NAME: SHALMON N. ANANDAS

CLASS: M.Sc. PART – I

COURSE: BIOINFORMATICS

ACADEMIC YEAR: 2021-2022

ROLL NO: 91

PAPER CODE: GNKPSBI202 (PAPER 2)

COURSE TITLE: Structural Biology & Bioinformatics

GURU NANAK KHALSA COLLEGE

MATUNGA, MUMBAI – 400019

DEPARTMENT OF BIOINFORMATICS

CERTIFICATE

This is to certify that Mr. **Shalmon N Anandas** of M.Sc. Part I Bioinformatics has satisfactorily completed the practical semester I course prescribed by the university of Mumbai during the academic year 2021-2022

TEACHER IN CHARGE **HEAD OF DEPARTMENT**

INDEX

	8c	FGENESB	172	19-03-22	
	8d	FGENES	178	19-03-22	
	8e	ORF finder – NCBI	183	19-03-22	
9		Introduction to Genomics & its various browsers (UCSC, ENSEMBL, GDV)	187	30-03-22	
	9a	UCSC Genome Browser	190	30-03-22	
	9b	ENSEMBL Genome Browser	208	30-03-22	
	9c	Genome Data Viewer	215	30-03-22	
10		A field guid to whole-genome sequencing, assembly and annotation	225	31-03-22	

DATE: 20-02-22

WEBLEM 1

SECONDARY STRUCTURE PREDICTION

Introduction:

- Secondary structure prediction is relatively accurate, and is in fact much easier to solve than three-dimensional structure prediction. The accuracy of assigning strand, helix or loops to a certain residue can go up to 80% with the most reliable methods. Typically, these methods use periodicity in the sequence combined with phi and psi angle preferences of certain amino acids types to come to accurate predictions. The real challenge lies in assembling the secondary structure element in a correct topology. Nevertheless, secondary structure prediction may be used to assess the quality of model built with a structure prediction method. Many methods also incorporated secondary structure information during homology detection.

Secondary structure prediction has three generations:

1. First Generation: Chou and Fasman

- CFSSP (Chou & Fasman Secondary Structure Prediction Server) is an online protein secondary structure prediction server. This server predicts regions of secondary structure from the protein sequence such as alpha helix, beta sheet, and turns from the amino acid sequence. The output of predicted secondary structure is also displayed in linear sequential graphical view based on the probability of occurrence of alpha helix, beta sheet, and turns. The method implemented in CFSSP is Chou-Fasman algorithm, which is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. CFSSP is freely accessible via ExPASy server.

2. Second Generation: GOR IV

- The GOR method is based on information theory and was developed by J.Garnier, D.Osguthorpe and B.Robson (J.Mol.Biol.120,97, 1978). The present version, GOR IV, uses all possible pair frequencies within a window of 17 amino acid residues and is reported by J. Garnier. J.F. Gibrat and B.Robson in Methods in Enzymology, vol 266, p 540-553 (1996). After cross validation on a data base of 267 proteins, the version IV of GOR has a mean accuracy of 64.4% for a three-state prediction (Q3). The program gives two outputs, one eye-friendly giving the sequence and the predicted secondary structure in rows, H=helix, E=extended or beta strand and C=coil; the second gives the probability values for each secondary structure at each amino acid position. The predicted secondary structure is the one of highest probability compatible with a predicted helix segment of at least four residues and a predicted extended segment of at least two residues.

3. Third Generation: PSIPRED

- The PSIPRED Workbench provides a range of protein structure prediction methods. The site can be used interactively via a web browser or programmatically via our REST API. For high- throughput analyses, downloads of all the algorithms are available. Amino acid sequences enable: secondary structure prediction, including regions of disorder and transmembrane helix packing; contact analysis; fold recognition; structure modelling; and prediction of domains and function. In addition, PDB Structure files allow prediction of protein-metal ion contacts, protein-protein hotspot residues, and membrane protein orientation.

References:

- Protein Secondary Structure - an overview | ScienceDirect Topics. (2016). Science Direct. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/protein-secondary-structure
- Edwards, Y. J., and Cottage, A. 2003. Bioinformatics methods to predict protein structure and function. A

practical approach. Mol. Biotechnol. 23:139-66.

- Ashok Kumar, T. (2013). CFSSP: Chou and Fasman Secondary Structure Prediction server. WIDE SPECTRUM: Research Journal. 1(9):15-19.
- Heringa J. 2002. Computational methods for protein secondary structure prediction using multiple sequence alignments. Curr. Protein Pept. Sci. 1:273-301.
- Lehnert, U., Xia, Y., Royce, T. E., Goh, C. S., Liu, Y., Senes, A., Yu. H., Zhang, Z. L., Engelman, D. M, and Gerstein M. 2004. Computational analysis of membrane proteins: Genomic occurrence,structure prediction and helix interactions. Q. Rev. Biophys. 37:121-46.
- Moller, S., Croning, M. D. R., and Apweiler, R. 2001. Evaluation of methods for the prediction of ¨membrane spanning regions. Bioinformatics 17:646-53.

DATE: 20-02-22

WEBLEM 1A
SECONDARY STRUCTURE PREDICTION

Aim:

➔ Secondary Structure prediction for CAD13_HUMAN (Cadherin) using various tools (CF, GOR IV, PSIPRED)

Introduction:

➔ Secondary Structure prediction has three generations:

1. First Generation: Chou and Fasman

➔ CFSSP (Chou & Fasman Secondary Structure Prediction Server) is an online protein secondary structure prediction server. This server predicts regions of secondary structure from the protein sequence such as alpha helix, beta sheet, and turns from the amino acid sequence. The output of predicted secondary structure is also displayed in linear sequential graphical view based on the probability of occurrence of alpha helix, beta sheet, and turns.

2. Second Generation: GOR IV

➔ The GOR method is based on information theory and was developed by J.Garnier, D.Osguthorpe and B.Robson. After cross validation on a data base of 267 proteins, the version IV of GOR has a mean accuracy of 64.4% for a three-state prediction. The program gives two outputs, one eye-friendly giving the sequence and the predicted secondary structure in rows, H=helix, E=extended or beta strand and C=coil; the second gives the probability values for each secondary structure at each amino acid position.

3. Third Generation: PSIPRED

➔ The PSIPRED Workbench provides a range of protein structure prediction methods. The site can be used interactively via a web browser or programmatically via our REST API. For high- throughput analyses, downloads of all the algorithms are available. Amino acid sequences enable: secondary structure prediction, including regions of disorder and transmembrane helix packing; contact analysis; fold recognition; structure modelling; and prediction of domains and function

Cadherin:

➔ Cadherins are a large family of cell-cell adhesion molecules acting in a homotypic, homophilic manner that play an important role in the maintenance of tissue integrity. In the human kidney several members of the cadherin family are expressed in a controlled spatiotemporal pattern. Cadherin-16 also called kidney-specific cadherin, is exclusively expressed in epithelial cells of the adult kidney.

Methodology:

➔ Open Homepage of (CF, GOR IV, PSIPRED)
➔ Enter the query Cadherin Fasta sequence in the search bar.
➔ Open the Result page.
➔ Interpret the Results.

Observations:

➔ **First Generation: Chou and Fasman**

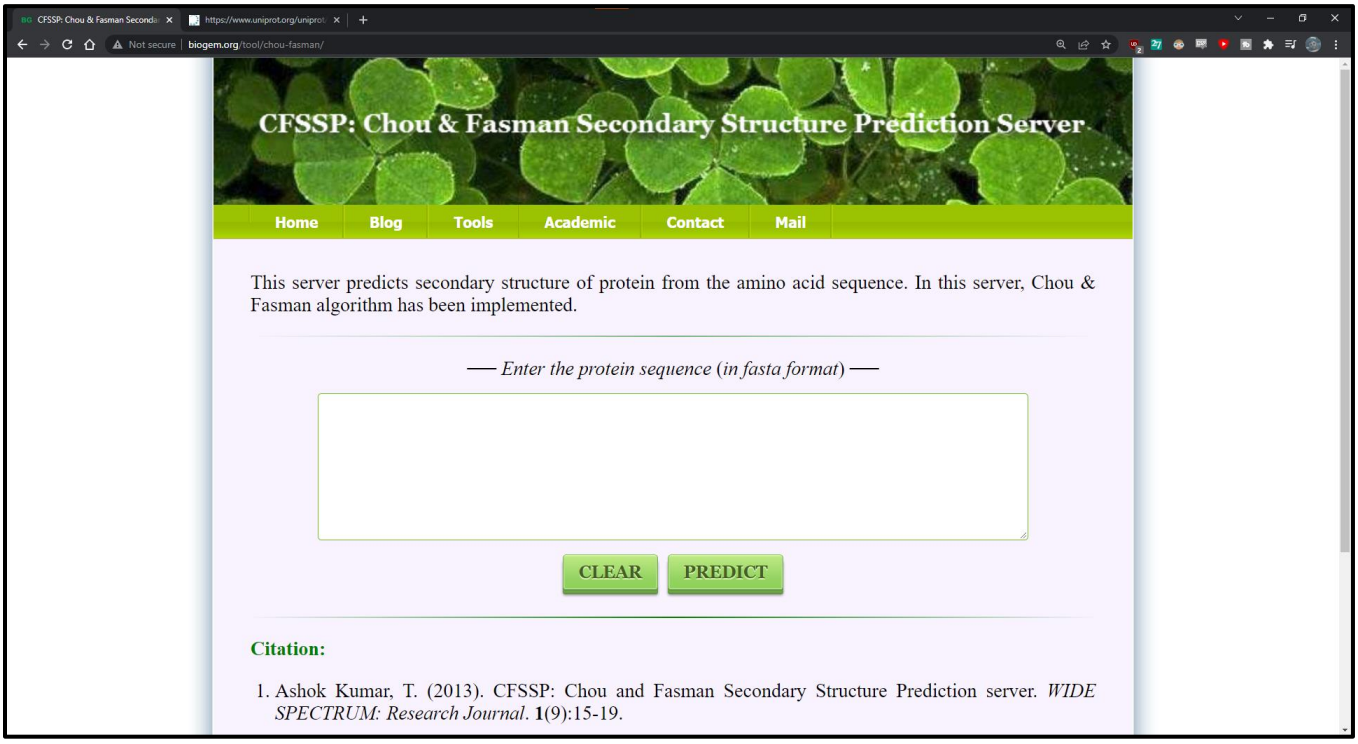


Fig1: Homepage of Chou and fasman

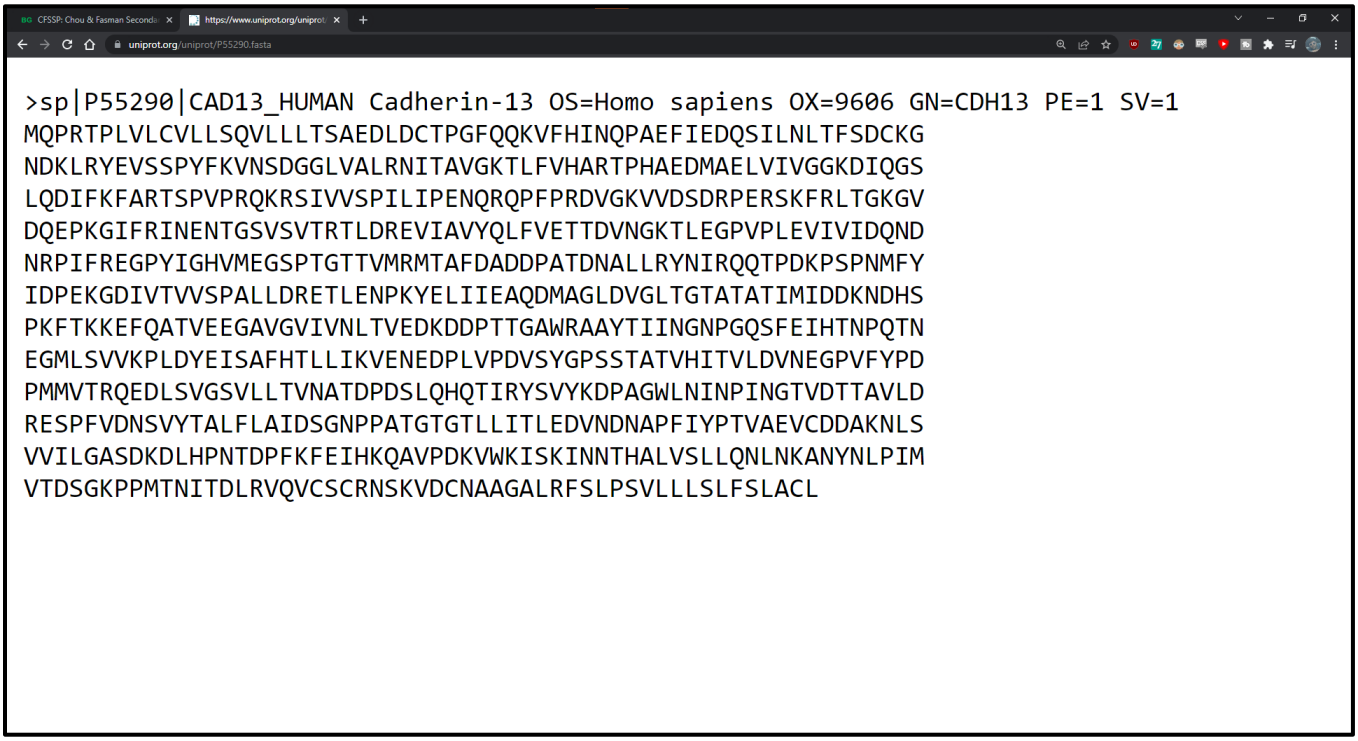


Fig2. FASTA sequence of CAD13_HUMAN Cadherin

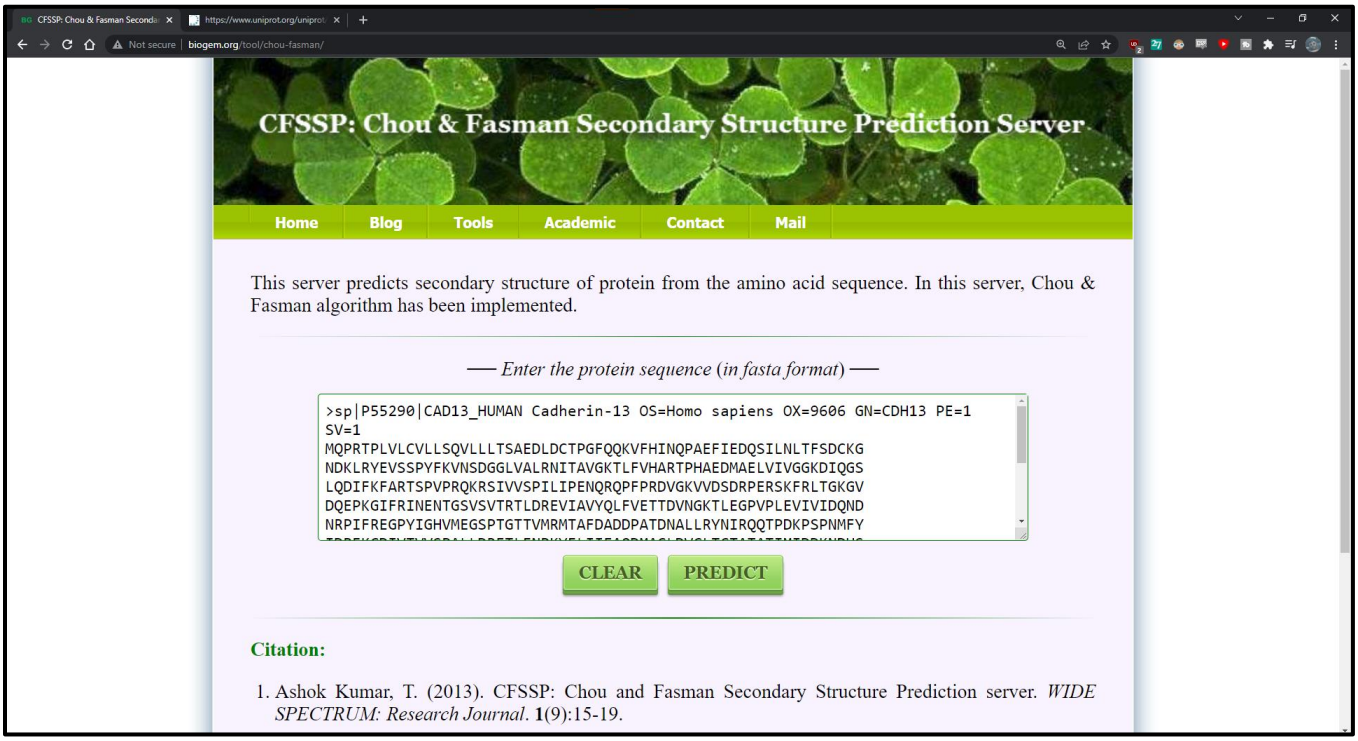


Fig3. Search bar with FASTA sequence of CAD13_HUMAN Cadherin

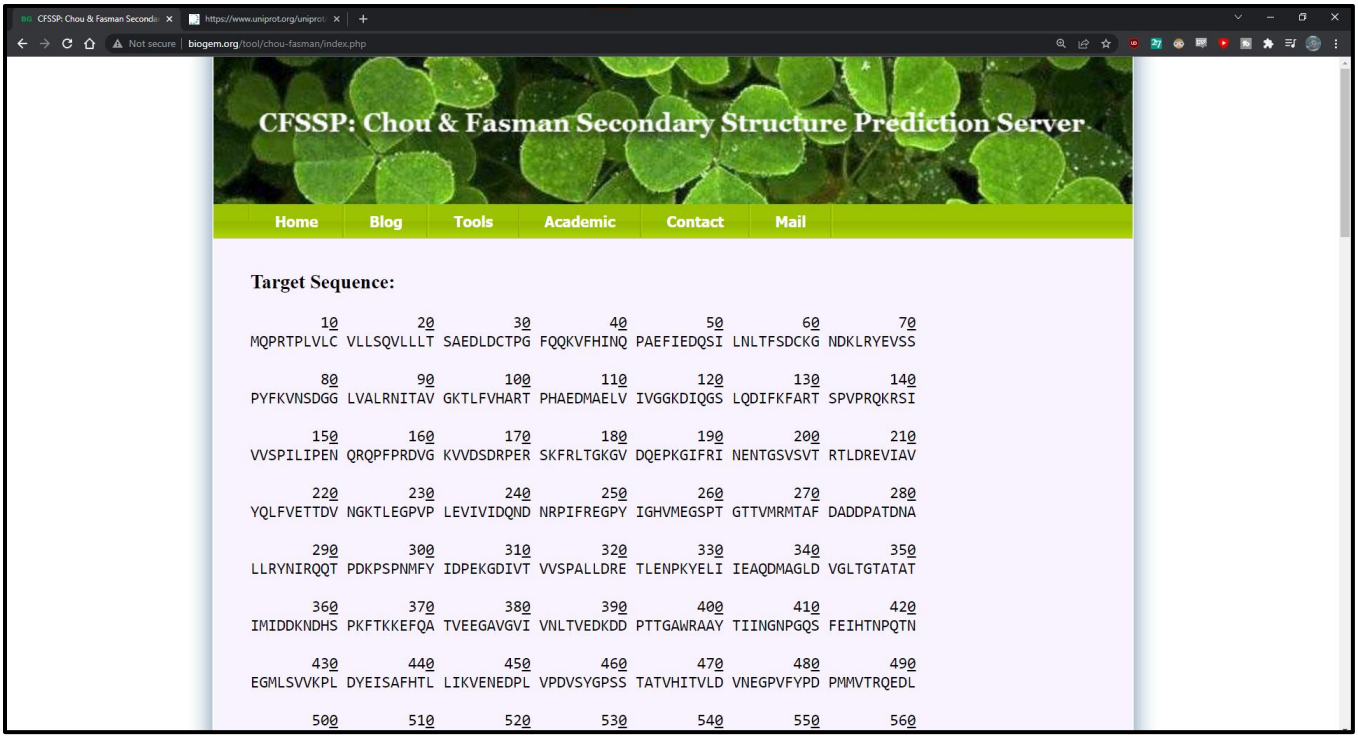


Fig4. Result page of Cadherin showing target sequence

Fig5. Result page of Cadherin showing Secondary structure

➔ **Secondary Generation: GOR IV**

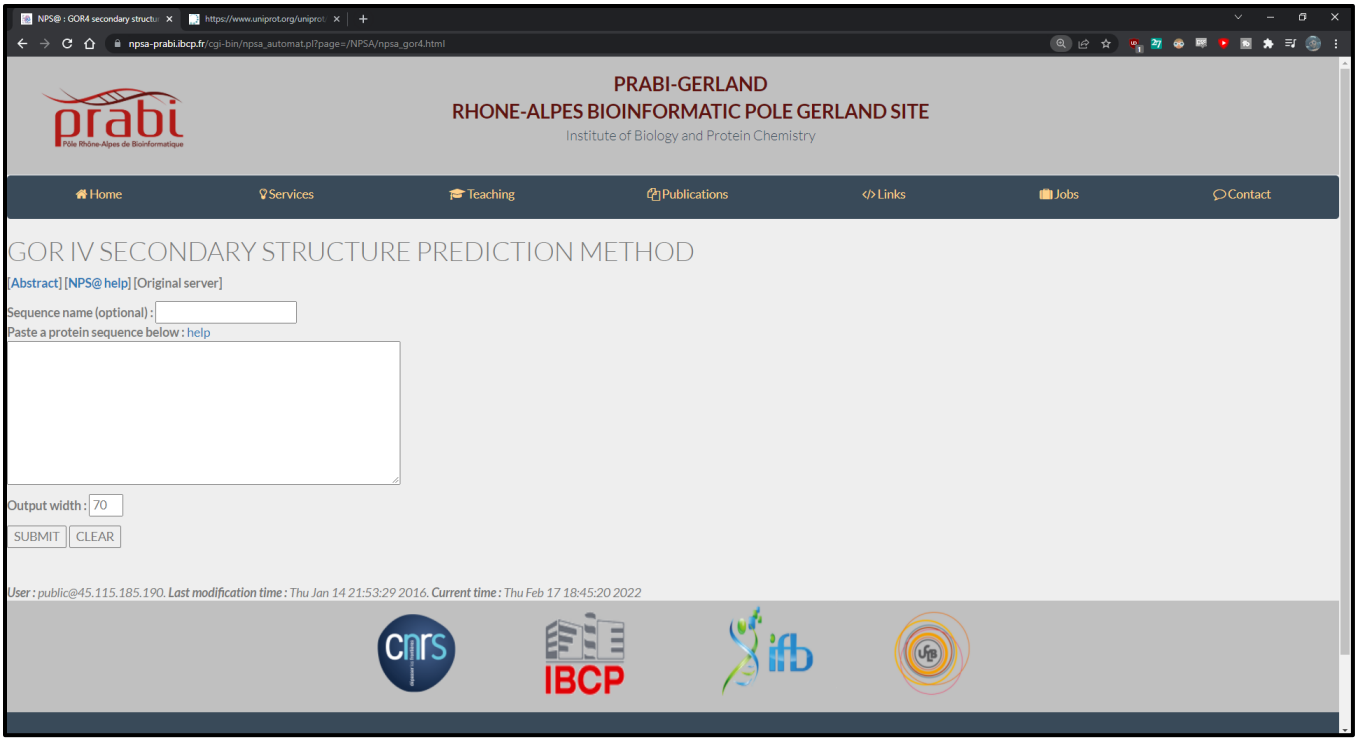


Fig1. Homepage of GOR IV

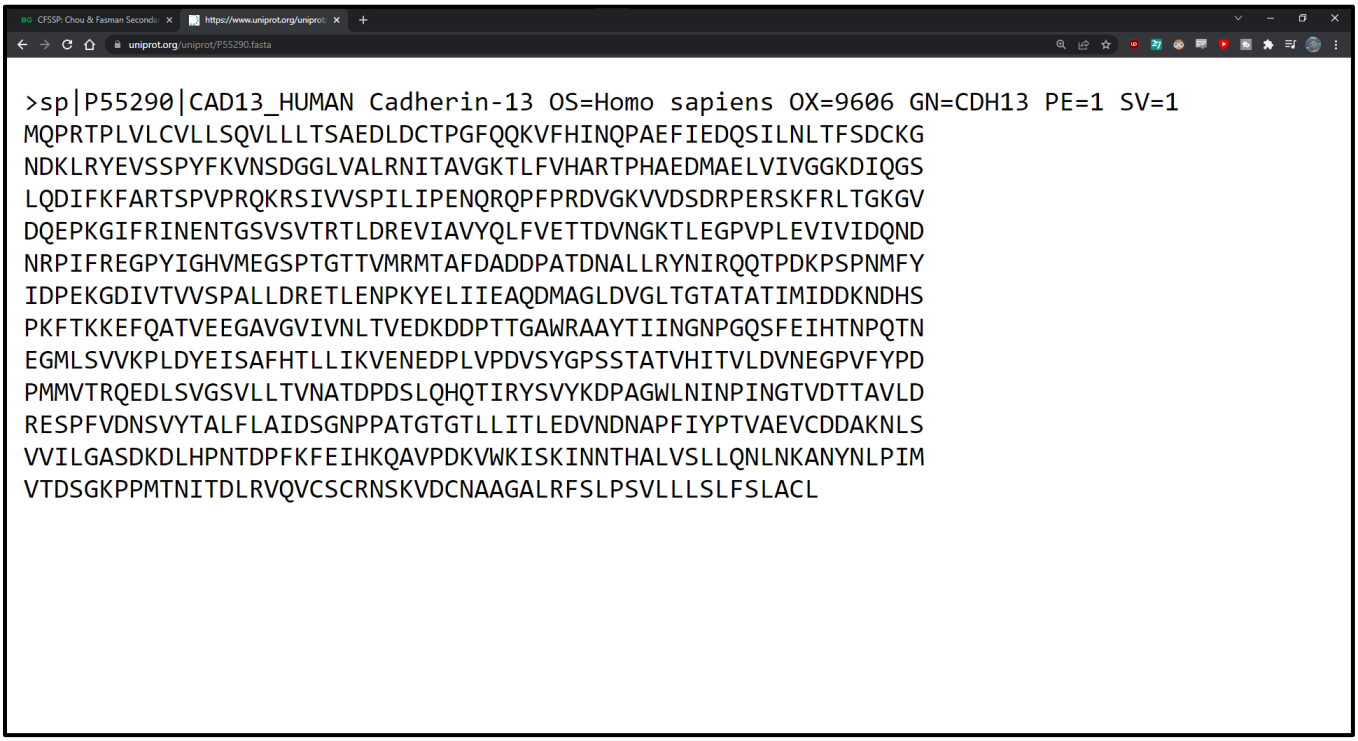


Fig2. FASTA sequence of CAD13_HUMAN Cadherin

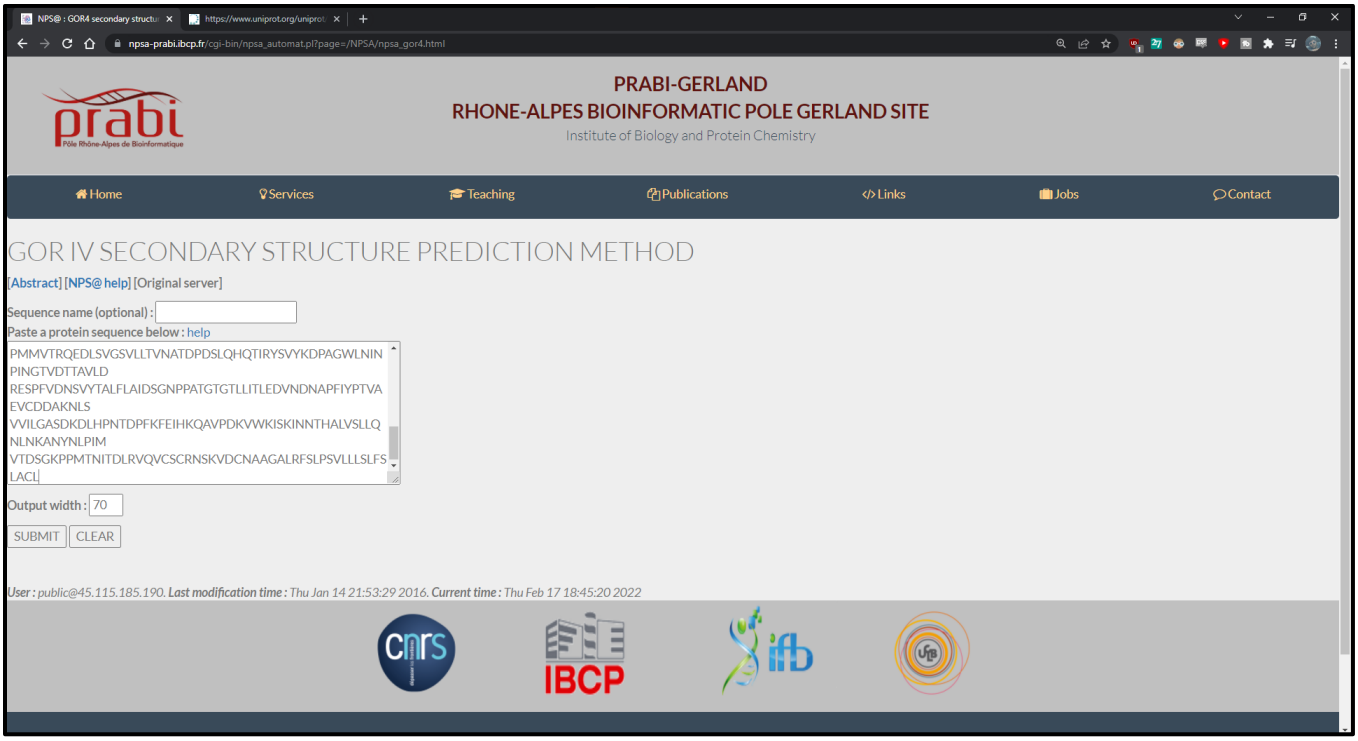


Fig3. Search bar with FASTA sequence of cadherin

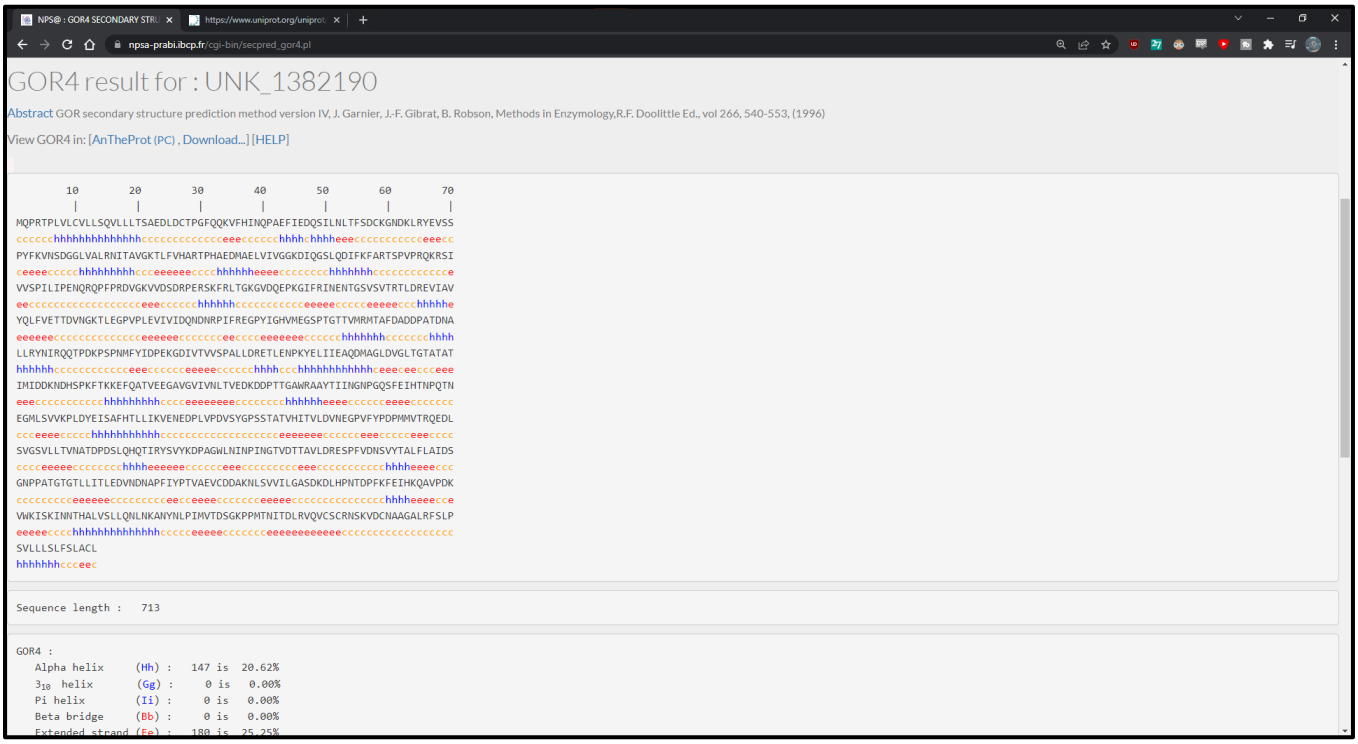


Fig4. Result page of GOR IV for Cadherin

Fig5. Frequency graph for Cadherin

➔ **Third Generation: PSIPRED**

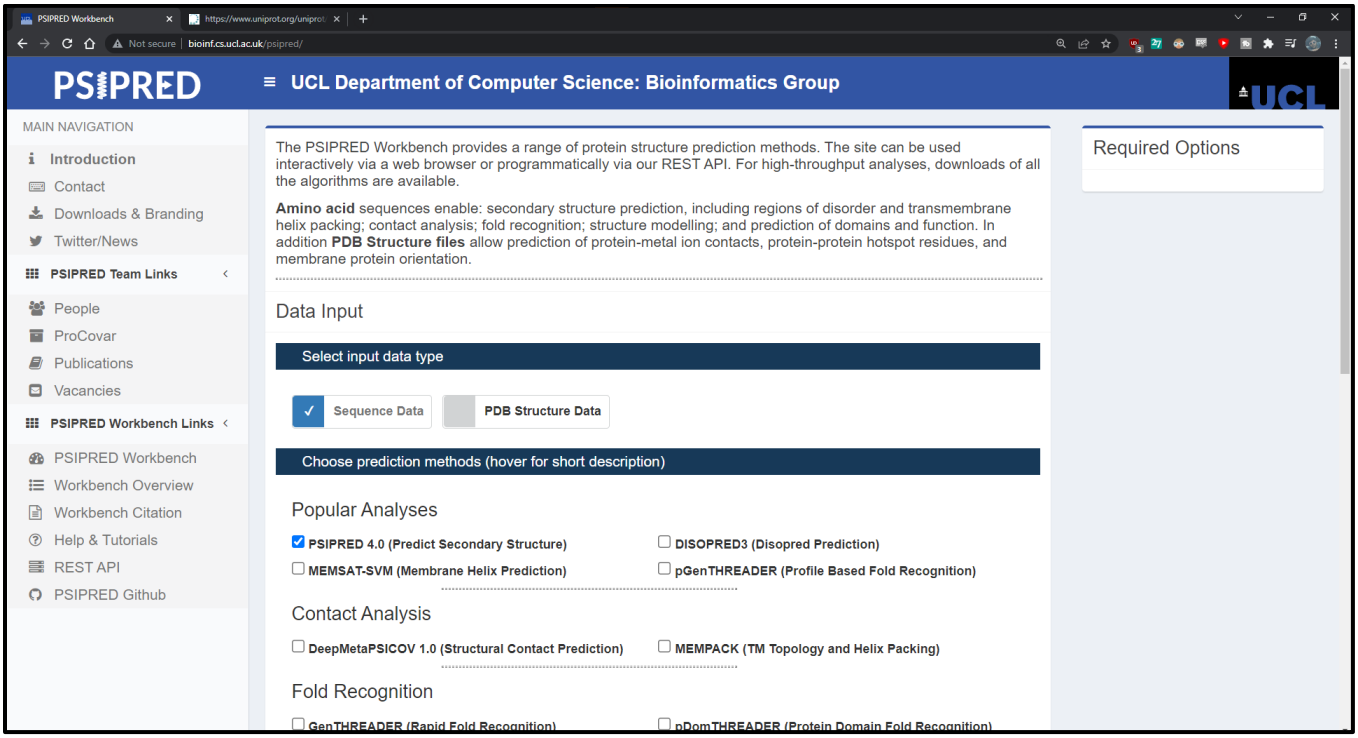


Fig1. Homepage for PSIPRED

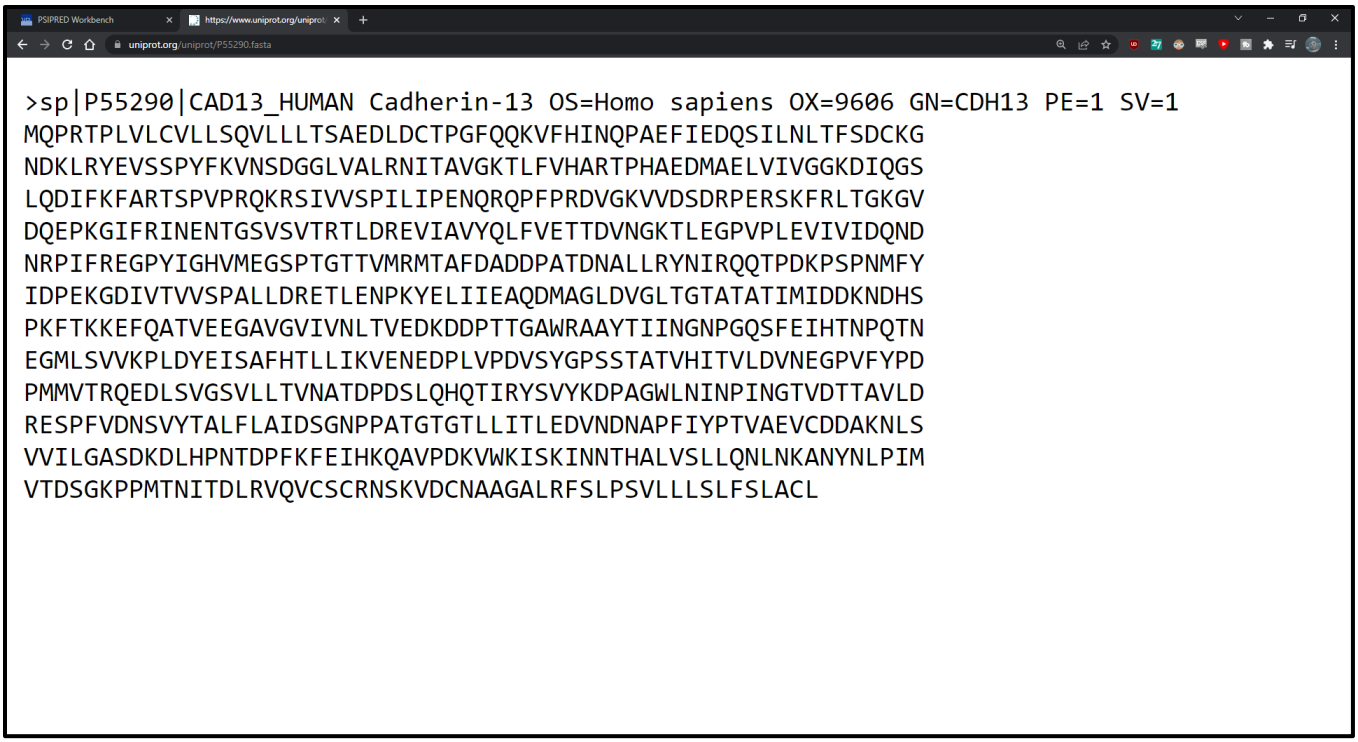


Fig2. FASTA sequence of CAD13_HUMAN Cadherin

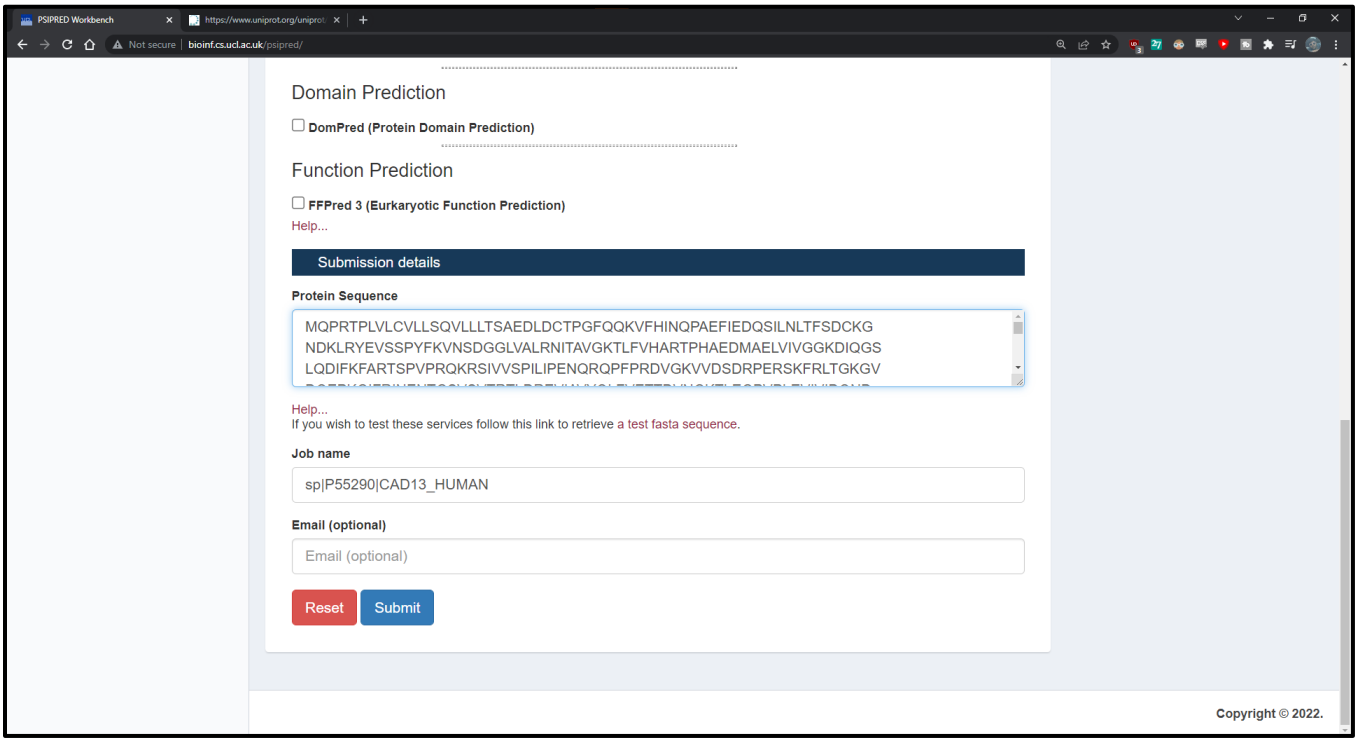


Fig3. Search bar of PSIPRED with FASTA sequence of Cadherin

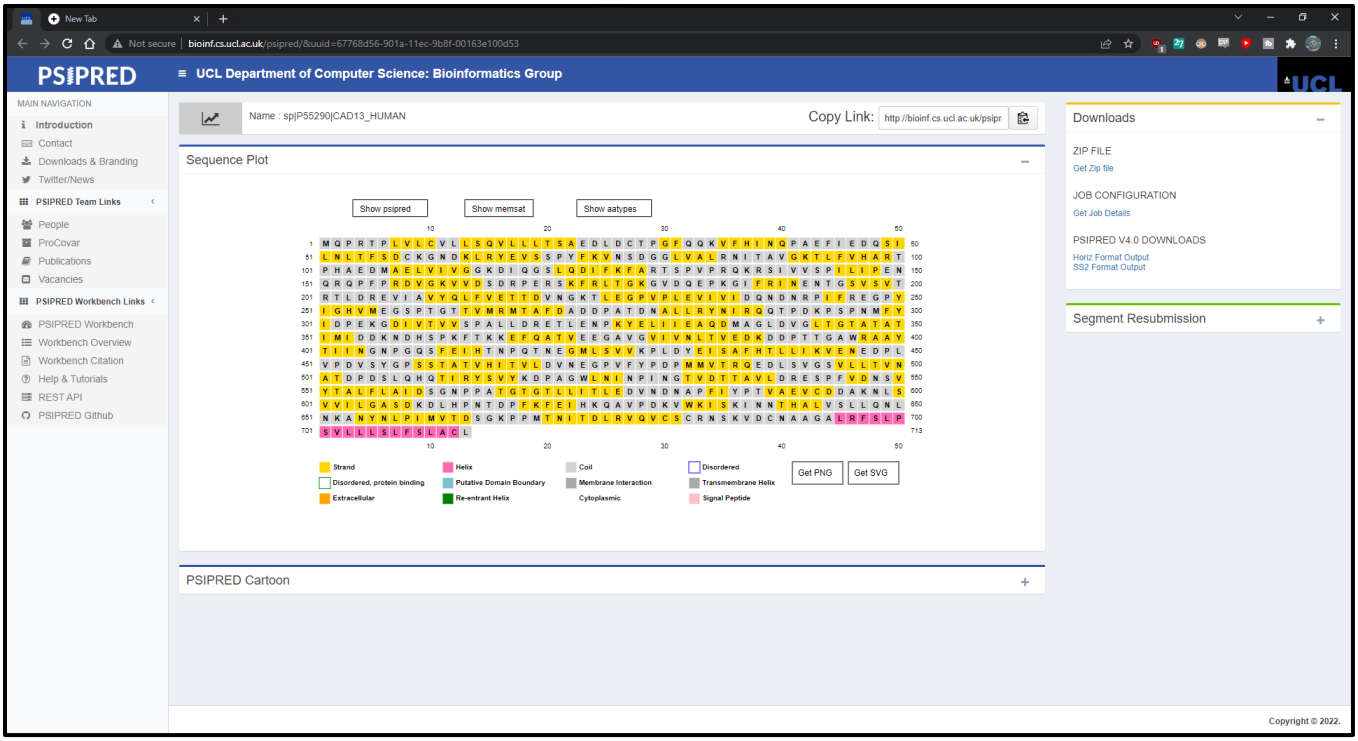


Fig4. Sequence plot of CAD13_HUMAN Cadherin in PSIPRED

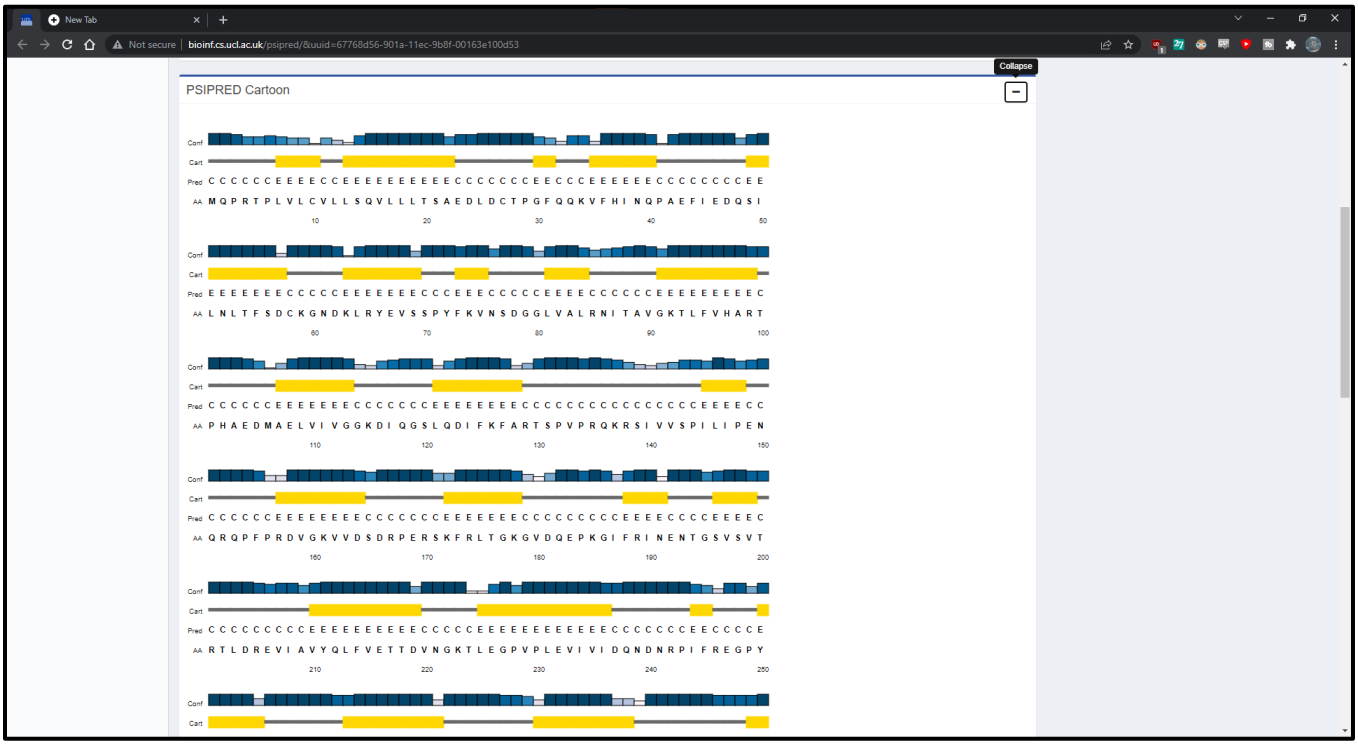


Fig5. Legend of Cadherin in PSIPRED

Results:

➔ **First Generation: Chou and Fasman**

- In secondary structure prediction using Chou and Fasman tools. The FASTA Sequence length of query Cadherin is 713 and their total residues are H: 455; E: 470; T:95; and the percentage is H: 63.8; E: 65.9; T:13.3;

➔ **Second Generation: GOR IV**

- In secondary structure prediction using GOR IV tools. The FASTA Sequence length of query Cadherin is 713 and tota residues are Alpha helix: 147 (20.62%); Extended Strand: 180 (25.25%); Random coil: 386 (54.14%).

➔ **Third Generation: PSIPRED**

- In secondary structure prediction using PSIPRED tools. The FASTA sequence length of query Cadherin is 713. Here the maximum helix residues are predicted and we can find legend of confidence of prediction, 3-state assignment cartoon and target sequence.

Conclusion:

➔ The secondary structure of the protein is Alpha helix and Beta sheet. Protein structure plays a key role in its function. Secondary structure prediction is important from the structural and functional point of view. Secondary structure is local interactions between stretches of a polypeptide chain and include alpha helix and beta pleated sheet structures.

References:

- Ranscht, B. (2010). Cadherin Regulation of Adhesive Interactions. Handbook of Cell Signaling, 1975–1988. https://doi.org/10.1016/b978-0-12-374145-5.00242-4
- CHOU-FASMAN. (2022, February 12). CHOU-FASMAN. Retrieved February 14, 2022, from http://www.biogem.org/tool/chou-fasman/index.php
- Ashok Kumar, T. (2013). CFSSP: Chou and Fasman Secondary Structure Prediction server. WIDE SPECTRUM: Research Journal. 1(9):15-19.
- GOR IV. (2022, February 12). GOR IV. Retrieved February 14, 2022, from https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html
- PSIPRED. (2016, September 12). Http://Bioinf.Cs.Ucl.Ac.Uk/Psipred. Retrieved February 14, 2022, from http://bioinf.cs.ucl.ac.uk/psipred

Date: 27-02-22

WEBLEM 2
INTRODUCTION TO PROTEIN CLASSIFICATION

Introduction:

One of the applications of protein structure comparison is structural classification. The ability to compare protein structure allows classification of the structure data and identification of relationships among structures. The reason to develop a protein structure classification system is to establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structure. Once a hierarchical classification system is established, a newly obtained protein structure can find its place in a proper category. As a result, its functions can be better understood based on association with other proteins.

CATH:

CATH (www.biochem.ucl.ac.uk/bsm/cath new/index.html) classifies proteins based on the automatic structural alignment program SSAP as well as manual comparison. Structural domain separation is carried out also as a combined effort of a human expert and computer programs. Individual domain structures are classified at five major levels: class, architecture, fold/topology, homologous superfamily, and homologous family. In the CATH release version 2.5.1 (January 2004), there are 4 classes, 37 architectures, 813 topologies, 1,467 homologous superfamilies, and 4,036 homologous families. The definition for class in CATH is similar to that in SCOP, and is based on secondary structure content. Architecture is a unique level in CATH, intermediate between fold and class. This level describes the overall packing and arrangement of secondary structures independent of connectivity between the elements. The topology level is equivalent to the fold level in SCOP, which describes overall orientation of secondary structures and takes into account the sequence connectivity between the secondary structure elements. The homologous superfamily and homologous family levels are equivalent to the superfamily and family levels in SCOP with similar evolutionary definitions, respectively.

SCOPe:

Structural Classification of Proteins-extended (SCOPe, http://scop.berkeley.edu) is a database of protein structural relationships that extends the SCOP database. SCOP is a manually curated ordering of domains from the majority of proteins of known structure in a hierarchy according to structural and evolutionary relationships. Development of the SCOP 1.x series concluded with SCOP 1.75. The ASTRAL compendium provides several databases and tools to aid in the analysis of the protein structures classified in SCOP, particularly through the use of their sequences. SCOPe extends version 1.75 of the SCOP database, using automated curation methods to classify many structures released since SCOP 1.75. We have rigorously benchmarked our automated methods to ensure that they are as accurate as manual curation, though there are many proteins to which our methods cannot be applied. SCOPe is also partially manually curated to correct some errors in SCOP. SCOPe aims to be backward compatible with SCOP, providing the same parseable files and a history of changes between all stable SCOP and SCOPe releases. SCOPe also incorporates and updates the ASTRAL database. The latest release of SCOPe, 2.03, contains 59 514 Protein Data Bank (PDB) entries, increasing the number of structures classified in SCOP by 55% and including more than 65% of the protein structures in the PDB.

References:

1. Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

2. Murzin, A. G. (1995). *Journal of Molecular Biology*, *247*(4), 536–540. https://doi.org/10.1006/jmbi.1995.0159

DATE: 27-02-22

WEBLEM 2A
(URL:https://www.cathdb.info/)

Aim:

To study the structural classification of proteins (Leucine) using **CATH** and **SCOPe** database.

Introduction:

- **CATH:**
  - CATH (www.biochem.ucl.ac.uk/bsm/cath new/index.html) classifies proteins based on the automatic structural alignment program SSAP as well as manual comparison. Structural domain separation is carried out also as a combined effort of a human expert and computer programs. Individual domain structures are classified at five major levels: class, architecture, fold/topology, homologous superfamily, and homologous family. In the CATH release version 2.5.1 (January 2004), there are 4 classes, 37 architectures, 813 topologies, 1,467 homologous superfamilies, and 4,036 homologous families. The definition for class in CATH is similar to that in SCOP, and is based on secondary structure content. Architecture is a unique level in CATH, intermediate between fold and class.
- **SCOPe:**
  - Structural Classification of Proteins-extended (SCOPe, http://scop.berkeley.edu) is a database of protein structural relationships that extends the SCOP database. SCOP is a manually curated ordering of domains from the majority of proteins of known structure in a hierarchy according to structural and evolutionary relationships. Development of the SCOP 1.x series concluded with SCOP 1.75. The ASTRAL compendium provides several databases and tools to aid in the analysis of the protein structures classified in SCOP, particularly through the use of their sequences. SCOPe extends version 1.75 of the SCOP database, using automated curation methods to classify many structures released since SCOP 1.75.
- **Leucine:**
  - Leucine is one of nine essential amino acids in humans (provided by food), Leucine is important for protein synthesis and many metabolic functions. Leucine contributes to regulation of blood-sugar levels; growth and repair of muscle and bone tissue; growth hormone production; and wound healing. Leucine also prevents breakdown of muscle proteins after trauma or severe stress and may be beneficial for individuals with phenylketonuria. Leucine is available in many foods and deficiency is rare.

- **Methodology:**
  - Open Homepage of CATH and SCOPe
  - Enter the query Leucine in the search bar
  - Open the result page of each category
  - Interpret the results.

Observation:

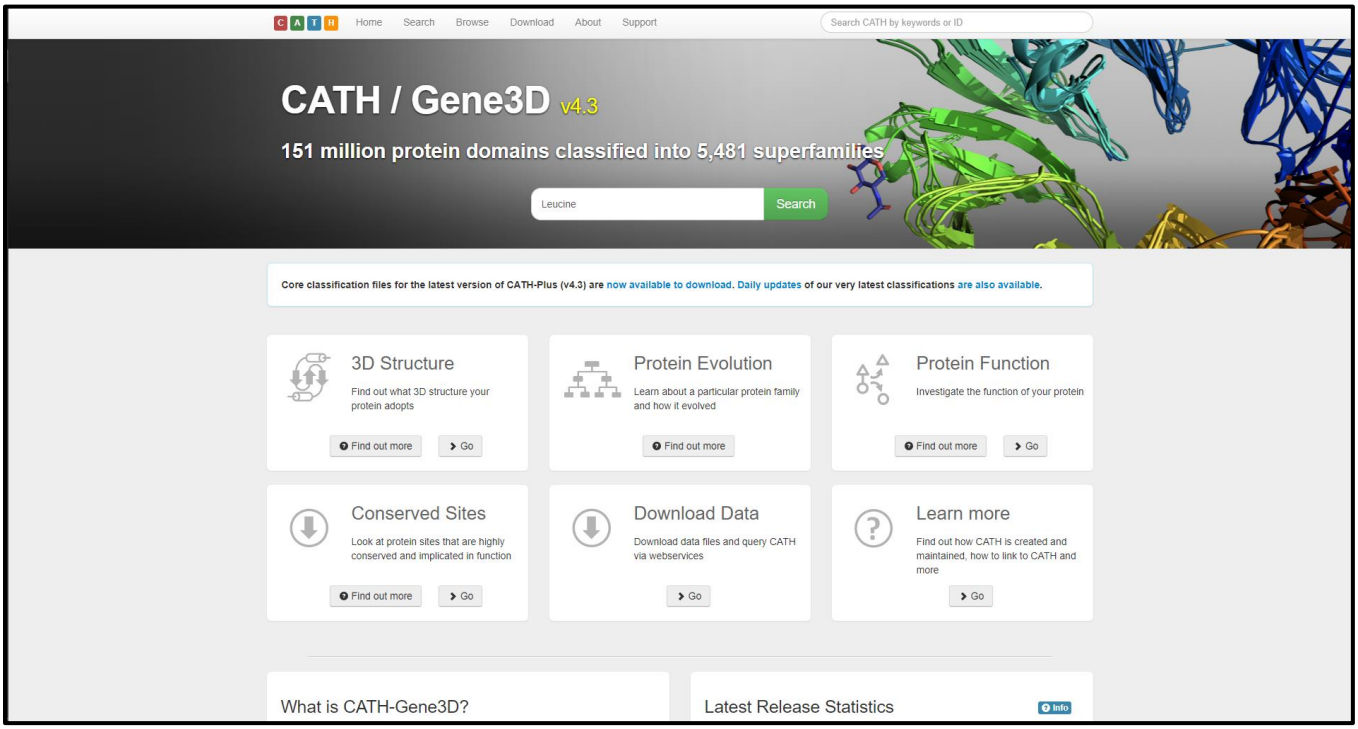


Fig1. Homepage of CATH database with query Leucine

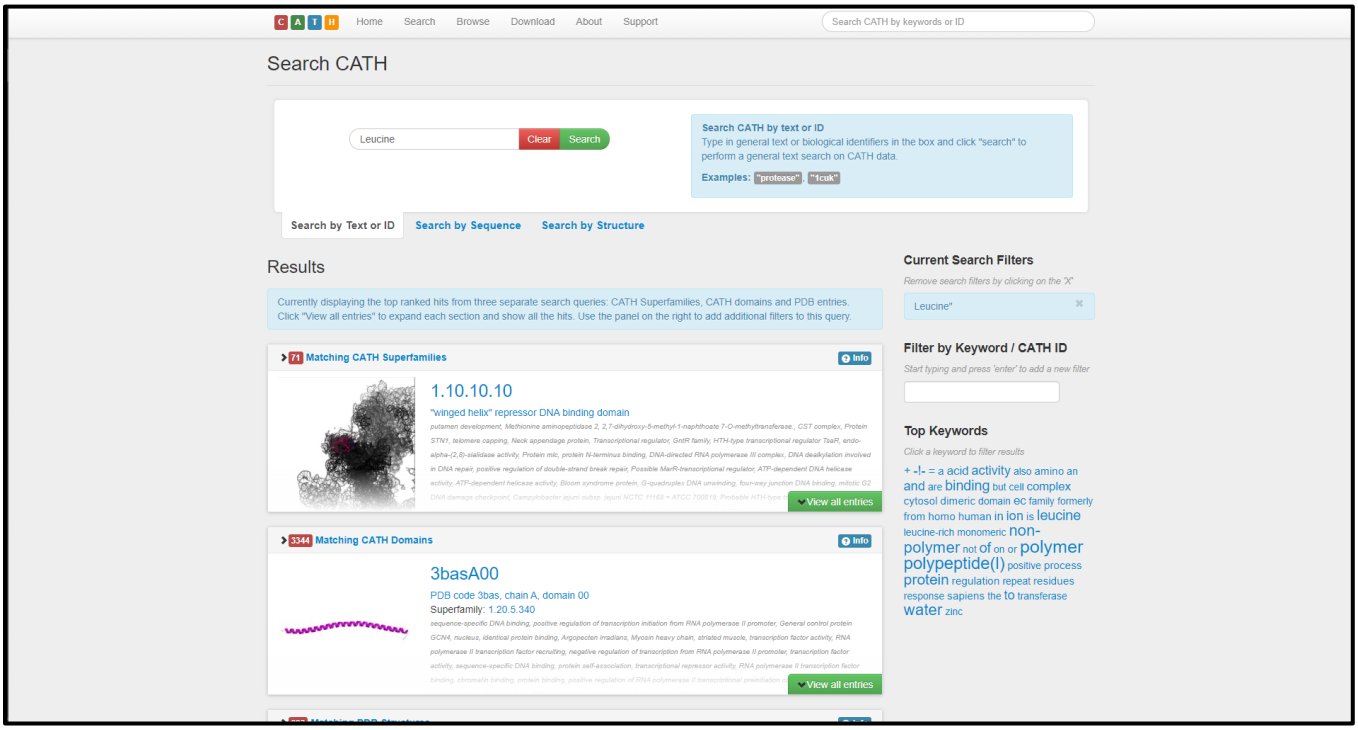


Fig2. Result page for leucine in CATH database

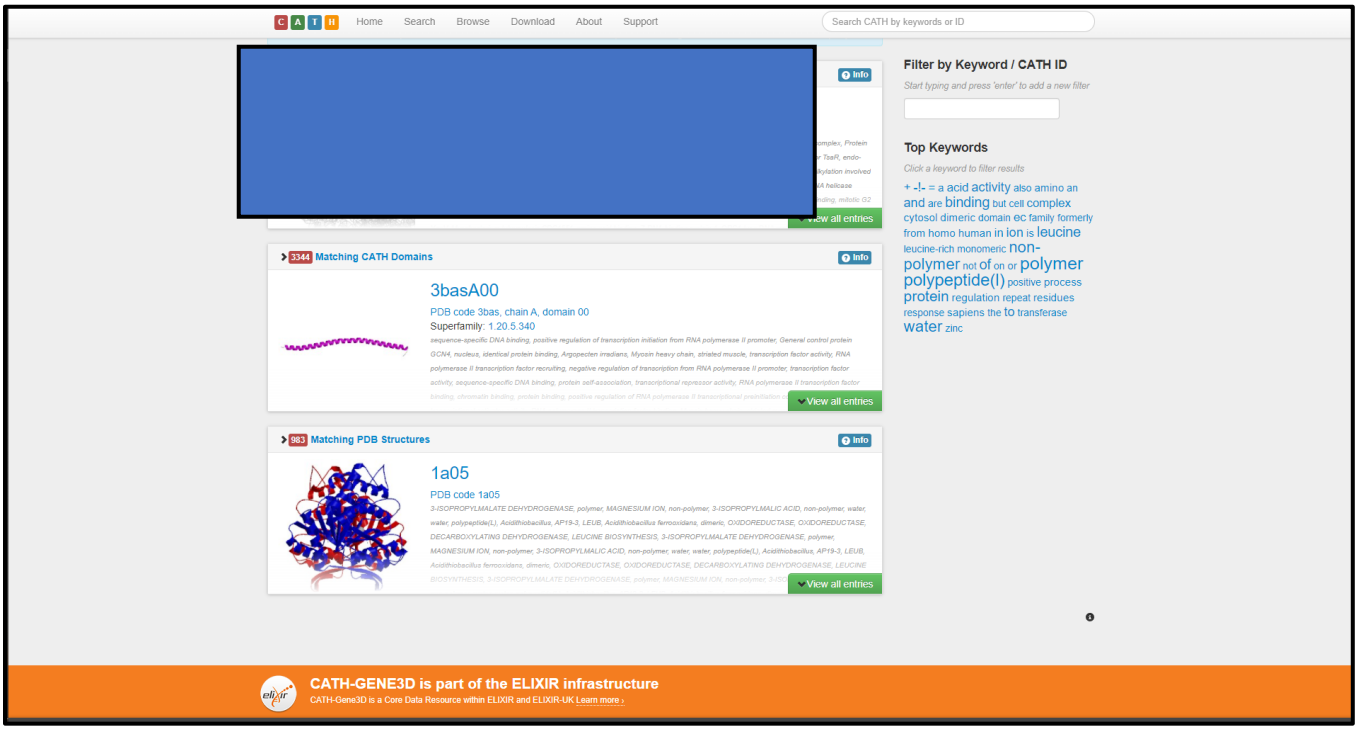


Fig3. Available CATH Superfamilies for leucine

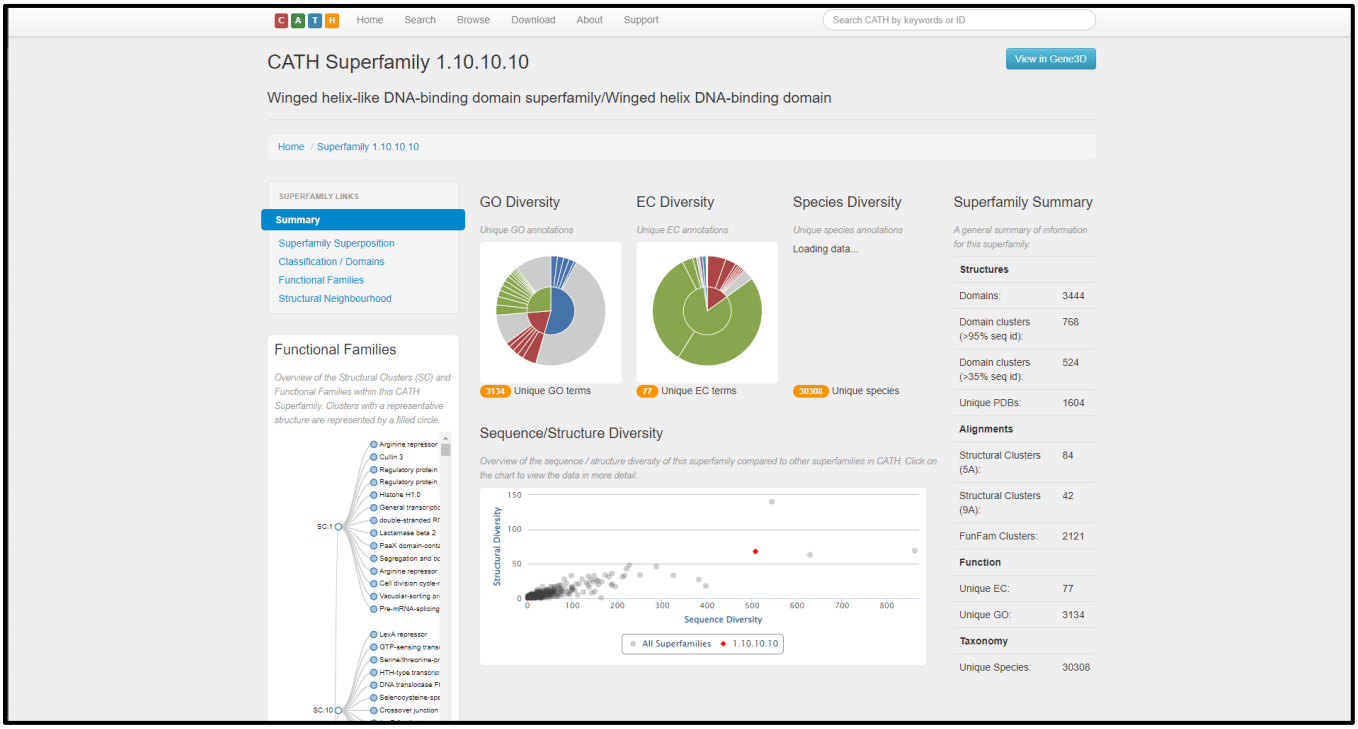


Fig3.1. Summary of CATH superfamily for Leucine

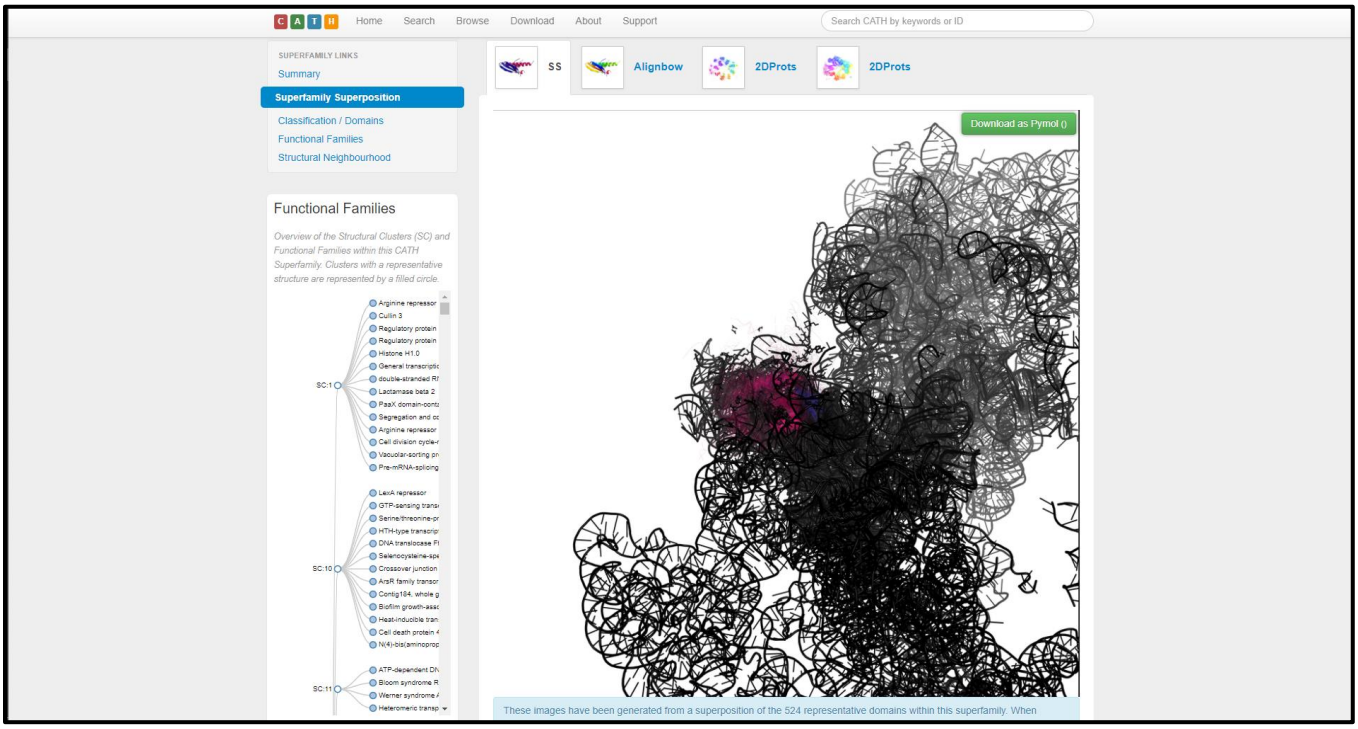


Fig3.2. Superfamily superposition under CATH superfamily for leucine

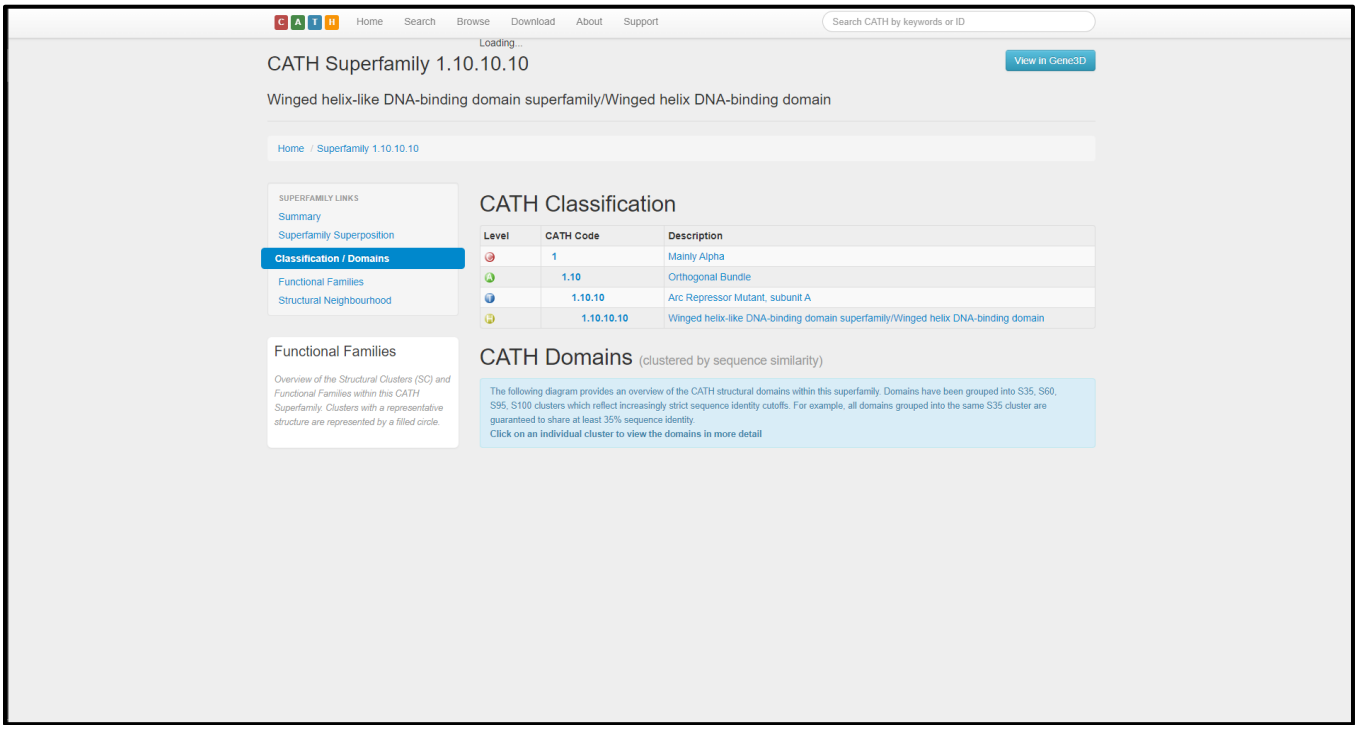


Fig3.3. Classification / Domain of CATH superfamilies for Leucine

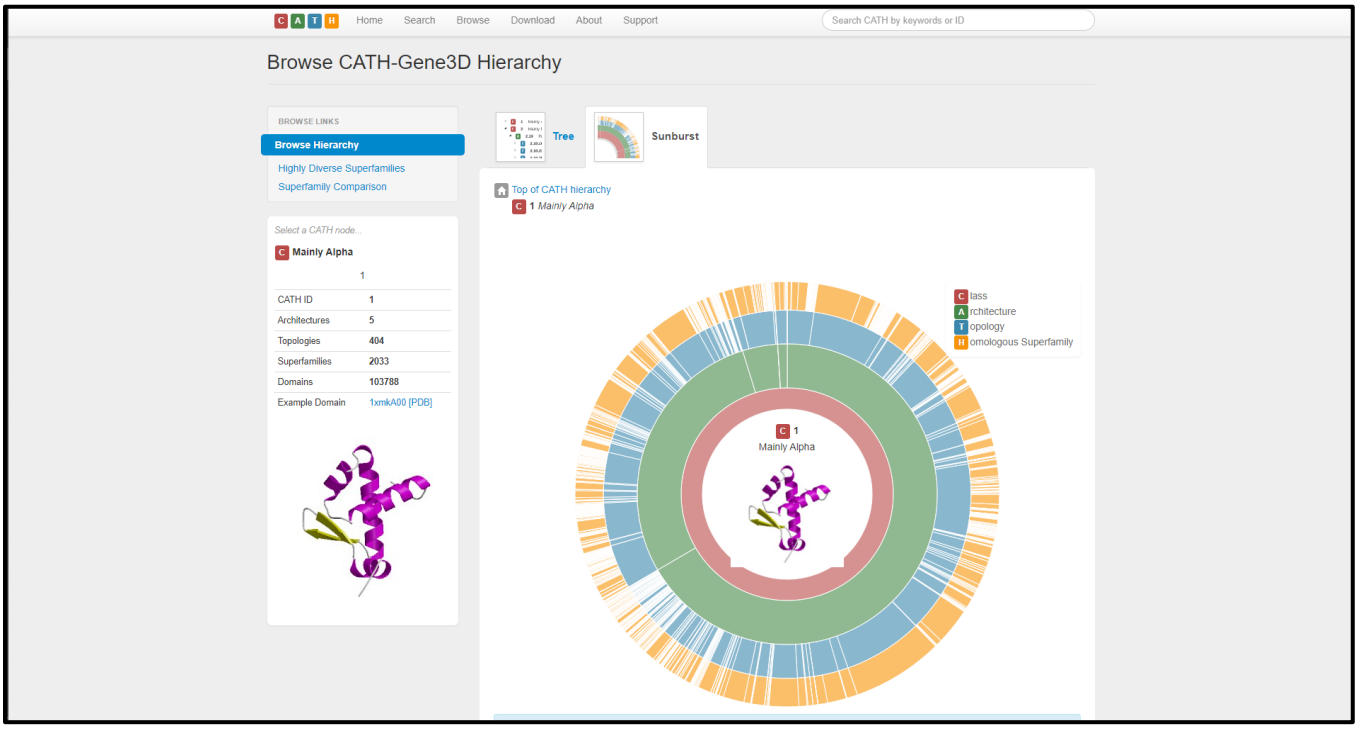


Fig3.4. Sunburst diagram of Alpha domain of CATH superfamilies for leucine

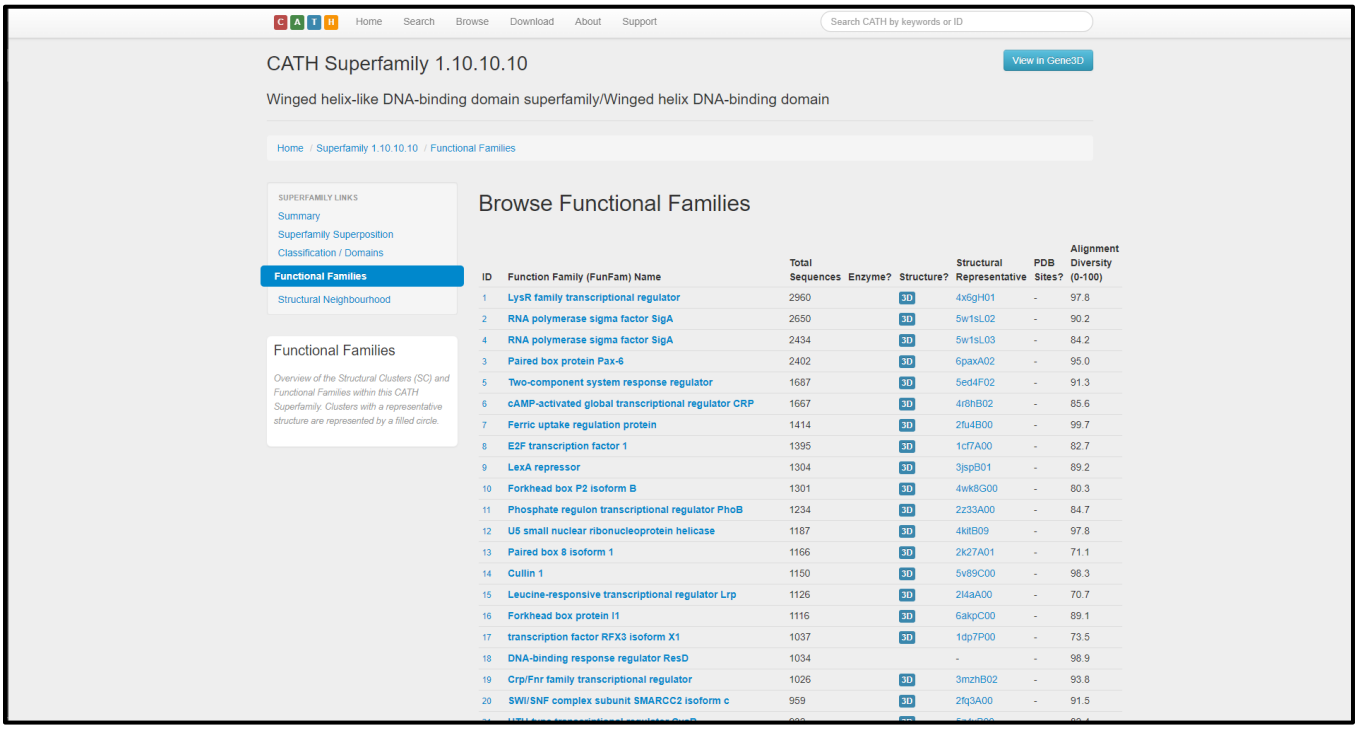


Fig3.5. Functional Families of CATH superfamilies for leucine

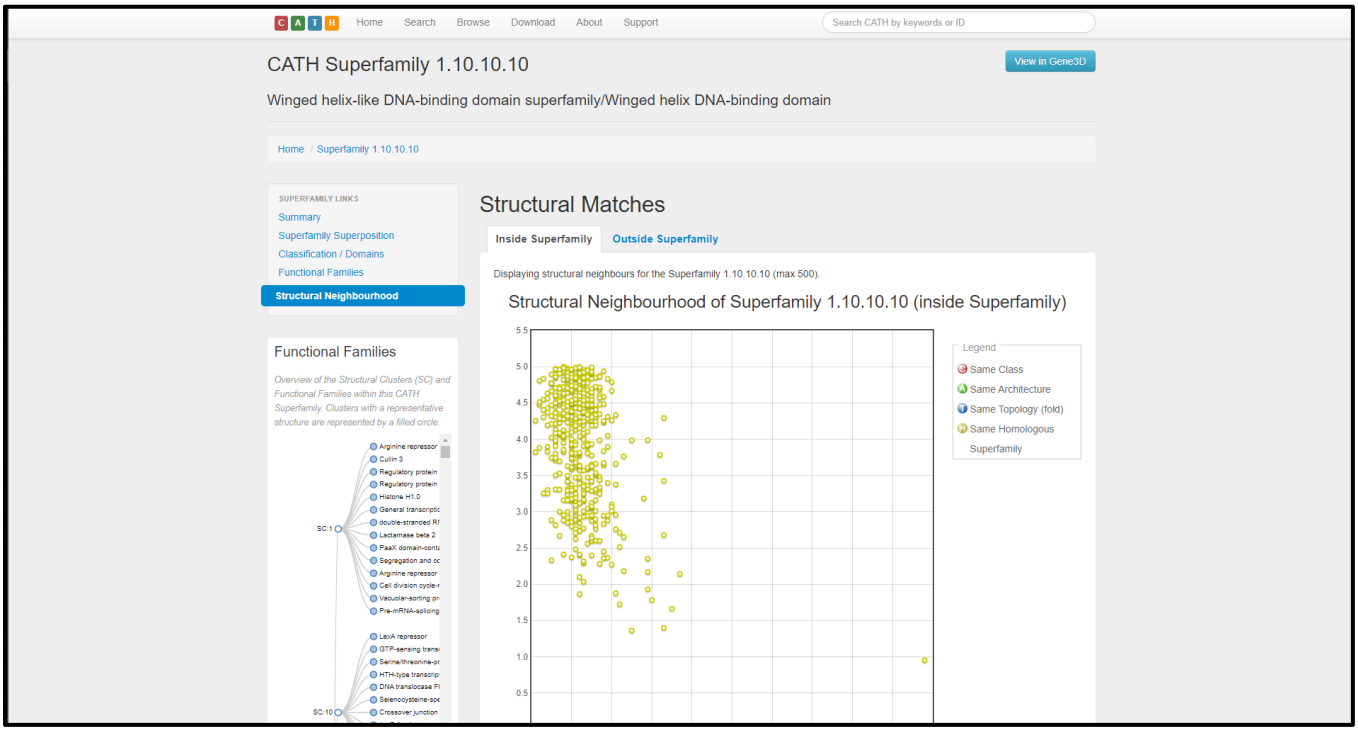


Fig3.6. Structural neighborhood of CATH superfamilies for leucine (Inside Superfamily)

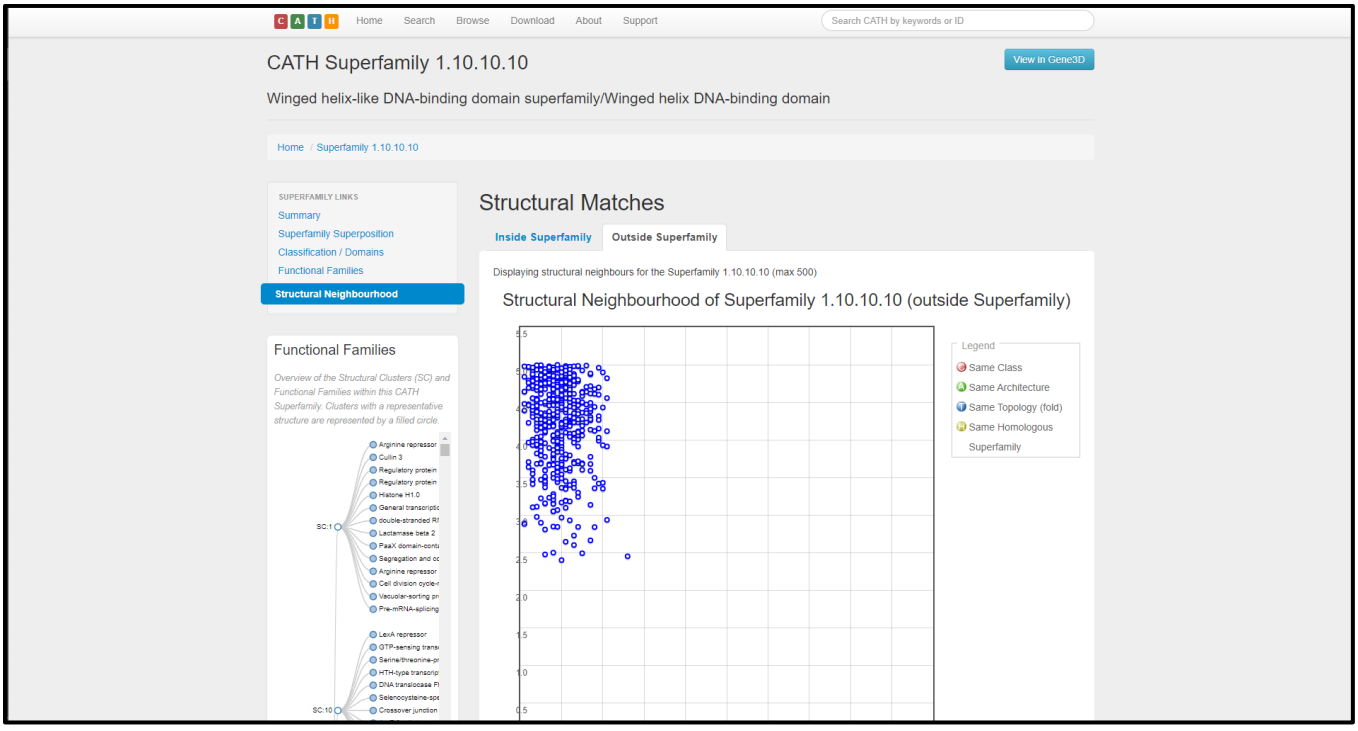


Fig3.7. Structural neighborhood of CATH superfamilies for leucine (Outside Superfamily)

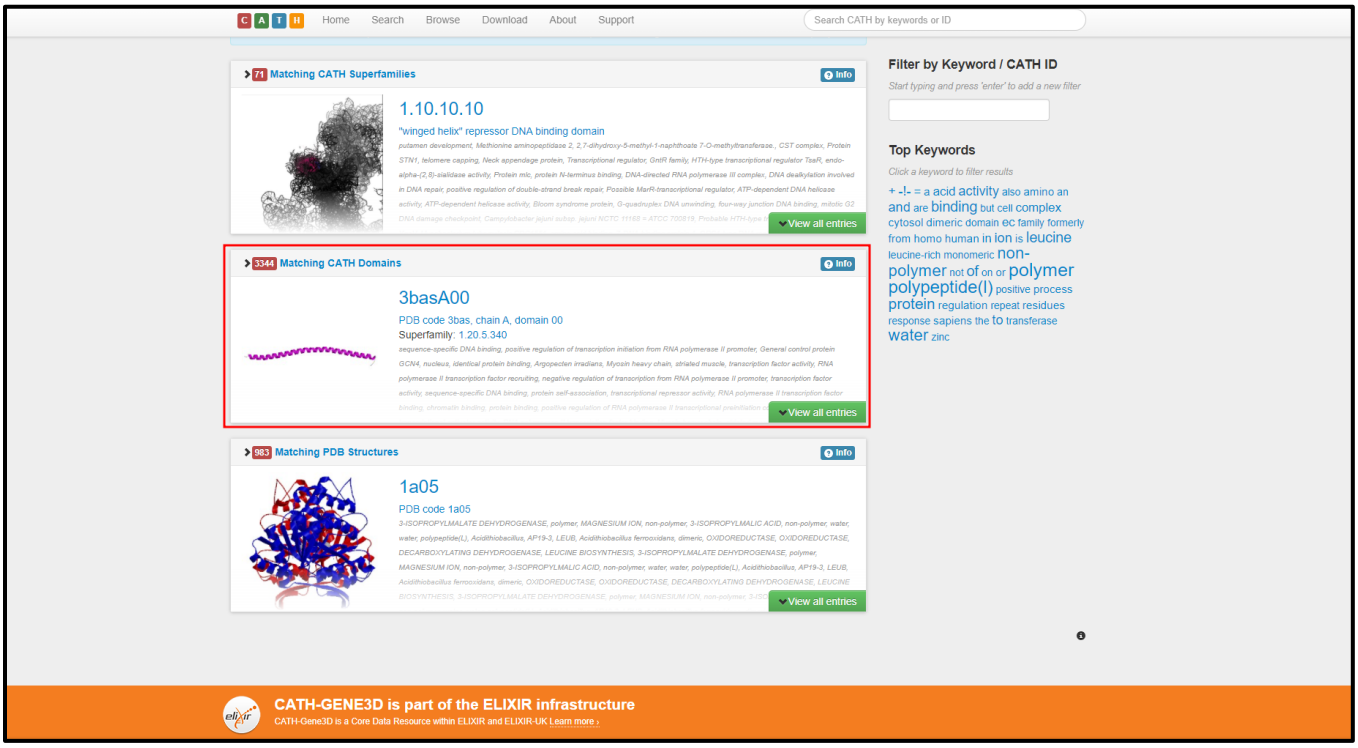


Fig4. Matching CATH domains for Leucine

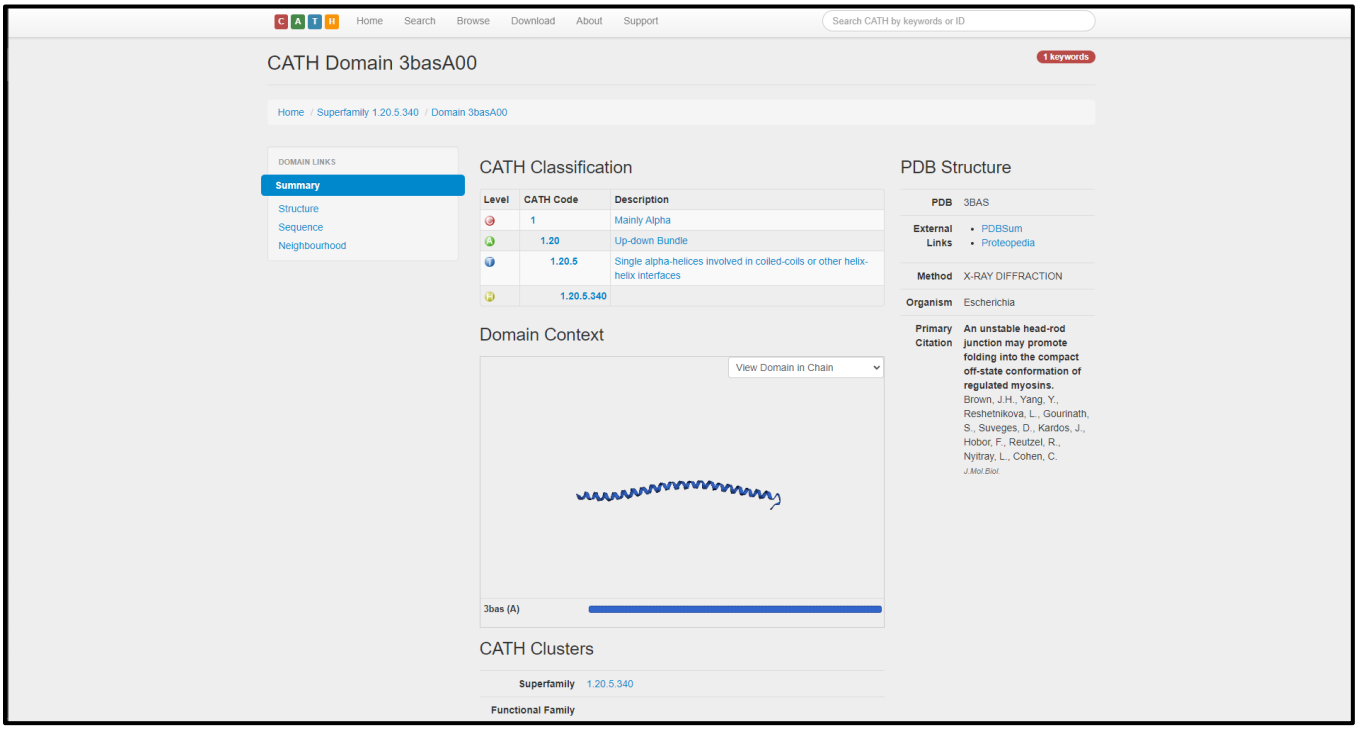


Fig4.1. Summary of matching CATH domains for Leucine

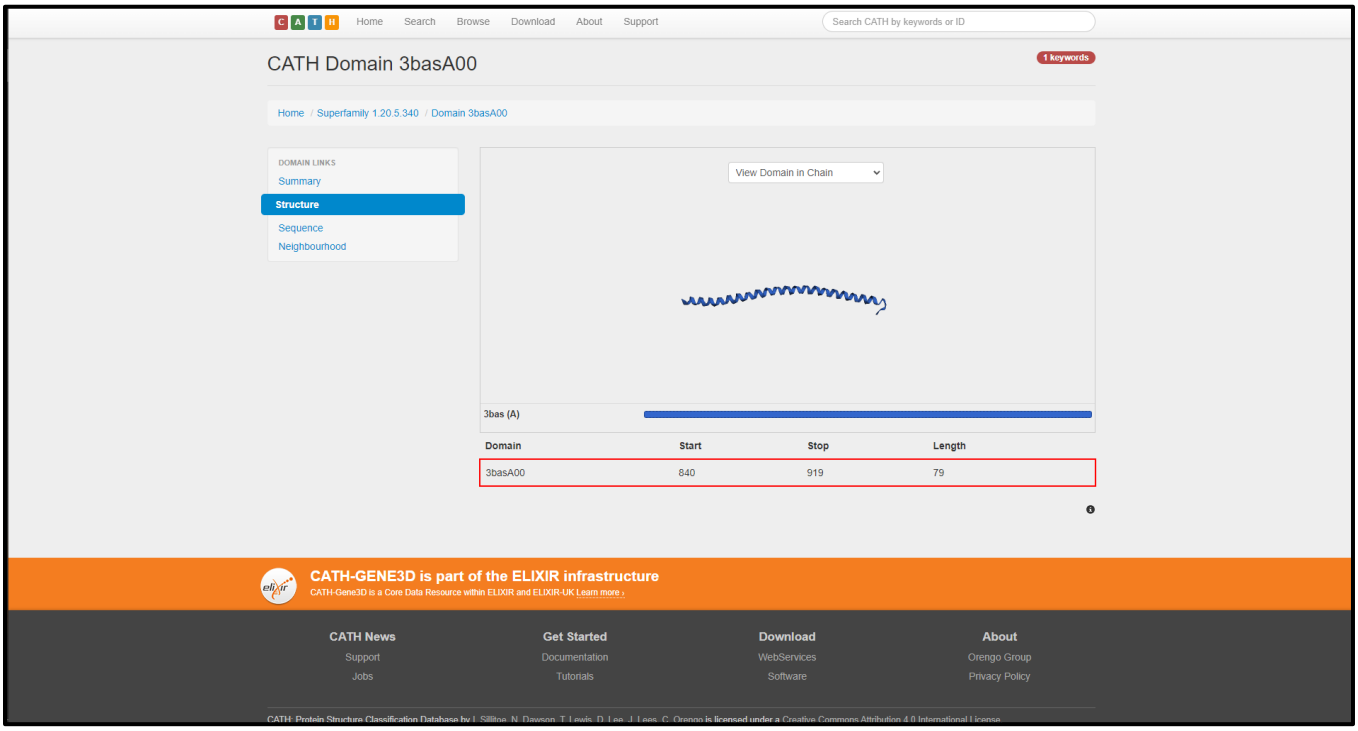


Fig4.2. Structure of matching CATH domains for Leucine

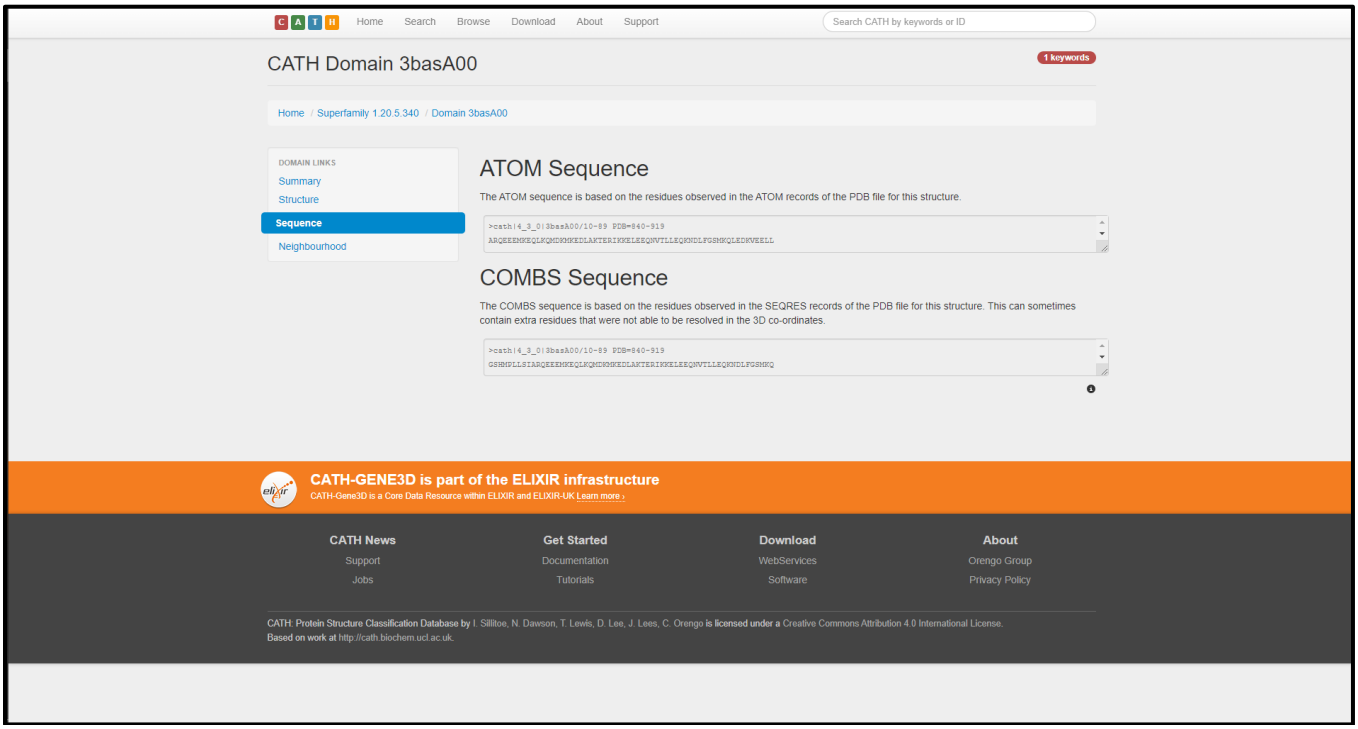


Fig4.3. ATOM and COMBS sequence of matching CATH domains for Leucine

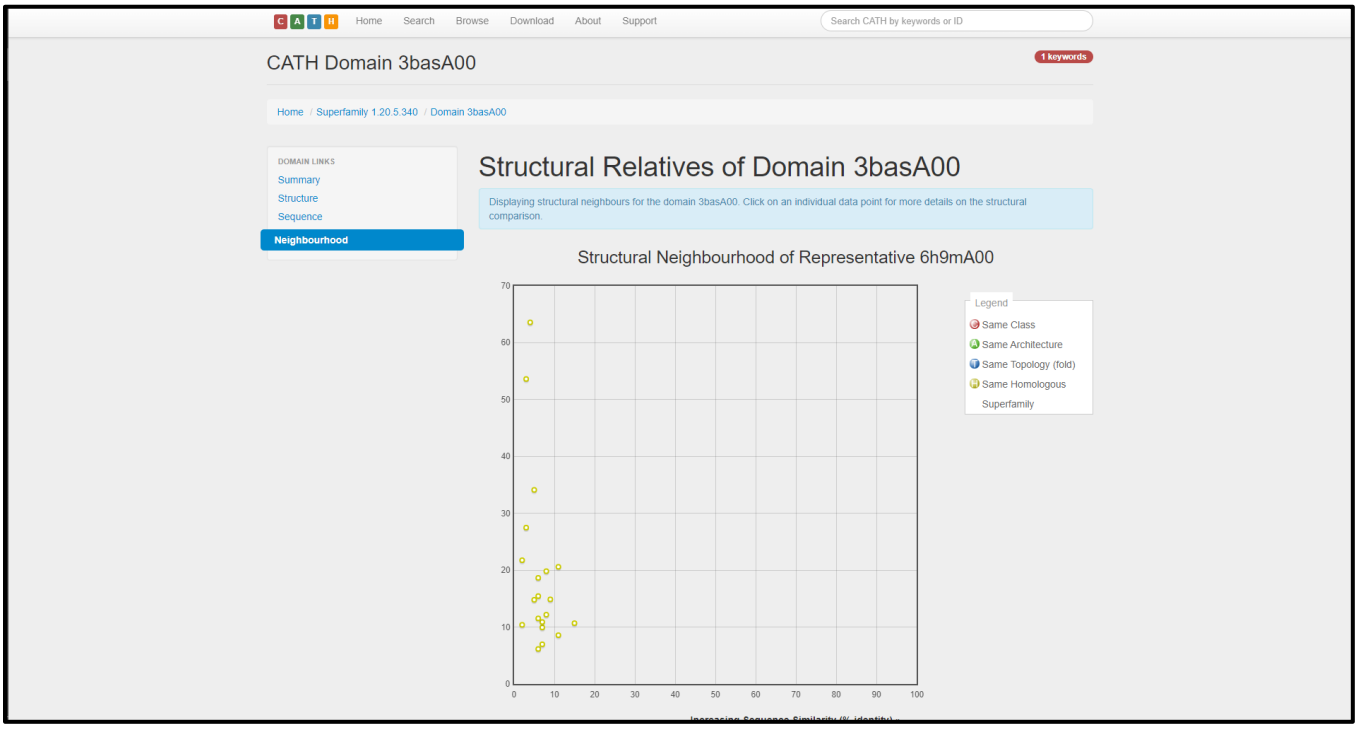


Fig4.4. Structural relatives of matching CATH domains for leucine

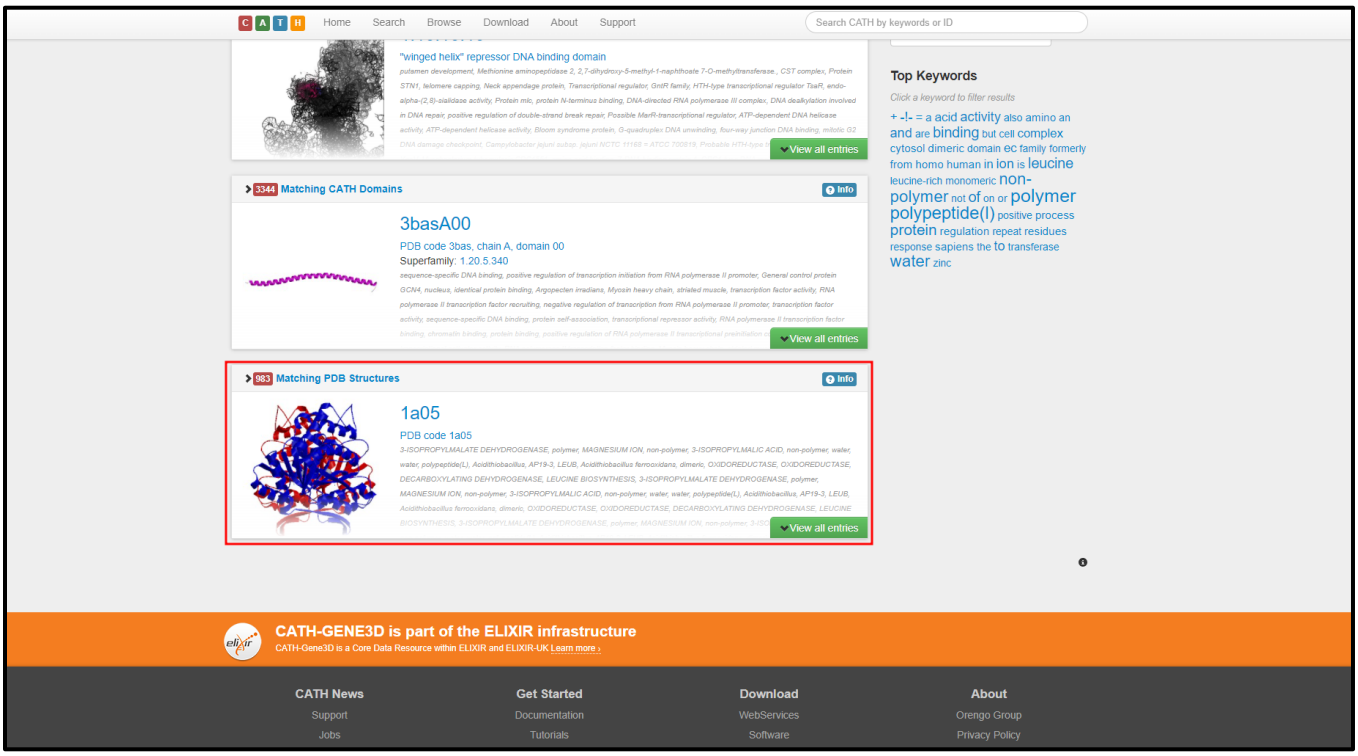


Fig5. Matching PDB structure for leucine

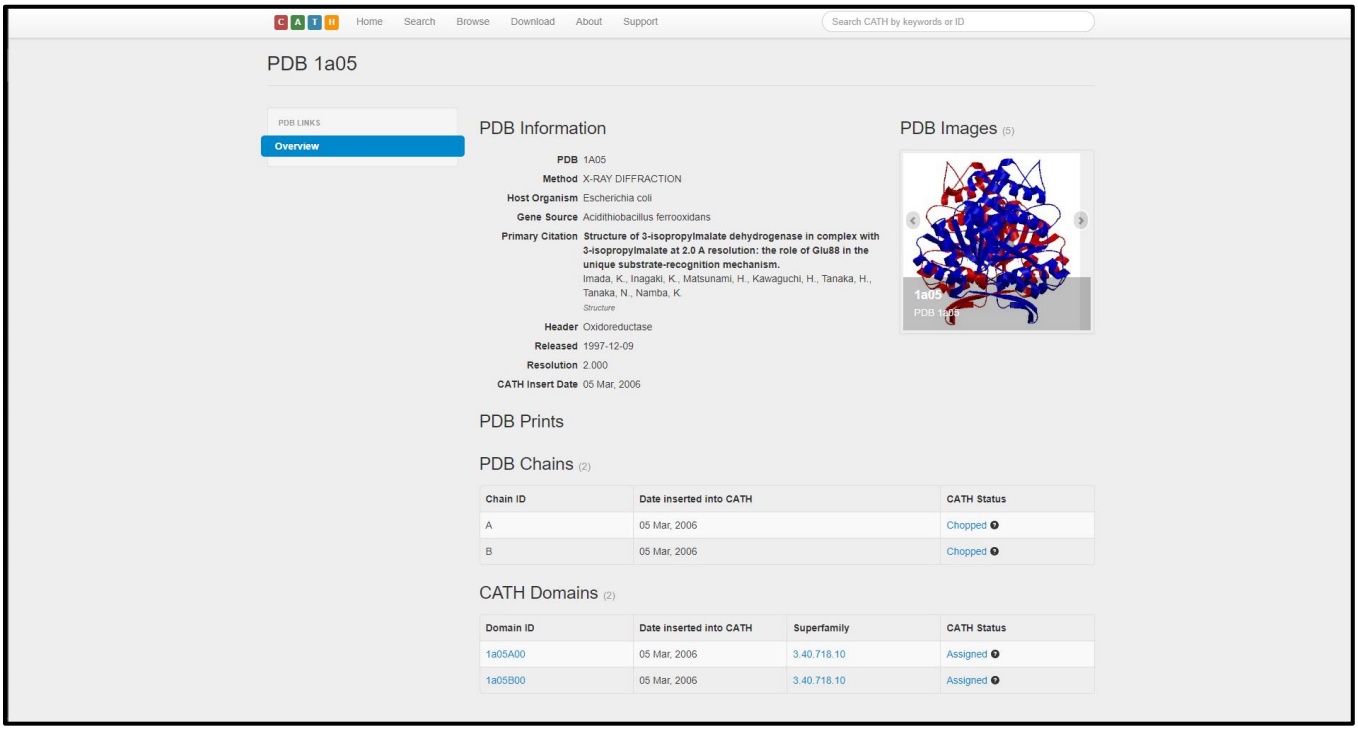


Fig5.1. Overview of matching PDB structures for leucine

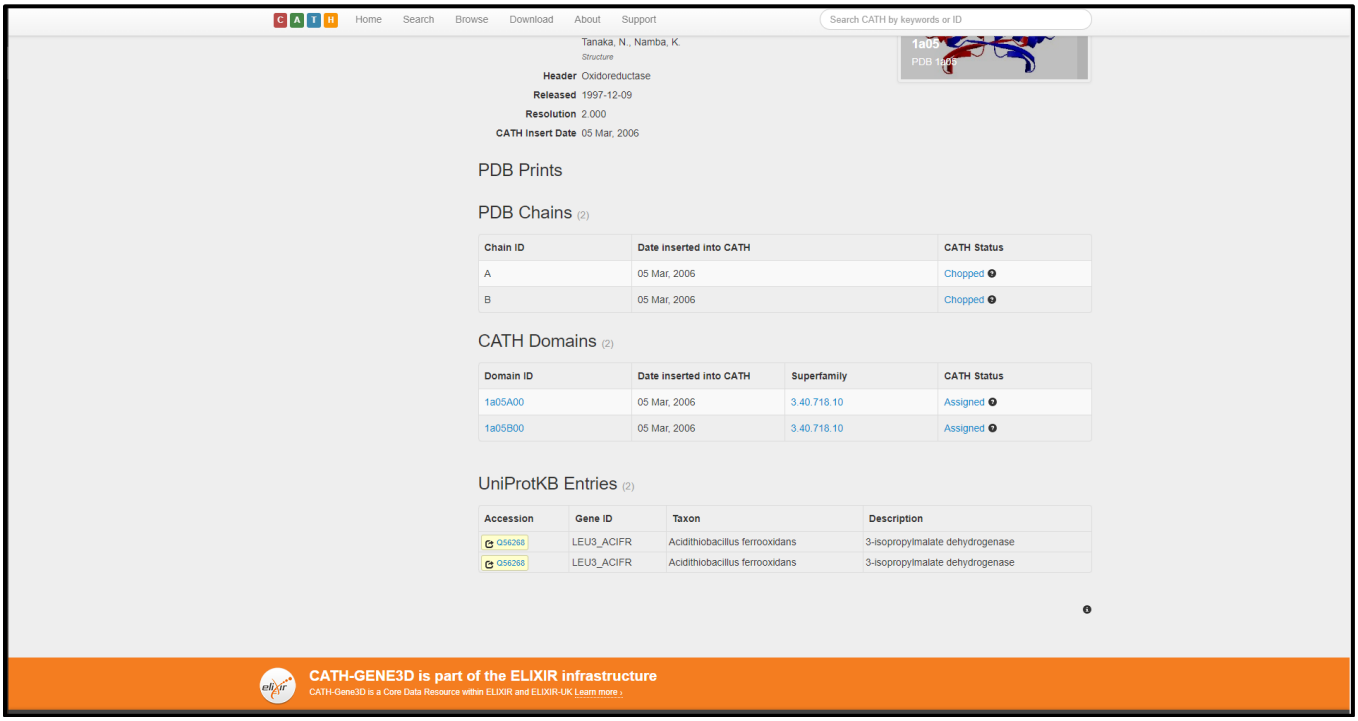


Fig5.2. Prints, Chains, CATH domains and UniProtKB entries for matching PDB structure for leucine

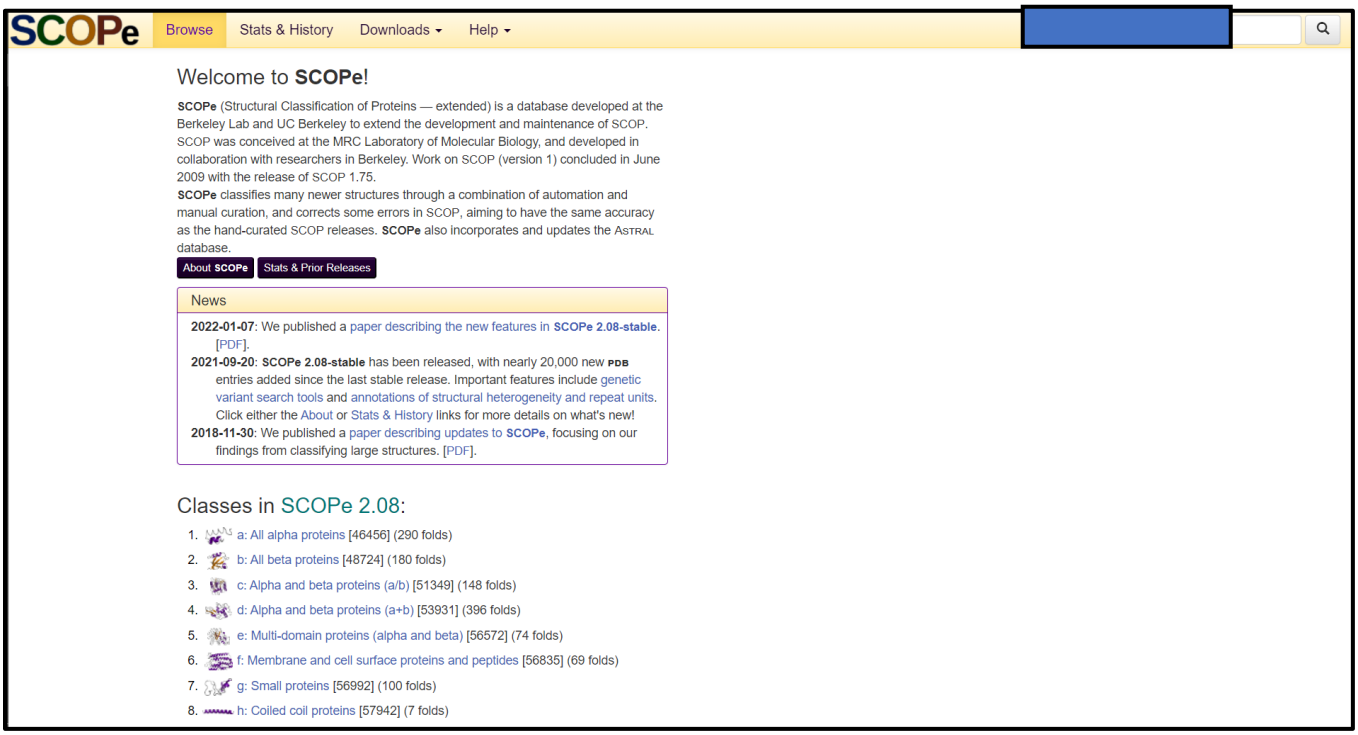


Fig6. Homepage of SCOPe Database with query leucine

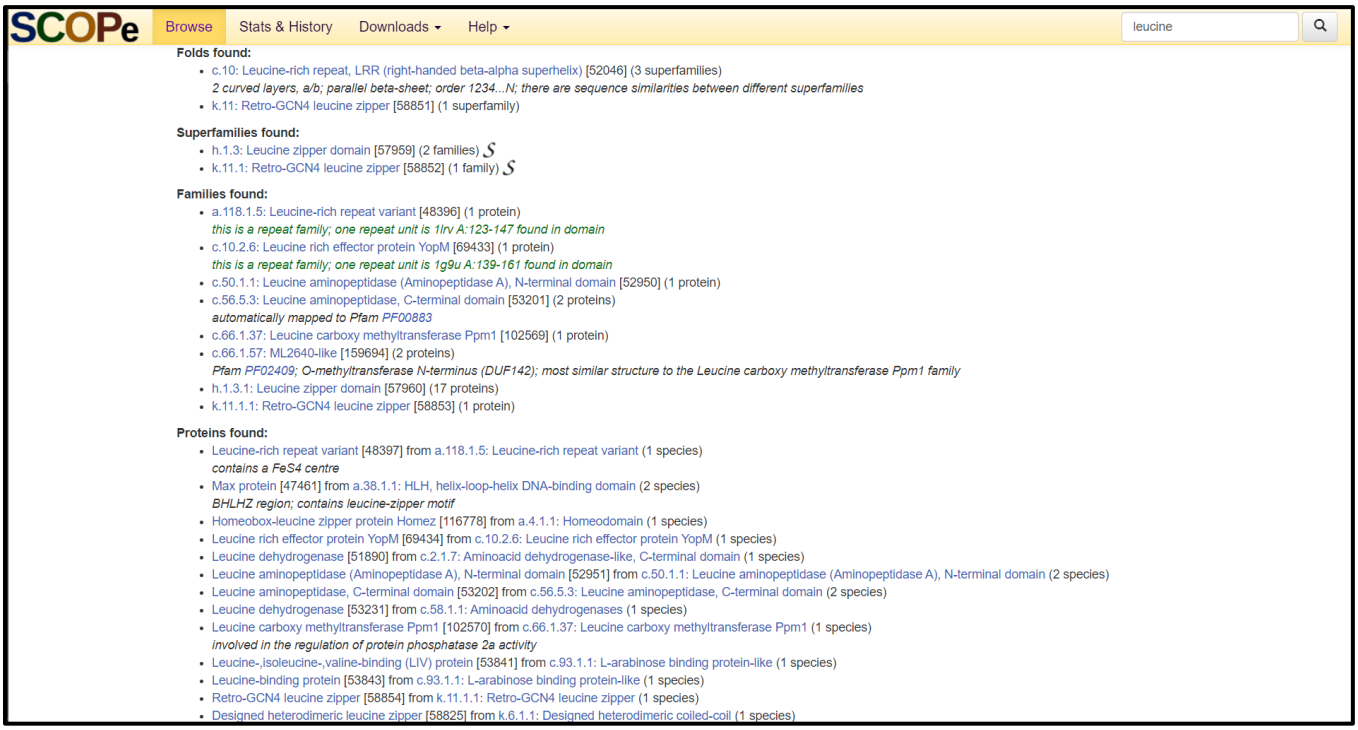


Fig7. Result page of SCOPe for query leucine

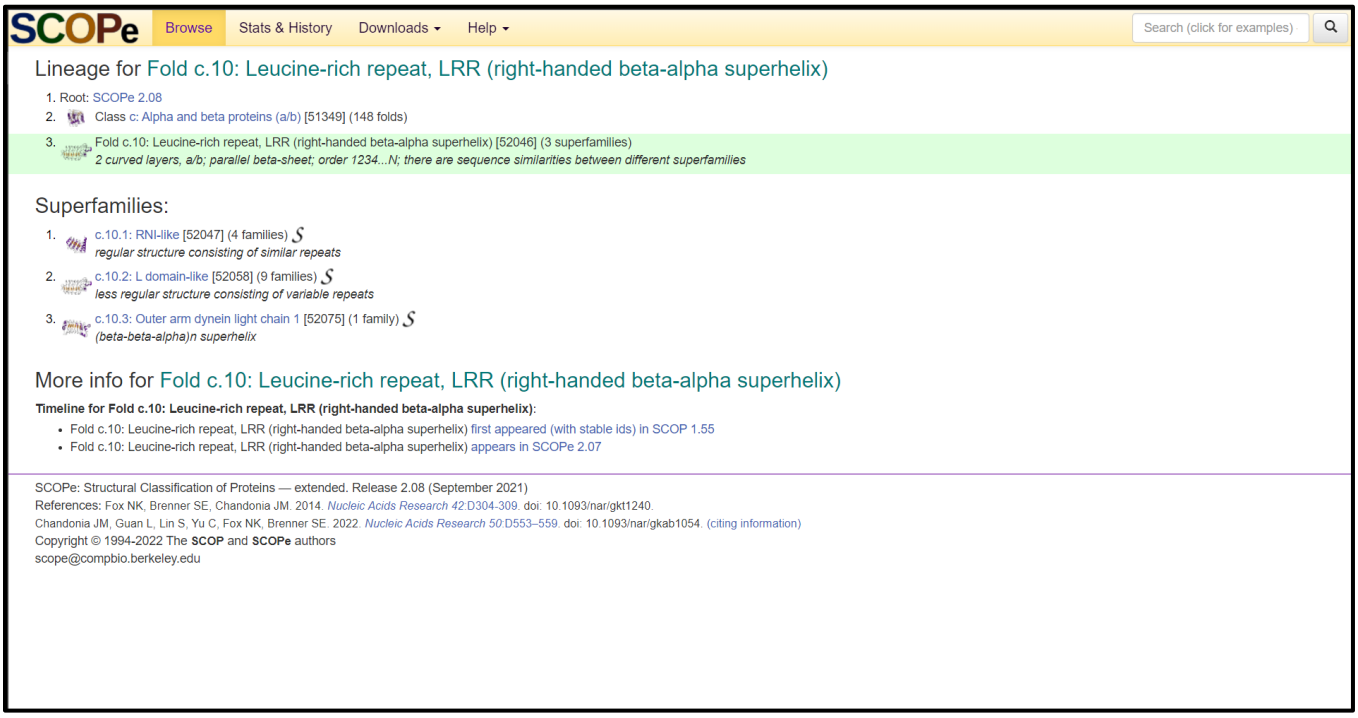


Fig8. Random result from "Folds found" of SCOPe database for leucine

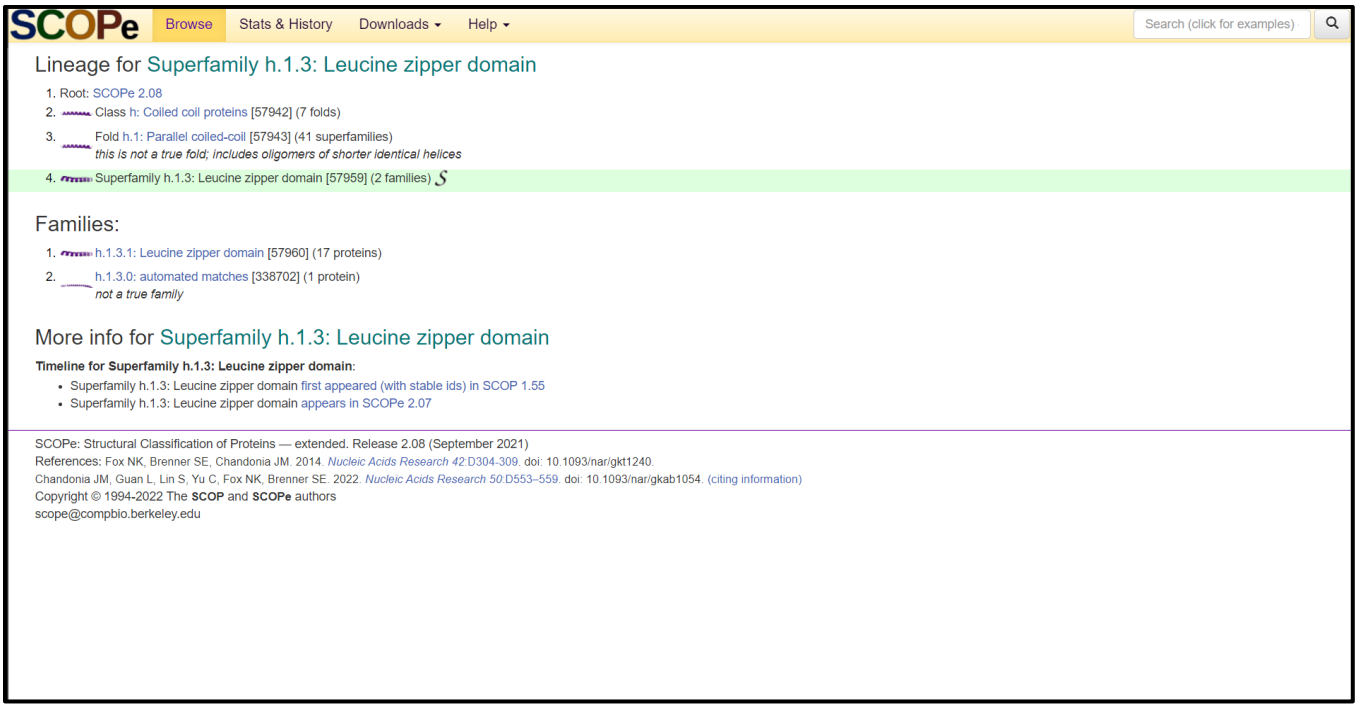


Fig8. Random result from "Superfamilies" of SCOPe database for leucine

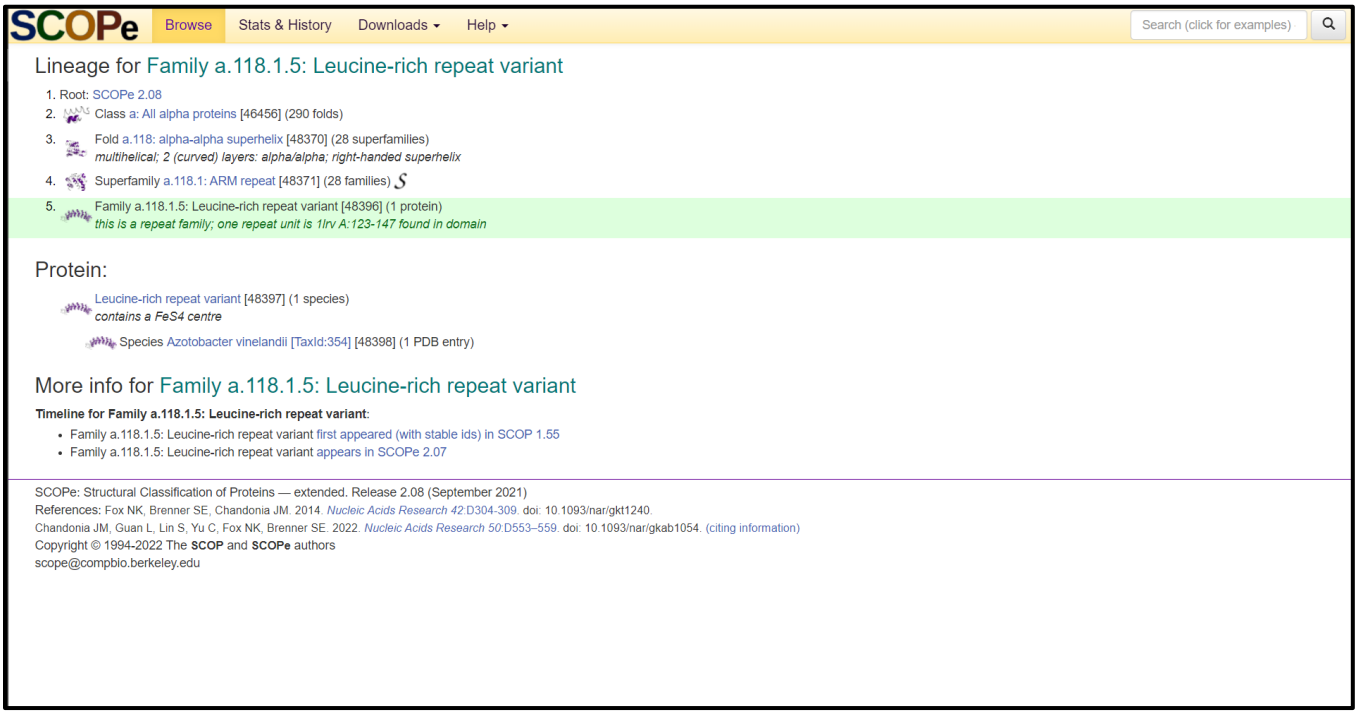


Fig9. Random result from "Families" of SCOPe database for leucine

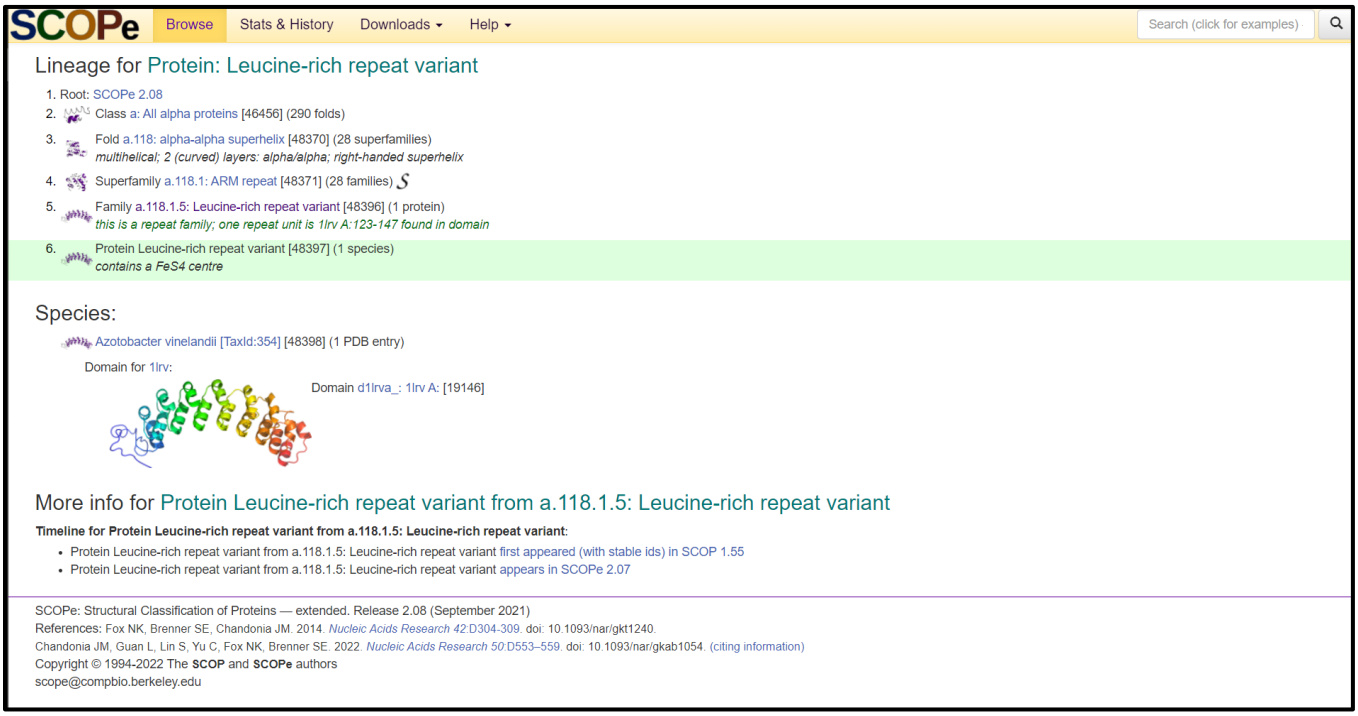


Fig10. Random result from "Proteins found" of SCOPe database for leucine

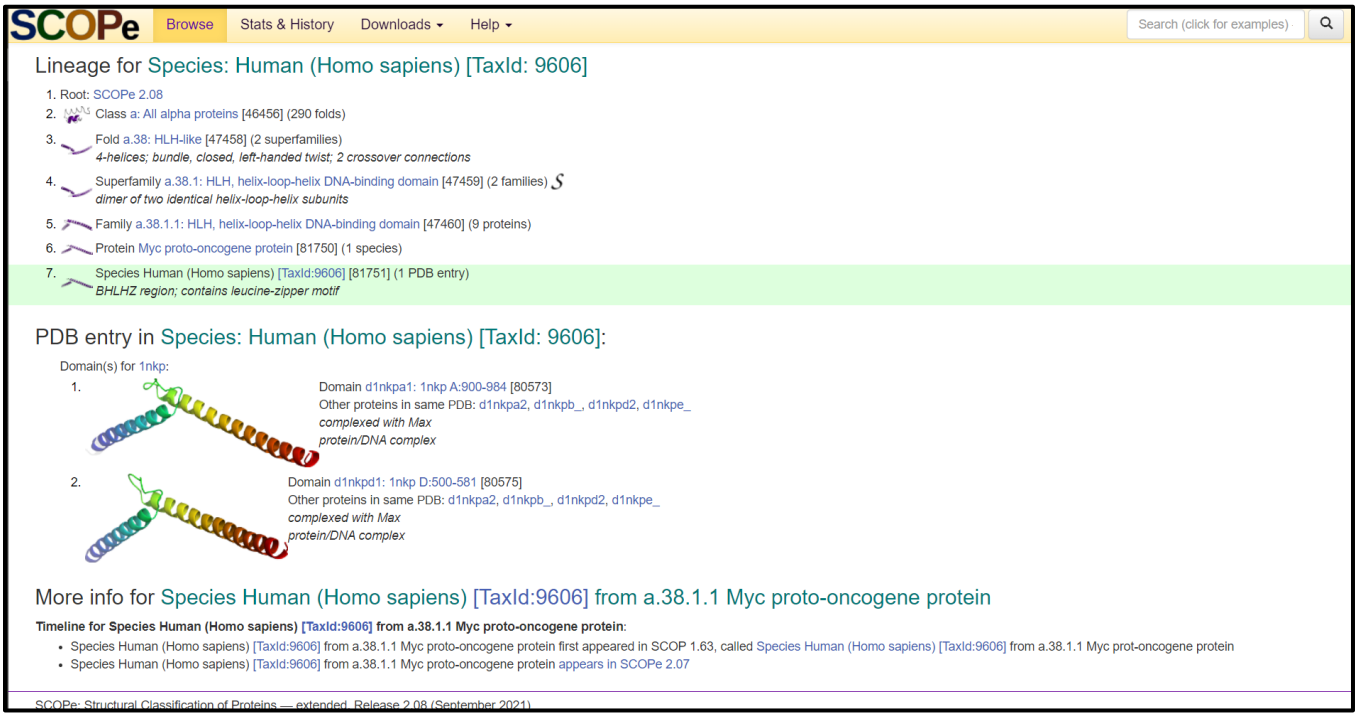


Fig11. Random result from "Species found" of SCOPe database for leucine

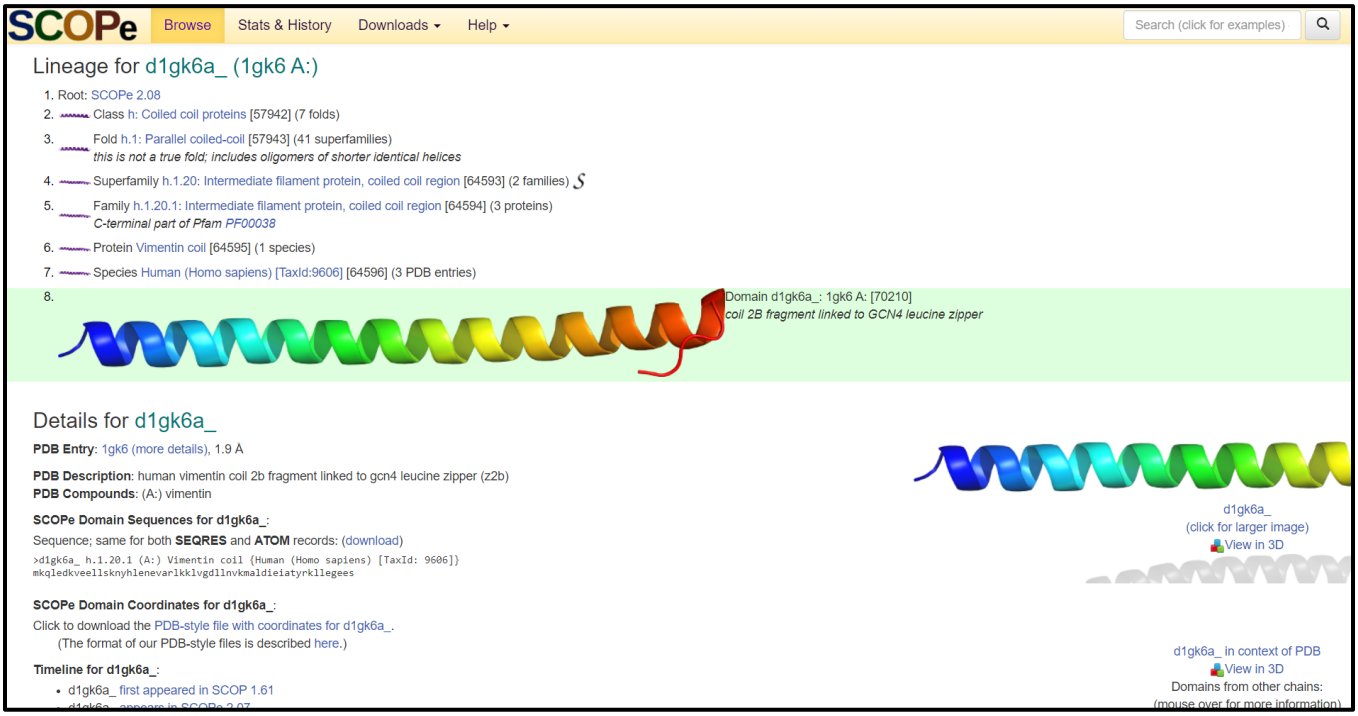


Fig11. Random result from "Domains found" of SCOPe database for leucine

Results:

CATH:

- ➔ In protein classification using CATH database for query leucine. It shows,
  - ➔ 71 matching CATH superfamilies.
  - ➔ 3344 matching CATH Domains
  - ➔ 983 matching PDB structures
- ➔ For CATH superfamilies we saw the Winged helix-like DNA-binding domain superfamily.
- ➔ For CATH Domains we saw CATH Domain 3basA00
- ➔ For PDB structures we saw PDB 1A05 from Escherichia Coli

SCOPe:

- ➔ In SCOPe database the result was divided into 6 sections:
  - ➔ Folds found [Fold c.10: Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)]
  - ➔ Superfamilies found [Superfamily h.1.3: Leucine zipper domain]
  - ➔ Family found [Family a.118.1.5: Leucine-rich repeat variant]
  - ➔ Protein Found [Protein: Leucine-rich repeat variant]
  - ➔ Species found [Species: Human (Homo sapiens) [TaxId: 9606]]
  - ➔ Domains found [d1gk6a_ (1gk6 A:)]

Conclusions:

- ➔ The CATH database is valuable for biologists and bioinformaticians alike.
- ➔ For biologists with very specific tasks, browsing for individual domains is made easy by the user-friendly web interface.
- ➔ For bioinformaticians with a focus on large-scale analyses can find complete datasets available for downloading.
- ➔ Thus, working with CATH is remarkably uncomplicated.
- ➔ Updates are frequent, and, given the significant upcoming extension with horizontal layers complementary to the hierarchical structure, CATH is likely to become an even more valuable resource in the future.
- ➔ Since it was created, the development of SCOPe has been always guided by its user's feedback and needs.
- ➔ The automation in crystallography and advances in cryo-electron microscopy open a new era in structural biology and with it come new demands for data suitable for modelling of large proteins and protein complexes.
- ➔ In addition to a range of new annotations, SCOPe has introduced new functionalities that support relatively easy retrieval and assembly of independently determined, structurally characterized parts of proteins of interest.
- ➔ They will continue updating the database and providing regular releases while working on steadily increasing the coverage of structural data and adding new functionalities to the web interface.

