

<b>Weblem No.</b>	<b>Weblem Title</b>	<b>Page No.</b>	<b>Date</b>	<b>Sign</b>
1.	Introduction to Sequence and Structure Database.	1	28/08/2024	
1A.	To explore the UniProt Database for further study of the query of Ig alpha chain C region (UniProt ID: P01878)	11	28/08/2024	
1B.	To study and explore the protein structure for the query Dimeric Immunoglobulin A (dIgA) (PDB ID: 7JG1) using the Protein Data Bank (PDB) Database.	17	28/08/2024	
2.	Structural Antibody Database (SAbDab).	25	12/09/2024	
2A.	To study the Antibody structure for the query 'Bovine anti-HIV Fab ElsE6' (PDB ID: 8VBL) using the Structural Antibody Database (SAbDab).	29	12/09/2024	
3.	To study antibody sequence using ABCD database.	39	12/09/2024	
3A.	To study Foralumab antibody sequence using ABCD Database	41	12/09/2024	
4	Antibody numbering using Kabat and Chothia method - Write up and include KabatMan database working and an AbRSA numbering tool as a demo.	44	12/09/2024	
5.	Introduction to STCRDab database.	56	24/09/2024	
5A.	To retrieve CDR position in query 2UWE using STCRDab database.	60	24/09/2024	
6.	Introduction To Yvis Database to study variable and constant domain along with topology diagram.	69	25/09/2024	
6A.	To study the variable and constant domain along with the Topology Diagram using Yvis Platform.	81	25/09/2024	
7.	Introduction to Ag-Ab Interaction Database (AgAbDb).	92	04/10/2024	
8.	Introduction to Discotope Server 1.1 (IDEB Database) from IEDB Database.	94	25/09/2024	
8A.	To predict B-Cell epitope for query AMA1 (PDB ID: 1Z40) using Discotope Server 1.1 from IEDB Database.	104	25/09/2024	
9.	Introduction to Molecular Descriptors and PaDEL Descriptor Software.	113	01/10/2024	

9A.	To study ID, 2D & 3D descriptors for “Gallic Acid” (PubChem CID: 370) using PaDELPy Software	123	01/10/2024	
10	To understand various Web-based tools for vaccine designing -Write Up	128	4/10/2024	
11	Introduction to IEDB Database for prediction of cytotoxic and helper T cell epitopes (MHC Class I epitopes and MHC Class II epitopes).	134	28/09/2024	
11A	To Predict MHC Class I and Class II Molecules for Query Dopamine (accession no: P09172) using TepiTool.	139	28/09/2024	

**WEBLEM: 1**

**Introduction to Sequence and Structure Database**

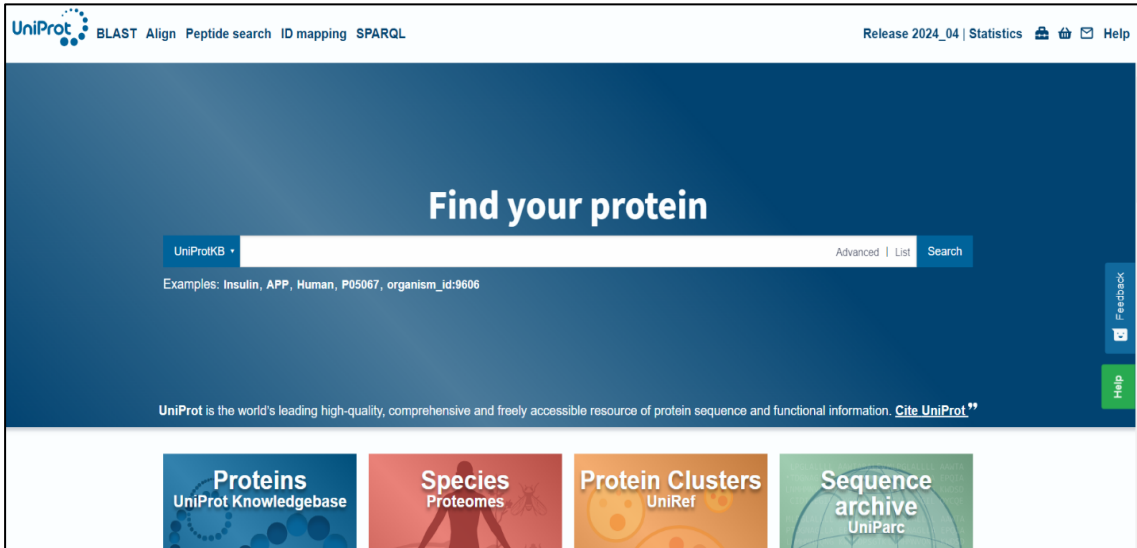
**INTRODUCTION:**

UniProt provides access to a vast collection of protein sequences, including those that are experimentally determined and those that are computationally predicted. The database includes extensive information about the function of proteins, such as their role in biological processes, molecular functions, and involvement in various pathways. UniProt integrates data on the 3D structures of proteins, where available, often linking to related resources like the Protein Data Bank (PDB). It provides cross-references to other biological databases, such as genomic, enzyme, and pathway databases, enabling a broad spectrum of data connectivity. UniProt includes information on different protein isoforms and variants, which are important for understanding protein diversity and function. The database distinguishes between manually curated entries (UniProtKB/Swiss-Prot) and automatically annotated entries (UniProtKB/TrEMBL), providing users with information on the reliability and source of the data.

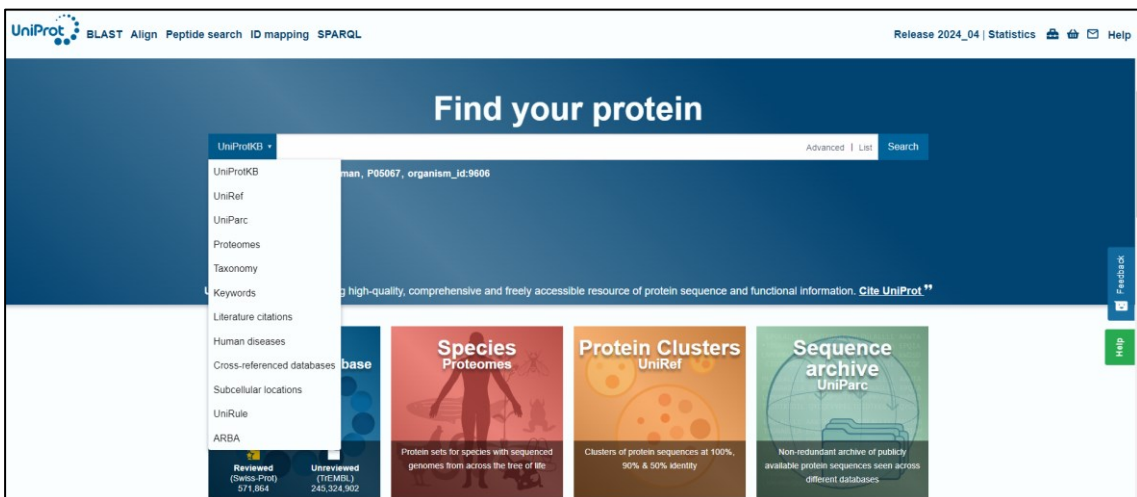
The UniProt databases support biological and biomedical research by providing a comprehensive collection of protein sequence data, along with functional information. UniProtKB combines expert-reviewed data (Swiss-Prot) with automated entries (TrEMBL). UniRef clusters sequences based on similarity, and UniParc stores all known sequences, including obsolete ones. UniProt links to 180 resources, ensuring data is findable, accessible, interoperable, and reusable (FAIR). Recognized for its data quality, UniProt received the ELIXIR Core Data Resource and CoreTrustSeal certifications. The database continually evolves, adding new sequences from projects like the Darwin Tree of Life, growing by over 65 million entries in two years.

UniProt is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation. It consists of two sections:

1. **Swiss-Prot (Reviewed):** Contains manually annotated records with data added by expert bio-curators giving information on protein function, structure, subcellular location, and molecular interactions. Each entry in UniProt/Swiss-Prot represents a single, non-redundant gene from a specific organism and all proteins and peptides transcribed by that gene are described within the record.
2. **TrEMBL (Unreviewed):** Contains computationally analyzed records with additional information transferred from related well annotated records in UniProt/Swiss-Prot (automatic annotation). There may be several separate UniProt/TrEMBL entries describing the proteins derived from a specific gene.

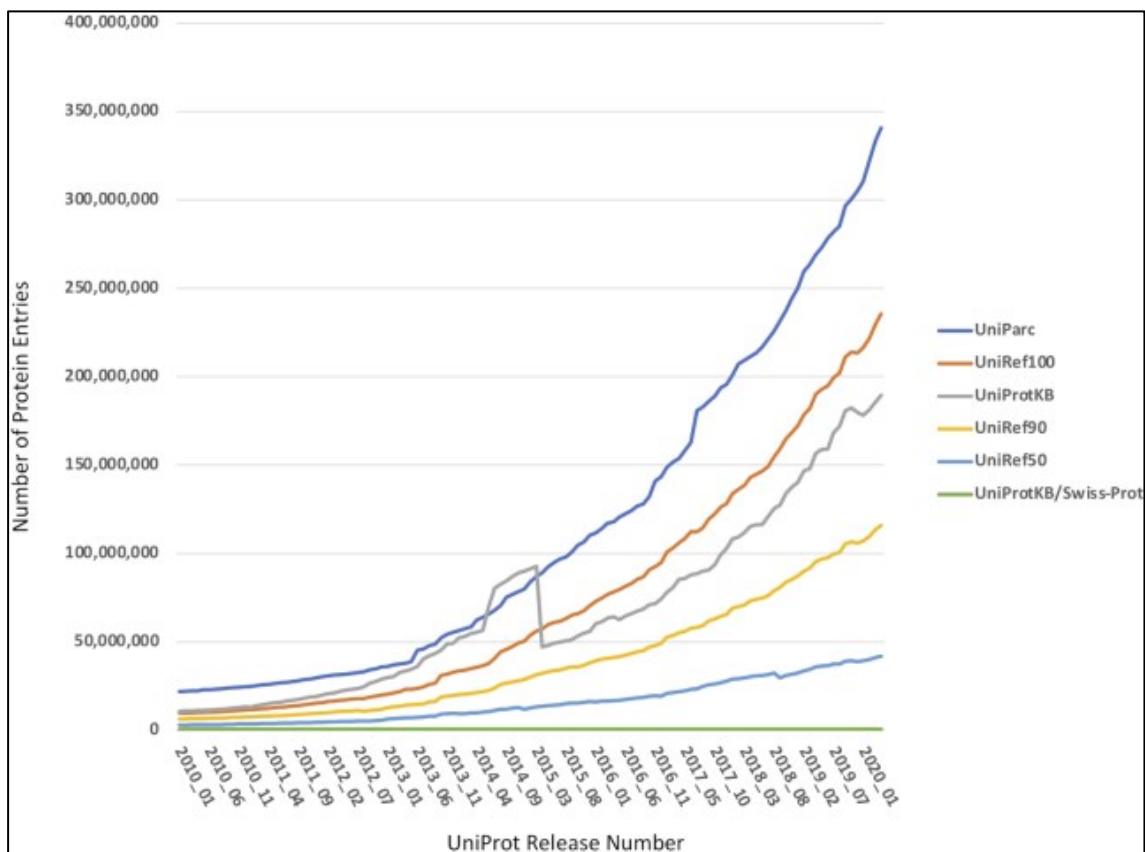


**Fig 1: Homepage of UniProt Database**



**Fig 2: Search options in UniProt Database**





**Fig 3: Growth in the number of entries in the UniProt databases over the last decade.**

### **APPLICATIONS:**

1. **Protein Function Prediction:** UniProt provides functional information about proteins, helping researchers predict the roles of unknown proteins based on sequence data.
2. **Drug Discovery:** By understanding protein structures and functions, researchers can identify potential drug targets and design therapies, especially in areas like cancer and infectious diseases.
3. **Genomics and Proteomics Research:** UniProt is essential for annotating genomes and identifying proteins in proteomics studies, aiding in the understanding of cellular functions.
4. **Evolutionary Studies:** Researchers use UniProt data to study protein evolution and compare protein sequences across species.
5. **Disease Research:** UniProt helps link genetic mutations to protein functions, assisting in the study of genetic disorders and personalized medicine.
6. **Metagenomics and Environmental Studies:** UniProt data is used to analyze microbial communities in environments, supporting research in biodiversity and ecosystem functioning.

## **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 most 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, because 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 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 as compared 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:

The Protein Data 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 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.
6. 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 like 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) Databases:**

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](http://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 Databases:**

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>
13. Shaik, N. A., Hakeem, K. R., Banaganapalli, B., & Elango, R. (2019). *Essentials of Bioinformatics, Volume I: Understanding Bioinformatics: Genes to Proteins.* Springer.
14. EMBL-EBI. (n.d.). *The UniProt databases.* UniProt. <https://www.ebi.ac.uk/training/online/courses/uniprot-exploring-protein-sequence-and-functional-info/what-is-uniprot/the-uniprot-databases/>
15. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
16. Garcia L., Bolleman J., Gehant S., Redaschi N., Martin M., Consortium UniProt. FAIR adoption, assessment, and challenges at UniProt. *Sci Data*. 2019



**WEBLEM: 1(A)**  
**UniProt Database**  
**(URL: <https://www.uniprot.org/>)**

**AIM:**

To explore the UniProt Database for further study of the query of Ig alpha chain C region (UniProt ID: P01878).

**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).

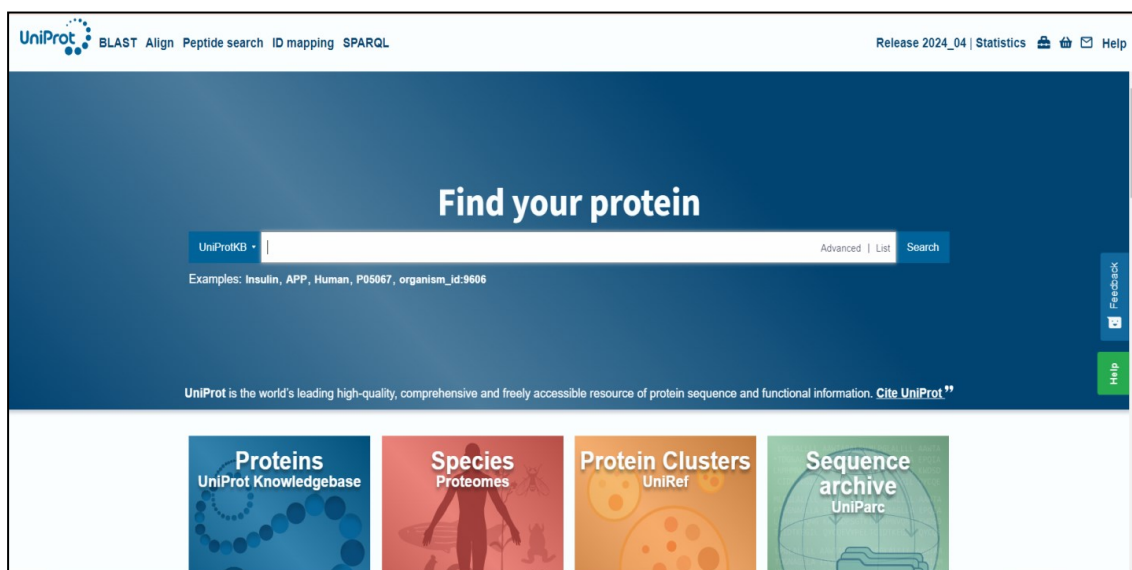
**Ig alpha chain C region:**

The Ig alpha chain C region, part of the heavy chain of immunoglobulins, plays a critical role in antibody structure and function, particularly in determining effector functions such as opsonization, complement activation, and antibody-dependent cellular cytotoxicity. It contributes to the stability and secretion of antibodies from B cells and is essential for maintaining the integrity of the immunoglobulin molecule during immune responses. Additionally, the C region is involved in B cell activation and class switching, allowing the immune system to produce different antibody types in response to various pathogens. Abnormalities in the C region can lead to autoimmune diseases, and its study is vital for developing monoclonal antibody therapies for treating cancers and immune disorders.

**METHODOLOGY:**

1. Go to the UniProt database homepage and type "Ig alpha chain C region" 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 Ig alpha chain C region, 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:



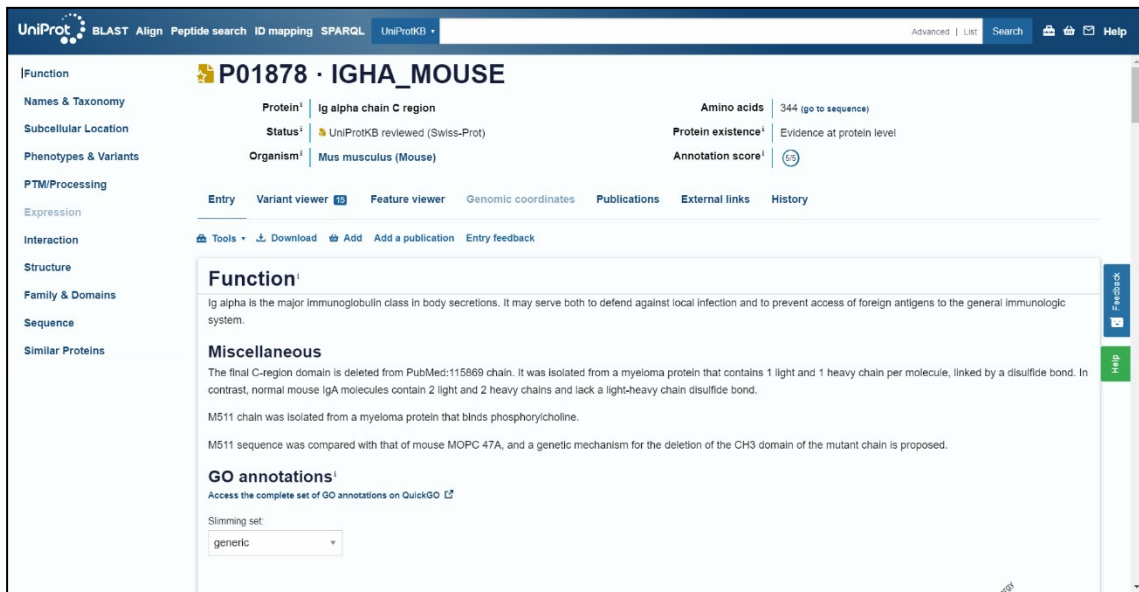
**Fig 1: Homepage of UniProt Database**

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.

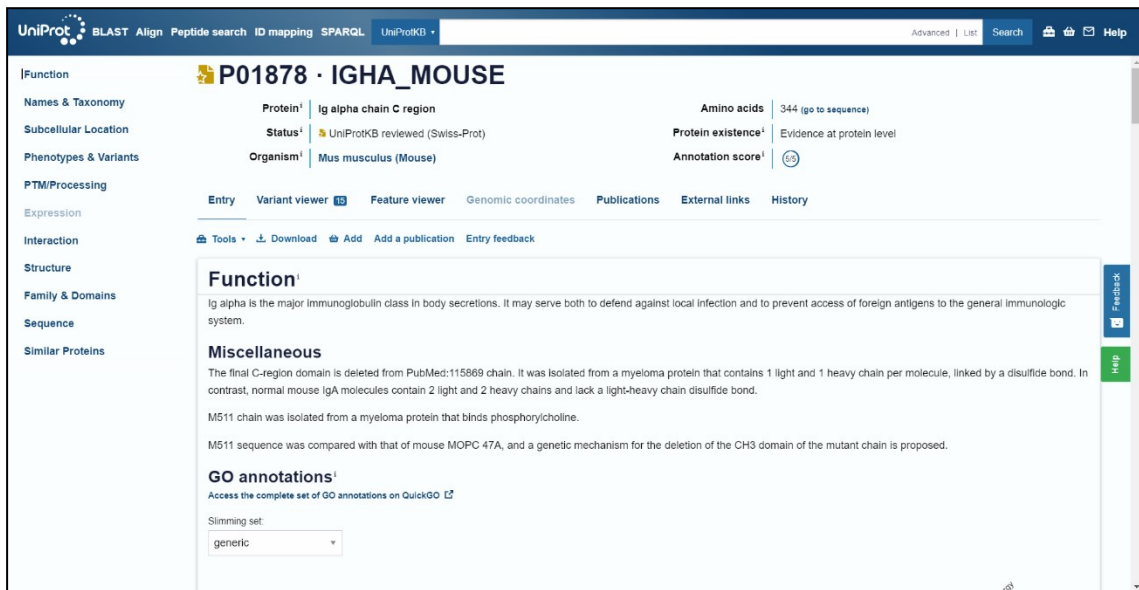
The screenshot shows the search results for "Ig alpha chain C region". The search bar contains "Ig alpha chain C region" and the results are displayed in a table. The table has columns for Entry, Entry Name, Protein Names, Gene Names, Organism, and Length. The results are filtered to show 43,866 results.

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P01878	IGHA_MOUSE	Ig alpha chain C region		Mus musculus (Mouse)	344 AA
P01877	IGHA2_HUMAN	Immunoglobulin heavy constant alpha 2[...]	IGHA2	Homo sapiens (Human)	391 AA
P01876	IGHA1_HUMAN	Immunoglobulin heavy constant alpha 1[...]	IGHA1	Homo sapiens (Human)	398 AA
P01880	IGHD_HUMAN	Immunoglobulin heavy constant delta[...]	IGHD	Homo sapiens (Human)	430 AA
P01859	IGHG2_HUMAN	Immunoglobulin heavy constant gamma 2[...]	IGHG2	Homo sapiens (Human)	395 AA
P01857	IGHG1_HUMAN	Immunoglobulin heavy constant gamma 1[...]	IGHG1	Homo sapiens (Human)	399 AA
P01854	IGHE_HUMAN	Immunoglobulin heavy constant epsilon[...]	IGHE	Homo sapiens (Human)	546 AA
P30517	1C01_SAGOE	Saee class I histocompatibility antigen, C alpha chain[...]		Saguinus oedipus (Cotton-top tamarin)	365 AA
P30686	1C01_PANTR	Patr class I histocompatibility antigen, C alpha chain[...]		Pan troglodytes (Chimpanzee)	366 AA
P30383	1C01_GORGO	Class I histocompatibility antigen, Gogo-C*0101/C*0102 alpha chain		Gorilla gorilla gorilla (Western lowland gorilla)	365 AA
P01768	HV330_HUMAN	Immunoglobulin heavy variable 3-30[...]	IGHV3-30	Homo sapiens (Human)	117 AA
P30386	1C03_GORGO	Class I histocompatibility antigen, Gogo-C*0202 alpha chain		Gorilla gorilla gorilla (Western lowland gorilla)	366 AA
P30385	1C02_GORGO	Class I histocompatibility antigen, Gogo-C*0201 alpha chain		Gorilla gorilla gorilla (Western lowland gorilla)	366 AA
P30387	1C04_GORGO	Class I histocompatibility antigen, Gogo-C*0203 alpha		Gorilla gorilla gorilla (Western lowland gorilla)	366 AA

**Fig 2: Ig alpha chain C region reviewed (SwissProt) search (1,398 results) and 42,468 hits are displayed in the search results.**



**Fig 3: The first result on a search for "Ig alpha chain C region (UniProt ID: P01878)" is protein with 693 amino acids.**



**Fig 4.1: P01878 protein present in *Mus musculus* searched serve both to defend against local infection and to prevent access of foreign antigens to the general immunologic system.**

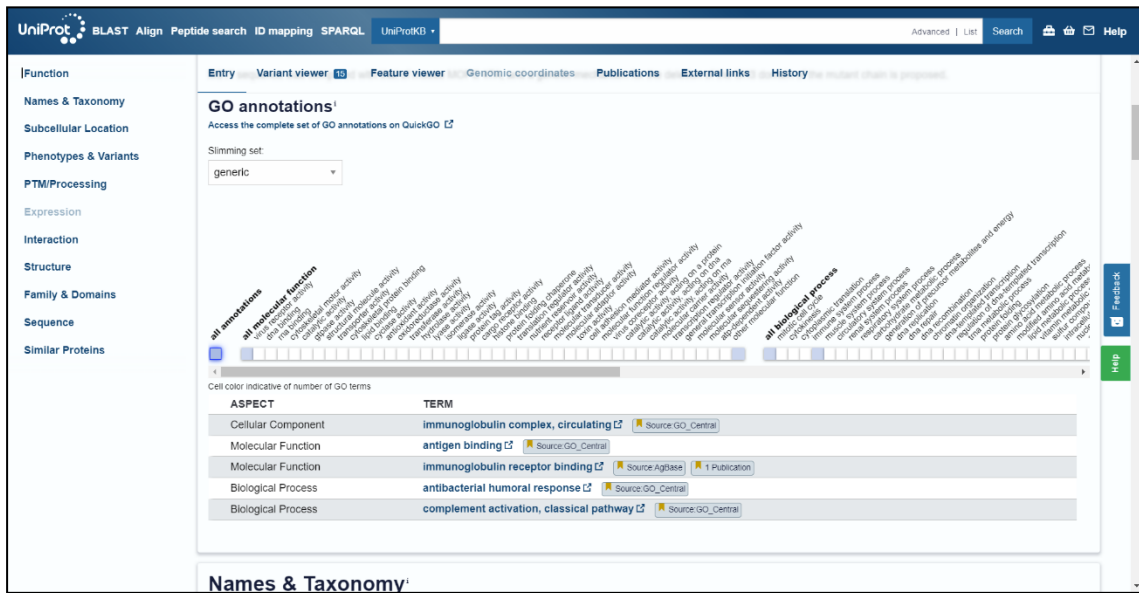


Fig 4.2: Number of Annotations and all molecular functions of site

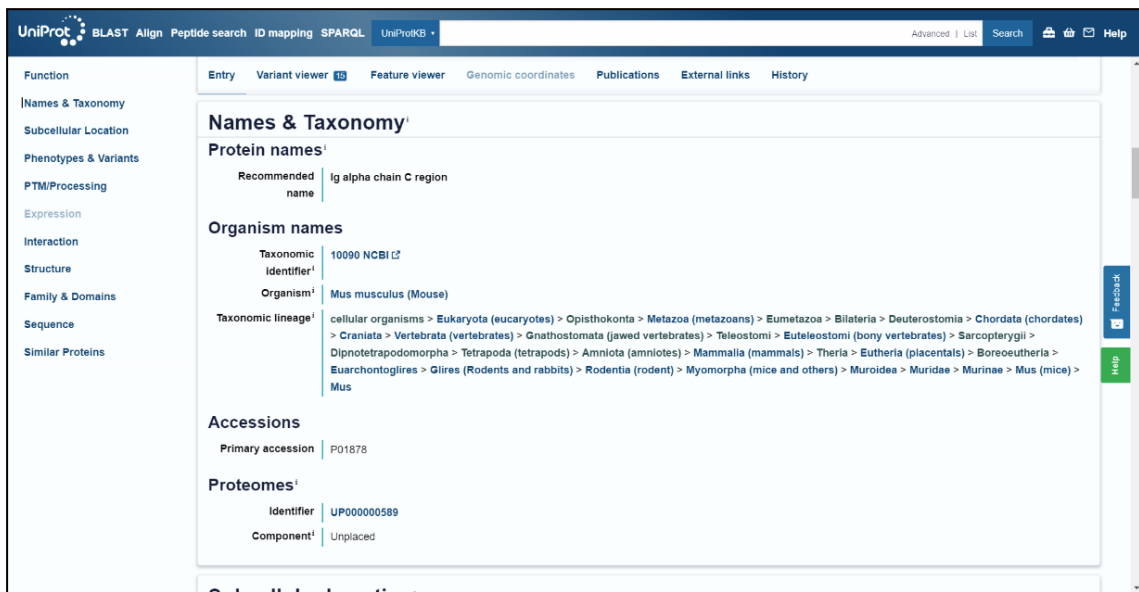
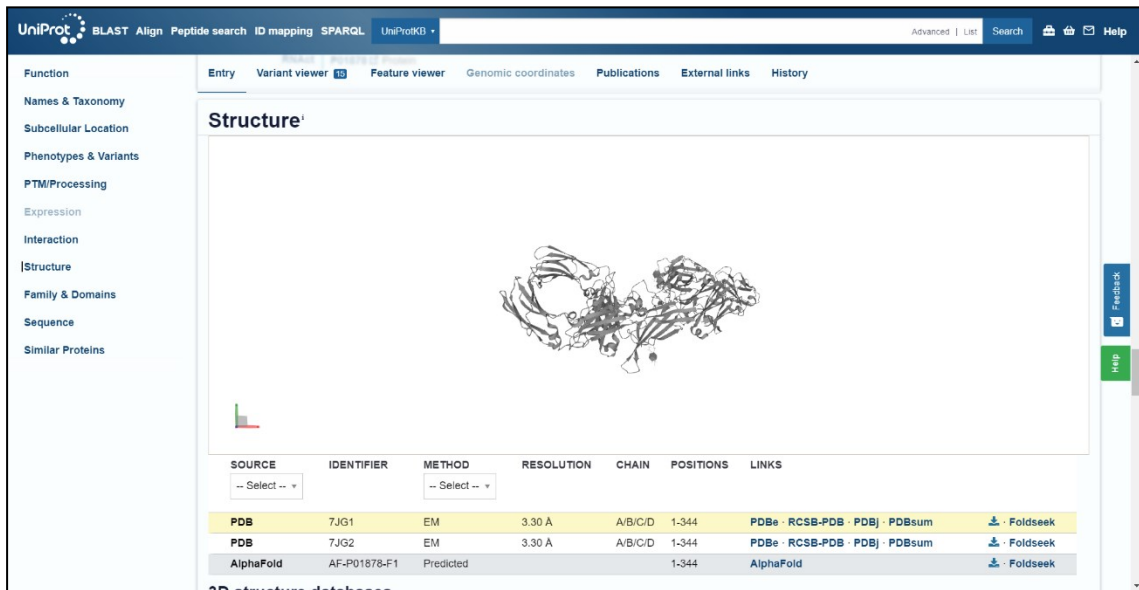
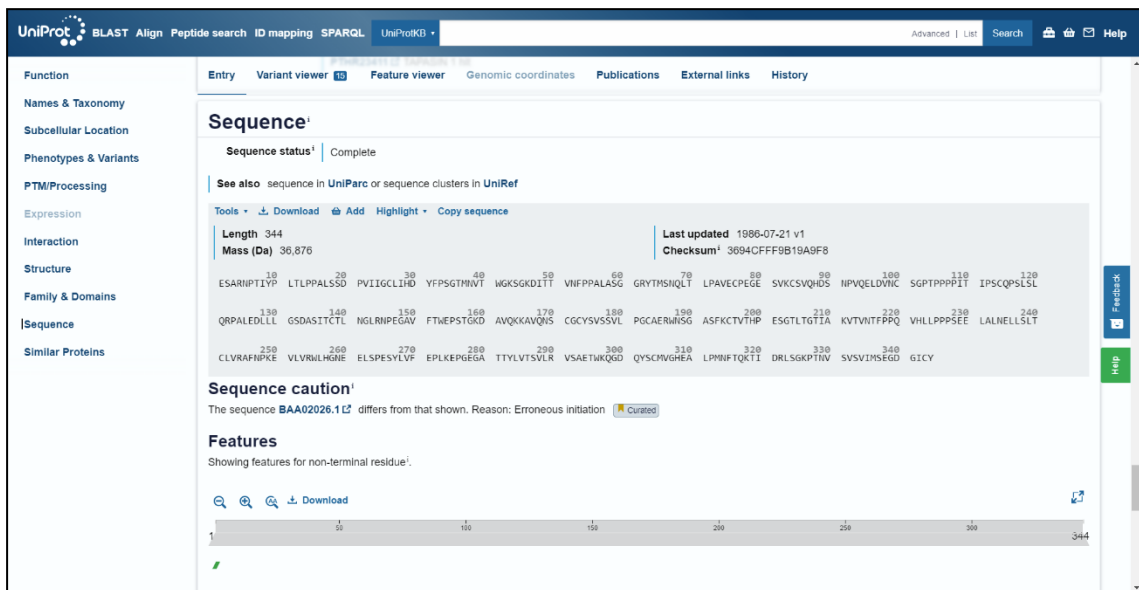


Fig 5: Name and Taxonomy of Ig alpha chain C region



**Fig 6: Structure of Ig alpha chain C region**



**Fig 7: Sequence of Ig alpha chain C region**

**RESULTS:**

The first entry for Ig alpha chain C region is a *Mus musculus* creature with 334 amino acids. Immunoglobulins, also known as antibodies, are specialized glycoproteins produced by B lymphocytes. Ig alpha is the major immunoglobulin class in body secretions. It may serve both to defend against local infection and to prevent access of foreign antigens to the general immunologic system.

**CONCLUSION:**

The UniProt, Swiss-Prot and TrEMBL databases were explored for the query Ig alpha chain C region (Accession ID: P01878) and related information was searched.

## **REFERENCES:**

1. UniProt. (n.d.). <https://www.uniprot.org/>
  2. Godwin, L., Sinawe, H., & Crane, J. S. (2022, September 24). Biochemistry, immunoglobulin e. StatPearls - NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK541058/>
  3. Sutton, B. J., Davies, A. M., Bax, H. J., & Karagiannis, S. N. (2019). IGE antibodies: From structure to function and clinical translation. *Antibodies*, 8(1), 19. <https://doi.org/10.3390/antib8010019>
  4. Miles, E. (2013). Adverse immune reactions to foods. In Elsevier eBooks (pp. 573–613). <https://doi.org/10.1533/9780857095749.4.573>
-

**WEBLEM: 1(B)**  
**Protein Data Bank (PDB) Database**  
**(URL: <https://www.rcsb.org/pdb/>)**

**AIM:**

To study and explore the protein structure for the query Dimeric Immunoglobulin A (dIgA) (PDB ID: 7JG1) using the Protein Data Bank (PDB) Database.

**INTRODUCTION:**

The Protein Data Bank (PDB) is a comprehensive database that houses three-dimensional structural data of biological macromolecules, including proteins and nucleic acids. Established in 1971, it is managed by the Worldwide Protein Data Bank (wwPDB), an international consortium responsible for overseeing the deposition, validation, curation, and open-access dissemination of 3D structural data.

The PDB is a vital resource for structural biology, particularly in fields like structural genomics, enabling scientists to study the 3D architecture of biological macromolecules. The archive contains atomic coordinates and other relevant information about proteins and key biological molecules, with the primary data being coordinate files that describe the atoms in each molecule and their spatial positions.

Noteworthy features of the PDB include its historical role as the first open-access digital platform for sharing protein structures, its importance in computational biology for applications such as structure-based drug design, and its continuous growth, reflecting the ongoing research in laboratories worldwide.

The PDB file format, a text-based format used to describe molecular structures, includes data on atomic coordinates, secondary structure assignments, and atomic connectivity. While the PDB format is a legacy system, the database now stores biological macromolecule data in the updated mmCIF format.

**Dimeric Immunoglobulin A (dIgA):**

Dimeric Immunoglobulin A (dIgA) is a key antibody involved in mucosal immunity, composed of two IgA monomers linked by a J chain. This structure enables dIgA to function effectively in mucosal tissues, where it is predominantly found in secretions like saliva, tears, and intestinal fluids. dIgA plays a crucial role in protecting mucosal surfaces by neutralizing pathogens and preventing their attachment to epithelial cells, thus inhibiting infection. It also supports immune responses by facilitating the clearance of antigens. Importantly, dIgA is transferred from mother to infant through breast milk, providing passive immunity in early life. Deficiencies in dIgA can lead to heightened susceptibility to infections and are linked to certain immunodeficiency disorders, making its study significant for developing vaccines and therapies.

**METHODOLOGY:**

1. Open the homepage of the Protein Data Bank (PDB) Database.
2. Enter the query 'Dimeric Immunoglobulin A' 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.

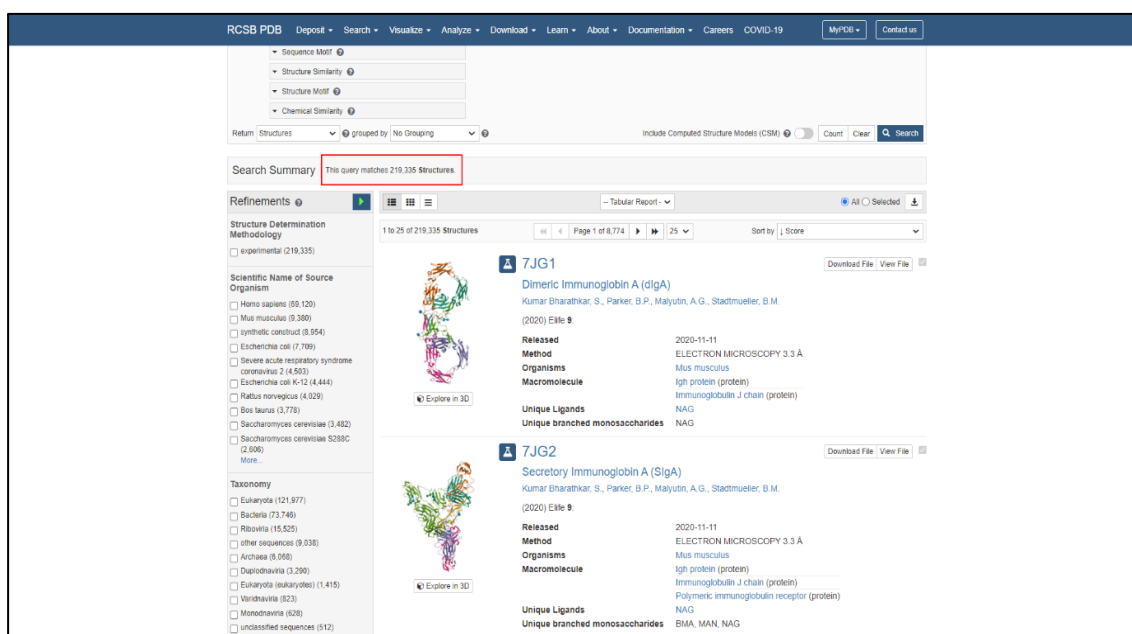


- Select a particular entry of interest [‘7JG1: Dimeric Immunoglobulin A (dIgA)] for further study in terms of its Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome, and Versions.
- To display and download the 3D structure of the protein, click on the ‘Display and Download’ option, and select the desired format.

## OBSERVATIONS:



**Fig 2: Homepage of the Protein Data Bank (PDB) Database**



**Fig 3: Number of hits obtained for Basic Search for the query**



RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MPOB Contact us

Return Structures grouped by No Grouping Include Computed Structure Models (CSM) Count Clear Search

Search Summary This query matches 219,335 Structures

Refinements

- Structure Determination Methodology
  - experimental (219,335)
- Scientific Name of Source Organism
  - Homo sapiens (98,120)
  - Mus musculus (9,380)**
  - synthetic construct (8,954)
  - Escherichia coli (2,709)
  - Severe acute respiratory syndrome coronavirus 2 (4,503)
  - Escherichia coli K-12 (4,444)
  - Rattus norvegicus (4,029)
  - Bos taurus (2,778)
  - Saccharomyces cerevisiae (2,482)
  - Staphylococcus cerevisiae 5298C (2,696)
  - More...
- Taxonomy
  - Eukaryota (121,977)
  - Bacteria (73,746)
  - Riboviria (15,525)
  - other sequences (9,038)
  - Archaea (8,366)
  - Diplomonada (3,390)
  - Eukaryota (eukaryotes) (1,415)
  - Virochordata (823)
  - Monodnaviria (628)
  - unclassified sequences (512)
  - More...

1 to 25 of 219,335 Structures Page 1 of 8,774 25 Sort by Score

**7JG1**  
Dimeric Immunoglobulin A (dIgA)  
Kumar Bharathkumar, S., Parker, B.P., Malayil, A.G., Stadtmueller, B.M.  
(2020) ELife 9  
Released: 2020-11-11  
Method: ELECTRON MICROSCOPY 3.3 A  
Organisms: Mus musculus  
Macromolecule: Igh protein (protein), Immunoglobulin J chain (protein)  
Unique Ligands: NAG  
Unique branched monosaccharides: NAG

**7JG2**  
Secretory Immunoglobulin A (SigA)  
Kumar Bharathkumar, S., Parker, B.P., Malayil, A.G., Stadtmueller, B.M.  
(2020) ELife 9  
Released: 2020-11-11  
Method: ELECTRON MICROSCOPY 3.3 A  
Organisms: Mus musculus  
Macromolecule: Igh protein (protein), Immunoglobulin J chain (protein), Polymeric immunoglobulin receptor (protein)  
Unique Ligands: NAG  
Unique branched monosaccharides: BMA, MAN, NAG

**Fig 4: List of Refinements (Filters) applied**

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MPOB Contact us

Return Structures grouped by No Grouping Include Computed Structure Models (CSM) Count Clear Search

Search Summary This query matches 9,380 Structures

Refinements

- Structure Determination Methodology
  - experimental (9,380)
- Scientific Name of Source Organism
  - Mus musculus (9,380)
  - Homo sapiens (1,832)
  - synthetic construct (208)
  - Rattus norvegicus (171)
  - Gallus gallus (116)
  - Severe acute respiratory syndrome coronavirus 2 (99)
  - Escherichia coli (87)
  - Bos taurus (85)
  - Human immunodeficiency virus 1 (77)
  - Streptomyces lividans (65)
  - More...
- Taxonomy
  - Eukaryota (9,366)
  - Riboviria (572)
  - Bacteria (348)
  - other sequences (312)
  - Eukaryota (eukaryotes) (149)
  - Diplomonada (55)
  - Riboviria (RNA viruses and viroids) (54)
  - Monodnaviria (34)
  - Virochordata (20)
  - Archaea (11)
  - More...

1 to 25 of 9,380 Structures Page 1 of 376 25 Sort by Score

**7JG1**  
Dimeric Immunoglobulin A (dIgA)  
Kumar Bharathkumar, S., Parker, B.P., Malayil, A.G., Stadtmueller, B.M.  
(2020) ELife 9  
Released: 2020-11-11  
Method: ELECTRON MICROSCOPY 3.3 A  
Organisms: Mus musculus  
Macromolecule: Igh protein (protein), Immunoglobulin J chain (protein)  
Unique Ligands: NAG  
Unique branched monosaccharides: NAG

**7JG2**  
Secretory Immunoglobulin A (SigA)  
Kumar Bharathkumar, S., Parker, B.P., Malayil, A.G., Stadtmueller, B.M.  
(2020) ELife 9  
Released: 2020-11-11  
Method: ELECTRON MICROSCOPY 3.3 A  
Organisms: Mus musculus  
Macromolecule: Igh protein (protein), Immunoglobulin J chain (protein), Polymeric immunoglobulin receptor (protein)  
Unique Ligands: NAG  
Unique branched monosaccharides: BMA, MAN, NAG

**Fig 5: Results obtained after applying refinements (filters) and select the query**

The screenshot shows the RCSB PDB website interface. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, About, Documentation, Careers, and COVID-19. Below this is a search bar and a header with the PDB logo and statistics. The main content area features a navigation menu where 'Structure Summary' is highlighted. The central focus is the entry for 7JG1, 'Dimeric Immunoglobulin A (dIgA)'. A 3D ribbon model of the protein is displayed on the left. To the right of the model, the entry ID '7JG1' is highlighted in a red box. Below the model, there are sections for 'Explore in 3D', 'Global Symmetry', 'Local Symmetry', and 'Biological Assembly Evidence'. On the right side, there is a detailed metadata section including 'Classification: IMMUNE SYSTEM', 'Organism(s): Mus musculus', 'Expression System: Homo sapiens', and 'Mutation(s): no'. An 'Experimental Data Snapshot' section provides details on the method (ELECTRON MICROSCOPY), resolution (3.30 Å), aggregation state (PARTICLE), and reconstruction method (SINGLE PARTICLE). A 'wwPDB Validation' chart is also visible, showing metrics like Clashscore, Ramachandran outliers, and Sidechain outliers. At the bottom, there is a 'Literature' section with a 'Download Primary Citation' button.

**Fig 6: Entry opened that displays the Structure Summary**

The screenshot shows the RCSB PDB website interface with the 'Structure' tab selected. The entry for 7JG1, 'Dimeric Immunoglobulin A (dIgA)', is displayed. A 3D ribbon model of the protein is shown at the bottom. The 'Structure' panel on the right is expanded, showing various interactive options: 'Type: Assembly', 'Asm Id: 1: Author Defined Asse...', 'Dynamic Bonds: Off', 'Nothing Focused', 'Measurements', 'Structure Motif Search', 'Components: 7JG1', 'Preset: + Add', 'Polymer: Carbon', 'Carbohydrate: 2 Heps', 'Density', 'Quality Assessment', 'Assembly Symmetry', 'Export Models', and 'Export Animation'. The 'Sequence of 7JG1' is also visible at the top of the structure view.

**Fig 7: 3D View of the structure**

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK 225,681 Structures from the PDB 1,868,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Help

Advanced Search Browse Annotations

PDB-101 PDB EMDataResource NAKB PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

7JG1 Display Files Download Files Data API

Dimeric Immunoglobulin A (dIgA)

External Resource: Annotation

- Domain Annotation: ECOD Classification
- IMGT Antibody Annotation
- Protein Family Annotation
- Gene Ontology: Gene Product Annotation
- InterPro: Protein Family Classification

Domain Annotation: ECOD Classification ECOD Database Homepage

Chain	Family Name	Domain Identifier	Architecture	Possible Homology	Homology	Topology	Family	Protein Source (Version)
A	C1-set_3	e7g1A2	A: beta sandwiches	X: Immunoglobulin-like beta-sandwich	H: Immunoglobulin-related	T: Immunoglobulin/Fibronectin type III/E set domains/PagD-like	F: C1-set_3	ECOD (1.6)
A	C1-set	e7g1A1	A: beta sandwiches	X: Immunoglobulin-like beta-sandwich	H: Immunoglobulin-related	T: Immunoglobulin/Fibronectin type III/E set domains/PagD-like	F: C1-set	ECOD (1.6)
B	C1-set_3	e7g1B2	A: beta sandwiches	X: Immunoglobulin-like beta-	H: Immunoglobulin-related	T: Immunoglobulin/Fibronectin type	F: C1-set_3	ECOD (1.6)

Fig 8: View of the Annotations Section

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK 225,681 Structures from the PDB 1,868,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Help

Advanced Search Browse Annotations

PDB-101 PDB EMDataResource NAKB PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

7JG1 Display Files Download Files Data API

Dimeric Immunoglobulin A (dIgA)

ELECTRON MICROSCOPY

Refinement

Key	Refinement Restraint Deviation
f_dihedral_angle_d	11.453
f_angle_d	2.02
f_chi1a1_restr	0.087
f_bond_d	0.018
f_plane_restr	0.017

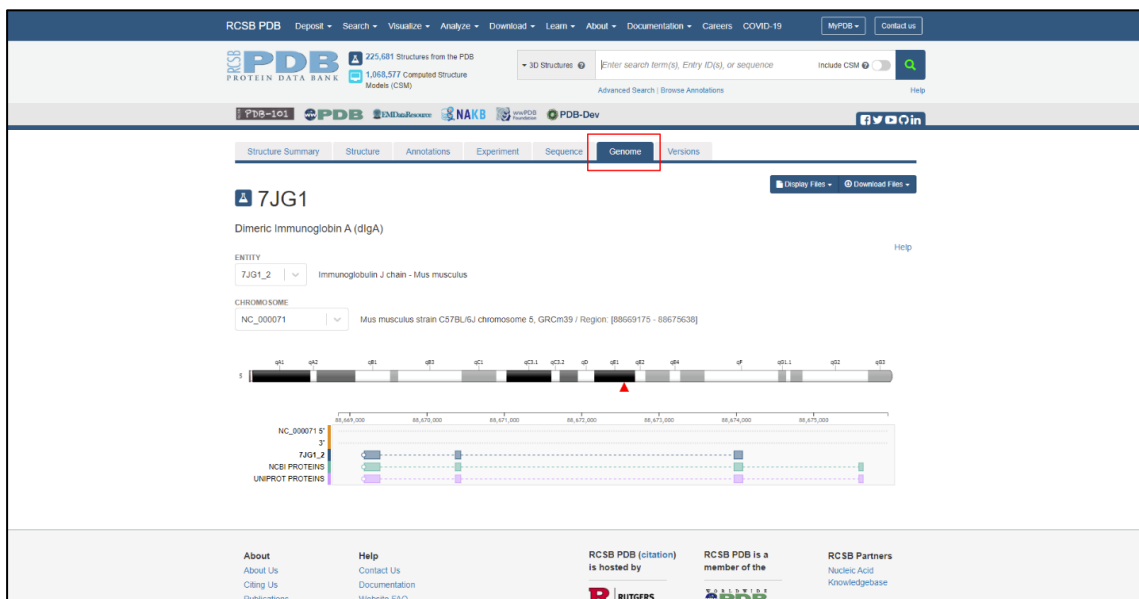
Sample	Data Acquisition
Secretory Immunoglobulin A	Detector Type: GATAN K3 BIOQUANTUM (6k x 4k)
	Electron Dose (electrons/Å <sup>2</sup> ): 60
	Imaging Experiment: 1
	Date of Experiment:
	Temperature (Kelvin):
	Microscope Model: FEI TITAN KRIOIS

Specimen Preparation	
Sample Aggregation State	PARTICLE
Vitification Instrument	FEI VITROBOT MARK IV
Cryogen Name	ETHANE
Sample Vitification Details	Wait time - 0s Drain time - 0s Blot time - 6s Blot Force - 5

Fig 9: View of the Experiment Section



**Fig 10: View of the Sequence Section**



**Fig 11: View of the Genome Section**

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK 225,681 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Help

Advanced Search | Browse Annotations

PDB-101 PDB EMDResource NAKB wwPDB PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome **Versions**

7JG1 Dimeric Immunoglobulin A (dIgA)

Changes made to a PDB entry after its initial release are considered to be either "major" or "minor". The latest minor version of each major version is available as a file download. More information about the PDB versioning is available.

Version Number	Version Date	Version Type/Reason	Version Change	Revised CIF Category
1.0	2020-11-11	Initial release		

Download

About: About Us, Citing Us, Publications, Team, Careers, Usage & Privacy

Help: Contact Us, Documentation, Website FAQ, Glossary, Service Status

RCSB PDB is hosted by: RUTGERS, UC San Diego SDSC, UCSF

RCSB PDB is a member of the: wwPDB

RCSB Partners: Nucleic Acid Knowledgebase

wwPDB Partners: RCSB PDB, PDBE, PDBJ, EMDB, EMDB

RCSB PDB Core Operations are funded by the U.S. National Science Foundation (DBI-2021666), the U.S. Department of Energy (DE-SC0019740), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM131196.

Fig 12: View of the Version Section

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK 225,681 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Help

Advanced Search | Browse Annotations

PDB-101 PDB EMDResource NAKB wwPDB PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

7JG1 Dimeric Immunoglobulin A (dIgA)

PDB DOI: https://doi.org/10.2210/pdb7JG1/pdb

Classification: IMMUNE SYSTEM

Organism(s): Mus musculus

Expression System: Homo sapiens

Mutation(s): No

Deposited: 2020-07-18 Released: 2020-11-11

Deposition Author(s): Kumar Bharathkar, S., Parker, B.P., Malutin, A.G., Stadtmueller, M.

