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:

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 *Neisseria gonorrhoeae* using BPROM tool	160-165	18/03/22	
8c.	To predict bacterial operon and gene for *Neisseria gonorrhoeae* 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 *Neisseria gonorrhoeae* 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	

DATE: 11/02/22

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 α-helices and β-sheets. It has been estimated that nearly 50% of residues of a protein fold into either α-helices and β-strands. As a review, an α-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 α-helices. A β-sheet consists of two or more β-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. β-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 clos eprotein 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. Insecondary 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
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

DATE: 11/02/22

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
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. Abinitio, 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 and http://bioinf.cs.ucl.ac.uk/psipred/
4. Observe and interpret the results from all three servers.

OBSERVATIONS:

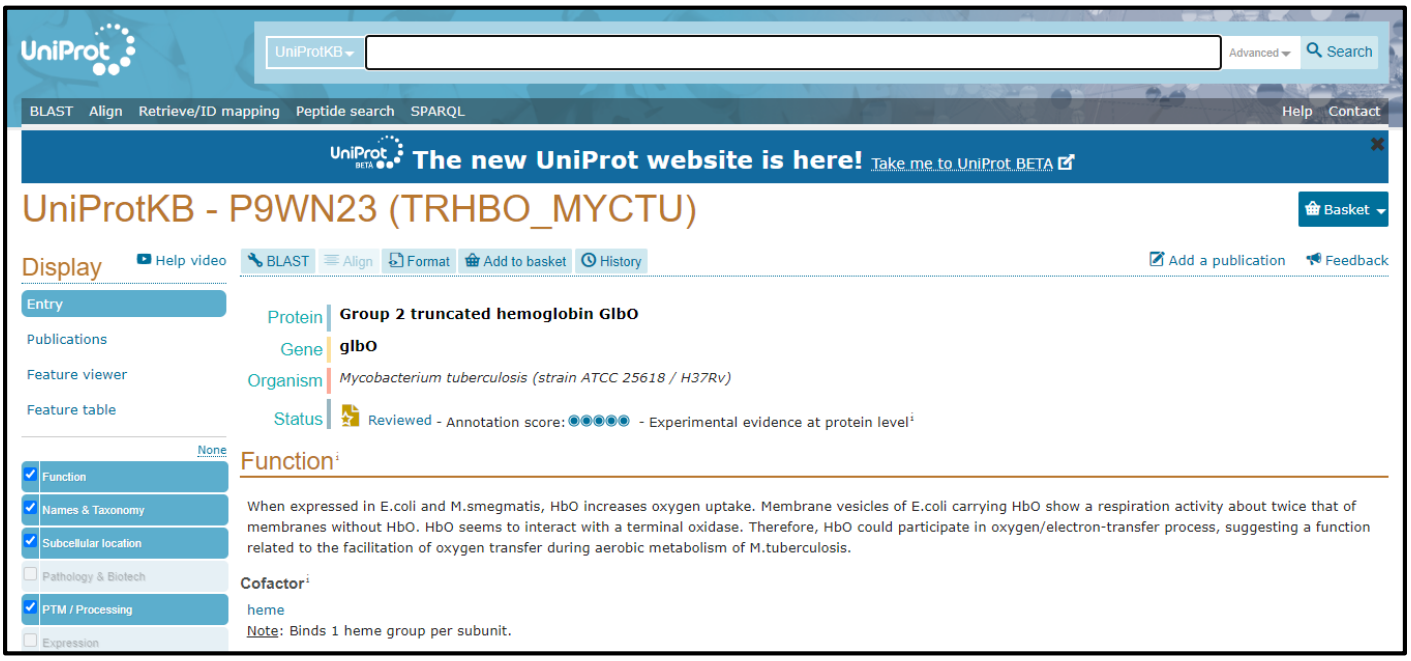


Fig1. UniProt result page for hemoglobin protein.

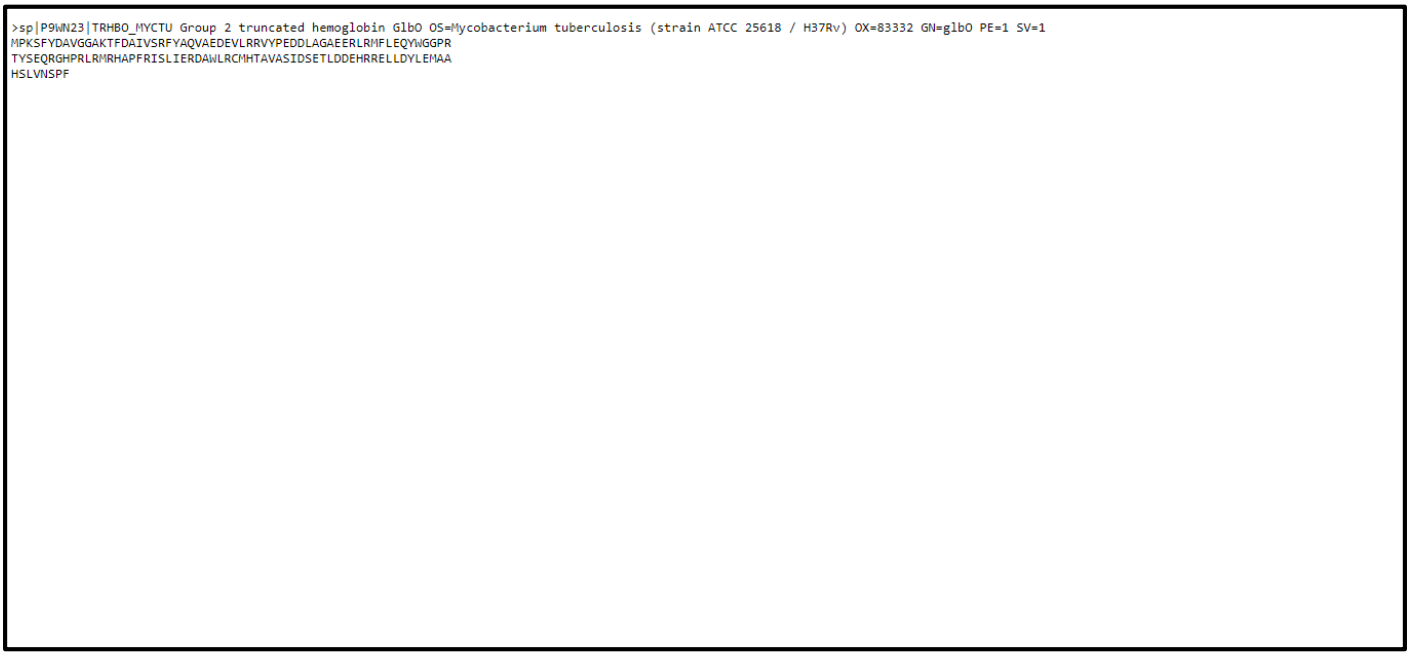


Fig2. Hemoglobin FASTA sequence from UniProt.

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

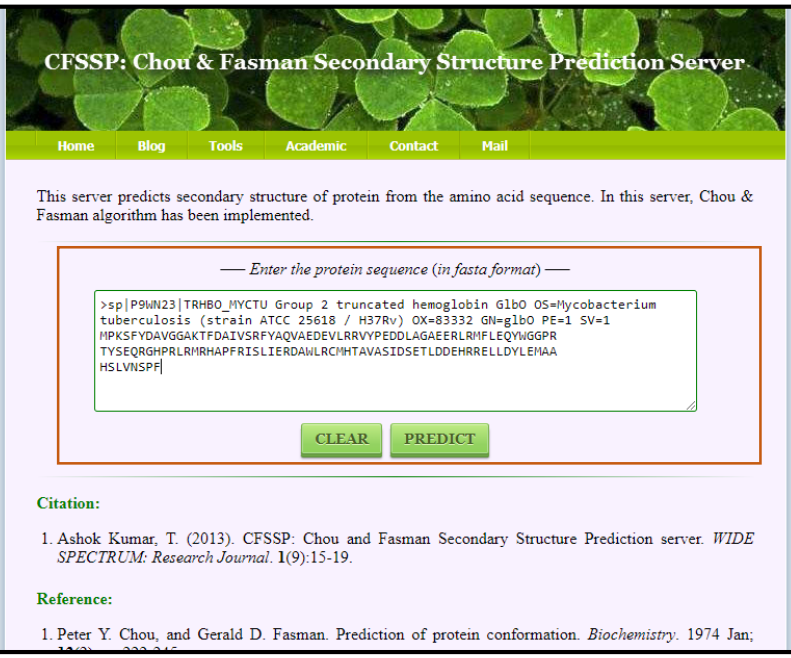


Fig4. Search page for Chou & Fasman Secondary Structure Prediction Server with protein FASTA sequence as input.

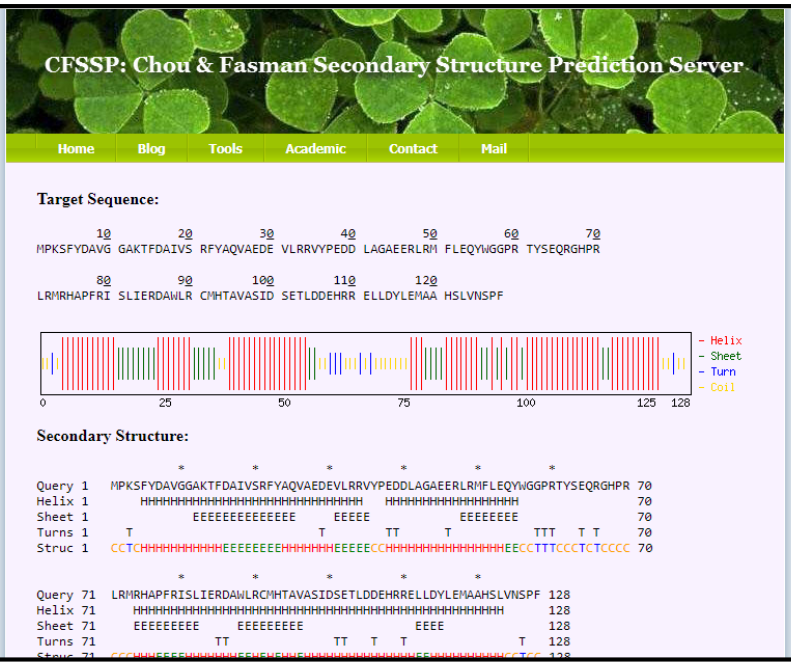


Fig5. Result page for Chou & Fasman Secondary Structure Prediction Server showing graphical representation.

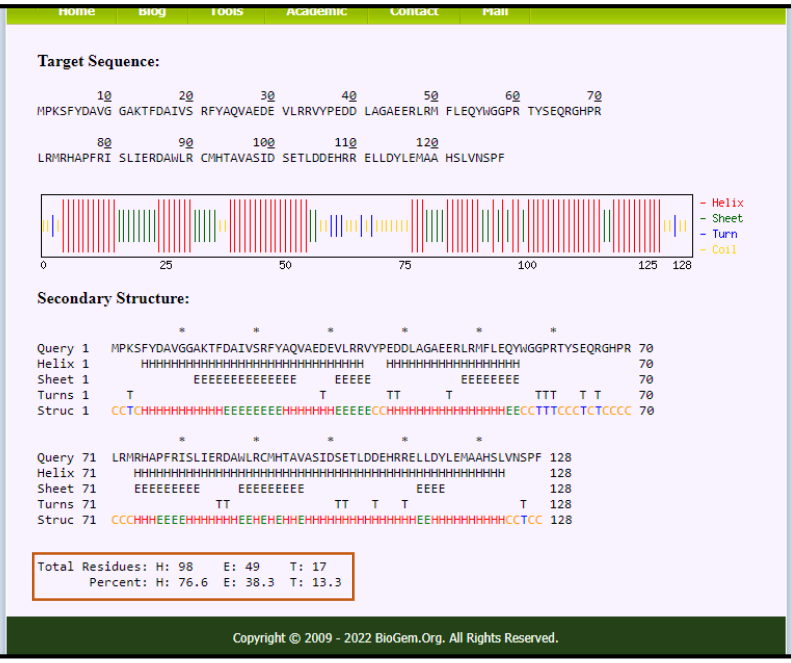


Fig5.1. Result page for Chou & Fasman Secondary Structure Prediction Server showing secondary structure information, total residues and their percentage in Hemoglobin.

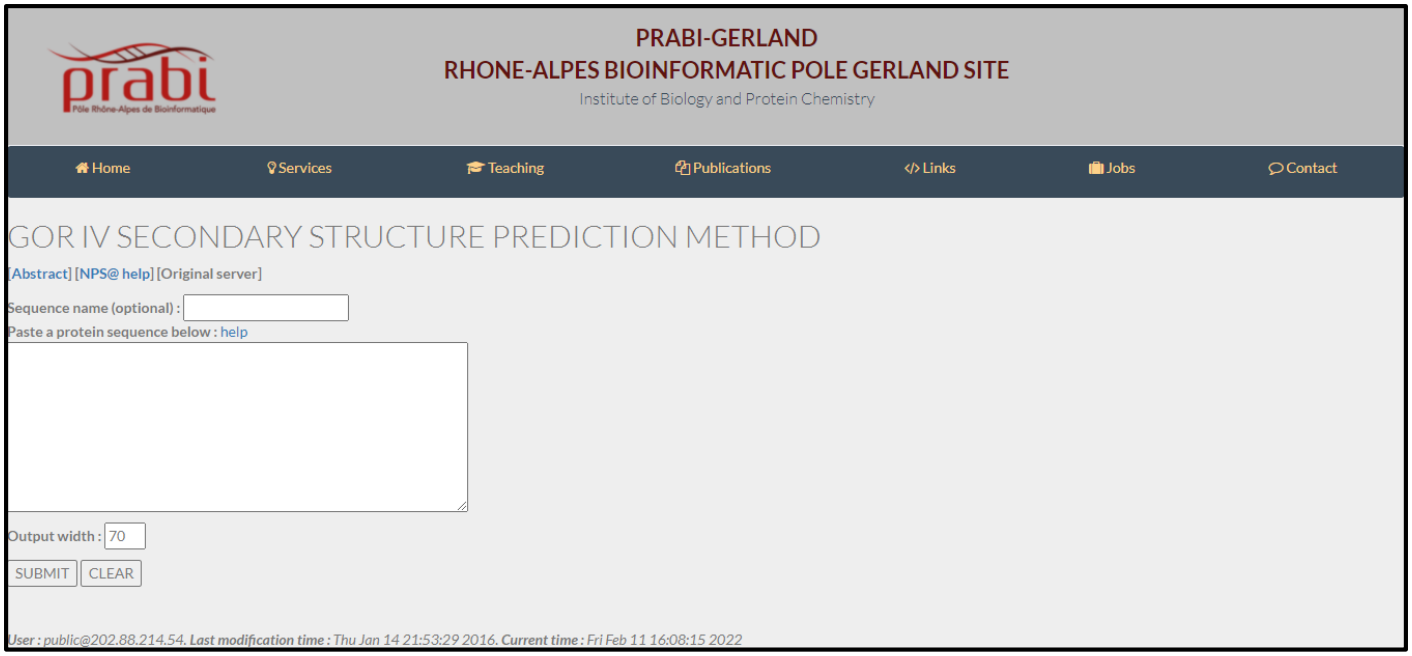


Fig6. Homepage for GOR IV Secondary Structure prediction method.

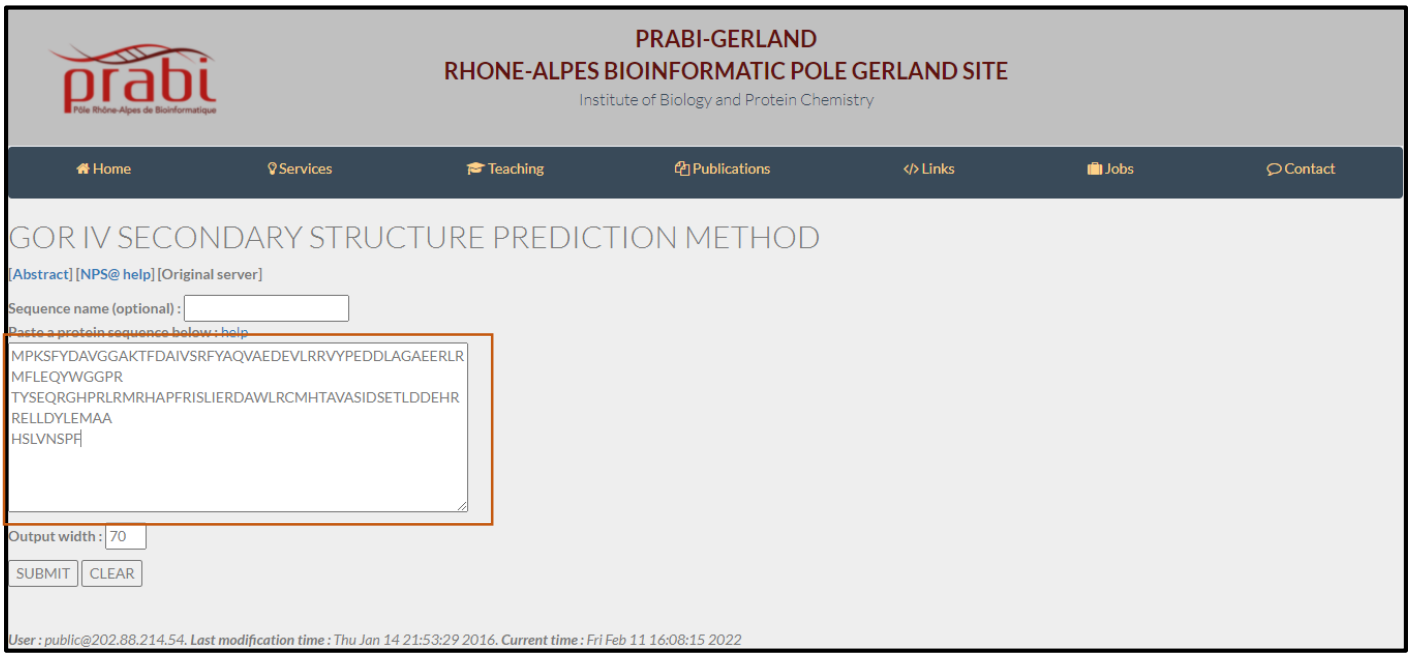


Fig7. Search page for GOR IV Secondary Structure prediction method with protein sequence as input.

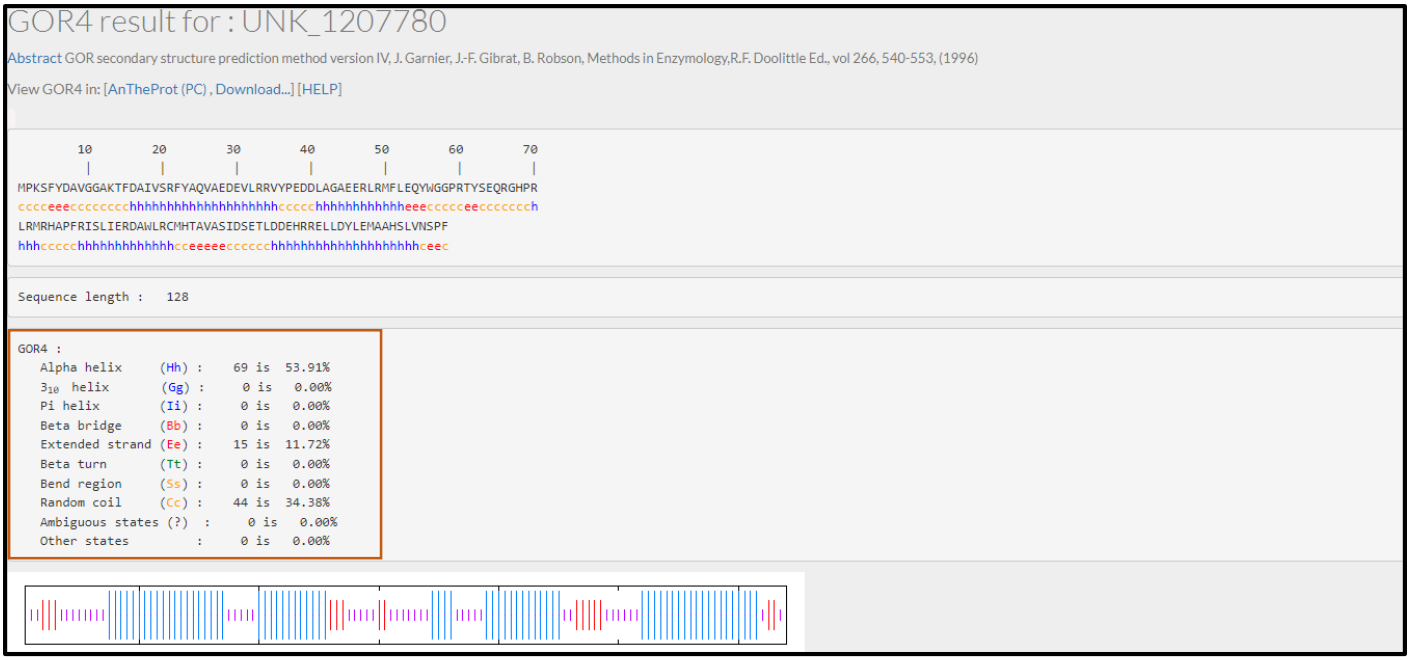


Fig8. Result page for GOR IV Secondary Structure prediction method showing total residues and their percentage in Hemoglobin.

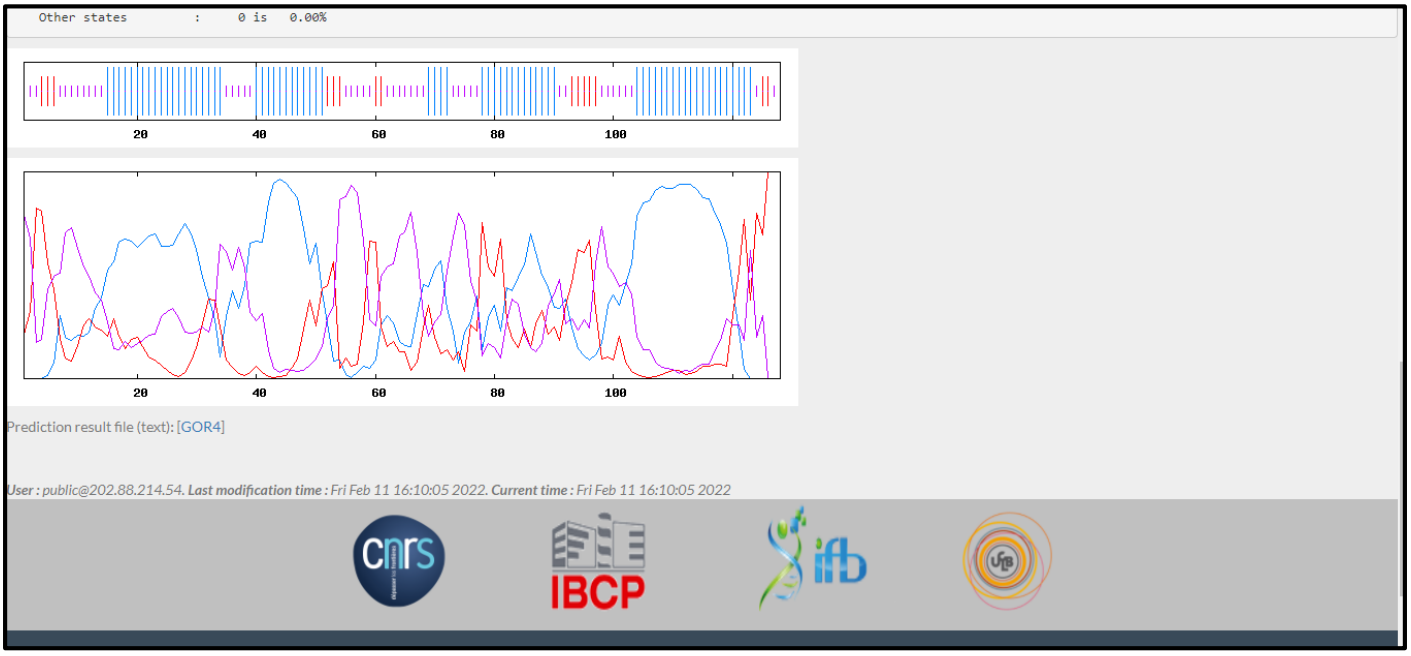


Fig8.1. Result page for GOR IV Secondary Structure prediction method showing graphical representation.

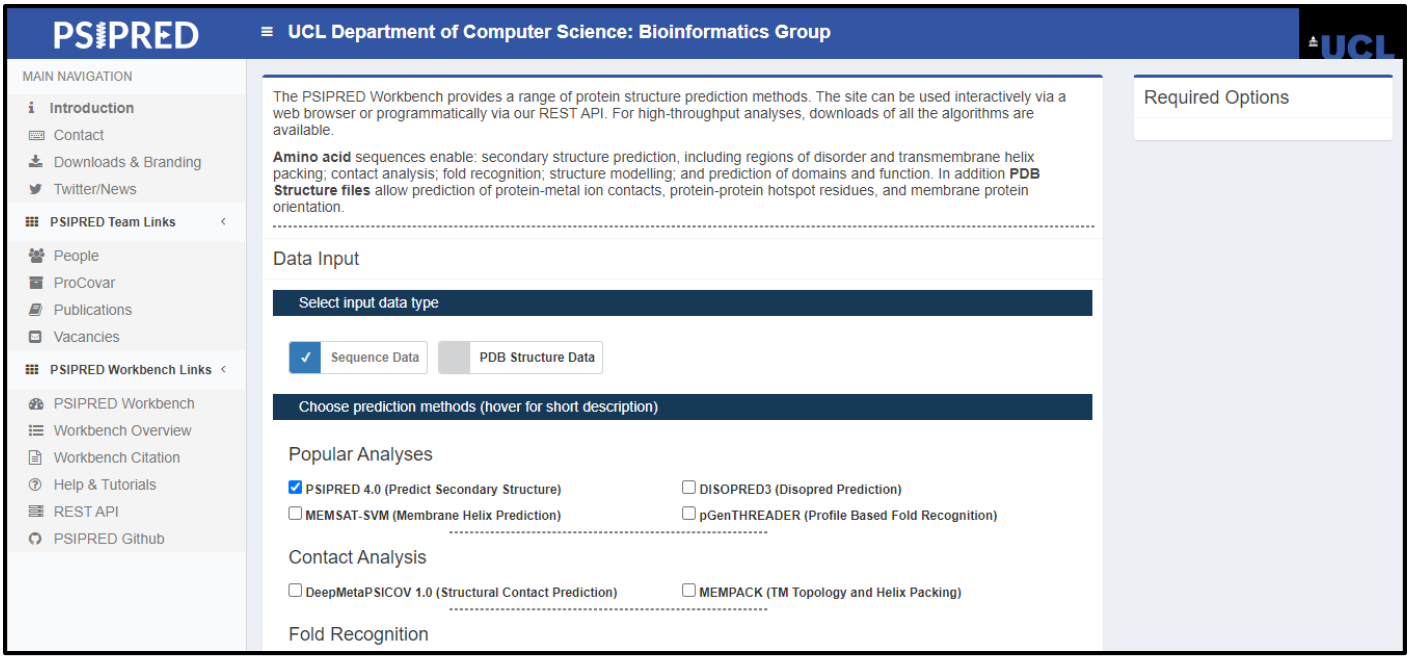


Fig9. Homepage for PSIPRED Protein Structure Prediction Server.

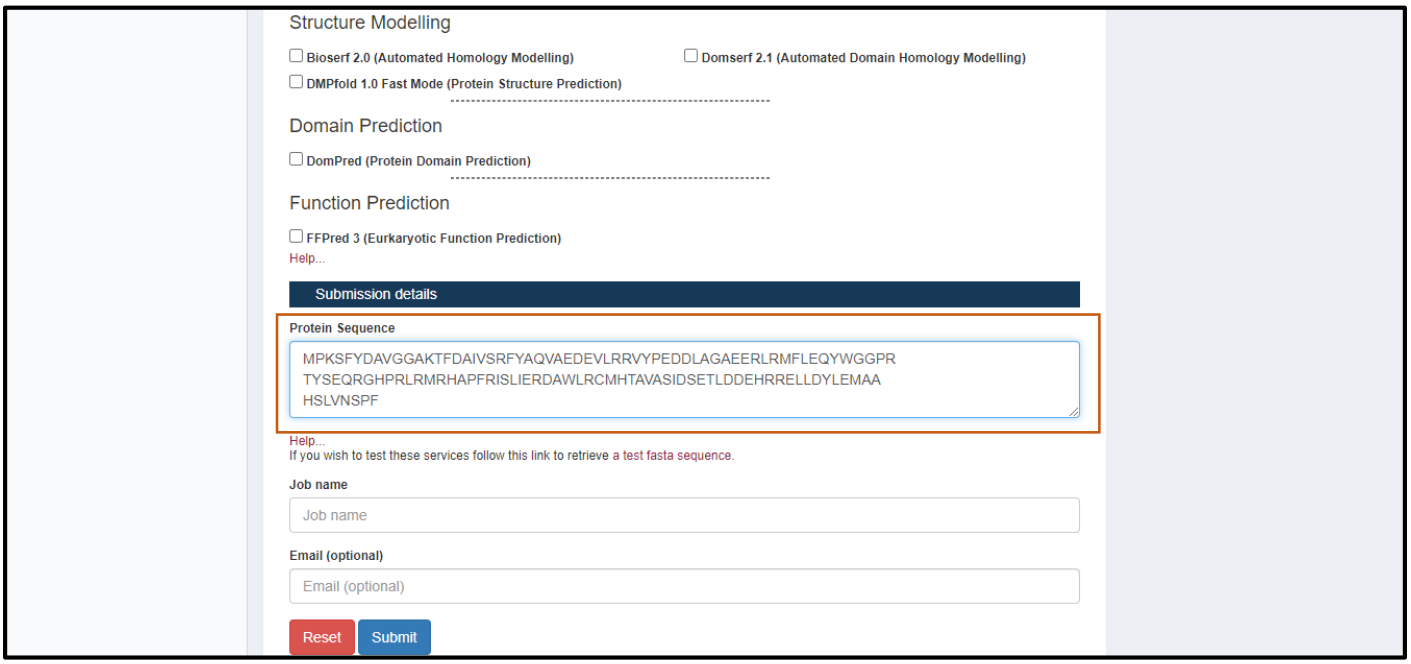


Fig10. Search page for PSIPRED Protein Structure Prediction Server with protein sequence as input.

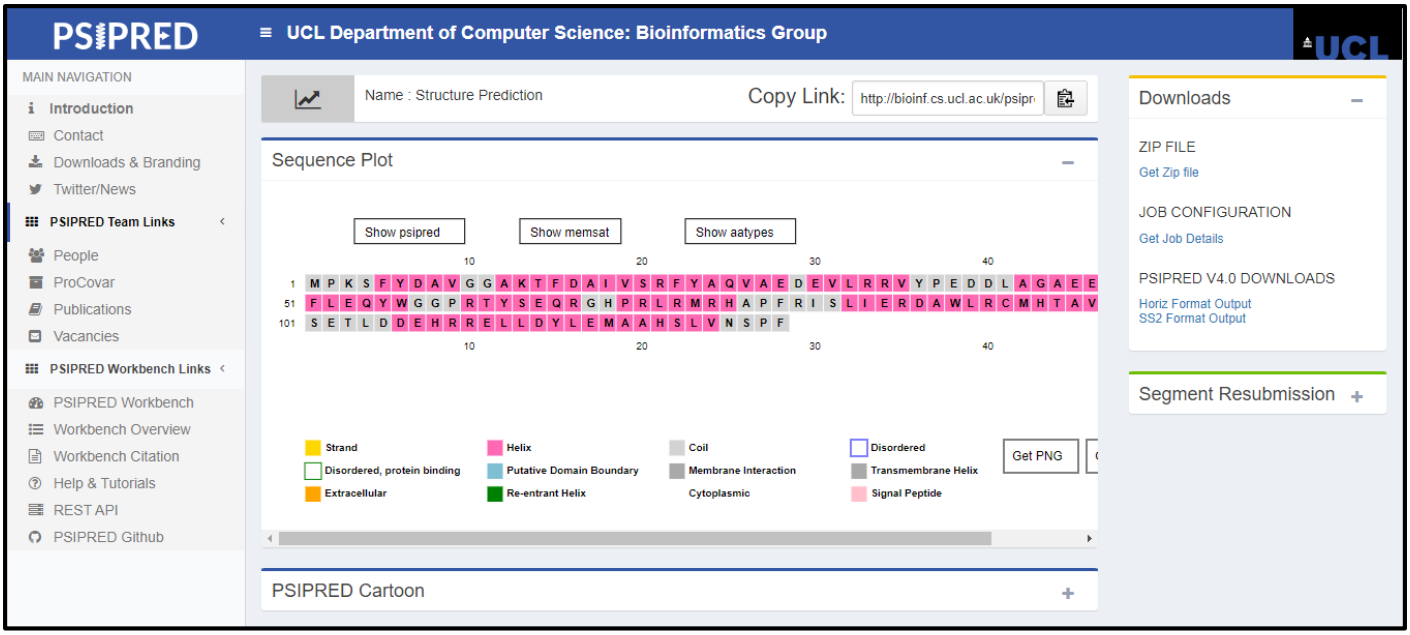


Fig11. Result page for PSIPRED Protein Structure Prediction Server.

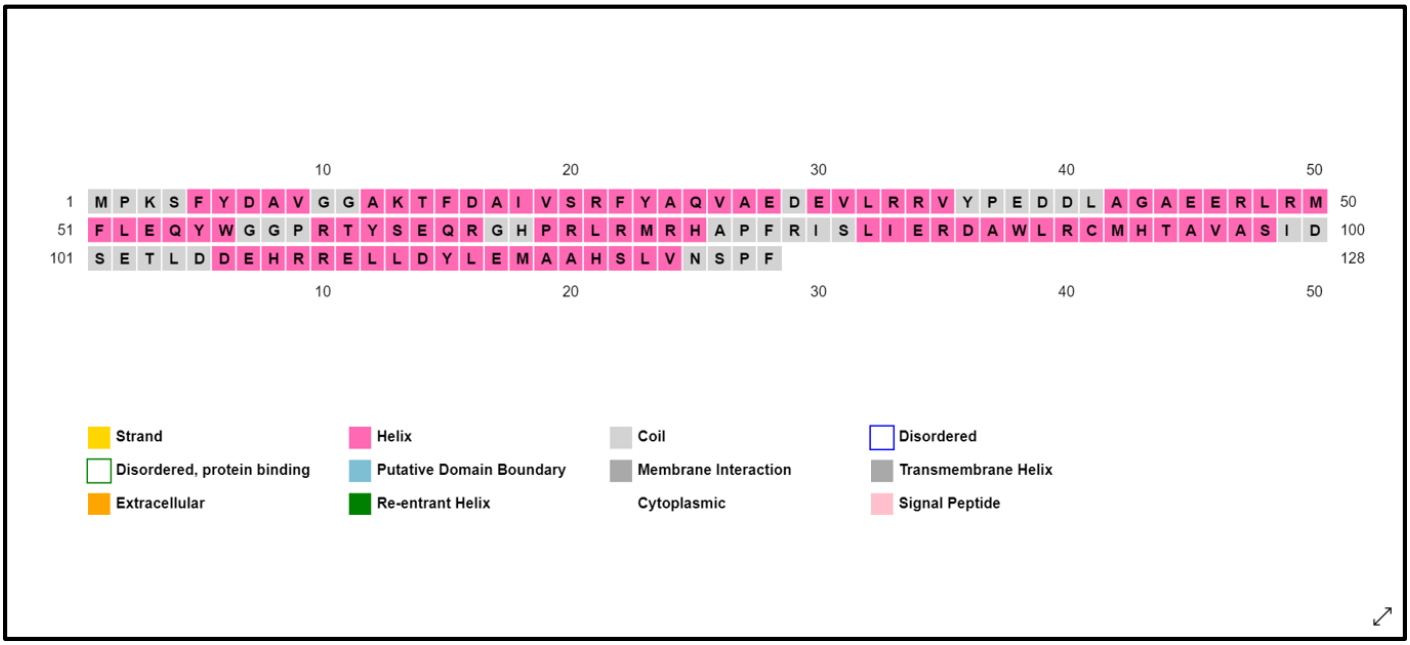


Fig11.1 Result page for PSIPRED Protein Structure Prediction Server showing helices and coils present in Hemoglobin.

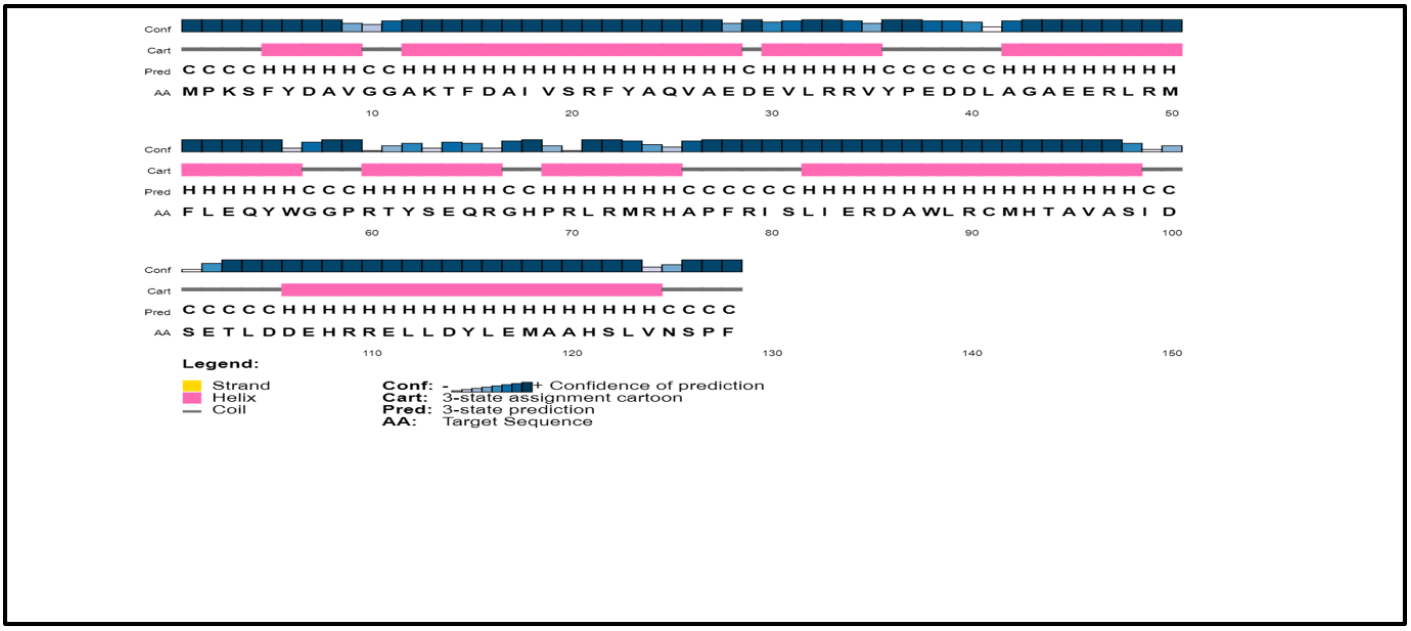


Fig11.2 Result page for PSIPRED Protein Structure Prediction Server showing helices and coils present in Hemoglobin with confidence of prediction.

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 3_{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
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. 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. 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
9. *NPS@ : GOR4 secondary structure prediction*. (2021). Ibcp.fr. Retrieved February 11, 2022, from 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

DATE: 16/02/22

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), 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 α, all β, α and β, 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

DATE: 16/02/22

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:

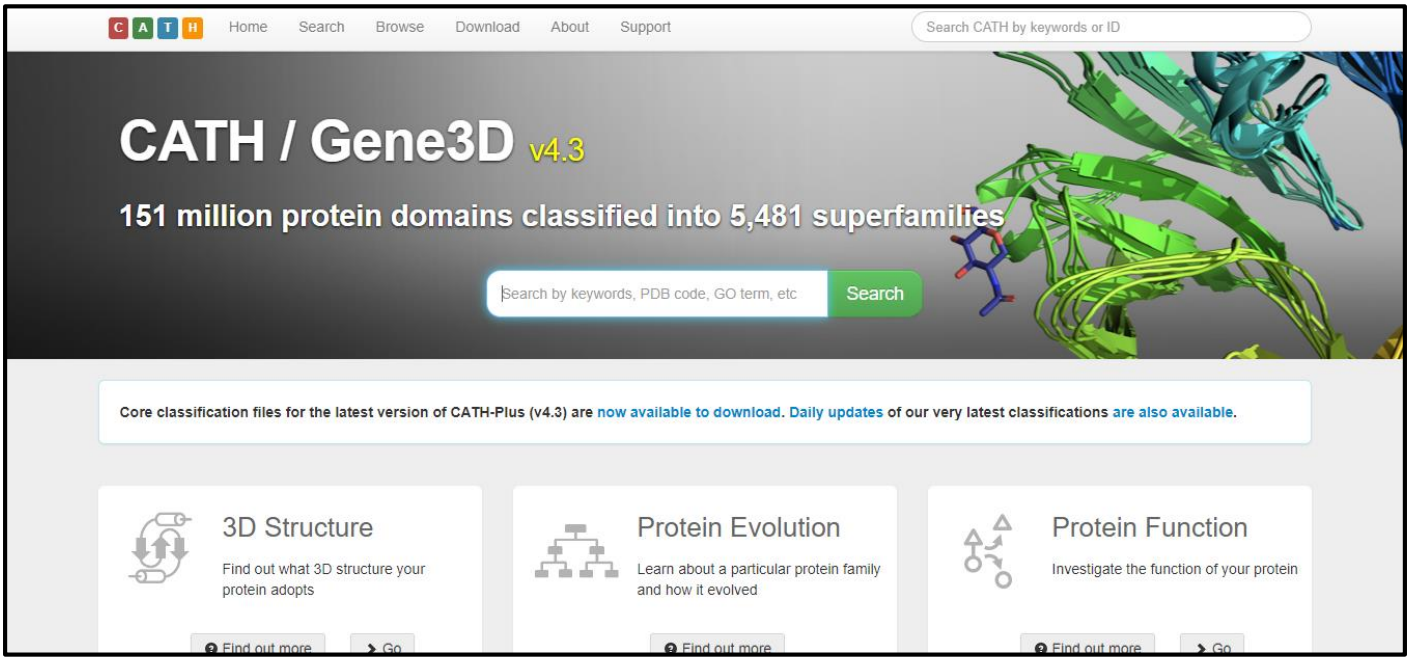


Fig1. Homepage for CATH database

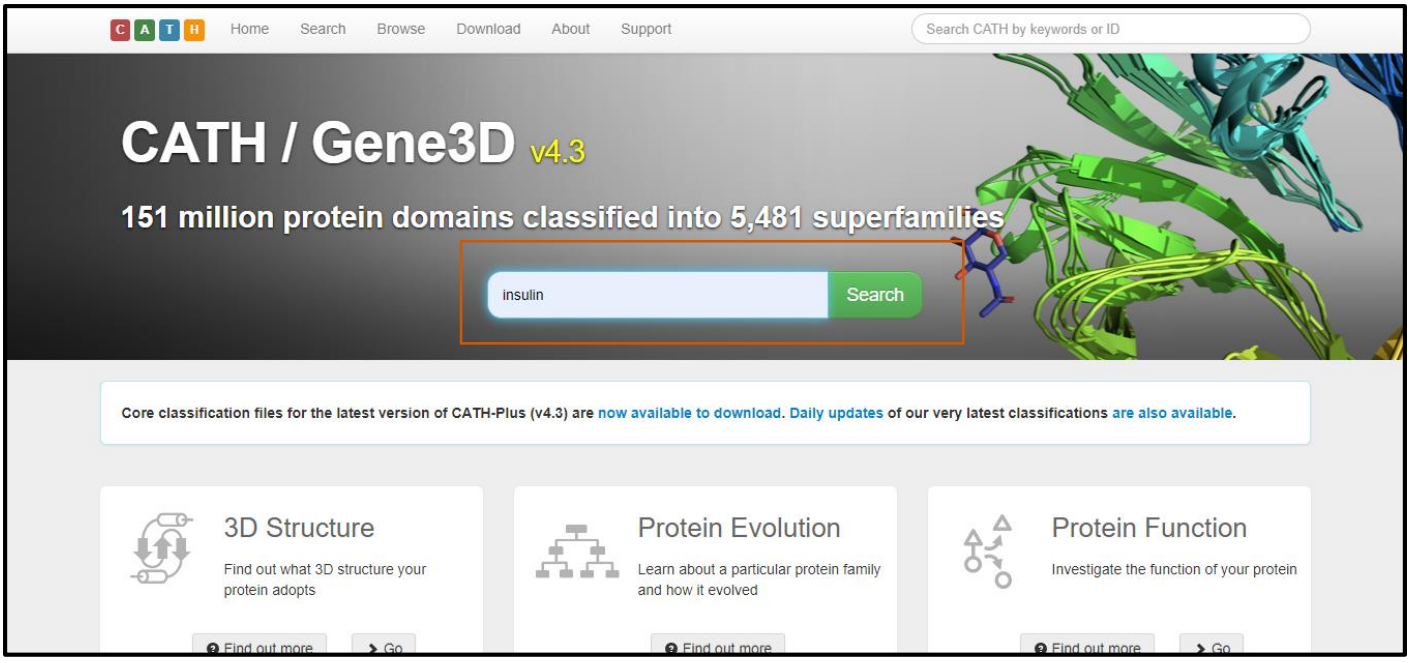


Fig2. Search for protein insulin

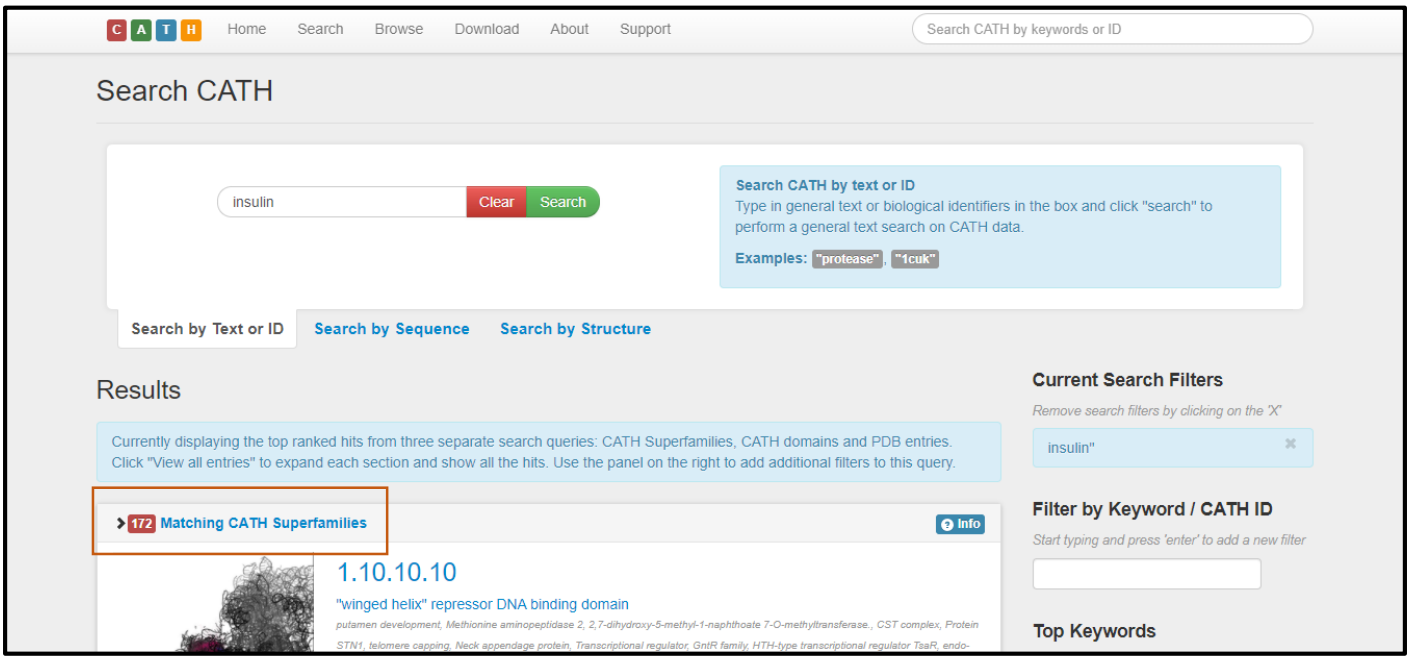


Fig3. Hit page for insulin showing matching CATH Superfamilies

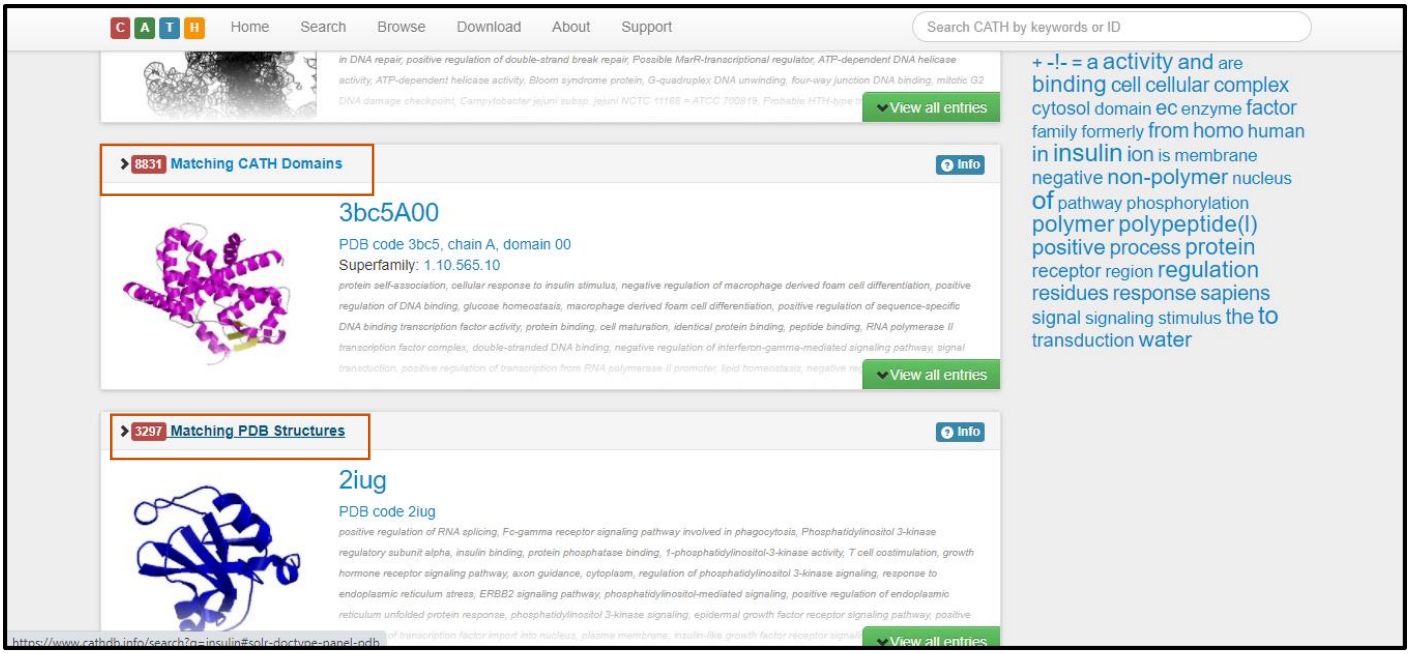


Fig3.1. Hit page for insulin showing matching CATH Domains and PDB structures

4) Insulin matching CATH Superfamilies

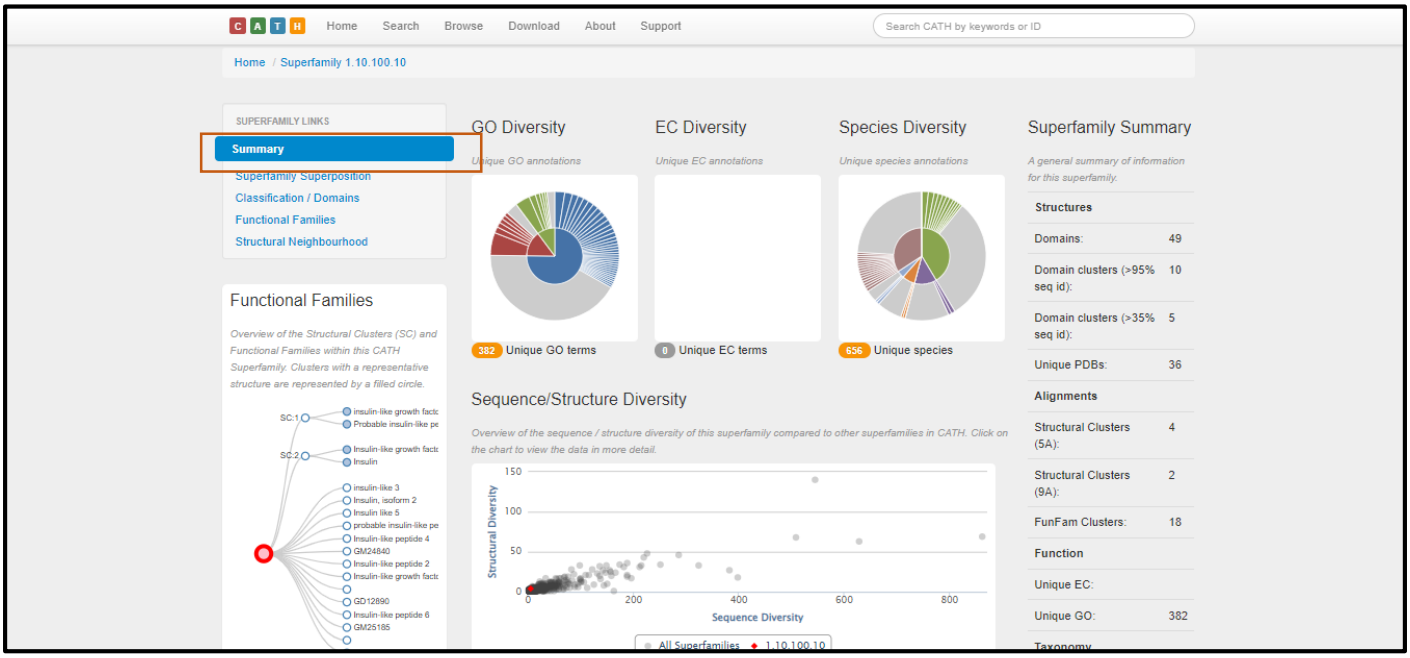


Fig4.1. Result page for Summary

Fig4.2. Result page for Superfamily Superposition

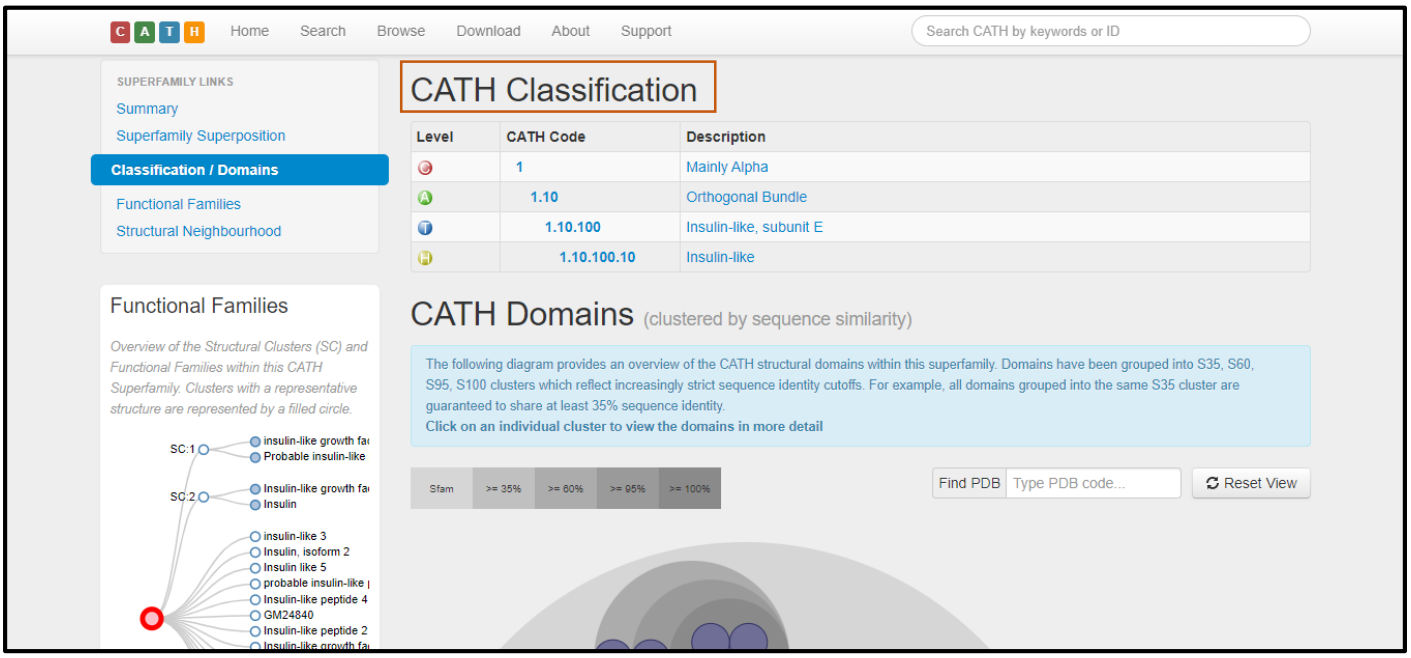


Fig4.3. Result page for Classification/Domains

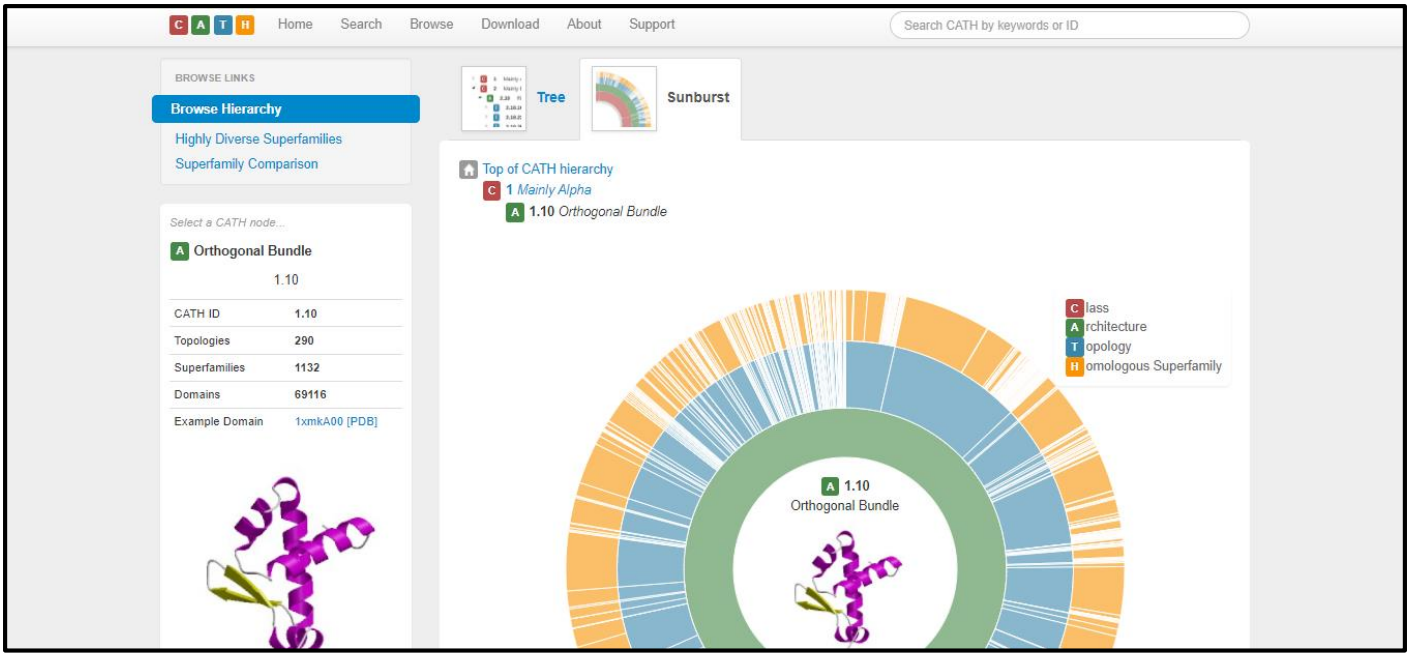


Fig4.4. Result page for Orthogonal Bundle (Sunburst) under CATH Classification

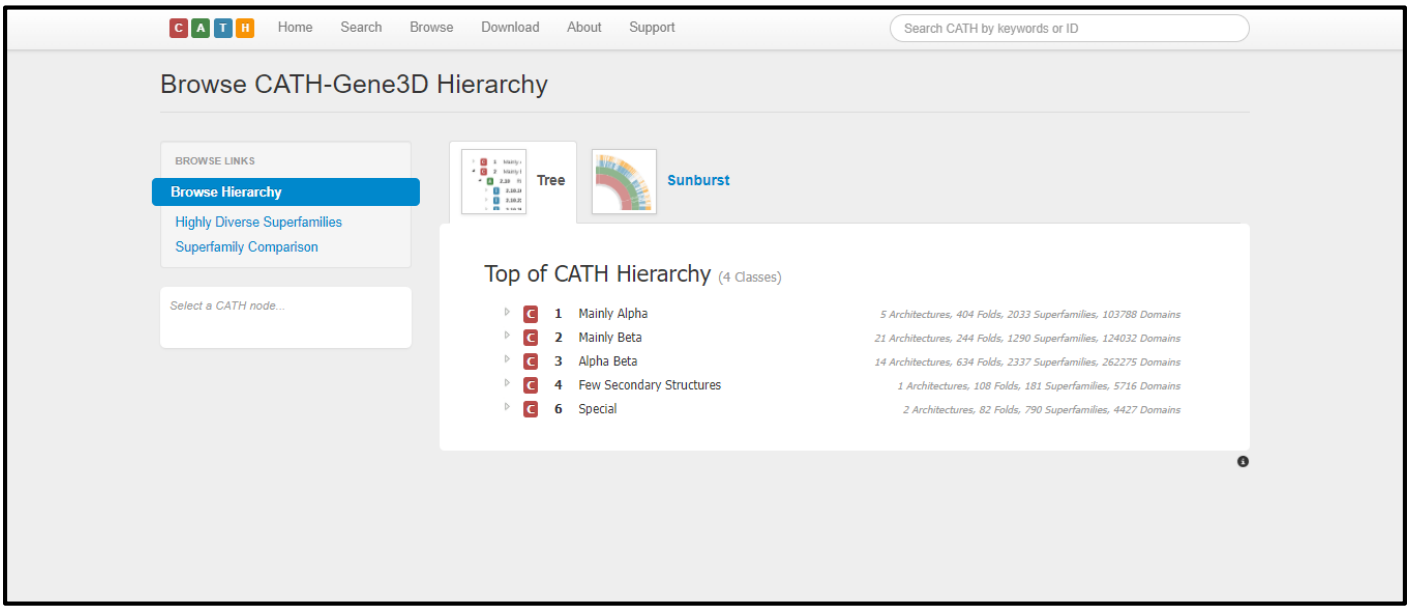


Fig4.5. Result page for Orthogonal Bundle (Tree) under CATH Classification

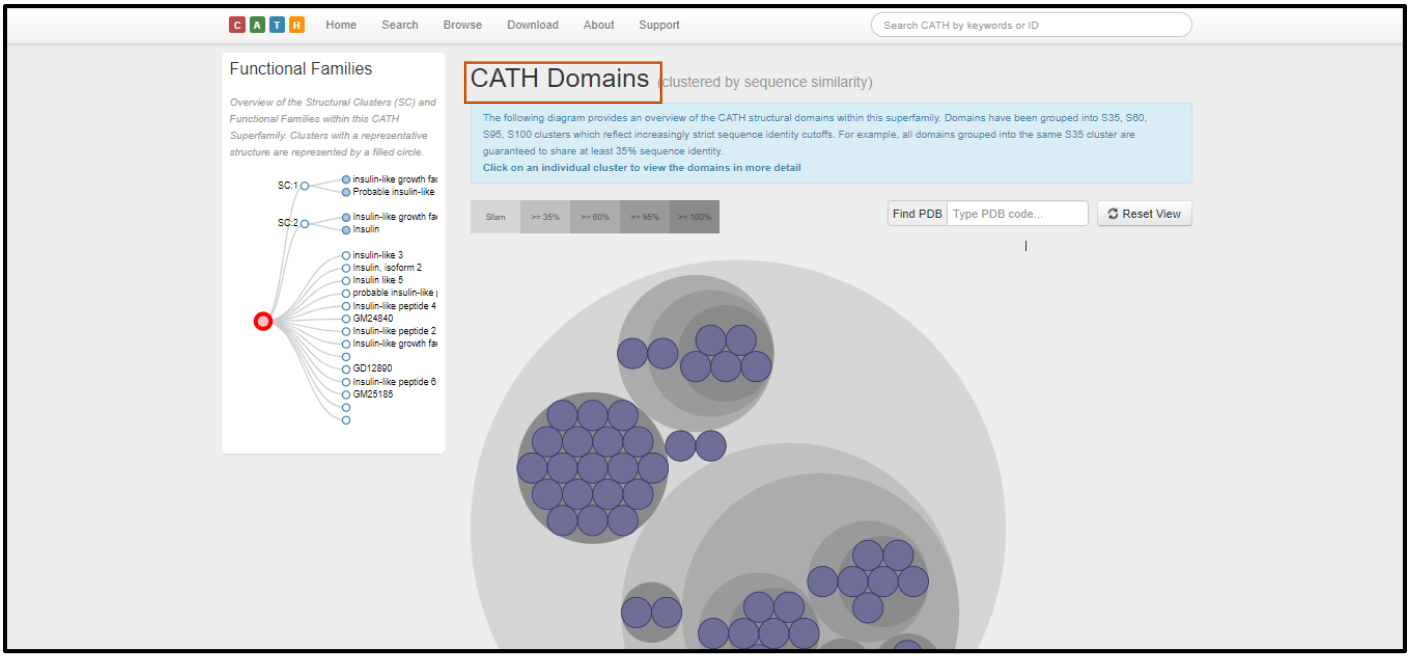
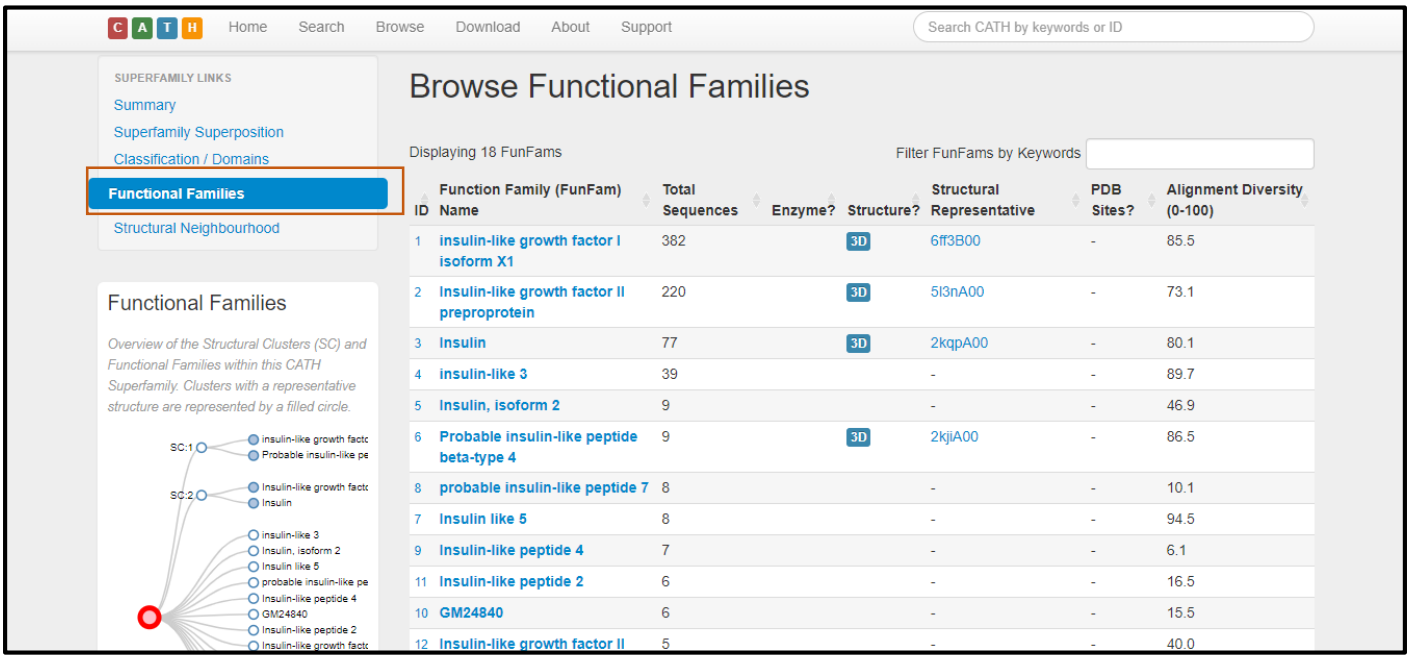


Fig4.6. Result page for CATH domains



ID	Function Family (FunFam) Name	Total Sequences	Enzyme?	Structure?	Structural Representative	PDB Sites?	Alignment Diversity (0-100)
1	insulin-like growth factor I isoform X1	382		3D	6ff3B00	-	85.5
2	Insulin-like growth factor II preproprotein	220		3D	5l3nA00	-	73.1
3	Insulin	77		3D	2kqpA00	-	80.1
4	insulin-like 3	39			-	-	89.7
5	Insulin, isoform 2	9			-	-	46.9
6	Probable insulin-like peptide beta-type 4	9		3D	2kjiA00	-	86.5
8	probable insulin-like peptide 7	8			-	-	10.1
7	Insulin like 5	8			-	-	94.5
9	Insulin-like peptide 4	7			-	-	6.1
11	Insulin-like peptide 2	6			-	-	16.5
10	GM24840	6			-	-	15.5
12	Insulin-like growth factor II	5			-	-	40.0

Fig4.7. Result page for Functional Families

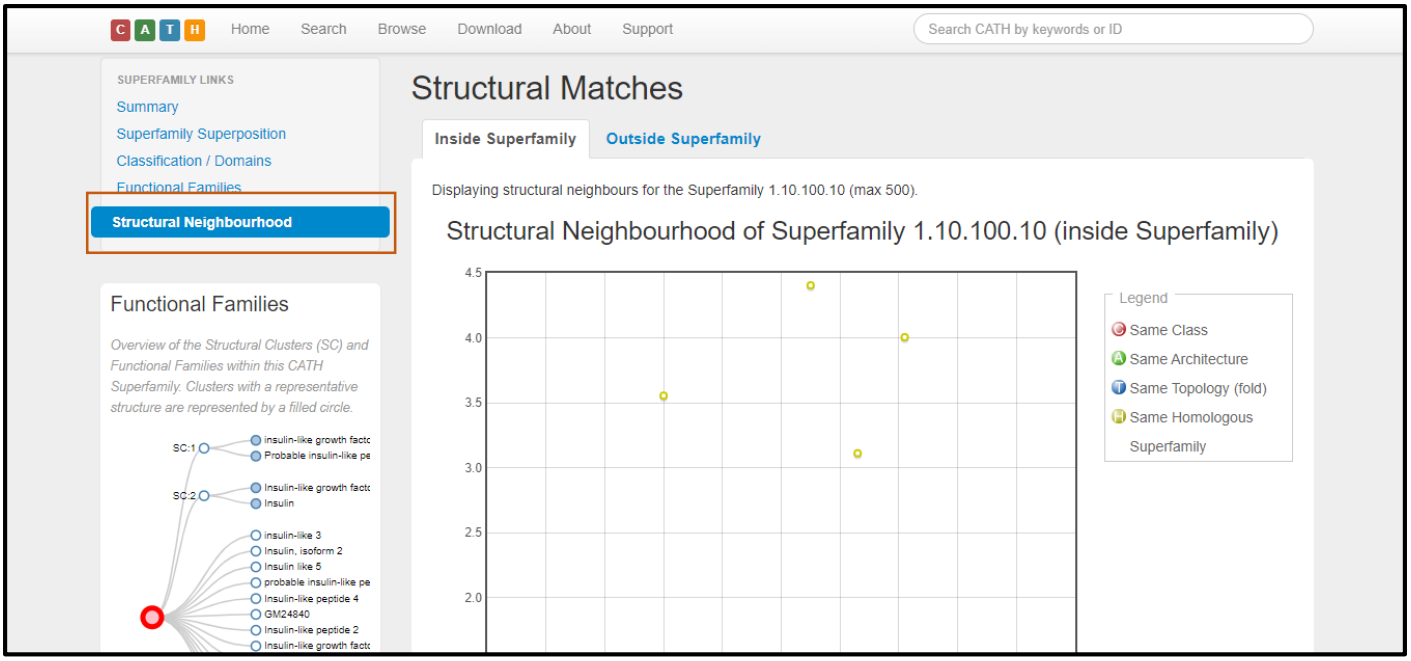


Fig4.8. Result page for Structural Neighbourhood

5) Insulin matching CATH domains

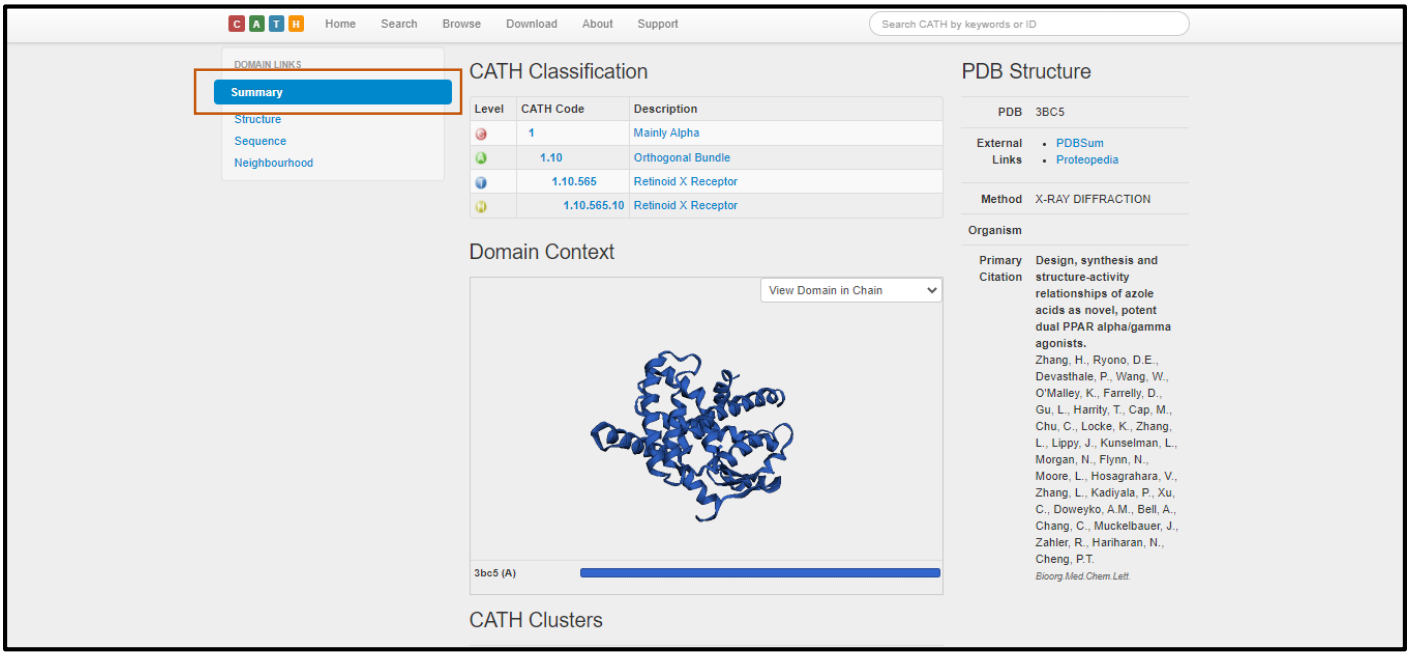


Fig5.1. Result page for Summary

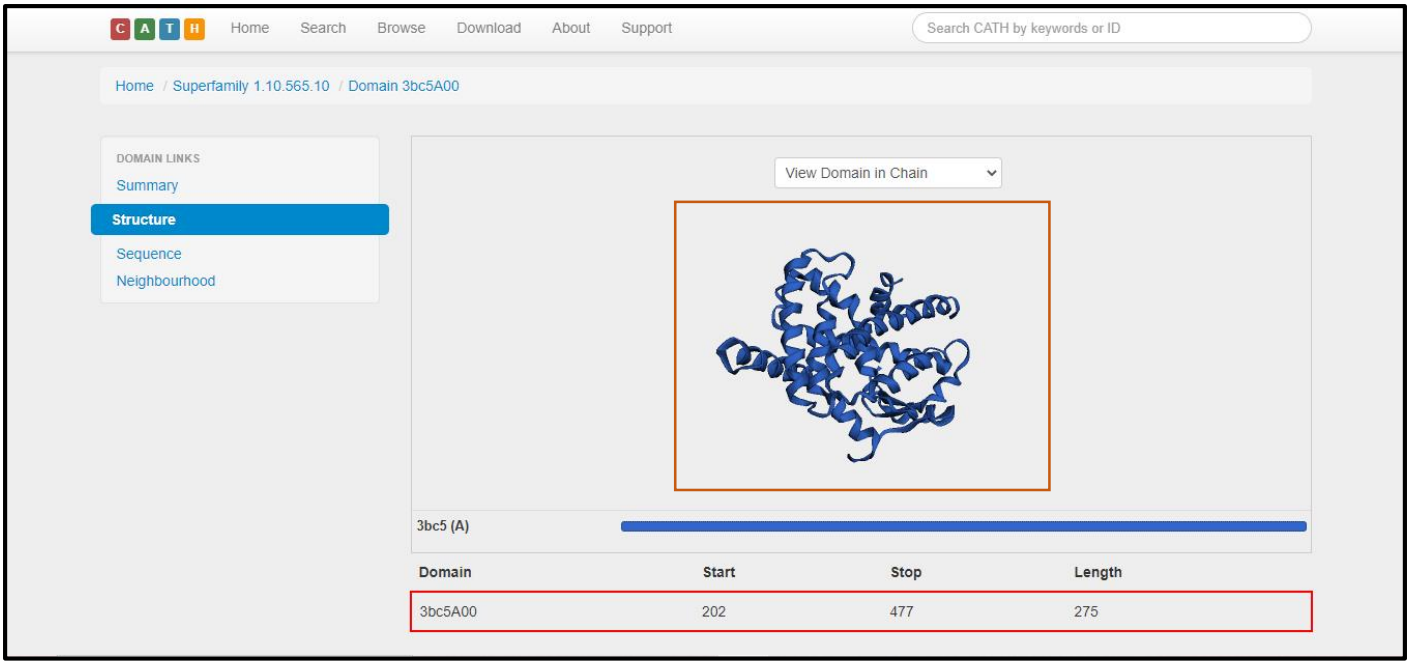


Fig5.2. Result page for Structure

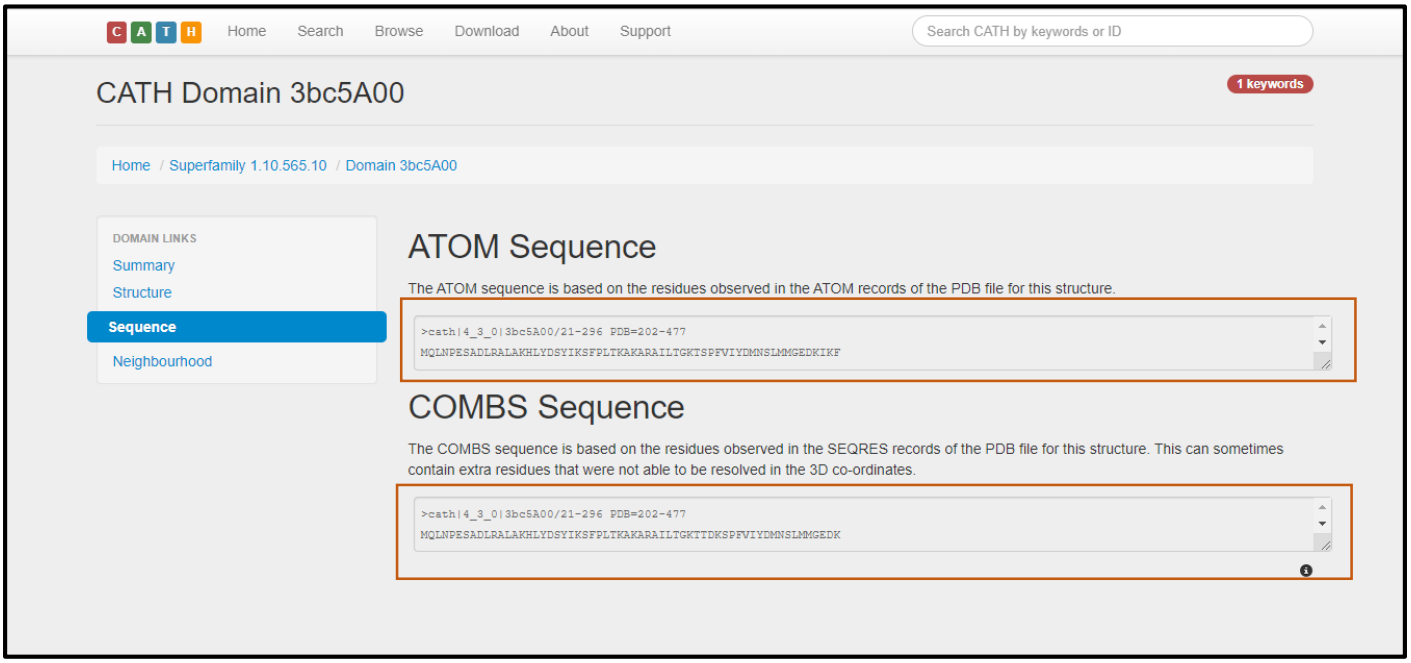


Fig5.3. Result page for Sequence

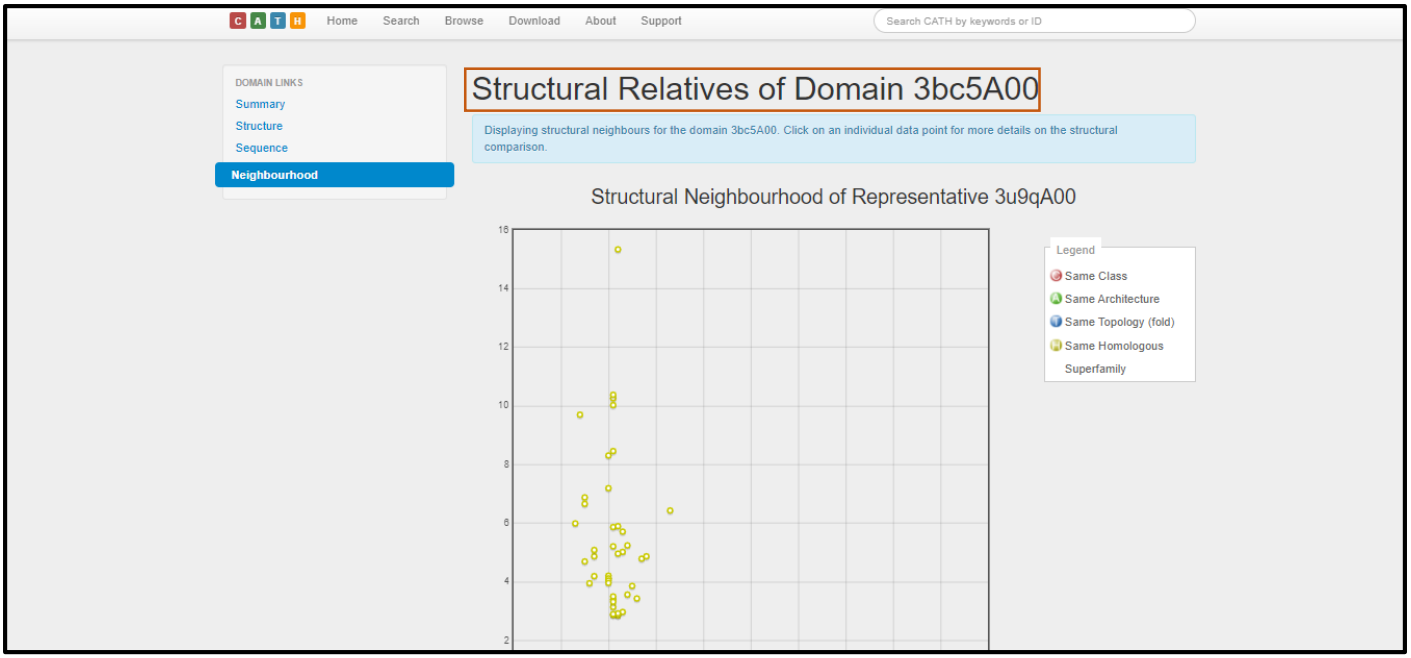


Fig5.4. Result page for Neighbourhood

6) Insulin matching PDB structures

Fig6.1. Result page for PDB Information

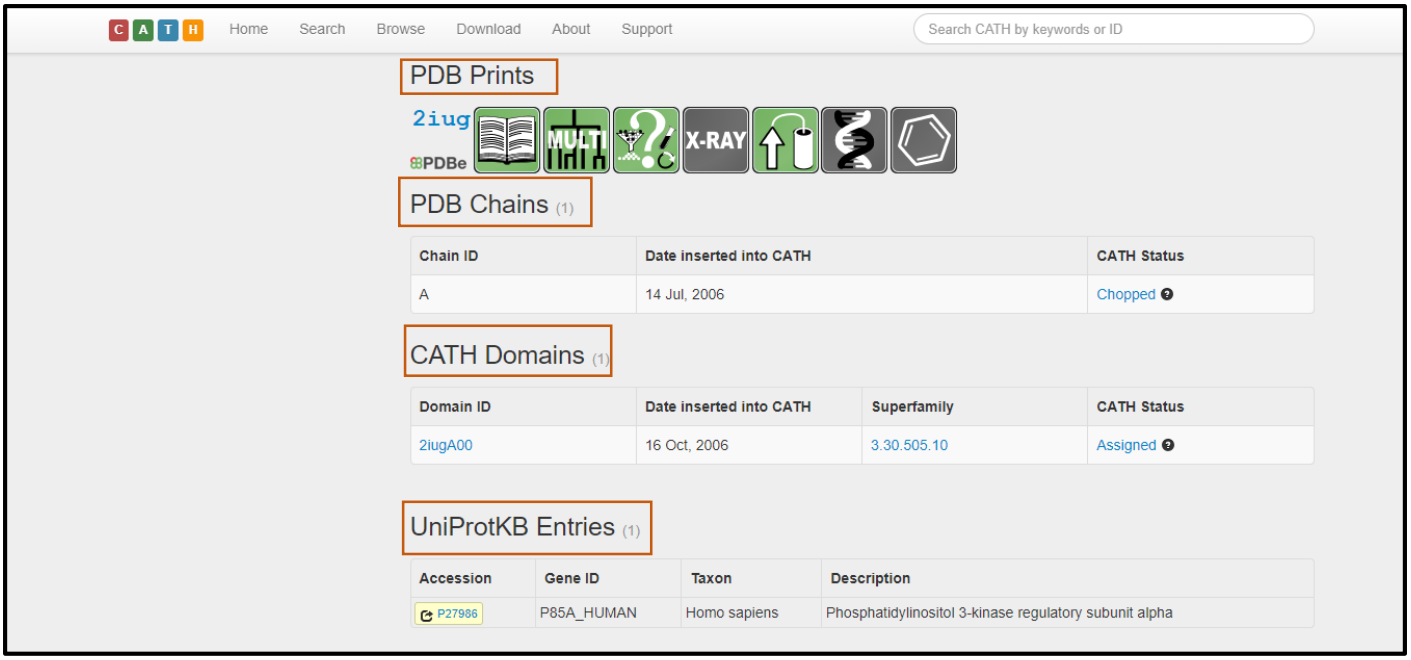


Fig6.2. Result page for PDB Prints, Chains, CATH Domains and UniPortKB Entries

SCOP database:

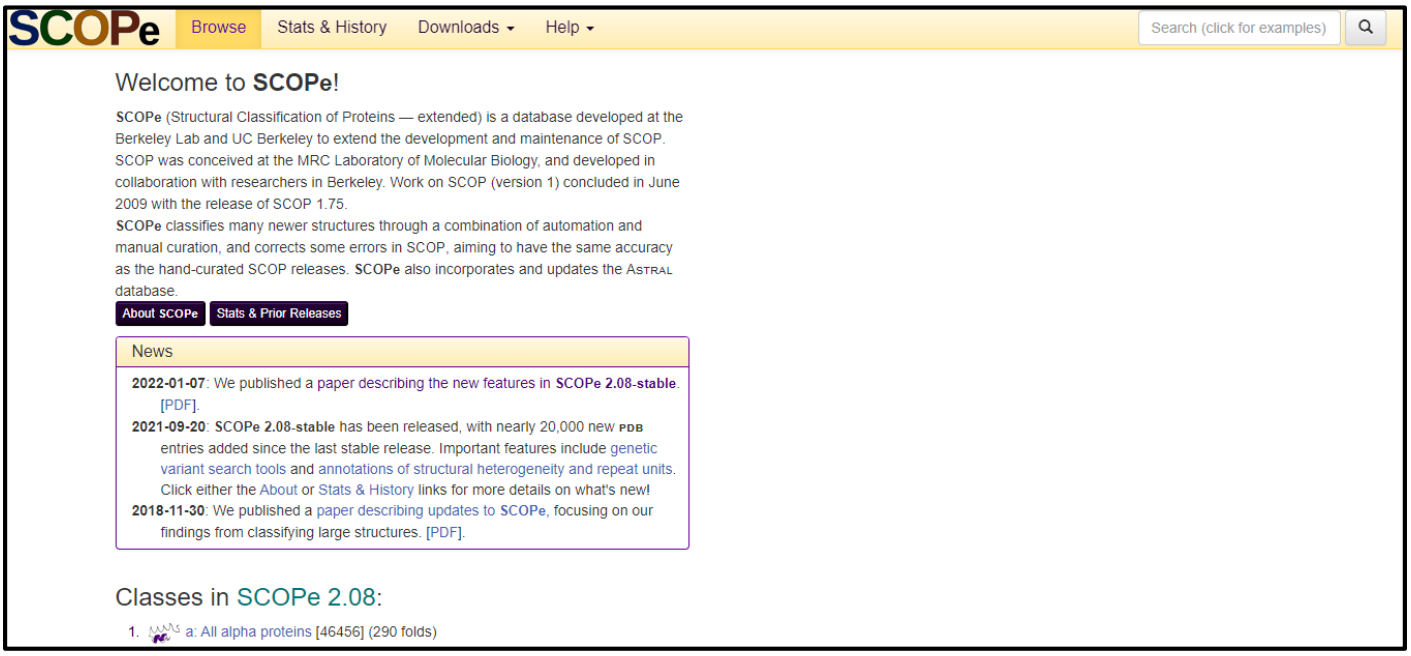


Fig1. Homepage for SCOPe database

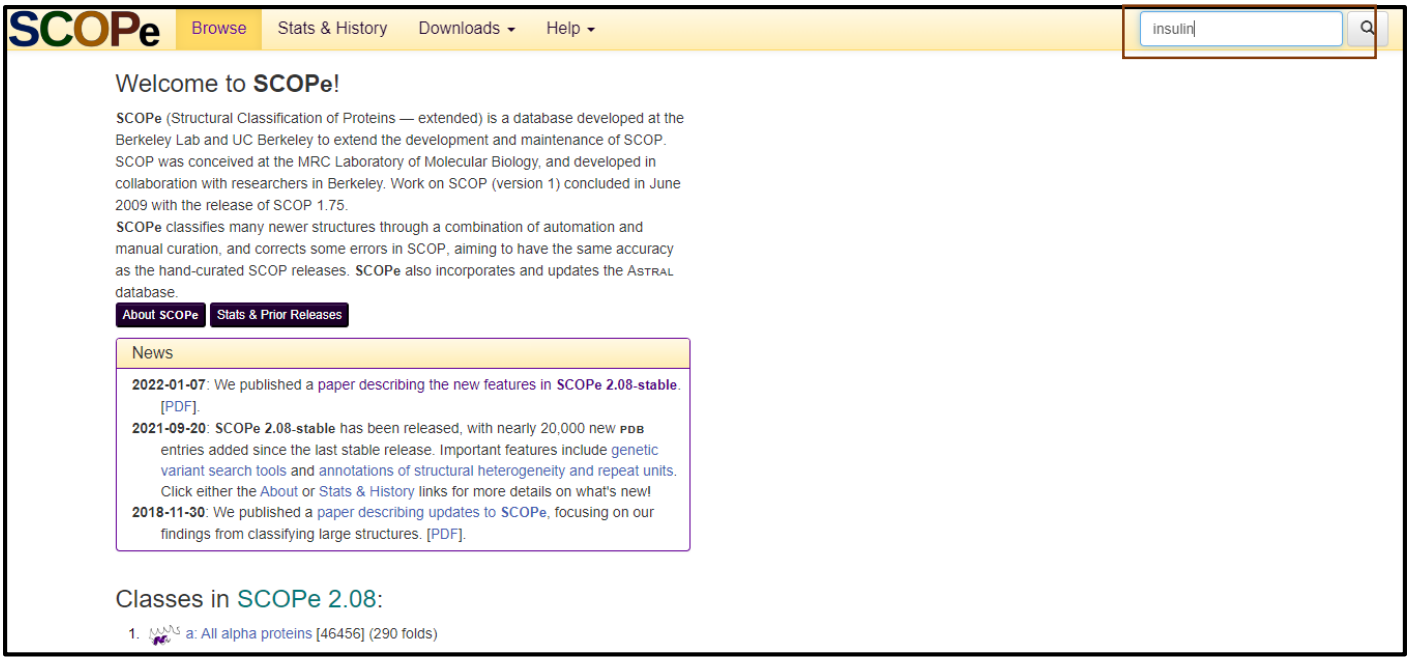


Fig2. Search for Insulin

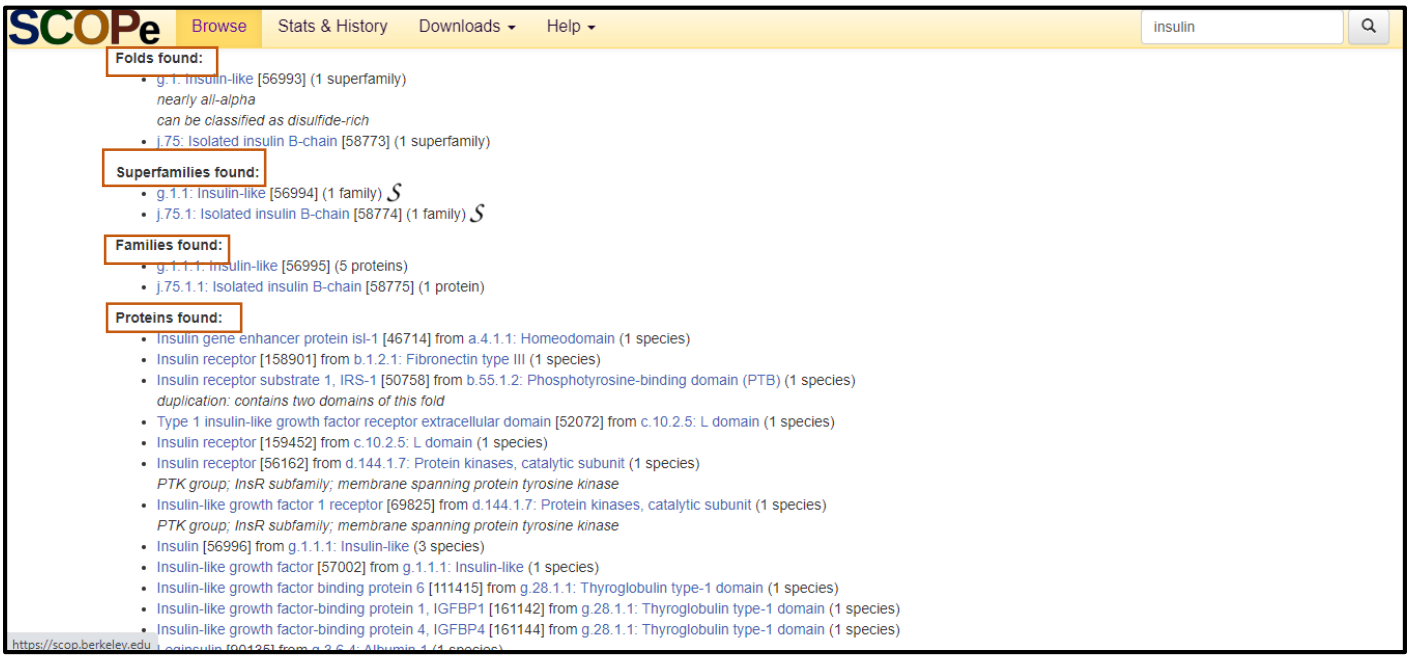


Fig3.1. Hit page for Insulin

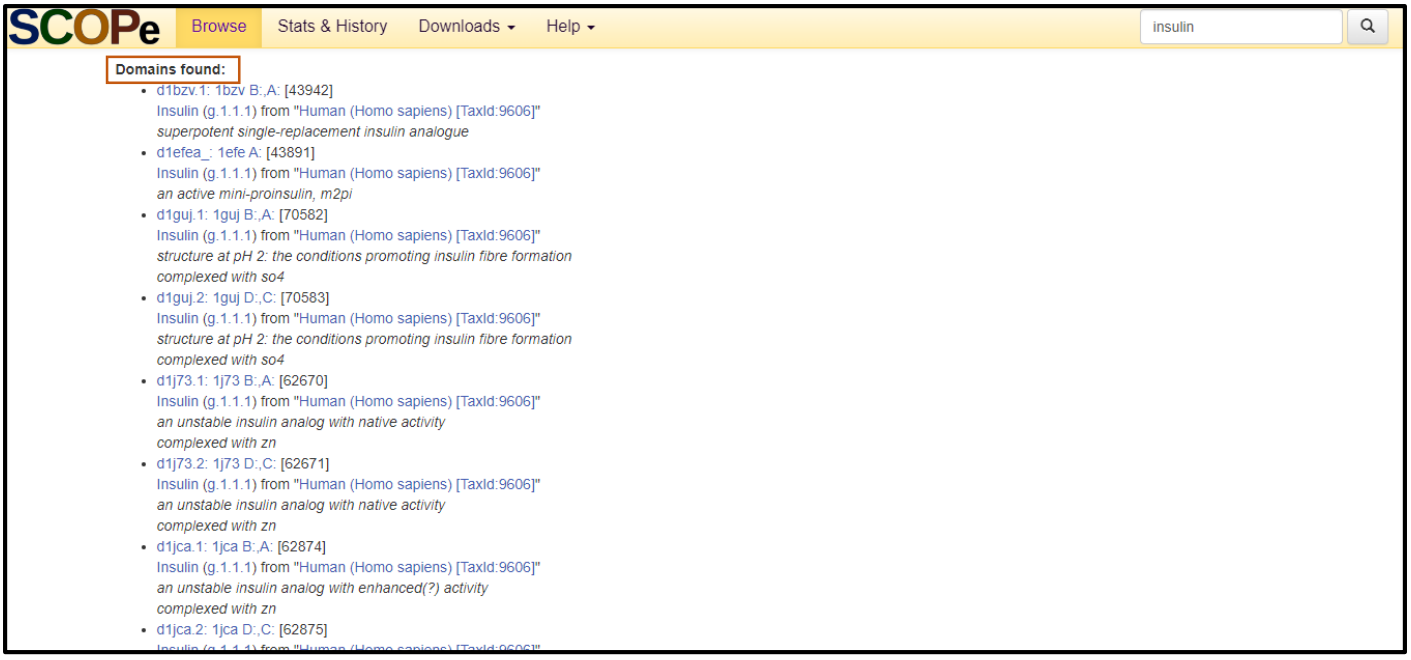


Fig3.2. Hit page for Insulin

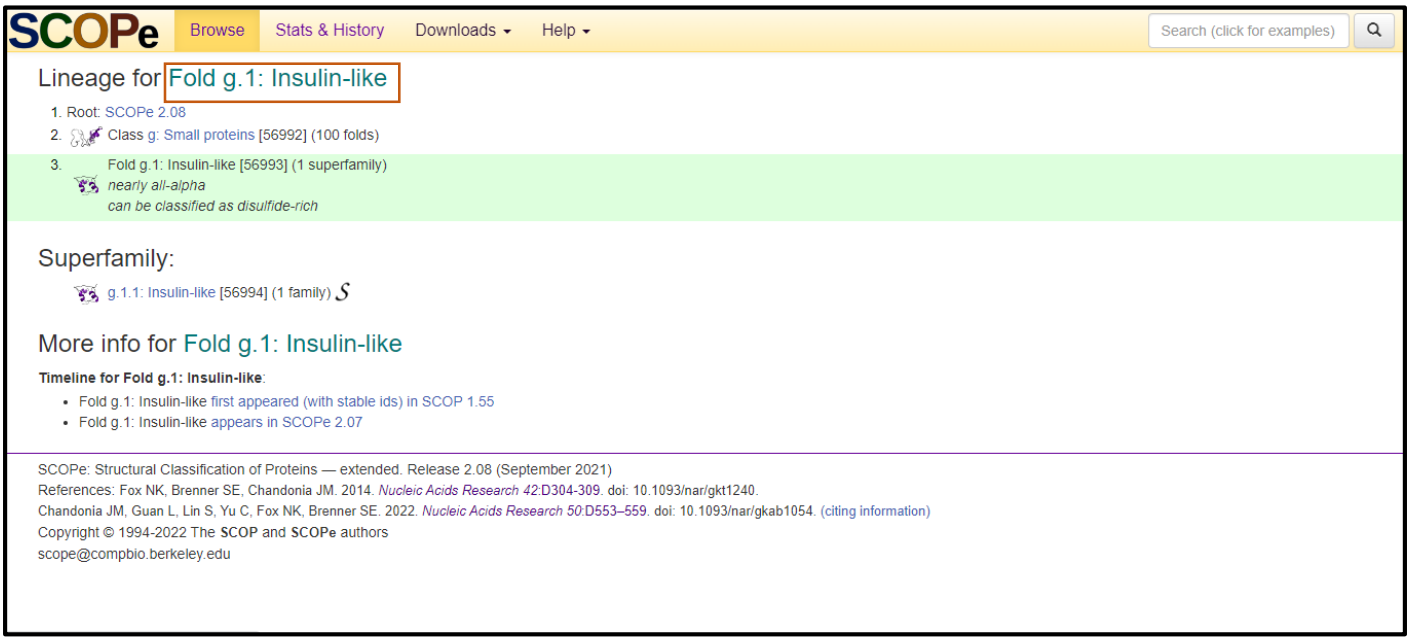


Fig4. Result page for Protein Fold

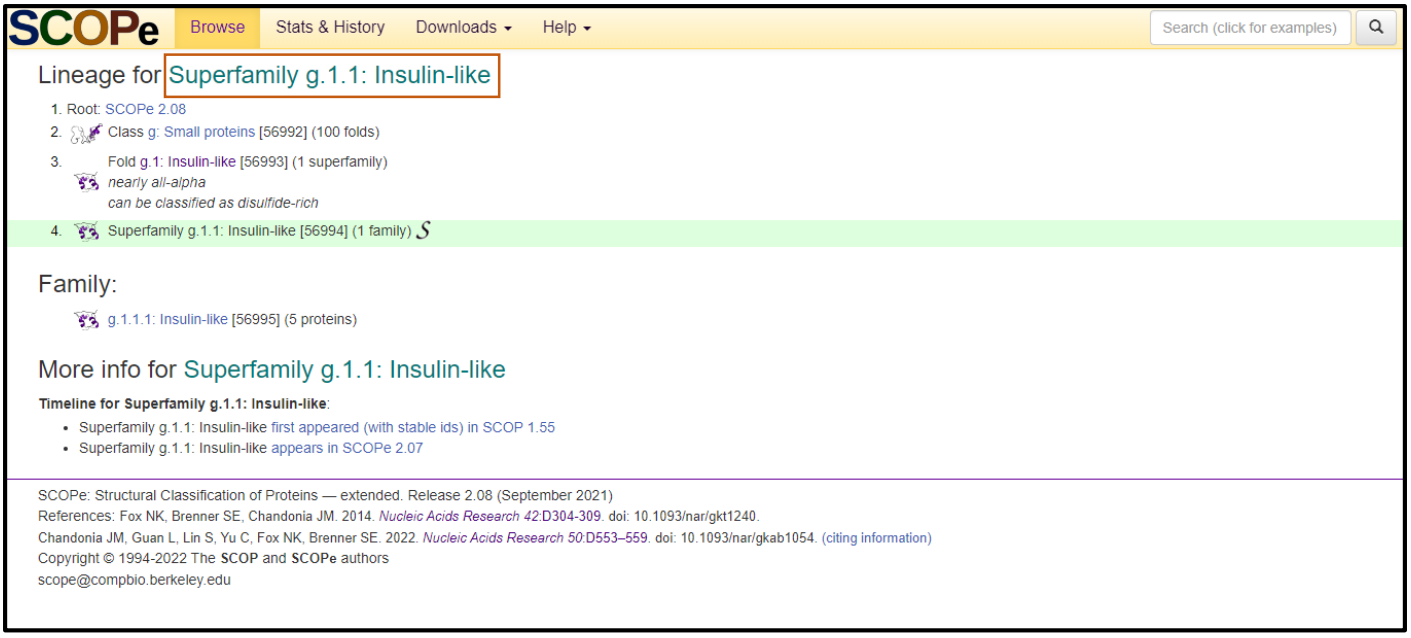


Fig5. Result page for Protein Superfamily

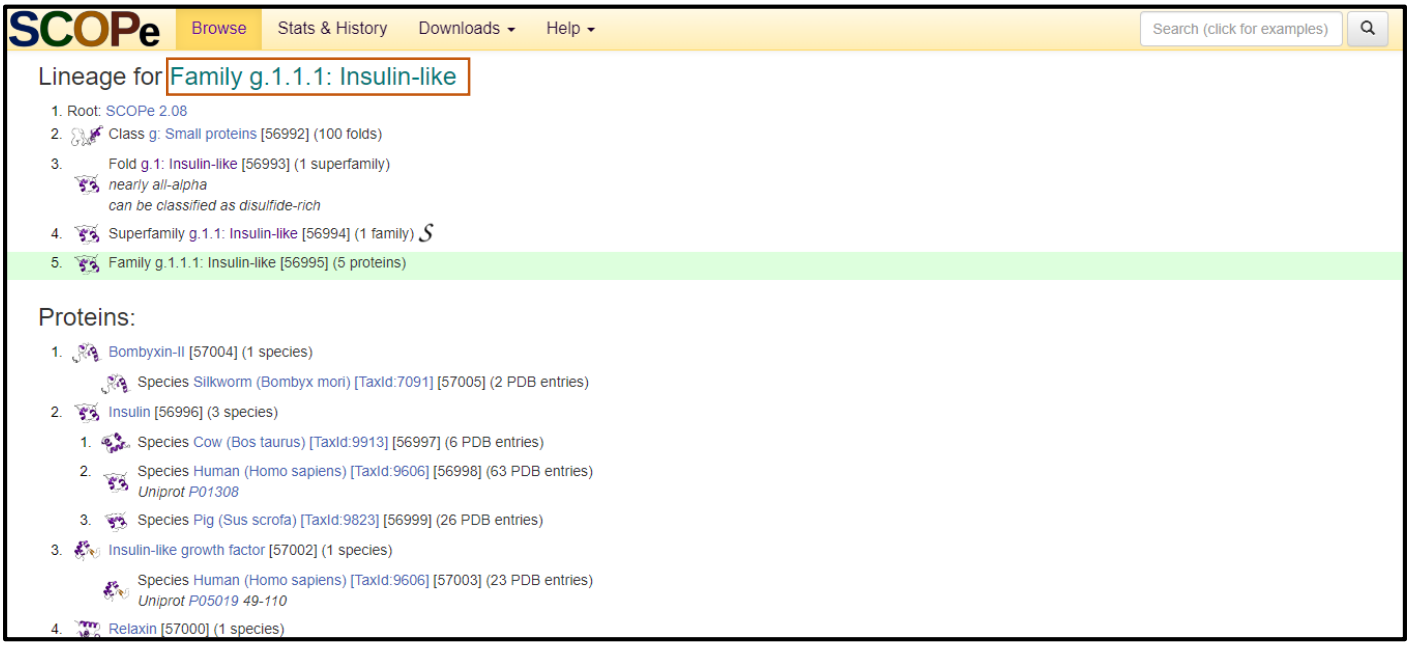


Fig6. Result page for Protein Family

Fig7. Result page for Protein

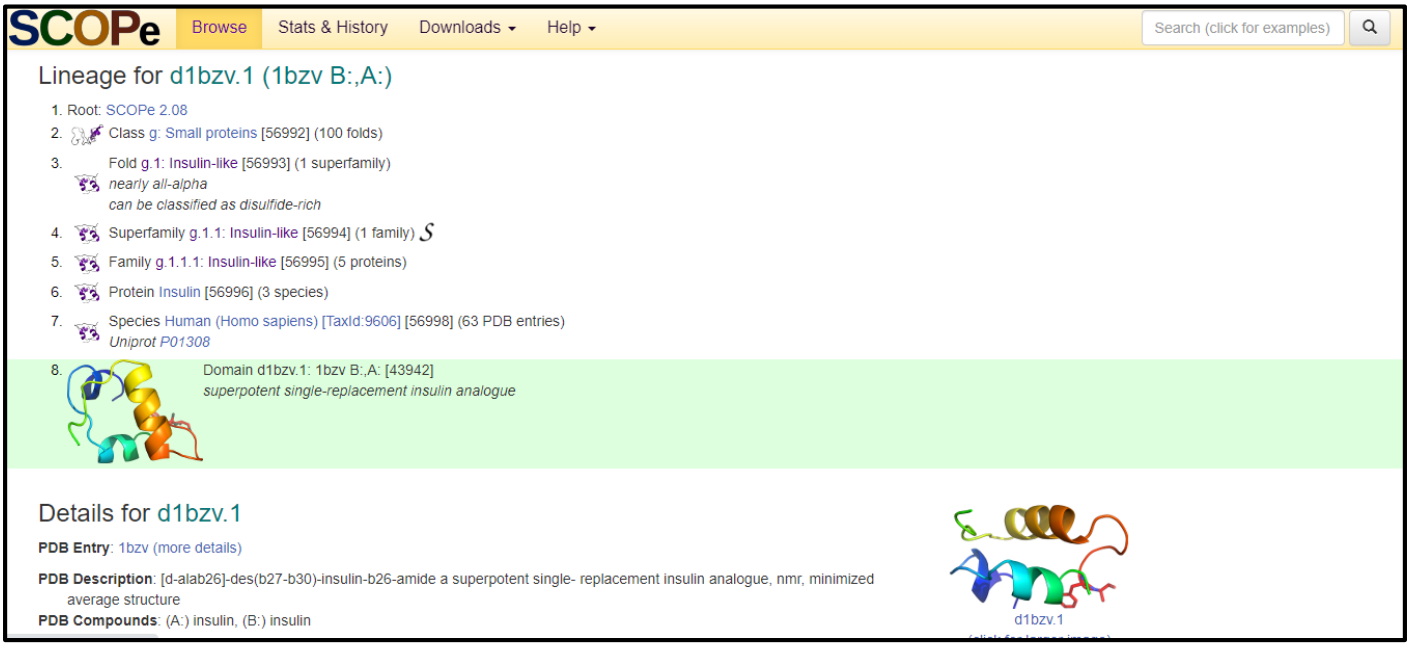


Fig8. Result page for Protein Domain

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. Retrieved February 16, 2022, from https://www.cathdb.info/
4. *CATH Search: Browse*. (n.d.). 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. 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. Retrieved February 16, 2022, from http://www.cathdb.info/version/latest/domain/3bc5A00c
7. *PDB 2iug*. (n.d.). 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. 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. 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. 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. 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. 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. Retrieved February 21, 2022, from https://scop.berkeley.edu/sunid=43942