References:

1. *Leucine*. Leucine - Health Encyclopedia - University of Rochester Medical Center. (n.d.). Retrieved February 27, 2022, from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=19&contentid=Leucine#:~:text=Leucine%20is%20one%20of%20the,enough%20of%20these%20amino%20acids.
2. Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235
3. Murzin, A. G. (1995). *Journal of Molecular Biology*, *247*(4), 536–540. https://doi.org/10.1006/jmbi.1995.0159
4. CATH Database. (2022b, February 21). CATH Database. Retrieved February 21, 2022, from https://www.cathdb.info/
5. *Search cath*. CATH Search: Browse. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/search?q=Leucine
6. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10
7. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10/superposition
8. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10/classification
9. Browse cath-gene3d hierarchy. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/browse/sunburst?from_cath_id=1
10. *Cath superfamily 1.10.10.10 - cathdb.info*. (n.d.). Retrieved February 26, 2022, from https://www.cathdb.info/version/latest/superfamily/1.10.10.10/alignments
11. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10/structure
12. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00
13. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00/structure
14. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00/neighbourhood
15. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00/sequence
16. PDB 1A05. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/pdb/1a05
17. *Scop.berkeley.edu*. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/search/?ver=2.08&key=leucine
18. *Scope 2.08: Domain d1gk6a_: 1GK6 A:* SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=70210
19. *Scope 2.08: Family A.118.1.5: Leucine-rich repeat variant*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=48396
20. *Scope 2.08: Fold c.10: Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=52046
21. *Scope 2.08: Protein: Leucine-rich repeat variant*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=48397
22. *Scope 2.08: Species: Human (homo sapiens) [taxid: 9606]*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=81751
23. *Scope 2.08: Superfamily H.1.3: Leucine zipper domain*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=57959

24. *Structural classification of proteins - extended. release 2.08 (September 2021)*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/

DATE: 20-03-22

WEBLEM 3

Introduction to tertiary structure prediction

Proteins are involved in many cell activities (e.g., molecular transport, mechanical functions, message exchange) thus **knowing their 3D structure is crucial** in order to understand their function. **Protein tertiary structure prediction** is a research field which aims to **create models and software tools** able to predict the **three-dimensional shape of protein molecules** by describing the spatial disposition of each of its atoms starting from the sequence of its amino acids. There exist exact methods to **resolve the molecular structure with high precision**, but they are both time and resource consuming. **Computational based software techniques** can predict the tertiary structure of a protein with **acceptable precision** for many applications with high efficiency allowing for **genome-wide investigations**, otherwise not feasible.

Having a **computer-generated three-dimensional model** of a protein of interest has many ramifications, assuming it is reasonably correct. It may be of use for the **rational design of biochemical experiments**, such as **site-directed mutagenesis, protein stability, or functional analysis**. In addition to serving as a **theoretical guide to design experiments** for protein characterization, the model can help to **rationalize the experimental results** obtained with the protein of interest. In short, the modelling study helps to advance our **understanding of protein functions**.

METHODS:

There are **three computational approaches** to protein three-dimensional structural modelling and prediction. They are **homology modelling, threading,** and **ab initio prediction.**

HOMOLOGY MODELLING:

As the name suggests, **homology modelling** predicts protein structures based on **sequence homology** with known structures. It is also known as comparative modelling. The principle behind it is that if two proteins share a **high enough sequence similarity**, they are likely to have very **similar three-dimensional structures**. If one of the protein sequences has a **known structure**, then the structure **can be copied to the unknown protein** with a high degree of confidence. Homology modelling produces an **all-atom model** based on **alignment with template proteins.**

The overall homology modelling procedure consists of six steps.

1. **Template Selection** which involves identification of homologous sequences in the protein structure database to be used as templates for modelling
2. **Alignment** of target and template sequences.
3. **Building a framework structure** for the target protein consisting of main chain atoms.
4. **Refine and optimize** the entire model according to energy criteria.
5. **Evaluation** of the overall quality of the model obtained.

A number of **comprehensive modelling programs** are able to perform the complete procedure of homology modelling in an automated fashion. The **automation requires assembling a pipeline** that includes target **selection, alignment, model generation, and model evaluation.**

MODELLER:

MODELLER is a computer program for **comparative protein structure modelling**. In the simplest case, the input is an **alignment of a sequence** to be modelled with the **template structures**, the atomic coordinates of the templates, and a simple script file. **MODELLER** then automatically **calculates a model** containing all **non-hydrogen atoms**, within minutes on a modern PC and with no user intervention. Apart from model building, **MODELLER** can perform additional auxiliary tasks, including **fold assignment**, **alignment of two protein sequences** or their profiles, **multiple alignment** of protein sequences and/or structures, **calculation of phylogenetic trees**, and **de novo modelling of loops** in protein structures.

THREADING AND FOLD RECOGNITION:

There are only **small number of protein folds available** (<1,000), compared to millions of protein sequences. This means that protein structures tend to be **more conserved** than protein sequences. Consequently, many proteins can share a **similar fold** even in the absence of **sequence similarities**. This allowed the development of computational methods to predict protein structures **beyond sequence similarities**. To determine whether a **protein sequence adopts** a known **three-dimensional structure** fold relies on **threading and fold recognition** methods. By definition, threading or structural fold recognition predicts the **structural fold** of an **unknown protein sequence** by fitting the sequence into a **structural database** and selecting the **best-fitting fold**. The comparison emphasizes matching of **secondary structures**, which are most evolutionarily conserved. Therefore, this approach can **identify structurally similar proteins** even without detectable sequence similarity.

The algorithms can be classified into two categories, **pairwise energy based** and **profile based**. The pairwise energy–based method was originally referred to as **threading** and the profile-based method was originally defined as **fold recognition**. However, the two terms are now often used **interchangeably without distinction** in the literature. A number of threading and fold recognition programs are available using **either or both prediction strategies.**

I-TASSER:

I-TASSER server is an on-line platform that implements the **I-TASSER based algorithms** for protein structure and function predictions. It allows academic users to **automatically generate high-quality model predictions** of 3D structure and **biological function** of protein molecules from their amino acid sequences**.** When user submits an amino acid sequence, the server **first tries to retrieve template proteins of similar folds** (or super-secondary structures) from the PDB library by LOMETS, a locally installed meta-threading approach.

In the **second step**, the continuous fragments excised from the PDB templates are reassembled into full-length models by **replica-exchange Monte Carlo simulations** with the threading unaligned regions (mainly loops) built by **ab initio modelling**. In cases where no appropriate template is identified by **LOMETS, I-TASSER** will build the whole structures by ab initio modelling. The low free-energy states are identified by **SPICKER** through clustering the simulation decoys.

In the **third step**, the fragment assembly simulation is performed again starting from the **SPICKER cluster centroids**, where the spatial restrains collected from **both** the **LOMETS** templates and the **PDB** structures by TM-align are used to **guide the simulations**. The purpose of the **second iteration** is to **remove the steric clash** as well as to **refine the global topology** of the cluster centroids. The decoys generated in the second simulations are then clustered and the lowest energy structures are selected. The **final full-atomic models** are obtained by **REMO** which builds the atomic details from the selected I-TASSER decoys through the optimization of the **hydrogen-bonding network.**

For **predicting the biological function** of the protein, the I-TASSER server matches the **predicted 3D models** to the **proteins in 3 independent libraries** which consist of proteins of known **enzyme classification (EC) number**, **gene ontology (GO) vocabulary**, and **ligand-binding sites.** The final results of function predictions are deduced from the consensus of **top structural matches** with the function scores calculated based on the confidence score of the I-TASSER structural models, the **structural similarity** between model and templates as **evaluated by TM-score**, and the sequence identity in the structurally aligned regions.

1. **What is C-score?**

   C-score is a **confidence score** for estimating the quality of predicted models by I-TASSER. It is calculated based on the **significance of threading template alignments** and the **convergence parameters** of the structure assembly simulations. C-score is typically in the **range of [-5,2],** where a C-score of **higher value** signifies a model with a **high confidence** and vice-versa.

2. **What is TM-score?**

TM-score is a recently proposed scale for measuring the **structural similarity between two structures**. The purpose of proposing TM-score is to **solve** the problem of **RMSD** which is sensitive to the local error. Because RMSD is an **average distance** of all residue pairs in **two structures**, a **local** error (e.g., a misorientation of the tail) will arise a **big RMSD value** although the global topology is correct. In TM-score, however, the **small distance** is **weighted stronger** than the **big distance** which makes the score **insensitive to the local modelling error**. A **TM-score >0.5** indicates a model of correct topology and a TM-score<0.17 means a random similarity. These cut-off does not depend on the protein length.

3. What is difference and relationship between C-score and TM-score?

TM-score (or RMSD) is a known **standard** for measuring **structural similarity** between two structures which are usually used to measure the **accuracy of structure modelling** when the native structure is known, while **C-score** is a metric that I-TASSER developed to **estimate the confidence of the modelling**. In case where the native structure is not known, it becomes necessary to predict the quality of the modelling prediction, i.e., what is the distance between the predicted model and the native structures? To answer this question, we tried **predicting** the **TM-score** and **RMSD** of the predicted models relative the native structures based on the **C-score**.

In a benchmark test set of 500 non-homologous proteins, we found that C-score is highly correlated with TM-score and RMSD. Correlation coefficient of C-score of the first model with TM-score to the native structure is 0.91, while the coefficient of C-score with RMSD to the native structure is 0.75. These data lay the base for the reliable prediction of the TM-score and RMSD using C-score. In the output section, I-TASSER only reports the quality prediction (TM-score and RMSD) for the first model, because it was found that the correlation between C-score and TM-score is weak for lower rank models. However, the C-score is listed for all models just for a reference.

AB INITIO PROTEIN STRUCTURAL PREDICTION

The limited knowledge of protein folding forms the basis of ab initio prediction. As the name suggests, the ab initio prediction method attempts to produce all-atom protein models based on sequence information alone without the aid of known protein structures. The perceived advantage of this method is that predictions are not restricted by known folds and that novel protein folds can be identified. However, because the physicochemical laws governing protein folding are not yet well understood, the energy functions used in the ab initio prediction are at present rather inaccurate. The folding problem remains one of the greatest challenges in bioinformatics today.

Current ab initio algorithms are not yet able to accurately simulate the protein folding process. They work by using some type of heuristics. Because the native state of a protein structure is near energy minimum, the prediction programs are thus designed using the energy minimization principle. These algorithms search for every possible conformation to find the one with the lowest global energy. However, searching for a fold with the absolute minimum energy may not be valid in reality. This contributes to one of the fundamental flaws of this approach. In addition, searching for all possible structural conformations is not yet computationally feasible. It has been estimated that, by using one of the world's fastest supercomputers (one trillion operations per second), it takes 10 20 years to sample all possible conformations of a 40-residue protein. Therefore, some type of heuristics must be used to reduce the conformational space to be searched. Some recent ab initio methods combine fragment search and threading to yield a model of an unknown protein. The following web program is such an example using the hybrid approach.

ROBETTA:

The ROBETTA server provides automated tools for protein structure prediction and analysis. For structure prediction, sequences submitted to the server are parsed into putative domains and structural models are generated using either comparative modelling order novo structure prediction methods. If a confident

match to a protein of known structure is found using BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative modelling. If no match is found, structure predictions are made using the de novo Rosetta fragment insertion method. Experimental nuclear magnetic resonance (NMR) constraints data can also be submitted with a query sequence for RosettaNMR de novo structure determination. Other current capabilities include the prediction of the effects of mutations on protein–protein interactions using computational interface alanine scanning. The Rosetta protein design and protein–protein docking methodologies will soon be available through the server as well.

INPUT AND OUTPUT:

Registration:

Users must register (http://robetta.bakerlab.org/register.jsp) before submitting jobs to Robetta.

Structure prediction server:

Sequences submitted to the structure prediction server must be in one-letter amino acid format. They can either be pasted into the submission form, or uploaded from a file. Users have the option to submit a sequence for either domain identification or full structure prediction. A user also has the option to specify the PDB id and chain for comparative modeling. For RosettaNMR submissions, a user must upload experimental NMR constraints data (chemical shifts, NOE data and/ or residual dipolar couplings). The required input format for each type of data is described at http://robetta.bakerlab.org/ documents/data_formats.jsp.

Results for a specific job are provided through the web interface by clicking on the job id listed in the queue table (http://robetta.bakerlab.org/queue.jsp). For full structure predictions, coordinates are also emailed to the user. For added insight, the following results are displayed along with the predicted models:

1. The prediction of transmembrane helices using TMHMM.
2. Low-complexity regions assigned by the program SEG
3. Coiled-coils prediction using COILS
4. The prediction of disordered regions using DISOPRED
5. Secondary structure predictions using PSIPRED, SAM-T99, Jufo and Jufo3D
6. The results listed above, domain predictions and the NR PSI-BLAST multiple sequence alignment used
7. For the last step in the domain prediction protocol condensed into an image to help corroborate the domain prediction results
8. Domain repeats prediction using REPRO predicted boundaries are given if repeats are detected
9. The top NR PSI-BLAST results and annotations for the top 20 species determined by lowest E-values.

The models for the full query are displayed as images at the bottom of the page. The coordinates for these models can be downloaded from the web site by clicking on the icons represented below each model image. Specific results are also provided for each domain by clicking on the domain number listed in the Ginzu domain prediction results table. For comparative models, the KSync alignment used for modelling is displayed. For de novo models, the Mammoth structure-model comparison results are displayed for the top 10 matches with Z-scores >4.5. The actual Mammoth structure-model alignment can be downloaded by clicking on the Z-score and viewed for further inspection using a molecular viewer such as RasMol. Users can download domain models by clicking on the icons below each domain model image.

Thus, modeller, I-TASSER and Robetta can be used to predict tertiary structures of proteins. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.

2. Tradigo, Giuseppe (2018). Reference Module in Life Sciences || Algorithms for Structure Comparison and Analysis: Prediction of Tertiary Structures of Proteins. , (), –. doi:10.1016/B978-0-12-809633- 8.20483-4
3. Bateman, Alex; Pearson, William R.; Stein, Lincoln D.; Stormo, Gary D.; Yates, John R. (2002). Current Protocols in Bioinformatics || Comparative Protein Structure Modeling Using MODELLER. , (), 5.6.1–5.6.37. doi:10.1002/cpbi.3
4. Tutorial. (n.d.). Salilab.org. Retrieved March 8, 2022, from https://salilab.org/modeller/tutorial/basic.html
5. I-TASSER server for protein structure and function prediction. (n.d.). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org/I-TASSER/about.html
6. Kim, D. E.; Chivian, D.; Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. , 32(0), 0–0. doi:10.1093/nar/gkh468

DATE: 20-03-22

WEBLEM 3a
MODELLER
(URL:https://salilab.org/modeller/)

AIM:

To perform tertiary structure prediction by comparative Modelling/Homology Modelling method using Modeller for query Rhodopsin

INTRODUCTION:

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye—specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.

MODELLER is a computer program for **comparative protein structure modelling**. In the simplest case, the input is an **alignment of a sequence** to be modelled with the **template structures**, the atomic coordinates of the templates, and a simple script file. **MODELLER** then automatically **calculates a model** containing all **non-hydrogen atoms**, within minutes on a modern PC and with no user intervention. Apart from model building, **MODELLER** can perform additional auxiliary tasks, including **fold assignment**, **alignment of two protein sequences** or their profiles, **multiple alignment** of protein sequences and/or structures, **calculation of phylogenetic trees**, and **de novo modelling of loops** in protein structures.

METHODOLOGY:

1. Install modeller. (URL: https://salilab.prg/modeller)
2. Retrieve FASTA sequence for enzyme rhodopsin
3. Follow the steps given in the tutorial section.
4. Run scripts for searching for structures related to query, selecting template target-template alignment and model building/
5. Observe and interpret the results.

OBSERVATION:

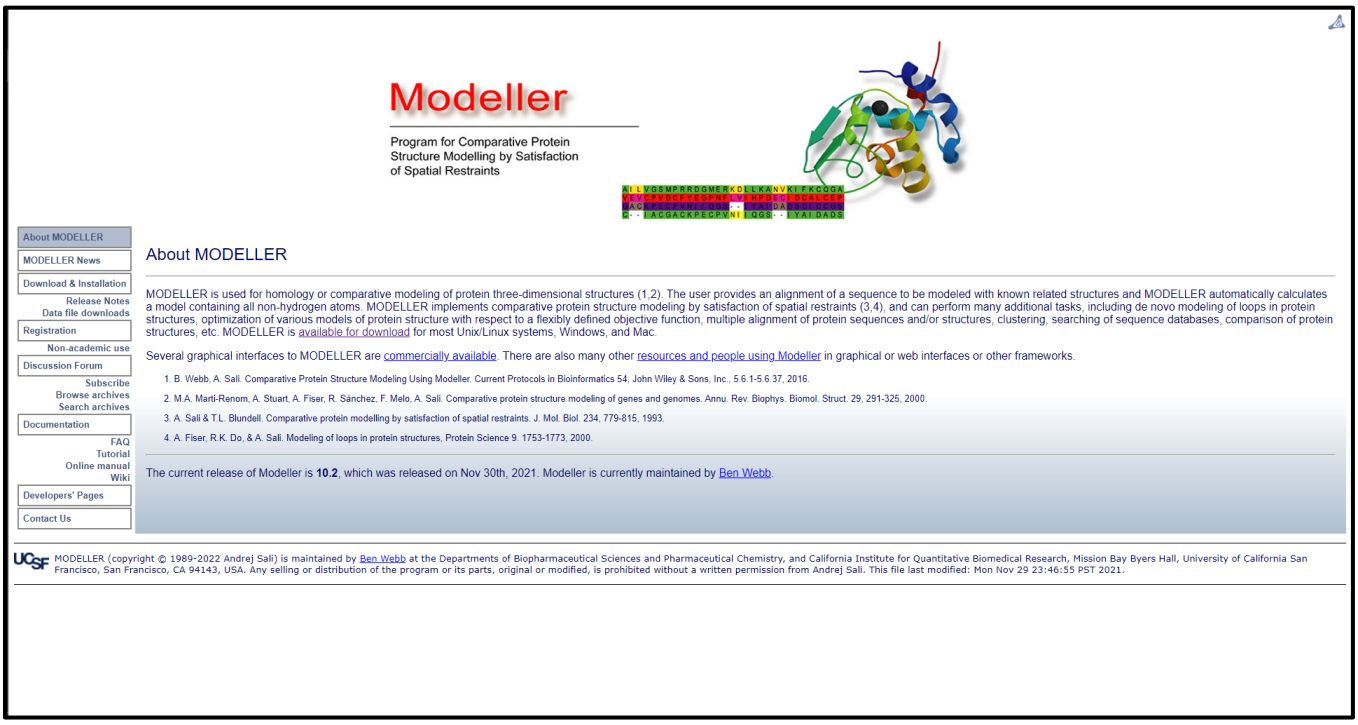


Fig1. Homepage for Modeller

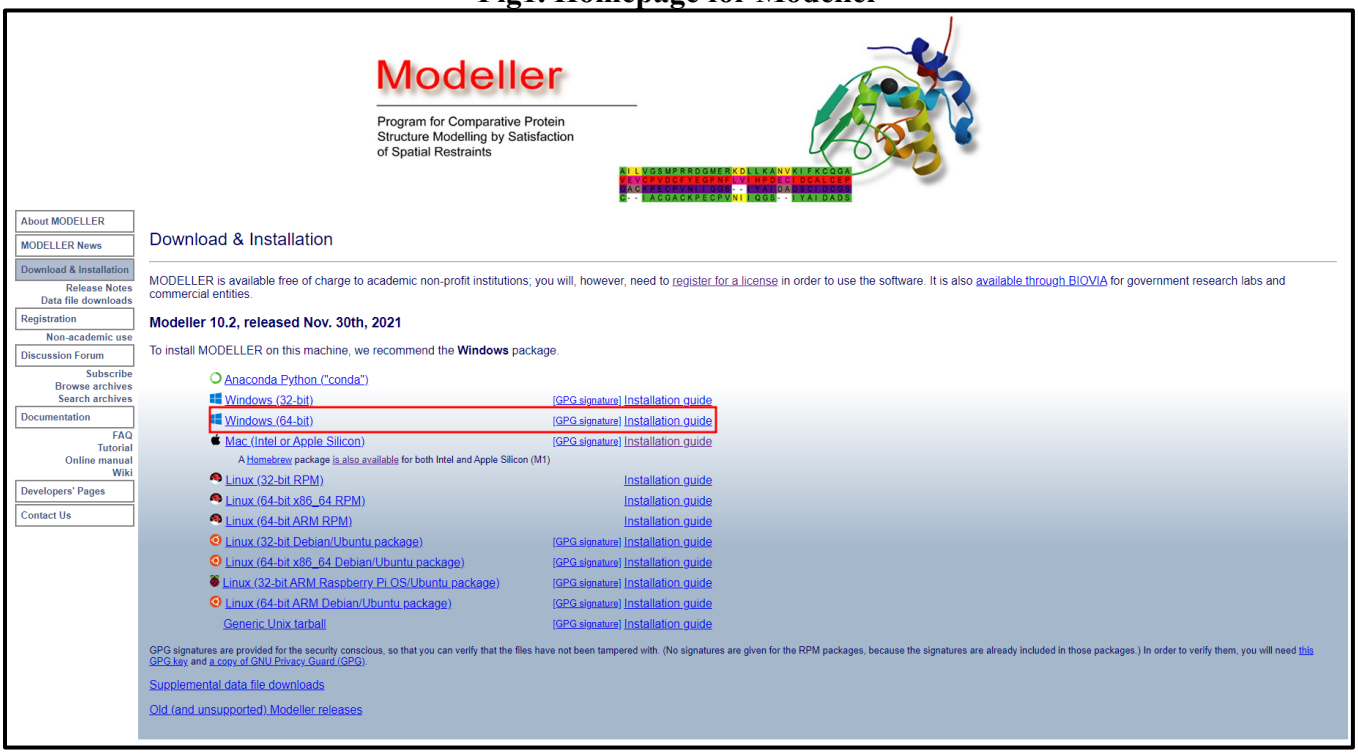


Fig2. Page to install Modeller

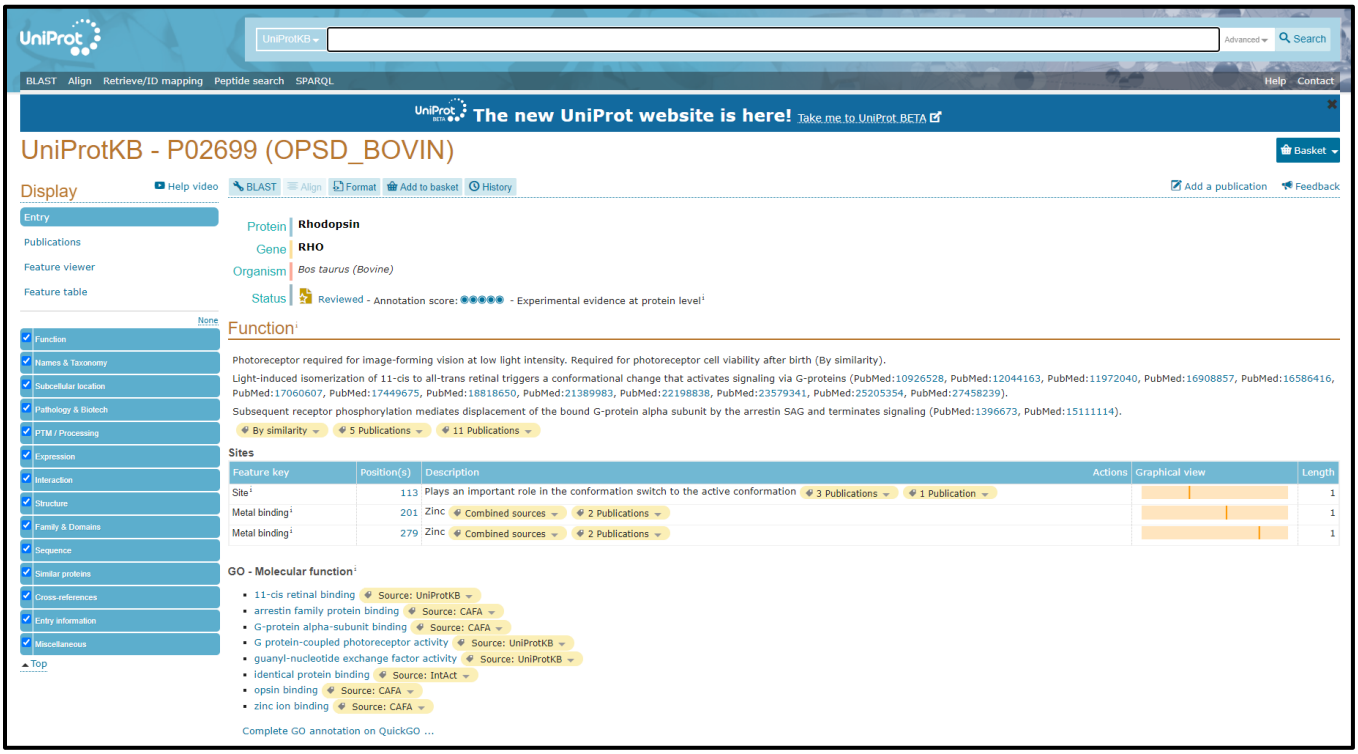


Fig3. Result page for Rhodopsin in UniProt database

```
>sp|Q15835|GRK1_HUMAN Rhodopsin kinase GRK1 OS=Homo sapiens OX=9606 GN=GRK1 PE=1
SV=1
MDFGSLETVVANSAFIAARGSFDGSSSQPSRDKKYLAKLKLPPLSKCESLRDSLSLEFES
VCLEQPIGKKLFQQFLQSAEKHLPALELWKDIEDYDTADNDLQPQKAQTILAQYLDPQAK
LFCSFLDEGIVAKFKEGPVEIQDGLFQPLLQATLAHLGQAPFQEYLGSLYFLRFLQWKWL
EAQPMGEDWFLDFRVLGKGGFGEVSACQMKATGKLYACKKLNKKRLKKRKGYQGAMVEKK
ILMKVHSRFIVSLAYAFETKADLCLVMTIMNGGDIRYHIYNVNEENPGFPEPRALFYTAQ
IICGLEHLHQRRIVYRDLKPENVLLDNDGNVRISDLGLAVELLDGQSKTKGYAGTPGFMA
PELLQGEEYDFSVDYFALGVTLYEMIAARGPFRARGEKVENKELKHRIISEPVKYPDKFS
QASKDFCEALLEKDPEKRLGFRDETCDKLRAHPLFKDLNWRQLEAGMLMPPFIPDSKTVY
AKDIQDVGAFSTVKGVAFDKTDTEFFQEFATGNCPIPWQEEMIETGIFGELNVWRSDGQM
PDDMKGISGGSSSSSKSGMCLVS
```

Fig4. FASTA sequence for Rhodopsin

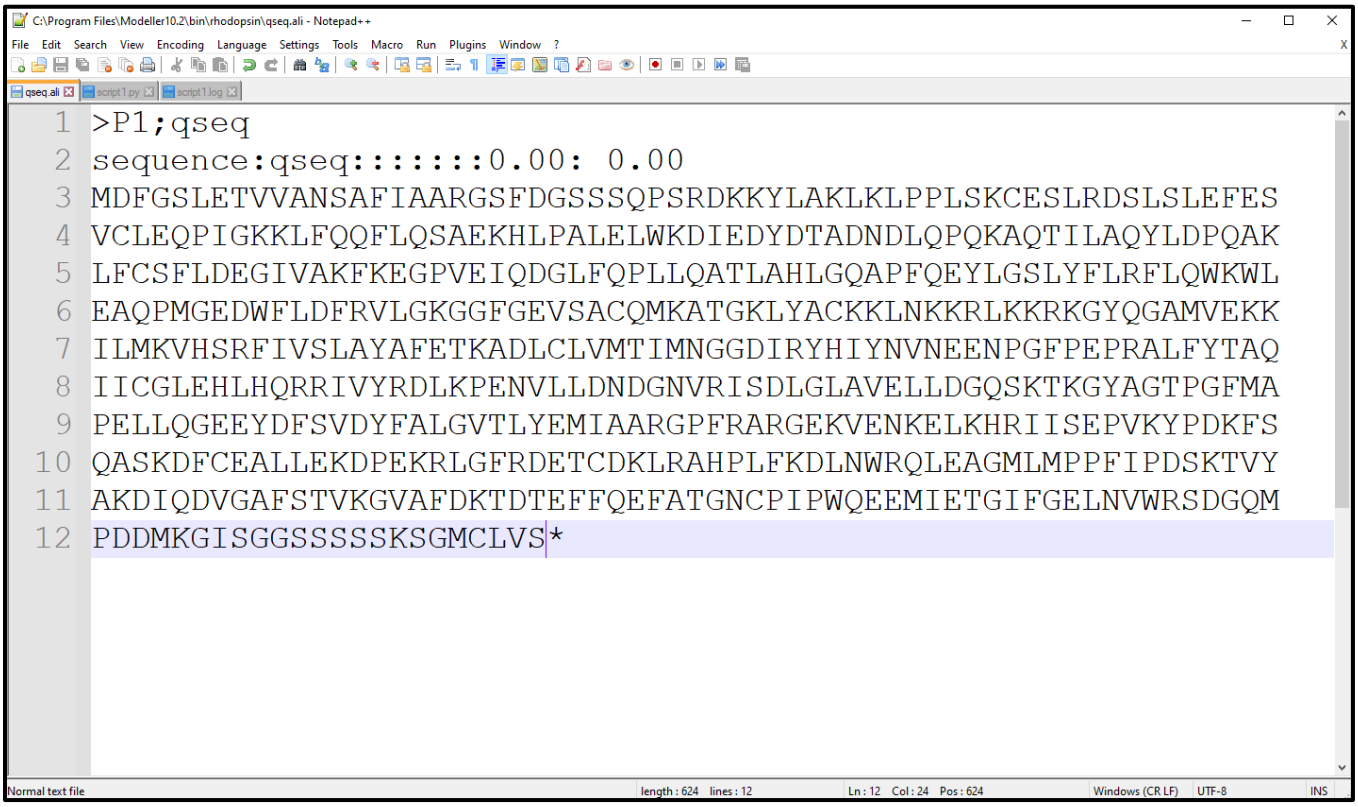


Fig5. FASTA sequence in PIR format

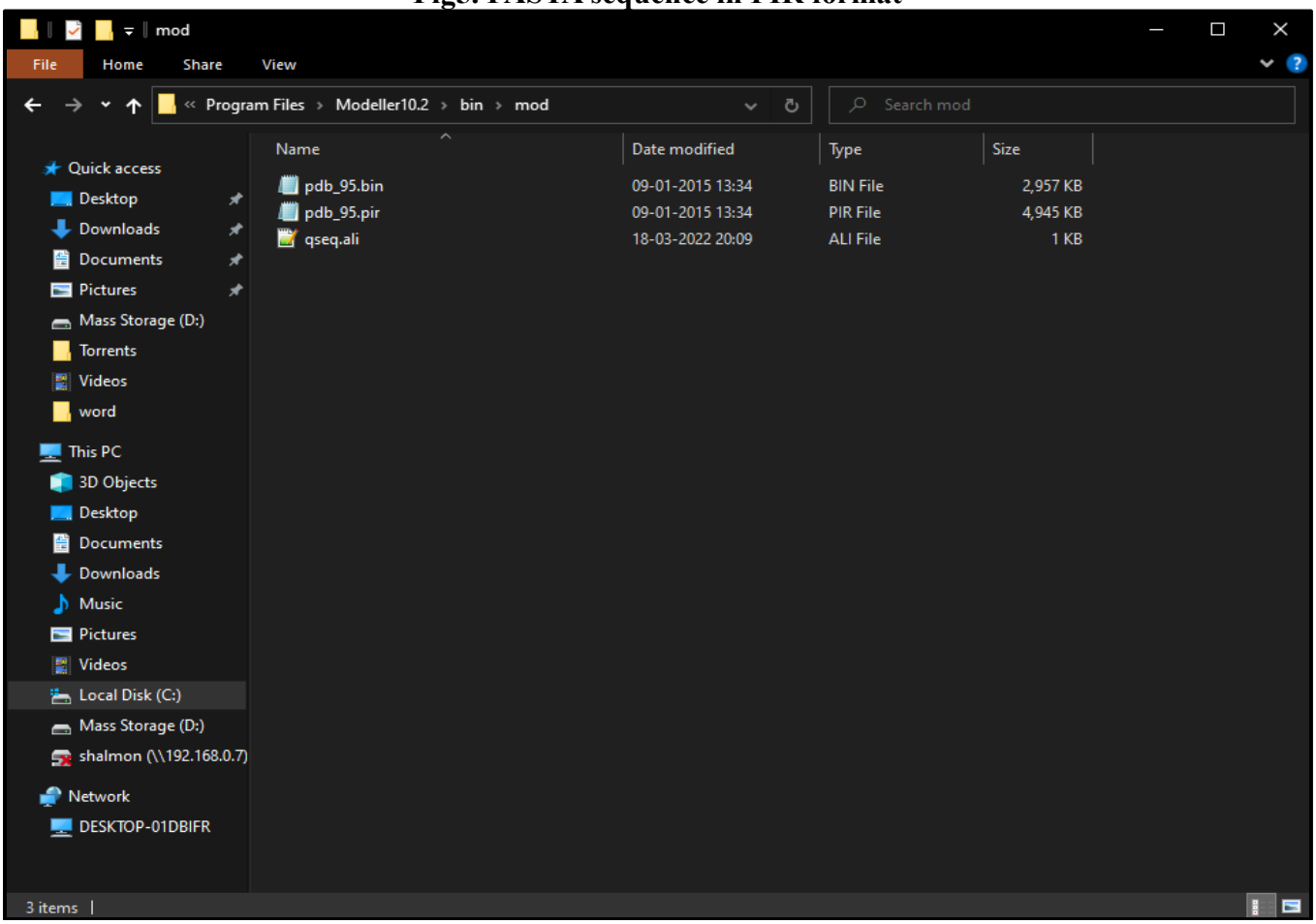


Fig6. Target sequence saved in .ali format

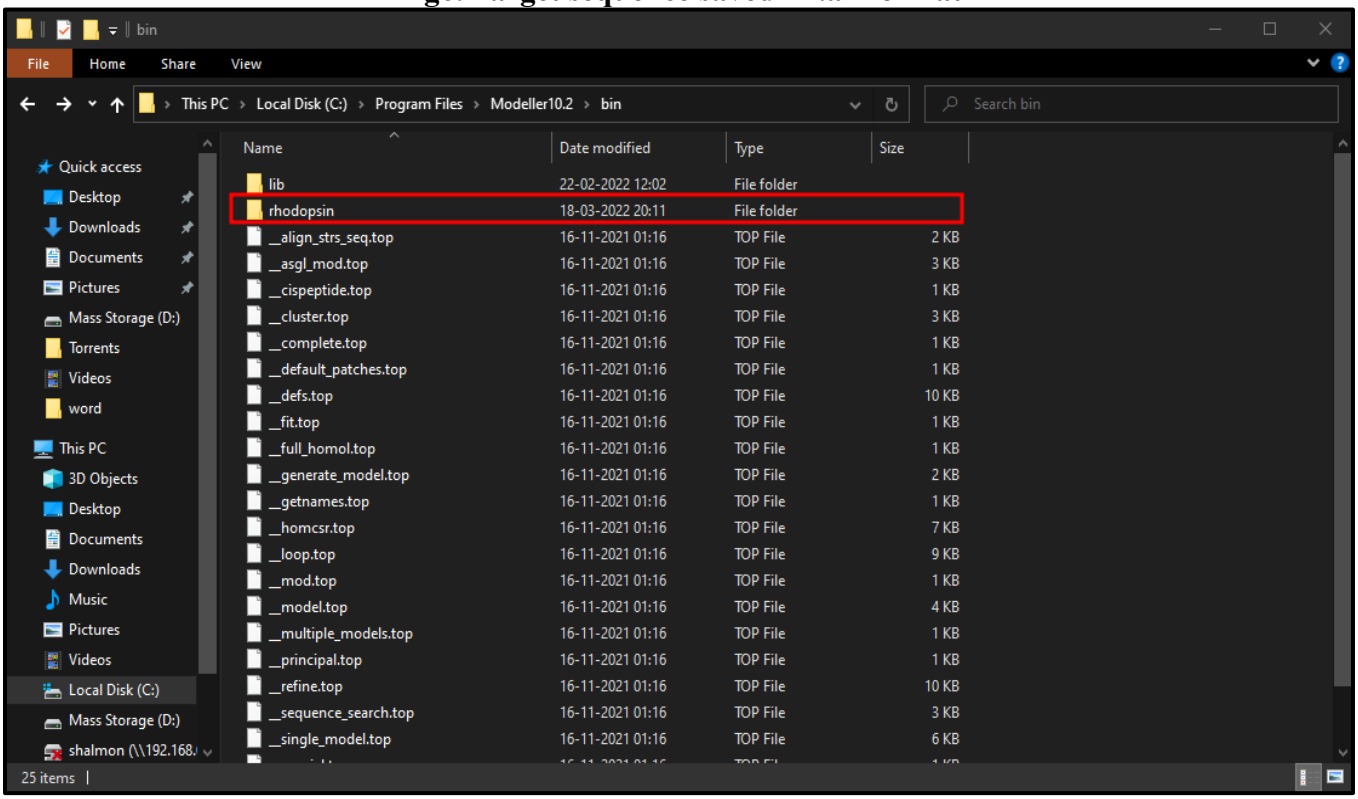


Fig7. Rhodopsin folder saved in the bin folder of modeller

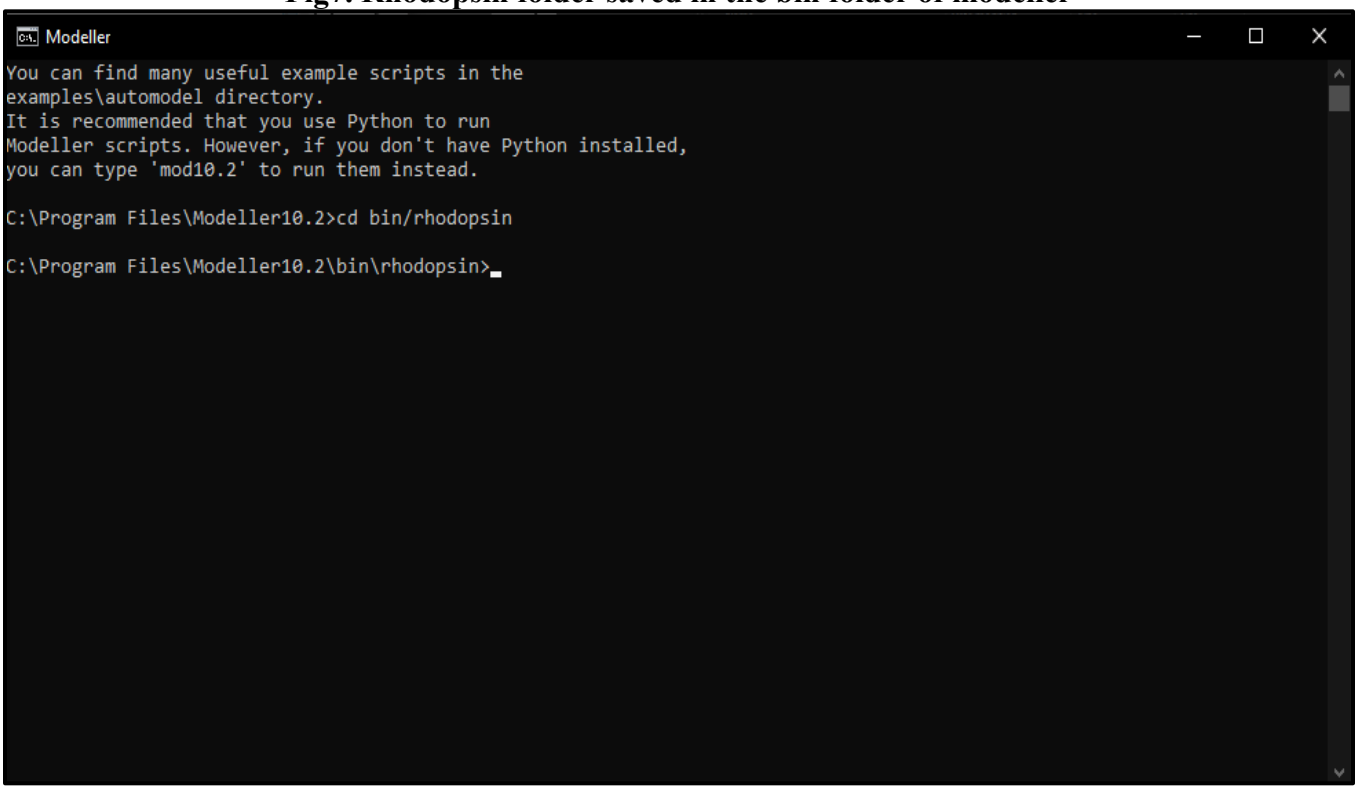


Fig8. Setting Working directory in Modeller command line

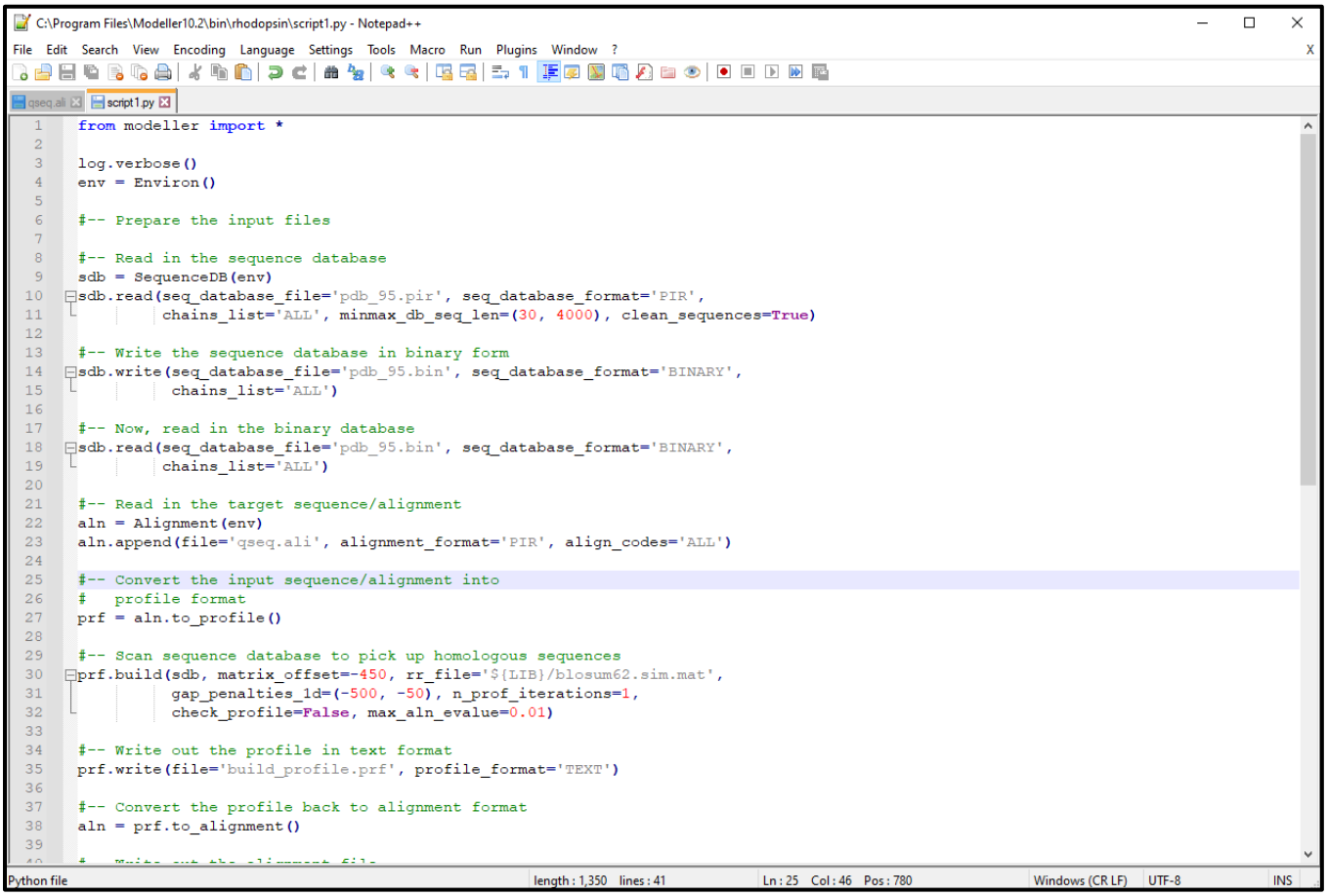


Fig9. Python script for searching for structures relation to rhodopsin

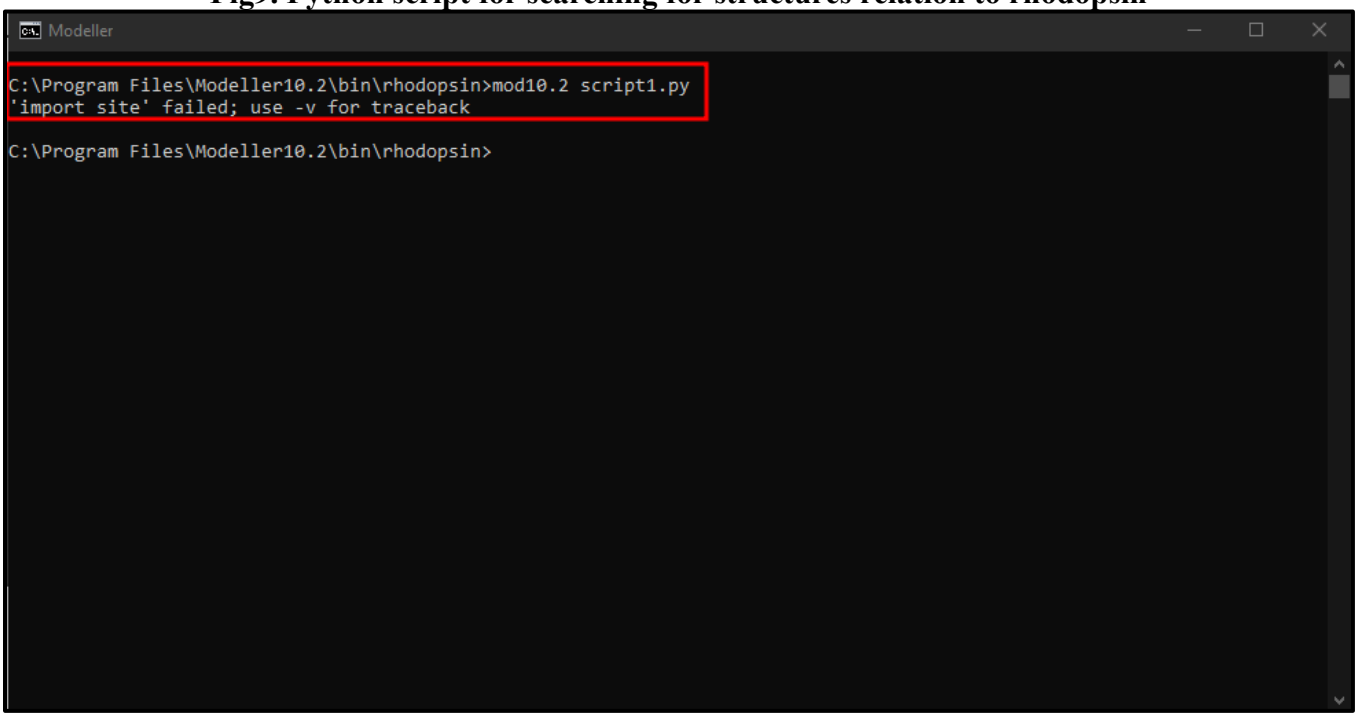
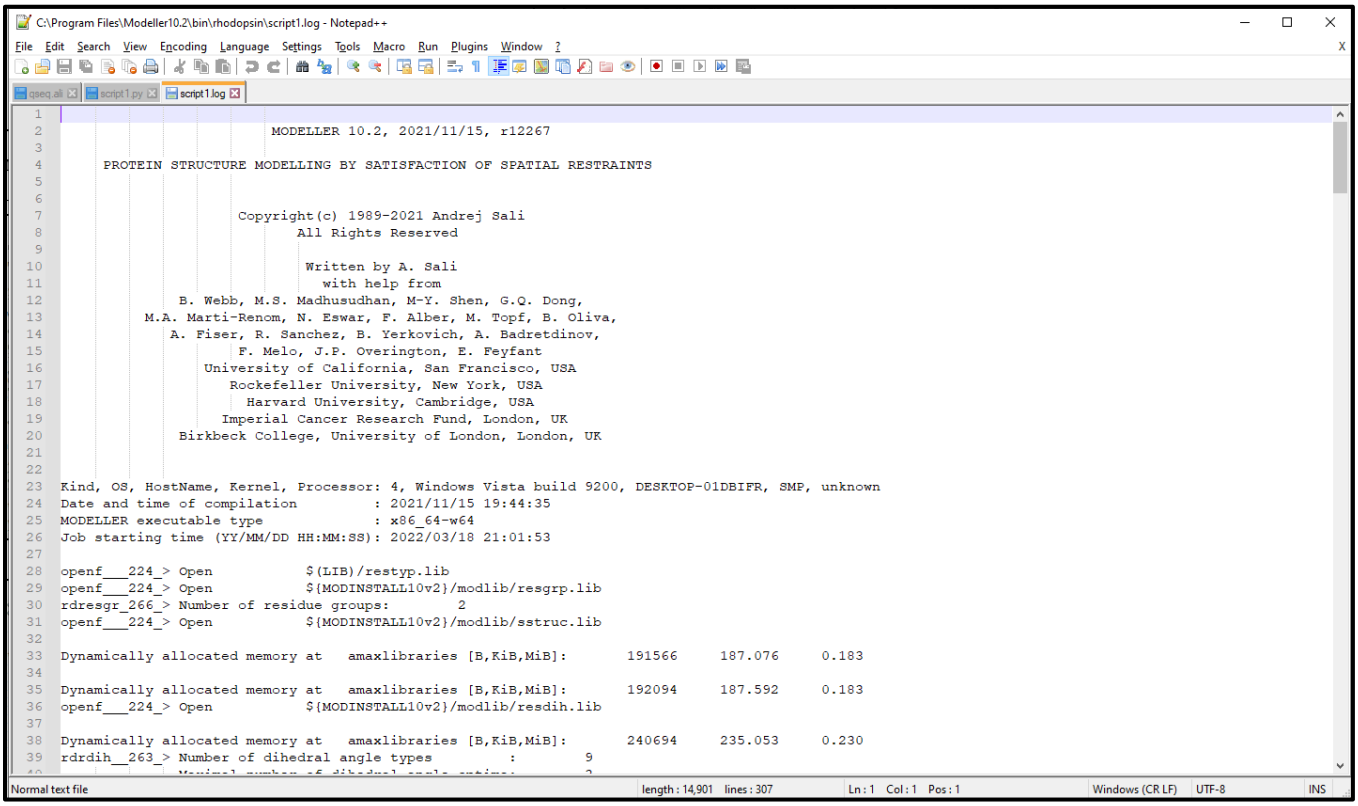
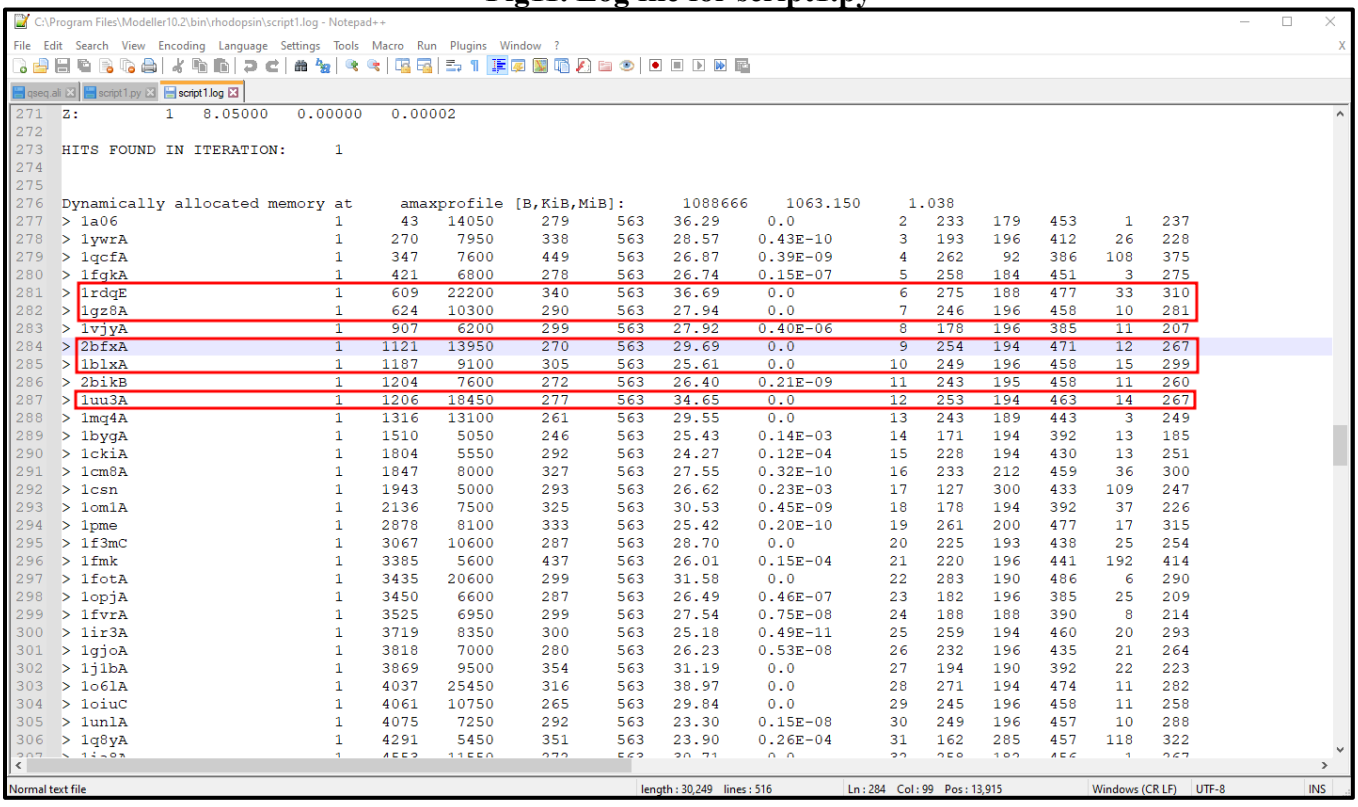


Fig10. Running script1.py



```
C:\Program Files\Modeller10.2\bin\rhodopsin\script1.log - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
qseq.ali | script1.py | script1.log

                    MODELLER 10.2, 2021/11/15, r12267

     PROTEIN STRUCTURE MODELLING BY SATISFACTION OF SPATIAL RESTRAINTS


                  Copyright(c) 1989-2021 Andrej Sali
                        All Rights Reserved

                        Written by A. Sali
                          with help from
            B. Webb, M.S. Madhusudhan, M-Y. Shen, G.Q. Dong,
          M.A. Marti-Renom, N. Eswar, F. Alber, M. Topf, B. Oliva,
             A. Fiser, R. Sanchez, B. Yerkovich, A. Badretdinov,
                    F. Melo, J.P. Overington, E. Feyfant
                University of California, San Francisco, USA
                   Rockefeller University, New York, USA
                     Harvard University, Cambridge, USA
                  Imperial Cancer Research Fund, London, UK
              Birkbeck College, University of London, London, UK


Kind, OS, HostName, Kernel, Processor: 4, Windows Vista build 9200, DESKTOP-01DBIFR, SMP, unknown
Date and time of compilation         : 2021/11/15 19:44:35
MODELLER executable type             : x86_64-w64
Job starting time (YY/MM/DD HH:MM:SS): 2022/03/18 21:01:53

openf___224_> Open           $(LIB)/restyp.lib
openf___224_> Open           ${MODINSTALL10v2}/modlib/resgrp.lib
rdresgr_266_> Number of residue groups:        2
openf___224_> Open           ${MODINSTALL10v2}/modlib/sstruc.lib

Dynamically allocated memory at   amaxlibraries [B,KiB,MiB]:       191566     187.076     0.183

Dynamically allocated memory at   amaxlibraries [B,KiB,MiB]:       192094     187.592     0.183
openf___224_> Open           ${MODINSTALL10v2}/modlib/resdih.lib

Dynamically allocated memory at   amaxlibraries [B,KiB,MiB]:       240694     235.053     0.230
rdrdih__263_> Number of dihedral angle types         :        9

Normal text file   length : 14,901   lines : 307   Ln : 1   Col : 1   Pos : 1   Windows (CR LF)   UTF-8   INS
```

Fig11. Log file for script1.py



```
C:\Program Files\Modeller10.2\bin\rhodopsin\script1.log - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
qseq.ali | script1.py | script1.log

Z:          1   8.05000   0.00000   0.00002

HITS FOUND IN ITERATION:      1


Dynamically allocated memory at     amaxprofile [B,KiB,MiB]:      1088666    1063.150     1.038
> 1a06                      1      43    14050      279      563   36.29     0.0          2   233   179   453     1   237
> 1ywrA                     1     270     7950      338      563   28.57     0.43E-10     3   193   196   412    26   228
> 1qcfA                     1     347     7600      449      563   26.87     0.39E-09     4   262    92   386   108   375
> 1fgkA                     1     421     6800      278      563   26.74     0.15E-07     5   258   184   451     3   275
> 1rdqE                     1     609    22200      340      563   36.69     0.0          6   275   188   477    33   310
> 1gz8A                     1     624    10300      290      563   27.94     0.0          7   246   196   458    10   281
> 1vjyA                     1     907     6200      299      563   27.92     0.40E-06     8   178   196   385    11   207
> 2bfxA                     1    1121    13950      270      563   29.69     0.0          9   254   194   471    12   267
> 1blxA                     1    1187     9100      305      563   25.61     0.0         10   249   196   458    15   299
> 2bikB                     1    1204     7600      272      563   26.40     0.21E-09    11   243   195   458    11   260
> 1uu3A                     1    1206    18450      277      563   34.65     0.0         12   253   194   463    14   267
> 1mq4A                     1    1316    13100      261      563   29.55     0.0         13   243   189   443     3   249
> 1bygA                     1    1510     5050      246      563   25.43     0.14E-03    14   171   194   392    13   185
> 1ckiA                     1    1804     5550      292      563   24.27     0.12E-04    15   228   194   430    13   251
> 1cm8A                     1    1847     8000      327      563   27.55     0.32E-10    16   233   212   459    36   300
> 1csn                      1    1943     5000      293      563   26.62     0.23E-03    17   127   300   433   109   247
> 1om1A                     1    2136     7500      325      563   30.53     0.45E-09    18   178   194   392    37   226
> 1pme                      1    2878     8100      333      563   25.42     0.20E-10    19   261   200   477    17   315
> 1f3mC                     1    3067    10600      287      563   28.70     0.0         20   225   193   438    25   254
> 1fmk                      1    3385     5600      437      563   26.01     0.15E-04    21   220   196   441   192   414
> 1fotA                     1    3435    20600      299      563   31.58     0.0         22   283   190   486     6   290
> 1opjA                     1    3450     6600      287      563   26.49     0.46E-07    23   182   196   385    25   209
> 1fvrA                     1    3525     6950      299      563   27.54     0.75E-08    24   188   188   390     8   214
> 1ir3A                     1    3719     8350      300      563   25.18     0.49E-11    25   259   194   460    20   293
> 1gjoA                     1    3818     7000      280      563   26.23     0.53E-08    26   232   196   435    21   264
> 1j1bA                     1    3869     9500      354      563   31.19     0.0         27   194   190   392    22   223
> 1o61A                     1    4037    25450      316      563   38.97     0.0         28   271   194   474    11   282
> 1oiuC                     1    4061    10750      265      563   29.84     0.0         29   245   196   458    11   258
> 1un1A                     1    4075     7250      292      563   23.30     0.15E-08    30   249   196   457    10   288
> 1q8yA                     1    4291     5450      351      563   23.90     0.26E-04    31   162   285   457   118   322

Normal text file   length : 30,249   lines : 516   Ln : 284   Col : 99   Pos : 13,915   Windows (CR LF)   UTF-8   INS
```

Fig11.1. Hits found for similar structures

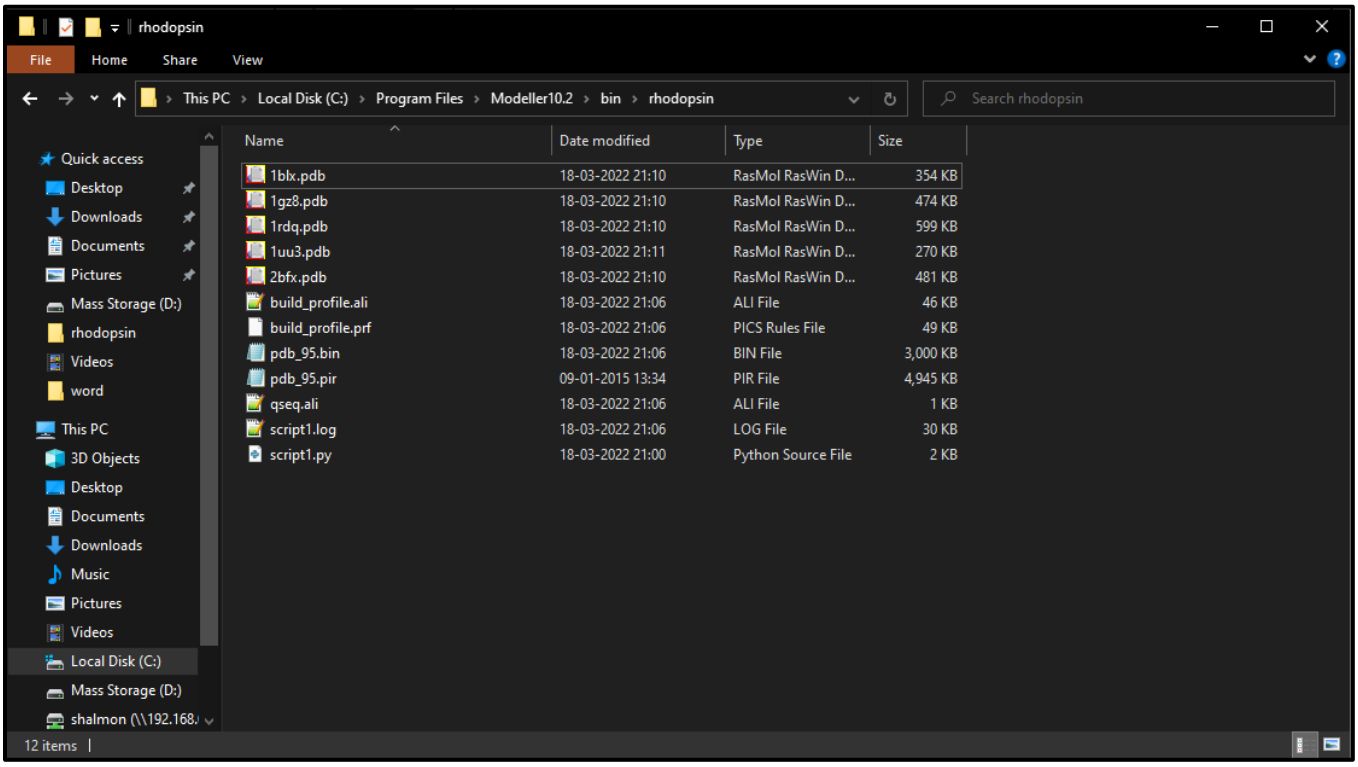


Fig12. Five structure download in PDB format

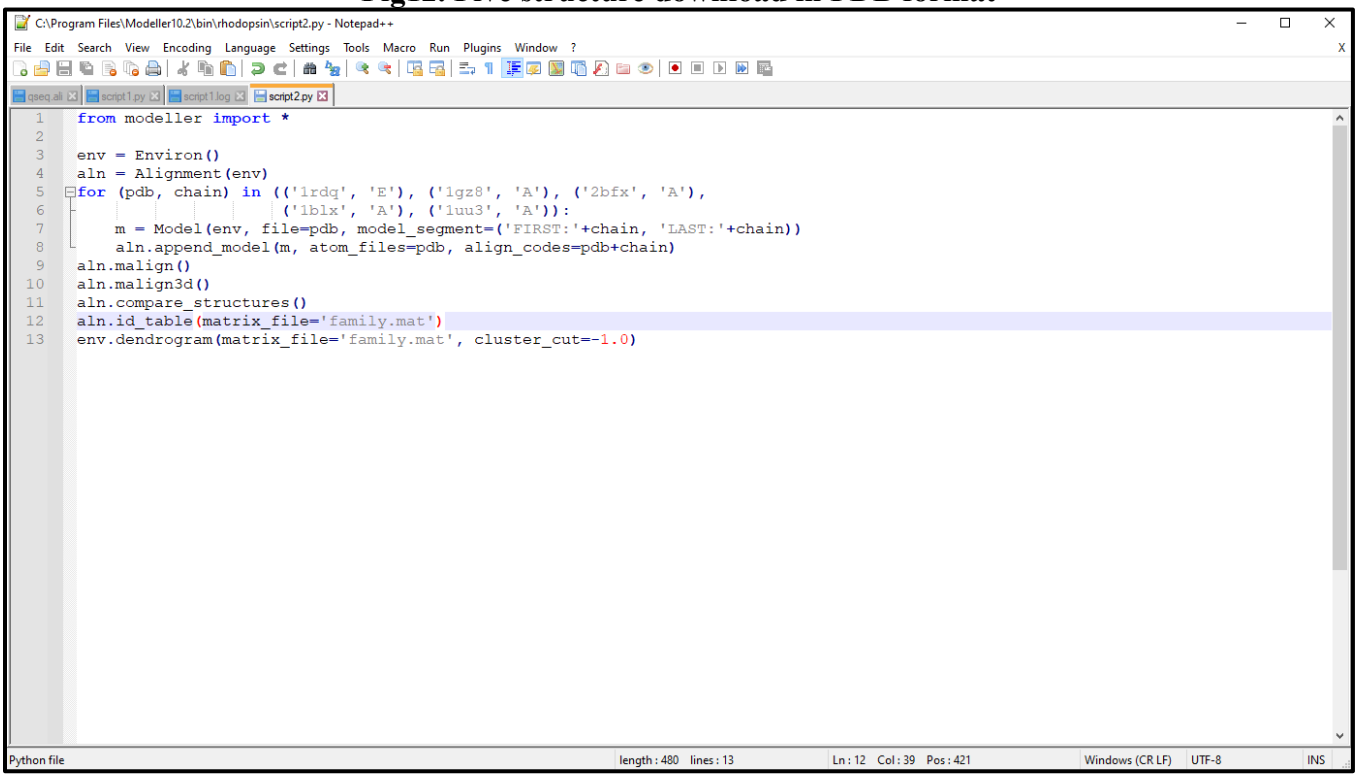


Fig13. Python script for selecting a template

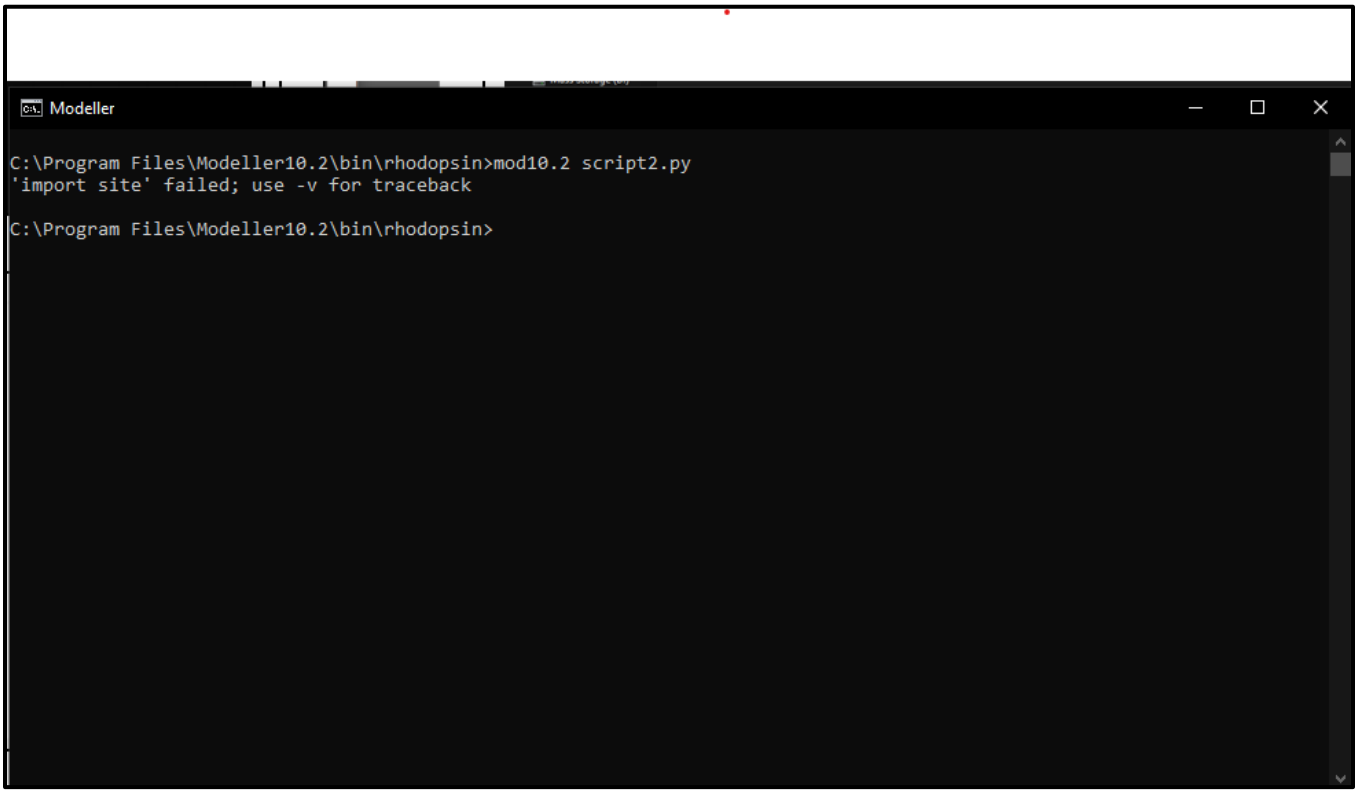


Fig14. Running script2.py

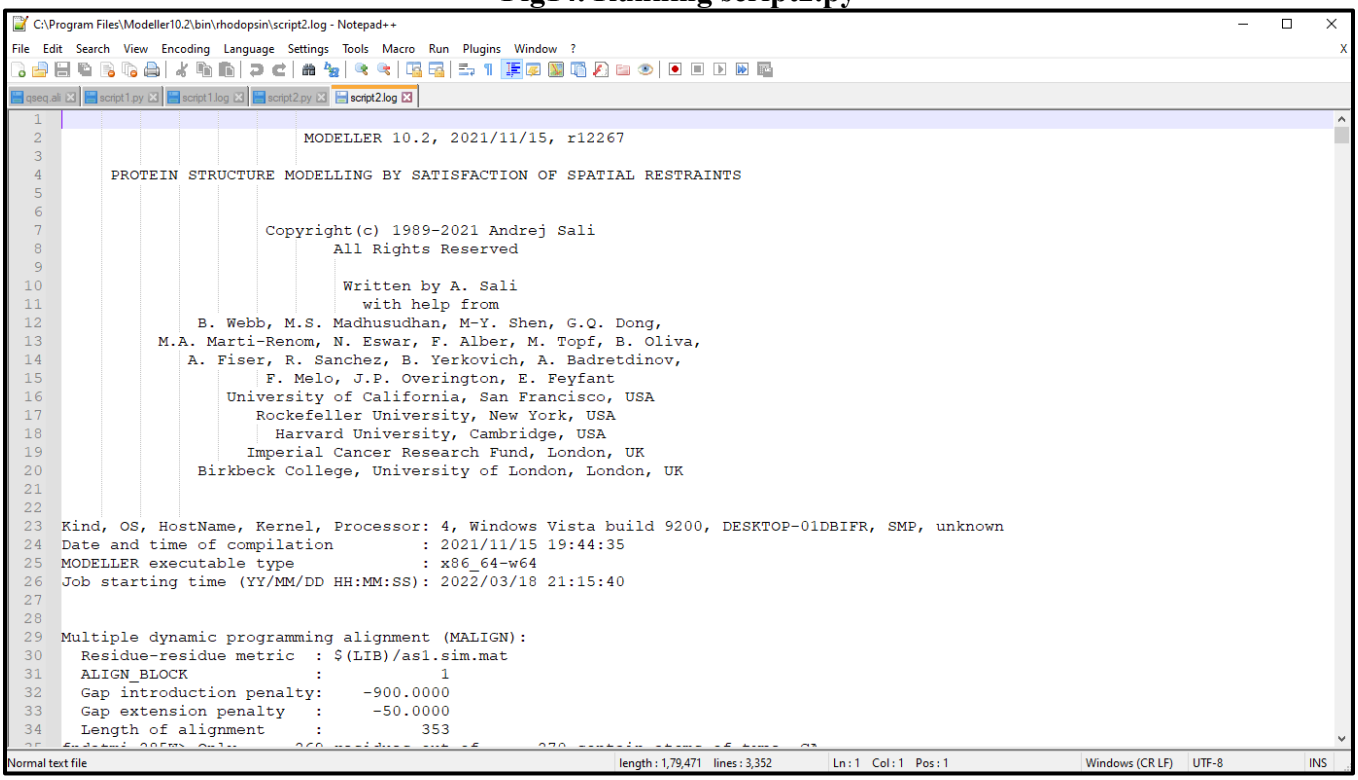


Fig15. Log file for script2

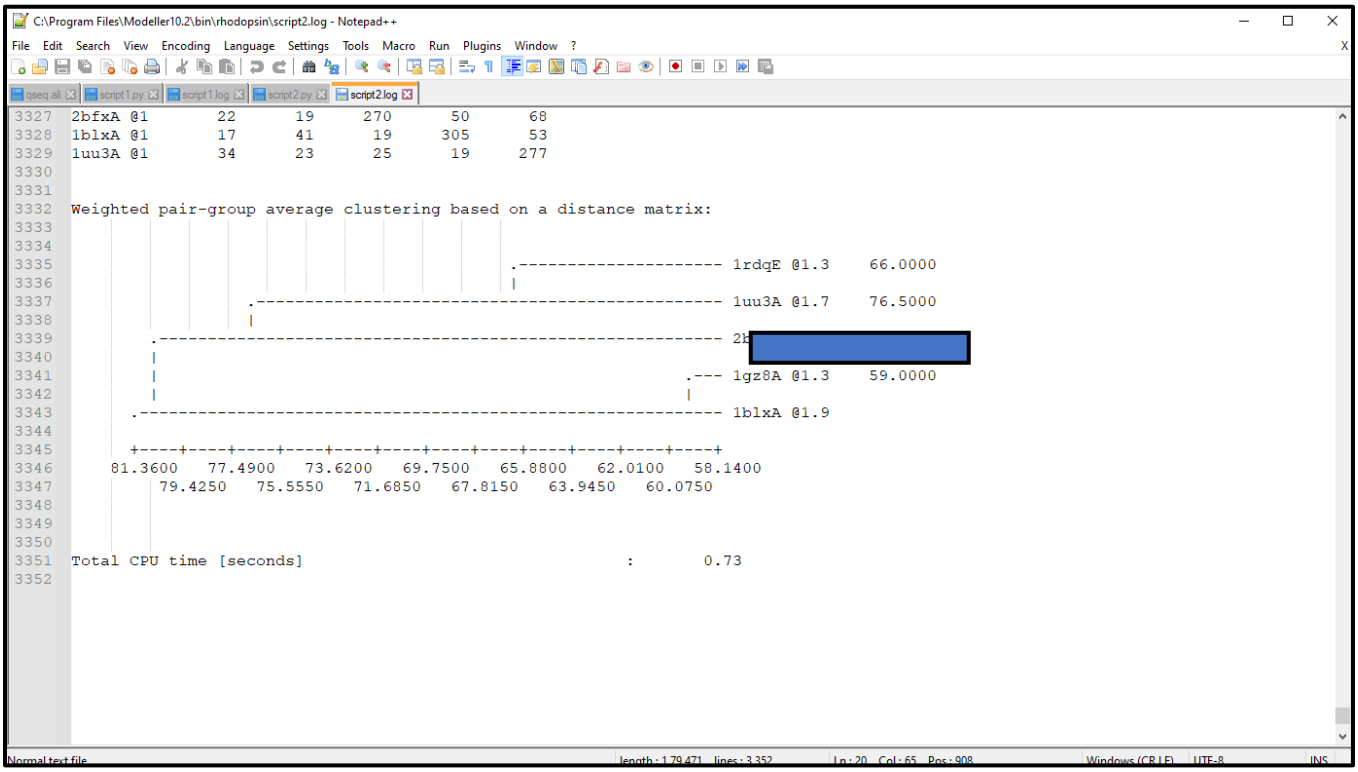


Fig15.1 Structure selected with low x=ray crystallography value and high NMR value

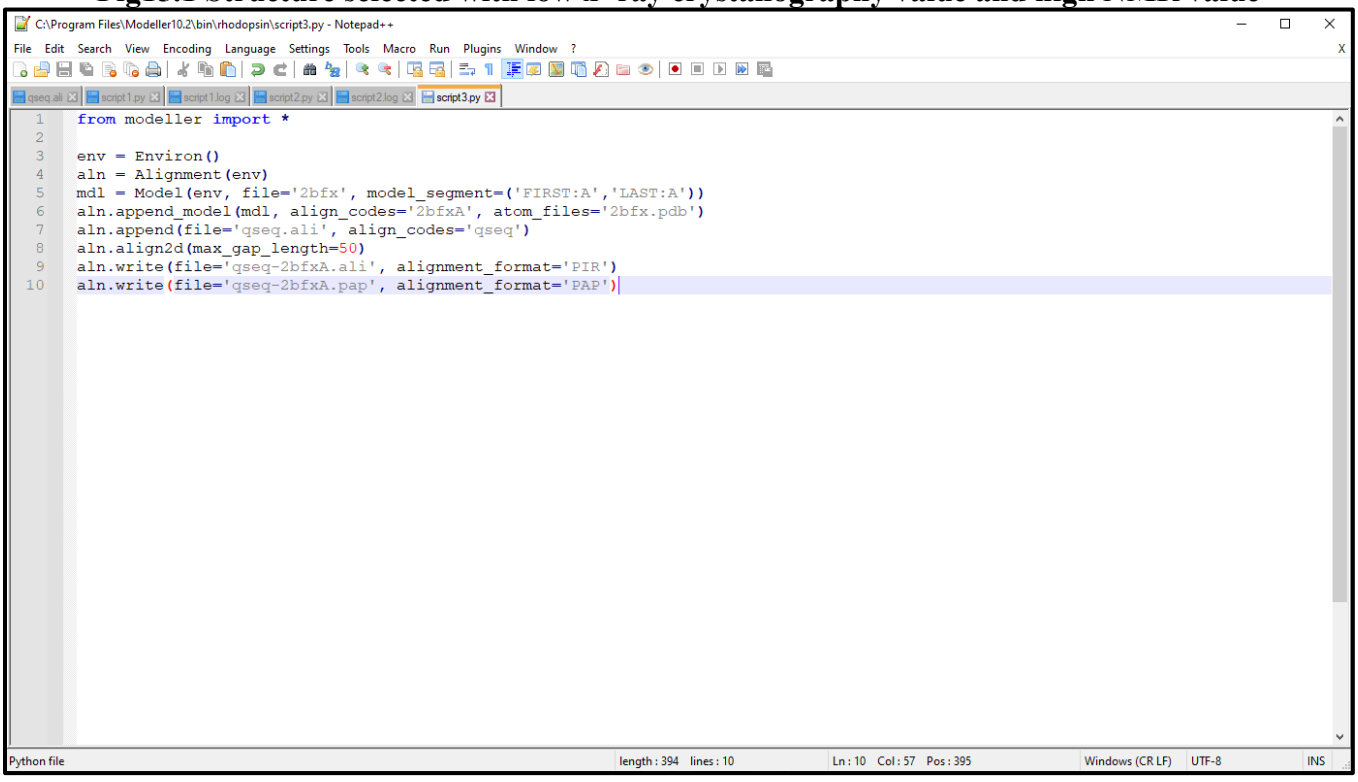


Fig16. Python script for aligning query with the template

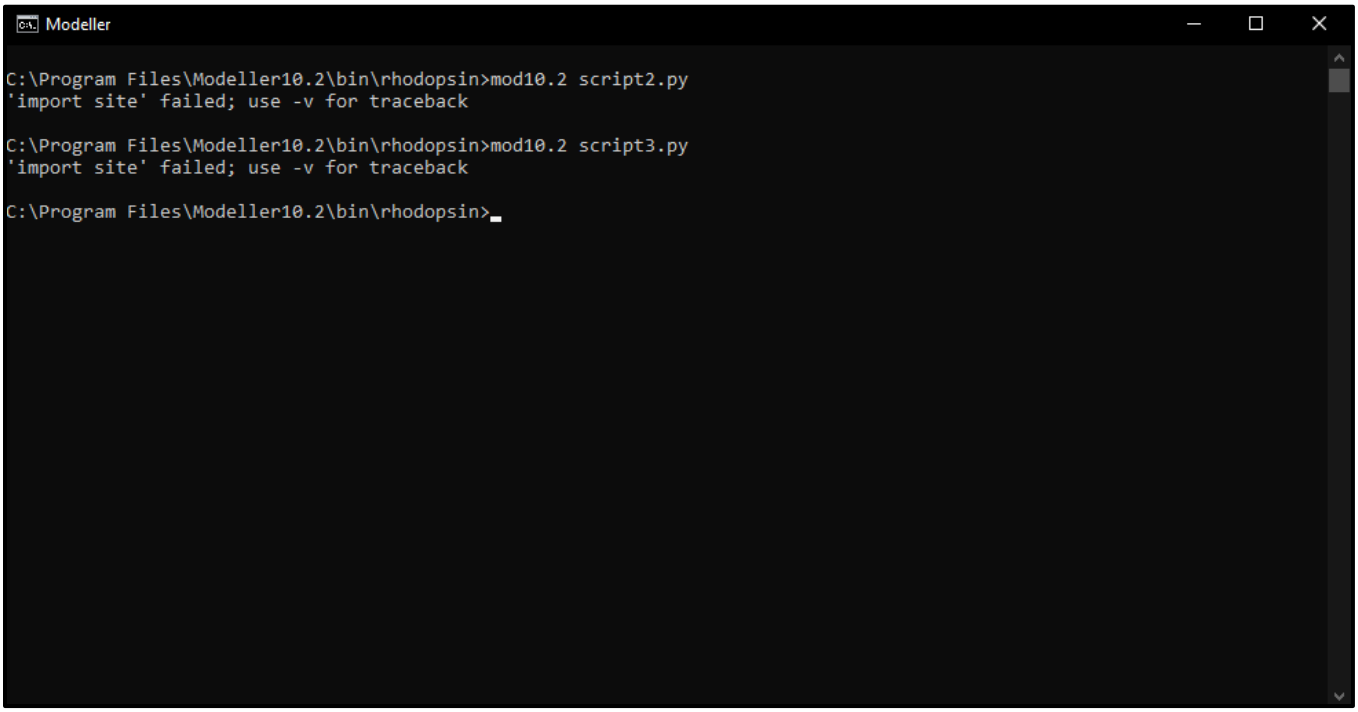


Fig17. Running script3.py

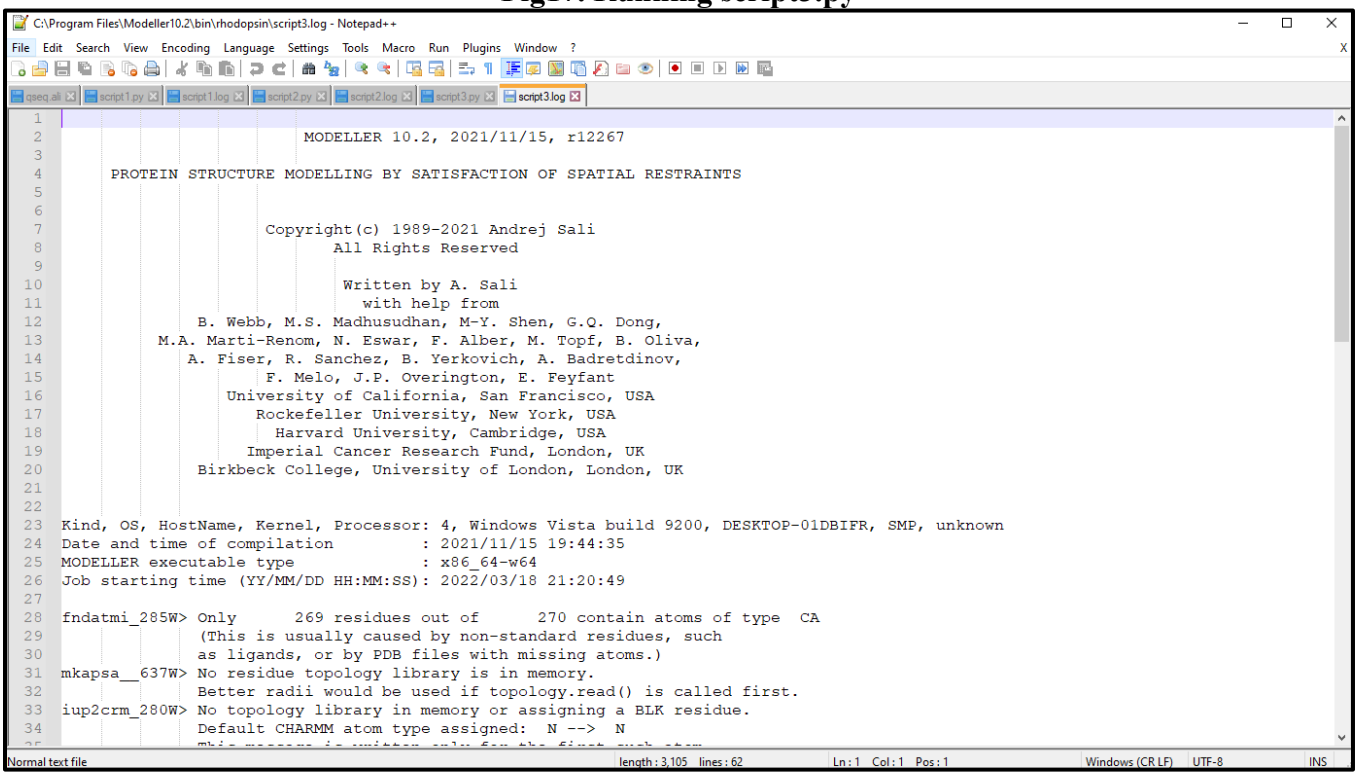


Fig18. Log file for script3

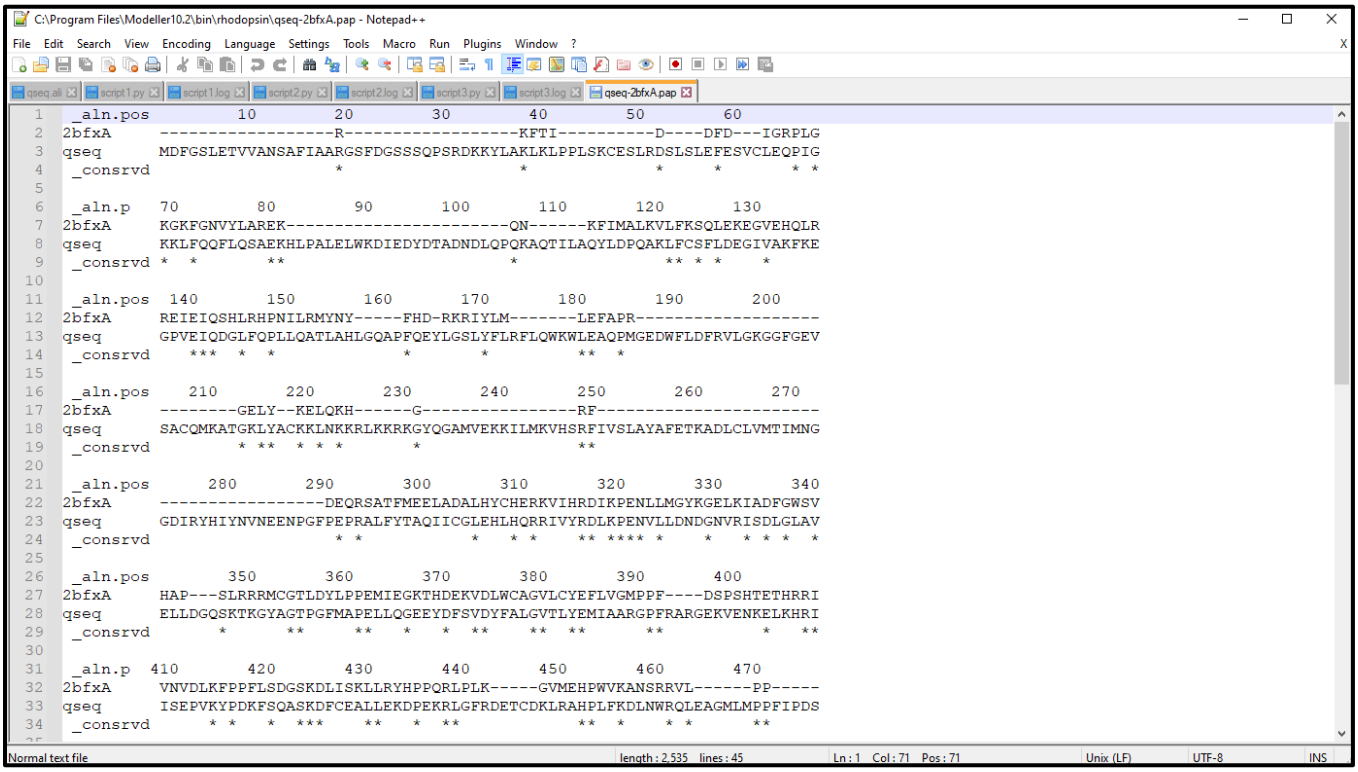


Fig19. Sequence alignment

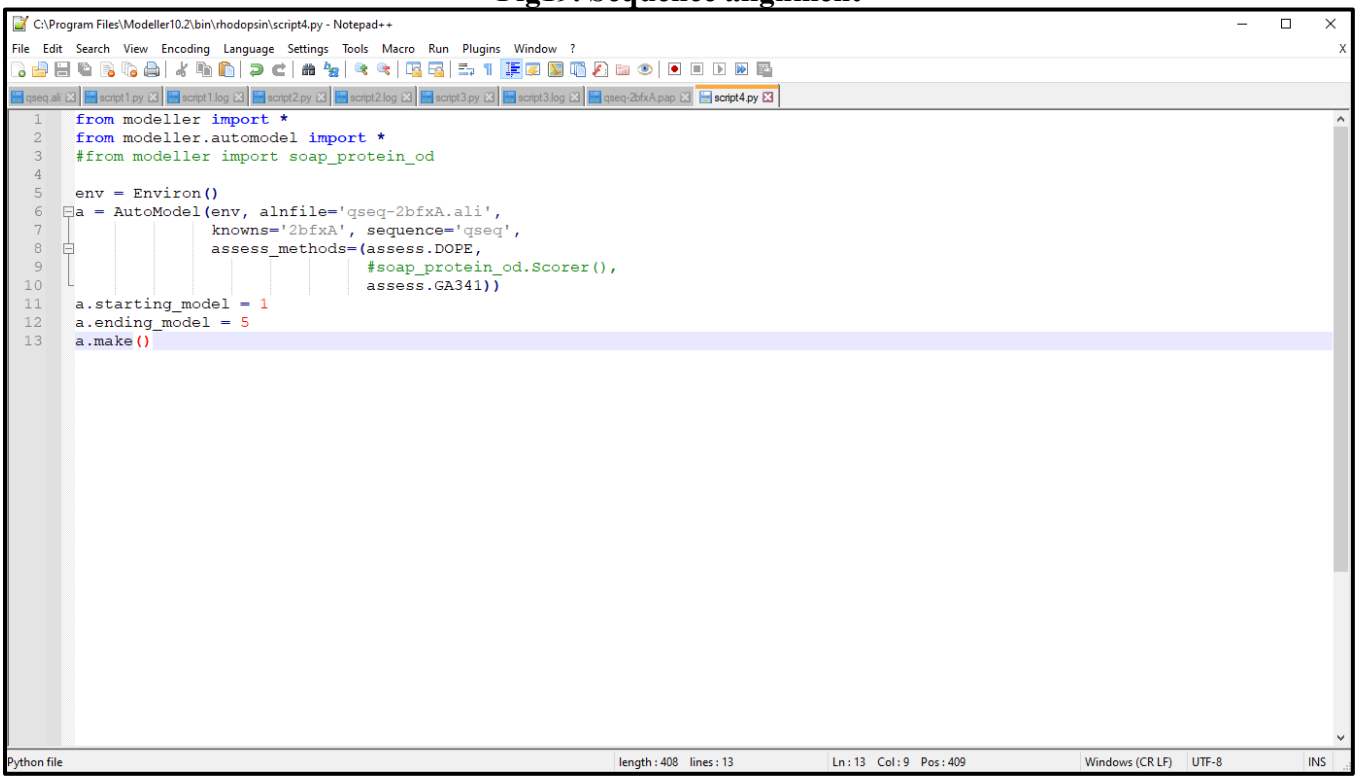


Fig20. Python script for model building

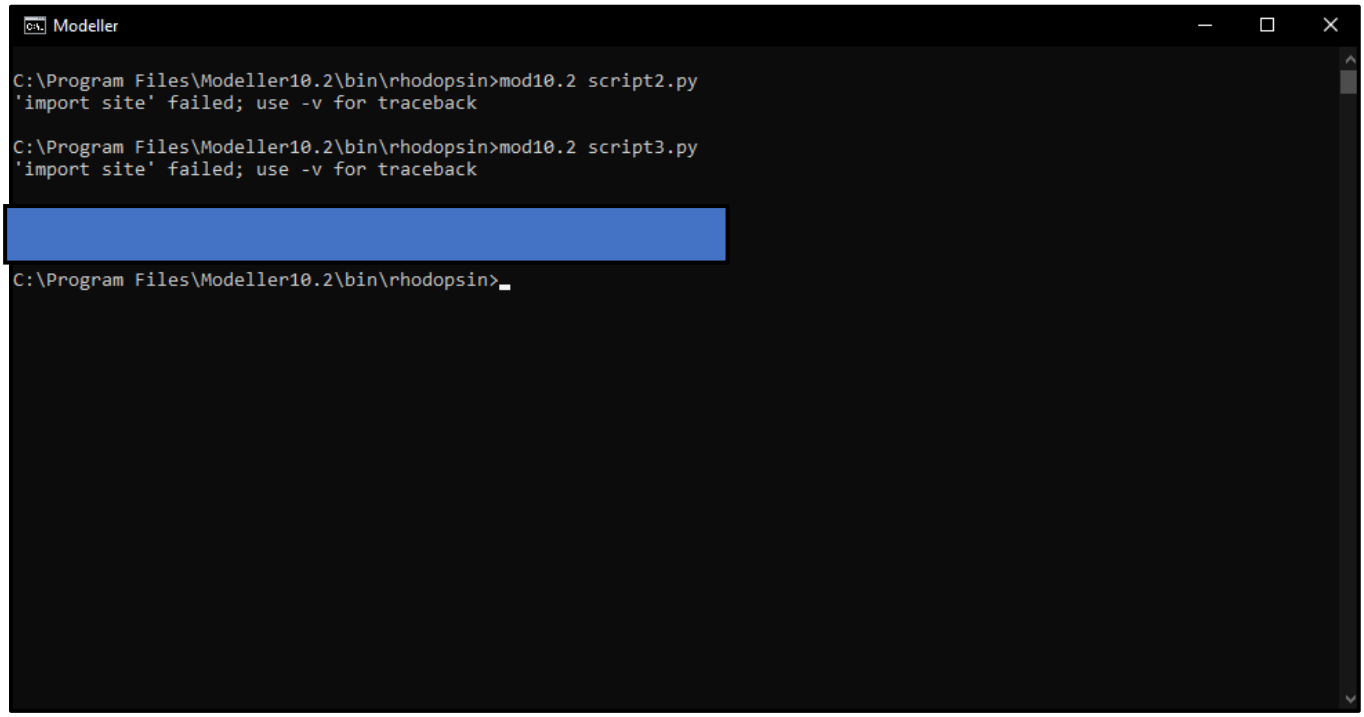


Fig21. Running script4.py

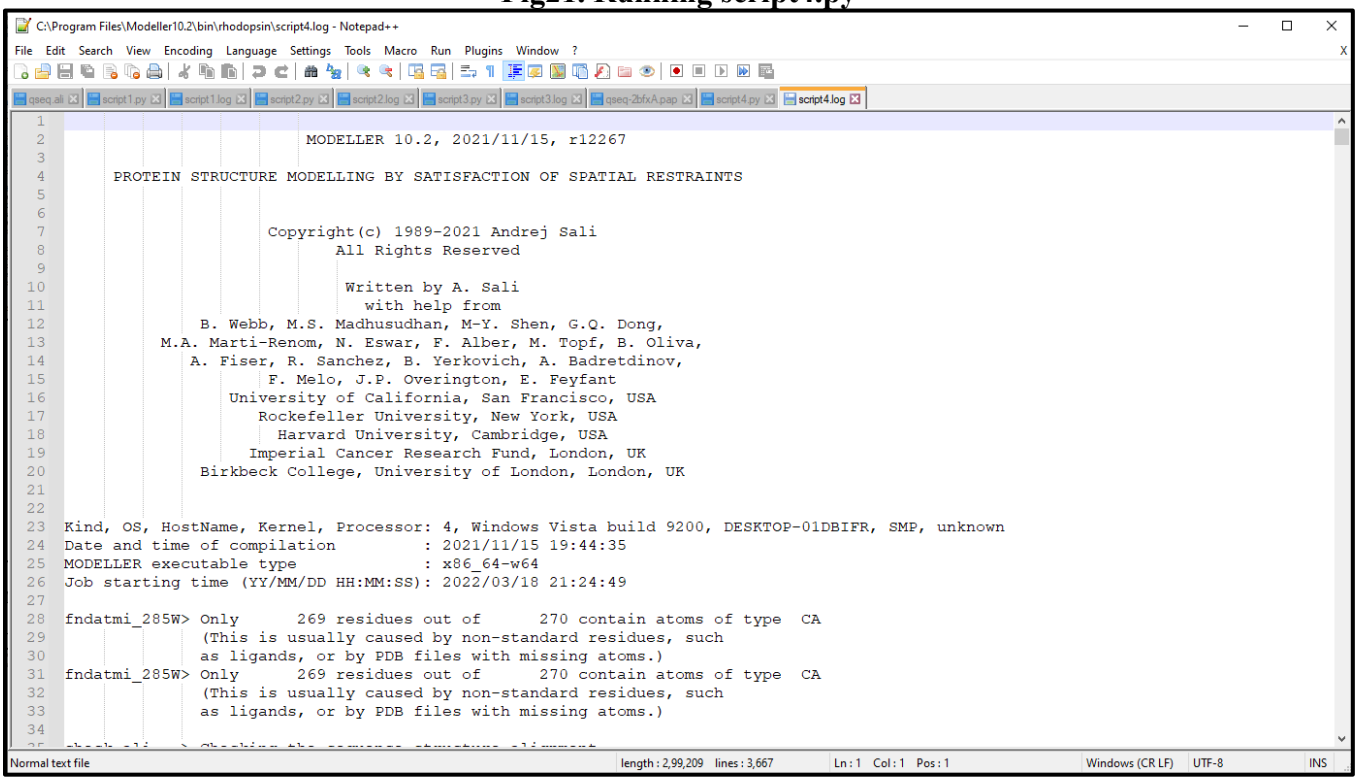


Fig22. Log file for script4

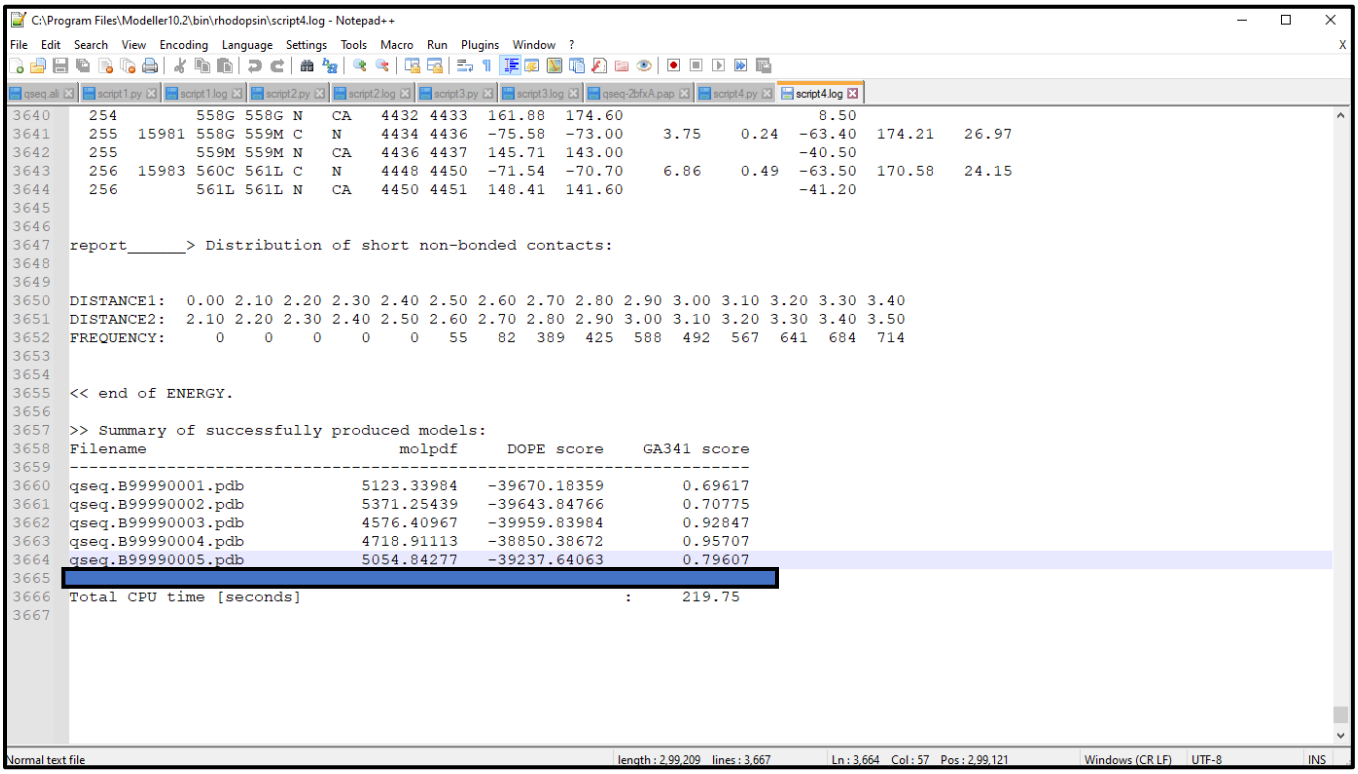


Fig23. Structure with lowest DOPE score selected as final model

RESULT:

Modeller was used to predict the tertiary structure of Rhodopsin

CONCLUSION:

Thus, modeller can be used to predict tertiary structures of proteins by comparative protein structure modelling. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions and drug designing.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.

2. Encyclopædia Britannica, inc. (n.d.). *Rhodopsin*. Encyclopædia Britannica. Retrieved March 18, 2022, from https://www.britannica.com/science/rhodopsin
3. Uniprot. (n.d.). Retrieved March 18, 2022, from https://www.uniprot.org/uniprot/Q15835.fasta

DATE: 20-03-22

WEBLEM 3b

I-TASSAR

(URL: https://zhanggroup.org/I-TASSER/)

AIM:

To perform tertiary structure prediction by threading approach using I-TASSER server for query rhodopsin.

INTRODUCTION:

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye—specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.

I-TASSER server is an on-line platform that implements the I-TASSER based algorithms for protein structure and function predictions. It allows academic users to automatically generate high-quality model predictions of 3D structure and biological function of protein molecules from their amino acid sequences.

METHODOLOGY:

1. Open homepage for I-TASSER. (URL: https://zhanggroup.org/I-TASSER/)
2. Complete registration.
3. Submit FASTA sequence for kinase.
4. Observe and interpret results.

