quick, precise access to any region of specific interest. The Genomics browser has been explored to gather information for the ABL1 gene, and it was depicted using the Ensembl genome browser. Information about genes, transcripts, and further annotation can be retrieved at the genome, gene, and protein level. This includes information on protein domains, genetic variation, homology, syntenic regions, and regulatory elements—the Ensembl genome browser imports genome sequences from consortia which keeps things consistent with various bioinformatics projects.

REFERENCES:

1. Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K., Raney, B. J., Powell, C. C., Nassar, L. R., Maulding, N., Lee, C. M., Lee, B. T., Hinrichs, A., Fyfe, A., Fernandes, J., Diekhans, M., Clawson, H., Casper, J., Benet-Pagès, A., Barber, G. P., . . . Kent, W. (2020). The UCSC Genome Browser database: 2021 update. Nucleic Acids Research, 49(D1), D1046–D1057. https://doi.org/10.1093/nar/gkaa1070
2. Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., . . . Kent, W. J. (2010). The UCSC Genome Browser database: update 2011. Nucleic Acids Research, 39(Database), D876–D882. https://doi.org/10.1093/nar/gkq963
3. Fernandes, J. D., Zamudio-Hurtado, A., Clawson, H., Kent, W. J., Haussler, D., Salama, S. R., & Haeussler, M. (2020). The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. Mobile DNA, 11, 13. https://doi.org/10.1186/s13100-020-00208-w
4. Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, José Carlos Marugán, Thomas Maurel, Aoife C McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P Sakthivel, Ahamed I Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, Ameya Chaubal, Nishadi De Silva, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth R IIsley, Nick Langridge, Jane E Loveland, Fergal J Martin, Jonathan M Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Fiona Cunningham, Andrew D Yates, Daniel R Zerbino, Paul Flicek, Ensembl 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D884–D891, https://doi.org/10.1093/nar/gkaa942
5. What is Ensembl? | Ensembl. (2021) https://www.ebi.ac.uk/training/online/courses/ensembl-browsing-genomes/what-is-ense mbl/
6. Academic.oup.com .29 October 2007. Genome browsing with Ensembl: a practical overview. Giulietta Spudich, Xosé M. Fernández-Suárez, Ewan Birney. [online]. Available at: https://academic.oup.com/bfg/article/6/3/202/237388 7

DATE: 08/11/23

WEBLEM 6(C)
MICROBIAL GENOME DATABASE FOR COMPARATIVE ANALYSIS (MBGD) DATABASE
(URL: https://mbgd.nibb.ac.jp/)

AIM:
To explore the Microbial Genome Database for Comparative Analysis (MGBD) for query *Escherichia coli.*

INTRODUCTION:
MBGD: Microbial Genome Database for Comparative Analysis. MBGD is a workbench system for comparative analysis of completely sequenced microbial genomes. The central function of MBGD is to create an orthologous gene classification table using precomputed all-against-all similarity relationships among genes in multiple genomes. In MBGD, an automated classification algorithm has been implemented so that users can create their own classification table by specifying a set of organisms and parameters. This feature is handy when the user focuses on some taxonomically related organisms. The created classification table is stored in the database and can be explored by combining individual genomes' data and similarity relationships among genomes. Using these data, users can carry out comparative analyses from various points of view, such as phylogenetic pattern analysis, gene order comparison, and detailed gene structure comparison. MBGD is accessible at http://mbgd.genome.ad.jp/.
The growth of the number of completed microbial genome sequences has accelerated recently and nearly a hundred genomes in various levels of relatedness have already been available today. Especially interesting are the recently available multiple genomes of some particular taxonomic groups such as proteobacteria gamma subdivision and Bacillus/Clostridium group in gram-positive bacteria. The role of comparative genomics becomes much more important to utilize these large numbers of sequences not only for elucidating commonality in all of life but also for understanding the evolutionary diversity within various groups and the evolutionary processes or mechanisms producing such diversity.
Ortholog identification is a crucial step for comparative genome analysis and several systems providing ortholog grouping have been developed. Clusters of Orthologous Groups (COG) is a representative of such a system, where comprehensive ortholog classification is manually maintained; each COG entry is well annotated and is assigned a stable accession number. In spite of its usefulness for genome annotation as well as for comparative genome analysis, however, ortholog grouping is not so simple task and a single classification table is not sufficient for every purpose of comparative analysis. Indeed, ortholog grouping can be considered as a mapping from a hierarchical structure representing gene phylogeny into a simple classification table, and different partitioning of the same set of genes may result when different sets of organisms are considered. In general, when one intends to compare genomes of some closely related organisms, orthologous groups are expected to contain more one-to-one relationships than those created from all organisms currently sequenced. MBGD provides functional annotations for genes within microbial genomes, helping researchers understand the biological roles of specific genes. One of the notable features of MBGD is its organization of genes into orthologous clusters. This helps in identifying genes with similar functions across different species. Databases like MBGD are often updated regularly to include new genomic data and improve the accuracy of comparative analyses.

***Escherichia coli*:**

Escherichia coli (*E. coli*) are facultative anaerobic gram-negative bacteria that are part of the normal gastrointestinal system. These organisms mainly are found within the large intestine and frequently are implicated as causes of bacterial infections. These infections can stem from disruption of the gut mucosal membrane leading to local tissue invasion and potential distant tissue seeding through bacteremia. Urinary tract infections are thought to occur via bacterial migration proximally up the ureter, causing colonization and potential infection of the bladder and more proximal structures. Common infections with E coli as a pathogen include cholecystitis, bacteremia, cholangitis, urinary tract infection (UTI), traveler's diarrhea, pneumonia, and neonatal meningitis.

METHODOLOGY:

1. Go to the MBGD database website.
2. In the search box, enter the keywords or terms you want to search for. These could be gene names, protein names, functional annotations, or other relevant terms. E.g., *Escherichia coli*. Click on the "Search" button to initiate the search process.
3. The result page appears. Click on individual entries to explore more details.
4. If the initial search doesn't yield the desired results, consider refining your keywords or modifying the search parameters to obtain more relevant information.

OBSERVATIONS:

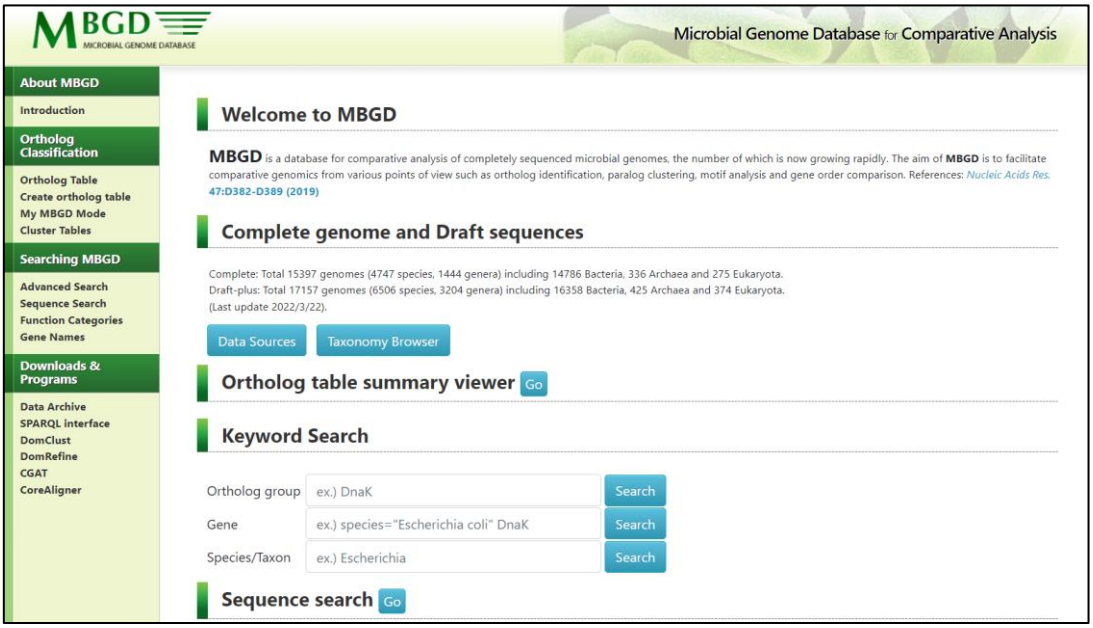


Figure 1: MBGD homepage

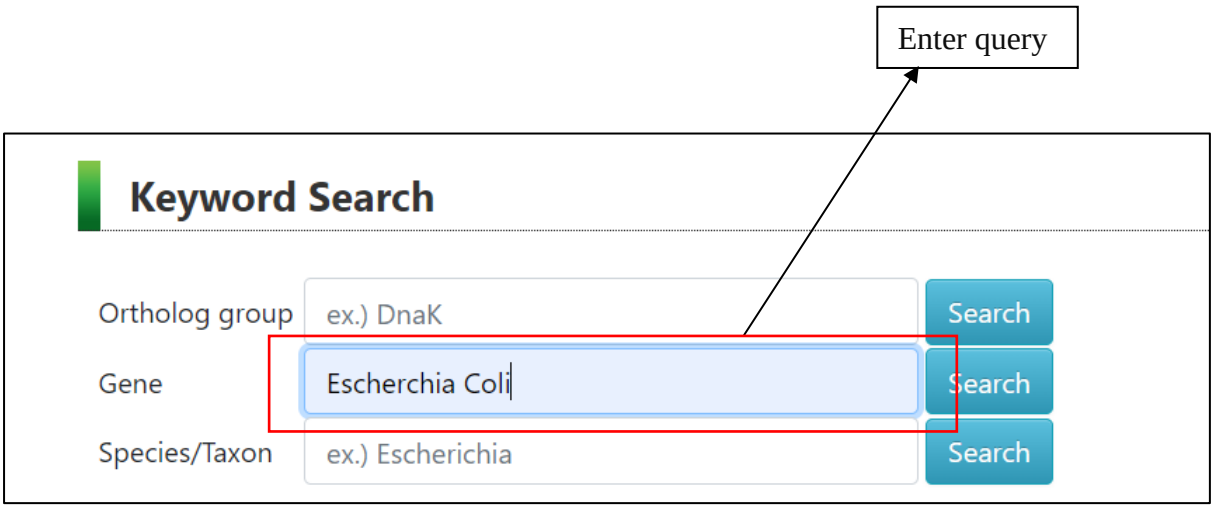
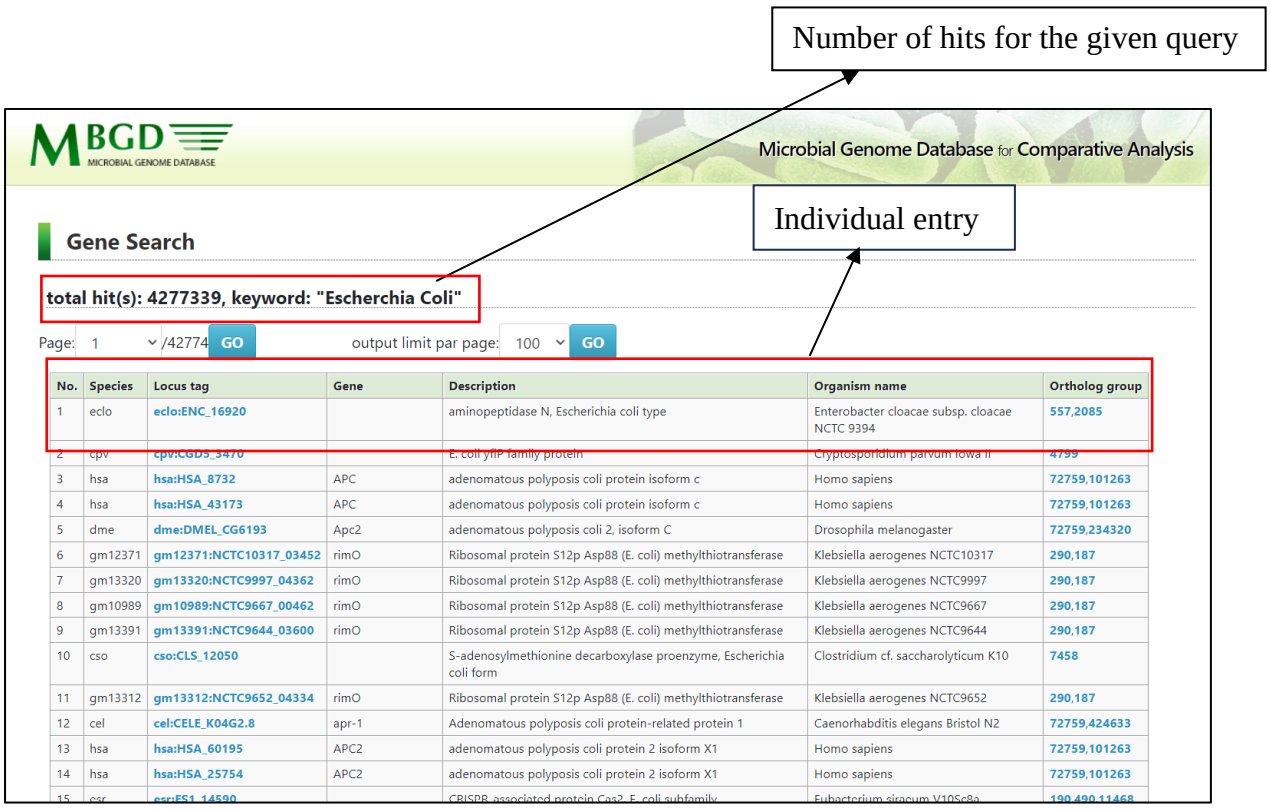


Figure 2: Enter query in keyword search



No.	Species	Locus tag	Gene	Description	Organism name	Ortholog group
1	eclo	eclo:ENC_16920		aminopeptidase N, Escherichia coli type	Enterobacter cloacae subsp. cloacae NCTC 9394	557,2085
2	cpv	cpv:CGD5_3470		E. coli yfiP family protein	Cryptosporidium parvum Iowa II	4799
3	hsa	hsa:HSA_8732	APC	adenomatous polyposis coli protein isoform c	Homo sapiens	72759,101263
4	hsa	hsa:HSA_43173	APC	adenomatous polyposis coli protein isoform c	Homo sapiens	72759,101263
5	dme	dme:DMEL_CG6193	Apc2	adenomatous polyposis coli 2, isoform C	Drosophila melanogaster	72759,234320
6	gm12371	gm12371:NCTC10317_03452	rimO	Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase	Klebsiella aerogenes NCTC10317	290,187
7	gm13320	gm13320:NCTC9997_04362	rimO	Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase	Klebsiella aerogenes NCTC9997	290,187
8	gm10989	gm10989:NCTC9667_00462	rimO	Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase	Klebsiella aerogenes NCTC9667	290,187
9	gm13391	gm13391:NCTC9644_03600	rimO	Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase	Klebsiella aerogenes NCTC9644	290,187
10	cso	cso:CLS_12050		S-adenosylmethionine decarboxylase proenzyme, Escherichia coli form	Clostridium cf. saccharolyticum K10	7458
11	gm13312	gm13312:NCTC9652_04334	rimO	Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase	Klebsiella aerogenes NCTC9652	290,187
12	cel	cel:CELE_K04G2.8	apr-1	Adenomatous polyposis coli protein-related protein 1	Caenorhabditis elegans Bristol N2	72759,424633
13	hsa	hsa:HSA_60195	APC2	adenomatous polyposis coli protein 2 isoform X1	Homo sapiens	72759,101263
14	hsa	hsa:HSA_25754	APC2	adenomatous polyposis coli protein 2 isoform X1	Homo sapiens	72759,101263
15	csr	csr:ES1_14590		CRISPR-associated protein Cas2, E. coli subfamily	Eubacterium siraeum V10Sc8a	190,490,11468

Figure 3: Results displayed after hitting search.

