Name:	Ms. Prarthi Hrishit Kothari
Class:	M. Sc. Bioinformatics (Part I)
Roll Number:	115
Course:	M. Sc. Bioinformatics
Department:	Department of Bioinformatics
Paper:	Mandatory Paper I
Paper Name and Code:	Advance Biology & Structural Bioinformatics (GNKPSBI1P502)
Academic Year:	2023-24

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

DEPARTMENT OF BIOINFORMATICS

CERTIFICATE

This is to certify that Ms. Prarthi Hrishit Kothari (Roll No: 115) of M. Sc. Bioinformatics (Part I) has satisfactorily completed the practical for Mandatory Paper I: Advance Biology & Structural Bioinformatics (GNKPSBI1P502) for Semester II course prescribed by the University of Mumbai during the academic year 2023-2024.

Teacher-in-Charge	Head of Department	External Examiner
(Dr. Gursimran Kaur Uppal) (Mrs. Aparna Patil Kose)	(Dr. Gursimran Kaur Uppal)	

INDEX

DATE: 03/01/2024

PRACTICAL NO.: 1(A)
ISOLATION OF PLASMID DNA

AIM:

To isolate plasmid pET21b from *E. coli* culture by alkaline lysis method.

THEORY:

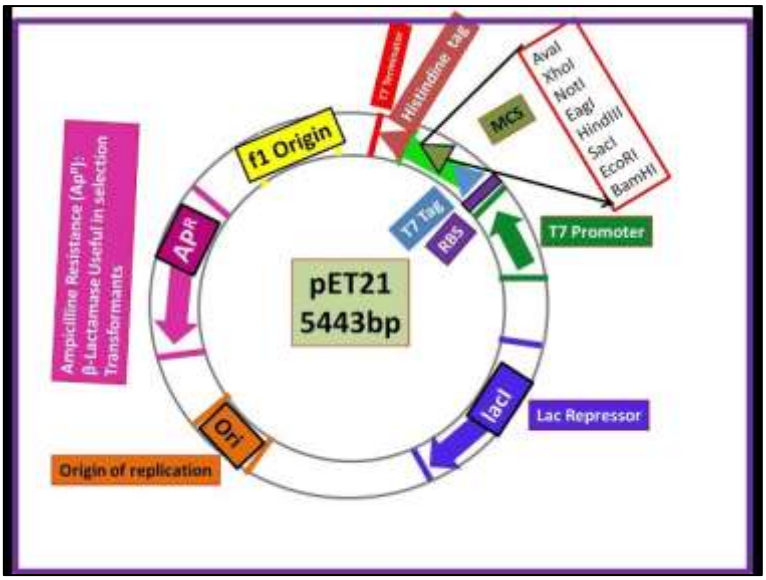


Figure 1: Plasmid DNA

A plasmid is a circular, double-stranded extrachromosomal DNA found in bacteria. It plays a crucial role in recombinant DNA experiments, enabling the cloning of genes from other organisms and the production of large quantities of their DNA. Plasmids vary in size, ranging from 1 to 1000 kilo base pairs. The term 'plasmid' was coined by American molecular biologist Joshua Lederberg. Plasmids are considered transferrable genetic elements or 'replicons,' essentially representing naked DNA.

There are two main types of plasmids:

1. **Conjugative plasmids**, which possess tra-genes (tra-transfer) and can undergo conjugation.
2. **Non-conjugative plasmids**, unable to perform conjugation.

An intermediate class of plasmid is known as mobilizable plasmid, capable of carrying only a subset of genes required for transfer.

Vectors come in two types: expressing vectors and multiplication vectors. E. coli serves as an expressing vector. The specific E. coli strain used is DH5α pET21b, where DH5α represents the E. coli strain, and pET21b is a plasmid. Successful gene uptake by E. coli requires making the bacteria competent. The DH5α culture used was 24 hours old and prepared in Luria-Bertani (LB) broth.

Plasmids, classified based on function, encompass five main types:

1. **F/Fertility Plasmid:** Primarily involved in conjugation, facilitating the transfer of genetic material between bacteria.
2. **R/Resistant Plasmid:** Contains genes that confer resistance to antibiotics. Additionally, it aids in the production of pili, hair-like structures used in bacterial adherence.
3. **Col Plasmid:** Carries genes responsible for the production of bacteriocins, toxins that inhibit the growth of similar or closely related bacterial strains.
4. **Degradative Plasmid:** Facilitates the breakdown of unusual substances, such as toluene.
5. **Virulence Plasmid:** Contributes to the pathogenicity of bacteria, enhancing their ability to cause disease.

Bacteria may harbor one or more plasmids, and these plasmids exist in various copies within each cell. Relaxed plasmids exhibit a high copy number, while stringent plasmids have a low copy number. Plasmid DNA can adopt one of five conformations, each migrating at different speeds during electrophoresis in a gel.

PRINCIPLE:

The alkaline lysis procedure is employed to lyse cells and denature DNA. This method involves raising the pH to a high level to achieve cell lysis and DNA denaturation, followed by neutralization. Plasmid DNA, being circular and supercoiled, returns to a double-stranded form when the pH is brought back to neutral. In contrast, genomic DNA, due to its larger size, is fragmented into linear pieces during this process. Various reagents in the alkaline lysis solution contribute to this procedure:

1. **Glucose:** acts to maintain osmotic pressure.
2. **Tris:** interacts with the lipopolysaccharides present on the outer membrane which helps to permeabilize and also acts as a buffering reagent.
3. **EDTA:** It chelates the divalent cations of DNases and hence keeps the DNA in the solution. It binds to Ca2+ and prevents joining of cadherins between cells, preventing clumping of cells grow in liquid suspension.
4. **NaOH:** It acts by loosening the rigid structures of a cell wall or a membrane, thereby releasing the DNA. Denatures both the chromosomal and plasmid DNA into single strands; the two strands of intact plasmid DNA remain intertwined. Na+ ions block the negative charge of DNA by forming an ionic bond with the negatively charged phosphates on the DNA, neutralizing the negative charges and allowing the DNA molecules to come together
5. **Triton X** is a detergent which destabilizes the phospholipid bilayer and also reduces the surface tension of water. It dissolves the lipid components of the cell membrane and cellular protein'
6. **Potassium Acetate:** Increases the salt concentration in the solution and brings about salting out of the DNA as it is insoluble at pH 5.3. The acetic acid solution brings the pH to neutral, and the DNA strands can renature. The large chromosomal strands cannot rehybridize perfectly, but instead become a partially-hybridized tangle. Potassium

acetate precipitates the SDS (with its lipids and proteins) from the solution. The SDS/lipid/protein precipitate traps the tangled chromosomal DNA. This creates the "white goop" that pellets after centrifugation. Only the plasmid DNA, small fragments of chromosomal DNA, and RNA remain in solution.

7. **Absolute ethanol:** Removes the water of hydration. So, the DNA cannot dissolve and sets free in the solution. An ethanol wash helps remove salts and any remaining SDS as these can interfere with a restriction digest.
8. **Isopropanol:** This alcohol rapidly precipitates nucleic acids. However, if allowed to sit for longer, proteins will also precipitate. Thus, we time this step for a quick precipitation before centrifugation.
9. **TE buffer** 10 mM Tris-Cl (desired pH) 1 mM EDTA (pH 8.0) Tris buffers the DNA solution. EDTA binds divalent cations (especially Mg+2 ions) that are a needed as cofactor for bacterial nucleases and thus limits DNA degradation.

REQUIREMENTS:

Chemicals:

1. Solution 1 (Re-suspension buffer)
2. Solution 2
3. Solution 3
4. Chloroform: isoamyl alcohol (24:1)
5. IPA – isopropyl alcohol
6. Chilled 70% alcohol
7. Absolute alcohol
8. TE buffer
9. Agarose
10. DNA ladder

Composition of the Solutions:

1. **Solution 1 (Re-Suspension Buffer):**
   a. 40% glucose
   b. 0.5 molar EDTA
   c. 1 molar tris
   d. Sterile water

2. **Solution 2 (Alkaline Lysis Buffer):**
   a. 0.4 N NaOH
   b. 10% Triton X

3. **Solution 3:**
   a. 5 M potassium acetate
   b. GAA – glacial acetic acid
   c. Water

For Making the Agarose Gel:

1. 0.6 g Agarose
2. 60 mL TAE buffer

Dyes:

1. EtBr – Ethidium Bromide
2. Bromophenol blue

Instruments, Glassware and Other:

1. Micro-pipette
2. Tips of the micropipette
3. Oven
4. Beaker
5. Weighing scale
6. Comb
7. Electrodes
8. Gel casting tray

PROCEDURE FOR EXTRACTING PLASMID DNA:

1. Take 2mL of *E. coli* culture in an eppendorf tube and centrifuge at 13500 rpm for.
2. Discard the supernatant and collect the pellet.
3. Add 100 µL of solution 1.
4. Add 200 µL of solution 2.
5. Mix it and if the string is visible on the lid, then it indicates that the plasmid isolation happened properly.
6. Add 150 µL of solution 3. Centrifuge the tube.
7. Collect the supernatant and add equal volume of chloroform: isoamyl alcohol (24:1). Centrifuge the tube.
8. Collect the upper layer and add 0.6 volume of chilled isopropyl alcohol (IPA). Centrifuge the tube.
9. Discard the supernatant. To the pellet, add 300 µL of chilled 70% ethanol. Centrifuge the tube.
10. Discard the supernatant, to the pellet add 300 µL of absolute ethanol and centrifuge the tube.
11. Discard the supernatant and dry the pellet.
12. Add 30 µL of TE buffer.

PROCEDURE FOR MAKING AGAROSE GEL AND LOADING THE SAMPLE:

1. Weigh 0.5 g of agarose on weighing scale.
2. Mix it with the 60 mL of TAE buffer.
3. Heat it until the mixture gets clear but don’t over boil it.
4. Add 5 µL of loading dye i.e. EtBr in it and mix.

5. Pour it in the gel casting tray, set the comb and allow it to solidify.
6. After solidifying, remove the comb, keep the gel in electrophoresis unit and mark the wells with bromophenol blue if needed.
7. Load DNA ladder at first well and then in the other wells load 10 µL of the extracted plasmid sample.
8. Run the gel until the dye line is approximately 80% way down the gel.
9. Under UV light observe the results.

OBSERVATIONS:

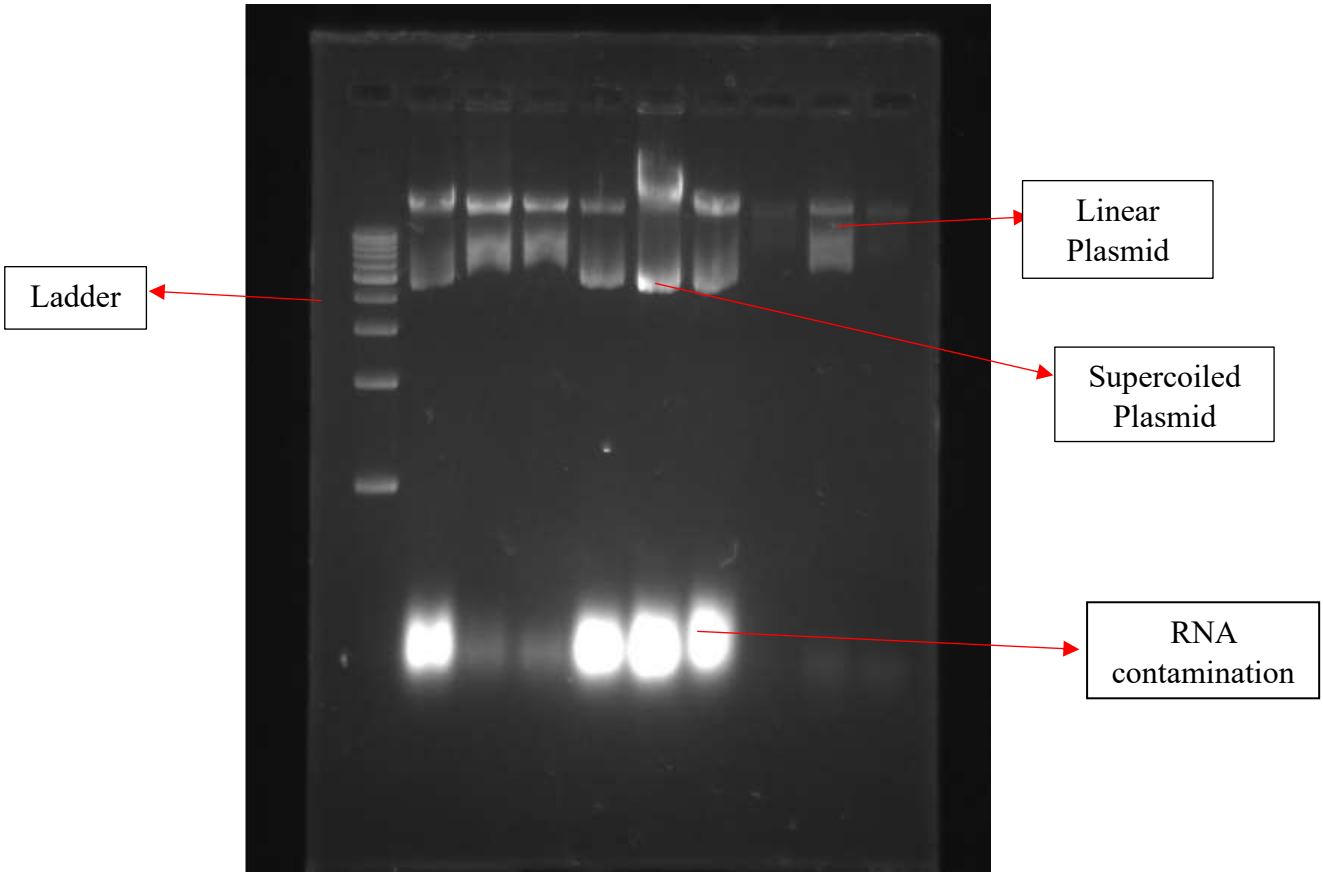


Figure 2: Gel image for Plasmid DNA extracted from *E. coli*

RESULTS:

In isolation of the plasmid DNA from *E. coli* culture by using alkaline lysis method, the plasmid sample exhibited distinct migration patterns, with linear structures appearing first, followed by partially coiled structures, and supercoiled structures at the end. This arrangement reflects the differing rates of migration: linear DNA, with its extended structure, migrated slowest, followed by partially coiled DNA (open-circular), and finally supercoiled DNA, which migrated fastest due to its compactness and extra twists in the double helix. The successful outcome underscores the effectiveness of gel separation based on both size and conformation of DNA molecules, facilitating the clear visualization of various forms of plasmid DNA.

CONCLUSION:

In conclusion, the successful isolation of plasmid DNA from an *E. coli* culture is a crucial step in molecular biology research, enabling various downstream applications such as cloning, gene expression studies, and genetic engineering. By employing established techniques such as alkaline lysis, neutralization, and precipitation, we can effectively separate plasmid DNA from chromosomal DNA and cellular debris. This purified plasmid DNA can then be further analyzed, manipulated, or utilized in experiments, contributing to advancements in fields ranging from biotechnology to medicine. The isolation process outlined not only provides a means to study plasmids in depth but also underscores their significance in genetic engineering and biotechnological applications.

DATE: 03/01/2024

PRACTICAL NO.: 1(B)
ISOLATION OF GENOMIC DNA

AIM:

To isolate genomic DNA from Bacterial cell (*Salmonella typhimurium)* sample.

THEORY:

Salmonella species are intracellular pathogens, of which certain serotypes cause illness such as salmonellosis. *Salmonella typhi* is a bacterium that causes typhoid fever, a systemic infection associated with inadequate hygiene and sanitation infrastructure in low-income settings. *Salmonella* is named after D. E. Salmon, an American bacteriologist, who first isolated the bacteria from a pig intestine in 1884.The *Salmonella* bacteria is a Gram-negative, motile, hydrogen sulfide producing, an acid-labile facultative intracellular microorganism that commonly causes gastroenteritis worldwide and causes cross-infection between humans and animals. The bacterium is transmitted through contaminated food and water, and symptoms include fever, headache, malaise, anorexia, constipation or diarrhea, and non-productive cough. Infections caused by *Salmonella typhi* are often characterized by insidious onset of sustained fever, and the clinical presentation varies, including mild and atypical infections. The QIAamp DNA mini kit is a widely used method for isolating high-quality genomic DNA from *Salmonella typhi*, as it provides a simple, efficient, and reproducible method for DNA extraction.

The isolation and purification of gDNA from cells is one of the most common procedures in molecular biology and embodies a transition from cell biology to the molecular biology; from in vivo to in vitro, as it were. DNA was first isolated as long ago as 1869 by Friedrich Miescher while he was a postdoctoral student at the University of Tubingen. Molecular biologists distinguish genomic DNA isolation from plasmid DNA isolation. In a genomic DNA isolation, the need is only to separate total DNA from RNA, protein, lipid, etc.

The CTAB extraction procedure is from Rogers and Bendich (1986). The magic bullet is supposed to be the separation of polysaccharides from nucleic acids by the use of CTAB. The technique capitalizes on the previous observations that nucleic acids can be selectively precipitated with CTAB. RNA and DNA are soluble in CTAB and NaCl.

PRINCIPLES OF CTAB METHOD (LYSIS, EXTRACTION AND PRECIPITATION):

Bacterial cells can be lysed with the ionic detergent cetyltrimethylammonium bromide (CTAB), which forms an insoluble complex with nucleic acids in a low-salt environment. Under these conditions, polysaccharides, phenolic compounds and other contaminants remain in the supernatant and can be washed away. The DNA complex is solubilized by raising the salt concentration and precipitated with ethanol or isopropanol. The principles of these three main steps, lysis of the cell membrane, extraction of the genomic DNA and its precipitation are as follows:

1. **Lysis of the cell membrane:**

As previously mentioned, the first step of the DNA extraction is the rupture of the cell and nucleus wall. For this purpose, the homogenized sample is first treated with the extraction buffer containing EDTA Tris/HCl and CTAB. All biological membranes have a common overall structure comprising lipid and protein molecules held together by non-covalent interactions.

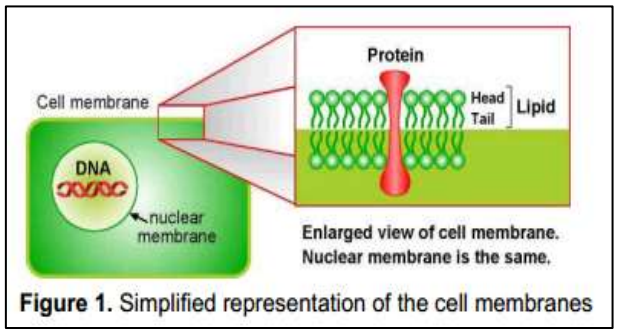


Figure 1. Simplified representation of the cell membranes

As shown in Figure 1, the lipid molecules are arranged as a continuous double layer in which the protein molecules are 'dissolved'. The lipid molecules are constituted by hydrophilic ends called heads and hydrophobic ends called tails. In the CTAB method the lysis of the membrane is accomplished by the detergent (CTAB) contained in the extraction buffer. Because of the similar composition of both the lipids and the detergent, the CTAB component of the extraction buffer has the function of capturing the lipids constituting the cell and nucleus membrane. The mechanism of solubilization of the lipids using a detergent is shown in Figure 2.

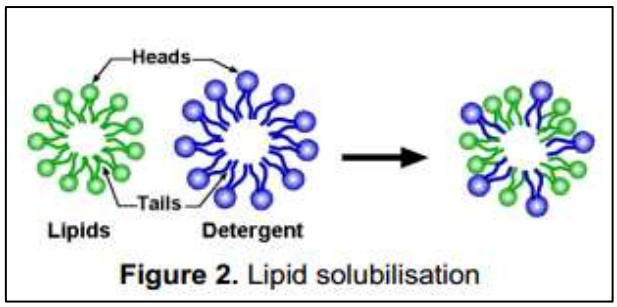


Figure 2. Lipid solubilisation

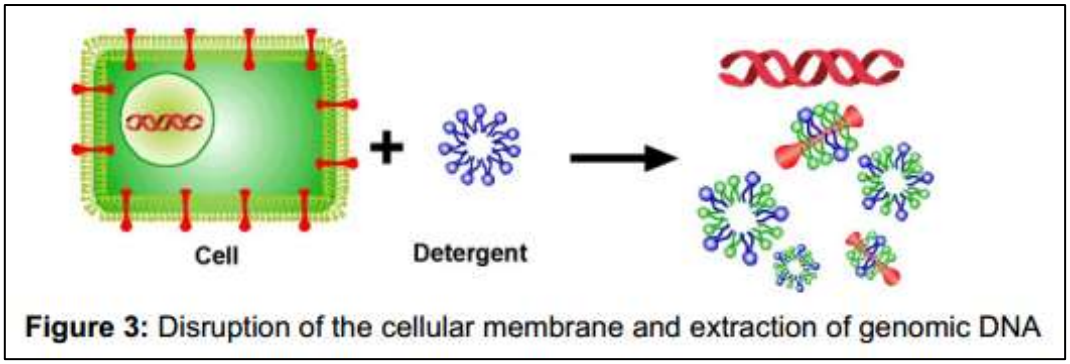


Figure 3: Disruption of the cellular membrane and extraction of genomic DNA

Figure 3 illustrates how, when the cell membrane is exposed to the CTAB extraction buffer, the detergent captures the lipids and the proteins allowing the release of the genomic DNA. In a

specific salt (NaCl) concentration, the detergent forms an insoluble complex with the nucleic acids. EDTA is a chelating component that among other metals binds magnesium. Magnesium is a cofactor for DNase. By binding Mg with EDTA, the activity of present DNase is decreased. Tris-HCl gives the solution a pH buffering capacity (a low or high pH damages DNA). It is important to notice that, since nucleic acids can easily degrade at this stage of the purification, the time between the homogenization of the sample and the addition of the CTAB buffer solution should be minimized. After the cell and the organelle membranes (such as those around the mitochondria and chloroplasts) have been broken apart, the purification of DNA is performed.

2. Extraction:

In this step, polysaccharides, phenolic compounds, proteins and other cell lysates dissolved in the aqueous solution are separated from the CTAB nucleic acid complex. The elimination of the polysaccharides as well as phenolic compounds is particularly important because of their capability to inhibit a great number of enzymatic reactions.
The Na+ ions (from NaCl) neutralize the negatively charged phosphates on DNA and facilitate DNA molecules coming together (the molecules are less hydrophilic). A high concentration of NaCl is required otherwise CTAB-nucleic acid precipitates can form junk (cell wall debris, denatured proteins, polysaccharides) complexed with CTAB and leave the bacterial DNA in solution.

The chloroform/IAA extraction (with microfuging) leaves these complexes in a whitish interface (chloroform at the bottom, DNA at the top; IAA helps stabilize the interface). If needed, the extraction with chloroform is performed two or three times in order to completely remove the impurities from the aqueous layer.
Once the nucleic acid complex has been purified, the last step of the procedure, precipitation, can be accomplished.