OBSERVATION:

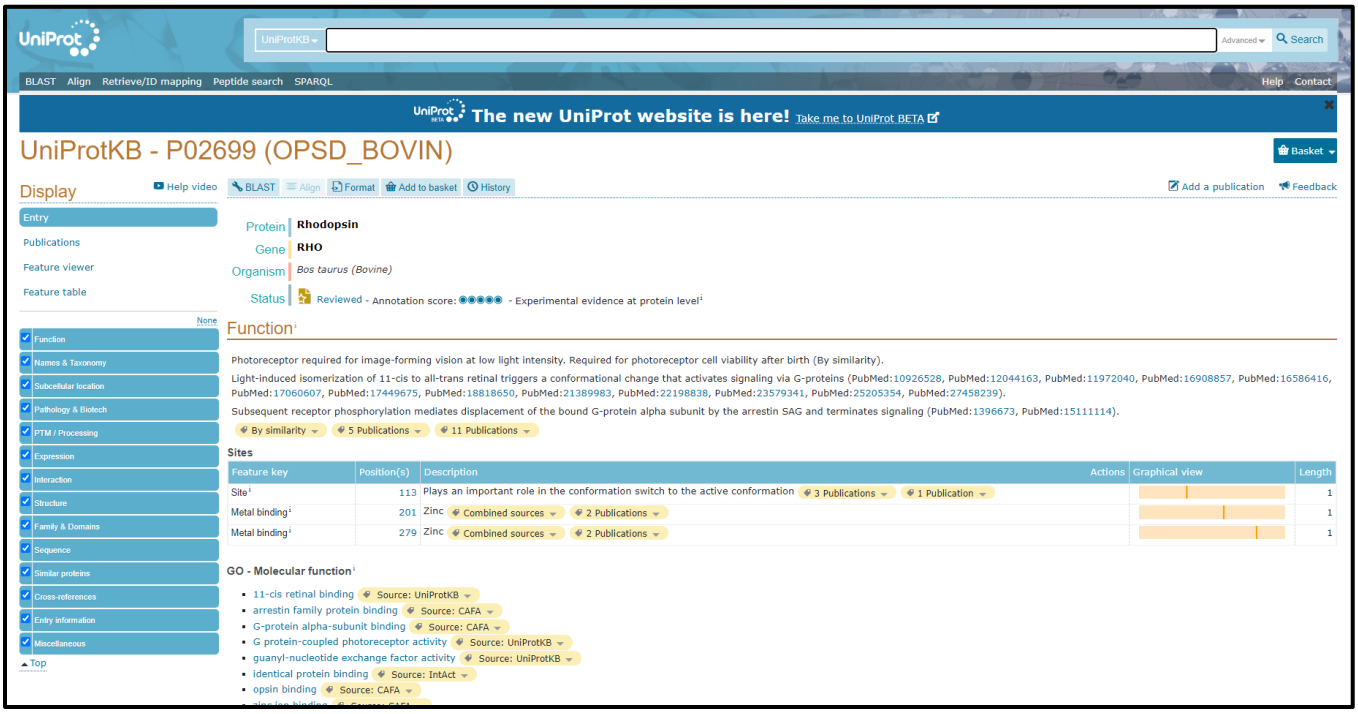


Fig1. Result page for Rhodopsin in UniProt database

```
>sp|P02699|OPSD_BOVIN Rhodopsin OS=Bos taurus OX=9913 GN=RHO PE=1 SV=1
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA
```

Fig2. FASTA sequence for Rhodopsin

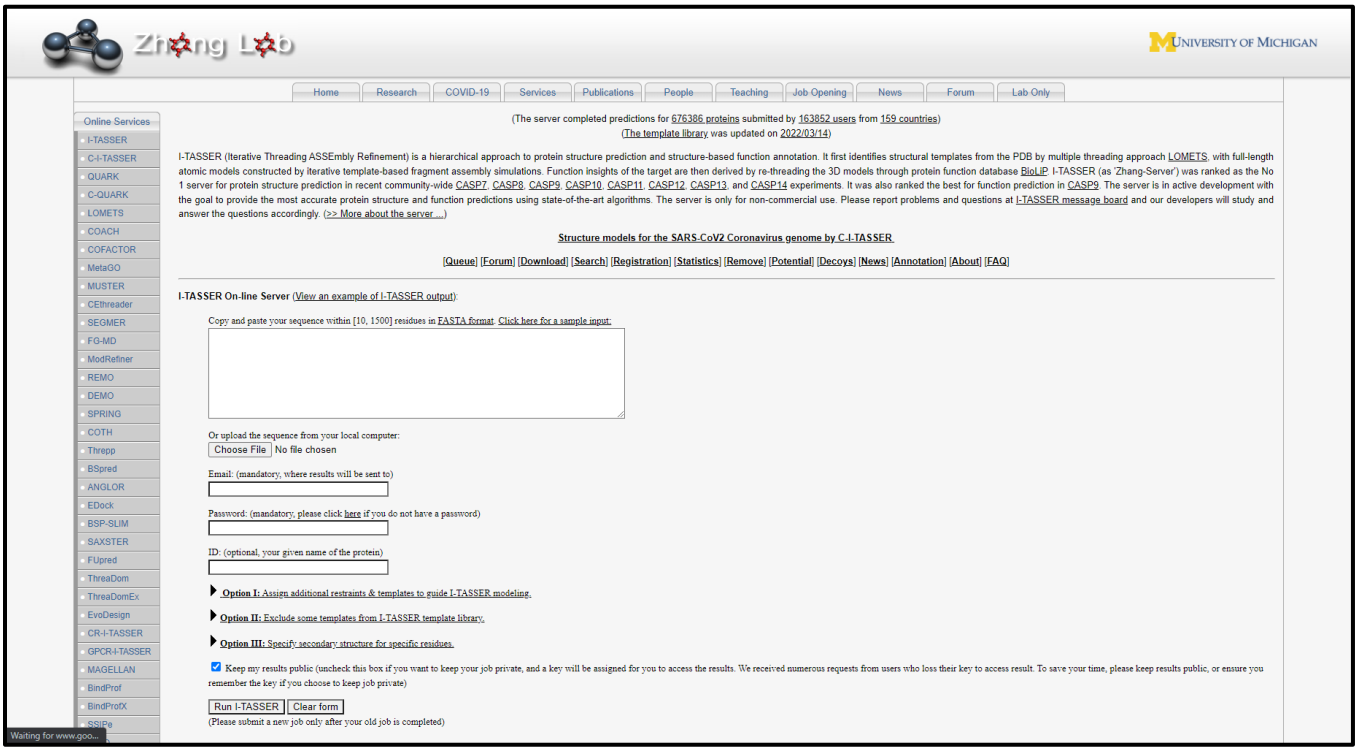


Fig3. Homepage for I-TASSAR

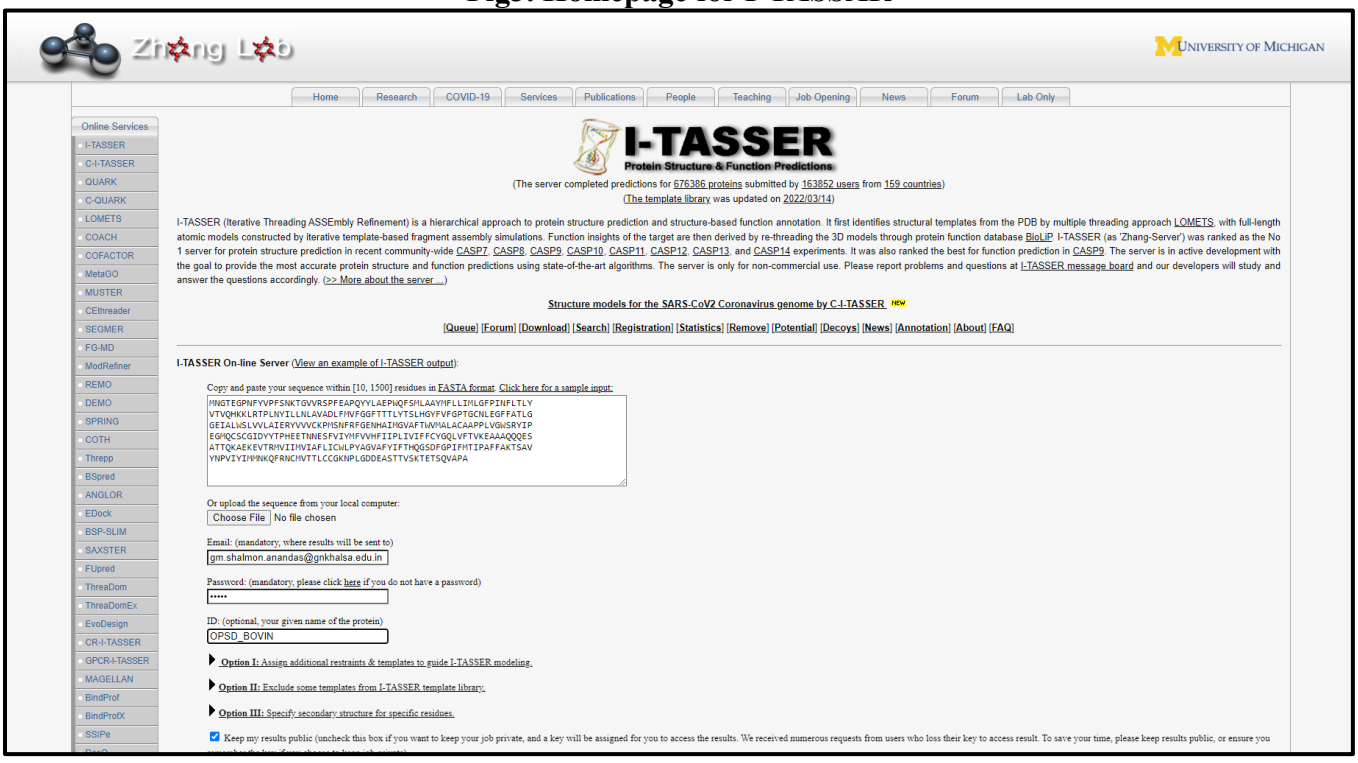


Fig4. Submission of query

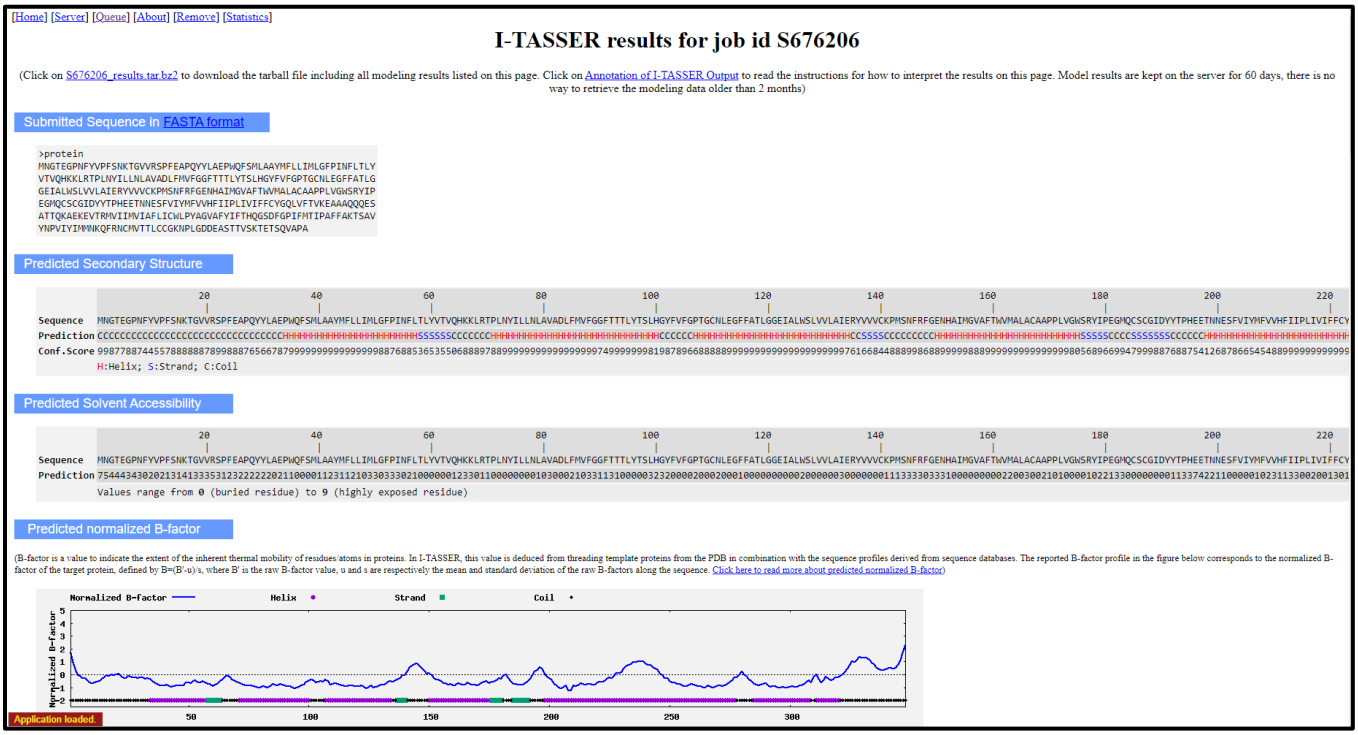


Fig5. Result for predicted secondary structure and normalized B-factor

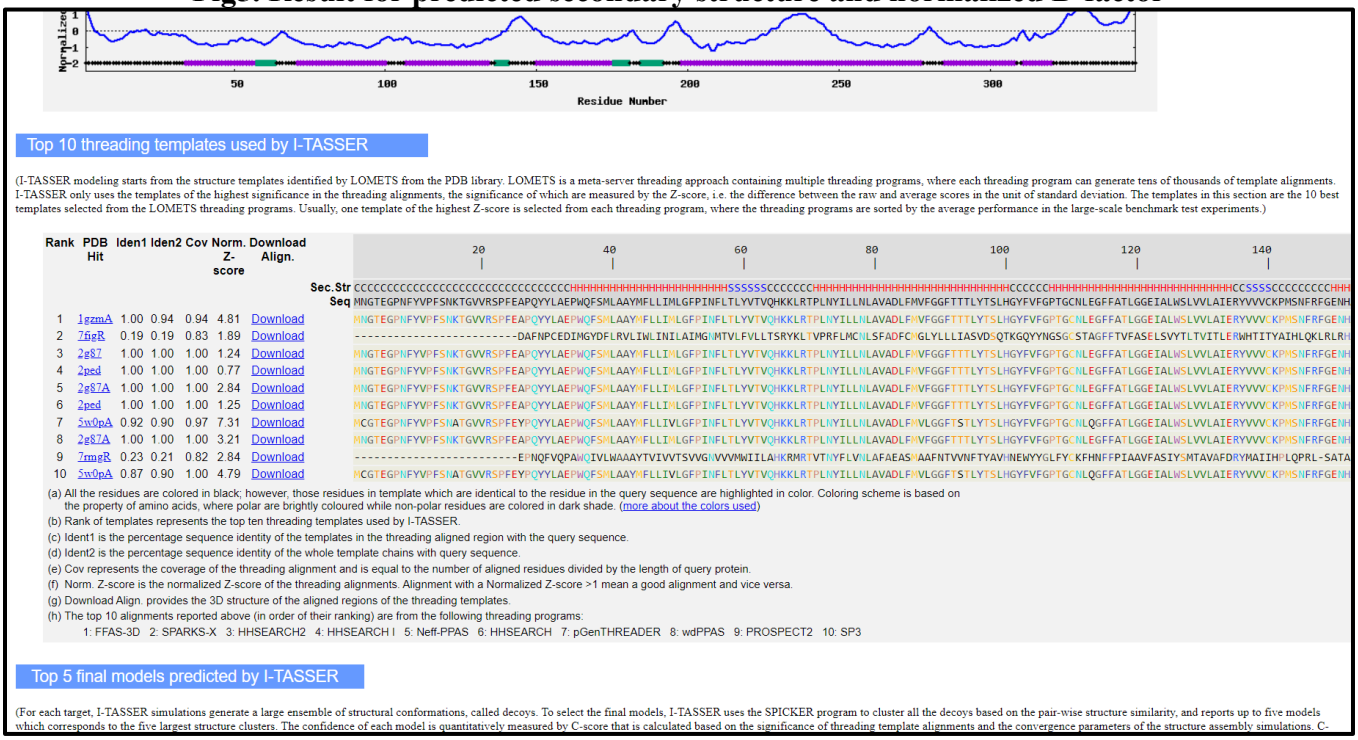


Fig7. Result for top 10 threading templates

a. All the residues are coloured in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in colour. Colouring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are coloured in dark shade. (More about the colours used)

b. Rank of templates represents the top ten threading templates used by I-TASSER.

c. Iden1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.

d. Iden2 is the percentage sequence identity of the whole template chains with query sequence.
e. Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein.
f. Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.
g. Download Align. provides the 3D structure of the aligned regions of the threading templates.
h. The top 10 alignments reported above (in order of their ranking) are from the following threading programs: 1: FFAS-3D 2: SPARKS-X 3: HHSEARCH2 4: HHSEARCH I 5: Neff-PPAS 6: HHSEARCH 7: pGenTHREADER 8: wdPPAS 9: PROSPECT2 10: SP3

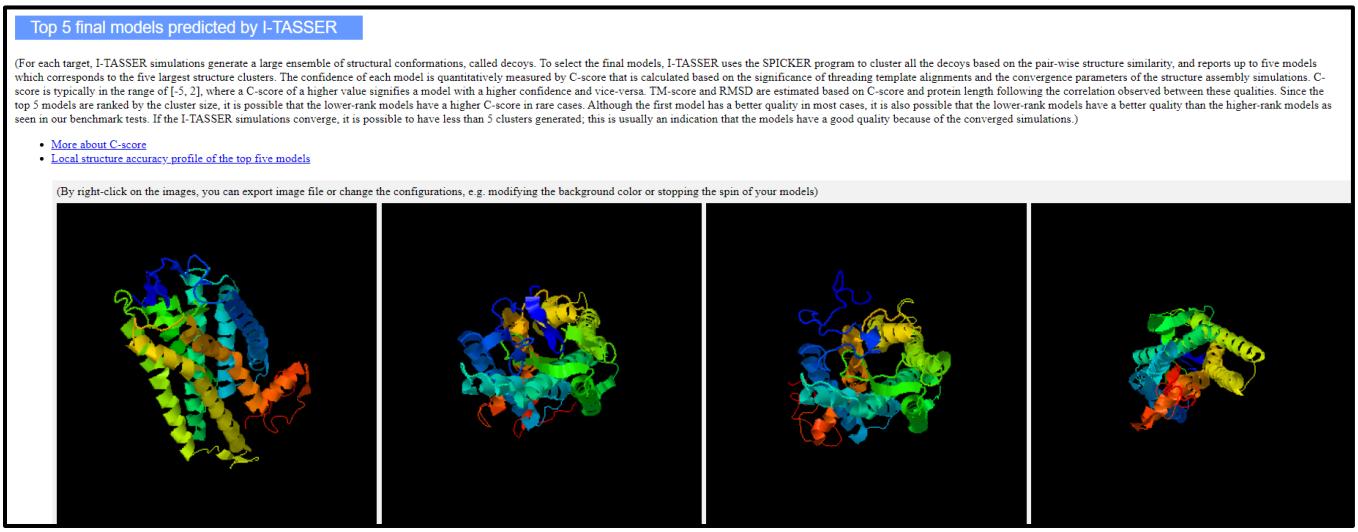


Fig8. Result for top 5 final models predicted

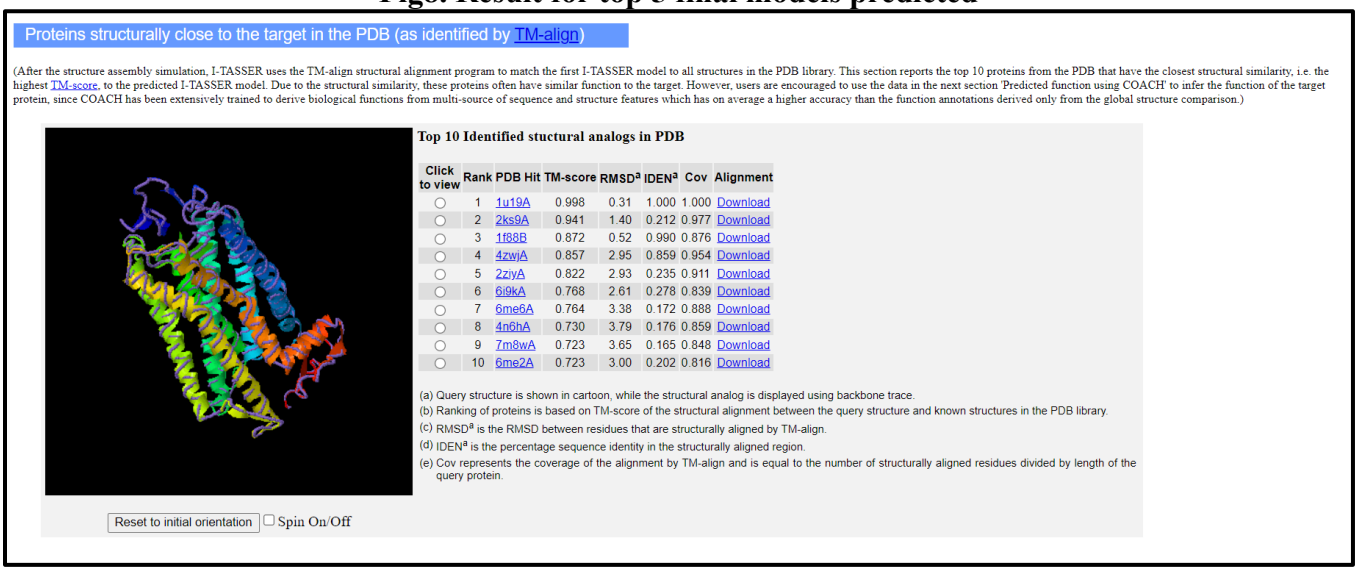


Fig9. Result for proteins that are structurally close to target

a. Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.
b. Ranking of proteins is based on TM-score of the structural alignment between the query structure and known structures in the PDB library
c. RMSDa is the RMSD between residues that are structurally aligned by TM-align
d. IDENa is the percentage sequence identity in the structurally aligned region.

e. Cov represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

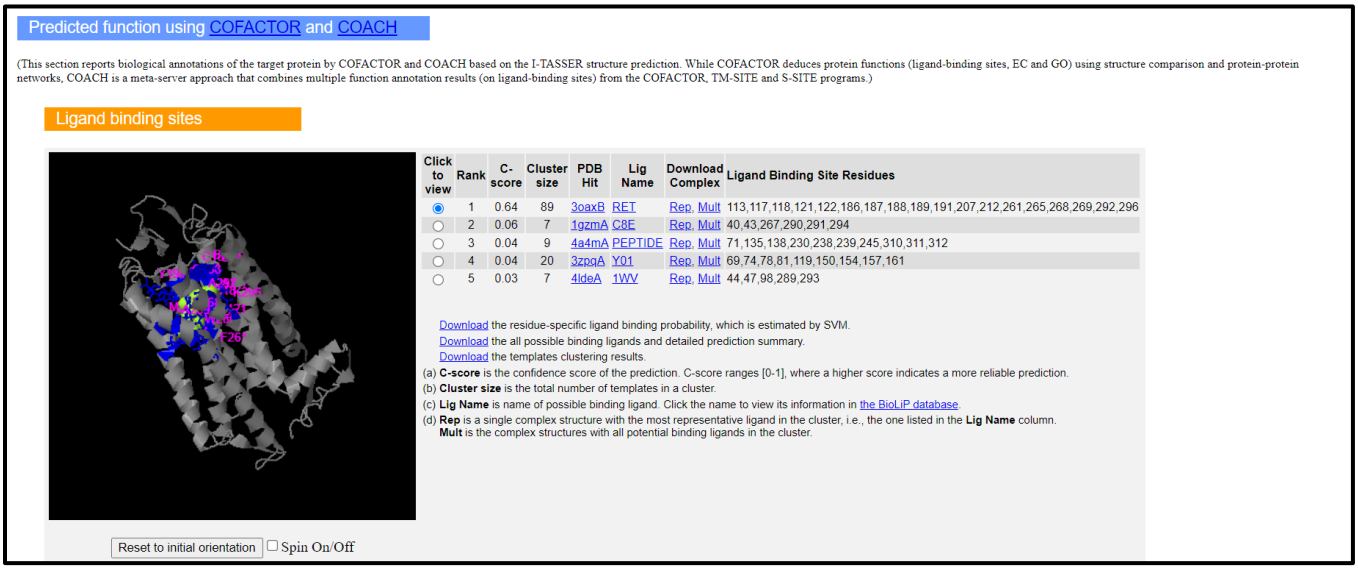


Fig10. Result for predicted fuctions

a. C-score is the confidence score of the prediction. C-scores ranges [0-1], where a higher score indicated a more reliable prediction.
b. Cluster size is the total number of templates in a cluster.
c. Lig Name is name of possible binding ligand. Click the name to view its information in the BioLiP database.
d. Rep is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the Lig Name column. Mult is the complex structures with all potential binding ligands in the cluster.

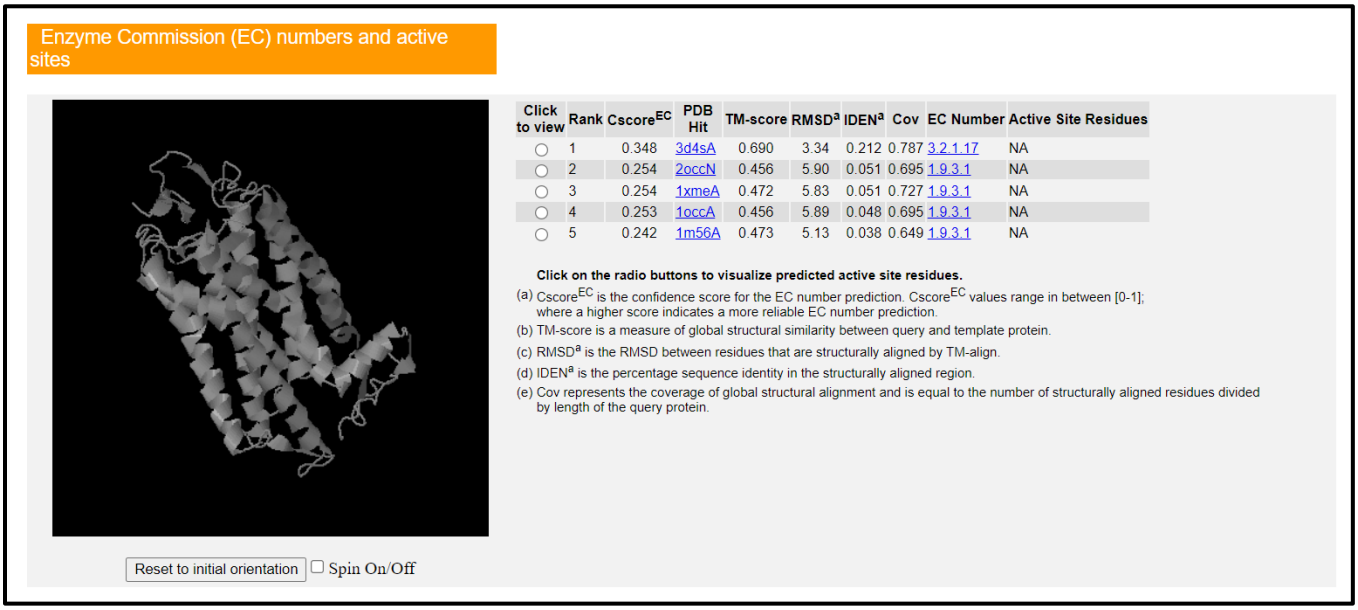
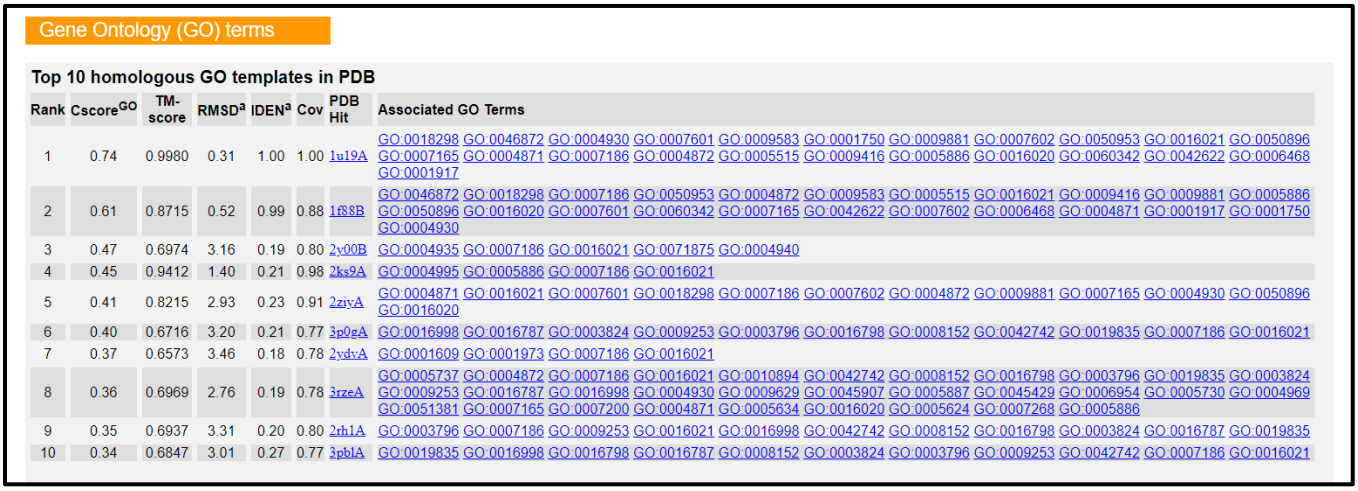


Fig11. Enzyme commission (EC) numbers and active sites

a. CscoreEC is the confidence score for the EC number prediction. CscoreEC values range in between [0-1]; where a higher score indicates a more reliable EC number prediction.
b. TM-score is a measure of global structural similarity between query and template protein.

c. RMSDa is the RMSD between residues that are structurally aligned by TM-align.
d. IDENa is the percentage sequence identity in the structurally aligned region.
e. Cov represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein.



Gene Ontology (GO) terms

Top 10 homologous GO templates in PDB

Rank	Cscore^{GO}	TM-score	RMSD^{a}	IDEN^{a}	Cov	PDB Hit	Associated GO Terms
1	0.74	0.9980	0.31	1.00	1.00	1u19A	GO:0018298 GO:0046872 GO:0004930 GO:0007601 GO:0009583 GO:0001750 GO:0009881 GO:0007602 GO:0050953 GO:0016021 GO:0050896 GO:0007165 GO:0004871 GO:0007186 GO:0004872 GO:0005515 GO:0009416 GO:0005886 GO:0016020 GO:0060342 GO:0042622 GO:0006468 GO:0001917
2	0.61	0.8715	0.52	0.99	0.88	1f88B	GO:0046872 GO:0018298 GO:0007186 GO:0050953 GO:0004872 GO:0009583 GO:0005515 GO:0016021 GO:0009416 GO:0009881 GO:0005886 GO:0050896 GO:0016020 GO:0007601 GO:0060342 GO:0007165 GO:0042622 GO:0007602 GO:0006468 GO:0004871 GO:0001917 GO:0001750 GO:0004930
3	0.47	0.6974	3.16	0.19	0.80	2v00B	GO:0004935 GO:0007186 GO:0016021 GO:0071875 GO:0004940
4	0.45	0.9412	1.40	0.21	0.98	2ks9A	GO:0004995 GO:0005886 GO:0007186 GO:0016021
5	0.41	0.8215	2.93	0.23	0.91	2ziyA	GO:0004871 GO:0016021 GO:0007601 GO:0018298 GO:0007186 GO:0007602 GO:0004872 GO:0009881 GO:0007165 GO:0004930 GO:0050896 GO:0016020
6	0.40	0.6716	3.20	0.21	0.77	3p0gA	GO:0016998 GO:0016787 GO:0003824 GO:0009253 GO:0003796 GO:0016798 GO:0008152 GO:0042742 GO:0019835 GO:0007186 GO:0016021
7	0.37	0.6573	3.46	0.18	0.78	2vdvA	GO:0001609 GO:0001973 GO:0007186 GO:0016021
8	0.36	0.6969	2.76	0.19	0.78	3rzeA	GO:0005737 GO:0004872 GO:0007186 GO:0016021 GO:0010894 GO:0042742 GO:0008152 GO:0016798 GO:0003796 GO:0019835 GO:0003824 GO:0009253 GO:0016787 GO:0016998 GO:0004930 GO:0009629 GO:0045907 GO:0005887 GO:0045429 GO:0006954 GO:0005730 GO:0004969 GO:0051381 GO:0007165 GO:0007200 GO:0004871 GO:0005634 GO:0016020 GO:0005624 GO:0007268 GO:0005886
9	0.35	0.6937	3.31	0.20	0.80	2rh1A	GO:0003796 GO:0007186 GO:0009253 GO:0016021 GO:0016998 GO:0042742 GO:0008152 GO:0016798 GO:0003824 GO:0016787 GO:0019835
10	0.34	0.6847	3.01	0.27	0.77	3pb1A	GO:0019835 GO:0016998 GO:0016798 GO:0016787 GO:0008152 GO:0003824 GO:0003796 GO:0009253 GO:0042742 GO:0007186 GO:0016021

Fig12. Gene Ontology (GO) terms

a. CscoreGO is a combined measure for evaluating global and local similarity between query and template protein. It's range is [0-1] and higher values indicate more confident predictions.
b. TM-score is a measure of global structural similarity between query and template protein.
c. RMSDa is the RMSD between residues that are structurally aligned by TM-align.
d. IDENa is the percentage sequence identity in the structurally aligned region.
e. Cov represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein.
f. The second table shows a consensus GO terms amongst the top scoring templates. The GO-Score associated with each prediction is defined as the average weight of the GO term, where the weights are assigned based on CscoreGO of the template.

RESULT:

I-TASSER was used to predict the tertiary structure of Kinase based on threading approach. The information regarding solvent accessibility, normalized B-factor, top 10 threading templates, top 5 final models, proteins that are structurally close to target, functions and active sites were predicted.

CONCLUSION:

Thus, I-TASSER can be used to predict tertiary structures of proteins by threading method. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions and drug designing.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.
2. kinase | Definition, Biology, & Function. (n.d.). Encyclopedia Britannica. Retrieved March 8, 2022, from https://www.britannica.com/science/kinase
3. I-TASSER server for protein structure and function prediction. (n.d.-b). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org/I-TASSER/

4. I-TASSER results. (n.d.). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org//I- TASSER/output/S673761/

DATE: 20-03-22

WEBLEM 3c
ROBETTA
(URL: https://robetta.bakerlab.org/)

AIM:

To perform tertiary structure prediction by Ab-Initio approach using ROBETTA server for query Rhodopsin.

INTRODUCTION:

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye—specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.

The Robetta server provides automated tools for protein structure prediction and analysis. For structure prediction, sequences submitted to the server are parsed into putative domains and structural models are generated using either comparative modeling orde novo structure prediction methods. If a confident match to a protein of known structure is found using BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative modeling. If no match is found, structure predictions are made using the de novo Rosetta fragment insertion method. Experimental nuclear magnetic resonance (NMR) constraints data can also be submitted with a query sequence for RosettaNMR de novo structure determination. Other current capabilities include the prediction of the effects of mutations on protein–protein interactions using computational interface alanine scanning. The Rosetta protein design and protein–protein docking methodologies will soon be available through the server as well.

METHODOLOGY:

1. Open homepage for Robetta (URL: https://robetta.bakerlab.org/)
2. Complete registration
3. Submit FASTA sequence for kinase.
4. Observe and interpret results.

OBSERVATION:

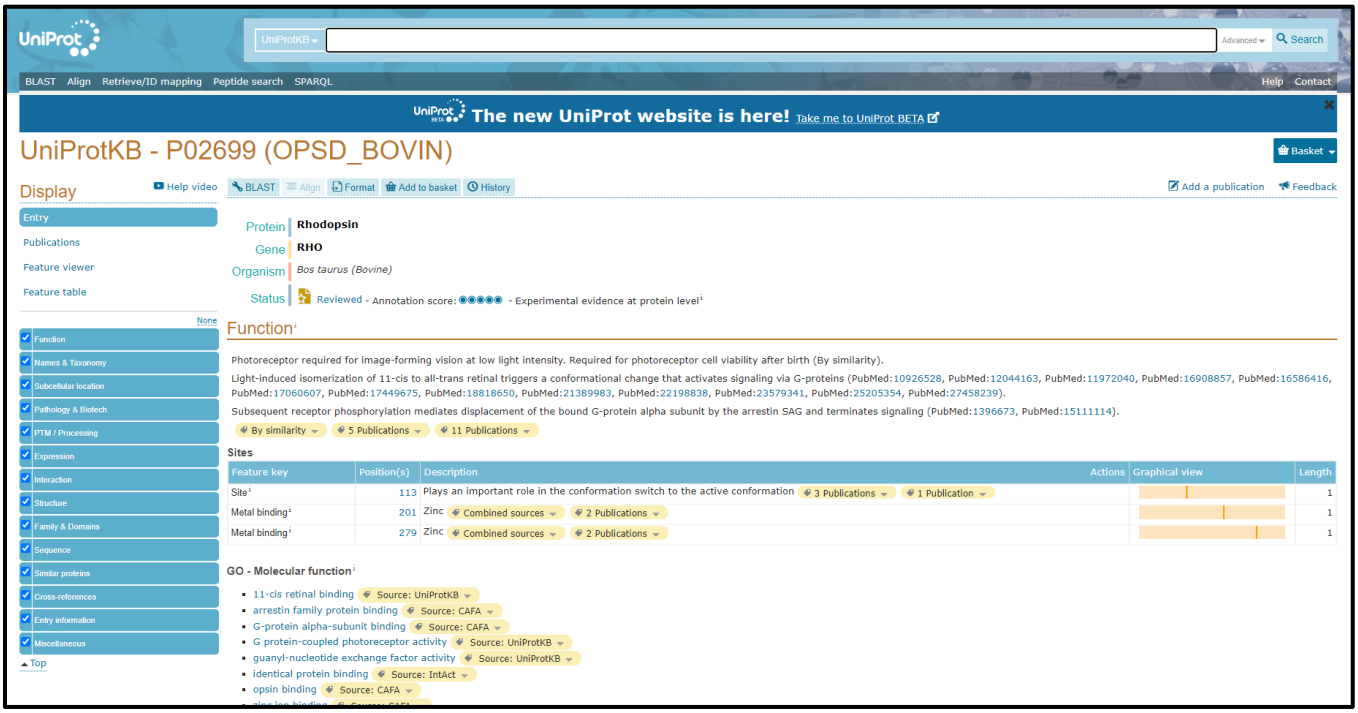


Fig1. Result page for Rhodopsin in UniProt database

```
>sp|P02699|OPSD_BOVIN Rhodopsin OS=Bos taurus OX=9913 GN=RHO PE=1 SV=1
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA
```

Fig2. FASTA sequence for Rhodopsin

Fig3. Model 1 with atom co-ordinates

Fig4. Model 2 with atom co-ordinates

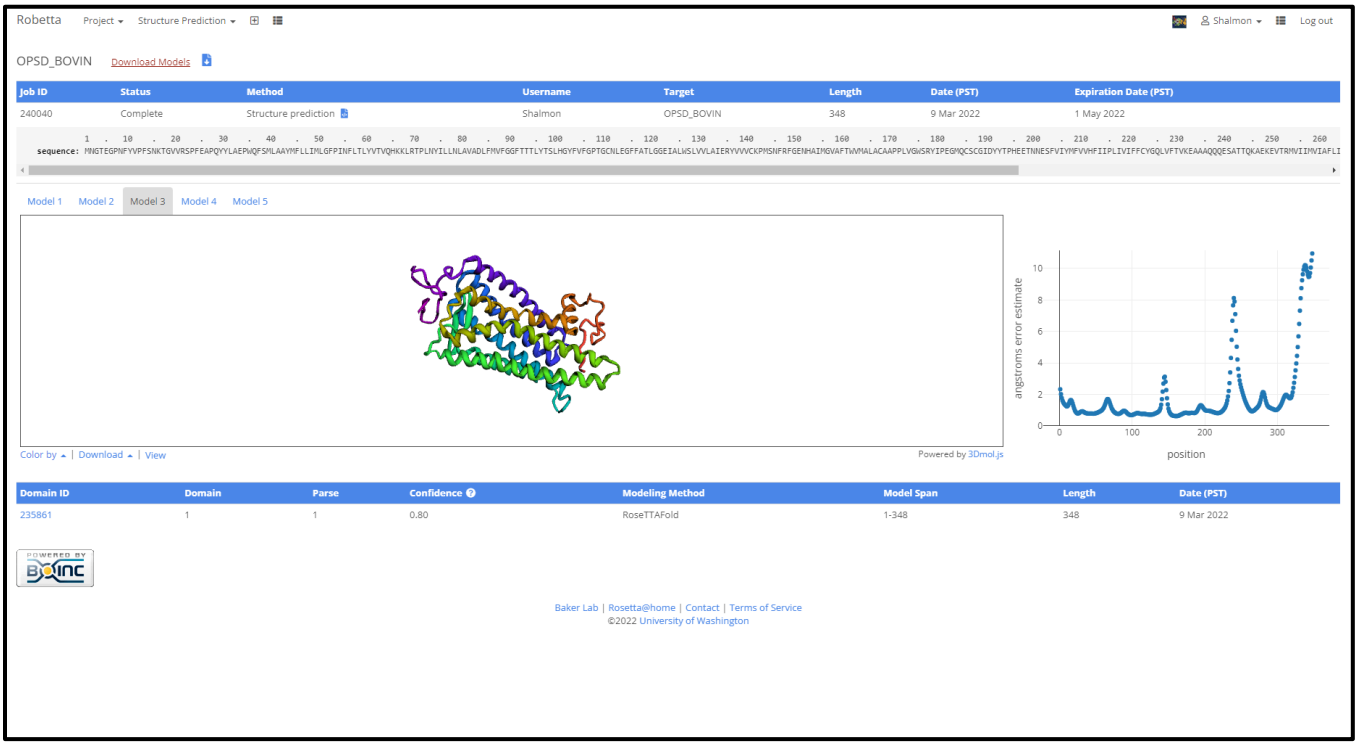


Fig5. Model 3 with atom co-ordinates

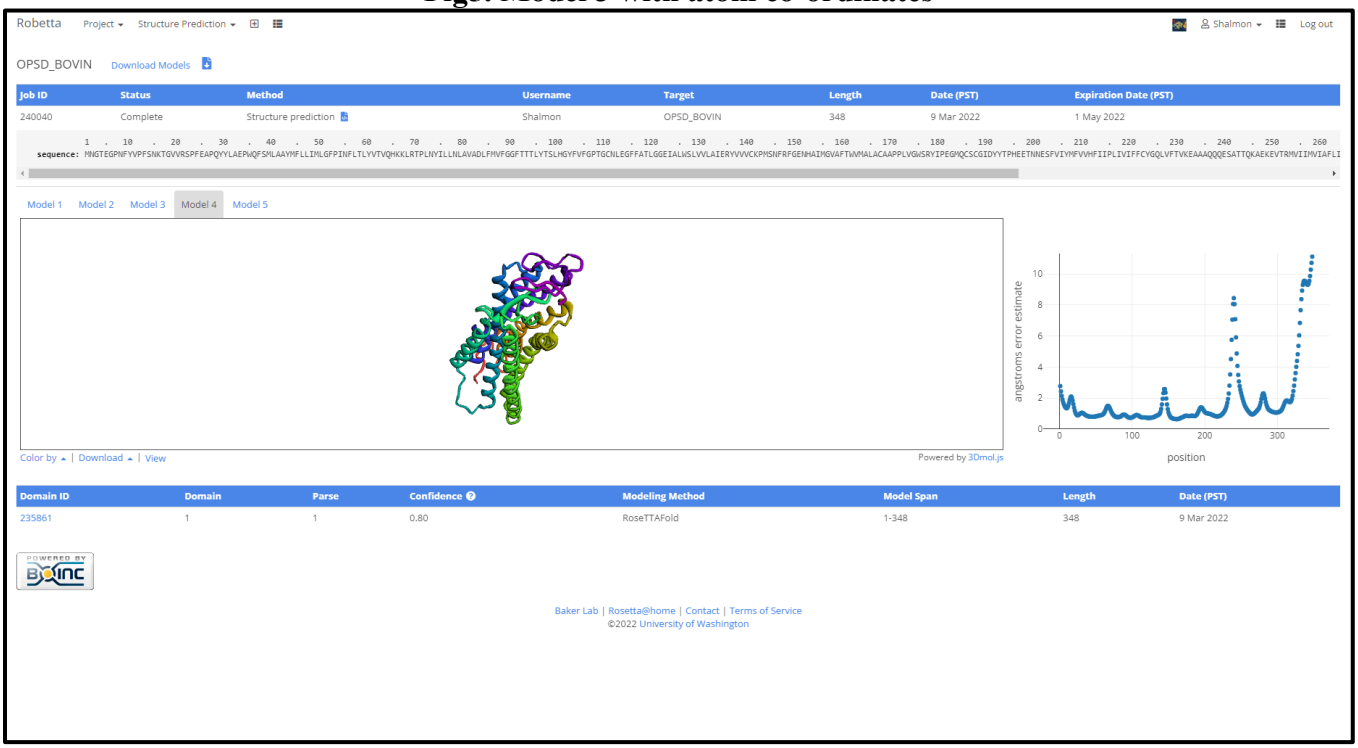


Fig6. Model 4 with atom co-ordinates

Fig7. Model 5 with atom co-ordinates

RESULT:

Robetta was used to predict the tertiary structure of Rhodopsin based on ab-initio approach.

CONCLUSION:

Thus, Robetta can be used to predict tertiary structures of proteins by ab-initio method. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions and drug designing.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.
2. Encyclopædia Britannica, inc. (n.d.). *Rhodopsin*. Encyclopædia Britannica. Retrieved March 18, 2022, from https://www.britannica.com/science/rhodopsin
3. Robetta (2021b). Bakerlab.org. Retrieved March 18, 2022, from https://robetta.bakerlab.org/results.php?id=240040

DATE: 10-03-22

WEBLEM 4

Introduction to Validation server – SAVES server

Homology model, threading method and ab-initio method of tertiary structure prediction all have to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules. This involves checking anomalies in φ–ψ angles, bond lengths, close contacts, and so on. Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account. This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures. By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

SAVES server:

SAVES server contains various tools for structure validation that are all integrated in one single server. The tool available are:

ERRAT:

A novel method for differentiating between correctly and incorrectly determined regions of protein structures based on characteristic atomic interaction is described. Different types of atoms are distributed nonrandomly with respect to each other in proteins. Errors in model building lead to more randomized distributions of the different atom types, which can be distinguished from correct distributions by statistical methods. Atoms are classified in one of three categories: carbon (C), nitrogen (N), and oxygen (O). This leads to six different combinations of pairwise noncovalently bonded interactions (CC, CN, CO, NN, NO, and OO). A quadratic error function is used to characterize the set of pairwise interactions from nine-residue sliding windows in a database of 96 reliable protein structures. Regions of candidate protein structures that are mistraced or misregistered can then be identified by analysis of the pattern of nonbonded interactions from each window.

Errat is a program for verifying protein structures determined by crystallography. Error values are plotted as a function of the position of a sliding 9-residue window. The error function is based on the statistics of non- bonded atom-atom interactions in the reported structure (compared to a database of reliable high-resolution structures).

A plot of an initial model and a final model is retrieved. Regions of the structure that can be rejected at the 95% confidence level are yellow; 5% of a good protein structure is expected to have an error value above this level. Regions that can be rejected at the 99% level are shown in red. Generally speaking, the method is sensitive to smaller errors than 3-D Profile analysis.

Verify3D:

It is another server using the statistical approach. It uses a precomputed database containing eighteen environmental profiles based on secondary structures and solvent exposure, compiled from high-resolution protein structures. To assess the quality of a protein model, the secondary structure and solvent exposure propensity of each residue are calculated. It determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures. If the parameters of a residue fall within one of the profiles, it receives a high score, otherwise a low score. The result is a two-dimensional graph illustrating the folding quality of each residue of the protein structure. The threshold value is normally set at zero. Residues with scores below zero are considered to have an unfavourable environment.

PROVES:

It calculates the volumes of atoms in macromolecules using an algorithm which treats the atoms like hard spheres and calculates a statistical Z-score deviation for the model from highly resolved (2.0 Å or better) and refined (R-factor of 0.2 or better) PDB-deposited structures.
Standard ranges of atomic and residue volumes are computed in 64 highly resolved and well-refined protein crystal structures using the classical Voronoi procedure. Deviations of the atomic volumes from the standard values, evaluated as the volume Z-scores, are used to assess the quality of protein crystal structures. To score a structure globally, we compute the volume Z-score root mean square deviation (Z-score rms), which measures the average magnitude of the volume irregularities in the structure. We find that the Z-score rms decreases as the resolution and R-factor improve, consistent with the fact that these improvements generally reflect more accurate models. From the Z-score rms distribution in structures with a given resolution or R- factor, we determine the normal limits in Z-score rms values for structures solved at that resolution or R- factor. Structures whose Z-score rms exceeds these limits are considered as outliers. Such structures also exhibit unusual stereochemistry, as revealed by other analyses. Absolute Z-scores of individual atoms are used to identify problems in specific regions within a protein model. These Z-scores correlate fairly well with the atomic B-factors, and atoms having absolute Z-scores > 3, occur at or near regions in the model where programs such as PROCHECK identify unusual stereochemistry. Atomic volumes, themselves not directly restrained in crystallographic refinement, can thus provide an independent, rather sensitive, measure of the quality of a protein structure.

WHAT_CHECK:

Derived from a subset of protein verification tools from the WHAT IF program, this does extensive checking of many sterochemical parameters of the residues in the model. WHAT IF is a comprehensive protein analysis server that validates a protein model for chemical correctness. It has many functions, including checking of planarity, collisions with symmetry axes (close contacts), proline puckering, anomalous bond angles, and bond lengths. It also allows the generation of Ramachandran plots as an assessment of the quality of the model.

PROCHECK:

It is a UNIX program that is able to check general physicochemical parameters such as φ–ψ angles, chirality, bond lengths, bond angles, and so on. It checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry. The parameters of the model are used to compare with those compiled from well-defined, high-resolution structures. If the program detects unusual features, it highlights the regions that should be checked or refined further.

CRYST:

This program searches the Protein Data Bank for entries that have a unit cell similar to your input file. CRYST1 record required. Use the standalone CRYST server for more options.
The assessment results can be different using different verification programs. Because no single method is clearly superior to any other, a good strategy is to use multiple verification methods and identify the consensus between them. It is also important to keep in mind that the evaluation tests performed by these programs only check the stereochemical correctness, regardless of the accuracy of the model, which may or may not have any biological meaning. Thus, SAVES server is an excellent platform that provides various validation methods to accurately validate the structures.

REFERENCES:

1. SAVESv6.0 - Structure Validation Server. (n.d.-b). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/
2. ERRAT – UCLA-DOE Institute. (n.d.). Retrieved March 8, 2022, from https://www.doembi.ucla.edu/errat/
3. Chris Colovos; Todd O. Yeates (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. , 2(9), 1511–1519. doi:10.1002/pro.5560020916

4. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 220-222.
5. Joan Pontius; Jean Richelle; Shoshana J. Wodak (1996). Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. , 264(1), 0–136. doi:10.1006/jmbi.1996.0628

DATE: 10-03-22

WEBLEM 4a
SAVES server
(URL: https://saves.mbi.ucla.edu/)

AIM:

To validate structure qseq.B99990005 generated from modeller.

Introduction:

qseq.B99990005 is the structure predicted using homology modelling using modeller. The structure has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules. This can be done using SAVES server.

SAVES is a structure validation server that has various tools like Errat, Verify3D, Prove, Whatcheck, Procheck.and Cryst integrated in one single platform. This involves checking anomalies in φ–ψ angles, bond lengths, close contacts, and so on. Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account. This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures. By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

METHODOLOGY:

1. Open homepage for SAVES server. (URL: https://saves.mbi.ucla.edu/)
2. Upload structure retrieved from Modeller in PDB format.
3. Obtain results for Errat, Verify3D, Prove, Whatcheck and Procheck.
4. Observe and interpret the results.

OBSERVATION:

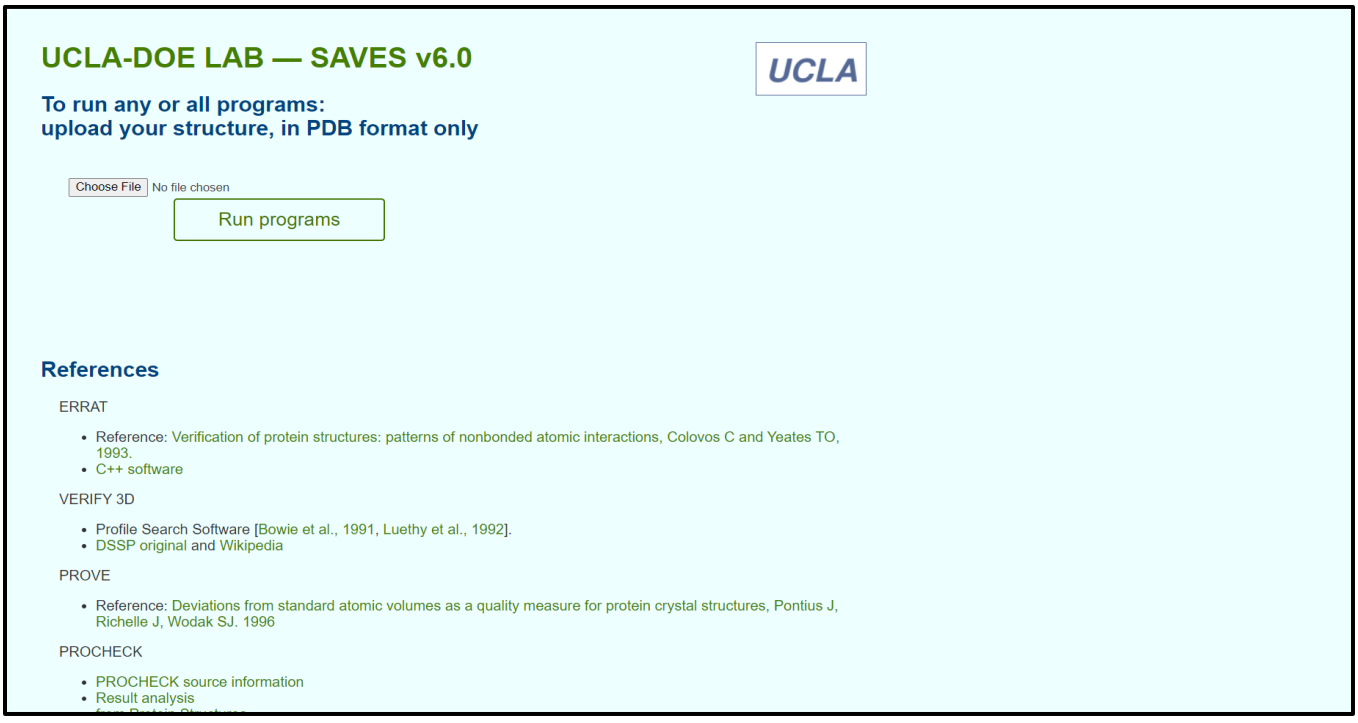


Fig1. Homepage for SAVES server

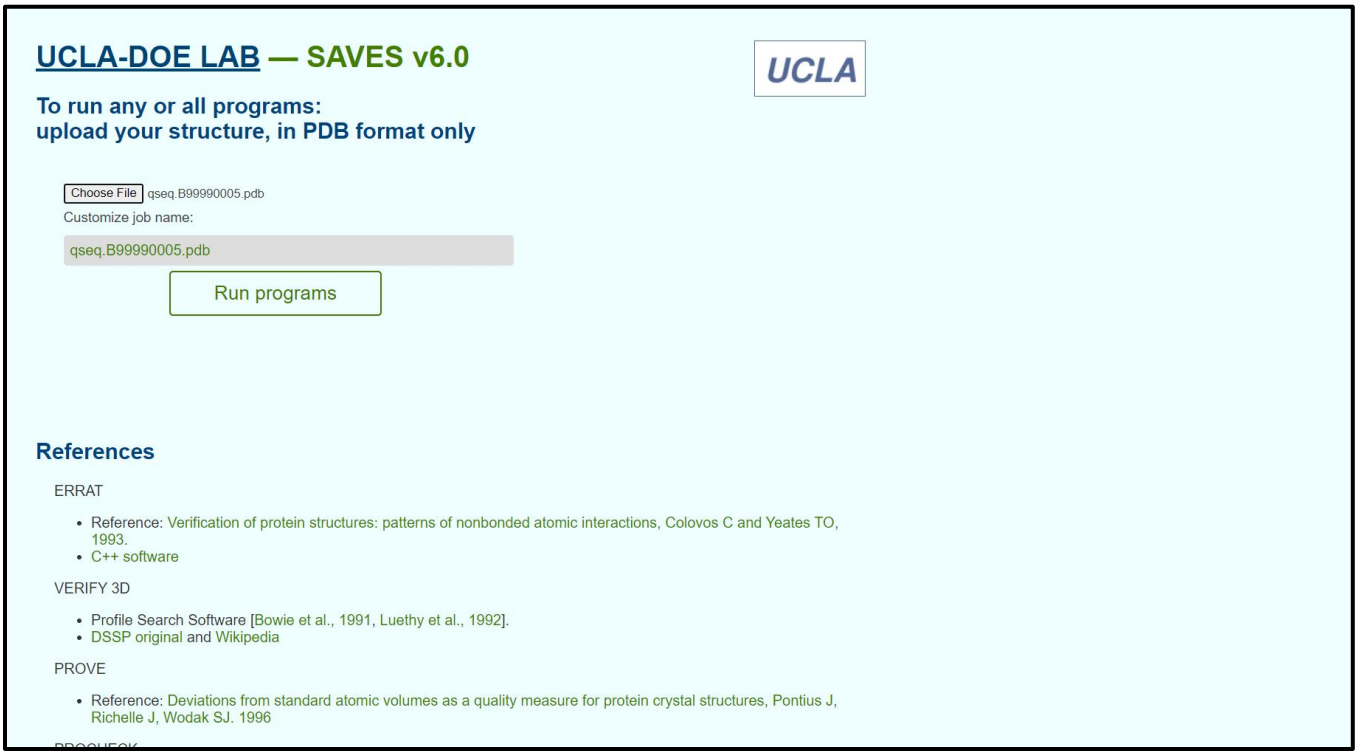


Fig2. Structure from Modeller for validation

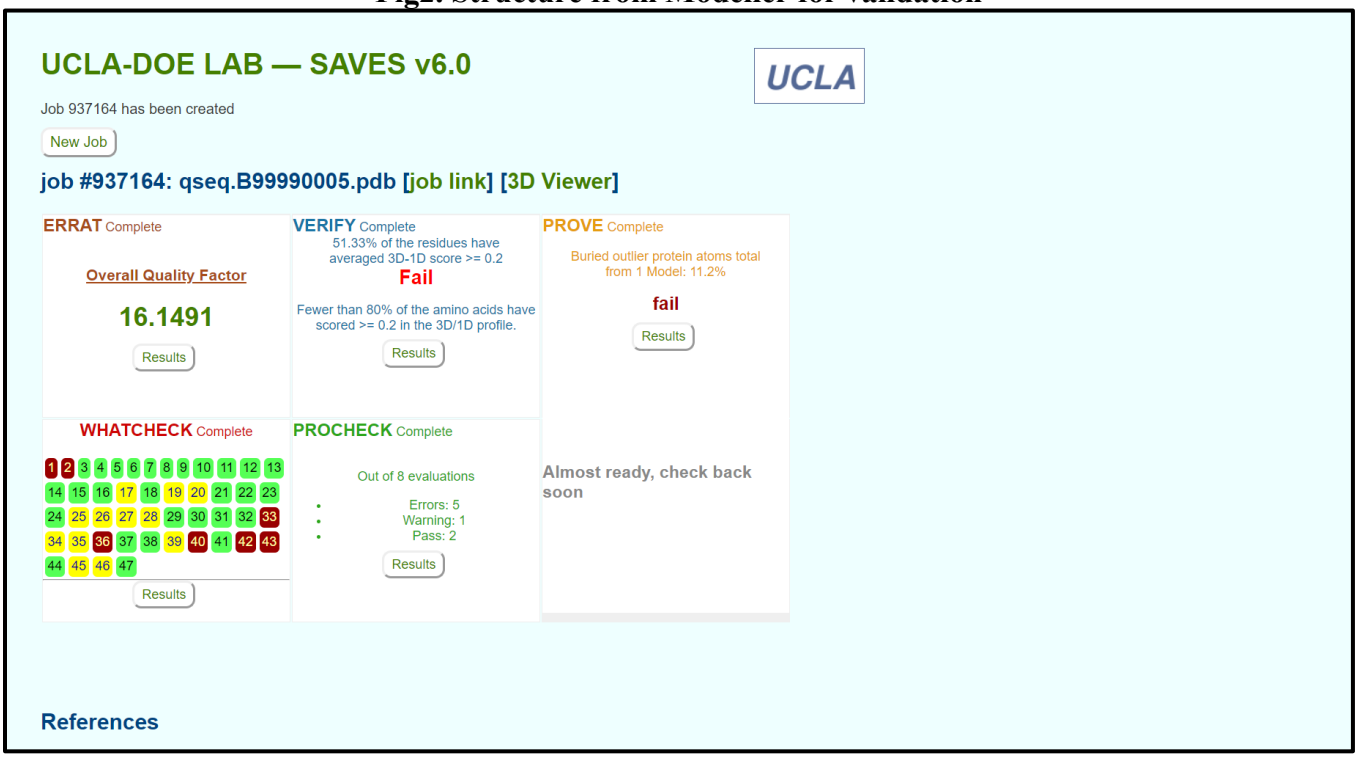


Fig3. Result page for structure validation for various servers

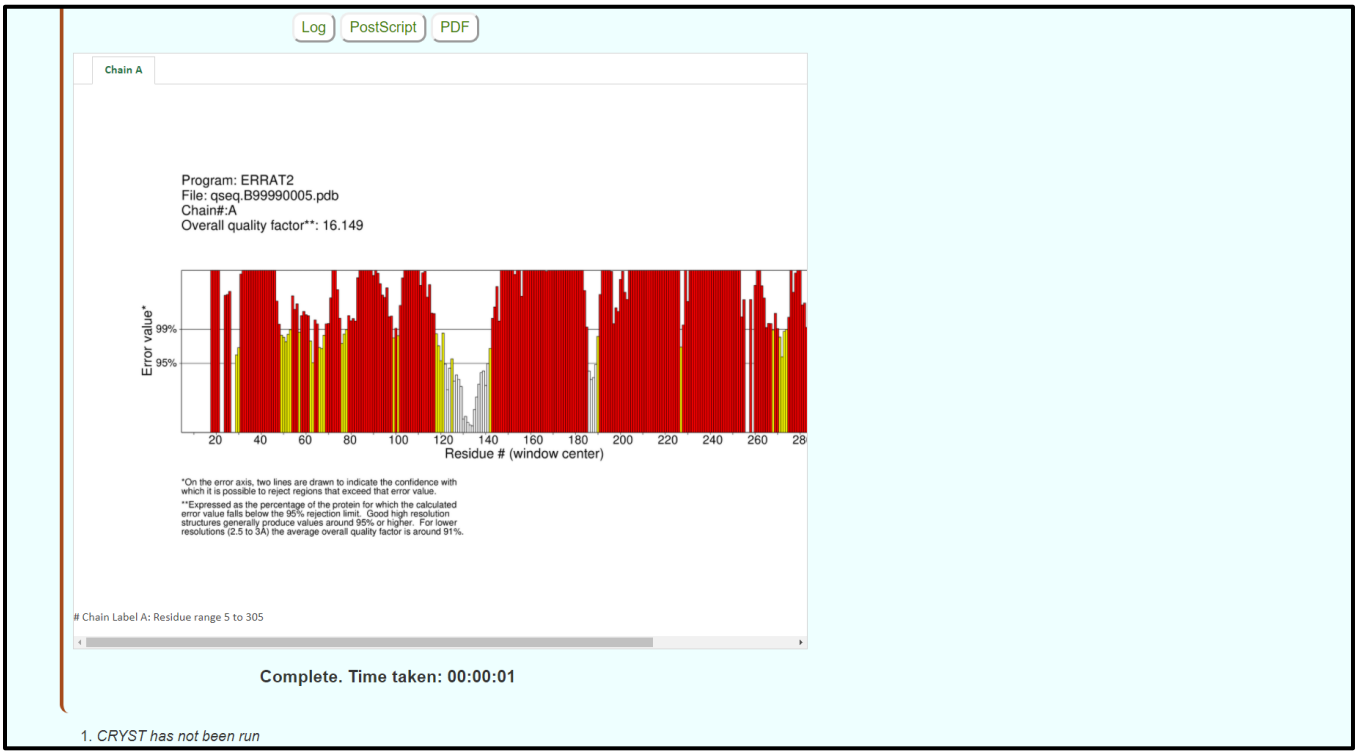


Fig4. Result page for ERRAT

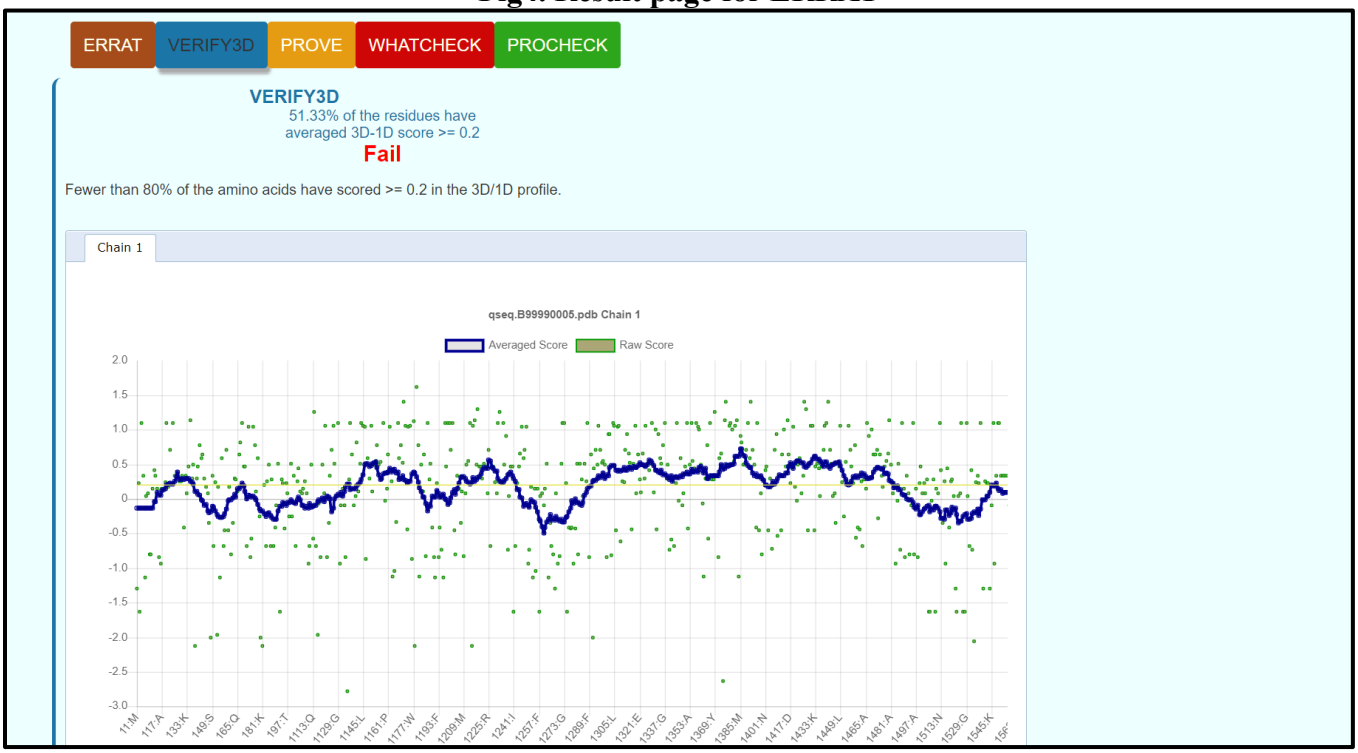


Fig5. Result page for Verify3D

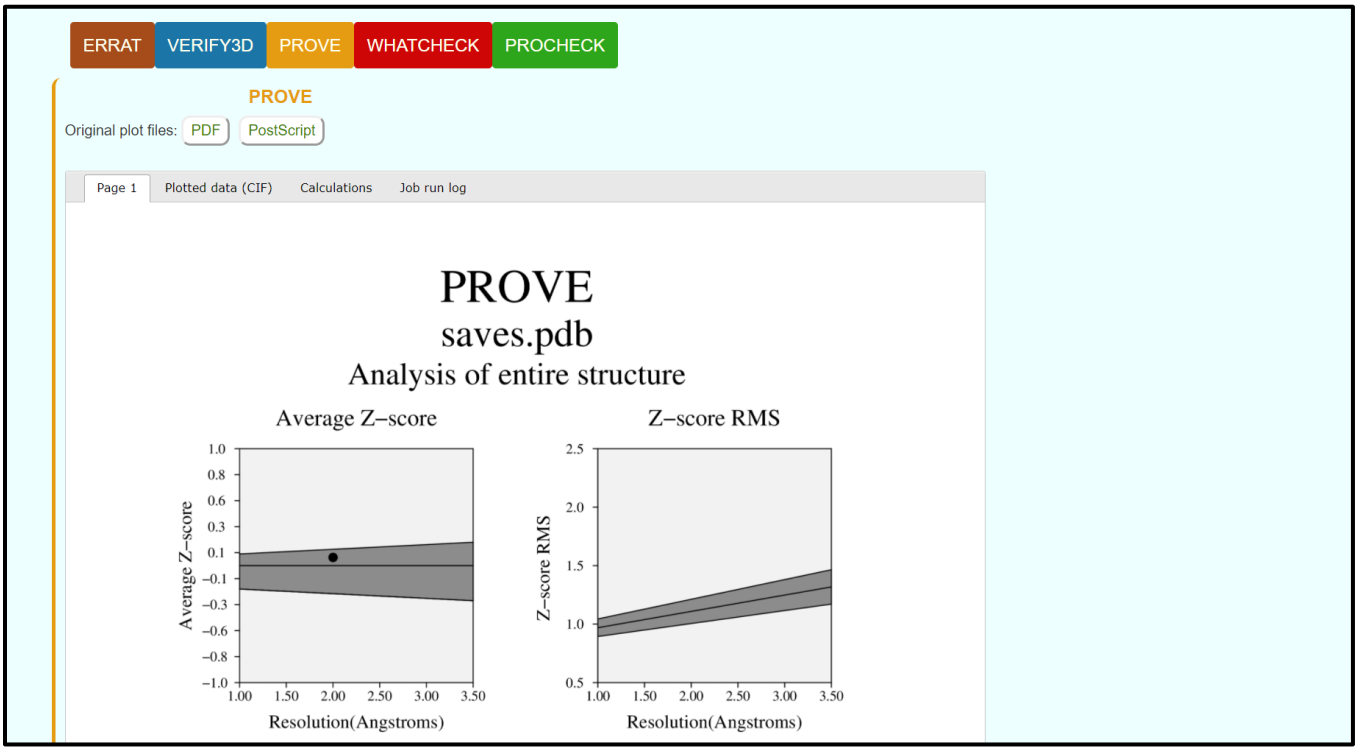


Fig6. Result page for Prove

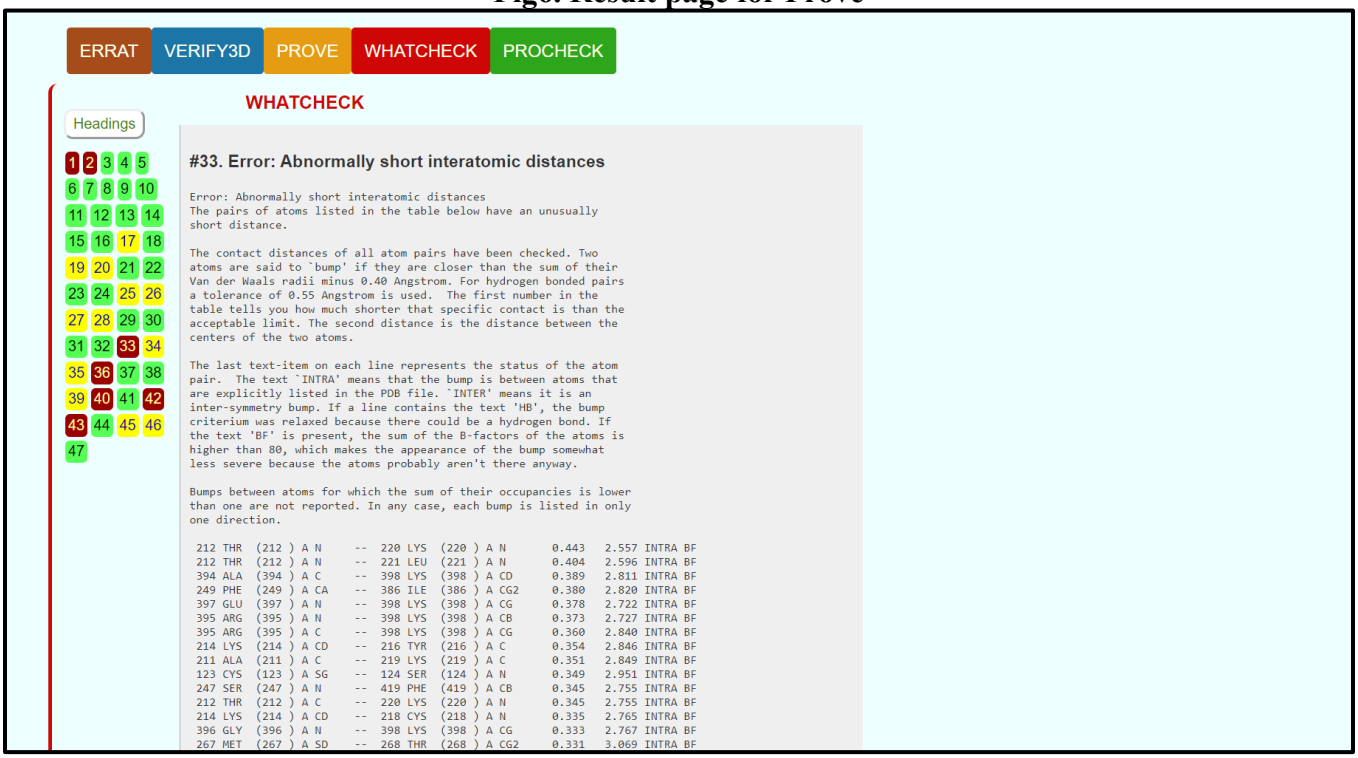


Fig7. Result page for Whatcheck

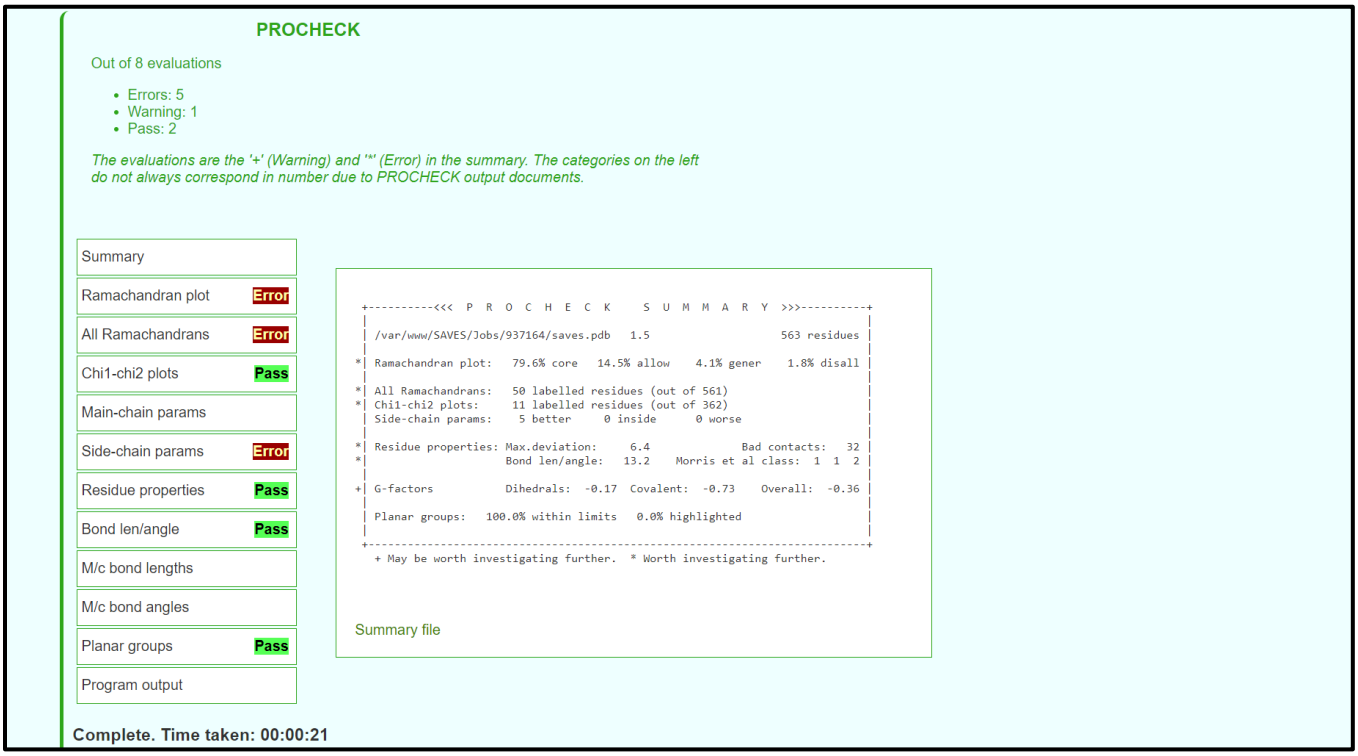


Fig8. Result page for procheck

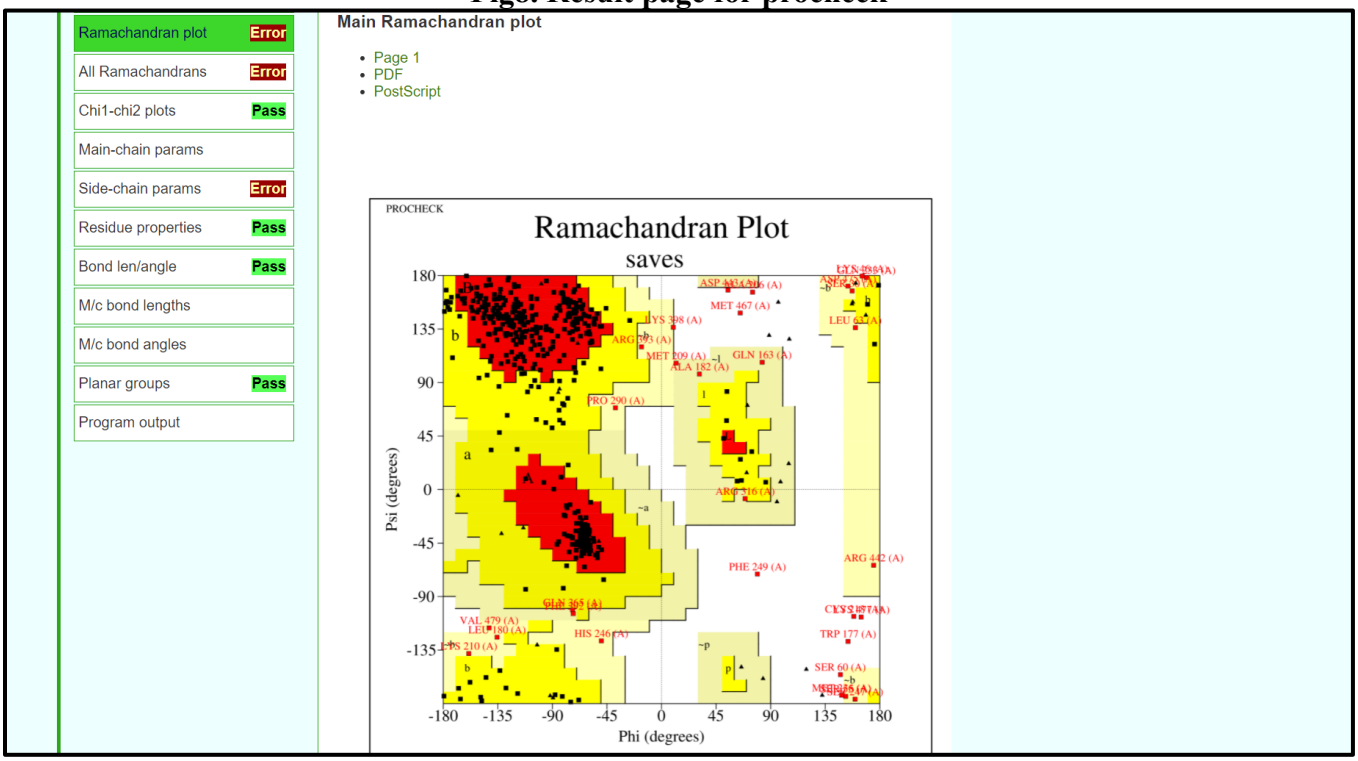


Fig9. Result page for Ramachandran plot

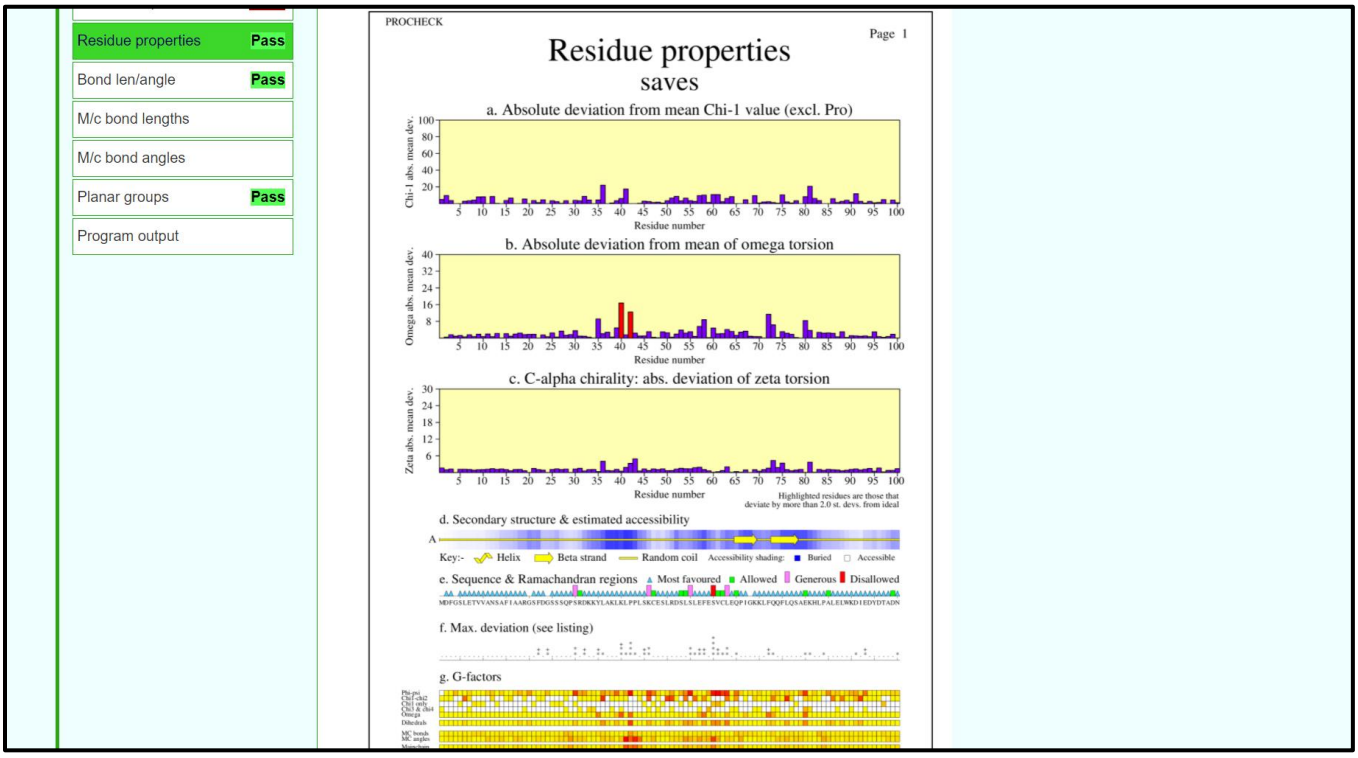


Fig10. Result page for residue properties

RESULT:

The structure predicted for enzyme kinase by homology modelling using modeller was validated using SAVES server.

CONCLUSION:

SAVES in an integrated server containing various tools on a single platform that can be used for tertiary structure validation. The predicted structure for rhodopsin by modeller failed the validation thus, I-TASSER based on threading approach will be used to predict a better structure and will be validated again using SAVES server.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 220-222.
2. SAVESv6.0 - Structure Validation Server. (n.d.). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/
3. SAVESv6.0 - Structure Validation Server. (n.d.). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/?job=924086

DATE: 10-03-22

WEBLEM 4b
SAVES server
(URL:https://saves.mbi.ucla.edu/)

AIM:

To validate structure model1 generated from I-TASSER server.

INTRODUCTION:

Model1 is the structure predicted using threading approach using I-TASSER. The structure has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules. This can be done using SAVES server.

SAVES is a structure validation server that has various tools like Errat, Verify3D, Prove, Whatcheck, Procheck.and Cryst integrated in one single platform. This involves checking anomalies in φ–ψ angles, bond lengths, close contacts, and so on. Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account. This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures. By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

METHODOLOGY:

1. Open homepage for SAVES server. (URL: https://saves.mbi.ucla.edu/)
2. Upload structure retrieved from I-TASSER in PDB format.
3. Obtain results for Errat, Verify3D, Prove, Whatcheck and Procheck.
4. Observe and interpret the results.

OBERSERVATION:

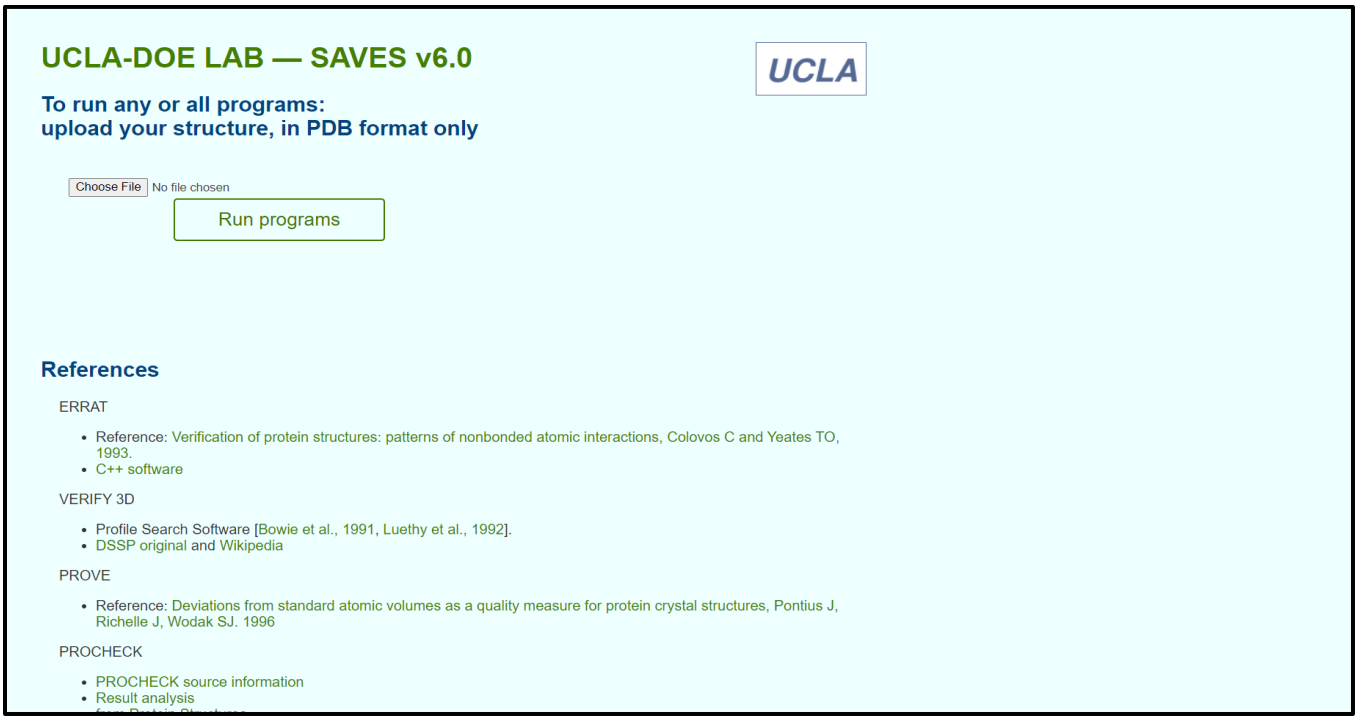


Fig1. Homepage for SAVES server

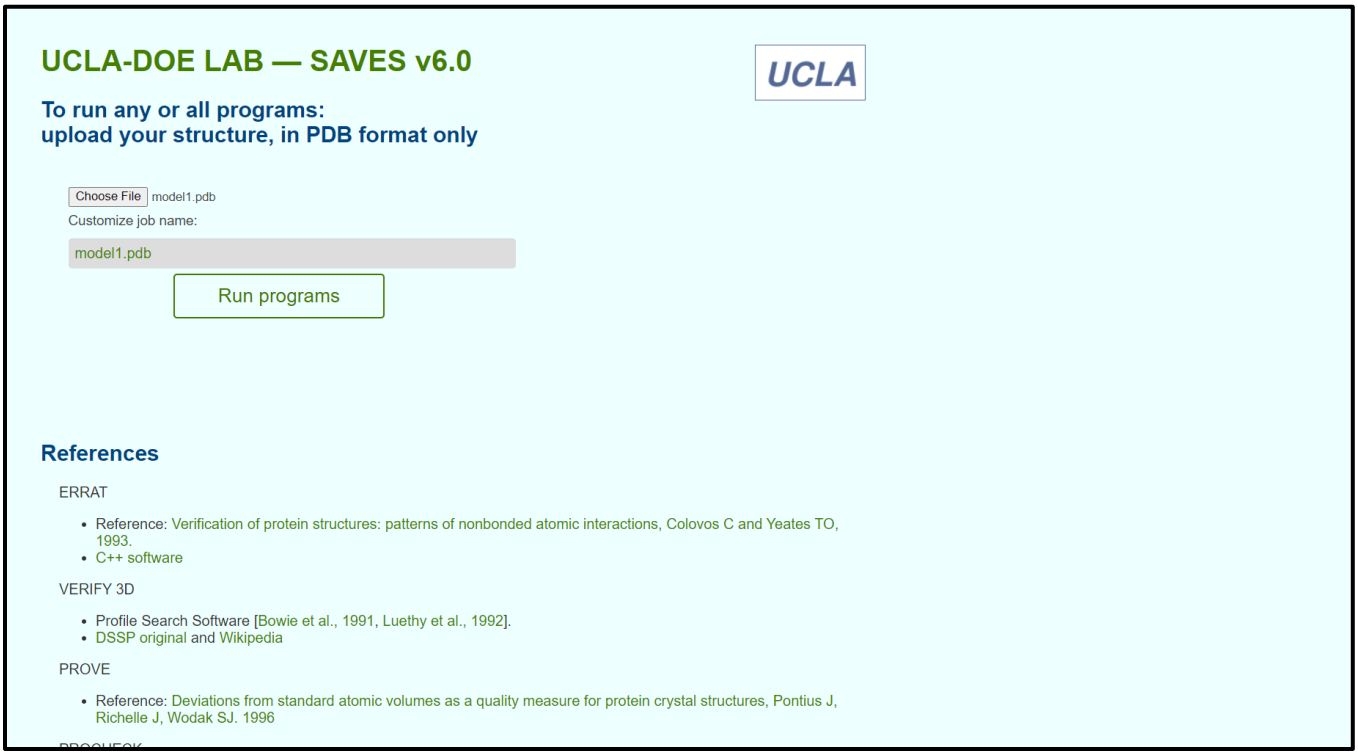


Fig2. Structure from Modeller for validation

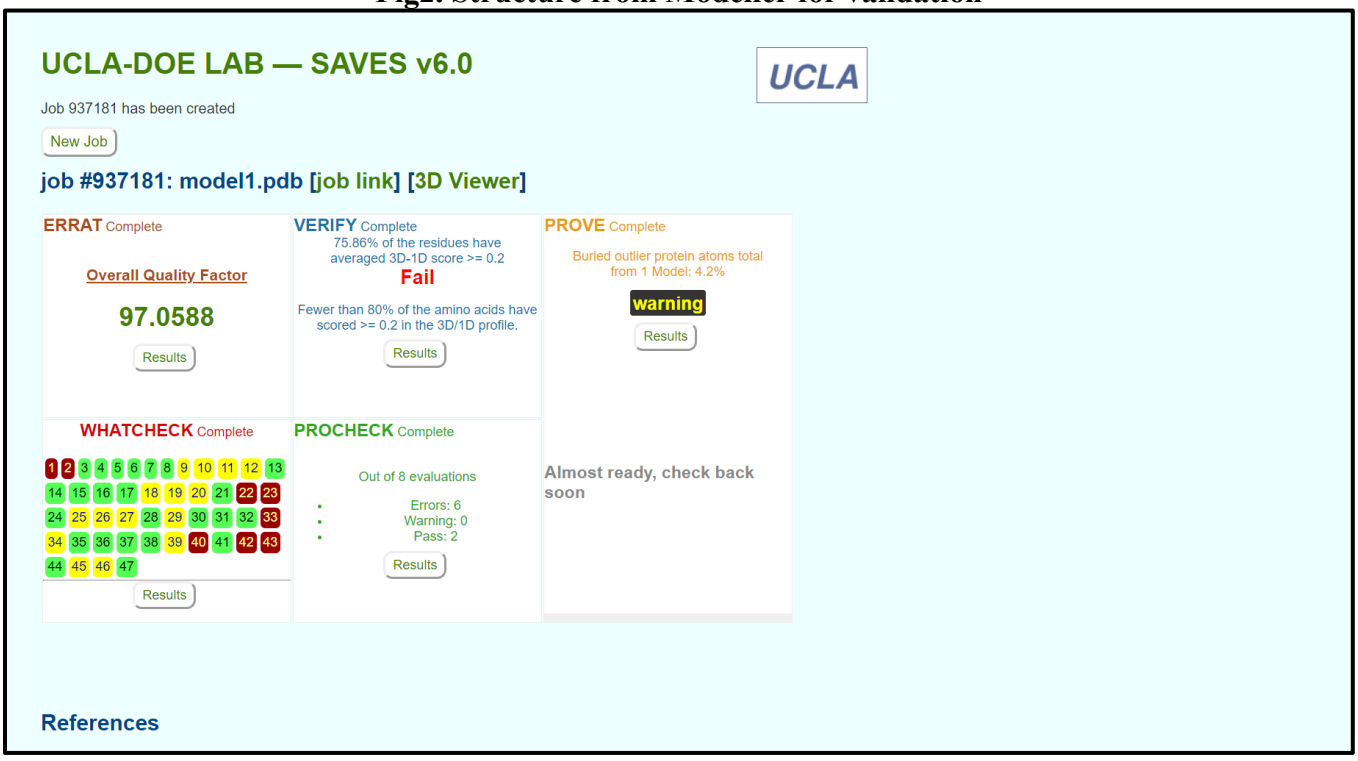


Fig3. Result page for structure validation for various servers

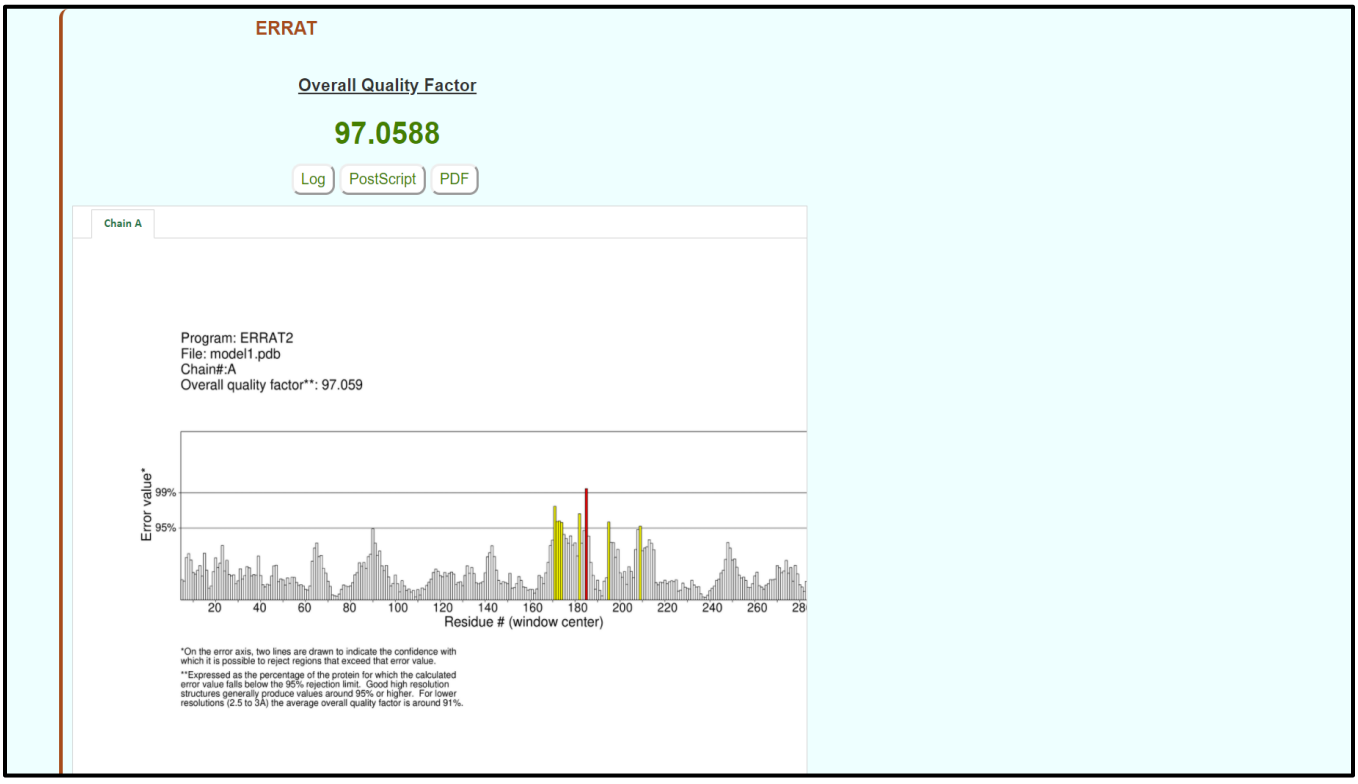


Fig4. Result page for Errat

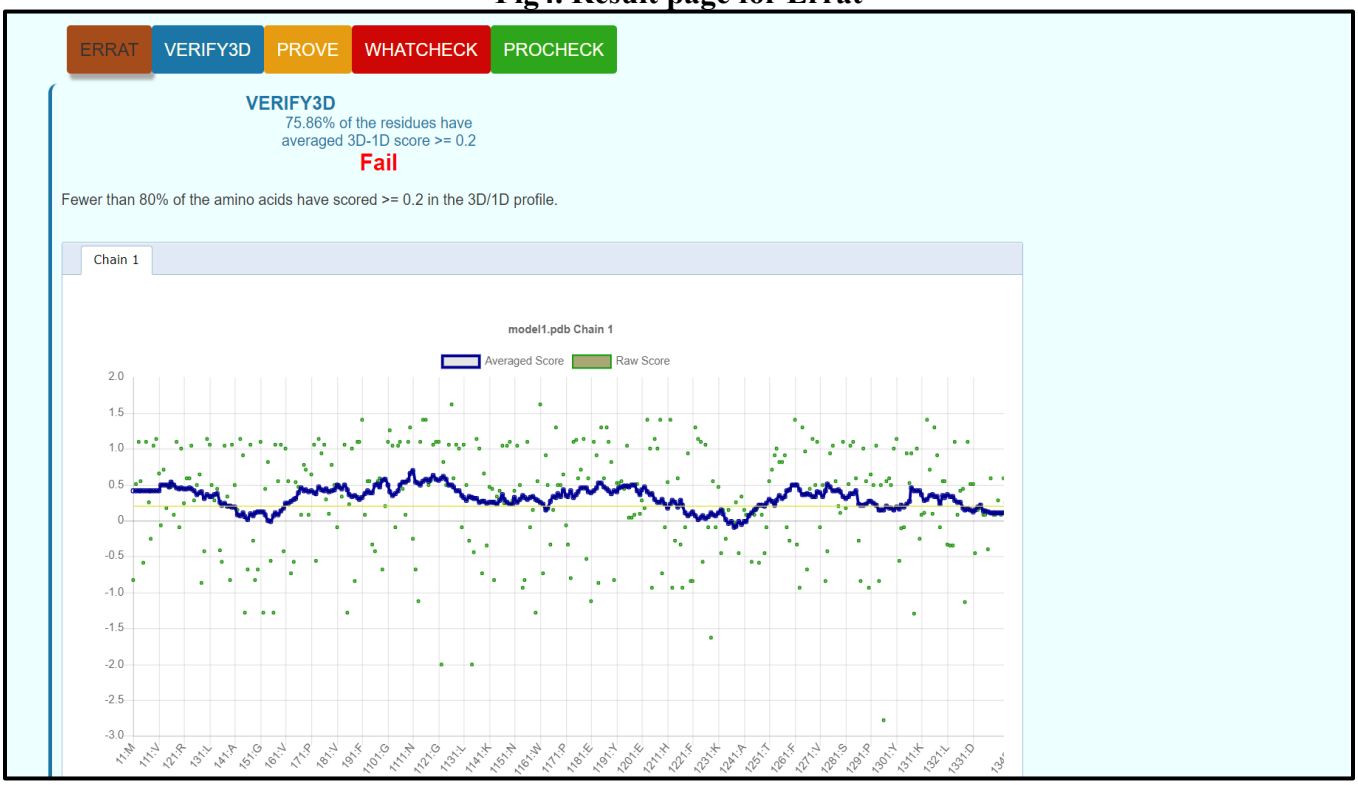


Fig5. Result page for Verify3D

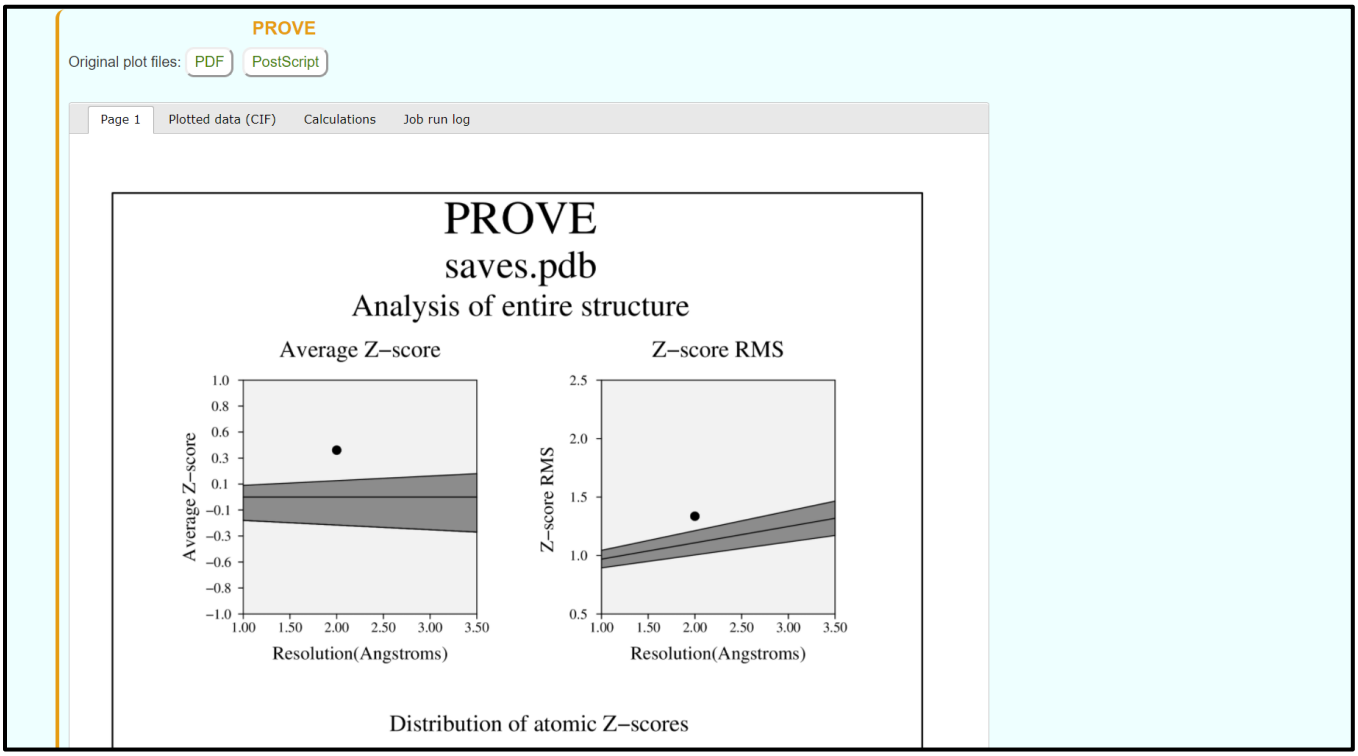


Fig6. Result page for Prove

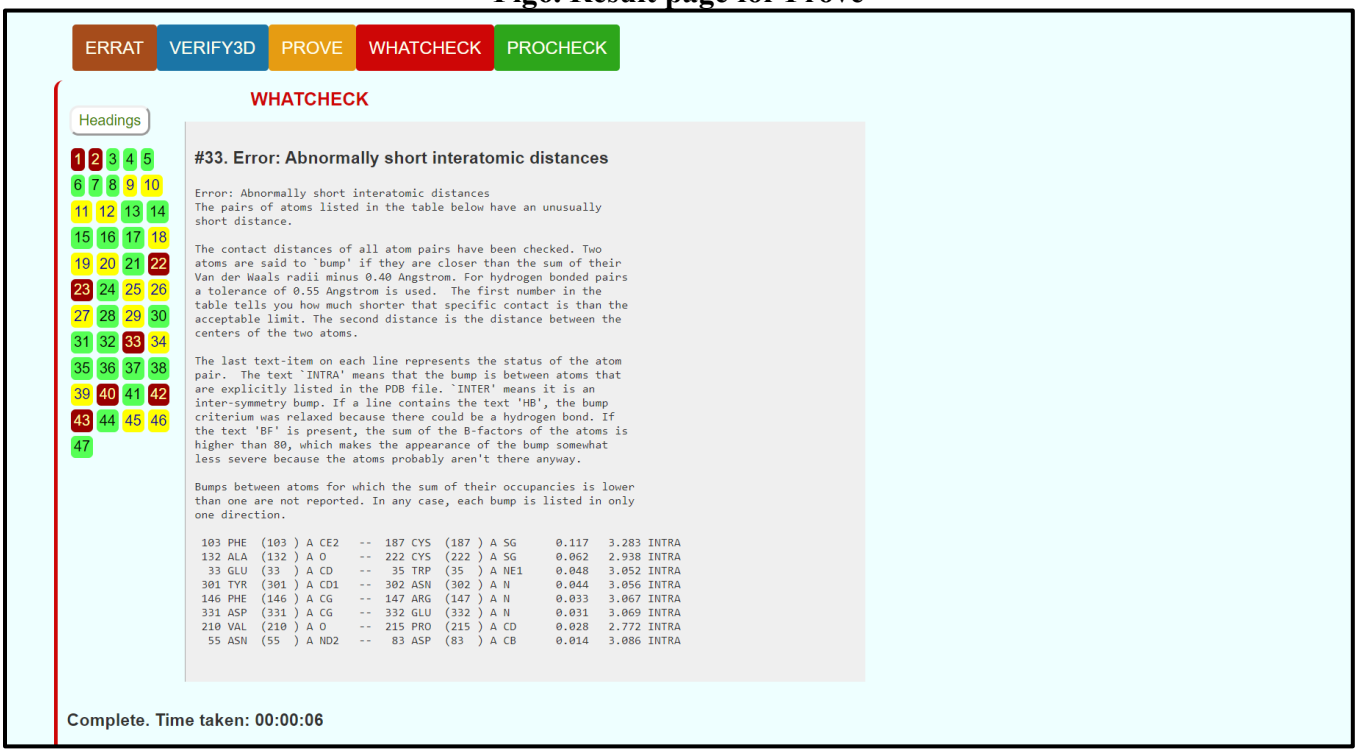


Fig7. Result page for Whatcheck

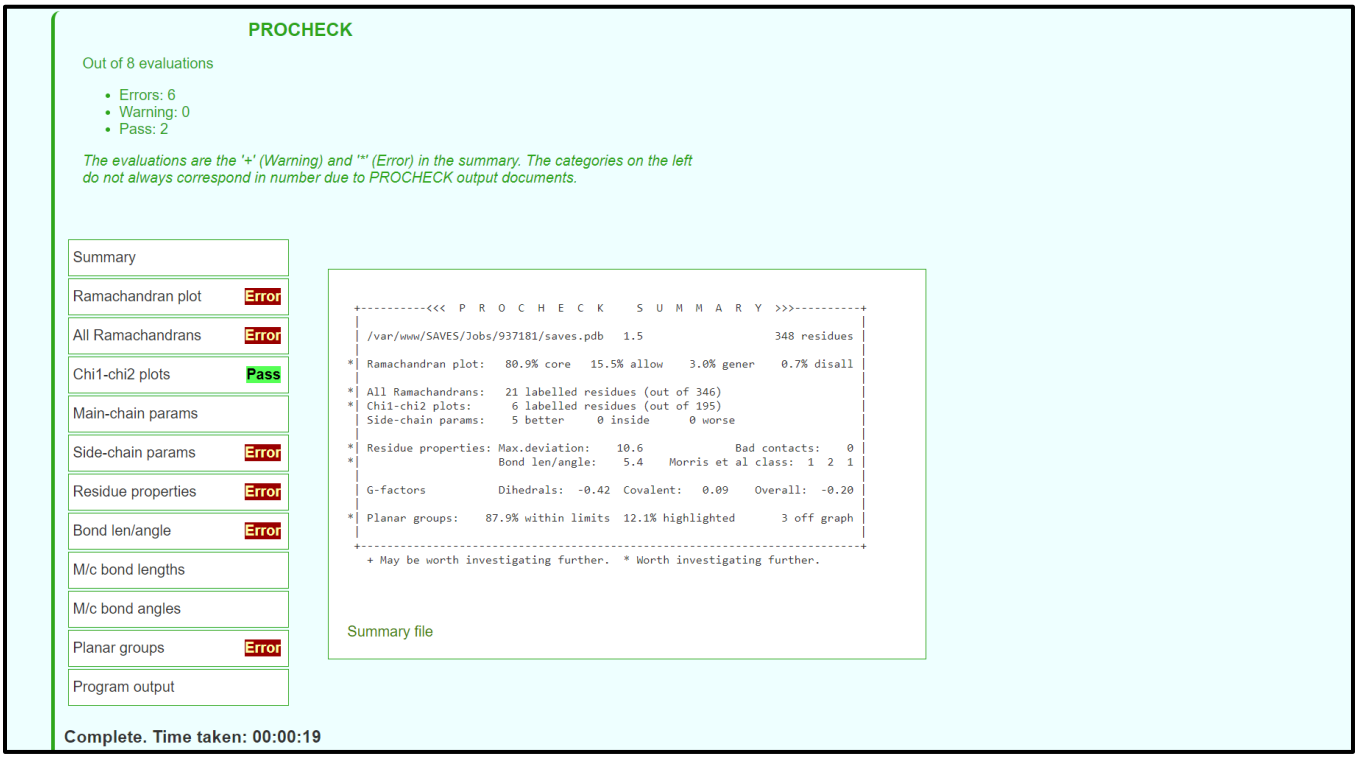


Fig8. Result page for Procheck

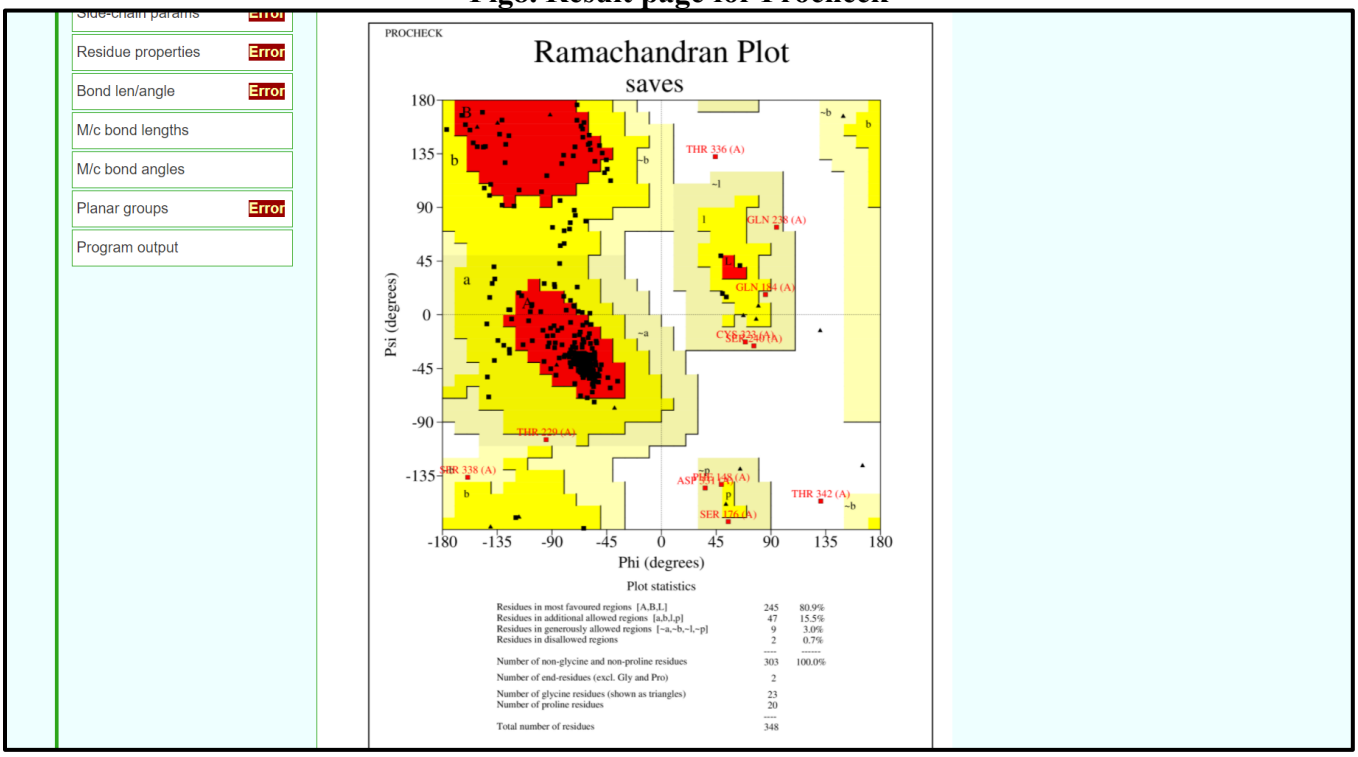


Fig9. Result page for Ramachandran plot

Fig10. Result page for reside properties

RESULT:

The structure predicted for enzyme kinase by threading approach using I-TASSER was validated using SAVES server.

CONCLUSION:

SAVES in an integrated server containing various tools on a single platform that can be used for tertiary structure validation. The predicted structure for kinase by I-TASSER passed only for ERRAT and did not give required results for the rest. Even though I-TASSER gave better predicted structure than modeller, Robetta based on ab-initio approach will be used to predict a better structure and will be validated again using SAVES server.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 220-222.
2. SAVESv6.0 - Structure Validation Server. (n.d.). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/
3. SAVESv6.0 - Structure Validation Server. (n.d.). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/?job=924117

DATE: 10-03-22

WEBLEM 4c
SAVES server
(URL: https://saves.mbi.ucla.edu/)

AIM:

To validate structure 240040 generated from Robetta server.

INTRODUCTION:

240040 is the structure predicted using Ab-initio approach using Robetta. The structure has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules. This can be done using SAVES server.

SAVES is a structure validation server that has various tools like Errat, Verify3D, Prove, Whatcheck, Procheck.and Cryst integrated in one single platform. This involves checking anomalies in φ–ψ angles, bond lengths, close contacts, and so on. Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account. This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures. By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

METHODOLOGY:

1. Open homepage for SAVES server. (URL: https://saves.mbi.ucla.edu/)
2. Upload structure retrieved from Robetta in PDB format.
3. Obtain results for Errat, Verify3D, Prove, Whatcheck and Procheck.
4. Observe and interpret the results.

OBSERVATION:

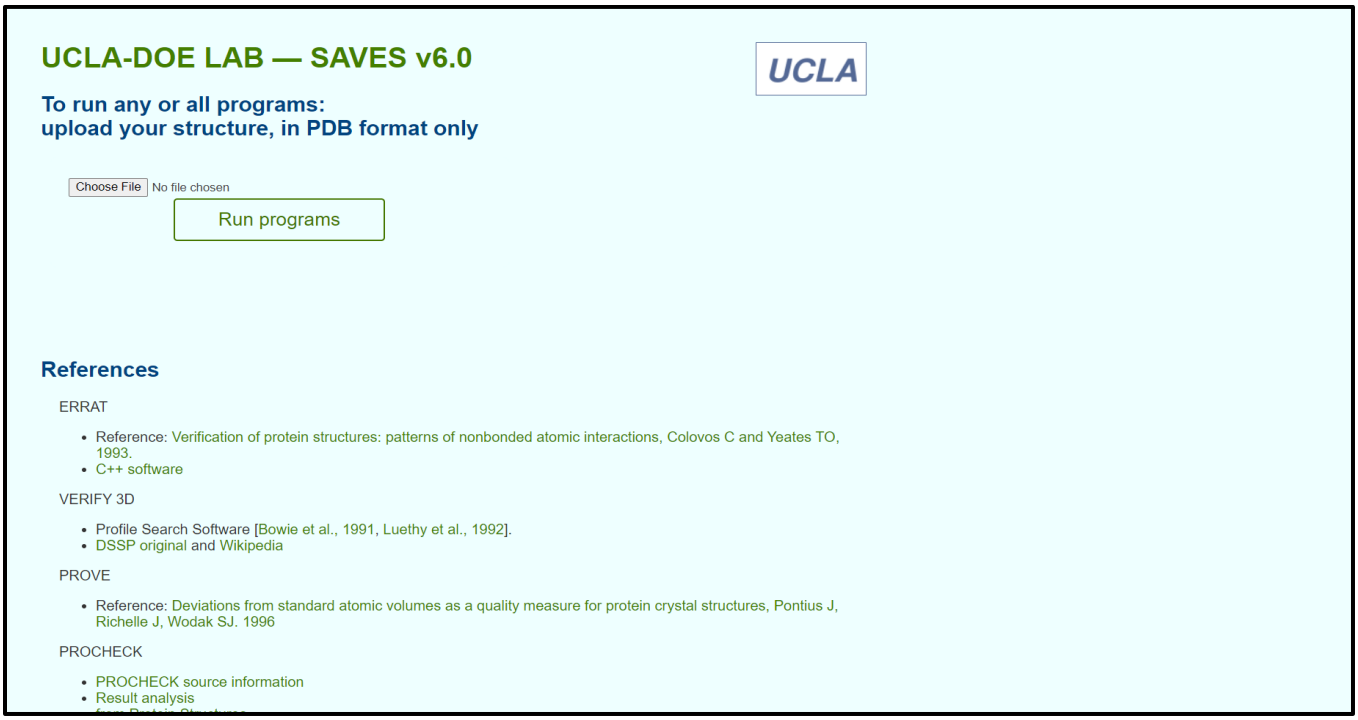


Fig1. Homepage for SAVES server

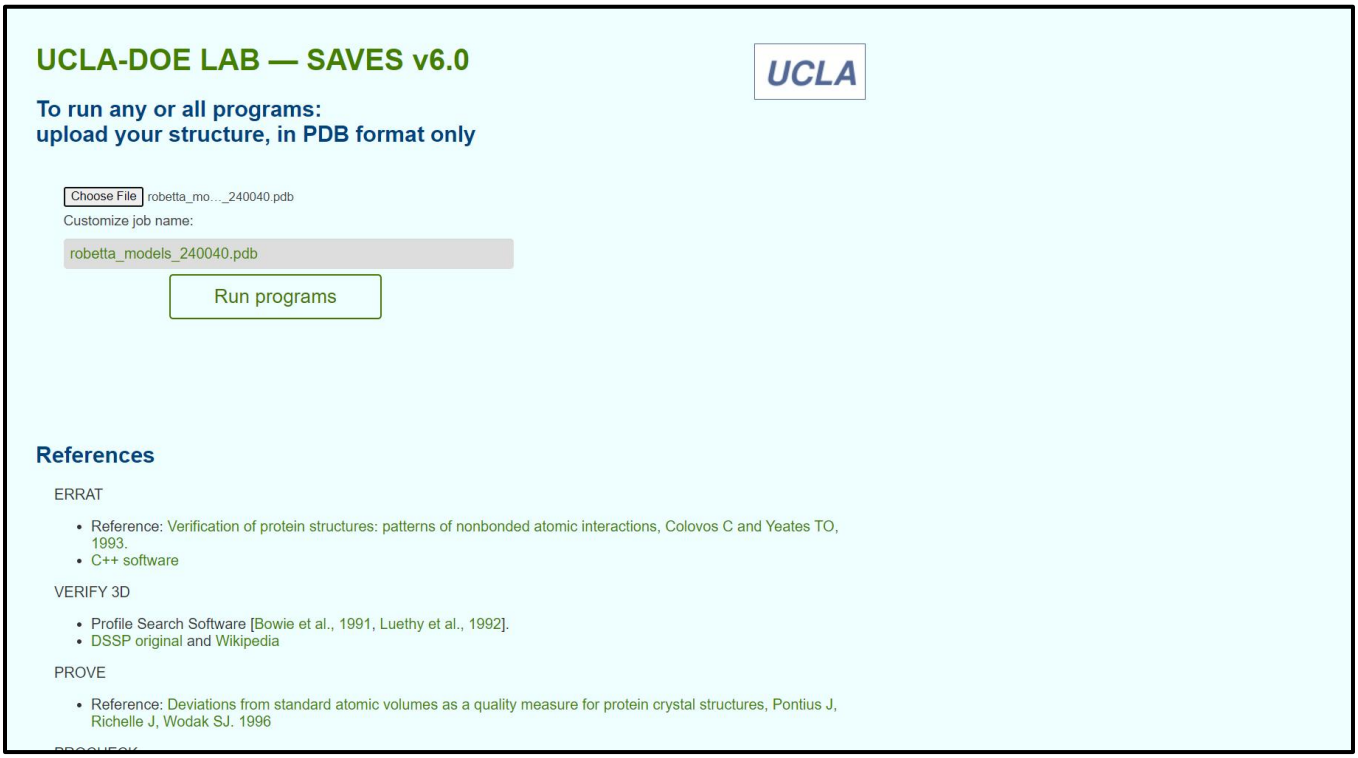


Fig2. Structure from Modeller for validation

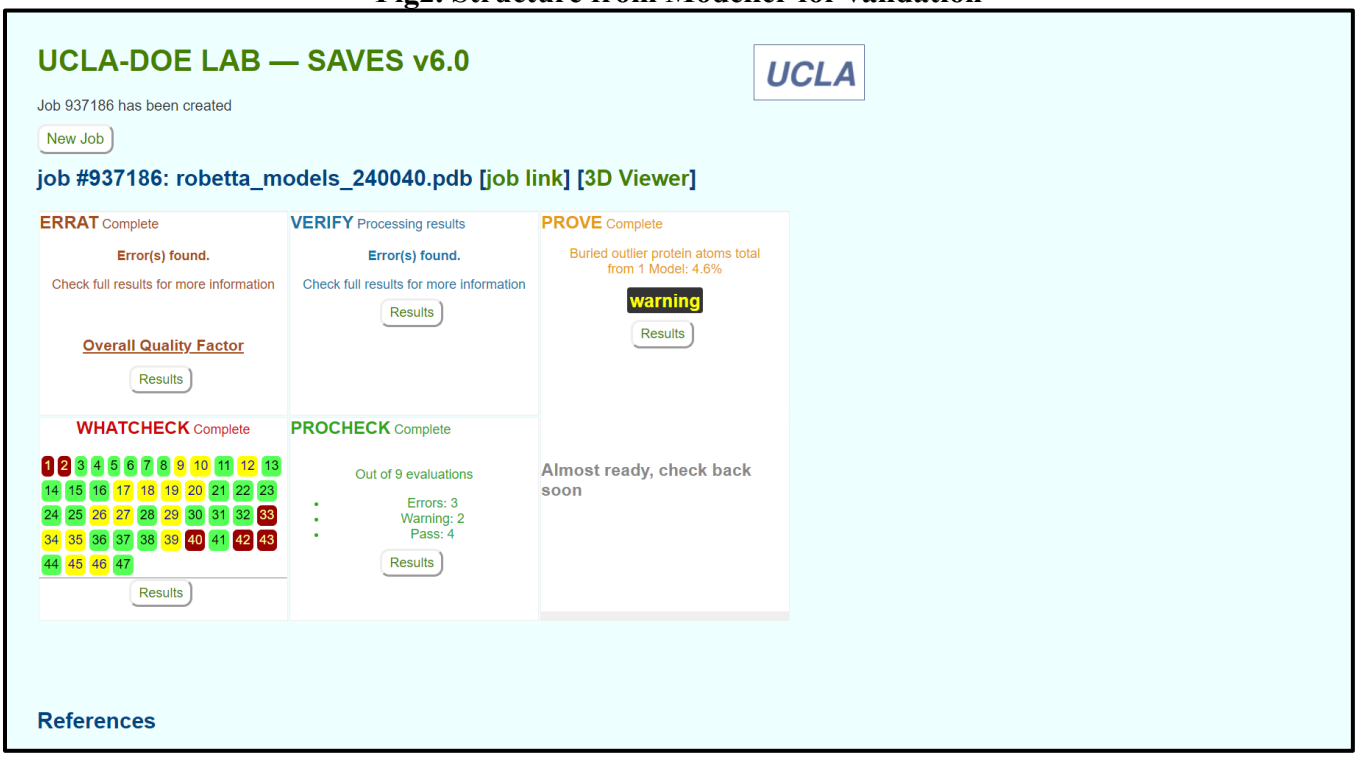


Fig3. Result page for structure validation for various servers

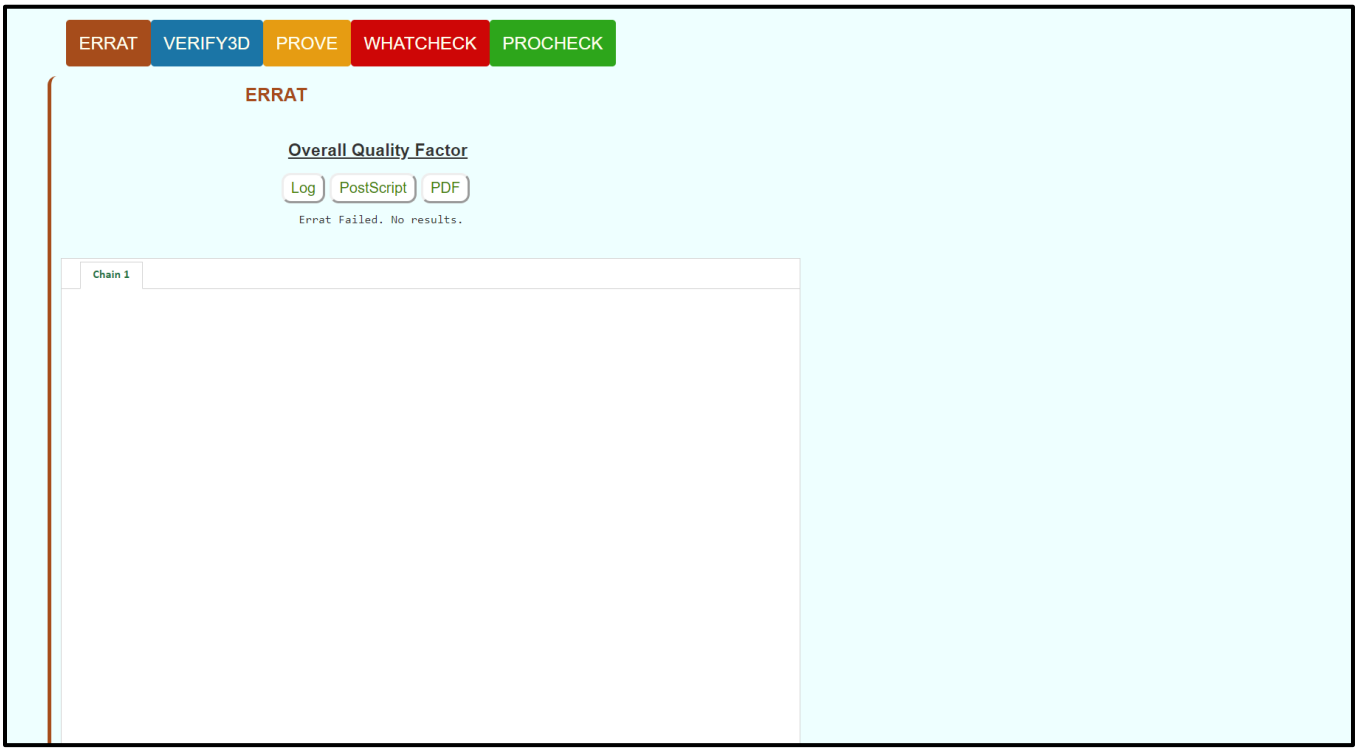


Fig4. Result page for Errat

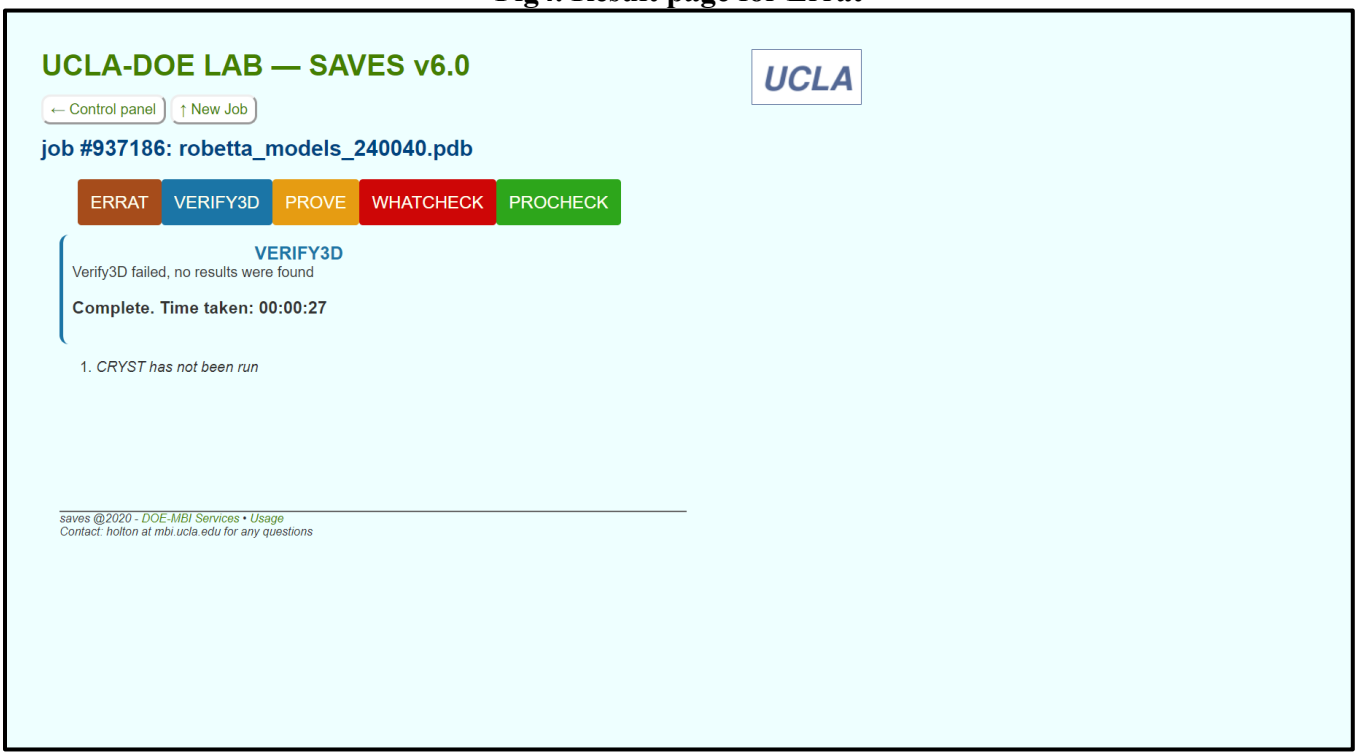


Fig5. Result page for Verify3D

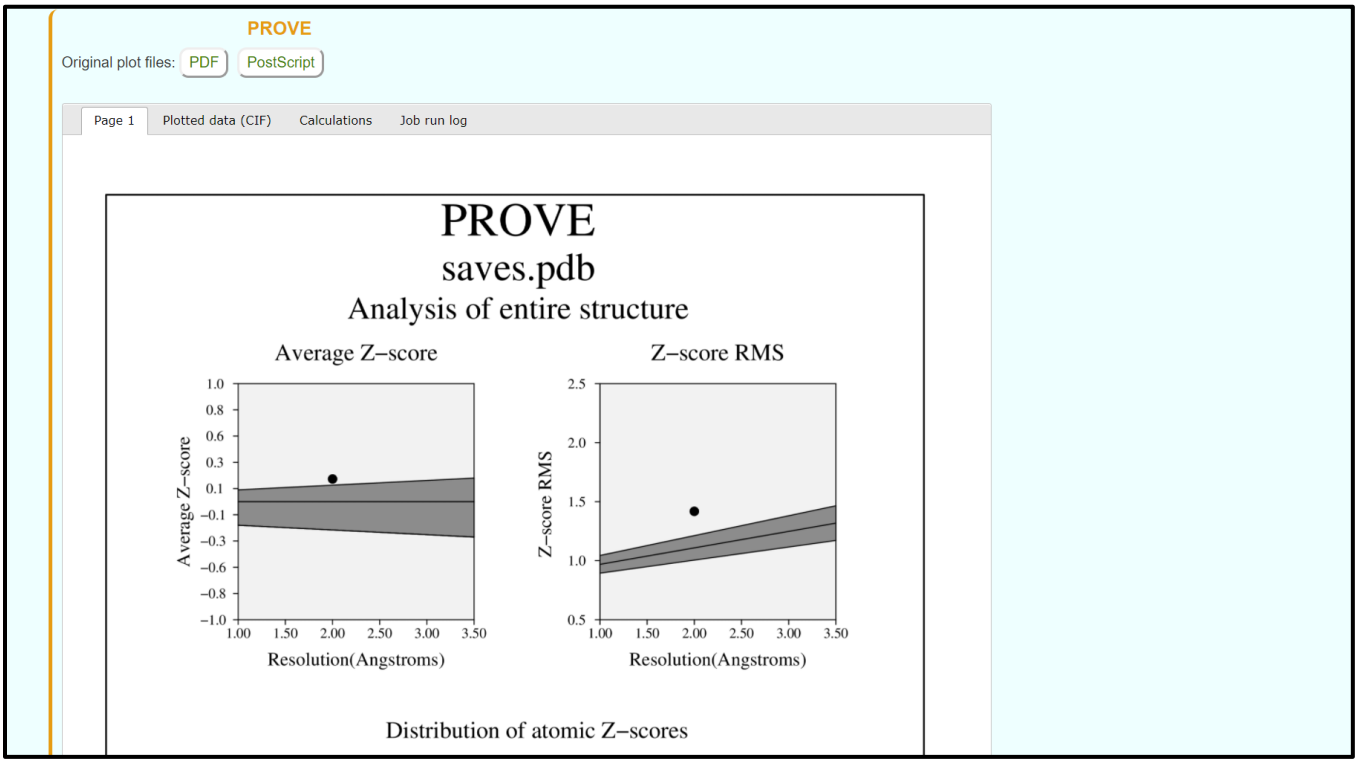


Fig6. Result page for PROVE

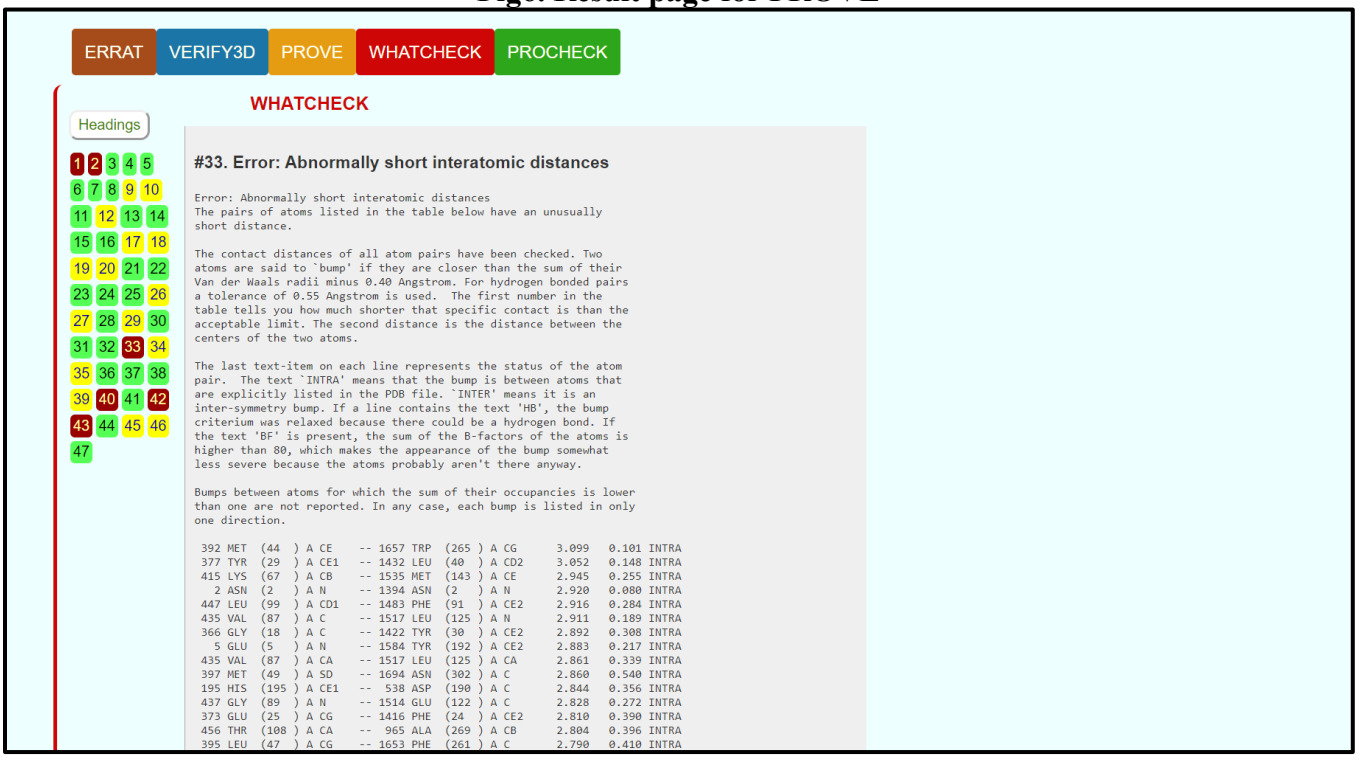


Fig7. Result page for Whatcheck

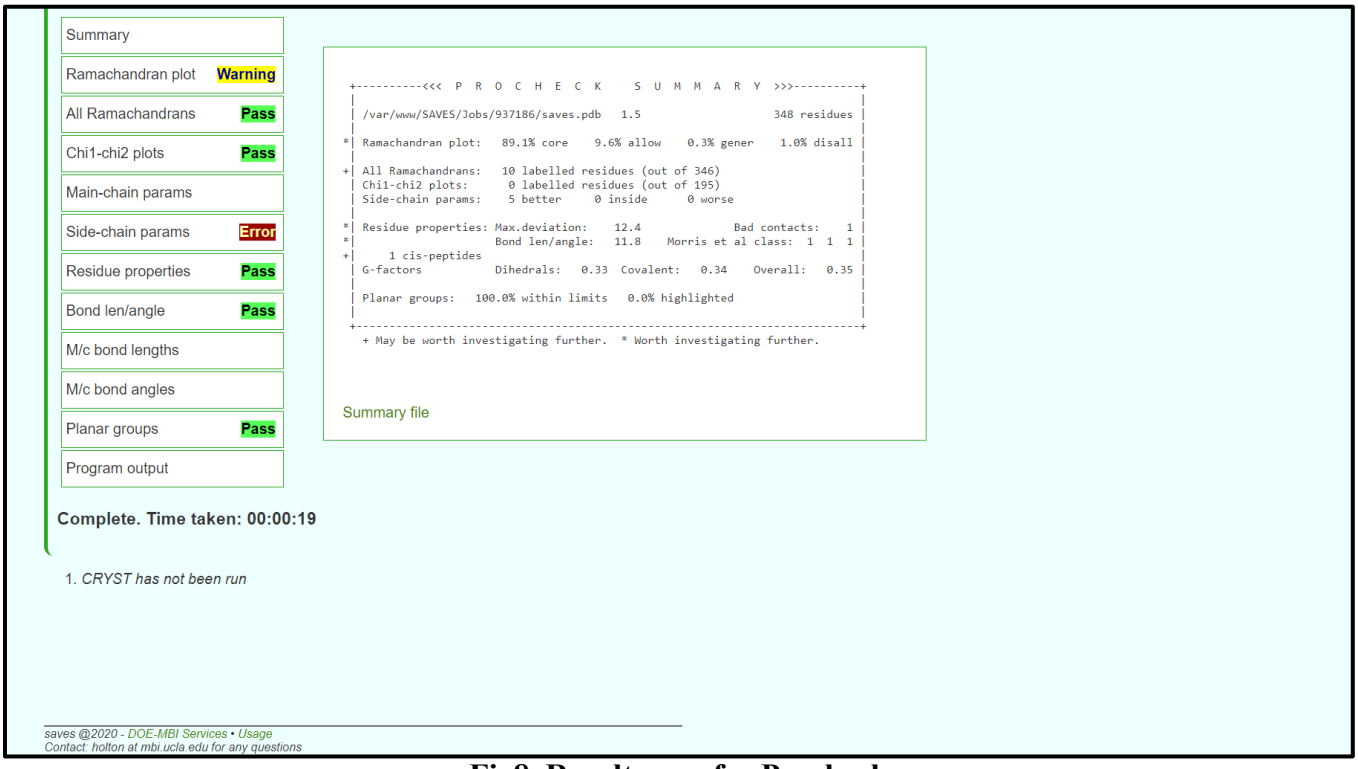


Fig8. Result page for Procheck

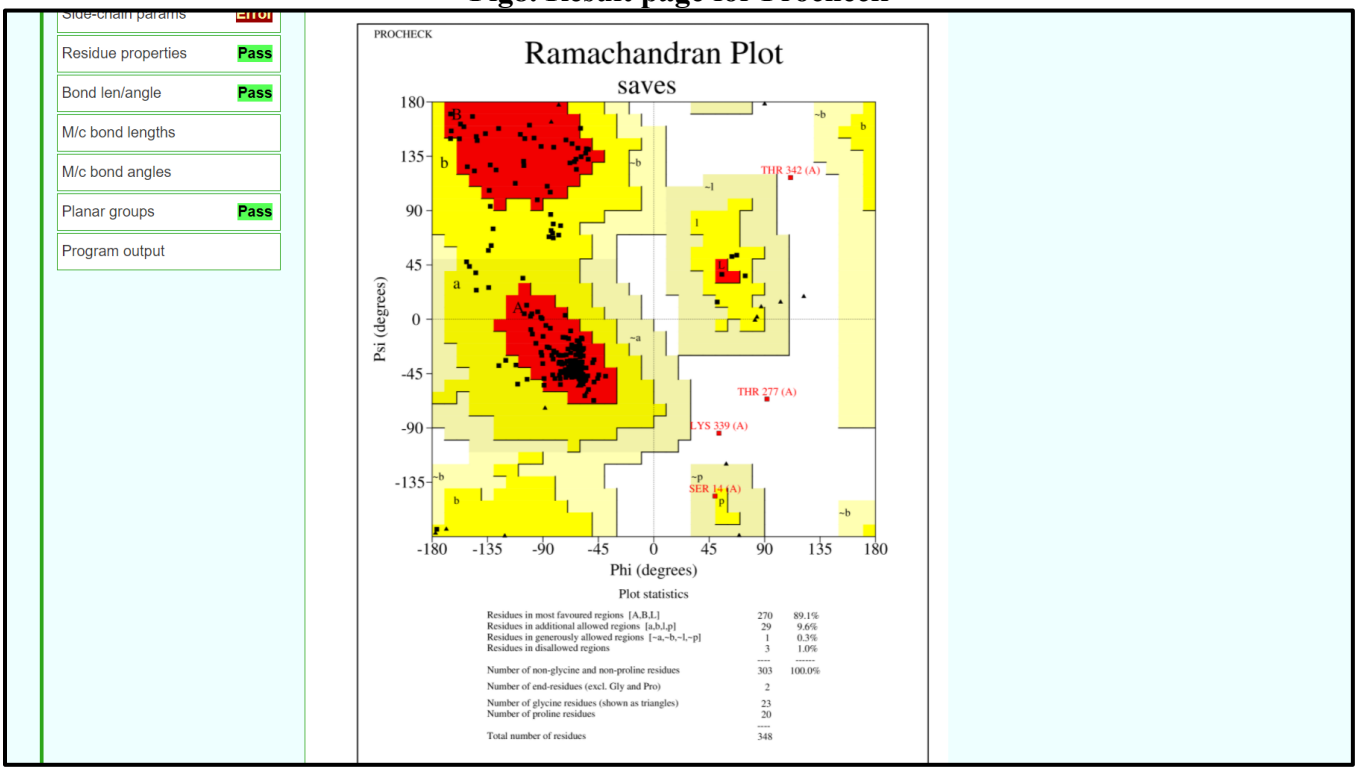


Fig9. Result page for Ramachandran plot

Fig10. Result page for residue properties

RESULT:

The structure predicted for enzyme kinase by abi-initio approach using Robetta was validated using SAVES server.