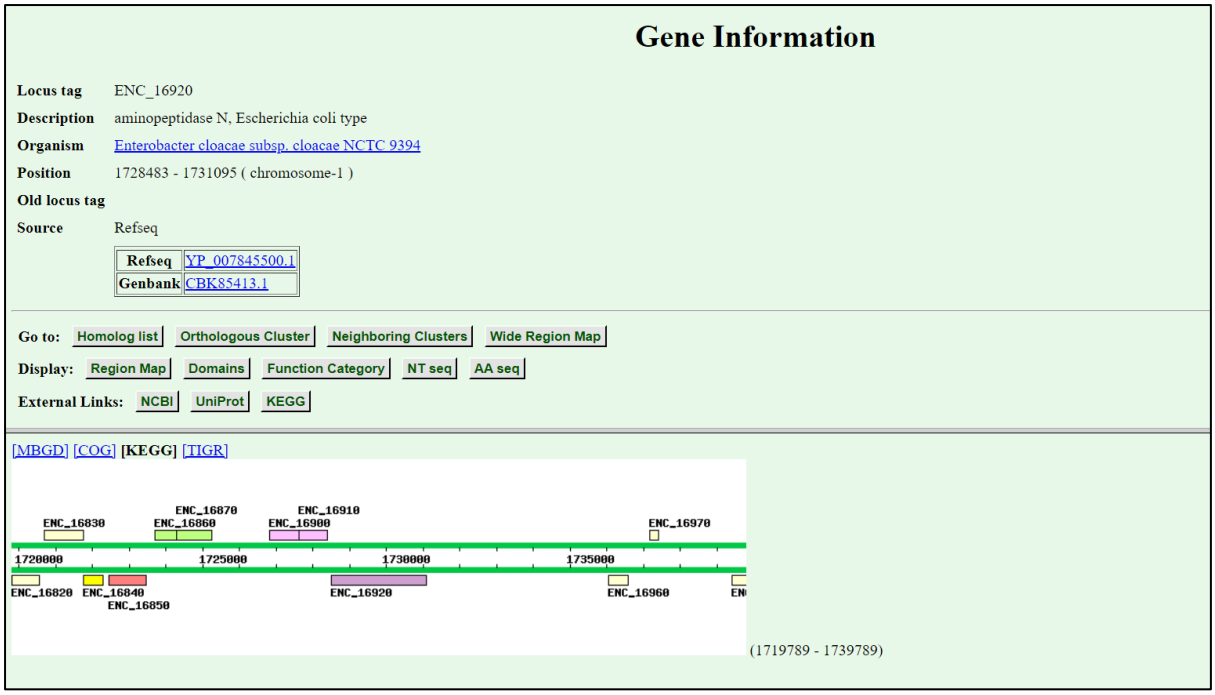


Figure 4: Gene information for selected individual entry using locus tag

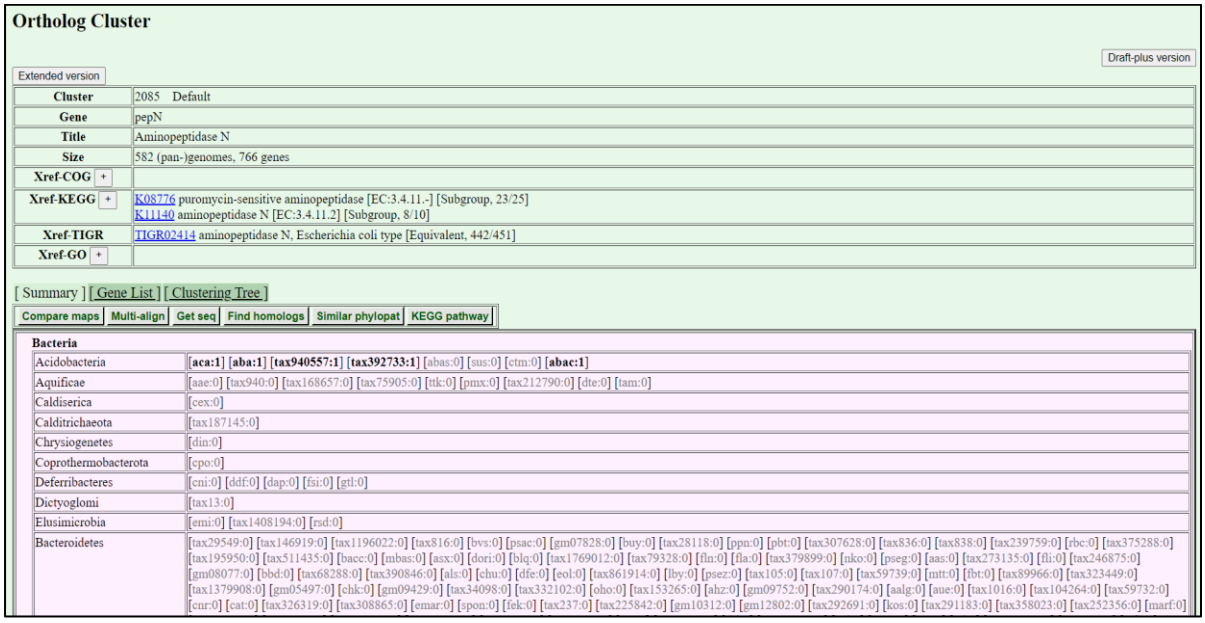


Figure 5: Ortholog cluster for selected individual entry using an ortholog group 2085

RESULTS:

MBGD database was explored for the query *Escherichia coli*. 4277339 hits were found for the given query. Gene information and Ortholog clusters for an individual entry were also explored.

CONCLUSION:

MBGD is a workbench system for comparative analysis of completely sequenced microbial genomes. The central function of MBGD is to create an orthologous gene classification table using precomputed all-against-all similarity relationships among genes in multiple genomes.

REFERENCES:

1. Collier, R. (n.d.). *Escherichia coli (E coli) Infections: Background, Pathophysiology, Epidemiology* (M. Bronze, Ed.). Medscape. https://emedicine.medscape.com/article/217485-overview
2. Uchiyama, I. (2003, January 1). MBGD: microbial genome database for comparative analysis. *Nucleic Acids Research*, *31*(1), 58–62. https://doi.org/10.1093/nar/gkg109

DATE: 08/11/2023

WEBLEM 6(D)
INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES (ICTV) DATABASE
(URL: https://ictv.global/)

AIM:
To explore the International Committee on Taxonomy of Viruses Database (ICTVdb) using the query 'Measles Virus'.

INTRODUCTION:
The International Committee on Taxonomy of Viruses (ICTV) is a committee which authorizes and organizes the taxonomic classification of viruses. The ICTV was established in 1966 as the International Committee on Nomenclature of Viruses, and was renamed the International Committee on Taxonomy of Viruses in 1977. They have developed a universal taxonomic scheme for viruses and aim to describe all the viruses of living organisms. Members of the committee are considered to be world experts on viruses. The committee formed from and is governed by the Virology Division of the International Union of Microbiological Societies. Detailed work such as delimiting the boundaries of species within a family is typically done by study groups, which consist of experts in the families. The committee also operates an authoritative database (ICTVdb)containing taxonomic information for 1,950 virus species, as of 2005. It is open to the public and is searchable by several different means.

Proposals for new names, name changes, and the establishment and taxonomic placement of taxa are handled by the Executive Committee of the ICTV in the form of proposals. All relevant ICTV subcommittees and study groups are consulted prior to a decision being made. The name of a taxon has no status until it has been approved by ICTV, and names will only be accepted if they are linked to approve hierarchical taxa. If no suitable name is proposed for a taxon, the taxon may be approved and the name be left undecided until the adoption of an acceptable international name, when one is proposed to and accepted by ICTV.

Measles Virus:
Measles is a childhood infection caused by a virus. Once quite common, measles can now almost always be prevented with a vaccine. Also called rubeola, measles spreads easily and can be serious and even fatal for small children. While death rates have been falling worldwide as more children receive the measles vaccine, the disease still kills more than 200,000 people a year, mostly children. As a result of high vaccination rates in general, measles hasn't been widespread in the United States in about two decades. Most recent measles cases in the U.S. originated outside the country and occurred in people who were unvaccinated or who didn't know whether or not they had been vaccinated.

METHODOLOGY:
1. Go to the ICTV database website.
2. Open the ICTV Taxonomy browser
3. Search for the query, 'Measles Virus'.
4. As the results are obtained, click on the measles virus taxon.

OBSERVATIONS:

Figure 1: Homepage of ICTV Database

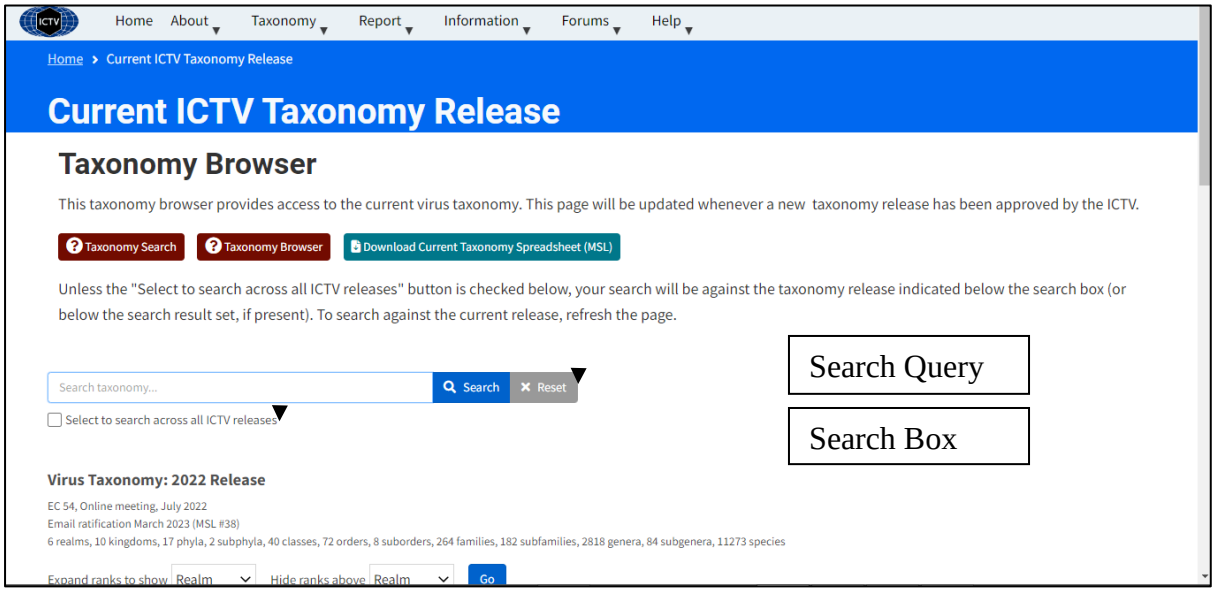


Figure 2: Taxonomy browser section

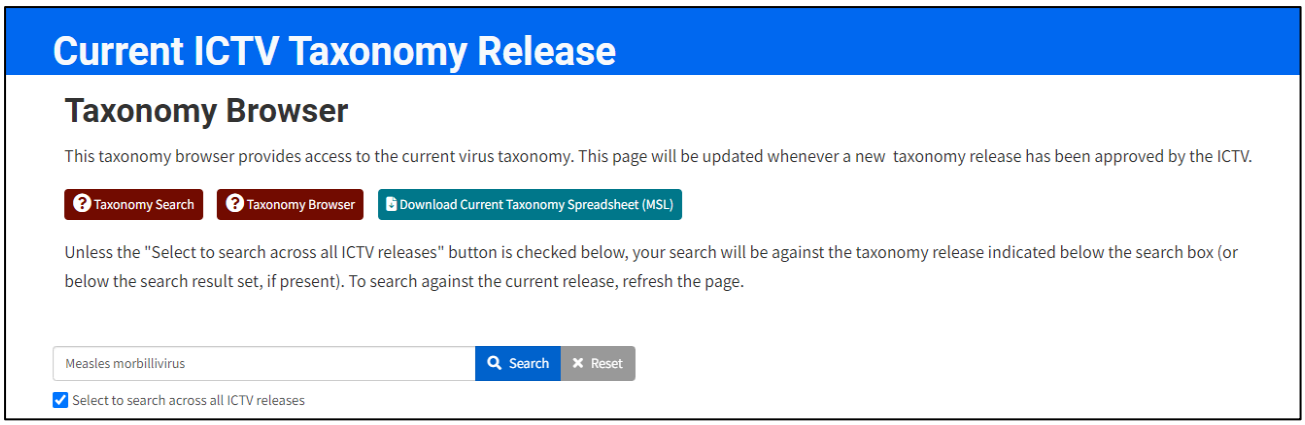


Figure 3: Taxonomy browser page after adding Query of measles virus along with select to search box ticked

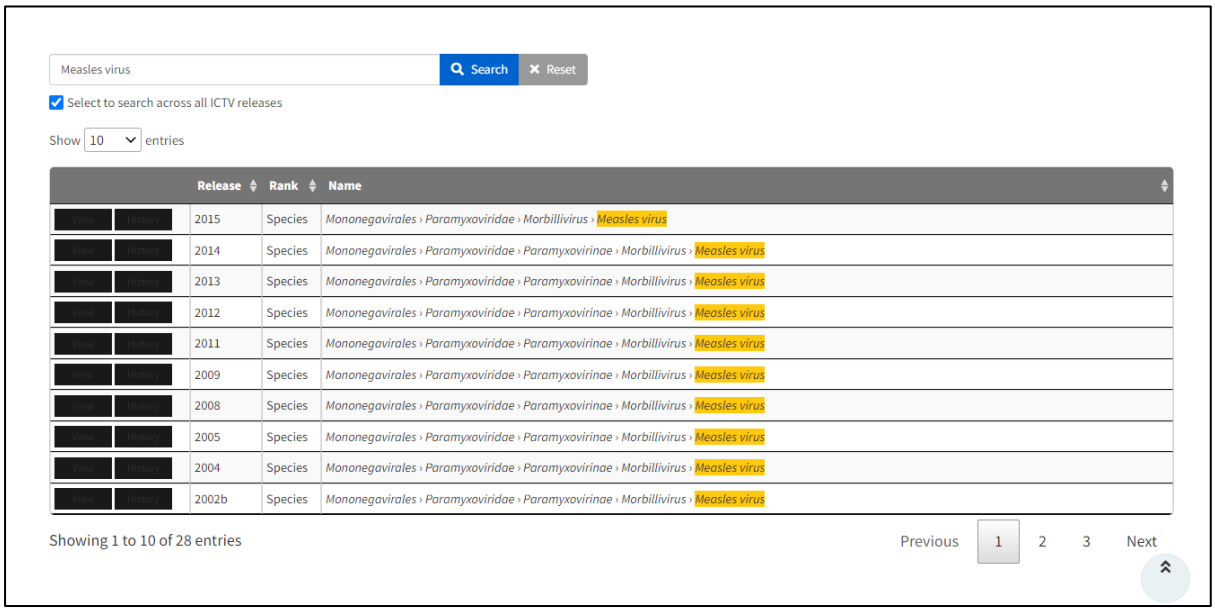


Figure 4: Result page of the ICTV Database for the query Measles virus

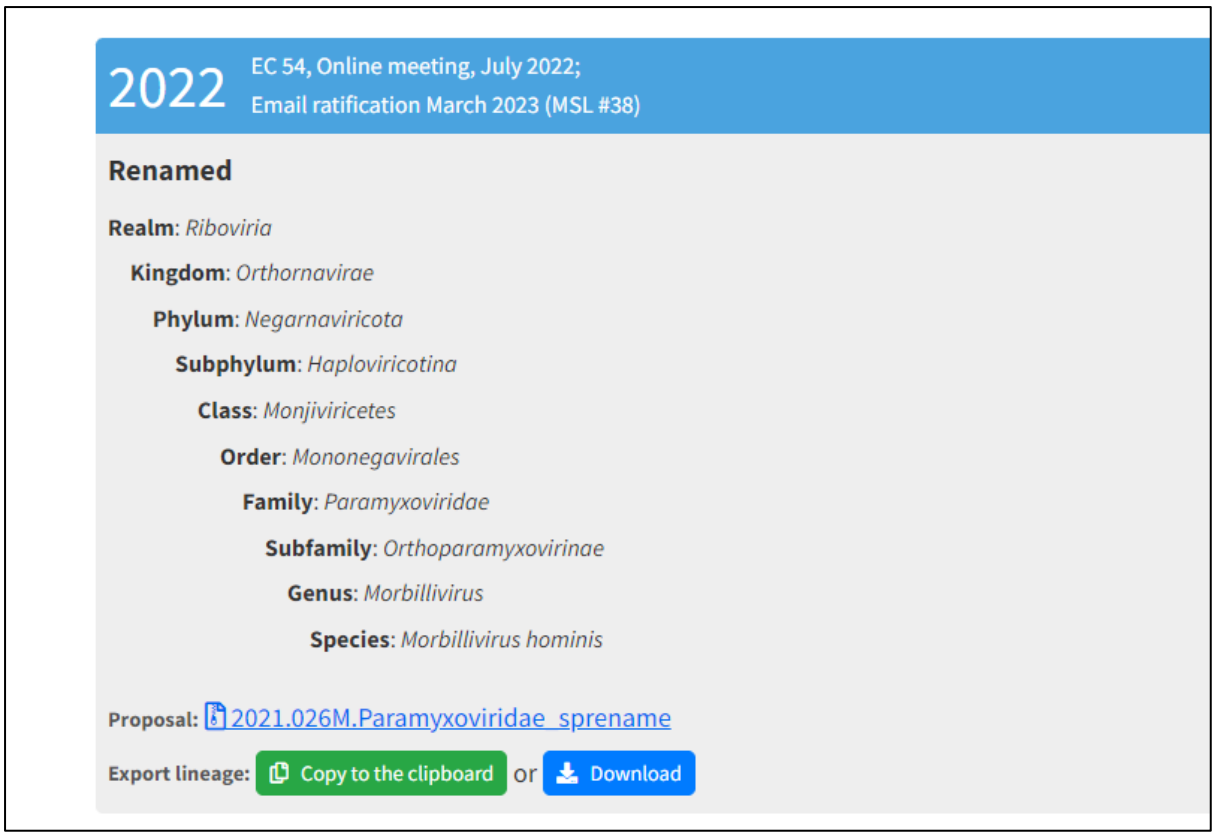


Figure 5: Updated taxon for the query Measles virus

RESULTS:

The International Committee on Taxonomy of Viruses (ICTV) was explored, the results were observed and studied for the query "Measles Virus". It resulted in 28 entries from which was selected and the updated taxon was studied.