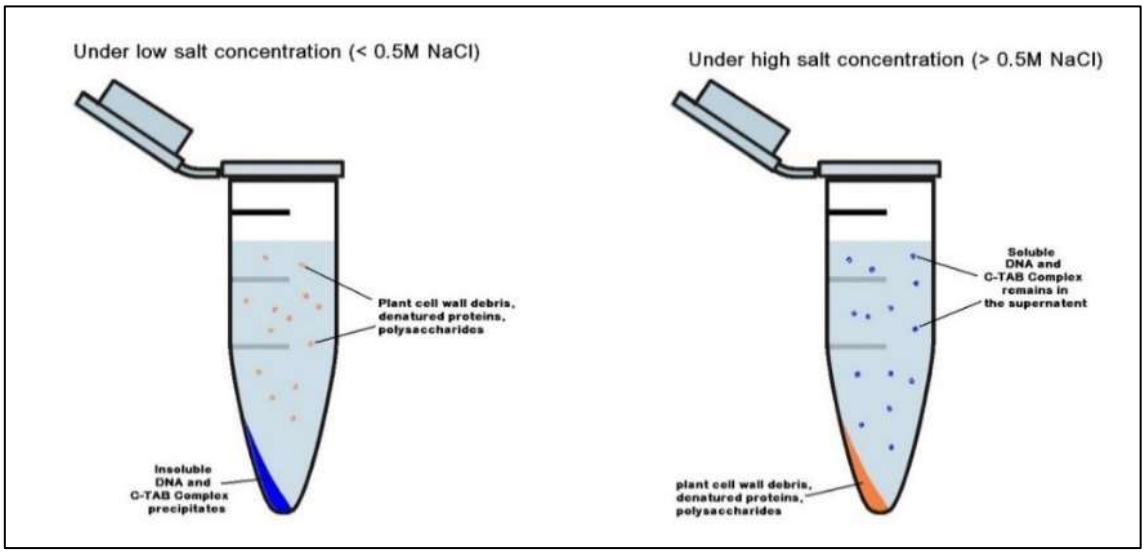


Figure 4: Effect of low and high salt concentration on gDNA isolation

3. **Precipitation:**

In this final stage, the nucleic acid is liberated from the detergent. To achieve the best recovery of nucleic acid, the organic phase may be back-extracted with an aqueous solution that is then added to the prior extract. The pellet is washed with EtOH to remove any residual salts. Under these conditions, the detergent, which is more soluble in alcohol than in water, can be washed out, while the nucleic acid precipitates. The successive treatment with ethanol allows an additional purification, or wash, of the nucleic acid from the remaining salt.

Role of each reagent:

1. **Tris:** Interacts with the lipopolysaccharides present on the outer membrane which helps to permeabilize and also acts as a buffering reagent.
2. **EDTA:** It chelates the divalent cations of DNases and hence keeps the DNA in the solution. It binds to Ca2+ and prevents joining of cadherins between cells, preventing clumping of cells grow in liquid suspension.
3. **Absolute ethanol:** Removes the water of hydration. So, the DNA cannot dissolve and sets free in the solution. An ethanol wash helps remove salts and any remaining SDS as these can interfere with a restriction digest.
4. **Chloroform:Iso-amyl alcohol (24:1):** Removes protein by precipitating it but leave DNA in aqueous solution.
5. **TE buffer:** 10 mM Tris-Cl (desired pH) 1 mM EDTA (pH 8.0) Tris buffers the DNA solution. EDTA binds divalent cations (especially Mg+2 ions) that are a needed as cofactor for bacterial nucleases and thus limits DNA degradation.

REQUIREMENTS:

Chemicals & reagents:

1. CTAB
2. β-Mercaptoethanol
3. EDTA
4. Tris-HCl
5. Absolute alcohol
6. Chloroform
7. Iso-amyl alcohol
8. TE buffer 10X assay buffer
9. Gel loading buffer (6X)
10. Sterile Distilled water
11. Agarose powder
12. 1X TAE buffer
13. DNA ladder 1kb
14. Ethidium Bromide Stock (1mg/ml)

Instruments:

1. Centrifuge
2. Electrophoresis unit
3. UV-trans illuminator
4. Microwave
5. Weighing balance
6. Power pack
7. Heating block/water bath at 37°C

Miscellaneous:

1. 2 mL Eppendorf
2. Crushed ice
3. Eppendorf rack
4. Micropipette and Sterile microtips
5. Gloves

PREPARATION:

1. **CTAB buffer composition:**

Sr. No.	Components	Volumes
1	2% CTAB (hexadecyltrimethylammonium bromide)	20 ml from 10% CTAB buffer
2	100mM Tris-HCl [pH 8.0]	10 ml from 1M Tris-Cl (pH 8.0)
3	20mM EDTA [pH 8.0]	4 ml from 0.5M EDTA (pH 8.0)
4	1.7M NaCl	34 ml from 5M NaCl
5	0.3% β-mercaptoethanol [add just before use]	300 µl from liquid β-ME
	Total	100ml

10% CTAB:

a. 10g of CTAB
b. Add 75ml of D/W.
c. Heat the solution at 60°C for 15 mins
d. Stir till CTAB dissolves completely.
e. Bring total volume to 100ml with D/W.

TE Buffer:

a. 1ml 1 M Tris HCl pH 8.0
b. 0.2ml 0.5 M EDTA pH 8.0
c. Bring total volume to 100ml with D/W.

1 M Tris HCl pH 8.0:

a. 121.1g Tris
b. Dissolve in about 700 ml of D/W.
c. Bring pH down to 8.0 by adding 10N HCl (about 50 ml).
d. Bring total volume to 1 L with D/W.

0.5 M EDTA pH 8.0:

a. 186.12g EDTA
b. Add about 700ml D/W.
c. Adjust the pH to 8.0 by 10N NaOH
d. EDTA won't dissolve until the pH is near 8.0
e. Bring total volume to 1L with D/W.

5M NaCl:

a. 292.2g of NaCl
b. 700ml D/W
c. Dissolve and bring to 1L.

PROCEDURE:

1. Weigh 400mg of plant material, make fine pieces.
2. Take weighed material in mortal and pestle and add 2ml of CTAB buffer. Grind the material till a slurry forms.
3. Transfer the mixture in 2ml eppendorf tube.
4. Keep the eppendorf at 60°C for 1 hour.
5. Centrifuge the tube at 10,000rpm for 10 minutes.
6. Transfer the supernatant into two tubes (1ml in each tube).
7. Add 1 ml of Chloroform: Iso-amyl alcohol (24:1) and mix slowly.
8. Centrifuge at 10000 rpm for 10 mins.
9. Collect upper aqueous layer into new 2 ml eppendorf tube without disturbing the bottom layer.
10. Add 2 volumes chilled 100% Ethanol and mix slowly.
11. Centrifuge at 10000 rpm for 10 minutes. Decant the supernatant without disturbing the pellet.
12. Add 300µl of 70% Ethanol.
13. Centrifuge for 5mins at 10000rpm. Decant the supernatant without disturbing the pellet.
14. Add 300µl of 100% Ethanol.
15. Centrifuge for 5mins at 10000rpm
16. Discard supernatant and dry the pellet at 37°C
17. Reconstitute with 30µl of TE buffer.
18. Load 10µl of DNA sample on 0.8% Agarose gel with TAE buffer.

OBSERVATIONS:

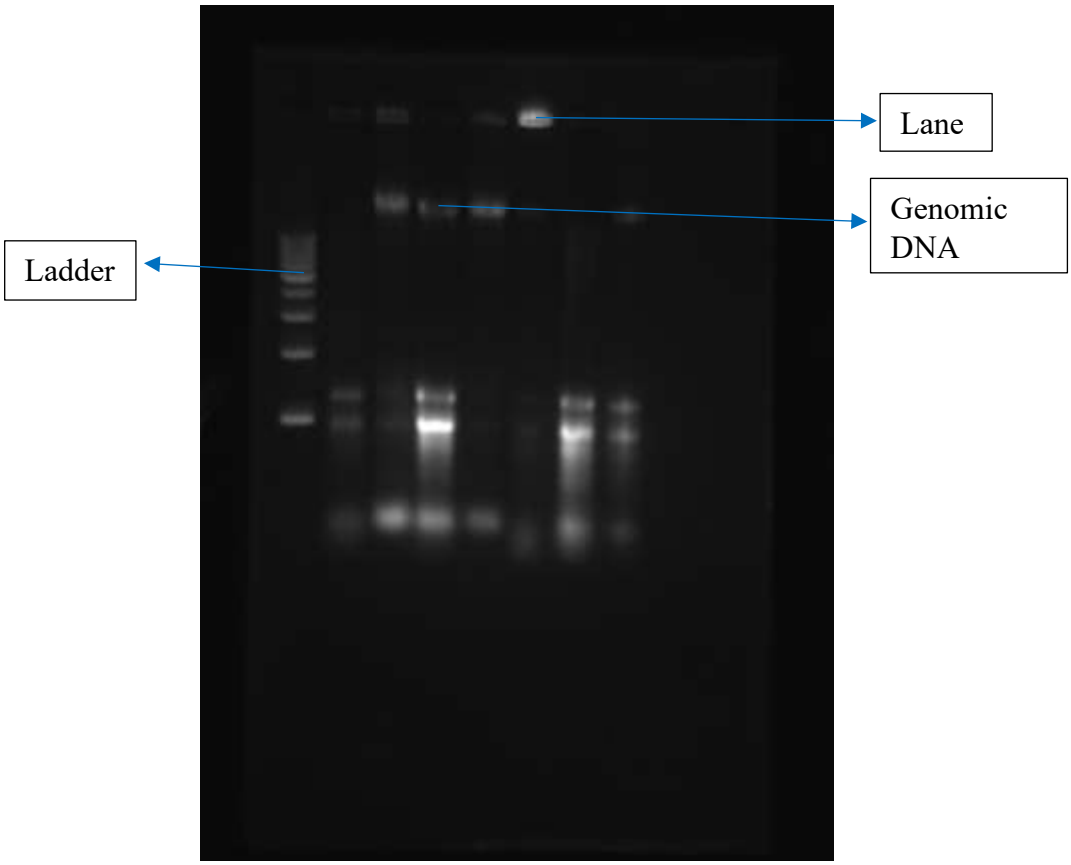


Figure 5: Gel image for Genomic DNA Isolation

RESULTS:

The presence of a DNA ladder in the first lane serves as a reference for determining the molecular weight of DNA fragments. The DNA ladder consists of DNA fragments of known sizes, which migrate through the gel at specific rates based on their size. The DNA samples in the other lanes, including the mixture of proteins, are observed under UV light after staining with ethidium bromide. The DNA fragments appear as distinct bands, while the proteins appear as darker bands due to their inherent absorption of UV light. The DNA bands are observed at positions corresponding to their sizes relative to the DNA ladder, while the protein bands are observed at different positions due to their distinct migration patterns.

CONCLUSIONS:

In conclusion, the CTAB method stands as a robust approach for isolating genomic DNA from *Salmonella typhimurium*, offering several advantages such as simplicity, cost-effectiveness, and scalability. By utilizing the CTAB method, researchers can efficiently lyse bacterial cells, precipitate proteins, and purify genomic DNA, yielding high-quality DNA samples suitable for a wide range of downstream applications in molecular biology, including PCR, sequencing, and genetic analysis. Moreover, the CTAB method's compatibility with various sample types and its capacity to produce DNA free from contaminants make it particularly valuable for research, diagnostic, and epidemiological studies focused on *Salmonella typhimurium*.

Embracing the CTAB method not only streamlines laboratory workflows but also enhances our understanding of the genetic characteristics, virulence factors, and antibiotic resistance profiles of this pathogenic bacterium, ultimately contributing to advancements in infectious disease management and public health initiatives.

DATE: 05/01/2024

PRACTICAL NO.: 2
RESTRICTION ENZYME DIGESTION

AIM:

To perform restriction enzyme digestion of the isolated plasmid.

THEORY:

A restriction enzyme, also known as a restriction endonuclease, is an enzyme that cleaves DNA at particular recognized nucleotide sequences called restriction sites, regardless of the fact that it is single-stranded or double-stranded. A restriction enzyme cuts the DNA twice: once through each strand of the DNA double helix, or sugar-phosphate backbone. These enzymes, which are present in bacteria and archaea, operate as a barrier that prevents the invasion of viruses.

A process termed as restriction occurs when restriction enzymes within bacterial hosts break up foreign DNA selectively; methylase, a modifying enzyme, methylates host DNA to shield it from the action of restriction enzymes. These two procedures come together to generate the restriction-modification system. This self-defense is made possible by a particular DNA methyltransferase enzyme, which transfers methyl groups to adenine or cytosine residues to form N6-methyladenine or 5-methylcytosine, thereby methylating the corresponding DNA sequence for the corresponding restriction enzyme.

Restrictions endonucleases frequently identify recognition sequences, which are primarily palindromes consisting of the same forward (5' to 3' on the top strand) and backward (5' to 3' on the bottom strand) sequences.

PRINCIPLE:

The DNA from bacteriophage is the most widely used substrate for screening restriction enzymes. A DNA is isolated as a linear molecule from the *E. coli* bacteriophage. It contains 48,502 base pairs and has 5 recognition sites for *Eco*RI, 7 for *Hind* III and 29 for *Bg*II.

*Eco*RI is a restriction enzyme, which is isolated from *E. coli* RX13 bacterial cells. *Eco*RI act as a dimer and recognizes six base pairs palindromic sequences (nucleotide pair sequences which are same when read forward or backward from a central axis of symmetry). It acts on phosphodiester bonds between two nucleotide and cuts at 5 recognition sites. This staggered cleavage of double stranded DNA results is sticky/cohesive ends, which are identical, complementary, single stranded projections and can base pair with each other.

In the experiment, DNA is subjected to digestion by *Eco*RI enzyme. The recognition sequences for *Eco*RI in DNA are 'GAATTC'. EcoRI enzyme protein binds itself to the DNA molecule and then recognizes specific restriction on the chromosome. *Eco*RI cleaves dsDNA strand between 'G' and 'A' nucleotides by hydrolyzing 2 phosphodiester bonds (1 per strand) within defined nucleotide sequences, and forms fragments with 5 terminal phosphate and 3 terminal hydroxyl residues. This results in living 5' overhangs of ‘TTAA’ on the complementary strand.

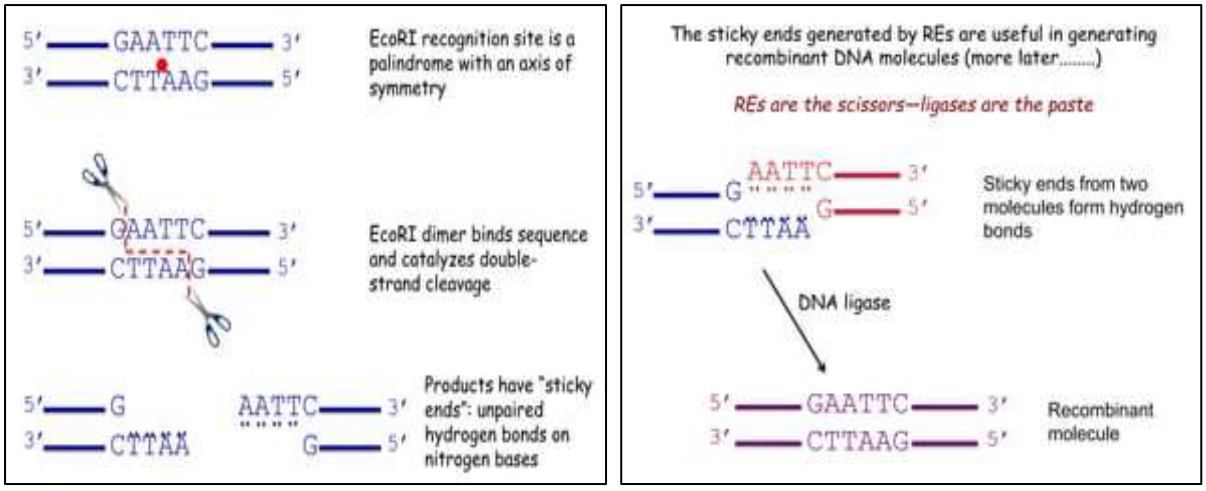


Figure 1: Restriction enzyme digestion

Restriction Enzyme	Strain of Origin	Recognition Site
*Eco*RI	*E,coli*	G A A T T C
Hind III	*H. influenza*	A A G C T T
*Bam*HI	*B. amyloliquefaciens*	G G A T C C

Based on their composition, the type of enzyme cofactor they require, the target sequence, and the location of their DNA cleavage site in relation to the target sequence, restriction endonucleases are divided into three broad classes (Types I, II, and III).

TYPE I ENZYMES	TYPE II ENZYMES	TYPE III ENZYMES
They cleave at sites remote from the recognition site.	They cleave within or at short specific distances from recognition site.	They recognize two separate non-palindromic sequences that are inversely oriented and cut DNA about 20-30 base pairs after the recognition site.
They require both ATP and S-adenosyl-Lmethionine to function.	They mostly require magnesium to function.	These enzymes contain more than one subunit and require S-Adenosyl methionine and ATP cofactors for their roles in DNA methylation and restriction, respectively.
Multifunctional protein with both restriction and methylase activities	Single function (restriction) enzymes independent of methylase.	

Factors Affecting Restriction Enzyme Activity:

1. **Temperature:** The majority of digestions occur at 37°C. There are, however, certain outliers. For example, digestion using Sma I occurs at a lower temperature (~25°C), however digestion using Taq I occurs at a higher temperature (65°C).
2. **Buffer Systems:** The most often utilized buffering agent in temperature-dependent incubation mixes is tris-HCl. The optimum pH range of 7.0 to 8.0 is where the majority of restriction enzymes are active.
3. **Ionic Conditions:** Mg^{2+} is an absolute requirement for all restriction endonucleases, but the requirement of other ions (Na+/K+) varies with different enzymes.
4. **Methylation of DNA:** Methylation of specific adenine or cytosine residues within the recognition sequence of the restriction enzyme affects the digestion of DNA.

Star Activity:

It is a variation in the specificity of DNA cleavage mediated by restriction enzymes that may occur under certain unusual circumstances that deviate from the enzyme's ideal state. The cleavage at non-specific locations is caused by this modification.

REQUIREMENTS:

Chemicals and Reagents:

1. Sterile distilled water
2. 10X Assay buffer
3. DNA sample
4. Restriction Enzyme 1
5. Restriction Enzyme 2

Instruments:

1. Dry bath
2. Micropipettes
3. Micro tips
4. PCR Vials

Miscellaneous:

1. Ice or frosty box
2. PCR vial racks
3. Vortex

PROCEDURE:

1. Arrange all the reagents on the ice box.
2. Arrange the PCR vials in the rack and label them accordingly.
3. Add all the reagents into the PCR vial according to the reaction mixture provided.

Reagents	Volume(µL)
Sterile distilled water	2
10X assay buffer	1
DNA sample	5
Restriction enzyme (*Bam*H1)	1
Restriction enzyme (*Pst*1)	1
Total volume	10

4. Mix the contents by tapping or gentle vortexing and incubate the tubes at 37°C for 30 mins in a heat block.
5. Load the sample on 1% agarose gel containing 0.1 µg/ mL of ethidium bromide.

OBSERVATIONS:

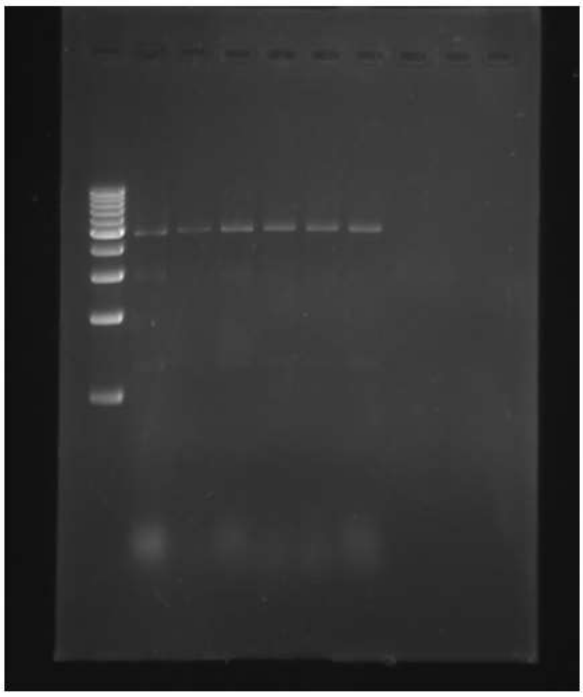


Figure 2: Evaluation of the Digestion by restriction enzyme was done and was evaluated on 1% Agarose gel and stained with 0.1µg/mL of Ethidium Bromide.

RESULTS:

Restriction Enzyme Digestion & Gel Electrophoresis of DNA demonstrates how DNA can be specifically cut into fragments by restriction enzymes and then can be separated by fragment size on an agarose gel. *Bam*HI is a type II restriction enzyme derived from *Bacillus amyloliquefaciens*. Like all Type II restriction endonucleases, it is a dimer and the recognition site is palindromic and 6 bases in length. It recognizes the DNA sequence of G'GATCC and

leaves an overhang of GATC which is compatible with many other enzymes. *Pst*I cleaves DNA at the recognition sequence 5′-CTGCA/G-3′ generating fragments with 3′-cohesive terminals. This cleavage yields sticky ends 4 base pairs long. *Pst*I is catalytically active as a dimer. The one fragment was observed ~between 4000 and 5000 and the other ~between 1000 and 1500.

CONCLUSION:

A unique class of enzymes known as restriction enzymes has the ability to fragment DNA at particular regions, or restriction sites. Bacteria use this defense mechanism to shield themselves from genetic or viral coding. Recombinant plasmid constructions can be quickly and cheaply restrictedly digested to yield indirect sequence information. The presence or absence of an insert, the insert's orientation, the size of the plasmid, and certain site-specific sequencing information can all be examined simultaneously for several plasmid constructs.

DATE: 06/01/2024

PRACTICAL NO.: 3
DNA LIGATION

AIM:

To perform ligation reaction using T4 DNA ligase.

THEORY:

Genetic engineering is a modern biotechnology technique which is used to modify the genetic makeup of an organism by adding new traits in to it and there by produce new variety of organisms. The desired trait or gene is cut from the donor gene and pasted to the carrying vector. Thus, formed recombinant DNA is transferred into the host organism which has to be modified genetically. The process of genetic engineering is made effective by the action of two enzymes namely, Restriction enzymes and ligases. Restriction enzymes, called molecular scissors are precisely cut the desired DNA and its carrying vector, at specific sites, whereas the ligase enzyme, called Molecular glue are pasted these DNA fragments into the carrying vector. The process of ligase enzyme for joining the two ends of DNA strands is called ligation. In this process, there occurs a synthesis of phosphodiester bond between the 3'hydroxyl of one nucleotide and 5'phosphate of another. In this situation we need to discuss about DNA and its Structure.