Funding Organization(s): National Institutes of Health/National Institute of General Medical Sciences, National Institutes of Health/National Institute of Allergy and Infectious Diseases (NIH)

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY

Resolution: 3.30 Å

Aggregation State: PARTICLE

Reconstruction Method: SINGLE PARTICLE

Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (NAG)

Global Symmetry: Asymmetric - C1

Global Stoichiometry: Hetero 5-mer - A4B1

Display Files Download Files Data API

- FASTA Sequence
- mmCIF Format
- mmCIF Format (Header)
- PDB Format
- PDB Format (Header)
- PDBML/XML Format (gz)
- EM Map EMD-22309 (map - gz)
- Validation Full PDF
- Validation (XML - gz)
- Validation (CIF - gz)
- Biological Assembly 1 (CIF - gz)
- Biological Assembly 1 (PDB - gz)

wwPDB Validation

Metric

Clashscore

Ramachandran outliers

Sidechain outliers

Fig 12: Display And Download Options

```

HEADER      IMMUNE SYSTEM                               18-JUL-20   7JG1
TITLE      DIMERIC IMMUNOGLOBULIN A (DIGA)
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: IGH PROTEIN;
COMPND     3 CHAIN: A, B, C, D;
COMPND     4 ENGINEERED: YES;
COMPND     5 OTHER_DETAILS: GENES ENCODING THE MUS MUSCULUS TGA HC CONSTANT REGION
COMPND     6 AND THE LAMBDA LC CONSTANT REGION WERE FUSED WITH HC AND LC VARIABLE
COMPND     7 REGION SEQUENCES TO CREATE COMPLETE HC AND LC SEQUENCES. THE HC AND
COMPND     8 LC VARIABLE REGION IS NOT MODELED IN THIS STRUCTURE. SINCE AUTHORS
COMPND     9 HAVE CHOSEN NOT TO PROVIDE COMPLETE SAMPLE SEQUENCE OF THE HC AND LC
COMPND     10 VARIABLE REGION, THE UNKS WERE NOT INCLUDED IN THE SEQUENCE.;
COMPND     11 MOL_ID: 2;
COMPND     12 MOLECULE: IMMUNOGLOBULIN J CHAIN;
COMPND     13 CHAIN: J;
COMPND     14 ENGINEERED: YES
SOURCE     MOL_ID: 1;
SOURCE     2 ORGANISM_SCIENTIFIC: MUS MUSCULUS;
SOURCE     3 ORGANISM_COMMON: MOUSE;
SOURCE     4 ORGANISM_TAXID: 10090;
SOURCE     5 GENE: IGH;
SOURCE     6 EXPRESSION_SYSTEM: HOMO SAPIENS;
SOURCE     7 EXPRESSION_SYSTEM_TAXID: 9606;
SOURCE     8 EXPRESSION_SYSTEM_CELL_LINE: EXP1293;
SOURCE     9 MOL_ID: 2;
SOURCE     10 ORGANISM_SCIENTIFIC: MUS MUSCULUS;
SOURCE     11 ORGANISM_COMMON: MOUSE;
SOURCE     12 ORGANISM_TAXID: 10090;
SOURCE     13 GENE: JCHAIN, IGJ;
SOURCE     14 EXPRESSION_SYSTEM: HOMO SAPIENS;
SOURCE     15 EXPRESSION_SYSTEM_TAXID: 9606;
SOURCE     16 EXPRESSION_SYSTEM_CELL_LINE: EXP1293
KEYWDS     IMMUNE SYSTEM
EXPDTA     ELECTRON MICROSCOPY
AUTHOR     S.KUMAR BHARATHAR,B.P.PARKER,A.G.MALYUTIN,B.M.STADTMUELLER
REVDAT    1   11-NOV-20  7JG1      B
JRNL      AUTH 2 K.E.HUEY-TUBMAN,E.TAKKORSHID,B.STADTMUELLER
JRNL      TITL THE STRUCTURES OF SECRETORY AND DIMERIC IMMUNOGLOBULIN A.
JRNL      REF  ELIFE                               V. 9
JRNL      REFN  ESN 2050-084X
JRNL      PMID 3310780
JRNL      DOI  10.7554/ELIFE.56998
REMARK    2   RESOLUTION: 3.30 ANGSTROMS.
REMARK    3
REMARK    3   REFINEMENT.
REMARK    3   SOFTWARE PACKAGES : SERIALLEN, CRYOSPARC, CRYOSPARC,
REMARK    3   CRYOSPARC, CRYOSPARC.
REMARK    3   RECONSTRUCTION SCHEMA : NULL
REMARK    3

```

**Fig 13: 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 ‘Dimeric Immunoglobulin A’ with the PDB ID: 7JG1. A total of 219,335 protein structure entries were initially obtained through a basic search. The results have been categorized into different sections, including Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome 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). 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](https://doi.org/10.1016/s0022-2836(77)80200-3)
3. Herr, A. B., Ballister, E. R., & Bjorkman, P. J. (2003). Insights into IgA-mediated immune responses from the crystal structures of human FcαRI and its complex with IgA1-Fc. *Nature*, 423(6940), 614-620. <https://doi.org/10.1038/nature01685>

**WEBLEM: 2**

**Structural Antibody Database (SAbDab)**

**(URL: <https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab>)**

**AIM:**

To study the antibody structure information using SAbDab Database.

**INTRODUCTION:**

Antibodies form the foundations of the vertebrate immune response. These proteins form complexes with potentially pathogenic molecules called antigens and inhibit their function or recruit other components of the immunological machinery to destroy them. In addition to the biological importance of antibodies, their ability to be raised against an almost limitless number of molecules has made them useful laboratory tools and increasingly useful as therapeutic agents in humans. This biopharmaceutical application has motivated the desire to understand how binding, stability and immunogenic properties of the antibody are determined and how they can be modified.

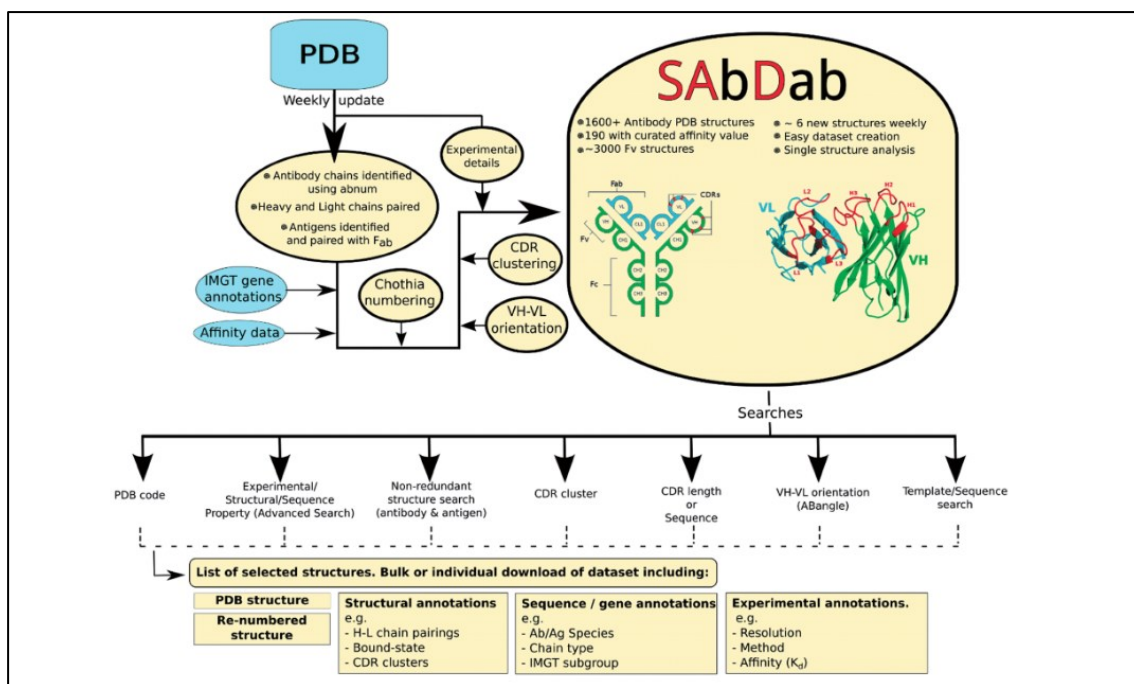
Computational analyses and tools are increasingly being employed to aid the antibody engineering process. Many of these tools now use only the antibody data, as opposed to general protein data, because this has been shown to increase performance. The publicly available structural data for most types of proteins are too sparse to merit protein-specific prediction methods. However, since the first antibody structure was deposited in 1976, the number of antibody structures in the protein data bank (PDB) has grown, and it now represents approximately 1.75% of the total 91939 entries.

Several databases that handle antibody data currently exist (7–13). Of these, most are sequence-based or are antibody discovery tools. The most recent, DIGIT, provides sequence information for immunoglobulins and has the advantage over earlier sequence databases [Kabat, IMGT, Vbase2] of providing heavy and light chain sequence pairings. However, it does not incorporate structural data. Antigen DB and IEDB-3D do include structural data. However, both focus on collecting epitope data and do not include unbound antibody structures. In comparison, both IMGT and the Aysis portal provide the ability to inspect and download individual bound and unbound antibody structures. Neither allow for the generation of bespoke datasets nor for the download of an ensemble of curated structural data.

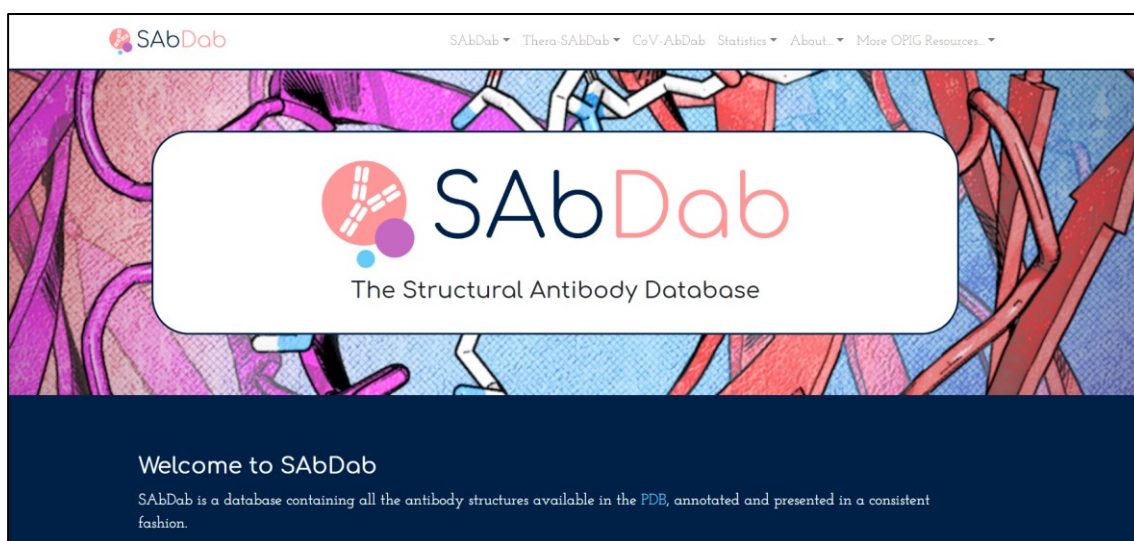
To address this problem, we have developed a Structural Antibody Database (SAbDab), a database devoted to automatically collecting, curating, and presenting antibody structural data in a consistent manner for both bulk analysis and individual inspection. SAbDab updates on a weekly basis and provides users with a range of methods to select sets of structures. For example, users can select by species, experimental details (e.g. method, resolution, and r-factor), similarity to a given antibody sequence, amino acid composition at certain positions and antibody–antigen affinity. Entries can also be selected using structural annotations including, for example, the canonical form of the complementarity determining regions (CDR), orientation between the antibody variable domains and the presence of constant domains in the structure. Structures can be inspected individually or downloaded en masse either as the original file from the PDB or as a structure that has been annotated using the Chothia numbering scheme. In all cases, a tab-separated file detailing heavy and light chain pairing, antibody–antigen pairing and all other annotations is generated.



Structural antibody database is an online resource containing all the publicly available antibody structures annotated and presented in a consistent fashion. The data are annotated with several properties including experimental information, gene details, correct heavy and light chain pairings, antigen details and, where available, antibody-antigen binding affinity. The user can select structures, according to these attributes as well as structural properties such as complementarity determining region loop conformation and variable domain orientation. Individual structures, datasets and the complete database can be downloaded.



**Fig 1: SABDab's workflow**



**Fig 2: Homepage of SABDab Database**



SAbDab Thera-SAbDab CoV-AbDab Statistics About More OPIG Resources

## About Thera-SAbDab

**About Thera-SAbDab**

The Therapeutic Structural Antibody Database (Thera-SAbDab) is a database of immunotherapeutic variable domain sequences and their corresponding structural representatives in [SAbDab](#) (which harvests data from the [PDB](#)). It updates structural mappings alongside SAbDab on a weekly basis. It detects not only exact sequence matches to known structures, but also close sequence matches (divided into two categories: 95-98% seqID, or 99% seqID).

We update Thera-SAbDab whenever a new [WHO International Non-proprietary Name \(INN\) list](#) is released, adding all therapeutics with an accompanying variable domain sequence. We also update the clinical trial status of all actively-developed therapeutics according to the latest updates on [AdisInsight](#). We host this up-to-date list of therapeutic sequences with metadata on the [Thera-SAbDab search page](#).

**Fig 3: Homepage of Thera-SAbDab**

UNIVERSITY OF OXFORD OPIG More OPIG Resources

## CoV-AbDab

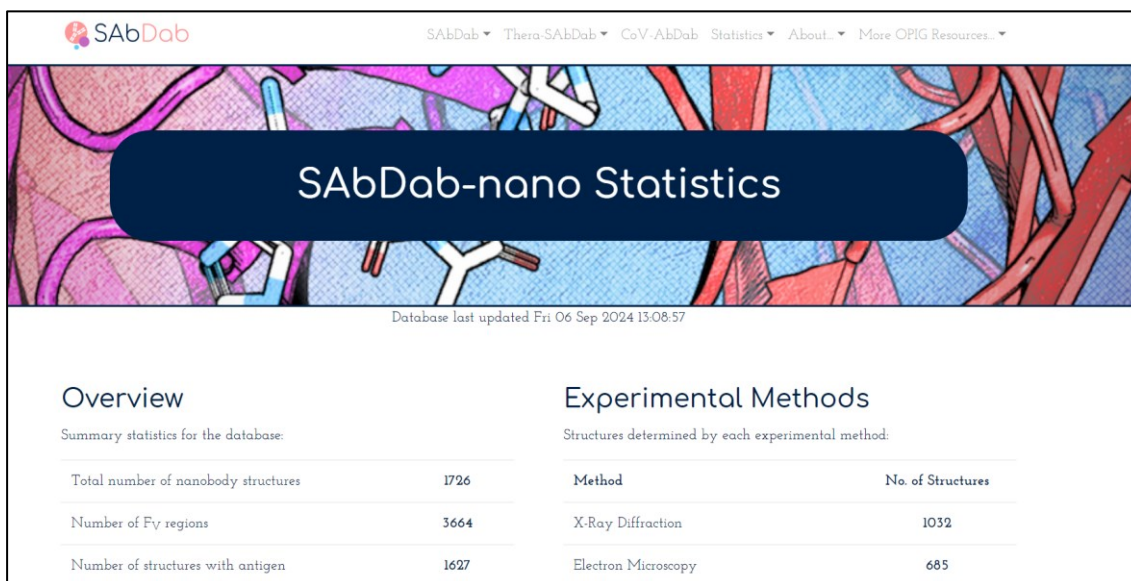
The Coronavirus Antibody Database

### Coronavirus-Binding Antibody Sequences & Structures

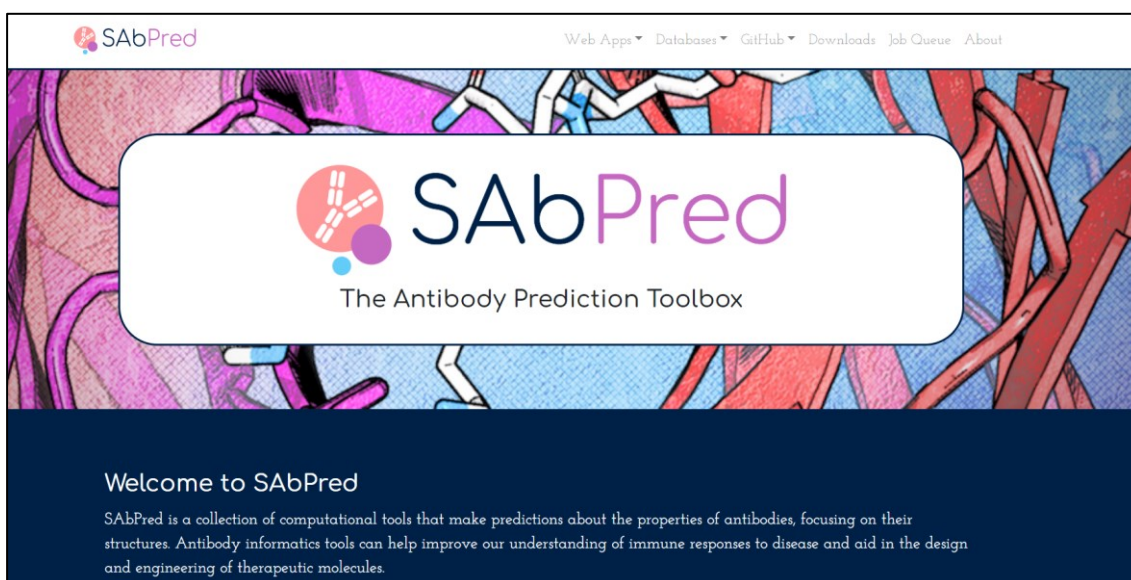
The [Oxford Protein Informatics Group](#) (Dept. of Statistics, University of Oxford) is collaborating in efforts to understand the immune response to SARS-CoV2 infection and vaccination. As part of our investigations, we are releasing and maintaining this public database to [document all published/patented antibodies and nonobodies able to bind to coronaviruses, including SARS-CoV2, SARS-CoV, and MERS-CoV.](#)

Explanations and a preliminary analysis of the database contents can be found in our [Applications Note](#) in Bioinformatics. Please consider citing it if you are making use of our database in your research. [BibTex Reference](#).

**Fig 4: Homepage of CoV-AbDab**



**Fig 5: SAbDab-nano Statistics**



**Fig 6: The antibody Prediction Toolkit**

## **REFERENCES:**

1. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., & Deane, C. M. (2014). SAbDab: the structural antibody database. *Nucleic acids research*, 42(Database issue), D1140–D1146. <https://doi.org/10.1093/nar/gkt1043>
2. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., & Deane, C. (2013). SAbDab: the structural antibody database. <https://www.semanticscholar.org/paper/SAbDab%3A-the-structural-antibody-database-Dunbar-Krawczyk/fefea2b9ed93a0c3163432c52a67cf34efa868f7>

**WEBLEM: 2(A)**

**The Structural Antibody Database (SAbDab)**

**(URL: <https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab>)**

**AIM:**

To study the Antibody structure for the query 'Bovine anti-HIV Fab ElsE6' (PDB ID: 8VBL) using the Structural Antibody Database (SAbDab).

**INTRODUCTION:**

**SAbDab Database**

The Structural Antibody Database (SAbDab) is a comprehensive online resource dedicated to the collection and curation of antibody structures. It provides researchers with access to all publicly available antibody structures, which are annotated and presented in a standardized format. This database is particularly valuable for those working in the fields of antibody structure prediction, docking, and therapeutic design.

**Key Features of SAbDab**

1. **Extensive Data Collection:** SAbDab includes a significant number of antibody structures, with around 7,184 variable domain structures recorded from 3,663 entries in the Protein Data Bank (PDB) as of August 2019.
2. **Detailed Annotations:** Each structure in the database is annotated with various properties, such as experimental details, gene information, heavy and light chain pairings, antigen details, and where available, antibody-antigen binding affinities. This comprehensive annotation allows users to filter and select structures based on specific criteria, including experimental methods and structural properties.
3. **User-Friendly Tools:** The database features several tools for users, such as:
  - a. **ABangle Tool:** This tool allows users to characterize the orientation between the antibody's variable domains (VH and VL) and visualize conformational changes.
  - b. **CDR Search and Clustering:** Users can select hyper-variable loops based on their length and type, facilitating the study of antibody variability.
  - c. **Template Search:** Users can submit antibody sequences to find structural templates suitable for homology modeling.
4. **Regular Updates:** SAbDab is updated weekly, ensuring that it reflects the latest entries from the PDB and includes new sequence data as it becomes available. This continuous updating process enhances the database's relevance for ongoing research.
5. **Accessibility:** The database is freely available for public use, encouraging collaboration and innovation in antibody research. Users can download individual structures or entire datasets for further analysis.

**Bovine anti-HIV Fab ElsE6**

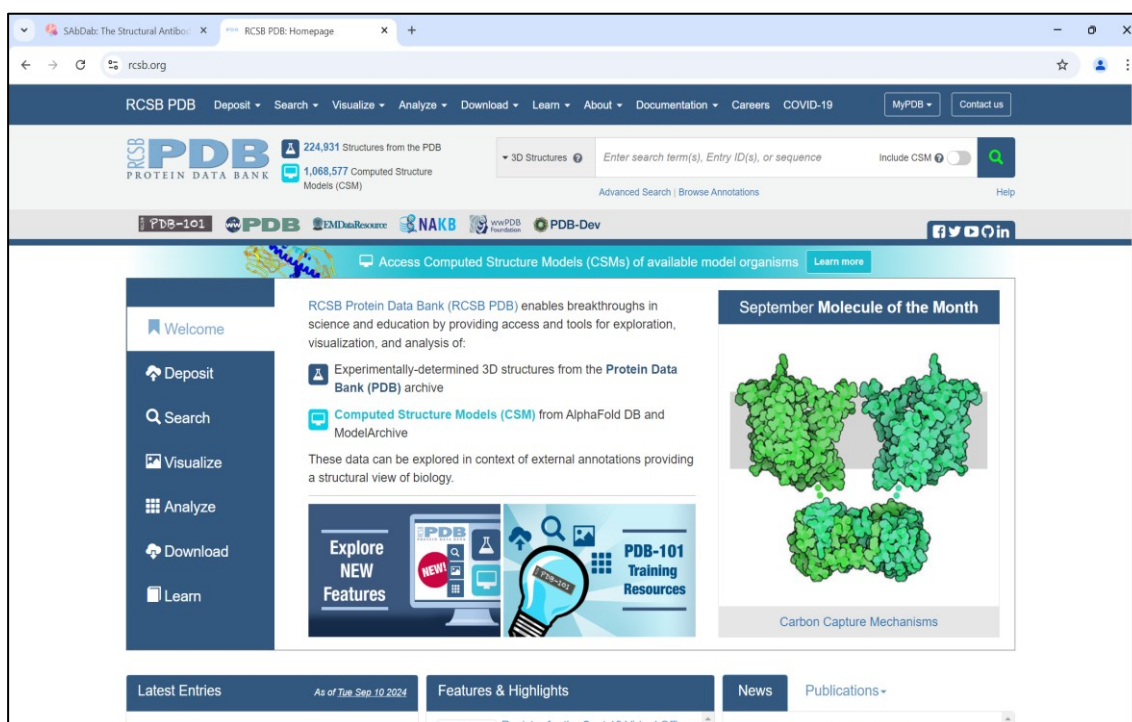
Bovine anti-HIV Fab ElsE6 is a specific antibody fragment derived from bovine sources, designed to target HIV-1 antigens, particularly the p24 protein. As a fragment antigen-binding (Fab) region, ElsE6 binds with high specificity to HIV p24, making it a key tool in the detection and quantification of the

virus. It is primarily used in immunoassays, such as enzyme immunoassays, to enhance the sensitivity of HIV detection, especially in early stages of infection before antibodies are detectable. Beyond diagnostics, ElsE6 also serves in HIV research, helping to study viral interactions with the immune system. While its primary use is in diagnostics, there is potential for therapeutic applications if modified for neutralizing HIV activity.

## **METHODOLOGY:**

1. Open the Protein Data Bank (PDB) website. (URL: <https://www.rcsb.org/>) to obtain the PDB ID of the structure (query).
2. Open the Structural Antibody Database (SAbDab). (URL: <https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab>) that contains structural information on antibodies and antibody-antigen complexes.
3. Select the 'Structure Search' option from the SAbDab portal.
4. Select the 'Search for a specific PDB entry' option to search for antibody structures by their PDB ID.
5. Enter the PDB ID of the query (PDB ID: 8JXR), obtained from the Protein Data Bank (PDB) and select the entry from the results obtained to view detailed information about the antibody structure.
6. Study the Structure Details, Structure Visualization, Fv regions. Structures can further be downloaded using the links provided in the 'Download' section.

## **OBSERVATIONS:**



**Fig 1: Homepage of the Protein Data Bank (PDB) database**



The screenshot shows the RCSB PDB website interface. At the top, there is a navigation bar with options like 'Deposit', 'Search', 'Visualize', 'Analyze', 'Download', 'Learn', 'About', 'Documentation', 'Careers', and 'COVID-19'. Below this, the PDB logo and statistics are displayed: '225,681 Structures from the PDB' and '1,668,577 Computed Structure Models (CSM)'. A search bar is present with the text 'Enter search term(s), Entry ID(s), or sequence'. The main content area is titled 'Structure Summary' and shows the entry '8VBL' with the title 'Structure of bovine anti-HIV Fab ElsE6'. A 3D ribbon diagram of the protein structure is shown on the left. To the right, there is a list of metadata including 'PDB DOI', 'Classification: IMMUNE SYSTEM', 'Organism(s): Bos taurus', 'Expression System: Homo sapiens', and 'Mutation(s): No'. Below this, there is an 'Experimental Data Snapshot' section with details like 'Method: X-RAY DIFFRACTION', 'Resolution: 2.35 Å', 'R-Value Free: 0.245', 'R-Value Work: 0.198', 'R-Value Observed: 0.200', and 'Starting Model: experimental'. A 'wwPDB Validation' section is also visible, showing a bar chart of various metrics like Rfree, Clashscore, Ramachandran outliers, Sidechain outliers, and RSRRZ outliers.

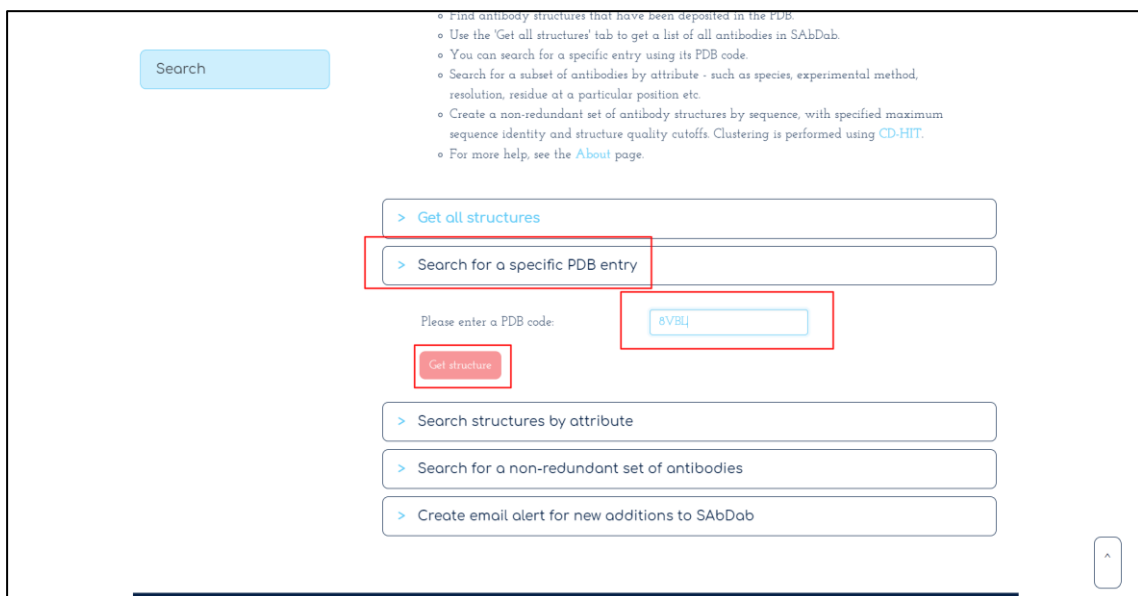
**Fig 2: Retrieving the query ‘Bovine anti-HIV Fab ElsE6’ (PDB ID: 8VBL) from the PDB database**

The screenshot shows the homepage of the Structural Antibody Database (SAbDab). The browser address bar shows the URL 'opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab'. The page has a navigation menu with 'SAbDab', 'Thera-SAbDab', 'CoV-AbDab', 'Statistics', 'About', and 'More OPIG Resources'. A large blue banner features the SAbDab logo and the text 'The Structural Antibody Database'. Below the banner, there is a 'Welcome to SAbDab' section with a paragraph: 'SAbDab is a database containing all the antibody structures available in the PDB, annotated and presented in a consistent fashion.' Another paragraph follows: 'Each structure is annotated with a number of properties including experimental details, antibody nomenclature (e.g. heavy-light pairings), curated affinity data and sequence annotations. You can use the database to inspect individual structures.' At the bottom, there is a cookie consent message: 'We use cookies to collect usage statistics for this website. By continuing to browse this site you agree to our use of cookies. For more details about cookies see our privacy policy. Continue'.

**Fig 3: Homepage of the Structural Antibody Database (SAbDab)**



**Fig 4: Selecting the ‘Structure Search’ option in the SABdab portal**



**Fig 5: Selecting the ‘Search for a specific PDB entry’ option and searching for the PDB code: ‘8JXR’**

SAbDab

SAbDab Thera-SAbDab CoV-AbDab Statistics About More OPIG Resources

## Search Structures

View results  
Downloads  
Search

### Search results

1 structure(s) fit your criteria. Click on the PDB code to view the structure.

PDB	Species	Method	Resolution	Chain Pairings	Antigens	Downloads
8vbl	BOS TAURUS	X-RAY DIFFRACTION	2.35 Å	Fv no.: 1 VH: H VL: L	None	<ul style="list-style-type: none"> <li>Structure (as PDB)</li> <li>Structure (Chothia)</li> <li>Structure (IMGT)</li> <li>Summary File</li> </ul>

Download results

**Fig 6: Search Results obtained for searching the PDB ID: 8JXR**

SAbDab

SAbDab Thera-SAbDab CoV-AbDab Statistics About More OPIG Resources

## Structure Viewer: 8vbl

Details  
Visualisation  
Fvs  
Downloads  
PDB

### > Structure details

Structure of Bovine anti-Hiv Fab Else6

PDB	8vbl
Species	BOS TAURUS
Method	X-RAY DIFFRACTION
Resolution	2.35Å
Number of Fvs	1
In complex	False
Light chain type	Lambda

**Fig 7: Results obtained for the entry**

Details

Visualisation

Fvs

Downloads

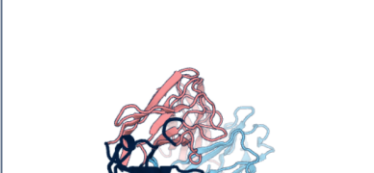
PDB ↗

> Structure details

Structure of Bovine anti-Hiv Fab Else6

PDB	8vbl
Species	BOS TAURUS
Method	X-RAY DIFFRACTION
Resolution	2.35Å
Number of Fvs	1
In complex	False
Light chain type	Lambda
Has constant region	True

> Structure visualisation



Key (Default Scheme):

- WI Chains
- VI Chains
- CDRs

Display options:

Spacefill

**Fig 8: Results obtained: ‘Structure Details’**

Details

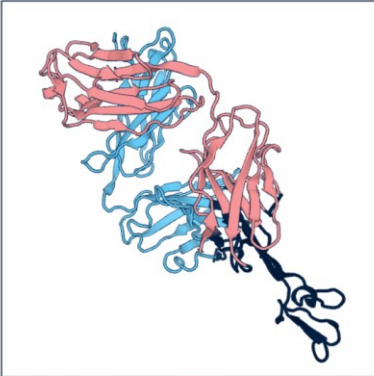
Visualisation

Fvs

Downloads

PDB ↗

> Structure visualisation



Please note the WebGL plugin needs to be enabled to use PV Viewer.

Key (Default Scheme):

- WI Chains
- VI Chains
- CDRs

Display options:

Spacefill

Wire

Ball&stick

Cartoon

Default colours

Colour by B-factor

Colour by chain

Colour by sec. structure

Colour by element

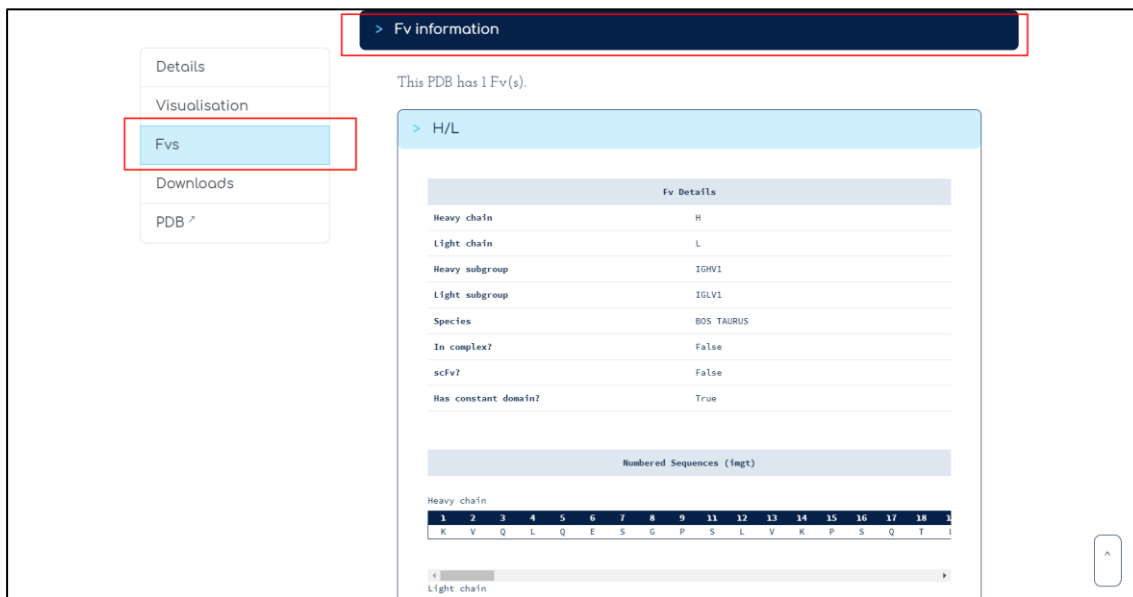
Spin on/off

> Fv information

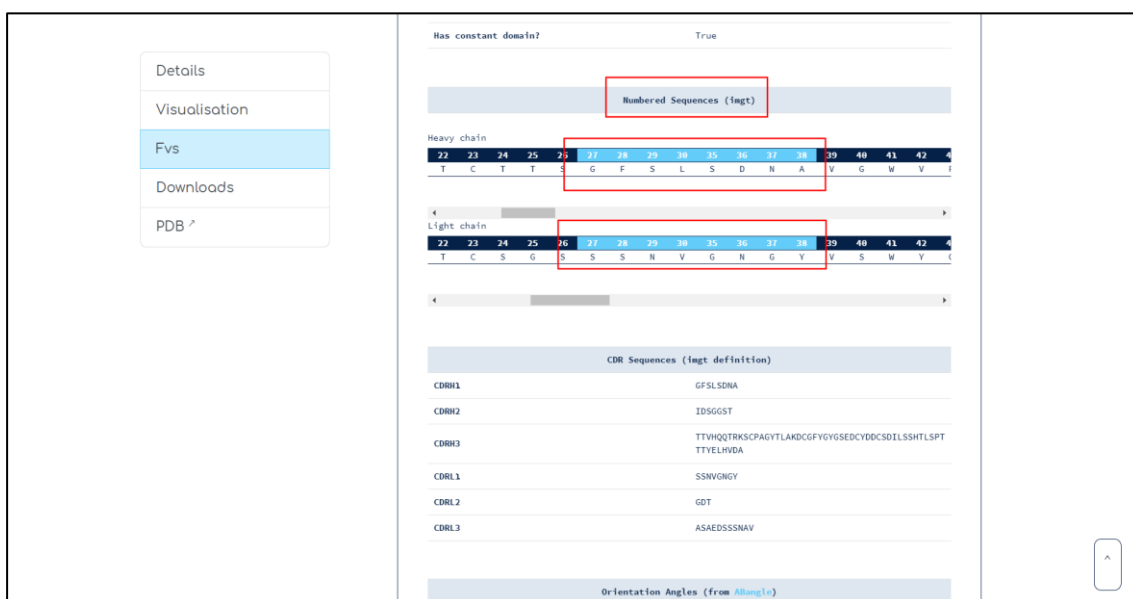
This PDB has 1 Fv(s).

**Fig 9: Results obtained: ‘Visualization’**

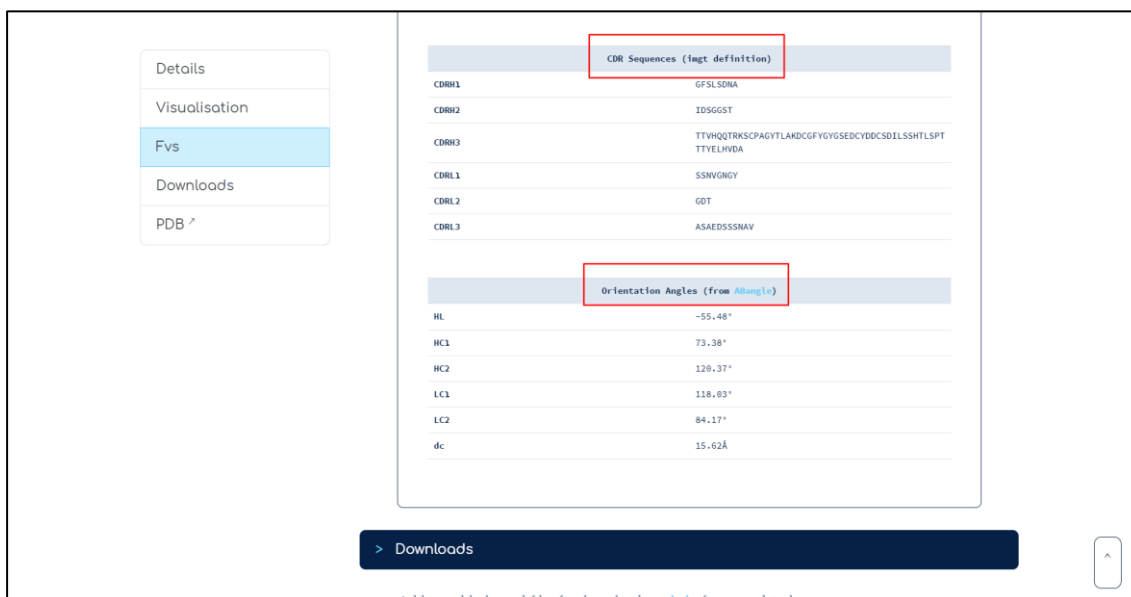




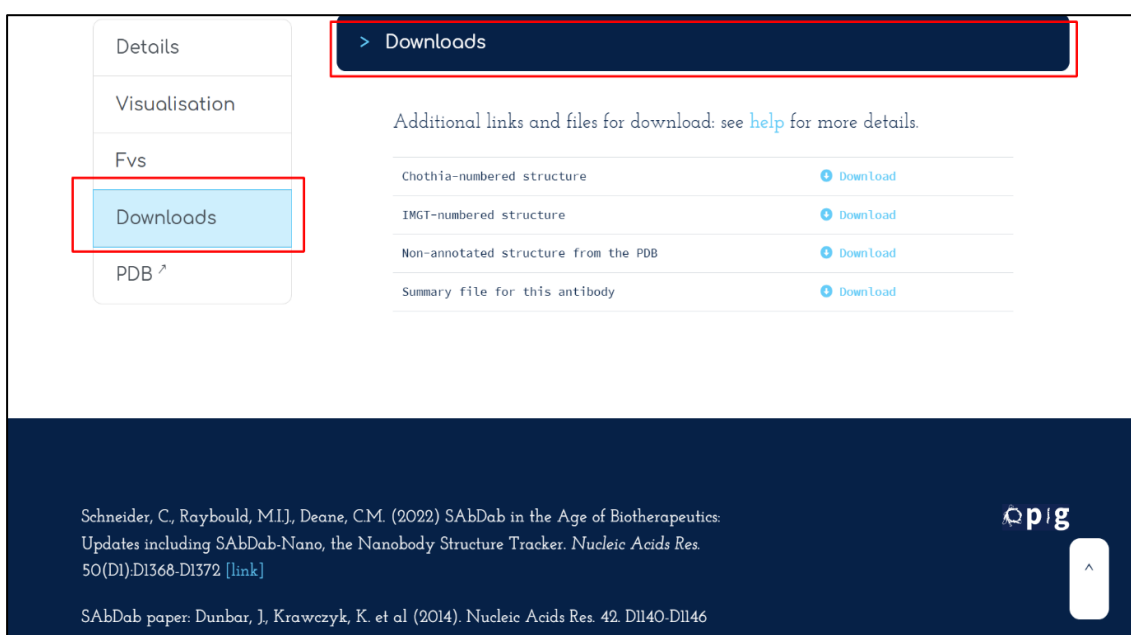
**Fig 10: Results obtained: 'Fv Information'  
[Header information for the Fv (H/L)]**



**Fig 10.a: Marked positions for the heavy chain and light chain for the Fv (H/L)  
through the chothia numbering system**



**Fig 10.b: CDR Sequences (IMGT definition) and Orientation angles for the Fv (H/L)**



**Fig 11: Links for downloading the structure under the 'Downloads' section**

## **RESULTS:**

The query 'Bovine anti-HIV Fab ElsE6' (PDB ID: 8VBL) was searched and studied using the Structural Antibody Database (SAbDab). Following information was studied for the selected entry:

### **1. Structure Details**

<b>Name</b>	Structure of bovine anti-HIV Fab ElsE6
<b>PDB</b>	8VBL

<b>Species</b>	<i>Bos taurus</i>
<b>Method</b>	X-RAY DIFFRACTION
<b>Resolution</b>	2.35Å
<b>Number of Fvs</b>	1
<b>In complex</b>	False
<b>Light chain type</b>	Lamba
<b>Has constant region</b>	True

## 2. Structure Visualization

The structure can be observed in terms of VH Chains, VL Chains and CDRs in the form of various display options (example: Wire) and colors.

Following information was studied for the Fvs:

<b>Fv</b>	<b>Header Information</b>		<b>Numbered Sequences (chothia)</b>
<b>H/L</b>	<b>Heavy chain</b>	H	<b>Heavy chain:</b> 27 – 38, 56 – 65, 105 – 117
	<b>Light chain</b>	L	
	<b>Heavy subgroup</b>	IGHV5	
	<b>Light subgroup</b>	IGKV8	
	<b>Species</b>	MUS MUSCULUS	<b>Light chain:</b> 27 – 38, 56 – 65, 105 – 117
	<b>In complex?</b>	True	
	<b>scFv?</b>	False	
	<b>Has constant domain?</b>	True	

Further information was studied about antigen details, CDR sequences (chothia definition) and orientation angles for each of the Fvs.

## 3. Downloads

Various links have been provided for the downloading:

- a. Chothia-numbered structure
- b. IMGT-numbered structure
- c. Non-annotated structure from the PDB
- d. Summary file for this antibody

## **CONCLUSION:**

The Antibody structure for the query ‘Bovine anti-Hiv Fab Else6’ (PDB ID: 8VBL) was studied using the Structural Antibody Database (SABDab). The SAbDab entry page provides extensive details about the antibody structure, including:

1. **Structure details:** resolution, R-factors, experimental method, etc.
2. **Visualization tools:** to view the antibody structure
3. Information on the **antibody variable (Fv) regions**
4. Options to **download** the structure coordinates in various formats

## **REFERENCES:**

1. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., & Deane, C. M. (2014). SABDab: the structural antibody database. *Nucleic acids research*, 42(Database issue), D1140–D1146. <https://doi.org/10.1093/nar/gkt1043>
  2. Schneider, C., Raybould, M.I.J., Deane, C.M. (2022) SABDab in the Age of Biotherapeutics: Updates including SABDab-Nano, the Nanobody Structure Tracker. *Nucleic Acids Res.* 50(D1): D1368-D1372. <https://doi.org/10.1093/nar/gkab1050>
  3. wwPDB.org. (n.d.). wwPDB: 8VBL. [https://www.wwpdb.org/pdb?id=pdb\\_00008vbl](https://www.wwpdb.org/pdb?id=pdb_00008vbl)
-

**WEBLEM: 3**

**AntiBodies Chemically Defined Database (ABCD)**

**(URL: <https://web.expasy.org/abcd/>)**

**AIM:**

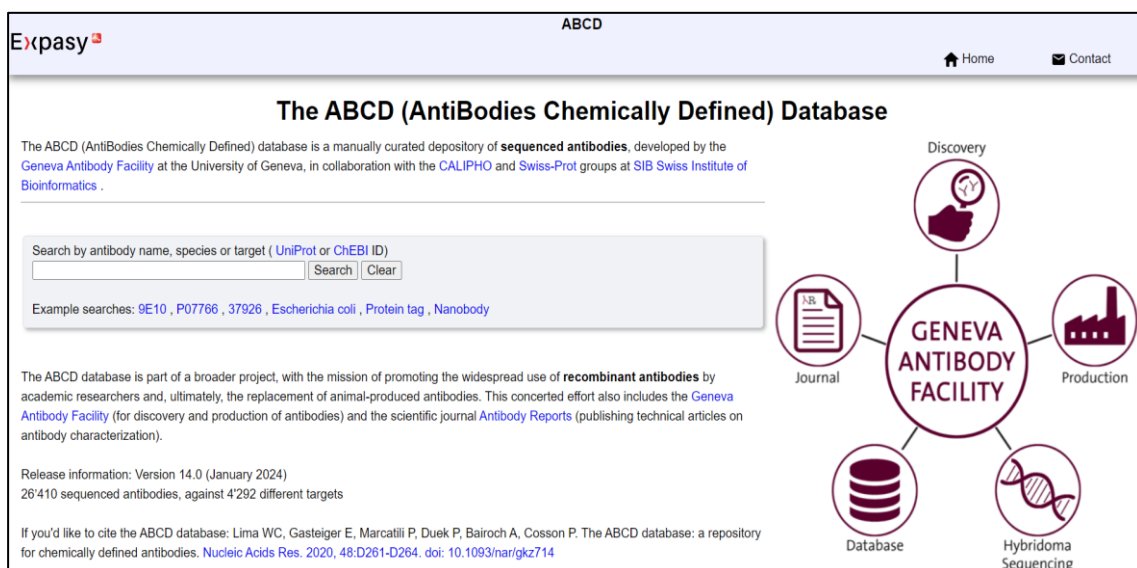
To study antibody sequence using ABCD database.

**INTRODUCTION:**

The ABCD (for AntiBodies Chemically Defined) database is a repository of sequenced antibodies, integrating curated information about the antibody and its antigen with cross-links to standardized databases of chemical and protein entities. It is freely available to the academic community, accessible through the ExpASY server (<https://web.expasy.org/abcd/>). The ABCD database aims at helping to improve reproducibility in academic research by providing a unique, unambiguous identifier associated to each antibody sequence. It also rapidly determines whether a sequenced antibody is available for a given antigen.

The ABCD (AntiBodies Chemically Defined) database is a manually curated repository of sequenced antibodies, developed by the Geneva Antibody Facility at the University of Geneva, in collaboration with the CALIPHO and Swiss-Prot groups at SIB Swiss Institute of Bioinformatics. The ABCD database is part of a broader project, aiming to promote the widespread use of recombinant antibodies by academic researchers and, ultimately, the replacement of animal-produced antibodies. This concerted effort also includes the Geneva Antibody Facility (for discovering and producing antibodies) and the scientific journal Antibody Reports (publishing technical articles on antibody characterization).

ABCD is a huge collection of AD-related data of molecular markers. The web interface contains information concerning the proteins, genes, transcription factors, SNPs, miRNAs, mitochondrial genes, and expressed genes implicated in AD pathogenesis. In addition to the molecular-level data, the database has information for animal models, medicinal candidates and pathways involved in the AD and some image data for AD patients.



**Fig 1: Homepage of ABCD Database**

## **REFERENCES:**

1. *ABCD - Database Commons*. (n.d.). <https://ngdc.cncb.ac.cn/databasecommons/database/id/6464>
  2. Lima, W. C., Gasteiger, E., Marcatili, P., Duek, P., Bairoch, A., & Cosson, P. (2019). The ABCD database: a repository for chemically defined antibodies. *Nucleic Acids Research*, 48(D1), D261–D264. <https://doi.org/10.1093/nar/gkz714>
  3. ABCD - SIB Swiss Institute of Bioinformatics | Expasy. (n.d.). <https://www.expasy.org/resources/abcd>
  4. ABCD - Database Commons. (n.d.-b). <https://ngdc.cncb.ac.cn/databasecommons/database/id/6771>
  5. Expasy - ABCD (AntiBodies Chemically Defined). (n.d.). <https://web.expasy.org/abcd/>
-

**WEBLEM: 3(A)**

**AntiBodies Chemically Defined Database (ABCD)**

**(URL: <https://web.expasy.org/abcd/>)**

**AIM:**

To study Foralumab antibody sequence using ABCD Database.

**INTRODUCTION:**

Antibodies are one of the most widespread tools used in biological sciences. However, they are currently deemed one of the major culprits in the reproducibility crisis plaguing bio-medical research. Problems include batch-to-batch variability, poorly characterized and/or non-validated antibodies that sometimes do not recognize the presumptive target, or recognize more than one target, lack of explicitly described procedures adapted to each antibody, decreasing scrutiny of results by scientists and misleading antibody nomenclature. The 2 million antibodies available on the market might represent as few as 250'000 actual clones.

The ABCD (for AntiBodies Chemically Defined) database is a repository of sequenced antibodies, integrating curated information about the antibody and its antigen with cross-links to standardized databases of chemical and protein entities. It is freely available to the academic community, accessible through the ExpASy server (<https://web.expasy.org/abcd/>). The ABCD database aims at helping to improve reproducibility in academic research by providing a unique, unambiguous identifier associated to each antibody sequence. It also allows to determine rapidly if a sequenced antibody is available for a given antigen.

