

**NAME: - LIZA KANJI PATEL**

**CLASS: - M.Sc. PART - I**

**COURSE: - BIOINFORMATICS**

**ACADEMIC YEAR: - 2021-2022**

**ROLL NO.: - 113**

**PAPER CODE: - GNKPSB|202**

**COURSE TITLE: - STRUCTURAL BIOLOGY & BIOINFORMATICS**

**GURU NANAK KHALSA COLLEGE**

**MATUNGA, MUMBAI-400 019.**

**DEPARTMENT OF BIOINFORMATICS**

**CERTIFICATE**

This is to certify that **Ms. Liza Kanji Patel (Roll.No.113)** of M.Sc. Part I Bioinformatics has satisfactorily completed the practical Semester II course prescribed by the University of Mumbai during the academic year 2021-2022.

**TEACHER INCHARGE**

**HEAD OF DEPARTMENT**

**INDEX:**

SR.NO.	EXPERIMENTS	PAGE NO.	DATE	SIGN
1.	Introduction to secondary structure prediction	1-3	11/02/22	
1a.	To predict secondary structure of Hemoglobin protein using various tools	4-13	11/02/22	
2.	Introduction to protein classification	14-16	16/02/22	
2a.	To study the structural classification of protein Insulin using CATH and SCOP database	17-32	16/02/22	
3.	Introduction to tertiary structure prediction	33-37	23/02/22	
3a.	To perform tertiary structure prediction by Comparative Modeling/Homology Modeling method using Modeller for query Kinase	38-51	08/03/22	
3b.	To perform tertiary structure prediction by Threading approach using I-TASSER server for query Kinase	52-58	08/03/22	
3c.	To perform tertiary structure prediction by Ab-Initio approach using ROBETTA server for query Kinase	59-63	08/03/22	
4.	Introduction to Validation server- SAVES server	64-66	08/03/22	
4a.	To validate structure kseq.B99990003 generated from Modeller	67-72	08/03/22	
4b.	To validate structure model1 generated from I-TASSAR server	73-78	08/03/22	
4c.	To validate structure 228776_1 generated from Robetta server	79-84	08/03/22	
5.	Introduction to Visualization of Tertiary structure using RASMOL & PyMOL	85-94	28/02/22	
5a.	To visualize 3D structure of Fibrin using RASMOL & PyMOL tool	95-106	28/02/22	
6.	Introduction to binding pocket prediction of protein with respect to PTM studies	107-110	25/02/22	
6a.	To predict binding pocket of protein Thrombin using CASTp server	111-117	25/02/22	
6b.	To predict binding pocket for Glycosylation sites in Thrombin using NetNGlyc 4.0 Server	118-121	25/02/22	
6c.	To predict binding pocket for Glycosylation sites in Thrombin using NetPhos 3.1 Server	122-126	25/02/22	
7.	Introduction to Structural Blast-VAST & DALI	127-131	14/03/22	
7a.	To perform structural Blast for Albumin using VAST tool	132-139	14/03/22	
7b.	To perform structural Blast for Albumin using DALI tool	140-149	14/03/22	
8.	Introduction to Gene Prediction and various elements in Prokaryotes and Eukaryotes	150-154	18/03/22	

8a.	To recognize the Protease human Pol III promoter region and start of transcription using TSSW tool	155-159	18/03/22	
8b.	To predict bacterial promoter for <i>Neisseria gonorrhoeae</i> using BPROM tool	160-165	18/03/22	
8c.	To predict bacterial operon and gene for <i>Neisseria gonorrhoeae</i> using FGENESB tool	166-170	18/03/22	
8d.	To predict exon signals in Protease using FGENES tool.	171-175	18/03/22	
8e.	To search for ORF region in <i>Neisseria gonorrhoeae</i> using ORF finder tool	176-179	18/03/22	
9.	Introduction to Genomics & its various browser (UCSC, ENSEMBL, GDV)	180-183	28/03/22	
9a.	To explore UCSC genome browser in order to understand the gene, its related studies & protein level information	184-202	28/03/22	
9b.	To explore Ensembl genome browser in order to gather information for annotated genes/genome/protein/transcript etc	203-210	28/03/22	
9c.	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)	211-220	28/03/22	
10.	A field guide to whole-genome sequencing, assembly and annotation	221-224	30/03/22	

## WEBLEM 1

### Introduction to secondary structure prediction

Proteins perform most essential biological and chemical functions in a cell. They play important roles in structural, enzymatic, transport, and regulatory functions. The protein functions are strictly determined by their structures. Therefore, protein structural bioinformatics is an essential element of bioinformatics. Protein three-dimensional structures are obtained using two popular experimental techniques, x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.

Protein secondary structures are stable local conformations of a polypeptide chain. They are critically important in maintaining a protein three-dimensional structure. The highly regular and repeated structural elements include  $\alpha$ -helices and  $\beta$ -sheets. It has been estimated that nearly 50% of residues of a protein fold into either  $\alpha$ -helices and  $\beta$ -strands. As a review, an  $\alpha$ -helix is a spiral-like structure with 3.6 amino acid residues per turn. The structure is stabilized by hydrogen bonds between residues  $I$  and  $i+4$ . Prolines normally do not occur in the middle of helical segments, but can be found at the end positions of  $\alpha$ -helices. A  $\beta$ -sheet consists of two or more  $\beta$ -strands having an extended zigzag conformation. The structure is stabilized by hydrogen bonding between residues of adjacent strands, which actually may be long-range interactions at the primary structure level.  $\beta$ -Strands at the protein surface show an alternating pattern of hydrophobic and hydrophilic residues; buried strands tend to contain mainly hydrophobic residues.

Protein secondary structure prediction refers to the prediction of the conformational state of each amino acid residue of a protein sequence as one of the three possible states, namely, helices, strands, or coils, denoted as H, E, and C, respectively. The prediction is based on the fact that secondary structures have a regular arrangement of amino acids, stabilized by hydrogen bonding patterns. The structural regularity serves the foundation for prediction algorithms.

Predicting protein secondary structures has a number of applications. It can be useful for the classification of proteins and for the separation of protein domains and functional motifs. Secondary structures are much more conserved than sequences during evolution. As a result, correctly identifying secondary structure elements (SSE) can help to guide sequence alignment or improve existing sequence alignment of distantly related sequences. In addition, secondary structure prediction is an intermediate step in tertiary structure prediction as in threading analysis

Efficient automatic methods for protein structure predictions are becoming increasingly important as a result of the influx of genomic data arising from sequencing projects. Methods for predicting topologies of both globular and membrane proteins have been publically available. Knowledge of protein three-dimensional (3D) structures is of major importance in providing insights into their molecular functions. Analysis of 3D structures assists identification of binding sites and thus facilitates design of new drugs. Also, structural investigations at atomic resolution help in conceiving how a single point mutation in a gene, and the possible subsequent amino acid substitution at the protein level can, for instance, lead to protein deficiency and disease. Moreover, biochemical methods to probe protein functions, or to enhance/inhibit some reactions, and the entire protein engineering field benefit tremendously when scientists have access to structural information prior to initiating research projects. Such data indeed tend to simplify the design of experiments and the rational processing of experimental results. Clearly, sophisticated methods used to compare genes and/or amino acid sequences can also provide significant amount of information about molecular functions, whether applied alone or in the context of a three-dimensional representative of a family member and facilitate rational engineering.

X-ray crystallography, NMR and biochemical methods have been essential to start developing concepts about how amino acid sequences relate to folding and function. Such experimental data have helped in the design of computer methods facilitating prediction and analysis of protein structures.

Bioinformatics encompasses tools and concepts aiming at the analysis of gene expression (e.g., cDNA microarrays immobilized on slides or gene chips) as well as approaches leading to the prediction of protein structures and functions. The term structural bioinformatics has been coined recently, and commonly refers to research studies dealing with structural predictions, simulations, sequence and structure analyses, family classification, and genome annotation. The field of structural bioinformatics and related concepts provide not only a means of organizing one's thinking about sequence-structure-function problems, but also a framework for predicting unobserved behaviour and suggesting novel experiments.

Secondary structure prediction methods are a set of techniques in bioinformatics that aim to study the local secondary structures of protein using the knowledge of their amino acid sequences. Algorithms used for secondary structure prediction are:

#### **Ab Initio-Based Methods:**

This type of method predicts the secondary structure based on a single query sequence. It measures the relative propensity of each amino acid belonging to a certain secondary structure element. The propensity scores are derived from known crystal structures.

#### **Homology-Based Methods:**

The third generation of algorithms were developed in the late 1990s by making use of evolutionary information. This type of method combines the ab initio secondary structure prediction of individual sequences and alignment information from multiple homologous sequences (>35% identity). The idea behind this approach is that closely related protein homologs should adopt the same secondary and tertiary structure. By aligning multiple sequences, information of positional conservation is revealed. Because residues in the same aligned position are assumed to have the same secondary structure, any inconsistencies or errors in prediction of individual sequences can be corrected using a majority rule. This homology based method has helped improve the prediction accuracy by another 10% over the second-generation methods.

#### **Neural Networks:**

The third-generation prediction algorithms also extensively apply sophisticated neural networks to analyze substitution patterns in multiple sequence alignments. As a review, a *neural network* is a machine learning process that requires a structure of multiple layers of interconnected variables or nodes. In secondary structure prediction, the input is an amino acid sequence and the output is the probability of a residue to adopt a particular structure. Between input and output are many connected hidden layers where the machine learning takes place to adjust the mathematical weights of internal connections. The neural network has to be first trained by sequences with known structures so it can recognize the amino acid patterns and their relationships with known structures. During this process, the weight functions in hidden layers are optimized so they can relate input to output correctly. When the sufficiently trained network processes an unknown sequence, it applies the rules learned in training to recognize particular structural patterns.

When multiple sequence alignments and neural networks are combined, the result is further improved accuracy. In this situation, a neural network is trained not by a single sequence but by a sequence profile derived from the multiple sequence alignment. This combined approach has been shown to improve the accuracy to above 75%, which is a breakthrough in secondary structure prediction. The improvement mainly comes from enhanced secondary structure signals through consensus drawing.

The three generations of secondary structure prediction are:

- First generation (Chou & Fasman Secondary Structure Prediction Server)
- Second generation (GOR IV)
- Third generation (PSIPRED)

**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 or directly from BioGem tools

**GOR:** The GOR method is based on information theory and was developed by J.Garnier, D.Osguthorpe and B.Robson. 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. 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.

**PSIPRED:** PSIPRED protein structure prediction server incorporates two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST. Using a stringent cross validation procedure to evaluate performance, PSIPRED has been shown to be capable of achieving an average Q3 score of 76.5%. This is the highest level of accuracy published for any method. A Java front-end application, PSIPREDView, has also been developed to interpret the output from PSIPRED and to provide users with publication quality graphical representations of secondary structure prediction. A hypertext link to the location of the images is included in the e-mail following the text representation of the prediction.

## REFERENCE:

1. Xiong, J. (2008). *Essential bioinformatics*. Cambridge: Cambridge University Press. 173-205.
2. McGuffin, L. J.; Bryson, K.; Jones, D. T. (2000). *The PSIPRED protein structure prediction server*. *Bioinformatics*, 16(4), 404–405. doi:10.1093/bioinformatics/16.4.404
3. Villoutreix, B. O. (2002). *Structural Bioinformatics: Methods, Concepts and Applications to Blood Coagulation Proteins*. *Current Protein and Peptide Science*, 3(3), 341–364. doi:10.2174/1389203023380657
4. Ashok Kumar, T. (2013). *CFSSP: Chou and Fasman Secondary Structure Prediction server*. WIDE SPECTRUM: Research Journal. 1(9):15-19.
5. *NPS@ REFERENCES*. (n.d.-b). Npsa-Prabi.ibcp.fr. Retrieved February 11, 2022, from [https://npsa-prabi.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_references.html](https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_references.html)
6. McGuffin, L. J.; Bryson, K.; Jones, D. T. (2000). *The PSIPRED protein structure prediction server*. *Bioinformatics*, 16(4), 404–405. doi:10.1093/bioinformatics/16.4.404

## WEBLEM 1a

### Secondary structure prediction using various tools

(URL: <http://www.biogem.org/tool/chou-fasman/index.php>

[https://npsa-prabi.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_gor4.html](https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html)

<http://bioinf.cs.ucl.ac.uk/psipred/> )

### AIM:

To predict secondary structure of Hemoglobin protein using various tools.

### INTRODUCTION:

Hemoglobin, also spelled haemoglobin, iron-containing protein in the blood of many animals in the red blood cells (erythrocytes) of vertebrates that transports oxygen to the tissues. Hemoglobin forms an unstable reversible bond with oxygen. In the oxygenated state, it is called oxyhemoglobin and is bright red; in the reduced state, it is purplish blue. The secondary structure prediction of hemoglobin protein can be done using various tools.

Secondary structure prediction methods are a set of techniques in bioinformatics that aim to study the local secondary structures of protein using the knowledge of their amino acid sequences. Ab initio, neural networks and homology modelling are some of the algorithms used for these methods.

The three generations of secondary structure prediction are:

- First generation (Chou & Fasman Secondary Structure Prediction Server)
- Second generation (GOR IV)
- Third generation (PSIPRED)

**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 or directly from BioGem tools

**GOR:** The GOR method is based on information theory and was developed by J.Garnier, D.Osguthorpe and B.Robson. 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. 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.

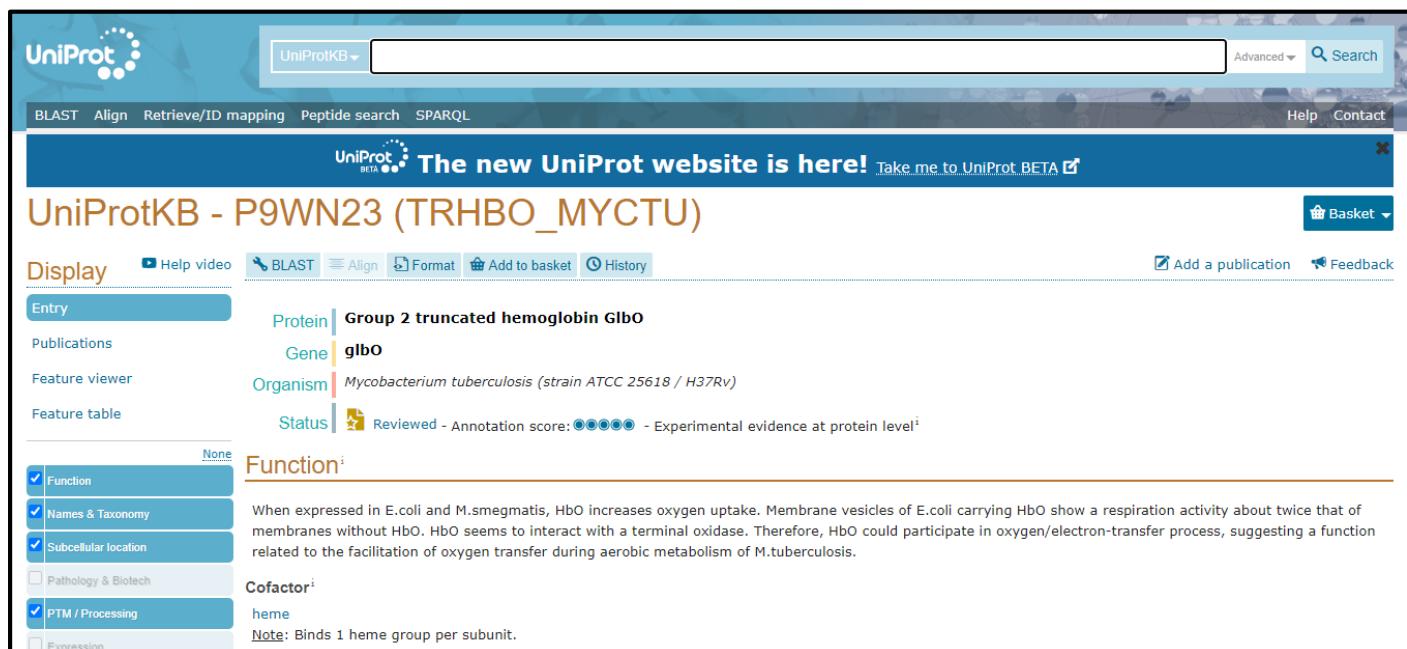
**PSIPRED:** PSIPRED protein structure prediction server incorporates two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST. Using a stringent cross validation procedure to evaluate performance, PSIPRED has been shown to be capable of achieving an average Q3 score of 76.5%. This is the highest level of accuracy published for any method. A Java front-end application, PSIPREDView, has also been developed to interpret the output from PSIPRED and to provide users with publication quality

graphical representations of secondary structure prediction. A hypertext link to the location of the images is included in the e-mail following the text representation of the prediction.

## METHODOLOGY:

1. Retrieve Hemoglobin protein sequence from UniProt database.
2. Use the protein FASTA sequence for secondary structure prediction in Chou & Fasman Secondary Structure Prediction Server. (URL: <http://www.biogem.org/tool/chou-fasman/index.php>)
3. Use protein sequence for secondary structure prediction in GOR IV and PSIPRED Secondary Structure Prediction Server. (URL: [https://npsa-prabi.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_gor4.html](https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html) and <http://bioinf.cs.ucl.ac.uk/psipred/>)
4. Observe and interpret the results from all three servers.

## OBSERVATIONS:



UniProtKB - P9WN23 (TRHBO\_MYCTU)

Display [Help video](#) [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

[Entry](#)

**Protein** [Group 2 truncated hemoglobin GlbO](#)

**Gene** [glbO](#)

**Organism** [Mycobacterium tuberculosis \(strain ATCC 25618 / H37Rv\)](#)

**Status** [Reviewed](#) - Annotation score: - Experimental evidence at protein level

**Function:**

[Function](#)

[Names & Taxonomy](#)

[Subcellular location](#)

[Pathology & Biotech](#)

[PTM / Processing](#)

[Expression](#)

**Cofactor:**

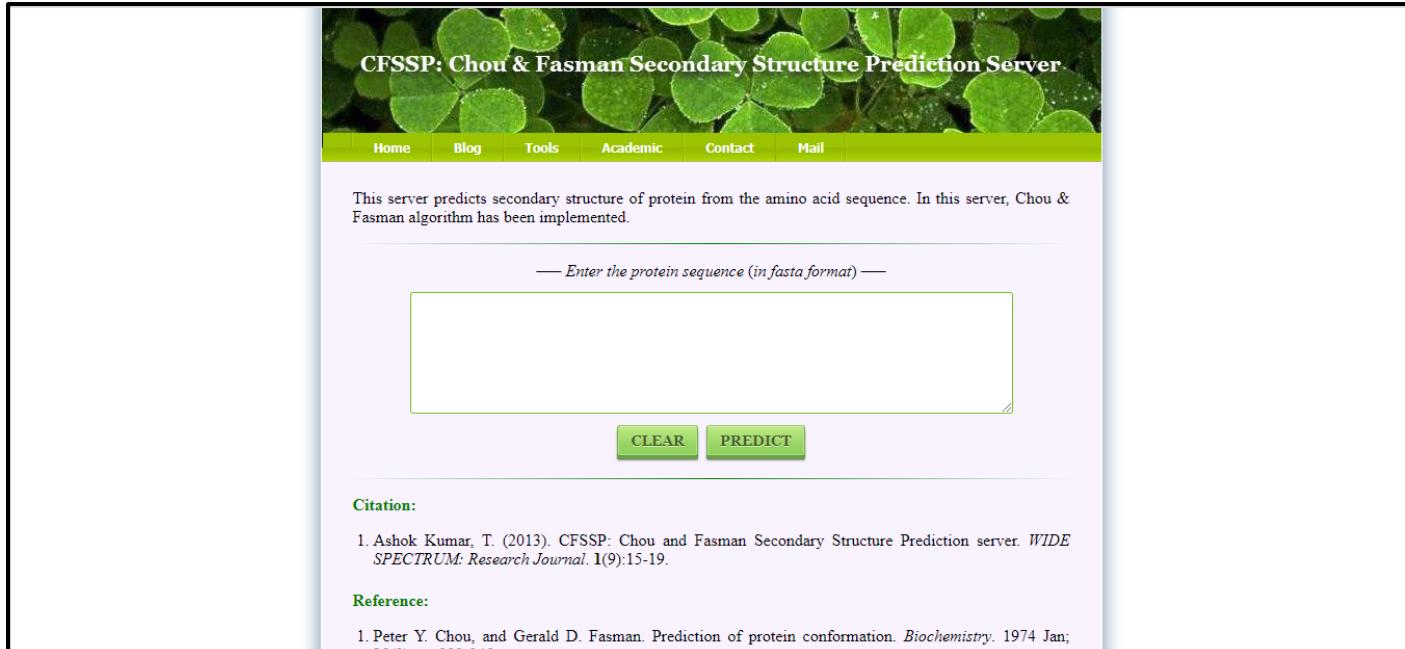
**heme**

**Note:** Binds 1 heme group per subunit.

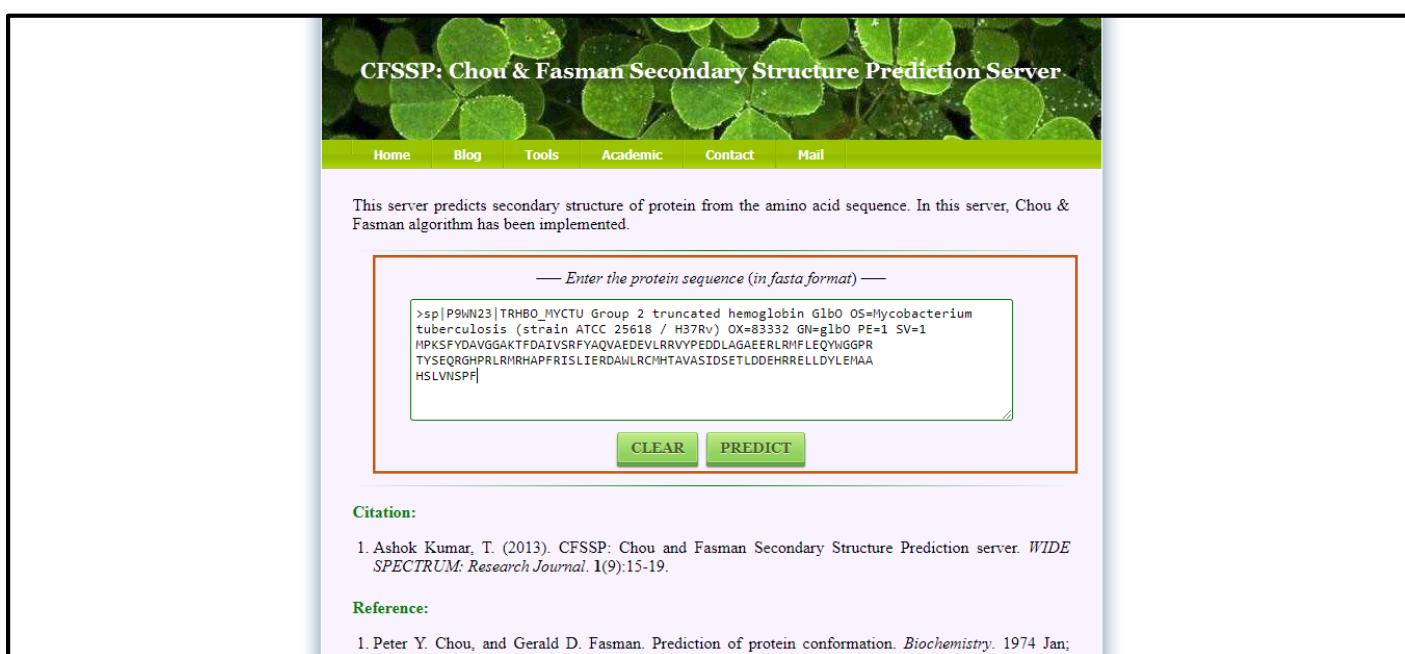
**Fig1. UniProt result page for hemoglobin protein.**



**Fig2. Hemoglobin FASTA sequence from UniProt.**



**Fig3.** Homepage for Chou & Fasman Secondary Structure Prediction Server.



**Fig4.** Search page for Chou & Fasman Secondary Structure Prediction Server with protein FASTA sequence as input.



**Fig5. Result page for Chou & Fasman Secondary Structure Prediction Server showing graphical representation.**



**Fig5.1. Result page for Chou & Fasman Secondary Structure Prediction Server showing secondary structure information, total residues and their percentage in Hemoglobin.**

**PRABI-GERLAND**  
**RHONE-ALPES BIOINFORMATIC POLE GERLAND SITE**  
Institute of Biology and Protein Chemistry

Home Services Teaching Publications Links Jobs Contact

## GOR IV SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

Paste a protein sequence below : [help](#)

Output width:

User : public@202.88.214.54. Last modification time : Thu Jan 14 21:53:29 2016. Current time : Fri Feb 11 16:08:15 2022

**Fig6. Homepage for GOR IV Secondary Structure prediction method.**

**PRABI-GERLAND**  
**RHONE-ALPES BIOINFORMATIC POLE GERLAND SITE**  
Institute of Biology and Protein Chemistry

Home Services Teaching Publications Links Jobs Contact

## GOR IV SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

Paste a protein sequence below : [help](#)

```
MPKSFYDAVGGAKTFDAIVSRFYAQVAEDEVLRVYPEDDLAGAEERLR
MFLEQYWGGPR
TYSEQRGHPRLRLMRHAPFRISLIERDAWLRCMHTAVASIDSETLDDHR
RELDDYLEMAA
HSLVNSP#
```

Output width:

User : public@202.88.214.54. Last modification time : Thu Jan 14 21:53:29 2016. Current time : Fri Feb 11 16:08:15 2022

**Fig7. Search page for GOR IV Secondary Structure prediction method with protein sequence as input.**

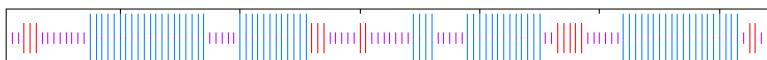
GOR4 result for : UNK\_1207780

Abstract GOR secondary structure prediction method version IV, J. Garnier, J.-F. Gibrat, B. Robson, Methods in Enzymology, R.F. Doolittle Ed., vol 266, 540-553, (1996)

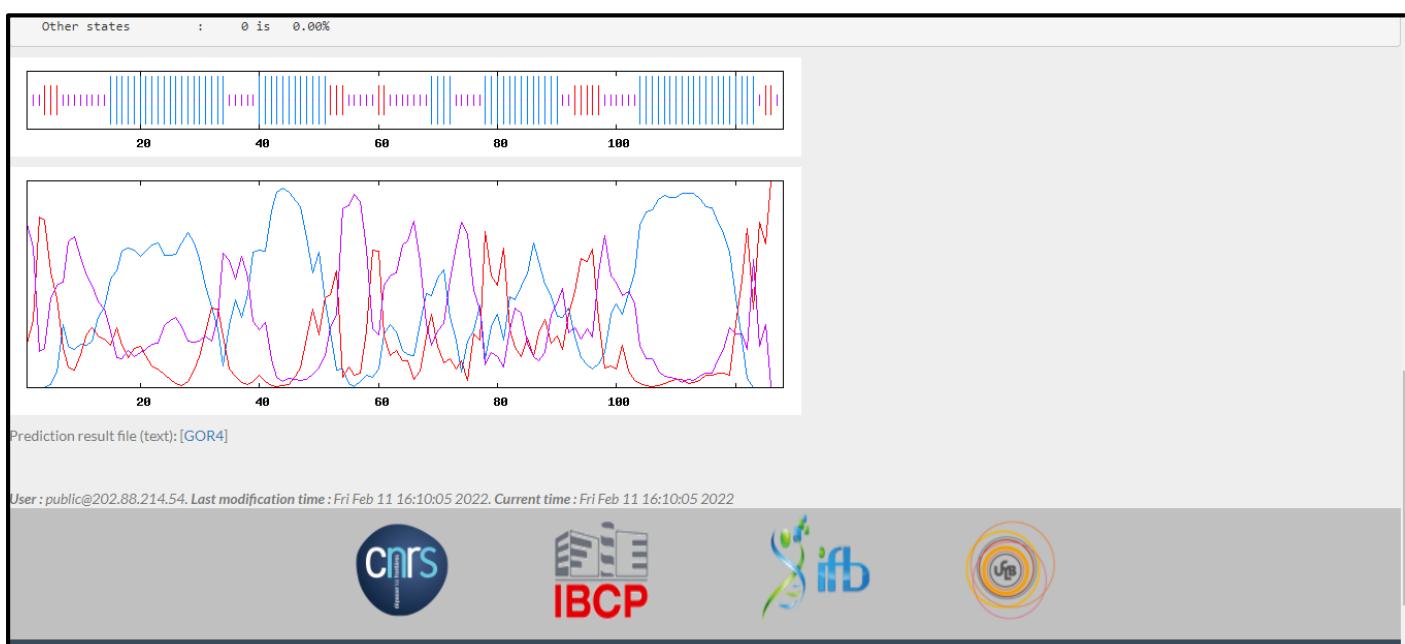
View GOR4 in: [AnTheProt \(PC\)](#) [Download...](#) [\[HELP\]](#)

Sequence length : 128

```
GOR4 :
  Alpha helix  (Hh) : 69 is 53.91%
  310 helix  (Gg) : 0 is 0.00%
  Pi helix    (Ii) : 0 is 0.00%
  Beta bridge (Bb) : 0 is 0.00%
  Extended strand (Ee) : 15 is 11.72%
  Beta turn   (Tt) : 0 is 0.00%
  Bend region (Ss) : 0 is 0.00%
  Random coil  (Cc) : 44 is 34.38%
  Ambiguous states (?) : 0 is 0.00%
  Other states  : 0 is 0.00%
```



**Fig8. Result page for GOR IV Secondary Structure prediction method showing total residues and their percentage in Hemoglobin.**



**Fig8.1. Result page for GOR IV Secondary Structure prediction method showing graphical representation.**

MAIN NAVIGATION

- [Introduction](#)
- [Contact](#)
- [Downloads & Branding](#)
- [Twitter/News](#)
- [PSIPRED Team Links](#)
  - [People](#)
  - [ProCovar](#)
  - [Publications](#)
  - [Vacancies](#)
- [PSIPRED Workbench Links](#)
  - [PSIPRED Workbench](#)
  - [Workbench Overview](#)
  - [Workbench Citation](#)
  - [Help & Tutorials](#)
  - [REST API](#)
  - [PSIPRED Github](#)

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.

#### Data Input

##### Select input data type

Sequence Data

PDB Structure Data

##### Choose prediction methods (hover for short description)

##### Popular Analyses

- PSIPRED 4.0 (Predict Secondary Structure)  DISOPRED3 (Disopred Prediction)  
 MEMSAT-SVM (Membrane Helix Prediction)  pGenTHREADER (Profile Based Fold Recognition)

##### Contact Analysis

- DeepMetaPSICOV 1.0 (Structural Contact Prediction)  MEMPACK (TM Topology and Helix Packing)

##### Fold Recognition

#### Required Options

**Fig9. Homepage for PSIPRED Protein Structure Prediction Server.**

Structure Modelling

Bioserf 2.0 (Automated Homology Modelling)  Domserf 2.1 (Automated Domain Homology Modelling)  
 DMPFold 1.0 Fast Mode (Protein Structure Prediction)

Domain Prediction

DomPred (Protein Domain Prediction)

Function Prediction

FFPred 3 (Eukaryotic Function Prediction)

[Help...](#)

**Submission details**

Protein Sequence

MPKSFYDAVGGAKTFDAIVSRFYAQVAEDEVLRRVYPEDDLAGAEERLRMFLEQYWGGPR  
TYSEQRGHPRRLMRHAPFRISLIERDAWLRCMHTAVASIDSETLDDEHRRELDDYLEMAA  
HSLVNSPF

[Help...](#)  
If you wish to test these services follow this link to retrieve a test fasta sequence.

Job name

Email (optional)

**Fig10. Search page for PSIPRED Protein Structure Prediction Server with protein sequence as input.**

MAIN NAVIGATION

- Introduction
- Contact
- Downloads & Branding
- Twitter/News

PSIPRED Team Links

- People
- ProCovar
- Publications
- Vacancies

PSIPRED Workbench Links

- PSIPRED Workbench
- Workbench Overview
- Workbench Citation
- Help & Tutorials
- REST API
- PSIPRED Github

Name : Structure Prediction

Copy Link: <http://bioinf.cs.ucl.ac.uk/psipred/>

Sequence Plot

Show psipred Show memsat Show aatypes

1 M P K S F Y D A V G G A K T F D A I V S R F Y A Q V A E D E V L R R V Y P E D D L A G A E E R L R M 50  
 51 F L E Q Y W G G P R T Y S E Q R G H P R L R M R H A P F R I S L I E R D A W L R C M H T A V 100  
 101 S E T L D D E H R R E L L D Y L E M A A H S L V N S P F 128

10 20 30 40 50

Strand Helix Coil Disordered Transmembrane Helix  
 Disordered, protein binding Putative Domain Boundary Membrane Interaction  
 Extracellular Re-entrant Helix Cytoplasmic Signal Peptide

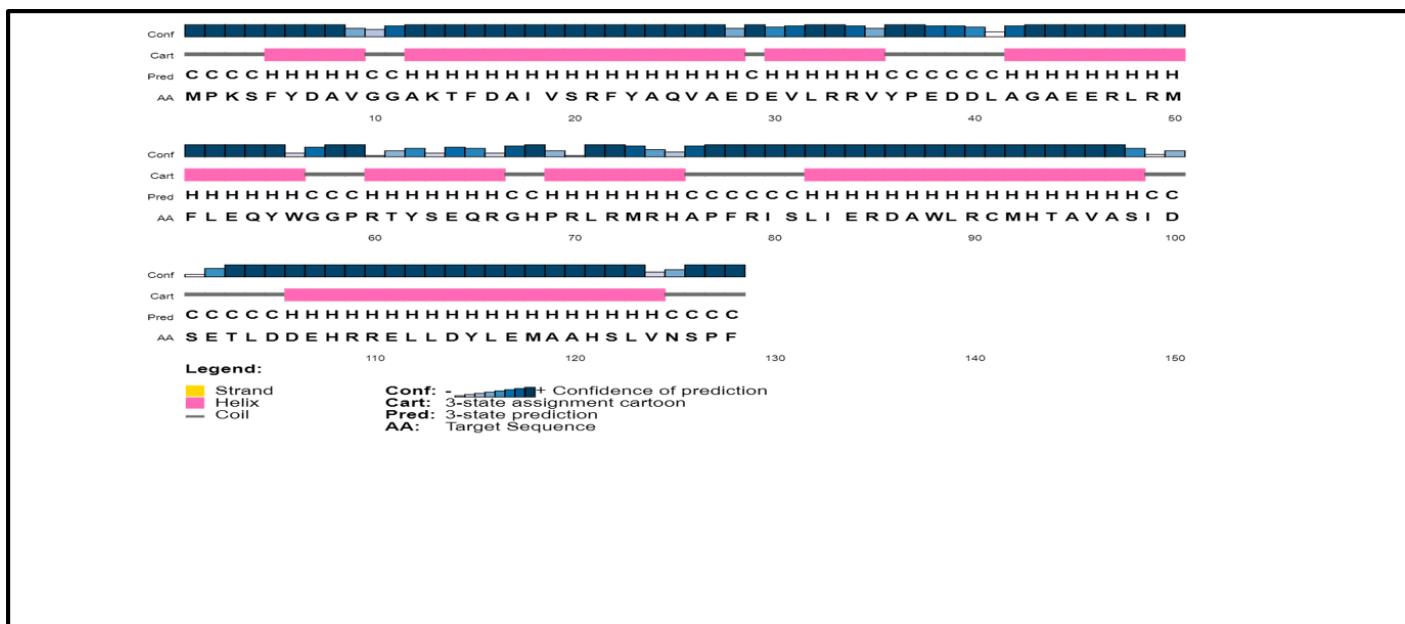
Get PNG

PSIPRED Cartoon

Fig11. Result page for PSIPRED Protein Structure Prediction Server.



Fig11.1 Result page for PSIPRED Protein Structure Prediction Server showing helices and coils present in Hemoglobin.



**Fig11.2 Result page for PSIPRED Protein Structure Prediction Server showing helices and coils present in Hemoglobin with confidence of prediction.**

## RESULTS:

First generation (Chou & Fasman Secondary Structure Prediction Server), Second generation (GOR IV) and Third generation (PSIPRED) were all used for secondary structure prediction of protein “Hemoglobin” which showed the structure mainly consists of helices and coils.

## CONCLUSION:

First generation (Chou & Fasman Secondary Structure Prediction Server), Second generation (GOR IV) and Third generation (PSIPRED) provides user with resources for secondary structure prediction. Chou & Fasman provides user with percentage of helix, turns and sheets while GOR IV provides user with more detailed information regarding  $\beta_{10}$  helix, Pi helix, beta bridges, beta turns, bend regions, etc. PSIPRED provides structural information along with confidence of prediction which tells user how reliable the information is. All these tools can be used by researchers to identify of binding sites and thus facilitates design of new drugs, study diseases that occur due to protein misfolding, etc.

## REFERENCES:

1. “Hemoglobin.” *Encyclopedia Britannica*, Encyclopedia Britannica, Inc., Retrieved February 11, 2022, from <https://www.britannica.com/science/hemoglobin>.
2. Ashok Kumar, T. (2013). *CFSSP: Chou and Fasman Secondary Structure Prediction server*. WIDE SPECTRUM: Research Journal. 1(9):15-19.
3. *NPS@ REFERENCES*. (n.d.-b). Npsa-Prabi.ibcp.fr. Retrieved February 11, 2022, from [https://npsa-prabi.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_references.html](https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_references.html)
4. McGuffin, L. J.; Bryson, K.; Jones, D. T. (2000). *The PSIPRED protein structure prediction server*. *Bioinformatics*, 16(4), 404–405. doi:10.1093/bioinformatics/16.4.404
5. *glbO - Group 2 truncated hemoglobin GlbO - Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv) - glbO gene & protein*. (n.d.). [Www.uniprot.org](http://www.uniprot.org/uniprot/P9WN23). Retrieved February 11, 2022, from <https://www.uniprot.org/uniprot/P9WN23>
6. *UniProt*. (2022). Uniprot.org. Retrieved February 11, 2022, from <https://www.uniprot.org/uniprot/P9WN23.fasta>
7. *CFSSP: Chou & Fasman Secondary Structure Prediction Server*. (n.d.). [Www.biogem.org](http://www.biogem.org/tool/chou-fasman/index.php). Retrieved February 11, 2022, from <http://www.biogem.org/tool/chou-fasman/index.php>

8. *NPS@ : GOR4 secondary structure prediction.* (n.d.). Npsa-Prabi.ibcp.fr. Retrieved February 11, 2022, from [https://npsa-prabi.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_gor4.html](https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html)
9. *NPS@ : GOR4 secondary structure prediction.* (2021). Ibcp.fr. Retrieved February 11, 2022, from [https://npsa-prabi.ibcp.fr/cgi-bin/secpred\\_gor4.pl](https://npsa-prabi.ibcp.fr/cgi-bin/secpred_gor4.pl)
10. *PSIPRED Workbench.* (n.d.). Bioinf.cs.ucl.ac.uk. Retrieved February 11, 2022, from <http://bioinf.cs.ucl.ac.uk/psipred/>
11. *PSIPRED Workbench.* (n.d.). Bioinf.cs.ucl.ac.uk. Retrieved February 11, 2022, from <http://bioinf.cs.ucl.ac.uk/psipred/&uuid=85a599ba-8b4d-11ec-97bc-00163e100d53>

## WEBLEM 2

### Introduction to protein classification

The current non-redundant protein sequence database contains over seven million entries and the number of individual functional domains is significantly larger than this value. The vast quantity of data associated with these proteins poses enormous challenges to any attempt at function annotation. Classification of proteins into sequence and structural groups has been widely used as an approach to simplifying the problem.

One of the applications of protein structure comparison is structural classification. The ability to compare protein structures 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 structures. 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. To date, several systems have been developed, the two most popular being Structural Classification of Proteins (SCOP) and Class, Architecture, Topology and Homologous (CATH). The following introduces the basic steps in establishing the systems to classify proteins.

The first step in structure classification is to remove redundancy from databases. Among the tens of thousands of entries in PDB, the majority of the structures are redundant as they correspond to structures solved at different resolutions, or associated with different ligands or with single-residue mutations. The redundancy can be removed by selecting representatives through a sequence alignment-based approach. The second step is to separate structurally distinct domains within a structure. Because some proteins are composed of multiple domains, they must be subdivided before a sensible structural comparison can be carried out. This domain identification and separation can be done either manually or based on special algorithms for domain recognition. Once multidomain proteins are split into separate domains, structure comparison can be conducted at the domain level, either through manual inspection, or automated structural alignment, or a combination of both. The last step involves grouping proteins/domains of similar structures and clustering them based on different levels of resemblance in secondary structure composition and arrangement of the secondary structures in space.

As mentioned, the two most popular classification schemes are SCOP and CATH, both of which contain a number of hierarchical levels in their systems.

#### **Class, Architecture, Topology and Homologous (CATH):**

The CATH database, originally developed in 1997, provides an up-to-date and systematic structural classification of protein 3D structures and is one of the Core Data Resources within ELIXIR, a major European distributed infrastructure for life-science information. CATH employs a semi-automated procedure to split 3D structures into their constituent domains (semi-independently folding globular units) and clusters these domains into homologous superfamilies where there is sufficient evidence of evolutionary ancestry. In addition to classifying domains in PDB structures, CATH assigns domains for protein sequences for which 3D structures are unknown. As well as providing this data in CATH, we also provide the data in our sister resource, Gene3D. Both CATH and Gene3D provide comprehensive structural domain assignments and functional annotation for protein sequences from major protein sequence databases such as UniProt and Ensembl. To obtain this predicted domain data we use a set of representative structural domains to 'seed' a set of protein sequence alignments, which are converted into hidden Markov models (HMMs). HMMs are then used to identify closely related domains within protein sequences from UniProt and ENSEMBL. Thus, by combining protein structure and sequence, CATH provides comprehensive structure-based domain superfamily assignments for over 82 million protein sequences (151 million protein domains).

The domains are classified into the following hierarchical levels: Class (C), Architecture (A), Topology (T) and Homologous superfamilies (H). For every superfamily, CATH provides structural superpositions of all representative protein domains using an in-house structure and sequence alignment program (SSAP).

The members of Homologous superfamilies (H) share a conserved structural core, however in large superfamilies they often tend to have diverse functions. To address this, CATH has developed a functional classification protocol (FunFHMMer) utilising a hierarchical agglomerative clustering algorithm, to further sub-classify Homologous superfamilies (H) into functionally coherent groups known as Functional Families (referred to as FunFams). FunFHMMer segregates functional families on the basis of specificity-determining positions as well as highly conserved positions in cluster alignments and calculates a functional coherence index in order to determine functionally coherent alignments. For each FunFam, CATH provides sequence alignments (generated using MAFFT, profile hidden Markov models (HMMs, generated using HMMER3), and a set of high-quality GO annotations from UniProt-GOA. As reported in the previous release, the CATH website provides a sequence-based search for identifying FunFams using query protein sequences ([cathdb.info/search/by sequence](http://cathdb.info/search/by sequence)), or through the API.

FunFams tend to be more functionally coherent than other domain-based approaches, making them useful for predicting functional sites as well as protein structure. Function prediction pipelines developed using FunFams are consistently ranked among the top performers for Molecular Function and Biological Process Gene Ontology terms in the Critical Assessment of Functional Annotation competition (CAFA).

Non-globular domains can cause problems during the initial domain chopping procedure. Since the release of CATH version 4.2, we have re-classified the non-globular superfamilies in a new Class 6 (6.x.x.x), separate from the main hierarchy. This special class now contains 790 superfamilies, and with continued curation efforts, we plan to include other special cases and architectures, such as short and synthetic peptides, fragments, linkers, nucleic acids and low resolution structures. A consequence of this reclassification brings down the number of SuperFamilies in the canonical 1–4 classification to a total of 5841.

The continuous deposition of structures and sequences in PDB and UniProt has led to significant expansions in the CATH superfamilies since the last release. Furthermore, superfamilies are unevenly populated and the 100 most populated CATH-Gene3D superfamilies contain around 54% of the >150 million sequences characterised in our resource. Among these, the top 11 ‘mega’ superfamilies contain millions of sequences, requiring novel approaches to reduce the computing time and processing power to properly classify them into functional families. Due to a newly redesigned functional classification pipeline, we can report an expansion of our functional families in CATH v4.3 to 212 872 families comprising 34 700 216 sequences, for which we can provide more accurate functional annotations. This article highlights improvements in our functional classification protocols, implemented to address the functional classification of superfamilies in general and of ‘mega-superfamilies’ in particular.

### **Structural Classification of Proteins (SCOP):**

The Structural Classification of Proteins—extended (SCOPe) knowledgebase aims to provide an accurate, detailed, and comprehensive description of the structural and evolutionary relationships amongst the majority of proteins of known structure, along with resources for analyzing the protein structures and their sequences. Structures from the PDB are divided into domains and classified using a combination of manual curation and highly precise automated methods. In the current release of SCOPe, 2.08, search and display tools for analysis of genetic variants are developed and mapped to structures classified in SCOPe. In order to improve the utility of SCOPe to automated methods such as deep learning classifiers that rely on multiple alignment of sequences of homologous proteins, new machine-parseable annotations that indicate aberrant structures as well as domains that are distinguished by a smaller repeat unit are developed. Structures from 74 of the largest Pfam families not previously classified in SCOPe were classified, and algorithm to remove N- and C-terminal cloning, expression and purification sequences from SCOPe domains were improved. SCOPe 2.08-stable classifies 106 976 PDB entries (about 60% of PDB entries).

By analogy with taxonomy, SCOP was created as a hierarchy of several levels where the fundamental unit of classification is a ‘domain’ in the experimentally determined protein structure. The hierarchy of SCOP domains comprises the following levels: ‘Species’ representing a distinct protein sequence and its naturally occurring or artificially created variants; ‘Protein’ grouping together similar sequences of essentially the same functions that either originate from different biological species or represent different isoforms within the same species; ‘Family’ containing proteins with similar sequences but typically distinct functions and ‘Superfamily’ bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor. Near the root, the basis of classification is purely structural: structurally similar superfamilies are grouped into ‘Folds’, which are further arranged into ‘Classes’ based mainly on their secondary structure content and organization.

It is believed that members of the same superfamily share a common ancestral origin, although the relationships between families are considered distant. Folds consist of superfamilies with a common core structure, which is determined manually. This level describes similar overall secondary structures with similar orientation and connectivity between them. Members within the same fold do not always have evolutionary relationships. Some of the shared core structure may be a result of analogy. Classes consist of folds with similar core structures. This is at the highest level of the hierarchy, which distinguishes groups of proteins by secondary structure compositions such as all  $\alpha$ , all  $\beta$ ,  $\alpha$  and  $\beta$ , and so on. Some classes are created based on general features such as membrane proteins, small proteins with few secondary structures and irregular proteins. Folds within the same class are essentially randomly related in evolution.

Class, Architecture, Topology and Homologous (CATH) and Structural Classification of Proteins (SCOP) thus are useful tools for classification of proteins. Classification of proteins can allow researchers to functionally annotate proteins and establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structures.

## REFERENCES:

1. Xiong, J. (2008). *Protein Structure Visualization, Comparison, and Classification. Essential bioinformatics*. Cambridge: Cambridge University Press. 195-197.
2. Petrey, D., & Honig, B. (2009, June). *Is protein classification necessary? Toward alternative approaches to function annotation*. Current Opinion in Structural Biology. <https://doi.org/10.1016/j.sbi.2009.02.001>
3. Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., ... Orengo, C. A. (2020). *CATH: increased structural coverage of functional space*. Nucleic Acids Research. doi:10.1093/nar/gkaa1079
4. Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N., & Brenner, S. (2021). SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. Nucleic Acids Research, 50(D1), D553–D559. <https://doi.org/10.1093/nar/gkab1054>

## WEBLEM 2a

### CATH and SCOP

(URL: <https://www.cathdb.info/>  
<https://scop.berkeley.edu/#:~:text=SCOPe%20>)

#### **AIM:**

To study the structural classification of protein Insulin using CATH and SCOP database.

#### **INTRODUCTION:**

Insulin is a hormone created by your pancreas that controls the amount of glucose in your bloodstream at any given moment. It also helps store glucose in your liver, fat, and muscles. Finally, it regulates your body's metabolism of carbohydrates, fats, and proteins. Structural classification information of protein insulin can be retrieved from CATH and SCOP database.

CATH (<https://www.cathdb.info>) identifies domains in protein structures from PDB and classifies these into evolutionary superfamilies, thereby providing structural and functional annotations. There are two levels: CATH-B, a daily snapshot of the latest domain structures and superfamily assignments, and CATH+, with additional derived data, such as predicted sequence domains, and functionally coherent sequence subsets (Functional Families or FunFams). The latest CATH+ release, version 4.3, significantly increases coverage of structural and sequence data, with an addition of 65,351 fully-classified domains structures (+15%), providing 500 238 structural domains, and 151 million predicted sequence domains (+59%) assigned to 5481 superfamilies. The FunFam generation pipeline has been re-engineered to cope with the increased influx of data. Three times more sequences are captured in FunFams, with a concomitant increase in functional purity, information content and structural coverage. FunFam expansion increases the structural annotations provided for experimental GO terms (+59%). CATHFunVar are web-pages displaying variations in protein sequences and their proximity to known or predicted functional sites. There are two case studies (1) putative cancer drivers and (2) SARS-CoV-2 proteins. Finally, there are improved links to and from CATH including SCOP, InterPro, Aquaria and 2DProt.

As well as providing this data in CATH, there is also a sister resource, Gene3D (available at <http://gene3d.biochem.ucl.ac.uk/Gene3D/> ). Both CATH and Gene3D provide comprehensive structural domain assignments and functional annotation for protein sequences from major protein sequence databases such as UniProt and Ensembl.

The Structural Classification of Proteins—extended (SCOPe) knowledgebase aims to provide an accurate, detailed, and comprehensive description of the structural and evolutionary relationships amongst the majority of proteins of known structure, along with resources for analyzing the protein structures and their sequences. Structures from the PDB are divided into domains and classified using a combination of manual curation and highly precise automated methods.

By analogy with taxonomy, SCOP was created as a hierarchy of several levels where the fundamental unit of classification is a 'domain' in the experimentally determined protein structure. The hierarchy of SCOP domains comprises the following levels: 'Species' representing a distinct protein sequence and its naturally occurring or artificially created variants; 'Protein' grouping together similar sequences of essentially the same functions that either originate from different biological species or represent different isoforms within the same species; 'Family' containing proteins with similar sequences but typically distinct functions and 'Superfamily' bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor. Near the root, the basis of classification is purely structural: structurally

similar superfamilies are grouped into ‘Folds’, which are further arranged into ‘Classes’ based mainly on their secondary structure content and organization.

## METHODOLOGY:

### CATH database:

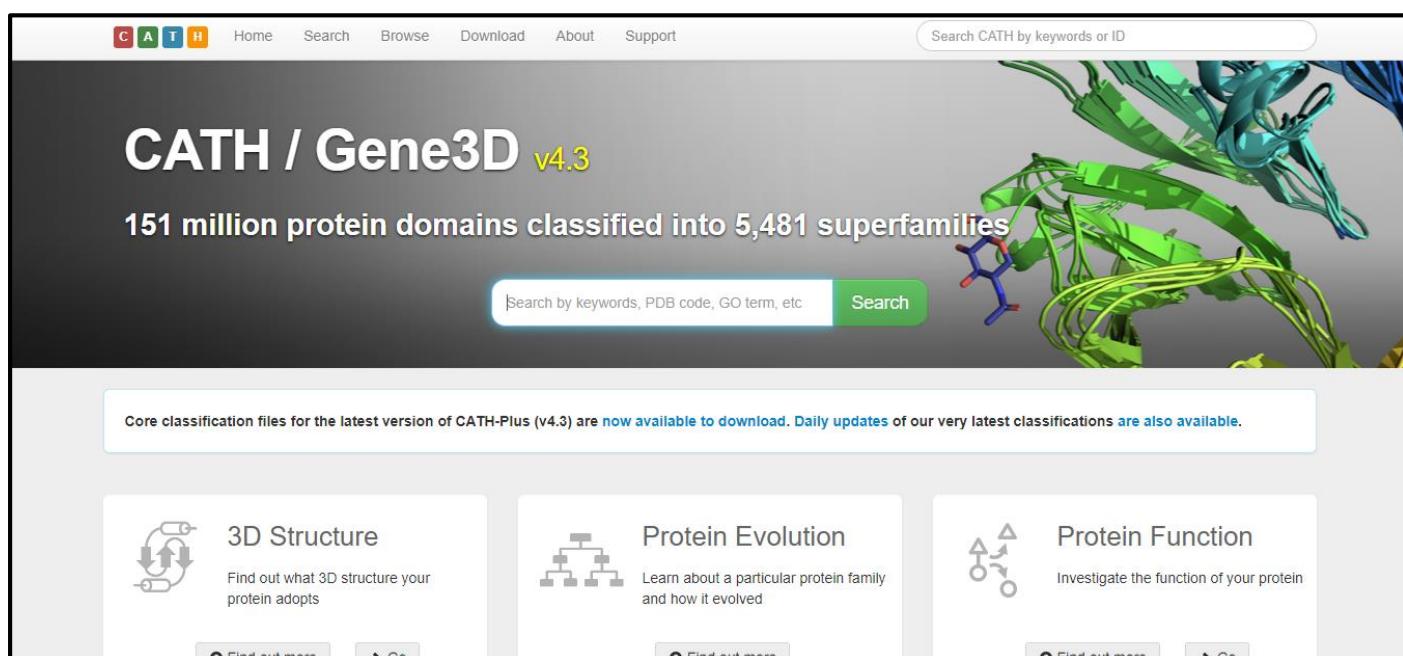
1. Open homepage for CATH. (URL: <https://www.cathdb.info/>)
2. Search for query “Insulin”.
3. Observe the results under all three sections: CATH superfamilies, CATH domains and PDB structures.
4. Interpret the results.

### SCOP database:

1. Open homepage for SCOPe database. (URL: <https://scop.berkeley.edu/#:~:text=SCOPe%20>)
2. Search for query “Insulin.”
3. Observe results for all categories. (folds, superfamilies, families, proteins, domains)
4. Interpret the results.

## OBSERVATIONS:

### CATH database:



**Fig1. Homepage for CATH database**

 Home Search Browse Download About Support

Search CATH by keywords or ID

# CATH / Gene3D v4.3

151 million protein domains classified into 5,481 superfamilies

insulin

Core classification files for the latest version of CATH-Plus (v4.3) are now available to download. Daily updates of our very latest classifications are also available.

**3D Structure**  
Find out what 3D structure your protein adopts

**Protein Evolution**  
Learn about a particular protein family and how it evolved

**Protein Function**  
Investigate the function of your protein

Fig2. Search for protein insulin

 Home Search Browse Download About Support

Search CATH by keywords or ID

## Search CATH

insulin

Search CATH by text or ID  
Type in general text or biological identifiers in the box and click "search" to perform a general text search on CATH data.

Examples: "protease", "1cuk"

### Results

Currently displaying the top ranked hits from three separate search queries: CATH Superfamilies, CATH domains and PDB entries. Click "View all entries" to expand each section and show all the hits. Use the panel on the right to add additional filters to this query.

**172 Matching CATH Superfamilies**

**1.10.10.10**  
"winged helix" repressor DNA binding domain  
putamen development, Methionine aminopeptidase 2, 2,7-dihydroxy-5-methyl-1-naphthoate 7-O-methyltransferase, CST complex, Protein STN1, telomere capping, Neck appendage protein, Transcriptional regulator, GntR family, HTH-type transcriptional regulator TsaR, endo-

**Current Search Filters**  
Remove search filters by clicking on the 'X'  
insulin

**Filter by Keyword / CATH ID**  
Start typing and press 'enter' to add a new filter

**Top Keywords**

Fig3. Hit page for insulin showing matching CATH Superfamilies

Fig3.1. Hit page for insulin showing matching CATH Domains and PDB structures

#### 4) Insulin matching CATH Superfamilies

**SUPERFAMILY LINKS**

**Summary** (selected)

[Superfamily Superposition](#)  
[Classification / Domains](#)  
[Functional Families](#)  
[Structural Neighbourhood](#)

**Functional Families**

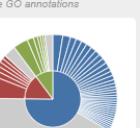
Overview of the Structural Clusters (SC) and Functional Families within this CATH Superfamily. Clusters with a representative structure are represented by a filled circle.



SC.1: [insulin-like growth fact](#), [Probable insulin-like pr](#)  
SC.2: [insulin-like growth fact](#), [insulin](#)

- [insulin-like 3](#)
- [Insulin, isoform 2](#)
- [Insulin 5](#)
- [probable insulin-like pe](#)
- [Insulin-like peptide 4](#)
- [GM2480](#)
- [Insulin-like peptide 2](#)
- [Insulin-like growth fact](#)
- [GD12890](#)
- [Insulin-like peptide 6](#)
- [GM25185](#)

**GO Diversity**  
Unique GO annotations



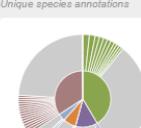
382 Unique GO terms

**EC Diversity**  
Unique EC annotations



656 Unique EC terms

**Species Diversity**  
Unique species annotations



656 Unique species

**Superfamily Summary**

A general summary of information for this superfamily.

**Structures**

Domains:	49
Domain clusters (>95% seq id):	10
Domain clusters (>35% seq id):	5

Unique PDBs: 36

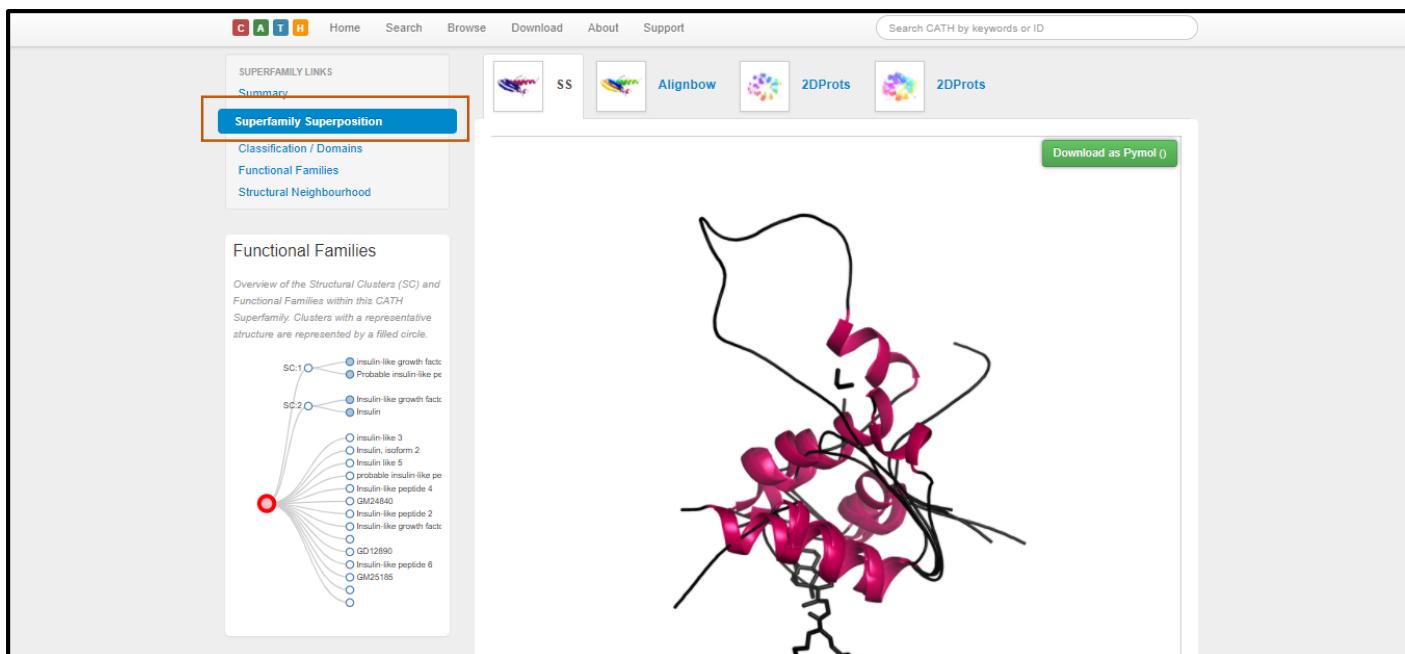
**Alignments**

Structural Clusters (5A):	4
Structural Clusters (9A):	2
FunFam Clusters:	18

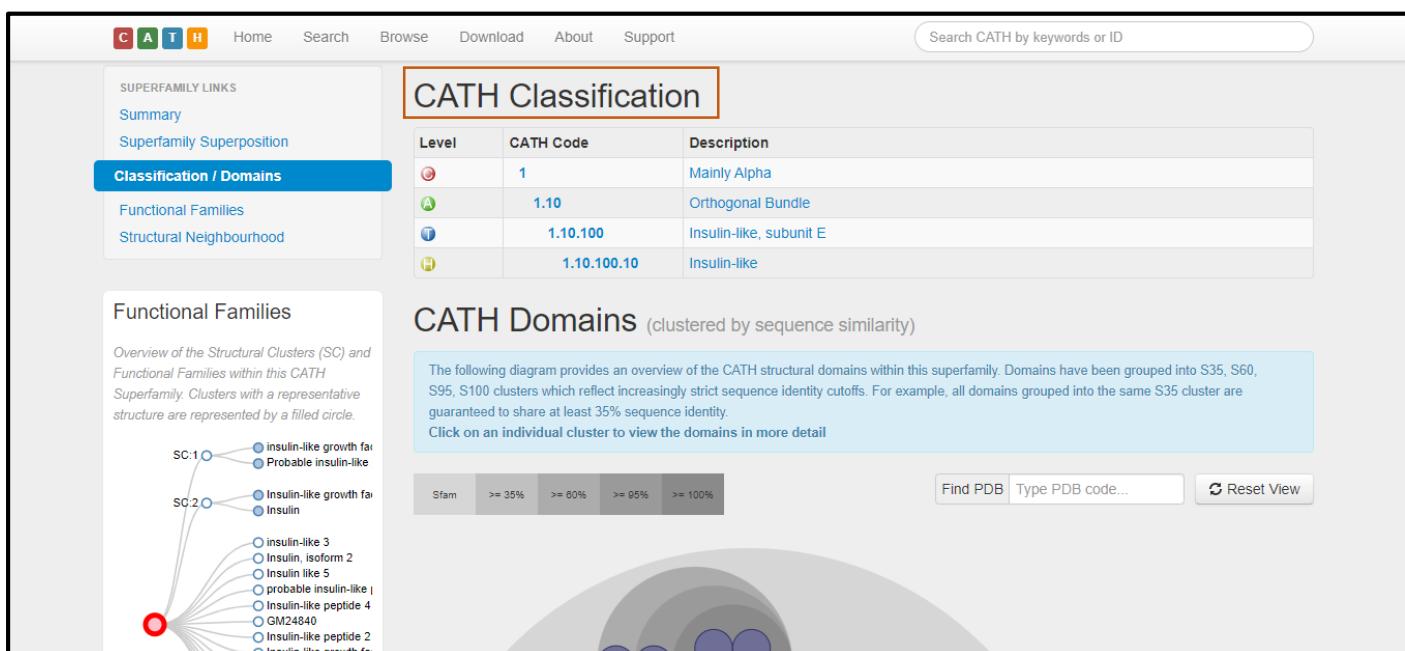
**Function**

Unique EC:	
Unique GO:	382

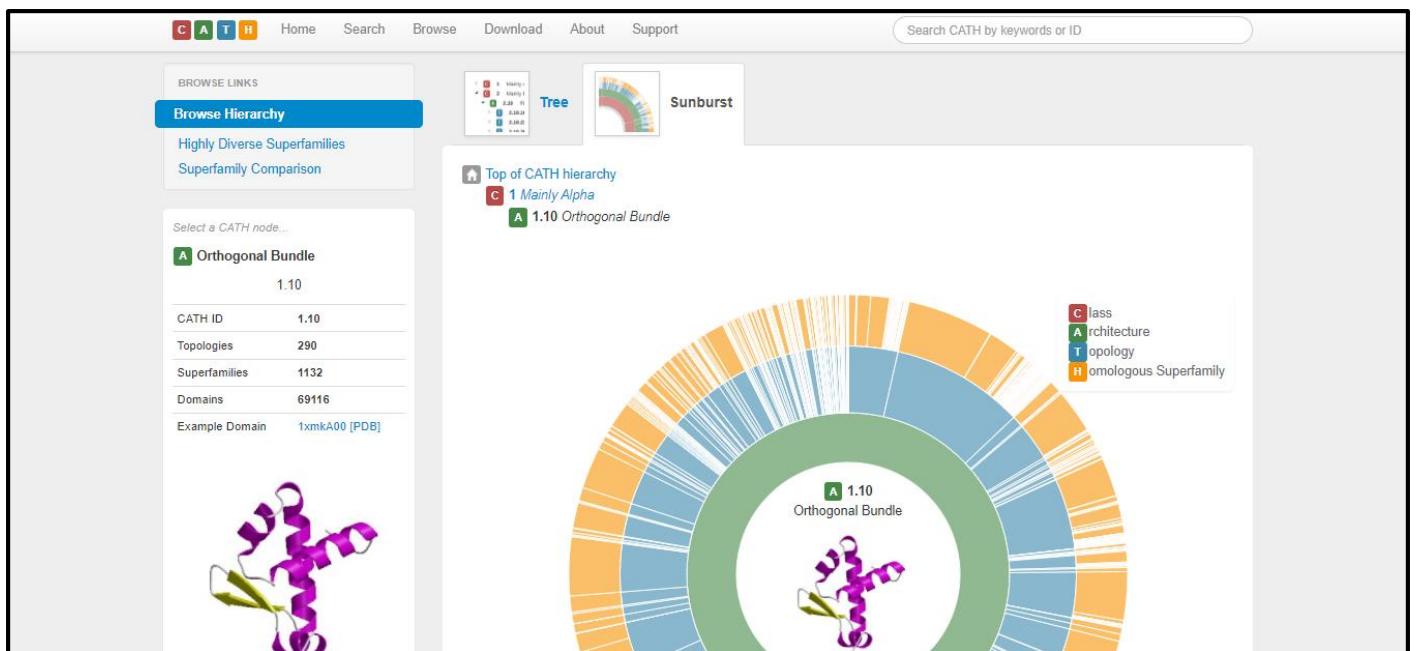
#### **Fig4.1. Result page for Summary**



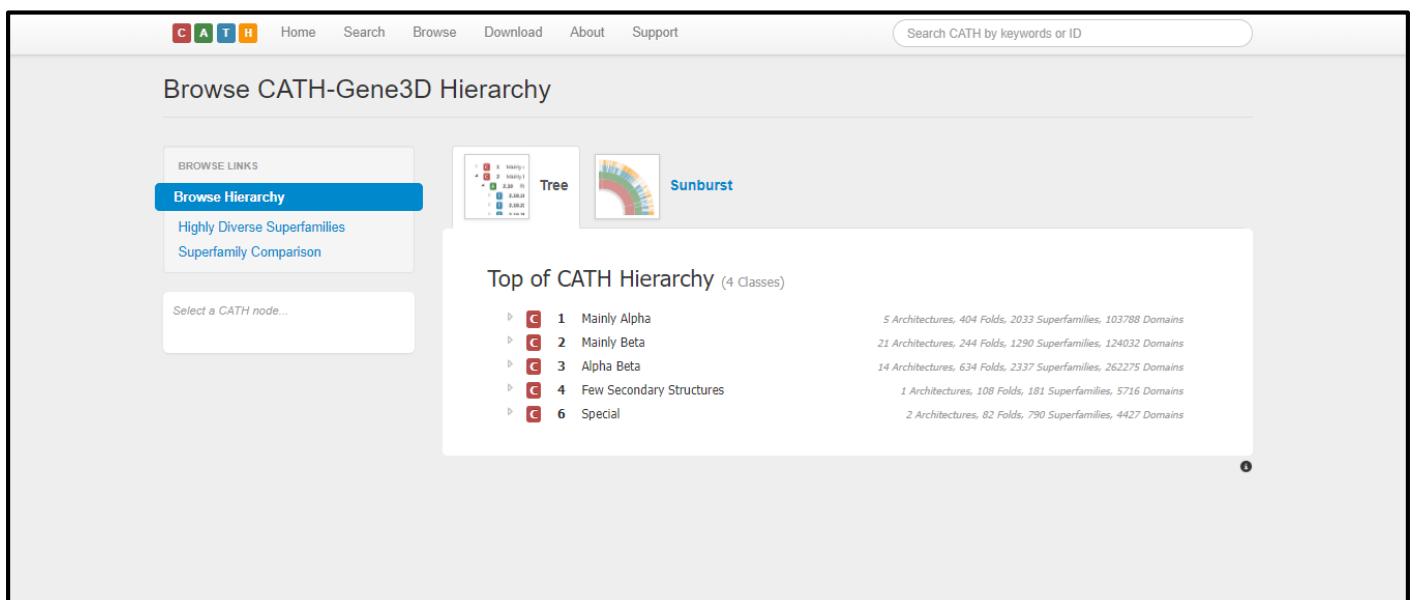
**Fig4.2. Result page for Superfamily Superposition**



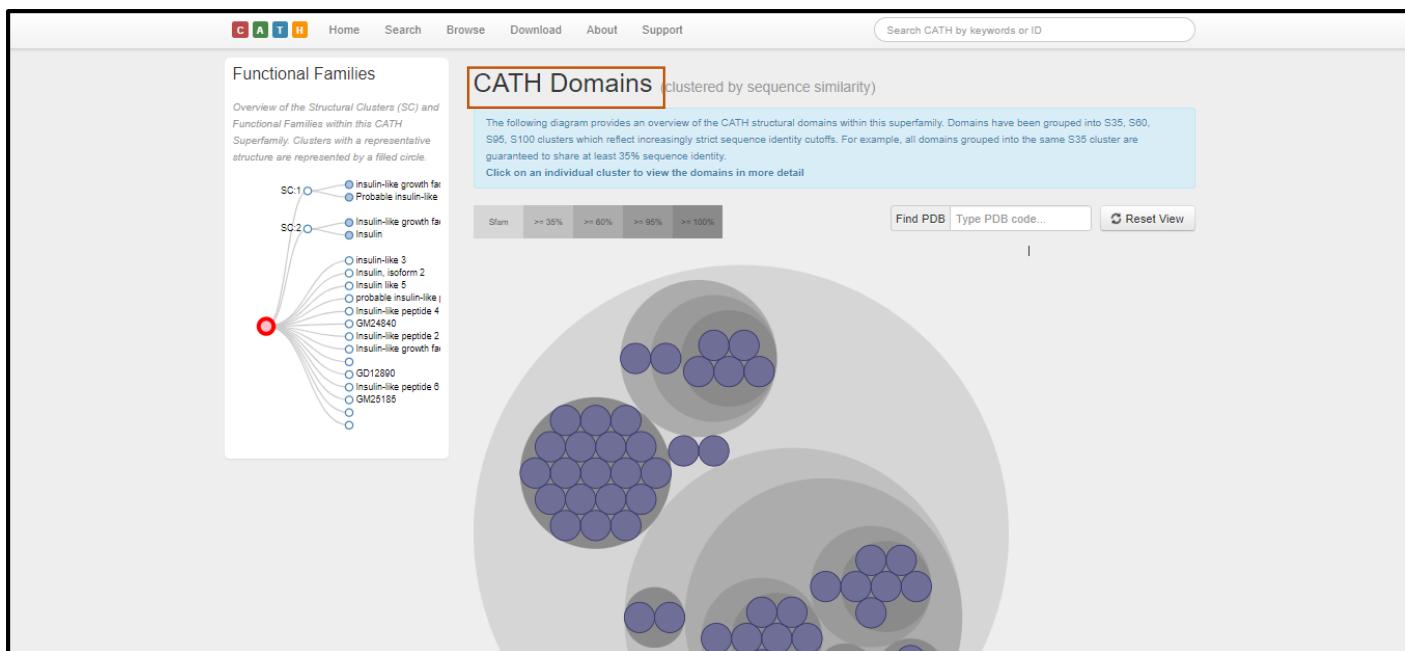
**Fig4.3. Result page for Classification/Domains**



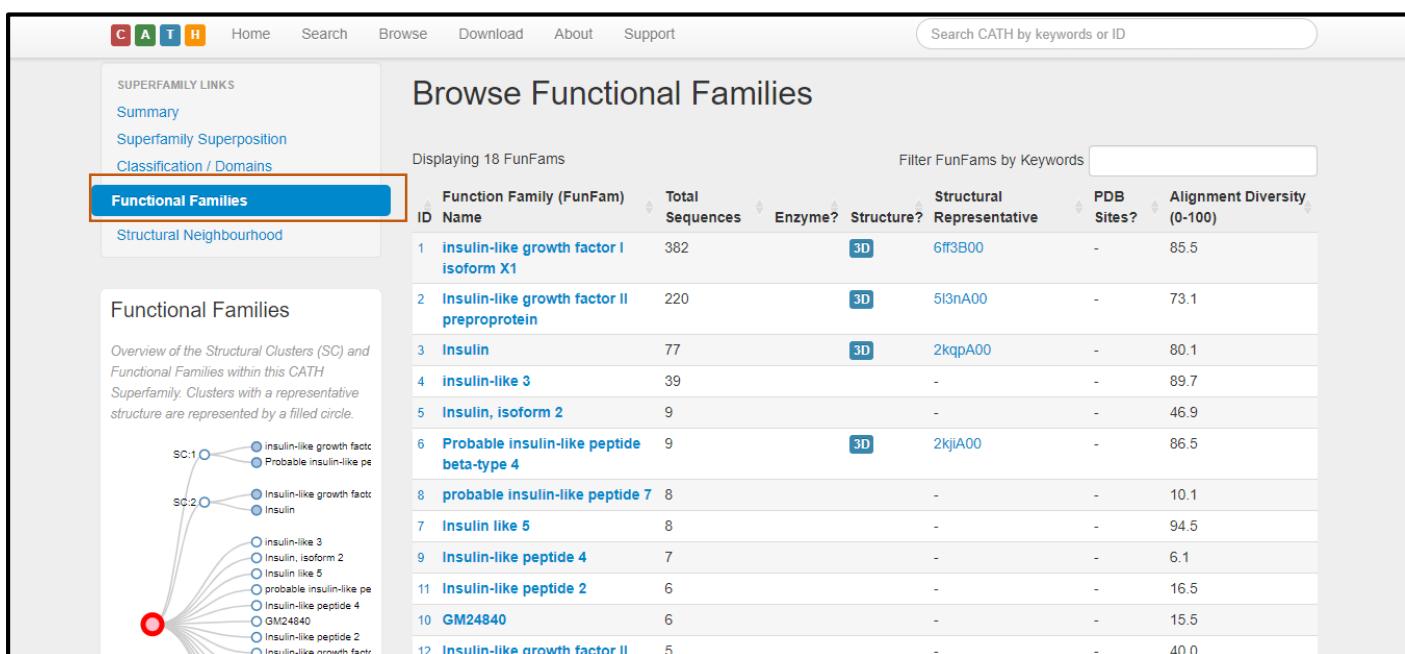
**Fig4.4. Result page for Orthogonal Bundle (Sunburst) under CATH Classification**



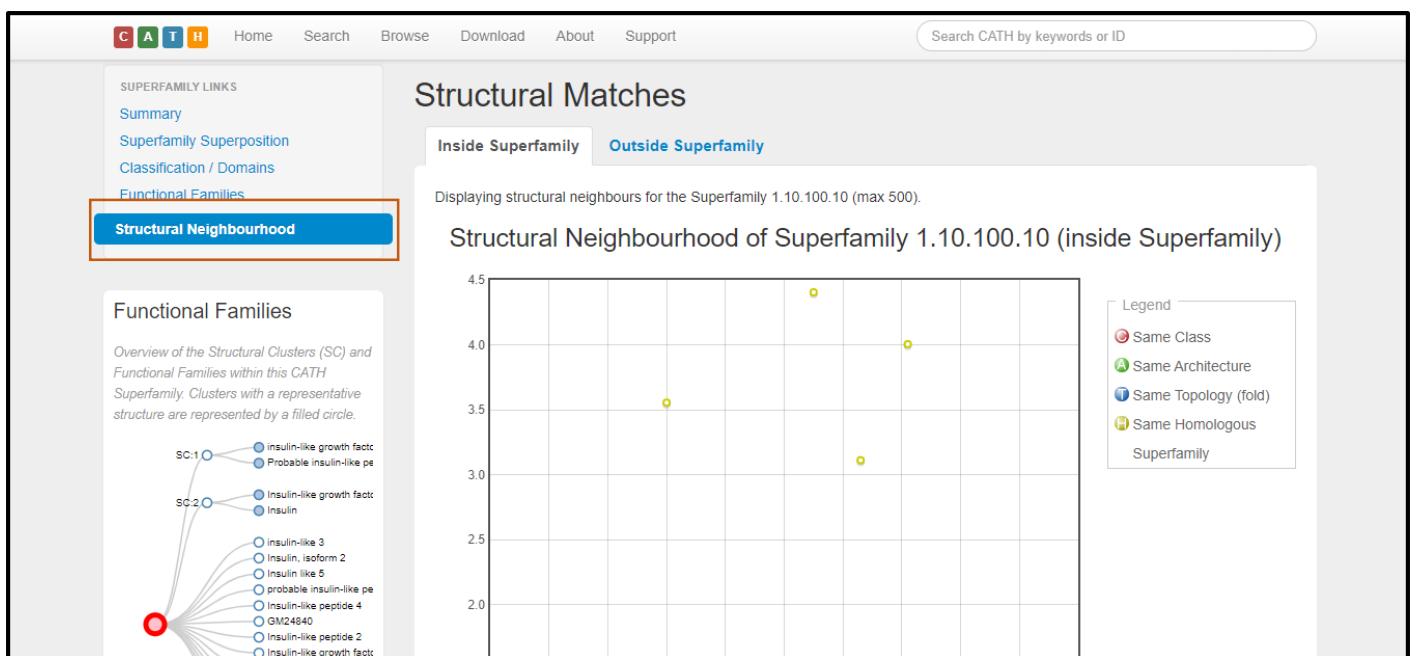
**Fig4.5. Result page for Orthogonal Bundle (Tree) under CATH Classification**



**Fig4.6. Result page for CATH domains**

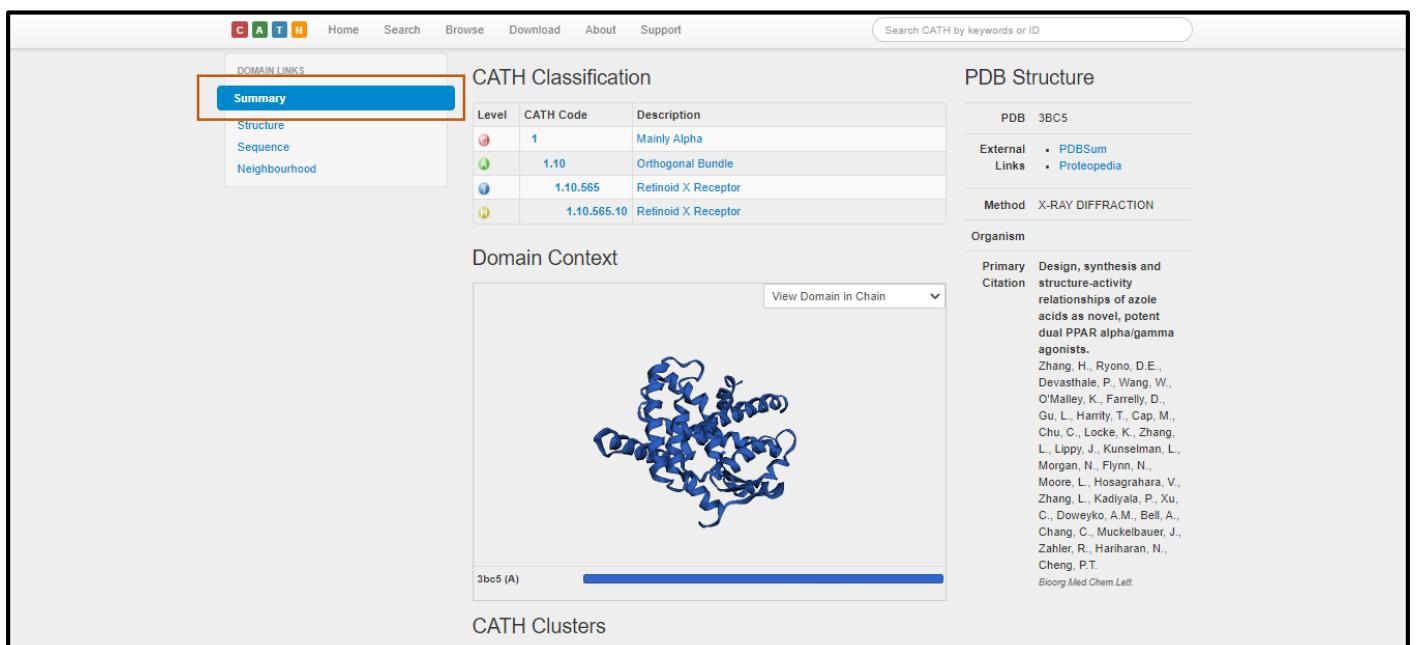


**Fig4.7. Result page for Functional Families**

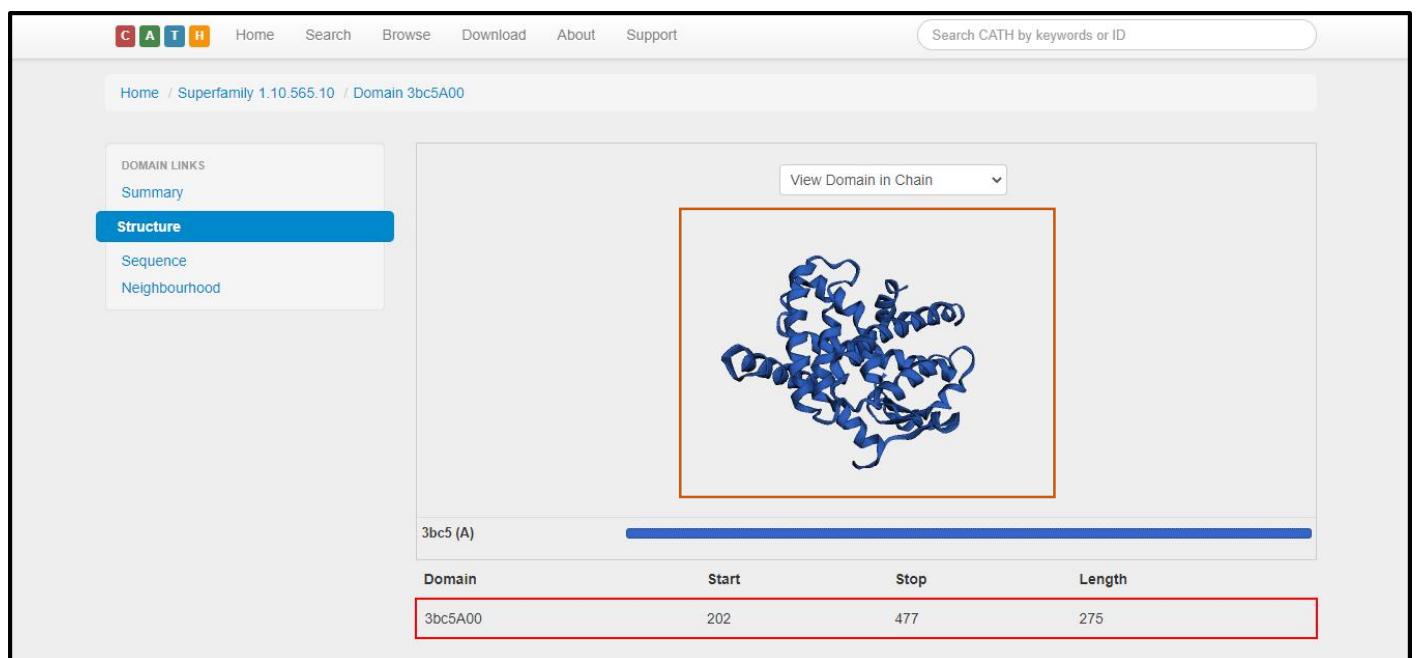


**Fig4.8. Result page for Structural Neighbourhood**

## 5) Insulin matching CATH domains



**Fig5.1. Result page for Summary**



**Fig5.2. Result page for Structure**

Home Search Browse Download About Support Search CATH by keywords or ID 1 keywords

CATH Domain 3bc5A00

Home / Superfamily 1.10.565.10 / Domain 3bc5A00

DOMAIN LINKS Summary Structure Sequence Neighbourhood

ATOM Sequence

The ATOM sequence is based on the residues observed in the ATOM records of the PDB file for this structure.

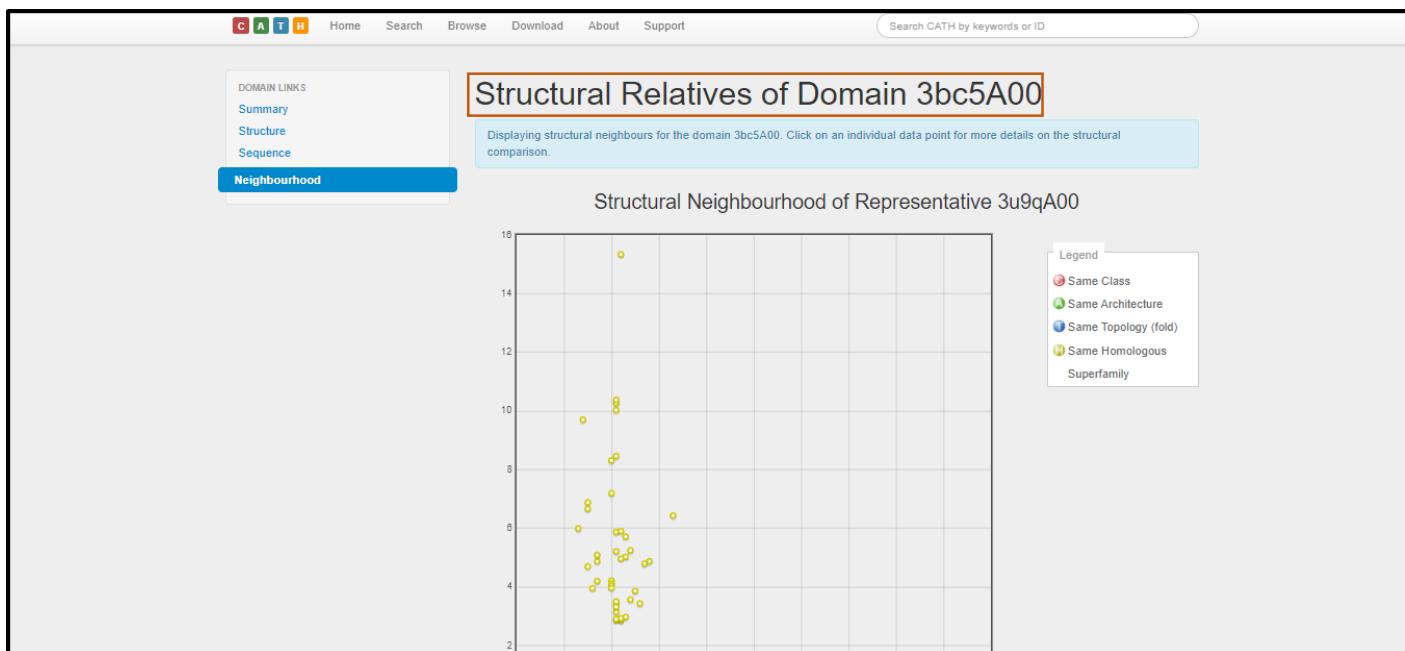
```
>cath14_3_0|3bc5A00/21-296 PDB=202-477
MQLNPEASDLRALAKHLYDSIYKSFPPLTKAKARAILTGKTSPPFVIYDMNSLMMGEDKIF
```

COMBS Sequence

The COMBS sequence is based on the residues observed in the SEQRES records of the PDB file for this structure. This can sometimes contain extra residues that were not able to be resolved in the 3D co-ordinates.

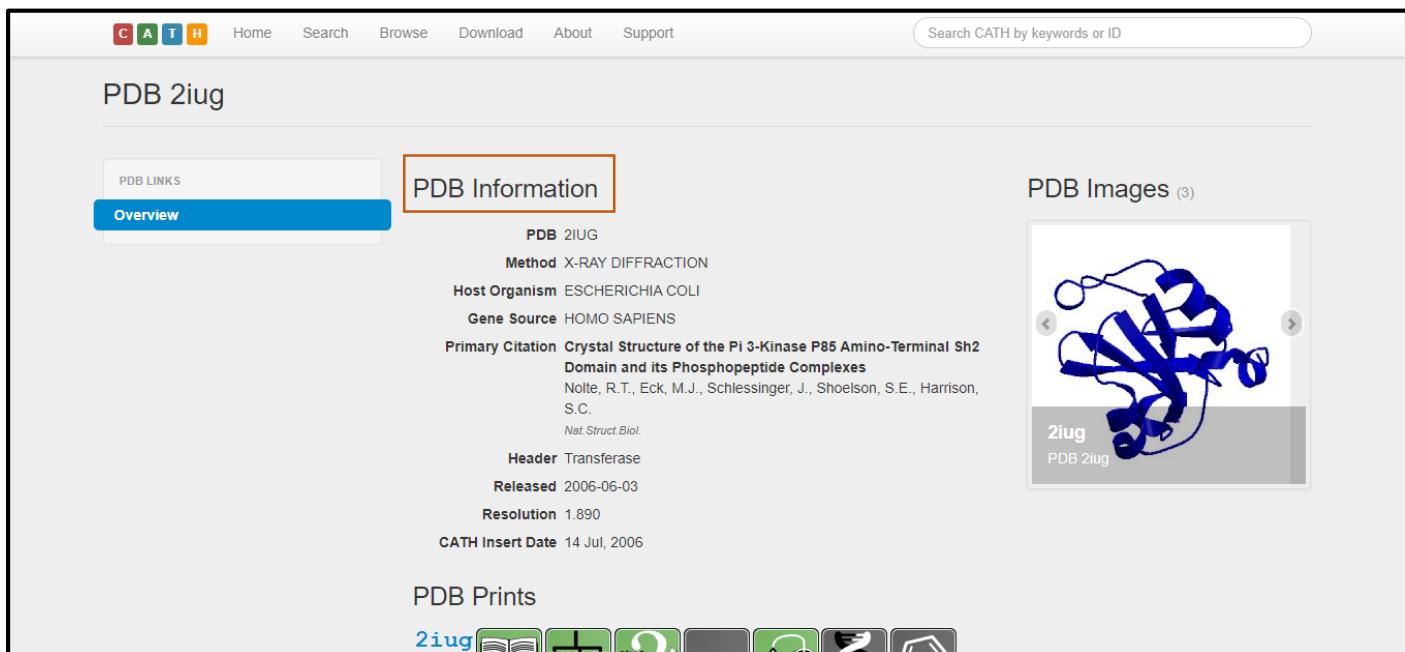
```
>cath14_3_0|3bc5A00/21-296 PDB=202-477
MQLNPEASDLRALAKHLYDSIYKSFPPLTKAKARAILTGKTTDKSPFVIYDMNSLMMGEDK
```

**Fig5.3. Result page for Sequence**



**Fig5.4. Result page for Neighbourhood**

## 6) Insulin matching PDB structures



**Fig6.1. Result page for PDB Information**

CATH Home Search Browse Download About Support Search CATH by keywords or ID

**PDB Prints**

2iug      

**PDB Chains (1)**

Chain ID	Date inserted into CATH	CATH Status
A	14 Jul, 2006	Chopped 

**CATH Domains (1)**

Domain ID	Date inserted into CATH	Superfamily	CATH Status
2iugA00	16 Oct, 2006	3.30.505.10	Assigned 

**UniProtKB Entries (1)**

Accession	Gene ID	Taxon	Description
 P27986	P85A_HUMAN	Homo sapiens	Phosphatidylinositol 3-kinase regulatory subunit alpha

**Fig6.2. Result page for PDB Prints, Chains, CATH Domains and UniPortKB Entries**

**SCOP database:**

**SCOPe** [Browse](#) [Stats & History](#) [Downloads](#) [Help](#) [Search \(click for examples\)](#) 

Welcome to SCOPe!