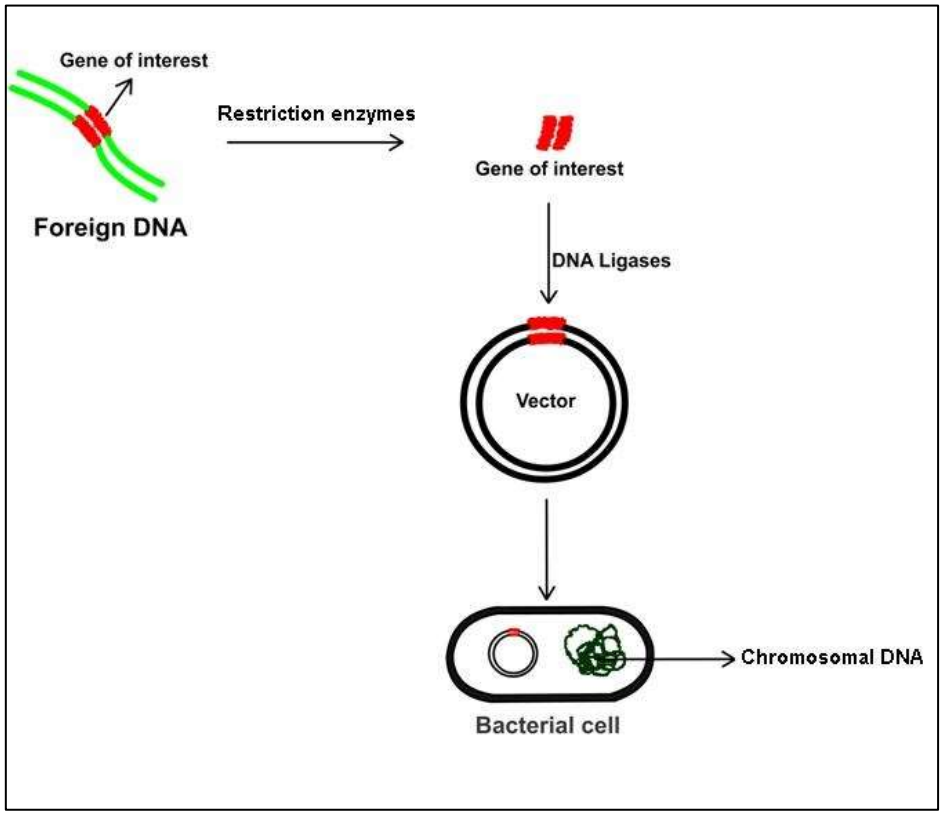


Figure 1: Vector cloning

Structure of DNA:

DNA is made up of nucleotides which are composed of a nitrogenous base, sugar (ribose) and phosphate group. The structure thus formed is called as Double helix. The sugar and the phosphate group form the backbone of this helix. The sugar is attached to one of the four bases namely adenine, thymine, guanine and cytosine. The sugar is a pentose sugar called as the 2'deoxyribose. This sugar is connected to the carbon atoms of the adjacent sugars by phosphate groups that make the phosphodiester bonds between the third carbon atoms of one sugar to the fifth carbon atom of the adjacent sugar.

Ligating an insert DNA into a plasmid requires complementary ends between the DNA and the plasmid vector. Sticky ends are produced by cutting the DNA in a staggered manner within the recognition site and there by produce short Single stranded DNA. These ends have identical nucleotide sequence and are sticky because they can bind to complementary tails of other DNA fragments cut by the same restriction enzyme. Both are useful in molecular genetics for making recombinant DNA and proteins. Blunt ends are generated by cutting both DNA strands in the middle of the recognition sequence. DNA ligase helps to join together the complementary ends of insert DNA and plasmid DNA.

Different parameters affect ligations such as the ratio of insert to vector, the quality and type of the DNA ends, the ligation temperature and the DNA concentration. Each of these factors is necessary for a successful ligation. The basic purpose in molecular cloning is the insertion of DNA fragment of interest (a segment of DNA) into a DNA molecule (called a vector) that has the capacity to replicate independently with in a host cell. The result is a recombinant molecule composed of the DNA insert joined to vector DNA sequences. Construction of these recombinant DNA molecules is dependent on the ability to covalently seal single stranded nicks in DNA. This process is accomplished both in vivo and in vitro by the enzyme DNA ligase.

The DNA fragments used to produce recombinant molecules are commonly generated by digestion with restriction endonucleases. Most of these enzymes cut their recognition sequences at staggered sites, giving overhanging or cohesive single-stranded tails. These are associate with each other by complementary base pairing. These paired complementary ends can be established permanently by DNA ligase treatment. Thus, two different fragments of DNA prepared by digestion with the same restriction endonuclease further joined to produce a recombinant DNA molecule.

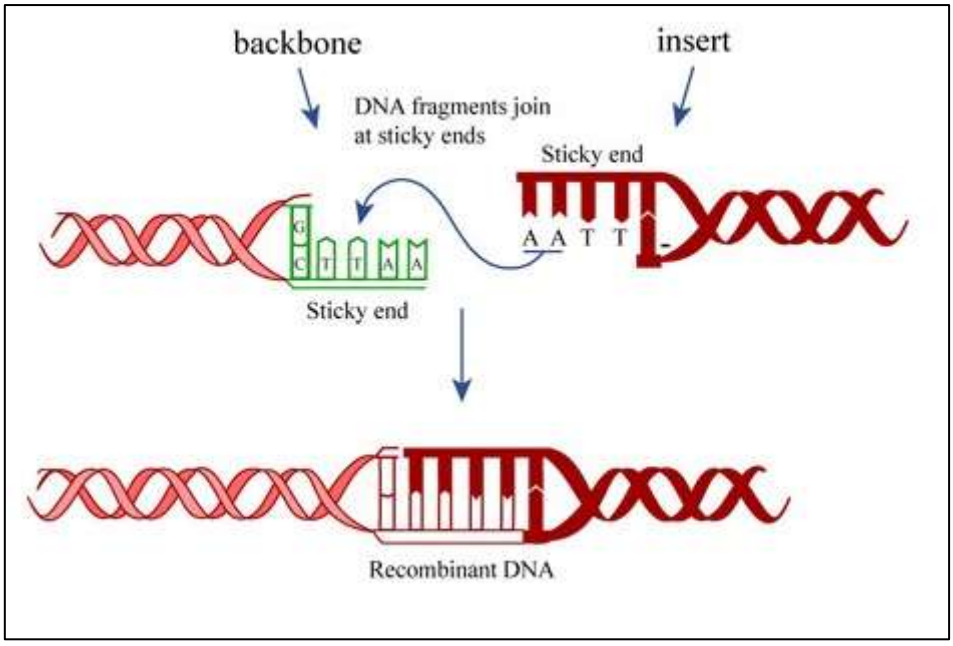


Figure 2: Recombinant DNA

PRINCIPLE:

DNA ligation is the act of joining together DNA strands with covalent bonds with the aim of making new viable DNA or plasmids. There are currently three methods for joining DNA fragments in vitro. The first of these is DNA ligase that covalently joins the annealed cohesive ends produced by certain restriction enzymes. The second depends upon the ability of DNA ligase from phage T4-infected *E. coli* to catalyze the formation of phosphodiester bonds between sticky or blunt-ended fragments. The third utilizes the enzyme terminal deoxynucleotidyl transferase to synthesize homopolymeric 3′ single-stranded tails at the ends of fragments. The most commonly used is the T4 DNA ligase method.

E. coli and phage T4 encode an enzyme, DNA ligase, which seals single-stranded nicks between adjacent nucleotides in a duplex DNA chain. Although the reactions catalyzed by the enzymes of *E. coli* and T4-infected *E. coli* are very similar, they differ in their cofactor requirements. The T4 enzyme requires ATP, while the *E. coli* enzyme requires NAD+. In each case the cofactor is split and forms an enzyme–AMP complex. The complex binds to the nick, which must expose a 5′ phosphate and 3′ OH group, and makes a covalent bond in the phosphodiester chain.

DNA fragments with either sticky ends or blunt ends can be inserted into vector DNA with the aid of DNA ligases. During normal DNA replication, DNA ligase catalyzes the end-to-end joining (ligation) of short fragments of DNA, called Okazaki fragments. For purposes of DNA cloning, purified DNA ligase is given to covalently join the ends of a restriction fragment and vector DNA that have complementary ends. The vector DNA and restriction fragment are covalently ligated together through the 3’ → 5’ phosphodiester bonds of DNA. When termini created by a restriction endonuclease that creates cohesive ends associate, the nicks in the joints have few base pairs apart in opposite strands. DNA ligase can then repair these nicks to form an intact duplex.

T4 DNA Ligase:

Bacteriophage T4 DNA ligase is a single polypeptide with a M.W. of 68,000 Dalton requiring ATP as energy source. The maximal activity pH range is 7.5-8.0. The enzyme exhibits 40% of its activity at pH 6.9 and 65% at pH 8.3. The presence of Mg++ ion is required and the optimal concentration is 10mM.

T4 DNA ligase has the unique ability to join sticky and blunt ended fragments. Cohesive end ligation is carried out at 12°C to 16°C to maintain a good balance between annealing of ends and activity of the enzyme. If reaction is set at higher temperatures annealing of the ends become difficult, while lower temperatures diminish the ligase activity. All T4 DNA ligase is inactivated by heating at 65°C for 10 minutes. Beside of these ligating complementary sticky ends, T4 ligase can ligate any two blunt DNA ends. Lack of cohesive termini makes blunt end ligation more complex and significantly slower. Since annealing of ends is not a factor, the reaction is done at 24˚C. However, 10 - 100 times more enzyme is required to achieve similar ligation efficiency as that of cohesive end ligation. The enzyme has involved in the catalysis of the joining of RNA to either an RNA or DNA strand in a duplex molecule but this will not involve in the joining of single stranded nucleic acids.

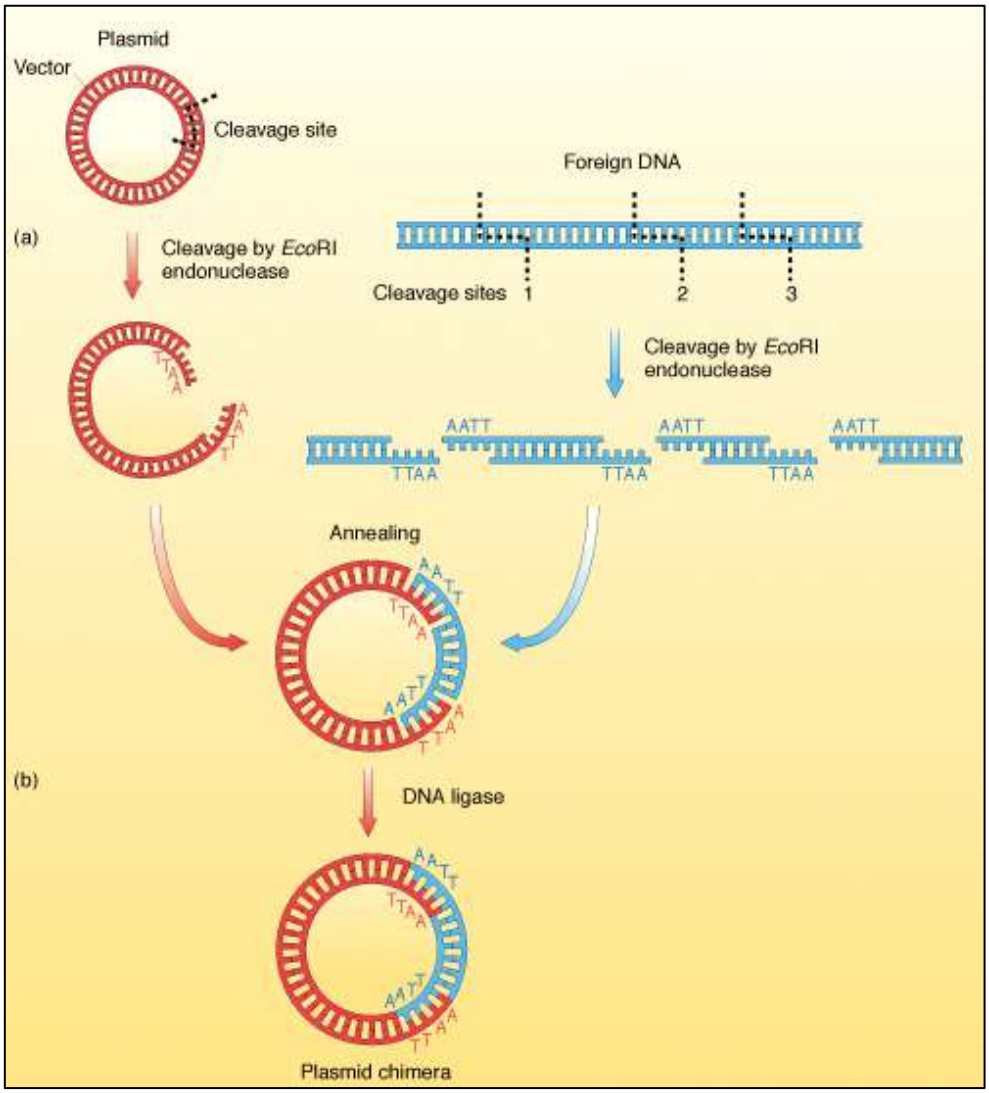


Figure 3: Plasmid

REQUIREMENTS:

The DNA Ligation Kit (Product No. LIGI) contains sufficient for 150 reactions.

Chemicals	Volume
10X Ligation Buffer 250 mM Tris-HCl, pH 7.8, 100 mM $MgCl_2$ 10 mM dithiothreitol	300 µL
T4 DNA Ligase 4 U/µL in 50% glycerol containing 10 mM Tris-HCl, pH 7.5 50 mM KCl 1 mM dithiothreitol	3 x 100 units
10 mM ATP	3 x 100 µL
Control DNA, pBR322 DNA HAE III Digest 0.5 µg/µL in 10 mM Tris-HCl pH 8.0, 1 mM EDTA	50 µL
24% (w/v) PEG Solution,	1.5 mL
Water	1.5 mL

Additional Reagents Required

1. Vector DNA (0.033-1 µg/µL)
2. DNA fragment to be inserted (0.033-1 µg/µL)
3. Microcentrifuge tubes, 0.5 mL and 1.5 mL
4. A low temperature water bath may also be required

Abbreviations

ATP = Adenosine 5'-triphosphate
PEG = Polyethylene glycol 8000
DTT = Dithiothreitol
EDTA = Ethylenediaminetetraacetic acid

PROCEDURE:

Reaction Volume: 10 µl

Chemicals	Volume (in µl)
Vector volume (µl)	5
Insert volume (µl)	5
D/W (µl)	8
10 X ligase buffer (µl)	1

T4 DNA ligase (5 Weiss U/mL) (µl)	1
Total volume (µl)	20

1. Incubate the vials at room temperature for two hours.
2. Load the samples on 1% agarose gel and observe for ligation.

OBSERVATIONS:

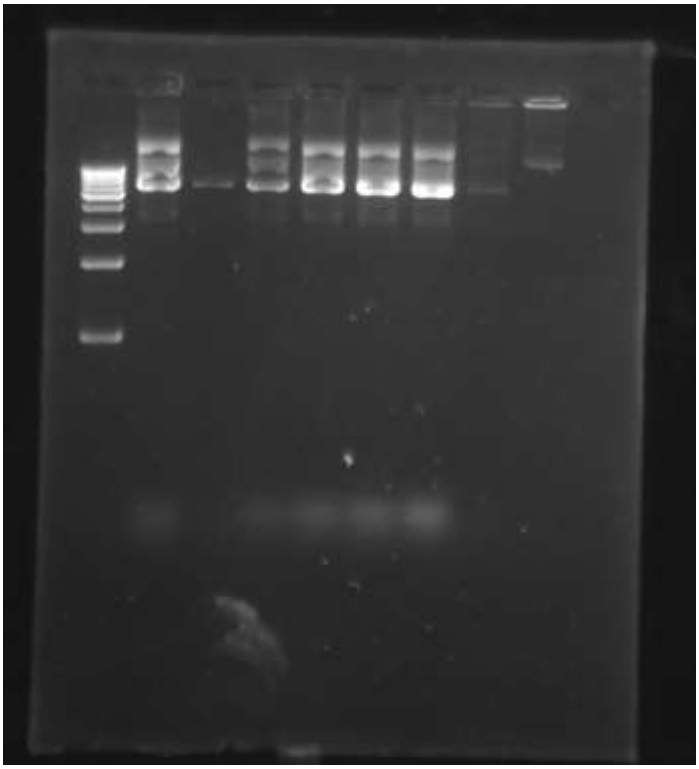


Figure 4: Ligation was done and was evaluated on 1% Agarose gel and stained with 0.1µg/mL of Ethidium Bromide.

RESULTS:

The construction of a recombinant plasmid is connecting the insert DNA (gene or fragment of interest) into a compatibly digested vector backbone. Then, the products are usually analyzed by running them on an agarose gel. By applying an electric current to the gel, DNA fragments will move through the gel at different rates based on their size and shape this visualizes and separate the different products of the ligation reaction. The resulting bands on the gel can then be analyzed to confirm the success of the ligation reaction and determine the size and quantity of the resulting DNA fragments. The two fragments of size 4000 and 1400bp were ligated and a resultant single band of 5400bp was seen on the gel.

CONCLUSION:

After ligation, the insert DNA is physically attached to the backbone and the complete plasmid can be transformed into bacterial cells for propagation. The majority of ligation reactions involve DNA fragments that have been generated by restriction enzyme digestion. DNA

ligation can be a useful tool in PCR for joining together DNA fragments to create longer DNA sequences. However, it is important to ensure that the ligation reaction is efficient and successful, and gel electrophoresis can be used to analyze the products of the ligation reaction in PCR.

DATE: 04/01/2024

PRACTICAL NO.: 4
SODIUM DODECYL SULPHATE POLYACRYLAMIDE GEL ELECTROPHORESIS (SDS–PAGE)

AIM:

To separate proteins on the basis of their molecular weights by the technique SDS-PAGE.

THEORY:

Electrophoresis is a method for separating macromolecules in an electric field, and one widely used approach for protein separation is sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The Laemmli method, named after U.K. Laemmli, is a commonly employed system in SDS-PAGE.

SDS, an anionic detergent also known as lauryl sulfate, imparts a negative charge to its molecules when dissolved, covering a broad pH range. SDS binds to polypeptide chains in proportion to their relative molecular mass. The negatively charged SDS disrupts the complex structure of proteins, drawing them towards the anode (positively charged electrode) in an electric field.

Polyacrylamide gels impede the migration of larger molecules, causing them to move more slowly than smaller ones. Since the charge-to-mass ratio is nearly uniform among SDS-denatured polypeptides, the final protein separation heavily relies on differences in their relative molecular masses. In a gel of uniform density, the relative migration distance of a protein (Rf, with 'f' as a subscript) is inversely proportional to the logarithm of its mass. Running known mass proteins alongside unknowns allows plotting the relationship between Rf and mass, aiding in the estimation of unknown protein masses.

SDS-PAGE for protein separation offers various applications, including estimating relative molecular mass, determining the abundance of major proteins, and assessing protein distribution in fractions. It aids in evaluating the purity of protein samples and tracking the progress of fractionation or purification procedures. Different staining methods help detect rare proteins and provide insights into their biochemical properties. Specialized techniques like Western blotting, two-dimensional electrophoresis, and peptide mapping are utilized to detect scarce gene products, identify similarities, and separate protein isoenzymes.

Polyacrylamide gels are crafted by reacting acrylamide and bis-acrylamide (N,N’-methylenebisacrylamide), resulting in a highly cross-linked gel matrix. Functioning as a sieve, the gel facilitates protein movement in response to the electric field. Proteins, having an overall positive or negative charge, migrate towards their isoelectric point where the molecule has no net charge. Denaturing the proteins, endowing them with a uniform negative charge, enables size-based separation as they move towards the positive electrode.

APPLICATIONS:

1. This is used to separate the HIV proteins during the HIV test.
2. The purity of the proteins is identified in this process.
3. SDS-PAGE is also used for peptide mapping.
4. This is used in measuring the molecular weight of the molecules.
5. The size of the protein is estimated by the process.
6. The polypeptide composition is compared in this process.

PRINCIPLE:

The SDS-PAGE principle hinges on the migration of charged molecules towards the electrode with the opposite sign when subjected to an electric field. The separation of these charged molecules relies on their relative mobility. Smaller molecules experience less resistance, leading to faster migration during electrophoresis. The protein's structure and charge also impact migration rates. The combination of sodium dodecyl sulfate (SDS) and polyacrylamide eliminates the influence of protein structure and charge, facilitating protein separation based on polypeptide chain length.

SDS disrupts the secondary, tertiary, and quaternary structure of proteins, resulting in a linear polypeptide chain enveloped by negatively charged SDS molecules. β-mercaptoethanol aids in protein denaturation by reducing all disulfide bonds. By subjecting the sample to denaturing and reducing conditions through heating, proteins unfold and bind with SDS detergent molecules, acquiring a substantial net negative charge proportional to the polypeptide chain length.

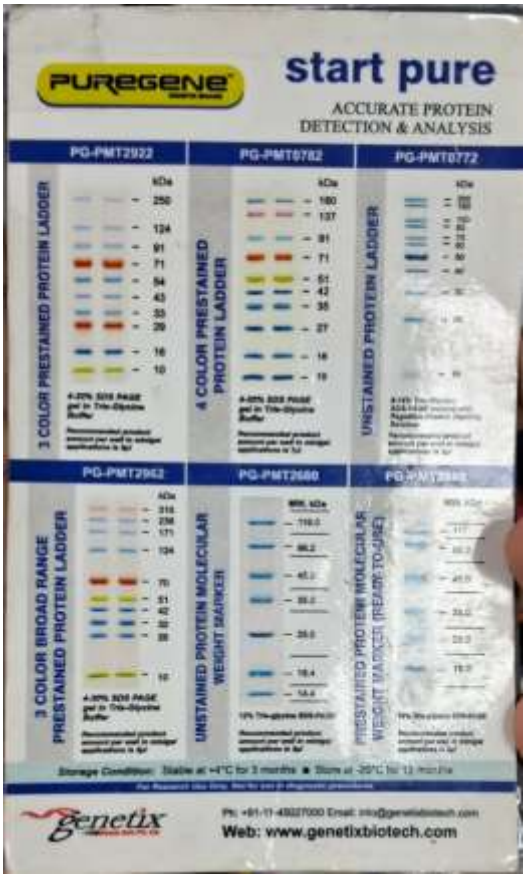


Figure 1: Protein ladder chart for accurate protein detection and analysis

REQUIREMENTS:

A) Reagents:

1. 30% Acrylamide
2. Tris HCl- pH: 8.8(1.5 M)
3. Tris HCl- pH: 6.8
4. 10% SDS
5. Gel loading Dye
   a. 0.2MpH 6.8 Tris Buffer
   b. 10% SDS
   c. 10mM Beta Mercaptoethanol
   d. 20% glycerol v/v
   e. 0.5% w/v Bromophenol Blue
6. 10% Ammonium persulphate
7. TEMED (it should be added at last)

Resolving gel composition: for 7.5 ml

Reagents	Stock	Working
1.5 M Tris pH: 8.8	1.5M	0.375 M
30% Acrylamide	30%	10%
SDS	10%	0.1%
APS 10%	10%	0.1%
TEMED	100%	0.07%
D/W (4ml)		

Stacking gel composition:

Reagents	Stock	Working
0.5 M Tris pH:.8	0.5M	0.125 M
30% Acrylamide	30%	5%
SDS	10%	0.1%
APS 10%	10%	0.1%
TEMED	100%	0.07%
D/W (4ml)		

B) Miscellaneous:

1. Comb (1mm)
2. Spacer (1mm)
3. Powerpack
4. Notch Plates

C) Staining dye:

1. Coomassie brilliant blue R250: 250mg
2. Glacial acetic acid: 10ml
3. Methanol: 40ml/100ml
4. D/W: 50ml

   Note: glacial acetic acid, methanol, D/W in ratio 1:4:5.

D) Destaining dye:

1. Glacial acetic acid :10ml
2. Methanol:40ml/100ml
3. D/W :50ml
4. destaining solution

   Note: glacial acetic acid, methanol, D/W in ratio 1:4:5.

Sample preparation:

Take 20 microlitre of sample in eppendorf tube

↓

Add 20 microlitre of D/W

↓

Add 10 microlitre of loading dye

↓

Heat at 95 ° C for 5 mins

↓

Load on gel

PROCEDURE:

1. Clean the glass plates and spacers of the gel casting unit with deionized water and ethanol.
2. Assemble the plates with the spacers on a stable, even surface.
3. Prepare the resolving gel according to the requirements.
4. Pour the gel solution in the plates assembled with spacers. To maintain an even and horizontal resolving gel surface, overlay the surface with water
5. Allow the gel to set for about 20-30 min at room temperature.
6. Prepare the staking gel according to the requirements.
7. Add the 5% stacking gel solution until it overflows. Insert the comb immediately ensuring no air bubbles are trapped in the gel or near the wells.
8. Let the gel set for 20-30 mins.
9. Remove the clips and transfer the prepared gel in the electrophoresis chamber
10. Load the ladder and sample in the wells
11. Perform electrophoresis.
12. Remove the gel wash it and immerse the gel in the solution of Coomassie brilliant blue dye and distilled water and keep the gel container on the rocking table for 45 mins.
13. Discard the solution and wash the gel with distilled water.