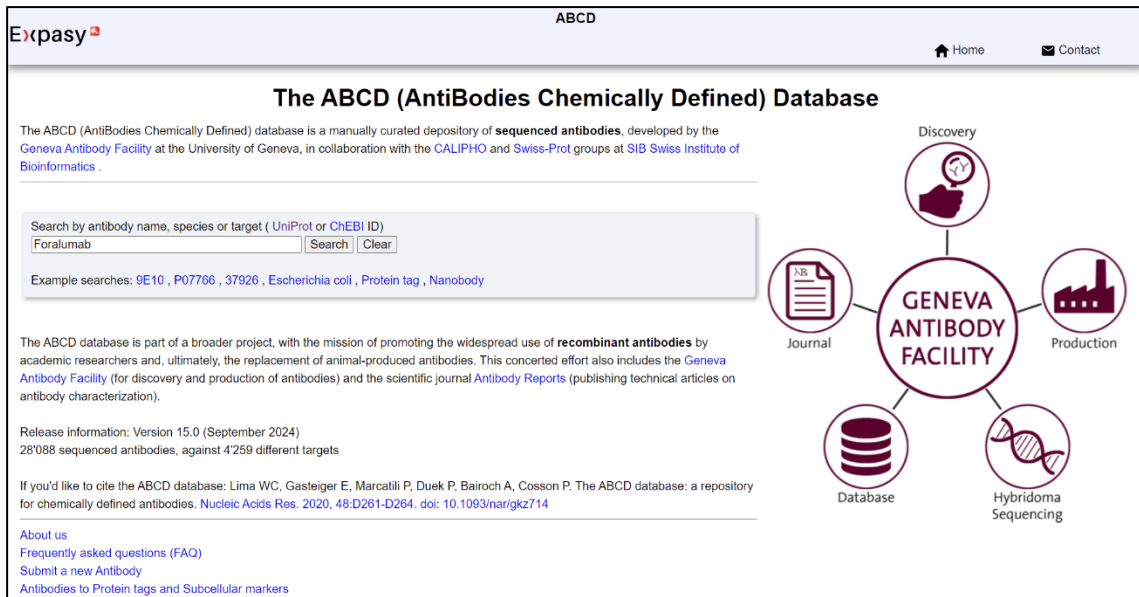
**Foralumab**

Foralumab is a fully human anti-CD3 monoclonal antibody that targets the CD3 complex on T cells, playing a significant role in modulating immune responses. By binding to CD3, it suppresses T cell activation and promotes the expansion of regulatory T cells, reducing inflammation and immune hyperactivity. Nasal administration of Foralumab has shown potential in reducing lung inflammation in COVID-19 patients, as evidenced by decreased pro-inflammatory cytokines. It is also being investigated for autoimmune diseases like multiple sclerosis and type 1 diabetes, as well as preventing transplant rejection. With an acceptable safety profile at lower doses, Foralumab represents a promising therapeutic agent for immune-related conditions.

**METHODOLOGY:**

1. Open the home page of ABCD Database (URL: <https://web.expasy.org/abcd/>)
2. Search for query Foralumab.
3. Open one entry (ID: ABCD\_AA611) from the obtained entries.
4. Interpret the results.

## OBSERVATION:



The screenshot shows the home page of the ABCD (AntiBodies Chemically Defined) Database. The page features a search bar with the text "Foralumab" and buttons for "Search" and "Clear". Below the search bar, there are example searches: "9E10", "P07766", "37926", "Escherichia coli", "Protein tag", and "Nanobody". The page also includes a diagram of the Geneva Antibody Facility, which is a central hub for antibody discovery, production, and sequencing. The diagram shows the facility's involvement in Discovery, Journal, Production, Database, and Hybridoma Sequencing. The page also contains information about the database's mission, release information (Version 15.0, September 2024), and a citation for the database.

**The ABCD (AntiBodies Chemically Defined) Database**

The ABCD (AntiBodies Chemically Defined) database is a manually curated depository of **sequenced antibodies**, developed by the **Geneva Antibody Facility** at the University of Geneva, in collaboration with the **CALIPHO** and **Swiss-Prot** groups at **SIB Swiss Institute of Bioinformatics**.

Search by antibody name, species or target ( UniProt or ChEBI ID)  
Foralumab Search Clear

Example searches: 9E10 , P07766 , 37926 , Escherichia coli , Protein tag , Nanobody

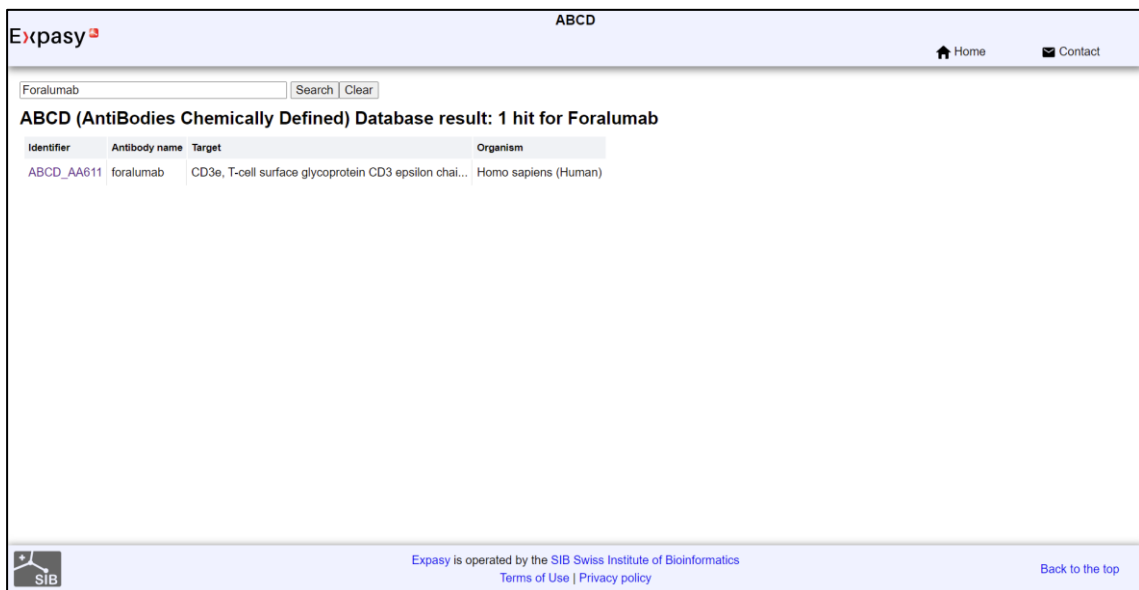
The ABCD database is part of a broader project, with the mission of promoting the widespread use of **recombinant antibodies** by academic researchers and, ultimately, the replacement of animal-produced antibodies. This concerted effort also includes the **Geneva Antibody Facility** (for discovery and production of antibodies) and the scientific journal **Antibody Reports** (publishing technical articles on antibody characterization).

Release information: Version 15.0 (September 2024)  
28'088 sequenced antibodies, against 4'259 different targets

If you'd like to cite the ABCD database: Lima WC, Gasteiger E, Marcatili P, Duek P, Bairoch A, Cosson P. The ABCD database: a repository for chemically defined antibodies. *Nucleic Acids Res.* 2020, 48:D261-D264. doi: 10.1093/nar/gkz714

[About us](#)  
[Frequently asked questions \(FAQ\)](#)  
[Submit a new Antibody](#)  
[Antibodies to Protein tags and Subcellular markers](#)

**Fig 1: Home Page of ABCD Database**



The screenshot shows the result page for a query of "Foralumab". The page displays a single hit for "Foralumab" with the following details:

Identifier	Antibody name	Target	Organism
ABCD_AA611	foralumab	CD3e, T-cell surface glycoprotein CD3 epsilon chai...	Homo sapiens (Human)

The page also includes a footer with the SIB logo, the text "Expasy is operated by the SIB Swiss Institute of Bioinformatics", and links for "Terms of Use" and "Privacy policy".

**Fig 2: Result page for Query Foralumab**



Antigen information	
Target type	Protein
Target link	UniProt: <a href="#">P07766</a> Homo sapiens (Human)
Target name	CD3e, T-cell surface glycoprotein CD3 epsilon chain, T-cell surface antigen T3/Leu-4 epsilon chain
Antibody information	
Antibody name	foralumab
Antibody synonyms	NI-0401, 28F11
Applications	ELISA, Flow cytometry, Surface plasmon resonance, Therapeutic
Cross-references	IMGT/mAb-DB: <a href="#">350</a>
Publications	Patent: <a href="#">US20060177896</a> Patent: <a href="#">WO2005118635</a> PMID: <a href="#">33649101</a> PMID: <a href="#">28096333</a> PMID: <a href="#">20848453</a>
Would you like to obtain this antibody?	
It can be produced at the <a href="#">Geneva Antibody facility</a> (for more information, please check here).	

**Fig 3: After Selecting an entry (ID: ABCD\_AA611)**

## **RESULTS:**

ABCD Database was explored to study the antigen-antibody information for query Foralumab the query was searched and 1 hit was obtained which was opened and studied. It contained information about target type, target link (UniProt: P07766), antibody name, cross references (IMGT/mAb-DB: 350) etc.

## **CONCLUSION:**

ABCD Database was explored for query Foralumab for antigen and antibody information. ABCD Database represents a vital resource for organizations seeking to optimize data management and analysis. Its combination of usability, analytical capabilities, and security makes it an invaluable tool for driving informed decisions and enhancing operational efficiency. As organizations increasingly rely on data-driven strategies, the ABCD Database is well-positioned to support their needs effectively.

## **REFERENCES:**

1. Moreira, T. G., Matos, K. T. F., De Paula, G. S., Santana, T. M. M., Da Mata, R. G., Pansera, F. C., Cortina, A. S., Spinola, M. G., Baecher-Allan, C. M., Keppeke, G. D., Jacob, J., Palejwala, V., Chen, K., Izzy, S., Healey, B. C., Rezende, R. M., Dedititis, R. A., Shailubhai, K., & Weiner, H. L. (2021). Nasal Administration of Anti-CD3 Monoclonal Antibody (Foralumab) Reduces Lung Inflammation and Blood Inflammatory Biomarkers in Mild to Moderate COVID-19 Patients: A Pilot Study. *Frontiers in immunology*, 12, 709861. <https://doi.org/10.3389/fimmu.2021.709861>
2. Lima, W. C., Gasteiger, E., Marcatili, P., Duek, P., Bairoch, A., & Cosson, P. (2019). The ABCD database: a repository for chemically defined antibodies. *Nucleic Acids Research*, 48(D1), D261–D264. <https://doi.org/10.1093/nar/gkz714>

**WEBLEM: 4**

**Antibody Numbering using KabatMan and Chothia Database and AbRSA  
Numbering Tool as Demo**

**AIM:**

To use KabatMan database and AbRSA numbering tool as demo.

**INTRODUCTION:**

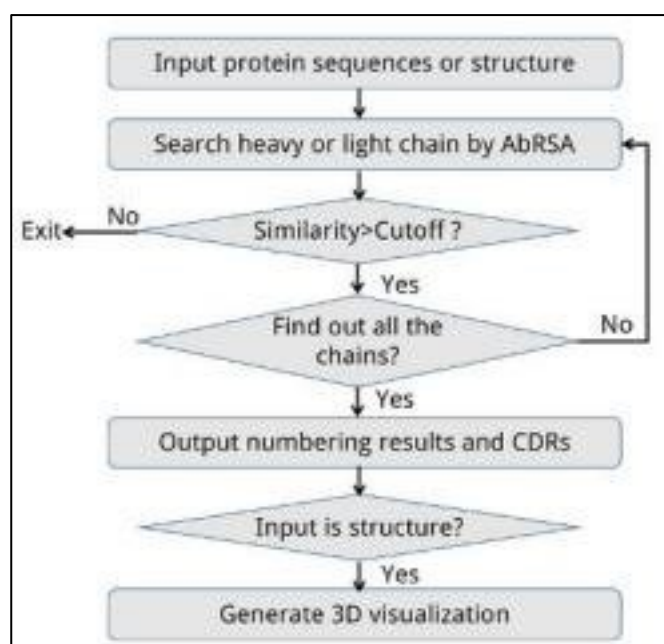
An important prerequisite for antibody humanization requires standardized numbering methods to define precisely complementary determining regions (CDR), frameworks and residues from the light and heavy chains that affect the binding affinity and/or specificity of the antibody-antigen interaction. The recently generated deep-sequencing data and the increasing number of solved three-dimensional structures of antibodies from human and non-human origins have led to the emergence of numerous databases. However, these different databases use different numbering conventions and CDR definitions. In addition, the large fluctuation of the variable chain lengths, especially in CDR3 of heavy chains (CDRH3), hardly complicates the comparison and analysis of antibody sequences and the identification of the antigen binding residues. This review compares and discusses the different numbering schemes and “CDR” definition that were established up to date. Furthermore, it summarizes concepts and strategies used for numbering residues of antibodies and CDR residues identification. Finally, it discusses the importance of specific sets of residues in the binding affinity and/or specificity of immunoglobulins.

Antibody engineering methods require precise identification of the residues that have an impact on the interaction or affinity of the antibody for its target antigen. CDR-grafting aims to decrease the immunogenicity of non-human antibodies by engineering the variable regions directed against the target antigen. This method requires an accurate identification of the CDRs and therefore an adequate alignment of antibody sequences from human and non-human species. Moreover, it has been shown that residues from the framework regions might also exert a strong impact on the antibody affinity. Thus, the precise identification of corresponding positions in human and animal immunoglobulin chains is essential. However, the use of different amino acid numbering schemes currently available in the literature is confusing and might lead to aberrant identification of framework and CDR residues. Therefore, it is of crucial importance to understand the different numbering schemes and, consequently, being able to compare them.

KabatMan database: To enter KabatMan database (<http://www.bioinf.org.uk/abs/simkab.html>). The purpose of maintaining the Kabat Database of aligned sequences of proteins of immunological interest, it provides useful correlations between structure and function for this special group of proteins from their nucleotide and amino acid sequences to their tertiary structures. The Kabat Database was initially started in 1970 to determine the combining site of antibodies based on the available amino acid sequences at that time. Bence Jones proteins, mostly from human, were aligned, using the now-known Kabat numbering system, and a quantitative measure, variability, was calculated for every position. Immunologists have extensively used it to derive useful structural and functional information from the primary sequences of these proteins. The Kabat Database may be accessed for searching, sequence retrieval and analysis by a few different methods: electronic mail, WWW, and ftp.

AbRSA tool: To enter AbRSA tool enter (<http://cao.labshare.cn/AbRSA/index.html>). Antibody sequence numbering and complementarity determining region (CDR) delimitation have wide applications in antibody engineering. They are generally obtained by mapping query sequences to the retrospective patterns. However, due to the enormous diversity of antibody sequences, novel patterns are often generated in antibody affinity maturation. They may not be recognized by the traditional methods. Antibody Region-Specific Alignment (AbRSA) integrates the biological insight of antibody region-specific feature with dynamic programming to improve the robustness of antibody numbering. Benchmarks show AbRSA is a powerful method in numbering unusual antibodies and distinguishing between antibody and non-antibody sequences.

Workflow: The pipeline of AbRSA web service is shown in the following Figure. The input could be either the protein sequence or structure. Multiple protein sequences are supported if the sequences are in FASTA format. The program judges whether it is a heavy chain, light chain or neither by comparing the sequence identities with consensus sequences. After all the possible heavy or light chains are found out, the program will output the numbering results and the location of FRs and CDRs in the sequences. If the input is a protein structure (PDB format), the web page will generate its interactive 3D visualization powered by 3Dmol JavaScript library. CDRs will be highlighted in colors. The 3D view can be rotated, translated, and re-sized by dragging, scrolling the mouse. We believe this feature could help to understand where and how antibody binding with antigen.



**Fig 1: Workflow of AbRSA tool**

## OBSERVATIONS:

### KabatMan database:

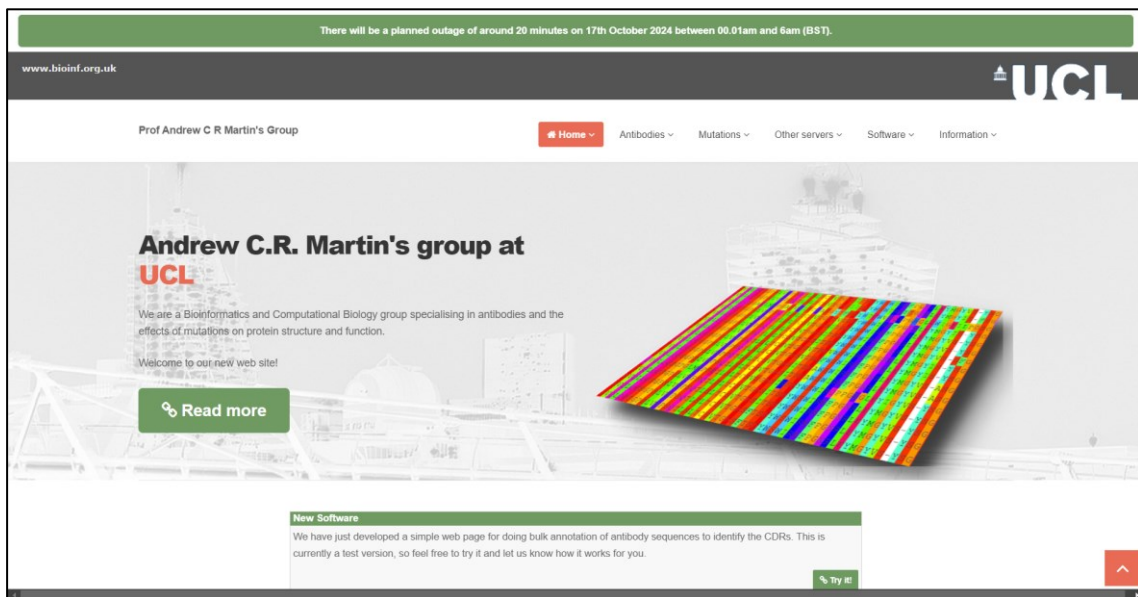


Fig 213: Main page to enter KabatMan database

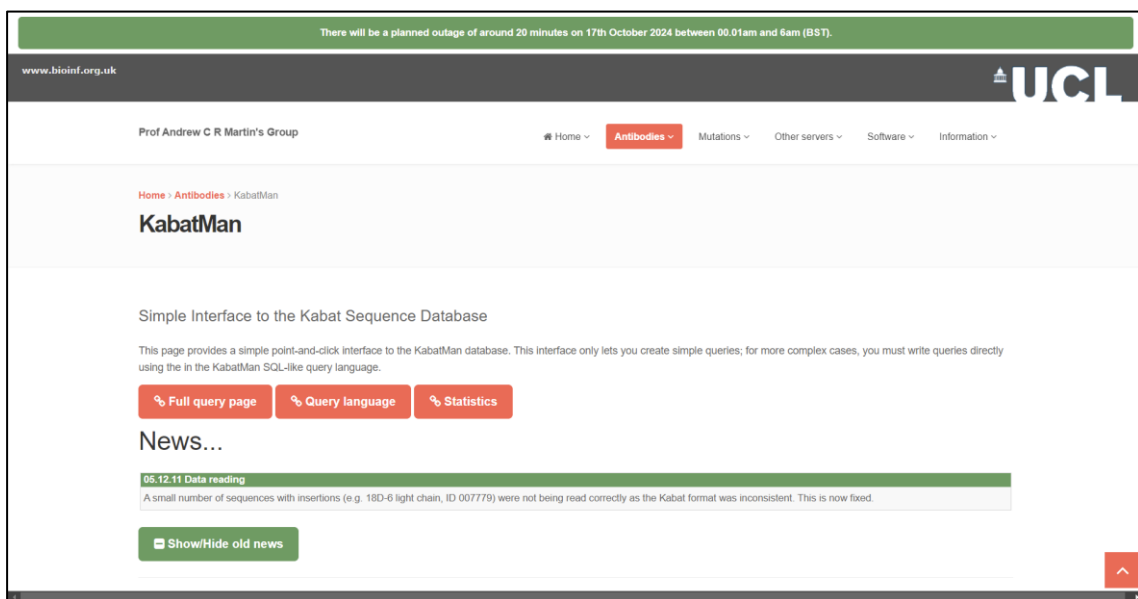
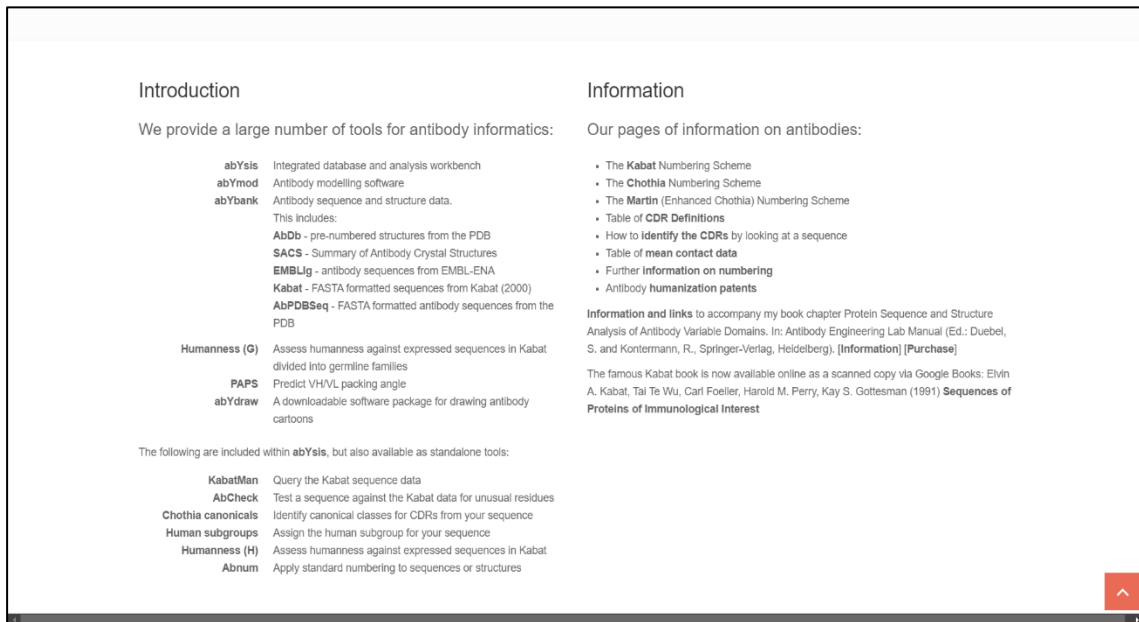
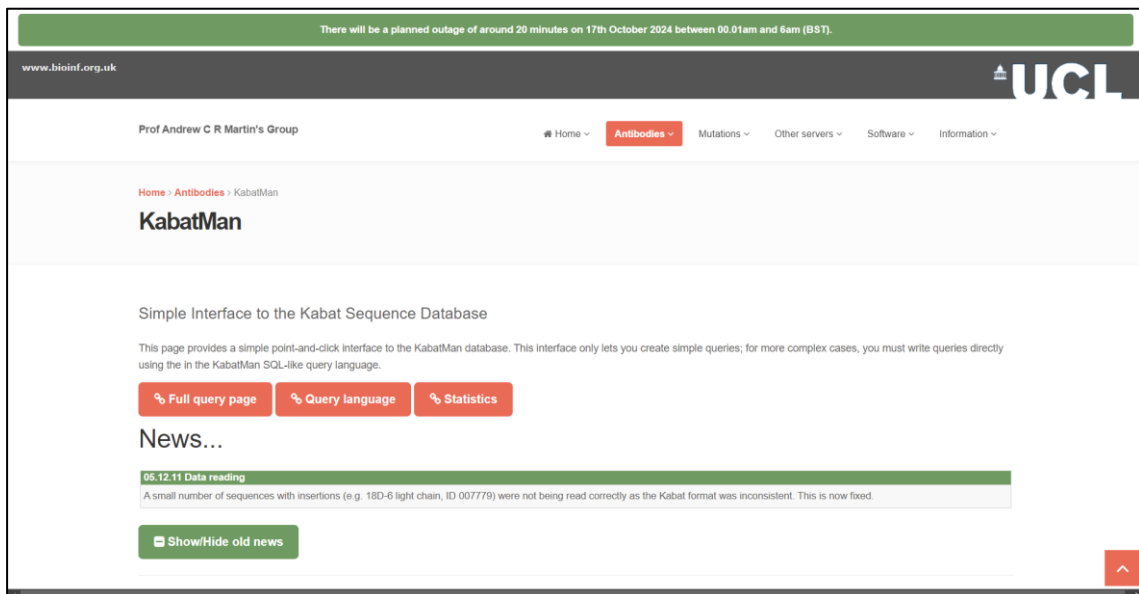


Fig 3: Antibodies page which gives the list of all available tools



**Fig 4: Antibodies page which shows KabatMan database**



**Fig 5: Homepage of KabatMan database**

### Examples

In the examples below, keywords are given in upper case, but this is only done for clarity. The database is completely case-insensitive.

1. Find all complete antibodies where the antigen is known with loop lengths:
 

```
SELECT name,antigen,length(11),length(12),length(13),
       length(h1),length(h2),length(h3)
WHERE antigen ne '' complete eq true AND
```
2. Get the sequences of all complete mouse antibodies which bind to lysozyme, display the results in PIR format:
 

```
SELECT pir
WHERE source includes mouse
       antigen includes lysozyme AND
       complete eq true AND
```
3. Find all antibodies with 11 residue CDR-L1s and a proline at the sixth position:
 

```
SELECT name, l1
WHERE len(11) eq 11 res(L29) eq P AND
```
4. Find all complete antibodies with the sequence Ser-Ala-Ser-Ser-Ser in the light chain:
 

Note that there must be no spaces in the sequence

```
SELECT name, light
WHERE complete = t
       light includes SASSS AND
```

**Fig 6: Example queries available in KabatMan database**

Home > Antibodies > KabatMan

## KabatMan

Simple Interface to the Kabat Sequence Database

This page provides a simple point-and-click interface to the KabatMan database. This interface only lets you create simple queries, for more complex cases, you must write queries directly using the in the KabatMan SQL-like query language.

### News...

**05-12-11 Data reading**  
 A small number of sequences with insertions (e.g. 18D-6 light chain, ID 007779) were not being read correctly as the Kabat format was inconsistent. This is now fixed.

What information do you wish to display?

Name  
 Antigen  
 Accession number (and link to raw data):  
 Light chain  Heavy chain  
 Sequence of:

**Fig 7: Click on full query on KabatMan homepage**



**Fig 8: Search bar to enter sequence taken from PDB database**



**Fig 9: Result page for the query**



Fig 10: Result page showing the number of hits for the entered query

**AbRSA tool:**

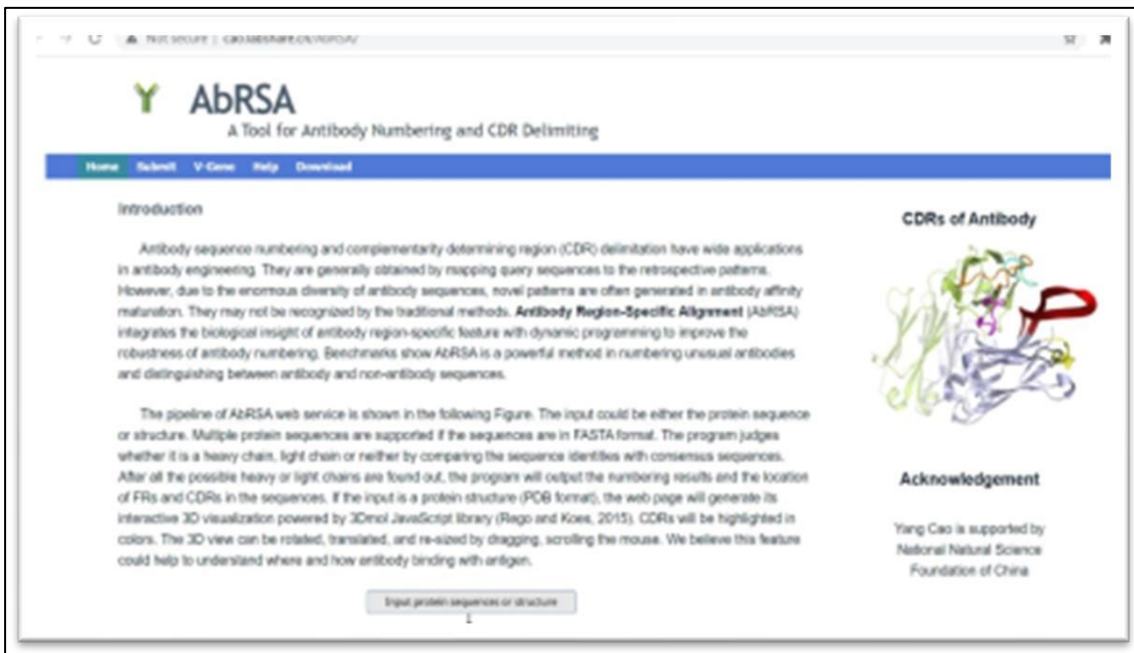
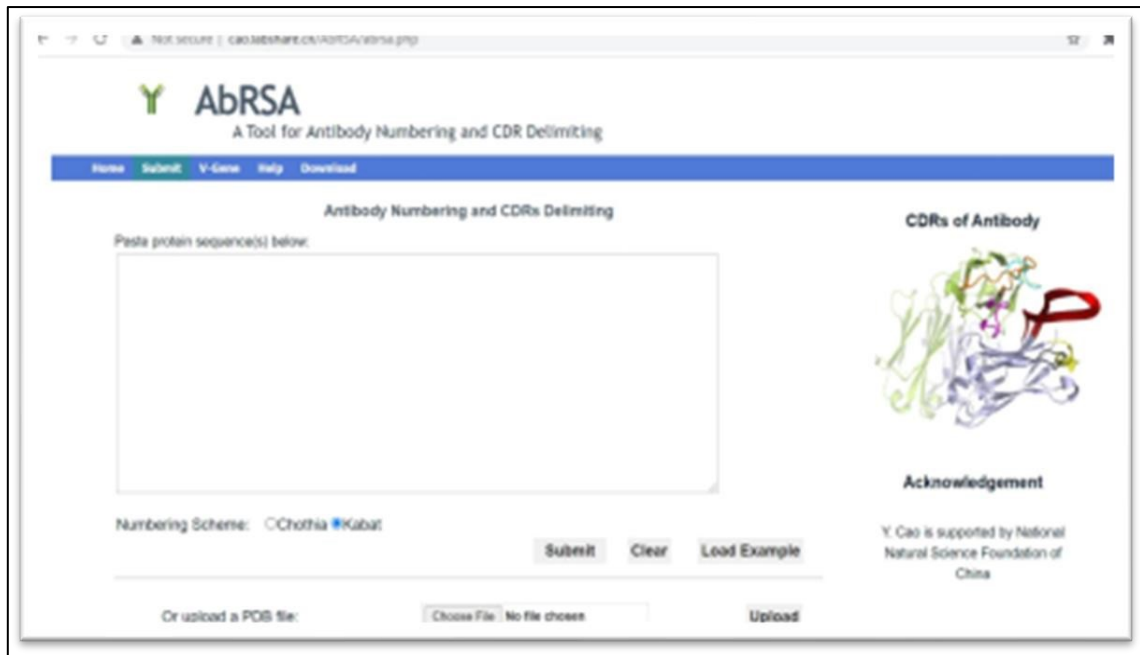
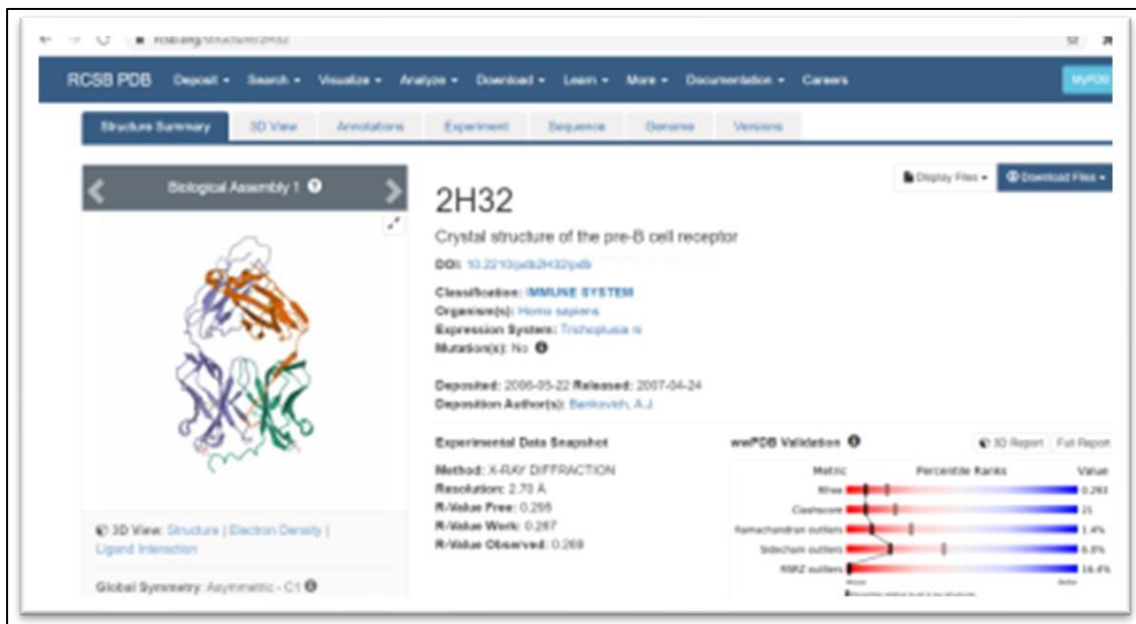


Fig 11: Homepage of AbRSA tool





**Fig 12: Paste the FASTA sequence taken from PDB database**



**Fig 13: Result page of PDB database**

```
2H32_1|Chain A|Immunoglobulin iota chain|Homo sapiens (9606)
PVUQPNWSSALETTERLCTLRIDQSDIVSYVGGPDPFPPFLLRVPSQDLSQDPQPPFSSDGLVWHDNLSZELQFDEKAVYCAWGRSSDLSRERDDEEKEPTAATFV
2H32_1|Chain B|Immunoglobulin omega chain|Homo sapiens (9606)
LTVWPESSDQTLVLSQPKATPSVTLFPPSSSEELQANKATLVCLMDFYPGILT
VYTKADGTPITQGVEMTTPSKQSNKYAASSYLSLTPEQWRSRSSYSCQVWHEG
STVEKTVAPAECS
2H32_1|Chain C|auth H|Immunoglobulin heavy chain|Homo sapiens (9606)
EVQLVQSGAEVVKKPGESLKISCKGSGYSFTSYWIGAVRQMPGKLEWNGIITYPG
DSDTRYSPSFQGGVITISADKSIESTAYLQWSSLKASDTANYCARHYYYYYGHEV
NGQGTTVTVSSMSASAPTLFPLVSCENSPDTSVAVGCLAQDFLPDSITFSWK
YKNSDLSSTRGFPVLRGGRYAATSQVLLPSKDVMOGTDEHVCKVQHPNGNK
EKVPLM
```

Fig 14: FASTA sequence received from PDB database

The screenshot shows the AbRSA web tool interface. The title is "AbRSA: A Tool for Antibody Numbering and CDR Delimiting". The main heading is "Antibody Numbering and CDRs Delimiting". On the left, there is a text area with the following FASTA sequence pasted in:

```
>2H32_2|Chain B|Immunoglobulin omega chain|Homo sapiens (9606)
SVTHVFGSGTQLTVLSQPKATPSVTLFPPSSSEELQANKATLVCLMDFYPGILT
VYTKADGTPITQGVEMTTPSKQSNKYAASSYLSLTPEQWRSRSSYSCQVWHEG
STVEKTVAPAECS
>2H32_3|Chain C|auth H|Immunoglobulin heavy chain|Homo sapiens (9606)
EVQLVQSGAEVVKKPGESLKISCKGSGYSFTSYWIGAVRQMPGKLEWNGIITYPG
DSDTRYSPSFQGGVITISADKSIESTAYLQWSSLKASDTANYCARHYYYYYGHEV
NGQGTTVTVSSMSASAPTLFPLVSCENSPDTSVAVGCLAQDFLPDSITFSWK
YKNSDLSSTRGFPVLRGGRYAATSQVLLPSKDVMOGTDEHVCKVQHPNGNK
EKVPLM
```

Below the text area, the "Numbering Scheme" is set to "Kabat". There are buttons for "Submit", "Clear", and "Load Example". At the bottom, there is an option to "Or upload a PDB file" with a "Choose File" button and an "Upload" button. On the right side, there is a 3D ribbon diagram of an antibody structure under the heading "CDRs of Antibody". Below that is an "Acknowledgement" section stating "Y. Cao is supported by National Natural Science Foundation of China".

Fig 15: FASTA sequence pasted in AbRSA tool page

AbRSA Result (Kabat)

Warning: No Antibody Variable Domain Sequence was Detected for 2H32\_2|CHAIN B|IMMUNOGLOBULIN OMEGA CHAIN|HOMO SAPIENS (9606)

**Summary of CDRs**

Name	Type	CDR1	CDR2	CDR3
2H32_3 CHAIN C AUTH H IMMUNOGLOBULIN HEAVY CHAIN HOMO SAPIENS (9606)	VH	SYWIG	IYPGDSDRYSPSPQG	HYYYYYGMDV
2H32_1 CHAIN A IMMUNOGLOBULIN IOTA CHAIN HOMO SAPIENS (9606)	VL	TLRNDHDIGVYSVY	YFSGSDKSQGP	ANGARSSEKEEREREWEE

Download Numbering Results: NumberingFile

CDRs of Antibody

Acknowledgement

Fig 16: Result page of AbRSA tool which shows summary of CDRs

Variable Domain

>2H32\_1|CHAIN A|IMMUNOGLOBULIN IOTA CHAIN|HOMO SAPIENS (9606)

```

1 QPVLHQPFAMSSALGTTIRLTCYLRNDHDIGVYSVYVYQQRPGPPRFLLRYPGSDRQD 60
61 QPQVPPFRPSGSKDVARRRGYLSISELQPEDEAMFYCANIGARSSEKERESEKEREKDKPTA 120
121 ARTKVP

```

>2H32\_2|CHAIN B|IMMUNOGLOBULIN OMEGA CHAIN|HOMO SAPIENS (9606)

```

1 SVTVVFGSGDTGLTVLSQPKATFSVTLFFPSSSELQAKKATLVCLMDFYNGILTVYWKAD 60
61 GYFITYQVDEMTFPSRQSNKCYAASYLELTFPQWHRKSYSCQVMSGGSTVETVAFAEC 120
121 H

```

>2H32\_3|CHAIN C|AUTH H|IMMUNOGLOBULIN HEAVY CHAIN|HOMO SAPIENS (9606)

```

1 KVVLSVCSGAEVKKKGRSLKISCEGGGYSRFTSYRIGVYVCMGKGLMMG-IYYPGSDRQD 60
61 IYYPGSDRQDPTISAEKSLSTAYLQSSSLKASCTAMTCARHYTYTDMQVWQDPTVYVSEW 120
121 SAKPTLFPPIVQCSNFPDTSNVAVGCLAQDFLDYIYFRMKTENKQDISSTRQFPVLR 180
181 GGRYAATSGVLLPSEKVMQPTDEIVVCKVQIFRFGREKQVYLF

```

- CDRs are highlighted in colors (CDR1, CDR2, CDR3).
- The gray letters indicate the non-variable-domain region.
- The underlined black letters indicate variable domain of heavy chain while the other black letters indicate variable domain of light chain.

Y. Cao is supported by National Natural Science Foundation of China

Fig 17: Variable domain results



Fig 18: Light chains for the query sequence entered



Fig 19: Heavy chain for the sequence entered

## **CONCLUSION:**

KabatMan database and AbRSA tool was used for antibody numbering.

## **REFERENCES:**

1. Abhinandan, K.R., and Andrew C.R. Martin. Analysis and Improvements to Kabat and Structurally Correct Numbering of Antibody Variable Domains. Molecular Immunology, vol. 45, no. 14, Aug. 2008, pp. 3832–3839, <https://doi.org/10.1016/j.molimm.2008.05.022>

2. AbRSA: A Tool for Antibody Numbering and CDR Delimiting. Cao.labshare.cn, <https://cao.labshare.cn/AbRSA/index.html>
  3. Bioinf.org.uk - Prof. Andrew C.R. Martin's Group at UCL. <https://www.bioinf.org.uk>
  4. Johnson, G. "Kabat Database and Its Applications: 30 Years after the First Variability Plot." Nucleic Acids Research, vol. 28, no. 1, 1 Jan. 2000, pp. 214–218, <https://doi.org/10.1093/nar/28.1.214>
-

**WEBLEM: 5**

**Introduction to STCRDab database**

**(URL: <https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/>)**

**INTRODUCTION:**

The Single-chain T-cell Receptor Database (STCRDab) is a specialized repository designed to support the study and development of single-chain T-cell receptors (scTCRs), which are key players in the immune response. T-cell receptors (TCRs) are essential in recognizing antigenic peptides presented by Major Histocompatibility Complex (MHC) molecules, allowing T-cells to identify and attack infected or abnormal cells. Single-chain T-cell receptors, which are engineered forms of natural TCRs, are increasingly used in research and immunotherapy, particularly in personalized cancer treatments and autoimmune disease research. The STCRDab provides a comprehensive collection of sequence, structure, and functional data on these receptors, thus facilitating their analysis, modification, and application in various biological and clinical contexts.

T-cell receptors are surface proteins found on T-lymphocytes, specialized white blood cells that play a critical role in the adaptive immune system. TCRs enable T-cells to detect and bind to specific antigens, such as viral or tumor-derived peptides, that are presented on the surface of infected or abnormal cells via MHC molecules. Once the TCR recognizes an antigen, the T-cell is activated and triggers an immune response to eliminate the threat.

TCRs are typically composed of two chains:  $\alpha$  (alpha) and  $\beta$  (beta) in most T-cells, or  $\gamma$  (gamma) and  $\delta$  (delta) in a smaller subset. These chains combine to form a unique antigen-binding site, providing the ability to recognize a vast array of foreign peptides. TCRs do not directly bind free-floating antigens, unlike antibodies. Instead, they recognize antigenic peptides that are presented by MHC molecules on the surface of cells.

This specific interaction between TCRs, MHC molecules, and peptides is at the heart of T-cell-mediated immunity, which is crucial for identifying and eliminating virus-infected, malignant, or abnormal cells.

**Single-Chain T-Cell Receptors (scTCRs)**

Single-chain T-cell receptors (scTCRs) are engineered constructs that combine the antigen-binding domains of TCRs into a single polypeptide chain. These constructs simplify the natural heterodimeric structure of TCRs into a single chain, retaining the binding specificity while making them more stable and easier to produce for therapeutic purposes. scTCRs have become highly valuable in immunotherapy, particularly in CAR-T cell therapies (Chimeric Antigen Receptor T-cells), where engineered TCRs are modified to recognize specific tumor antigens, enabling precision targeting of cancer cells.

The primary advantage of scTCRs lies in their ability to be custom-designed for antigens, making them powerful tools for personalized medicine. They can be used in treating cancers, viral infections, and autoimmune diseases by enabling highly targeted immune responses.

The STCRDab database was created to provide a centralized platform for the accumulation, sharing, and analysis of single-chain T-cell receptor data. Its goal is to advance the understanding of TCRs and their therapeutic applications by providing a comprehensive resource for researchers engaged in T-cell biology, immunotherapy, and vaccine development.

STCRDab serves as a multi-functional database offering several key features to facilitate research:

1. **TCR Sequences and Annotations:** The database contains a vast array of TCR sequences, focusing on both natural and engineered single-chain TCRs. Each sequence entry is accompanied by relevant annotations, such as the antigen it binds to, the MHC class involved, and structural features.
2. **Structural Data:** For many TCRs and scTCRs, structural data is available, allowing researchers to explore the three-dimensional conformation of these receptors. The structure-function relationship is critical for designing scTCRs with enhanced binding properties or specificity, making this feature of great importance to those involved in rational drug design or antigen-targeted therapy.
3. **Functional Information:** In addition to sequence and structural data, STCRDab includes functional characterizations of TCRs, such as antigen specificity, binding affinity, and T-cell activation strength. This data helps researchers understand how different TCRs function in diverse immunological contexts, which is crucial for developing therapies targeting specific diseases like cancer or autoimmune disorders.
4. **Mutational Analysis:** STCRDab offers information on mutated TCRs and their effects on binding affinity, antigen specificity, and MHC interaction. This feature supports the engineering of optimized scTCRs for use in experimental models or therapeutic applications.
5. **Cross-Species Data:** The database includes TCR data from various species, allowing researchers to perform comparative studies that might offer insights into evolutionary conservation and diversity of T-cell receptors. This cross-species comparison is particularly valuable for preclinical studies where animal models are used to test TCR-based therapies.
6. **MHC-TCR-Peptide Complex Information:** Given the critical role of MHC in TCR-antigen recognition, STCRDab provides information on MHC-bound peptides and their interactions with TCRs. This helps in understanding the specificity of TCRs for certain MHC alleles and the potential for cross-reactivity with other peptides.
7. **Search and Analysis Tools:** The database offers a suite of bioinformatics tools that allow users to search, compare, and analyze TCR data. This includes sequence alignment, structural modeling, and antigenic epitope prediction, which are crucial for researchers looking to design novel TCRs for therapeutic use.

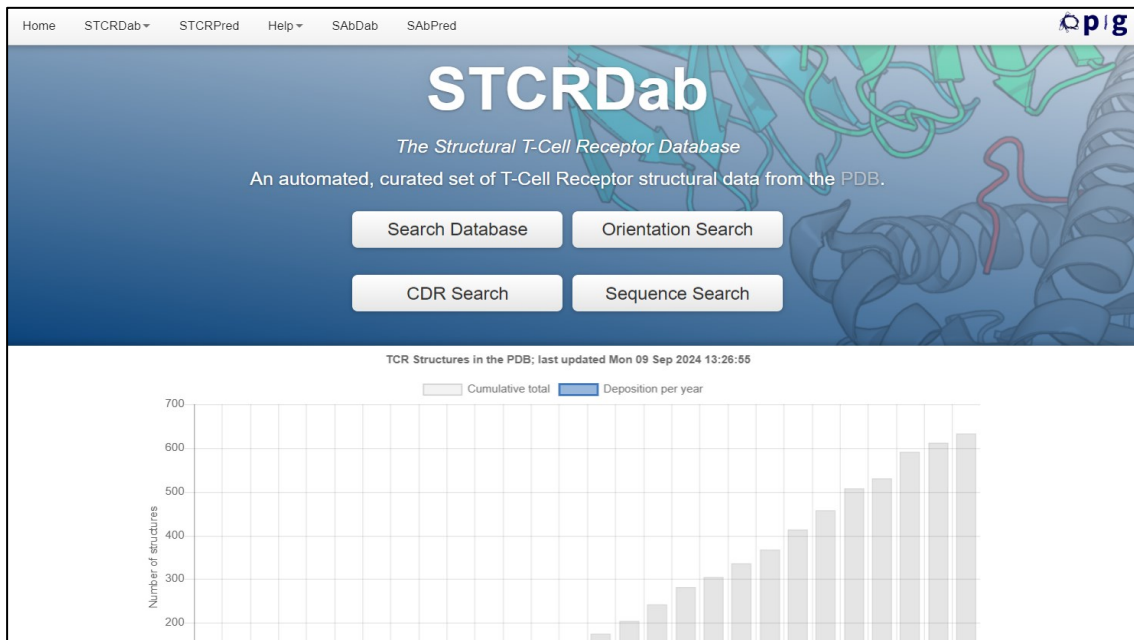
### **STCRPred:**

STCRPred is a bioinformatics tool used for modeling the structure of T-cell receptors (TCRs). It includes several computational tools designed to predict and analyze TCR structures, which are essential for understanding immune responses and designing therapeutic TCR-related proteins.

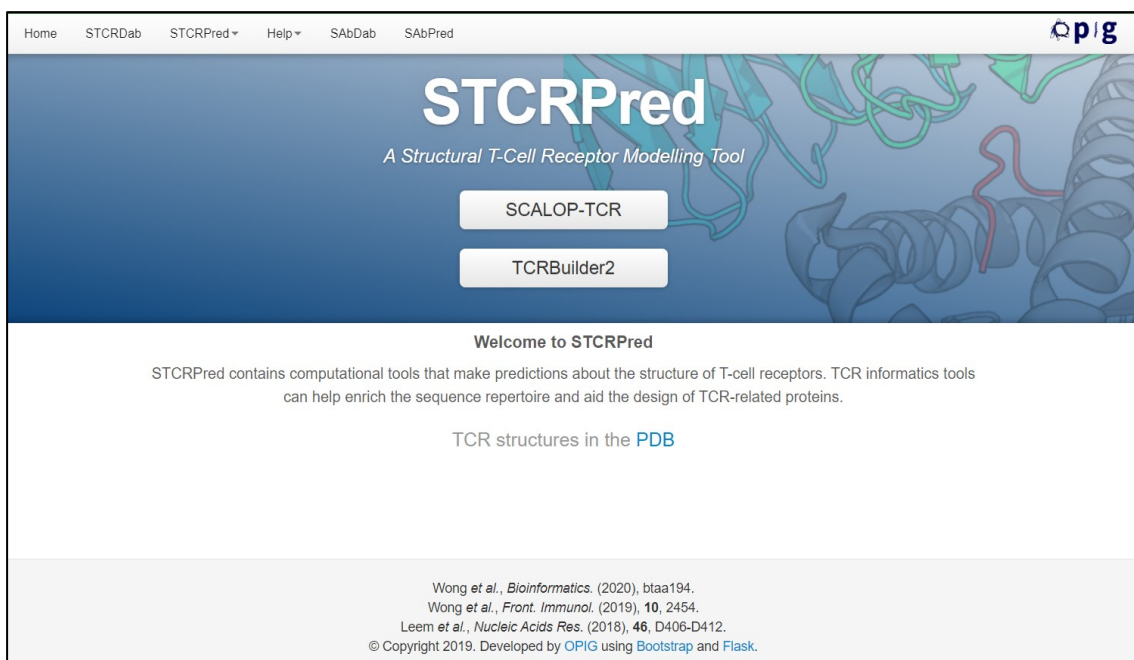
One of its main functions is the **sequence-based prediction** of complementarity-determining regions (CDRs), which are crucial for TCR recognition of antigens. STCRPred uses a tool called SCALOP-TCR to predict the canonical form of CDRs based on sequence data. The tool



assigns sequences to clusters by analyzing the structural information available from databases like the Protein Data Bank (PDB).



**Fig 1: Homepage of STCRDab**



**Fig 2: Homepage of STCRPred**

## **REFERENCES:**

1. Li, J., & Zhang, C. (2020). Advances in T-cell receptor engineering for cancer immunotherapy. *Frontiers in Immunology*, 11, 1082. <https://doi.org/10.3389/fimmu.2020.01082>



2. Morris, G. P., & Allen, P. M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature Immunology*, 13(2), 121–128. <https://doi.org/10.1038/ni.2190>
  3. Smith, J. A., & June, C. H. (2019). CAR T cell therapies: The road to universal T cells. *Nature Reviews Drug Discovery*, 18(7), 481–493. <https://doi.org/10.1038/s41573-019-0025-2>
-

**DATE: 24/09/2024**

**WEBLEM: 5(A)**

**Structural T-cell Receptor Database (STCRDab)**

**(URL: <https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/>)**

**AIM:**

To retrieve CDR position in query 2UWE using STCRDab database.