CONCLUSION:

The International Committee on Taxonomy of Viruses (ICTV) was explored and studied using the query "Measles Virus".

REFERENCES:

1. Virology Division News. (2002, May). *Archives of Virology*, *147*(5), 1071–1076. https://doi.org/10.1007/s007050200036
2. Simmonds, P., Becher, P., Bukh, J., Gould, E. A., Meyers, G., Monath, T., Muerhoff, S., Pletnev, A., Rico-Hesse, R., Smith, D. B., & Stapleton, J. T. (2017, January 1). ICTV Virus Taxonomy Profile: Flaviviridae. *Journal of General Virology*, *98*(1), 2–3. https://doi.org/10.1099/jgv.0.000672

DATE: 06/11/23

WEBLEM 7

INTRODUCTION TO MULTIPLE SEQUENCE ALIGNMENT USING DIFFERENT TOOLS: CLUSTAL OMEGA, T – COFFEE AND MUSCLE

(URLs: 1. Clustal Omega: http://www.ebi.ac.uk/Tools/msa/clustalo/
2. T-Coffee: http://www.ebi.ac.uk/Tools/msa/tcoffee/
3. MUSCLE: http://www.ebi.ac.uk/Tools/msa/muscle/)

INTRODUCTION:

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. MSA is used to detect key functional residues, predict secondary or tertiary structures, and infer the evolutionary history of a protein family.

MSA is also used to identify conserved regions in sequences, which can provide insights into the function of the sequences studied. By aligning multiple sequences, researchers can identify regions that are conserved across different species, indicating that these regions are important for the function of the sequence. Furthermore, MSA is instrumental in identifying new members of a protein family by comparing them with similar sequences. Accurate alignments show homology and can help identify new members of a protein family. Some widely used MSA tools are Clustal Omega, T-Coffee and MUSCLE.

The MSA Tools are available through the EMBL-EBI Bioinformatics web and programmatic tools framework. These tools are accessible through the EMBL-EBI Portal (http://www.ebi.ac.uk/), providing researchers with valuable resources for multiple sequence alignment. The EMBL-EBI framework has been providing free access to a range of mainstream sequence analysis applications, including MSA tools, since 2009. Therefore, researchers can access and utilize these tools for their sequence alignments and analysis needs.

MSA TOOLS:

(A) Clustal Omega:

Clustal Omega is a multiple sequence alignment program that can align three or more sequences together in a computationally efficient and accurate manner. Clustal Omega uses seeded guide trees and HMM profile-profile techniques to generate alignments between sequences. It is widely used for carrying out multiple sequence alignment and has been benchmarked against other alignment tools. Clustal Omega can be run online at the EMBL-EBI website or downloaded and installed on a local machine.

Method:

1. Clustal Omega uses a progressive approach for multiple sequence alignment. It builds the alignment step by step, starting with the two most similar sequences and progressively adding others.
2. The algorithm employs a guide tree to determine the order of sequence alignment, which helps in efficiently aligning the sequences.
3. Clustal Omega uses a combination of pairwise and multiple sequence alignments to achieve the final alignment.

Special Information:

1. **Speed:** Clustal Omega is designed to be fast and scalable, making it suitable for large-scale sequence alignments.
2. **Accuracy:** While it may sacrifice some accuracy compared to more computationally intensive methods, it strikes a good balance between speed and alignment quality.
3. **Web Interface:** Clustal Omega provides a user-friendly web interface, making it accessible to users who may not be familiar with command-line tools.

(B) <u>T-Coffee (Tree-based Consistency Objective Function for Alignment Evaluation):</u>

T-Coffee is a multiple sequence alignment (MSA) program that can align protein, DNA and RNA sequences using structural information and homology extension. It is a consistency-based MSA program that can combine the output of various alignment methods, such as Clustal, Mafft, Probcons, and Muscle, into one alignment. T-Coffee has a new regressive mode that allows large-scale alignments, making it suitable for handling a large number of sequences efficiently. It also provides tools for evaluating alignments and outputs colored versions indicating the local reliability.

Method:

1. T-Coffee integrates information from multiple sequence alignment methods using a consistency-based approach.
2. It builds a library of pairwise alignments and constructs a library-specific scoring function to evaluate the consistency of each pairwise alignment with the multiple sequence alignment.
3. The final alignment is produced by optimizing the objective function based on the consistency scores.

Special Information:

1. **Versatility:** T-Coffee can incorporate alignments from various sources, including structure-based alignments and profile-profile alignments, to improve accuracy.

2. **Web Server:** T-Coffee is available through a web server, making it accessible for users who prefer a graphical interface.
3. **Consistency:** The method of using consistency scores helps in producing more accurate alignments, especially in regions of high variability.

(C) MUSCLE (MUltiple Sequence Comparison by Log-Expectation):

Method:

1. MUSCLE employs a progressive method similar to Clustal Omega. It starts with pairwise alignments and builds a guide tree to guide the progressive alignment of sequences.
2. It uses a log-expectation scoring scheme, which considers the likelihood of observing the observed residues in the sequences given their evolutionary relationship.
3. MUSCLE allows for refinement iterations to improve the initial alignment.

Special Information:

1. **Accuracy:** MUSCLE is known for producing highly accurate alignments and is often used when high precision is crucial.
2. **Speed:** While not as fast as Clustal Omega, MUSCLE is still efficient and can handle larger datasets with good performance.
3. **Command-Line and Web Interface:** MUSCLE can be used through both command-line tools and a web interface for user convenience.

Maximum input file size for multiple sequence alignment:

Each tool has its own limit.

Tool	Sequence Limit
Clustal Omega	4000 sequences and 4 MB of data
T - Coffee	500 sequences or a maximum file size of 1 MB
MUSCLE	500 sequences and 1 MB of data

Representation of different colors in Protein Alignments:

Residue	Color	Property
AVFPMILW	RED	Small [small + hydrophobic (including aromatic – Y)]
DE	BLUE	Acidic
RK	MAGENTA	Basic – H
STYHCNGQ	GREEN	Hydroxyl + Sulfhydryl + Amine + G
Others	GREY	Unusua amino / imino acids, etc.

Representation of consensus symbols in multiple sequence alignment:

An asterisk (*) indicates positions which have a single, fully conserved residue. A colon (:) indicates conservation between groups of strongly similar properties. A period (.) indicates conservation between groups of weakly similar properties.

SIGNIFICANCE OF MULTIPLE SEQUENCE ALIGNMENT:

1. **Homology Inference:** MSA is crucial for inferring homology between biological sequences. By aligning multiple sequences, researchers can identify conserved regions, which are indicative of functional and structural importance. This aids in understanding the evolutionary relationships and functional implications of the aligned sequences.
2. **New Member Identification:** MSA is instrumental in identifying new members of protein families by comparing them with similar sequences. Accurate alignments facilitate the recognition of homologous sequences and the discovery of new members within a protein family.
3. **Evolutionary Analysis:** MSA methods are essential for evolutionary analysis, as they consider evolutionary events such as mutations, insertions, deletions, and rearrangements. This allows researchers to study the evolutionary history and relationships between sequences, providing valuable insights into the genetic and functional evolution of biological entities.
4. **Benchmarking and Efficiency:** MSA programs are critical for benchmarking and assessing the efficiency of alignment methods. The accuracy and computational costs of MSA programs are essential indicators for selecting the most suitable program for specific datasets. Finding the right balance between speed and accuracy is crucial, and MSA tools provide the means to achieve this balance.
5. **Handling Large Datasets:** MSA methods are the only feasible solution for handling large datasets, especially in the era of high-throughput sequencing. They enable the alignment of numerous sequences, allowing for comprehensive comparative analyses and evolutionary studies.

REFERENCES:

1. Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A., Abeln, S., & Heringa, J. (2017). Multiple sequence alignment. In J. M. Keith (Ed.), *Bioinformatics* (Vol. 1525, pp. 167–189). Springer New York. https://doi.org/10.1007/978-1-4939-6622-6_8
2. Anderson, C. L., Strope, C. L., & Moriyama, E. N. (2011). SuiteMSA: Visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC Bioinformatics*, *12*(1), 184. https://doi.org/10.1186/1471-2105-12-184
3. Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290
4. Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F., & Notredame, C. (2011). T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, *39*(suppl), W13–W17. https://doi.org/10.1093/nar/gkr245
5. Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
6. Korak, T., Aşir, F., Işik, E., & CengiZ, N. (2021). Multiple sequence alignment quality comparison in T-Coffee, MUSCLE and M-Coffee based on different benchmarks. *Cumhuriyet Science Journal*, *42*(3), 526–535. https://doi.org/10.17776/csj.842265
7. Hubley, R., Wheeler, T. J., & Smit, A. F. A. (2022). Accuracy of multiple sequence alignment methods in the reconstruction of transposable element families. *NAR Genomics and Bioinformatics*, *4*(2), lqac040. https://doi.org/10.1093/nargab/lqac040
8. Ahola, V., Aittokallio, T., Vihinen, M., & Uusipaikka, E. (2006). A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics*, *7*(1), 484. https://doi.org/10.1186/1471-2105-7-484

DATE: 06/11/23

WEBLEM: 7 (A)
MULTIPLE SEQUENCE ALIGNMENT USING DIFFERENT TOOLS: CLUSTAL OMEGA, T-COFFEE AND MUSCLE

(URLs: 1. Clustal Omega: http://www.ebi.ac.uk/Tools/msa/clustalo/
2. T-Coffee: http://www.ebi.ac.uk/Tools/msa/tcoffee/
3. MUSCLE: http://www.ebi.ac.uk/Tools/msa/muscle/)

AIM:

To explore Multiple Sequence Alignment Tools, namely Clustal Omega, T-Coffee and MUSCLE for aligning 'cytochrome c oxidase subunit 1' protein sequences from five different species. The species used in this study and their UniProt IDs are as follows: *Homo sapiens* (UniProt ID: P00395), *Mus musculus* (UniProt ID: P00397), *Rattus norvegicus* (UniProt ID: P05503), *Bos taurus* (UniProt ID: P00396), *Ovis aries* (UniProt ID: 078749).

INTRODUCTION:

Multiple Sequence Alignment (MSA) is a method used to align three or more biological sequences (protein or nucleic acid) of similar length. It is employed to detect key functional residues, predict secondary or tertiary structures, and infer the evolutionary history of a protein family. MSA is also used to identify conserved regions in sequences, providing insights into their function, and to identify new members of a protein family by comparing them with similar sequences. Some widely used MSA tools include Clustal Omega, T-Coffee, and MUSCLE.

The EMBL-EBI Bioinformatics web and programmatic tools framework provides access to various mainstream sequence analysis applications, including MSA tools, through the EMBL-EBI Portal (www.ebi.ac.uk/). These tools have been freely accessible since 2009, offering researchers valuable resources for multiple sequence alignment.

Clustal Omega, known for its computational efficiency and accuracy, utilizes seeded guide trees and HMM profile-profile techniques to generate alignments, making it suitable for large-scale sequence alignments. T-Coffee, on the other hand, is a consistency-based MSA program that integrates information from various alignment methods using a consistency-based approach. MUSCLE, which stands for MUltiple Sequence Comparison by Log-Expectation, is recognized for its high accuracy and efficiency, particularly when precision is crucial. MUSCLE allows for refinement iterations to improve the initial alignment.

These tools are available through the EMBL-EBI Bioinformatics web and programmatic tools framework, providing valuable resources for researchers' sequence alignment and analysis needs since.

Cytochrome C Oxidase subunit 1 (COX1):

Cytochrome c oxidase subunit 1 (COX1) is a protein encoded by the MT-CO1 gene in eukaryotes. It is a core subunit of the cytochrome c oxidase complex, which is the terminal enzyme of the mitochondrial electron transport chain. COX1 contributes to cytochrome-c oxidase activity and is involved in mitochondrial electron transport, facilitating the transfer of electrons from cytochrome c to oxygen. The structure of the core subunits of cytochrome c oxidase is conserved from α-proteobacteria, the ancestors of mitochondria, to human COX. COX1 and COX3 are highly conserved across species, highlighting their evolutionary significance. COX1 is a crucial component of the mitochondrial electron transport chain, playing a central role in cellular respiration and energy production across eukaryotic organisms.

The MSA of COX1 protein sequences can aid in understanding the molecular evolution and functional implications of this protein across different species, providing valuable insights into the conservation and variation of this essential mitochondrial protein.

Here, COX1 protein of the following five species were studied: *Homo sapiens, Mus musculus, Rattus norvegicus, Bos taurus* and *Ovis aries*

Cytochrome c oxidase subunit 1 (COX1) is a crucial protein found in multiple species, including *Bos taurus* (cattle), *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Ovis aries* (sheep). In *Bos taurus*, the COX1 protein is encoded by the MT-CO1 gene and consists of 514 amino acids. It is a component of the cytochrome c oxidase, the last enzyme in the mitochondrial electron transport chain that drives oxidative phosphorylation. The structure of the core subunits of cytochrome c oxidase is conserved from α-proteobacteria, the ancestors of mitochondria, to bovine COX, highlighting its evolutionary significance. In *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, and *Ovis aries*, COX1 plays a crucial role in aerobic metabolism and energy production within cells by facilitating the transfer of electrons from cytochrome c to oxygen. The conservation of COX1 across species underscores its evolutionary significance and its essential role in cellular respiration and energy production.

METHODOLOGY:

1. Visit the UniProt website.
2. In the Search bar, type “Cytochrome c oxidase subunit 1” or “COX1” and analyze the Results to identify the entries for the five species you’re interested in (e.g., *Homo sapiens, Bos taurus, Mus musculus, Rattus norvegicus, Ovis aries*)
3. Select the entries of the five species simultaneously and click on the “Download” button and save the sequence in “FASTA (canonical)” format.
4. Go to EMBL-EBI Website.
5. Click on the “Clustal W”, “T-Coffee” and “MUSCLE” Tools available under EMBL-EBI Tools section.

6. In each tool (Clustal W, T-Coffee, MUSCLE) input the FASTA Sequence for all five species.
7. Set any relevant parameters or options according to the tool's requirements and your Analysis goals.
8. Submit the queries in each tool.
9. Note the job submission and identifiers or URLs provided by the tools for later retrieval of Results.
10. For each tool, observe the generated Multiple sequence Alignments.
11. Interpret the results by analyzing the quality of alignments, identifying conserved regions and considering any observed variations.