14. Keep the gel on the rocking table for 40 mins
15. Add destaining solution and keep the gel on the rocking table for 40 mins
16. Wash the gel and interpret the results.

OBSERVATIONS:

Figure 2: Destained gel

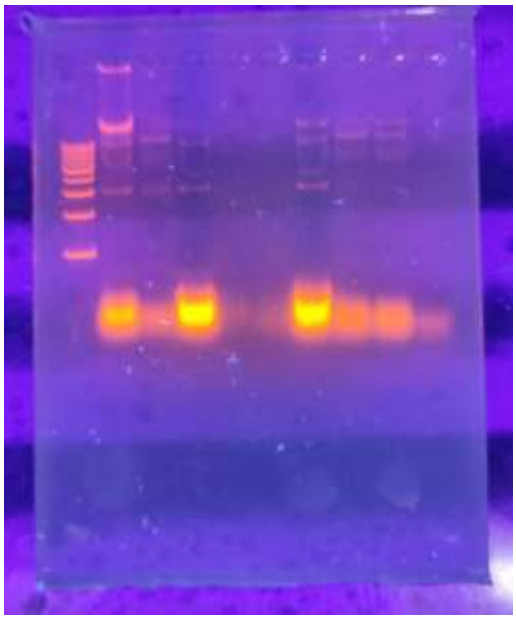


Figure 3: Gel analysed under UV light

RESULTS:

SDS-PAGE Technique was performed to separate the protein samples. The protein samples were loaded in the wells present on the gel. The electrophoresis was performed were the molecules which has higher molecular weight were seen on the top of the gel were as the those which were having lower molecular weight were at the bottom.

CONCLUSION:

Protein samples were separated on the basis of molecular weight by the technique of SDS-PAGE electrophoresis. Staining dye and destaining dye were used, The molecules present in the protein samples were separated on the basis of molecular weight, higher molecular weight molecules were present on the top and those which were having lower molecular weight were present at the bottom Polyacrylamide gel electrophoresis of SDS-treated proteins allows researchers to separate proteins based on their length in an easy, inexpensive, and relatively accurate manner were the proteins are separated on the basis of polypeptide chain length.

DATE: 09/01/2024

PRACTICAL NO.:5
POLYMERASE CHAIN REACTION (PCR)

AIM:

To amplify the given DNA sample by using specific primers in thermal cycler using PCR machine.

THEORY:

Kary B. Mullis developed PCR in the early 1980s; for this achievement, he shared the Chemistry Nobel Prize. Since then, PCR has developed into a common and crucial procedure in the field of molecular biology. It is employed in many scientific methods, including molecular cloning and molecular diagnostics. Little segments of DNA or a gene can be amplified (made numerous copies of) using PCR.

The technique has numerous uses in various industries, including genetic testing in medicine, disease detection in water supplies, and person identification in forensic science. A particular DNA segment can be quickly produced (amplified) in millions to billions of copies using the PCR process, allowing for more in-depth analysis. In PCR, a section of the genome to be amplified is chosen using short synthetic DNA fragments known as primers. That segment is then amplified using several cycles of DNA synthesis. It is conventional lab equipment used in biological and medical research.

It is employed in the initial phases of DNA processing for sequencing, in the identification of pathogens during infection by determining the presence or lack of a gene, and in the creation of forensic DNA profiles from microscopic DNA samples.

17S rRNA gene:

17S RNA, also known as 17S ribosomal RNA, is a type of RNA molecule found in prokaryotic cells, specifically in the ribosome, which is the cellular structure responsible for protein synthesis. It's a component of the small ribosomal subunit, which is essential for the translation of messenger RNA (mRNA) into proteins.

The 17S RNA molecule is single-stranded and relatively short compared to other RNA molecules. It folds into a complex three-dimensional structure, forming loops and stems, which are crucial for its function within the ribosome. These structural features enable the 17S RNA to interact with other ribosomal components and participate in the decoding of mRNA during translation.

Function of 17S RNA:

The primary function of 17S RNA is to help assemble the ribosome and facilitate protein synthesis. It plays a crucial role in the initiation of translation by binding to the mRNA and guiding the assembly of the ribosomal subunits. Additionally, 17S RNA is involved in maintaining the correct reading frame during translation, ensuring that the amino acids are added to the growing polypeptide chain in the proper order.

17S RNA is a key player in the process of gene expression, which involves the transcription of DNA into RNA and the subsequent translation of RNA into proteins. By participating in translation, 17S RNA helps to regulate the expression of genes by controlling the rate at which proteins are synthesized.

Understanding the structure and function of 17S RNA has significant implications in biotechnology and medicine. Researchers utilize this knowledge to develop antibiotics that target bacterial ribosomes, inhibiting protein synthesis and ultimately killing the bacteria. Additionally, the study of ribosomal RNA has led to advancements in genetic engineering techniques, such as CRISPR-Cas9, which allows for precise editing of genes in various organisms.

In summary, 17S RNA is a vital component of the ribosome involved in protein synthesis in prokaryotic cells. Its structure and function are fundamental to the process of translation and gene expression, with implications ranging from basic research to biotechnological applications. Understanding the intricacies of 17S RNA provides valuable insights into cellular biology and offers opportunities for innovation in medicine and biotechnology.

Components of PCR:

1. **DNA template:** The sample DNA that contains the target sequence. At the beginning of the reaction, high temperature is applied to the original double – stranded DNA molecule to separate the strands from each other.
2. **DNA polymerase:** a type of enzyme that synthesizes new strands of DNA complementary to the target sequence. The first and most commonly used of these enzymes is TaqDNA polymerase (from *Thermis aquaticus*), whereas PfuDNA polymerase (from *Pyrococcus furiosus*) is used widely because of its higher fidelity when copying DNA. Although these enzymes are subtly different, they both have two capabilities that make them suitable for PCR:
   a. They can generate new strands of DNA using a DNA template and primers.
   b. They are heat resistant.
3. **Primers:** short pieces of single-stranded DNA that are complementary to the target sequence. The polymerase begins synthesizing new DNA from the end of the primer.
4. **Nucleotides (dNTPs or deoxynucleotide triphosphates):** single units of the bases A, T, G, and C, which are essentially "building blocks" for new DNA strands.

RT-PCR(Reverse Transcription PCR) is PCR preceded with conversion of sample RNA into cDNA with enzyme reverse transcriptase.

Applications of PCR:

1. **DNA Amplification:** PCR allows the amplification of specific DNA sequences, making it invaluable for various downstream applications such as sequencing, cloning, and genetic analysis.
2. **Diagnostic Testing:** PCR is extensively used in medical diagnostics for the detection of infectious diseases, genetic disorders, and cancer. It enables the detection of pathogens like viruses and bacteria in clinical samples with high sensitivity and specificity.

3. **Cloning:** PCR may be used to amplify selected sequences for insertion into a vector. These sequences can be modified to include specific regions (tails) for cloning enzyme recognition. Primers directed to the vector are used to isolate fragments that have already been cloned into vectors. A low error rate DNA polymerase enzyme is necessary for the PCR steps.
4. **Genetic Engineering:** PCR facilitates the manipulation and modification of DNA sequences in genetic engineering applications. Techniques like site-directed mutagenesis and gene cloning heavily rely on PCR amplification.

PRINCIPLE:

Enzymatic DNA replication forms the basis of the PCR process. Using primer-mediated enzymes, a short DNA segment is amplified during the PCR process. New DNA strands that are complementary to the template DNA are created by DNA polymerase. Only the pre-existing 3'-OH group can have a nucleotide added to it by the DNA polymerase. Consequently, a primer is needed. The 3' prime end of the DNA polymerase thus receives more nucleotides.

A sequence of temperature cycles is used in the PCR process in order to denature, anneal, and lengthen DNA strands. A typical PCR reaction proceeds as follows:

1. **Denaturation:** The double-stranded DNA is denatured into single strands by heating the DNA sample to a high temperature (often between 94 – 96°C).
2. **Annealing:** Short oligonucleotide primers can anneal (bind) to the complementary portions of the single-stranded DNA template when the temperature is decreased to about 50–60°C. The primers serve as the initial points for the synthesis of new DNA strands by the enzyme DNA polymerase.
3. **Elongation:** At this step, the temperature is raised to 72-80℃. The bases are added to the 3’ end of the primer by the Taq polymerase enzyme. This elongates the DNA in the 5’ to 3’ direction. The DNA polymerase adds about 1000bp/minute under optimum conditions. Taq Polymerase can tolerate very high temperatures. It attaches to the primer and adds DNA bases to the single strand. As a result, a double-stranded DNA molecule is obtained.

These three steps are repeated 20-40 times in order to obtain a number of sequences of DNA of interest in a very short time period.

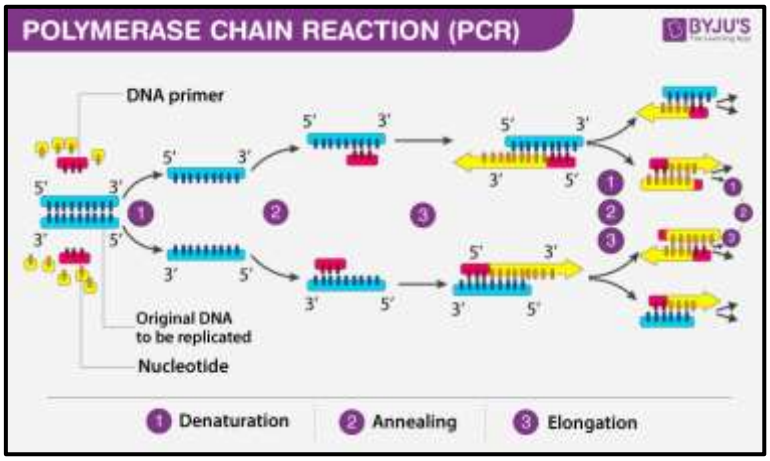


Figure 1: PCR Steps

REQUIREMENTS:

Chemicals & Reagents:

1. **Template DNA:** The desired sequence is contained in the DNA template. This could come from any source of DNA that has the target sequence, including plasmid DNA, cDNA, and genomic DNA.
2. **Primers:** synthesized, short, single-stranded DNA molecules that complement the target region's flanking regions. They serve as a binding site for the Taq polymerase enzyme and identify the beginning and ending points of the amplified region.
3. **Taq polymerase:** A heat-stable DNA polymerase enzyme that is resistant to the high temperatures required for the PCR cycle's denaturation stage.
4. **Deoxynucleoside triphosphates (dNTPs):** The building blocks of DNA that are incorporated into the growing DNA strand during PCR.
5. **Buffer:** A solution that provides the optimal pH and salt conditions for the PCR reaction.

For Result analysis: Agarose gel (1.5%)**,** Ethidium Bromide, Gel loading Buffer, DNALadder

Instruments: Thermocycler, Electrophoresis equipment, UV-transilluminator, Microwave,Weighing balance, Power pack.

Miscellaneous: Gloves, Micropipettes and tips, PCR tubes, Cello tape.

PROCEDURE:

A. Preparation of reaction mixture for PCR:

1. Add Reagents given in the kit as shown in order given in the Table.

Reagents	**Volume for 25µl reaction mixture (µL)**
Distilled water	15.5
10X Taq polymerase assay Buffer ($MgCl_2$)	2.5
dNTPs	0.5
Forward primers	2.5
Reverse primers	2.5
Taq DNA Polymerase	0.5
DNA Template	1
Total Reaction Mixture Volume	**25**

2. Mix the contents gently and tap spin the master mixture. Place the PCR tubes in the thermal cycler machine for PCR amplification.

Steps	**Temperature**	**Time**
Initial denaturation	95°C	1 min
Denaturation	94°C	30 sec
Annealing	55°C	30 sec
Extension	72°C	30 sec
Final extension	72°C	5min

3. Carry out the amplification using the following conditions and perform at least 35 cycles for amplification of the DNA fragment.
4. Load the sample on 1.5% agarose gel.

OBSERVATIONS:

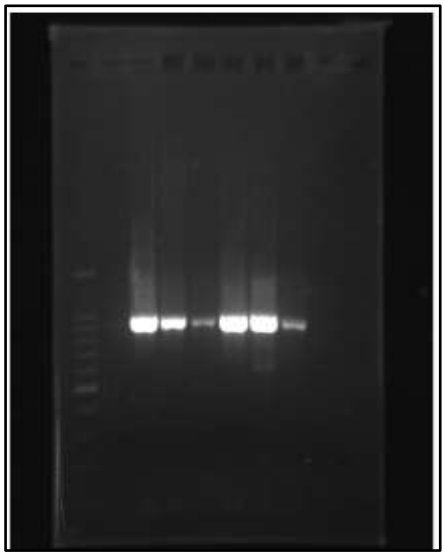


Figure 2: PCR was performed on a gene of interest 17S rRNA and was evaluated on 1.5% Agarose gel and stained with 0.1µg/mL of Ethidium Bromide

RESULTS:

PCR of gene '17S rRNA' was done and analyzed by 1.5% Agarose gel electrophoresis. In order to compare and examine the base pairs of the amplified gene, the first well on the agarose gel was loaded with the molecular ladder, and the second well contains the relevant sample gene. The amplified bands were around 900 base pairs in size, according to the electrophoresis data. The amplified gene or band's intensity suggested that the sample was of high quality and quantity. On the other hand, no band implies that the amplification was ineffective.

CONCLUSION:

Polymerase Chain Reaction (PCR) technique was performed using a 17S rRNA sample. It is a highly accurate and rapid method used for replicating genetic material to make multiple copies of samples. The results obtained indicate PCR conditions are optimized for efficient amplification. The amplified product was visualized on the Agarose Gel and the size was confirmed using the DNA ladder.

PCR experiment is dependent upon many factors such as quality of reagents, primers, and PCR conditions. The PCR has ability to amplify a specific DNAs fragment from a complex mixture of the DNA. The use of PCR has numerous applications in molecular biology, gene expression analysis etc. PCR is the valuable technique for diagnosis disease and variation in the gene.

DATE: 08/01/2024

PRACTICAL NO.: 6
DEMONSTRATION OF DNA SEQUENCER

AIM:

To demonstrate the working of a DNA Sequencer.

THEORY:

DNA sequencers are instruments designed to read a DNA sample and produce an electronic file containing symbols representing the sequence of nitrogen bases – A, C, G, T – in the sample. These scientific instruments automate the DNA sequencing process, crucial for understanding genetic information. When provided with a DNA sample, a DNA sequencer identifies the order of the four bases: G (guanine), C (cytosine), A (adenine), and T (thymine), presenting the information as a text string called a "read". Some DNA sequencers can be classified as optical instruments, as they analyze light signals emitted by fluorochromes attached to nucleotides.

The first automated DNA sequencer, introduced by Applied Biosystems in 1987 and invented by Lloyd M. Smith, utilized the Sanger sequencing method. This technology marked the inception of the "first generation" of DNA sequencers, playing a pivotal role in completing the Human Genome Project in 2001. These initial sequencers were essentially automated electrophoresis systems that detected the migration of labeled DNA fragments. Besides DNA sequencing, they were also employed in genotyping genetic markers, particularly in scenarios where only the length of DNA fragments (e.g., microsatellites, AFLPs) needed determination. The field witnessed a significant advancement with the advent of next-generation sequencing (NGS) technologies, commercialized around 2005. These technologies have since evolved rapidly, enabling DNA sequencing on platforms capable of generating information about millions of base pairs in a single run.

Two primary DNA sequencing methods include:

1. Sanger Sequencing (Chain Termination Sequencing Method)
2. Maxam and Gilbert Sequencing (Chemical Sequencing Method)

Sanger Sequencing

Sanger sequencing is a DNA sequencing method that involves electrophoresis and relies on the incorporation of chain-terminating dideoxynucleoides by DNA polymerase during in vitro DNA replication. Developed by Frederick Sanger and colleagues in 1977, it served as the predominant sequencing method for approximately 40 years and was first commercially introduced by Applied Biosystems in 1986. In recent years, next-generation sequencing methods have largely supplanted higher volume Sanger sequencing, especially in large-scale automated genome analyses.

Despite its diminished role in high-throughput projects, the Sanger method continues to find use in smaller-scale initiatives and for validating deep sequencing results. One notable advantage it retains over short-read sequencing technologies like Illumina is its ability to generate DNA sequence reads exceeding 500 nucleotides while maintaining a remarkably low

error rate, approximately 99.99% accuracy. Sanger sequencing remains actively employed in various public health initiatives, including sequencing the spike protein of SARS-CoV-2 and participating in the surveillance of norovirus outbreaks through the Center for Disease Control and Prevention's (CDC) CaliciNet surveillance network.

While Sanger sequencing has been a widely utilized DNA sequencing method, it is not without limitations such as cost, time-consumption, and susceptibility to artifacts. These drawbacks have prompted the development of newer sequencing technologies.

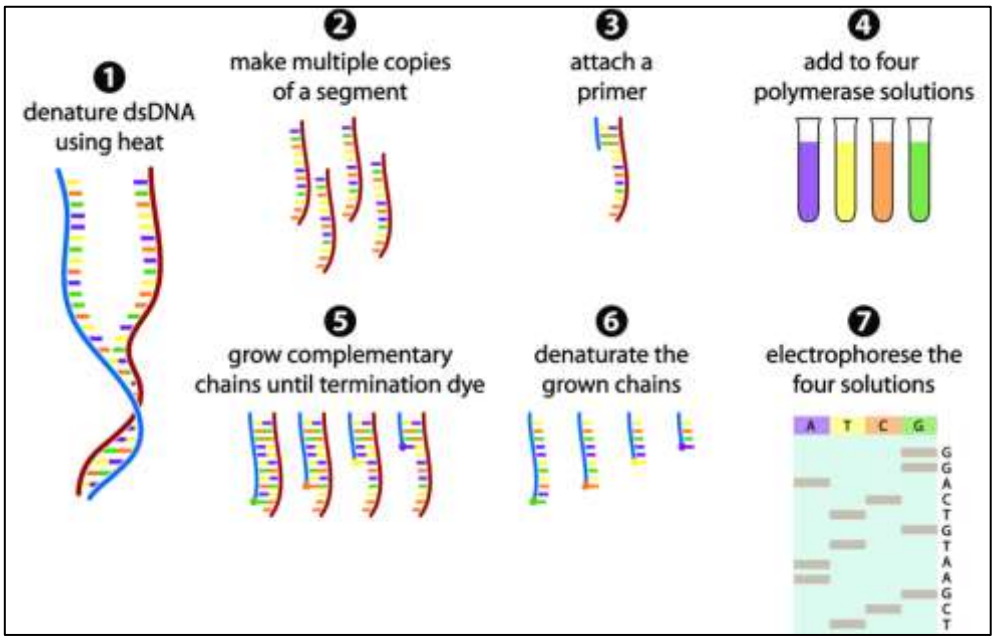


Figure 1: Steps of Sanger Sequencing

Maxam-Gilbert Sequencing

Maxam–Gilbert sequencing, developed by Allan Maxam and Walter Gilbert in 1976–1977, is a DNA sequencing method that relies on nucleobase-specific partial chemical modification of DNA followed by the cleavage of the DNA backbone at sites adjacent to the modified nucleotides. This approach marked a significant breakthrough in the early days of DNA sequencing and, along with the Sanger dideoxy method, constitutes the first generation of DNA sequencing methods.

Initially, Maxam–Gilbert sequencing was widely adopted as the first method for DNA sequencing. However, with the advent of next-generation sequencing methods, it has gradually fallen out of widespread use. Sanger sequencing, along with newer technologies like next-generation sequencing (NGS) and third-generation sequencing (TGS), has largely supplanted Maxam & Gilbert's method. While Maxam–Gilbert sequencing played a crucial role in the history of DNA sequencing, its utilization has diminished over time, giving way to more advanced and efficient sequencing techniques.

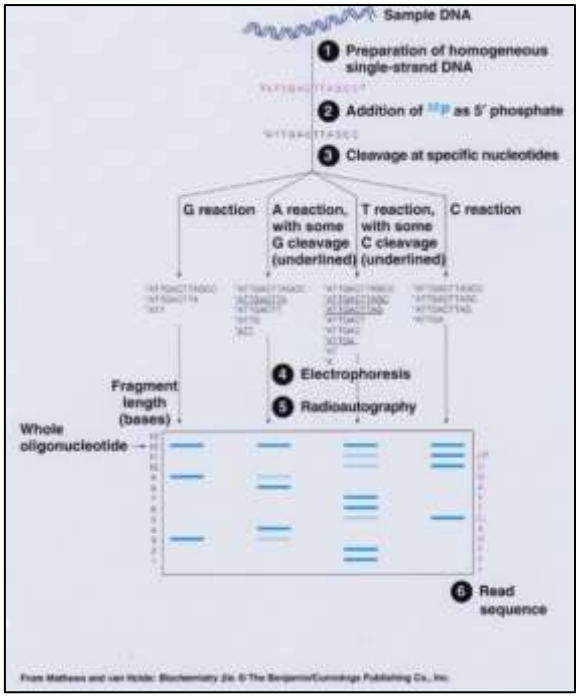


Figure 2: Steps of Maxam-Gilbert Sequencing

ADVANTAGES OF DNA SEQUENCER:

1. **High throughput:** Modern DNA sequencers can process millions of DNA fragments in parallel, allowing researchers to sequence entire genomes in a matter of days or weeks.
2. **Accuracy:** Advances in sequencing technology have led to higher accuracy rates and lower error rates, making it easier to generate high-quality sequencing data.
3. **Speed:** DNA sequencers can generate results quickly, allowing researchers to rapidly analyze and interpret the data.
4. **Versatility:** DNA sequencers can be used to sequence a wide range of DNA samples, including human genomes, microbiomes, and environmental DNA.
5. **Accessibility:** With the increasing availability of affordable DNA sequencers, more researchers and institutions have access to this technology, making it easier to conduct genomic research.
6. **Advancements in research:** DNA sequencing has revolutionized the field of genomics, leading to important discoveries about the genetic basis of disease, evolution, and biodiversity.

DISADVANTAGES OF DNA SEQUENCER:

1. **Cost:** DNA sequencing technology can be expensive, particularly for high-throughput sequencing applications.
2. **Limited read length:** The read length of DNA sequencers is limited, which can make it challenging to assemble genomes and identify structural variants.
3. **Errors:** Sequencing errors can occur due to a variety of factors, such as base-calling errors, PCR errors, and sample contamination, which can affect the accuracy of the sequencing data.

4. **Sample preparation:** DNA sequencing requires careful sample preparation, which can be time-consuming and prone to errors.
5. **Bioinformatics analysis:** DNA sequencing generates massive amounts of data that require sophisticated bioinformatics tools and expertise to analyze, which can be a challenge for many researchers.
6. **Data storage and management:** DNA sequencing generates large amounts of data that require significant storage and computational resources to manage and analyze.