**INTRODUCTION:**

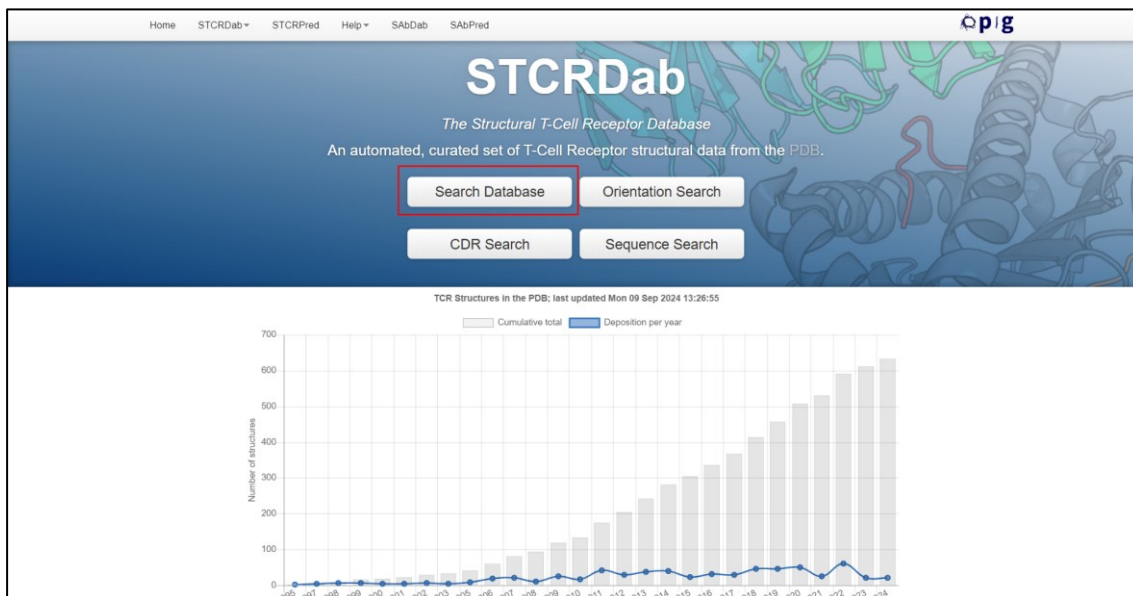
The STCRDab Database is an essential online resource designed to assist researchers working on single-chain T-cell receptors (scTCRs), which play a pivotal role in the immune system's ability to recognize and respond to antigens. scTCRs are engineered constructs that simplify the natural T-cell receptor, allowing for greater stability and targeted therapeutic use. This database offers a rich collection of information on TCR sequences, structural data, and functional characterizations, facilitating in-depth studies into how these receptors interact with antigens presented by MHC molecules.

The database serves as a valuable tool for advancing research in areas such as cancer immunotherapy, where engineered TCRs are used to precisely target tumor antigens, and in the study of infectious diseases, where TCRs recognize pathogen-derived peptides. Additionally, STCRDab provides insights into autoimmune conditions by analyzing how TCRs interact with self-antigens. Key features of STCRDab include a robust set of tools for sequence alignment, structural modeling, and mutational analysis, as well as comprehensive data on TCR-antigen specificity and MHC interactions. The database supports cross-species comparisons, helping researchers understand evolutionary aspects of TCR diversity and function. By centralizing TCR-related data, STCRDab accelerates discoveries in immunology, fosters the development of novel therapies, and plays a critical role in the growing field of personalized medicine.

**METHODOLOGY:**

1. Visit the homepage of the STCRDab Database by accessing the URL: <https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/>
2. Retrieve the antibody PDB ID: 2UWE from the PDB Database.
3. Input the retrieved PDB ID (2UWE) into the PDB search option in STCRDab.
4. Examine each section of the results displayed.
5. Provide an interpretation of the observed results.

## OBSERVATIONS:




**Fig 1: Homepage of STCRDab database**

The screenshot shows the "CDR search" interface on the STCRDab website. The header includes the same navigation bar as Fig 1. The main heading is "CDR search" with the subtitle "Search CDRs based on specific criteria." Below this is a search input field with a "Search" button. To the right, there are several search options: "Get all structures.", "Search for a specific PDB.", and "2. Get the CDR loops of a particular PDB in STCRDab." The second option is selected, and a form is displayed with the label "Please enter the PDB Code:". The input field contains the text "2uwe" (highlighted with a red box). Below the input field is a "Get CDR structures" button. Further down, there are links for "Advanced search of CDR structures.", "Search for unique CDR structures.", and "About the canonical forms".

At the bottom of the page, there is a footer with the text: "Leem et al., *Nucleic Acids Res.* (2018), 46, D406-D412. © Copyright 2017. Developed by OPIG using Bootstrap and Flask."

**Fig 2: Enter the PDB ID: 2UWE in CDR search of STCRDab**

Home STCRDab STCRPred Help SABDab SABPred 

## CDR search

Search CDRs based on specific criteria.

[View results](#) >

[Downloads](#) >

[Search](#) >


1 structure(s) fit your criteria.

PDB	Species	Method	Resolution (Å)	View CDR loop Structure	View TCR Structure	Downloads
2uwe	mouse	X-RAY DIFFRACTION	2.4	<b>TCR FE:</b> CDRA1: STYSPP CDRA2: SFTDNKR CDRA3: ALFLASSFSKLV CDRB1: NNHDY CDRB2: SYADS CDRB3: ASSDWSYEQY <b>TCR ML:</b> CDRA1: STYSPP CDRA2: SFTDNKR CDRA3: ALFLASSFSKLV CDRB1: NNHDY CDRB2: SYADS CDRB3: ASSDWSYEQY	<a href="#">View Structure</a>	<ul style="list-style-type: none"> <li>• <a href="#">IMGT-numbered Structure</a></li> <li>• <a href="#">Summary file</a></li> </ul>

### Download results

- [Download the summary file](#) for this search. See [help](#) for more details on file formats.
- [Download an archived zip file](#) for this search. See [help](#) for more details on file formats.

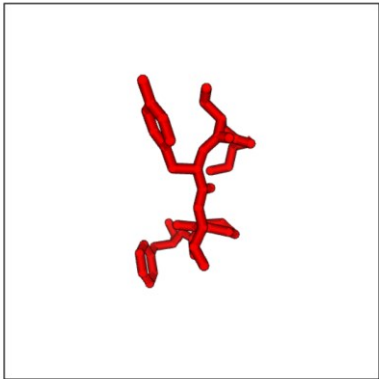
**Fig 3: Result page of PDB ID: 2UWE**

Home STCRDab STCRPred Help SABDab SABPred 

## Details for 2uwe CDRA1

Click on any of the tabs below to see the detailed information about the structure.

Structure visualisation



**Key (Default Scheme):**

IMGT-defined CDR loop

**Display options:**

[Cartoon model](#)

[Spacefill model](#)

[Wire model](#)

[Ball&stick model](#)

[Default colouring](#)

[Color by atom](#)

[Color by B-factor](#)


[Color by element](#)

Spin: on off

Show anchor residues (?):

[Get anchors](#)

**Fig 4: CDRA1 loop structure summary of 2UWE**

Home STCRDab STCRPred Help SABDab SABPred 

Basic parent structure information

Large Cdr3a Loop alteration as a Function of MHC Mutation

Item	Info
PDB	2uwe
Organism	MUS MUSCULUS
Method	X-RAY DIFFRACTION
Resolution	2.4Å

CDR loop information

IMGT Numbered sequence


A27	A28	A29	A36	A37	A38
S	T	Y	S	P	F

Comparison to antibody CDR loops

Antibody PDB	Antibody CDR loop	Antibody CDR sequence	Backbone RMSD (Å)	SABDab link to Antibody PDB
4lj6A	CDRL1	SIGSRA	1.39	<a href="#">View structure</a>
4fqL	CDRL1	SLGSRA	1.46	<a href="#">View structure</a>

Available downloads


**Fig 4.1: CDRA1 details of 2UWE (TCR F/E)**

Home STCRDab STCRPred Help SABDab SABPred 

**Structure summary for 2uwe**  
Detailed entry information.

**Details for 2uwe**  
Click on any of the tabs below to see the detailed information about the structure.

Structure visualisation



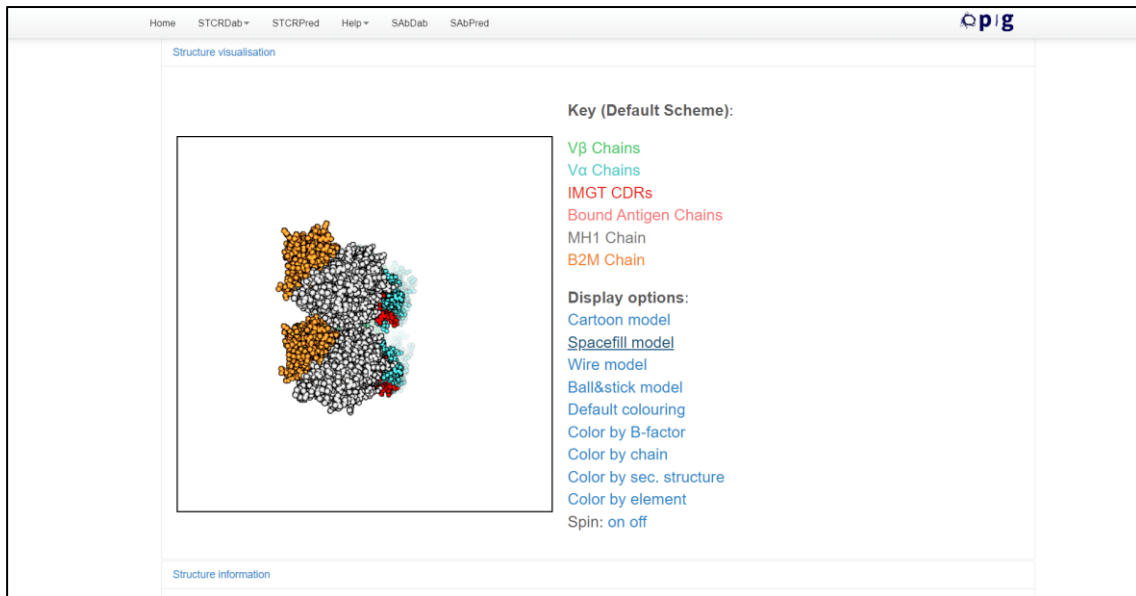
**Key (Default Scheme):**

- Vβ Chains
- Vα Chains
- IMGT CDRs
- Bound Antigen Chains
- MH1 Chain
- B2M Chain

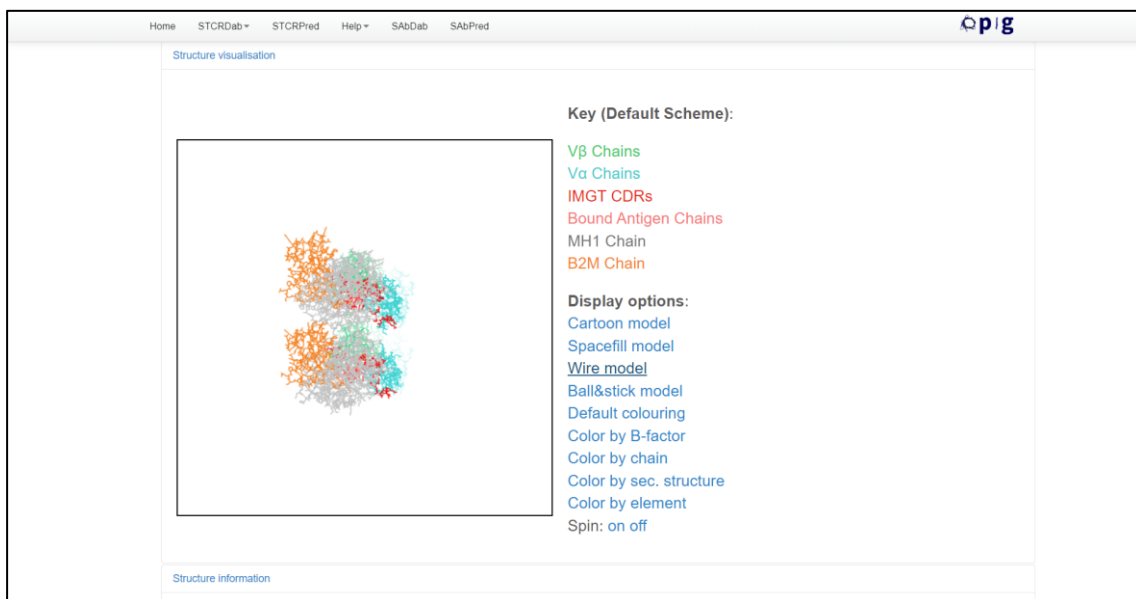
**Display options:**

- Cartoon model
- Spacefill model
- Wire model

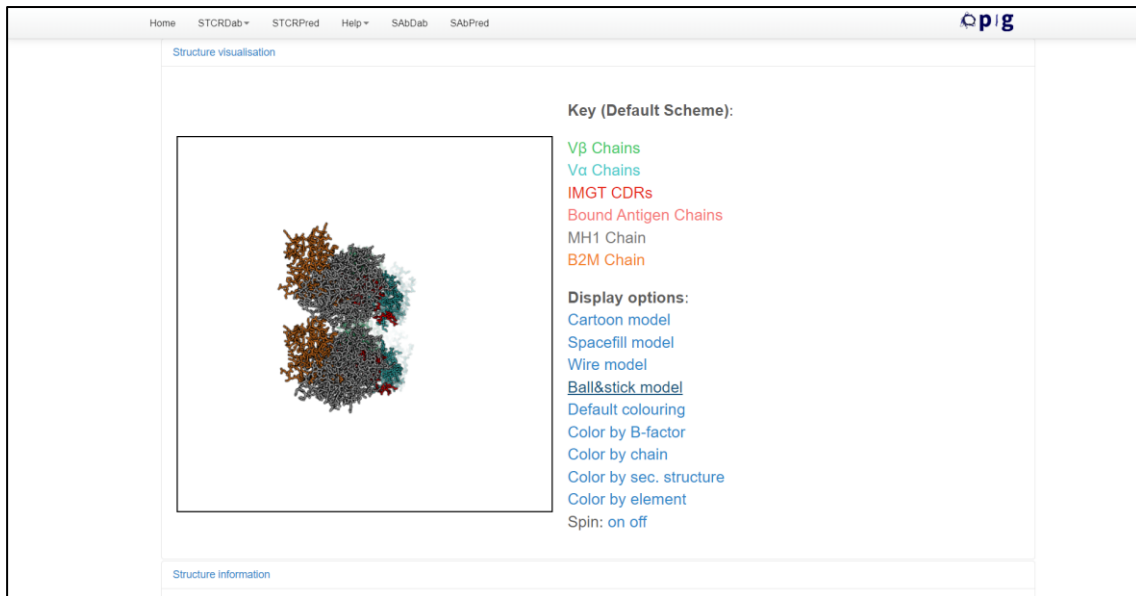
**Fig 5: Structure summary for 2UWE**



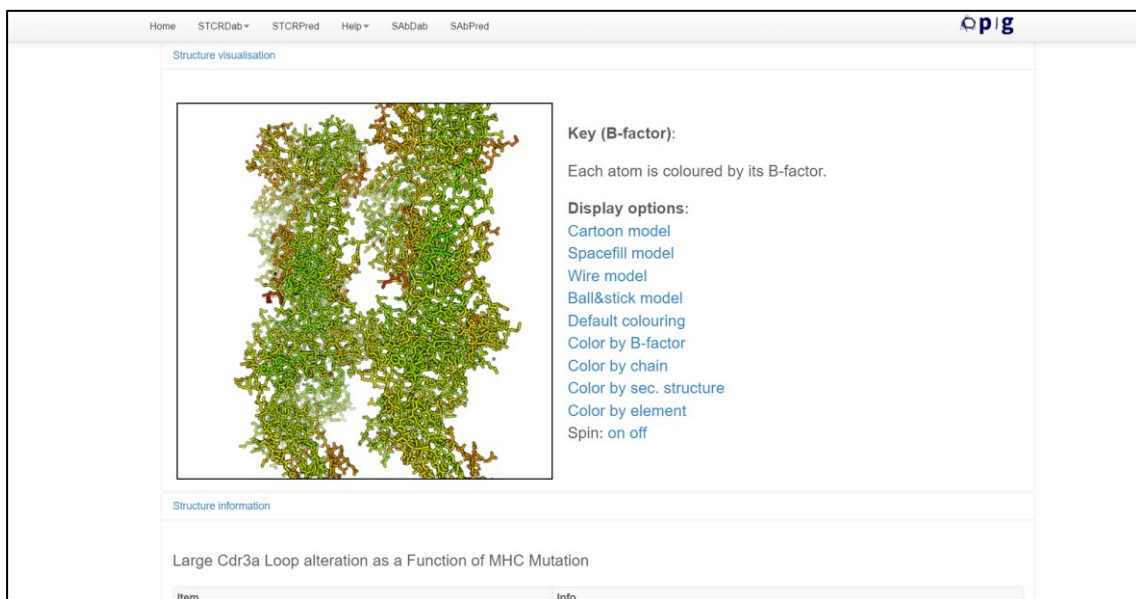
**Fig 5.1: Visualization of structure in space fill model**



**Fig 5.2: Visualization of structure in wire model**




**Fig 5.3: Visualization of structure in Ball and stick model**

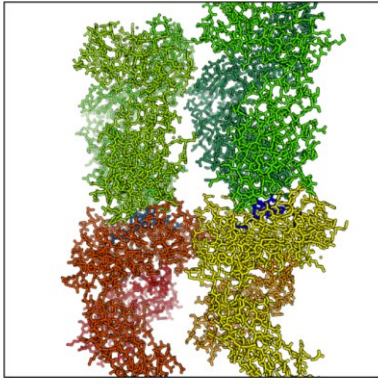


**Fig 5.4: Visualization of structure in color by B-factor**



Home STCRDab STCRPred Help SABDab SABPred 

Structure visualisation



**Key (Chain):**  
Each chain is coloured by its chain ID.


**Display options:**  
[Cartoon model](#)  
[Spacefill model](#)  
[Wire model](#)  
[Ball&stick model](#)  
[Default colouring](#)  
[Color by B-factor](#)  
[Color by chain](#)  
[Color by sec. structure](#)  
[Color by element](#)  
 Spin: on off

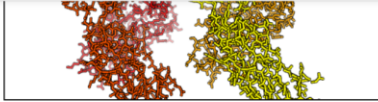
Structure information

Large Cdr3a Loop alteration as a Function of MHC Mutation

Item	Info
PDB	2uwe
Organism	MUS MUSCULUS
Method	X-RAY DIFFRACTION
Resolution	2.4Å
Number of TCRs	2

**Fig 5.5: 15 Visualization of structure in color by chain**

Home STCRDab STCRPred Help SABDab SABPred 



[Color by sec. structure](#)  
[Color by element](#)  
 Spin: on off

Structure information

Large Cdr3a Loop alteration as a Function of MHC Mutation


Item	Info
PDB	2uwe
Organism	MUS MUSCULUS
Method	X-RAY DIFFRACTION
Resolution	2.4Å
Number of TCRs	2

[Paired chains information](#)

[Available downloads](#)

Leem *et al.*, *Nucleic Acids Res.* (2018), **46**, D406-D412. © Copyright 2017. Developed by OPIG using Bootstrap and Flask.

**Fig 6: Structure information details for 2UWE**

Home STCRDab STCRPred Help SABDab SABPred 

Paired chains information

This PDB has 2 TCR(s).

[F/E](#)

**TCR Details:**

Item	Info
VB chain	F
VA chain	E
VB IMGT details	TRBV13/TRBJ2
VA IMGT details	TRAJ12/TRAJ50
Species	mouse


**Antigen Details:**

Item	Info
Antigen Chain	C
Antigen Type	Peptide
Antigen Organism	HOMO SAPIENS
Antigen Sequence	ALWGFPPVL
Antigen Length	9

**MHC details:**

Item	Info
MHC Chain	A B
MHC Type	MH1
MHC Species	human

**Fig 7: Pair chain information of F/E TCR showing TCR, Antigen details and MHC details**

Home STCRDab STCRPred Help SABDab SABPred 

**CDR Sequences:**

Loop	Sequence	Predicted canonical form	CDR Length
<a href="#">CDRB3</a>	ASSDWVSYEQY	None	11
<a href="#">CDRB2</a>	SYVADS	None	6
<a href="#">CDRB1</a>	NNHDY	None	5
<a href="#">CDRA3</a>	ALFLASSSFSKLV	A3-13-D	13
<a href="#">CDRA2</a>	SFTDNKR	None	7
<a href="#">CDRA1</a>	STYSFP	None	6

**Orientation and docking angles:**

Angle	Value
BC2	109.04°
BC1	74.17°
BA	-66.62°
AC2	74.12°
AC1	124.21°
dc	16.65Å
Docking angle	68.54°

**TCRs with similar orientations:**

TCR PDB	BC2	BC1	BA	AC2	AC1	dc	TAngle Distance
<a href="#">1p9_FE</a>	108.68°	73.78°	-66.70°	74.14°	123.92°	16.54Å	0.6
<a href="#">2j8u_FE</a>	108.31°	74.02°	-66.29°	74.42°	123.85°	16.69Å	0.9
<a href="#">2jcc_FE</a>	109.08°	74.58°	-66.28°	75.06°	124.04°	16.62Å	1.1
<a href="#">1p9_ML</a>	108.14°	73.36°	-66.36°	74.25°	124.79°	16.60Å	1.4
<a href="#">2j8u_ML</a>	107.93°	72.98°	-66.36°	73.99°	124.42°	16.81Å	1.7

**Fig 7.1: Pair chain information of M/L TCR showing CDR sequences, orientation and docking angles and TCRs with similar orientations**

## **RESULTS:**

The CDR search for the query PDB ID-2UWE was conducted in the STCRDab database to identify the CDR position for drug design purposes. STCRDab is a database of T-cell receptor (TCR) structures. The results section showed one structure, along with its species name, method, and resolution. One of the CDR loop structures, 'CDRA1,' was visualized, and both its parent structure information and CDR loop details were explored. Following this, the TCR structure of 2UWE was visualized using different display options. The paired chain information of the 2UWE structure revealed two TCRs, named F/E and M/L, displaying details such as TCR, antigen, MHC information, numbered sequences, CDR sequences, orientation and docking angles, and TCRs with similar orientations.

## **CONCLUSION:**

The retrieval of CDR positions in query 2UWE using the STCRDab database provided valuable structural insights into the antigen-binding mechanisms of T-cell receptors. This information was crucial for understanding the CDR loops, which played an essential role in drug design and therapeutic development.

## **REFERENCES:**

1. Davis, M. M., & Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181), 395–402. <https://doi.org/10.1038/334395a0>
  2. Garcia, K. C., Adams, J. J., Feng, D., & Ely, L. K. (2009). The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nature Immunology*, 10(2), 143–147. <https://doi.org/10.1038/ni.1698>
  3. Smith-Garvin, J. E., Koretzky, G. A., & Jordan, M. S. (2009). T cell activation. *Annual Review of Immunology*, 27, 591–619. <https://doi.org/10.1146/annurev.immunol.021908.132706>
-

**WEBLEM: 6**

**Introduction to Yvis Database**

**(URL: <http://bioinfo.icb.ufmg.br/yvis/>)**

**INTRODUCTION:**

Yvis Database is a modern, versatile data management platform that addresses the growing needs of today's data-centric environments. It supports a wide range of data types, including relational, non-relational, and semi-structured formats, making it ideal for applications ranging from IoT data streams and big data analytics to cloud-native solutions. One of its key strengths lies in its high performance and scalability, enabling it to handle large-scale, high-volume transactions with ease. Its distributed architecture ensures efficient processing, while resources are dynamically adjusted to maintain peak performance.

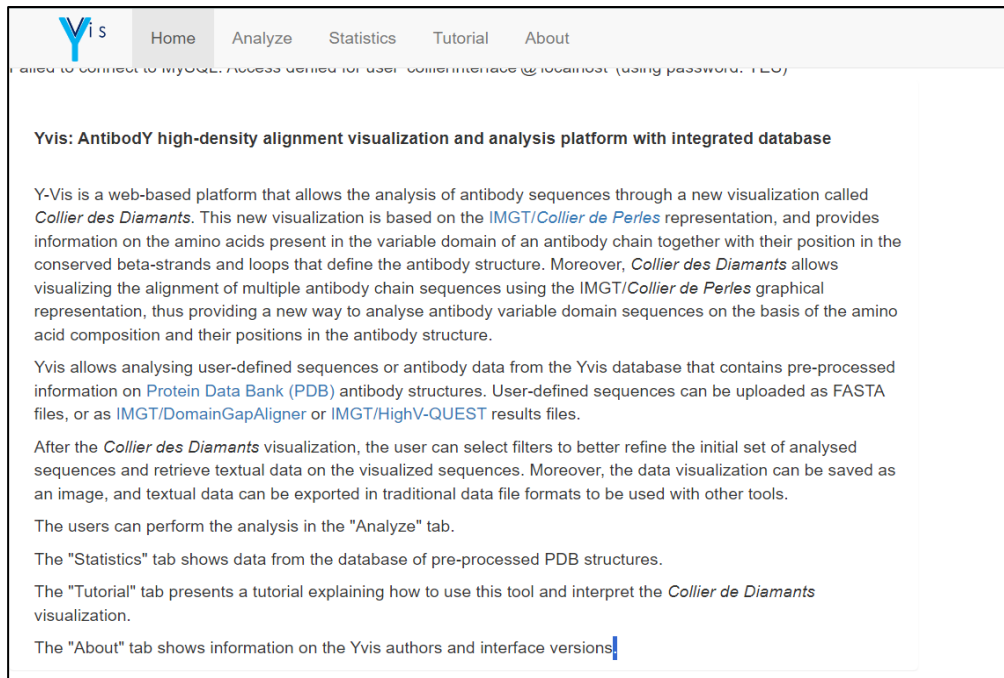
Security is a core focus for Yvis, with robust encryption, role-based access control (RBAC), and compliance with major standards such as GDPR, HIPAA, and SOC2. This ensures that sensitive data always remains secure. Yvis also integrates real-time analytics and machine learning, allowing users to process and analyze data as it's ingested and apply predictive analytics or anomaly detection without exporting data to separate systems.

The platform is designed for flexibility, supporting seamless integration with various data sources, including traditional SQL databases, cloud services, and data lakes. Continuous data ingestion from APIs and streaming services like Kafka is also easily managed. Users benefit from both an intuitive graphical interface and comprehensive API support, enabling efficient data management regardless of technical proficiency. Yvis Database is a powerful solution that combines speed, security, and flexibility for the most demanding data-driven applications.

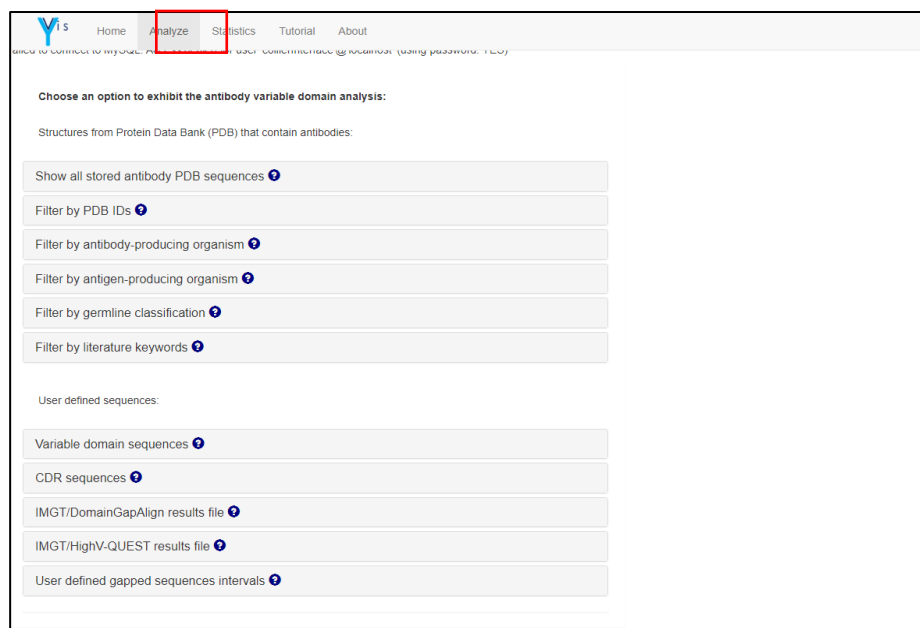
Antibodies or immunoglobulins are vertebrate immune system proteins that are produced by B cells and can bind to antigens with high specificity and affinity. For this reason, antibodies are an important tool in diagnosis, therapy, and experimental biology. To elucidate the antibody characteristics, large numbers of antibody structures and sequences have been generated in the last years. The number of antibodies or antibody fragment structures deposited in Protein Data Bank (PDB) has increased exponentially, leading to the development of databases of antibody structures. Moreover, many antibody sequences have been obtained by high-throughput sequencing of the B-cell receptor repertoire. This extraordinary and still increasing number of antibody structures and sequences demands integrative data organization and tools for their analysis, comparison, and visualization. One of the major bottlenecks in this field is the concomitant visualization of a large amount of antibody data. AbYsis and IMGT/3Dstructure-DB allow antibody visualization, but only a limited number of sequences can be analyzed at a time. Indeed, abYsis presents a classical multiple sequence alignment (MSA) that displays a limited number of sequences and positions each time. IMGT/3Dstructure-DB display only one antibody sequence using the IMGT/Collier des Perles representation that allows sequence analysis related to the antibody structure. To fill this gap, we developed the antibody high-density alignment visualization and analysis (Yvis) platform that includes:

1. an updated weekly and curated antibody structure database (Yvis database)

- integrated antibody analysis resources, such as an antibody high-density alignment visualization called *Collier de Diamants*, and multiple filter options to analyze data from user files or from the Yvis database.

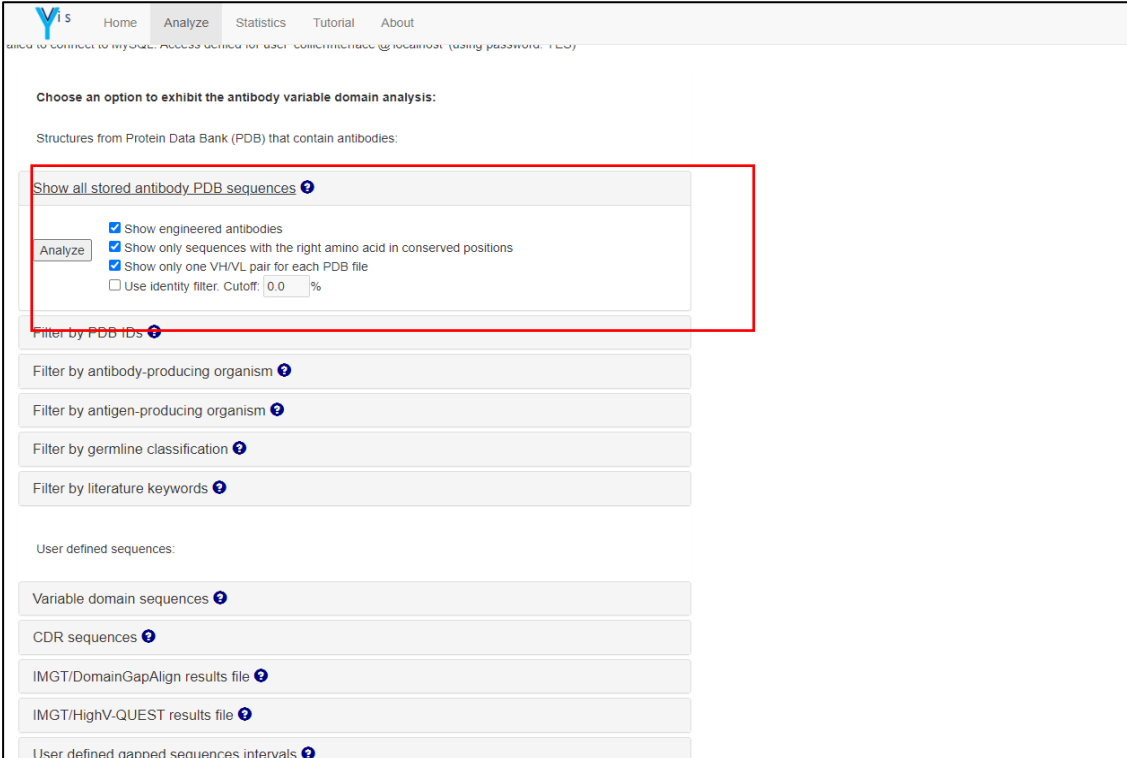


**Fig 1: Homepage of Yvis Database**



**Fig 2: Analysis options in Yvis Database**

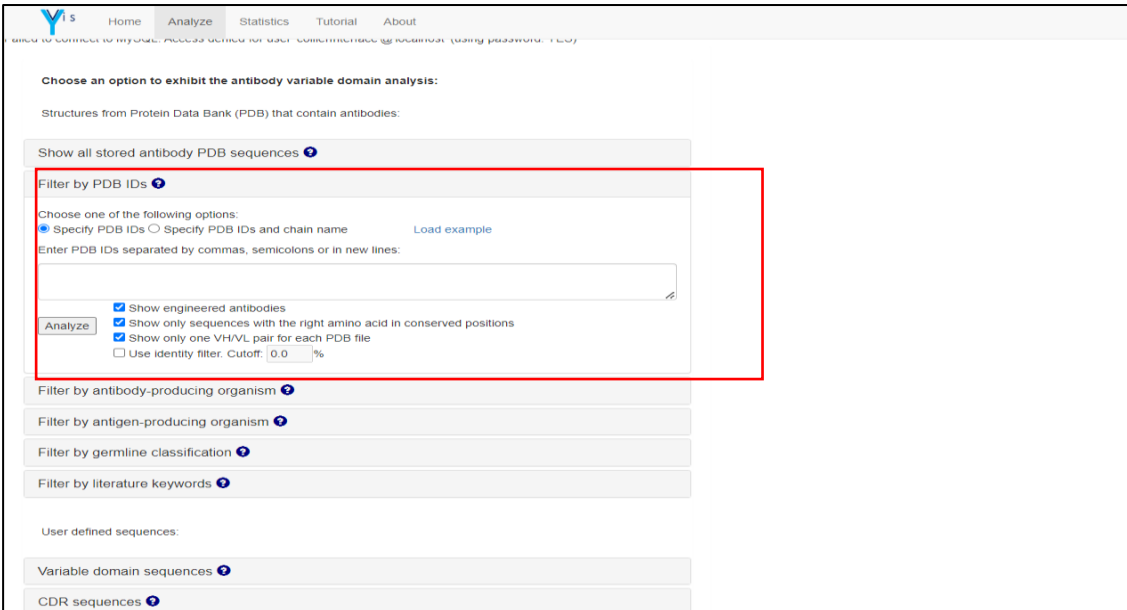
## 1. Structures from Protein Data Bank (PDB) that contain antibodies:



The screenshot shows the Yvis web interface. At the top, there is a navigation bar with 'Home', 'Analyze', 'Statistics', 'Tutorial', and 'About'. Below the navigation bar, there is a header section with the text 'Choose an option to exhibit the antibody variable domain analysis:'. Underneath, there is a section titled 'Structures from Protein Data Bank (PDB) that contain antibodies:'. A red box highlights the 'Show all stored antibody PDB sequences' button, which is selected. Below this button, there is an 'Analyze' button and several checkboxes: 'Show engineered antibodies' (checked), 'Show only sequences with the right amino acid in conserved positions' (checked), 'Show only one VH/VL pair for each PDB file' (checked), and 'Use identity filter. Cutoff: 0.0 %' (unchecked). Below the 'Analyze' button, there are several filter buttons: 'Filter by PDB IDs', 'Filter by antibody-producing organism', 'Filter by antigen-producing organism', 'Filter by germline classification', and 'Filter by literature keywords'. At the bottom, there is a section titled 'User defined sequences:' with buttons for 'Variable domain sequences', 'CDR sequences', 'IMGT/DomainGapAlign results file', 'IMGT/HighV-QUEST results file', and 'User defined gapped sequences intervals'.

**Fig 3: Show all Stored antibody PDB sequences**

Select this option to show information on all antibody sequences from PDB and stored in Yvis database.



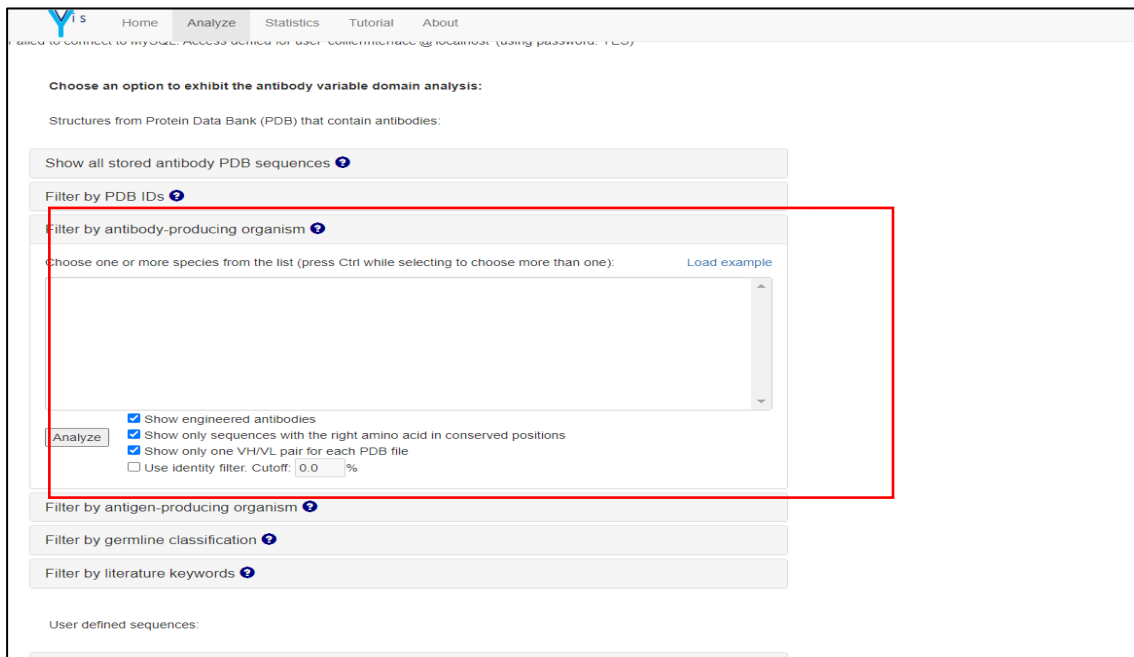
The screenshot shows the Yvis web interface. At the top, there is a navigation bar with 'Home', 'Analyze', 'Statistics', 'Tutorial', and 'About'. Below the navigation bar, there is a header section with the text 'Choose an option to exhibit the antibody variable domain analysis:'. Underneath, there is a section titled 'Structures from Protein Data Bank (PDB) that contain antibodies:'. A red box highlights the 'Filter by PDB IDs' button, which is selected. Below this button, there is a section titled 'Choose one of the following options:' with two radio buttons: 'Specify PDB IDs' (selected) and 'Specify PDB IDs and chain name'. There is a 'Load example' link next to the second option. Below this, there is a text input field with the placeholder text 'Enter PDB IDs separated by commas, semicolons or in new lines:'. Below the input field, there is an 'Analyze' button and several checkboxes: 'Show engineered antibodies' (checked), 'Show only sequences with the right amino acid in conserved positions' (checked), 'Show only one VH/VL pair for each PDB file' (checked), and 'Use identity filter. Cutoff: 0.0 %' (unchecked). Below the 'Analyze' button, there are several filter buttons: 'Filter by antibody-producing organism', 'Filter by antigen-producing organism', 'Filter by germline classification', and 'Filter by literature keywords'. At the bottom, there is a section titled 'User defined sequences:' with buttons for 'Variable domain sequences' and 'CDR sequences'.

**Fig 4: Filter by PDB IDs.**

Select this option to show only chains from structures of a user-defined list of PDB identifiers, with or without chain specification.

You can specify a list of PDB IDs by selecting the "Specify PDB IDs" option and inserting in the textbox the PDB IDs separated by commas, semicolons, or by putting each ID in a new line. In this case, Yvis will show the chains stored in the Yvis database that are part of the indicated structures.

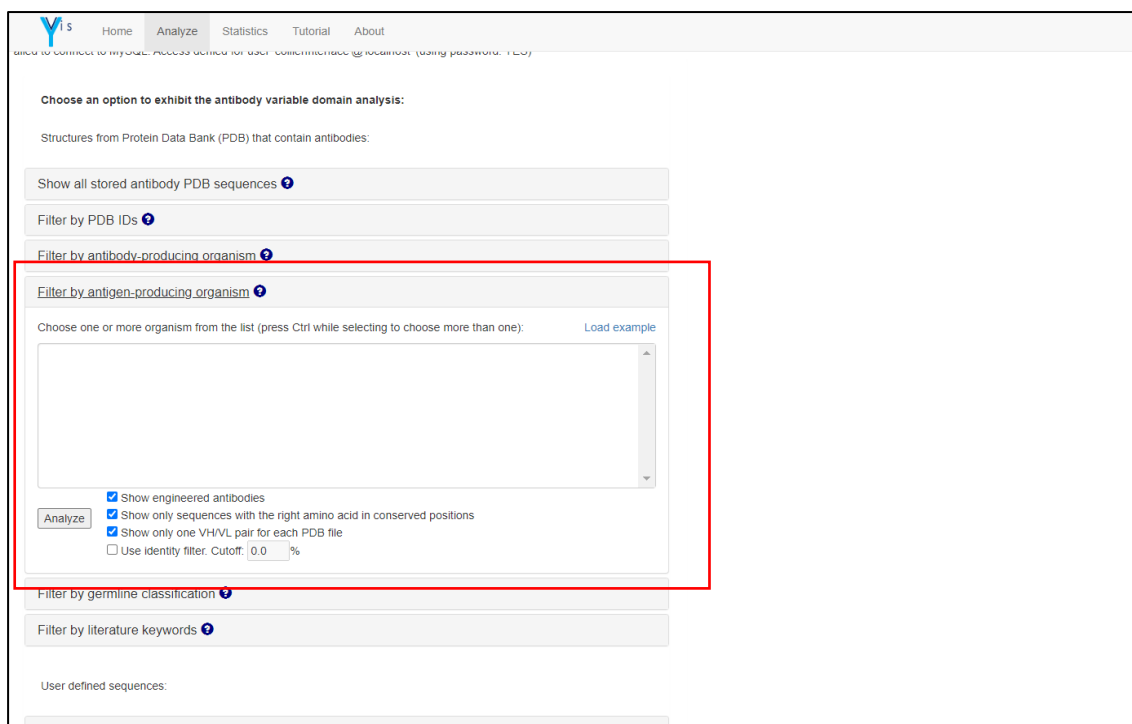
If you want to restrict the analysis to specific chains, you should select the "Specify PDB IDs and chain name" option and insert in the textbox a list of chains separated by commas, semicolons, or in new lines. Each chain must be specified by the PDB ID followed by a colon and the chain name.



**Fig 5: Filter by antibody producing organism.**

Select this option to show only chains of antibodies produced by specific organisms. Choose one or more species from the list of all antibody-producing organisms stored in the database, standardized by species, based on the UniProt Taxonomy Database.





**Fig 6: Filter by antigen producing organism.**

Select this option to show only chains from antibody structures that presents an antibody-antigen complex, or choose the "None" option to show only chains of antibodies that are not in complex with antigens. The list presents non-protein antigens type (carbohydrate, hapten, and nucleic acid) and, for proteins or peptide antigens, the antigen-producing organisms stored in the database, standardized by species, based on the UniProt Taxonomy Database. Choose one or more items from this list.

Y1S Home Analyze Statistics Tutorial About

Choose an option to exhibit the antibody variable domain analysis:

Structures from Protein Data Bank (PDB) that contain antibodies:

Show all stored antibody PDB sequences

Filter by PDB IDs

Filter by antibody-producing organism

Filter by antigen-producing organism

**Filter by germline classification**

Choose one or more options from lists (press Ctrl while selecting to choose more than one): [Load example](#)

Assigned organism: V gene allele name: J gene allele name:

Show engineered antibodies

Show only sequences with the right amino acid in conserved positions

Use identity filter. Cutoff: 0.0 %

Analyze

Filter by literature keywords

User defined sequences:

Variable domain sequences

**Fig 7: Filter by germline classification**

Select this option to show only chains assigned to specific germline alleles by IMGT/DomainGapAlign. You can restrict the assigned species and V or J alleles by choosing one or more options from the lists. If you do not want to restrict the analysis, select all options.

Y1S Home Analyze Statistics Tutorial About

Filter by antigen-producing organism

Filter by germline classification

**Filter by literature keywords**

Enter the expressions to search for in each field of literature information. The operators AND, OR and NOT can be used in any fields. [Load example](#)

Title keywords:

Summary keywords:

Authors:

Publication/Year:

Article identifier (DOI, PMID or PMCID):

Show engineered antibodies

Show only sequences with the right amino acid in conserved positions

Show only one VH/VL pair for each PDB file

Use identity filter. Cutoff: 0.0 %

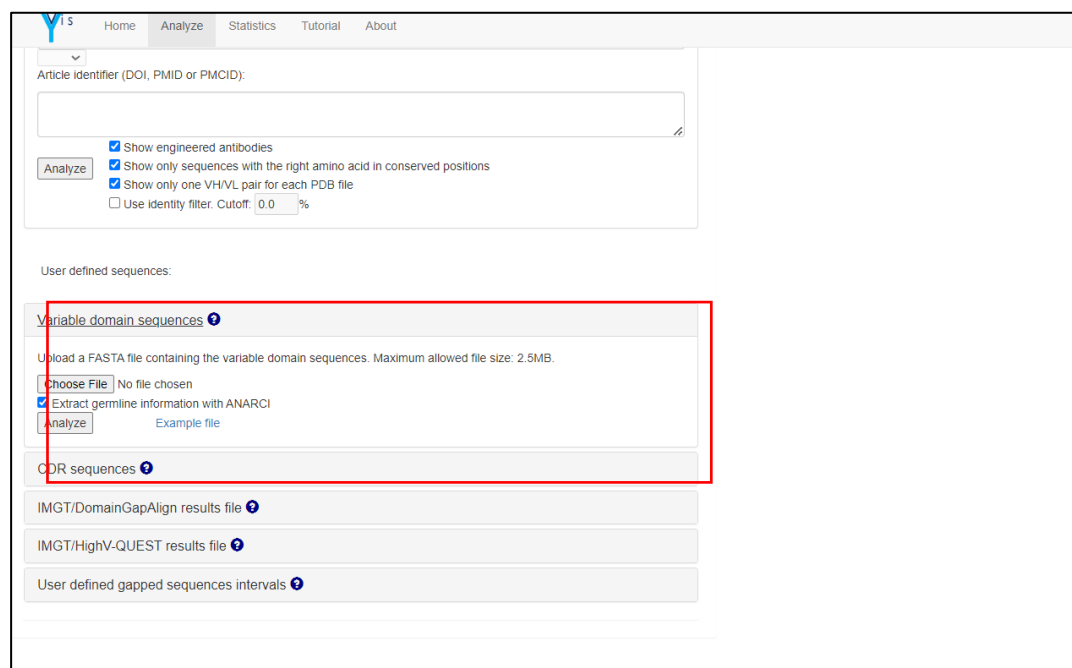
Analyze

User defined sequences:

## Fig 8: Filter by literature Keywords

Select this option to show only antibody chains from PDB structures filtered based on literature information. You can specify paper title or summary, authors' names, publication year, or article identifier (DOI, PMID or PMCID). These fields accept multiple keywords and can be defined with Boolean operators (AND, OR and NOT).

### 2. User Defined sequences:



## Fig 9: Variable domain sequences.

Select this option to insert a FASTA file that contains amino acid sequences of variable domains of antibody chains. The Yvis server uses ANARCI to gap sequences.

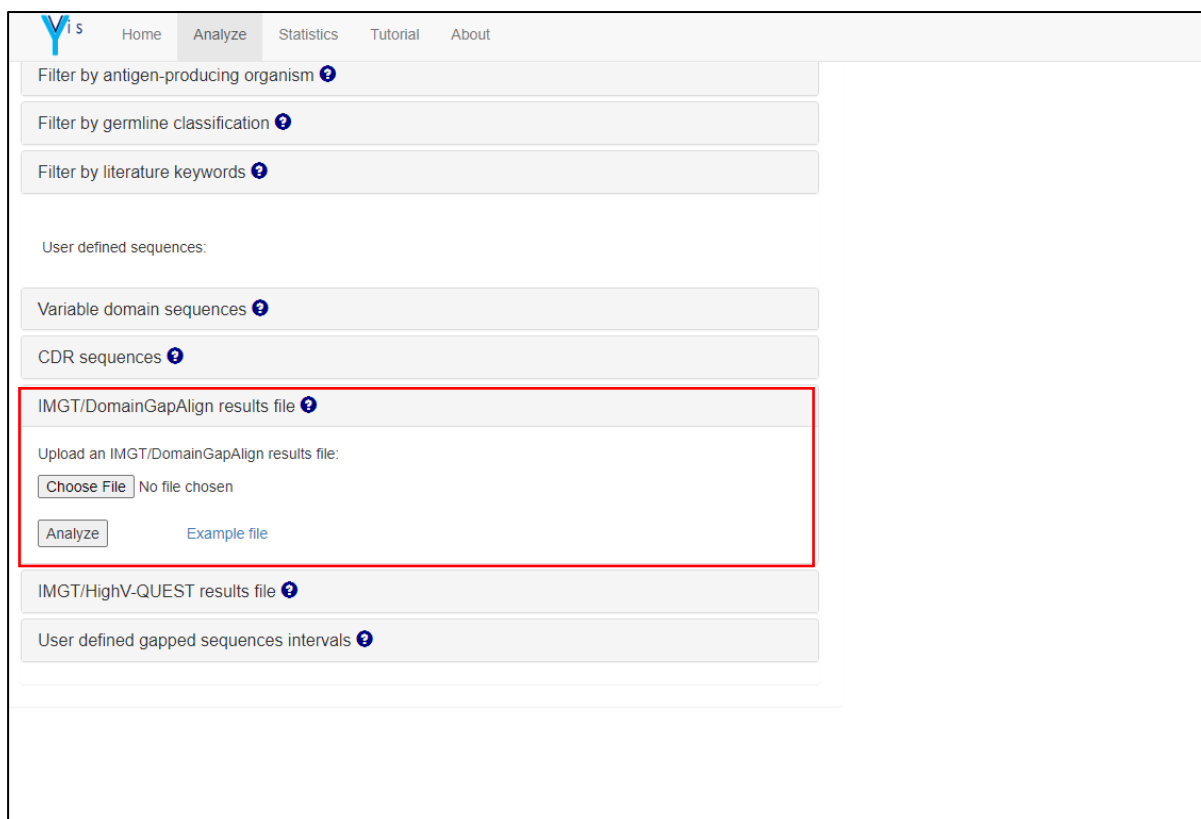
In sequence name and comments line (line starting with ">"), the user can insert the following information, separated by "|": PDB/identification source, chain identification, chain type (it is overwritten if ANARCI finds a different type), antibody-producing organism, engineered antibody information (engineered or not), antigen-producing organism, antigen molecule description, assigned germline species, V gene, percentage of V gene identity, J gene, and percentage of J gene identity. These are suggested information and could be used in analysis filters; however, only the sequence identification is mandatory (and will be used in the PDB identification field). Optionally, you can select the option "Extract germline information with ANARCI" to obtain germline information from ANARCI instead of getting them from the user's file.