CONCLUSION:

SAVES in an integrated server containing various tools on a single platform that can be used for tertiary structure validation. The predicted structure for Rhodopsin by Robetta passed maximum requirements of validation. Hence, it can be concluded that the structure predicting by Robetta was the most accurate out of all three methods used for prediction.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 220-222.
2. SAVESv6.0 - Structure Validation Server. (n.d.). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/
3. SAVESv6.0 - Structure Validation Server. (n.d.). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/?job=928054

Date: 05-03-22

WEBLEM 5

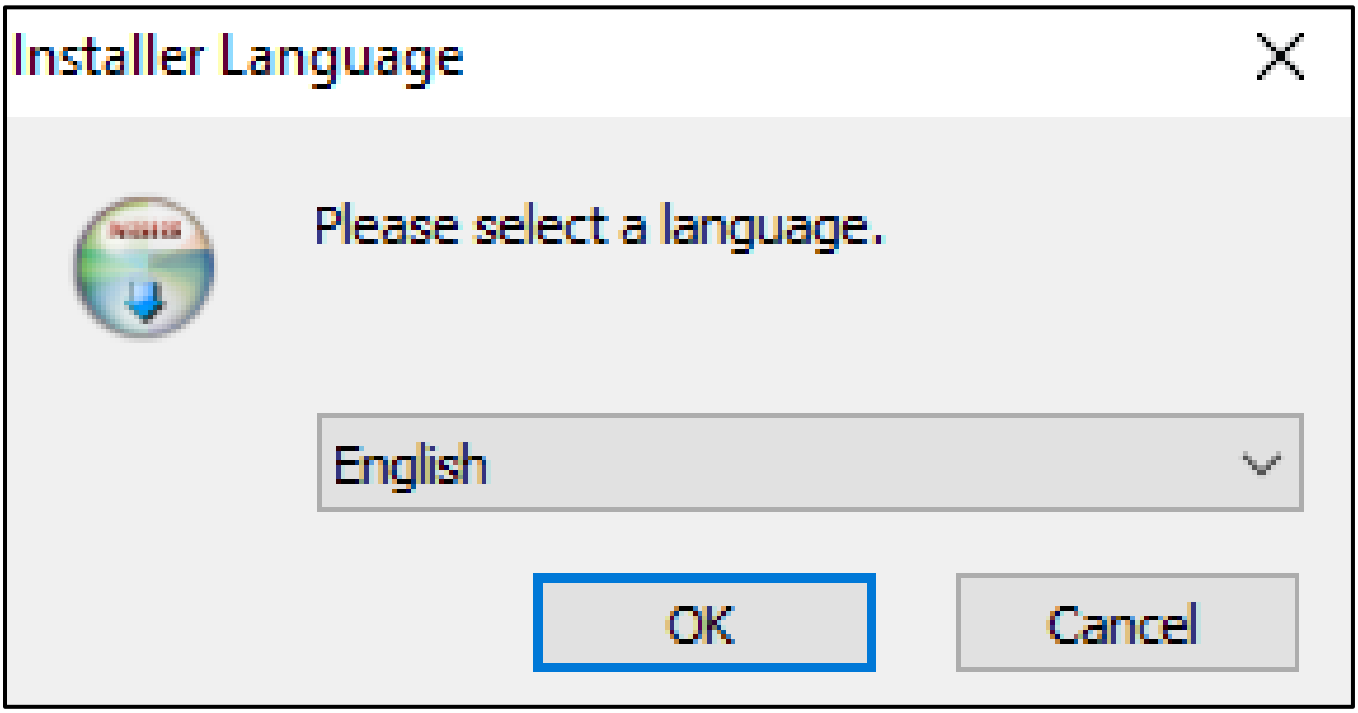
Introduction to Visualization of Tertiary structure using RASMOL & PyMOL

We can obtain all the available information about the 3D structure of this enzyme by browsing through the links or download the PDB file to a local directory in our computer and work with our preferred molecular modelling and visualization package.

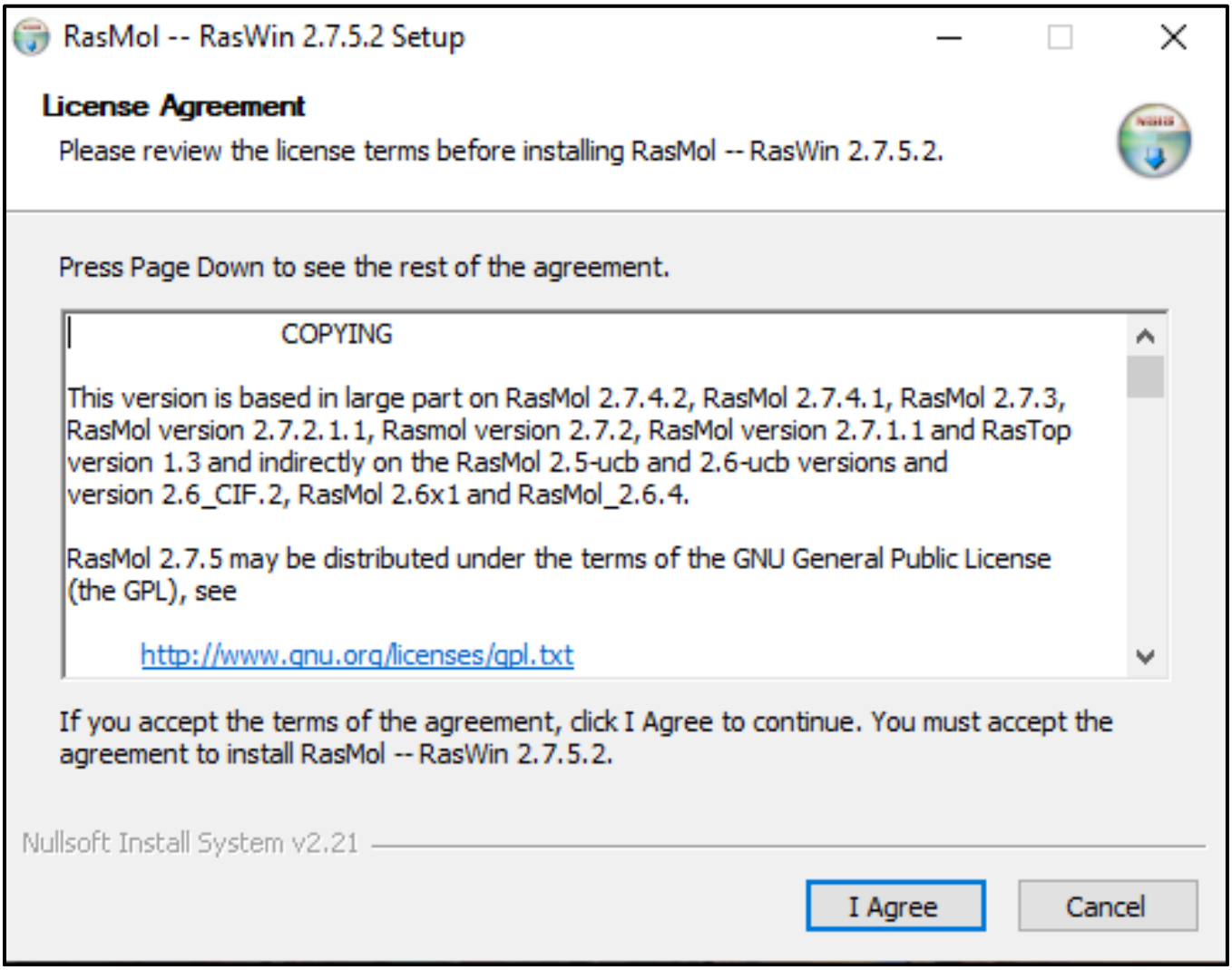
RasMol:

RasMol is a computer program written for molecular graphics visualization intended and used primarily for the depiction and exploration of biological macromolecule structures, such as those found in the Protein Data Bank. It was originally developed by Roger Sayle in the early 90s. Historically, it was an important tool for molecular biologists since the extremely optimized program allowed the software to run on (then) modestly powerful personal computers. Before RasMol, visualization software ran on graphics workstations that, due to their expense, were less accessible to scholars. RasMol has become an important educational tool as well as continuing to be an important tool for research in structural biology.

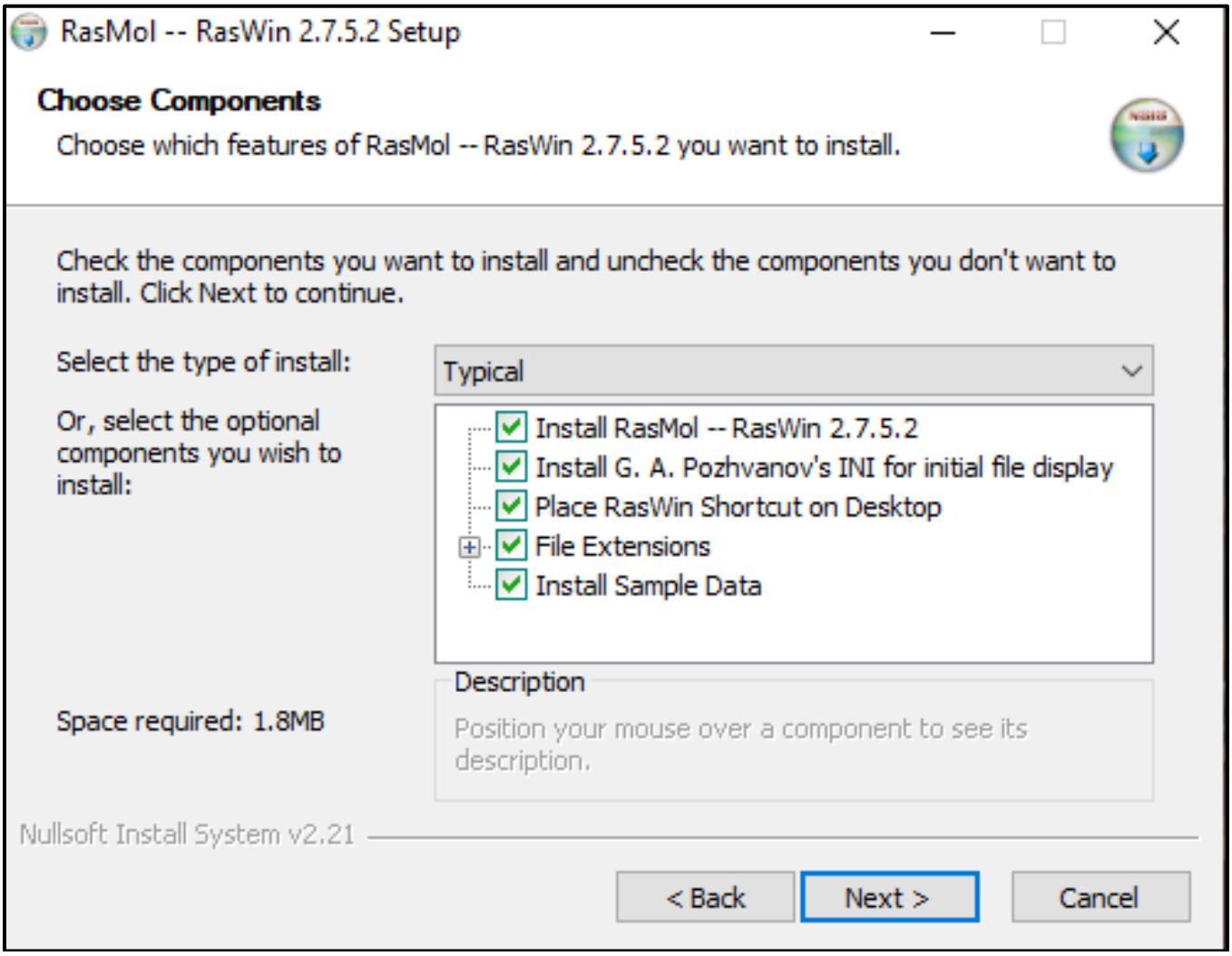
Installation:



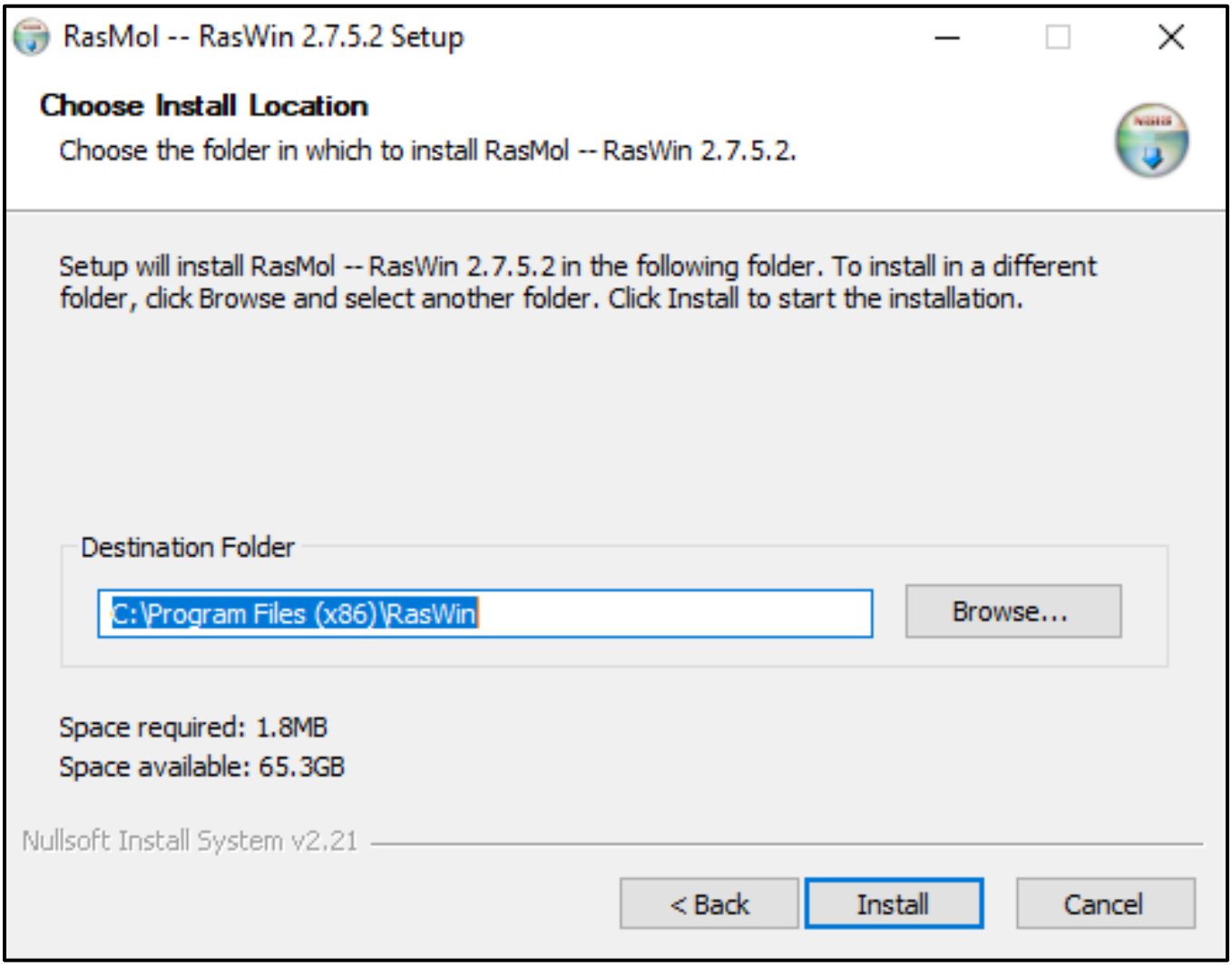
Select language and proceed



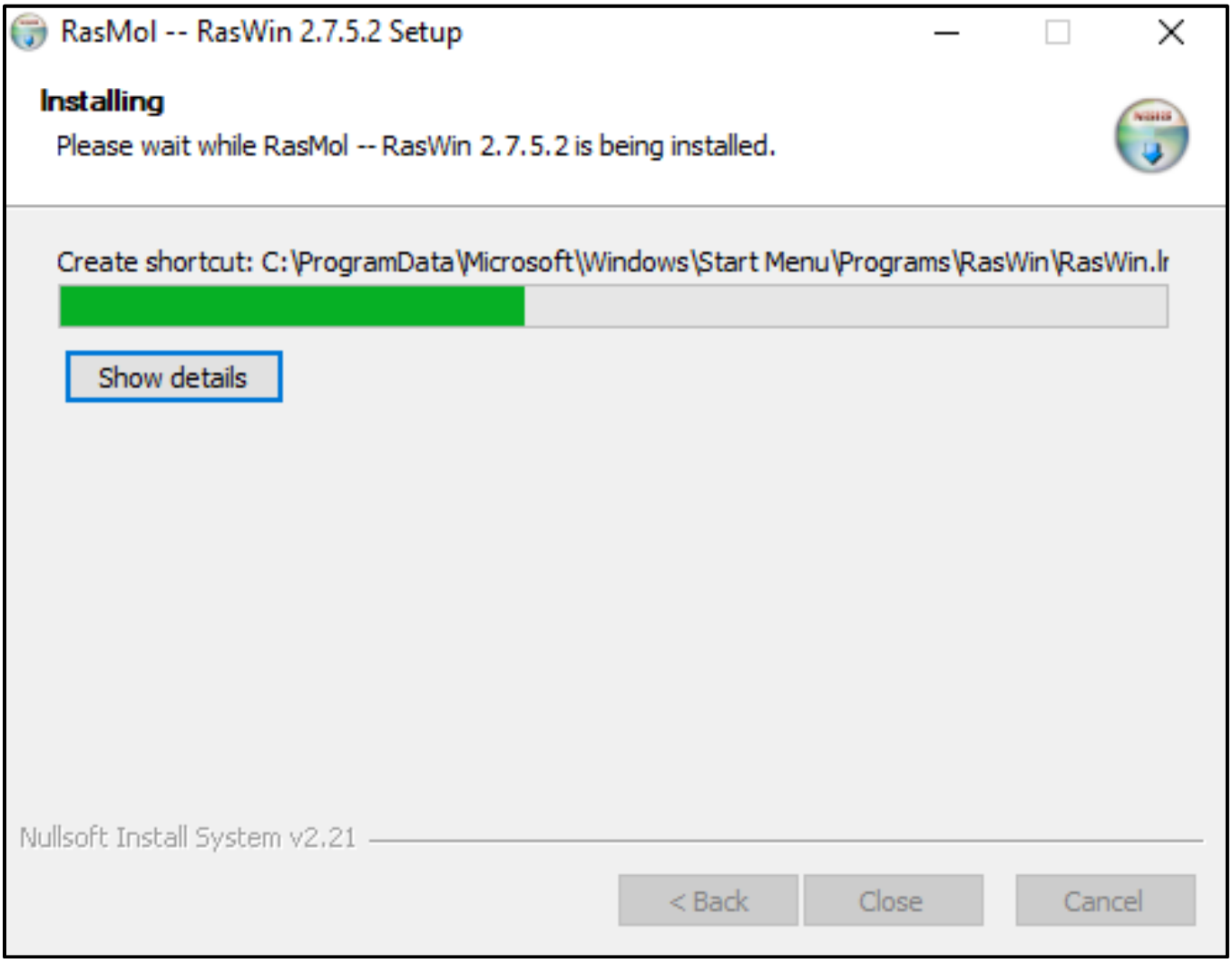
Read the License Agreement and click "I Agree" to proceed



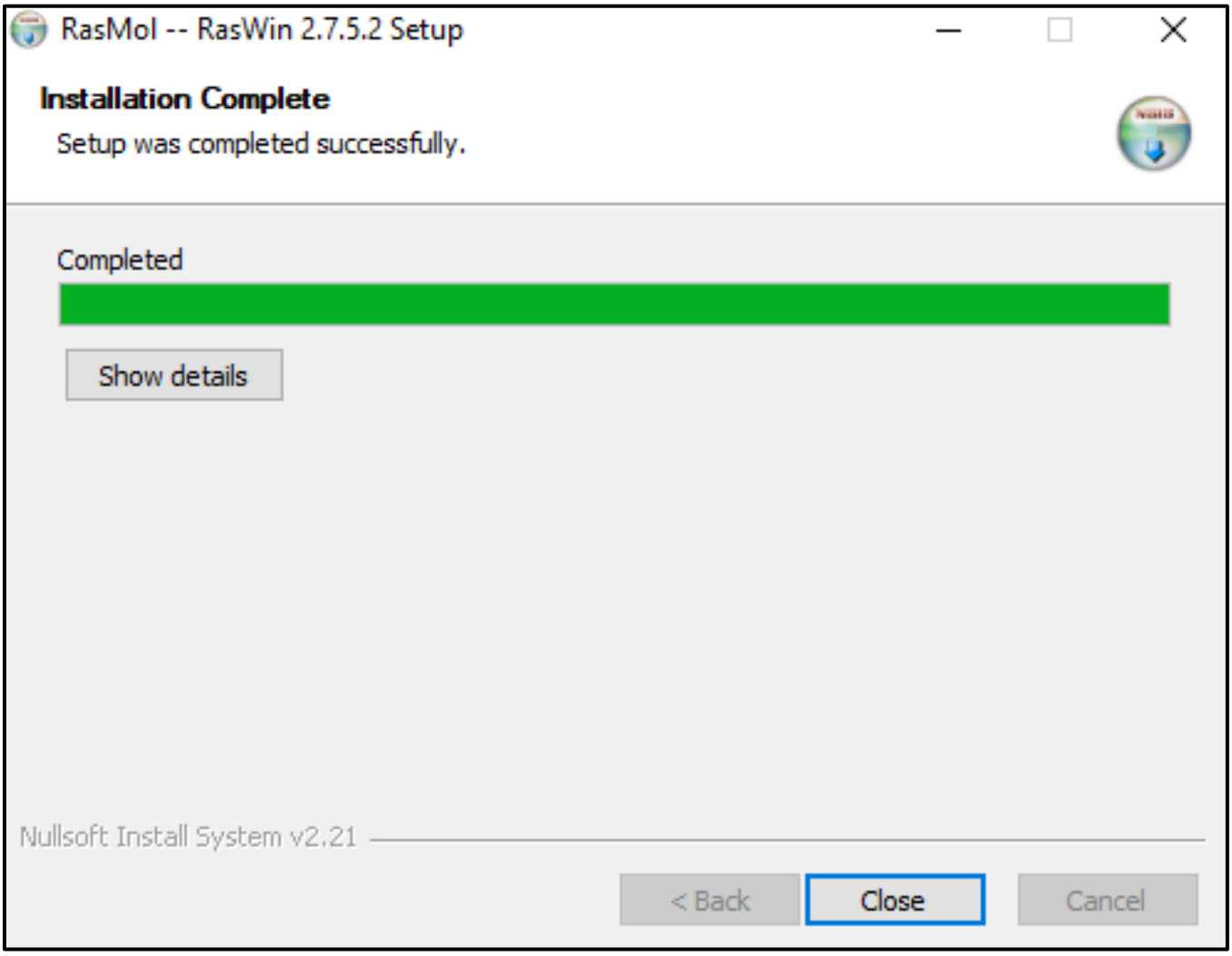
Select the components of RasMol you want to install and click "Next" to proceed



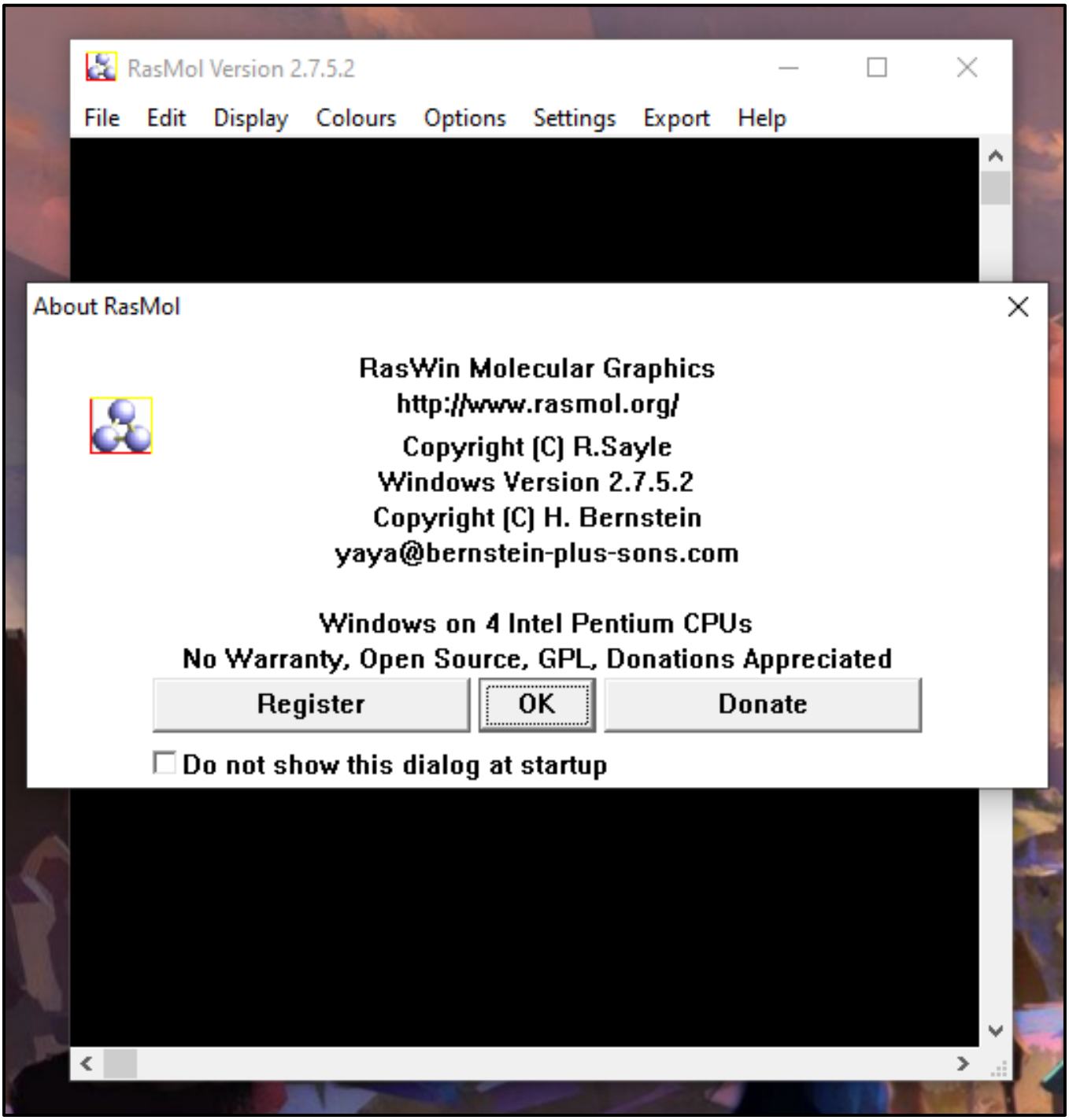
Select the directory in which you want to install RasMol and click "Install" to proceed



The software will now install



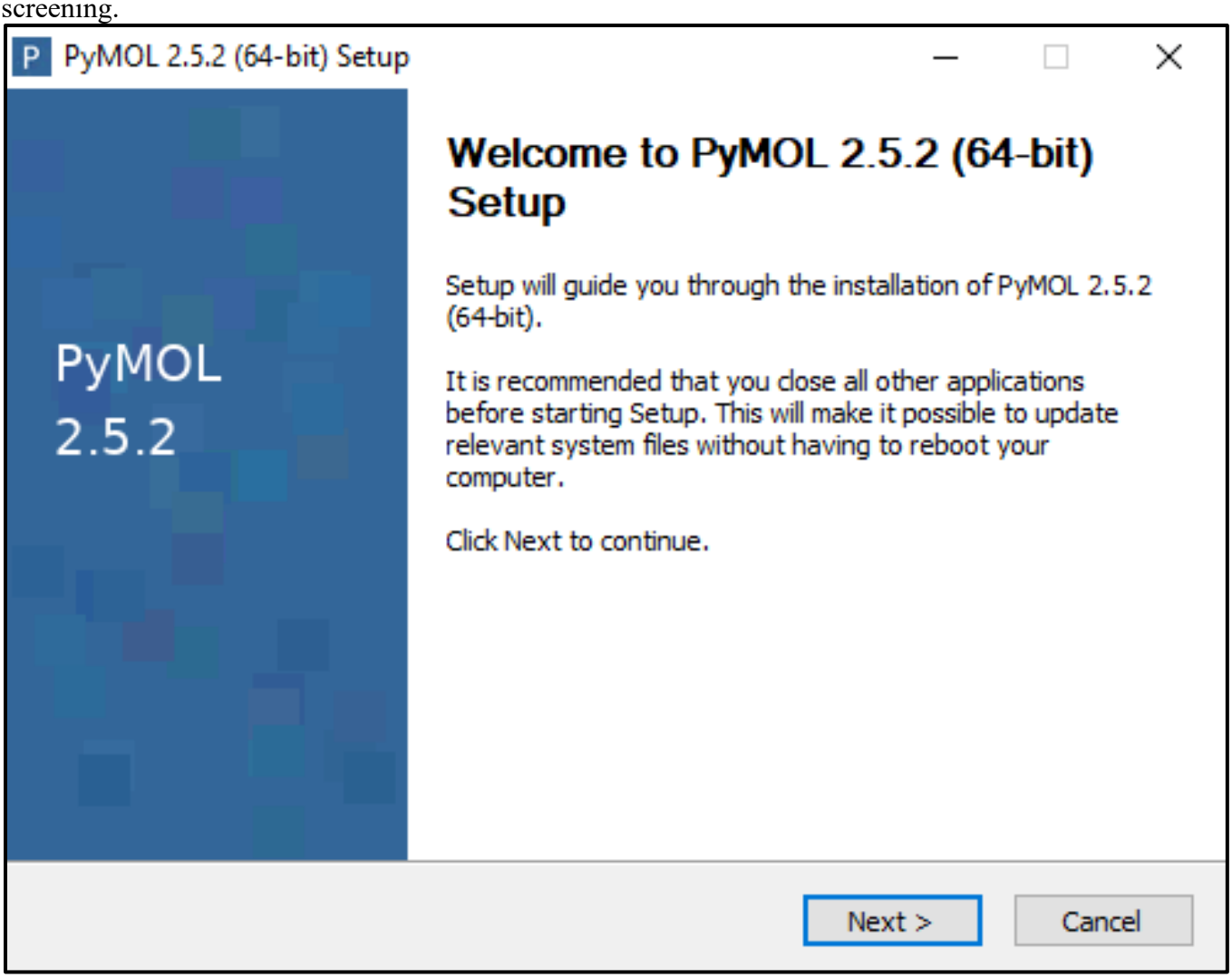
Close the installer after the install is completed



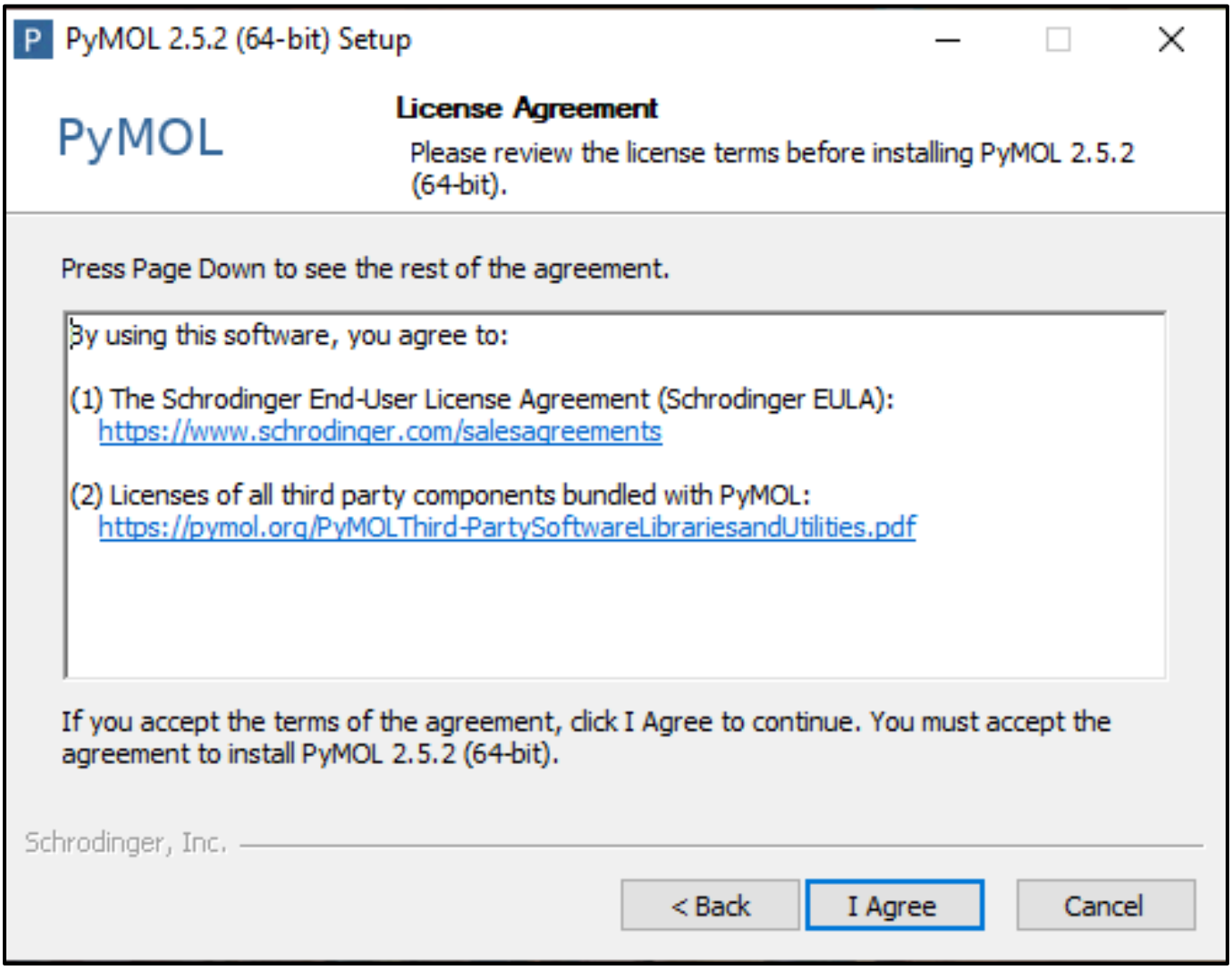
RasMol is now installed

PyMol

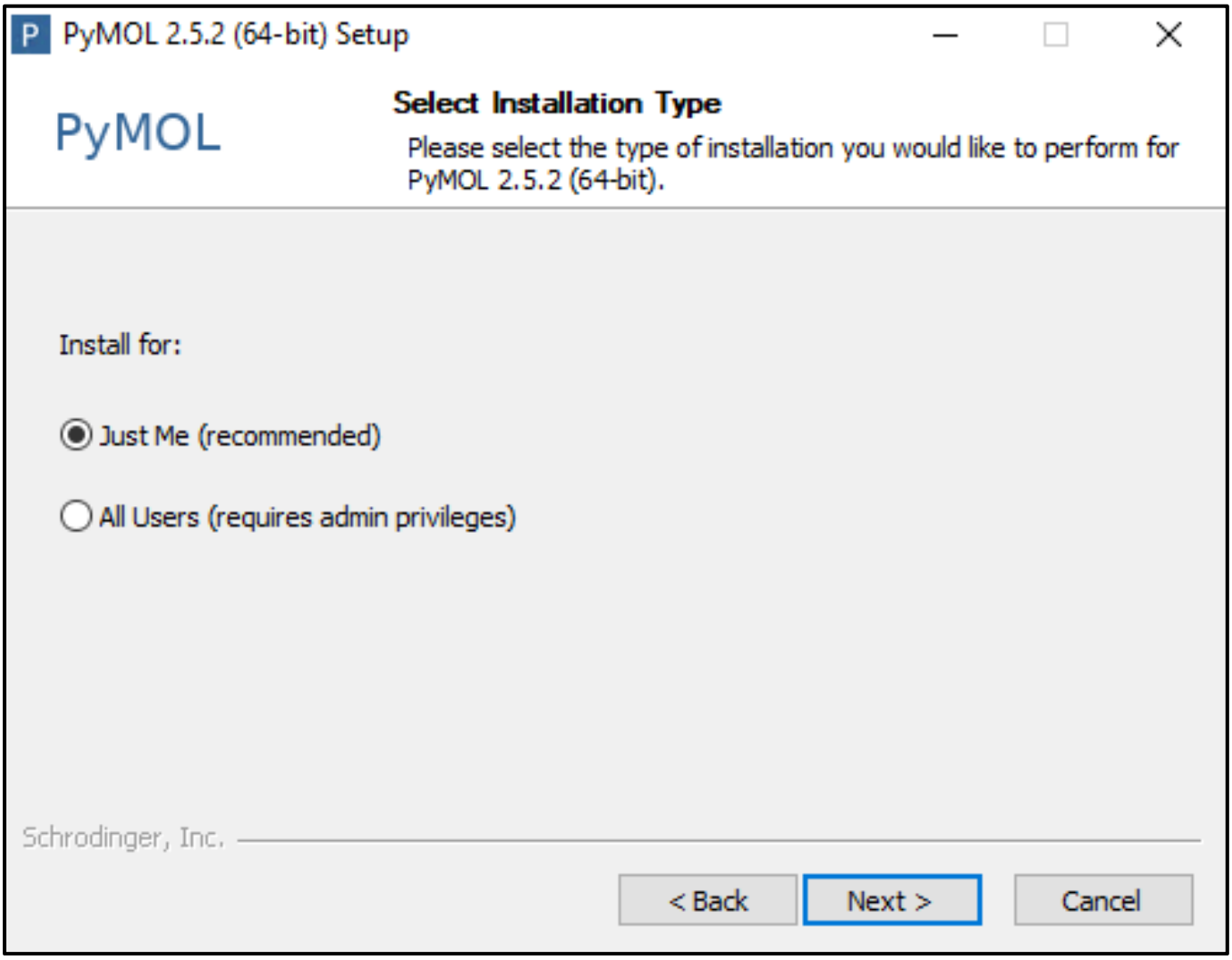
PyMOL, a cross-platform molecular graphics tool, has been widely used for three-dimensional (3D) visualization of proteins, nucleic acids, small molecules, electron densities, surfaces, and trajectories. It is also capable of editing molecules, ray tracing, and making movies. This Python-based software, alongside many Python plugin tools, has been developed to enhance its utilities and facilitate the drug design in PyMOL. To gain an insightful view of useful drug design tools and their functions in PyMOL, we present an extensive discussion on various molecular modeling modules in PyMOL, covering those for visualization and analysis enhancement, protein–ligand modeling, molecular simulations, and drug screening.



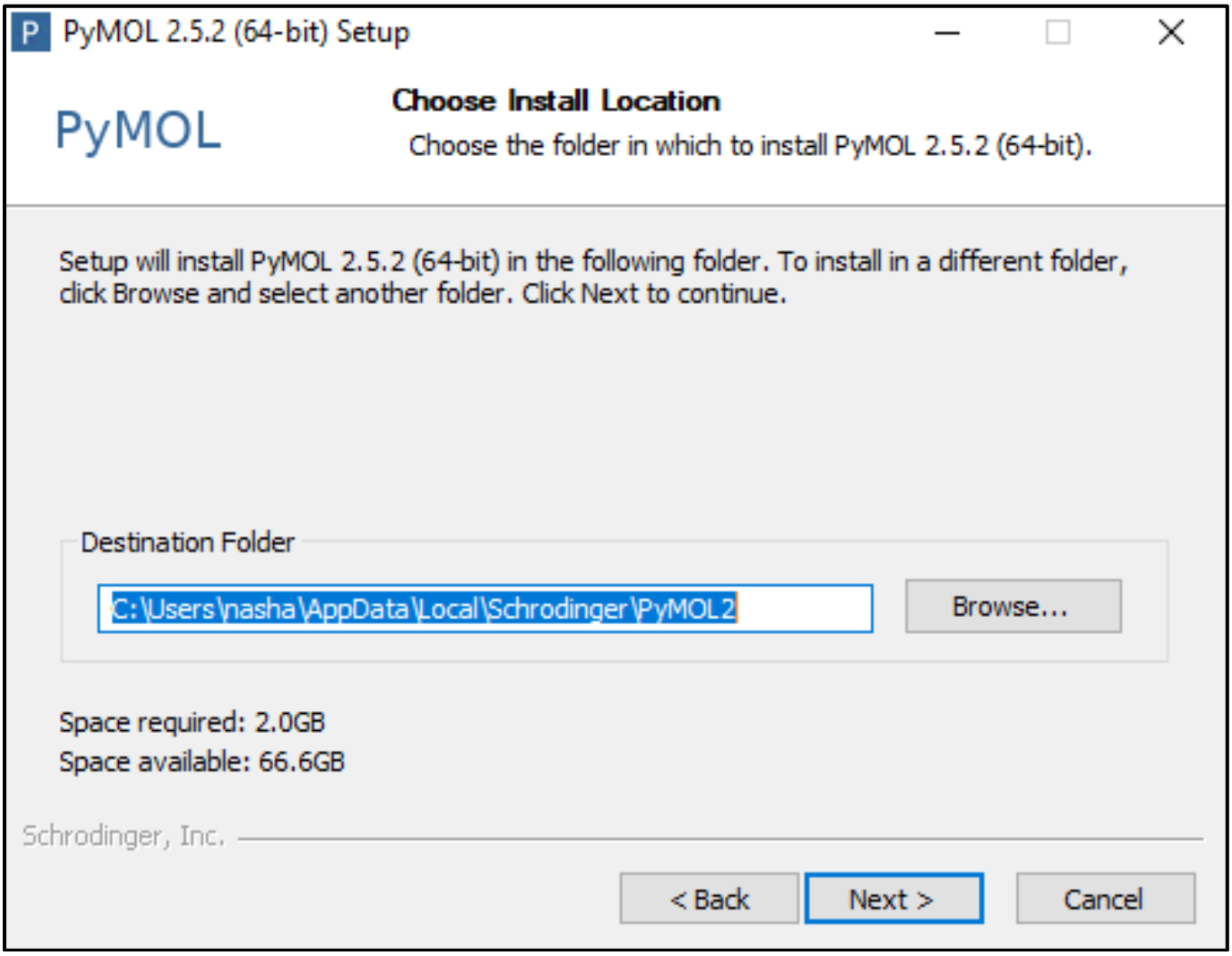
Download and run the installer. Click next to proceed



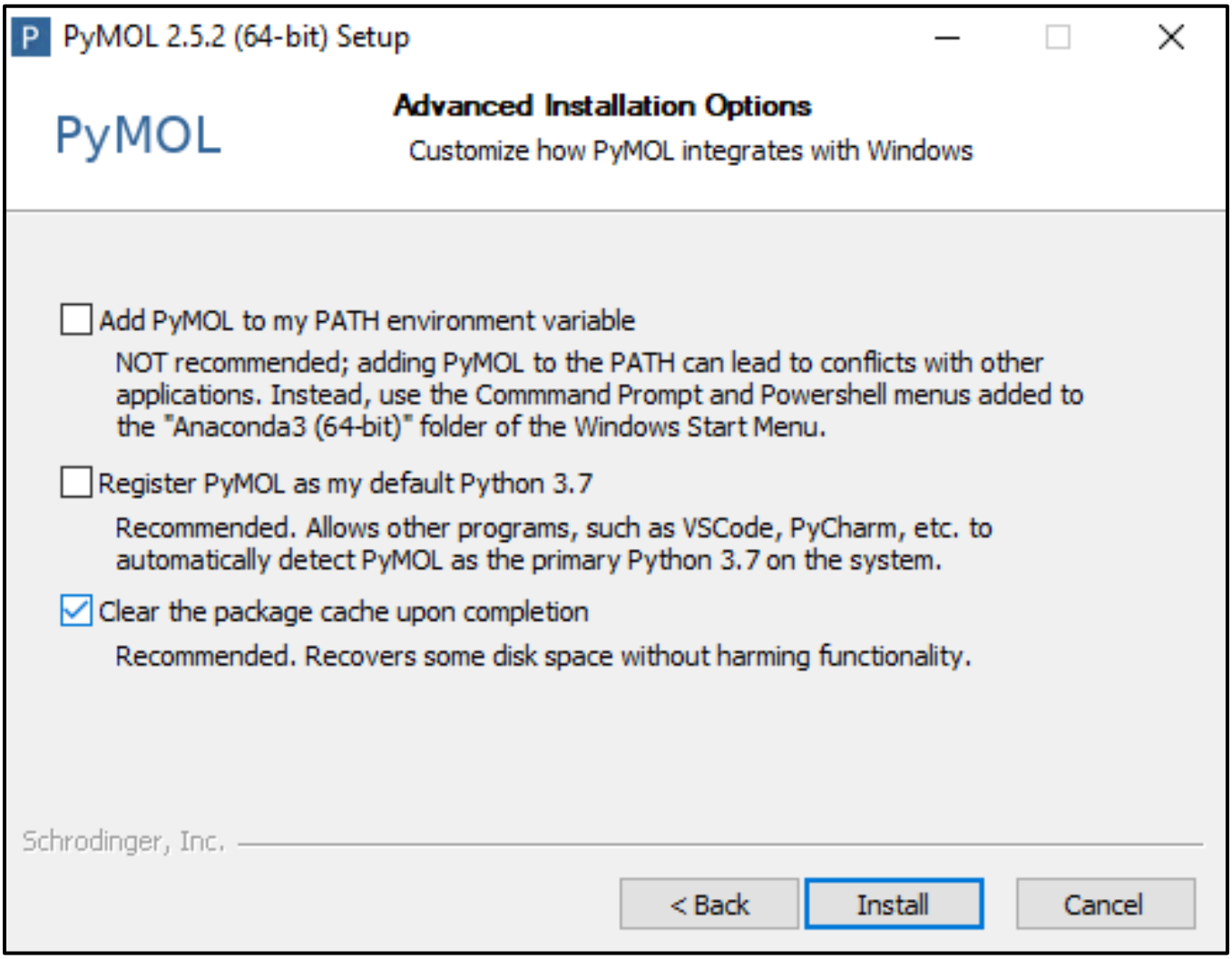
Read the License Agreement and click "I Agree" to proceed



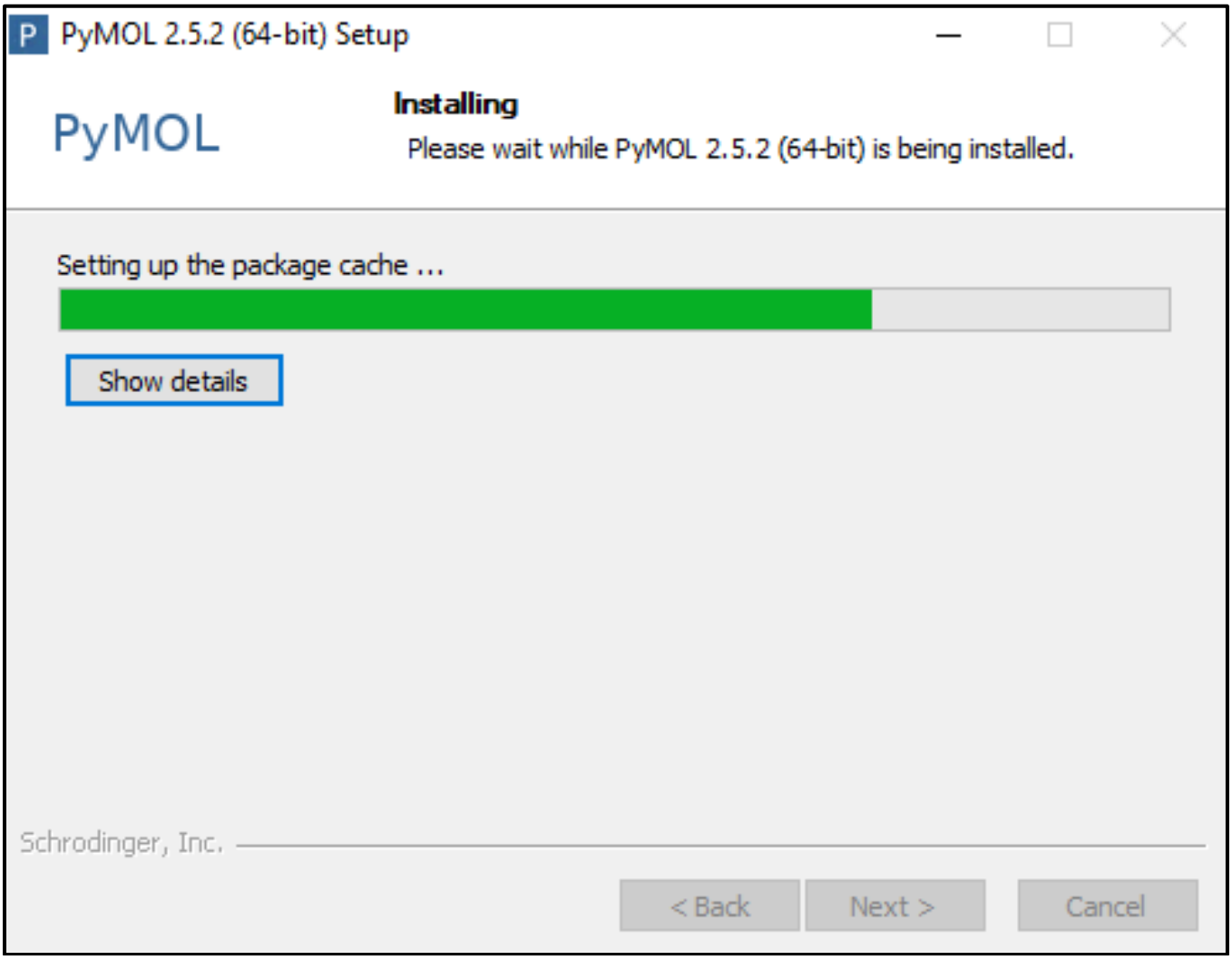
Select the Installation type and click "Next" to proceed



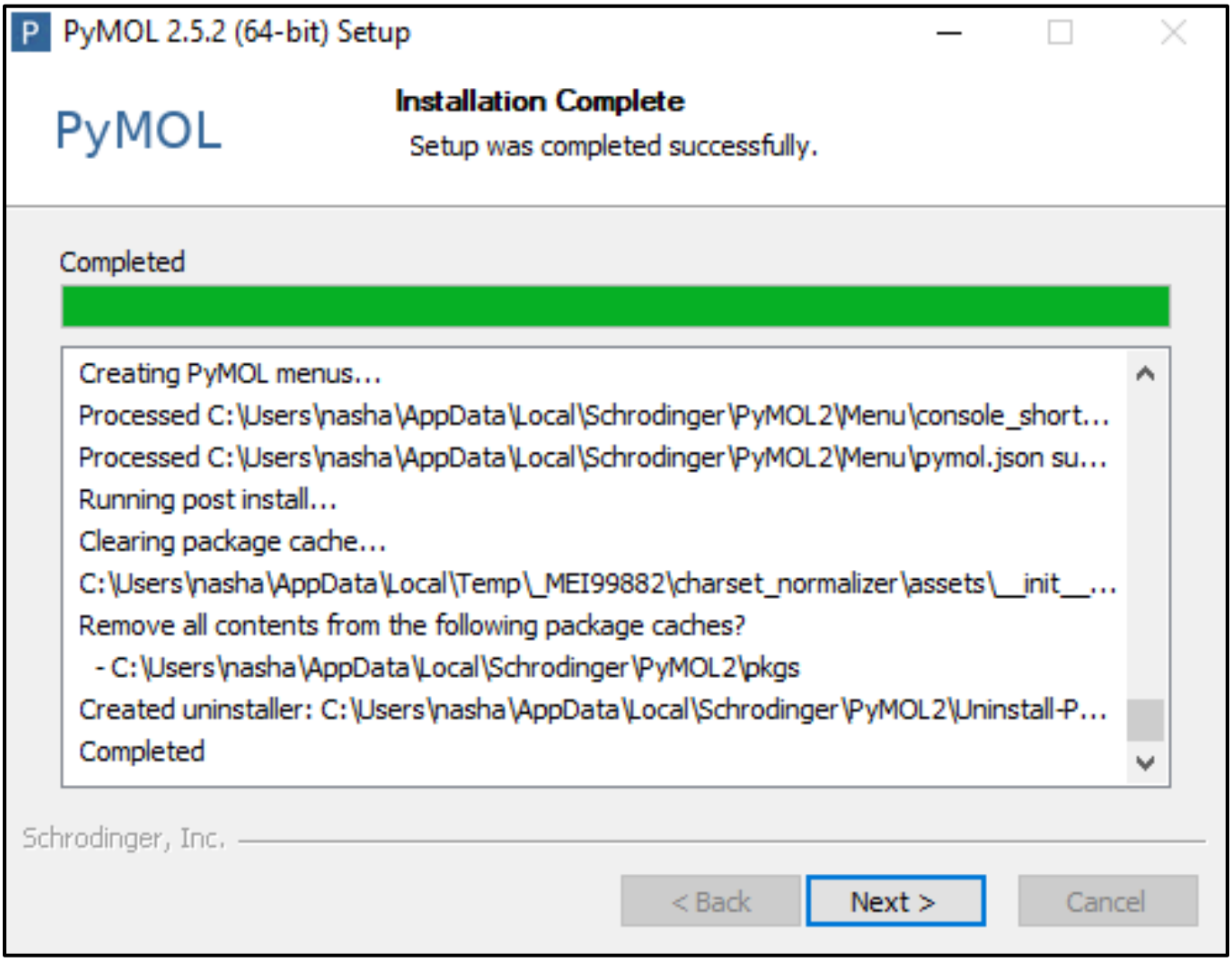
Select the Directory in which you want to install PyMol



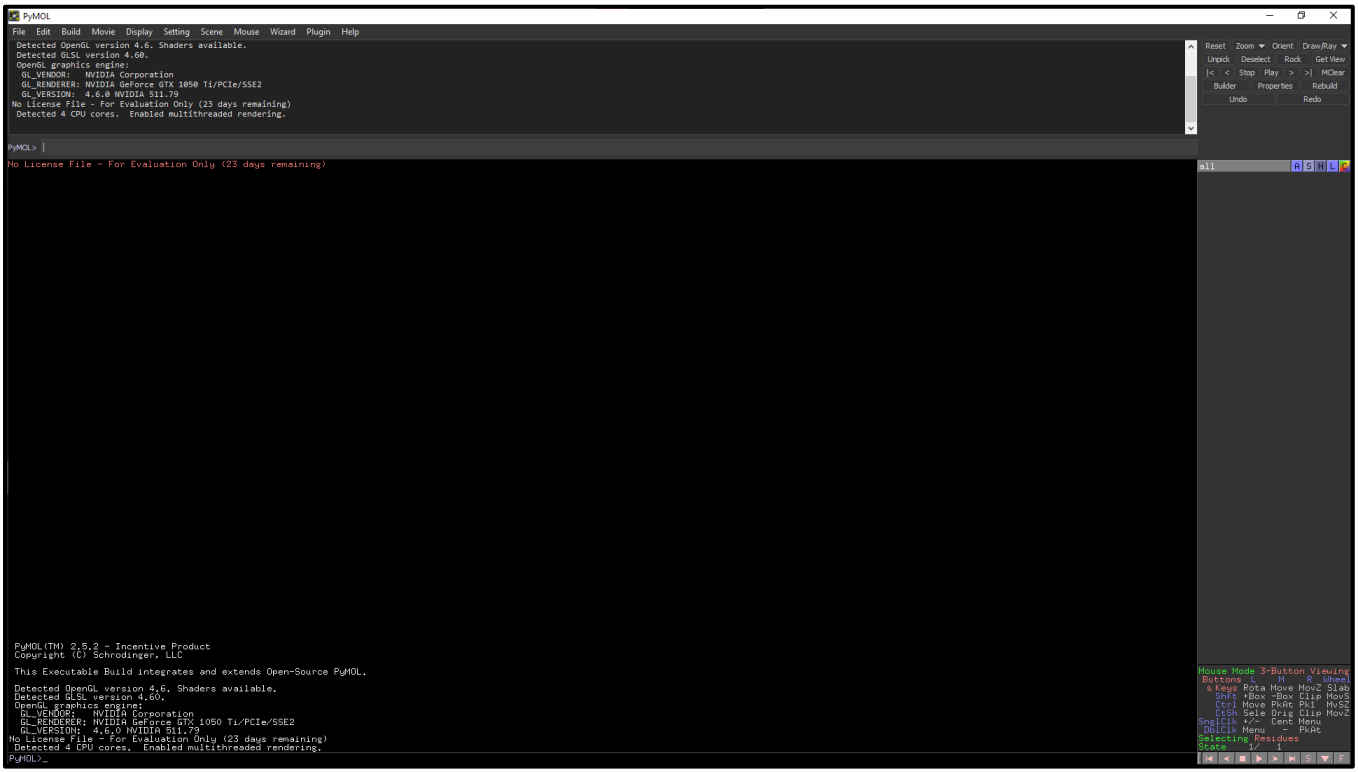
Select the additional options if you want to and click "Install" to start the installation of PyMol



Let the installer install PyMol



After completion click on "Next" to finalize installation



Pymol is now installed

Thus, RASMOL and PyMOL tools can be used for protein structure visualisation which is helpful in understanding protein–ligand modeling, molecular simulations, and drug screening. It provides an essential support for presenting results, reasoning on and formulating hypotheses related to molecular structure. It also helps to analyze and compare protein structures to gain insight to functions of the proteins.

REFERENCES:

1. Xiong, J. (2008). Protein Structure Visualization, Comparison, and Classification. Essential bioinformatics. Cambridge: Cambridge University Press. 187-188.
2. Yuan, S., Chan, H. C. S., & Hu, Z. (2017). Using PyMOL as a platform for computational drug design. WIREs Computational Molecular Science, 7(2). https://doi.org/10.1002/wcms.1298
3. RasMol and OpenRasMol. (n.d.). Www.openrasmol.org. Retrieved March 4, 2022, from http://www.openrasmol.org/
4. PyMOL | pymol.org. (2019). Pymol.org. Retrieved March 4, 2022, from https://pymol.org/2/

Date: 05-03-22

WEBLEM 6A

To visualize 3D structure of Hemoglobin (1SI4) using RASMOL & PyMOL tool

RASMOL:

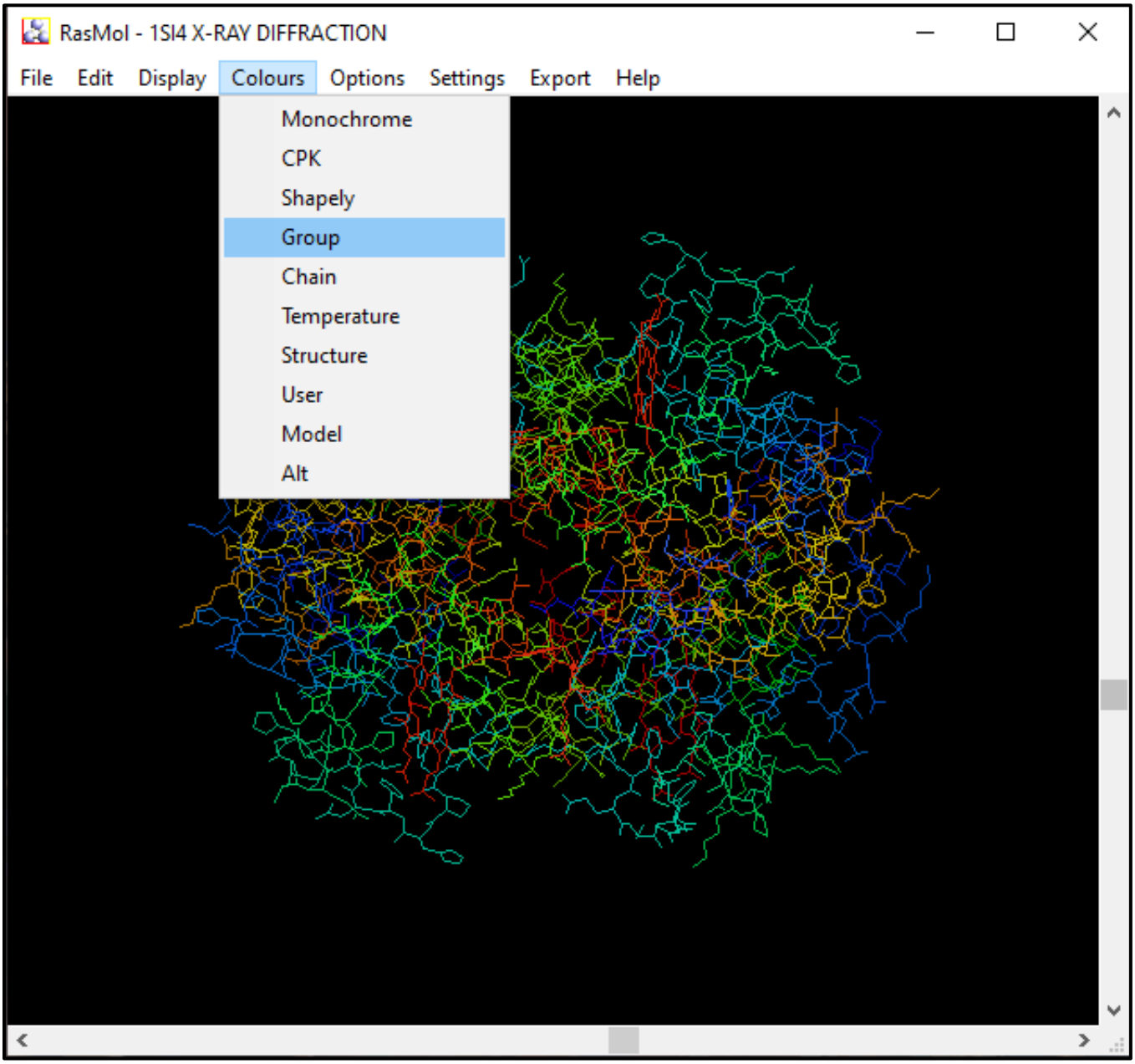


Fig1. PDB Structure of Hemoglobin (1SI4) loaded in RasMol to show it in 3D space with the colour scheme set to Groups

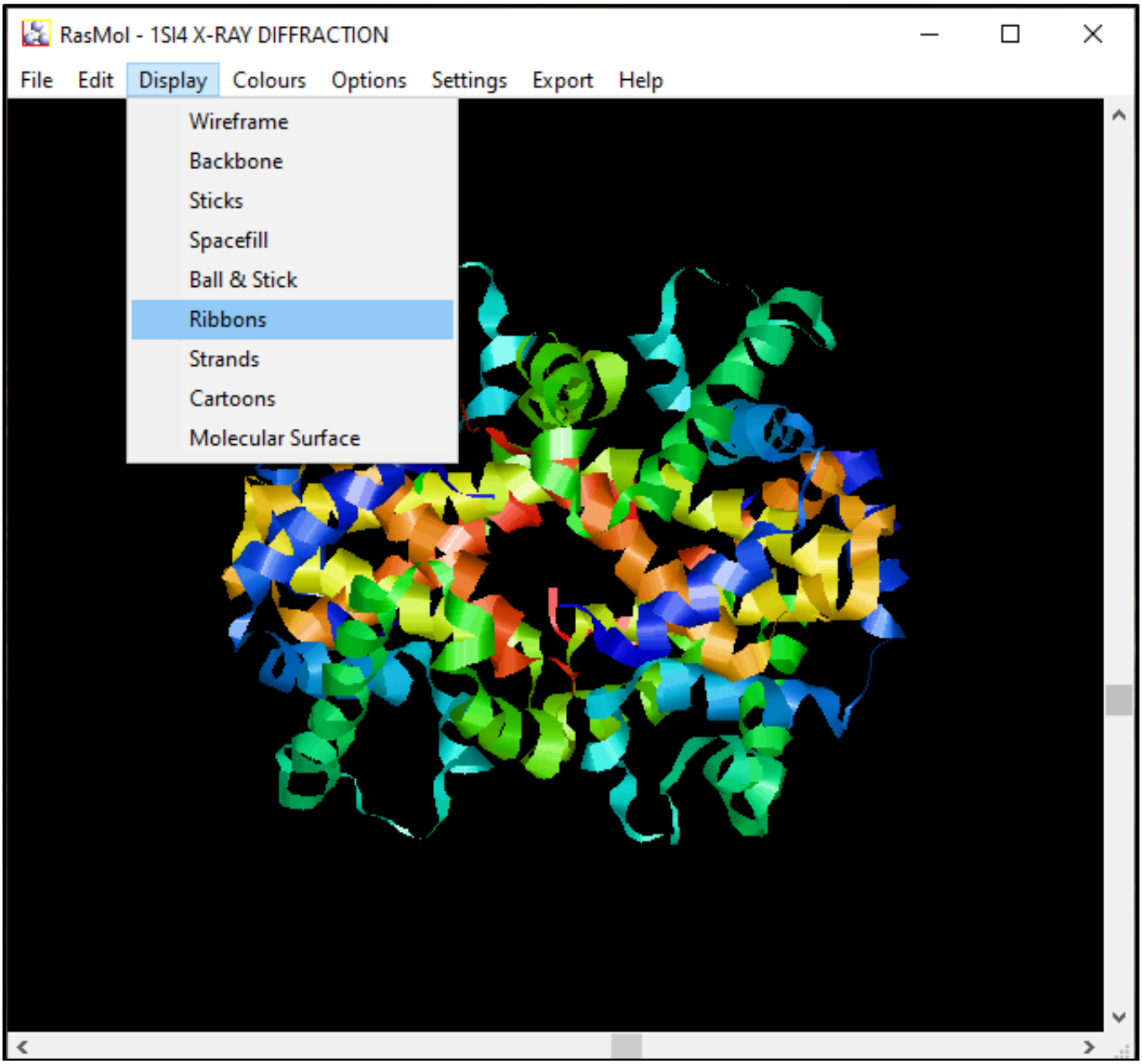


Fig2. PDB structure of Hemoglobin (1SI4) Shown with the display mode set to ribbons

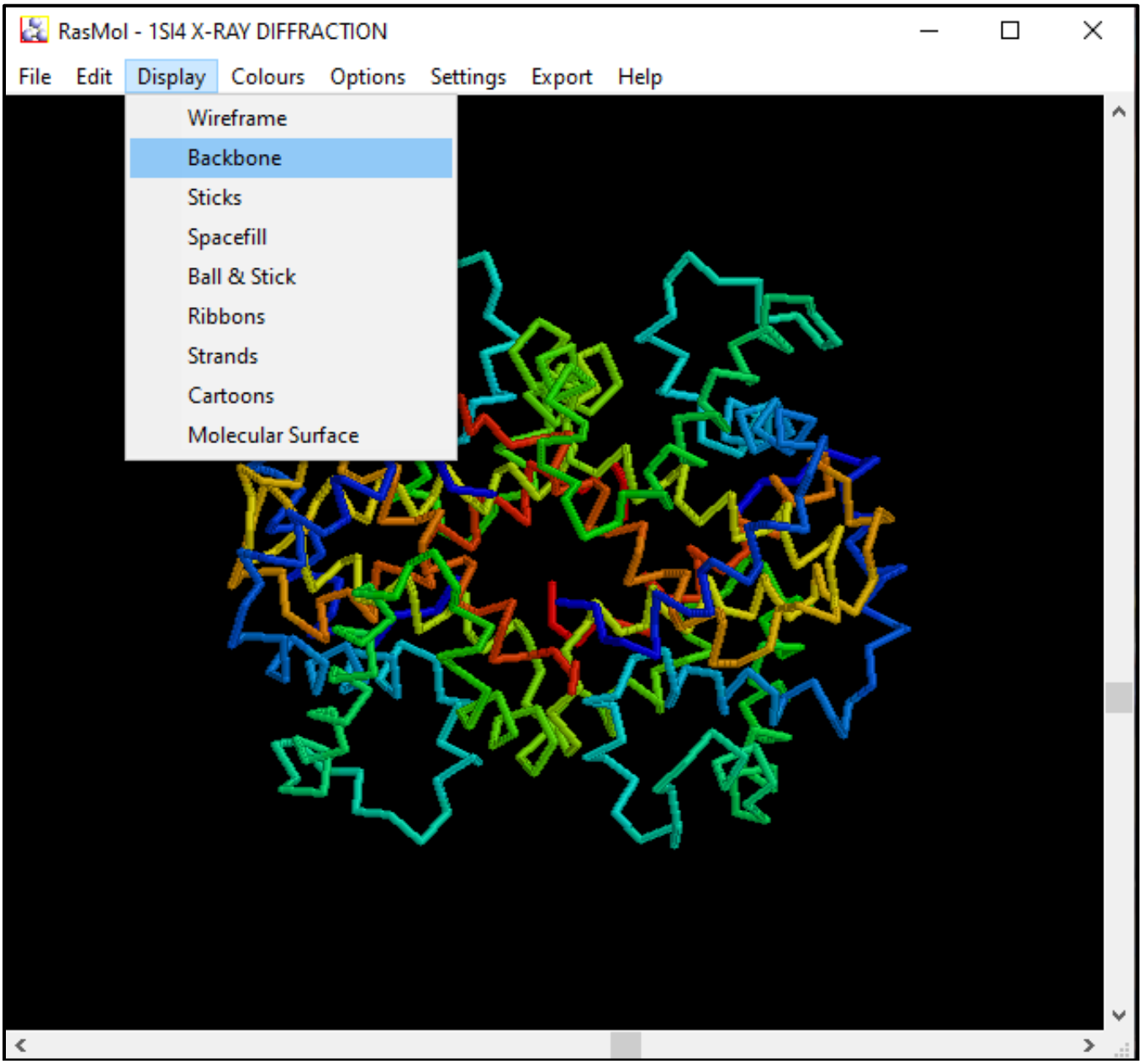


Fig3. PDB structure of Hemoglobin (1SI4) shown with display type set to Backbone

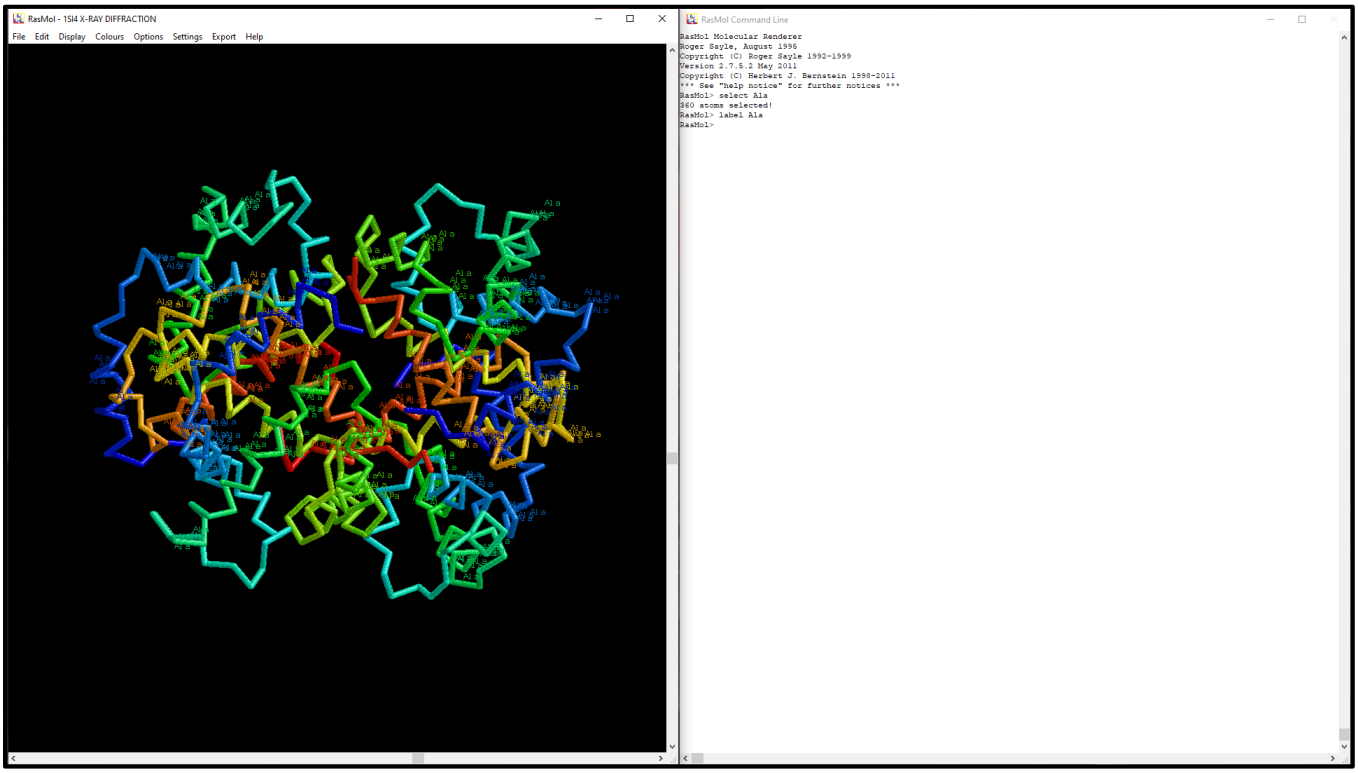


Fig.4 PDB Structure of Hemoglobin (1SI4) visualizing Alanine in the structure.

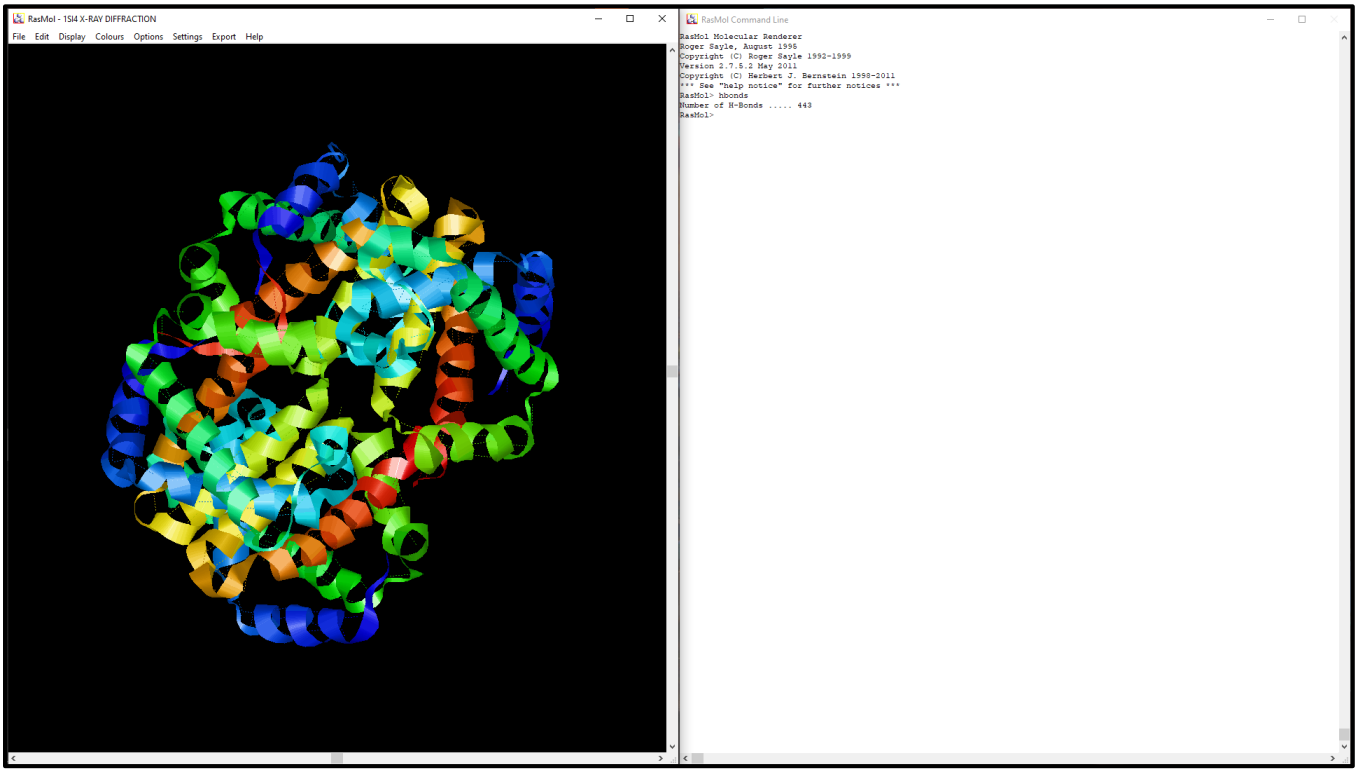


Fig5. PDB structure of Hemoglobin (1SI4) with Hydrogen Bonds visualized

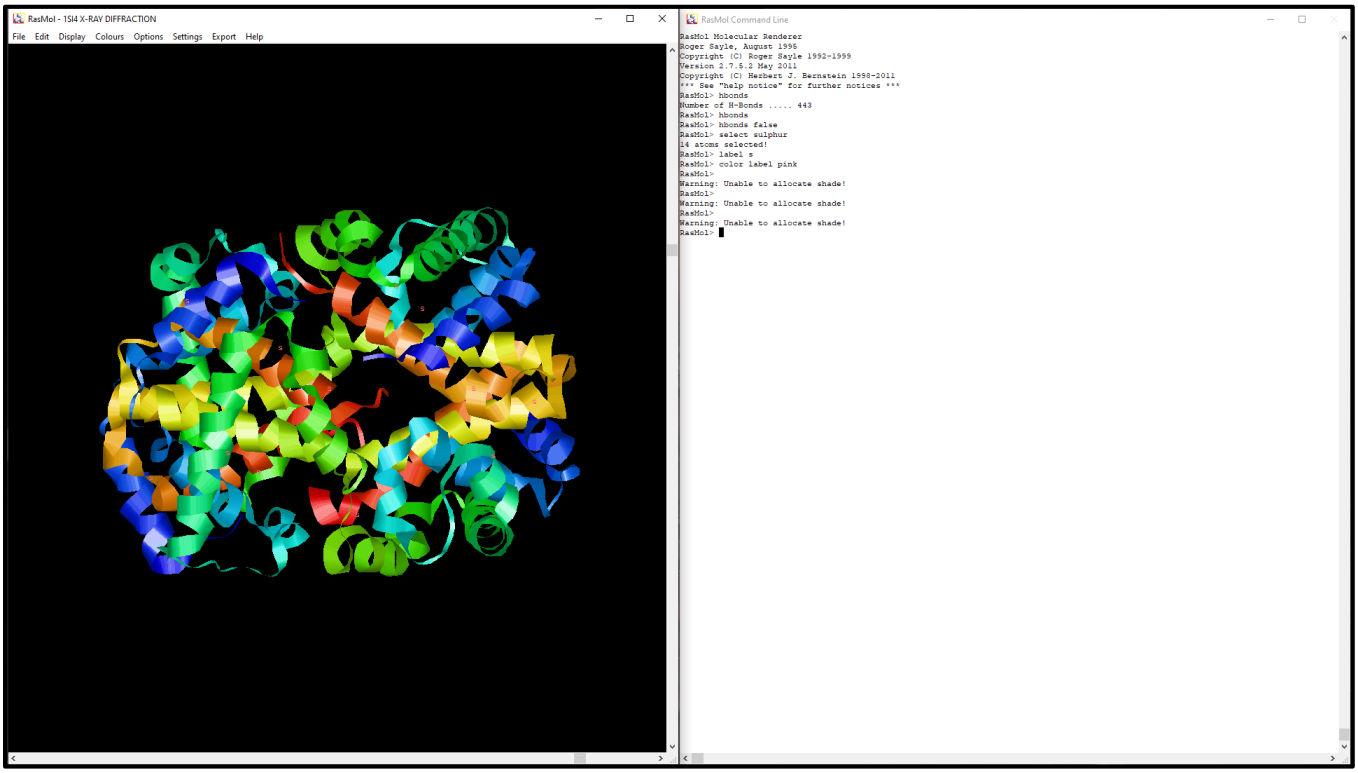


Fig6. PDB structure of Hemoglobin (1SI4) with sulphur atoms visualized with label "s"

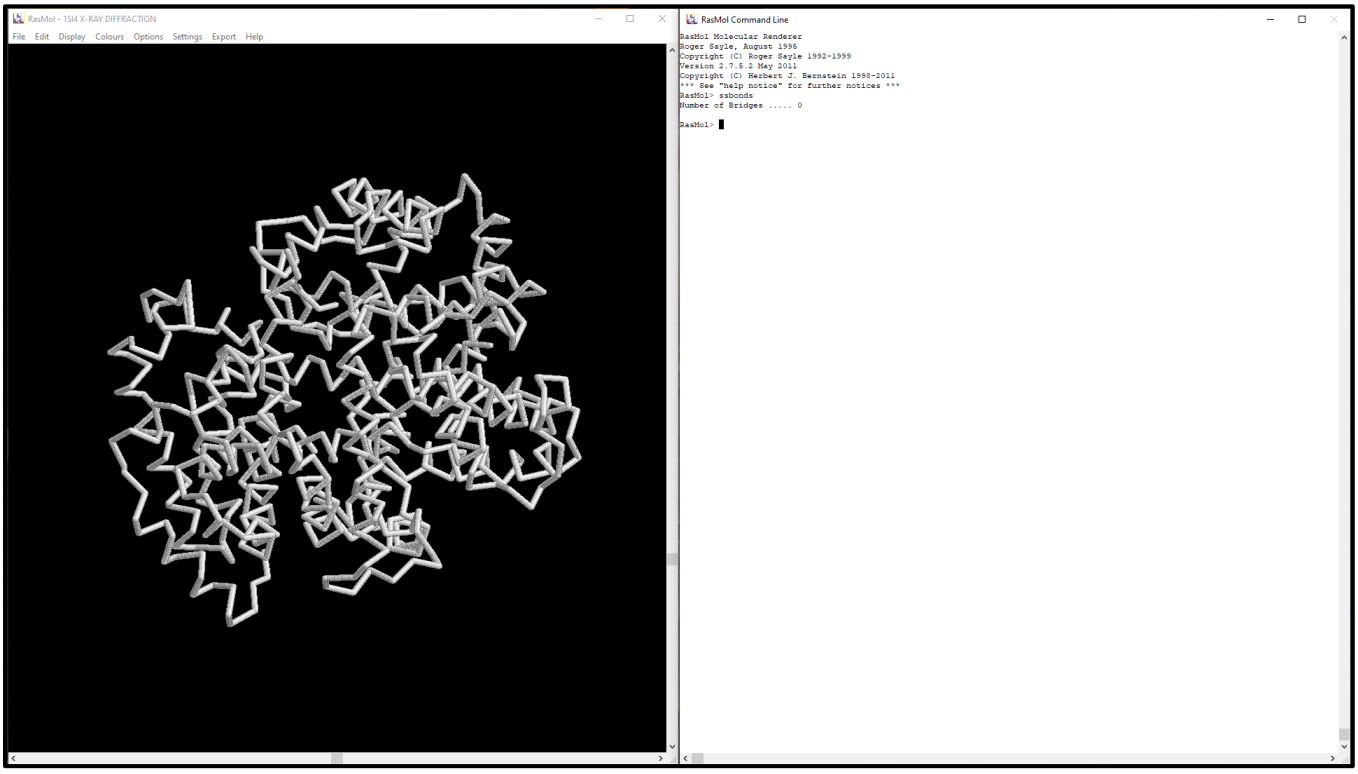


Fig7. PDB structure of Hemoglobin (1SI4) visualizing sulphuer-sulphur bonds (0 in this case)

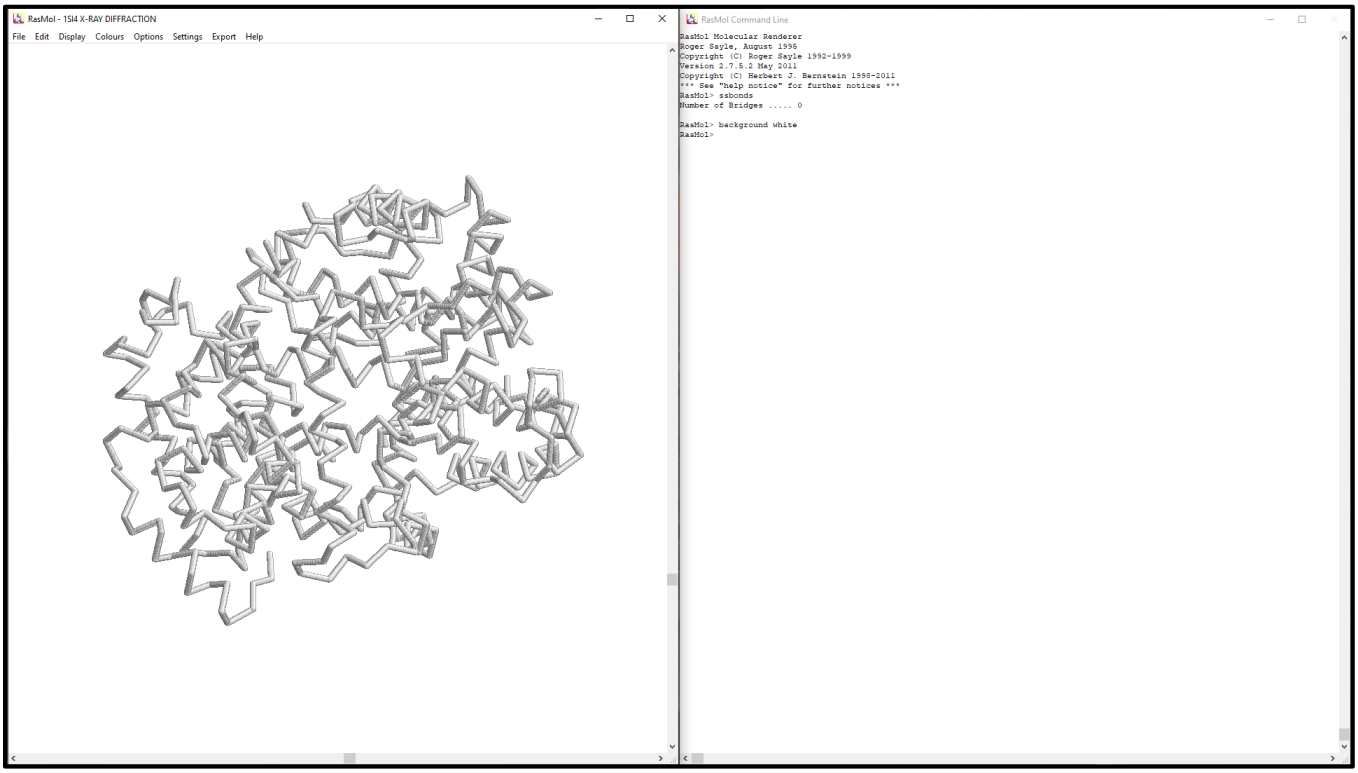


Fig8. PDB structure of Hemoglobin (1SI4) with the background color set to white to clearly see the sulphur-sulphur bonds (0 in this case)

PYMOL:

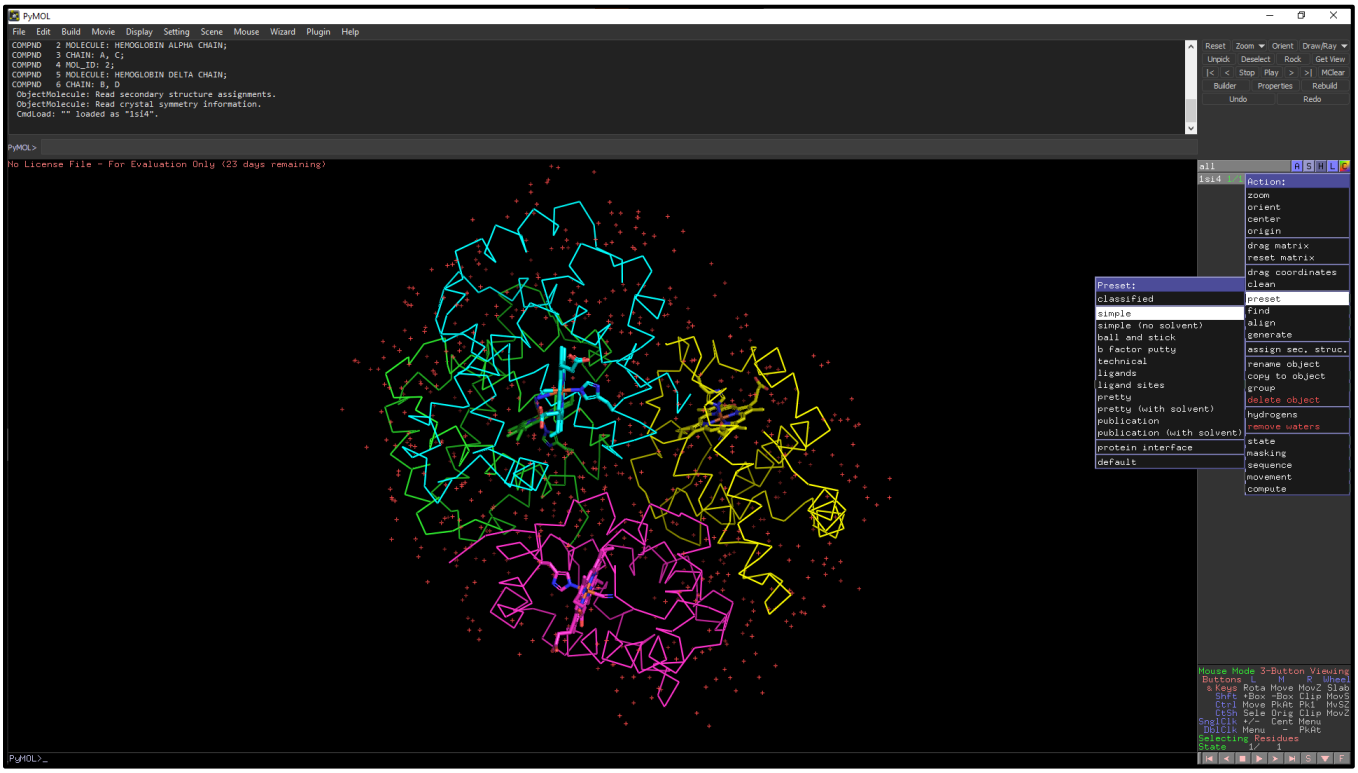


Fig9. PDB structure Hemoglobin (1SI4) loaded in PyMol and set to simple view preset

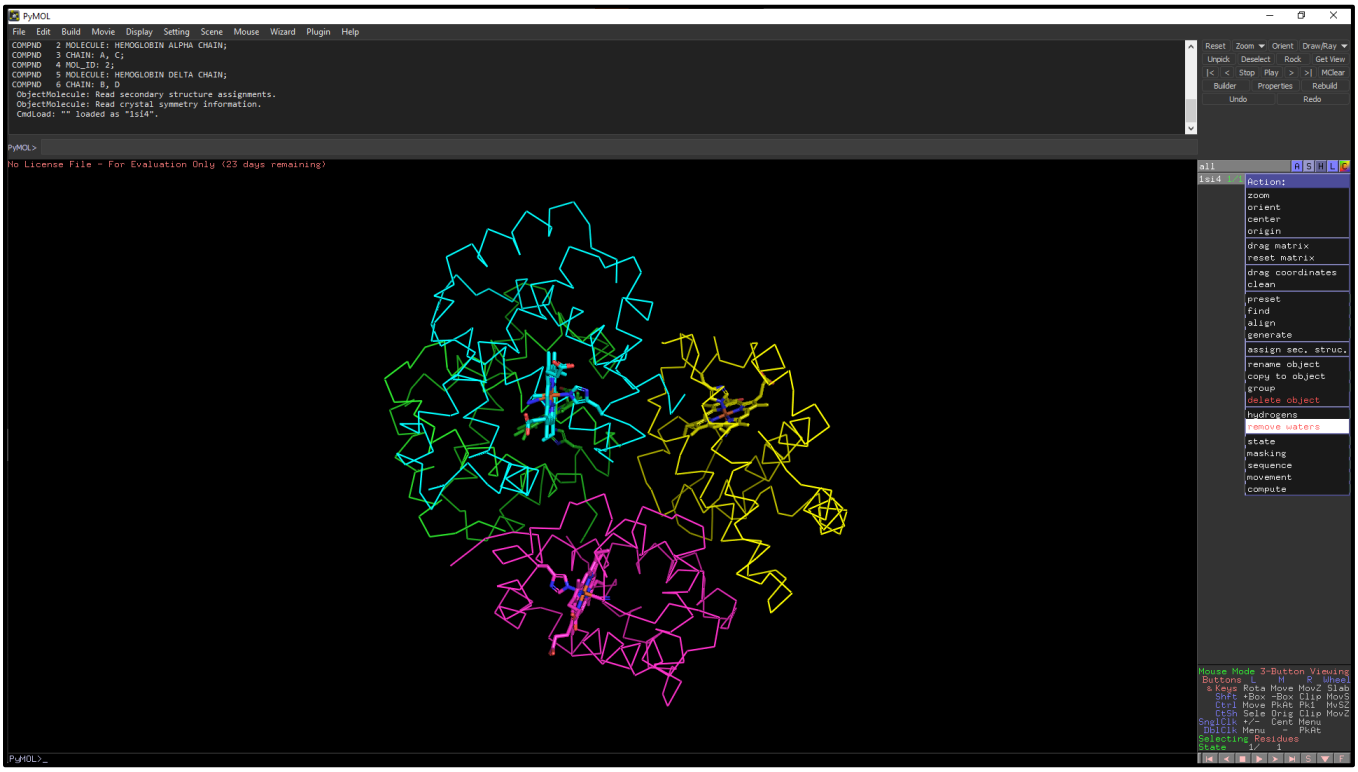


Fig10. PDB structure of Hemoglobin (1SI4) with waters removed

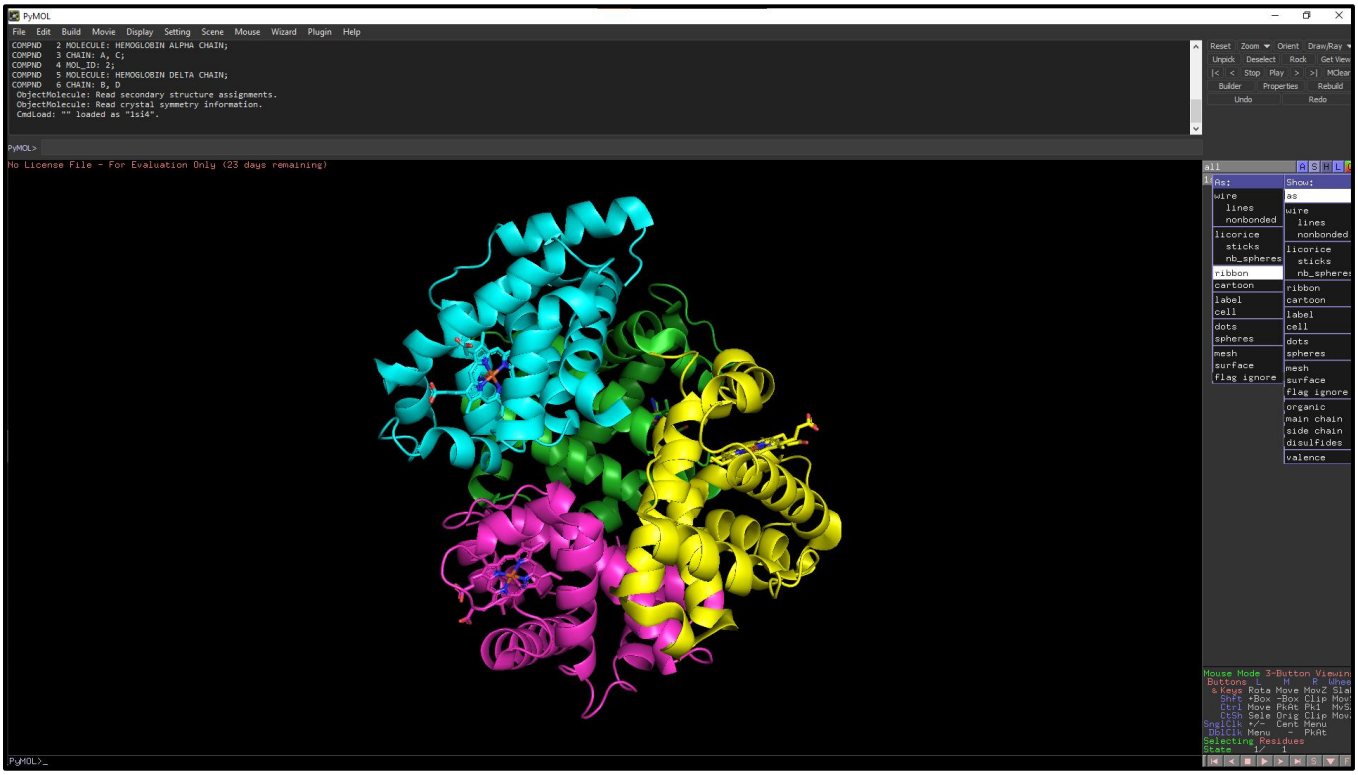


Fig11. PDB structure of Hemoglobin (1SI4) with visual style set to ribbon

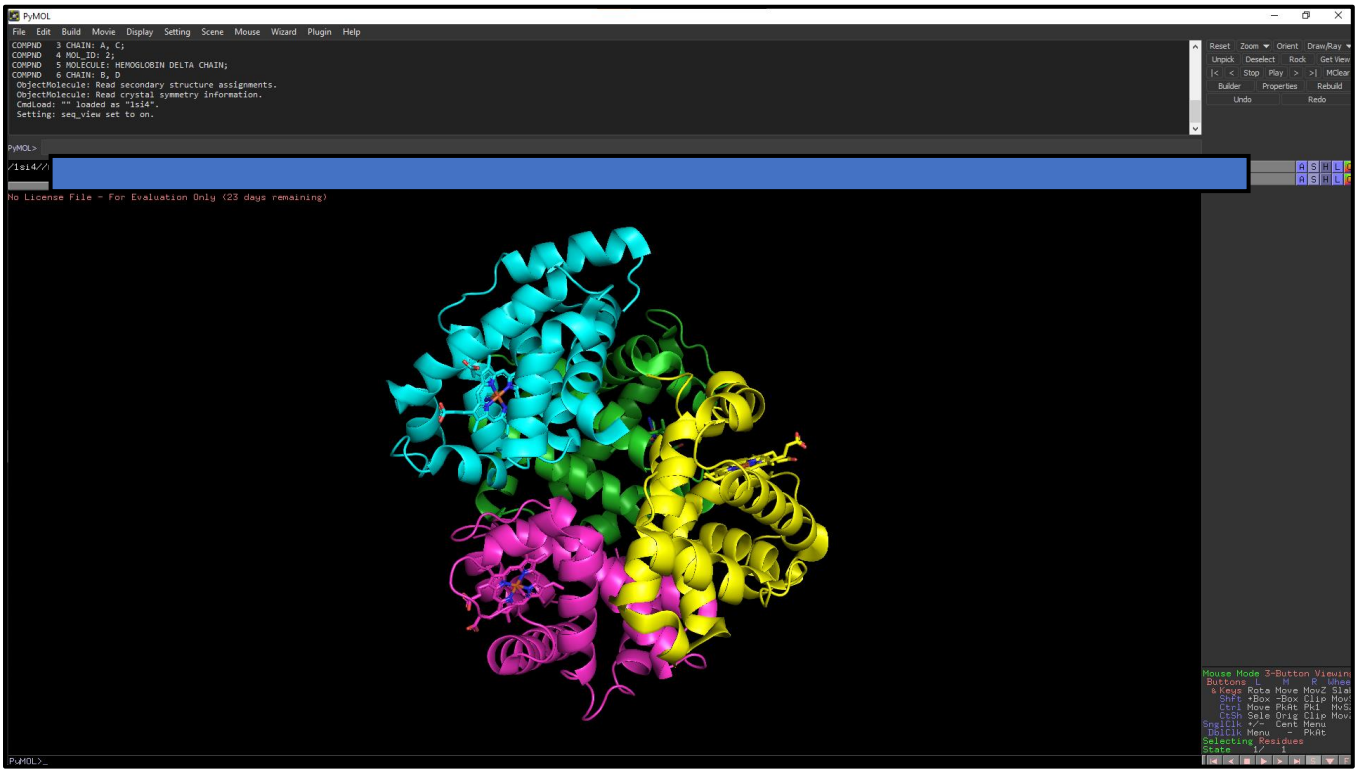


Fig12. PDB structure of Hemoglobin (1SI4) with the sequence displayed

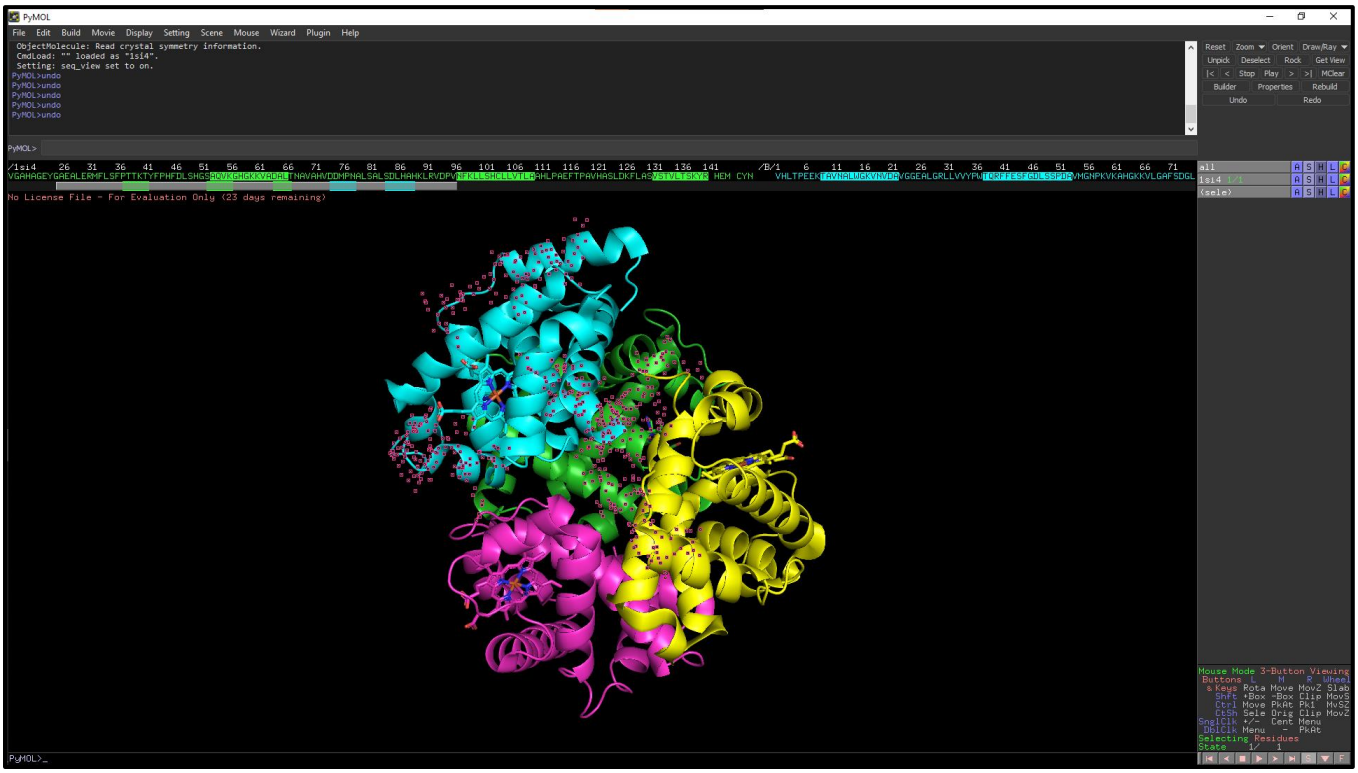


Fig13. PDB structure of Hemoglobin (1SI4) with random segments from sequence selected and visualized

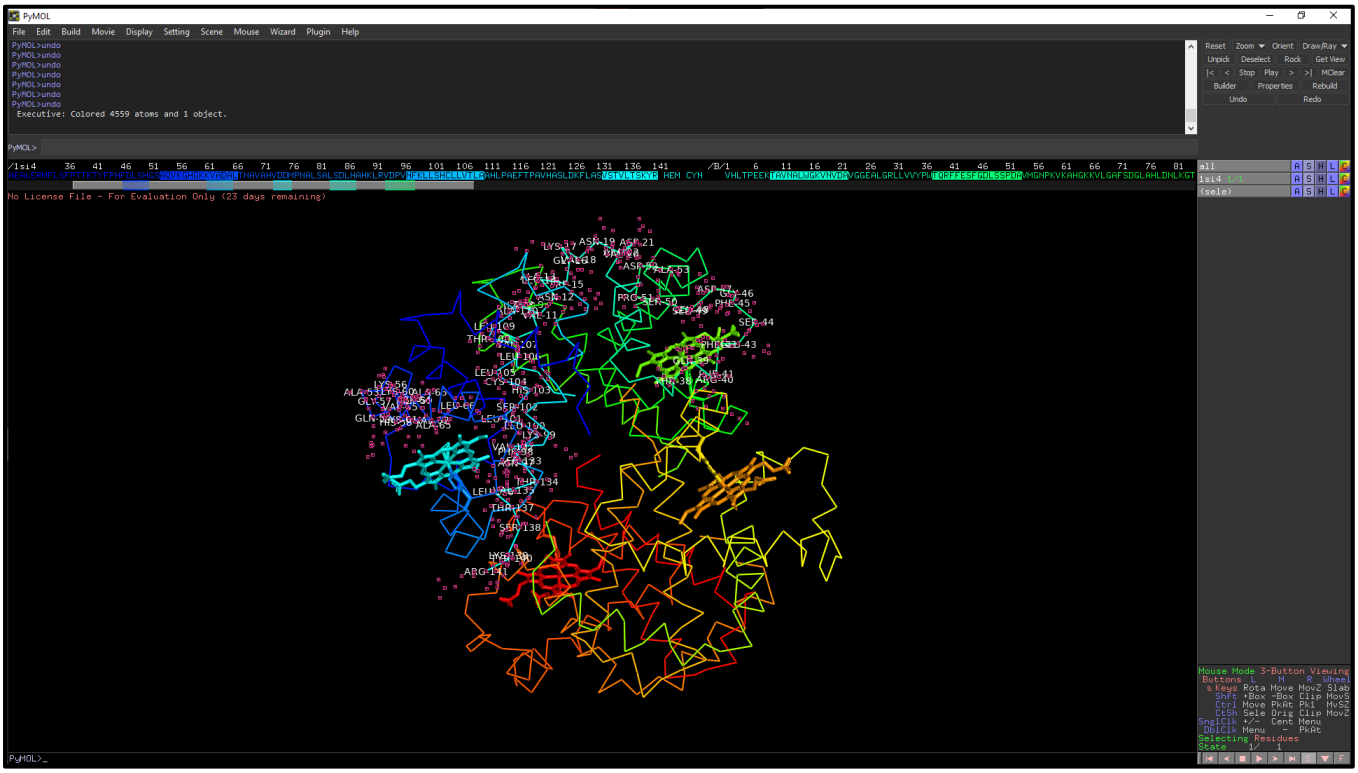


Fig14. PDB structure of Hemoglobin (1SI4) with the residues of the randomly selected segments from sequence visualized

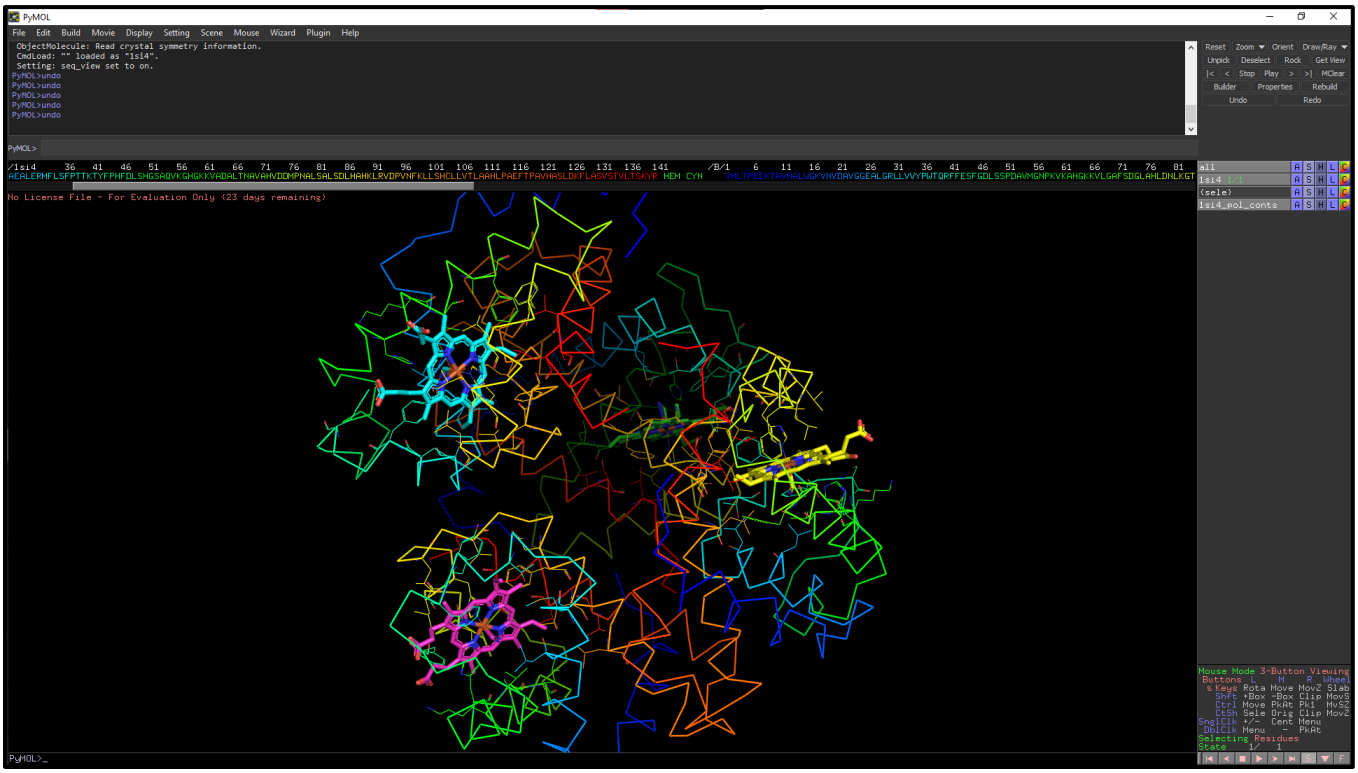


Fig15. PDB structure of Hemoglobin (1SI4) with it set to preset ligand view

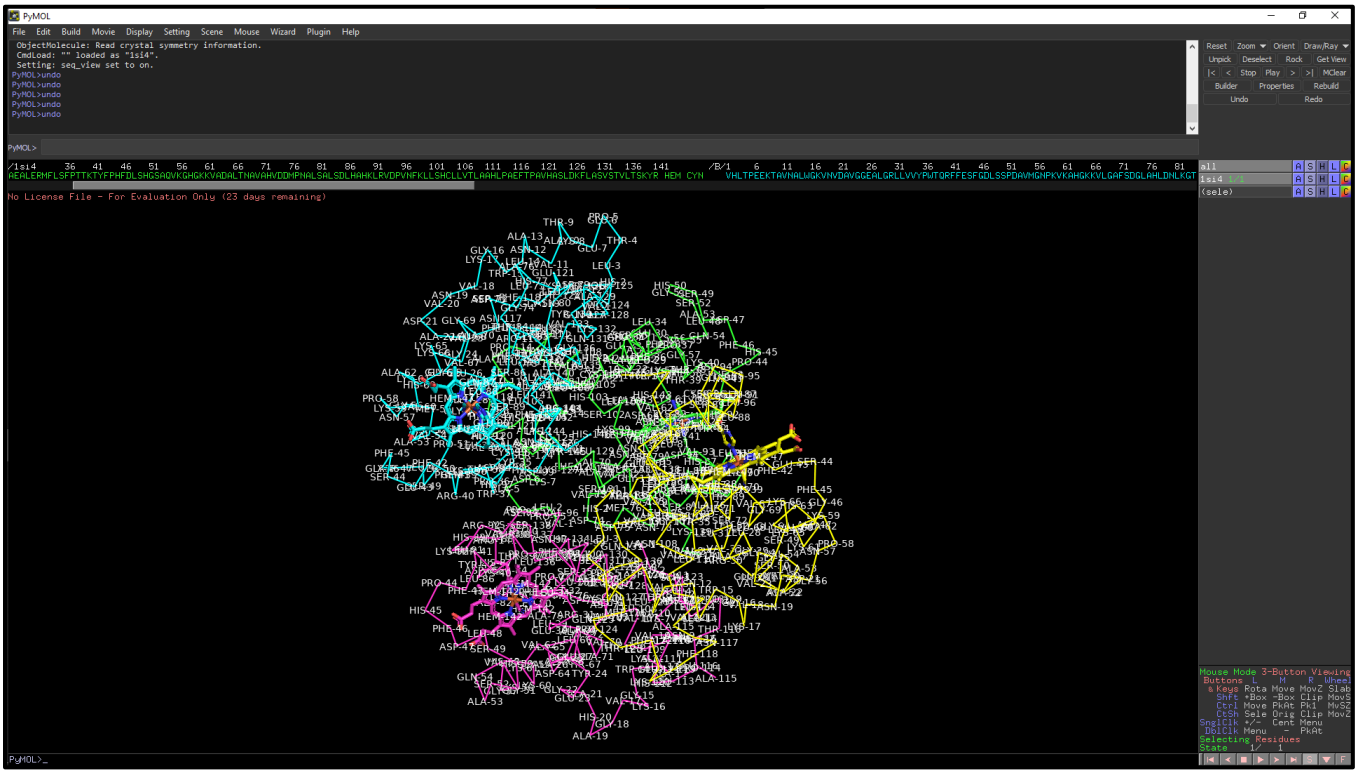


Fig16. PDB structure of Hemoglobin (1SI4) with all its residues labelled

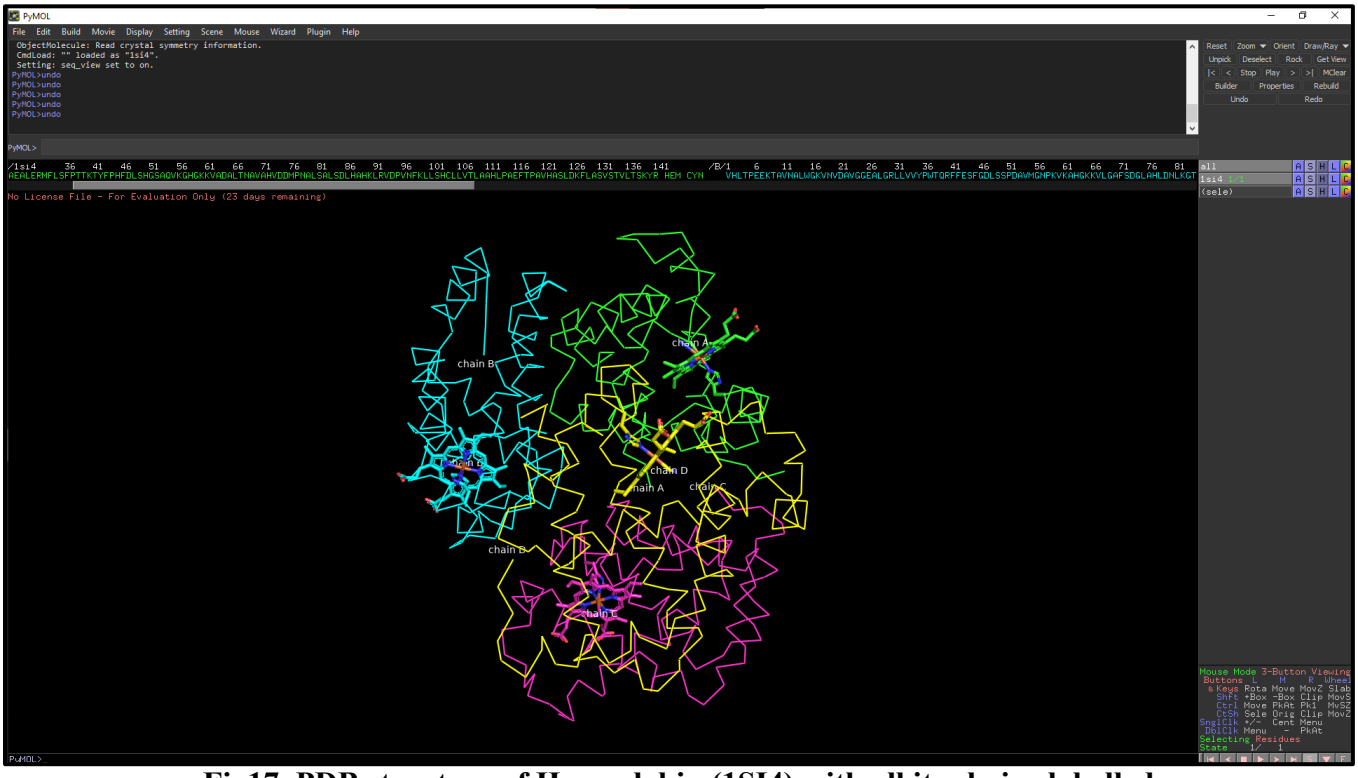


Fig17. PDB structure of Hemoglobin (1SI4) with all its chains labelled

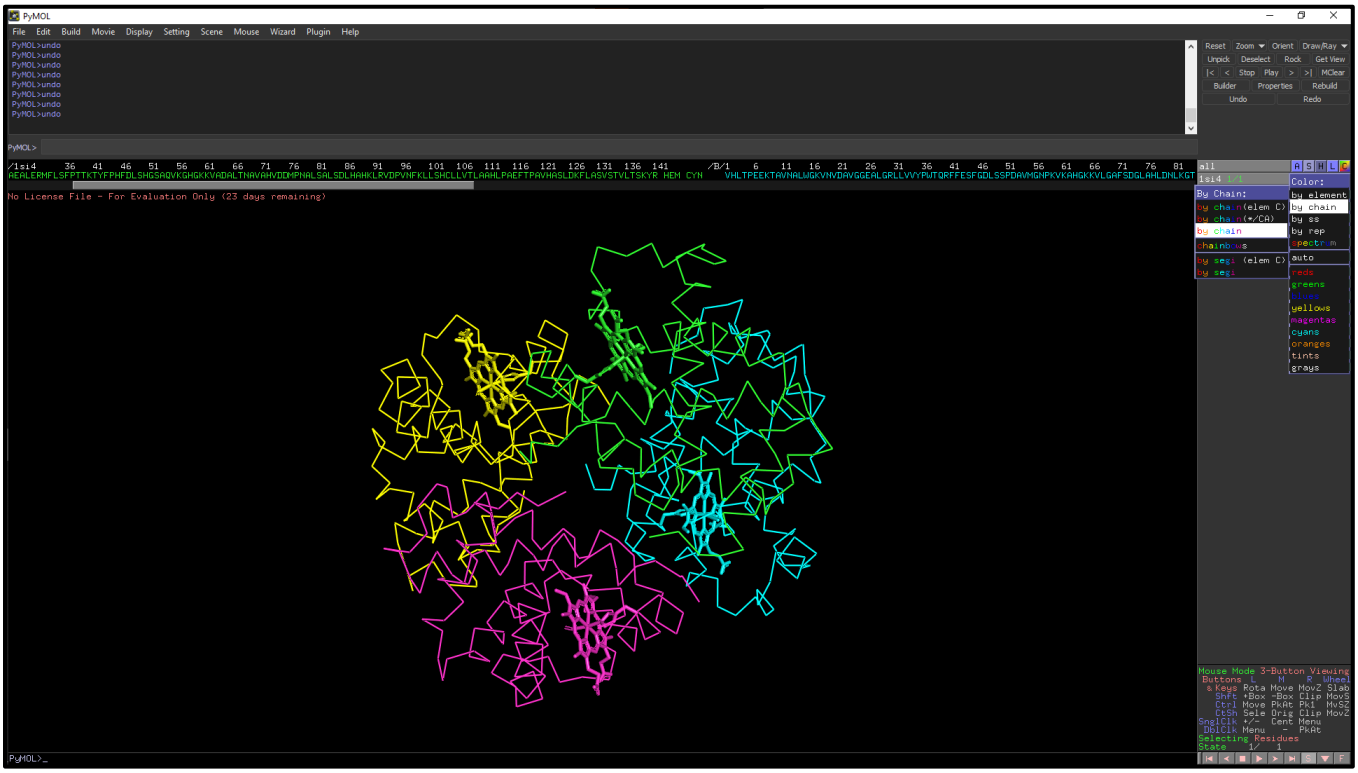


Fig18. PDB structure of Hemoglobin (1SI4) visualized with the chain color scheme set to rainbow

RESULT:

Hemoglobin structure was visualised with various formats along with its residues, bonds and chains using RASMOL and PyMOL tools.