SCOPe (Structural Classification of Proteins — extended) is a database developed at the Berkeley Lab and UC Berkeley to extend the development and maintenance of SCOP. SCOP was conceived at the MRC Laboratory of Molecular Biology, and developed in collaboration with researchers in Berkeley. Work on SCOP (version 1) concluded in June 2009 with the release of SCOP 1.75. SCOPe classifies many newer structures through a combination of automation and manual curation, and corrects some errors in SCOP, aiming to have the same accuracy as the hand-curated SCOP releases. SCOPe also incorporates and updates the ASTRAL database.

[About SCOPe](#) [Stats & Prior Releases](#)

**News**

2022-01-07: We published a paper describing the new features in SCOPe 2.08-stable. [\[PDF\]](#)

2021-09-20: SCOPe 2.08-stable has been released, with nearly 20,000 new PDB entries added since the last stable release. Important features include genetic variant search tools and annotations of structural heterogeneity and repeat units. Click either the About or Stats & History links for more details on what's new!

2018-11-30: We published a paper describing updates to SCOPe, focusing on our findings from classifying large structures. [\[PDF\]](#)

**Classes in SCOPe 2.08:**

1.  All alpha proteins [46456] (290 folds)

**Fig1. Homepage for SCOPe database**

## Welcome to SCOPe!

SCOPe (Structural Classification of Proteins — extended) is a database developed at the Berkeley Lab and UC Berkeley to extend the development and maintenance of SCOP. SCOP was conceived at the MRC Laboratory of Molecular Biology, and developed in collaboration with researchers in Berkeley. Work on SCOP (version 1) concluded in June 2009 with the release of SCOP 1.75.

SCOPe classifies many newer structures through a combination of automation and manual curation, and corrects some errors in SCOP, aiming to have the same accuracy as the hand-curated SCOP releases. SCOPe also incorporates and updates the ASTRAL database.

[About SCOPe](#) [Stats & Prior Releases](#)

## News

2022-01-07: We published a paper describing the new features in SCOPe 2.08-stable. [\[PDF\]](#)

2021-09-20: SCOPe 2.08-stable has been released, with nearly 20,000 new PDB entries added since the last stable release. Important features include genetic variant search tools and annotations of structural heterogeneity and repeat units. Click either the About or Stats & History links for more details on what's new!

2018-11-30: We published a paper describing updates to SCOPe, focusing on our findings from classifying large structures. [\[PDF\]](#)

## Classes in SCOPe 2.08:

1.  a: All alpha proteins [46456] (290 folds)

## Fig2. Search for Insulin

## Folds found:

- g.1: insulin-like [56993] (1 superfamily)  
nearly all-alpha  
can be classified as disulfide-rich
- j.75: Isolated insulin B-chain [58773] (1 superfamily)

## Superfamilies found:

- g.1.1: Insulin-like [56994] (1 family) 
- j.75.1: Isolated insulin B-chain [58774] (1 family) 

## Families found:

- g.1.1.1: insulin-like [56995] (5 proteins)
- j.75.1.1: Isolated insulin B-chain [58775] (1 protein)

## Proteins found:

- Insulin gene enhancer protein isl-1 [46714] from a.4.1.1: Homeodomain (1 species)
- Insulin receptor [158901] from b.1.2.1: Fibronectin type III (1 species)
- Insulin receptor substrate 1, IRS-1 [50758] from b.55.1.2: Phosphotyrosine-binding domain (PTB) (1 species)  
*duplication: contains two domains of this fold*
- Type 1 insulin-like growth factor receptor extracellular domain [52072] from c.10.2.5: L domain (1 species)
- Insulin receptor [159452] from c.10.2.5: L domain (1 species)
- Insulin receptor substrate 1, IRS-1 [50758] from b.55.1.2: Phosphotyrosine-binding domain (PTB) (1 species)  
*PTK group: InsR subfamily; membrane spanning protein tyrosine kinase*
- Insulin-like growth factor 1 receptor [69825] from d.144.1.7: Protein kinases, catalytic subunit (1 species)  
*PTK group: InsR subfamily; membrane spanning protein tyrosine kinase*
- Insulin [56996] from g.1.1.1: Insulin-like (3 species)
- Insulin-like growth factor [57002] from g.1.1.1: Insulin-like (1 species)
- Insulin-like growth factor binding protein 6 [111415] from g.28.1.1: Thyroglobulin type-1 domain (1 species)
- Insulin-like growth factor-binding protein 1, IGFBP1 [161142] from g.28.1.1: Thyroglobulin type-1 domain (1 species)
- Insulin-like growth factor-binding protein 4, IGFBP4 [161144] from g.28.1.1: Thyroglobulin type-1 domain (1 species)

<https://scop.berkeley.edu> | [1001251](https://scop.berkeley.edu/1001251) from a.2.6.1: Albumin (1 species)

## Fig3.1. Hit page for Insulin

SCOPe [Browse](#) [Stats & History](#) [Downloads](#) [Help](#)  [Search](#)

Domains found:

- d1bzv.1: 1bzv B,A: [43942]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
superpotent single-replacement insulin analogue
- d1feaa\_1efe A: [43891]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
an active mini-proinsulin, m2pi
- d1guj.1: 1guj B,A: [70582]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
structure at pH 2: the conditions promoting insulin fibre formation  
complexed with so4
- d1guj.2: 1guj D,C: [70583]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
structure at pH 2: the conditions promoting insulin fibre formation  
complexed with so4
- d1j73.1: 1j73 B,A: [62670]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
an unstable insulin analog with native activity  
complexed with zn
- d1j73.2: 1j73 D,C: [62671]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
an unstable insulin analog with native activity  
complexed with zn
- d1jca.1: 1jca B,A: [62874]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"  
an unstable insulin analog with enhanced(?) activity  
complexed with zn
- d1jca.2: 1jca D,C: [62875]  
Insulin (g.1.1.1) from "Human (Homo sapiens) [TaxId:9606]"

Fig3.2. Hit page for Insulin

SCOPe [Browse](#) [Stats & History](#) [Downloads](#) [Help](#)  [Search](#)

Lineage for **Fold g.1: Insulin-like**

1. Root: SCOPe 2.08
2. Class g: Small proteins [56992] (100 folds)
3. Fold g.1: insulin-like [56993] (1 superfamily)  
nearly all-alpha  
can be classified as disulfide-rich

Superfamily:

g.1.1: Insulin-like [56994] (1 family)

More info for **Fold g.1: Insulin-like**

Timeline for Fold g.1: Insulin-like:

- Fold g.1: Insulin-like first appeared (with stable ids) in SCOP 1.55
- Fold g.1: Insulin-like appears in SCOPe 2.07

SCOPe: Structural Classification of Proteins — extended. Release 2.08 (September 2021)  
References: Fox NK, Brenner SE, Chandonia JM. 2014. *Nucleic Acids Research* 42:D304-309. doi: 10.1093/nar/gkt1240.  
Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. 2022. *Nucleic Acids Research* 50:D553–559. doi: 10.1093/nar/gkab1054. (citing information)  
Copyright © 1994-2022 The SCOP and SCOPe authors  
scope@compbio.berkeley.edu

Fig4. Result page for Protein Fold

**SCOPe** [Browse](#) [Stats & History](#) [Downloads](#) [Help](#) [Search \(click for examples\)](#) 

Lineage for **Superfamily g.1.1: Insulin-like**

1. Root: SCOPe 2.08
2.  Class g: Small proteins [56992] (100 folds)
3.  Fold g.1: Insulin-like [56993] (1 superfamily)
  -  nearly all-alpha  
can be classified as disulfide-rich
4.  Superfamily g.1.1: Insulin-like [56994] (1 family) 

Family:

 g.1.1.1: Insulin-like [56995] (5 proteins)

More info for **Superfamily g.1.1: Insulin-like**

Timeline for **Superfamily g.1.1: Insulin-like**:

- Superfamily g.1.1: Insulin-like first appeared (with stable ids) in SCOP 1.55
- Superfamily g.1.1: Insulin-like appears in SCOPe 2.07

SCOPe: Structural Classification of Proteins — extended. Release 2.08 (September 2021)  
 References: Fox NK, Brenner SE, Chandonia JM. 2014. *Nucleic Acids Research* 42:D304-309. doi: 10.1093/nar/gkt1240.  
 Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. 2022. *Nucleic Acids Research* 50:D553–559. doi: 10.1093/nar/gkab1054. (citing information)  
 Copyright © 1994-2022 The SCOP and SCOPe authors  
 scope@compbio.berkeley.edu

**Fig5. Result page for Protein Superfamily**

**SCOPe** [Browse](#) [Stats & History](#) [Downloads](#) [Help](#) [Search \(click for examples\)](#) 

Lineage for **Family g.1.1.1: Insulin-like**

1. Root: SCOPe 2.08
2.  Class g: Small proteins [56992] (100 folds)
3.  Fold g.1: Insulin-like [56993] (1 superfamily)
  -  nearly all-alpha  
can be classified as disulfide-rich
4.  Superfamily g.1.1: Insulin-like [56994] (1 family) 
5.  Family g.1.1.1: Insulin-like [56995] (5 proteins)

Proteins:

1.  Bombyxin-II [57004] (1 species)
  -  Species Silkworm (*Bombyx mori*) [TaxId:7091] [57005] (2 PDB entries)
2.  Insulin [56996] (3 species)
  1.  Species Cow (*Bos taurus*) [TaxId:9913] [56997] (6 PDB entries)
  2.  Species Human (*Homo sapiens*) [TaxId:9606] [56998] (63 PDB entries)
    -  Uniprot P01308
  3.  Species Pig (*Sus scrofa*) [TaxId:9823] [56999] (26 PDB entries)
3.  Insulin-like growth factor [57002] (1 species)
  -  Species Human (*Homo sapiens*) [TaxId:9606] [57003] (23 PDB entries)
    -  Uniprot P05019 49-110
4.  Relaxin [57000] (1 species)

**Fig6. Result page for Protein Family**

## Lineage for Protein: Type 1 insulin-like growth factor receptor extracellular domain

1. Root: SCOPe 2.08
2. Class c: Alpha and beta proteins (a/b) [51349] (148 folds)
3. Fold c.10: Leucine-rich repeat, LRR (right-handed beta-alpha superhelix) [52046] (3 superfamilies)
  - 2 curved layers, a/b; parallel beta-sheet; order 1234...N; there are sequence similarities between different superfamilies
4. Superfamily c.10.2: L domain-like [52058] (9 families) 
  - less regular structure consisting of variable repeats
5. Family c.10.2.5: L domain [52071] (6 proteins)
  - this is a repeat family; one repeat unit is 1n8z C:42-66 found in domain
6. Protein Type 1 insulin-like growth factor receptor extracellular domain [52072] (1 species)

## Species:

 Human (Homo sapiens) [TaxId:9606] [52073] (1 PDB entry)  
L1 and L2 domains

Domains for 1igr:

1.  Domain d1igra1: 1igr A:1-149 [30877]
  - Other proteins in same PDB: d1igra3 complexed with nag, so4
2.  Domain d1igra2: 1igr A:300-478 [30878]
  - Other proteins in same PDB: d1igra3 complexed with nag, so4

## Fig7. Result page for Protein

## Lineage for d1bzv.1 (1bzv B:,A:)

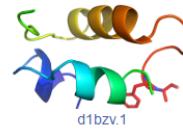
1. Root: SCOPe 2.08
2. Class g: Small proteins [56992] (100 folds)
3. Fold g.1: Insulin-like [56993] (1 superfamily)
  -  nearly all-alpha can be classified as disulfide-rich
4. Superfamily g.1.1: Insulin-like [56994] (1 family) 
5. Family g.1.1.1: Insulin-like [56995] (5 proteins)
6. Protein Insulin [56996] (3 species)
7. Species Human (Homo sapiens) [TaxId:9606] [56998] (63 PDB entries)
  - Uniprot P01308
8.  Domain d1bzv.1: 1bzv B:,A: [43942]
  - superpotent single-replacement insulin analogue

## Details for d1bzv.1

PDB Entry: 1bzv (more details)

PDB Description: [d-alab26]-des(b27-b30)-insulin-b26-amide a superpotent single-replacement insulin analogue, nmr, minimized average structure

PDB Compounds: (A:) insulin, (B:) insulin



## Fig8. Result page for Protein Domain

## RESULT:

## CATH database:

CATH database was searched for query “Insulin” and 172 matching CATH superfamilies, 8831 matching CATH domains and 3297 matching PDB structures were retrieved.

## SCOP database:

SCOPe database was searched for query “Insulin” and information for protein folds, superfamilies, families and domains was retrieved.

## CONCLUSION:

CATH and SCOP database provides user with information regarding protein folds, superfamilies, domains and structures. It is a useful tool for classification of proteins. Classification of proteins can allow researchers

to functionally annotate proteins and establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structures.

## REFERENCES:

### CATH database:

1. Jaffe, L., & Hess-Fischl Ms, R. A. D. (n.d.). *What Is Insulin?* EndocrineWeb. Retrieved February 16, 2022, from <https://www.endocrineweb.com/conditions/type-1-diabetes/what-insulin>
2. Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., ... Orengo, C. A. (2020). *CATH: increased structural coverage of functional space.* *Nucleic Acids Research.* doi:10.1093/nar/gkaa1079
3. *CATH: Protein Structure Classification Database at UCL.* (n.d.). [Www.cathdb.info](http://www.cathdb.info). Retrieved February 16, 2022, from <https://www.cathdb.info>
4. *CATH Search: Browse.* (n.d.). [Www.cathdb.info](http://www.cathdb.info). Retrieved February 16, 2022, from <https://www.cathdb.info/search?q=insulin>
5. *CATH Superfamily 1.10.100.10.* (n.d.). [Www.cathdb.info](http://www.cathdb.info). Retrieved February 16, 2022, from <https://www.cathdb.info/version/latest/superfamily/1.10.100.10>
6. *CATH Domain 3bc5A00.* (n.d.). [Www.cathdb.info](http://www.cathdb.info). Retrieved February 16, 2022, from <http://www.cathdb.info/version/latest/domain/3bc5A00c>
7. *PDB 2iug.* (n.d.). [Www.cathdb.info](http://www.cathdb.info). Retrieved February 16, 2022, from <https://www.cathdb.info/pdb/2iug>

### SCOP database:

8. Jaffe, L., & Hess-Fischl Ms, R. A. D. (n.d.). *What Is Insulin?* EndocrineWeb. Retrieved February 16, 2022, from <https://www.endocrineweb.com/conditions/type-1-diabetes/what-insulin>
9. Xiong, J. (2008). *Protein Structure Visualization, Comparison, and Classification. Essential bioinformatics.* Cambridge: Cambridge University Press. 195-197.
10. Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N., & Brenner, S. (2021). SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research*, 50(D1), D553–D559. <https://doi.org/10.1093/nar/gkab1054>
11. *SCOPe 2.08: Structural Classification of Proteins — extended.* (n.d.). [Scop.berkeley.edu](http://scop.berkeley.edu). Retrieved February 21, 2022, from <https://scop.berkeley.edu/search/?ver=2.08&key=insulin>
12. *SCOPe 2.08: Fold g.1: Insulin-like.* (n.d.). [Scop.berkeley.edu](http://scop.berkeley.edu). Retrieved February 21, 2022, from <https://scop.berkeley.edu/sunid=56993>
13. *SCOPe 2.08: Superfamily g.1.1: Insulin-like.* (n.d.). [Scop.berkeley.edu](http://scop.berkeley.edu). Retrieved February 21, 2022, from <https://scop.berkeley.edu/sunid=56994>
14. *SCOPe 2.08: Family g.1.1.1: Insulin-like.* (n.d.). [Scop.berkeley.edu](http://scop.berkeley.edu). Retrieved February 21, 2022, from <https://scop.berkeley.edu/sunid=56995>
15. *SCOPe 2.08: Protein: Type 1 insulin-like growth factor receptor extracellular domain.* (n.d.). [Scop.berkeley.edu](http://scop.berkeley.edu). Retrieved February 21, 2022, from <https://scop.berkeley.edu/sunid=52072>
16. *SCOPe 2.08: Domain d1bzv.1: 1bzv B:,A:* (n.d.). [Scop.berkeley.edu](http://scop.berkeley.edu). Retrieved February 21, 2022, from <https://scop.berkeley.edu/sunid=43942>

## 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.

Currently, it takes 1 to 3 years to solve a protein structure. Certain proteins, especially membrane proteins, are extremely difficult to solve by x-ray or NMR techniques. There are many important proteins for which the sequence information is available, but their three-dimensional structures remain unknown. The full understanding of the biological roles of these proteins requires knowledge of their structures. Hence, the lack of such information hinders many aspects of the analysis, ranging from protein function and ligand binding to mechanisms of enzyme catalysis. Therefore, it is often necessary to obtain approximate protein structures through computer modelling.

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 modeling and prediction. They are homology modeling, threading, and ab initio prediction.

### HOMOLOGY MODELING

As the name suggests, homology modeling predicts protein structures based on sequence homology with known structures. It is also known as comparative modeling. 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 modeling produces an all-atom model based on alignment with template proteins.

The overall homology modeling procedure consists of six steps. The first step is template selection, which involves identification of homologous sequences in the protein structure database to be used as templates for modeling. The second step is alignment of the target and template sequences. The third step is to build a framework structure for the target protein consisting of main chain atoms. The fourth step of model building includes the addition and optimization of side chain atoms and loops. The fifth step is to refine and optimize the entire model according to energy criteria. The final step involves evaluating of the overall quality of the model obtained. If necessary, alignment and model building are repeated until a satisfactory result is obtained.

A number of comprehensive modeling programs are able to perform the complete procedure of homology modeling 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 modeled 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 modeling of loops in protein structures.

Comparative modeling consists of five main steps: Searching for structures related to query, selecting template target-template alignment, model building, and model evaluation.

## 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 modeling. In cases where no appropriate template is identified by LOMETS, I-TASSER will build the whole structures by ab initio modeling. 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 restraints 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.

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.

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 modeling error. A TM-score  $>0.5$  indicates a model of correct topology and a TM-score  $<0.17$  means a random similarity. These cutoff does not depends on the protein length.

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 modeling when the native structure is known, while C-score is a metric that I-TASSER developed to estimate the confidence of the modeling. In case where the native structure is not known, it becomes necessary to predict the quality of the modeling prediction, i.e. what is the distance between the predicted model and the native structures? To answer this question, we tried predicted 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 modeling or de 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.

## **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](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:

- (i) the prediction of transmembrane helices using TMHMM
- (ii) low-complexity regions assigned by the program SEG
- (iii) coiled-coils prediction using COILS
- (iv) the prediction of disordered regions using DISOPRED
- (v) secondary structure predictions using PSIPRED, SAM-T99, Jufo and Jufo3D
- (vi) the results listed above, domain predictions and the NR PSI-BLAST multiple sequence alignment used for the last step in the domain prediction protocol condensed into an image to help corroborate the domain prediction results;
- (vii) domain repeats prediction using REPRO predicted boundaries are given if repeats are detected;
- (viii) 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 modeling 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

**WEBLEM 3a****MODELLER**

(URL: <https://salilab.org/modeller/>)

**AIM:**

To perform tertiary structure prediction by Comparative Modeling/Homology Modeling method using Modeller for query Kinase.

**INTRODUCTION:**

Kinase, an enzyme that adds phosphate groups (PO<sub>4</sub>3<sup>-</sup>) to other molecules. For protein targets, kinases can phosphorylate the amino acids serine, threonine, and tyrosine. Phosphorylation of lipid molecules by kinases is important for controlling the molecular composition of membranes in cells, which helps to specify the physical and chemical properties of the different membranes. Nucleotides, the fundamental units of RNA (ribonucleic acid) and DNA (deoxyribonucleic acid), contain a phosphate molecule attached to a nucleoside, a compound made up of a ribose moiety and a purine or pyrimidine base. The tertiary structure of kinase can be predicted using Modeller.

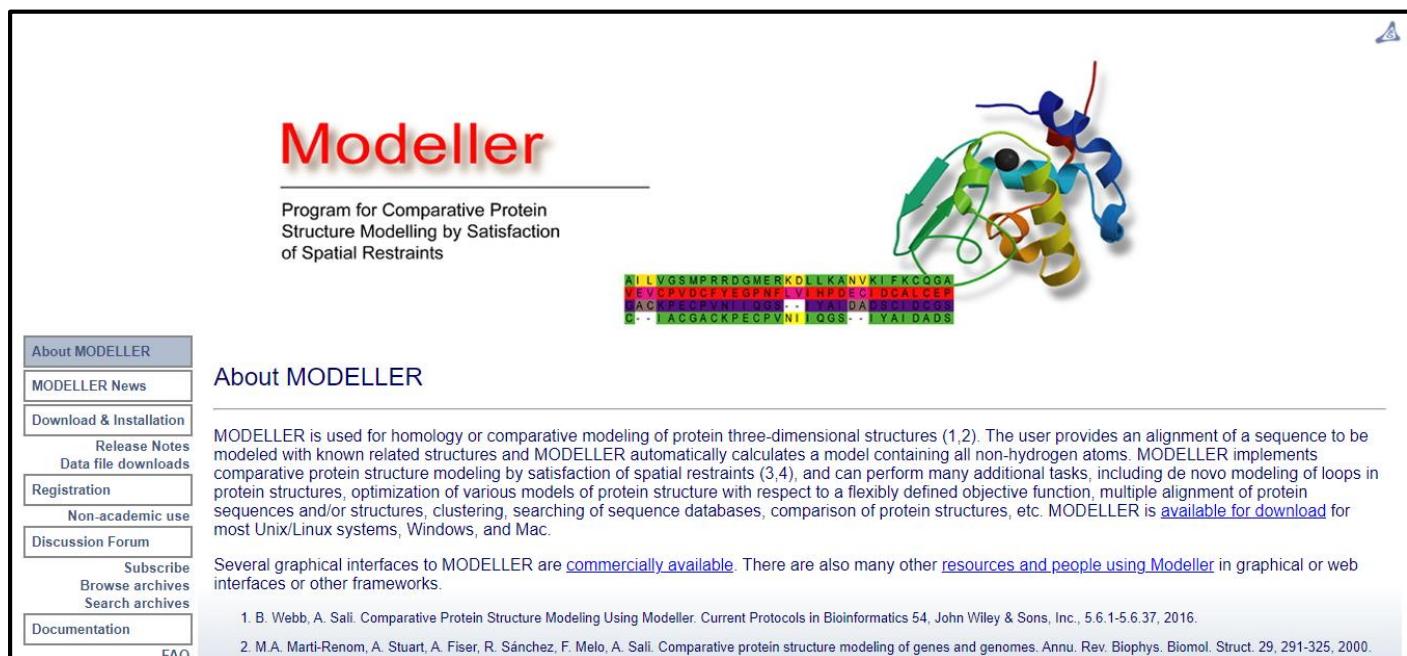
MODELLER is a computer program for comparative protein structure modelling. In the simplest case, the input is an alignment of a sequence to be modeled with the template structures, the atomic coordinates of the templates, and a simple script file. MODELLER then automatically calculates a model containing 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 modeling of loops in protein structures.

Comparative modeling consists of five main steps: Searching for structures related to query, selecting template target-template alignment, model building, and model evaluation.

**METHODOLOGY:**

1. Install modeller. (URL: <https://salilab.org/modeller/>)
2. Retrive FASTA sequence for enzyme kinase.
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:



**Modeller**  
Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

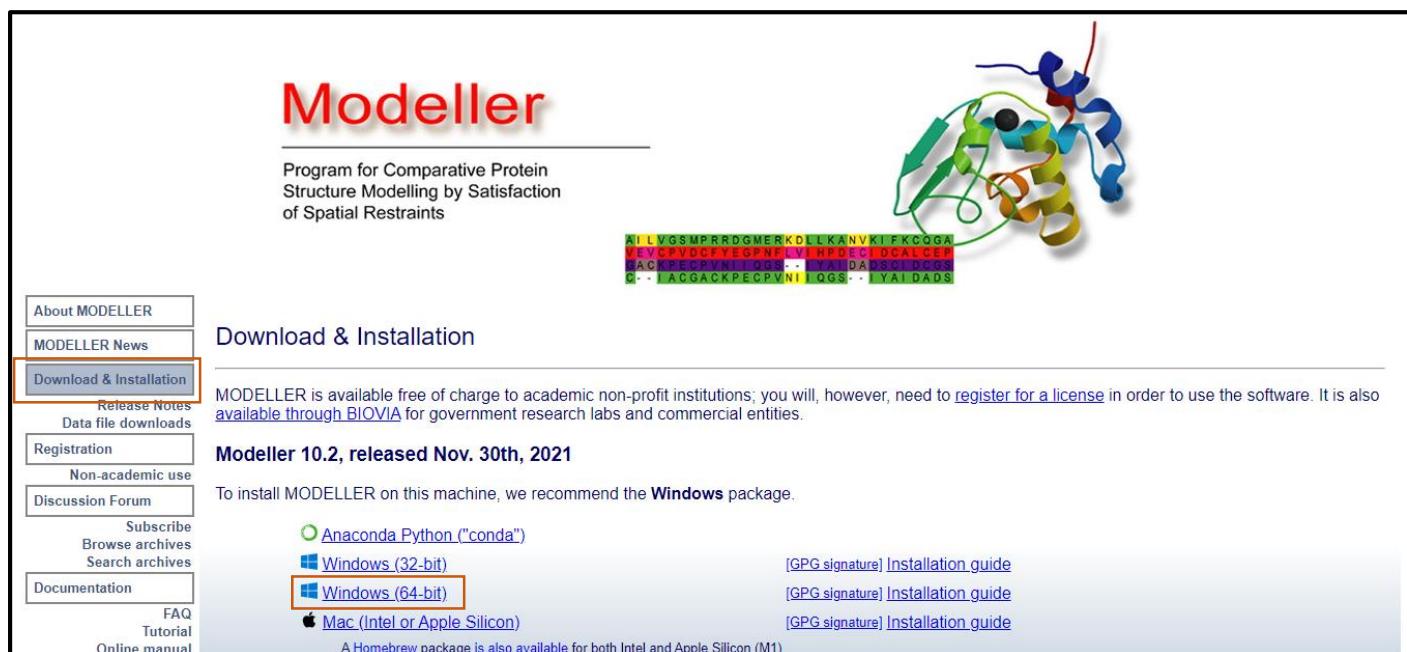
**About MODELLER**

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is [available for download](#) for most Unix/Linux systems, Windows, and Mac.

Several graphical interfaces to MODELLER are [commercially available](#). There are also many other [resources and people using Modeller](#) in graphical or web interfaces or other frameworks.

1. B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics 54. John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.  
2. M.A. Martí-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.

**Fig1. Homepage for Modeller**



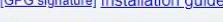
**Modeller**  
Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

**Download & Installation**

MODELLER is available free of charge to academic non-profit institutions; you will, however, need to [register for a license](#) in order to use the software. It is also [available through BIOVIA](#) for government research labs and commercial entities.

**Modeller 10.2, released Nov. 30th, 2021**

To install MODELLER on this machine, we recommend the **Windows** package.

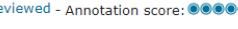
 <a href="#">Anaconda Python ("conda")</a>	 <a href="#">[GPG signature] Installation guide</a>
 <a href="#">Windows (64-bit)</a>	 <a href="#">[GPG signature] Installation guide</a>
 <a href="#">Mac (Intel or Apple Silicon)</a>	 <a href="#">[GPG signature] Installation guide</a>

A Homebrew package is also available for both Intel and Apple Silicon (M1)

**Fig2. Page to install Modeller**

UniProtKB - P52564 (MP2K6\_HUMAN)

Display [Help video](#) [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#) [Basket](#) [Add a publication](#) [Feedback](#)

**Protein** Dual specificity mitogen-activated protein kinase kinase 6  
**Gene** MAP2K6  
**Organism** Homo sapiens (Human)  
**Status**  Reviewed - Annotation score:  - Experimental evidence at protein level<sup>1</sup>

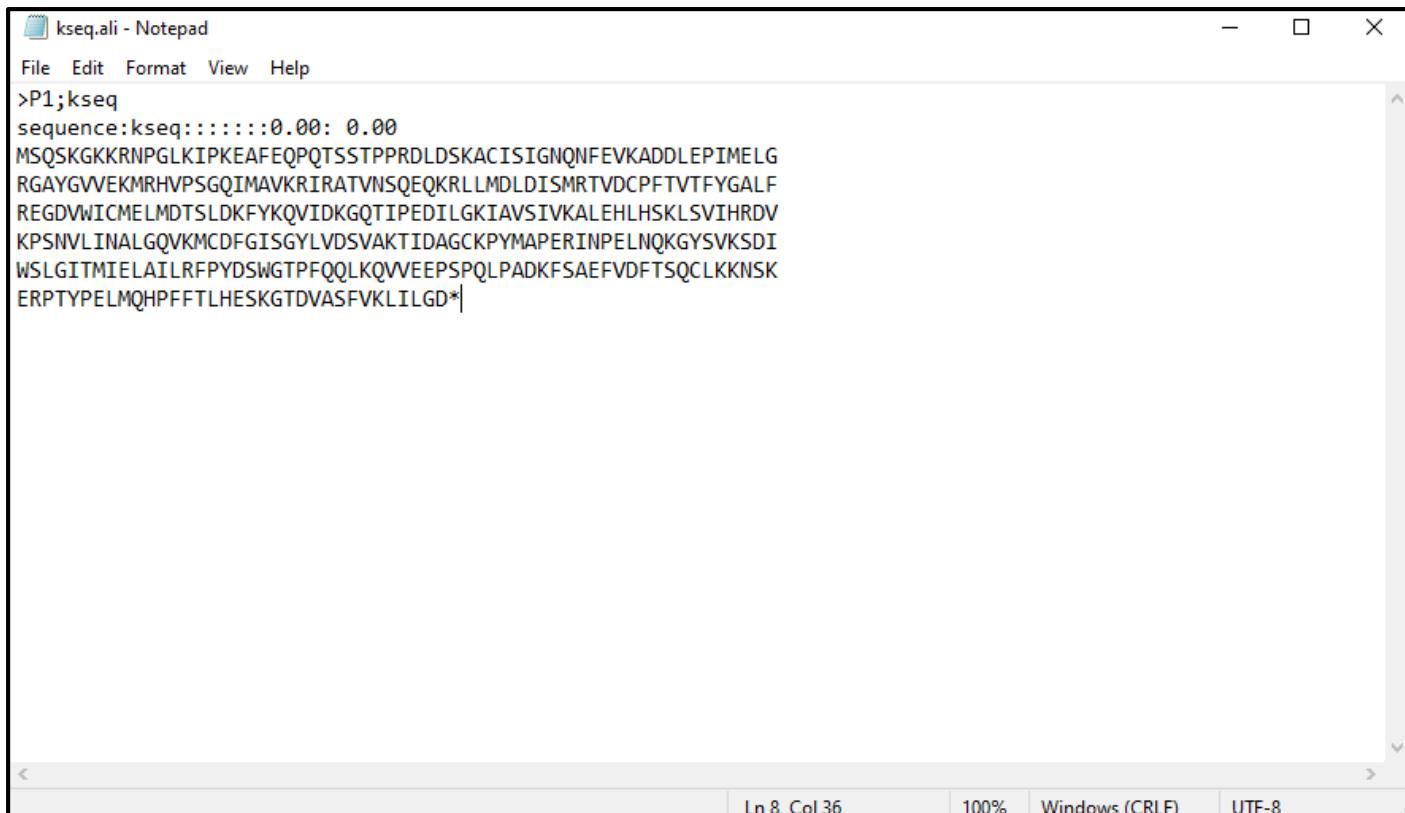
**Function**<sup>1</sup>

Dual specificity protein kinase which acts as an essential component of the MAP kinase signal transduction pathway. With MAP3K3/MKK3, catalyzes the concomitant phosphorylation of a threonine and a tyrosine residue in the MAP kinases p38 MAPK11, MAPK12, MAPK13 and MAPK14 and plays an important role in the regulation of cellular responses to cytokines and all kinds of stresses. Especially, MAP2K3/MKK3 and MAP2K6/MKK6 are both essential for the activation of MAPK11 and MAPK13 induced by environmental stress, whereas MAP2K6/MKK6 is the major MAPK11 activator in response to TNF. MAP2K6/MKK6 also phosphorylates and activates PAK6. The p38 MAP kinase signal transduction pathway leads to direct activation of transcription factors. Nuclear targets of p38 MAP kinase include the transcription factors ATF2 and ELK1. Within the p38 MAPK signal transduction pathway, MAP3K6/MKK6 mediates phosphorylation of STAT4 through MAPK14 activation, and is therefore required for STAT4 activation and STAT4-regulated gene expression in response to IL-12 stimulation. The pathway is also crucial for IL-6-induced SOCS3 expression and down-regulation of IL-6-mediated gene induction; and for IFNG-dependent gene transcription. Has a role in osteoclast differentiation through NF- $\kappa$ B.

Fig3. Result page for kinase in UniProt database

```
>sp|P52564|MP2K6_HUMAN Dual specificity mitogen-activated protein kinase kinase 6 OS=Homo sapiens OX=9606 GN=MAP2K6 PE=1 SV=1
MSQSKGKRNPGKLKIPKEAFAEQPOTSSPPRDLDSKACISIGHQNEFEKADDLEPTMELG
RGAYGVVEKIRHVPSSQIMAVKRIRATVNSQEQRLLWLDISHRTDCCPTVTFVGALF
REGDWVICHELMDTSLDKFYKQVIDKGQTIPEDILGKIAVSVKALELHHSKLSVIHRDV
KPSNVLINALQVKMDFGIGSYLVDSVAKTIDAGCKPYMAPERINPELNQKGSVKSDI
WSLGTTMIELA1LRFPYDSIGTPQLQKVEEPSPQLPADKFSAEFVDFTSQCLKKNSK
ERPTYPELIMQHPFTLHESKGTDVASFVKLILGD
```

Fig4. FASTA sequence for kinase



kseq.ali - Notepad

File Edit Format View Help

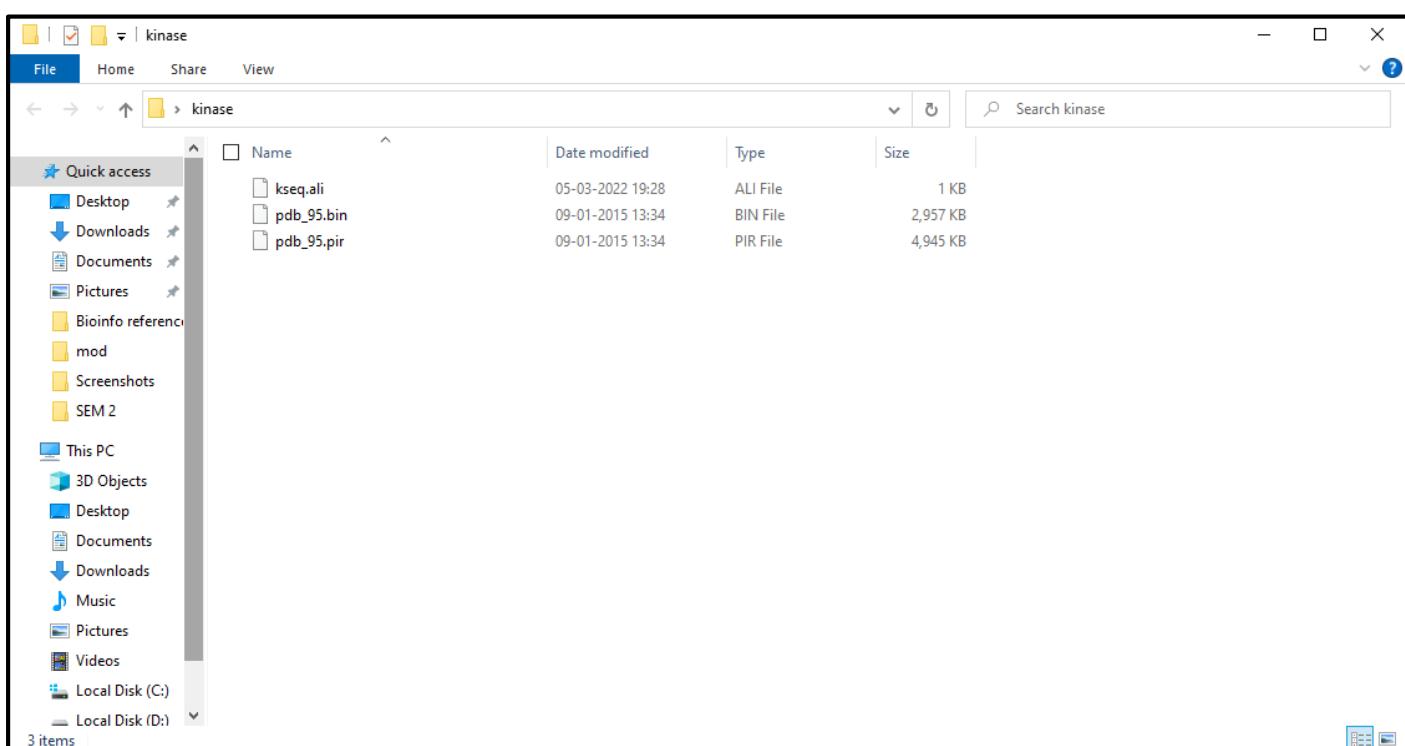
>P1;kseq

sequence:kseq::::::::::0.00: 0.00

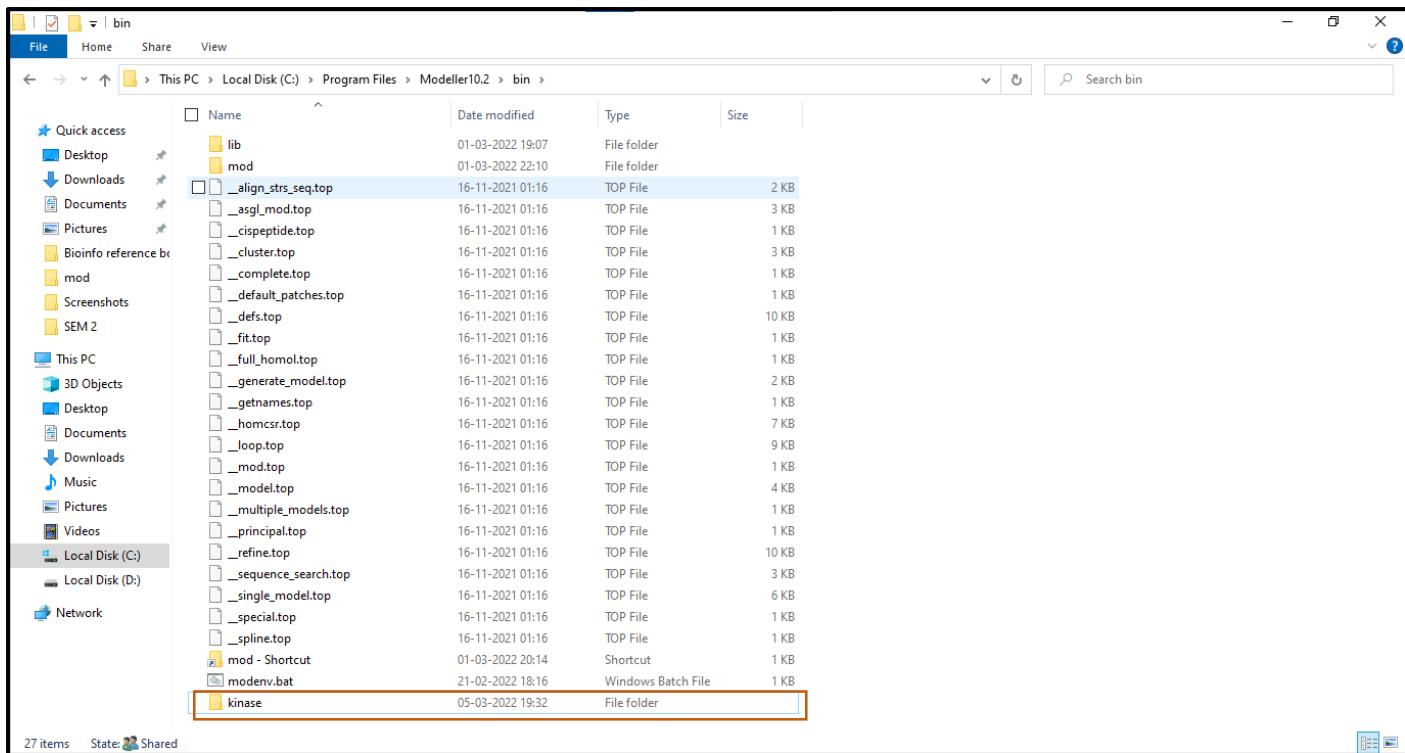
MSQSKGKKRNPGKIPKEAFEQPQTSTPPRDLDSKACISIGNQNFEVKADDLEPIMELG  
RGAYGVVEKMRHVPSGQIMAVKRIRATVNSQEQQKRLLMDLDISMRDVTDCPFTVTFYGALF  
REGDWICMELMDTSLDKFYKQVIDKGQTIPEDILGKIAVSIVKALEHLHSKLSVIHRDV  
KPSNVLINALGQVKMCDFGISGYLVDSVAKTIDAGCKPYMAPERINPELNQKGYSVKSIDI  
WSLGITMIELAILRFPYDSWGTFFQQLKQWEEPSPQLPADKFSAEFVDFTSQLKKNSK  
ERPTYPELMQHPFTLHESKGTDVASFVKLILGD\*

Ln 8, Col 36 100% Windows (CRLF) UTF-8

**Fig5. Target sequence in PIR format**



**Fig6. Target sequence saved in .ali format**



**Fig7. Kinase folder saved in bin folder of Modeller**

A screenshot of a Modeller command line interface. The text shows the following commands being run:

```

Modeller
You can find many useful example scripts in the
examples\automerol directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.2' to run them instead.

C:\Program Files\Modeller10.2>cd bin
C:\Program Files\Modeller10.2\bin>cd kinase
C:\Program Files\Modeller10.2\bin\kinase>_

```

**Fig8. Setting working directory on Modeller command line**

```

script1.py - Notepad
File Edit Format View Help

--- Read in the sequence database
sdb = SequenceDB(env)
sdb.read(seq_database_file='pdb_95.pin', seq_database_format='PIR',
         chains_list='ALL', minmax_db_seq_len=(30, 4000), clean_sequences=True)

--- Write the sequence database in binary form
sdb.write(seq_database_file='pdb_95.bin', seq_database_format='BINARY',
          chains_list='ALL')

--- Now, read in the binary database
sdb.read(seq_database_file='pdb_95.bin', seq_database_format='BINARY',
          chains_list='ALL')

--- Read in the target sequence/alignment
aln = Alignment(env)
aln.append(file='kseq.ali', alignment_format='PIR', align_codes='ALL')

--- Convert the input sequence/alignment into
# profile format
prf = aln.to_profile()

--- Scan sequence database to pick up homologous sequences
prf.build(sdb, matrix_offset=-450, rr_file='${LIB}/blosum62.sim.mat',
          gap_penalties_1d=(-500, -50), n_prof_iterations=1,
          check_profile=False, max_aln_evalue=0.01)

--- Write out the profile in text format
prf.write(file='build_profile.prf', profile_format='TEXT')

--- Convert the profile back to alignment format
aln = prf.to_alignment()

--- Write out the alignment file
aln.write(file='build_profile.ali', alignment_format='PIR')

```

**Fig9. Python script for searching for structures related to kinase**

```

Modeller
You can find many useful example scripts in the
examples\automer directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.2' to run them instead.

C:\Program Files\Modeller10.2>cd bin
C:\Program Files\Modeller10.2\bin>cd kinase
C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script1.py
import site' failed; use -v for traceback
C:\Program Files\Modeller10.2\bin\kinase>

```

**Fig10. Running script1.py**

```

script1.log - Notepad
File Edit Format View Help
MODELLER 10.2, 2021/11/15, r12267
PROTEIN STRUCTURE MODELLING BY SATISFACTION OF SPATIAL RESTRAINTS

Copyright(c) 1989-2021 Andrey 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-EP80I3R, 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): 2022/03/05 19:41:50

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

```

Fig11. Log file for script1

```

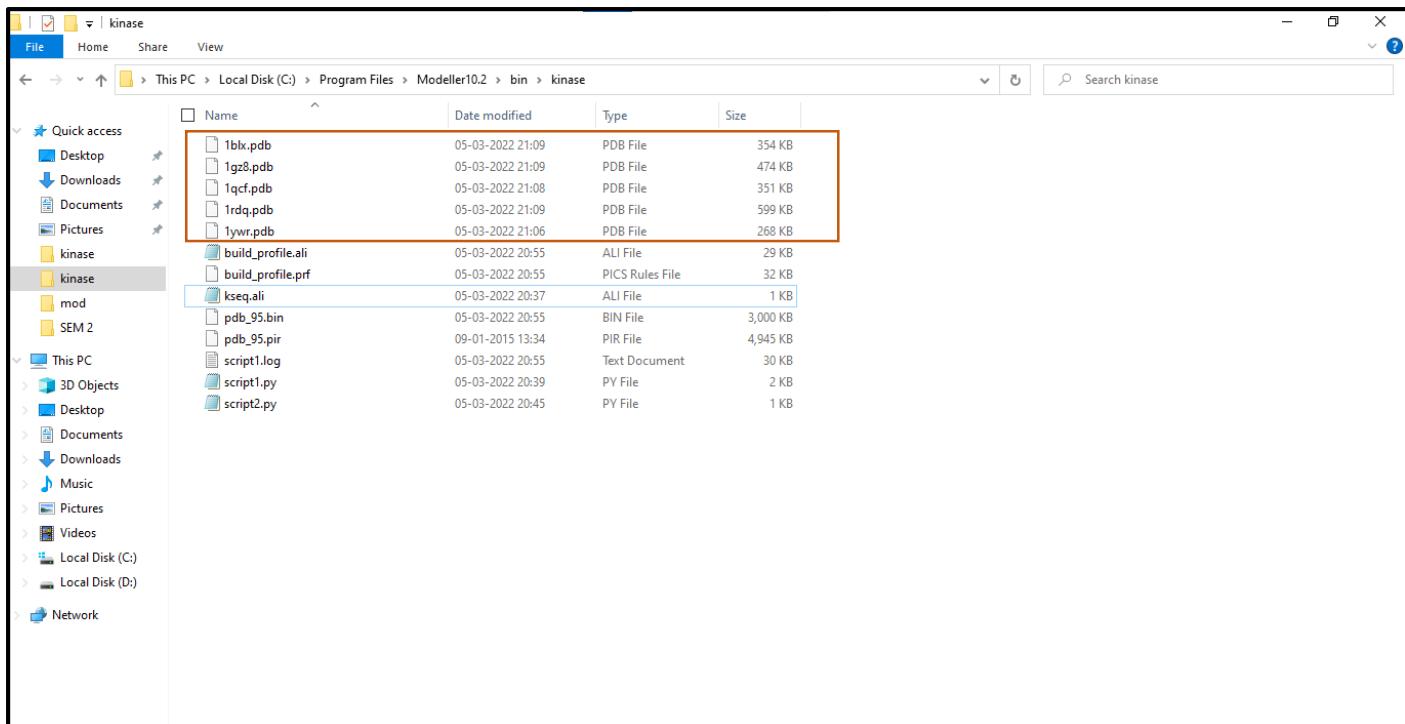
script1.log - Notepad
File Edit Format View Help
Z: 1 7.95000 0.00000 0.00003
Z: 1 8.05000 0.00000 0.00002

HITS FOUND IN ITERATION: 1

Dynamically allocated memory at amaxprofile [B,KiB,MiB]: 892034 871.127 0.851
> 1a06 1 43 7900 279 334 28.81 0.28E-10 2 228 49 313 3 238
> 1ywrA 1 270 8700 338 334 29.12 0.0 3 245 52 325 22 306
> 1qcfA 1 347 10100 449 334 28.81 0.0 4 286 7 320 151 445
> 1fgkA 1 421 8300 278 334 25.22 0.37E-11 5 220 79 310 39 268
> 1rdqE 1 609 8800 340 334 29.29 0.0 6 227 51 302 33 271
> 1gz8A 1 624 11150 290 334 29.96 0.0 7 250 51 314 2 278
> 1vqyA 1 907 7200 299 334 32.74 0.13E-08 8 102 149 251 96 208
> 2bfxA 1 1121 8200 270 334 26.29 0.62E-11 9 248 51 313 6 256
> 1blxA 1 1187 11300 305 334 28.22 0.0 10 252 54 314 10 296
> 2bikB 1 1204 7900 272 334 24.49 0.28E-10 11 238 59 313 12 256
> 1uu3A 1 1206 8900 277 334 26.23 0.0 12 237 59 314 16 259
> 1mg4A 1 1316 11450 261 334 29.30 0.0 13 252 51 319 6 261
> 1bygA 1 1510 6950 246 334 28.38 0.38E-08 14 225 52 310 8 236
> 1ck1A 1 1804 5950 292 334 25.10 0.93E-06 15 233 58 310 14 264
> 1cm8A 1 1847 8050 327 334 29.17 0.16E-10 16 187 124 323 91 306
> 1om1A 1 2136 7600 325 334 23.96 0.17E-09 17 249 51 314 31 318
> 1pme 1 2878 8700 333 334 34.67 0.0 18 138 124 277 82 231
> 1f3mC 1 3067 14300 287 334 33.47 0.0 19 239 58 316 27 268
> 1fmk 1 3385 7750 437 334 28.69 0.11E-09 20 243 46 318 179 429
> 1fotA 1 3435 10550 299 334 30.65 0.0 21 197 52 259 5 203
> 1opjA 1 3450 9450 287 334 28.02 0.0 22 248 46 310 12 268
> 1fvra 1 3525 7200 299 334 26.80 0.13E-08 23 226 59 309 18 267
> 1ir3A 1 3719 8600 300 334 25.00 0.12E-11 24 258 46 314 9 288
> 1gjoA 1 3818 8300 280 334 25.00 0.37E-11 25 255 46 310 8 271
> 1j1bA 1 3869 9050 354 334 24.73 0.0 26 247 59 314 28 306
> 1o61A 1 4037 9650 316 334 27.00 0.0 27 256 47 314 1 263
> 1oiuC 1 4061 11850 265 334 30.04 0.0 28 250 51 314 3 255
> 1unlA 1 4075 10550 292 334 26.15 0.0 29 250 54 315 5 287
> 1q8vA 1 4291 5750 351 334 25.12 0.34E-05 30 197 125 332 99 301

```

Fig11.1. Hits found for similar structures



**Fig12. Five structures download in PDB format**

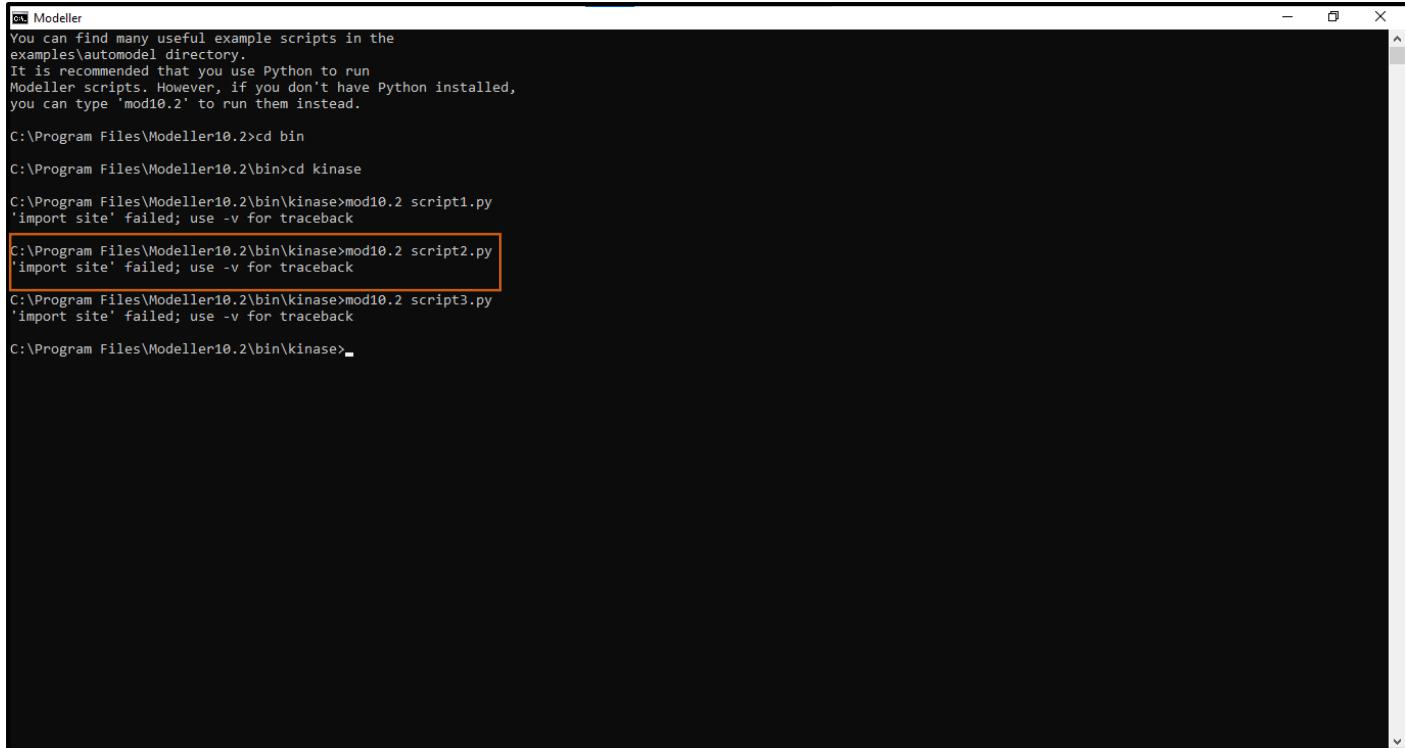
```

script2.py - Notepad
File Edit Format View Help
from modeller import *

env = Environ()
aln = Alignment(env)
for (pdb, chain) in (('1ywr', 'A'), ('1qcf', 'A'), ('1rdq', 'E'),
                     ('1gzb', 'A'), ('1bbx', 'A')):
    m = Model(env, file=pdb, model_segment=('FIRST:'+chain, 'LAST:' + chain))
    aln.append_model(m, atom_files=pdb, align_codes=pdb+chain)
aln.align()
aln.align3d()
aln.compare_structures()
aln.id_table(matrix_file='family.mat')
env.dendrogram(matrix_file='family.mat', cluster_cut=-1.0)

```

**Fig13. Python script for selecting a template**



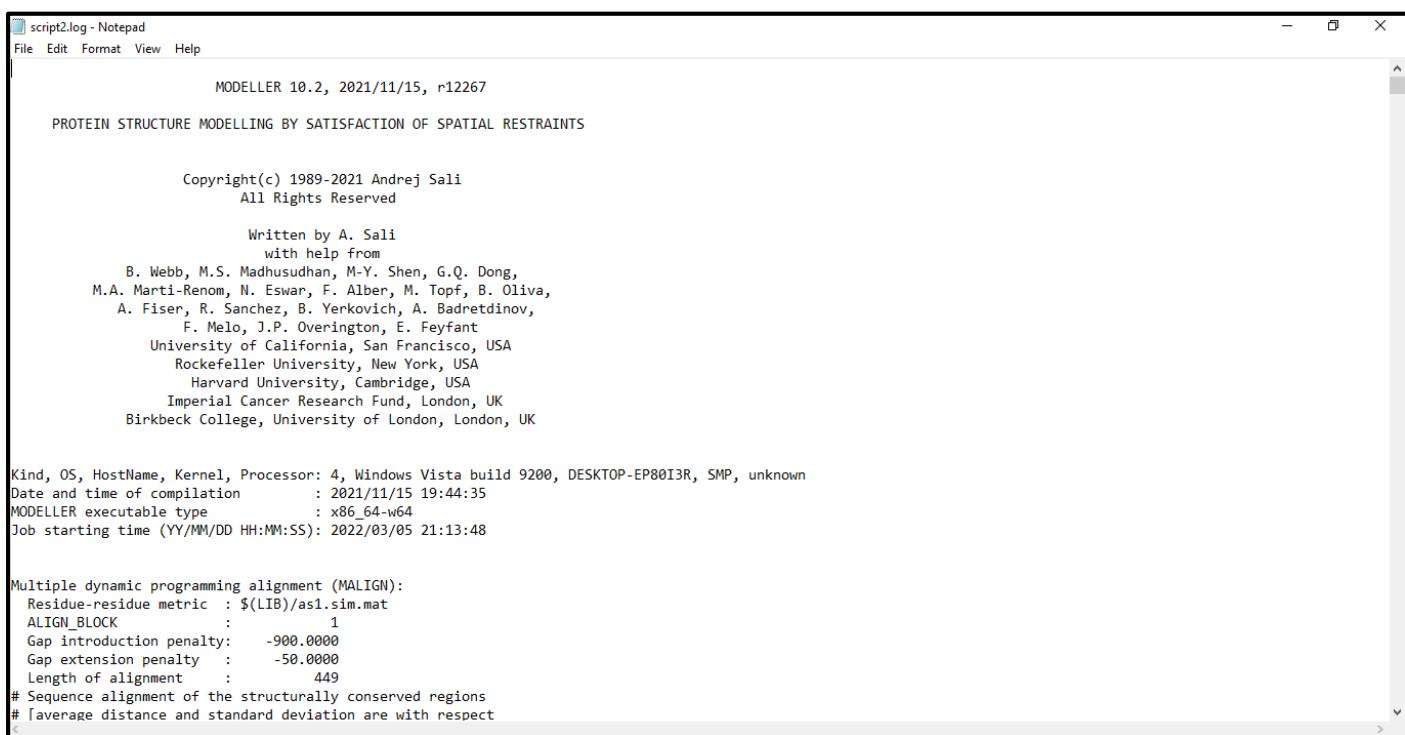
```

Modeller
You can find many useful example scripts in the
examples\automodel directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.2' to run them instead.

C:\Program Files\Modeller10.2>cd bin
C:\Program Files\Modeller10.2\bin>cd kinase
C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script1.py
'import site' failed; use -v for traceback
C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script2.py
'import site' failed; use -v for traceback
C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script3.py
'import site' failed; use -v for traceback
C:\Program Files\Modeller10.2\bin\kinase>

```

**Fig14. Running script2**



```

script2.log - Notepad
File Edit Format View Help
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-EP80I3R, 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): 2022/03/05 21:13:48

Multiple dynamic programming alignment (MALIGN):
Residue-residue metric : ${LIB}/asl1.sim.mat
ALIGN_BLOCK      :      1
Gap introduction penalty: -900.0000
Gap extension penalty  :  -50.0000
Length of alignment   :      449
# Sequence alignment of the structurally conserved regions
# (average distance and standard deviation are with respect
<

```

**Fig15. Log file for script2**

```

script2.log - Notepad
File Edit Format View Help
Sequence identity comparison (ID_TABLE):

Diagonal ... number of residues;
Upper triangle ... number of identical residues;
Lower triangle ... % sequence identity, id/min(length).

1ywrA @21qcfA @21rdqE @11gz8A @11blxA @1
1ywrA @2      338      1     14     80     77
1qcfA @2      0      449      0      2      3
1rdqE @1      4      0     340      6      6
1gz8A @1      28      1      2    290    108
1blxA @1      25      1      2     37    305

Weighted pair-group average clustering based on a distance matrix:

      ----- 1ywrA @2.0  73.5000
      |----- 1gz8A @1.3  63.0000
      |----- 1blxA @1.9  97.0000
      |----- 1rdqE @1.3  99.7500
      ----- 1qcfA @2.0

+-----+-----+-----+-----+-----+-----+
101.2200 94.6050 87.9900 81.3750 74.7600 68.1450 61.5300
97.9125 91.2975 84.6825 78.0675 71.4525 64.8375

Total CPU time [seconds] : 0.83

```

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

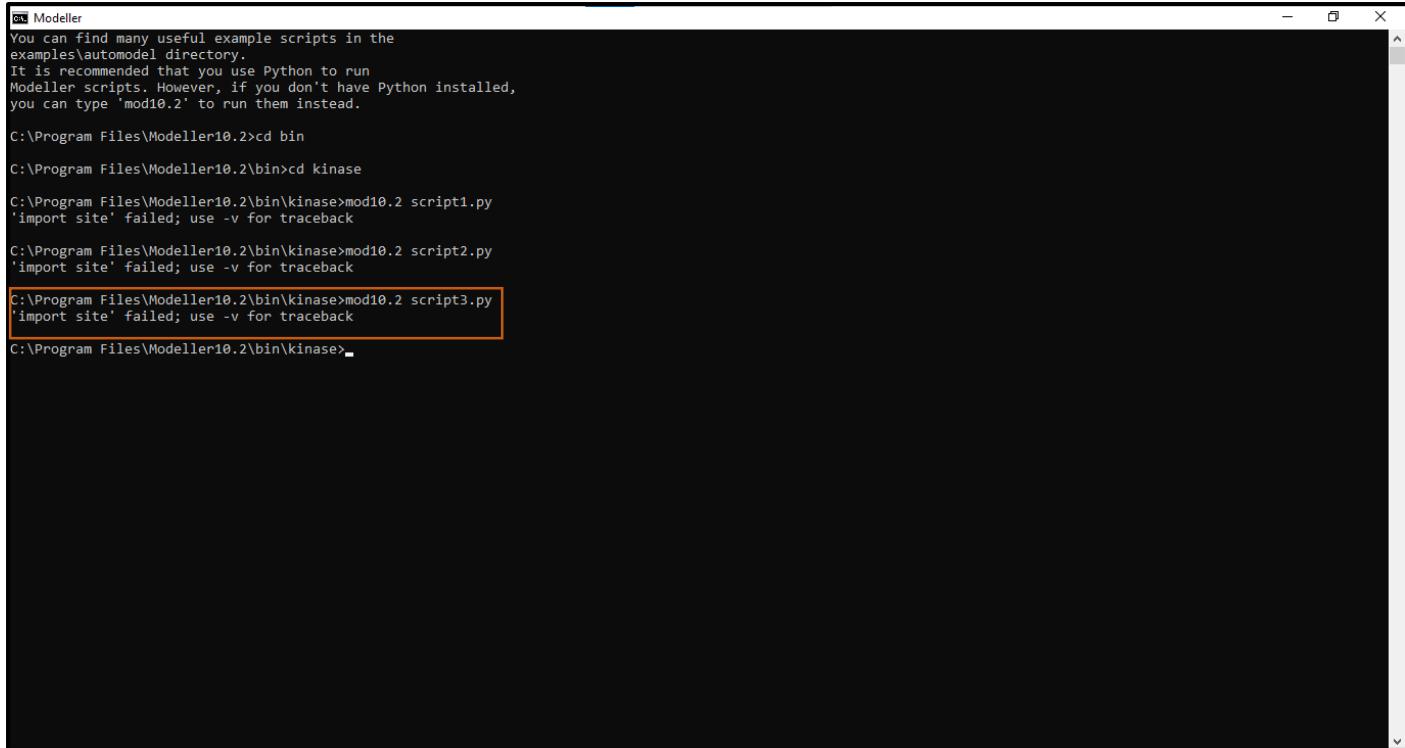
```

script3.py - Notepad
File Edit Format View Help
from modeller import *

env = Environ()
aln = Alignment(env)
mdl = Model(env, file='1rdq', model_segment=('FIRST:E','LAST:E'))
aln.append_model(mdl, align_codes='1rdqE', atom_files='1rdq.pdb')
aln.append(file='kseq.ali', align_codes='kseq')
aln.align2d(max_gap_length=50)
aln.write(file='kseq-1rdqE.ali', alignment_format='PIR')
aln.write(file='kseq-1rdqE.pap', alignment_format='PAP')

```

**Fig16. Python script for aligning query with the template**



```
Modeller
You can find many useful example scripts in the
examples\automodel directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.2' to run them instead.

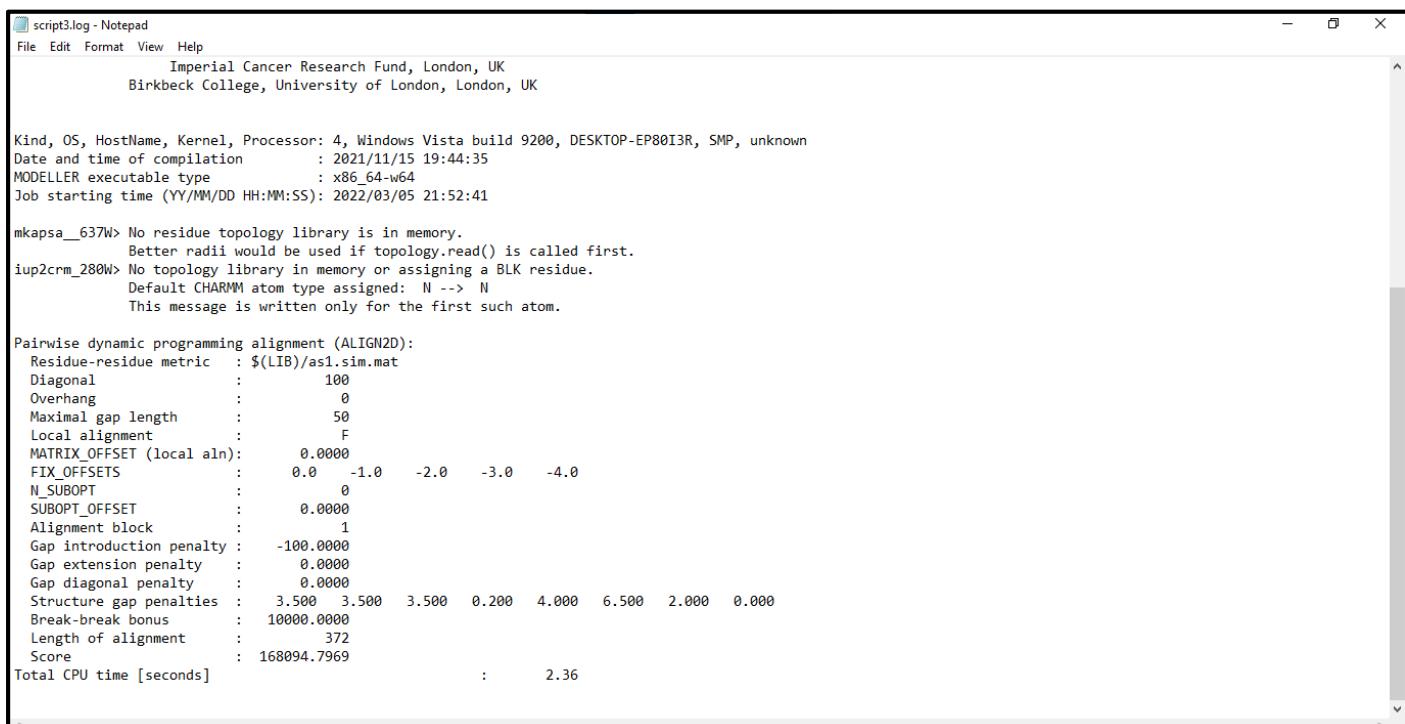
C:\Program Files\Modeller10.2>cd bin
C:\Program Files\Modeller10.2\bin>cd kinase
C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script1.py
'import site' failed; use -v for traceback

C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script2.py
'import site' failed; use -v for traceback

C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script3.py
'import site' failed; use -v for traceback

C:\Program Files\Modeller10.2\bin\kinase>
```

Fig17. Running script3



```
script3.log - Notepad
File Edit Format View Help
Imperial Cancer Research Fund, London, UK
Birkbeck College, University of London, London, UK

Kind, OS, HostName, Kernel, Processor: 4, Windows Vista build 9200, DESKTOP-EP80I3R, 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): 2022/03/05 21:52:41

mkapsa_637W> No residue topology library is in memory.
      Better radii would be used if topology.read() is called first.
iup2crm_280W> No topology library in memory or assigning a BLK residue.
      Default CHARMM atom type assigned: N --> N
      This message is written only for the first such atom.

Pairwise dynamic programming alignment (ALIGN2D):
  Residue-residue metric : $(LIB)/asl.sim.mat
  Diagonal : 100
  Overhang : 0
  Maximal gap length : 50
  Local alignment : F
  MATRIX_OFFSET (local aln): 0.0000
  FIX_OFFSETS : 0.0 -1.0 -2.0 -3.0 -4.0
  N_SUBOPT : 0
  SUBOPT_OFFSET : 0.0000
  Alignment block : 1
  Gap introduction penalty : -100.0000
  Gap extension penalty : 0.0000
  Gap diagonal penalty : 0.0000
  Structure gap penalties : 3.500 3.500 3.500 0.200 4.000 6.500 2.000 0.000
  Break-break bonus : 10000.0000
  Length of alignment : 372
  Score : 168094.7969
  Total CPU time [seconds] : 2.36
```

Fig17. Log file for script3

```

kseq-1rdqE.pap - Notepad
File Edit Format View Help
_pos 10 20 30 40 50 60
_1rdqE  GNAAASVKE--FLAKAKEDFLK--KWETP----S-----QN-TAQLDQFDRIKTLGTGSGRVM
kseq  MSQSKGKKRNPGLKIPKEAEQPOQTSSTPPRDLDSKACISIGNQFEVKADDLEPIMELGRGAYGVVE
_consrvd * * *** * ** * ** * * * *** * *
_pos 70 80 90 100 110 120 130
1rdqE  LVKHKESGNHAYAMKILDQKVVKLQIEHTLNNEKRILQAVNFPFLVKLEFSFKDNSNLYMMVEYVAGG
kseq  KMRHVPMSGQIMAVKRI-RATVNSQEQRLLMDLDISMRTVDCPFTVTFYGALFREGDWICMELMDTS
_consrvd * * *** * * * * * * *
_pos 140 150 160 170 180 190 200
1rdqE  --EMFSHLRRIG-RFSEPHARFYAAQIVLTFEYLHS-LDLIYRDLKPENLLIDQGGYIQVTDGFAKR
kseq  LDKFYKQVIDKGQTIPEDILGKIAVSITVKALEHLHSKLSVIRDVKPSNVLINALQVVKMCFGISGY
_consrvd * * *** * * * * * * *
_pos 210 220 230 240 250 260 270
1rdqE  VK--GRTWLCGTPEALAPEIILS---KGYNKAVDWALGVLIYEMAAGYPPFA-DQPIQIYEKIV
kseq  LVDSVAKTIDAGCKPYMAPERINPELNQKGYSVKSDIWSLGITMIELA1LRFPYDSWGTPEQQLQVW
_consrvd * * *** * * * * * *
_pos 280 290 300 310 320 330 340
1rdqE  SG-KVRFPS-HFSSDLKDLRNLQVDLTKRFGNLKNGVNDIKNHKWFAATTDWIAIYQRKVEAPFIPK
kseq  EEPSPQLPADKESAEFVDFTSQCLKKNISKERPT---YPELMQHPFTLHESKGT---DV-ASFVKL
_consrvd * * * * *
_pos 350 360 370
1rdqE  FKGPGDTSNFDDYEEEIRVINEKGKEFTF
kseq  ILG--D-----
_consrvd * *

```

Fig18. Sequence alignment

```

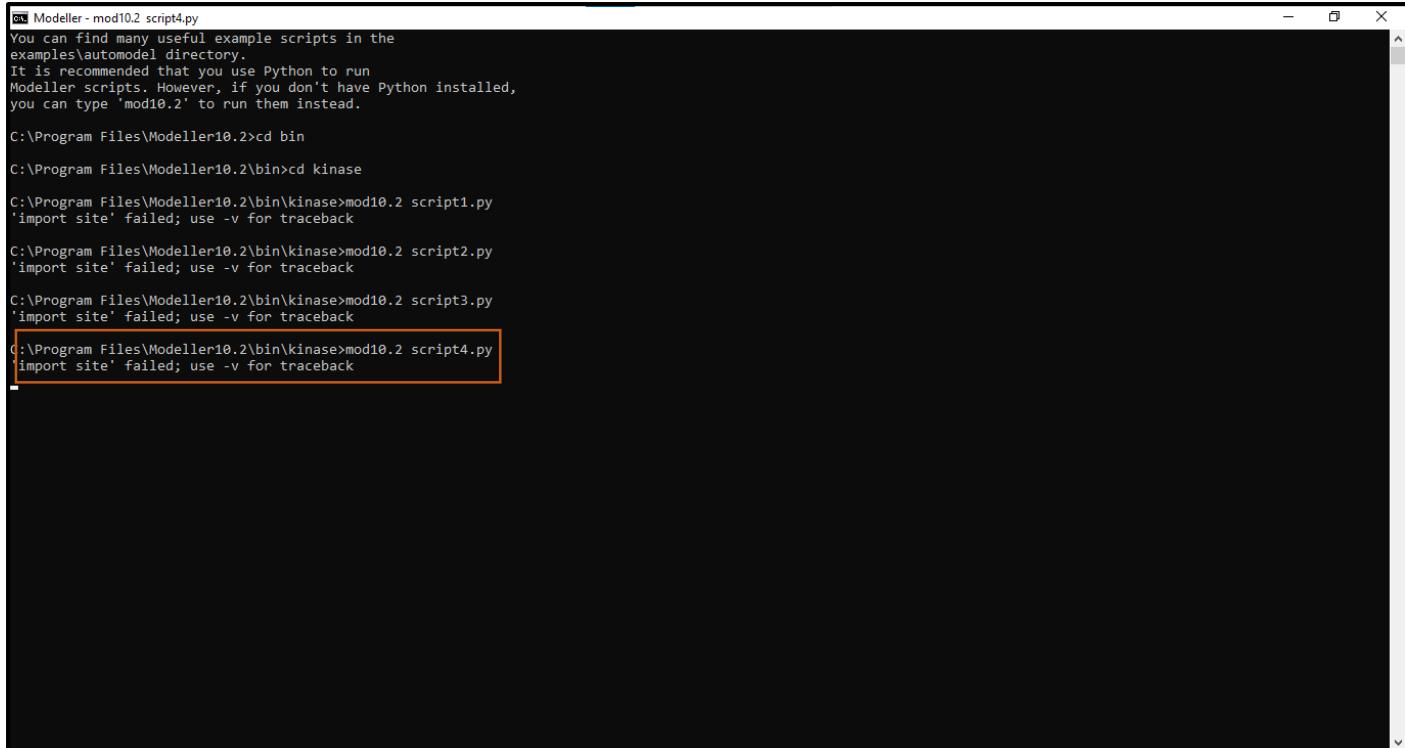
script4.py - Notepad
File Edit Format View Help
from modeller import *
from modeller.automodel import *
#from modeller import soap_protein_od

env = Environ()
a = AutoModel(env, alnfile='kseq-1rdqE.ali',
              knowns='1rdqE', sequence='kseq',
              assess_methods=(assess.DOPE,
                             #soap_protein_od.Scorer(),
                             assess.GA341))

a.starting_model = 1
a.ending_model = 5
a.make()

```

Fig19. Python script for model building



```
Modeller - mod10.2 script4.py
You can find many useful example scripts in the
examples\automodel directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.2' to run them instead.

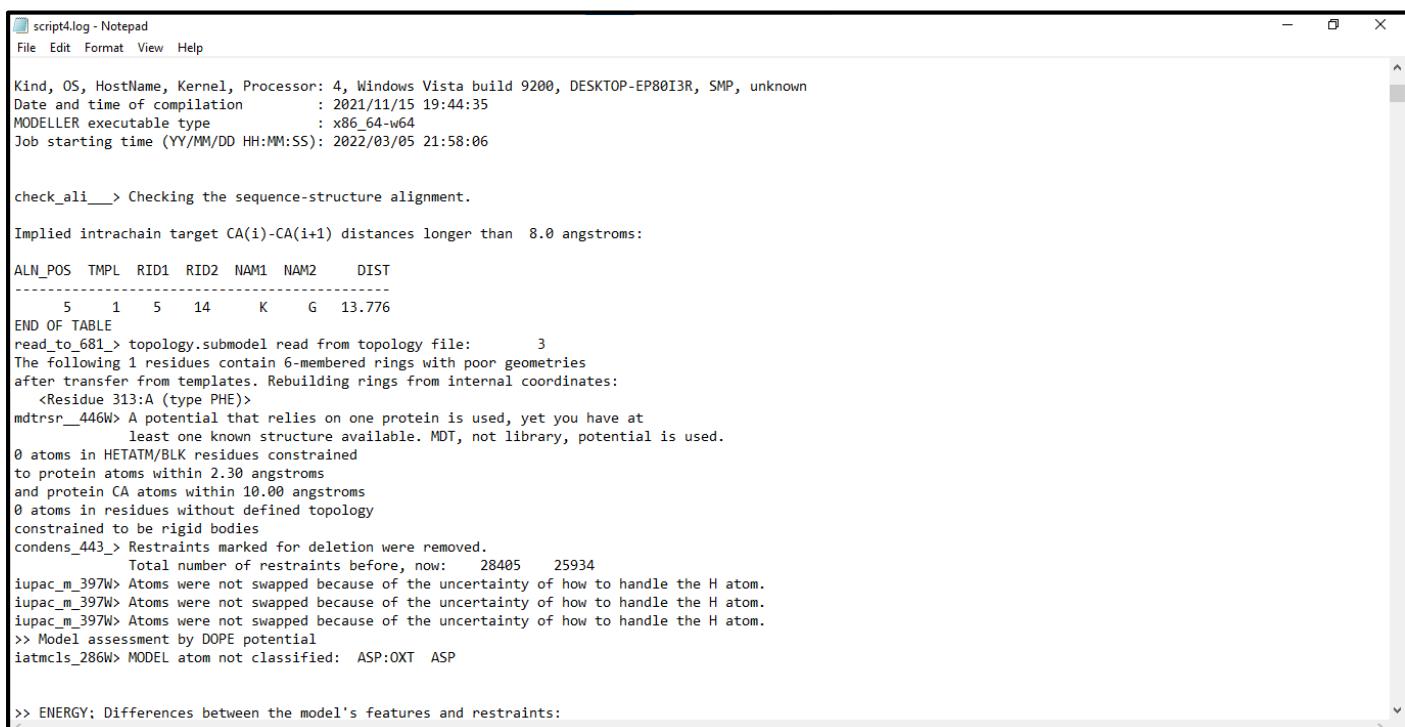
C:\Program Files\Modeller10.2>cd bin
C:\Program Files\Modeller10.2\bin>cd kinase
C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script1.py
'import site' failed; use -v for traceback

C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script2.py
'import site' failed; use -v for traceback

C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script3.py
'import site' failed; use -v for traceback

C:\Program Files\Modeller10.2\bin\kinase>mod10.2 script4.py
'import site' failed; use -v for traceback
```

Fig20. Running script4



```
script4.log - Notepad
File Edit Format View Help

Kind, OS, HostName, Kernel, Processor: 4, Windows Vista build 9200, DESKTOP-EP80I3R, 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): 2022/03/05 21:58:06

check_ali__> Checking the sequence-structure alignment.

Implied intrachain target CA(i)-CA(i+1) distances longer than 8.0 angstroms:

ALN_POS TMPL RID1 RID2 NAM1 NAM2 DIST
----- 5 1 5 14 K G 13.776
END OF TABLE
read_to_681_> topology.submodel read from topology file: 3
The following 1 residues contain 6-membered rings with poor geometries
after transfer from templates. Rebuilding rings from internal coordinates:
  <Residue 313:A (type PHE)>
mdtrsr_446W> A potential that relies on one protein is used, yet you have at
              least one known structure available. MDT, not library, potential is used.
0 atoms in HETATM/BLK residues constrained
to protein atoms within 2.30 angstroms
and protein CA atoms within 10.00 angstroms
0 atoms in residues without defined topology
constrained to be rigid bodies
condens_443_> Restraints marked for deletion were removed.
  Total number of restraints before, now: 28405 25934
iupac_m_397W> Atoms were not swapped because of the uncertainty of how to handle the H atom.
iupac_m_397W> Atoms were not swapped because of the uncertainty of how to handle the H atom.
iupac_m_397W> Atoms were not swapped because of the uncertainty of how to handle the H atom.
>> Model assessment by DOPE potential
iatmcls_286W> MODEL atom not classified: ASP:OXT ASP

>> ENERGY: Differences between the model's features and restraints:
<
```

Fig21. Log file for script4

```

script4.log - Notepad
File Edit Format View Help
51 9349 303P 304T C N 2380 2382 -116.22 -63.20 63.99 11.03 -63.20 63.99 11.03
51 304T 304T N CA 2382 2383 -77.92 -42.10 -42.10
52 9350 304T 305Y C N 2387 2389 -126.92 -63.50 76.50 14.48 -63.50 76.50 14.48
52 305Y 305Y N CA 2389 2390 -86.18 -43.40 -43.40
53 9363 317H 318E C N 2504 2506 -67.43 -63.60 3.85 0.62 -117.80 -175.61 8.17
53 318E 318E N CA 2506 2507 -40.58 -40.30 -136.80
54 9364 318E 319S C N 2513 2515 80.68 -64.10 159.30 16.89 -64.10 159.30 16.89
54 319S 319S N CA 2515 2516 -101.45 -35.00 -35.00
55 9365 319S 320K C N 2519 2521 -60.37 -70.20 10.56 0.85 -62.90 174.96 22.58
55 320K 320K N CA 2521 2522 144.26 149.40 -40.80
56 9367 321G 322T C N 2532 2534 -80.36 -78.10 83.64 3.65 -63.20 109.64 13.00
56 322T 322T N CA 2534 2535 66.19 149.80 -42.10

report_____> Distribution of short non-bonded contacts:

DISTANCE1: 0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.40
DISTANCE2: 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.40 3.50
FREQUENCY: 0 0 0 0 20 42 179 218 286 313 327 428 476 437

<< end of ENERGY.

>> Summary of successfully produced models:
Filename molpdf DOPE score GA341 score
-----
kseq.B99990001.pdb 2175.30371 -33837.22656 1.00000
kseq.B99990002.pdb 2077.15063 -33663.25000 1.00000
kseq.B99990003.pdb 2211.22461 -33844.13281 1.00000
kseq.B99990004.pdb 2191.18750 -32973.11719 1.00000
kseq.B99990005.pdb 2449.14087 -33137.49609 1.00000

Total CPU time [seconds] : 158.08

```

**Fig22. Structure with low DOPE score selected as final model**

## RESULT:

Modeller was used to predict the tertiary structure of Kinase.

## 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. *MAP2K6 - Dual specificity mitogen-activated protein kinase 6 - Homo sapiens (Human) - MAP2K6 gene & protein.* (n.d.). [Www.uniprot.org](http://www.uniprot.org/uniprot/P52564). Retrieved March 8, 2022, from <https://www.uniprot.org/uniprot/P52564>
3. *kinase / Definition, Biology, & Function.* (n.d.). Encyclopedia Britannica. Retrieved March 8, 2022, from <https://www.britannica.com/science/kinase>

## WEBLEM 3b

### I-TASSER

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

#### AIM:

To perform tertiary structure prediction by Threading approach using I-TASSER server for query Kinase.

#### INTRODUCTION:

Kinase, an enzyme that adds phosphate groups (PO<sub>4</sub><sup>3-</sup>) to other molecules. For protein targets, kinases can phosphorylate the amino acids serine, threonine, and tyrosine. Phosphorylation of lipid molecules by kinases is important for controlling the molecular composition of membranes in cells, which helps to specify the physical and chemical properties of the different membranes. Nucleotides, the fundamental units of RNA (ribonucleic acid) and DNA (deoxyribonucleic acid), contain a phosphate molecule attached to a nucleoside, a compound made up of a ribose moiety and a purine or pyrimidine base. The tertiary structure of kinase can be predicted using 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.

#### 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:

The screenshot shows the UniProtKB result page for P52564 (MP2K6\_HUMAN). The top navigation bar includes links for UniProtKB, Advanced search, and a search bar. Below the navigation is a banner for the new UniProt website. The main content area displays the following information:

- Entry:** P52564 (MP2K6\_HUMAN)
- Protein:** Dual specificity mitogen-activated protein kinase kinase 6
- Gene:** MAP2K6
- Organism:** Homo sapiens (Human)
- Status:** Reviewed - Annotation score: 5/5 - Experimental evidence at protein level

**Function:**

Dual specificity protein kinase which acts as an essential component of the MAP kinase signal transduction pathway. With MAP3K3/MKK3, catalyzes the concomitant phosphorylation of a threonine and a tyrosine residue in the MAP kinases p38 MAPK11, MAPK12, MAPK13 and MAPK14 and plays an important role in the regulation of cellular responses to cytokines and all kinds of stresses. Especially, MAP2K3/MKK3 and MAP2K6/MKK6 are both essential for the activation of MAPK11 and MAPK13 induced by environmental stress, whereas MAP2K6/MKK6 is the major MAPK11 activator in response to TNF. MAP2K6/MKK6 also phosphorylates and activates PAK6. The p38 MAP kinase signal transduction pathway leads to direct activation of transcription factors. Nuclear targets of p38 MAP kinase include the transcription factors ATF2 and ELK1. Within the p38 MAPK signal transduction pathway, MAP3K6/MKK6 mediates phosphorylation of STAT4 through MAPK14 activation, and is therefore required for STAT4 activation and STAT4-regulated gene expression in response to IL-12 stimulation. The pathway is also crucial for IL-6-induced SOCS3 expression and down-regulation of IL-6-mediated gene induction; and for IFNG-dependent gene transcription. Has a role in osteoclast differentiation through NF-kappa-B.

**Fig1. Result page for kinase in UniProt database**

```

>sp|P52564|MAP2K6_HUMAN Dual specificity mitogen-activated protein kinase kinase 6 OS=Homo sapiens OX=9606 GN=MAP2K6 PE=1 SV=1
MSQSGKKRNPGLKIPKEAFEQPQTSSTPRDLDSKACISIGNQNFVEKADDLEPTMELG
RGAYGVVEKIRHVPSGQTMAVKRIRATVNSQEQKRLLIDLDISWRTVDCPFTVTFVGAFL
REGDGVWICHELDITSLDKFYKQVIDKGQTIPEDILGKIAVSIVKALELHSKLSVIRHDV
KPSNWLINALQGVKNCDFGIGSYLVDSVAKTIDAGCKPYMAPERINPENQKGSVSKSDI
WSLGTTMIELA1LRFFYDSWGTTFQLQKVVEEPSPOLPADKFSAEFVDFTSQCLKKNISK
ERTPTYPELIQHPFFTILHESKGTDVASFVKLLLD

```

Fig2. FASTA sequence for kinase




Home Research COVID-19 Services Publications People Teaching Job Opening News Forum Lab Only

**I-TASSER**  
Protein Structure & Function Predictions

(The server completed predictions for 674755 proteins submitted by 163285 users from 158 countries)  
(The template library was updated on 2022/03/06)

I-TASSER (Iterative Threading ASSEMBly Refinement) is a hierarchical approach to protein structure prediction and structure-based function annotation. It first identifies structural templates from the PDB by multiple threading approach **LOMETS**, with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through protein function database **BioLiP**. I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide **CASP7**, **CASP8**, **CASP9**, **CASP10**, **CASP11**, **CASP12**, **CASP13**, and **CASP14** experiments. It was also ranked the best for function prediction in **CASP9**. The server is in active development with the goal to provide the most accurate protein structure and function predictions using state-of-the-art algorithms. The server is only for non-commercial use. Please report problems and questions at [I-TASSER message board](#) and our developers will study and answer the questions accordingly. [\(More about the server...\)](#)

[Structure models for the SARS-CoV2 Coronavirus genome by C-I-TASSER](#) new

[\[Queue\]](#) [\[Forum\]](#) [\[Download\]](#) [\[Search\]](#) [\[Registration\]](#) [\[Statistics\]](#) [\[Remove\]](#) [\[Potential\]](#) [\[Decoys\]](#) [\[News\]](#) [\[Annotation\]](#) [\[About\]](#) [\[FAQ\]](#)

I-TASSER On-line Server ([View an example of I-TASSER output](#)):

Copy and paste your sequence within [10, 1500] residues in **FASTA** format. [Click here for a sample input](#).