As the time of execution of this analysis is in function of the number of uploaded/chosen sequences, in the case of large files this time will be long (the user's browser might present a slow script dialog). In the case of a huge number of sequences, users are invited first to submit them to IMGT/DomainGapAlign and then upload the results page into Yvis, using the "IMGT/DomainGapAligner results file" input option.

The screenshot shows the Yvis web interface. At the top, there is a navigation menu with 'Home', 'Analyze', 'Statistics', 'Tutorial', and 'About'. Below the menu, there is a section for 'Article Identifier (DOI, PMID or PMCID)' with a text input field and an 'Analyze' button. To the right of the input field are four checkboxes: 'Show engineered antibodies' (checked), 'Show only sequences with the right amino acid in conserved positions' (checked), 'Show only one VH/VL pair for each PDB file' (checked), and 'Use identity filter. Cutoff: 0.0 %' (unchecked). Below this is a section for 'User defined sequences' with a 'Variable domain sequences' button. The 'CDR sequences' section is highlighted with a red box and contains three radio buttons: 'CDR1 (Max 12 aa)', 'CDR2 (Max 10 aa)', and 'CDR3 (unlimited)'. Below the radio buttons is the text 'Upload a fasta file containing the CDR sequences.' and a 'Choose File' button with the text 'No file chosen'. There is also an 'Analyze' button and an 'Example file' link. Below the 'CDR sequences' section are three buttons: 'IMGT/DomainGapAlign results file', 'IMGT/HighV-QUEST results file', and 'User defined gapped sequences intervals'.

**Fig 10: CDR sequences.**

Select this option to insert a FASTA file containing complementarity-determining region (CDR) amino acid sequences. Choose the type of CDR sequences (CDR1, CDR2, or CDR3; heavy and light chain are treated in the same way). The sequence length must be at most equal to the number of amino acids indicated in each CDR. The Yvis platform will gap sequences according to the chosen CDR. In sequence name and comments line (line starting with ">"), the user can insert the following information, separated by "|": PDB/identification source, chain identification, chain type (H or L, otherwise the information will be ignored), antibody-producing organism, engineered antibody information (engineered or not), antigen-producing organism, antigen molecule description, assigned germline species, V gene, percentage of V gene identity, J gene, and percentage of J gene identity. These are suggested data and could be used in analysis filters; however, only the sequence identification is mandatory (and will be used in the PDB identification field).



**Fig 11: IMGT/DomainGapAlign results file.**

Select this option to insert an IMGT/DomainGapAlign results file. IMGT/Domain Gap Align allows aligning amino acid sequences, gapping uploaded sequences, and indicating the closest germline V and J genes. When uploading sequences into IMGT/DomainGapAlign, it is recommended to choose 1 as input in the "Displayed alignments" option, because all displayed alignment sequences will be analyzed by Yvis, even if there are multiple alignments of the same sequence. After submitting sequences to IMGT/DomainGapAlign, save the webpage that presents the results in your computer (HTML file: .htm or .html extension). Then submit this file to Yvis.

Yvis will process the submitted file on the user's web browser, extracting the chain identification, chain type, and antibody numbering and germline information. As IMGT/DomainGapAlign ignores the additional information passed on the sequence headers from the FASTA file, some information will be missing in the data table (e.g., engineered, antigen and antibody species, and molecule description).

The screenshot shows the Yvis web interface. At the top, there is a navigation menu with 'Home', 'Analyze', 'Statistics', 'Tutorial', and 'About'. Below the menu are several filter options: 'Filter by antigen-producing organism', 'Filter by germline classification', and 'Filter by literature keywords'. Under the heading 'User defined sequences:', there are four more options: 'Variable domain sequences', 'CDR sequences', 'IMGT/DomainGapAlign results file', and 'IMGT/HighV-QUEST results file'. The 'IMGT/HighV-QUEST results file' option is highlighted with a red rectangular box. Below this option, there is a text prompt: 'Upload an IMGT/HighV-QUEST gapped AA results file:'. This is followed by a 'Choose File' button and the text 'No file chosen'. Below that is an 'Analyze' button and a blue link labeled 'Example file'. At the bottom of the visible interface, there is another option: 'User defined gapped sequences intervals'.

**Fig 12: IMGT/HighV-QUEST results file.**

Select this option to insert an IMGT/HighV-QUEST results file. IMGT/HighV-QUEST analyses next-generation sequencing (NGS) data on antigen receptors. Users must submit a FASTA file containing the nucleotide sequences to IMGT/HighV-QUEST. This tool will generate a set of files that can be downloaded as a compressed file. After decompressing the file, submit the gapped amino acid file, identified as “4\_IMGT-gapped-AA-sequences.txt” to Yvis. This file has a header row followed by several antibody chain rows. Each row has the following fields, as described in IMGT/V-QUEST Documentation, separated by tabs:

Yvis will present the *Collier de Diamants* visualization of sequences that are marked as productive in “V-DOMAIN Functionality” and do not have ambiguous amino acids.

As the time of execution of this analysis is in function of the number of inputted sequences, users should be patient in the case of large files, even if their browser presents a slow script dialog.

Yvis 1.5 Home Analyze Statistics Tutorial About

Filter by antigen-producing organism ?

Filter by germline classification ?

Filter by literature keywords ?

User defined sequences:

Variable domain sequences ?

CDR sequences ?

IMGT/DomainGapAlign results file ?

IMGT/HighV-QUEST results file ?

**User defined gapped sequences intervals ?**

Define the positions of variable domain gapped sequences intervals indicating the number of CDR3 insertions and the first and last represented positions (1-128):

CDR3 insertions number:

First position:

Last position:

Upload a FASTA file containing the variable domain gapped sequences:

No file chosen

**Fig 13: User defined gapped sequences intervals.**

Select this option to upload a FASTA file containing gapped amino acid sequences of variable domains of antibodies chains. As Yvis will not change the sequences, they must be aligned in the submitted file. If the uploaded sequences have CDR3 insertions, the user must indicate the number of insertions in the corresponding field. It is also possible to insert a sequence of only one part of the variable domain. In this case, the first and last positions in the corresponding fields must be changed.

In sequence name and comments line (line starting with ">"), the user can insert the following information, separated by "|": PDB/source identification, chain identification, chain type (H or L, otherwise the information will be ignored), antibody-producing organism, engineered antibody information (engineered or not), antigen-producing organism, antigen molecule description, assigned germline species, V gene, percentage of V gene identity, J gene, and percentage of J gene identity. These are suggested information and could be used in analysis filters. However, only the sequence identification is mandatory (and will be used in the PDB identification field).



## **REFERENCES:**

1. Yvis: Antibody high-density alignment visualization and analysis platform with integrated database. (n.d.). <http://bioinfo.icb.ufmg.br/yvis/>
  2. Carvalho, M., Molina, F., & Felicori, L. L. (2019). Yvis: antibody high-density alignment visualization and analysis platform with an integrated database. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkz387>
  3. YVIS - Database Commons. (n.d.). <https://ngdc.cncb.ac.cn/databasecommons/database/id/6818>
-

**DATE: 25/09/2024**

**WEBLEM: 6(A)**

**Yvis: Antibody high-density alignment visualization and analysis platform  
with integrated database**

**(URL: <http://bioinfo.icb.ufmg.br/yvis/>)**

**AIM:**

To study the variable and constant domain along with the Topology Diagram using Yvis Platform.

**INTRODUCTION:**

Yvis: Antibody high-density alignment visualization and analysis platform with integrated database. Yvis is a web-based platform designed for the analysis of antibody sequences through a novel visualization method called Collier des Diamants, an adaptation of the IMGT/Collier de Perles representation. This platform enables users to examine amino acids in the variable domains of antibody chains, aligned with their structural positions in conserved beta-strands and loops. Yvis facilitates the analysis of multiple antibody chain sequences using this graphical representation, offering insights into their composition and structural arrangement. Users can upload sequences or use pre-processed data from the Yvis database, and the platform supports exportable visual and textual data for further analysis.

The users can perform the analysis in the "Analyze" tab.

The "Statistics" tab shows data from the database of pre-processed PDB structures.

The "Tutorial" tab presents a tutorial explaining how to use this tool and interpret the Collier de Diamants visualization.

The "About" tab shows information on the Yvis authors and interface versions.

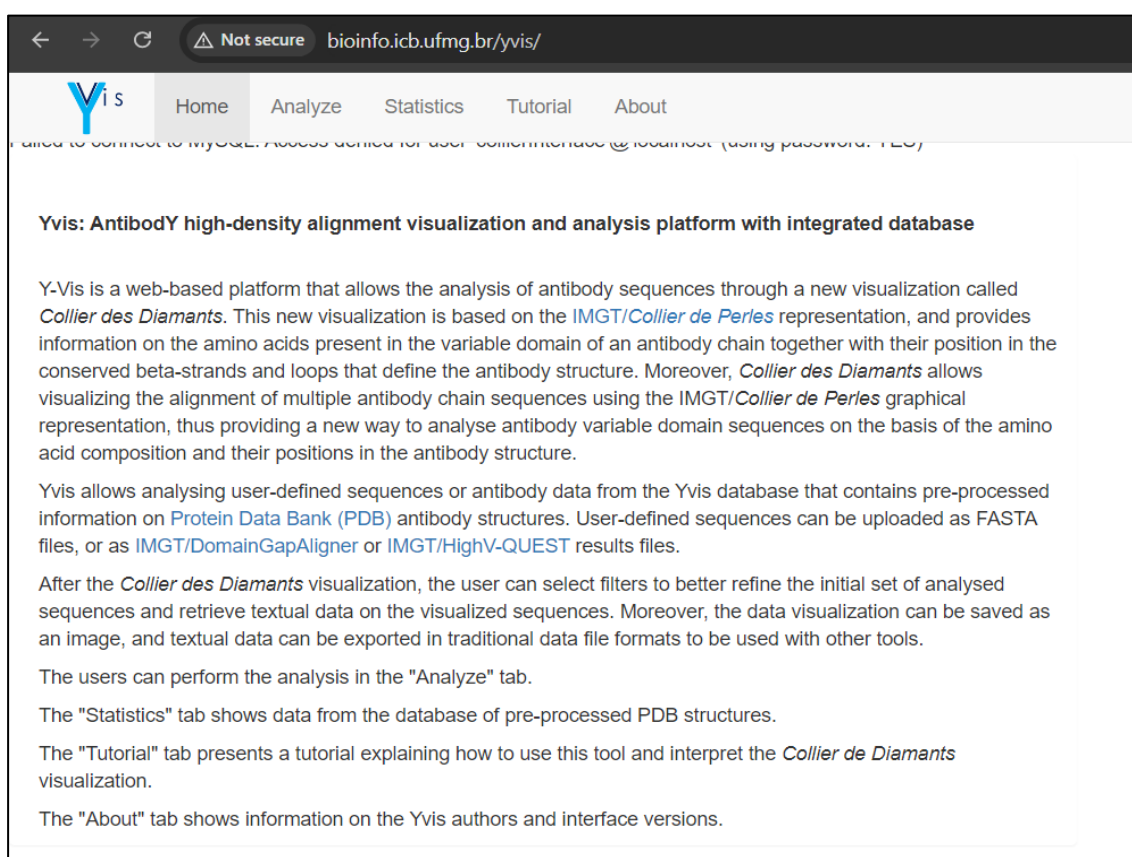
**Zika virus (ZIKV):**

Several emerging and re-emerging infections have taken a heavy toll on the public health around the globe. ZIKV was first identified, almost 70 years ago, in rhesus monkeys during a yellow fever surveillance in the Zika Forest in Uganda and was initially reported in humans in 1952. ZIKV is one of the re-emerging arboviruses (arthropod borne) which is transmitted by *Aedes* mosquito. It is a single-stranded RNA virus belonging to the genus *Flavivirus* of the *Flaviviridae* family and has been related to the other Flaviviruses including yellow fever virus, dengue virus (DENV), chikungunya virus, and West Nile virus. ZIKV virus belongs to two phylogenetic types: Asian and African. ZIKV in Africa is maintained in a life cycle (sylvatic transmission) that mainly includes monkeys and apes with humans as occasional hosts, but on the other hand, the Asian lineage of ZIKV includes humans as the main host. In most people, infection by the Zika virus is mild and self-limiting. Diseases caused by Zika virus are predominately arboviral and transmitted by the bite of female *Aedes aegypti* and *Aedes albopictus* mosquitoes. Besides a mosquito bite, the virus can also be transmitted sexually.

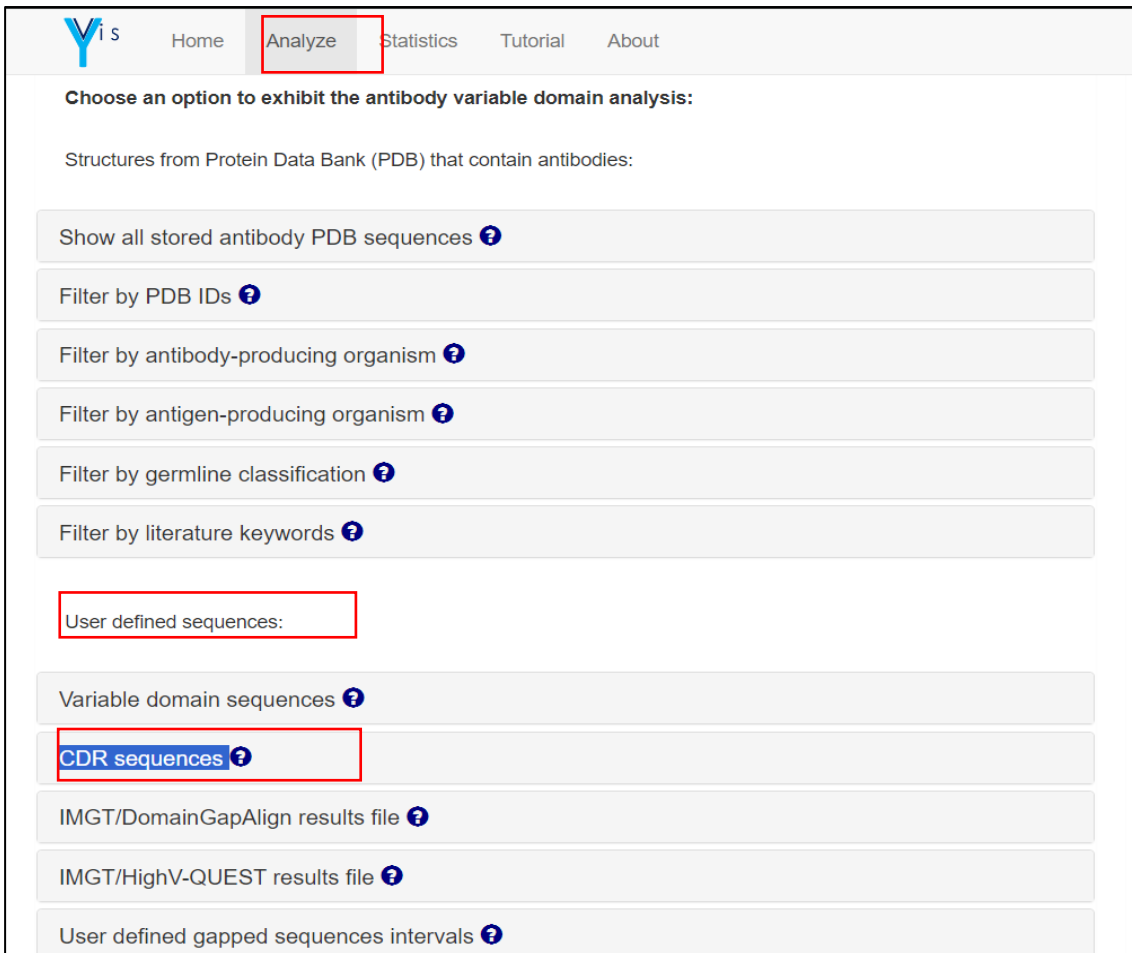
## **METHODOLOGY:**

1. Open a web browser and navigate to the homepage of the YVis Database.
2. Select 'Analyze' Option from the available menu on the Homepage.
3. Select Input Options -Select User-defined sequences, followed by choosing the 'CDR sequences' option.
4. (To proceed with the analysis, download a sample FASTA file containing complementarity-determining region (CDR) sequences.)
5. Click on 'Example file' and download the file containing 68 CDR3 sequences of anti-Zika virus antibody heavy chains isolated from four infected donors.
6. Click on the 'Choose File' button and select the downloaded FASTA file containing the CDR sequences.
7. Choose the 'CDR3' option to analyze the complementarity-determining region 3 (CDR3).
8. Click on the 'Analyze' button to initiate the analysis of the antibody variable domain sequences.

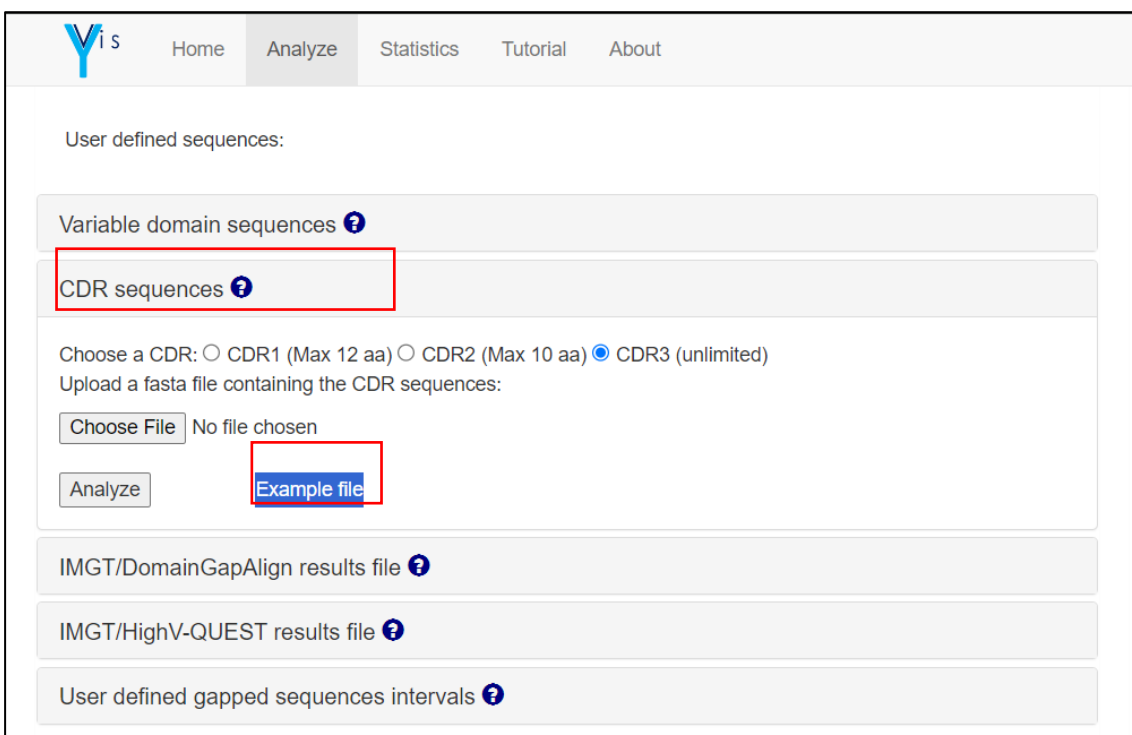
## **OBSERVATIONS:**



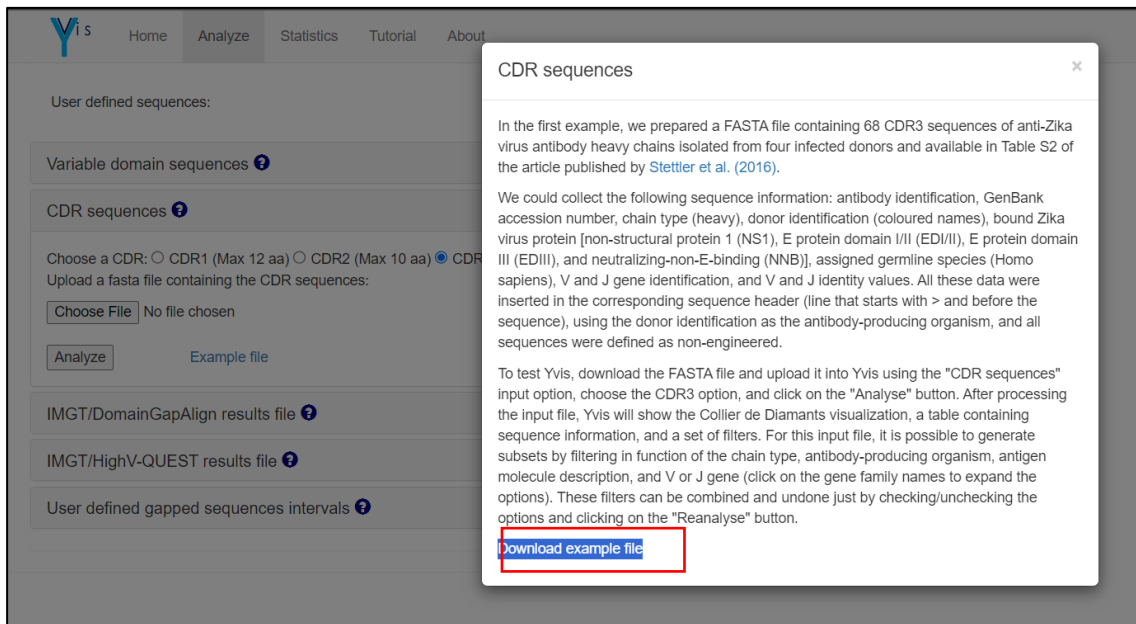
**Fig 1: Homepage Yvis: Antibody high-density alignment visualization and analysis platform with integrated Databases**



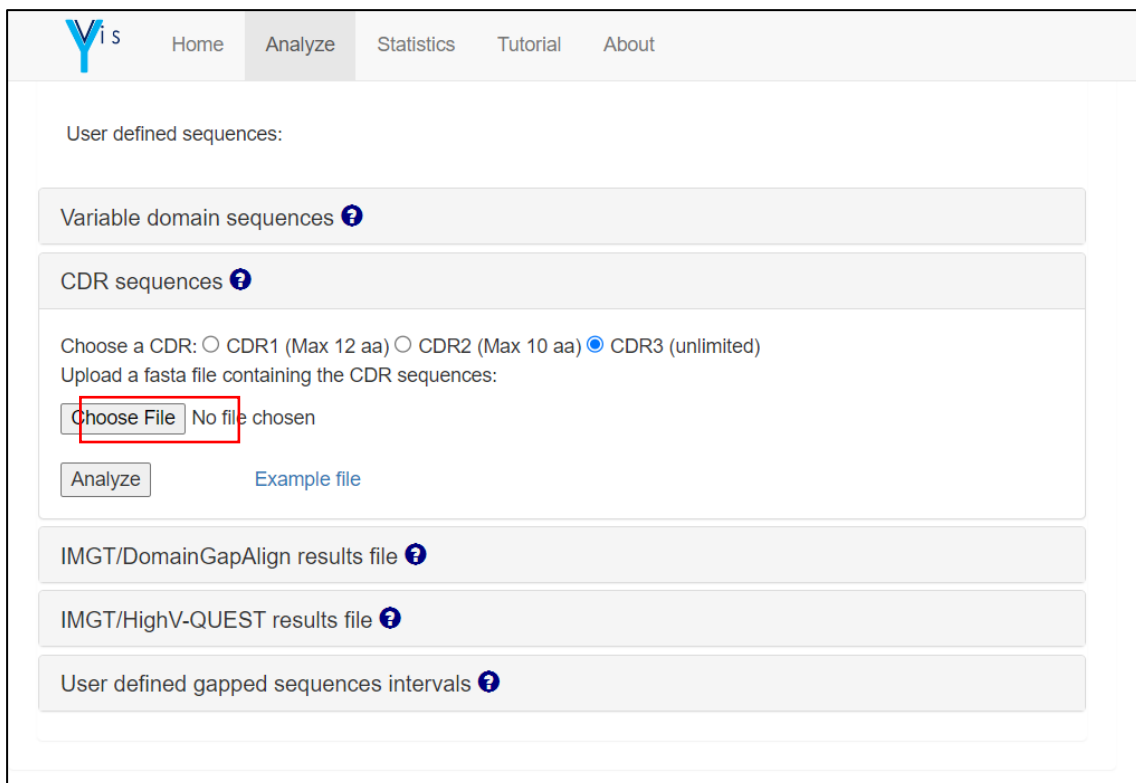
**Fig 2: Selecting the ‘Analyze’ option in the YVis portal and selecting the ‘CDR sequences’ option**



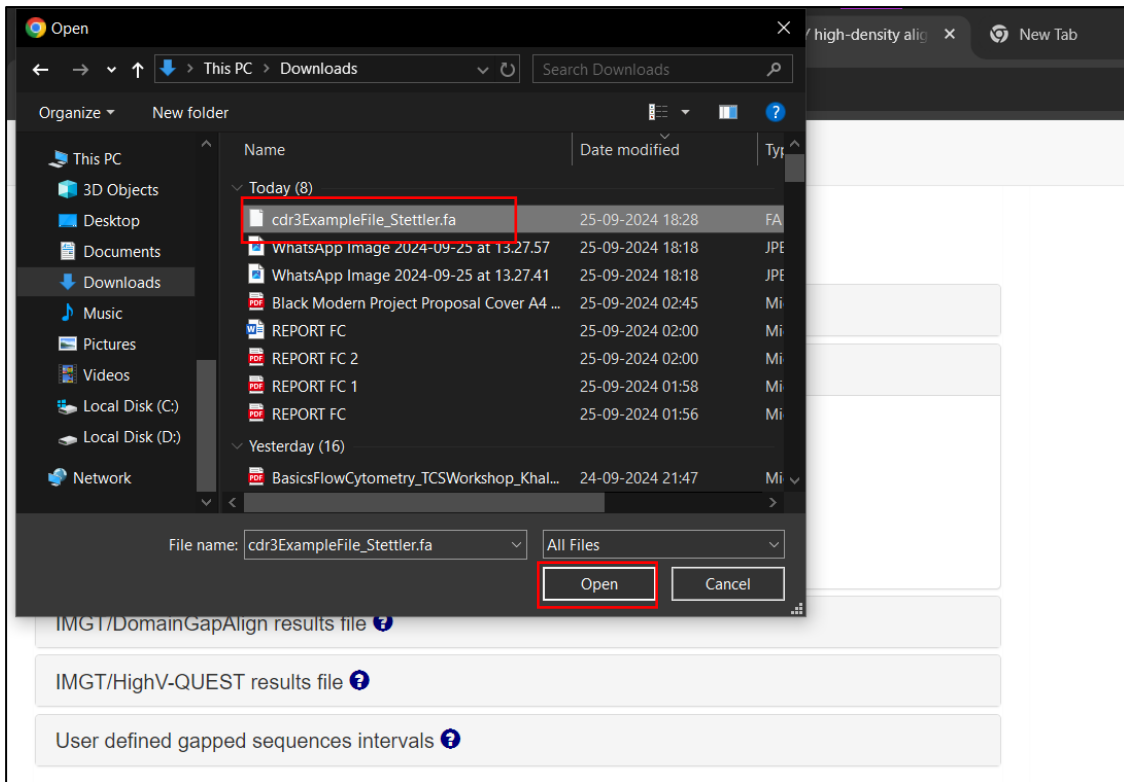
**Fig 3: Selecting the Example file**



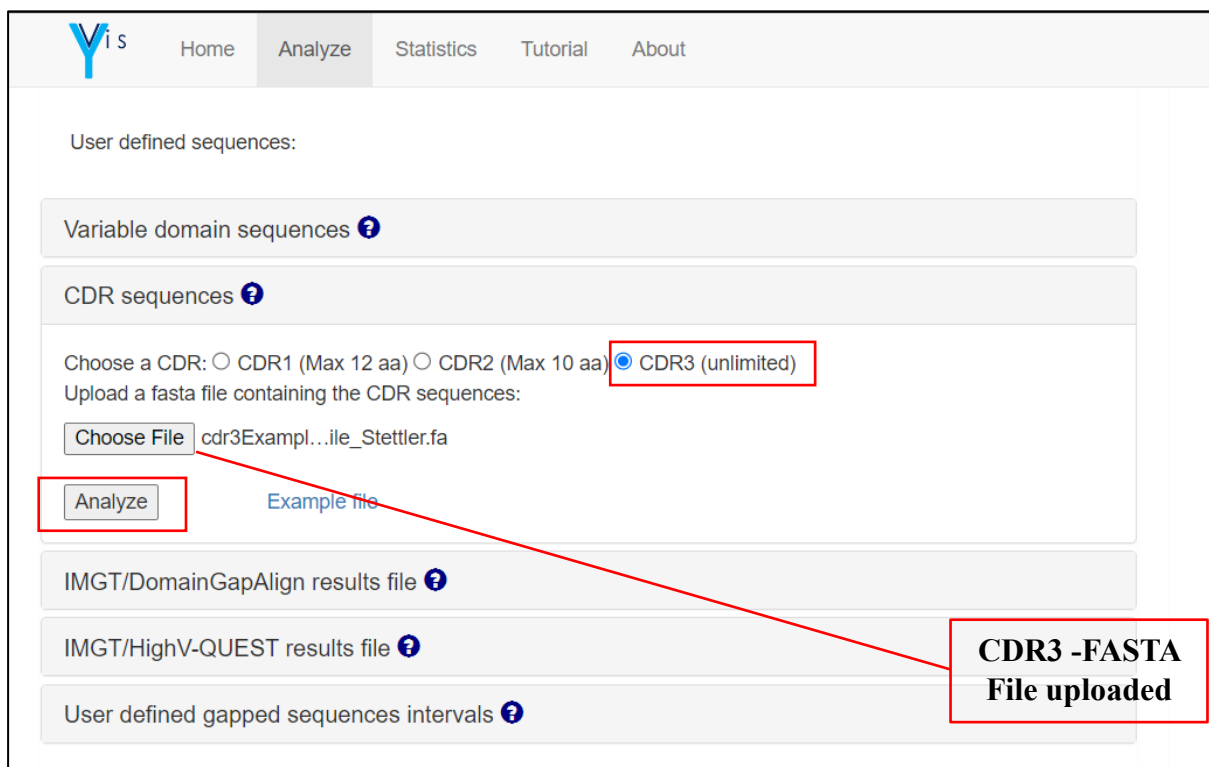
**Fig 4: Click on Download Example File Option**



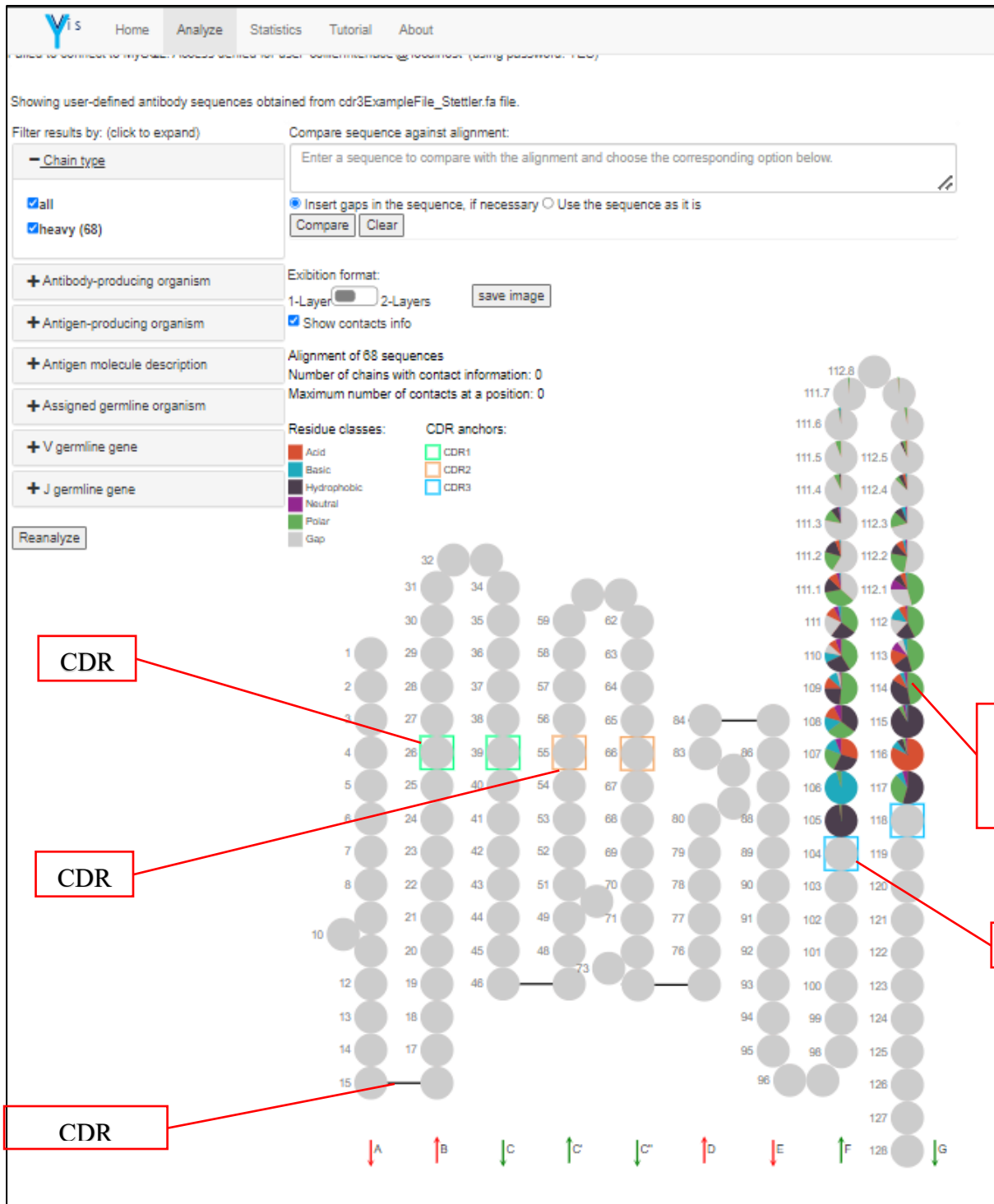
**Fig 5: Click on Choose File Option**



**Fig 6: Select the FASTA File Downloaded and Click on Open**



**Fig 7: Choose the option 'CDR3(Unlimited)'**

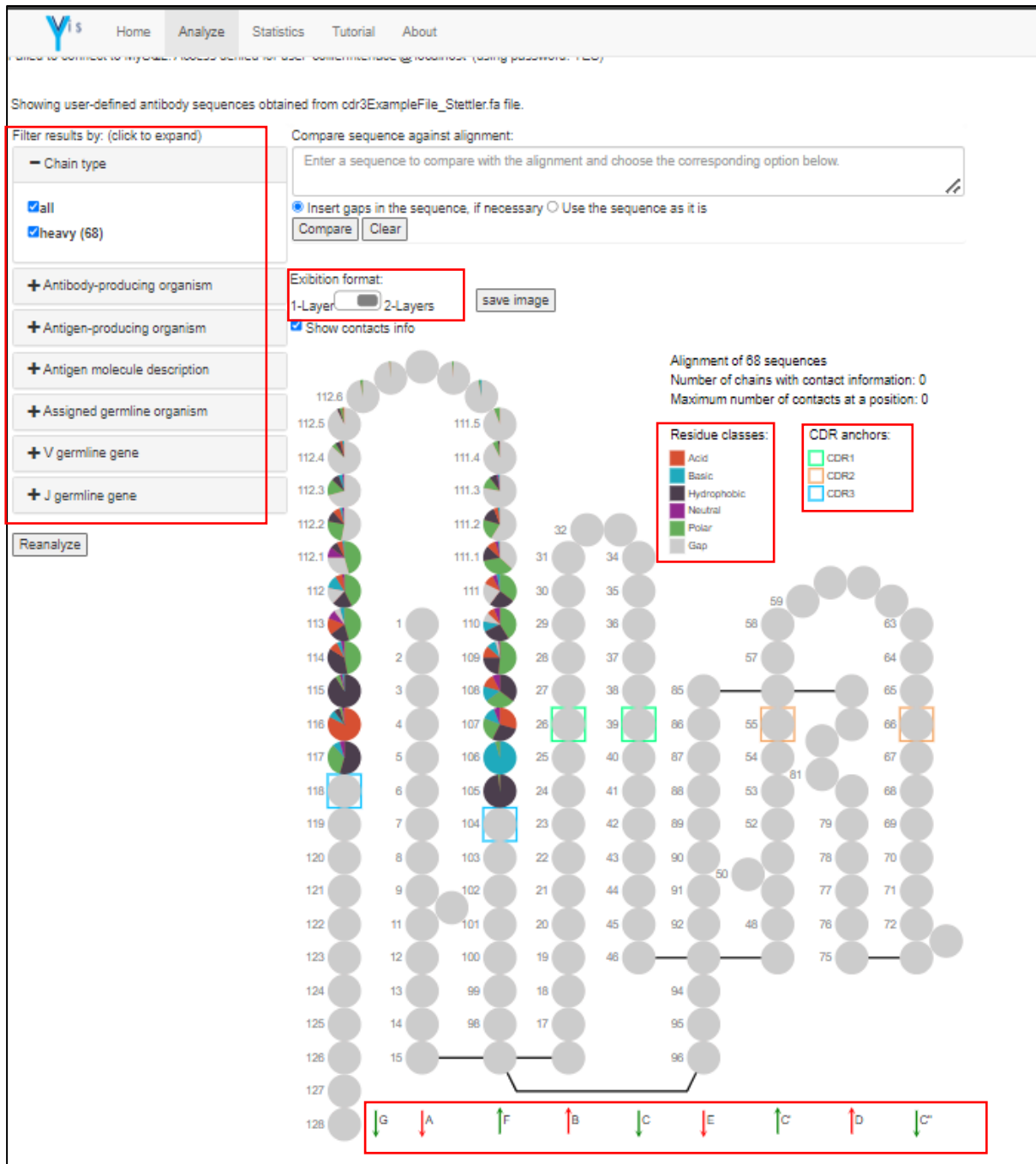


**Fig 8: Result page Showing user-defined antibody sequences obtained from cdr3ExampleFile\_Stettler (1). fa file**

**a. IMGT/Collier de Perles (Pearl Necklace) visualization in one layer**

Squares indicate the CDR anchors, one position before the CDR start regions and one after the CDR end regions (i.e., green for CDR1, orange for CDR2, and blue for CDR3)





**Fig 8(a): b. IMGT/Collier de Perles visualization in two layers**

The “pie slice” (sector) represents the number of sequences with an amino acid of a specific class (defined by a color) in that position. Green represents polar amino acids (G, S, T, Y, C, Q, N), blue represents basic (K, R, H), red represents acidic (D, E), and black represents hydrophobic amino acids (A, V, L, I, P, W, F, M). In Yvis, gaps are in grey.

The strands of the variable domain are identified by letters (A-G) and arrows at the bottom of *Collier de perles* visualization. Filter option, is to analyze a subset of the initial dataset.

PDB Id	Chain Id	Antibody Chain Type	Antibody Species	Engineered Antibody	Antigen Organism	Antigen Molecule Description	Gapped Sequence	CDR highlights: CDR1 CDR2 CDR3	Putative contact highlights:
ZKA10	KX496835	Heavy	Blue	No	Zika virus	NS1	.....	.....	.....
ZKA117	KX496861	Heavy	Blue	No	Zika virus	EDIII	.....	.....	.....
ZKA134	KX496852	Heavy	Blue	No	Zika virus	EDIII	.....	.....	.....
ZKA160	KX496843	Heavy	Blue	No	Zika virus	NNB	.....	.....	.....
ZKA172	KX496833	Heavy	Blue	No	Zika virus	NNB	.....	.....	.....
ZKA174	KX496850	Heavy	Blue	No	Zika virus	NNB	.....	.....	.....
ZKA18	KX496830	Heavy	Blue	No	Zika virus	NS1	.....	.....	.....
ZKA185	KX496858	Heavy	Blue	No	Zika virus	NNB	.....	.....	.....
ZKA189	KX496825	Heavy	Blue	No	Zika virus	NNB	.....	.....	.....
ZKA190	KX496868	Heavy	Blue	No	Zika virus	EDIII	.....	.....	.....

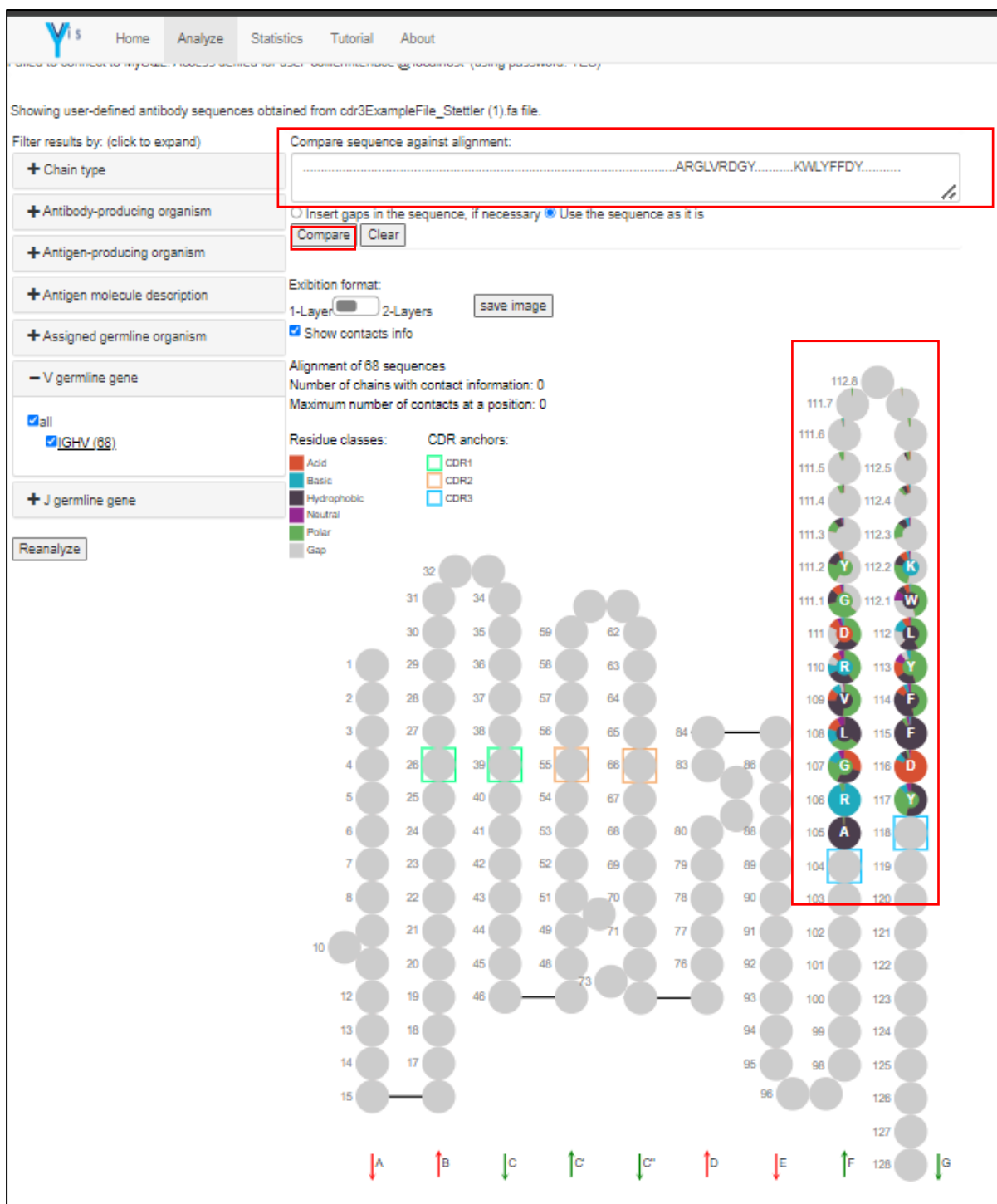
Showing 1 to 10 of 68 entries

Previous 1 2 3 4 5 6 7 Next

**Fig 9: Data table-Yvis presents a table with information on each analyzed sequence.**

**Click the PDB Id to compare with the multiple sequence alignment presented in the *Collier de perles***

A table containing the information available for all chains presented in the multiple sequence alignment. Information table contains the following fields: PDB/source identification, chain identification, chain type (heavy or light), antibody-producing organism, engineered antibody information (if the chain was marked as engineered), antigen-producing organism, antigen molecule description, gapped and ungapped chain sequences, assigned germline species, V gene, percentage of V gene identity, J gene, and percentage of J gene identity. At the gapped sequence column, the CDR positions are highlighted (green for CDR1, orange for CDR2, and blue for CDR3) as well as the putative contacts (salmon).



**Fig 10: Result page for comparison the selected sequence with the multiple sequence alignment**

Centre of each pie chart that represents a position, a small circle with the inputted sequence amino acid corresponding to that position.

## **RESULTS:**

Yvis platform were explored and the *Collier de perles* visualization presents multiple sequence alignments using pie charts, where each sector represents amino acid classes by color. Conserved positions show dominant sectors, while variable ones display multiple sectors. It

highlights key structural regions like CDRs: CDR1 corresponds to positions 26-39, CDR2 corresponds to positions 55-66, and CDR3 corresponds to position 104-118, allowing visualization of residues involved in antigen binding.

The IMGT/*Collier de Perles*, presented in one or two layers were observed. The two-layers version presents the variable domain strands in a position closer to the 3D structure, while the one-layer version has a representation closer to the variable domain sequence.

The result for the comparing the multiple sequence alignment with using the sequence as it is by selecting PDB Id from the data table containing the information available for all chains presented in the multiple sequence alignment, were observed.

## **CONCLUSIONS:**

The Yvis platform was explored and a detailed study of both the variable and constant regions, along with the topology diagram, offering a comprehensive analysis of the antibody's structure and function were studied. The Yvis platform, used for high-density antibody alignment and analysis, effectively assessed the similarity between the query and amino acid sequences, aiding in determining the relevance of associated antigens. By applying relevant filters, the platform enabled an in-depth investigation into sequence alignments.

## **REFERENCES:**

1. Wolford RW, Schaefer TJ. Zika Virus. [Updated 2023 Aug 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK430981/>
  2. Rawal, G., Yadav, S., & Kumar, R. (2016). Zika virus: An overview. *Journal of family medicine and primary care*, 5(3), 523–527. <https://doi.org/10.4103/2249-4863.197256>
  3. Masmajan, S., Musso, D., Vouga, M., Pomar, L., Dashraath, P., Stojanov, M., Panchaud, A., & Baud, D. (2020). Zika Virus. *Pathogens (Basel, Switzerland)*, 9(11), 898. <https://doi.org/10.3390/pathogens9110898>
-

**WEBLEM: 7**  
**AgAbDb DATABASE**

**AIM:**

Introduction to Ag-Ab Interaction Database (AgAbDb)

**INTRODUCTION:**

The function of antibodies (Abs) involves specific binding to antigens (Ags) and activation of other components of the immune system to fight pathogens. The six hypervariable loops within the variable domains of Abs, commonly termed complementarity determining regions (CDRs), are widely assumed to be responsible for Ag recognition, while the constant domains are believed to mediate effector activation. Recent studies and analyses of the growing number of available Ab structures, indicate that this clear functional separation between the two regions may be an oversimplification. Some positions within the CDRs have been shown to never participate in Ag binding and some off CDRs residues often contribute critically to the interaction with the Ag. Moreover, there is now growing evidence for non-local and even allosteric effects in Ab-Ag interaction in which Ag binding affects the constant region and vice versa. The CDRs have different approaches for their identification and their relationship to the Ag interface. We also review what is currently known about the contribution of non- CDRs regions to Ag recognition, namely the framework regions (FRs) and the constant domains. The suggested mechanisms by which these regions contribute to Ag binding are discussed. Beyond improving the understanding of immunity, characterization of the functional role of different parts of the Ab molecule may help in Ab engineering, design of CDR-derived peptides, and epitope prediction.

Antibodies are produced by vertebrates in response to antigens. Antigens are usually foreign molecules of invading pathogens. Antibodies are produced in billions of forms by B cells and are collectively referred to as immunoglobulins (abbreviated as Ig). The clonal selection theory states that all the antibodies produced by an individual B cell have the same antigen-binding site. Furthermore, every B cell produces a single species of antibody having a unique antigen-binding site.

Antigen–Antibody Interaction Database (AgAbDb) is an immunoinformatics resource developed at the Bioinformatics Centre, University of Pune, and is available online at <http://bioinfo.net.in/AgAbDb.htm>. Antigen–antibody interactions are a special class of protein– protein interactions that are characterized by high affinity and strict specificity of antibodies towards their antigens. Several co-crystal structures of antigen–antibody complexes have been solved and are available in the Protein Data Bank (PDB). AgAbDb is a derived knowledge base developed with an objective to compile, curate, and analyze determinants of interactions between the respective antigen– antibody molecules. AgAbDb lists not only the residues of binding sites of antigens and antibodies, but also interacting residue pairs. It also helps in the identification of interacting residues and buried residues that constitute antibody-binding sites of protein and peptide antigens. The Antigen–Antibody Interaction Finder (AAIF), a program developed in-house, is used to compile the molecular interactions, viz. van der Waals interactions, salt bridges, and hydrogen bonds.

A module for curating water-mediated interactions has also been developed. In addition, various residue level features, viz. accessible surface area, data on epitope segment, and secondary structural state of binding site residues, are also compiled. Apart from the PDB numbering, Wu–Kabat numbering and explicit definitions of complementarity-determining regions are provided for residues of antibodies. The molecular interactions can be visualized using the program Jmol. AgAbDb can be used as a benchmark dataset to validate algorithms for prediction of B-cell epitopes. It can as well be used to improve accuracy of existing algorithms and to design new algorithms. AgAbDb can also be used to design mimotopes representing antigens as well as aid in designing processes leading to humanization of antibodies.

### **REFERENCES:**

1. Merck. “Blue-White Screening & Protocols for Colony Selection.” Sigmaaldrich.com, 2021, [www.sigmaaldrich.com/MX/en/technical-documents/technical-article/genomics/cloning-and-expression/blue-white-screening](http://www.sigmaaldrich.com/MX/en/technical-documents/technical-article/genomics/cloning-and-expression/blue-white-screening)
2. Qiu, Tianyi, et al. “Proteochemometric Modeling of the Antigen-Antibody Interaction: New Fingerprints for Antigen, Antibody and Epitope-Paratope Interaction.” PLOS ONE, vol. 10, no. 4, 22 Apr. 2015, p. e0122416, <https://doi.org/10.1371/journal.pone.0122416>

**NOTE: The portal is not working for AgAbDb database**

---

**WEBLEM: 8**

**Immune Epitope Database (IEDB)**

**(URL: <https://www.iedb.org/homev3.php>)**

**AIM:**

Introduction to IEDB Database for the prediction of the Cytotoxic and Helper T cell epitopes (MHC Class 1 epitopes and MHC Class 2 epitopes)

**INTRODUCTION:**

The Immune Epitope Database (IEDB) is a comprehensive and freely accessible resource that provides detailed information on immune epitopes, which are crucial for understanding adaptive immune responses. Established in 2003 and continually updated, the IEDB serves as a vital tool for researchers studying various aspects of immunology, including vaccine development, allergy research, and autoimmune diseases.