CONCLUSION:

RASMOL and PyMOL tools can be used for protein structure visualisation which is helpful in understanding protein–ligand modeling, molecular simulations, and drug screening. It provides an essential support for presenting results, reasoning on and formulating hypotheses related to molecular structure. It also helps to analyze and compare protein structures to gain insight to functions of the proteins.

REFERENCES:

1. The Editors of Encyclopedia Britannica. (2018). Fibrin | biochemistry. In Encyclopædia Britannica. Retrieved March 4, 2022, from https://www.britannica.com/science/fibrin
2. Yuan, S., Chan, H. C. S., & Hu, Z. (2017). Using PyMOL as a platform for computational drug design. WIREs Computational Molecular Science, 7(2). https://doi.org/10.1002/wcms.1298
3. RasMol and OpenRasMol. (n.d.). Www.openrasmol.org. Retrieved March 4, 2022, from http://www.openrasmol.org/
4. PyMOL | pymol.org. (2019). Pymol.org. Retrieved March 4, 2022, from https://pymol.org/2/

Date: 04-03-22

WEBLEM 6
Introduction to binding pocket prediction of protein w.r.t to PTM studies

Protein structures are complex and are sculpted with numerous surface pockets, internal cavities and cross channels. These topographic features provide structural basis and micro-environments for proteins to carry out their functions such as ligand binding, DNA interaction and enzymatic activity. Identification and quantification of these topographic features of proteins are therefore of fundamental importance for understanding the structure–function relationship of proteins, in engineering proteins for desired properties and in developing therapeutics against protein targets.

CASTp:

- The CASTp server aims to provide comprehensive and detailed quantitative characterization of topographic features of proteins. Since its release 15 years ago, the CASTp server has ∼45,000 visits and fulfills ∼33,000 calculation requests annually. It has been proven to be a useful tool for a wide range of studies, including investigations of signaling receptors, discoveries of cancer therapeutics, understanding of mechanism of drug actions, studies of immune disorder diseases, analysis of protein–nanoparticle interactions, inference of protein functions and development of high-throughput computational tools.
- The CASTp server takes protein structures in the PDB format and a probe radius as input for topographic computation. Through the intuitive interface, users can either search for pre-computed results using a four-letter PDB ID, or submit their own protein structures to request customized computation. For pre-computed results, a default probe radius of 1.4 Å is used, which is the standard value for computing solvent accessible surface area. For customized computation request, users can specify any probe radius desired.
- The CASTp server identifies all surface pockets, interior cavities and cross channels in a protein structure and provides detailed delineation of all atoms participating in their formation. It also measures their exact volumes and areas, as well as sizes of the mouth openings if exist. These metrics are calculated analytically, using both the solvent accessible surface model (Richards' surface)and the molecular surface model (Connolly's surface). In addition, the CASTp server also provides imprints of topographic features. These results can be directly downloaded from CASTp server, which can be visualized using either the UCSF Chimera or our PyMOL plugin, CASTpyMOL.

NetOGlyc – 4.0

- Glycosylation is the most abundant and diverse posttranslational modification of proteins. While several types of glycosylation can be predicted by the protein sequence context, and substantial knowledge of these glycoproteomes is available, our knowledge of the GalNAc-type O-glycosylation is highly limited. This type of glycosylation is unique in being regulated by 20 polypeptide GalNAc-transferases attaching the initiating GalNAc monosaccharides to Ser and Thr (and likely some Tyr) residues.
- The finding of unique subsets of O-glycoproteins in each cell line provides evidence that the O-glycoproteome is differentially regulated and dynamic. The greatly expanded view of the O-glycoproteome should facilitate the exploration of how site-specific O-glycosylation regulates protein function.
- The output conforms to the GFF version 2 format. For each input sequence the server prints a list of potential glycosylation sites, showing their positions in the sequence and the prediction confidence scores. Only the sites with scores higher than 0.5 are predicted as glycosylated and marked with the string "#POSITIVE" in the comment field.

NetPhos - 3.1

- Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes. How is specificity in substrate recognition and phosphorylation by protein kinases achieved?

- In addition, serine and threonine residues in p300/CBP that can be modified by O-linked glycosylation with N-acetylglucosamine are identified. Glycosylation may prevent phosphorylation at these sites, a mechanism named yin-yang regulation.
- The results can be interpreted as:
    - **Sequence** - the sequence name;
    - # - the position of the residue in the sequence;
    - **x** - the residue in one-letter code;
    - **Context** - the sequence context of the residue, shown as a 9-residue subsequence centered on the residue;
    - **Score** - the prediction score (a value in the range [0.000-1.000]; the scores above **0.500** indicate positive predictions);
    - **Kinase** - the active kinase or the string "unsp" for non-specific prediction (as in NetPhos 2.0);
    - **Answer** - the string "**YES**" for positive predictions, else a dot.

References:

- Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: Computed atlas of surface topography of proteins. Nucleic Acids Research, 46(W1). https://doi.org/10.1093/nar/gky473 Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology.
- Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Bennett EP, Mandel U, Brunak S, Wandall HH, Levery SB, Clausen H. EMBO J, 32(10):1478-88, May 15, 2013. (doi: 10.1038/emboj.2013.79. Epub 2013 Apr 12)
- Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. Blom, N., Gammeltoft, S., and Brunak, S. Journal of Molecular Biology: 294(5): 1351-1362, 1999.

Date: 04-03-22

WEBLEM 6A
To predict binding pocket of protein Glutamine using Castp server.

Introduction:

- The CASTp server aims to provide comprehensive and detailed quantitative characterization of topographic features of proteins. Since its release 15 years ago, the CASTp server has ~45,000 visits and fulfills ~33,000 calculation requests annually. It has been proven to be a useful tool for a wide range of studies.
- **Glutamine**
    - Glutamine is the most abundant and versatile amino acid in the body. In health and disease, the rate of glutamine consumption by immune cells is similar or greater than glucose. For instance, in vitro and in vivo studies have determined that glutamine is an essential nutrient for lymphocyte proliferation and cytokine production, macrophage phagocytic plus secretory activities, and neutrophil bacterial killing. Glutamine release to the circulation and availability is mainly controlled by key metabolic organs, such as the gut, liver, and skeletal muscles.

Methodology:

- Take a PDB id from a protein structure of Glutamine
- Enter the PDB id on the webpage of CASTp
- Interpret the results

Observation:

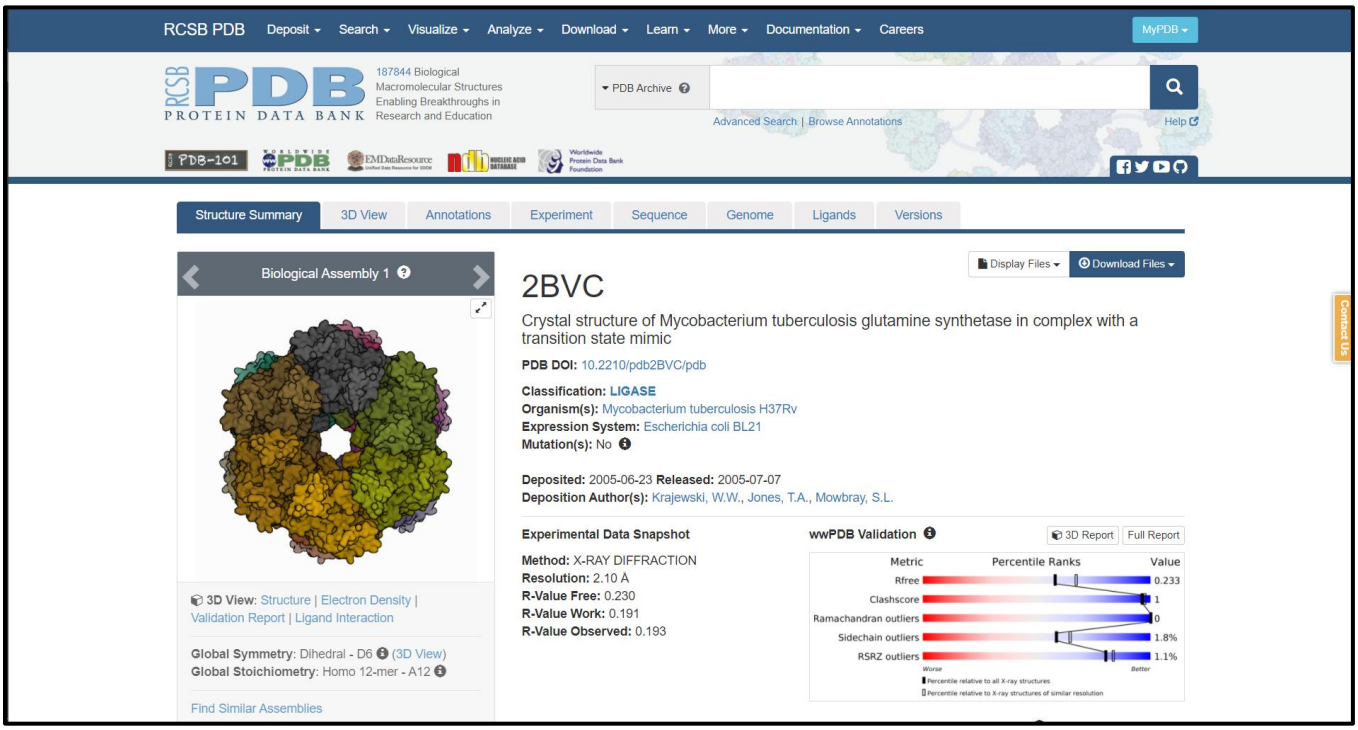


Fig1. PDB page for query Glutamine

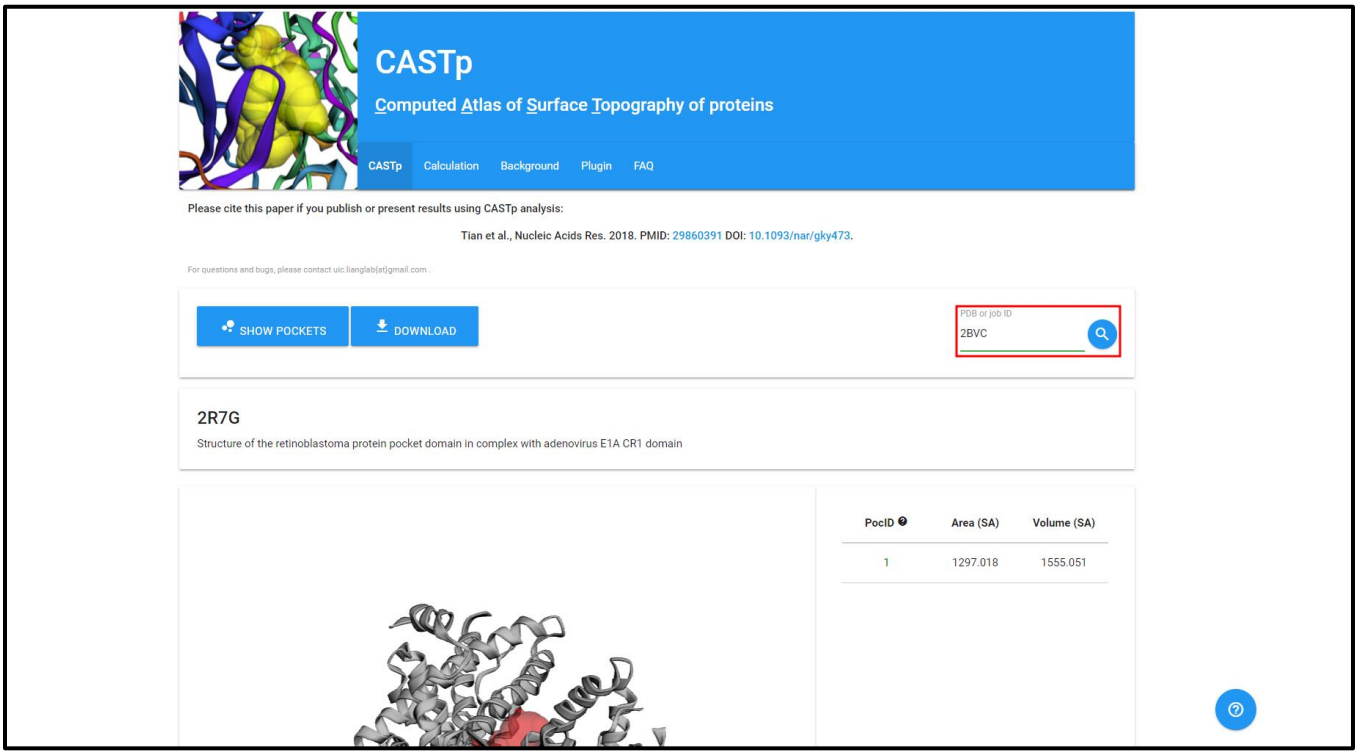


Fig2. CASTp page with the PDB id of my query entered

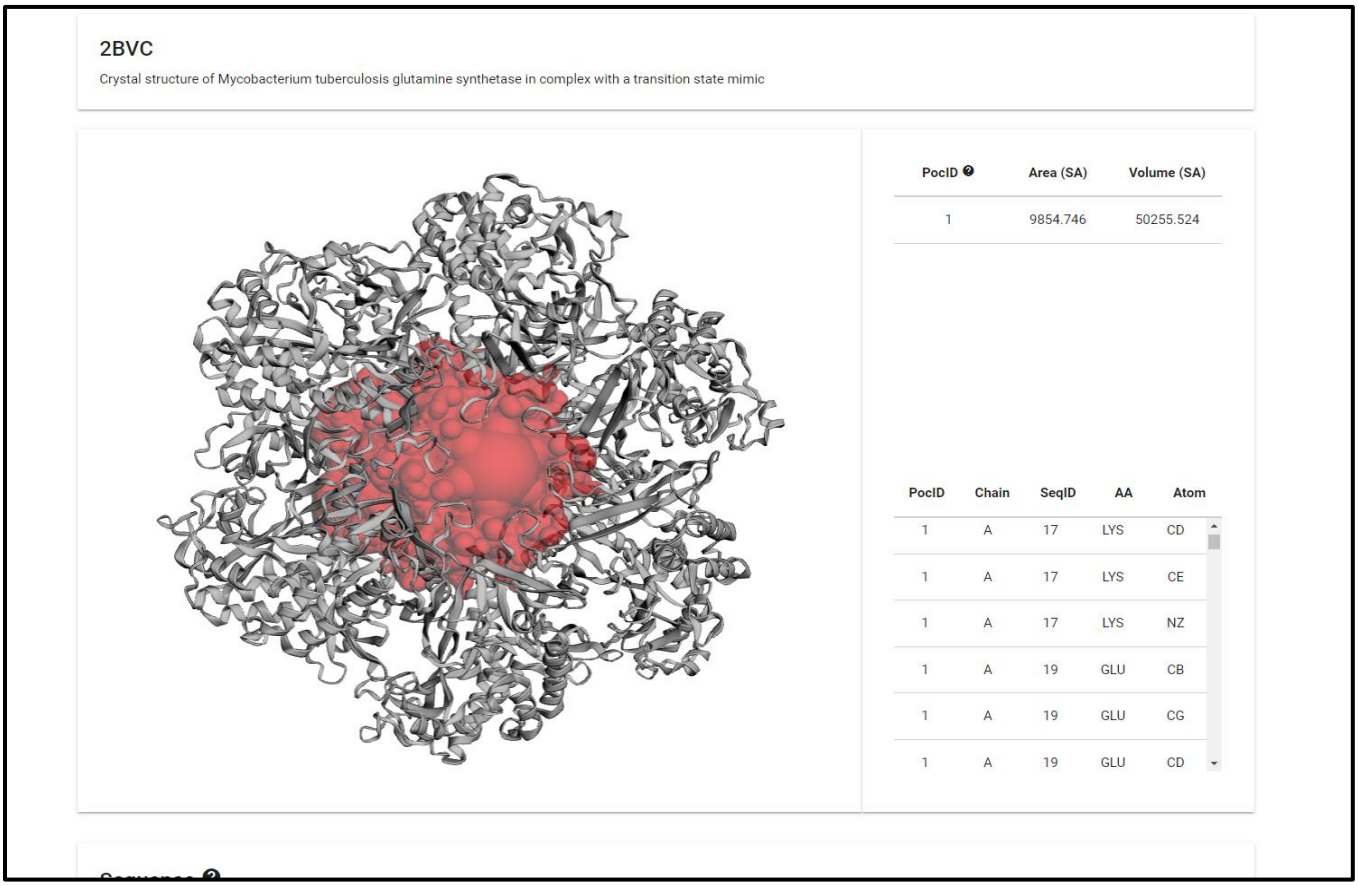


Fig3. Result of my query on CASTp

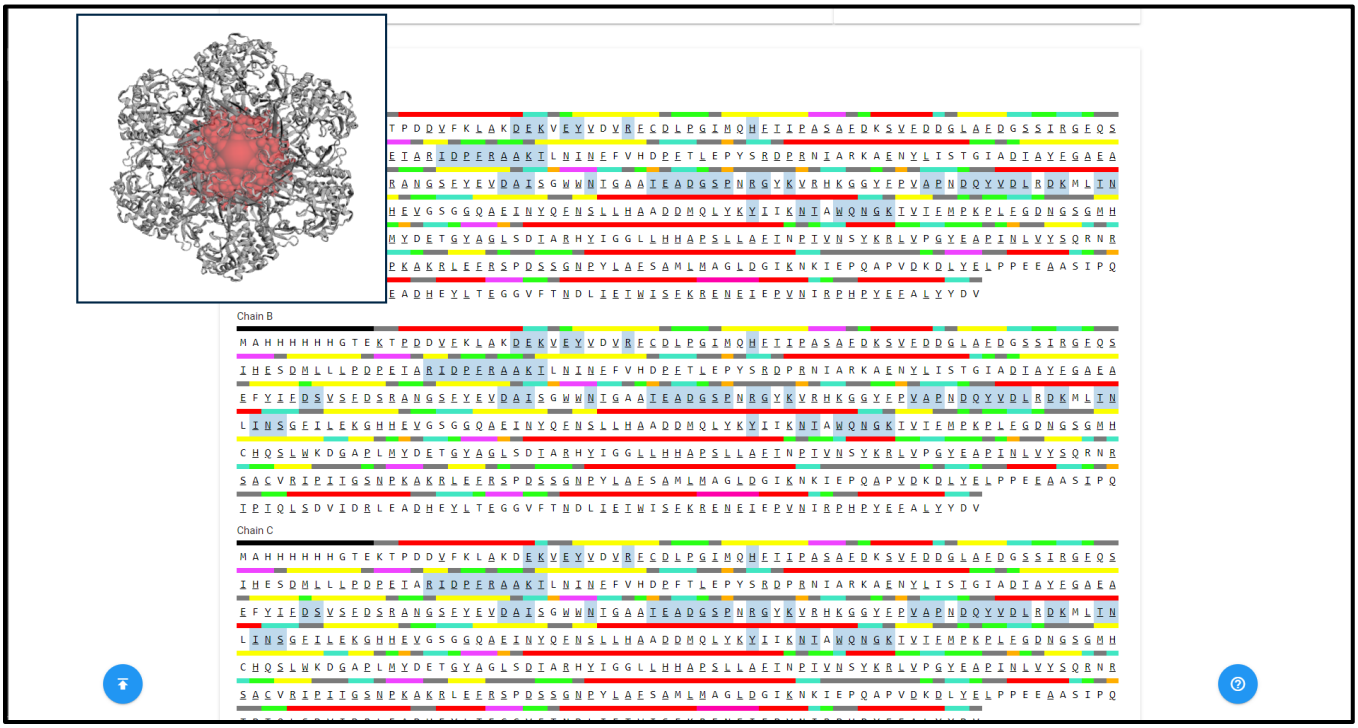


Fig4. Sequence data of the query on CASTp

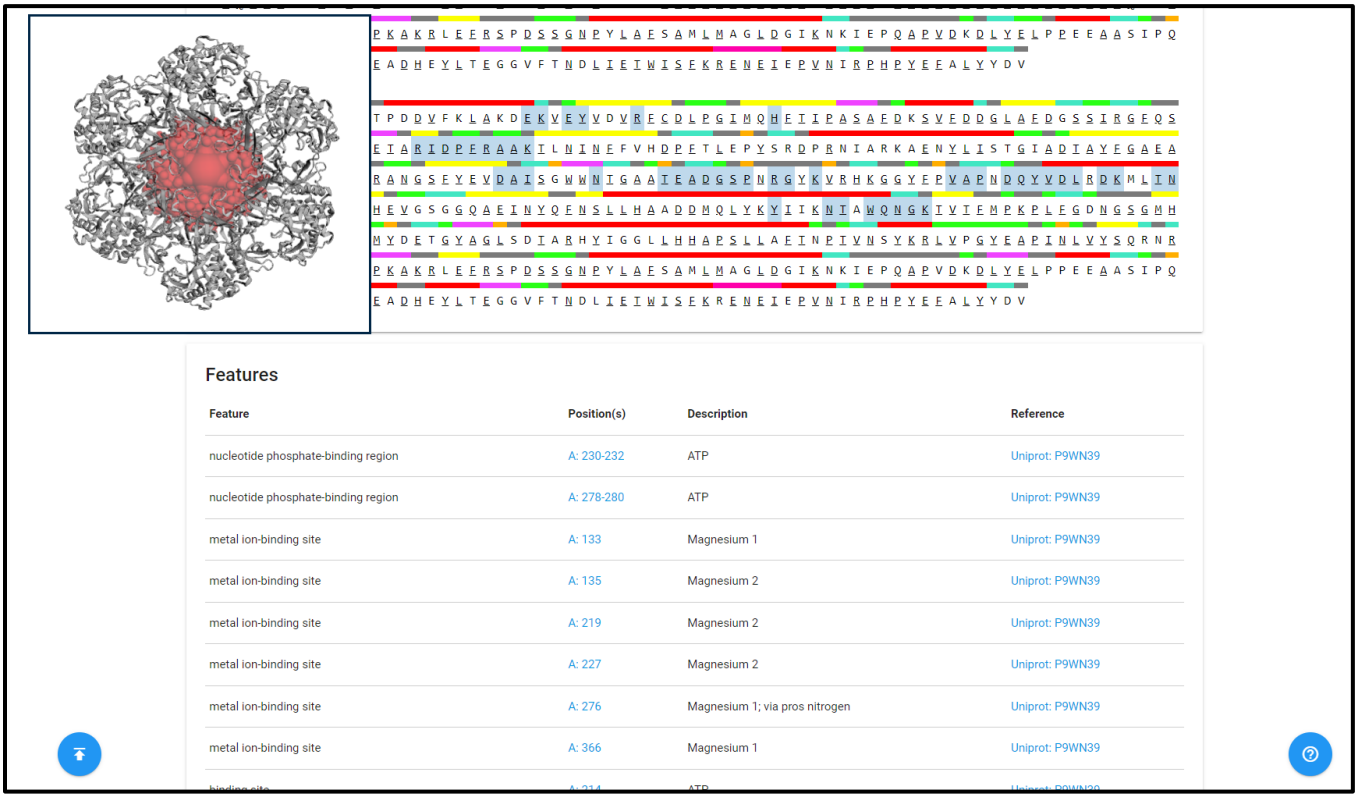


Fig5. Features data of my query on CASTp

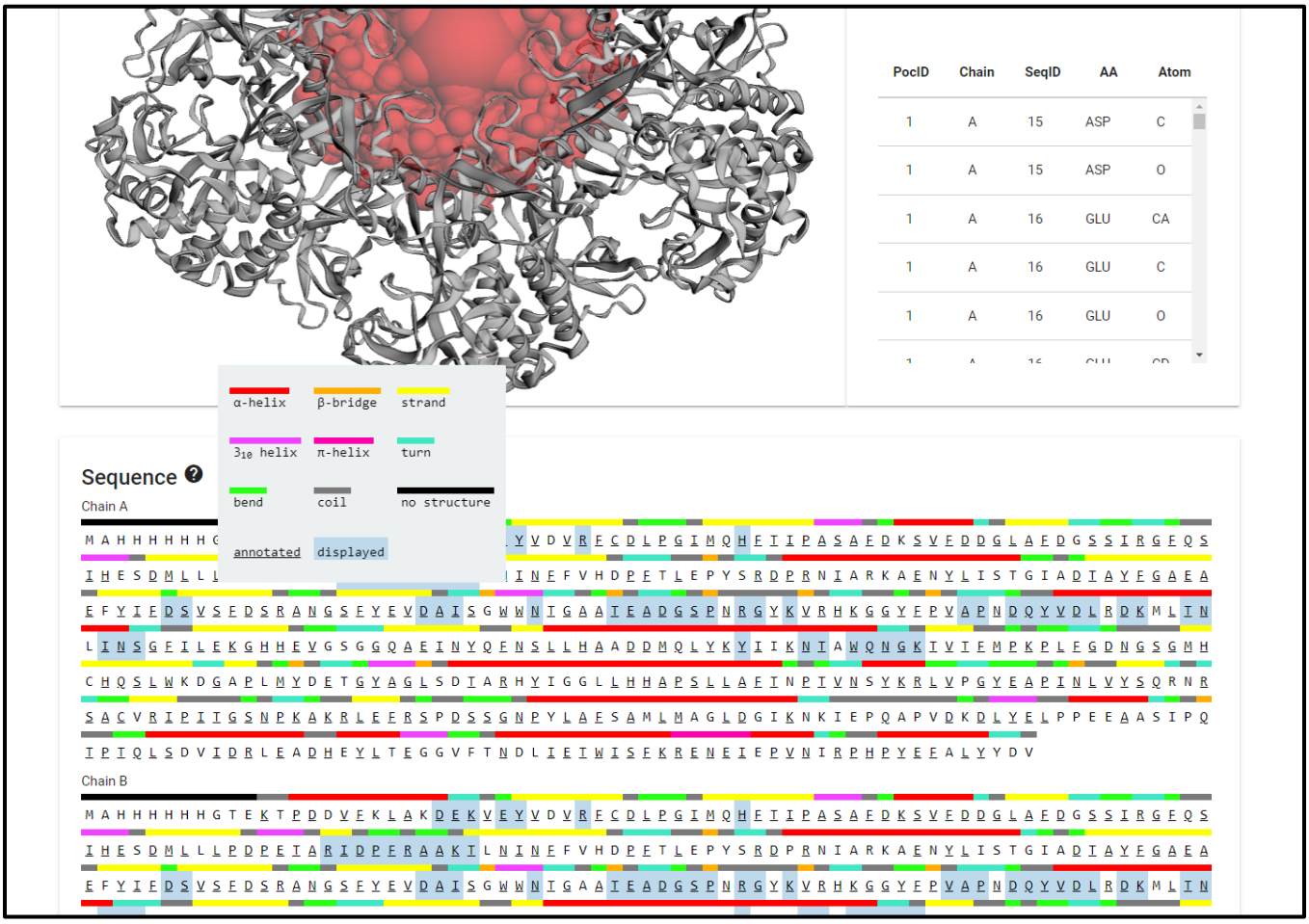


Fig6. Legend to refer to interpret data

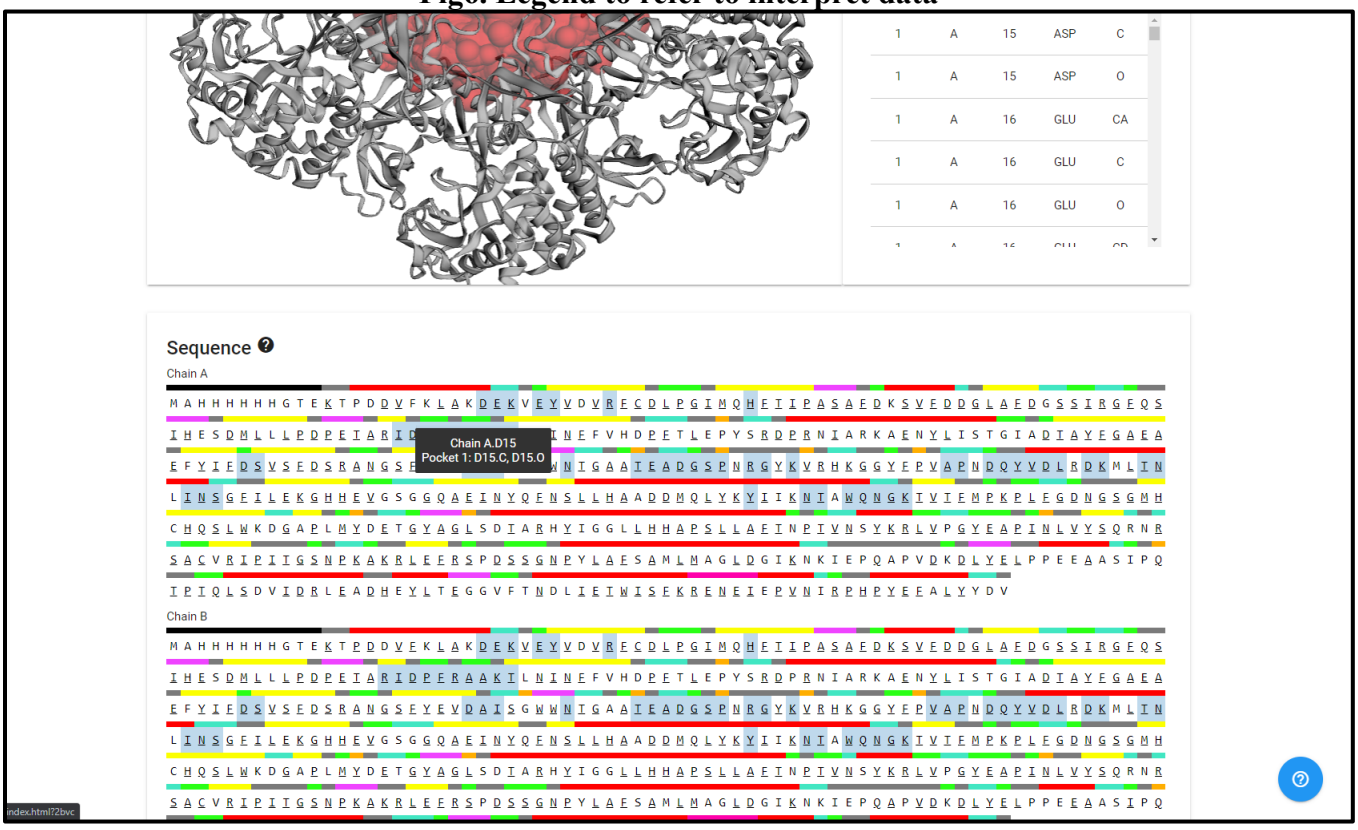


Fig7. Pocket info for Chain A.D15, Pocket 1: D15.C, D15.O

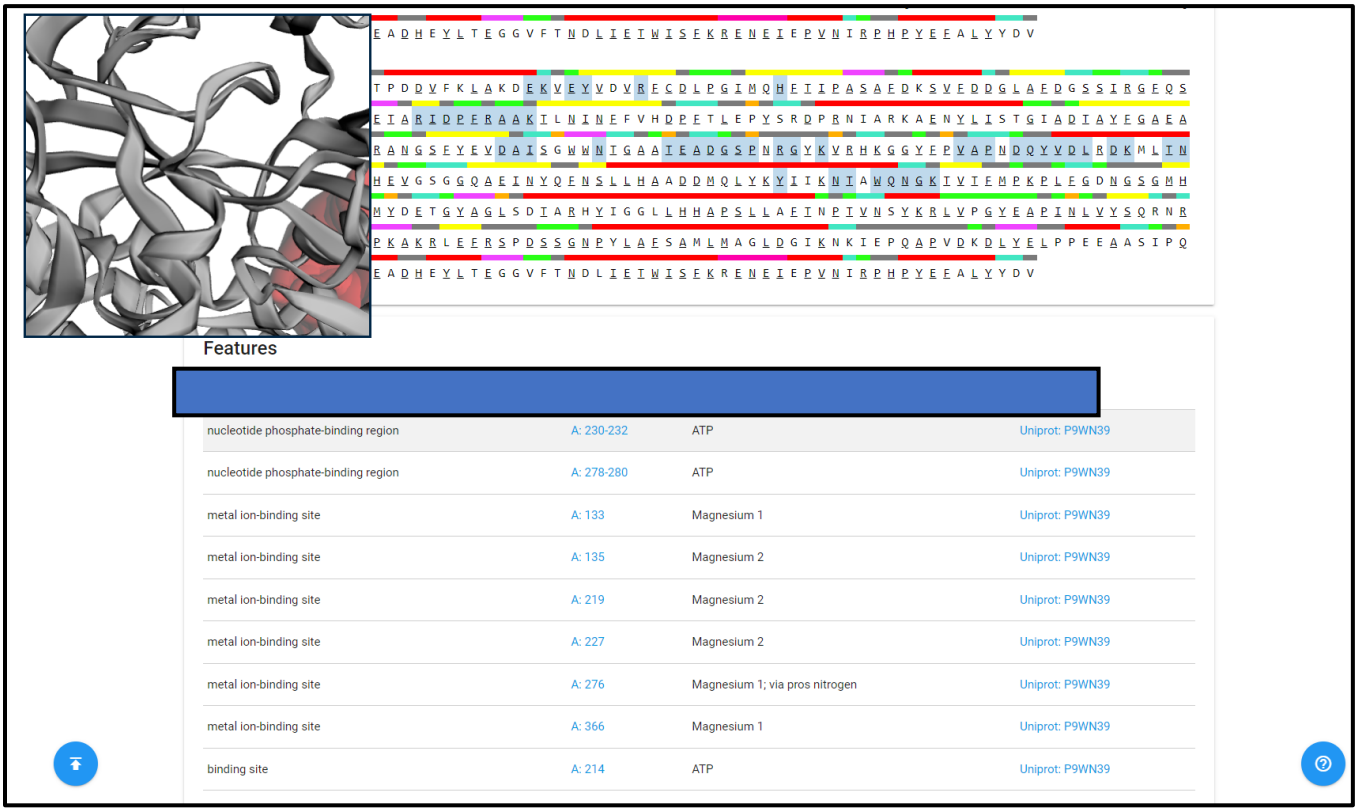


Fig8. Position in 3D space of nucleotide phosphate binding region A:230-232

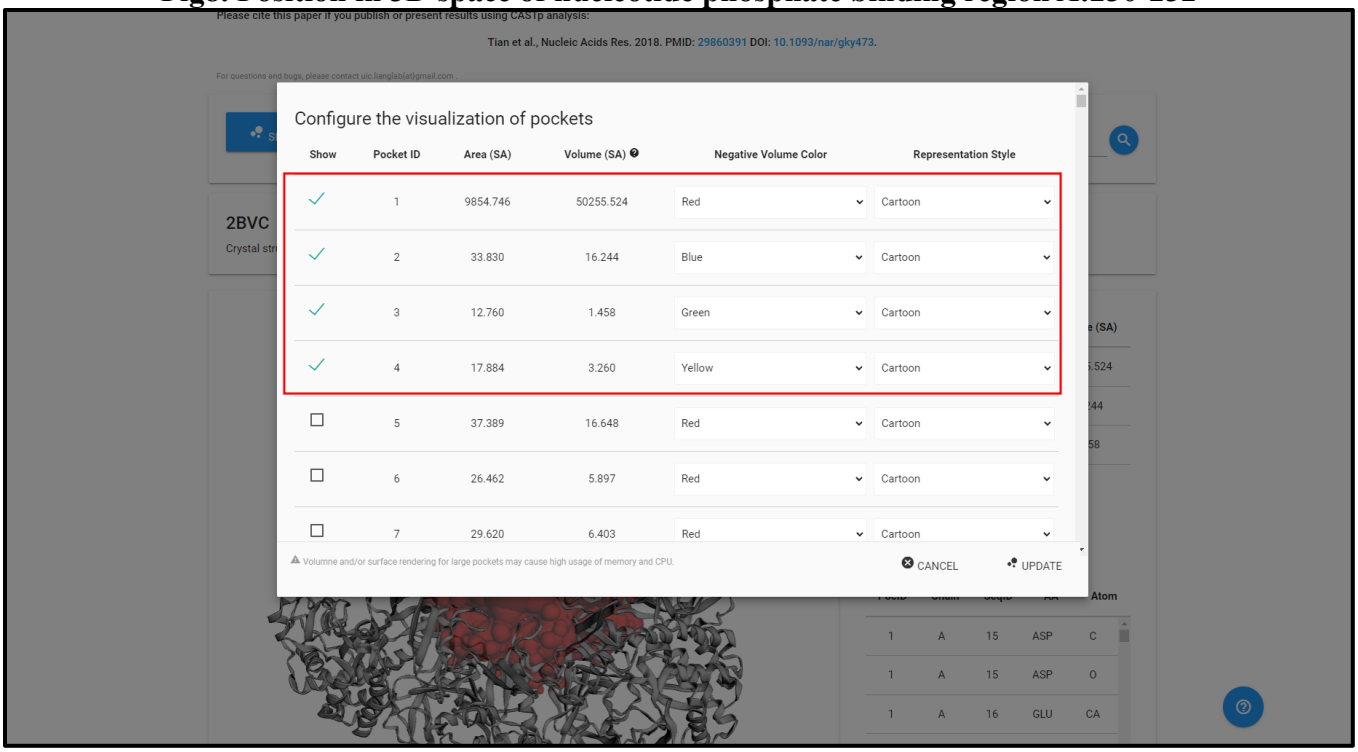


Fig9. Customizing the visualization of pocket in CASTp

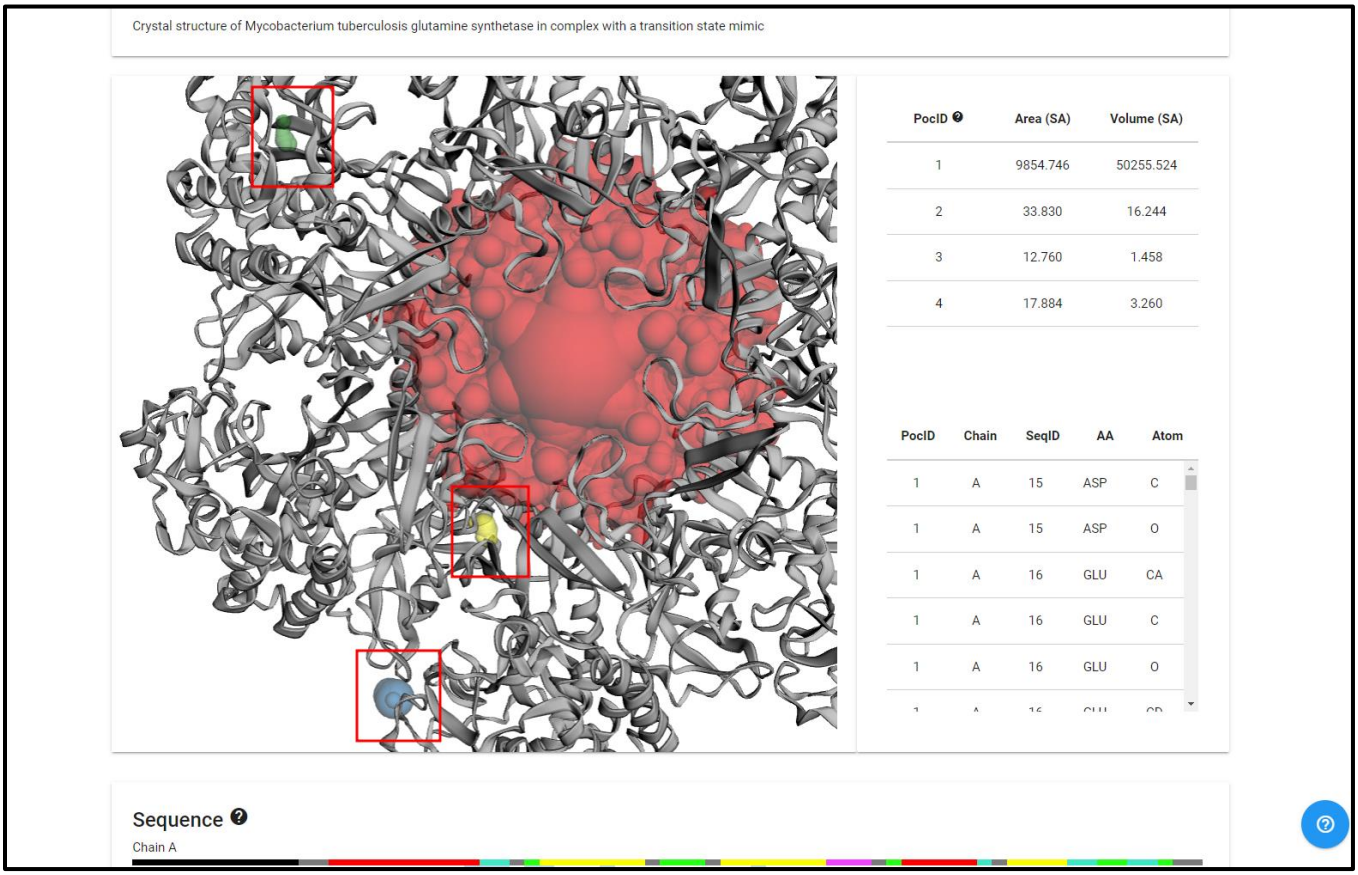


Fig10. Customized view of pockets for query in CASTp

Result:

1. The Results are categorically divided:
    a. **Structure:** Here we can customize the visualization of the protein structure.
    b. **Sequence:** Here we can see the amino acid sequences of the given protein structure.
    c. **Features:** Here we see the features of the provided protein structure
2. In the given structure the Area is 9854.746 SA and the volume is 50255.524 SA

Observation:

- Castp is a great tool to find out the specific sites on the protein in 3D space.
- The colors of the site can be changed according to us and be seen clearly and boldly.

References:

- Bank, R. C. S. B. P. D. (n.d.). *2BVC: Crystal structure of mycobacterium tuberculosis glutamine synthetase in complex with a transition state mimic*. RCSB PDB. Retrieved March 3, 2022, from https://www.rcsb.org/structure/2BVC
- CASTp 3.0: Computed atlas of surface topography of proteins. (n.d.). Retrieved March 3, 2022, from http://sts.bioe.uic.edu/castp/index.html?2bvc
- Cruzat, V., Macedo Rogero, M., Noel Keane, K., Curi, R., & Newsholme, P. (2018). Glutamine: Metabolism and immune function, supplementation and clinical translation. *Nutrients*, *10*(11), 1564. https://doi.org/10.3390/nu10111564

Date: 04-03-22

WEBLEM 6B
To predict binding pocket for Glycosylation sites in (query name) using NetOGlyc 4.0 Server

Introduction:

- Glycosylation is the most abundant and diverse posttranslational modification of proteins.
- The output conforms to the GFF version 2 format. For each input sequence the server prints a list of potential glycosylation sites, showing their positions in the sequence and the prediction confidence scores. Only the sites with scores higher than 0.5 are predicted as glycosylated and marked with the string "#POSITIVE" in the comment field.
- **Glutamine**
    - Glutamine is the most abundant and versatile amino acid in the body. In health and disease, the rate of glutamine consumption by immune cells is similar or greater than glucose. For instance, in vitro and in vivo studies have determined that glutamine is an essential nutrient for lymphocyte proliferation and cytokine production, macrophage phagocytic plus secretory activities, and neutrophil bacterial killing. Glutamine release to the circulation and availability is mainly controlled by key metabolic organs, such as the gut, liver, and skeletal muscles.

Methodology:

- Take a FASTA sequence from uniport
- Enter the sequence into the submission box
- Interpret the result according to the output tab

Observation:

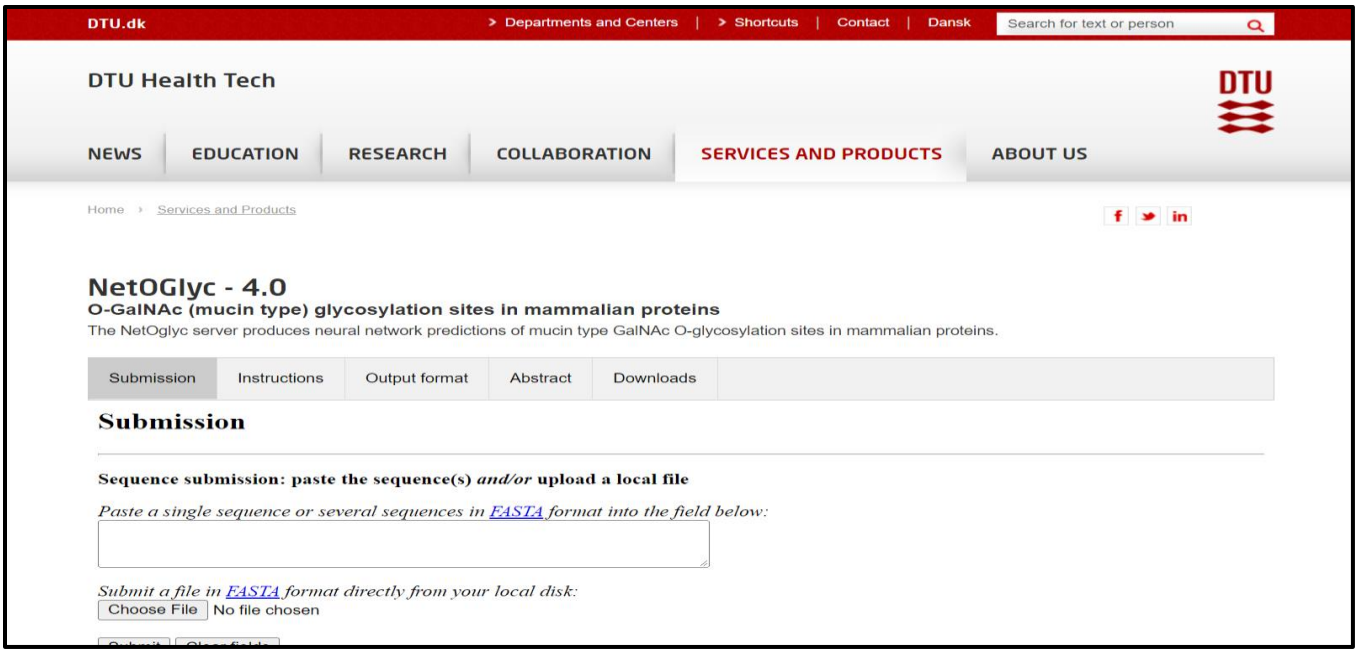


Fig1. Homepage of NetOGlyc – 1.0

```
>sp|P21980|TGM2_HUMAN Protein-glutamine gamma-glutamyltransferase 2 OS=Homo sapiens
OX=9606 GN=TGM2 PE=1 SV=2
MAEELVLERCDLELETNGRDHHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFS
VVTGPAPSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTPANAPIGLYRLSLE
ASTGYQGSSFVLGHFILLFNAWCPADAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPW
NFGQFEDGILDICLILLDVNPKFLKNAGRDCSRRSSPVYVGRVVSGMVNCNDDQGVLLGR
WDNNYGDGVSPMSWIGSVDILRRWKNHGCQRVKYGQCWVFAAVACTVLRCLGIPTRVVTN
YNSAHDQNSNLLIEYFRNEFGEIQGDKSEMIWNFHCWVESWMTRPDLQPGYEGWQALDPT
PQEKSEGTYCCGPVPVRAIKEGDLSTKYDAPFVFAEVNADVVDWIQQDDGSVHKSINRSL
IVGLKISTKSVGRDEREDITHTYKYPEGSSEEREAFTRANHLNKLAEKEETGMAMRIRVG
QSMNMGSDFDVFAHITNNTAEEYVCRLLLCARTVSYNGILGPECGTKYLLNLNLEPFSEK
SVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAERDLYLENPEIKIRILGEPKQK
RKLVAEVSLQNPLPVALEGCTFTVEGAGLTEEQKTVEIPDPVEAGEEVKVRMDLLPLHMG
LHKLVVNFESDKLKAVKGFRNVIIGPA
```

Fig2. FASTA Sequence for query Glutamine

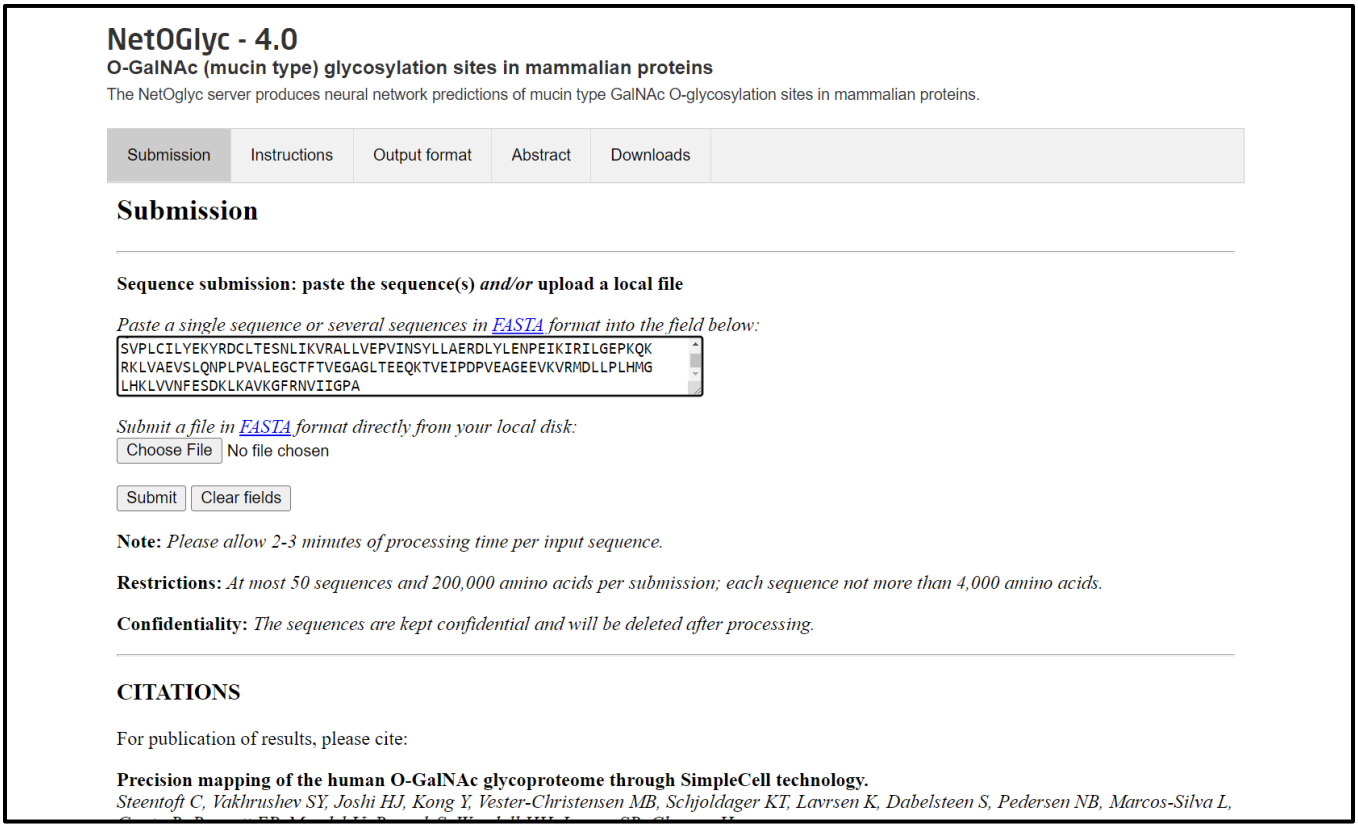


Fig3. Homepage of NetOGlyc with the FASTA sequence of query pasted in the search box

NetOGlyc - 4.0

O-GalNAc (mucin type) glycosylation sites in mammalian proteins

The NetOglyc server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins.

Submission | Instructions | Output format | Abstract | Downloads

NetOGlyc-4.0 Server Output - DTU Health Tech

```
##gff-version 2
##source-version NetOGlyc 4.0.0.13
##date 22-3-3
##Type Protein
#seqname        source  feature start   end     score   strand  frame   comment
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        16      16      0.0787162       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        23      23      0.137768        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        42      42      0.0592758       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        53      53      0.033419        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        56      56      0.0444774       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        58      58      0.0335661       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        60      60      0.203501        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        63      63      0.285525        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        68      68      0.352325        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        73      73      0.355121        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        89      89      0.051151        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        91      91      0.086065        .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        99      99      0.0260473       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        101     101     0.0235857       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        105     105     0.0208268       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        106     106     0.0140802       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        118     118     0.0247396       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        122     122     0.0498245       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        123     123     0.0239158       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        128     128     0.0189996       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        129     129     0.0207258       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        152     152     0.00668514      .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        162     162     0.0333635       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        171     171     0.0196308       .       .
SP_P21980_TGM2_HUMAN    netOGlyc-4.0.0.13       CARBOHYD        212     212     0.226317
```

Fig4. Result page of NetNGlyc showing the submission data for query

NetOGlyc - 4.0

O-GalNAc (mucin type) glycosylation sites in mammalian proteins

The NetOglyc server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins.

Submission | Instructions | Output format | Abstract | Downloads

Output format

DESCRIPTION

The output conforms to the GFF version 2 format. For each input sequence the server prints a list of potential glycosylation sites, showing their positions in the sequence and the prediction confidence scores. Only the sites with scores higher than **0.5** are predicted as glycosylated and marked with the string "#POSITIVE" in the comment field.

The example below shows the output for human granulocyte-macrophage colony-stimulating factor, The example below shows the output for human granulocyte-macrophage colony-stimulating factor, taken from the UniProt entry CSF2_HUMAN. Currently, 4 sites have been experimentally annotated for this protein, and NetOGlyc predicts that two of these are glycosylated. Additionally, it predicts an additional site is glycosylated at site 108. Occupancy of O-glycosylation sites can vary in-vivo depending on the cells that are expressing the protein. The interactions between sites of initial O-Glycosylation with subsequent sites of glycosylation are yet to be fully elucidated, while our capability to precisely predict the substrate specificity of individual GalNAc-Ts remains limited. The combination of these factors mean that although NetOGlyc will attempt to predict individual sites of glycosylation, a safe interpretation of a positive prediction is that the protein in that local region is more likely to carry O-GalNAc modifications.

EXAMPLE OUTPUT

```
##gff-version 2
##source-version NetOGlyc 4.0.0.12
##date 13-7-15
##Type Protein
#seqname        source  feature start   end     score   strand  frame   comment
CSF2_HUMAN      netOGlyc-4.0.0.12       CARBOHYD        5       5       0.04656 .       .
```

Fig5. Output format for NetOGlyc used to refer to interpret the results of our query

Result:

- After submitting the FASTA sequence for query Glutamine it gives 3 potential glycolysis sites.
- This interpretation can be made by looking at the scores higher than 0.5 and a comment that has #POSITIVE

Conclusion:

- To conclude NetOGlyc is a great source to gather data about the potential Glycolysis sites quickly and accurately
- We can submit one or multiple sequences upto 50 sequences and 200,000 amino acids at once which makes the processing of sequences more efficient

References:

- Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Bennett EP, Mandel U, Brunak S, Wandall HH, Levery SB, Clausen H. EMBO J, 32(10):1478-88, May 15, 2013. (doi: 10.1038/emboj.2013.79. Epub 2013 Apr 12)
- Uniprot. (n.d.). Retrieved March 3, 2022, from https://www.uniprot.org/uniprot/P21980.fasta
- Cruzat, V., Macedo Rogero, M., Noel Keane, K., Curi, R., & Newsholme, P. (2018). Glutamine: Metabolism and immune function, supplementation and clinical translation. *Nutrients*, *10*(11), 1564. https://doi.org/10.3390/nu10111564

Date: 04-03-22

WEBLEM 6C
To predict binding pocket for Phosphorylation site in (query name) using NetPhos 3.1 server

Introduction:

- Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes. How is specificity in substrate recognition and phosphorylation by protein kinases achieved? In addition, serine and threonine residues in p300/CBP that can be modified by O-linked glycosylation with N-acetylglucosamine are identified. Glycosylation may prevent phosphorylation at these sites, a mechanism named yin-yang regulation.
- **Glutamine**
    - Glutamine is the most abundant and versatile amino acid in the body. In health and disease, the rate of glutamine consumption by immune cells is similar or greater than glucose. For instance, in vitro and in vivo studies have determined that glutamine is an essential nutrient for lymphocyte proliferation and cytokine production, macrophage phagocytic plus secretory activities, and neutrophil bacterial killing. Glutamine release to the circulation and availability is mainly controlled by key metabolic organs, such as the gut, liver, and skeletal muscles.

Methodology:

- Take FASTA sequence from uniprot
- Enter it in the submission box for NetPhos
- Interpret the result according to the output page of NetPhos

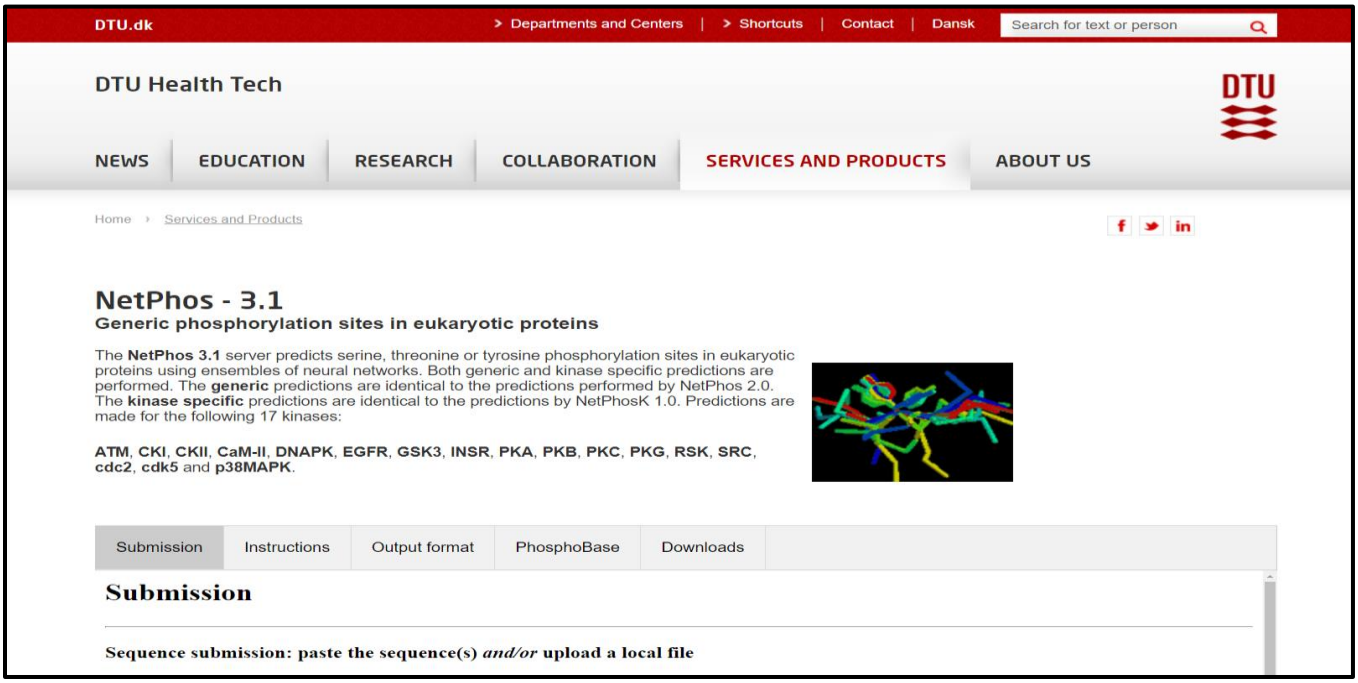


Fig1. Homepage of NetPhos 3.1

```
>sp|P21980|TGM2_HUMAN Protein-glutamine gamma-glutamyltransferase 2 OS=Homo sapiens
OX=9606 GN=TGM2 PE=1 SV=2
MAEELVLERCDLELETNGRDHHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFS
VVTGPAPSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTPANAPIGLYRLSLE
ASTGYQGSSFVLGHFILLFNAWCPADAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPW
NFGQFEDGILDICLILLDVNPKFLKNAGRDCSRRSSPVYVGRVVSGMVNCNDDQGVLLGR
WDNNYGDGVSPMSWIGSVDILRRWKNHGCQRVKYGQCWVFAAVACTVLRCLGIPTRVVTN
YNSAHDQNSNLLIEYFRNEFGEIQGDKSEMIWNFHCWVESWMTRPDLQPGYEGWQALDPT
PQEKSEGTYCCGPVPVRAIKEGDLSTKYDAPFVFAEVNADVVDWIQQDDGSVHKSINRSL
IVGLKISTKSVGRDEREDITHTYKYPEGSSEEREAFTRANHLNKLAEKEETGMAMRIRVG
QSMNMGSDFDVFAHITNNTAEEYVCRLLLCARTVSYNGILGPECGTKYLLNLNLEPFSEK
SVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAERDLYLENPEIKIRILGEPKQK
RKLVAEVSLQNPLPVALEGCTFTVEGAGLTEEQKTVEIPDPVEAGEEVKVRMDLLPLHMG
LHKLVVNFESDKLKAVKGFRNVIIGPA
```

Fig2. FASTA Sequence for Query Glutamine

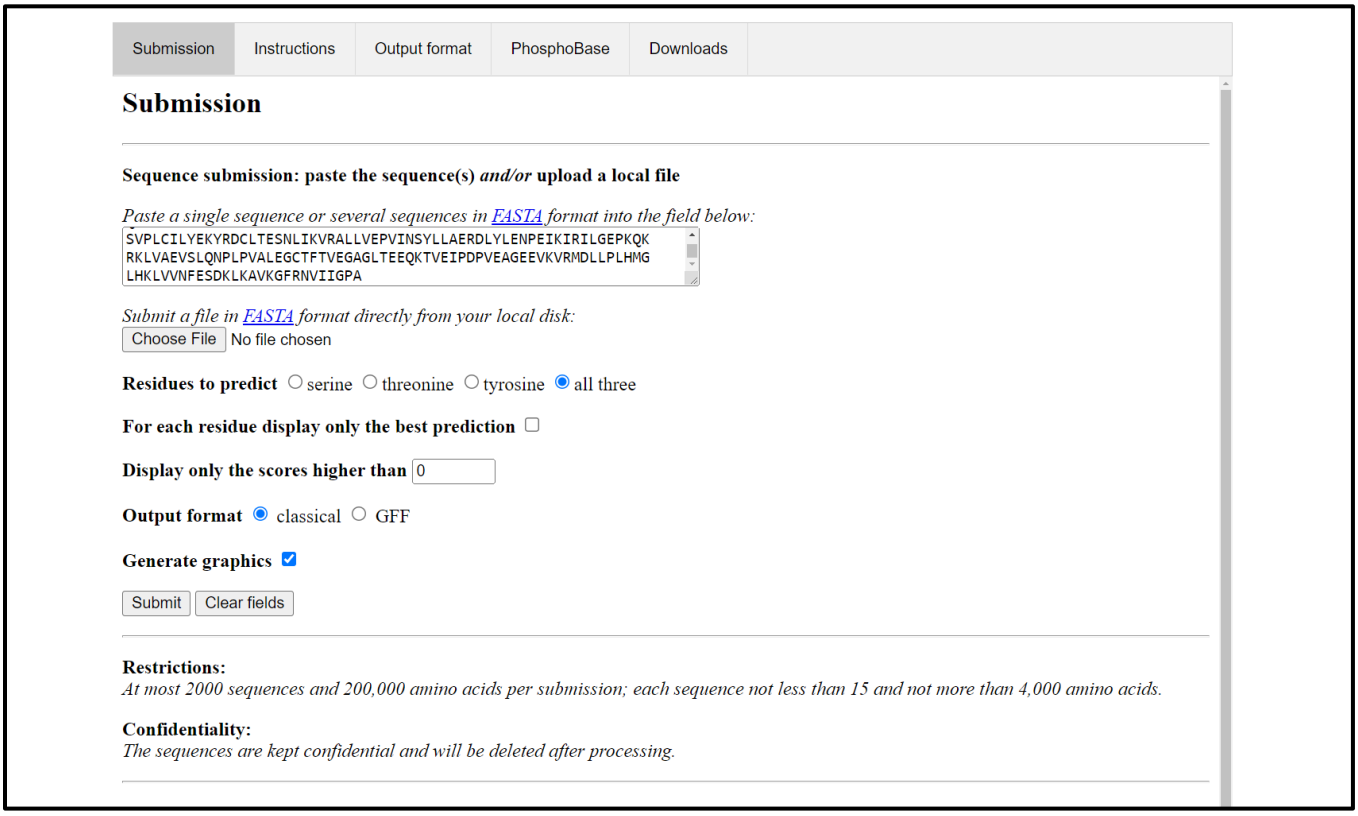


Fig3. FASTA sequence pasted in the submission box on NetPhos 3.1

Submission | Instructions | Output format | PhosphoBase | Downloads

NetPhos-3.1 Server Output - DTU Health Tech

```
>sp_P21980_TGM2_HUMAN   687 amino acids
#
# netphos-3.1b prediction results
#
# Sequence                 # x   Context     Score   Kinase    Answer
# ---------------------------------------------------------------------
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.486   CKII        .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.461   GSK3        .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.432   CaM-II      .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.411   cdc2        .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.367   CKI         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.342   DNAPK       .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.302   p38MAPK     .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.267   PKC         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.240   ATM         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.237   PKG         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.192   RSK         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.171   cdk5        .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.088   PKA         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.085   PKB         .
# sp_P21980_TGM2_HUMAN     16 T   LELETNGRD   0.029   unsp        .
#
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.568   CKII       YES
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.452   CaM-II      .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.448   GSK3        .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.388   cdc2        .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.361   CKI         .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.343   PKG         .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.342   DNAPK       .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.325   p38MAPK     .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.260   ATM         .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.224   PKA         .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.201   RSK         .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.153   cdk5        .
# sp_P21980_TGM2_HUMAN     23 T   RDHHTADLC   0.090   PKC         .
```

Fig4. Server output for the FASTA sequence showing amino acid prediction results

```
%1  .....S.T....T....S....T.................T.........   #    100
%1  S....T...........S...ST.....S...................Y.   #    150
%1  .S......Y..T........S.............................   #    200
%1  ..........S..SS..Y.....S...................Y....S   #    250
%1  ..S..............................T........T...T.   #    300
%1  ..S.........................S.....................   #    350
%1  .........T....S...Y...............ST..............   #    400
%1  ..........S...S...S.......S..S...........T......SS   #    450
%1  ..............................................T....   #    500
%1  ..Y...........S............Y.........S..S.........   #    550
%1  .......S...............S........Y..................   #    600
%1  ......................T......T.....................   #    650
%1  ....................................
```

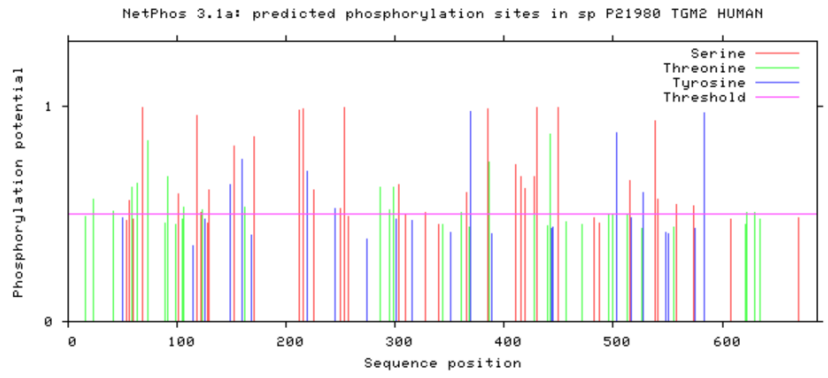


Fig5. Predicted phosphorylation sites in the sequence according to NetPhos 3.1

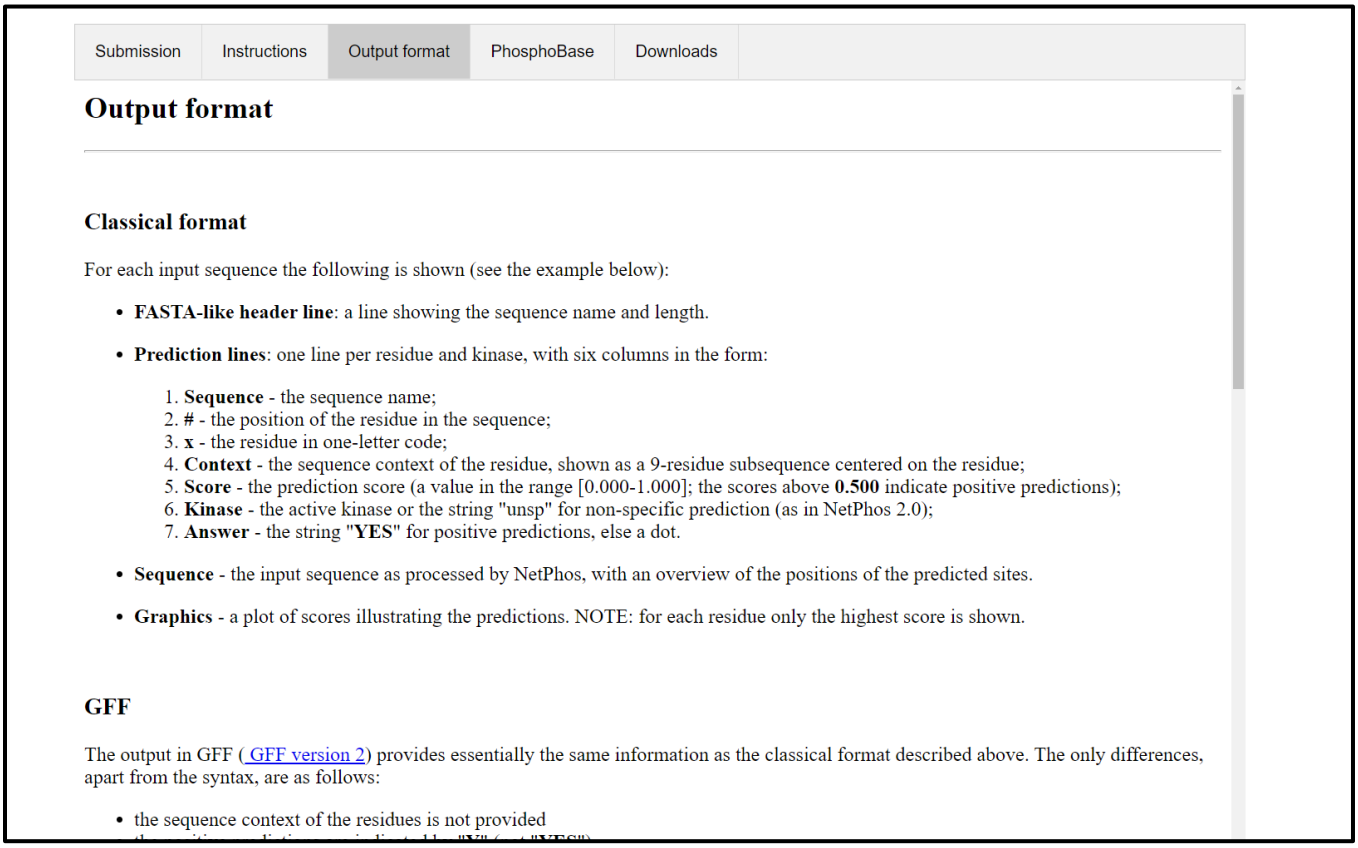


Fig6. Output format of NetPhos 3.1 used to interpret results for query accordingly

Result:

- After submitting the result we can determine that there are 56 potential phosphorylation sites in the given protein sequence
- This interpretation was made on the basis that 56 results have a score about 0.5 and answer is YES
- The active kinase is also given and for non-specific prediction it is "**unsp**"

Observation:

- NetPhos is a great tool to find the phosphorylation sites of a given protein sequence.
- Multiple sequences can be submitted at once and the results can be interpreted in bulk making it easier to interpret.

References:

- Uniprot. (n.d.). Retrieved March 3, 2022, from https://www.uniprot.org/uniprot/P21980.fasta
- Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. Blom, N., Gammeltoft, S., and Brunak, S. Journal of Molecular Biology: 294(5): 1351-1362, 1999.
- Cruzat, V., Macedo Rogero, M., Noel Keane, K., Curi, R., & Newsholme, P. (2018). Glutamine: Metabolism and immune function, supplementation and clinical translation. *Nutrients*, *10*(11), 1564. https://doi.org/10.3390/nu10111564

DATE: 18-03-22

WEBLEM 7

Introduction to Structural Blast – VAST and DALI

The protein structures that populate the PDB have provided crucial insights at the atomic level as to the molecular mechanisms that underlie protein function. Indeed, structural studies have had, and continue to have, a significant and sometimes revolutionary impact in all areas of biology. However, structural biology has tended to focus on single proteins or biological systems and, despite significant advances in the general area of structural bioinformatics, the horizontal integration of the vast quantity of structural information available in the PDB has had little or no impact in the larger biological community. This is in contrast to protein sequence information, which is more routinely, automatically and broadly used. Given that structure is more conserved than sequence, structural similarity has the potential to yield a great deal of functional information that sequence relationships cannot provide and to identify relationships between many more pairs of proteins. In this article we argue that the exploitation of statistical and machine learning techniques combined with the vast amount of high-throughput experimental data constantly being generated enable a significant expansion in the scale and diversity of application of structural information to biological problems.

The ultimate potential impact of both global and local structural relationships in inferring function is highlighted by the observation that, given a suitably "loose" definition of structural similarity, the repertoire of structures currently in the structural databases is nearly complete at the domain level. Thus, it can be expected that most newly solved protein structures will have both near and remote structural neighbors which can provide clues as to their function. Programs such as BLAST use local sequence relationships to quickly scan sequence databases. Since structure-based scans of protein structural databases can be carried out very quickly with current technology (typically minutes for a database of tens of thousands of structures), a similar strategy can be used for structural relationships as well, essentially defining a "structural BLAST"

Comparative analyses of protein sequences and structures play a fundamental role in understanding proteins and their functions. Assuming an evolutionary continuity of structure and function, describing the structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered proteins. The most widespread purpose of structural alignment has been to identify homologous residues (encoded by the same codon in the genome of a common ancestor). Mutations manifest in plastic deformations, shifts and rotations of the secondary structure elements (SSEs). A wide spectrum of structural alignment methods exist, which differ in their treatment of structural variations, scoring functions and optimization algorithms.

There are aware of half a dozen web servers that provide structure comparisons against the current weekly updated Protein Data Bank (PDB). Each server is unique because they employ different structure comparison methods.

VAST:

The VAST search database and database of precomputed structure alignments have been maintained as complete and redundant collections since their launch, with automated updates occurring on a weekly basis. This was made possible by implementing a fast heuristic that uses a model for the statistical significance of initial alignments of secondary structure vectors (which can be computed quickly), so that the database searches can avoid costly alignment refinements for the large majority of insignificant and uninteresting similarities. The drawbacks are that a heuristic will miss some potentially interesting similarities. The VAST algorithm will not, for example, report similarities between structures deemed to have secondary structure elements. Searches for structural similarity can and should be complemented with searches for sequence similarity, as flexibility of molecular structure and limitations of the structure comparison method may preclude the detection of matches between structures of homologous polypeptides. In general, though, structure comparison methods will pick up many subtle similarities that evade detection by sequence

comparison strategies, and there is no natural cutoff point for a ranked list of similar structures, unlike in the sequence comparison scenario, where matches to non-homologous gene products are considered accidental and uninformative, for the most part.

Results computed by the VAST algorithm have been compared against other approaches a number of times. Although there are subtle differences in retrieval sensitivity and alignment accuracy, it appears fair to state that the large majority of extensive structural similarities, which are indicative of common evolutionary descent and could be used to infer functional similarities, are reported by VAST (and by most if not all of the alternative approaches to detect common substructures).

As structure similarity search strategies have been developed to also detect distant relationships that might not be evident from sequence analysis, most if not all of the current approaches have been implemented so that they use a single protein molecule or rather a single domain as the unit of comparison. This has been true for VAST, in particular. However, the Protein Data Bank is continuing to accumulate structures of larger macromolecular complexes and has started to provide data on what constitutes functionally or biologically relevant macromolecular complexes or biological assemblies. Such assemblies range from simple homo-oligomers to intricate arrangements of many different components, revealing details on specific molecular interactions and on how these might constrain sequence variation. A small number of approaches have been published in the past few years that examine structural similarity of macromolecular complexes. Here we present a simple strategy that builds on the existing database of pairwise structure alignments computed by VAST and supports the first (to our knowledge) comprehensive and regularly updated collection of macromolecular complex similarities.

VAST+ as an extension to existing protein structure comparison

As information characterizing biological assemblies in macromolecular structure data has become available, it seemed that the biological assembly would be a convenient and informative unit of comparison between individual entries in the structure database. If the goal is to list structures most similar to any particular query, one would have to consider that the query itself may contain a macromolecular complex with a given stoichiometry, and that matching complexes with matching stoichiometry might be more informative ‘structure neighbors’ than, for example, the structures that happen to contain molecules with the strongest local similarity to the query, irrespective of the context.

VAST+ builds on the existing VAST database to generate such a report of structure neighbors. Its goal is to find the largest set of pairs of matching macromolecules between two biological assemblies and to characterize that match and compute instructions for a global superimposition that can be used to visualize the structural similarity. For each pair of structures in MMDB, VAST+ examines pre-computed structure alignments stored in the VAST database that were computed for the full-length protein molecule components of the default biological assemblies. If such pairwise alignments are found, the alignments between individual protein components of the biological assemblies are compared with each other for compatibility, and compatible/matching alignments are clustered into sets of alignments that together constitute a biological assembly match. Pairwise alignments are compatible (i) if they do not share the same macromolecules, i.e. a protein molecule from one assembly cannot be aligned to two molecules from the other assembly at the same time and (ii) if they generate similar instructions (spatial transformation matrices) for the superpositions of coordinate sets. A simple distance metric can be used to compare transformation matrices and it lends itself to cluster alignment sets efficiently.

Each set of compatible pairwise alignments can be characterized by (i) the number of pairwise matches, i.e. the total number of pairs of protein molecules from the query and subject biological assemblies, that are simultaneously aligned with each other; (ii) the RMSD of the superposition obtained from considering all alignments in the set; (iii) the total length of all pairwise alignments, i.e. the total number of amino acids that are aligned in 3D space; and (iv) percentage of identical residues in the alignments. For each pairwise comparison of two biological assemblies, only the match with the highest number of aligned molecules and the highest number of aligned residues is recorded and reported.

Currently, 53% of polypeptide-containing structures in MMDB have >1 polypeptide chain. The histogram plotted in Figure 1 breaks down the numbers by oligomer size and indicates that large fractions of the oligomeric assemblies have, in general, structure neighbors that match the entire assemblies. It should be noted that the fractions might be somewhat exaggerated, as exact duplicates of a structure would be counted as biological assembly matches, and no attempt was made to remove redundant structures or classify biological assembly matches as informative versus uninformative.

DALI:

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

User can perform three types of database searches:

- Heuristic PDB search - compares one query structure against those in the Protein Data Bank
- Exhaustive PDB25 search - compares one query structure against a representative subset of the Protein Data Bank
- Hierarchical AF-DB search - compares one query structure against a species subset of the AlphaFold Database

There are two types of structure comparisons of user selected structures:

- Pairwise structure comparison - compares one query structure against those specified by the user
- All against all structure comparison - returns a structural similarity dendrogram for a set of structures specified by the user

DESCRIPTION OF THE SERVER

Inputs

The input to the server is one or two protein structures in PDB format. The query structure can be specified as a PDB identifier plus chain identifier, or a PDB file uploaded by the user. There are three cross-linked query forms for the Dali server, Dali Database and pairwise comparison, respectively. For example, the entry point to the Dali server is http://ekhidna.biocenter.helsinki.fi/dali_server.

All backbone atoms (N, CA, C, O) are required and the minimum chain length is 30 amino acids. Backbone atoms may be reconstructed from a CA trace using the MaxSprout server at.

External links to the Dali database should use , where 1nnn represents a PDB identifier and chainid is optional. Meta-servers may link to, which directly returns the match list and alignment data as plain text.

Processing

Queries to the Dali Database and pairwise comparison are processed interactively; the result is usually returned within a minute. The Dali server processes up to eight PDB searches in parallel, others are queued. Most PDB- search queries are processed in less than an hour. Results are stored on the server for two weeks. The results of identical queries are retrieved instantly from cache.

The Dali server and Dali database return only the best match of the query to each PDB structure. The pairwise comparison returns also suboptimal matches. The pairwise comparison is based on a systematic branch-and- bound search that returns non-overlapping solutions in decreasing order of alignment score. Suboptimal matches can be of interest in cases of internal symmetries or repeated domains.

Dali Database is updated twice a year and contains precomputed structural alignments of PDB90 against the full PDB. The query structure is mapped to the closest representative in PDB90 and the structure comparison scores are recomputed using the transitive alignment via the representative.

The Dali server aims to retrieve a list of 500 structural neighbors of the query structure with the highest Z-scores. Most query structures have strong similarity to a structure already in the PDB. We use fast filters to identify a shortlist of about 100 promising candidates. If these produce strong matches, the search proceeds

by walking. Otherwise, the query structure is compared with PDB90 in one versus all fashion, followed by a walk to collect matches to redundant PDB structures (which are over 90% sequence identical to PDB90 representatives).

Walking selects targets for structural comparison from the neighbours of neighbours found so far. The second shell of neighbors is known because all structures in the PDB are stored in a precomputed network of similarities. The pairwise alignments (Q,P) and (P,R) induce a transitive alignment (Q,R), which is used as the starting point of refinement rather than optimizing the alignment from scratch. There are many possible choices of intermediate structure P en route from Q to R. We select the 'high road', in other words, the minimum of the Z-scores Z(Q,P) and Z(P,R) should be as high as possible. The 'high road' may change as more structures are added to the first neighbour shell. To avoid redundant comparisons, we only test induced alignments which are longer than previously obtained ones. When the alignment (Q,R) has been refined, R is added to the first neighbour shell. The walk ends when either there are no new neighbours in the second shell, a specified number of hits (1000) have been reported, or a maximum number of comparisons (1000) have been performed.

Outputs

The Dali server, Dali Database and pairwise comparison use a common output format and share interactive analysis tools.

The result consists of (i) a list of structural neighbours, ranked by Z-score, and (ii) the alignment data. The results are presented as plain text for downloading by downstream application, and as hypertext for interactive analysis. The default results page reports the top 500 matches to all chains in the PDB. A subset of matches to PDB90, filtered at 90% sequence identity, is provided for convenience.

Selected subsets of matches can be visualized (i) as multiple sequence alignments, or (ii) in multiple 3D superimposition. While sophisticated tools with integrated sequence alignment and structure superimposition views are available, we have chosen Jmol, an open source Java viewer for molecular graphics, because it was most easily accessible to the casual user. Each neighbour is aligned (superimposed) against the query structure in a star-like tree topology. Active sites can be recognized by clusters of conserved residues and ligands. Sequence and structure conservation are calculated within the selected subset of matches.

VAST and DALI are thus very useful structure similarity BLAST tool. VAST provides user with similar structures to their query along with its molecular components and chemicals and non-standard biopolymers, aligned sequences and 3D structure superimposition information which includes information regarding H-bonds, interactions, buried surface area, 2D interaction network and much more. DALI provides user with similar structures to their query along with its pairwise alignment, coordinates information, 3Dsuperimposition results. Describing the structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered proteins.

REFERENCES:

1. Dey, Fabian; Cliff Zhang, Qiangfeng; Petrey, Donald; Honig, Barry (2013). Toward a "Structural BLAST": Using structural relationships to infer function. Protein Science, 22(4), 359– 366. doi:10.1002/pro.2225
2. Madej, T.; Lanczycki, C. J.; Zhang, D.; Thiessen, P. A.; Geer, R. C.; Marchler-Bauer, A.; Bryant, S. H. (2014). MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Research, 42(D1), D297–D303. doi:10.1093/nar/gkt1208
3. Dali server. (n.d.). Ekhidna2.Biocenter.helsinki.fi. Retrieved March 14, 2022, from http://ekhidna2.biocenter.helsinki.fi/dali/
4. Holm, L.; Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Research, 38(Web Server), W545–W549. doi:10.1093/nar/gkq366

DATE: 19-03-22

WEBLEM 7a
VAST
(URL: https://www.ncdi.nlm.nih.gov/Structure/vastplus/vastplus.cgi)

AIM:

To perform structure BLAST for tubulin using VAST tool.

INTRODUCTION:

Tubulin is the protein that polymerizes into long chains or filaments that form microtubules, hollow fibers which serve as a skeletal system for living cells. Microtubules have the ability to shift through various formations which is what enables a cell to undergo mitosis or to regulate intracellular transport.

The computational detection of similarities between protein 3D structures has become an indispensable tool for the detection of homologous relationships, the classification of protein families and functional inference. Consequently, numerous algorithms have been developed that facilitate structure comparison, including rapid searches against a steadily growing collection of protein structures. To this end, NCBI's Molecular Modeling Database (MMDB), which is based on the Protein Data Bank (PDB), maintains a comprehensive and up-to- date archive of protein structure similarities computed with the Vector Alignment Search Tool (VAST). These similarities have been recorded on the level of single proteins and protein domains, comprising in excess of 1.5 billion pairwise alignments. VAST+, an extension to the existing VAST service, which summarizes and presents structural similarity on the level of biological assemblies or macromolecular complexes. VAST+ simplifies structure neighboring results and shows, for macromolecular complexes tracked in MMDB, lists of similar complexes ranked by the extent of similarity. VAST+ replaces the previous VAST service as the default presentation of structure neighboring data in NCBI's Entrez query and retrieval system.

METHODOLOGY:

1. Open homepage for VAST. (URL: https://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi)
2. Retrieve PDB ID for Albumin.
3. Search for similar structures on VAST for the PDB ID.
4. Observe and interpret the results.

OBSERVATION:

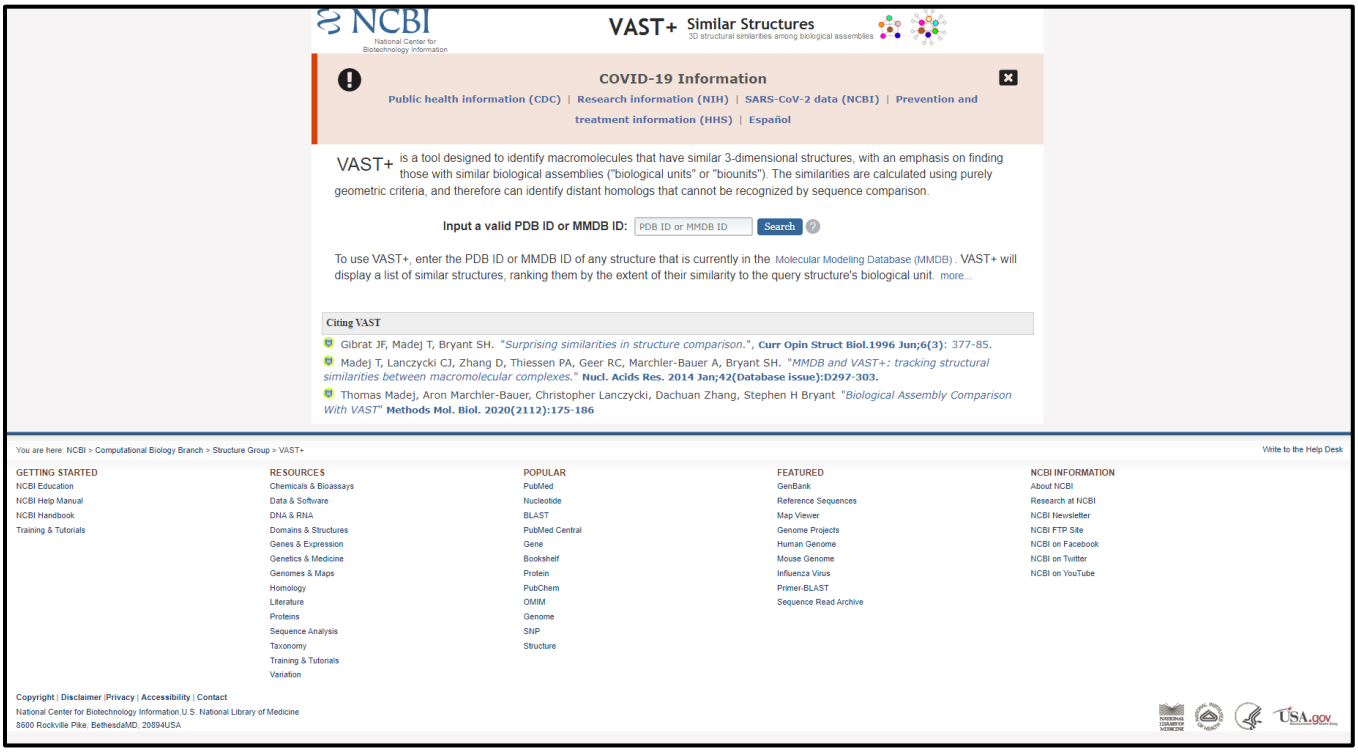


Fig1. Homepage for VAST

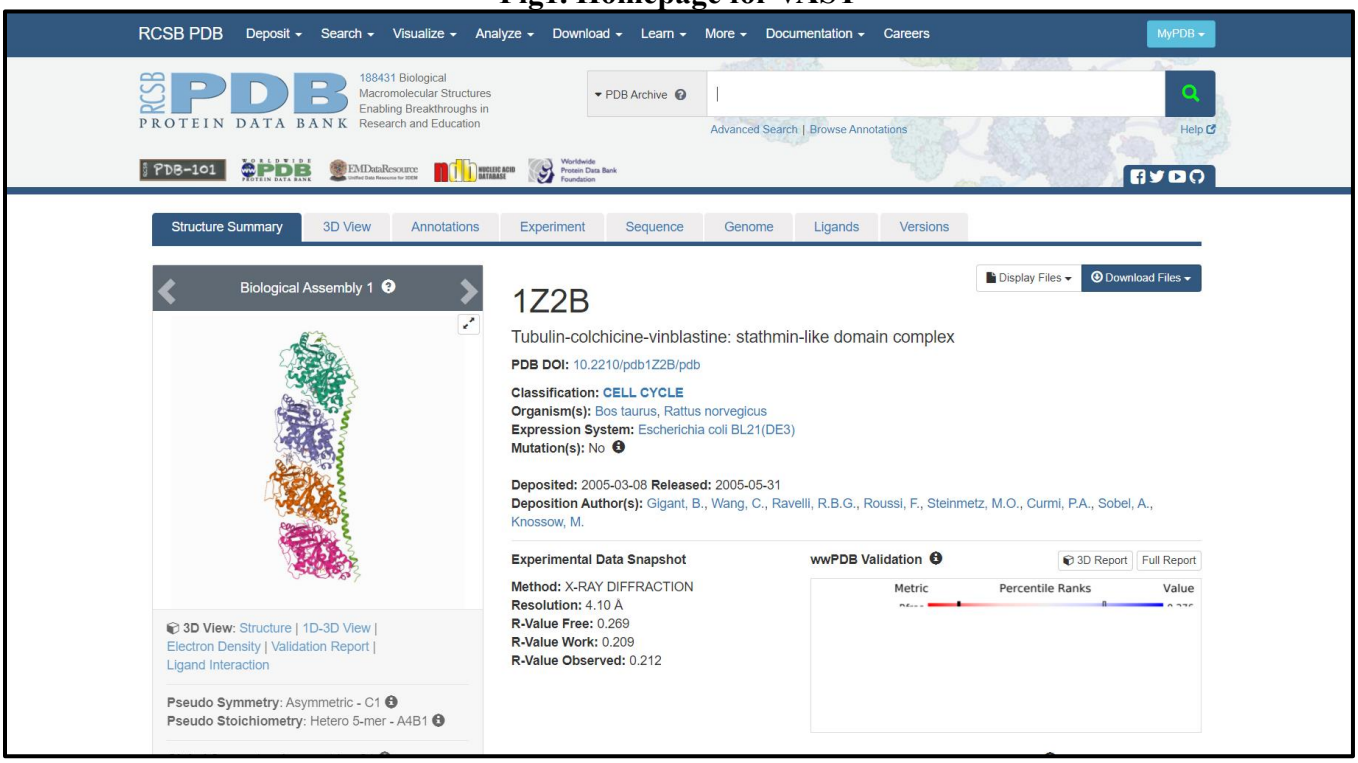


Fig2. Tubulin PDB structure

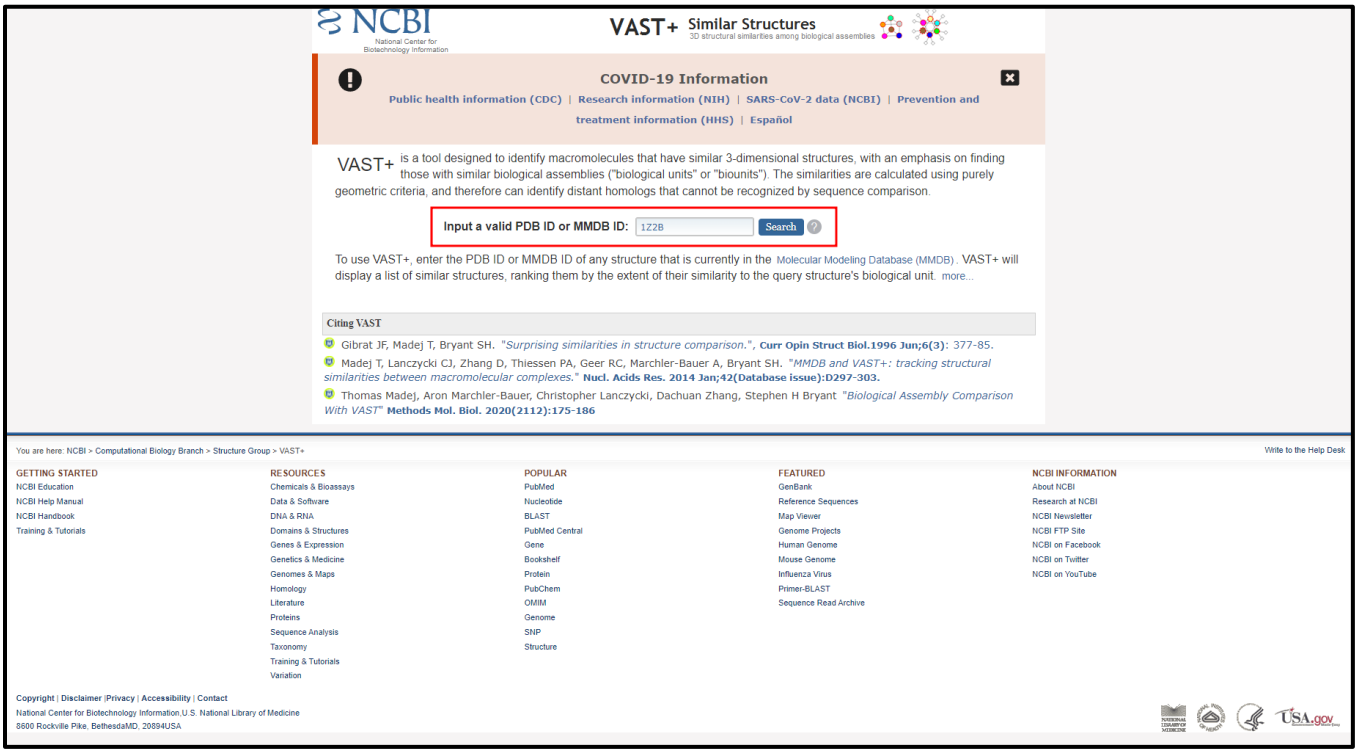


Fig3. Search for tubulin PDB structure

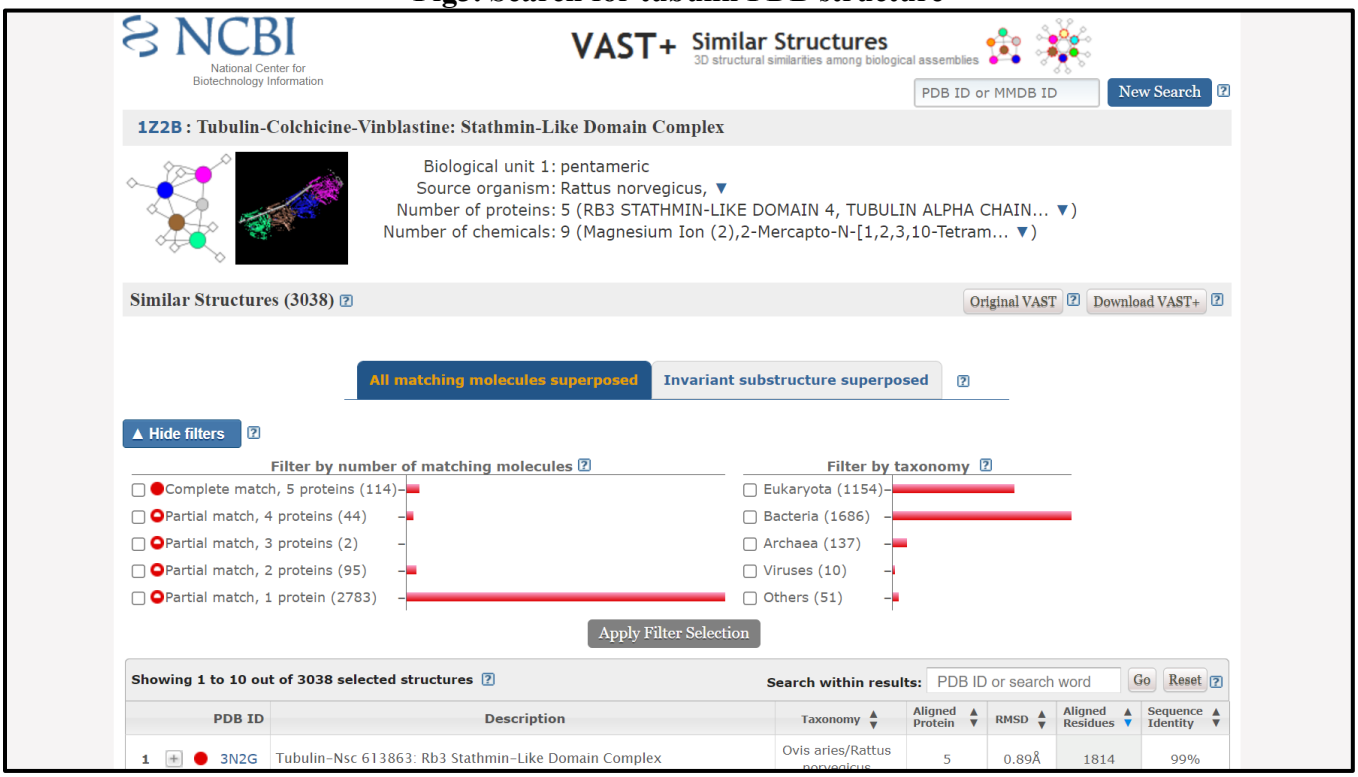
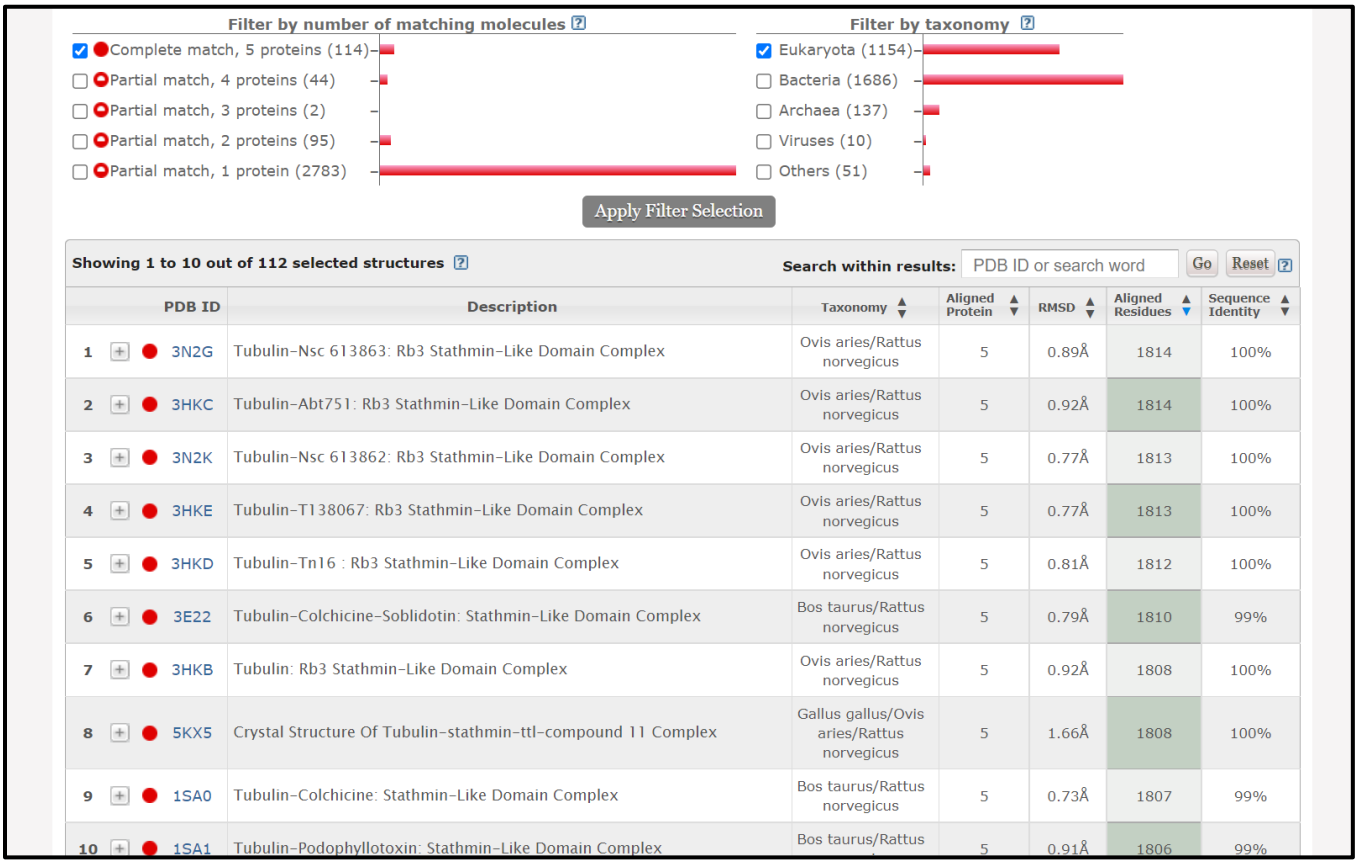
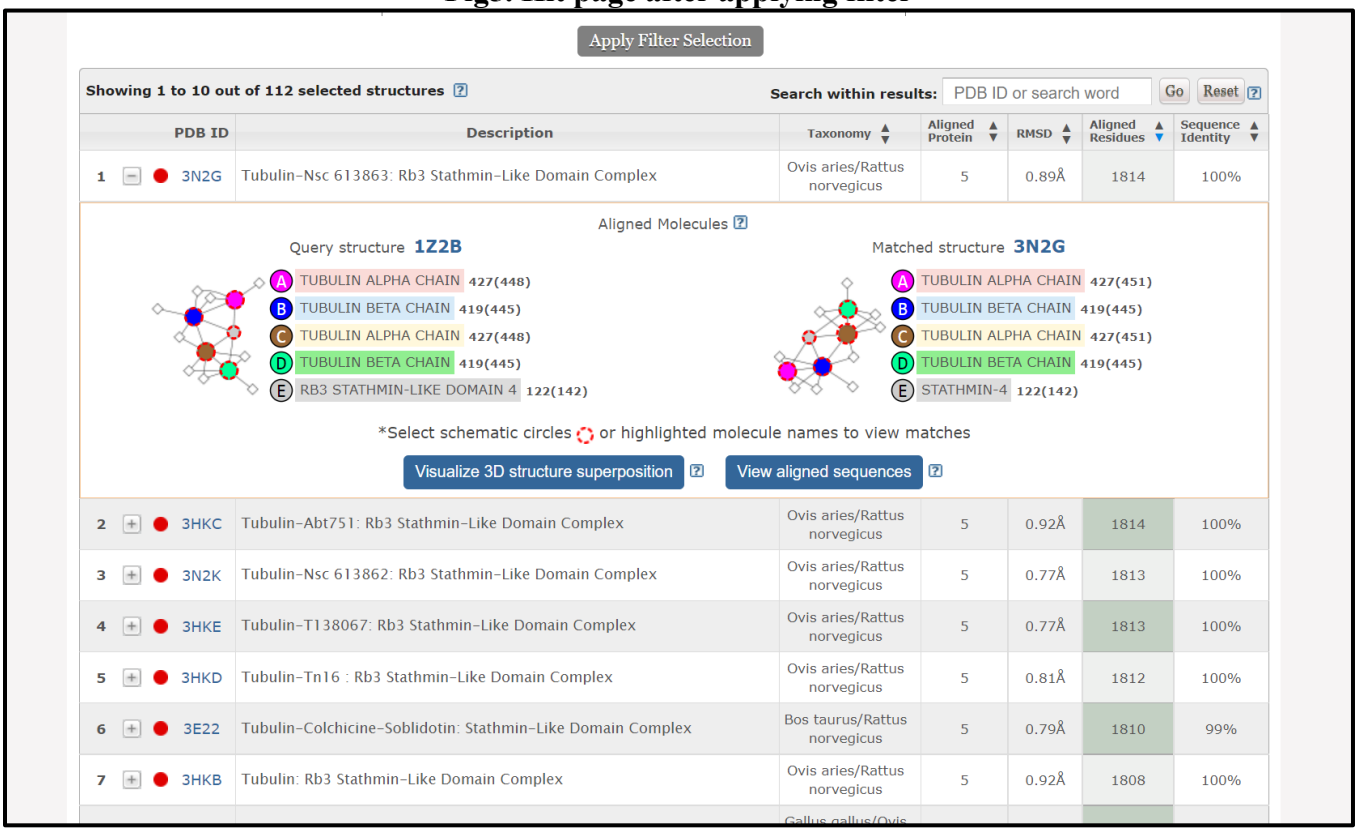


Fig4. Hit page for Tubulin



	PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1	3N2G	Tubulin-Nsc 613863: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.89Å	1814	100%
2	3HKC	Tubulin-Abt751: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.92Å	1814	100%
3	3N2K	Tubulin-Nsc 613862: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.77Å	1813	100%
4	3HKE	Tubulin-T138067: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.77Å	1813	100%
5	3HKD	Tubulin-Tn16 : Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.81Å	1812	100%
6	3E22	Tubulin-Colchicine-Soblidotin: Stathmin-Like Domain Complex	Bos taurus/Rattus norvegicus	5	0.79Å	1810	99%
7	3HKB	Tubulin: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.92Å	1808	100%
8	5KX5	Crystal Structure Of Tubulin-stathmin-ttl-compound 11 Complex	Gallus gallus/Ovis aries/Rattus norvegicus	5	1.66Å	1808	100%
9	1SA0	Tubulin-Colchicine: Stathmin-Like Domain Complex	Bos taurus/Rattus norvegicus	5	0.73Å	1807	99%
10	1SA1	Tubulin-Podophyllotoxin: Stathmin-Like Domain Complex	Bos taurus/Rattus	5	0.91Å	1806	99%

Fig5. Hit page after applying filter



	PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1	3N2G	Tubulin-Nsc 613863: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.89Å	1814	100%
2	3HKC	Tubulin-Abt751: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.92Å	1814	100%
3	3N2K	Tubulin-Nsc 613862: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.77Å	1813	100%
4	3HKE	Tubulin-T138067: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.77Å	1813	100%
5	3HKD	Tubulin-Tn16 : Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.81Å	1812	100%
6	3E22	Tubulin-Colchicine-Soblidotin: Stathmin-Like Domain Complex	Bos taurus/Rattus norvegicus	5	0.79Å	1810	99%
7	3HKB	Tubulin: Rb3 Stathmin-Like Domain Complex	Ovis aries/Rattus norvegicus	5	0.92Å	1808	100%

Fig6. Result for aligned molecules

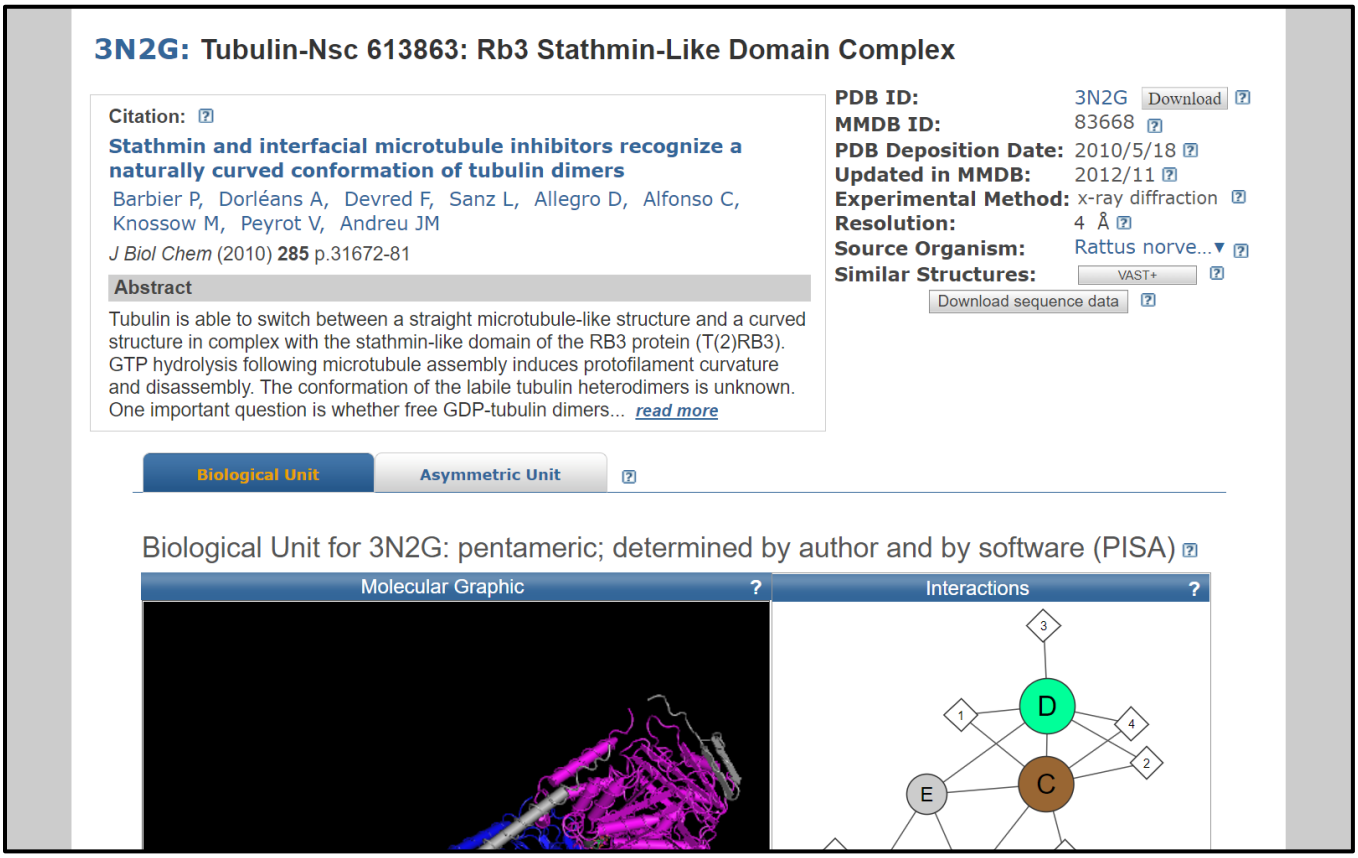


Fig7. Result page for structure (3N2G) similar to Tubulin

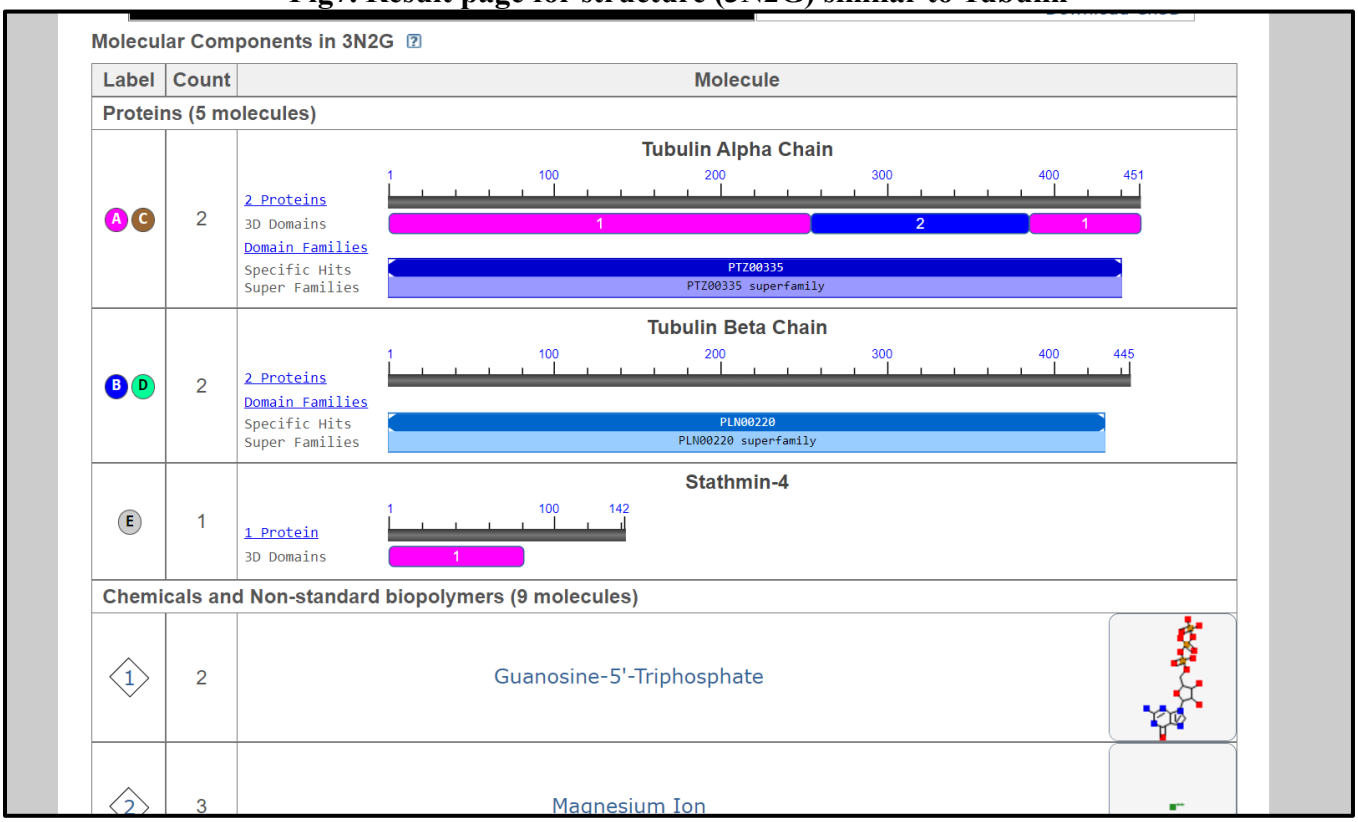


Fig8. Molecular components and chemical and non-standard biopolymers for 1N5U

Fig9. Result page for aligned sequences

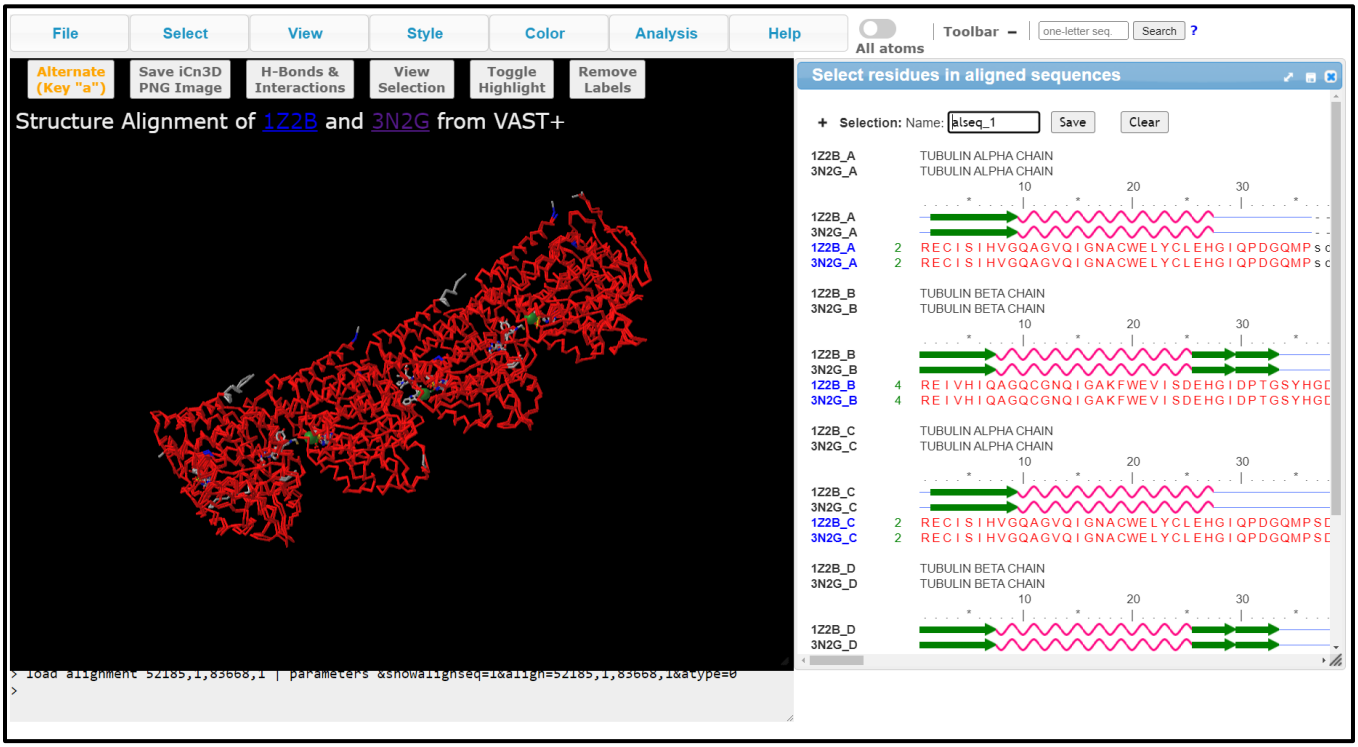


Fig10. Result page for 3D structure superimposition

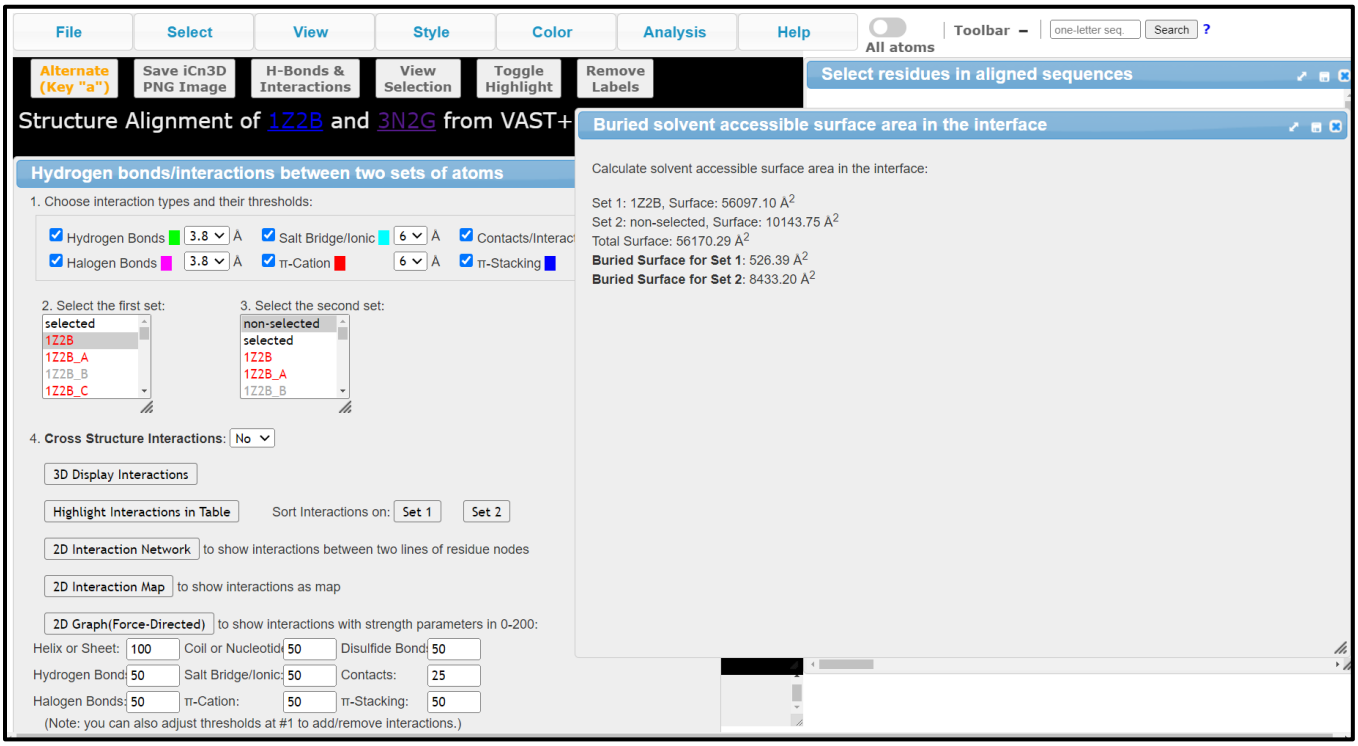


Fig11. Buried surface information

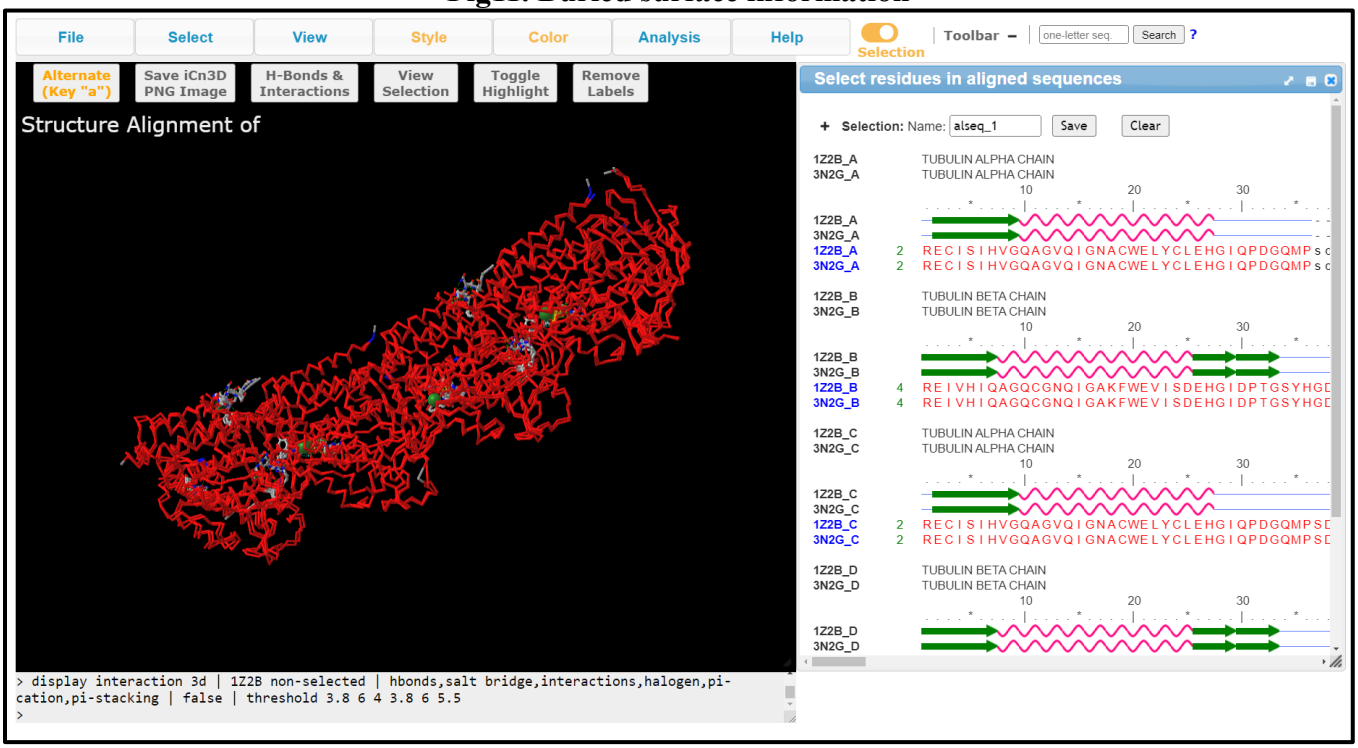


Fig12. Result for itneractions

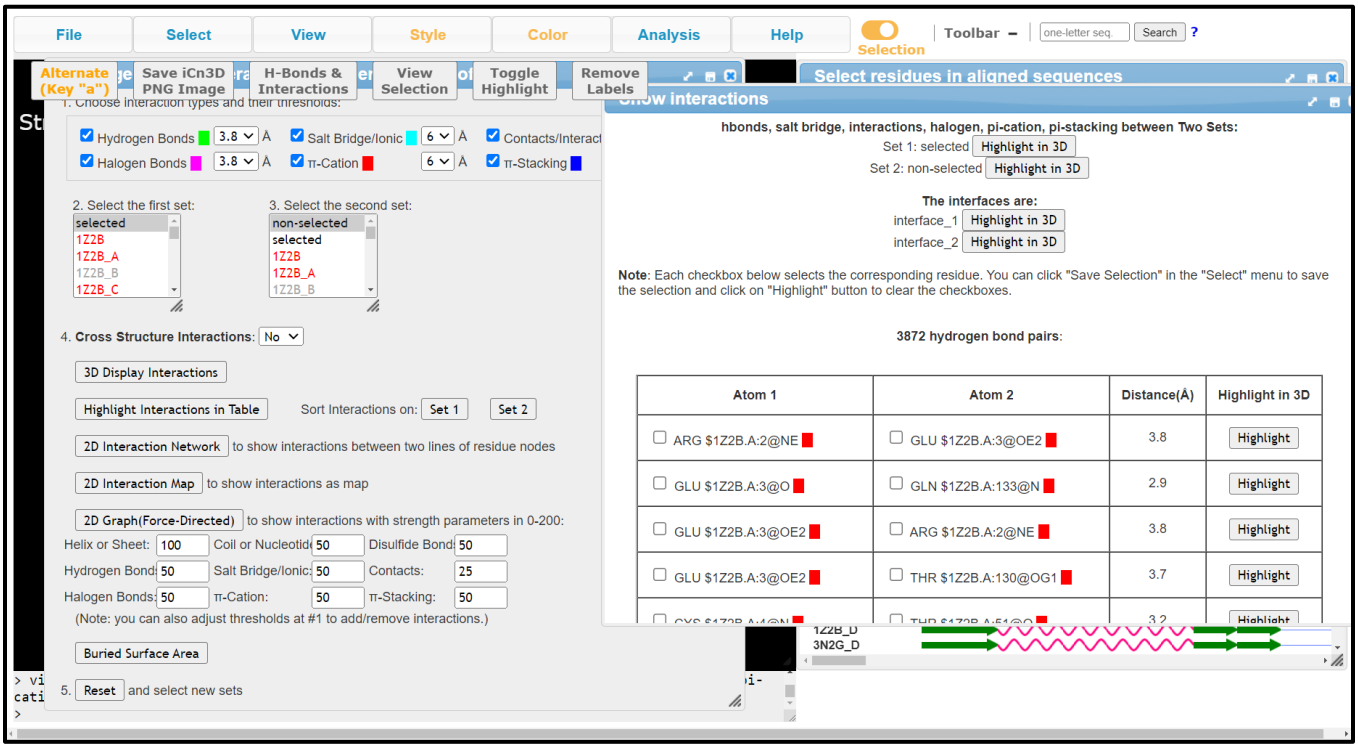


Fig13. Result for interactions

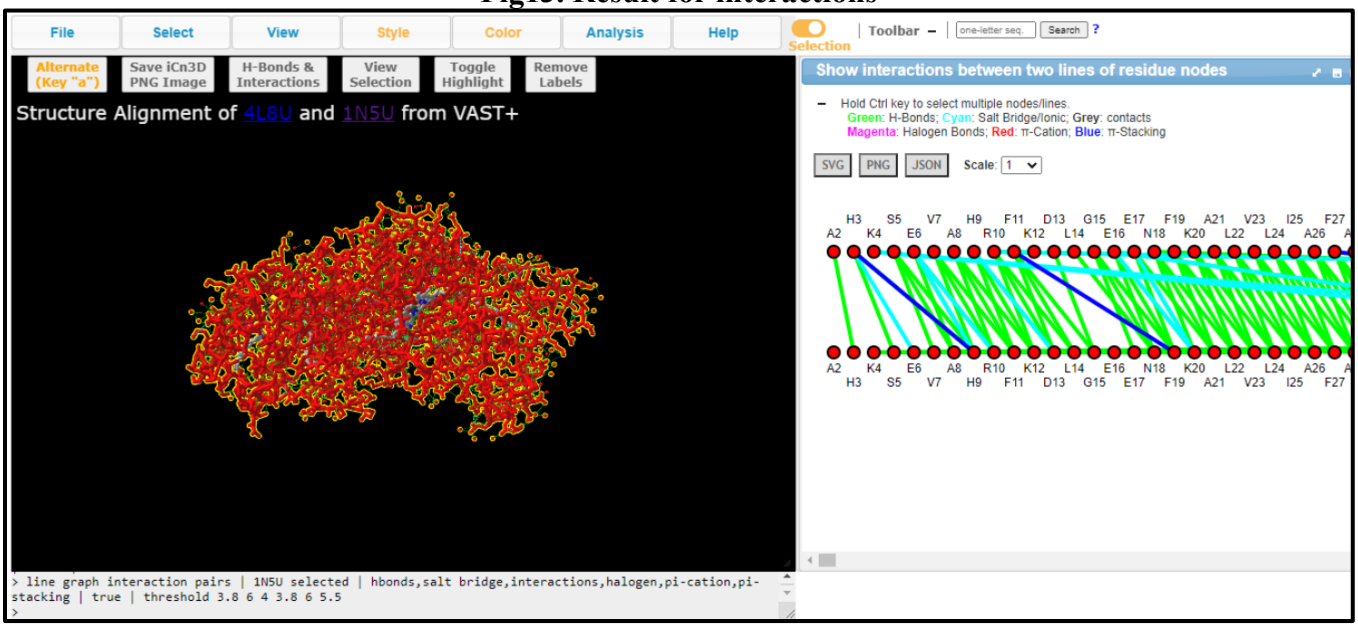


Fig14. Result for 2D interaction network

RESULT:

PDB ID for tubulin structure was searched in structure similarity BLAST tool, VAST and 152 similar structures were retrieved.