Fig3. Homepage for I-TASSER

- C-QUARK
- LOMETS
- COACH
- COFACTOR
- MetaGO
- MUSTER
- CETHreader
- SEGMER
- FG-MD
- ModRefiner
- REMO
- DEMO
- SPRING
- COTH
- Threpp
- BSpred
- ANGLOR
- EDock
- BSP-SLIM
- SAXSTER
- Upred
- Threedom
- ThreedomEx
- EvoDesign

(The template library was updated on 2022/03/06)

I-TASSER (Iterative Threading ASSEmby Refinement) is a hierarchical approach to protein structure prediction and structure-based function annotation. It first identifies structural templates from the PDB by multiple threading approach [LOMETS](#), with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through protein function database [BioLiP](#). I-TASSER (as Zhang-Server) was ranked as the No 1 server for protein structure prediction in recent community-wide [CASP7](#), [CASP8](#), [CASP9](#), [CASP10](#), [CASP11](#), [CASP12](#), [CASP13](#), and [CASP14](#) experiments. It was also ranked the best for function prediction in [CASP9](#). The server is in active development with the goal to provide the most accurate protein structure and function predictions using state-of-the-art algorithms. The server is only for non-commercial use. Please report problems and questions at [I-TASSER message board](#) and our developers will study and answer the questions accordingly. ([More about the server...](#))

---

[Structure models for the SARS-CoV2 Coronavirus genome by C-I-TASSER](#) NEW

[\[Queue\]](#) [\[Forum\]](#) [\[Download\]](#) [\[Search\]](#) [\[Registration\]](#) [\[Statistics\]](#) [\[Remove\]](#) [\[Potential\]](#) [\[Decoys\]](#) [\[News\]](#) [\[Annotation\]](#) [\[About\]](#) [\[FAQ\]](#)

---

**I-TASSER On-line Server** ([View an example of I-TASSER output](#)):

Copy and paste your sequence within [10, 1500] residues in [FASTA format](#). [Click here for a sample input](#).

```
>sp|P52564|MP2K6_HUMAN Dual specific mitogen-activated protein kinase
kinase 6 OS=Homo sapiens OX=9606 GN=MAP2K6 PE=1 SV=1
MSQSKGKKNRPGKLIPKAEFPQQTSSTPPRDLKSCAISIGNQNFEVKADDLEPIMELG
RGAYGVEKIRHVPSPQIZIAVKIRKATAVNSQEQKRLNLDLDIRHTVDCPFTVFGALF
REGDWICHIELMDTSLDFKFYKQVIDKGQTIPEDILGKIAVSVKALEHLHSKLSVIRHDV
KPSNVLINALQVKMCDFGISGVLDSQVAKTIDAGCKPYWAPERINPFLNGQGYSVSKSDI
WSLGITHMELAIIQLRFPYDSWGTPOQQLKQVVEEPSPQLPADKFSAEFVDFTSQCLKNSK
ERPTYPELMQHPFFTLHESKGTDASFVKLILGD
```

Or upload the sequence from your local computer:

[Choose file](#) [No file chosen](#)

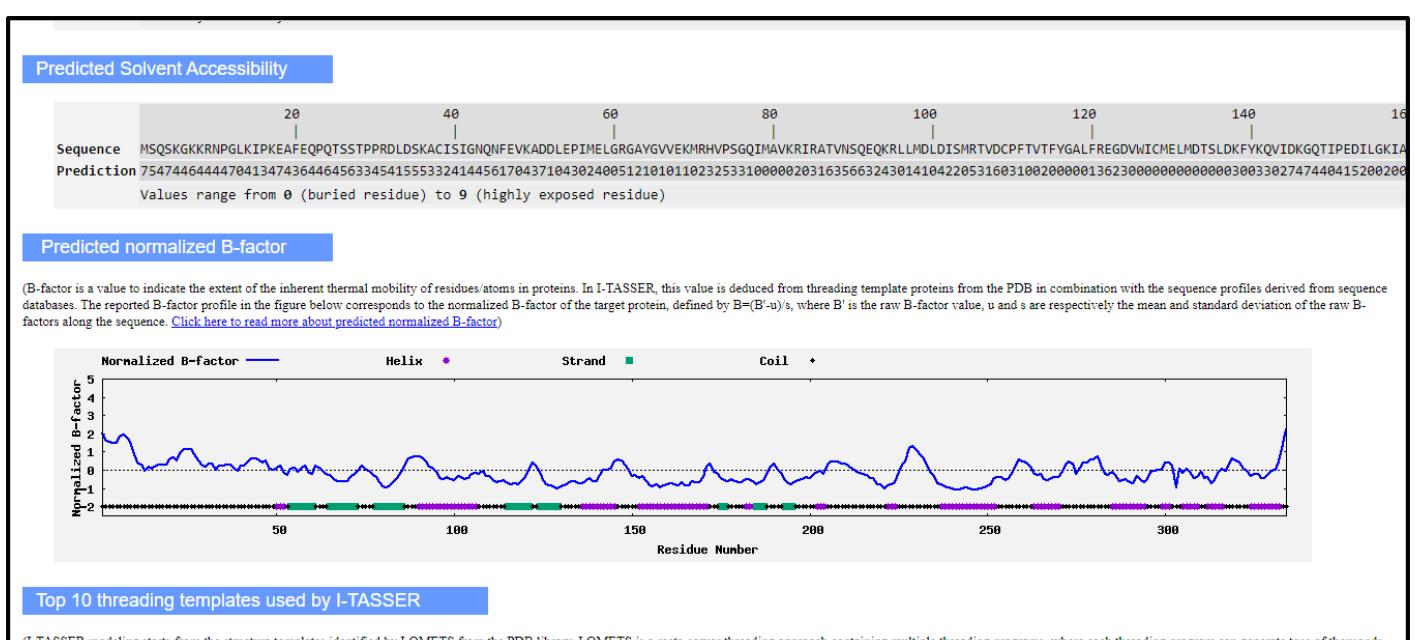
Email: (mandatory, where results will be sent to)

Password: (mandatory, please [click here](#) if you do not have a password)

<https://zhangserver.org/I-TASSER/news.html>

#### **Fig4. Submission of query**

**Fig5. Result for predicted secondary structure**



**Fig6. Result for predicted solvent accessibility and normalized B-factor**



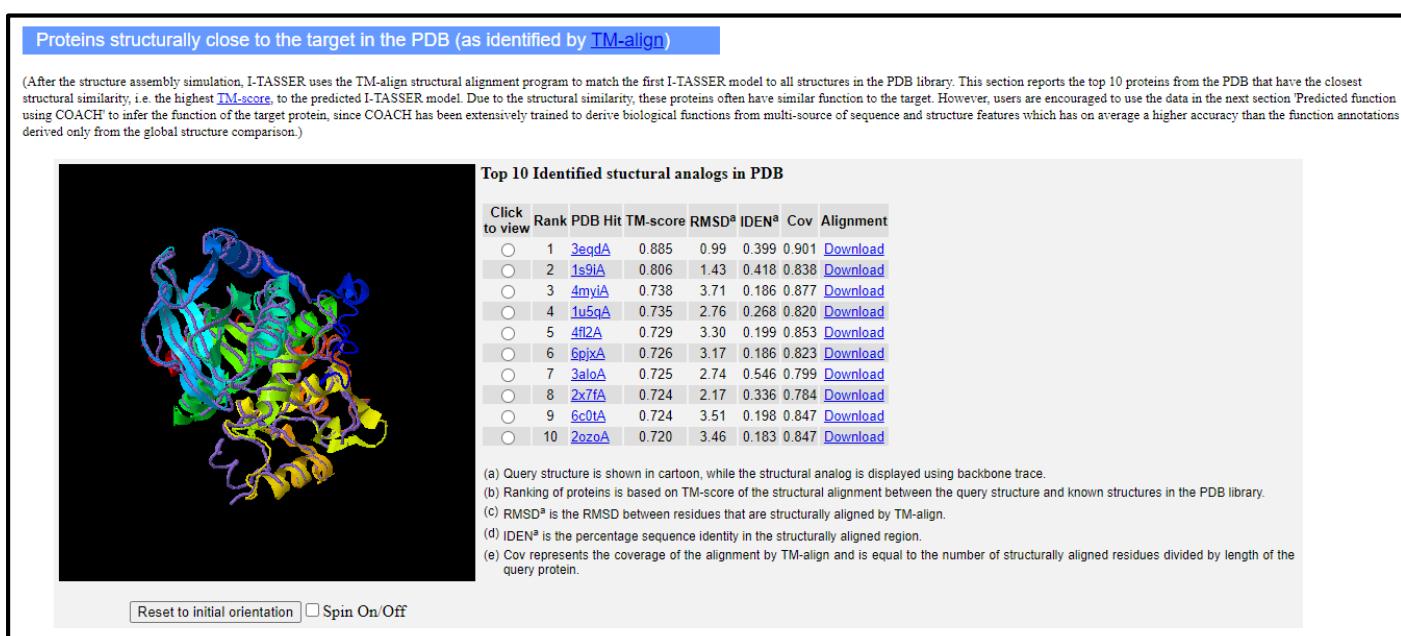
**Fig7. Result for top 10 threading templates**

- a) All the residues are colored in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in color. Coloring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are colored in dark shade. (more about the colors used)
  - b) Rank of templates represents the top ten threading templates used by I-TASSER.
  - c) Ident1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.
  - d) Ident2 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**



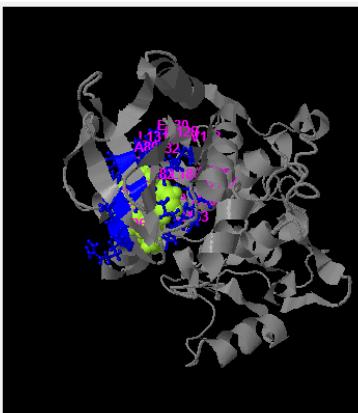
**Fig9. Result for proteins that are structurally close to target**

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

## Predicted function using COFACTOR and COACH

(This section reports biological annotations of the target protein by COFACTOR and COACH based on the I-TASSER structure prediction. While COFACTOR deduces protein functions (ligand-binding sites, EC and GO) using structure comparison and protein-protein networks, COACH is a meta-server approach that combines multiple function annotation results (on ligand-binding sites) from the COFACTOR, TM-SITE and S-SITE programs.)

### Ligand binding sites



Spin On/Off

Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Ligand Binding Site Residues
<input checked="" type="radio"/>	1	0.93	3000	5e8yA STU	<a href="#">Rep. Mult</a>	59,60,61,67,80,82,113,129,130,131,132,133,134,135,183,184,186,196,197	
<input type="radio"/>	2	0.04	72	4mneA 573	<a href="#">Rep. Mult</a>	63,64,65,82,84,100,113,127,129,196,197,198,199,200,201,204,205	
<input type="radio"/>	3	0.04	107	4wb7B PEPTIDE	<a href="#">Rep. Mult</a>	61,62,63,64,135,137,141,149,181,182,183,216,217,218,220,249,254,255,258,259,261,266,267	
<input type="radio"/>	4	0.02	58	5bmlA 4TW	<a href="#">Rep. Mult</a>	59,61,62,64,65,66,67,80,82,84,129,130,131,132,196,197	
<input type="radio"/>	5	0.01	28	3zobB 0LJ	<a href="#">Rep. Mult</a>	59,80,81,82,100,103,104,107,112,113,127,129,130,131,132,169,175,176,177,178,186,195,196,197	

[Download](#) the residue-specific ligand binding probability, which is estimated by SVM.

[Download](#) the all possible binding ligands and detailed prediction summary.

[Download](#) the templates clustering results.

(a) C-score is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates 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.

## Fig10. Result for predicted functions

- C-score is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.
- Cluster size is the total number of templates in a cluster.
- Lig Name is name of possible binding ligand. Click the name to view its information in the BioLiP database.
- 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.

Spin On/Off

### Enzyme Commission (EC) numbers and active sites



Spin On/Off

Click to view	Rank	Cscore <sup>EC</sup>	PDB Hit	TM-score	RMSD <sup>a</sup>	IDEN <sup>a</sup>	Cov	EC Number	Active Site Residues
<input type="radio"/>	1	0.334	3eqdA	0.885	0.99	0.399	0.901	2.7,12.2	NA
<input type="radio"/>	2	0.328	1s9IA	0.806	1.43	0.418	0.838	2.7,12.2,2.7,1.37	NA
<input type="radio"/>	3	0.318	3dcA	0.653	1.91	0.262	0.698	2.7,11.25	NA
<input type="radio"/>	4	0.311	1sm2A	0.649	2.54	0.219	0.725	2.7,10.2	NA
<input type="radio"/>	5	0.307	2ofvA	0.626	2.83	0.249	0.710	2.7,10.2	NA

Click on the radio buttons to visualize predicted active site residues.

(a) Cscore<sup>EC</sup> is the confidence score for the EC number prediction. Cscore<sup>EC</sup> 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) RMSD<sup>a</sup> is the RMSD between residues that are structurally aligned by TM-align.

(d) IDEN<sup>a</sup> 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

## Fig11. Enzyme commission (EC) numbers and active sites

- Cscore<sup>EC</sup> is the confidence score for the EC number prediction. Cscore<sup>EC</sup> values range in between [0-1]; where a higher score indicates a more reliable EC number prediction.
- TM-score is a measure of global structural similarity between query and template protein.
- RMSD<sup>a</sup> is the RMSD between residues that are structurally aligned by TM-align.
- IDEN<sup>a</sup> is the percentage sequence identity in the structurally aligned region.
- 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 <sup>GO</sup>	TM-score	RMSD <sup>a</sup>	IDEN <sup>a</sup>	Cov	PDB Hit	Associated GO Terms															
1	0.57	0.8850	0.99	0.40	0.90	<a href="#">3eqdA</a>	GO:0034134	GO:0006979	GO:0019901	GO:0031435	GO:0004674	GO:0002756	GO:0045087	GO:0051403	GO:0005515	GO:000165	GO:0045597					
2	0.56	0.7051	2.25	0.50	0.76	<a href="#">2qdmA</a>	GO:0004672	GO:0004674	GO:0005524	GO:0006468	GO:0016772	GO:0005625	GO:0030182	GO:002320	GO:0004728	GO:0048313	GO:0032402	GO:0008063	GO:0005794	GO:0047496	GO:0005737	GO:0090398
3	0.56	0.8064	1.43	0.42	0.84	<a href="#">1sq1A</a>	GO:0004672	GO:0004674	GO:0005524	GO:0006468	GO:0016772	GO:0003056	GO:0007173	GO:0004672	GO:000187	GO:0005886	GO:0006468	GO:0048870	GO:0006928	GO:0034111	GO:000166	GO:0004713
4	0.53	0.7251	2.74	0.55	0.80	<a href="#">3ak0A</a>	GO:0004672	GO:0004674	GO:0005524	GO:0006468	GO:0016772	GO:00051291	GO:0016749	GO:0017016	GO:0043204	GO:0030216	GO:0005874	GO:0032839	GO:0007165	GO:0002224	GO:0034142	GO:0007050
5	0.44	0.6624	3.02	0.24	0.76	<a href="#">2vraA</a>	GO:0004672	GO:0004713	GO:0005524	GO:0006468	GO:0016772	GO:0008285	GO:0048471	GO:0006711	GO:0060674	GO:0002755	GO:0032968	GO:0005524	GO:0000186	GO:0004708	GO:0005938	GO:0008283
6	0.44	0.6916	2.96	0.42	0.79	<a href="#">2dy1A</a>	GO:0004672	GO:0004674	GO:0005524	GO:0006468	GO:0016772	GO:00030335	GO:0030425	GO:0048812	GO:0051384	GO:0007067	GO:016310	GO:0007264	GO:0016772			
7	0.44	0.6511	3.51	0.25	0.78	<a href="#">3pxA</a>	GO:0004672	GO:0004713	GO:0005524	GO:0006468	GO:0016772											
8	0.43	0.6748	2.70	0.21	0.76	<a href="#">3mj2A</a>	GO:0004672	GO:0004713	GO:0005524	GO:0006468	GO:0016772											
9	0.42	0.6531	1.91	0.26	0.70	<a href="#">3dtcA</a>	GO:0004672	GO:0004674	GO:0004709	GO:0005524	GO:0006468	GO:0016772										
10	0.41	0.6165	2.95	0.25	0.71	<a href="#">3pj1A</a>	GO:0004672	GO:0004713	GO:0005524	GO:0006468	GO:0016772											

Consensus prediction of GO terms

Molecular Function	<a href="#">GO:0005524</a>	<a href="#">GO:0004674</a>	<a href="#">GO:0004713</a>	<a href="#">GO:0004728</a>	<a href="#">GO:0031435</a>	<a href="#">GO:0004708</a>	<a href="#">GO:0017016</a>			
GO-Score	0.98	0.96	0.76	0.57	0.57	0.57	0.57			
Biological Process	<a href="#">GO:0030216</a>	<a href="#">GO:0006979</a>	<a href="#">GO:0007411</a>	<a href="#">GO:0048011</a>	<a href="#">GO:0008285</a>	<a href="#">GO:0032402</a>	<a href="#">GO:0051384</a>	<a href="#">GO:0090398</a>	<a href="#">GO:0048678</a>	<a href="#">GO:0051291</a>
GO-Score	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Cellular Component	<a href="#">GO:0033267</a>	<a href="#">GO:0005874</a>	<a href="#">GO:0032839</a>	<a href="#">GO:0048471</a>	<a href="#">GO:005938</a>	<a href="#">GO:0005794</a>	<a href="#">GO:0043204</a>	<a href="#">GO:0005886</a>	<a href="#">GO:0005625</a>	<a href="#">GO:0005829</a>
GO-Score	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57

**Fig12. Gene Ontology (GO) terms**

- 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.
- TM-score is a measure of global structural similarity between query and template protein.
- RMSD<sup>a</sup> is the RMSD between residues that are structurally aligned by TM-align.
- IDEN<sup>a</sup> is the percentage sequence identity in the structurally aligned region.
- 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.
- 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/>

## WEBLEM 3c

### Robetta

(URL: <https://robbetta.bakerlab.org/>)

#### **AIM:**

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

#### **INTRODUCTION:**

Kinase, an enzyme that adds phosphate groups (PO<sub>4</sub><sup>3-</sup>) to other molecules. For protein targets, kinases can phosphorylate the amino acids serine, threonine, and tyrosine. Phosphorylation of lipid molecules by kinases is important for controlling the molecular composition of membranes in cells, which helps to specify the physical and chemical properties of the different membranes. Nucleotides, the fundamental units of RNA (ribonucleic acid) and DNA (deoxyribonucleic acid), contain a phosphate molecule attached to a nucleoside, a compound made up of a ribose moiety and a purine or pyrimidine base. The tertiary structure of kinase can be predicted using 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 modeling or de 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:**

5. Open homepage for Robetta (URL: <https://robbetta.bakerlab.org/>)
6. Complete registration.
7. Submit FASTA sequence for kinase.
8. Observe and interpret results.

## OBSERVATION:

The new UniProt website is here! [Take me to UniProt BETA](#)

**UniProtKB - P52564 (MP2K6\_HUMAN)**

**Display** [Help video](#) [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

**Entry** [Protein](#) **Dual specificity mitogen-activated protein kinase kinase 6**

**Publications** [Gene](#) **MAP2K6**

**Feature viewer** [Organism](#) **Homo sapiens (Human)**

**Feature table** [Status](#) **Reviewed** - Annotation score: - Experimental evidence at protein level<sup>i</sup>

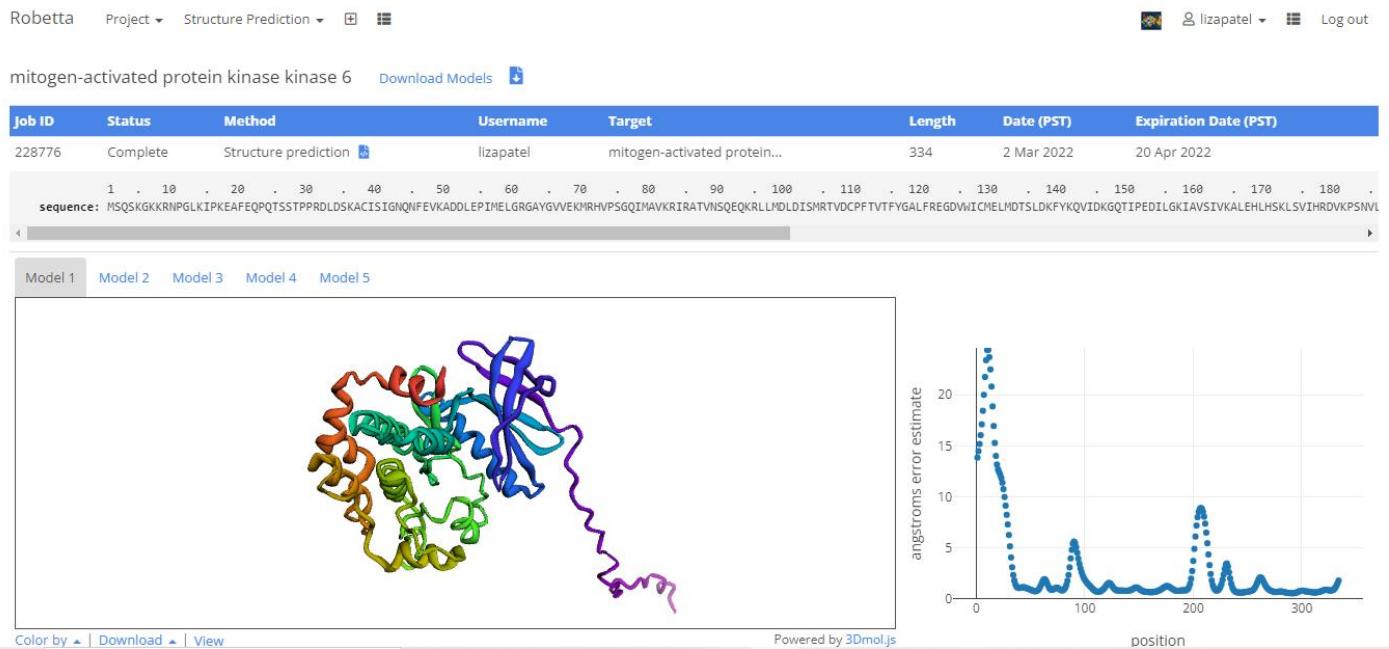
**Function**  Function

Dual specificity protein kinase which acts as an essential component of the MAP kinase signal transduction pathway. With MAP3K3/MKK3, catalyzes the concomitant phosphorylation of a threonine and a tyrosine residue in the MAP kinases p38 MAPK11, MAPK12, MAPK13 and MAPK14 and plays an important role in the regulation of cellular responses to cytokines and all kinds of stresses. Especially, MAP2K3/MKK3 and MAP2K6/MKK6 are both essential for the activation of MAPK11 and MAPK13 induced by environmental stress, whereas MAP2K6/MKK6 is the major MAPK11 activator in response to TNF. MAP2K6/MKK6 also phosphorylates and activates PAK6. The p38 MAP kinase signal transduction pathway leads to direct activation of transcription factors. Nuclear targets of p38 MAP kinase include the transcription factors ATF2 and ELK1. Within the p38 MAP kinase signal transduction pathway, MAP3K6/MKK6 mediates phosphorylation of STAT4 through MAPK14 activation, and is therefore required for STAT4 activation and STAT4-regulated gene expression in response to IL-12 stimulation. The pathway is also crucial for IL-6-induced SOCS3 expression and down-regulation of IL-6-mediated gene induction; and for IFNG-regulated gene transcription. Has a role in osteoclast differentiation through NF-kappa-B

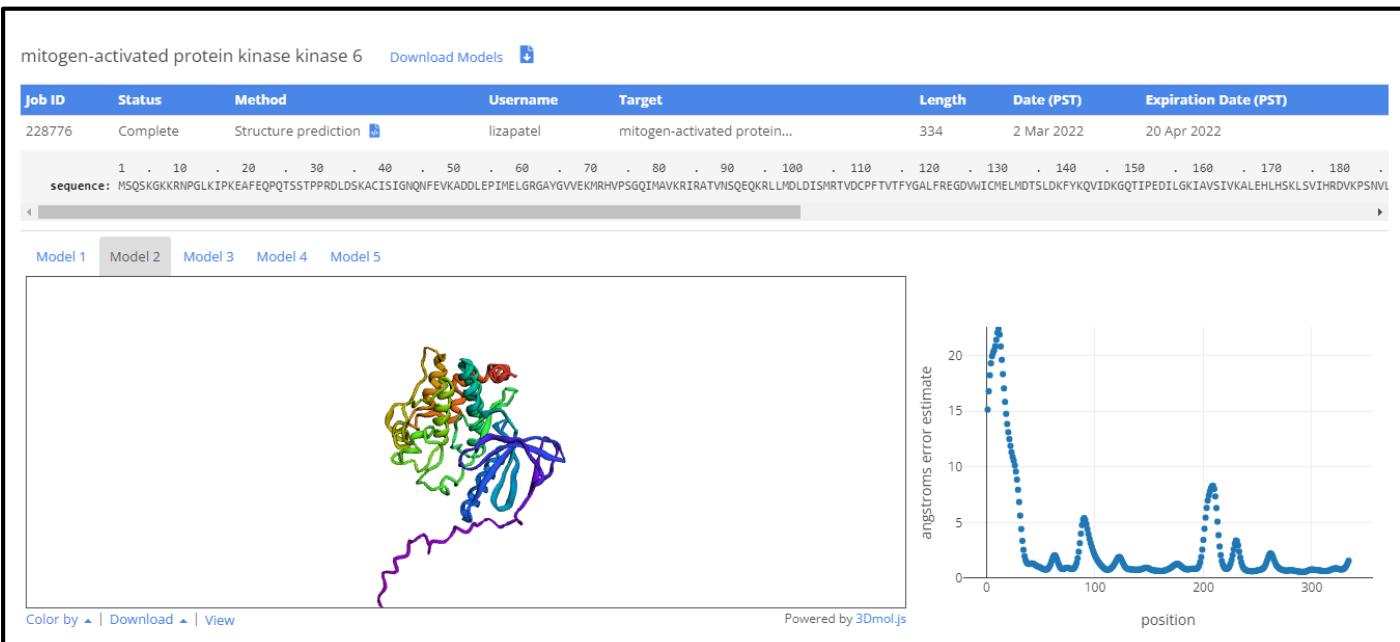
Fig1. Result page for kinase in UniProt database

```
>sp|P52564|MP2K6_HUMAN Dual specificity mitogen-activated protein kinase kinase 6 OS=Homo sapiens OX=9606 GN=MAP2K6 PE=1 SV=1
MSQSGKKRNPGLKIPKEAEPQQTSSTPRDLDSKACISIGNQNEVKADDLEPIMELG
RGAYGVVEKVRHVPVSGQTMAVKRIRATVNSQEQKRLLMDLISMRVTPCPFTVTFYALF
REGDWHICHELMIDTSLDKFYKQVQIDKGQTIPEDIIGKIAVSIVKALEHLHSKLSVIRHDV
KPSNVLINALQGVKICDFGIGSGYLVDSVAKTIDAGCKPVVAPERINPELNQKGYSVKSIDI
ISLGITMIEELAILRFFPYDSWGTTPQQLKQVVEEPSPQLPADKFSAEFVDFTSQCLKKNSK
ERPTYPELNLQHPFFTLHESKGTDVASFVKLILGD
```

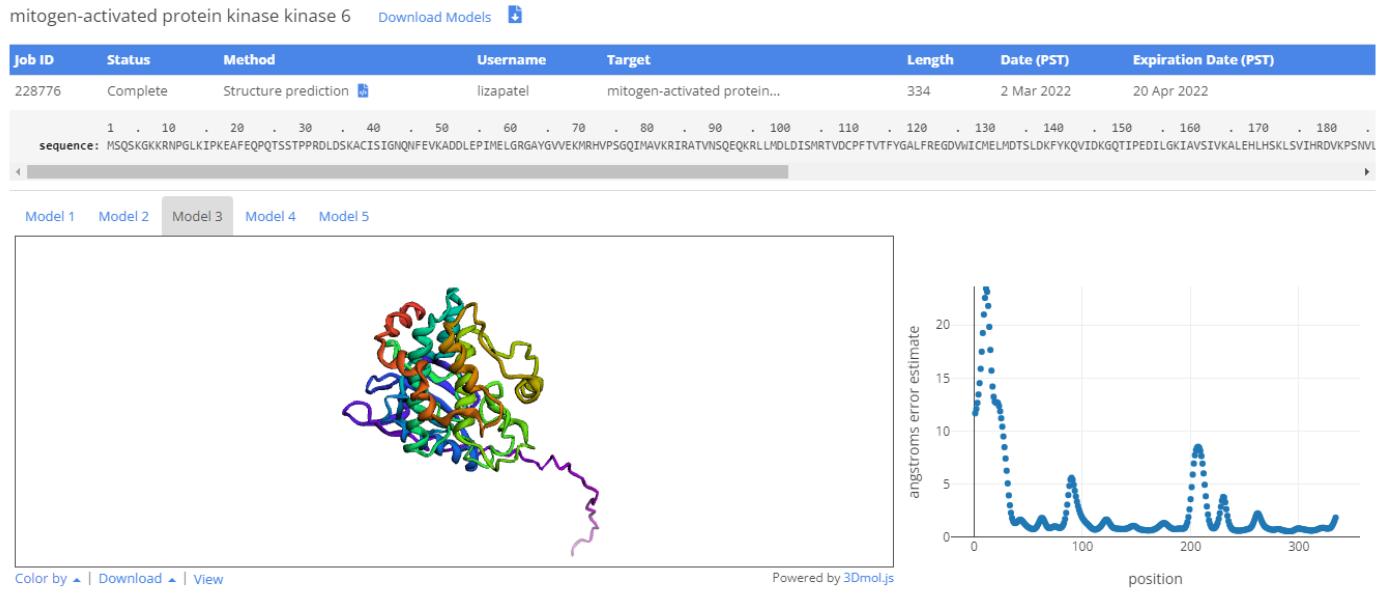
Fig2. FASTA sequence for kinase



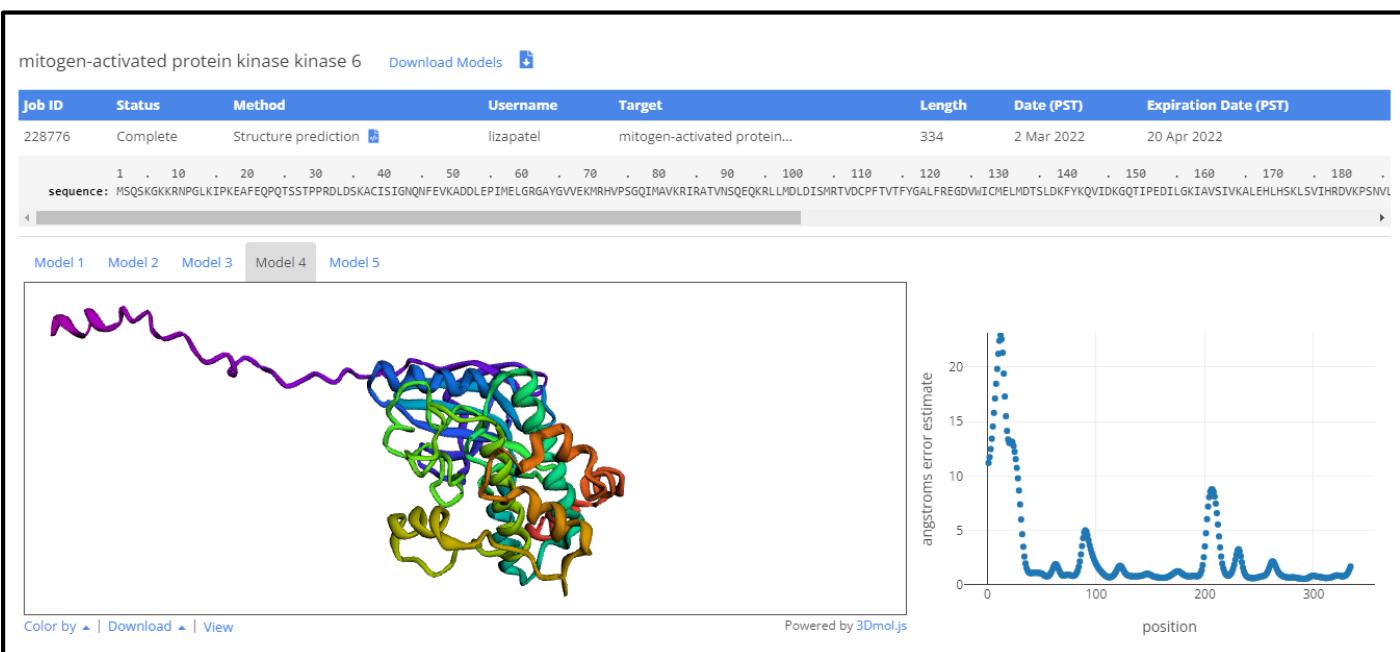
**Fig3. Model 1 with atom co-ordinates**



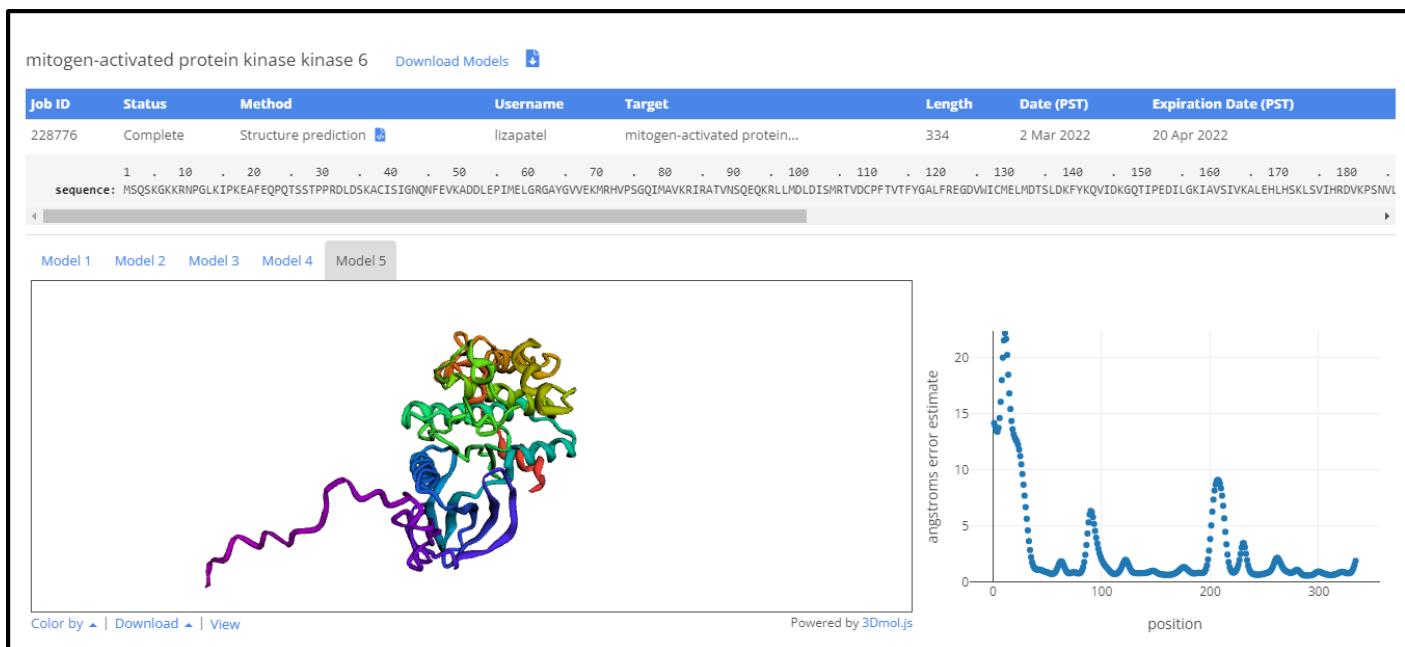
**Fig4. Model 2 with atom co-ordinates**



**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 Kinase 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. *kinase / Definition, Biology, & Function*. (n.d.). Encyclopedia Britannica. Retrieved March 8, 2022, from <https://www.britannica.com/science/kinase>
3. Robetta (2021b). Bakerlab.org. Retrieved March 8, 2022, from <https://robbetta.bakerlab.org/>
4. *mitogen-activated protein kinase kinase 6*. (n.d.). Robetta.bakerlab.org. Retrieved March 8, 2022, from <https://robbetta.bakerlab.org/results.php?id=228776>

## 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  $\varphi$ - $\psi$  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 unfavorable 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 stereochemical 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  $\varphi-\psi$  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.doe-mbi.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

## WEBLEM 4a

### SAVES server (URL: <https://saves.mbi.ucla.edu/>)

#### AIM:

To validate structure kseq.B99990003 generated from Modeller.

#### INTRODUCTION:

Kseq.B99990003 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  $\varphi$ - $\psi$  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:

**UCLA-DOE LAB — SAVES v6.0**

**UCLA**

To run any or all programs:  
upload your structure, in PDB format only

No file chosen

**References**

ERRAT

- Reference: Verification of protein structures: patterns of nonbonded atomic interactions, Colovos C and Yeates TO, 1993.
- C++ software

VERIFY 3D

- Profile Search Software [Bowie et al., 1991, Luethy et al., 1992].
- DSSP original and Wikipedia

**Fig1. Homepage for SAVES server**

## UCLA-DOE LAB — SAVES v6.0

UCLA

To run any or all programs:  
upload your structure, in PDB format only

Choose file kseq.B99990003.pdb

Customize job name:

kseq.B99990003.pdb

Run programs

### References

#### ERRAT

- Reference: Verification of protein structures: patterns of nonbonded atomic interactions, Colovos C and Yeates TO, 1993.
- C++ software

Fig2. Structure from Modeller for validation

## UCLA-DOE LAB — SAVES v6.0

UCLA

Job 924086 has been created

New Job

job #924086: kseq.B99990003.pdb [job link] [3D Viewer]

ERRAT Complete

Overall Quality Factor

**57.362**

Results

VERIFY Complete

67.37% of the residues have averaged 3D-1D score  $\geq 0.2$

**Fail**

Fewer than 80% of the amino acids have scored  $\geq 0.2$  in the 3D/1D profile.

Results

PROVE Complete

Buried outlier protein atoms total from 1 Model: 7.3%

**fail**

Results

WHATCHECK Complete

1	2	3	4	5	6	7	8	9	10	11	12	13
14	15	16	17	18	19	20	21	22	23			
24	25	26	27	28	29	30	31	32	33			
34	35	36	37	38	39	40	41	42	43			
44	45	46	47									

Results

PROCHECK Complete

Out of 9 evaluations  
Errors: 4  
Warning: 2  
Pass: 3

Results

Almost ready, check back soon

Fig3. Result page for structure validation for various servers

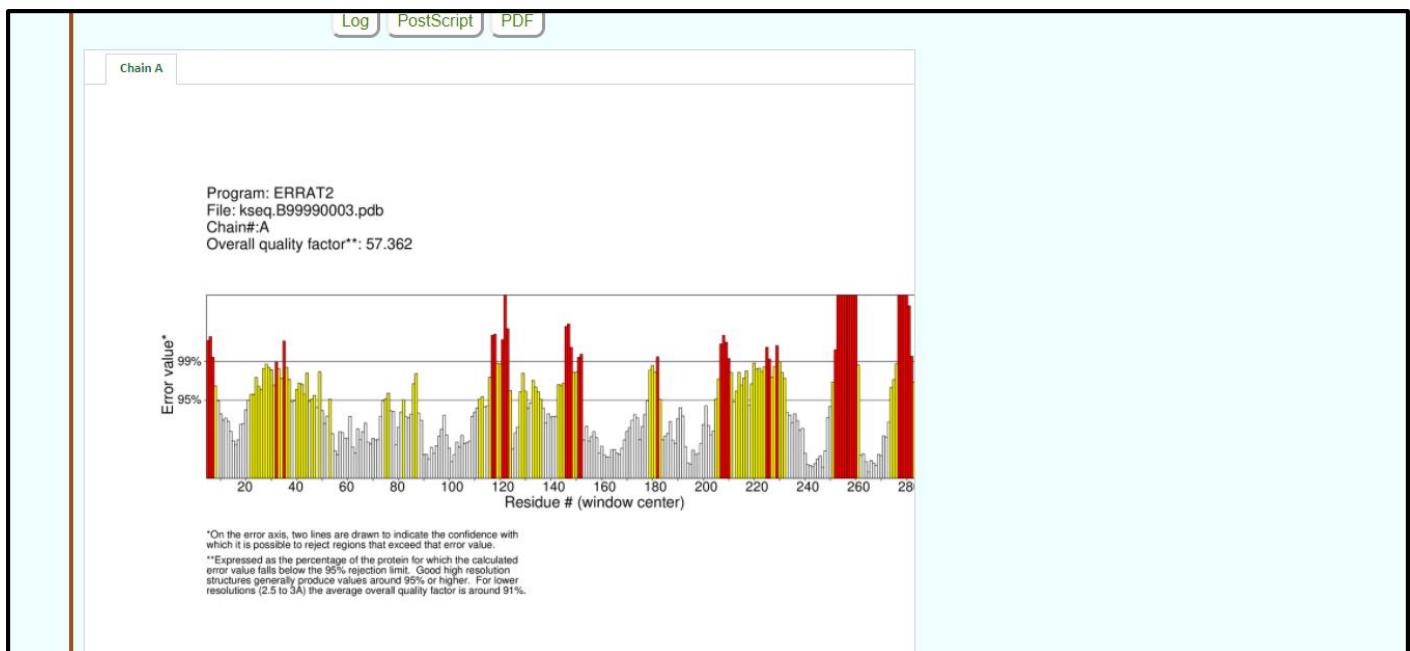


Fig4. Result page for Errat

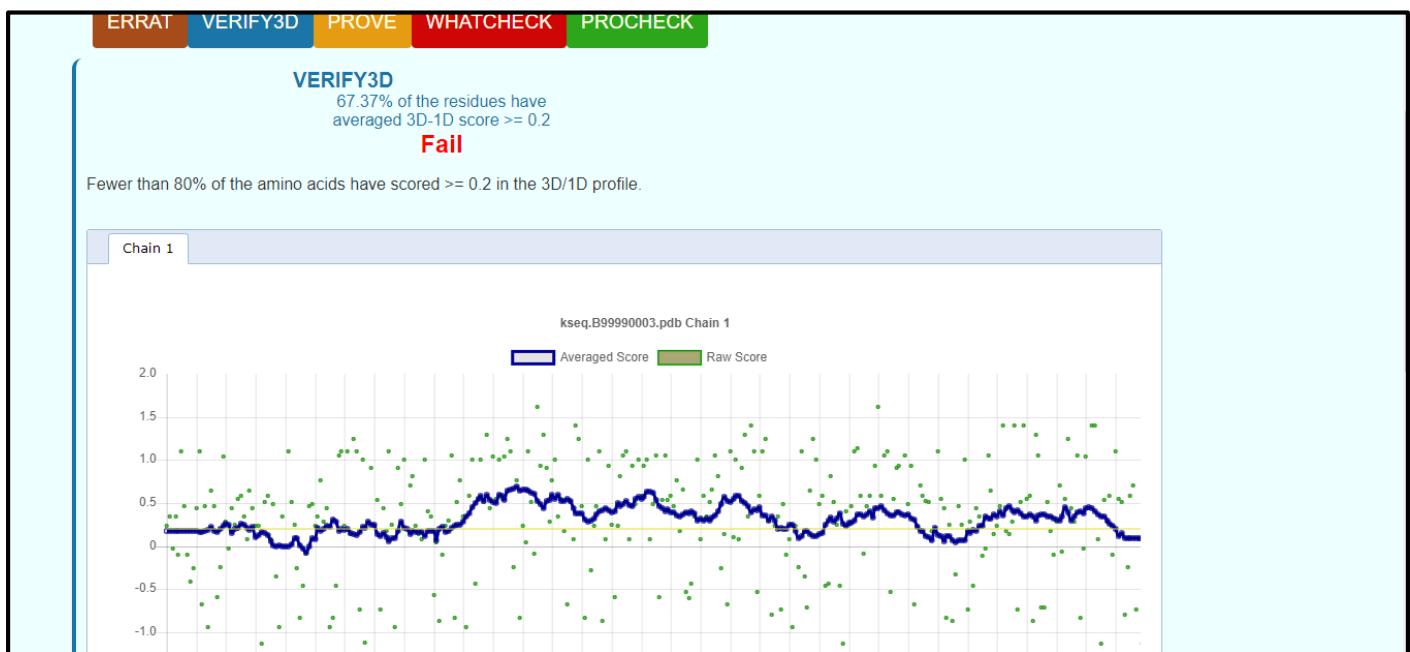


Fig5. Result page for Verify3D

# PROVE

## saves.pdb

- Analysis of entire structure

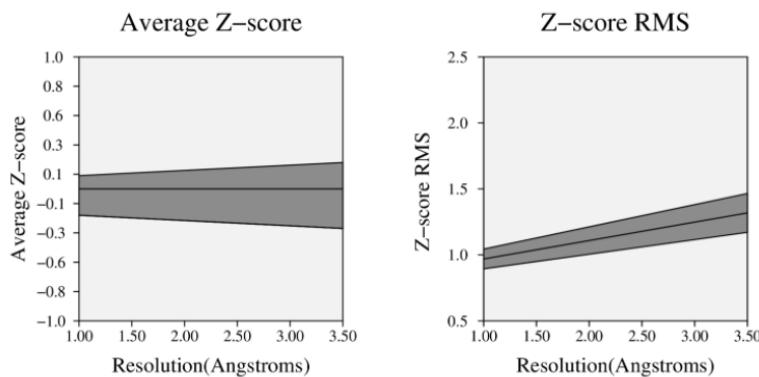


Fig6. Result page for Prove

ERRAT VERIFY3D PROVE WHATCHECK PROCHECK

### WHATCHECK

Headings

1 2 3 4 5  
6 7 8 9 10  
11 12 13 14  
15 16 17 18  
19 20 21 22  
23 24 25 26  
27 28 29 30  
31 32 33 34  
35 36 37 38  
39 40 41 42  
43 44 45 46  
47

#### #33. Error: Abnormally short interatomic distances

Error: Abnormally short interatomic distances  
The pairs of atoms listed in the table below have an unusually short distance.

The contact distances of all atom pairs have been checked. Two atoms are said to 'bump' if they are closer than the sum of their Van der Waals radii minus 0.40 Angstrom. For hydrogen bonded pairs a tolerance of 0.55 Angstrom is used. The first number in the table tells you how much shorter that specific contact is than the acceptable limit. The second distance is the distance between the centers of the two atoms.

The last text-item on each line represents the status of the atom pair. The text 'INTRA' means that the bump is between atoms that are explicitly listed in the PDB file. 'INTER' means it is an inter-symmetry bump. If a line contains the text 'HB', the bump criterium was relaxed because there could be a hydrogen bond. If the text 'BF' is present, the sum of the B-factors of the atoms is higher than 80, which makes the appearance of the bump somewhat less severe because the atoms probably aren't there anyway.

Bumps between atoms for which the sum of their occupancies is lower than one are not reported. In any case, each bump is listed in only one direction.

109 CYS (109 ) A SG	--	112 THR (112 ) A CG2	0.340	3.060	INTRA	BF
132 MET (132 ) A SD	--	194 LYS (194 ) A CD	0.339	3.061	INTRA	BF
70 MET (70 ) A SD	--	81 VAL (81 ) A CG2	0.332	3.068	TNTBA	BF

Fig7. Result page for Whatcheck

## PROCHECK

Out of 9 evaluations

- Errors: 4
- Warning: 2
- Pass: 3

The evaluations are the '+' (Warning) and '\*' (Error) in the summary. The categories on the left do not always correspond in number due to PROCHECK output documents.

Summary
Ramachandran plot <b>Error</b>
All Ramachandrans <b>Warning</b>
Chi1-chi2 plots <b>Pass</b>
Main-chain params
Side-chain params <b>Error</b>
Residue properties <b>Pass</b>
Bond len/angle <b>Pass</b>

```
+-----<< P R O C H E C K S U M M A R Y >>-----+
| /var/www/SAVES/Jobs/924086/saves.pdb 1.5          334 residues |
* Ramachandran plot: 84.9% core 13.7% allow 1.0% gener 0.3% disall
* All Ramachandrans: 13 labelled residues (out of 332)
+ Chi1-chi2 plots: 6 labelled residues (out of 205)
Side-chain params: 5 better 0 inside 0 worse
* Residue properties: Max.deviation: 12.3      Bad contacts: 13
* Bond len/angle: 6.2      Morris et al. class: 1 1 2
+ 1 cis-peptides
G-factors      Dihedrals: -0.12  Covalent: -0.36  Overall: -0.20
Planar groups: 100.0% within limits 0.0% highlighted
```

Fig8. Result page for Procheck

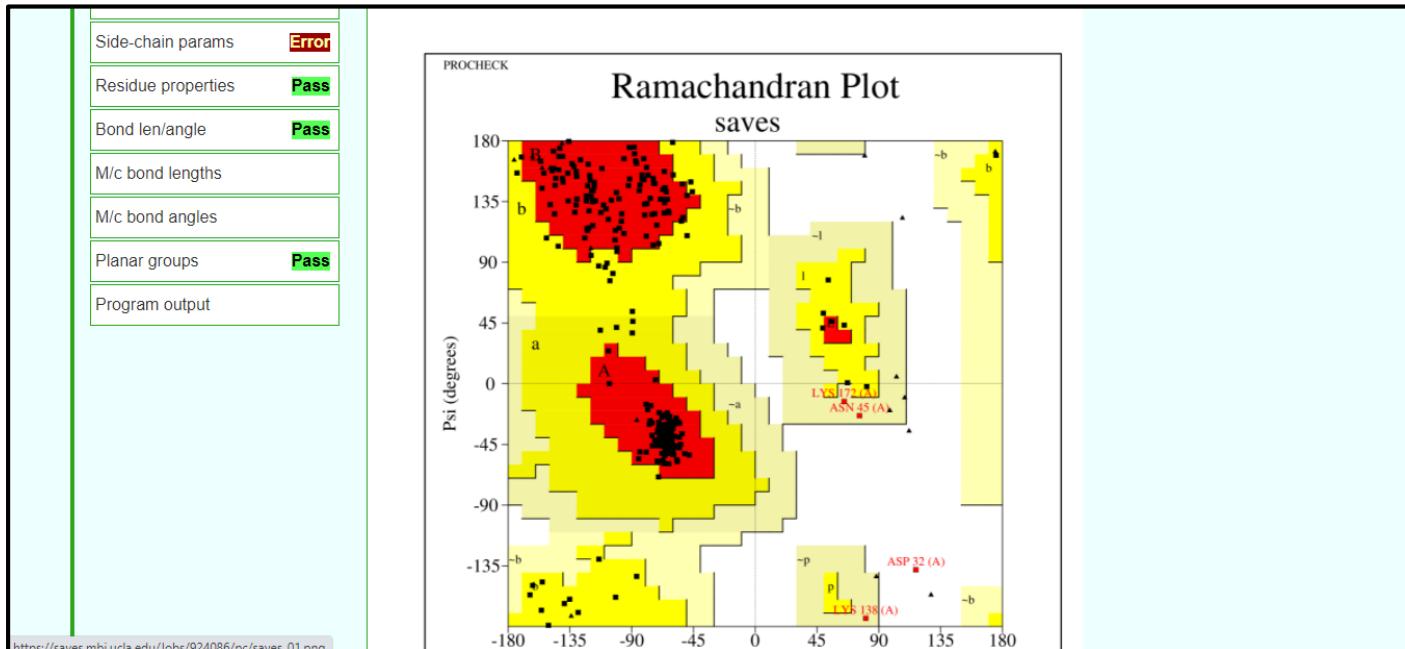
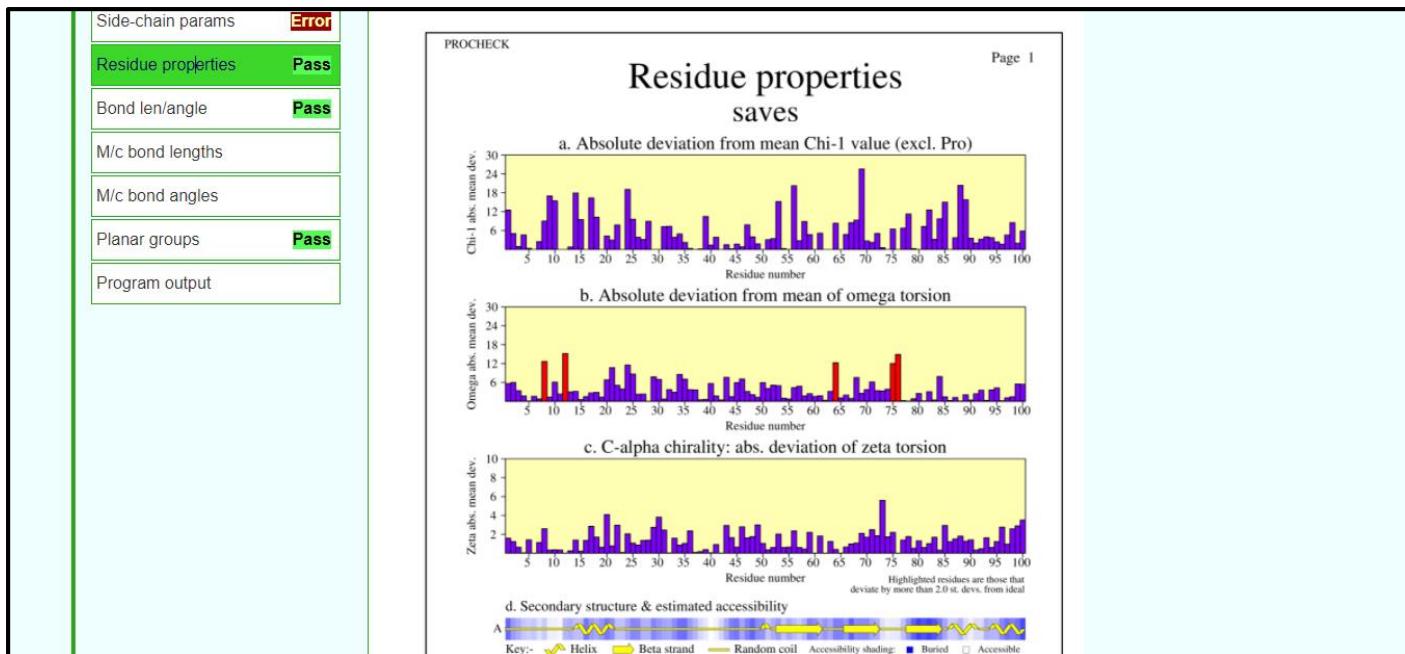


Fig8.1. Result page for Ramachandran Plot



**Fig8.2. Result page for residue properties**

## RESULT:

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

## CONCLUSION:

SAVES is an integrated server containing various tools on a single platform that can be used for tertiary structure validation. The predicted structure for kinase 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>

## WEBLEM 4b

### SAVES server (URL: <https://saves.mbi.ucla.edu/>)

#### AIM:

To validate structure model1 generated from I-TASSAR 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  $\varphi$ - $\psi$  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:

5. Open homepage for SAVES server. (URL: <https://saves.mbi.ucla.edu/>)
6. Upload structure retrieved from I-TASSER in PDB format.
7. Obtain results for Errat, Verify3D, Prove, Whatcheck and Procheck.
8. Observe and interpret the results.

#### OBSERVATION:

**UCLA-DOE LAB — SAVES v6.0**

To run any or all programs:  
upload your structure, in PDB format only



No file chosen

**References**

ERRAT

- Reference: Verification of protein structures: patterns of nonbonded atomic interactions, Colovos C and Yeates TO, 1993.
- C++ software

VERIFY 3D

- Profile Search Software [Bowie et al., 1991, Luethy et al., 1992].
- DSSP original and Wikipedia

**Fig1. Homepage for SAVES server**

## UCLA-DOE LAB — SAVES v6.0



To run any or all programs:  
upload your structure, in PDB format only

Choose file model1.pdb

Customize job name:

model1.pdb

Run programs

## References

### ERRAT

- Reference: Verification of protein structures: patterns of nonbonded atomic interactions, Colovos C and Yeates TO, 1993.
- C++ software

Fig2. Structure from Modeller for validation

## UCLA-DOE LAB — SAVES v6.0



Job 924117 has been created

New Job

### job #924117: model1.pdb [job link] [3D Viewer]

#### ERRAT Complete

##### Overall Quality Factor

**93.2515**

[Results](#)

#### VERIFY Complete

79.64% of the residues have averaged 3D-1D score  $\geq 0.2$

**Fail**

Fewer than 80% of the amino acids have scored  $\geq 0.2$  in the 3D/1D profile.

[Results](#)

#### PROVE Complete

Buried outlier protein atoms total from 1 Model: 4.9%

**warning**

[Results](#)

#### WHATCHECK Complete

1	2	3	4	5	6	7	8	9	10	11	12	13
14	15	16	17	18	19	20	21	22	23			
24	25	26	27	28	29	30	31	32	33			
34	35	36	37	38	39	40	41	42	43			
44	45	46	47	48								

[Results](#)

#### PROCHECK Complete

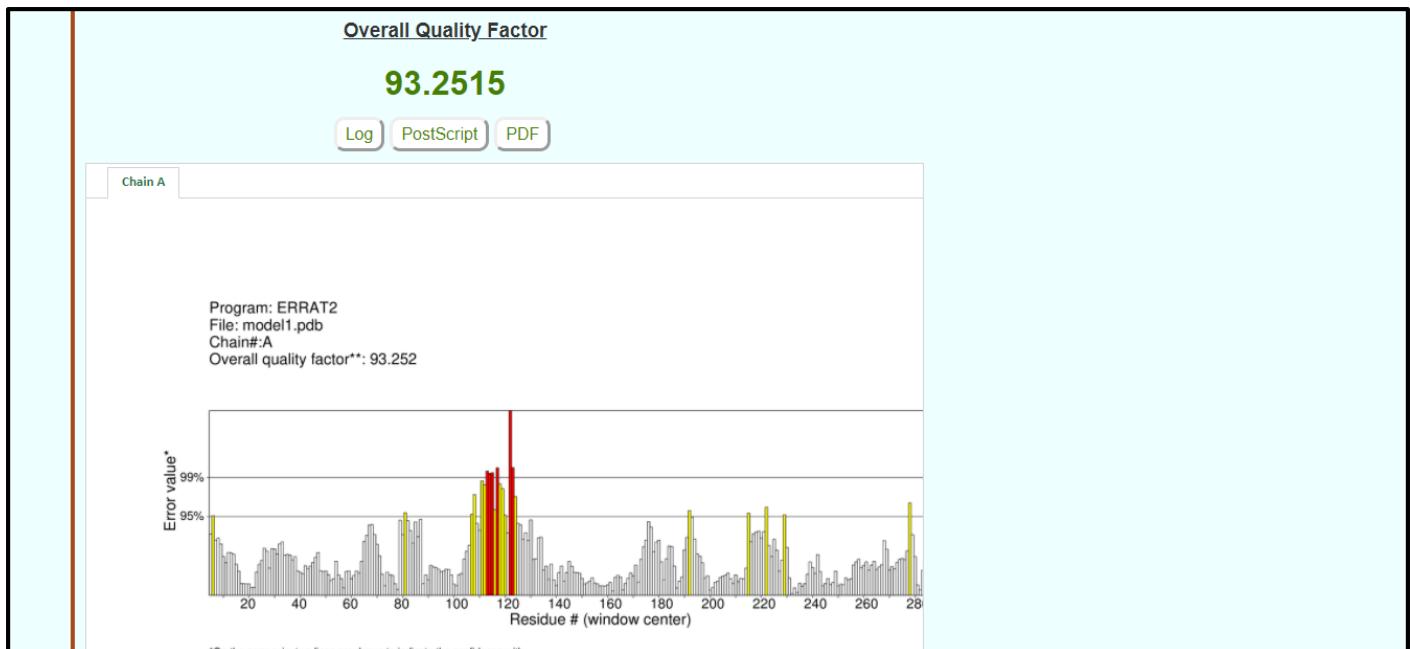
Out of 9 evaluations

Errors: 6  
Warning: 3  
Pass: 0

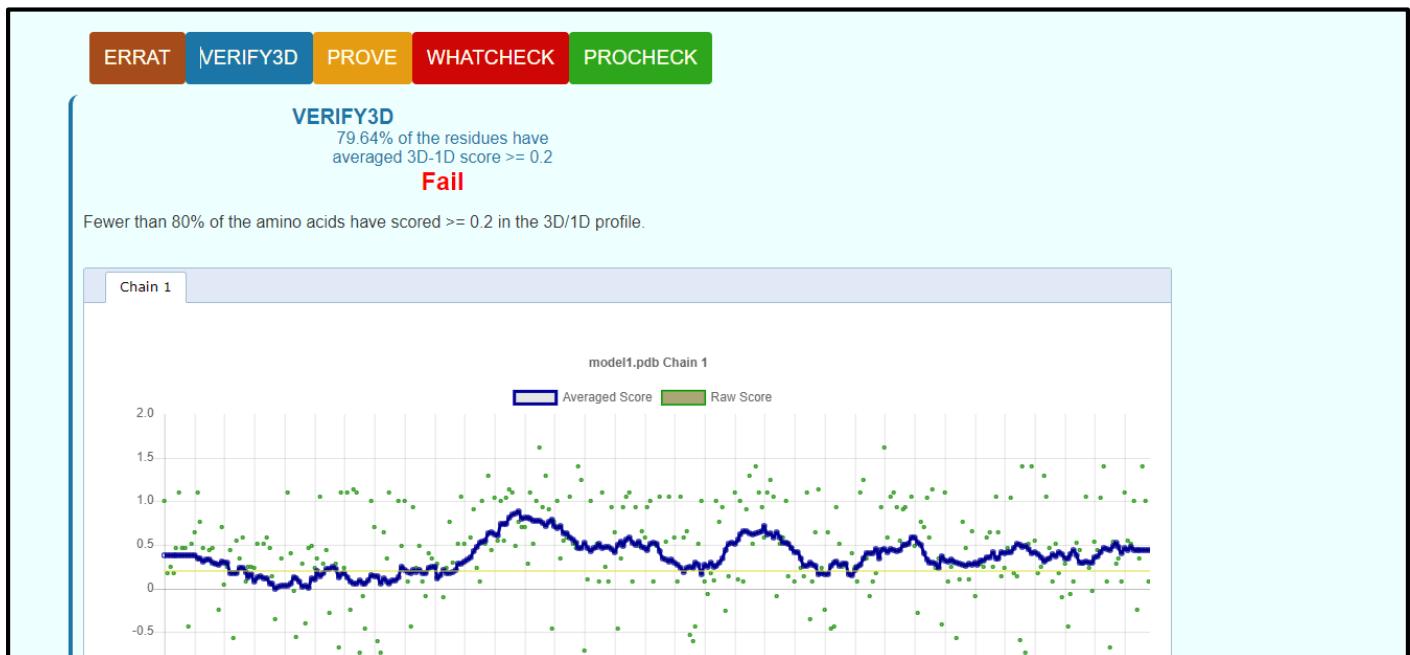
[Results](#)

Almost ready, check back soon

Fig3. Result page for structure validation for various servers



#### Fig4. Result page for Errat



## Fig5. Result page for Verify3D

# PROVE

## saves.pdb

### Analysis of entire structure

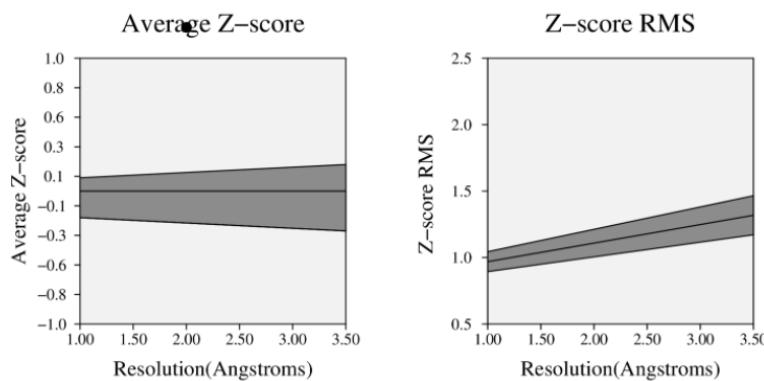


Fig6. Result page for Prove

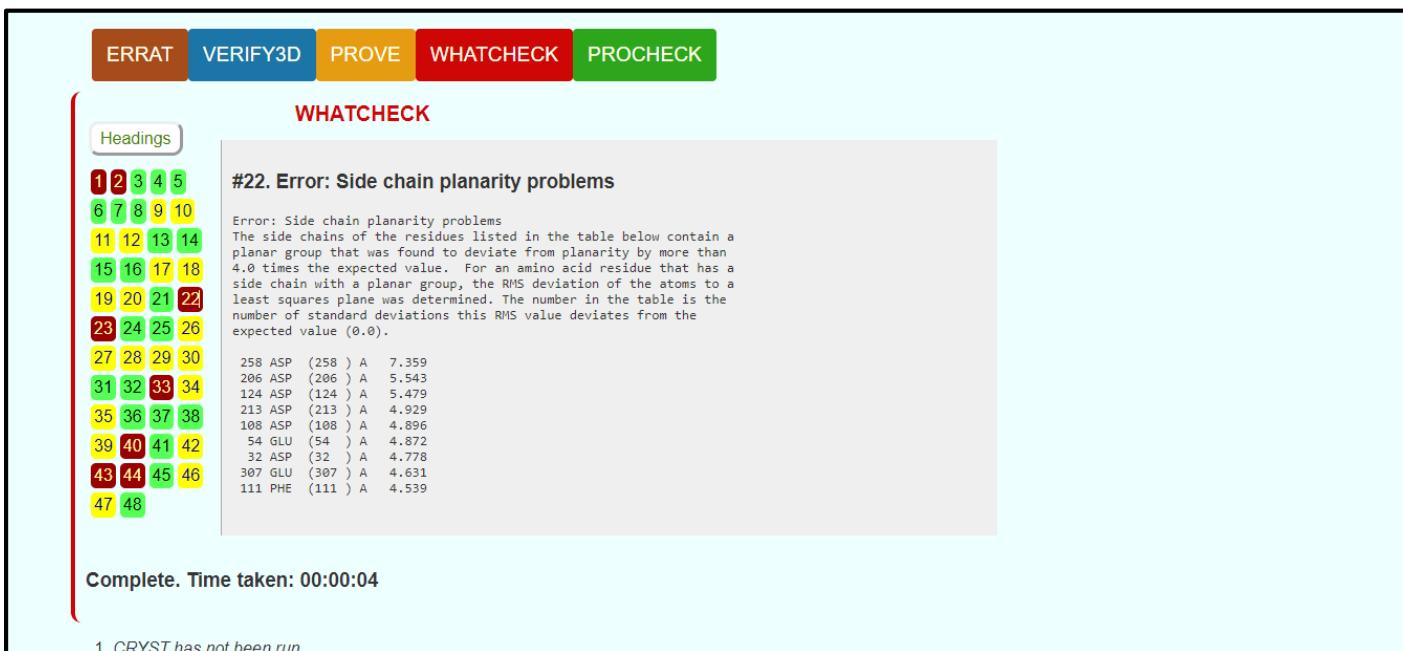


Fig7. Result page for Whatcheck

PROCHECK

Out of 9 evaluations

- Errors: 6
- Warning: 3
- Pass: 0

The evaluations are the '+' (Warning) and '\*' (Error) in the summary. The categories on the left do not always correspond in number due to PROCHECK output documents.

Summary
Ramachandran plot <span style="color: red;">Error</span>
All Ramachandrans <span style="color: red;">Error</span>
Chi1-chi2 plots <span style="color: yellow;">Warning</span>
Main-chain params
Side-chain params <span style="color: red;">Error</span>
Residue properties <span style="color: red;">Error</span>

```
-----<<< P R O C H E C K   S U M M A R Y >>>-----+
/var/www/SAVES/Jobs/924117/saves.pdb  1.5          334 residues
* Ramachandran plot:  78.0% core   18.6% allow   2.4% gener   1.0% disall
* All Ramachandrans: 33 labelled residues (out of 332)
* Chi1-chi2 plots:  11 labelled residues (out of 205)
+ Side-chain params: 3 better     0 inside     2 worse
* Residue properties: Max.deviation: 19.2          Bad contacts: 3
* Bond len/angle: 10.8          Morris et al class: 1 3 2
+ 2 cis-peptides
+ G-factors          Dihedrals: -0.71 Covalent: -0.12 Overall: -0.45
* Planar groups: 77.0% within limits 23.0% highlighted 6 off graph
```

Fig8. Result page for Procheck

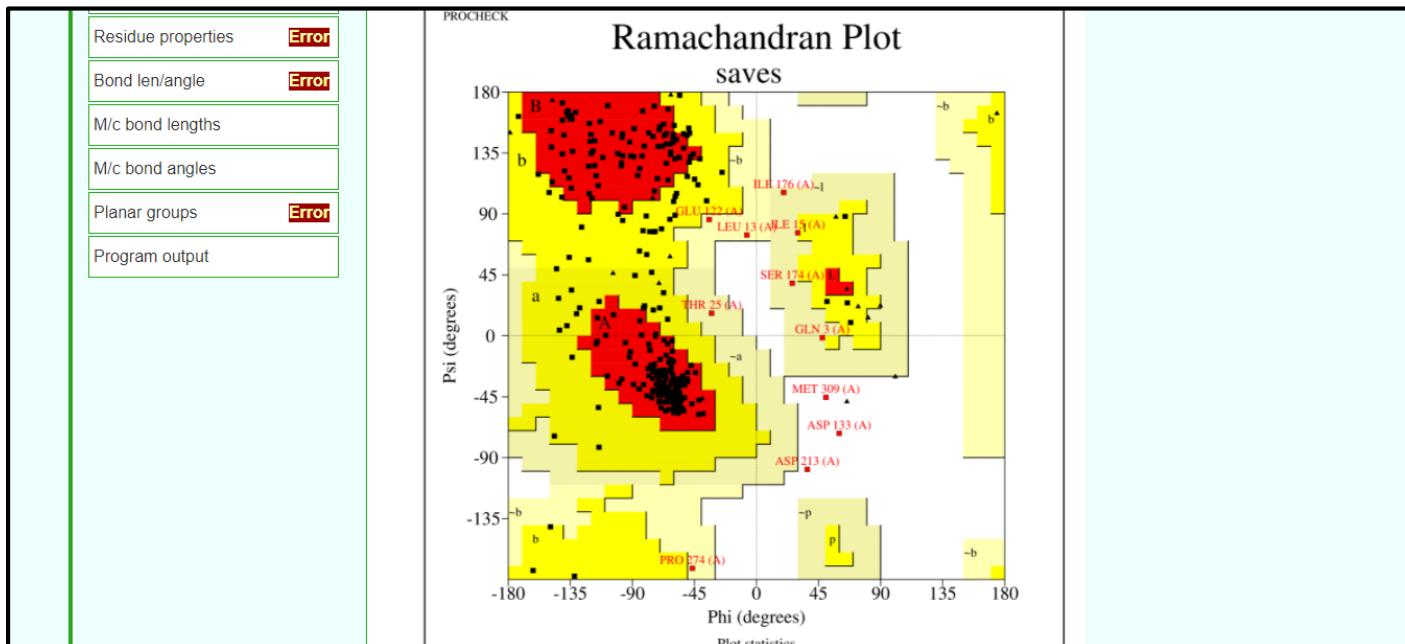
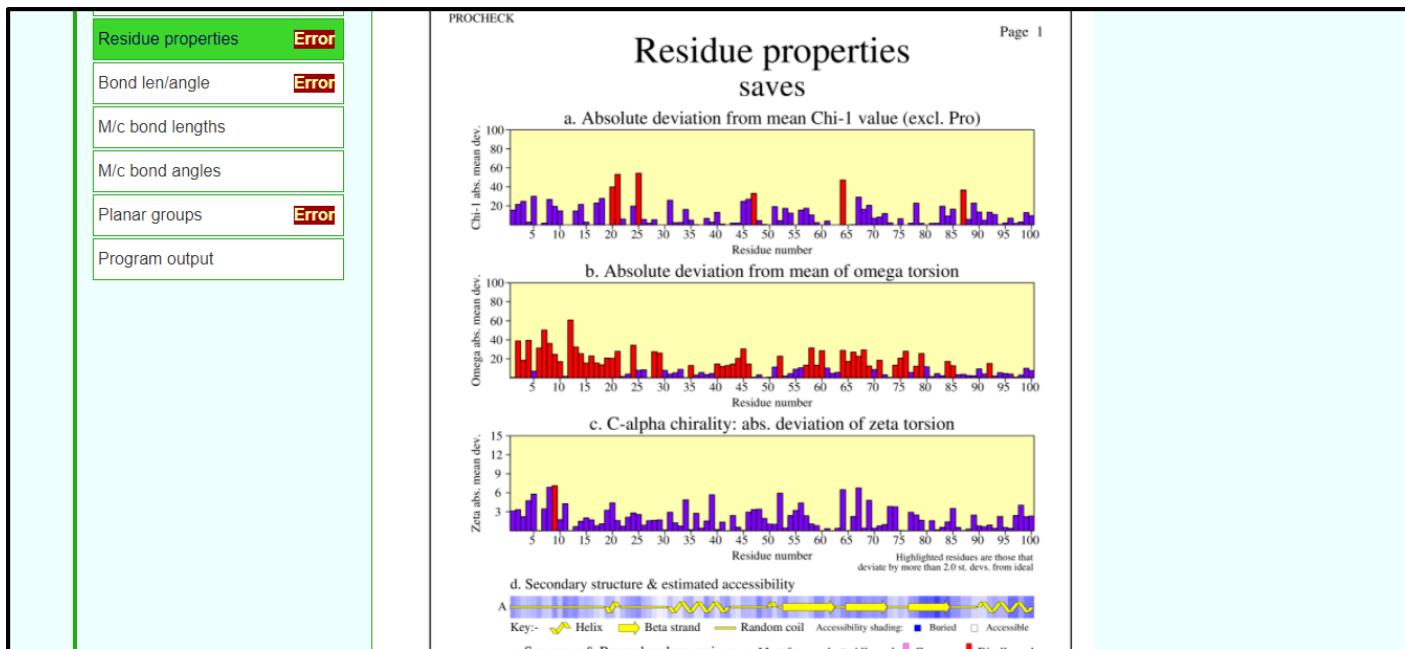


Fig8.1. Result page for Ramachandran Plot



**Fig8.2. Result page for residue properties**

## RESULT:

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

## CONCLUSION:

SAVES is 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.). <https://saves.mbi.ucla.edu/> Retrieved March 8, 2022, from <https://saves.mbi.ucla.edu/>
3. SAVESv6.0 - Structure Validation Server. (n.d.). <https://saves.mbi.ucla.edu/?job=924117>

## WEBLEM 4c

### SAVES server (URL: <https://saves.mbi.ucla.edu/>)

#### AIM:

To validate structure 228776\_1 generated from Robetta server.

#### INTRODUCTION:

228776\_1 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  $\varphi$ - $\psi$  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:

9. Open homepage for SAVES server. (URL: <https://saves.mbi.ucla.edu/>)
10. Upload structure retrieved from Robetta in PDB format.
11. Obtain results for Errat, Verify3D, Prove, Whatcheck and Procheck.
12. Observe and interpret the results.

#### OBSERVATION:

**UCLA-DOE LAB — SAVES v6.0**

To run any or all programs:  
upload your structure, in PDB format only



Choose file  No file chosen

**Run programs**

**References**

ERRAT

- Reference: [Verification of protein structures: patterns of nonbonded atomic interactions](#), Colovos C and Yeates TO, 1993.
- [C++ software](#)

VERIFY 3D

- [Profile Search Software](#) [Bowie et al., 1991, Luethy et al., 1992].
- [DSSP original and Wikipedia](#)

**Fig1. Homepage for SAVES server**

## UCLA-DOE LAB — SAVES v6.0

UCLA

To run any or all programs:  
upload your structure, in PDB format only

Choose file full\_model\_228776\_1.pdb

Customize job name:

full\_model\_228776\_1.pdb

Run programs

## References

### ERRAT

- Reference: Verification of protein structures: patterns of nonbonded atomic interactions, Colovos C and Yeates TO, 1993.
- C++ software

Fig2. Structure from Modeller for validation

## UCLA-DOE LAB — SAVES v6.0

UCLA

Job 928054 has been created

New Job

job #928054: full\_model\_228776\_1.pdb [job link] [3D Viewer]

ERRAT Complete

Overall Quality Factor

**97.7848**

Results

VERIFY Complete

84.73% of the residues have averaged 3D-1D score  $\geq 0.2$

**Pass**

At least 80% of the amino acids have scored  $\geq 0.2$  in the 3D/1D profile.

Results

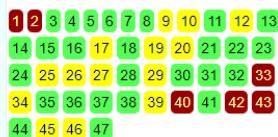
PROVE Complete

Buried outlier protein atoms total from 1 Model: 4.1%

**warning**

Results

WHATCHECK Complete



Results

PROCHECK Complete

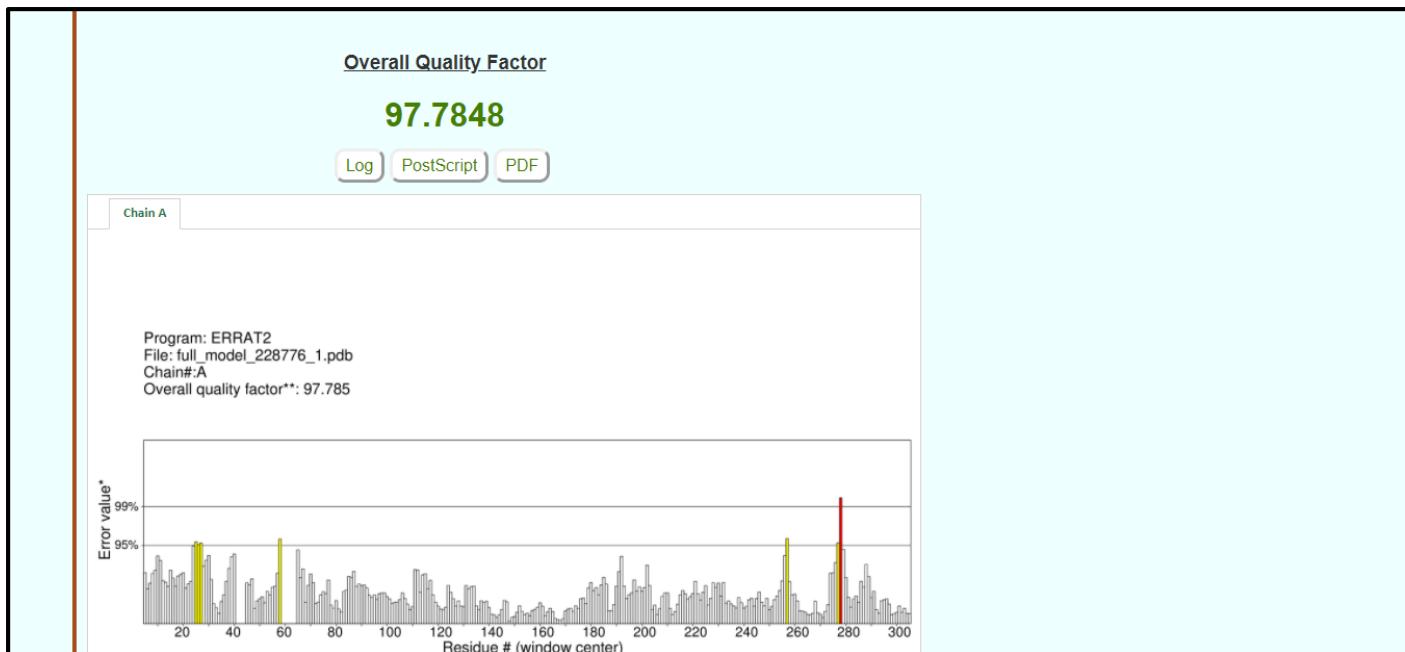
Out of 9 evaluations

Errors: 3  
Warning: 2  
Pass: 4

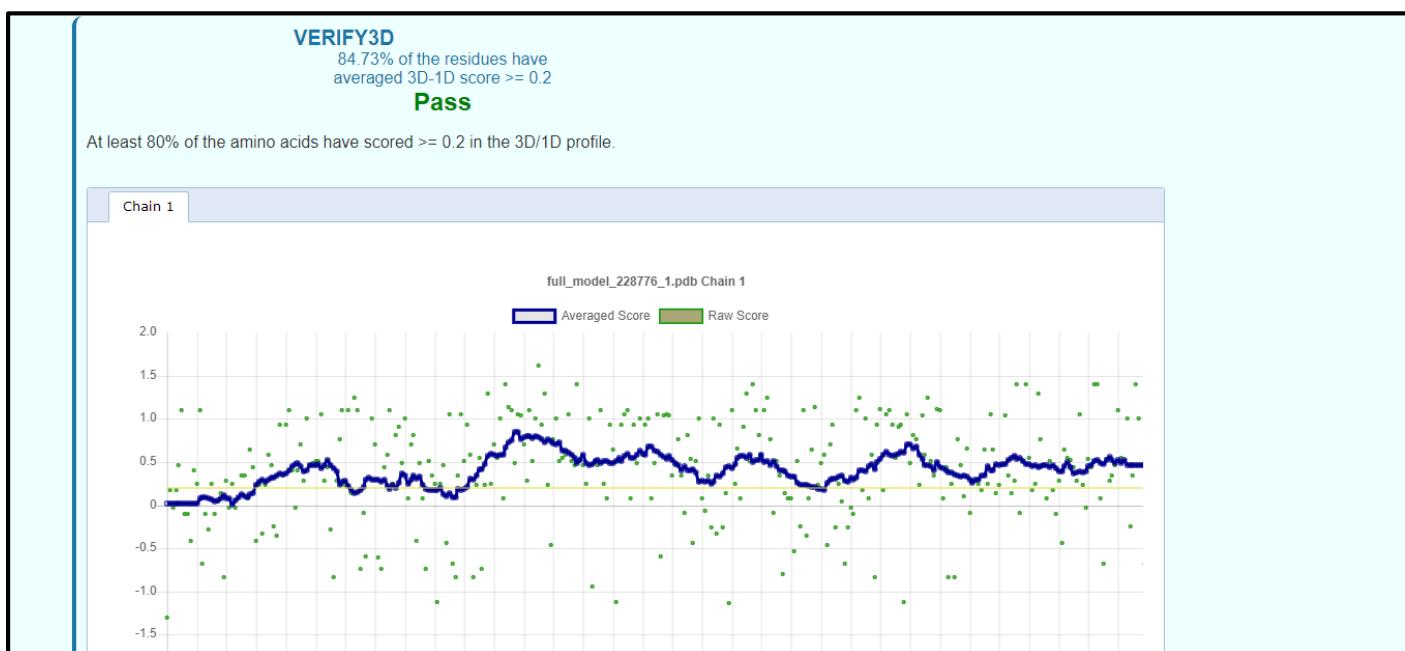
Results

Almost ready, check back soon!

Fig3. Result page for structure validation for various servers



**Fig4. Result page for Errat**



**Fig5. Result page for Verify3D**

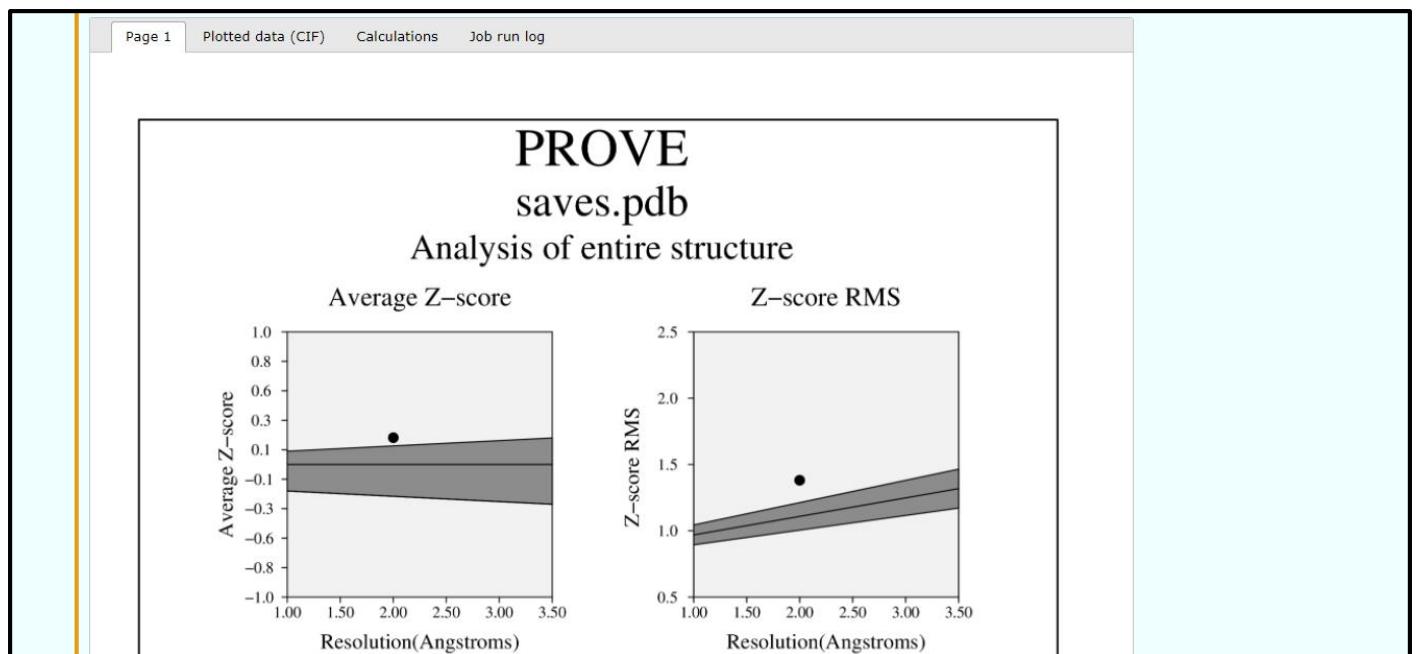


Fig6. Result page for Prove



Fig7. Result page for Whatcheck

## PROCHECK

Out of 9 evaluations

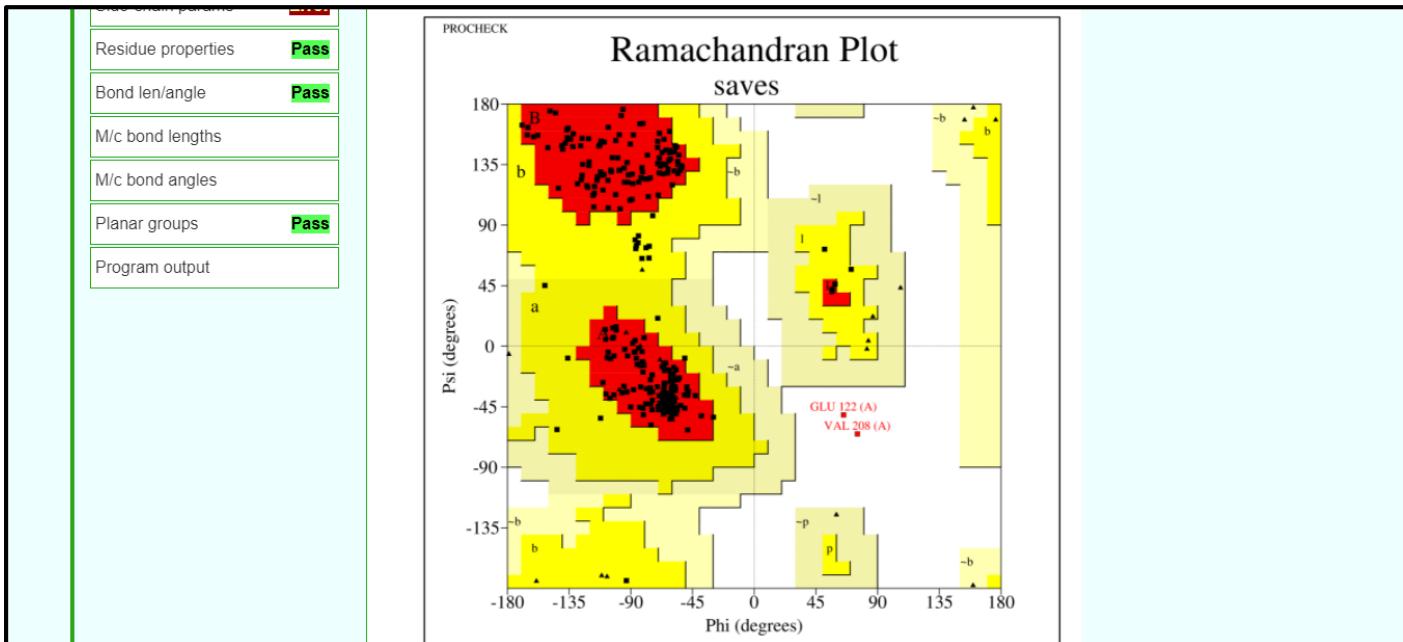
- Errors: 3
- Warning: 2
- Pass: 4

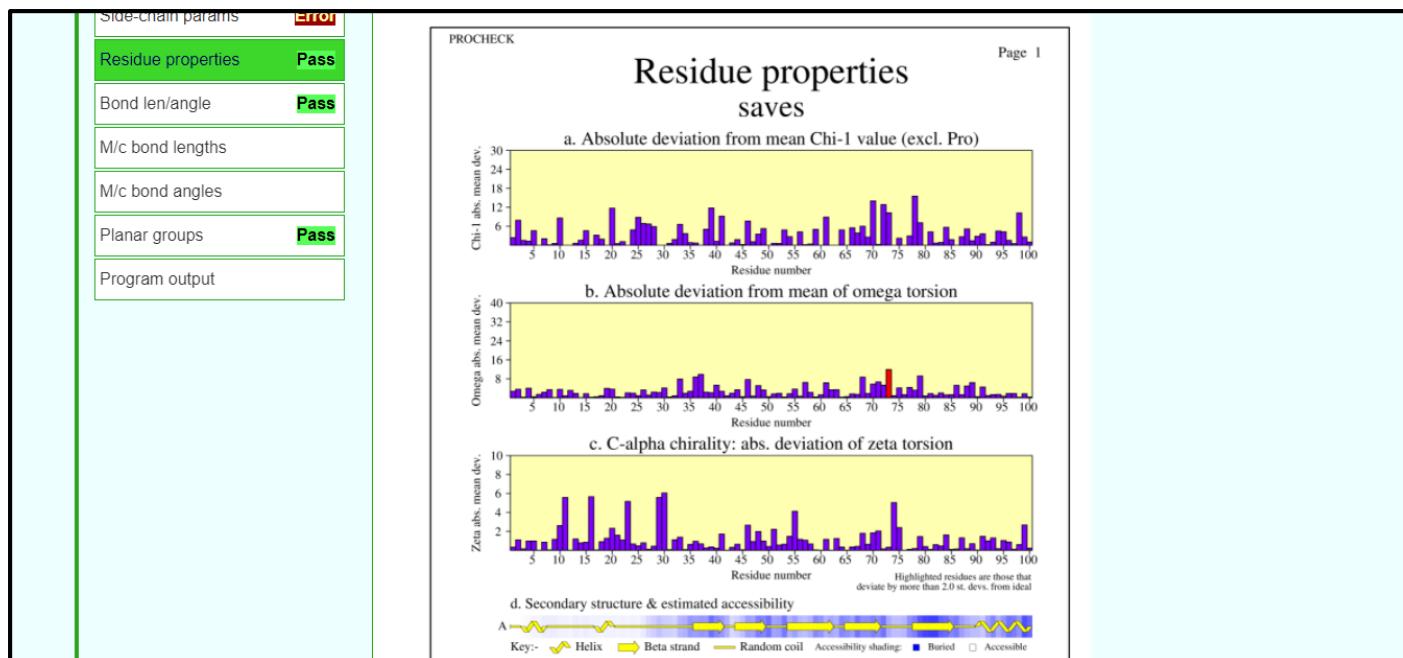
The evaluations are the '+' (Warning) and '\*' (Error) in the summary. The categories on the left do not always correspond in number due to PROCHECK output documents.