**Key Features of IEDB**

**1. Extensive Data Repository:**

- a. The IEDB contains information on over 2.2 million epitopes related to infectious diseases, allergies, autoimmunity, and transplantation.
- b. It includes curated data from more than 20,000 published manuscripts and covers both T cell and B cell epitopes across multiple species, including humans and non-human primates.

**2. User-Friendly Interface:**

- a. The database features a web portal ([www.iedb.org](http://www.iedb.org)) that allows users to easily search and access epitope data. This includes tools for predicting and analysing B cell and T cell epitopes.
- b. Users can download data in various formats, including Microsoft Excel, XML, or MySQL.

**3. Curation Process:**

- a. Data is meticulously curated from scientific literature and submissions by researchers. The IEDB employs rigorous automated validation processes to ensure data accuracy and relevance. The curation process has evolved to reflect the increasing complexity of immune epitope data, accommodating advancements in scientific techniques and standards.

**4. Analysis Resources:**

- a. The IEDB Analysis Resource (IEDB-AR) is a companion site offering computational tools for epitope prediction and analysis. These tools include epitope clustering, sequence conservancy analysis, and predictions of T cell receptor (TCR) and B cell receptor (BCR) structures. New tools are regularly added to enhance functionality, such as those for predicting naturally processed ligands for MHC class I and II.



## **Accessibility**

The IEDB is funded by the National Institute of Allergy and Infectious Diseases (NIAID) and is available to the public without any cost. Its continuous updates ensure that it remains a relevant resource for ongoing research in immunology.

## **T Cell Epitopes**

T cell epitopes are specific peptide fragments derived from proteins (antigens) that are recognized by T cells, a crucial component of the adaptive immune system. These epitopes are presented on the surface of antigen-presenting cells (APCs) bound to major histocompatibility complex (MHC) molecules, allowing T cells to initiate an immune response.

## **Types of T Cell Epitopes**

### **1. CD8 T Cell Epitopes:**

- a. Recognized by CD8+ T cells and presented by MHC class I molecules.
- b. These epitopes typically originate from intracellular proteins, including viral or mutated proteins, allowing CD8+ T cells to identify and destroy infected or cancerous cells.

### **2. CD4 T Cell Epitopes:**

- a. Recognized by CD4+ T cells and presented by MHC class II molecules.
- b. These epitopes are usually derived from extracellular proteins and play a vital role in orchestrating the immune response by helping other immune cells.

## **Mechanism of Recognition**

The recognition of T cell epitopes involves several key steps:

- 1. Antigen Processing:** Proteins are broken down into smaller peptides within the APC.
- 2. Peptide-MHC Binding:** The processed peptides bind to MHC molecules, which then transport these complexes to the cell surface.
- 3. T Cell Activation:** The T cell receptor (TCR) on T cells recognizes the peptide-MHC complex, leading to T cell activation and proliferation.

## **Importance of T Cell Epitope Prediction**

Identifying T cell epitopes is essential for various applications, including:

- 1. Vaccine Development:** Understanding which epitopes can elicit a strong immune response aid in designing effective vaccines.
- 2. Immunotherapy:** Predicting neoepitopes (cancer-specific peptides) can enhance personalized cancer treatment strategies.
- 3. Disease Understanding:** Epitope mapping helps in elucidating mechanisms of autoimmune diseases and allergies, where inappropriate immune responses occur.

## **T Cell prediction**

### **T Cell Epitopes - MHC Binding Prediction**

#### **1. Peptide binding to MHC class I molecules**

This tool will take in an amino acid sequence, or set of sequences and determine each subsequence's ability to bind to a specific MHC class I molecule.

## **2. Peptide binding to MHC class II molecules**

This tool employs different methods to predict MHC Class II epitopes, including a consensus approach which combines NN-align, SMM-align and Combinatorial library methods.

## **3. TepiTool**

The Tepitool provides prediction of peptides binding to MHC class I and class II molecules. Tool is designed as a wizard with 6 steps.

### **T Cell Epitopes - Processing Prediction:**

These tools predict epitope candidates based upon the processing of peptides in the cell.

#### **1. Proteasomal cleavage/TAP transport/MHC class I combined predictor**

This tool combines predictors of proteasomal processing, TAP transport, and MHC binding to produce an overall score for each peptide's intrinsic potential of being a T cell epitope.

#### **2. Neural network-based prediction of proteasomal cleavage sites (NetChop) and T cell epitopes (NetCTL and NetCTLpan)**

NetChop is a predictor of proteasomal processing based upon a neural network. NetCTL and NetCTLpan are predictors of T cell epitopes along a protein sequence. It also employs a neural network architecture.

#### **3. MHC-NP**

Prediction of peptides naturally processed by the MHC.

MHC-NP employs data obtained from MHC elution experiments to assess the probability that a given peptide is naturally processed and binds to a given MHC molecule.

#### **4. MHCII-NP**

This tool utilizes MHC II ligand elution data to predict naturally processed MHC II ligands by scanning the given peptide sequences.

### **T Cell Epitopes - Immunogenicity Prediction**

These tools make predictions about the relative ability of a peptide/MHC complex to elicit an immune response.

#### **1. T cell class I pMHC immunogenicity predictor**

This tool uses amino acid properties as well as their position within the peptide to predict the immunogenicity of a class I peptide MHC (pMHC) complex.

#### **2. Deimmunization**

The deimmunization tool is attempt to identify immunodominant regions in each therapeutically important protein, and suggest amino-acid substitutions that create non-immunogenic versions of the proteins.

#### **3. CD4 T cell immunogenicity prediction**

The server is developed to predict the allele independent CD4 T cell immunogenicity at population level. User can predict the T cell immunogenicity using 7-allele method, immunogenicity method and combined method (IEDB recommended). The combined method predicts the final score that combines the predictions from 7-allele method and immunogenicity method.

#### **4. AXEL-F (Antigen Expression based Epitope Likelihood-Function)**

AXEL-F incorporates antigen abundance estimates with MHC binding predictions to enhance epitope predictions.

## **TCR Analysis**

### **1. TCRMatch**

TCRMatch compares input CDR3b sequences against curated CDR3b sequences in the IEDB to find matches that are predicted to share epitope specificity. Matches are determined by sequence similarity, which is scored using a comprehensive k-mer comparison.

## **Structure Tools**

### **1. LYRA (Lymphocyte Receptor Automated Modelling):**

The LYRA server predicts structures for either T-Cell Receptors (TCR) or B-Cell Receptors (BCR) using homology modelling. Framework templates are selected based on BLOSUM score, and complementary determining regions (CDR) are then selected if needed based on a canonical structure model and grafted onto the framework templates.

### **2. SCEptRe: Structural Complexes of Epitope Receptor**

SCEptRe provides weekly updated, non-redundant, user customized benchmark datasets with information on the immune receptor features for receptor-specific epitope predictions.

### **3. Docktope**

DockTope is a web-based tool, based on D1-EM-D2 approach, intended to allow the pMHC-I modelling. This tool has been developed by Gustavo Fioravanti Vieira's group and has been deployed to the IEDB-AR servers with minimal modification by the IEDB team.

## **B Cell Epitopes**

B cell epitopes are the specific regions of an antigen that are recognized by B cell receptors or secreted antibodies. These epitopes can be classified into two main categories based on their structure:

### **1. Linear (Continuous) Epitopes**

Consist of contiguous amino acid residues in the primary sequence of the antigen.

Represent about 10% of all identified epitopes.

Can be recognized by antibodies out of the remaining protein context and can replace the whole protein for antibody production.

### **2. Conformational (Discontinuous) Epitopes**

Include amino acid residues that are not sequential in the primary structure but are close in space due to the three-dimensional folding of the antigen.

Make up the majority (about 90%) of B cell epitopes.

The minimal amino acid sequence required for proper folding may range from 20 to 400 residues.

## **Importance of B Cell Epitope Mapping**

Identifying B cell epitopes is crucial for various applications:

- 1. Development of epitope-based vaccines:** Epitopes can be used to replace the entire antigen for antibody production.
- 2. Design of therapeutic antibodies:** Knowledge of epitopes aids in developing antibody-based therapeutics.
- 3. Improvement of immunodiagnostic tools:** Epitopes can be used in serodiagnostic assays for disease detection.

## **B Cell Epitope Prediction**

### **1. Prediction of linear epitopes from protein sequence**

A collection of methods to predict linear B cell epitopes based on sequence characteristics of the antigen using amino acid scales and HMMs.

### **2. DiscoTope - Prediction of epitopes from protein structure**

This method incorporates solvent-accessible surface area calculations, as well as contact distances into its prediction of B cell epitope potential along the length of a protein sequence.

### **3. ElliPro - Epitope prediction based upon structural protrusion**

This method predicts epitopes based upon solvent-accessibility and flexibility.

Methods for modelling and docking of antibody and protein 3D structures

This page provides information on available methods for modelling and docking of antibody and protein 3D structures.

## **Structure Tools**

### **1. LYRA (Lymphocyte Receptor Automated Modelling)**

The LYRA server predicts structures for either T-Cell Receptors (TCR) or B-Cell Receptors (BCR) using homology modelling.

### **2. SCEptRe: Structural Complexes of Epitope Receptor**

SCEptRe provides weekly updated, non-redundant, user customized benchmark datasets with information on the immune receptor features for receptor-specific epitope predictions.

This tool extracts weekly updated 3D complexes of antibody-antigen, TCR-pMHC and MHC-ligand from the Immune Epitope Database (IEDB) and clusters them based on antigens, receptors, and epitopes to generate benchmark datasets.

## **DiscoTope in IEDB**

DiscoTope is a specialized tool within the Immune Epitope Database (IEDB) designed for the prediction of B cell epitopes based on the three-dimensional (3D) structures of proteins. This tool is crucial for researchers aiming to identify potential epitopes that can elicit an immune response, particularly in the context of vaccine development and therapeutic antibody design.

## **Key Features of DiscoTope**

### **1. Structure-Based Prediction**

DiscoTope employs a structure-based approach to predict discontinuous (conformational) B cell epitopes. It utilizes 3D structural data to assess surface accessibility and calculate contact numbers, which are essential for determining how well an epitope can be recognized by antibodies.

### **2. Improved Prediction Algorithms**

The latest version, DiscoTope-3.0, features significant advancements over previous iterations. It incorporates inverse folding structure representations and utilizes a positive-unlabelled learning strategy, enabling it to predict epitopes from both solved and predicted protein structures. This enhances its applicability across various datasets and reduces dependency on experimentally solved structures.

### **3. High Performance**

DiscoTope-3.0 has demonstrated improved predictive performance compared to earlier methods, achieving high accuracy in identifying both linear and conformational epitopes across multiple independent datasets. This is particularly beneficial for large-scale predictions involving numerous proteins.

#### **4. Accessibility**

The tool is accessible through the IEDB Analysis Resource, allowing users to input PDB (Protein Data Bank) IDs or upload their own PDB files for analysis. Users can select different versions of DiscoTope for their predictions, ensuring flexibility based on their specific research needs.

#### **5. Integration with Other Databases**

DiscoTope interfaces with databases such as RCSB PDB and AlphaFold DB, facilitating large-scale predictions across a vast catalog of proteins. This integration allows researchers to leverage structural data from multiple sources for more comprehensive epitope mapping. DiscoTope is a tool used for predicting discontinuous epitopes from protein 3D structures. The transition from version 1.1 to version 2.0 introduced several significant changes in methodology and performance.

### **Algorithm Enhancements**

#### **1. Proximity Scoring Function**

**Version 1.1** utilizes a non-weighted proximity scoring function that evaluates the full-sphere neighbour count to determine the likelihood of nearby epitopes.

**Version 2.0** introduces a modified, weighted proximity scoring function that focuses on an upper half-sphere neighbour count. This adjustment improves prediction accuracy by concentrating on residues that are more likely to be exposed on the protein surface.

#### **2. Surface Exposure Measurement**

**Version 1.1** measures surface exposure based on a neighbour count within a 10 Å radius.

**Version 2.0** expands this radius to 14 Å and limits the measurement to the upper half-sphere, providing a more precise evaluation of residues that are accessible for antibody binding.

**Welcome**

The Immune Epitope Database (IEDB) is a freely available resource funded by NIAID. It catalogs experimental data on antibody and T cell epitopes studied in humans and other animal species in the context of infectious disease, allergy, autoimmunity and transplantation. The IEDB also hosts epitope prediction and analysis tools, and has a companion site, CEDAR (funded by NCI), which houses cancer epitopes.

**Upcoming Events & News**

Virtual User Workshop	Nov 5-7, 2024
Festival of Biologics	Apr 23-25, 2025
AACR 2025	Apr 25-30, 2025
Immunology 2025	May 3-7, 2025

**Summary Metrics**

Peptidic Epitopes	1,620,158
Non-Peptidic Epitopes	3,188
T Cell Assays	539,898
B Cell Assays	1,409,409
MHC Ligand Assays	4,881,364
Epitope Source Organisms	4,540
Restricting MHC Alleles	1,011
References	25,157

**START YOUR SEARCH HERE**

**Epitope**

- Any
- Linear peptide
- Discontinuous
- Non-peptidic

Exact M: [Ex: SIINFEKL]

**Assay**

- T Cell
- B Cell
- MHC Ligand

Ex: neutralization | Find

Outcome:  Positive  Negative

**Epitope Source**

Organism: [Ex: influenza, peanut] | Find

Antigen: [Ex: core, capsid, myosin] | Find

**Host**

- Any
- Human
- Mouse
- Non-human primate

Ex: dog, camel | Find

**MHC Restriction**

- Any
- Class I
- Class II
- Non-classical

Ex: HLA-A\*02:01 | Find

**Disease**

- Any
- Infectious
- Allergic
- Autoimmune

Ex: asthma | Find

Reset Search

**Epitope Analysis Resource**

**T Cell Epitope Prediction**

Scan an antigen sequence for amino acid patterns indicative of:

- MHC I Binding
- MHC II Binding
- MHC I Processing (Proteasome, TAP)
- MHC I Immunogenicity

**B Cell Epitope Prediction**

Predict linear B cell epitopes using:

- Antigen Sequence Properties

Predict discontinuous B cell epitopes using antigen structure via:

- Discoptoe
- ElliPro

**Epitope Analysis Tools**

Analyze epitope sets of:

- Population Coverage
- Conservation Across Antigens
- Clusters with Similar Sequences

**Fig 1: Homepage of the IEDB database**

**IEDB Analysis Resource**

Overview | T Cell Tools | B Cell Tools | Analysis Tools | Tools-API | Usage | Download | Datasets | Contribute Tools | References

**Epitope Prediction and Analysis Tools**

Welcome to the Immune Epitope Database Analysis Resource. This site provides a collection of tools for the prediction and analysis of immune epitopes. It serves as a companion site to the Immune Epitope Database (IEDB), a manually curated database of experimentally characterized immune epitopes.

The tools contained fall into the following categories:

**T Cell Epitope Prediction Tools**

This set of tools includes MHC class I & II binding predictions, as well as peptide processing predictions and immunogenicity predictions.

**B Cell Epitope Prediction Tools**

The tools here are intended to predict regions of proteins that are likely to be recognized as epitopes in the context of a B cell response.

**Analysis Tools**

The epitope analysis tools are intended for the detailed analysis of a known epitope sequence or group of sequences.

**IEDB-AR News**

- Next-generation Tools site available!**  
Head over to <https://nextgen-tools.iedb.org> where we are re-implementing our tools to provide intuitive but powerful user interface features most requested by our user community. Try it out!

**IEDB-AR Release Notes**

- IEDB Analysis Resource v2.28 release notes (26 Apr 2024)**  
2024-05-08  
NetMHCIIpan 4.2 and NetMHCIIpan 4.3 incorporated into mhcii binding web, API, and standalone tools. [Bugfix] Axel-F support...
- IEDB Analysis Resource v2.27 release notes (25 May 2023)**
- IEDB Analysis Resource v2.26 release notes (24 Feb 2022)**

© 2005-2024 | [IEDB Home](#) | [Help](#) | [Contact](#)  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 2: Overview of the IEDB database**

The screenshot shows the IEDB (Immune Epitope Database & Tools) homepage. At the top, there is a navigation bar with 'Home', 'Specialized Searches', and 'Analysis Resource'. A dropdown menu is open under 'Analysis Resource', listing options like 'Analysis Resource Overview', 'T Cell Epitope Prediction', 'B Cell Epitope Prediction', 'Epitope Analysis Tools', and 'Tool Licensing Information'. A red banner at the top reads: 'Check out our new IEDB updates! (1) Learn how to customize your database exports and (2) to visit our new analysis tools site for all your analysis and prediction needs.'

The main content area is divided into several sections:

- Welcome:** A brief introduction to the IEDB as a freely available resource funded by NIAID, cataloging experimental data on antibody and T cell epitopes.
- Upcoming Events & News:** A list of events including a Virtual User Workshop (Nov 5-7, 2024), Festival of Biologics (Apr 23-25, 2025), AACR 2025 (Apr 25-30, 2025), and Immunology 2025 (May 3-7, 2025).
- Summary Metrics:** A table showing the following data:
 

Peptidic Epitopes	1,620,158
Non-Peptidic Epitopes	3,188
T Cell Assays	539,898
B Cell Assays	1,409,409
MHC Ligand Assays	4,881,364
Epitope Source Organisms	4,540
Restricting MHC Alleles	1,011
References	25,157
- START YOUR SEARCH HERE:** A central search interface with filters for:
  - Epitope:** Any (selected), Linear peptide, Exact M (Ex: SIINFEKL), Discontinuous, Non-peptidic.
  - Assay:** T Cell (checked), B Cell (checked), MHC Ligand (checked). Outcome: Positive (checked), Negative.
  - Epitope Source:** Organism (Ex: influenza, peanut), Antigen (Ex: core, capsid, myosin).
  - MHC Restriction:** Any (selected), Class I, Class II, Non-classical (Ex: HLA-A\*02:01).
  - Host:** Any (selected), Human, Mouse, Non-human primate (Ex: dog, camel).
  - Disease:** Any (selected), Infectious, Allergic, Autoimmune (Ex: asthma).
- Epitope Analysis Resource:** A sidebar with sections for:
  - T Cell Epitope Prediction:** Scan an antigen sequence for amino acid patterns indicative of: MHC I Binding, MHC II Binding, MHC I Processing (Proteasome, TAP), MHC I Immunogenicity.
  - B Cell Epitope Prediction:** Predict linear B cell epitopes using: Antigen Sequence Properties. Predict discontinuous B cell epitopes using antigen structure via: Discotope, ElliPro.
  - Epitope Analysis Tools:** Analyze epitope sets of: Population Coverage, Conservation Across Antigens, Clusters with Similar Sequences.

**Fig 3: Different resources in the IEDB Database**

The screenshot shows the 'IEDB Analysis Resource' page, specifically the 'T Cell Epitope Prediction Tools' section. The navigation bar includes 'Overview', 'T Cell Tools', 'B Cell Tools', 'Analysis Tools', 'Tools-API', 'Usage', 'Download', 'Datasets', 'Contribute Tools', and 'References'. The 'T Cell Tools' tab is active.

The main content area is titled 'T Cell Epitope Prediction Tools' and contains the following information:

- T Cell Epitopes - MHC Binding Prediction:** These tools predict IC50 values for peptides binding to specific MHC molecules. Note that binding to MHC is necessary but not sufficient for recognition by T cells.
  - Peptide binding to MHC class I molecules: This tool will take in an amino acid sequence, or set of sequences and determine each subsequence's ability to bind to a specific MHC class I molecule.
  - Peptide binding to MHC class II molecules: This tool employs different methods to predict MHC Class II epitopes, including a consensus approach which combines NN-align, SMM-align and Combinatorial library methods.
  - TepiTool: The TepiTool provides prediction of peptides binding to MHC class I and class II molecules. Tool is designed as a wizard with 6 steps as described below. Each field (except sequences and alleles) is filled with default recommended settings for prediction and selection of optimum peptides. The input parameters can be adjusted as per your specific needs. You can go back to previous steps to change your selection before submission of the job. Once you submit the job (at the end of step-6), you will not be able to make any more changes and will have to start the prediction all over again with updated input parameters.
- T Cell Epitopes - Processing Prediction:** These tools predict epitope candidates based upon the processing of peptides in the cell.
  - Proteasomal cleavage/TAP transport/MHC class I combined predictor: This tool combines predictors of proteasomal processing, TAP transport, and MHC binding to produce an overall score for each peptide's intrinsic potential of being a T cell epitope.
  - Neural network based prediction of proteasomal cleavage sites (NetChop) and T cell epitopes (NetCTL and NetCTLpan): NetChop is a predictor of proteasomal processing based upon a neural network. NetCTL and NetCTLpan are predictors of T cell epitopes along a protein sequence. It also employs a neural network architecture.
  - MHC-NP: Prediction of peptides naturally processed by the MHC: MHC-NP employs data obtained from MHC elution experiments in order to assess the probability that a given peptide is naturally processed and binds to a given MHC molecule. This tool was the winner of the 2nd Machine Learning Competition in Immunology.

**Fig 4: T cell Prediction Tool**

**IEDB Analysis Resource**

Overview | T Cell Tools | **B Cell Tools** | Analysis Tools | Tools-API | Usage | Download | Datasets | Contribute Tools | References

### B Cell Epitope Prediction Tools

#### B Cell Epitope Prediction

[Prediction of linear epitopes from protein sequence](#)  
A collection of methods to predict linear B cell epitopes based on sequence characteristics of the antigen using amino acid scales and HMMs.

[DiscoTope - Prediction of epitopes from protein structure](#)  
This method incorporates solvent-accessible surface area calculations, as well as contact distances into its prediction of B cell epitope potential along the length of a protein sequence.

[ElliPro - Epitope prediction based upon structural protrusion](#)  
This method predicts epitopes based upon solvent-accessibility and flexibility.

[Methods for modeling and docking of antibody and protein 3D structures](#)  
This page provides information on available methods for modeling and docking of antibody and protein 3D structures.

#### Structure Tools

[LYRA \(Lymphocyte Receptor Automated Modelling\)](#):  
The LYRA server predicts structures for either T-Cell Receptors (TCR) or B-Cell Receptors (BCR) using homology modelling. Framework templates are selected based on BLOSUM score, and complementary determining regions (CDR) are then selected if needed based on a canonical structure model and grafted onto the framework templates.

[SCEptRe: Structural Complexes of Epitope Receptor](#)

**Fig 5: B cell prediction tool**

**IEDB Analysis Resource**

Overview | T Cell Tools | B Cell Tools | **Analysis Tools** | Tools-API | Usage | Download | Datasets | Contribute Tools | References

### Analysis Tools

#### Analysis Tools

The tools below are intended for the detailed analysis of a known epitope sequence or group of sequences.

[Population Coverage](#)  
This tool calculates the fraction of individuals predicted to respond to a given set of epitopes with known MHC restrictions. This calculation is made on the basis of HLA genotypic frequencies assuming non-linkage disequilibrium between HLA loci.

[Epitope Conservancy Analysis](#)  
This tool calculates the degree of conservancy of an epitope within a given protein sequence set at different degrees of sequence identity. The degree of conservation is defined as the fraction of protein sequences containing the epitope at a given identity level.

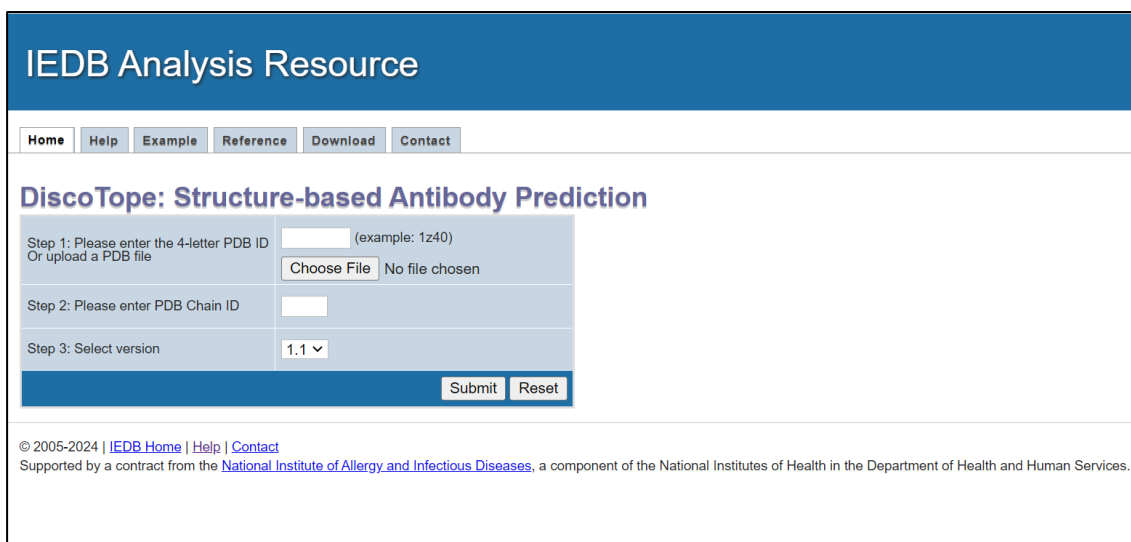
[Epitope Cluster Analysis](#)  
This tool groups epitopes into clusters based on sequence identity. A cluster is defined as a group of sequences which have a sequence similarity greater than the minimum sequence identity threshold specified.

[Computational Methods for Mapping Mimotopes to Protein Antigens](#)  
This page provides information on available methods for mimotope mapping, how to search the IEDB for mimotopes, and an example of a mimotope dataset and the results of its mapping, using the available web servers hosted outside the IEDB.

[RATE \(Restrictor Analysis Tool for Epitopes\)](#)  
The RATE is an automated method that can infer HLA restriction for a set of given epitopes from large datasets of T cell responses in HLA typed subjects. The tool takes two data files, one containing the alleles expressed by the subjects and the other containing the response of the peptides in the subjects. The tool calculates the odds ratio and estimates its significance using Fisher's exact test. It also calculates a parameter called relative frequency similar to odds ratio. The tool was developed with a focus on class II alleles but can also be applied to class I alleles.

**Fig 6: Analysis tools**





**IEDB Analysis Resource**

Home Help Example Reference Download Contact

**DiscoTope: Structure-based Antibody Prediction**

Step 1: Please enter the 4-letter PDB ID (example: 1z40)  
Or upload a PDB file

Choose File No file chosen

Step 2: Please enter PDB Chain ID

Step 3: Select version 1.1

Submit Reset

© 2005-2024 | [IEDB Home](#) | [Help](#) | [Contact](#)  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 7: Homepage of DiscoTope program**

## **REFERENCES:**

1. Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and Methods for T- and B-Cell Epitope Prediction. *Journal of immunology research*, 2017, 2680160. <https://doi.org/10.1155/2017/2680160>
2. Konstantinou G. N. (2017). T-Cell Epitope Prediction. *Methods in molecular biology (Clifton, N.J.)*, 1592, 211–222. [https://doi.org/10.1007/978-1-4939-6925-8\\_17](https://doi.org/10.1007/978-1-4939-6925-8_17)
3. Kohlgruber, A. C., Dezfulian, M. H., Sie, B. M., Wang, C. I., Kula, T., Laserson, U., Larman, H. B., & Elledge, S. J. (2024). High-throughput discovery of MHC class I- and II-restricted T cell epitopes using synthetic cellular circuits. *Nature Biotechnology*. <https://doi.org/10.1038/s41587-024-02248-6>
4. *T cell tools*. (n.d.). <http://tools.iedb.org/main/tcell/>
5. Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and Methods for T- and B-Cell Epitope Prediction. *Journal of immunology research*, 2017, 2680160. <https://doi.org/10.1155/2017/2680160>
6. Høie, M. H., Gade, F. S., Johansen, J. M., Würtzen, C., Winther, O., Nielsen, M., & Marcatili, P. (2024). DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. *Frontiers in immunology*, 15, 1322712. <https://doi.org/10.3389/fimmu.2024.1322712>
7. Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and Methods for T- and B-Cell Epitope Prediction. *Journal of immunology research*, 2017, 2680160. <https://doi.org/10.1155/2017/2680160>
8. Martini, S., Nielsen, M., Peters, B., & Sette, A. (2020). The Immune Epitope Database and Analysis Resource Program 2003-2018: reflections and outlook. *Immunogenetics*, 72(1-2), 57–76. <https://doi.org/10.1007/s00251-019-01137-6>

**DATE: 25/09/2024**

**WEBLEM: 8(A)**

**Immune Epitope Database (IEDB)**

**(URL: <https://www.iedb.org/>)**

**AIM:**

To predict B-Cell epitope for query AMA1 (PDB ID: 1Z40) using DiscoTope Server 1.1 from IEDB Database.

**INTRODUCTION:**

The Immune Epitope Database (IEDB) is a free, publicly accessible resource established in 2004 that catalogs experimental data on antibody and T cell epitopes. It serves as a comprehensive platform for researchers to access curated epitope data from scientific literature, currently housing over 1.6 million experiments related to various fields, including infectious diseases, allergies, autoimmunity, and transplantation. The IEDB allows users to easily search for epitope information and integrates data from multiple external resources, enhancing usability and accessibility. The IEDB also hosts epitope prediction and analysis tools, and has a companion site, CEDAR (funded by NCI), which houses cancer epitopes.

**IEDB Analysis Resource (IEDB-AR) and DiscoTope**

Accompanying the IEDB is the IEDB Analysis Resource (IEDB-AR), which provides computational tools for predicting and analyzing B and T cell epitopes. Among its various tools is **DiscoTope**, specifically designed to predict B cell epitopes based on structural data. DiscoTope utilizes amino acid statistics, surface accessibility, and spatial information to identify potential epitopes on the surface of antigens. DiscoTope is a method for predicting discontinuous epitopes from 3D structures of proteins in PDB format. This tool has become invaluable in antibody engineering and vaccine design, allowing researchers to pinpoint B cell epitopes that can be targeted by antibodies. The IEDB-AR continues to evolve, offering enhanced features and improved performance for epitope prediction and analysis.

**AMA1 (PDB ID: 1Z40)**

AMA1 (Apical Membrane Antigen 1) is a critical type I transmembrane protein found on the merozoite stage of the Plasmodium parasite, which causes malaria. Its primary function is facilitating the invasion of red blood cells by interacting with host cells, making it essential for the parasite's life cycle. AMA1 is also known for its role in immune evasion due to antigenic variation. Given its prominence in the invasion process, AMA1 is a key target for malaria vaccine development. Antibodies against AMA1 can inhibit erythrocyte invasion, highlighting its importance in generating protective immunity. Consequently, AMA1 remains a significant focus in malaria research, offering valuable insights into infection dynamics and vaccine design.

**METHODOLOGY:**

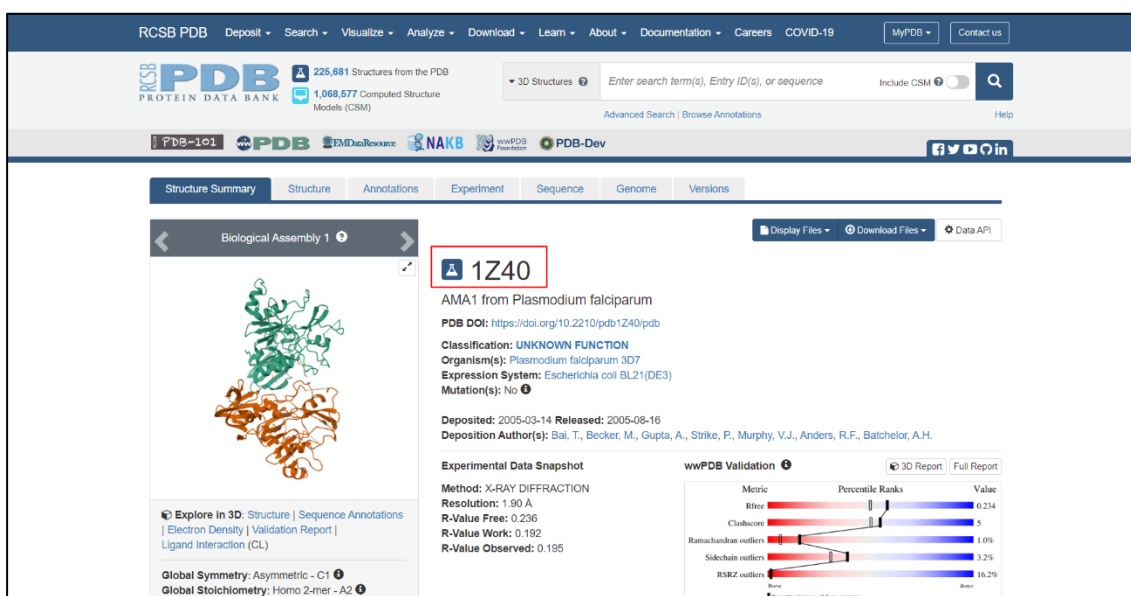
1. Open the Protein Data Bank (PDB) website. (URL: <https://www.rcsb.org/>) to obtain the PDB ID of the structure (query).

2. Open the Immune Epitope Database and Tools (IEDB) (<https://www.iedb.org/>) that contains experimental data on antibody and T-cell epitope, also host epitope prediction.
3. Select the 'DiscoTope' option under B Cell Epitope prediction from Epitope Analysis Resource section.
4. Enter the 4-letter PDB ID or upload a PDB file for the query (PDB ID:1BBJ), Chain ID for protein chain of interest and select the DiscoTope version 1.1.
5. Click on Submit.
6. Analyse the results in Chart view, Table view and 3D View.
7. The prediction can be saved in .csv extension.

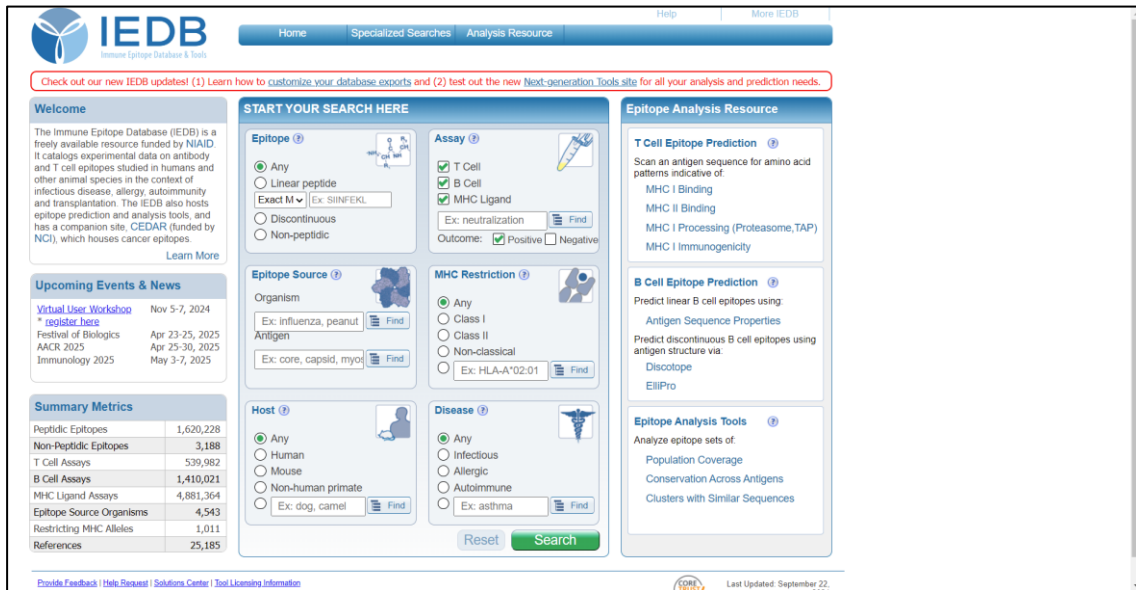
## **OBSERVATIONS:**



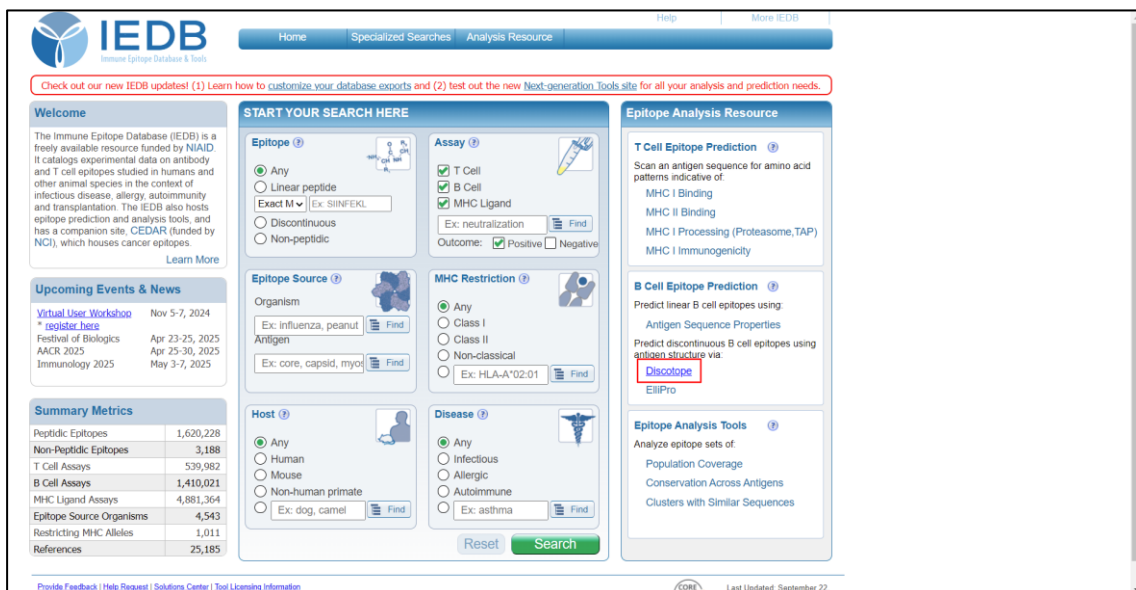
**Fig 1: Homepage of the Protein Data Bank (PDB) database**



**Fig 2: Retrieving the query ‘AMA1’ (PDB ID: 1Z40) from the PDB database**



**Fig 3: Homepage of the Immune Epitope Database (IEDB)**



**Fig 4: Selecting the ‘Discoptope’ option in the Epitope Analysis Resource section**

**IEDB Analysis Resource**

[Home](#) [Help](#) [Example](#) [Reference](#) [Download](#) [Contact](#)

### DiscoTope: Structure-based Antibody Prediction

Step 1: Please enter the 4-letter PDB ID  (example: 1z40)  
Or upload a PDB file  No file chosen

Step 2: Please enter PDB Chain ID

Step 3: Select version

© 2005-2024 | [IEDB Home](#) | [Help](#) | [Contact](#)  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 5: Homepage of DiscoTope Program for Prediction**

**IEDB Analysis Resource**

[Home](#) [Help](#) [Example](#) [Reference](#) [Download](#) [Contact](#)

### DiscoTope: Structure-based Antibody Prediction

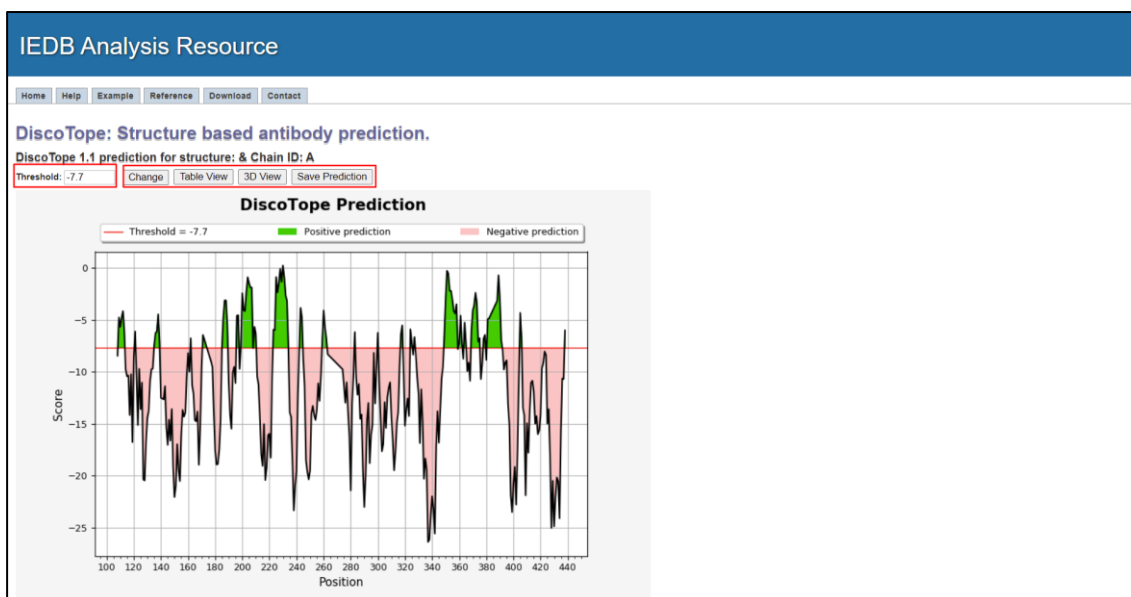
Step 1: Please enter the 4-letter PDB ID  (example: 1z40)  
Or upload a PDB file  No file chosen

Step 2: Please enter PDB Chain ID

Step 3: Select version

© 2005-2024 | [IEDB Home](#) | [Help](#) | [Contact](#)  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 6: Searching for the PDB code: '1BBJ', Chain ID: A, select version 1.1 and click on submit**



**Fig 7: Search Results obtained can be analysed in Chart View, Table View and 3D View**

### Description

To change the threshold value, enter a different threshold and click on 'Change'. The default value for version 1.1 is -7.7 and version 2.0 is -3.7, which corresponds to a specificity of 75%. Higher values correspond to higher specificity. A specificity of 0.75 means that 25% of the non-epitope residues were predicted as part of epitopes. A sensitivity of 0.47 means that 47% of the epitope residues were predicted as part of epitopes. In the chart, predictions above the threshold (red line) are positive predictions (displayed in green) and predictions below the threshold are negative prediction (displayed in orange).

Chain ID	Residue ID	Residue Name	Contact Number	Propensity Score	DiscoTope Score
A	108	ASN	14	-1.459	-8.459
A	109	PRO	11	0.724	-4.776
A	110	TRP	13	0.804	-5.696
A	111	THR	12	1.211	-4.789
A	112	GLU	11	1.331	-4.189
A	113	TYR	14	0.929	-6.071
A	114	MET	18	-0.779	-9.779
A	115	ALA	20	-0.444	-10.444
A	116	LYS	21	0.122	-10.378
A	117	TYR	24	-2.172	-14.172
A	118	ASP	21	0.267	-10.243
A	119	ILE	32	-0.783	-16.783
A	120	GLU	21	1.954	-8.546
A	121	GLU	15	1.366	-6.134
A	122	VAL	20	-0.374	-10.374
A	123	HIS	28	-1.144	-15.144
A	124	GLY	22	1.274	-9.726
A	125	SER	29	0.867	-13.813
A	126	GLY	28	2.951	-11.949
A	127	ILE	35	-2.881	-20.381
A	128	ARG	29	-5.973	-20.473
A	129	VAL	30	-1.817	-16.817
A	130	ASP	31	1.848	-14.852
A	131	LEU	31	1.727	-13.773
A	132	GLY	25	1.617	-10.883
A	133	GLU	19	-0.26	-5.76
A	134	ASP	18	-0.714	-9.714

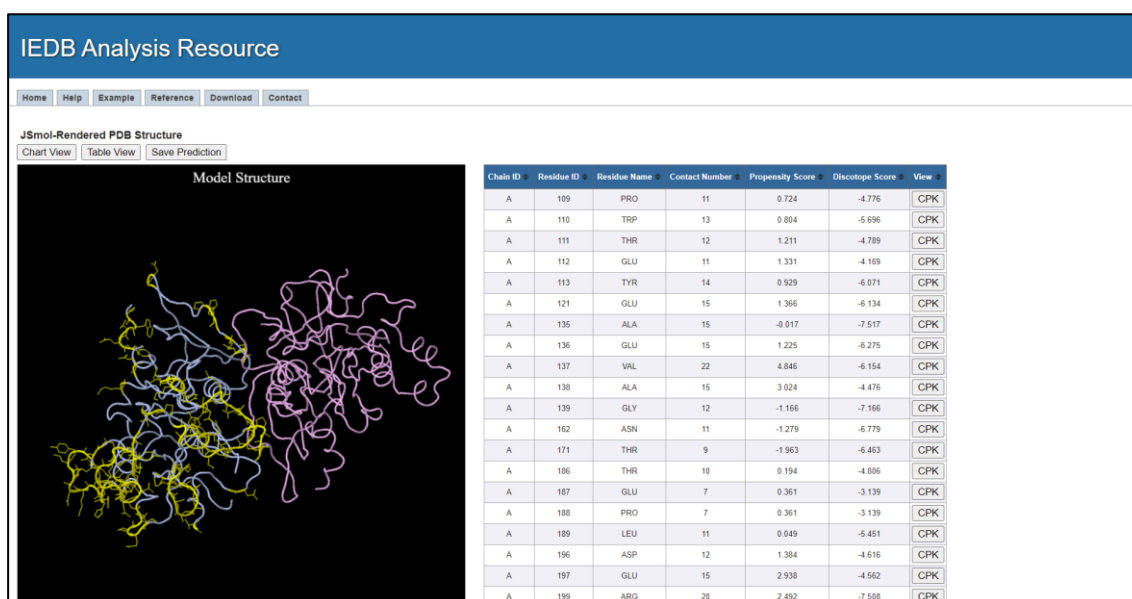
**Fig 8: Results obtained in Table View**

### Description

1. Chain ID: The chain id of the protein chain used in prediction (specified by the user)

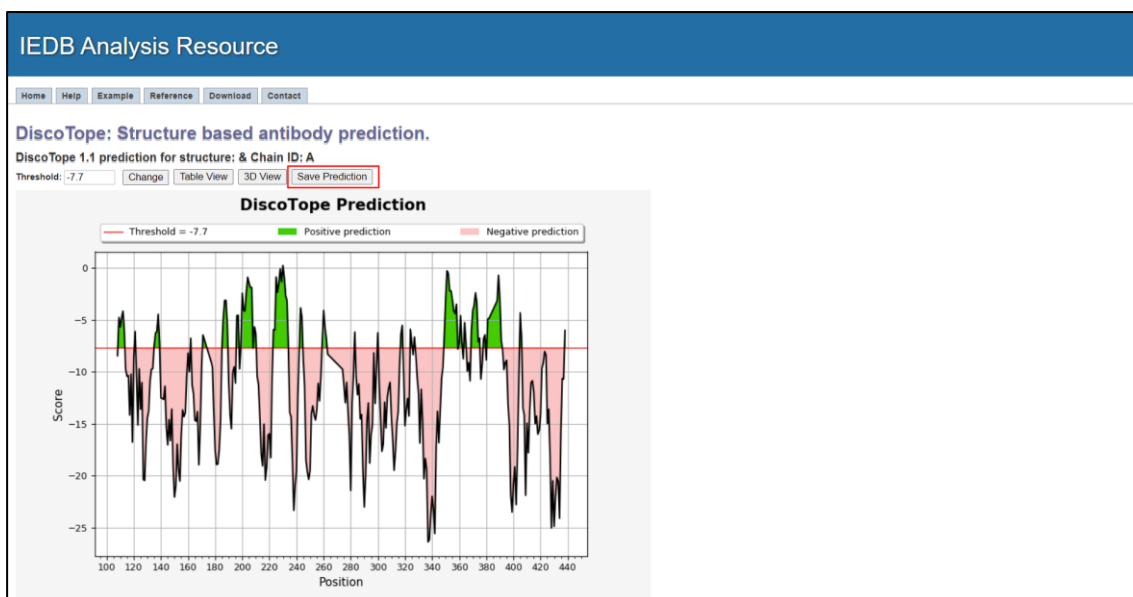
2. Residue ID: PDB Residue id
3. Residue Name: Name of the residue
4. Contact Number: The residue contact number is the number of C $\alpha$  atoms in the antigen within 10 Å of the residue's C $\alpha$  atom. A low contact number correlates with localization of the residue close to the surface or in protruding regions of the antigen's structures.
5. Propensity Score: This score tells you about the probability/tendency of being part of an epitope for that residue. The propensity is reflected in amino acid epitope log-odds ratios, which were calculated on a set of 75 antigens. The propensity score is calculated by sequentially averaging epitope log-odds ratios within a window of 9 residues. Then the scores are summed up based on the proximity in the 3D structure of the antigen. For any given residue, the sequentially averaged log-odds scores from all residues within 10Å are summed to give the propensity score.
6. Discotope Score: This score is calculated by combining the contact numbers with propensity score. DiscoTope score above the threshold value indicates positive predictions and that below the threshold value indicates negative predictions.

Positive predictions are displayed in green. Click on header to sort column.



**Fig 9: Results obtained in 3D Viewer**





**Fig 10: Save the prediction in .csv Extension**

## **RESULTS:**

The DiscoTope analysis of the AMA1 antibody (PDB ID: 1Z40) provided several critical insights regarding potential B cell epitopes:

1. **Chart View:** The predictions indicated that several residues had DiscoTope scores above the threshold value i.e., -7.7, which are considered positive predictions, displayed in green. This score suggests a high likelihood of these residues being part of an epitope, indicating their potential accessibility for antibody binding.
2. **Table View:** The detailed results included the following key metrics:
  - a. **Chain ID:** The chain ID for the analyzed protein, which is Chain A.
  - b. **Residue ID:** The specific identifier for the lysine residue (LYS).
  - c. **Contact Number:** The contact number for this proline residue is **11**, indicating that there are eleven C $\alpha$  atoms from other residues within 10 Å of this proline's C $\alpha$  atom. A low contact number suggests that this residue is well-exposed on the surface of the protein and accessible for antibody binding. In this context, a contact number of 11 indicates reasonable accessibility, although it may not be as optimal as residues with lower contact numbers. Generally, low contact numbers (e.g., 1 or 2) are preferred for predicting epitopes, as they correlate with increased exposure on the protein surface.
  - d. **Propensity Score:** The propensity score for this proline is **5.717**, reflecting a strong likelihood that this residue is part of an epitope based on its amino acid characteristics and statistical analysis from known antigens.
  - e. **DiscoTope Score:** The DiscoTope score for this residue is **0.217**. This positive score indicates a higher probability of this residue being involved in immune recognition, despite its contact number and favorable propensity score.
3. **3D View:** The spatial visualization allowed for examination of predicted epitopes on the AMA1 structure, facilitating insights into their interactions with antibodies.



## **CONCLUSION:**

The analysis highlighted a proline residue with a contact number of 11, indicating reasonable accessibility, a high propensity score of 5.717 suggesting it has a strong likelihood of being part of an epitope, but a DiscoTope score of 0.217 indicating that it may be as strongly favored for immune recognition compared to other residues with negative scores.

Additionally, the findings underscore the importance of utilizing tools like DiscoTope within the Immune Epitope Database (IEDB) to predict B cell epitopes effectively. The IEDB serves as a comprehensive resource for researchers, offering curated data and analytical tools that enhance our understanding of immune responses and facilitate vaccine development and therapeutic design. Understanding the range of contact scores is crucial; while lower scores indicate better surface exposure and accessibility for antibody binding, higher scores can still suggest reasonable accessibility depending on the context. Overall, these insights contribute significantly to advancing research in immunology and related fields by providing a clearer picture of how structural features influence epitope prediction and antibody interactions.