APPLICATIONS:

1. **Genomic analysis:** DNA sequencers can be used to sequence the entire genome of an organism, providing insight into the genetic makeup of the organism.
2. **Evolutionary studies:** DNA sequencing can be used to study the evolution of species by comparing the DNA sequences of different organisms.
3. **Disease diagnosis:** DNA sequencing can be used to identify genetic mutations or variations associated with specific diseases, allowing for early detection and personalized treatment.
4. **Biotechnology and genetic engineering:** DNA sequencing can be used to identify genes that code for specific traits, allowing researchers to engineer organisms with desired traits.
5. **Forensics:** DNA sequencing can be used to identify suspects in criminal investigations by comparing DNA samples from crime scenes to DNA samples from potential suspects.
6. **Microbial diversity studies:** DNA sequencing can be used to study the diversity of microbial communities in various environments, including soil, water, and the human body.
7. **Personalized medicine:** DNA sequencing can be used to develop personalized treatment plans based on an individual's genetic makeup.
8. **Agriculture:** The mapping and sequencing of a genome of microorganisms has helped to make them useful for crops and food plants.

Overall, DNA sequencers have revolutionized the field of genetics and have numerous applications in research, medicine, and beyond.

PRINCIPLE:

Sequencing is the process of determining the order of nucleotide bases along a DNA strand. In 1977, two distinct methods for DNA sequencing emerged: the chain termination method and the chemical degradation method. Initially equally popular, the chain termination method has become more prevalent for various reasons and is widely utilized today. This method relies on the principle that single-stranded DNA molecules differing in length by just a single nucleotide can be separated using polyacrylamide gel electrophoresis, as previously explained.

To initiate sequencing, the DNA to be sequenced, known as template DNA, is first prepared as a single-stranded molecule. Subsequently, a short oligonucleotide is annealed to the same position on each template strand, acting as a primer for the synthesis of a new DNA strand complementary to the template DNA. This approach necessitates performing four nucleotide-specific reactions, one each for G, A, C, and T, on four identical samples of DNA.

The four sequencing reactions require the addition of all the components necessary to synthesize and label new DNA, including:

1. A DNA template;
2. A primer tagged with a mildly radioactive molecule or a light-emitting chemical;
3. DNA polymerase--an enzyme that drives the synthesis of DNA;
4. Four deoxynucleotides (G, A, C, T);
5. One dideoxynucleotide, either ddG, ddA, ddC, or ddT.

Following the addition of the first deoxynucleotide to the growing complementary sequence, DNA polymerase proceeds along the template, continuously adding base after base. The synthesis of the strand persists until a dideoxynucleotide is introduced, preventing further elongation due to the absence of a 3'-hydroxyl group required for connecting with the next nucleotide.

In each reaction, a small quantity of a dideoxynucleotide is incorporated, enabling reactions to proceed for various durations. Eventually, by chance, DNA polymerase inserts a dideoxynucleotide, terminating the reaction. Consequently, a collection of new DNA chains of varying lengths is produced.

To decipher the newly generated sequence, the four reactions are conducted side-by-side on a polyacrylamide sequencing gel. The molecules generated in the presence of ddATP are loaded into one lane, while the other three families, generated with ddCTP, ddGTP, and ddTTP, are loaded into three adjacent lanes. Post-electrophoresis, the DNA sequence can be directly read from the positions of the bands in the gel.

Several variations of this method have been devised for automated sequencing machines. In one such method, known as cycle sequencing, the dideoxynucleotides—rather than the primers—are tagged with differently colored fluorescent dyes. This allows all four reactions to occur in the same tube and be separated in a single lane on the gel. As each labeled DNA fragment passes a detector at the bottom of the gel, the color is recorded, and the sequence is reconstructed from the pattern of colors representing each nucleotide in the sequence.

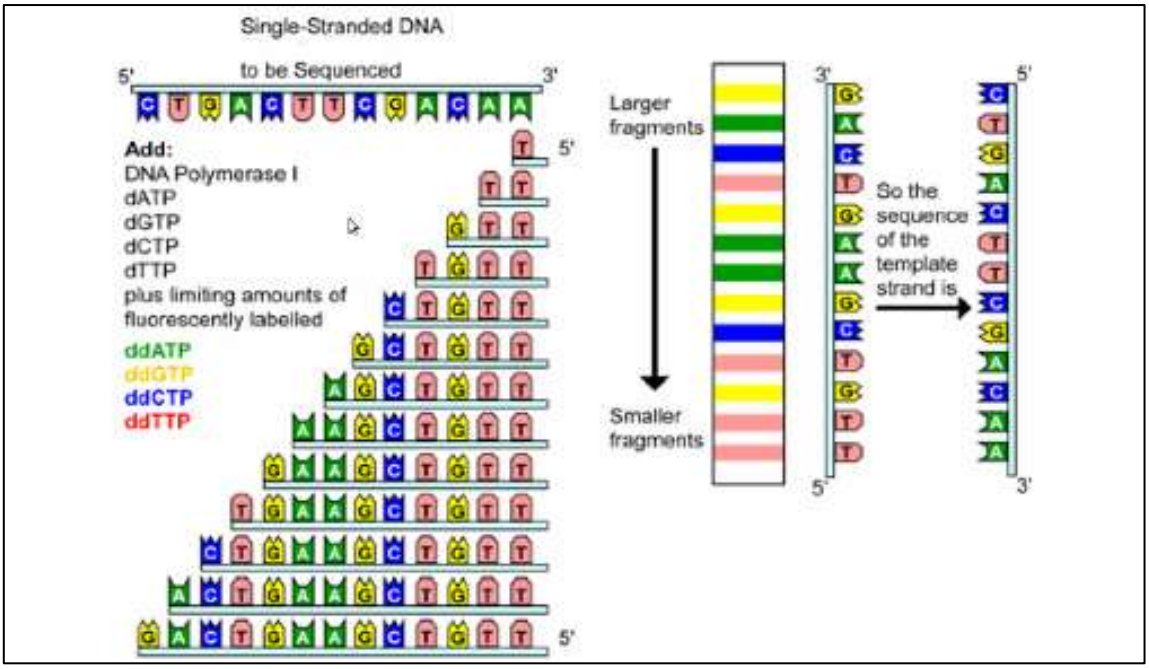


Figure 3: Principle of DNA Sequencing

REQUIREMENTS:

1. **DNA sample:** This can be obtained from a wide range of sources, including cells, tissues, blood, saliva, or environmental samples.
2. **Library preparation kit:** This kit contains enzymes and reagents needed to fragment and amplify the DNA, add adapters, and prepare the library for sequencing.
3. **Sequencing reagents:** This includes the reagents needed for sequencing, such as the nucleotides (A, C, G, T), sequencing polymerases, and fluorescent dyes.
4. **Sequencing instrument:** This is the actual DNA sequencer that reads the DNA sequence and generates the data.
5. **Computer and software:** This is used to process and analyze the raw sequencing data, including base calling, sequence alignment, and variant calling.

STEPS OF DNA SEQUENCING:

1. **Sample preparation:** DNA is extracted from the sample and purified to remove contaminants.
2. **Library preparation:** The DNA is fragmented into small pieces and adapters are added to each end to enable the DNA to be amplified and sequenced.
3. **Amplification:** The DNA fragments are amplified using PCR (polymerase chain reaction) to create multiple copies of each fragment.
4. **Sequencing:** The amplified DNA is loaded onto the sequencer, which reads the DNA sequence. There are different sequencing technologies available, including Illumina, PacBio, and Oxford Nanopore, among others.
5. **Base calling:** After the sequencing is complete, the raw data is processed by a computer to identify the base calls (A, C, G, or T) for each read.
6. **Assembly:** The reads are then assembled into longer contiguous sequences (contigs) using specialized software.
7. **Analysis:** The assembled sequences are compared to reference genomes or other databases to identify genes, mutations, or other features of interest.

CONCLUSION:

The DNA Sequencer practical provided valuable insights into the fascinating world of genomics and molecular biology. By utilizing DNA sequencing technology, we were able to decipher the genetic code with precision and efficiency. The practical showcased modern sequencing techniques and also emphasized their practical applications in various scientific domains, including genetics, medicine and biotechnology.

DATE: 08/01/2024

PRACTICAL NO.: 7
DEMONSTRATION OF FLOW CYTOMETRY

AIM:

To study the working and principle of Flow Cytometry.

THEORY:

Flow cytometry is a technique designed for counting, examining, and sorting microscopic particles suspended in a fluid stream. It enables the simultaneous multiparametric analysis of the physical and/or chemical characteristics of individual cells as they flow through an optical and/or electronic detection apparatus.

The inception of fluorescence-based flow cytometry can be attributed to the development of the ICP 11 device in 1968 by Wolfgang Göhde from the University of Münster, Germany. Subsequently, it was first commercialized in 1968-69 by the German developer and manufacturer Partec through Phywe AG in Göttingen. Originally known as "Pulse Cytophotometry," this technology marked the pioneering steps into the field of flow cytometry.

ADVANTAGES OF FLOW CYTOMETRY:

1. **High-throughput analysis:** Flow cytometers can rapidly analyze thousands of cells per second, facilitating efficient data acquisition.
2. **Multiparametric analysis:** Modern flow cytometers can simultaneously measure up to 30 parameters, offering detailed insights into various cell characteristics.
3. **Single-cell analysis:** Flow cytometry enables the examination of individual cells, crucial for understanding heterogeneous cell populations.
4. **Cell sorting:** Flow cytometers can sort cells based on specific characteristics, facilitating the isolation of distinct cell populations for further study.
5. **Quantitative measurements:** Flow cytometry allows for the quantitative assessment of cell populations and measurement of cellular characteristics, such as size, cell cycle status, and protein expression.
6. **High sensitivity:** Modern flow cytometers exhibit high sensitivity, enabling the detection of low-abundance cell populations and the identification of rare cell subsets.
7. **Automation:** Flow cytometry is an automated technique, reducing the risk of human error and ensuring result consistency.
8. **Non-destructive analysis:** Flow cytometry does not damage cells, allowing for the analysis of live cells and the potential for additional studies.

DISADVANTAGES OF FLOW CYTOMETRY:

1. **Requires fresh samples:** Incorrect storage or prolonged storage can lead to apoptosis, reducing the specificity of flow cytometric analysis.
2. **Cannot be used on fixed tissue:** Flow cytometry is incompatible with formalin-fixed tissue, lacking morphological correlation.

3. **Technical issues:** The complexity of equipment necessitates trained operators, and potential errors include mislabeled samples, incorrect antibody addition, and reagent issues.
4. **Analytical issues:** Data interpretation can be challenging due to the absence of morphological correlation, issues in excluding debris/doublets, and errors in gating strategies.
5. **Expensive and sophisticated instruments:** Flow cytometers are costly, requiring highly trained specialists and ongoing maintenance.
6. **Limited information on intracellular distributions:** While providing data on cell surface markers, flow cytometry lacks detailed information on intracellular distributions.
7. **Slow for some applications:** High-speed sorters may be too slow for certain experiments, discarding cell pairs due to the inability to distinguish between them quickly.
8. **Not suitable for all cell types:** Flow cytometry is not ideal for studying tissue structure and is unsuitable for some cell types that do not form single-cell suspensions.

APPLICATIONS OF FLOW CYTOMETRY:

In molecular biology, flow cytometry, particularly when utilizing fluorescence-tagged antibodies, is valuable for studying specific characteristics of target cells. These antibodies bind to the target cells, providing essential information when analyzed in the cytometer. Its applications extend to various medical fields such as transplantation, hematology, tumor immunology, chemotherapy, genetics, and sperm sorting for sex pre-selection. Additionally, flow cytometry finds utility in marine biology, where it leverages the auto-fluorescent properties of photosynthetic plankton to characterize abundance and community structure. In protein engineering, flow cytometry is employed in conjunction with yeast and bacterial displays to identify cell surface-displayed protein variants with desired properties.

PRINCIPLE:

In flow cytometry, a focused stream of fluid is illuminated by a single-wavelength beam of light, typically generated by a laser. Several detectors are strategically positioned at the point where the fluid stream intersects with the light beam. Among these detectors are one aligned with the light beam, known as Forward Scatter (FSC), and several positioned perpendicular to it, including Side Scatter (SSC) and one or more fluorescent detectors.

As suspended particles ranging from 0.2 to 150 mm pass through the light beam, they scatter the light in various ways. Additionally, fluorescent chemicals within the particles or attached to them may be excited, emitting light at a higher wavelength than the light source. The detectors capture this combination of scattered and fluorescent light. By analyzing fluctuations in brightness at each detector (with one for each fluorescent emission peak), diverse information about the physical and chemical structure of each individual particle can be derived.

Forward Scatter (FSC) is correlated with the cell volume, while Side Scatter (SSC) depends on the inner complexity of the particle. This complexity includes factors such as the shape of the

nucleus, the quantity and type of cytoplasmic granules, or the roughness of the membrane. Some flow cytometers on the market have eliminated the necessity for fluorescence and rely solely on light scatter for measurement. Others go a step further by generating images of each cell's fluorescence, scattered light, and transmitted light. This comprehensive approach allows for a detailed analysis of the characteristics of individual particles in the fluid stream.

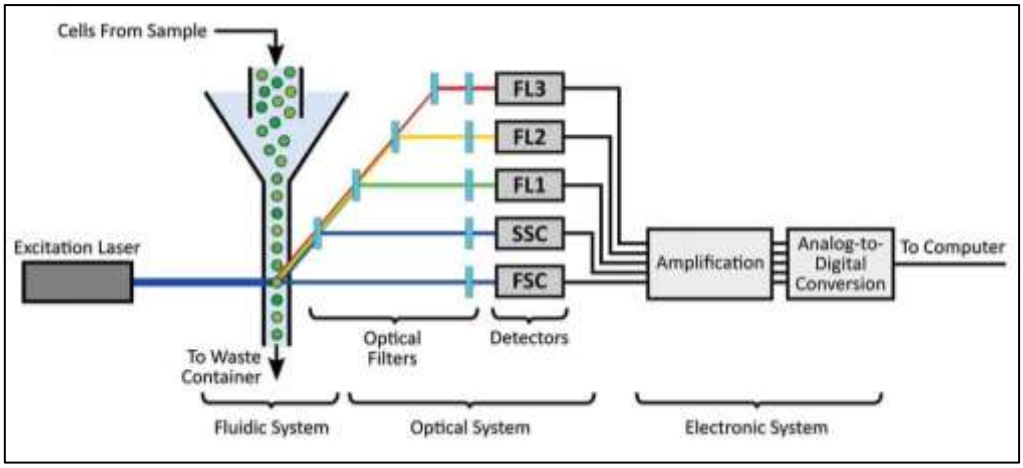


Figure 1: Component diagram of Flow Cytometer

Components of a Flow Cytometer:

A flow cytometer has 5 main components:

1. **Flow cell:** liquid stream (sheath fluid) carries and aligns the cells so that they pass single file through the light beam for sensing.
2. **Optical system:** commonly used are lamps (mercury, xenon); high power water-cooled lasers (argon, krypton, dye laser); low power air-cooled lasers (argon (488nm), red-HeNe (633mm), green-HeNe, HeCd (UV)); diode lasers (blue, green, red, violet) resulting in light signals.
3. **Detector and Analogue to Digital Conversion (ADC) system:** generating FSC and SSC as well as fluorescence signals from light into electrical signals that can be processed by a computer.
4. **Amplification system**
5. **Computational analysis of the signals.**

Figure 2: Flow Cytometer

Initially, flow cytometers were primarily experimental devices. However, recent technological advancements have spurred a significant market for both the instrumentation and the reagents essential for analysis, such as fluorescently labeled antibodies and specialized analysis software.

Modern flow cytometry instruments have evolved to include multiple lasers and fluorescence detectors. Notably, the current record for a commercial instrument boasts 4 lasers and 18 fluorescence detectors. This increase in the number of lasers and detectors enables the simultaneous use of multiple antibodies for labeling, allowing for more comprehensive and intricate analyses of cellular characteristics. The enhanced capabilities of modern flow cytometers contribute to their widespread use in various fields, including immunology, cell biology, and clinical diagnostics.

Fluorescence-Activated Cell Sorting (FACS):

Fluorescence-activated cell sorting (FACS) is a specialized form of flow cytometry that allows for the sorting of a heterogeneous mixture of biological cells into two or more containers based on their specific light scattering and fluorescent characteristics. This sorting occurs one cell at a time.

In FACS, the cell suspension is directed to the center of a narrow, swiftly flowing liquid stream. The flow is meticulously arranged to maintain a substantial separation between cells relative to their diameter. A vibrating mechanism induces the stream of cells to break into individual droplets. To minimize the probability of more than one cell being in a droplet, the system is finely adjusted. Just before the stream forms droplets, it passes through a fluorescence measuring station where the fluorescent characteristics of each cell are assessed.

An electrical charging ring is positioned precisely where the stream breaks into droplets. Based on the preceding fluorescence intensity measurement, a charge is applied to the ring, and the opposite charge is retained on the droplet as it separates from the stream. These charged droplets then pass through an electrostatic deflection system, guiding them into containers based on their charge. In some systems, the charge is directly applied to the stream, and the droplet retains a charge of the same sign as the stream. The stream is subsequently returned to a neutral state after the droplet separates. This sorting mechanism allows for the precise isolation of cells based on their unique characteristics, offering a powerful tool in various fields of research and clinical applications.

Antibody Staining in Flow Cytometry:

To conduct antibody staining in flow cytometry, it is imperative that the cells under analysis are in a single-cell suspension. In cases where cells are clumped or present in solid organs, enzymatic digestion or mechanical dissociation of the tissue is employed to convert them into a single-cell suspension. Mechanical filtration follows to prevent instrument clogs, ensuring higher quality flow data. The resulting cells are then incubated with either mislabelled or fluorescently conjugated antibodies in test tubes or microtiter plates before analysis using a flow cytometer.

Running Samples:

Once the sample is prepared, cells are coated with fluorochrome-conjugated antibodies specific to surface markers on various cells. This can be achieved through direct, indirect, or intracellular staining.

1. **Direct Staining:** Cells are incubated with antibodies directly conjugated to a fluorophore.
2. **Indirect Staining:** In this method, cells are first incubated with antibodies directly conjugated to a fluorophore. Subsequently, a fluorophore-conjugated secondary antibody is introduced to detect the primary antibody.
3. **Intracellular Staining:** This procedure allows for the direct measurement of antigens within the cell cytoplasm or nucleus. To achieve this, cells are first made permeable, and then they are stained with antibodies in a permeabilization buffer.

These staining methods enable the identification and characterization of specific cell populations based on surface or intracellular markers. Flow cytometry provides a powerful tool for understanding the composition and properties of complex cell populations in biological samples.

CONCLUSION:

The theory, principle and working of flow cytometry was studied.

DATE: 08/01/2024

PRACTICAL NO.: 8
DEMONSTRATION OF REAL-TIME POLYMERASE CHAIN REACTION (RT-PCR)

AIM:

To understand the technique of Real-Time Polymerase Chain Reaction (RT-PCR).

THEORY:

Real-time Polymerase Chain Reaction (PCR) is a widely utilized molecular biology technique for the real-time detection and quantification of nucleic acid sequences. Built upon the PCR method, it involves the amplification of a specific DNA sequence using primers and a DNA polymerase enzyme. Unlike conventional PCR, real-time PCR allows the measurement of PCR products as they are being amplified, in real-time. It incorporates a fluorescent dye or probe that binds specifically to the PCR products during synthesis, with emitted fluorescence proportional to the amount of generated PCR product, enabling quantification.

The process begins with reverse transcription (RT), where RNA is converted into complementary DNA (cDNA) using reverse transcription enzymes. After denaturing the RNA-cDNA hybrid, PCR amplification is performed to generate multiple copies of the target cDNA sequence. This involves repeated cycles of denaturation, annealing, and extension. The denaturation step separates DNA strands, annealing allows primers to bind, and extension synthesizes new DNA strands.

Various methods are employed for detecting amplified DNA products in real-time PCR, including fluorescent dyes that intercalate into double-stranded DNA or sequence-specific probes that hybridize to the target DNA sequence. Probe-based real-time PCR utilizes fluorescent probes designed to hybridize with the PCR product, containing a fluorescent reporter and quencher. The polymerase cleaves the probe during amplification, resulting in increased fluorescence proportional to the PCR product amount.

Real-time PCR, or quantitative PCR (qPCR), has diverse applications in molecular biology, such as gene expression analysis, pathogen detection, and genetic variation analysis. It is a powerful and sensitive technique capable of detecting low amounts of DNA or RNA, making it valuable in research and clinical applications.

The PCR process can be divided into four phases: linear ground phase, early exponential phase, log-linear (exponential) phase, and plateau phase. The linear ground phase initiates PCR, with fluorescence emission yet to rise above background. Baseline fluorescence is calculated during this phase. The early exponential phase sees fluorescence significantly higher than background, defining the cycle threshold (Ct). The log-linear phase is the optimal amplification period, with the PCR product doubling after each cycle in ideal conditions. The plateau phase occurs when reaction components become limited, and fluorescence intensity is no longer useful for data calculation.

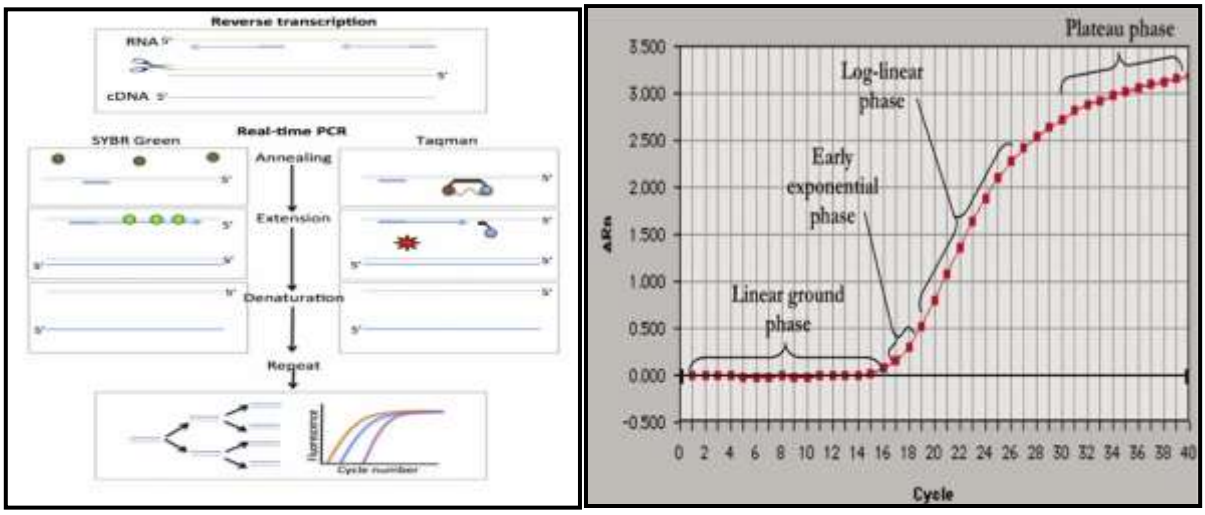


Figure 1: Real-time Polymerase Chain Reaction(RT-PCR) and its graphical representation to interpret the results

Figure 2: Real-time Polymerase Chain Reaction (RT-PCR)

APPLICATIONS:

1. **Gene expression analysis:** RT-PCR allows researchers to quantify the expression levels of specific genes in different tissues, cells, or experimental conditions.
2. **Cancer research:** RT-PCR enables the detection of cancer-specific gene mutations or alterations, aiding in early diagnosis, prognosis, and monitoring of treatment response.
3. **Drug research:** RT-PCR assists in studying drug metabolism, identifying drug targets, and assessing individual responses to medications through pharmacogenetic studies.
4. **Agricultural Applications:** RT-PCR is used in crop improvement programs to study gene expression patterns related to plant development, stress responses, and disease resistance.
5. **Food Safety:** RT-PCR can detect foodborne pathogens such as Salmonella, Listeria, and E. coli, ensuring the safety of food products.