Summary
Ramachandran plot <b>Warning</b>
All Ramachandrans <b>Pass</b>
Chi1-chi2 plots <b>Pass</b>
Main-chain params
Side-chain params <b>Error</b>
Residue properties <b>Pass</b>
Bond len/angle <b>Pass</b>
M/c bond lengths

```
+-----<< P R O C H E C K S U M M A R Y >>-----+
| /var/www/SAVES/Jobs/928054/saves.pdb 1.5 334 residues |
* Ramachandran plot: 92.1% core 7.2% allow 0.0% gener 0.7% disall
+ All Ramachandrans: 6 labelled residues (out of 332)
Chi1-chi2 plots: 0 labelled residues (out of 205)
Side-chain params: 5 better 0 inside 0 worse
* Residue properties: Max.deviation: 5.4 Bad contacts: 0
* Bond len/angle: 5.8 Morris et al class: 1 1 1
+ 2 cis-peptides
G-factors Dihedrals: 0.23 Covalent: 0.33 Overall: 0.28
Planar groups: 100.0% within limits 0.0% highlighted
+
+ May be worth investigating further. * Worth investigating further.
```

**Fig8. Result page for Procheck**





**Fig8.2. 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 is an integrated server containing various tools on a single platform that can be used for tertiary structure validation. The predicted structure for kinase 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>

## WEBLEM 5

### Introduction to Visualization of Tertiary structure using RASMOL & PyMOL

Once a protein structure has been solved, the structure has to be presented in a three dimensional view on the basis of the solved Cartesian coordinates. Before computer visualization software was developed, molecular structures were represented by physical models of metal wires, rods, and spheres. With the development of computer hardware and software technology, sophisticated computer graphics programs have been developed for visualizing and manipulating complicated three-dimensional structures. The computer graphics help to analyze and compare protein structures to gain insight to functions of the proteins.

The main feature of computer visualization programs is interactivity, which allows users to visually manipulate the structural images through a graphical user interface. At the touch of a mouse button, a user can move, rotate, and zoom an atomic model on a computer screen in real time, or examine any portion of the structure in great detail, as well as draw it in various forms in different colors. Further manipulations can include changing the conformation of a structure by protein modeling or matching a ligand to an enzyme active site through docking exercises.

Because a Protein Data Bank (PDB) data file for a protein structure contains only *x*, *y*, and *z* coordinates of atoms, the most basic requirement for a visualization program is to build connectivity between atoms to make a view of a molecule. The visualization program should also be able to produce molecular structures in different styles, which include wire frames, balls and sticks, space-filling spheres, and ribbons.

A wire-frame diagram is a line drawing representing bonds between atoms. The wire frame is the simplest form of model representation and is useful for localizing positions of specific residues in a protein structure, or for displaying a skeletal form of a structure when  $C\alpha$  atoms of each residue are connected. Balls and sticks are solid spheres and rods, representing atoms and bonds, respectively. These diagrams can also be used to represent the backbone of a structure. In a space-filling representation (or Corey, Pauling, and Koltan [CPK]), each atom is described using large solid spheres with radii corresponding to the van der Waals radii of the atoms. Ribbon diagrams use cylinders or spiral ribbons to represent  $\alpha$ -helices and broad, flat arrows to represent  $\beta$ -strands. This type of representation is very attractive in that it allows easy identification of secondary structure elements and gives a clear view of the overall topology of the structure. The resulting images are also visually appealing.

Different representation styles can be used in combination to highlight a certain feature of a structure while deemphasizing the structures surrounding it. For example, a cofactor of an enzyme can be shown as space-filling spheres while the rest of the protein structure is shown as wire frames or ribbons. RASMOL and PyMOL are tools used for visualisation of tertiary structure of proteins.

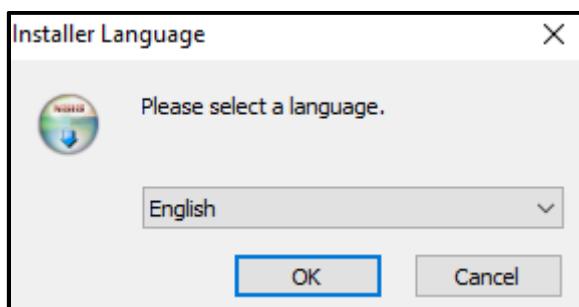
#### **RASMOL:**

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.

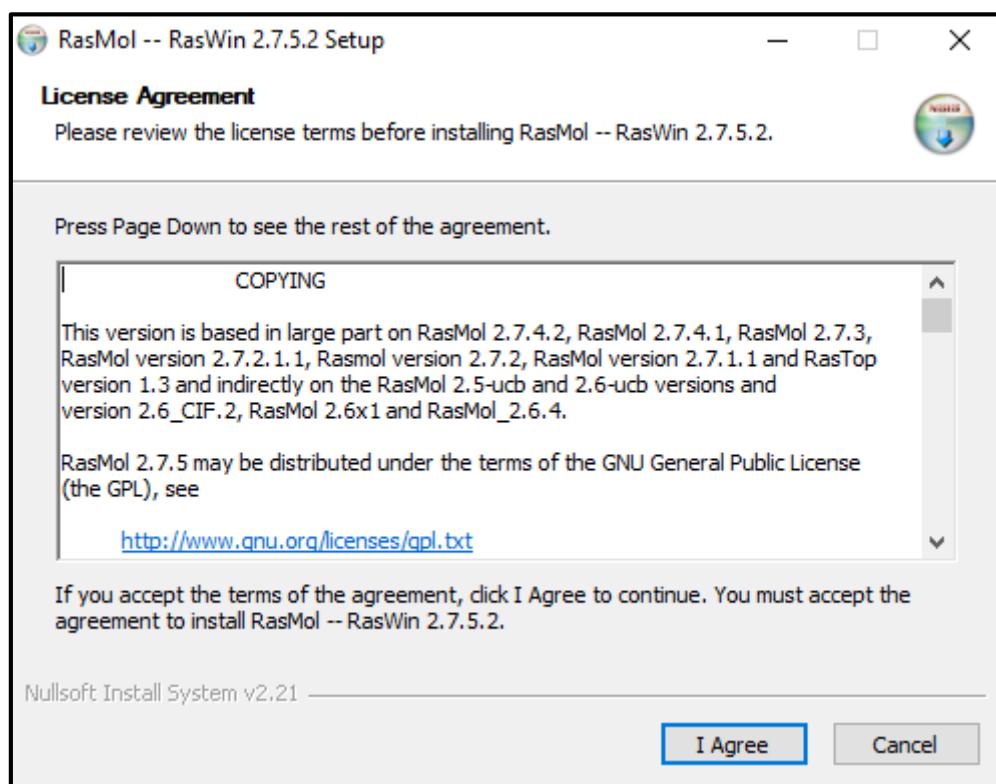
RasMol (<http://rutgers.rcsb.org/pdb/help-graphics.html#rasmol> download) is a command-line-based viewing program that calculates connectivity of a coordinate file and displays wireframe, cylinder, stick bonds,  $\alpha$ -carbon trace, space-filling (CPK) spheres, and ribbons. It reads both PDB and mmCIF formats and can display a whole molecule or specific parts of it. It is available in multiple platforms: UNIX, Windows, and Mac.

RasTop([www.geneinfinity.org/rastop/](http://www.geneinfinity.org/rastop/)) is a new version of RasMol for Windows with a more enhanced user interface.

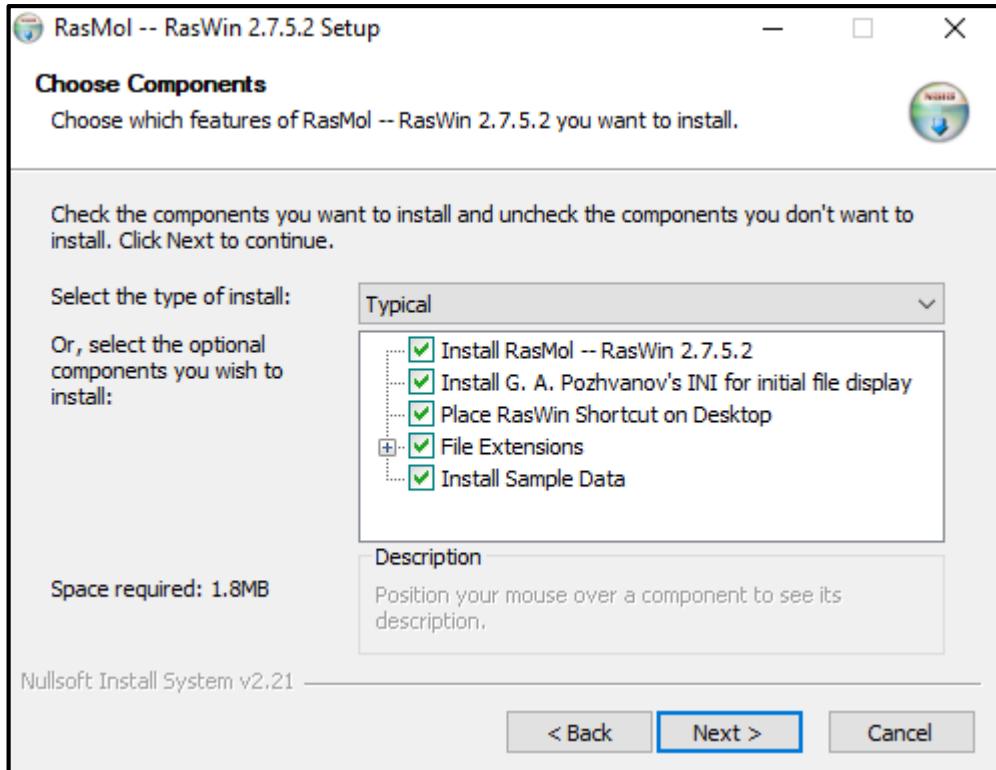
## Installation:



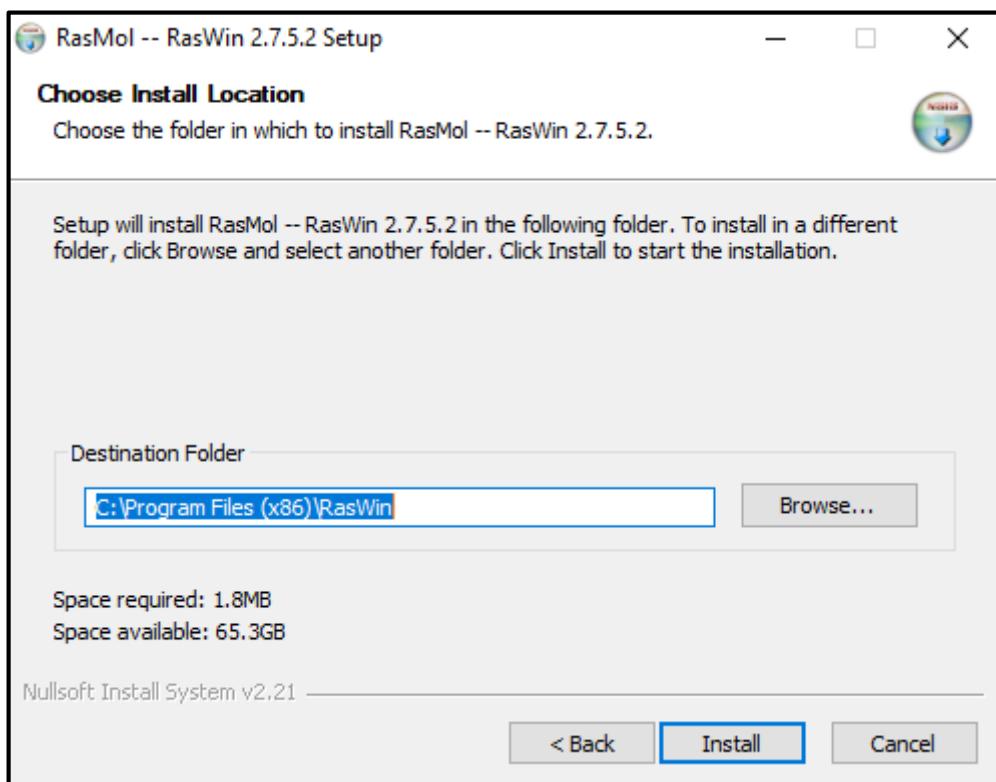
**Fig1. Select language and proceed**



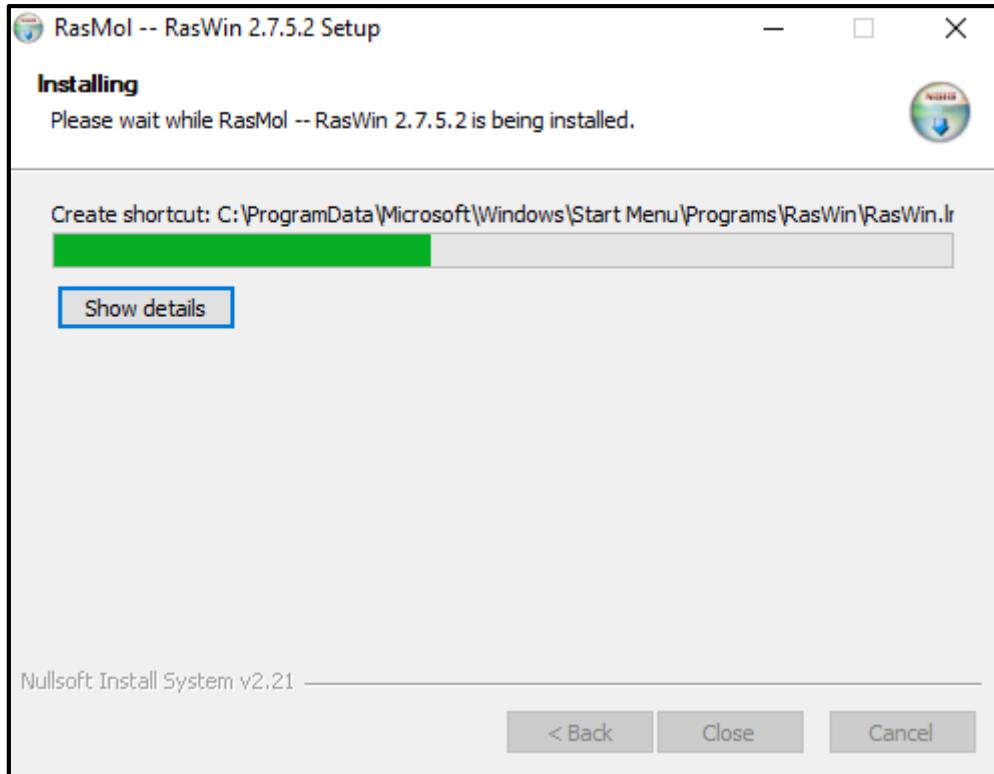
**Fig2. Read the License Agreement and click "I Agree" to proceed**



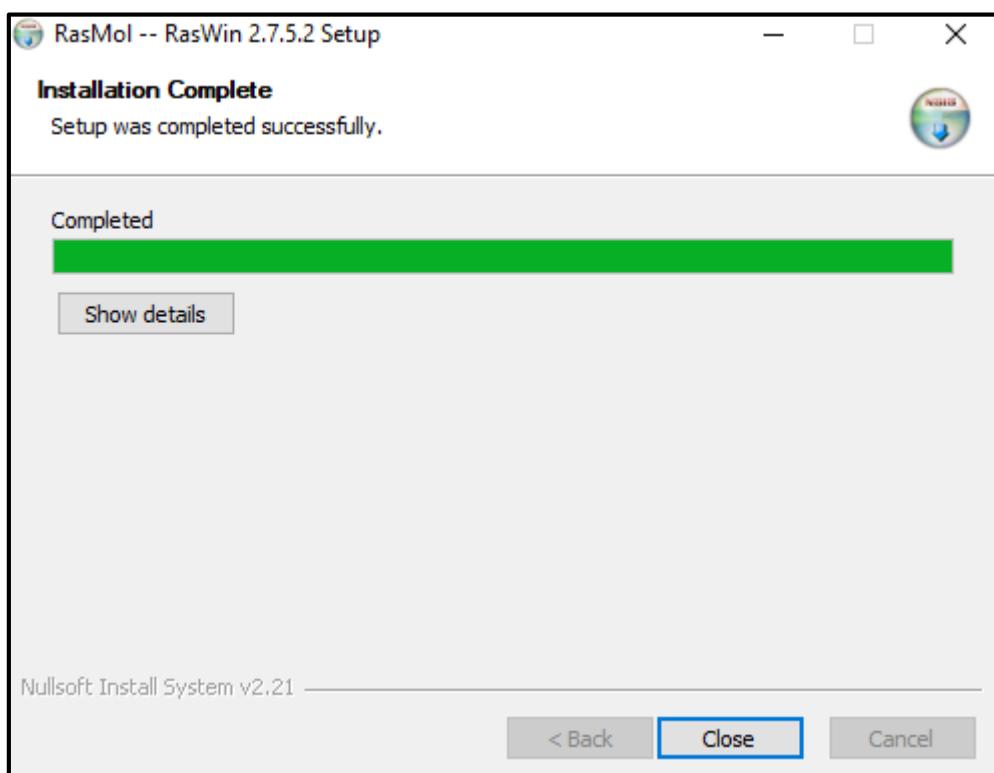
**Fig3.** Select the components of RasMol you want to install and click “Next” to proceed



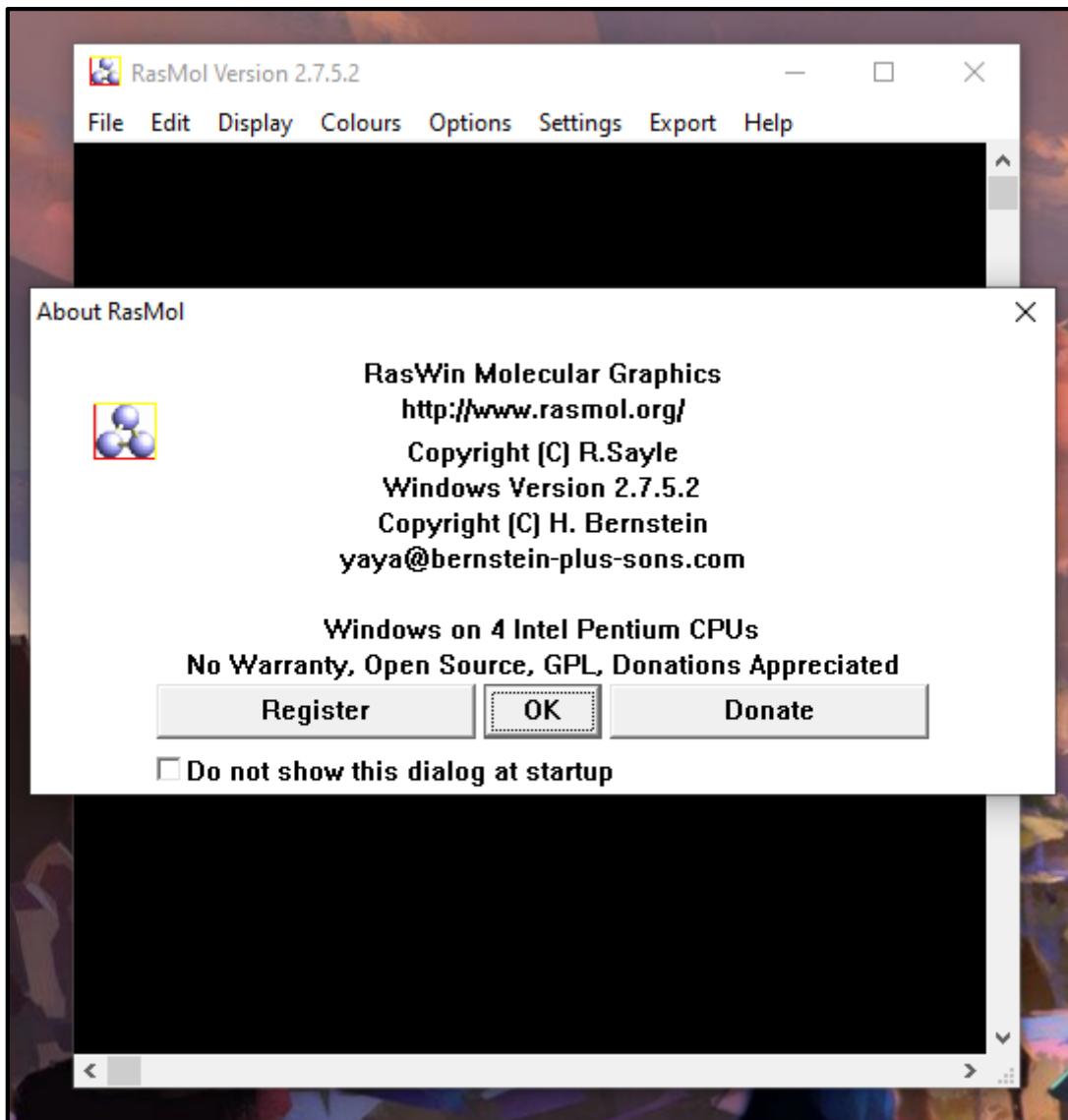
**Fig4.** Select the directory in which you want to install RasMol and click “Install” to proceed



**Fig5.** The software will now install



**Fig6.** Close the installer after the install is completed

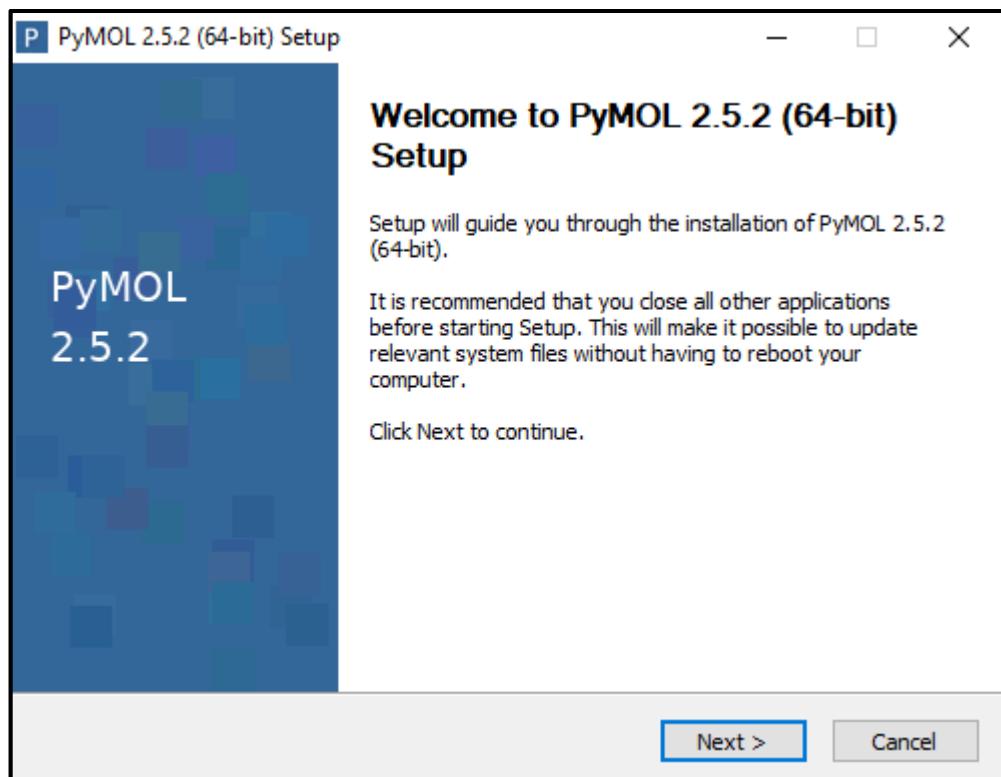


**Fig7. RasMol is now installed**

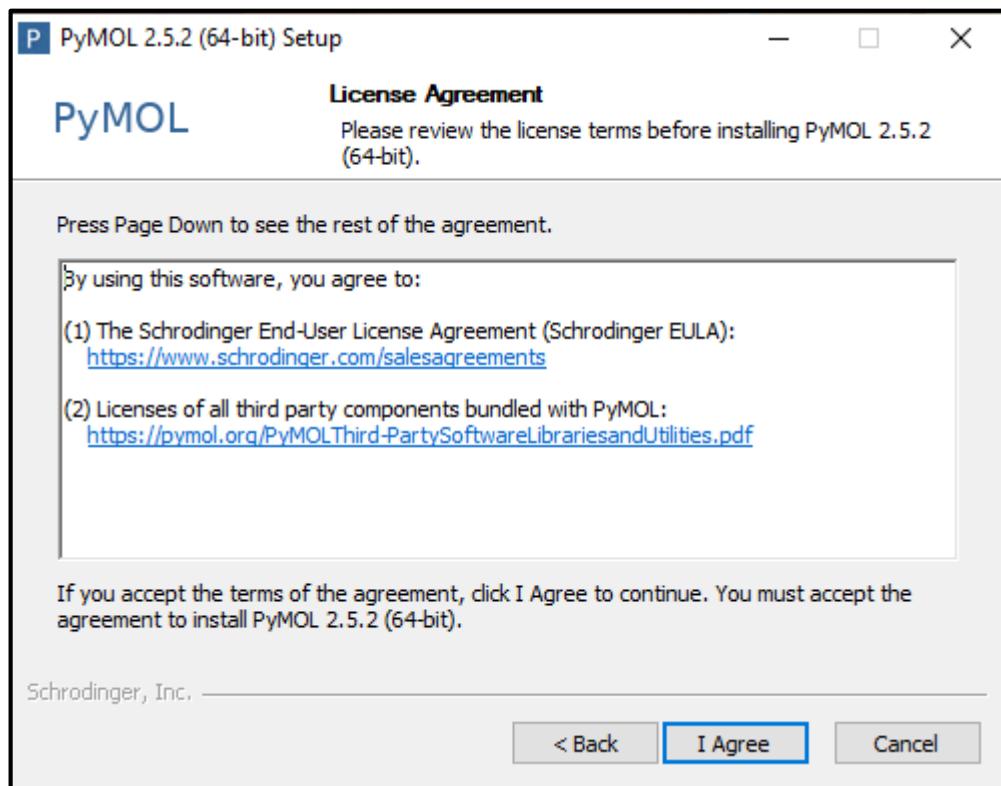
### PyMOL:

PyMOL is an open-source molecular visualization system created by Warren Lyford DeLano and commercialized initially by DeLano Scientific LLC. In 2010, Schrödinger Inc. reached an agreement to acquire PyMOL. It is 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. PyMOL can produce high-quality movies and images of macro-molecules in different representations including ribbons, cartoons, dots, surfaces, spheres, sticks, and lines. 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, there are various molecular modeling modules in PyMOL, covering those for visualization and analysis enhancement, protein–ligand modeling, molecular simulations, and drug screening.

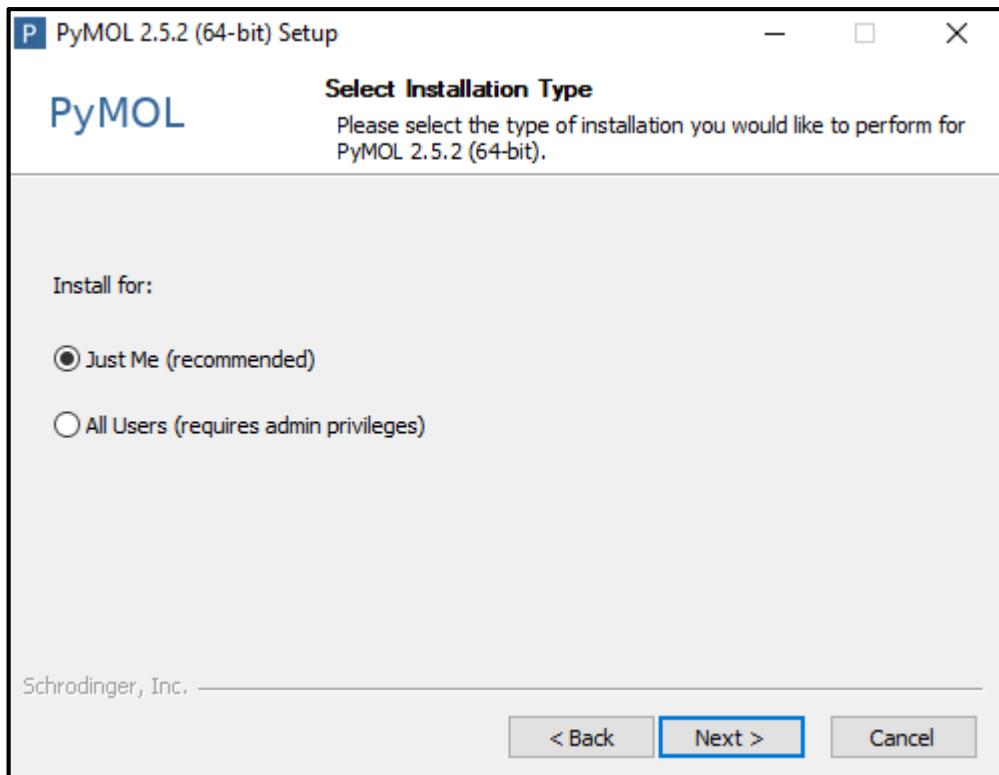
## Installation:



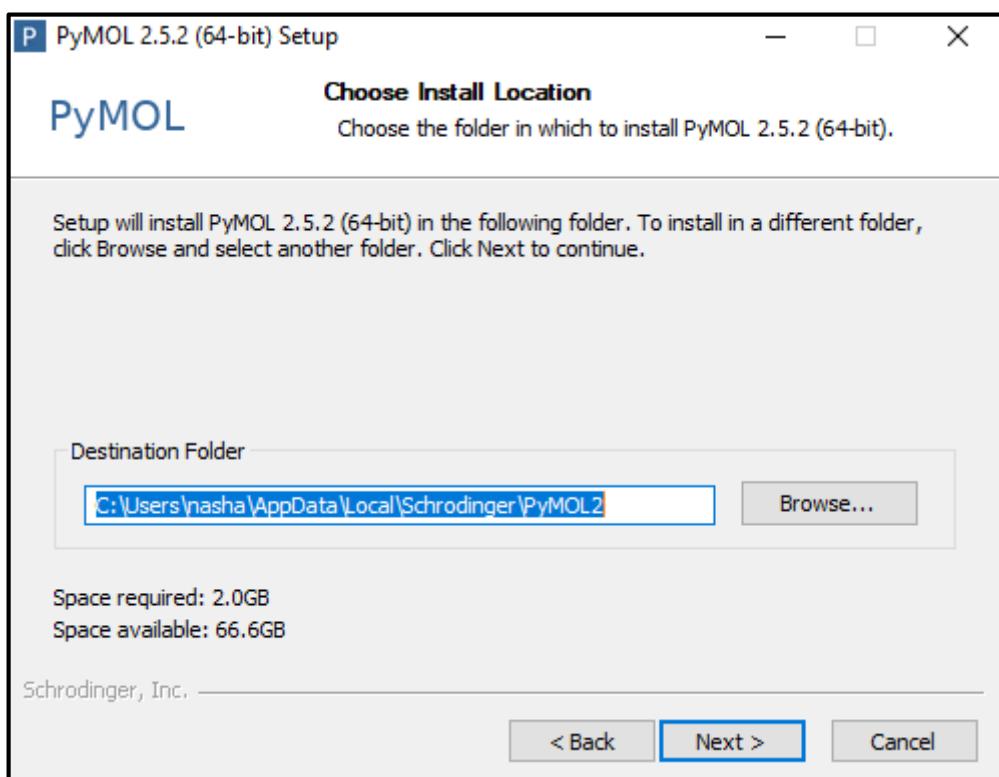
**Fig1.** Download and run the installer. Click next to proceed



**Fig2.** Read the License Agreement and click "I Agree" to proceed



**Fig3.** Select the Installation type and click “Next” to proceed



**Fig4.** Select the Directory in which you want to install PyMol

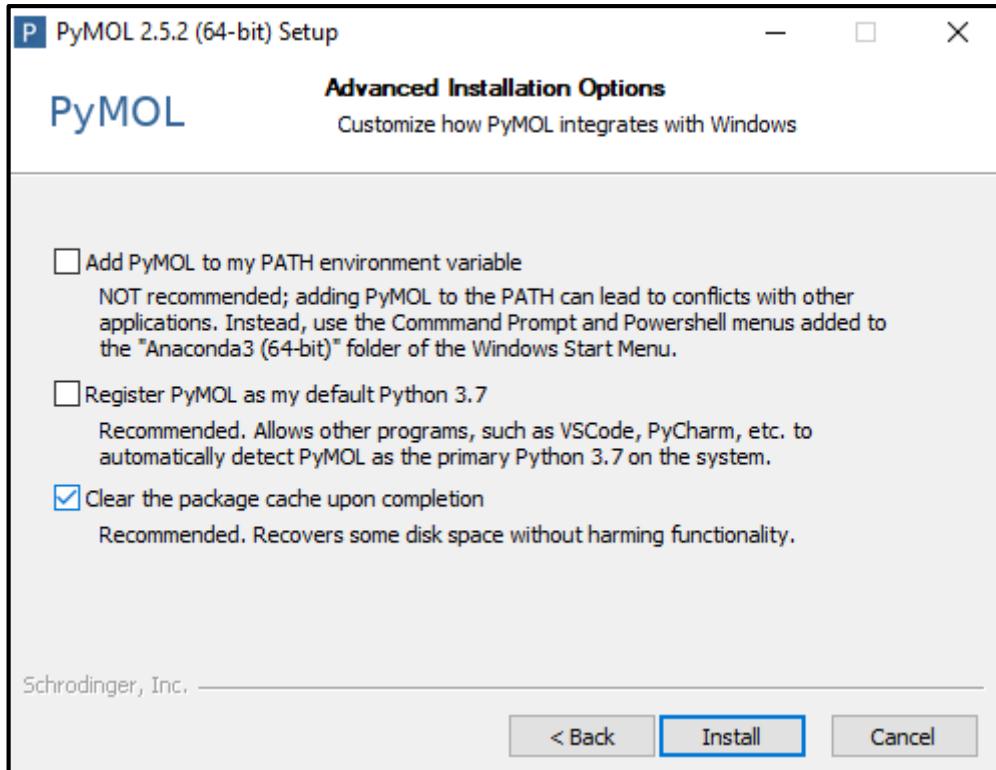


Fig5. Select the additional options if you want to and click “Install” to start the installation of PyMol

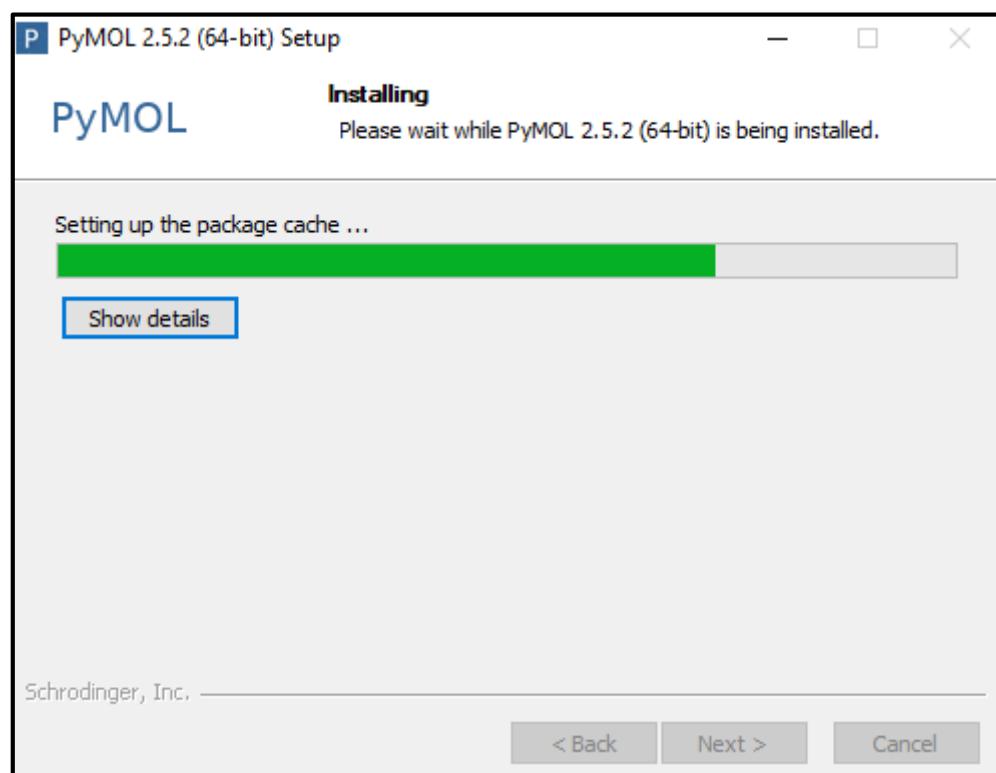
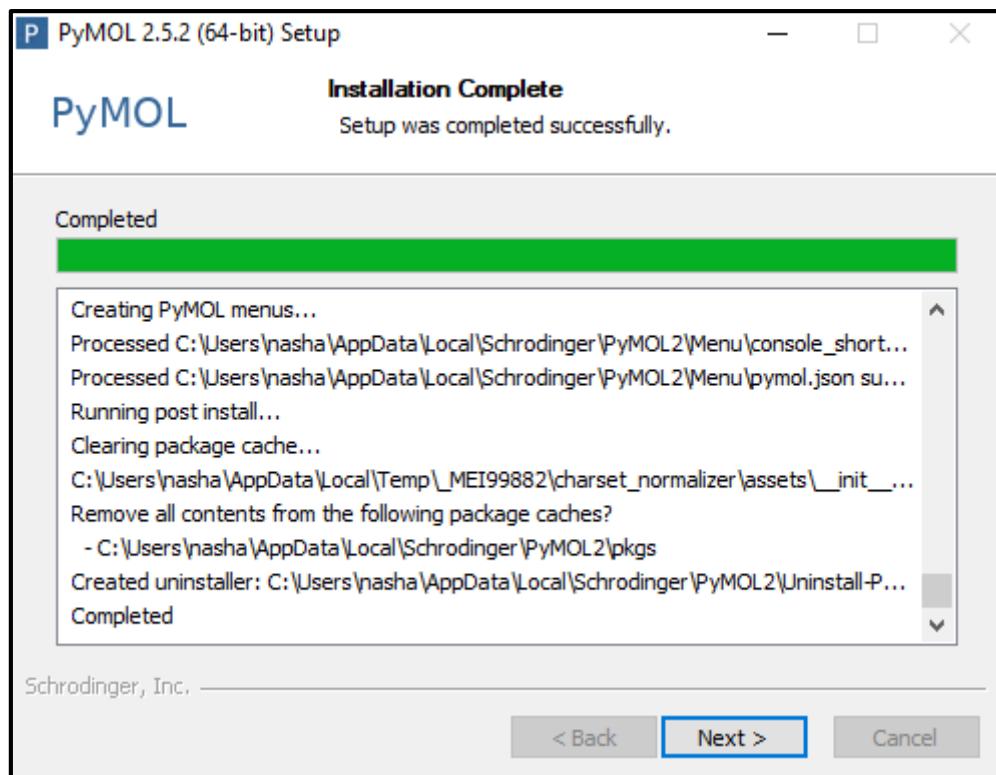
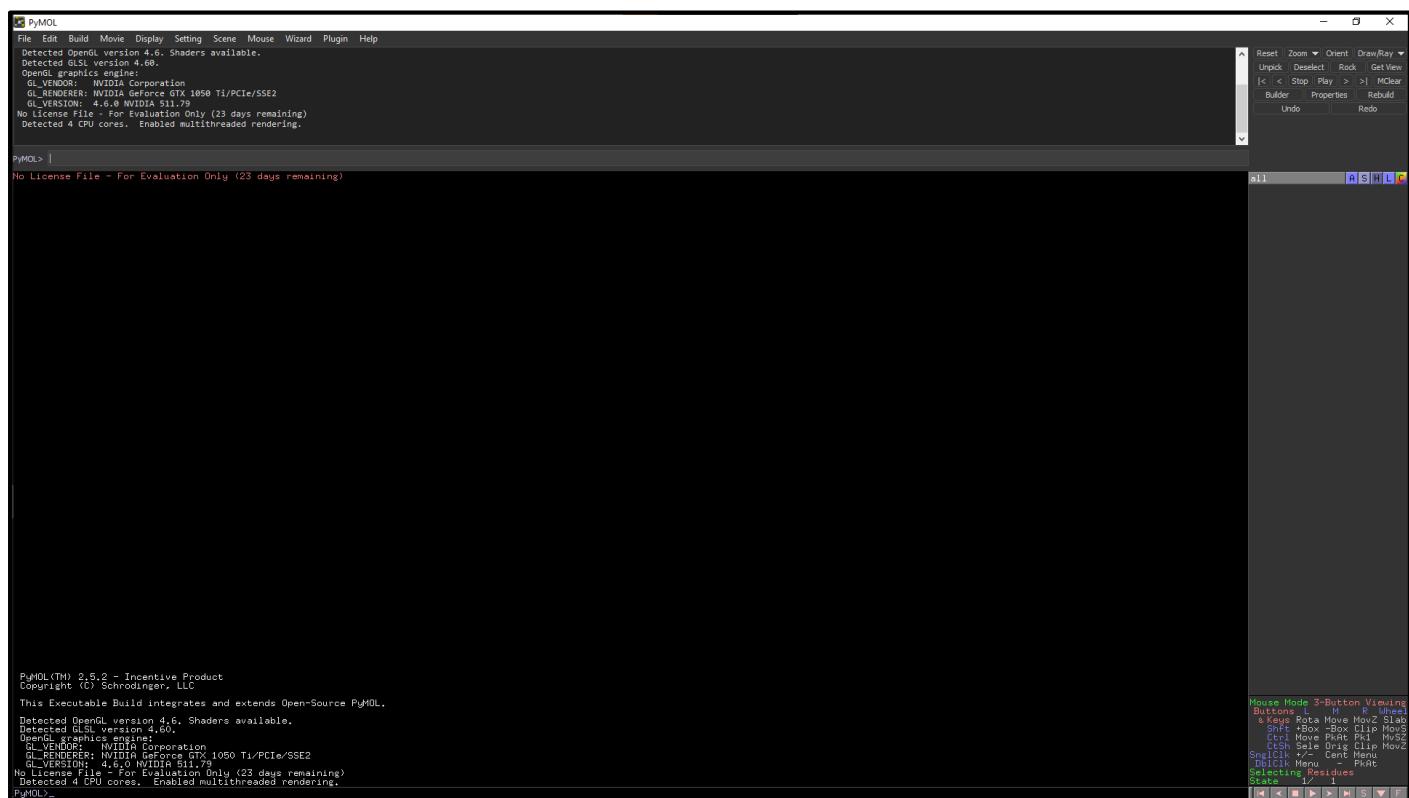


Fig6. Let the installer install PyMol



**Fig7. After completion click on “Next” to finalize installation**



## Fig8. 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](http://www.openrasmol.org). Retrieved March 4, 2022, from <http://www.openrasmol.org/>
4. *PyMOL / pymol.org*. (2019). [Pymol.org](https://pymol.org/2/). Retrieved March 4, 2022, from <https://pymol.org/2/>

## WEBLEM 5a

### RASMOL & PyMOL

(URL: <http://www.openrasmol.org/> and <https://pymol.org/2/> )

#### **AIM:**

To visualize 3D structure of Fibrin using RASMOL & PyMOL tool.

#### **INTRODUCTION:**

Fibrin, an insoluble protein that is produced in response to bleeding and is the major component of the blood clot. Fibrin is a tough protein substance that is arranged in long fibrous chains; it is formed from fibrinogen, a soluble protein that is produced by the liver and found in blood plasma. When tissue damage results in bleeding, fibrinogen is converted at the wound into fibrin by the action of thrombin, a clotting enzyme. Fibrin molecules then combine to form long fibrin threads that entangle platelets, building up a spongy mass that gradually hardens and contracts to form the blood clot. This hardening process is stabilized by a substance known as fibrin-stabilizing factor, or factor XIII. The PDB structure of fibrin can be visualised using RASMOL and PyMOL tools.

#### **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. RasMol (<http://rutgers.rcsb.org/pdb/help-graphics.html#rasmol> download) is a command-line-based viewing program that calculates connectivity of a coordinate file and displays wireframe, cylinder, stick bonds,  $\alpha$ -carbon trace, space-filling (CPK) spheres, and ribbons. It reads both PDB and mmCIF formats and can display a whole molecule or specific parts of it. It is available in multiple platforms: UNIX, Windows, and Mac. RasTop([www.geneinfinity.org/rastop/](http://www.geneinfinity.org/rastop/)) is a new version of RasMol for Windows with a more enhanced user interface.

#### **PyMol**

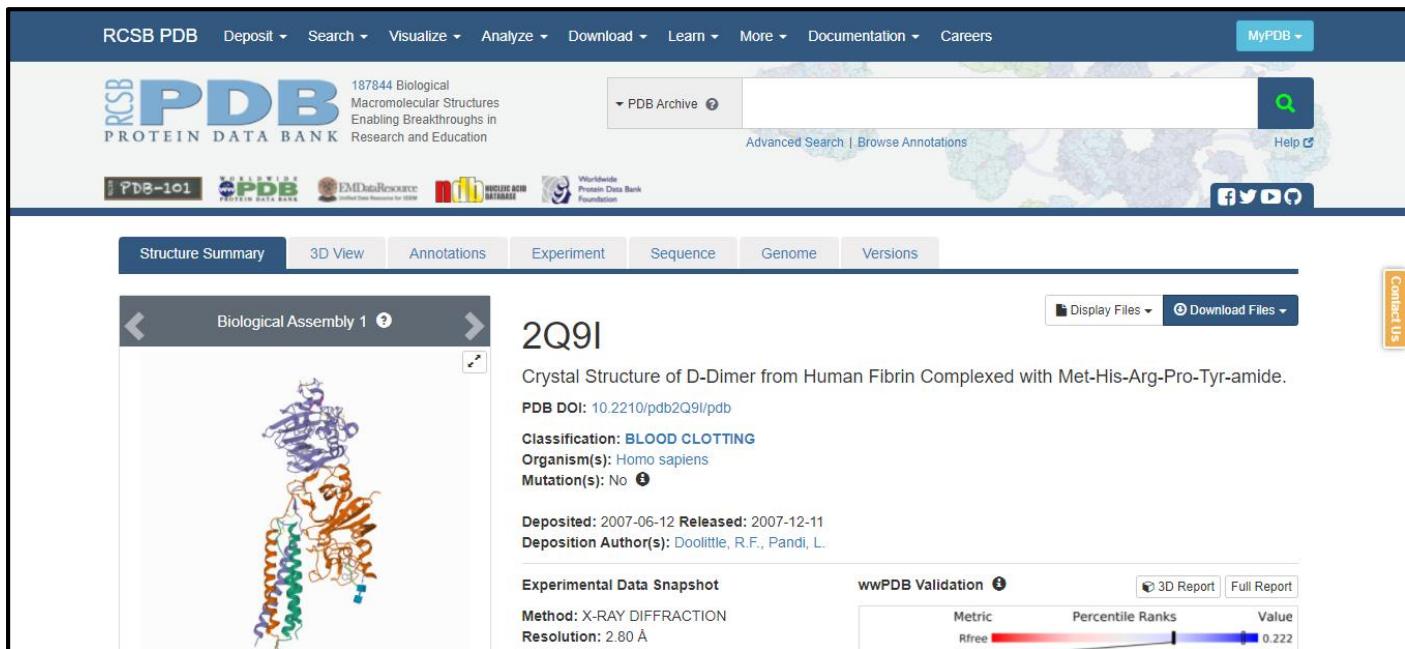
PyMOL is an open-source molecular visualization system created by Warren Lyford DeLano and commercialized initially by DeLano Scientific LLC. In 2010, Schrödinger Inc. reached an agreement to acquire PyMOL. It is a cross-platform molecular graphics tool, has been widely used for three-dimensional (3D) visualization of proteins, nucleic acids, small molecules, electron densities, surfaces, and trajectories. PyMOL can produce high-quality movies and images of macro-molecules in different representations including ribbons, cartoons, dots, surfaces, spheres, sticks, and lines. 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, there are various molecular modeling modules in PyMOL, covering those for visualization and analysis enhancement, protein–ligand modeling, molecular simulations, and drug screening.

#### **METHODOLOGY:**

1. Retrieve Fibrin structure from PDB.
2. Install RASMOL and PyMOL tools. (URL: <http://www.openrasmol.org/> and <https://pymol.org/2/>)
3. Visualise the PDB structure on both tools.
4. Observe and interpret the results.

## OBSERVATION:

### RASMOL:



RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB

187844 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB Archive Advanced Search | Browse Annotations Help

PDB-101 Worldwide Protein Data Bank EMDataResource Molecule of the Month Worldwide Protein Data Bank Foundation

Structure Summary 3D View Annotations Experiment Sequence Genome Versions

2Q9I

Crystal Structure of D-Dimer from Human Fibrin Complexed with Met-His-Arg-Pro-Tyr-amide.

PDB DOI: 10.2210/pdb2Q9I/pdb

Classification: BLOOD CLOTTING

Organism(s): Homo sapiens

Mutation(s): No

Deposited: 2007-06-12 Released: 2007-12-11

Deposition Author(s): Doolittle, R.F., Pandi, L.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION Resolution: 2.80 Å

wwPDB Validation 3D Report Full Report

Metric Percentile Ranks Value Rfree 0.222

Fig1. Fibrin structure retrieved from PDB

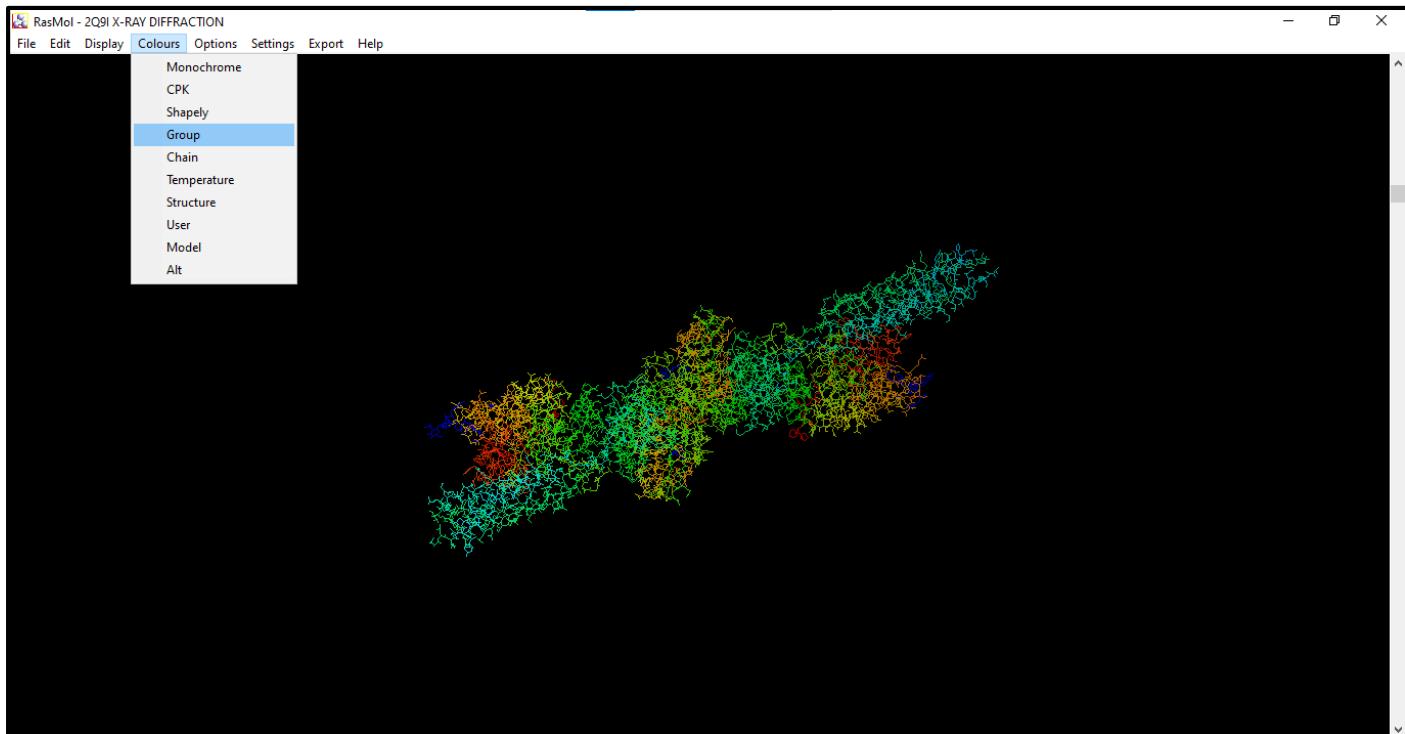
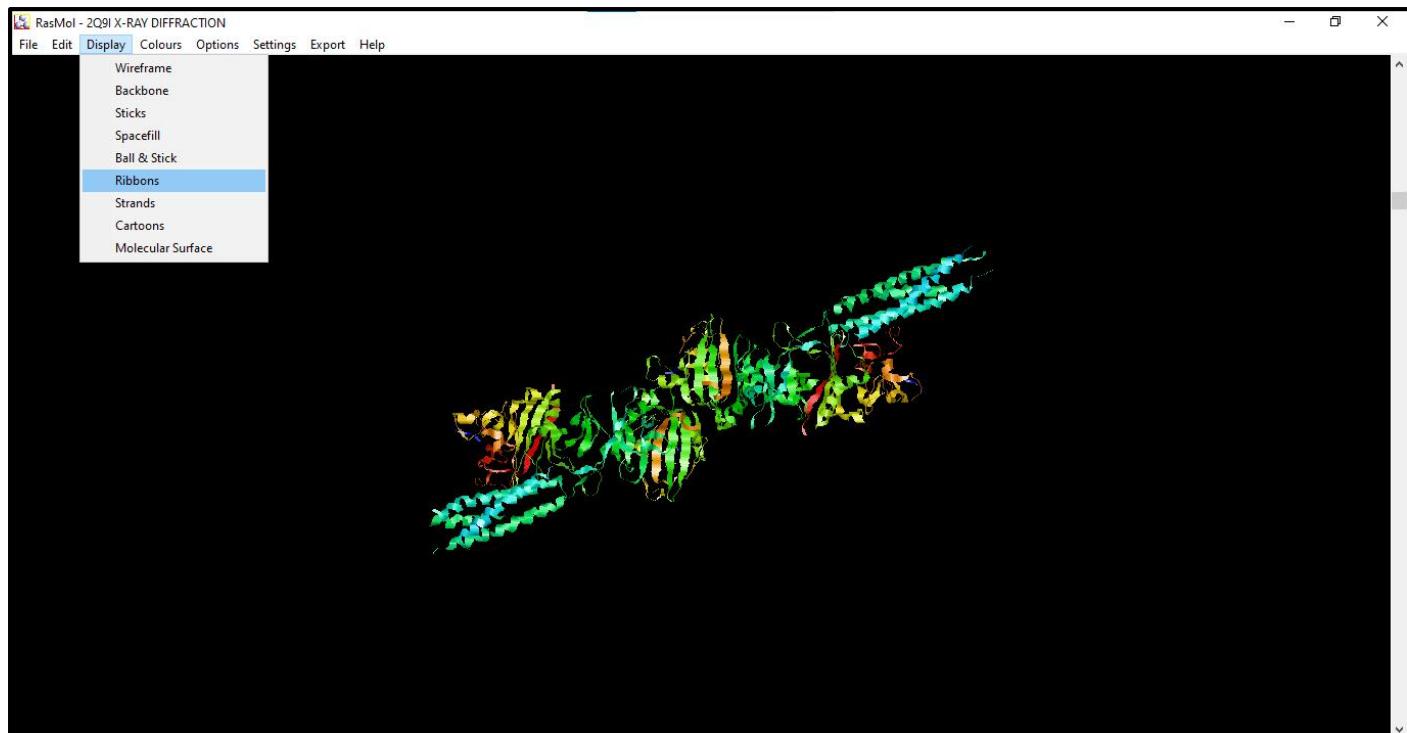
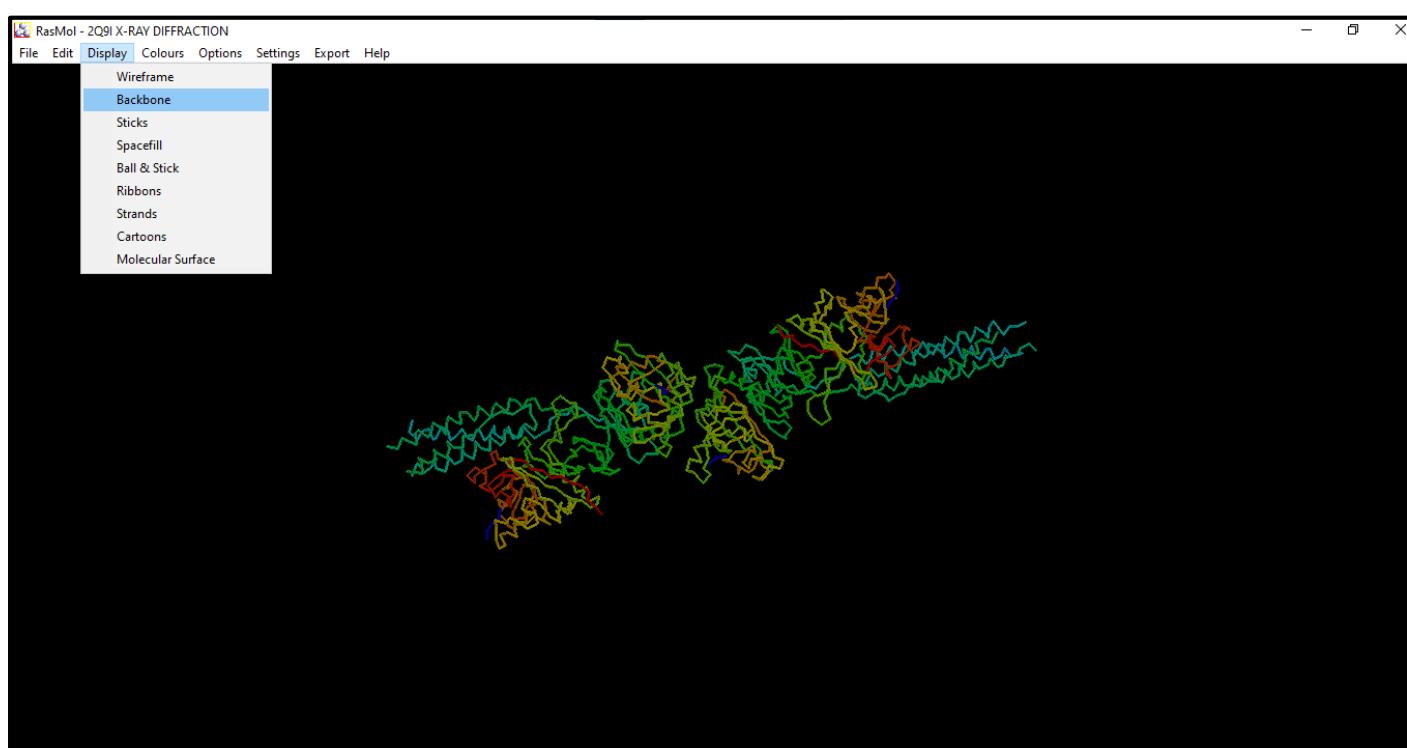


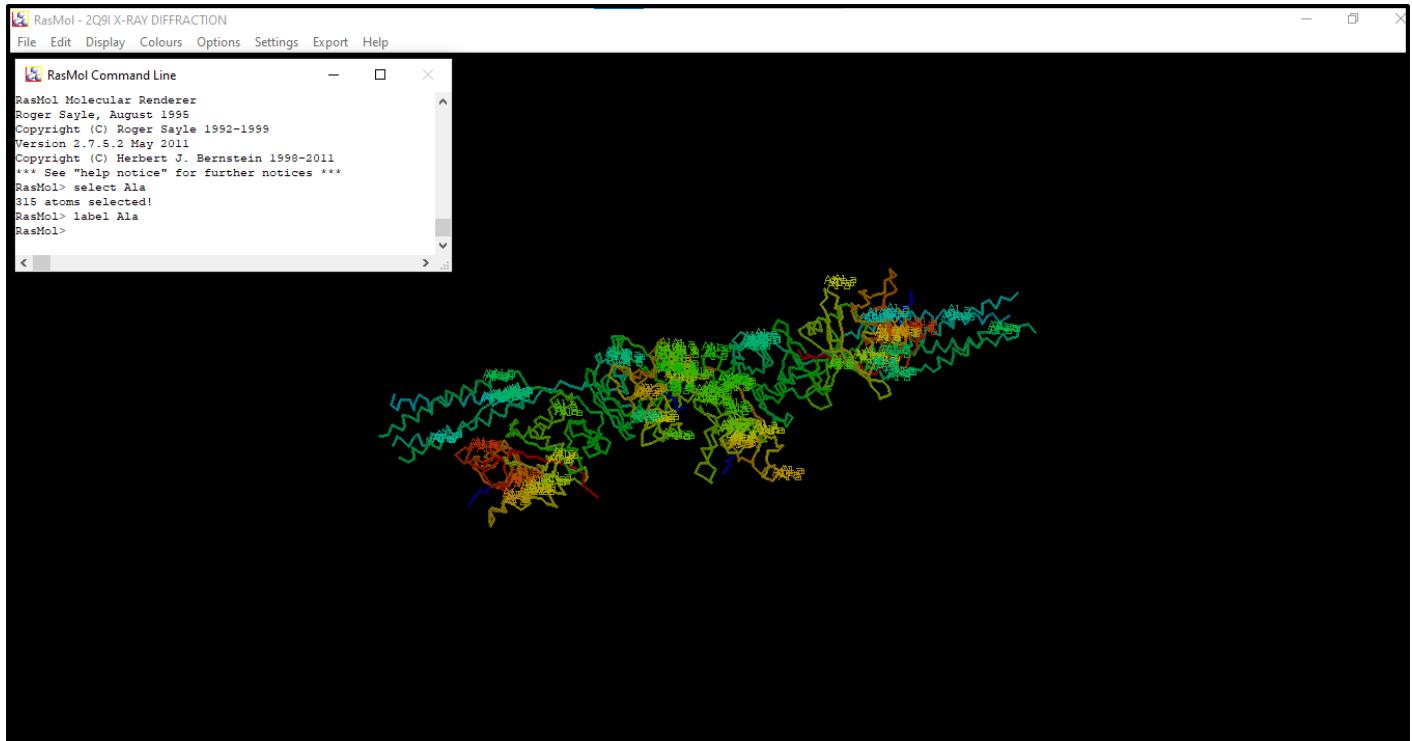
Fig2. Structure of Fibrin in group colour format



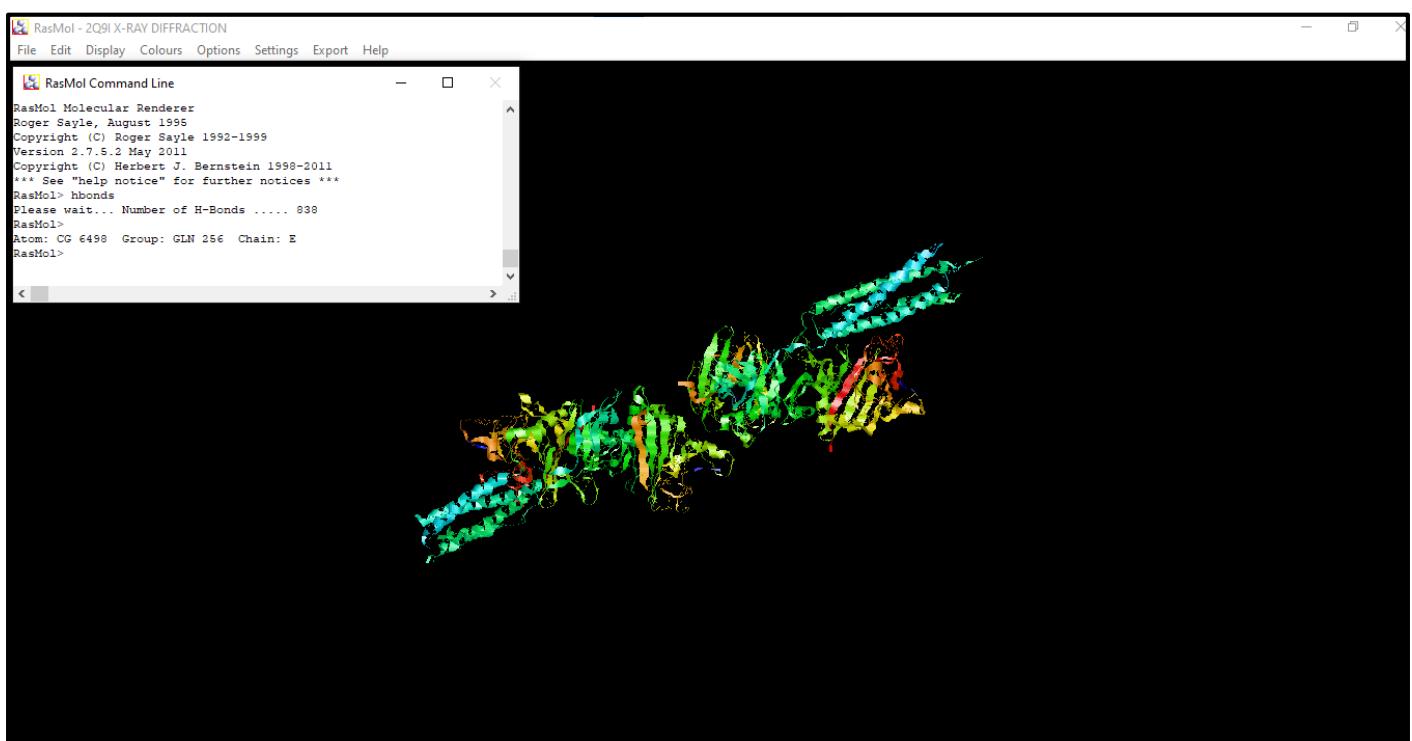
**Fig3. Fibrin structure in ribbons format**



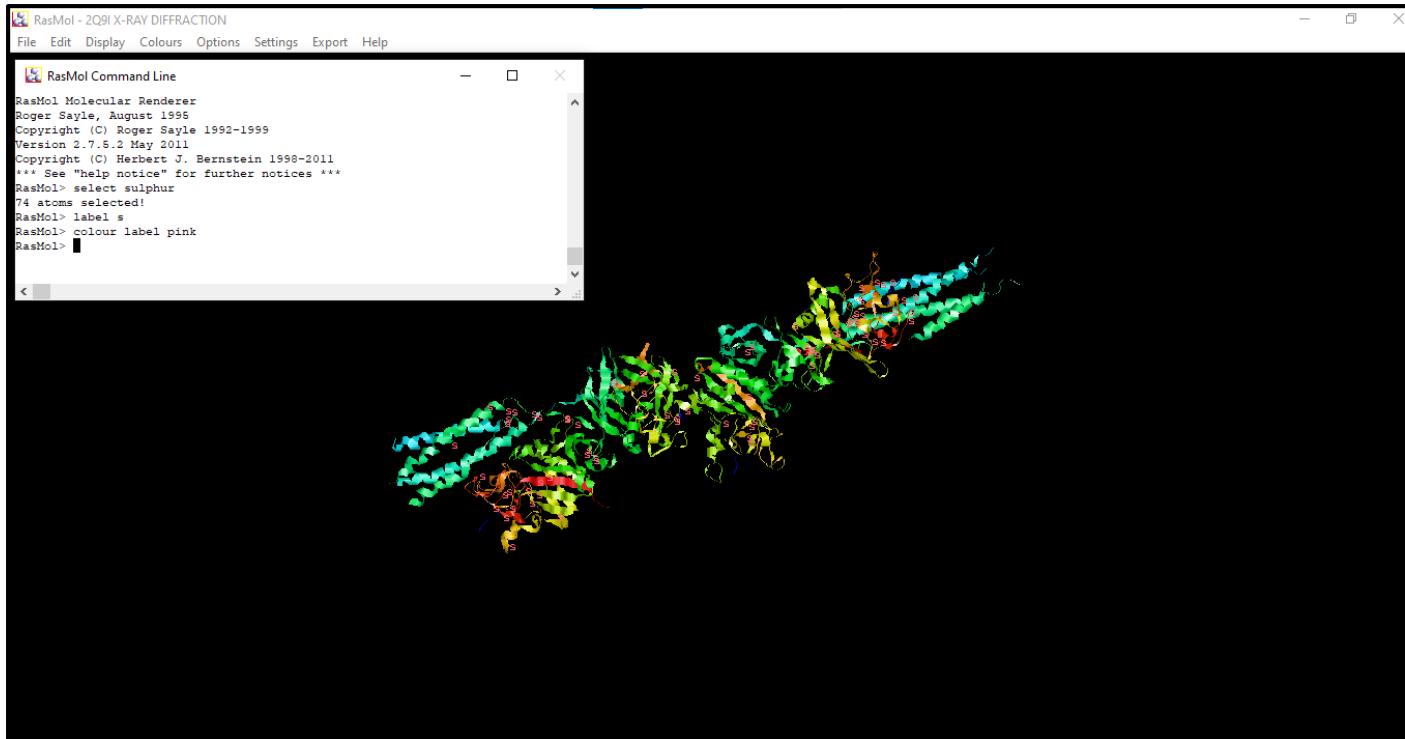
**Fig4. Fibrin structure in backbone format**



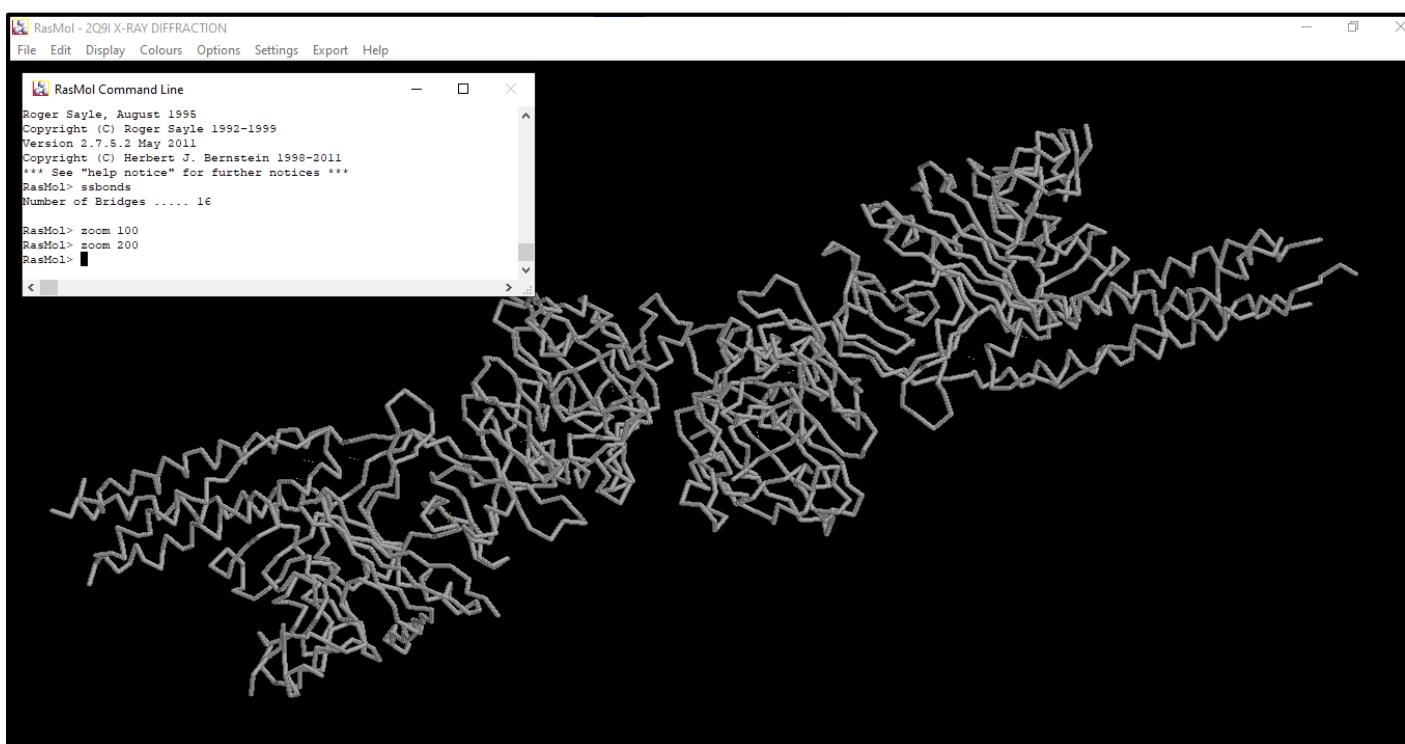
**Fig5. Visualisation of Alanine residues in Fibrin**



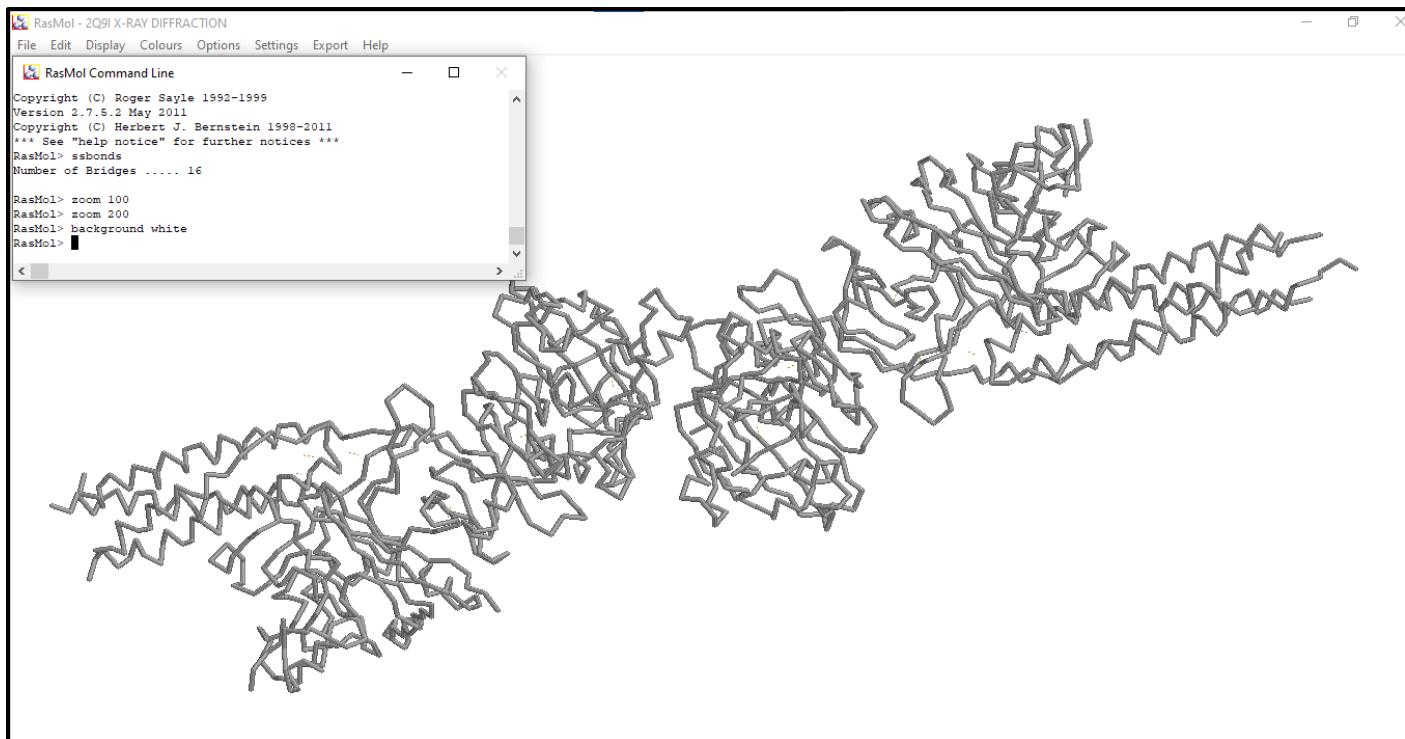
**Fig6. Visualisation of H-bonds in Fibrin**



**Fig7. Visualisation of sulphur residues in Fibrin**



**Fig8. Visualisation of disulphide bonds in Fibrin**



**Fig9. Visualisation of structure with white background**

**PyMOL:**

Structure Summary    3D View    Annotations    Experiment    Sequence    Genome    Versions

**2Q9I**

Crystal Structure of D-Dimer from Human Fibrin Complexed with Met-His-Arg-Pro-Tyr-amide.