## **REFERENCES:**

1. Martini, S., Nielsen, M., Peters, B., & Sette, A. (2020). The Immune Epitope Database and Analysis Resource Program 2003-2018: reflections and outlook. *Immunogenetics*, 72(1-2), 57–76. <https://doi.org/10.1007/s00251-019-01137-6>
  2. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic acids research*, 47(D1), D339–D343. <https://doi.org/10.1093/nar/gky1006>
  3. *IEDB.org: Free epitope database and prediction resource*. (n.d.). <https://www.iedb.org/>
  4. Treeck, M., Zacherl, S., Herrmann, S., Cabrera, A., Kono, M., Struck, N. S., Engelberg, K., Haase, S., Frischknecht, F., Miura, K., Spielmann, T., & Gilberger, T. W. (2009b). Functional Analysis of the Leading Malaria Vaccine Candidate AMA-1 Reveals an Essential Role for the Cytoplasmic Domain in the Invasion Process. *PLoS Pathogens*, 5(3), e1000322. <https://doi.org/10.1371/journal.ppat.1000322>
-

**DATE: 01/10/2024**

**WEBLEM: 9**

**PaDELPy**

**(URL: <https://github.com/ecrl/padelpy>)**

**AIM:**

Introduction to molecular Descriptors and PADEL Descriptor software.

**INTRODUCTION:**

PaDELPy is a powerful and versatile Python package that serves as a wrapper for PaDEL-Descriptor, a widely-used software in the field of cheminformatics and computational chemistry. This package bridges the gap between the Java-based PaDEL-Descriptor and the Python ecosystem, enabling researchers and data scientists to seamlessly integrate molecular descriptor calculations into their Python workflows. By providing a convenient Python interface to PaDEL-Descriptor, PaDELPy facilitates the calculation of a wide array of molecular descriptors and fingerprints, which are essential for various cheminformatics applications.

The primary purpose of PaDELPy is to simplify the process of calculating molecular descriptors and fingerprints, making these crucial tools more accessible to researchers working in Python environments. It offers a comprehensive set of features that make it an invaluable asset in computational chemistry. PaDELPy can compute a vast array of molecular descriptors, including 1D, 2D, and 3D descriptors. These encompass constitutional descriptors, which provide basic information about the molecule's composition; topological descriptors, which capture the connectivity and shape of molecules; geometrical descriptors, which describe the three-dimensional structure of molecules; electronic descriptors, which relate to the distribution of charge in molecules; and hybrid descriptors, which combine multiple types of molecular information.

In addition to descriptor calculation, PaDELPy supports the generation of various types of molecular fingerprints, such as MACCS keys, PubChem fingerprints, and substructure fingerprints. These fingerprints are crucial for tasks like similarity searching and machine learning applications in drug discovery. The package accepts multiple input formats, including SMILES strings, SDF files, and MOL2 files, providing flexibility in handling different molecular representations. This input flexibility allows researchers to work with their preferred molecular formats without the need for additional conversion steps.

One of the key strengths of PaDELPy is its customization options. Users can tailor their calculations by selecting specific descriptors or fingerprints and adjusting calculation parameters to suit their research needs. This level of control allows researchers to focus on the molecular features most relevant to their studies, potentially improving the efficiency and relevance of their analyses. To use PaDELPy, researchers must have both Python and Java installed on their system. The package can be easily installed using pip, the Python package installer, with a simple command: "pip install padel-py". This straightforward installation process makes PaDELPy readily accessible to researchers and developers working in Python environments. Once installed, using PaDELPy in a Python script is straightforward. For

example, to calculate descriptors for a molecule represented as a SMILES string, one would first import the PaDELDescriptor class from `padel_py`, initialize it, and then use the `calculate_descriptors` method with the SMILES string as input.

## **MOLECULAR DESCRIPTORS**

Molecular descriptors are numerical values that describe the chemical structure of molecules. They can represent various properties such as molecular weight, atom counts, functional groups, and 3D spatial information. These descriptors are essential for quantitative structure-activity relationship (QSAR) modeling, virtual screening, and other cheminformatics applications. Here are the main types of molecular descriptors up to 6D:

### **1. 0D Descriptors (Constitutional Descriptors)**

These descriptors are simple counts of atoms, bonds, or molecular fragments without considering the molecule's connectivity or spatial arrangement.

#### **Examples:**

Molecular weight, number of atoms, number of bonds, and atom type counts.

### **2. 1D Descriptors (Structural Descriptors)**

1D descriptors represent the sequence of atoms or specific chemical groups in a molecule without taking molecular topology into account.

#### **Examples:**

Molecular formulas, number of specific functional groups (e.g., hydroxyl groups, halogens).

### **3. 2D Descriptors (Topological Descriptors):**

These descriptors represent the connectivity or topology of a molecule's structure in two dimensions. They are derived from the molecular graph, where atoms are represented as nodes and bonds as edges.

#### **Examples:**

- a. Topological indices like Wiener index, Balaban index.
- b. Atom connectivity indices (degree of atoms, valence values).
- c. Fragment-based descriptors (counts of specific substructures or functional groups).

### **4. 3D Descriptors (Geometric/Spatial Descriptors)**

3D descriptors represent the three-dimensional arrangement of atoms in space, capturing molecular shape, size, and volume. These are crucial for understanding stereochemistry and interactions in docking studies.

#### **Examples:**

- a. Molecular surface area (van der Waals or solvent-accessible surface).
- b. Molecular volume.
- c. Dipole moment.
- d. Shape indices (e.g., radius of gyration).

## 5. 4D Descriptors (Molecular Dynamics Descriptors)

4D descriptors incorporate time-dependent information to represent the dynamic behavior of molecules in different environments (e.g., solution, gas phase). These descriptors are generated from molecular dynamics (MD) simulations and capture conformational changes over time.

### Examples:

- a. Time-averaged molecular properties (e.g., average distances between atoms).
- b. Conformational flexibility measures.

## 6. 5D Descriptors (Quantum Descriptors)

5D descriptors account for quantum mechanical properties and interactions, often used in quantum chemistry. These descriptors consider the electronic structure of molecules and how they change under different conditions.

### Examples:

- a. Electron density distribution.
- b. Molecular orbitals (HOMO, LUMO).
- c. Quantum mechanical energy levels.

## 7. 6D Descriptors (Pharmacophore Descriptors)

6D descriptors are used to represent the pharmacophoric properties of molecules. A pharmacophore is a set of features (like hydrogen bond donors, acceptors, hydrophobic regions) that are responsible for the biological activity of a molecule. These descriptors aim to capture the spatial arrangement and dynamic nature of pharmacophoric features in a molecule.

### Examples:

- a. Pharmacophoric patterns based on 3D alignments of molecular features.
- b. Dynamic pharmacophore models (time-dependent movements of pharmacophoric features).

### Significance of Molecular Descriptors:

1. Drug Design and Discovery: Aid in predicting biological activity, toxicity, and pharmacokinetic properties for identifying potential drug candidates.
2. Structure-Activity Relationship (SAR): Help correlate molecular structures with biological activity, enabling QSAR modeling to understand structure-function relationships.
3. Predictive Modeling: Serve as inputs for machine learning models to predict chemical properties like toxicity, solubility, and binding affinity.
4. Chemical Property Analysis: Provide insights into molecular properties like hydrophobicity, polarity, molecular weight, etc., essential for understanding molecular interactions.

### **Key Molecular Descriptors and Their Symbols:**

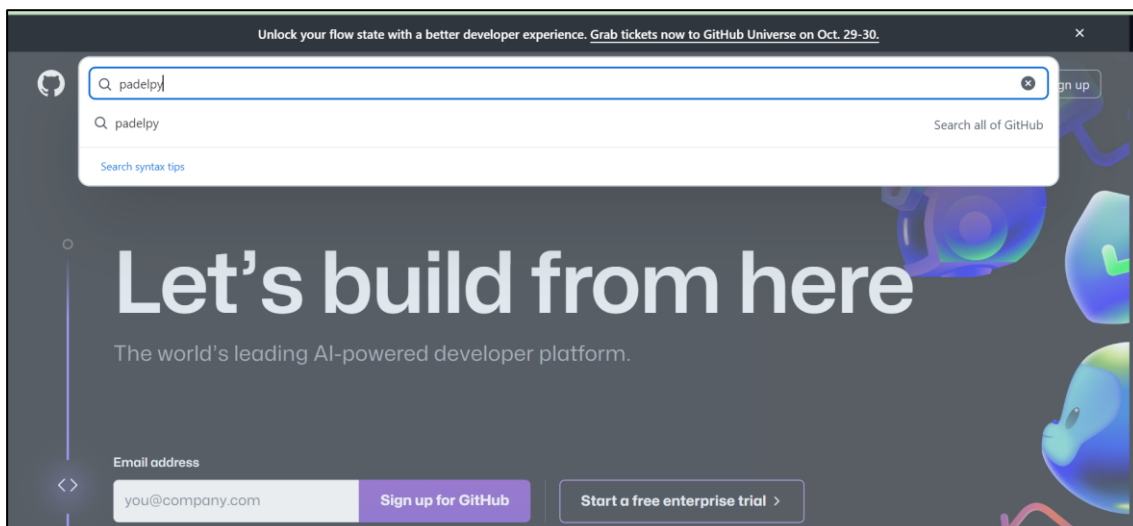
Descriptor Type	Symbol	Definition
Molecular Weight	MW	The sum of the atomic weights of all atoms in a molecule, indicating its size.
Log P (Partition Coefficient)	Log P	A measure of lipophilicity, indicating how a compound partitions between water and octanol.
Topological Polar Surface Area (TPSA)	TPSA	The surface area of polar atoms in a molecule, affecting solubility and permeability.
Molecular Volume	V <sub>m</sub>	The 3D space occupied by a molecule, often computed from van der Waals radii.
Molecular Surface Area	A <sub>s</sub>	The total surface area of a molecule, which can impact interactions with biological targets.
Hydrogen Bond Donor Count	HBD	The number of hydrogen bond donors in a molecule, influencing solubility and interaction.
Hydrogen Bond Acceptor Count	HBA	The number of hydrogen bond acceptors in a molecule, affecting its interaction properties.
Rotatable Bonds	R <sub>B</sub>	The number of rotatable bonds in a molecule, indicating flexibility and conformational change.
Log D (Distribution Coefficient)	Log D	A measure of the distribution of a compound between two phases, considering pH and ionization.
Molecular Shape Index	-	Describes the overall shape of a molecule, which can influence biological activity.
Dipole Moment	μ	A vector quantity that represents the polarity of a molecule, indicating charge distribution.
Polarizability	α	The ability of a molecule to have its electron cloud distorted by an external electric field.
Solvent Accessible Surface Area (SASA)	SASA	The surface area of a molecule that is accessible to solvent, relevant for solubility studies.
Constitutional Index	CI	A measure that considers the connectivity of atoms in a molecule, providing insight into stability.
3D-Radius of Gyration	R <sub>g</sub>	A measure of the distribution of atoms around the molecule's centre of mass, indicating compactness.

### **PaDEL-DESCRIPTOR**

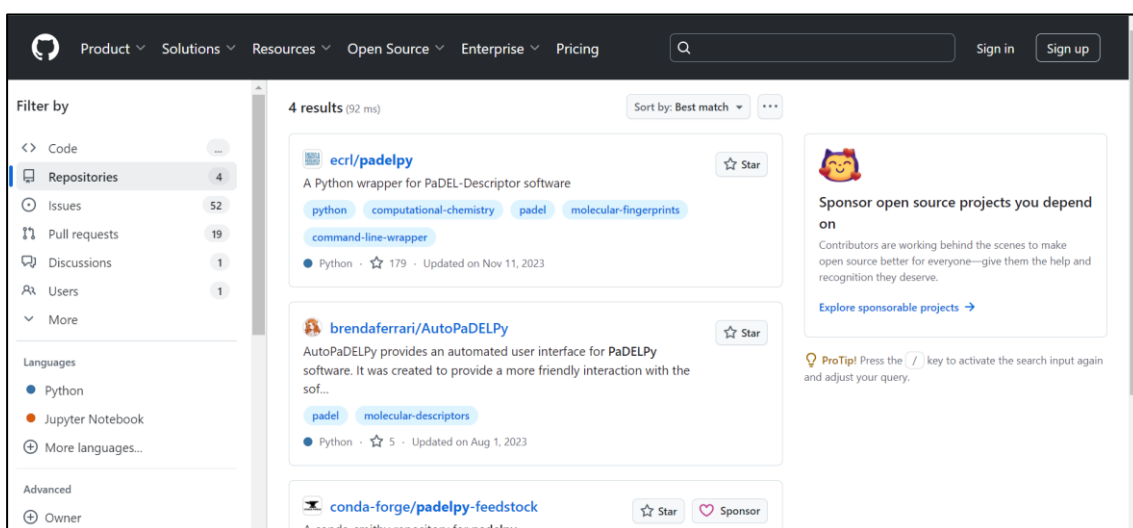
PaDEL-Descriptor is a software for calculating molecular descriptors and fingerprints. The software currently calculates 797 descriptors (663 1D, 2D descriptors, and 134 3D descriptors) and 10 types of fingerprints. These descriptors and fingerprints are calculated mainly using The Chemistry Development Kit. Some additional descriptors and fingerprints were added, which include atom type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures identified by Laggner,

and binary fingerprints and count of chemical substructures identified by Klekota and Roth. Although many descriptors can be calculated using various descriptor calculation software, considering the information available in a PubChem fingerprint, only PubChem fingerprints of the compounds were used to train the model. PaDELPy, a Python wrapper for PaDEL-Descriptor software, was used for calculating the PubChem fingerprints.

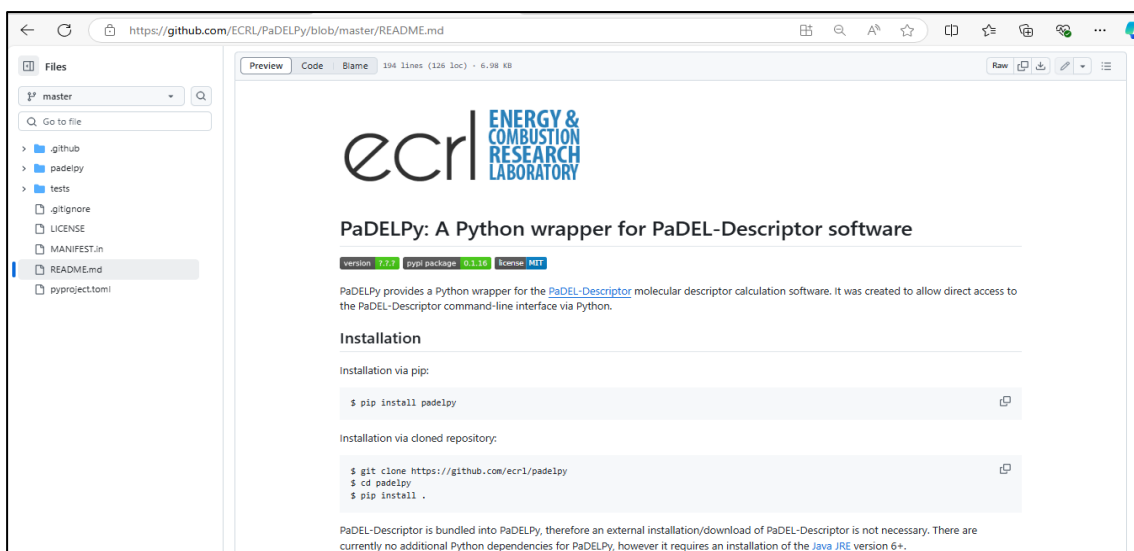
## **INSTALLATION:**



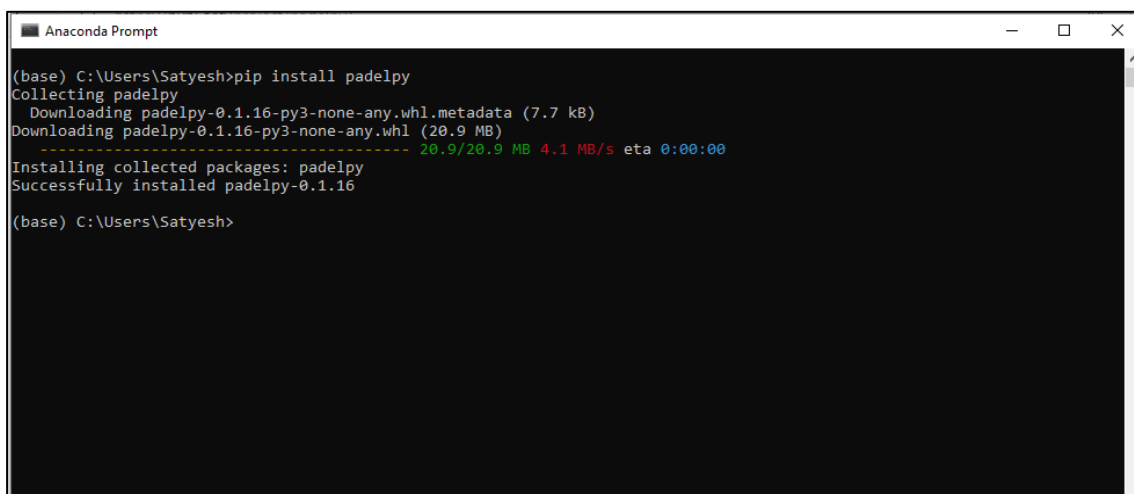
**Fig 1: Open GitHub and search Padelpy**



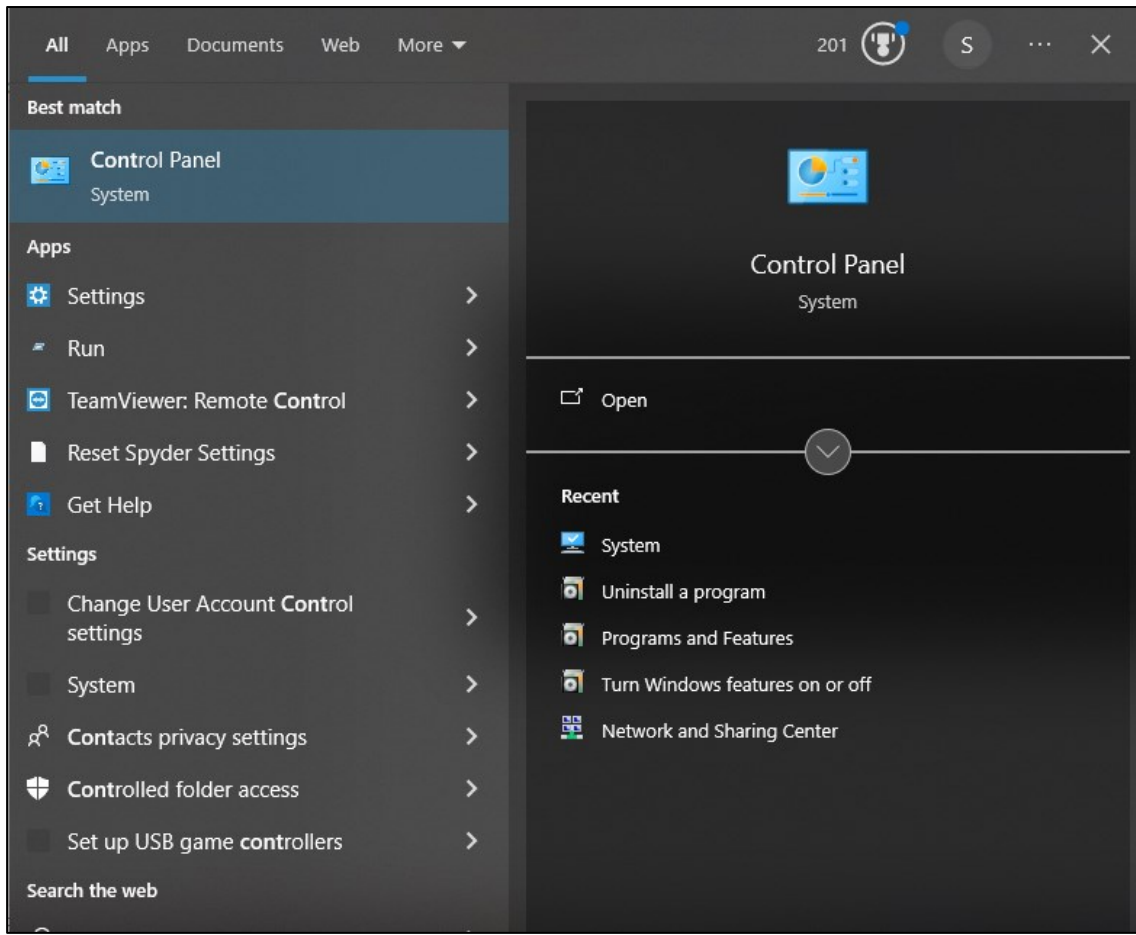
**Fig 2: Open “ercl/Padelpy”**



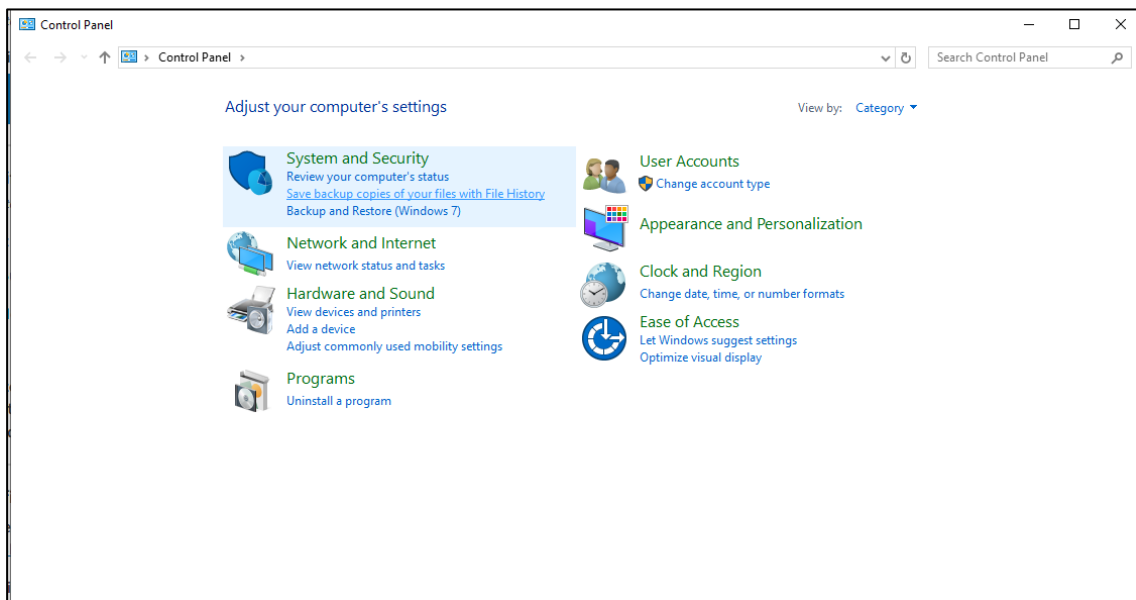
**Fig 3.a: Python Wrapper for PaDEL-Descriptor Software**



**Fig 4: Installing PaDELPy via Anaconda Prompt**

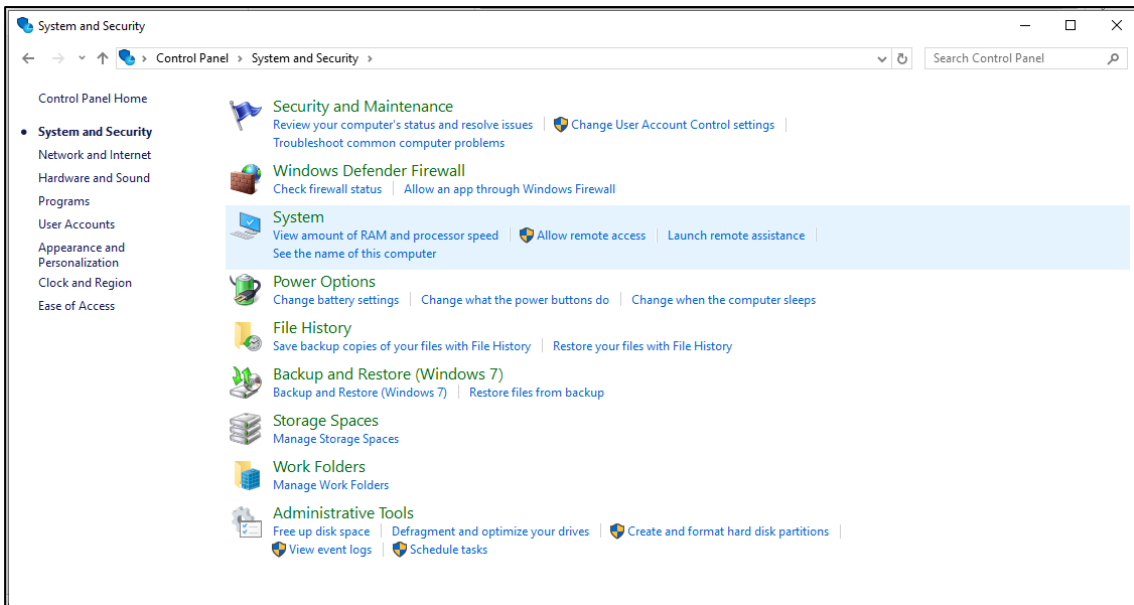


**Fig 5: Open Control Open**

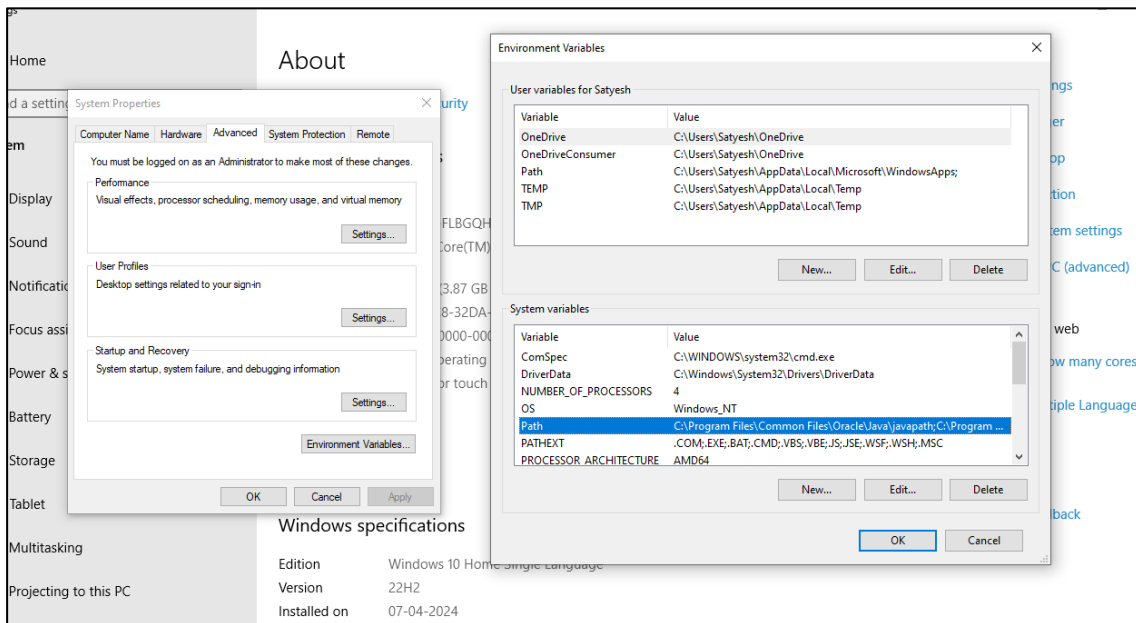


**Fig 6: Select “System and Security”**

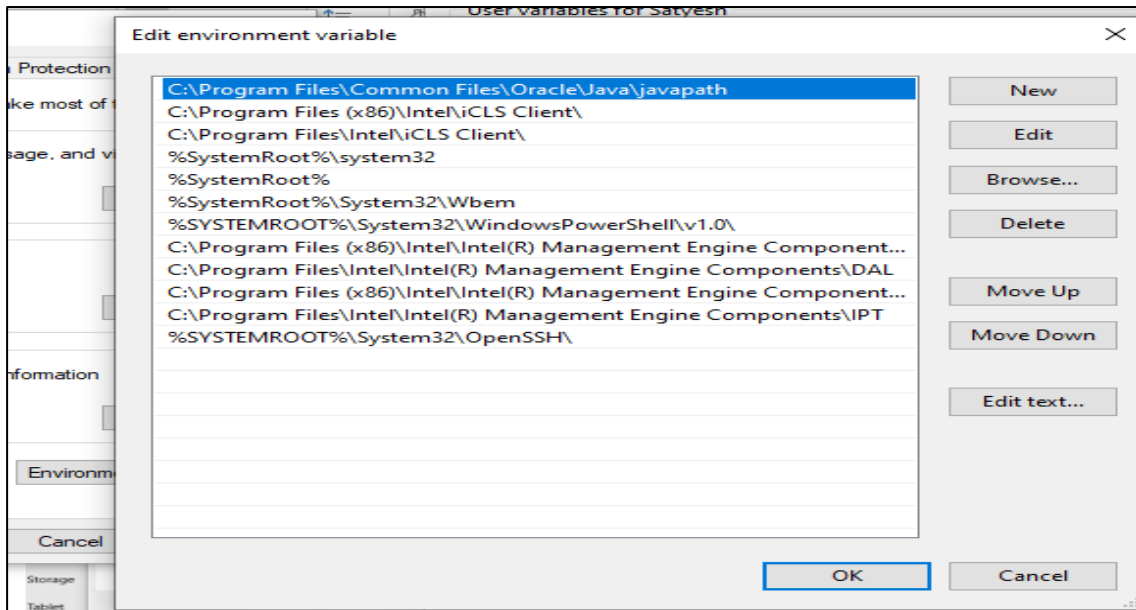




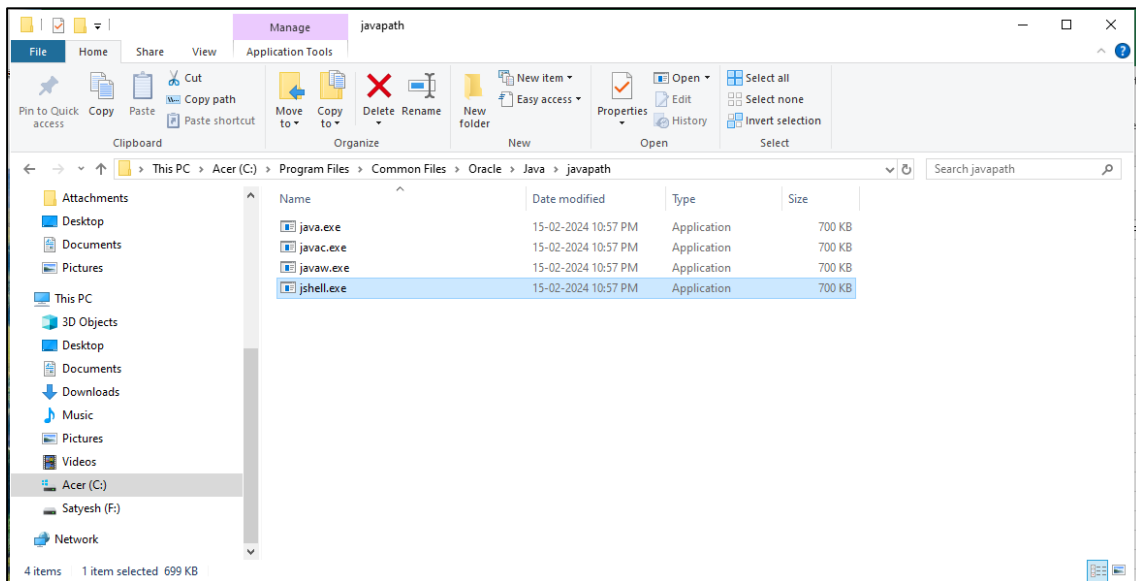
**Fig 7: Select “System”**



**Fig 8: Click on “Environment Variables” and then under “System variables” click on path**



**Fig 9: Add “Python” and “JDK” files path over here and click “OK”**



**Fig 10: PaDELPy installed successfully in assigned path**

## **REFERENCES:**

1. PaDELPy: A Python wrapper for PaDEL-Descriptor software. (2022, April 30). GitHub. <https://github.com/ECRL/PaDELPy>
  2. Danishuddin, & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today*, 21(8), 1291–1302. <https://doi.org/10.1016/j.drudis.2016.06.013>
  3. Yap, C. W. (2010). PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474. <https://doi.org/10.1002/jcc.21707>
-

**DATE: 01/10/2024**

**WEBLEM: 9(A)**

**PaDELPy: A Python wrapper for PaDEL-Descriptor software**

**(URL: <https://github.com/ecrl/padelpy>)**

### **AIM:**

To study ID, 2D & 3D descriptors for “Gallic Acid” (PubChem CID: 370) using PaDELPy Software.

### **INTRODUCTION:**

PaDELPy simplifies the process of calculating molecular descriptors and fingerprints by providing a Python interface for the PaDEL-Descriptor software. PaDEL-Descriptor, written in Java, generates molecular descriptors that are used to describe molecular properties in various computational chemistry and cheminformatics applications. It helps correlate molecular structures with biological activities, which is essential for building machine learning models in scientific research, particularly in drug discovery, toxicology, and other biological studies.

PaDELPy eliminates the need for manual handling of the Java-based PaDEL-Descriptor. This means users no longer must install and execute the Java `.jar` file separately. Instead, the library automates the process, allowing users to calculate molecular fingerprints directly in Python. This reduces the complexity of installation and streamlines the workflow for researchers and data scientists working on chemical data and machine learning model creation.

### **Gallic Acid:**

Gallic acid (3,4,5-trihydroxybenzoic acid) is a naturally occurring polyphenolic compound found in various fruits, vegetables, and herbs. It exhibits strong antioxidant, anti-inflammatory, antimicrobial, and anticancer properties, making it beneficial for health and industrial applications. Gallic acid helps in scavenging free radicals, reducing oxidative stress, and modulating inflammatory responses, with promising effects on gut health, cancer prevention, and managing infections. It is also used in food preservation for its antioxidant activity, as well as in cosmetics and pharmaceuticals for its skin-protective and therapeutic benefits. Ongoing research continues to explore its broader therapeutic potential.

### **METHODOLOGY:**

1. Search for your query in the PubChem database.
2. Retrieve canonical SMILES of the best match.
3. Open PaDELPy in GitHub and copy the code for “SMILES to Descriptors/Fingerprints”.
4. Using Python IDLE:
  - a. Install PaDELPy via pip: `pip install padelpy`
  - b. Paste the code copied and change the name of the query and input the SMILES.
  - c. Run the code and interpret the results in an excel file containing the data for the descriptor.
5. Using Google Colab:

- Install PaDELPy via pip: `pip install padelphy`
- Paste the code copied and change the name of the query and input the SMILES.
- Run the code and interpret the results in table containing the data for the descriptor.

### CODE:

```
from padelphy import from_smiles
```

```
# Calculate molecular descriptors for propane
```

```
descriptors = from_smiles('C1=C(C=C(C(=C1O)O)O)C(=O)O')
```

```
# In addition to descriptors, calculate PubChem fingerprints
```

```
desc_fp = from_smiles('C1=C(C=C(C(=C1O)O)O)C(=O)O', fingerprints=True)
```

```
# Only calculate fingerprints
```

```
fingerprints = from_smiles('C1=C(C=C(C(=C1O)O)O)C(=O)O', fingerprints=True,
descriptors=False)
```

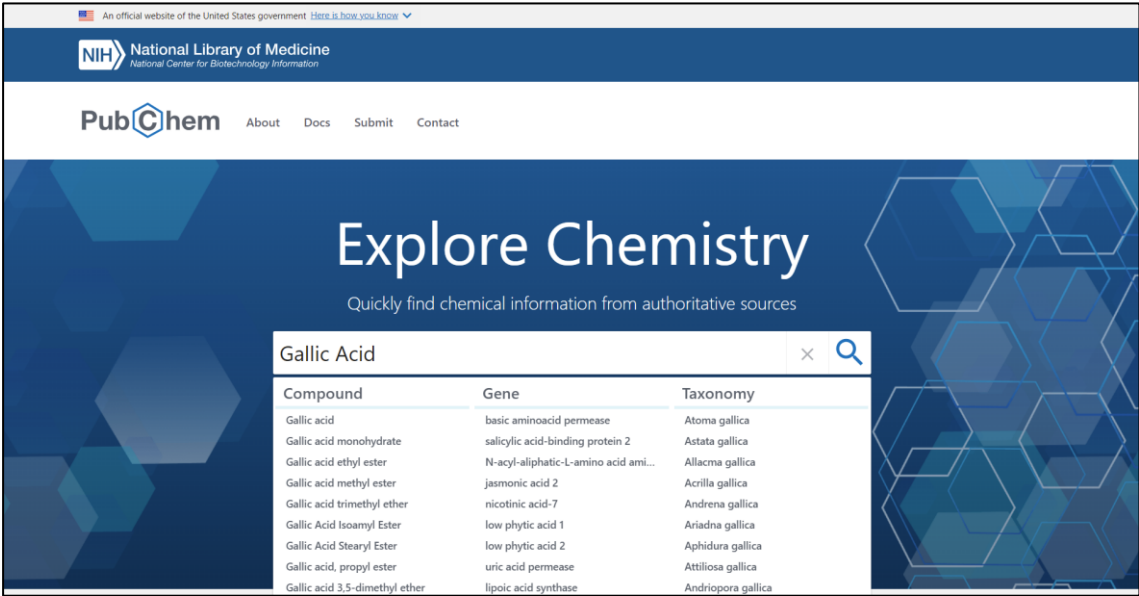
```
# Setting the number of threads, this uses one CPU thread to compute descriptors
```

```
descriptors = from_smiles(['C1=C(C=C(C(=C1O)O)O)C(=O)O'], threads = 1)
```

```
# Save descriptors to a CSV file
```

```
_ = from_smiles('C1=C(C=C(C(=C1O)O)O)C(=O)O', output_csv='descriptors.csv')
```

### OBSERVATIONS:



The screenshot shows the PubChem homepage with a search bar containing "Gallic Acid". Below the search bar is a table with three columns: Compound, Gene, and Taxonomy. The table lists various derivatives of Gallic Acid and their associated genes and taxonomic classifications.

Compound	Gene	Taxonomy
Gallic acid	basic aminoacid permease	Atoma gallica
Gallic acid monohydrate	salicylic acid-binding protein 2	Astata gallica
Gallic acid ethyl ester	N-acyl-aliphatic-L-amino acid ami...	Allacma gallica
Gallic acid methyl ester	jasmonic acid 2	Acrilla gallica
Gallic acid trimethyl ether	nicotinic acid-7	Andrena gallica
Gallic Acid Isoamyl Ester	low phytic acid 1	Ariadna gallica
Gallic Acid Stearyl Ester	low phytic acid 2	Aphidura gallica
Gallic acid, propyl ester	uric acid permease	Attilosa gallica
Gallic acid 3,5-dimethyl ether	lipoic acid synthase	Andriopora gallica

**Fig 14: Homepage for PubChem**

An official website of the United States government <https://www.ssa.gov/privacy>


**NIH** National Library of Medicine  
National Center for Biotechnology Information

**PubChem** About Docs Submit Contact

SEARCH FOR  
**Gallic Acid**

Treating this as a text search.

**BEST MATCH**

 Gallic acid; 149-91-7; 3,4,5-Trihydroxybenzoic acid; gallate; Benzoic acid, 3,4,5-trihydroxy-; Gallic acid, tech.; Kyselina gallova; Pyrogallol-5-carboxylic acid; ...

Compound CID: 370  
 MF: C<sub>7</sub>H<sub>6</sub>O<sub>5</sub> MW: 170.12g/mol  
 IUPAC Name: 3,4,5-trihydroxybenzoic acid  
 Isomeric SMILES: C1=C(C(=C(C(=C1O)O)O)C(=O)O)  
 InChIKey: LNTHITQWFMADLM-UHFFFAOYSA-N  
 InChI: InChI=1S/C7H6O5/c8-4-1-3/7(11)12(2-5)(6(4)10)/h1-2,8-10H,(H,11,12)  
 Create Date: 2004-09-16



Summary Similar Structures Search Related Records PubMed (MeSH Keyword)



Compounds (460) Substances (936) BioAssays (964) Literature (13,576) Patents (852)



**Fig 15: Best Match for the Query- Phencyclidine (CID:6468)**





**PubChem** Gallic Acid (Compound) TOP

PubChem

**2.1.3 InChIKey**    
 LNTHITQWFMADLM-UHFFFAOYSA-N  
 Computed by InChI 1.0.6 (PubChem release 2021.10.14)  
 PubChem

**2.1.4 SMILES**    
C1=C(C(=C(C(=C1O)O)O)C(=O)O  
 Computed by OEChem 2.3.0 (PubChem release 2021.10.14)  
 PubChem

**2.2 Molecular Formula**    
 C<sub>7</sub>H<sub>6</sub>O<sub>5</sub>  
 Computed by PubChem 2.2 (PubChem release 2021.10.14)  
 Australian Industrial Chemicals Introduction Scheme (AICIS); CAMEO Chemicals; PubChem

**2.3 Other Identifiers**    
**2.3.1 CAS**    
 149-91-7

**CONTENTS**

- Title and Summary
- 1 Structures
- 2 Names and Identifiers**
- 3 Chemical and Physical Properties
- 4 Spectral Information
- 5 Related Records
- 6 Chemical Vendors
- 7 Drug and Medication Information
- 8 Food Additives and Ingredients
- 9 Pharmacology and Biochemistry
- 10 Use and Manufacturing
- 11 Identification
- 12 Safety and Hazards
- 13 Toxicity
- 14 Associated Disorders and Diseases
- 15 Literature
- 16 Patents
- 17 Interactions and Pathways
- 18 Biological Test Results
- 19 Taxonomy
- 20 Classification

**Fig 3: SMILES format of the query**

```

README MIT license
SMILES to Descriptors/Fingerprints
The "from_smiles" function accepts a SMILES string or list of SMILES strings as an argument, and returns a Python dictionary with descriptor/fingerprint names/values as keys/values respectively - if multiple SMILES strings are supplied, "from_smiles" returns a list of dictionaries.

from padelpy import from_smiles

# calculate molecular descriptors for propane
descriptors = from_smiles('CCC')

# calculate molecular descriptors for propane and butane
descriptors = from_smiles(['CCC', 'CCCC'])

# in addition to descriptors, calculate PubChem fingerprints
desc_fp = from_smiles('CCC', fingerprints=True)

# only calculate fingerprints
fingerprints = from_smiles('CCC', fingerprints=True, descriptors=False)

# setting the number of threads, this uses one cpu thread to compute descriptors
descriptors = from_smiles(['CCC', 'CCCC'], threads = 1)

# save descriptors to a CSV file
_ = from_smiles('CCC', output_csv='descriptors.csv')

MDL MolFile to Descriptors/Fingerprints
The "from_md1" function accepts a filepath (to an MDL MolFile) as an argument, and returns a list. Each list element is a dictionary with descriptors/fingerprints corresponding to each supplied molecule (indexed as they appear in the MolFile).

from padelpy import from_md1

```

Fig 4: SMILES to Descriptors code on PaDELPy GitHub

```

[1] !pip install padelpy
Collecting padelpy
  Downloading padelpy-0.1.14-py2.py3-none-any.whl.metadata (7.7 kB)
  Downloading padelpy-0.1.14-py2.py3-none-any.whl (20.9 MB)
    20.9/20.9 MB 18.5 MB/s eta 0:00:00
Installing collected packages: padelpy
Successfully installed padelpy-0.1.14

```

Fig 5: Installing padelpy using pip in Google Colab

```

from padelpy import from_smiles

# calculate molecular descriptors for propane
descriptors = from_smiles('C1=C(C=C(C(=C1)O)O)C(=O)O')

# in addition to descriptors, calculate PubChem fingerprints
desc_fp = from_smiles('C1=C(C=C(C(=C1)O)O)C(=O)O', fingerprints=True)

# only calculate fingerprints
fingerprints = from_smiles('C1=C(C=C(C(=C1)O)O)C(=O)O', fingerprints=True, descriptors=False)

# setting the number of threads, this uses one cpu thread to compute descriptors
descriptors = from_smiles(['C1=C(C=C(C(=C1)O)O)C(=O)O'], threads = 1)

# save descriptors to a CSV file
_ = from_smiles('C1=C(C=C(C(=C1)O)O)C(=O)O', output_csv='descriptors.csv')

```

Fig 6: Code for SMILES to Descriptor/Fingerprints

Name	nAcid	ALogP	ALogp2	AMR	apol	naAromAtom	nAromBond	nAtom	nHeavyAtom	nH	nB	nC	nN	nO	nS	nP	nF	r
AUTOGEN_20241004144917322715181604859	1	-0.7213	0.5202736900000001	41.7715	20.330757999999999	6	6	18	12	6	0	7	0	5	0	0	0.0	0

Show 10 per page

**Fig 7: Output in csv format**

## **RESULTS:**

Using PaDELPy, a Python wrapper for the PaDEL-Descriptor, molecular descriptors for the compound "Gallic Acid" were calculated. These descriptors provide key numerical values representing the chemical and structural properties of the molecule. For example:

1. **AlogP:** Represents the lipophilicity (hydrophobicity) of the molecule.
2. **nAtom:** The total number of atoms in the molecule.
3. **nHeavyAtom:** The number of non-hydrogen atoms (heavy atoms) in the molecule.
4. **nH:** The number of hydrogen atoms.
5. **nC:** The number of carbon atoms.

These descriptors, along with many others, help quantify the chemical characteristics of molecules, which is crucial for predicting molecular behaviour in biological systems. This information is particularly useful in computational tasks such as molecular docking simulations, where the binding affinities and interactions between ligands (such as drugs) and biological targets (like proteins) are predicted.

## **CONCLUSION:**

PaDEL-Descriptor, an open-source and multithreaded molecular descriptor calculation software, provides a powerful and efficient tool for extracting molecular descriptors. Its ability to handle large datasets and compute a wide range of molecular descriptors quickly makes it a valuable tool in cheminformatics, drug discovery, and computational biology. With its cross-platform compatibility and ease of use, PaDELPy enhances productivity in scientific research by facilitating the integration of molecular descriptor calculations into machine learning models and other data analysis workflows.

## **REFERENCES:**

1. Dey, V. (2021). Hands-On Guide to PaDELPy for ML Model Building. Analytics India Magazine. <https://analyticsindiamag.com/hands-on-guide-to-padelpy-for-ml-model-building/>
2. Kolmar, S. S., & Grulke, C. M. (2021). The effect of noise on the predictive limit of QSAR 4. models. Journal of Cheminformatics, 13(1). <https://doi.org/10.1186/s13321-021-00571-7>
3. Su, A., & Rajan, K. (2021). A database framework for rapid screening of structure-function relationships in PFAS chemistry. Scientific Data, 8(1). <https://doi.org/10.1038/s41597-021-00798-x>
4. Ecr1. (n.d.). GitHub ecr1/padelpy: A Python wrapper for PaDEL-Descriptor software. GitHub. <https://github.com/ecr1/padelpy>



**WEBLEM:10**

**WEB-BASED TOOLS FOR VACCINE DESIGNING**

**AIM:**

To understand various web-based tools for vaccine designing.

**INTRODUCTION:**

Immunization is a cornerstone of public health policy and is demonstrably highly cost effective when used to protect child health. Although it could be argued that immunology has not thus far contributed much to vaccine development, in that most of the vaccines we use today were developed and tested empirically, there are major challenges ahead to develop new vaccines for difficult-to target pathogens, for which we urgently need a better understanding of protective immunity. Moreover, recognition of the huge potential and challenges for vaccines to control disease outbreaks and protect the older population, together with the availability of an array of new technologies, make it the perfect time for immunologists to be involved in designing the next generation of powerful immunogens. This Review provides an introductory overview of vaccines, immunization and related issues and thereby aims to inform a broad scientific audience about the underlying immunological concepts.

Define Vaccines: A vaccine is a biological product that can be used to safely induce an immune response that confers protection against infection and/or disease on subsequent exposure to a pathogen. To achieve this, the vaccine must contain antigens that are either derived from the pathogen or produced synthetically to represent components of the pathogen. The essential component of most vaccines is one or more protein antigens that induce immune responses that provide protection. However, polysaccharide antigens can also induce protective immune responses and are the basis of vaccines that have been developed to prevent several bacterial infections, such as pneumonia and meningitis caused by *Streptococcus pneumoniae*, since the late 1980s. Protection conferred by a vaccine is measured in clinical trials that relate immune responses to the vaccine antigen to clinical end points. Vaccines are generally classified as live or non-live (sometimes loosely referred to as 'inactivated') to distinguish those vaccines that contain attenuated replicating strains of the relevant pathogenic organism from those that contain only components of a pathogen or killed whole organisms. In addition to the 'traditional' live and non-live vaccines, several other platforms have been developed over the past few decades, including viral vectors, nucleic acid-based RNA and DNA vaccines, and virus like particles.

**HISTORY:**

Epidemics of smallpox swept across Europe in the seventeenth and eighteenth centuries, accounting for as much as 29% of the death rate of children in London<sup>137</sup>. Initial efforts to control the disease led to the practice of variolation, which was introduced to England by Lady Mary Wortley Montagu in 1722, having been used in the Far East since the mid-1500s. In variolation, material from the scabs of smallpox lesions was scratched into the skin to provide protection against the disease. Variolation did seem to induce protection, reducing the attack rate during epidemics, but sadly some of those who were variolated developed the disease and

sometimes even died. It was in this context that Edward Jenner wrote an Inquiry into the Causes and Effects of the Variole Vaccine in 1798. His demonstration, undertaken by scratching material from cowpox lesions taken from the hands of a milkmaid, Sarah Nelms, into the skin of an 8-year-old boy, James Phipps, who he subsequently challenged with smallpox, provided early evidence that vaccination could work.

Jenner's contribution to medicine was thus not the technique of inoculation but his startling observation that milkmaids who had had mild cowpox infections did not contract smallpox, and the serendipitous assumption that material from cowpox lesions might immunize against smallpox. Furthermore, Jenner brilliantly predicted that vaccination could lead to the eradication of smallpox; in 1980, the World Health Assembly declared the world free of naturally occurring smallpox. Almost 100 years after Jenner, the work of Louis Pasteur on rabies vaccine in the 1880s heralded the beginning of a frenetic period of development of new vaccines, so that by the middle of the twentieth century, vaccines for many different diseases (such as diphtheria, pertussis, and typhoid) had been developed as inactivated pathogen products or toxoid vaccines. However, it was the coordination of immunization as a major public health tool from the 1950s onwards that led to the introduction of comprehensive vaccine programs and their remarkable impact on child health that we enjoy today. In 1974, the World Health Organization launched the Expanded Program on Immunization and a goal was set in 1977 to reach every child in the world with vaccines for diphtheria, pertussis, tetanus, poliomyelitis, measles, and tuberculosis by 1990. Unfortunately, that goal has still not been reached; although global coverage of 3 doses of the diphtheria–tetanus–pertussis vaccine has risen to more than 85%, there are still more than 19 million children who did not receive basic vaccinations in 2019.