DATE: 23/02/22

WEBLEM 3

Introduction to tertiary structure prediction

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

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

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

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

INPUT AND OUTPUT:

Registration:

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

Structure prediction server:

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

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

(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

DATE: 08/03/22

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_4^{3-}) 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:

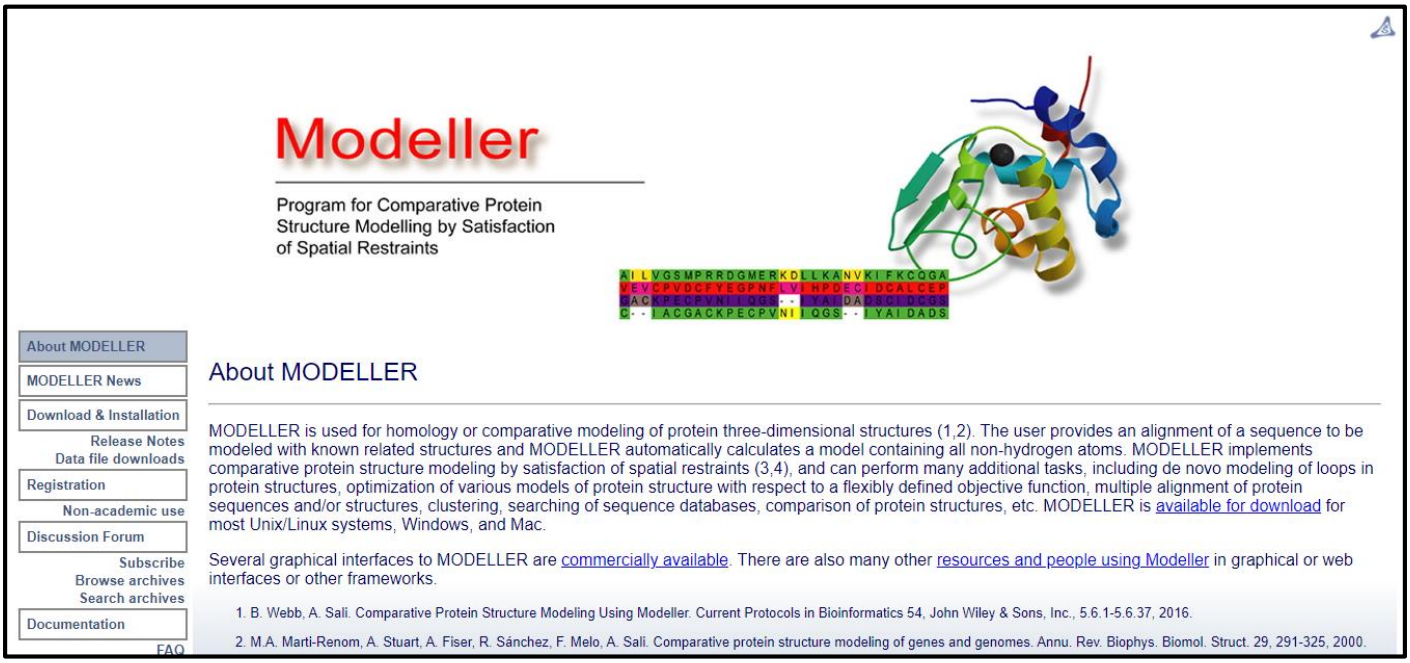


Fig1. Homepage for Modeller

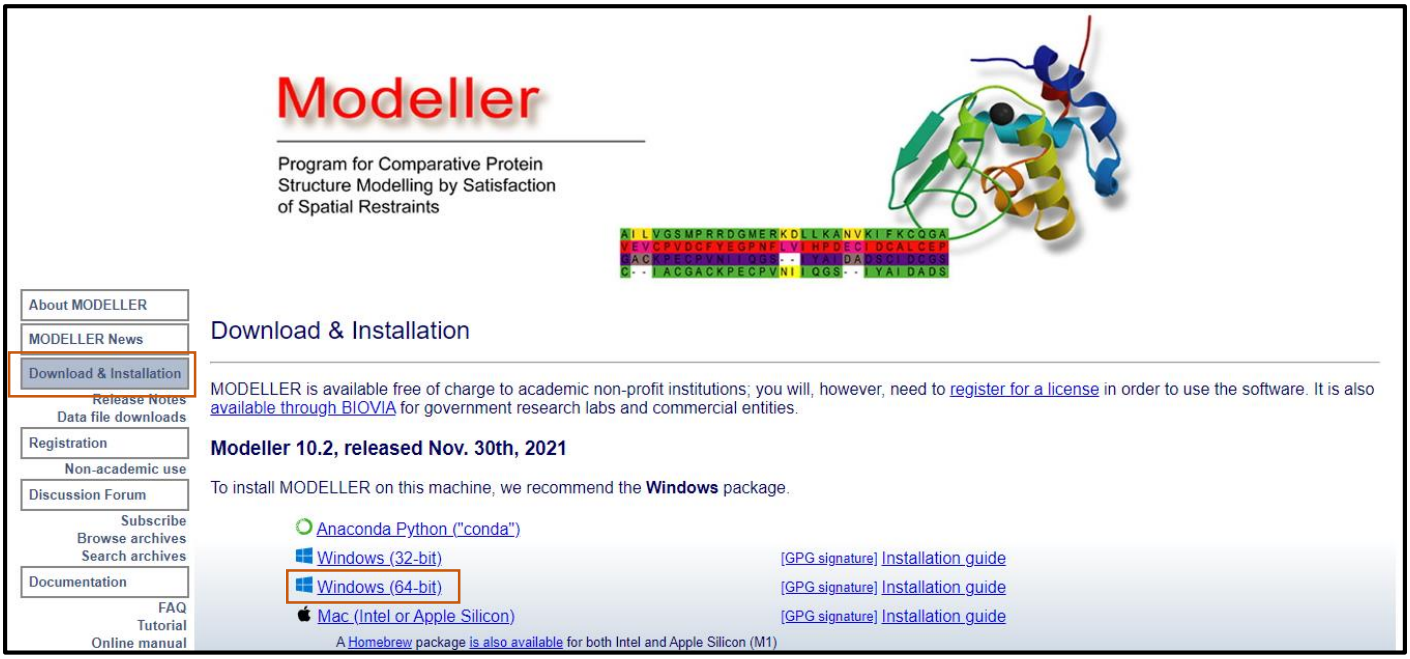


Fig2. Page to install Modeller

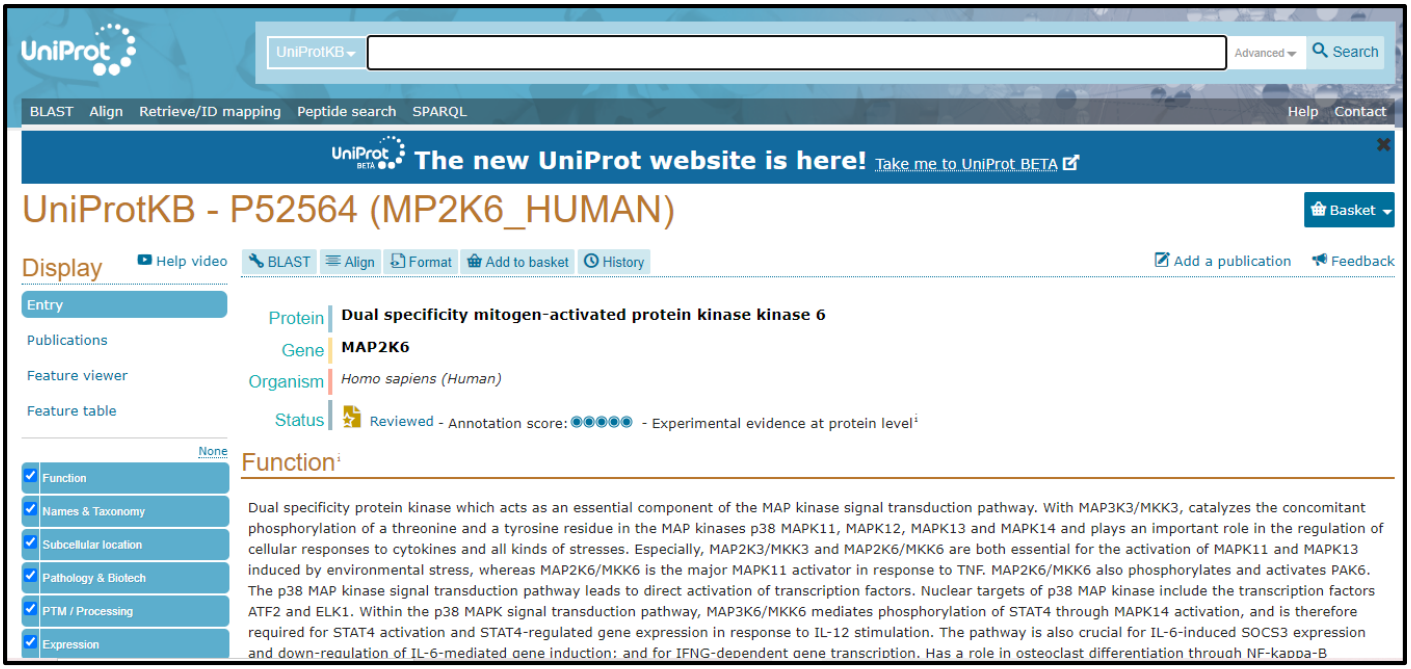


Fig3. Result page for kinase in UniProt database

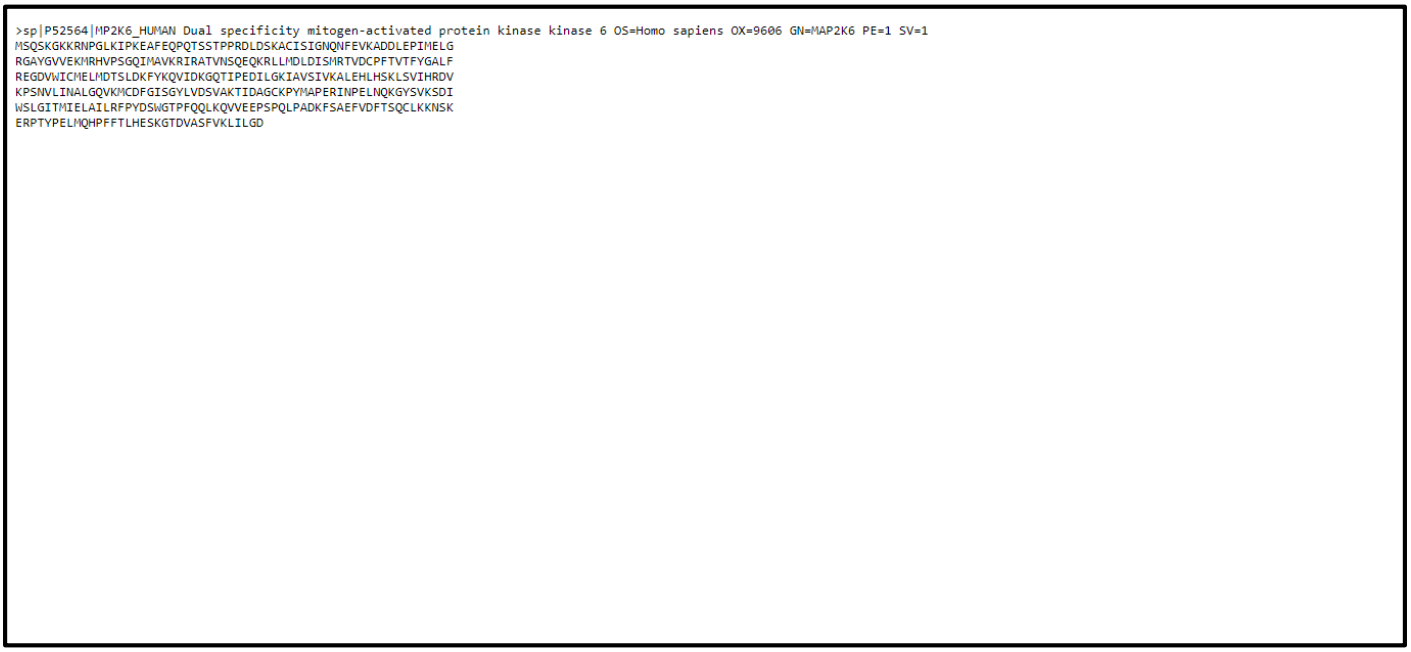


Fig4. FASTA sequence for kinase

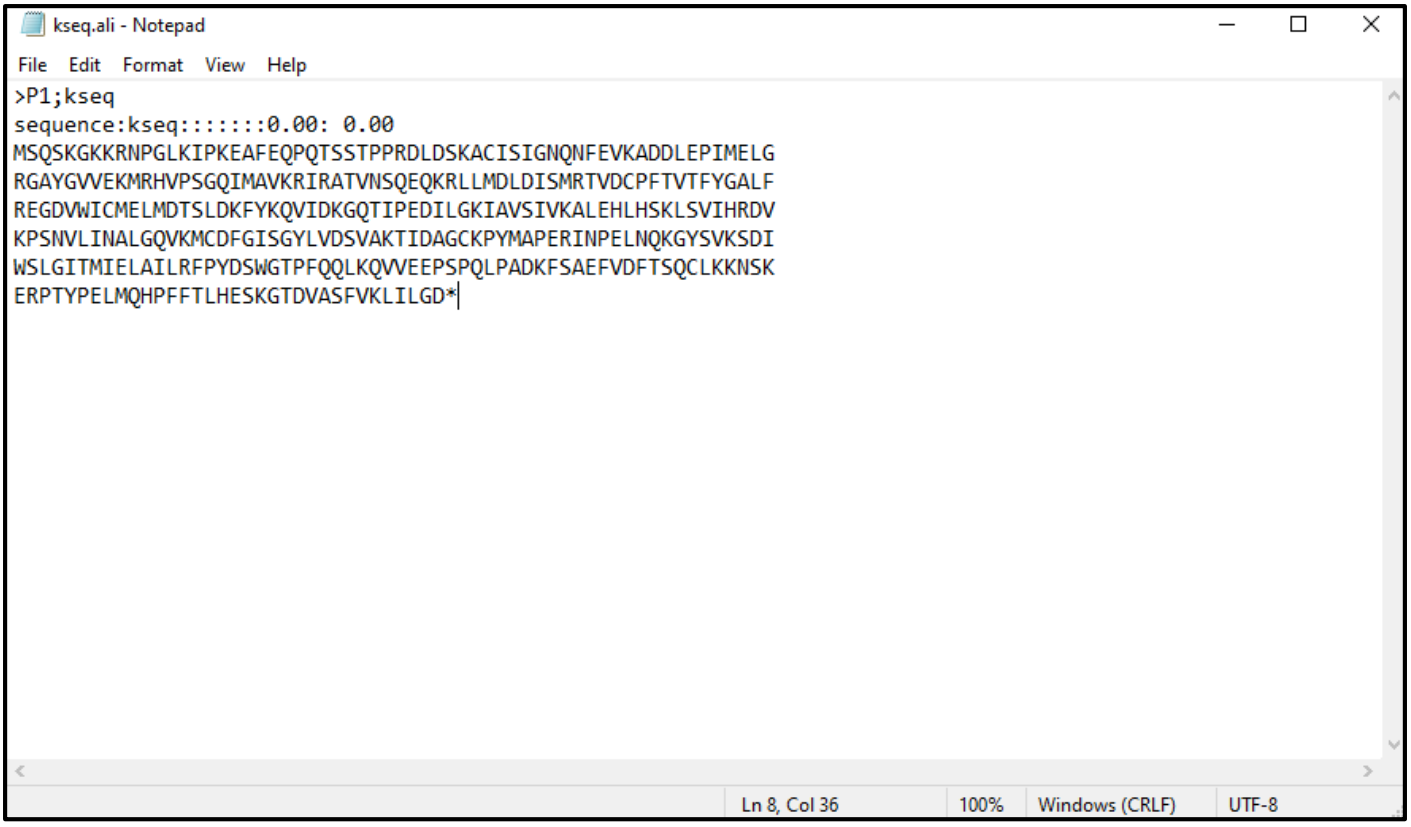


Fig5. Target sequence in PIR format

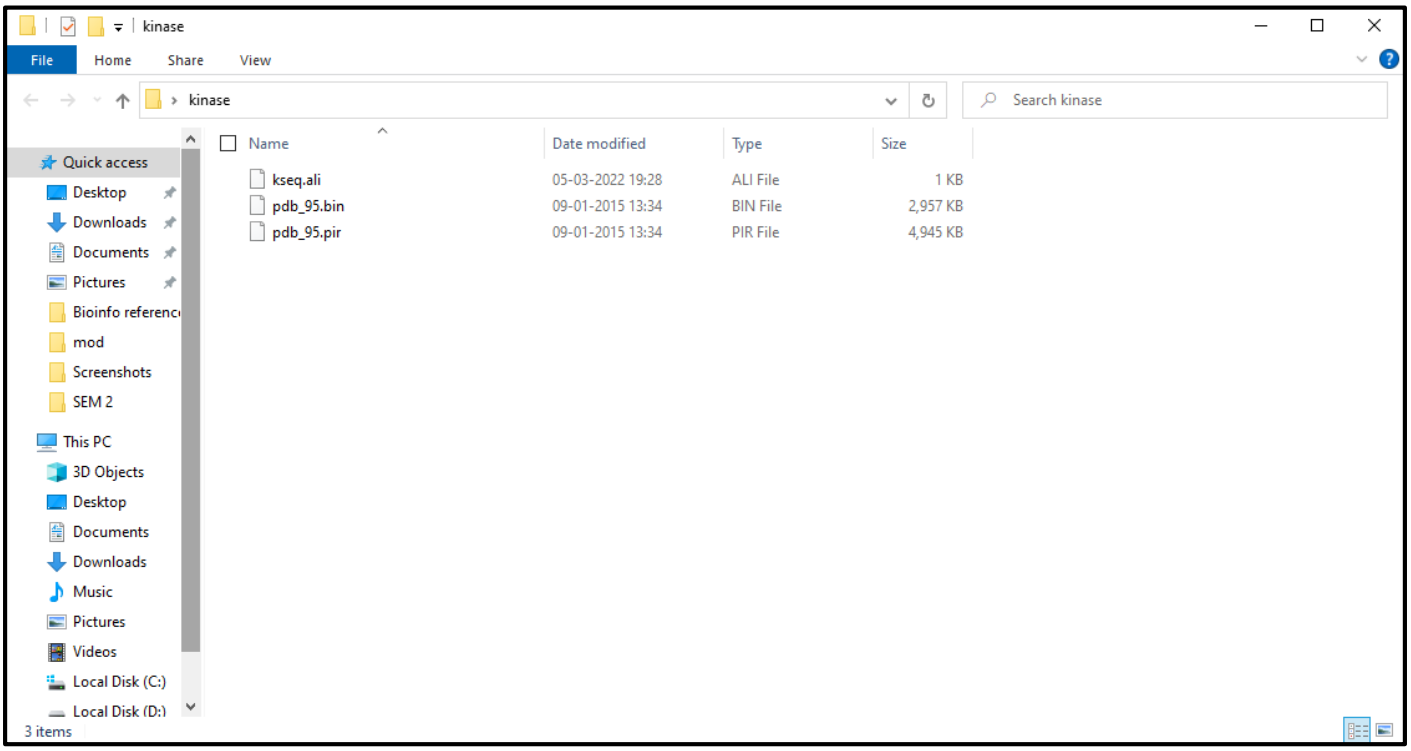


Fig6. Target sequence saved in .ali format

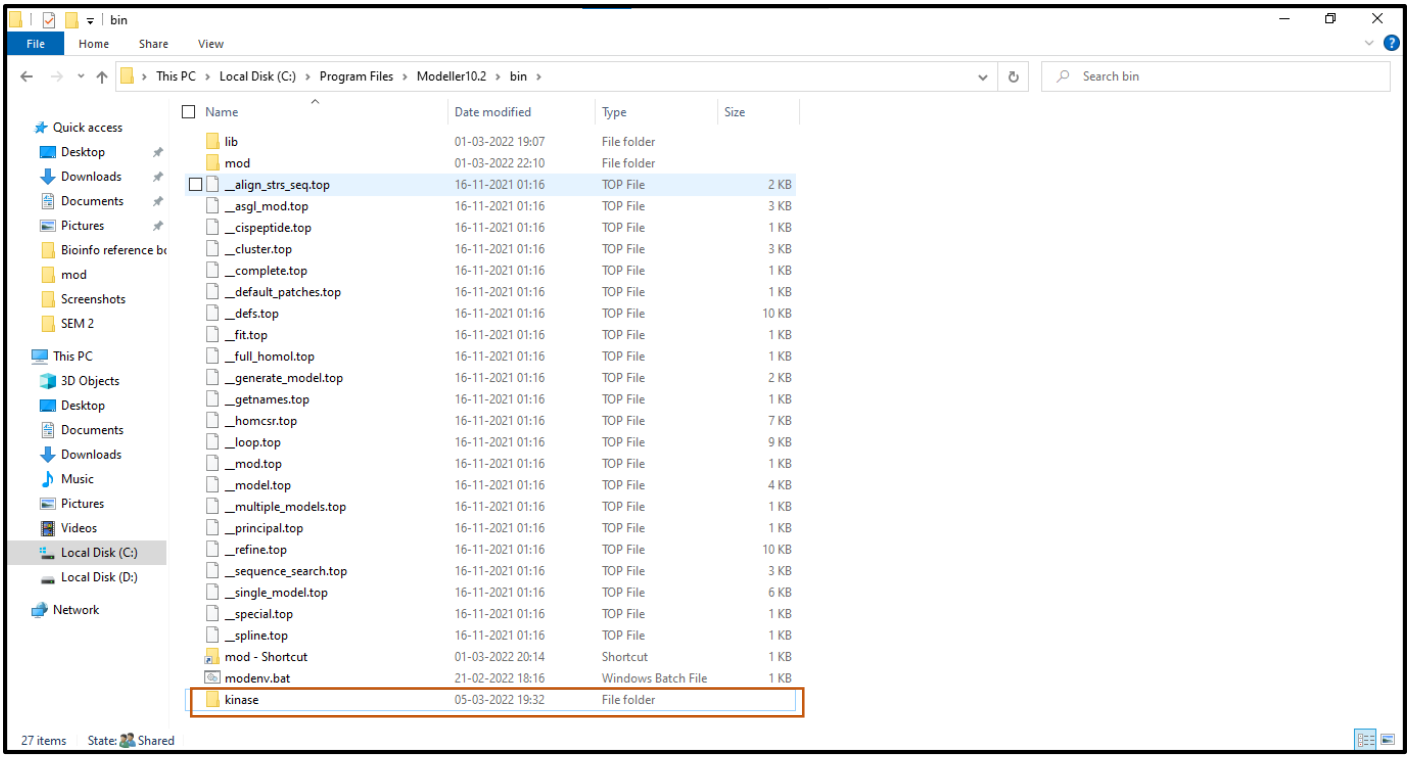


Fig7. Kinase folder saved in bin folder of Modeller

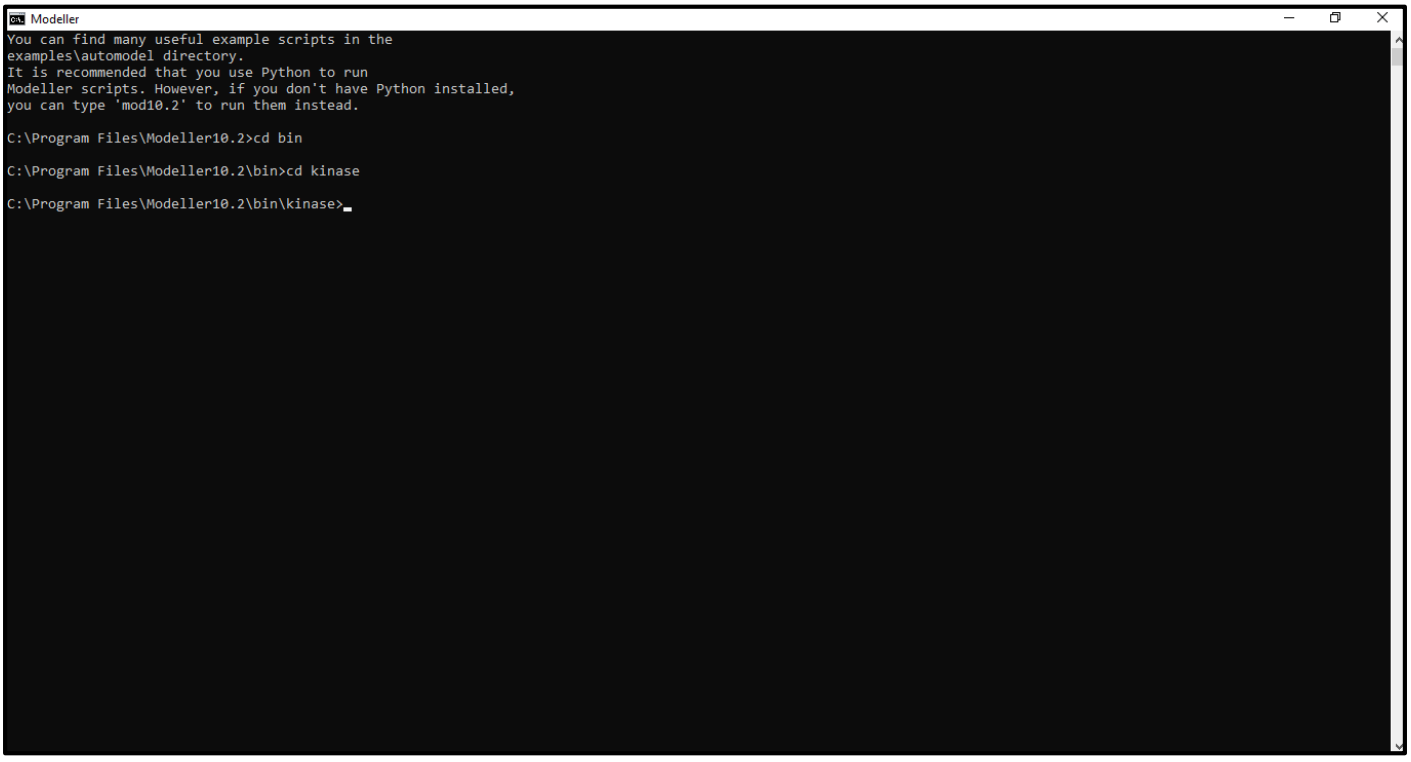


Fig8. Setting working directory on Modeller command line

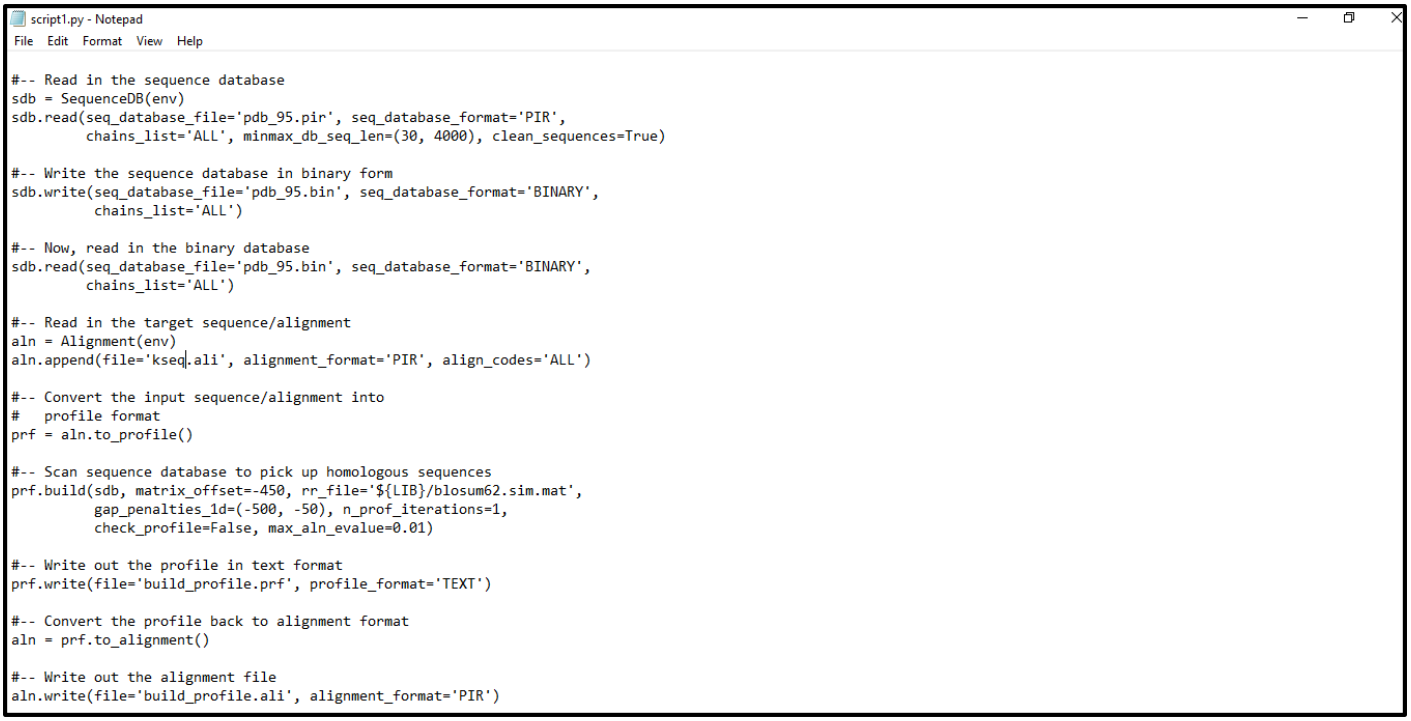


Fig9. Python script for searching for structures related to kinase

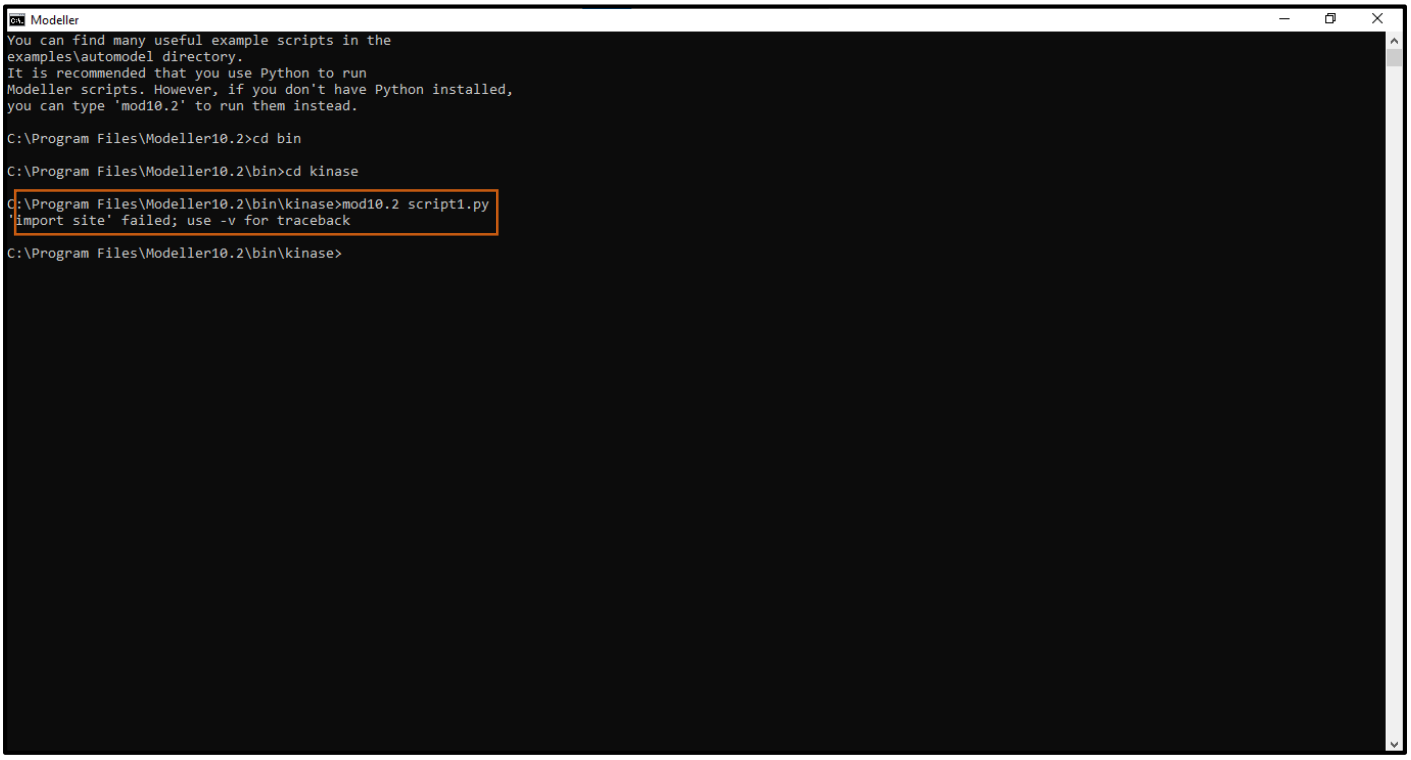
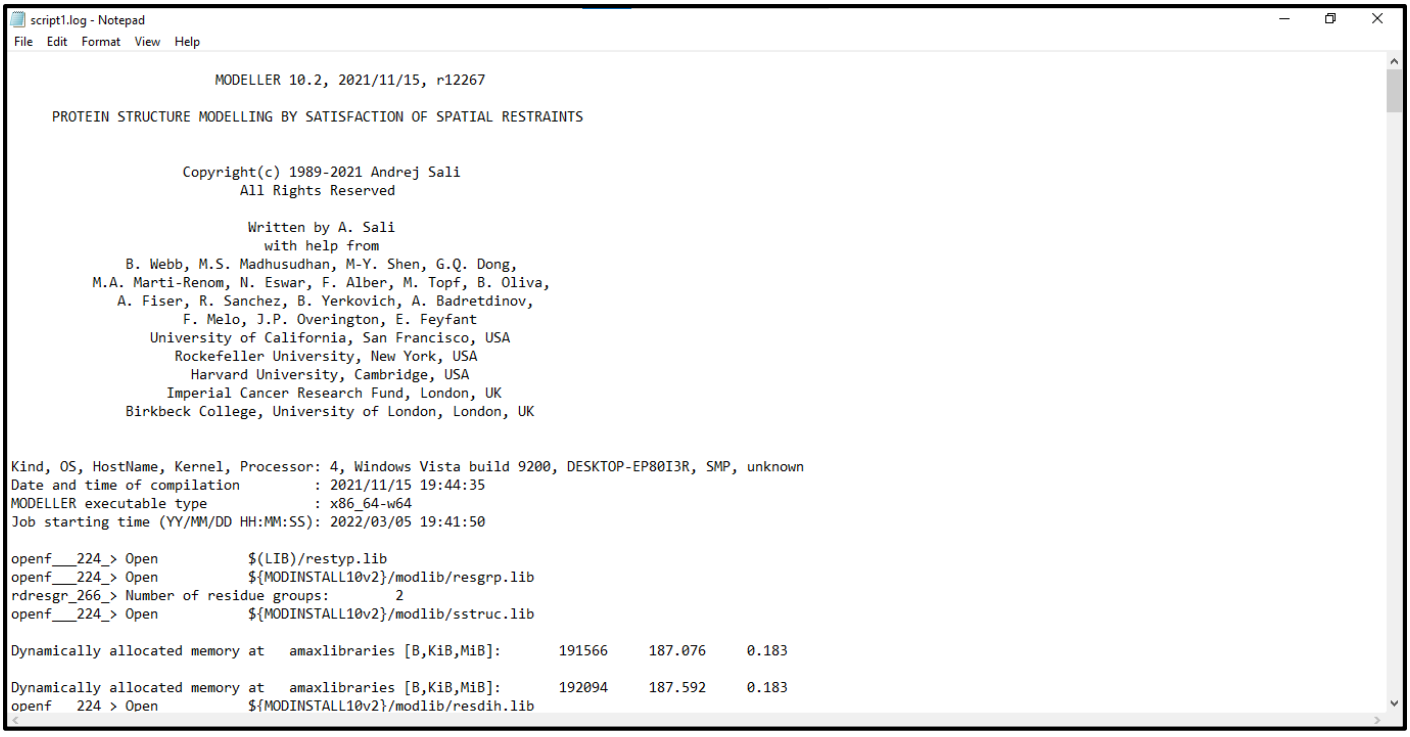
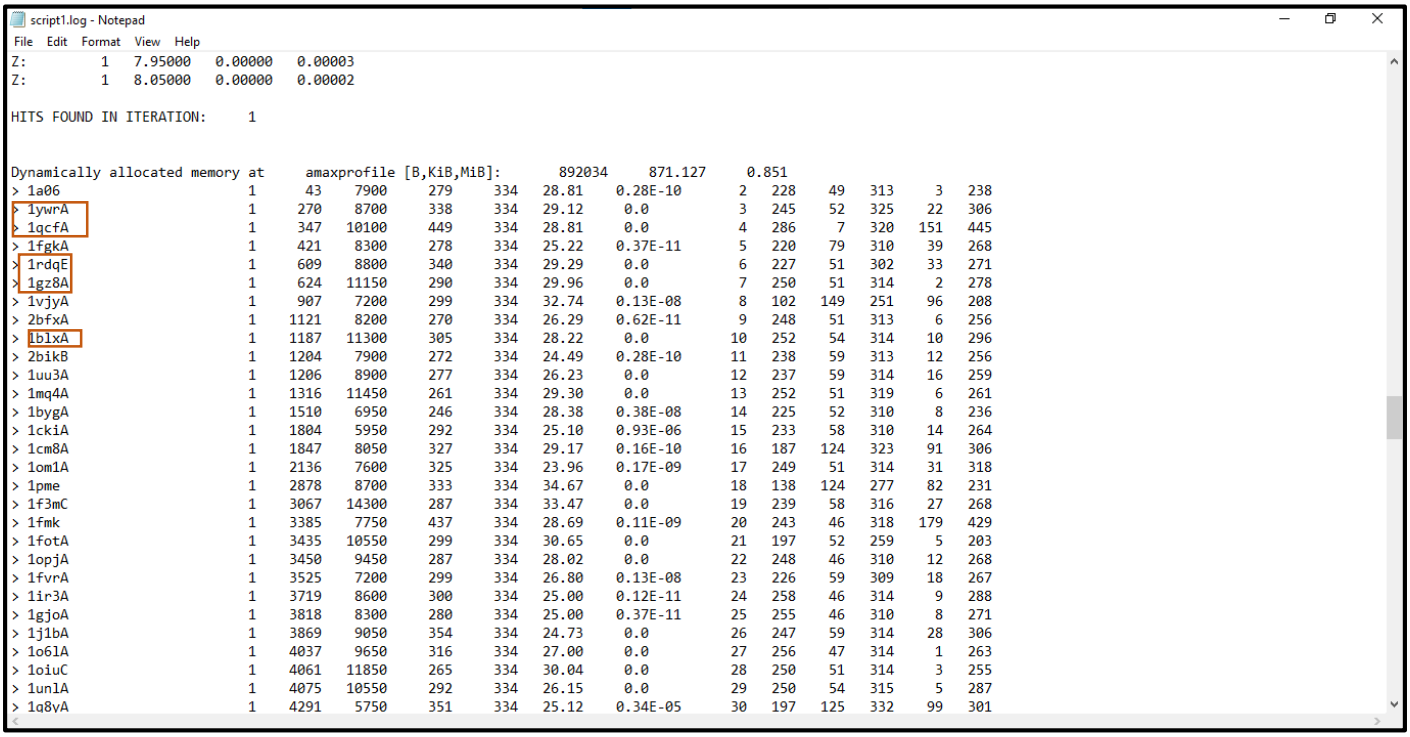


Fig10. Running script1.py



```
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 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 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
> 1vjyA                      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
> 1b1xA                      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
> 1mq4A                      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
> 1ckiA                      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
> 1q8yA                      1   4291    5750     351     334   25.12    0.34E-05      30   197   125   332    99   301
```

Fig11.1. Hits found for similar structures

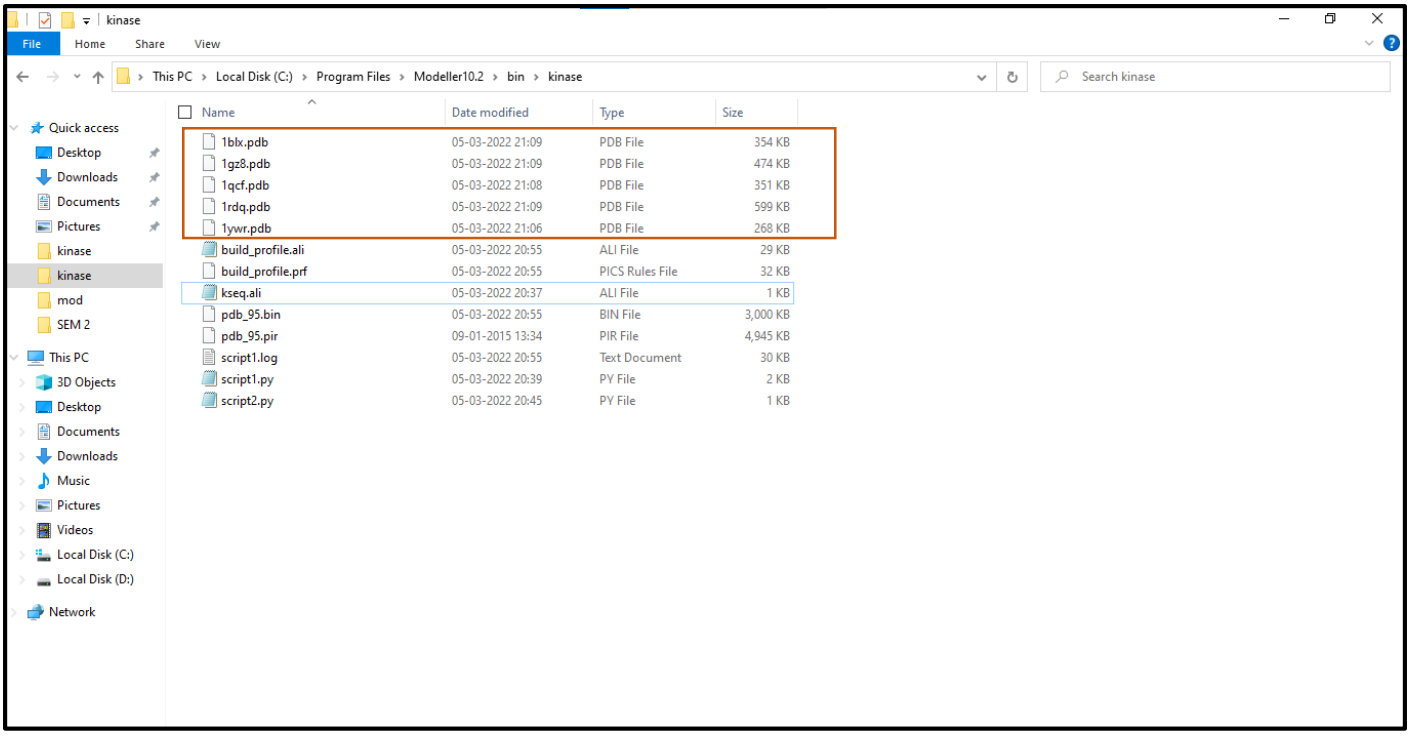


Fig12. Five structures download in PDB format

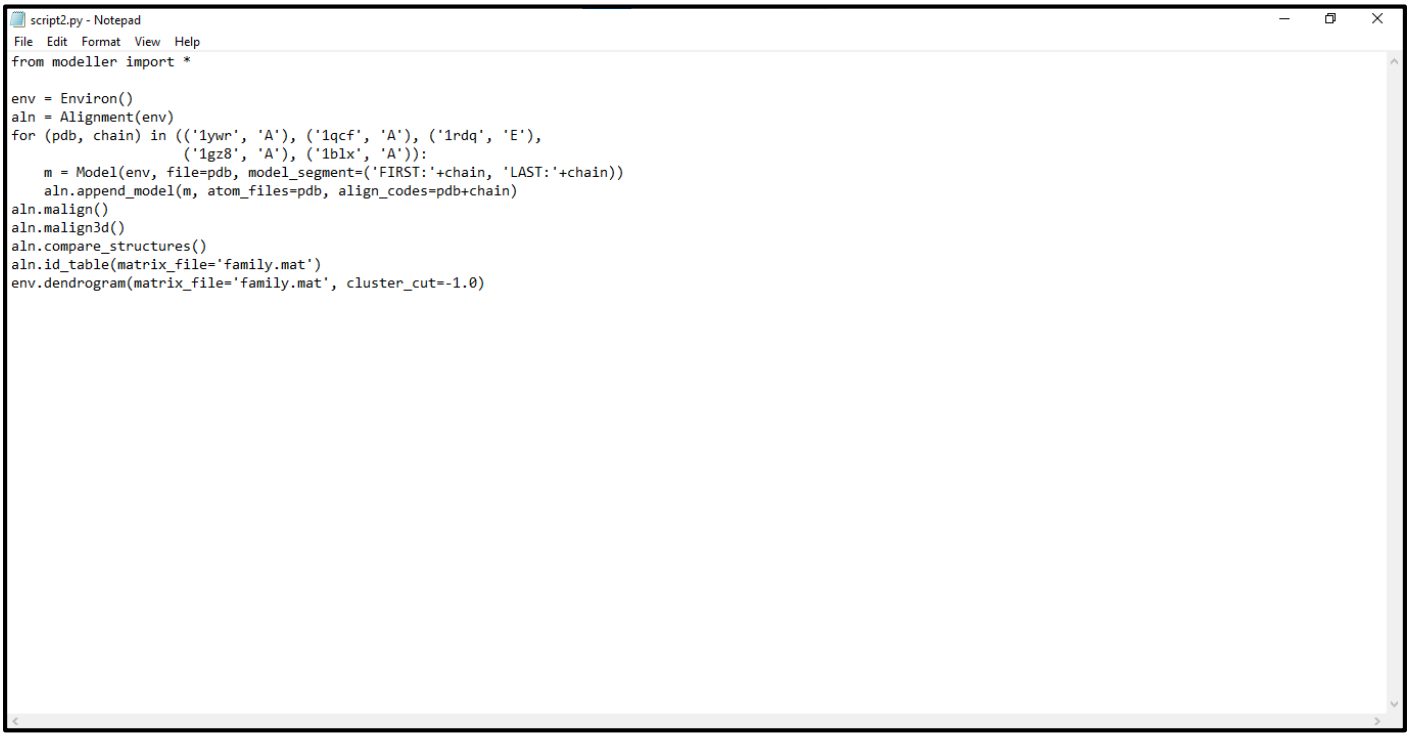


Fig13. Python script for selecting a template

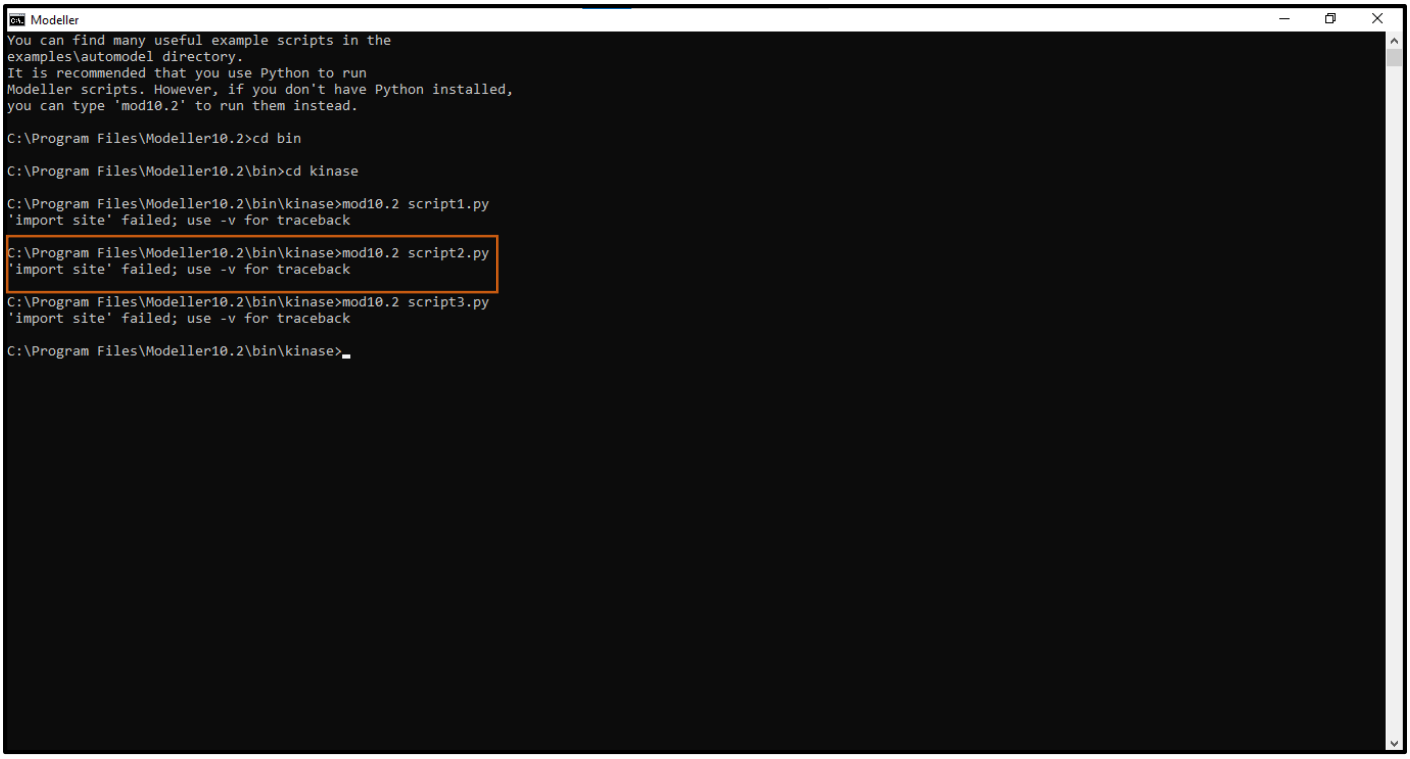


Fig14. Running script2

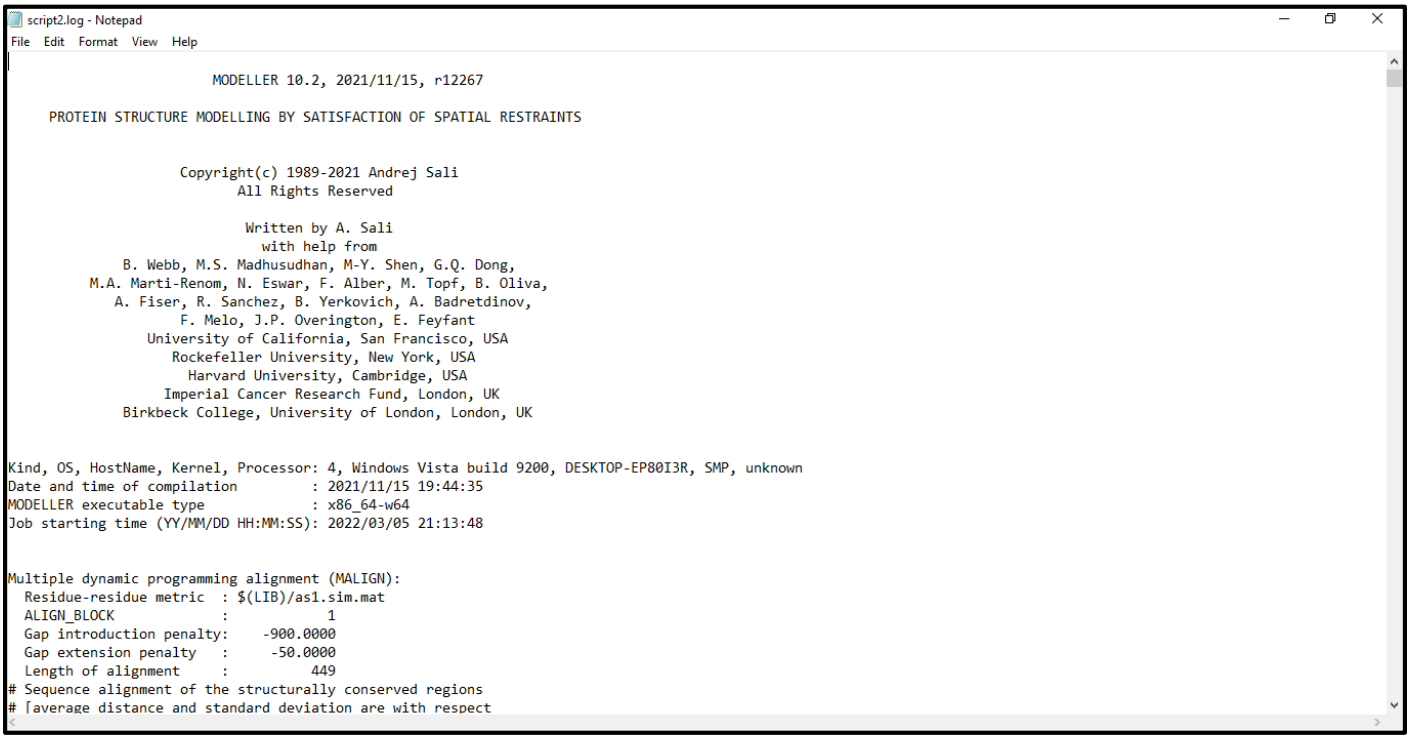


Fig15. Log file for script2

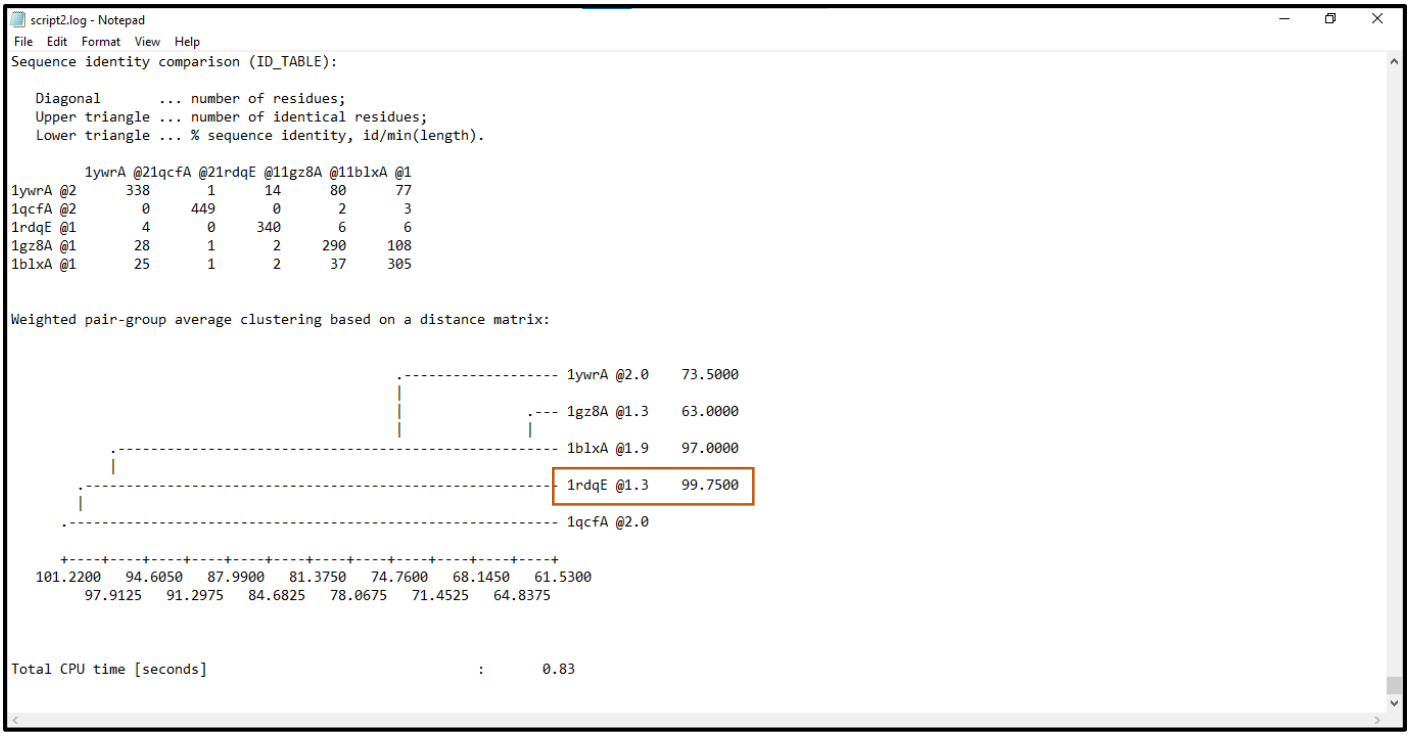


Fig15.1. Structure selected with low x-ray crystallography value and high NMR value

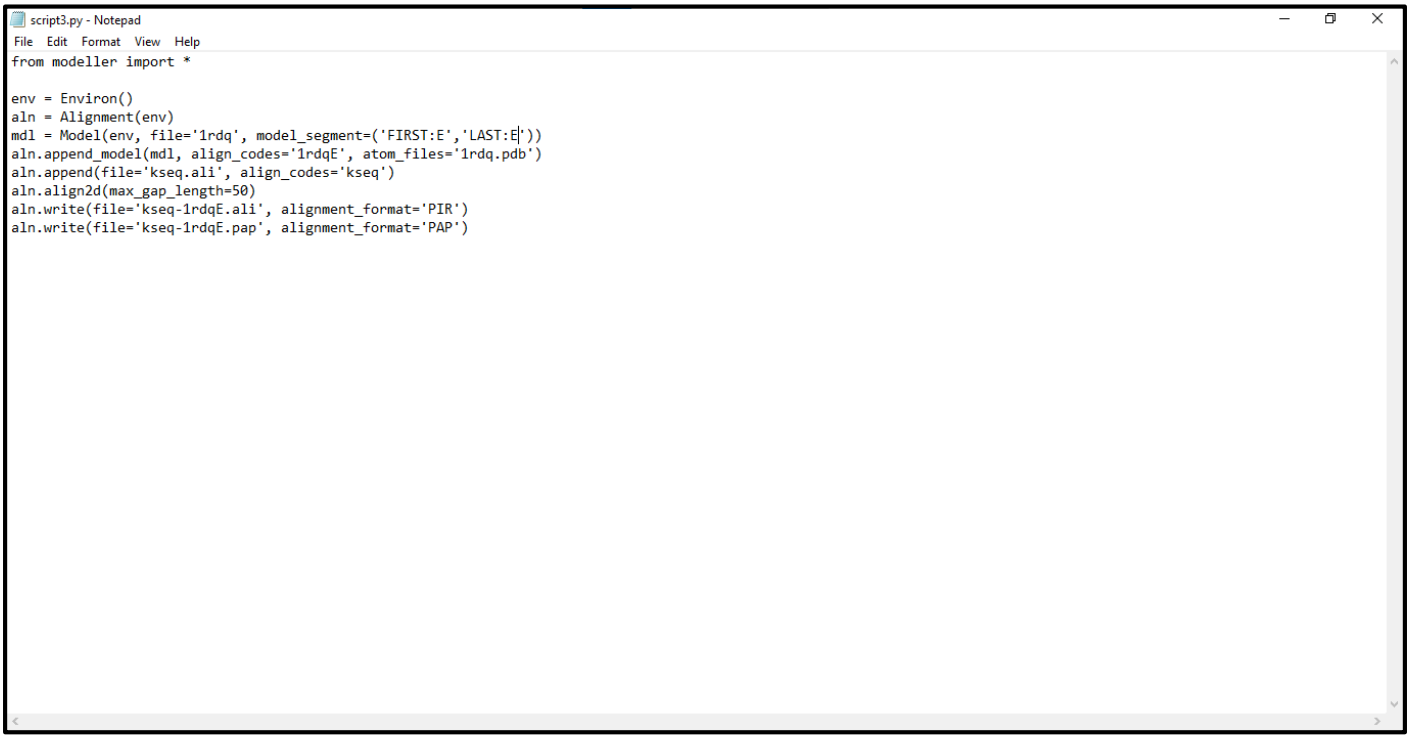


Fig16. Python script for aligning query with the template

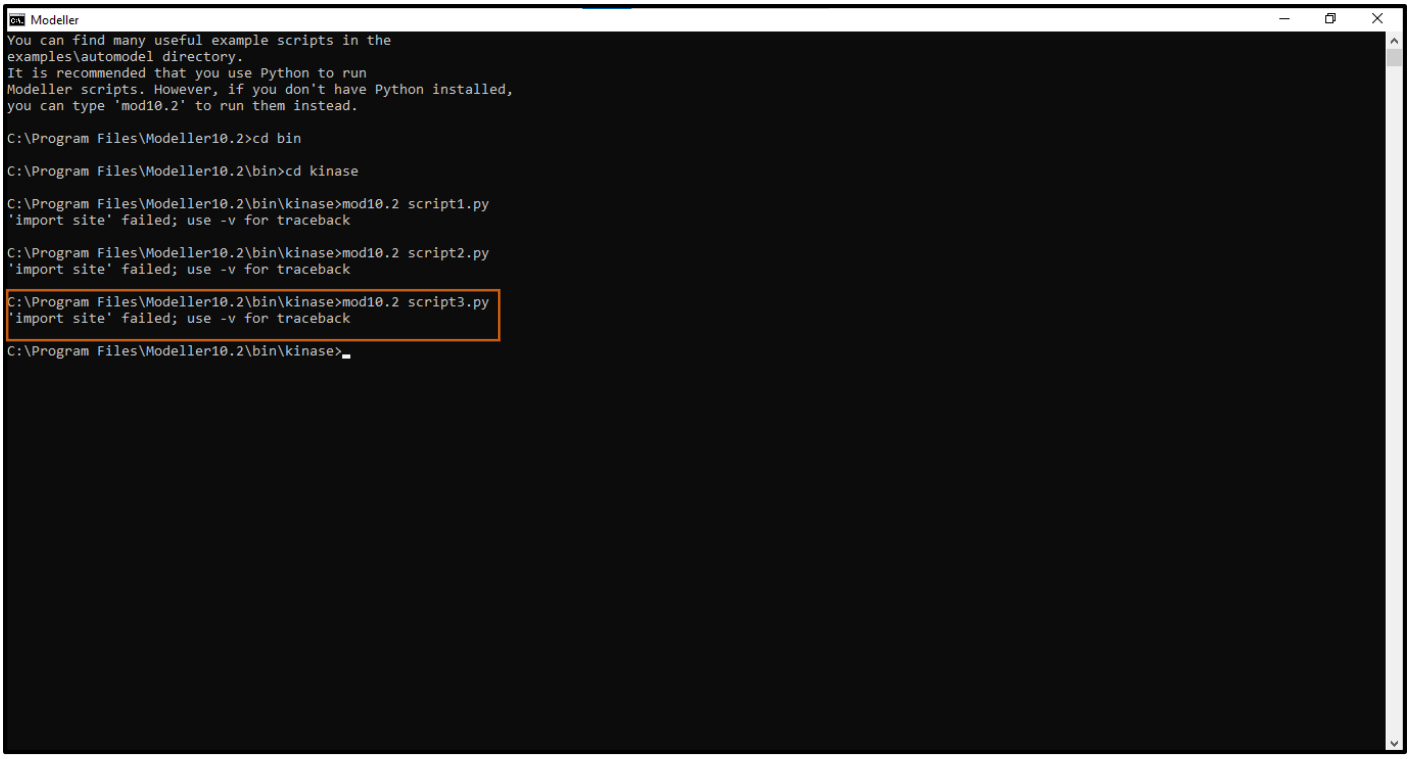


Fig17. Running script3

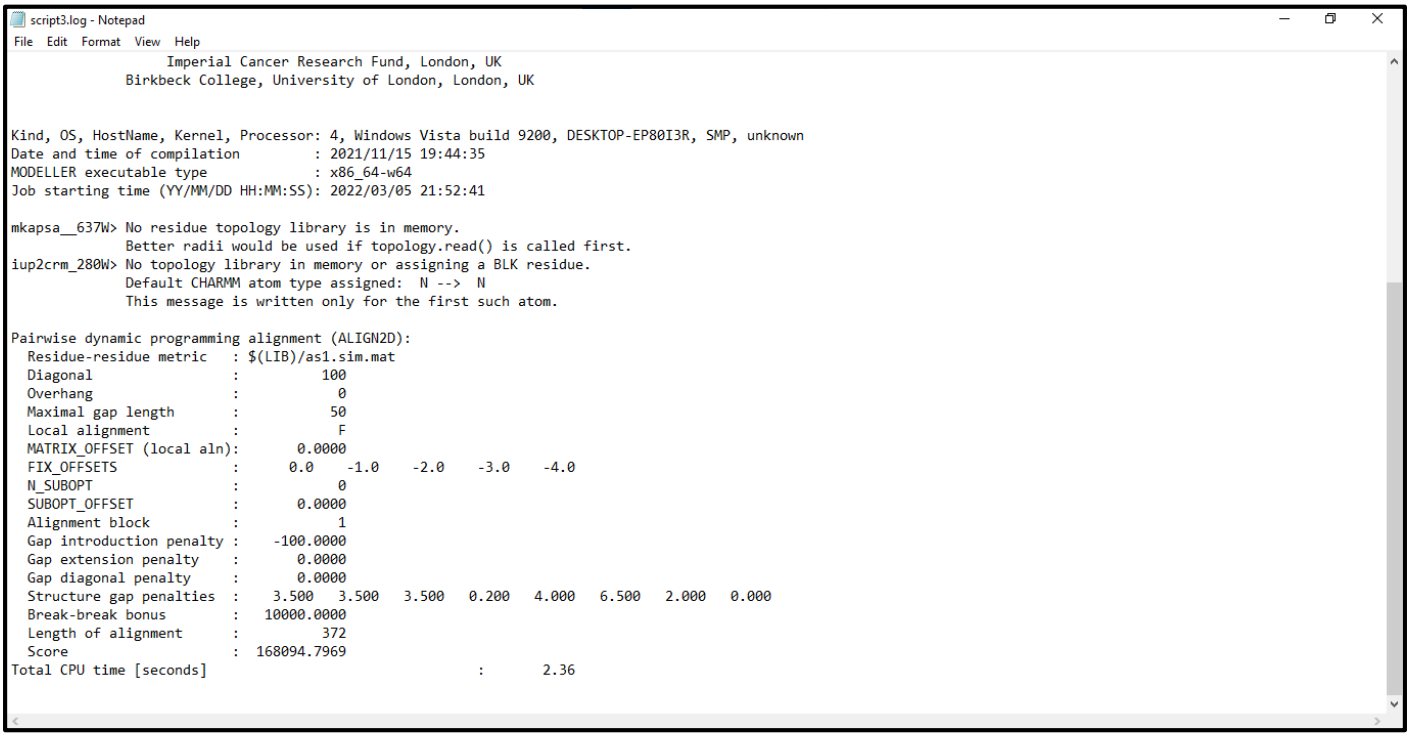


Fig17. Log file for script3

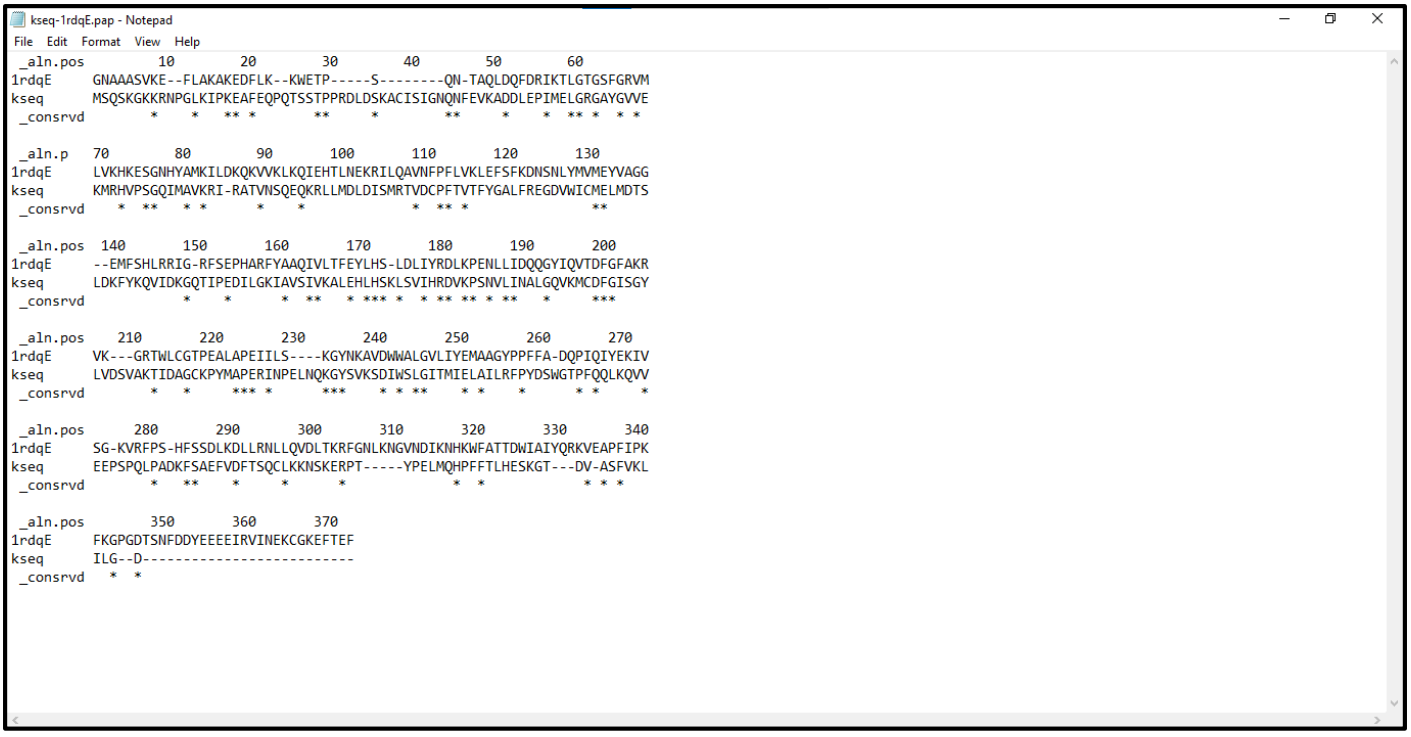


Fig18. Sequence alignment

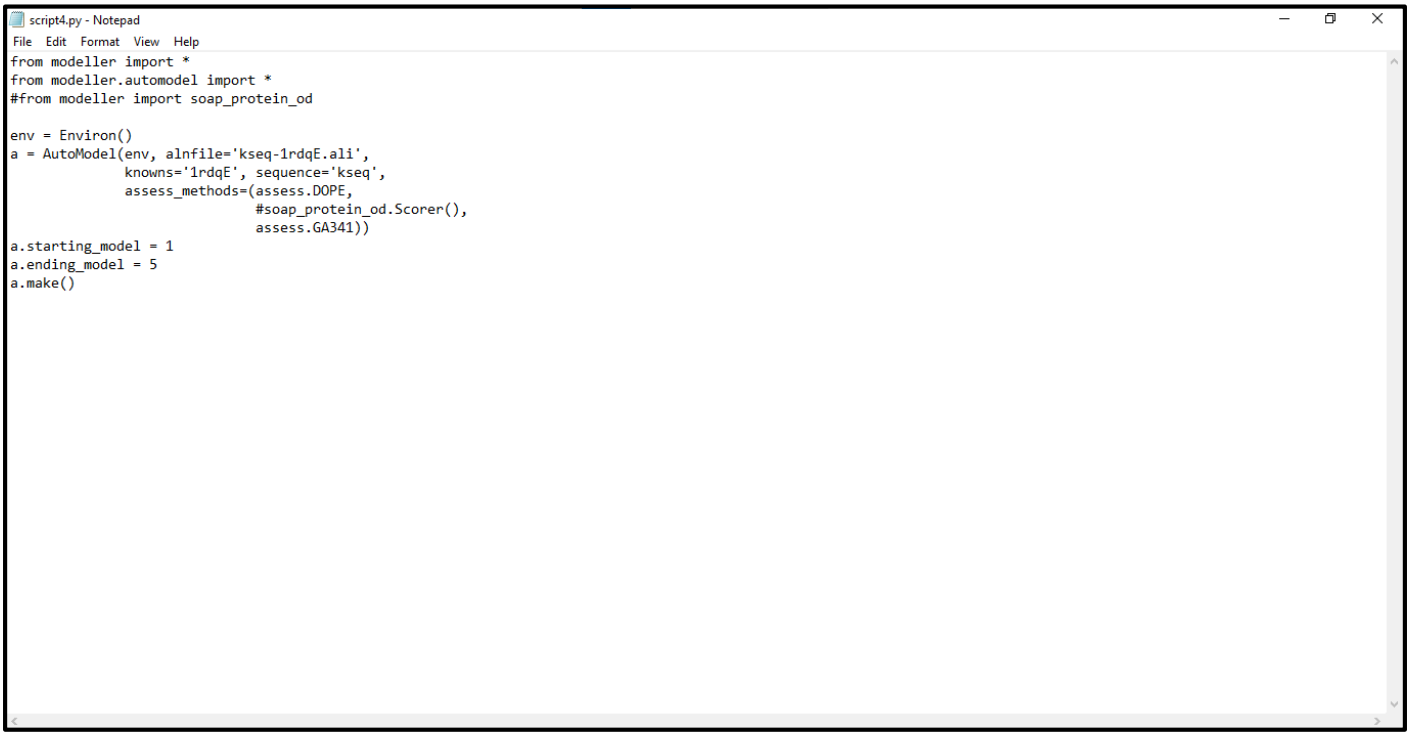


Fig19. Python script for model building

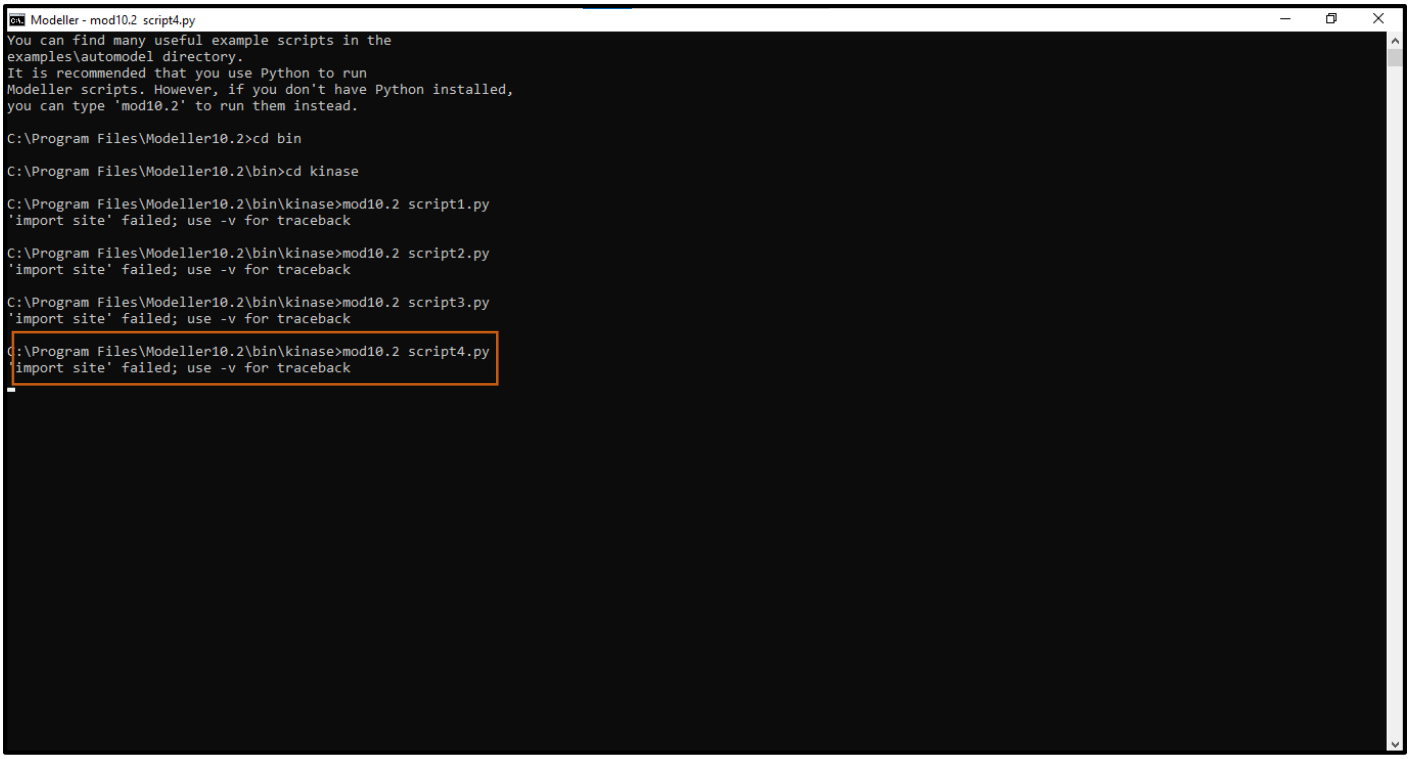


Fig20. Running script4

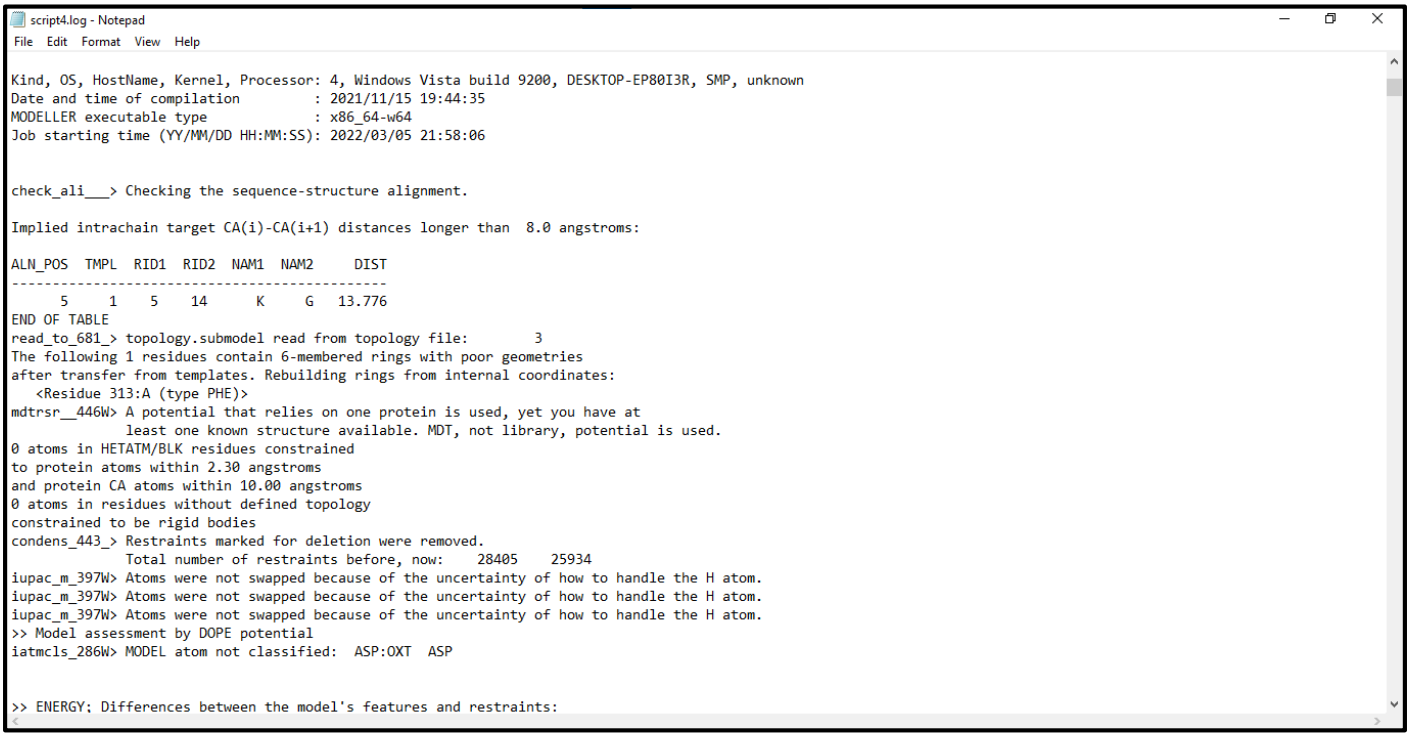


Fig21. Log file for script4

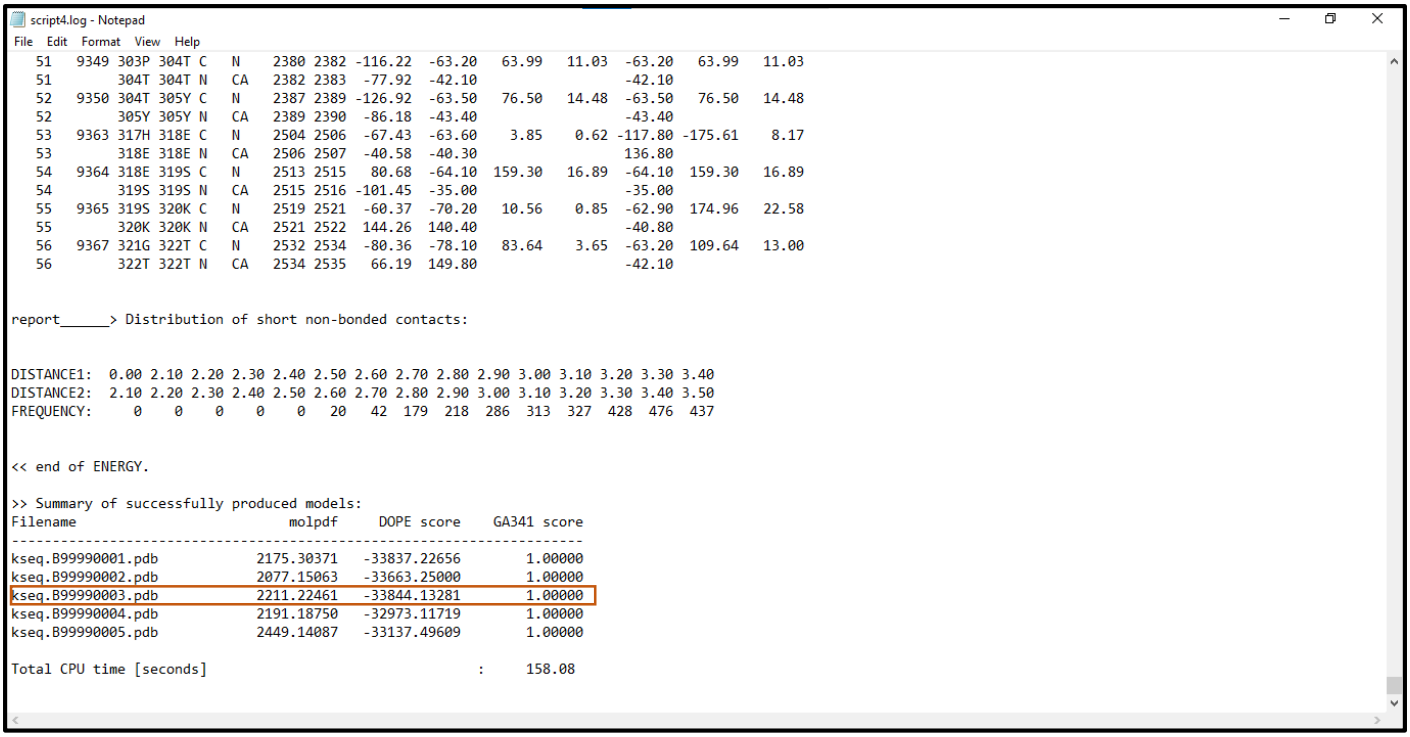


Fig22. Structure with low DOPE score selected as final model

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 kinase 6 - Homo sapiens (Human) - MAP2K6 gene & protein*. (n.d.). Www.uniprot.org. 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

DATE: 08/03/22

WEBLEM 3b

I-TASSAR

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

AIM:

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

INTRODUCTION:

Kinase, an enzyme that adds phosphate groups (PO43−) 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:

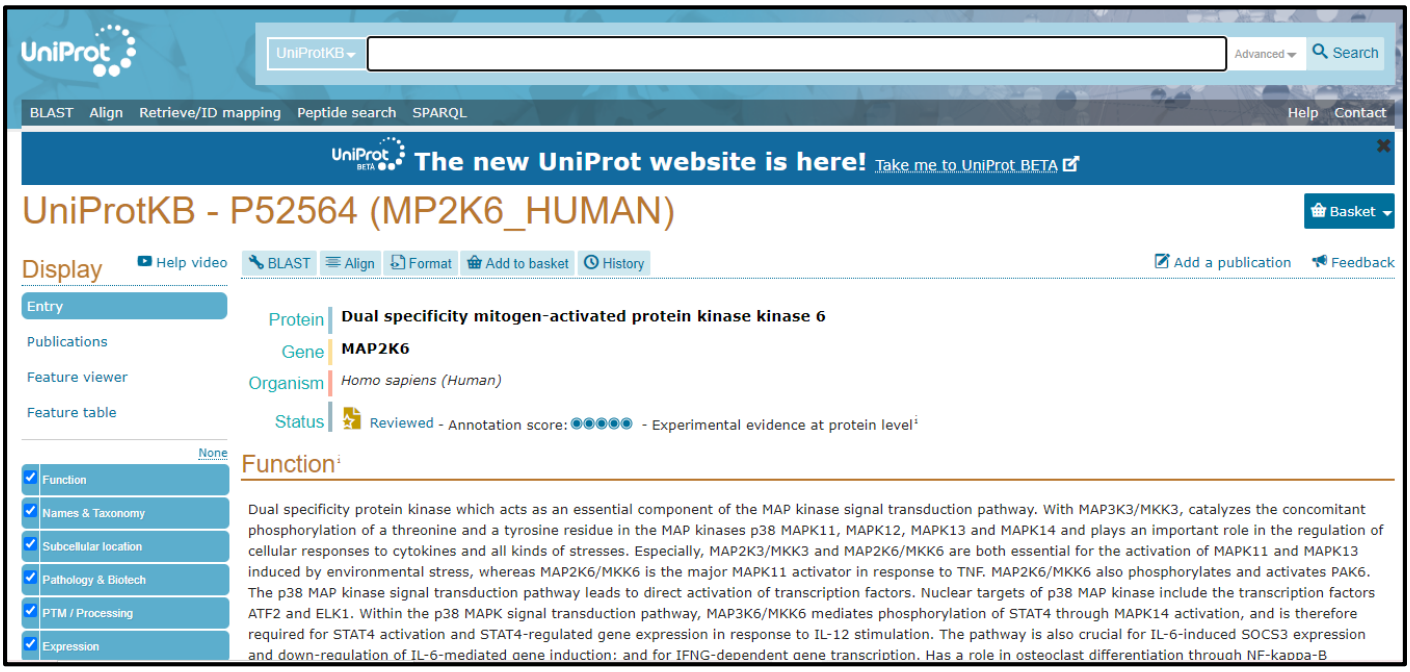


Fig1. Result page for kinase in UniProt database

```
>sp|P52564|MP2K6_HUMAN Dual specificity mitogen-activated protein kinase kinase 6 OS=Homo sapiens OX=9606 GN=MAP2K6 PE=1 SV=1
MSQSKGKKRNPGLKIPKEAFEQPQTSSTPPRDLDSKACISIGNQNFEVKADDLEPIMELG
RGAYGVVEKMRHVPSGQIMAVKRIRATVNSQEQKRLLMDLDISMRTVDCPFTVTFYGALF
REGDVWICMELMDTSLDKFYKQVIDKGQTIPEDILGKIAVSIVKALEHLHSKLSVIHRDV
KPSNVLINALGQVKMCDFGISGYLVDSVAKTIDAGCKPYMAPERINPELNQKGYSVKSDI
WSLGITMIELAILRFPYDSWGTPFQQLKQVVEEPSPQLPADKFSAEFVDFTSQCLKKNSK
ERPTYPELMQHPFFTLHESKGTDVASFVKLILGD
```

Fig2. FASTA sequence for kinase

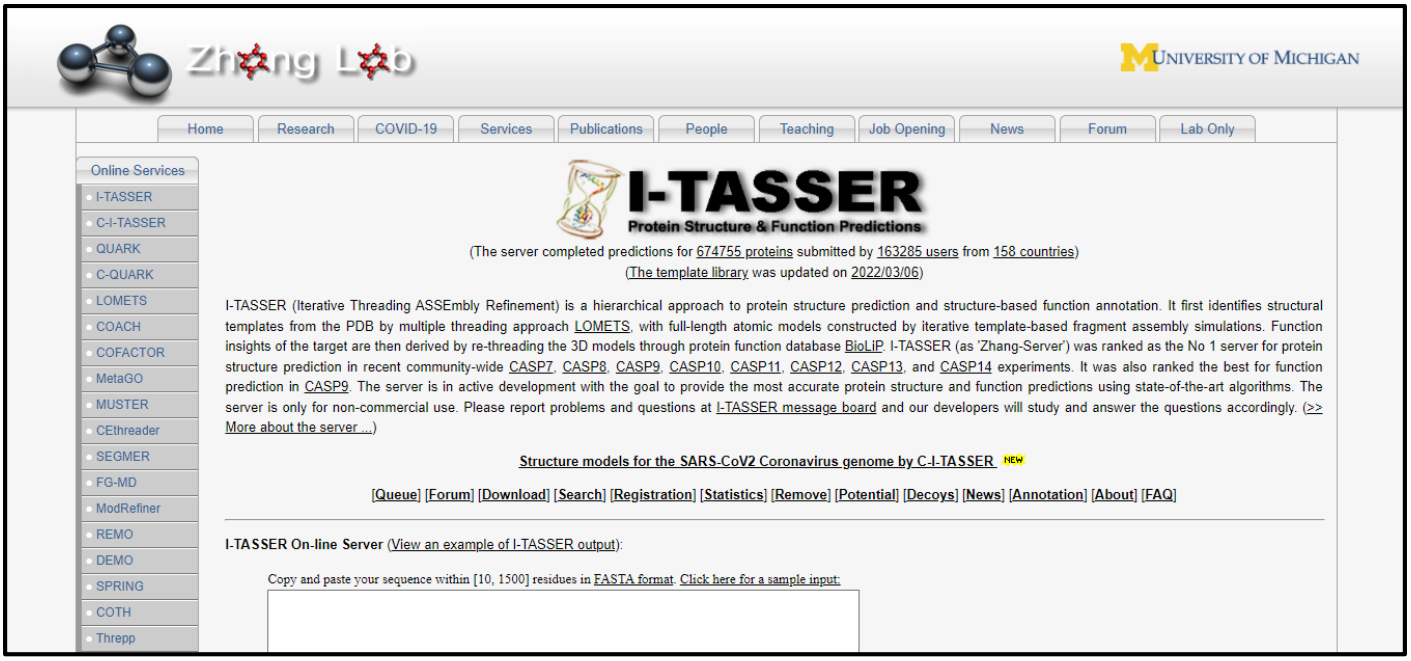


Fig3. Homepage for I-TASSER

Fig4. Submission of query

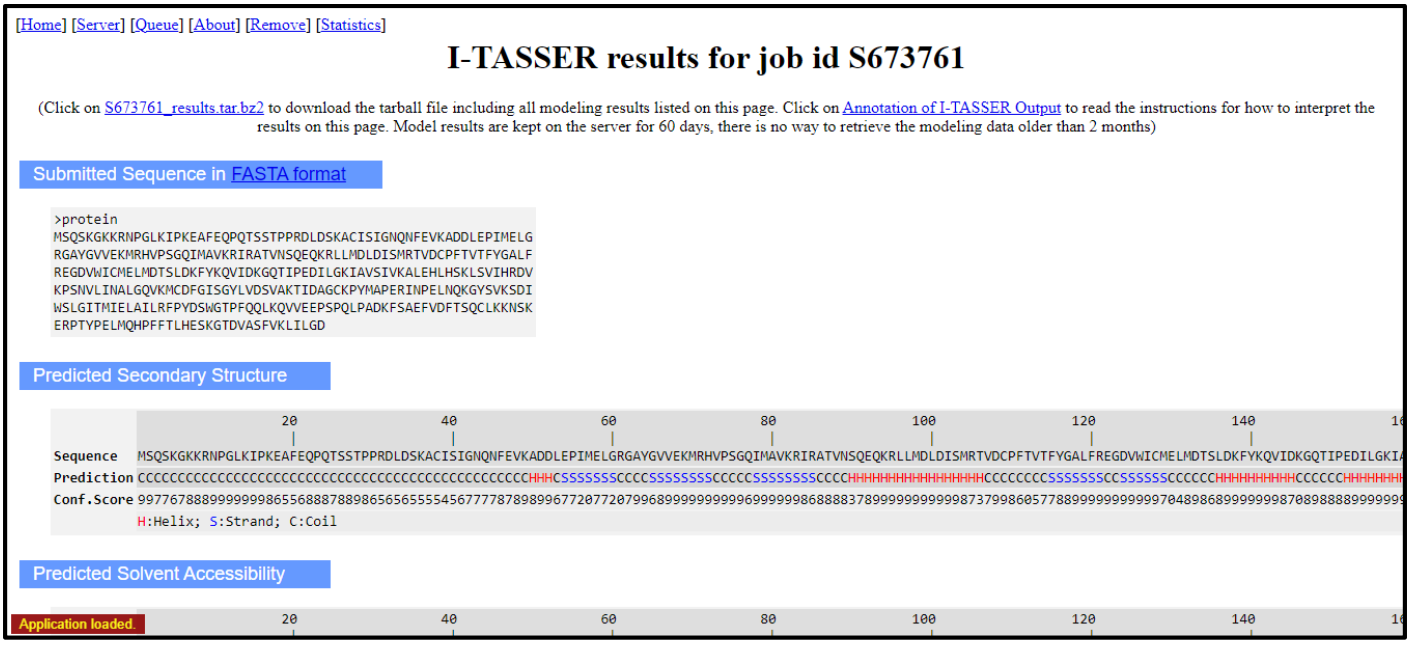


Fig5. Result for predicted secondary structure

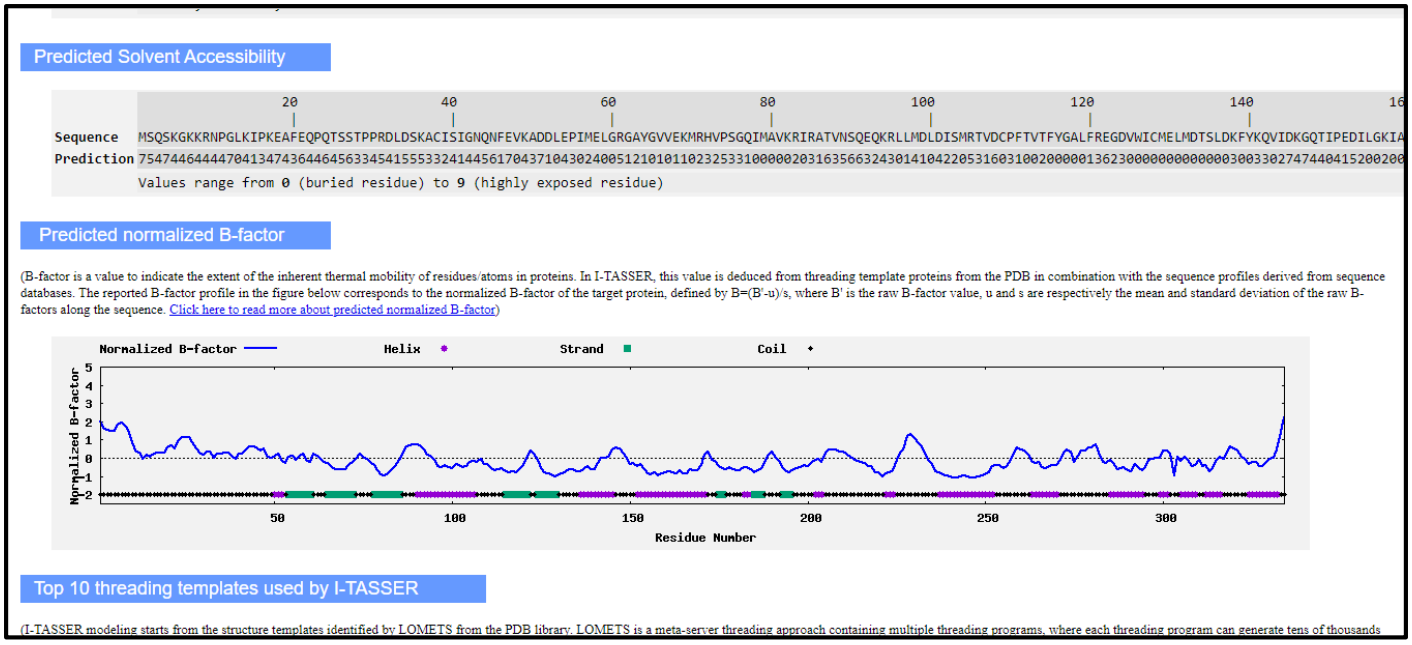


Fig6. Result for predicted solvent accessibility and normalized B-factor

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

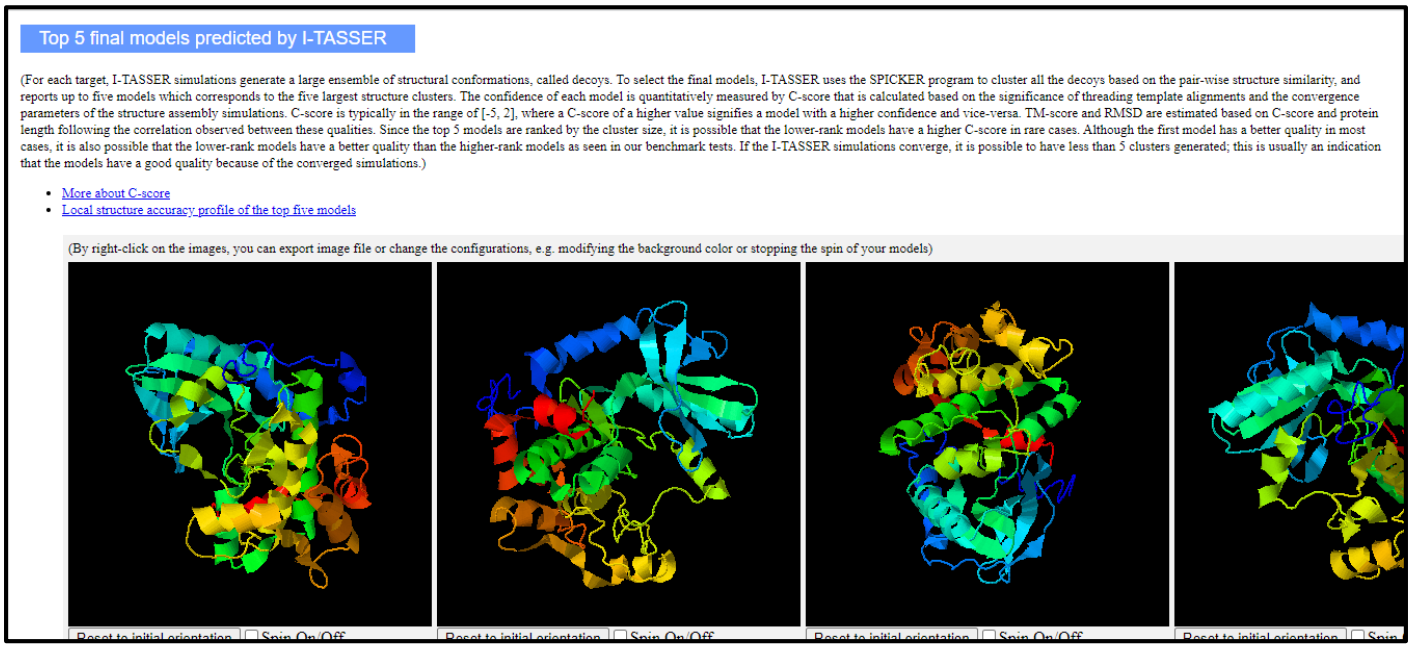


Fig8. Result for top 5 final models predicted

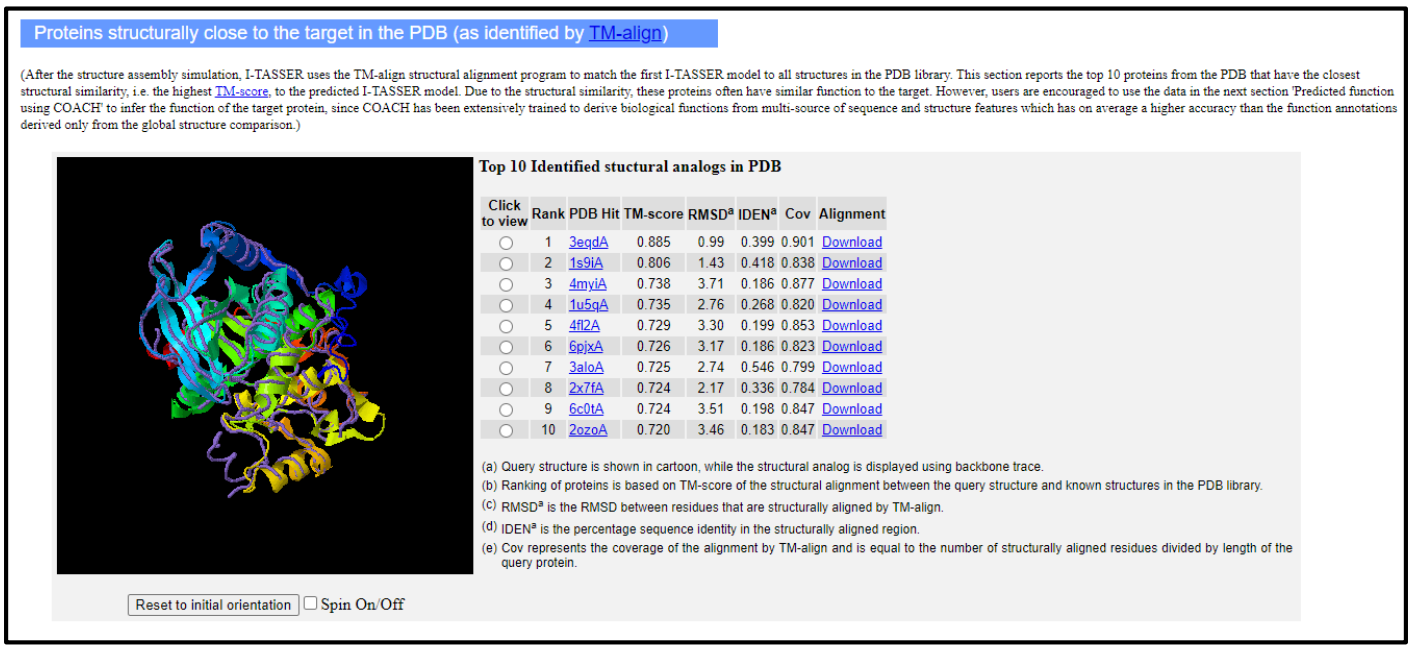


Fig9. Result for proteins that are structurally close to target

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

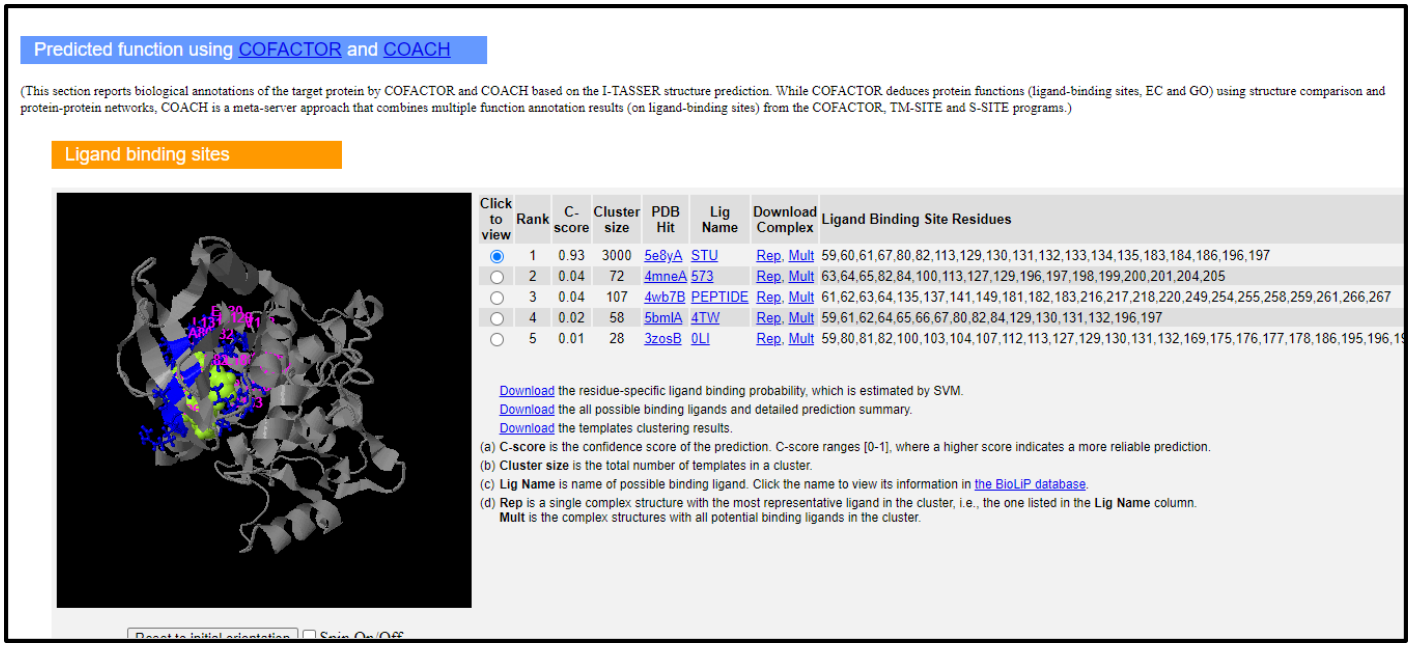


Fig10. Result for predicted functions

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.