CONCLUSION:

VAST is a useful structure similarity BLAST tool which provides user with similar structures to their query along with its molecular components and chemicals and non-standard biopolymers, aligned sequences and 3D structure superimposition information which includes information regarding H-bonds, interactions, buried surface area, 2D interaction network and much more. Describing the structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered

proteins.

REFERENCES:

1. Madej, T.; Lanczycki, C. J.; Zhang, D.; Thiessen, P. A.; Geer, R. C.; Marchler-Bauer, A.; Bryant, S. H. (2014). MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Research, 42(D1), D297–D303. doi:10.1093/nar/gkt1208
2. Tubulin: Mystery of vital cell protein solved after 30 years. (n.d.). Retrieved March 19, 2022, from https://www2.lbl.gov/Science-Articles/Archive/3D-tubulin.html#:~:text=Tubulin%20is%20the%20protein%20that,or%20to%20regulate%20intracellular%20transport.
3. Similar Protein Structure Assemblies. (2014). Nih.gov. Retrieved March 19, 2022, from https://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi
4. Bank, R. P. D. (n.d.-b). RCSB PDB - 4L8U: X-ray study of human serum albumin complexed with 9 amino camptothecin. Www.rcsb.org. Retrieved March 19, 2022, from https://www.rcsb.org/structure/4L8U
5. Bank, R. C. S. B. P. D. (n.d.). *1Z2B: Tubulin-colchicine-vinblastine: Stathmin-like domain complex*. RCSB PDB. Retrieved March 19, 2022, from https://www.rcsb.org/structure/1Z2B
6. U.S. National Library of Medicine. (n.d.). *3N2G: Tubulin-NSC 613863: RB3 stathmin-like domain complex*. National Center for Biotechnology Information. Retrieved March 19, 2022, from https://www.ncbi.nlm.nih.gov/Structure/pdb/3N2G
7. iCn3D: Web-based 3D Structure Viewer. (n.d.). Www.ncbi.nlm.nih.gov. Retrieved March 19, 2022, from https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html?showalignseq=1&align=112122

DATE: 19-03-22

WEBLEM 7b
DALI
(URL: http://ekhidna2.biocenter.helsinki.fi/dali/)

AIM:

To erform structural blast for tubulin using DALI tool

Introduction:

Tubulin is the protein that polymerizes into long chains or filaments that form microtubules, hollow fibers which serve as a skeletal system for living cells. Microtubules have the ability to shift through various formations which is what enables a cell to undergo mitosis or to regulate intracellular transport.

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

User can perform three types of database searches:

- **Heuristic PDB search** - compares one query structure against those in the Protein Data Bank
- **Exhaustive PDB25 search** - compares one query structure against a representative subset of the Protein Data Bank
- **Hierarchical AF-DB search -** compares one query structure against a species subset of the AlphaFold Database

METHODOLOGY:

1. Open homepage for DALI. (URL: http://ekhidna2.biocenter.helsinki.fi/dali/)
2. Enter Albumin PDB ID.
3. Observe similar structures matches against PDB25, PDB50, PDB90 and all PDB structures.
4. Interpret the results.

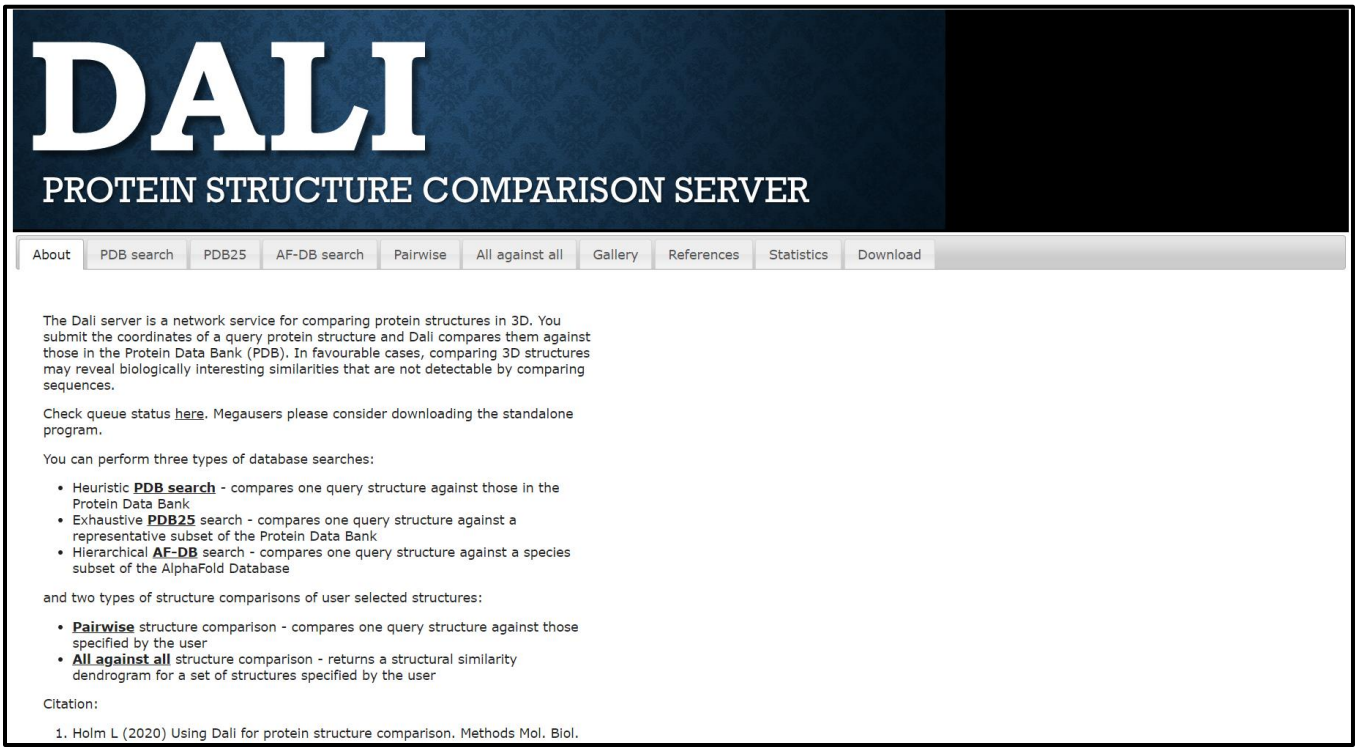


Fig1. Homepage for DALI

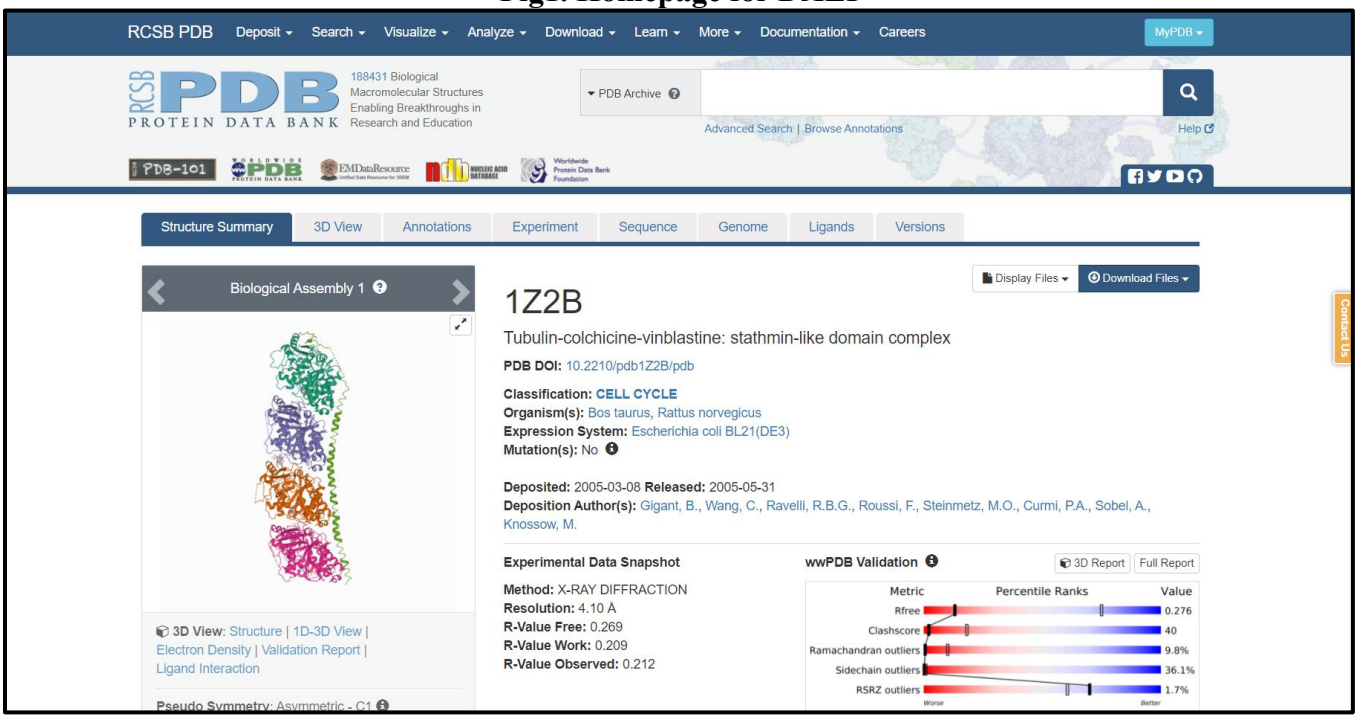


Fig2. Tubulin PDB structure

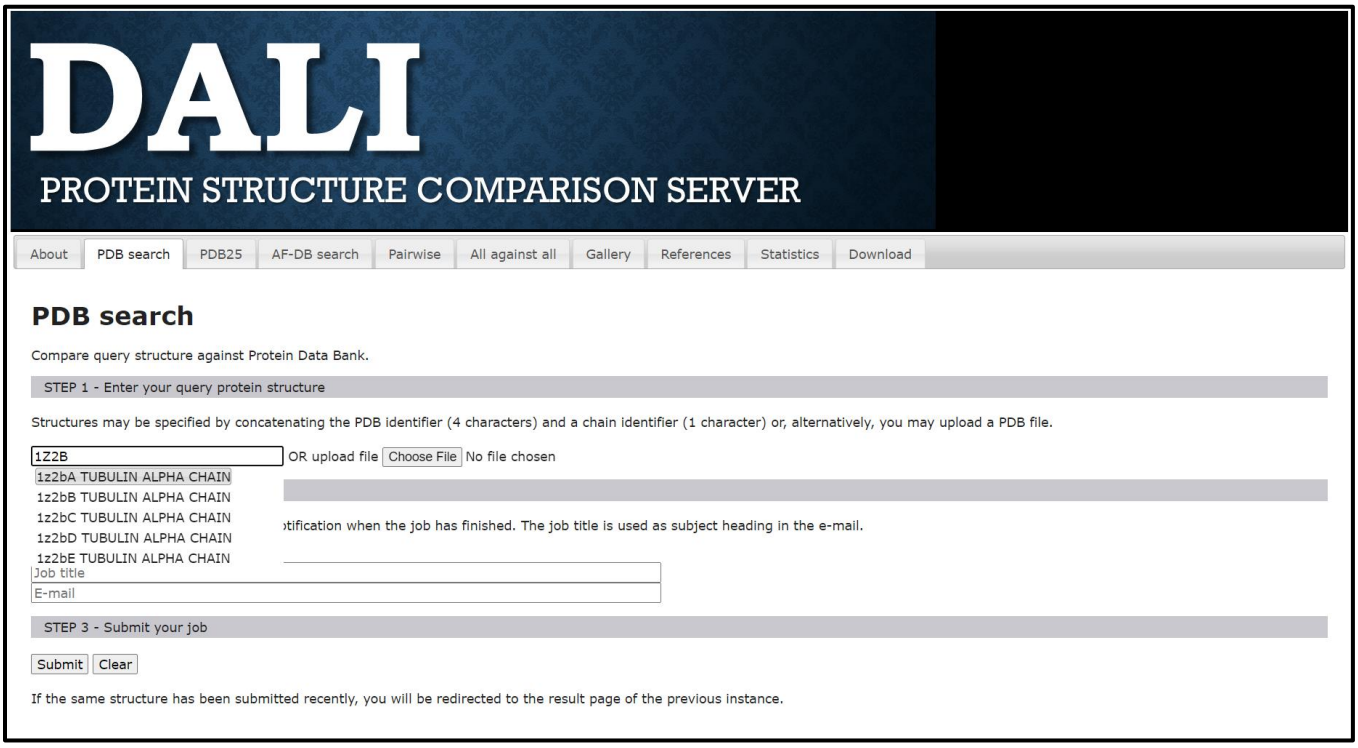


Fig3. PDB search for tubulin

Results: lmao

Chain: 1z2bA

- Matches against PDB25. Correlation matrix
- Matches against PDB50
- Matches against PDB90
- Matches against full PDB
- Download matches against PDB25
- Download matches against PDB50
- Download matches against PDB90
- Download matches against full PDB

Results will be deleted after one week.

Fig4. Result page of Tubulin

Matches against PDB25:

Results: lmao

Query: 1z2bA

MOLECULE: TUBULIN ALPHA CHAIN;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | Expand gaps | 3D Superimposition (PV) | SANS | PANZ | Pfam | Reset Selection

Summary

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	7m2w-A	47.6	1.8	408	453	30	PDB	MOLECULE: TUBULIN GAMMA CHAIN;
2:	3zid-A	32.4	2.7	317	360	16	PDB	MOLECULE: TUBULIN/FTSZ, GTPASE;
3:	1w59-A	24.7	3.1	291	350	12	PDB	MOLECULE: CELL DIVISION PROTEIN FTSZ HOMOLOG 1;
4:	4xcq-A	23.9	3.5	280	303	11	PDB	MOLECULE: TUBZ;
5:	4ei7-A	23.6	4.0	313	382	11	PDB	MOLECULE: PLASMID REPLICATION PROTEIN REPX;
6:	2xka-F	23.2	3.8	313	414	12	PDB	MOLECULE: FTSZ/TUBULIN-RELATED PROTEIN;
7:	3zbq-A	19.8	3.5	263	315	13	PDB	MOLECULE: PHIKZ039;
8:	4fkz-A	7.8	3.9	174	384	10	PDB	MOLECULE: UDP-N-ACETYLGLUCOSAMINE 2-EPIMERASE;
9:	4zht-A	7.3	3.7	167	384	9	PDB	MOLECULE: BIFUNCTIONAL UDP-N-ACETYLGLUCOSAMINE 2-EPIMERASE/

Fig5. Result for similar structures

Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence. Structurally equivalent residues are in uppercase, structurally non-equivalent residues (e.g. in loops) are in lowercase. Amino acid identities are marked by vertical bars.

No 1: Query=1z2bA Sbjct=7m2wA Z-score=47.6

back to top

```
DSSP  --LEEEEEEEHHHHHHHHHHHHHHHHHHHLLLLLLLLL---------lLLLLLLLLLLLL
Query --RECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMP---------dSFNTFFSETGAG   49
ident    | |    || |   |   |     || |  ||                || |
Sbjct mgGEIITLQAGQCGNHVGKFLWSQLAKEHAIGTDGLSQlpdssterddDTKPFFRENCRN   60
DSSP  llLLEEEEEELHHHHHHHHHHHHHHHHHLLLLLLLLLLllllllllllLLLLLEEELLLL


DSSP  LLLLLEEEEELLLLHHHHHLLLlLLLLLLHHHEEEL-LLLL-LLLHHHHHLlLHHHHHHH
Query KHVPRAVFVDLEPTVIDEVRTGtYRQLFHPEQLITG-KEDA-ANNYARGHYtIGKEIIDL  107
ident |  |||   | || ||  |     |  | |             |  | |   ||    |
Sbjct KFTPRAIMMDSEPSVIADVENT-FRGFFDPRNTWVAsDGASaGNSWANGYD-IGTRNQDD  118
DSSP  LEEELEEEELLLHHHHHHHHHH-LLLLLLHHHEELLlLLLLlLLLHHHHHH-HHHHLHHH


DSSP  HHHHHHHHLLLLLLEEEEEEEELLLHHHHHHHHHHHHHHHHHLLLLEEEEEEELLLLll
Query VLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPqv  167
ident  |  | |  |      ||   ||  ||||||  | | | |   | ||     |  ||
Sbjct ILNKIDKEIDSTDNFEGFQLLHSVAGGTGSGLGSNLLEALCDRYPKKILTTYSVFPAR--  176
DSSP  HHHHHHHHHHLLLLLLEEEEELLLLHHHHHHHHHHHHHHHHHHLLLLEEEEEEELLL--
```

Fig6. Results for pairwise structural alignment

```
REMARK  Coordinates of 7m2w rotated and translated as follows:
REMARK  |  0.40520  0.53216 -0.74338 |   | x |   |  173.000 |
REMARK  | -0.07305 -0.79168 -0.60655 | * | y | + |  510.000 |
REMARK  | -0.91131  0.30008 -0.28191 |   | z |   |  313.000 |
REMARK
REMARK  HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER    CELL CYCLE                              17-MAR-21   7M2W
TITLE     ENGINEERED DISULFIDE CROSS-LINKED CLOSED CONFORMATION OF THE YEAST
TITLE    2 GAMMA-TURC(SS)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: TUBULIN GAMMA CHAIN;
COMPND   3 CHAIN: B, A, C, D;
COMPND   4 SYNONYM: GAMMA-TUBULIN;
COMPND   5 ENGINEERED: YES;
COMPND   6 MUTATION: YES;
COMPND   7 MOL_ID: 2;
COMPND   8 MOLECULE: SPINDLE POLE BODY COMPONENT SPC97;
COMPND   9 CHAIN: E, G;
COMPND  10 ENGINEERED: YES;
COMPND  11 MOL_ID: 3;
COMPND  12 MOLECULE: SPINDLE POLE BODY COMPONENT SPC98;
COMPND  13 CHAIN: F, H;
COMPND  14 ENGINEERED: YES;
COMPND  15 MOL_ID: 4;
COMPND  16 MOLECULE: SPINDLE POLE BODY COMPONENT 110;
COMPND  17 CHAIN: U, K, X, Y;
COMPND  18 SYNONYM: EXTRAGENIC SUPPRESSOR OF CMD1-1 MUTANT PROTEIN 1,NUCLEAR
COMPND  19 FILAMENT-RELATED PROTEIN 1,SPINDLE POLE BODY SPACER PROTEIN SPC110;
COMPND  20 ENGINEERED: YES
AUTHOR    A.F.BRILOT,A.S.LYON,A.ZELTER,S.VISWANATH,A.MAXWELL,M.J.MACCOSS,
AUTHOR   2 E.G.MULLER,A.SALI,T.N.DAVIS,D.A.AGARD
```

Fig7. Result for coordinates of similar structure

Matches against PDB50:

Results: lmao

Query: 1z2bA

MOLECULE: TUBULIN ALPHA CHAIN;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | ☑ Expand gaps | 3D Superimposition (PV) | SANS | PANZ | Pfam | Reset Selection

Summary

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
☐ 1:	2btq-A	54.0	1.5	410	436	38	PDB	MOLECULE: TUBULIN BTUBA;
☐ 2:	7pqc-B	54.0	1.6	426	451	99	PDB	MOLECULE: TUBULIN BETA CHAIN;
☐ 3:	2btq-B	53.5	1.8	388	391	41	PDB	MOLECULE: TUBULIN BTUBA;
☐ 4:	7pqc-A	51.5	1.8	418	445	41	PDB	MOLECULE: TUBULIN BETA CHAIN;
☐ 5:	3cb2-A	51.4	1.8	407	432	32	PDB	MOLECULE: TUBULIN GAMMA-1 CHAIN;
☐ 6:	7m2w-A	47.6	1.8	408	453	30	PDB	MOLECULE: TUBULIN GAMMA CHAIN;
☐ 7:	7anz-B	44.6	1.8	375	410	28	PDB	MOLECULE: TUBULIN GAMMA CHAIN;
☐ 8:	3zid-A	32.4	2.7	317	360	16	PDB	MOLECULE: TUBULIN/FTSZ, GTPASE;
☐ 9:	4b46-A	30.6	2.7	307	330	18	PDB	MOLECULE: CELL DIVISION PROTEIN FTSZ;

Fig8. Result for similar structures

Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence. Structurally equivalent residues are in uppercase, structurally non-equivalent residues (e.g. in loops) are in lowercase. Amino acid identities are marked by vertical bars.

No 1: Query=1z2bA Sbjct=2btqA Z-score=54.0

back to top

```
DSSP  --LEEEEEEEHHHHHHHHHHHHHHHHHHLLL-LLLLLL------lLLLLLLLLLLL---
Query --RECISIHVGQAGVQIGNACWELYCLEHGIQ-PDGQMP------dSFNTFFSETGA---   48
ident      |    |||| ||    |   ||||||    ||            |||  |
Sbjct kvNNTIVVSIGQAGNQIAASFWKTVCLEHGIDpLTGQTApgvaprgNWSSFFSKLGEsss   60
DSSP  llLEEEEEEELHHHHHHHHHHHHHHHHHHLLLlLLLLLLllllllLLLLLEEEELLlll


DSSP  LLLLLLEEEEELLLLHHHHHLLLlLLLLLLHHHEEELLLLLLLLHHHHHLLLHHHHHHHH
Query GKHVPRAVFVDLEPTVIDEVRTGtYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLV  108
ident |  ||||  ||||| ||| |       || |  ||   | |  | | |    | |    |
Sbjct GSYVPRAIMVDLEPSVIDNVKAT-SGSLFNPANLISRTEGAGGNFAVGYLGAGREVLPEV  119
DSSP  LEEEELEEEEELLHHHHHHHHHH-LLLLLLHHHEEELLLLLLLLHHHHHLLHHHHHHHH


DSSP  HHHHHHHLLLLLLEEEEEEEELLLHHHHHHHHHHHHHHHHHHLLLLEEEEEEELLLLLLL
Query LDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVS  168
ident   |     | |    |  | |  ||||||||  || | |   ||    |     | ||||
Sbjct MSRLDYEIDKCDNVGGIIVLHAIGGGTGSGFGALLIESLKEKYGEIPVLSCAVLPSPQVS  179
DSSP  HHHHHHHHHLLLLEEEEEEEELLLLLHHHHHHHHHHHHHHHHHLLLLEEEEEEELLHHHL
```

Fig9. Result for similar structures

```
REMARK  Coordinates of 2btq rotated and translated as follows:
REMARK  |  0.28475 -0.95786  0.03764 |    | x |    |   92.000 |
REMARK  | -0.37807 -0.07614  0.92264 | * | y | + |  137.000 |
REMARK  | -0.88090 -0.27695 -0.38382 |    | z |    |   57.000 |
REMARK
REMARK  HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER    STRUCTURAL PROTEIN                      06-JUN-05   2BTQ
TITLE     STRUCTURE OF BTUBAB HETERODIMER FROM PROSTHECOBACTER DEJONGEII
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: TUBULIN BTUBA;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES;
COMPND   5 MOL_ID: 2;
COMPND   6 MOLECULE: TUBULIN BTUBB;
COMPND   7 CHAIN: B;
COMPND   8 ENGINEERED: YES
AUTHOR    D.SCHLIEPER,J.LOWE
HELIX    1   1 GLY A   12  GLY A   31  1                                  20
HELIX    2   2 GLU A   74  SER A   85  1                                  12
HELIX    3   3 ASN A   90  ALA A   92  5                                   3
HELIX    4   4 ASN A  104  LEU A  110  1                                   7
HELIX    5   5 GLY A  111  LYS A  130  1                                  20
HELIX    6   6 GLY A  146  TYR A  163  1                                  18
HELIX    7   7 SER A  176  SER A  180  5                                   5
HELIX    8   8 THR A  184  ALA A  200  1                                  17
HELIX    9   9 ASN A  208  ARG A  217  1                                  10
HELIX   10  10 THR A  225  PHE A  246  1                                  22
HELIX   11  11 SER A  255  VAL A  264  1                                  10
HELIX   12  12 GLY A  291  PHE A  300  1                                  10
HELIX   13  13 SER A  310  GLY A  314  5                                   5
HELIX   14  14 ASP A  329  LEU A  344  1                                  16
```

Fig10. Result for coordinates of similar structure

Query: 1z2bA

MOLECULE: TUBULIN ALPHA CHAIN;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | Expand gaps | 3D Superimposition (PV) | SANS | PANZ | Pfam | Reset Selection

Summary

	No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
☐	1:	5w3f-A	55.4	1.5	424	440	74	PDB	MOLECULE: TUBULIN ALPHA-1 CHAIN;
☐	2:	2btq-A	54.0	1.5	410	436	38	PDB	MOLECULE: TUBULIN BTUBA;
☐	3:	7pqc-B	54.0	1.6	426	451	99	PDB	MOLECULE: TUBULIN BETA CHAIN;
☐	4:	2btq-B	53.5	1.8	388	391	41	PDB	MOLECULE: TUBULIN BTUBA;
☐	5:	5ubq-A	52.2	1.7	425	441	87	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐	6:	5mlv-C	52.0	1.7	418	430	40	PDB	MOLECULE: TUBULIN ALPHA-1 CHAIN;
☐	7:	7pqc-A	51.5	1.8	418	445	41	PDB	MOLECULE: TUBULIN BETA CHAIN;
☐	8:	3cb2-A	51.4	1.8	407	432	32	PDB	MOLECULE: TUBULIN GAMMA-1 CHAIN;
☐	9:	5w3f-B	51.3	1.7	416	427	40	PDB	MOLECULE: TUBULIN ALPHA-1 CHAIN;
☐	10:	7m2w-A	47.6	1.8	408	453	30	PDB	MOLECULE: TUBULIN GAMMA CHAIN;
☐	11:	6v5v-g	45.8	1.6	356	369	33	PDB	MOLECULE: TUBULIN GAMMA-1 CHAIN;
☐	12:	7anz-B	44.6	1.8	375	410	28	PDB	MOLECULE: TUBULIN GAMMA CHAIN;

Fig11. Result for similar structures

Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence. Structurally equivalent residues are in uppercase, structurally non-equivalent residues (e.g. in loops) are in lowercase. Amino acid identities are marked by vertical bars.

No 1: Query=1z2bA Sbjct=5w3fA Z-score=55.4

back to top

```
DSSP  -LEEEEEEEHHHHHHHHHHHHHHHHHHLLLLLLLLL---------LLLL-LLLLLLLLL
Query -RECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMP---------DSFN-TFFSETGAG   49
ident  || ||| ||||| |||||||||| ||||| |||                 ||| ||| |
Sbjct mREVISINVGQAGCQIGNACWELYSLEHGIKPDGHLEdglskpkggEEGFsTFFHETGYG   60
DSSP  lLLEEEEEEHHHHHHHHHHHHHHHHHHLLLLLLLLLLllllllllLLLLlLLEELLLLL


DSSP  LLLLLEEEEELLLHHHHHLLLLLLLLLLHHHEEELLLLLLLHHHHHLLLHHHHHHHHH
Query KHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVL  109
ident | ||||  ||||| |||||| | |  |||||||| ||||||||||||||| | ||   ||
Sbjct KFVPRAIYVDLEPNVIDEVRNGPYKDLFHPEQLISGKEDAANNYARGHYTVGREILGDVL  120
DSSP  LLEELEEEELLLHHHHHHHHLLLLHHHLLHHHEELLLLLLLLHHHHLLLLHHHLHHHHH


DSSP  HHHHHHLLLLLLEEEEEEEELLLHHHHHHHHHHHHHHHHHLLLLEEEEEEELLL-LLLL
Query DRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPA-PQVS  168
ident |||||||||| ||||||  || |||||||  ||| | ||  |||||||||  |||
Sbjct DRIRKLADQCDGLQGFLFTHSLGGGTGSGLGSLLLEELSAEYGKKSKLEFAVYPApQVST  180
DSSP  HHHHHHHHHLLLLLEEEEEEELLLHHHHHHHHHHHHHHHHHLLLLEEEEEEELLlLLLL
```

Fig12. Result for pairwise structural alignment

```
REMARK  Coordinates of 5w3f rotated and translated as follows:
REMARK  |  0.39885 -0.50887  0.76286 |   | x |   |  -11.000 |
REMARK  | -0.13412  0.79058  0.59749 | * | y | + | -390.000 |
REMARK  | -0.90715 -0.34062  0.24708 |   | z |   |  366.000 |
REMARK
REMARK  HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER    HYDROLASE                               07-JUN-17   5W3F
TITLE     YEAST TUBULIN POLYMERIZED WITH GTP IN VITRO
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: TUBULIN ALPHA-1 CHAIN;
COMPND   3 CHAIN: A;
COMPND   4 MOL_ID: 2;
COMPND   5 MOLECULE: TUBULIN BETA CHAIN;
COMPND   6 CHAIN: B;
COMPND   7 SYNONYM: BETA-TUBULIN
AUTHOR    S.C.HOWES,E.A.GEYER,B.LAFRANCE,R.ZHANG,E.H.KELLOGG,S.WESTERMANN,
AUTHOR   2 L.M.RICE,E.NOGALES
HELIX    1 AA1 GLY A   10  HIS A   28  1                                  19
HELIX    2 AA2 GLU A   72  ASN A   81  1                                  10
HELIX    3 AA3 TYR A   84  PHE A   88  5                                   5
HELIX    4 AA4 HIS A   89  GLU A   91  5                                   3
HELIX    5 AA5 ASN A  103  HIS A  108  1                                   6
HELIX    6 AA6 VAL A  111  GLU A  114  5                                   4
HELIX    7 AA7 ILE A  115  CYS A  130  1                                  16
HELIX    8 AA8 GLY A  144  TYR A  162  1                                  19
HELIX    9 AA9 VAL A  183  ALA A  199  1                                  17
HELIX   10 AB1 ASN A  207  ASN A  217  1                                  11
HELIX   11 AB2 SER A  224  VAL A  239  1                                  16
HELIX   12 AB3 THR A  240  ARG A  244  5                                   5
HELIX   13 AB4 ASN A  254  VAL A  261  1                                   8
HELIX   14 AB5 SER A  288  GLU A  298  1                                  11
```

Fig13. Result for coordinates of similar structure

Query: 1z2bA

MOLECULE: TUBULIN ALPHA CHAIN;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | ☑ Expand gaps | 3D Superimposition (PV) | SANS | PANZ | Pfam | Reset Selection

Summary

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
☐ 1:	1z2b-A	75.0	0.0	427	427	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 2:	1sa0-A	69.5	0.3	427	427	99	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 3:	3hkc-A	69.1	0.5	426	426	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 4:	3hke-A	69.0	0.5	427	427	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 5:	3n2k-A	68.9	0.5	427	428	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 6:	3e22-A	68.9	0.5	427	427	98	PDB	MOLECULE: TUBULIN ALPHA-1C CHAIN;
☐ 7:	3hkb-A	68.6	0.5	427	427	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 8:	1z2b-C	68.5	0.6	420	427	99	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 9:	1sa0-C	68.5	0.5	421	421	99	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 10:	3edl-F	68.4	0.5	421	421	99	PDB	MOLECULE: ALPHA-TUBULIN;
☐ 11:	3n2g-A	68.1	0.5	427	428	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;
☐ 12:	3hkd-A	68.1	0.5	427	428	100	PDB	MOLECULE: TUBULIN ALPHA CHAIN;

Fig14. Result for similar structures

Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence. Structurally equivalent residues are in uppercase, structurally non-equivalent residues (e.g. in loops) are in lowercase. Amino acid identities are marked by vertical bars.

No 1: Query=1z2bA Sbjct=1z2bA Z-score=75.0

back to top

```
DSSP  LEEEEEEEHHHHHHHHHHHHHHHHHHHLLLLLLLLLLLLLLLLLLLLLLLLLLLEEEEEL
Query RECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPDSFNTFFSETGAGKHVPRAVFVDL   60
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct RECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPDSFNTFFSETGAGKHVPRAVFVDL   60
DSSP  LEEEEEEEHHHHHHHHHHHHHHHHHHHLLLLLLLLLLLLLLLLLLLLLLLLLLLEEEEEL


DSSP  LLLHHHHHLLLLLLLLLLHHHEEELLLLLLLLHHHHHLLLHHHHHHHHHHHHHHHLLLLL
Query EPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCT  120
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct EPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCT  120
DSSP  LLLHHHHHLLLLLLLLLLHHHEEELLLLLLLLHHHHHLLLHHHHHHHHHHHHHHHLLLLL


DSSP  LEEEEEEEEELLLHHHHHHHHHHHHHHHHHLLLLEEEEEEELLLLLLLLLLLHHHHHHHH
Query GLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILT  180
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct GLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILT  180
DSSP  LEEEEEEEEELLLHHHHHHHHHHHHHHHHHLLLLEEEEEEELLLLLLLLLLLHHHHHHHH
```

Fig15. Result for pairwise structural alignment

```
REMARK  Coordinates of 1z2b rotated and translated as follows:
REMARK  |  1.00000  0.00000 -0.00000 |   | x |   |   -0.000 |
REMARK  | -0.00000  1.00000  0.00000 | * | y | + |    0.000 |
REMARK  |  0.00000 -0.00000  1.00000 |   | z |   |    0.000 |
REMARK
REMARK  HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER    CELL CYCLE                              08-MAR-05   1Z2B
TITLE     TUBULIN-COLCHICINE-VINBLASTINE: STATHMIN-LIKE DOMAIN COMPLEX
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: TUBULIN ALPHA CHAIN;
COMPND   3 CHAIN: A, C;
COMPND   4 MOL_ID: 2;
COMPND   5 MOLECULE: TUBULIN BETA CHAIN;
COMPND   6 CHAIN: B, D;
COMPND   7 MOL_ID: 3;
COMPND   8 MOLECULE: RB3 STATHMIN-LIKE DOMAIN 4;
COMPND   9 CHAIN: E;
COMPND  10 SYNONYM: STATHMIN-LIKE PROTEIN B3, RB3-SLD;
COMPND  11 ENGINEERED: YES
AUTHOR    B.GIGANT,C.WANG,R.B.G.RAVELLI,F.ROUSSI,M.O.STEINMETZ,P.A.CURMI,
AUTHOR   2 A.SOBEL,M.KNOSSOW
HELIX    1   1 GLY A   10  GLY A   29  1                                  20
HELIX    2   2 THR A   73  ARG A   79  1                                   7
HELIX    3   3 HIS A   88  GLU A   90  5                                   3
HELIX    4   4 ASN A  102  TYR A  108  1                                   7
HELIX    5   5 ILE A  110  ALA A  126  1                                  17
HELIX    6   6 GLY A  143  TYR A  161  1                                  19
HELIX    7   7 VAL A  182  GLU A  196  1                                  15
HELIX    8   8 ASP A  205  ASN A  216  1                                  12
HELIX    9   9 THR A  223  ALA A  240  1                                  18
HELIX   10  10 ALA A  240  ASP A  245  1                                   6
```

Fig16. Result for coordinates of similar structure

RESULT:

PDB ID for albumin structure was searched in structure similarity BLAST tool, DALI and similar structures matches against PDB25, PDB50, PDB90 and all PDB structures were retrieved.

CONCLUSION:

DALI is a useful structure similarity BLAST tool which provides user with similar structures to their query along with its pairwise alignment, coordinates information, 3D superimposition results. Describing the

structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered proteins.

REFERENCES:

1. Tubulin: Mystery of vital cell protein solved after 30 years. (n.d.). Retrieved March 19, 2022, from https://www2.lbl.gov/Science-Articles/Archive/3D-tubulin.html#:~:text=Tubulin%20is%20the%20protein%20that,or%20to%20regulate%20intracellular%20transport.
2. Dali server. (n.d.). Ekhidna2.Biocenter.helsinki.fi. Retrieved March 14, 2022, from http://ekhidna2.biocenter.helsinki.fi/dali/
3. Bank, R. P. D. (n.d.-b). RCSB PDB - 4L8U: X-ray study of human serum albumin complexed with 9 amino camptothecin. Www.rcsb.org. Retrieved March 14, 2022, from https://www.rcsb.org/structure/4L8U
4. Dali server. (n.d.). Ekhidna2.Biocenter.helsinki.fi. Retrieved March 14, 2022, from http://ekhidna2.biocenter.helsinki.fi/barcosel/tmp//4l8uA/

DATE: 19-03-22

WEBLEM 8

Introduction to Gene Prediction and various elements in Prokaryotes and Eukaryotes

With the rapid accumulation of genomic sequence information, there is a pressing need to use computational approaches to accurately predict gene structure. Computational gene prediction is a prerequisite for detailed functional annotation of genes and genomes. The process includes detection of the location of open reading frames (ORFs) and delineation of the structures of introns as well as exons if the genes of interest are of eukaryotic origin. The ultimate goal is to describe all the genes computationally with near 100% accuracy. The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

CATEGORIES OF GENE PREDICTION PROGRAMS

The current gene prediction methods can be classified into two major categories, ab initio–based and homology-based approaches. The ab initio–based approach predicts genes based on the given sequence alone. It does so by relying on two major features associated with genes. The first is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites. In addition, the triplet codon structure limits the coding frame length to multiples of three, which can be used as a condition for gene prediction. The second feature used by ab initio algorithms is gene content, which is statistical description of coding regions. It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the noncoding regions. The unique features can be detected by employing probabilistic models such as Markov models or hidden Markov models to help distinguish coding from noncoding regions.

The homology-based method makes predictions based on significant matches of the query sequence with sequences of known genes. For instance, if a translated DNA sequence is found to be similar to a known protein or protein family from a database search, this can be strong evidence that the region codes for a protein. Alternatively, when possible exons of a genomic DNA region match a sequenced cDNA, this also provides experimental evidence for the existence of a coding region.

Some algorithms make use of both gene-finding strategies. There are also a number of programs that actually combine prediction results from multiple individual programs to derive a consensus prediction. This type of algorithms can therefore be considered as consensus based.

Gene Prediction Using Markov Models and Hidden Markov Models

Markov models and HMMs can be very helpful in providing finer statistical description of a gene. A Markov model describes the probability of the distribution of nucleotides in a DNA sequence, in which the conditional probability of a particular sequence position depends on k previous positions. In this case, k is the order of a Markov model. A zero-order Markov model assumes each base occurs independently with a given probability. This is often the case for noncoding sequences. A first-order Markov model assumes that the occurrence of a base depends on the base preceding it. A second-order model looks at the preceding two bases to determine which base follows, which is more characteristic of codons in a coding sequence.

The use of Markov models in gene finding exploits the fact that oligonucleotide distributions in the coding regions are different from those for the noncoding regions. These can be represented with various orders of Markov models. Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous k nucleotides, the longer the oligomer unit, the more non randomness can be describedfor the coding region. Therefore, the higher the order of a Markov model, the more accurately it can predict a gene.

FGENESB is a web-based program that is also based on fifth-order HMMs for detecting coding regions. The program is specifically trained for bacterial sequences. It uses the Vertibi algorithm to find an optimal match for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to

further distinguish coding signals from noncoding signals.

Step-by-Step Description of FGENESB annotation.

STEP 1. Finds all potential ribosomal RNA genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected rRNA genes.

STEP 2. Predicts tRNA genes using tRNAscan-SE program (Washington University) and masks detected tRNA genes.

STEP 3. Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes. Generates parameters such as 5th-order in-frame Markov chains for coding regions, 2nd-order Markov models for region around start codon and upstream RBS site, Stop codon and probability distributions of ORF lengths.

STEP 4. Predicts operons based only on distances between predicted genes.

STEP 5. Runs BLASTP for predicted proteins against COG database, cog.pro.

STEP 6. Uses information about conservation of neighboring gene pairs in known genomes to improve operon prediction.

STEP 7. Runs BLASTP against NR for proteins having no COGs hits.

STEP 8. Predicts potential promoters (BPROM program) or terminators (BTERM) in upstream and downstream regions, correspondingly, of predicted genes. BTERM is the program predicting bacterial - independent terminators with energy scoring based on discriminant function of hairpin elements.

STEP 9. Refines operon predictions using predicted promoters and terminators as additional evidences.

Prediction Using Discriminant Analysis.

Some gene prediction algorithms rely on discriminant analysis, either LDA or quadratic discriminant analysis (QDA), to improve accuracy. LDA works by plotting a two-dimensional graph of coding signals versus all potential 3_ splice site positions and drawing a diagonal line that best separates coding signals from noncoding signals based on knowledge learned from training data sets of known gene structures. QDA draws a curved line based on a quadratic function instead of drawing a straight line to separate coding and noncoding features. This strategy is designed to be more flexible and provide a more optimal separation between the data points.

FGENES is a web-based program that uses LDA to determine whether a signal is an exon. In addition to FGENES, there are many variants of the program. Some programs, such as FGENESH, make use of HMMs. There are others, such as FGENESH C, that are similarity based. Some programs, such as FGENESH+, combine both ab initio and similarity-based approaches.

Output format:

- G - predicted gene number, starting from start of sequence
- Str - DNA strand (+ for direct or - for complementary)
- Feature - type of coding sequence: CDSf - First (Starting with Start codon), CDSi - internal (internal exon), CDSl - last coding segment, ending with stop codon)
- TSS - Position of transcription start (TATA-box position and score)
- TSS - position of transcription start
- TATA - position of TATA-box
- wTATA - Discriminant function score for TATA box
- Start and End - Position of the Feature
- Weight - Discriminant function score for the feature
- ORF - start/end positions where the first complete codon starts and the last codon ends

An issue related to gene prediction is promoter prediction. Promoters are DNA elements located in the vicinity of gene start sites (which should not be confused with the translation start sites) and serve as binding sites for the gene transcription machinery, consisting of RNA polymerases and transcription

factors. Therefore, these DNA elements directly regulate gene expression. Promoters and regulatory elements are traditionally determined by experimental analysis. The process is extremely time consuming and laborious. Computational prediction of promoters and regulatory elements is especially promising because it has the potential to replace a great deal of extensive experimental analysis.

PREDICTION ALGORITHMS

Current algorithms for predicting promoters and regulatory elements can be categorized as either ab initio based, which make de novo predictions by scanning individual sequences; or similarity based, which make predictions based on alignment of homologous sequences; or expression profile based using profiles constructed from a number of co-expressed gene sequences from the same organism. The similarity type of prediction is also called phylogenetic foot-printing.

PREDICTION FOR PROKARYOTES

One of the unique aspects in prokaryotic promoter prediction is the determination of operon structures, because genes within an operon share a common promoter located upstream of the first gene of the operon. Thus, operon prediction is the key in prokaryotic promoter prediction. Once an operon structure is known, only the first gene is predicted for the presence of a promoter and regulatory elements, whereas other genes in the operon do not possess such DNA elements.

There are a number of methods available for prokaryotic operon prediction. The most accurate is a set of simple rules developed by Wang et al. This method relies on two kinds of information: gene orientation and intergenic distances of a pair of genes of interest and conserved linkage of the genes based on comparative genomic analysis. A scoring scheme is developed to assign operons with different levels of confidence. This method is claimed to produce accurate identification of an operon structure, which in turn facilitates the promoter prediction.

This newly developed scoring approach is, however, not yet available as a computer program. The prediction can be done manually using the rules, however. The few dedicated programs for prokaryotic promoter prediction do not apply the Wang et al. rule for historical reasons. The most frequently used program is BPROM.

BPROM is a web-based program for prediction of bacterial promoters. It uses a linear discriminant function combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites. This programfirst predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than theWang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

Output format:

- First line - name of your sequence;
- Second and Third lines - LDF threshold and the length of presented sequence
- 4th line - The number of predicted promoters
- Next lines - positions of predicted promoters, and their scores with 'weights' of two conserved promoter boxes. Promoter position assign to the first nucleotide of the transcript (Transcription Start Site position).
- After that we present elements of Transcriptional factor binding sites for each predicted promoter (if they found).

PREDICTION FOR EUKARYOTES

The ab initio method for predicting eukaryotic promoters and regulatory elements also relies on searching the input sequences for matching of consensus patterns of known promoters and regulatory elements. The consensus patterns are derived from experimentally determined DNA binding sites which are compiled into

profiles and stored in a database for scanning an unknown sequence to find similar conserved patterns. However, this approach tends to generate very high rate of false positives owing to nonspecific matches with the short sequence patterns. Furthermore, because of the high variability of transcription factor binding sites, the simple sequence matching often misses true promoter sites, creating false negatives.

To increase the specificity of prediction, a unique feature of eukaryotic promoter is employed, which is the presence of CpG islands. It is known that many vertebrate genes are characterized by a high density of CG dinucleotides near the promoter region overlapping the transcription start site. By identifying the CpG islands, promoters can be traced on the immediate upstream region from the islands. By combining CpG islands and other promoter signals, the accuracy of prediction can be improved. Several programs have been developed based on the combined features to predict the transcription start sites in particular.

The eukaryotic transcription initiation requires cooperation of a large number of transcription factors. Cooperativity means that the promoter regions tend to contain a high density of protein-binding sites. Thus, finding a cluster of transcription factor binding sites often enhances the probability of individual binding site prediction. TSSW is a web program that distinguishes promoter sequences from non-promoter sequences based on a combination of unique content information such as hexamer/trimer frequencies and signal information such the TATA box in the promoter region. The values are fed to a linear discriminant function to separate true motifs from background noise.

Output format:

- First line - name of your sequence;
- Second and Third lines - LDF threshold and the length of presented sequence
- 4th line - The number of predicted promoter regions
- Next lines - positions of predicted sites, their 'weights' and TATA box position (if found)
- Position shows the first nucleotide of the transcript (TSS position)
- After that functional motifs are given for each predicted region; (+) or (-) reflects the direct or complementary chain; S... means a particular motif identificator from the Wingender data base.
- Lower cased letters mean non-conserved nucleotides in the site consensus
- The letters except (A,T,G,C) describe ambiguous sites in a given DNA sequence motif, where a single character may represent more than one nucleotide using Standard IUPAC Nucleotide code.

ORF finder:

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP. This web version of the ORF finder is limited to the sub range of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation.

Thus, TSSW and BPROM are a useful tool for the recognition promoter region and start of transcription. Understanding the regulation of gene expression is an important aspect of understanding the gene function, thus this tool will be useful in elucidating regulation and expression mechanisms of genes and may even shed light on the function of novel and uncharacterized genes with the help of knowledge of promoters. FGENESB tool is useful for prediction of bacterial operon and gene and FGENES for prediction of exons. Identifying the genes that are grouped together into operons may enhance our knowledge of gene regulation and function, and such information is an important addition to genome annotation. All this can be done with the help of FGENESB. ORF finder can used to predict open reading frames in the genome. This information of long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence. Small Open Reading Frames (small ORFs/sORFs/smORFs) are important sources of putative peptides previously dismissed as being non-functional or junk DNA.

REFERENCES:

1. Xiong, J. (2008).Gene Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 97-111.
2. Xiong, J. (2008). Promoter and Regulatory Element Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 113-119.
3. TSSW - Recognition of human PolII promoter region and start of transcription. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter
4. BPROM - Prediction of bacterial promoters. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb
5. FGENESB - Bacterial Operon and Gene Prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022 from http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb
6. FGENES - pattern-based gene structure prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind
7. Home - ORFfinder – NCBI. (2019). Nih.gov. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/orffinder/

DATE: 19-03-22

WEBLEM 8a
TSSW

(URL: http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter)

AIM:

To recognize the Saccharomyces cerevisiae kinase and start of transcription using TSSW tool.

INTRODUCTION:

kinase, an enzyme that adds phosphate groups (PO_4^{3-}) to other molecules. A large number of kinases exist—the human genome contains at least 500 kinase-encoding genes. Included among these enzymes' targets for phosphate group addition (phosphorylation) are proteins, lipids, and nucleic acids. Saccharomyces cerevisiae kinase and start of transcription can to recognized using TSSW.

TSSW is a web program that distinguishes promoter sequences from non-promoter sequences based on a combination of unique content information such as hexamer/trimer frequencies and signal information such the TATA box in the promoter region. The values are fed to a linear discriminant function to separate true motifs from background noise.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under search for promotor/functional motifs select TSSW. (URL: http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter)
3. Retrieve nucleotide FASTA sequence for protease from GenBank.
4. Process the FASTA sequence on TSSW.
5. Observe and interpret the results.

OBSERVATION:

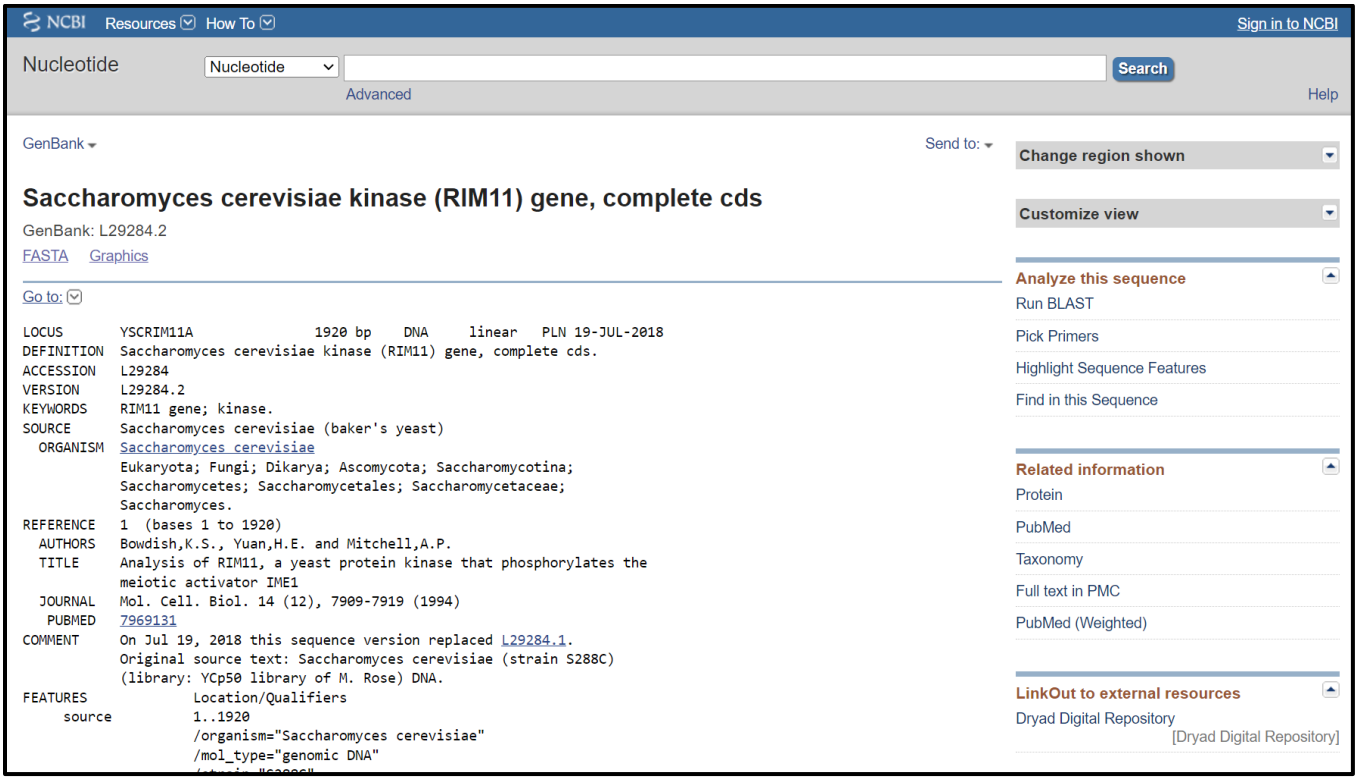


Fig1. GenBank result for Human kinase

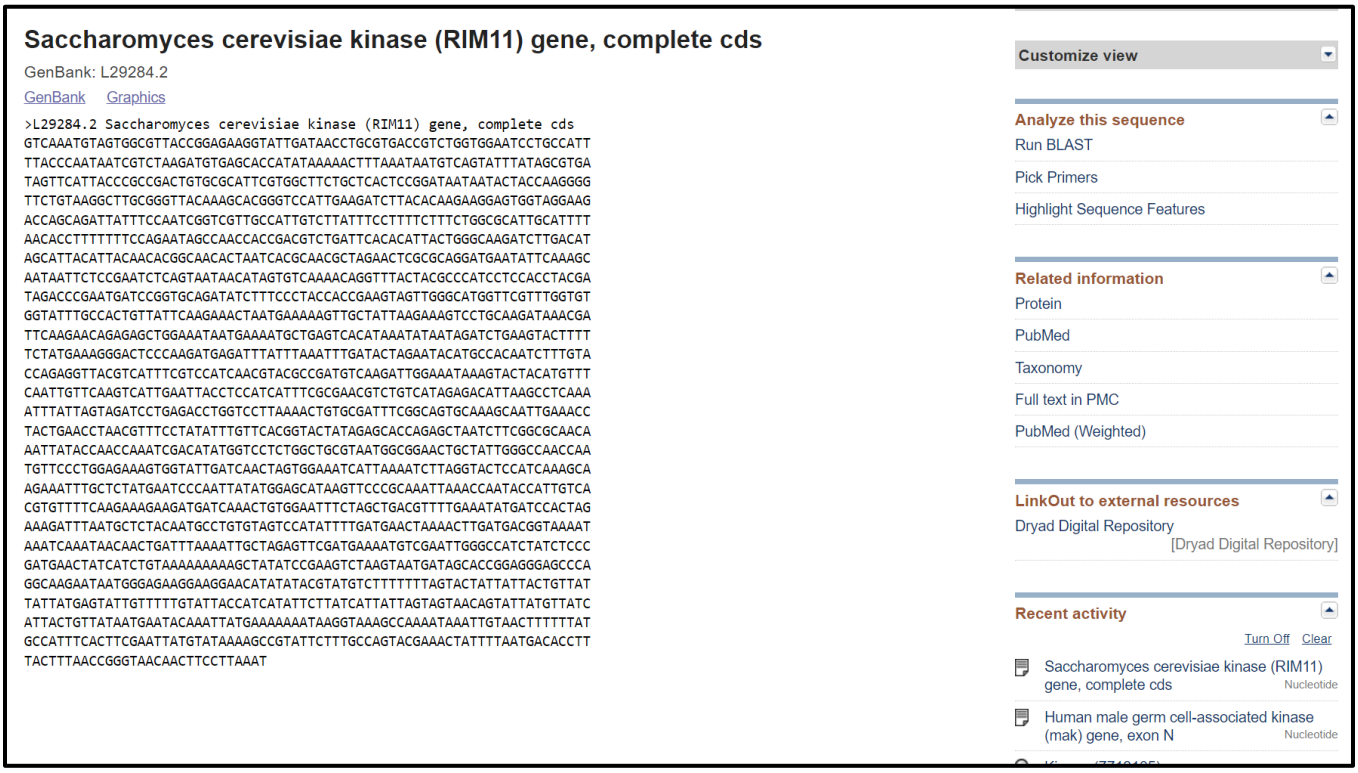


Fig2. Nucleotide FASTA sequence for Protease

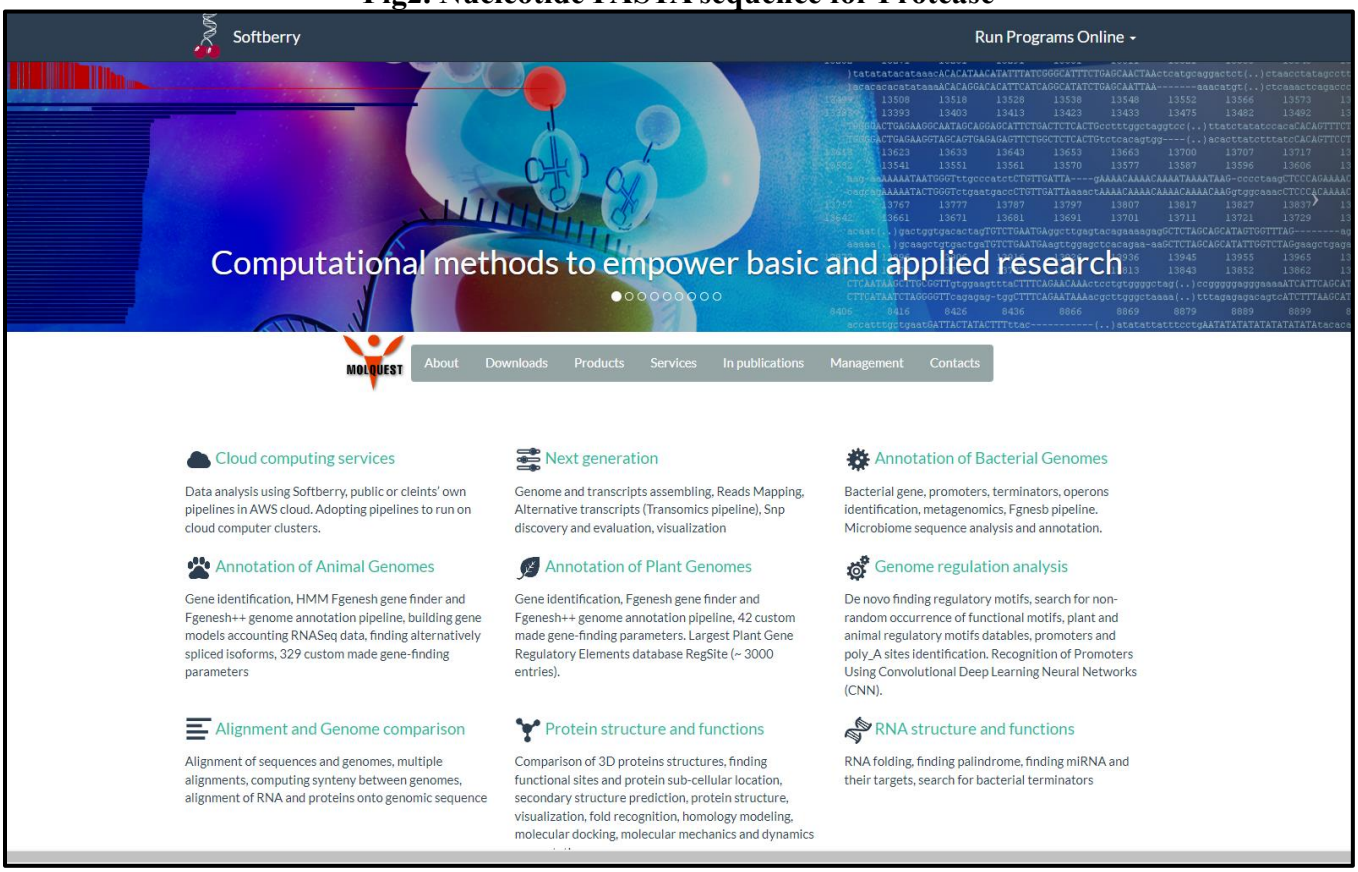


Fig3. Homepage for Softberry

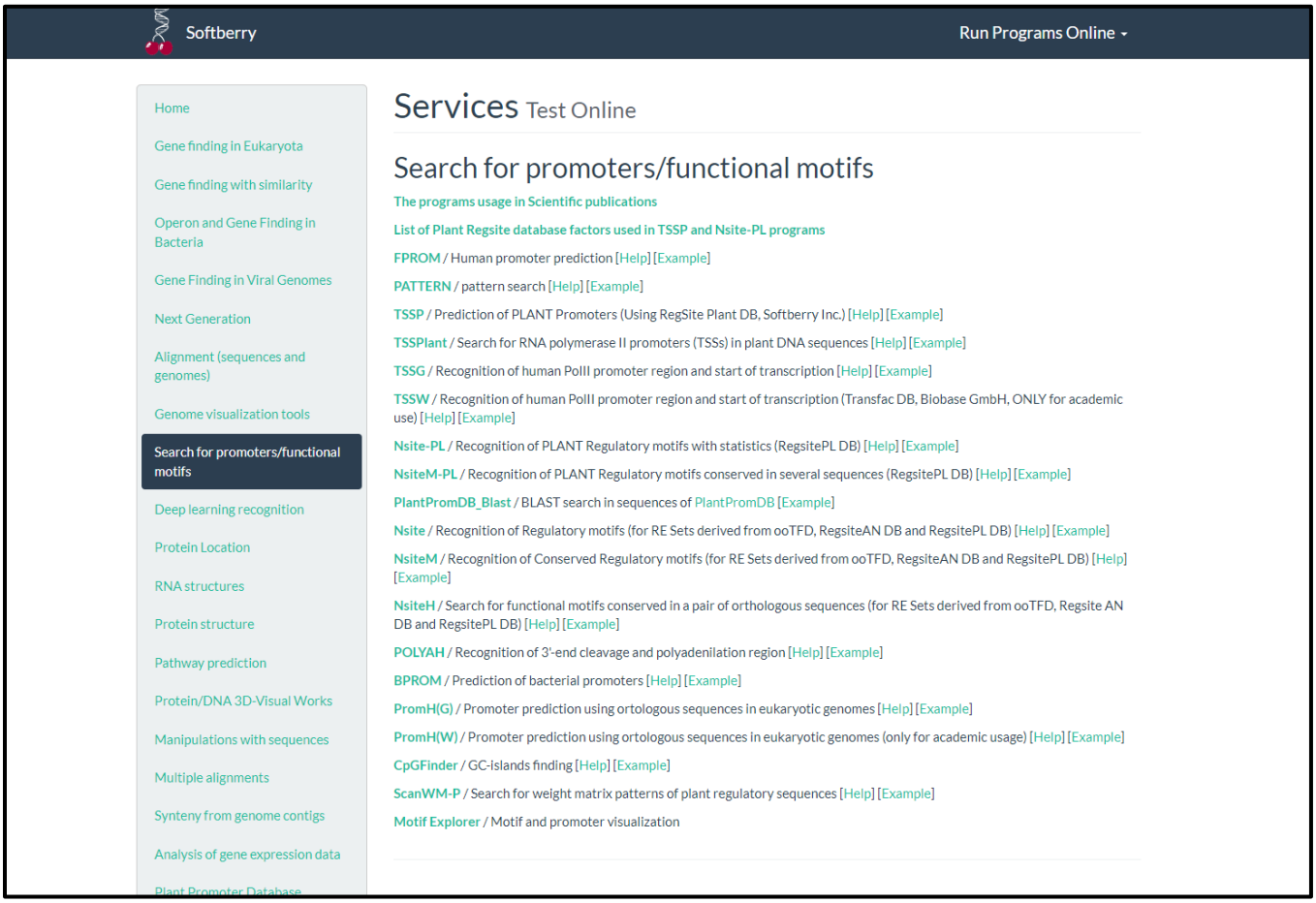


Fig4. Tools for promoters/functional motifs

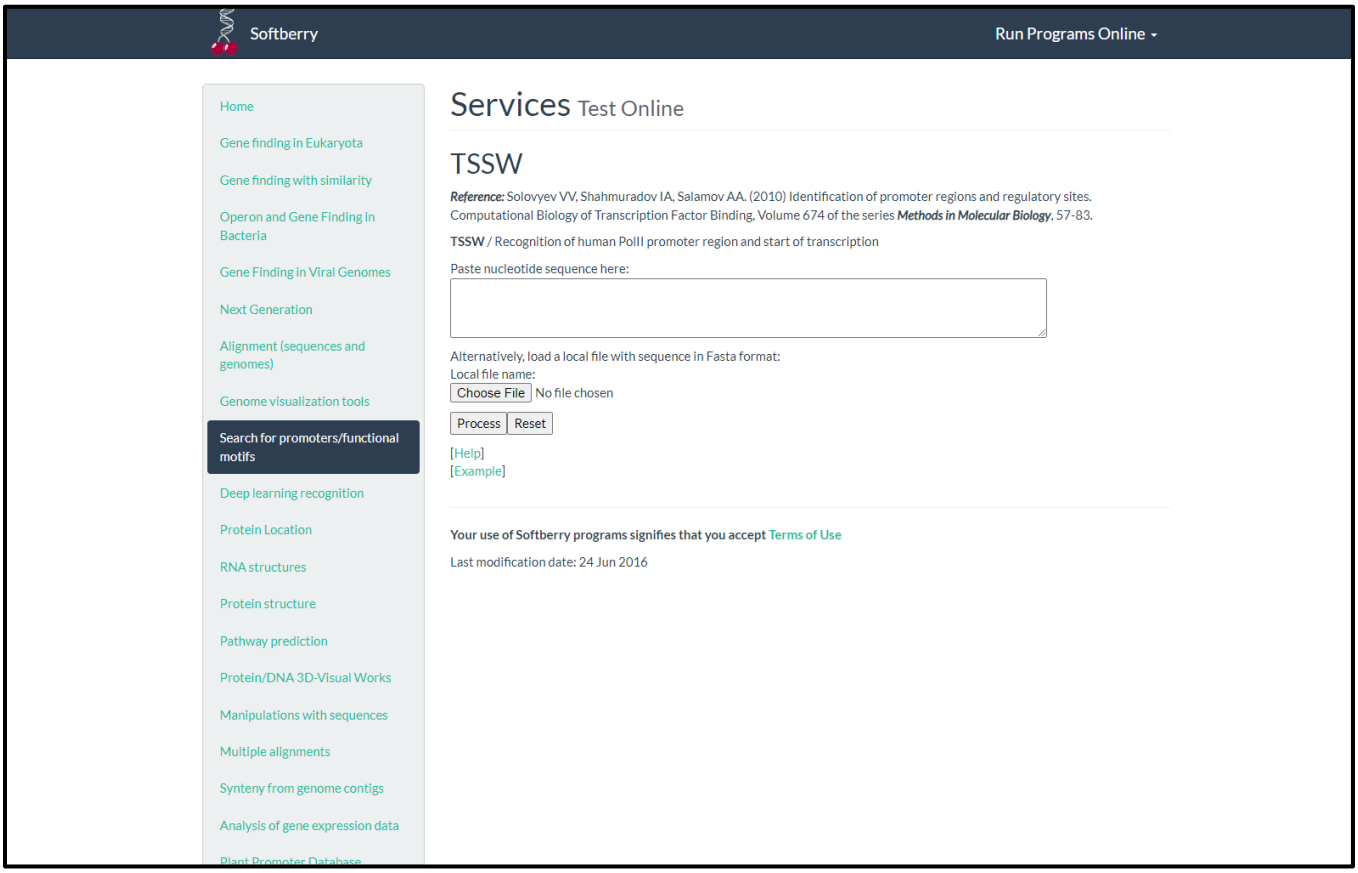


Fig5. Homepage for TSSW

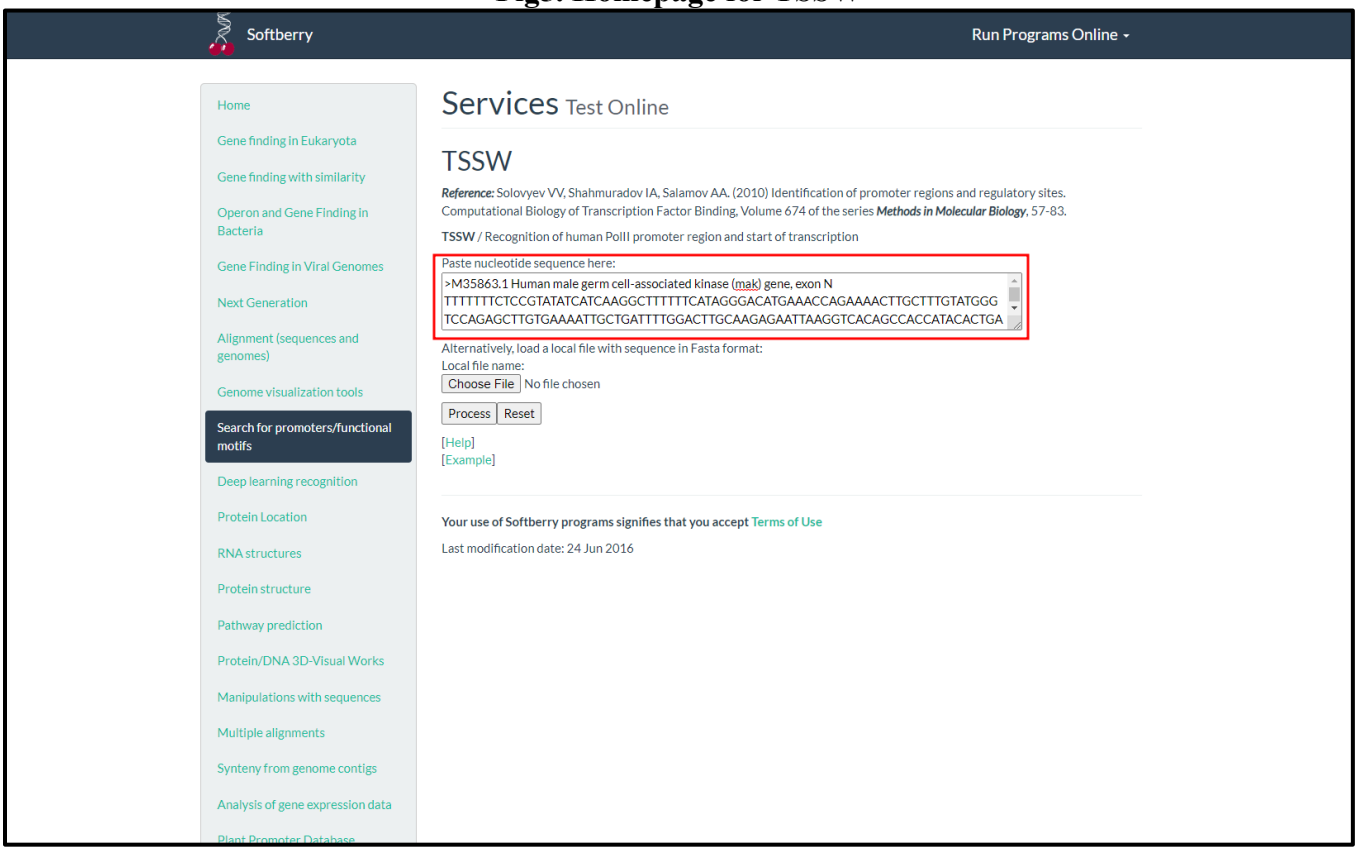


Fig6. Search for protease nucleotide FASTA sequence

```
>L29284.2 Saccharomyces cerevisiae kinase (RIM11) gene, complete cds
 Length of sequence-     1920
 Thresholds for TATA+ promoters -  0.45, for TATA-/enhancers -  3.70
    3 promoter/enhancer(s) are predicted
 Enhancer Pos:   1391 LDF-  5.79
 Promoter Pos:   1390 LDF-  5.75
 Promoter Pos:   1870 LDF-  0.84 TATA box at   1841    21.05
 Transcription factor binding sites:
for promoter at position -    1391
  1187 (+) HS$BAC_03    CCAAT
  1281 (+) HS$BAC_03    CCAAT
  1316 (+) HS$BAC_03    CCAAT
  1181 (-) HS$BAC_03    CCAAT
  1289 (-) CHICK$BAC_   TATAA
  1127 (-) CHICK$BAC_   TATAA
  1203 (-) Y$ADH2_01    TCTCC
  1185 (+) RAT$ALBU_2   AACCAAT
  1314 (+) RAT$ALBU_2   AACCAAT
  1327 (+) Y$CFES_01    GTCACGTG
  1177 (+) MOUSE$A21C   ATTGG
  1320 (-) MOUSE$A21C   ATTGG
  1285 (-) MOUSE$A21C   ATTGG
  1191 (-) MOUSE$A21C   ATTGG
  1164 (+) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1333 (-) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1298 (-) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1204 (-) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1201 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
  1146 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
  1297 (-) AD$E3_06     gggcagggTATAActcacctga
  1135 (-) AD$E3_06     gggcagggTATAActcacctga
  1380 (-) AD$E4_16     ACGTCA
  1215 (-) HS$EGFR_15   TCAAT
  1378 (-) RAT$EAI_09   GTCAG
  1283 (+) Y$GAL1_02    AcTTATAT
  1290 (-) Y$GAL1_12    ATATAA
  1110 (-) HS$BG_05     AAGATTAG
  1239 (-) MOUSE$BMG_   cagtagtTGATTgagca
  1099 (+) MOUSE$BMG_   aggggccAATCTgctc
  1231 (+) MOUSE$BMG_   aggggccAATCTgctc
  1187 (+) HS$GG_17     CCAATag
  1281 (+) HS$GG_17     CCAATag
  1316 (+) HS$GG_17     CCAATag
  1181 (-) HS$GG_17     CCAATag
  1286 (+) RAT$GLU_04   TATAT
  1231 (+) HS$GMCSF_0   CATTA
  1165 (-) HS$GMCSF_0   CATTA
  1389 (-) HS$GMCSF_0   TATTT
  1340 (-) HS$HO_01     ctggccCACGTGACccgc
  1231 (-) HS$HH4_01    GATTTC
  1146 (+) HS$HH4_02    GGTCC
  1335 (-) HA$HMGCR_0   gttCGTGACgtaggccg
  1137 (-) HS$IGH_04    ATTTGcgT
  1229 (-) HS$IGKL_01   TTTCCA
  1223 (+) RAT$INS_01   GTGGAAA
```

Fig7. Result for recognised promoter regions

RESULT:

Nucleotide FASTA sequence for Homo sapiens neutral protease of length 3298bps was submitted. With LDF threshold of 0.45 for promoters and 3.70 for enhancers, 3 promoters at position 809, 323 and 2884 with 9.40, 6.96 and 3.70 LDF values were recognised. 1 enhancer at position 319 with 7.38 LDF was also recognised.

CONCLUSION:

TSSW is a useful tool for the recognition of human Pol III promoter region and start of transcription. Understanding the regulation of gene expression is an important aspect of understanding the gene function, thus this tool will be useful in elucidating regulation and expression mechanisms of genes and may even shed light on the function of novel and uncharacterized genes with the help of knowledge of promoters.

REFERENCE:

1. Xiong, J. (2008). Promoter and Regulatory Element Prediction. Essential bioinformatics. Cambridge Cambridge University Press. 113-119.
2. PROTEOLYTIC ENZYMES (PROTEASES): Overview, Uses, Side Effects, Precautions, Interactions Dosing and Reviews. (n.d.). Www.webmd.com. Retrieved March 18, 2022, from https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20(proteases)%20are%20enzymes
3. Homo sapiens neutral protease alpha subunit gene, complete cds. (2016). NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/AH001431.2?report=genbank
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
5. TSSW - Recognition of human PolII promoter region and start of transcription. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter

6. Softberry - TSSW result. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/promoter/tssw.pl

DATE: 19-03-22

WEBLEM 8b
BPROM

(URL: http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb)

AIM:

To predict bacterial promoter for Enterococcus Alcedinis using BPROM tool.

INTRODUCTION:

Two Gram-positive, catalase-negative bacterial strains were isolated from the cloaca of common kingfishers (Alcedo atthis). Repetitive sequence-based PCR fingerprinting using the (GTG)5 primer grouped these isolates into a single cluster separated from all known enterococcal species. The two strains revealed identical 16S rRNA gene sequences placing them within the genus Enterococcus with Enterococcus aquimarinus LMG 16607(T) as the closest relative (97.14 % similarity).

BPROM is a web-based program for prediction of bacterial promoters. It uses a linear discriminant function combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites. This program first predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than the Wang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under operon and gene finding select BPROM. (URL:http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb)
3. Retrieve bacterial nucleotide FASTA sequence from GenBank.
4. Process the FASTA sequence on BPROM.
5. Observe and interpret the results.

OBSERVATION:

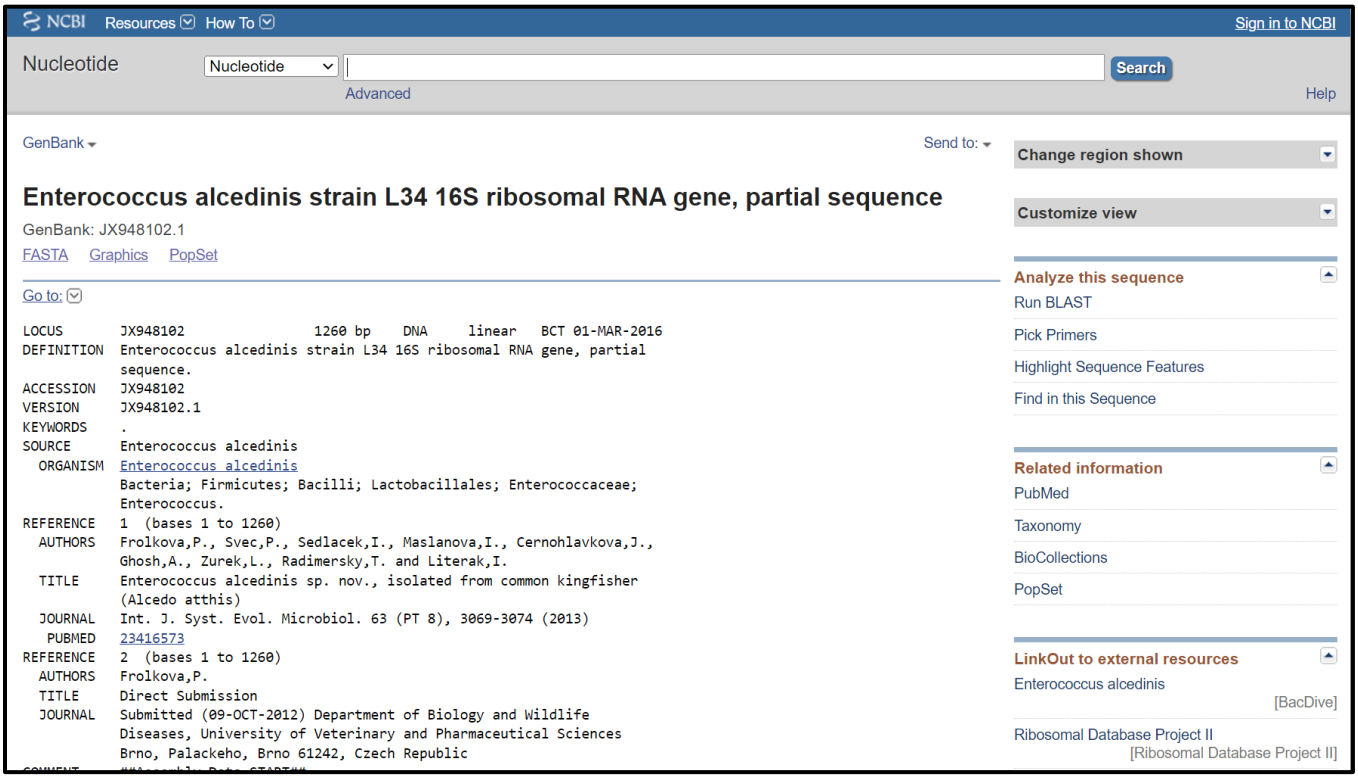


Fig1. GenBank result for Neisseria gonorrhoeae

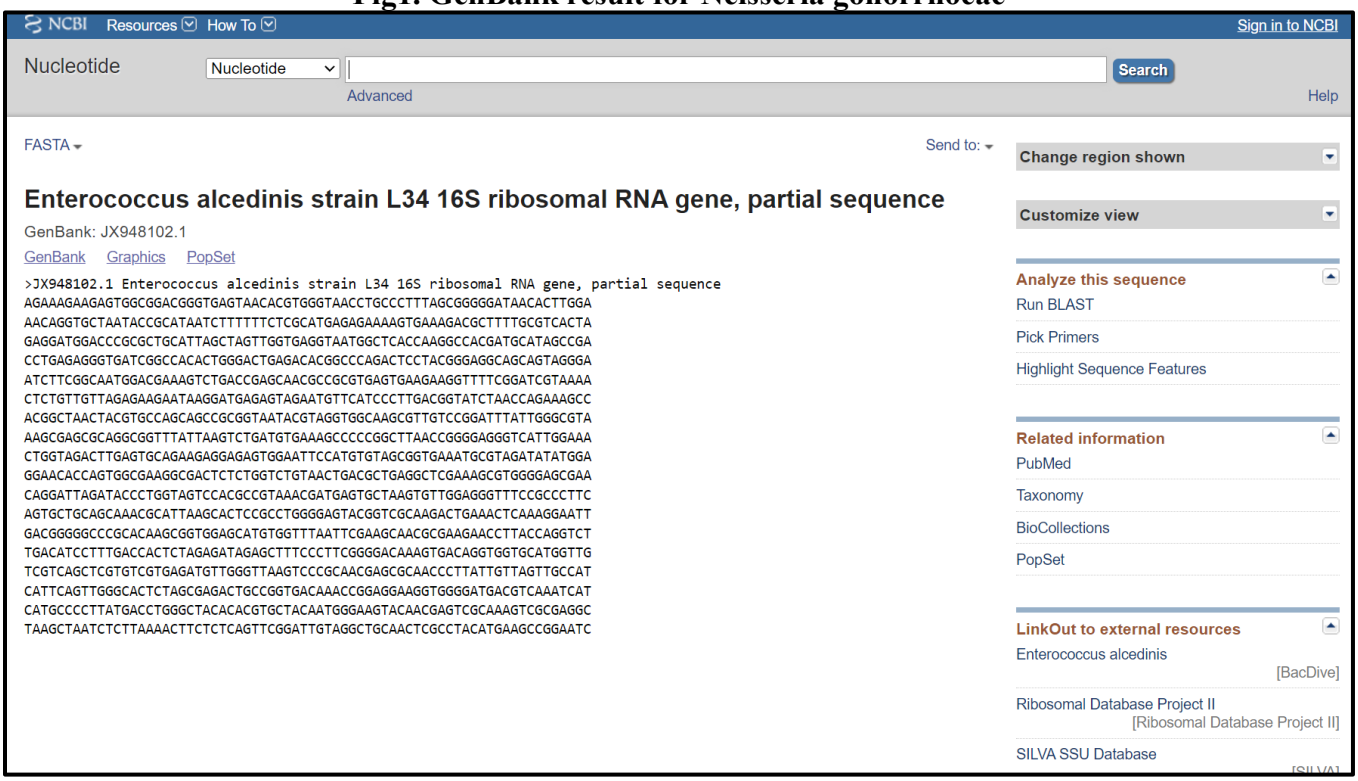


Fig2. Nucleotide FASTA sequence

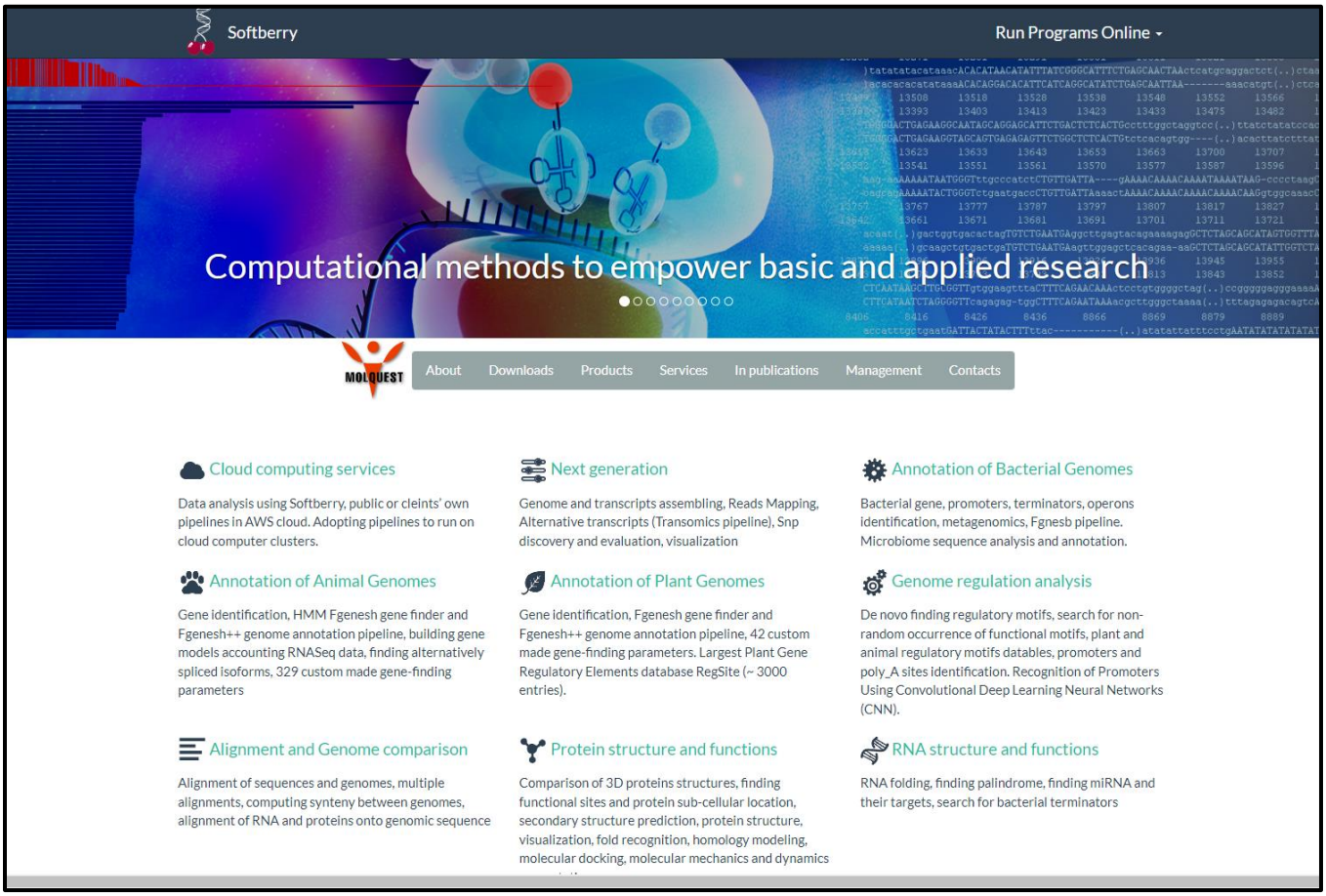


Fig3. Homepage for softberry

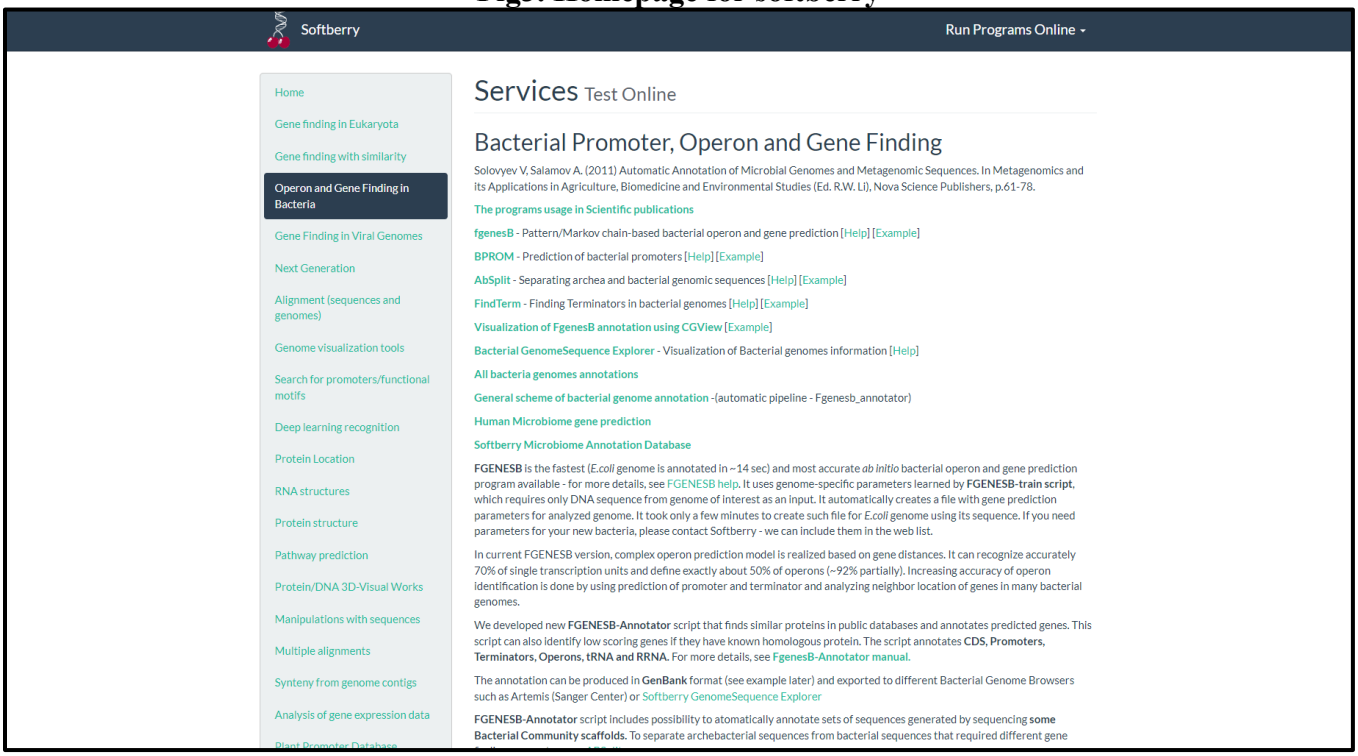


Fig4. Tools for bacterial promoter, operon and gene finding

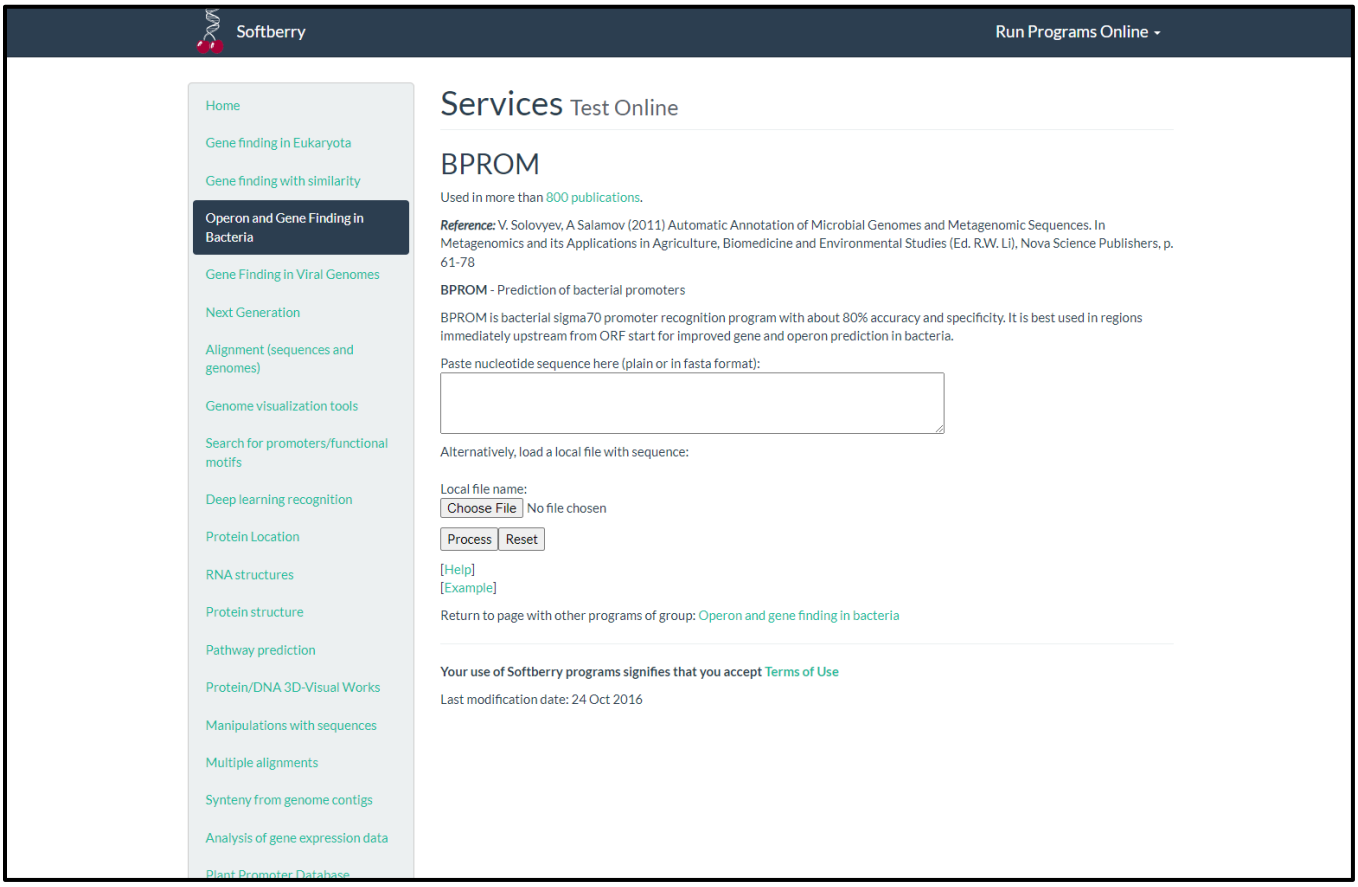


Fig5. Homepage for BPROM

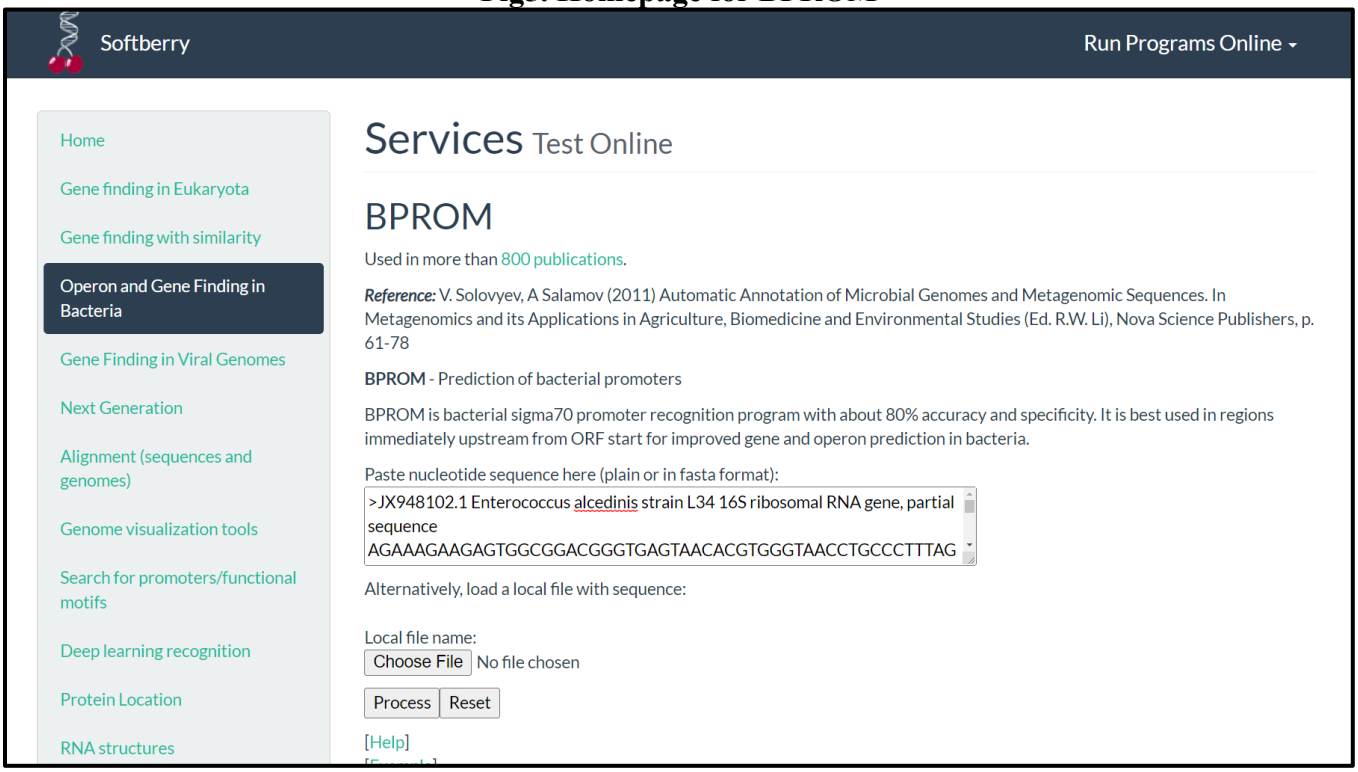


Fig6. Homepage for nucleotide FASTA sequence

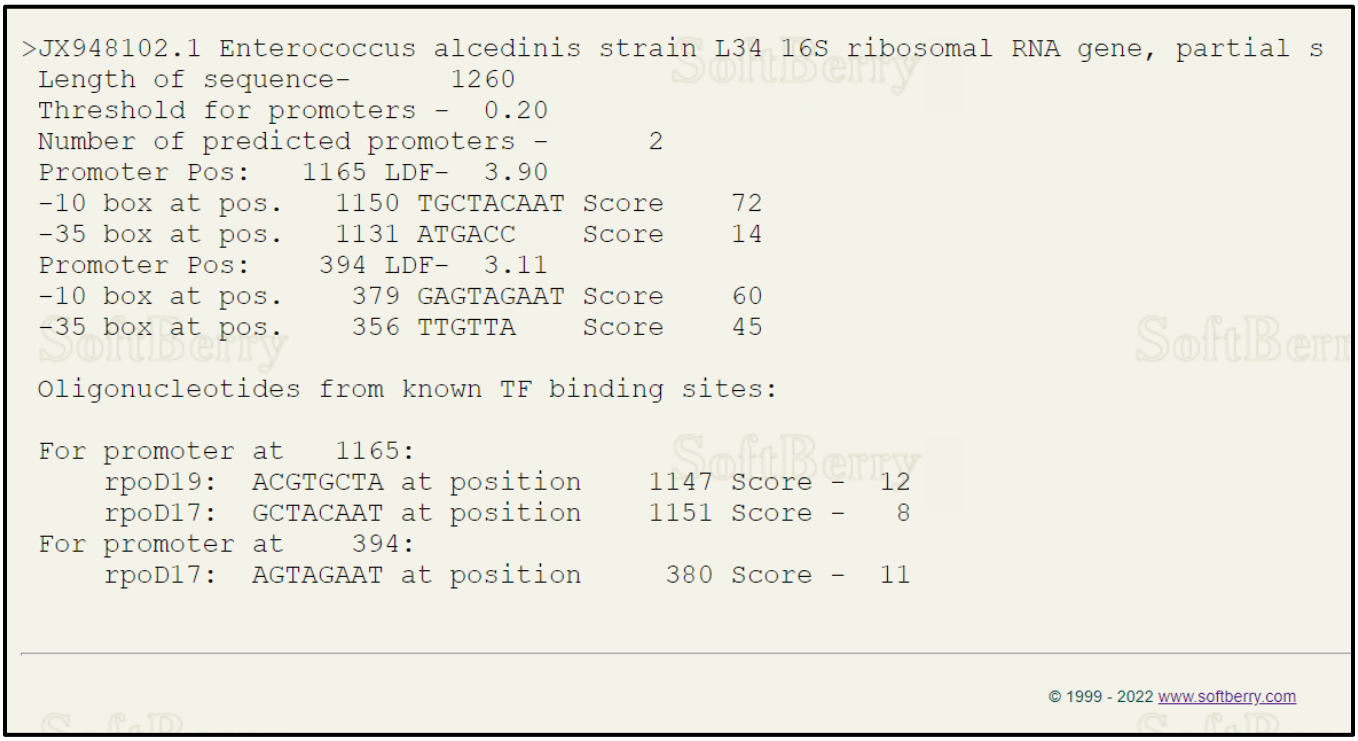


Fig7. Result for predicted promoters and known TF binding sites

RESULT:

Nucleotide FASTA sequence for Enterococcus alcedinis strain L34 16S ribosomal RNA gene, partial sequence with length 1260 was submitted. With threshold LDF value of 0.20, 9 promoters were predicted using BPROM with their TF binding sites.

CONCLUSION:

BPROM is a useful tool for the recognition of bacterial promoter region. Understanding the regulation of gene expression is an important aspect of understanding the gene function, thus this tool will be useful in elucidating regulation and expression mechanisms of genes and may even shed light on the function of novel and uncharacterized genes with the help of knowledge of promoters.

REFERENCES:

1. Xiong, J. (2008). Promoter and Regulatory Element Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 113-119.
2. Neisseria gonorrhoeae - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com. Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
3. Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome shotgun sequence. (2022).NCBI Nucleotide. Retrieved March 18, 2022, fromhttps://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, fromhttp://www.softberry.com/
5. BPROM - Prediction of bacterial promoters. (n.d.). Www.softberry.com. Retrieved March 18, 2022,from http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb
6. Softberry - BPROM result. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfindb/bprom.pl

DATE: 19-03-22

WEBLEM 8c
FGENESB
(URL: http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb)

AIM:

To predict bacterial operon and gene for *Enterococcus alcedinis* using FGENESB tool.

INTRODUCTION:

Two Gram-positive, catalase-negative bacterial strains were isolated from the cloaca of common kingfishers (Alcedo atthis). Repetitive sequence-based PCR fingerprinting using the (GTG)5 primer grouped these isolates into a single cluster separated from all known enterococcal species. The two strains revealed identical 16S rRNA gene sequences placing them within the genus Enterococcus with Enterococcus aquimarinus LMG 16607(T) as the closest relative (97.14 % similarity).

FGENESB is a web-based program that is also based on fifth-order HMMs for detecting coding regions. The program is specifically trained for bacterial sequences. It uses the Vertibi algorithm to find an optimal match for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to further distinguish coding signals from noncoding signals.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under operon and gene finding in bacteria select FGENESB. (URL:http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb)
3. Retrieve bacterial nucleotide FASTA sequence from GenBank.
4. Process the FASTA sequence on FGENESB.
5. Observe and interpret the results.

OBSERVATION:

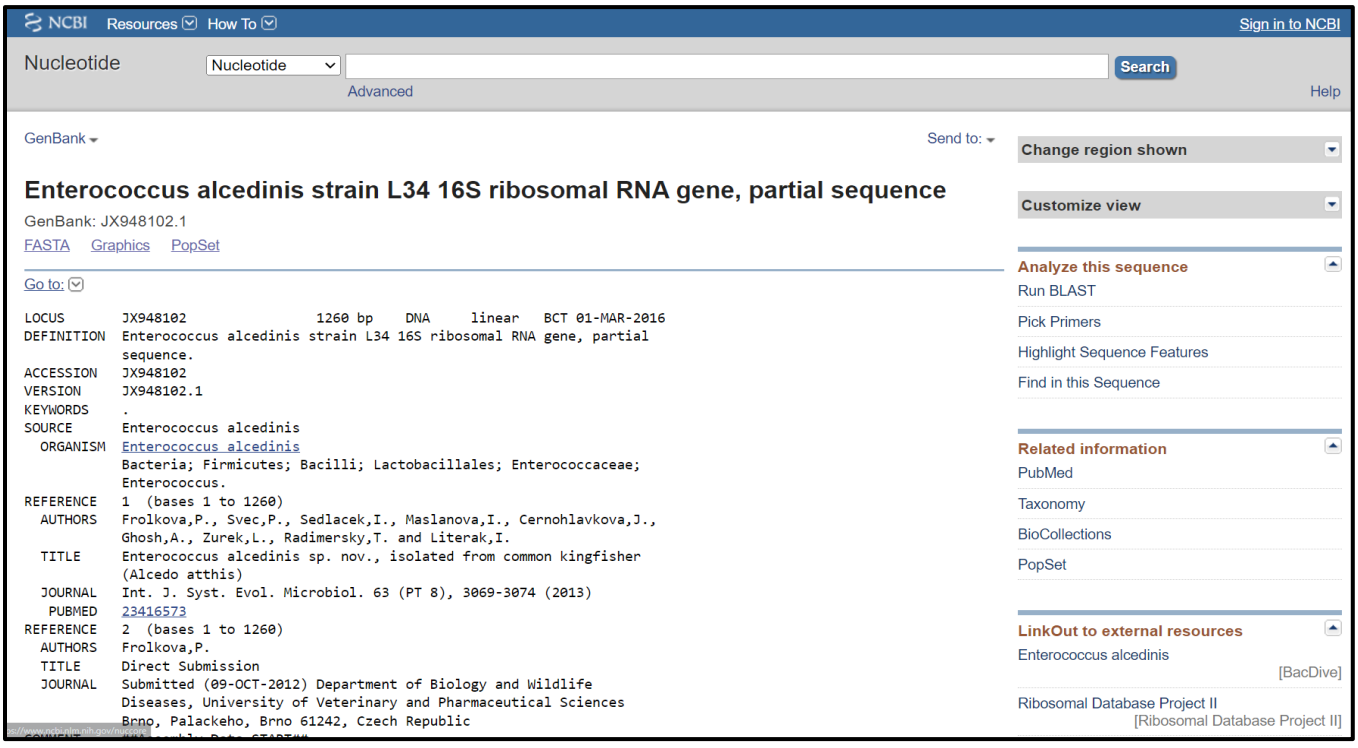


Fig1. GenBank result for *Enterococcus alcedinis*

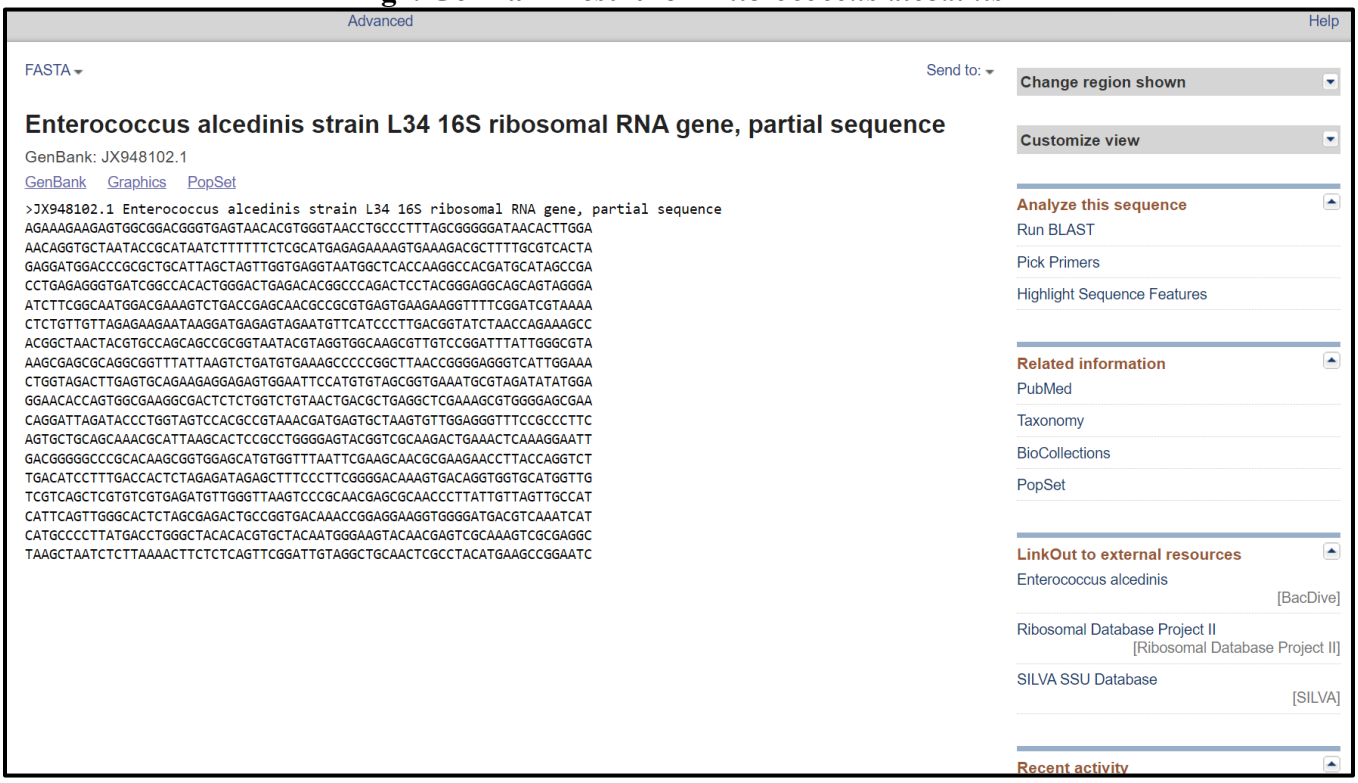


Fig2. Nucleotide FASTA sequence

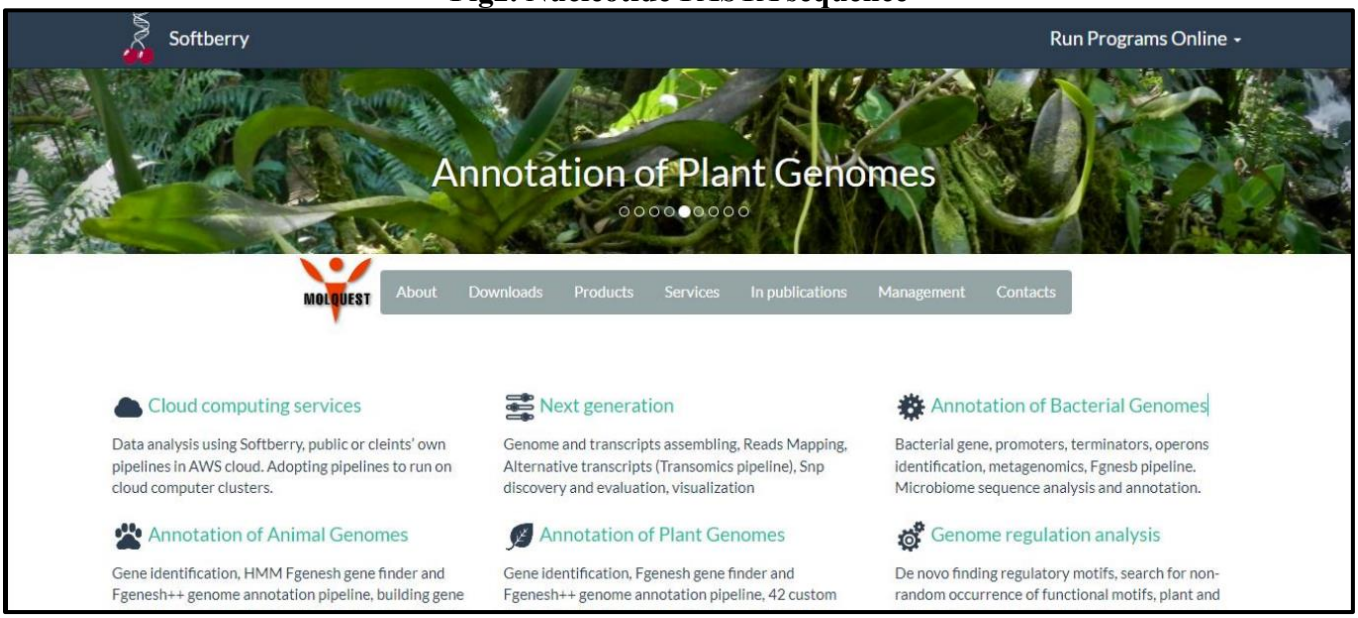


Fig3. Homepage for softberry

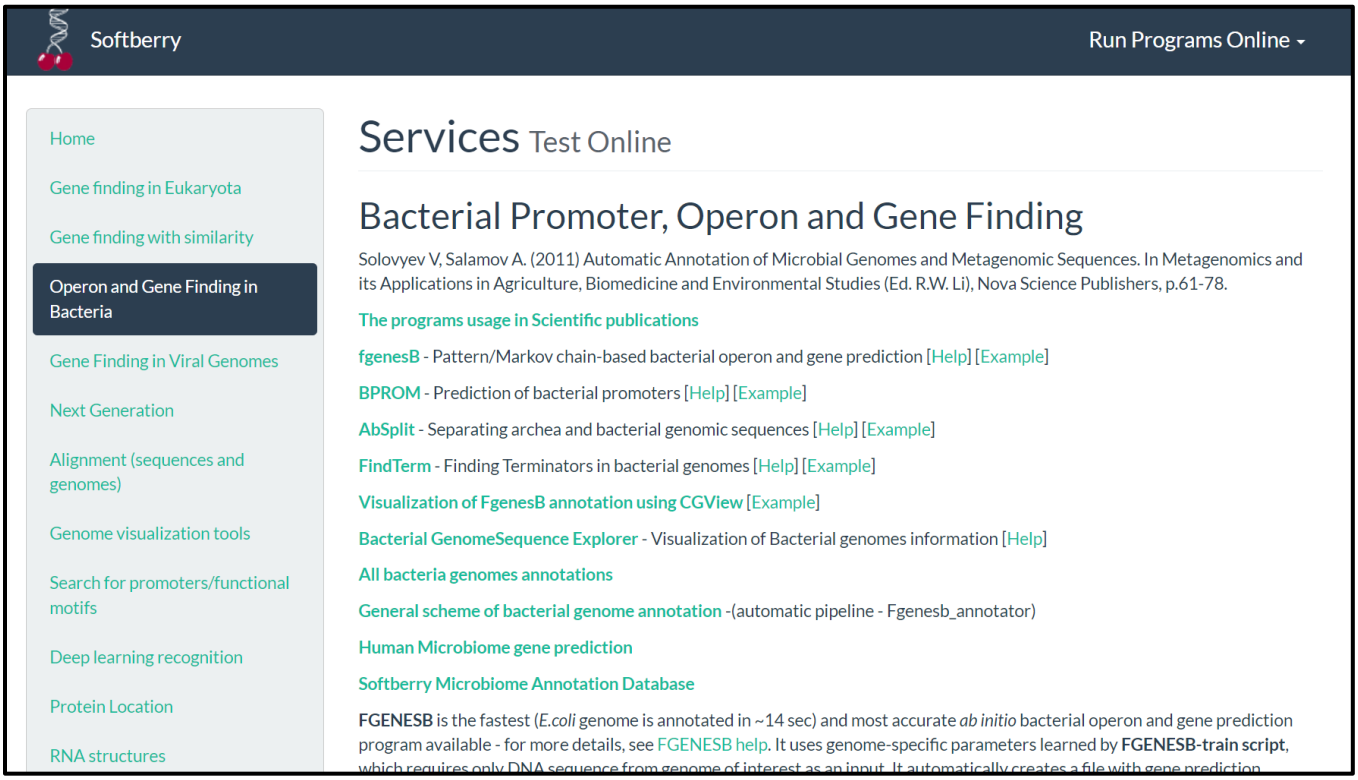


Fig4. Tools for bacterial promoter, operon and gene finding

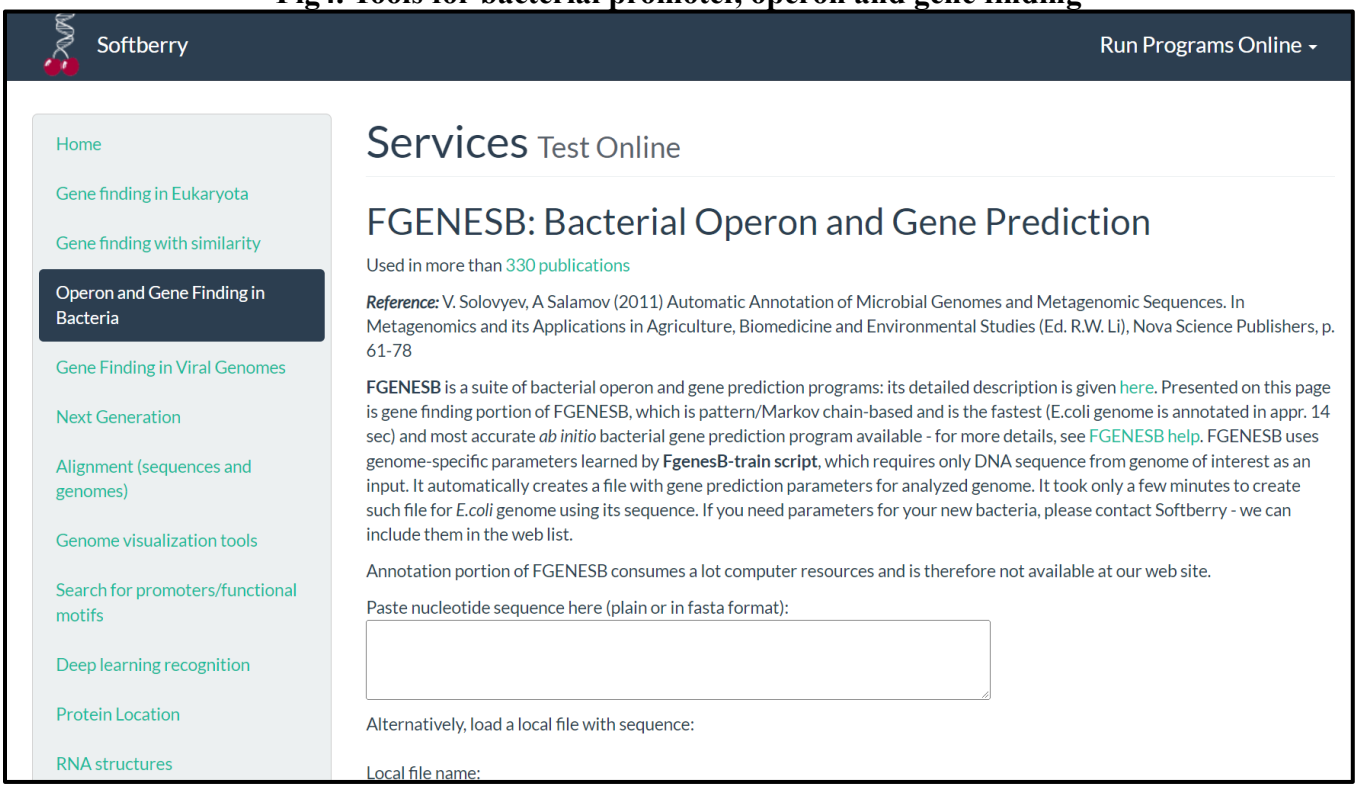


Fig5. Homepage for FGENESB

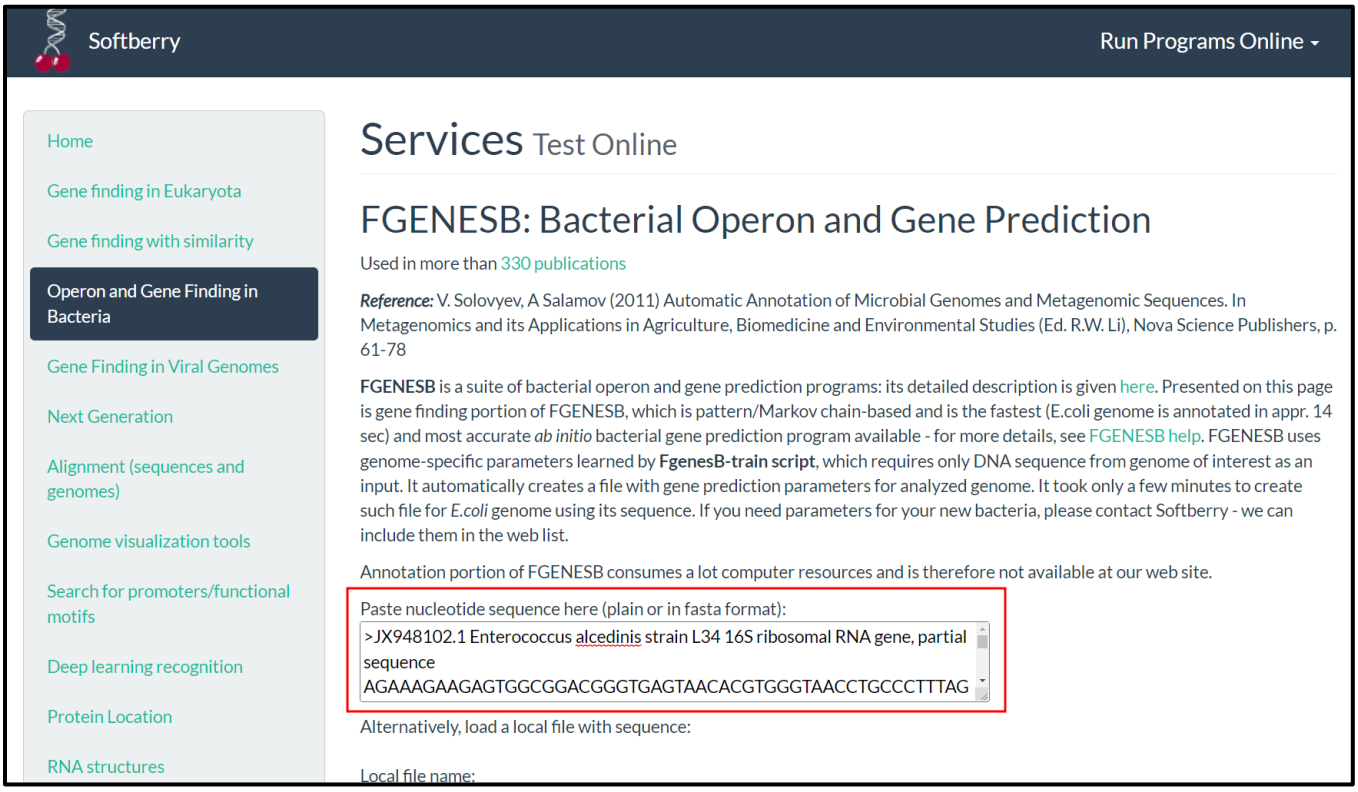


Fig6. Search for nucleotide FASTA sequence

Fig7. Result for predicted bacterial operon and genes

RESULT:

Nucleotide FASTA sequence for Enterococcus alcedinis strain L34 16S ribosomal RNA gene, partial sequence with length of 1260 was submitted and 9 genes, 5 transcriptional units and 2 operons were predicted using FGENESB.

CONCLUSION:

FGENESB tool is useful for prediction of bacterial operon and gene. Gene prediction information is a prerequisite for detailed functional annotation of genes and genomes. Identifying the genes that are grouped together into operons may enhance our knowledge of gene regulation and function, and such information is an important addition to genome annotation. All this can be done with the help of FGENESB.

REFERENCES:

1. Xiong, J. (2008).Gene Prediction. Essential bioinformatics. Cambridge: Cambridge University Press.97111.
2. Neisseria gonorrhoeae - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com.Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
3. Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome shotgun sequence. (2022).NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
5. FGENESB - Bacterial Operon and Gene Prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb
6. Softberry - fgenesB results. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfindb/fgenesb.pl

DATE: 19-03-22

WEBLEM 8d
FGENES

(URL: http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind)

AIM:

To predict exon signals in Kinase using FGENES tool.

INTRODUCTION:

Kinase, an enzyme that adds phosphate groups (PO_4^{3-}) to other molecules. A large number of kinases exist—the human genome contains at least 500 kinase-encoding genes. Included among these enzymes' targets for phosphate group addition (phosphorylation) are proteins, lipids, and nucleic acids. Saccharomyces cerevisiae kinase and start of transcription can to recognized using FGENES.

FGENES is a web-based program that uses LDA to determine whether a signal is an exon. In addition to FGENES, there are many variants of the program. Some programs, such as FGENESH, make use of HMMs. There are others, such as FGENESH C, that are similarity based. Some programs, such as FGENESH+, combine both ab initio and similarity-based approaches.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under Gene for Eukaryotes select FGENES. (URL:http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind)
3. Retrieve nucleotide FASTA sequence for protease from GenBank.
4. Process the FASTA sequence on FGENES.
5. Observe and interpret the results.