**PDB DOI:** 10.2210/pdb2Q9I/pdb

**Classification:** BLOOD CLOTTING  
**Organism(s):** Homo sapiens  
**Mutation(s):** No

**Deposited:** 2007-06-12 **Released:** 2007-12-11  
**Deposition Author(s):** Doolittle, R.F., Pandi, L.

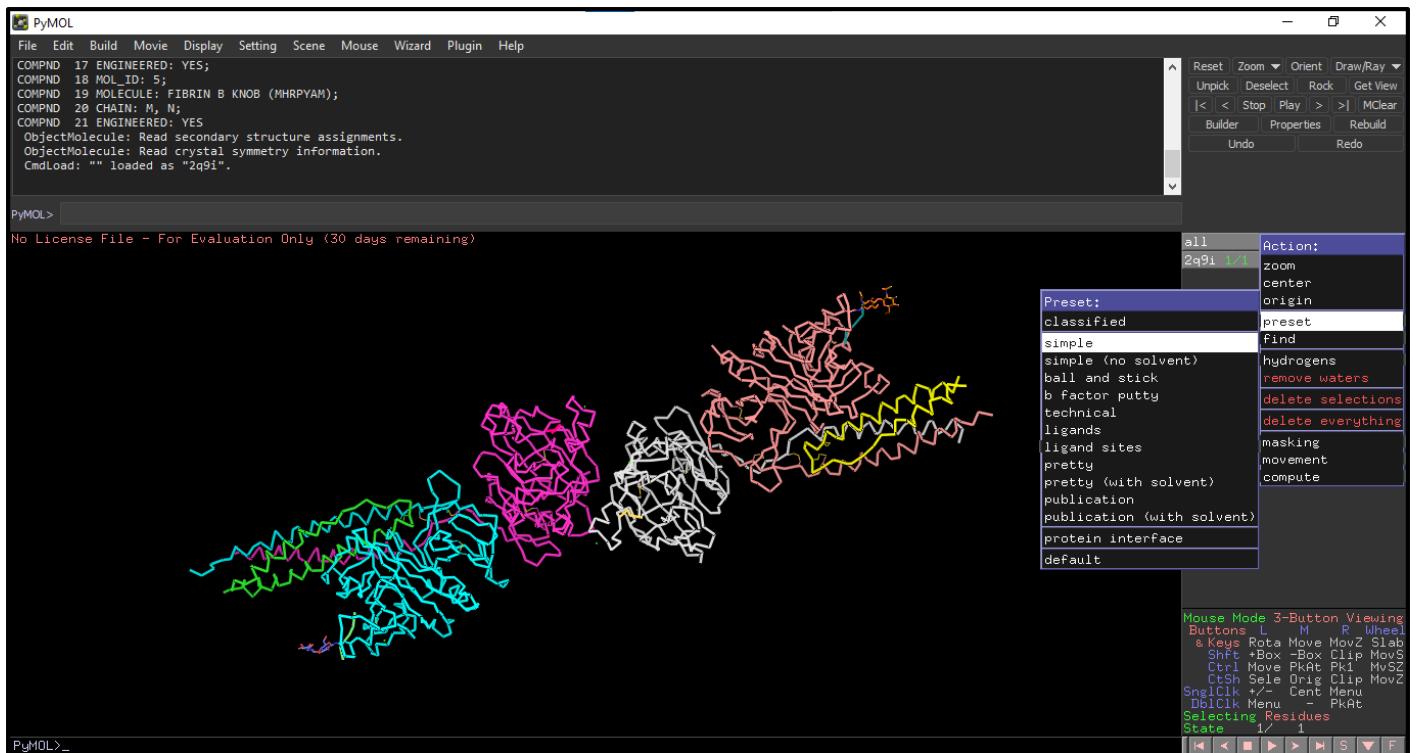
**Experimental Data Snapshot**

**Method:** X-RAY DIFFRACTION  
**Resolution:** 2.80 Å

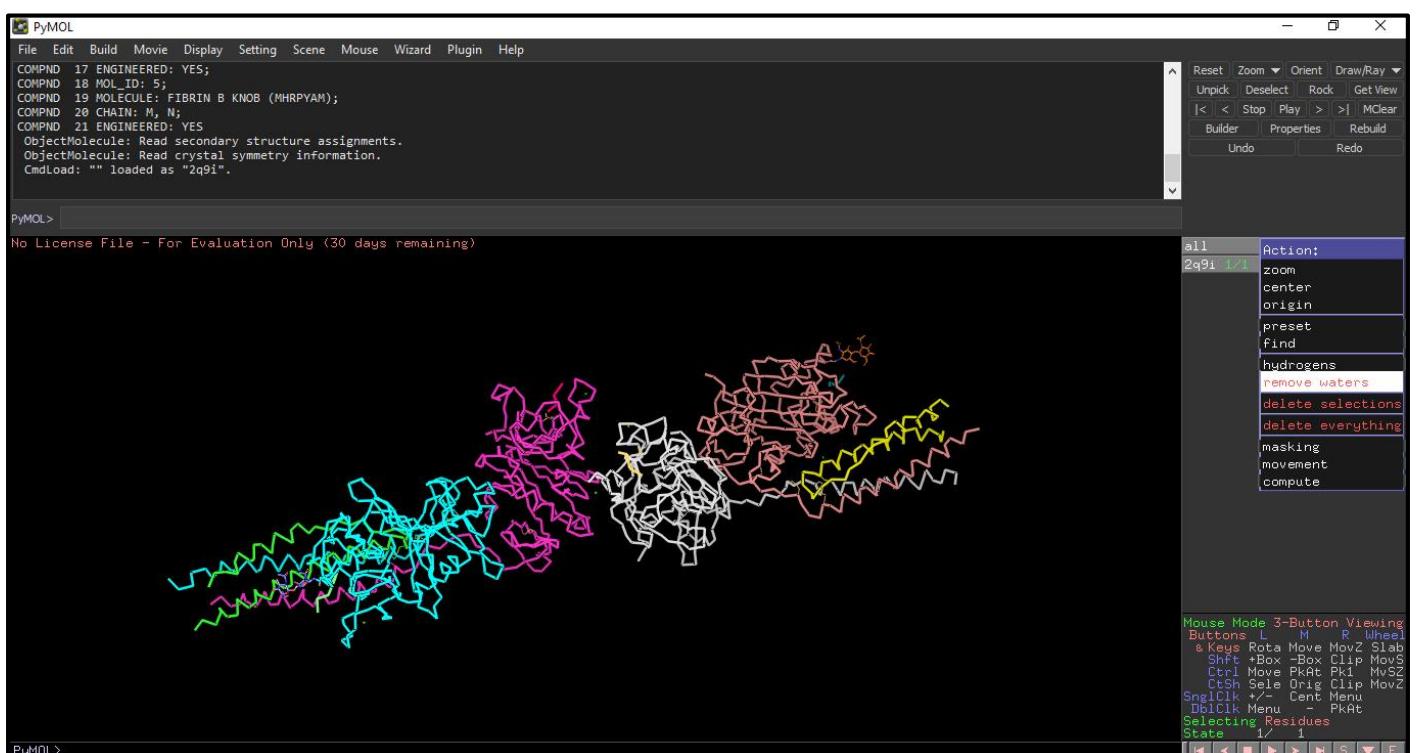
**wwPDB Validation**

Metric	Percentile Ranks	Value
Rfree	0.222	0.222

**Fig1. Fibrin structure retrieved from PDB**



**Fig2. Visualisation of Fibrin in simple format**



**Fig3. Removal of water from Fibrin**

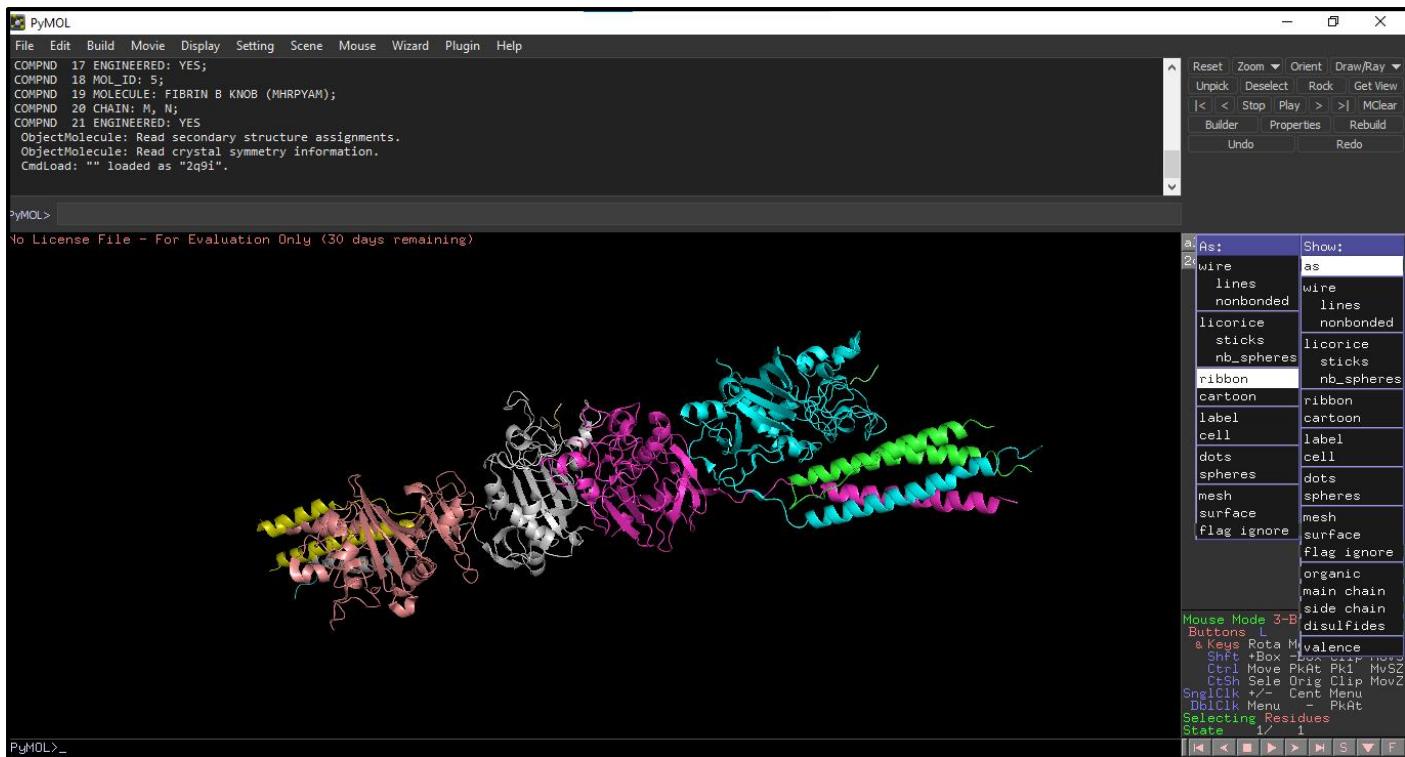


Fig4. Visualisation of Fibrin in ribbon format

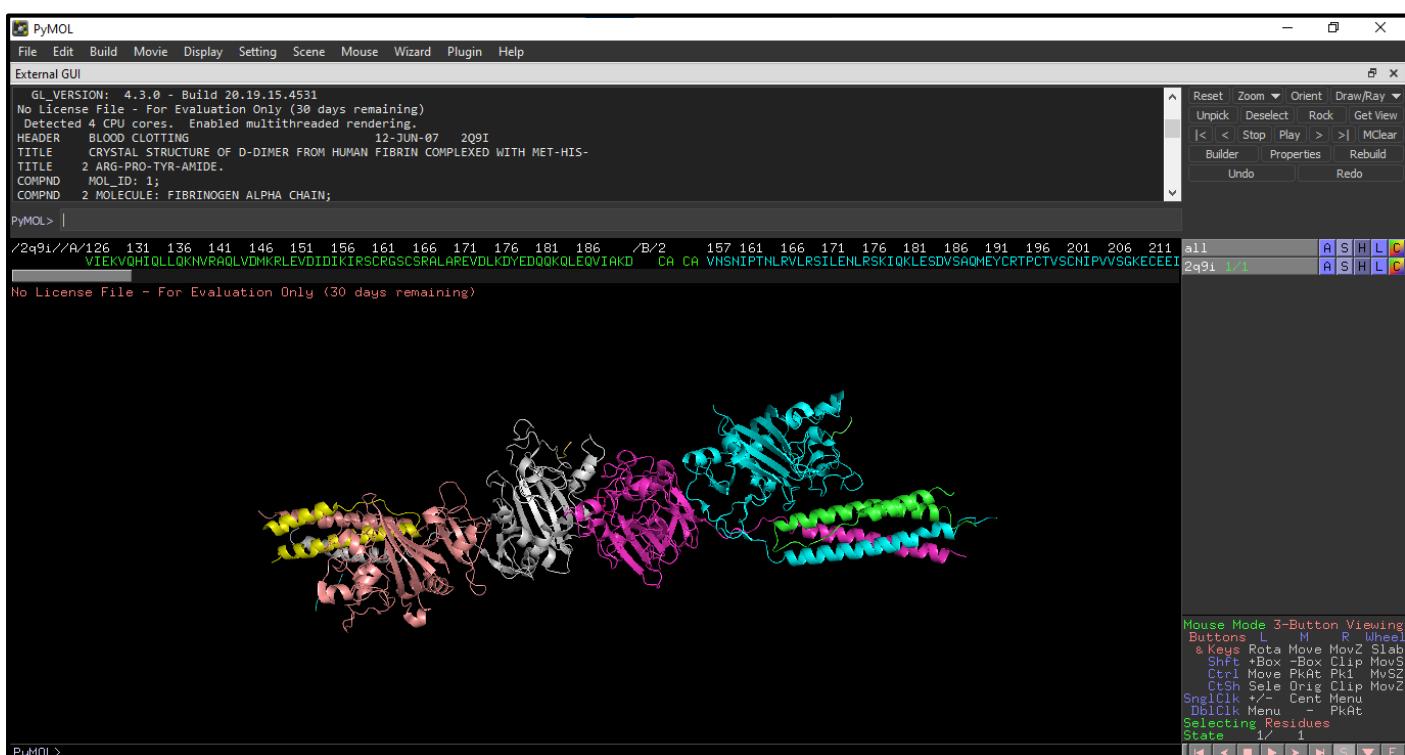


Fig5. Displaying sequence for Fibrin

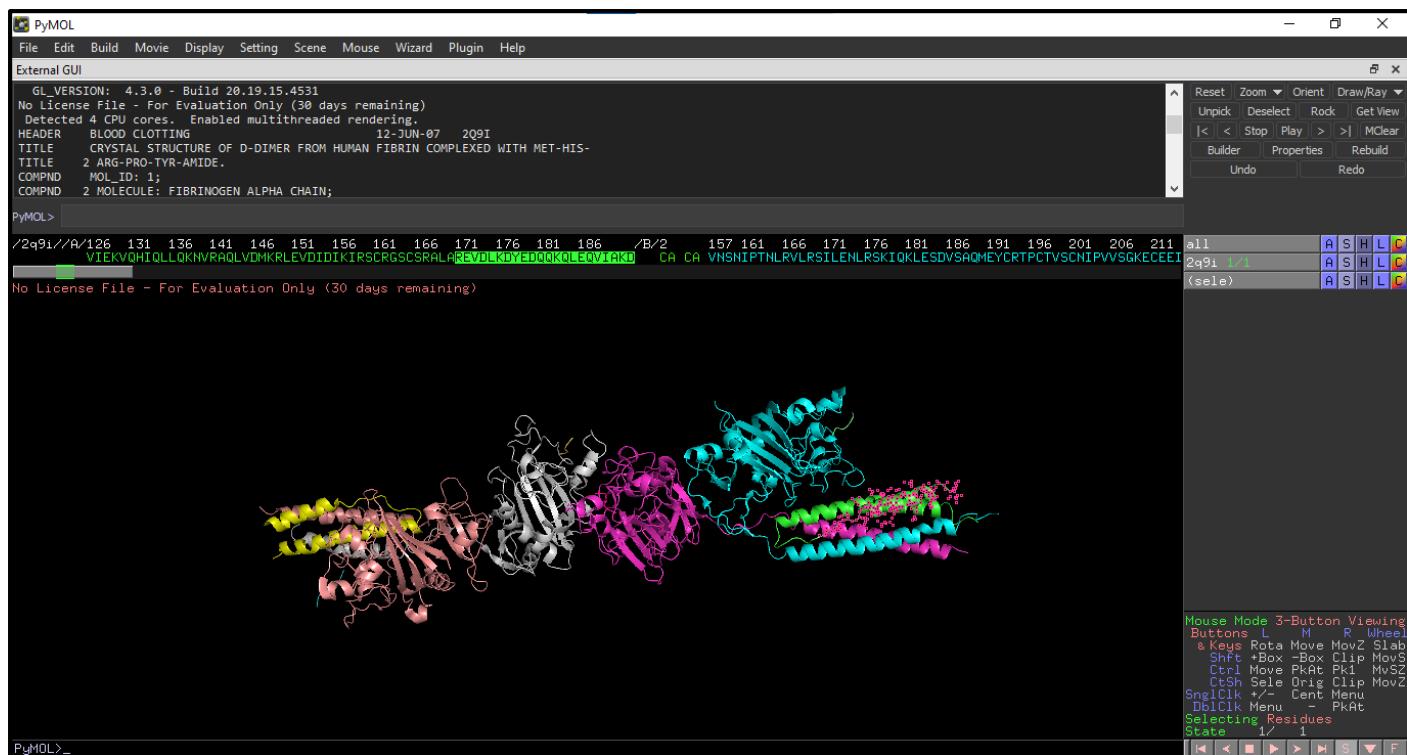


Fig6. Displaying specific residues from sequence

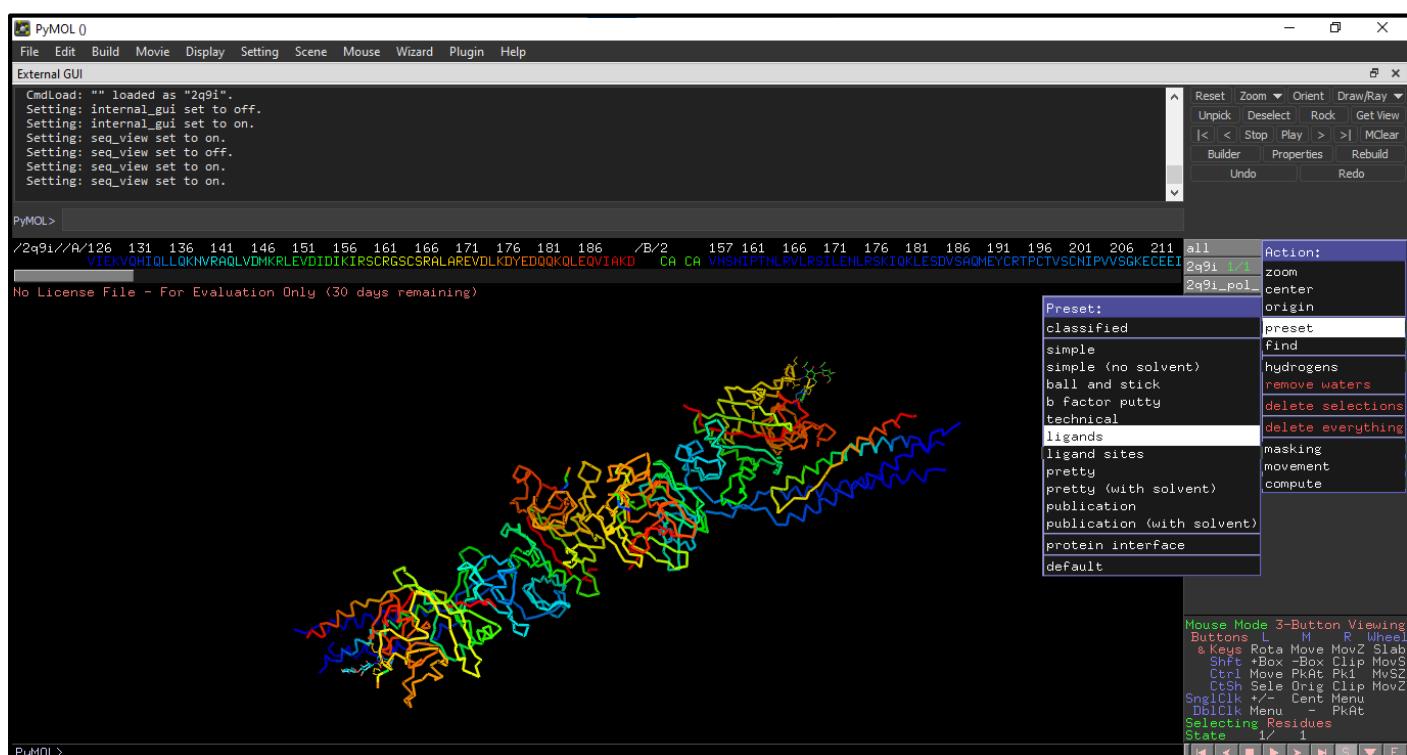


Fig7. Visualisation of ligands

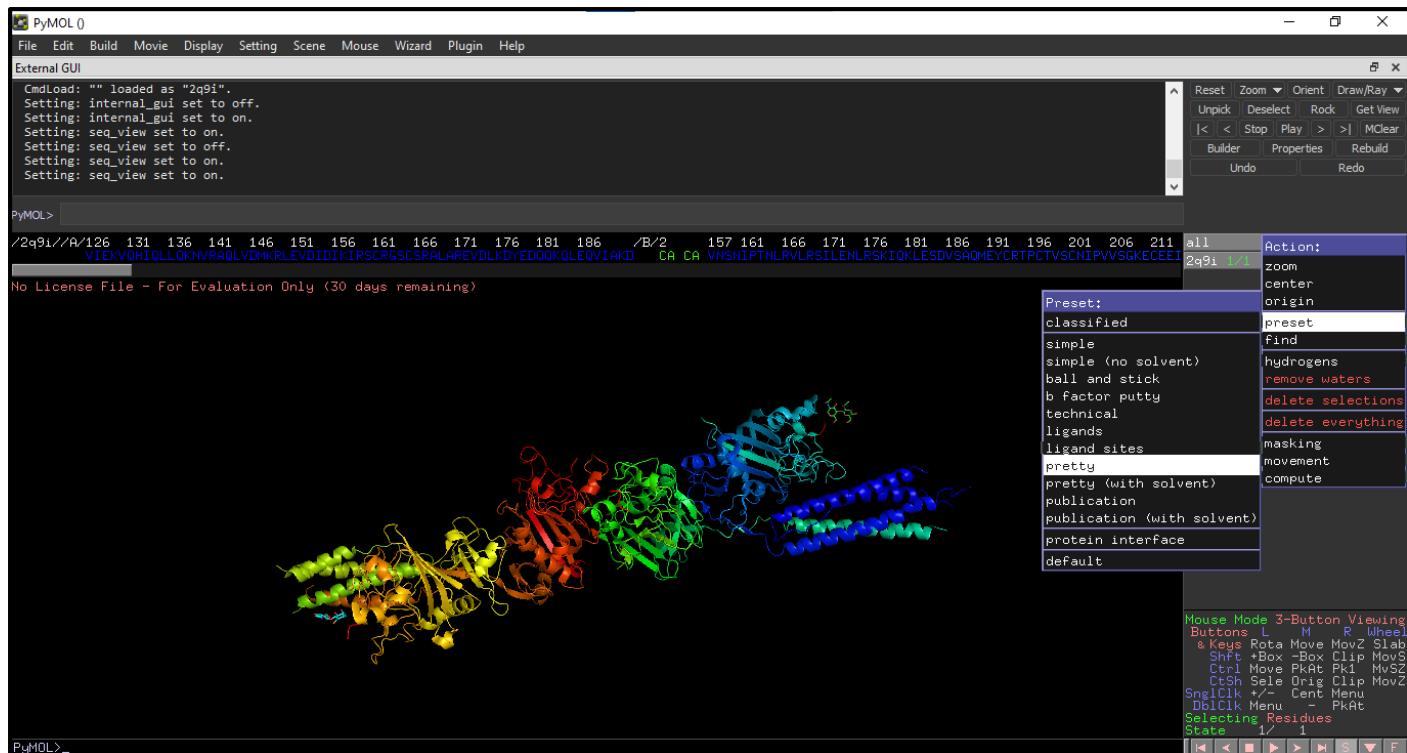


Fig8. Visualisation of Fibrin structure in pretty format

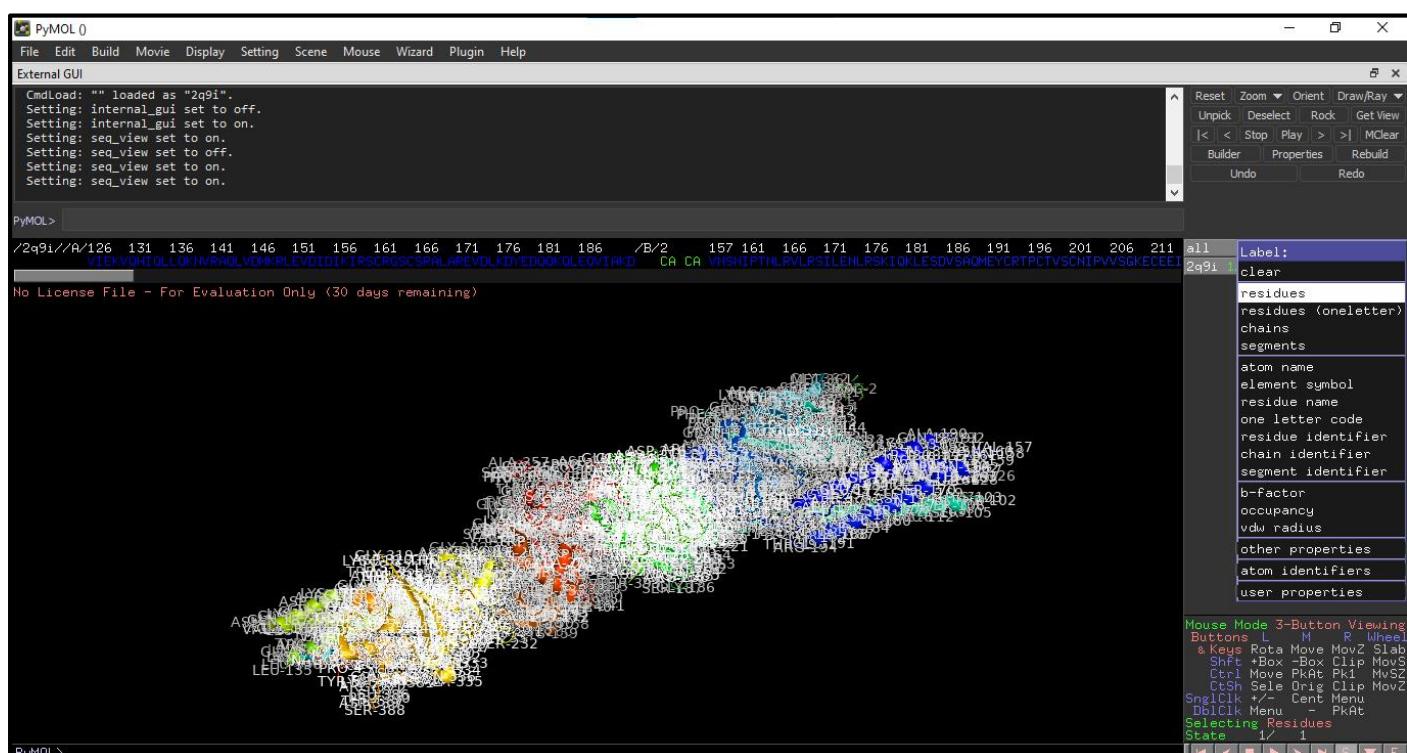
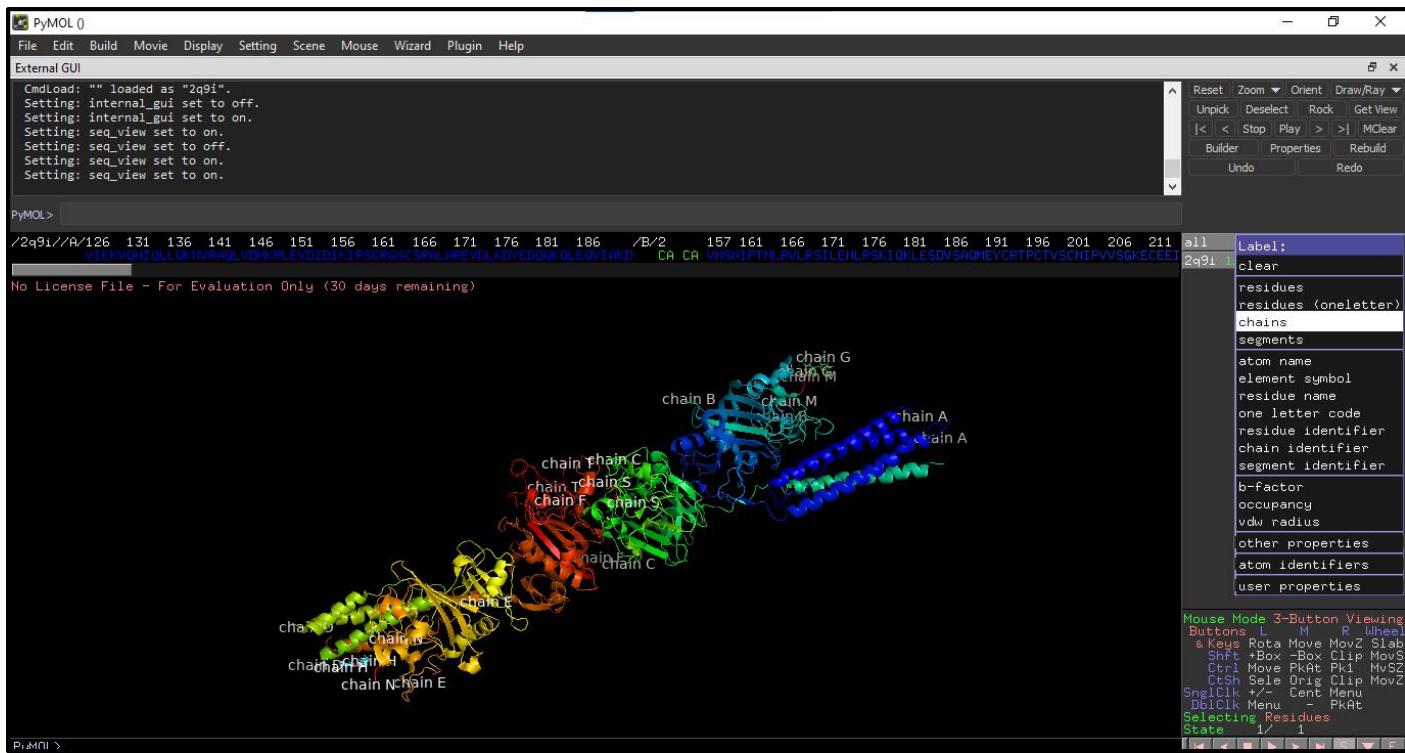
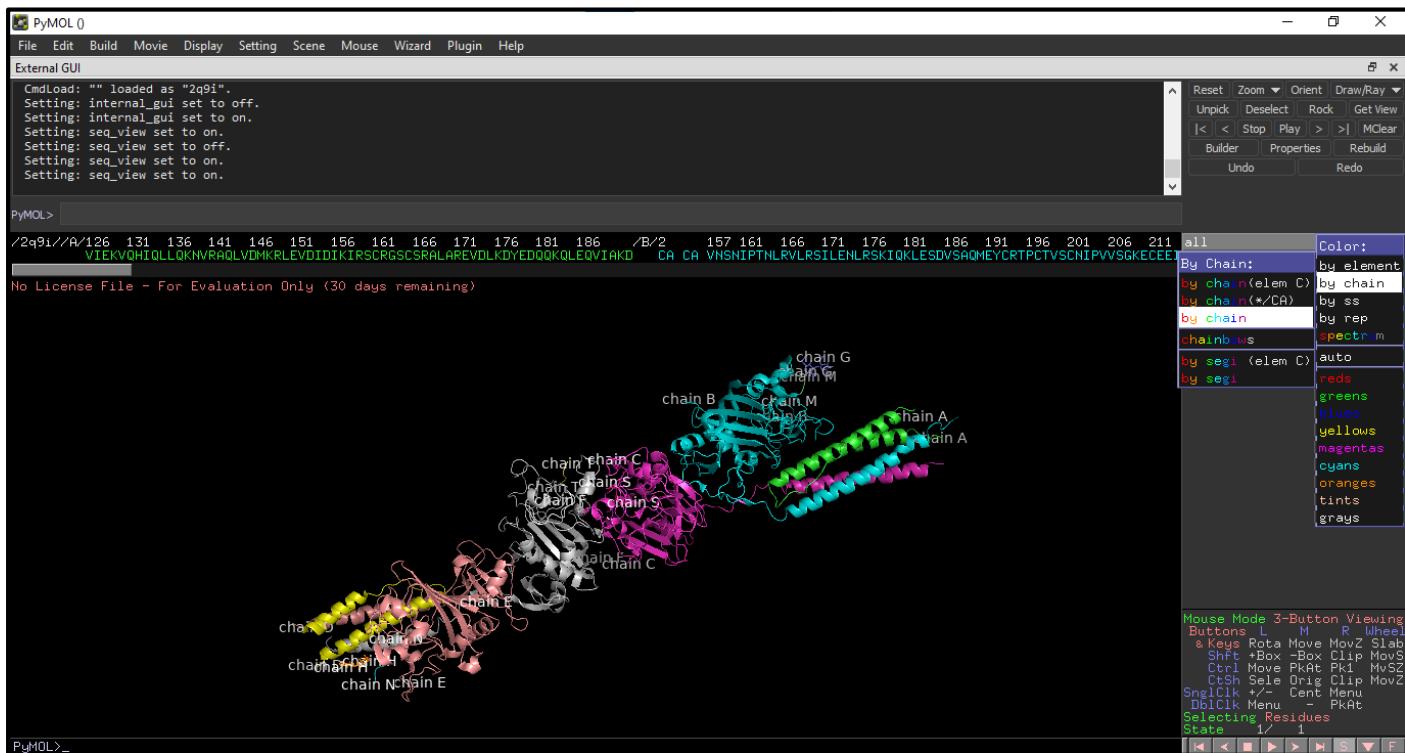


Fig9. Labelling residues of Fibrin



**Fig10. Labelling all chains of Fibrin**



**Fig11. Visualisation of all chains of Fibrin in different colours.**

## RESULT:

Fibrin 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:

5. The Editors of Encyclopedia Britannica. (2018). Fibrin | biochemistry. In *Encyclopædia Britannica*. Retrieved March 4, 2022, from <https://www.britannica.com/science/fibrin>
6. 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>
7. *RasMol and OpenRasMol*. (n.d.). [Www.openrasmol.org](http://www.openrasmol.org/). Retrieved March 4, 2022, from <http://www.openrasmol.org/>
8. *PyMOL / pymol.org*. (2019). [Pymol.org](https://pymol.org/2/). Retrieved March 4, 2022, from <https://pymol.org/2/>

## WEBLEM 6

### **Introduction to binding pocket prediction of protein with respect 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.

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. To provide additional useful information and to deliver improved user experience, we introduce here an updated server called CASTp 3.0. All important features of the previous versions of the server are retained, including detecting and characterizing cavities, pockets and channels of protein structures. In addition, we have substantially extended its functions by providing pre-computed topographic features of biological assemblies in the PDB database, as well as imprints of negative volumes of these topographic features. Furthermore, the user interface has been redesigned, making it more intuitive and informative.

#### **New features of CASTp 3.0**

Pre-computed results for biological assemblies. The atomic coordinates of a PDB ID in the Protein Data Bank describe an asymmetric unit, which is the minimum structure that can produce the unit cell of the crystal through duplication and crystal symmetry operations. For many PDB IDs, the asymmetric unit differs from the biological assembly, which is the unit that has either been shown or is thought to be the functional form of the protein. This difference can be significant, and can make the computation results of a PDB entry biologically irrelevant. For example, PDB ID: 2iww of the outer membrane porin OmpG contains four units as deposited. Its direct computation leads to the detection of a giant artificial pocket, which obscures the actual functional channel. To uncover topographic features that are biologically most relevant, the new CASTp server pre-computes topographic features for the biological assemblies of PDB IDs. Users now are able to navigate effortlessly between results of the asymmetric unit and results of biological assemblies of a PDB ID.

#### **Imprints of negative volumes of topographic features:**

In the previous versions of CASTp servers, geometric and topological features, such as pockets, cavities and channels, were shown only through the representation of surface atoms participating in their formation. The new CASTp server has added imprints of the negative volumes as the default visualization option. The negative volume is the space encompassed by the atoms that form these geometric and topological features, which can give users direct and intuitive understanding of these important structural features of proteins.

#### **Improved user interface:**

The visualization techniques employed in previous versions of CASTp servers over 12-year ago are out-of-date and are incompatible with the modern browsers. The new CASTp server now uses 3Dmol.js for structural visualization, which allows users to view, to interact with protein structures, and to examine computation results in modern web browsers such as Chrome and Firefox. Users can choose the representation style of the atoms that form each topographic feature. The imprint of these features can also be shown with user-selected colors.

An intuitive sequence panel is also presented to users with secondary structures color-coded, where residues in user-selected pockets are highlighted. Residues annotated from UniProt are also labeled. Both information on topographic feature-forming atoms and UniProt annotations are conveniently displayed when the cursor hovers over the relevant residues. The annotations are also summarized in the annotation panel. In addition, both the sequence panel and the annotation panel are linked to the structure viewer, so users can conveniently click on one residue on the sequence map to have the structure viewer zoomed into that specific residue of interest.

Furthermore, a floating structure viewer has been added. When a user inspects information in the sequence panel or the annotation panel, the structure viewer will automatically follow the web page scrolling of the user, which saves the user from unnecessary and unproductive efforts in scrolling up and down.

## **INPUT AND OUTPUT:**

### **Input**

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.

### **Output**

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 (26) or our PyMOL plugin, CASTpyMOL.

Most proteins do not perform their molecular function as unmodified folded polypeptides. In most cases, proteins need to acquire permanent or transient molecular features in order to function as they should. Post-translational modifications (PTMs) of proteins most often come in the form of proteolytic cleavage events or covalent modifications at specific amino acid residues. Proteolytic cleavage is of course an irreversible modification, while covalent modifications may be reversible.

After being synthesized (translated), the polypeptide chain is subject to many different types of post-translational processing in different cellular compartments, including the nucleus, cytosol, endoplasmic reticulum and Golgi apparatus. This happens during or after folding, and involves enzymatic processing including removal of one or more amino acids from the amino terminus, proteolytic cleavage, or addition of acetyl, phosphoryl, glycosyl, methyl or other groups to certain amino acid residues. These modifications may confer various structural and functional properties to the affected proteins.

Post-translational modifications (PTMs) occur on almost all proteins analyzed to date. The function of a modified protein is often strongly affected by these modifications and therefore increased knowledge about the potential PTMs of a target protein may increase our understanding of the molecular processes in which it takes part. High-throughput methods for the identification of PTMs are being developed, in particular within the fields of proteomics and mass spectrometry. However, these methods are still in their early stages, and it is indeed advantageous to cut down on the number of experimental steps by integrating computational approaches into the validation procedures. Many advanced methods for the prediction of PTMs exist and many are made publicly available.

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. A genetic engineering approach using human cell lines to simplify O-glycosylation (SimpleCells) that enables proteome-wide discovery of O-glycan sites using 'bottom-up' ETD-based mass spectrometric analysis was developed. This was implemented on 12 human cell lines from different organs, and present a first map of the human O-glycoproteome with almost 3000 glycosites in over 600 O-glycoproteins as well as an improved NetOGlyc4.0 model for prediction of O-glycosylation. 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.

### 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.

#### Output format:

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.

Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes. Phosphorylation sites predicted by neural networks.

### NetPhos - 3.1

Generic phosphorylation sites in eukaryotic proteins

The NetPhos 3.1 server predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks. Both generic and kinase specific predictions are performed. The generic predictions are identical to the predictions performed by NetPhos 2.0. The kinase specific predictions are identical to the predictions by NetPhosK 1.0. Predictions are made for the following 17 kinases:

ATM, CKI, CKII, CaM-II, DNAPK, EGFR, GSK3, INSR, PKA, PKB, PKC, PKG, RSK, SRC, cdc2, cdk5 and p38MAPK.

#### Output format

For each input sequence the following is shown:

- **FASTA-like header line:** a line showing the sequence name and length.
- **Prediction lines:** one line per residue and kinase, with six columns in the form:
- **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.

- **Sequence** - the input sequence as processed by NetPhos, with an overview of the positions of the predicted sites.
- **Graphics** - a plot of scores illustrating the predictions. NOTE: for each residue only the highest score is shown.

## GFF

The output in GFF ( GFF version 2) provides essentially the same information as the classical format described above. The only differences, apart from the syntax, are as follows:

- the sequence context of the residues is not provided
- the positive predictions are indicated by "Y" (not "YES")

This option has been provided for the benefit of the users who have access to software parsing GFF.

Thus, CASTP, NetOGlyc - 4.0 and NetPhos - 3.1 servers can be used to predict protein binding pockets with respect to post transcriptional modification. This information can be used by researchers 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.

## REFERENCES:

1. Tian, Wei; Chen, Chang; Lei, Xue; Zhao, Jieling; Liang, Jie (2018). *CASTP 3.0: computed atlas of surface topography of proteins.* *Nucleic Acids Research*, 46(W1), W363–W367. doi:10.1093/nar/gky473
2. Steentoft, Catharina; Vakhrushev, Sergey Y; Joshi, Hiren J; Kong, Yun; Vester-Christensen, Malene B; Schjoldager, Katrine T-B G; Lavrsen, Kirstine; Dabelsteen, Sally; Pedersen, Nis B; Marcos-Silva, Lara; Gupta, Ramneek; Paul Bennett, Eric; Mandel, Ulla; Brunak, Søren; Wandall, Hans H; Levery, Steven B; Clausen, Henrik (2013). *Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology.* *The EMBO Journal*, 32(10), 1478–1488. doi:10.1038/emboj.2013.79
3. Nikolaj Blom; Steen Gammeltoft; Søren Brunak (1999). *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.* , 294(5), 0–1362. doi:10.1006/jmbi.1999.3310
4. Nikolaj Blom; Thomas Sicheritz-Pontén; Ramneek Gupta; Steen Gammeltoft; Søren Brunak (2004). *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.* , 4(6), 1633–1649. doi:10.1002/pmic.200300771

**WEBLEM 6a****CASTp**(URL: <http://sts.bioe.uic.edu/castp/>)**AIM:**

To predict binding pocket of protein Thrombin using CASTp server.

**INTRODUCTION:**

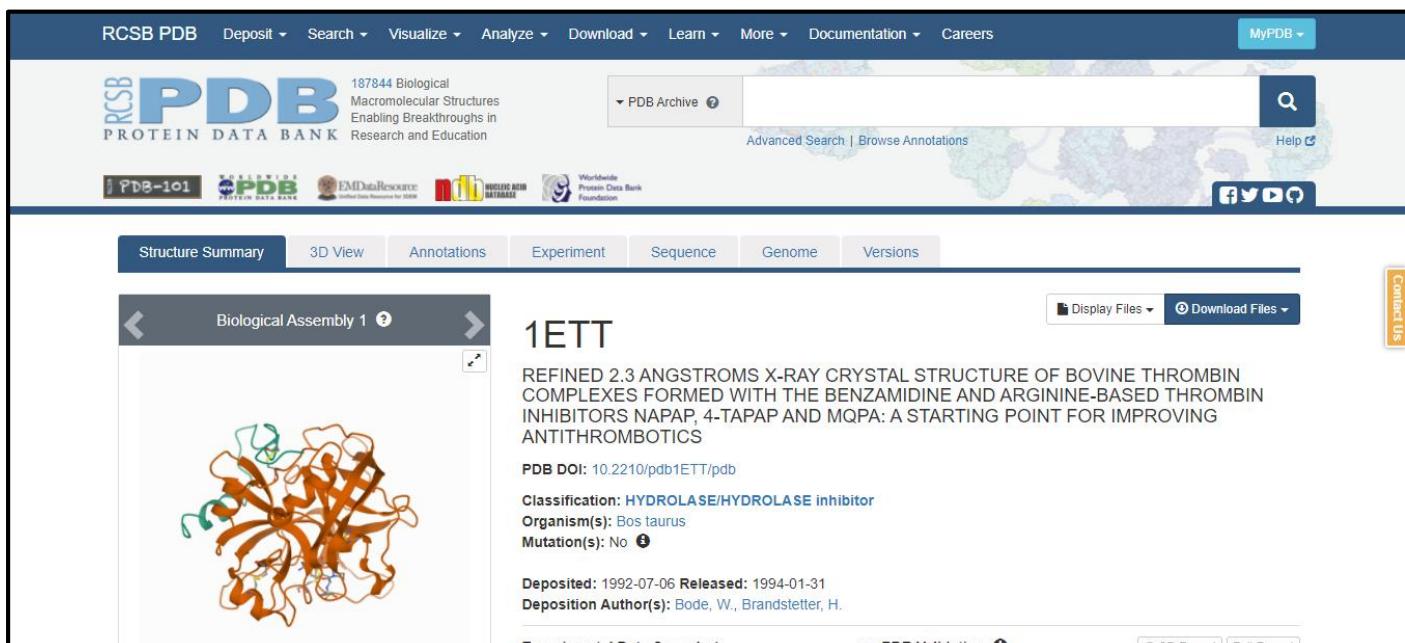
Thrombin is a multifunctional serine protease which plays a central role in haemostasis by regulating platelet aggregation and blood coagulation. It is formed from its precursor prothrombin following tissue injury and converts fibrinogen to fibrin in the final step of the clotting cascade. It also promotes numerous cellular effects including chemotaxis, proliferation, extracellular matrix turnover and release of cytokines. The binding pocket information of thrombin can be retrieved from CASTp server.

Geometric and topological properties of protein structures, including surface pockets, interior cavities and cross channels, are of fundamental importance for proteins to carry out their functions. Computed Atlas of Surface Topography of proteins (CASTp) is a web server that provides online services for locating, delineating and measuring these geometric and topological properties of protein structures. It has been widely used since its inception in 2003. CASTp 3.0 continues to provide reliable and comprehensive identifications and quantifications of protein topography. In addition, it now provides: (i) imprints of the negative volumes of pockets, cavities and channels, (ii) topographic features of biological assemblies in the Protein Data Bank, (iii) improved visualization of protein structures and pockets, and (iv) more intuitive structural and annotated information, including information of secondary structure, functional sites, variant sites and other annotations of protein residues.

**METHODOLOGY:**

1. Open homepage for CASTp server. (URL: <http://sts.bioe.uic.edu/castp/>)
2. Search for query “Thrombin” using PDB ID.
3. Configure binding pockets.
4. Observe and interpret the results.

## OBSERVATION:



187844 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB Archive Advanced Search | Browse Annotations Help

Structure Summary 3D View Annotations Experiment Sequence Genome Versions

Display Files Download Files

1ETT

REFINED 2.3 ANGSTROMS X-RAY CRYSTAL STRUCTURE OF BOVINE THROMBIN COMPLEXES FORMED WITH THE BENZAMIDINE AND ARGININE-BASED THROMBIN INHIBITORS NAPAP, 4-TAPAP AND MQPA: A STARTING POINT FOR IMPROVING ANTITHROMBOTICS

PDB DOI: 10.2210/pdb1ETT/pdb

Classification: HYDROLASE/HYDROLASE inhibitor

Organism(s): Bos taurus

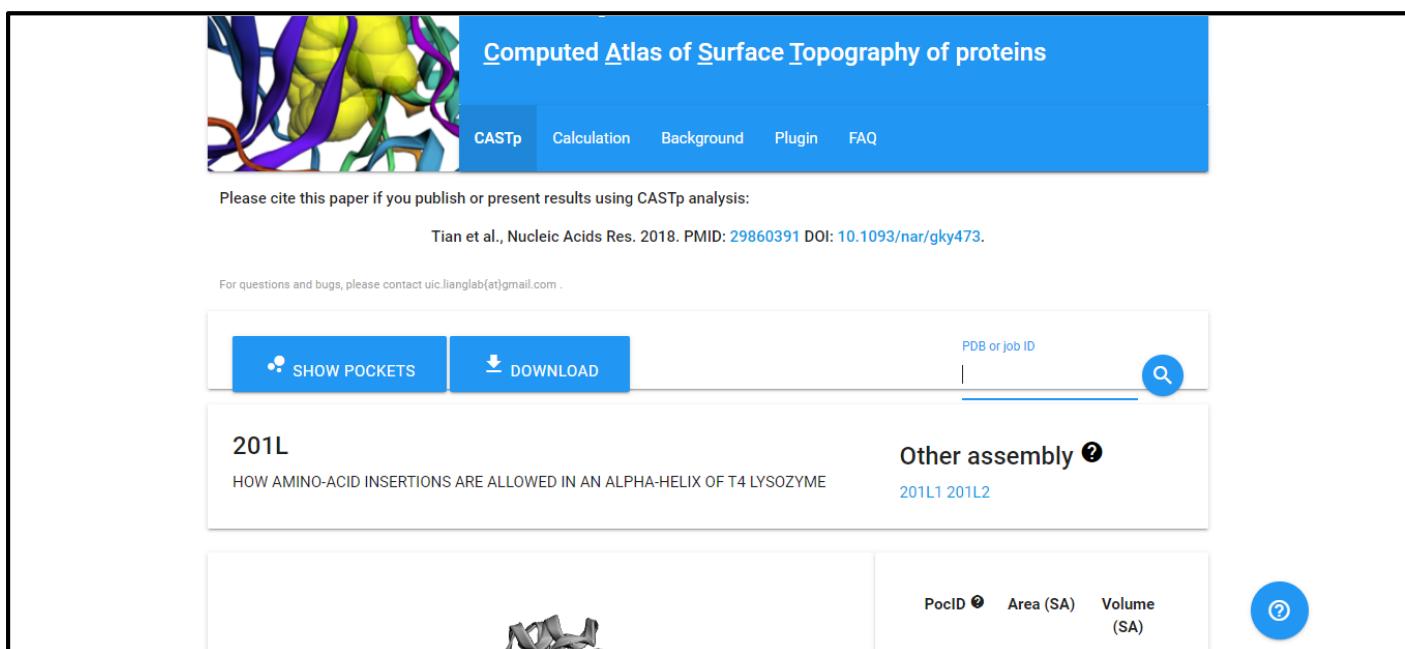
Mutation(s): No

Deposited: 1992-07-06 Released: 1994-01-31

Deposition Author(s): Bode, W., Brandstetter, H.

Experimental Data Snapshot wwPDB Validation 3D Report Full Report

Fig1. PDB structure selected for Thrombin



Computed Atlas of Surface Topography of proteins

CASTp Calculation Background Plugin FAQ

Please cite this paper if you publish or present results using CASTp analysis:

Tian et al., Nucleic Acids Res. 2018. PMID: 29860391 DOI: 10.1093/nar/gky473.

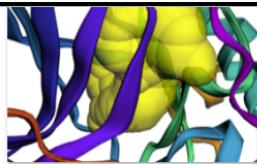
For questions and bugs, please contact uic.lianglab(at)gmail.com .

SHOW POCKETS DOWNLOAD PDB or job ID

201L HOW AMINO-ACID INSERTIONS ARE ALLOWED IN AN ALPHA-HELIX OF T4 LYSOZYME Other assembly 201L1 201L2

PocID Area (SA) Volume (SA)

Fig2. Homepage for CASTp server.



## Computed Atlas of Surface Topography of proteins

CASTp   Calculation   Background   Plugin   FAQ

Please cite this paper if you publish or present results using CASTp analysis:

Tian et al., Nucleic Acids Res. 2018. PMID: 29860391 DOI: 10.1093/nar/gky473.

For questions and bugs, please contact uic.lianglab(at)gmail.com .

 SHOW POCKETS

 DOWNLOAD

PDB or job ID

1ETT



201L

HOW AMINO-ACID INSERTIONS ARE ALLOWED IN AN ALPHA-HELIX OF T4 LYSOZYME

Other assembly 

201L1 201L2



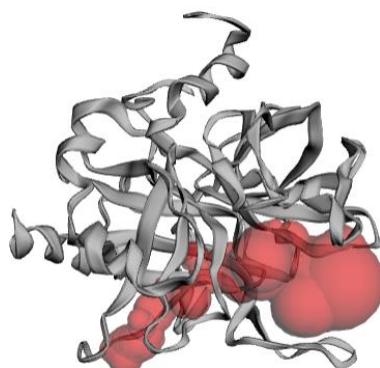
PocID  Area (SA)   Volume (SA)



**Fig3. Search for Thrombin structure (PDB ID: 1ETT)**

1ETT

REFINED 2.3 ANGSTROMS X-RAY CRYSTAL STRUCTURE OF BOVINE THROMBIN COMPLEXES FORMED WITH THE BENZAMIDINE AND ARGININE-BASED THROMBIN INHIBITORS NAPAP, 4-TAPAP AND MQPA: A STARTING POINT FOR IMPROVING ANTITHROMBOTICS



PocID  Area (SA)   Volume (SA)

1      780.524      832.908

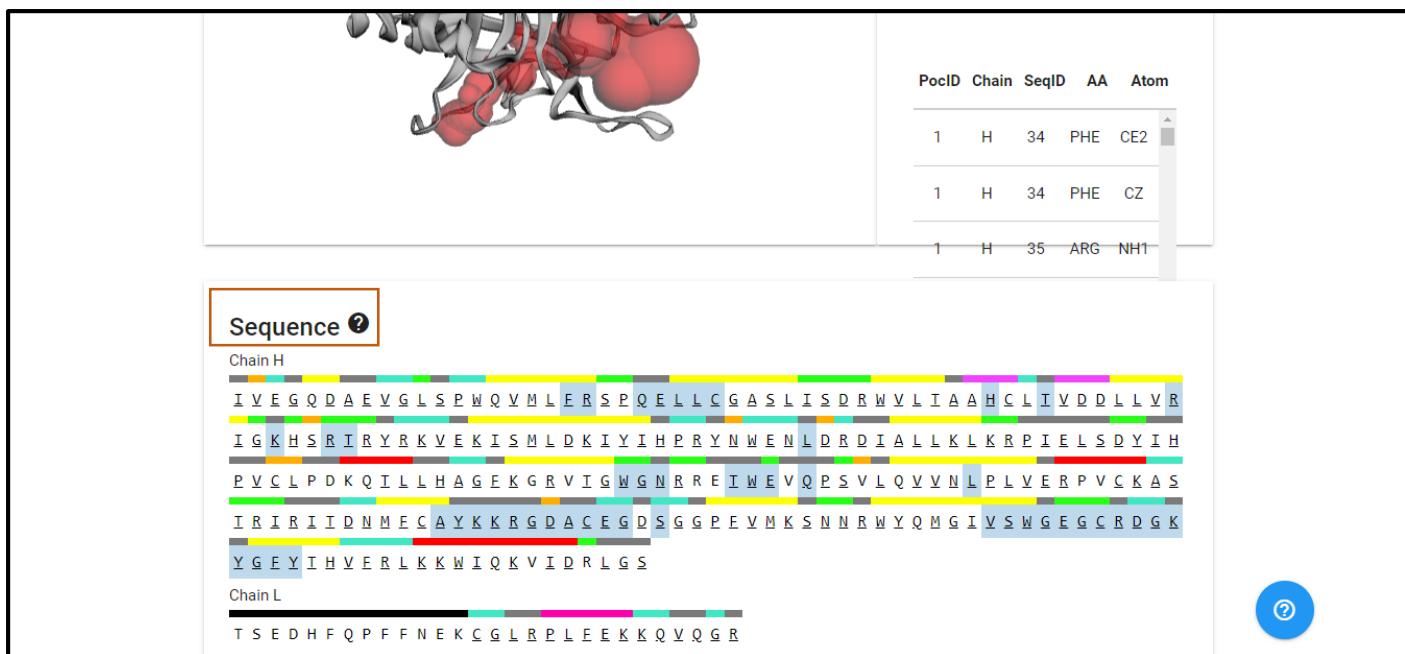
PocID   Chain   SeqID   AA   Atom

PocID	Chain	SeqID	AA	Atom
1	H	34	PHE	CE2

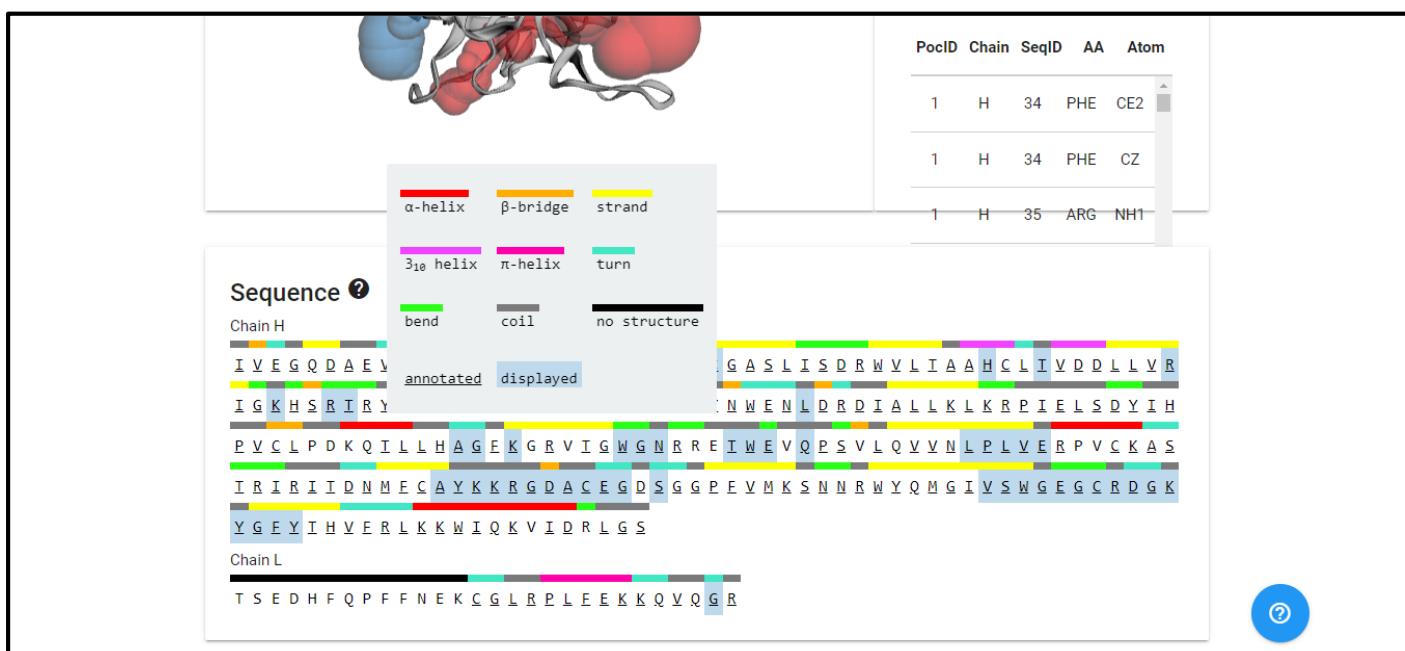
PocID	Chain	SeqID	AA	Atom
1	H	34	PHE	CZ



**Fig4. Result page for Thrombin**



**Fig4.1 Result for sequence and structure information.**



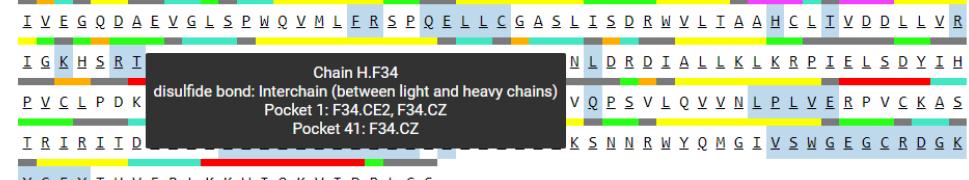
**Fig4.2 Legends for sequence information**



PocID	Chain	SeqID	AA	Atom
1	H	34	PHE	CE2
1	H	34	PHE	CZ
1	H	35	ARG	NH1

**Sequence ?**

Chain H



disulfide bond: Interchain (between light and heavy chains)

Chain H.F34

Pocket 1: F34.CE2, F34.CZ

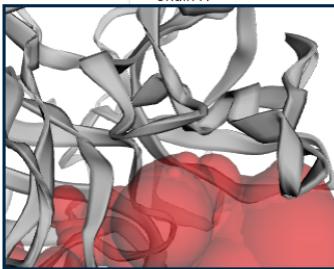
Pocket 41: F34.CZ

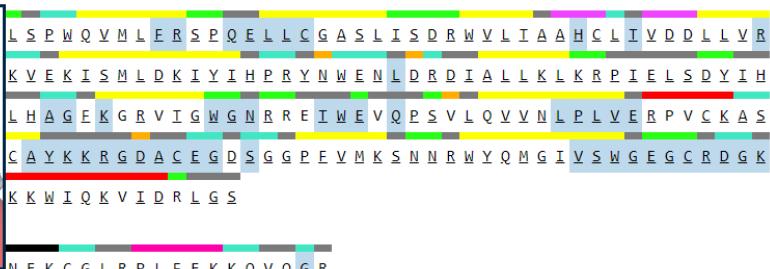
Chain L



<sts.bioe.uic.edu/castp/index.html?1ett>

**Fig4.3 Result for binding pocket information**

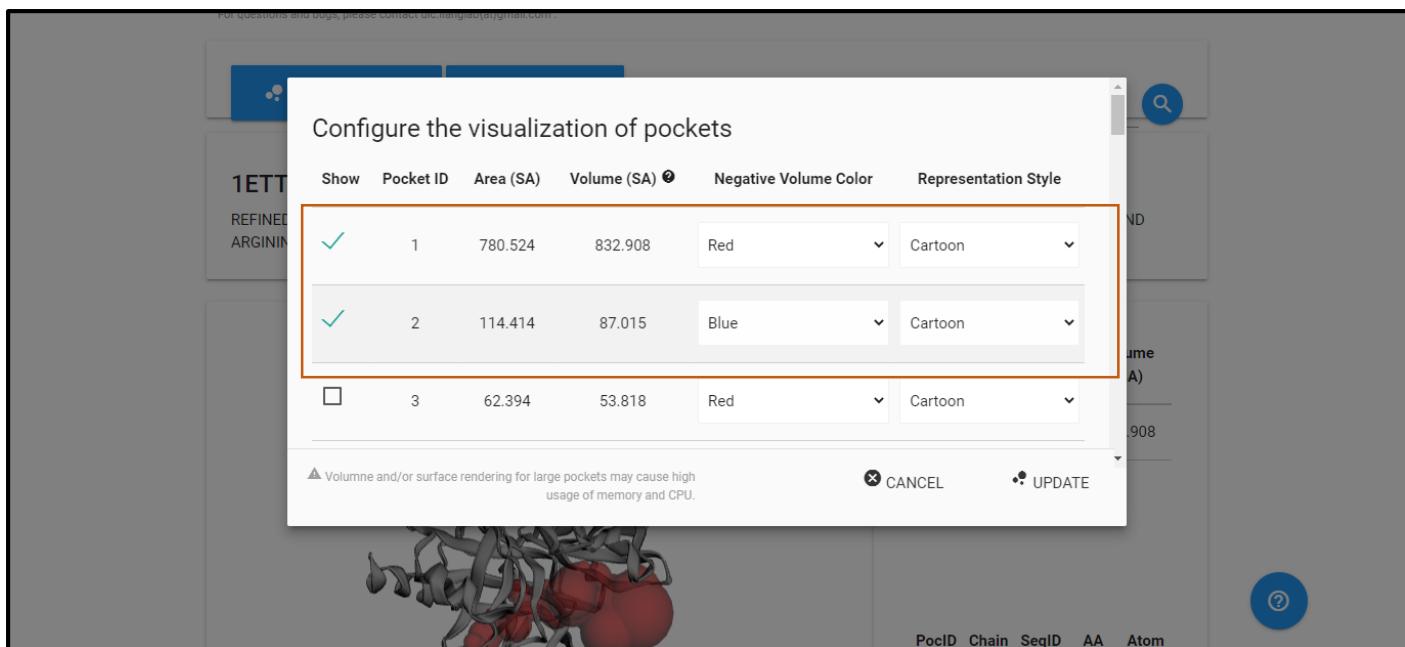




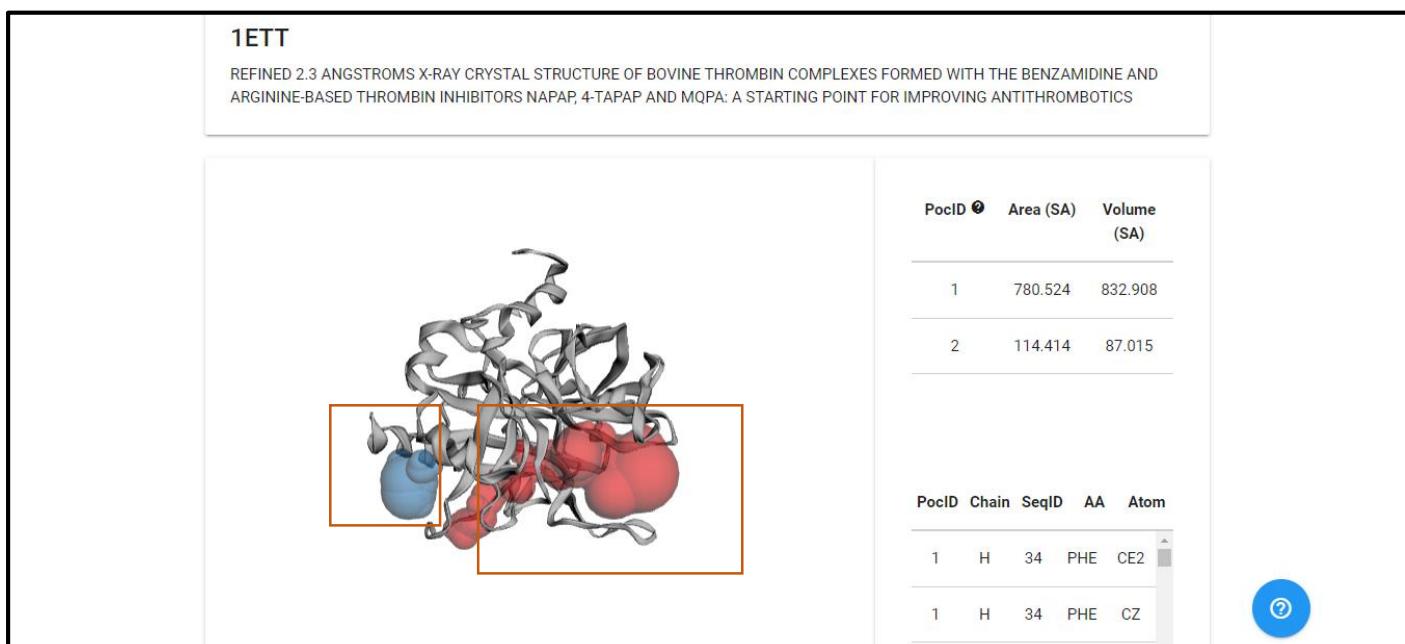
**Features**

Feature	Position(s)	Description	Reference
active site	H: 57	Charge relay system	<a href="#">Uniprot: P00735</a>
active site	H: 102	Charge relay system	<a href="#">Uniprot: P00735</a>
active site	H: 195	Charge relay system	<a href="#">Uniprot: P00735</a>
disulfide bond	H: 42-58		<a href="#">Uniprot: P00735</a>

**Fig4.4 Result for features**



**Fig5. Configuration for visualisation of binding pockets**



**Fig5.1 Result for binding pocket visualisation**

## RESULT:

On passing Thrombin PDB ID, the binding pockets were visualised, sequence and features which includes active site and disulphide bond information was retrieved.

## CONCLUSION:

CASTp server can be used to predict binding pockets of protein. This information can be used for a wide range of studies, including investigations of signalling 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.

## REFERENCES:

- Goldsack, N. R., Chambers, R. C., Dabbagh, K., & Laurent, G. J. (1998, June). *Molecules in focus Thrombin*. The International Journal of Biochemistry & Cell Biology. [https://doi.org/10.1016/s1357-2725\(98\)00011-9](https://doi.org/10.1016/s1357-2725(98)00011-9)
- Tian, Wei; Chen, Chang; Lei, Xue; Zhao, Jieling; Liang, Jie (2018). *CASTp 3.0: computed atlas of surface topography of proteins*. *Nucleic Acids Research*, 46(W1), W363–W367. doi:10.1093/nar/gky473
- *CASTp 3.0: Computed Atlas of Surface Topography of proteins*. (n.d.). Sts.bioe.uic.edu. Retrieved March 3, 2022, from <http://sts.bioe.uic.edu/castp/index.html?1ett>
- Bank, R. P. D. (n.d.). *RCSB PDB - 1ETT: REFINED 2.3 ANGSTROMS X-RAY CRYSTAL STRUCTURE OF BOVINE THROMBIN COMPLEXES FORMED WITH THE BENZAMIDINE AND ARGININE-BASED THROMBIN INHIBITORS NAPAP, 4-TAPAP AND MQPA: A STARTING POINT FOR IMPROVING ANTITHROMBOTICS*. Www.rcsb.org. Retrieved March 3, 2022, from <https://www.rcsb.org/structure/1ETT>

## WEBLEM 6b

### NetNGlyc 4.0

(URL: <https://services.healthtech.dtu.dk/service.php?NetOGlyc-4.0>)

#### AIM:

To predict binding pocket for Glycosylation sites in Thrombin using NetNGlyc 4.0 Server.

#### INTRODUCTION:

Thrombin is a multifunctional serine protease which plays a central role in haemostasis by regulating platelet aggregation and blood coagulation. It is formed from its precursor prothrombin following tissue injury and converts fibrinogen to fibrin in the final step of the clotting cascade. It also promotes numerous cellular effects including chemotaxis, proliferation, extracellular matrix turnover and release of cytokines. The binding pocket information of thrombin can be retrieved from CASTp server.

A genetic engineering approach using human cell lines to simplify O-glycosylation (SimpleCells) that enables proteome-wide discovery of O-glycan sites using ‘bottom-up’ ETD-based mass spectrometric analysis was developed. This was implemented on 12 human cell lines from different organs, and present a first map of the human O-glycoproteome with almost 3000 glycosites in over 600 O-glycoproteins as well as an improved NetOGlyc4.0 model for prediction of O-glycosylation. 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.

#### METHODOLOGY:

1. Open homepage for NetNGlyc 4.0 server. (URL: <https://services.healthtech.dtu.dk/service.php?NetOGlyc-4.0>)
2. Search for query “Thrombin” using FASTA sequence.
3. Observe and interpret the results.

#### OBSERVATION:

UniProtKB - K4LLQ2 (VSP\_BOTBA)

Protein: Thrombin-like enzyme barnettobin  
 Gene: N/A  
 Organism: Bothrops barnetti (Barnett's lancehead) (Trimeresurus barnetti)  
 Status: Reviewed - Annotation score: 5/5 - Experimental evidence at protein level

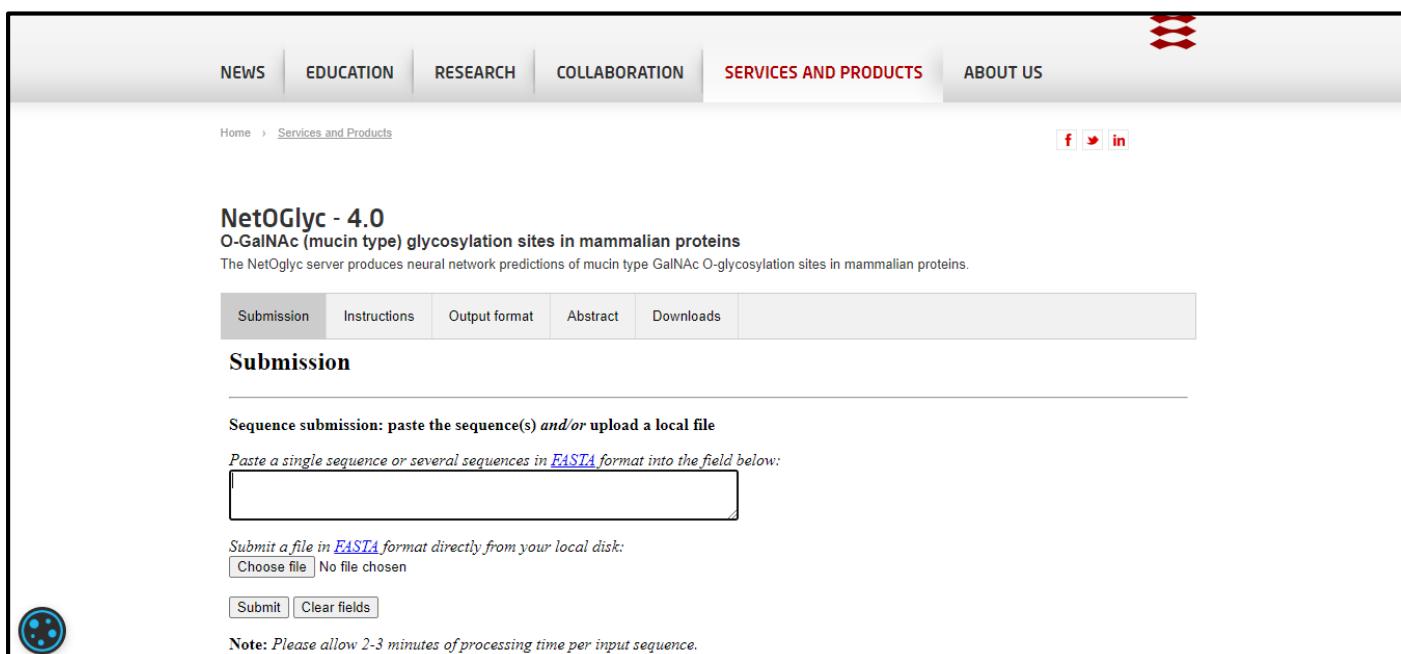
Function: Thrombin-like snake venom serine protease that releases only fibrinopeptide A from human Alpha chain of fibrinogen (specific coagulant activity was 251.7 NIH thrombin units/mg). Also shows fibrinolytic activities in vitro and defibrinogenating effects in vivo.

Miscellaneous: Does not degrade beta- and gamma-chains of fibrinogen (FGB). Does not hydrolyze the plasmin substrate S2251 (D-Val-Leu-Lys-pNA).

Fig1. Result page for Thrombin in UniProt

```
>sp|K4LLQ2|VSP_BOTBA Thrombin-like enzyme barnettobin (Fragment) OS=Bothrops barnetti OX=1051630 PE=1 SV=1
APKELOQSYAHKSSELVIGGDECINHEPFLAFYSRGNFCGLTLINQEWLTAHCDRR
FMPYIYLGIHTLSVPNDDEVRYPKDNFICPNNNIIDEKDKDINHILRLNRPVKNSEHIAPI
SLPSNLPSVGSCRVVIGIGSITAPNDTFPDVPHCANINLFLNDTVCHGAYKRFPVKSRTLC
AGVLQGGDKCMGDSGGPLICHGPFHGILFWGDDPCALPRKPALYTKGFEYPPWIQSIIA
KNTTETCPP
```

**Fig2. FASTA sequence for Thrombin**



NEWS EDUCATION RESEARCH COLLABORATION **SERVICES AND PRODUCTS** ABOUT US

Home > Services and Products [f](#) [t](#) [in](#)

**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

**Submission**

Sequence submission: paste the sequence(s) and/or upload a local file

Paste a single sequence or several sequences in [FASTA](#) format into the field below:

Submit a file in [FASTA](#) format directly from your local disk:  
 Choose file | No file chosen

**Note:** Please allow 2-3 minutes of processing time per input sequence.

**Fig3. Homepage for NetOGlyc-4.0 server.**

**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](#)

**Submission**

**Sequence submission: paste the sequence(s) and/or upload a local file**

*Paste a single sequence or several sequences in [FASTA](#) format into the field below:*  
 >sp|K4LLQ2|VSP\_BOTBA Thrombin-like enzyme barnettobin (Fragment)  
 OS=Bothrops barnetti OX=1051630 PE=1 SV=1  
 APKELQVSYAHKSSFLVIGGDECINEHPLAFLYSQRNFCLGLTLINQEHVVLTAACDRR

*Submit a file in [FASTA](#) format directly from your local disk:*  
 Choose file No file chosen

Submit  Clear fields

**Note:** Please allow 2-3 minutes of processing time per input sequence.

**Restrictions:** At most 50 sequences and 200,000 amino acids per submission; each sequence not more than 4,000 amino acids.

**Confidentiality:** The sequences are kept confidential and will be deleted after processing.

**CITATIONS**

**Fig4. Search for Thrombin using FASTA sequence.**

**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_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 8 8 0.643952 . . #POSITIVE
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 13 13 0.410675 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 14 14 0.336267 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 36 36 0.00519118 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 44 44 0.00313484 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 53 53 0.00393464 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 70 70 0.0202752 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 72 72 0.0163823 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 114 114 0.196203 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 121 121 0.107666 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 124 124 0.200306 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 128 128 0.102044 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 131 131 0.0185613 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 140 140 0.16095 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 142 142 0.0388771 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 147 147 0.0274213 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 163 163 0.00998127 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 176 176 0.0706452 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 178 178 0.14117 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 195 195 0.0188447 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 226 226 0.113172 . . .
SP_K4LLQ2_VSP_BOTBA netOGlyc-4.0.0.13 CARBOHYD 237 237 0.218237 . . .

```

**Fig5. Result page for Thrombin showing predicted glycosylation sites.**

**RESULT:**

After submitting Thrombin FASTA sequence, one glycosylation site were predicted.

Only the sites with scores higher than 0.5 are predicted as glycosylated and marked with the string "#POSITIVE" in the comment field.

**CONCLUSION:**

NetNGlyc 4.0 Server can be used to predict the glycosylation sites of proteins. This information can be used by researchers 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.

## REFERENCES:

1. Goldsack, N. R., Chambers, R. C., Dabbagh, K., & Laurent, G. J. (1998, June). *Molecules in focus Thrombin*. The International Journal of Biochemistry & Cell Biology. [https://doi.org/10.1016/s1357-2725\(98\)00011-9](https://doi.org/10.1016/s1357-2725(98)00011-9)
2. Steentoft, Catharina; Vakhrushev, Sergey Y; Joshi, Hiren J; Kong, Yun; Vester-Christensen, Malene B; Schjoldager, Katrine T-B G; Lavrsen, Kirstine; Dabelsteen, Sally; Pedersen, Nis B; Marcos-Silva, Lara; Gupta, Ramneek; Paul Bennett, Eric; Mandel, Ulla; Brunak, Søren; Wandall, Hans H; Levery, Steven B; Clausen, Henrik (2013). *Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology*. *The EMBO Journal*, 32(10), 1478–1488. doi:10.1038/embj.2013.79
3. *Thrombin-like enzyme barnettobin precursor - Bothrops barnetti (Barnett's lancehead)*. (n.d.). [Www.uniprot.org](https://www.uniprot.org/uniprot/K4LLQ2). Retrieved March 3, 2022, from <https://www.uniprot.org/uniprot/K4LLQ2>
4. *Services*. (n.d.). [Https://Www.healthtech.dtu.dk.](https://www.healthtech.dtu.dk/) Retrieved March 3, 2022, from <https://services.healthtech.dtu.dk/service.php?NetOGlyc-4.0>

**WEBLEM 6c****NetPhos 3.1**

(URL: <https://services.healthtech.dtu.dk/service.php?NetPhos-3.1>)

**AIM:**

To predict binding pocket for Glycosylation sites in Thrombin using NetPhos 3.1 Server.

**INTRODUCTION:**

Thrombin is a multifunctional serine protease which plays a central role in haemostasis by regulating platelet aggregation and blood coagulation. It is formed from its precursor prothrombin following tissue injury and converts fibrinogen to fibrin in the final step of the clotting cascade. It also promotes numerous cellular effects including chemotaxis, proliferation, extracellular matrix turnover and release of cytokines. The binding pocket information of thrombin can be retrieved from NetPhos - 3.1 server.

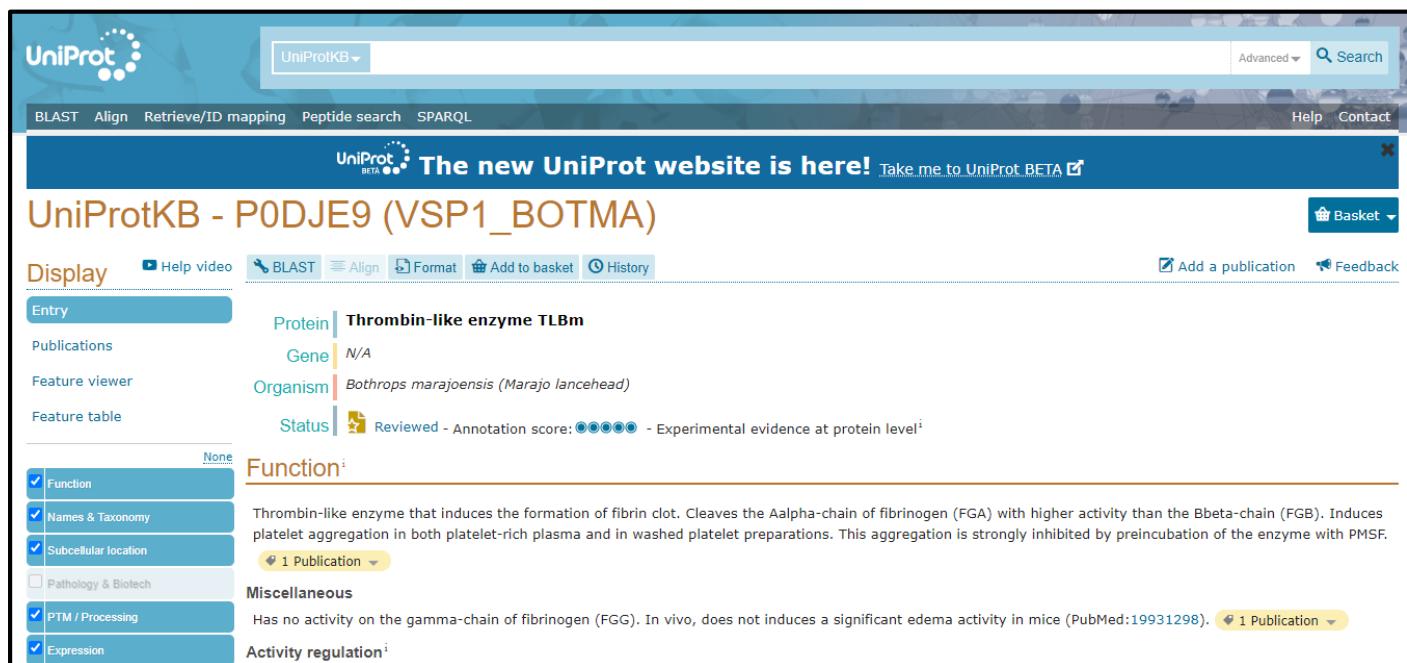
Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes. Phosphorylation sites predicted by neural networks. NetPhos - 3.1 server (Generic phosphorylation sites in eukaryotic proteins) is used. The NetPhos 3.1 server predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks. Both generic and kinase specific predictions are performed. The generic predictions are identical to the predictions performed by NetPhos 2.0. The kinase specific predictions are identical to the predictions by NetPhosK 1.0. Predictions are made for the following 17 kinases:

ATM, CKI, CKII, CaM-II, DNAPK, EGFR, GSK3, INSR, PKA, PKB, PKC, PKG, RSK, SRC, cdc2, cdk5 and p38MAPK

**METHODOLOGY:**

1. Open homepage for NetPhos 3.1 server.  
(URL:<https://services.healthtech.dtu.dk/service.php?NetPhos-3.1>)
2. Search for query “Thrombin” using FASTA sequence.
3. Observe and interpret the results.

## OBSERVATION:



UniProtKB - P0DJE9 (VSP1\_BOTMA)

Display [Help video](#) [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

[Basket](#)

**Entry**

**Protein** Thrombin-like enzyme TLBm

**Gene** N/A

**Organism** *Bothrops marajoensis* (Marajo lancehead)

**Status** Reviewed - Annotation score: Experimental evidence at protein level<sup>i</sup>

**Function**<sup>i</sup>

Thrombin-like enzyme that induces the formation of fibrin clot. Cleaves the Aalpha-chain of fibrinogen (FGA) with higher activity than the Bbeta-chain (FGB). Induces platelet aggregation in both platelet-rich plasma and in washed platelet preparations. This aggregation is strongly inhibited by preincubation of the enzyme with PMSF.

1 Publication

**Miscellaneous**

Has no activity on the gamma-chain of fibrinogen (FGG). In vivo, does not induces a significant edema activity in mice (PubMed:19931298). 1 Publication

**PTM / Processing**

**Expression**

Fig1. Result page for Thrombin in UniProt

```
>sp|P0DJE9|VSP1_BOTMA Thrombin-like enzyme TLBm OS=Bothrops marajoensis OX=157554 PE=1 SV=1
VIGGDECNINESPFLAFLYSQLLSSRRYFCGHTLINOEWVLTAHCNLYPDRKDMMWILL
IKLGKHSGSTRRIWANYDEQVRVNPKEKFIMWCPNKKDVINNVVWVWWDKDTLLWELW
MILRLNRPVKYSEHIAPLSLPSSPPSAKWHVGSVCRIMHNGQITETWINSEDTLPDVPR
CANINLFNYEVCRAYNQRMWIRGLPAKTLCAAGDEIIRGGHDTCVGDGGPLICD6QYQG
IAYWGSKPCAEPDDEPAAYSKVFDLDSQSV1AGGTWWRGDDTCP
```

Fig2. FASTA sequence for Thrombin

DTU.dk > Departments and Centers | > Shortcuts | Contact | Dansk Search for text or person 

DTU Health Tech

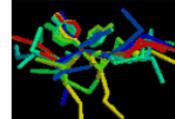
NEWS EDUCATION RESEARCH COLLABORATION SERVICES AND PRODUCTS ABOUT US

Home > Services and Products 

**NetPhos - 3.1**  
Generic phosphorylation sites in eukaryotic proteins

The NetPhos 3.1 server predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks. Both generic and kinase specific predictions are performed. The generic predictions are identical to the predictions performed by NetPhos 2.0. The kinase specific predictions are identical to the predictions by NetPhosK 1.0. Predictions are made for the following 17 kinases:

ATM, CKI, CKII, CaM-II, DNAPK, EGFR, GSK3, INSR, PKA, PKB, PKC, PKG, RSK, SRC, cdc2, cdk5 and p38MAPK.



Submission Instructions Output format PhosphoBase Downloads

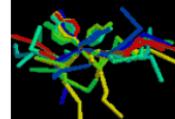
**Submission**

**Fig3. Homepage for NetPhos-3.1 server**

NetPhos - 3.1  
Generic phosphorylation sites in eukaryotic proteins

The NetPhos 3.1 server predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks. Both generic and kinase specific predictions are performed. The generic predictions are identical to the predictions performed by NetPhos 2.0. The kinase specific predictions are identical to the predictions by NetPhosK 1.0. Predictions are made for the following 17 kinases:

ATM, CKI, CKII, CaM-II, DNAPK, EGFR, GSK3, INSR, PKA, PKB, PKC, PKG, RSK, SRC, cdc2, cdk5 and p38MAPK.



Submission Instructions Output format PhosphoBase Downloads

**Submission**

Sequence submission: paste the sequence(s) and/or upload a local file

Paste a single sequence or several sequences in **FASTA** format into the field below:  
`>sp|P0DJE9|VSP1_BOTMA Thrombin-like enzyme TLM OS=Bothrops marajoensis OX=157554 PE=1 SV=1 VIGGDECNINESPFLAFLYSQLLSSRRYFCGHLINQENVLTAACCNLYPDRKDMMWLL`

Submit a file in **FASTA** format directly from your local disk:  
 No file chosen

Residues to predict  serine  threonine  tyrosine  all three

For each residue display only the best prediction

**Fig4. Search page for Thrombin using FASTA sequence.**

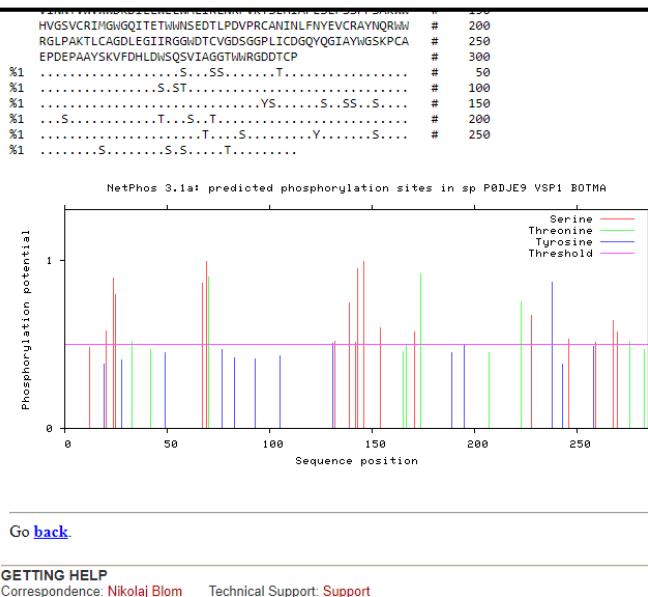
## NetPhos-3.1 Server Output - DTU Health Tech