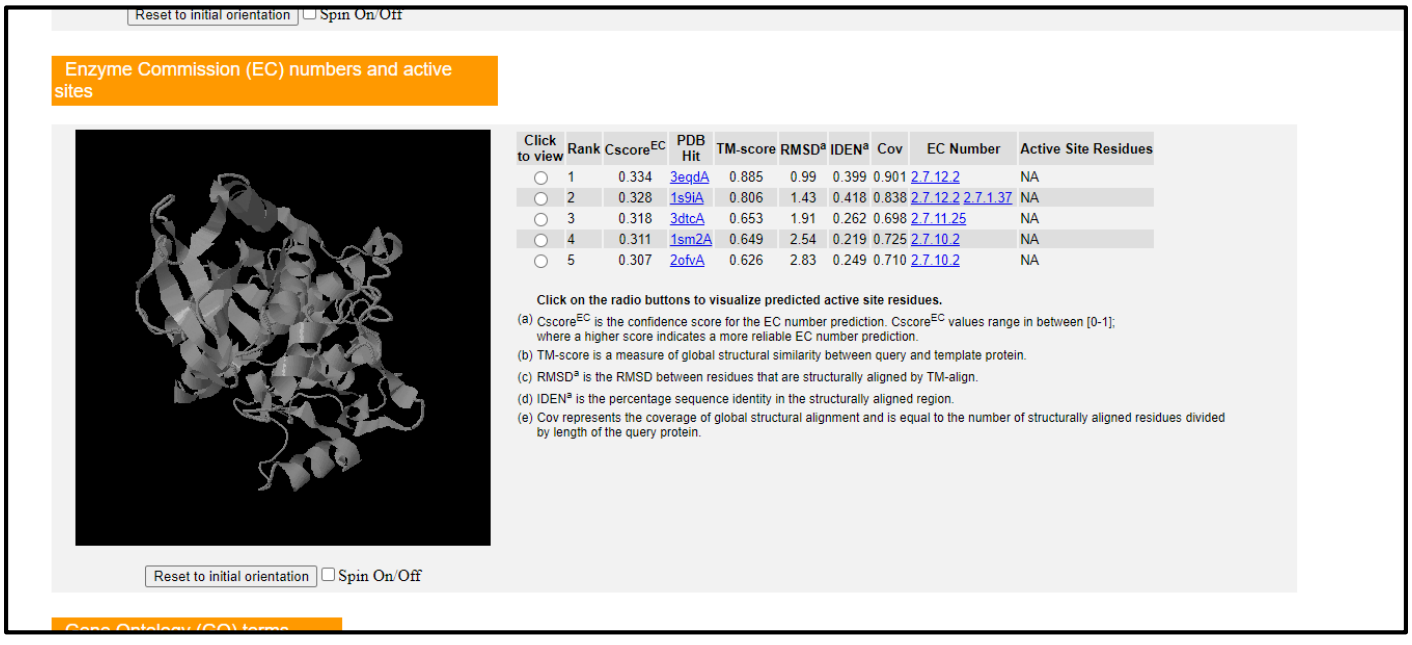
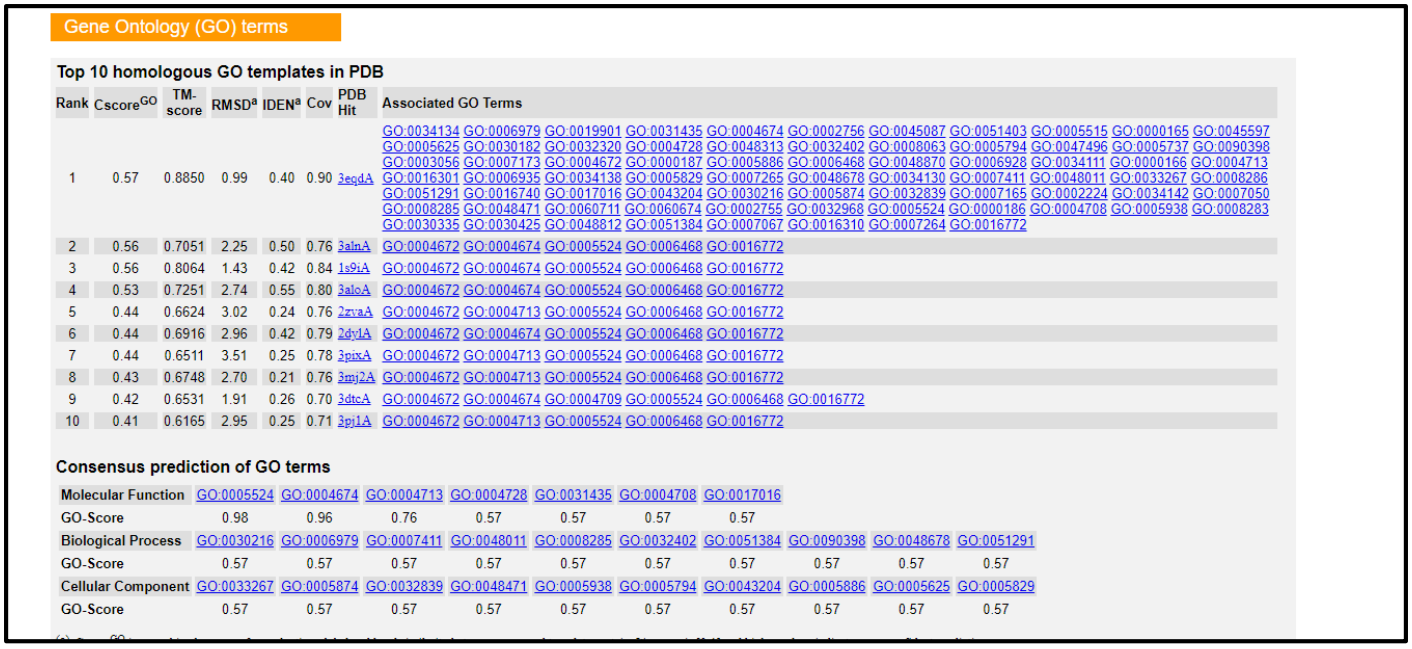


Fig11. Enzyme commission (EC) numbers and active sites

a) CscoreEC is the confidence score for the EC number prediction. CscoreEC values range in between [0-1]; where a higher score indicates a more reliable EC number prediction.
b) TM-score is a measure of global structural similarity between query and template protein.
c) RMSDa is the RMSD between residues that are structurally aligned by TM-align.
d) IDENa is the percentage sequence identity in the structurally aligned region.
e) Cov represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein.



Gene Ontology (GO) terms

Top 10 homologous GO templates in PDB

Rank	CscoreGO	TM-score	RMSDa	IDENa	Cov	PDB Hit	Associated GO Terms
1	0.57	0.8850	0.99	0.40	0.90	3eqdA	GO:0034134 GO:0006979 GO:0019901 GO:0031435 GO:0004674 GO:0002756 GO:0045087 GO:0051403 GO:0005515 GO:0000165 GO:0045597 GO:0005625 GO:0030182 GO:0032320 GO:0004728 GO:0048313 GO:0032402 GO:0008063 GO:0005794 GO:0047496 GO:0005737 GO:0090398 GO:0003056 GO:0007173 GO:0004672 GO:0000187 GO:0005886 GO:0006468 GO:0048870 GO:0006928 GO:0034111 GO:0000166 GO:0004713 GO:0016301 GO:0006935 GO:0034138 GO:0005829 GO:0007265 GO:0048678 GO:0034130 GO:0007411 GO:0048011 GO:0033267 GO:0008286 GO:0051291 GO:0016740 GO:0017016 GO:0043204 GO:0030216 GO:0005874 GO:0032839 GO:0007165 GO:0002224 GO:0034142 GO:0007050 GO:0008285 GO:0048471 GO:0060711 GO:0060674 GO:0002755 GO:0032968 GO:0005524 GO:0000186 GO:0004708 GO:0005938 GO:0008283 GO:0030335 GO:0030425 GO:0048812 GO:0051384 GO:0007067 GO:0016310 GO:0007264 GO:0016772
2	0.56	0.7051	2.25	0.50	0.76	3alnA	GO:0004672 GO:0004674 GO:0005524 GO:0006468 GO:0016772
3	0.56	0.8064	1.43	0.42	0.84	1s9iA	GO:0004672 GO:0004674 GO:0005524 GO:0006468 GO:0016772
4	0.53	0.7251	2.74	0.55	0.80	3alcA	GO:0004672 GO:0004674 GO:0005524 GO:0006468 GO:0016772
5	0.44	0.6624	3.02	0.24	0.76	2zvaA	GO:0004672 GO:0004713 GO:0005524 GO:0006468 GO:0016772
6	0.44	0.6916	2.96	0.42	0.79	2dyIA	GO:0004672 GO:0004674 GO:0005524 GO:0006468 GO:0016772
7	0.44	0.6511	3.51	0.25	0.78	3pixA	GO:0004672 GO:0004713 GO:0005524 GO:0006468 GO:0016772
8	0.43	0.6748	2.70	0.21	0.76	3mj2A	GO:0004672 GO:0004713 GO:0005524 GO:0006468 GO:0016772
9	0.42	0.6531	1.91	0.26	0.70	3dtcA	GO:0004672 GO:0004674 GO:0004709 GO:0005524 GO:0006468 GO:0016772
10	0.41	0.6165	2.95	0.25	0.71	3pj1A	GO:0004672 GO:0004713 GO:0005524 GO:0006468 GO:0016772

Consensus prediction of GO terms

Molecular Function	GO:0005524	GO:0004674	GO:0004713	GO:0004728	GO:0031435	GO:0004708	GO:0017016			
GO-Score	0.98	0.96	0.76	0.57	0.57	0.57	0.57			
Biological Process	GO:0030216	GO:0006979	GO:0007411	GO:0048011	GO:0008285	GO:0032402	GO:0051384	GO:0090398	GO:0048678	GO:0051291
GO-Score	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Cellular Component	GO:0033267	GO:0005874	GO:0032839	GO:0048471	GO:0005938	GO:0005794	GO:0043204	GO:0005886	GO:0005625	GO:0005829
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

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

RESULT:

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

CONCLUSION:

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

REFERENCES:

1. Xiong, J. (2008).*Tertiary structure prediction*. *Essential bioinformatics*. Cambridge: Cambridge University Press. 214-228.
2. *kinase | Definition, Biology, & Function*. (n.d.). Encyclopedia Britannica. Retrieved March 8, 2022, from https://www.britannica.com/science/kinase
3. *I-TASSER server for protein structure and function prediction*. (n.d.-b). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org/I-TASSER/
4. *I-TASSER results*. (n.d.). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org//I-TASSER/output/S673761/

DATE: 08/03/22

WEBLEM 3c

Robetta

(URL: https://robetta.bakerlab.org/)

AIM:

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

INTRODUCTION:

Kinase, an enzyme that adds phosphate groups (PO_4^{3-}) 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 orde novo structure prediction methods. If a confident match to a protein of known structure is found using BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative modeling. If no match is found, structure predictions are made using the de novo Rosetta fragment insertion method. Experimental nuclear magnetic resonance (NMR) constraints data can also be submitted with a query sequence for RosettaNMR de novo structure determination. Other current capabilities include the prediction of the effects of mutations on protein–protein interactions using computational interface alanine scanning. The Rosetta protein design and protein–protein docking methodologies will soon be available through the server as well.

METHODOLOGY:

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

OBSERVATION:

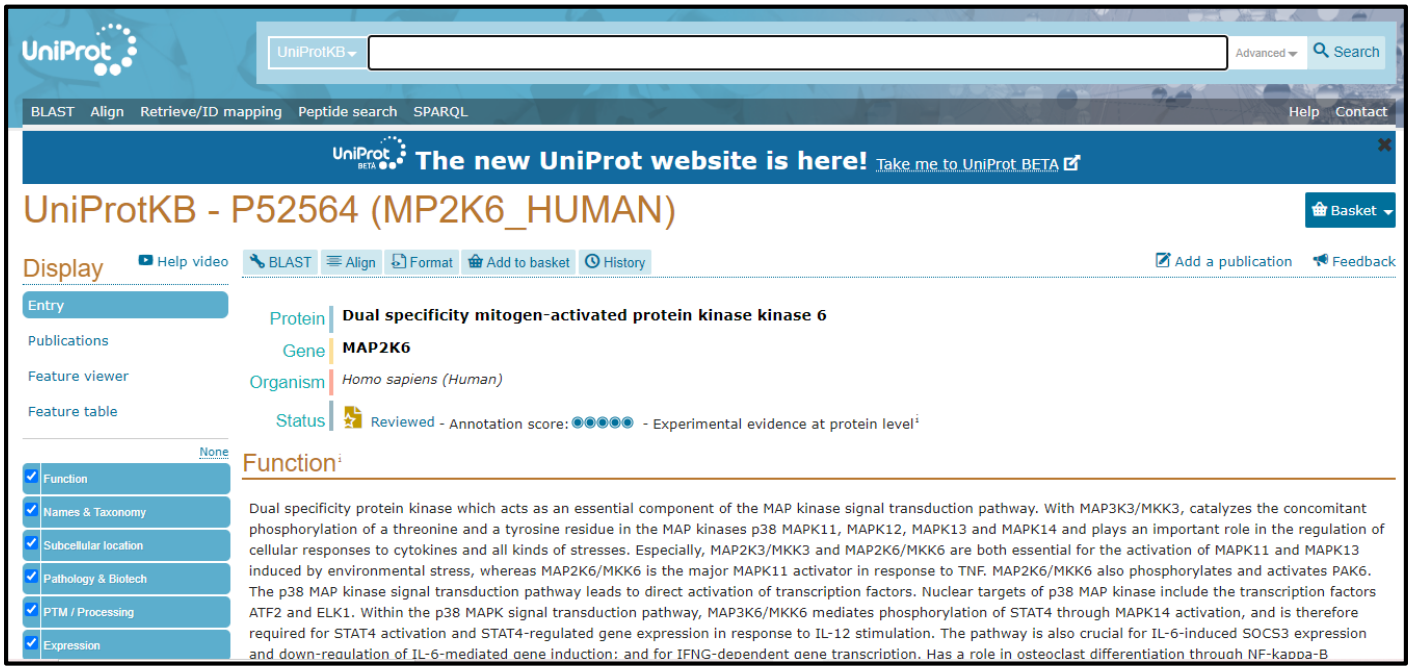
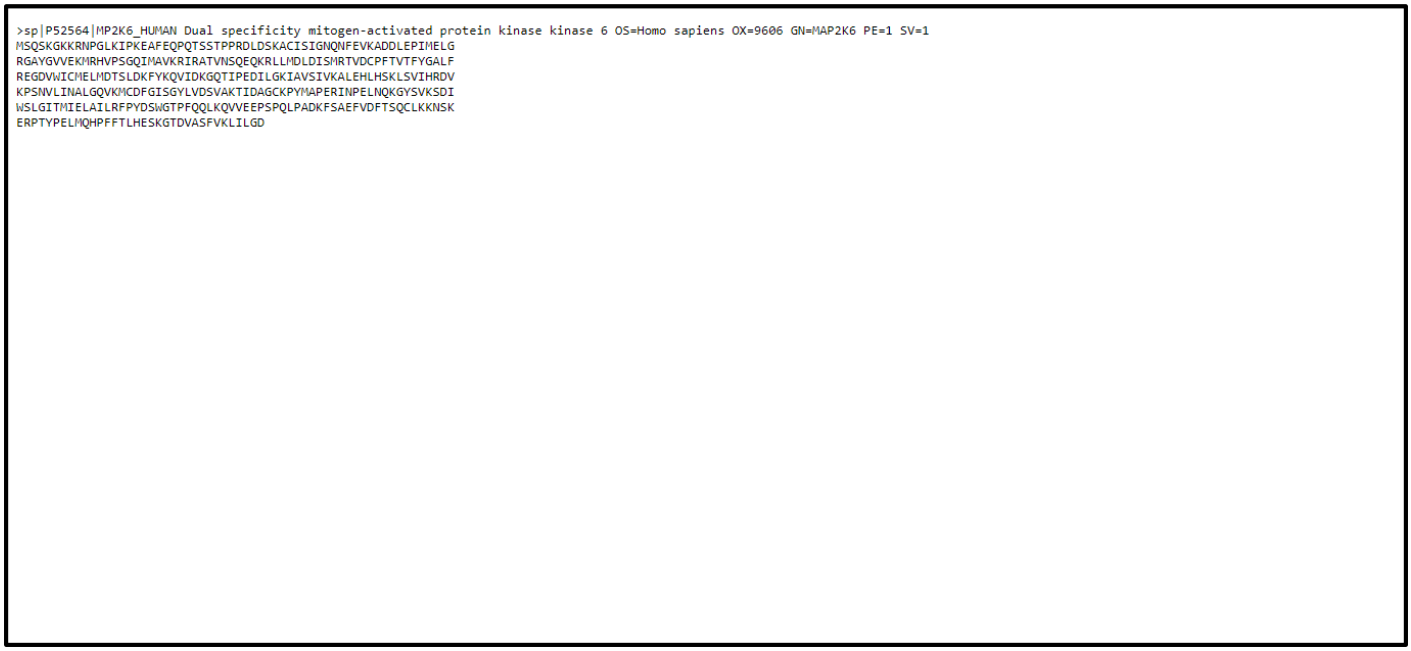


Fig1. Result page for kinase in UniProt database



```
>sp|P52564|MP2K6_HUMAN Dual specificity mitogen-activated protein kinase kinase 6 OS=Homo sapiens OX=9606 GN=MAP2K6 PE=1 SV=1
MSQSKGKKRNPGLKIPKEAFEQPQTSSTPPRDLDSKACISIGNQNFEVKADDLEPIMELG
RGAYGVVEKMRHVPSGQIMAVKRIRATVNSQEQKRLLMDLDISMRTVDCPFTVTFYGALF
REGDVWICMELMDTSLDKFYKQVIDKGQTIPEDILGKIAVSIVKALEHLHSKLSVIHRDV
KPSNVLINALGQVKMCDFGISGYLVDSVAKTIDAGCKPYMAPERINPELNQKGYSVKSDI
WSLGITMIELAILRFPYDSWGTPFQQLKQVVEEPSPQLPADKFSAEFVDFTSQCLKKNSK
ERPTYPELMQHPFFTLHESKGTDVASFVKLILGD
```

Fig2. FASTA sequence for kinase

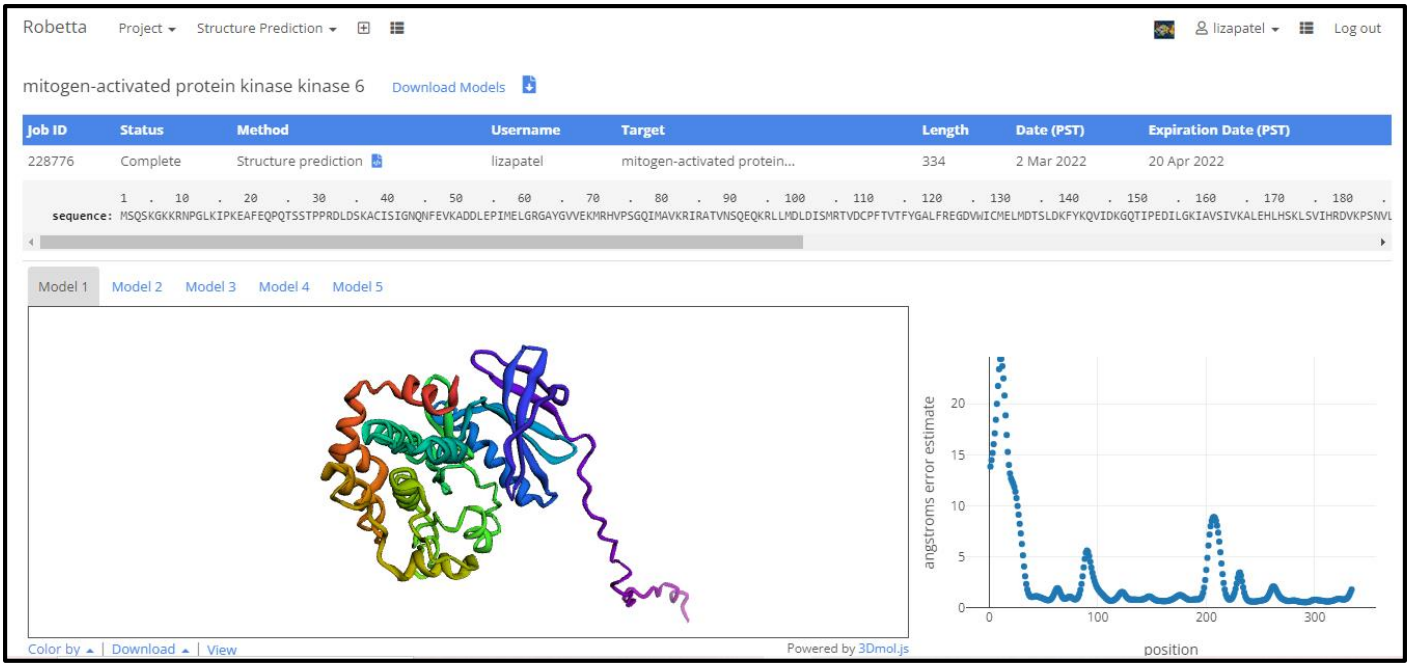


Fig3. Model 1 with atom co-ordinates

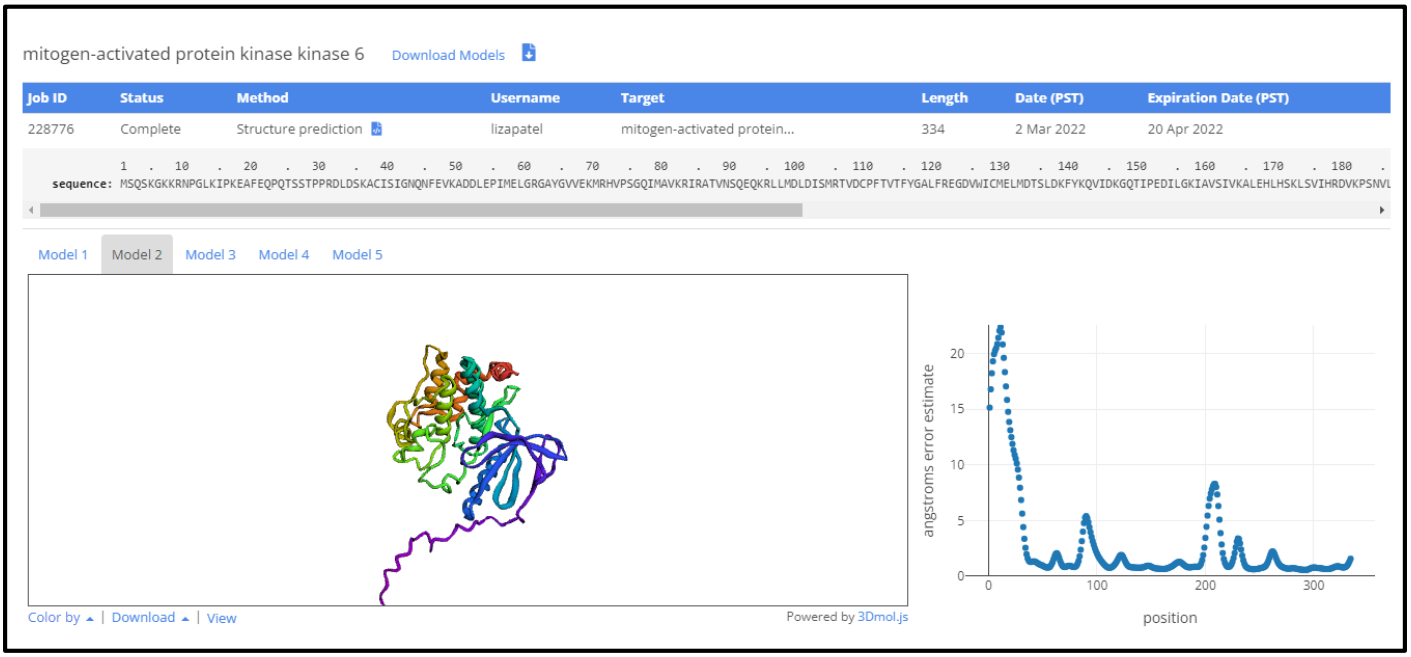


Fig4. Model 2 with atom co-ordinates

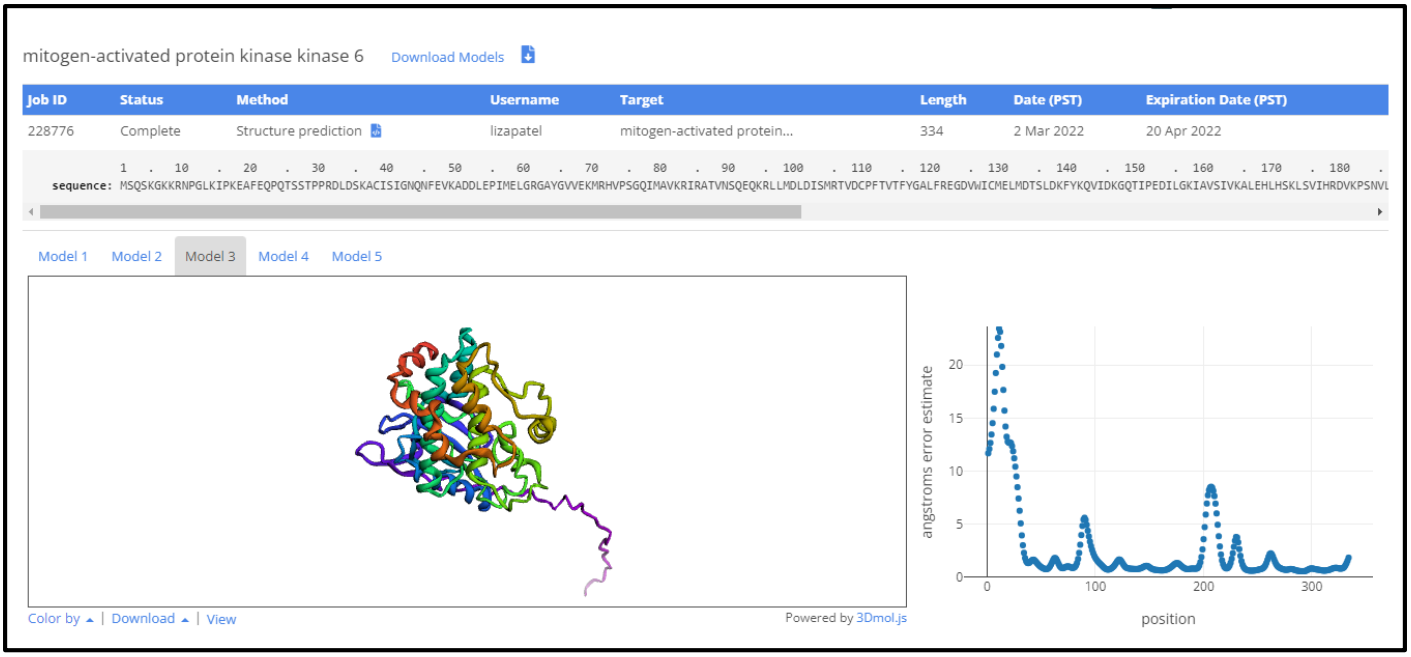


Fig5. Model 3 with atom co-ordinates

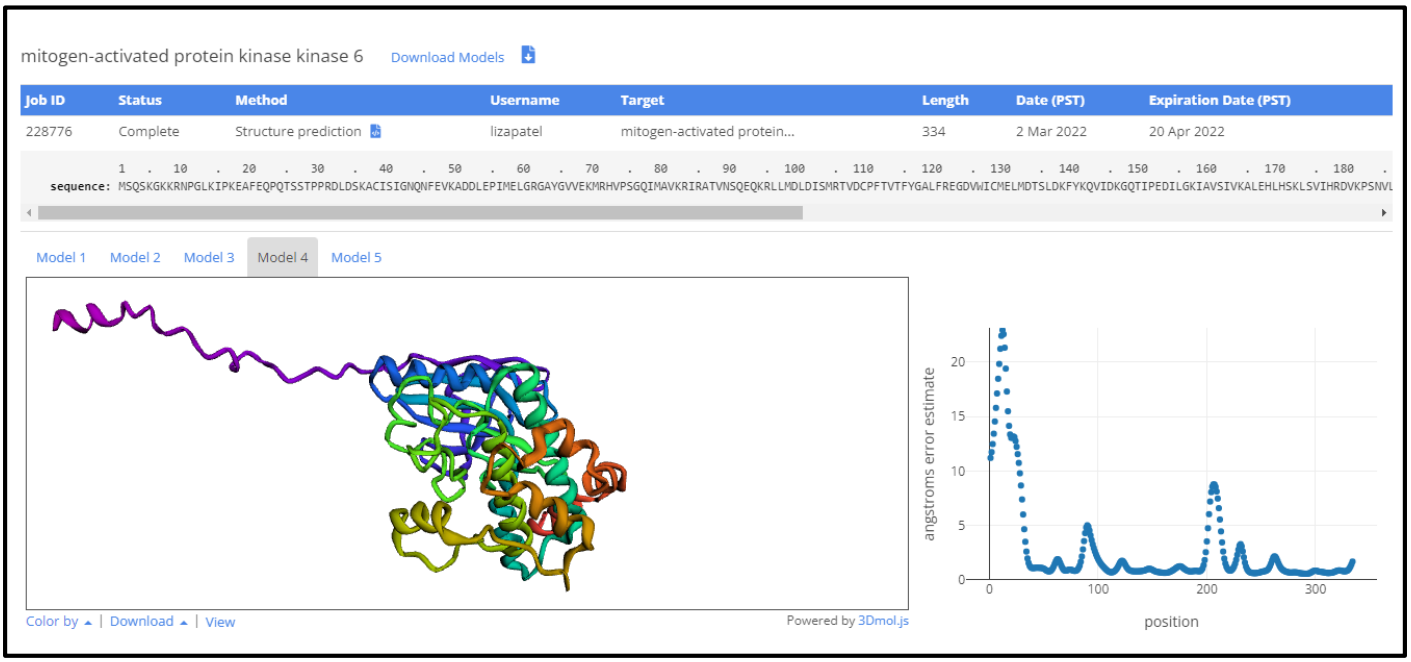


Fig6. Model 4 with atom co-ordinates

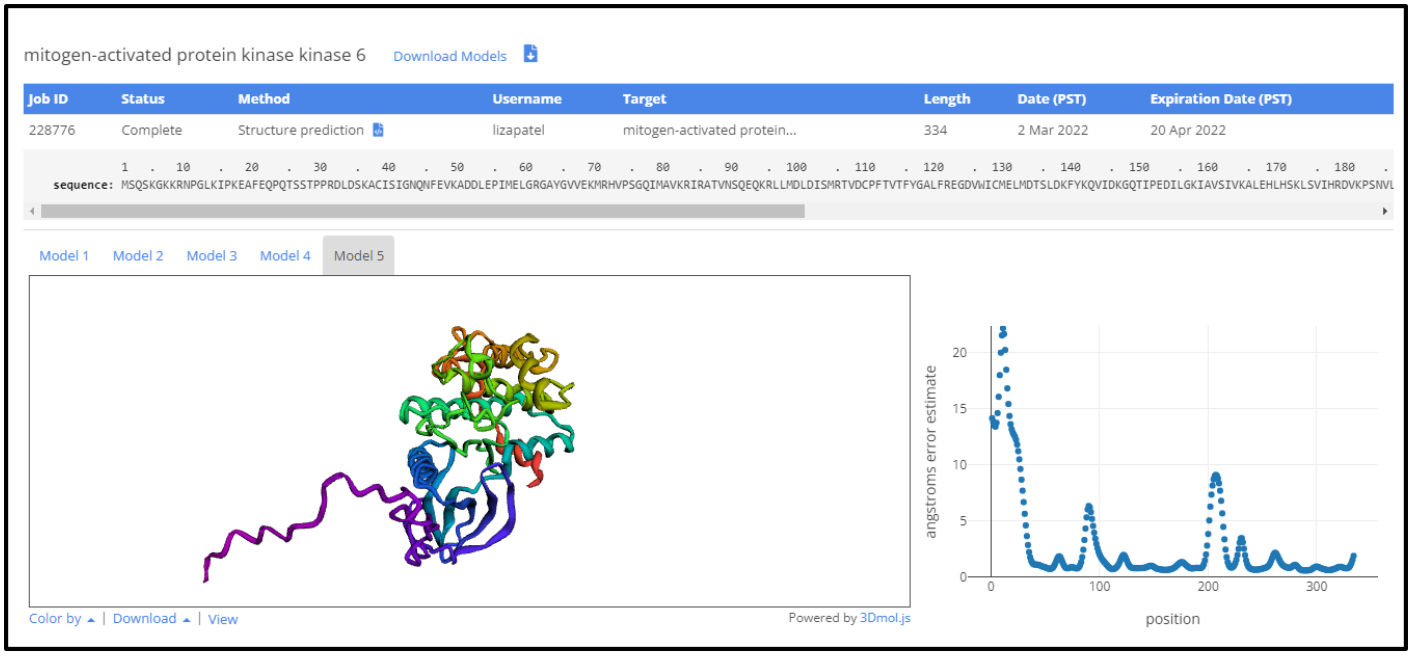


Fig7. Model 5 with atom co-ordinates

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://robetta.bakerlab.org/
4. *mitogen-activated protein kinase kinase 6*. (n.d.). Robetta.bakerlab.org. Retrieved March 8, 2022, from https://robetta.bakerlab.org/results.php?id=228776

DATE: 08/03/22

WEBLEM 4

Introduction to Validation server- SAVES server

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

SAVES server:

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

ERRAT:

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

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

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

Verify3D:

It is another server using the statistical approach. It uses a precomputed database containing eighteen environmental profiles based on secondary structures and solvent exposure, compiled from high-resolution protein structures. To assess the quality of a protein model, the secondary structure and solvent exposure propensity of each residue are calculated. It determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures. If the parameters of a residue fall within one of the profiles, it receives a high score, otherwise a low score. The result is a two-dimensional graph illustrating the folding quality of each residue of the protein structure. The threshold value is normally set at zero. Residues with scores below zero are considered to have an 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 sterochemical parameters of the residues in the model. WHAT IF is a comprehensive protein analysis server that validates a protein model for chemical correctness. It has many functions, including checking of planarity, collisions with symmetry axes (close contacts), proline puckering, anomalous bond angles, and bond lengths. It also allows the generation of Ramachandran plots as an assessment of the quality of the model.

PROCHECK:

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

CRYST:

This program searches the Protein Data Bank for entries that have a unit cell similar to your input file. CRYST1 record required. Use the standalone CRYST server for more options.

The assessment results can be different using different verification programs. Because no single method is clearly superior to any other, a good strategy is to use multiple verification methods and identify the consensus between them. It is also important to keep in mind that the evaluation tests performed by these programs only check the stereochemical correctness, regardless of the accuracy of the model, which may or may not have any biological meaning. Thus, SAVES server is an excellent platform that provides various validation methods to accurately validate the structures.

REFERENCES:

1. *SAVESv6.0 - Structure Validation Server*. (n.d.-b). Saves.mbi.ucla.edu. Retrieved March 8, 2022, from https://saves.mbi.ucla.edu/
2. *ERRAT – UCLA-DOE Institute*. (n.d.). Retrieved March 8, 2022, from https://www.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

DATE: 08/03/22

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

METHODOLOGY:

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

OBSERVATION:

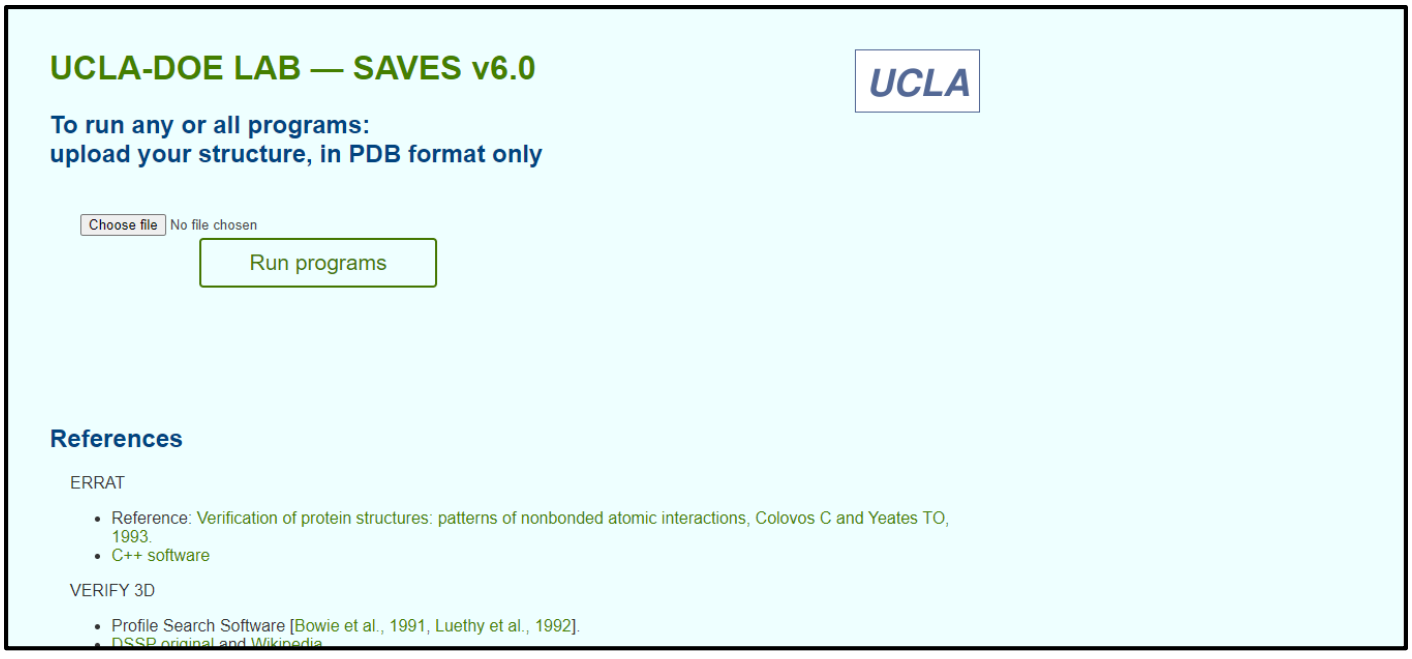


Fig1. Homepage for SAVES server

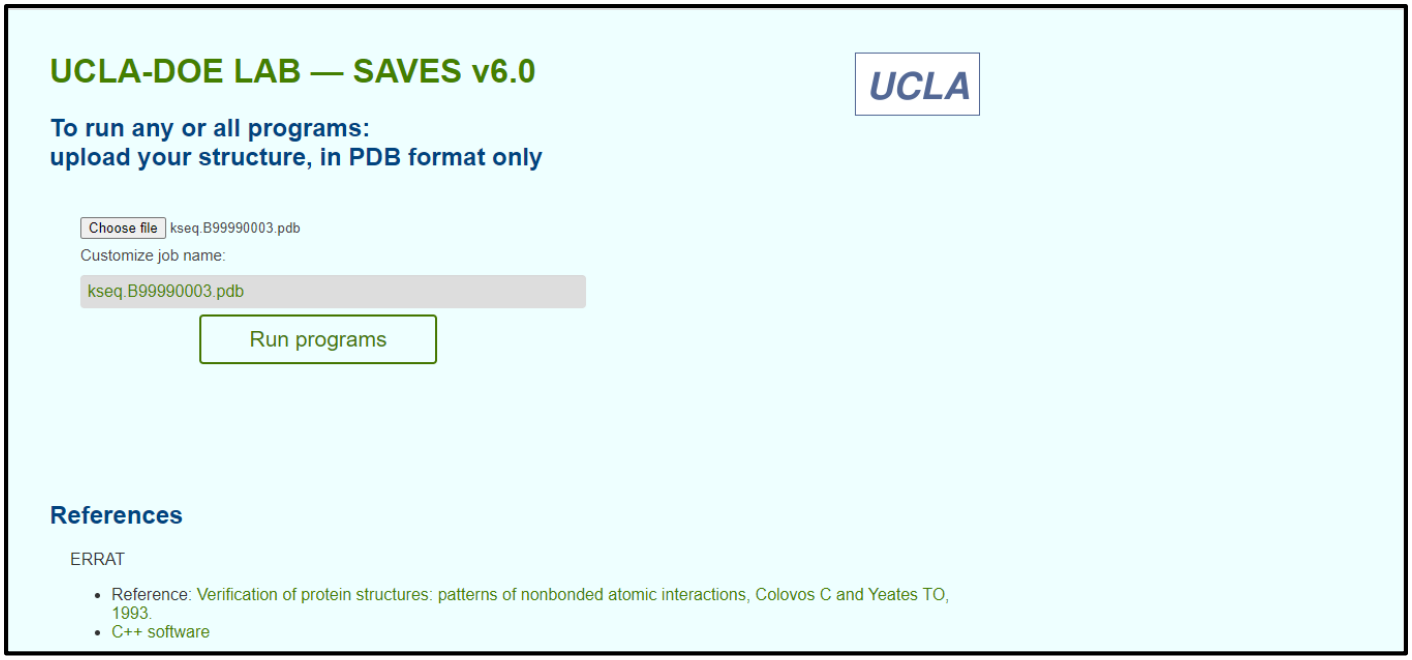


Fig2. Structure from Modeller for validation

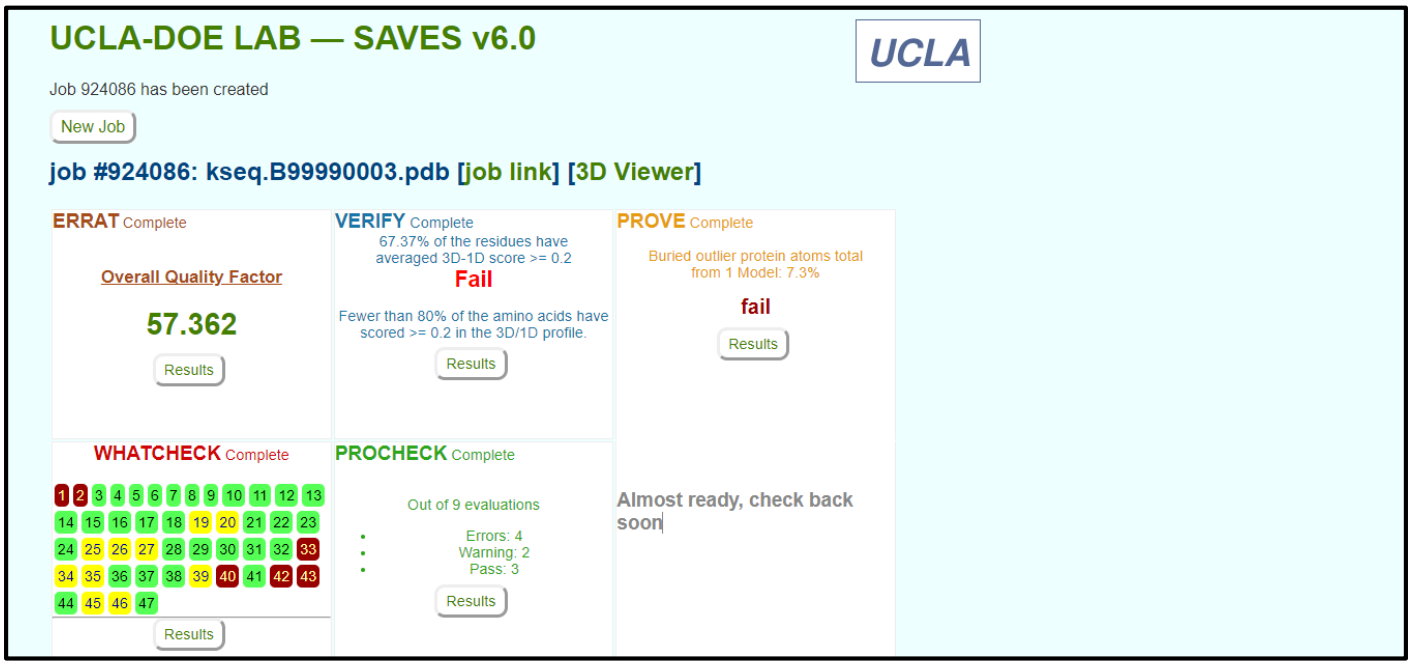


Fig3. Result page for structure validation for various servers

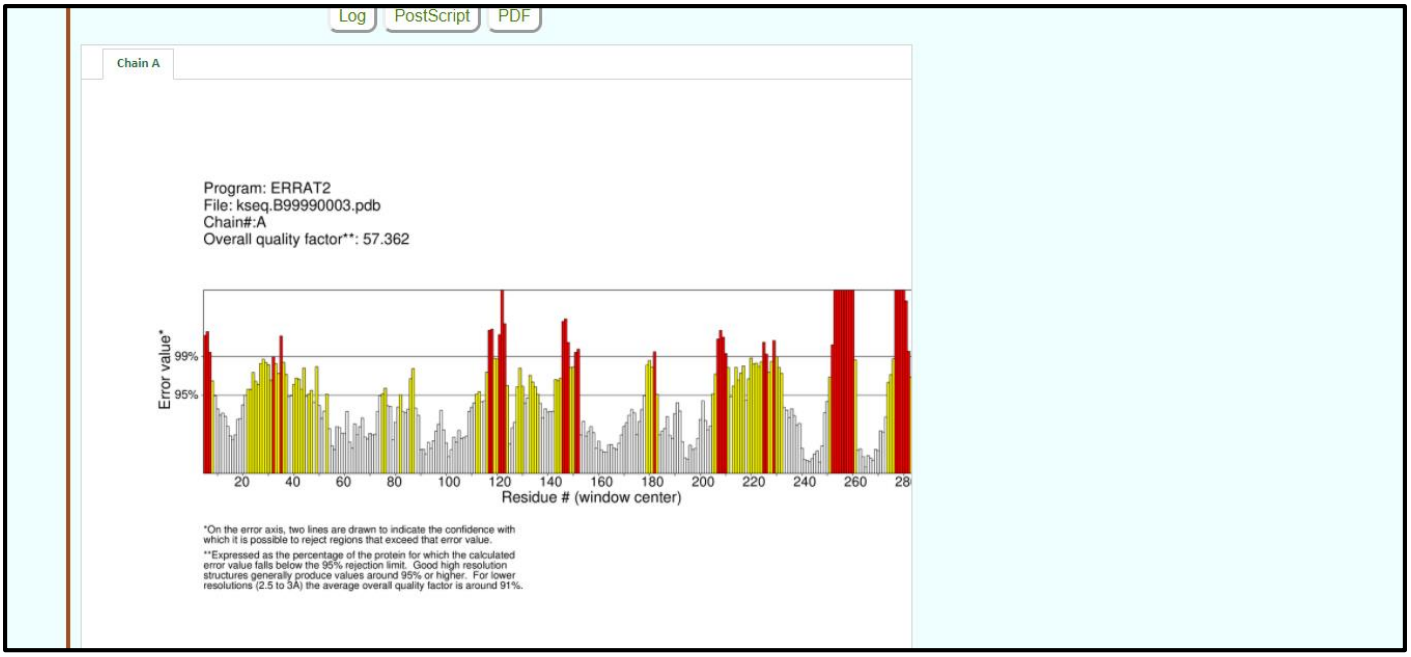


Fig4. Result page for Errat

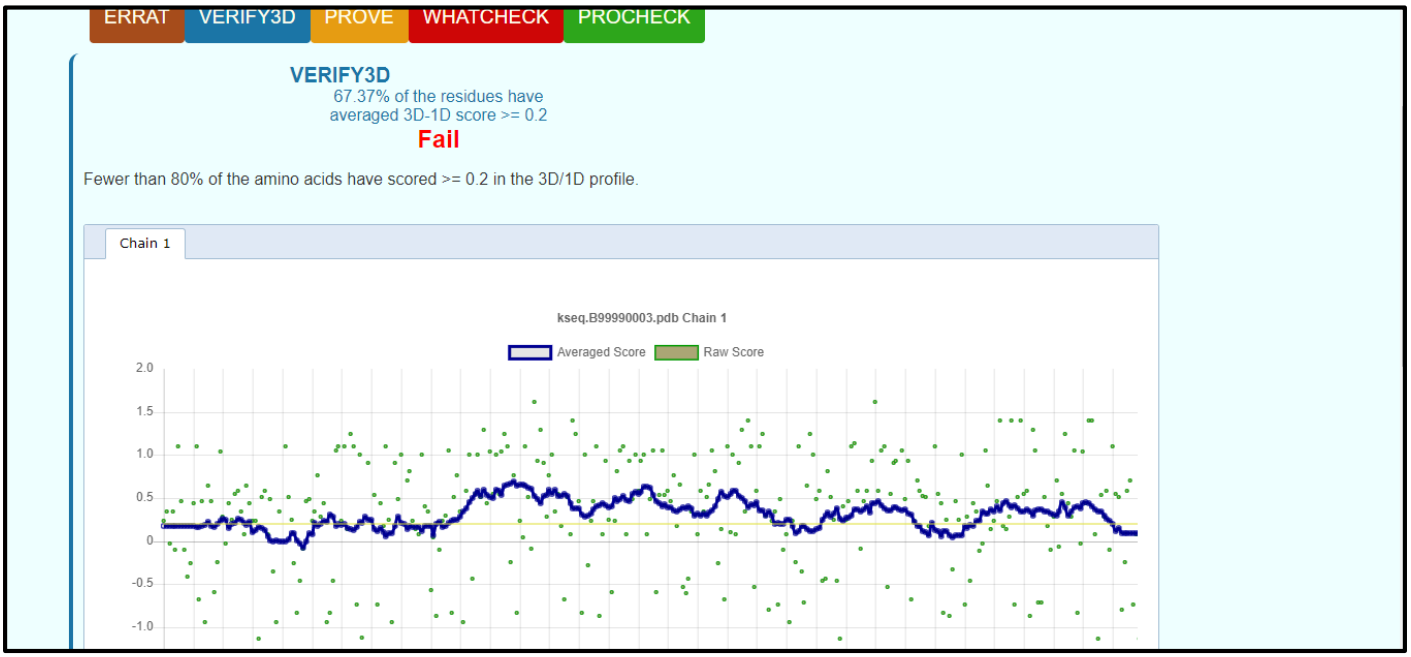


Fig5. Result page for Verify3D

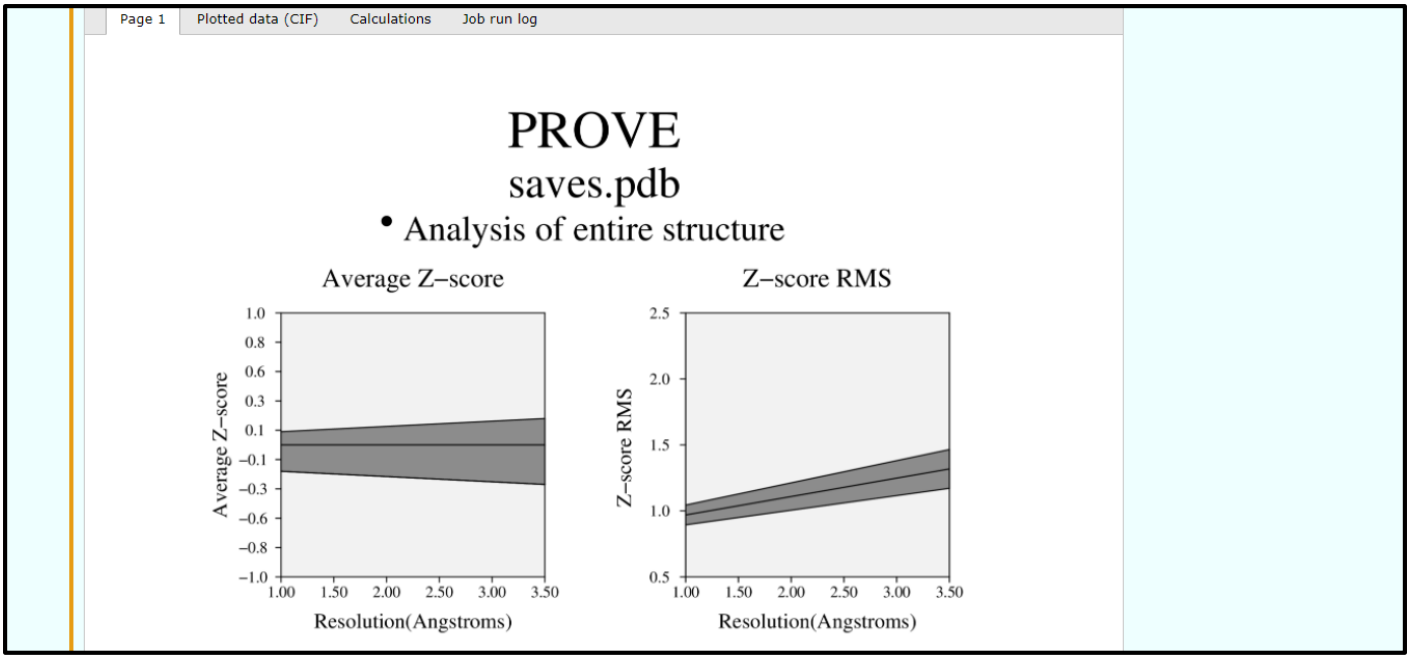


Fig6. Result page for Prove

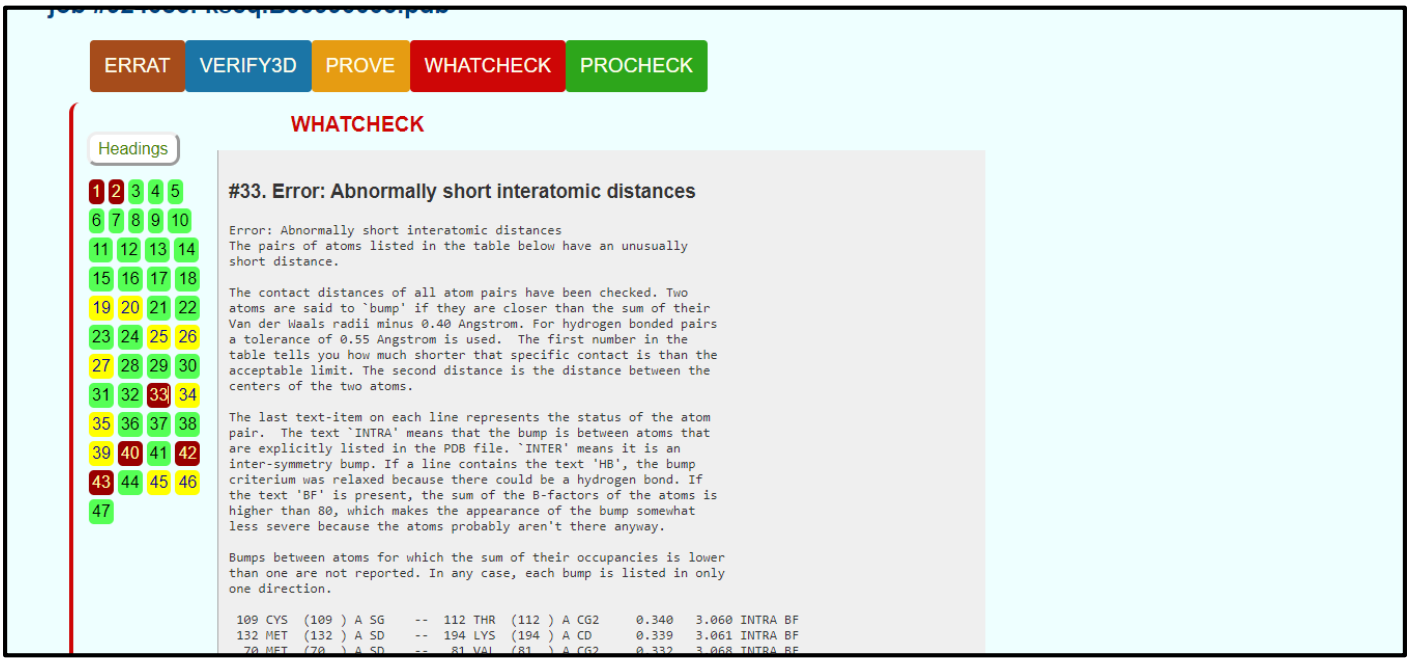


Fig7. Result page for Whatcheck

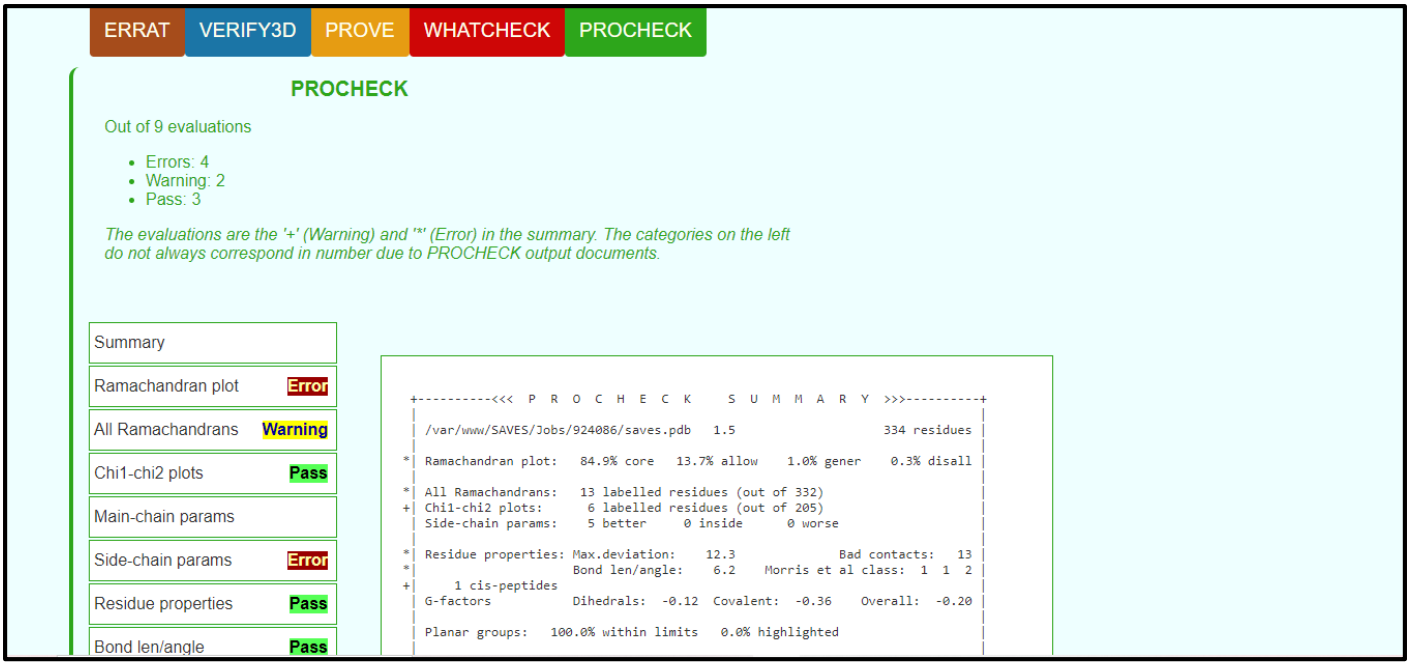


Fig8. Result page for Procheck

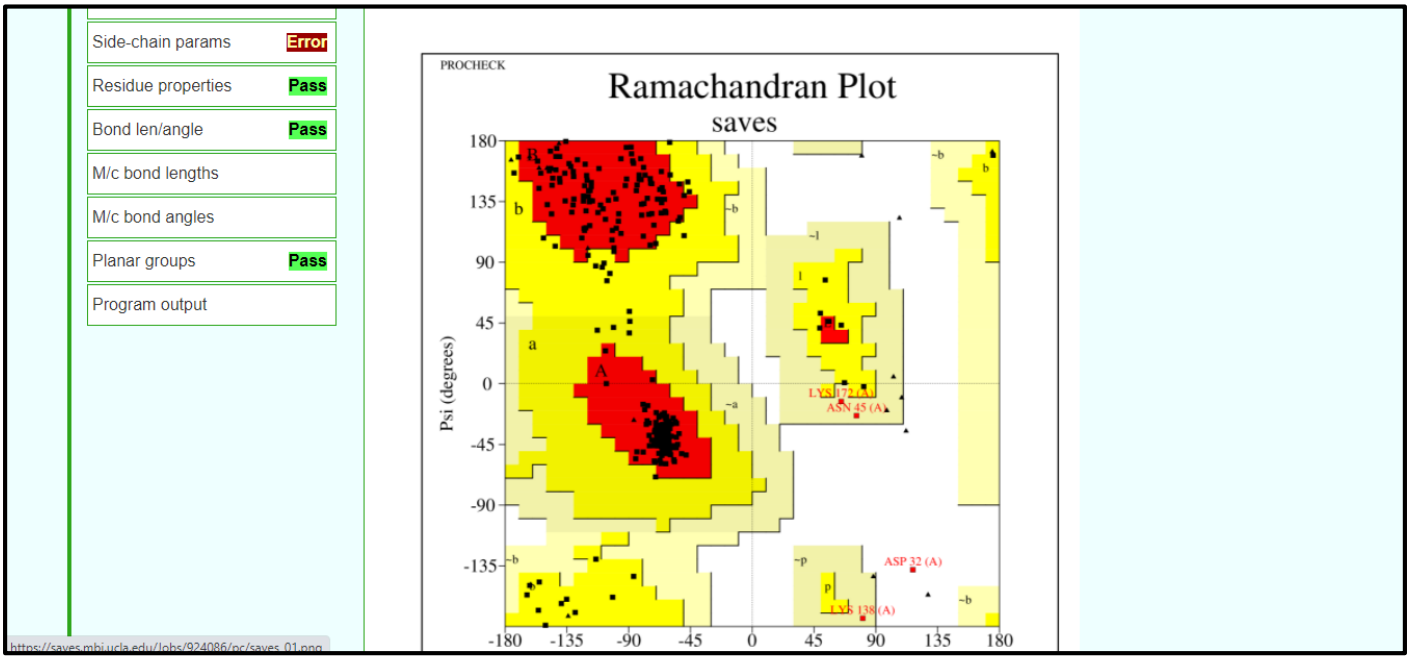


Fig8.1. Result page for Ramachandran Plot

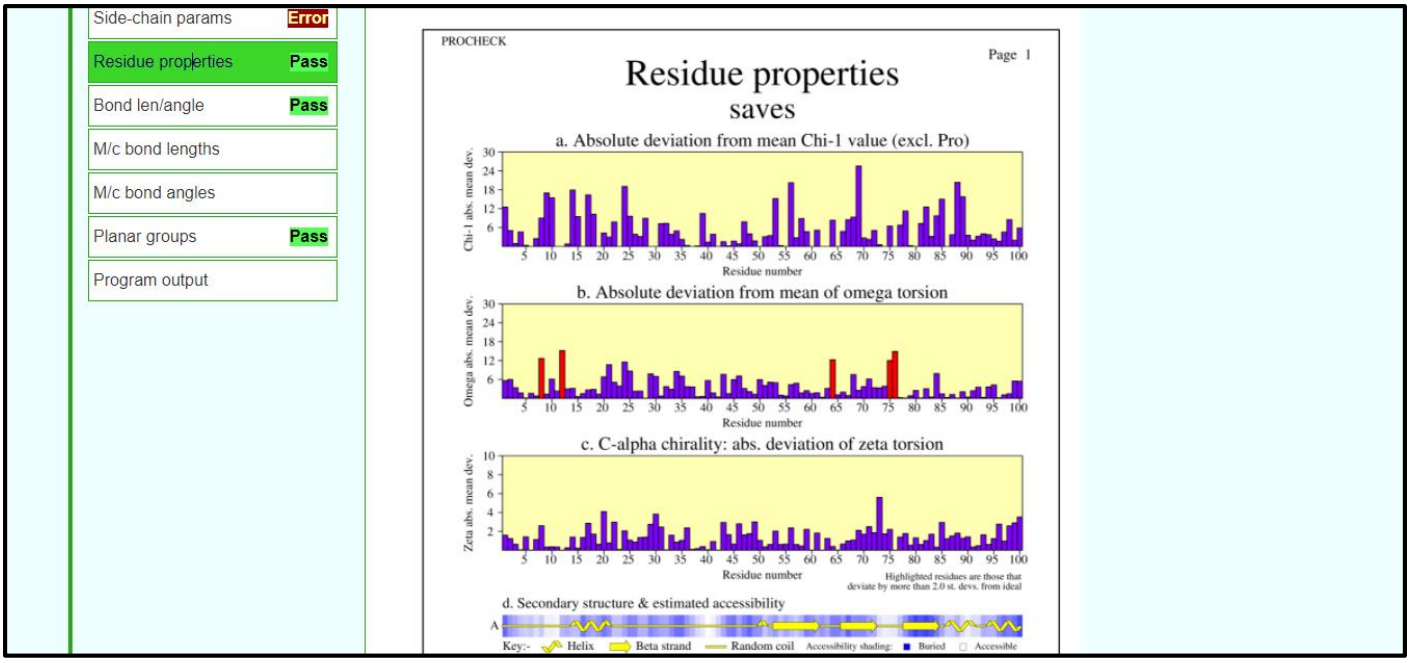


Fig8.2. Result page for residue properties

RESULT:

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

CONCLUSION:

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

REFERENCES:

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

DATE: 08/03/22

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

METHODOLOGY:

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:

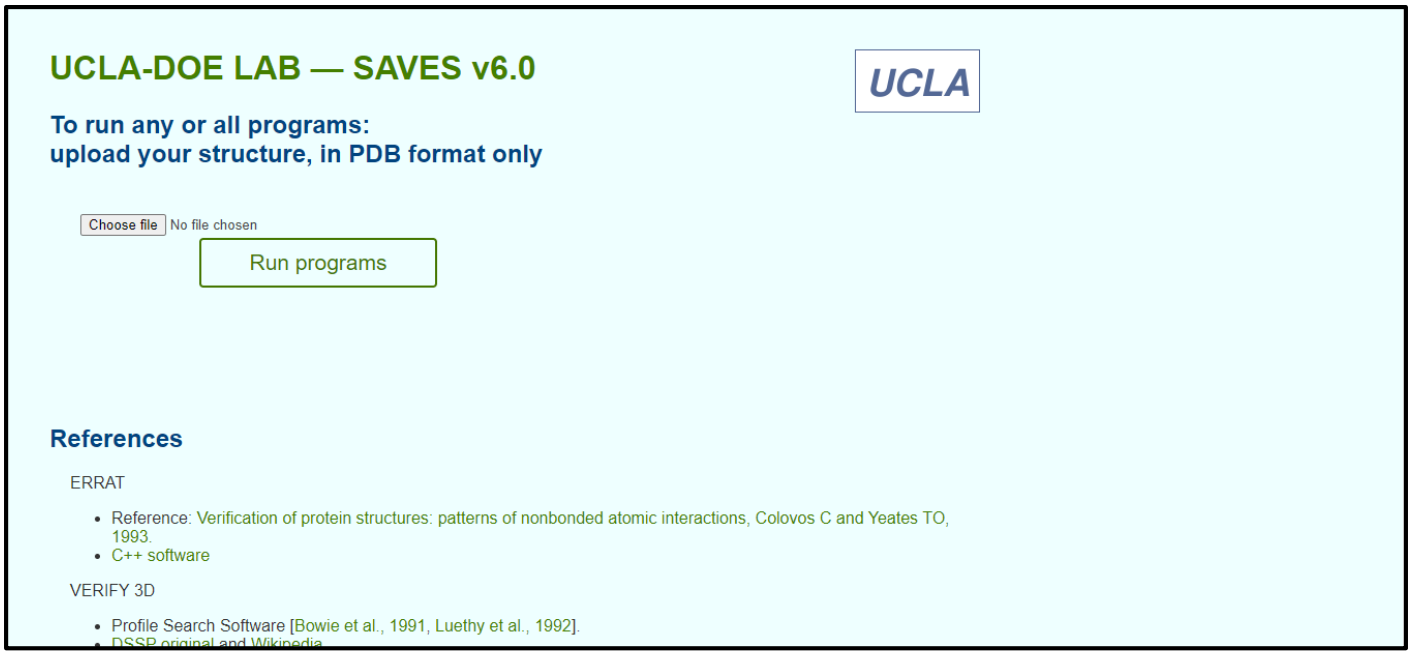


Fig1. Homepage for SAVES server

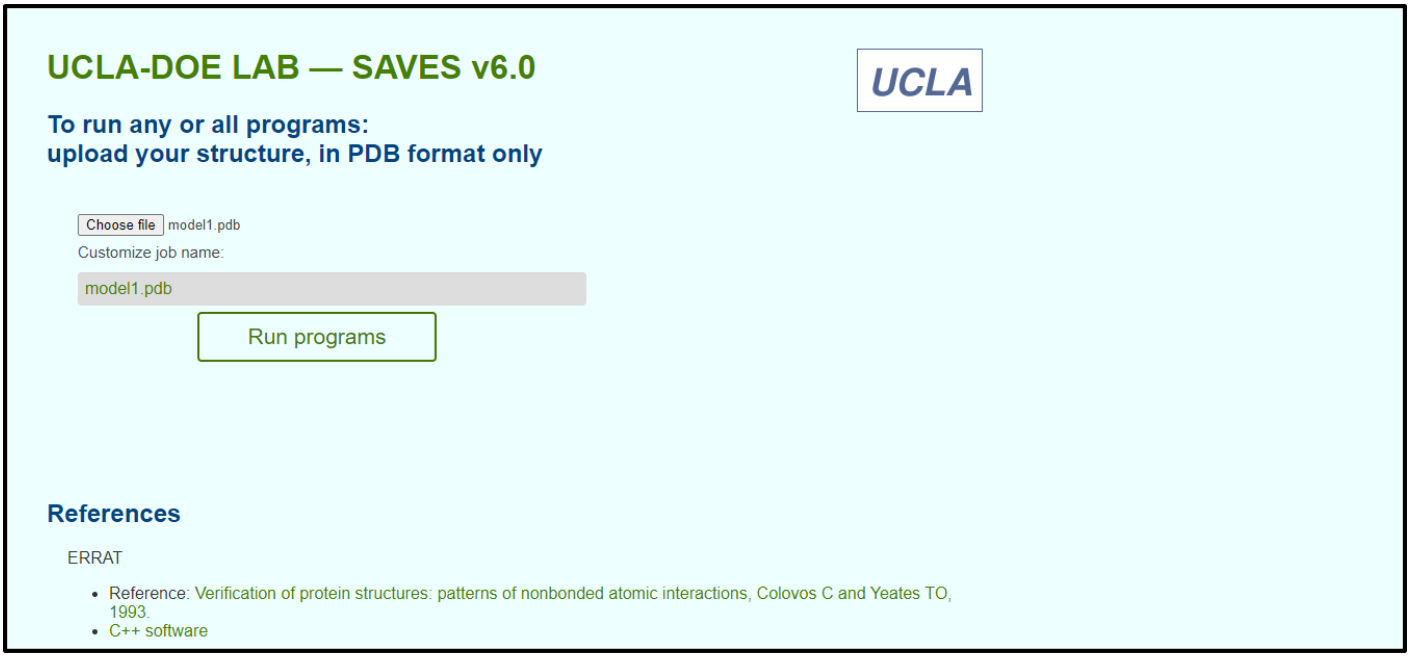


Fig2. Structure from Modeller for validation

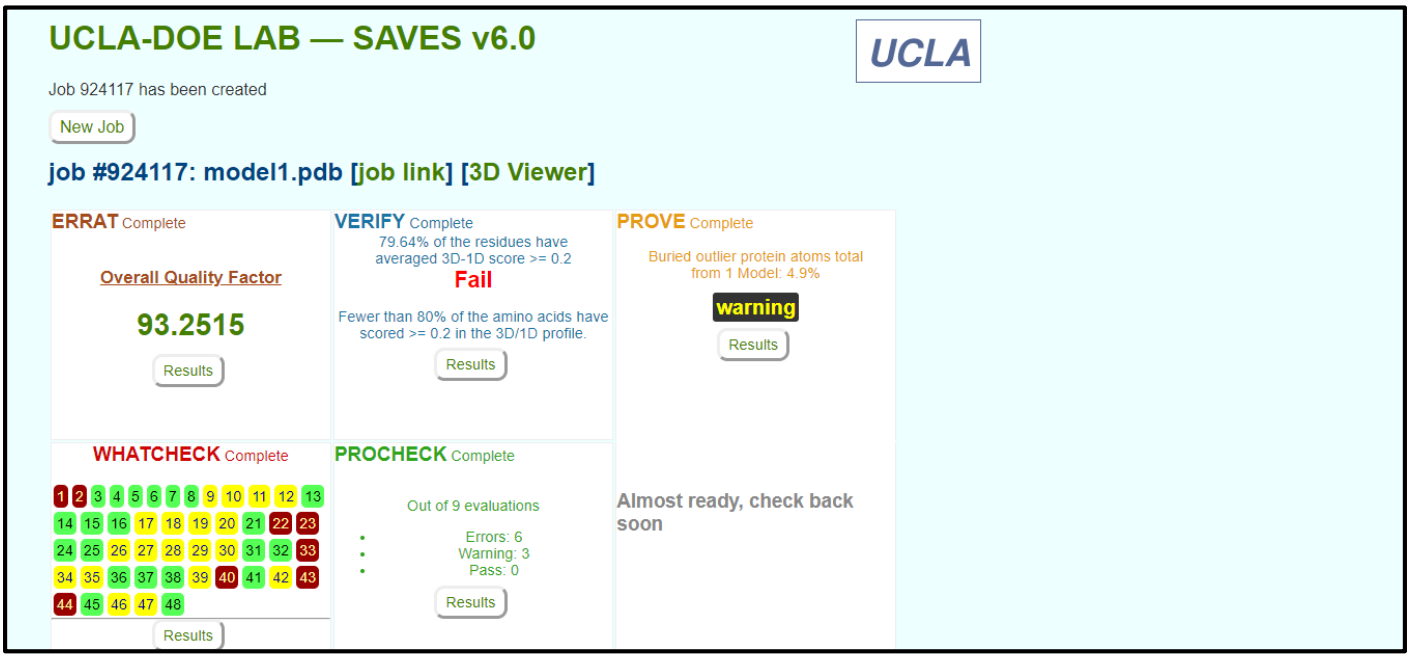


Fig3. Result page for structure validation for various servers

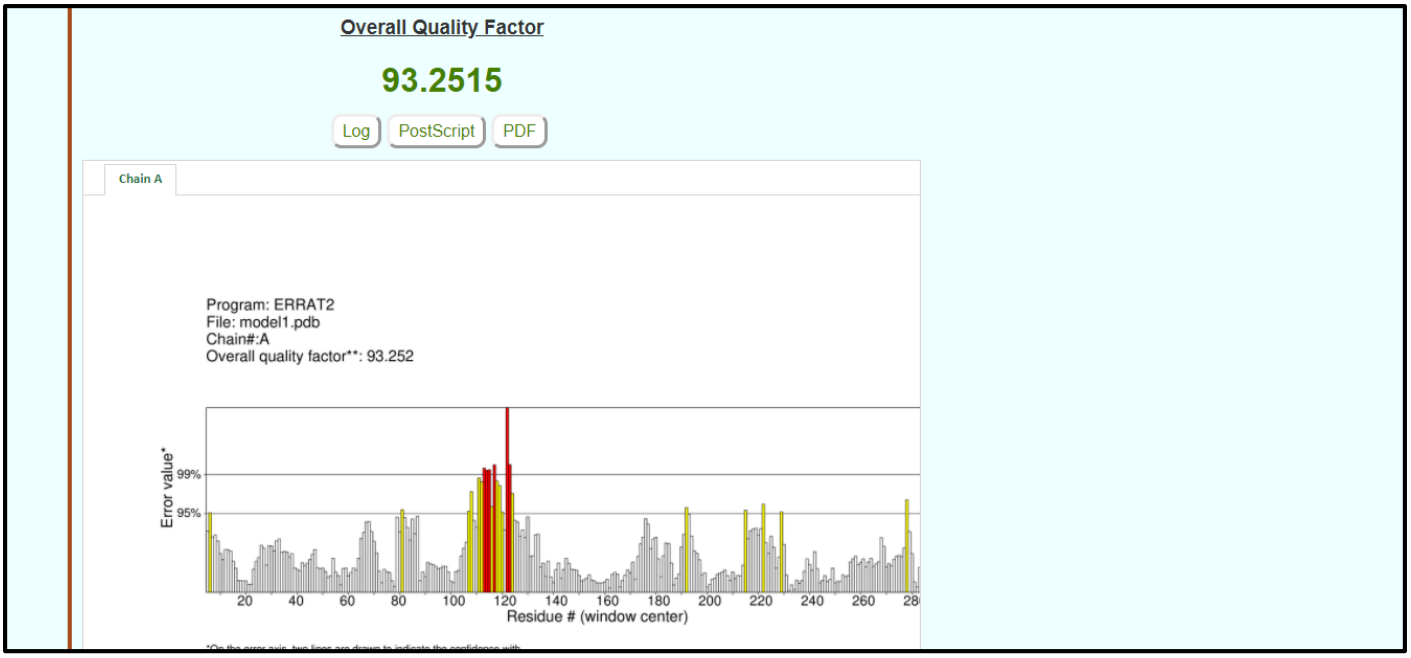


Fig4. Result page for Errat

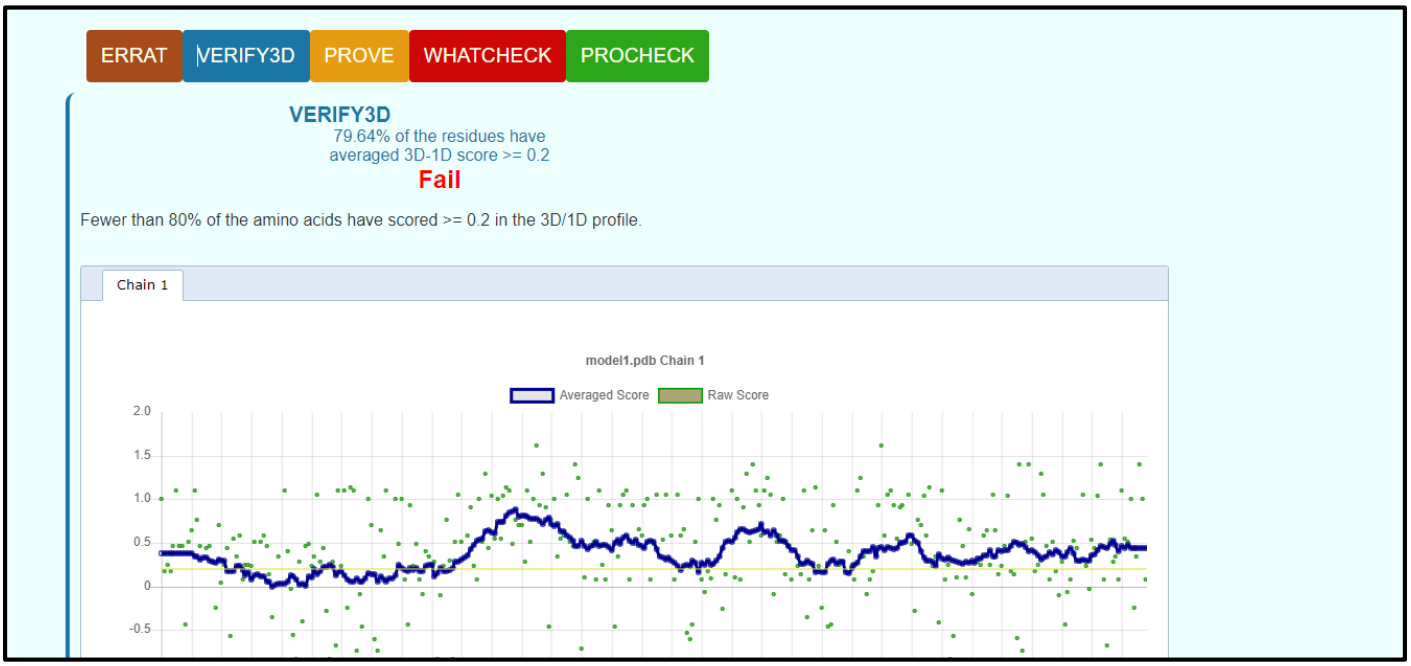


Fig5. Result page for Verify3D

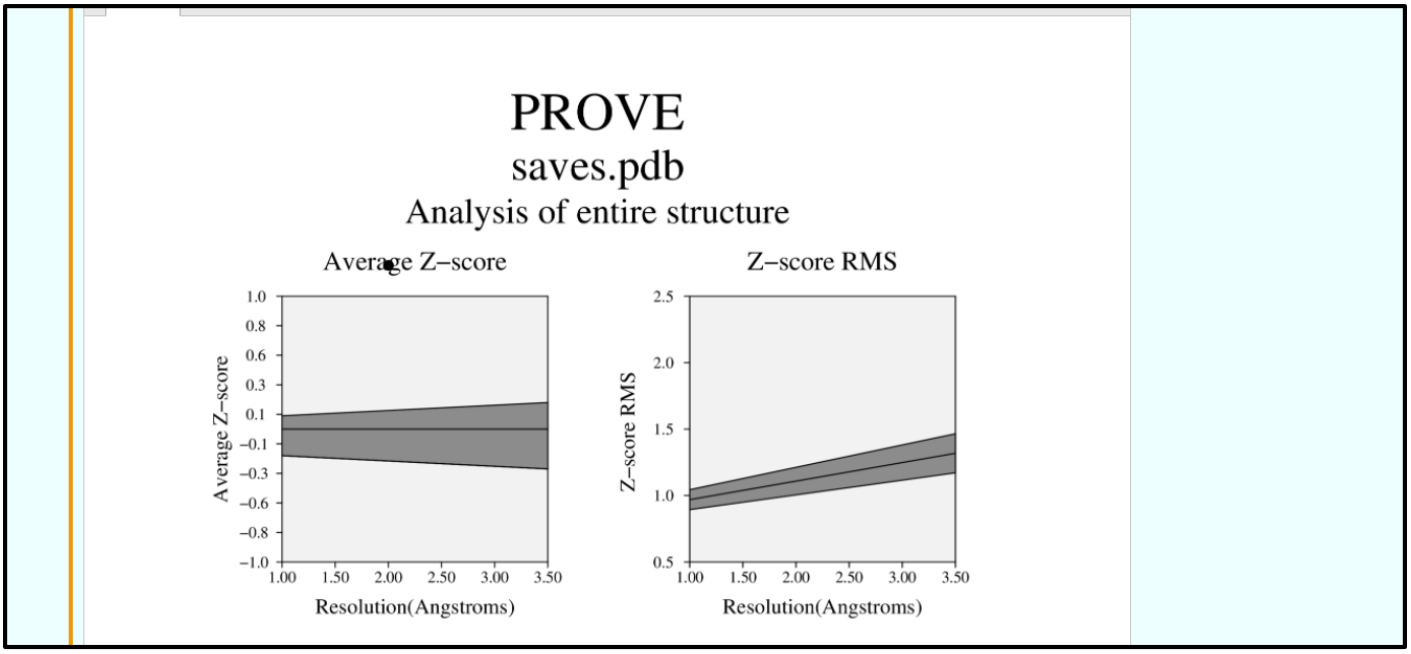


Fig6. Result page for Prove

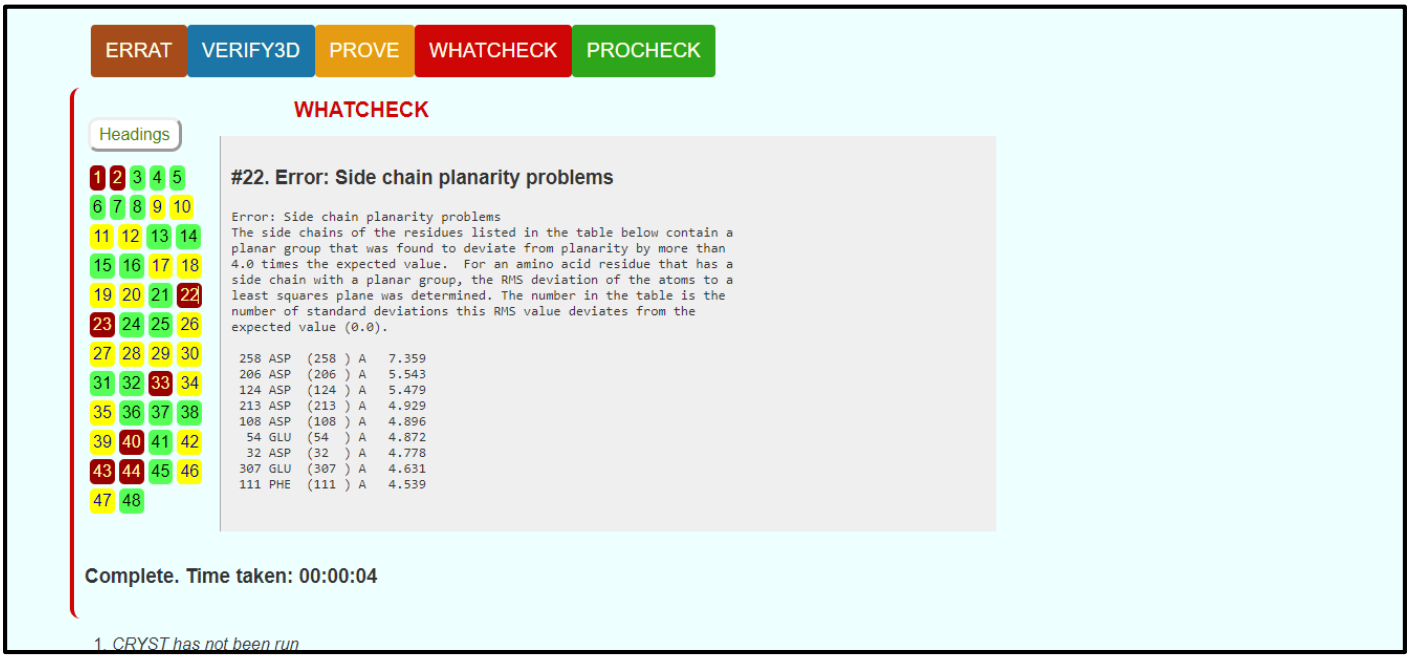


Fig7. Result page for Whatcheck

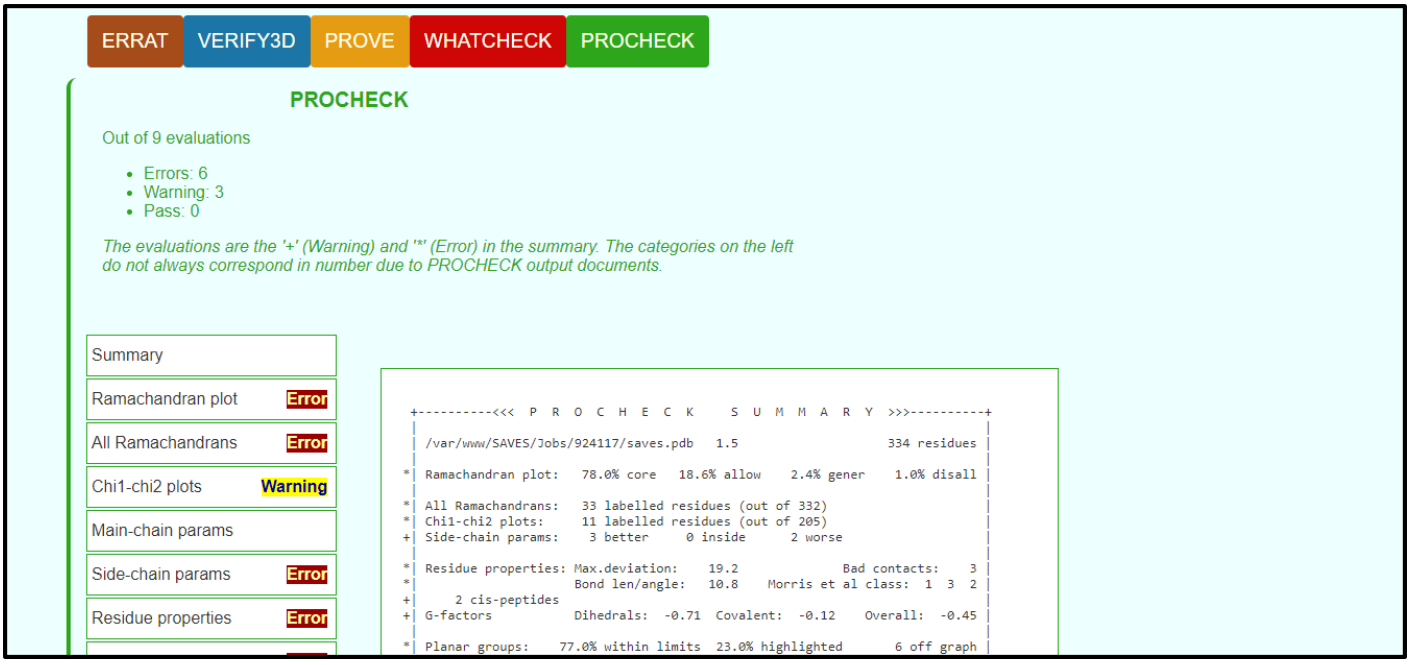


Fig8. Result page for Procheck

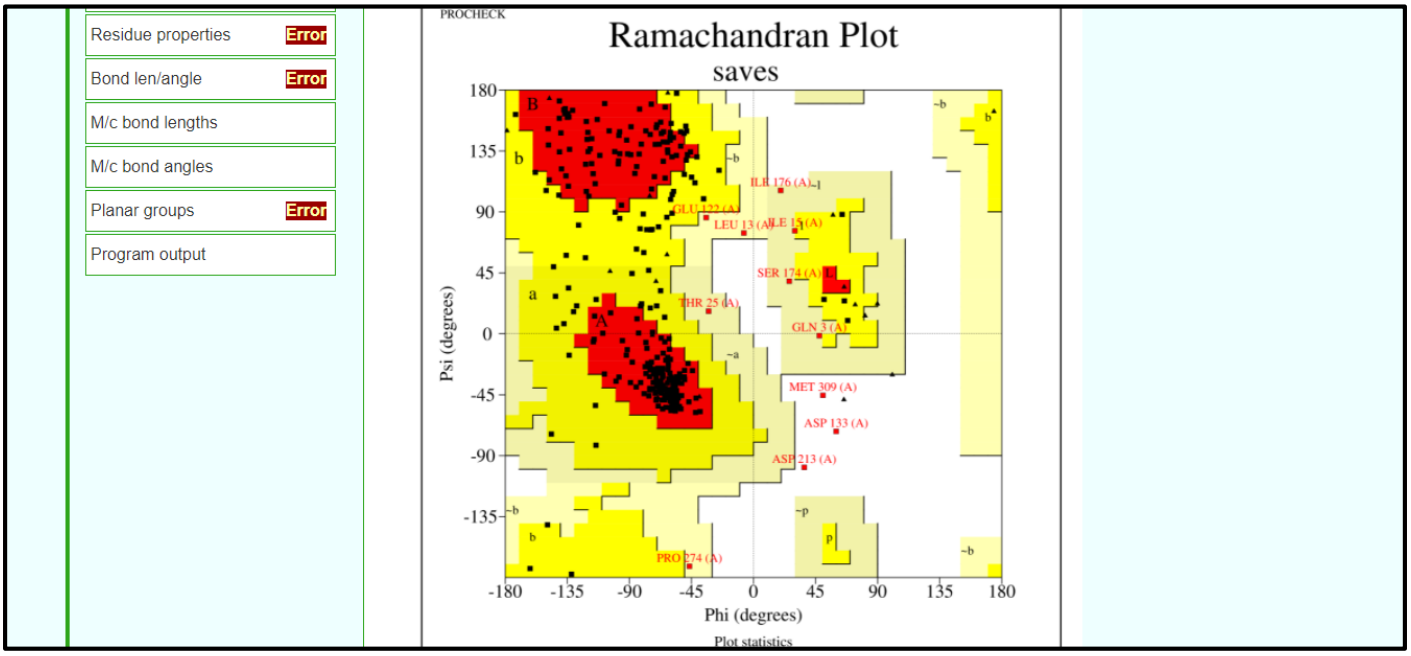


Fig8.1. Result page for Ramachandran Plot

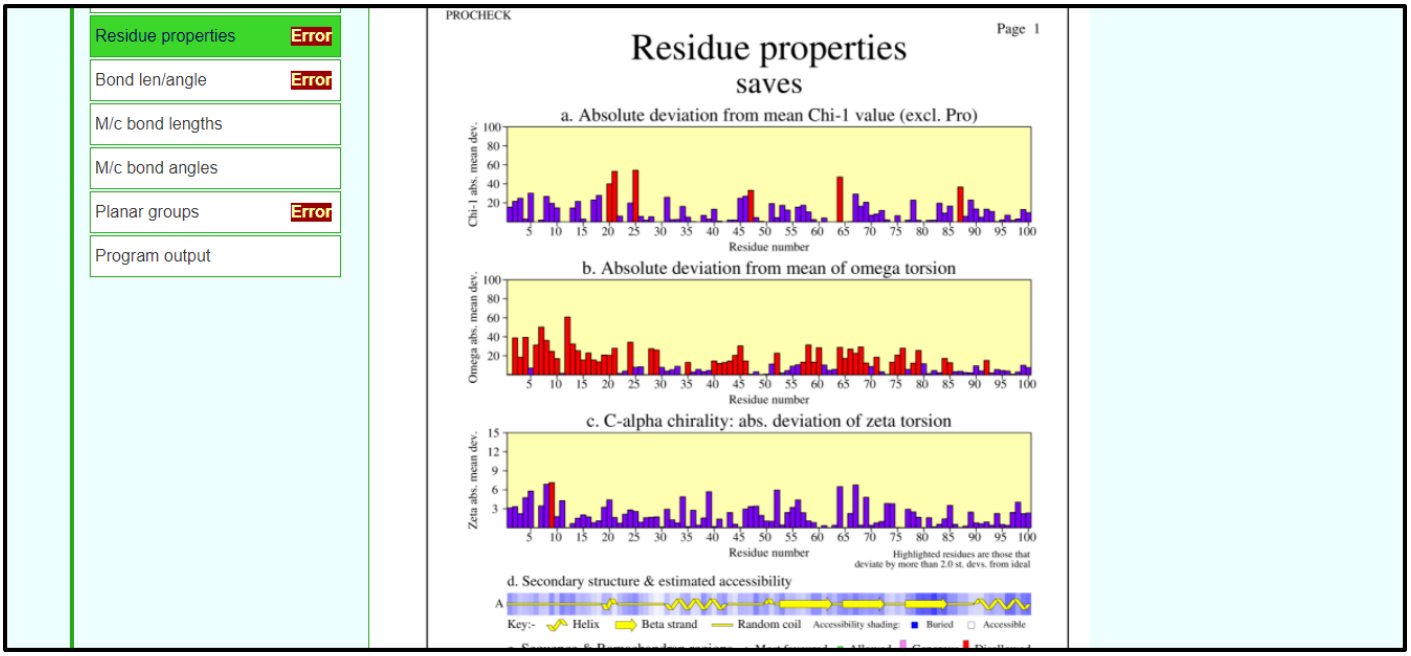


Fig8.2. Result page for residue properties

RESULT:

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

CONCLUSION:

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

REFERENCES:

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

DATE: 08/03/22

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

METHODOLOGY:

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:

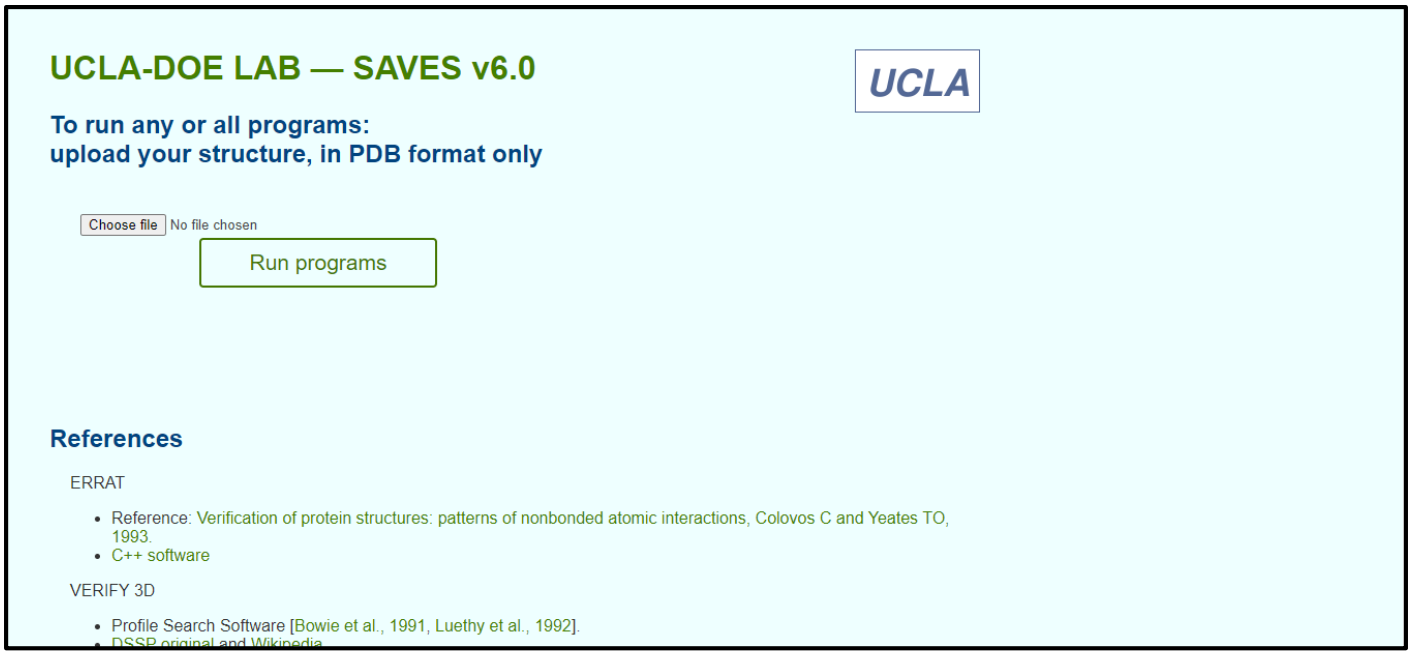


Fig1. Homepage for SAVES server

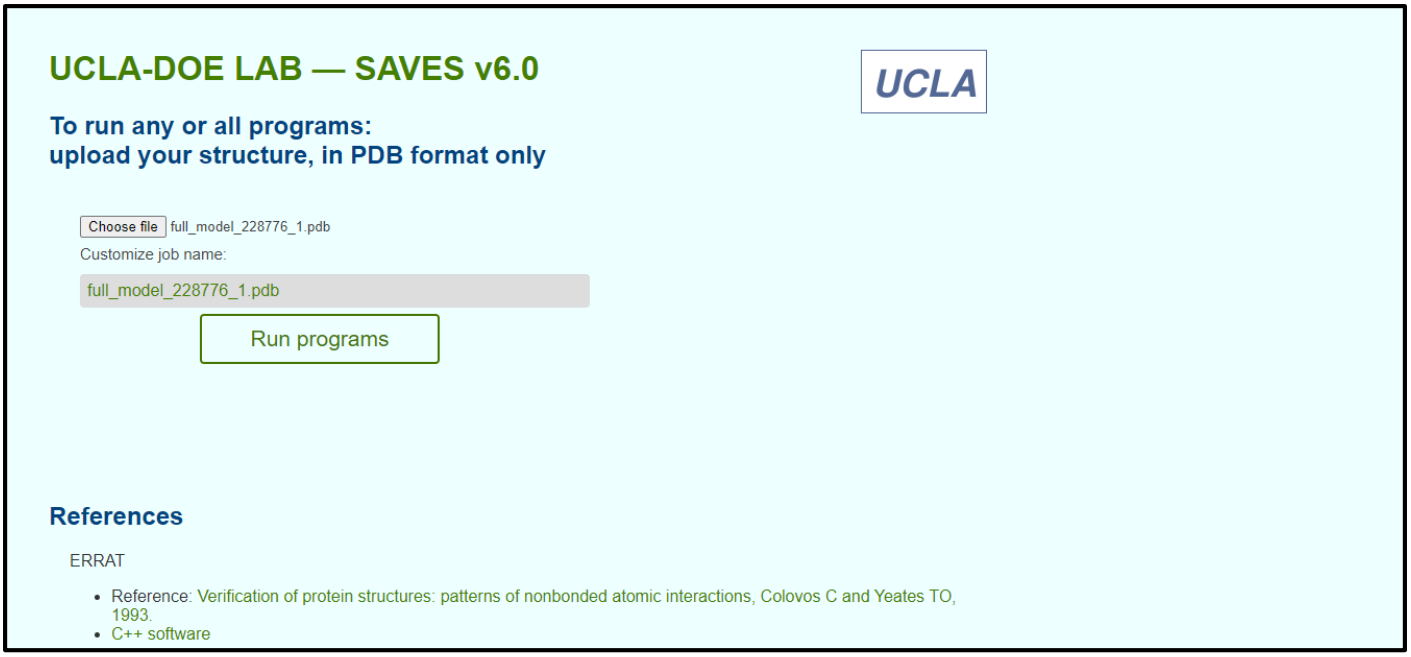


Fig2. Structure from Modeller for validation

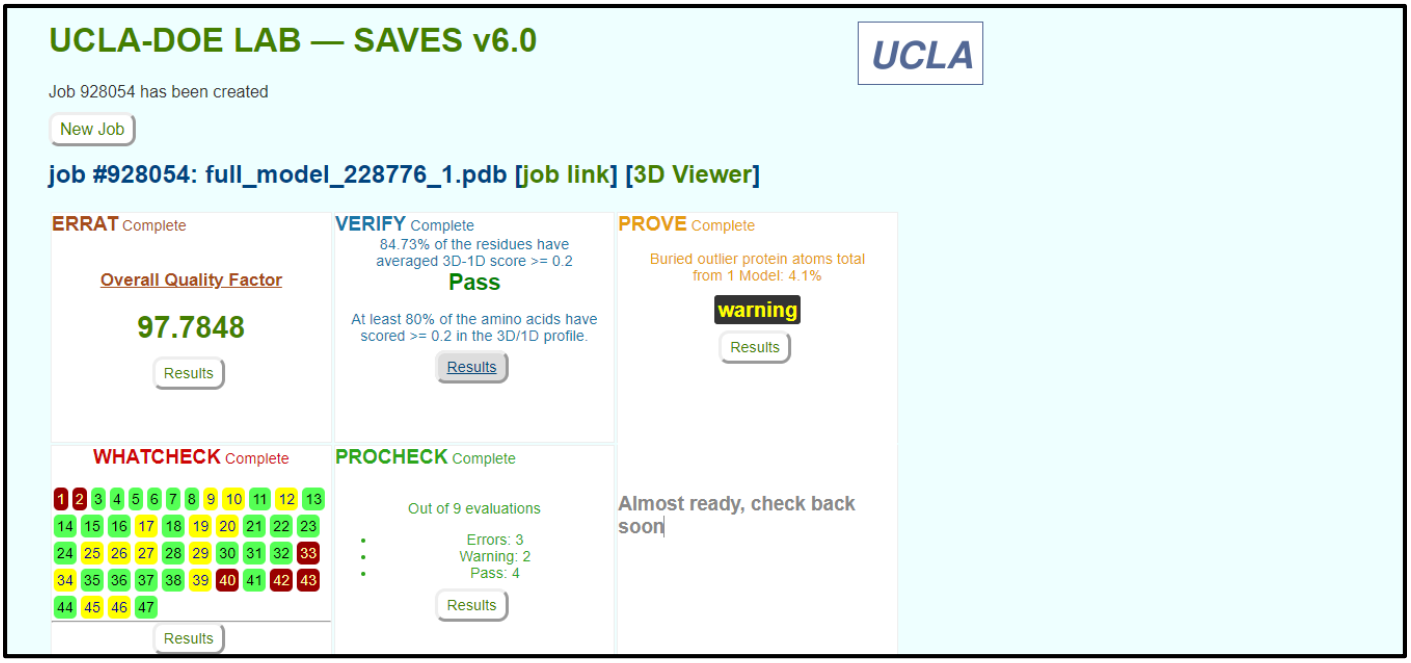


Fig3. Result page for structure validation for various servers

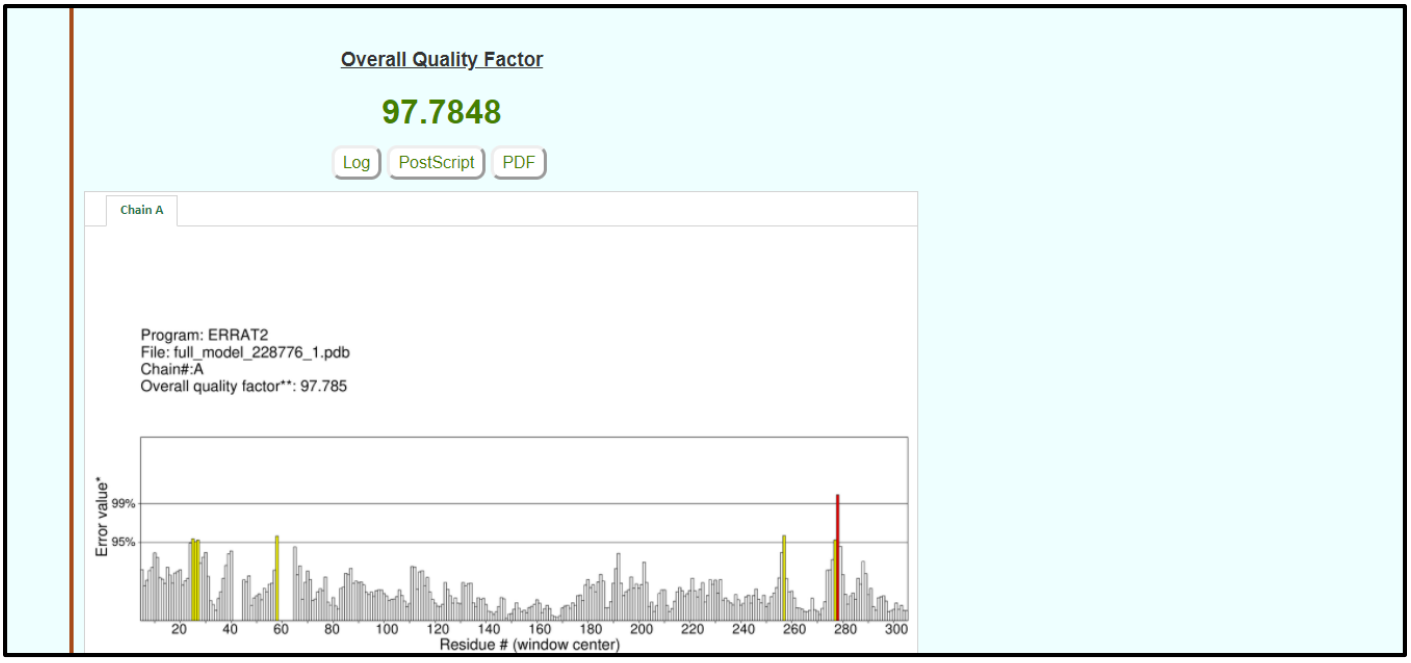


Fig4. Result page for Errat

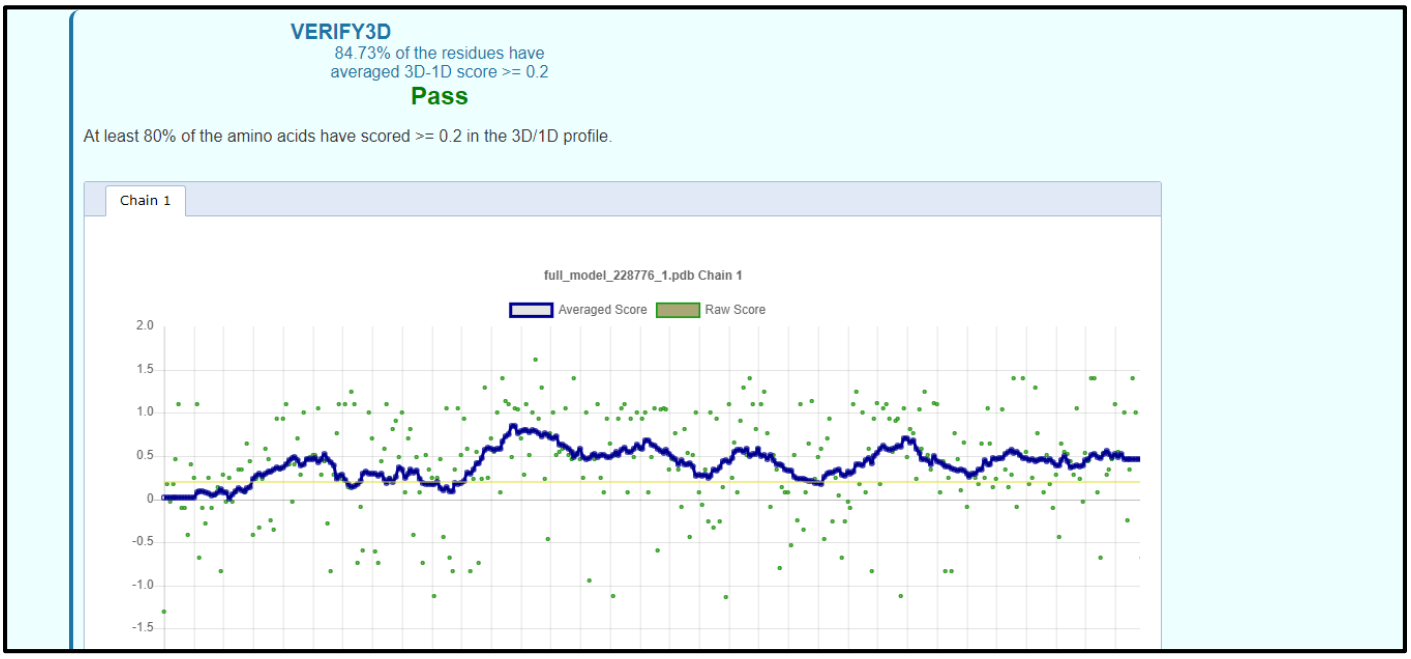


Fig5. Result page for Verify3D

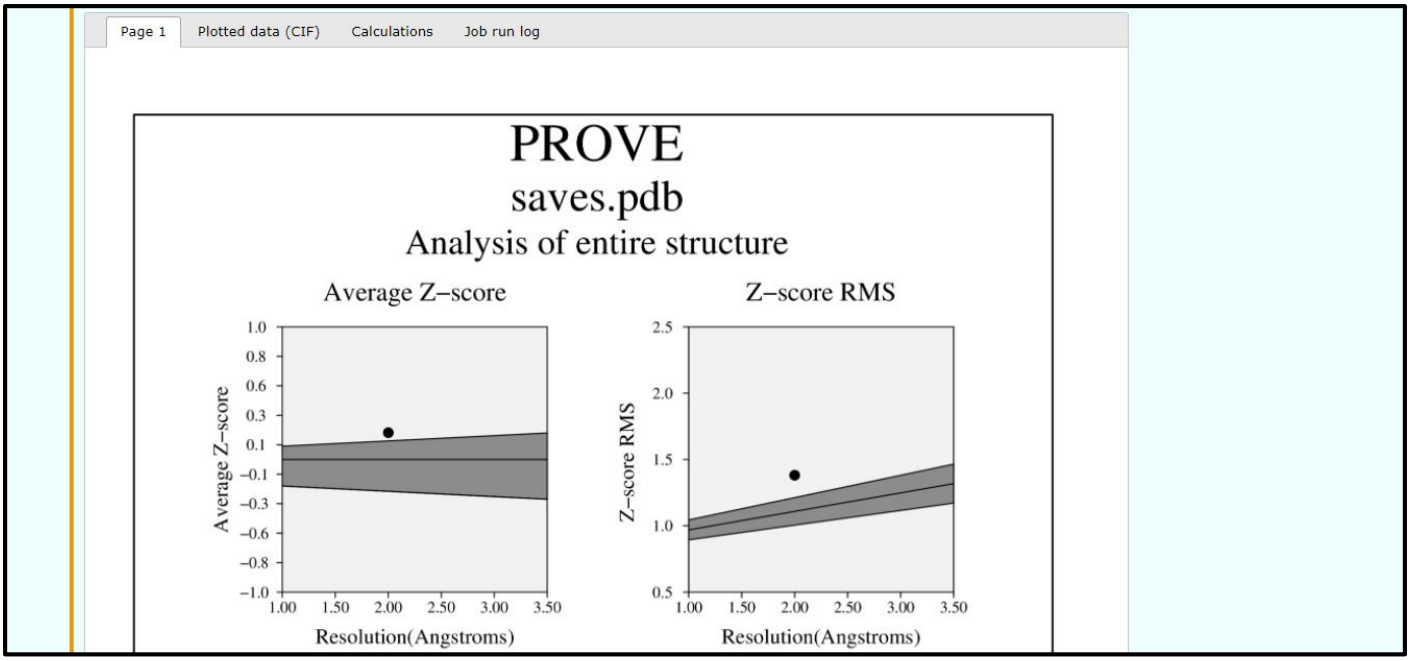


Fig6. Result page for Prove

Fig7. Result page for Whatcheck

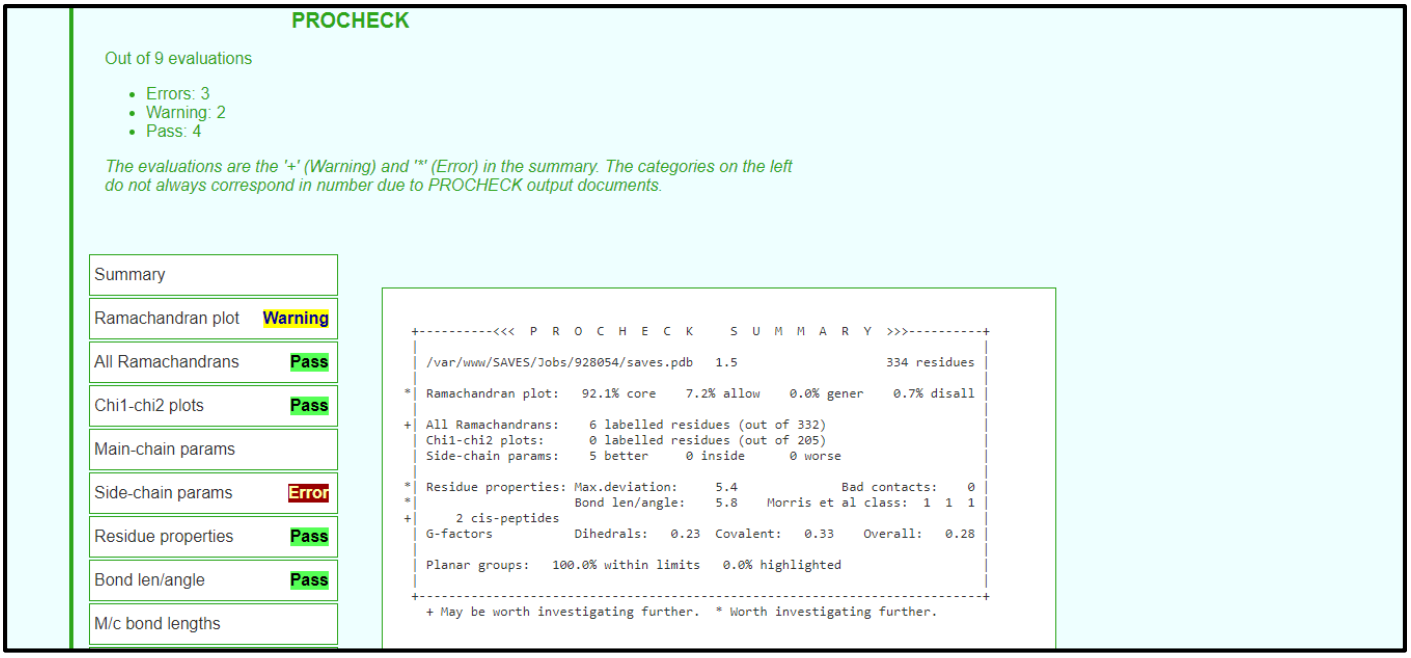


Fig8. Result page for Procheck

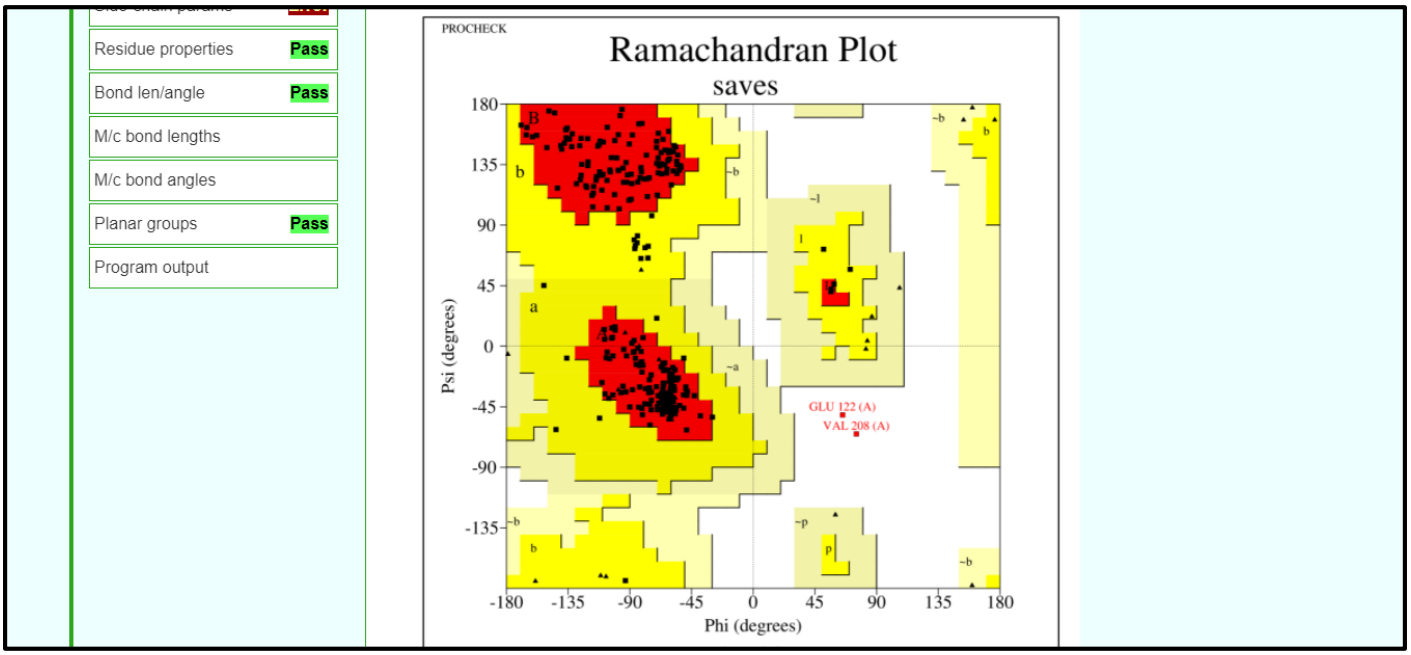


Fig8.1. Result page for Ramachandran Plot

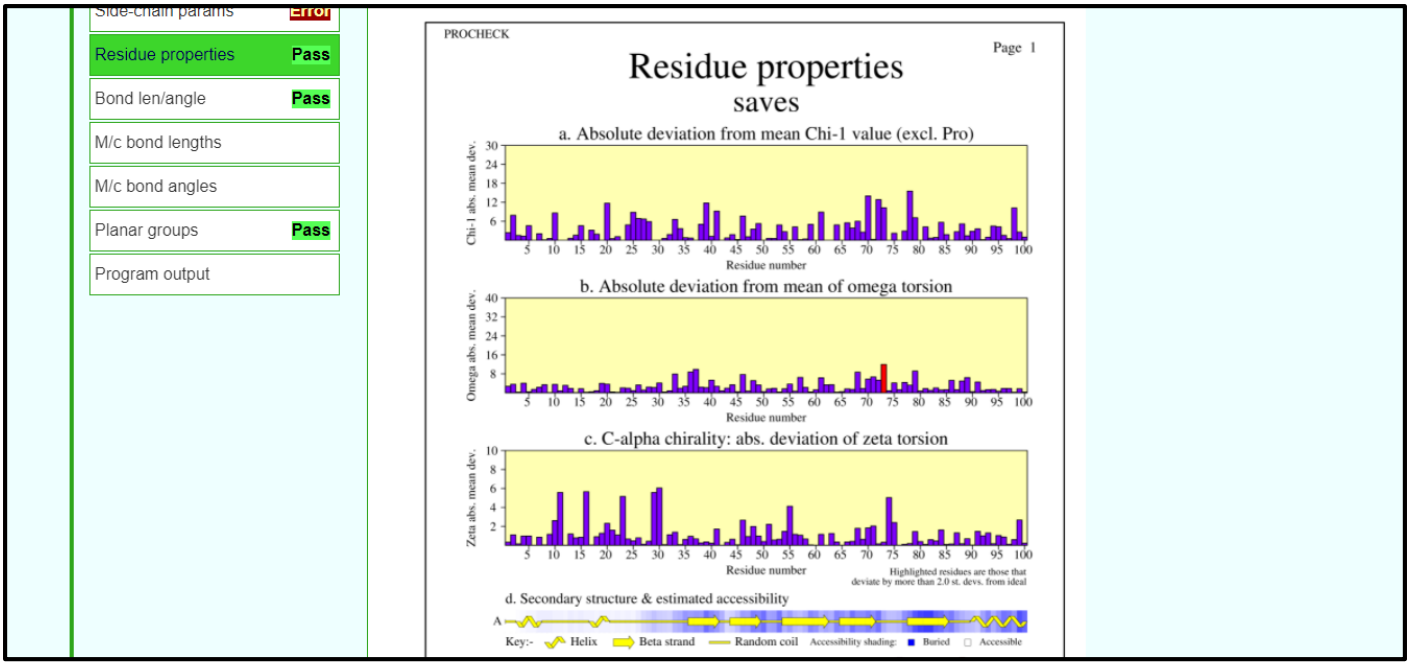


Fig8.2. Result page for residue properties

RESULT:

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

CONCLUSION:

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

REFERENCES:

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

Date: 28/02/22

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 Cα 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 α-helices and broad, flat arrows to represent β-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, α-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/) is a new version of RasMol for Windows with a more enhanced user interface.