OBSERVATIONS:

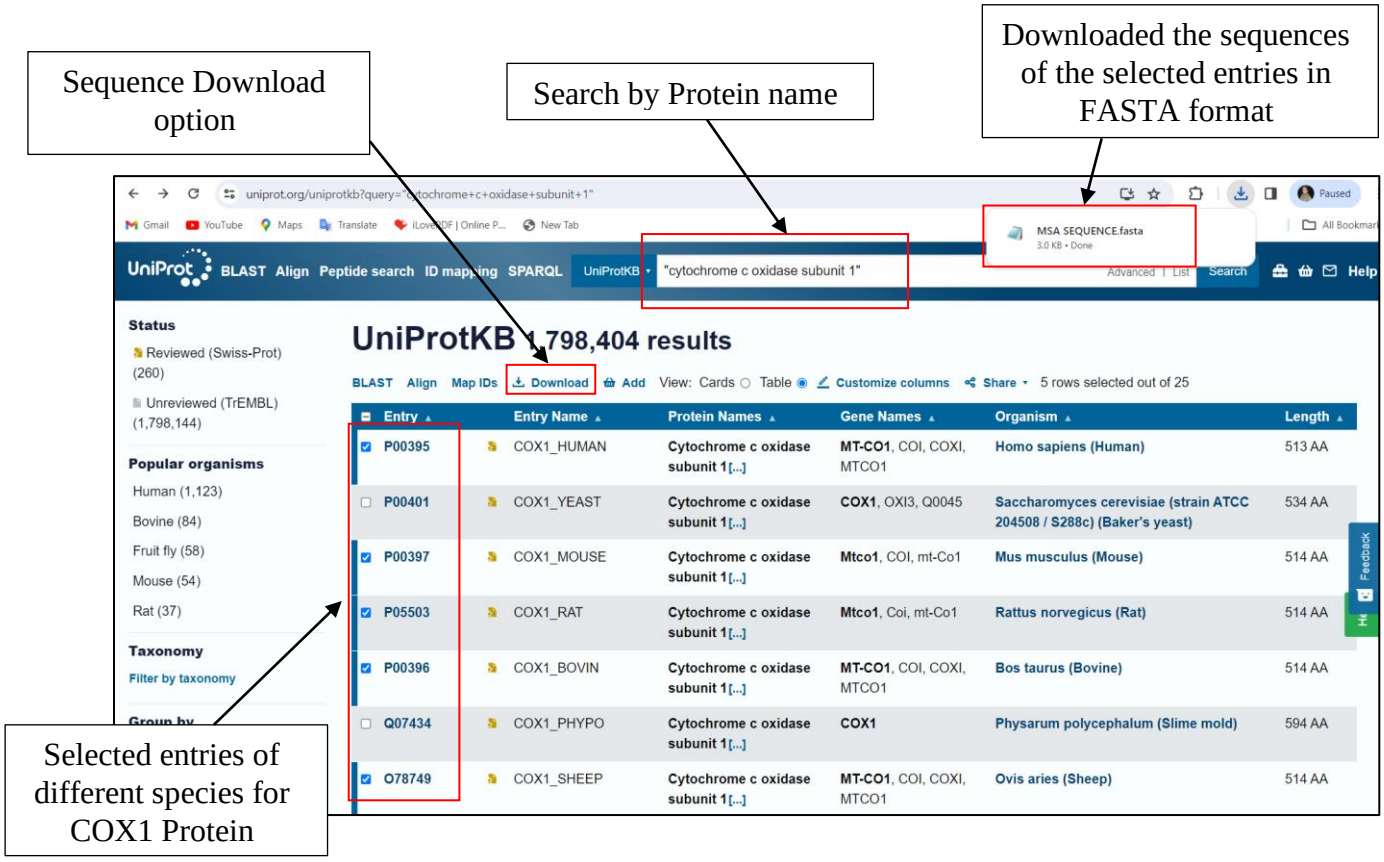
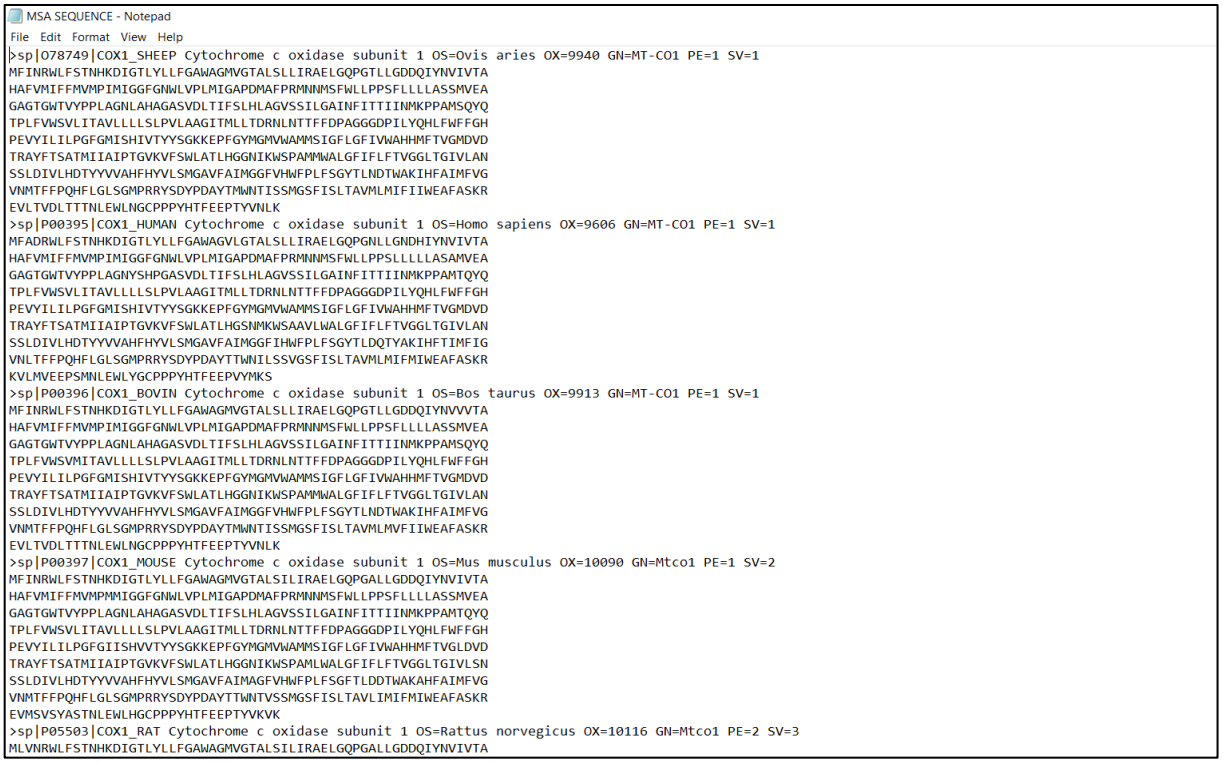


Figure 1: Selected entries of five species for protein Cytochrome c oxidase subunit 1 (COX1) from the UniProt database and downloaded the sequences in FASTA format



```
MSA SEQUENCE - Notepad
File  Edit  Format  View  Help
>sp|O78749|COX1_SHEEP Cytochrome c oxidase subunit 1 OS=Ovis aries OX=9940 GN=MT-CO1 PE=1 SV=1
MFINRWLFSTNHKDIGTLYLLFGAWAGMVGTALSLLIRAELGQPGTLLGDDQIYNVIVTA
HAFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSFLLLLASSMVEA
GAGTGWTVYPPLAGNLAHAGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMSQYQ
TPLFVWSVLITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
PEVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVD
TRAYFTSATMIIAIPTGVKVFSWLATLHGGNIKWSPAMMWALGFIFLFTVGGLTGIVLAN
SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFVHWFPLFSGYTLNDTWAKIHFAIMFVG
VNMTFFPQHFLGLSGMPRRYSDYPDAYTMWNTISSMGSFISLTAVMLMIFIIWEAFASKR
EVLTVDLTTTNLEWLNGCPPPYHTFEEPTYVNLK
>sp|P00395|COX1_HUMAN Cytochrome c oxidase subunit 1 OS=Homo sapiens OX=9606 GN=MT-CO1 PE=1 SV=1
MFADRWLFSTNHKDIGTLYLLFGAWAGVLGTALSLLIRAELGQPGNLLGNDHIYNVIVTA
HAFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSLLLLASAMVEA
GAGTGWTVYPPLAGNYSHPGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMTQYQ
TPLFVWSVLITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
PEVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVD
TRAYFTSATMIIAIPTGVKVFSWLATLHGSNMKWSAAVLWALGFIFLFTVGGLTGIVLAN
SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLDQTYAKIHFTIMFIG
VNLTFFPQHFLGLSGMPRRYSDYPDAYTTWNILSSVGSFISLTAVMLMIFMIWEAFASKR
KVLMVEEPSMNLEWLYGCPPPYHTFEEPVYMKS
>sp|P00396|COX1_BOVIN Cytochrome c oxidase subunit 1 OS=Bos taurus OX=9913 GN=MT-CO1 PE=1 SV=1
MFINRWLFSTNHKDIGTLYLLFGAWAGMVGTALSLLIRAELGQPGTLLGDDQIYNVVVTA
HAFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSFLLLLASSMVEA
GAGTGWTVYPPLAGNLAHAGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMSQYQ
TPLFVWSVMITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
PEVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVD
TRAYFTSATMIIAIPTGVKVFSWLATLHGGNIKWSPAMMWALGFIFLFTVGGLTGIVLAN
SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFVHWFPLFSGYTLNDTWAKIHFAIMFVG
VNMTFFPQHFLGLSGMPRRYSDYPDAYTMWNTISSMGSFISLTAVMLMVFIIWEAFASKR
EVLTVDLTTTNLEWLNGCPPPYHTFEEPTYVNLK
>sp|P00397|COX1_MOUSE Cytochrome c oxidase subunit 1 OS=Mus musculus OX=10090 GN=Mtco1 PE=1 SV=2
MFINRWLFSTNHKDIGTLYLLFGAWAGMVGTALSILIRAELGQPGALLGDDQIYNVIVTA
HAFVMIFFMVMPMMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSFLLLLASSMVEA
GAGTGWTVYPPLAGNLAHAGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMTQYQ
TPLFVWSVLITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
PEVYILILPGFGIISHVVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGLDVD
TRAYFTSATMIIAIPTGVKVFSWLATLHGGNIKWSPAMLWALGFIFLFTVGGLTGIVLSN
SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMAGFVHWFPLFSGFTLDDTWAKAHFAIMFVG
VNMTFFPQHFLGLSGMPRRYSDYPDAYTTWNTVSSMGSFISLTAVLIMIFMIWEAFASKR
EVMSVSYASTNLEWLHGCPPPYHTFEEPTYVKVK
>sp|P05503|COX1_RAT Cytochrome c oxidase subunit 1 OS=Rattus norvegicus OX=10116 GN=Mtco1 PE=2 SV=3
MLVNRWLFSTNHKDIGTLYLLFGAWAGMVGTALSILIRAELGQPGALLGDDQIYNVIVTA
```

Figure 2: FASTA sequences for 'cytochrome c oxidase subunit 1 (COX1)' protein of the five selected species shown in Notepad

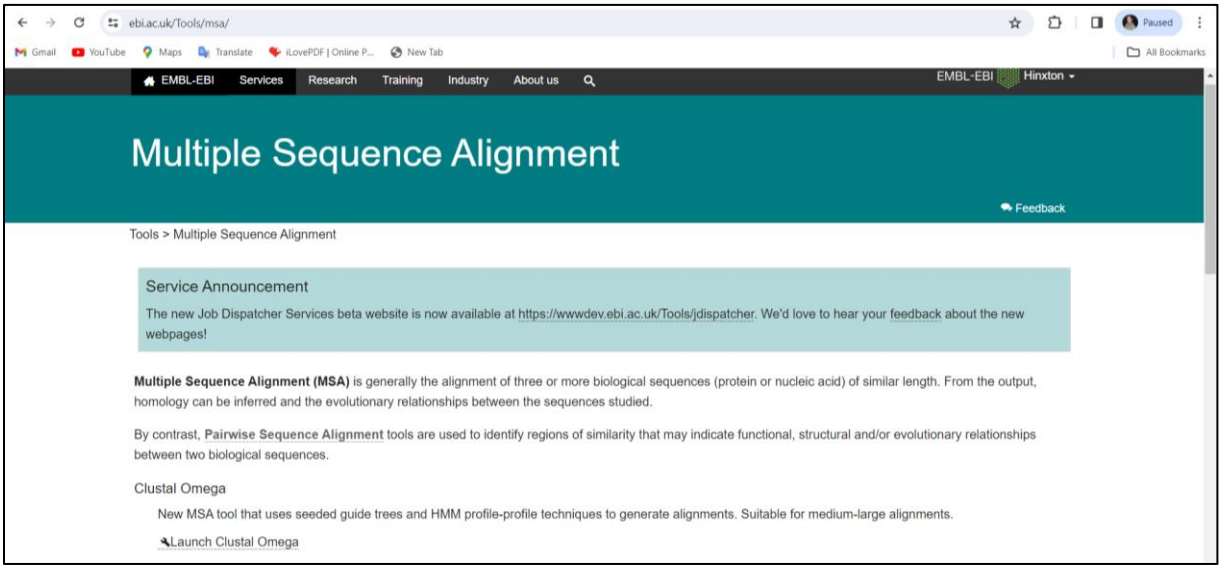


Figure 3: EMBL-EBI Portal showing list of Multiple Sequence Alignment tools

TOOL 1: CLUSTAL OMEGA

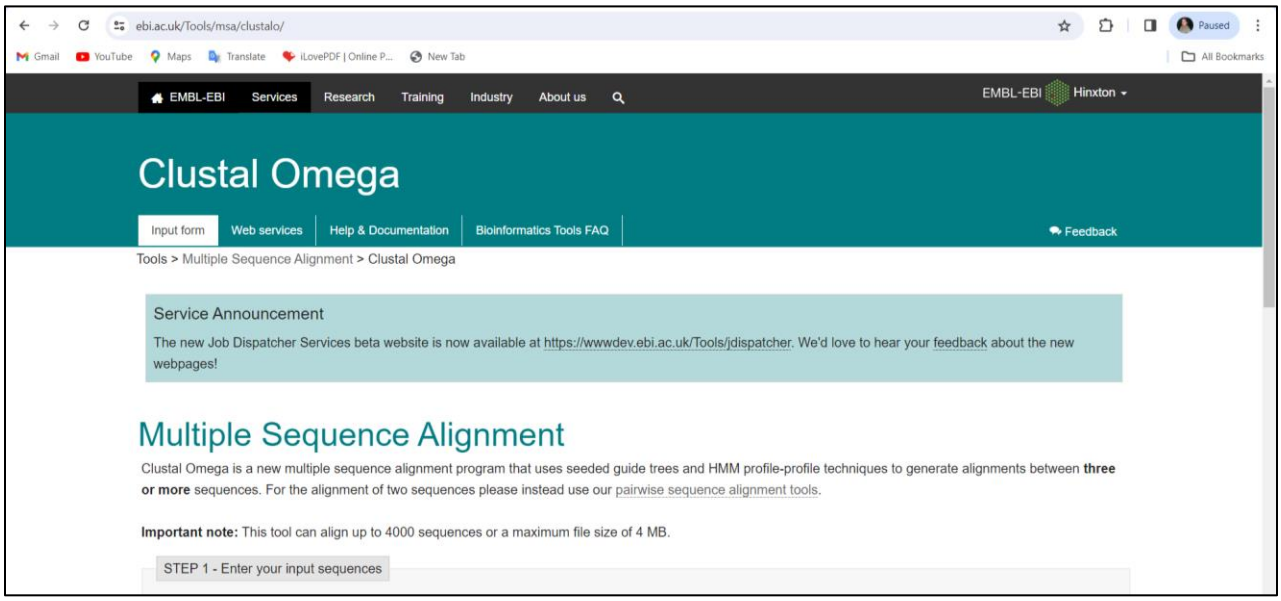


Figure 4: Homepage for Clustal Omega tool

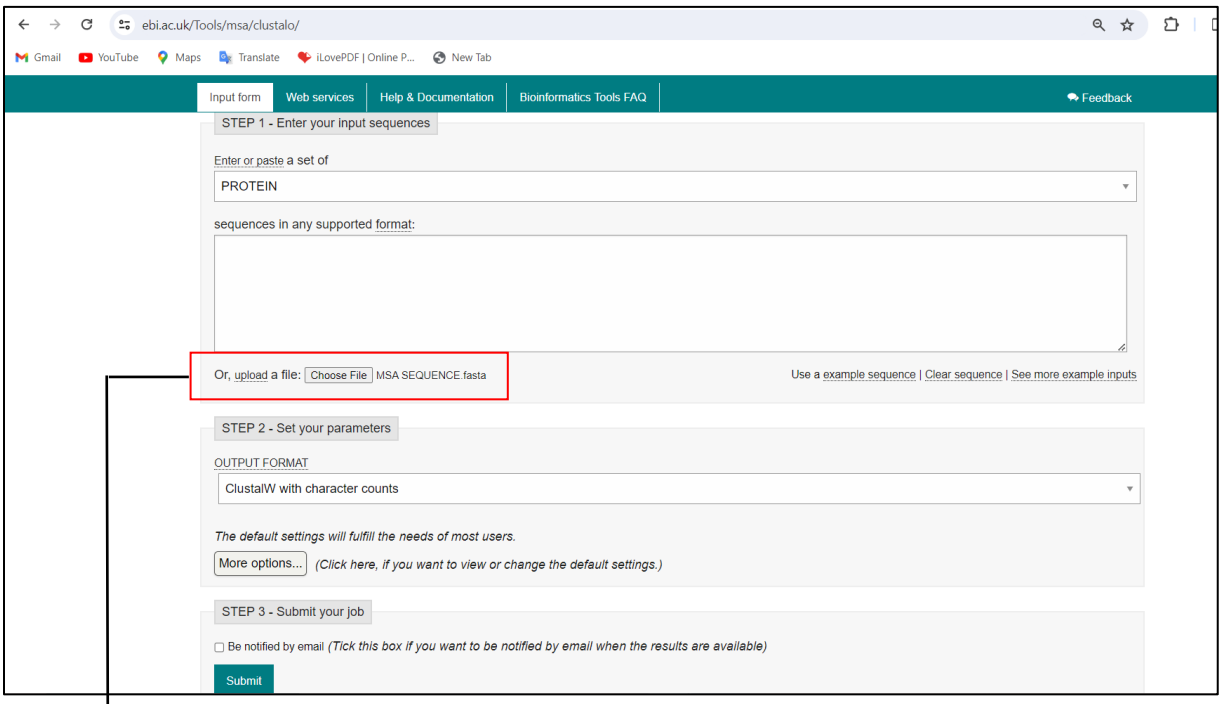
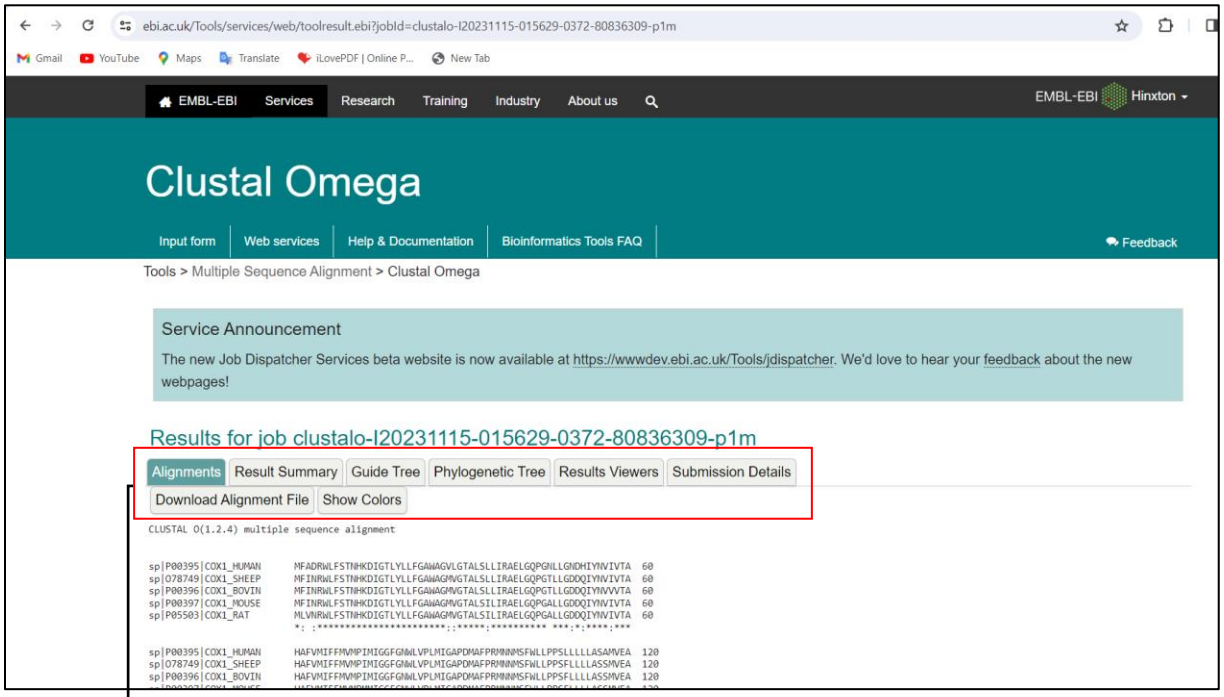