```

>sp_P00JE9_VSP1_BOTMA 285 amino acids
#
# netphos-3.1b prediction results
#
# Sequence      # x  Context      Score  Kinase   Answer
# -----
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.479 GSK3   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.477 p38MAPK  .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.436 CKII   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.415 CaM-II  .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.406 cdc2   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.370 CKI   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.353 DNAPK  .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.350 cdk5   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.277 ATM   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.255 RSK   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.255 PKG   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.223 PKA   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.077 PKB   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.055 PKC   .
# sp_P00JE9_VSP1_BOTMA 12 S NINESPFLA 0.052 unsp  .
#
# sp_P00JE9_VSP1_BOTMA 19 Y LAFLYSQLL 0.384 INSR  .
# sp_P00JE9_VSP1_BOTMA 19 Y LAFLYSQLL 0.330 EGFR  .
# sp_P00JE9_VSP1_BOTMA 19 Y LAFLYSQLL 0.305 SRC   .
# sp_P00JE9_VSP1_BOTMA 19 Y LAFLYSQLL 0.029 unsp  .
#
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.579 ATM   YES
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.475 cdc2  .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.464 CaM-II  .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.446 GSK3  .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.395 DNAPK  .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.374 CKI   .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.359 p38MAPK  .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.331 CKII  .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.284 RSK   .
# sp_P00JE9_VSP1_BOTMA 20 S AFLYSQLLS 0.263 PKG   .

```

**Fig5.** Result page for Thrombin showing predicted phosphorylation sites.



**Fig6.** Graphical representation of predicted phosphorylation sites.

## RESULT:

After submitting Thrombin FASTA sequence, a total of 41 phosphorylation sites were predicted.

**Prediction lines:** one line per residue and kinase, with six columns in the form:

1. **Sequence** - the sequence name;
2. **#** - the position of the residue in the sequence;
3. **x** - the residue in one-letter code;
4. **Context** - the sequence context of the residue, shown as a 9-residue subsequence centered on the residue;
5. **Score** - the prediction score (a value in the range [0.000-1.000]; the scores above **0.500** indicate positive predictions);
6. **Kinase** - the active kinase or the string "unsp" for non-specific prediction (as in NetPhos 2.0);
7. **Answer** - the string "**YES**" for positive predictions, else a dot.

## CONCLUSION:

NetPhos - 3.1 server can be used to predict the phosphorylation sites of proteins. This information can be used by researchers 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.

## REFERENCES:

1. Goldsack, N. R., Chambers, R. C., Dabbagh, K., & Laurent, G. J. (1998, June). *Molecules in focus Thrombin*. The International Journal of Biochemistry & Cell Biology. [https://doi.org/10.1016/s1357-2725\(98\)00011-9](https://doi.org/10.1016/s1357-2725(98)00011-9)
2. Nikolaj Blom; Steen Gammeltoft; Søren Brunak (1999). *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites*. , 294(5), 0–1362. doi:10.1006/jmbi.1999.3310
3. Nikolaj Blom; Thomas Sicheritz-Pontén; Ramneek Gupta; Steen Gammeltoft; Søren Brunak (2004). *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence*. , 4(6), 1633–1649. doi:10.1002/pmic.200300771
4. *Thrombin-like enzyme TLBm - Bothrops marajoensis (Marajo lancehead)*. (n.d.). [Www.uniprot.org](http://www.uniprot.org). Retrieved March 3, 2022, from <https://www.uniprot.org/uniprot/P0DJF9>
5. *Services*. (n.d.). [Https://Www.healthtech.dtu.dk](https://Www.healthtech.dtu.dk). Retrieved March 3, 2022, from <https://services.healthtech.dtu.dk/service.php?NetPhos-3.1>

## WEBLEM 7

### Introduction to Structural Blast-VAST & 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 homooligomers 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 MMDDB, 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](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, 3D superimposition 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

**WEBLEM 7a****VAST**(URL: <https://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi>)**AIM:**

To perform structural Blast for Albumin using VAST tool.

**INTRODUCTION:**

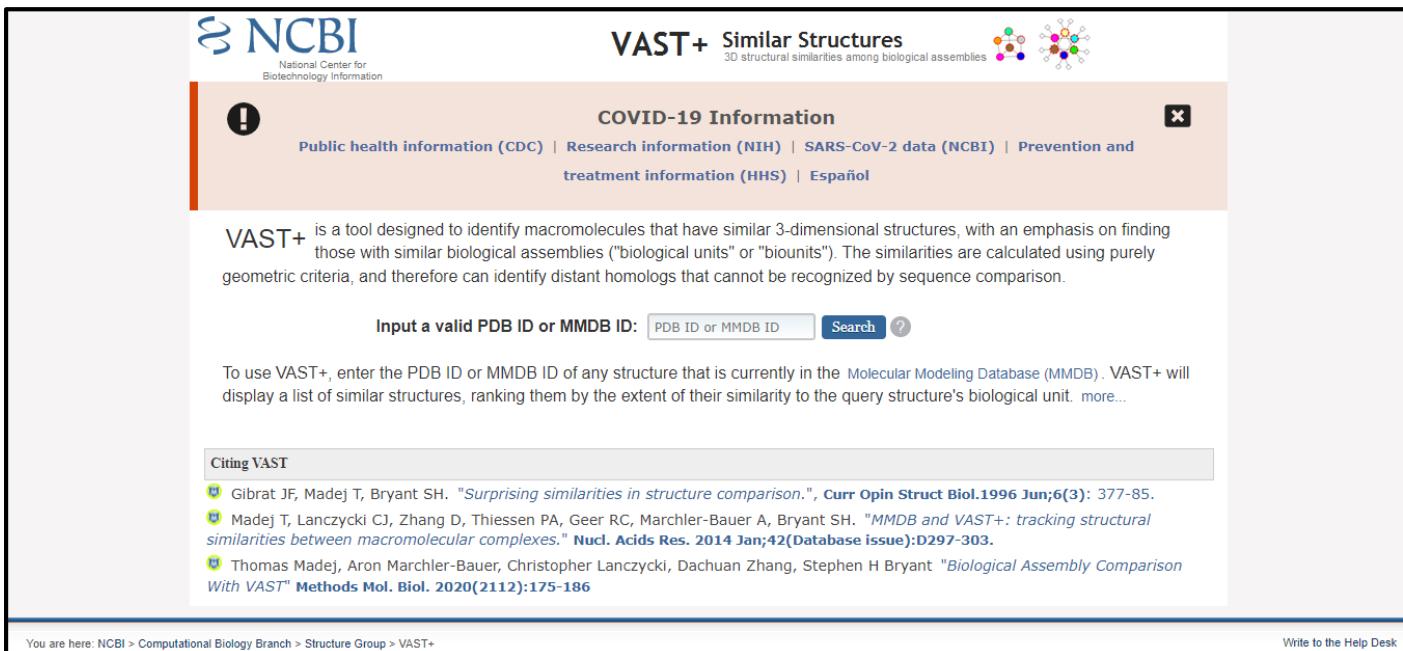
Albumin is a protein made by your liver. Albumin helps keep fluid in your bloodstream so it doesn't leak into other tissues. It is also carries various substances throughout your body, including hormones, vitamins, and enzymes. Low albumin levels can indicate a problem with your liver or kidneys. Structures similar to albumin can be retrieved using VAST tool.

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. Retrive PDB ID for Albumin.
3. Search for similar structures on VAST for the PDB ID.
4. Observe and interpret the results.

## OBSERVATION:



**VAST+ Similar Structures**  
3D structural similarities among biological assemblies

**COVID-19 Information**

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

**VAST+** is a tool designed to identify macromolecules that have similar 3-dimensional structures, with an emphasis on finding those with similar biological assemblies ("biological units" or "biounits"). The similarities are calculated using purely geometric criteria, and therefore can identify distant homologs that cannot be recognized by sequence comparison.

**Input a valid PDB ID or MMDB ID:**  [Search](#) [?](#)

To use VAST+, enter the PDB ID or MMDB ID of any structure that is currently in the Molecular Modeling Database (MMDB). VAST+ will display a list of similar structures, ranking them by the extent of their similarity to the query structure's biological unit. [more...](#)

**Citing VAST**

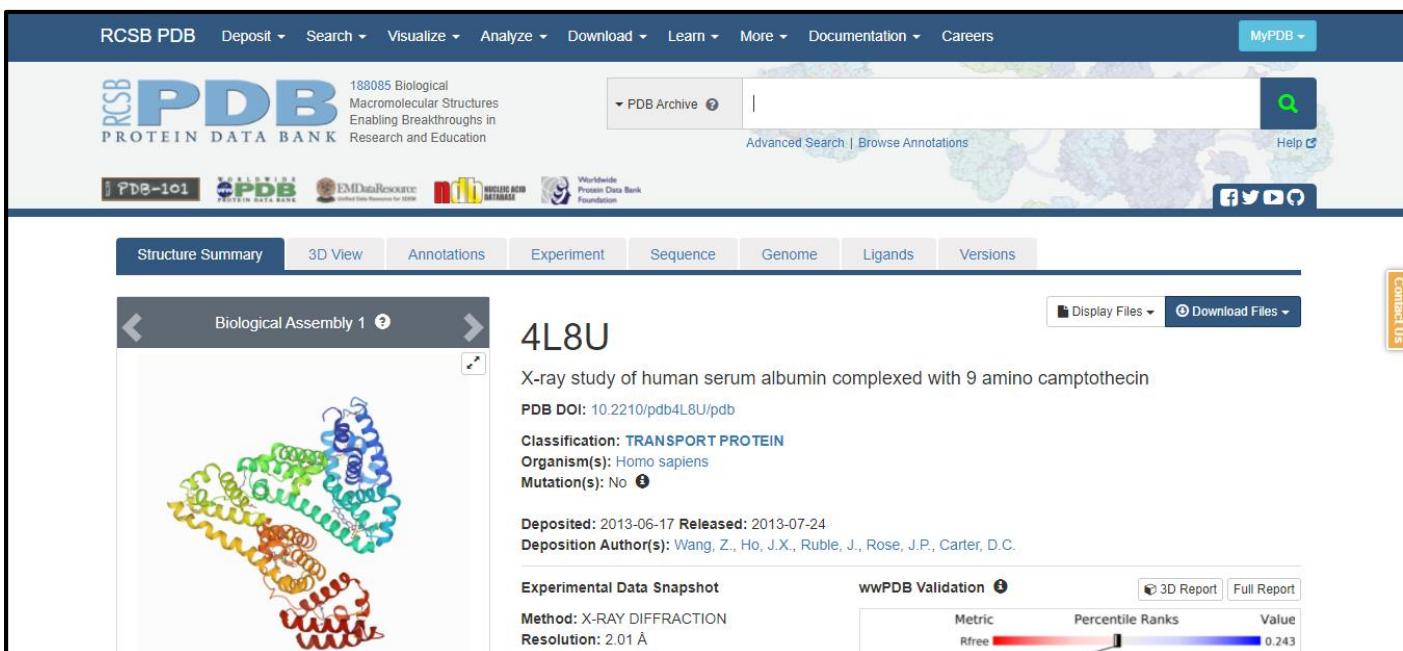
1. Gibrat JF, Madej T, Bryant SH. "Surprising similarities in structure comparison.", *Curr Opin Struct Biol.* 1996 Jun;6(3): 377-85.

2. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH. "MMDB and VAST+: tracking structural similarities between macromolecular complexes." *Nucl. Acids Res.* 2014 Jan;42(Database issue):D297-303.

3. Thomas Madej, Aron Marchler-Bauer, Christopher Lanczycki, Dachuan Zhang, Stephen H Bryant "Biological Assembly Comparison With VAST" *Methods Mol. Biol.* 2020(2112):175-186

You are here: NCBI > Computational Biology Branch > Structure Group > VAST+ Write to the Help Desk

Fig1. Homepage for VAST



**RCsb PDB** Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ More ▾ Documentation ▾ Careers [MyPDB ▾](#)

**PDB** PROTEIN DATA BANK 188085 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB-101 Worldwide PDB EMDataResource Worldwide Protein Data Bank Foundation

Advanced Search | Browse Annotations [Help](#)

Structure Summary 3D View Annotations Experiment Sequence Genome Ligands Versions

**4L8U**  
X-ray study of human serum albumin complexed with 9 amino camptothecin  
PDB DOI: 10.2210/pdb4L8U/pdb  
Classification: TRANSPORT PROTEIN  
Organism(s): Homo sapiens  
Mutation(s): No

Deposited: 2013-06-17 Released: 2013-07-24  
Deposition Author(s): Wang, Z., Ho, J.X., Ruble, J., Rose, J.P., Carter, D.C.

Experimental Data Snapshot [wwPDB Validation](#) [3D Report](#) [Full Report](#)  
Method: X-RAY DIFFRACTION Resolution: 2.01 Å  
Rfree: 0.243

Fig2. Albumin PDB structure



### COVID-19 Information



[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

VAST+ is a tool designed to identify macromolecules that have similar 3-dimensional structures, with an emphasis on finding those with similar biological assemblies ("biological units" or "biounits"). The similarities are calculated using purely geometric criteria, and therefore can identify distant homologs that cannot be recognized by sequence comparison.

Input a valid PDB ID or MMDB ID:  X Search ?

To use VAST+, enter the PDB ID or MMDB ID of any structure that is currently in the Molecular Modeling Database (MMDB). VAST+ will display a list of similar structures, ranking them by the extent of their similarity to the query structure's biological unit. [more...](#)

#### Citing VAST

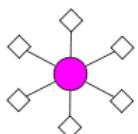
- [Gibrat JF, Madej T, Bryant SH. "Surprising similarities in structure comparison.", \*Curr Opin Struct Biol.\* 1996 Jun;6\(3\): 377-85.](#)
- [Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH. "MMDB and VAST+: tracking structural similarities between macromolecular complexes." \*Nucl. Acids Res.\* 2014 Jan;42\(Database issue\):D297-303.](#)
- [Thomas Madej, Aron Marchler-Bauer, Christopher Lanczycki, Dachuan Zhang, Stephen H Bryant "Biological Assembly Comparison With VAST" \*Methods Mol. Biol.\* 2020\(2112\):175-186](#)

You are here: NCBI > Computational Biology Branch > Structure Group > VAST+

[Write to the Help Desk](#)

### Fig3. Search for Albumin PDB structure

**4L8U : X-ray study of human serum albumin complexed with 9 amino camptothecin**




Biological unit 1: monomeric  
Source organism: **Homo sapiens**  
Number of proteins: 1 (Serum albumin)  
Number of chemicals: 6 ((2S)-2-[1-amino-8-(hydroxymethyl)-9-oxo-9,11-di... ▾)

**Similar Structures (152) ?** Original VAST ? Download VAST+ ?

**All matching molecules superposed** ?
**Invariant substructure superposed** ?

▲ Hide filters ?

**Filter by number of matching molecules ?**

Complete match, 1 proteins (152) ?

**Filter by taxonomy ?**

Eukaryota (149) ?

Others (3) ?

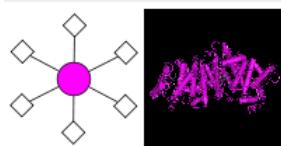
**Apply Filter Selection**

Showing 1 to 10 out of 152 selected structures ?
Search within results:  PDB ID or search word Go Reset ?

PDB ID	Description	Taxonomy <span style="color: #0070C0;">?</span>	Aligned Protein <span style="color: #0070C0;">?</span>	RMSD <span style="color: #0070C0;">?</span>	Aligned Residues <span style="color: #0070C0;">?</span>	Sequence Identity <span style="color: #0070C0;">?</span>
1 <span style="color: #0070C0;">+</span> <span style="color: #0070C0;">● 1N5U</span>	X-Ray Study Of Human Serum Albumin Complexed With Heme	Homo sapiens	1	0.36 Å	583	100%
2 <span style="color: #0070C0;">+</span> <span style="color: #0070C0;">● 4LB9</span>	X-ray Study Of Human Serum Albumin Complexed With Etoposide	Homo sapiens	1	1.22 Å	583	100%

### Fig4. Hit page for Albumin

### 4L8U: X-ray study of human serum albumin complexed with 9 amino camptothecin



Biological unit 1: monomeric  
 Source organism: Homo sapiens  
 Number of proteins: 1 (Serum albumin)  
 Number of chemicals: 6 ((2S)-2-[1-amino-8-(hydroxymethyl)-9-oxo-9,11-di... ▾)

Similar Structures (152) [?](#)

[Original VAST](#) [?](#) [Download VAST+](#) [?](#)

[All matching molecules superposed](#)

[Invariant substructure superposed](#) [?](#)

[▲ Hide filters](#) [?](#)

[Filter by number of matching molecules](#) [?](#)

[Filter by taxonomy](#) [?](#)

● Complete match, 1 proteins (152) [-](#)

Eukaryota (149) [-](#)

Others (3) [-](#)

[Apply Filter Selection](#)

Showing 1 to 10 out of 149 selected structures [?](#)

Search within results:  PDB ID or search word [Go](#) [Reset](#) [?](#)

PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1 <a href="#">+</a> ● 1N5U	X-Ray Study Of Human Serum Albumin Complexed With Heme	Homo sapiens	1	0.36 Å	583	100%
2 <a href="#">+</a> ● 4LB9	X-ray Study Of Human Serum Albumin Complexed With Etoposide	Homo sapiens	1	1.22 Å	583	100%

**Fig4.1. Hit page after applying filter**

[Apply Filter Selection](#)

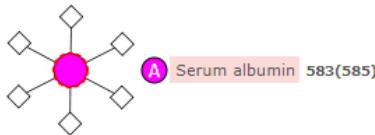
Showing 1 to 10 out of 149 selected structures [?](#)

Search within results:  PDB ID or search word [Go](#) [Reset](#) [?](#)

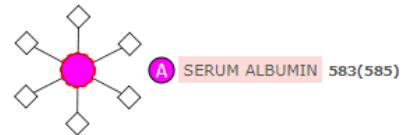
PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1 <a href="#">-</a> ● 1N5U	X-Ray Study Of Human Serum Albumin Complexed With Heme	Homo sapiens	1	0.36 Å	583	100%

Aligned Molecules [?](#)

Query structure **4L8U**



Matched structure **1N5U**



\*Select schematic circles or highlighted molecule names to view matches

[Visualize 3D structure superposition](#) [?](#) [View aligned sequences](#) [?](#)

2 <a href="#">+</a> ● 4LB9	X-ray Study Of Human Serum Albumin Complexed With Etoposide	Homo sapiens	1	1.22 Å	583	100%
3 <a href="#">+</a> ● 3SQJ	Recombinant Human Serum Albumin From Transgenic Plant	Homo sapiens	1	0.55 Å	582	100%
4 <a href="#">+</a> ● 3CX9	Crystal Structure Of Human Serum Albumin Complexed With Myristic Acid And Lysophosphatidylethanolamine	Homo sapiens	1	0.60 Å	582	100%
5 <a href="#">+</a> ● 3UIV	Human Serum Albumin–Myristate–Amantadine Hydrochloride Complex	Homo sapiens	1	0.60 Å	582	100%
6 <a href="#">+</a> ● 2XVW	Human serum albumin complexed with dansyl-L-arginine and myristic acid	Homo sapiens	1	0.68 Å	582	100%
7 <a href="#">+</a> ● 2BXK	Human serum albumin complexed with myristate, azapropazone and indomethacin	Homo sapiens	1	0.68 Å	582	100%

**Fig5. Result for aligned molecules**

Citation: [? 2](#)**The atomic structure of human methemalbumin at 1.9 Å**Wardell M, Wang Z, Ho JX, Robert J, Ruker F, Ruble J, Carter DC  
Biochem Biophys Res Commun (2002) 291 p.813-9

» All references (5)

**Abstract**

The high resolution structure of hemalalbumin was determined by single crystal X-ray diffraction to 1.9 Å. The structure revealed the protoporphyrin IX bound to a single site within a hydrophobic cavity in subdomain 1B, one of the principal binding sites for long chain fatty acid. The iron is penta coordinated with the fifth ligand comprised of the hydroxyl oxygen of Tyr-161... [read more](#)

PDB ID:

1N5U [Download](#) [?](#)

MMDB ID:

23424 [?](#)

PDB Deposition Date:

2002/11/7 [?](#)

Updated in MMDB:

2012/10 [?](#)

Experimental Method:

x-ray diffraction [?](#)

Resolution:

1.9 Å [?](#)

Source Organism:

Homo sapiens [?](#)

Similar Structures:

VAST+ [?](#)[Download sequence data](#) [?](#)**Biological Unit****Asymmetric Unit**[?](#)Biological Unit for 1N5U: monomeric; determined by author [?](#)

Molecular Graphic

[?](#)

Interactions

[?](#)**Fig5.1. Result page for structure (1N5U) similar to Albumin****3D view****full-featured 3D viewer**[Download Cn3D](#)**Molecular Components in 1N5U** [?](#)

Label	Count	Molecule
<b>Protein (1 molecule)</b>		
A	1	<p>Serum Albumin (Gene symbol: <a href="#">ALB</a>)</p>
		<p><a href="#">1 Protein</a> <a href="#">3D Domains</a> <a href="#">Domain Families</a> <a href="#">Specific Hits</a> <a href="#">Super Families</a></p>
<b>Chemicals and Non-standard biopolymers (6 molecules)</b>		
<a href="#">1</a>	5	Myristic Acid
<a href="#">2</a>	1	Protoporphyrin IX Containing Fe

\* Click molecule labels to explore molecular sequence information.

**Citing MMDB**

[Medeiros T, Lanzczuk CJ, Zheng D, Thiessen PA, Carter DC, Marchler-Bauer A, Bryant SH. " MMDB and VAST+: tracking](#)

**Fig5.2. Molecular components and chemicals and non-standard biopolymers for 1N5U**

Aligned Sequences <a href="#">?</a>		<a href="#">Close</a>
4L8U_A: Serum albumin	1N5U_A: SERUM ALBUMIN	<a href="#">Visualize 3D structure superposition</a> <a href="#">?</a>
<a href="#">4L8U_A</a>	2 AHKSEVAHFKDLGEENFKALVLI <del>A</del> FAQYLQQCPFEDHV <del>K</del> V <del>N</del> E <del>V</del> TEFAKTCVADESAEN 61	
<a href="#">1N5U_A</a>	2 AHKSEVAHFKDLGEENFKALVLI <del>A</del> FAQYLQQCPFEDHV <del>K</del> V <del>N</del> E <del>V</del> TEFAKTCVADESAEN 61	
<a href="#">4L8U_A</a>	62 CDKSLHTLFGOKLCTVATLRETYGEMADCCAKQEPERNECFLQHKODNPNLPRLVRPEVD 121	
<a href="#">1N5U_A</a>	62 CDKSLHTLFGOKLCTVATLRETYGEMADCCAKQEPERNECFLQHKODNPNLPRLVRPEVD 121	
<a href="#">4L8U_A</a>	122 VMCTAFHDNEETFLKKYLYEIAARRH <del>P</del> YF <del>A</del> P <del>E</del> LFFAKRYKA <del>A</del> F <del>E</del> CCQAA <del>D</del> KAA <del>C</del> LLPK 181	
<a href="#">1N5U_A</a>	122 VMCTAFHDNEETFLKKYLYEIAARRH <del>P</del> YF <del>A</del> P <del>E</del> LFFAKRYKA <del>A</del> F <del>E</del> CCQAA <del>D</del> KAA <del>C</del> LLPK 181	
<a href="#">4L8U_A</a>	182 LDELRDEGKASSAKQRLK <del>C</del> ASLQ <del>K</del> GERAFKA <del>A</del> VARLSQ <del>R</del> PKA <del>E</del> FA <del>V</del> SKLV <del>T</del> DLTKV 241	
<a href="#">1N5U_A</a>	182 LDELRDEGKASSAKQRLK <del>C</del> ASLQ <del>K</del> GERAFKA <del>A</del> VARLSQ <del>R</del> PKA <del>E</del> FA <del>V</del> SKLV <del>T</del> DLTKV 241	
<a href="#">4L8U_A</a>	242 HTECCHGDLLE <del>C</del> AD <del>A</del> DLAKY <del>I</del> CENQDS <del>I</del> SSKL <del>K</del> KE <del>C</del> CE <del>K</del> PL <del>E</del> K <del>H</del> SCIAEV <del>E</del> ND <del>E</del> MPAD 301	
<a href="#">1N5U_A</a>	242 HTECCHGDLLE <del>C</del> AD <del>A</del> DLAKY <del>I</del> CENQDS <del>I</del> SSKL <del>K</del> KE <del>C</del> CE <del>K</del> PL <del>E</del> K <del>H</del> SCIAEV <del>E</del> ND <del>E</del> MPAD 301	
<a href="#">4L8U_A</a>	302 LPSLAADF <del>V</del> ESK <del>V</del> DKVCKN <del>Y</del> AEAKD <del>V</del> FLG <del>M</del> FL <del>Y</del> EYARRH <del>P</del> DSV <del>V</del> LLRLAKTYET <del>T</del> LEKCC 361	
<a href="#">1N5U_A</a>	302 LPSLAADF <del>V</del> ESK <del>V</del> DKVCKN <del>Y</del> AEAKD <del>V</del> FLG <del>M</del> FL <del>Y</del> EYARRH <del>P</del> DSV <del>V</del> LLRLAKTYET <del>T</del> LEKCC 361	
<a href="#">4L8U_A</a>	362 AAADPH <del>E</del> CYAKV <del>F</del> DEF <del>K</del> PLVEEP <del>Q</del> N <del>L</del> IKQ <del>N</del> EL <del>F</del> EQ <del>L</del> GEY <del>K</del> FQ <del>N</del> ALLV <del>R</del> Y <del>T</del> KKV <del>P</del> Q <del>V</del> STP 421	
<a href="#">1N5U_A</a>	362 AAADPH <del>E</del> CYAKV <del>F</del> DEF <del>K</del> PLVEEP <del>Q</del> N <del>L</del> IKQ <del>N</del> EL <del>F</del> EQ <del>L</del> GEY <del>K</del> FQ <del>N</del> ALLV <del>R</del> Y <del>T</del> KKV <del>P</del> Q <del>V</del> STP 421	
<a href="#">4L8U_A</a>	422 TLVEVSRNLGKVGS <del>K</del> CKHPEAKR <del>N</del> PA <del>E</del> DYLSV <del>L</del> NQ <del>L</del> CVL <del>H</del> E <del>K</del> TPV <del>S</del> DR <del>V</del> TKC <del>T</del> ESL 481	
<a href="#">1N5U_A</a>	422 TLVEVSRNLGKVGS <del>K</del> CKHPEAKR <del>N</del> PA <del>E</del> DYLSV <del>L</del> NQ <del>L</del> CVL <del>H</del> E <del>K</del> TPV <del>S</del> DR <del>V</del> TKC <del>T</del> ESL 481	
<a href="#">4L8U_A</a>	482 VNR <del>PC</del> FSALE <del>V</del> ET <del>Y</del> P <del>K</del> EFNAET <del>T</del> TFHADIC <del>T</del> LSE <del>K</del> ERQ <del>I</del> KKQTAL <del>V</del> EL <del>V</del> KH <del>K</del> PKATK 541	
<a href="#">1N5U_A</a>	482 VNR <del>PC</del> FSALE <del>V</del> ET <del>Y</del> P <del>K</del> EFNAET <del>T</del> TFHADIC <del>T</del> LSE <del>K</del> ERQ <del>I</del> KKQTAL <del>V</del> EL <del>V</del> KH <del>K</del> PKATK 541	
<a href="#">4L8U_A</a>	542 EQLKAVMDDFAAF <del>V</del> E <del>K</del> CKKA <del>D</del> KET <del>C</del> FA <del>E</del> E <del>G</del> KKLV <del>A</del> ASQ <del>A</del> LG 584	
<a href="#">1N5U_A</a>	542 EQLKAVMDDFAAF <del>V</del> E <del>K</del> CKKA <del>D</del> KET <del>C</del> FA <del>E</del> E <del>G</del> KKLV <del>A</del> ASQ <del>A</del> LG 584	

Fig6. Result page for aligned sequences

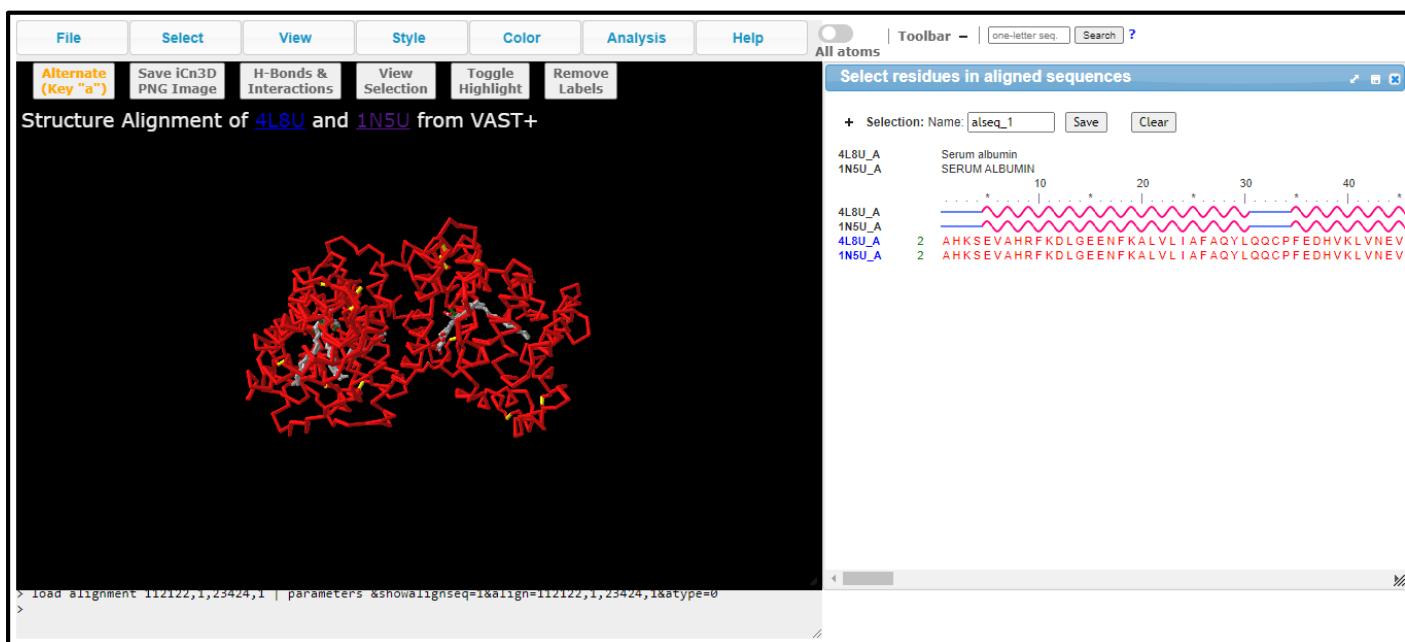


Fig7. Result page for 3D structure superimposition

Hydrogen bonds/... | one-letter seq. | Search | in two sets of atoms

1. Choose interaction types and their thresholds:

Hydrogen Bonds  Å  Salt Bridge/Ionic  Å  Contacts/Interactions  Å  
 Halogen Bonds  Å  π-Cation  Å  π-Stacking  Å

2. Select the first set:  
 selected  
 1N5U  
 1N5U\_A  
 1N5U\_Misc  
 1N5U\_cons

3. Select the second set:  
 non-selected  
 selected  
 1N5U  
 1N5U\_A  
 1N5U\_Misc

4. Cross Structure Interactions:

3D Display Interactions  
 Highlight Interactions in Table Sort Interactions on: Set 1 Set 2  
 2D Interaction Network to show interactions between two lines of residue nodes  
 2D Interaction Map to show interactions as map  
 2D Graph(Force-Directed) to show interactions with strength parameters in 0-200:  
 Helix or Sheet:  Coil or Nucleotide:  Disulfide Bond:   
 Hydrogen Bond:  Salt Bridge/Ionic:  Contacts:   
 Halogen Bonds:  π-Cation:  π-Stacking:   
 (Note: you can also adjust thresholds at #1 to add/remove interactions.)

5.  and select new sets

> buried  
> calc b...  
>

Fig7.1. Buried surface information

File Select View Style Color Analysis Help | Toolbar | one-letter seq. | Search ?

Selection

Hydrogen bonds/... | one-letter seq. | Search | in two sets of atoms

1. Choose interaction types and their thresholds:

Hydrogen Bonds  Å  Salt Bridge/Ionic  Å  Contacts/Interactions  Å  
 Halogen Bonds  Å  π-Cation  Å  π-Stacking  Å

2. Select the first set:  
 selected  
 1N5U  
 1N5U\_A  
 1N5U\_Misc  
 1N5U\_cons

3. Select the second set:  
 non-selected  
 selected  
 1N5U  
 1N5U\_A  
 1N5U\_Misc

4. Cross Structure Interactions:

3D Display Interactions  
 Highlight Interactions in Table Sort Interactions on: Set 1 Set 2  
 2D Interaction Network to show interactions between two lines of residue nodes  
 2D Interaction Map to show interactions as map  
 2D Graph(Force-Directed) to show interactions with strength parameters in 0-200:  
 Helix or Sheet:  Coil or Nucleotide:  Disulfide Bond:   
 Hydrogen Bond:  Salt Bridge/Ionic:  Contacts:   
 Halogen Bonds:  π-Cation:  π-Stacking:   
 (Note: you can also adjust thresholds at #1 to add/remove interactions.)

5.  and select new sets

> buried  
> calc b...  
>

Structure Alignment of 4L8U and 1N5U from VAST+

Show interactions

hbonds, salt bridge, interactions, halogen, π-cation, π-stacking between Two Sets:  
 Set 1: 1N5U   
 Set 2: selected

The interfaces are:  
 interface\_1   
 interface\_2

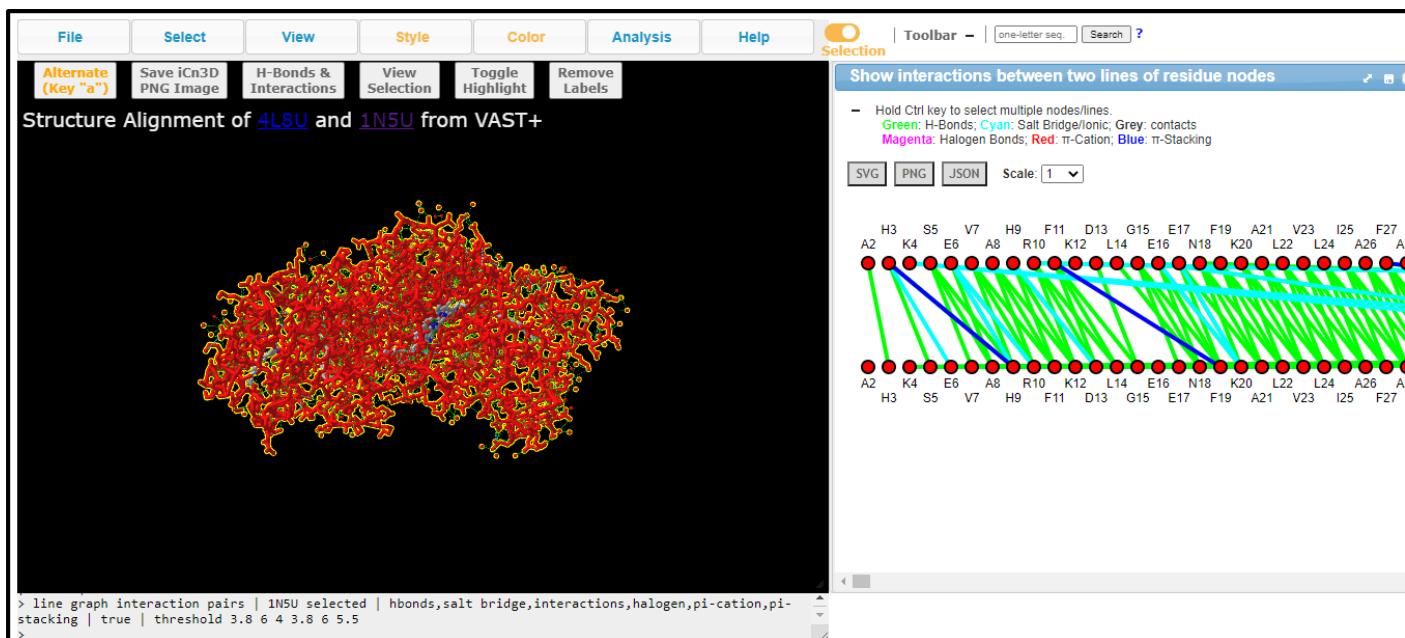
Note: Each checkbox below selects the corresponding residue. You can click "Save Selection" in the "Select" menu to the selection and click on "Highlight" button to clear the checkboxes.

2009 hydrogen bond pairs:

Atom 1	Atom 2	Distance(Å)	Highlight
<input type="checkbox"/> ALA \$1N5U.A:2@O	<input type="checkbox"/> HIS \$1N5U.A:3@ND1	3.5	<input type="button" value="Highlight"/>
<input type="checkbox"/> HIS \$1N5U.A:3@ND1	<input type="checkbox"/> ALA \$1N5U.A:2@O	3.5	<input type="button" value="Highlight"/>
<input type="checkbox"/> HIS \$1N5U.A:3@ND1	<input type="checkbox"/> HIS \$1N5U.A:9@ND1	3.6	<input type="button" value="Highlight"/>
<input type="checkbox"/> HIS \$1N5U.A:3@NE2	<input type="checkbox"/> HIS \$1N5U.A:9@NE2	3.3	<input type="button" value="Highlight"/>
<input type="checkbox"/> HIS \$1N5U.A:2@O	<input type="checkbox"/> GLD \$1N5U.A:6@N	3.5	<input type="button" value="Highlight"/>

> view interaction pairs | 1N5U selected | hbonds,salt bridge,interactions,halogen,pi-cation,pi-stacking  
| false | threshold 3.8 6 4 3.8 6 5.5  
>

Fig7.2. Results for interactions



**Fig7.3. Result for 2D interaction network**

## RESULT:

PDB ID for albumin 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. *Albumin Blood Test: MedlinePlus Medical Test*. (n.d.). Medlineplus.gov. Retrieved March 14, 2022, from <https://medlineplus.gov/lab-tests/albumin-blood-test/#:~:text=Albumin%20is%20a%20protein%20made>
3. *Similar Protein Structure Assemblies*. (2014). Nih.gov. Retrieved March 14, 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*. Wwww.rcsb.org. Retrieved March 14, 2022, from <https://www.rcsb.org/structure/4L8U>
5. *4L8U: VAST+ Similar Structure Assemblies and Proteins*. (n.d.). Wwww.ncbi.nlm.nih.gov. Retrieved March 14, 2022, from <https://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi?uid=4L8U>
6. *1N5U: X-Ray Study Of Human Serum Albumin Complexed With Heme*. (n.d.). Wwww.ncbi.nlm.nih.gov. Retrieved March 14, 2022, from <https://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbbsrv.cgi?uid=1N5U>
7. *iCn3D: Web-based 3D Structure Viewer*. (n.d.). Wwww.ncbi.nlm.nih.gov. Retrieved March 14, 2022, from <https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html?showalignseq=1&align=112122>

## WEBLEM 7b

### DALI

(URL: <http://ekhidna2.biocenter.helsinki.fi/dali/>)

#### AIM:

To perform structural Blast for Albumin using DALI tool.

#### INTRODUCTION:

Albumin is a protein made by your liver. Albumin helps keep fluid in your bloodstream so it doesn't leak into other tissues. It is also carries various substances throughout your body, including hormones, vitamins, and enzymes. Low albumin levels can indicate a problem with your liver or kidneys. Structures similar to albumin can be retrieved using DALI tool.

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.

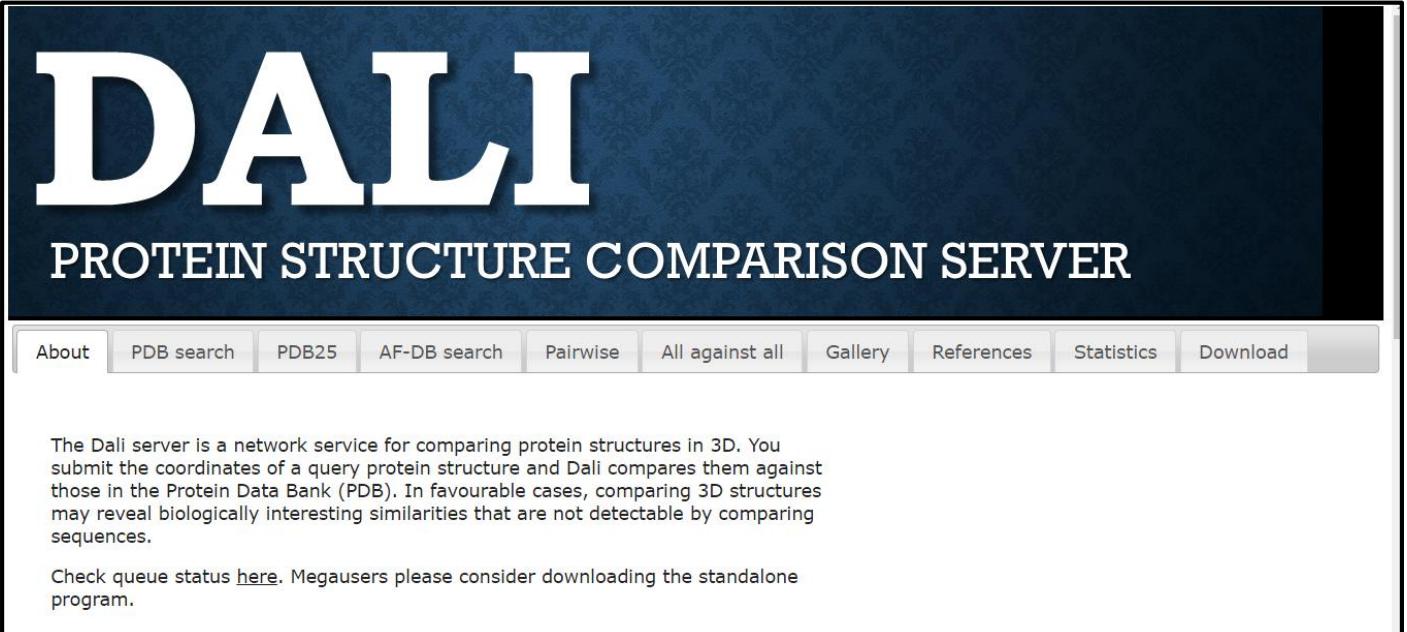
User can perform three types of database searches:

- 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:

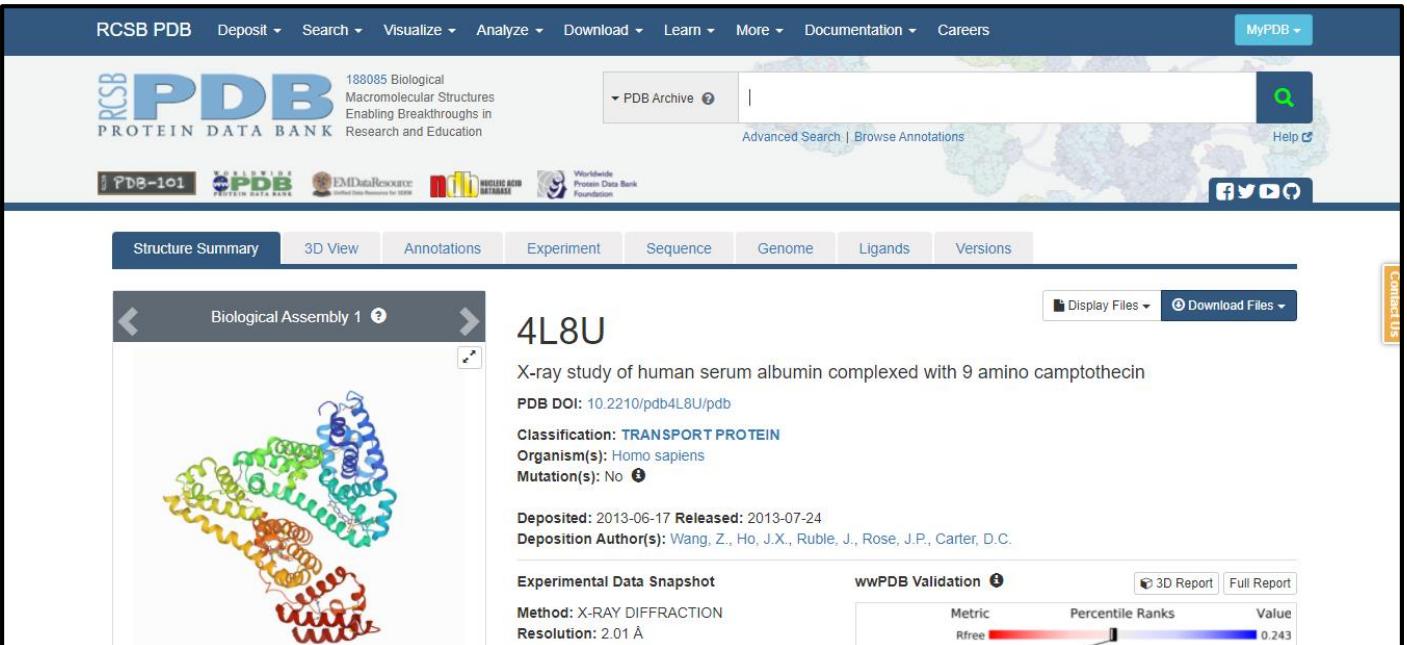
1. Open homepage for DALI. (URL: <http://ekhidna2.biocenter.helsinki.fi/dali/>)
2. Enter Albumin PDB ID.
3. Observe similar structures matches against PDB25, PDB50, PDB90 and all PDB structures.
4. Interpret the results.

## OBSERVATION:



The image shows the homepage of the DALI Protein Structure Comparison Server. The header features a large, stylized 'DALI' logo in white on a dark blue background, with the text 'PROTEIN STRUCTURE COMPARISON SERVER' below it. A navigation bar at the top includes links for 'About', 'PDB search', 'PDB25', 'AF-DB search', 'Pairwise', 'All against all', 'Gallery', 'References', 'Statistics', and 'Download'. Below the navigation bar, a text box explains the service: '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.' It also mentions a 'Check queue status [here](#)' and a note for 'Megauers please consider downloading the standalone program.' A footer at the bottom of the page states 'You can perform three types of database searches'.

Fig1. Homepage for DALI



The image shows the RSCB PDB (Protein Data Bank) page for the structure 4L8U. The header includes links for 'Structure Summary', '3D View', 'Annotations', 'Experiment', 'Sequence', 'Genome', 'Ligands', and 'Versions'. The main content area displays a 3D ribbon model of the protein structure, labeled '4L8U'. Below the model, text provides details: 'X-ray study of human serum albumin complexed with 9 amino camptothecin', 'PDB DOI: 10.2210/pdb4L8U/pdb', 'Classification: TRANSPORT PROTEIN', 'Organism(s): Homo sapiens', 'Mutation(s): No', 'Deposited: 2013-06-17 Released: 2013-07-24', 'Deposition Author(s): Wang, Z., Ho, J.X., Ruble, J., Rose, J.P., Carter, D.C.', 'Experimental Data Snapshot', 'Method: X-RAY DIFFRACTION', 'Resolution: 2.01 Å', 'wwPDB Validation', and a '3D Report' and 'Full Report' button. A color scale bar for 'Rfree' is shown at the bottom right.

Fig2. Albumin PDB structure

About PDB search PDB25 AF-DB search Pairwise All against all Gallery References Statistics Download

## PDB search

Compare query structure against Protein Data Bank.

STEP 1 - Enter your query protein structure

Structures may be specified by concatenating the PDB identifier (4 characters) and a chain identifier (1 character) or, alternatively, you may upload a PDB file.

PDB identifier + chain identifier OR upload file Choose file No file chosen

STEP 2 - Optional data

You may leave an e-mail address for notification when the job has finished. The job title is used as subject heading in the e-mail.

Job title  
E-mail

STEP 3 - Submit your job

ekhidna2.biocenter.helsinki.fi/dali/#tabs-1

**Fig3. PDB search**

About PDB search PDB25 AF-DB search Pairwise All against all Gallery References Statistics Download

## PDB search

Compare query structure against Protein Data Bank.

STEP 1 - Enter your query protein structure

Structures may be specified by concatenating the PDB identifier (4 characters) and a chain identifier (1 character) or, alternatively, you may upload a PDB file.

4L8U OR upload file Choose file No file chosen  
4L8uA SERUM ALBUMIN

STEP 2 - Optional data

You may leave an e-mail address for notification when the job has finished. The job title is used as subject heading in the e-mail.

Job title  
E-mail

STEP 3 - Submit your job

**Fig3.1. PDB search for Albumin**

## Results:

Chain: 4l8uA

- [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 for Albumin**

### Matches against PDB25:

## Results: 4l8uA

### Query: 4l8uA

MOLECULE: SERUM ALBUMIN;

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:	5orf-A	38.1	4.4	581	586	73	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
2:	1kw2-A	17.2	13.0	262	455	20	<a href="#">PDB</a>	MOLECULE: VITAMIN D-BINDING PROTEIN;
3:	5yje-B	4.7	5.6	91	318	13	<a href="#">PDB</a>	MOLECULE: PROTEIN HIRA;
4:	7k0m-D	4.7	5.1	49	153	10	<a href="#">PDB</a>	MOLECULE: SERINE PALMITOYLTRANSFERASE 1;
5:	2apl-A	4.6	7.5	75	149	11	<a href="#">PDB</a>	MOLECULE: HYPOTHETICAL PROTEIN PG0816;
6:	4mk6-A	4.6	3.6	72	188	4	<a href="#">PDB</a>	MOLECULE: PROBABLE DIHYDROXYACETONE KINASE REGULATOR DHSK_R
7:	2n5n-A	4.6	2.8	52	86	10	<a href="#">PDB</a>	MOLECULE: CHROMODOMAIN-HELICASE-DNA-BINDING PROTEIN 4;
8:	2a7g-A	4.5	3.5	68	88	9	<a href="#">PDB</a>	MOLECULE: PROBABLE RNA POLYMERASE SIGMA-C FACTOR;
9:	6cnb-R	4.4	11.6	106	522	8	<a href="#">PDB</a>	MOLECULE: DNA-DIRECTED RNA POLYMERASE III SUBUNIT RPC1;
10:	3kdw-A	4.4	3.8	71	206	7	<a href="#">PDB</a>	MOLECULE: PUTATIVE SUGAR BINDING PROTEIN;
11:	2l1rm-A	4.4	2.9	67	84	13	<a href="#">PDB</a>	MOLECULE: UNCHARACTERIZED PROTEIN YNGO;
12:	5g5p-B	4.3	3.5	90	455	3	<a href="#">PDB</a>	MOLECULE: NUCLEAR mRNA EXPORT PROTEIN SAC3;
13:	3h36-A	4.2	3.1	52	78	12	<a href="#">PDB</a>	MOLECULE: POLYRIBONUCLEOTIDE NUCLEOTIDYLTRANSFERASE;
14:	5w7p-A	4.1	10.0	79	397	14	<a href="#">PDB</a>	MOLECULE: OXAC;
15:	4cc9-B	4.1	8.7	55	98	5	<a href="#">PDB</a>	MOLECULE: PROTEIN VPRBP;
16:	6iac-A	4.0	5.8	73	301	3	<a href="#">PDB</a>	MOLECULE: PORTAL PROTEIN;
17:	6ue2-A	4.0	4.5	60	154	10	<a href="#">PDB</a>	MOLECULE: 3' RNA

**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=4l8uA Sbjct=5orfA Z-score=38.1

[back to top](#)

## Fig5.1. Result for pairwise structural alignment

```

REMARK Coordinates of 5orf rotated and translated as follows:
REMARK | 0.89773 -0.44037 -0.01257 | x | -5.000 |
REMARK | -0.40732 -0.84054 0.35719 | * | y | + | 36.000 |
REMARK | -0.16786 -0.31554 -0.93395 | z | | 85.000 |
REMARK
REMARK HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER TRANSPORT PROTEIN 16-AUG-17 5ORF
TITLE STRUCTURE OF OVINE SERUM ALBUMIN IN P1 SPACE GROUP
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: SERUM ALBUMIN;
COMPND 3 CHAIN: A, B, C, D
AUTHOR J. A. TALAJ, A. BUJACZ, G. BUJACZ, A. J. PIETRZYK-BRZEZINSKA
HELIX 1 A A1 SER A 5 GLY A 15 1 11
HELIX 2 A A2 GLY A 15 LEU A 31 1 17
HELIX 3 A A3 PRO A 35 ASP A 56 1 22
HELIX 4 A A4 SER A 65 LYS A 76 1 12
HELIX 5 A A5 ASP A 89 LYS A 93 5 5
HELIX 6 A A6 PRO A 96 HIS A 185 1 10
HELIX 7 A A7 GLU A 118 ASP A 129 1 12
HELIX 8 A A8 ASP A 129 HIS A 145 1 17
HELIX 9 A A9 TYR A 149 CYS A 168 1 20
HELIX 10 AB1 ASP A 172 GLY A 206 1 35
HELIX 11 AB2 GLY A 206 PHE A 222 1 17
HELIX 12 AB3 ASP A 226 HIS A 246 1 21
HELIX 13 AB4 ASP A 246 HIS A 266 1 19
HELIX 14 AB5 HIS A 266 SER A 271 1 6
HELIX 15 AB6 PRO A 281 GLU A 291 1 11
HELIX 16 AB7 LEU A 304 ALA A 309 1 6
HELIX 17 AB8 GLU A 313 ALA A 321 1 9
HELIX 18 AB9 ALA A 321 ARG A 335 1 15
HELIX 19 AC1 ALA A 341 CYS A 360 1 20
HELIX 20 AC2 ASP A 364 ALA A 370 1 7
HELIX 21 AC3 THR A 371 VAL A 380 1 10
HELIX 22 AC4 VAL A 380 ALA A 414 1 35
HELIX 23 AC5 SER A 418 CYS A 437 1 20
HELIX 24 AC6 PRO A 440 GLU A 464 1 25
HELIX 25 AC7 SER A 469 GLU A 478 1 10
HELIX 26 AC8 ASN A 482 LEU A 490 1 9
HELIX 27 AC9 ASP A 502 PHE A 506 5 5
HELIX 28 AC10 ALA A 502 IGLU A 515 1 2

```

Fig5.2. Result for coordinates of similar structure

## Matches against PDB50:

## Results: 4l8uA

Query: 4l8uA

### MOLECULE: SERUM ALBUMIN;

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:	5orf-A	38.1	4.4	581	586	73	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 2:	6faf-A	35.7	4.4	501	575	33	<a href="#">PDB</a>	MOLECULE: AFAMIN;
□ 3:	1kw2-A	17.2	13.0	262	455	20	<a href="#">PDB</a>	MOLECULE: VITAMIN D-BINDING PROTEIN;
□ 4:	513w-A	5.3	5.7	103	297	2	<a href="#">PDB</a>	MOLECULE: SIGNAL RECOGNITION PARTICLE RECEPTOR FTSY;
□ 5:	5yje-B	4.7	5.6	91	318	13	<a href="#">PDB</a>	MOLECULE: PROTEIN HIRA;
□ 6:	7k0m-D	4.7	5.1	49	153	10	<a href="#">PDB</a>	MOLECULE: SERINE PALMITOYLTRANSFERASE 1;
□ 7:	2apl-A	4.6	7.5	75	149	11	<a href="#">PDB</a>	MOLECULE: HYPOTHETICAL PROTEIN PG0816;
□ 8:	4mk6-A	4.6	3.6	72	188	4	<a href="#">PDB</a>	MOLECULE: PROBABLE DIHYDROXYACETONE KINASE REGULATOR DHSK_R;
□ 9:	2n5n-A	4.6	2.8	52	86	10	<a href="#">PDB</a>	MOLECULE: CHROMODOMAIN-HELICASE-DNA-BINDING PROTEIN 4;
□ 10:	4he8-I	4.5	5.8	99	427	13	<a href="#">PDB</a>	MOLECULE: NADH-QUINONE OXIDOREDUCTASE SUBUNIT 7;
□ 11:	207g-A	4.5	3.5	68	88	9	<a href="#">PDB</a>	MOLECULE: PROBABLE RNA POLYMERASE SIGMA-C FACTOR;
□ 12:	4hfq-A	4.4	4.6	73	203	11	<a href="#">PDB</a>	MOLECULE: MUTT/NUDIX FAMILY PROTEIN;
□ 13:	6cnb-R	4.4	11.6	106	522	8	<a href="#">PDB</a>	MOLECULE: DNA-DIRECTED RNA POLYMERASE III SUBUNIT RPC1;
□ 14:	3kdw-A	4.4	3.8	71	206	7	<a href="#">PDB</a>	MOLECULE: PUTATIVE SUGAR BINDING PROTEIN;
□ 15:	2lrm-A	4.4	2.9	67	84	13	<a href="#">PDB</a>	MOLECULE: UNCHARACTERIZED PROTEIN YMGD;
□ 16:	2o4t-A	4.3	6.4	67	90	6	<a href="#">PDB</a>	MOLECULE: BH3976 PROTEIN;

## Fig6. 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=4l8uA Sbjct=5orfA Z-score=38.1

[back to top](#)

DSSP	-L L L H L L L L	59
Query	-A H K S E V A R H F K D L G E E N K A L V L I A F Q A Y Q L Q C P F E D H V K L V N E T E F A K T C V A D E S A E	59
ident		
subjct	d T H K S E A T H R F N D L G E E N Q L G V L I A F Q S Y L Q Q C P F E D H V K L V K E L T E F A K T C V A D E S H	60
psr		

```

DSSP  LLLHHHHHHHHHHHL - LLLLHHHHHHHHHHHHHL LHHHHHHHHHHHL LLLL LLLL LLLL1
Query NCDKSLSLHTLFGDKLCL-TVATLRETGYEMADCCAKQEPPERNECFQLHHKDNPNLPRLVRPe 118
ident
Sbjct GCDKSLSLHTLFGDLCKVATLRET - YGMDADCCAKQEPPERNECFLNHHKDNPDLPLKLPKPE- 118
DSSP  LLLLHHHHHHHHHHHL - LLLLHHHHHHHHHHHHHL LHHHHHHHHHHHL LLLL LLLL LLLL1

```

Fig6.1. Result for pairwise structural alignment

REMARK Coordinates of Sorf rotated and translated as follows:  
REMARK | 0.89773 -0.44037 -0.01257 | | x | | -5.000 |  
REMARK | -0.40732 -0.84054 0.35719 | \* | y | + | 36.000 |  
REMARK | -0.16786 -0.31554 -0.93395 | | z | | 85.000 |  
REMARK  
REMARK HOH and TIP residues excluded. Only first MODEL passed through.  
REMARK  
HEADER TRANSPORT PROTEIN 16-AUG-17 SORF  
TITLE STRUCTURE OF OVINE SERUM ALBUMIN IN P1 SPACE GROUP  
COMPND MOL\_ID: 1;  
COMPND 2 MOLECULE: SERUM ALBUMIN;  
COMPND 3 CHAIN: A, B, C, D  
AUTHOR J.A.TALAJ,A.BUJACZ,G.BUJACZ,A.J.PIETRZYK-BRZEZINSKA  
HELIX 1 AA1 SER A 5 GLY A 15 1 11  
HELIX 2 AA2 GLY A 15 LEU A 31 1 17  
HELIX 3 AA3 PRO A 35 ASP A 56 1 22  
HELIX 4 AA4 SER A 65 LYS A 76 1 12  
HELIX 5 AA5 ASP A 89 LYS A 93 5 5  
HELIX 6 AA6 PRO A 96 HIS A 105 1 10  
HELIX 7 AA7 GLU A 118 ASP A 129 1 12  
HELIX 8 AA8 ASP A 129 HIS A 145 1 17  
HELIX 9 AA9 TYR A 149 CYS A 168 1 20  
HELIX 10 AB1 ASP A 172 GLY A 206 1 35  
HELIX 11 AB2 GLY A 206 PHE A 222 1 17  
HELIX 12 AB3 ASP A 226 HIS A 246 1 21  
HELIX 13 AB4 ASP A 248 HIS A 266 1 19  
HELIX 14 AB5 HIS A 266 SER A 271 1 6  
HELIX 15 AB6 PRO A 281 GLU A 291 1 11  
HELIX 16 AB7 LEU A 304 ALA A 309 1 6  
HELIX 17 AB8 GLU A 313 ALA A 321 1 9  
HELIX 18 AB9 ALA A 321 ARG A 335 1 15  
HELIX 19 AC1 ALA A 341 CYS A 360 1 20  
HELIX 20 AC2 ASP A 364 ALA A 370 1 7  
HELIX 21 AC3 THR A 371 VAL A 380 1 10  
HELIX 22 AC4 VAL A 380 ALA A 414 1 35  
HELIX 23 AC5 SER A 418 CYS A 437 1 20  
HELIX 24 AC6 PRO A 440 GLU A 464 1 25  
HELIX 25 AC7 SER A 469 GLU A 478 1 10  
HELIX 26 AC8 ASN A 482 LEU A 490 1 9  
HELIX 27 AC9 ASP A 502 PHE A 506 5 5  
HELIX 28 AD1 ALA A 510 LEU A 515 1 6

**Fig6.2. Result for coordinates of similar structure**

## Matches against PDB90:

### Results: 4l8uA

#### Query: 4l8uA

MOLECULE: SERUM ALBUMIN;

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:	2xsi-A	46.8	0.8	582	584	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
2:	4f5t-A	42.0	3.8	581	583	76	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
3:	5yxe-A	40.2	4.5	572	572	82	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
4:	4f5s-B	39.0	4.9	581	583	76	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
5:	6m58-B	38.9	4.3	492	550	95	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
6:	6z11-A	38.3	3.6	553	555	98	<a href="#">PDB</a>	MOLECULE: ALBUMIN;
7:	5orf-A	38.1	4.4	581	586	73	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
8:	4f5v-A	37.9	5.1	581	584	73	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
9:	6fak-A	35.7	4.4	501	575	33	<a href="#">PDB</a>	MOLECULE: AFAMIN;
10:	5ghk-A	35.0	4.1	542	563	80	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
11:	6rq7-B	31.8	4.9	473	479	36	<a href="#">PDB</a>	MOLECULE: AFAMIN;
12:	1kv2-A	17.2	13.0	262	455	20	<a href="#">PDB</a>	MOLECULE: VITAMIN D-BINDING PROTEIN;
13:	1lot-A	15.8	8.5	237	436	15	<a href="#">PDB</a>	MOLECULE: VITAMIN D-BINDING PROTEIN;
14:	6pv7-A	5.3	8.9	92	387	10	<a href="#">PDB</a>	MOLECULE: FUSION PROTEIN OF NEURONAL ACETYLCHOLINE RECEPTOR
15:	5l3w-A	5.3	5.7	103	297	2	<a href="#">PDB</a>	MOLECULE: SIGNAL RECOGNITION PARTICLE RECEPTOR FTSY;
16:	5yje-B	4.7	5.6	91	318	13	<a href="#">PDB</a>	MOLECULE: PROTEIN HIRA;
17:	7k0m-D	4.7	5.1	49	153	10	<a href="#">PDB</a>	MOLECULE: SERINE PALMITOYLTRANSFERASE 1;

**Fig7. 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=4l8uA Sbjct=2xsiA Z-score=46.8

[back to top](#)

## Fig7.1. Result for pairwise structural alignment

```

REMARK Coordinates of 2xsi rotated and translated as follows:
REMARK | 0.99987 0.01571 0.00300 | x | -0.000 |
REMARK | -0.01569 0.99986 -0.00614 | * y | + -3.000 |
REMARK | -0.00310 0.00610 0.99998 | z | -0.000 |
REMARK
REMARK HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER TRANSPORT PROTEIN 29-OCT-10 2XSI
TITLE HUMAN SERUM ALBUMIN COMPLEXED WITH DANSYL-L-GLUTAMATE AND
TITLE 2 MYRISTIC ACID
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: SERUM ALBUMIN;
COMPND 3 CHAIN: A;
COMPND 4 ENGINEERED: YES
AUTHOR J.GHUMAN,S.CURRY
HELIX 1 1 SER A 5 GLY A 15 1 11
HELIX 2 2 GLY A 15 LEU A 31 1 17
HELIX 3 3 PRO A 35 ASP A 56 1 22
HELIX 4 4 SER A 65 VAL A 77 1 13
HELIX 5 5 THR A 79 GLY A 85 1 7
HELIX 6 6 GLU A 86 ALA A 92 5 7
HELIX 7 7 PRO A 96 GLN A 104 1 9
HELIX 8 8 GLU A 119 ASN A 130 1 12
HELIX 9 9 ASN A 130 HIS A 146 1 17
HELIX 10 10 TYR A 150 CYS A 169 1 20
HELIX 11 11 ASP A 173 PHE A 206 1 34
HELIX 12 12 PHE A 205 PHE A 223 1 18
HELIX 13 13 GLU A 227 GLY A 248 1 22
HELIX 14 14 ASP A 249 ASN A 267 1 19
HELIX 15 15 GLN A 268 ILE A 271 5 4
HELIX 16 16 LEU A 275 LYS A 281 1 7
HELIX 17 17 PRO A 282 VAL A 293 1 12
HELIX 18 18 LEU A 305 VAL A 310 1 6
HELIX 19 19 ASP A 314 ALA A 322 1 9
HELIX 20 20 ALA A 322 ARG A 337 1 16
HELIX 21 21 SER A 342 CYS A 361 1 20
HELIX 22 22 ASP A 365 ALA A 371 1 7
HELIX 23 23 VAL A 373 LEU A 398 1 26
HELIX 24 24 GLY A 393 VAL A 415 1 17
HELIX 25 25 SER A 419 CYS A 438 1 20
HELIX 26 26 PRO A 441 LYS A 466 1 26

```

Fig7.2. Result for coordinates of similar structure

## Matches against all PDB structures:

## Results: 4l8uA

Query: 4l8uA

### MOLECULE: SERUM ALBUMIN;

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:	418u-A	56.2	0.0	583	583	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 2:	1n5u-A	53.4	0.4	583	583	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 3:	6hsc-B	52.4	0.7	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 4:	4269-I	51.5	0.6	581	581	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 5:	7vr9-B	51.2	0.8	578	578	99	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 6:	6uwu-A	51.1	0.7	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 7:	6uwu-B	51.1	0.9	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 8:	3a73-A	50.9	1.0	576	576	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 9:	3uiv-A	50.6	0.6	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 10:	2bxm-A	50.5	0.7	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 11:	5id7-B	50.5	1.3	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 12:	1e7f-A	50.0	0.8	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 13:	3cx9-A	50.0	0.6	581	581	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 14:	2i30-A	49.6	0.8	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 15:	5vnw-A	49.5	1.7	583	583	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 16:	1e7g-A	49.5	0.9	582	582	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;
□ 17:	418u-A	49.4	0.7	581	581	100	<a href="#">PDB</a>	MOLECULE: SERUM ALBUMIN;

## Fig8. Result for similar structures

No 2: Query=418uA Sbct=1u5uA Z-score=53.4

[back to top](#)

Fig8.1. Result for pairwise structural alignment

```

REMARK  Coordinates of 1n5u rotated and translated as follows:
REMARK  | 1.00000 -0.00131 -0.00077 | | x | | 0.000 |
REMARK  | 0.00131 1.00000 -0.00264 | * | y | + | 0.000 |
REMARK  | 0.00078 0.00264 1.00000 | | z | | -0.000 |
REMARK
REMARK HOH and TIP residues excluded. Only first MODEL passed through.
REMARK
HEADER PLASMA PROTEIN          07-NOV-02  1N5U
TITLE  X-RAY STUDY OF HUMAN SERUM ALBUMIN COMPLEXED WITH HEME
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: SERUM ALBUMIN;
COMPND 3 CHAIN: A
AUTHOR M.WARDELL,Z.WANG,J.X.HO,J.ROBERT,F.RUKER,J.RUBLE,D.C.CARTER
HELIX 1 1 SER A 5 GLY A 15 1 11
HELIX 2 2 GLY A 15 LEU A 31 1 17
HELIX 3 3 PRO A 35 ASP A 56 1 22
HELIX 4 4 SER A 65 THR A 76 1 12
HELIX 5 5 THR A 79 GLY A 85 1 7
HELIX 6 6 GLU A 86 LYS A 93 5 8
HELIX 7 7 GLN A 94 HIS A 105 1 12
HELIX 8 8 GLU A 119 ASN A 130 1 12
HELIX 9 9 ASN A 130 HIS A 146 1 17
HELIX 10 10 TYR A 150 CYS A 169 1 20
HELIX 11 11 ASP A 173 GLY A 207 1 35
HELIX 12 12 GLY A 207 PHE A 223 1 17
HELIX 13 13 GLU A 227 GLY A 248 1 22
HELIX 14 14 ASP A 249 GLU A 266 1 18
HELIX 15 15 ASN A 267 ILE A 271 5 5
HELIX 16 16 LEU A 275 GLU A 280 1 6
HELIX 17 17 PRO A 282 GLU A 292 1 11
HELIX 18 18 LEU A 305 VAL A 310 1 6
HELIX 19 19 ASP A 314 ALA A 322 1 9
HELIX 20 20 ALA A 322 ARG A 337 1 16
HELIX 21 21 SER A 342 CYS A 361 1 20
HELIX 22 22 ASP A 365 ALA A 371 1 7
HELIX 23 23 LYS A 372 GLY A 399 1 28
HELIX 24 24 GLY A 399 VAL A 415 1 17
HELIX 25 25 SER A 419 CYS A 438 1 20
HELIX 26 26 PRO A 441 GLU A 465 1 25
HELIX 27 27 SER A 470 GLU A 479 1 10
HELIX 28 28 ASN A 483 ALA A 490 1 8

```

**Fig8.2. 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:

8. *Albumin Blood Test: MedlinePlus Medical Test.* (n.d.). Medlineplus.gov. Retrieved March 14, 2022, from <https://medlineplus.gov/lab-tests/albumin-blood-test/#:~:text=Albumin%20is%20a%20protein%20made>
9. *Dali server.* (n.d.). Ekhidna2.Biocenter.helsinki.fi. Retrieved March 14, 2022, from <http://ekhidna2.biocenter.helsinki.fi/dali/>
10. Bank, R. P. D. (n.d.-b). *RCSB PDB - 4L8U: X-ray study of human serum albumin complexed with 9 amino camptothecin.* Wwww.rcsb.org. Retrieved March 14, 2022, from <https://www.rcsb.org/structure/4L8U>
11. *Dali server.* (n.d.). Ekhidna2.Biocenter.helsinki.fi. Retrieved March 14, 2022, from <http://ekhidna2.biocenter.helsinki.fi/barcosel/tmp//4l8uA/>

## 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 described for 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 Viterbi 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 coexpressed gene sequences from the same organism. The similarity type of prediction is also called phylogenetic footprinting.

### 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 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.

### 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 as 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 identifier 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 of 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 be 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](http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter). 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](http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb). 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](http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb). 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](http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind). Retrieved March 18, 2022, from <http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind>
7. Home - *ORFfinder* – NCBI. (2019). [Nih.gov](https://www.ncbi.nlm.nih.gov/orffinder/). Retrieved March 18, 2022, from <https://www.ncbi.nlm.nih.gov/orffinder/>

## WEBLEM 8a

### TSSW

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

#### AIM:

To recognize the Protease human Pol III promoter region and start of transcription using TSSW tool.

#### INTRODUCTION:

Proteolytic enzymes (proteases) are enzymes that break down protein. These enzymes are made by animals, plants, fungi, and bacteria. Proteolytic enzymes break down proteins in the body or on the skin. This might help with digestion or with the breakdown of proteins involved in swelling and pain. Protease human Pol III promoter region and start of transcription can be 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 as 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:

**Homo sapiens neutral protease alpha subunit gene, complete cds**

GenBank: AH001431.2

[FASTA](#) [Graphics](#)

[Go to: ▾](#)

**LOCUS** AH001431 3298 bp DNA linear PRI 10-JUN-2016

**DEFINITION** Homo sapiens neutral protease alpha subunit gene, complete cds.

**ACCESSION** AH001431 M31501 M31502 M31503 M31504 M31505 M31506 M31507 M31508 M31509 M31510 M31511

**VERSION** AH001431.2

**KEYWORDS** neutral protease.

**SOURCE** Homo sapiens (human)

**ORGANISM** *Homo sapiens*  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

**REFERENCE** 1 (bases 1 to 3298)

**AUTHORS** Miyake,S., Emori,Y. and Suzuki,K.

**TITLE** Gene organization of the small subunit of human calcium-activated neutral protease

**JOURNAL** Nucleic Acids Res. 14 (22), 8805-8817 (1986)

**PUBMED** 3024120

**COMMENT** On or before Jun 10, 2016 this sequence version replaced [M31501.1](#), [M31502.1](#), [M31503.1](#), [M31504.1](#), [M31505.1](#), [M31506.1](#), [M31507.1](#), [M31508.1](#), [M31509.1](#), [M31510.1](#), [M31511.1](#), [AH001431.1](#).

**FEATURES** Location/Qualifiers

**SOURCE** 1 3298

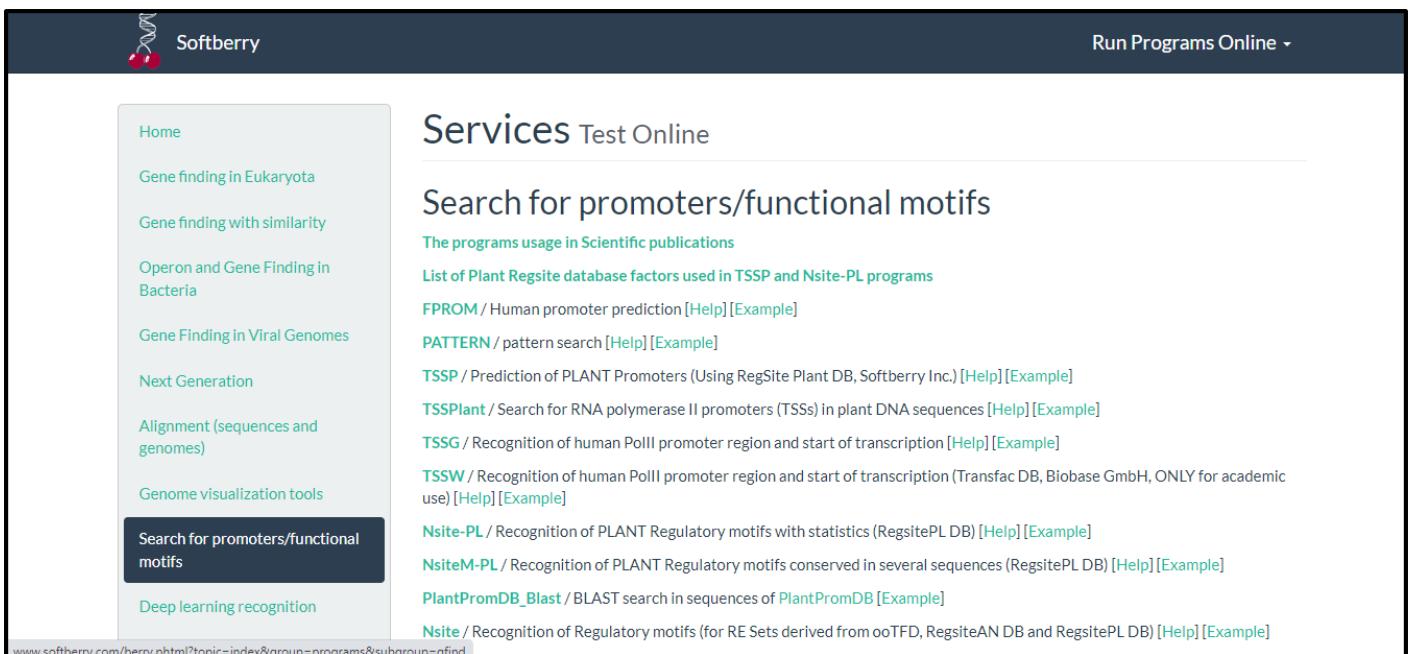
**Articles about the CAPNS1 gene**  
Dual proteome-scale networks reveal cell-specific remodeling of the human inte [Cell. 2021]  
Overexpression of Capns1 Predicts Poor Prognosis and Correlates with Tur [Urol Int. 2021]  
Circular RNA ABCB10 promotes cell proliferation and invasion, but inhibits ap [Mol Med Rep. 2021]

[See all...](#)

**Reference sequence information**  
DefSeq alternative policies

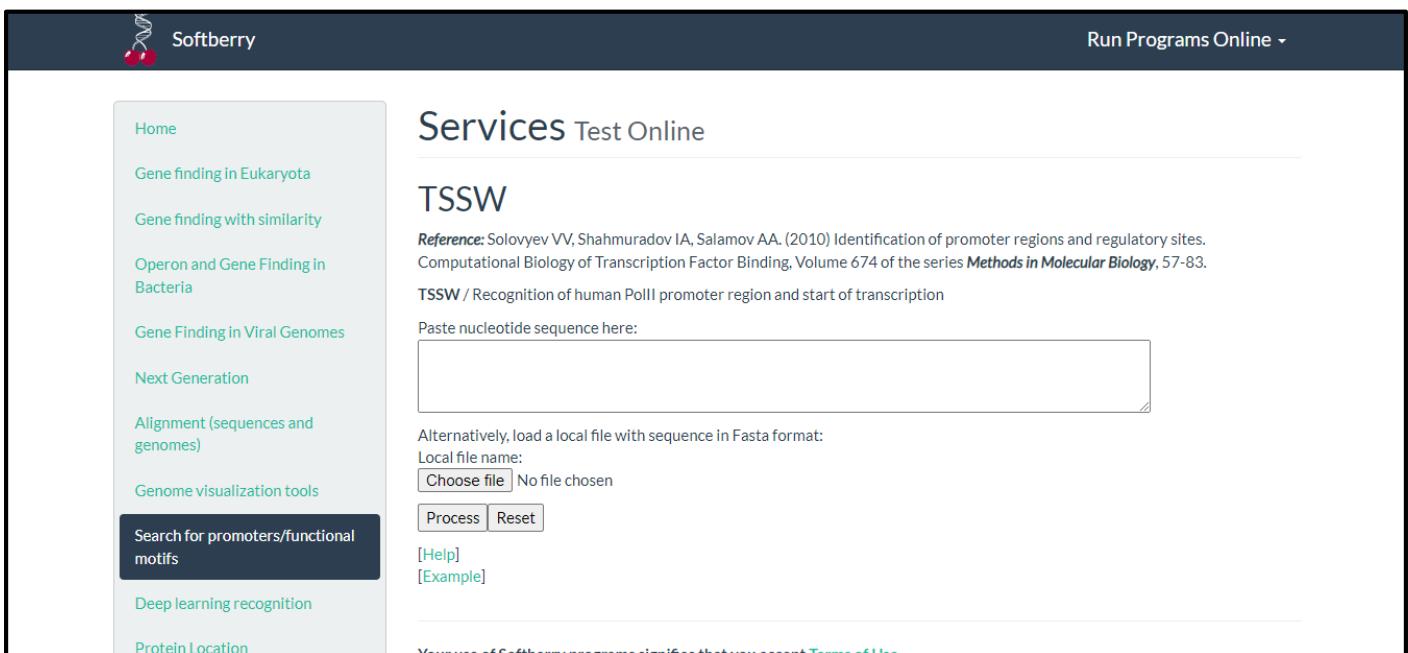
**Fig1. GenBank result for Human Protease**

### Fig3. Homepage for Softberry



The screenshot shows the Softberry Services Test Online interface. The left sidebar contains links to various services: Home, Gene finding in Eukaryota, Gene finding with similarity, Operon and Gene Finding in Bacteria, Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs (which is highlighted in a dark blue box), and Deep learning recognition. The main content area is titled "Search for promoters/functional motifs" and lists several tools: The programs usage in Scientific publications, List of Plant Regsite database factors used in TSSP and Nsite-PL programs, FPROM / Human promoter prediction [Help] [Example], PATTERN / pattern search [Help] [Example], TSSP / Prediction of PLANT Promoters (Using RegSite Plant DB, Softberry Inc.) [Help] [Example], TSSPlant / Search for RNA polymerase II promoters (TSSs) in plant DNA sequences [Help] [Example], TSSG / Recognition of human PolII promoter region and start of transcription [Help] [Example], TSSW / Recognition of human PolII promoter region and start of transcription (Transfac DB, Biobase GmbH, ONLY for academic use) [Help] [Example], Nsite-PL / Recognition of PLANT Regulatory motifs with statistics (RegsitePL DB) [Help] [Example], NsiteM-PL / Recognition of PLANT Regulatory motifs conserved in several sequences (RegsitePL DB) [Help] [Example], PlantPromDB\_Blast / BLAST search in sequences of PlantPromDB [Example], and Nsite / Recognition of Regulatory motifs (for RE Sets derived from ooTFD, RegsiteAN DB and RegsitePL DB) [Help] [Example]. A URL at the bottom of the sidebar is [www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=qfind](http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=qfind).

**Fig4. Tools for promoters/functional motifs**



The screenshot shows the Softberry Services Test Online interface, specifically for the TSSW tool. The left sidebar contains links to Home, Gene finding in Eukaryota, Gene finding with similarity, Operon and Gene Finding in Bacteria, Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs, Deep learning recognition, and Protein Location. The main content area is titled "Services Test Online" and "TSSW". It includes a reference section: "Reference: Solovyev VV, Shahmuradov IA, Salamov AA. (2010) Identification of promoter regions and regulatory sites. Computational Biology of Transcription Factor Binding, Volume 674 of the series *Methods in Molecular Biology*, 57-83." Below this is a section for "TSSW / Recognition of human PolII promoter region and start of transcription". It features a text input field for "Paste nucleotide sequence here:", a file input field for "Alternatively, load a local file with sequence in Fasta format: Local file name: Choose file No file chosen", and buttons for "Process" and "Reset". Below these are links for "[Help]" and "[Example]". A small note at the bottom states: "Your use of Softberry programs signifies that you accept Terms of Use".

**Fig5. Homepage for TSSW**



Home  
 Gene finding in Eukaryota  
 Gene finding with similarity  
 Operon and Gene Finding in Bacteria  
 Gene Finding in Viral Genomes  
 Next Generation  
 Alignment (sequences and genomes)  
 Genome visualization tools  
**Search for promoters/functional motifs**  
 Deep learning recognition  
 Protein Location

## Services Test Online

### TSSW

Reference: Solovyev VV, Shahmuradov IA, Salamov AA. (2010) Identification of promoter regions and regulatory sites. Computational Biology of Transcription Factor Binding, Volume 674 of the series *Methods in Molecular Biology*, 57-83.

TSSW / Recognition of human PolII promoter region and start of transcription

Paste nucleotide sequence here:

>AH001431.2 Homo sapiens neutral protease alpha subunit gene, complete cds  
 CTGCAGAGGGCCCGTGGAGTCCTTAGTGAGCGGACCGAAAACGCCACCTGGAGGATATTGG  
 CAT

Alternatively, load a local file with sequence in Fasta format:

Local file name:

No file chosen

[\[Help\]](#)

[\[Example\]](#)

**Fig6. Search for protease nucleotide FASTA sequence**

>AH001431.2 Homo sapiens neutral protease alpha subunit gene, complete cds  
 Length of sequence- 3298  
 Thresholds for TATA+ promoters - 0.45, for TATA-/enhancers - 3.70  
 4 promoter/enhancer(s) are predicted  
 Promoter Pos: 809 LDF- 9.40  
 Enhancer Pos: 319 LDF- 7.38  
 Promoter Pos: 323 LDF- 6.96  
 Promoter Pos: 2884 LDF- 3.70  
 Transcription factor binding sites:  
 for promoter at position - 809  
 730 (-) CHICK\$ACRA CCGCCC  
 646 (-) CHICK\$ACRA CCGCCC  
 726 (-) MAIZE\$ADH1 CCCCGG  
 756 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 753 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 750 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 747 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 744 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 741 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 660 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 657 (-) DROME\$ANTP CGCCGCCGCCGCCG  
 641 (+) HS\$APOE\_08 GGGCGG  
 725 (+) HS\$APOE\_08 GGGCGG  
 684 (-) HS\$BG\_01 ccaCACCCg  
 571 (+) HS\$GG\_13 CACCC  
 681 (-) HS\$GG\_13 CACCC  
 684 (-) HS\$GG\_14 ctcCACCCatggg  
 696 (-) HS\$GMCSF\_0 CATT  
 805 (+) YSHIS3\_02 GACTCA  
 579 (+) HS\$HH4\_02 GGTCC  
 657 (+) HSV1\$IE3\_0 GGCGGG  
 666 (+) HSV1\$IE3\_0 GGCGGG  
 790 (+) MOUSE\$MT1\_ TGCAC  
 795 (-) MOUSE\$MT1\_ TGCGC

**Fig7. Results 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:

8. Xiong, J. (2008). *Promoter and Regulatory Element Prediction. Essential bioinformatics*. Cambridge: Cambridge University Press. 113-119.
9. *PROTEOLYTIC ENZYMES (PROTEASES): Overview, Uses, Side Effects, Precautions, Interactions, Dosing and Reviews.* (n.d.). [https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20\(proteases\)%20are%20enzymes](https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20(proteases)%20are%20enzymes)
10. 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>
11. *Softberry Home Page.* (n.d.). <http://www.softberry.com/>
12. *TSSW - Recognition of human PolII promoter region and start of transcription.* (n.d.). <http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter>
13. *Softberry - TSSW result.* (n.d.). <http://www.softberry.com/cgi-bin/programs/promoter/tssw.pl>

**WEBLEM 8b****BPROM**

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

**AIM:**

To predict bacterial promoter for *Neisseria gonorrhoeae* using BPROM tool.

**INTRODUCTION:**

*Neisseria gonorrhoeae* infects primarily columnar epithelium, because stratified squamous epithelium is relatively resistant to invasion. Mucosal invasion by gonococci results in a local inflammatory response that produces a purulent exudate consisting of PMNs, serum, and desquamated epithelium. Bacterial promoter region can be recognized using 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 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:**

6. Open homepage for softberry. (URL: <http://www.softberry.com/>)
7. Under operon and gene finding select BPROM. (URL: <http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>)
8. Retrieve bacterial nucleotide FASTA sequence from GenBank.
9. Process the FASTA sequence on BPROM.
10. Observe and interpret the results.

## OBSERVATION:

NCBI Resources ▾ How To ▾ Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

GenBank ▾ Send to: ▾ Change region shown

**Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence**

NCBI Reference Sequence: NZ\_CM003348.1

FASTA Graphics

Go to: ▾

Locus NZ\_CM003348 4104 bp DNA circular CON 25-FEB-2022  
 Definition Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence.  
 Accession NZ\_CM003348 NZ\_LFJW01000000  
 Version NZ\_CM003348.1  
 DBLINK BioProject: PRJNA224116  
 BioSample: SAMN03782447  
 Assembly: GCF\_001039435.1  
 Keywords WGS; RefSeq.  
 Source Neisseria gonorrhoeae  
 Organism Neisseria gonorrhoeae  
 Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales; Neisseriaceae; Neisseria.  
 Reference 1 (bases 1 to 4104)  
 Authors Ang, G.Y., Yu, C.Y., Yong, D.A., Cheong, Y.M., Yin, W.F. and Chan, K.G.  
 Title Draft Genome Sequence of Neisseria gonorrhoeae Strain NG\_869 with Penicillin, Tetracycline and Ciprofloxacin Resistance Determinants Isolated from Malaysia  
 eintercept.qualtrics.com... 36 (2), 225-227 (2016)

Analyze this sequence Run BLAST Pick Primers Highlight Sequence Features Find in this Sequence

Related information Assembly BioProject BioSample PubMed Taxonomy Components (Core) Full text in PMC

Fig1. GenBank result for *Neisseria gonorrhoeae*

NCBI Resources ▾ How To ▾ Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

FASTA ▾ Send to: ▾ Change region shown

**Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence**

NCBI Reference Sequence: NZ\_CM003348.1

GenBank Graphics

>NZ\_CM003348.1 Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence  
 GGGCGGGAAATGCCGAAGTGTCCACGGTTATCGCGGCTGATAGAGTTTCGGCTAGGTAGTCGGCG  
 CGGGATTGCGACCATGCCGAAACGGTTCTTGTGGAAACAGCGGAAGCGTTGGCAGTTCTTAGCCG  
 TAAGCTCTGCTTGGTTTGGATAGGGCATATTCCAAGCTCCGTGATGATTTTTGGTCT  
 GCTGCAATTGATGGTTGTTCCGCAAACATTCCGATGAGCGCTGCAATTCCGAAACCGTGTGATTGCTG  
 GCTGCAATATCAAACCTGTACCAAAACCGATTAGCGATTGATTTGGAGATACATTCACTTCTTTTCCC  
 ACATCGGACAGGGAAATCCCCATTACCCGATACACTGTAATGCACTTCTTCAGGCCAGCCTTGCAC  
 CGTGTGCGCTGCTGTTGAACGCCAACAGCTTGTGGCCCAACGGCTCACAAAGGTTAACACCGTG  
 CAATTCTGCTTAATCGACCGCTTCCACGTTAGCGGACCGATAGGGCTTAACATTGCTAAATCCA  
 CGAAATCGCAAGATACTCAAATCGTAGCCCCGTGTCGTCAGGAACGCCGCGCAGCGTCAAGCCATG  
 CGGATGTTGCGGATTTCGTAATCAGCGATAAAAGCCCATACCCCGCAATTTCGCTTATGCT  
 GCTTCAAGTGCAGCAAGATAAGCGCAAGGGCTTTTGCTCCACCGTATTCCGCGTCAAGCACAGGCCG  
 AAAGCGCATAGGCAAGGGTGTGCGCCCGCTTTCCCTGTTGATGCGCCCAAGCAGGCATAGGCAGATT  
 ATTGCTTCCAAAGCCAACCCGCCCTTGTAACTCCAAGCTAAAGAGCATAAACACAGCAGATGCCG  
 GGATTGACTGGATGTAGCGGCGTTGTAGGGCGCAGCGTAAGAGCGCAGCGCATAGGGCTTCTTGA  
 AATCTTGCAGTGGCTTGTGGATACGTTCTTGTCAAGAGAGGGGGTTTGCTTGTCAAGATTGCT  
 CATGTTGATCTCGAAACCCCGTGCAGATTGGCGTTGGCGGGGTTTGCTTGTCAAGATTGCT  
 ATTGATGTTGTTTAAAGATGATACAAACTATGTCAAATAACCATACAGATAACAGCCGATAGGG  
 GTTCTTATTCAAATTTCCAATCGCAATTAGCGAAAGCCAGCGGAAGCGGTAAGCTGGAGCG  
 CAGCAGCGCAGCTAAGCCGGCCAGCAGGGCGCTTTGGGGAAACATGAAACCAAGTCCGACAGGGC  
 GGGGTGCGTGTCTTCCCGGAGTCTCATGGATATGGGAGATGGCTGATGAAATGCGTTTTT

Analyze this sequence Run BLAST Pick Primers Highlight Sequence Features

Related information Assembly BioProject BioSample PubMed Taxonomy Components (Core) Full text in PMC

Genome Identical GenBank Sequence

Fig2. Nucleotide FASTA sequence

**Fig3. Homepage for Softberry**

**Fig4. Tools for bacterial promoter, operon and gene finding**

The screenshot shows the Softberry Services Test Online interface. The top navigation bar includes the Softberry logo and a "Run Programs Online" dropdown. The left sidebar lists various bioinformatics tools: Home, Gene finding in Eukaryota, Gene finding with similarity, Operon and Gene Finding in Bacteria (highlighted in a dark box), Gene Finding in Viral Genomes, Next Generation, Alignment (sequences and genomes), Genome visualization tools, Search for promoters/functional motifs, Deep learning recognition, and Protein Location. The main content area is titled "Services Test Online" and "BPROM". It states that BPROM is used in over 800 publications and is a bacterial sigma70 promoter recognition program. A text input field is provided for pasting nucleotide sequences, and an "Alternatively, load a local file with sequence" section includes a "Choose file" button. At the bottom are "Process" and "Reset" buttons.

Fig5. Homepage for BPROM

The screenshot shows the Softberry Services Test Online interface, identical to Fig5, but with a search result displayed in the nucleotide sequence input field. The sequence is: >NZ\_CM003348.1 Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence GGCCGGGAAATGCCGAAGTGTCCACGGTTATCGCGGCTGATAGAGTTTCGG. The rest of the interface is the same, with the "Process" and "Reset" buttons at the bottom.

Fig6. Search for nucleotide FASTA sequence

NZ\_CM003348.1 *Neisseria gonorrhoeae* strain NG\_869 plasmid pNG869\_3, whole geno  
 Length of sequence- 4104  
 Threshold for promoters - 0.20  
 Number of predicted promoters - 9

Promoter Pos.	Sequence	Score
1052	LDF- 5.65	
-10 box at pos. 1037	GTGTATAAT	84
-35 box at pos. 1015	TTGCAA	55
Promoter Pos: 1439	LDF- 3.02	
-10 box at pos. 1424	GGGTATT	57
-35 box at pos. 1399	TTGAGC	21
Promoter Pos: 4037	LDF- 1.77	
-10 box at pos. 4022	CGTTA	48
-35 box at pos. 4005	TTTCAA	36
Promoter Pos: 2067	LDF- 1.69	
-10 box at pos. 2052	GTGTATT	50
-35 box at pos. 2032	CTGATG	17
Promoter Pos: 190	LDF- 1.49	
-10 box at pos. 175	GGCCATATT	39
-35 box at pos. 152	TTGGTG	24
Promoter Pos: 556	LDF- 1.45	
-10 box at pos. 540	GCTTAAACT	59
-35 box at pos. 521	TTAGAG	3
Promoter Pos: 2443	LDF- 1.04	
-10 box at pos. 2428	TTTGAAAT	41
-35 box at pos. 2406	TGGAAA	23
Promoter Pos: 3307	LDF- 0.92	
-10 box at pos. 3291	GGCTAGCCT	40
-35 box at pos. 3275	TCGCCA	24
Promoter Pos: 2888	LDF- 0.47	
-10 box at pos. 2873	TTTGACACT	24
-35 box at pos. 2853	TTGGAA	32

Oligonucleotides from known TF binding sites:

**Fig7. Results for predicted promoters**

Promoter Pos.	Sequence	Score
556	LDF- 1.45	
-10 box at pos. 540	GCTTAAACT	59
-35 box at pos. 521	TTAGAG	3
Promoter Pos: 2443	LDF- 1.04	
-10 box at pos. 2428	TTTGAAAT	41
-35 box at pos. 2406	TGGAAA	23
Promoter Pos: 3307	LDF- 0.92	
-10 box at pos. 3291	GGCTAGCCT	40
-35 box at pos. 3275	TCGCCA	24
Promoter Pos: 2888	LDF- 0.47	
-10 box at pos. 2873	TTTGACACT	24
-35 box at pos. 2853	TTGGAA	32

Oligonucleotides from known TF binding sites:

For promoter at 1052:  
 rpoD16: TGTATAAT at position 1038 Score - 13  
 soxS: AACCCCG at position 1066 Score - 10  
 For promoter at 1439:  
 purR: CGTTTTT at position 1392 Score - 8  
 fnr: TTTTTGA at position 1395 Score - 9  
 No such sites for promoter at 4037  
 No such sites for promoter at 2067  
 For promoter at 190:  
 dnaA: TTTGGATA at position 167 Score - 6  
 No such sites for promoter at 556  
 No such sites for promoter at 2443  
 For promoter at 3307:  
 rpoD17: CCCTAAAA at position 3308 Score - 11  
 No such sites for promoter at 2888

**Fig8. Results for known TF binding sites**

## RESULT:

Nucleotide FASTA sequence for *Neisseria gonorrhoeae* strain NG\_869 plasmid pNG869\_3 with length of 4104 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](http://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](https://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1)
4. *Softberry Home Page*. (n.d.). [Www.softberry.com](http://www.softberry.com). Retrieved March 18, 2022, from <http://www.softberry.com/>
5. *BPROM - Prediction of bacterial promoters*. (n.d.). [Www.softberry.com](http://www.softberry.com). Retrieved March 18, 2022, from <http://www.softberry.com/berry.phtml?topic=beprom&group=programs&subgroup=gfindb>
6. *Softberry - BPROM result*. (n.d.). [Www.softberry.com](http://www.softberry.com). Retrieved March 18, 2022, from <http://www.softberry.com/cgi-bin/programs/gfindb/beprom.pl>

## WEBLEM 8c

### FGENESB

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

#### AIM:

To predict bacterial operon and gene for *Neisseria gonorrhoeae* using FGENESB tool.

#### INTRODUCTION:

*Neisseria gonorrhoeae* infects primarily columnar epithelium, because stratified squamous epithelium is relatively resistant to invasion. Mucosal invasion by gonococci results in a local inflammatory response that produces a purulent exudate consisting of PMNs, serum, and desquamated epithelium. Bacterial operon and gene can be recognized using FGENESB.

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 Viterbi 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:

11. Open homepage for softberry. (URL: <http://www.softberry.com/>)
12. Under operon and gene finding in bacteria select FGENESB. (URL: <http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb>)
13. Retrieve bacterial nucleotide FASTA sequence from GenBank.
14. Process the FASTA sequence on FGENESB.
15. Observe and interpret the results.

#### OBSERVATION:

Fig1. GenBank result for *Neisseria gonorrhoeae*

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

FASTA ▾ Send to: ▾ Change region shown ▾

Customize view ▾

**Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence**

NCBI Reference Sequence: NZ\_CM003348.1

[GenBank](#) [Graphics](#)

>NZ\_CM003348.1 Neisseria gonorrhoeae strain NG\_869 plasmid pNG869\_3, whole genome shotgun sequence

```
GGCGCAGGAAATGCCGAAGTGTCCACGGTTTATCGGGCTGATAGAGTTTGGCTAGGTAGTCGGCG
CGGGATTGACCATGCCGAAACCGTTCTTGTGAAACAGCGGAACGGTTGGCAGTTGACCGCTG
TAACGTCTCGTTGGTGGGTTTCAGTTGGATAGGCCATATTCAAGCTCGTGTATTTTGTCT
GCCATTGATGGTGGGTTTCAGCTTGGATAGGCCATGGCAAGTGGAAACACGGTGTAGTGTCTG
GCTGCAATCAAACCTGTACCAAACCCAGTTAGCGATTGGAGATACATTCTACTTCACTTCCC
ACATCGGGACAGGGAAATCCCCATTACCGCATACACTTGTATGCACTCTTCAGCCAGCCTGAC
CGTGTAGCCCTGCTGTTGAACGCCAACAGCTTGGCGCCAAACGGCTCACAAAGGTAAACACCGTG
CAATTCTGCTTAATCCGACGCCAGTGTAGCGACCCGTAGGGCTTAAACTGTGCTAAATCCA
CGAAATCCGCAAGAGTACTCAAATCGTAGCCCCGTAGGGCTCAGGAACGCCGGCAGCGTCAAGT
CGGATGTTGGGATTTTGTGTAATCAGCGATACAAGCCCATCAGCCGCAATTTCGCTTATAGCT
GCCATTGAGTGGGAAAGTGGGATTTTGTGTAATCAGCGATACAAGCCCATCAGCCGCAATTTCGCTTATAGCT
AAAGCGCATAGGCAAGGTTGCGCCCGTTTCCCTGTGATTTGGCGCCAAAGCAGGCATAGGCGATT
ATTGTCTTCCCAAGCCAACCCGCCCTCTGTAATCCAAGTCAAAGAGCATAAACACAGCAGATGCGC
GGATTGACTGTGATGGGCTTGTGATGGCGCGACGGCTAAAGGGCCACCAAGCAGGCTTCTTGA
AAATCTTGCAGTATGGCTTGTGATGGGATACGTTCTGCAAGAAGAGGTGGGTTGGGTGTATAATTGGCT
CATGTTGATCTGAAACCCCGTGCAGATTGGCGTTGGGGGGTTTGCTTGTCTAAGATTGCG
ATTGATGCTGTTTTAAAGTGTACAAACTATGTCAAATAACCATAAATCAGGATAACAGCCGATAGGG
GTTCTTATTCAAATTTCCAATCCGCAATTAGCGAAGCCAGCAGGGGAAGGGTAAAGCTTGGCG
CAGCAGCGCAGCTAAGCGGGCAGCAGGGGGGTTGGGGAAACATGAAACAGTCCGACAGGGC
GGCGTGTGTTCTCCGGAGTCTTGTGATGGGAAATGCGGTGATGAAATGCGTTTTT
```

Analyze this sequence ▾

Run BLAST

Pick Primers

Highlight Sequence Features

Related information ▾

Assembly

BioProject

BioSample

PubMed

Taxonomy

Components (Core)

Full text in PMC

Genome

Identical GenBank Sequence

**Fig2. Nucleotide FASTA sequence**

Softberry Run Programs Online ▾

**Annotation of Plant Genomes**

○○○○●○○○

**MOLQUEST** About Downloads Products Services In publications Management Contacts

**Cloud computing services**  
Data analysis using Softberry, public or clients' own pipelines in AWS cloud. Adopting pipelines to run on cloud computer clusters.

**Annotation of Animal Genomes**  
Gene identification, HMM Fgenesh gene finder and Fgenesh++ genome annotation pipeline, building gene

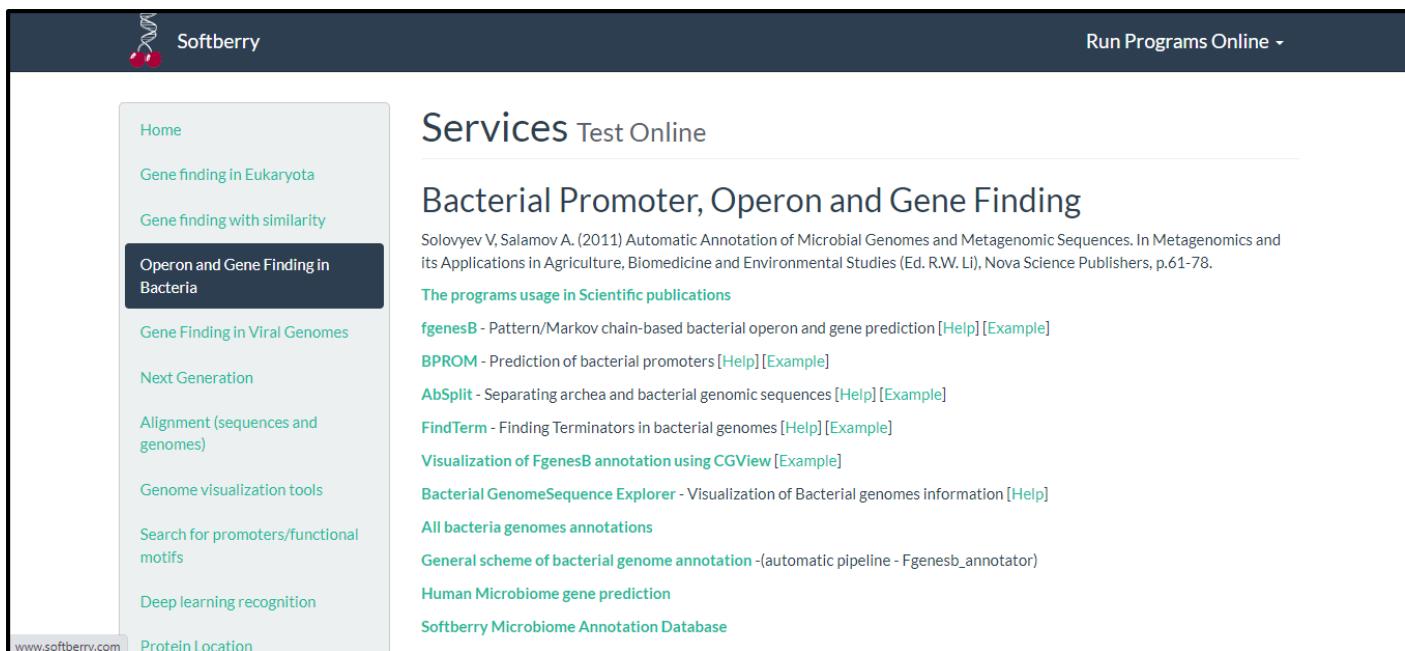
**Next generation**  
Genome and transcripts assembling, Reads Mapping, Alternative transcripts (Transomics pipeline), Snp discovery and evaluation, visualization

**Annotation of Plant Genomes**  
Gene identification, Fgenesh gene finder and Fgenesh++ genome annotation pipeline, 42 custom

**Annotation of Bacterial Genomes**  
Bacterial gene, promoters, terminators, operons identification, metagenomics, Fgenesh pipeline, Microbiome sequence analysis and annotation.

**Genome regulation analysis**  
De novo finding regulatory motifs, search for non-random occurrence of functional motifs, plant and

**Fig3. Homepage for Softberry**

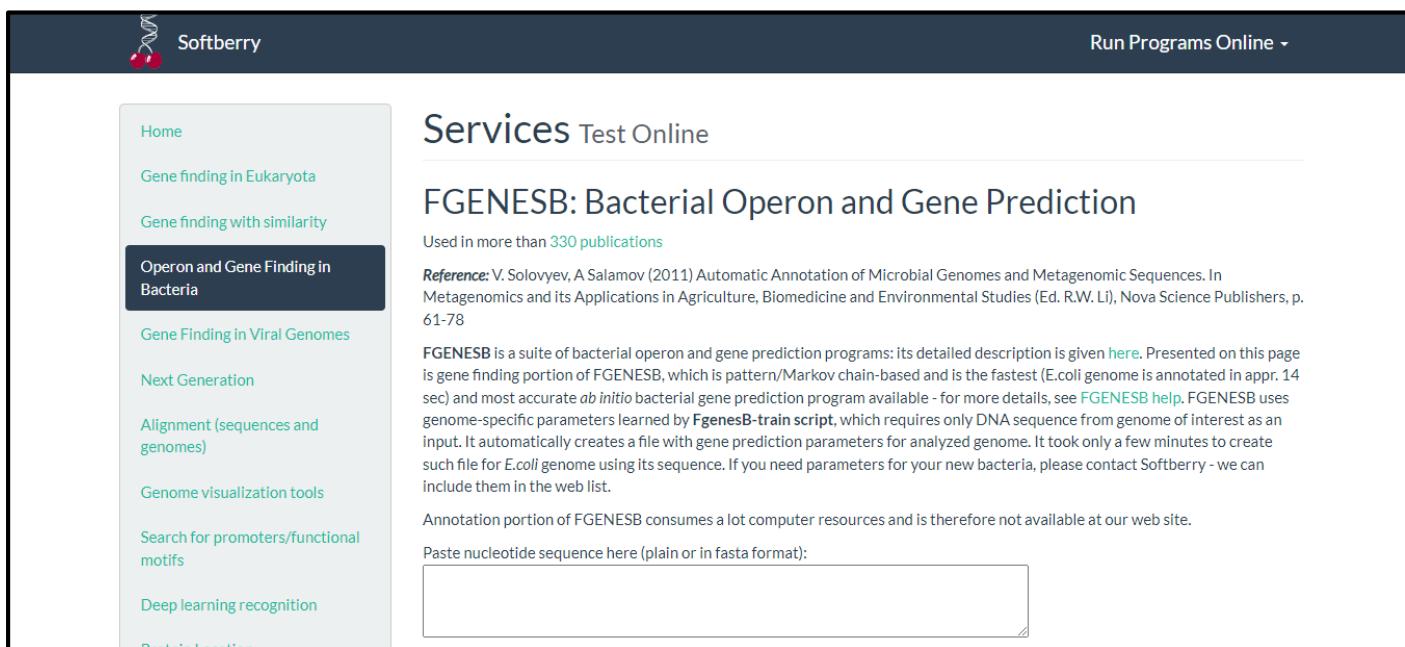


The screenshot shows the Softberry Services Test Online interface. The main content area is titled "Bacterial Promoter, Operon and Gene Finding". Below the title, a reference is cited: "Solovyev V, Salamov A. (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p.61-78." A list of tools is provided, each with a link to "Help" and "Example".

- fgenesB - Pattern/Markov chain-based bacterial operon and gene prediction [Help] [Example]
- BPROM - Prediction of bacterial promoters [Help] [Example]
- AbSplit - Separating archaea and bacterial genomic sequences [Help] [Example]
- FindTerm - Finding Terminators in bacterial genomes [Help] [Example]
- Visualization of FgenesB annotation using CGView [Example]
- Bacterial GenomeSequence Explorer - Visualization of Bacterial genomes information [Help]
- All bacteria genomes annotations
- General scheme of bacterial genome annotation -(automatic pipeline - Fgenesb\_annotator)
- Human Microbiome gene prediction
- Softberry Microbiome Annotation Database

The sidebar on the left lists various services, with "Operon and Gene Finding in Bacteria" highlighted in a dark blue box. The footer contains the URL "www.softberry.com".

**Fig4. Tools for bacterial promoter, operon and gene finding**



The screenshot shows the Softberry Services Test Online interface. The main content area is titled "FGENESB: Bacterial Operon and Gene Prediction". A note states "Used in more than 330 publications". A reference is cited: "Reference: V. Solovyev, A Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78". The text describes FGENESB as a suite of bacterial operon and gene prediction programs, mentioning its speed and accuracy. It also notes that the annotation portion consumes a lot of computer resources. A text input field is provided for pasting nucleotide sequences.

The sidebar on the left lists various services, with "Operon and Gene Finding in Bacteria" highlighted in a dark blue box. The footer contains the URL "www.softberry.com".

**Fig5. Homepage for FGENESB**



- [Home](#)
- [Gene finding in Eukaryota](#)
- [Gene finding with similarity](#)
- Operon and Gene Finding in Bacteria**
- [Gene Finding in Viral Genomes](#)
- [Next Generation](#)
- [Alignment \(sequences and genomes\)](#)
- [Genome visualization tools](#)
- [Search for promoters/functional motifs](#)
- [Deep learning recognition](#)
- [Protein Location](#)

## Services Test Online

### FGENESB: Bacterial Operon and Gene Prediction

Used in more than 330 publications

**Reference:** V. Solovyev, A. Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78

FGENESB is a suite of bacterial operon and gene prediction programs: its detailed description is given [here](#). Presented on this page is gene finding portion of FGENESB, which is pattern/Markov chain-based and is the fastest (*E.coli* genome is annotated in appr. 14 sec) and most accurate *ab initio* bacterial gene prediction program available - for more details, see [FGENESB help](#). FGENESB uses genome-specific parameters learned by *FgenesB-train* script, which requires only DNA sequence from genome of interest as an input. It automatically creates a file with gene prediction parameters for analyzed genome. It took only a few minutes to create such file for *E.coli* genome using its sequence. If you need parameters for your new bacteria, please contact Softberry - we can include them in the web list.

Annotation portion of FGENESB consumes a lot computer resources and is therefore not available at our web site.

Paste nucleotide sequence here (plain or in fasta format):