Installation:

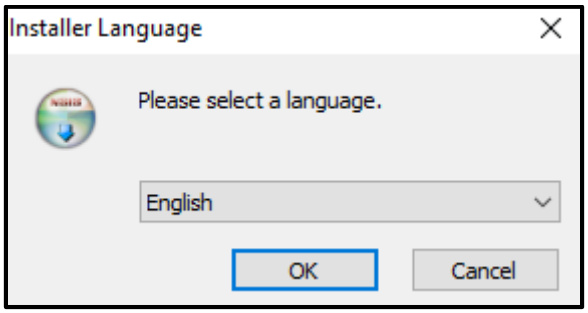


Fig1. Select language and proceed

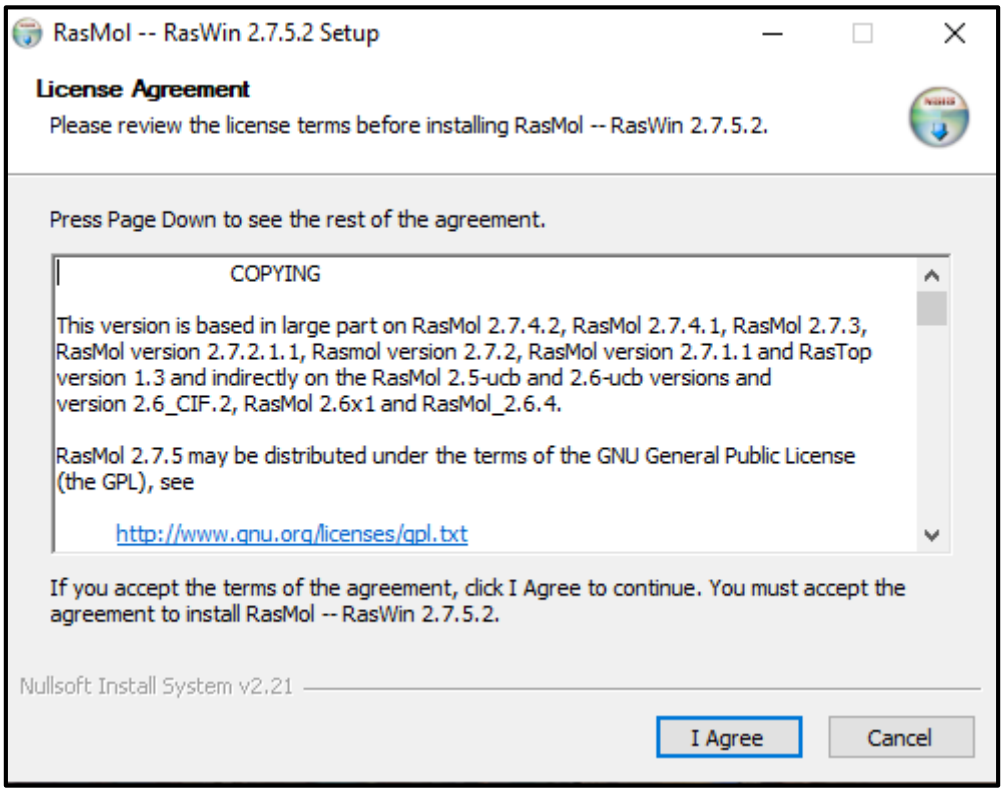


Fig2. Read the License Agreement and click "I Agree" to proceed

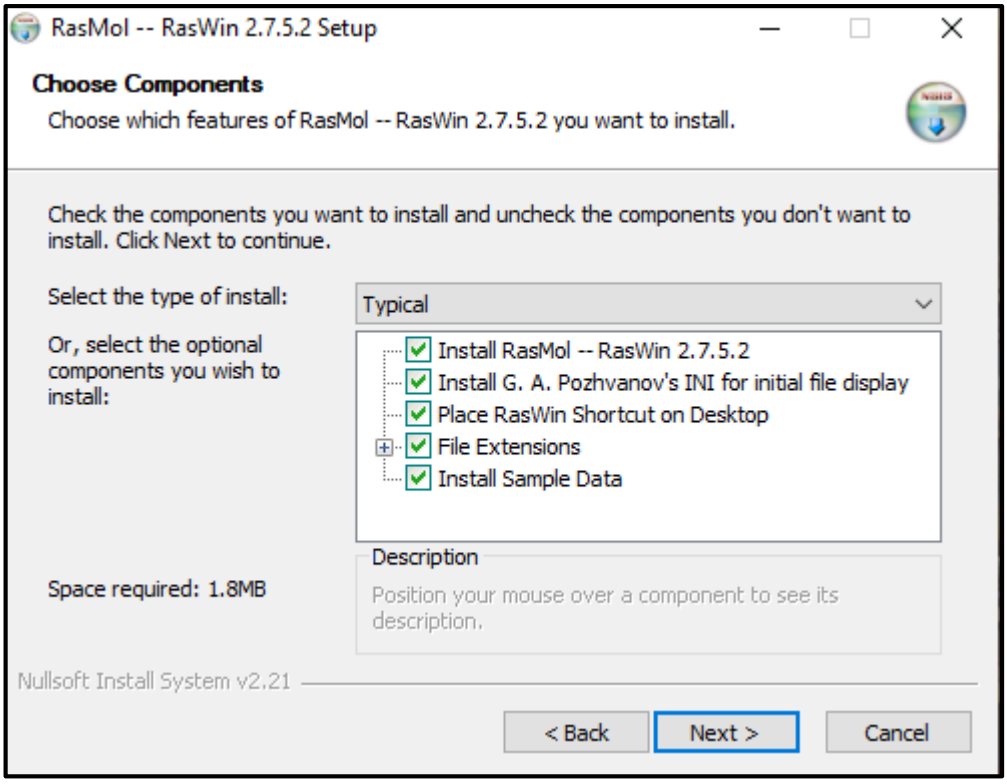


Fig3. Select the components of RasMol you want to install and click "Next" to proceed

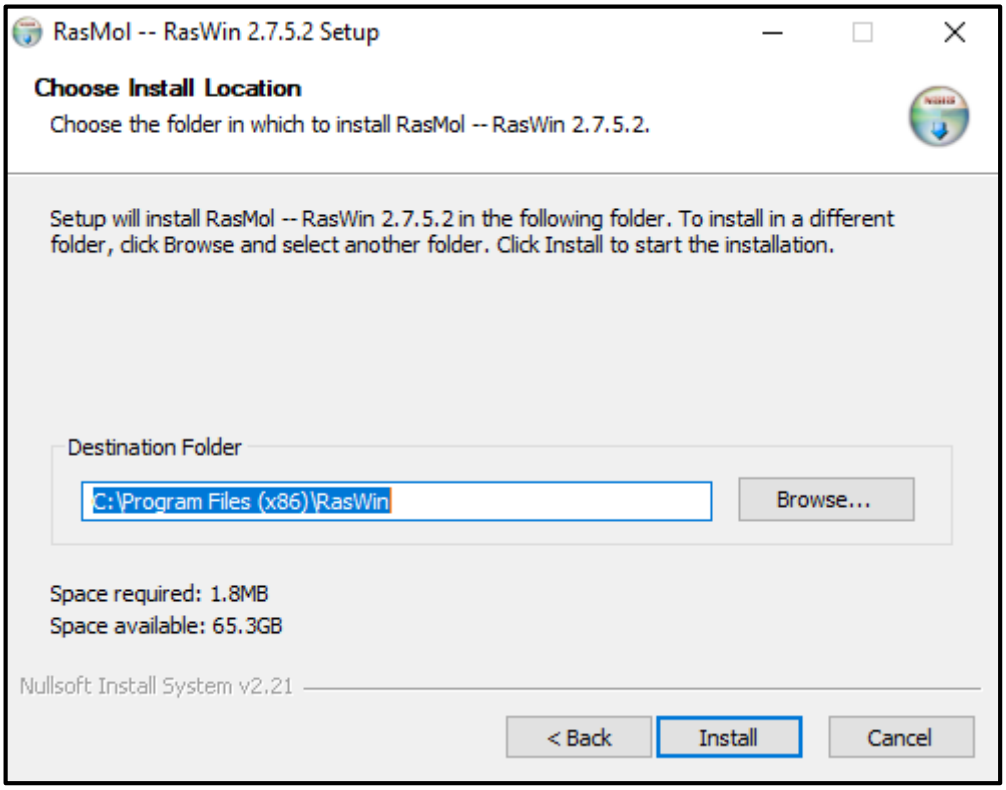


Fig4. Select the directory in which you want to install RasMol and click "Install" to proceed

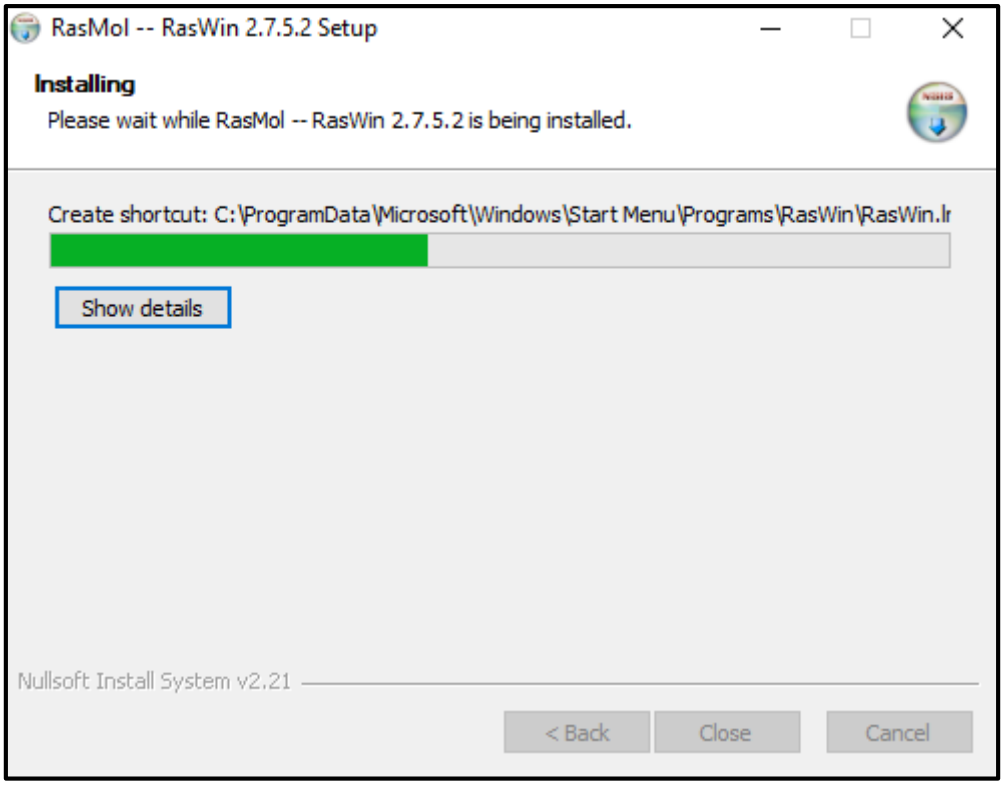


Fig5. The software will now install

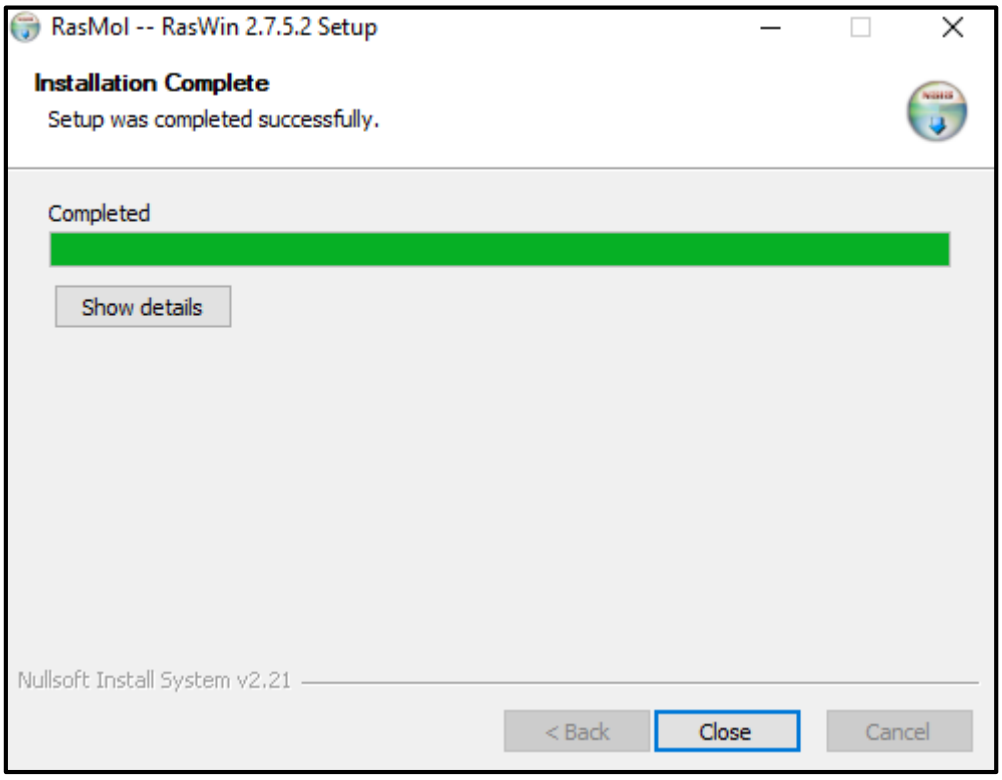


Fig6. Close the installer after the install is completed

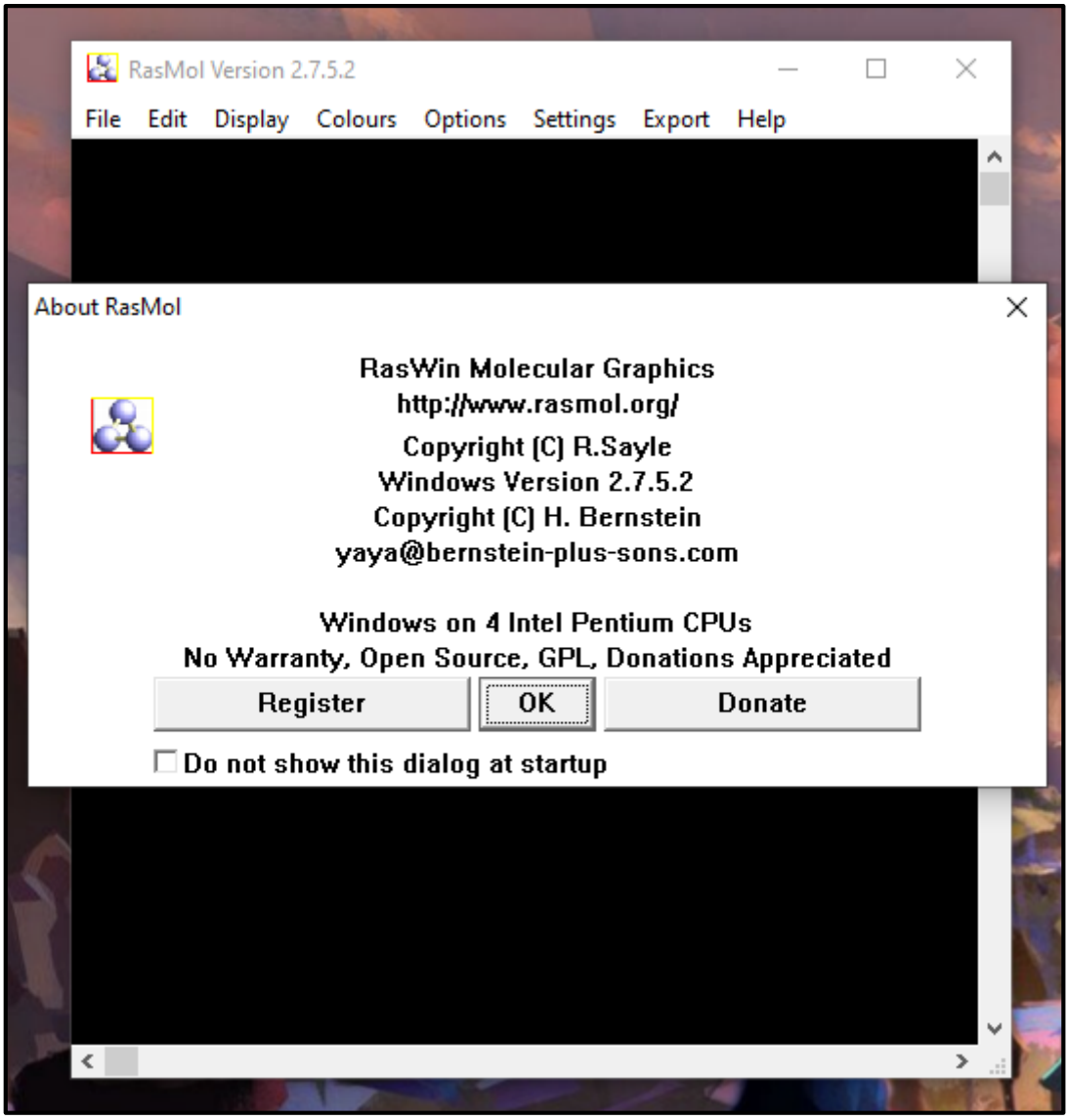


Fig7. RasMol is now installed

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:

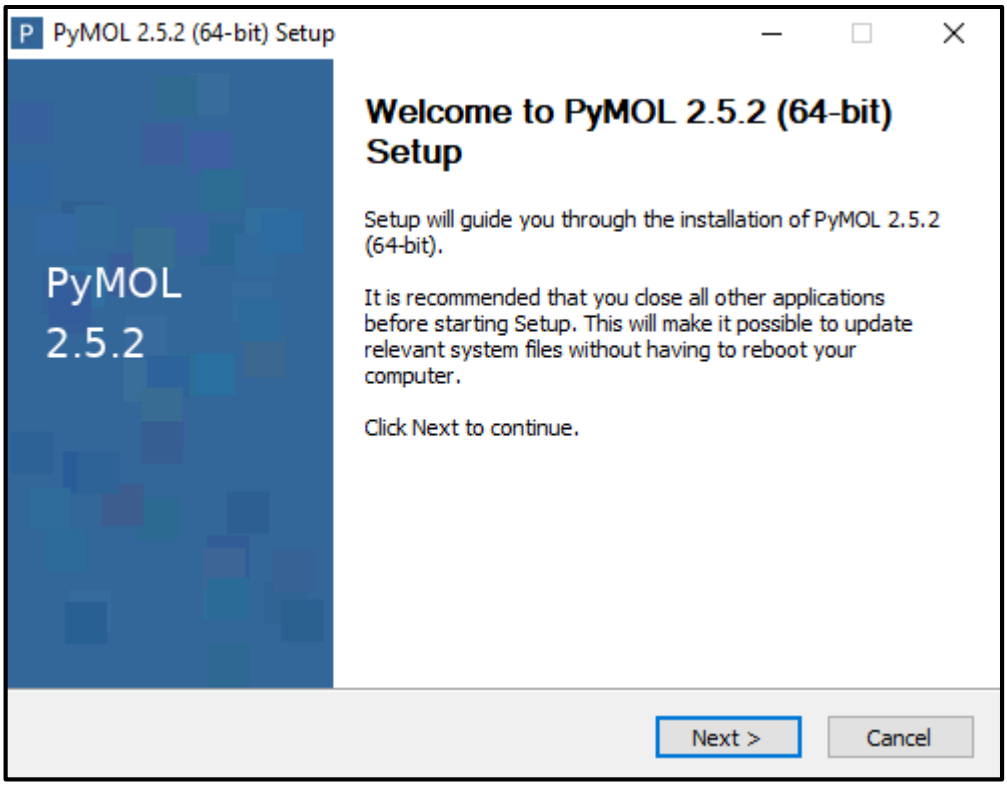


Fig1. Download and run the installer. Click next to proceed

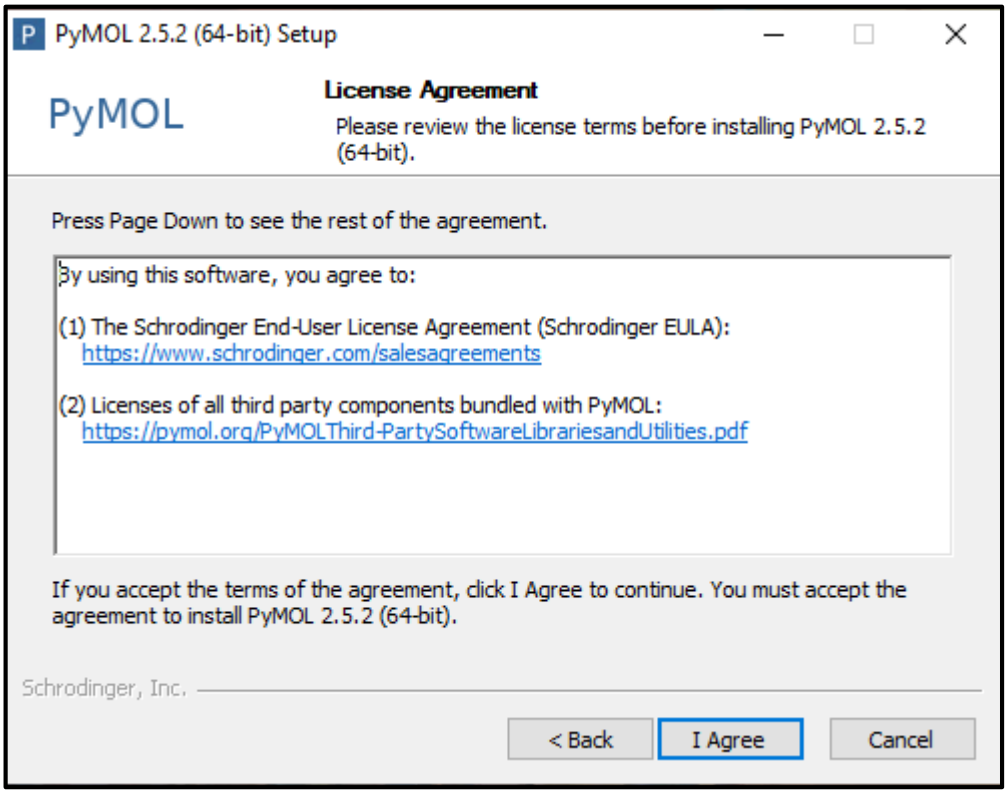


Fig2. Read the License Agreement and click "I Agree" to proceed

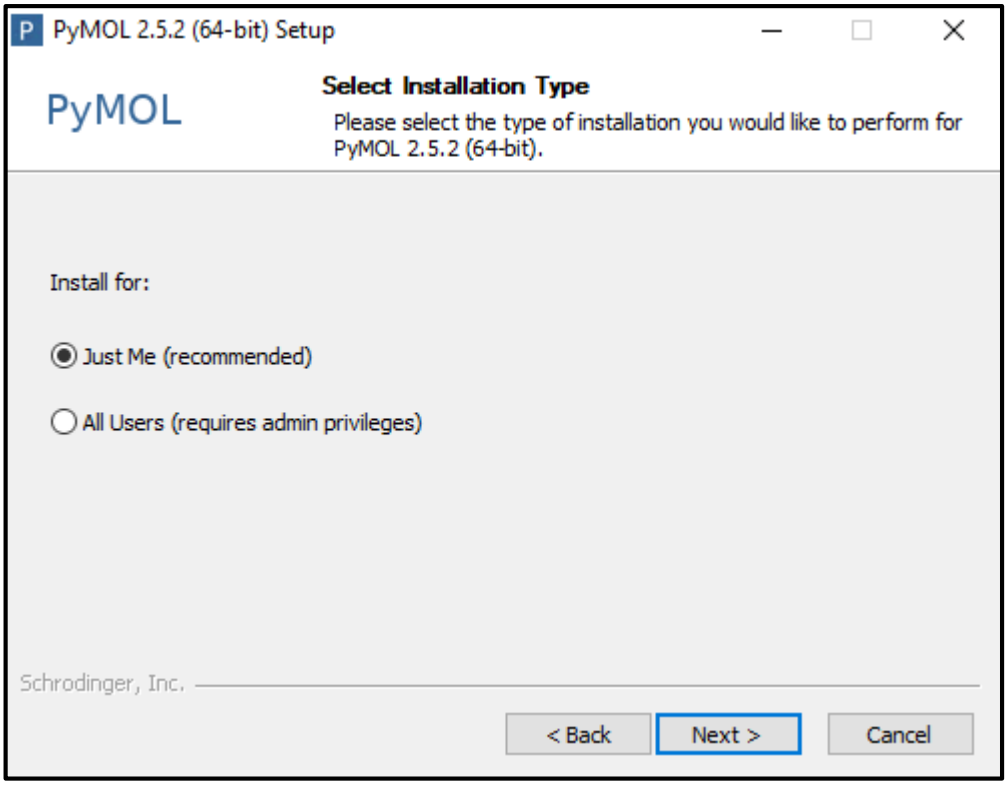


Fig3. Select the Installation type and click "Next" to proceed

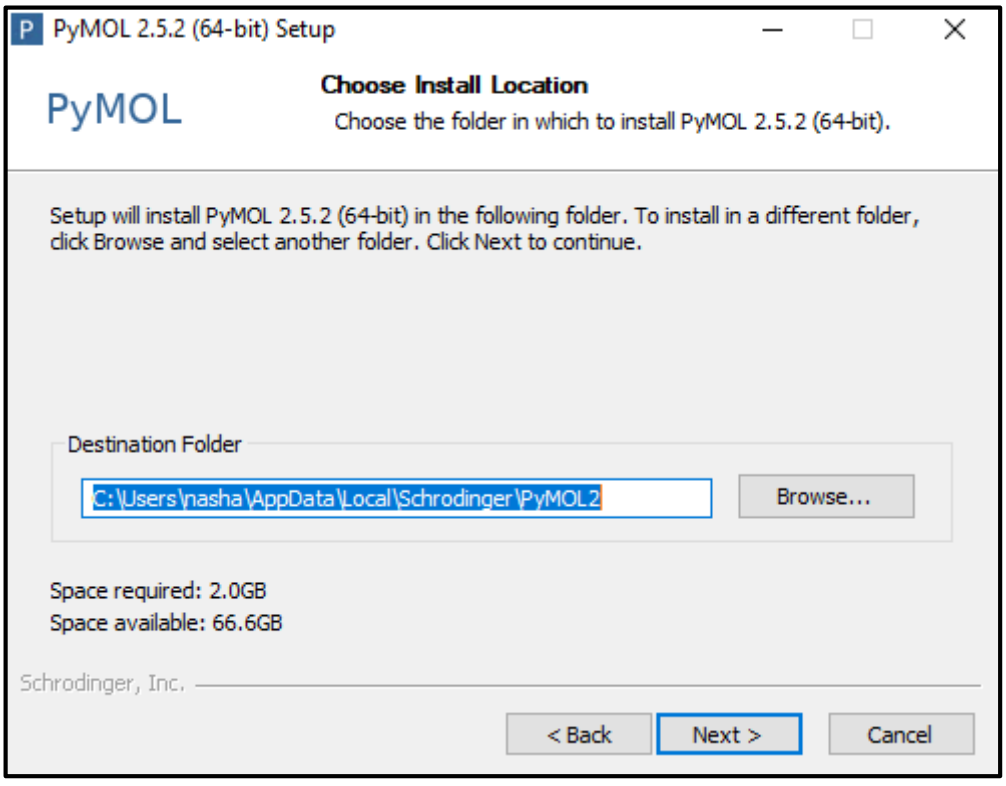


Fig4. Select the Directory in which you want to install PyMol

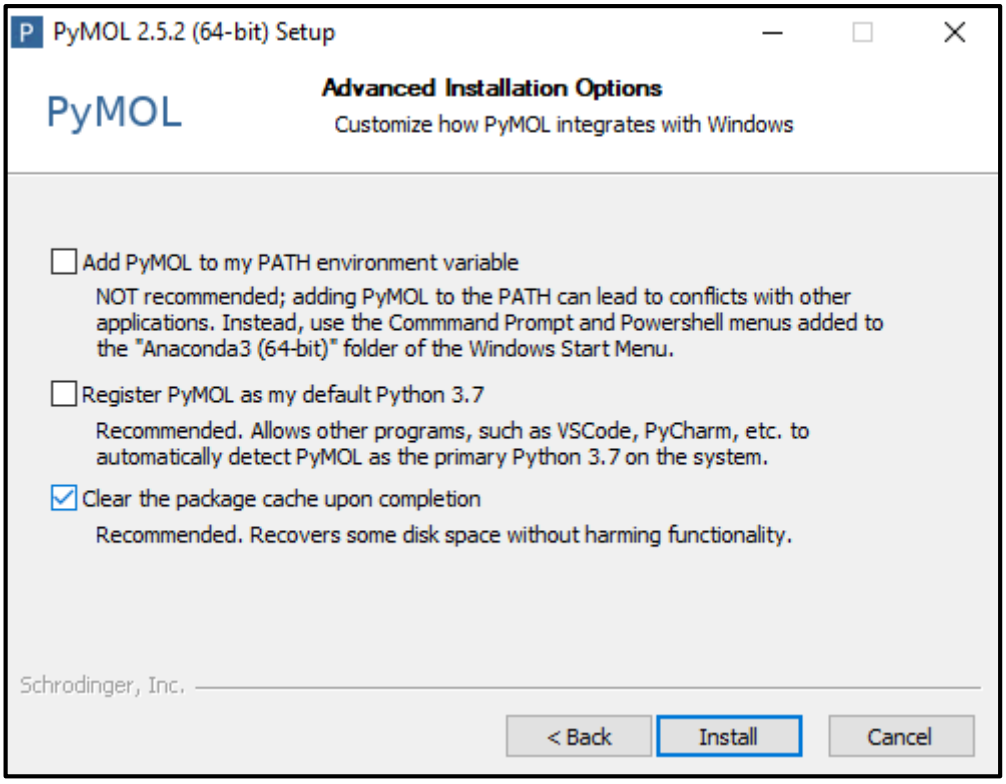


Fig5. Select the additional options if you want to and click "Install" to start the installation of PyMol

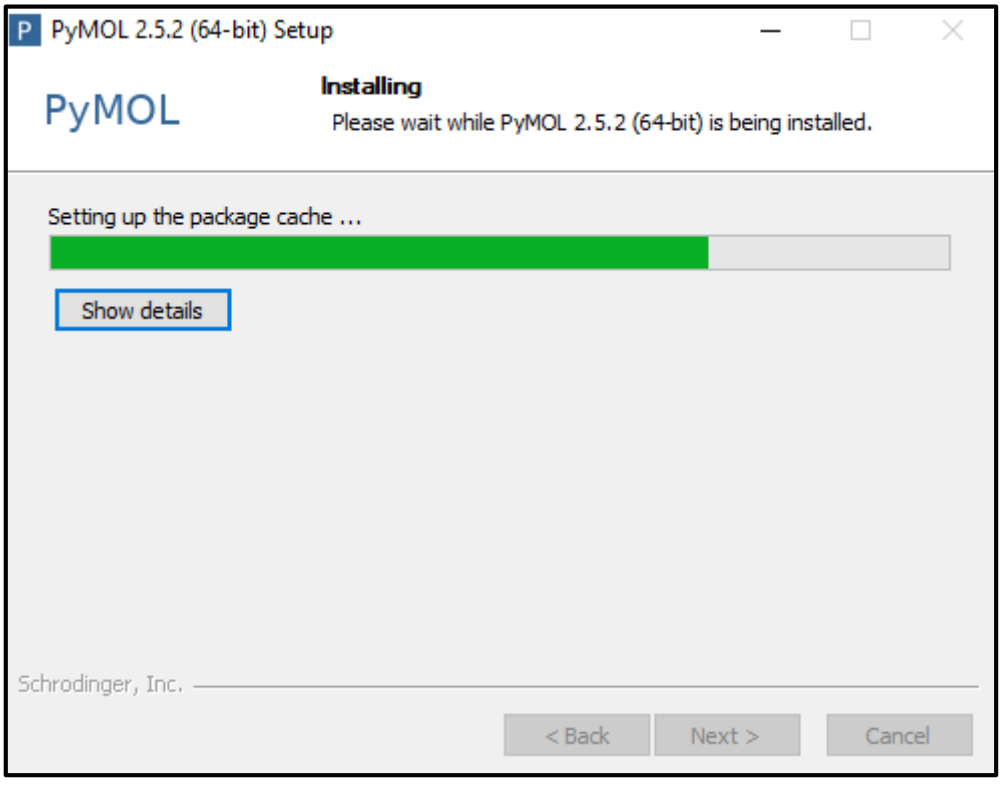


Fig6. Let the installer install PyMol

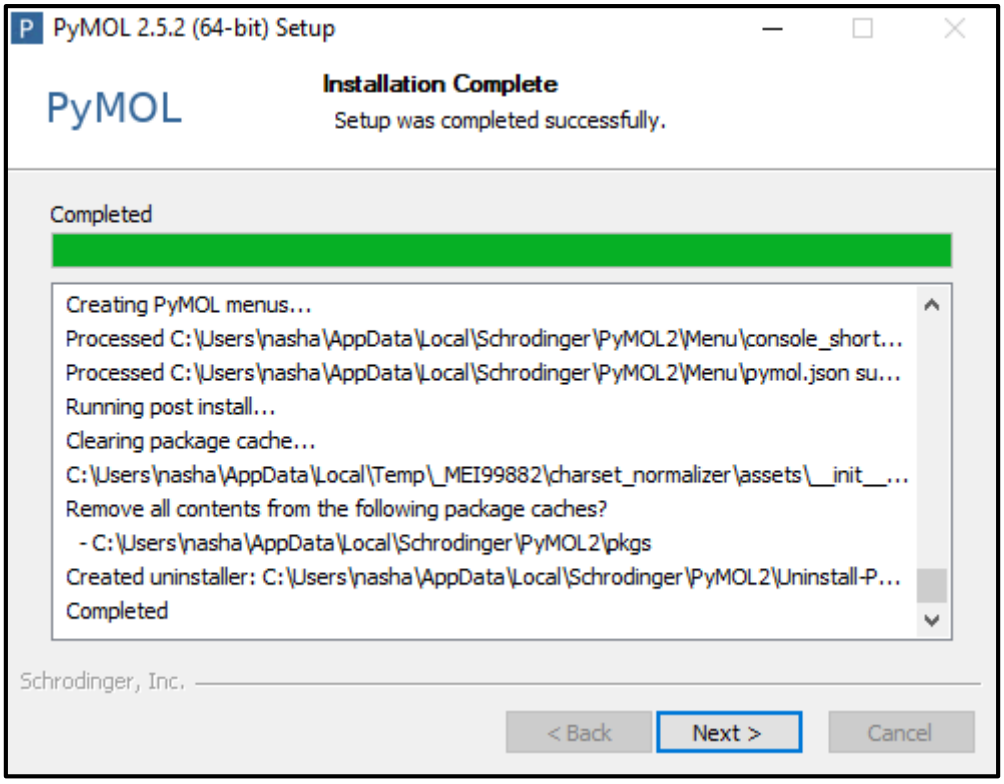


Fig7. After completion click on "Next" to finalize installation

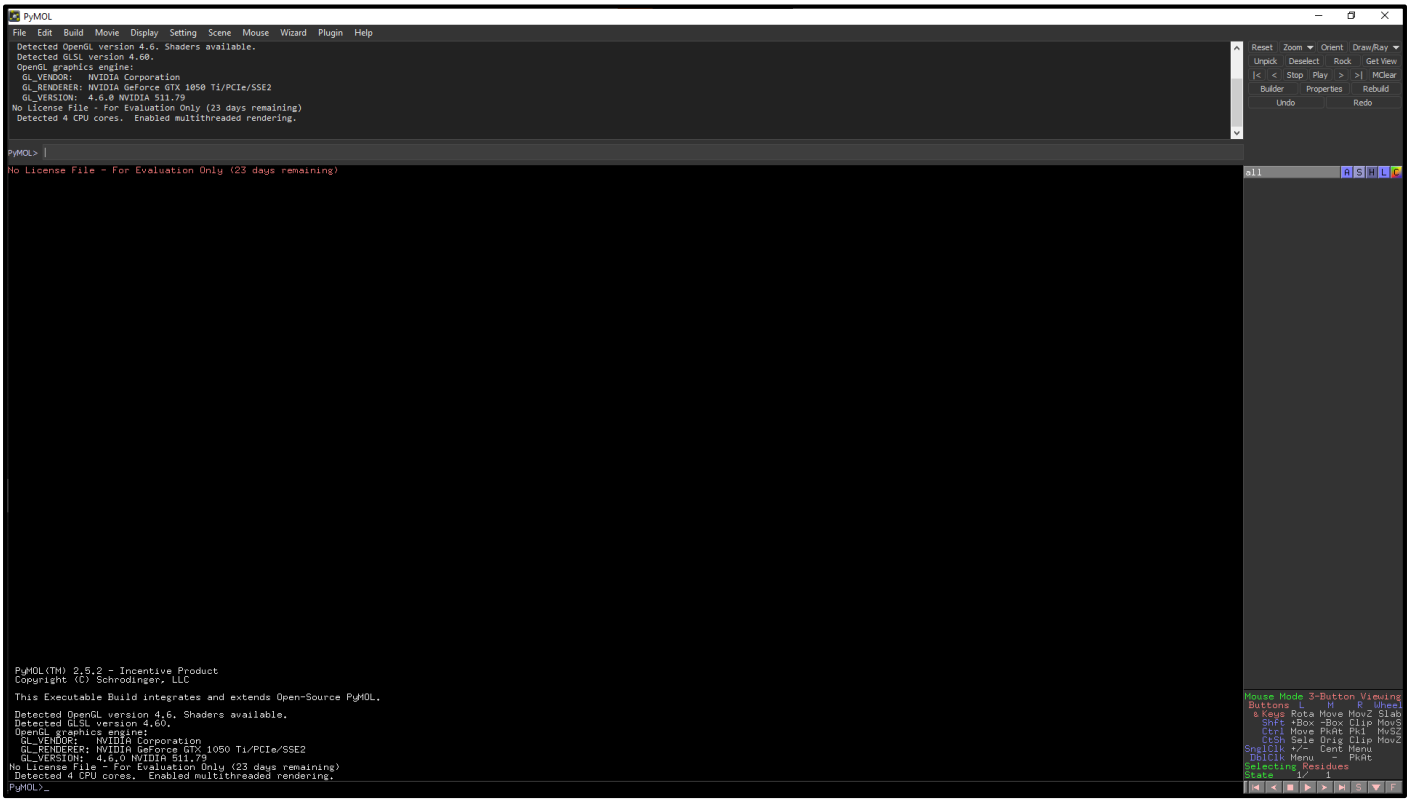


Fig8. Pymol is now installed

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. Retrieved March 4, 2022, from http://www.openrasmol.org/
4. *PyMOL | pymol.org*. (2019). Pymol.org. Retrieved March 4, 2022, from https://pymol.org/2/

DATE: 28/02/22

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, α-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/) 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:

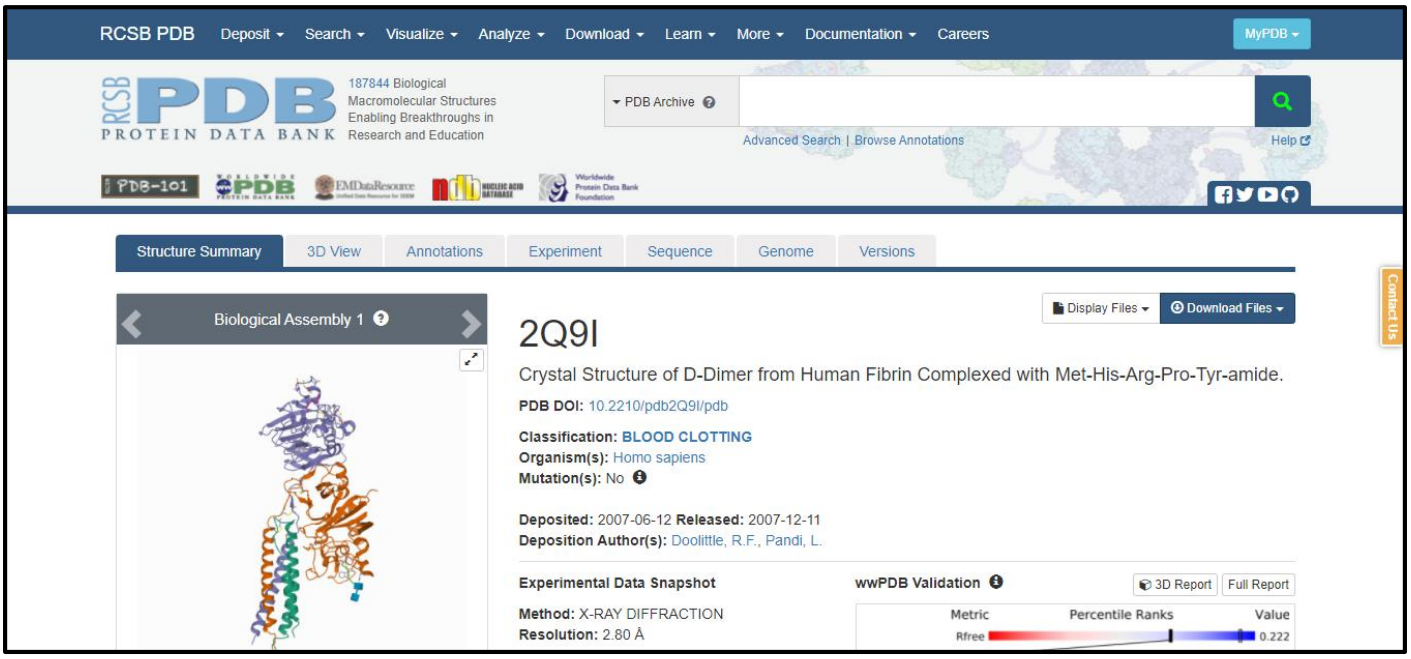


Fig1. Fibrin structure retrieved from PDB

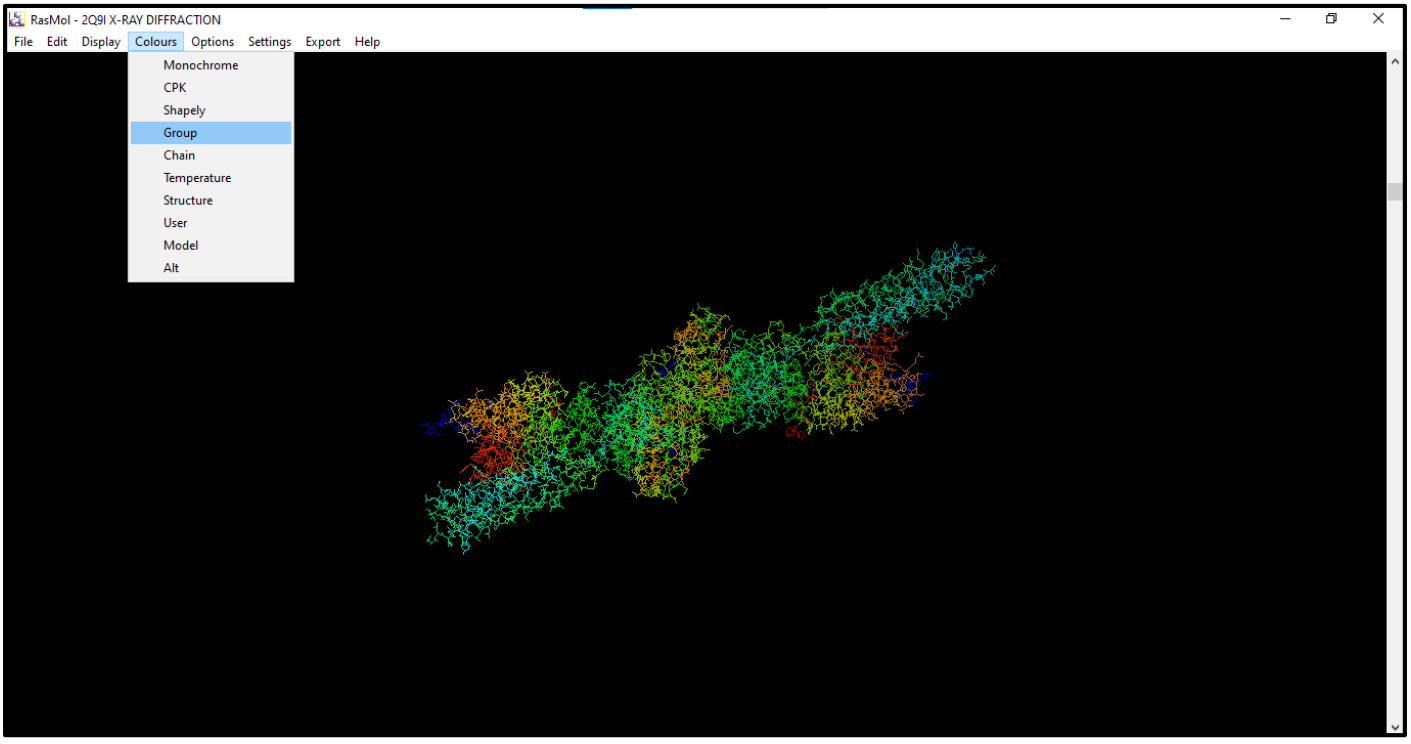
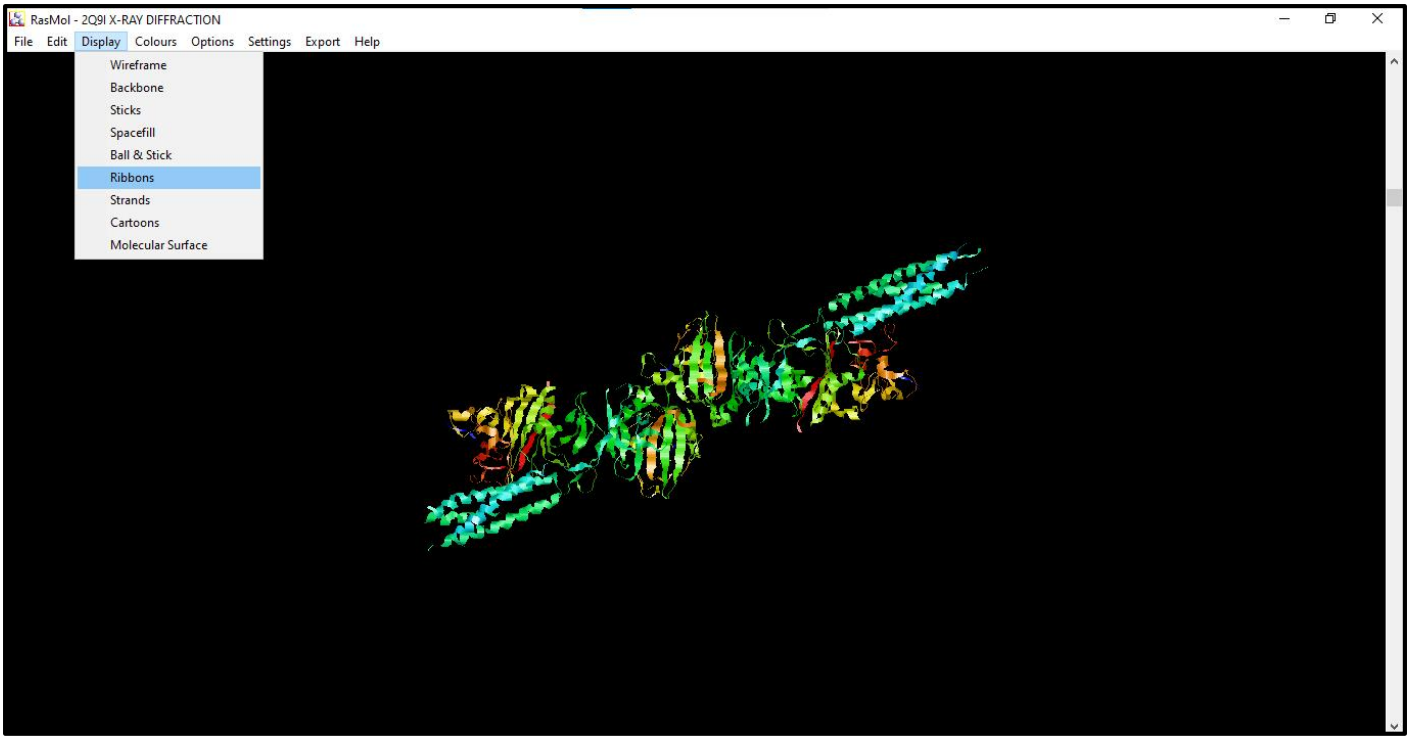


Fig2. Structure of Fibrin in group colour format



Fiig3. Fibrin structure in ribbons format

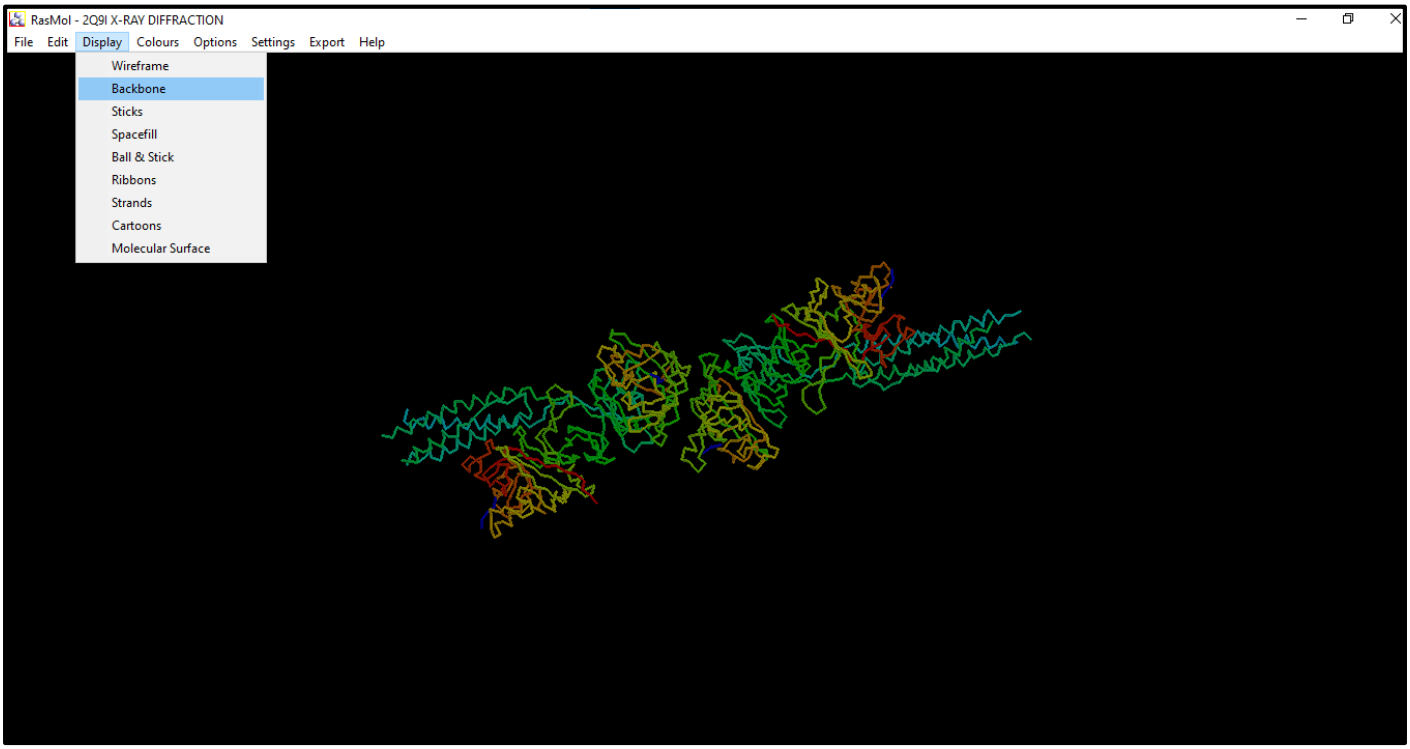


Fig4. Fibrin structure in backbone format

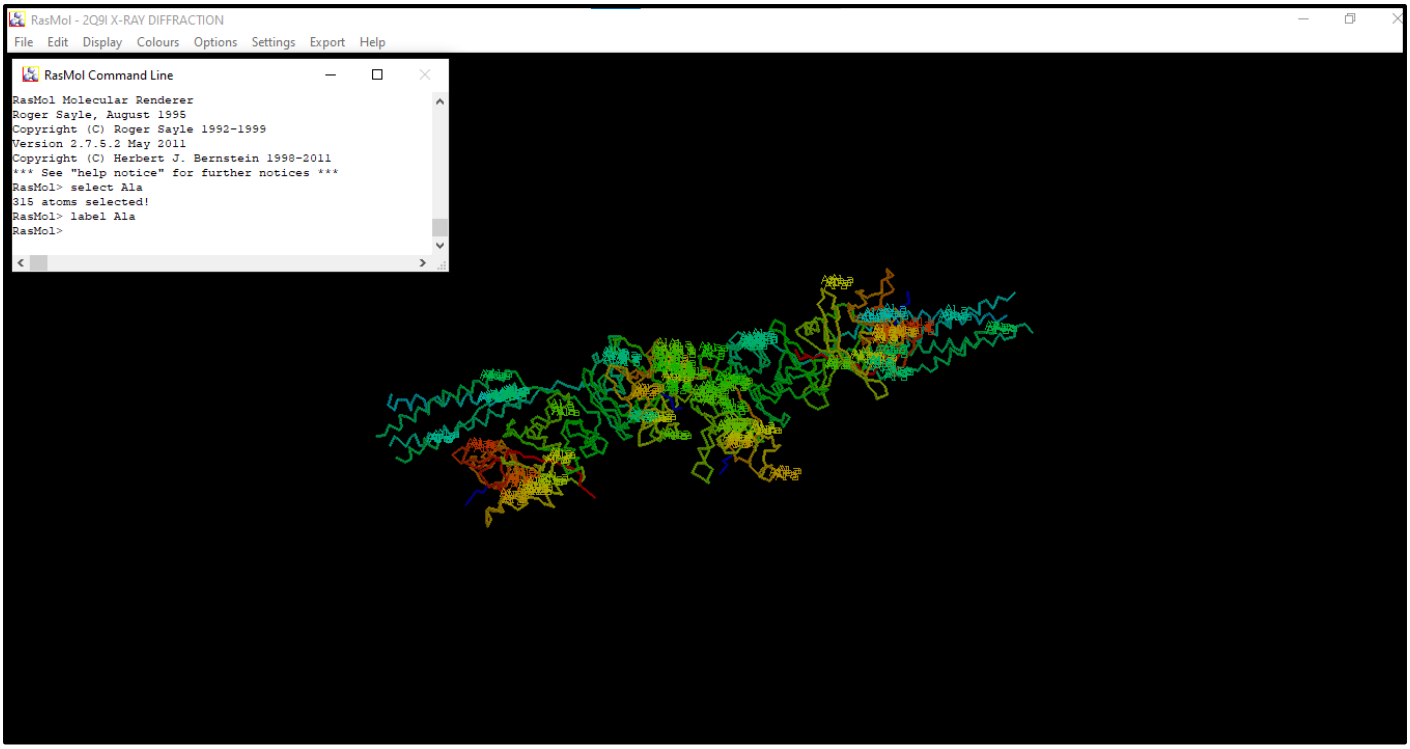


Fig5. Visualisation of Alanine residues in Fibrin

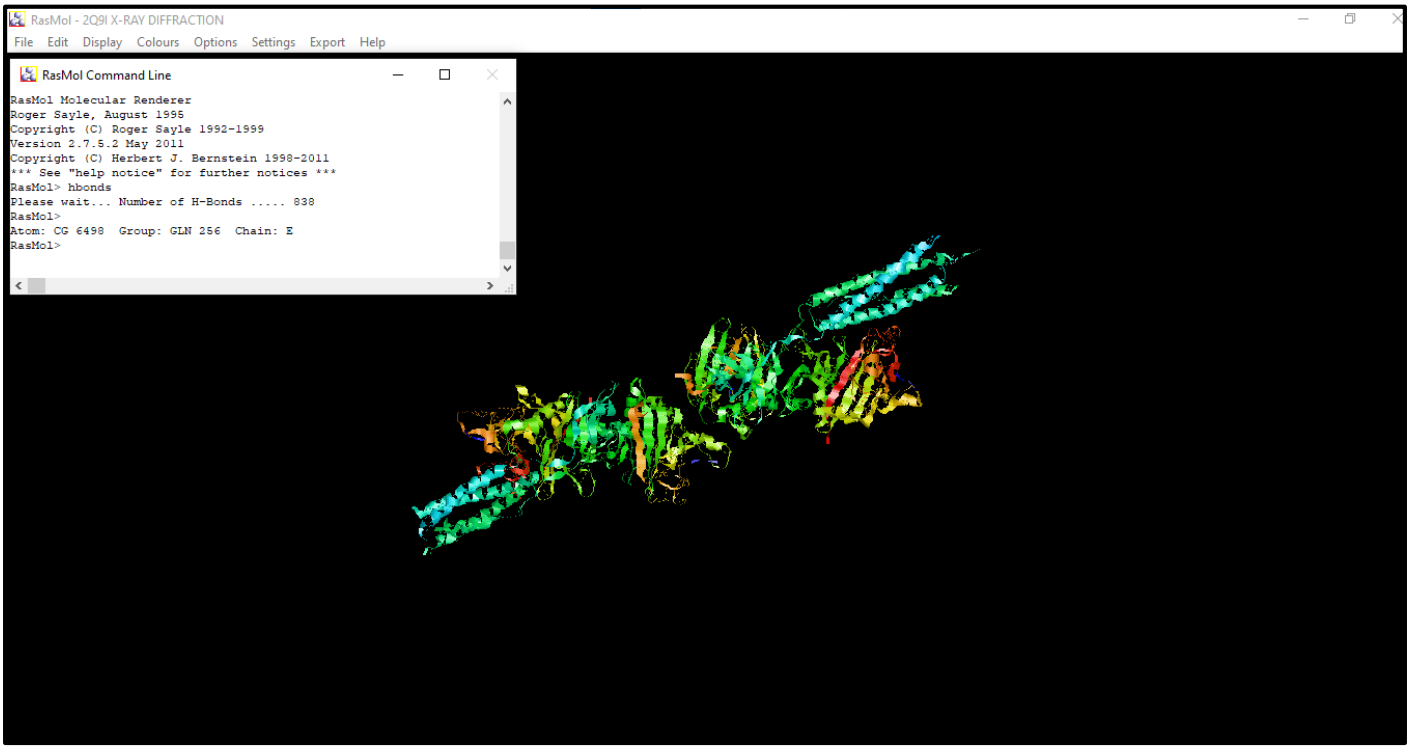


Fig6. Visualisation of H-bonds in Fibrin

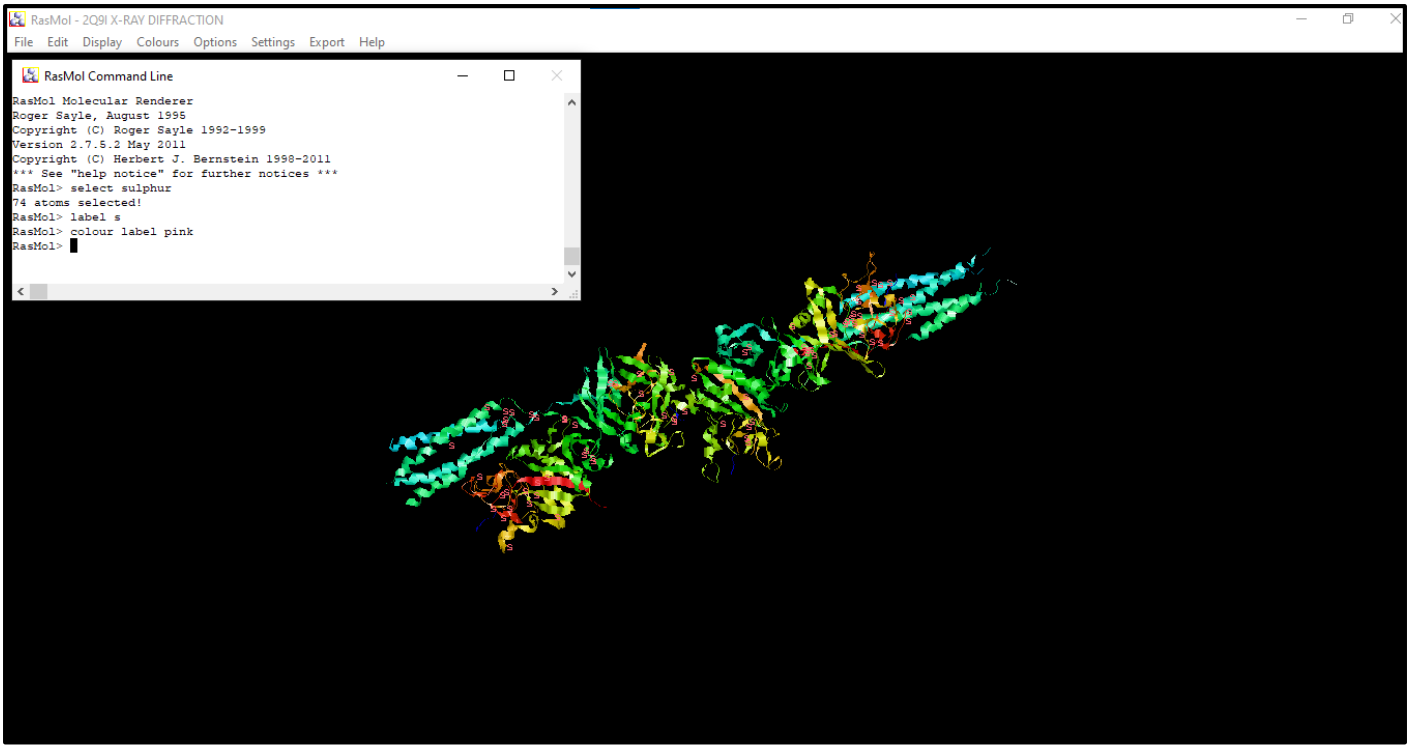


Fig7. Visualisation of sulphur residues in Fibrin

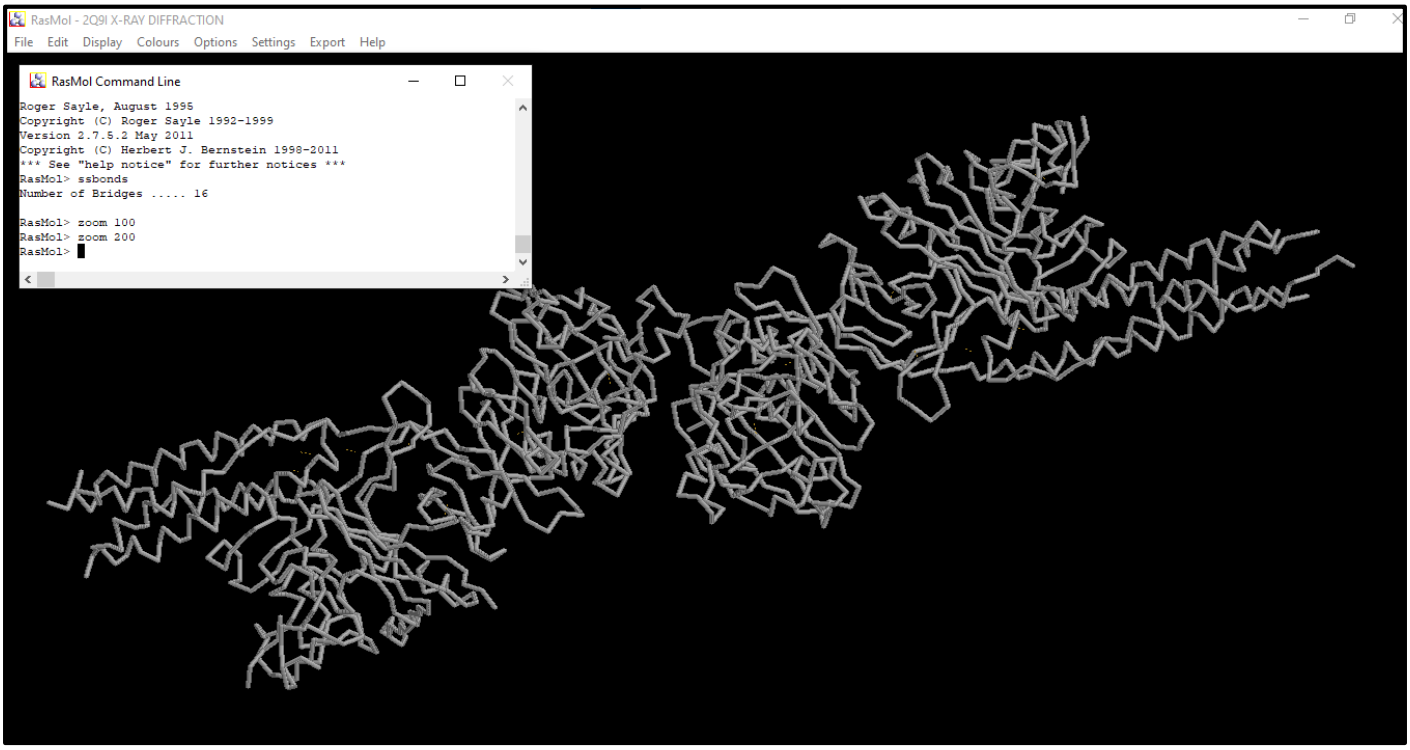


Fig8. Visualisation of disulphide bonds in Fibrin

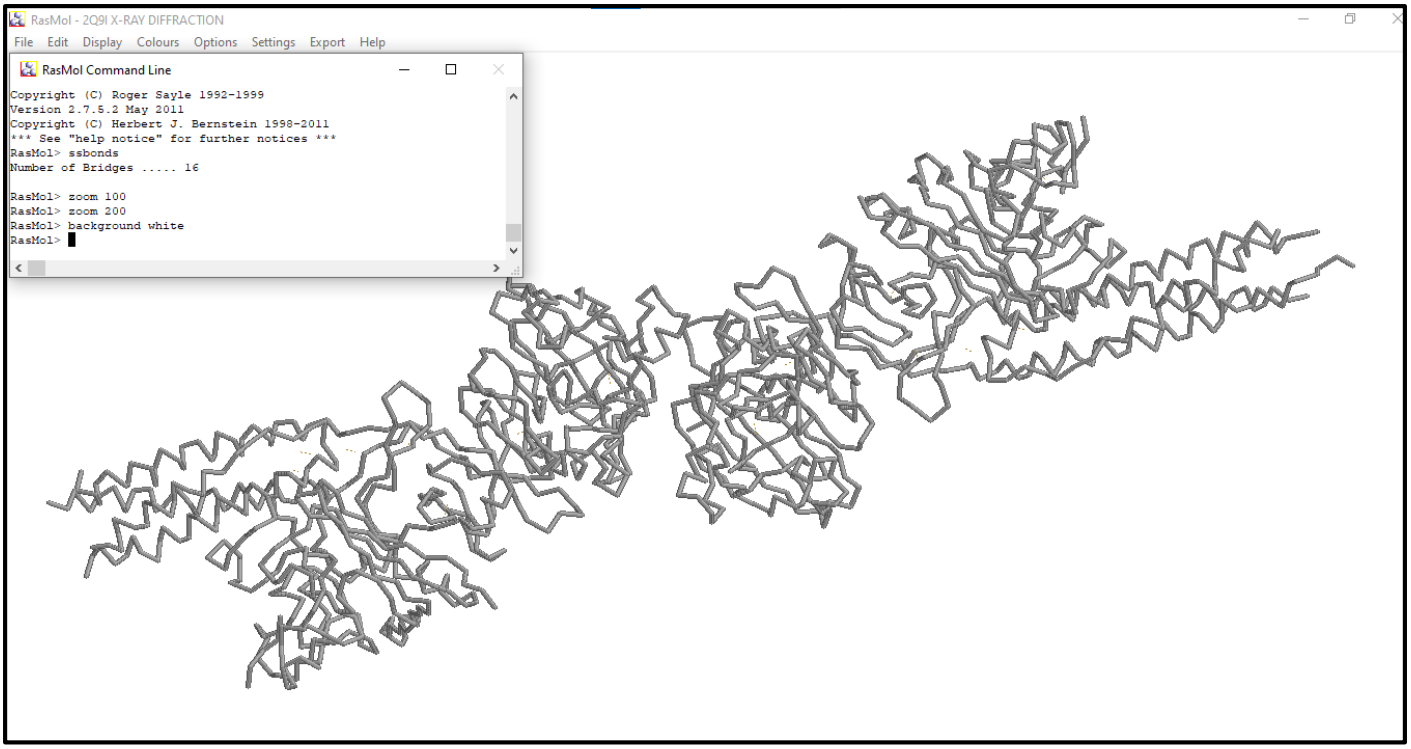


Fig9. Visualisation of structure with white background

PyMOL:

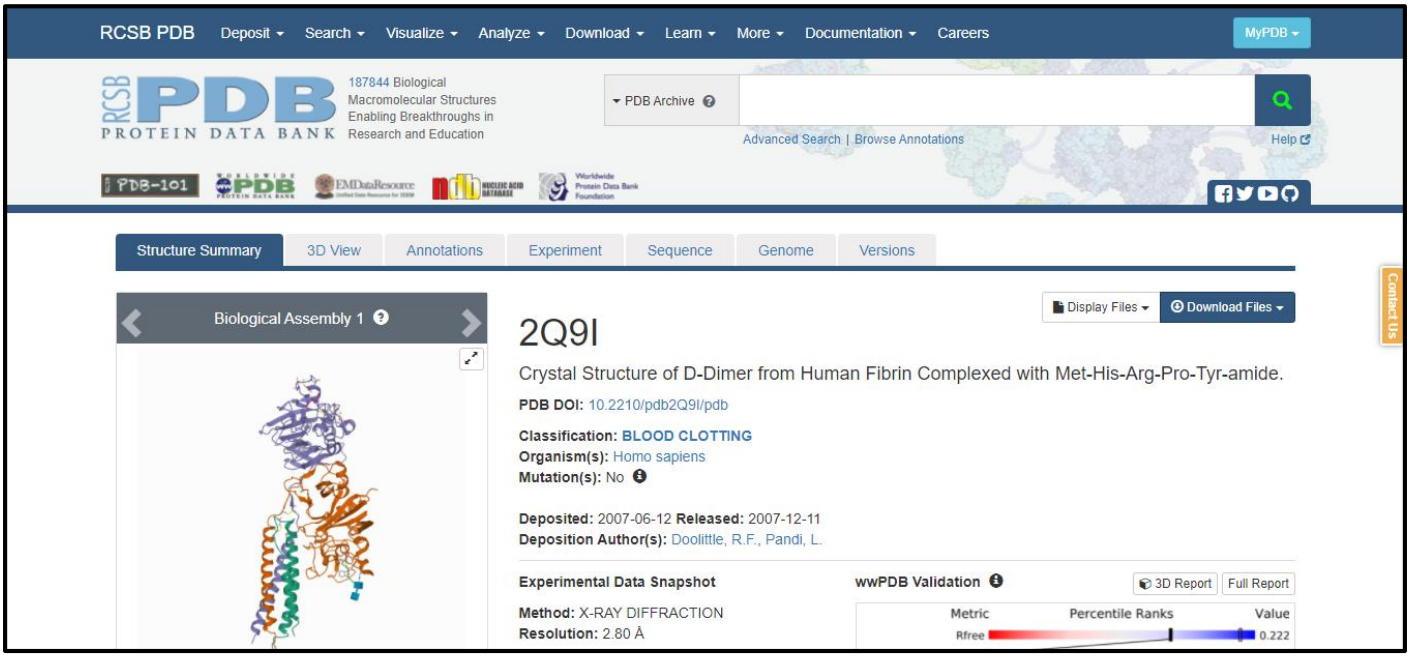


Fig1. Fibrin structure retrieved from PDB

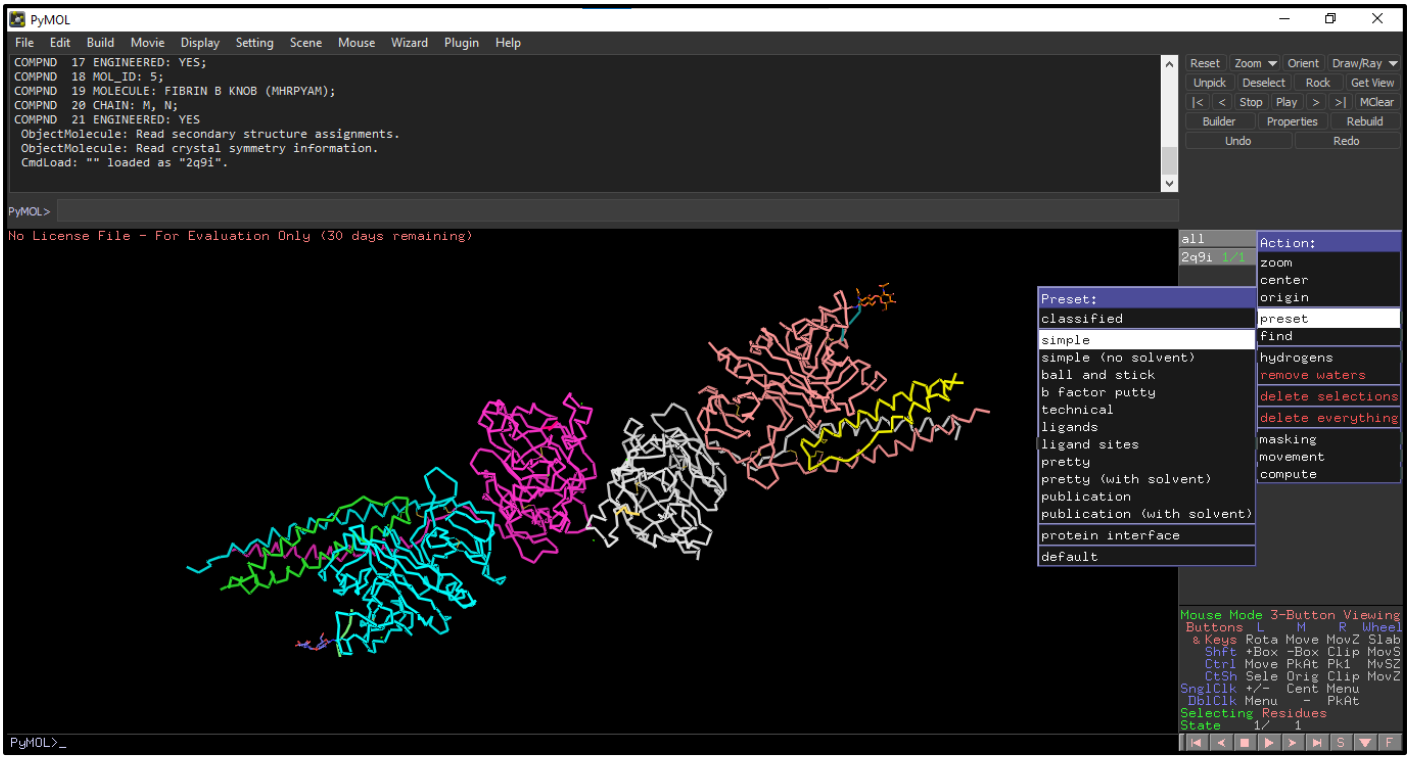


Fig2. Visualisation of Fibrin in simple format

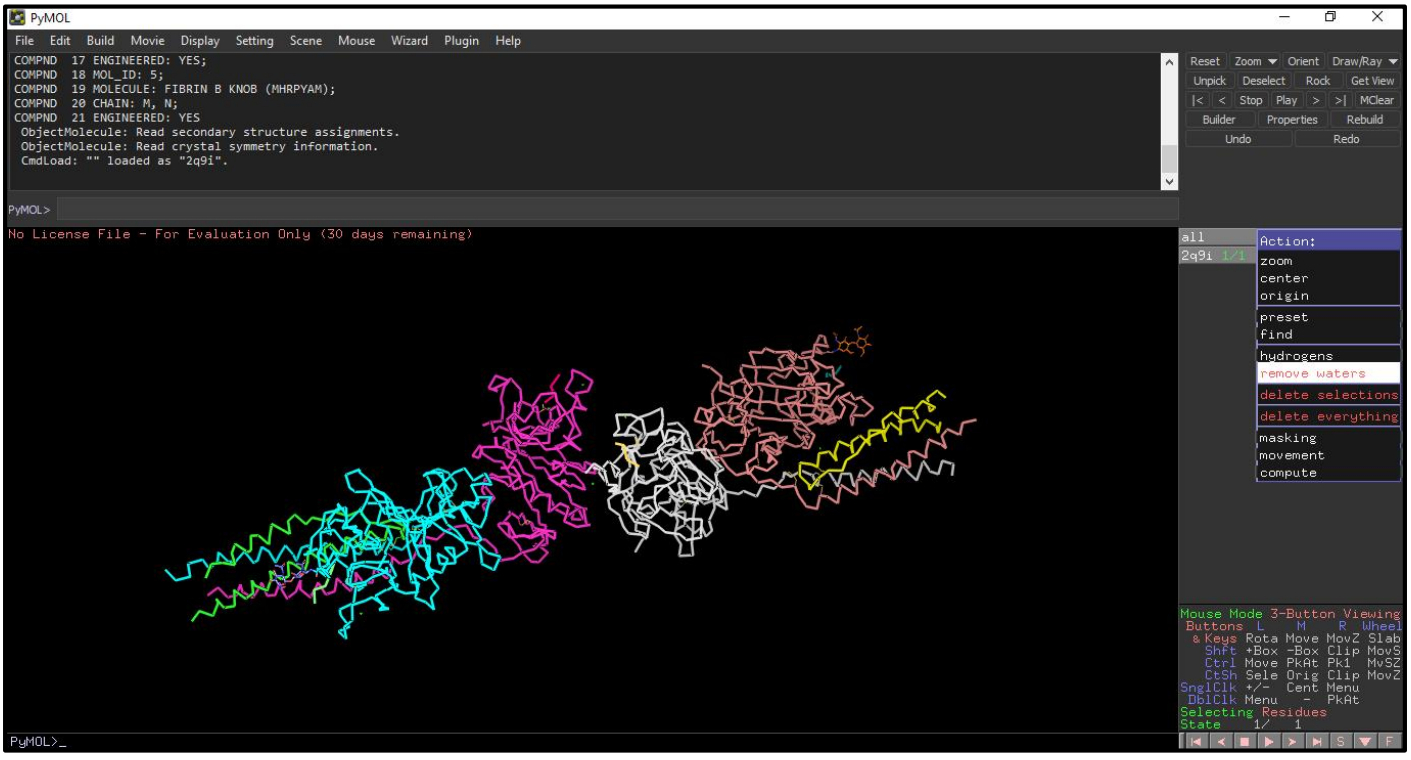


Fig3. Removal of water from Fibrin

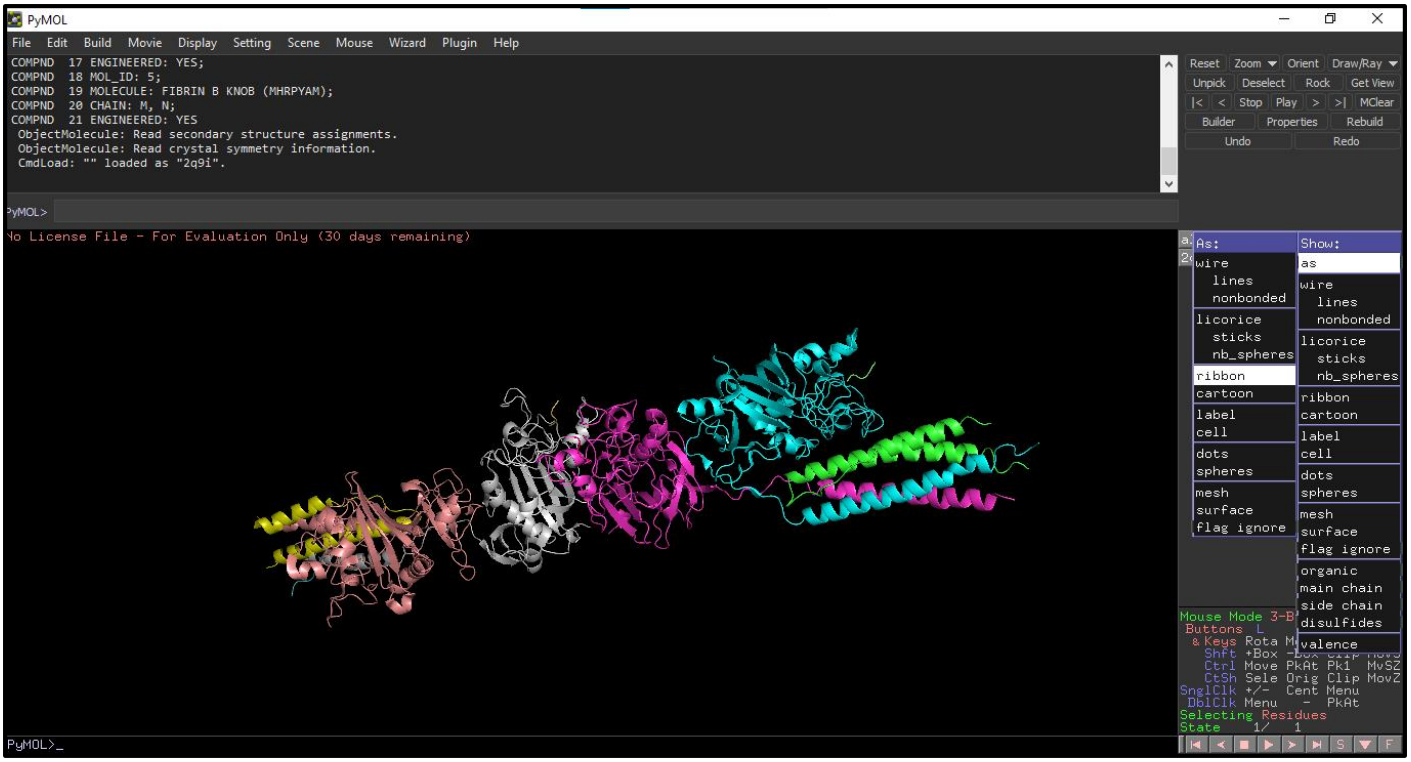


Fig4. Visualisation of Fibrin in ribbon format

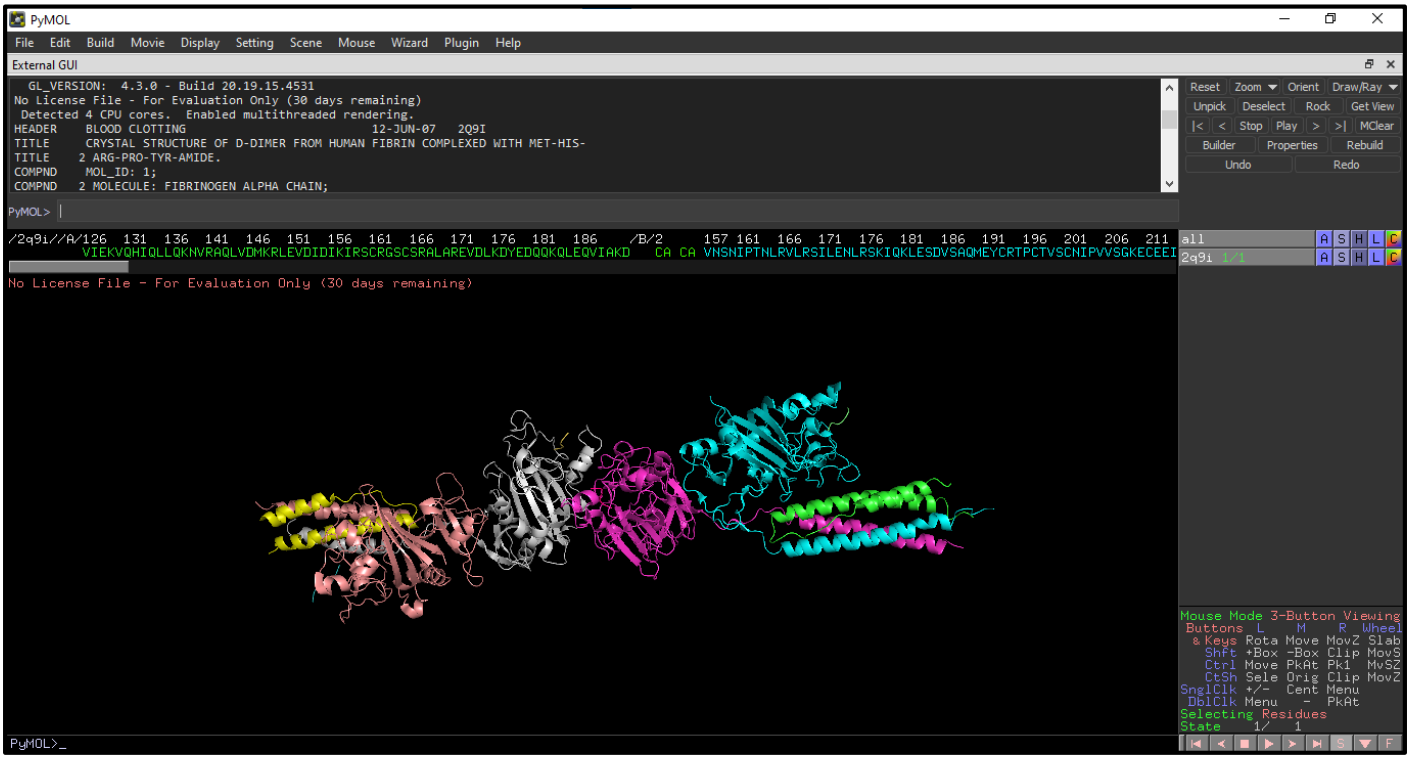


Fig5. Displaying sequence for Fibrin

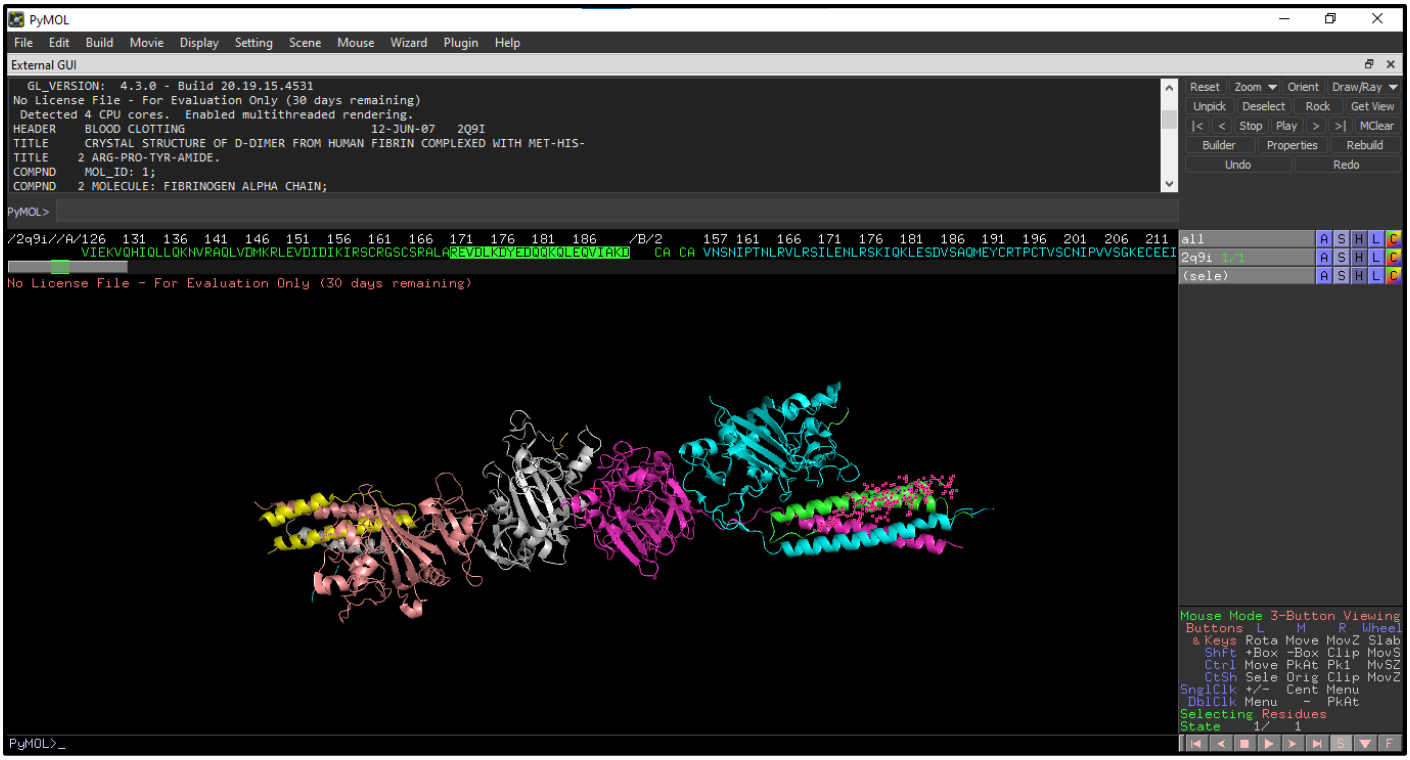


Fig6. Displaying specific residues from sequence

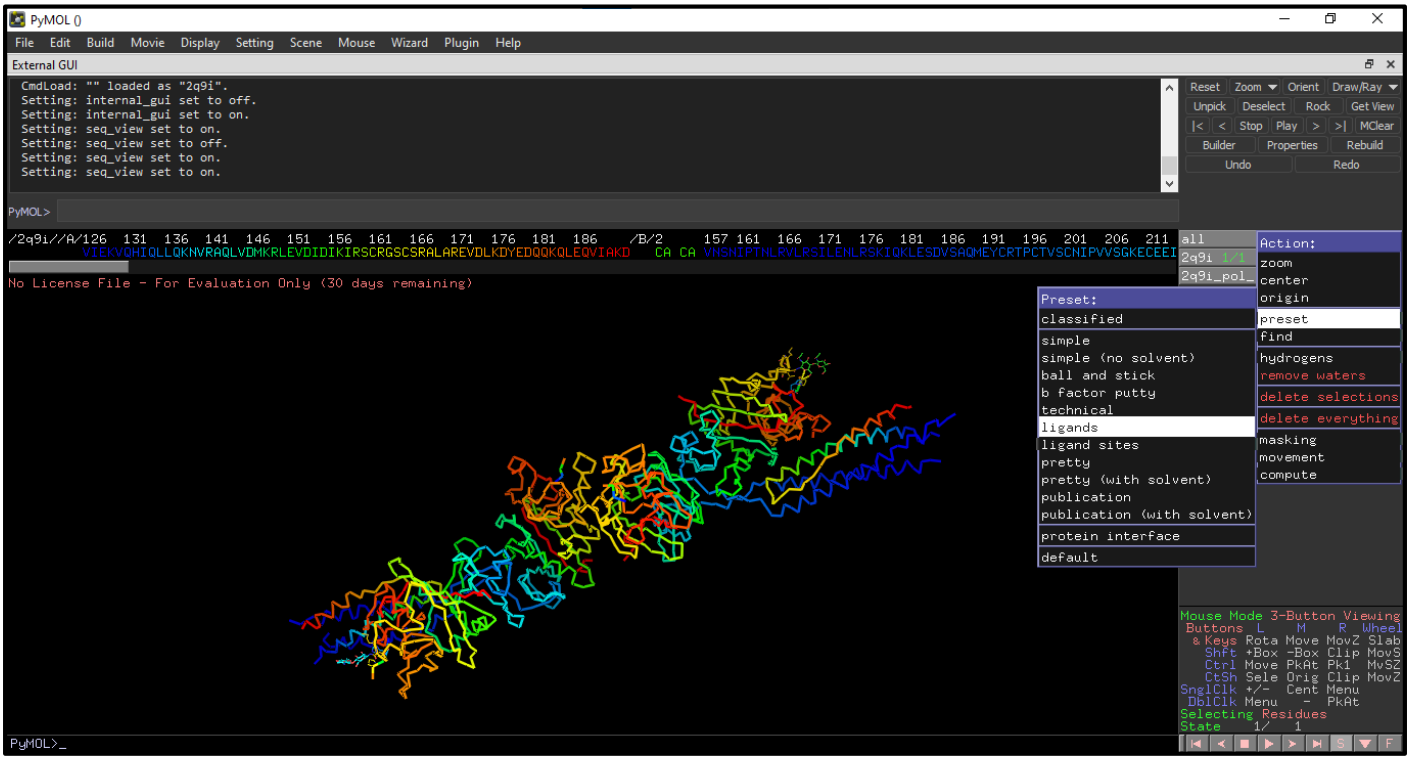


Fig7. Visualisation of ligands

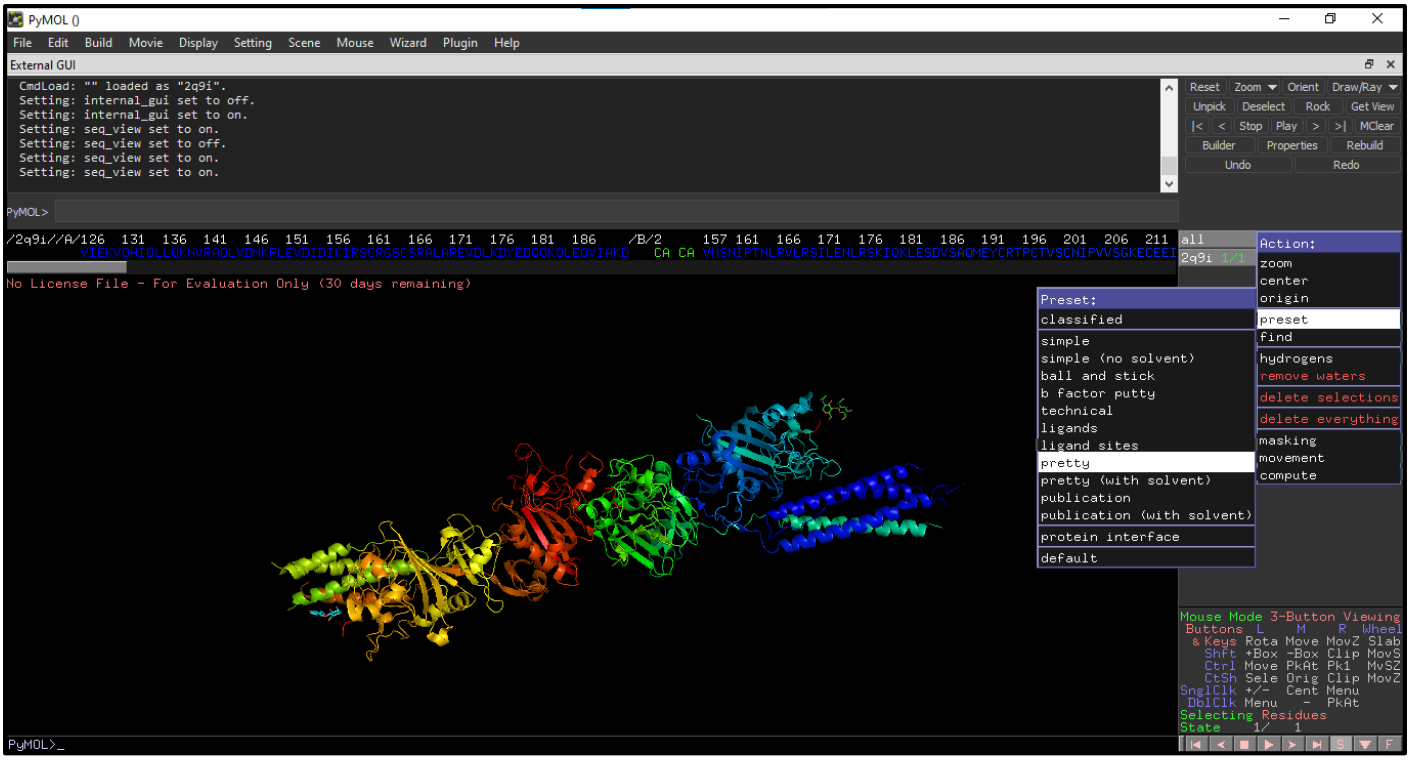


Fig8. Visualisation of Fibrin structure in pretty format

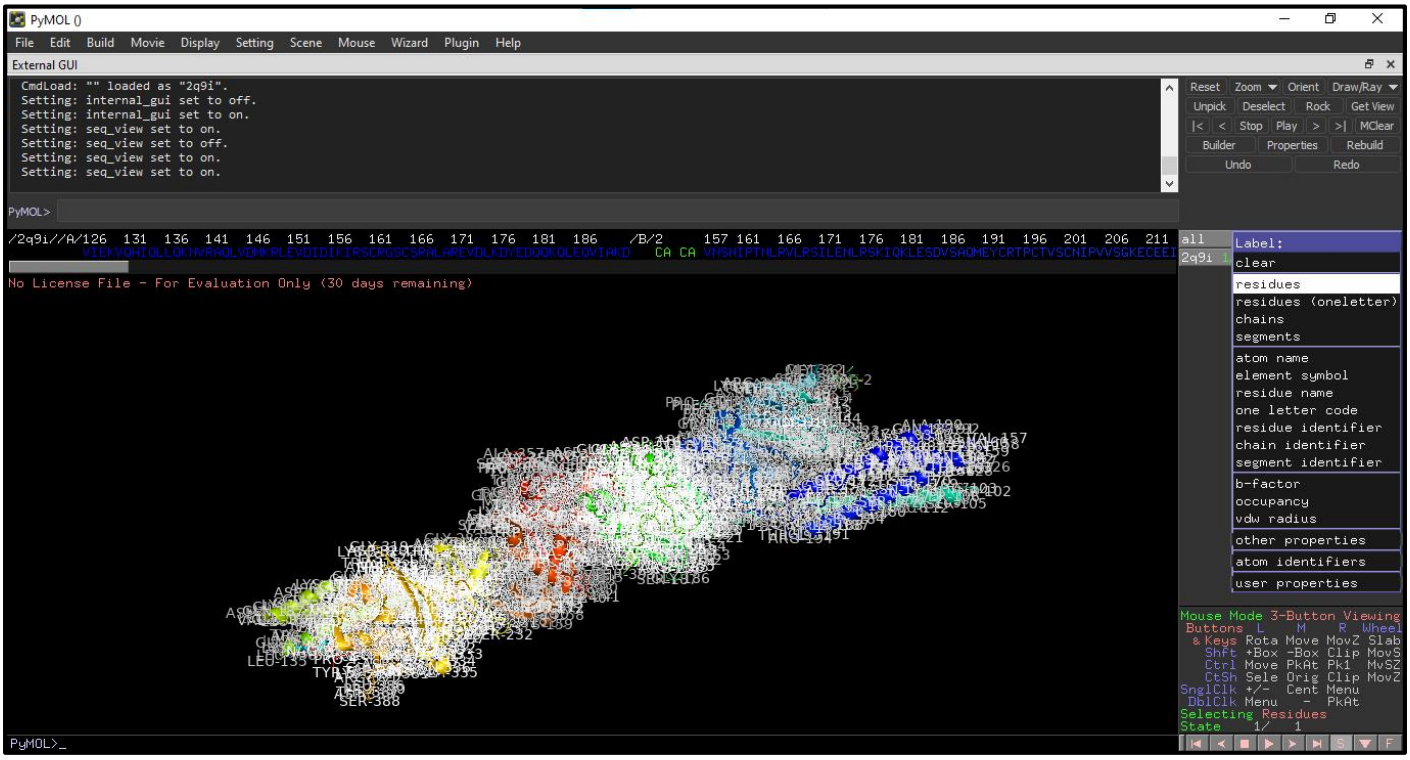


Fig9. Labelling residues of Fibrin

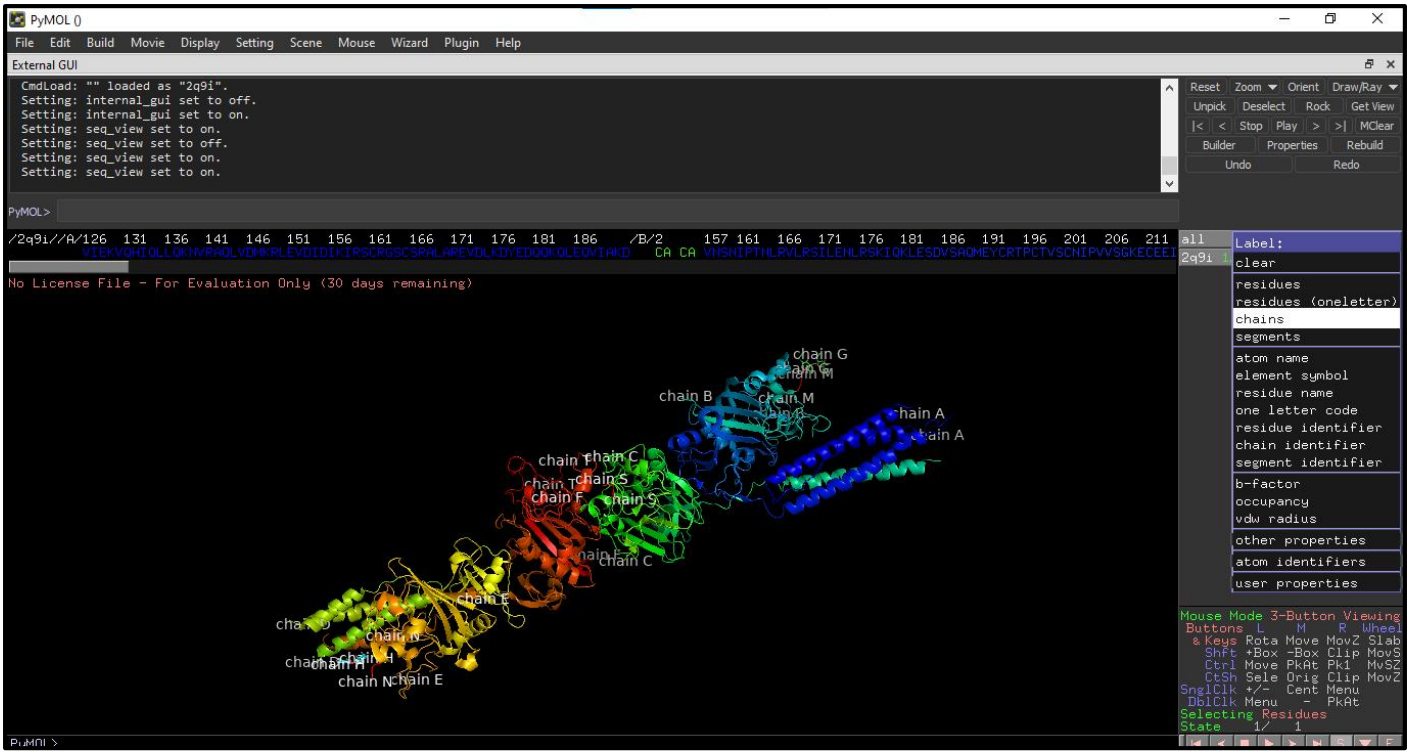


Fig10. Labelling all chains of Fibrin

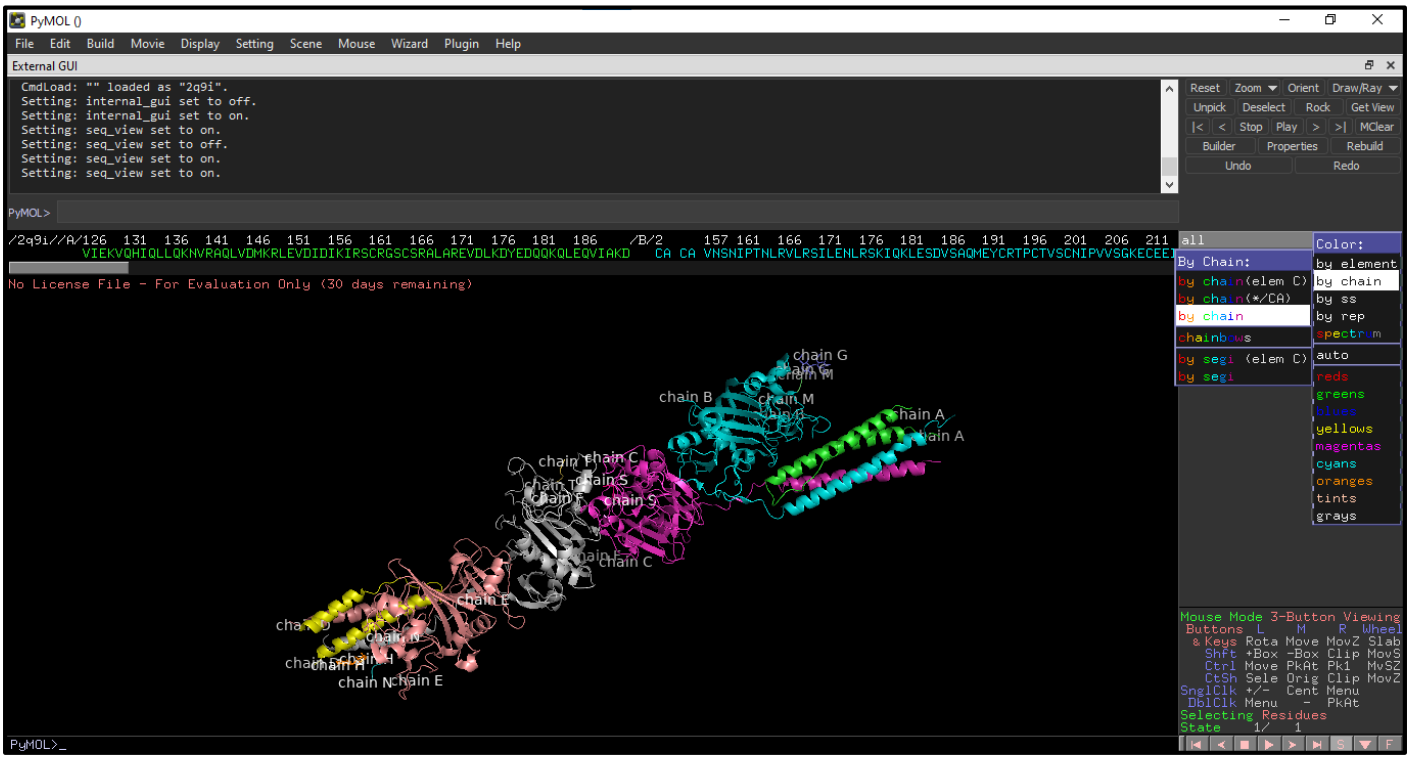


Fig11. Visualisation of all chains of Fibrin in different colours.