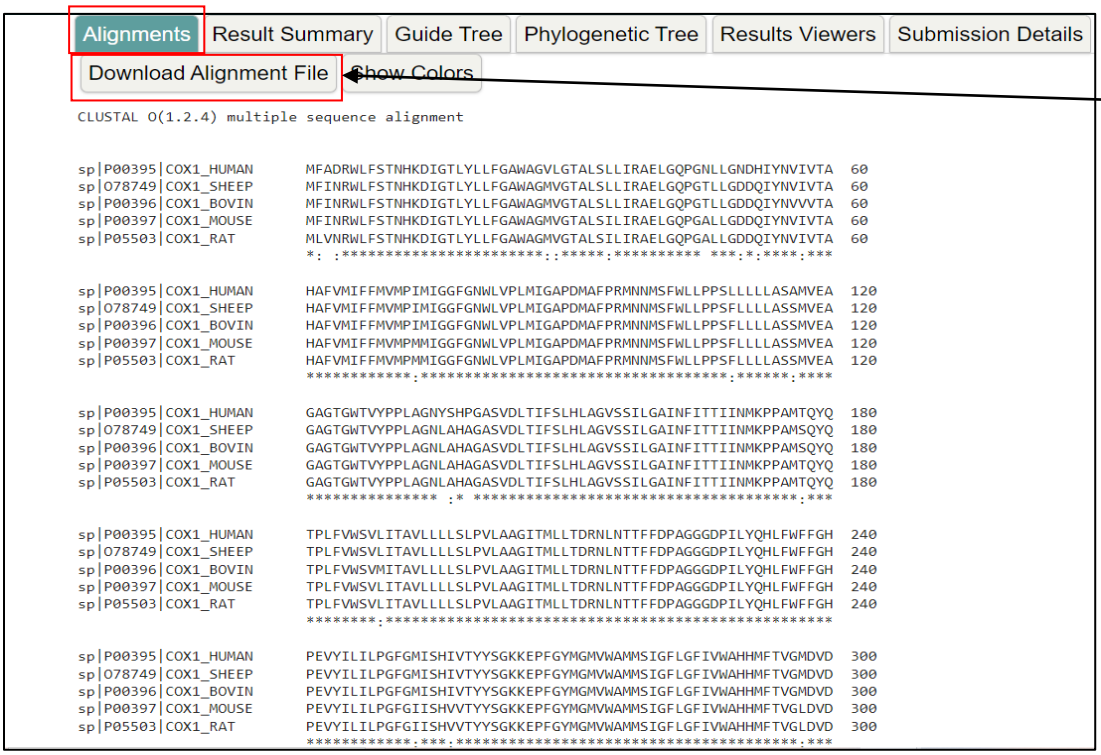
Figure 4a: Input form for submitting data to Clustal Omega tool