OBSERVATION:

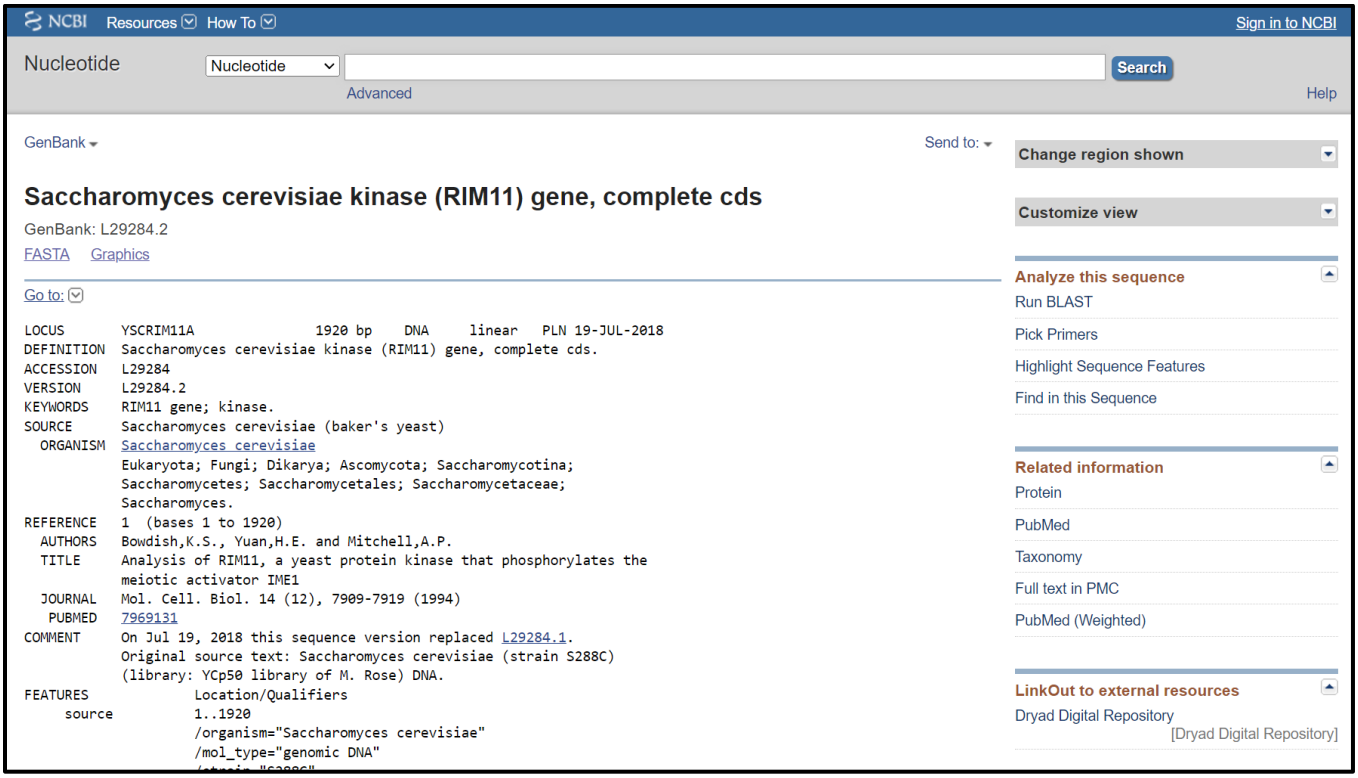


Fig1. GenBank result for Saccharomyces cerevisiae kinase

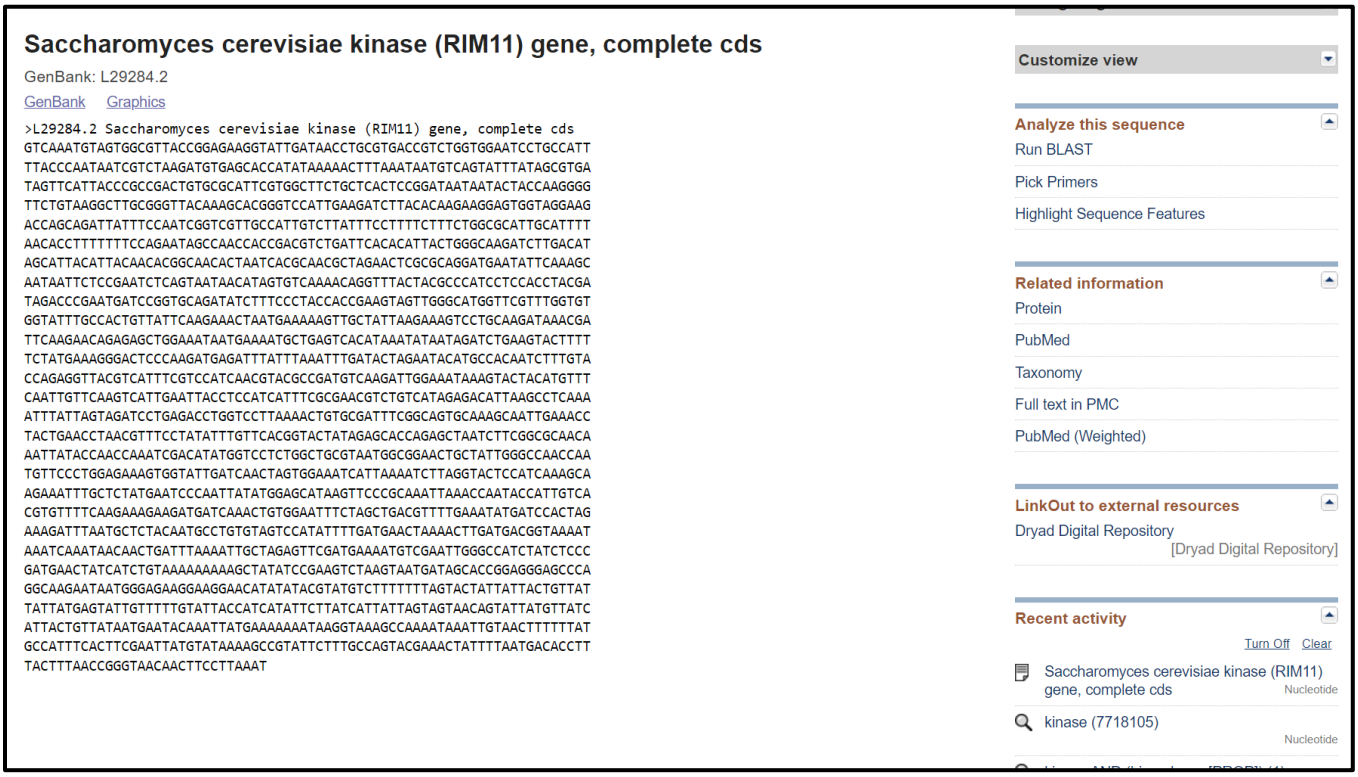


Fig2. Nucleotide FASTA sequence for Saccharomyces cerevisiae kinase

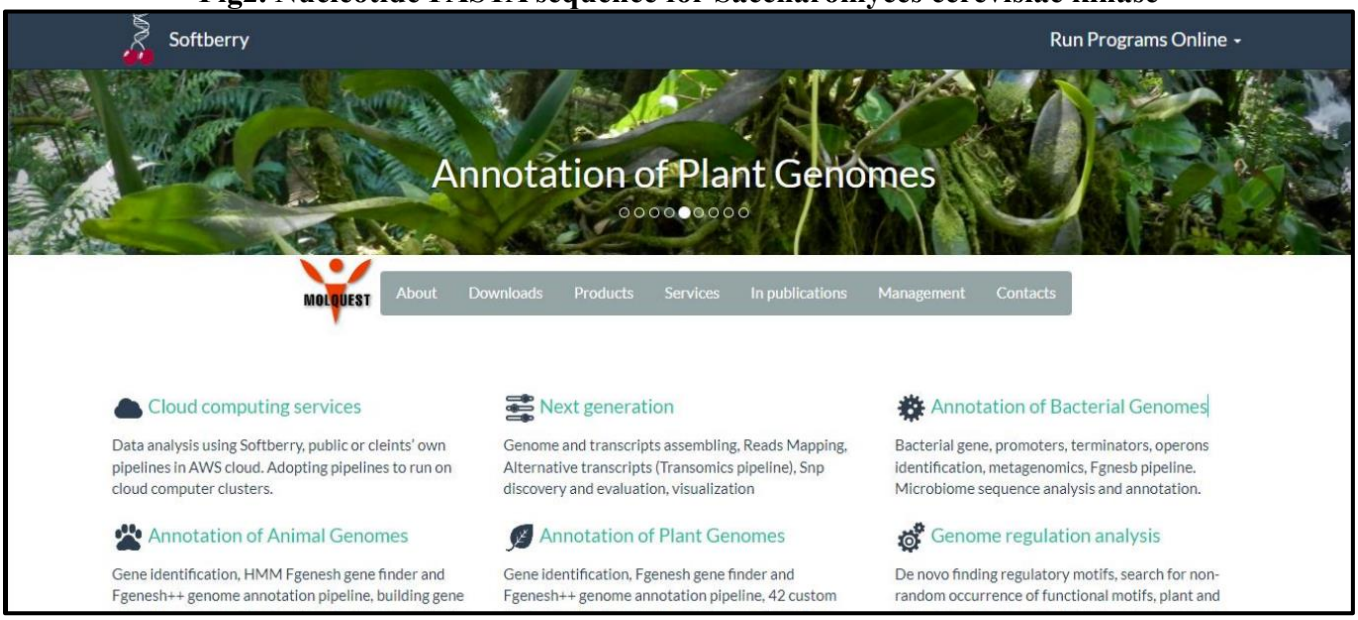


Fig3. Homepage for softberry

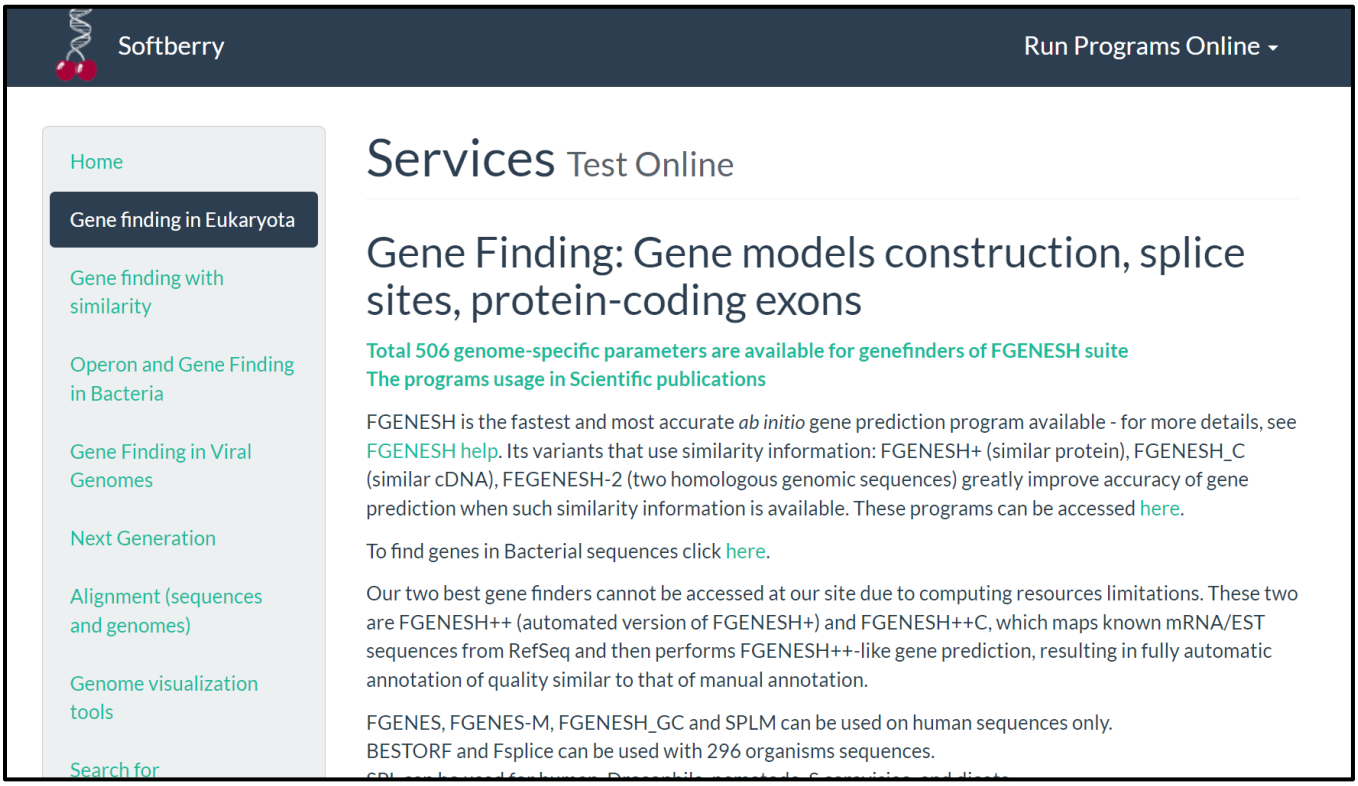


Fig4. Tools for gene finding in Eukaryota

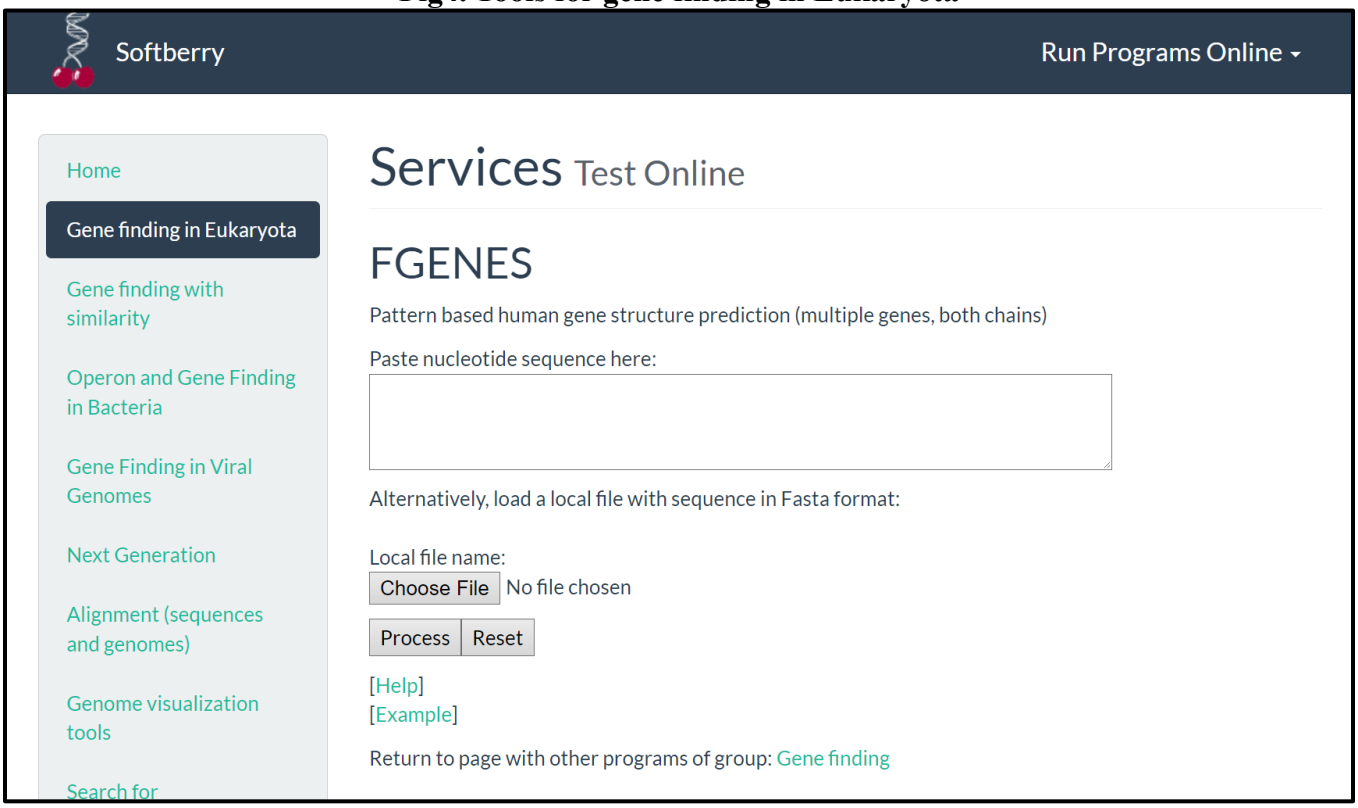


Fig5. Homepage for FGENES

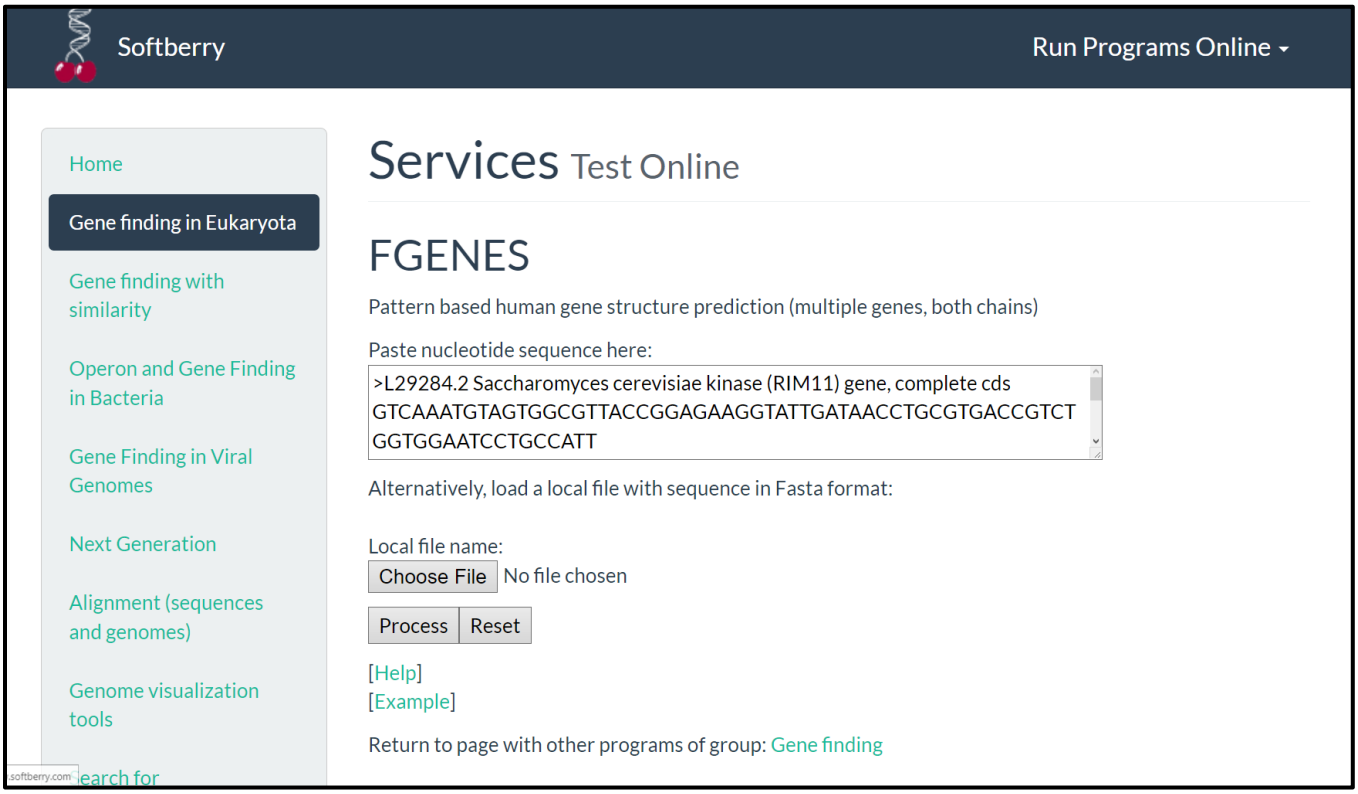
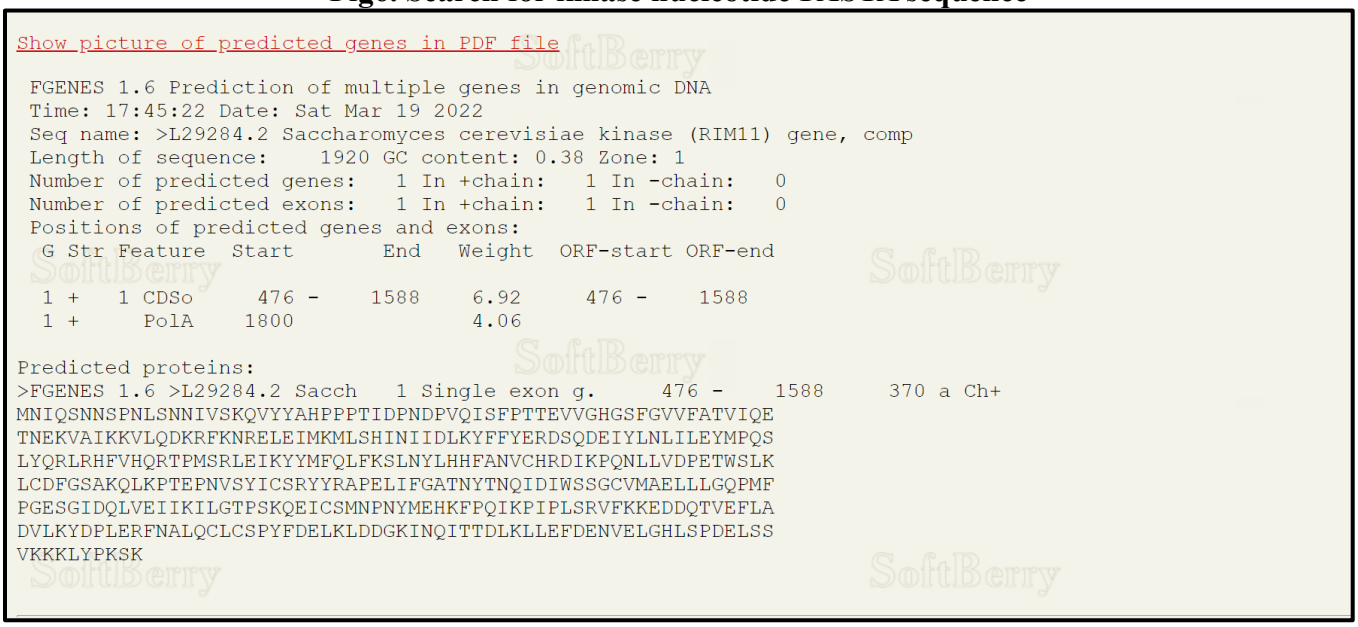


Fig6. Search for kinase nucleotide FASTA sequence



Show picture of predicted genes in PDF file

```
 FGENES 1.6 Prediction of multiple genes in genomic DNA
 Time: 17:45:22 Date: Sat Mar 19 2022
 Seq name: >L29284.2 Saccharomyces cerevisiae kinase (RIM11) gene, comp
 Length of sequence:    1920 GC content: 0.38 Zone: 1
 Number of predicted genes:   1 In +chain:   1 In -chain:   0
 Number of predicted exons:   1 In +chain:   1 In -chain:   0
 Positions of predicted genes and exons:
  G Str Feature  Start       End   Weight  ORF-start ORF-end

  1 +   1 CDSo     476 -    1588    6.92     476 -    1588
  1 +     PolA    1800             4.06

Predicted proteins:
>FGENES 1.6 >L29284.2 Sacch   1 Single exon g.     476 -    1588     370 a Ch+
MNIQSNNSPNLSNNIVSKQVYYAHPPPTIDPNDPVQISFPTTEVVGHGSFGVVFATVIQE
TNEKVAIKKVLQDKRFKNRELEIMKMLSHINIIDLKYFFYERDSQDEIYLNLILEYMPQS
LYQRLRHFVHQRTPMSRLEIKYYMFQLFKSLNYLHHFANVCHRDIKPQNLLVDPETWSLK
LCDFGSAKQLKPTEPNVSYICSRYYRAPELIFGATNYTNQIDIWSSGCVMAELLLGQPMF
PGESGIDQLVEIIKILGTPSKQEICSMNPNYMEHKFPQIKPIPLSRVFKKEDDQTVEFLA
DVLKYDPLERFNALQCLCSPYFDELKLDDGKINQITTDLKLLEFDENVELGHLSPDELSS
VKKKLYPKSK
```

Fig7. Result for predicted exon signals

RESULT:

Nucleotide FASTA sequence Saccharomyces cerevisiae kinase of length 1260 was submitted. With GC content of 0.57, 1 gene and 10 exons were predicted on the + chain. ORF start and end of each exon was predicted.

CONCLUSION:

FGENES tool is useful for prediction of exon signals. Exon prediction information can be used to predict genes and annotate genes and genomes. It can help in undering the gene function.

REFERENCES:

1. Xiong, J. (2008).Gene Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 97-111.
2. PROTEOLYTIC ENZYMES (PROTEASES): Overview, Uses, Side Effects, Precautions, Interactions, Dosing and Reviews. (n.d.). Www.webmd.com. Retrieved March 18, 2022, from https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20(proteases)%20are%20enzymes
3. Homo sapiens neutral protease alpha subunit gene, complete cds. (2016). NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/AH001431.2?report=genbank
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
5. FGENES - pattern-based gene structure prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind
6. Softberry - FGENES result. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfind/fgenes.pl

DATE: 19-03-22

WEBLEM 8e
ORF finder-NCBI
(URL: https://www.ncbi.nlm.nih.gov/orffinder/)

AIM:
To search for ORF region in Neisseria gonorrhoeae using ORF finder tool.

INTRODUCTION:
Two Gram-positive, catalase-negative bacterial strains were isolated from the cloaca of common kingfishers (Alcedo atthis). Repetitive sequence-based PCR fingerprinting using the (GTG)5 primer grouped these isolates into a single cluster separated from all known enterococcal species. The two strains revealed identical 16S rRNA gene sequences placing them within the genus Enterococcus with Enterococcus aquimarinus LMG 16607(T) as the closest relative (97.14 % similarity).
ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP. This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation.

METHODOLOGY:
1. Open homepage for ORF finder in NCBI. (URL: https://www.ncbi.nlm.nih.gov/orffinder/)
2. Retrieve nucleotide FASTA sequence for Neisseria gonorrhoeae from GenBank.
3. Submit the FASTA sequence.
4. Observe and interpret the results.

OBSERVATION:

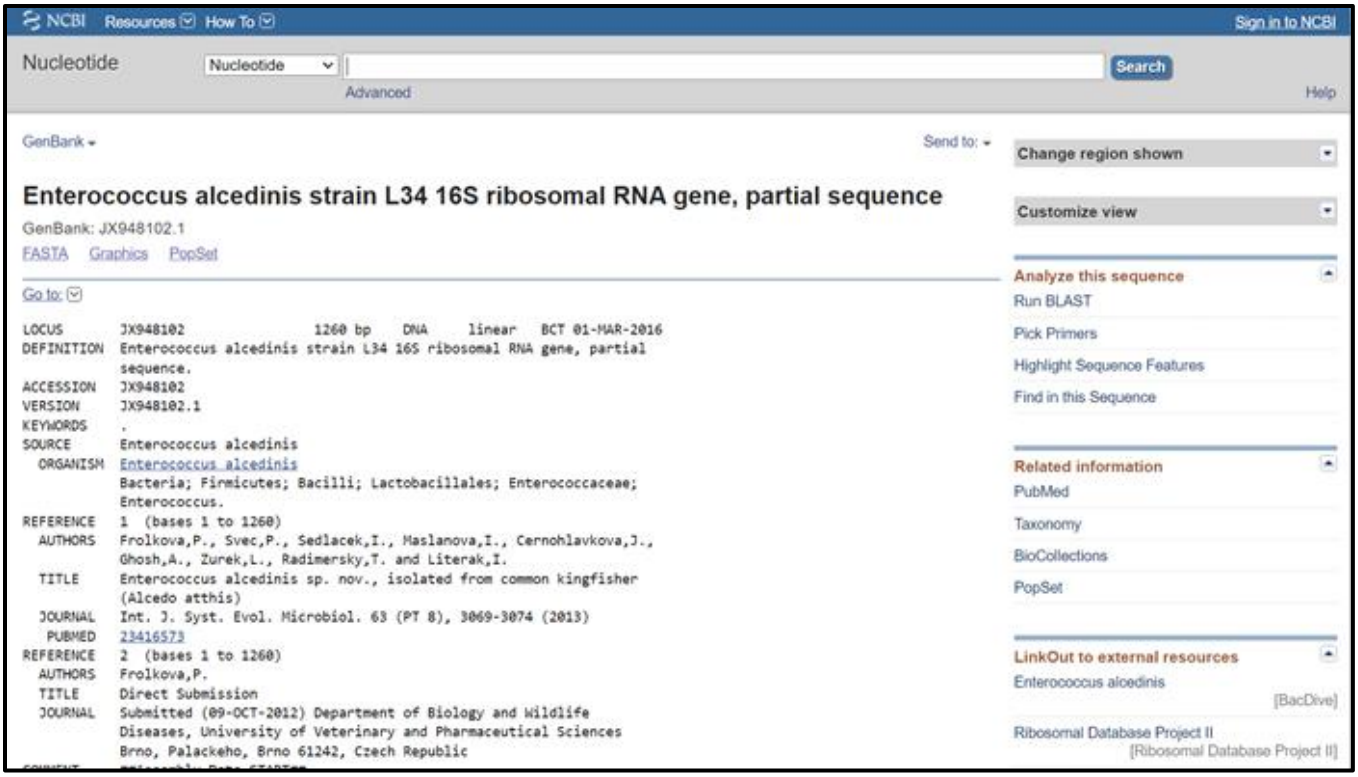


Fig1. Result for Enterococcus alcedinis

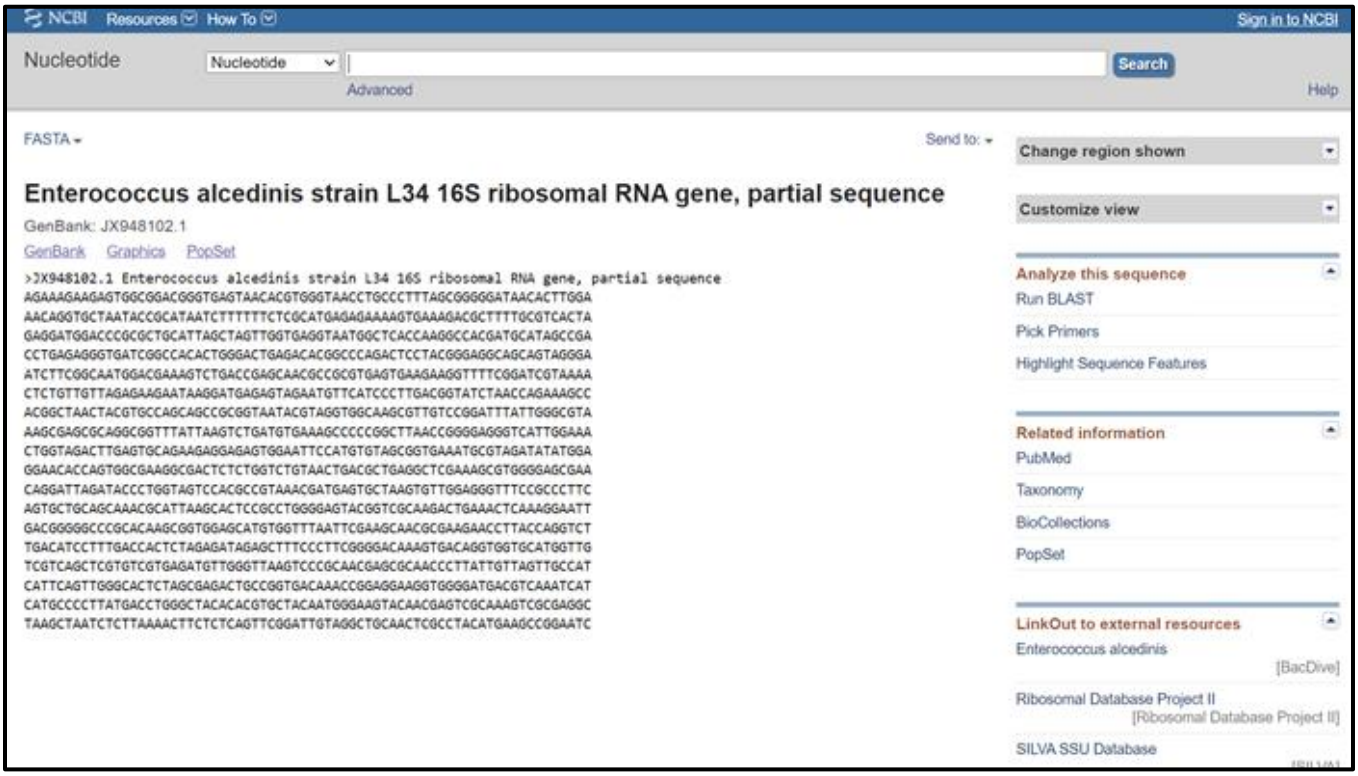


Fig2. Nucleotide FASTA sequence for Enterococcus alcedinis

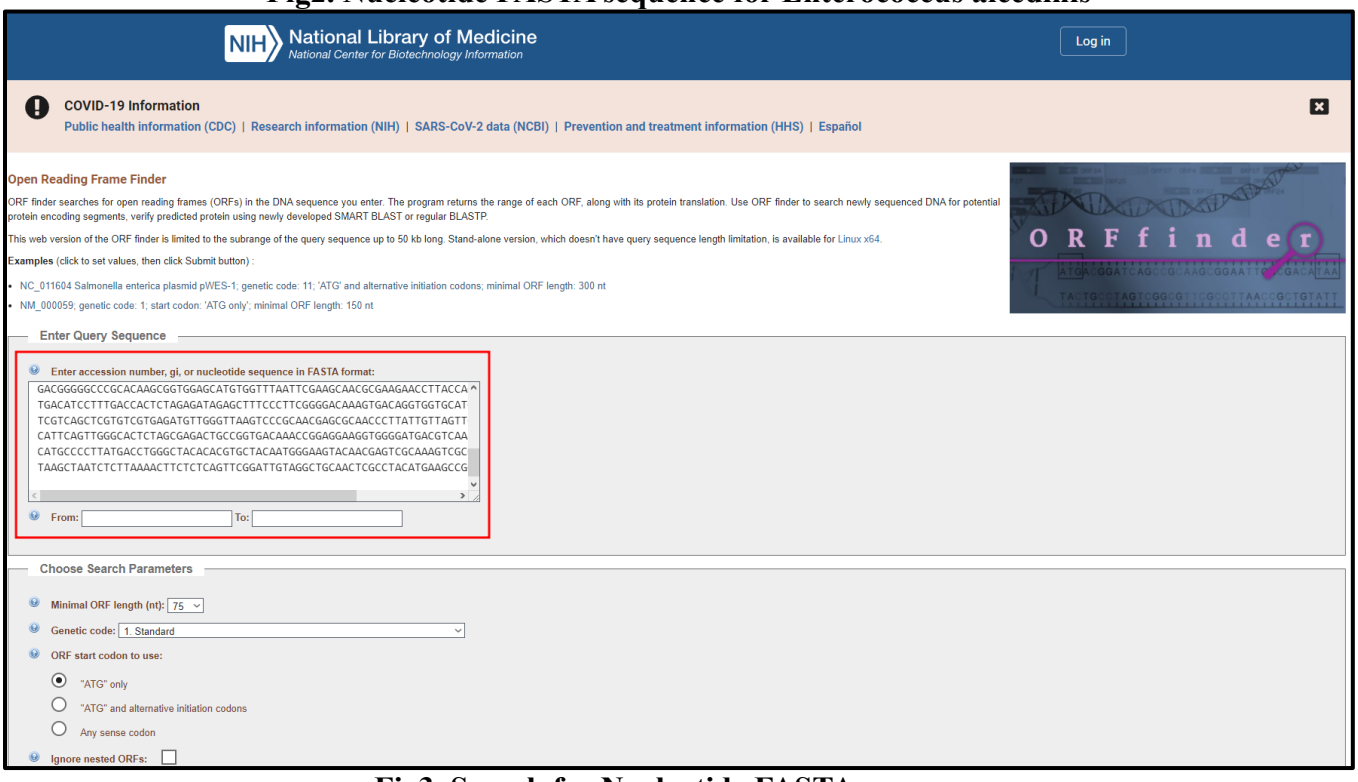


Fig3. Search for Nucleotide FASTA sequence

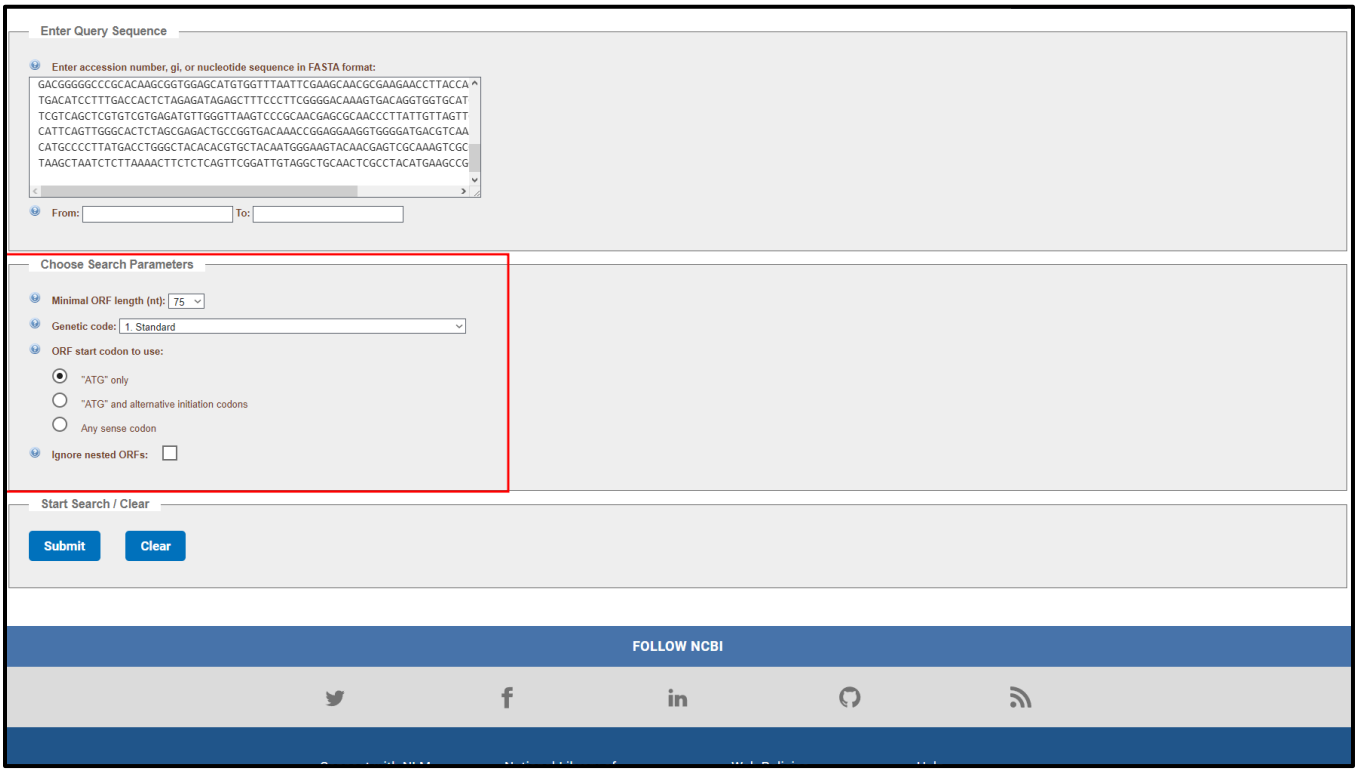


Fig4. Search parameters

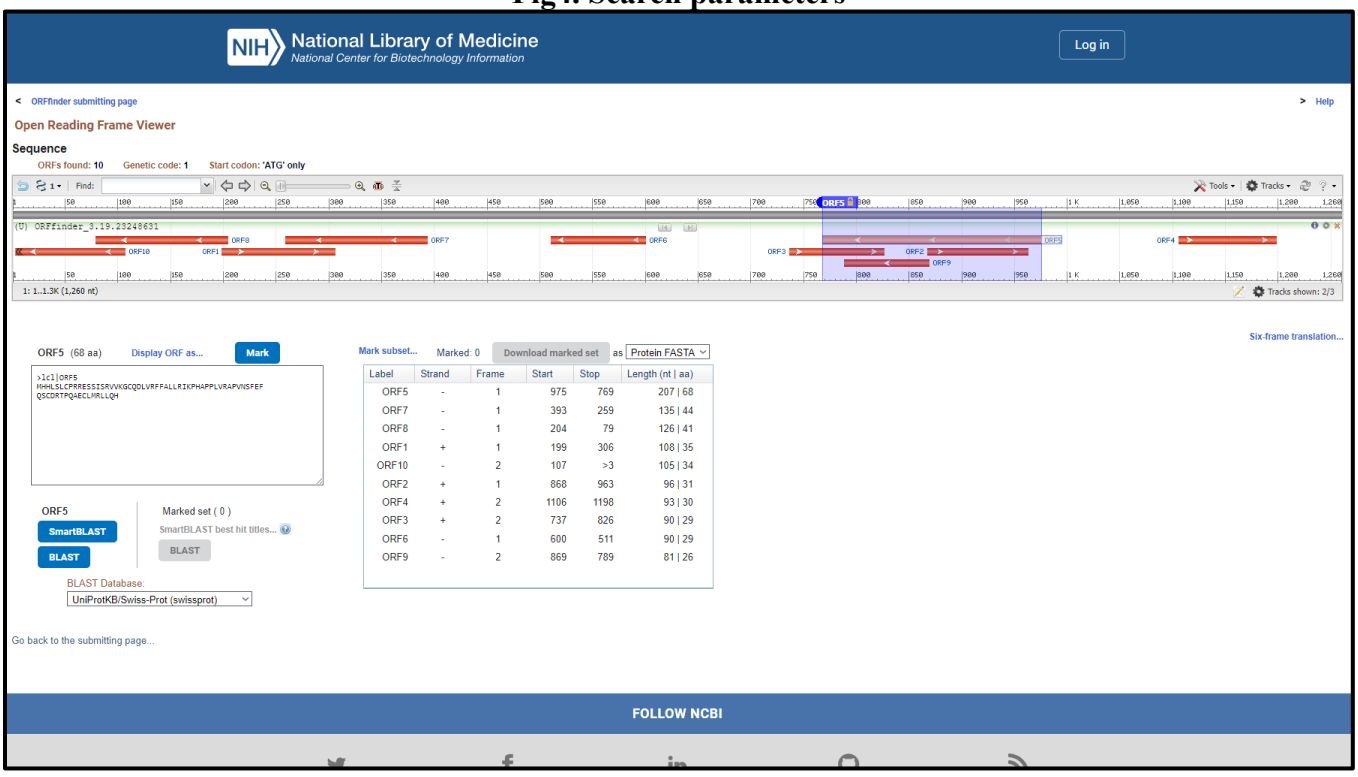


Fig5. Result for recognised ORFs and tracks information

Results:

After submitting nucleotide FASTA sequence for complete genome of Enterococcus alcedinis on NCBI's ORF finder, 360 ORFs were predicted.

CONCLUSION:

ORF finder can used to predict open reading frames in the genome. This information of long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA- coding regions in a DNA sequence. Small Open Reading Frames (small ORFs/sORFs/smORFs) are important sources of putative peptides previously dismissed as being non-functional or junk DNA.

REFERENCES:

1. Neisseria gonorrhoeae - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com.Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
2. Neisseria gonorrhoeae strain NJ189125 chromosome, complete genome. (2022). NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP041586.1
3. Home - ORFfinder - NCBI. (2019). Nih.gov. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/orffinder/

DATE: 30-03-22

WEBLEM 9

Introduction to Genomics & its various browser (UCSC, ENSEMBL, GDV)

Genomics is the **study of genomes**. Genomic studies are characterized by **simultaneous analysis of a large number** of genes using **automated data gathering tools**. The topics of **genomics range** from genome **mapping sequencing**, and **functional genomic analysis** to comparative **genomic analysis**. The **advent of genomics** and the ensuing explosion of **sequence information** are the main driving force behind the **rapid development** of bioinformatics today.

Genomic study can be **tentatively divided** into **structural genomics** and **functional genomics**. Structural genomics refers to the **initial phase of genome analysis**, which includes construction of **genetic and physical maps** of a genome, identification of genes, **annotation of gene features**, and **comparison of genome structures**. Functional genomics refers to the analysis of **global gene expression** and gene functions in a genome.

Genome browsers are resources that **integrate data at the genomic level**, thereby allowing visualization of related **genomic information** in one space. These data can **include genes**, **noncoding elements** that regulate **gene expression**, **genetic variation** and the results of **comparative genomics analyses**, among other forms of annotation. Commonly used genome **browsers include Ensembl**, the **UCSC Genome Browser** and **GDV**.

1. UCSC Genome Browser

The **University of California Santa Cruz** (UCSC) **Genome Browser** (genome.ucsc.edu) is a popular **Web-based tool** for quickly displaying a requested **portion of a genome** at any scale, accompanied by a series of **aligned annotation "tracks".** The annotations generated by the **UCSC Genome Bioinformatics Group** and external collaborators display **gene predictions**, **mRNA** and **expressed sequence tag alignments**, **simple nucleotide polymorphisms**, **expression** and **regulatory data**, **phenotype** and **variation data**, and **pairwise** and **multiple-species comparative** genomics data. All **information relevant** to a region is presented **in one window**, facilitating **biological analysis** and **interpretation**. The database tables **underlying the Genome Browser tracks** can be **viewed**, **downloaded**, and **manipulated** using another **Web-based application**, the UCSC Table Browser. Users can upload data as **custom annotation tracks** in both browsers for research or educational use.

The **vast size** of vertebrate genome data sets **presents challenges** in **efficient data storage** and **retrieval**. In addition, the **burgeoning number** of versions of a **particular genome demands** a process that can **rapidly integrate new data** and annotations into the database while **implementing creative solutions** for maintaining and **enhancing views of the data**. Through **software algorithmic refinements** and optimizations to **both the database and hardware**, the UCSC Genome Browser viewer maintains the **same interactive** response time on the **large Homo sapiens** and **Mus musculus genomes** that its predecessor had on the **much smaller Caenorhabditis** elegans genome.

Sequence and annotation data for each genome assembly are **stored in a MySQL relational database**, which is **quite efficient at retrieving data** from indexed files. The **database is loaded** in large batches and is used **primarily as a read-only database**. To improve performance, each of the Genome Browser web servers has a **copy of the database on its local disk**. UCSC **generates several annotations** based on **mRNA alignments**. The **mRNA and EST sequences** are extracted from GenBank, and are **aligned** against the genome **using the BLAST-like Alignment Tool** (BLAT), a fast sequence alignment tool developed by Jim Kent. The **data is filtered based on percentage identity** and near best in genome to select only those alignments that **best match the sequence**. The **spliced EST** annotation is computed from the **filtered data by analyzing** the EST alignments for **evidence of splicing**.

2. ENSEMBL gene browser

The Ensembl project was initially launched in **1999** with the aim of **developing methodologies** for **automatic annotation** of (human) **genomic sequence** with genes and their **constituent transcripts**. Since that time, the project has broadened substantially in scope; the **Ensembl Genome Browser**, which came online in **2000**, now includes **reference genomic sequence** and **annotation** for nearly **100 chordate organisms**. Ensembl is **rapidly incorporating** new data, including **whole clades** of new species' genomes and reference sequence for **multiple strains of existing species**, such as mouse. In addition, existing annotation is **regularly augmented** by the **inclusion of new data sets**. Ensembl's sister site, **Ensembl Genomes**, provides access to **nonvertebrate genomes** through dedicated portals for **Bacteria, Fungi, Plants, Metazoa, and Protists.**

Ensembl data, **annotations**, and analyses are updated **every 2−3 months**, alongside **software updates** to both the **public-facing website** and the **underlying databases**. A dedicated site is also maintained for the **GRCh37 reference human genome assembly**, which is annotated with **new data on a limited basis**; partial data from **ongoing genome annotation** can be accessed via the preview Pre! site.

Data from Ensembl can be **accessed at multiple scales**. Data can be accessed through the browser web pages and **via BioMart**, a web-based tool that allows **customized retrieval** of data from the **Ensembl databases**. However, data can also be **accessed programmatically** via our **Perl and REST APIs.** Files containing **genome-wide data** are available for all species **represented in Ensembl** via an FTP site; data from all releases of **Ensembl** can be **retrieved** from the **FTP site**, or from our databases via the **Perl APIs, in perpetuity.**

3. Genome Data Viewer

GDV is composed of an **embedded instance of SV** that displays **sequence and track data**, along with **additional page elements** that allow a user to **search within an entire genome** assembly and efficiently **narrow in on their chromosome**, **sequence**, **region**, or **gene of interest**. GDV replaced the NCBI Map Viewer, **NCBI's previous** tool for whole-genome display. Researchers using GDV **can go directly** to the **NCBI BLAST** service from the browser and **load BLAST results** as alignment tracks that can be viewed **side by side** with **gene annotation** and other data. **Variation Viewer**, a related browser **associated with NCBI's** variation resources, is **functionally similar to GDV** and also **incorporates an instance of SV** but is **configured with features specifically intended for analyzing human variation data**. GDV and Variation Viewer can both **display the same types of NCBI** variation track data.

The GDV **can be accessed** from its own **home page** and can also be **found via links** from other **NCBI resources, including gene, assembly, GEO, and dbGaP record pages**. GDV provides users a graphical gateway to data at the NCBI, **especially RefSeq and refSNP** annotation. Below, we highlight **some of the functions** of GDV and other instances of the **NCBI SV** and provide context for **GDV's** features with respect to the **broader collection** of publicly available **genome browsers**, including the **UCSC** and **Ensembl genome browsers**, **JBrowse**, and **IGV**. **GDV** was designed specifically to **support visualization** and **analysis** of the wide range of genomes and assemblies annotated at the NCBI. **RefSeq gene annotation** data tracks are shown by default in the **graphical view** for these assemblies. **NCBI refSNP data tracks** are also shown by default for **human assemblies**. **Gene and SNP tracks** are **automatically updated** in **GDV** and **SV** embedded instances upon **new releases** of the NCBI databases, so that **users of the NCBI graphical viewers** always have immediate access to the **latest versions of RefSeq and SNP annotation.**

GDV offers users the ability to **customize the displays of individual tracks**. Users can **hide** or **configure** tracks from the **track configuration panel** or by using the icons at the right end of each track. Different **public genome browsers provide conceptually similar**, but somewhat **distinct options**, for **visualizing gene, graphical, and alignment data**. In this section, we highlight **track data**

visualizations in the GDV browser and other instances of the **SV graphical** view component that support various **analysis scenarios**.

REFERENCES:

1. Xiong, J. (2008). Genome Mapping, Assembly, and Comparison. Essential bioinformatics. Cambridge: Cambridge University Press. 243.
2. Baxevanis, Andreas D.; Petsko, Gregory A.; Stein, Lincoln D.; Stormo, Gary D. (2002). Current Protocols in Bioinformatics || The UCSC Genome Browser. , (), –. doi:10.1002/0471250953.bi0104s28
3. Karolchik, D. (2003). The UCSC Genome Browser Database. , 31(1), 51–54. doi:10.1093/nar/gkg129
4. Birney, E. (2004). An Overview of Ensembl. Genome Research, 14(5), 925–928. doi:10.1101/gr.1860604
5. Kollmar, Martin (2018). [Methods in Molecular Biology] Eukaryotic Genomic Databases Volume 1757 || The Ensembl Genome Browser: Strategies for Accessing Eukaryotic Genome Data. , 10.1007/978-1-4939-7737-6(Chapter 6), 115–139. doi:10.1007/978-1-4939-7737-6_6
6. Rangwala, S. H., Kuznetsov, A., Ananiev, V., Asztalos, A., Borodin, E., Evgeniev, V., Joukov, V., Lotov, V., Pannu, R., Rudnev, D., Shkeda, A., Weitz, E. M., & Schneider, V. A. (2020). Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Research, gr.266932.120. https://doi.org/10.1101/gr.266932.120

DATE: 30-03-22

WEBLEM 9a
UCSC Genome Browser
(URL: https://genome.ucsc.edu/)

AIM:

To explore UCSC genome browser in order to understand the gene, its related studies & protein level information.

INTRODUCTION:

The **University of California Santa Cruz** (UCSC) **Genome Browser** (genome.ucsc.edu) is a popular **Web-based tool** for quickly displaying a requested **portion of a genome** at any scale, accompanied by a series of **aligned annotation "tracks".** The annotations generated by the **UCSC Genome Bioinformatics Group** and external collaborators display **gene predictions**, **mRNA** and **expressed sequence tag alignments**, **simple nucleotide polymorphisms**, **expression** and **regulatory data**, **phenotype** and **variation data**, and **pairwise** and **multiple-species comparative** genomics data. All **information relevant** to a region is presented **in one window**, facilitating **biological analysis** and **interpretation**. The database tables **underlying the Genome Browser tracks** can be **viewed**, **downloaded**, and **manipulated** using another **Web-based application**, the UCSC Table Browser. Users can upload data as **custom annotation tracks** in both browsers for research or educational use.

METHODOLOGY:

1. Open homepage for UCSC browser (URL: https://genome.ucsc.edu/)
2. Select genome browser.
3. Select human assembly (GRCh3/hg38)
4. Navigate results through gene name, SNP id, Ref_Seq, OMIM id, coordinates and cytological band.
5. Use tools for zooming tracks in and out, configuration by right click, drag and select and various option available at bottom of the page.
6. Observer and interpret the results.

OBSERVATION:

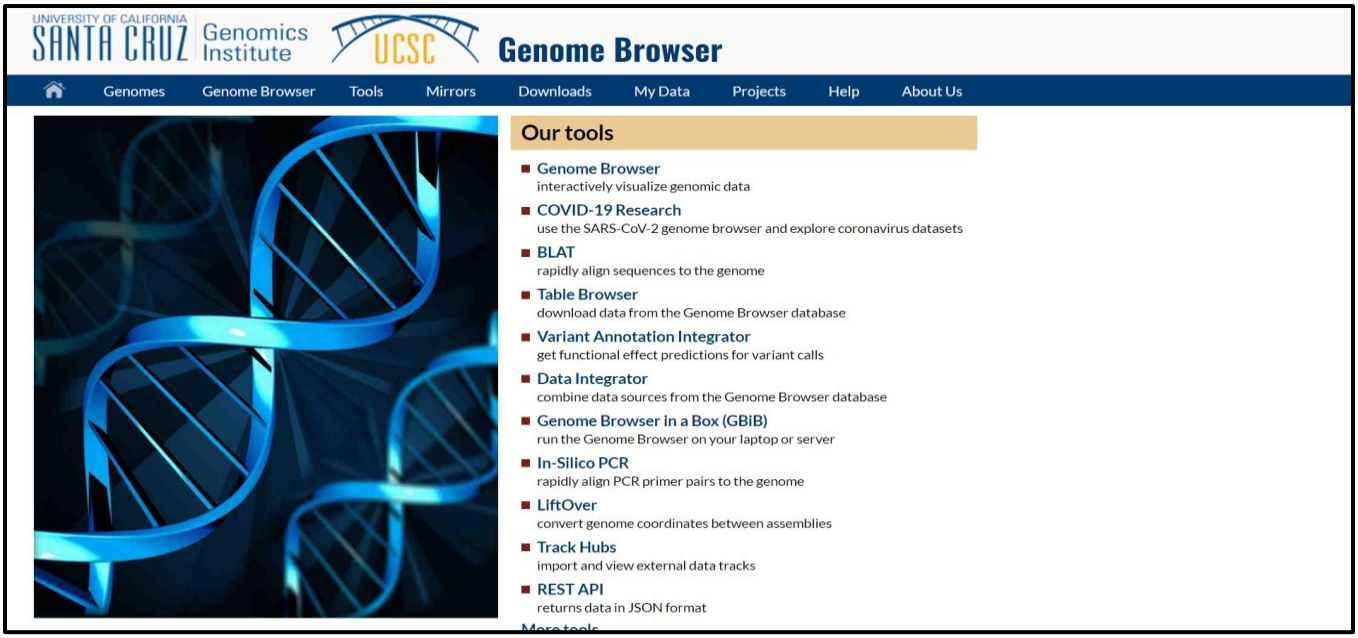


Fig1. Homepage of UCSC browser

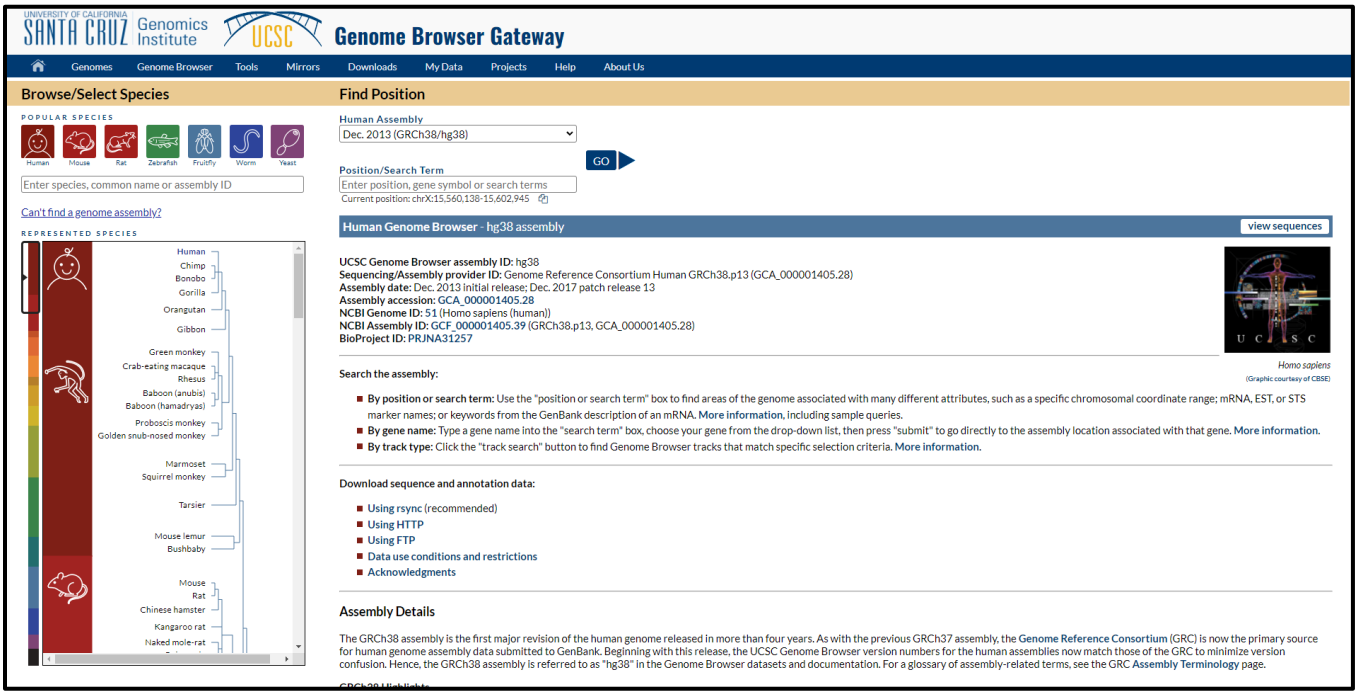


Fig2. Genome browser gateway

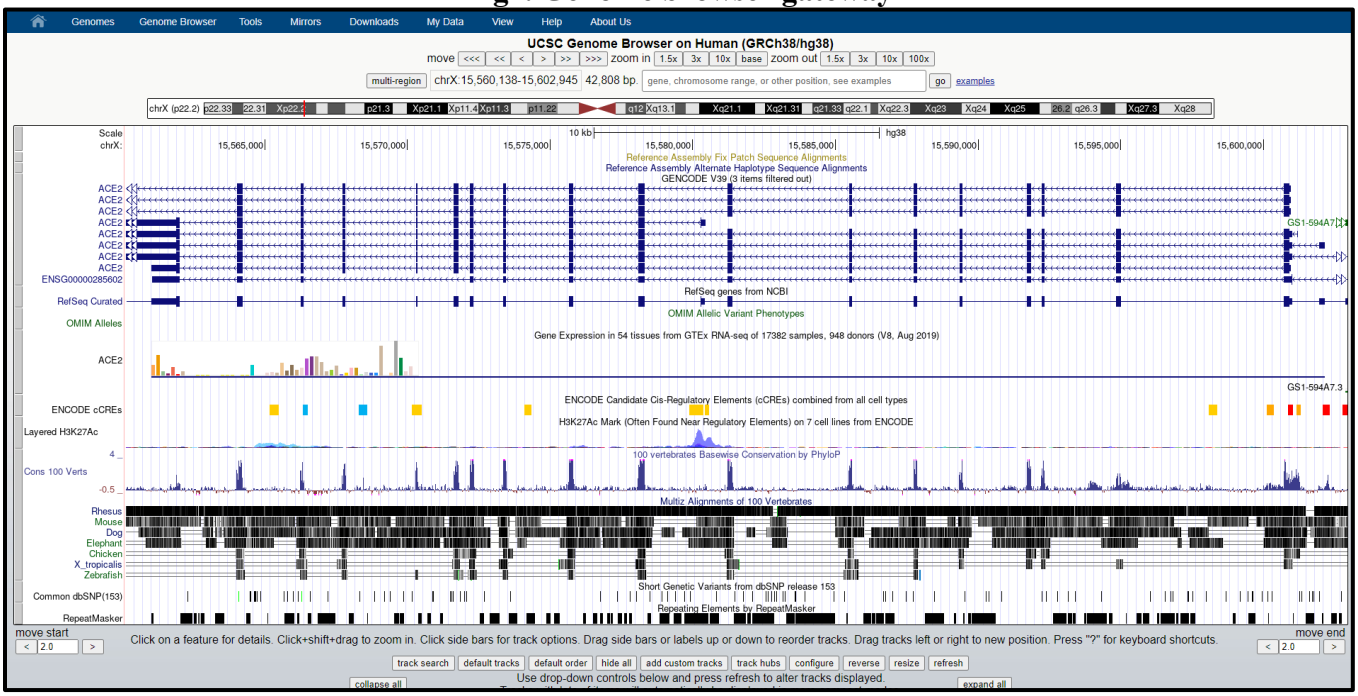


Fig3. UCSC genome browser on human (GRCh38/hg38)

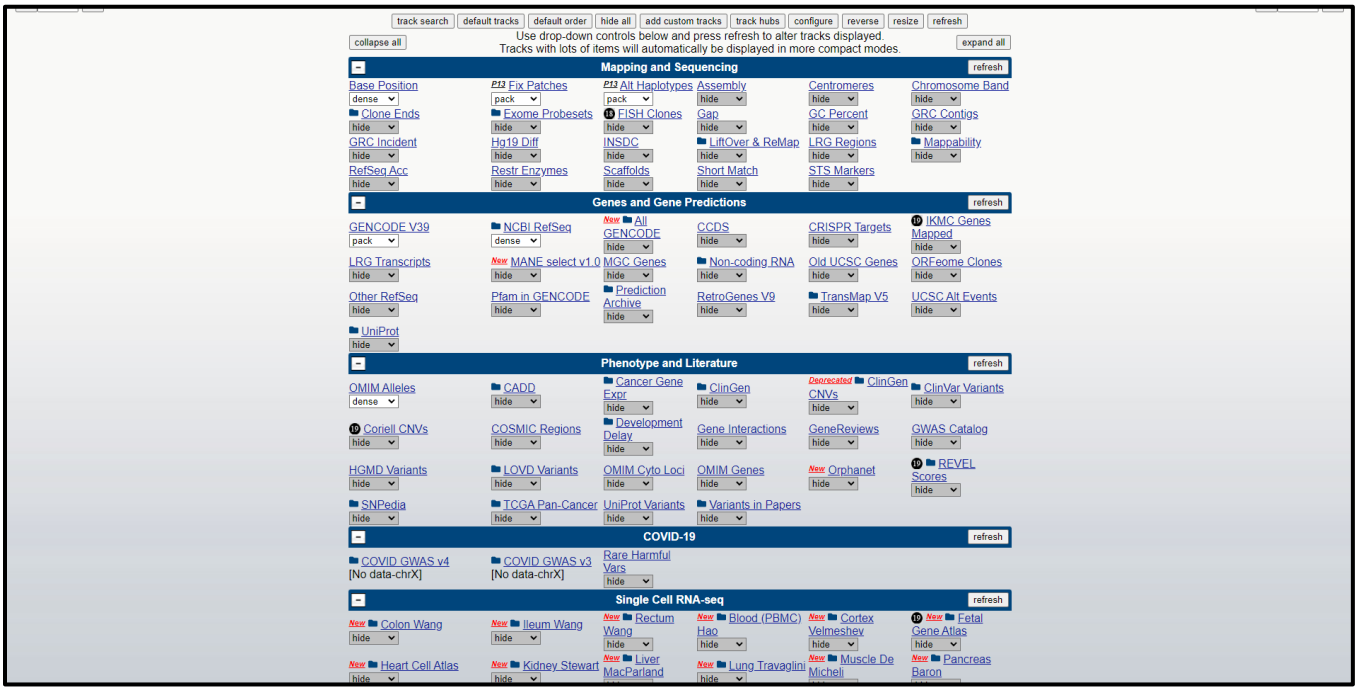


Fig4. Options for customization

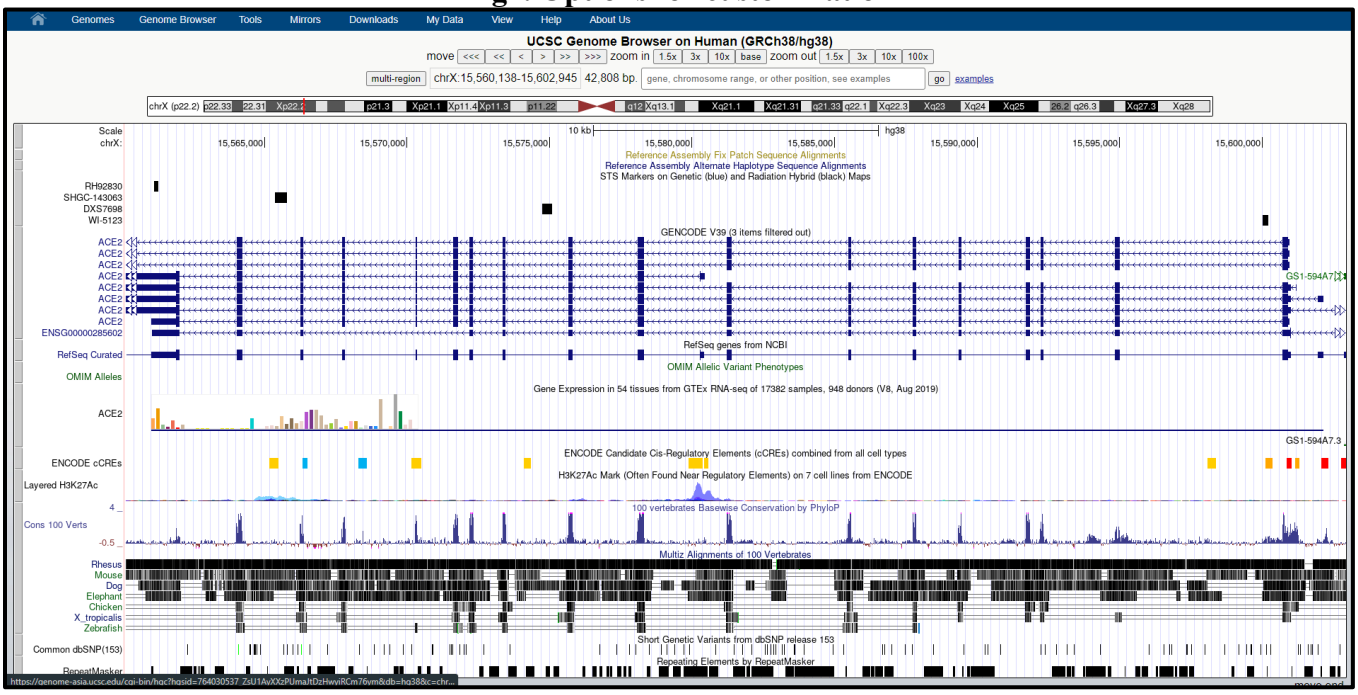


Fig5. Results after customization

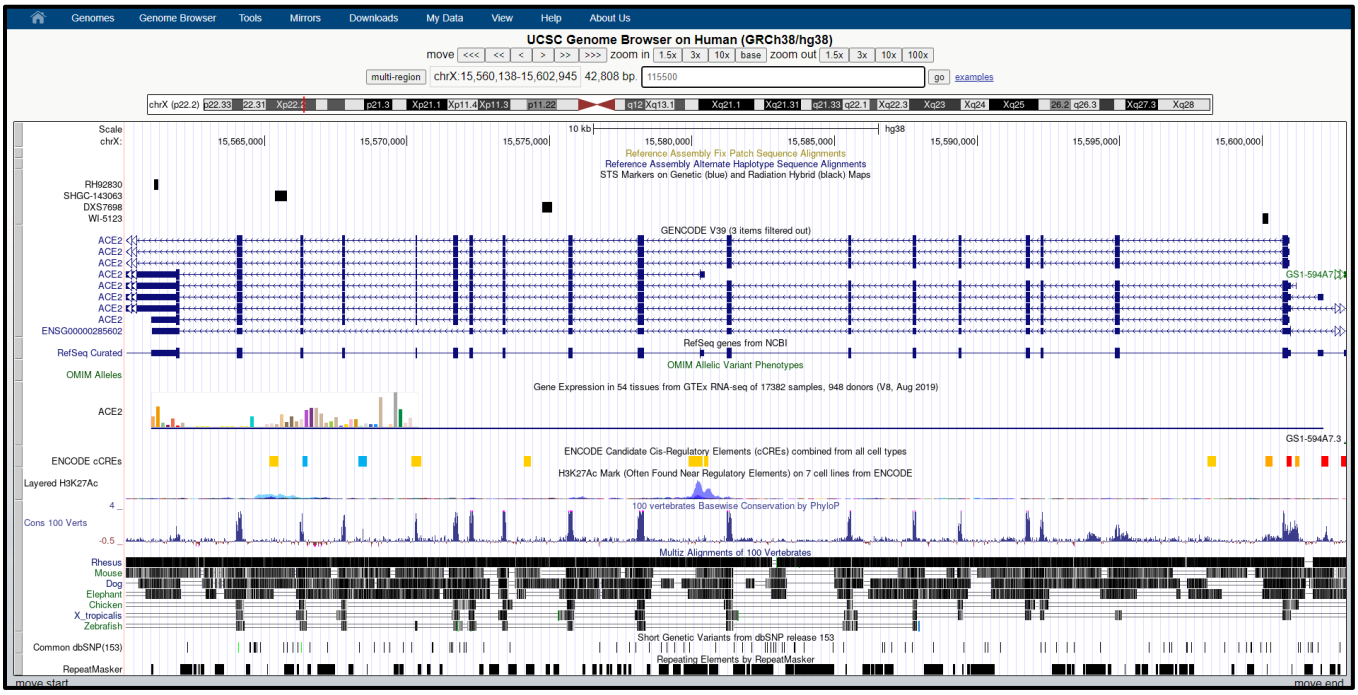


Fig6. Navigation by OMIM id: 115500

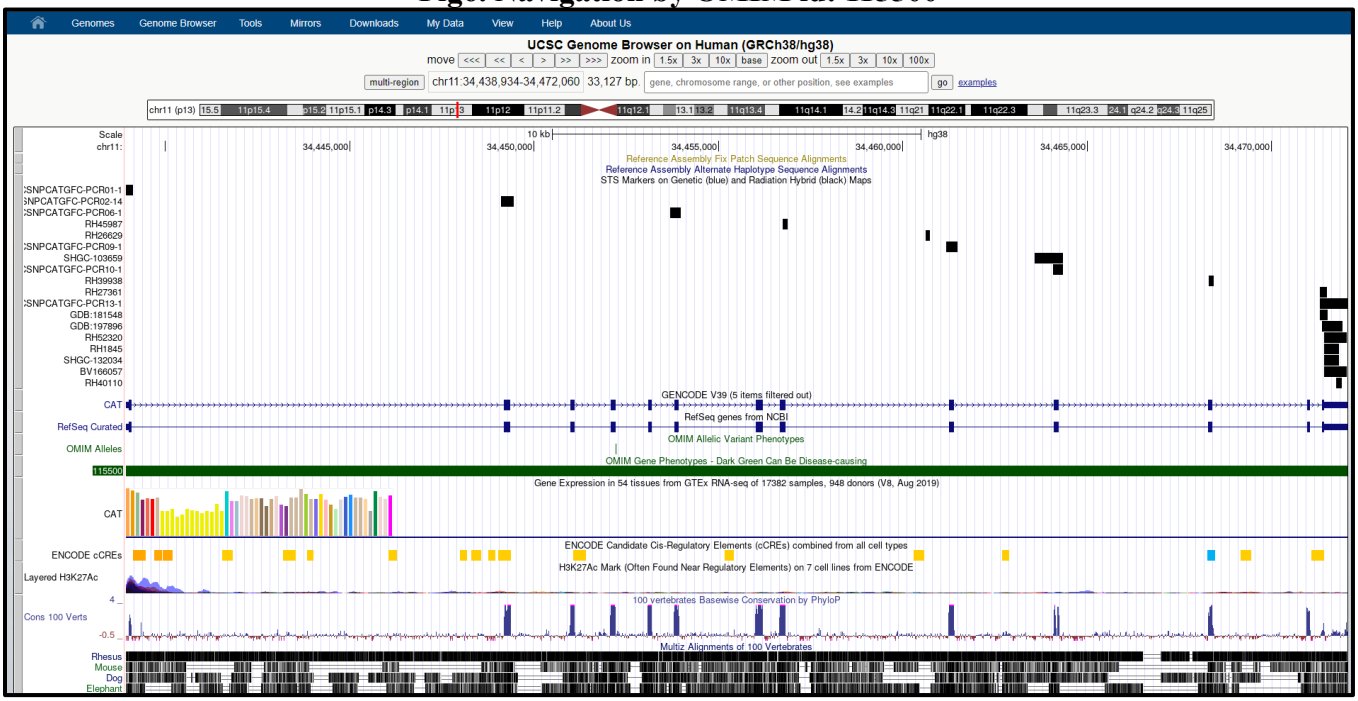
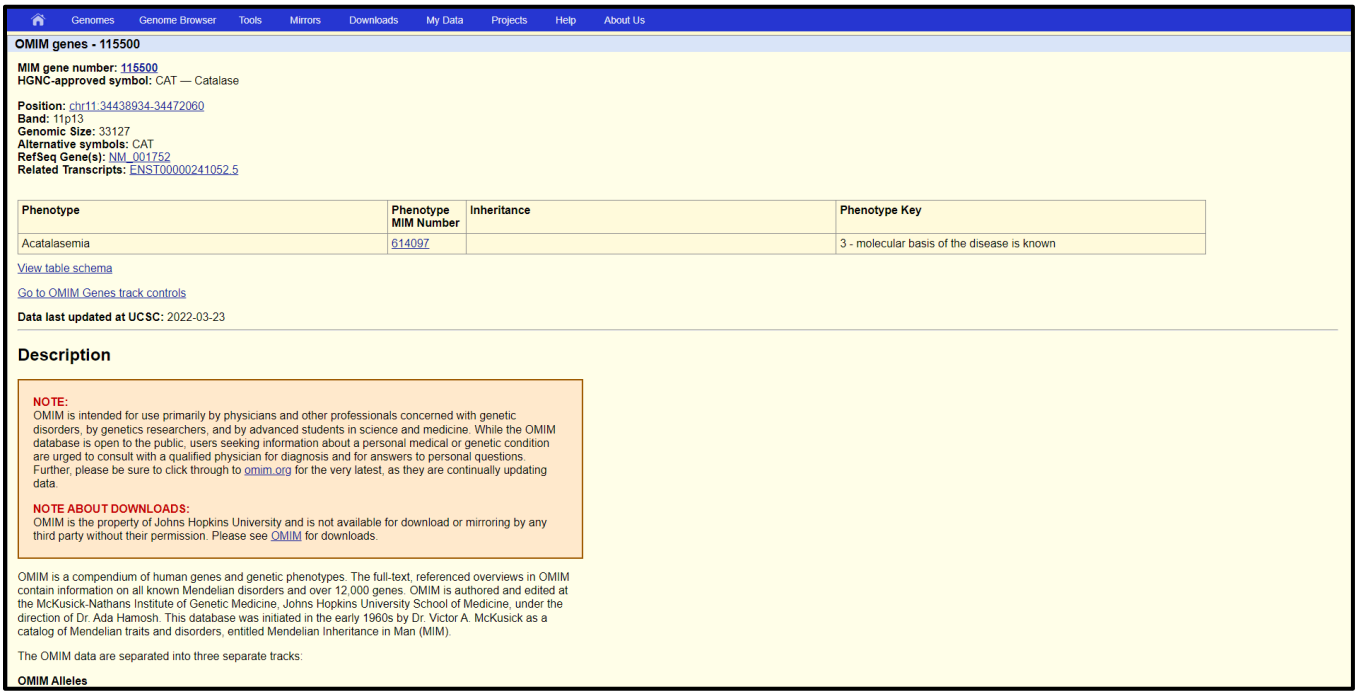


Fig7. Results for OMIM id: 115500



Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

OMIM genes - 115500

MIM gene number: 115500
HGNC-approved symbol: CAT — Catalase

Position: chr11:34438934-34472060
Band: 11p13
Genomic Size: 33127
Alternative symbols: CAT
RefSeq Gene(s): NM_001752
Related Transcripts: ENST00000241052.5

Phenotype	Phenotype MIM Number	Inheritance	Phenotype Key
Acatalasemia	614097		3 - molecular basis of the disease is known

View table schema

Go to OMIM Genes track controls

Data last updated at UCSC: 2022-03-23

Description

NOTE:
OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions. Further, please be sure to click through to omim.org for the very latest, as they are continually updating data.

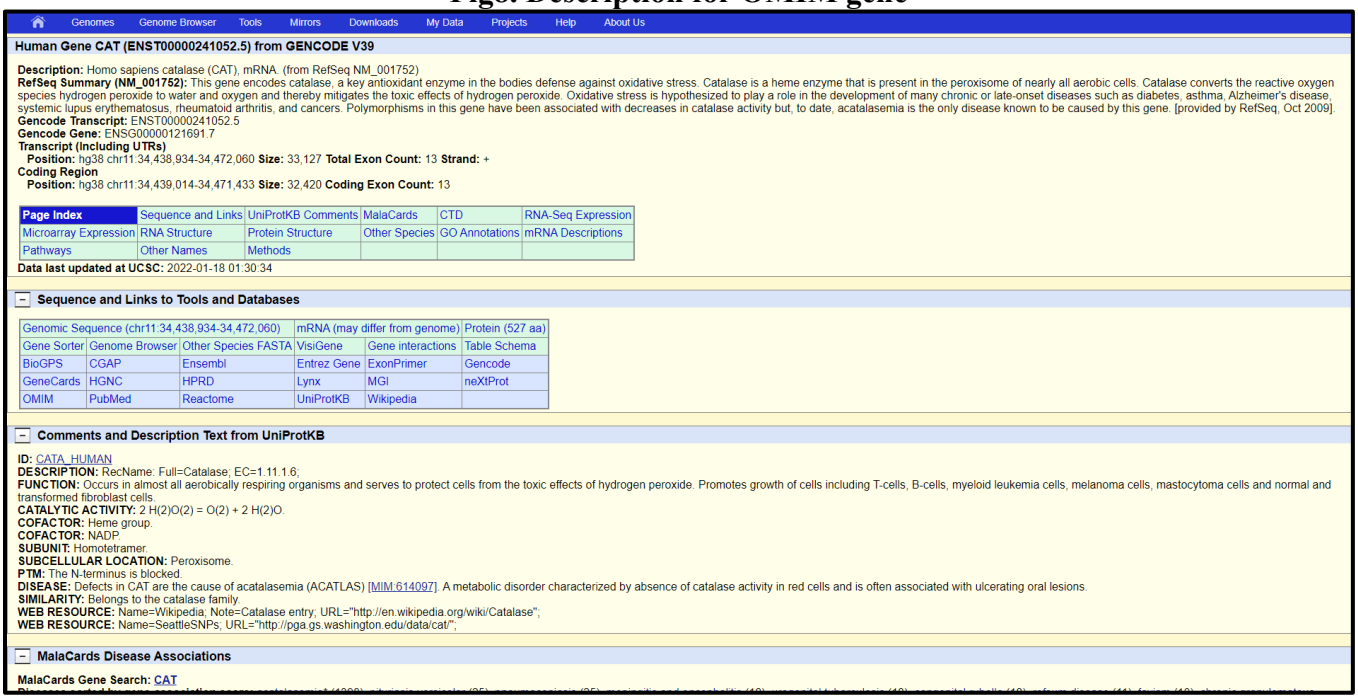
NOTE ABOUT DOWNLOADS:
OMIM is the property of Johns Hopkins University and is not available for download or mirroring by any third party without their permission. Please see OMIM for downloads.

OMIM is a compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of Mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM).

The OMIM data are separated into three separate tracks:

OMIM Alleles

Fig8. Description for OMIM gene



Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Human Gene CAT (ENST00000241052.5) from GENCODE V39

Description: Homo sapiens catalase (CAT), mRNA. (from RefSeq NM_001752)
RefSeq Summary (NM_001752): This gene encodes catalase, a key antioxidant enzyme in the bodies defense against oxidative stress. Catalase is a heme enzyme that is present in the peroxisome of nearly all aerobic cells. Catalase converts the reactive oxygen species hydrogen peroxide to water and oxygen and thereby mitigates the toxic effects of hydrogen peroxide. Oxidative stress is hypothesized to play a role in the development of many chronic or late-onset diseases such as diabetes, asthma, Alzheimer's disease, systemic lupus erythematosus, rheumatoid arthritis, and cancers. Polymorphisms in this gene have been associated with decreases in catalase activity but, to date, acatalasemia is the only disease known to be caused by this gene. [provided by RefSeq, Oct 2009].
Gencode Transcript: ENST00000241052.5
Gencode Gene: ENSG00000121691.7
Transcript (Including UTRs)
Position: hg38 chr11:34,438,934-34,472,060 **Size:** 33,127 **Total Exon Count:** 13 **Strand:** +
Coding Region
Position: hg38 chr11:34,439,014-34,471,433 **Size:** 32,420 **Coding Exon Count:** 13

Page Index	Sequence and Links	UniProtKB Comments	MalaCards	CTD	RNA-Seq Expression
Microarray Expression	RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions
Pathways	Other Names	Methods			

Data last updated at UCSC: 2022-01-18 01:30:34

Sequence and Links to Tools and Databases

Genomic Sequence (chr11:34,438,934-34,472,060)			mRNA (may differ from genome)		Protein (527 aa)
Gene Sorter	Genome Browser	Other Species FASTA	VisiGene	Gene interactions	Table Schema
BioGPS	CGAP	Ensembl	Entrez Gene	ExonPrimer	Gencode
GeneCards	HGNC	HPRD	Lynx	MGI	neXtProt
OMIM	PubMed	Reactome	UniProtKB	Wikipedia	

Comments and Description Text from UniProtKB

ID: CATA_HUMAN
DESCRIPTION: RecName: Full=Catalase; EC=1.11.1.6;
FUNCTION: Occurs in almost all aerobically respiring organisms and serves to protect cells from the toxic effects of hydrogen peroxide. Promotes growth of cells including T-cells, B-cells, myeloid leukemia cells, melanoma cells, mastocytoma cells and normal and transformed fibroblast cells.
CATALYTIC ACTIVITY: 2 H(2)O(2) = O(2) + 2 H(2)O.
COFACTOR: Heme group.
COFACTOR: NADP.
SUBUNIT: Homotetramer.
SUBCELLULAR LOCATION: Peroxisome.
PTM: The N-terminus is blocked.
DISEASE: Defects in CAT are the cause of acatalasemia (ACATLAS) [MIM:614097]. A metabolic disorder characterized by absence of catalase activity in red cells and is often associated with ulcerating oral lesions.
SIMILARITY: Belongs to the catalase family.
WEB RESOURCE: Name=Wikipedia; Note=Catalase entry; URL="http://en.wikipedia.org/wiki/Catalase";
WEB RESOURCE: Name=SeattleSNPs; URL="http://pga.gs.washington.edu/data/cat/";

MalaCards Disease Associations

MalaCards Gene Search: CAT

Fig9. Related transcripts information

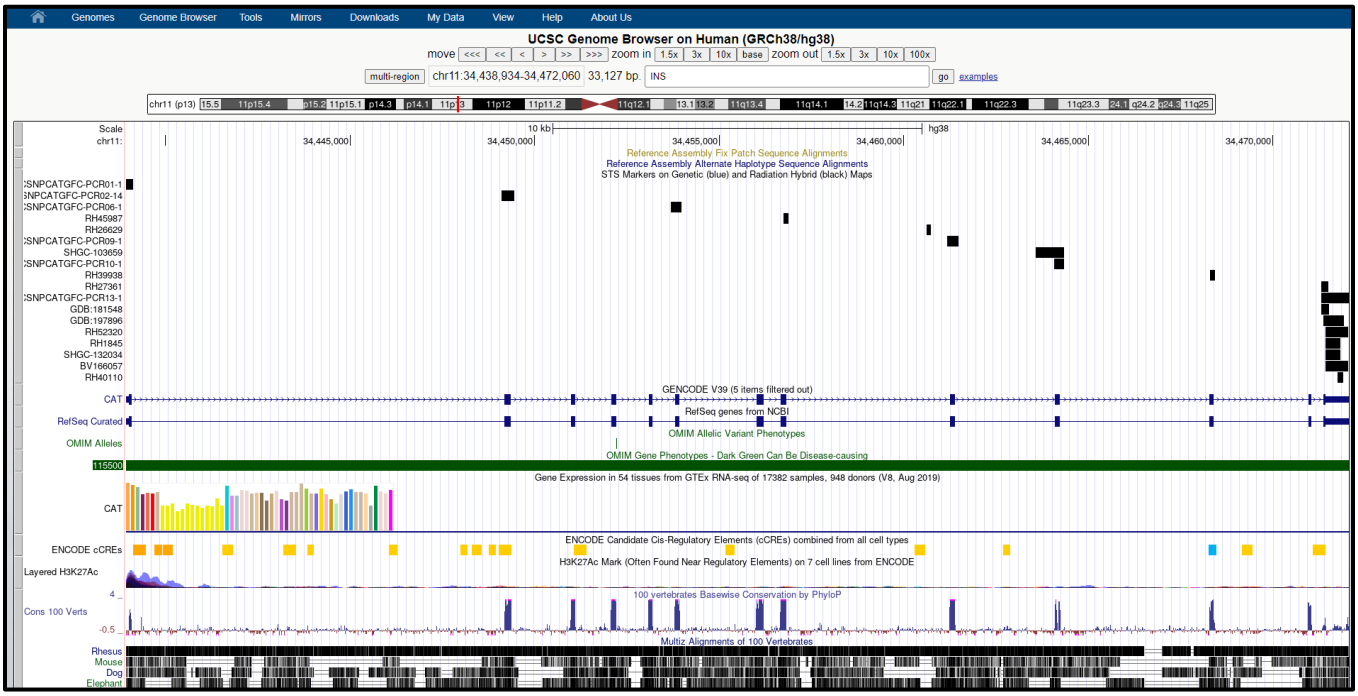


Fig10. Navigation by gene name: INS

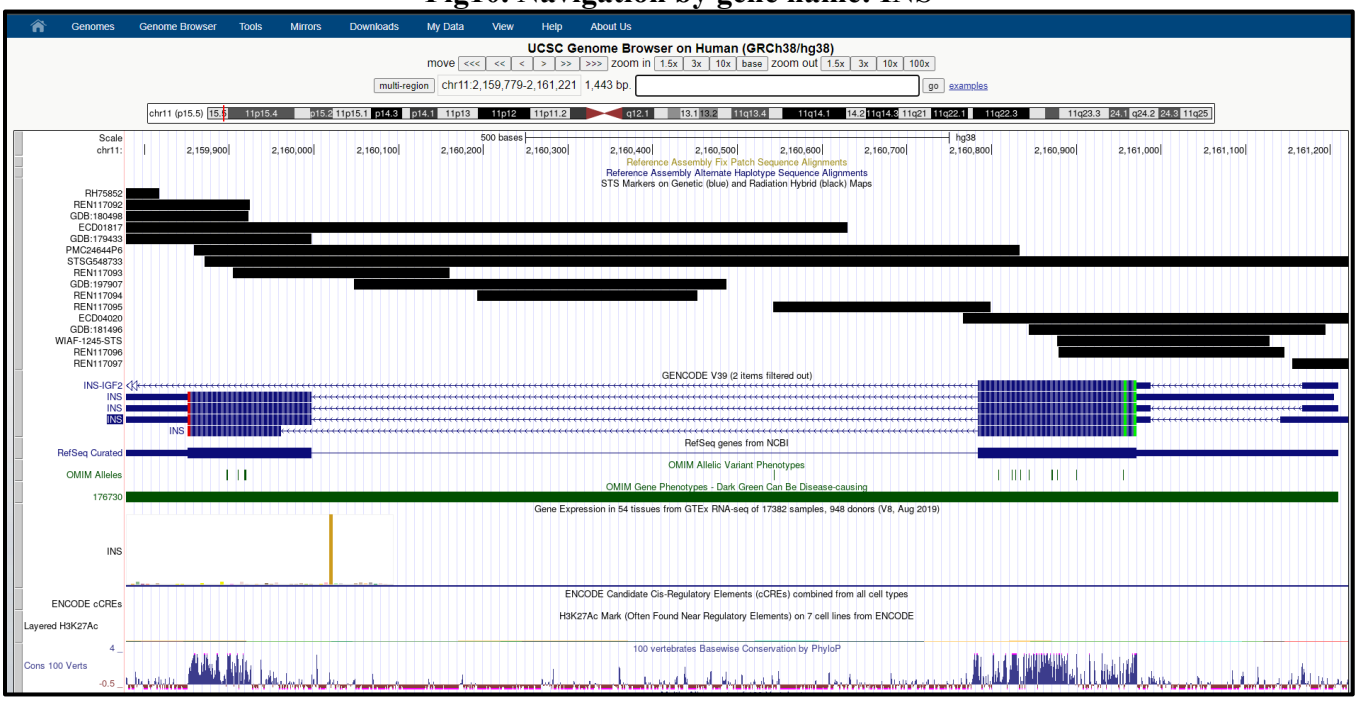


Fig11. Result for INS gene

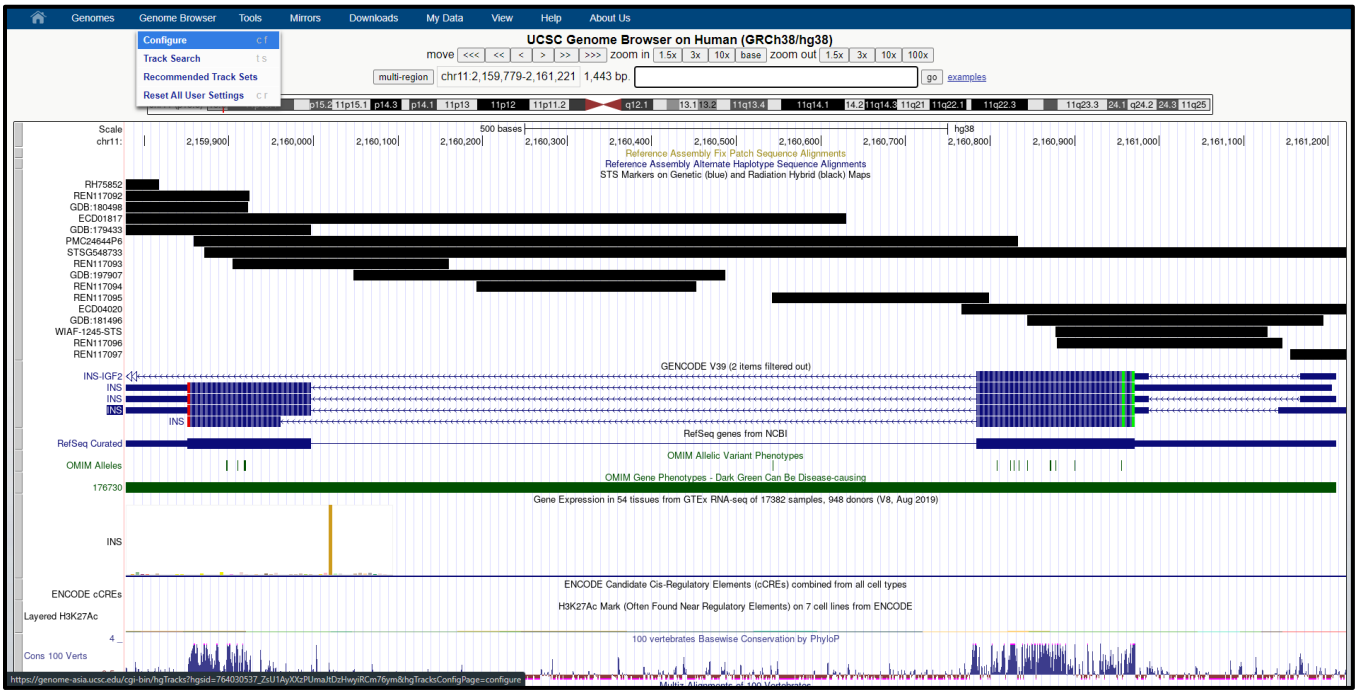


Fig12. Option to configure tracks

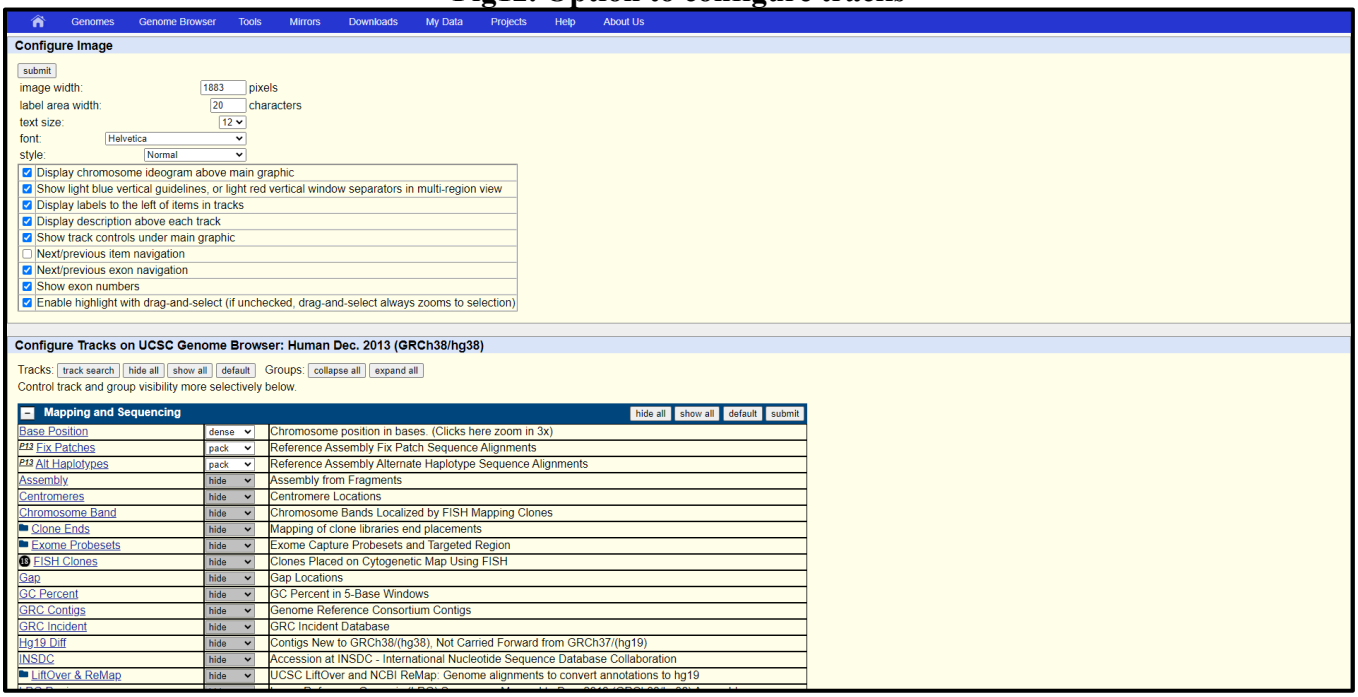


Fig13. Configuration settings

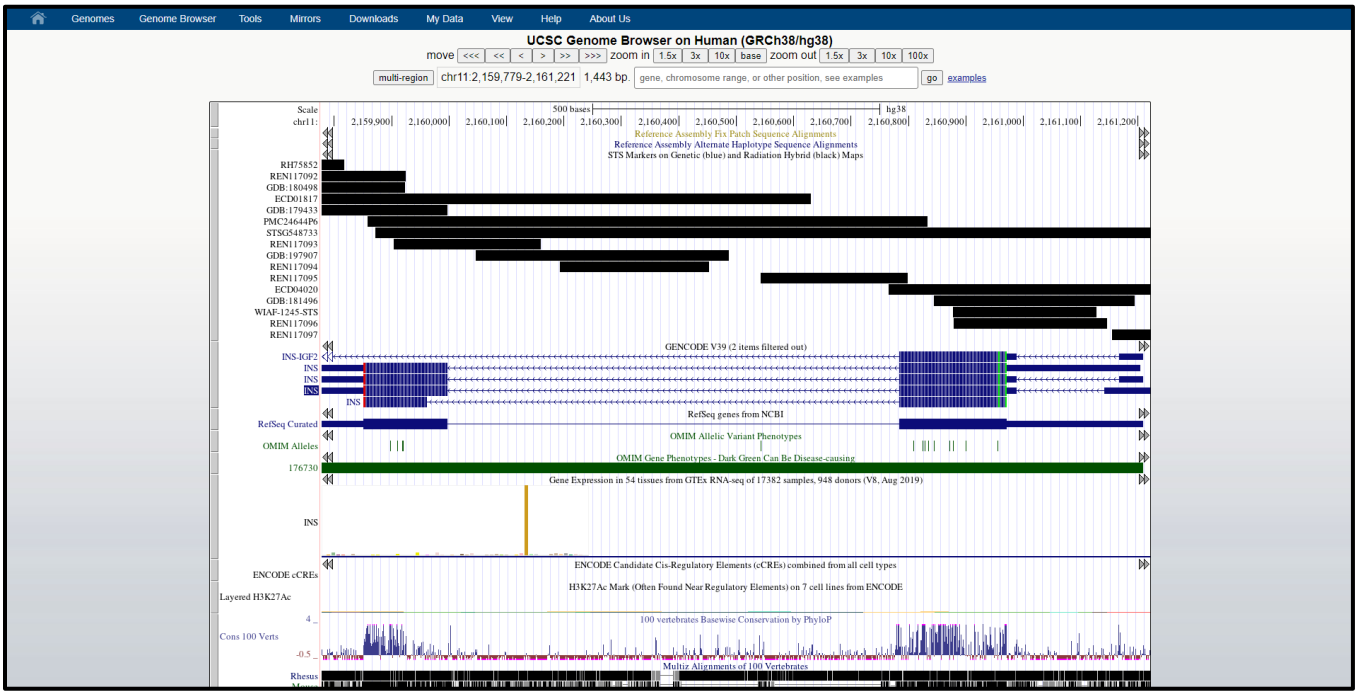


Fig14. Configured tracks

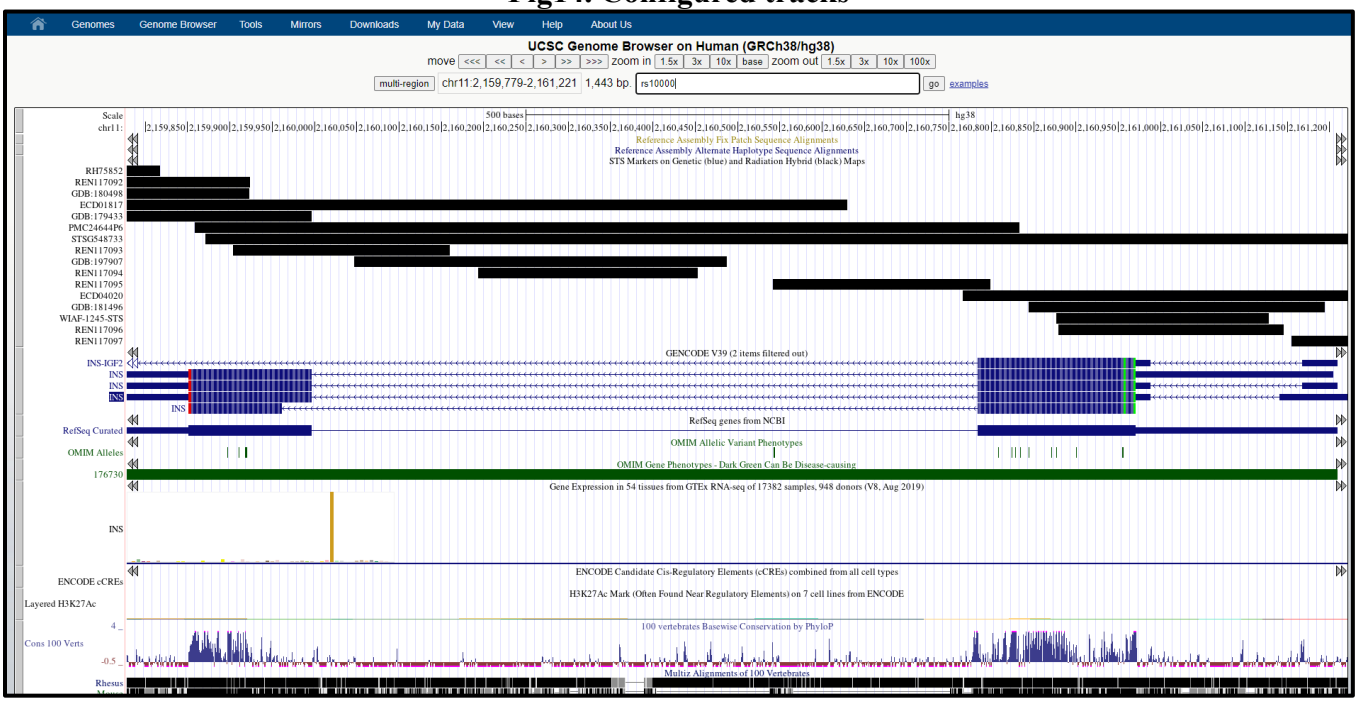


Fig15. Navigation by SNP id: rs10000

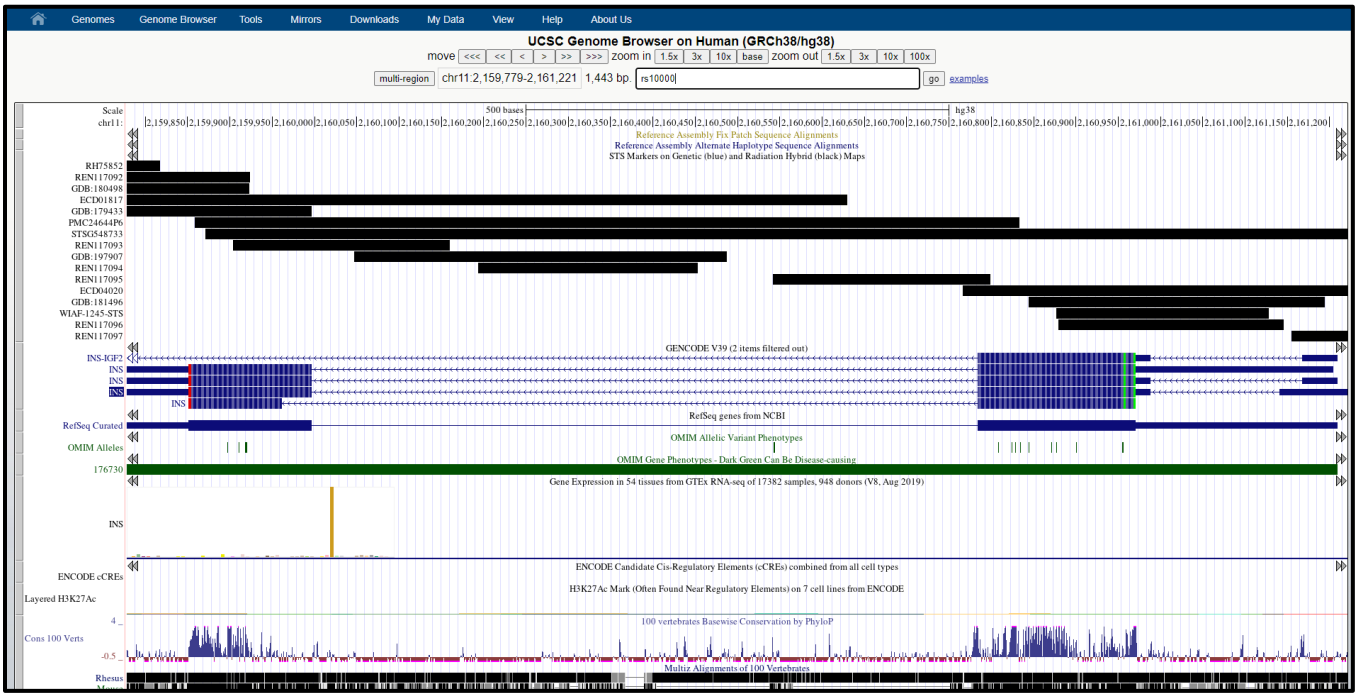


Fig16. Result for SNP id: rs10000

Fig17. Description for SNP id: rs10000

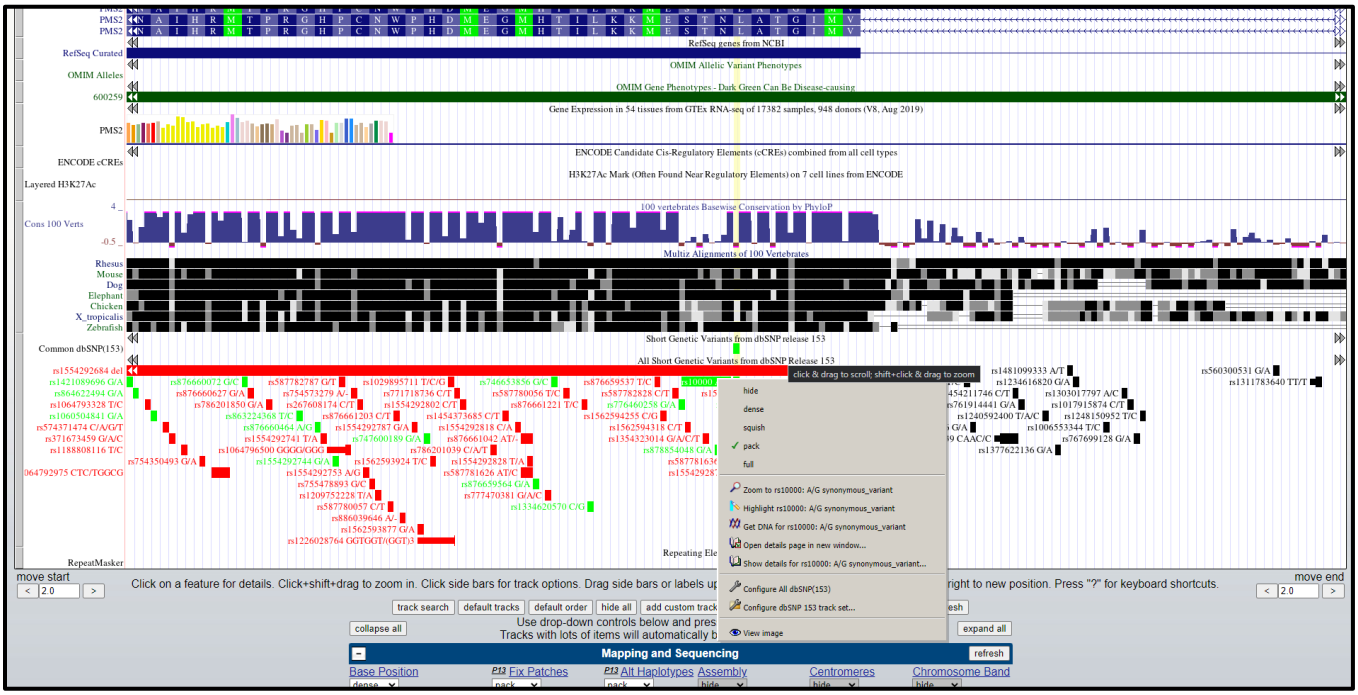


Fig18. Configuration option by right click

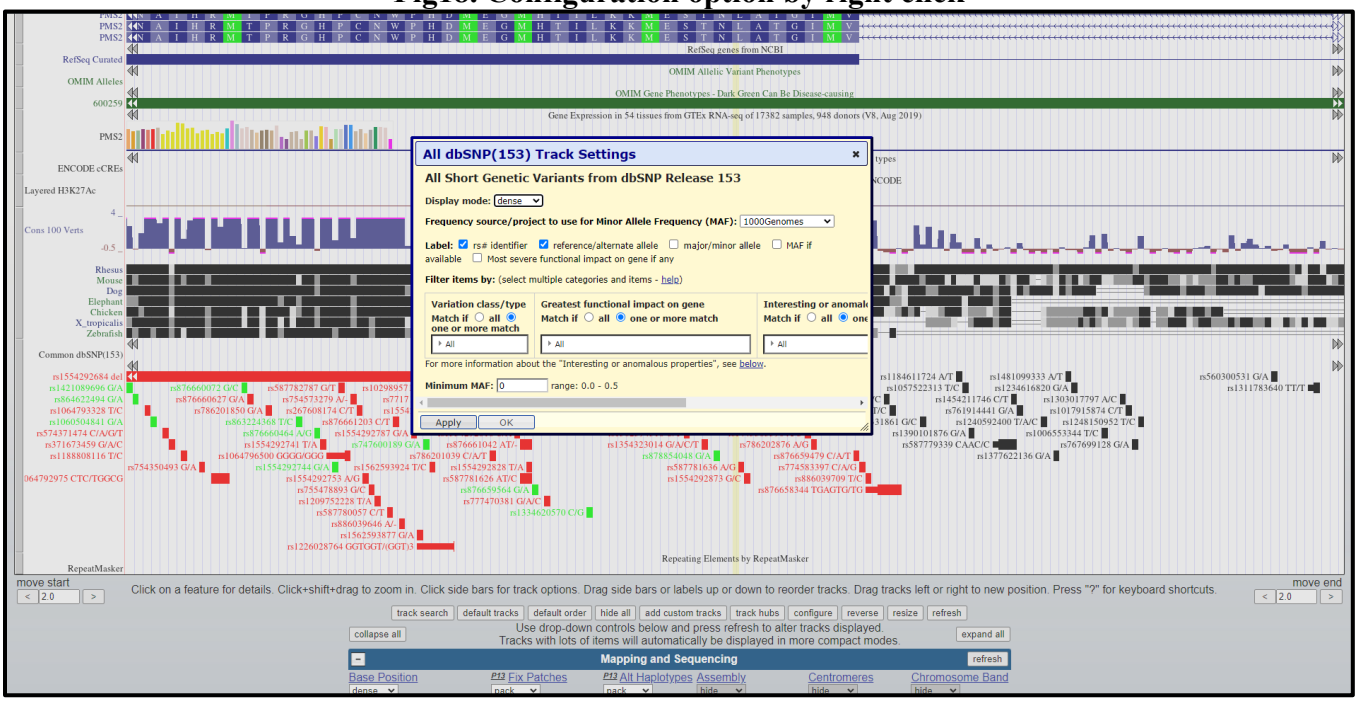


Fig19. Configuration applied

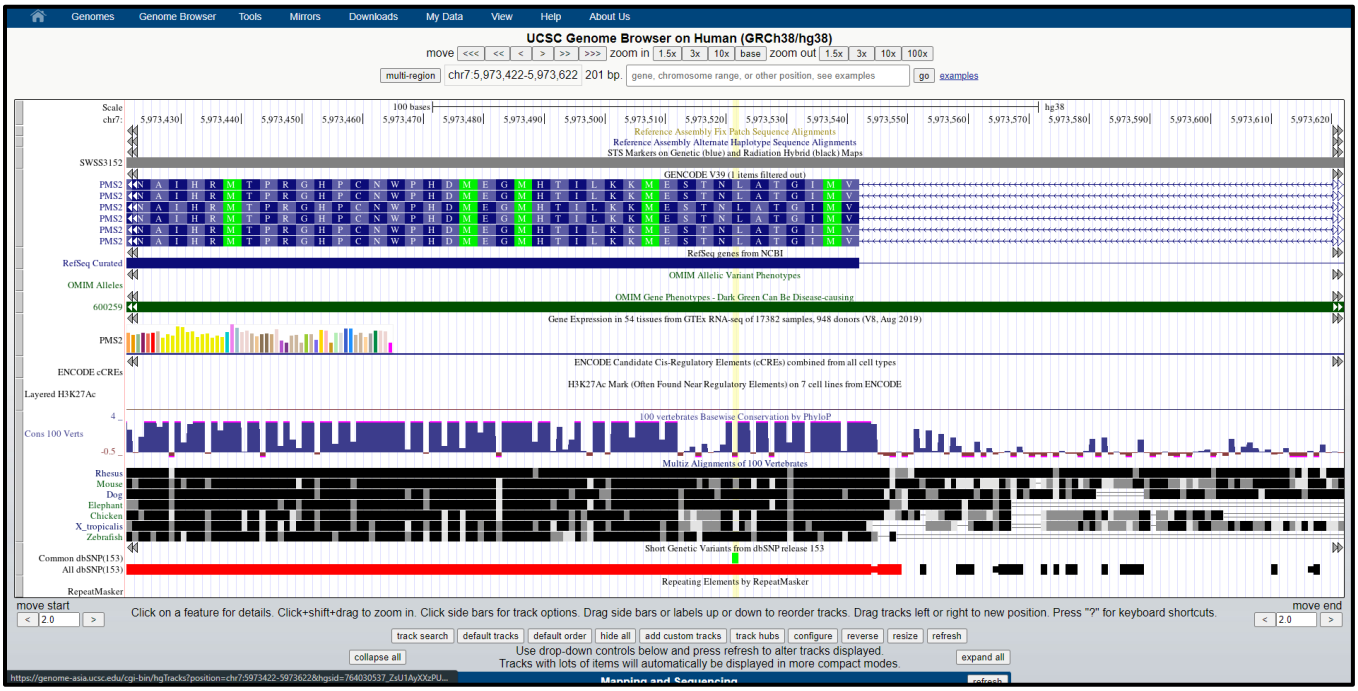


Fig20. Result after configuration

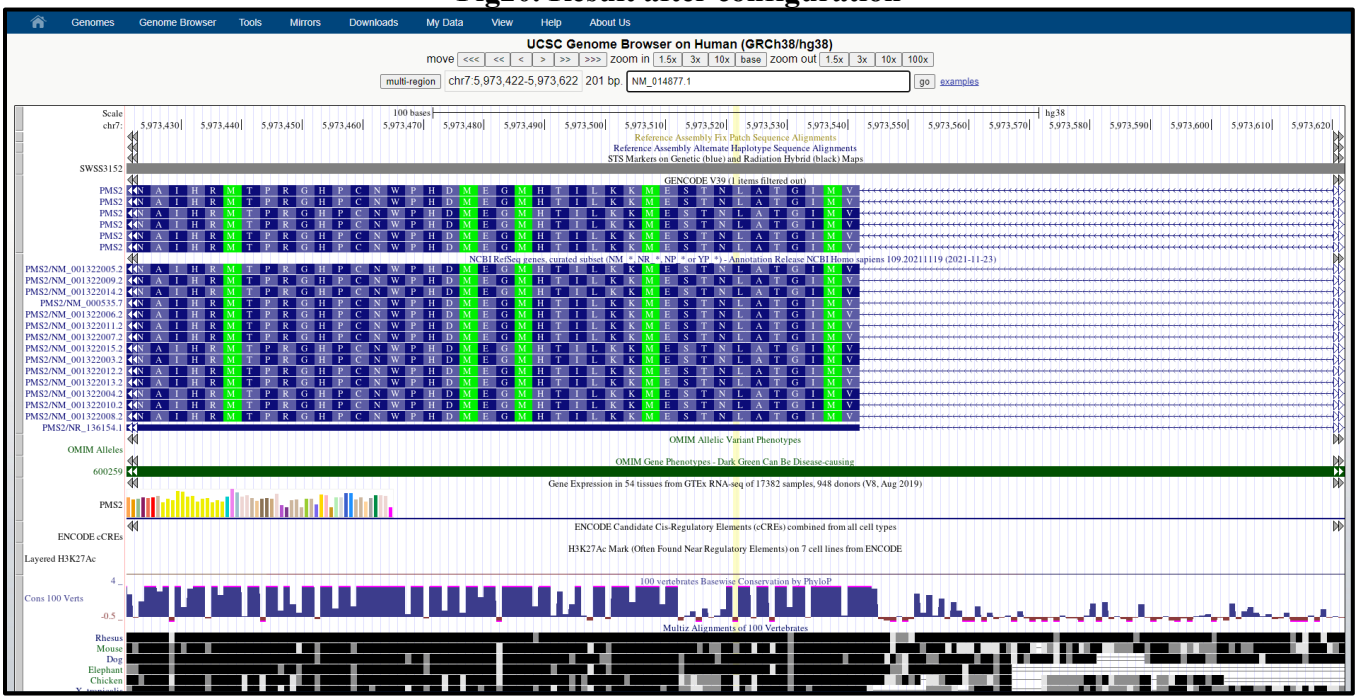


Fig21. Navigation by Ref_Seq: NM_014877.1 (HEL gene)

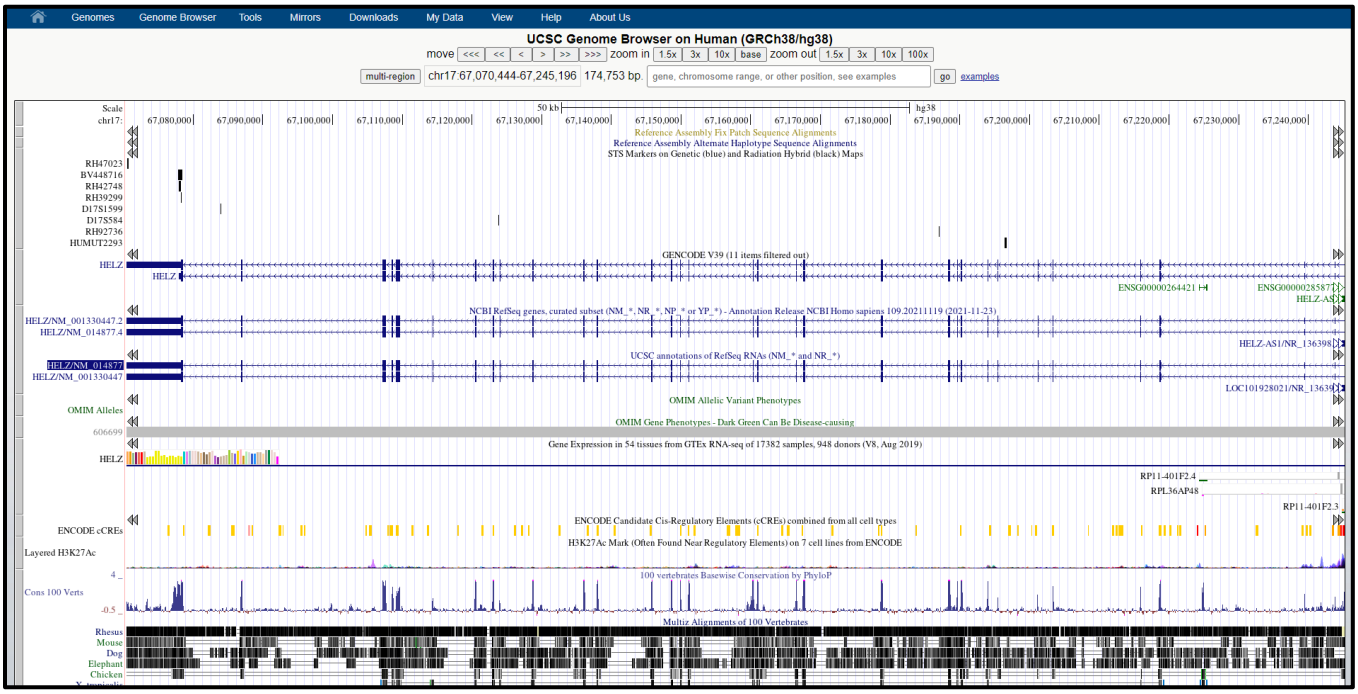


Fig22. Result for Ref_Seq: NM_014977.1 (HELZ gene)