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. Retrieved March 4, 2022, from http://www.openrasmol.org/
8. *PyMOL | pymol.org*. (2019). Pymol.org. Retrieved March 4, 2022, from https://pymol.org/2/

DATE: 25/02/22

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: 2iwv 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 A 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 GalNActransferases 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

DATE: 25/02/22

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:

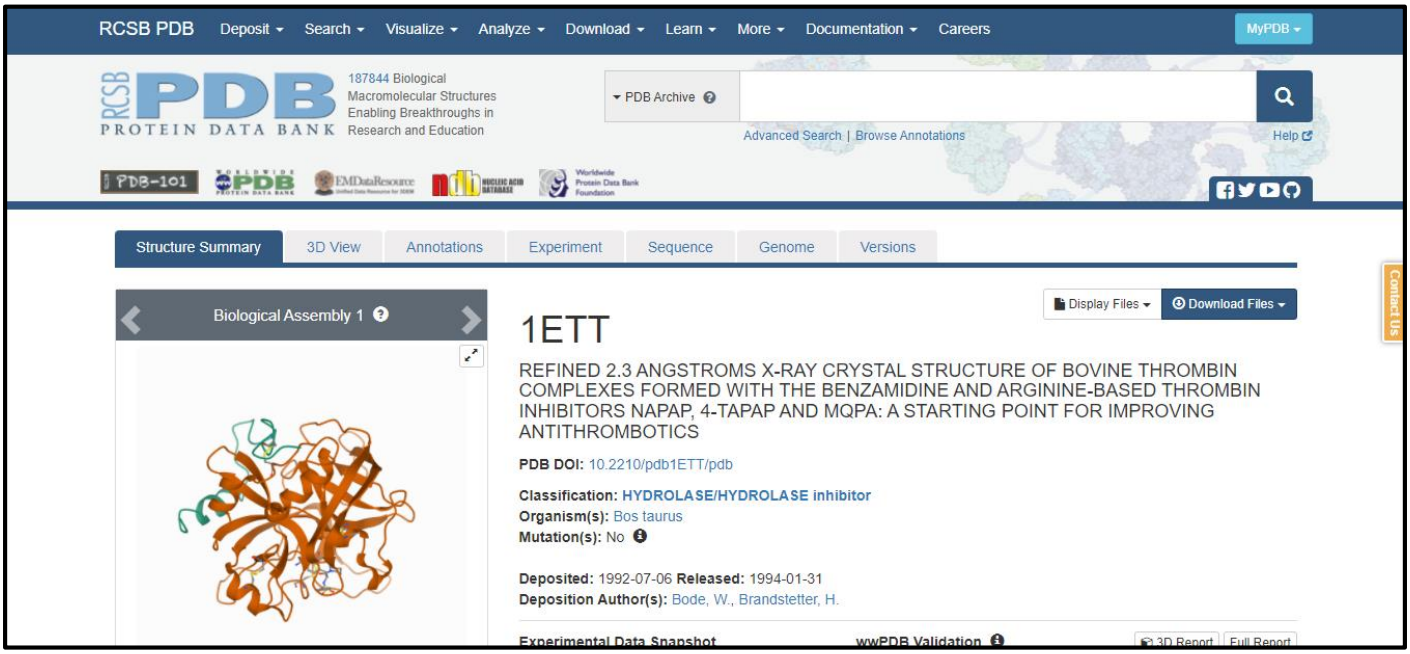


Fig1. PDB structure selected for Thrombin

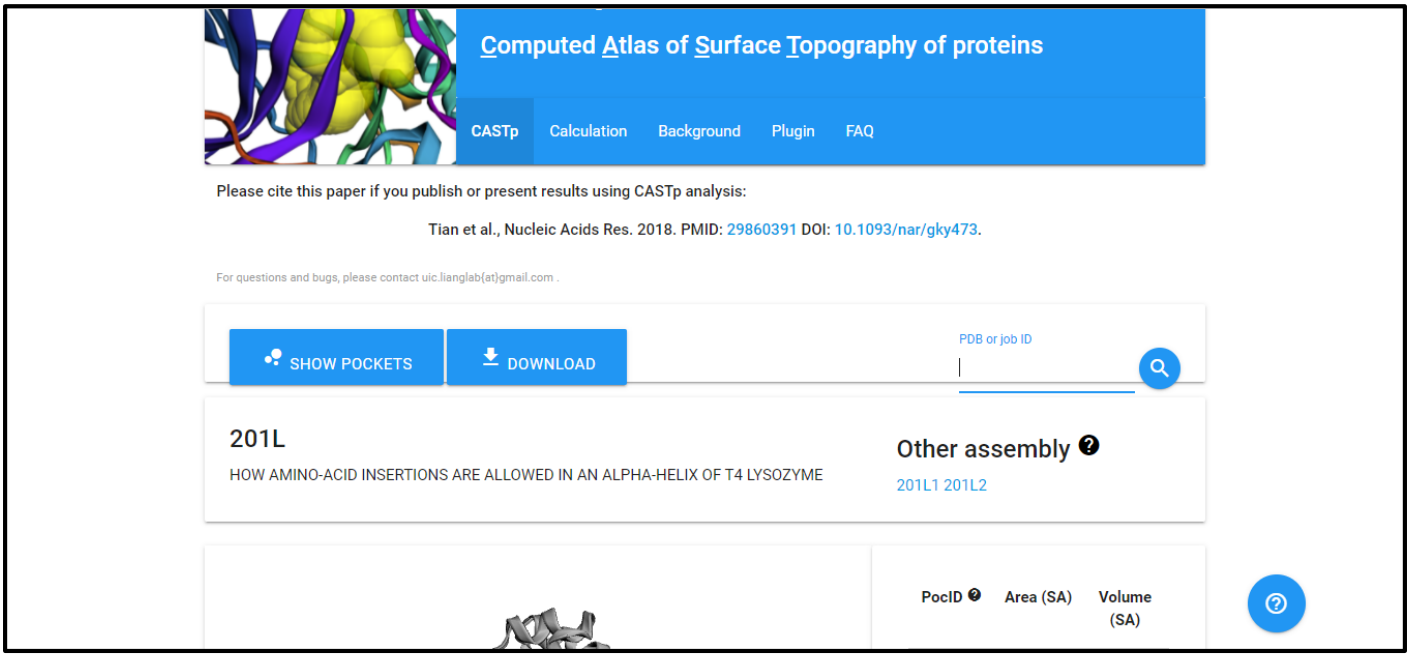


Fig2. Homepage for CASTp server.

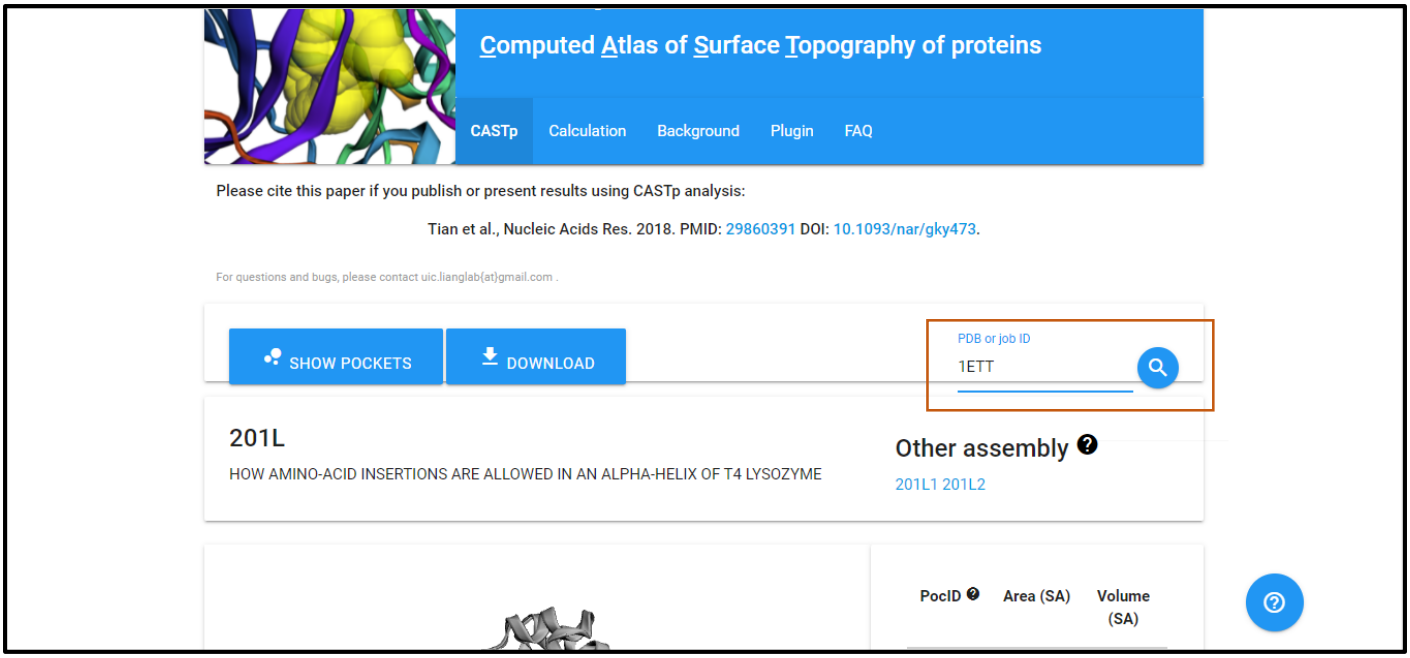


Fig3. Search for Thrombin structure (PDB ID: 1ETT)

Fig4. Result page for Thrombin

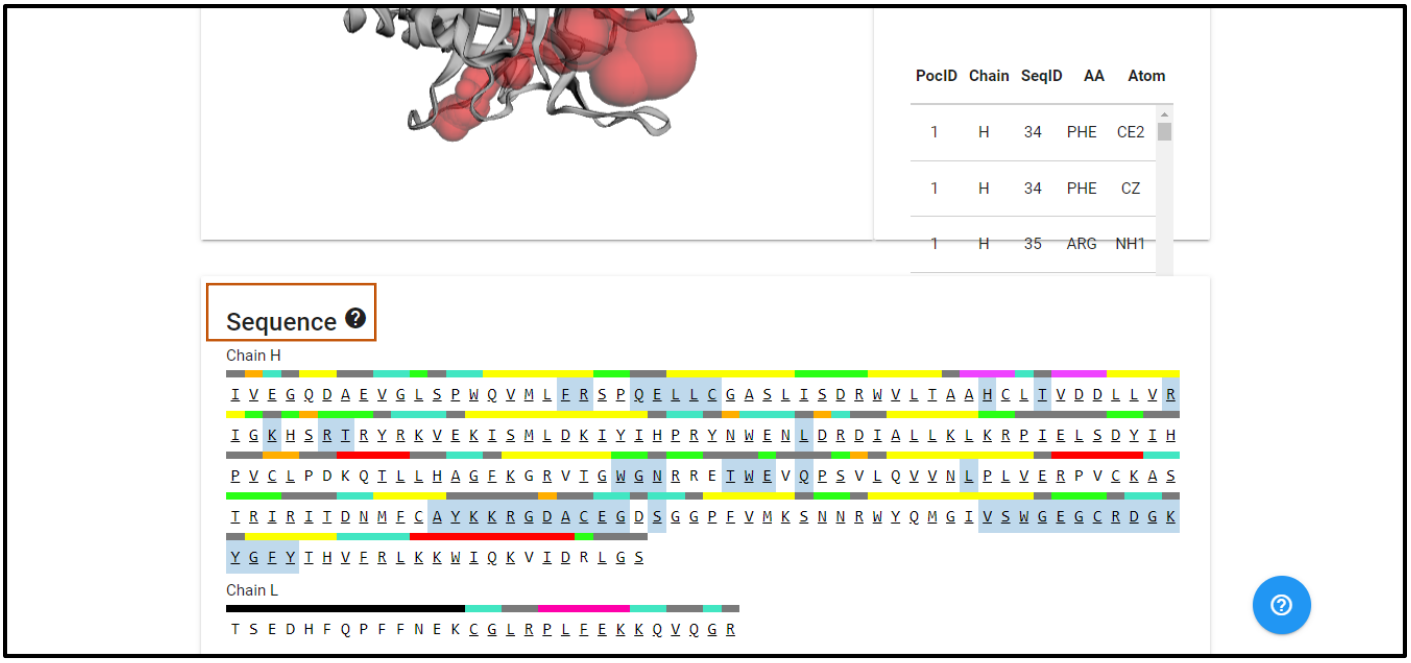


Fig4.1 Result for sequence and structure information.

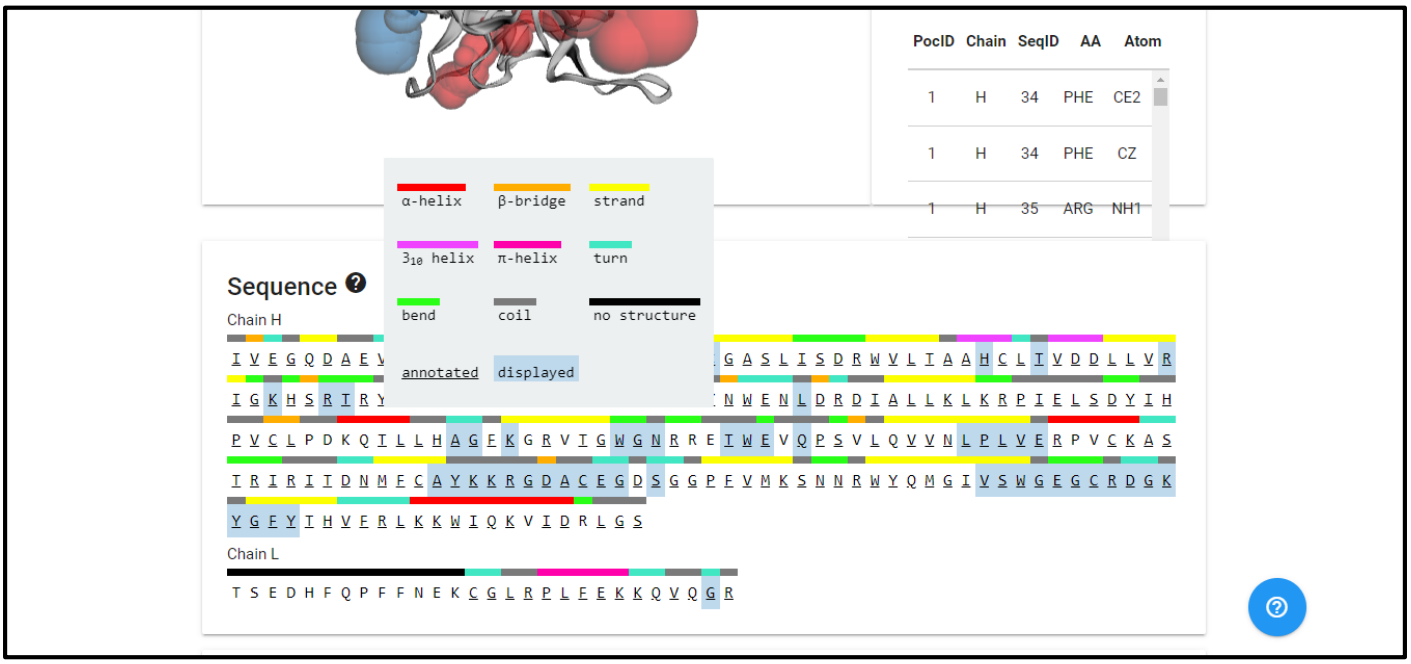


Fig4.2 Legends for sequence information

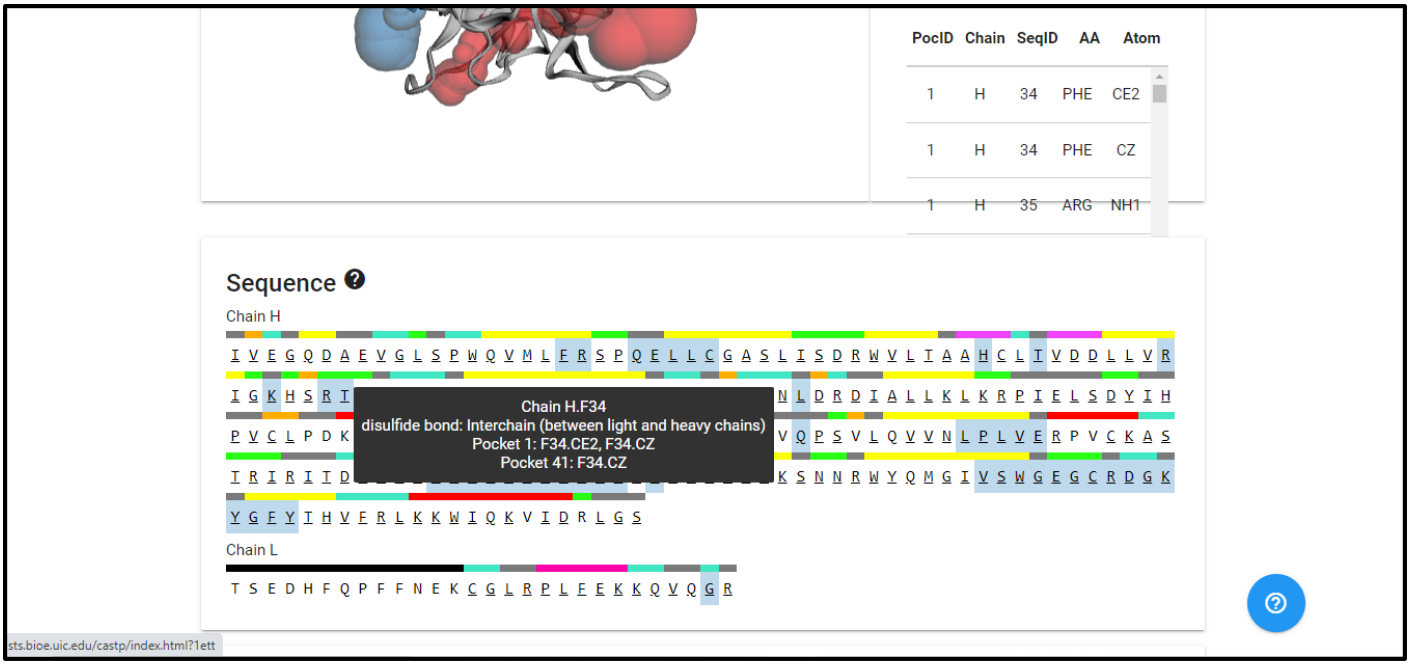


Fig4.3 Result for binding pocket information

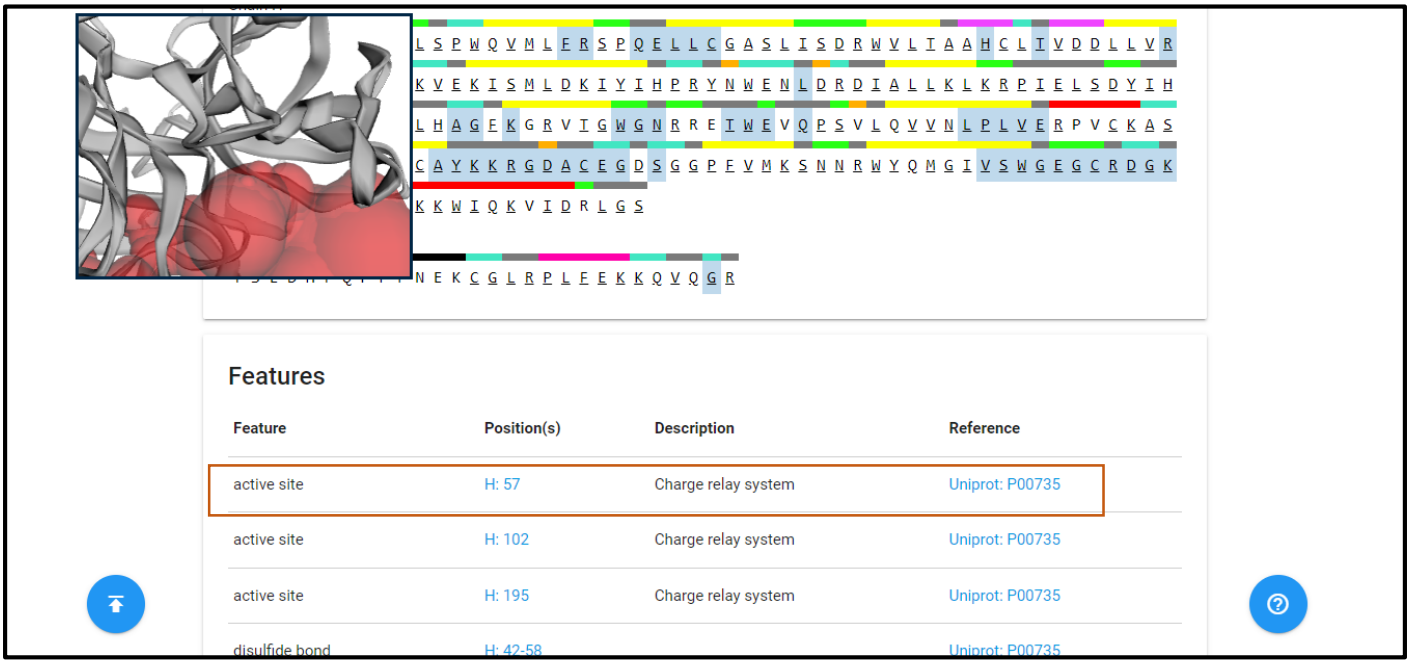


Fig4.4 Result for features

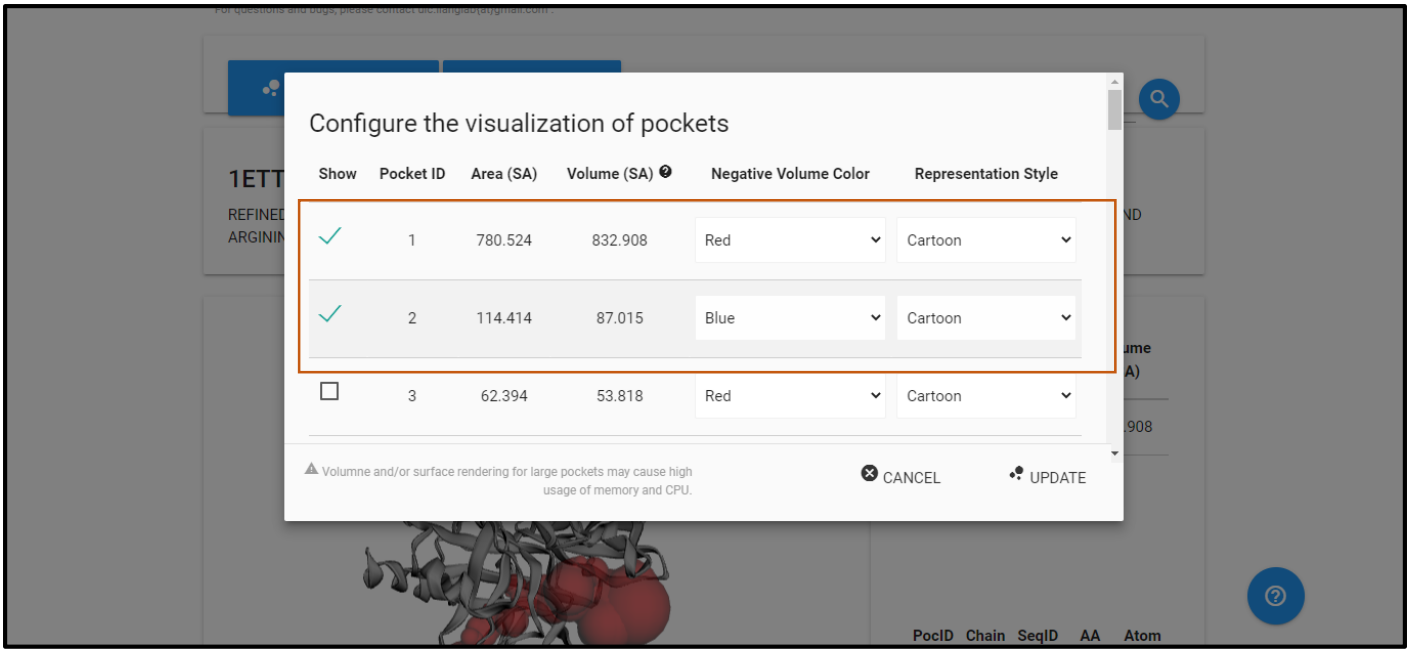


Fig5. Configuration for visualisation of binding pockets

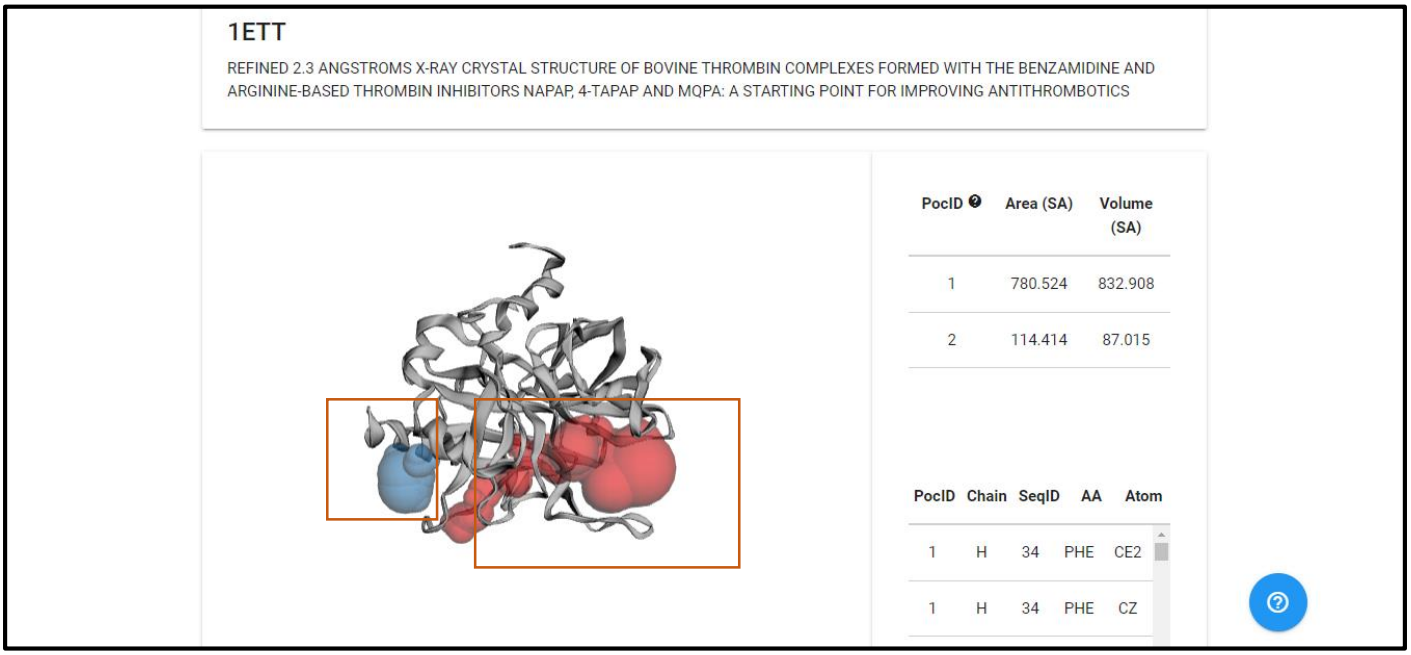


Fig5.1 Result for binding pocket visualisation

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 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:

- 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
- 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

DATE: 25/02/22

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:

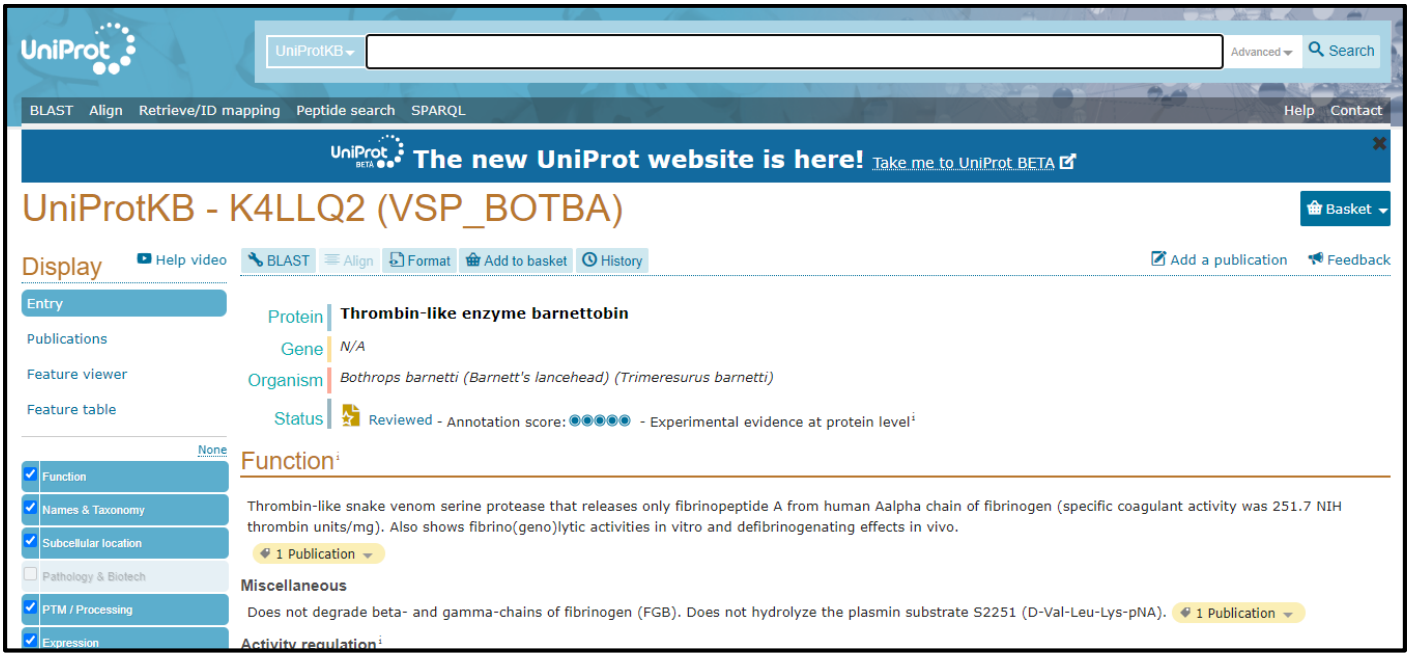


Fig1. Result page for Thrombin in UniProt

```
>sp|K4LLQ2|VSP_BOTBA Thrombin-like enzyme barnettobin (Fragment) OS=Bothrops barnetti OX=1051630 PE=1 SV=1
APKELQVSYAHKSSELVIGGDECDINEHPFLAFLYSRGNFCGLTLINQEWVLTAAHCDRR
FMPIYLGIHTLSVPNDDEVIRYPKDNFICPNNNIIDEKDKDIMLIRLNRPVKNSEHIAPI
SLPSNLPSVGSVCRVMGWGSITAPNDTFPDVPHCANINLFNDTVCHGAYKRFPVKSRTLC
AGVLQGGKDKCMGDSGGPLICNGPFHGILFWGDDPCALPRKPALYTKGFEYPPWIQSIIA
KNTTETCPP
```

Fig2. FASTA sequence for Thrombin

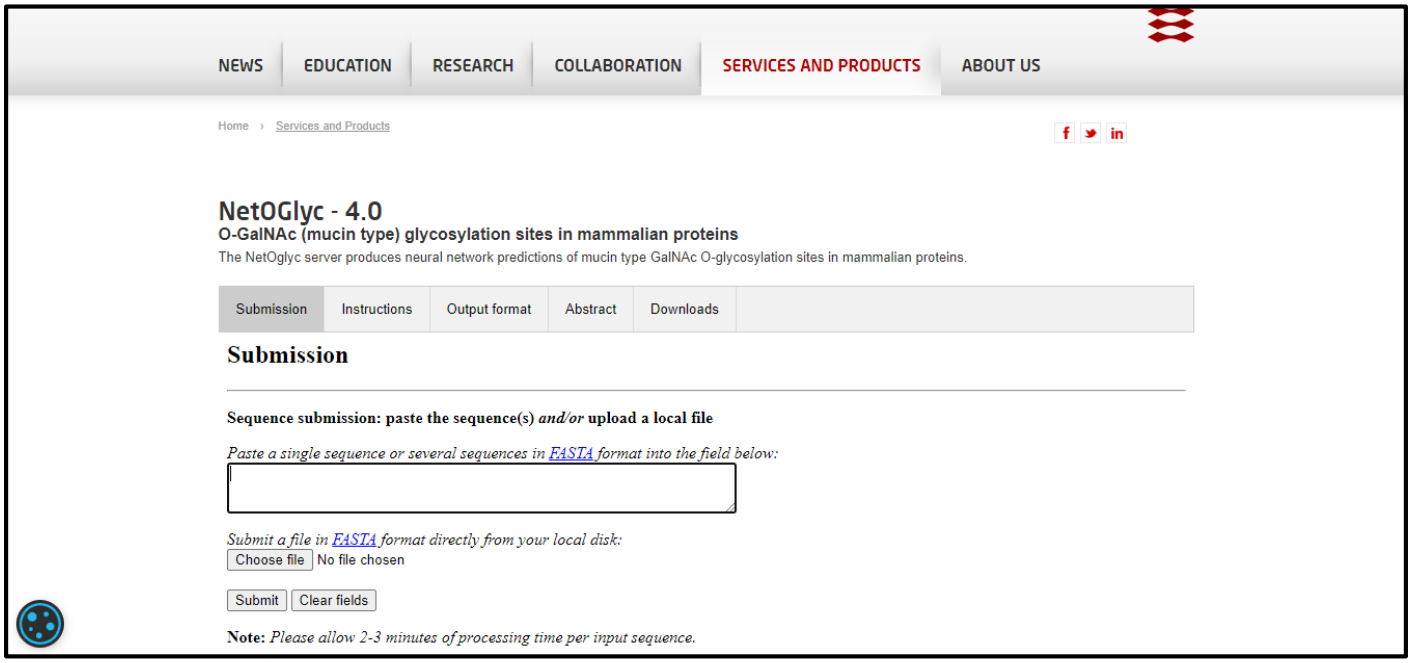


Fig3. Homepage for NetOGlyc-4.0 server.

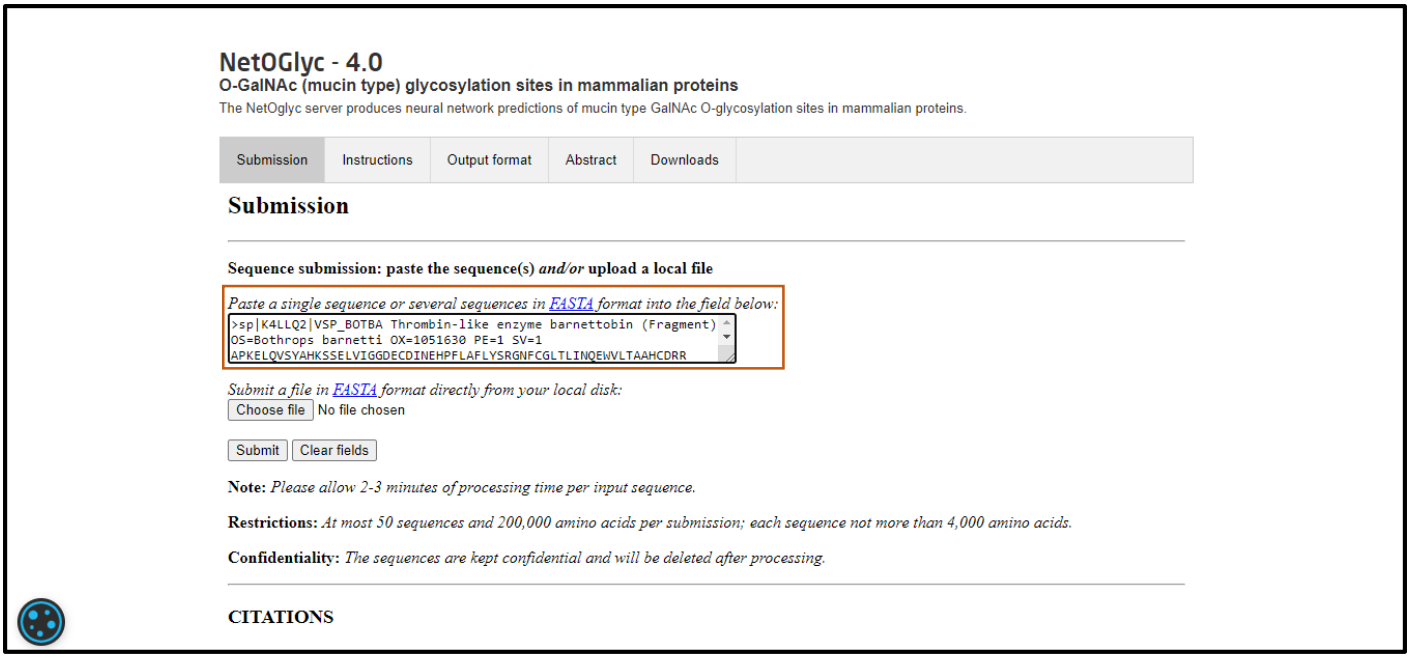


Fig4. Search for Thrombin using FASTA sequence.

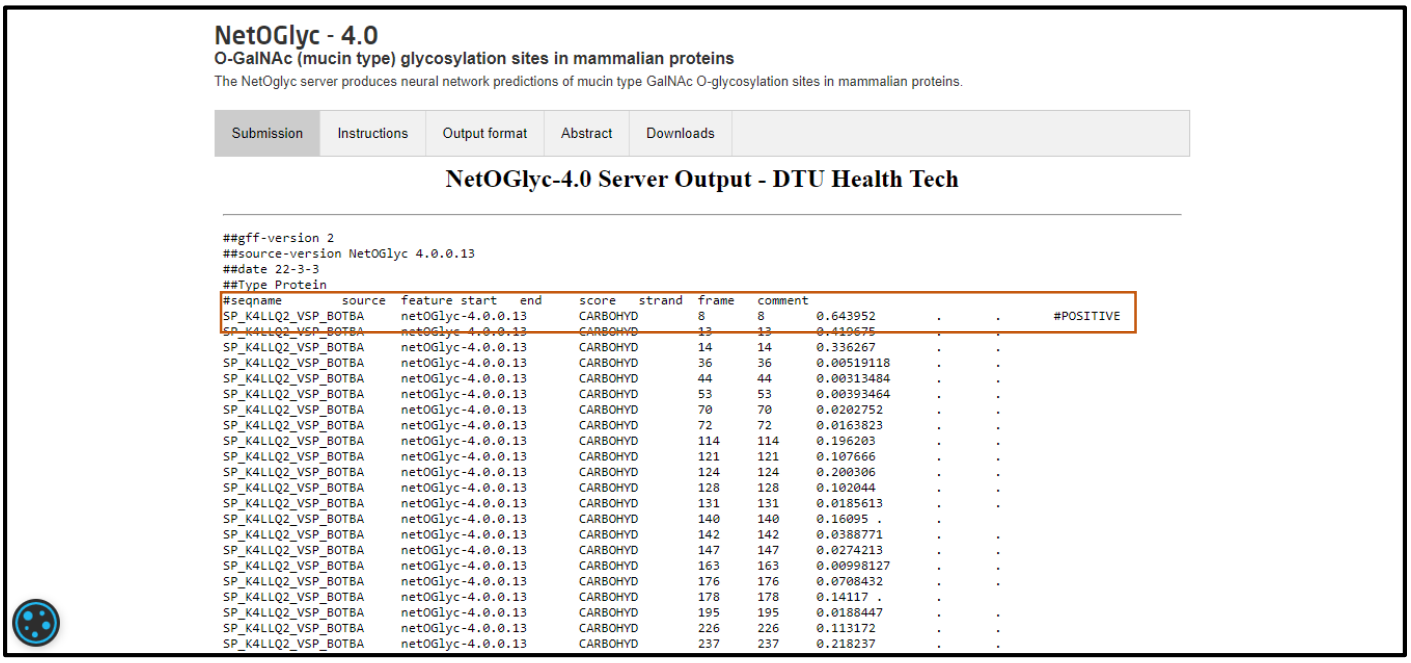


Fig5. Result page for Thrombin showing predicted glycosylation sites.

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
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. *Thrombin-like enzyme barnettobin precursor - Bothrops barnetti (Barnett's lancehead)*. (n.d.). Www.uniprot.org. Retrieved March 3, 2022, from https://www.uniprot.org/uniprot/K4LLQ2
4. *Services*. (n.d.). Https://Www.healthtech.dtu.dk. Retrieved March 3, 2022, from https://services.healthtech.dtu.dk/service.php?NetOGlyc-4.0

DATE: 25/02/22

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:

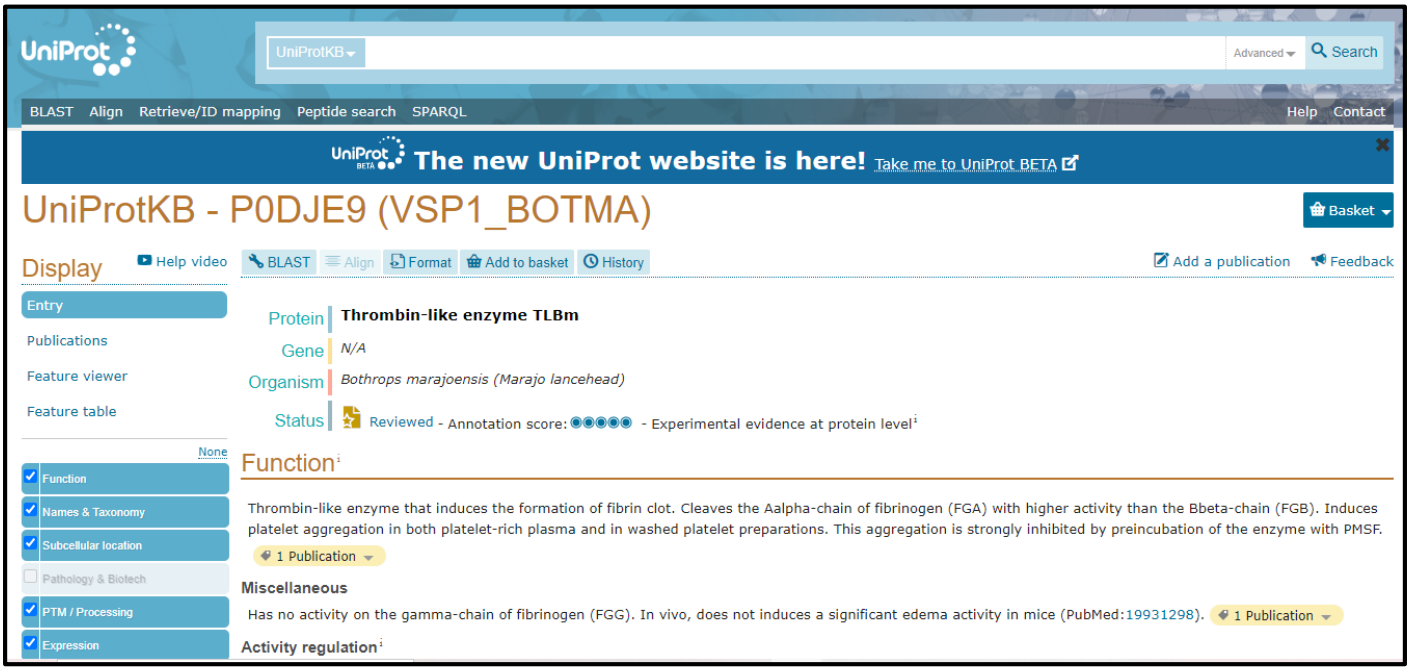


Fig1. Result page for Thrombin in UniProt

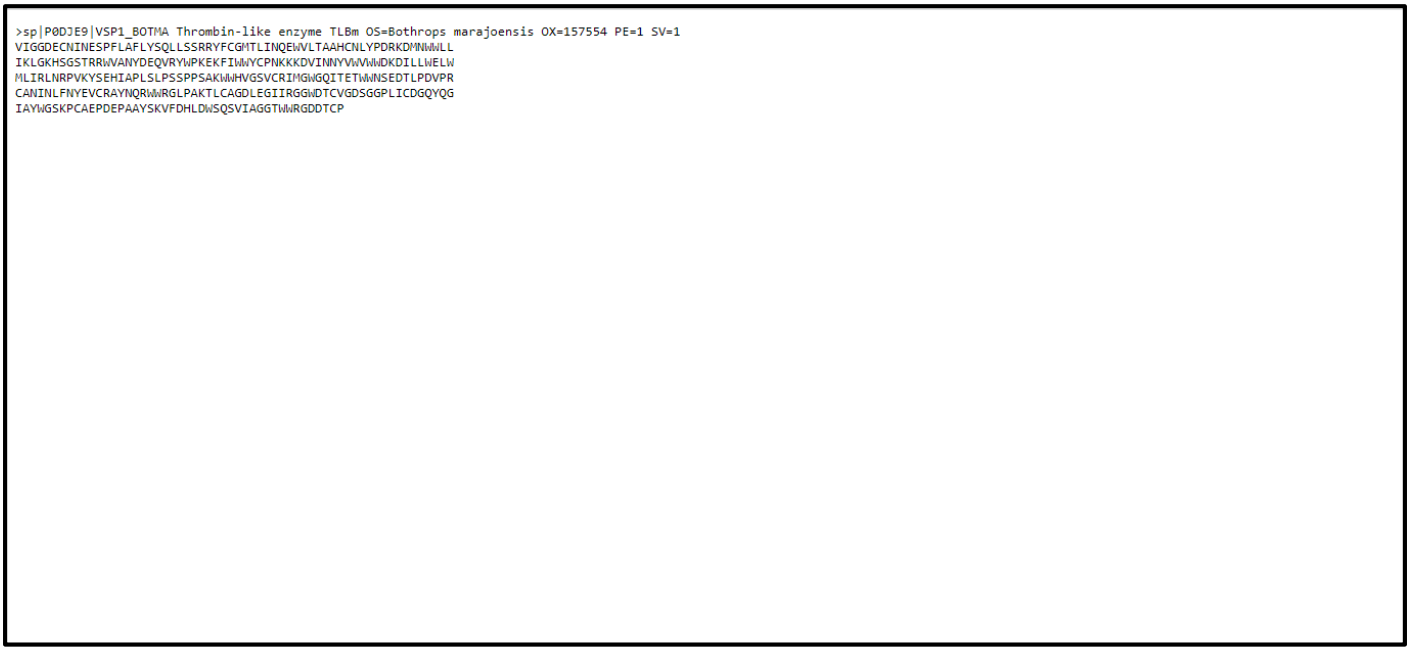


Fig2. FASTA sequence for Thrombin

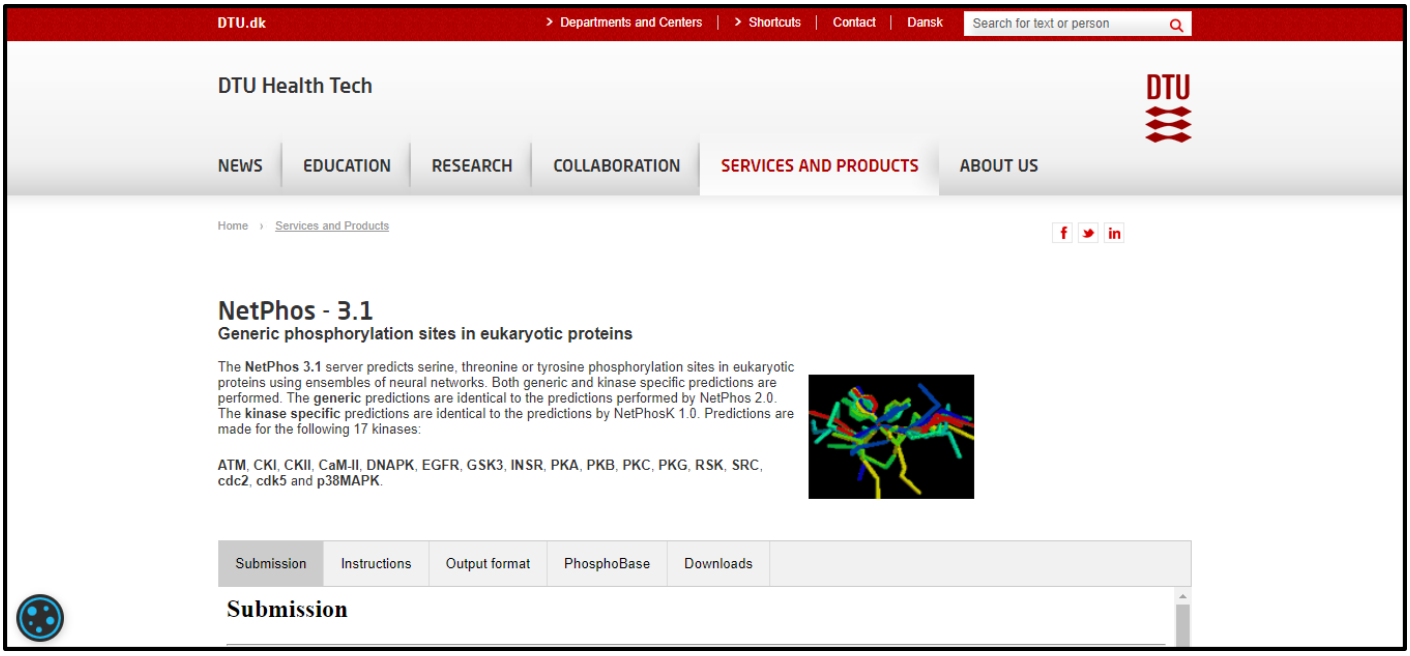


Fig3. Homepage for NetPhos-3.1 server

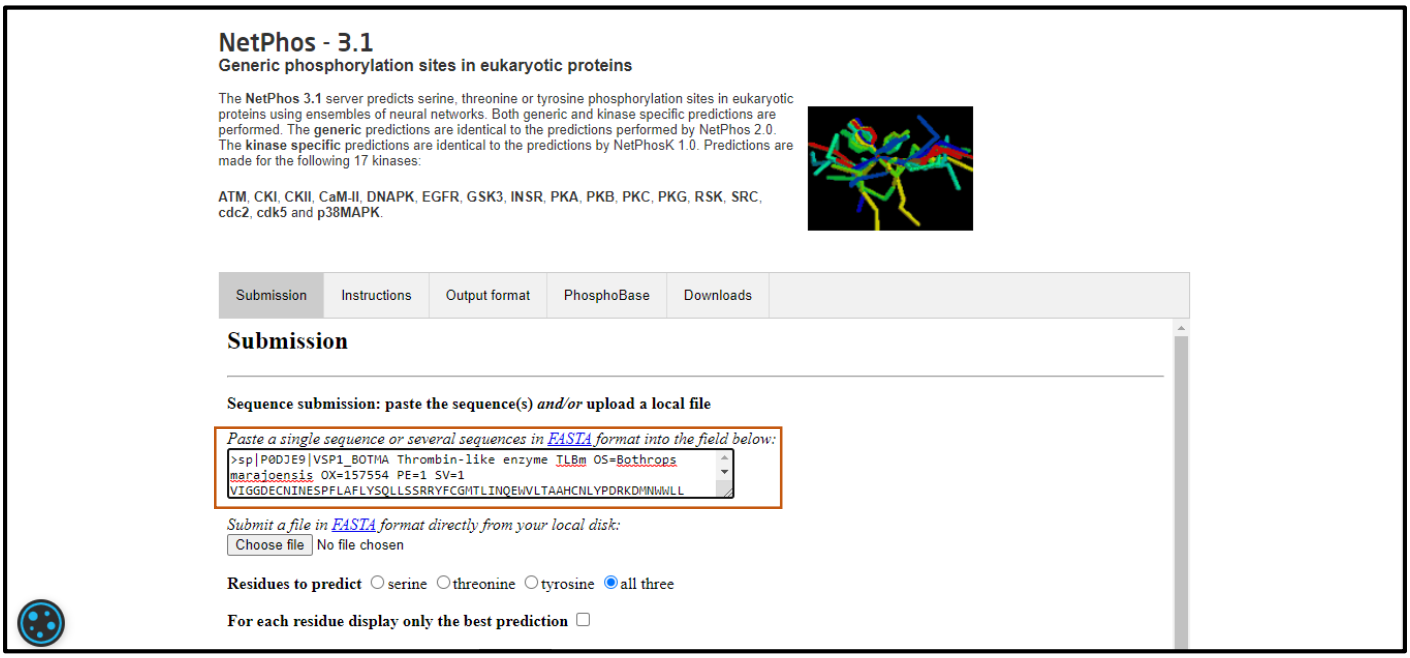


Fig4. Search page for Thrombin using FASTA sequence.

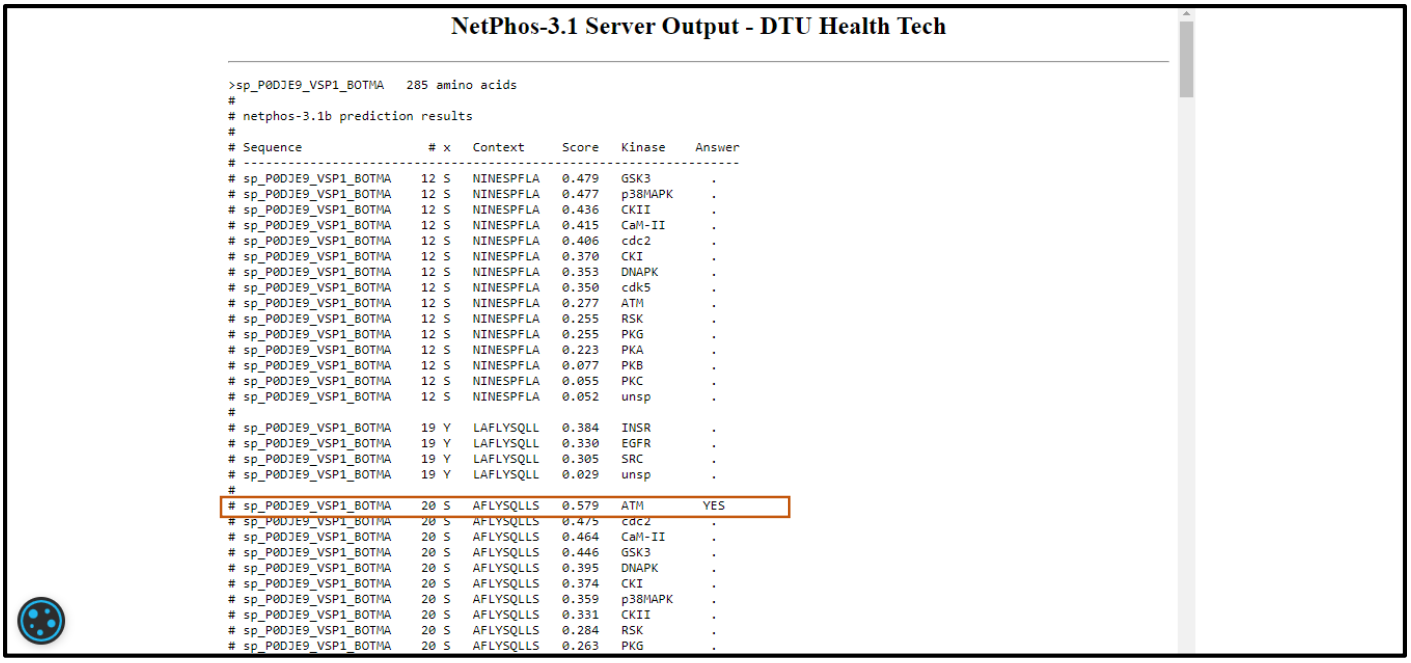


Fig5. Result page for Thrombin showing predicted phosphorylation sites.

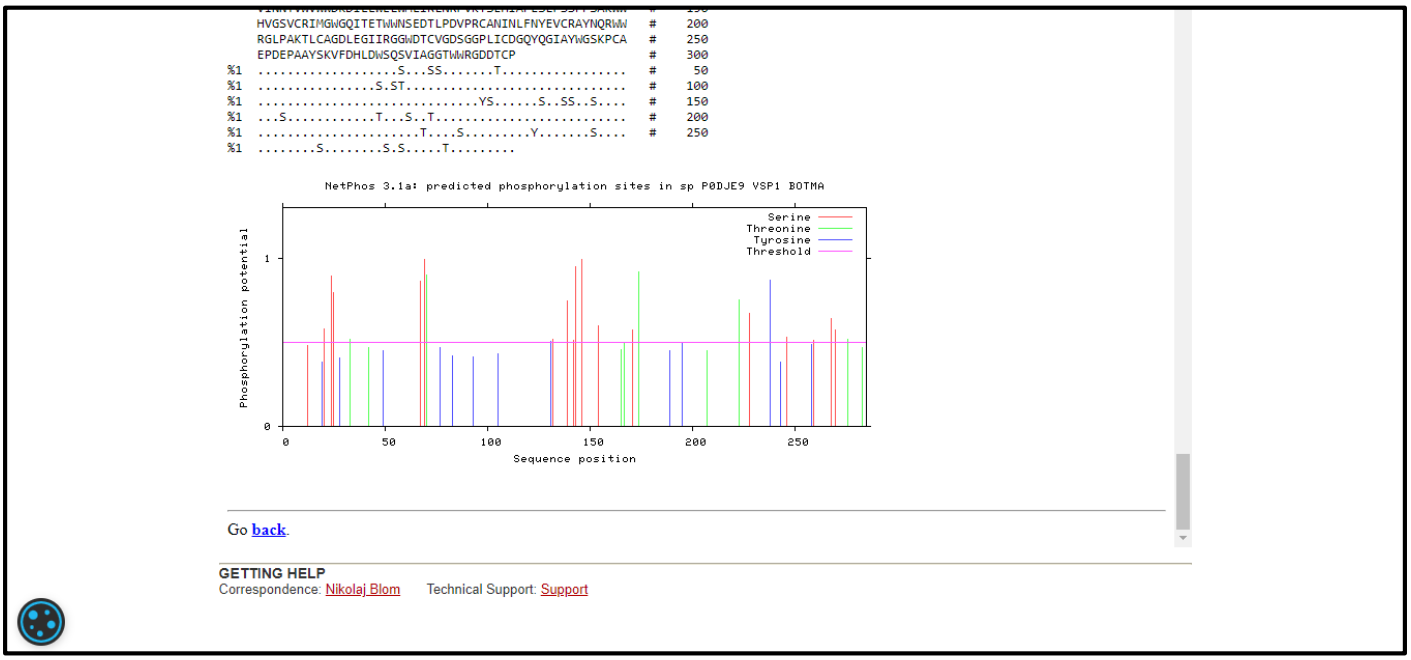


Fig6. Graphical representation of predicted phosphorylation sites.

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
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. Retrieved March 3, 2022, from https://www.uniprot.org/uniprot/P0DJE9
5. *Services*. (n.d.). Https://Www.healthtech.dtu.dk. Retrieved March 3, 2022, from https://services.healthtech.dtu.dk/service.php?NetPhos-3.1

DATE: 14/03/22

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 homo-oligomers to intricate arrangements of many different components, revealing details on specific molecular interactions and on how these might constrain sequence variation. A small number of approaches have been published in the past few years that examine structural similarity of macromolecular complexes. Here we present a simple strategy that builds on the existing database of pairwise structure alignments computed by VAST and supports the first (to our knowledge) comprehensive and regularly updated collection of macromolecular complex similarities.

VAST+ AS AN EXTENSION TO EXISTING PROTEIN STRUCTURE COMPARISON

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

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

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

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

DALI:

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

User can perform three types of database searches:

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

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

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

DESCRIPTION OF THE SERVER

Inputs

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

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

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

Processing

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

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

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

The Dali server aims to retrieve a list of 500 structural neighbors of the query structure with the highest *Z*-scores. Most query structures have strong similarity to a structure already in the PDB. We use fast filters to identify a shortlist of about 100 promising candidates. If these produce strong matches, the search proceeds by walking. Otherwise, the query structure is compared with PDB90 in one versus all fashion, followed by a walk to collect matches to redundant PDB structures (which are over 90% sequence identical to PDB90 representatives).

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

Outputs

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

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

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

VAST and DALI are thus very useful structure similarity BLAST tool. VAST provides user with similar structures to their query along with its molecular components and chemicals and non-standard biopolymers, aligned sequences and 3D structure superimposition information which includes information regarding H-bonds, interactions, buried surface area, 2D interaction network and much more. DALI provides user with similar structures to their query along with its pairwise alignment, coordinates information, 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

DATE: 14/03/22

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:

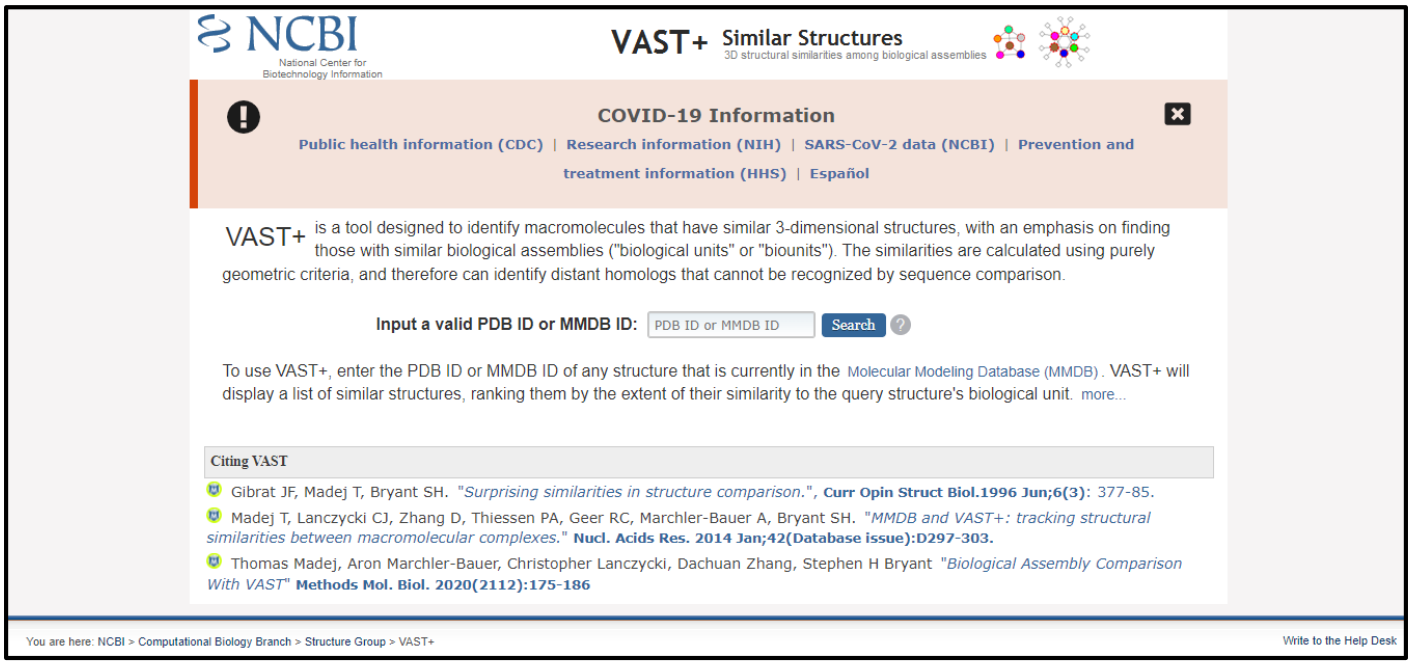


Fig1. Homepage for VAST

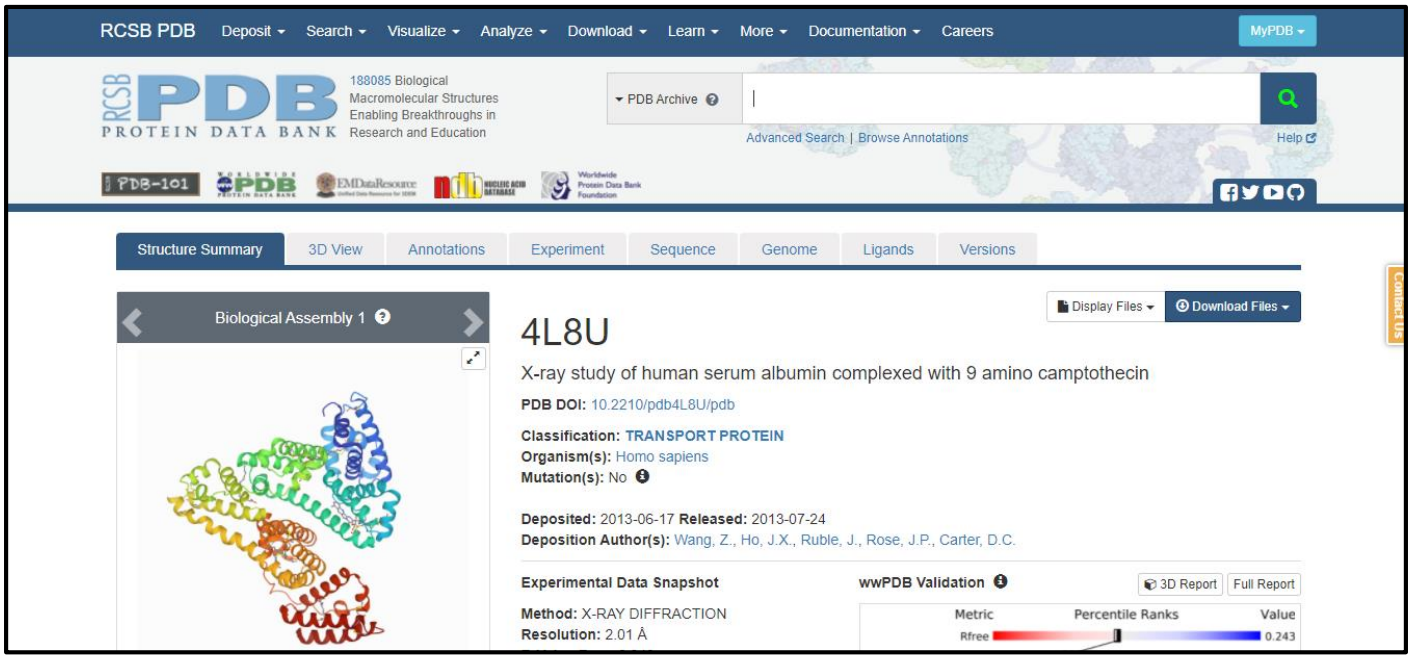


Fig2. Albumin PDB structure

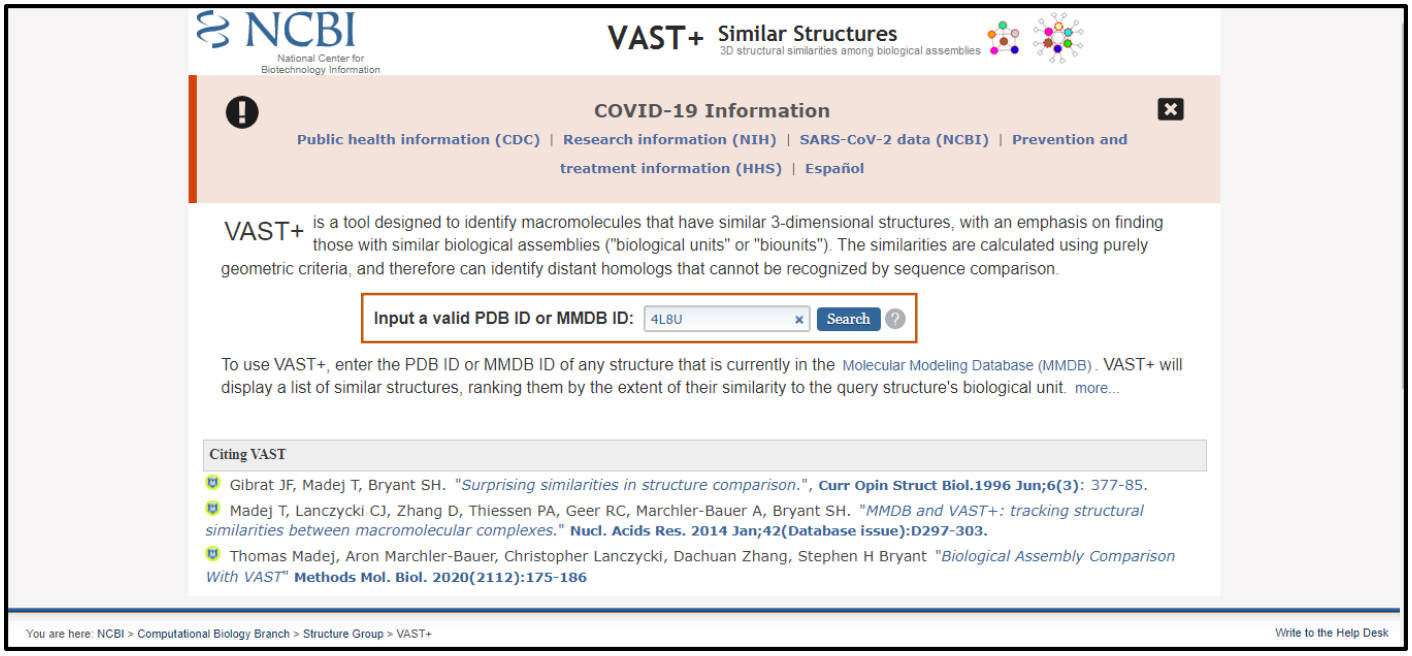


Fig3. Search for Albumin PDB structure

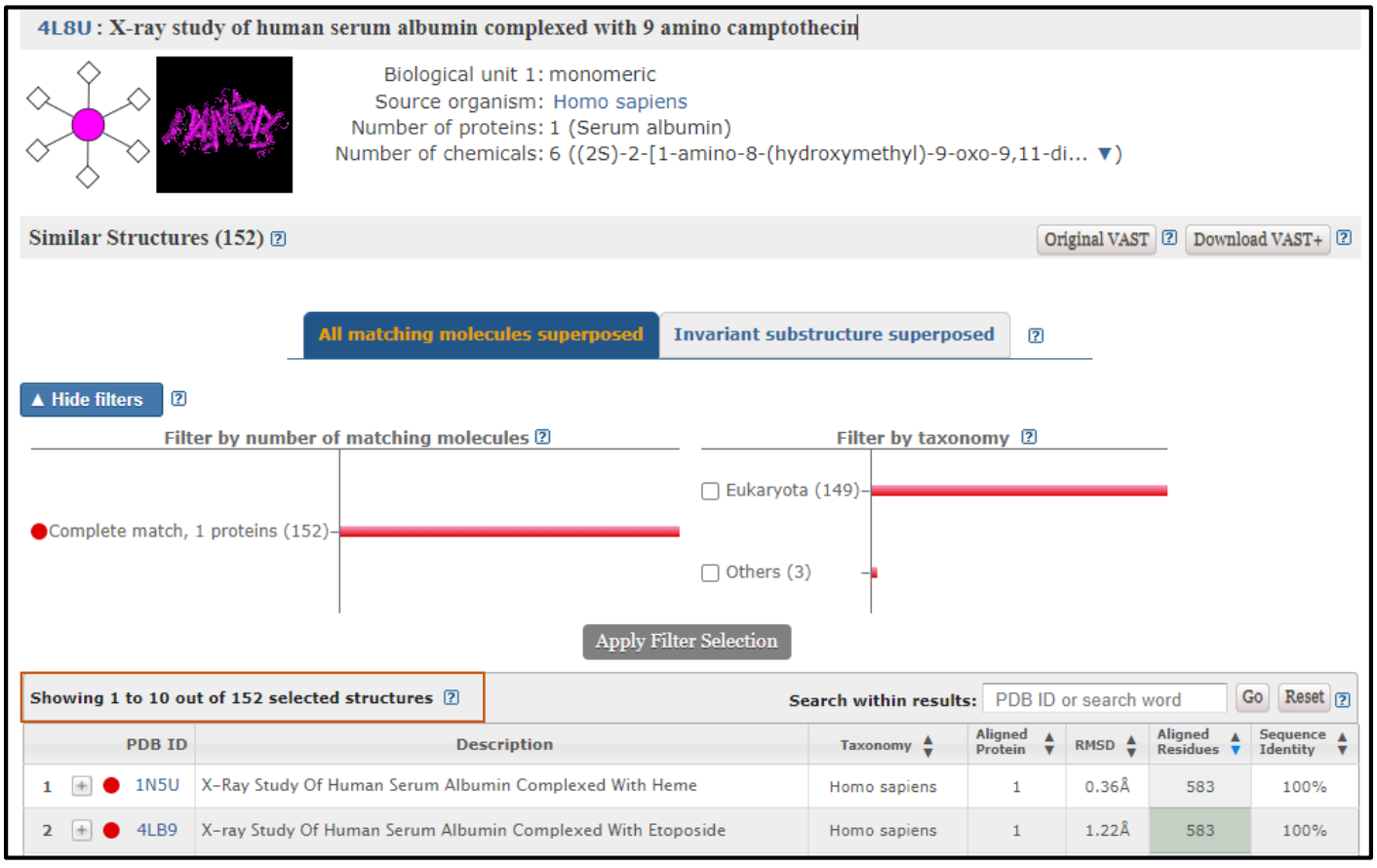


Fig4. Hit page for Albumin

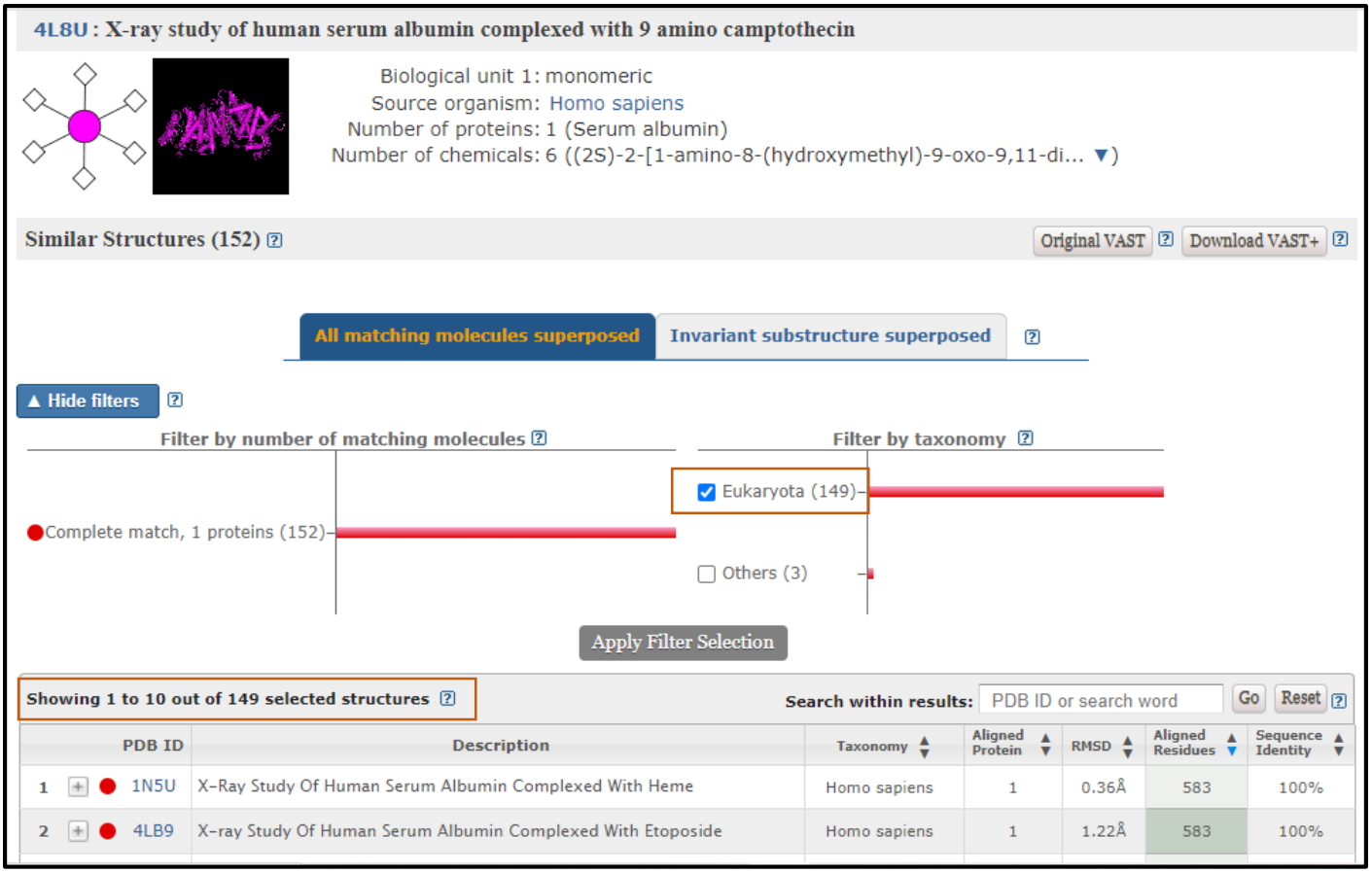


Fig4.1. Hit page after applying filter

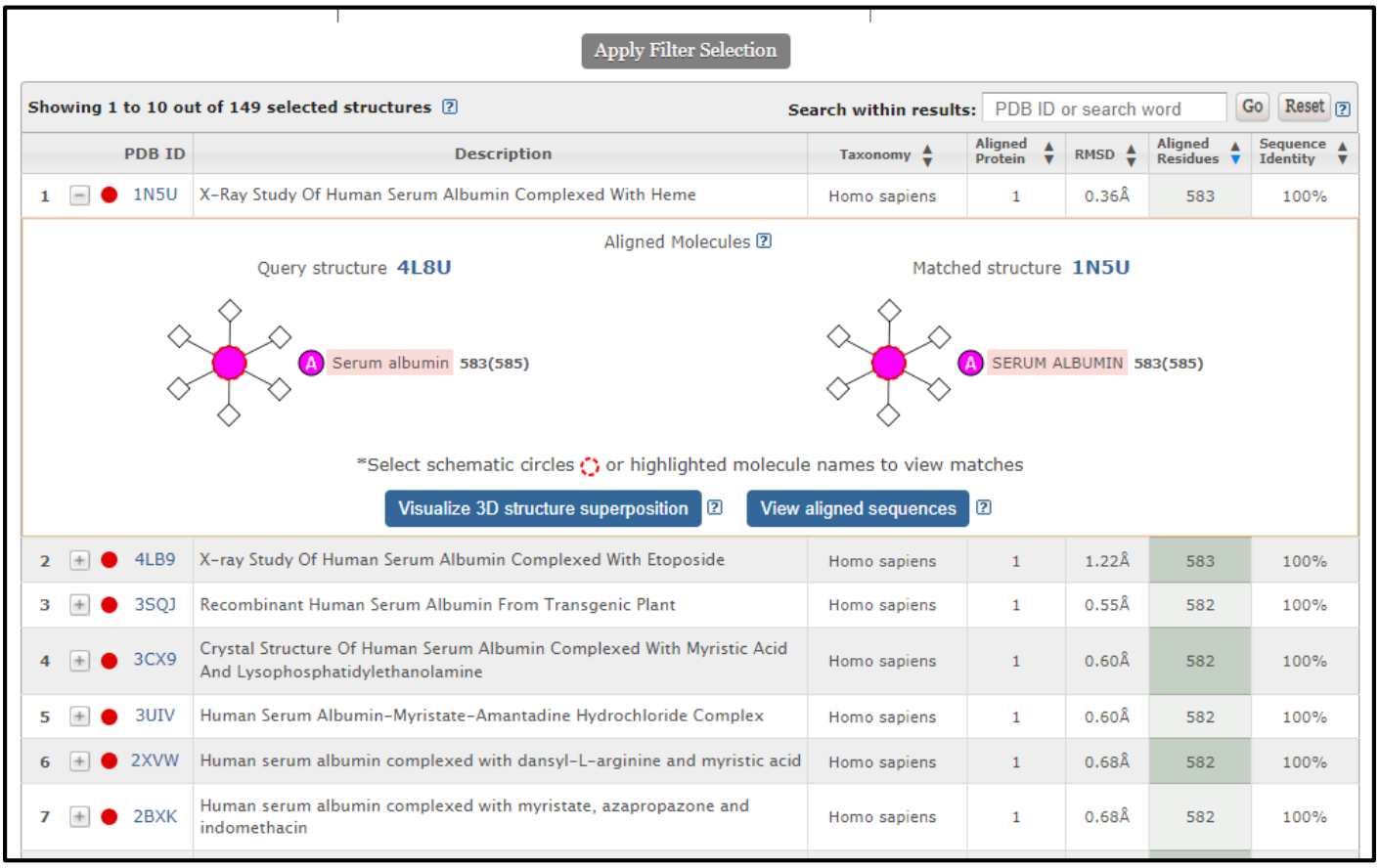


Fig5. Result for aligned molecules

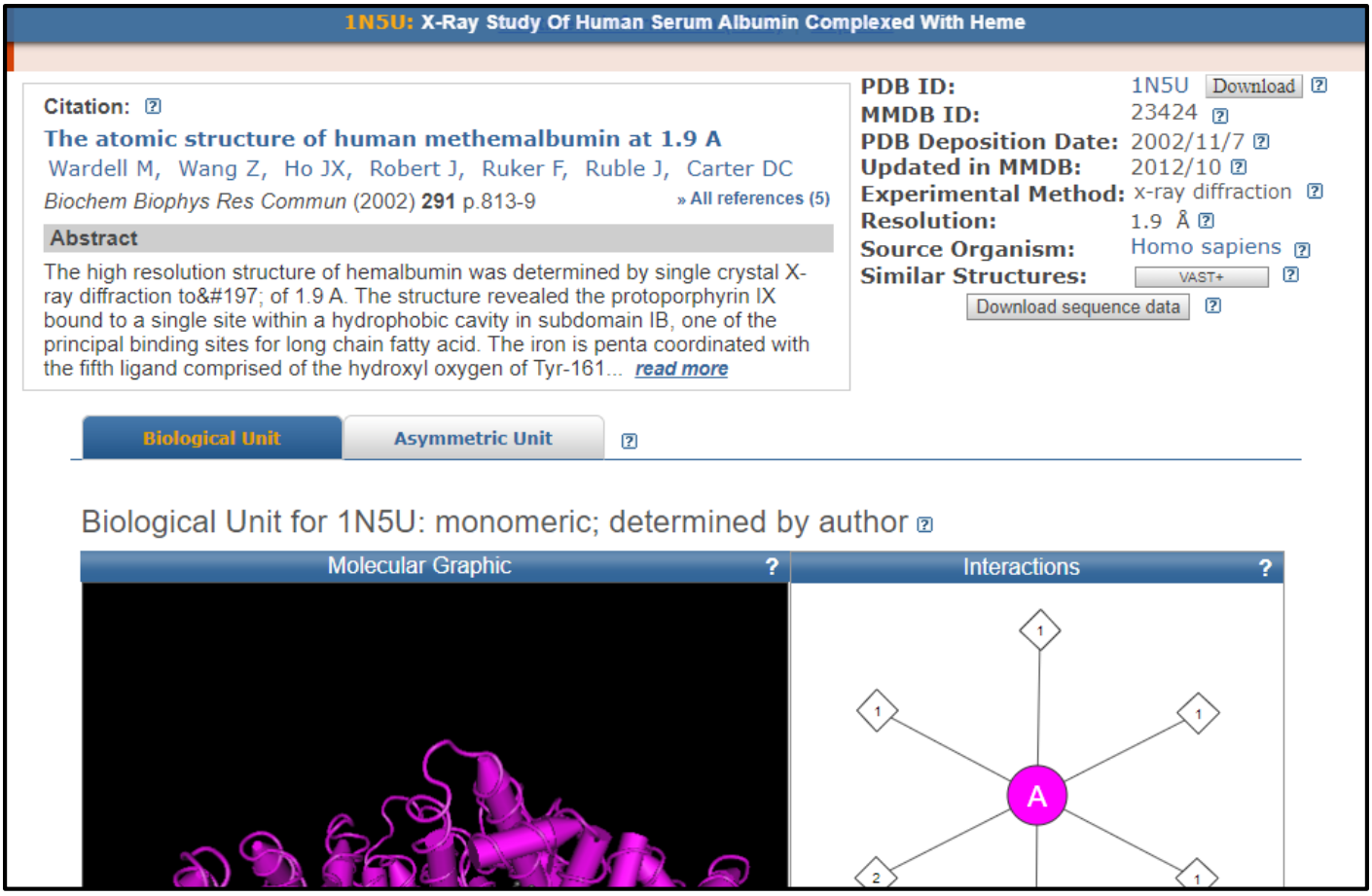


Fig5.1. Result page for structure (1N5U) similar to Albumin

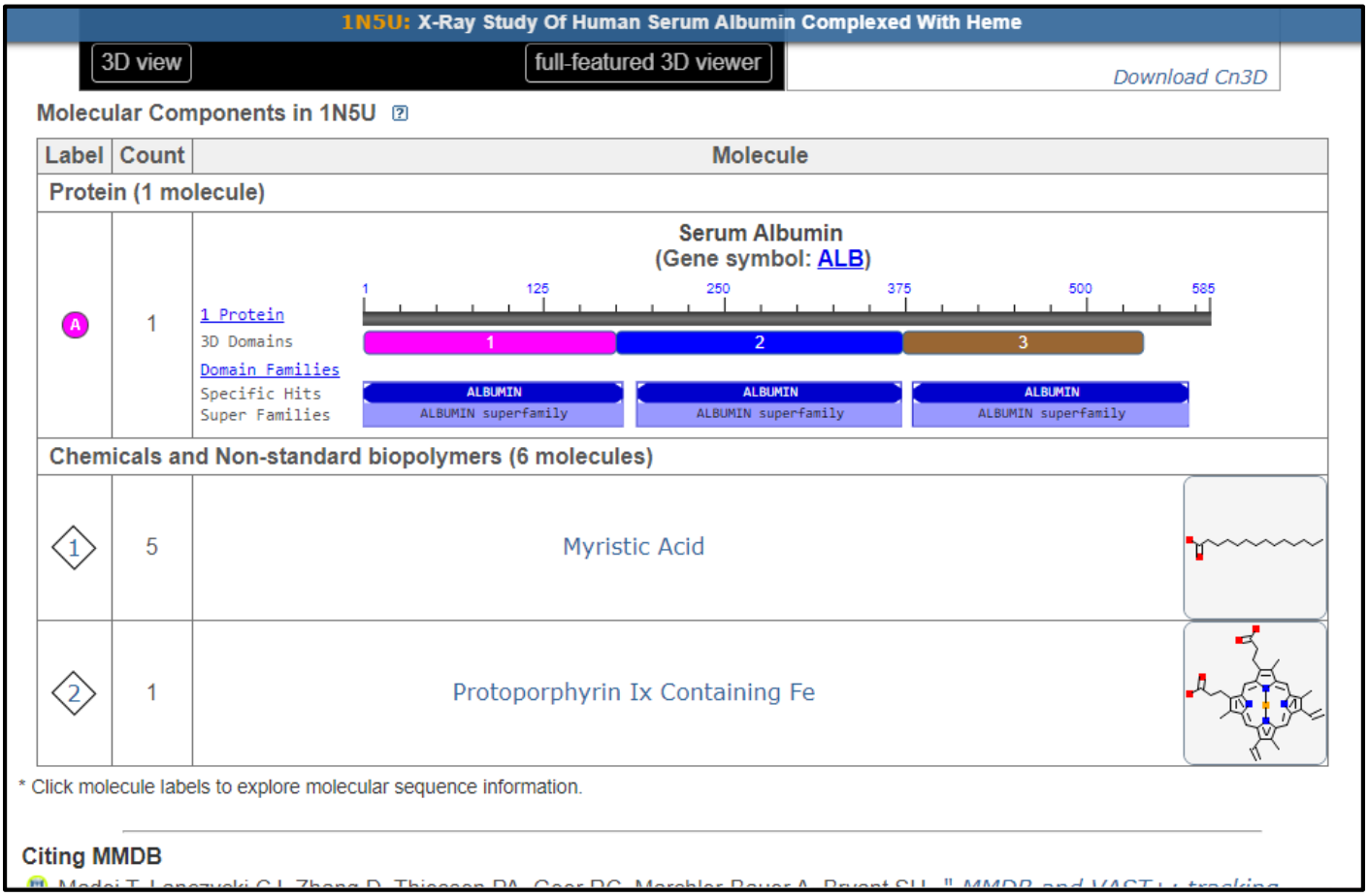


Fig5.2. Molecular components and chemicals and non-standard biopolymers for 1N5U

Fig6. Result page for aligned sequences

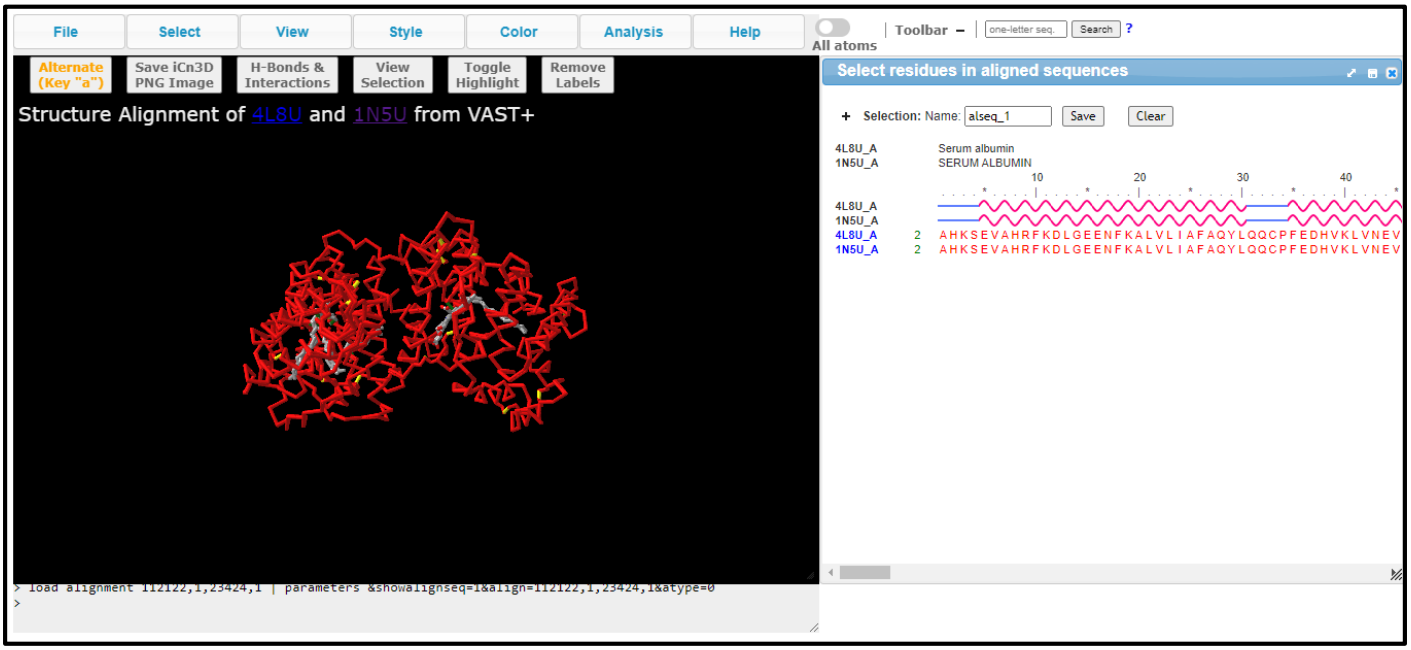


Fig7. Result page for 3D structure superimposition

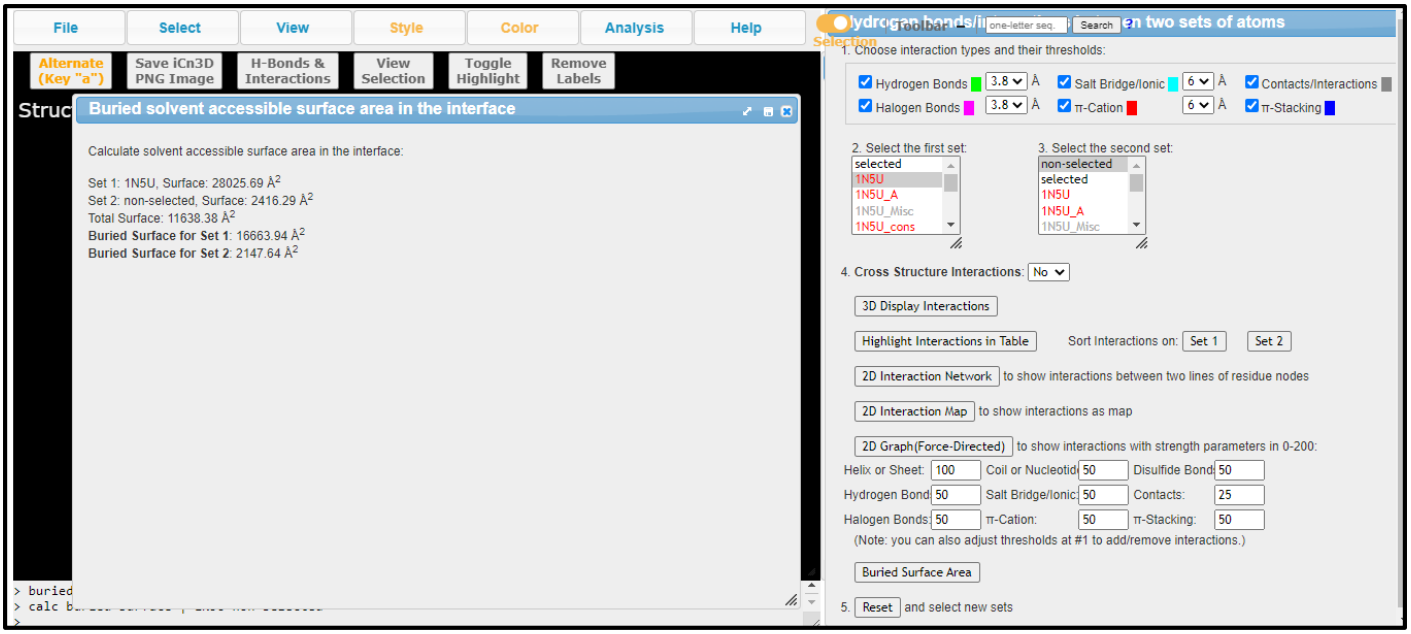


Fig7.1. Buried surface information

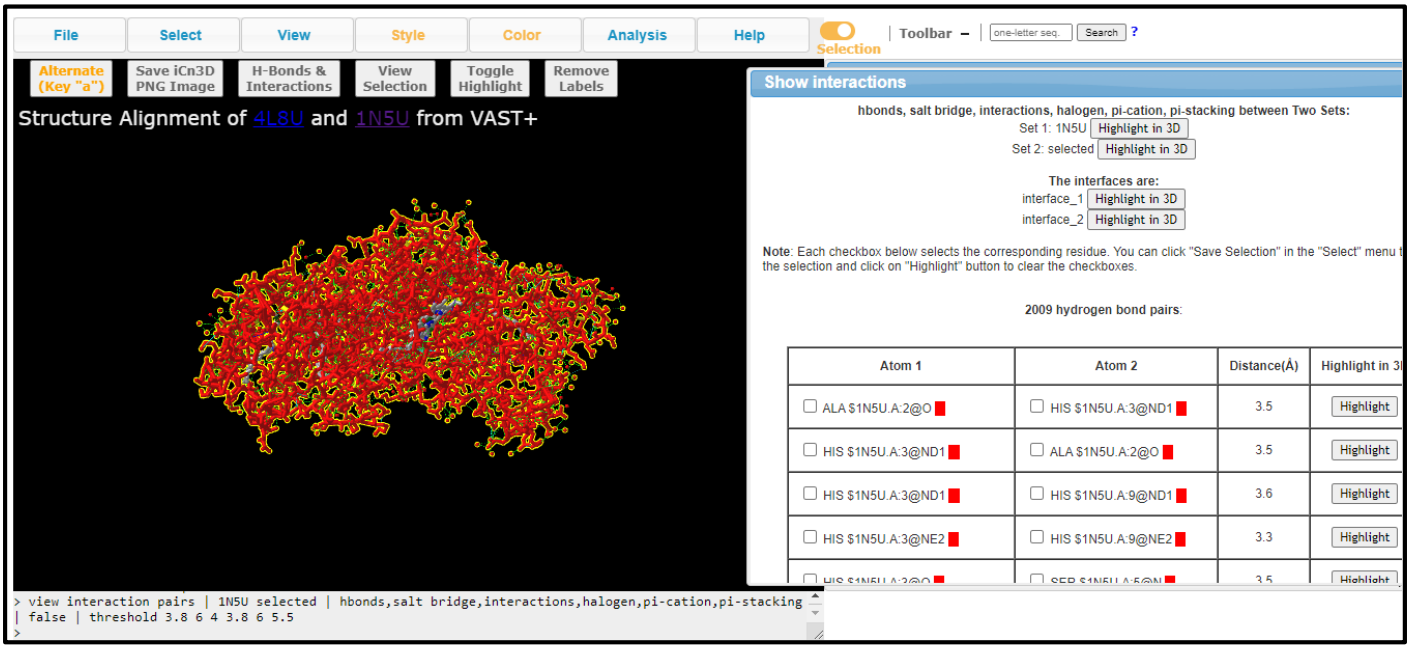


Fig7.2. Results for interactions

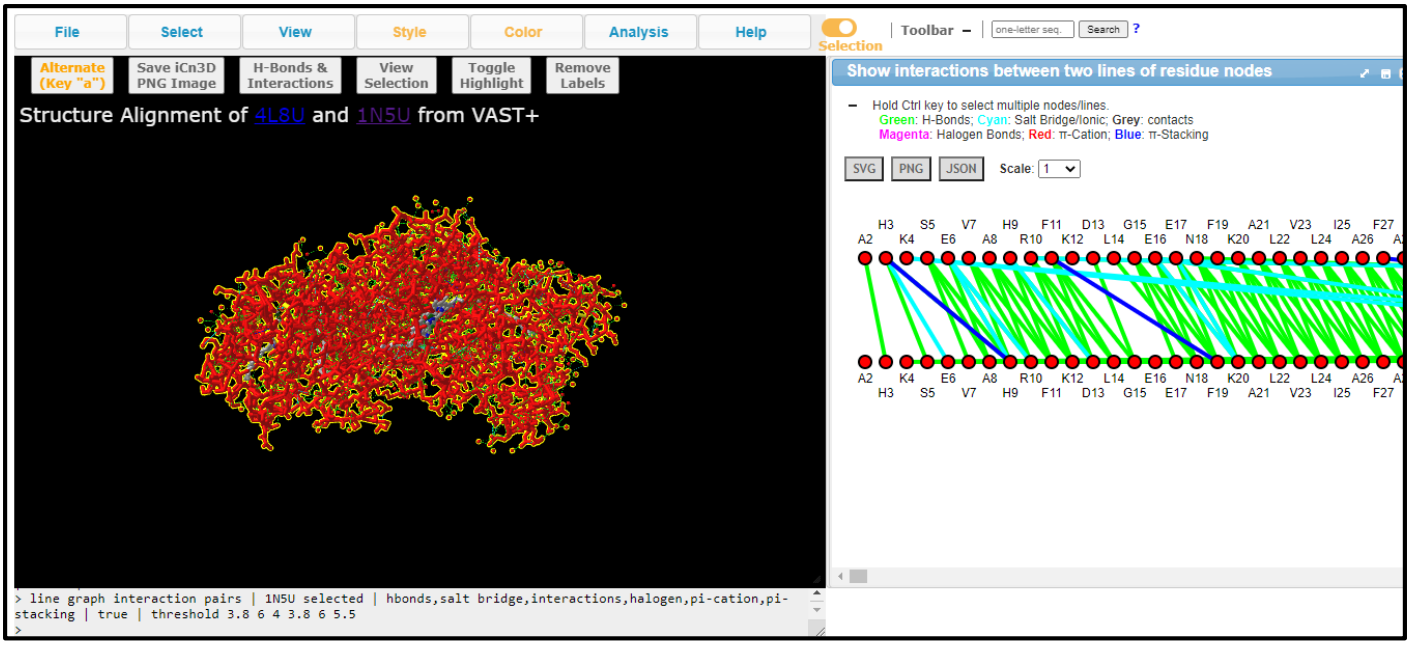


Fig7.3. Result for 2D interaction network

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*. Www.rcsb.org. Retrieved March 14, 2022, from https://www.rcsb.org/structure/4L8U
5. *4L8U: VAST+ Similar Structure Assemblies and Proteins*. (n.d.). Www.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.). Www.ncbi.nlm.nih.gov. Retrieved March 14, 2022, from https://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=1N5U
7. *iCn3D: Web-based 3D Structure Viewer*. (n.d.). Www.ncbi.nlm.nih.gov. Retrieved March 14, 2022, from https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html?showalignseq=1&align=112122