PRINCIPLE:

Reverse Transcription Polymerase Chain Reaction (RT-PCR) is a laboratory technique employed for the amplification and detection of RNA molecules. It operates by converting RNA into complementary DNA (cDNA) through reverse transcription, followed by the amplification of the cDNA using the polymerase chain reaction (PCR) method. RT-PCR finds widespread use in molecular biology research and diagnostic applications, particularly in studying gene expression, quantifying viral loads, and detecting RNA viruses.

Real-time PCR, a variant of RT-PCR, relies on the incorporation of a fluorescent dye. The underlying principle involves quantifying the amount of nucleic acid in the sample using fluorescent dye or fluorescent-labeled oligos. In this method, when the dye or probe binds to the target template, it releases fluorochrome, emitting fluorescence that is detected by the fluorometer. The detector captures the emitted signal as an indication of positive template amplification.

REQUIREMENTS:

1. **Thermal Cycler:** The thermal cycler, also known as a PCR machine or thermocycler, is a pivotal component in RT-PCR. It orchestrates precise temperature changes during the repeated cycles required for different PCR steps. This device ensures accurate temperature control throughout the PCR cycles.
2. **Fluorescent Dye or Probe:** RT-PCR commonly employs fluorescent dyes, such as SYBR Green, which intercalate into double-stranded DNA during amplification. When excited by a suitable light source, these dyes emit fluorescence. While SYBR Green does not necessitate sequence-specific probes, it can bind nonspecifically to any double-stranded DNA, including primer dimers or other non-specific PCR products, potentially causing false-positive signals. Both fluorescent dye and probe-based methods enable real-time monitoring of PCR amplification, allowing quantification of target RNA or DNA based on emitted fluorescence during each PCR cycle.
3. **PCR Tubes or Plates:** PCR tubes, typically made of polypropylene, are thin-walled tubes designed to fit into thermal cyclers. They can hold reaction volumes ranging from 0.1 mL to 0.5 mL. Available in individual tube formats or strips, these tubes facilitate efficient heat transfer during thermal cycling. PCR plates, made of materials like polypropylene or polycarbonate, are compatible with robotic liquid handling systems, making them suitable for automated PCR workflows.
4. **Optical Detection System:** The optical detection system monitors fluorescence during PCR amplification. It detects and quantifies fluorescent signals from dyes or probes binding to amplified DNA products in real-time.
5. **Computer and Software:** PCR instrument software analyzes collected fluorescence data from the optical detection system. It calculates cycle threshold (Ct) values, representing the cycle number when fluorescence crosses a predefined threshold, indicating exponential target DNA amplification. The software also quantifies the target DNA amount based on fluorescence signal intensity.

6. **Reagents:** Essential reagents for real-time PCR include primers, dNTPs, Taq polymerase, and buffer solutions. These components are crucial for the success of the PCR reaction and accurate detection of nucleic acids.

PROCEDURE:

1. **DNA Extraction:** The initial step involves extracting DNA from the sample using established techniques.
2. **Primer Design:** Specific primers are meticulously designed to anneal exclusively to the target DNA sequence of interest, ensuring selective amplification.
3. **Amplification:**
   a. **Denaturation:** Double-stranded DNA is melted into single strands at a high temperature, typically 95°C, to disrupt secondary structures in single-stranded DNA. Denaturation time can be adjusted based on template GC content.
   b. **Annealing:** Complementary sequences have an opportunity to hybridize during annealing. An appropriate temperature, determined by the calculated melting temperature (Tm) of the primers (5°C below the Tm of the primer), is utilized.
   c. **Extension:** Optimal DNA polymerase activity occurs at 70-72°C, facilitating primer extension at rates up to 100 bases per second. In real-time PCR, when the amplicon is small, extension may be combined with annealing at 60°C.
4. **Real-time Monitoring:** Detection relies on fluorescence technology. The sample is placed in a well and subjected to a thermal cycle similar to traditional PCR. In real-time PCR, a tungsten or halogen source induces fluorescence in the marker added to the sample. The signal is amplified with the increasing copy number of sample DNA. The emitted signal is detected by a detector, converted into a digital signal, and displayed on a computer screen. Detection occurs as the signal surpasses the threshold level.
5. **Data Analysis:** The data generated during the real-time PCR reaction is analyzed using software. This analysis determines the quantity of DNA present in the original sample.

CONCLUSION:

By exploring the real-time PCR machine, we get the proper detection and quantification of nucleic acid sequence in real-time.

INDEX

12	**Visualization of 3D Protein Structure using RasMol and Chimera tool**	122	12/02/2024	
12(A)	To visualize 3D structure of protein ‘Insulin’ (PDB ID: 1GZ8) using RasMol tool.	136	12/02/2024	
12(B)	To visualization of 3D structure of protein ‘Insulin’ (PDB ID: 1GZ8) using Chimera.	142	12/02/2024	
13	**Introduction to Binding Pocket Predictions**	151	12/02/2024	
13(A)	To predict binding pockets located on protein surface for query ‘Insulin’(PDB ID: 1H59) using CASTp Tool.	154	12/02/2024	
13(B)	To predict and study the N-glycosylation sites present in the protein for query ‘Insulin’ (UniProt ID: P06213) using NetNGlyc – 1.0 tool.	159	12/02/2024	
13(C)	To predict and study the phosphorylation sites present in the protein ‘Insulin’ (UniProt ID: P06213) using Netphos3.1 program.	166	09/02/2024	
14	**Introduction to Structural BLAST**	173	13/02/2024	
14(A)	To identify similar structure for query ‘trypsin’ (PDB ID: 1P2J) using VAST+ tool.	179	13/02/2024	
14(B)	To perform structure comparison for query ‘trypsin’ (PDB ID: 1P2J) by using Dali server.	188	13/02/2024	

DATE: 15/01/2024

WEBLEM 9
INTRODUCTION TO SECONDARY STRUCTURE PREDICTIONS AND ITS VARIOUS TOOLS

INTRODUCTION:

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.

The secondary structure prediction methods can be either ab initio based, which make use of single sequence information only, or homology based, which make use of multiple sequence alignment information. The ab initio methods, which belong to early generation methods, predict secondary structures based on statistical calculations of the residues of a single query sequence. The homology-based methods do not rely on statistics of residues of a single sequence, but on common secondary structural patterns conserved among multiple homologous sequences.

The secondary structure prediction methods can be either:

1. Ab-initio based
2. Homology based
3. Neural networks based

1. Ab-initio method:

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. Examples of ab initio prediction are the Chou–Fasman and Garnier, Osguthorpe, Robson (GOR) methods. The ab initio methods were developed in the 1970s when protein structural data were very limited. The statistics derived from the limited data sets can therefore be rather inaccurate. However, the methods are simple enough that they are often used to illustrate the basics of secondary structure prediction.

2. Homology based method:

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 close protein homologs should adopt the same secondary and tertiary structure. evolutionary conservation dictates that there should be no major variations for their secondary structure elements. Therefore, 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.

3. **Neural networks-based methods:**
   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 inter connected variables or nodes. In secondary structure prediction, the input is an amino acid sequence and the output is the probability of a residue to adopt a particular structure. 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.

There are three generations of secondary structure prediction methods. Each generation has overall accuracy of about 10% higher than the earlier generation.

Generation	**1st**	**2nd**	**3rd**
Method	Chou-Fasman	GOR IV	PHD
Algorithm	Ab-initio	Ab-initio	Neural Network
		Neural Network	Nearest Neighbor
		Nearest Neighbor	Homology based

1st generation: Chou-Fasman:

The Chou–Fasman algorithm determines the propensity or intrinsic tendency of each residue to be in the helix, strand, and β-turn conformation using observed frequencies found in protein crystal structures (conformational values for coils are not considered). Prediction with the Chou–Fasman method works by scanning through a sequence with a certain window size to find regions with a stretch of contiguous residues each having a favored SSE score to make a prediction. The Chou–Fasman which is the first-generation methods developed in the 1970s, suffer from the fact that the prediction rules are somewhat arbitrary. They are based on single sequence statistics without clear relation to known protein-folding theories. The predictions solely rely on local sequence information and fail to take into account long range interactions. A Chou-Fasman–based prediction does not even consider the short-range environmental information. These reasons, combined with unreliable statistics derived from a very small structural database, limit the prediction accuracy of these methods to about 50%.

Fig 1: Homepage of Chou-Fasman tool

2nd generation: GOR IV:

The three alphabets GOR were derived from the first letter of their names (Garnier-Osguthorpe-Robson. This is the second-generation prediction algorithm developed in the 1980s and early 1990s. In used version of GOR method, database of 267 proteins is used which contains 63,000 residues It is based on information theory and Bayesian statistics. Information theory approaches are popular in secondary structure prediction and these approaches are mathematical probability based. GOR method works on window of 17 residues, eight nearest neighboring residues are included in calculations for a given residue. The conformational state among three states will be predicted and depends upon the type of amino acid R as well as neighboring residue along window. Information theory helps to retrieve the information function. GOR method calculates information from residue within sliding window. They have improved accuracy over the first generation by about 10%. The GOR method has improved from about 55% up to 64.4%.

Fig 2: Homepage of GOR IV Tool

3rd generation: PHD:

PHD is a web-based program that combines neural network with multiple sequence alignment. It first performs a BLASTP of the query sequence against a non-redundant protein sequence database to find a set of homologous sequences, which are aligned with the MAXHOM program (a weighted dynamic programming algorithm performing global alignment). The resulting alignment in the form of a profile is fed into a neural network that contains three hidden layers. The first hidden layer makes raw prediction based on the multiple sequence alignment by sliding a window of thirteen positions. The second layer refines the raw prediction by sliding a window of seventeen positions, which takes into account more flanking positions. This step makes adjustments and corrections of unfeasible predictions from the previous step. The third hidden layer is called the jury network, and contains network strained in various ways. It makes final filtering by deleting extremely short helices (one or two residues long) and converting them in to coils. After the correction, the highest scored state defines the conformational state of the residue. When multiple sequence alignments and neural networks are combined, the result is further improved accuracy. This combined approach has been shown to improve the accuracy to above 75%, which is a breakthrough in secondary structure prediction.

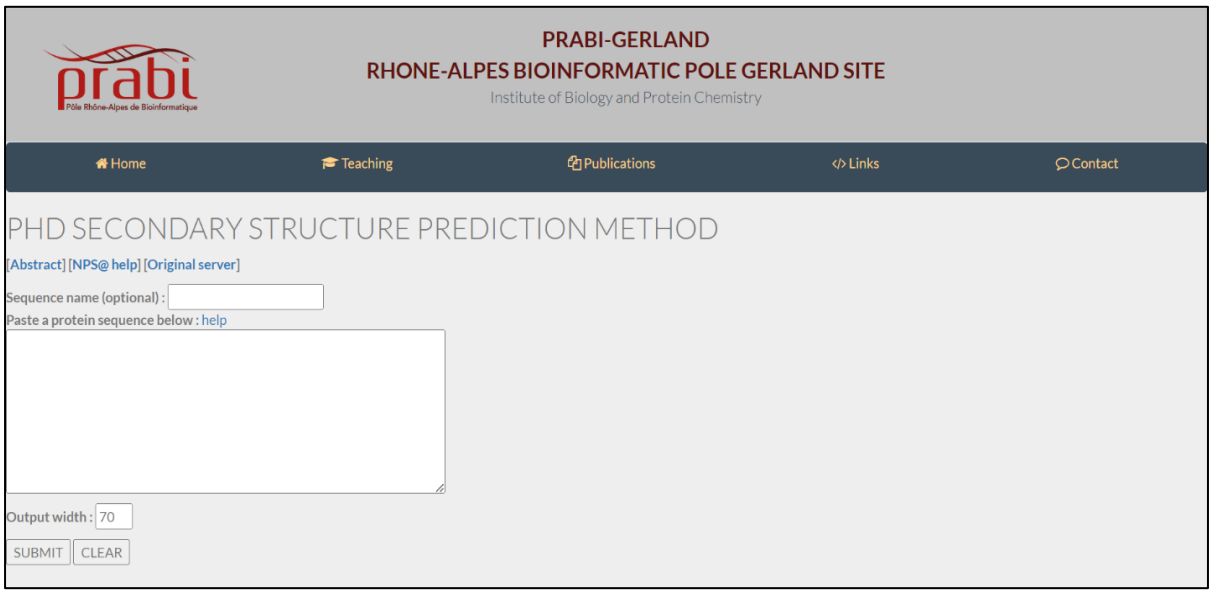


Fig 3: Homepage of PHD Tool

REFERENCES:

1. Xiong, J. (2006). Essential Bioinformatics. Cambridge: Cambridge University Press.
2. Singh, R., Jain, N., & Kaur, D. P. (2013). GOR Method for Protein Structure Prediction using Cluster Analysis. International Journal of Computer Applications, 73(1), 1–6. https://doi.org/10.5120/12702-9495
3. Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and computational tools for Next-Generation sequencing analysis in clinical genetics. Journal of Clinical Medicine, 9(1), 132. https://doi.org/10.3390/jcm9010132
4. Shiragannavar, S. (n.d.). Methods of prediction of secondary structures of proteins. https://biotecharticles.com/Bioinformatics-Article/Methods-of-Prediction-of-Secondary-Structures-of-Proteins-3759.html

DATE: 15/01/2024

WEBLEM 9(A)
CHOU AND FASMAN
(URL: https://www.biogem.org/tool/chou-fasman/)

AIM:

To predict secondary structure for query 'Biotin' (Accession No: P06709) using Chou and Fasman method.

INTRODUCTION:

The first protein structure prediction algorithm was developed by Chou and Fasman in 1974. The Chou-Fasman method is an early but important discovery for predicting the secondary structure of proteins. The basic idea behind the Chou-Fasman method is that the secondary structure of a protein is determined by the sequence of its amino acids.

The Chou-Fasman algorithm is based on the observation that certain amino acids have a higher propensity to form specific secondary structures. By assigning a set of parameters called the "propensities" to each amino acid, the algorithm calculates a score for each position in a protein sequence. This score reflects the likelihood of that position adopting a particular secondary structure.

The propensities in the Chou-Fasman algorithm were determined through an analysis of the frequencies of different amino acids in experimentally determined protein structures. The algorithm relies on the assumption that these frequencies correlate with the ability of each amino acid to stabilize a particular secondary structure.

The Chou-Fasman method is a relatively simple and computationally efficient method. It is typically around 50-60% accurate. Despite its limitations, the Chou-Fasman method remains an important historical milestone in the field of protein structure prediction. It helped to lay the foundation for more advanced methods that are used today.

Biotin:

Biotin, also known as vitamin B7 or vitamin H, is a water-soluble vitamin that plays a crucial role in various metabolic processes in the body. It is essential for metabolizing fats, carbohydrates, and proteins, and it contributes to the maintenance of a healthy nervous system, nails, hair, and skin. Biotin is a coenzyme that supports carboxylase enzymes involved in synthesizing fatty acids, amino acids like isoleucine and valine, and generating glucose through gluconeogenesis.

The secondary structure of biotin involves a sulfur-containing tetrahydrothiophene ring fused to a ureido group, with a C5-carboxylic acid side chain appended to the former ring. This structure classifies biotin as a heterocyclic compound. The ureido ring in biotin, containing the −N−CO−N− group, serves as the carbon dioxide carrier in carboxylation reactions.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Biotin'.
2. Open the protein of interest (amino acid sequence less than 1000 aa) and copy its FASTA sequence.
3. Open Chou and Fasman server and paste the copied FASTA sequence of 'Biotin' in the 'Enter the protein Sequence' box and then click on 'Predict'.
4. Interpret the results displayed for Tubulin on Chou and Fasman server.

OBSERVATIONS:

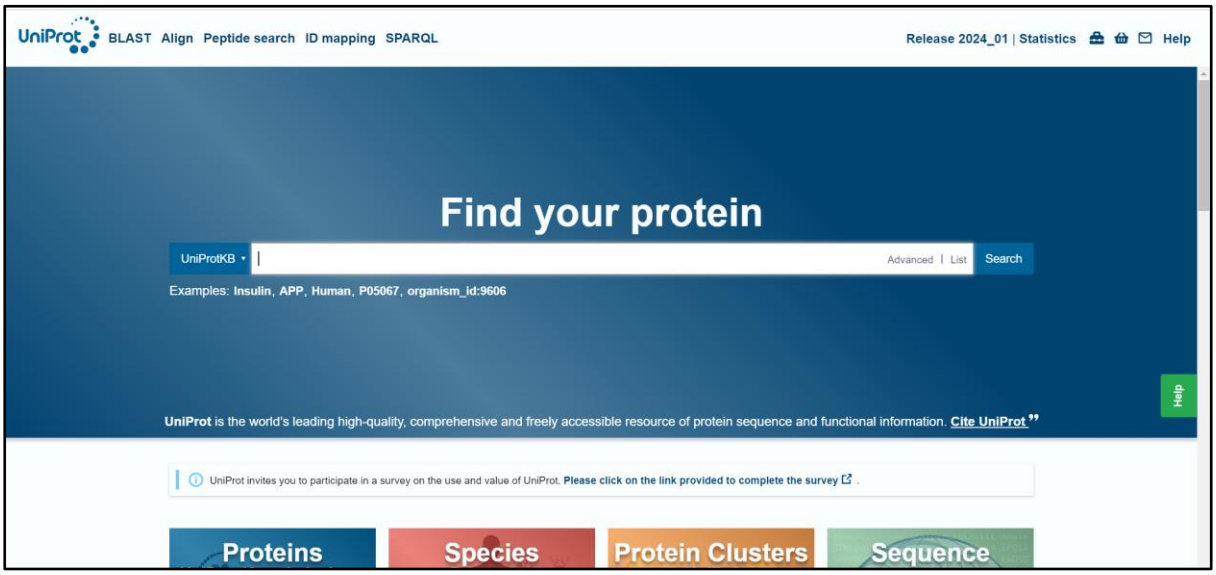


Fig 1: Homepage of UniProt Database

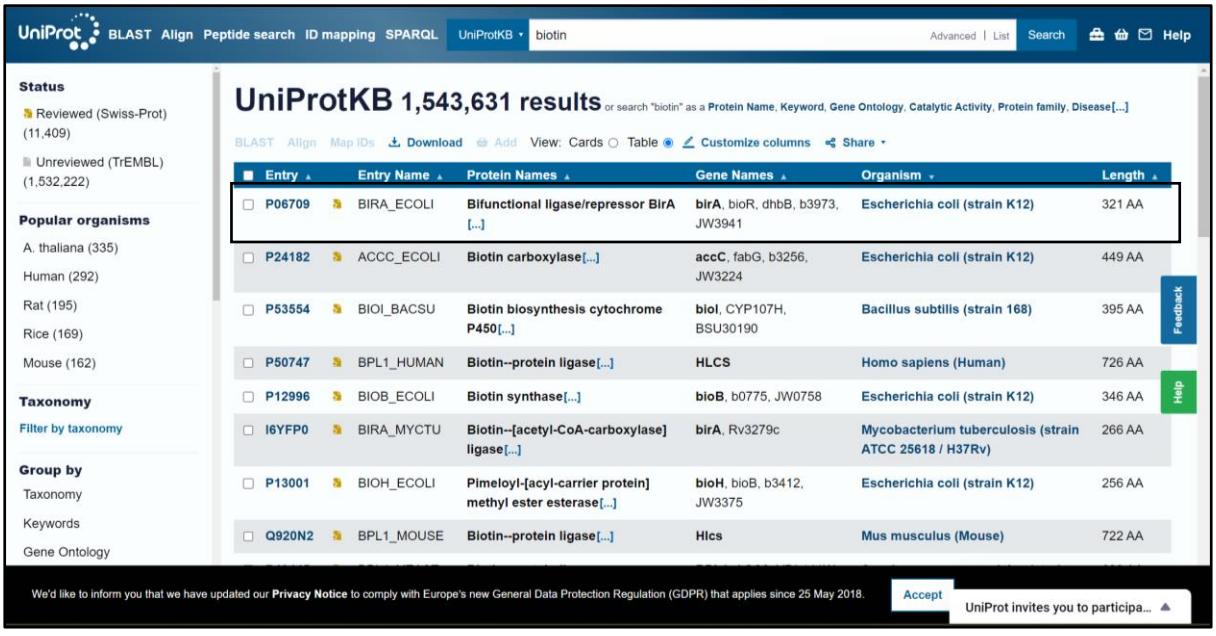


Fig 2: UniProt Database search for the query 'Biotin' (Accession No.: P06709)

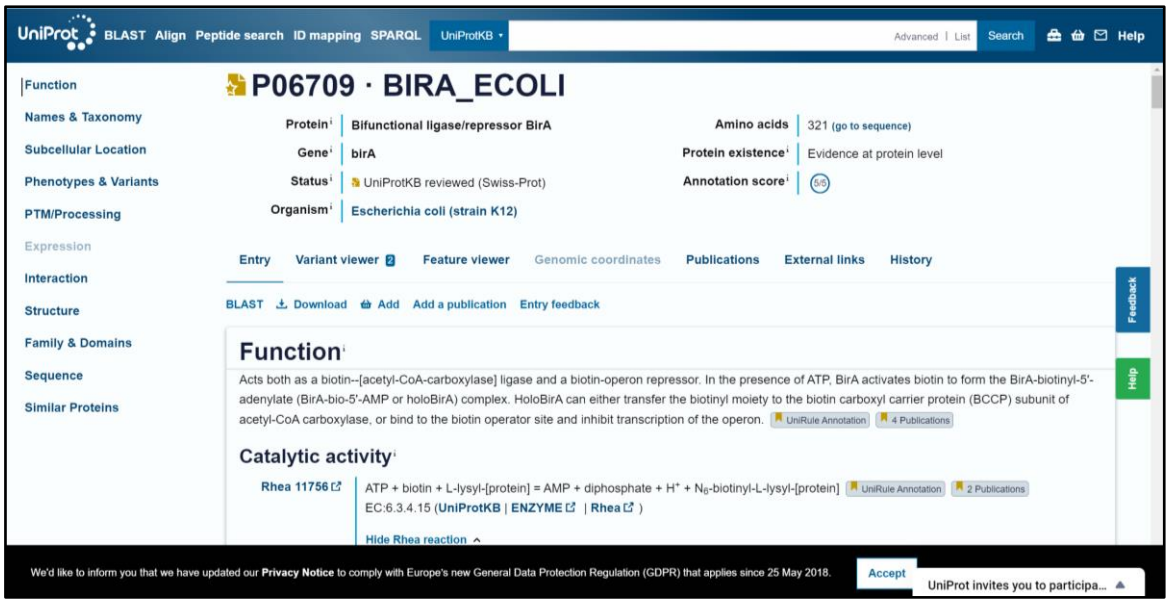


Fig 3: Information for the query 'Biotin' (Accession No.: P06709)

```
>sp|P06709|BIRA_ECOLI Bifunctional ligase/repressor BirA OS=Escherichia coli (strain K12) OX=83333 GN=birA PE=1 SV=1
MKDNTVPLKLIALLANGEFHSGEQLGETLGMSRAAINKHIQTLRDWGVDVFTVPGKGYSL
PEPIQLLNAKQILGQLDGGSVAVLPVIDSTNQYLLDRIGELKSGDACIAEYQQAGRGRRG
RKWFSPFGANLYLSMFWRLEQGPAAAIGLSLVIGIVMAEVLRKLGADKVRVKWPNDLYLQ
DRKLAGILVELTGKTGDAAQIVIGAGINMAMRRVEESVVNQGWITLQEAGINLDRNTLAA
MLIRELRAALELFEQEGLAPYLSRWEKLDNFINRPVKLIIGDKEIFGISRGIDKQGALLL
EQDGIIKPWMGGEISLRSAEK
```

Fig 4: FASTA sequence for the query 'Biotin' (Accession No.: P06709)

Fig 5: Homepage of CFSSP: Chou and Fasman Secondary Structure Prediction Server with the pasted FASTA sequence for the query 'Biotin' (Accession No.: P06709)

Fig 6.1: Results – Target sequence information and secondary structure details along with graphical representation

Fig 6.2: Results – Secondary Structure Details

```
Total Residues: H: 262   E: 137   T: 42
       Percent: H: 81.6  E: 42.7  T: 13.1
```

Fig 6.3: Statistics of secondary structure predicted elements

RESULTS:

The tabular data given below describes the statistics of secondary structure predicted elements which are as follows:

	Helix (H)	Coils (E)	Sheets (T)
Total Residues	262	137	42
Percent (%)	81.6	42.7	13.1

The predicted secondary structure for the sequence has different colors like blue for turns, red for helix, green for sheets and yellow for coils. The graph indicated the amount of helix, sheets and coils.