Uploaded Sequence in FASTA Format



Aggregated Results

Figure 4b: Result obtained for the query in Clustal Omega tool



The Alignment file can be downloaded in the sequence input format

Figure 4c: Sequence Alignments in Clustal Omega tool

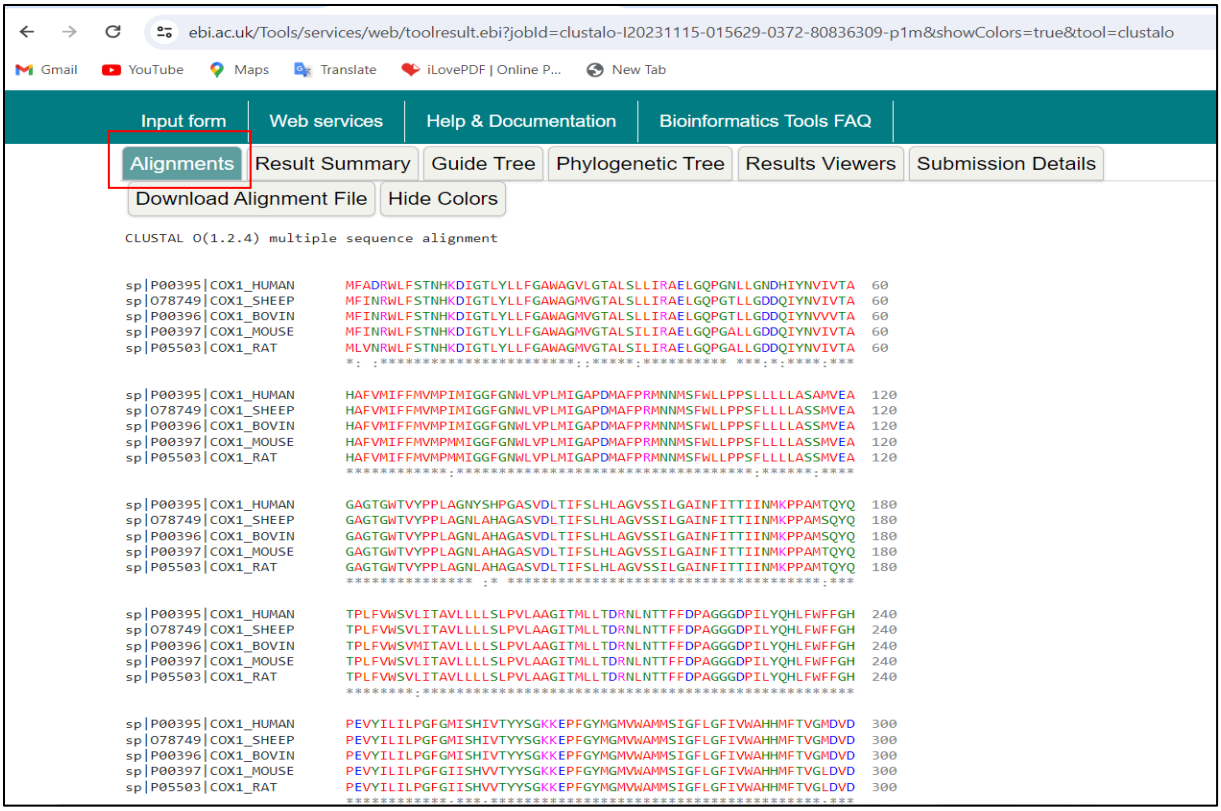


Figure 4d: Color key/Show Colors for Alignments in Clustal Omega tool

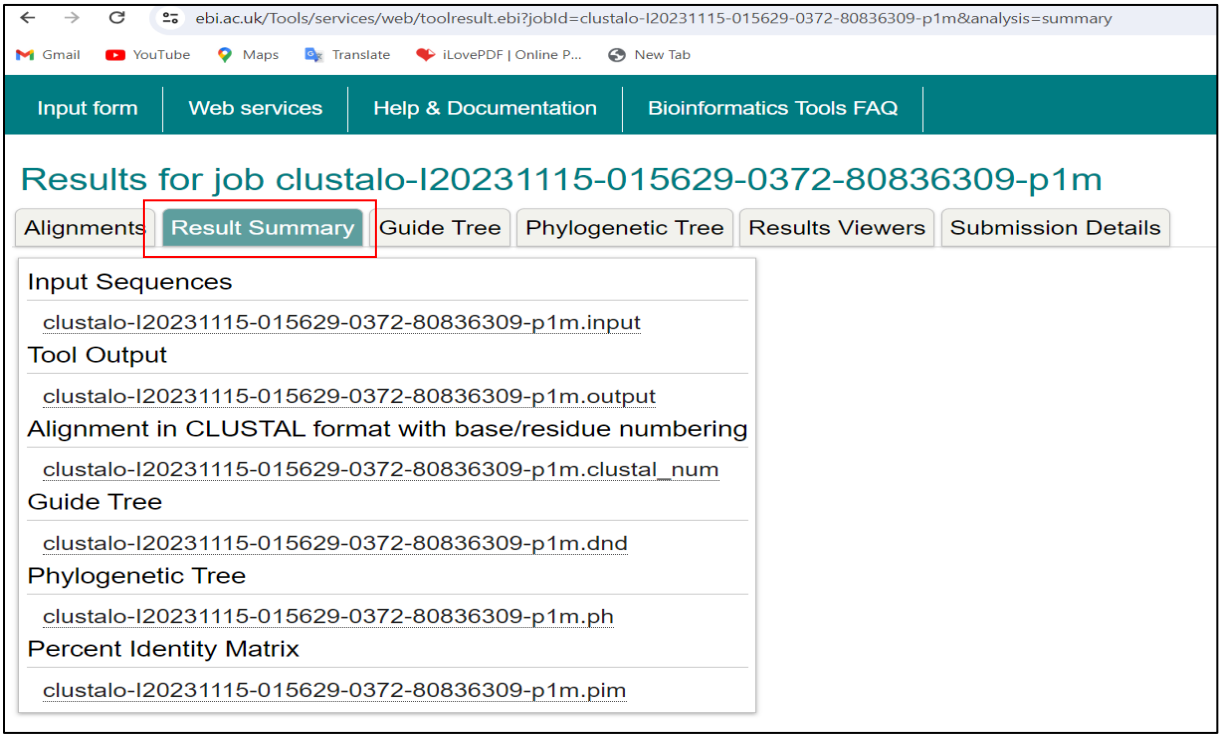


Figure 4e: Summary of the results in Clustal Omega tool

The Guide Tree data can be downloaded in Newick format

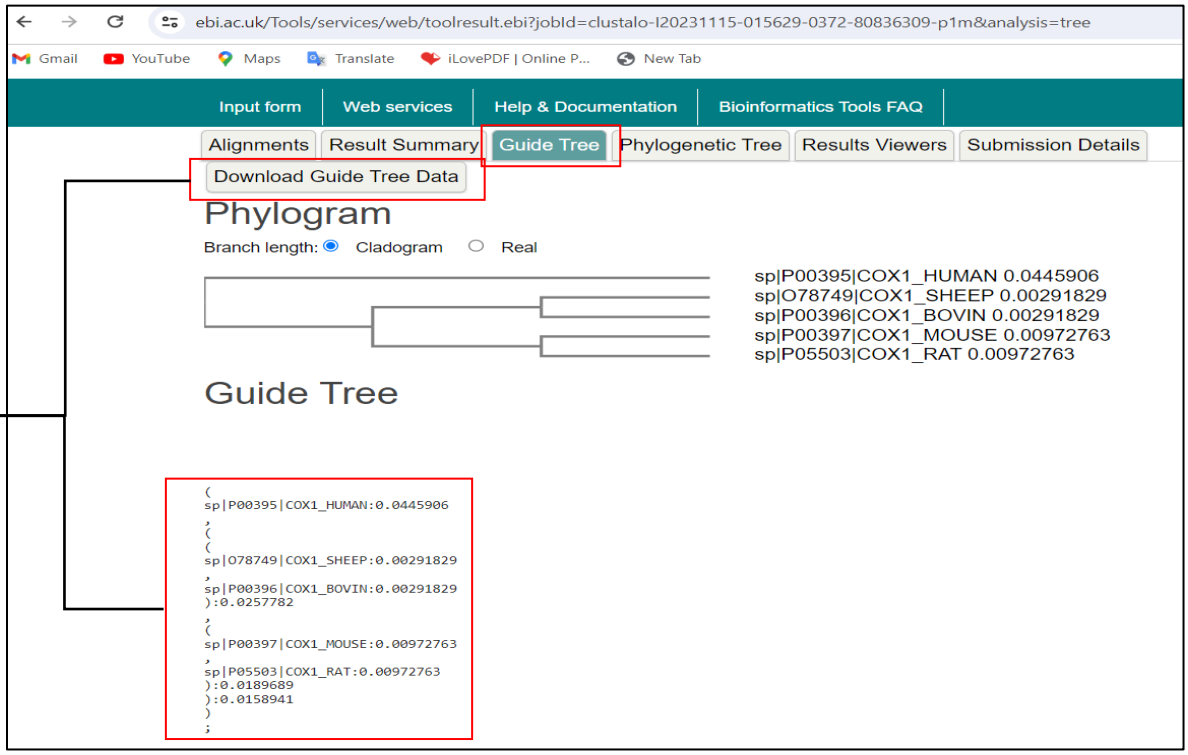


Figure 4f: Guide Tree from Clustal Omega tool

The Phylogenetic Tree Data can be downloaded in Newick format

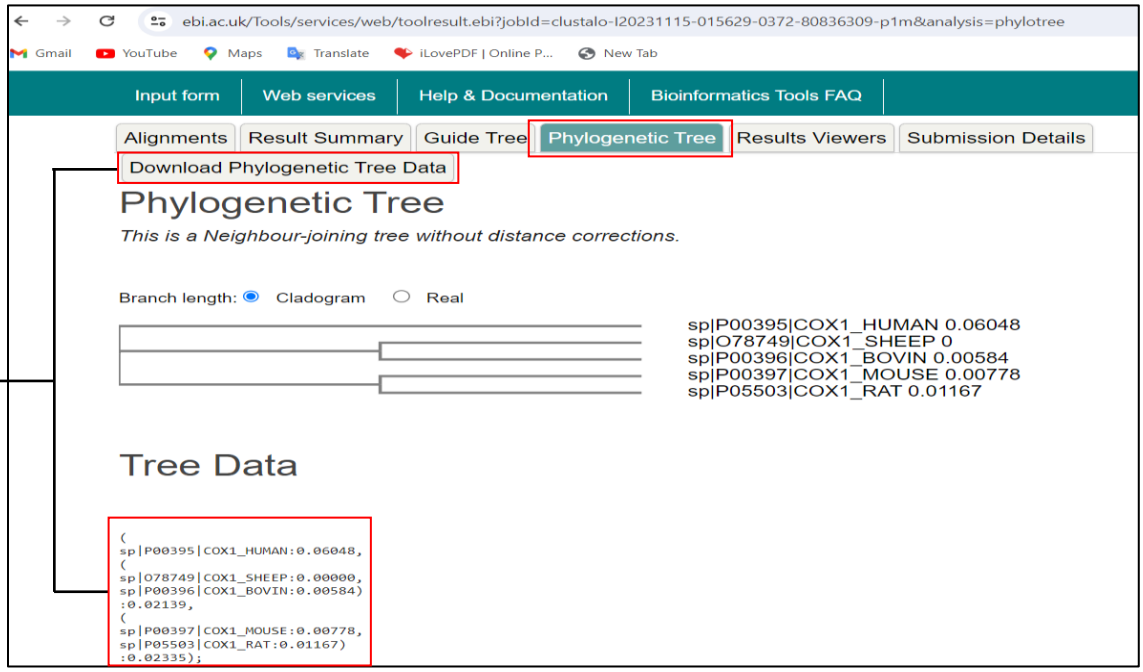


Figure 4g: Phylogenetic Tree from Clustal Omega tool

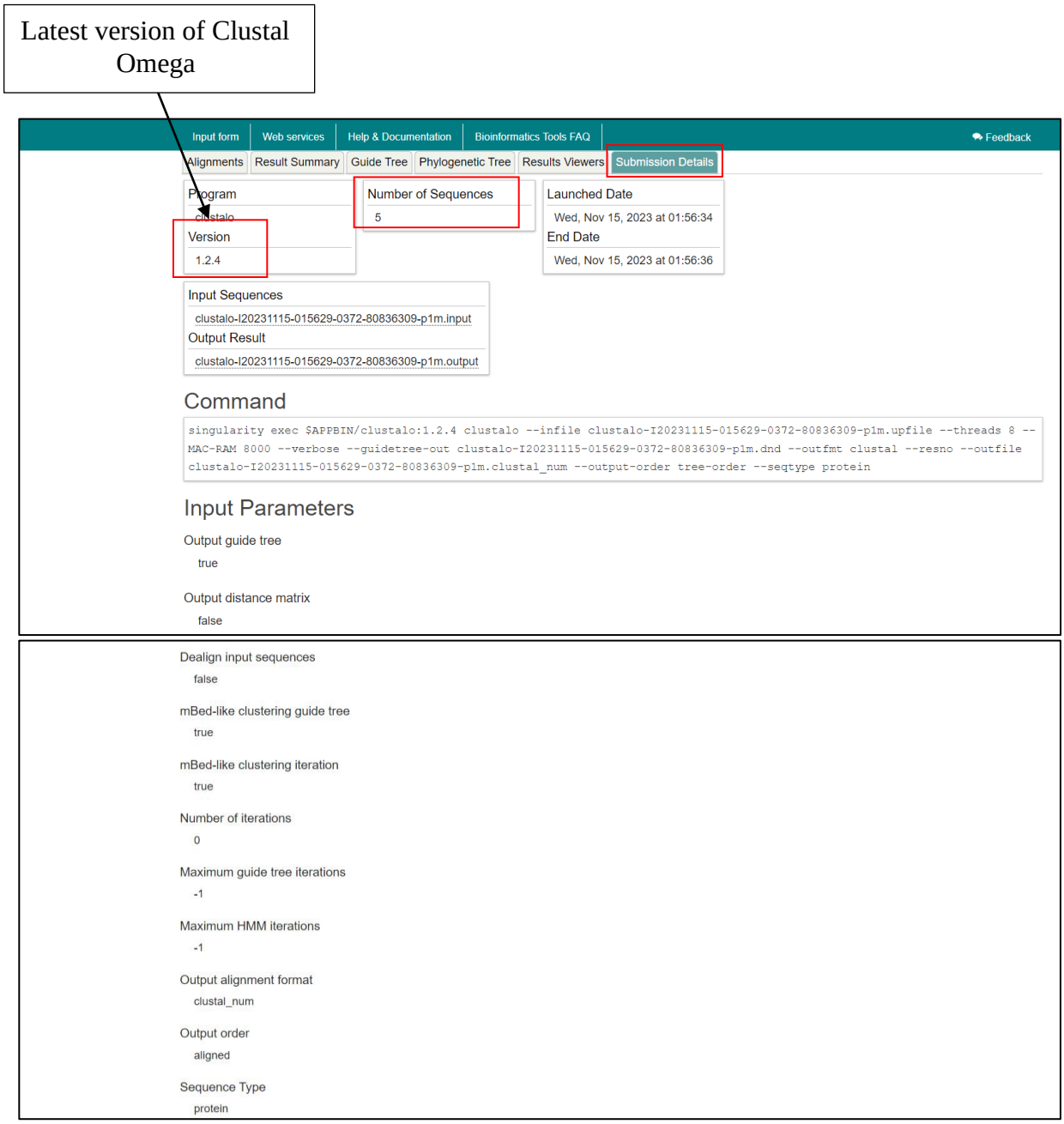


Figure 4h: Submission Details in Clustal Omega tool

TOOL 2: T-COFFEE

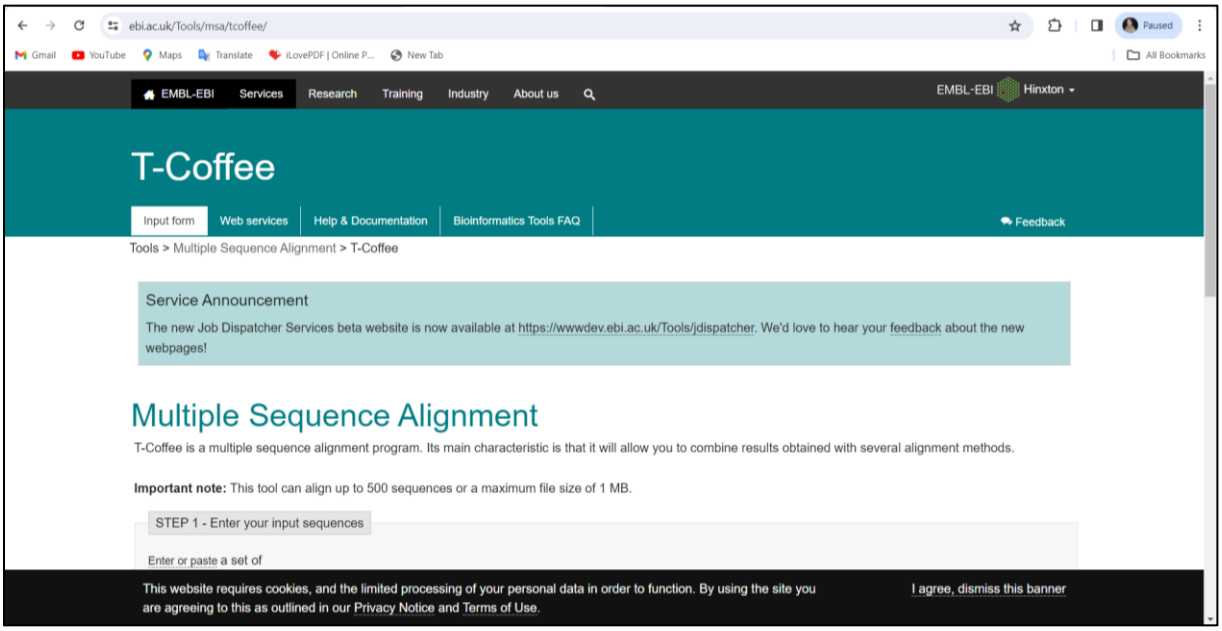


Figure 5: Homepage for T-Coffee

Uploaded sequence in FASTA Format

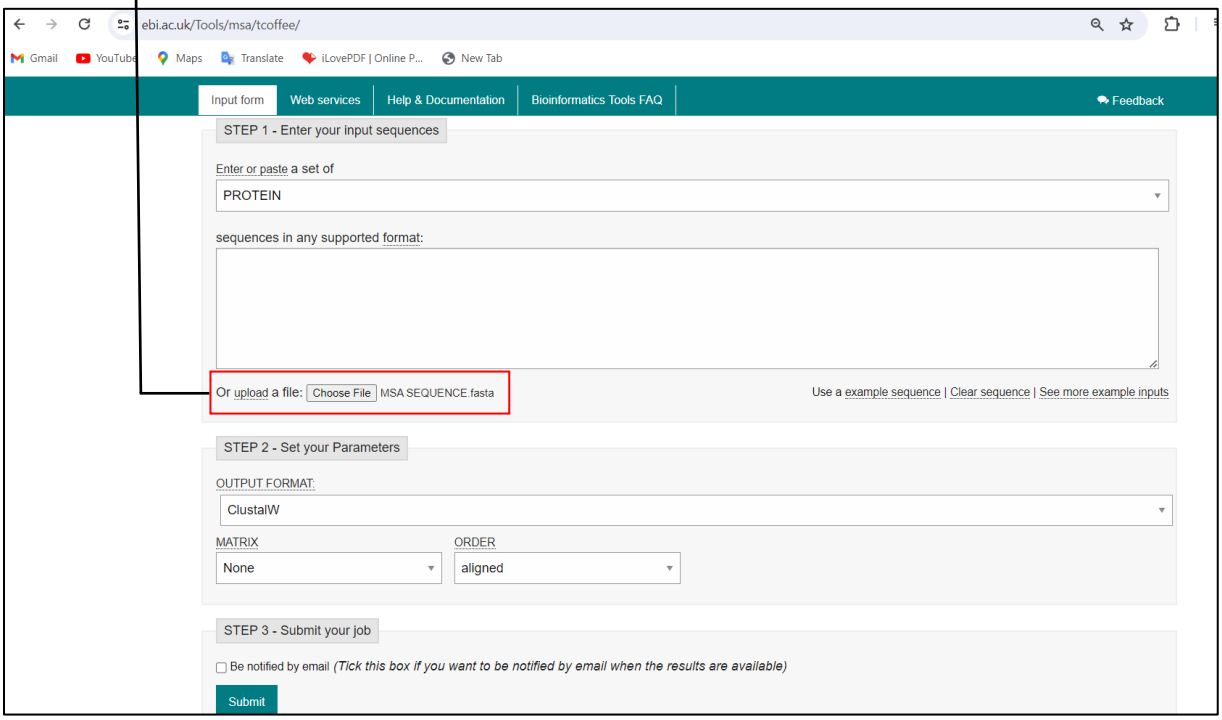


Figure 5a: Input form for submitting the data to T-Coffee tool

Set of Aggregate Results

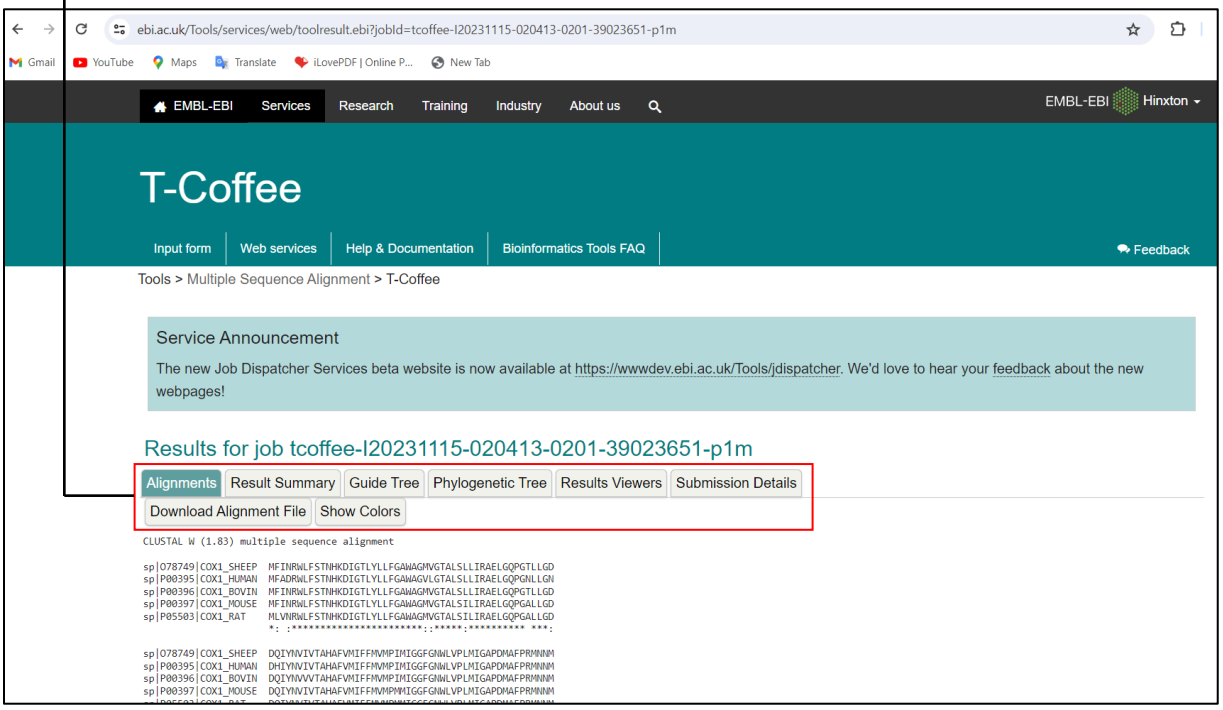


Figure 5b: Result for the query in T-Coffee tool

TOOL 3: MUSCLE

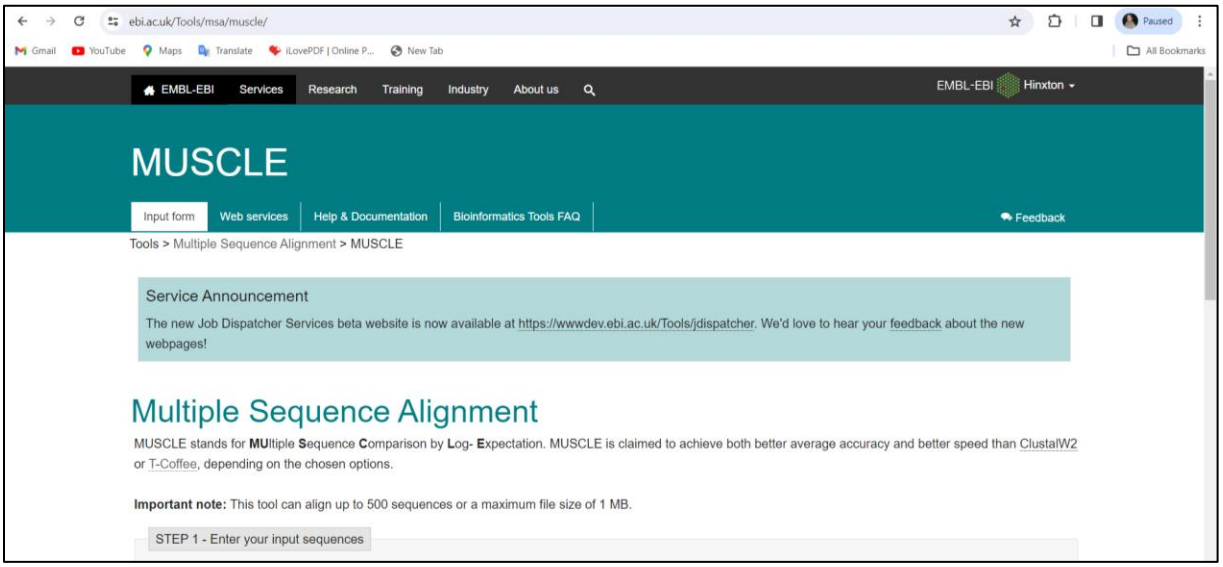


Figure 6: Homepage for MUSCLE tool

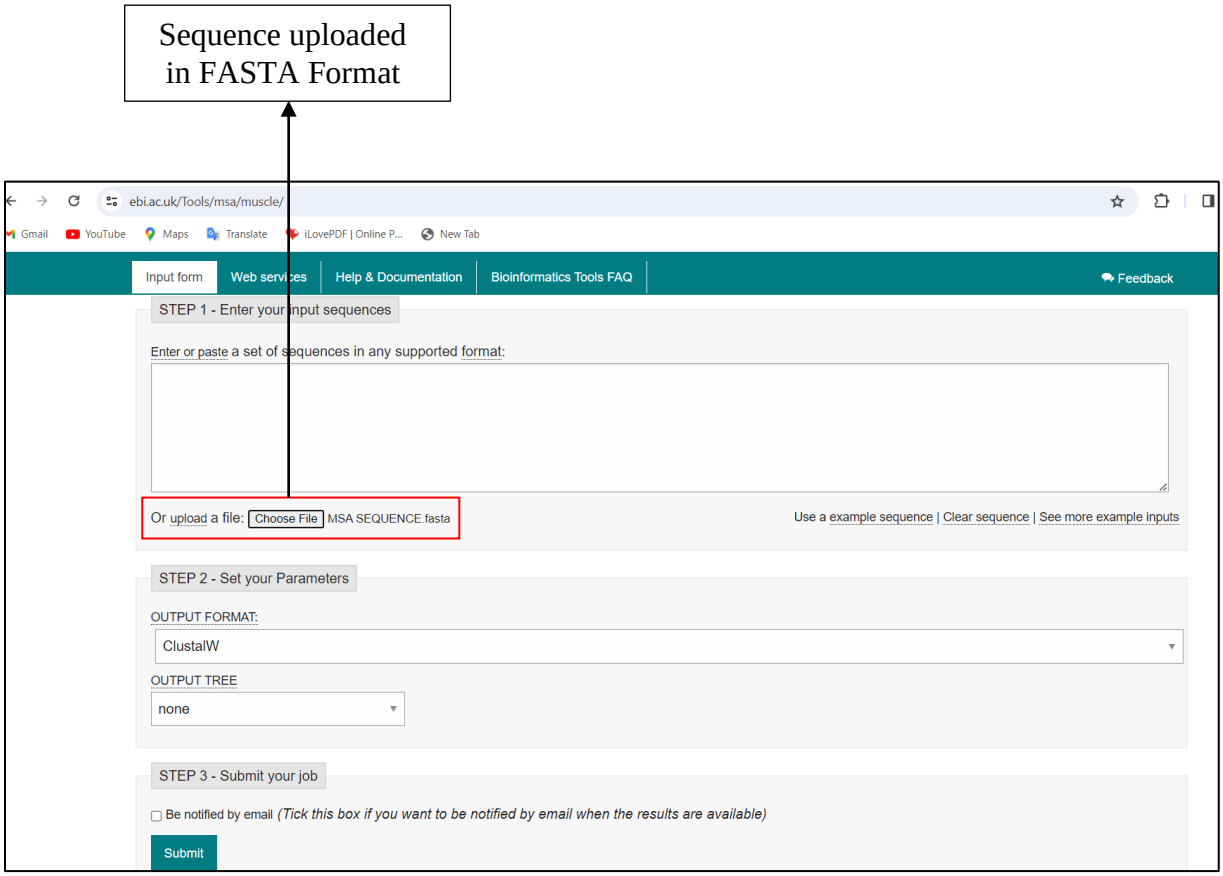


Figure 6a: Input form for submitting the data to MUSCLE tool

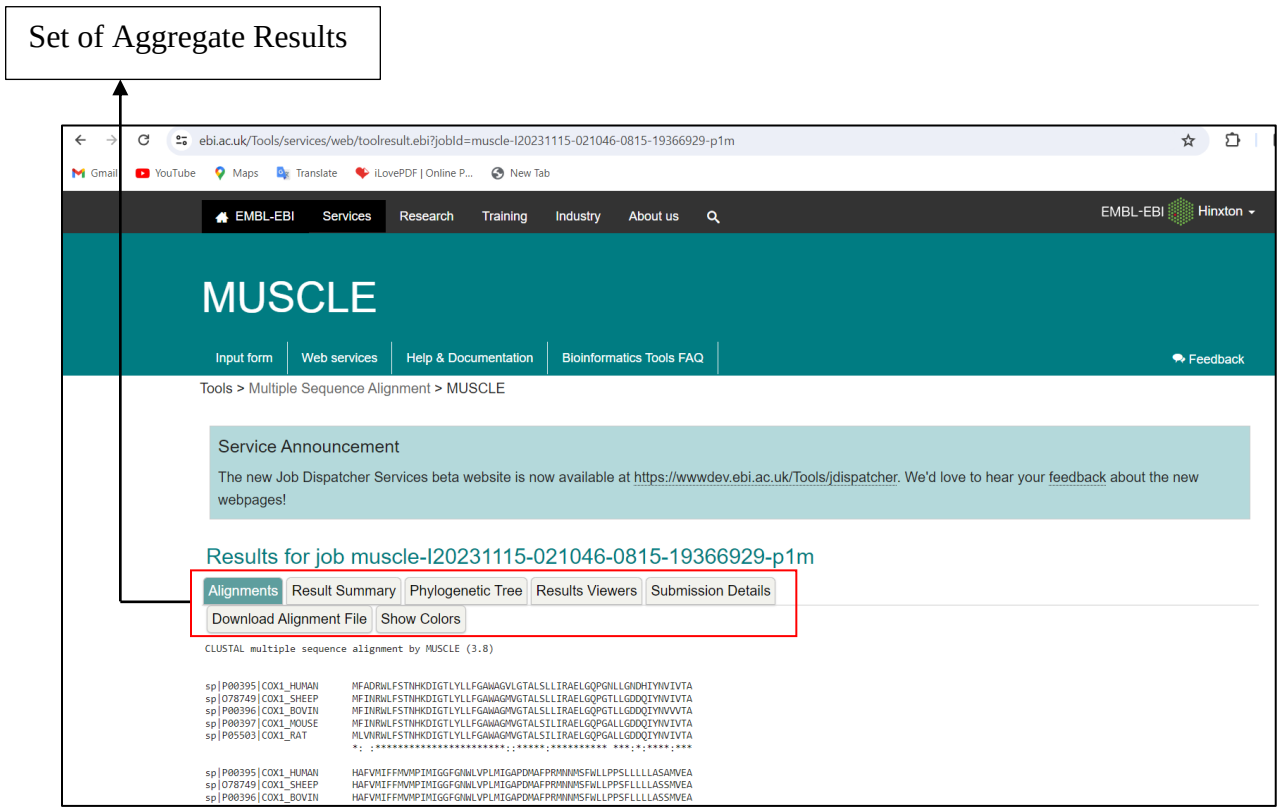


Figure 6b: Result for the query in MUSCLE tool

RESULTS:

In this comprehensive study, we conducted multiple sequence alignments (MSA) using three prominent tools: Clustal Omega, T-Coffee, and MUSCLE under default parameters to analyze the genetic relationships among *Homo sapiens* (UniProt ID: P00395), *Bos taurus* (UniProt ID: P00396), *Ovis aries* (UniProt ID: 078749), *Mus musculus* (UniProt ID: P00397), and *Rattus norvegicus* (UniProt ID: P05503). In the phylogenetic tree analysis, specific scores were observed across all three tools, providing insights into the genetic relationships among the analyzed species, as detailed in the interpretive table below.

Species	Clustal Omega	T-Coffee	MUSCLE
Homo sapiens (Human)	0.06048	0.06048	0.06048
Bos taurus (Cattle)	0.00584	0.00584	0.00584
Ovis aries (Sheep)	0	0	0
Rattus norvegicus (Rat)	0.01167	0.01167	0.01167
Mus musculus (Mouse)	0.00778	0.00778	0.00778

Homo sapiens is positioned on a distinct branch, demonstrating a genetic divergence with a consistent score of 0.06048 across all three tools. *Bos taurus* (cattle) and *Ovis aries* (sheep) form sister taxa, with *Bos taurus* achieving a sequence similarity score of 0.00584, and *Ovis aries* obtaining a perfect match with a score of 0. Similarly, *Mus musculus* (house mouse) and *Rattus norvegicus* (brown rat) constitute sister taxa, exhibiting identical scores of 0.00778 and 0.01167, respectively.

The scores for all the five species are consistent in all three tools. The uniformity in scores across all three tools highlights the robustness and reliability of the observed genetic similarities. These scores depict the genetic similarity between the sequences of each species, with higher scores indicating greater similarity.

The cladogram visually portrays evolutionary relationships based on alignment scores, emphasizing Homo sapiens' genetic distinctiveness, while other species exhibit varying degrees of genetic similarity. Alignments were displayed to further elucidate these relationships. Phylogenetic and guide trees were generated for each tool based on the aligned sequences.

CONCLUSION:

In conclusion, our practical employed three MSA tools: Clustal Omega, T-Coffee, and MUSCLE to align the “cytochrome c oxidase subunit 1” protein across five species: *Homo sapiens, Bos taurus, Ovis aries, Mus musculus, and Rattus norvegicus*. The comprehensive results, including Alignments, phylogenetic and guide trees, result summary, and submission details, offered a holistic understanding of genetic relationships.

Figures of Results from Clustal Omega were exclusively showcased due to its widespread recognition, user-friendly output, and consistent alignment scores, reinforcing its reliability.

Clustal Omega is fast and scalable. The uniform scores across tools underscored robust sequence similarity assessments.

Notably, *Homo sapiens* stood out as genetically distinct with a score of 0.06048, elucidated by the phylogenetic tree depicting inferred evolutionary history. The guide tree played a pivotal role in guiding the alignment process, influencing the creation of the final MSA. This study contributes insights into conserved regions and evolutionary dynamics, emphasizing the reliability of genetic similarities across MSA tools.

REFERENCES:

1. Institute, E. B. (n.d.). *EMBL-EBI homepage*. Retrieved November 23, 2023, from https://www.ebi.ac.uk/
2. *Bioinformatics tools for multiple sequence alignment < embl-ebi*. (n.d.). Retrieved November 23, 2023, from https://www.ebi.ac.uk/Tools/msa/
3. *Clustal omega < multiple sequence alignment < embl-ebi*. (n.d.). Retrieved November 23, 2023, from https://www.ebi.ac.uk/Tools/msa/clustalo/
4. *Muscle < multiple sequence alignment < embl-ebi*. (n.d.). Retrieved November 23, 2023, from https://www.ebi.ac.uk/Tools/msa/muscle/
5. Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., & Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, *38*(Web Server issue), W695-699. https://doi.org/10.1093/nar/gkq3
6. Shullia, N. I., Kuswati, K., Kurniawan, A., & Fiarani, H. S. (2023). In Silico Primer Design for geographical detection of *Apis florea* using Cytochrome c oxidase subunit 1 (Cox1) gene. *Life Science and Biotechnology*, *1*(1), 27. https://doi.org/10.19184/lsb.v1i1.40052
7. Timón-Gómez, A., Nývltová, E., Abriata, L. A., Vila, A. J., Hosler, J., & Barrientos, A. (2018). Mitochondrial cytochrome c oxidase biogenesis: Recent developments. *Seminars in Cell & Developmental Biology*, *76*, 163–178. https://doi.org/10.1016/j.semcdb.2017.08.055