DATE: 14/03/22

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:

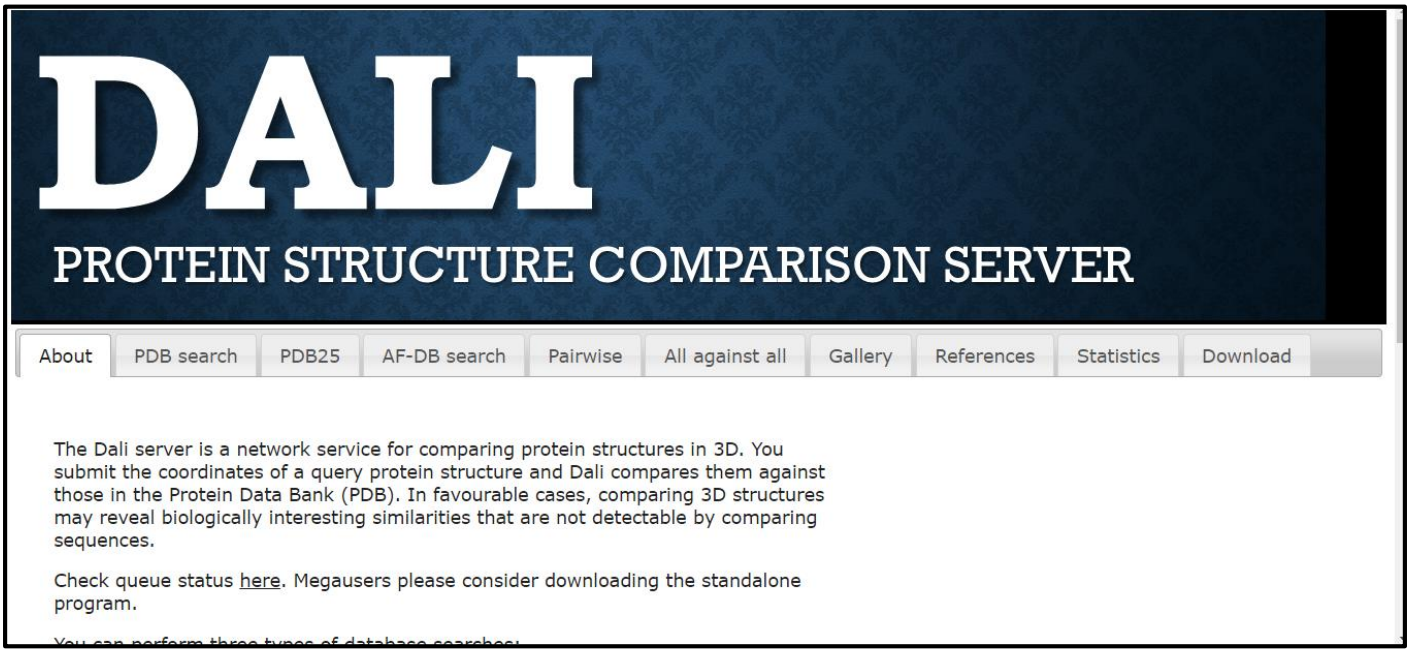


Fig1. Homepage for DALI

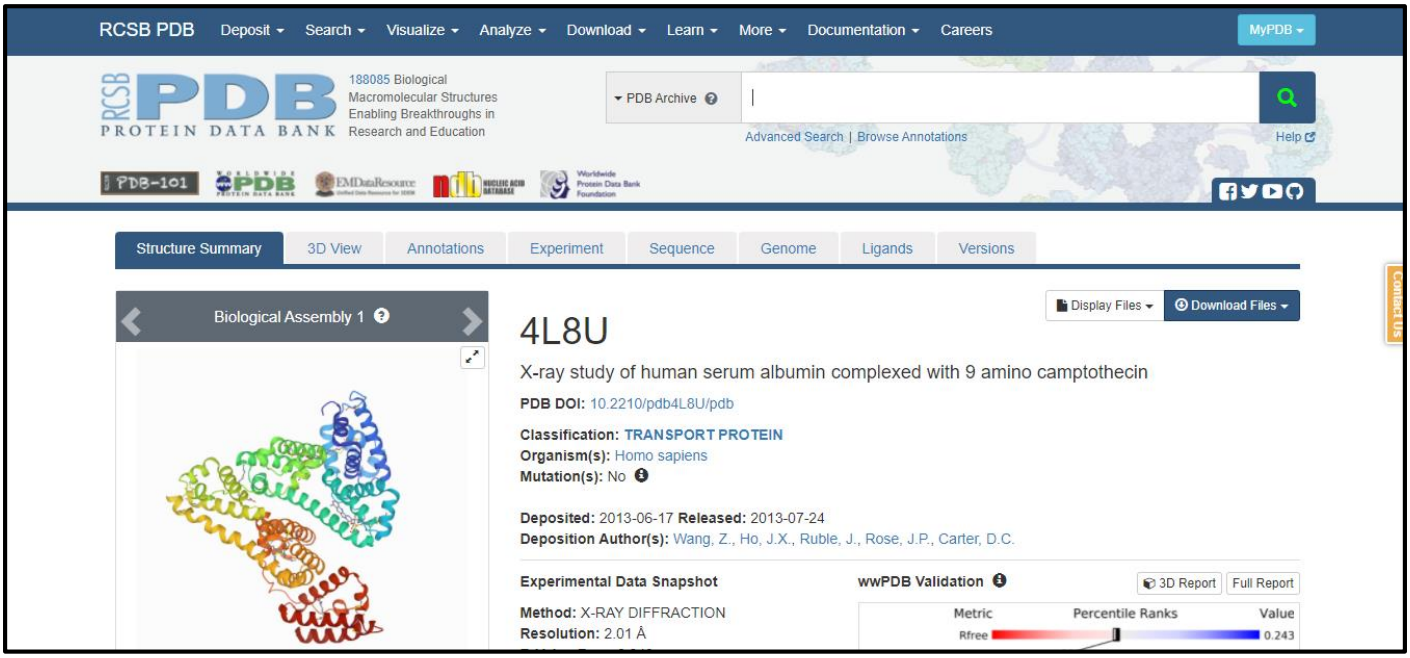


Fig2. Albumin PDB structure

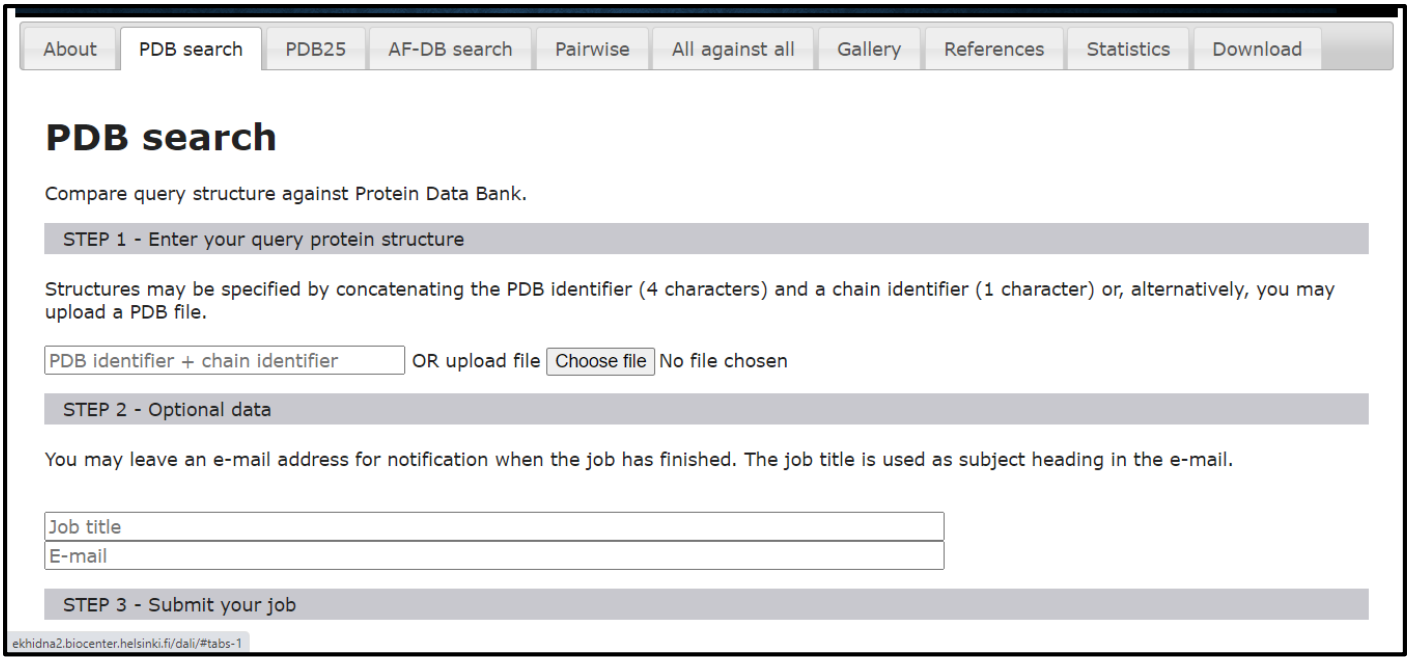


Fig3. PDB search

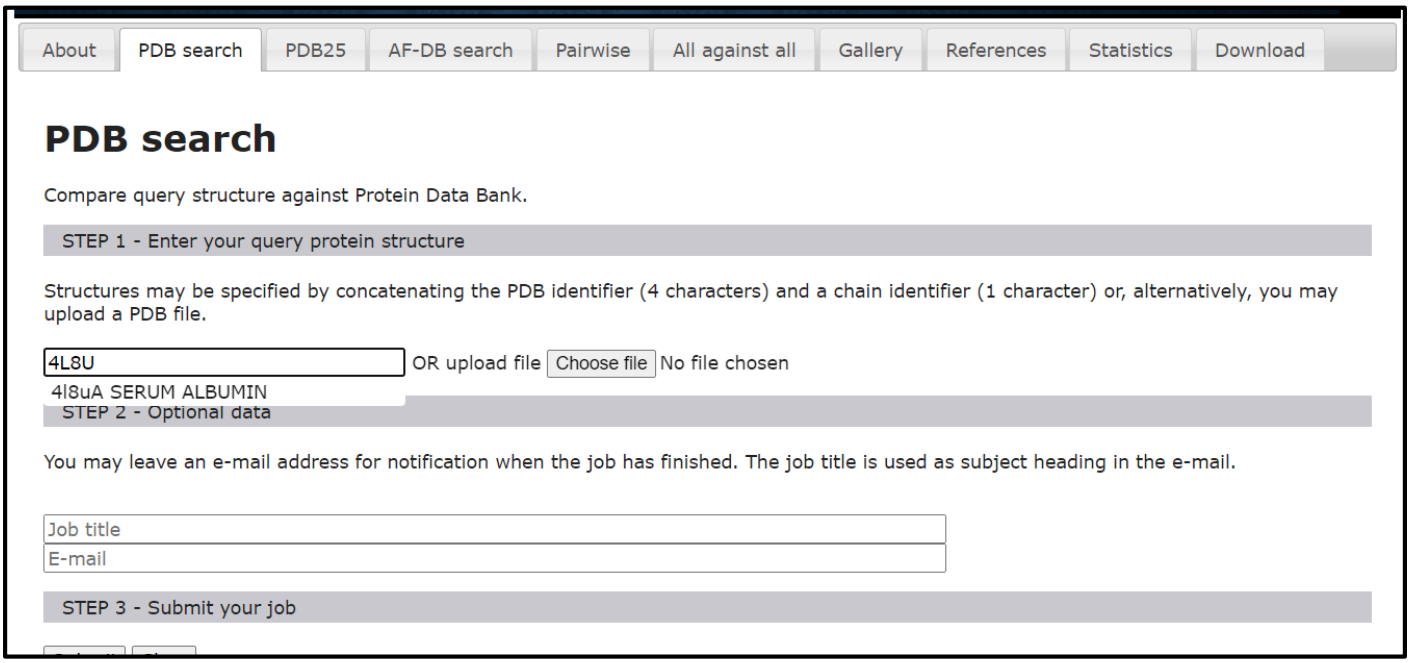


Fig3.1. PDB search for Albumin

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	PDB	MOLECULE: SERUM ALBUMIN;
2:	1kw2-A	17.2	13.0	262	455	20	PDB	MOLECULE: VITAMIN D-BINDING PROTEIN;
3:	5yje-B	4.7	5.6	91	318	13	PDB	MOLECULE: PROTEIN HIRA;
4:	7k0m-D	4.7	5.1	49	153	10	PDB	MOLECULE: SERINE PALMITOYLTRANSFERASE 1;
5:	2apl-A	4.6	7.5	75	149	11	PDB	MOLECULE: HYPOTHETICAL PROTEIN PG0816;
6:	4mk6-A	4.6	3.6	72	188	4	PDB	MOLECULE: PROBABLE DIHYDROXYACETONE KINASE REGULATOR DHSK_R
7:	2n5n-A	4.6	2.8	52	86	10	PDB	MOLECULE: CHROMODOMAIN-HELICASE-DNA-BINDING PROTEIN 4;
8:	2o7g-A	4.5	3.5	68	88	9	PDB	MOLECULE: PROBABLE RNA POLYMERASE SIGMA-C FACTOR;
9:	6cnb-R	4.4	11.6	106	522	8	PDB	MOLECULE: DNA-DIRECTED RNA POLYMERASE III SUBUNIT RPC1;
10:	3kdw-A	4.4	3.8	71	206	7	PDB	MOLECULE: PUTATIVE SUGAR BINDING PROTEIN;
11:	2lrm-A	4.4	2.9	67	84	13	PDB	MOLECULE: UNCHARACTERIZED PROTEIN YMGD;
12:	5g5p-B	4.3	3.5	90	455	3	PDB	MOLECULE: NUCLEAR MRNA EXPORT PROTEIN SAC3;
13:	3h36-A	4.2	3.1	52	78	12	PDB	MOLECULE: POLYRIBONUCLEOTIDE NUCLEOTIDYLTRANSFERASE;
14:	5w7p-A	4.1	10.0	79	397	14	PDB	MOLECULE: OXAC;
15:	4cc9-B	4.1	8.7	55	98	5	PDB	MOLECULE: PROTEIN VPRBP;
16:	6iac-A	4.0	5.8	73	301	3	PDB	MOLECULE: PORTAL PROTEIN;
17:	6wc2-A	4.0	4.5	69	154	10	PDB	MOLECULE: A-DNA;

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

```
DSSP  -LLLLHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHLLLLL
Query -AHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE   59
ident   |||| |||| |||||||  |||||| ||||||||  ||||| | ||||||||||||
Sbjct dTHKSEIAHRFNDLGEENFQGLVLIAFSQYLQQCPFDEHVKLVKELTEFAKTCVADESHA   60
DSSP  lLLLLHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHLLLLL


DSSP  LLLLLHHHHHHHHHL-LLLLHHHHhHHHHHHHHLLHHHHHHHHHLLLLLLLLLLLLLLl
Query NCDKSLHTLFGDKLC-TVATLREtYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPe  118
ident  |||||||||| ||         || ||||| |||||||||| |||| | || |
Sbjct GCDKSLHTLFGDELCkVATLRET-YGDMADCCEKQEPERNECFLNHKDDSPDLPKLKPE-  118
DSSP  LLLLLHHHHHHHHHHlLLLLLLL-LLLHHHHLLLLLLHHHHHHHLLLLLLLLLLLLL-


DSSP  HHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHLLLLLHHHHHH
Query VDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLL  178
ident  |  |  |    |  |  ||||| |||||||||||||  |  |    | ||||| || ||||
Sbjct PDTLCAEFKADEKKFWGKYLYEVARRHPYFYAPELLYYANKYNGVFQECCQAEDKGACLL  178
DSSP  HHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHLLLLLHHHHHH


DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHH
Query PKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLT  238
ident || |  |    |||| ||| ||| ||||||| ||| |||||| |||| |  | | |||||
Sbjct PKIDAMREKVLASSARQRLRCASIQKFGERALKAWSVARLSQKFPKADFTDVTKIVTDLT  238
DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHHHLHHHHHHHHHHHHHHLLLLLHHHHHHHHHH


DSSP  HHHHHHHLLHHHHHHHHHHHHHHHHHLHHHLLLLLHHHHHLLHHHHHHHHLLLLLLLL
Query KVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMP  298
```

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 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
```

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	PDB	MOLECULE: SERUM ALBUMIN;
2:	6fak-A	35.7	4.4	501	575	33	PDB	MOLECULE: AFAMIN;
3:	1kw2-A	17.2	13.0	262	455	20	PDB	MOLECULE: VITAMIN D-BINDING PROTEIN;
4:	5l3w-A	5.3	5.7	103	297	2	PDB	MOLECULE: SIGNAL RECOGNITION PARTICLE RECEPTOR FTSY;
5:	5yje-B	4.7	5.6	91	318	13	PDB	MOLECULE: PROTEIN HIRA;
6:	7k0m-D	4.7	5.1	49	153	10	PDB	MOLECULE: SERINE PALMITOYLTRANSFERASE 1;
7:	2apl-A	4.6	7.5	75	149	11	PDB	MOLECULE: HYPOTHETICAL PROTEIN PG0816;
8:	4mk6-A	4.6	3.6	72	188	4	PDB	MOLECULE: PROBABLE DIHYDROXYACETONE KINASE REGULATOR DHSK_R
9:	2n5n-A	4.6	2.8	52	86	10	PDB	MOLECULE: CHROMODOMAIN-HELICASE-DNA-BINDING PROTEIN 4;
10:	4he8-I	4.5	5.8	99	427	13	PDB	MOLECULE: NADH-QUINONE OXIDOREDUCTASE SUBUNIT 7;
11:	2o7g-A	4.5	3.5	68	88	9	PDB	MOLECULE: PROBABLE RNA POLYMERASE SIGMA-C FACTOR;
12:	4hfq-A	4.4	4.6	73	203	11	PDB	MOLECULE: MUTT/NUDIX FAMILY PROTEIN;
13:	6cnb-R	4.4	11.6	106	522	8	PDB	MOLECULE: DNA-DIRECTED RNA POLYMERASE III SUBUNIT RPC1;
14:	3kdw-A	4.4	3.8	71	206	7	PDB	MOLECULE: PUTATIVE SUGAR BINDING PROTEIN;
15:	2lrm-A	4.4	2.9	67	84	13	PDB	MOLECULE: UNCHARACTERIZED PROTEIN YMGD;
16:	2o4t-A	4.3	6.4	67	90	6	PDB	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  -LLLLHHHHHHHHHLHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHHLLLLL
Query -AHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE   59
ident   |||| |||| |||||||  |||||| ||||||||  ||||| | ||||||||||||
Sbjct dTHKSEIAHRFNDLGEENFQGLVLIAFSQYLQQCPFDEHVKLVKELTEFAKTCVADESHA   60
DSSP  lLLLLHHHHHHHHHLHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHHLLLLL


DSSP  LLLLLHHHHHHHHHL-LLLLHHHhHHHHHHHHLLHHHHHHHHHHLLLLLLLLLLLLLLl
Query NCDKSLHTLFGDKLC-TVATLREtYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPe  118
ident  |||||||||||  ||         || ||||| |||||||||||| |||| | || |
Sbjct GCDKSLHTLFGDELCkVATLRET-YGDMADCCEKQEPERNECFLNHKDDSPDLPKLKPE-  118
DSSP  LLLLLHHHHHHHHHHlLLLLLLLL-LLLHHHHLLLLLLHHHHHHHLLLLLLLLLLLLLL-


DSSP  HHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHLLLLLHHHHHH
Query VDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLL  178
ident  |  |  |   |  |  ||||| |||||||||||||  |  |   | ||||| || ||||
Sbjct PDTLCAEFKADEKKFWGKYLYEVARRHPYFYAPELLYYANKYNGVFQECCQAEDKGACLL  178
DSSP  HHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHLLLLLHHHHHH


DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHH
Query PKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLT  238
ident || |  |    |||| ||| ||| ||||||| ||| |||||| |||| |  | | |||||
Sbjct PKIDAMREKVLASSARQRLRCASIQKFGERALKAWSVARLSQKFPKADFTDVTKIVTDLT  238
DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHHHHLHHHHHHHHHHHHHHLLLLLHHHHHHHHHH
```

Fig6.1. 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 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	PDB	MOLECULE: SERUM ALBUMIN;
☐ 2:	4f5t-A	42.0	3.8	581	583	76	PDB	MOLECULE: SERUM ALBUMIN;
☐ 3:	5yxe-A	40.2	4.5	572	572	82	PDB	MOLECULE: SERUM ALBUMIN;
☐ 4:	4f5s-B	39.0	4.9	581	583	76	PDB	MOLECULE: SERUM ALBUMIN;
☐ 5:	6m58-B	38.9	4.3	492	550	95	PDB	MOLECULE: SERUM ALBUMIN;
☐ 6:	6zl1-A	38.3	3.6	553	555	98	PDB	MOLECULE: ALBUMIN;
☐ 7:	5orf-A	38.1	4.4	581	586	73	PDB	MOLECULE: SERUM ALBUMIN;
☐ 8:	4f5v-A	37.9	5.1	581	584	73	PDB	MOLECULE: SERUM ALBUMIN;
☐ 9:	6fak-A	35.7	4.4	501	575	33	PDB	MOLECULE: AFAMIN;
☐ 10:	5ghk-A	35.0	4.1	542	563	80	PDB	MOLECULE: SERUM ALBUMIN;
☐ 11:	6rq7-B	31.8	4.9	473	479	36	PDB	MOLECULE: AFAMIN;
☐ 12:	1kw2-A	17.2	13.0	262	455	20	PDB	MOLECULE: VITAMIN D-BINDING PROTEIN;
☐ 13:	1lot-A	15.8	8.5	237	436	15	PDB	MOLECULE: VITAMIN D-BINDING PROTEIN;
☐ 14:	6pv7-A	5.3	8.9	92	387	10	PDB	MOLECULE: FUSION PROTEIN OF NEURONAL ACETYLCHOLINE RECEPTOR
☐ 15:	5l3w-A	5.3	5.7	103	297	2	PDB	MOLECULE: SIGNAL RECOGNITION PARTICLE RECEPTOR FTSY;
☐ 16:	5yje-B	4.7	5.6	91	318	13	PDB	MOLECULE: PROTEIN HIRA;
☐ 17:	7k0m-D	4.7	5.1	49	153	10	PDB	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

```
DSSP  lLLLHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHLLLLLL
Query aHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAEN   60
ident  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct -HKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAEN   59
DSSP  -LLLHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHLLLLLL


DSSP  LLLLHHHHHHHHHLLLLLHHHHHHHHHHHHHLLHHHHHHHHLLLLLLLLLLLLLLHH
Query CDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVD  120
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct CDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVD  119
DSSP  LLLLHHHHHHHHHHHLLLHHHHHHHHHHHHLLLLLHHHHHHLLLLLLLLLLLLLLLHH


DSSP  HHHHHHHLHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHLLLLLHHHHHHH
Query VMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPK  180
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct VMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPK  179
DSSP  HHHHHHHLHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHLLLLLHHHHHHH


DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHH
Query LDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKV  240
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct LDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKV  239
DSSP  HHHHHHHHHHHHHHHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHH


DSSP  HHHHHLLHHHHHHHHHHHHHHHHHLHHHLLLLLHHHHLLHHHHHHLLLLLLLLL
```

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  206  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  399  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:	4l8u-A	56.2	0.0	583	583	100	PDB	MOLECULE: SERUM ALBUMIN;
2:	1n5u-A	53.4	0.4	583	583	100	PDB	MOLECULE: SERUM ALBUMIN;
3:	6hsc-B	52.4	0.7	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
4:	4z69-I	51.5	0.6	581	581	100	PDB	MOLECULE: SERUM ALBUMIN;
5:	7vr9-B	51.2	0.8	578	578	99	PDB	MOLECULE: SERUM ALBUMIN;
6:	6wuw-A	51.1	0.7	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
7:	6wuw-B	51.1	0.9	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
8:	3a73-A	50.9	1.0	576	576	100	PDB	MOLECULE: SERUM ALBUMIN;
9:	3uiv-A	50.6	0.6	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
10:	2bxm-A	50.5	0.7	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
11:	5id7-B	50.5	1.3	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
12:	1e7f-A	50.0	0.8	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
13:	3cx9-A	50.0	0.6	581	581	100	PDB	MOLECULE: SERUM ALBUMIN;
14:	2i30-A	49.6	0.8	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
15:	5vnw-A	49.5	1.7	583	583	100	PDB	MOLECULE: SERUM ALBUMIN;
16:	1e7g-A	49.5	0.9	582	582	100	PDB	MOLECULE: SERUM ALBUMIN;
17:	1bke-A	49.4	0.7	581	581	100	PDB	MOLECULE: SERUM ALBUMIN;

Fig8. Result for similar structures

No 2: Query=4l8uA Sbjct=1n5uA Z-score=53.4

back to top

```
DSSP  LLLLHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHLLLLLL
Query AHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAEN   60
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct AHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAEN   60
DSSP  LLLLHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHHLLLLLL


DSSP  LLLLHHHHHHHHHLLLLLHHHHHHHHHHHHHLLHHHHHHHHHLLLLLLLLLLLLLLHH
Query CDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVD  120
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct CDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVD  120
DSSP  LLLLHHHHHHHHHLLLLHHHHHHHHHHHHHLLHHHHHHHHHLLLLLLLLLLLLLLHH


DSSP  HHHHHHHLHHHHHHHHHHHHHHHLLLLLLHHHHHHHHHHHHHHHHHHLLLLLHHHHHHH
Query VMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPK  180
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct VMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPK  180
DSSP  HHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHHHHHHHHHLLLLLHHHHHHH


DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHH
Query LDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKV  240
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct LDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKV  240
DSSP  HHHHHHHHHHHHHHHHHHHHHHHHHLHHHHHHHHHHHHHHHLLLLLHHHHHHHHHHH


DSSP  HHHHHHLLHHHHHHHHHHHHHHHHHLHHHLLLLLHHHHLLHHHHHHHHLLLLLLLLL
Query HTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPAD  300
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct HTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPAD  300
DSSP  HHHHHHLLHHHHHHHHHHHHHHHHLLHHHLLLLLHHHHLLHHHHHHHHLLLLLLLLL
```

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*. Www.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/

DATE: 18/03/22

WEBLEM 8

Introduction to Gene Prediction and various elements in Prokaryotes and Eukaryotes

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

CATEGORIES OF GENE PREDICTION PROGRAMS

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

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

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

Gene Prediction Using Markov Models and Hidden Markov Models

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

The use of Markov models in gene finding exploits the fact that oligonucleotide distributions in the coding regions are different from those for the noncoding regions. These can be represented with various orders of Markov models. Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous *k* nucleotides, the longer the oligomer unit, the more non randomness can be 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 Vertibi algorithm to find an optimal match

for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to further distinguish coding signals from noncoding signals.

Step-by-Step Description of FGENESB annotation.

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

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

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

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

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

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

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

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

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

Prediction Using Discriminant Analysis.

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

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

Output format:

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

- ORF - start/end positions where the first complete codon starts and the last codon ends

An issue related to gene prediction is promoter prediction. Promoters are DNA elements located in the vicinity of gene start sites (which should not be confused with the translation start sites) and serve as binding sites for the gene transcription machinery, consisting of RNA polymerases and transcription factors. Therefore, these DNA elements directly regulate gene expression. Promoters and regulatory elements are traditionally determined by experimental analysis. The process is extremely time consuming and laborious. Computational prediction of promoters and regulatory elements is especially promising because it has the potential to replace a great deal of extensive experimental analysis.

PREDICTION ALGORITHMS

Current algorithms for predicting promoters and regulatory elements can be categorized as either ab initio based, which make de novo predictions by scanning individual sequences; or similarity based, which make predictions based on alignment of homologous sequences; or expression profile based using profiles constructed from a number of 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 programfirst predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than theWang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

Output format:

- First line - name of your sequence;
- Second and Third lines - LDF threshold and the length of presented sequence
- 4th line - The number of predicted promoters
- Next lines - positions of predicted promoters, and their scores with 'weights' of two conserved promoter boxes. Promoter position assign to the first nucleotide of the transcript (Transcription Start Site position).

- After that we present elements of Transcriptional factor binding sites for each predicted promoter (if they found).

Prediction for Eukaryotes

The ab initio method for predicting eukaryotic promoters and regulatory elements also relies on searching the input sequences for matching of consensus patterns of known promoters and regulatory elements. The consensus patterns are derived from experimentally determined DNA binding sites which are compiled into profiles and stored in a database for scanning an unknown sequence to find similar conserved patterns. However, this approach tends to generate very high rate of false positives owing to nonspecific matches with the short sequence patterns. Furthermore, because of the high variability of transcription factor binding sites, the simple sequence matching often misses true promoter sites, creating false negatives.

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

The eukaryotic transcription initiation requires cooperation of a large number of transcription factors. *Cooperativity* means that the promoter regions tend to contain a high density of protein-binding sites. Thus, finding a cluster of transcription factor binding sites often enhances the probability of individual binding site prediction.

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

Output format:

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

ORF finder:

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

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

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

REFERENCES:

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

DATE: 18/03/22

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 to recognized using TSSW.

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

METHODOLOGY:

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

OBSERVATION:

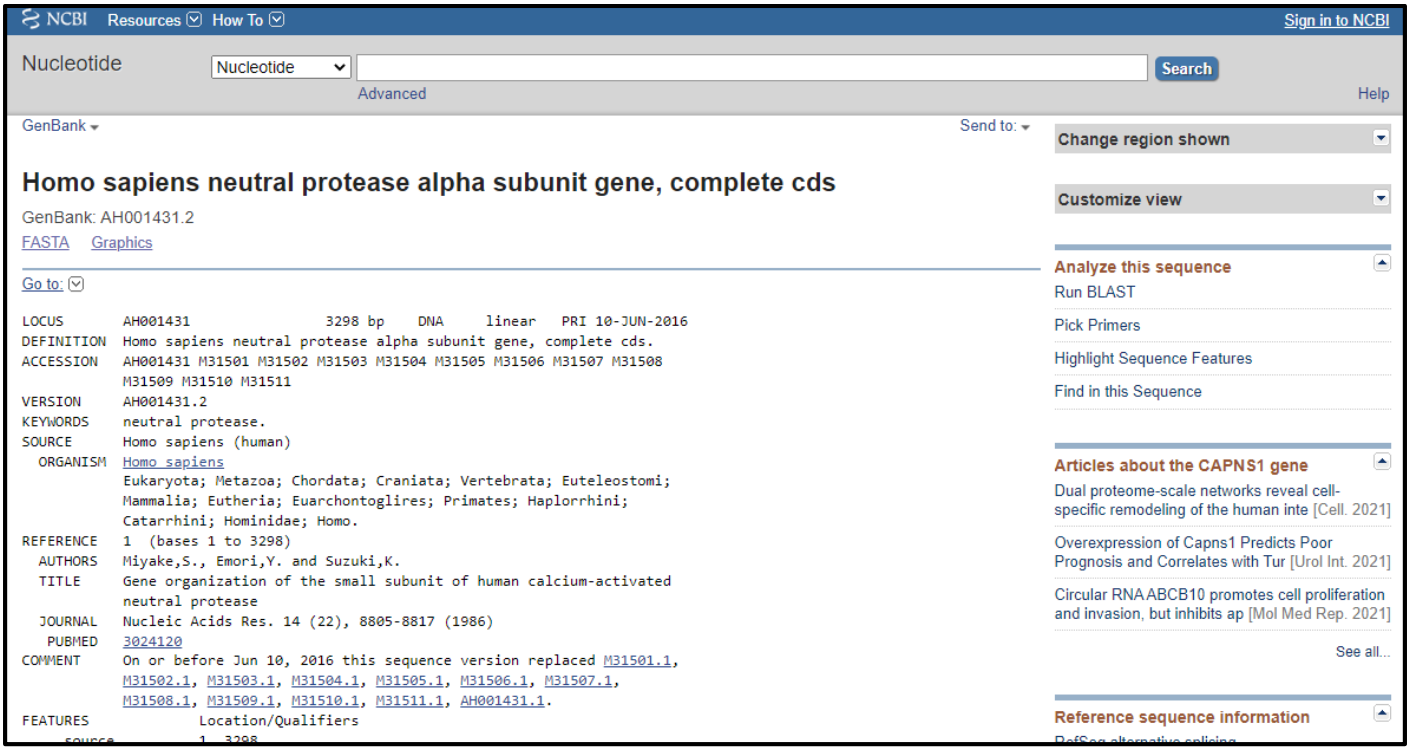


Fig1. GenBank result for Human Protease

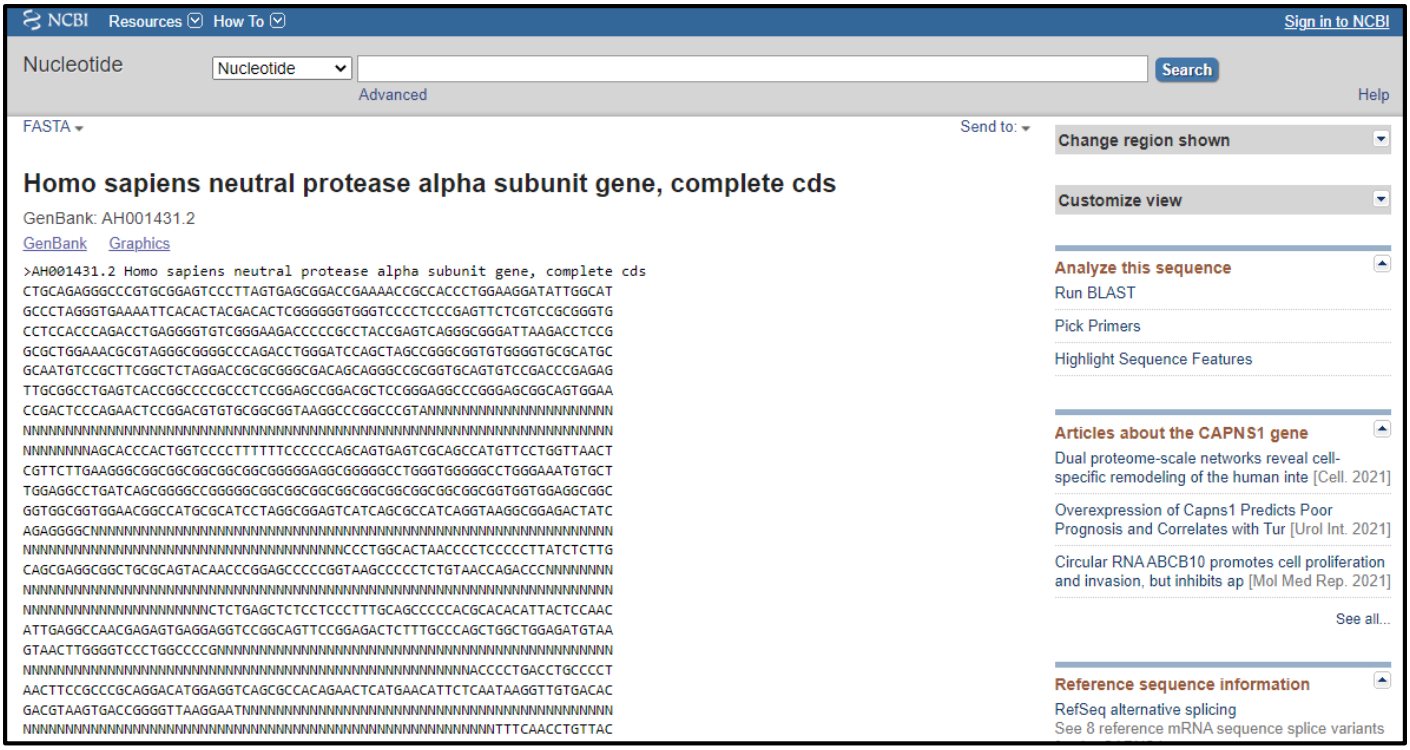


Fig2. Nucleotide FASTA sequence for Protease

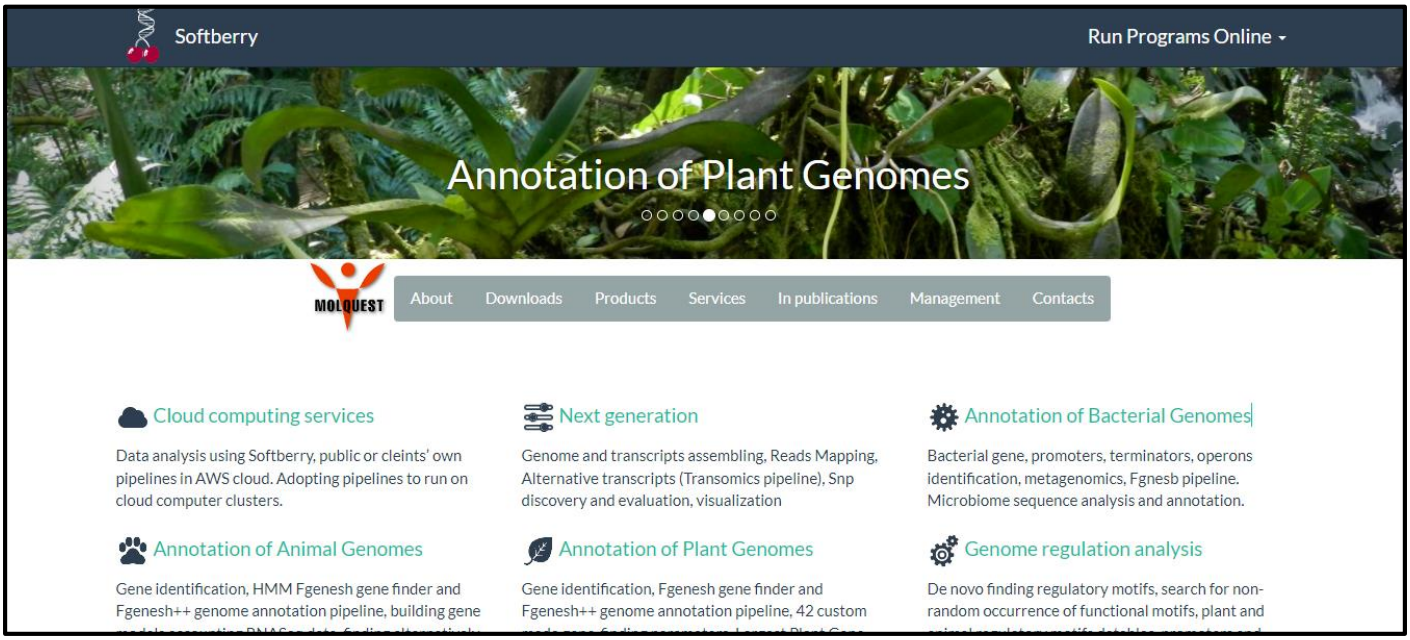


Fig3. Homepage for Softberry

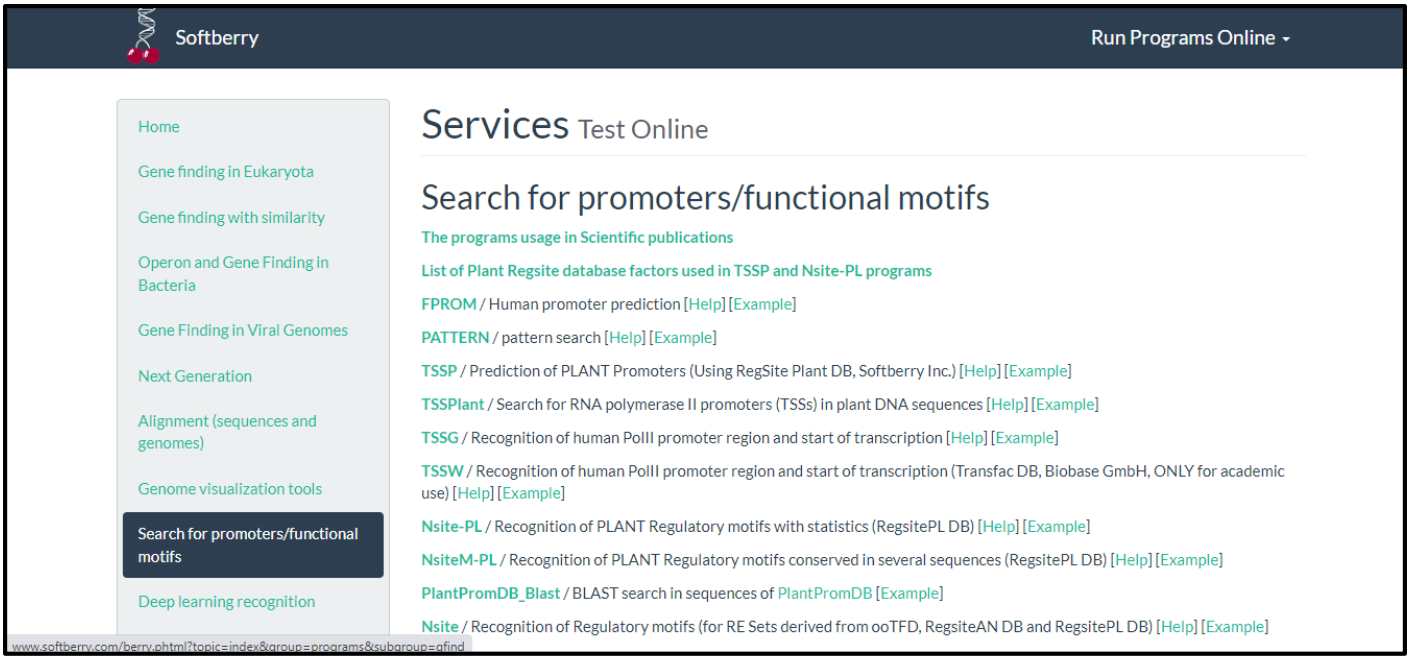


Fig4. Tools for promoters/functional motifs

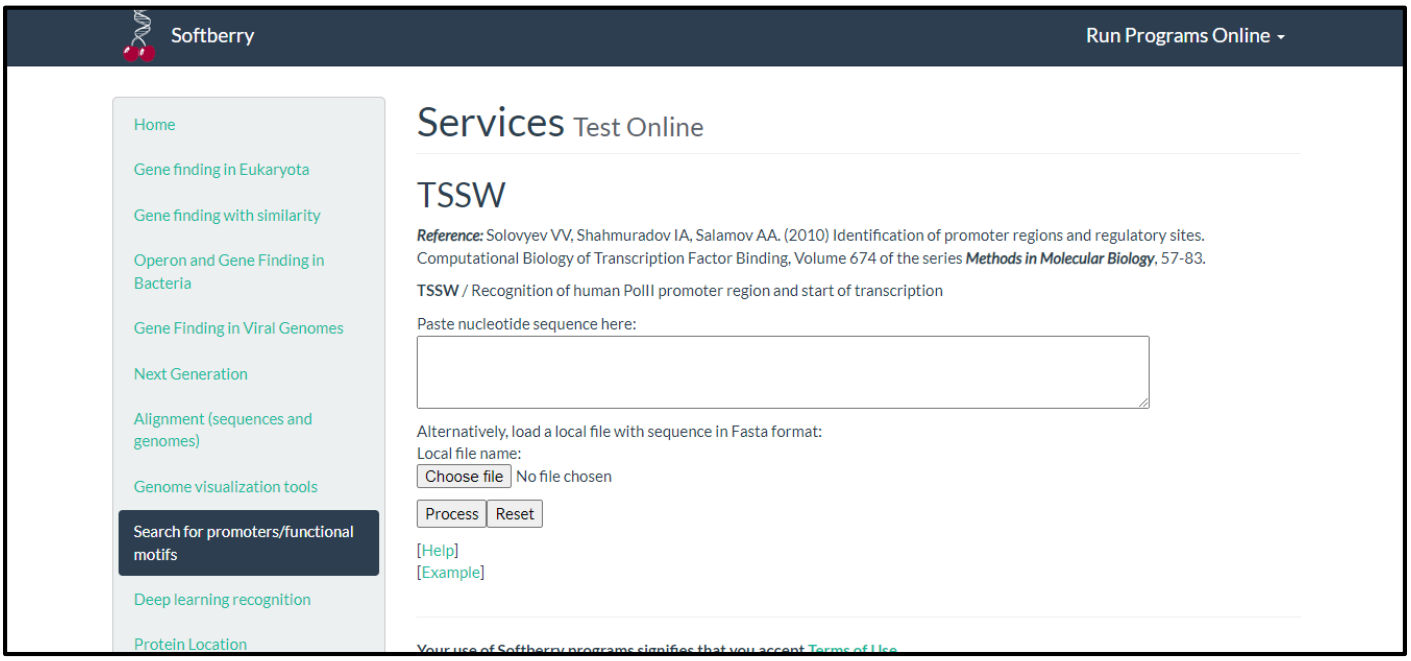


Fig5. Homepage for TSSW

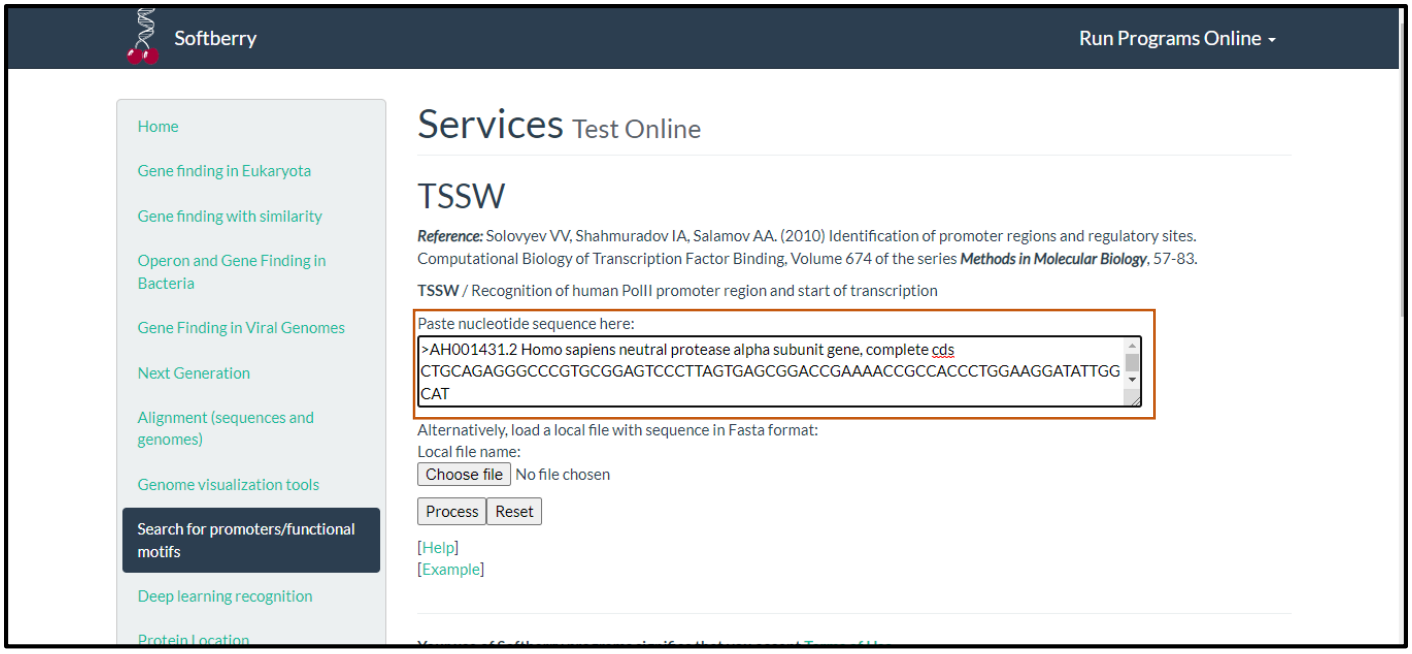


Fig6. Search for protease nucleotide FASTA sequence

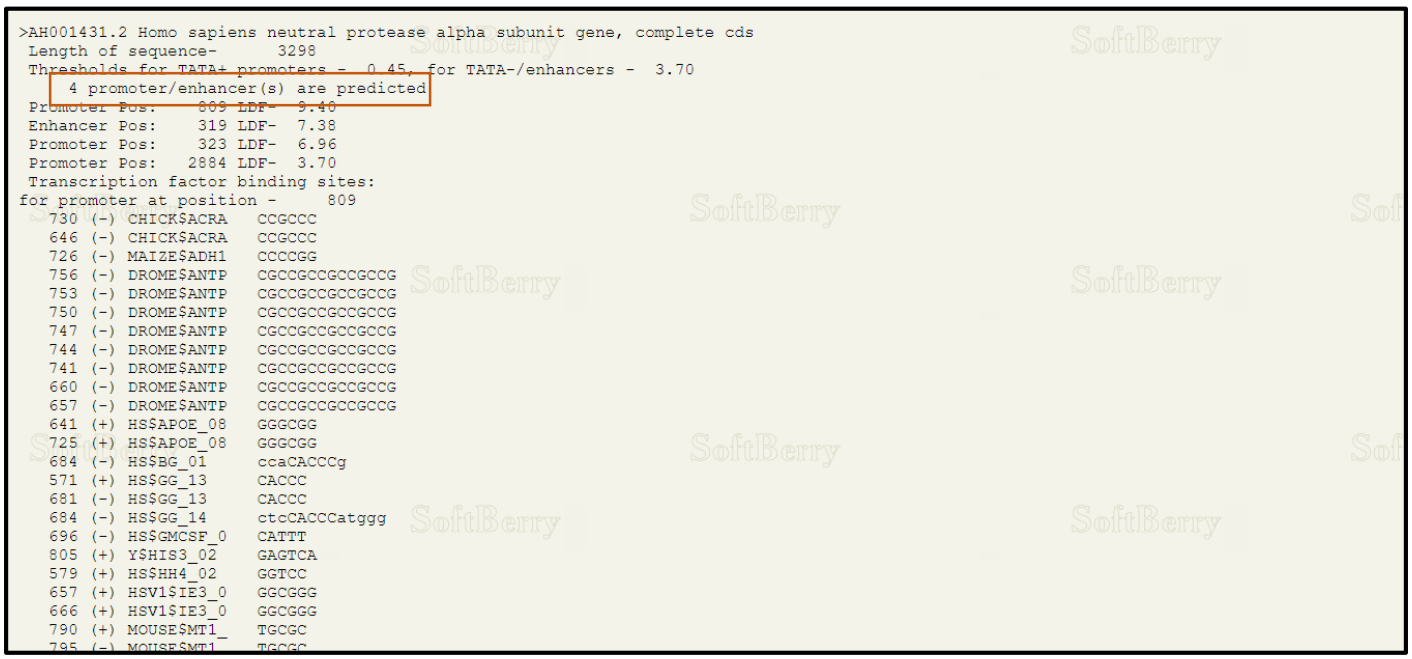


Fig7. Results for recognised promoter regions

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.). Www.webmd.com. Retrieved March 18, 2022, from https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20(proteases)%20are%20enzymes
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.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
12. *TSSW - Recognition of human PolII promoter region and start of transcription*. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter
13. *Softberry - TSSW result*. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/promoter/tssw.pl

DATE: 18/03/22

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:

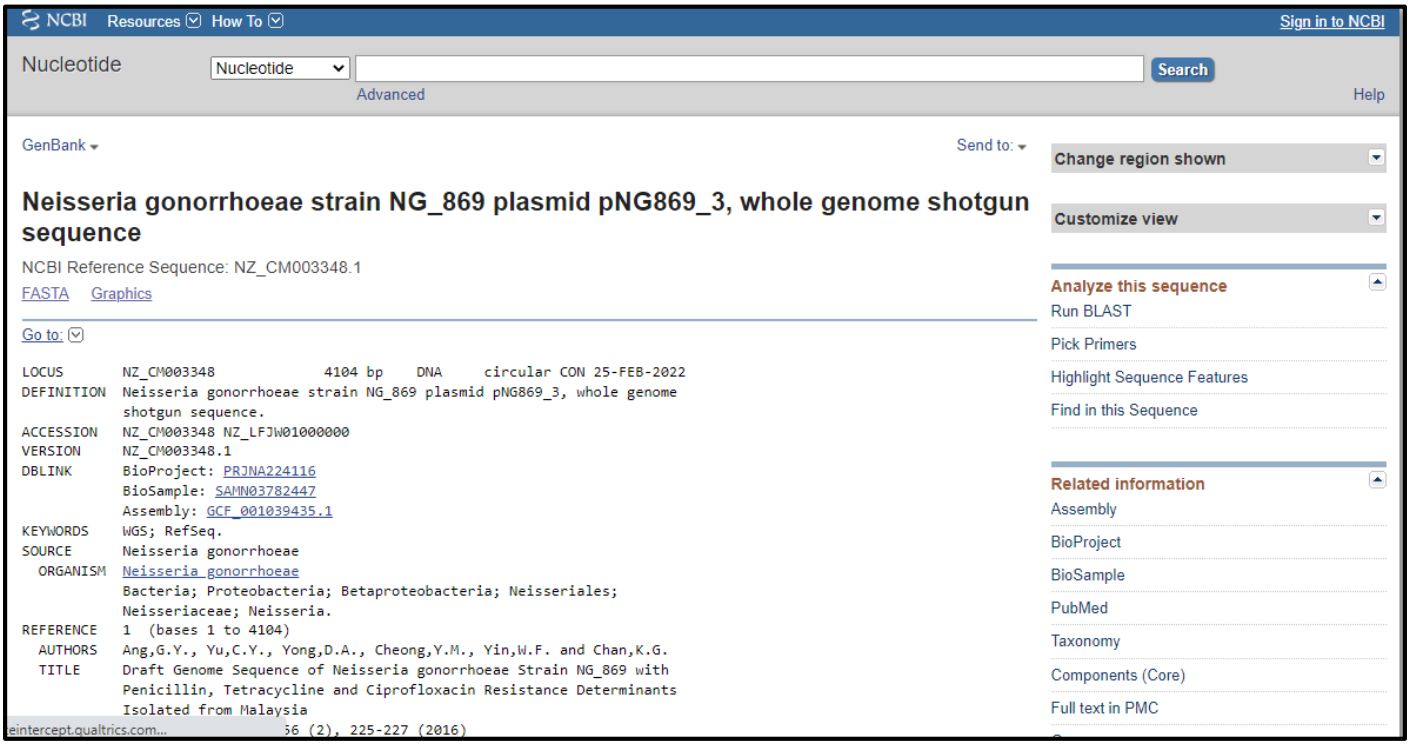


Fig1. GenBank result for *Neisseria gonorrhoeae*

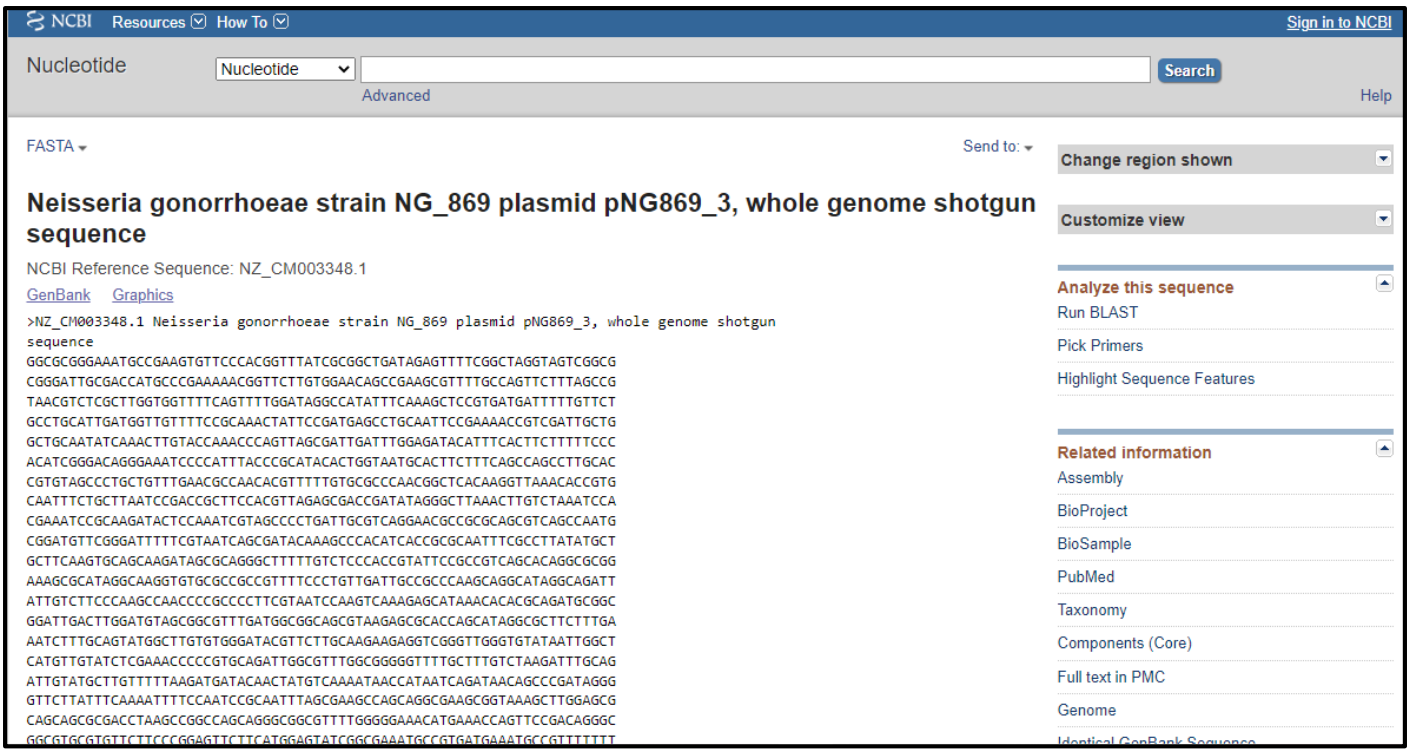


Fig2. Nucleotide FASTA sequence

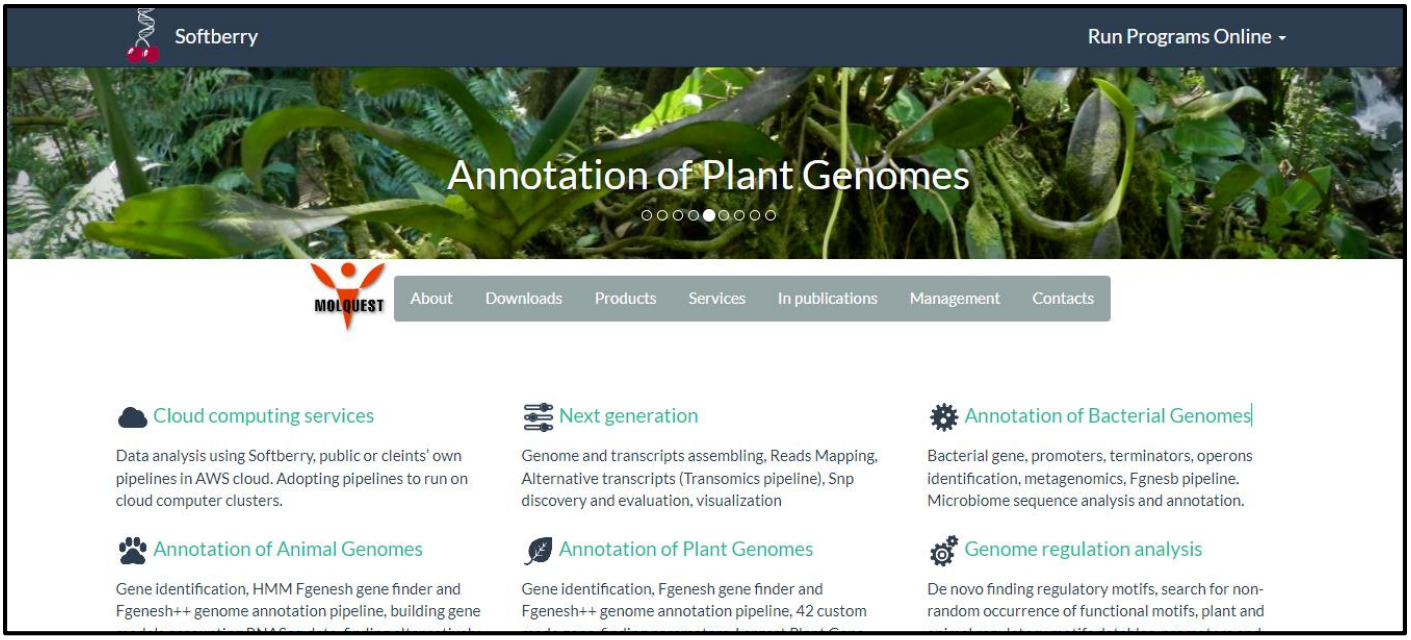


Fig3. Homepage for Softberry

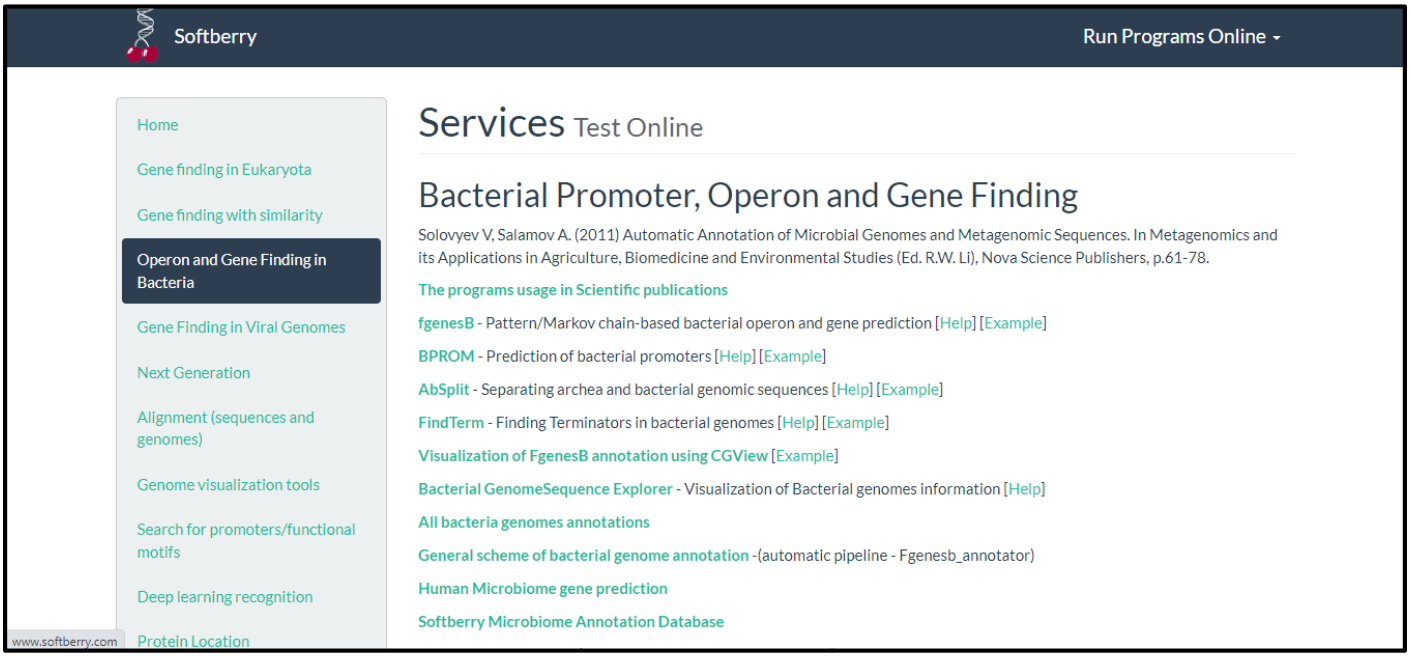


Fig4. Tools for bacterial promoter, operon and gene finding

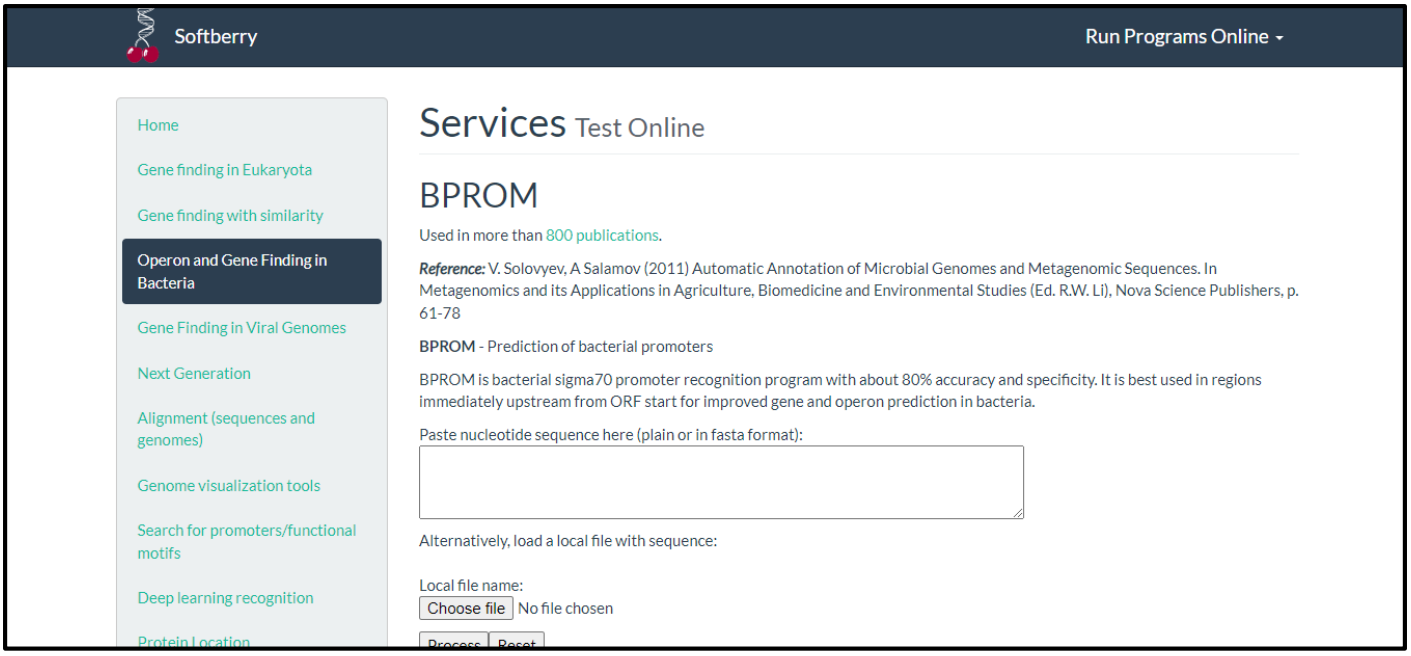


Fig5. Homepage for BPROM

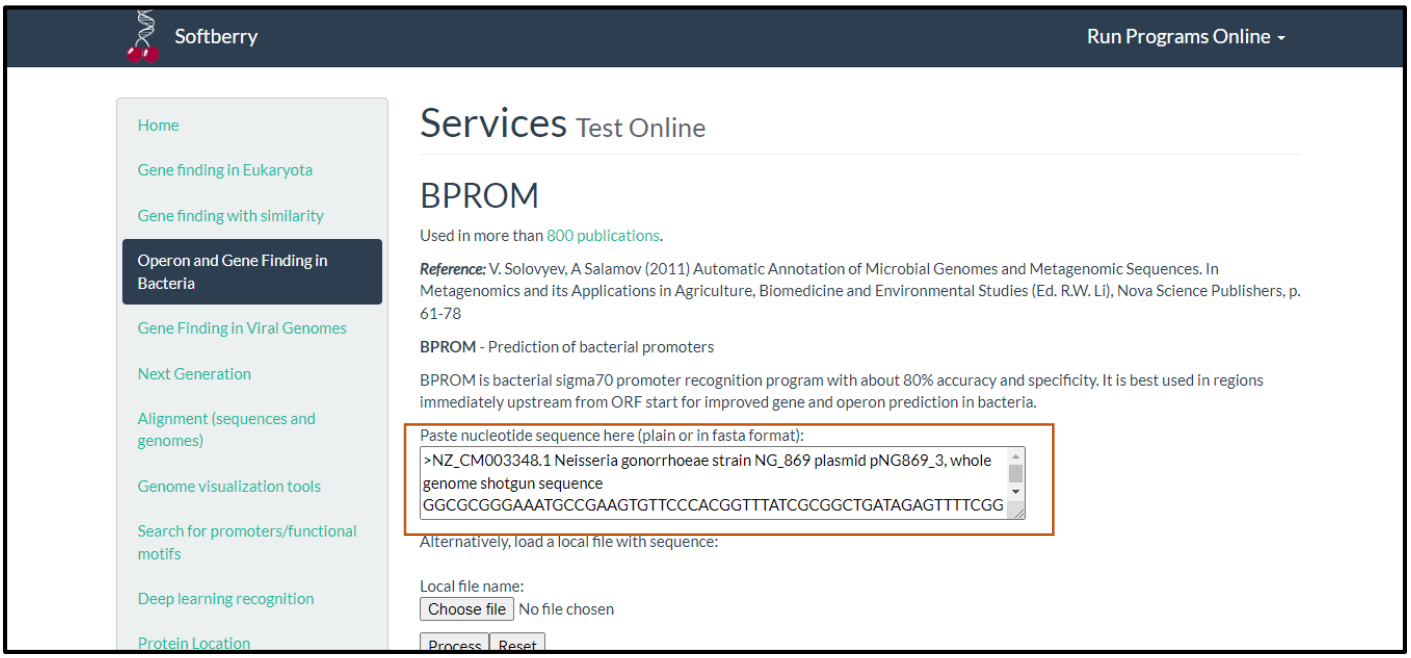


Fig6. Search for nucleotide FASTA sequence

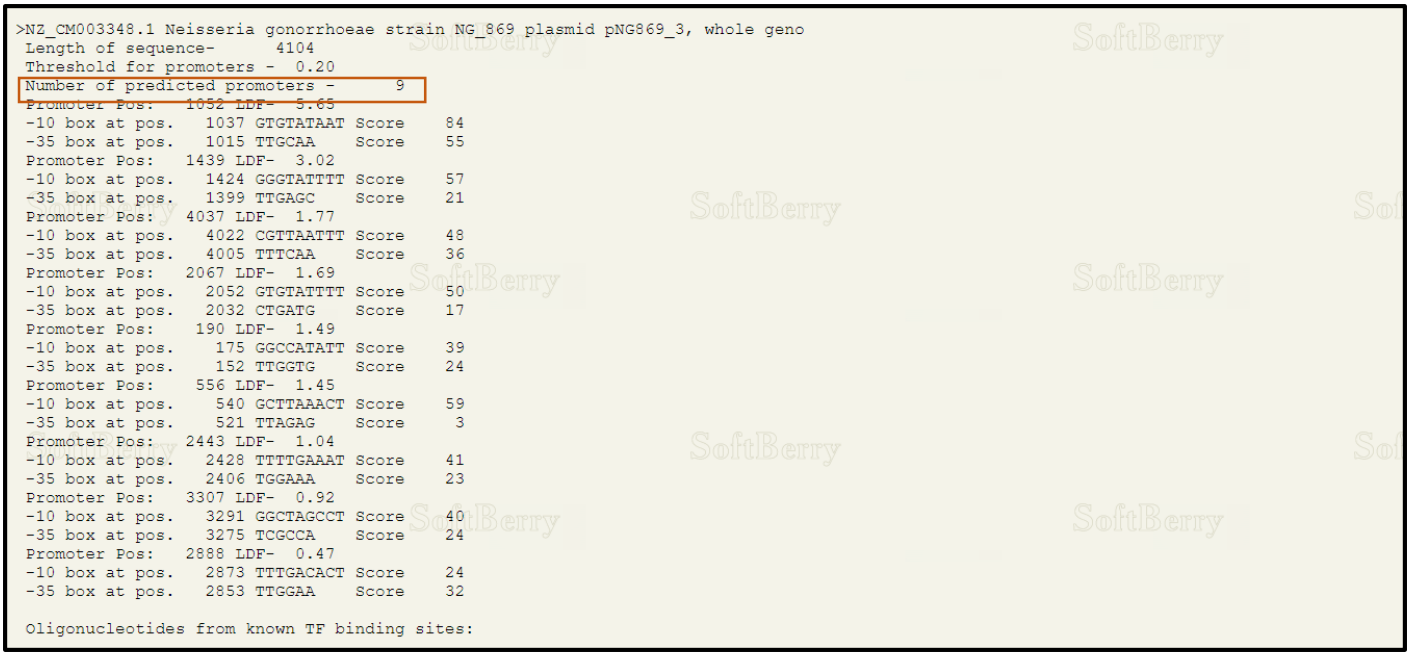


Fig7. Results for predicted promoters

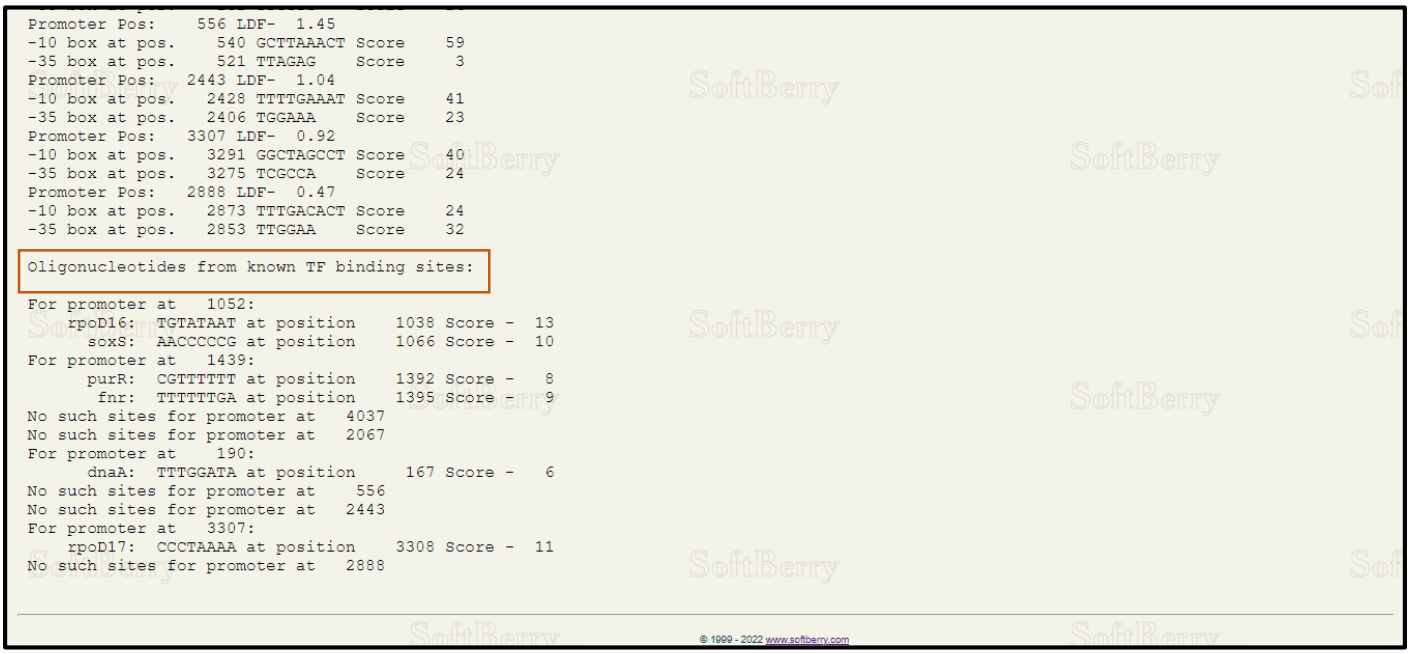


Fig8. Results for known TF binding sites

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. Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
3. Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome shotgun sequence. (2022). *NCBI Nucleotide*. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1
4. *Softberry Home Page*. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
5. *BPROM - Prediction of bacterial promoters*. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb
6. *Softberry - BPROM result*. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfindb/bprom.pl

DATE: 18/03/22

WEBLEM 8c

FGENESB

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

AIM:

To predict bacterial operon and gene for *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 Vertibi algorithm to find an optimal match for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to further distinguish coding signals from noncoding signals.

METHODOLOGY:

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:

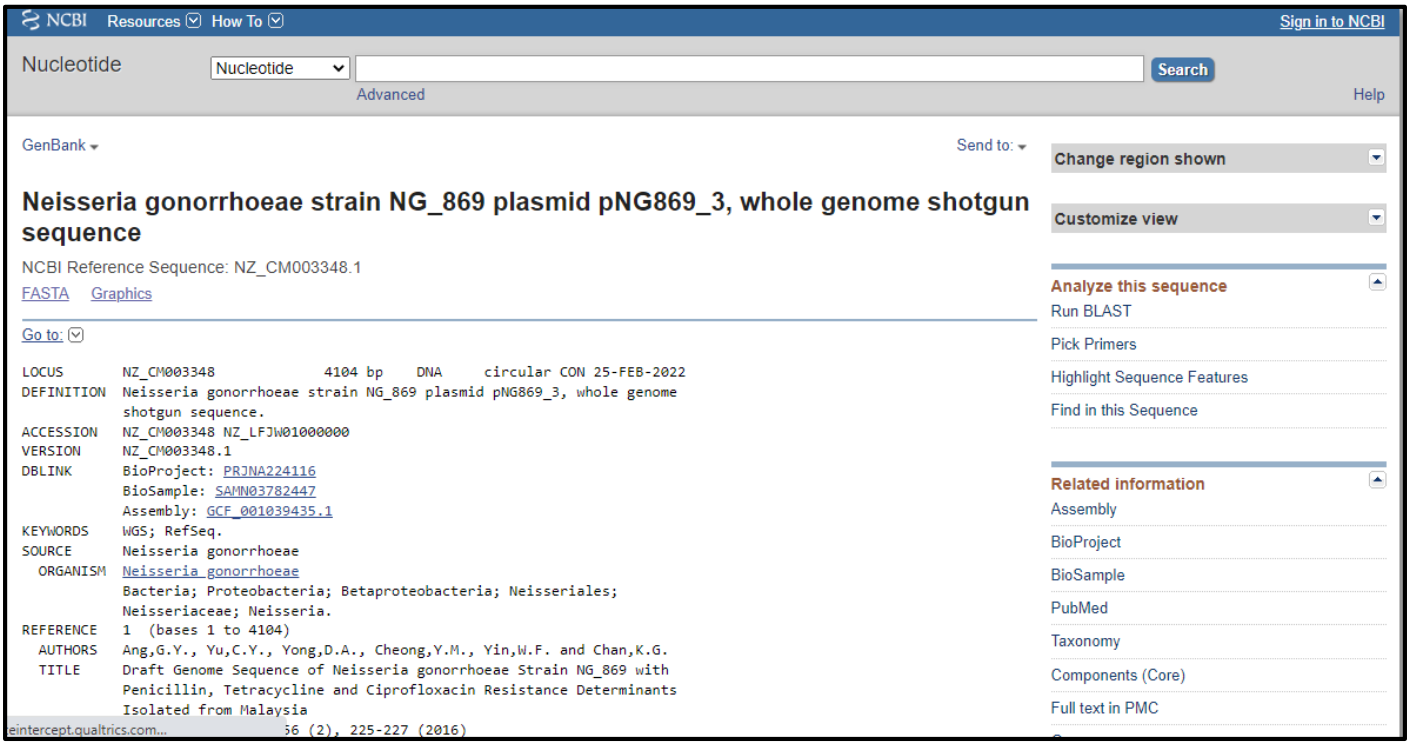


Fig1. GenBank result for *Neisseria gonorrhoeae*

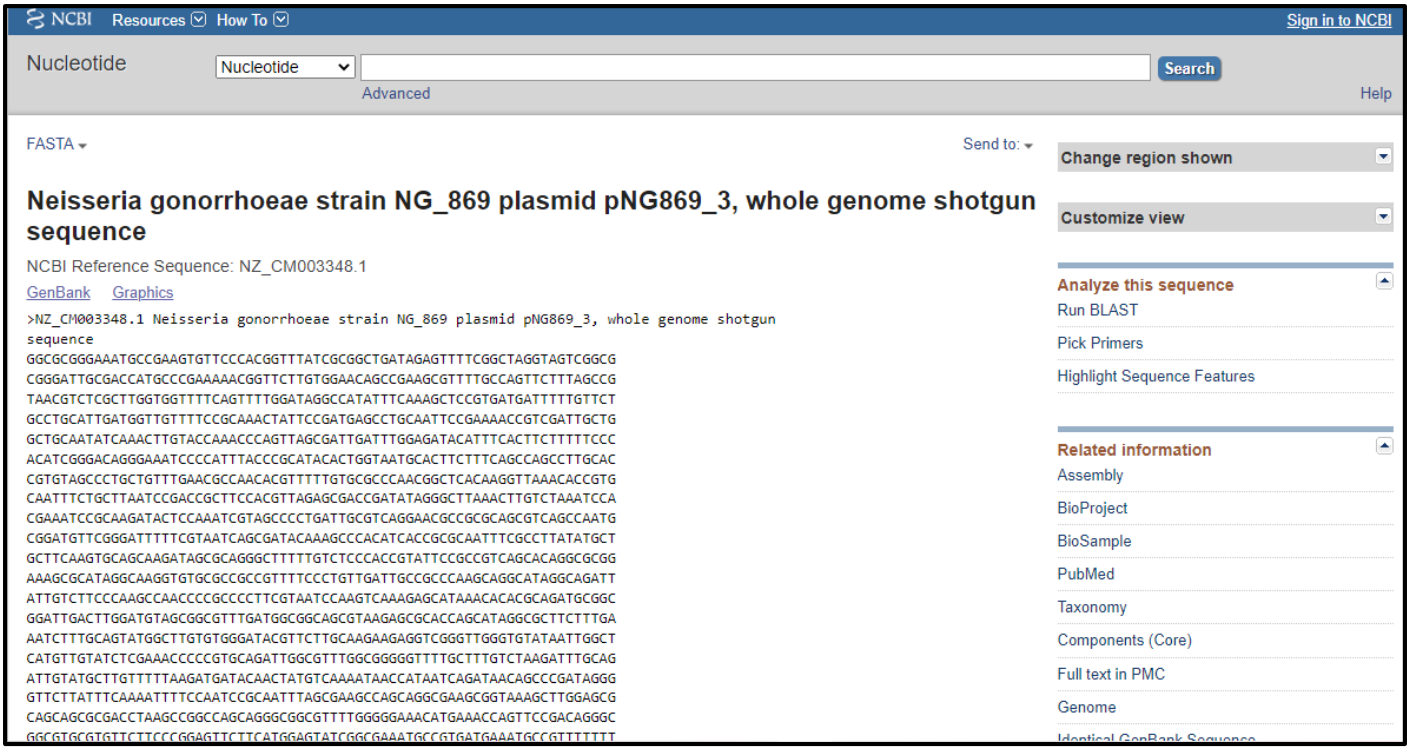


Fig2. Nucleotide FASTA sequence

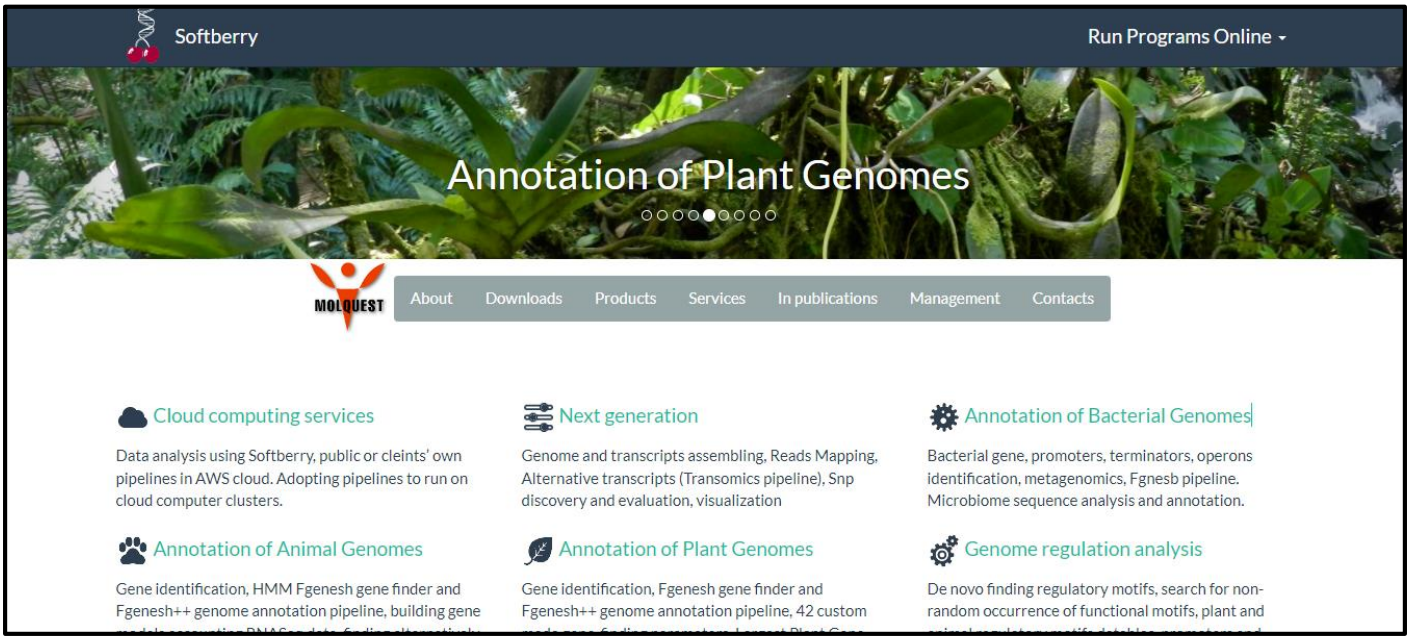


Fig3. Homepage for Softberry

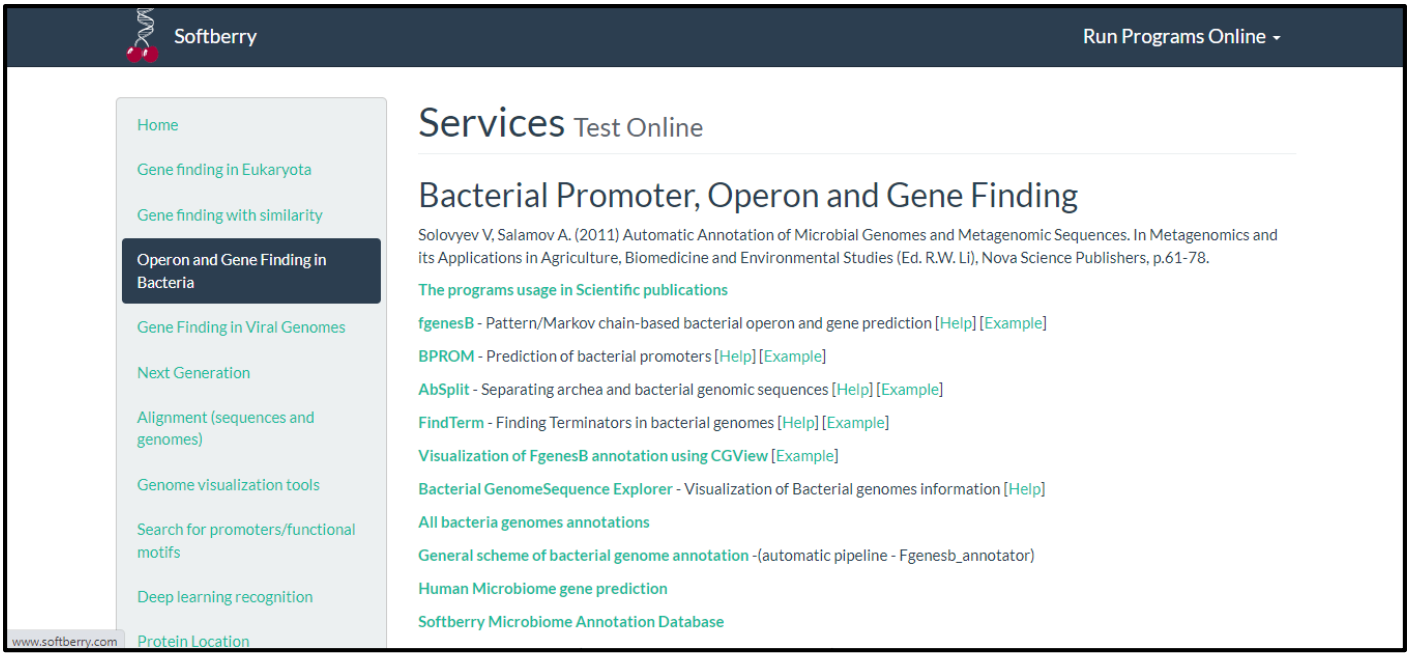


Fig4. Tools for bacterial promoter, operon and gene finding

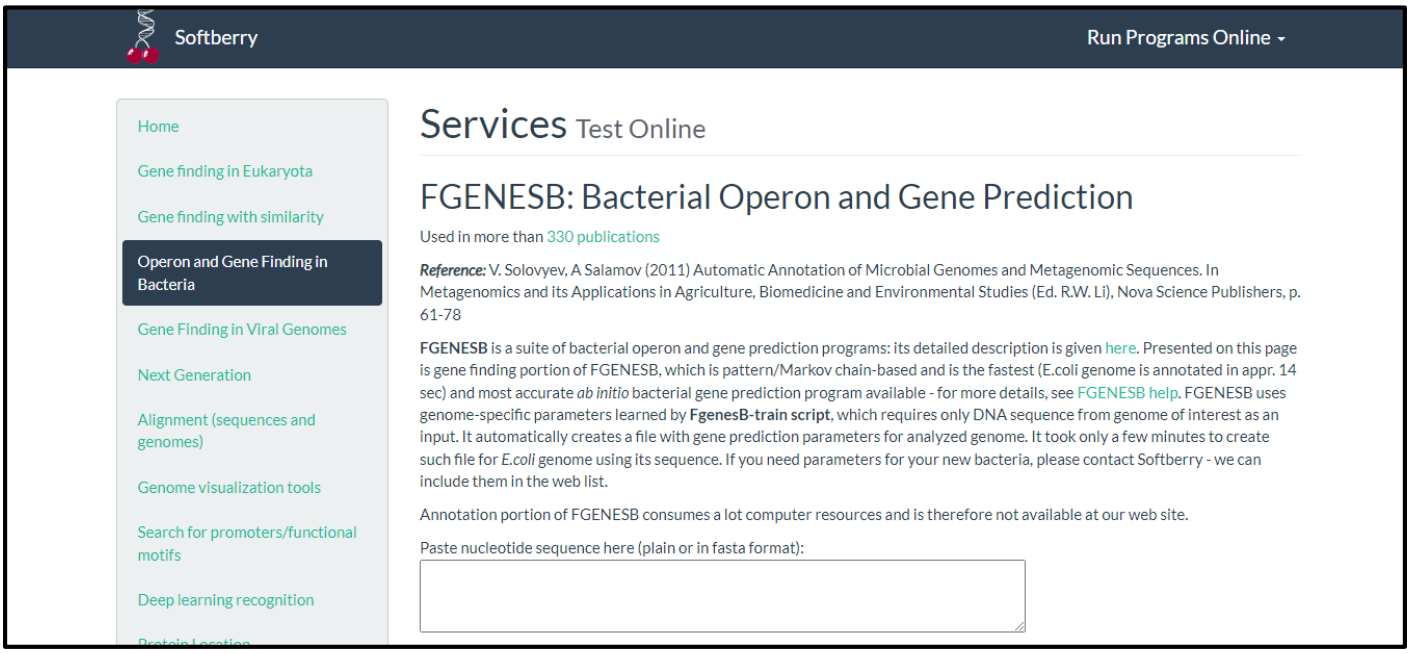


Fig5. Homepage for FGENESB

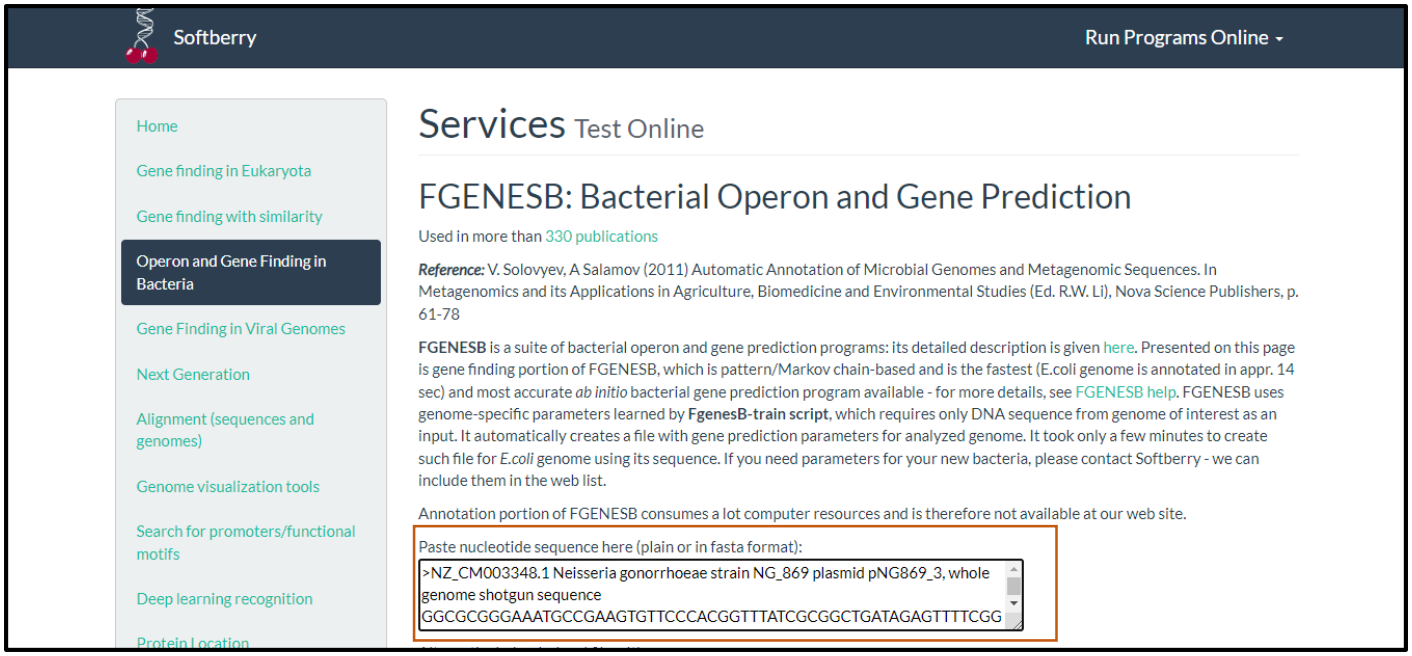
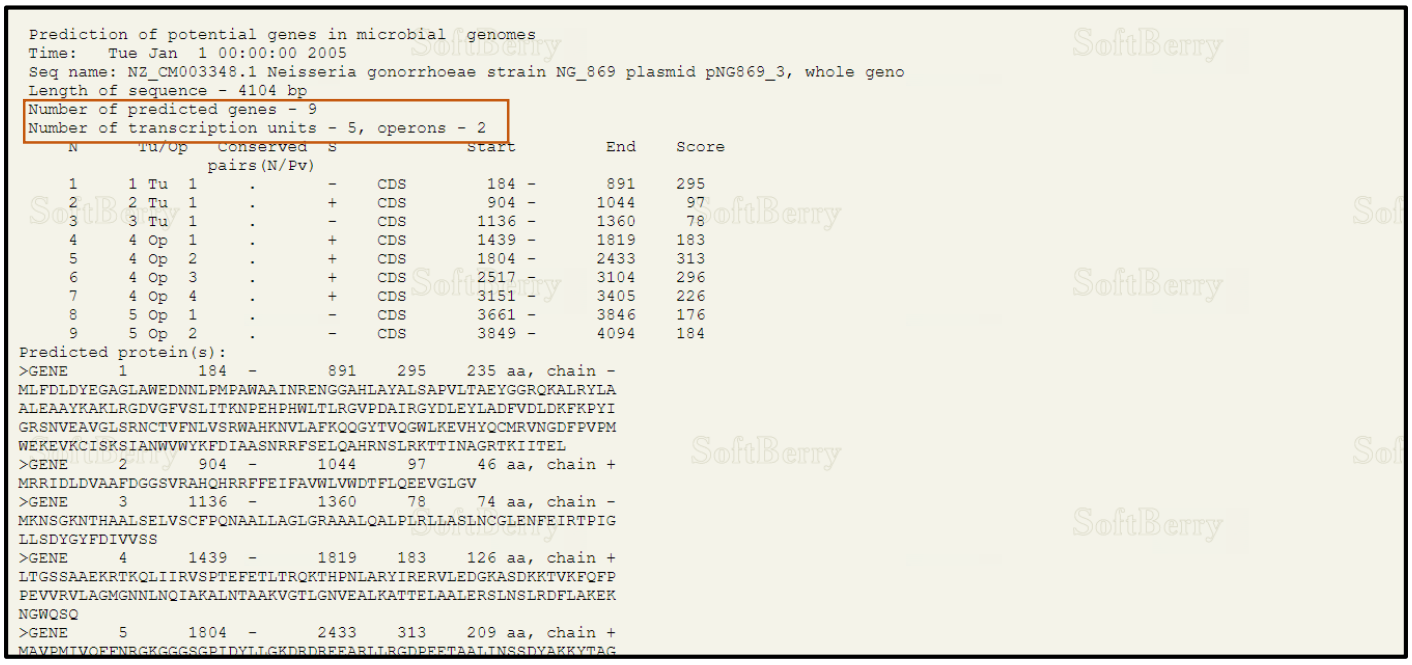


Fig6. Search for nucleotide FASTA sequence



```
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 geno
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
                    pairs(N/Pv)
     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 -
MLFDLDYEGAGLAWEDNNLPMPAWAAINRENGGAHLAYALSAPVLTAEYGGRQKALRYLA
ALEAAYKAKLRGDVGFVSLITKNPEHPHWLTLRGVPDAIRGYDLEYLADFVDLDKFKPYI
GRSNVEAVGLSRNCTVFNLVSRWAHKNVLAFKQQGYTVQGWLKEVHYQCMRVNGDFPVPM
WEKEVKCISKSIANWVWYKFDIAASNRRFSELQAHRNSLRKTTINAGRTKIITEL
>GENE     2       904  -      1044     97     46 aa, chain +
MRRIDLDVAAFDGGSVRAHQHRRFFEIFAVWLVWDTFLQEEVGLGV
>GENE     3      1136  -      1360     78     74 aa, chain -
MKNSGKNTHAALSELVSCFPQNAALLAGLGRAAALQALPLRLLASLNCGLENFEIRTPIG
LLSDYGYFDIVVSS
>GENE     4      1439  -      1819    183    126 aa, chain +
LTGSSAAEKRTKQLIIRVSPTEFETLTRQKTHPNLARYIRERVLEDGKASDKKTVKFQFP
PEVVRVLAGMGNNLNQIAKALNTAAKVGTLGNVEALKATTELAALERSLNSLRDFLAKEK
NGWQSQ
>GENE     5      1804  -      2433    313    209 aa, chain +
MAVPMIVQFFNRGKGGGSGPIDYLLGKDRDREFARLLRGDPETAALINSSDYAKKYTAG
```

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

DATE: 18/03/22

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 to recognized using FGENES.