### **MATERIALS AND METHODS:**

Vaccines induce antibodies: The adaptive immune response is mediated by B cells that produce antibodies (humoral immunity) and by T cells (cellular immunity). All vaccines in routine use, except BCG which is believed to induce T cell responses that prevent severe disease and innate immune responses and are thought to mainly confer protection through the induction of antibodies. There is considerable supportive evidence that various types of functional antibody are important in vaccine induced protection, and this evidence comes from three main sources: immunodeficiency states, studies of passive protection and immunological data.

Vaccines need T cell help: The role of T cells in protection is poorly characterized, except for their role in providing help for B cell development and antibody production in lymph nodes. From studies of individuals with inherited or acquired immunodeficiency, it is clear that whereas antibody deficiency increases susceptibility to acquisition of infection, T cell deficiency results in failure to control a pathogen after infection. For example, T cell deficiency results in uncontrolled and fatal varicella zoster virus infection, whereas individuals with antibody deficiency readily develop infection but recover in the same way as immunocompetent individuals. The relative suppression of T cell responses that occurs at the end of pregnancy increases the severity of infection with influenza and varicella zoster viruses. Studies show that sterilizing immunity against carriage of *S. pneumoniae* in mice can be achieved by the transfer of T cells from donor mice exposed to *S. pneumoniae*, which indicates that further investigation of T cell-mediated immunity is warranted to better understand the nature of T cell responses that could be harnessed to improve protective immunity.

Although somewhat simplistic, the evidence therefore indicates that antibodies have the major role in prevention of infection (supported by TH cells), whereas cytotoxic T cells are required to control and clear established infection.

Epitope-based vaccines: Epitopes are of particular interest to both clinical and basic biomedical researchers as they hold huge potential for vaccine design, disease prevention, diagnosis, and treatment. Using rDNA technologies, we can isolate specific epitopes which can replace the whole pathogen in a vaccine. However, within the diversity of epitopes in a pathogen, it is important to notice that not all the epitopes, even those that seem to be dominant, are equal in their ability to elicit antibody production. The proteins that contain many epitopes recognized by the common MHC alleles are known as promiscuous binders. The human leukocyte antigen (HLA) supertype refers to a set of HLA alleles with overlapping peptide binding specificities. The alleles in the given HLA supertype often represent the same epitope, which refers to the region on the surface of an antigen capable of eliciting an immune response for T cell recognition. On the other hand, elicitation of humoral responses relies on the recognition of linear epitopes and conformational epitopes. The latter constitute a challenge for chimeric vaccine design as they must retain their native conformation to be functional. Therefore, knowledge on the whole antigen structure is necessary to aid in the rational design of vaccines targeting conformational B cell epitopes.

3. Bioinformatics tools to prediction of potential T cell binding-epitopes: The first step on applying bioinformatics to vaccine development consists of discriminating epitopes that are potentially immune-protective from epitopes that are not. Since T-cell epitopes are bound in a linear form to MHCs, the interface between ligands and T-cells can be modeled with accuracy. It is currently well known that epitopes link together into the binding groove of MHC Class I and Class II molecules through interactions between their R group side chains and pockets located on the floor of the MHC. Based on this knowledge, many T-cell epitope-mapping algorithms have been established and used to develop tools to rapidly identify putative T-cell epitopes. MHC-I binding predictors are currently very efficient and have wide allelic coverage, a prediction accuracy in the range of 90–95% positive predictive value has been estimated.

Among the numerous servers for MHC-I alleles is RANKPEP, which predicts peptide binders to MHC-I and MHC-II molecules from protein sequences or sequence alignments using Position Specific Scoring Matrices (PSSMs). In addition, it predicts those MHC-I ligands whose C-terminal end is likely to be the result of proteasomal cleavage. This is a friendly platform which offers the widest allelic coverage to MHC-I and MHC-II alleles for humans and mice. To search epitopes sequences for MHC-I ligands using PSSMs, a dynamic algorithm written in Python is used; which scores all protein segments with the length of the PSSM width and sorts them accordingly. Scoring starts at the beginning of each sequence and the PSSM is slid over the sequence one residue at a time until reaching the end of the sequence. Furthermore, to narrow down the potential binders from the list of ranked peptides, a binding threshold is defined as the score value that includes 90% of the peptides within the PSSM. This binding threshold is built into each matrix, delineating the range of putative binders among the top scoring peptides.

Bioinformatics tools for predicting potential B cell binding epitopes: B cell epitopes are recognized by B cell receptors or antibodies in their native structure. Continuous B cell epitope prediction is very similar to T cell epitope prediction, which has mainly been based on the

amino acid properties such as hydrophilicity, charge, exposed surface area and secondary structure. Discontinuous B cell epitope prediction requires 3D structure of the antigen. Some specific resources to predict continuous or discontinuous B-cell epitopes are available on the Web. To predict linear B-cell epitopes, the Bcepred tool is based on physicochemical properties such as hydrophilicity, flexibility, polarity, and exposed surface on a non-redundant dataset. The dataset consists of 1029 B-cell epitopes obtained from the Bcipep database and an equal number of non-epitopes obtained randomly from the Swiss-Prot database. The prediction accuracy for models based on these properties varies from 52.92% to 57.53%.

The ABCpred server, which is based on neural networks, has an estimated accuracy of 65.93%. Another server called BepiPred predicts the location of linear B-cell epitopes using a combination of a hidden Markov model and a propensity scale method. The servers mentioned above are easy to use and properly organized. Among the tools used to predict discontinuous B cell epitopes we can mention DiscoTope, which uses the three-dimensional structure of proteins to determine the surface accessibility and a novel epitope propensity amino acid score. The final scores are calculated by combining the propensity scores of residues in spatial proximity and the contact numbers. This server also predicts epitopes in complexes of multiple chains. This tool along with BEpro (formerly known as PEPITO) and SEPPA (Spatial Epitope Prediction of Protein Antigens) requires a 3-D structure as input, specifically, in PDB format. Using SEPPA, each residue in the query protein will be given a score according to information from its neighborhood residues. The higher score corresponds to the higher probability of the residue to be involved in an epitope. One of the most complete tools in this field is ElliPro. This server predicts linear and discontinuous epitopes based on a protein antigen's 3D structure. ElliPro associates each predicted epitope with a score, defined as a PI (Protrusion Index) value. Compared with databases mentioned earlier, in ElliPro the input is a protein sequence. A 3-D structure will be predicted for the input protein sequence by homology modeling based on user selected structural template. Afterwards, linear, and discontinuous epitopes will be computed based on the predicted protein structure. All these integrative tools represent an opportunity for the development of new vaccines, in special those that aim at the elicitation of humoral responses.

Bioinformatics strategies for emergent peptide-based vaccines against hypervariable viruses. Historically, most of the known successful vaccines have been developed empirically. However, the emergence of highly sophisticated viruses, such as HIV and influenza characterized by having a high degree of genetic and antigenic diversity, has impeded the development of effective, broad-coverage vaccines using traditional methods. The rapid emergence of these viral pathogens underscores the need for improved and accelerated processes to develop and produce vaccines, a need that can be addressed by the methods described above allowing a rapid, in silico-based approach to formulate vaccine candidates. This section briefly discusses some approaches developed for the case of the human immunodeficiency (HIV) and influenza viruses as examples on how successful candidate vaccine design can be achieved in the case of hypervariable viruses using bioinformatic tools.

### **PERSPECTIVES:**

Bioinformatics tools have enabled the capability of selecting potential epitopes without running the risks involved in cultivating the pathogen of interest. This kind of methodology represents a huge advantage over conventional vaccinology techniques, including faster outputs and lower costs. The application of omics technologies to this field has also

revolutionized the way in which potential vaccine candidates can be identified. Proteomics and transcriptomics have been used as complementary approaches to genomics and are often more useful in identifying surface proteins during host pathogen interaction. Despite that numerous epitope prediction methods are available, developing a systematic assessment of different methods on standard benchmark datasets is still a need.

Launching a Critical Assessment of Techniques for Epitope Prediction will indeed benefit the field. It has been proposed that computational methods will be used to perform blinded de novo epitope prediction from query proteins previously screened experimentally. Comparison of different methods is yet a complex task due to many aspects including the following:

1. inadequate documentation of datasets and prediction methods
2. unavailability of the benchmark dataset used to evaluate the methods
3. unavailability of the code that implements the method
4. the lack of a unified output format, which complicates the process of combining the results of several servers to obtain consensus predictions.

Therefore, it is necessary to develop standardized data representations; this will enable the evaluation of different prediction methods on a standardized benchmark datasets to compare the methods and develop meta-servers combining the predictions of multiple prediction tools.

## **CONCLUSIONS AND FUTURE DIRECTIONS:**

Immunization protects populations from diseases that previously claimed the lives of millions of individuals each year, mostly children. Under the United Nations Convention on the Rights of the Child, every child has the right to the best possible health, and by extrapolation a right to be vaccinated. Despite the outstanding success of vaccination in protecting the health of our children, there are important knowledge gaps and challenges to be addressed. An incomplete understanding of immune mechanisms of protection and the lack of solutions to overcome antigenic variability have hampered the design of effective vaccines against major diseases such as HIV/AIDS and TB. Huge efforts have resulted in the licensing of a partially effective vaccine against malaria, but more effective vaccines will be needed to defeat this disease. Moreover, it is becoming clear that variation in host response is an important factor to consider.

New technologies and analytical methods will aid the delineation of the complex immune mechanisms involved, and this knowledge will be important to design effective vaccines for the future. Apart from the scientific challenges, sociopolitical barriers stand in the way of safe and effective vaccination for all. Access to vaccines is one of the greatest obstacles, and improving infrastructure, continuing education, and enhancing community engagement will be essential to improve this, and novel delivery platforms that eliminate the need for a cold chain could have great implications. There is a growing subset of the population who are skeptical about vaccination and this requires a response from the scientific community to provide transparency about the existing knowledge gaps and strategies to overcome these.

Constructive collaboration between scientists and between scientific institutions, governments and industry will be imperative to move forwards. The COVID-19 pandemic has indeed shown that, in the case of an emergency, many parties with different incentives can come together to ensure that vaccines are being developed at unprecedented speed but has also highlighted some of the challenges of national and commercial interests. As immunologists, we have a responsibility to create an environment where immunization is

normal, the science is accessible and robust, and access to vaccination is a right and expectation.

### **REFERENCES:**

1. Pollard, Andrew J., and Else M. Bijker. “Publisher Correction: A Guide to Vaccinology: From Basic Principles to New Developments.” *Nature Reviews Immunology*, vol. 21, no. 2, 5 Jan. 2021, pp. 129–129, <https://doi.org/10.1038/s41577-020-00497-5>
  2. Soria-Guerra, Ruth E., et al. “An Overview of Bioinformatics Tools for Epitope
  3. Prediction: Implications on Vaccine Development.” *Journal of Biomedical Informatics*, vol. 53, Feb. 2015, pp. 405–414, <https://doi.org/10.1016/j.jbi.2014.11.003>
-

**DATE: 28/09/2024**

**WEBLEM: 11**

**INTRODUCTION TO TEPITOOL**

**(URL: <http://tools.iedb.org/tepitool/>)**

**AIM:**

Introduction to IEDB Database for prediction of cytotoxic and helper T cell epitopes (MHC Class I epitopes and MHC Class II epitopes).

**INTRODUCTION:**

TepiTool is a powerful online platform developed as part of the Immune Epitope Database (IEDB) to facilitate the prediction of T-cell epitope candidates through the analysis of peptide binding to Major Histocompatibility Complex (MHC) class I and class II molecules. Accurate prediction of peptide-MHC binding is critical for understanding T-cell responses, which play a central role in the immune system's ability to recognize and respond to pathogens, cancer cells, and other antigens. This capability is particularly relevant for applications in vaccine design, immunotherapy, and diagnostics. TepiTool was created to address the growing need for accessible, accurate, and easy-to-use tools in epitope prediction, helping immunologists and researchers identify peptides that can potentially elicit immune responses. MHC molecules present peptides to T cells, activating an immune response when the peptide fits well in the MHC molecule's binding groove. MHC class I molecules bind shorter peptides (8-11 amino acids) and are recognized by CD8+ T cells, while MHC class II molecules bind longer peptides (12-20 amino acids) and are recognized by CD4+ T cells. Predicting which peptides will successfully bind to these MHC molecules is a vital step in identifying epitopes—regions of antigens that are recognized by T cells. This information is crucial in designing vaccines and therapies that target specific immune responses.

TepiTool simplifies this complex task by integrating state-of-the-art computational algorithms for MHC binding prediction, ensuring that researchers can efficiently identify candidate epitopes for further experimental validation. A key advantage of TepiTool is its user-friendly, step-by-step interface, which allows users to easily input amino acid sequences and specify parameters such as the species of interest, MHC class, and peptide length. This is particularly helpful for researchers unfamiliar with computational prediction tools, as it guides them through the entire process, ensuring accurate results without requiring extensive technical expertise. Furthermore, TepiTool supports predictions for hundreds of MHC alleles across multiple species, including humans and common model organisms like mice and pigs, making it highly versatile for a wide range of immunological studies. In addition to its simplicity and accessibility, TepiTool is equipped with some of the most advanced MHC binding prediction algorithms, including artificial neural networks and machine learning techniques. These algorithms have been refined to provide accurate and reliable predictions of peptide-MHC interactions, ensuring that researchers can quickly and effectively identify the most promising epitopes for further investigation. TepiTool also allows users to customize input parameters, such as MHC allele selection and peptide binding thresholds, offering flexibility to meet the specific needs of various research projects. TepiTool has become an essential resource for immunology research, offering applications in vaccine development, cancer immunotherapy,

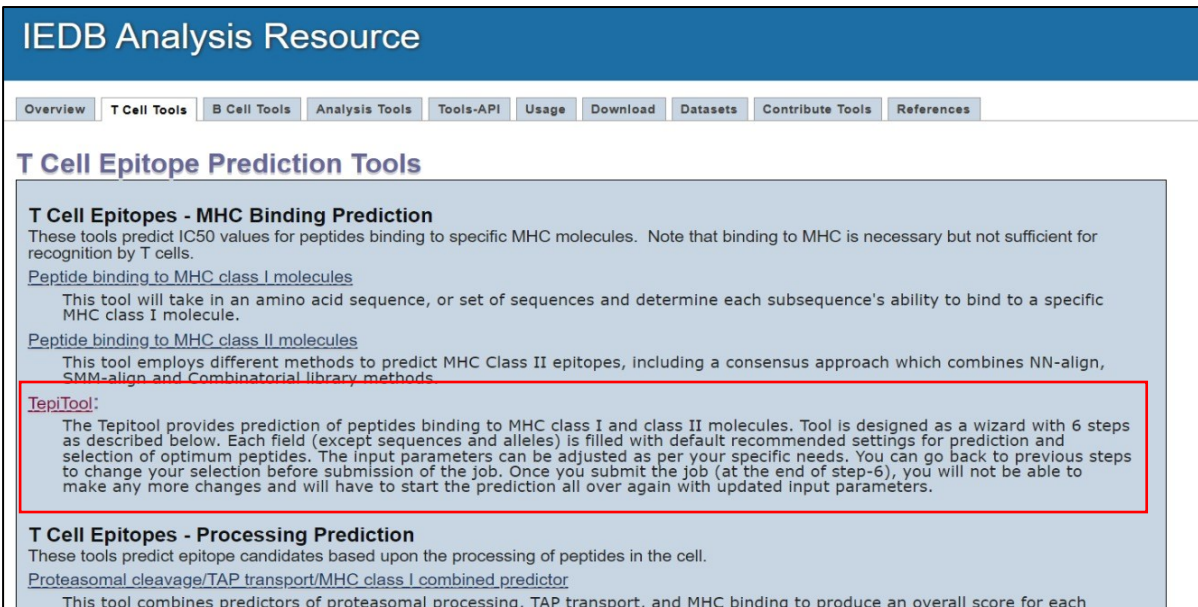
autoimmune disease studies, and diagnostics. By facilitating the prediction of T-cell epitopes, TepiTool enables researchers to accelerate the discovery of novel immunotherapies and vaccines that can precisely target disease mechanisms. Its combination of cutting-edge algorithms and an intuitive interface ensures that it remains a key tool for both experienced researchers and those new to the field of epitope prediction.

The screenshot shows the IEDB homepage with a navigation bar (Home, Specialized Searches, Analysis Resource) and a main content area. A red box highlights the 'Epitope Analysis Resource' section on the right, which includes 'T Cell Epitope Prediction' and 'B Cell Epitope Prediction' options. The 'START YOUR SEARCH HERE' section contains filters for Epitope, Assay, Epitope Source, MHC Restriction, Host, and Disease. A 'Summary Metrics' table is located in the bottom left.

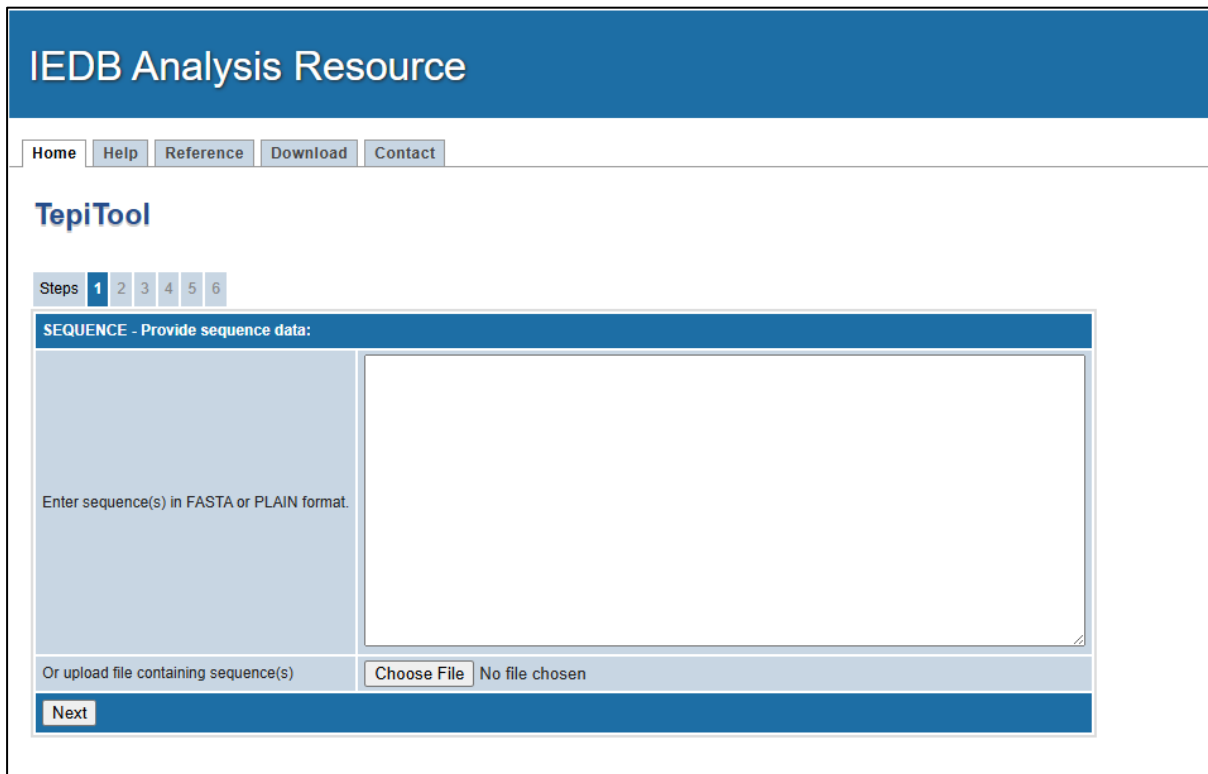
Metric	Count
Peptidic Epitopes	1,620,228
Non-Peptidic Epitopes	3,188
T Cell Assays	539,982
B Cell Assays	1,410,021
MHC Ligand Assays	4,881,364
Epitope Source Organisms	4,543
Restricting MHC Alleles	1,011
References	25,185

**Fig 1: Homepage of IEDB and Select Epitope analysis resource**





**Fig 2: Selection of Tepi Tool from T cell Epitope of IEDB analysis resource**



**Fig 3: Homepage of TEPI TOOL**

**Basic Protocol:**

Computational Prediction of Peptides Binding to MHC Class I AND Class II Molecules.

This protocol explains prediction of T cell epitope candidates from a given set of amino acid sequences, based on predicted peptide binding to MHC class I and class II molecules, using the online computational MHC binding prediction tool called TepiTool. The tool is designed as a wizard where the user is led through a series of well-defined steps to complete the task. Each step is a client-side web form that takes user input data that is in turn processed at the server-

side when the user submits the entire form. All fields except sequences and alleles are filled with default recommended settings for prediction and selection of optimum peptides. The input parameters can be adjusted as per the user's specific needs, and the user can go back to previous steps to change the selection before final submission of the job. The TepiTool has six steps as described below.

### **Protocol steps to be followed:**

1. Provide sequence data
2. Select the host species and MHC allele class
3. Select the alleles for binding prediction
4. Select peptides to be included in prediction
5. Select preferred methods for binding prediction and peptide selection and cutoff values
6. Review selections, enter job details and submit data

### **Step 1: Provide sequence data**

Users input protein sequences in single-letter amino acid code, either by direct entry or by uploading a FASTA file. This step is foundational as it determines the specific proteins to be analyzed for potential T cell epitopes, setting the stage for all subsequent predictions.

### **Step 2: Select Host Species and MHC Allele Class**

Users choose the species (e.g., human, mouse) and MHC class (I or II) relevant to their research. This selection is crucial because it dictates which MHC alleles will be considered in binding predictions, impacting the relevance of the results to the specific biological context.

### **Step 3: Select Alleles for Binding Prediction**

Users specify which MHC alleles from the chosen species will be analyzed. Tailoring the analysis to specific alleles allows for more precise predictions that are relevant to the target population or experimental model.

### **Step 4: Select Peptides to Include in Prediction**

Users can define peptide lengths and whether to include duplicates in their analysis. This step enables users to control the dataset's size and composition, which can significantly influence the quality and interpretability of the prediction results.

### **Step 5: Select Preferred Methods for Binding Prediction**

Users choose algorithms for predicting binding affinities and set parameters like cutoff values for peptide selection. Customizing prediction methods allows users to optimize results based on their specific research needs or hypotheses, enhancing the accuracy of epitope identification.

### **Step 6: Review Selections and Submit Data**

Users review all inputs, enter job details, and submit data for processing. This final step confirms that all parameters are correct before running predictions, ensuring that users have control over their analysis and can avoid errors before submission.

## **REFERENCE:**

1. Paul, S., Sidney, J., Sette, A., & Peters, B. (2016). TEPiTool: A pipeline for Computational Prediction of T cell epitope Candidates. *Current Protocols in Immunology*, 114(1). <https://doi.org/10.1002/cpim.12>
  2. Lefranc, M., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., & Duroux, P. (2008). IMGT(R), the international ImMunoGeneTics information system(R). *Nucleic Acids Research*, 37(Database), D1006–D1012. <https://doi.org/10.1093/nar/gkn838>
-

**DATE: 28/9/2024**

**WEBLEM: 11(A)**

**TEPITOOL**

**(URL:<http://tools.iedb.org/tepitool/>)**

**AIM :**

To Predict MHC Class I and Class II Molecules for Query Dopamine (accession no: P09172) using TepiTool.

**INTRODUCTION:**

T-cell epitope prediction plays a crucial role in a variety of applications, including vaccine discovery, diagnostic development, and mitigating immune responses against therapeutic proteins. Despite ongoing advancements in MHC binding prediction tools, their widespread adoption among immunologists has been slow. This is primarily due to the lack of intuitive interfaces and clear guidance regarding key aspects such as allele selection, peptide lengths, and appropriate cutoff values. Current tools often provide minimal advice on these important factors, leaving users without the necessary insights to optimize their predictions.

To address these challenges, TepiTool—a newly developed online resource available through the Immune Epitope Database (IEDB)—offers a user-friendly interface and integrates top-performing MHC binding prediction algorithms. Designed to simplify the prediction process, TepiTool supports multiple species, including humans, chimpanzees, bovines, gorillas, macaques, mice, and pigs, making it a versatile tool for researchers. With step-by-step instructions and built-in recommendations, TepiTool streamlines the identification of optimal T-cell epitope candidates. Freely accessible at TepiTool, this tool enables immunologists and researchers to efficiently predict and analyze T-cell epitopes, enhancing applications in immunotherapy, vaccine development, and immune response modulation.

**Dopamine**

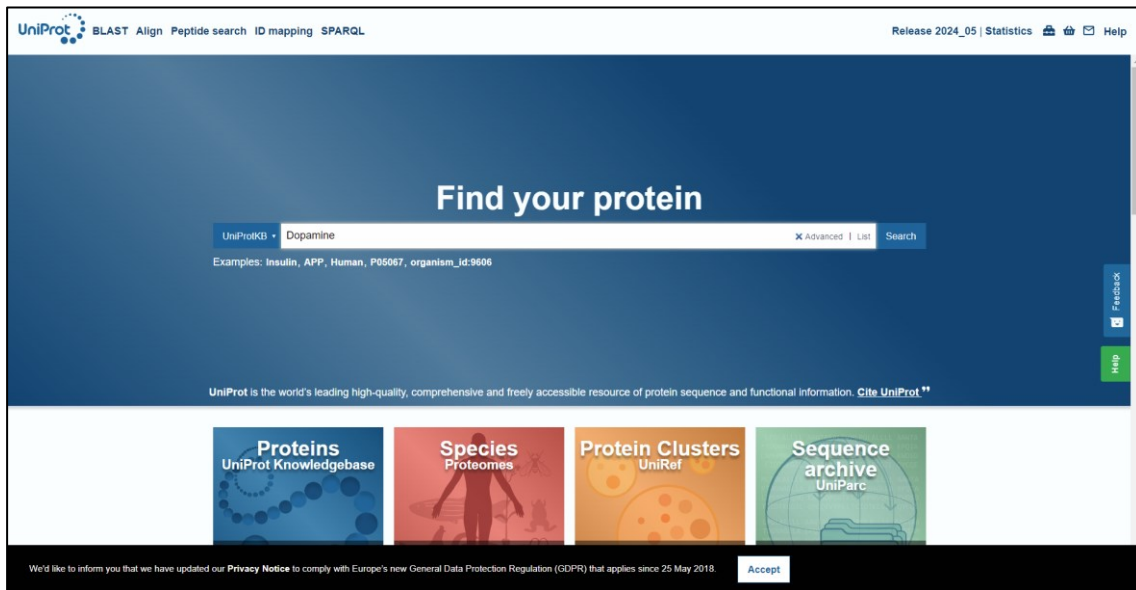
Dopamine is a key neurotransmitter derived from the amino acid tyrosine, playing a critical role in mood regulation, motivation, motor control, and cognitive functions. It transmits signals in the brain, particularly in pathways associated with reward and pleasure, influencing behaviors related to reinforcement and decision-making. Dopamine is essential for smooth motor control, and its deficiency is linked to movement disorders such as Parkinson's disease, while its dysregulation contributes to mental health conditions like depression, ADHD, and schizophrenia. Additionally, dopamine is central to addiction mechanisms, where its elevated levels reinforce drug-seeking behavior. Understanding dopamine's functions and imbalances is crucial for addressing various neurological and psychological disorders.

**METHODOLOGY:**

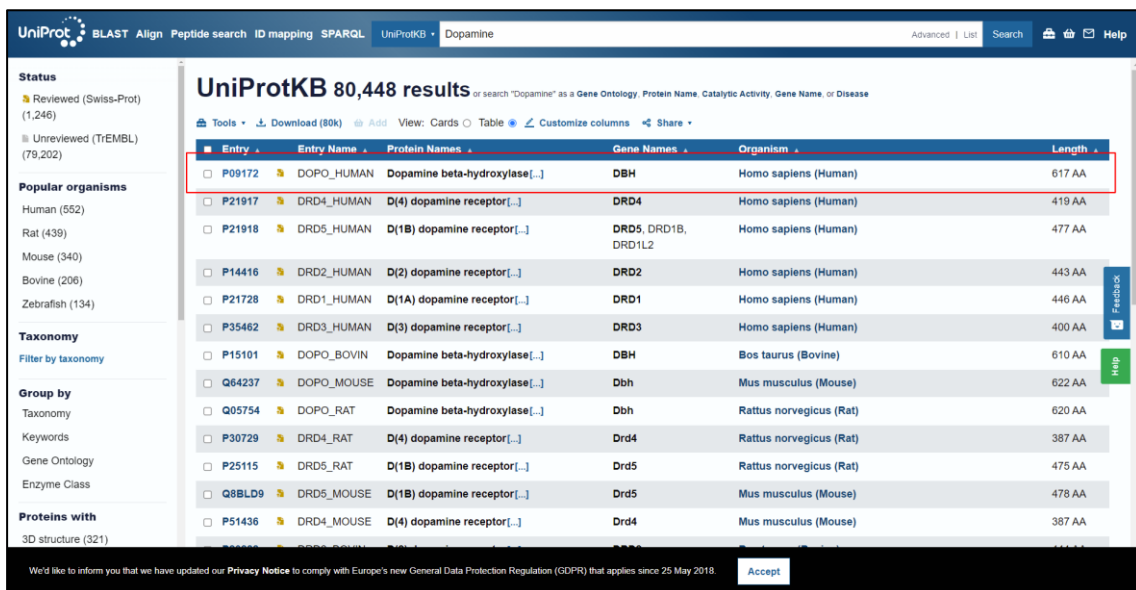
1. Access the UniProt database and search for the Dopamine.
2. Locate the epitope of interest and copy its FASTA sequence.
3. Navigate to the TepiTool server and paste the copied FASTA sequence of the Dopamine  
Provide the sequence data.
4. Select the host species and MHC allele class.

5. Choose the alleles for prediction.
6. Select the peptides to be included in the prediction.
7. Select for preferred methods for binding prediction, peptide selection strategy, and cutoff values.
8. Review your selections, enter job details, and submit the data.

**OBSERVATIONS:**



**Fig 1: Homepage of UniProt Database with entered query Dopamine**



**Fig 2: Selected the entry of Accession ID: P09172**

```
>sp|P09172|DOPO_HUMAN Dopamine beta-hydroxylase OS=Homo sapiens OX=9606 GN=DBH PE=1 SV=3
MPALSRWASLPGPSMREAAFMYSTAVAI FLVILVAALQGSAPRESPLPYHIPLDPEGSLE
LSWNVSYTQEA IHFQLLVRR LKAGV LFGMSDRGELENADLVVLWTDGDTAYFADAWSDQK
GQIHLDPQQDYQL LQVQRTPEGLTLLFKRPFGTCDPKDYLI EDGTVHLVYGI LEEPFRSL
EAINGSG LQMGLQRVQL LKPN IPEPELPSDACTMEVQAPNIQIP SQETTYWCYIKELPKG
FSRHIIKYEP IVTKGNEALVHHMEVFQCAPEMDSVPHFSGPCDSKMKPDRLNYCRH VLA
AWALGAKAFYYPEEAGLAFGGPGSSRYLRLEVHYHNPLVIEGRNDSSGIRLYYTAKLRRF
NAGIMELGLVYTPVMAIPPRETAFIL TGYCTDKCTQLALPPSGIHFASQLHHTL TGRKV
VTVLVRDGREWEIVNQDNHYS PHFQEIRMLKKVVS VHPGDVLTISCTYNTEDRELATVGG
FGILEEMCVNYVHYYPQTQLELCKSAVDAGFLQKYFHLINRFNNE DVCTCPQASVSQQFT
SVPWN SFNRDVLKALYSFAPISMHCNKSSAVRFQGEWNLQPLPKVISTLEEPTQCPTSQ
GRSPAGPTVVSIGGGK
```

**Fig 3: Copy the FASTA sequence**

IEDB Analysis Resource

Home Help Reference Download Contact

### TepiTool

Steps 1 2 3 4 5 6

SEQUENCE - Provide sequence data:

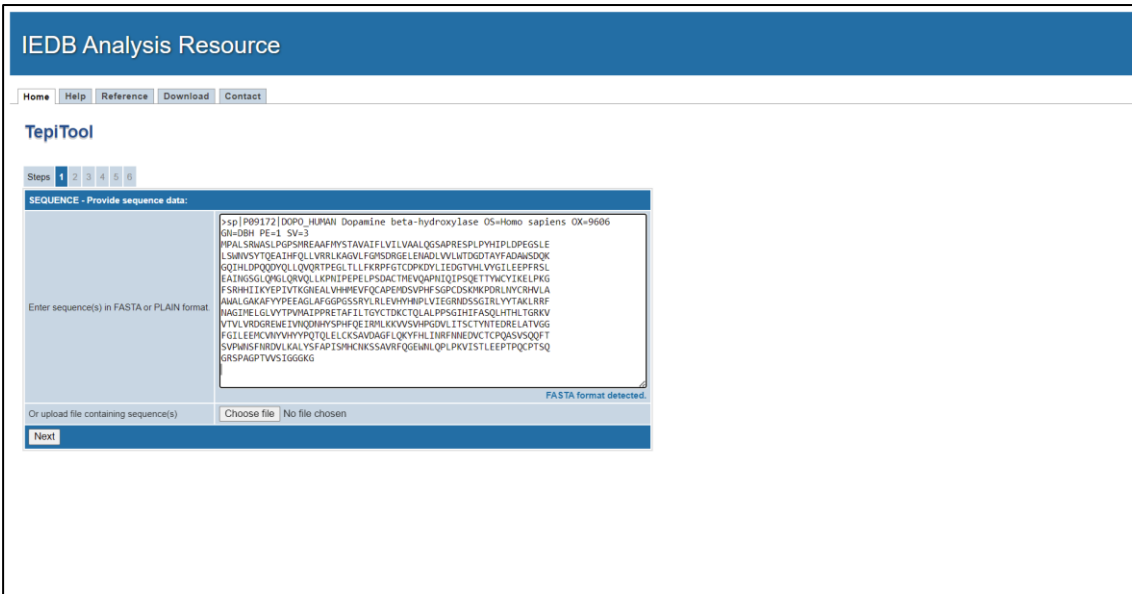
Enter sequence(s) in FASTA or PLAIN format.

No format detected.

Or upload file containing sequence(s) Choose File No file chosen

Next

**Fig 4: Homepage of TepiTool Database**



**Fig 5: Enter the FASTA sequence**

## MHC CLASS I

**IEDB Analysis Resource**

Home Help Reference Download Contact

### TepiTool

Steps 1 **2** 3 4 5 6

**SPECIES & ALLELE CLASS - Select the host species and MHC allele class:**

Host species	Human
Allele class	Class I

Current selections:  
No. of sequences 1

Start Over Back Next

**Fig 6: Select the Host species and MHC allele class I**

**IEDB Analysis Resource**

Home Help Reference Download Contact

### TepiTool

Steps 1 2 **3** 4 5 6

**ALLELES - Specify alleles:**

Human - Class I

- Select from list of frequently occurring alleles (Frequency > 1%)
- Select from list of all available alleles
- Select from list of representative alleles from different HLA supertypes
- Use panel of 27 most frequent A & B alleles
- Upload allele file

Alleles

- A\*01:01
- A\*02:01
- A\*02:06
- A\*03:01
- A\*11:01
- A\*23:01
- A\*24:02
- A\*25:01
- A\*26:01
- A\*29:02
- A\*30:01
- A\*30:02

Current selections:

No. of sequences	1
Host species	Human
Allele class	Class I
Selected alleles	1 A*01:01 2 A*02:01 3 A*02:06

[Reset alleles](#)

Save

Start Over Back Next

**Fig 7: Select the 3 alleles for prediction**



**IEDB Analysis Resource**

Home Help Reference Download Contact

**TepiTool**

Steps 1 2 3 **4** 5 6

**PEPTIDES - Select peptides to be included in prediction:**

Peptides to be included in prediction

- Apply default settings for low number of peptides
- Apply default settings for moderate number of peptides
- Apply default settings for high number of peptides
- Custom selection - Select your own settings

Handling of duplicate peptides:  
- Duplicate peptides will be removed.

Peptide lengths to be considered in prediction:  
- Only peptide length 9 will be included  
9mers = 609

Conservancy analysis  
(Uses only peptides conserved in specified % of sequences)  
N/A (You have only 1 sequence)

Start Over Back Next

**Current selections:**

No. of sequences	1
Host species	Human
Allele class	Class I
Selected alleles	1.A*01:01 2.A*02:01 3.A*02:06

**Fig 8: Choose peptides to be included in prediction**

**IEDB Analysis Resource**

Home Help Reference Download Contact

**TepiTool**

Steps 1 2 3 4 **5** 6

**METHOD - Select prediction & peptide selection methods and cutoff values:**

Prediction method to use: IEDB recommended

Selection of predicted peptides: Select peptides based on predicted percentile rank

Select peptides with predicted consensus percentile rank  $\leq$  1

Start Over Back Next

**Current selections:**

No. of sequences	1
Host species	Human
Allele class	Class I
Selected alleles	1.A*01:01 2.A*02:01 3.A*02:06
Duplicate peptides	Removed
Peptide lengths selected	9mers
No. of peptides included (Not considering conservancy analysis)	609
Conservancy analysis	Peptides conserved in at least % sequences

**Fig 9: Select prediction method, peptide selection strategy & cutoff values**

# IEDB Analysis Resource

[Home](#) [Help](#) [Reference](#) [Download](#) [Contact](#)

## TepiTool

Steps 1 2 3 4 5 **6**

### REVIEW: Review selections, enter job details & submit data:

#### Summary:

No. of sequences	1
Host species	Human
Allele class	Class I
Alleles	1.A*01:01 2.A*02:01 3.A*02:06
Duplicate peptides	Removed
Peptide lengths selected	9mers
Approx no. of peptides included	609
Peptide overlap	N/A (all possible nmers are included in class I)
Conservancy analysis	Peptides conserved in at least % sequences
Prediction method	IEDB recommended
Peptide selection criterion	Based on predicted consensus percentile rank (Cutoff selected = 1)

#### Job details:

Job name (optional)	<input type="text"/>
Email (optional - will notify when job is finished)	<input type="text"/>

[Start Over](#) [Back](#) [Submit](#)

[Save](#)

(Please note that you will not be able to make any more changes once submitted. You will have to start again if you want to do so.)

**Fig 10: Review summary, enter job details & submit data**

IEDB Analysis Resource

Home Help Reference Download Contact

TepiTool

Prediction results - concise (Download table):

Seq #	Peptide start	Peptide end	Peptide	Percentile rank	Allele
1	448	456	RMLKRVVSV	0.02	HLA-A*02:01
1	448	456	RMLKRVVSV	0.04	HLA-A*02:06
1	345	353	DSGIRLRY	0.06	HLA-A*01:01
1	104	112	WTDGDTAYF	0.07	HLA-A*01:01
1	534	542	SVSQPTTSV	0.07	HLA-A*02:06
1	511	519	FLQVFNLI	0.1	HLA-A*02:01
1	57	65	GSLELRNV	0.11	HLA-A*02:06
1	128	136	QDDYQLGV	0.11	HLA-A*02:06
1	381	389	ETAFILTV	0.14	HLA-A*01:01
1	187	195	QLQMLQNV	0.17	HLA-A*02:01
1	172	180	IIEEPPRL	0.18	HLA-A*02:01
1	511	519	FLQVFNLI	0.19	HLA-A*02:06
1	483	491	IIEEPCNV	0.21	HLA-A*01:01
1	172	180	IIEEPPRL	0.21	HLA-A*02:06
1	398	406	ALPFGIHI	0.23	HLA-A*02:06
1	398	406	ALPFGIHI	0.25	HLA-A*02:01
1	241	249	FRHILRY	0.26	HLA-A*01:01
1	225	233	QETTIVCV	0.26	HLA-A*01:01
1	138	146	RFFGDTLL	0.26	HLA-A*02:06
1	319	327	FGFGSSRY	0.29	HLA-A*01:01
1	20	28	PMYSAVAL	0.3	HLA-A*02:01
1	160	168	LIEDQVHL	0.33	HLA-A*02:06
1	534	542	SVSQPTTSV	0.36	HLA-A*02:01

Fig 11: predicted result for MHC Class I

- Number (Seq #):** Indicates the order of the predicted peptides Sequence.
- Peptide Start:** The position in the P53 sequence where the predicted peptide begins.
- Peptide End:** The position in the P53 sequence where the predicted peptide ends.
- Peptide:** The amino acid sequence of the predicted peptide.
- Percentile Rank:** Represents the predicted binding affinity of the peptide to a specific MHC Class I allele. A lower percentile rank indicates a stronger binding affinity and therefore a higher likelihood of the peptide functioning as a T-cell epitope.
- Allele:** Refers to the specific MHC Class I molecules (HLA alleles) that the peptide is predicted to bind.

Download results details:

Complete results Prediction results of all peptides

Citation information:

If you use these predictions in a manuscript, please include the following in the method section:  
 For complete list of references please click here: [References](#)

Input sequences:

Seq #	Seq title	Sequence
1	sp P09172 DOPD_HUMAN Dopamine beta-hydroxylase OS:Homo sapiens OX:9606 ON:FDIH PE:1 SV:3	MPALSRWASLPGPSMREAFMYSTAWAIFVLVLAQLGSSAPRESPLFYHFLDPEGLSELSWVNSYTGAIHFGLLVRRLKAGVLFQMSDRQELLENADLVLTDDG

Other input parameters:

Input summary:	
No. of sequences	1
Host species	Human
Allele class	Class I
Alleles	A*01:01 A*02:01 A*02:06
Duplicate peptides	Removed
Peptide lengths selected	9mers
Peptide overlap	NA
Consensancy analysis	No
Prediction method	IEDB recommended
Peptide selection criterion	Predicted percentile rank
Cutoff for peptide selection criterion	1
Job name	
Email	

© 2005-2024 IEDB  
 Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

Fig 12: Input Sequences and Other parameters

## MHC CLASS II

IEDB Analysis Resource

Home | Help | Reference | Download | Contact

### TepiTool

Steps: 1 2 3 4 5 6

**SPECIES & ALLELE CLASS - Select the host species and MHC allele class:**

Host species: Human

Allele class: Class II

Current selections:  
No. of sequences: 1

Start Over | Back | Next

Fig 13: Select the Host species and MHC allele class II

TepiTool

Steps: 1 2 3 4 5 6

**ALLELES - Specify alleles:**

Human - Class II

Predict for custom allele set  
 Predict for pre-selected panel of alleles  
 Predict using pre-selected allele sets & methods

Options:  
 Select from list of alleles  
 Upload allele file

Select  $\alpha$  and  $\beta$  chains separately when applicable

Alleles: DR

- DRB1\*01:01
- DRB1\*01:02
- DRB1\*01:03
- DRB1\*01:04
- DRB1\*01:05
- DRB1\*01:06
- DRB1\*01:07
- DRB1\*01:08
- DRB1\*01:09
- DRB1\*01:10
- DRB1\*01:11
- DRB1\*01:12

Current selections:  
No. of sequences: 1  
Host species: Human  
Allele class: Class II  
Selected alleles: DRB1\*01:01, DRB1\*01:02, DRB1\*01:03  
Reset alleles

Start Over | Back | Next

© 2005-2024 IEDB  
Supported by a contract from the National Institute of Allergy and Infectious Diseases, a component of the National Institutes of Health in the Department of Health and Human Services.

Fig 14: Select the alleles for prediction

**IEDB Analysis Resource**

Home | Help | Reference | Download | Contact

**TepiTool**

Steps 1 2 3 4 5 6

**PEPTIDES - Select peptides to be included in prediction:**

Apply default settings for low number of peptides  
 Apply default settings for moderate number of peptides  
 Apply default settings for high number of peptides  
 Custom selection - Select your own settings

Peptides to be included in prediction

Handling of duplicate peptides  
- Duplicate peptides will be removed.

Desired no. of overlapping residues for 15mers  
- No. of overlapping residues fixed at 10

Approximate no. of peptides to be considered for prediction = 122

Conservancy analysis (Uses only peptides conserved in specified % of sequences)  
N/A (You have only 1 sequence)

Start Over | Back | Next

**Current selections:**

No. of sequences	1
Host species	Human
Allele class	Class II
Selected alleles	1.DRB1*01:01 2.DRB1*01:02 3.DRB1*01:03

© 2005-2024 IEDB  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 15: Choose peptides to be included in prediction**

**IEDB Analysis Resource**

Home | Help | Reference | Download | Contact

**TepiTool**

Steps 1 2 3 4 5 6

**METHOD - Select prediction & peptide selection methods and cutoff values:**

Prediction method to use: IEDB recommended

Selection of predicted peptides: Select peptides based on predicted percentile rank

Select peptides with predicted percentile rank ≤ 10

Start Over | Back | Next

**Current selections:**

No. of sequences	1
Host species	Human
Allele class	Class II
Alleles selected	1.DRB1*01:01 2.DRB1*01:02 3.DRB1*01:03
Duplicate peptides	Removed
Peptide overlap	10 AA residues
Approx no. of peptides included (Not considering conservancy analysis)	122
Conservancy analysis	Peptides conserved in at least % sequences

© 2005-2024 IEDB  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 16: Select prediction method, peptide selection strategy & cutoff values**

Home | Help | Reference | Download | Contact

## TepiTool

Steps 1 2 3 4 5 6

**REVIEW: Review selections, enter job details & submit data.**

**Summary:**

No. of sequences	1
Host species	Human
Allele class	Class II
Alleles	1 DRB1*01:01 2 DRB1*01:02 3 DRB1*01:03
Duplicate peptides	Removed
Peptide lengths selected	15mers (Only one length for class II)
Approx no. of peptides included	122
Peptide overlap	10 AA residues
Consistency analysis	Peptides conserved in at least % sequences
Prediction method	IEDB recommended
Peptide selection criterion	Based on predicted consensus percentile rank (Cutoff selected = 10)

**Job details:**

Job name (optional):

Email (optional - will notify when job is finished):

[Start Over](#) [Back](#) [Submit](#)

(Please note that you will not be able to make any more changes once submitted. You will have to start again if you want to do so.)

© 2005-2024 IEDB  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services

**Fig 17: Review summary, enter job details & submit data**

IEDB Analysis Resource

Home | Help | Reference | Download | Contact

## TepiTool

Prediction results - concise ([Download table](#)):

Seq #	Peptide start	Peptide end	Peptide sequence	Percentile rank	Allele
1	174	188	EEFFSLEAINGSGL	0.01	HLA-DRB1*01:01
1	128	142	QQDVLLGQRTFED	0.2	HLA-DRB1*01:01
1	307	321	RAFTTFEASLAPFD	0.56	HLA-DRB1*01:01
1	4	18	LSPWALGPPGSRRA	0.56	HLA-DRB1*01:01
1	512	526	LQYFHLINRNNED	4.7	HLA-DRB1*01:01
1	156	170	REYLLDDGTHLVY	4.8	HLA-DRB1*01:01
1	368	382	GLVYTFVDAIFPRT	4.8	HLA-DRB1*01:01
1	75	89	QLVRRLEAGVLFDM	5.2	HLA-DRB1*01:01
1	192	206	LQVLLKSNIPSEF	5.7	HLA-DRB1*01:01
1	15	29	NRDALPHYSTVAIF	5.7	HLA-DRB1*01:01
1	245	259	RIKLYEPIYKQNEA	8.4	HLA-DRB1*01:01
1	472	486	DRRLATVGGFILLE	8.8	HLA-DRB1*01:01
1	258	272	EALVRRHVFQCAPE	9.1	HLA-DRB1*01:01
1	323	337	GSSVYLELVYDHF	9.7	HLA-DRB1*01:01
1	430	444	REIYVQGNHYFHF	9.8	HLA-DRB1*01:01
1	174	188	EEFFSLEAINGSGL	0.81	HLA-DRB1*01:02
1	192	206	QQVLLKSNIPSEF	2.3	HLA-DRB1*01:02
1	128	142	QQDVLLGQRTFED	2.5	HLA-DRB1*01:02
1	75	89	QLVRRLEAGVLFDM	2.9	HLA-DRB1*01:02
1	450	464	LKVVVRFQGVLLIT	4.8	HLA-DRB1*01:02
1	430	444	REIYVQGNHYFHF	4.9	HLA-DRB1*01:02
1	472	486	DRRLATVGGFILLE	4.9	HLA-DRB1*01:02

**Fig 18: Predicted result for MHC Class II**

Download results details:

Non-redundant results Prediction results with redundant peptides within each sequence removed - Includes positives and negatives  
 Complete results Prediction results of all peptides

Citation information:

If you use these predictions in a manuscript, please include the following in the method section:

For complete list of references please click here: [References](#)

Input sequences:

Seq #	Seq title	Sequence
1	sp P09172 DOPD_HUMAN Dopamine beta-hydroxylase OS: Homo sapiens OX:9606 GN:DBH PE:1 SV:3	MPALSRWASLPGPSMREAAFMYSTAVAFVLVLAALQSSAPRESFLPYHPLDPEGSLSELSWVSYTQEAHFQLLVRRKAGVLFGMSDRGELENAQLVLTGDT

Other input parameters:

Input summary:	
No. of sequences	1
Host species	Human
Allele class	Class II
Alleles	DRB1*01:01 DRB1*01:02 DRB1*01:03
Duplicate peptides	Removed
Peptide lengths selected	15mers (Only one length for class II)
Peptide overlap	10 AA residues
Conservancy analysis	No
Prediction method	IEDB recommended
Peptide selection criterion	Predicted percentile rank
Cutoff for peptide selection criterion	10
Job name	
Email	

© 2005-2024 IEDB  
Supported by a contract from the [National Institute of Allergy and Infectious Diseases](#), a component of the National Institutes of Health in the Department of Health and Human Services.

**Fig 19: Input Sequences and Other parameter**

## **RESULTS:**

The prediction results for MHC Class I and II molecules for the Dopamine (Accession No: P09172) were obtained using IEDB's TepiTool. The predicted result in concise table shows the percentile rank. The low percentile ranks (less than 1) indicate these peptides are strong candidates for further experimental validation as potential MHC Class I and II epitopes for Dopamine. Best percentile ranks 0.02 for the allele peptides are predicted to bind well to the corresponding MHC Class I molecules and for MHC Class II the percentile rank is 1.01.

## **CONCLUSION:**

TepiTool was used to perform T cell epitope predictions on the IEDB database. It identified potential epitopes for both Class I and Class II MHC molecules. TepiTool is a tool designed to predict peptide sequences that can bind to MHC molecules, aiding in the identification of T cell epitopes crucial for immune responses. It provides a ranking based on binding affinity, helping prioritize peptides for further research.

## **REFERENCES:**

1. Paul, S., Sidney, J., Sette, A., & Peters, B. (2016). TEPiTool: A pipeline for Computational Prediction of T cell epitope Candidates. *Current Protocols in Immunology*, 114(1). <https://doi.org/10.1002/cpim.12>
2. Kim, Y., Ponomarenko, J., Zhu, Z., Tamang, D., Wang, P., Greenbaum, J., Lundegaard, C., Sette, A., Lund, O., Bourne, P. E., Nielsen, M., & Peters, B. (2012). Immune epitope database analysis resource. *Nucleic acids research*, 40(Web Server issue), W525–W530. <https://doi.org/10.1093/nar/gks438>
3. Wang, H., Guo, M., Wei, H., & Chen, Y. (2023). Targeting p53 pathways: mechanisms, structures, and advances in therapy. *Signal Transduction and Targeted Therapy*, 8(1). <https://doi.org/10.1038/s41392-023-01347-1>