```
>NZ_CM003348.1 Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome shotgun sequence
GGCGCGGAAATGCCGAAGTGTCCACGGTTATCGCGCTGATAGAGTTTCGG
```

**Fig6. Search for nucleotide FASTA sequence**

Prediction of potential genes in microbial genomes						
Time: Tue Jan 1 00:00:00 2005						
Seq name: NZ_CM003348.1 Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome						
Length of sequence - 4104 bp						
Number of predicted genes - 9						
Number of transcription units - 5, operons - 2						
N	Tu/Op	conserved	s	start	End	Score
1	1 Tu 1	.	-	CDS	184 -	891 295
2	2 Tu 1	.	+	CDS	904 -	1044 97
3	3 Tu 1	.	-	CDS	1136 -	1360 78
4	4 Op 1	.	+	CDS	1439 -	1819 183
5	4 Op 2	.	+	CDS	1804 -	2433 313
6	4 Op 3	.	+	CDS	2517 -	3104 296
7	4 Op 4	.	+	CDS	3151 -	3405 226
8	5 Op 1	.	-	CDS	3661 -	3846 176
9	5 Op 2	.	-	CDS	3849 -	4094 184
Predicted protein(s):						
>GENE 1 184 - 891 295 235 aa, chain -						
MLFDLDYEGAGLAWEADNNLMPAWAAINRENGGAHLAYALSAPVLTAEYGGRKALRYLA						
ALEAAKYAKLRLGDVGFWSLITKNEPHWLTLRGVPAIRGYDLEYLADFVLDKFKYI						
GRSNVEAVGLSRSRCTVENLVSRAWHKNVLAFKQQGYTVQGWLKEVHYQCMRVNGDFVPM						
WEKEVRCISKSIAANWWYKEDIAASNRFFSELQAHRSNLSLRKTTINAGRTKITEL						
>GENE 2 904 - 1044 97 46 aa, chain +						
MRRIDLDDVAAFDDGGSVRQHRRFFEIFAWLWVDTFLQEEVGLGV						
>GENE 3 1136 - 1360 78 74 aa, chain -						
MRNSGKNTAHLSELIVSCFPQNAALLAGLGRAAALQALPRLLLASLNGLENFEIRTPIG						
LLSDGYFDIVVSS						
>GENE 4 1439 - 1819 183 126 aa, chain +						
LTGSSAAEKRTKQLIIRVSPTEFETLIRQKTHPNLARYIRERVLEDGKASDKKTVKFQFP						
PEVVRVLAGMGNLNQIAKALNTAAKGVTGLGNVEALKATTELAALERSLNSLRDFLAKEK						
NGWQSQ						
>GENE 5 1804 - 2433 313 209 aa, chain +						
MVAPEMTVOFFENRGKGGGSGPTDVITLGKDRDRREPRTIRGDPEETABLTINSSDVAKKVTAG						

**Fig7. Result for predicted bacterial operon and genes**

## RESULT:

Nucleotide FASTA sequence for *Neisseria gonorrhoeae* strain NG\_869 plasmid pNG869\_3 with length of 4104bps 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. 97-111.

2. *Neisseria gonorrhoeae - an overview / ScienceDirect Topics.* (n.d.). [Www.sciencedirect.com](http://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](https://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1)
4. *Softberry Home Page.* (n.d.). [Www.softberry.com](http://www.softberry.com). Retrieved March 18, 2022, from <http://www.softberry.com/>
5. *FGENESB - Bacterial Operon and Gene Prediction.* (n.d.). [Www.softberry.com](http://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](http://www.softberry.com). Retrieved March 18, 2022, from <http://www.softberry.com/cgi-bin/programs/gfindb/fgenesb.pl>

## WEBLEM 8d

### FGENES

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

#### AIM:

To predict exon signals in Protease using FGENES tool.

#### INTRODUCTION:

Proteolytic enzymes (proteases) are enzymes that break down protein. These enzymes are made by animals, plants, fungi, and bacteria. Proteolytic enzymes break down proteins in the body or on the skin. This might help with digestion or with the breakdown of proteins involved in swelling and pain. Human protease exon signals can be 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:

16. Open homepage for softberry. (URL: <http://www.softberry.com/>)
17. Under Gene for Eukaryotes select FGENES. (URL: <http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind>)
18. Retrieve nucleotide FASTA sequence for protease from GenBank.
19. Process the FASTA sequence on FGENES.
20. Observe and interpret the results.

#### OBSERVATION:

**Homo sapiens neutral protease alpha subunit gene, complete cds**

GenBank: AH001431.2

[FASTA](#) [Graphics](#)

**Go to:** [Nucleotide](#)

**LOCUS** AH001431 3298 bp DNA linear PRI 10-JUN-2016

**DEFINITION** Homo sapiens neutral protease alpha subunit gene, complete cds.

**ACCESSION** AH001431 M31501 M31502 M31503 M31504 M31505 M31506 M31507 M31508 M31509 M31510 M31511

**VERSION** AH001431.2

**KEYWORDS** neutral protease.

**SOURCE** Homo sapiens (human)

**ORGANISM** *Homo sapiens*  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

**REFERENCE** 1 (bases 1 to 3298)

**AUTHORS** Miyake,S., Emori,Y. and Suzuki,K.

**TITLE** Gene organization of the small subunit of human calcium-activated neutral protease

**JOURNAL** Nucleic Acids Res. 14 (22), 8805-8817 (1986)

**PUBLMED** 3024120

**COMMENT** On or before Jun 10, 2016 this sequence version replaced [M31501.1](#), [M31502.1](#), [M31503.1](#), [M31504.1](#), [M31505.1](#), [M31506.1](#), [M31507.1](#), [M31508.1](#), [M31509.1](#), [M31510.1](#), [M31511.1](#), [AH001431.1](#).

**FEATURES** Location/Qualifiers

source 1 3298

**Articles about the CAPNS1 gene**  
Dual proteome-scale networks reveal cell-specific remodeling of the human inte [Cell. 2021]  
Overexpression of Capns1 Predicts Poor Prognosis and Correlates with Tur [Urol Int. 2021]  
Circular RNA ABCB10 promotes cell proliferation and invasion, but inhibits ap [Mol Med Rep. 2021]

**See all...**

**Reference sequence information**  
RefSeq alternative splicing

**Fig1. GenBank result for Human Protease**

### Fig3. Homepage for Softberry

Softberry

Run Programs Online ▾

Home

Gene finding in Eukaryota

Gene finding with similarity

Operon and Gene Finding in Bacteria

Gene Finding in Viral Genomes

Next Generation

Alignment (sequences and genomes)

Genome visualization tools

Search for promoters/functional motifs

Deep learning recognition

Protein Location

RNA

Services Test Online

## Gene Finding: Gene models construction, splice sites, protein-coding exons

Total 506 genome-specific parameters are available for genefinders of FGENESH suite  
[The programs usage in Scientific publications](#)

FGENESH is the fastest and most accurate *ab initio* gene prediction program available - for more details, see [FGENESH help](#). Its variants that use similarity information: FGENESH+ (similar protein), FGENESH\_C (similar cDNA), FGENESH-2 (two homologous genomic sequences) greatly improve accuracy of gene prediction when such similarity information is available. These programs can be accessed [here](#).

To find genes in Bacterial sequences click [here](#).

Our two best gene finders cannot be accessed at our site due to computing resources limitations. These two are FGENESH++ (automated version of FGENESH+) and FGENESH++C, which maps known mRNA/EST sequences from RefSeq and then performs FGENESH++-like gene prediction, resulting in fully automatic annotation of quality similar to that of manual annotation.

FGENES, FGENES-M, FGENESH\_GC and SPLM can be used on human sequences only. BESTORF and Fsplice can be used with 296 organisms sequences. SPL can be used for human, Drosophila, nematode, S.cerevisiae, and dicots.

[FGENESH](#) - HMM-based gene structure prediction (multiple genes, both chains) [\[Help\]](#) [\[Example\]](#)

[FGENES](#) - Pattern based human gene structure prediction (multiple genes, both chains) [\[Help\]](#) [\[Example\]](#)

[FGENES-M](#) - Pattern-based human multiple variants of gene structure prediction) [\[Help\]](#) [\[Example\]](#)

**Fig4. Tools for gene finding in Eukaryota**

Softberry

Run Programs Online ▾

Home

Gene finding in Eukaryota

Gene finding with similarity

Operon and Gene Finding in Bacteria

Gene Finding in Viral Genomes

Next Generation

Alignment (sequences and genomes)

Genome visualization tools

Search for promoters/functional motifs

Deep learning recognition

Protein Location

Services Test Online

## FGENES

Pattern based human gene structure prediction (multiple genes, both chains)

Paste nucleotide sequence here:

Alternatively, load a local file with sequence in Fasta format:

Local file name:  No file chosen

[\[Help\]](#) [\[Example\]](#)

Return to page with other programs of group: [Gene finding](#)

Your use of Softberry programs signifies that you accept [Terms of Use](#)

**Fig5. Homepage for FGENES**

[Home](#)[Gene finding in Eukaryota](#)[Gene finding with similarity](#)[Operon and Gene Finding in Bacteria](#)[Gene Finding in Viral Genomes](#)[Next Generation](#)[Alignment \(sequences and genomes\)](#)[Genome visualization tools](#)[Search for promoters/functional motifs](#)[Deep learning recognition](#)[Protein Location](#)

## Services Test Online

### FGENES

Pattern based human gene structure prediction (multiple genes, both chains)

Paste nucleotide sequence here:

```
>AH001431.2 Homo sapiens neutral protease alpha subunit gene, complete cds
CTGCAGAGGGCCCGTGCAGTCCCTAGTGAGCGGACCGAAAACGCCACCT
GGAAGGATATTGGCAT
```

Alternatively, load a local file with sequence in Fasta format:

Local file name:

 No file chosen [\[Help\]](#)[\[Example\]](#)Return to page with other programs of group: [Gene finding](#)Your use of Softberry programs signifies that you accept [Terms of Use](#)**Fig6. Search for protease nucleotide FASTA sequence**

Show picture of predicted genes in PDF file

softBerry

FGENES 1.6 Prediction of multiple genes in genomic DNA

Time: 13:37:11 Date: Fri Mar 18 2022

Seq name: >AH001431.2 Homo sapiens neutral protease alpha subunit gene

Length of sequence: 3298 GC content: 0.57 Zone: 4

Number of predicted genes: 1 In +chain: 1 In -chain: 0

Number of predicted exons: 10 In +chain: 10 In -chain: 0

Positions of predicted genes and exons:

G	Str	Feature	Start	End	Weight	ORF-start	ORF-end
1	+	1 CDS	277	451	7.07	277	450
1	+	2 CDS	706	823	3.66	708	821
1	+	3 CDS	984	1017	3.03	985	1017
1	+	4 CDS	1167	1256	3.91	1167	1256
1	+	5 CDS	1416	1473	3.58	1416	1472
1	+	6 CDS	1855	1910	2.09	1857	1910
1	+	7 CDS	2052	2130	2.29	2052	2129
1	+	8 CDS	2287	2403	2.90	2289	2402
1	+	9 CDS	2548	2606	3.87	2550	2606
1	+	10 CDS	2763	2789	2.71	2763	2786

Predicted proteins:

```
>FGENES 1.6 >AH001431.2 Hom 1 Multiexon gene 277 - 2789 270 a Ch+
MRNVRFGSRTARATAGPRCSVRESCGLSHRPRPEPDAPGGPGAVERPTPRTPDVCGGL
ISGAGGGGGGGGGGGGGGGGGGGTAMRILGGVISAISEAAQYNEPEPPPRTHYSNIEA
NESEEVRQFRRLFAQLAGDDMVEVSATELMNILNKVVTRRKLGFEEFKYLWNNN1KRWQAIY
KQFDTDRSGTICSSSELPGAFEARGFHLNEHLYNMIIIRYSDESGNMDFDNFISCLVRLDA
MFRAFKSLDKDGTGQIQVNQIWEWLQLTMYS
```

© 1999 - 2022 [www.softberry.com](http://www.softberry.com)

**Fig7. Result for predicted exon signals**

## RESULT:

Nucleotide FASTA sequence for Homo sapiens neutral protease of length 3298bps 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 understanding 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.). [https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20\(proteases\)%20are%20enzymes](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](http://www.softberry.com/). Retrieved March 18, 2022, from <http://www.softberry.com/>
5. *FGENES - pattern-based gene structure prediction.* (n.d.). [Www.softberry.com](http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind). 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](http://www.softberry.com/cgi-bin/programs/gfind/fgenes.pl). Retrieved March 18, 2022, from <http://www.softberry.com/cgi-bin/programs/gfind/fgenes.pl>

## 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:

*Neisseria gonorrhoeae* infects primarily columnar epithelium, because stratified squamous epithelium is relatively resistant to invasion. Mucosal invasion by gonococci results in a local inflammatory response that produces a purulent exudate consisting of PMNs, serum, and desquamated epithelium. ORF region in *Neisseria gonorrhoeae* can be recognised using ORF finder tool.

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:

NCBI Resources ▾ How To ▾ Sign in to NCBI

Nucleotide Nucleotide ▾ Advanced Search Help

GenBank ▾ Send to: ▾ Change region shown ▾

Customize view ▾ Analyze this sequence ▾ Run BLAST

Go to: ▾

**Neisseria gonorrhoeae strain NJ189125 chromosome, complete genome**

NCBI Reference Sequence: NZ\_CP041586.1

[FASTA](#) [Graphics](#)

**LOCUS** NZ\_CP041586 2185626 bp **DNA** **circular** CON 04-MAR-2022

**DEFINITION** *Neisseria gonorrhoeae* strain NJ189125 chromosome, complete genome.

**ACCESSION** NZ\_CP041586

**VERSION** NZ\_CP041586.1

**DBLINK** BioProject: PRJNA224116  
BioSample: SAMN12252303  
Assembly: GCF\_007107365.1

**KEYWORDS** RefSeq.

**SOURCE** *Neisseria gonorrhoeae*

**ORGANISM** *Neisseria gonorrhoeae*

Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales; Neisseriaceae; *Neisseria*.

**REFERENCE** 1 (bases 1 to 2185626)

**AUTHORS** Zhao,Y., Le,W., Lou,X., Genco,C.A., Rice,P. and Su,X.

**TITLE** Identification of multidrug resistant *Neisseria gonorrhoeae* FC428 clone in Nanjing, China

**JOURNAL** Unpublished

**REFERENCE** 2 (bases 1 to 2185626)

**AUTHORS** Zhao,Y., Le,W., Lou,X., Genco,C.A., Rice,P. and Su,X.

**Related information**

- Assembly
- BioProject
- BioSample
- Protein
- Taxonomy
- Components (Core)
- Genome
- Identical GenBank Sequence
- PubMed (Unweighted)

Fig1. Result for *Neisseria gonorrhoeae* in GenBank

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

FASTA ▾ Send to: ▾ Change region shown ▾

**Neisseria gonorrhoeae strain NJ189125 chromosome, complete genome**

NCBI Reference Sequence: NZ\_CP041586.1

GenBank Graphics

>NZ\_CP041586.1 Neisseria gonorrhoeae strain NJ189125 chromosome, complete genome

GATTTCGCCGTCGCCAGCGGACAAATAGGAAGTCCGGTATCGACGGCGAGGACGGCGG  
 TTGAATCGGCTTCGATGGTGTGGTCTCGTGGTCTAGACGGCATTATAGTGAATCGGCTTCGCTG  
 CCGTGCCTGCTCTAGGGCTATGGCGCAAAATCGCGTCAACGGTAAATTATCGTGGCTCG  
 GGCATTTTCAAATACTCTGTTGCGGATTCGTGAACCTTTCCCTATCTCACGGT  
 TGCCGAGCCAGCGATTCTGGCGGGTCCGCAAAGGGCGAGGGGTAAGTGCCTTCTTCGCG  
 GTAGCGCTGTCGCCAAAATGACGGCTGTGATTTCATCTGCTTGAAGTGCCTGGAGCG  
 ACATCTGCTCGGGTGTTCAGGGTTCGATTTCCGCTGCTCTGGCGGGAATCTATGGTAT  
 AGGTAGTTCGCGGATCGTGGAGCTGAATCTCGGCTTTCGCGTCAAGCG  
 TTGTTGAAGCGCGATGTCGCCAAATCGCTGACAAAGGAAGGACGTGTTTATCAGGCTGCG  
 ATTTTGCGGGTGTGATTCTGCCAAATCGCTGACAAAGGAAGGACGTGTTTATCAGGCTGCG  
 GAAATTGCGCTTCGCACTGGTGTGATTGCGCTGCTGCACTTCGATTTTGCAAGTGCAGCG  
 GATTTGCGCTTCGCACTGGTGTGATTGCGCTGCTGCACTTCGATTTTGCAAGTGCAGCG  
 GCACCTTGCGCCAGGGTCAGTTGGCTGGTGGGAAGGCTTCAATGTTAGGATTTCAAC  
 CTTTGCTTATGGTCGAGCGCGCGGTTAATGCATCGCAAGCCCGTGCCTGCGATCGGCG  
 ACCGCTCCGCCCTCGCTGCAAGGGTCAAGAACGGCCGGCGTCAAGAAAGGCAAGCAAG  
 GTAGTCGGTTCAAGGTTACTCTAGTCATACAGAGAATAGATAATATAAACGTTTGTATGGT  
 ATCTTTGCGTTCAAGGTTACTCTAGTCATACAGAGAAGCAAATCAATGCGCTGCAAGCG  
 TCAGACGGCATTTGTTACAGGCAACCTGTTATGGTCAATTGCGCTGACCTTCGCAACCG  
 CTGGCATTTCGCAACAGAAAGGGTGGCTGACCTTCGCAACCGAACGCAAGATAAAC  
 GGTCGAGGGCAGATTGCTTCATGCTTCAATGGTTGGCCAAAGTATTGGCGGATCAAAGTC  
 GTTTGGCAAGGGTAGCGCGTACCGCGCGGTTGAAGCGAGGCGCAGGTCGATCGATGTT  
 CGCACCGGTGTTGATGCCCTGACGATTCTCACCGCAACGCCGTAAGGTTTGCCTGATATT  
 CGCGATATTGCG

Analyze this sequence ▾

Run BLAST

Pick Primers

Highlight Sequence Features

Related information ▾

Assembly

BioProject

BioSample

Protein

Taxonomy

Components (Core)

Genome

Identical GenBank Sequence

Detailed ▾

**Fig2. Nucleotide FASTA sequence for *Neisseria gonorrhoeae***

Open Reading Frame 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 subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button):

- NC\_011604 *Salmonella enterica* plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM\_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

TGTGCGCCCTGTTGGGGCGGGGACGGGTTCTGGGGGGAACTTTTCCAGCGATCAT  
 CCAGCGCGATGGCAAAACGCCGTGAGCTTTAAATCCATATAGGTTCTGAATGGTGG  
 CCGCAGGTGCGTGTCCGATAGACGGGAATAAGGGCGGCTTCATACCGACGGCGTCCGC  
 CGAAGGTTAGACGCCATTATAGAACCGATGGGAAATAAAGGAAAGCGTCAATTGAATA  
 TCGGGTCAGGGACGGGTGCTGTCGTCCGCCGAAGGGTGGCAGCGTCAATGGCTTGA  
 GCCGGAGGCCTTGA|

From: To:

Choose Search Parameters



**Fig3. Search for Nucleotide FASTA sequence**

CGAAGGTTAGACGGCATTATAGAACCGATGGAAATAAAGGAAAGCGTCAATTGAA  
TCGGGTCAGGAACGGGTCTGTCCGTCCGCCAAAGGGTGCAGCGTCAATCGCTTGAT  
GCGGAGGCCTTGAA

From:  To:

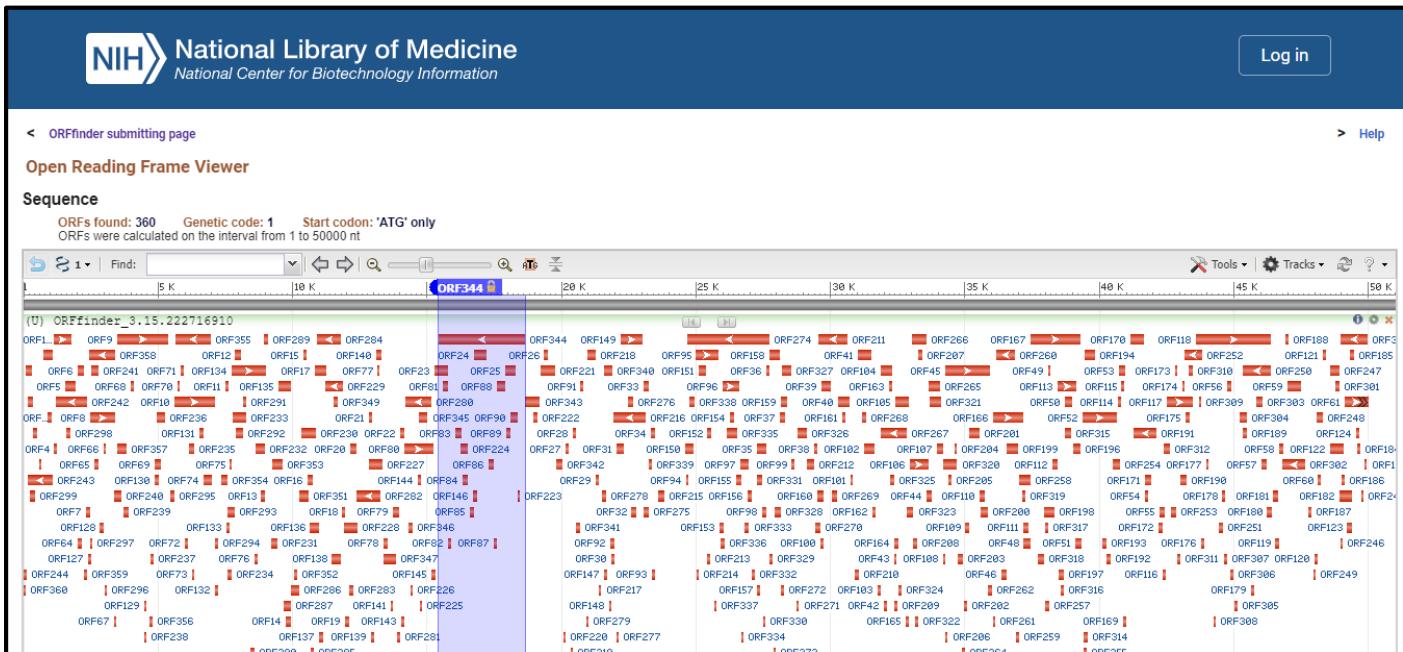
**Choose Search Parameters**

Minimal ORF length (nt):  Genetic code:  ORF start codon to use:  "ATG" only  "ATG" and alternative initiation codons  Any sense codon Ignore nested ORFs:

Start Search / Clear

**Submit** **Clear**

**Fig4. Search parameters**



**Fig5. Result for recognised ORFs**

1: 1..51K (51,000 nt) Tracks shown: 2/3  
Six-frame translation...

ORF344 (1071 aa) [Display ORF as...](#) [Mark](#)