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

METHODOLOGY:

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:

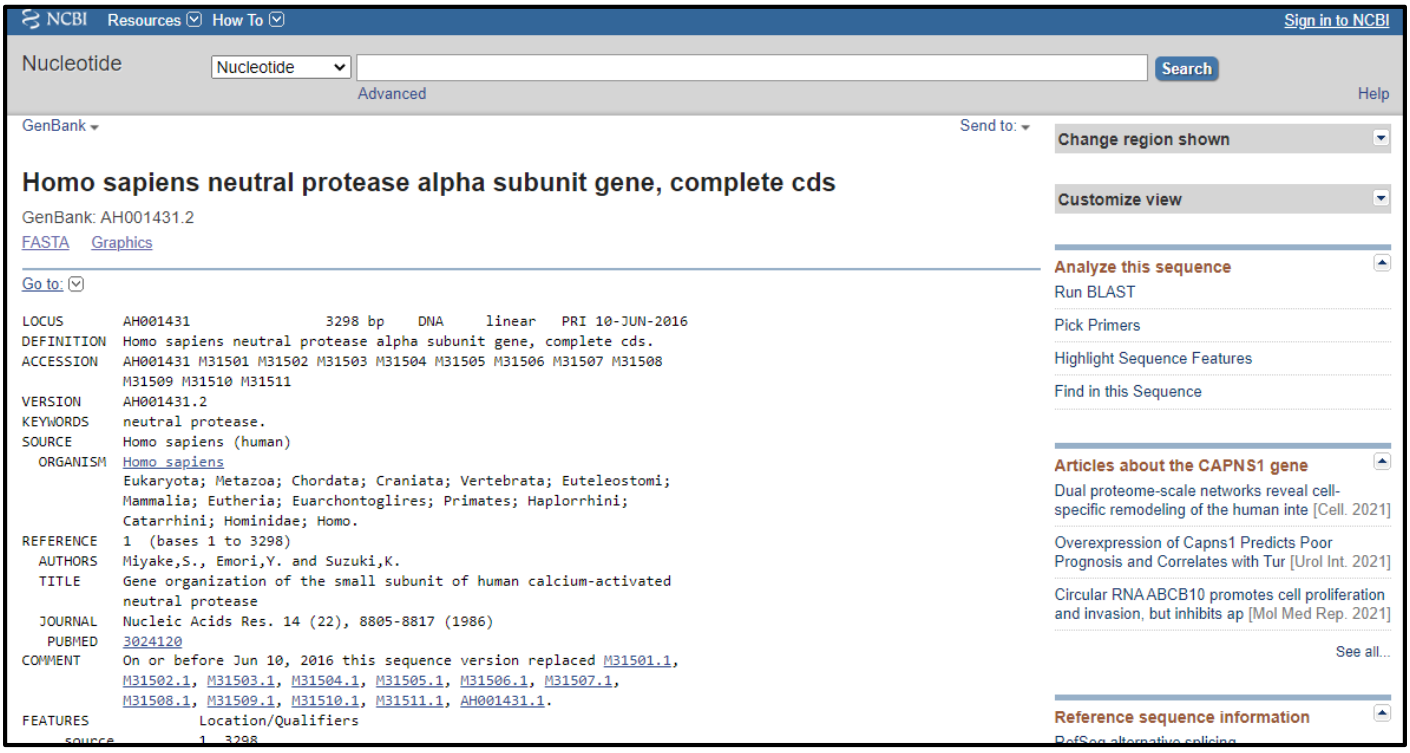


Fig1. GenBank result for Human Protease

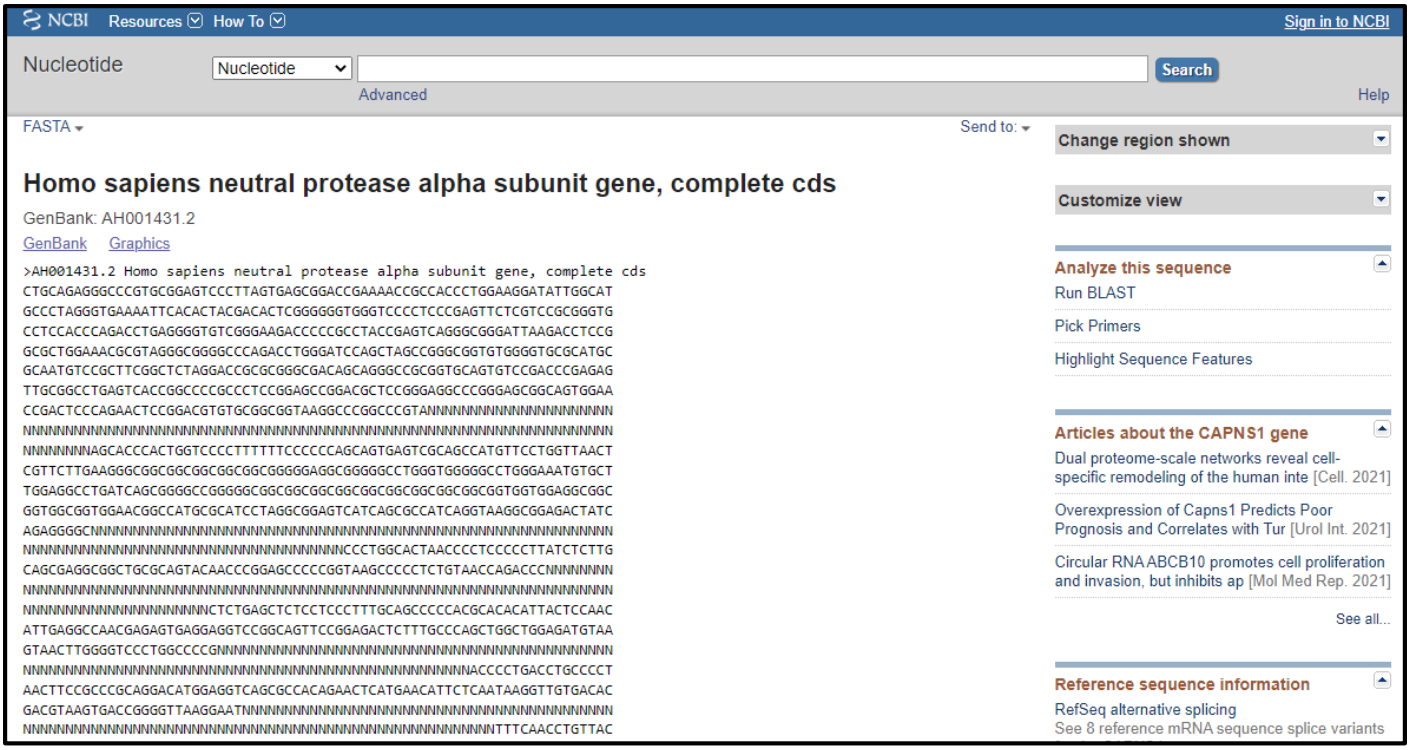


Fig2. Nucleotide FASTA sequence for Protease

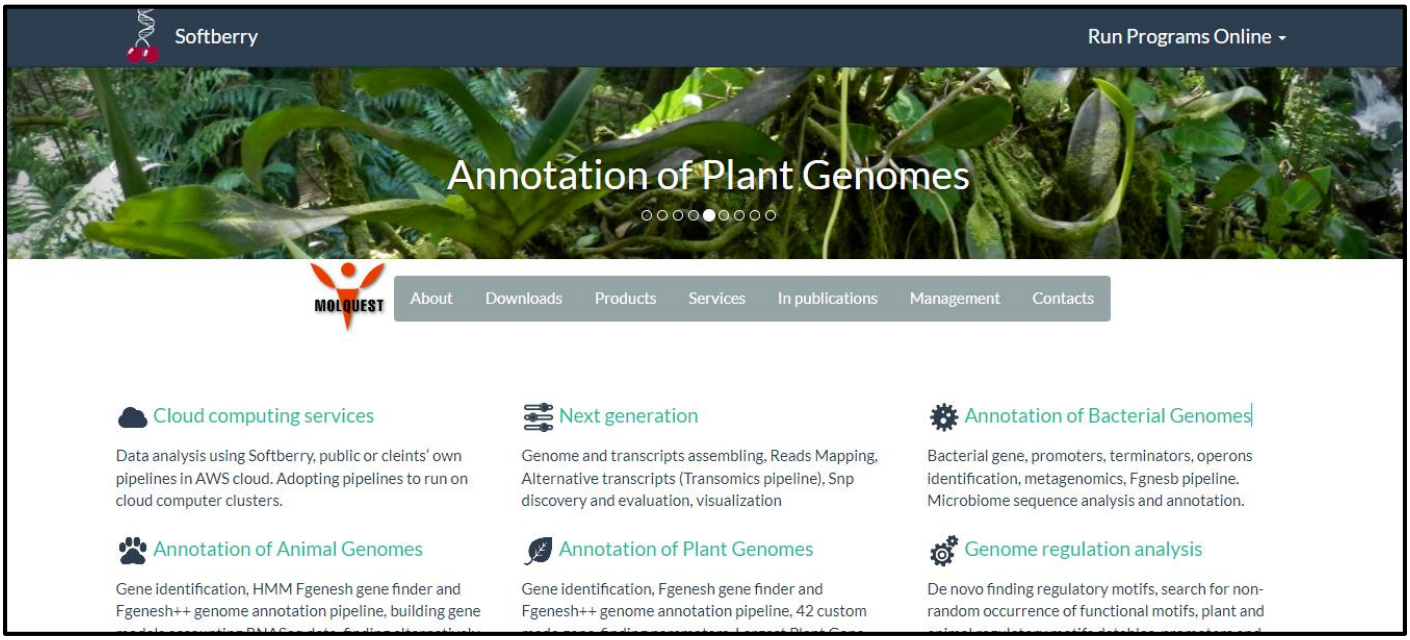


Fig3. Homepage for Softberry

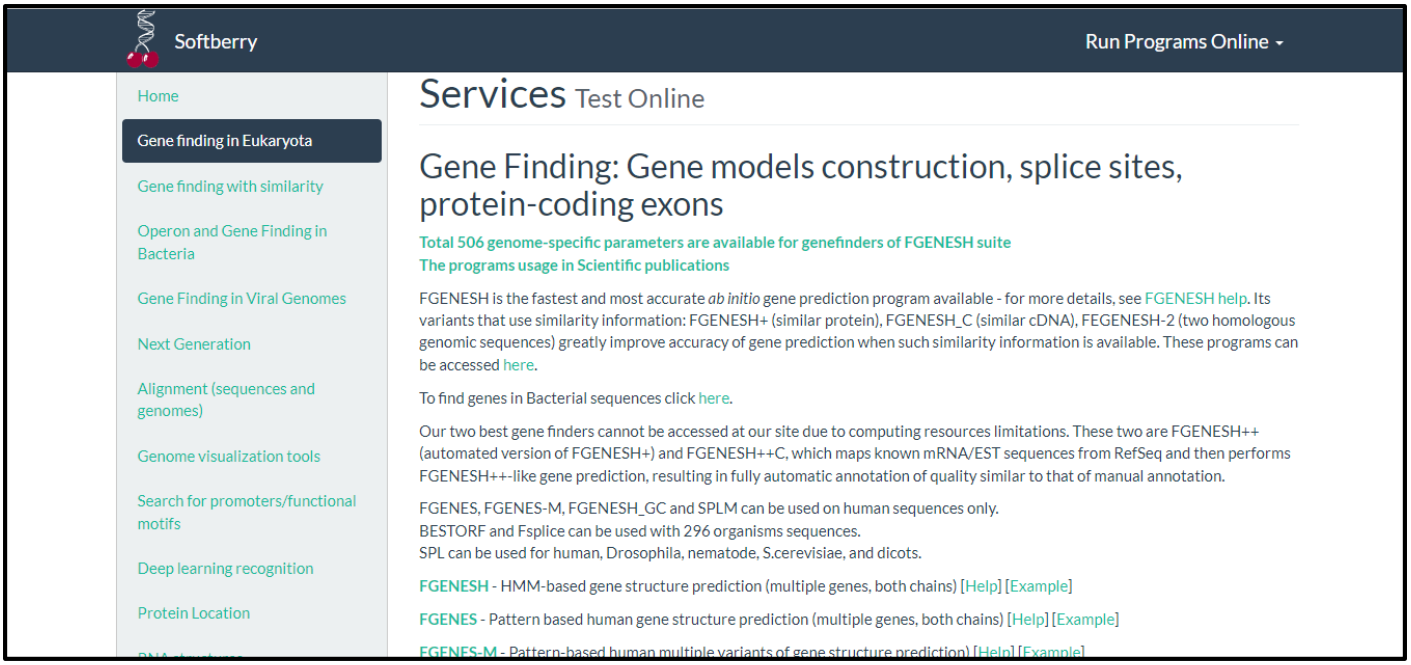


Fig4. Tools for gene finding in Eukaryota

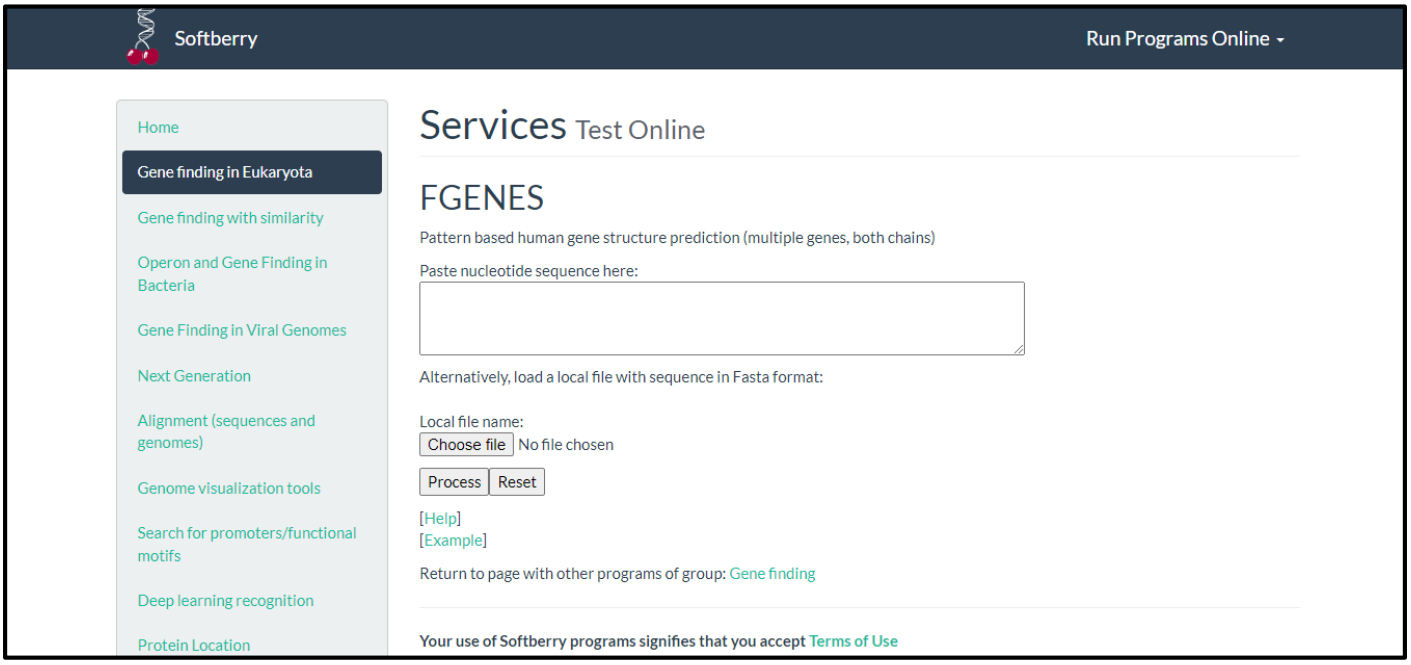


Fig5. Homepage for FGENES

Fig6. Search for protease nucleotide FASTA sequence

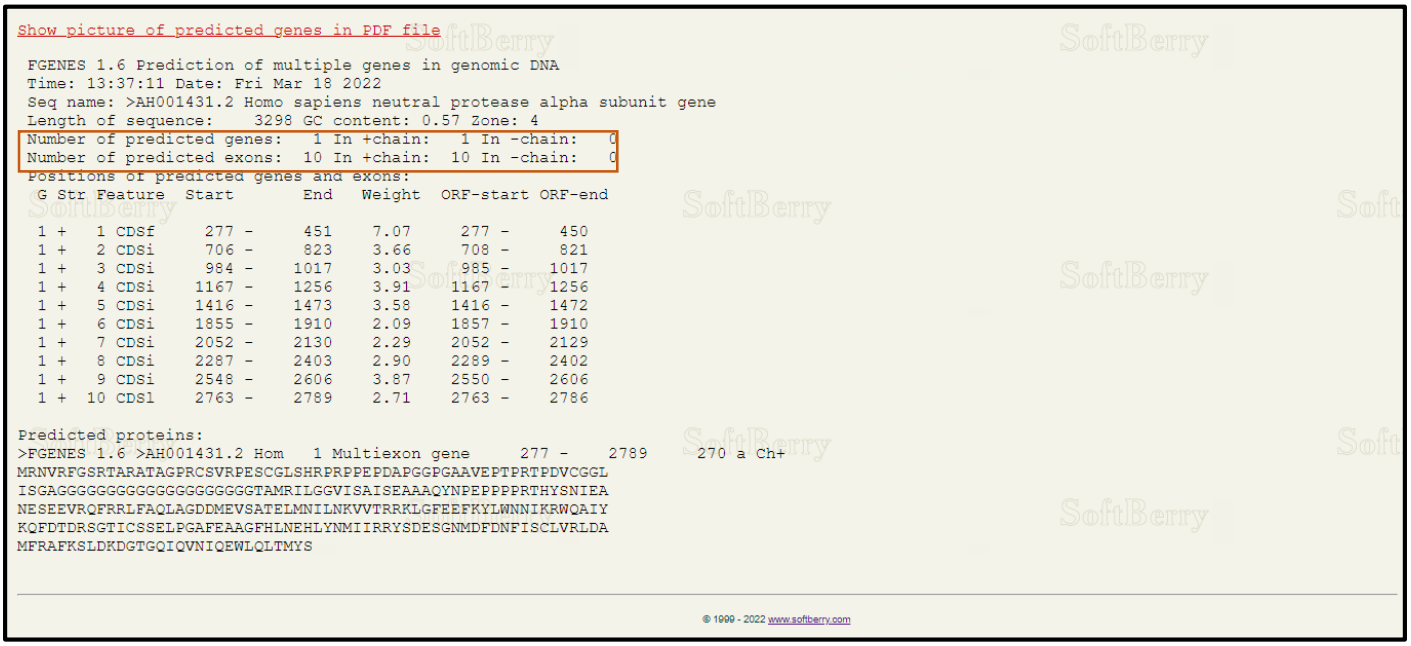


Fig7. Result for predicted exon signals

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 undering the gene function.

REFERENCES:

1. Xiong, J. (2008).*Gene Prediction*. *Essential bioinformatics*. Cambridge: Cambridge University Press. 97-111.

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

DATE: 18/03/22

WEBLEM 8e

ORF finder- NCBI

(URL: https://www.ncbi.nlm.nih.gov/orffinder/)

AIM:

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

INTRODUCTION:

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:

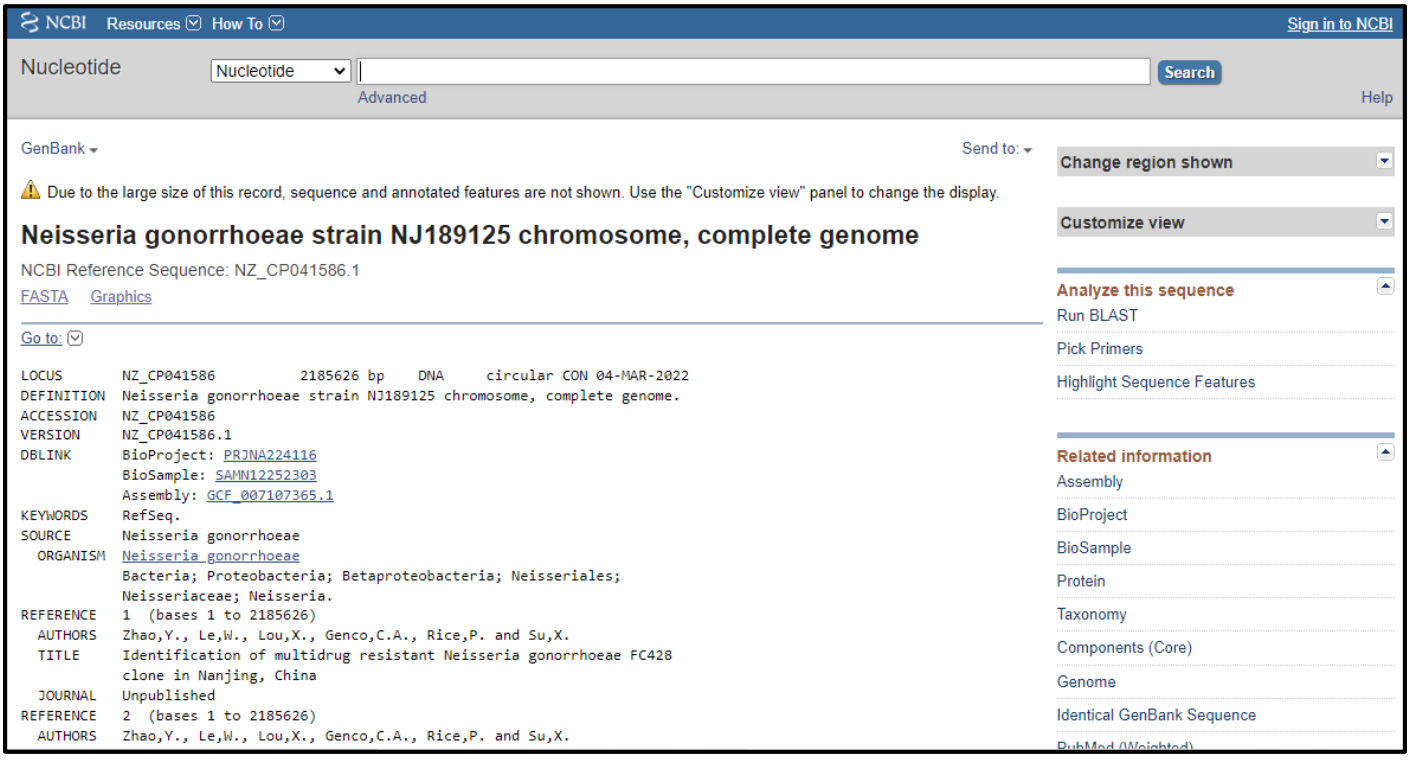


Fig1. Result for *Neisseria gonorrhoeae* in GenBank

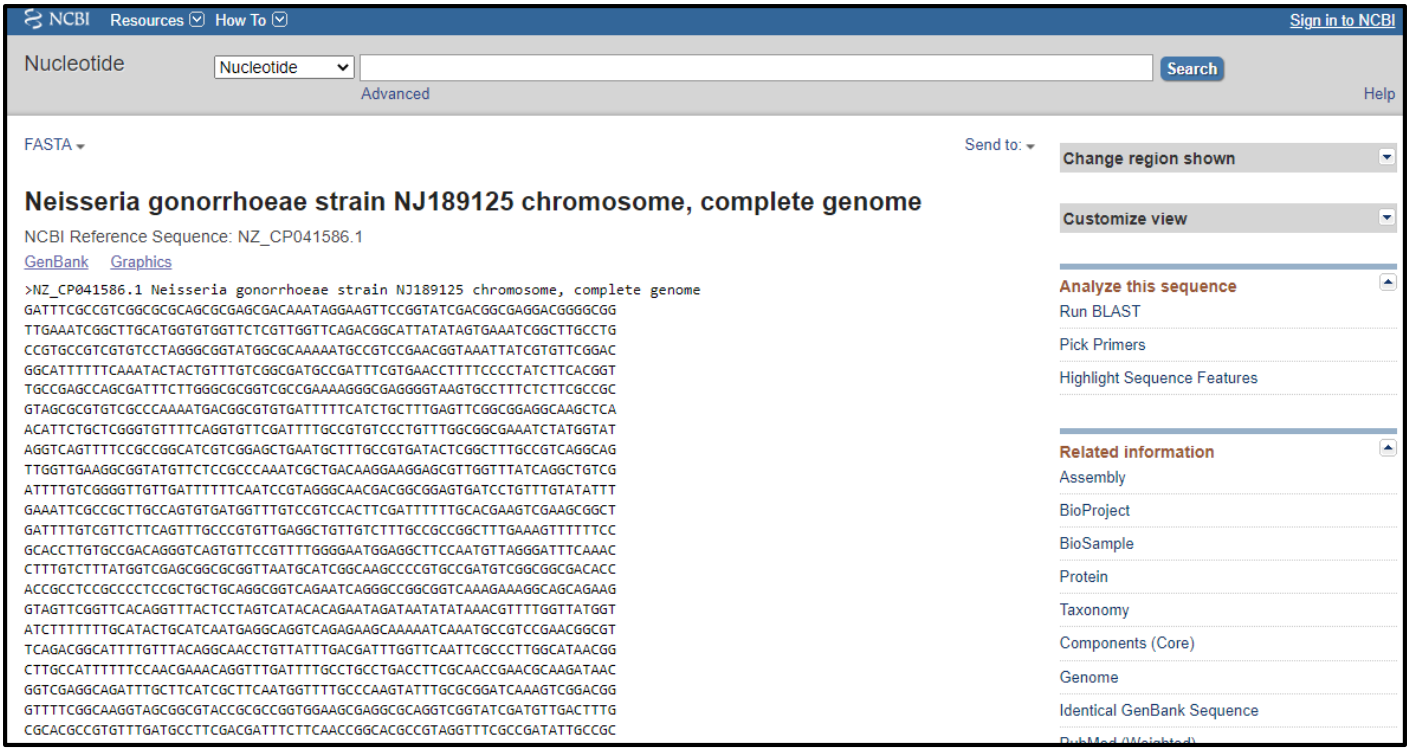


Fig2. Nucleotide FASTA sequence for *Neisseria gonorrhoeae*

Fig3. Search for Nucleotide FASTA sequence

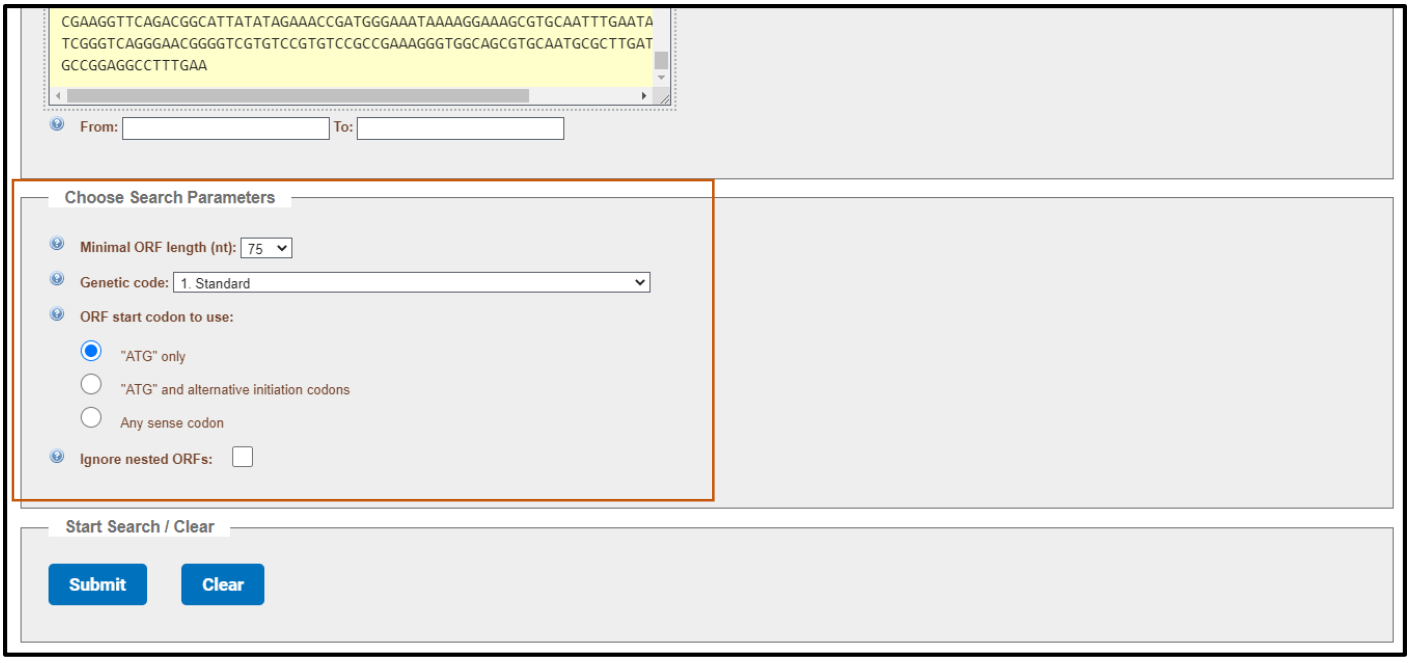


Fig4. Search parameters

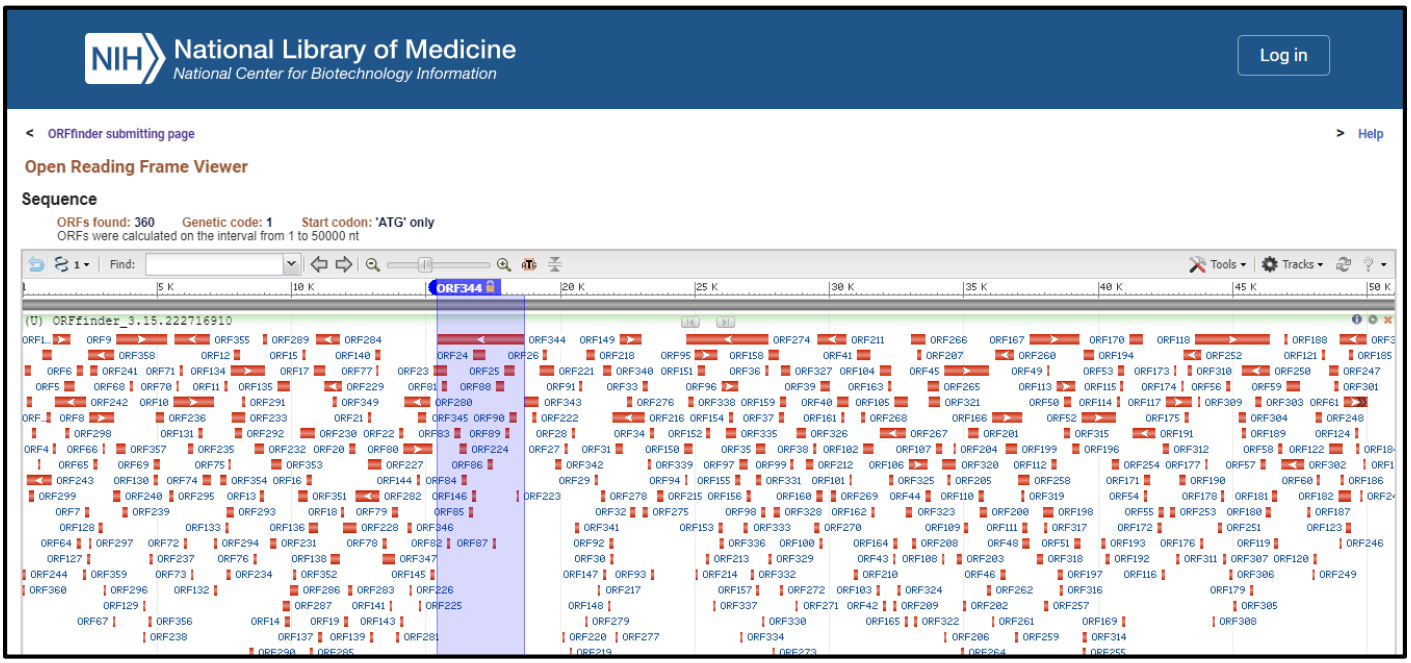


Fig5. Result for recognised ORFs

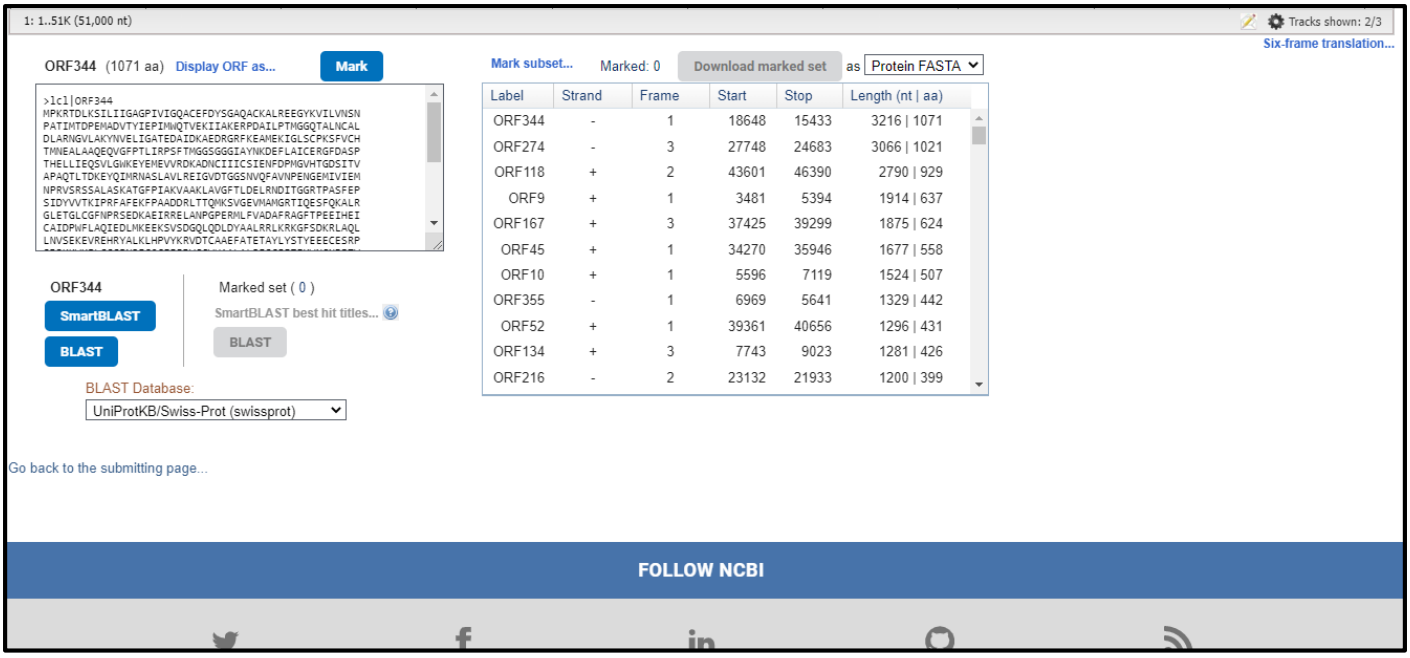


Fig6. Results for tracks information

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

DATE: 28/03/22

WEBLEM 9

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

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

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

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

1. UCSC Genome Browser

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

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

Sequence and annotation data for each genome assembly are stored in a MySQL relational database, which is quite efficient at retrieving data from indexed files. The database is loaded in large batches and is used primarily as a read-only database. To improve performance, each of the Genome Browser web servers has a copy of the database on its local disk.

UCSC generates several annotations based on mRNA alignments. The mRNA and EST sequences are extracted from GenBank, and are aligned against the genome using the BLAST-like Alignment Tool (BLAT), a fast sequence alignment tool developed by Jim Kent. The data is filtered based on percentage identity and near best in genome to select only those alignments that best match the sequence. The spliced EST annotation is computed from the filtered data by analyzing the EST alignments for evidence of splicing.

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

DATE: 28/03/22

WEBLEM 9a

UCSC Genome Browser

(URL: https://genome.ucsc.edu/)

AIM:

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

INTRODUCTION:

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

METHODOLOGY:

1. Open homepage for UCSC browser. (URL: https://genome.ucsc.edu/)
2. Select genome browser.
3. Select human assembly (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:

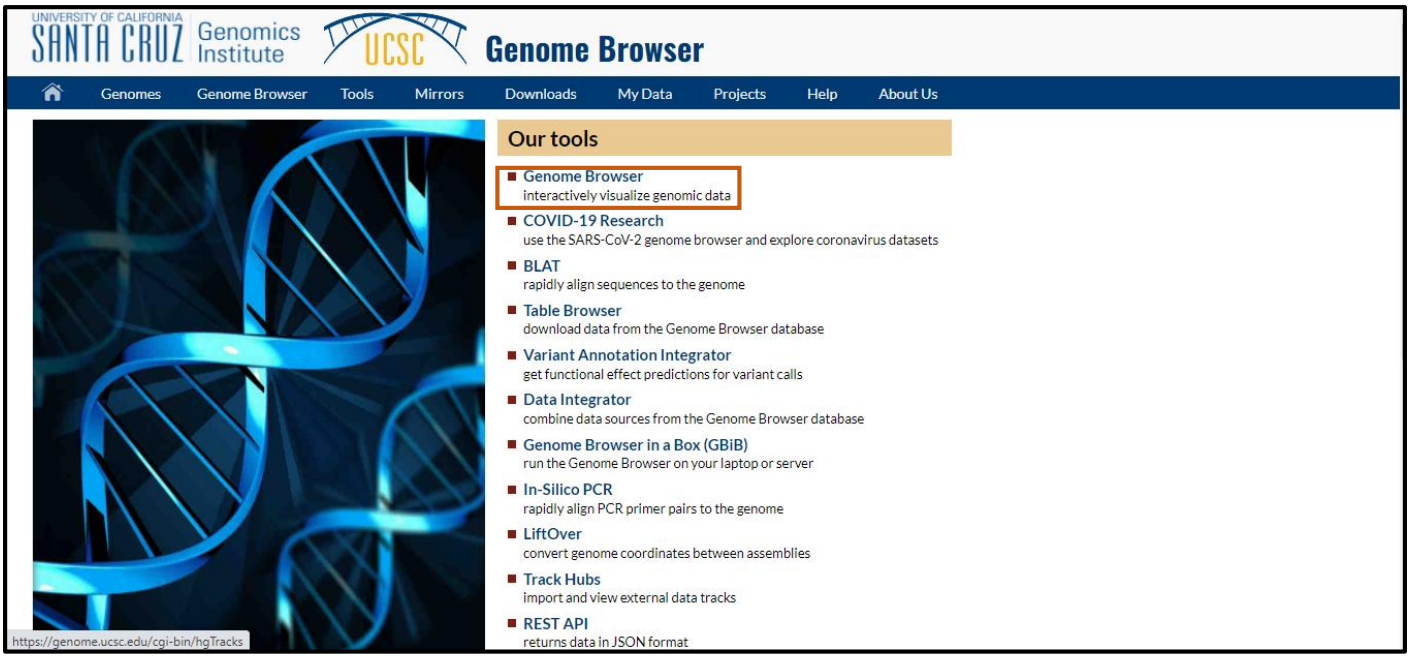


Fig1. Homepage for UCSE browser

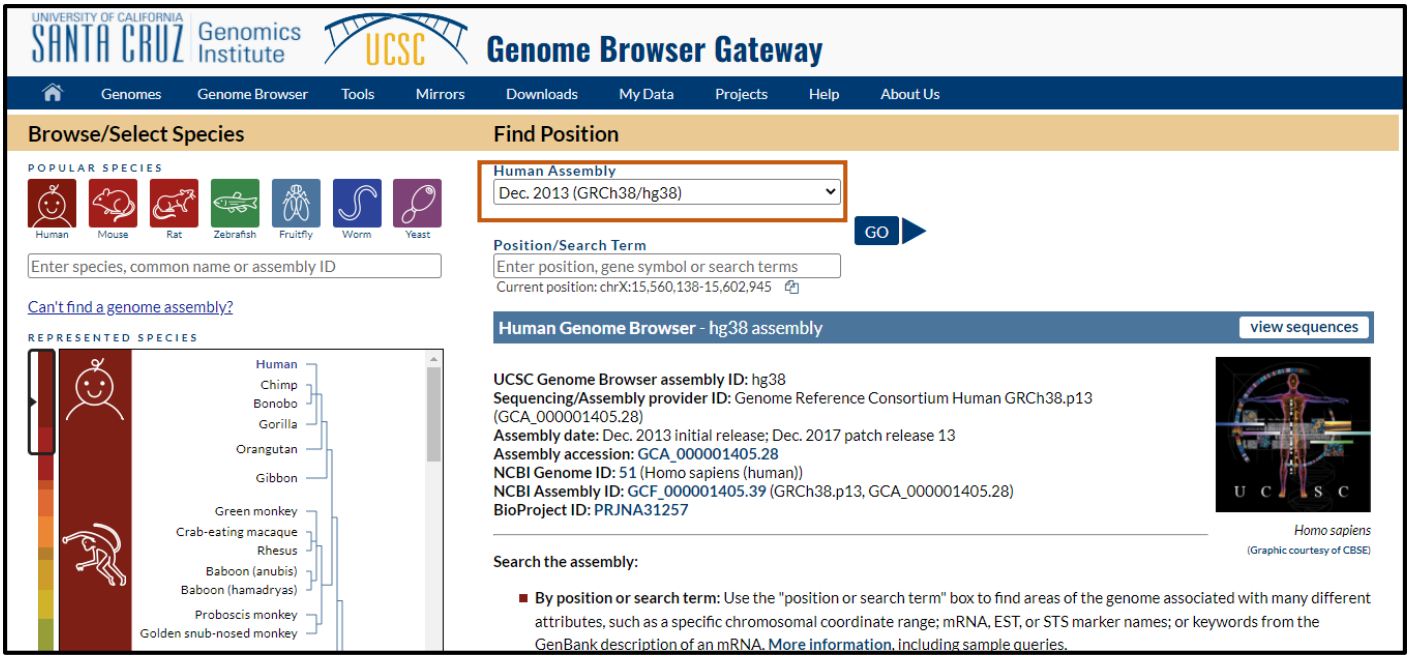


Fig2. Genome browser gateway

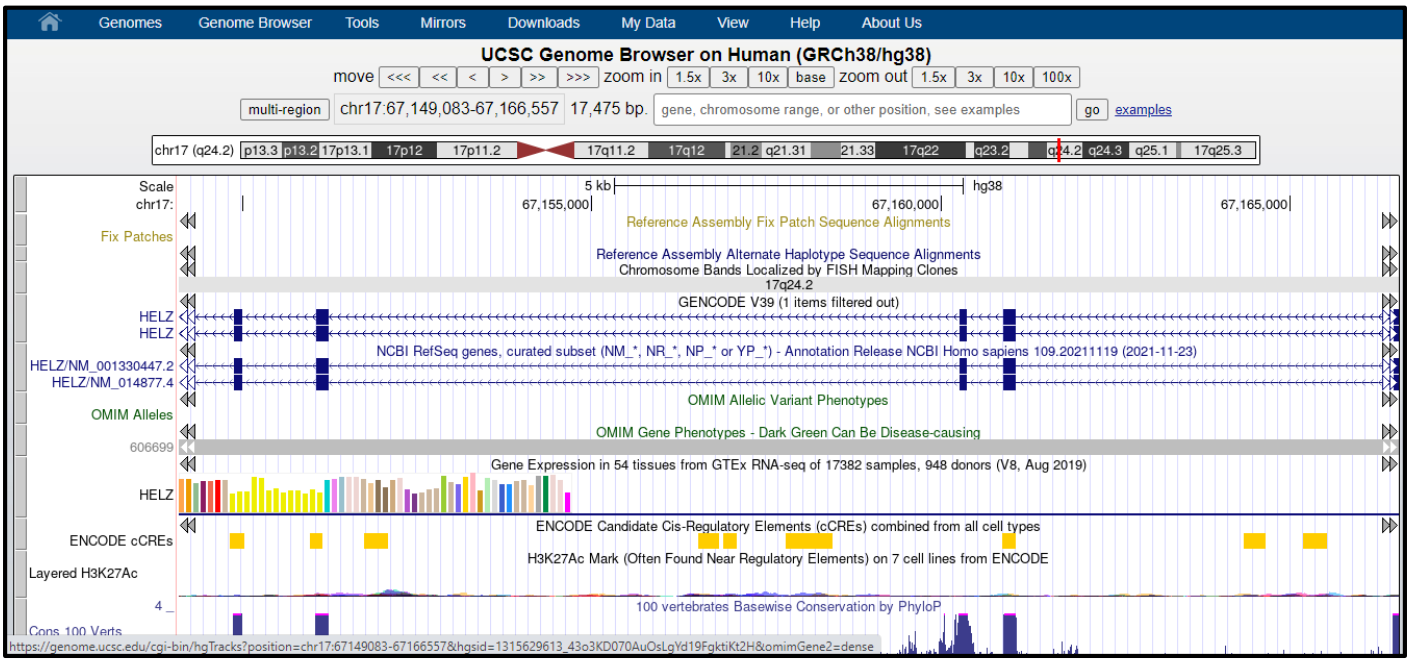


Fig3. UCSC genome browser on human (GRCh38/hg38)

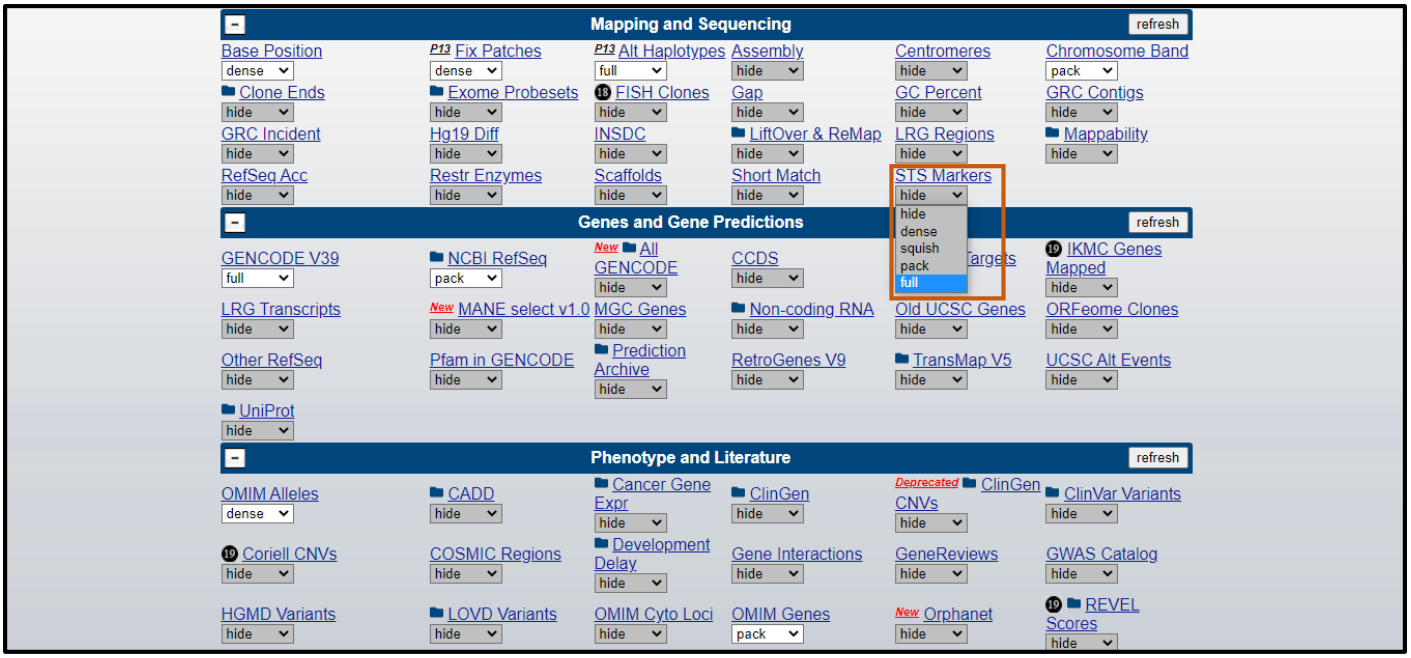


Fig4. Options for configuration

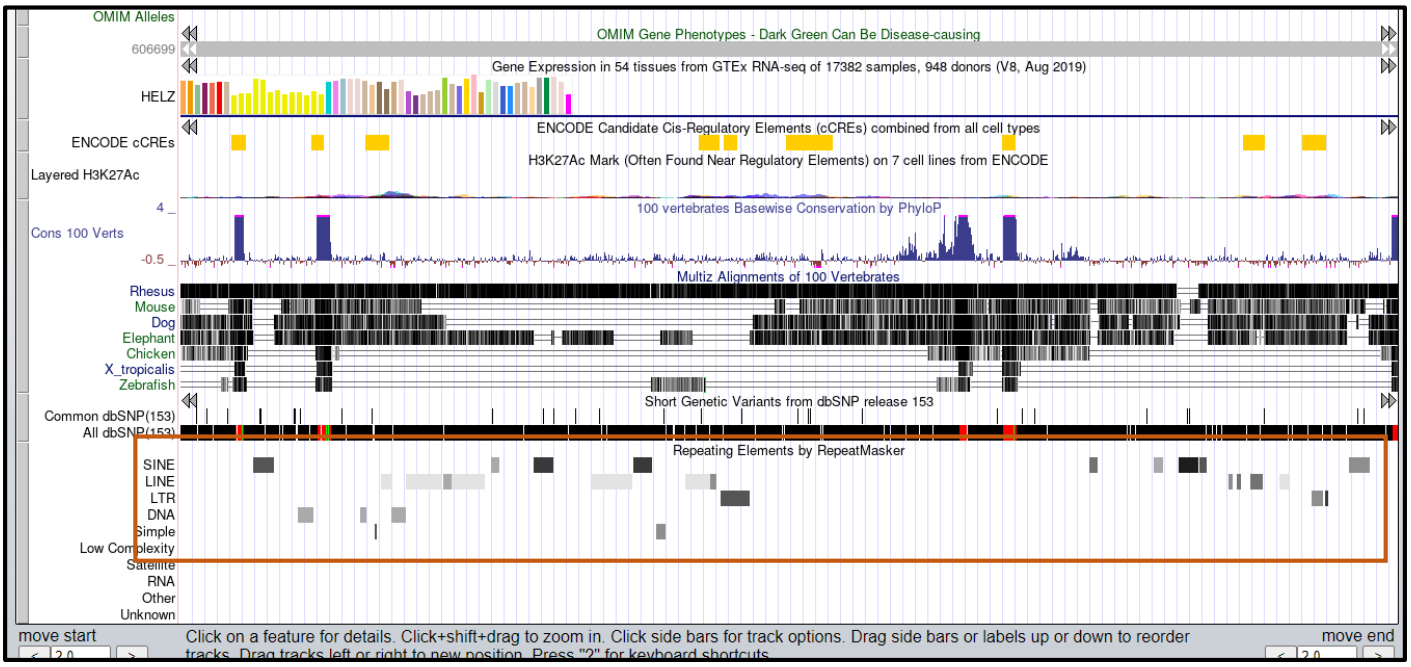


Fig5. Result after configuration

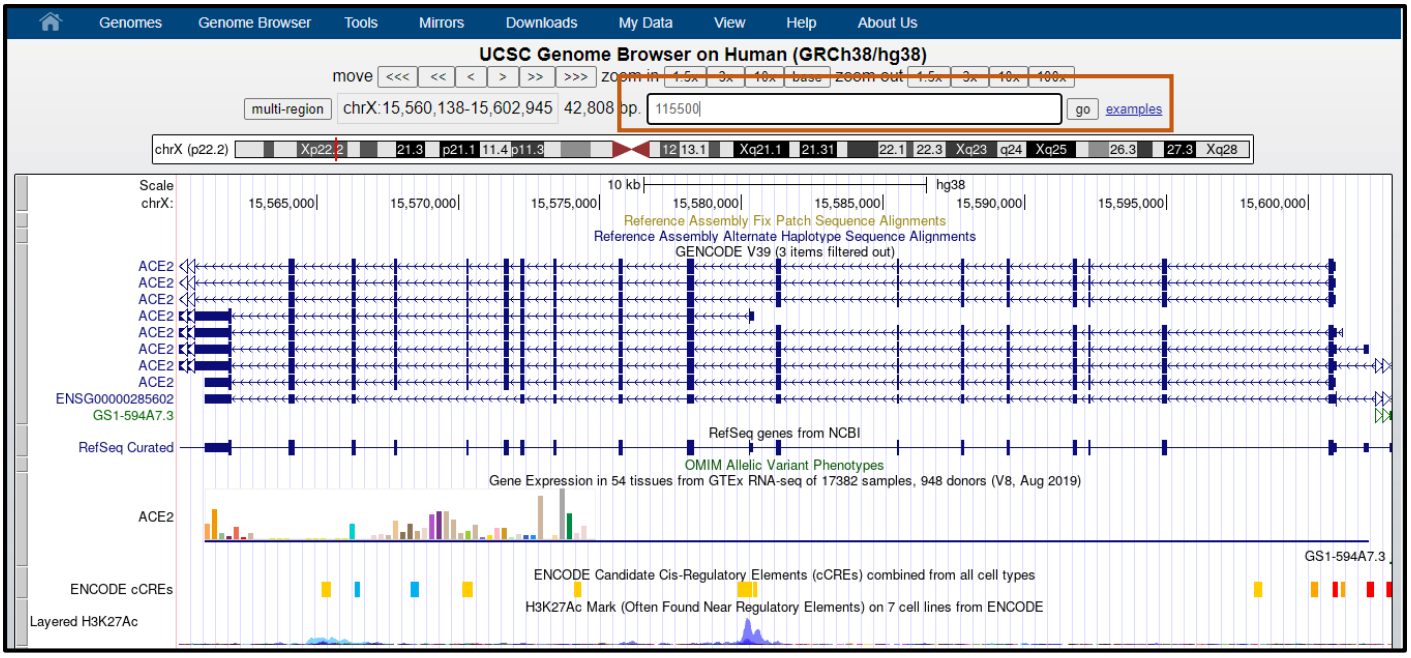


Fig6. Navigation by OMIM Id: 115500

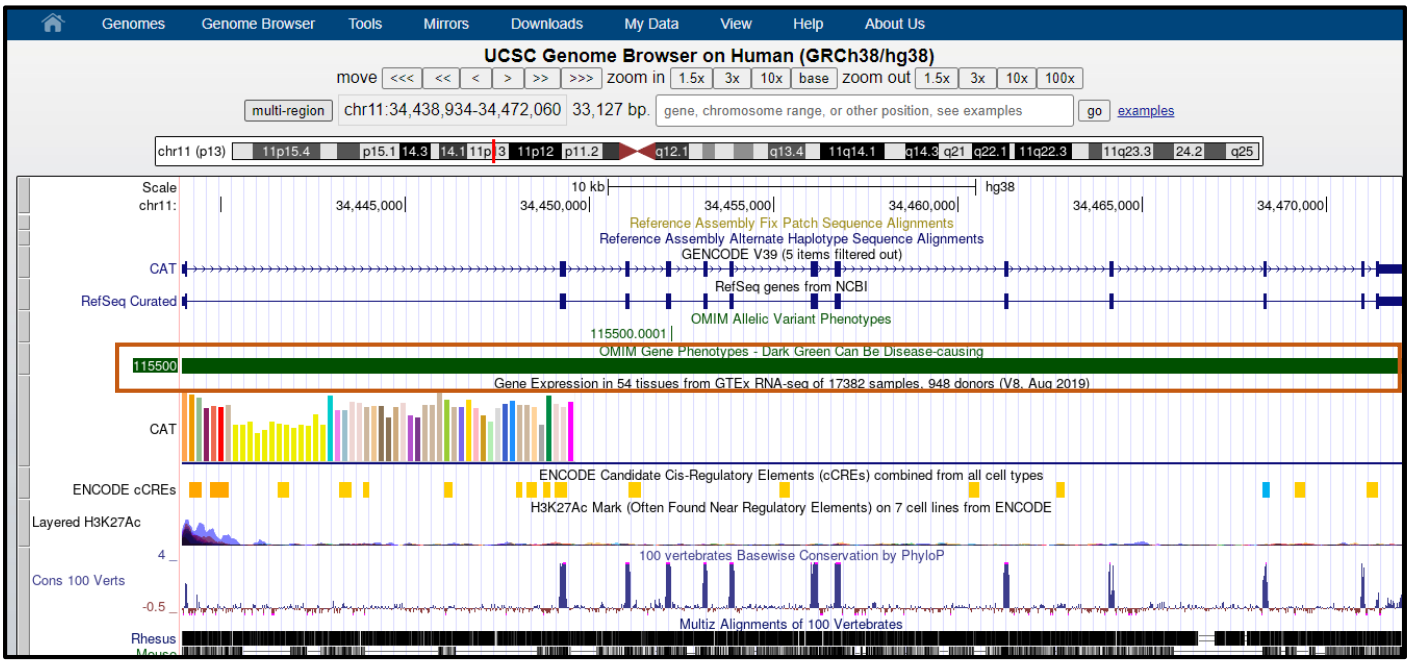


Fig6.1. Result for OMIM Id: 115500

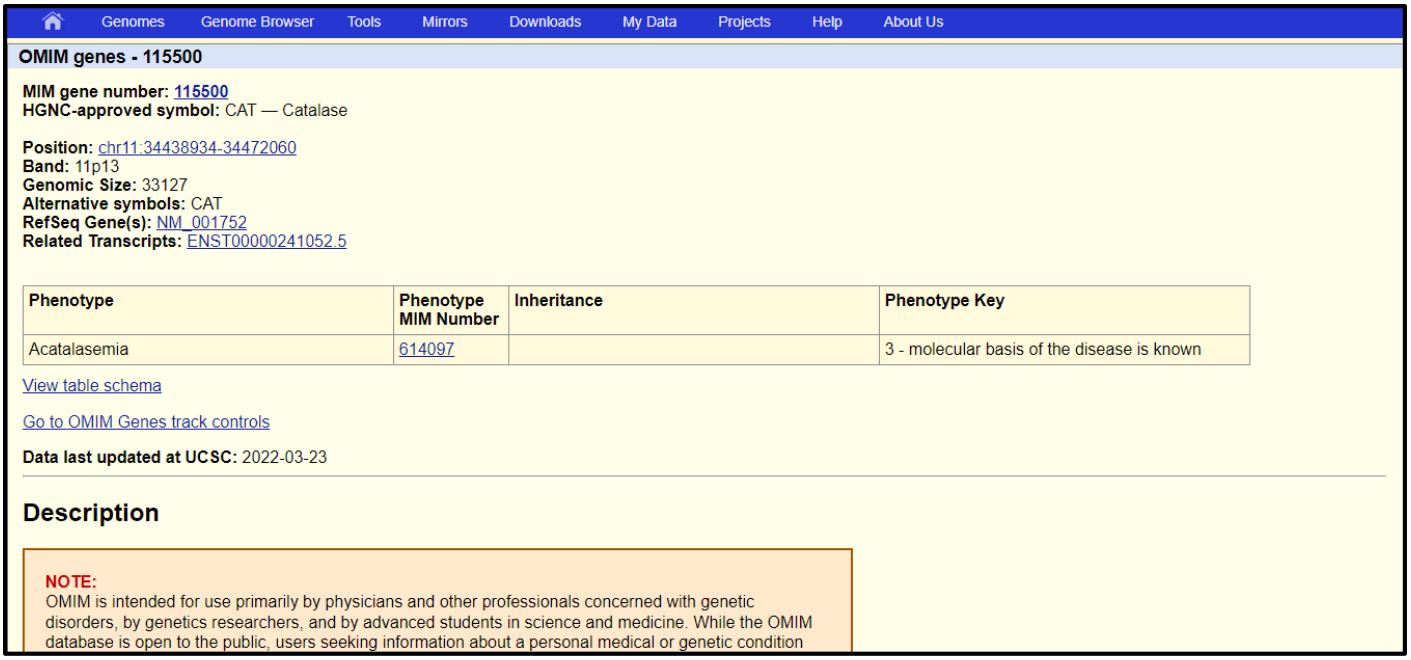


Fig6.2. Description for OMIM gene

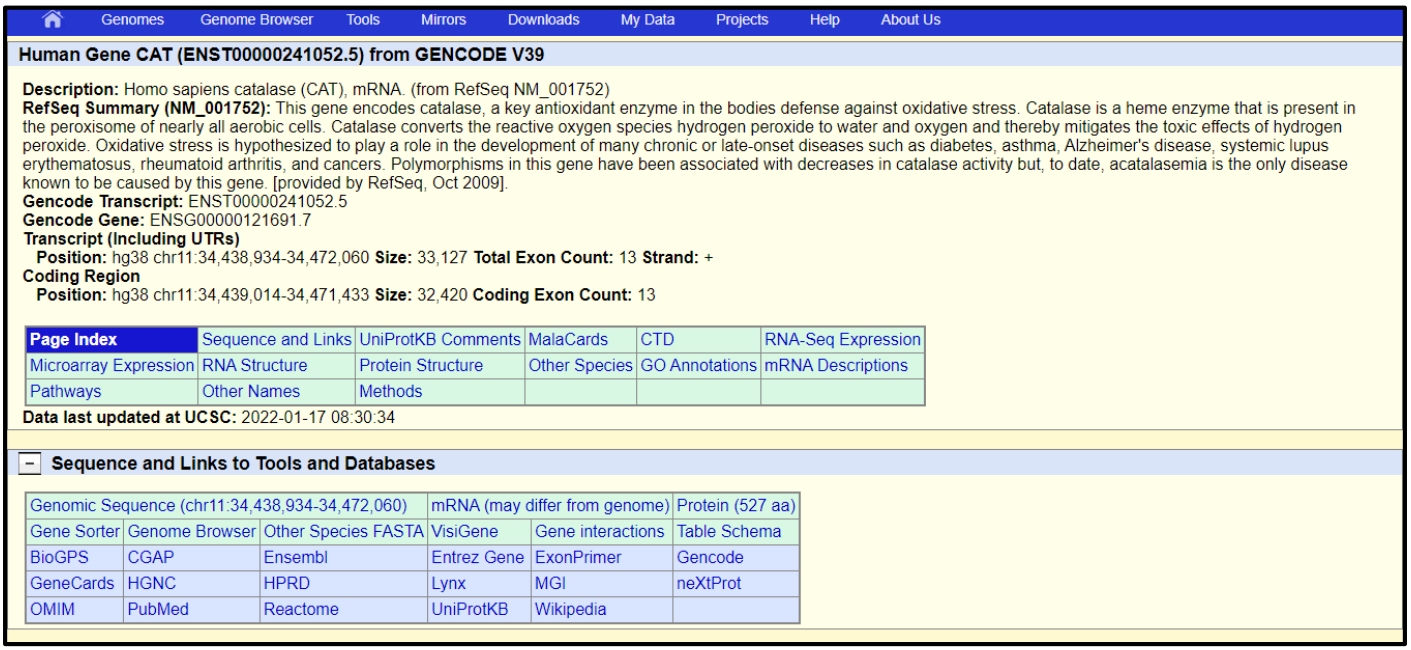


Fig6.3. Related transcripts information

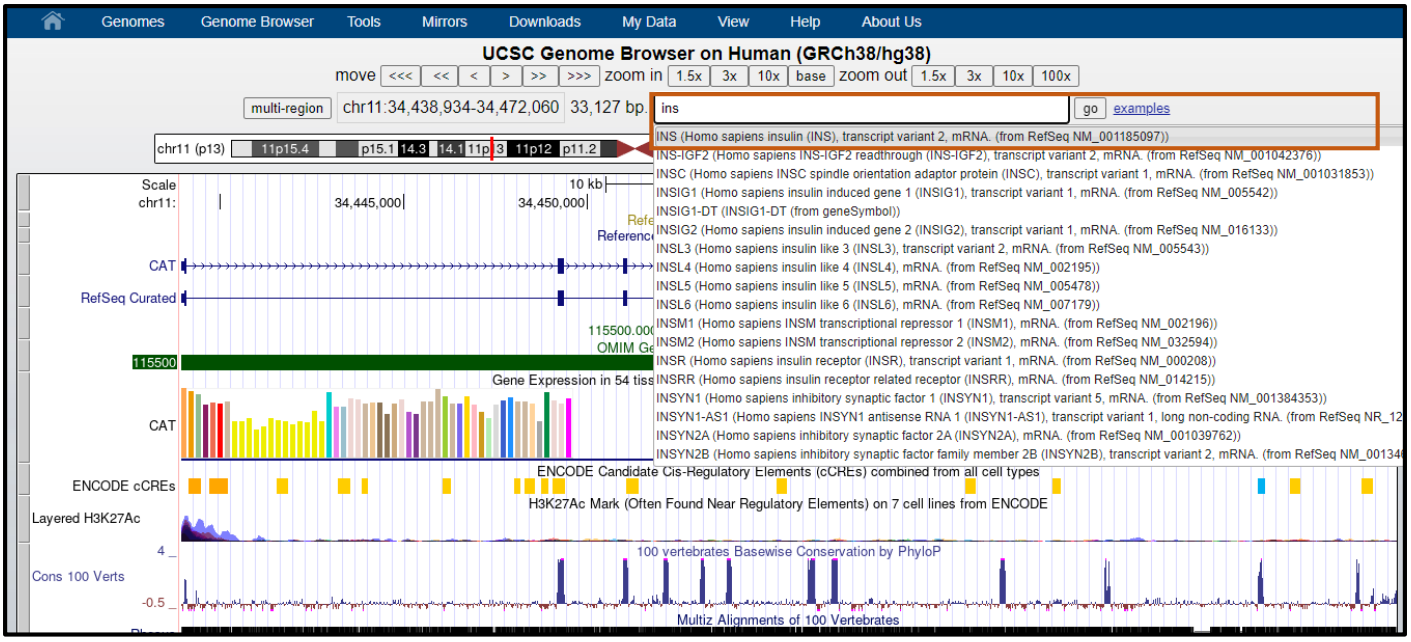


Fig7. Navigation by gene name: INS

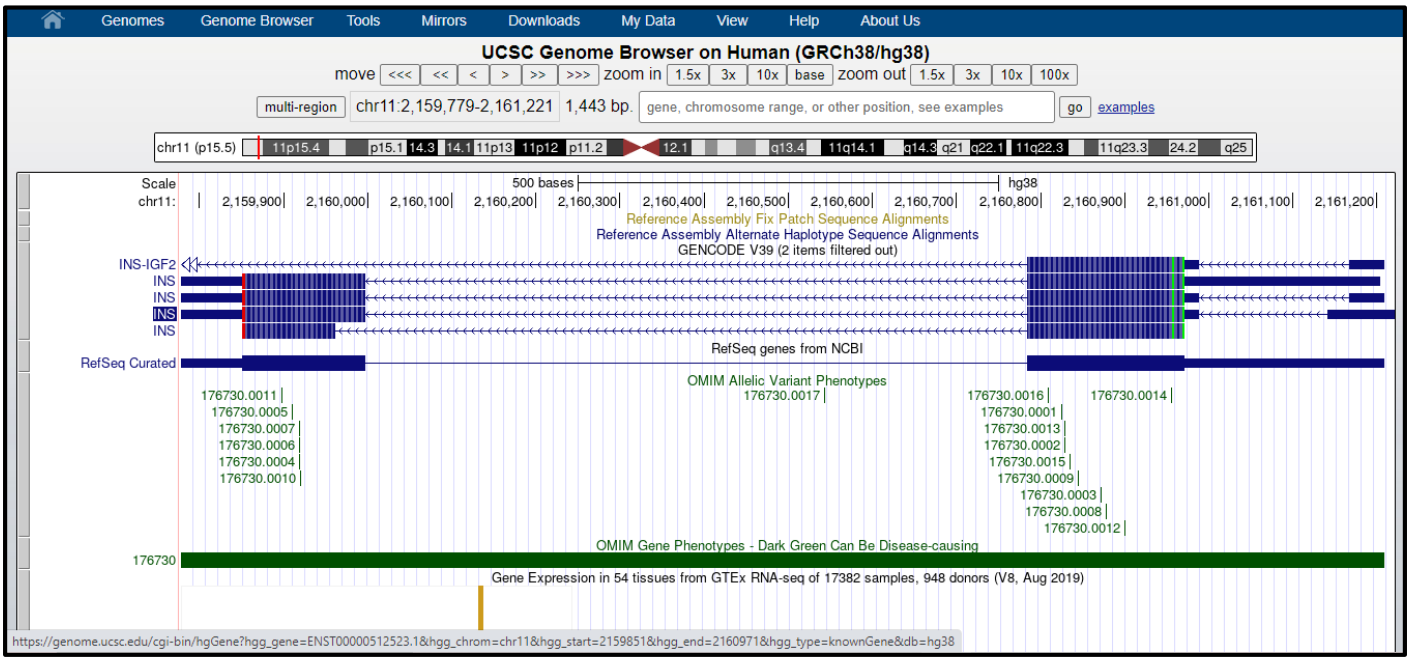


Fig7.1. Result for INS gene

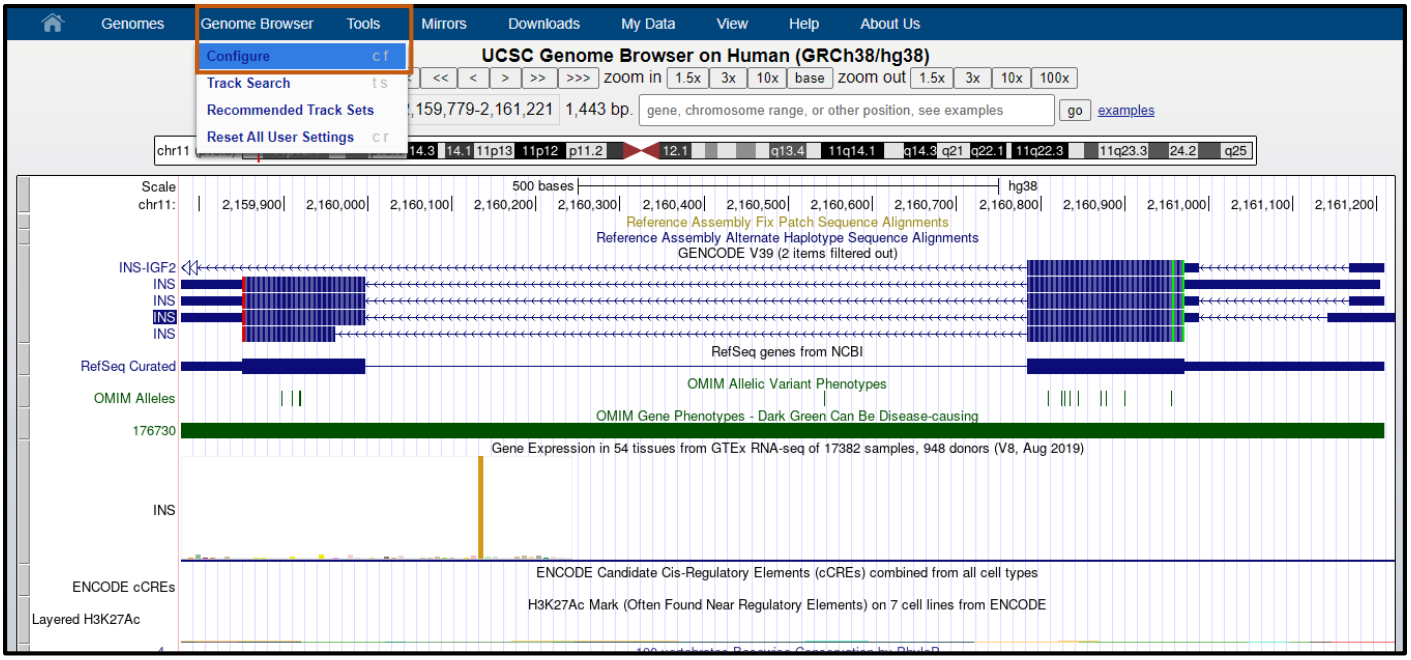


Fig7.2. Option to configure tracks

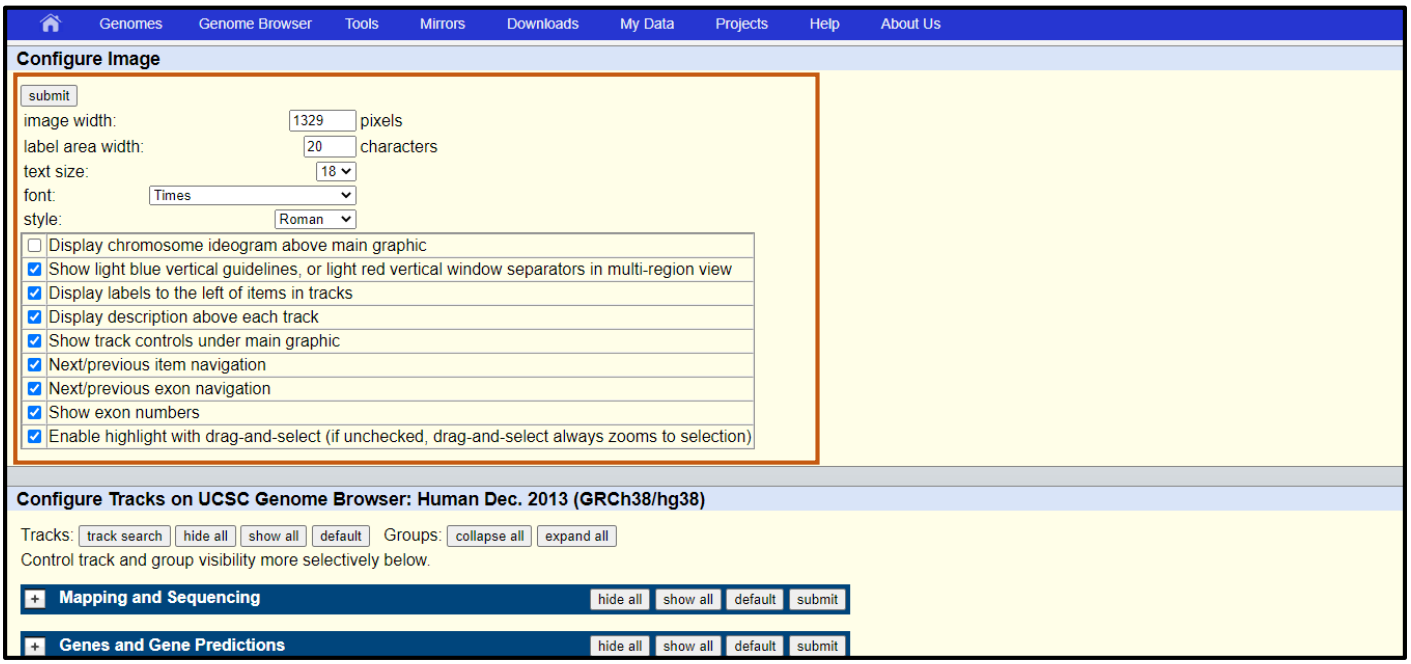


Fig7.3. Configuration settings

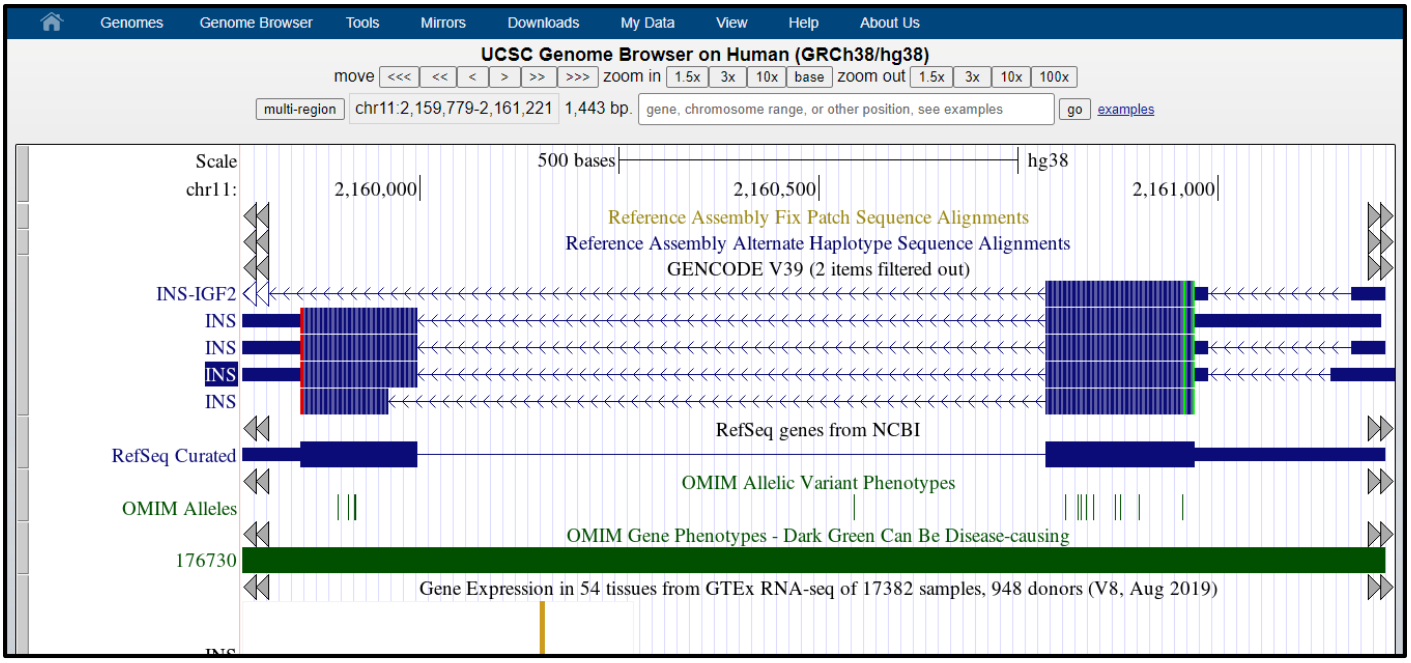


Fig7.4. Configured tracks

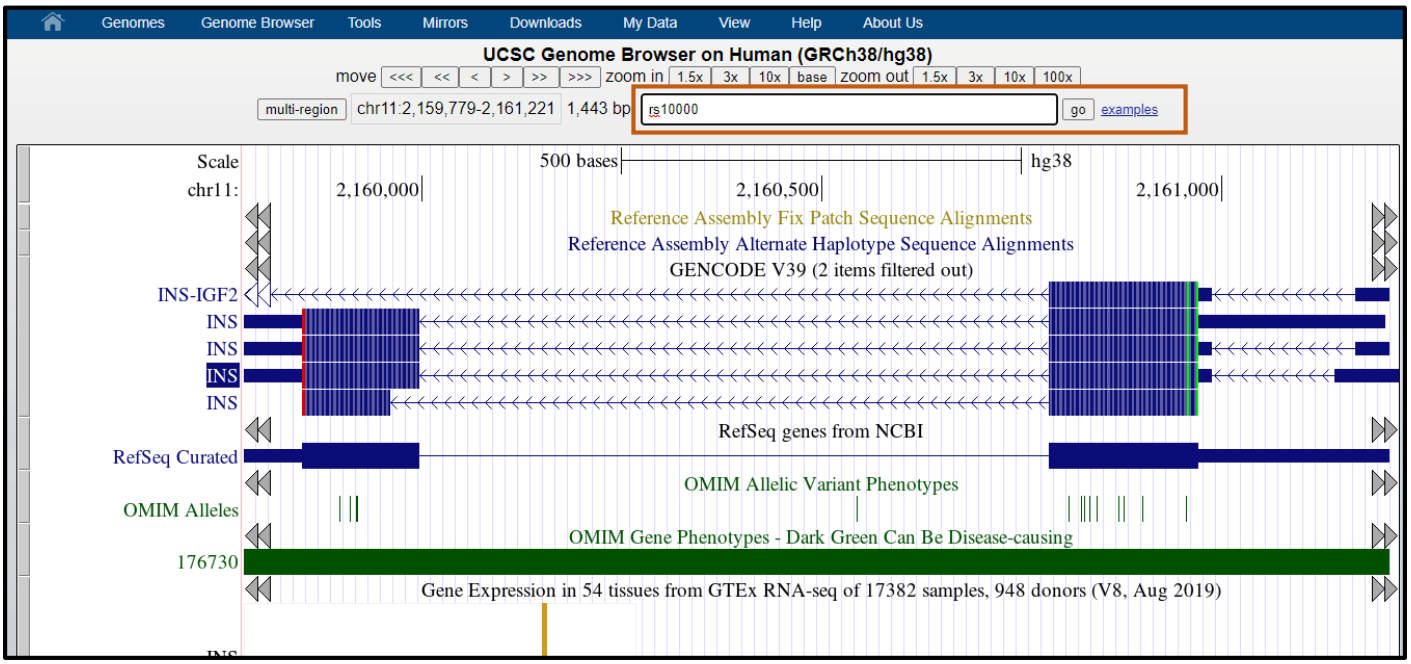


Fig8. Navigation by SNP id: rs10000

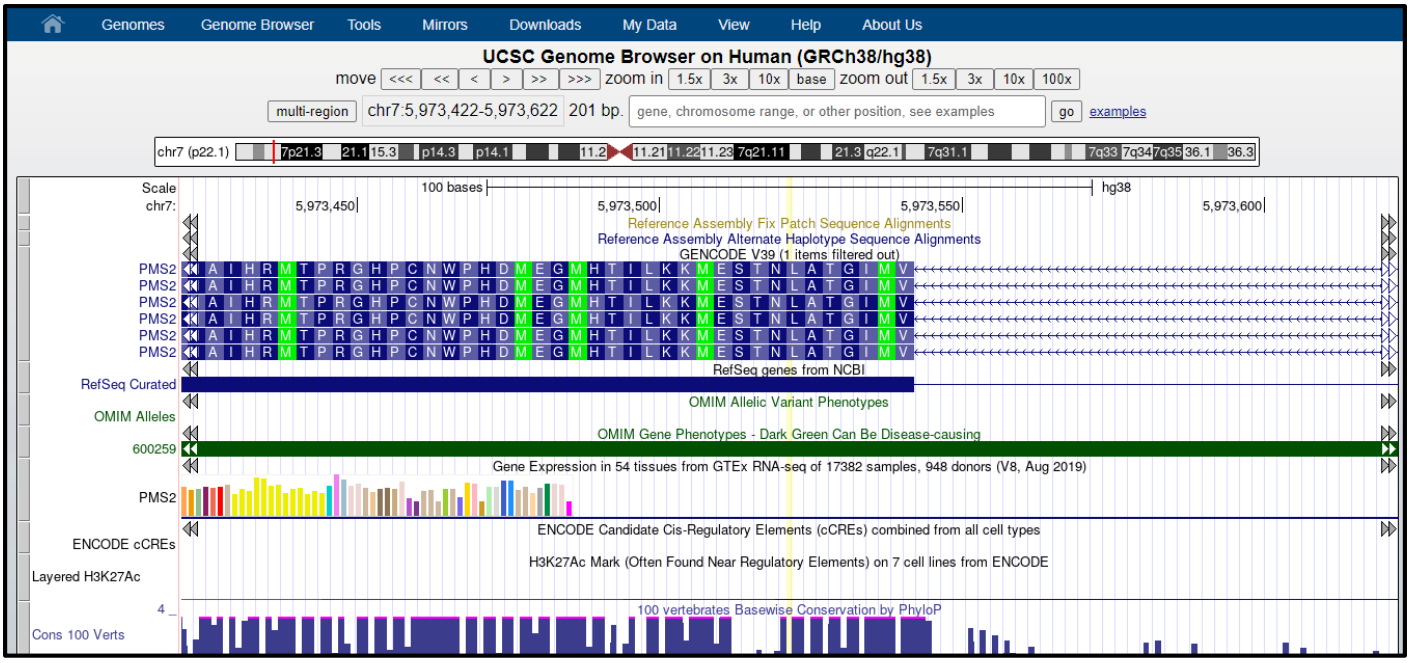


Fig8.1. Result for SNP id: rs10000

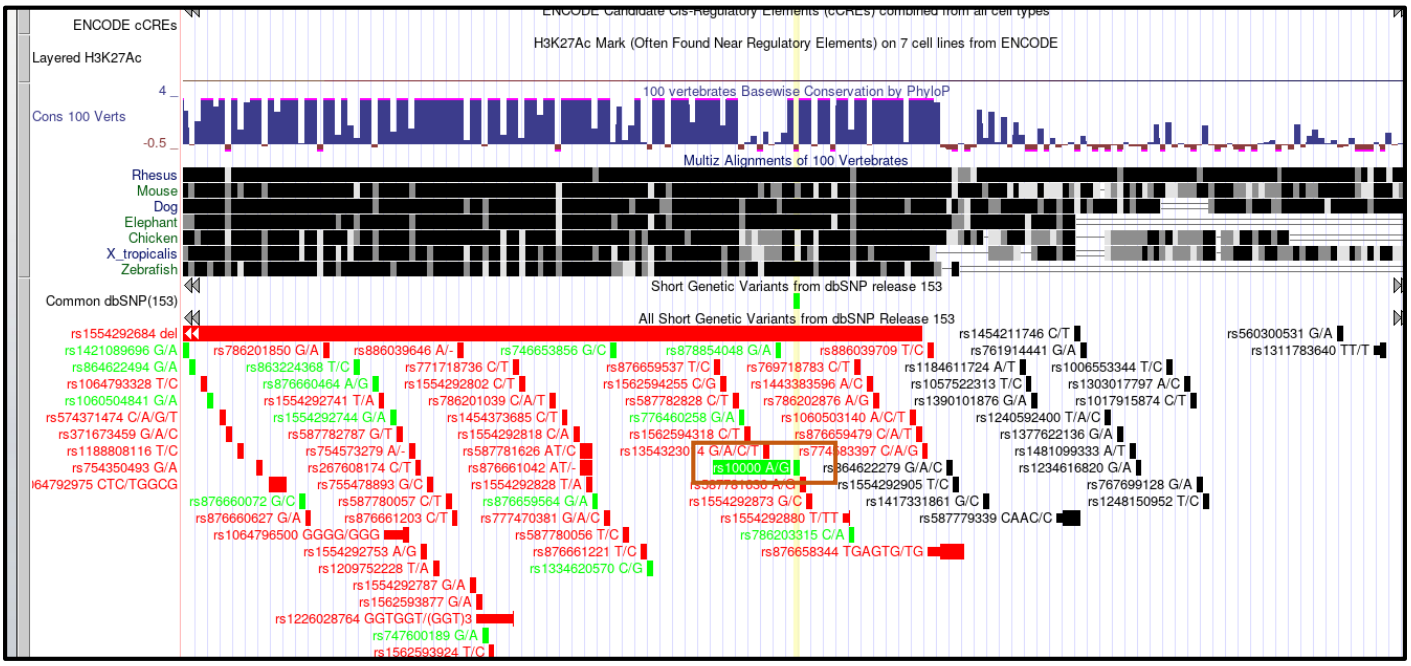


Fig8.2. Result for SNP id: rs10000

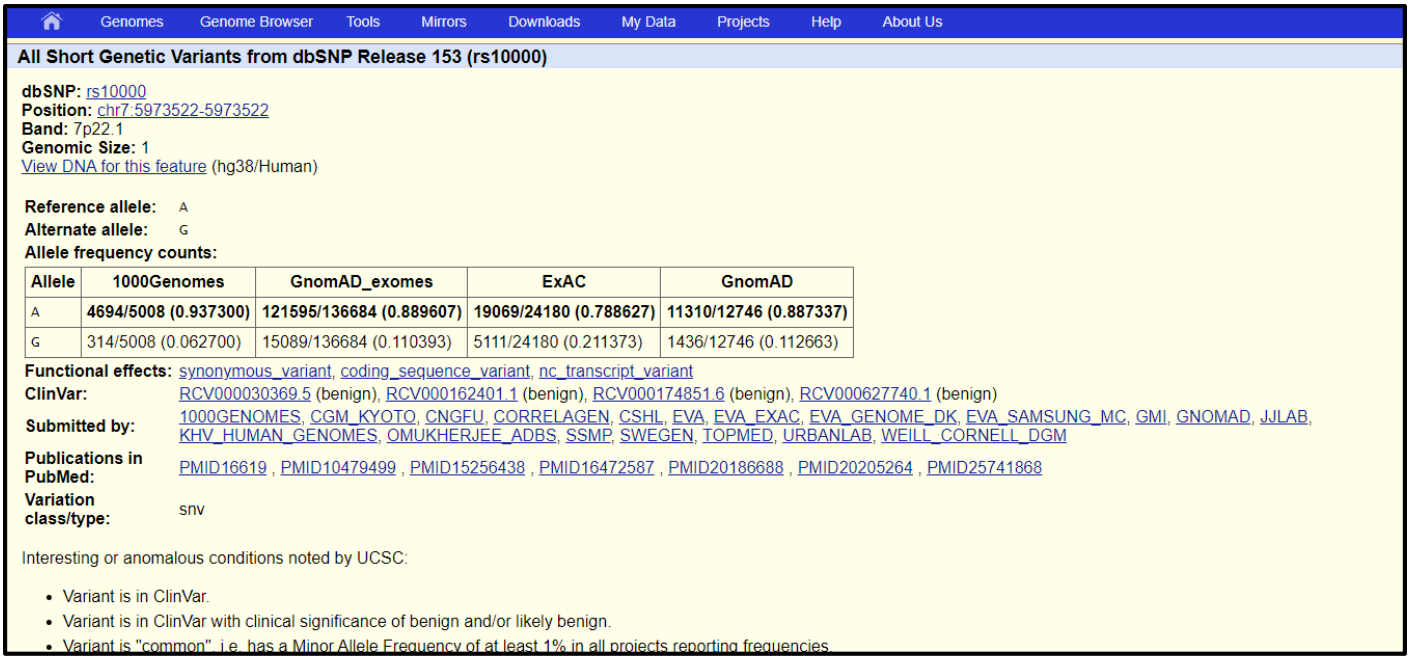


Fig8.3 Description for SNP id: rs10000

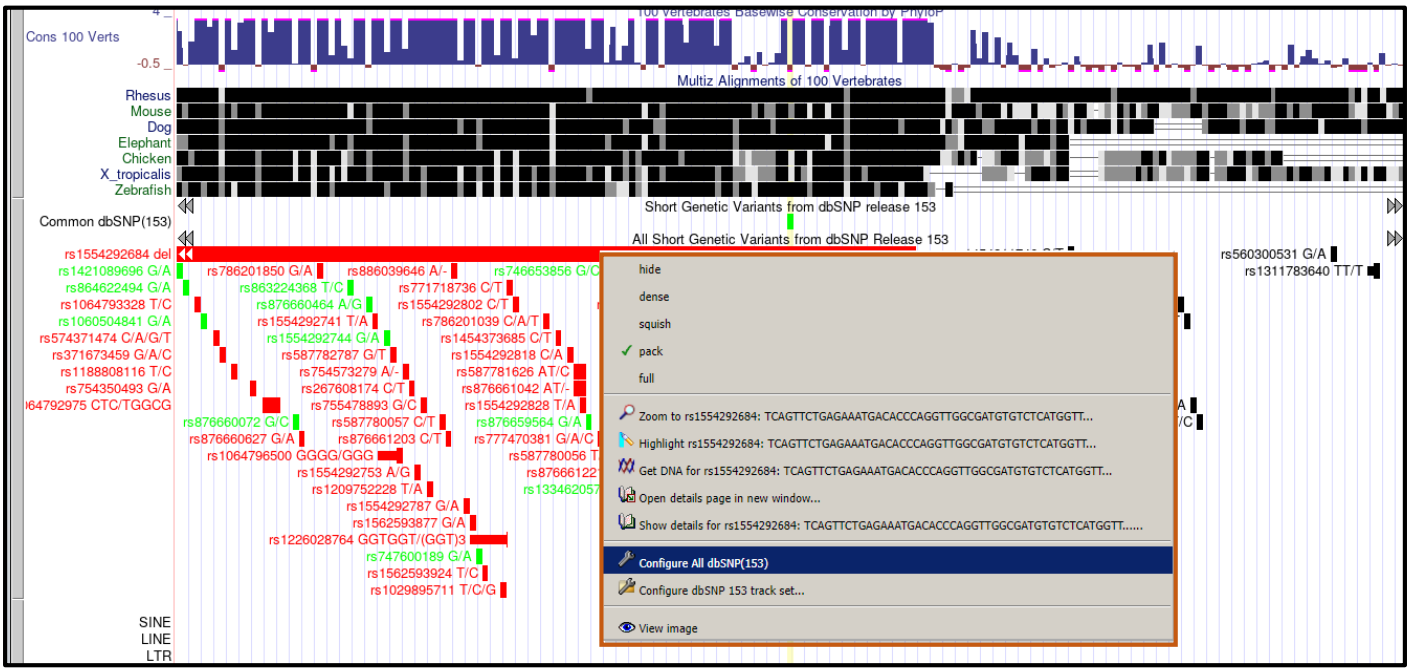


Fig8.4. Configuration option by right click

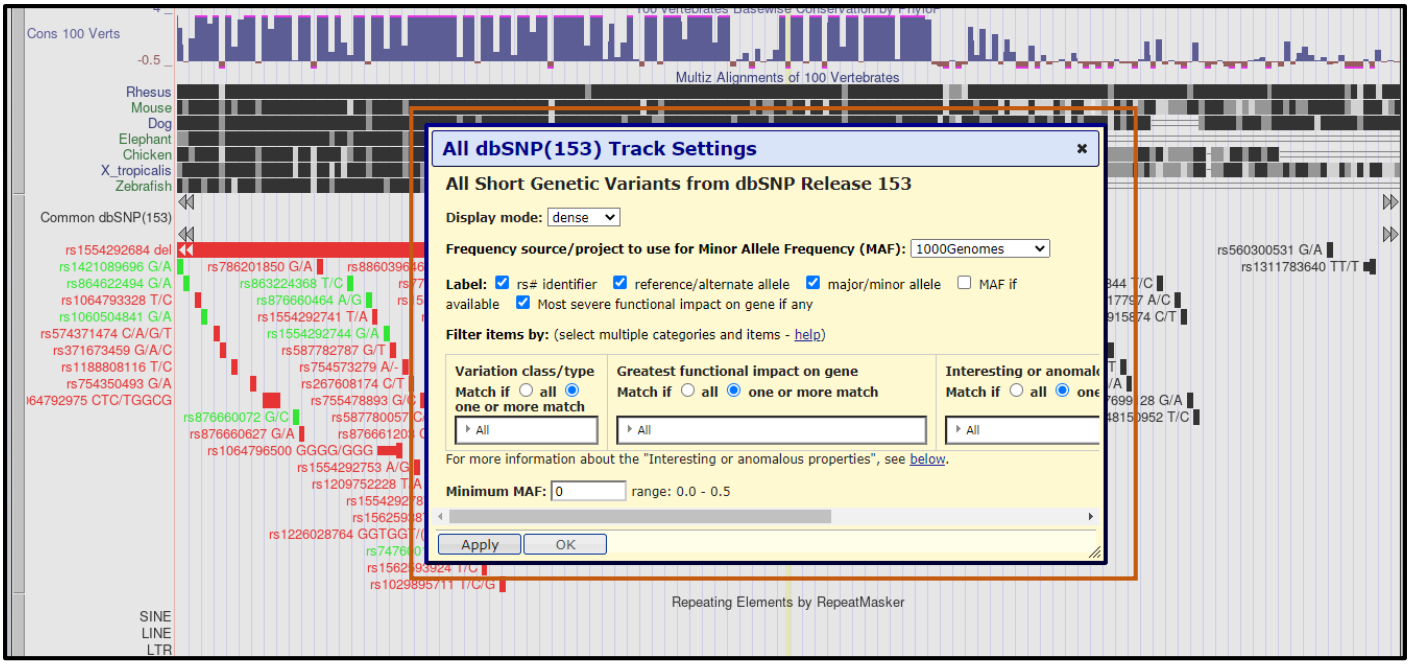


Fig8.5. Configurations applied

8

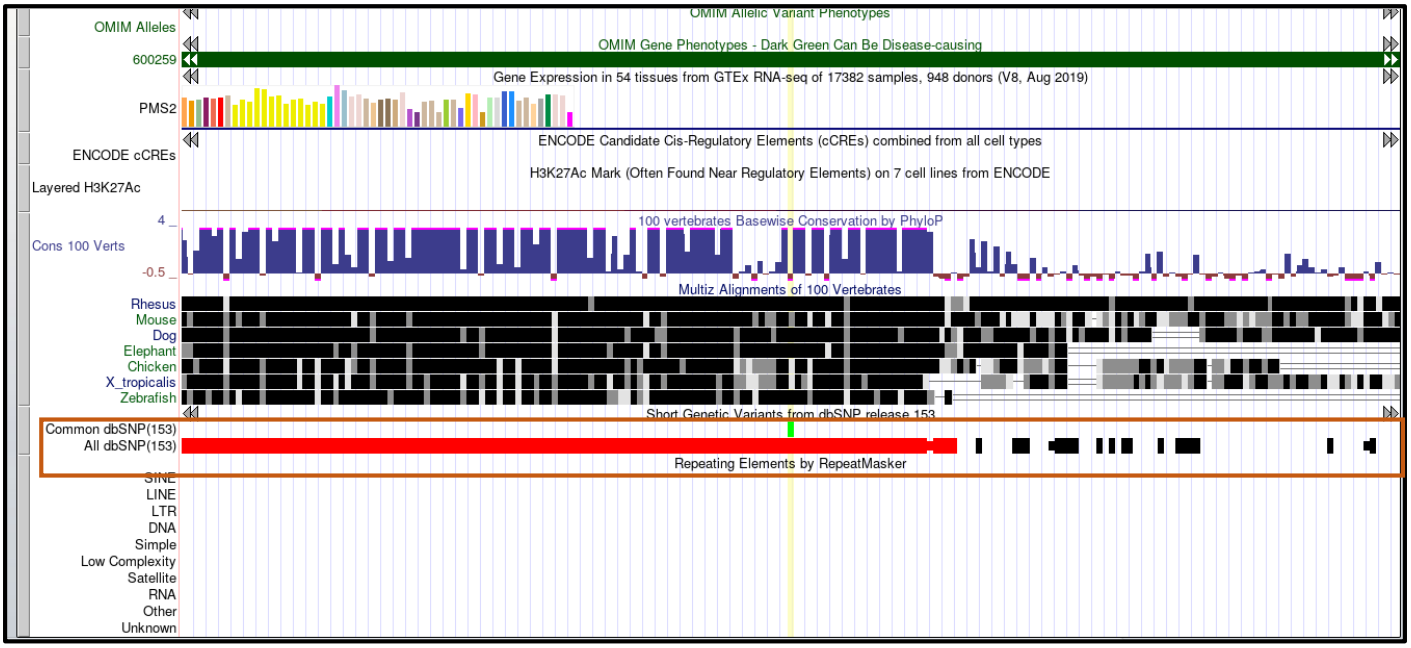


Fig.8.6. Result after configuration

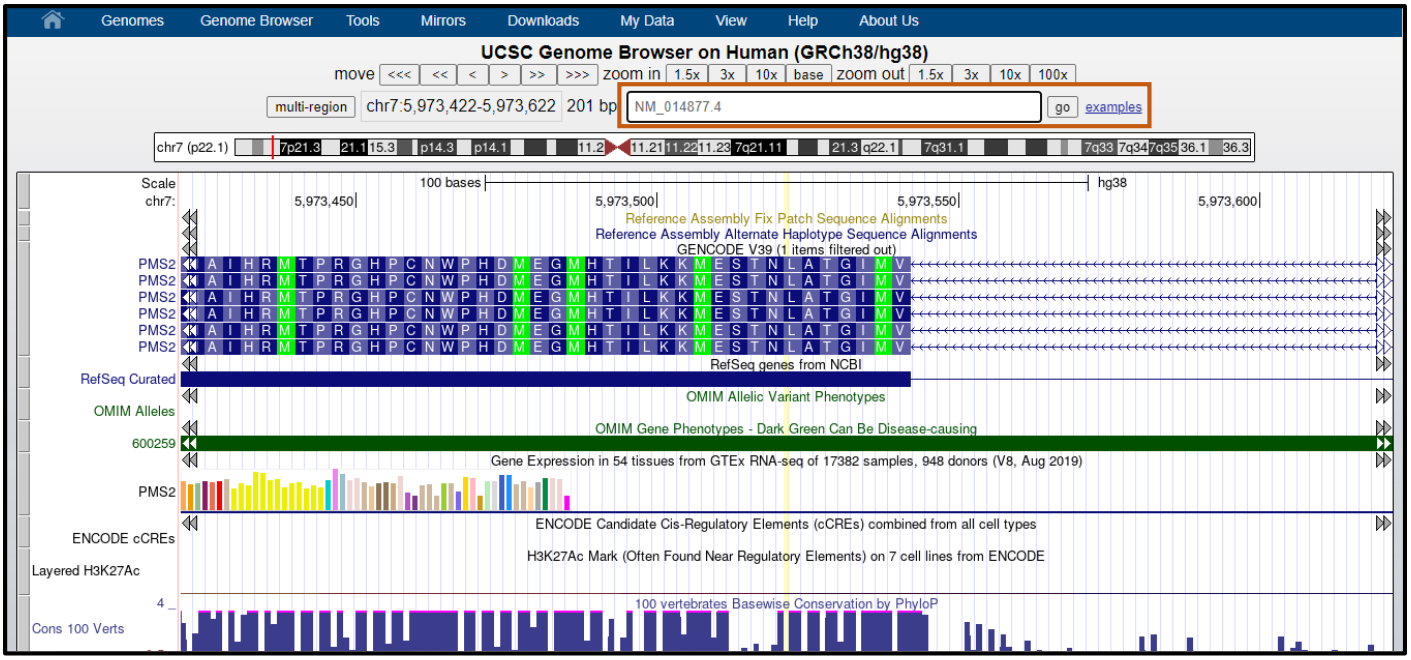


Fig9. Navigation by Ref_Seq: NM_014877.1 (HEL gene)

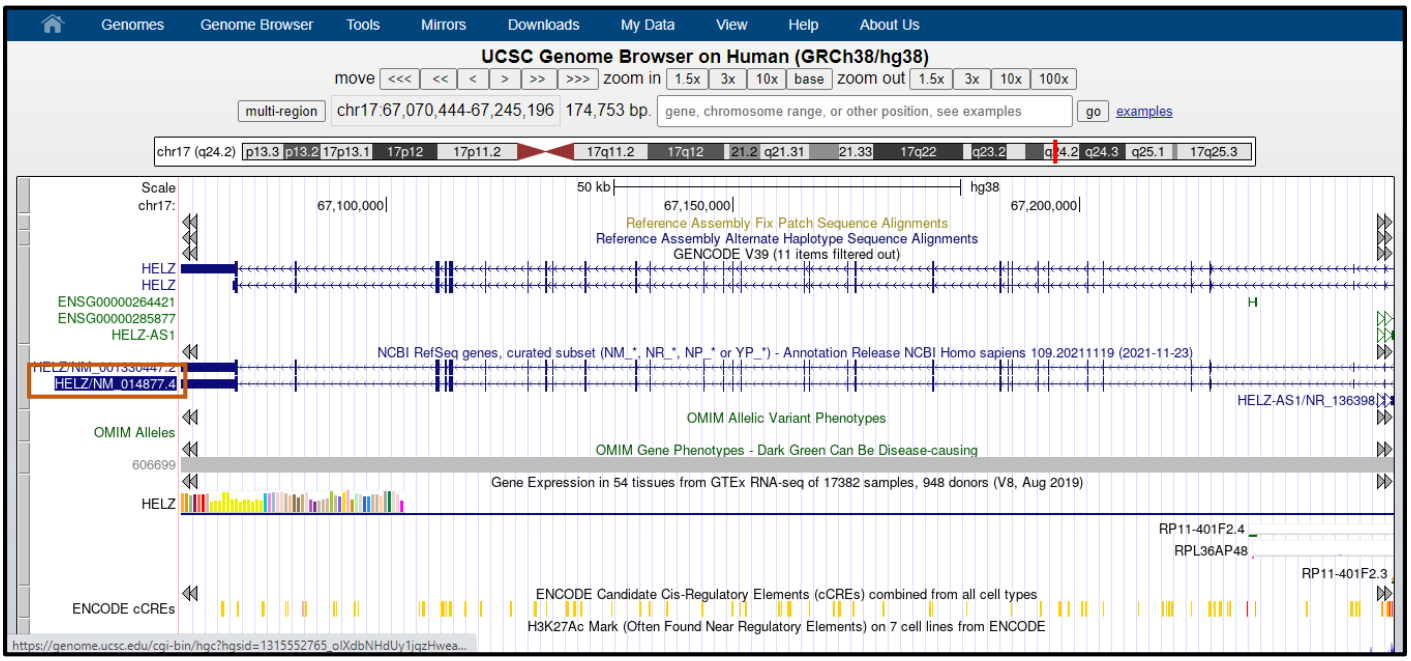


Fig9.1. Result for Ref_Seq: NM_014877.1 (HELZ gene)