8. Bank, R. P. D. (n.d.). *RCSB PDB - 5XDQ: Bovine heart cytochrome c oxidase in the fully oxidized state with pH 7.3 at 1.77 angstrom resolution*. Retrieved November 14, 2023, from https://www.rcsb.org/structure/5xdq

DATE: 10/11/2023

WEBLEM 8
INTRODUCTION TO RESCTRICTION ENZYME DATABASE (REBASE)
(URL: http://rebase.neb.com/rebase/rebase.html)

REBASE—a database for DNA restriction and modification: enzymes, genes and genomes which is a comprehensive database of information about restriction enzymes, DNA methyltransferases and related proteins involved in the biological process of restriction–modification (R–M). It contains fully referenced information about recognition and cleavage sites, isoschizomers, neoschizomers, commercial availability, methylation sensitivity,crystal and sequence data. Experimentally characterized homing endonucleases are also included. The fastest growing segment of REBASE contains the putative R–M systems found in the sequence databases. Comprehensive descriptions of the R–M content of all fully sequenced genomes are available including summary schematics. The contents of REBASE may be browsed from the web (http://rebase.neb.com) and selected compilations can be downloaded by ftp (ftp.neb.com).

The REBASE web site (http://rebase.neb.com) summarizes all information known about every restriction enzyme and any associated proteins. This includes the recognition sequences, cleavage sites, source, commercial availability, sequence data, crystal structure information, isoschizomers and methylation sensitivity. Within the reference section of REBASE, links are maintained to the full text of all papers whenever they are readily available on the web. Also, there is extensive reciprocal cross-referencing between REBASE and NCBI, including links to GenBank and PubMed and NCBI's LinkOut utility. Links to other major databases such as UniProt, PDB and Pfam are also maintained. There are currently 3945 biochemically or genetically characterized restriction enzymes in REBASE and of the 3834 Type II restriction enzymes, 299 distinct specificities are known. Six hundred and forty-one restriction enzymes are commercially available, including 235 distinct specificities.

Figure 1: Homepage of REBASE Database

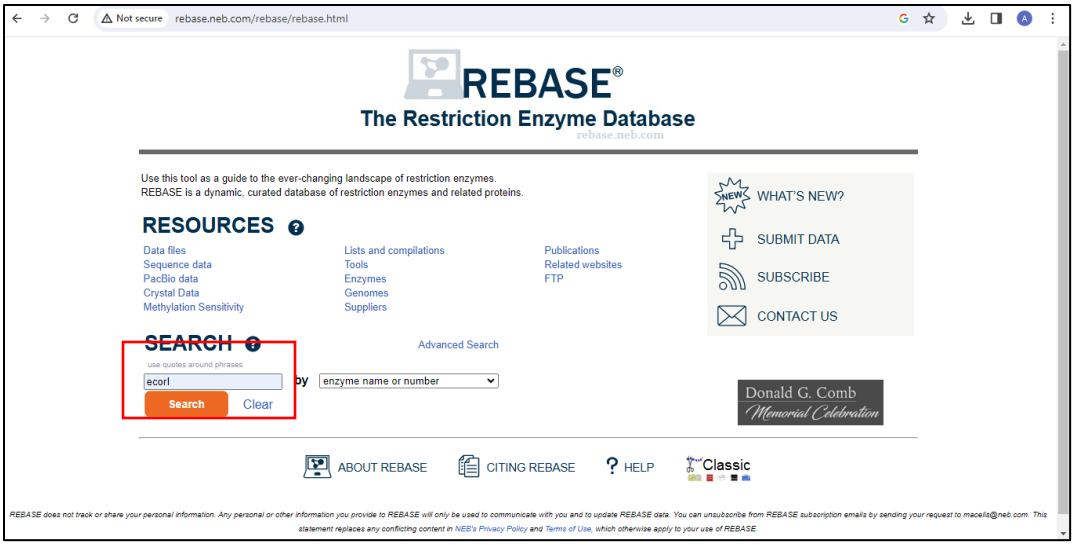


Figure 2: Search for EcoRI restriction endonuclease enzyme isolated from species E. coli

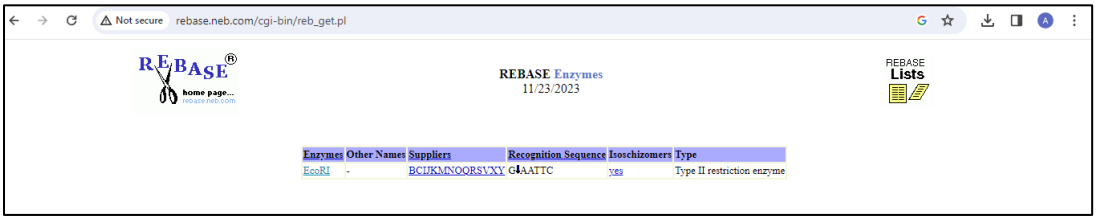


Figure 3: Result for EcoRI restriction endonuclease enzyme

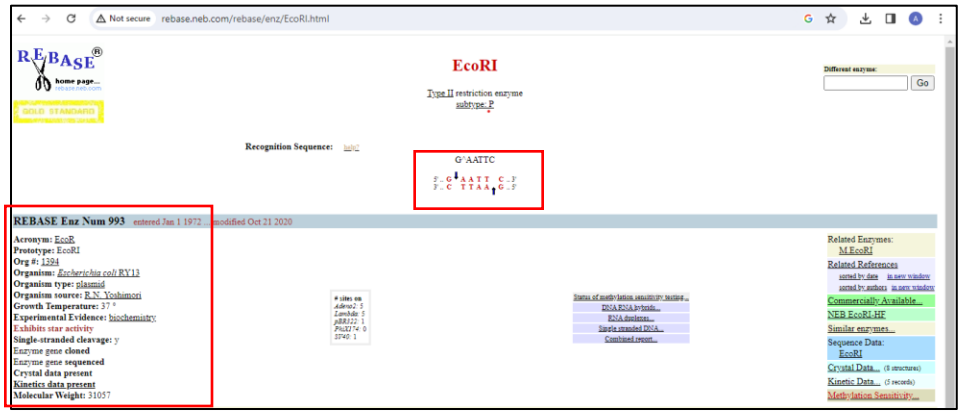


Figure 3a: Detailed result for EcoRI restriction endonuclease enzyme

(Red Marker Box: indicates the cleavage site for EcoRI and Description for restriction enzyme)

The rate of discovery of new putative restriction and modification genes is rising rapidly. In contrast, the rate at which candidates are being characterized biochemically has actually dropped to the level it was three decades ago. Nevertheless, because of the large number of sequenced examples of biochemically characterized restriction systems, the putative recognition sequences of predicted restriction enzymes and DNA methyltransferases can be inferred. Currently, all new sequences entering GenBank are checked using data mining techniques for the presence of R–M

systems and, following extensive manual checking, the resulting inferences are all included within REBASE where they are clearly marked as predictions. When analyzing DNA sequence data, it is the DNA methyltransferase genes that are the more reliable indicators of an R–M system and the presence, proper order and characteristic spacing of well-conserved motifs that are used to suggest likely candidates. It should be noted that at the present time it is not possible to distinguish DNA methyltransferases reliably enough to be completely confident in the assignments.

Some RNA and protein methyltransferases can sometimes be confused for DNA methyltransferases as is widely reflected by the annotations found in GenBank files. In general, REBASE takes a liberal approach and includes all likely candidates until it becomes clear that non-DNA methyltransferases have been included erroneously and then these are culled from the database. The more widely divergent genes that encode the restriction enzymes always reside close to the genes for their cognate methyltransferases, but often they cannot be recognized directly because they are a rapidly evolving set of genes and frequently lack any sequence similarity to any other genes in GenBank. However, other methods can sometimes be used to infer their presence such as the analysis of shotgun sequence data from which missing clones can be inferred to be caused by the presence of active restriction enzyme genes.

Given the wealth of experimental data, both published and unpublished, contained within REBASE, it can be an especially valuable resource during the annotation of bacterial and archaeal genomes. With the plethora of restriction systems that occur in all sequenced microbial genomes, annotators are encouraged to use the resources of the REBASE database or to contact the REBASE staff if help is needed. Custom analyses of unpublished genome sequence data are carried out upon request. From the REBASE web site users have a variety of resources available that facilitate the analysis of sequence information including tools for analyzing sequences (REBASE tools) that allow restriction enzyme recognition sites to be found in submitted sequences (NEBcutter) and an implementation of BLAST to allow searching against all sequences in REBASE. Specialty lists of sequence data (REBASE lists) such as all known Type II restriction enzyme genes, all known Type I specificity subunit genes, etc., are available for download. The coming year will see some major additions to REBASE in terms of new sequence acquisitions, such as the inclusion of all metagenomics sequence data (only partially analyzed to date) and a tool to permit users to perform their own analysis of newly sequenced genomes.