CONCLUSION:

In summary, the practical application of the Chou-Fasman algorithm for predicting the query 'Biotin' (Accession No.: P06709) revealed insights into its potential secondary structure elements. We could identify regions likely to form helices, sheets, or coils based on the propensity of specific amino acids to participate in such structures.

REFERENCES:

1. M. (2023, July 19). The Science of Protein Folding: Exploring the Basics. Medium. https://blogsbymoleculeai.medium.com/the-science-of-protein-folding-exploring-the-basics-d6cc5886700e
2. Zempleni, J., Wijeratne, S. S., & Hassan, Y. I. (2009). Biotin. BioFactors (Oxford, England), 35(1), 36–46. https://doi.org/10.1002/biof.8

DATE: 09/02/2024

WEBLEM 9(B)
GOR IV
(URL: https://npsa-prabi.ibcp.fr/NPSA/npsa_gor4.html)

AIM:

To predict the secondary structure for query 'Fibrinogen' (UniProt ID: P02675) by using GOR IV secondary structure prediction method.

INTRODUCTION:

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. 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. The secondary structure prediction methods can be either:

1. Ab-initio based
2. Homology based
3. Neural networks based

The secondary structure prediction GOR method is one of the first major methods proposed for prediction of structure from sequence. The three alphabets GOR were derived from the first letter of their names (Garnier-Osguthorpe-Robson). In used version of GOR method, database of 267 proteins is used which contains 63,000 residues. In prediction method for secondary structure of protein determines the accuracy in terms of present percentage of helix, sheet and coil. Formation of α-helix, β- sheet and coils are predicted with respect to each amino acid residue present in a sequence of amino acids residues. Result of the prediction of all secondary structure elements are combined to obtain the result of prediction of secondary structure of protein. Rather than considering propensities for a single residue, position-dependent propensities have been calculated for all residue types. GOR method work on various types of sequences formats which uses the information theory to generate the code that relates amino acids sequence and secondary structure of proteins. Three scoring matrices are prepared in GOR method to calculate the probability of each amino acids present in every positions. One matrix corresponds to the central amino acid being found in α helix, the second for the amino acid being in a β strand, the third a coil.

Through the successive incorporation of observed frequencies of single, then pairs of residues on a local sequence of 17 residues, the accuracy of the GOR method has improved from about 55% up to 64.4%. The GOR method has the advantage over neural network-based methods or nearest-neighbor methods in that it clearly identifies what is taken into account for the prediction and what is neglected. The method provides estimates of probabilities for the three secondary structures at each residue position that can be useful for further application of the method.

Fibrinogen:

Fibrinogen is a thrombin-coagulable glycoprotein occurring in the blood of vertebrates. The primary structure of the alpha, beta, and gamma polypeptide chains of human fibrinogen is known from amino acid and nucleic acid sequencing. The intact molecule has a trinocular, dimeric structure and is functionally bivalent. Thrombin cleaves short peptides from the amino termini of the alpha and beta chains exposing polymerization sites that are responsible for the formation of fibrin fibers and appearance of a clot. The major physiological function of fibrinogen is the formation of fibrin that binds together platelets and some plasma proteins in a hemostatic plug. In pathological situations, the network entraps large numbers of erythrocytes and leukocytes forming a thrombus that may occlude a blood vessel. Fibrinogen is a multifunctional protein. Fibrinogen is indispensable for platelet aggregation; it also binds to several plasma proteins.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Fibrinogen'.
2. From the results page, open the protein entry of interest. Here, FIBB_HUMAN (UniProt ID: P02675).
3. Copy the fibrinogen protein sequence.
4. Open the homepage of GOR IV secondary structure prediction method and paste the sequence in the query box and set the desired parameters. Click on 'SUBMIT' button to submit the query.
5. The results page of GOR IV displays the predicted secondary structure, percentage of helix, turns and coils. Interpret the results.

OBSERVATIONS:

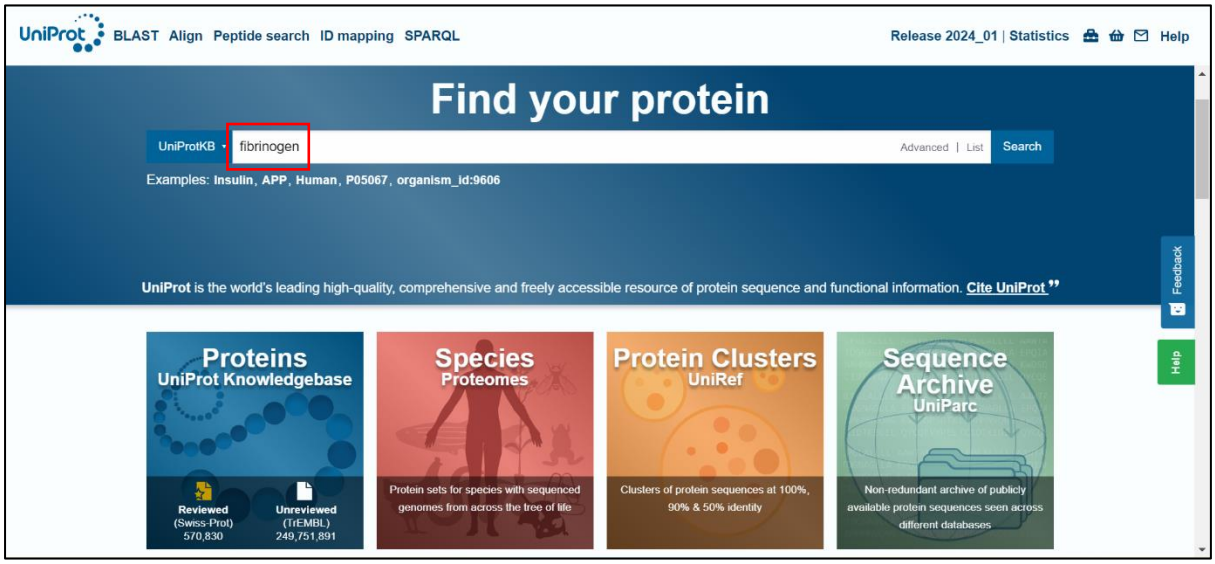


Fig 1: Search for query 'Fibrinogen' in UniProt database

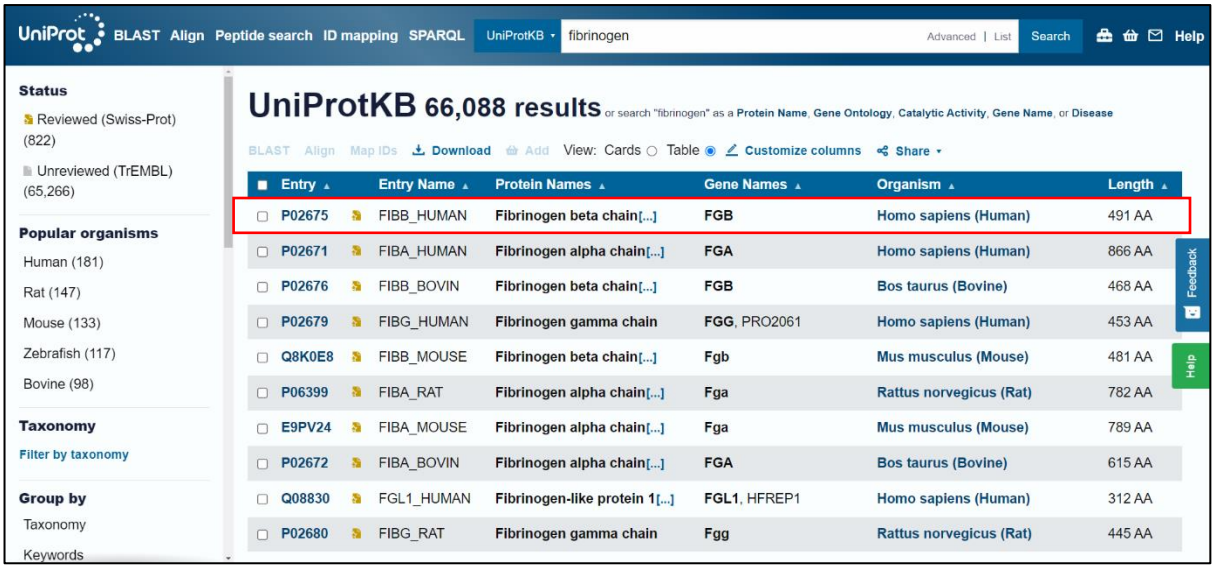


Fig 2: Open the protein entry of interest
Here, FIBB_HUMAN (UniProt ID: P02675)

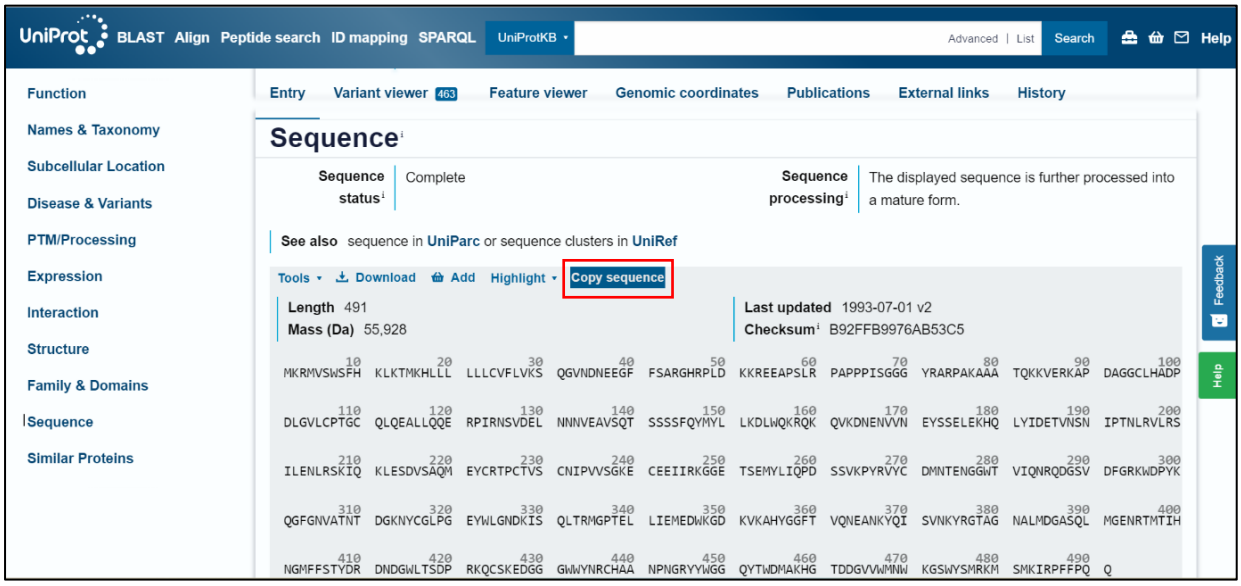


Fig 3: Copy the fibrinogen protein sequence

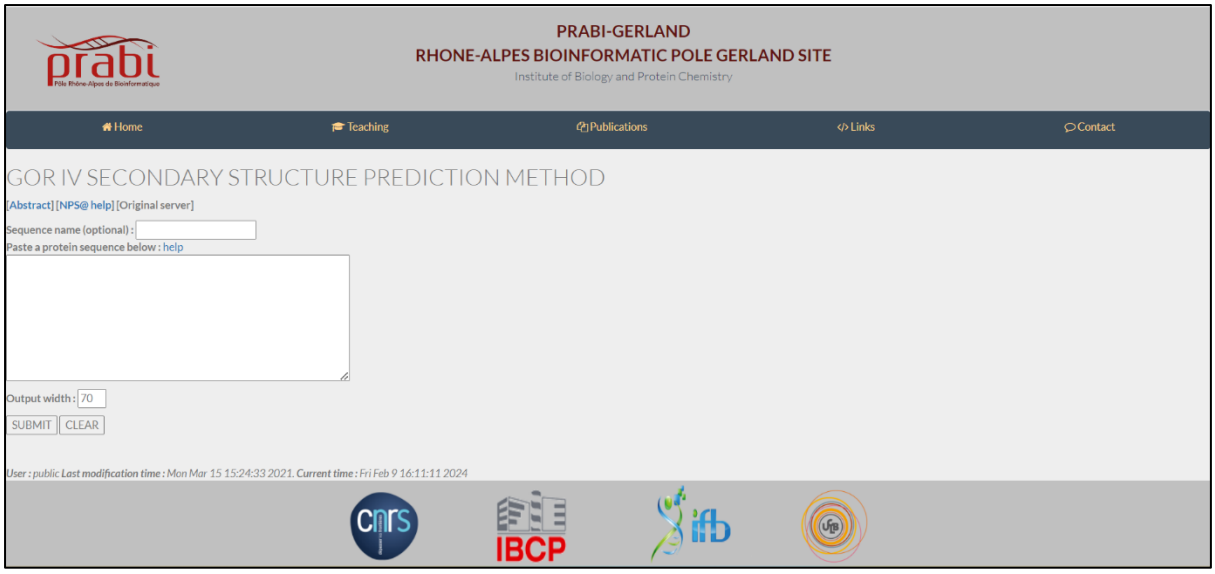


Fig 4: Homepage of GOR IV Tool

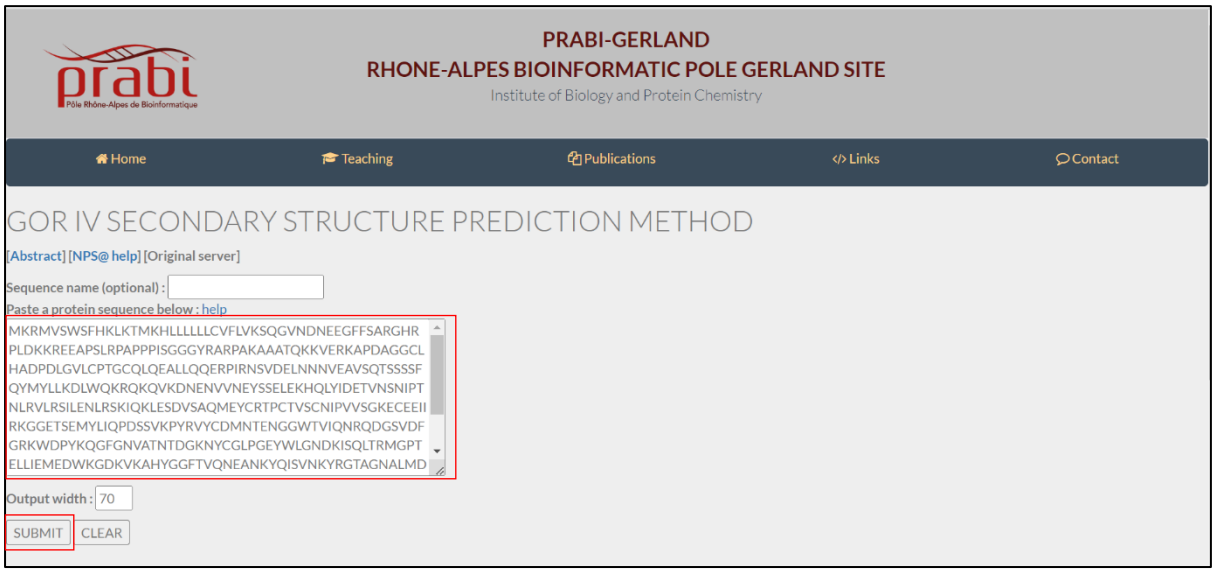


Fig 5: Paste the sequence in the query box and submit the query

Fig 6: Predicted secondary structure along with different colors for each of the query sequence

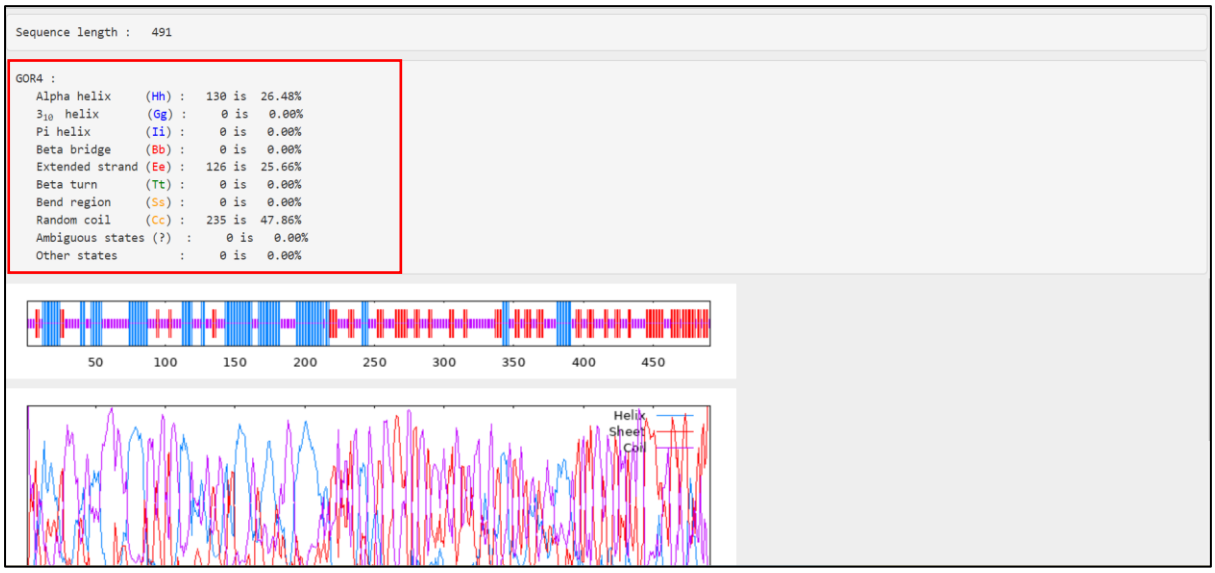


Fig 7: Percentage of Alpha helix, Beta turn and random coil

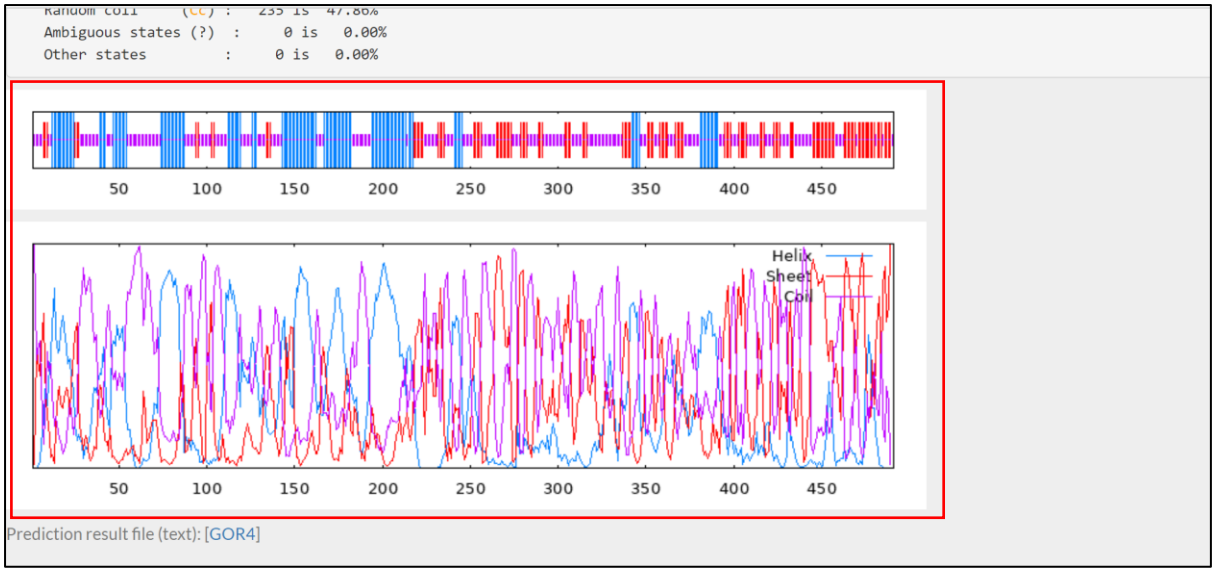


Fig 8: Graph represents the amount of helix, turn and coil

RESULTS:

The amount and percentage of predicted secondary structures for the query 'Fibrinogen' (UniProt ID: P02675) are:

Secondary structure	Amount	Percentage
Alpha helix	130	26.48%
Extended strand	126	25.66%
Random coil	235	47.86%

The predicted secondary structure for the sequence has different colors like blue for helix, red for sheets and yellow for coils. Random coil was seen more in the sequence. The graph indicated the amount of helix, sheet and coil.

CONCLUSION:

Secondary structure was predicted for query 'Fibrinogen' (UniProt ID: P02675) using first generation tool that is GOR IV. 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.

REFERENCES

1. Singh, R., Jain, N., & Kaur, D. P. (2013). GOR Method for Protein Structure Prediction using Cluster Analysis. International Journal of Computer Applications, 73(1), 1–6. https://doi.org/10.5120/12702-9495
2. Jean Garnier, Jean-François Gibrat, Barry Robson,[32] GOR method for predicting protein secondary structure from amino acid sequence, Methods in Enzymology, Academic Press, Volume 266,1996, Pages 540-553, ISSN 0076-6879, ISBN 9780121821678, https://doi.org/10.1016/S0076-6879(96)66034-0.
3. Budzynski A. Z. (1986). Fibrinogen and fibrin: biochemistry and pathophysiology. Critical reviews in oncology/hematology, 6(2), 97–146. https://doi.org/10.1016/s1040-8428(86)80019-1

DATE: 09/02/2024

WEBLEM 9(C)
PHD
(URL: https://npsa-prabi.ibcp.fr/NPSA/npsa_phd.html)

AIM:
To predict the secondary structure for query 'Rhodopsin' (UniProt ID: Q8WTQ7) by using PHD secondary structure prediction method.

INTRODUCTION:
The PhD method for protein secondary structure prediction is a second-generation approach introduced by Rost and Sander in the 1990s. It is based on the idea of using multiple sequence alignments and neural networks to predict the secondary structure of proteins. The PhD system uses a combination of multiple sequence alignments and neural networks to build a profile using a MSA and the profile techniques. It then employs a feed-forward neural network with three levels: a sequence-to-structure net, a structure-structure net, and a jury decision. The jury decision level takes the results from each of the nets and averages them, and the secondary structure with the highest average score is output as the prediction. The method has been shown to have a prediction accuracy of 73.5% when using multiple sequence alignments as input. The PhD method has significantly improved the accuracy of secondary structure predictions, making it a valuable tool in the field of computational biology.