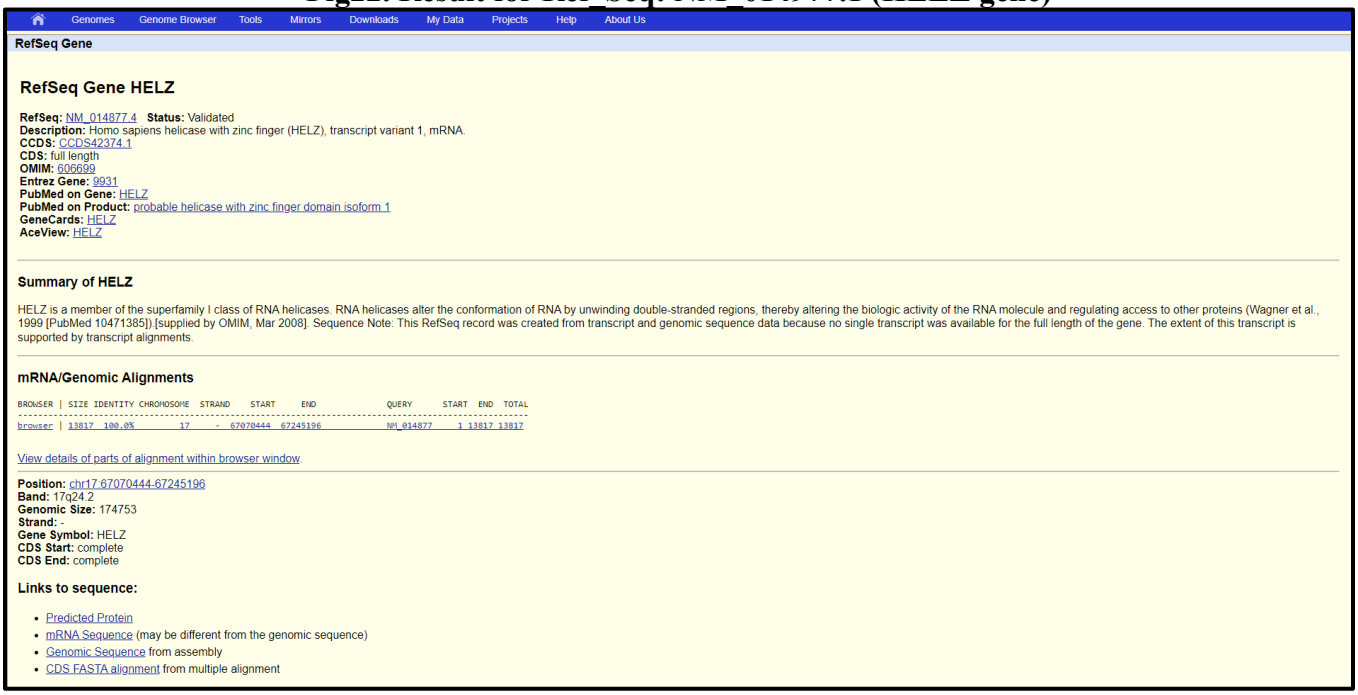


Fig23. Description for Gene HELZ

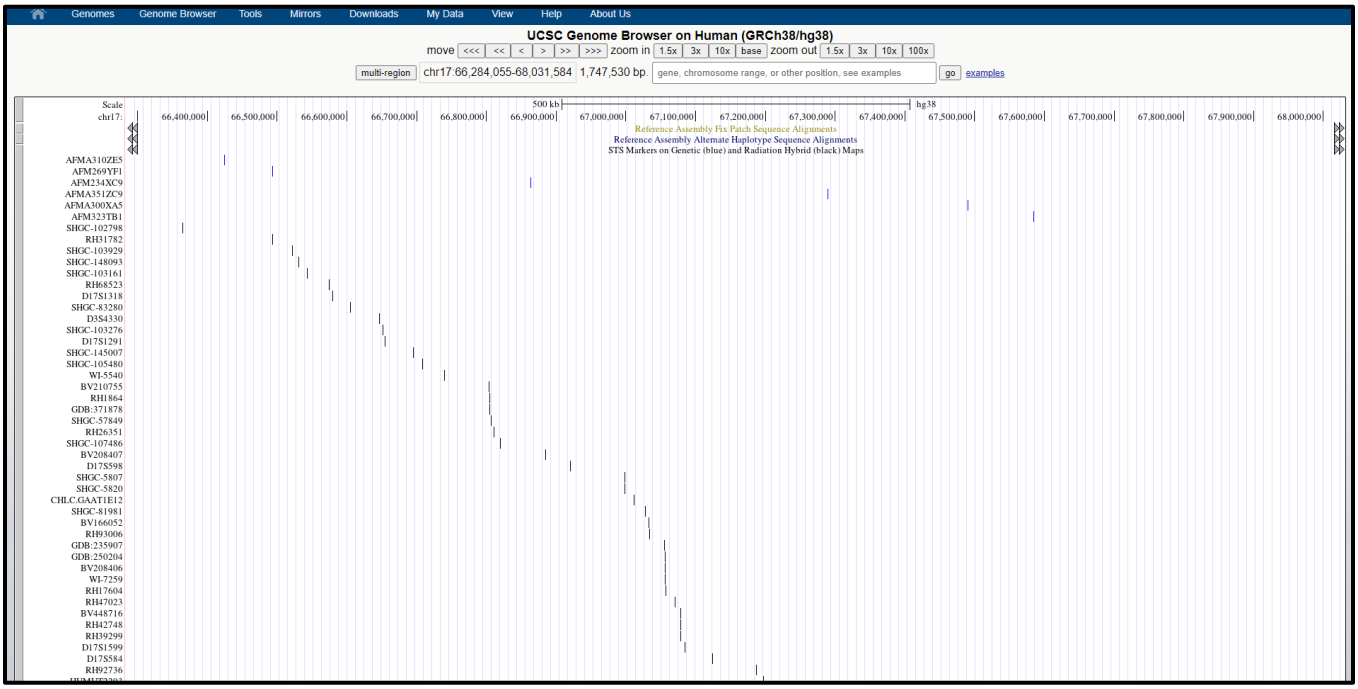


Fig24. Result after zooming 10x

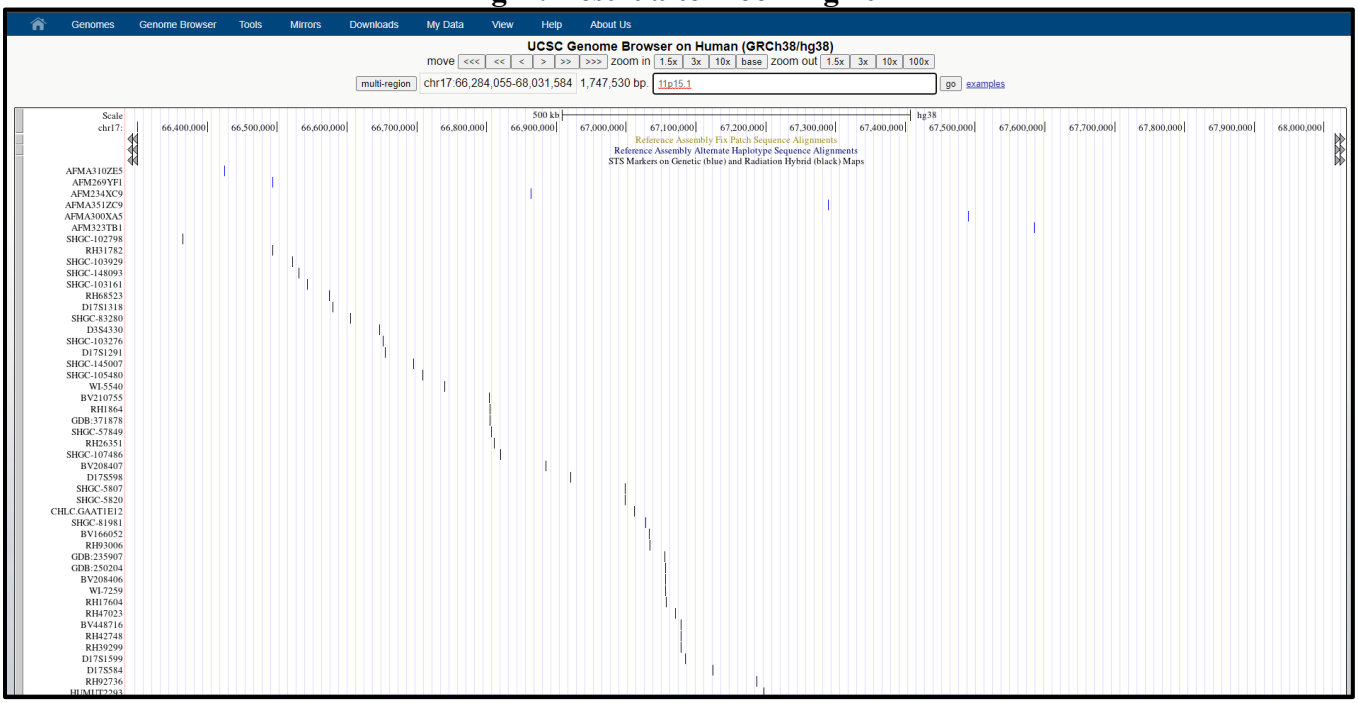


Fig25. Navigation by cytological band: 11p15.1

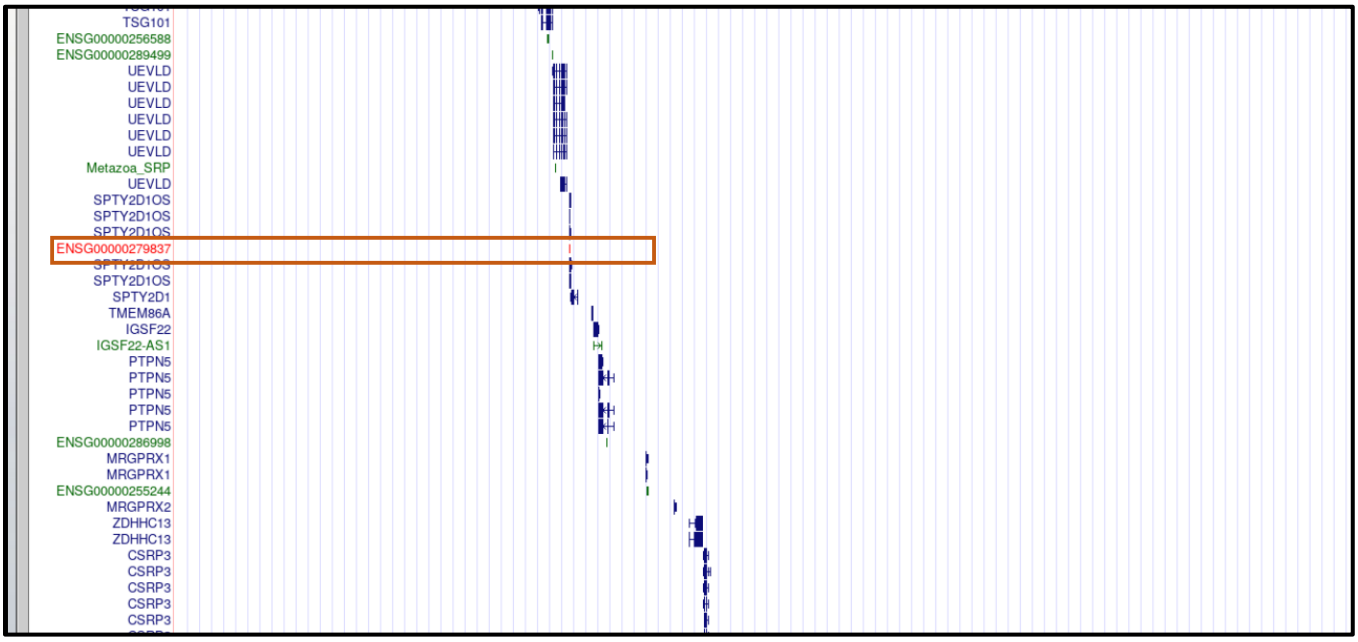


Fig26. Result for cytological band

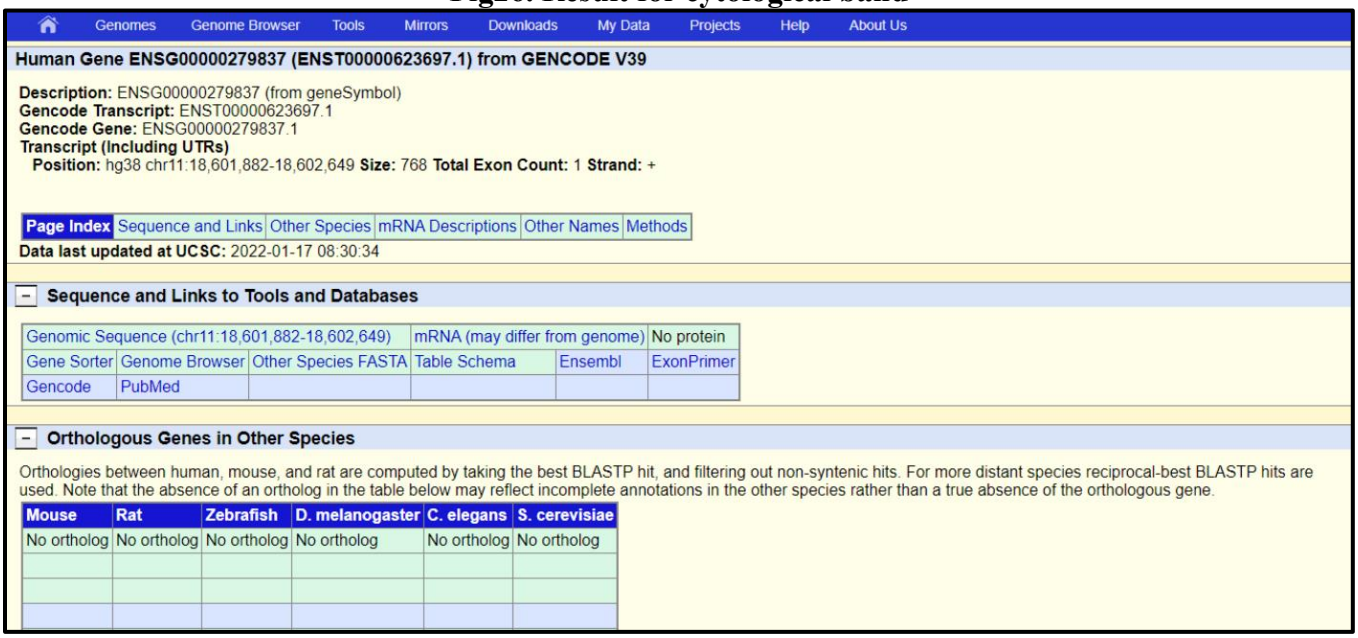


Fig27. Description for cytological ban: 11p15.1

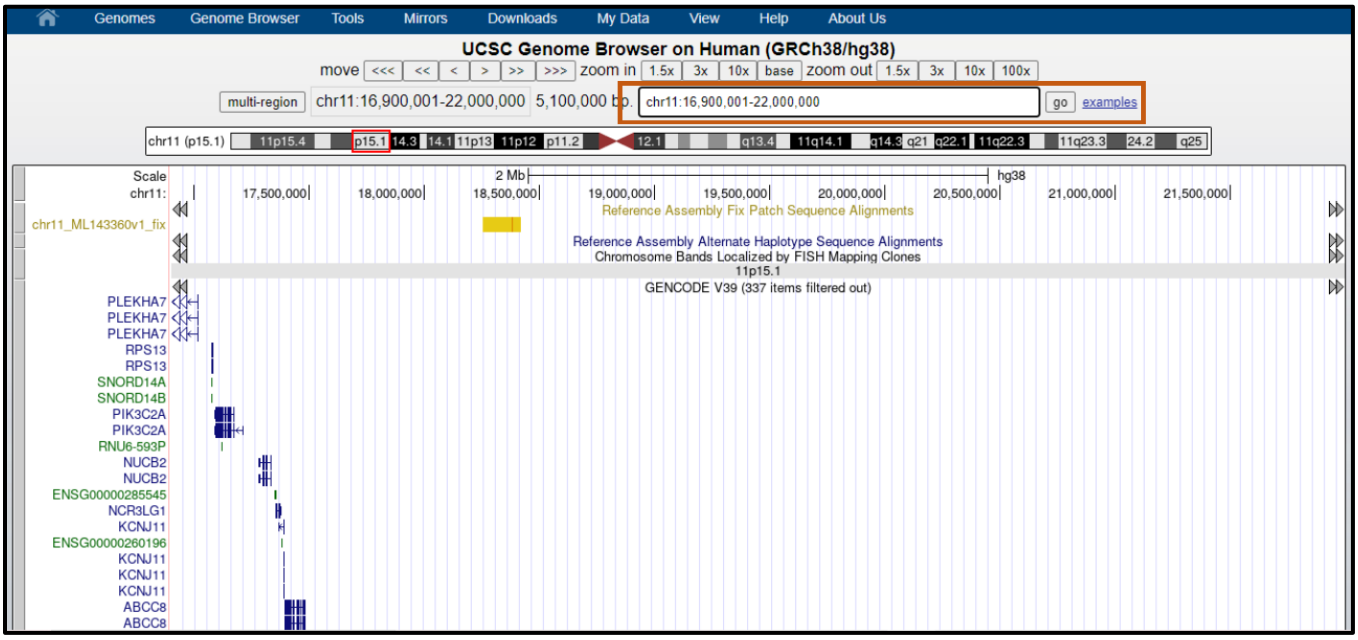


Fig28. Navigation by coordinates

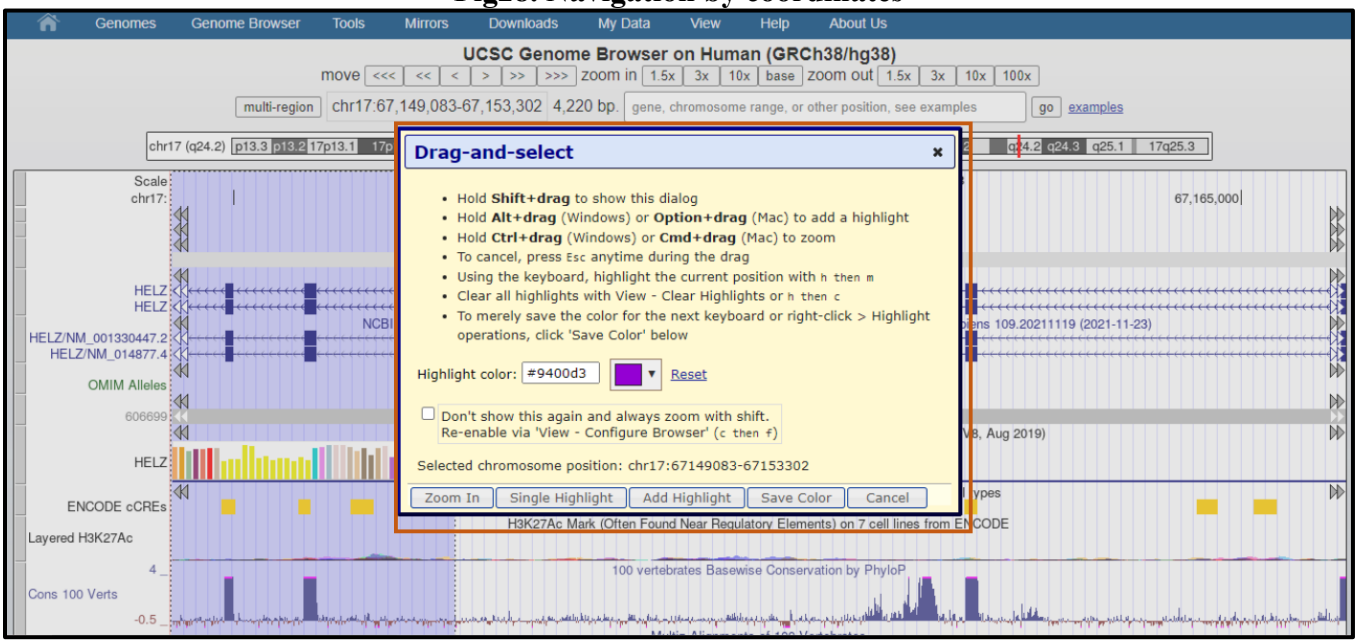


Fig29. Drag and select option for configuration

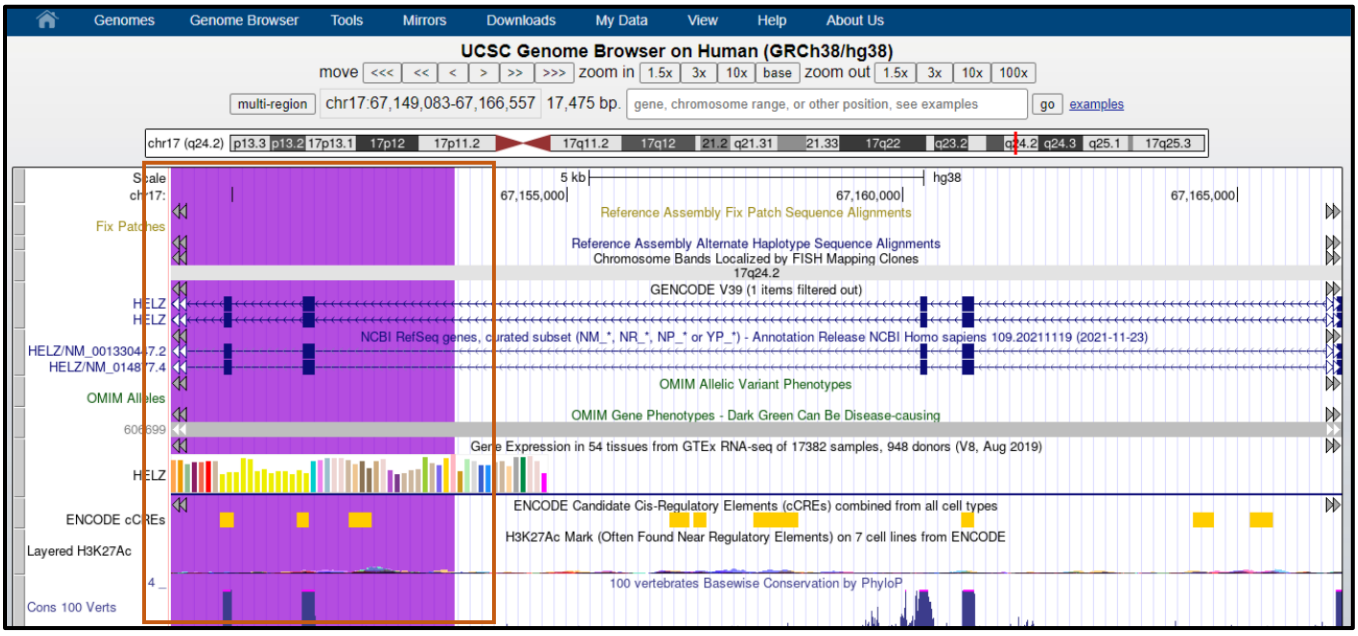


Fig30. Result after configuration

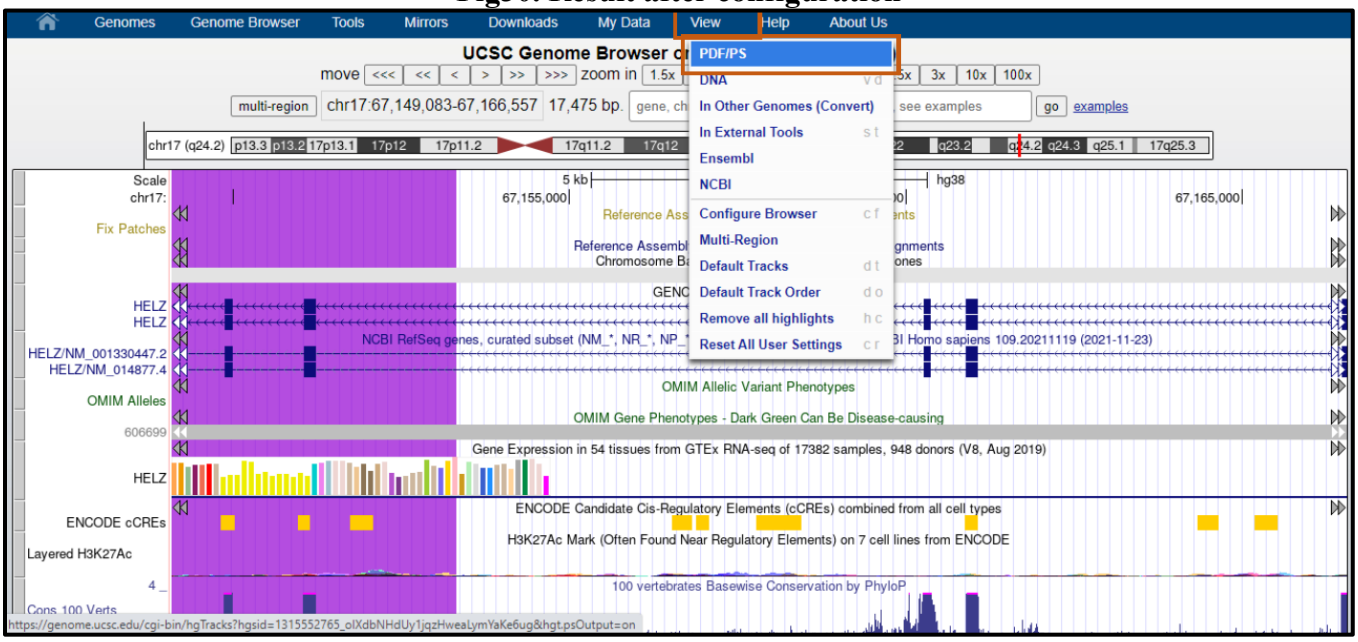


Fig32. Steps to export result as PDF

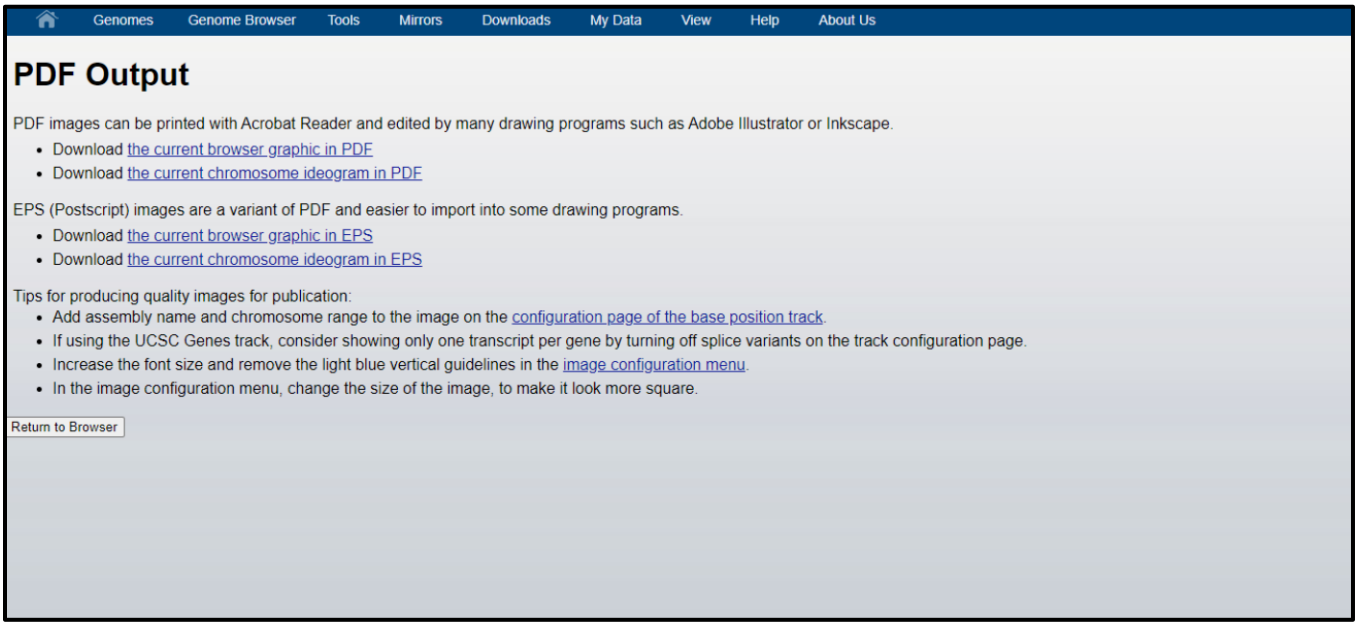


Fig33. Option to download PDF

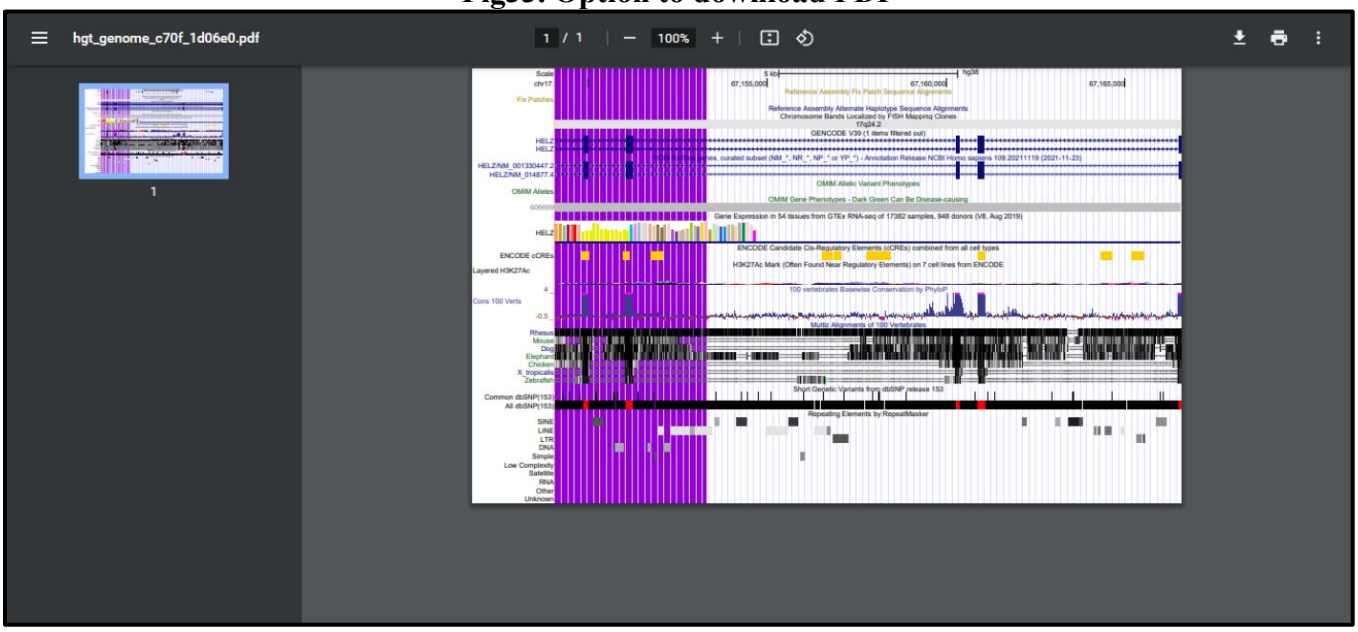


Fig34. Result in PDF format

RESULT:

UCSC genome browser was used for setting for GRCH/hg38 browser and search option used were:

- Navigation by gene name
- Navigation by SNP id
- Navigation by Ref_Seq: NM_014877.4
- Navigation by OMIM Id: 115500
- Navigation by cytological band: 11p15.1

Various options for configuration of tracks, zooming in and out the results, saving results in PDF format, etc were also used.

CONCLUSION:

UCSC genome browser can be used for gene predictions, mRNA and expressed sequence tag alignments,

simple nucleotide polymorphisms, expression and regulatory data, phenotype and variation data, and pairwise and multiple-species comparative genomics data. All information relevant to a region is presented in one window, facilitating biological analysis and interpretation. It also provides various tools for configuration of tracks and refining the results. Options are available for zooming in and out the results and downloading the results in PDF format.

REFERENCES:

1. Baxevanis, Andreas D.; Petsko, Gregory A.; Stein, Lincoln D.; Stormo, Gary D. (2002). Current Protocols in Bioinformatics || The UCSC Genome Browser. , (), –. doi:10.1002/0471250953.bi0104s28
2. UCSC Genome Browser Home. (2019). Ucsc.edu. Retrieved March 28, 2022, from https://genome.ucsc.edu/
3. UCSC Genome Browser Gateway. (2018). Ucsc.edu. Retrieved March 28, 2022, from https://genome.ucsc.edu/cgi-bin/hgGateway
4. Human hg38 chr11%3A34438934%2D34472060 UCSC Genome Browser v428. (n.d.).Genome.ucsc.edu. Retrieved March 28, 2022, from https://genome.ucsc.edu/cg/bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A34438934%2D34472060&hgsid=1315552765_oIXdbNHdUy1jqzHweaLymYaKe6ug
5. **Navigation by OMIM id:** OMIM genes - 115500. (n.d.). Genome.ucsc.edu. Retrieved March 28, 2022, from https://genome.ucsc.edu/cgi-bin/hgc?hgsid=1315629613_43o3KD070AuOsLgYd19FgktiKt2H&db=hg38&c=chr11&l=34438933&r=34472060&o=34438933&t=34472060&g=omimGene2&i=115500
6. **Navigation by SNP id:** All Short Genetic Variants from dbSNP Release 153 (rs10000). (n.d.). Genome.ucsc.edu. Retrieved March 28 2022, from https://genome.ucsc.edu/cgi-bin/hgc?hgsid=1315552765_oIXdbNHdUy1jqzHweaLymYaKe6ug&db=hg38&c=chr7&l=5973421&r=5973622&o=5973521&t=5973522&g=dbSnp153&i=rs10000
7. **Navigation by Ref_Seq:** NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - NM_014877.4. (n.d.).Genome.ucsc.edu. Retrieved March 28, 2022, from https://genome.ucsc.edu/cgibin/hgc?hgsid=1315552765_oIXdbNHdUy1jqzHweaLymYaKe6ug&db=hg38&c=chr17&l=67070443&r=67245196&o=67070443&t=67245196&g=ncbiRefSeqCurated&i=NM_014877.4
8. **Navigation by Cytological band:** Human Gene ENSG00000279837 (ENST00000623697.1) from GENCODE V39. (n.d.)Genome.ucsc.edu. Retrieved March 28, 2022, from https://genome.ucsc.edu/cgibin/hgGene?hgg_gene=ENST00000623697.1&hgg_chrom=chr11&hgg_start=18601881&hgg_end=18602649&hgg_type=knownGene&db=hg38

DATE: 30-03-22

WEBLEM 9b
Ensemble Genome Browser
(URL: https://asia.ensemble.org/index.html)

AIM:

To explore Ensembl genome browser in order to gather information for annotated genes/genome/protein/transcript etc.

INTROCUTION:

The Ensembl project was initially launched in **1999** with the aim of **developing methodologies** for **automatic annotation** of (human) **genomic sequence** with genes and their **constituent transcripts**. Since that time, the project has broadened substantially in scope; the **Ensembl Genome Browser**, which came online in **2000**, now includes **reference genomic sequence** and **annotation** for nearly **100 chordate organisms**. Ensembl is **rapidly incorporating** new data, including **whole clades** of new species' genomes and reference sequence for **multiple strains of existing species**, such as mouse. In addition, existing annotation is **regularly augmented** by the **inclusion of new data sets**. Ensembl's sister site, **Ensembl Genomes**, provides access to **nonvertebrate genomes** through dedicated portals for **Bacteria, Fungi, Plants, Metazoa, and Protists.**

METHODOLOGY:

1. Open homepage for Ensembl genome browser. (URL: https://asia.ensembl.org/index.html)
2. Select human (GRCH38.p13) genome assembly
3. Search for helad1 gene.
4. Observer the results.
5. Use the configuration tools for the tracks.
6. Interpret the results.

OBSERVATION:

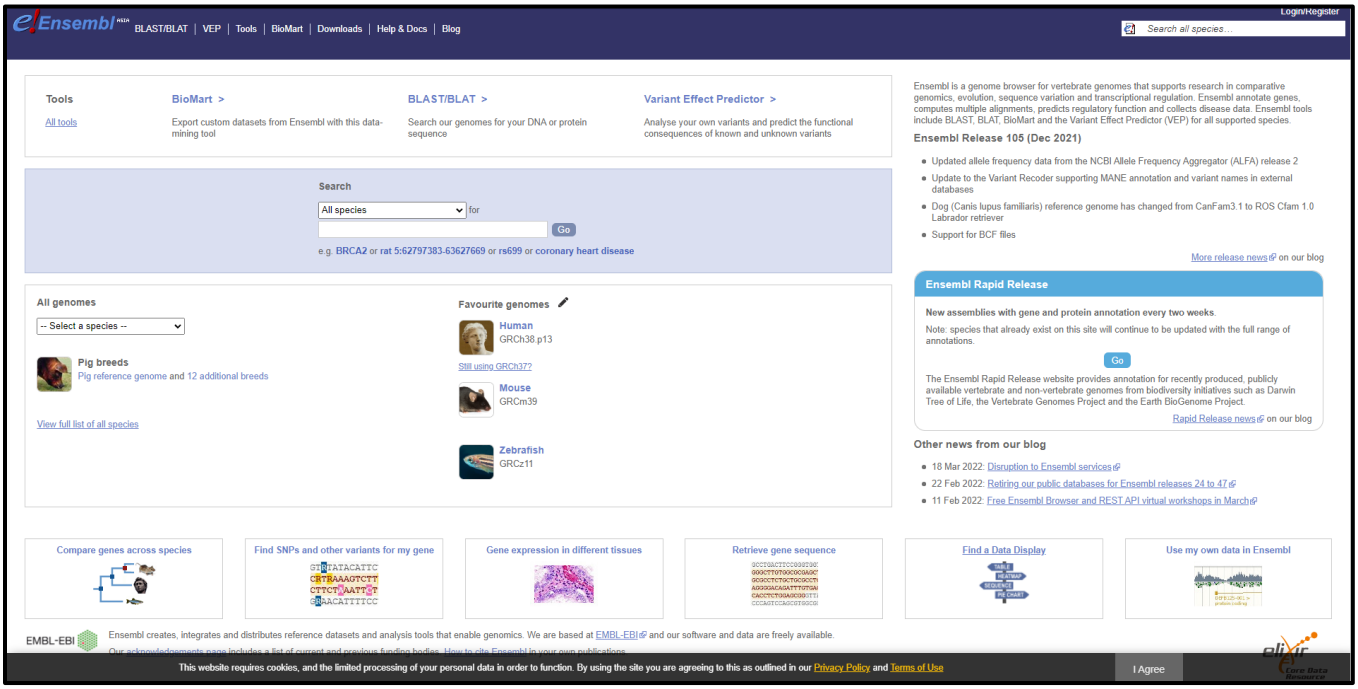


Fig1. Homepage for Ensembl

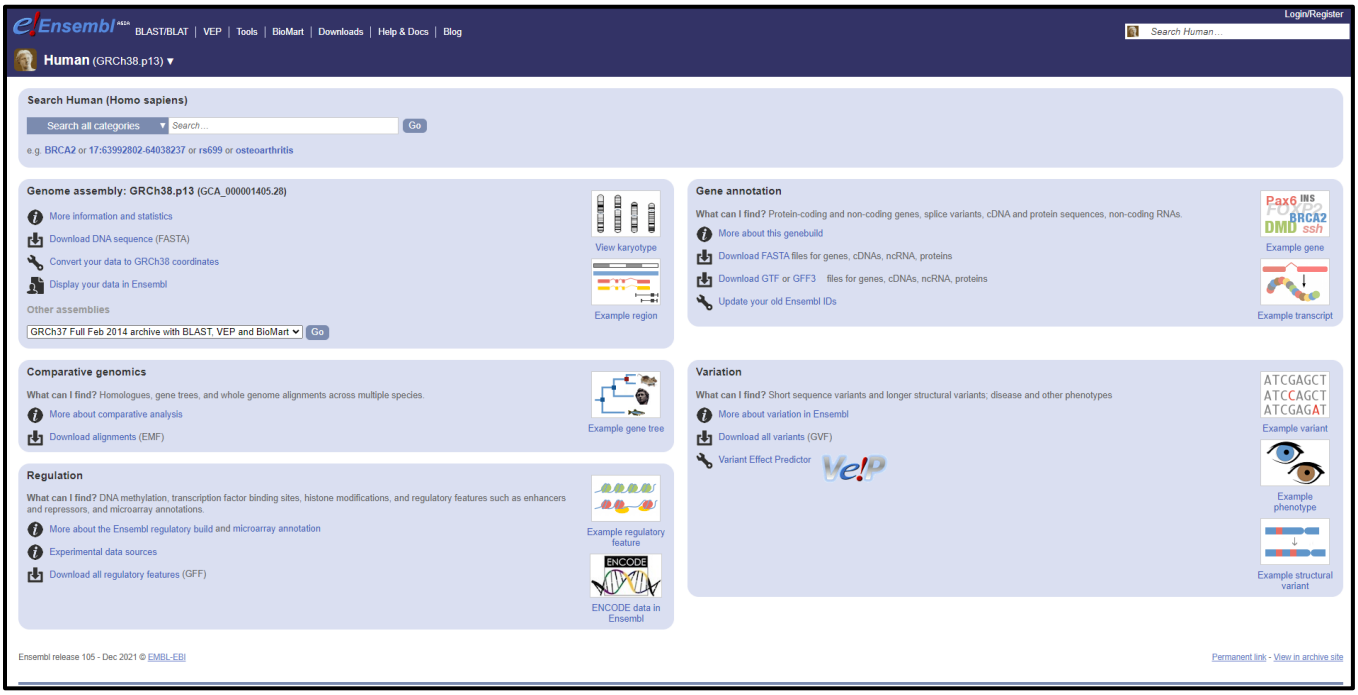


Fig2. Homepage for Human (GRCH28.p13) genome assembly

Fig3. Search for helad1 gene

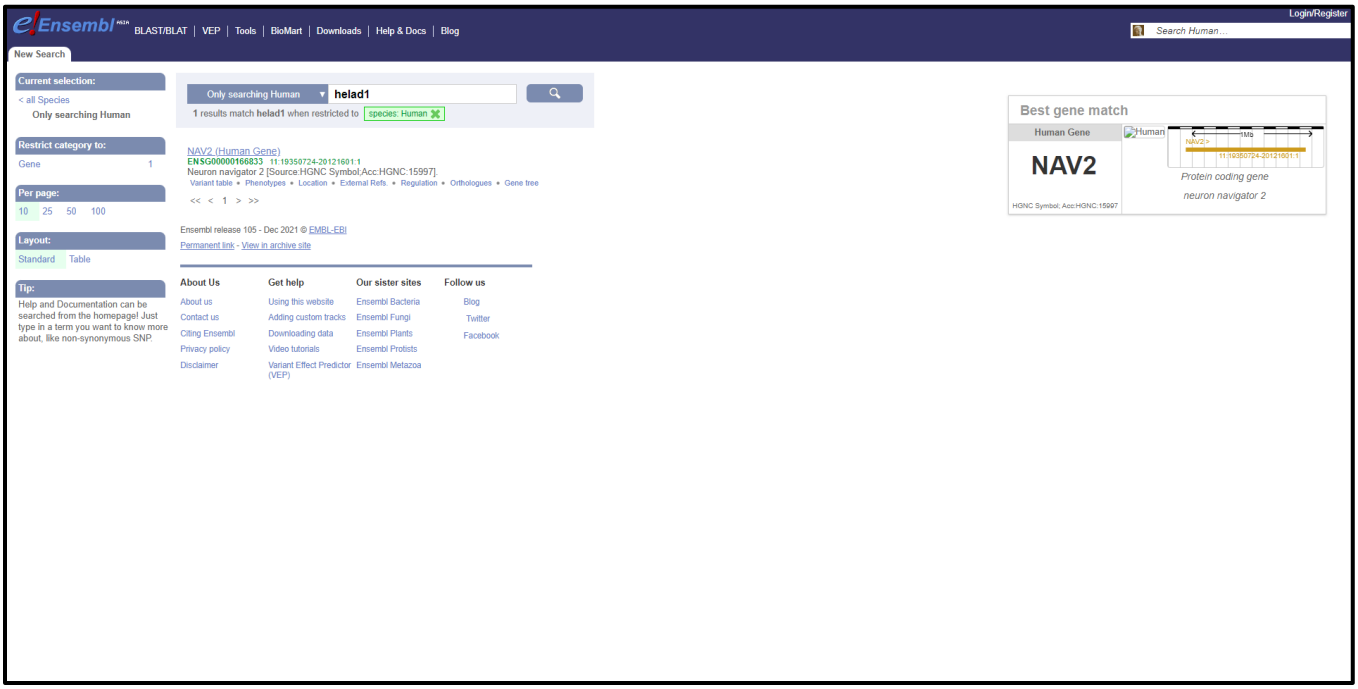


Fig4. Hit page for helad1 gene

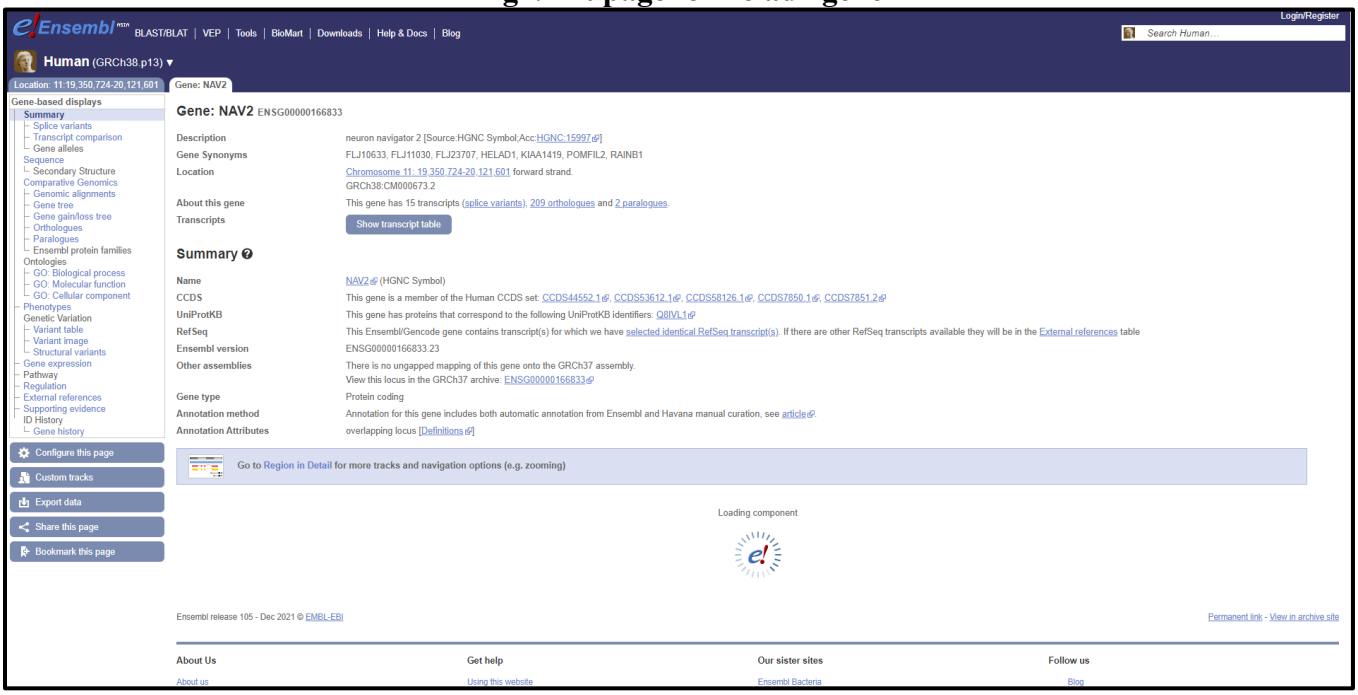
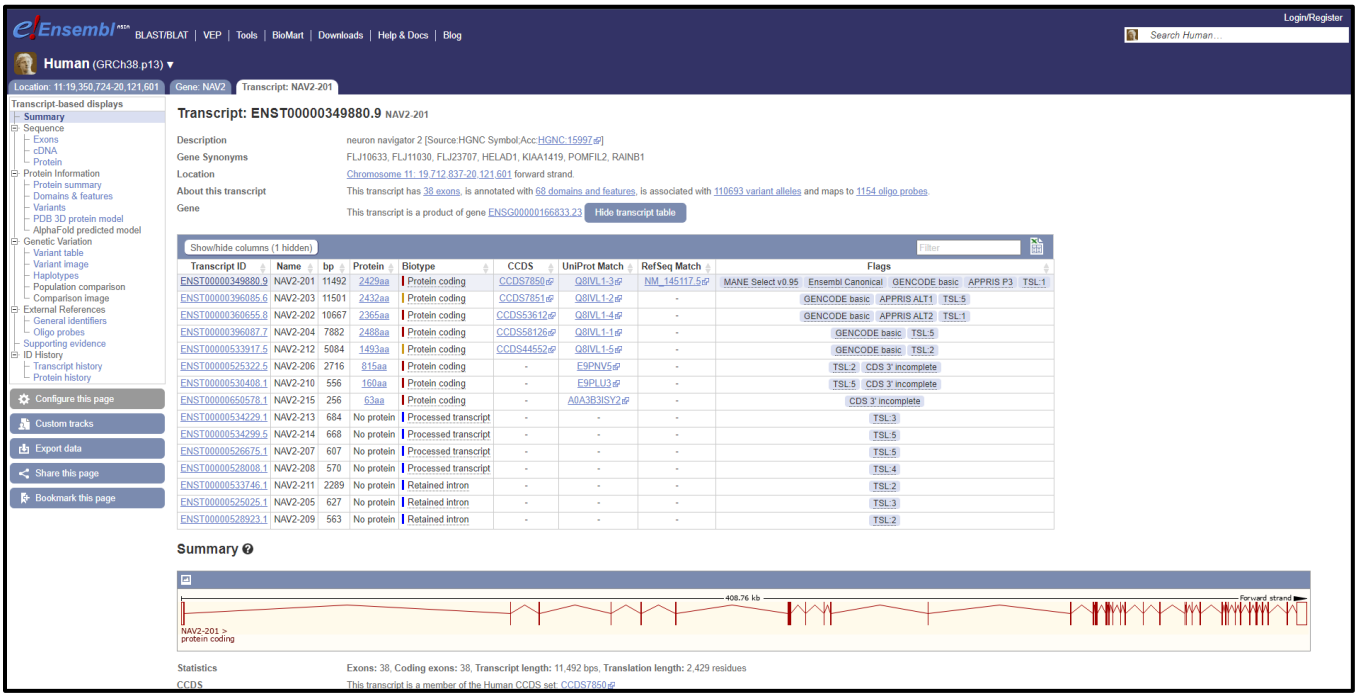


Fig5. Result for NAV2 gene



Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000349880.9	NAV2-201	11492	2429aa	Protein coding	CCDS7850	Q8IVL1-3	NM_145117.5	MANE Select v0.95 Ensembl Canonical GENCODE basic APPRIS P3 TSL:1
ENST00000396085.6	NAV2-203	11501	2432aa	Protein coding	CCDS7851	Q8IVL1-2	-	GENCODE basic APPRIS ALT1 TSL:5
ENST00000360655.8	NAV2-202	10667	2365aa	Protein coding	CCDS53612	Q8IVL1-4	-	GENCODE basic APPRIS ALT2 TSL:1
ENST00000396087.7	NAV2-204	7882	2488aa	Protein coding	CCDS58126	Q8IVL1-1	-	GENCODE basic TSL:5
ENST00000533917.5	NAV2-212	5084	1493aa	Protein coding	CCDS44552	Q8IVL1-5	-	GENCODE basic TSL:2
ENST00000525322.5	NAV2-206	2716	815aa	Protein coding	-	E9PNV5	-	TSL:2 CDS 3' incomplete
ENST00000530408.1	NAV2-210	556	160aa	Protein coding	-	E9PLU3	-	TSL:5 CDS 3' incomplete
ENST00000650578.1	NAV2-215	256	63aa	Protein coding	-	A0A3B3ISY2	-	CDS 3' incomplete
ENST00000534229.1	NAV2-213	684	No protein	Processed transcript	-	-	-	TSL:3
ENST00000534299.5	NAV2-214	668	No protein	Processed transcript	-	-	-	TSL:5
ENST00000526675.1	NAV2-207	607	No protein	Processed transcript	-	-	-	TSL:5
ENST00000528008.1	NAV2-208	570	No protein	Processed transcript	-	-	-	TSL:4
ENST00000533746.1	NAV2-211	2289	No protein	Retained intron	-	-	-	TSL:2
ENST00000525025.1	NAV2-205	627	No protein	Retained intron	-	-	-	TSL:3
ENST00000528923.1	NAV2-209	563	No protein	Retained intron	-	-	-	TSL:2

Fig6. Result for NAV2-201

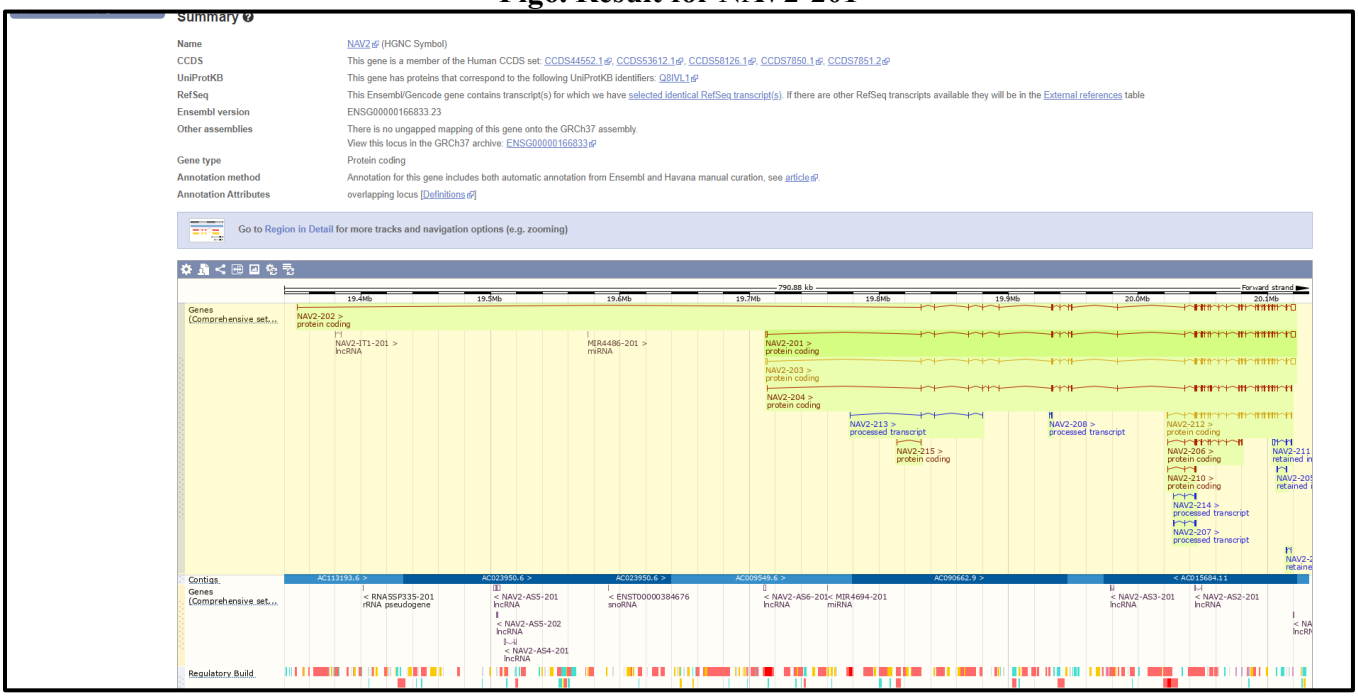


Fig7. Result for NAV2

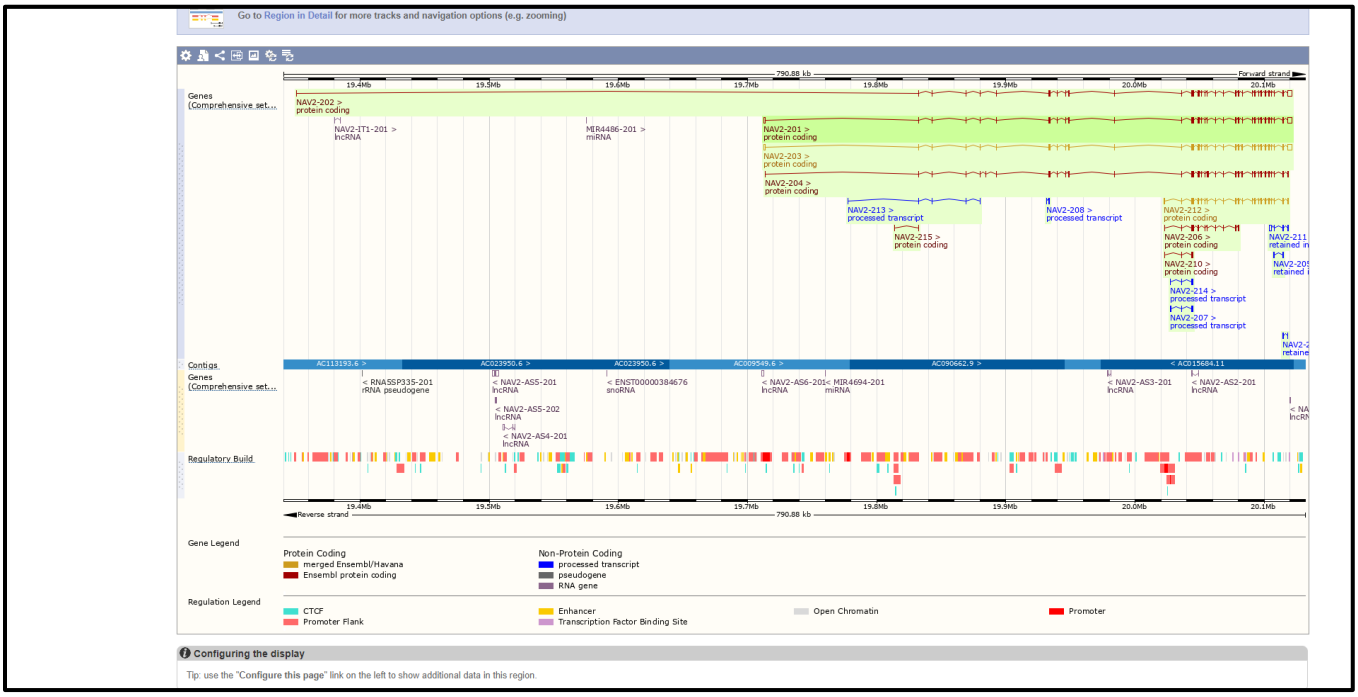


Fig6. Tracks information

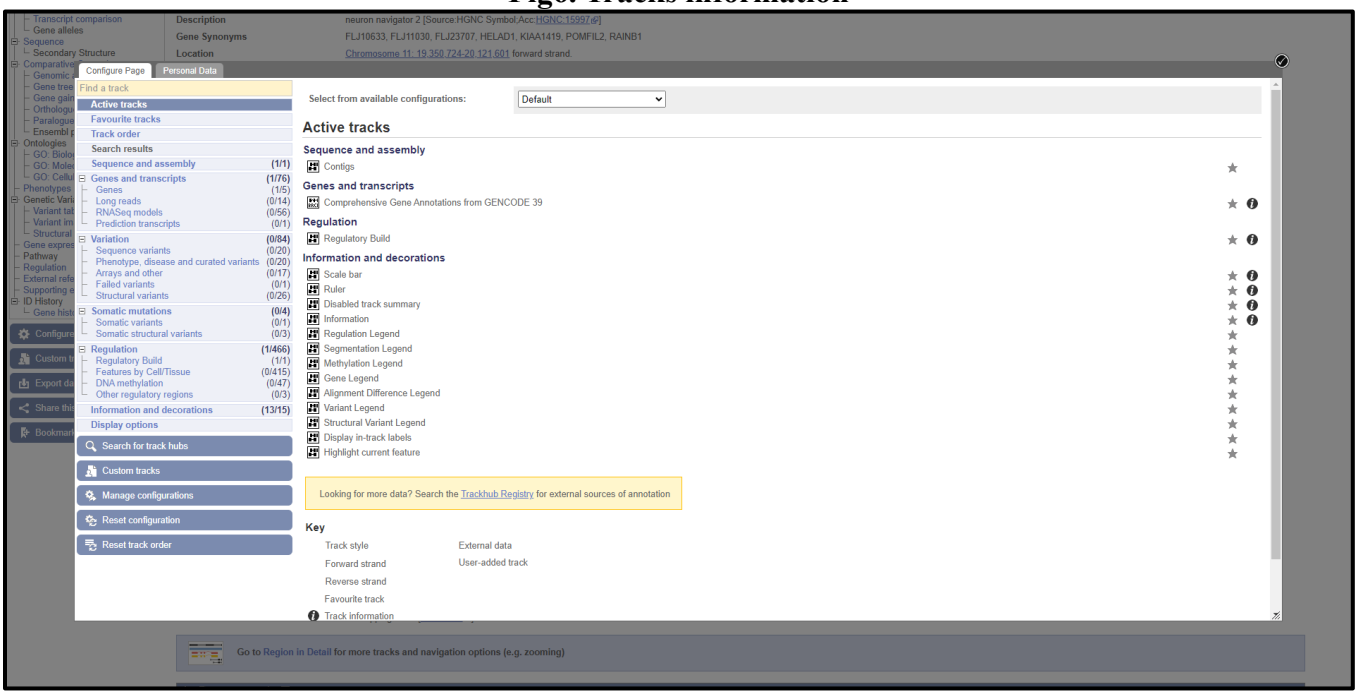


Fig7. Option for tracks configuration

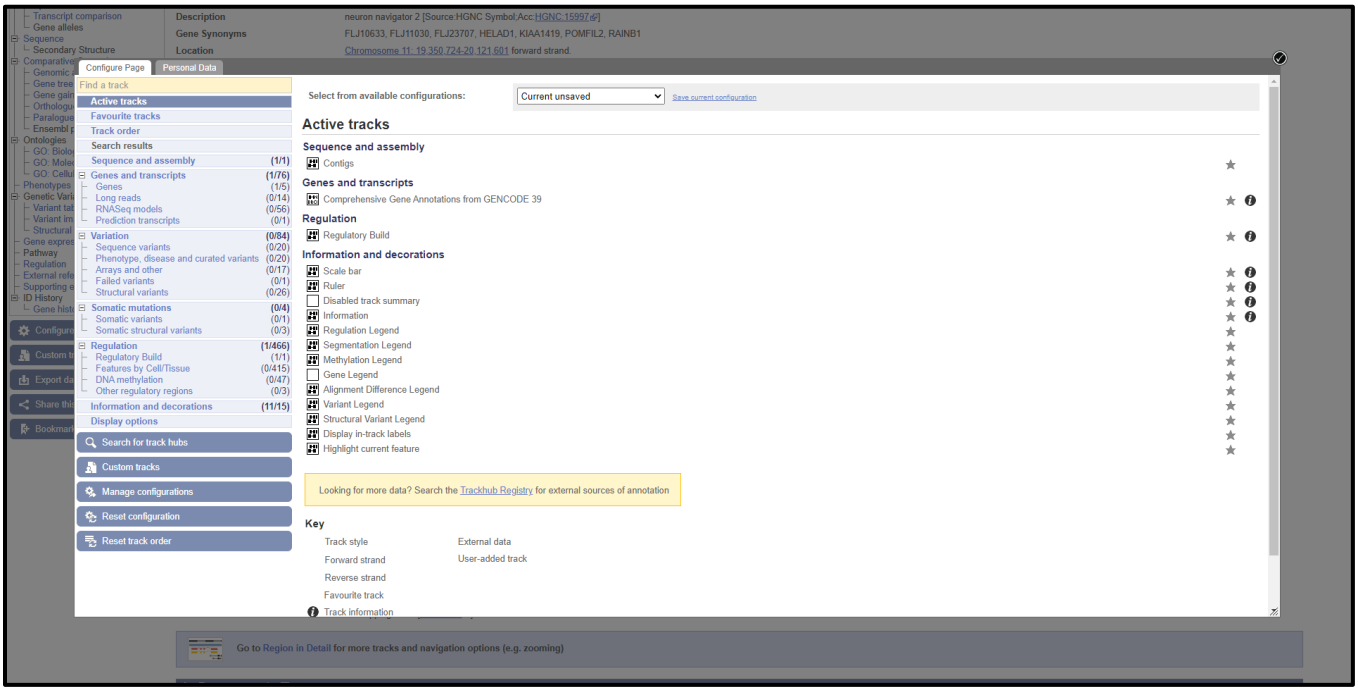


Fig8. Tracks configuration

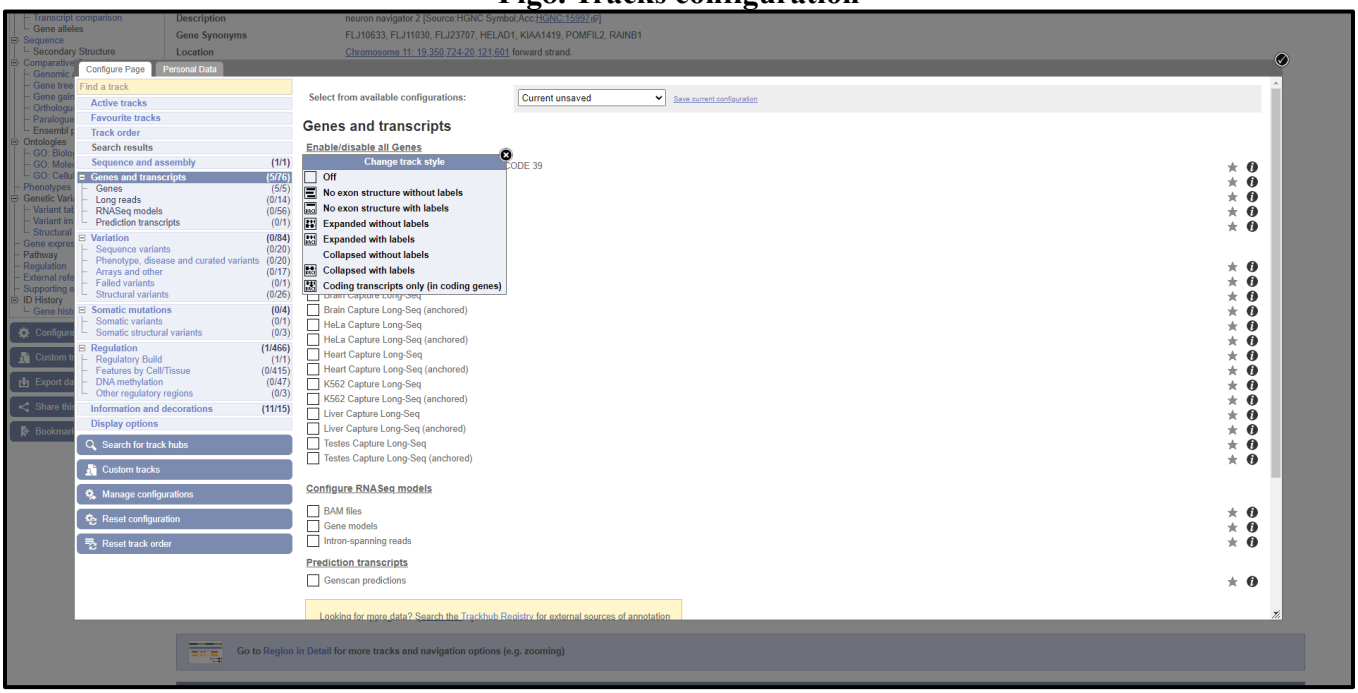


Fig9. Tracks configuration

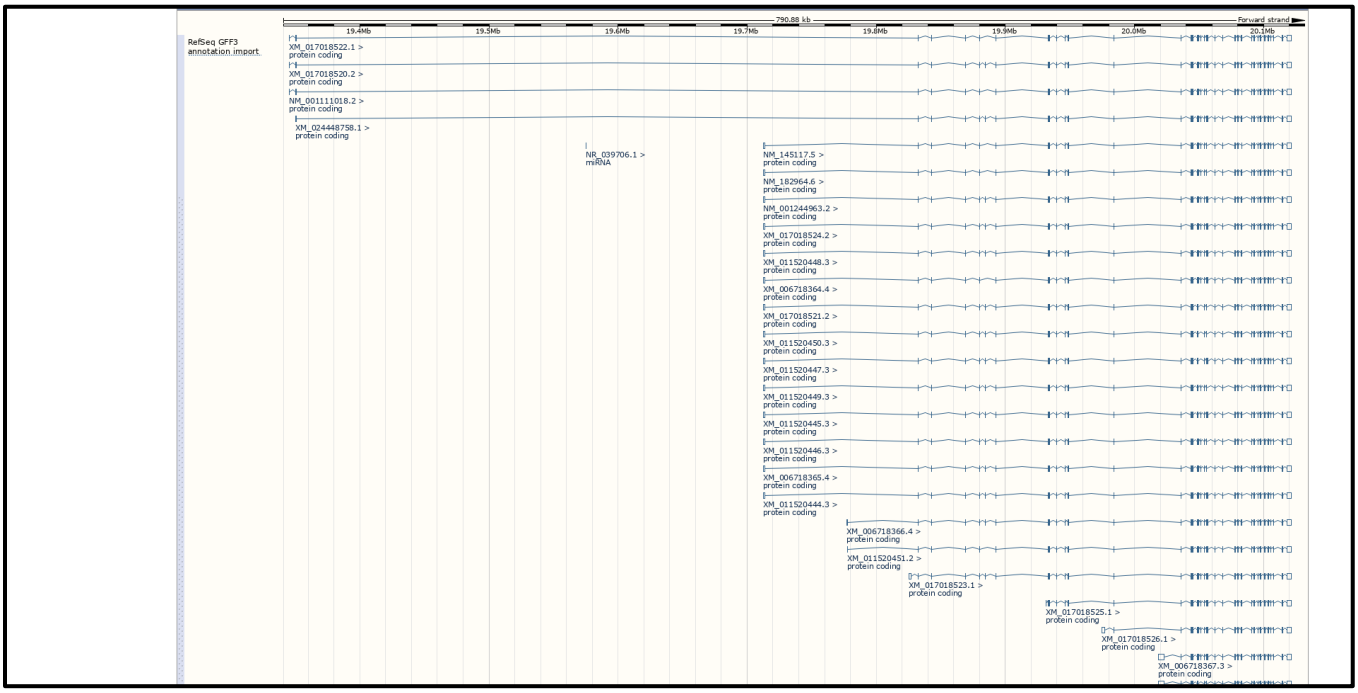


Fig10. Updated results after track configuration

RESULT:

Emsembl genome browser was used to search for helad1 gene under human genome assembly and was explored for various tracks configuration options.

CONCLUSTION:

Ensembl genome browser provides annotation of (human) genomic sequence with genes and their constituent transcripts. Beyond providing access to data related to publicly available genome annotation, Ensembl integrates a number of tools designed to process or analyze your own data. Sequence alignment using BLAST and BLAT against Ensembl genes, genomes and proteins is also available, along with a suite of tools developed as part of the 1000 Genomes Project that can be accessed on the dedicated GRCh37 browser site.

REFERENCES:

1. Karolchik, D. (2003). The UCSC Genome Browser Database. , 31(1), 51–54. doi:10.1093/nar/gkg129
2. Ensembl genome browser 100. (n.d.-b). Uswest.ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/index.html
3. Homo_sapiens - Ensembl genome browser 104. (2014). Ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Homo_sapiens/Info/Index
4. helad1 - Search - Homo_sapiens - Ensembl genome browser 105. (2021b). Ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Human/Search/Results?q=helad1
5. Summary - Homo sapiens - Ensembl genome browser 100. (n.d.). Uswest.ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Homo_sapiens/Gene/Summary?db=core
6. Summary - Homo sapiens - Ensembl genome browser 100. (n.d.). Uswest.ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000166833;r=11:19350724-20121601;t=ENST00000349880
7. Summary - Homo sapiens - Ensembl genome browser 100. (n.d.). Uswest.ensembl.org. Retrieved March 28, 2022, from

https://asia.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000166833;r=11:19350724-20121601;t=ENST00000349880

DATE: 30-03-22

WEBLEM 9c
Genome Data Viewer
(URL: https://www.ncbi.nlm.nih.gov/genome/gdv/)

AIM:

To explore graphical displays of features on NCBI's assembly of human genomic sequence data as well as cytogenetic, genetic, physical, and radiation hybrid maps using Genome Data Viewer (GDV).

INTRODUCTION:

GDV is composed of an **embedded instance of SV** that displays **sequence and track data**, along with **additional page elements** that allow a user to **search within an entire genome** assembly and efficiently **narrow in on their chromosome**, **sequence**, **region**, or **gene of interest**. GDV replaced the NCBI Map Viewer, **NCBI's previous** tool for whole-genome display. Researchers using GDV **can go directly** to the **NCBI BLAST** service from the browser and **load BLAST results** as alignment tracks that can be viewed **side by side** with **gene annotation** and other data. **Variation Viewer**, a related browser **associated with NCBI's** variation resources, is **functionally similar to GDV** and also **incorporates an instance of SV** but is **configured with features specifically intended for analyzing human variation data**. GDV and Variation Viewer can both **display the same types of NCBI** variation track data.

METHODOLOGY:

1. Open homepage for GDV genome browser (URL: https://www.ncbi.nlm.nih.gov/genome/gdv/)
2. Select human genome assembly
3. Search for DNA repair in genome
4. Select BRCA1 gene
5. Observe the results
6. Use various configuration options
7. Interpret the results.

OBSERVATION:

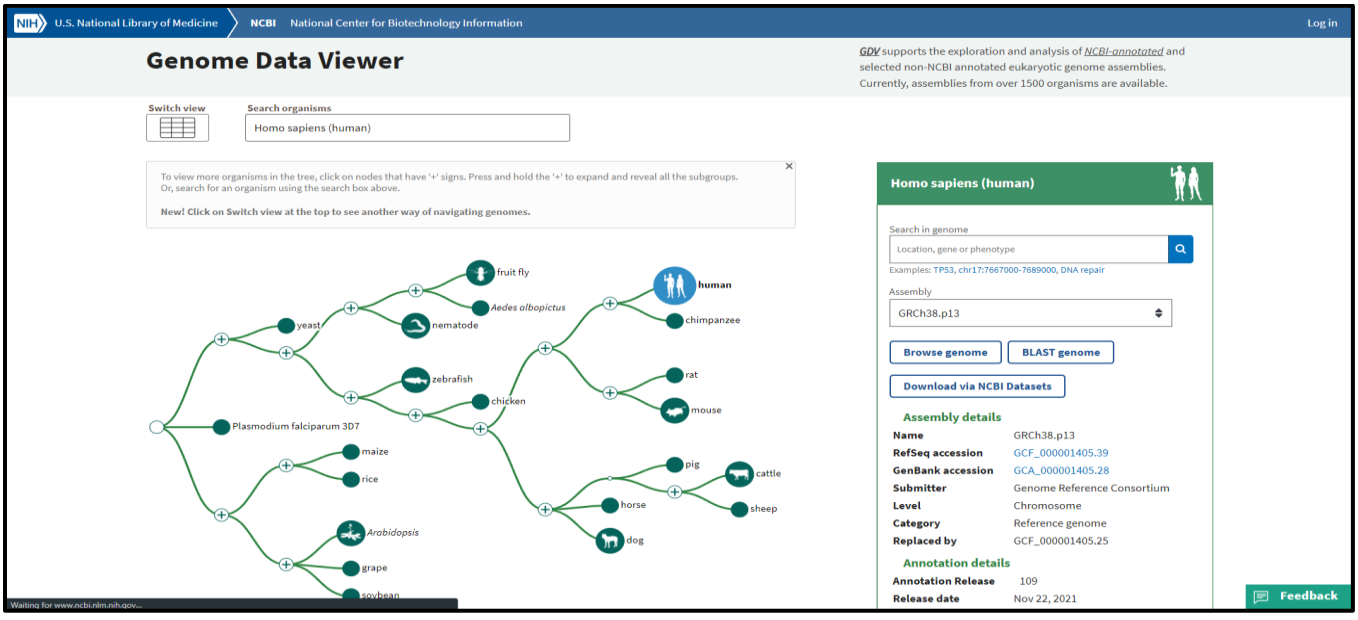


Fig1. Homepage for Genome Data Viewer

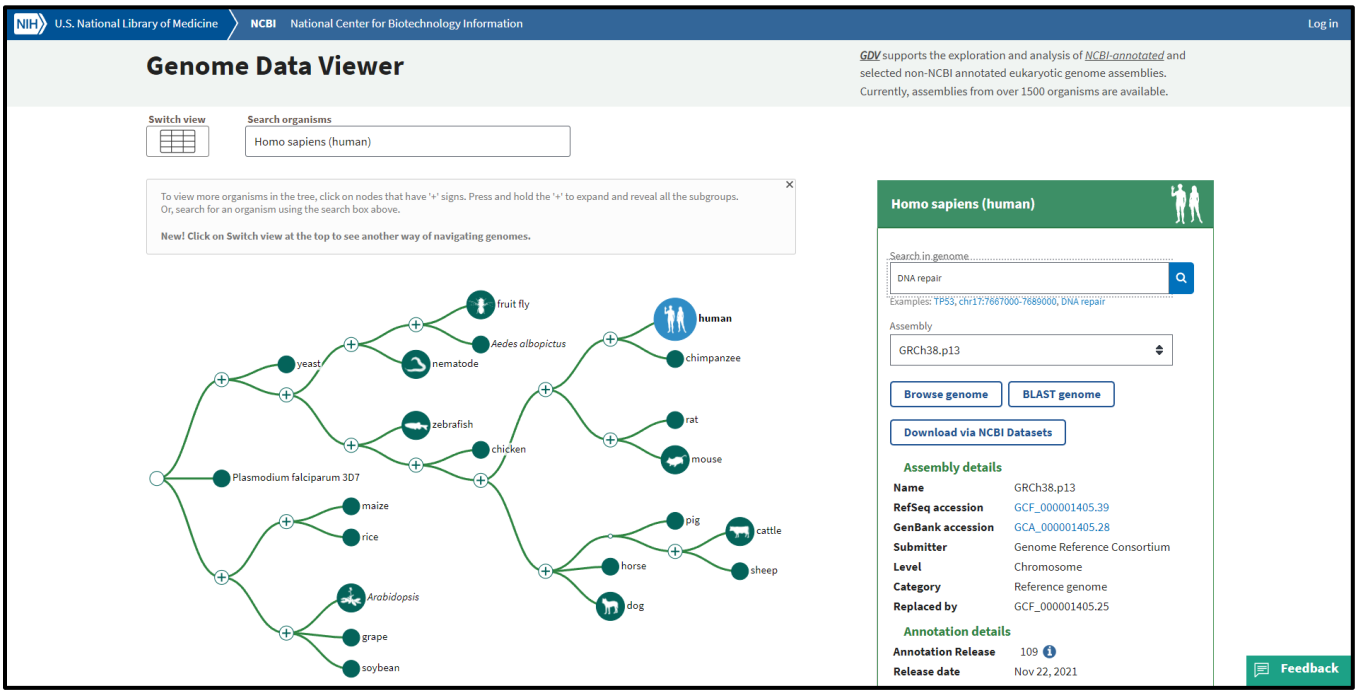


Fig2. Search for DNA repair in GRCh38.p13 assembly

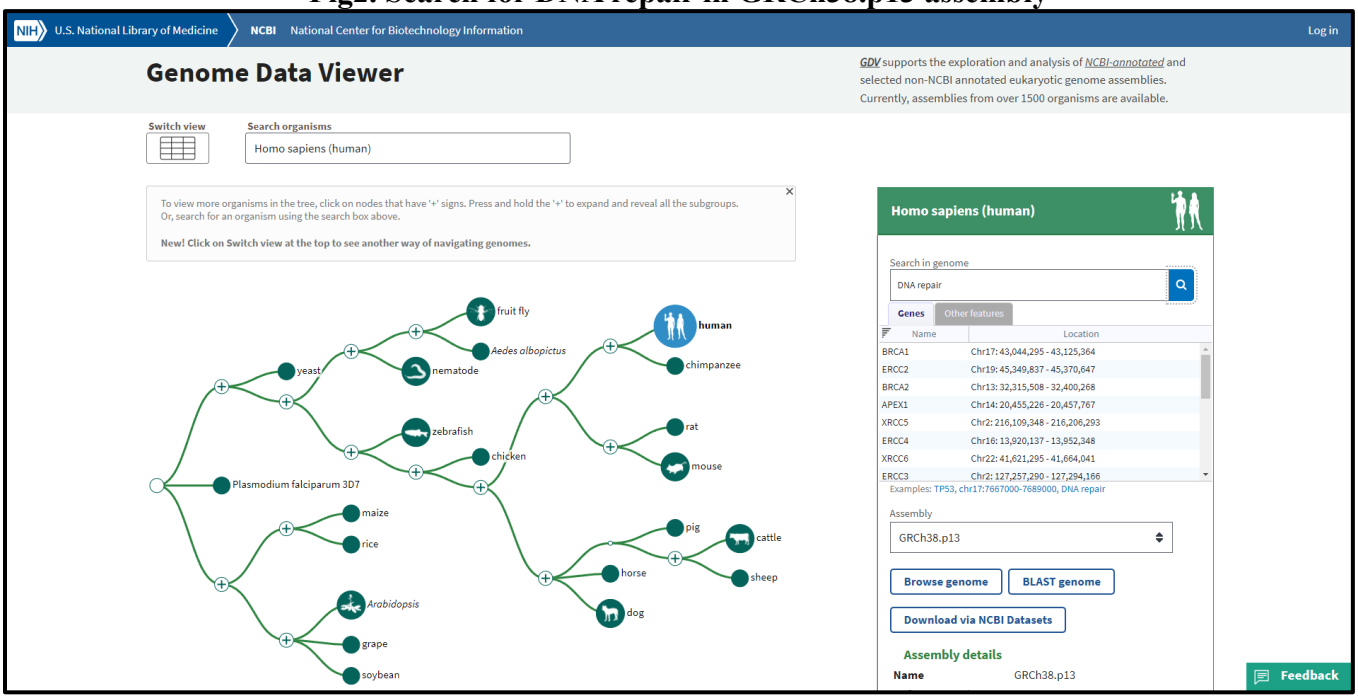


Fig3. Results for DNA repair genes

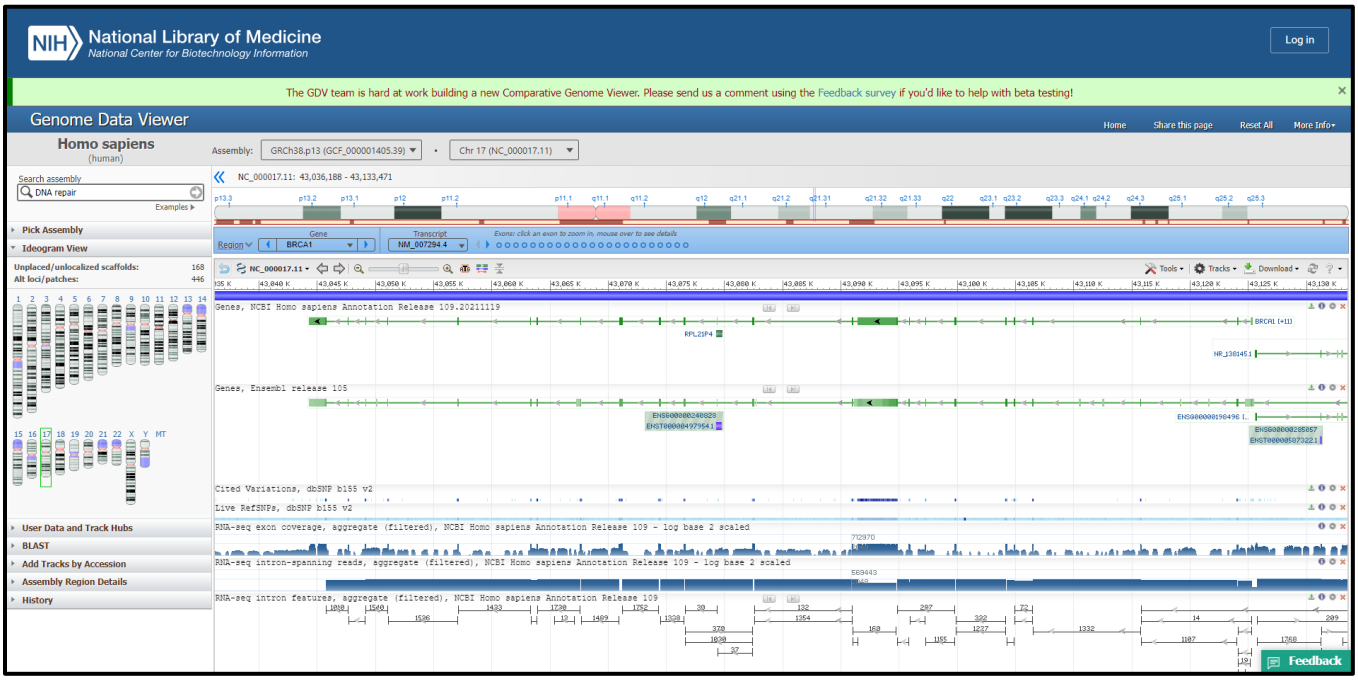


Fig4. Result for BRCA1 gene

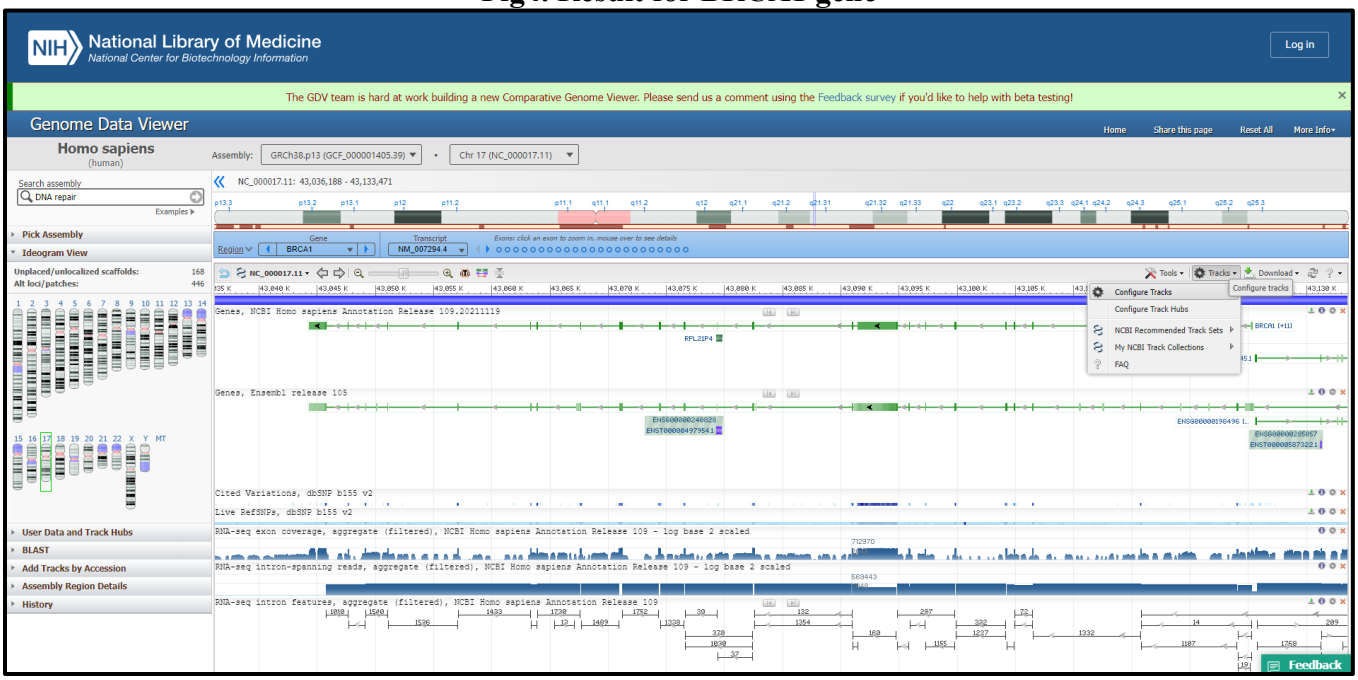


Fig5. Steps to configure tracks

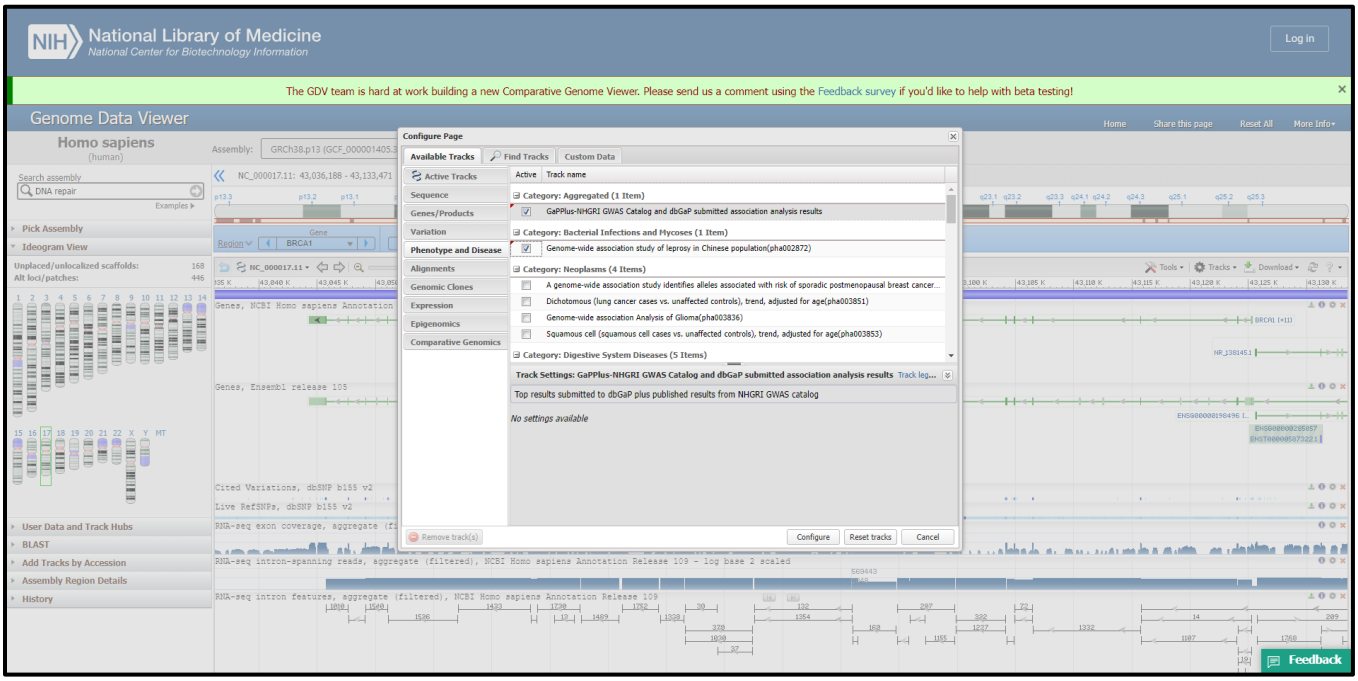


Fig6. Configuration page for phenotype and disease

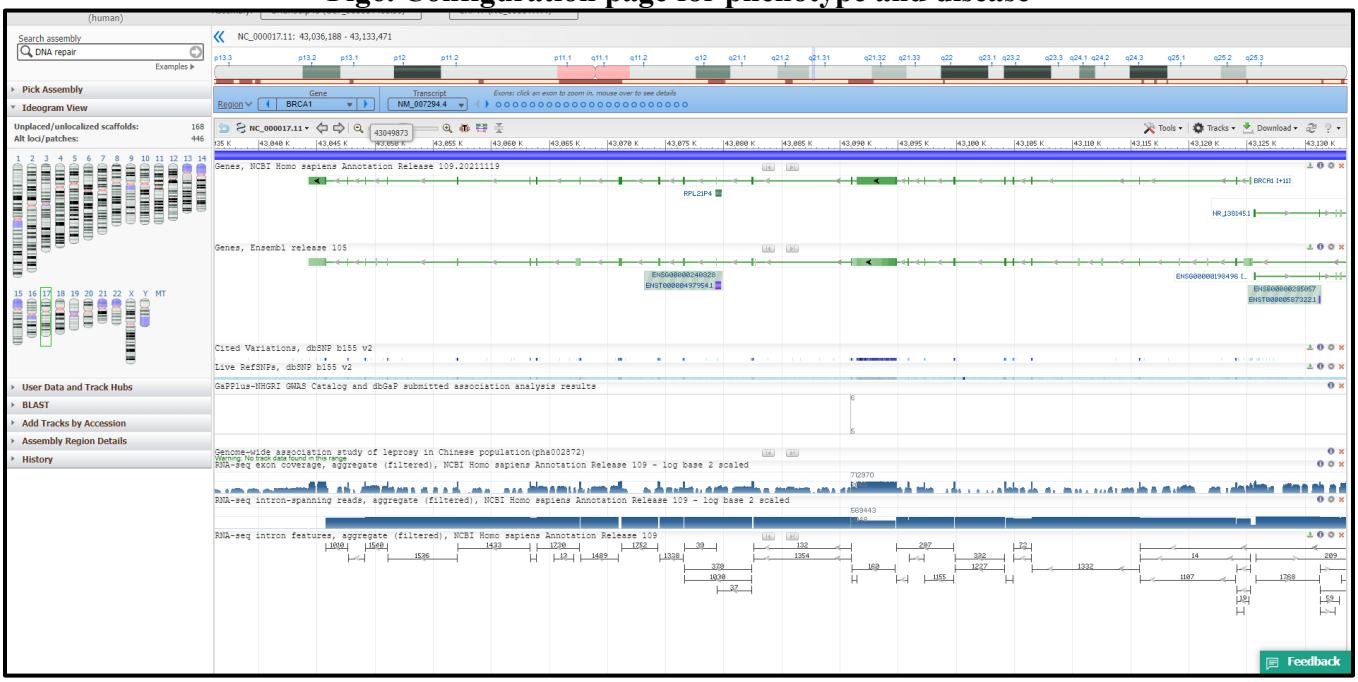


Fig7. Updated result after configuration

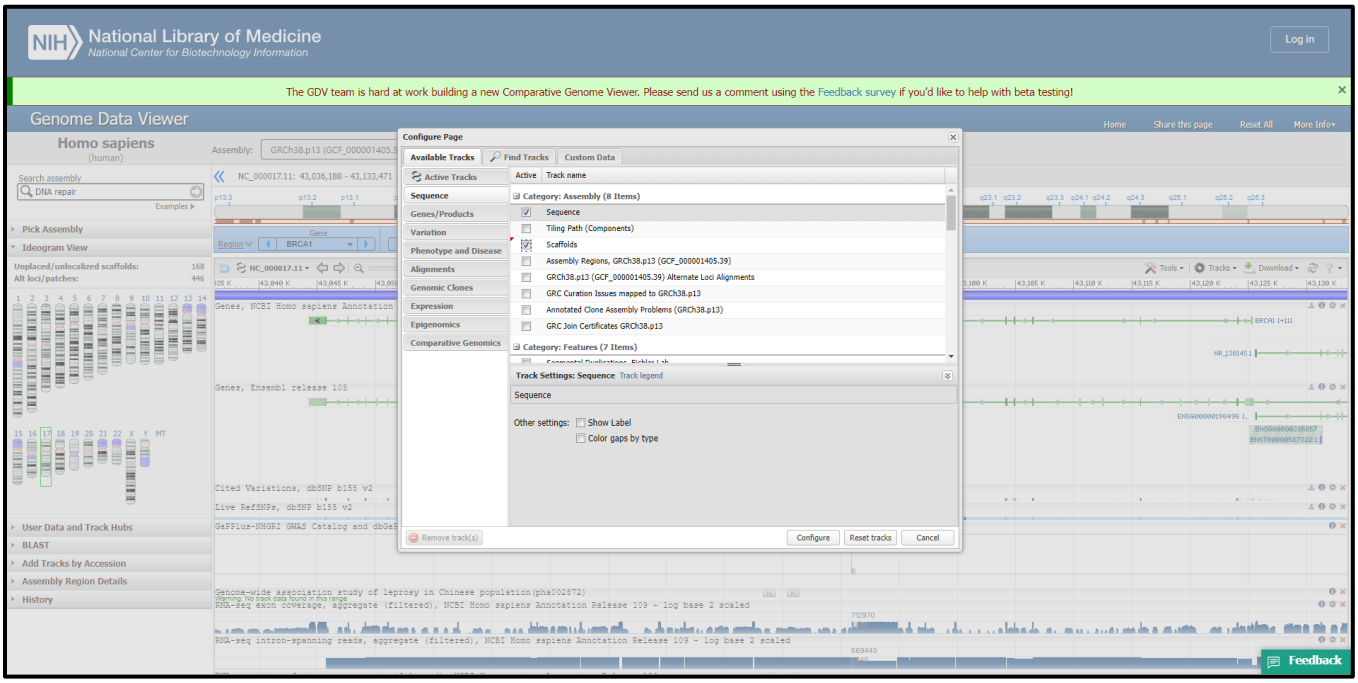


Fig8. Configuration page for sequence

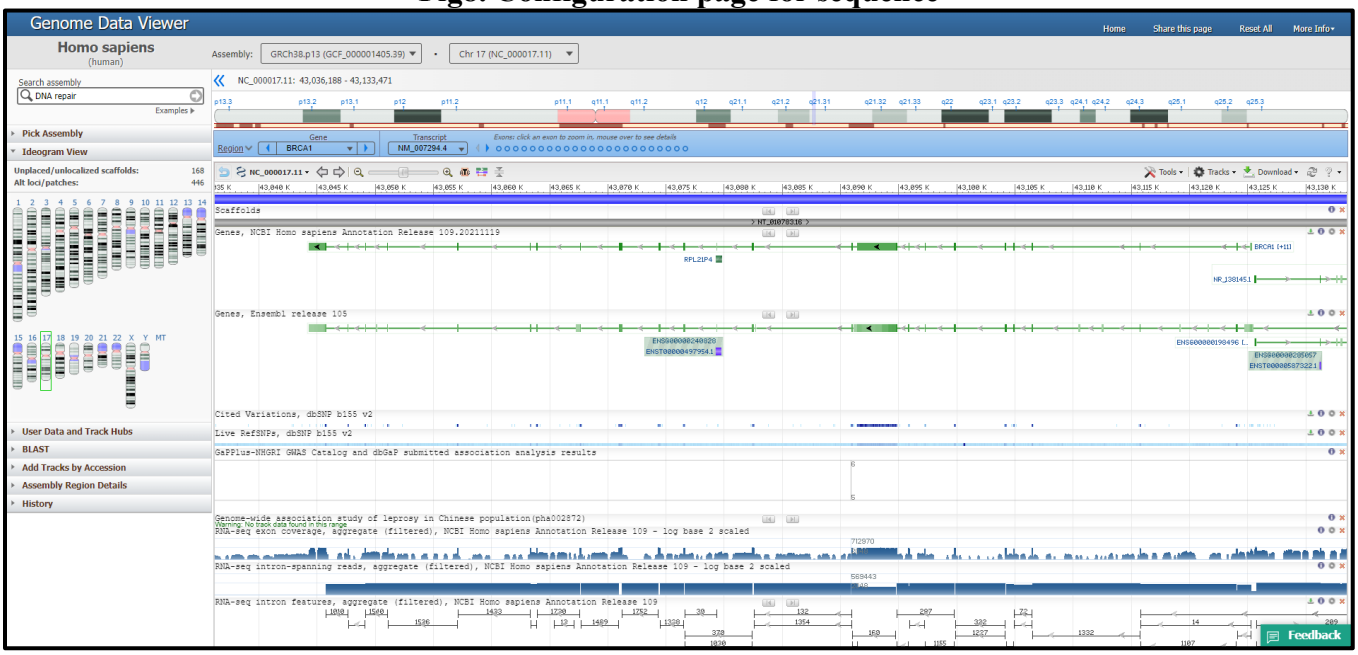


Fig9. Updated result after configuration

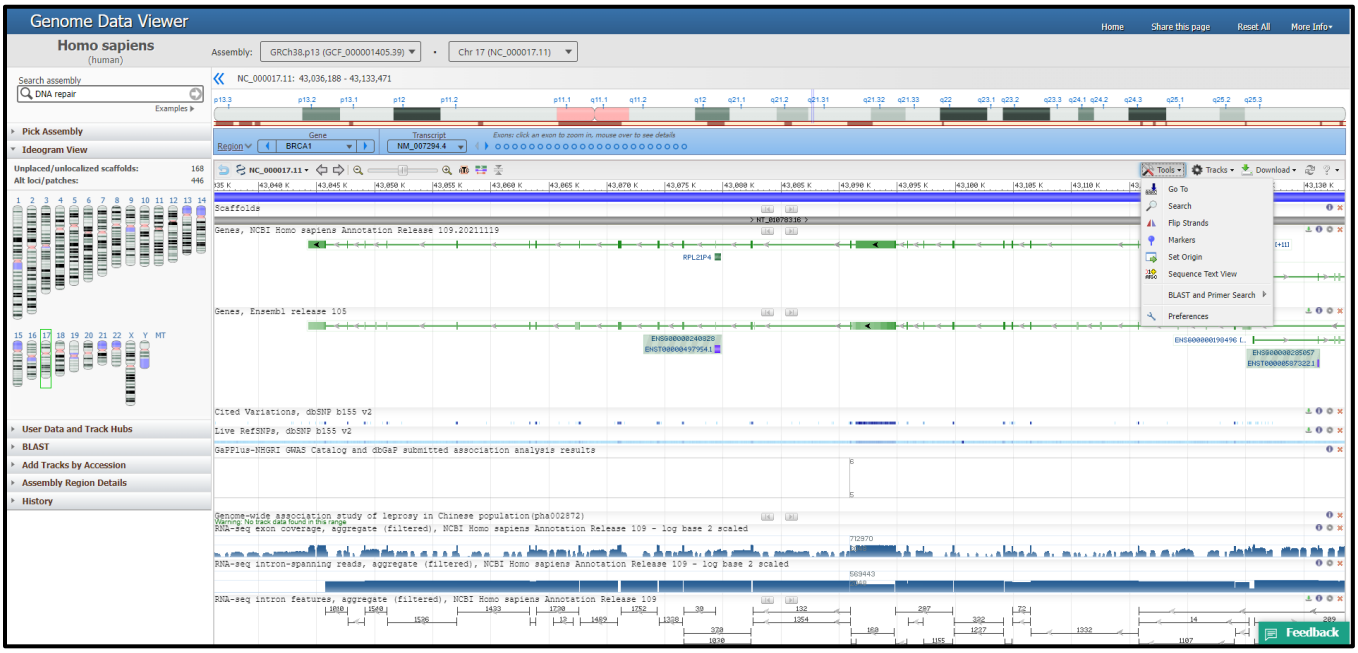


Fig10. Steps for sequence text view

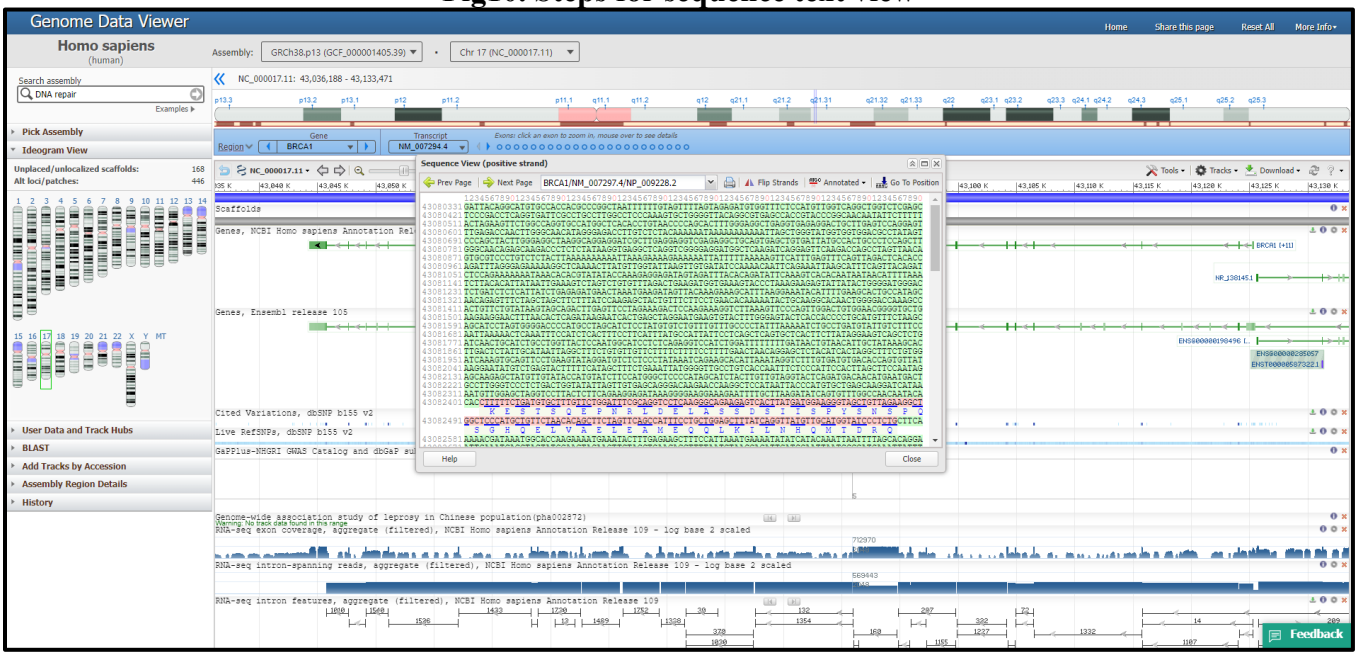


Fig11. Result for sequence text view

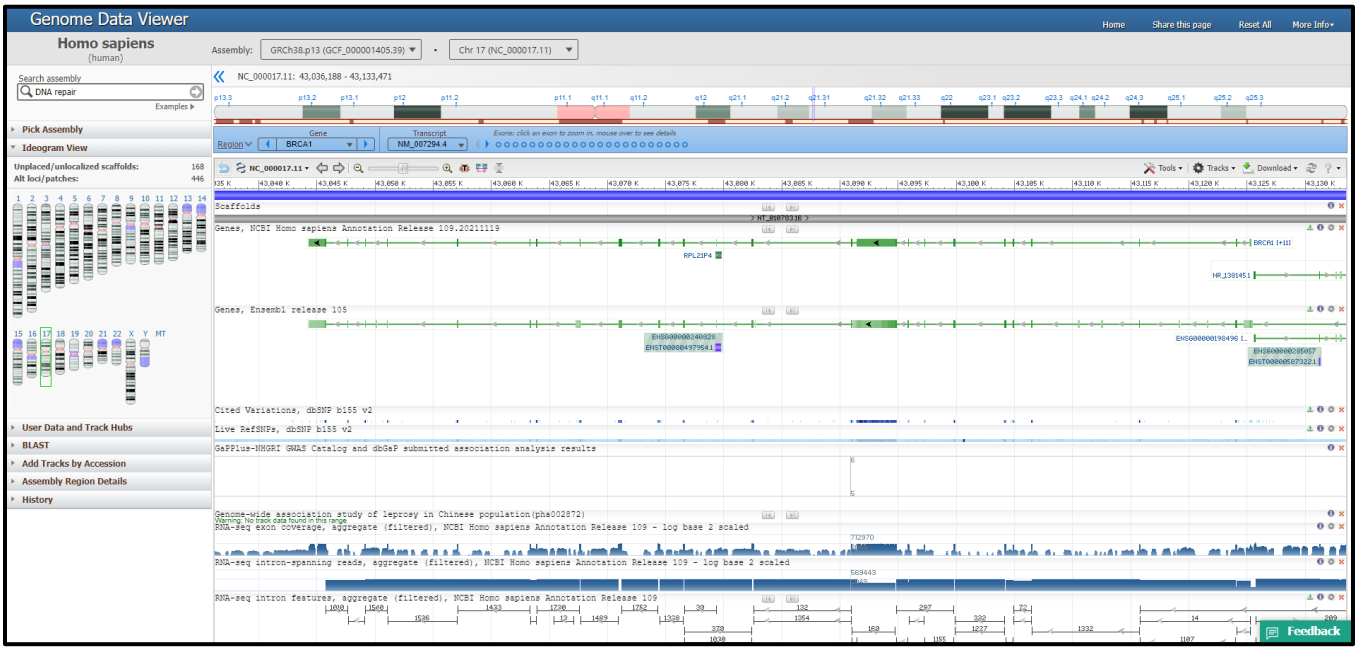


Fig12. Options to view exon information

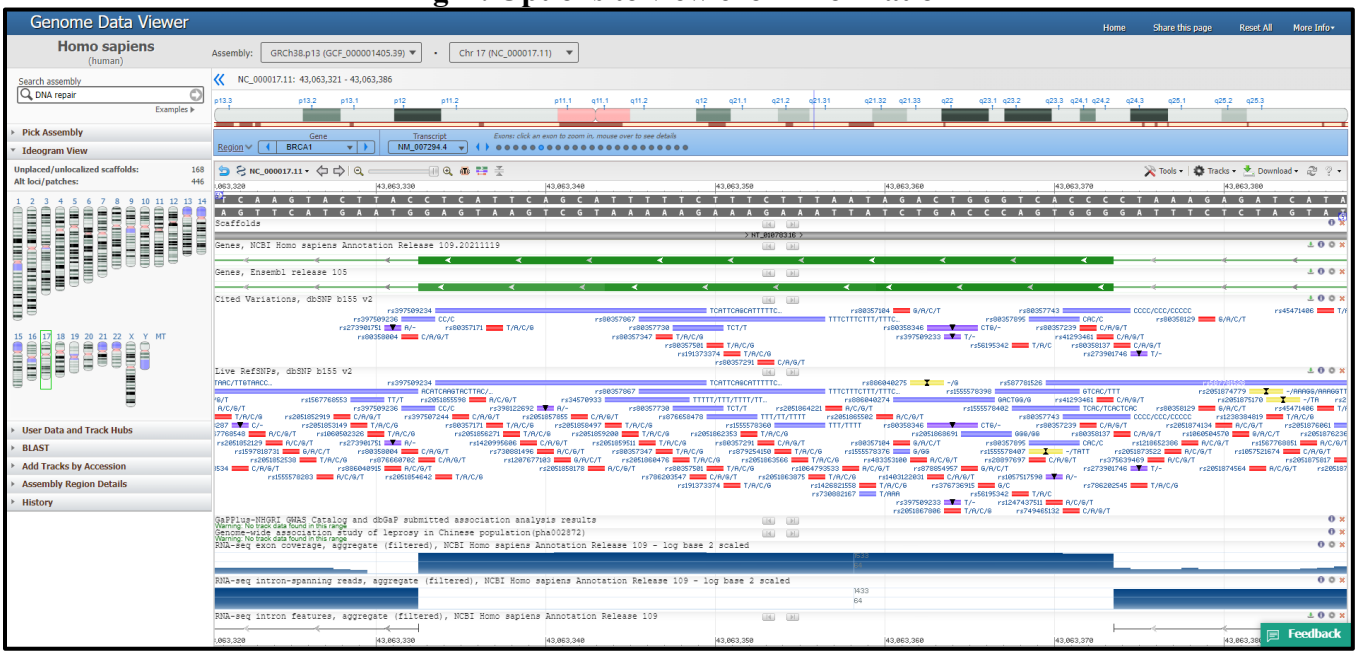


Fig13. Result for exon 18

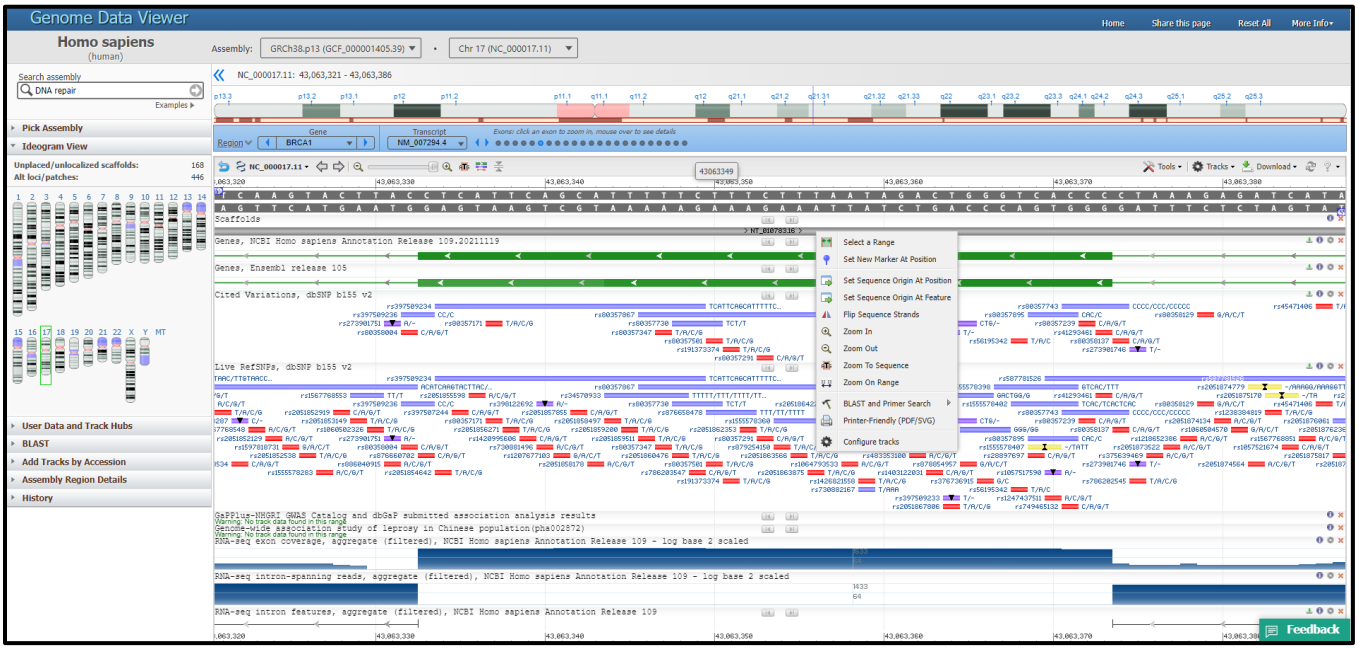


Fig14. Drag and select option for configuring tracks

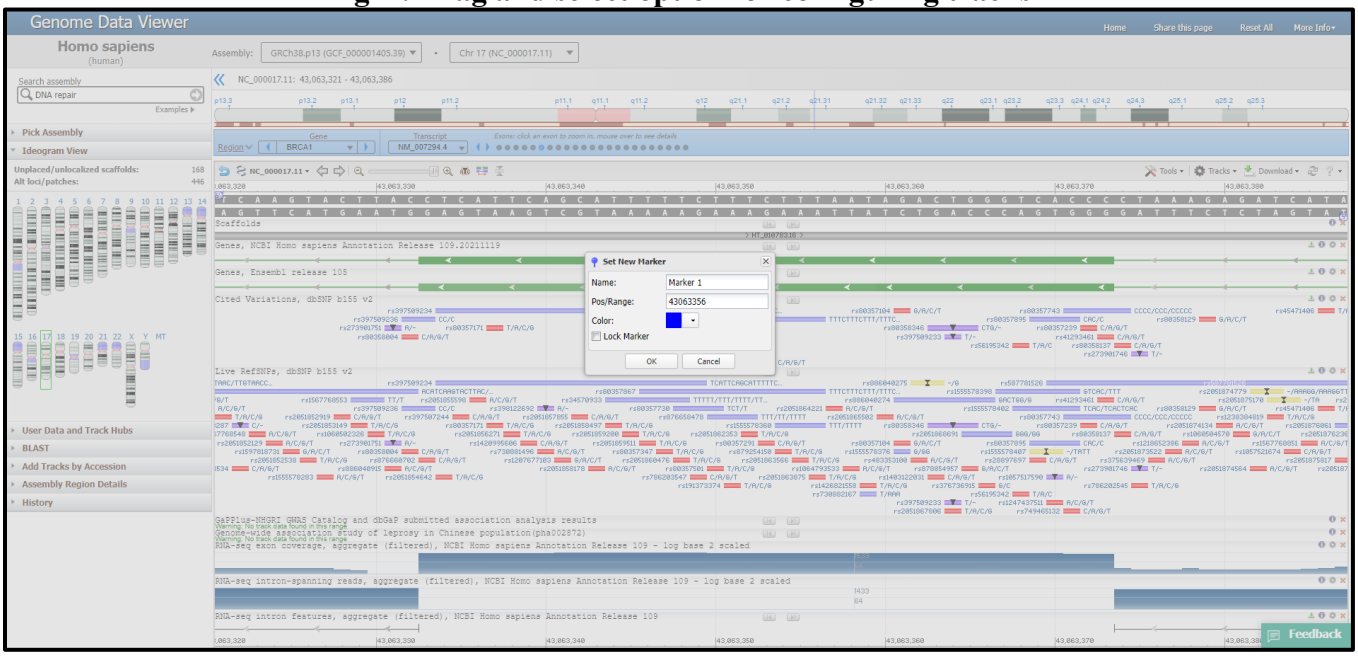


Fig15. Option to set new marker

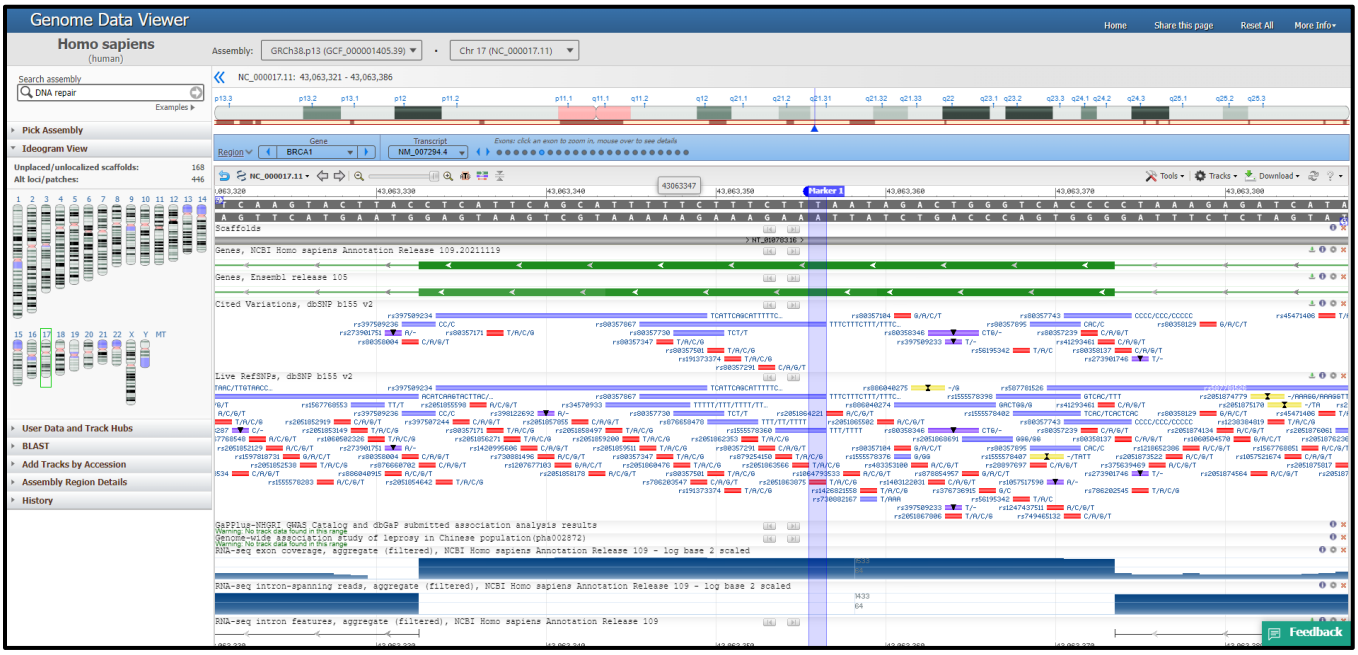


Fig16. Result for new marker

RESULT:

GDV genome browser was used to search for DNA repair under human genome assembly and results were observed for BRCA1 gene. Various options for tracks configuration were explored and information regarding sequence and exons were also viewed.

CONCLUSION:

GDV can be used for visualization and analysis of the wide range of genomes and assemblies annotated at the NCBI. RefSeq gene annotation data tracks are shown by default in the graphical view for these assemblies. NCBI ref SNP data tracks are also shown by default for human assemblies. GDV offers users the ability to customize the displays of individual tracks. Users can hide or configure tracks from the track configuration panel or by using the icons at the right end of each track.

REFERENCES:

1. Rangwala, S. H., Kuznetsov, A., Ananiev, V., Asztalos, A., Borodin, E., Evgeniev, V., Joukov, V., Lotov, V., Pannu, R., Rudnev, D., Shkeda, A., Weitz, E. M., & Schneider, V. A. (2020). Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Research, gr.266932.120. https://doi.org/10.1101/gr.266932.120
2. NCBI Genome Data Viewer. (n.d.). Www.ncbi.nlm.nih.gov. Retrieved March 28, 2022, from https://www.ncbi.nlm.nih.gov/genome/gdv/
3. Genome Data Viewer. (n.d.). Www.ncbi.nlm.nih.gov. Retrieved March 28, 2022, from https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39

DATE: 31-03-22

WEBLEM 10

A field guide to whole-genome sequencing, assembly and annotation

Introduction:

Genome sequencing projects were long confined to biomedical model organisms and required the concerted effort of large consortia. Rapid progress in high-throughput sequencing technology and the simultaneous development of bioinformatic tools have democratized the field. It is now within reach for individual research groups in the eco-evolutionary and conservation community to generate de novo draft genome sequences for any organism of choice. Because of the cost and considerable effort involved in such an endeavour, the important first step is to thoroughly consider whether a genome sequence is necessary for addressing the biological question at hand. Once this decision is taken, a genome project requires careful planning with respect to the organism involved and the intended quality of the genome draft.

The Steps involved in this are Genome Sequencing, Genome Assembly and Genome annotation. We will discuss these in detail ahead.

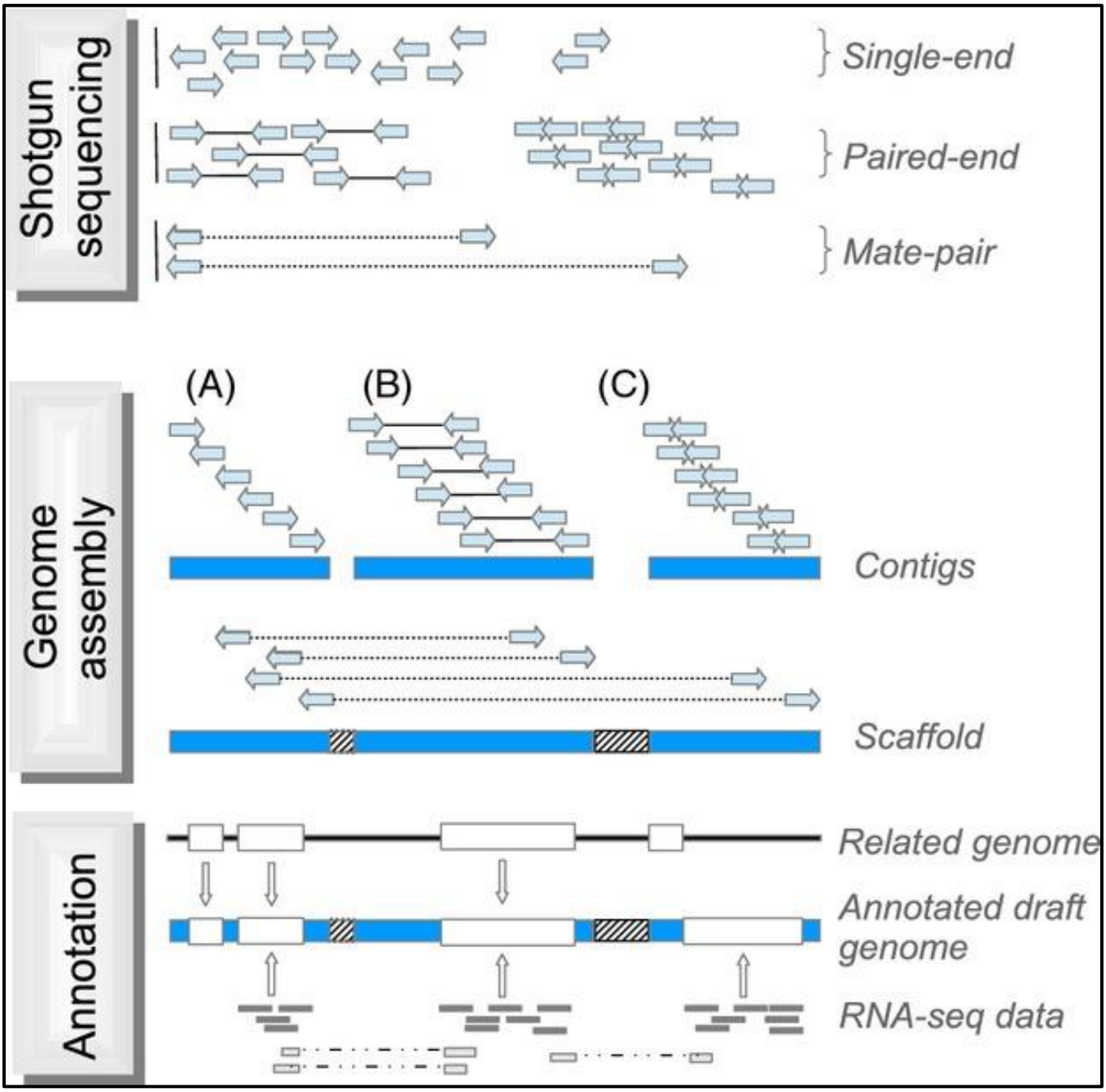


Fig1. Simplified illustration of the assembly process

GENOME SEQUENCING:

- The highest resolution genome map is the genomic DNA sequence that can be considered as a type of physical map describing a genome at the single base-pair level.
- DNA sequencing is now routinely carried out using the Sanger method. This involves the use of DNA polymerases to synthesize DNA chains of varying lengths.
- There are two major strategies for whole genome sequencing:
    - Shotgun approach
        - The shotgun approach randomly sequences clones from both ends of cloned DNA. This approach generates a large number of sequenced DNA fragments.

        - The number of random fragments has to be very large, so large that the DNA fragments overlap sufficiently to cover the entire genome.
        - This approach does not require knowledge of physical mapping of the clone fragments, but rather a robust computer assembly program to join the pieces of random fragments into a single, whole-genome sequence.
        - Generally, the genome has to be redundantly sequenced in such a way that the overall length of the fragments covers the entire genome multiple times.
        - This is designed to minimize sequencing errors and ensure correct assembly of a contiguous sequence.
        - Despite the multiple coverage, sometimes certain genomic regions remain un-sequenced, mainly owing to cloning difficulties.
        - In such cases, the remainder gap sequences can be obtained through extending sequences from regions of known genomic sequences using a more traditional PCR technique, which requires the use of custom primers and performs genome walking in a stepwise fashion.
    - Hierarchical approach
        - The hierarchical genome sequencing approach is similar to the shotgun approach, but on a smaller scale.
        - The chromosomes are initially mapped using the physical mapping strategy. Longer fragments of genomic DNA (100 to 300 kB) are obtained and cloned into a high-capacity bacterial vector called bacterial artificial chromosome (BAC).
        - Based on the results of physical mapping, the locations and orders of the BAC clones on a chromosome can be determined.
        - By successively sequencing adjacent BAC clone fragments, the entire genome can be covered. The complete sequence of each individual BAC clone can be obtained using the shotgun approach.
        - Overlapping BAC clones are subsequently assembled into an entire genome sequence.

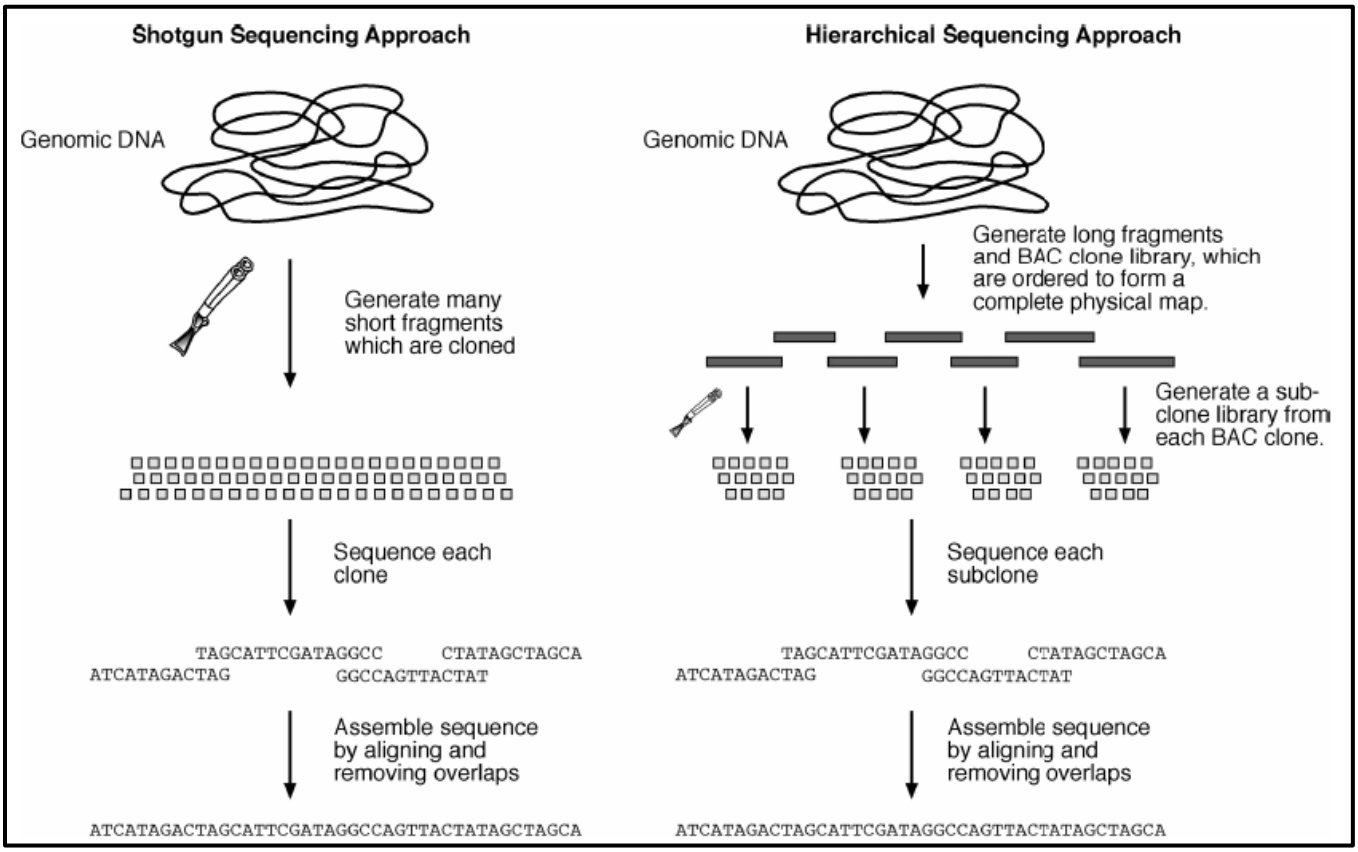


Fig2. Differences between Shotgun sequencing and hierarchical sequencing

GENOME ASSEMBLY:

- As described, initial DNA sequencing reactions generate short sequence reads from DNA clones. The average length of the reads is about 500 bases.
- To assemble a whole genome sequence, these short fragments are joined to form larger fragments after removing overlaps.
- These longer, merged sequences are termed contigs, which are usually 5,000 to 10,000 bases long. A number of overlapping contigs can be further merged to form scaffolds (30,000–50,000 bases, also called super contigs), which are unidirectionally oriented along a physical map of a chromosome.
- Overlapping scaffolds are then connected to create the final highest resolution map of the genome.
- Correct identification of overlaps and assembly of the sequence reads into contigs are like joining jigsaw puzzles, which can be very computationally intensive when dealing with data at the whole-genome level.
- The major challenges in genome assembly are sequence errors, contamination by bacterial vectors, and repetitive sequence regions. Sequence errors can often be corrected by drawing a consensus from an alignment of multiple overlapped sequences.
- Bacterial vector sequences can be removed using filtering programs prior to assembly. To overcome the problem of sequence repeats, programs such as RepeatMasker can be used to detect and mask repeats. Additional constraints on the sequence reads can be applied to avoid misassembly caused by repeat sequences.

- A commonly used constraint to avoid errors caused by sequence repeats is the so- called forward–reverse constraint.
- When a sequence is generated from both ends of a single clone, the distance between the two opposing fragments of a clone is fixed to a certain range, meaning that they are always separated by a distance defined by a clone length (normally 1,000 to 9,000 bases).
- When the constraint is applied, even when one of the fragments has a perfect match with a repetitive element outside the range, it is not able to be moved to that location to cause miss assembly.
- The first step toward genome assembly is to derive base calls and assign associated quality scores. The next step is to assemble the sequence reads into contiguous sequences.
- This step includes identifying overlaps between sequence fragments, assigning the order of the fragments and deriving a consensus of an overall sequence. Assembling all shotgun fragments into a full genome is a computationally very challenging step.
- There are a variety of programs available for processing the raw sequence data.
- The following is a selection of base calling and assembly programs commonly used in genome sequencing projects:
    - **Phred** (https://www.phrap.org/) is a UNIX program for base calling. It uses a Fourier analysis to resolve fluorescence traces and predict actual peak locations of bases
    - **Phrap** (https://www.phrap.org/) is a UNIX program for sequence assembly. It takes Phred base-call files with quality scores as input and aligns individual fragments in a pairwise fashion using the Smith–Waterman algorithm.
    - **VecScreen** (https://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) is a web-based program that helps detect contaminating bacterial vector sequences.
    - **TIGR Assembler** (https://www.tigr.org/) is a UNIX program from TIGR for assembly of large shotgun sequence fragments. It treats the sequence input as clean reads without consideration of the sequence quality.
    - **ARACHNE** (https://www.genome.wi.mit.edu/wga/) is a free UNIX program for the assembly of whole-genome shotgun reads. Its unique features include using a heuristic approach similar to FASTA to align overlapping fragments, evaluating alignments using statistical scores, correcting sequencing errors based on multiple sequence alignment, and using forward–reverse constraints.
    - **EULER** (http://nbcr.sdsc.edu/euler/) is an assembly algorithm that uses a Eulerian Superpath approach, which is a polynomial algorithm for solving puzzles such as the famous "traveling salesman problem": finding the shortest path of visiting a given number of cities exactly once and returning to the starting point.

GENE ANNOTATION

- Before the assembled sequence is deposited into a database, it has to be analyzed for useful biological features. The genome annotation process provides comments for the features.
- This involves two steps:
    - Gene prediction.
    - Functional assignment.
- As a real-world example, gene annotation of the human genome employs a com- bination of theoretical prediction and experimental verification.
- Gene structures are first predicted by ab initio exon prediction programs such as GenScan or FgenesH.

- The predictions are verified by BLAST searches against a sequence database. The predicted genes are further compared with experimentally determined cDNA and EST sequences using the pairwise alignment programs such as GeneWise, Spidey, SIM4, and EST2Genome.
- All predictions are manually checked by human curators. Once open reading frames are determined, functional assignment of the encoded proteins is carried out by homology searching using BLAST searches against a protein database.
- Further functional descriptions are added by searching protein motif and domain databases such as Pfam and InterPro as well as by relying on published literature.

REFERENCES:

1. JB;, E. R. W. (n.d.). *A field guide to whole-genome sequencing, assembly and Annotation*. Evolutionary applications. Retrieved March 31, 2022, from https://pubmed.ncbi.nlm.nih.gov/25553065/
2. Xiong, J. (2008). *Essential bioinformatics*. Cambridge University Press.