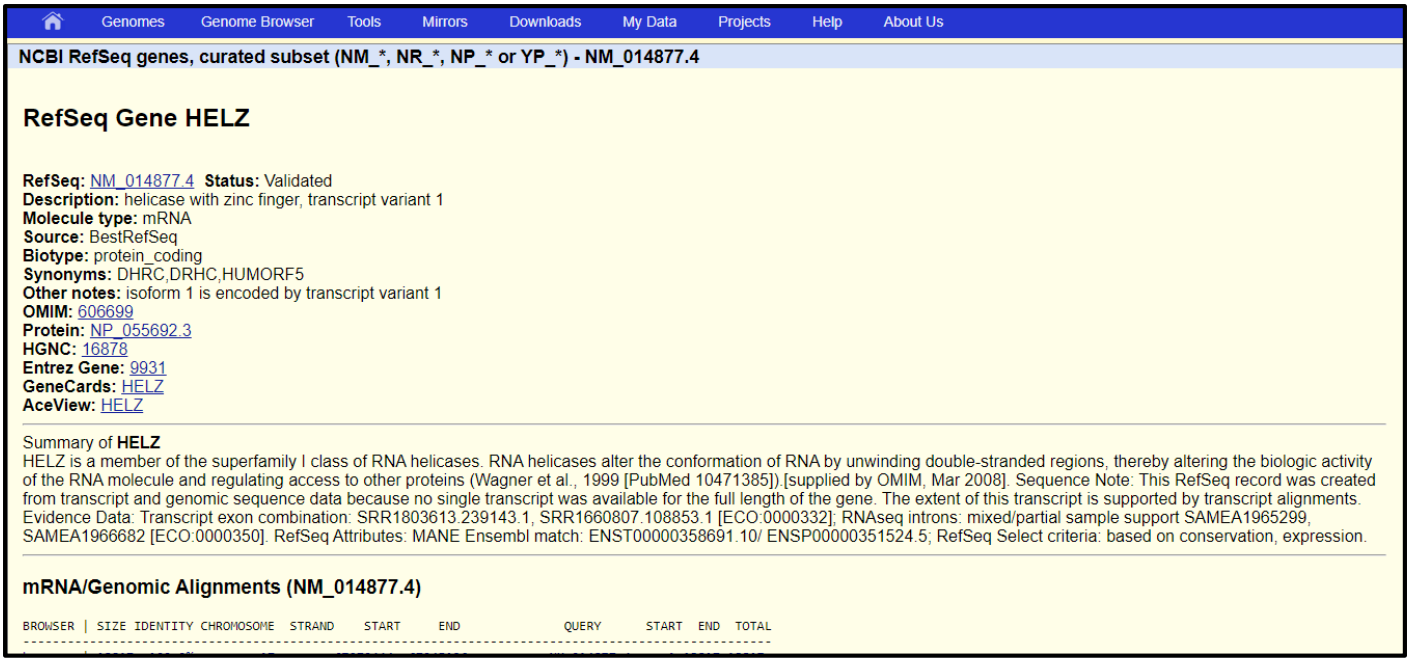


Fig9.2. Description for Gene HELZ

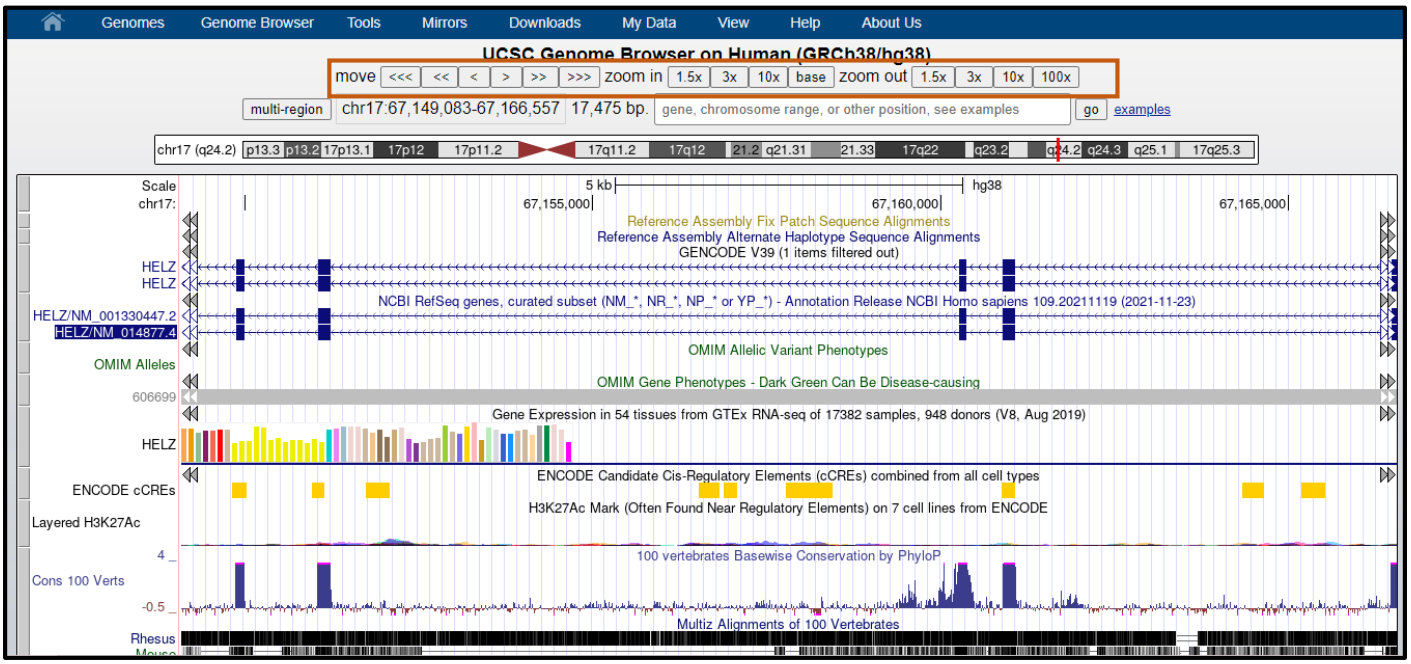


Fig9.3. Result after zooming 10x

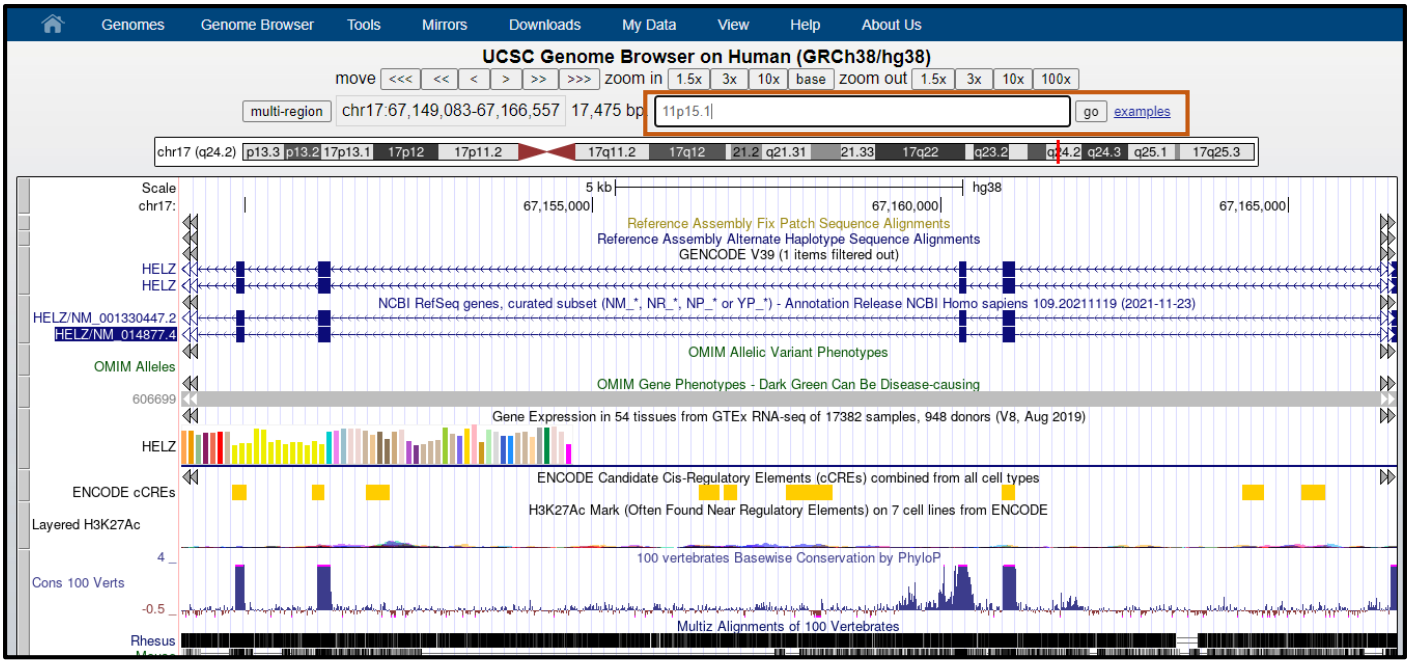


Fig10. Navigation by cytological band: 11p15.1

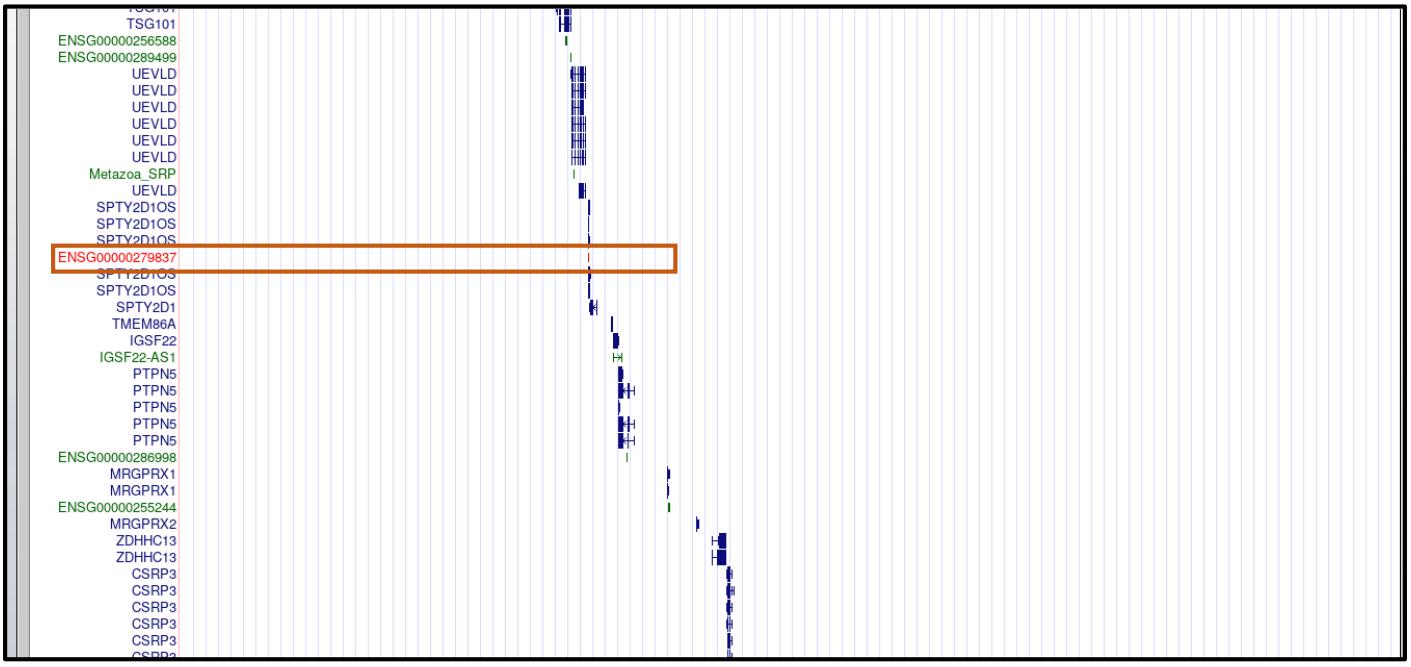


Fig10.1. Result for cytological band: 11p15.1

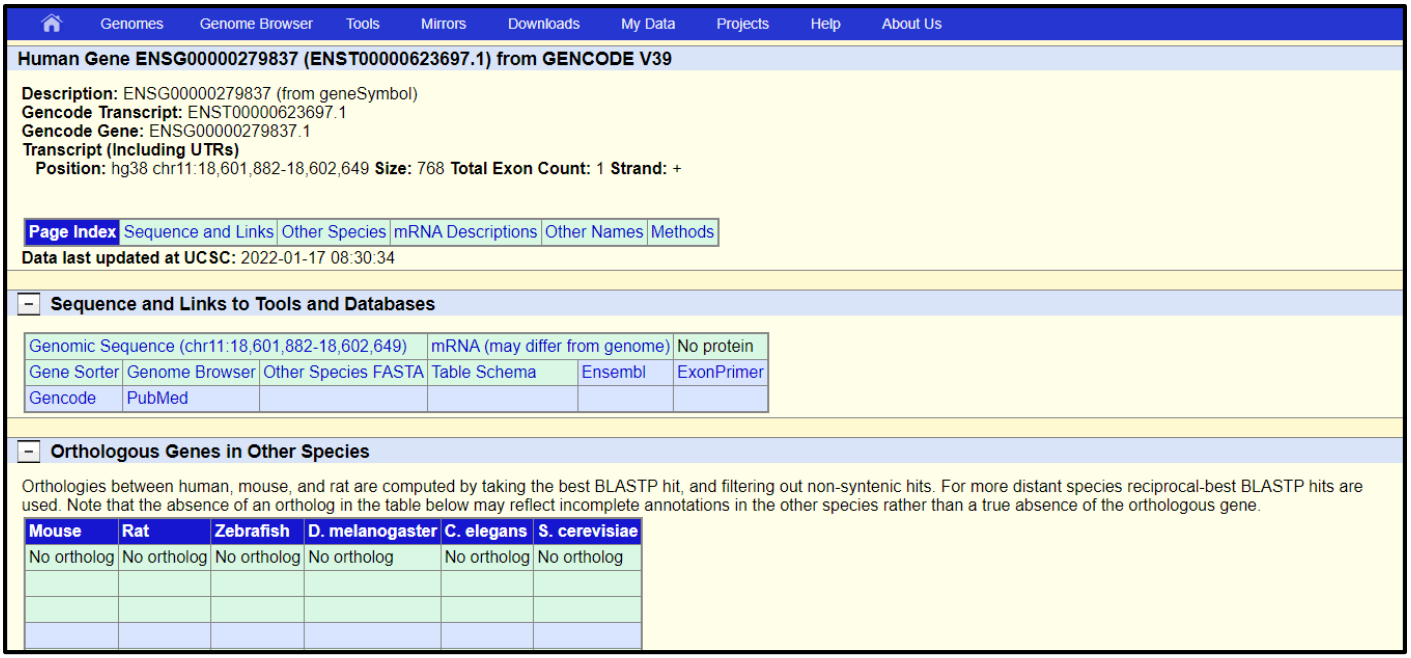


Fig10.2. Description for cytological band: 11p15.1

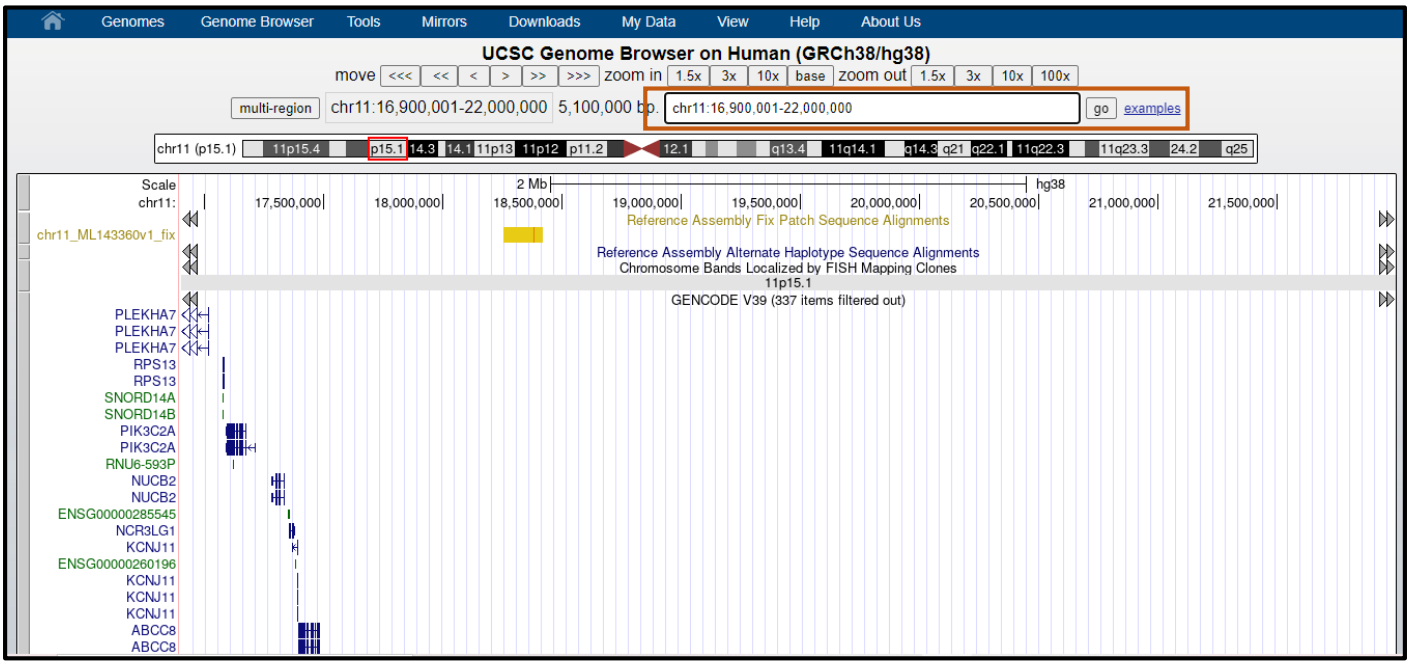


Fig11. Navigation by coordinates

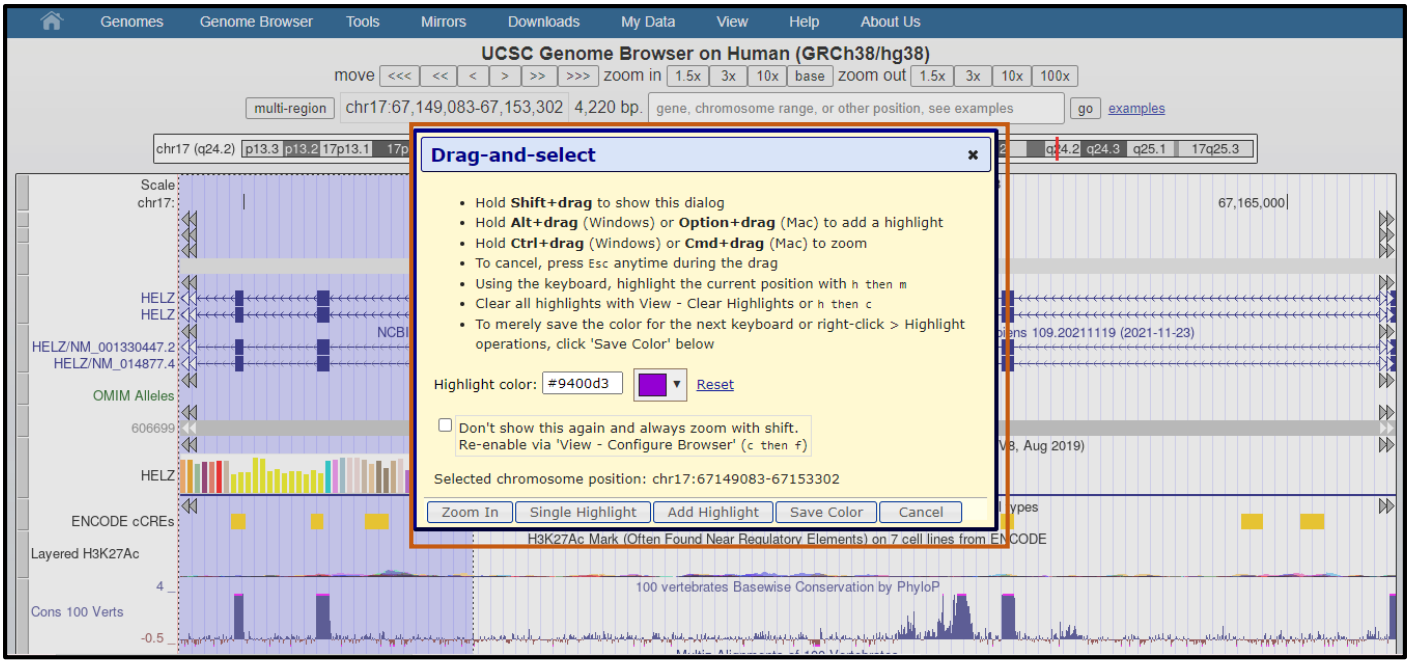


Fig12. Drag and select option for configuration

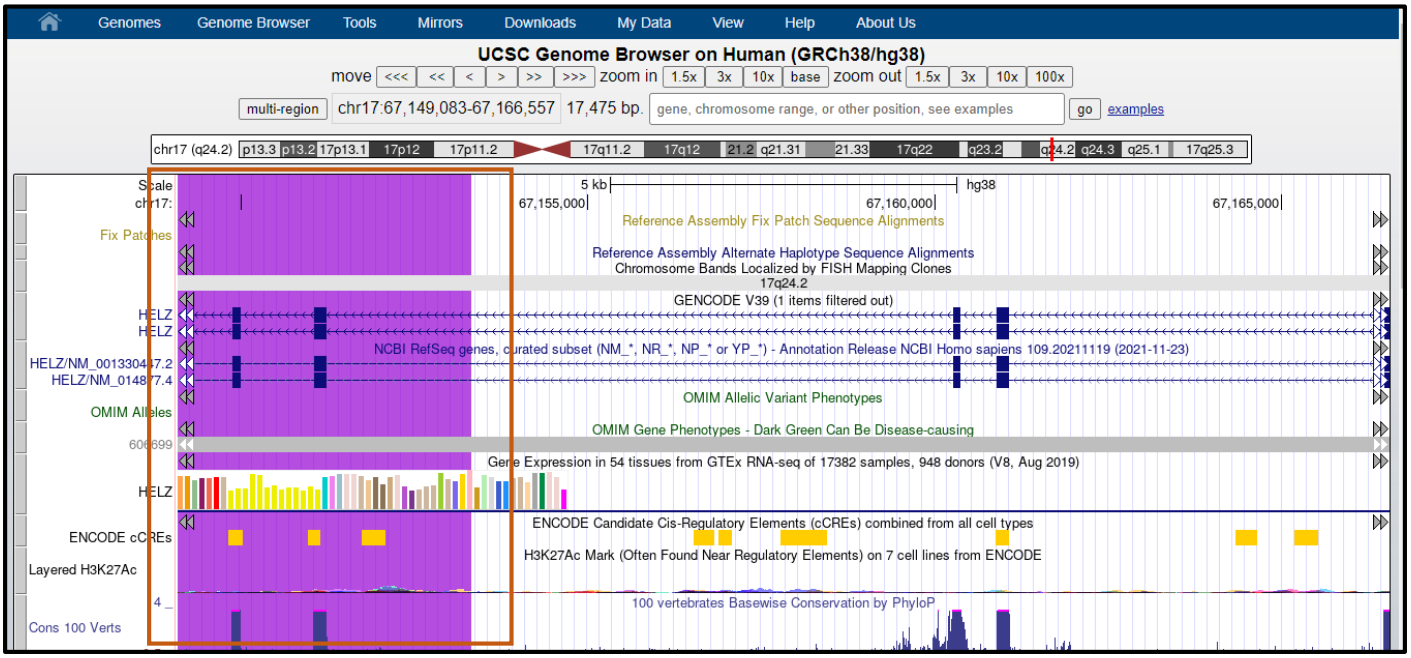


Fig12.1. Result after configuration

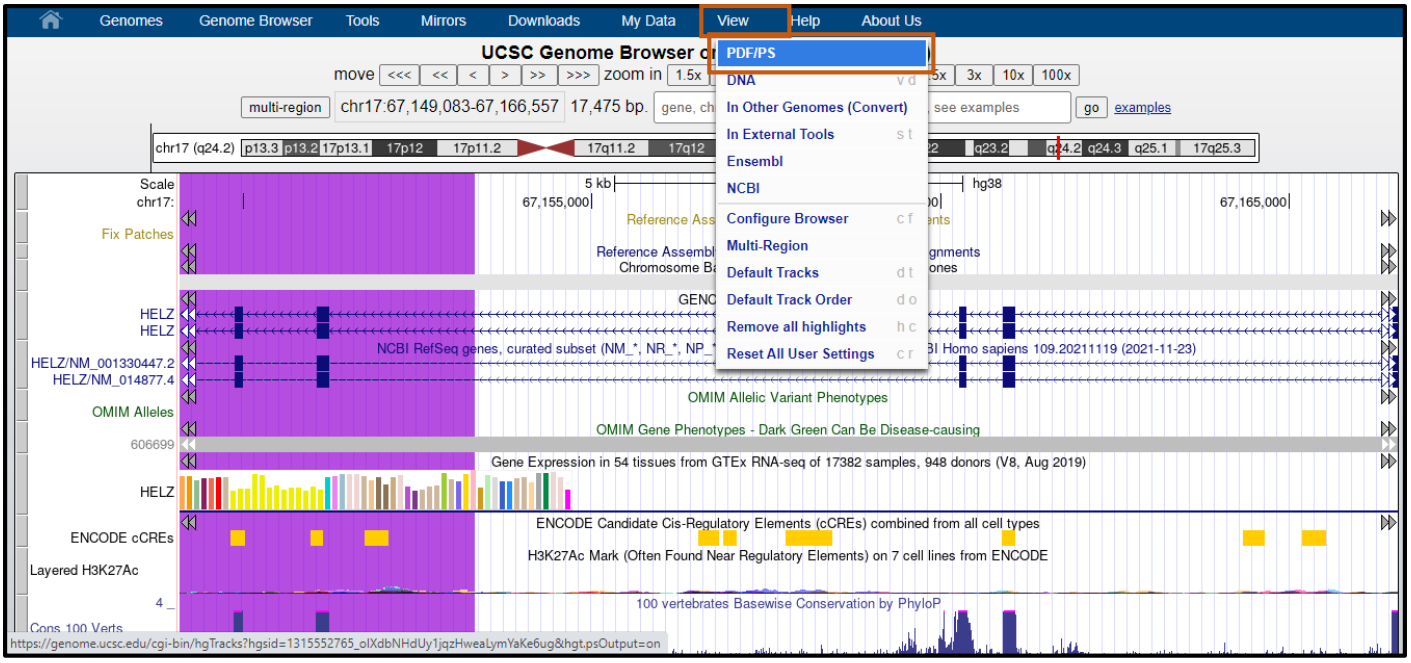


Fig13. Steps to export result as PDF

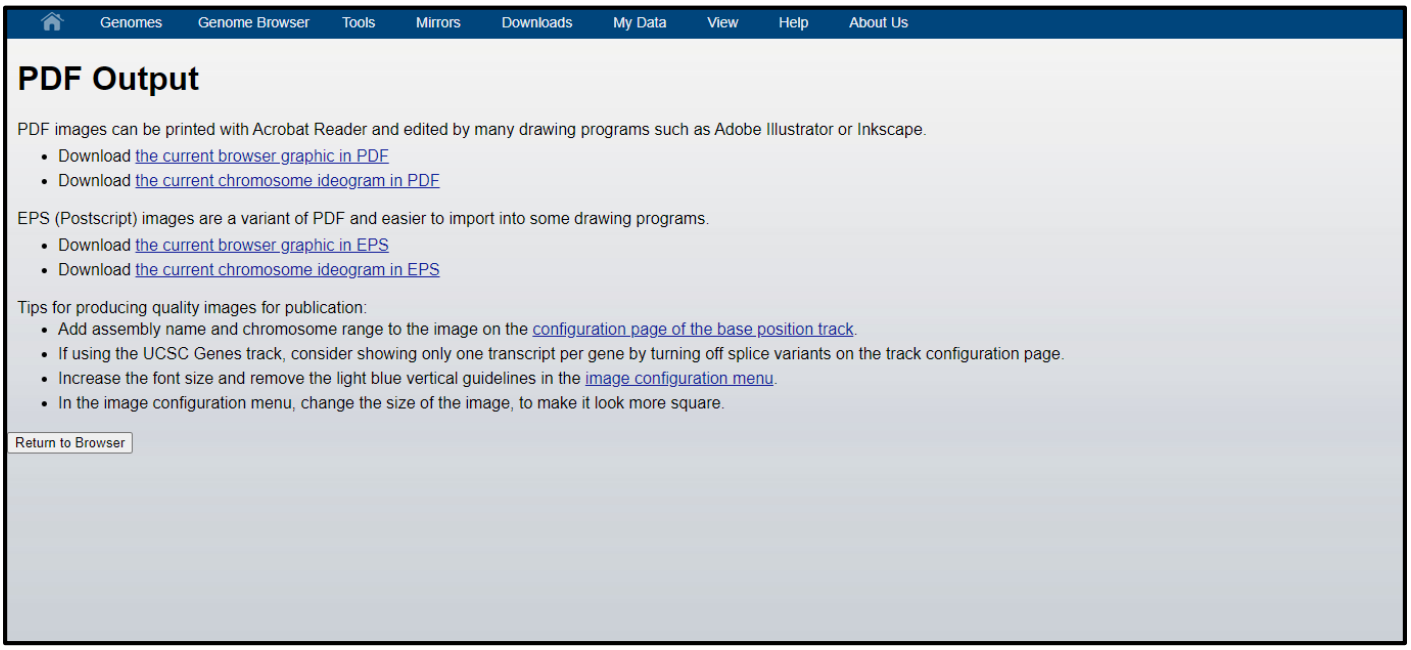


Fig13.1. Option to download PDF

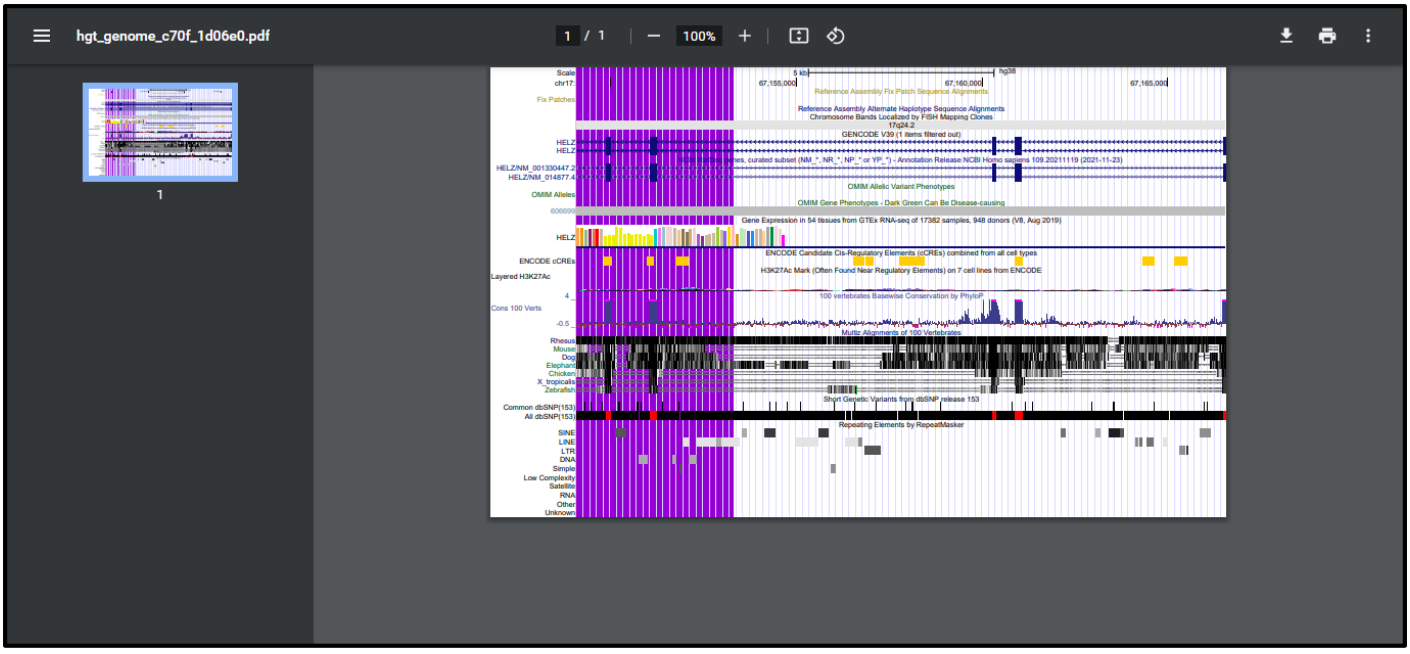


Fig13.2. Result in PDF format

RESULT:

UCSC genome browser was used for setting for GRCH/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=1315552765_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

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

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

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

DATE: 28/03/22

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:

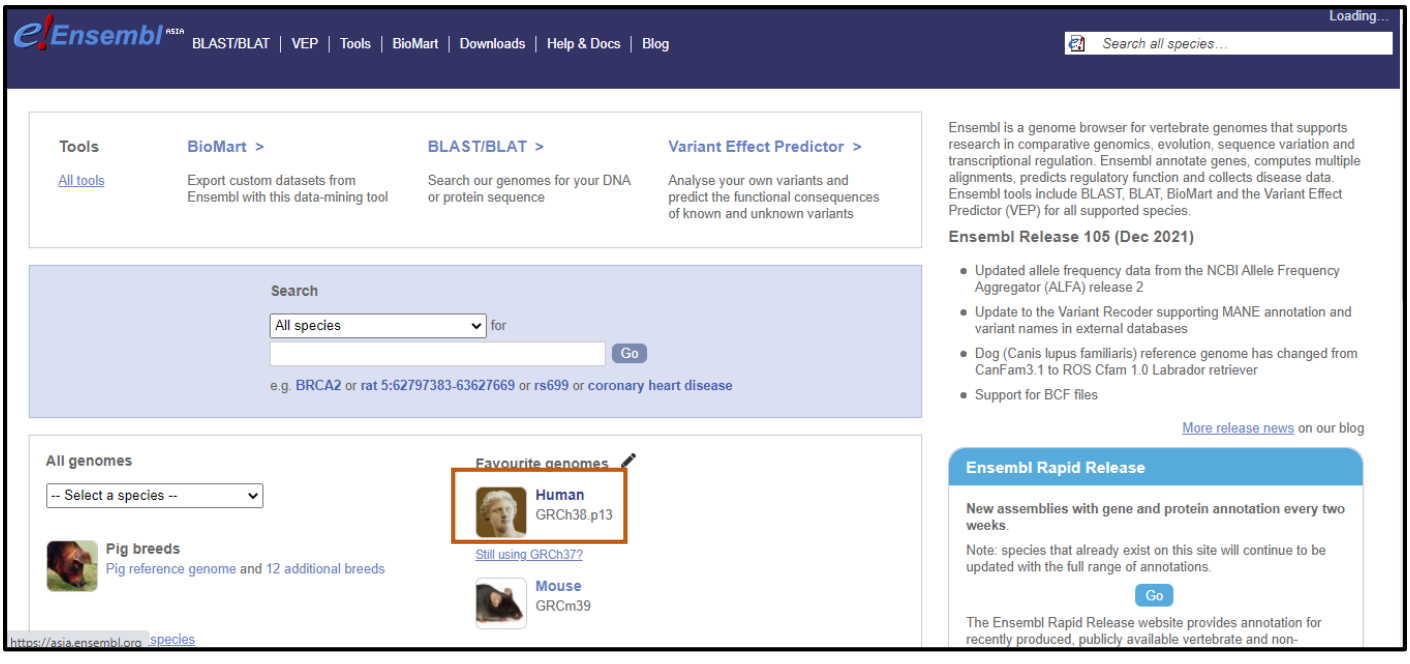


Fig1. Homepage for Ensembl

Fig2. Homepage for Human (GRCh38.p13) genome assembly

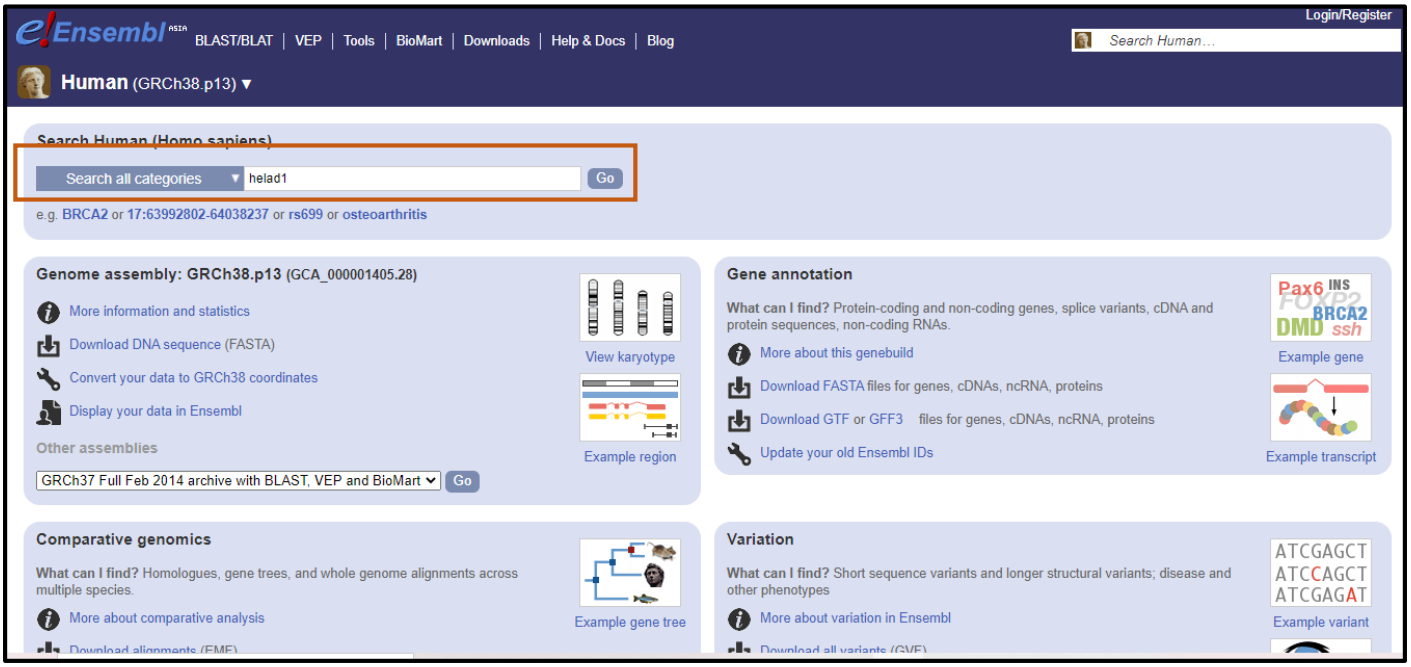


Fig3. Search for helad1 gene

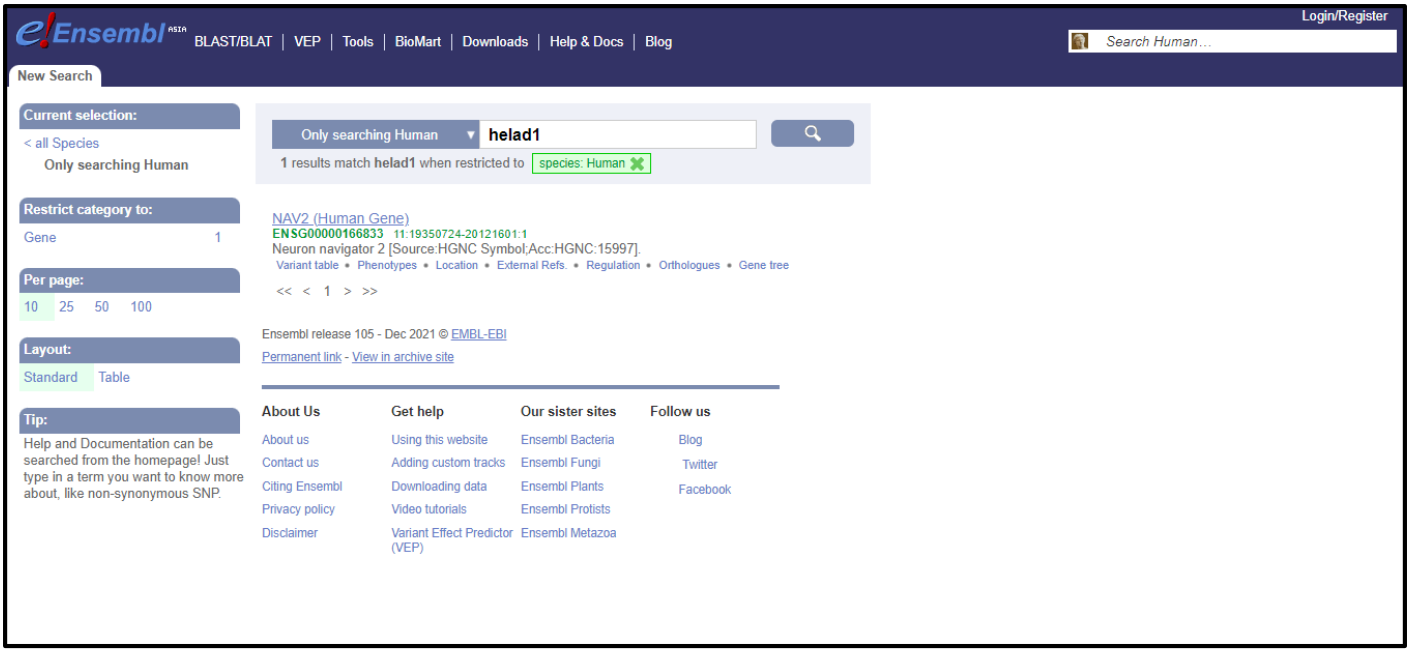


Fig4. Hit page for helad1 gene

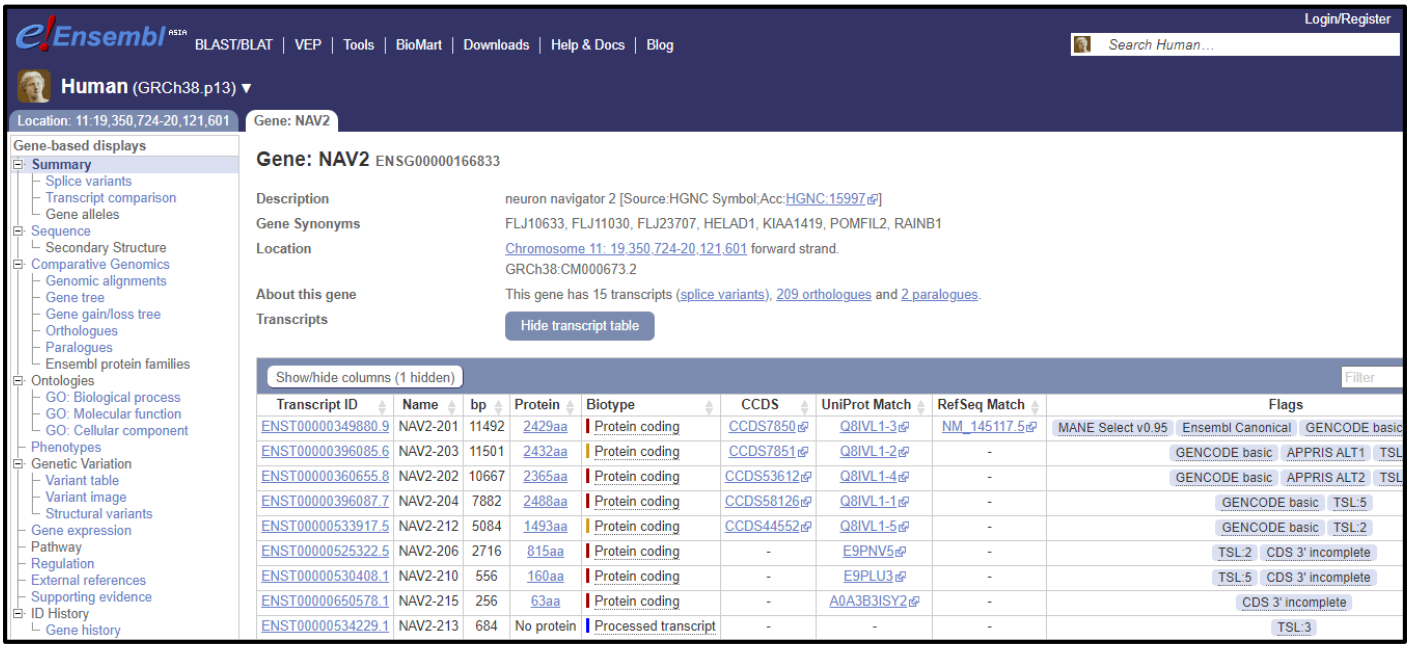


Fig5. Result for NAV2 gene

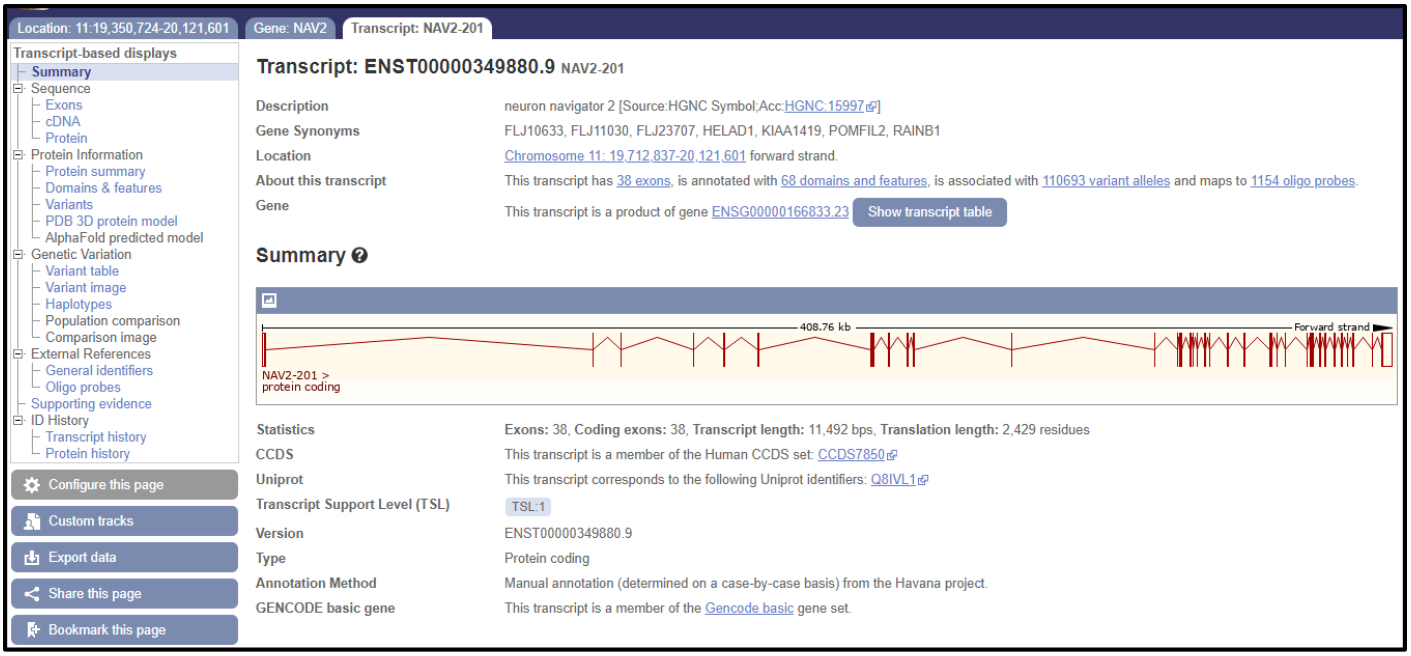
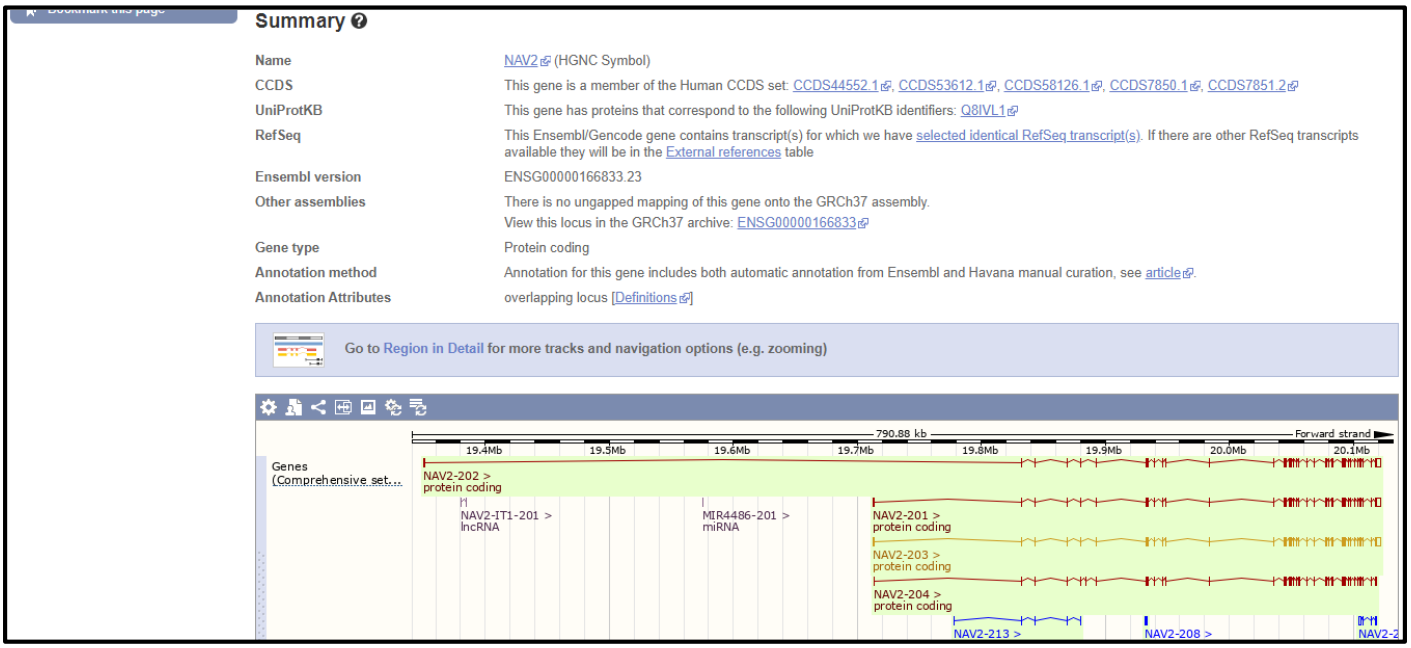


Fig5.1. Result for NAV2 gene



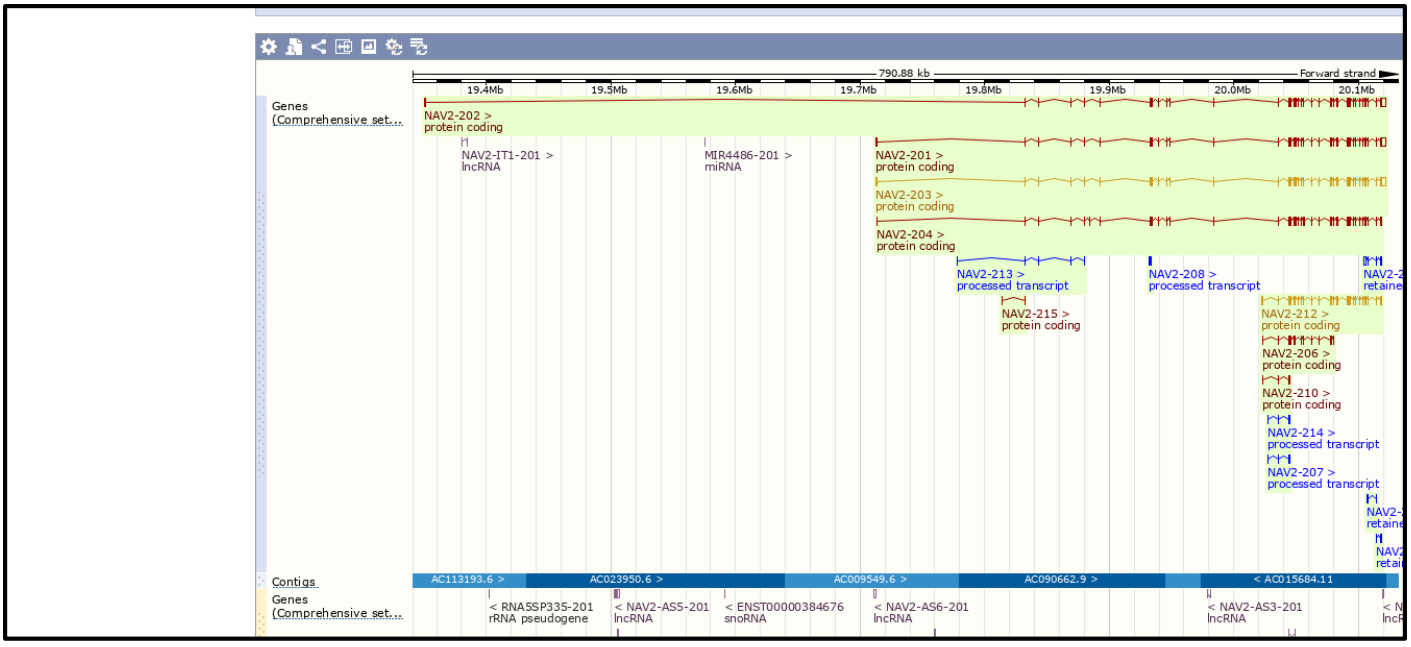


Fig5.2. Tracks information

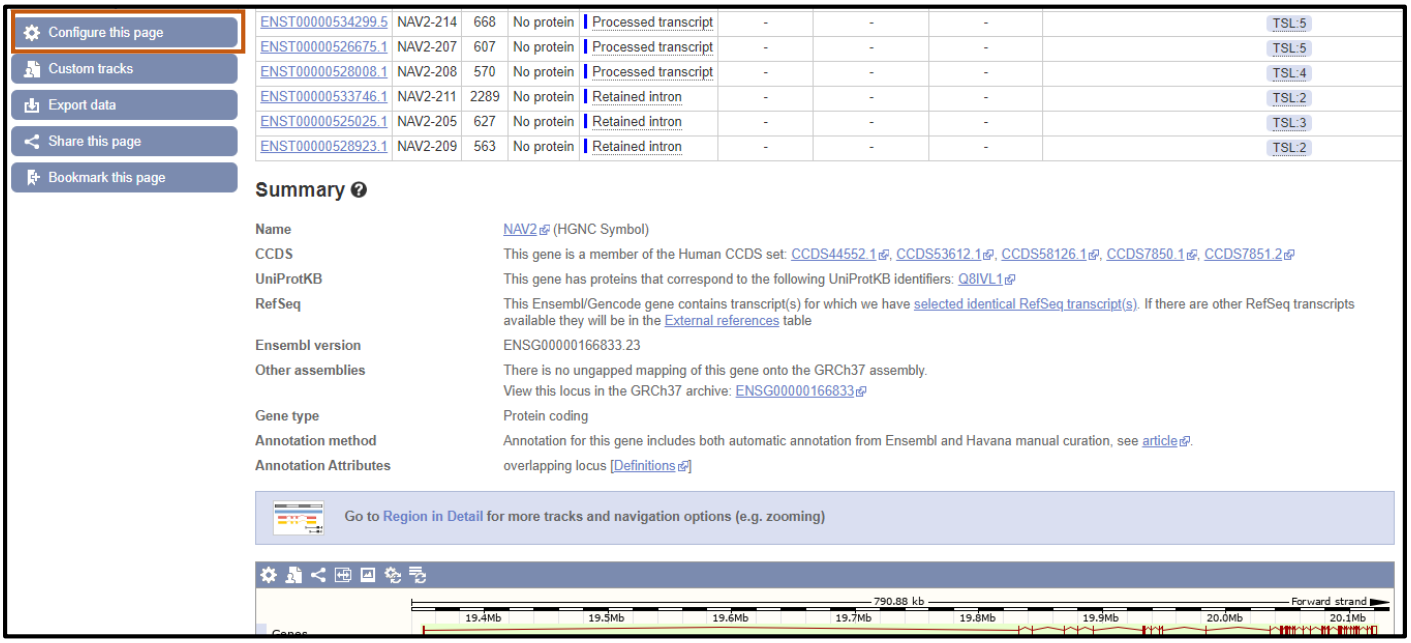


Fig6. Steps to configure tracks

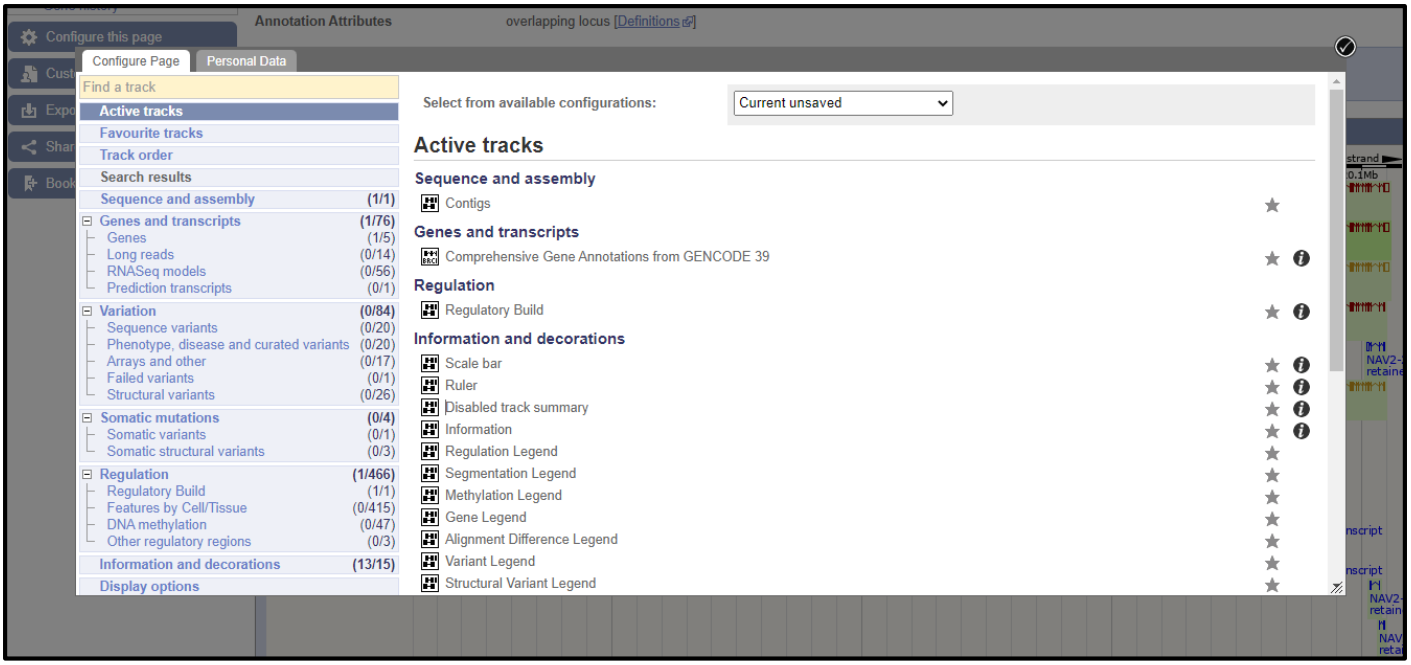


Fig6.1. Option for tracks configuration

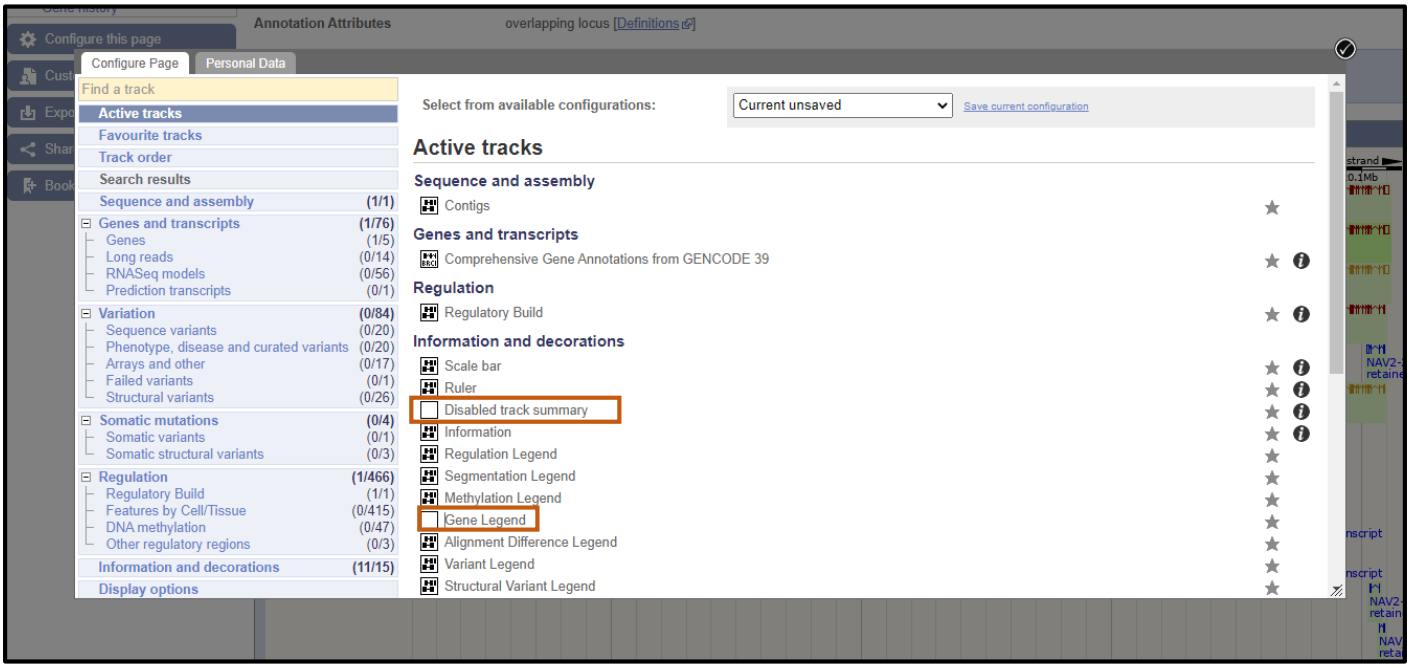


Fig6.2. Tracks configuration

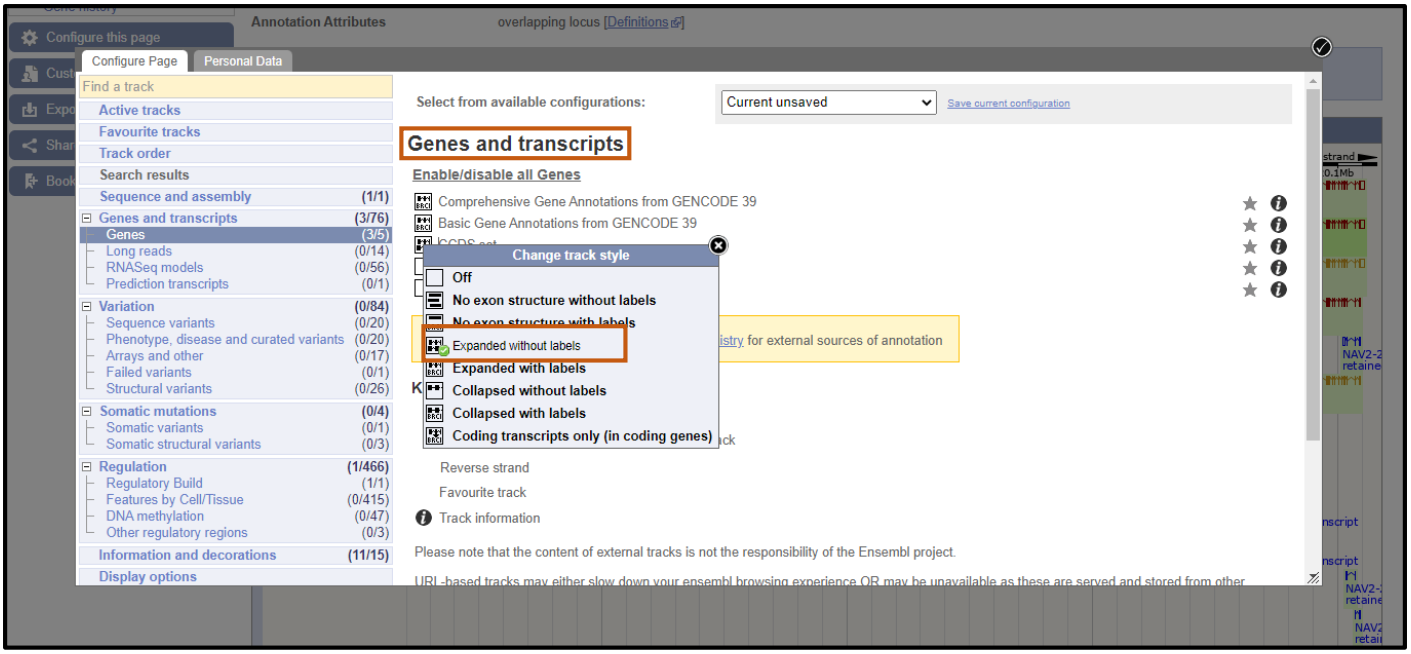


Fig6.3. Tracks configuration

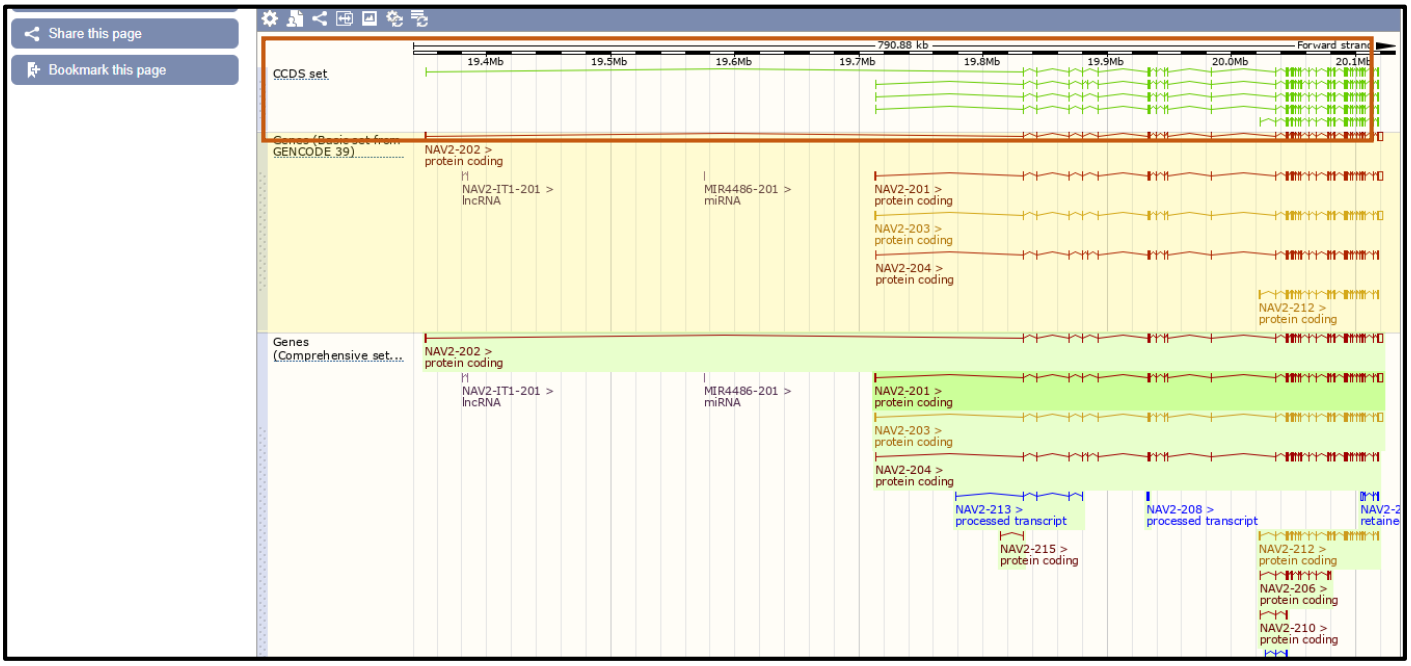


Fig6.4. Updated result after track configuration

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
4. *helad1 - Search - Homo_sapiens - Ensembl genome browser 105*. (2021b). Ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Human/Search/Results?q=helad1
5. *Summary - Homo sapiens - Ensembl genome browser 100*. (n.d.). Uswest.ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Homo_sapiens/Gene/Summary?db=core
6. *Summary - Homo sapiens - Ensembl genome browser 100*. (n.d.). Uswest.ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000166833;r=11:19350724-20121601;t=ENST00000349880
7. *Summary - Homo sapiens - Ensembl genome browser 100*. (n.d.). Uswest.ensembl.org. Retrieved March 28, 2022, from https://asia.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000166833;r=11:19350724-20121601;t=ENST00000349880

DATE: 28/03/22

WEBLEM 9c

Genome Data Viewer

(URL: https://www.ncbi.nlm.nih.gov/genome/gdv/)

AIM:

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

INTRODUCTION:

GDV 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:

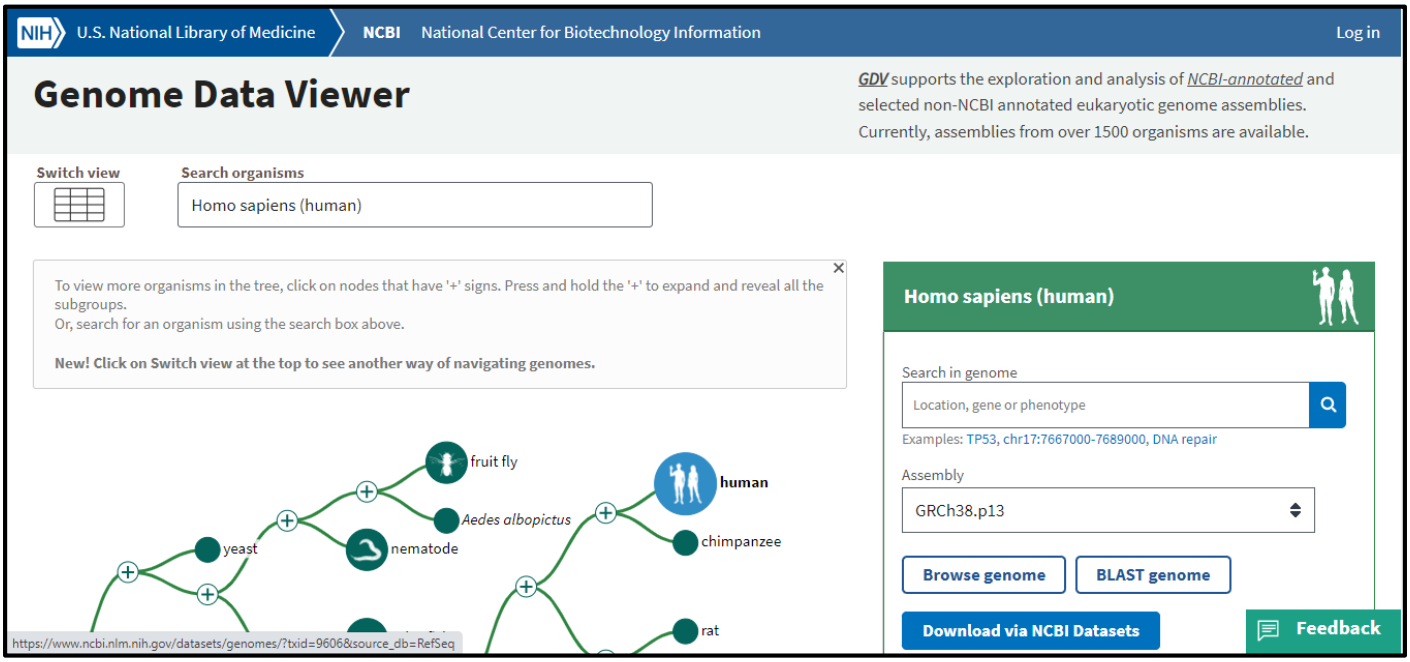


Fig1. Homepage for Genome Data Viewer

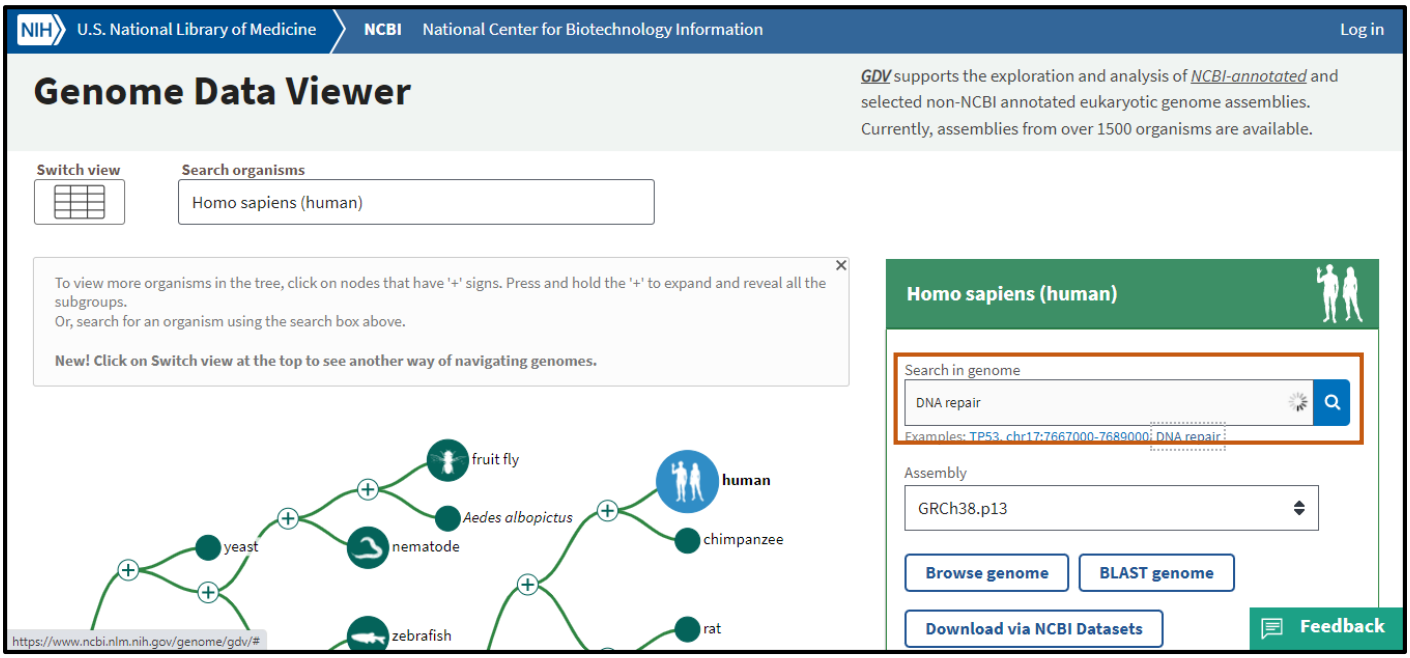


Fig2. Search for DNA repair in GRCh38.p13 assembly

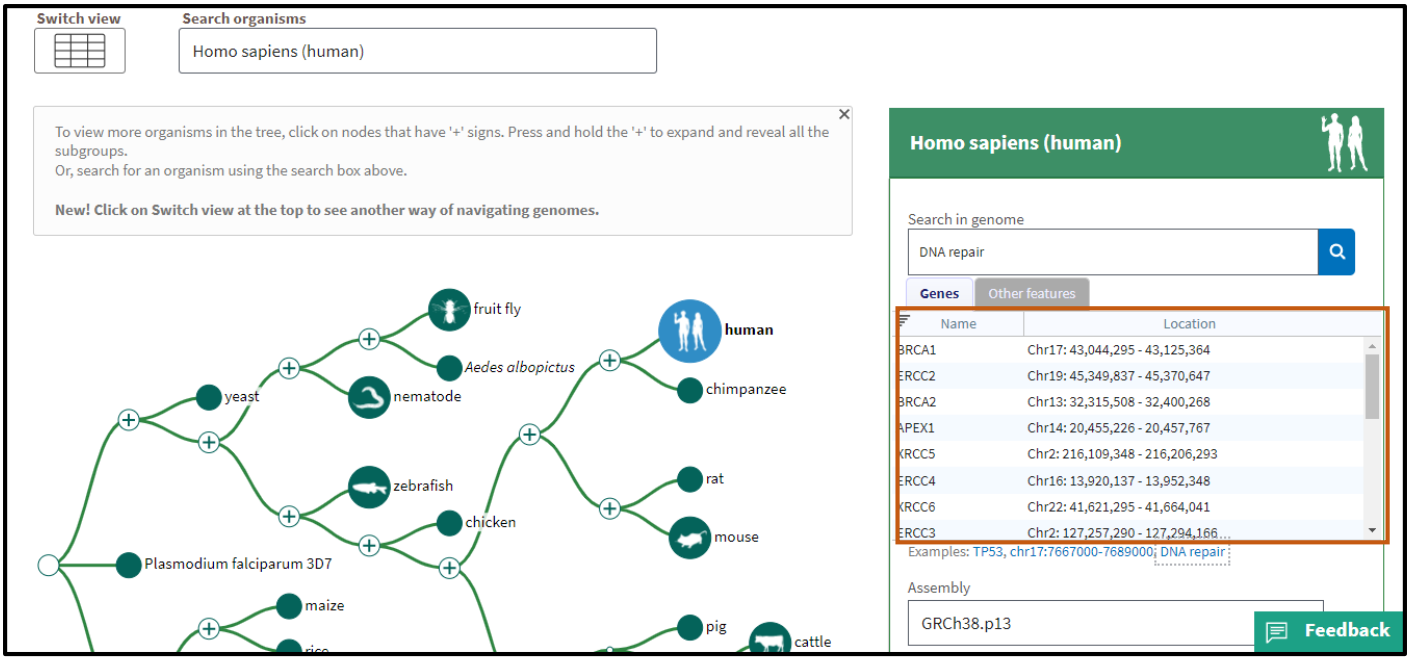


Fig3. Results for DNA repair genes

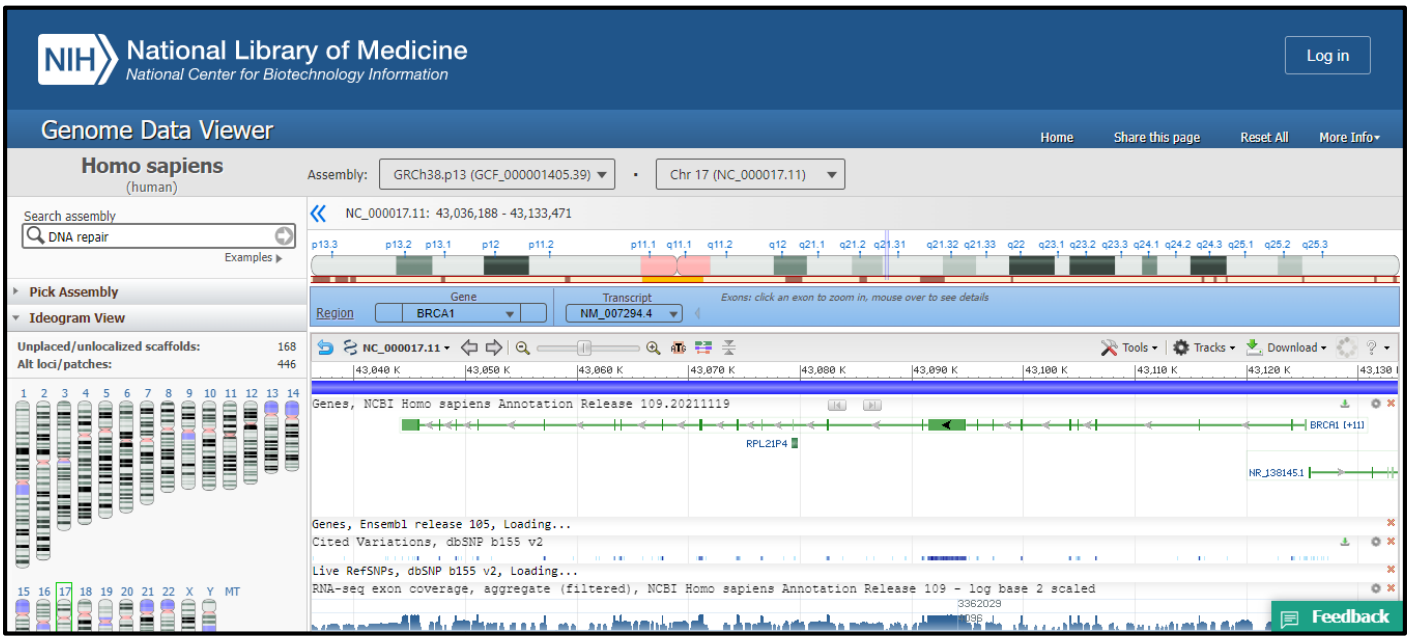


Fig4. Result for BRCA1 gene

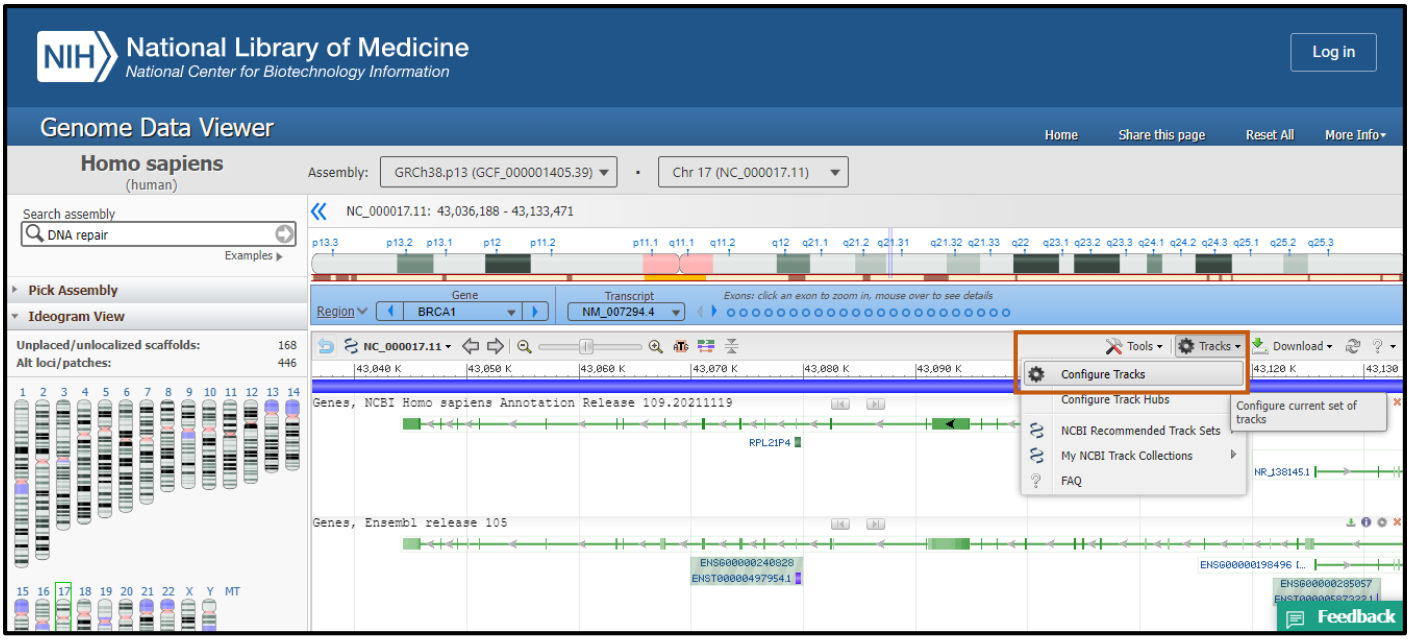


Fig5. Steps to configure tracks

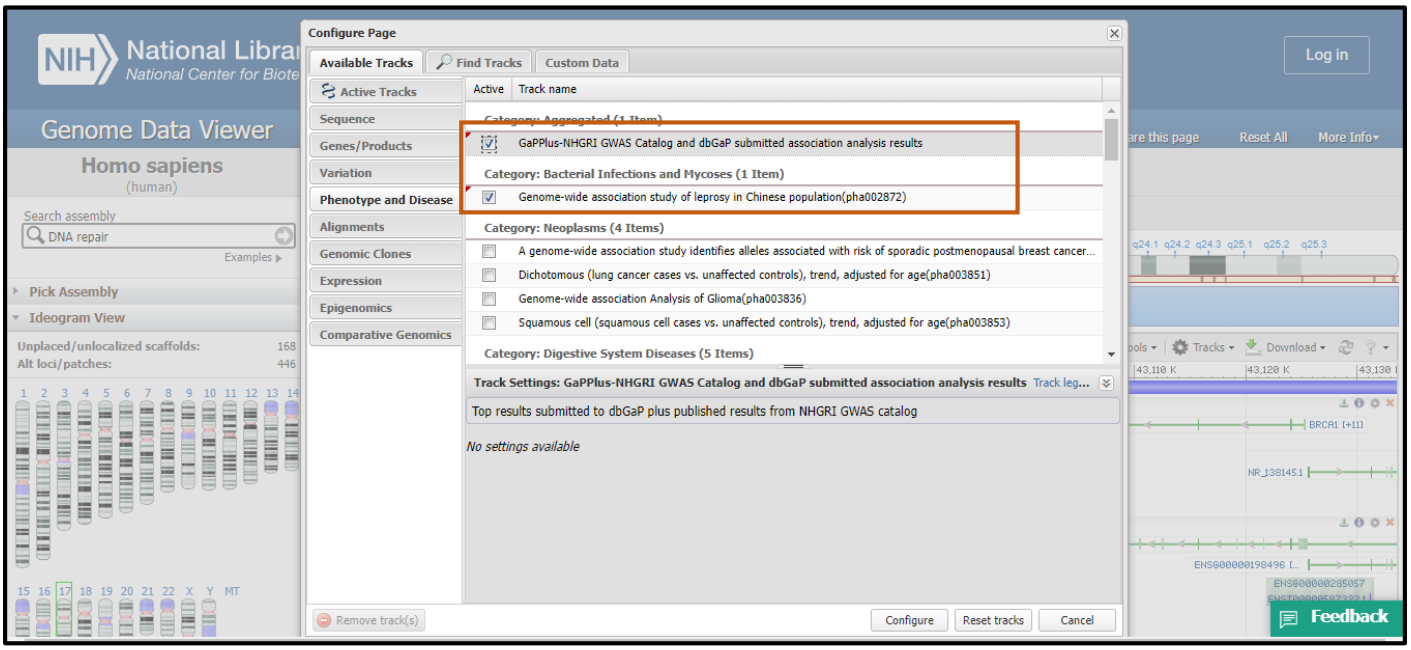


Fig5.1. Configuration page for phenotype and disease

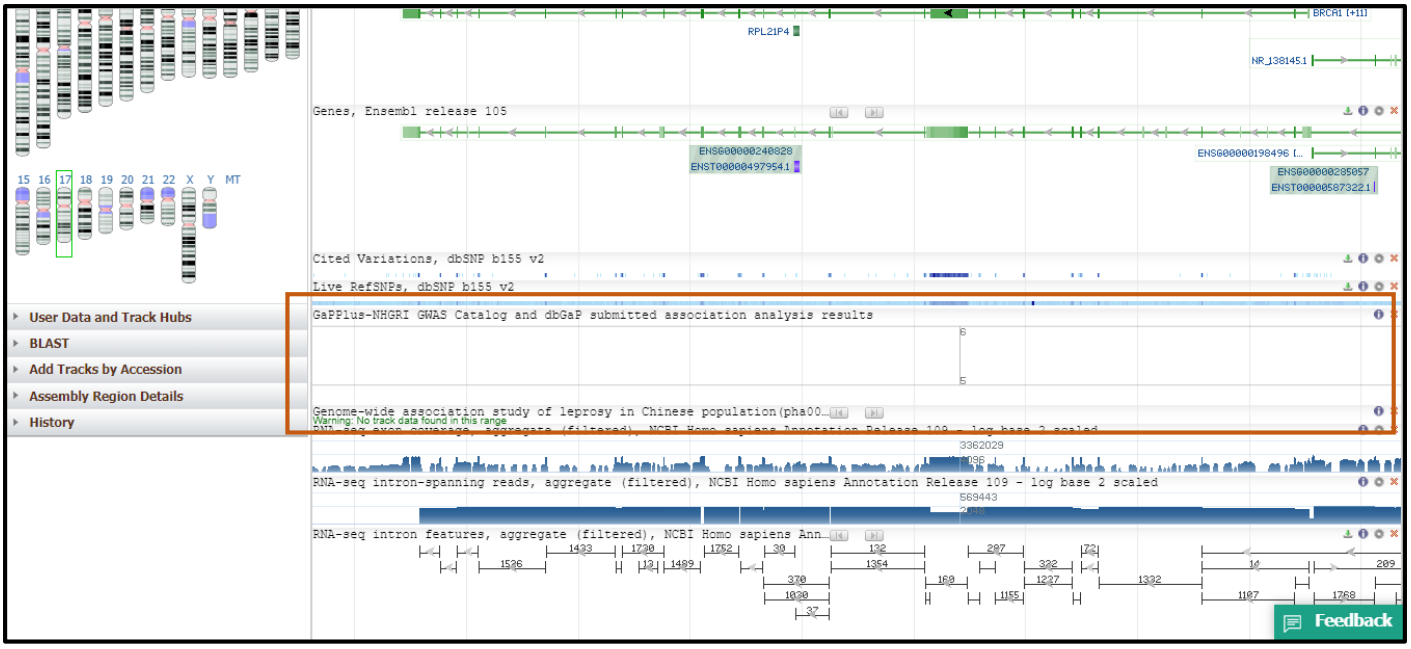


Fig5.2. Updated result after configuration

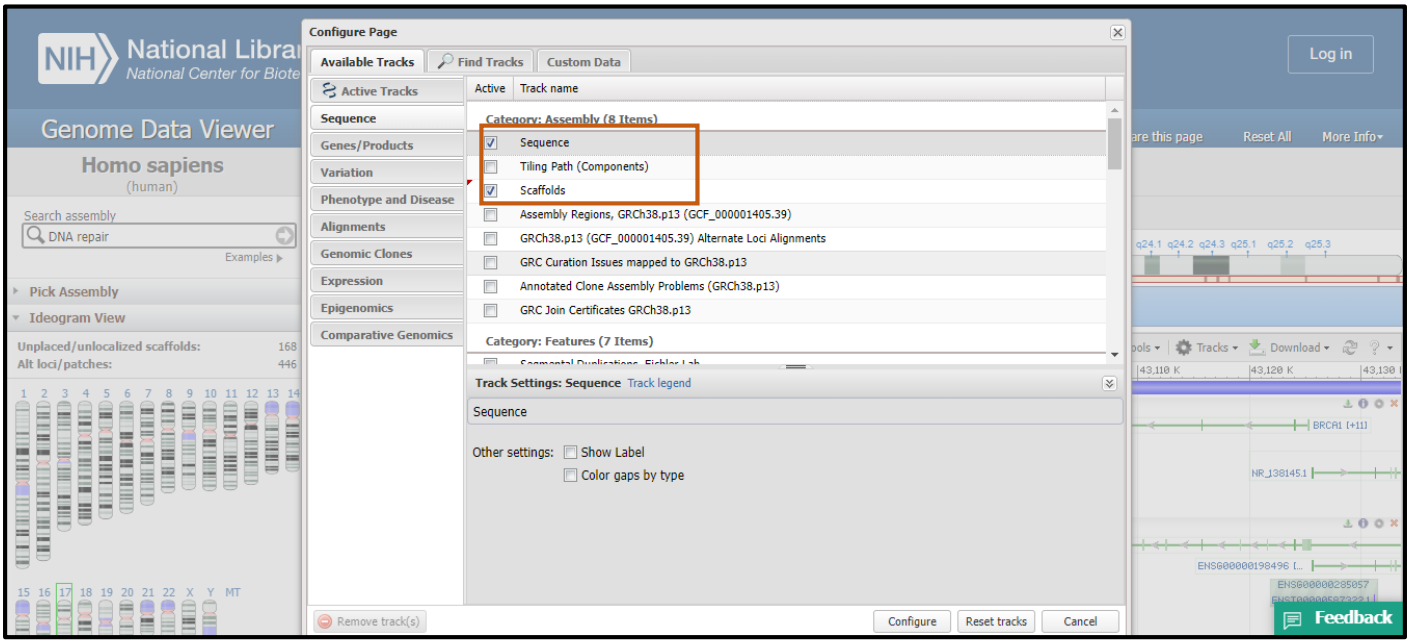


Fig5.3. Configuration page for sequence

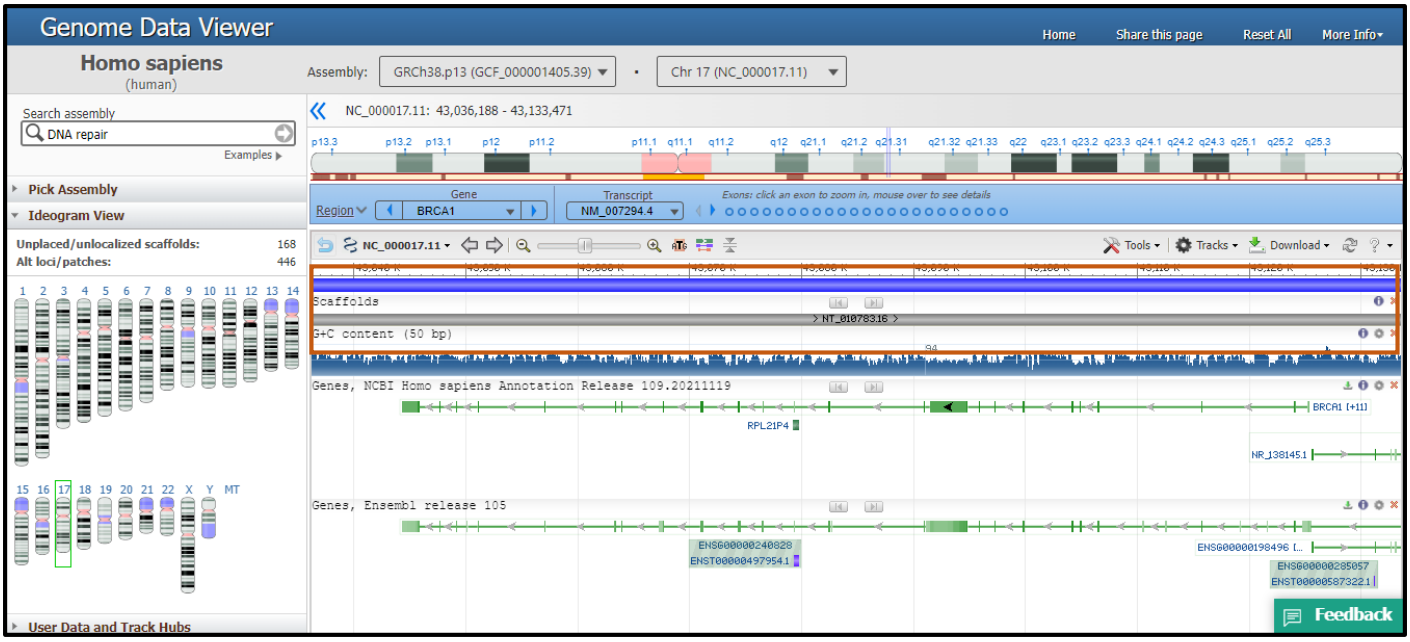


Fig5.4. Updated result after configuration

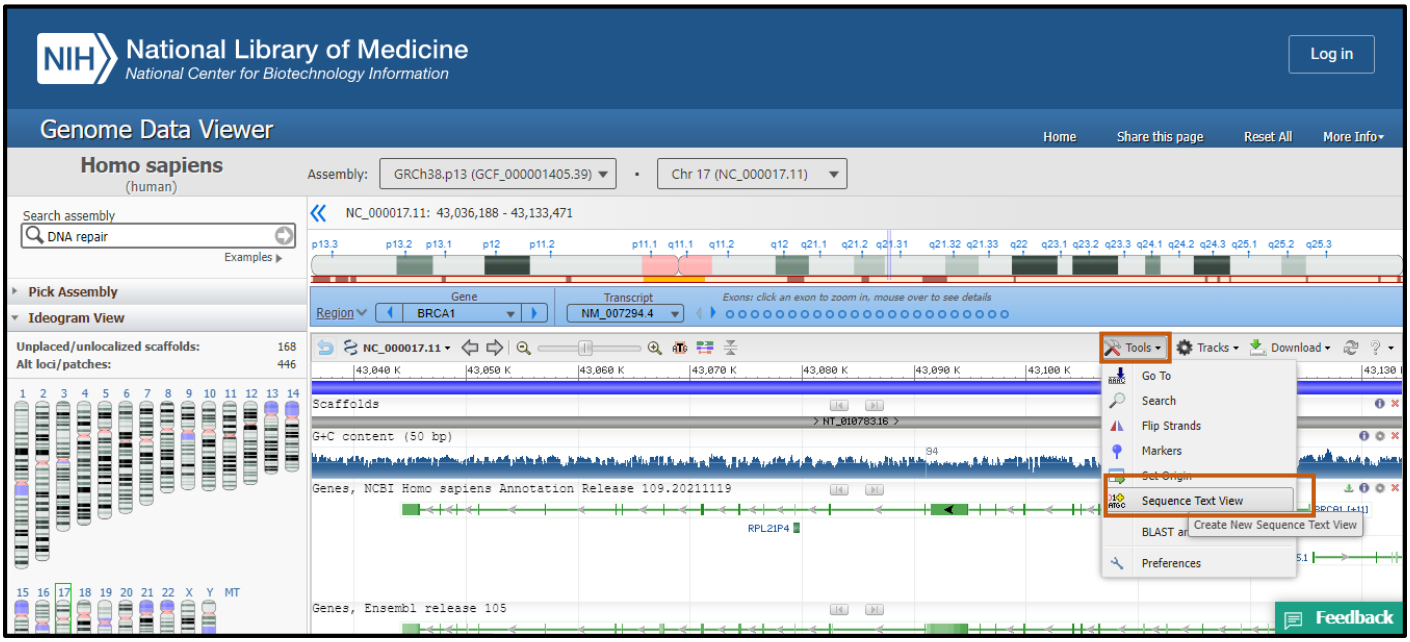


Fig6. Steps for sequence text view

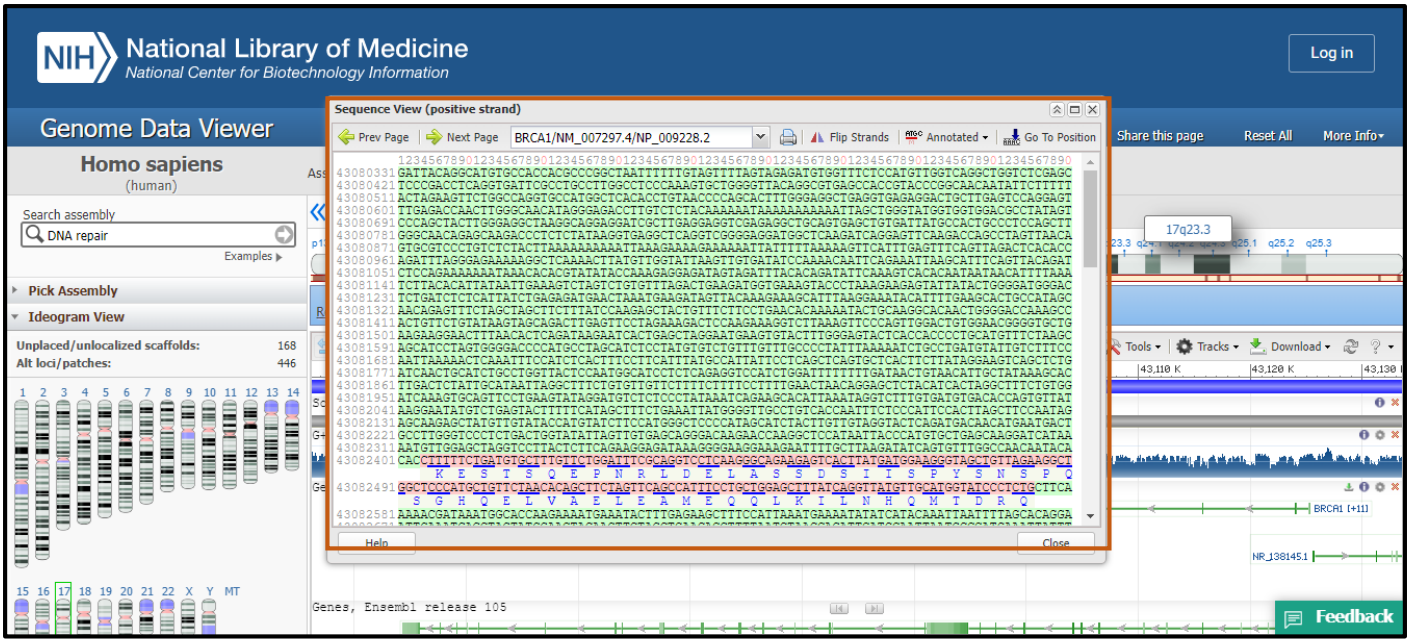


Fig6.1. Result for sequence text view

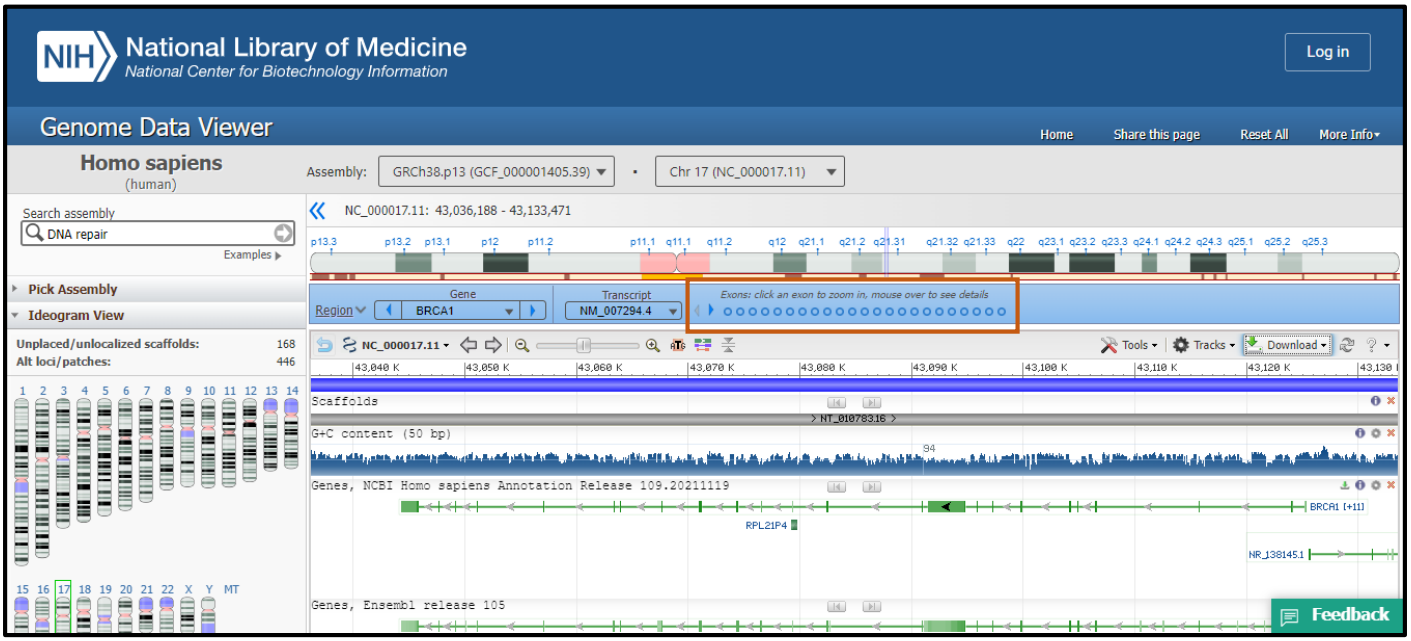


Fig7. Options to view exon information

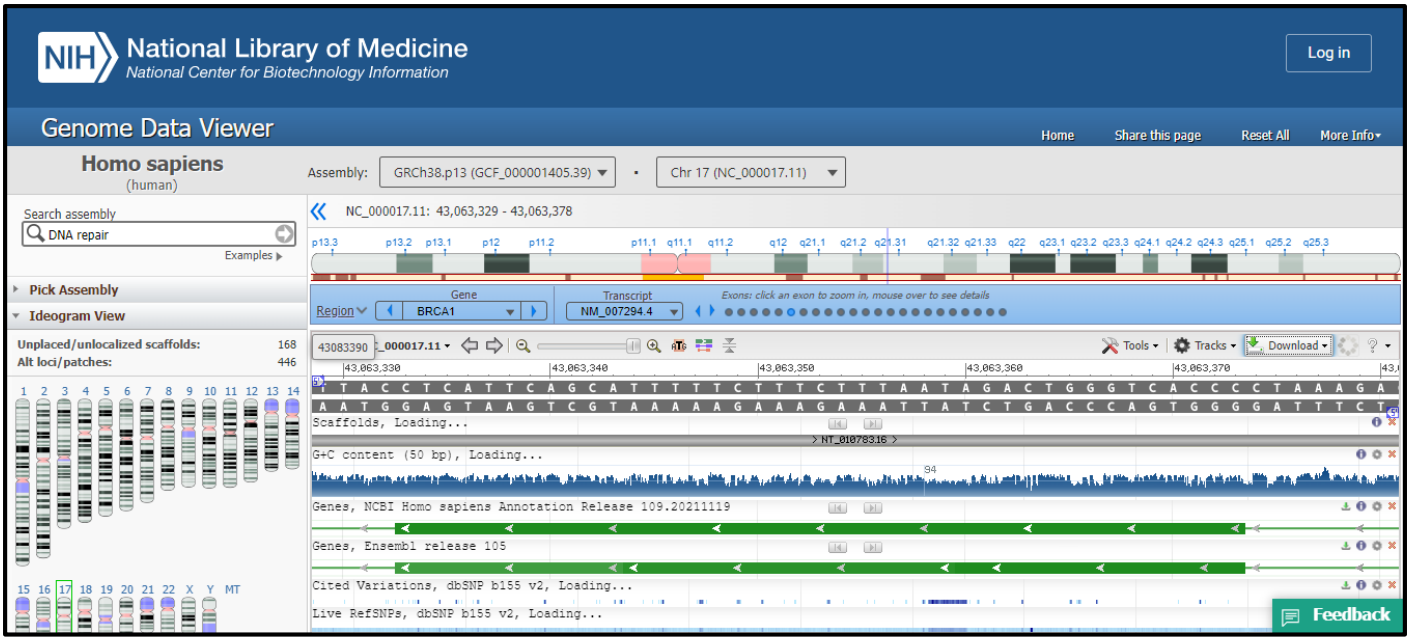


Fig7.1. Result for exon 18

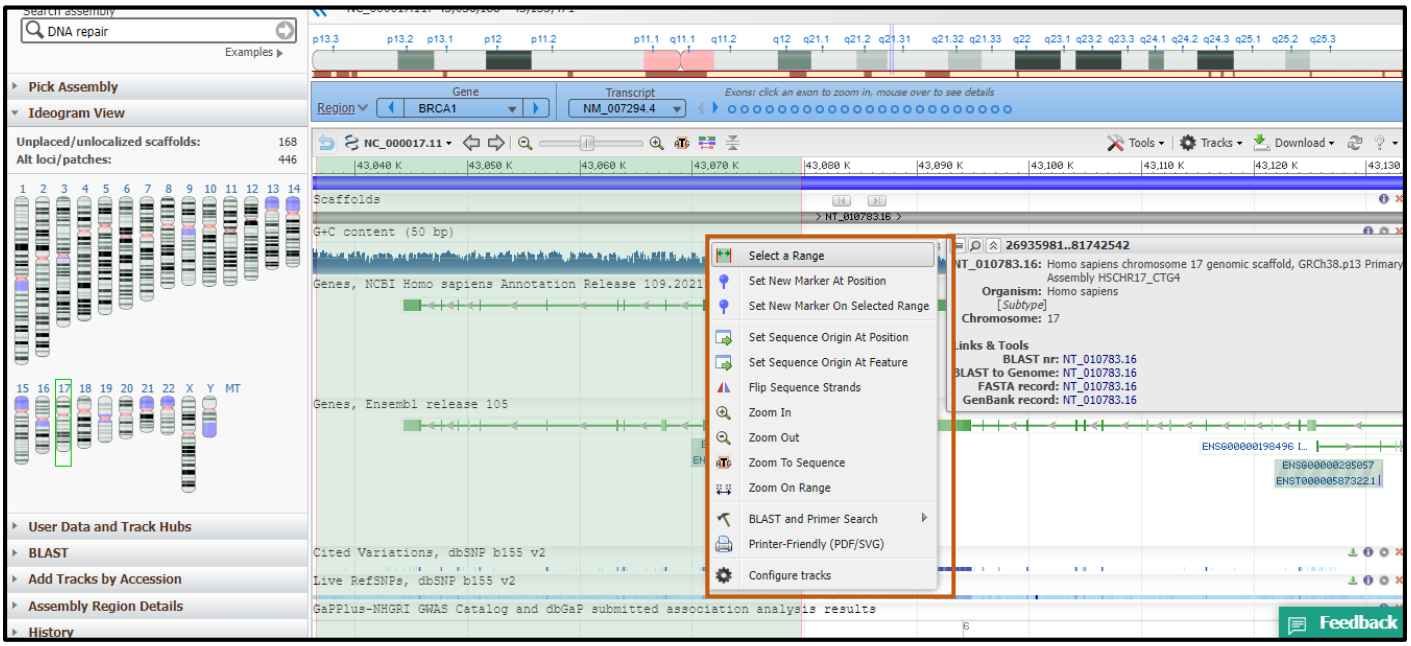


Fig8. Drag and select option for configuring tracks

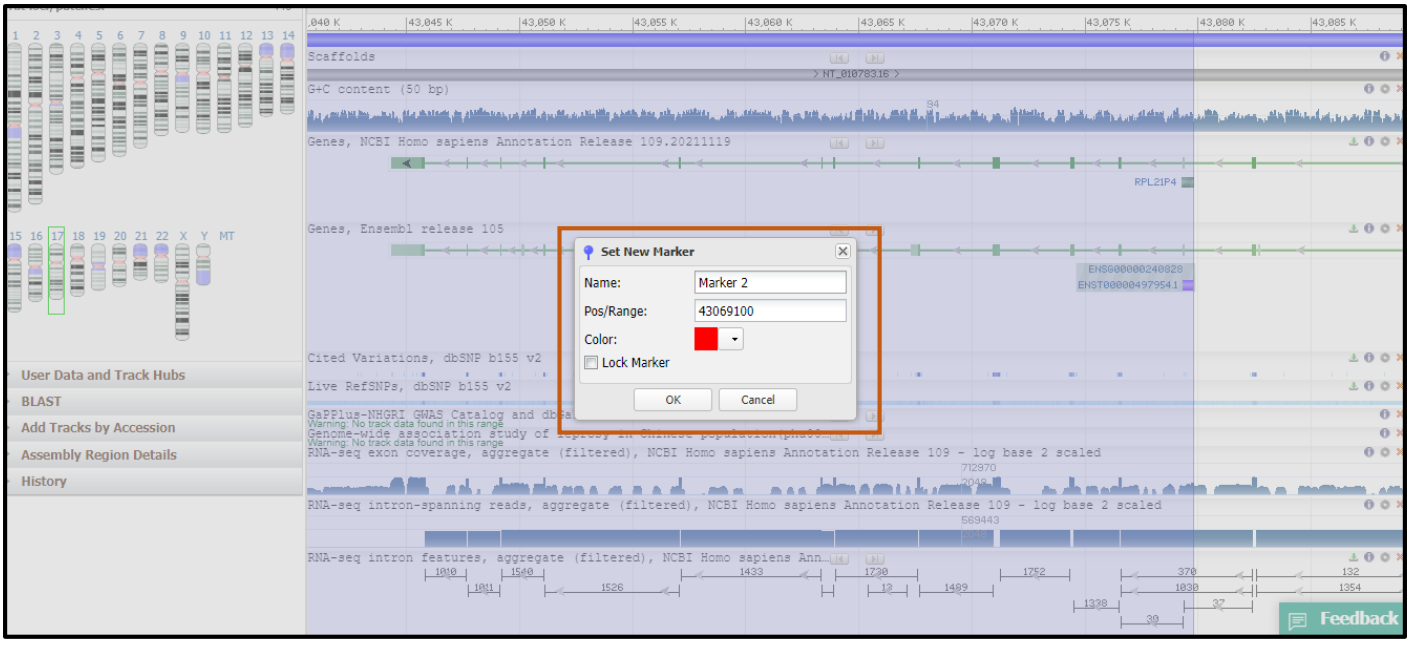


Fig8.1. Option to set new marker

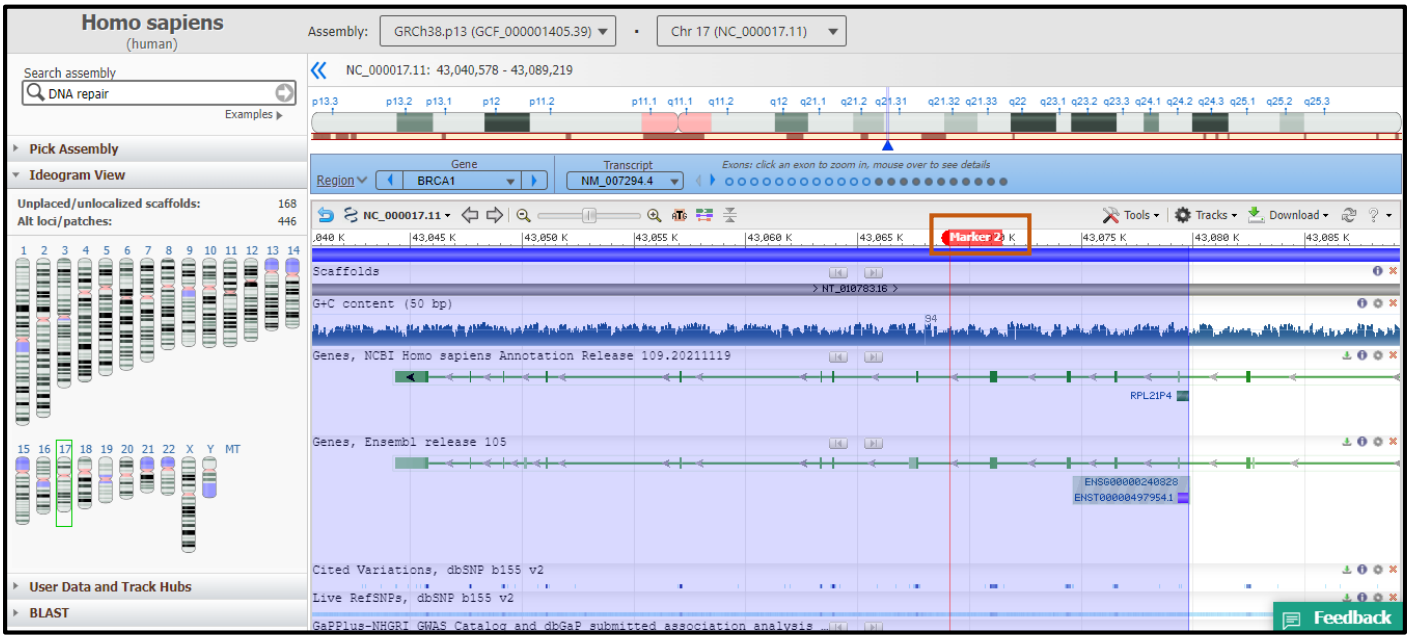


Fig8.2. Result for new marker

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. Retrieved March 28, 2022, from https://www.ncbi.nlm.nih.gov/genome/gdv/
3. *Genome Data Viewer*. (n.d.). Www.ncbi.nlm.nih.gov. Retrieved March 28, 2022, from https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39

DATE: 30/03/22

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 misasembly 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 usergenerated 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