Key features and information provided by REBASE include:

1. **Enzyme Information:** Details about various restriction enzymes, including their names, sources (organisms where they are found), recognition sequences (the specific DNA sequence they recognize and cut), and other properties like cleavage patterns, isoschizomers (enzymes recognizing the same sequence), and methylation sensitivity.
2. **DNA Recognition Sequences:** The database includes information on the specific DNA sequences recognized by restriction enzymes. This is crucial in molecular biology experiments for cloning, DNA manipulation, and other techniques where precise DNA cutting is required.

3. **Reference Information:** REBASE provides relevant references, citations, and links to scientific literature where the properties and functions of these enzymes have been documented and characterized.
4. **Search and Analysis Tools:** The database offers search functionalities allowing users to query specific enzymes or DNA sequences, enabling researchers to find enzymes that recognize particular sequences or properties that fit their experimental needs.
5. **Updates and Annotations:** REBASE is regularly updated with new discoveries, additions, and modifications to ensure that researchers have access to the most current information on restriction enzymes.

Thus, this database serves as a valuable resource for researchers in molecular biology, genetics, genomics, and related fields, aiding them in the design and execution of experiments involving DNA manipulation, gene editing, recombinant DNA technology, and more.

REFERNCES:

1. Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic acids research*, *38*(Database issue), D234–D236. https://doi.org/10.1093/nar/gkp874
2. Richard J. Roberts, Tamas Vincze, Janos Posfai, *et al;* REBASE—a database for DNA restriction and modification: enzymes, genes and genomes; Nucleic Acids Res. 2010 Jan; 38(Database issue): D234–D236.Published online 2009 Oct 21. https://doi.org/10.1093/nar/gkp874

DATE: 10/11/2023

WEBLEM 9
INTRODUCTION TO OMICS AND APPLICATIONS OF BIOINFORMATICS

The human history has witnessed the rapid development of technologies such as high-throughput sequencing and mass spectrometry that led to the concept of "omics" and methodological advancement in systematically interrogating a cellular system. Yet, the ever-growing types of molecules and regulatory mechanisms being discovered have been persistently transforming our understandings on the cellular machinery. This renders cell omics seemingly, like the universe, expand with no limit and our goal toward the complete harness of the cellular system merely impossible. Therefore, it is imperative to review what has been done and is being done to predict what can be done toward the translation of omics information to disease control with minimal cell perturbation. With a focus on the "four big omics," i.e., genomics, transcriptomics, proteomics, metabolomics, this hierarchies of these omics together with their epiomics and interactomics, and review technologies developed for interrogation.

The branches of science known informally as omics are various disciplines in biology whose names end in the suffix -omics, such as genomics, proteomics, metabolomics, metagenomics, phenomics and transcriptomics. Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. Thus, "OMICS" is defined as probing and analyzing large amount of data representing the structure and function of an entire makeup of a given biological system at a particular level, which has substantially revolutionized the methodologies in interrogating biological systems.

Various disciplines of OMICS are as follow:

1. **Genomics:** Genomics involves the study of an organism's complete set of DNA, including its genes and their functions. It explores the structure, function, evolution, mapping, and editing of genomes.
2. **Proteomics:** Proteomics is the study of the entire set of proteins expressed by a cell, tissue, or organism. It involves the identification, quantification, structure, function, and interactions of proteins.
3. **Transcriptomics:** Transcriptomics focuses on studying the complete set of RNA transcripts (messenger RNA, non-coding RNA, etc.) produced by the genome under specific conditions or in specific cells. It provides insights into gene expression patterns and regulation.
4. **Metabolomics:** Metabolomics deals with the comprehensive analysis of all metabolites present within a biological system. It aims to identify and quantify small molecules (metabolites) to understand metabolic pathways and their regulation.

5. **Epigenomics:** It explores modifications to the DNA that do not change the underlying genetic code but affect gene expression. It involves the study of epigenetic modifications like DNA methylation, histone modifications, and chromatin structure.
6. **Pharmacogenomics:** This field examines how an individual's genetic makeup influences their response to drugs. It involves identifying genetic variations that impact drug efficacy, toxicity, and dosage requirements.

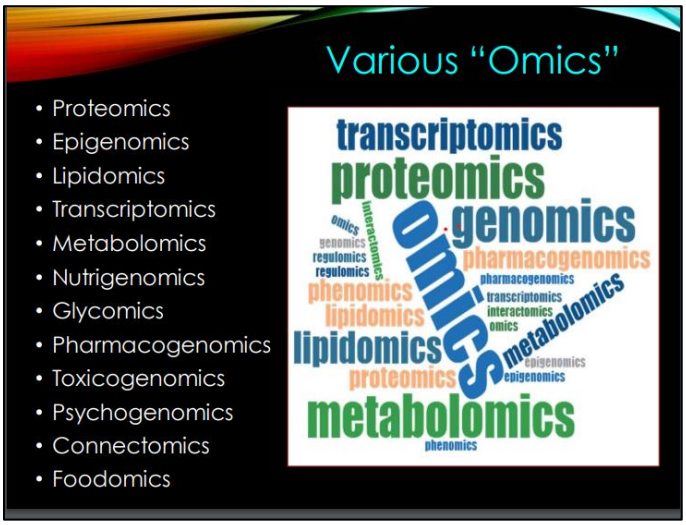


Figure 1: Various disciplines of OMICS studies

Ever since the establishment of the first high-throughput technology, DNA microarray, technologies for omics exploration have been developed by leaps and bounds. Following the central dogma, omics technologies have been used to capture the static genomic alterations, temporal transcriptomic perturbations and alternative splicing, as well as spatio-temporal proteomic dynamics and post translational modifications (PTMs). Beyond this, omics technologies have been expanded to analyze various omics at the epi-level (such as epigenome, epitranscriptome, epiproteome that are defined as the collection of all modifications of the referred omics beyond information it covered in a single cell), molecular interactions (i.e., varied levels of interactome), and disease associated hallmarks as metabolome and immunome.

Multi-omics integration has become a prevailing trend for constructing a comprehensive causal relationship between molecular signatures and phenotypic manifestations of a particular disease, and single cell sequencing offers additional resolving power that enables investigations at a single cell level. This rapidly-developing and ever-growing field, omics, has empowered us to uncover the intricate molecular mechanism underlying different phenotypic manifestations of disordered traits in an overwhelming and systematic manner at a high accuracy. However, the complexity of the cellular behavior and its decision-making system may persistently drive the establishment of novel omics and associated techniques.

The impact of omics is most apparent in medicine. Sequencing of the human genome, for example, has fueled advances in personalized medicine, in which decisions about disease prevention,

diagnosis, and treatment are tailored to patients based on information derived from genetic and genomic research. In particular, genomic data have played key roles in the development of predictive models of disease and in informing therapeutic decisions, such as in the treatment of cancer. Similar links between omics and personalized medicine have emerged from metabolomics with the discovery of new biomarkers of disease. An example is the investigation of disturbances in metabolic pathways that affect levels of substances such as fatty acids and bile acids; this work has led to the identification of biomarkers with the potential to improve early diagnosis of hepatocellular carcinoma. Nonetheless, significant challenges remain in the omics sciences, especially concerning data complexity, data management, and the integration of data from omics studies with data from other sources, such as clinical data gathered during routine physician visits. Other challenges are more fundamental, such as in assay development and refinement. In large-scale proteomic analysis, for instance, agents designed to bind to specific proteins often are lacking in sensitivity and specificity, decreasing their affinity for the proteins of interest and resulting in suboptimal protein capture.

Bioinformatics, a well-established multidisciplinary branch of sciences, often known as computational biology, is gaining immense importance in the era of omics marked by generation of huge biological data constantly. The advances in molecular biology led to the genesis of bioinformatics for the sole purpose of storing, retrieving and analyzing nucleotide and protein sequences to get an insight into life processes. The bioinformatics primarily deals with data-curation, developing tools, assisting data interpretation and analysis using web-based resources in a biologically meaningful manner. The computational expertise along with a good understating of biological processes are associated with developing appropriate algorithms for sequence comparisons, phylogenetic/evolutionary tree construction, specific pattern recognitions, sequence-structure-function elucidation, annotating sequences, deciphering metabolic pathways, gene regulation and expression, drug designing etc.

The science of ‘omics’ reflects characterization and quantification of pools of diverse biological molecules associated with the structure, function, and dynamics of organisms The new approach for real time understanding of biology is system biology which combines the information of different field to simulate and analyze the networks, pathways, the spatial and temporal relations that exist in biological systems. The extensive data being generated by experimentation related with diverse fields of omics is being successfully managed by bioinformatics experts through the development of appropriate user-friendly biological databases with provision for open access to researchers globally.

The data management and data mining are two important bottlenecks for omics-based research and this demands bioinformatics intervention. Bioinformatics aims to establish standard formats by using algorithms based on mathematical and statistical models and developing efficient methods for storing, retrieving and sharing high-throughput data in the era of omics. The data analysis, molecular modeling, predictions, simulation, phylogenetic analysis, sequence comparison are in the purview of bioinformatics and accordingly there has been development of appropriate in-silico tools. Thus, Bioinformatics acts as a bridge between biological sciences and

computational analysis, enabling researchers to derive meaningful insights from the vast amount of data produced by OMICS technologies. Its integration is fundamental in advancing our understanding of biological systems, diseases, drug development, and personalized healthcare. One of the widely used OMICS open-source platform that can be used to access, discover and disseminate omics datasets is OMICSDI (The omics discovery REST interface).

OMICSDI:

The Omics Discovery Index is an open-source platform that can be used to access, discover and disseminate omics datasets. OmicsDI integrates proteomics, genomics, metabolomics, models and transcriptomics datasets. Using an efficient indexing system, OmicsDI integrates different biological entities including genes, transcripts, proteins, metabolites and the corresponding publications from PubMed. In addition, it implements a group of pipelines to estimate the impact of each dataset by tracing the number of citations, reanalysis and biological entities reported by each dataset.

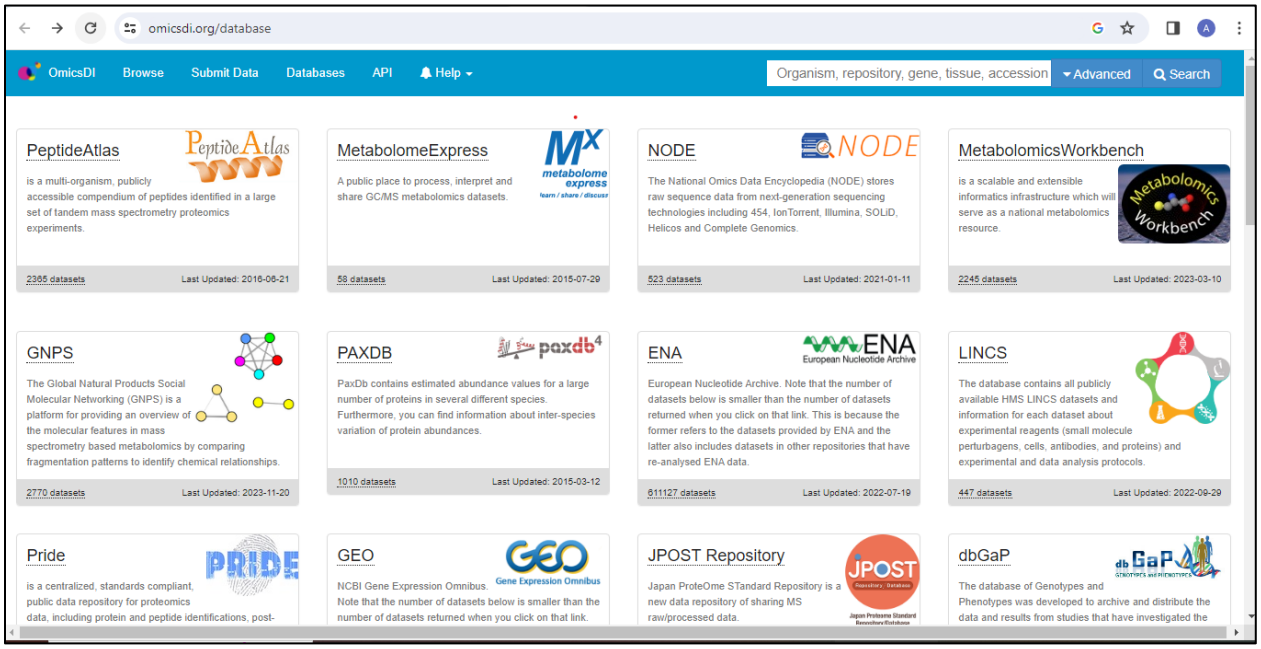


Figure 2: Homepage of OMICSDI server

Various bioinformatics database for OMICS studies i.e. Database resources for genomics, proteomics, and other omics data studies:

A) Genomics Databases:

1. **NCBI (National Center for Biotechnology Information):** Provides a wide range of genomic resources including GenBank, RefSeq, dbSNP, and more.
2. **Ensembl:** A genome browser for vertebrate genomes, offering comprehensive genomic annotations and comparative genomics.

3. **UCSC Genome Browser:** Allows visualization and analysis of genomes, including comparative genomics and epigenomic data.
4. **1000 Genomes Project:** Provides a comprehensive map of human genetic variation.
5. **ExAC (Exome Aggregation Consortium):** Aggregates exome sequencing data from over 60,000 individuals across diverse populations.
6. **dbSNP:** A database of single nucleotide polymorphisms (SNPs) and other variations in the human genome.

B) Proteomics Databases:

1. **Expasy server:** A comprehensive a resource portal that provide information about genomics, proteomics, structure analysis, systems biology, evolutionary biology, population genetics, transcriptomics, glycomics, medicinal chemistry, etc.
2. **UniProt:** A comprehensive resource for protein sequence and functional information.
3. **PeptideAtlas:** A repository for mass spectrometry-based proteomics data.
4. **PRIDE Database:** A centralized repository for mass spectrometry-based proteomics data.
5. **Human Protein Atlas:** Provides information on the expression and localization of proteins in a wide range of human tissues and cells.

C) Transcriptomics Databases:

1. **GTEx (Genotype-Tissue Expression):** Offers gene expression data across multiple human tissues.
2. **ArrayExpress:** An archive of functional genomics data, including microarray and next-generation sequencing experiments.
3. **TCGA (The Cancer Genome Atlas):** Provides multi-dimensional maps of the key genomic changes in various types of cancer.

D) Metabolomics Databases:

1. **HMDB (Human Metabolome Database):** Contains comprehensive information about metabolites found in the human body.
2. **MetaboLights:** An open-access database for metabolomics experiments and data.
3. **RefMet:** A Reference list of Metabolite names.
4. **MetaCyc:** A broad metabolic pathways and enzymes database.
5. **Kyoto Encyclopedia of Genes and Genomes (KEGG):** A collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances

E) Systems Biology and Integrated Omics Databases:

1. **STRING:** A database for protein-protein interactions and functional network analysis.
2. **Reactome:** A curated knowledgebase for biological pathways and reactions.
3. **BioGRID:** A database of protein and genetic interactions.

These databases serve as invaluable resources for researchers in the field of genomics, proteomics, transcriptomics, metabolomics, and other omics-related studies, providing access to vast amounts of data and tools for analysis and interpretation.

REFERENCES:

1. Xiaofeng Dai , Li Shen; Advances and Trends in Omics Technology Development; Front. Med., 01 July 2022 ; Sec. Translational Medicine; Volume 9 - 2022 | https://doi.org/10.3389/fmed.2022.911861 https://www.frontiersin.org/articles/10.3389/fmed.2022.911861/full
2. Introduction to OMICS: https://nursing.wayne.edu/ohr/j_pierce_introduction_to_omics_092220-opt.pdf
3. OMICS: https://www.britannica.com/science/omics
4. Serena Dato, Paolina Crocco, Nicola Rambaldi Miglior, et al;Review article; Omics in a Digital World: The Role of Bioinformatics in Providing New Insights Into Human Aging; Front. Genet., 10 June 2021, Sec. Genetics of Aging, Volume 12 - 2021 https://doi.org/10.3389/fgene.2021.689824
5. Bioinformatics & OMICS: https://link.springer.com/chapter/10.1007/978-0-387-72430-0_6
6. Dinesh Yadav; Relevance of Bioinformatics in the Era of Omics Driven Research; Yadav, Next Generat Sequenc & Applic 2015, 2:1 DOI: 10.4172/2469-9853.1000e102 https://www.iomcworld.com/open-access/relevance-of-bioinformatics-in-the-era-of-omics-driven-research-jngsa-1000e102.pdf
7. The omics discovery REST interface: https://academic.oup.com/nar/article/48/W1/W380/5831190