```
>1c1|ORF344
MPKRTDLKSILIIAGAPIVIGQACEFDYSGAQACKALREEGYKVILVNSN
PATIMTDPEHADVTYIEPIIMQTVKIAKERPDAILPTMGGTALCAL
DLARNGVLAKYIVVELGATEADIKAEADRGRKFAEMEKIGLSPCKSFVCH
TMNEALAAQEQVQGPTLIRPSFTMGGSGGAYADEFALACERGPDPSP
THEEATGQSVLQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
PAQGTTTAAQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQ
HPPVVRSSALAAKATGPTIAKVAAMALAVGFTTDLQELNDITTGRTTAFCEP
SIDYVVTKIPRFAPEVFPAAQDLRTQWKSVSGEVMAMGRTIQQSFQALKR
GLETGLCGFIPPSDEKAEIRRELAPGPEPMFLVADAFRAGFPTEEHEI
CAZDPWFLAQIEDUNKEEKSVSQGQLODQYALRLLRKGFSDKRLAQAL
LNVSEKEVREHRYALKLHPVVKRVTCAAEFATEAYLTYEEECESRP
```

ORF344 Marked set ( 0 ) [SmartBLAST best hit titles...](#) [BLAST](#)

[BLAST Database:](#) [UniProtKB/Swiss-Prot \(swissprot\)](#)

Mark subset... Marked: 0 Download marked set as [Protein FASTA](#)

Label	Strand	Frame	Start	Stop	Length (nt   aa)
ORF344	-	1	18648	15433	3216   1071
ORF274	-	3	27748	24683	3066   1021
ORF118	+	2	43601	46390	2790   929
ORF9	+	1	3481	5394	1914   637
ORF167	+	3	37425	39299	1875   624
ORF45	+	1	34270	35946	1677   558
ORF10	+	1	5596	7119	1524   507
ORF355	-	1	6969	5641	1329   442
ORF52	+	1	39361	40656	1296   431
ORF134	+	3	7743	9023	1281   426
ORF216	-	2	23132	21933	1200   399

Go back to the submitting page...

FOLLOW NCBI

**Fig6. Results for tracks information**

## RESULT:

After submitting nucleotide FASTA sequence for complete genome of *Neisseria gonorrhoeae* 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](http://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](https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP041586.1)
3. *Home - ORFfinder - NCBI*. (2019). [Nih.gov](http://www.ncbi.nlm.nih.gov/orffinder/). Retrieved March 18, 2022, from <https://www.ncbi.nlm.nih.gov/orffinder/>

## 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 IGV.

#### 1. UCSC Genome Browser

The University of California Santa Cruz (UCSC) Genome Browser ([genome.ucsc.edu](http://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.

The database contains a large collection of gene prediction annotations. The RefSeq Genes annotation is computed at UCSC from RefSeq mRNAs that have been aligned against the genome using BLAT and then

filtered. The protein-coding portion of the mRNA is mapped to the genome and blocks separated by gaps of 5 or fewer bases are merged into exons.

Several cross-species homology annotations are computed by UCSC and its collaborators, including BLAT and BLASTZ alignments of human and mouse genomes against the target genome, mRNA and EST alignments from other species, and synteny annotations. In cross-species annotations, RepeatMasker and Tandem Repeats Finder are first applied to the target genome to mask repetitive elements before the alignments are generated.

The database includes several high-level map annotations. Some examples of these include the Chromosome Bands annotation that approximates the locations of Giemsa-stained chromosomes bands at an 800-band resolution, the Sequence-Tagged Site (STS) Markers annotation showing the positions of several markers—many of which were used in constructing genome-wide genetic and physical maps—and the Fluorescent *In Situ* Hybridization (FISH) Clones annotation showing the location of FISH-mapped BAC clones from the BAC Resource consortium on the genome.

## 2. Ensembl Genome 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.

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. The ID History Converter converts Ensembl IDs from a previous release into their current equivalents, while the Assembly Converter maps genomic coordinates from one version of a genome assembly to another. The Variant Effect Predictor predicts the functional consequences of a set of known and/or novel variants. 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.

## 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.

Thus, 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. 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. 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:

5. Xiong, J. (2008). *Genome Mapping, Assembly, and Comparison. Essential bioinformatics*. Cambridge: Cambridge University Press. 243.
6. 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
7. Karolchik, D. (2003). *The UCSC Genome Browser Database*. , 31(1), 51–54. doi:10.1093/nar/gkg129
8. Birney, E. (2004). *An Overview of Ensembl. Genome Research*, 14(5), 925–928. doi:10.1101/gr.1860604
9. 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
10. 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>

## 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 (GRCh38/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. Observe and interpret the results.

#### OBSERVATION:

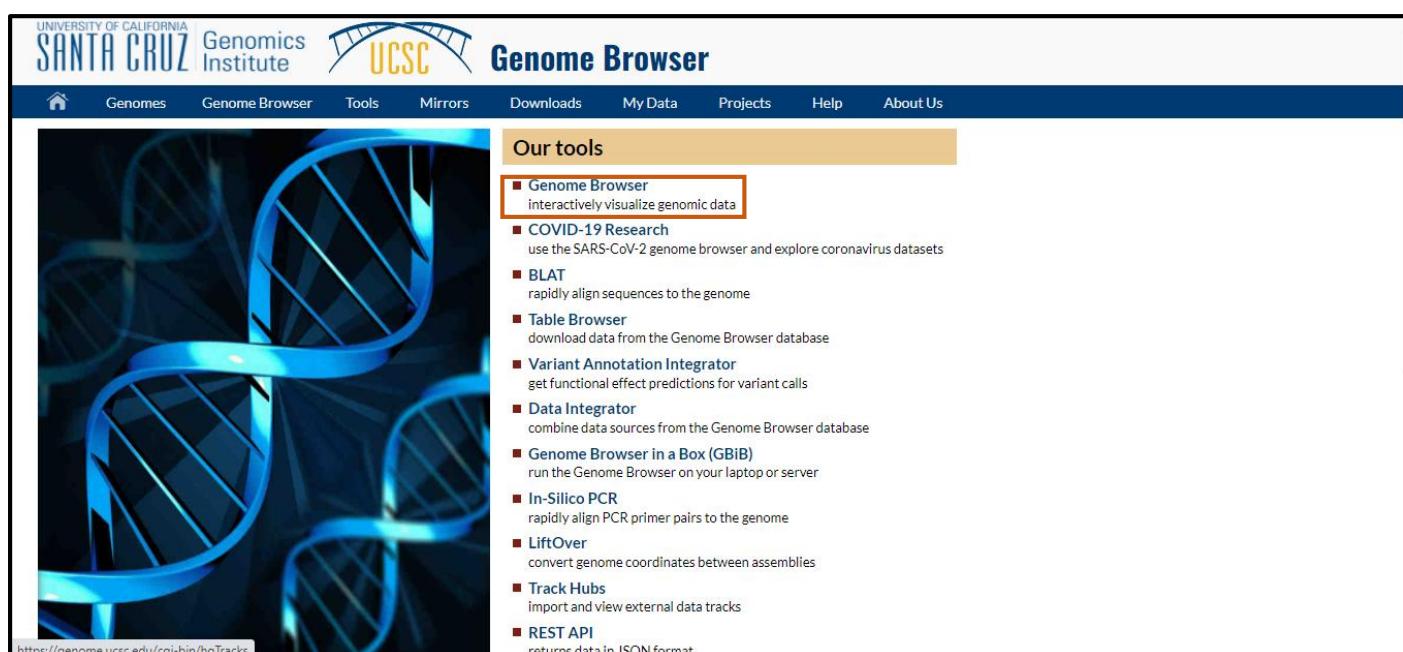


Fig1. Homepage for UCSE browser

UNIVERSITY OF CALIFORNIA  
SANTA CRUZ Genomics Institute 

Genome Browser Gateway

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

**Browse/Select Species**

**Find Position**

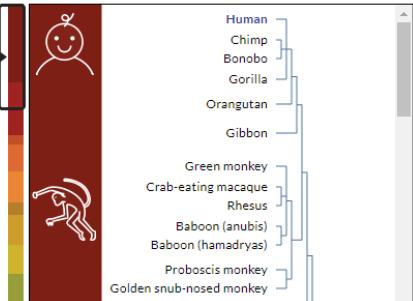
Human Assembly  
Dec. 2013 (GRCh38/hg38)

Position/Search Term  
Enter position, gene symbol or search terms  
Current position: chrX:15,560,138-15,602,945

GO 

**Human Genome Browser - hg38 assembly** 

**REPRESENTED SPECIES**



UCSC Genome Browser assembly ID: hg38  
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p13 (GCA\_000001405.28)  
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 13  
Assembly accession: GCA\_000001405.28  
NCBI Genome ID: 51 (Homo sapiens (human))  
NCBI Assembly ID: GCF\_000001405.39 (GRCh38.p13, GCA\_000001405.28)  
BioProject ID: PRJNA31257

  
Homo sapiens  
(Graphic courtesy of CBSE)

Search the assembly:

■ By position or search term: Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. [More information](#), including sample queries.

Fig2. Genome browser gateway

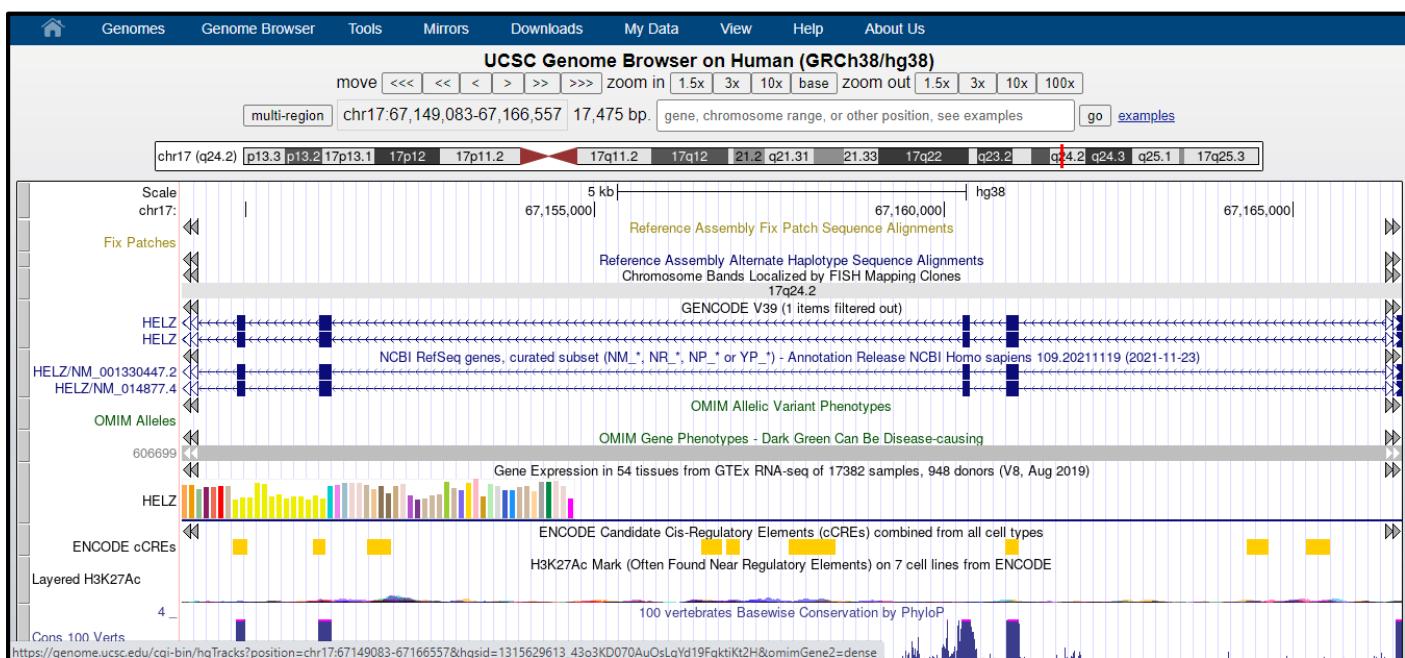


Fig3. UCSC genome browser on human (GRCh38/hg38)

**Mapping and Sequencing**

- Base Position: **dense**
- Clone Ends: **hide**
- GRC Incident: **hide**
- RefSeq Acc: **hide**
- P13 Fix Patches: **dense**
- Exome Probesets: **hide**
- Hg19 Diff: **hide**
- Restr Enzymes: **hide**
- P13 Alt Haplotypes Assembly: **full**
- FISH Clones: **hide**
- INSDC: **hide**
- STS Markers: **hide**
- Centromeres: **hide**
- GC Percent: **hide**
- Liftover & ReMap: **hide**
- LRG Regions: **hide**
- Chromosome Band: **pack**
- GRC Contigs: **hide**
- Mappability: **hide**

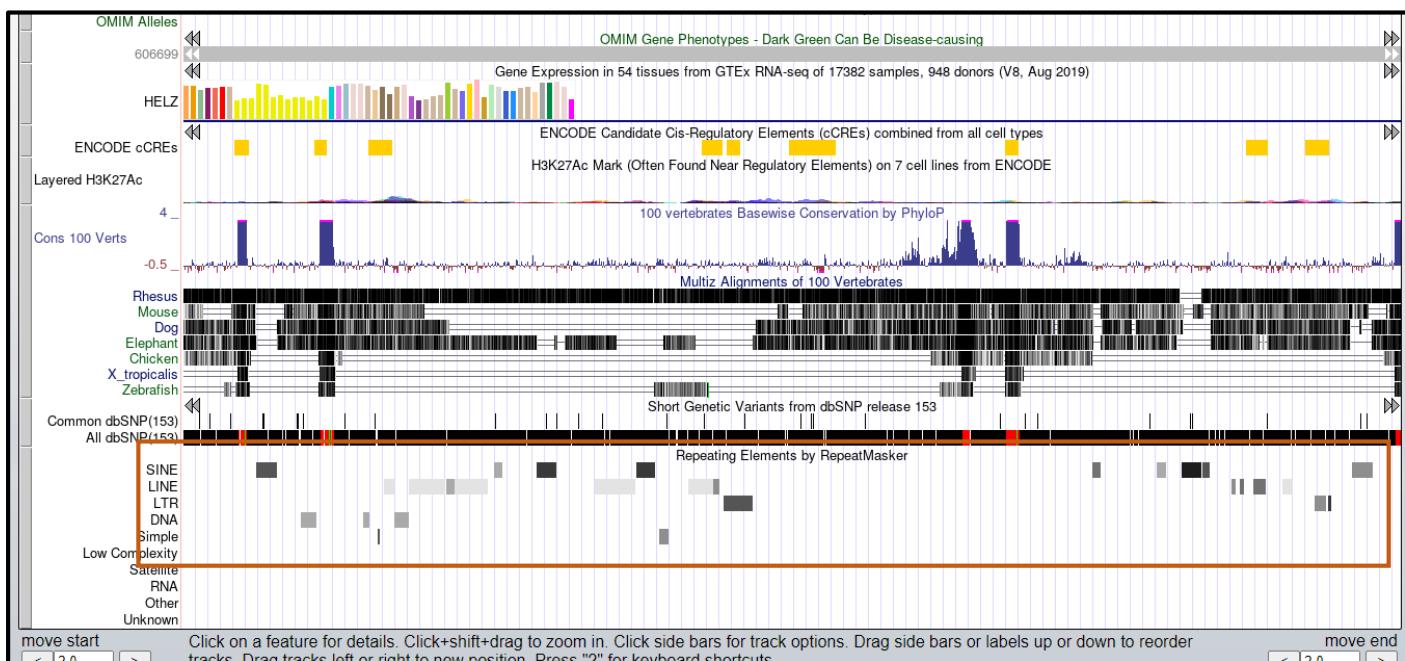
**Genes and Gene Predictions**

- GENCODE V39: **full**
- LRG Transcripts: **hide**
- Other RefSeq: **hide**
- UniProt: **hide**
- NCBI RefSeq: **pack**
- GENCODE: **hide**
- Pfam in GENCODE: **hide**
- CCDS: **hide**
- Non-coding RNA: **hide**
- Prediction Archive: **hide**
- RetroGenes V9: **hide**
- Old UCSC Genes: **hide**
- IKMC Genes Mapped: **hide**
- ORFeome Clones: **hide**
- TransMap V5: **hide**
- UCSC Alt Events: **hide**

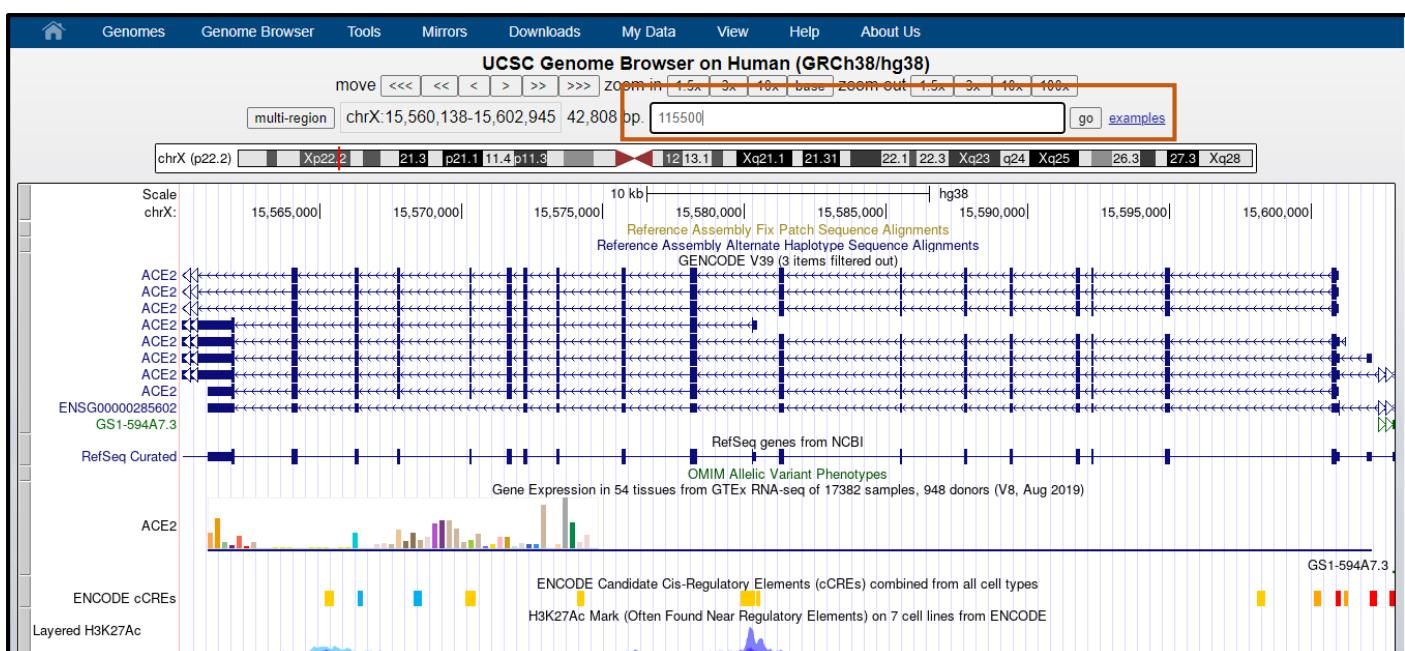
**Phenotype and Literature**

- OMIM Alleles: **dense**
- Coriell CNVs: **hide**
- HGMD Variants: **hide**
- CADD: **hide**
- COSMIC Regions: **hide**
- LOVD Variants: **hide**
- Cancer Gene Expr: **hide**
- Development Delay: **hide**
- OMIM Cyto Loci: **hide**
- OMIM Genes: **pack**
- ClinGen: **hide**
- Gene Interactions: **hide**
- GeneReviews: **hide**
- Orphanet: **hide**
- ClinVar Variants: **hide**
- GWAS Catalog: **hide**
- REVEL Scores: **hide**

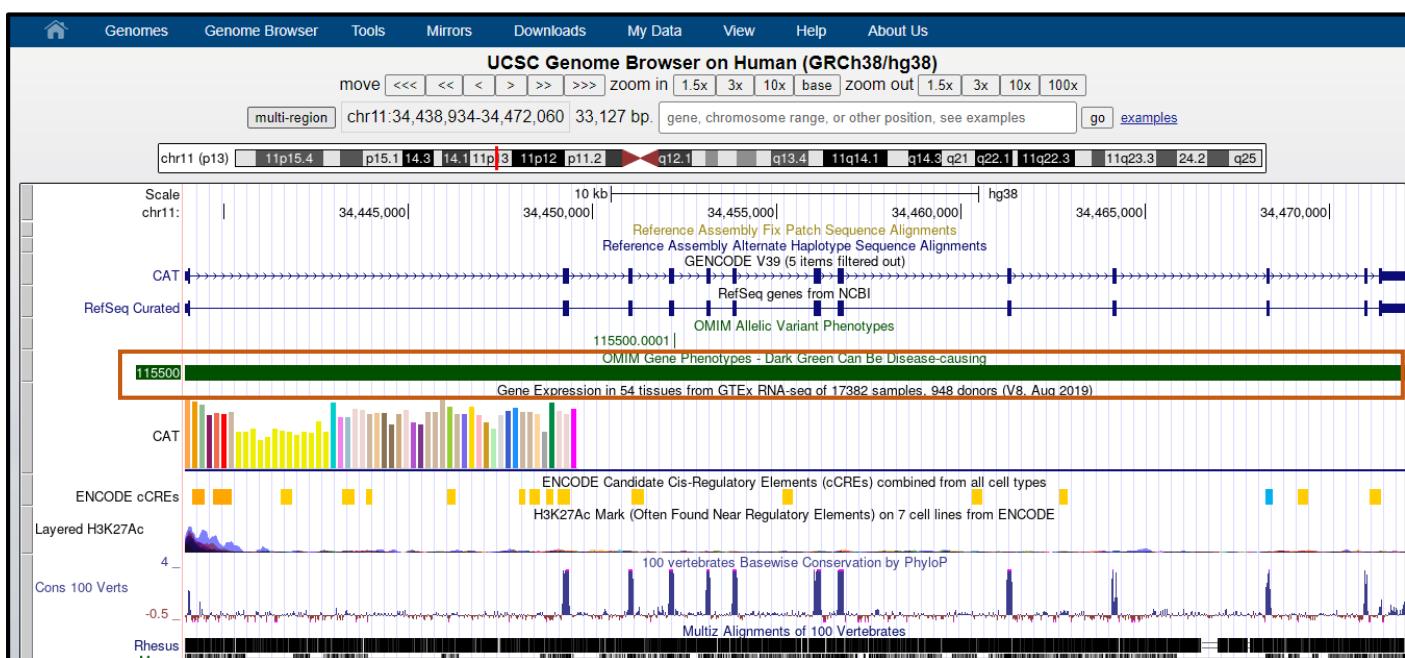
**Fig4. Options for configuration**



**Fig5. Result after configuration**



## Fig6. Navigation by OMIM Id: 115500



### Fig6.1. Result 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

## Fig6.2. 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

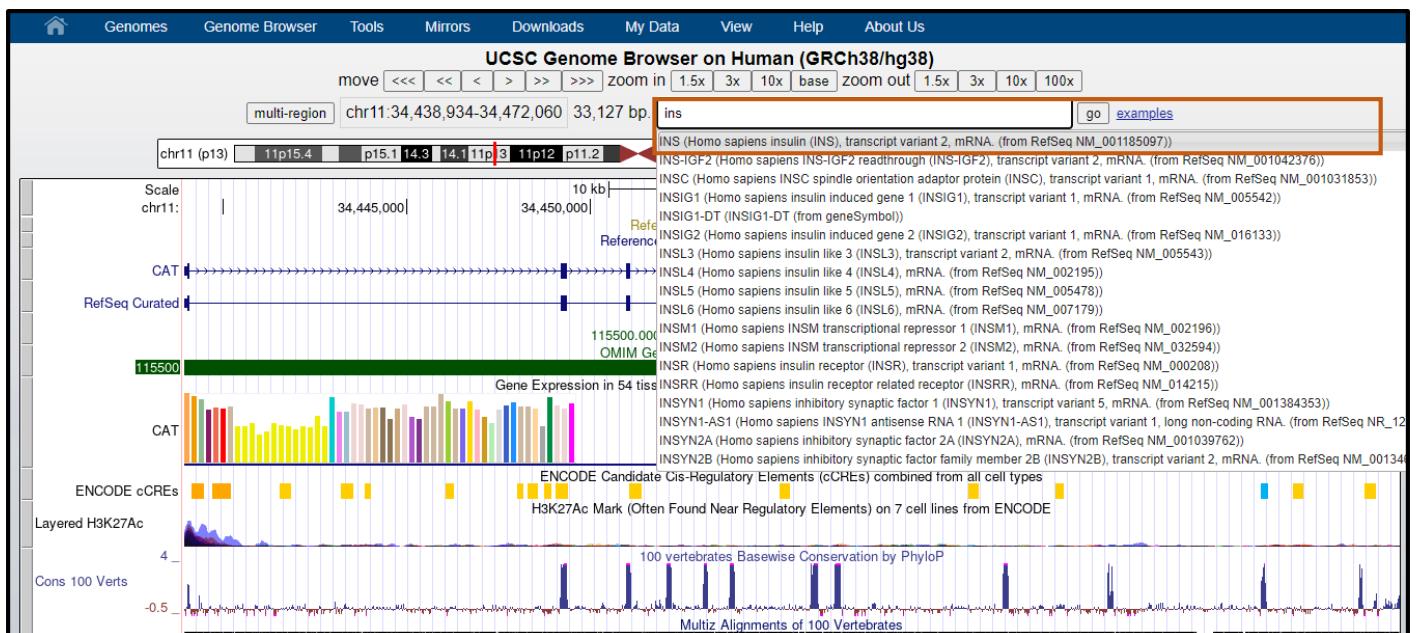
Page Index	Sequence and Links	UniProtKB Comments	MalaCards	CTD	RNA-Seq Expression
Microarray Expression	RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions
Pathways	Other Names	Methods			

Data last updated at UCSC: 2022-01-17 08:30:34

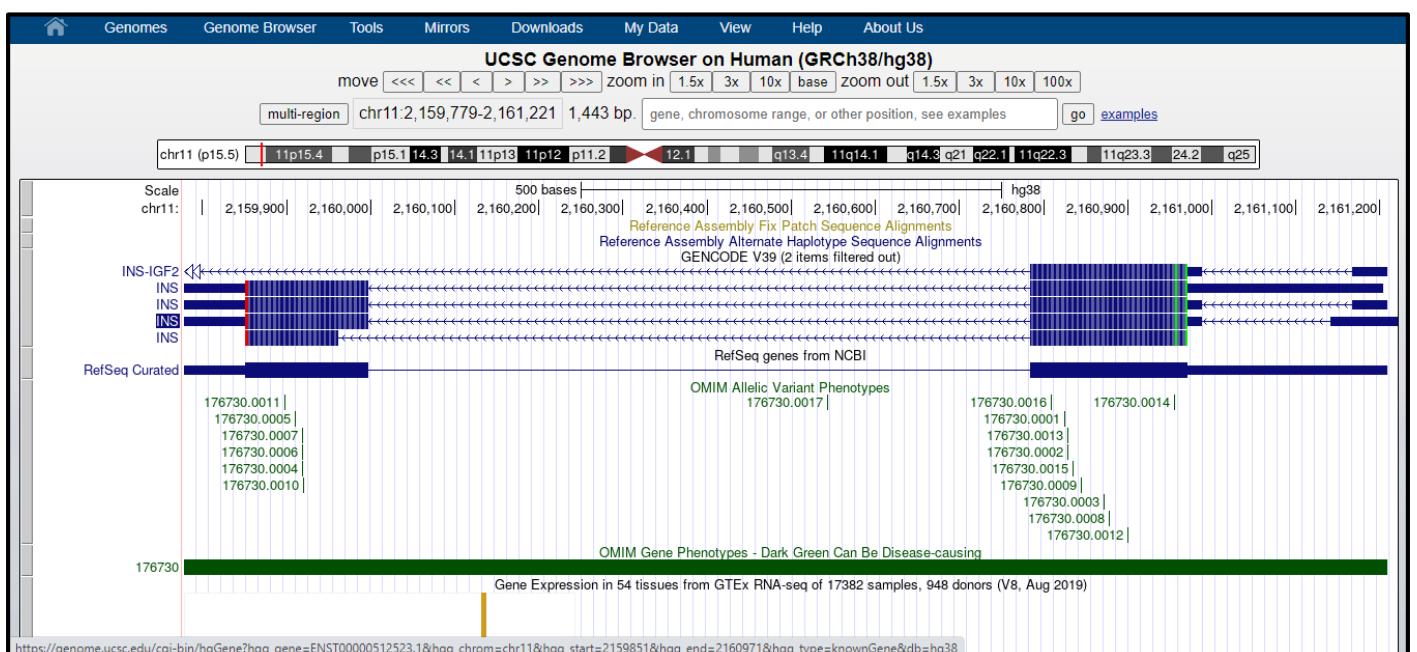
**Sequence and Links to Tools and Databases**

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	

## Fig6.3. Related transcripts information



**Fig7. Navigation by gene name: INS**



**Fig7.1. Result for INS gene**

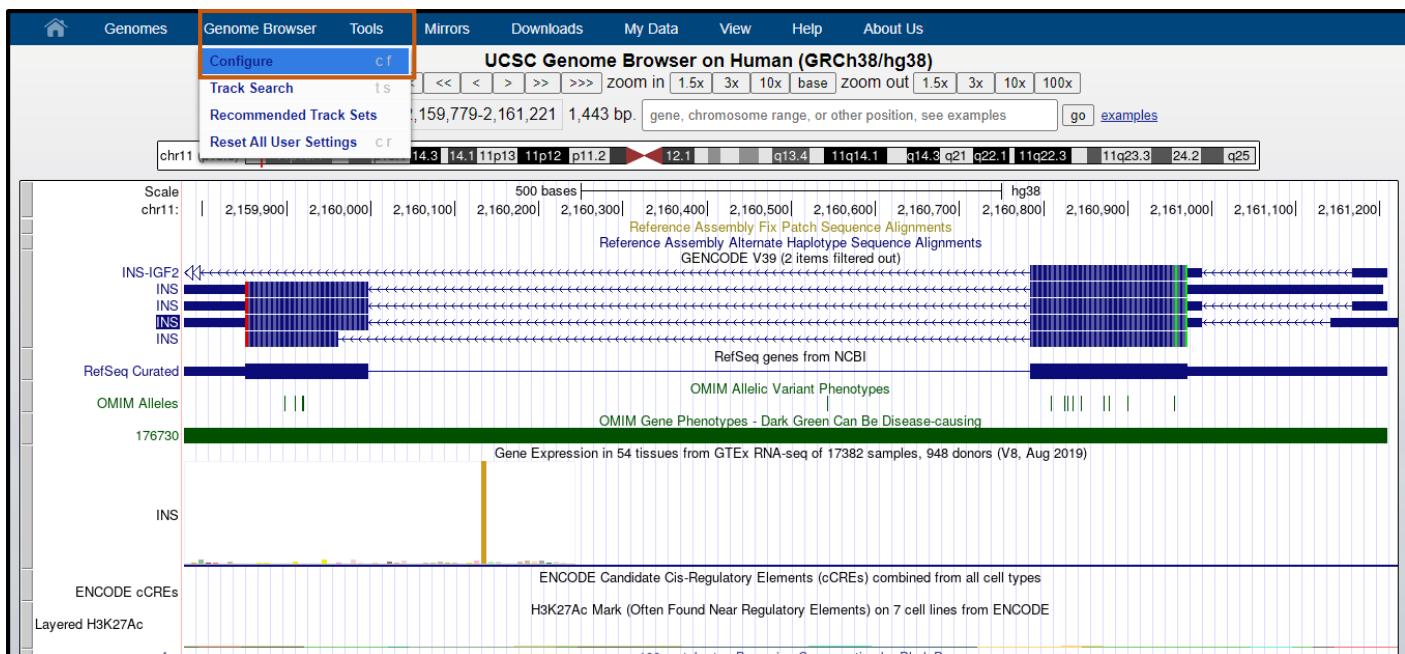


Fig7.2. Option to configure tracks

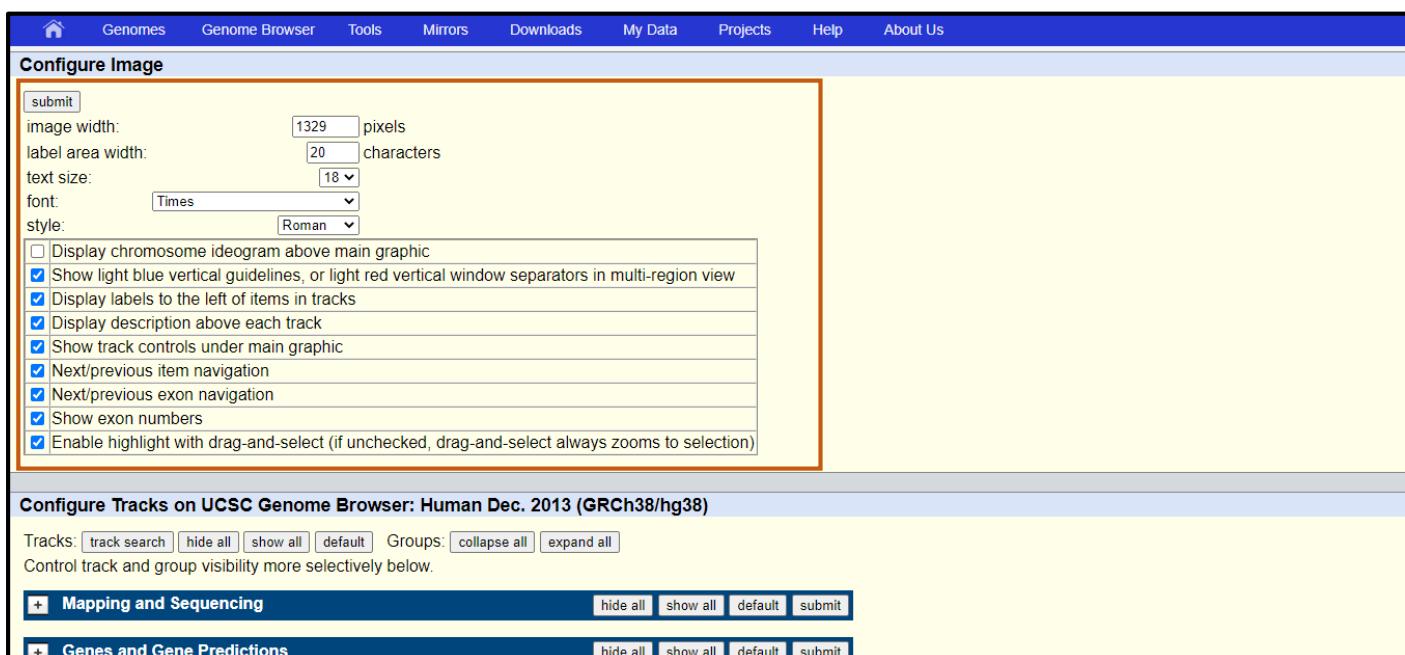


Fig7.3. Configuration settings

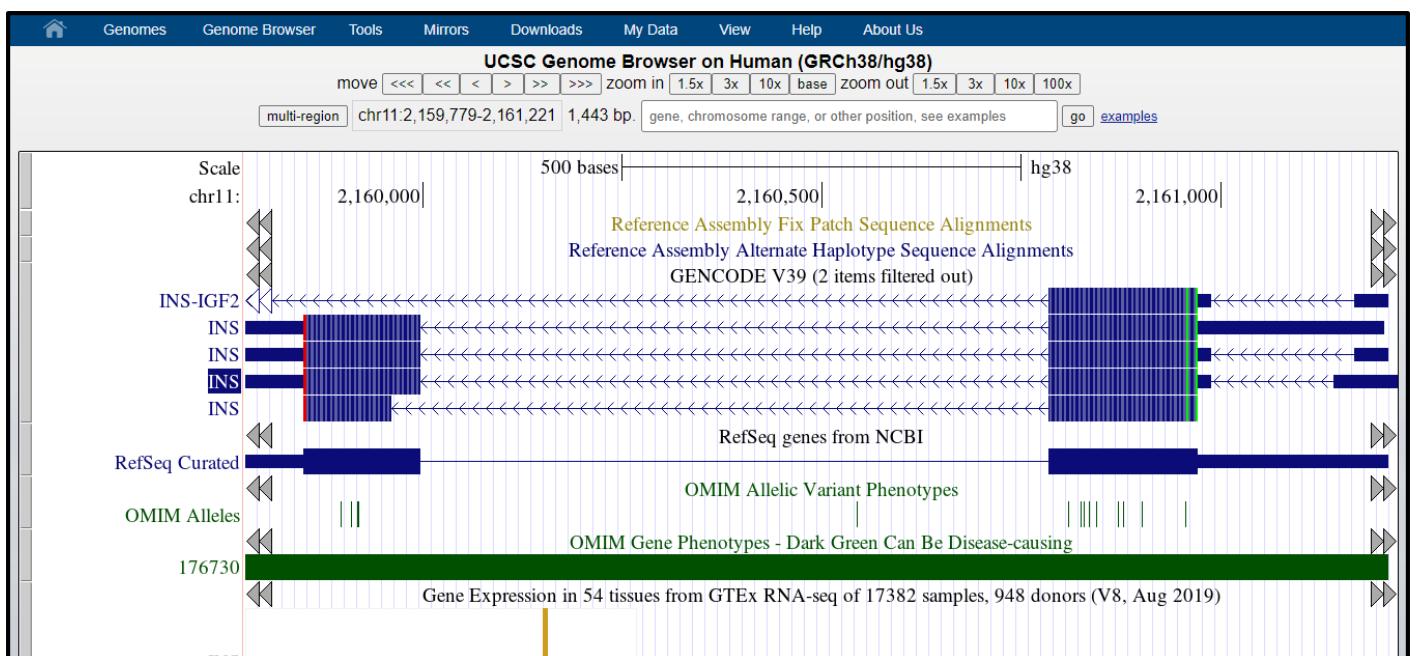


Fig7.4. Configured tracks

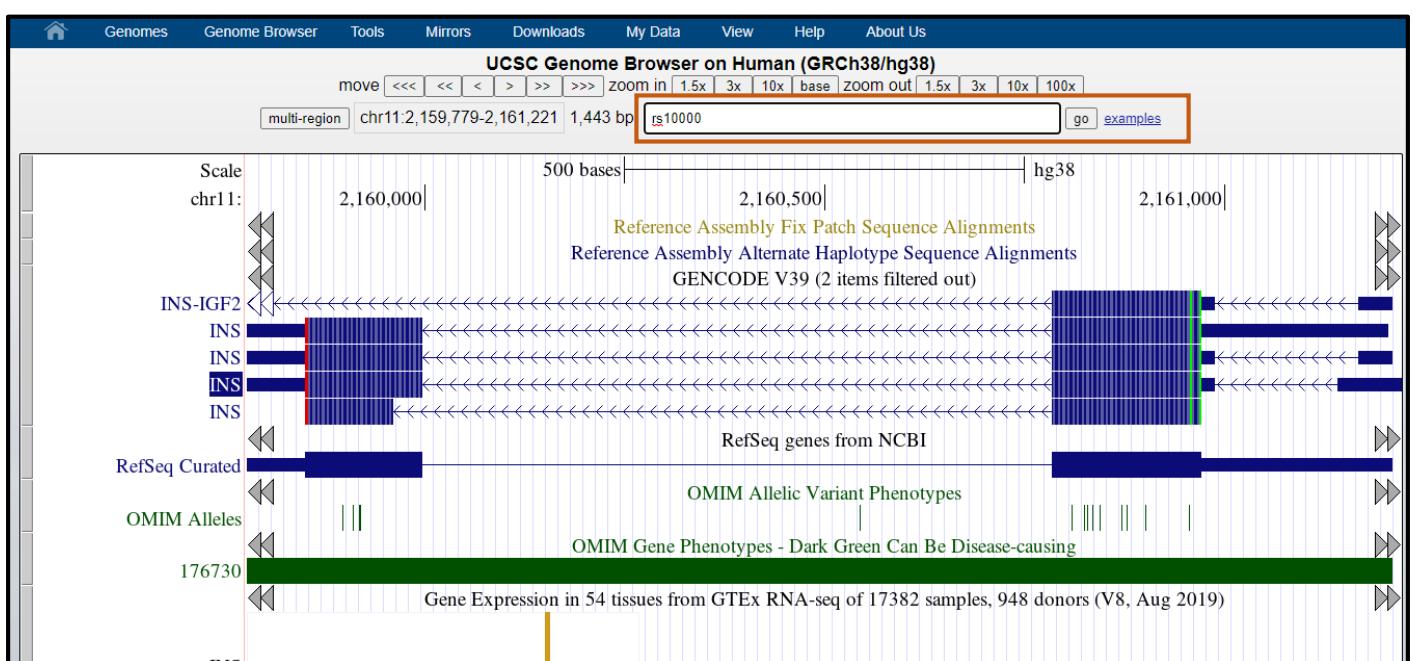


Fig8. Navigation by SNP id: rs10000

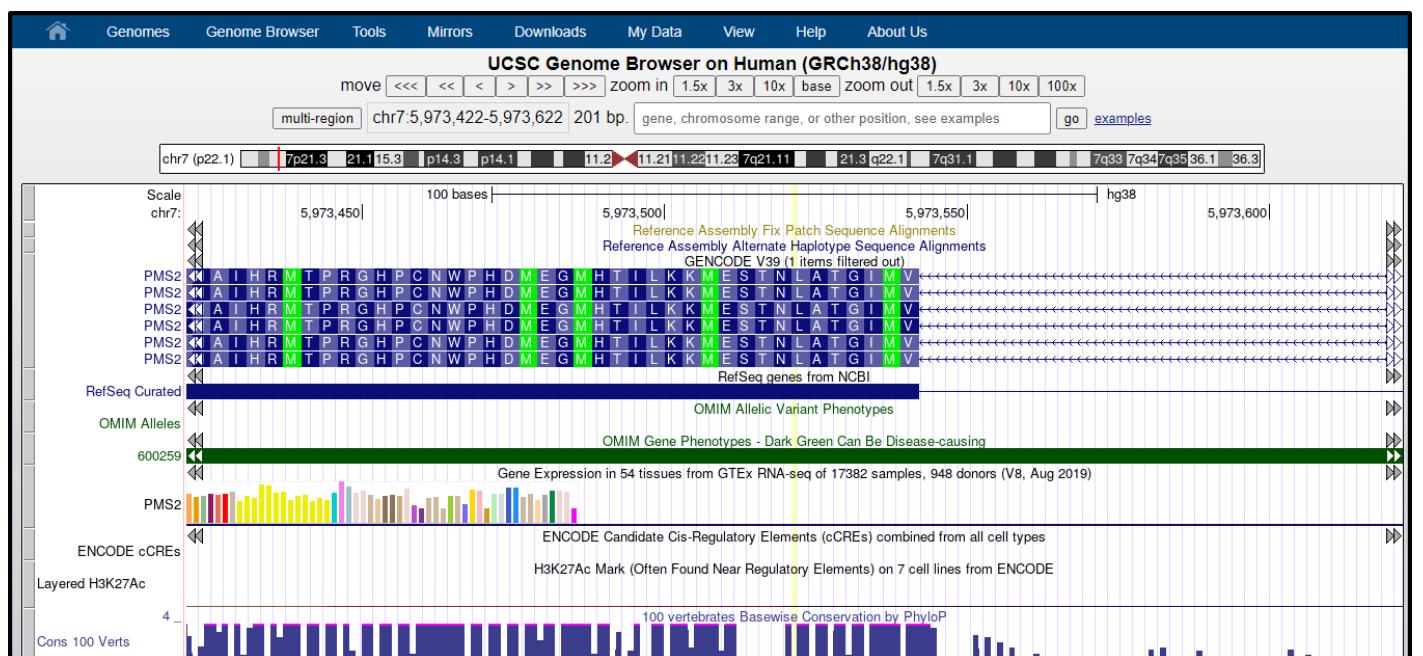


Fig8.1. Result for SNP id: rs10000

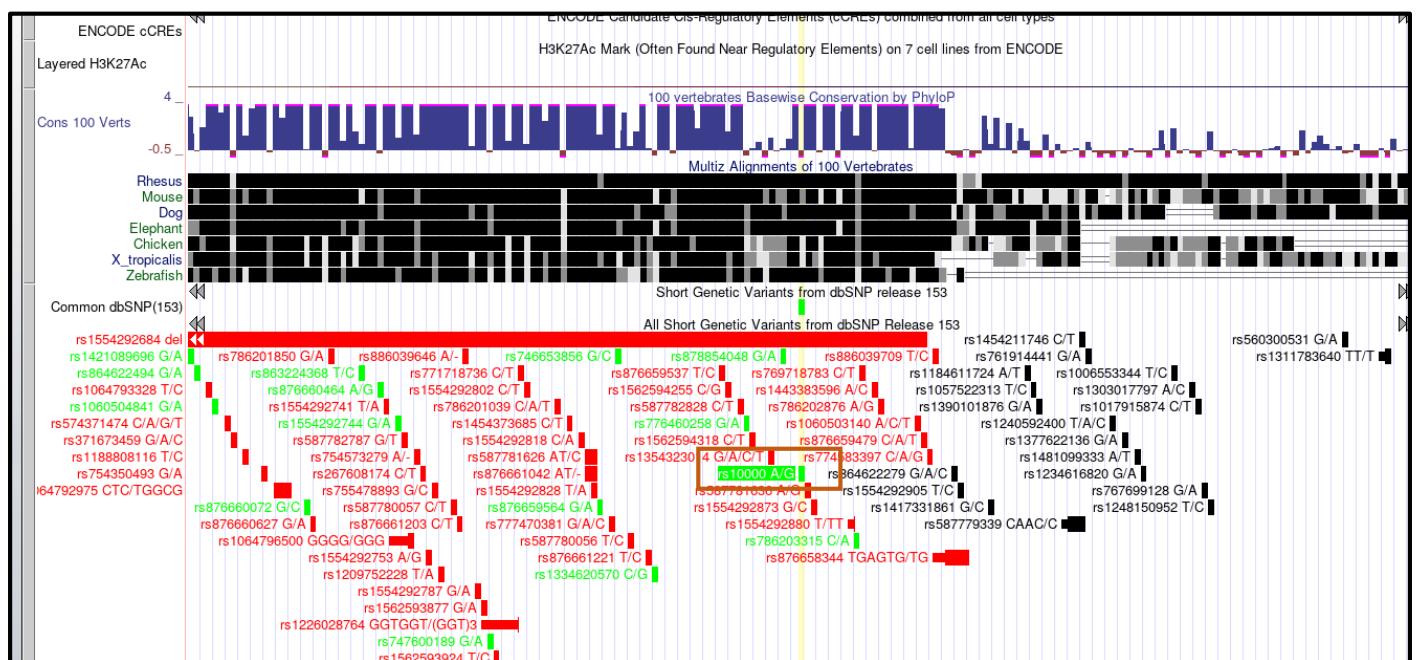


Fig8.2. Result for SNP id: rs10000

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

All Short Genetic Variants from dbSNP Release 153 (rs10000)

dbSNP: rs10000  
 Position: chr7:5973522-5973522  
 Band: 7p22.1  
 Genomic Size: 1  
[View DNA for this feature \(hg38/Human\)](#)

Reference allele: A  
 Alternate allele: G  
 Allele frequency counts:

Allele	1000Genomes	GnomAD_exomes	ExAC	GnomAD
A	4694/5008 (0.937300)	121595/136684 (0.889607)	19069/24180 (0.788627)	11310/12746 (0.887337)
G	314/5008 (0.062700)	15089/136684 (0.110393)	5111/24180 (0.211373)	1436/12746 (0.112663)

Functional effects: synonymous\_variant, coding\_sequence\_variant, nc\_transcript\_variant  
 ClinVar: RCV000030369.5 (benign), RCV000162401.1 (benign), RCV000174851.6 (benign), RCV000627740.1 (benign)  
 Submitted by: 1000GENOMES, CGM\_KYOTO, CNGFU, CORRELAGEN, CSHL, EVA, EVA\_EXAC, EVA\_GENOME\_DK, EVA\_SAMSUNG\_MC, GMI, GNOMAD, JJLAB, KHV\_HUMAN\_GENOMES, OMUKHERJEE\_ADBS, SSMP, SWEGEN, TOPMED, URBANLAB, WEILL\_CORNELL\_DGM  
 Publications in PubMed: PMID16619, PMID10479499, PMID15256438, PMID16472587, PMID20186688, PMID20205264, PMID25741868  
 Variation class/type: snv

Interesting or anomalous conditions noted by UCSC:

- Variant is in ClinVar.
- Variant is in ClinVar with clinical significance of benign and/or likely benign.
- Variant is "common", i.e. has a Minor Allele Frequency of at least 1% in all projects reporting frequencies.

Fig8.3 Description for SNP id: rs10000

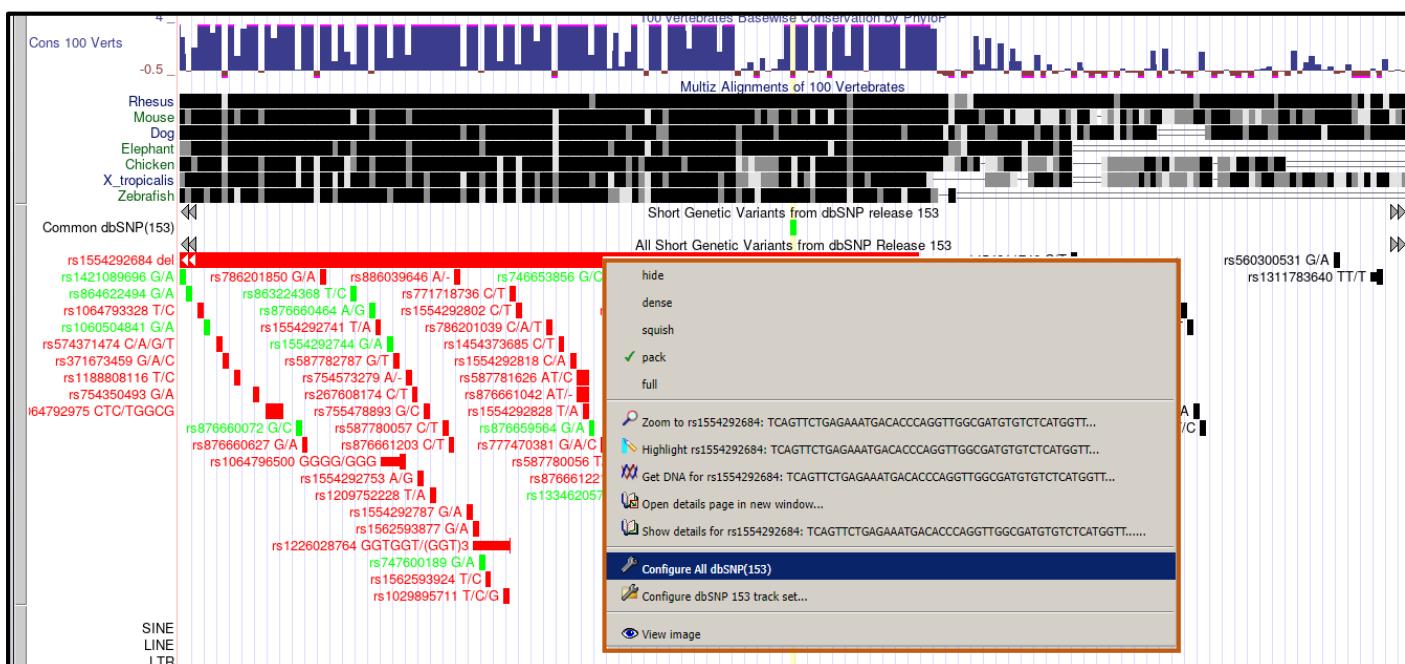


Fig8.4. Configuration option by right click

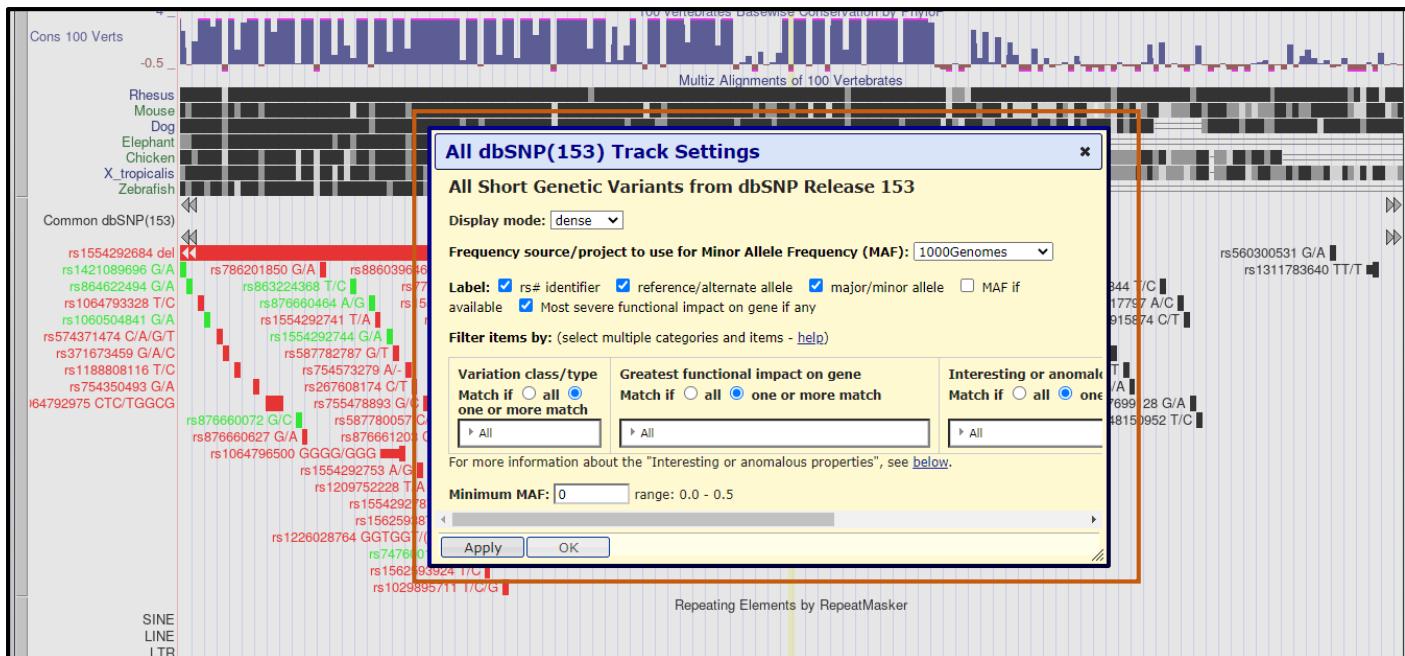


Fig8.5. Configurations applied

8

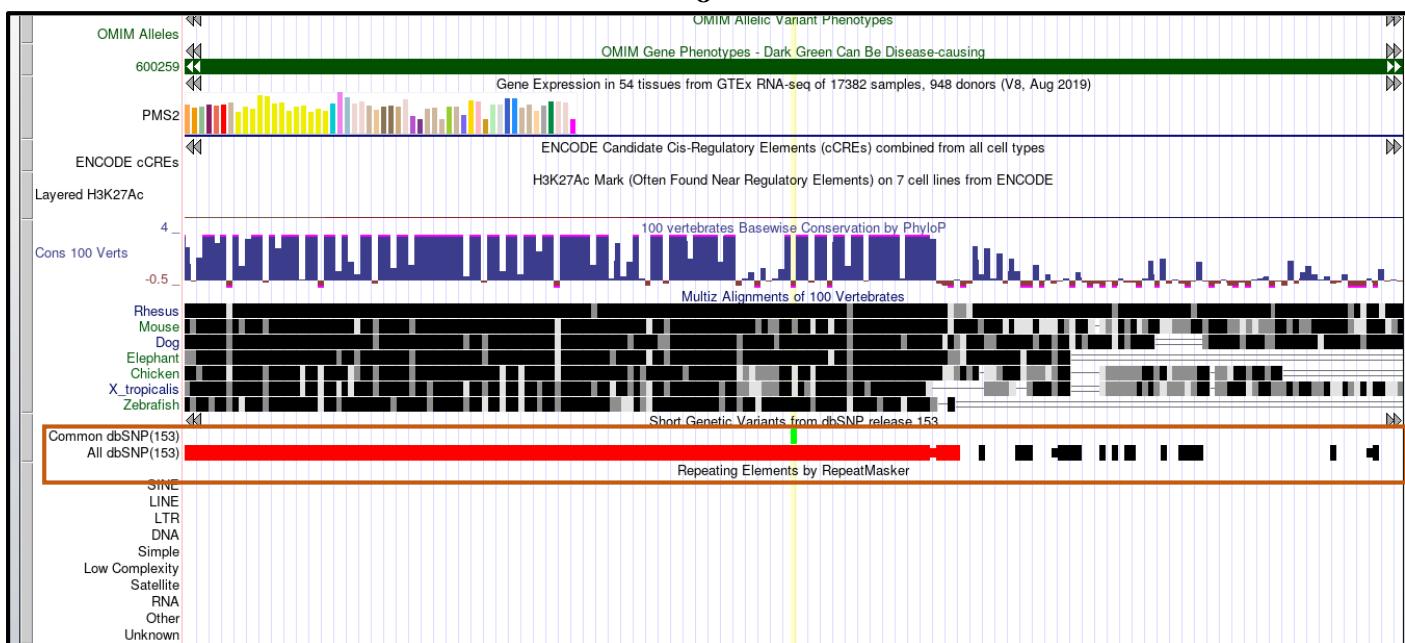
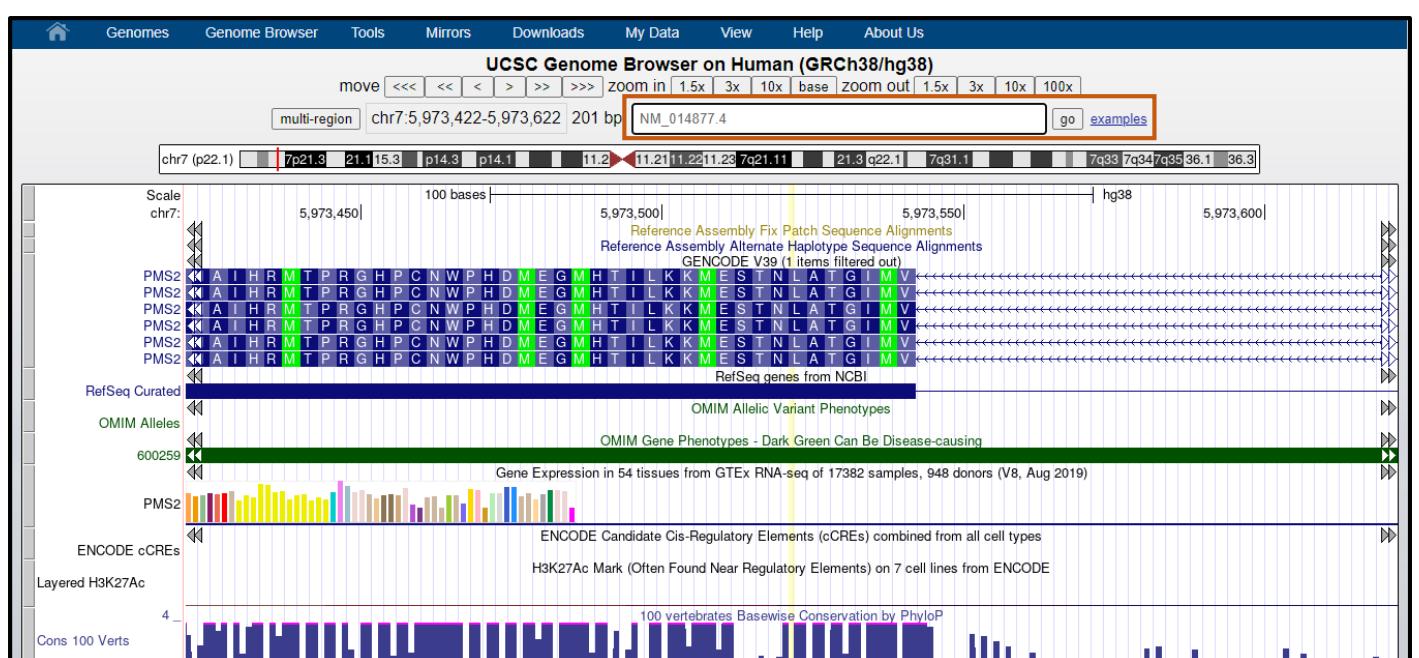


Fig8.6. Result after configuration



**Fig9. Navigation by Ref\_Seq: NM\_014877.1 (HEL gene)**

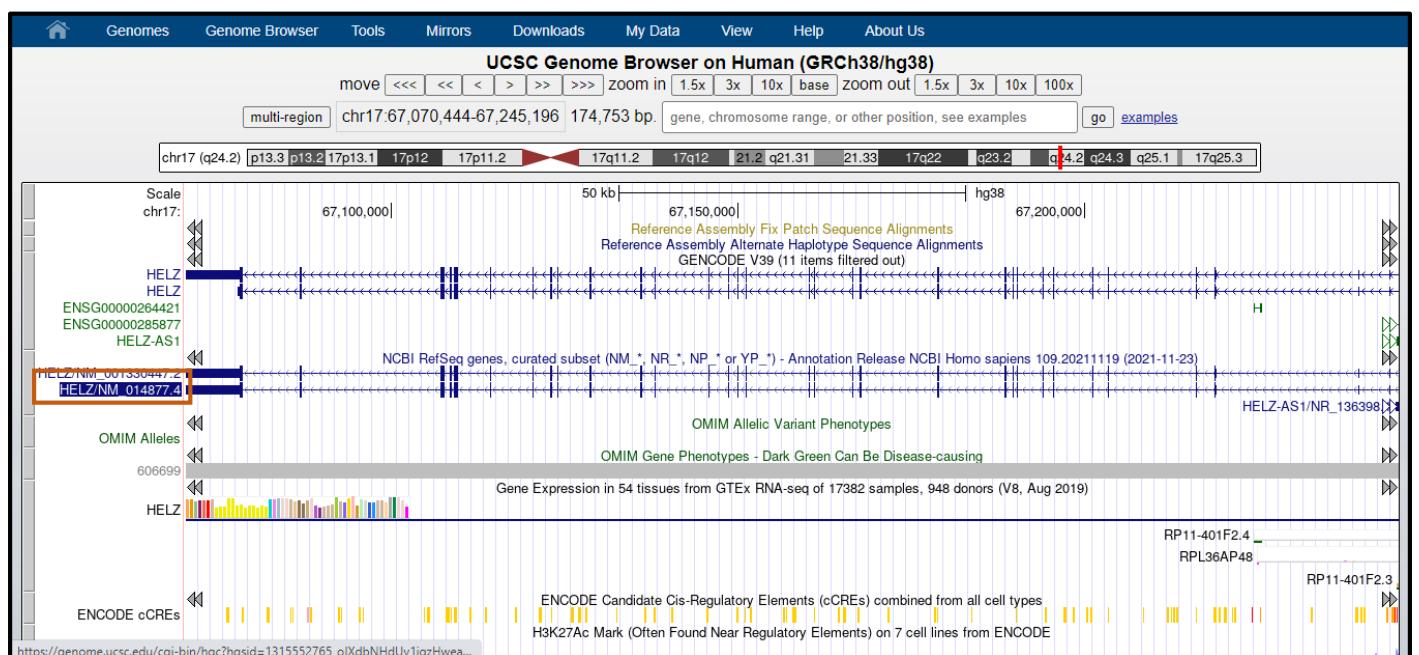


Fig9.1. Result for Ref Seq: NM\_014877.1 (HELZ gene)

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

NCBI RefSeq genes, curated subset (NM\_\*, NR\_\*, NP\_\* or YP\_\*) - NM\_014877.4

## RefSeq Gene HELZ

RefSeq: NM\_014877.4 Status: Validated  
 Description: helicase with zinc finger, transcript variant 1  
 Molecule type: mRNA  
 Source: BestRefSeq  
 Biotype: protein\_coding  
 Synonyms: DHRC,DRHC,HUMORF5  
 Other notes: isoform 1 is encoded by transcript variant 1  
 OMIM: 606699  
 Protein: NP\_055692.3  
 HGNC: 16878  
 Entrez Gene: 9931  
 GeneCards: HELZ  
 AceView: HELZ

Summary of HELZ  
 HELZ is a member of the superfamily I class of RNA helicases. RNA helicases alter the conformation of RNA by unwinding double-stranded regions, thereby altering the biologic activity of the RNA molecule and regulating access to other proteins (Wagner et al., 1999 [PubMed 10471385]). [supplied by OMIM, Mar 2008]. Sequence Note: This RefSeq record was created from transcript and genomic sequence data because no single transcript was available for the full length of the gene. The extent of this transcript is supported by transcript alignments. Evidence Data: Transcript exon combination: SRR1803613.239143.1, SRR1660807.108853.1 [ECO:0000332]; RNAseq introns: mixed/partial sample support SAMEA1965299, SAMEA1966682 [ECO:0000350]. RefSeq Attributes: MANE Ensembl match: ENST00000358691.10/ ENSP00000351524.5; RefSeq Select criteria: based on conservation, expression.

mRNA/Genomic Alignments (NM\_014877.4)

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
---------	------	----------	------------	--------	-------	-----	-------	-------	-----	-------

Fig9.2. Description for Gene HELZ

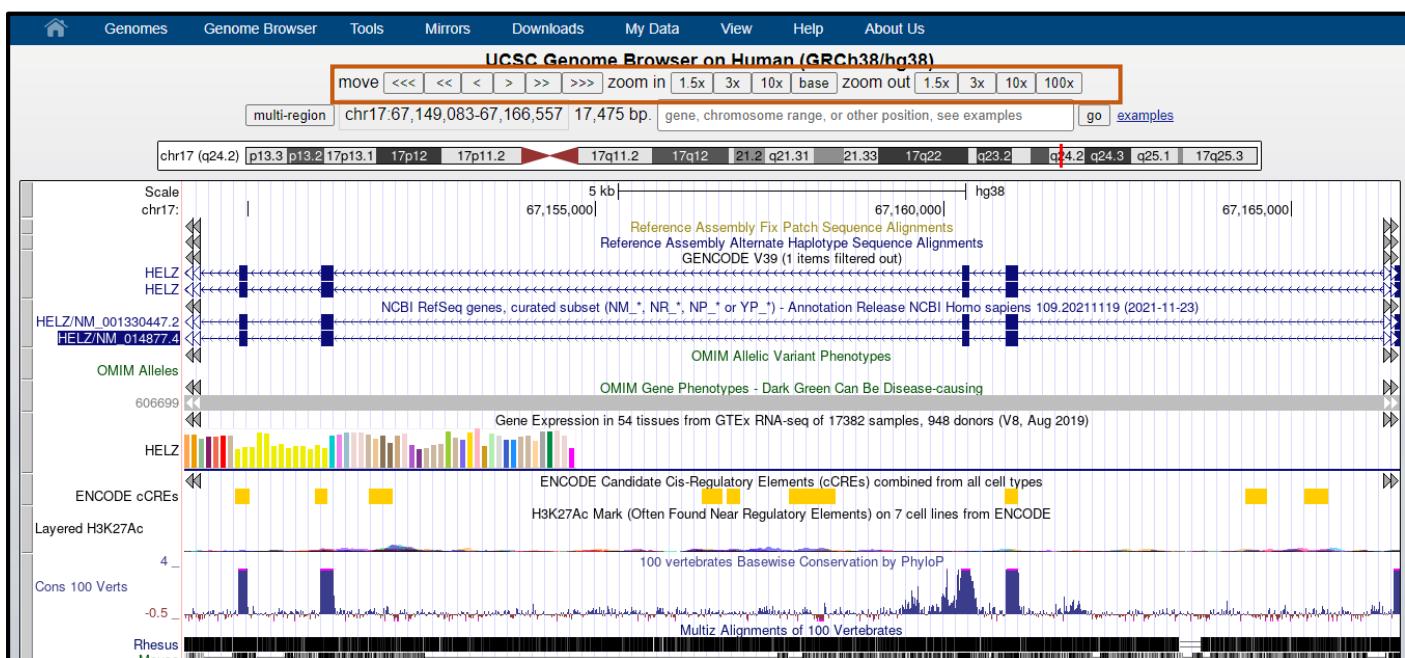
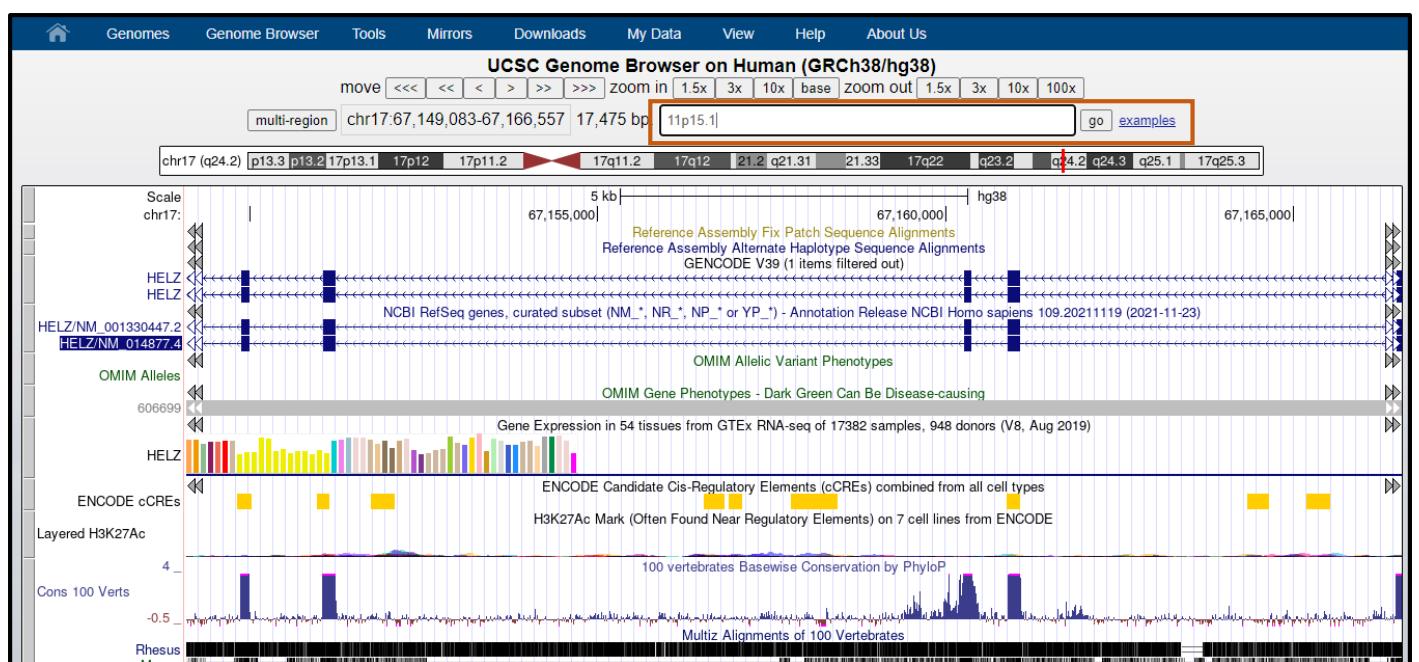
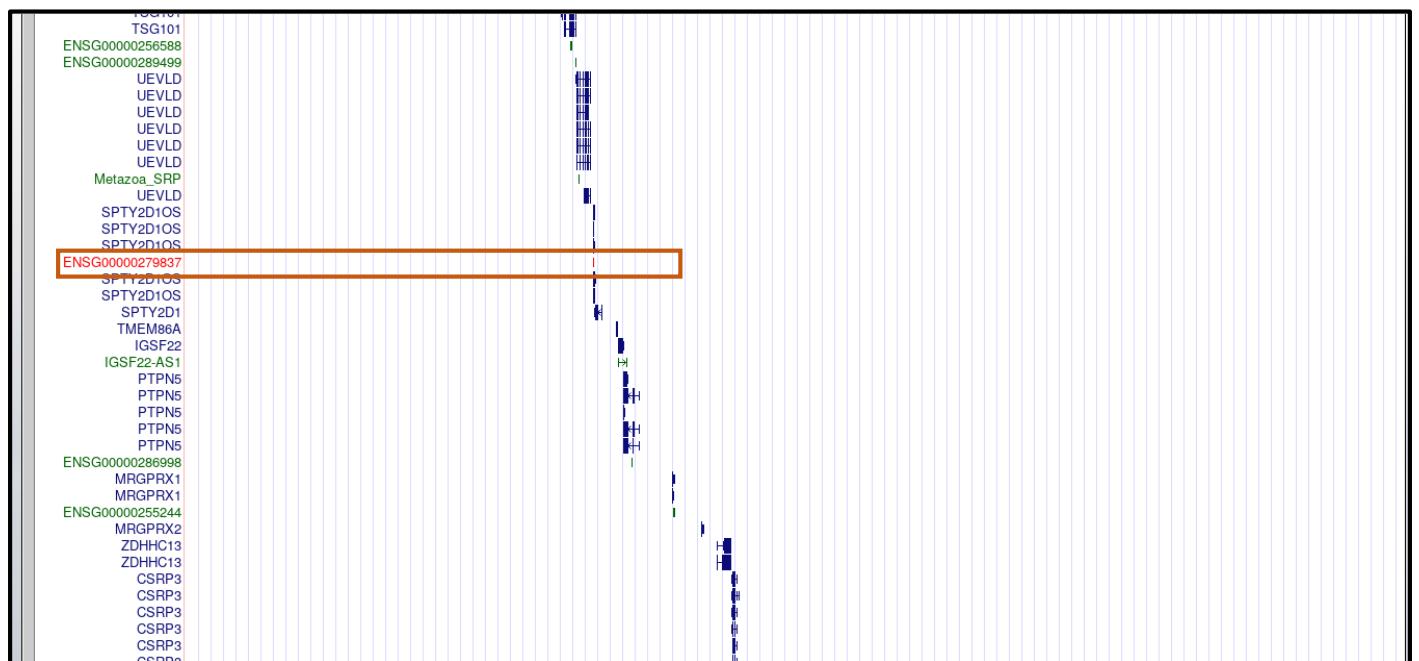


Fig9.3. Result after zooming 10x



**Fig10. Navigation by cytological band: 11p15.1**



**Fig10.1. Result for cytological band: 11p15.1**

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

**Human Gene ENSG00000279837 (ENST00000623697.1) from GENCODE V39**

Description: ENSG00000279837 (from geneSymbol)  
 Gencode Transcript: ENST00000623697.1  
 Gencode Gene: ENSG00000279837.1  
 Transcript (Including UTRs)  
 Position: hg38 chr11:18,601,882-18,602,649 Size: 768 Total Exon Count: 1 Strand: +

**Page Index** Sequence and Links Other Species mRNA Descriptions Other Names Methods  
 Data last updated at UCSC: 2022-01-17 08:30:34

**Sequence and Links to Tools and Databases**

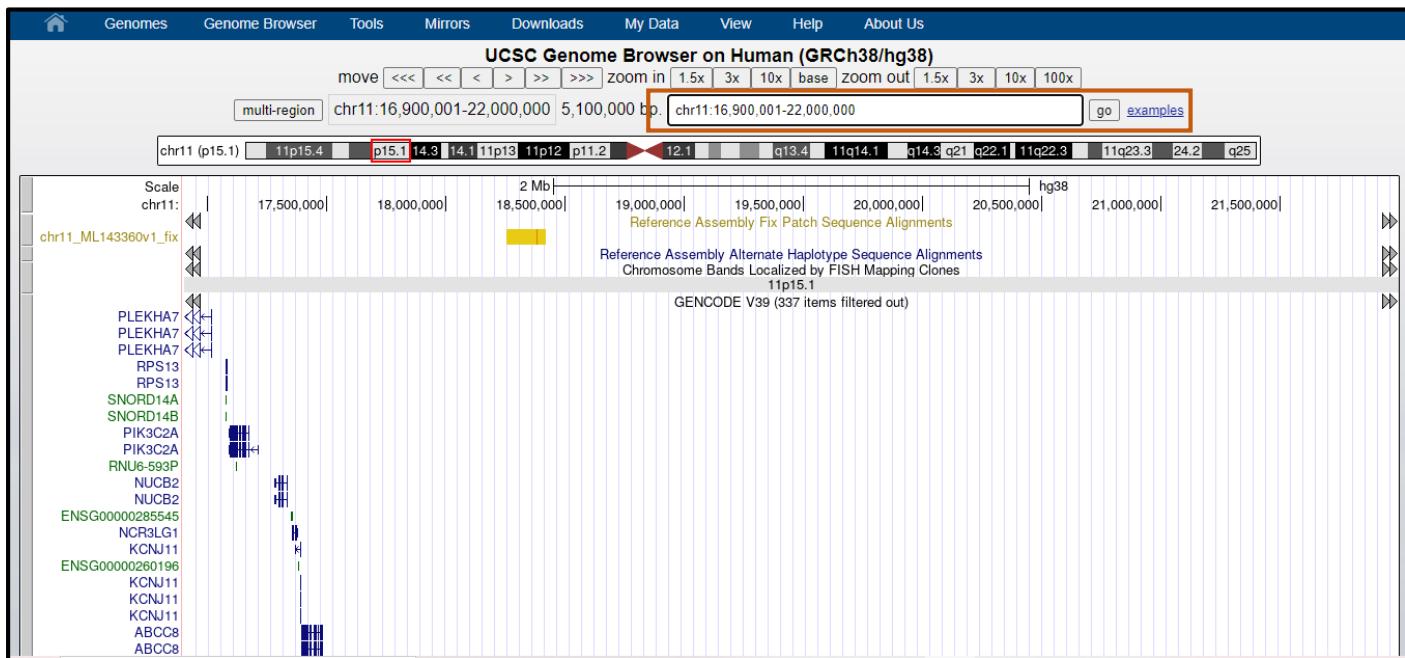
Genomic Sequence (chr11.18,601,882-18,602,649)	mRNA (may differ from genome)	No protein			
Gene Sorter	Genome Browser	Other Species FASTA	Table Schema	Ensembl	ExonPrimer
Gencode	PubMed				

**Orthologous Genes in Other Species**

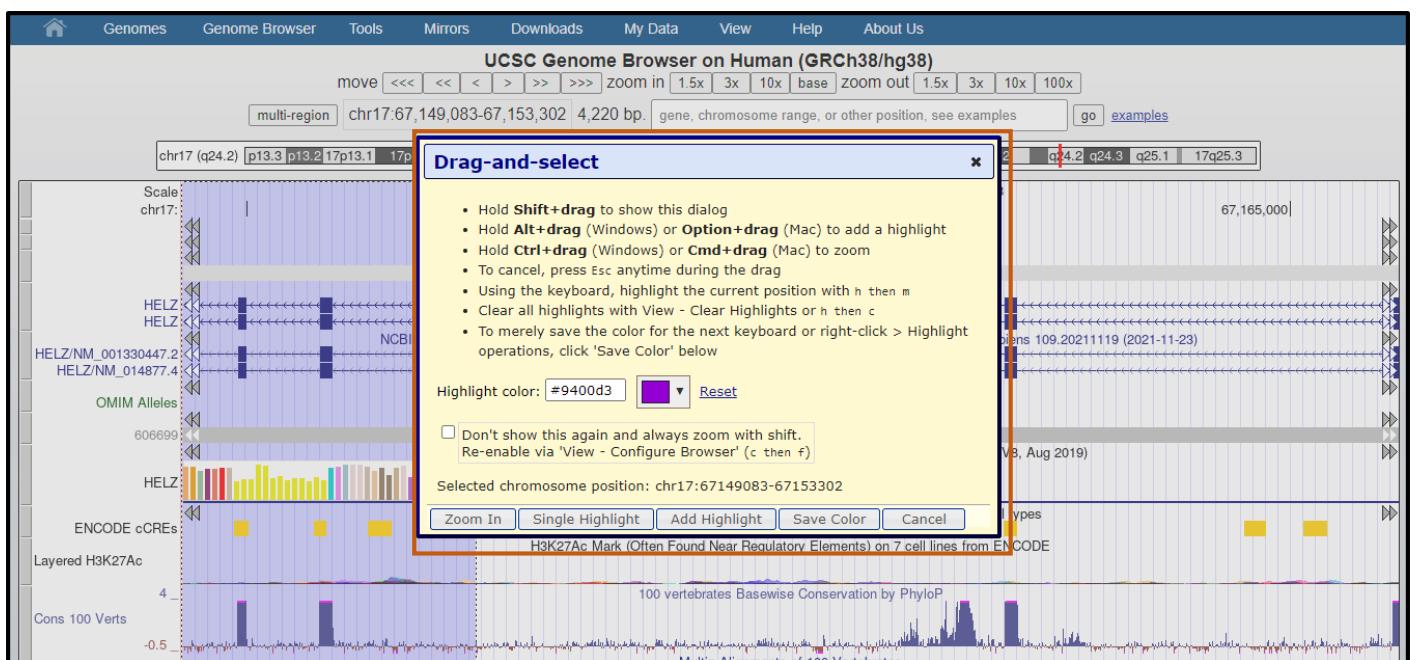
Orthologies between human, mouse, and rat are computed by taking the best BLASTP hit, and filtering out non-syntenic hits. For more distant species reciprocal-best BLASTP hits are used. Note that the absence of an ortholog in the table below may reflect incomplete annotations in the other species rather than a true absence of the orthologous gene.

Mouse	Rat	Zebrafish	D. melanogaster	C. elegans	S. cerevisiae
No ortholog	No ortholog	No ortholog	No ortholog	No ortholog	No ortholog

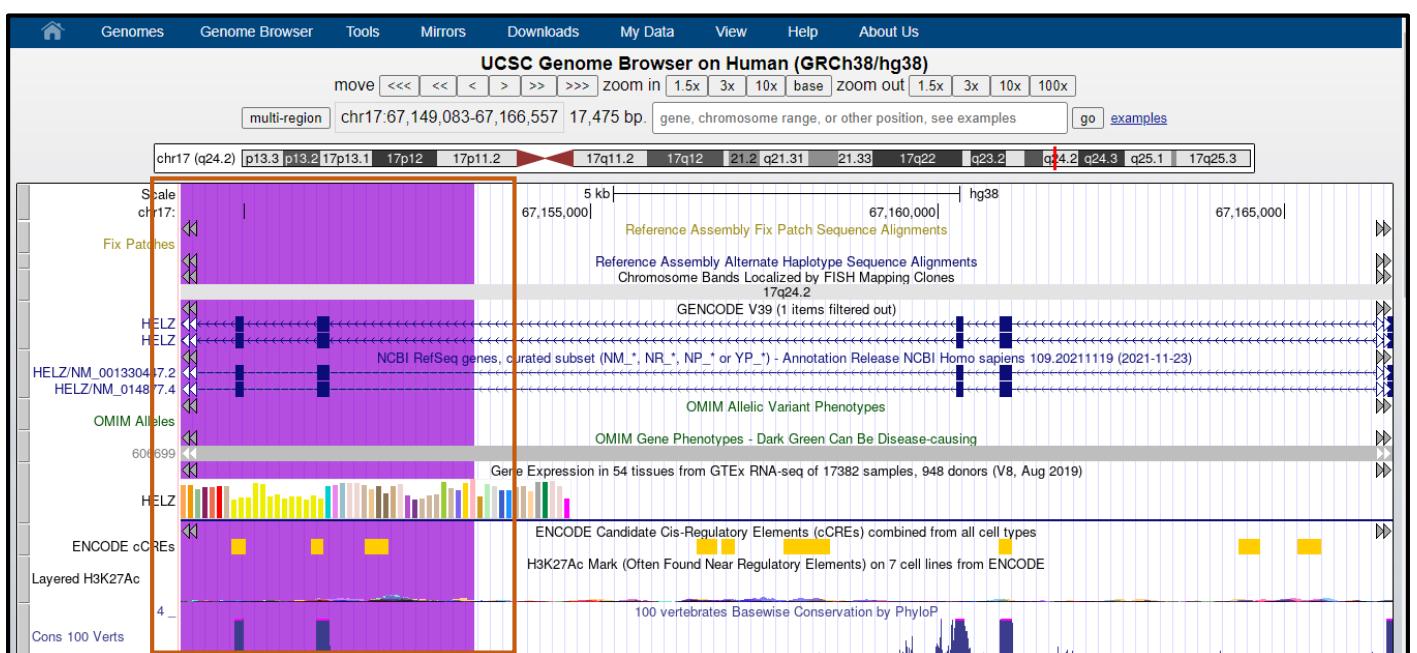
**Fig10.2. Description for cytological band: 11p15.1**



**Fig11. Navigation by coordinates**



**Fig12. Drag and select option for configuration**



### Fig12.1. Result after configuration

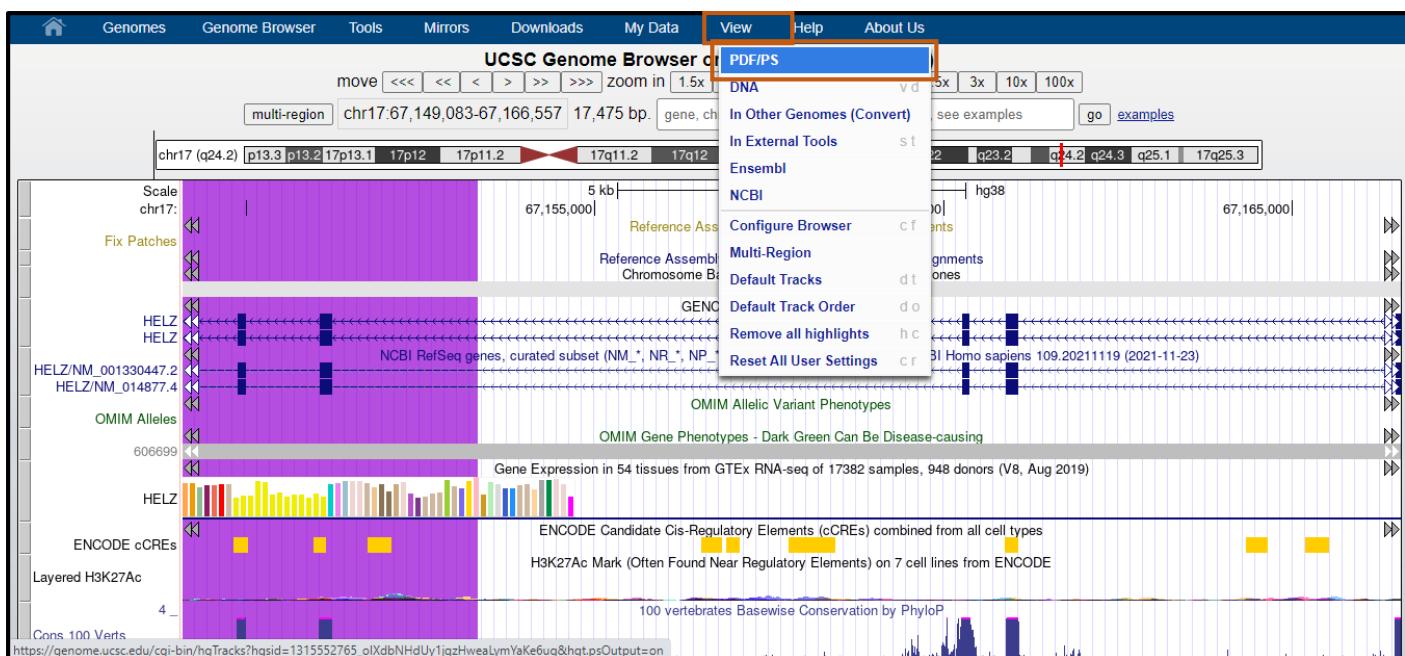


Fig13. Steps to export result as PDF

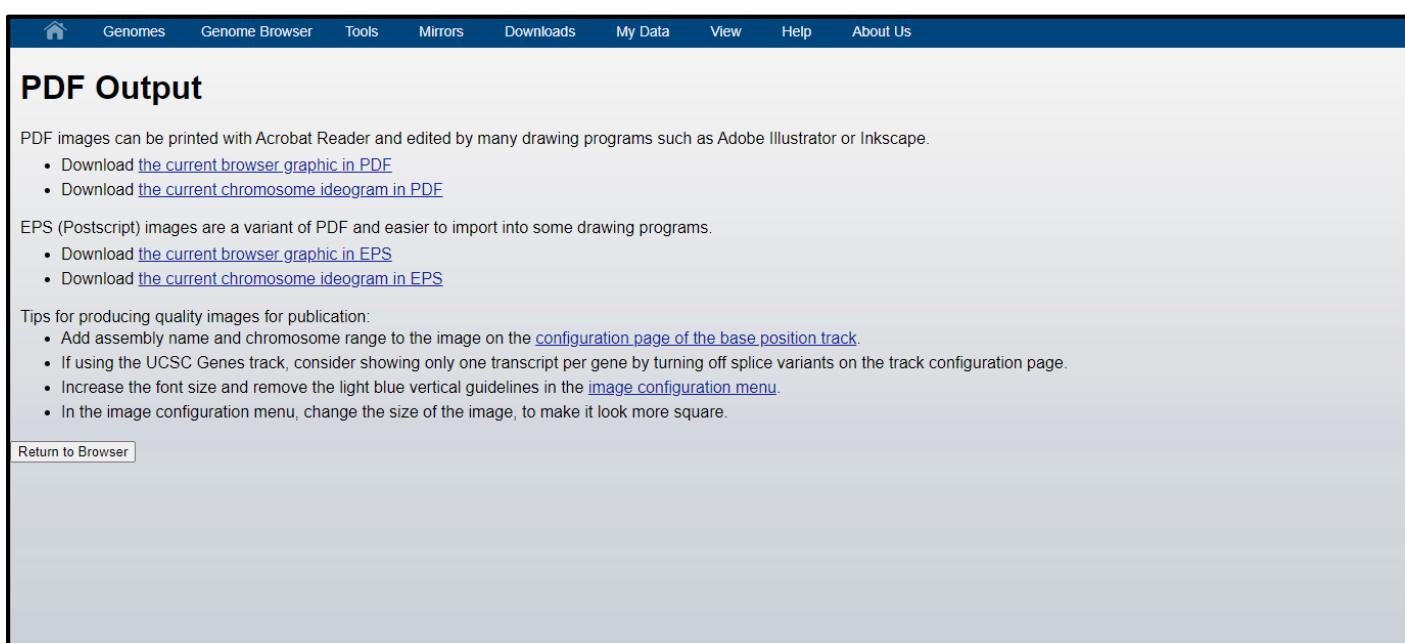
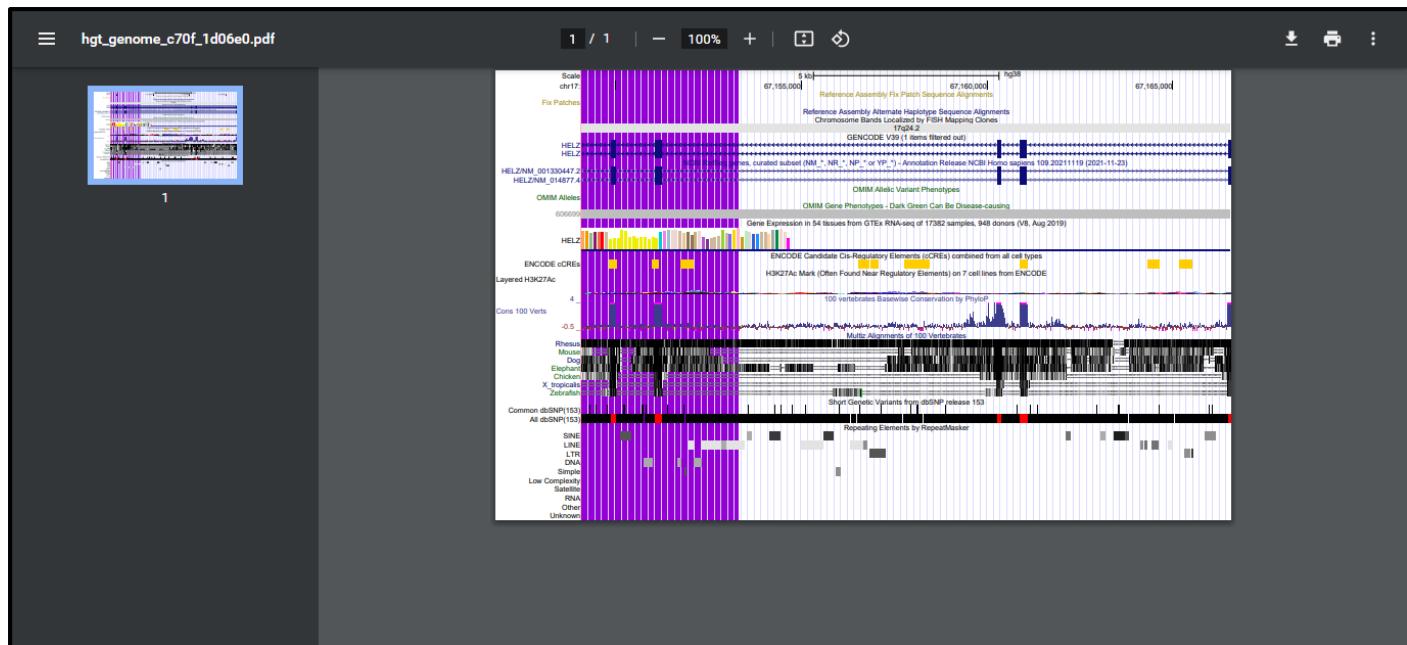


Fig13.1. Option to download PDF



**Fig13.2. Result in PDF format**

## RESULT:

UCSC genome browser was used for setting for GRCh38/hg38 browser and search options used were:

- Navigation by gene name
- Navigation by SNP id
- Navigation by Ref\_Seq: NM\_014877.4 (HEL gene)
- 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/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=defa>

[ult&virtMode=0&nonVirtPosition=&position=chr11%3A34438934%2D34472060&hgsid=131552765\\_oIXdbNHdUy1jqzHweaLymYaKe6ug](https://genome.ucsc.edu/cgi-bin/hgc?hgsid=131552765_oIXdbNHdUy1jqzHweaLymYaKe6ug&virtMode=0&nonVirtPosition=&position=chr11%3A34438934%2D34472060&hgsid=131552765_oIXdbNHdUy1jqzHweaLymYaKe6ug)

**Navigation by OMIM id:**

5. *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](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)

**Navigation by SNP id:**

6. *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](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)

**Navigation by Ref\_Seq:**

7. *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/cgi-bin/hgc?hgsid=1315552765\\_oIXdbNHdUy1jqzHweaLymYaKe6ug&db=hg38&c=chr17&l=67070443&r=67245196&o=67070443&t=67245196&g=ncbiRefSeqCurated&i=NM\\_014877.4](https://genome.ucsc.edu/cgi-bin/hgc?hgsid=1315552765_oIXdbNHdUy1jqzHweaLymYaKe6ug&db=hg38&c=chr17&l=67070443&r=67245196&o=67070443&t=67245196&g=ncbiRefSeqCurated&i=NM_014877.4)

**Navigation by Cytological band:**

8. *Human Gene ENSG00000279837 (ENST00000623697.1) from GENCODE V39*. (n.d.). Genome.ucsc.edu. Retrieved March 28, 2022, from [https://genome.ucsc.edu/cgi-bin/hgGene?hgg\\_gene=ENST00000623697.1&hgg\\_chrom=chr11&hgg\\_start=18601881&hgg\\_end=18602649&hgg\\_type=knownGene&db=hg38](https://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000623697.1&hgg_chrom=chr11&hgg_start=18601881&hgg_end=18602649&hgg_type=knownGene&db=hg38)

## WEBLEM 9b

### Ensembl Genome Browser (URL: <https://asia.ensembl.org/index.html>)

#### AIM:

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

#### INTRODUCTION:

Ensembl is a bioinformatics project to organize biological information around the sequences of large genomes. It is a comprehensive source of stable automatic annotation of individual genomes, and of the synteny and orthology relationships between them. It is also a framework for integration of any biological data that can be mapped onto features derived from the genomic sequence. Ensembl is available as an interactive Web site, a set of flat files, and as a complete, portable open source software system for handling genomes. All data are provided without restriction, and code is freely available. Ensembl's aims are to continue to "widen" this biological integration to include other model organisms relevant to understanding human biology as they become available; to "deepen" this integration to provide an ever more seamless linkage between equivalent components in different species; and to provide further classification of functional elements in the genome that have been previously elusive.

#### 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. Observe the results.
5. Use the configuration tools for the tracks.
6. Interpret the results.

#### OBSERVATION:

Fig1. Homepage for Ensembl

**e!Ensembl** ASIA BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Search Human (Homo sapiens)

Search all categories ▾ Search ... Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

**Genome assembly: GRCh38.p13 (GCA\_000001405.28)**

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

**Gene annotation**

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

**Example gene**

**Pax6** INS  
**FOXP2**  
**BRCA2**  
**DMD**  
**ssh**

**Example transcript**

**Comparative genomics**

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

**Example gene tree**

**Variation**

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVE)

**Example variant**

ATCGAGCT  
ATCAGCT  
ATCGAGAT

**Fig2. Homepage for Human (GRCh38.p13) genome assembly**

**e!Ensembl** ASIA BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Search Human (Homo sapiens)

Search all categories ▾ helad1 Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

**Genome assembly: GRCh38.p13 (GCA\_000001405.28)**

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

**Gene annotation**

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

**Example gene**

**Pax6** INS  
**FOXP2**  
**BRCA2**  
**DMD**  
**ssh**

**Example transcript**

**Comparative genomics**

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

**Example gene tree**

**Variation**

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVE)

**Example variant**

ATCGAGCT  
ATCAGCT  
ATCGAGAT

**Fig3. Search for helad1 gene**

Ensembl ASIA BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

New Search

Current selection: Only searching Human ▾ helad1

1 results match helad1 when restricted to species: Human

NAV2 (Human Gene)  
ENSG00000166833 11:19350724-20121601 1  
Neuron navigator 2 [Source:HGNC Symbol;Acc:HGNC:15997].  
Variant table • Phenotypes • Location • External Refs • Regulation • Orthologues • Gene tree

Per page: 10 25 50 100

Layout: Standard Table

Tip: Help and Documentation can be searched from the homepage! Just type in a term you want to know more about, like non-synonymous SNP.

About Us Get help Our sister sites Follow us

About us Using this website Ensembl Bacteria Blog

Contact us Adding custom tracks Ensembl Fungi Twitter

Citing Ensembl Downloading data Ensembl Plants Facebook

Privacy policy Video tutorials Ensembl Protists

Disclaimer Variant Effect Predictor Ensembl Metazoa (VEP)

Fig4. Hit page for helad1 gene

Ensembl ASIA BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Location: 11:19,350,724-20,121,601 Gene: NAV2

Gene-based displays

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
- Secondary Structure
- Comparative Genomics
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues
- Paralogues
- Ensembl protein families
- Ontologies
- GO: Biological process
- GO: Molecular function
- GO: Cellular component
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
- Gene history

Gene: NAV2 ENSG00000166833

Description neuron navigator 2 [Source:HGNC Symbol;Acc:HGNC:15997]

Gene Synonyms FLJ10633, FLJ11030, FLJ23707, HELAD1, KIAA1419, POMFIL2, RAINB1

Location Chromosome 11: 19,350,724-20,121,601 forward strand. GRCh38:CM000673.2

About this gene This gene has 15 transcripts (splice variants), 209 orthologues and 2 paralogues.

Transcripts Hide transcript table

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
ENST00000396085.6	NAV2-203	11501	2432aa	Protein coding	CCDS7851	Q8IVL1-2	-	GENCODE basic APPRIS ALT1 TSL
ENST00000360655.8	NAV2-202	10667	2365aa	Protein coding	CCDS53612	Q8IVL1-4	-	GENCODE basic APPRIS ALT2 TSL
ENST00000396087.7	NAV2-204	7882	2488aa	Protein coding	CCDS58126	Q8IVL1-1	-	GENCODE basic TSL5
ENST00000533917.5	NAV2-212	5084	1493aa	Protein coding	CCDS44532	Q8IVL1-5	-	GENCODE basic TSL2
ENST00000525322.5	NAV2-206	2716	815aa	Protein coding	-	E9PNV5	-	TSL2 CDS 3' incomplete
ENST00000530408.1	NAV2-210	556	160aa	Protein coding	-	E9PLU3	-	TSL5 CDS 3' incomplete
ENST00000650578.1	NAV2-215	256	63aa	Protein coding	-	A0A3B3ISY2	-	CDS 3' incomplete
ENST00000534229.1	NAV2-213	684	No protein	Processed transcript	-	-	-	TSL3

Fig5. Result for NAV2 gene

Location: 11:19,350,724-20,121,601 Gene: NAV2 Transcript: NAV2-201

**Transcript-based displays**

- Summary
- Sequence
  - Exons
  - cDNA
  - Protein
- Protein Information
  - Protein summary
  - Domains & features
  - Variants
  - PDB 3D protein model
  - AlphaFold predicted model
- Genetic Variation
  - Variant table
  - Variant image
  - Haplotypes
  - Population comparison
  - Comparison image
- External References
  - General identifiers
  - Oligo probes
- Supporting evidence
- ID History
  - Transcript history
  - Protein history

**Summary**

**Statistics**

CCDS

Uniprot

Transcript Support Level (TSL)

Version

Type

Annotation Method

GENCODE basic gene

Exons: 38, Coding exons: 38, Transcript length: 11,492 bps, Translation length: 2,429 residues

This transcript is a member of the Human CCDS set: [CCDS7850](#)

This transcript corresponds to the following Uniprot identifiers: [Q8IVL1](#)

TSL:1

ENST00000349880.9

Protein coding

Manual annotation (determined on a case-by-case basis) from the Havana project.

This transcript is a member of the [Gencode basic gene set](#).

**Configure this page**

**Custom tracks**

**Export data**

**Share this page**

**Bookmark this page**

**Fig5.1. Result for NAV2 gene**

**Summary**

Name: [NAV2](#) (HGNC Symbol)

CCDS: This gene is a member of the Human CCDS set: [CCDS44552.1](#), [CCDS53612.1](#), [CCDS58126.1](#), [CCDS7850.1](#), [CCDS7851.2](#)

UniProtKB: This gene has proteins that correspond to the following UniProtKB identifiers: [Q8IVL1](#)

RefSeq: This Ensembl/Gencode gene contains transcript(s) for which we have selected identical RefSeq transcript(s). If there are other RefSeq transcripts available they will be in the [External references](#) table

Ensembl version: ENSG00000166833.23

Other assemblies: There is no ungapped mapping of this gene onto the GRCh37 assembly. View this locus in the GRCh37 archive: [ENSG00000166833](#)

Gene type: Protein coding

Annotation method: Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see [article](#)

Annotation Attributes: overlapping locus [\[Definitions\]](#)

Go to [Region in Detail](#) for more tracks and navigation options (e.g. zooming)

**Genes (Comprehensive set...)**

19.4Mb 19.5Mb 19.6Mb 19.7Mb 19.8Mb 19.9Mb 20.0Mb 20.1Mb

790.88 kb

Forward strand

NAV2-202 > protein coding

NAV2-201 > protein coding

NAV2-203 > protein coding

NAV2-204 > protein coding

NAV2-213 > NAV2-208 > NAV2-2

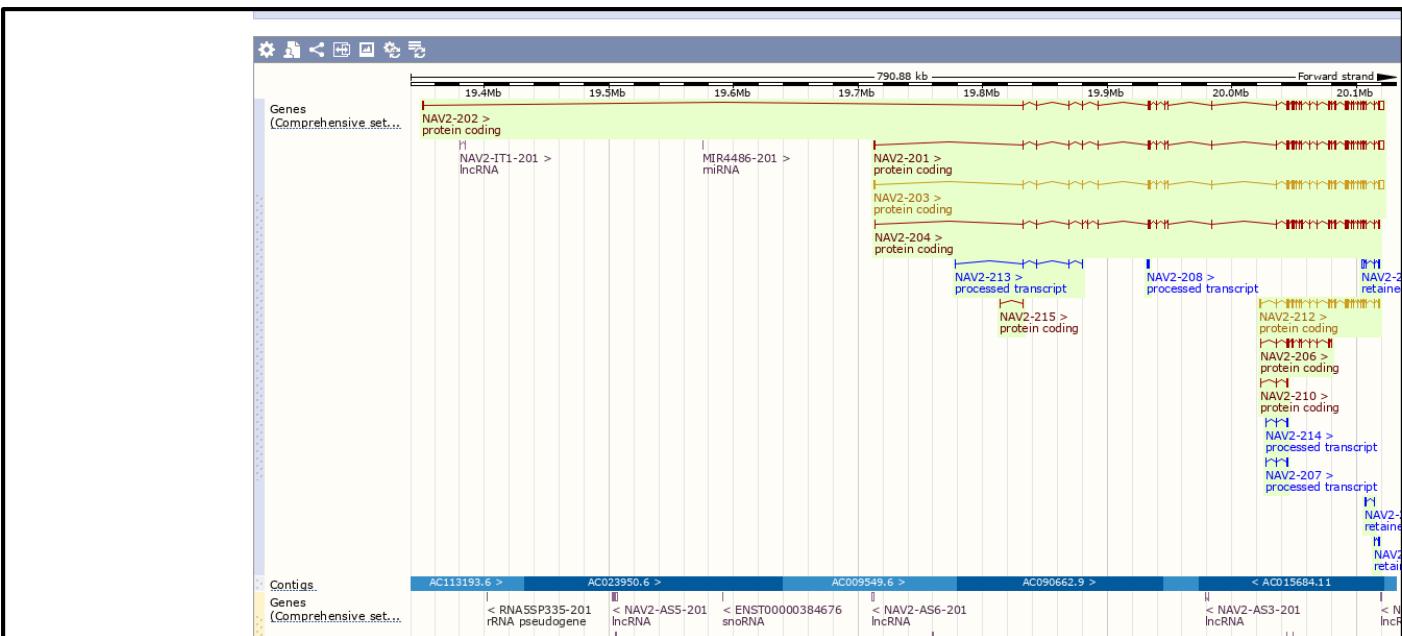


Fig5.2. Tracks information

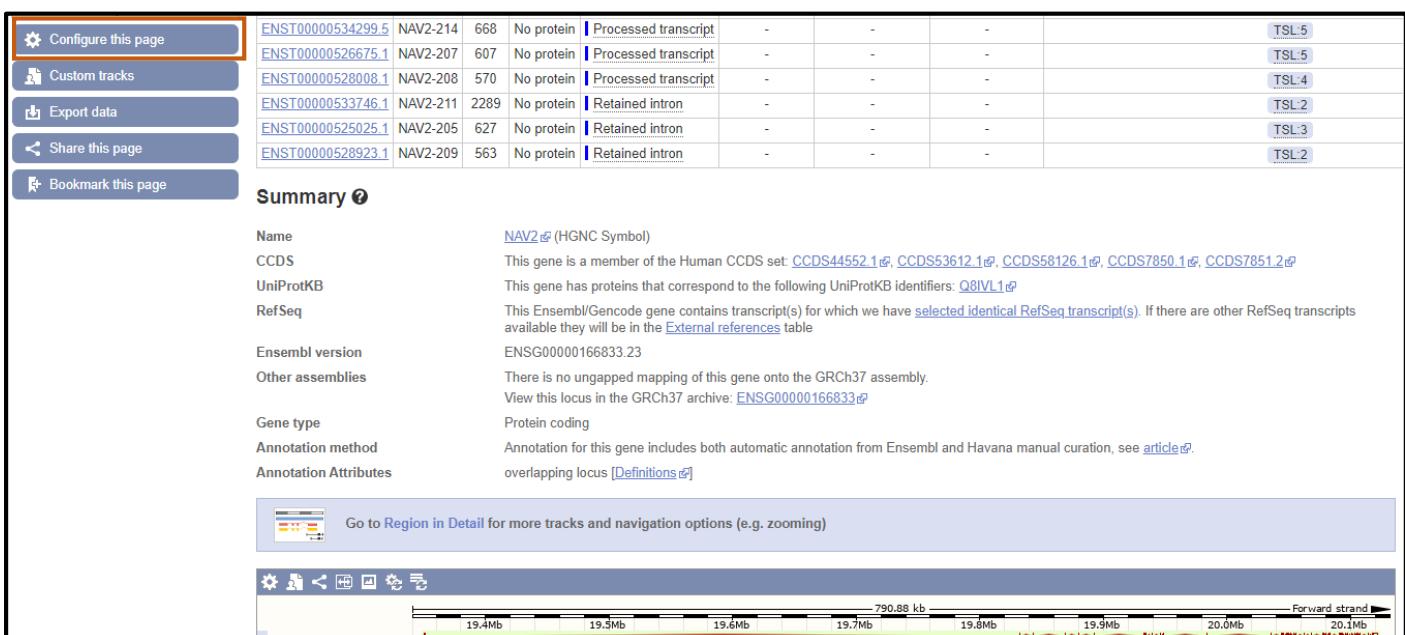


Fig6. Steps to configure tracks

Annotation Attributes

overlapping locus [Definitions

Configure this page

Find a track

Active tracks

Sequence and assembly

Genes and transcripts

Regulation

Information and decorations

Select from available configurations:

Current unsaved

Fig6.1. Option for tracks configuration

Annotation Attributes

overlapping locus [Definitions

Configure this page

Find a track

Active tracks

Sequence and assembly

Genes and transcripts

Regulation

Information and decorations

Select from available configurations:

Current unsaved [Save current configuration](#)

Active tracks

Sequence and assembly

Genes and transcripts

Regulation

Information and decorations

Scale bar

Ruler

Disabled track summary

Information

Regulation Legend

Segmentation Legend

Methylation Legend

Gene Legend

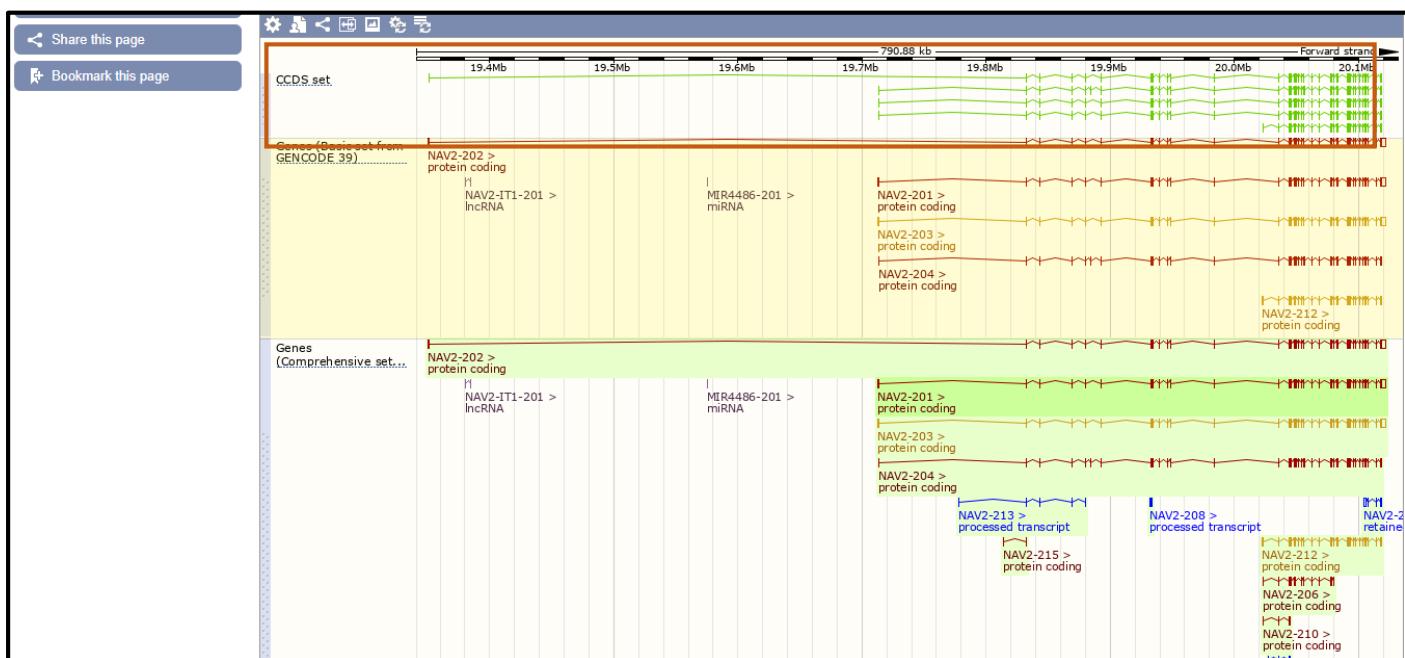
Alignment Difference Legend

Variant Legend

Structural Variant Legend

Fig6.2. Tracks configuration

**Fig6.3. Tracks configuration**



**Fig6.4. Updated result 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.

## CONCLUSION:

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

**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 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.

**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:

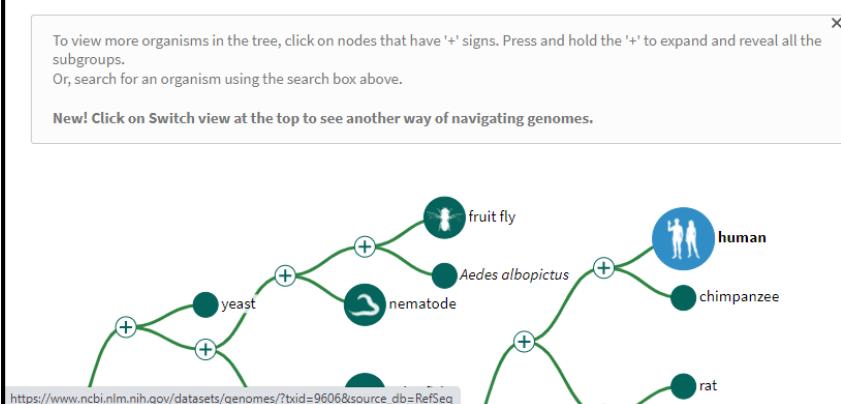
NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Log in

### Genome Data Viewer

**Switch view** **Search organisms** **Homo sapiens (human)**

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.



**Homo sapiens (human)**

Search in genome

Examples: TP53, chr17:7667000-7689000, DNA repair

Assembly

**Browse genome** **BLAST genome**

**Download via NCBI Datasets** **Feedback**

Fig1. Homepage for Genome Data Viewer

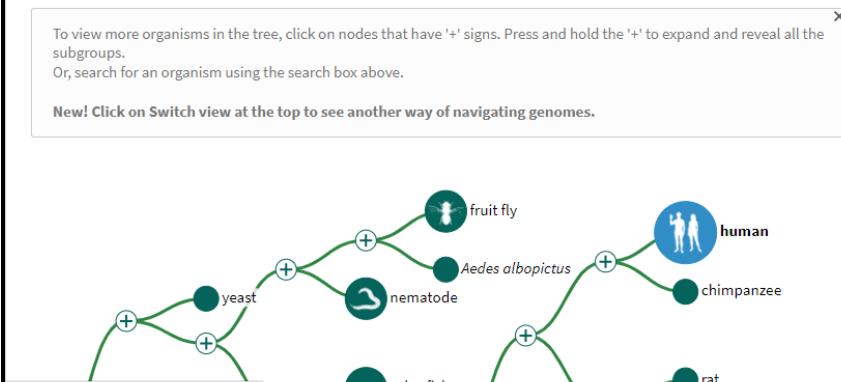
NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Log in

### Genome Data Viewer

**Switch view** **Search organisms** **Homo sapiens (human)**

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.



**Homo sapiens (human)**

Search in genome

Examples: TP53, chr17:7667000-7689000; DNA repair

Assembly

**Browse genome** **BLAST genome**

**Download via NCBI Datasets** **Feedback**

Fig2. Search for DNA repair in GRCh38.p13 assembly

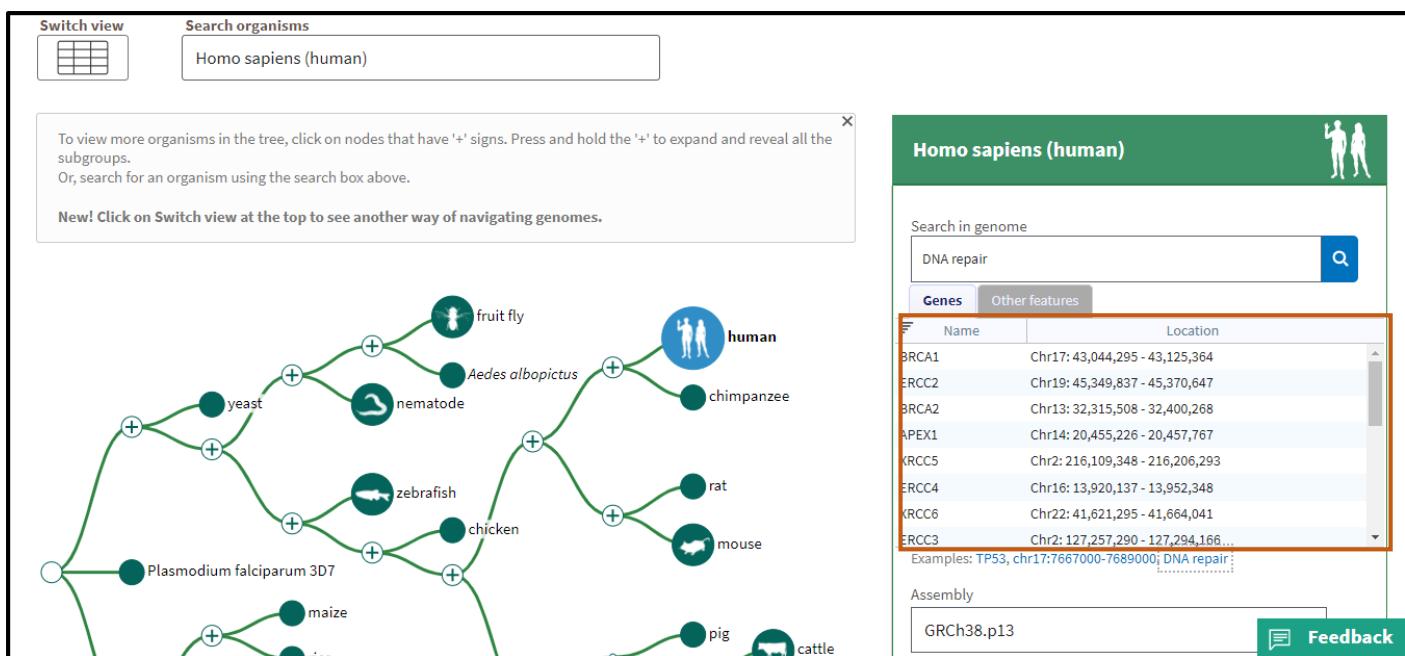


Fig3. Results for DNA repair genes

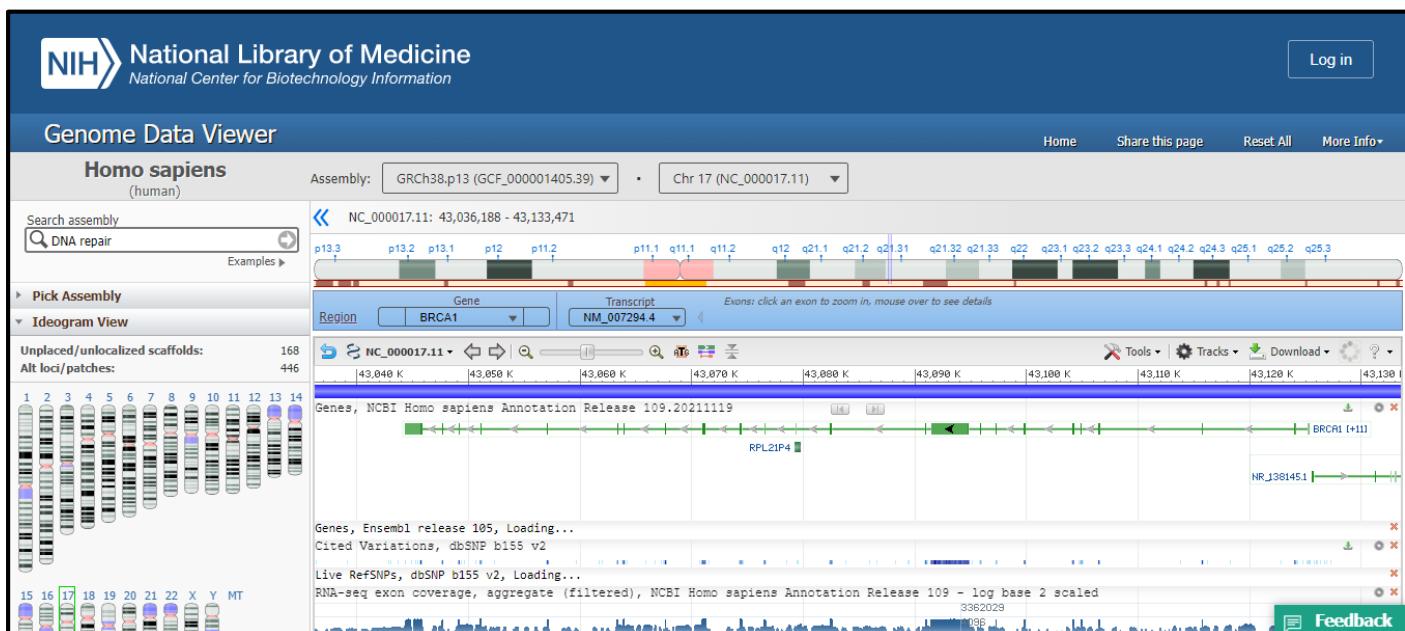


Fig4. Result for BRCA1 gene

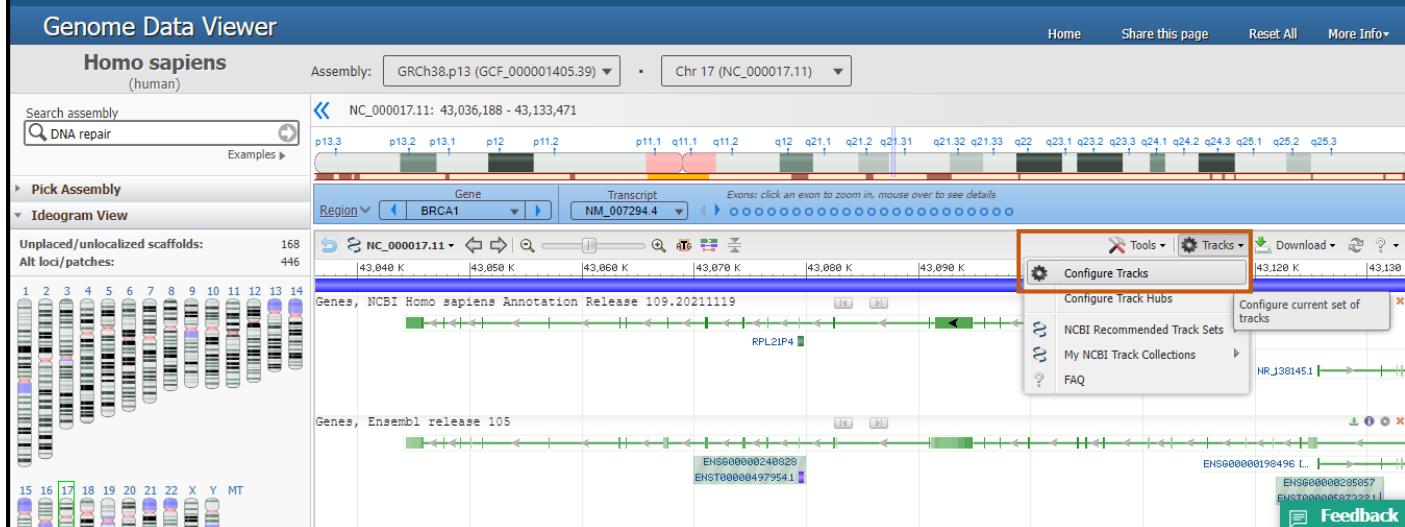


Fig5. Steps to configure tracks

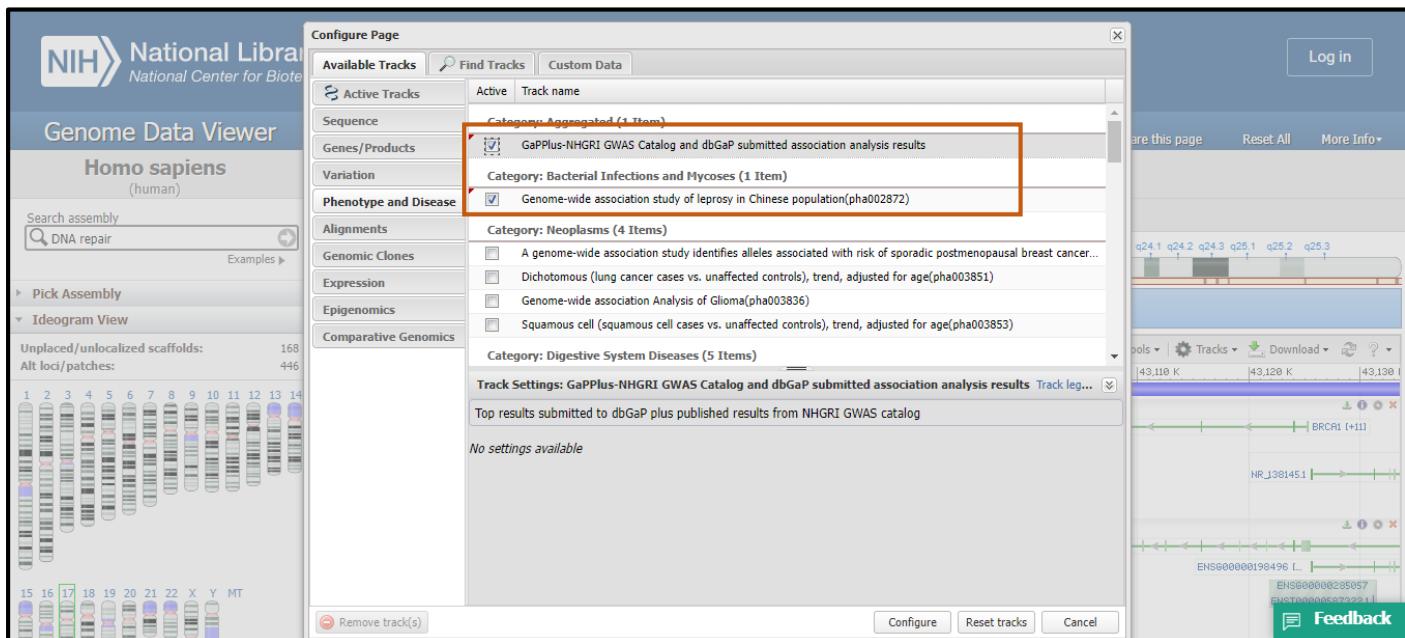


Fig5.1. Configuration page for phenotype and disease



Fig5.2. Updated result after configuration

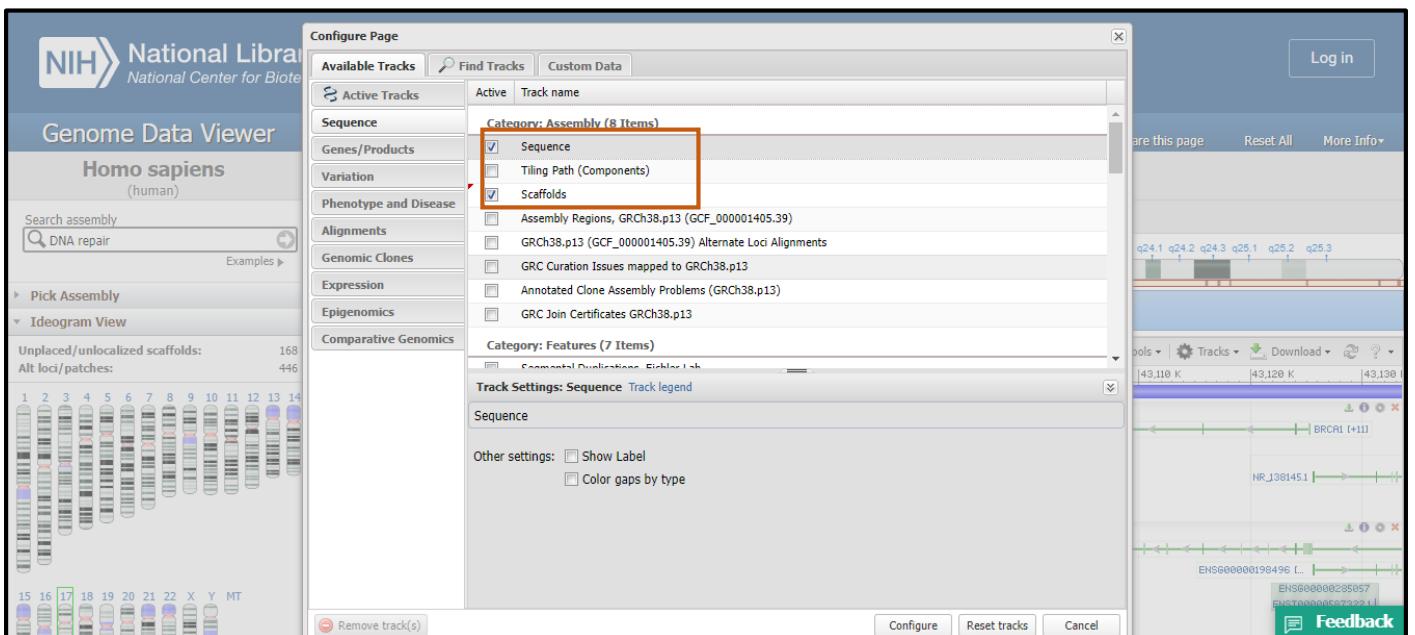


Fig5.3. Configuration page for sequence

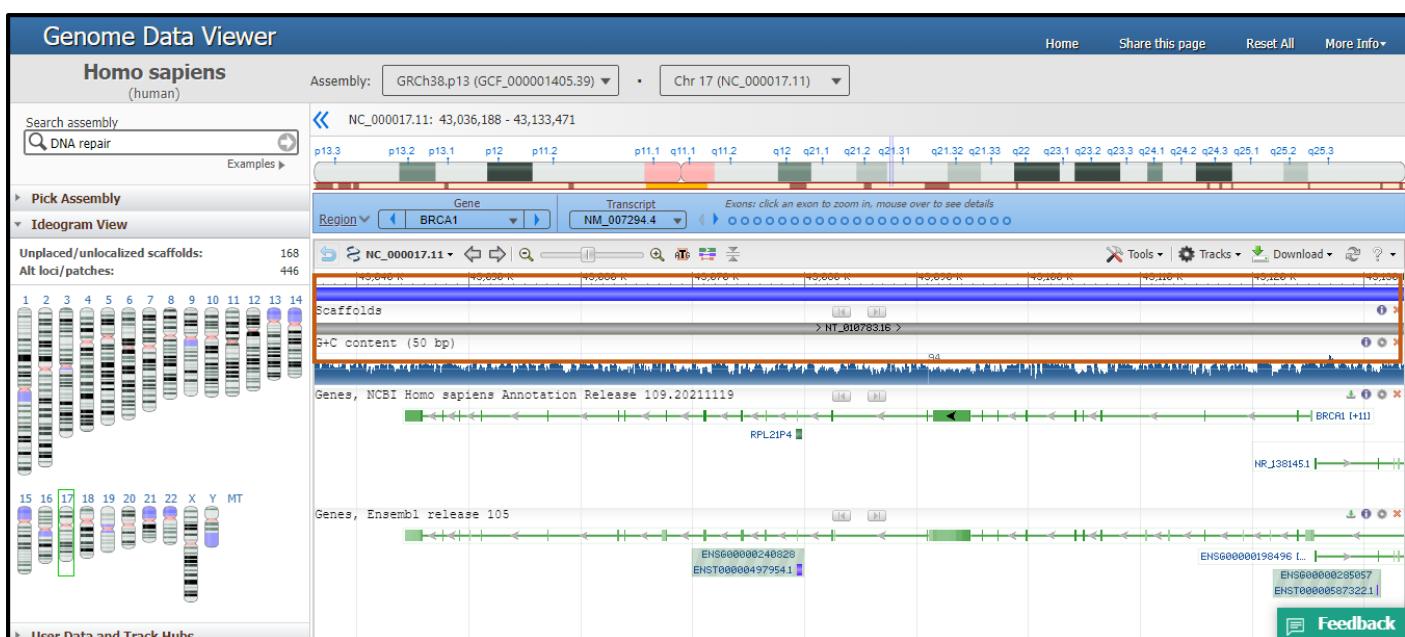


Fig5.4. Updated result after configuration

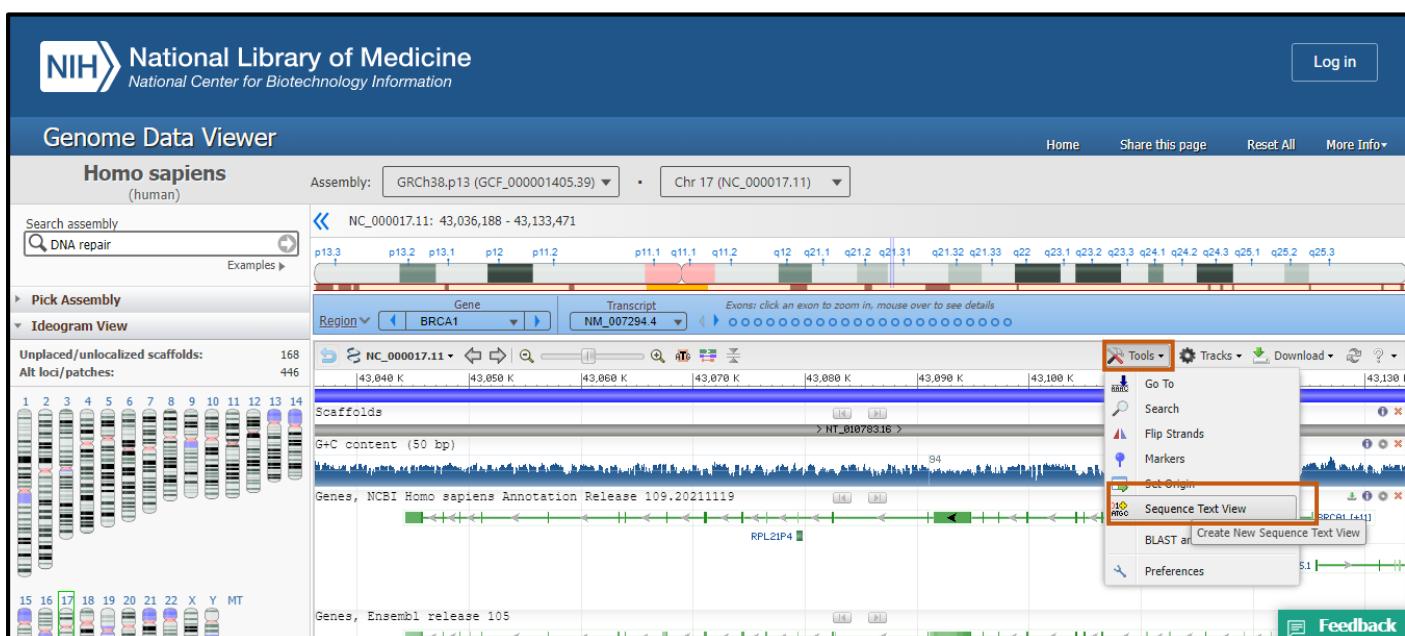


Fig6. Steps for sequence text view



Fig6.1. Result for sequence text view



Fig7. Options to view exon information

Genome Data Viewer

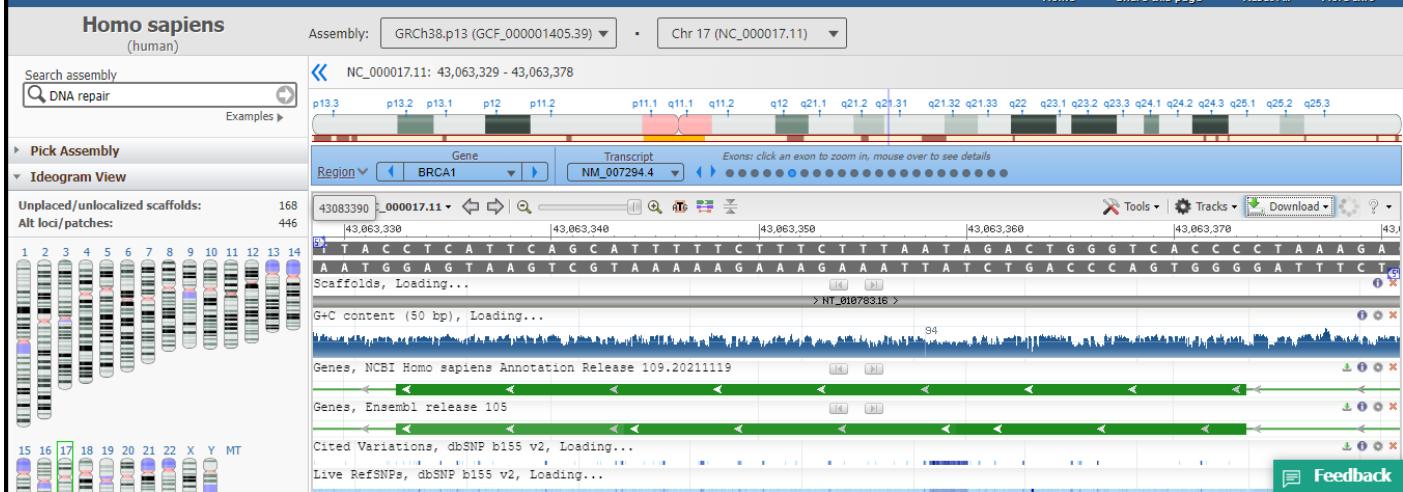


Fig7.1. Result for exon 18

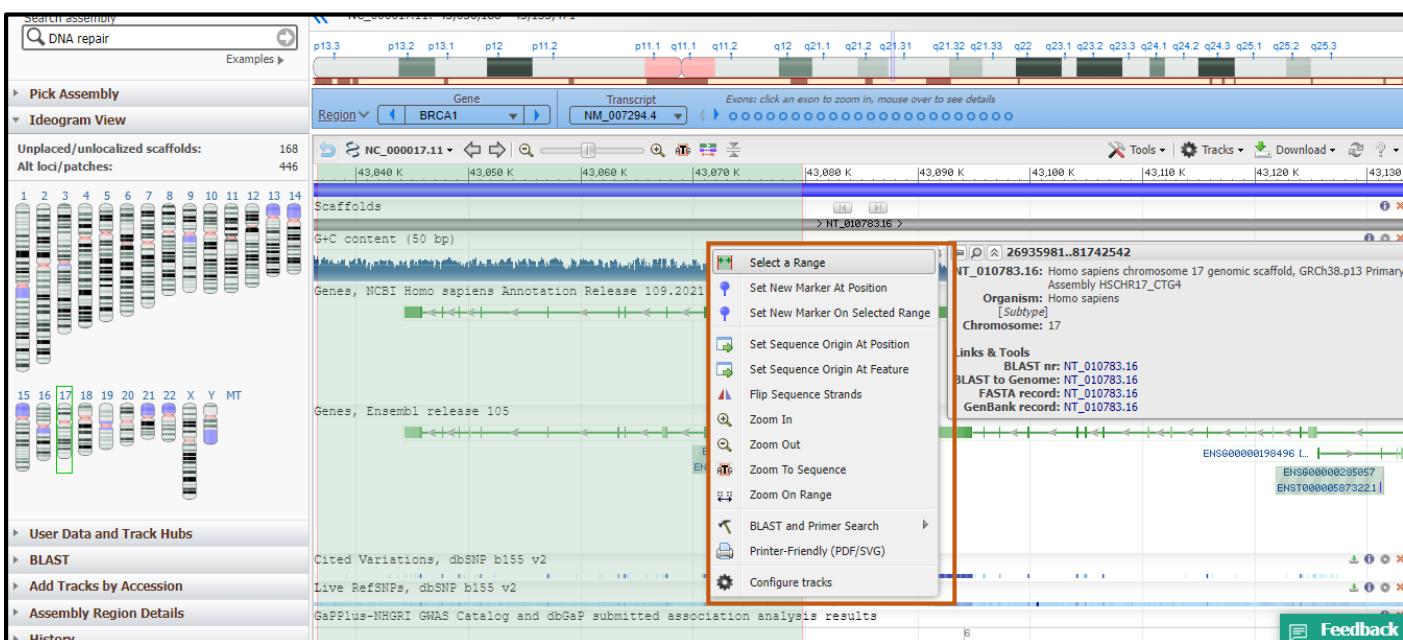
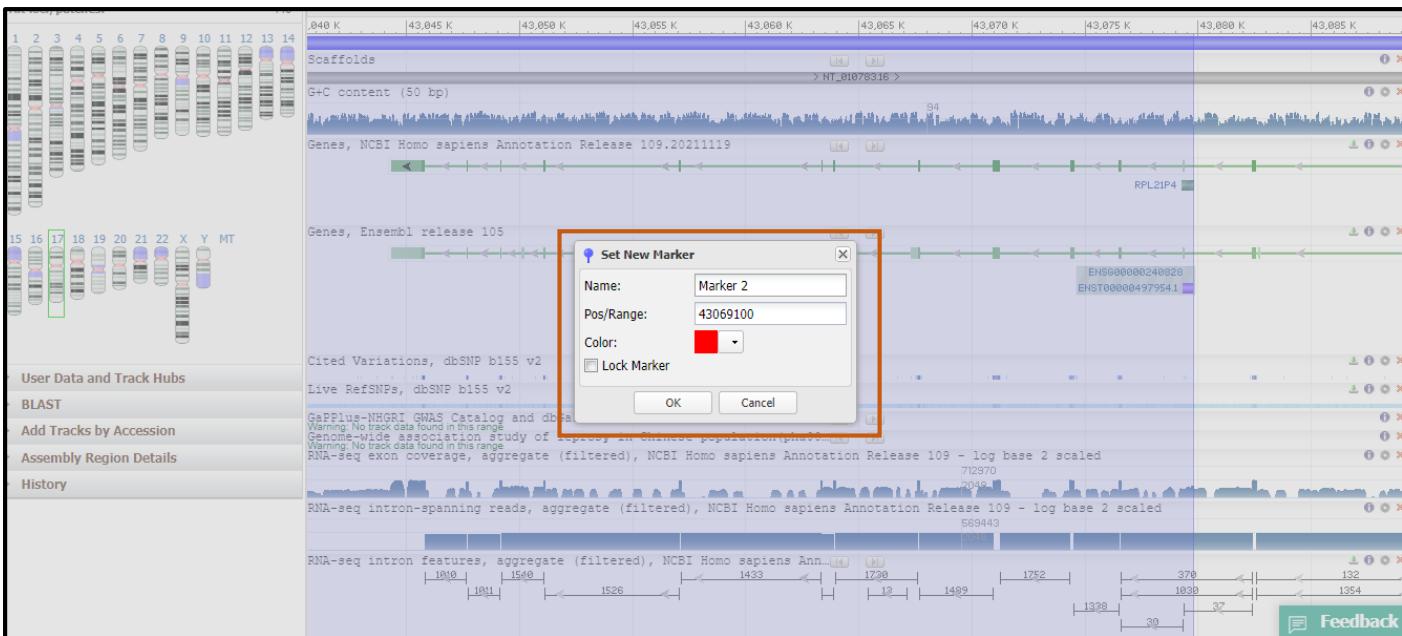
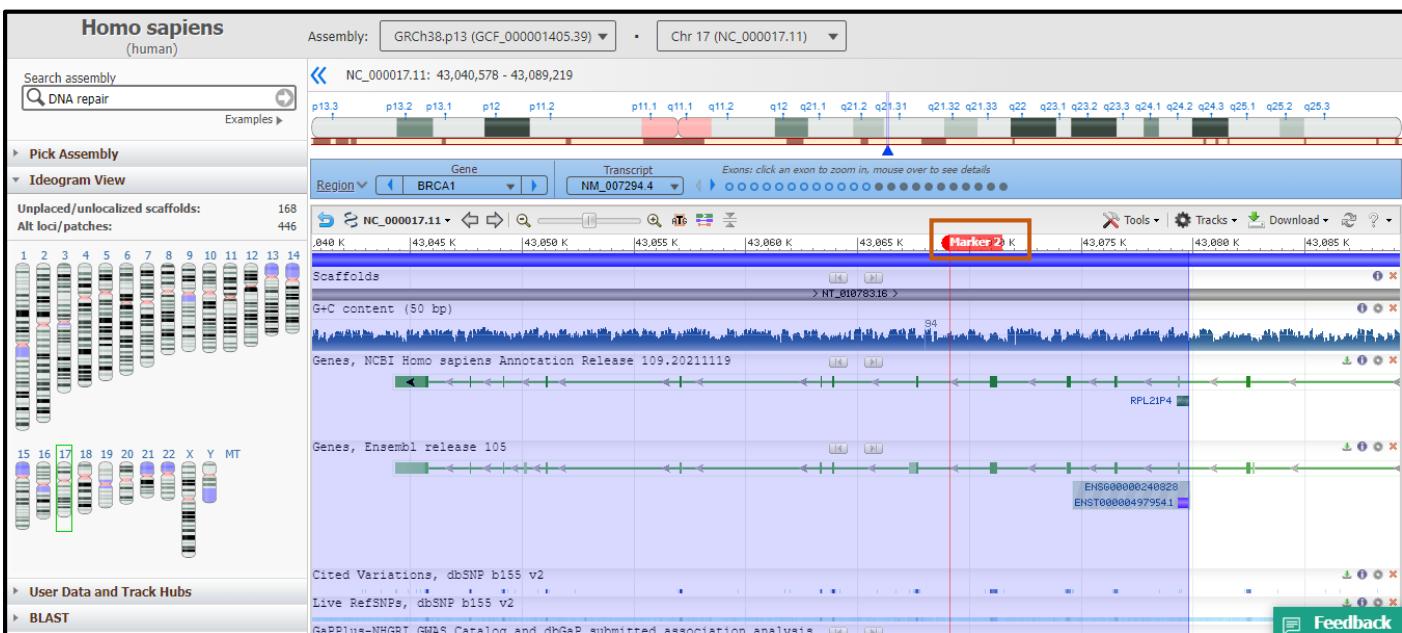


Fig8. Drag and select option for configuring tracks



**Fig8.1. Option to set new marker**



**Fig8.2. 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](http://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](http://www.ncbi.nlm.nih.gov). Retrieved March 28, 2022, from [https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF\\_000001405.39](https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39)

## WEBLEM 10

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

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.

Genome projects employ state-of-the-art DNA sequencing, mapping, and computational technologies (including cross-disciplinary experimental designs) to expand our knowledge and understanding of molecular/cellular mechanisms, gene repertoires, genome architecture, and evolution. The revolution in new sequencing technologies and computational developments has allowed researchers to drive advances in genome assembly and annotation to make the process better, faster, and cheaper with key model organisms.

Such technical advantages and established recommendations and strategies have been widely applied in humans, terrestrial animals, and plants and crops. Genomic applications in aquatic species that could be potentially important for aquaculture are slower compared with human, livestock, and crops, compounded by larger diversity, lack of reference genomes, and more novice aquaculture industries. Given that aquaculture is the most rapidly expanding food sector, with the widest diversity of species cultured, it is poised for rapid adoption of genomics applications as these become more accessible.

Following is the state of the art within this field and provide a step-by-step introduction to the workflow involved in genome sequencing, assembly and annotation.

#### **GENOME SEQUENCING:**

Before genome sequencing, a must-have step involves RNA sequencing (RNA-seq) that has provided significant insights into the biological functions. RNA-seq plays a key role in genome annotation through the identification of protein-coding genes based on transcriptome sequencing data and *ab initio* or homology-based prediction. However, the use of RNA-seq for genome assembly is limited to genome scaffolding. While RNA-seq is a powerful technology that will likely remain a key asset in the biologist's toolkit, recent single-molecule mRNA sequencing approaches (e.g., Pacific Bioscience [PacBio] and Oxford Nanopore Technology [ONT]) have provided significant improvements in gene and genome annotation, making them appealing alternatives or complementary techniques for genome annotation.

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. The DNA synthesis is stopped by adding dideoxynucleotides. The dideoxynucleotides are labeled with fluorescent dyes, which terminate the DNA synthesis at positions containing all four bases, resulting in nested fragments that vary in length by a single base. When the labeled DNA is subjected to electrophoresis, the banding patterns in the gel reveal the DNA sequence.

The fluorescent traces of the DNA sequences are read by a computer program that assigns bases for each peak in a chromatogram. This process is called *base calling*. Automated base calling may generate errors and human intervention is often required to correct the sequence calls.

There are two major strategies for whole genome sequencing: the shotgun approach and the hierarchical 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. Overlapping sequences with an overall length of six to ten times the genome size are normally obtained for this purpose.

Despite the multiple coverage, sometimes certain genomic regions remain unsequenced, 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. This step of genome sequencing is also known as *finishing*, which is followed by computational assembly of all the sequence data into a final complete genome.

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.

During the era of human genome sequencing, there was a heated debate on the merits of each of the two strategies. In fact, there are advantages and disadvantages in either. The hierarchical approach is slower and more costly than the shotgun approach because it involves an initial clone-based physical mapping step. However, once the map is generated, assembly of the whole genome becomes relatively easy and less error prone. In contrast, the whole genome shotgun approach can produce a draft sequence very rapidly because it is based on the direct sequencing approach. However, it is computationally very demanding to assemble the short random fragments. Although the approach has been successfully employed in sequencing small microbial genomes, for a complex eukaryotic genome that contains high levels of repetitive sequences, such as the human genome, the full shotgun approach becomes less accurate and tends to leave more “holes” in the final assembled sequence than the hierarchical approach. Current genome sequencing of large organisms often uses a combination of both approaches.

### **GENOME SEQUENCE 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 *supercontigs*), 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 Repeat Masker 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 missassembly.

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.

### **GENOME ANNOTATION:**

As a real-world example, gene annotation of the human genome employs a combination 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.

### **Gene Ontology**

A problem arises when using existing literature because the description of a gene function uses natural language, which is often ambiguous and imprecise. Researchers working on different organisms tend to apply different terms to the same type of genes or proteins. Alternatively, the same terminology used in different organisms may actually refer to different genes or proteins. Therefore, there is a need to standardize protein functional descriptions. This demand has spurred the development of the gene ontology (GO) project, which uses a limited vocabulary to describe molecular functions, biological processes, and cellular components. The controlled vocabulary is organized such that a protein function is linked to the cellular function through a hierarchy of descriptions with increasing specificity. The top of the hierarchy provides an overall picture of the functional class, whereas the lower level in the hierarchy specifies more precisely the functional role. This way, protein functionality can be defined in a standardized and unambiguous way.

A GO description of a protein provides three sets of information: *biological process*, *cellular component*, and *molecular function*, each of which uses a unique set of nonoverlapping vocabularies. The standardization of the names, activities, and associated pathways provides consistency in describing overall protein functions and facilitates grouping of proteins of related functions. A database searching using GO for a particular protein can easily bring up other proteins of related functions in much the same way as using a thesaurus. Using GO, a genome annotator can assign functional properties of a gene product at different hierarchical levels, depending on how much is known about the gene product.

At present, the GO databases have been developed for a number of model organisms by an international consortium, in which each gene is associated with a hierarchy of GO terms. These have greatly facilitated genome annotation efforts.

### **Automated Genome Annotation**

With the genome sequence data being generated at an exponential rate, there is a need to develop fast and automated methods to annotate the genomic sequences. The automated approach relies on homology detection, which is essentially heuristic sequence similarity searching. If a newly sequenced gene or its gene

product has significant matches with a database sequence beyond a certain threshold, a transfer of functional assignment is taking place. In addition to sequence matching at the full length, detection of conserved motifs often offers additional functional clues. Because using a single database searching method is often incomplete and error prone, automated methods have to mimic the manual process, which takes into consideration multiple lines of evidence in assigning a gene function, to minimize errors. The following algorithm is an example that goes a step beyond examining sequence similarity and provides functional annotations based on multiple protein characteristics. GeneQuiz is a web server for protein sequence annotation.

### **Publishing the genome:**

Draft genome sequences are now being produced at an ever-increasing rate. Traditional databases such as ENSEMBL from the European Molecular Biology Labs (EMBL) and the Wellcome Trust Sanger Institute, or genomic databases from the National Center for Biotechnology Information (NCBI) providing access to genomes and meta-information can no longer annotate and curate all incoming genomes. NCBI therefore already provides the possibility to upload draft genome sequences and user-generated annotation. To allow other users to improve the assembly and its annotation, all available raw data should be uploaded, together with the assembled genome and all relevant meta-data, for example as a BioProject on NCBI.

Thus, a genome can be described at the highest resolution by a complete genome sequence. Whole-genome sequencing can be carried out using full shotgun or hierarchical approaches. The former requires more extensive computational power in the assembly step, and the latter is inefficient because of the physical mapping process required. Among the genome sequence assembly programs, ARACHNE and EULER are the best performers. Genome annotation includes gene finding and assignment of function to these genes. Functional assignment depends on homology searching and literature information. GO projects aim to facilitate automated annotation by standardizing the descriptions used for gene functions.

### **REFERENCES:**

1. Xiong, J. (2008). *Genome Mapping, Assembly, and Comparison. Essential bioinformatics*. Cambridge: Cambridge University Press. 243-259.
2. Ekblom, Robert; Wolf, Jochen B. W. (2014). *A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications*, 7(9), 1026–1042. doi:10.1111/eva.12178
3. Jung, H., Ventura, T., Chung, J. S., Kim, W.-J., Nam, B.-H., Kong, H. J., Kim, Y.-O., Jeon, M.-S., & Eyun, S. (2020). Twelve quick steps for genome assembly and annotation in the classroom. *PLOS Computational Biology*, 16(11), e1008325. <https://doi.org/10.1371/journal.pcbi.1008325>