The PhD method for protein secondary structure prediction offers several advantages, such as:

1. **High Prediction Accuracy:** The method has a high prediction accuracy, with a reported accuracy of 73.5% when using multiple sequence alignments as input.
2. **Utilization of Aligned Protein Families:** It uses an aligned family of proteins to predict the secondary structure, which has significantly improved its performance compared to earlier methods.
3. **Incorporation of Neural Networks:** The method employs a feed-forward neural network with three levels, which enhances its predictive capabilities.

On the other hand, the PhD method has some limitations, including:

1. **Dependence on Homologous Proteins:** Its high accuracy is limited to proteins without known homologs, as homology alignment predicts secondary structure better than the PhD method for proteins with known homologs.
2. **Definition of Secondary Structure:** The accuracy of secondary structure prediction depends on how secondary structure is defined, which can impact the method's performance.
3. **Prediction of Beta-Sheets:** While the PhD method improved the prediction of beta-sheets, there may still be limitations in accurately predicting this secondary structure element.

Rhodopsin:

Rhodopsin is a light-sensitive G protein-coupled receptor (GPCR) found in the retinal rod cells of the eye, playing a fundamental role in the visual system. Composed of opsin, a colorless protein, and 11-cis-retinal, a light-absorbing molecule, rhodopsin undergoes a conformational change upon light exposure, initiating the photo transduction cascade. This cascade leads to the generation of electrical signals, which are then transmitted to the brain, enabling vision. Rhodopsin's structure, particularly its seven transmembrane domains and the retinal chromophore, has been extensively studied, providing insights into its molecular mechanism of photoreception and signal transduction.

Due to its significance in vision and its representation of the GPCR superfamily, rhodopsin serves as a crucial model for understanding the function and dysfunction of GPCRs and their associated hereditary eye diseases.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Rhodopsin'.
2. From the results page, open the protein entry of interest. Here, GRK7_HUMAN (UniProt ID: Q8W2Q7).
3. Copy the Rhodopsin protein sequence.
4. Open the homepage of PHD secondary structure prediction method and paste the sequence in the query box and set the desired parameters. Click on 'SUBMIT' button to submit the query.
5. The results page of PHD displays the predicted secondary structure, percentage of helices, turns and coils. Interpret the results.

OBSERVATIONS:

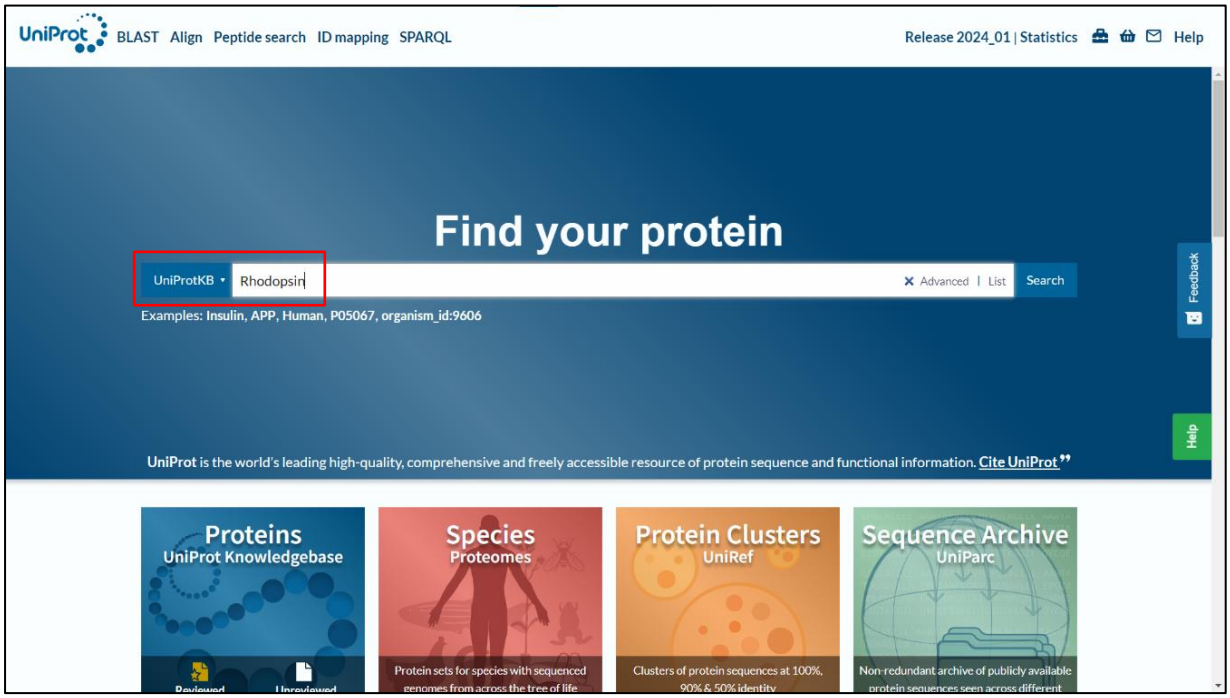


Fig 1: Search for query 'Rhodopsin' in UniProt Database

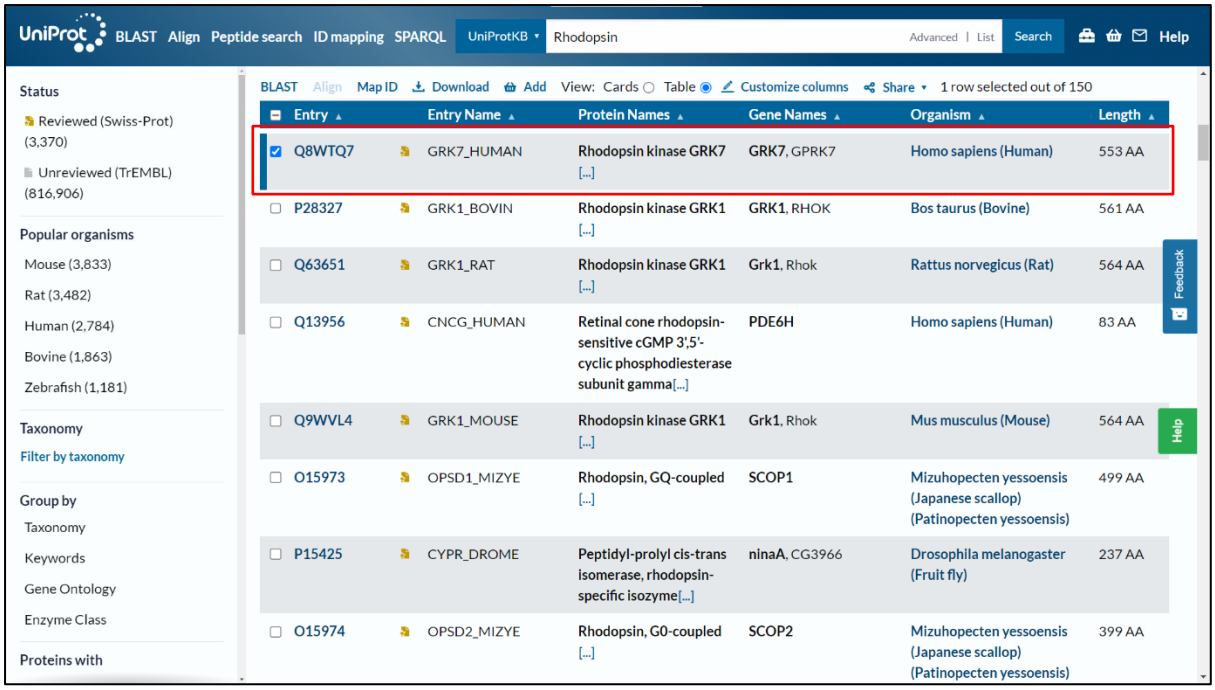


Fig 2: Open the protein entry of interest. Here, GRK7_HUMAN (UniProt ID: Q8WTQ7)

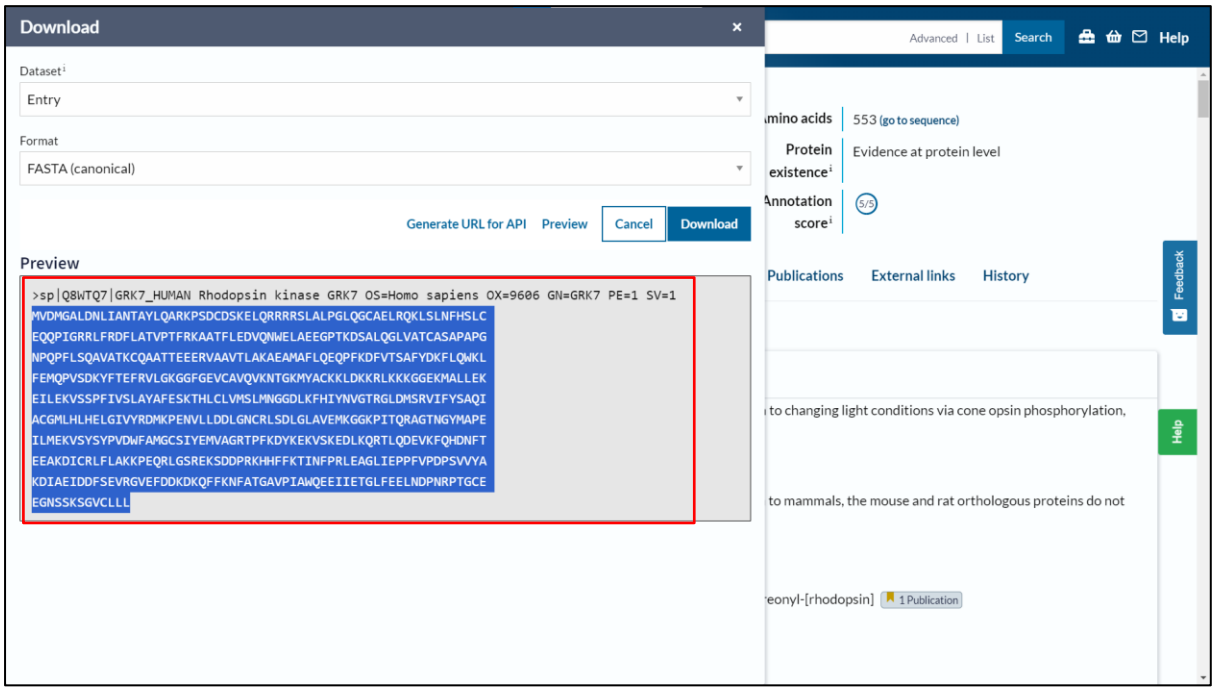


Fig 3: Copy the sequence of selected entry

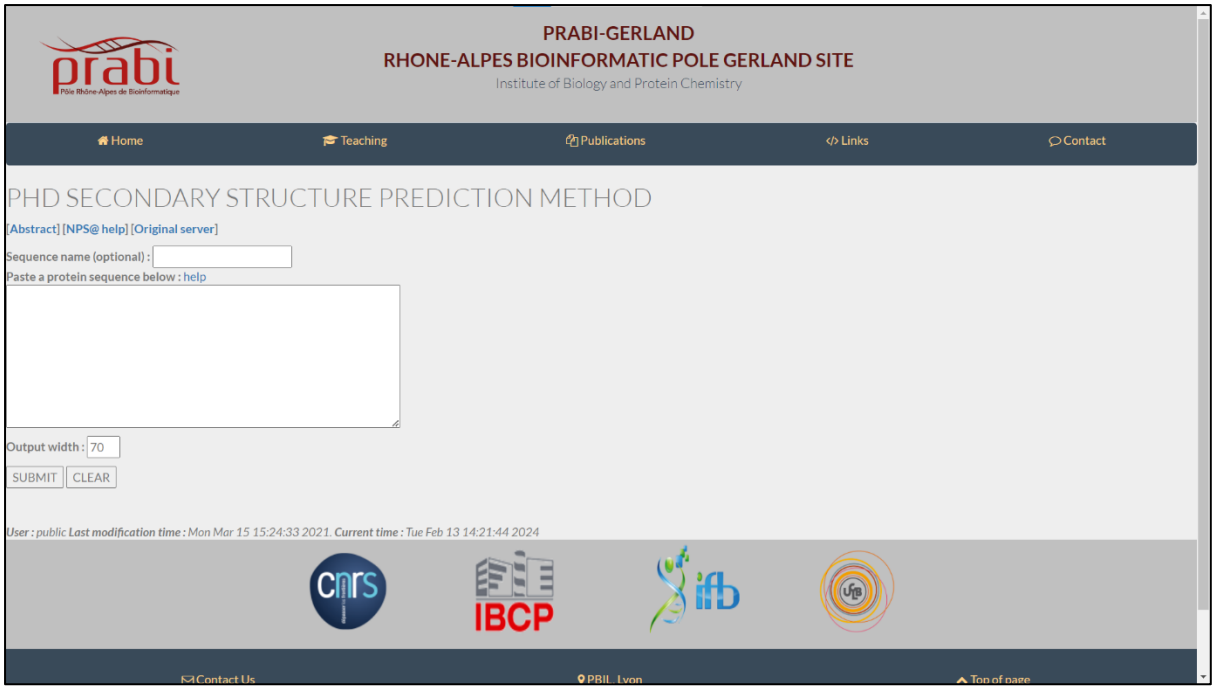


Fig 4: Homepage of PHD Tool

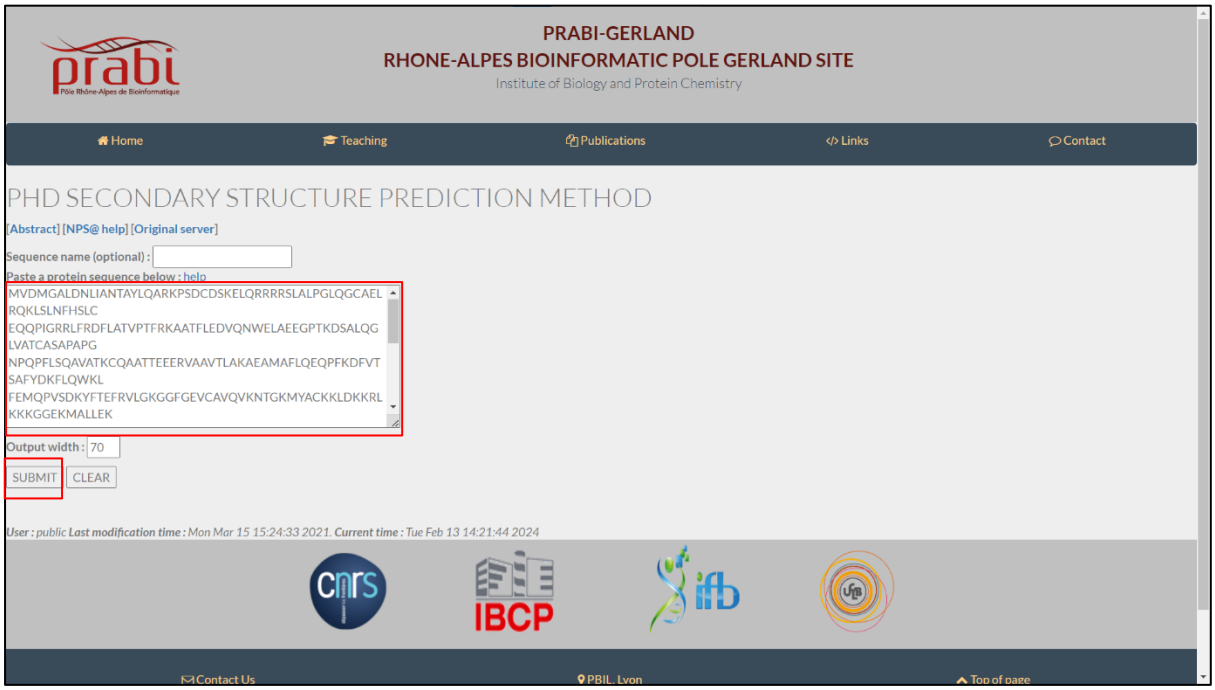


Fig 5: Paste the sequence in the query box and submit the query

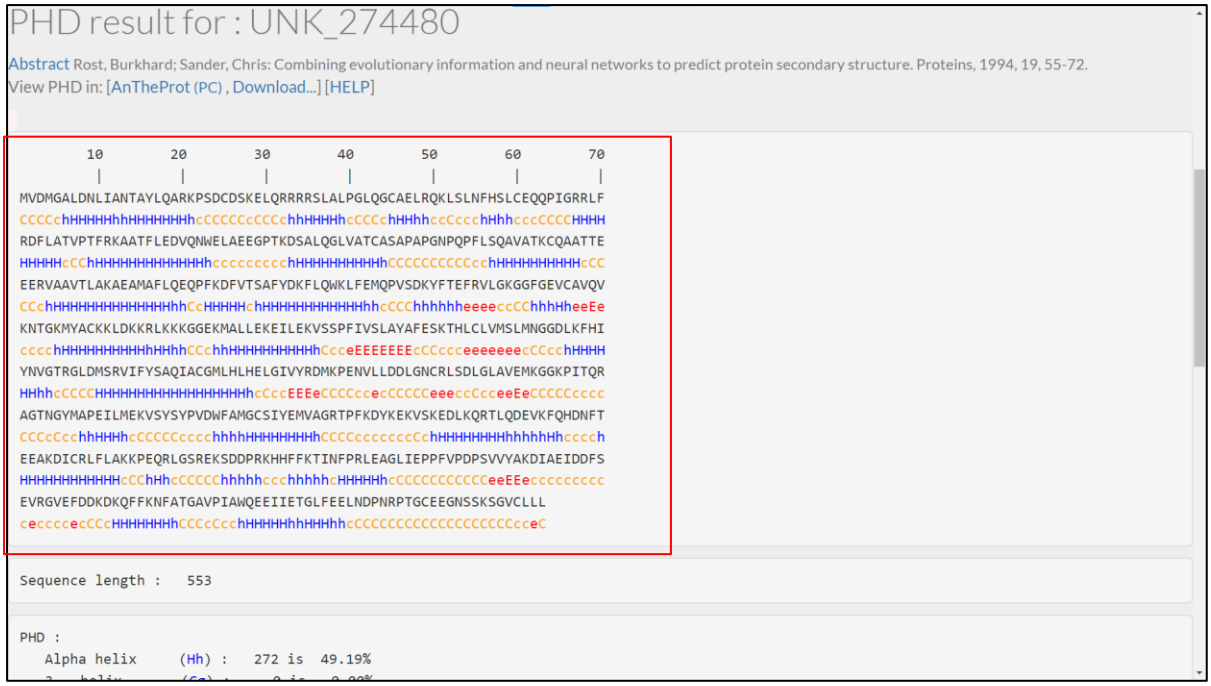


Fig 6: Predicted secondary structure Elements: Helices, Strands & Coils

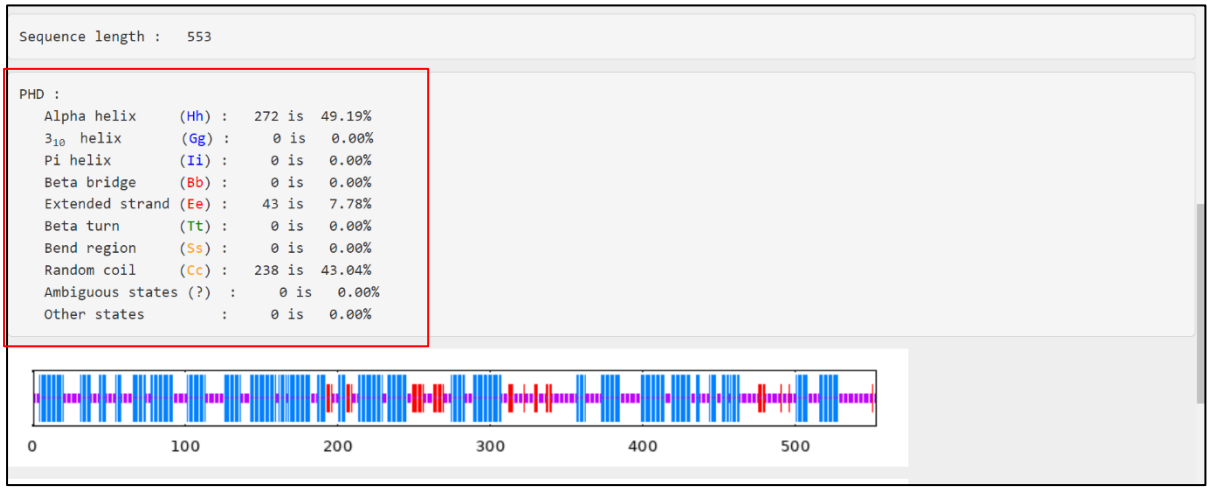


Fig 7: Percentage of Alpha Helix, Extended Strand and Random Coil in the Secondary Structure

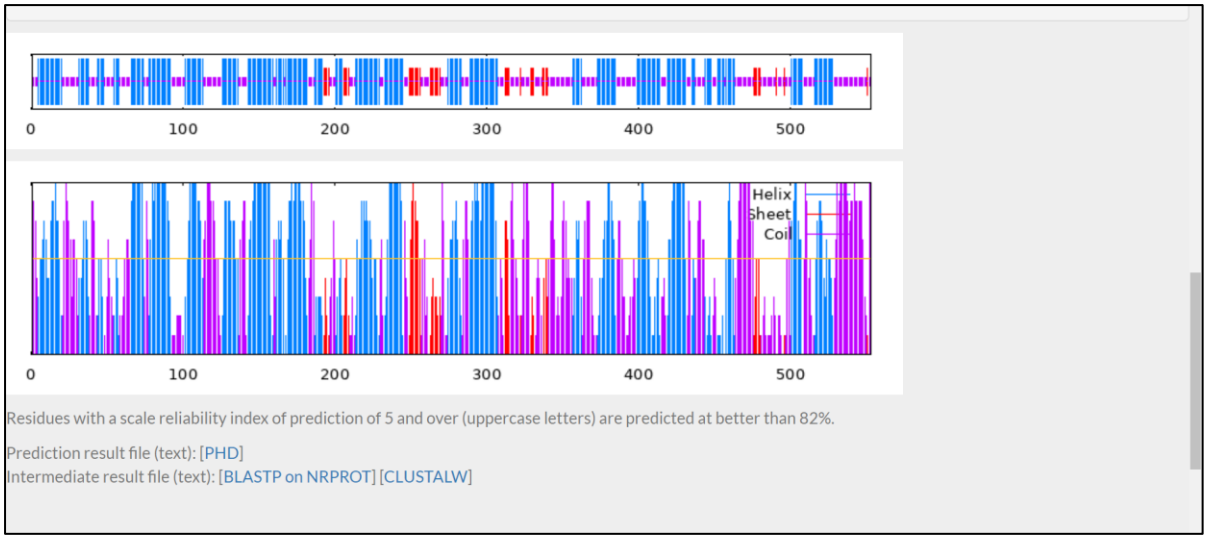


Fig 8: Graph represents the amount of Secondary Structure elements viz Helix, Sheet and Coil

RESULTS:

The amount and percentage of predicted secondary structures for the query 'Rhodopsin' (UniProt ID: Q8WTQ7) are:

Secondary structure	Amount	Percentage
Alpha helix	272	49.19%
Extended strand	43	7.78%
Random coil	238	43.04%

The color codes typically used to represent different secondary structure elements in protein visualization are as follows:

1. Helices are represented by Blue Color.
2. Strands are represented by Red Color.
3. Coils are represented by Orange Color.

These color representations aid in the visual identification and differentiation of the various secondary structure elements within a protein.

The graph resulting from the PhD secondary structure prediction method represented the predicted secondary structure elements, such as helices, turns, and coils, along the horizontal axis. The length of the lines in the graph indicated the confidence or probability of the predicted secondary structure at each position along the protein sequence. A longer line or a higher bar for a specific secondary structure element, such as a helix, indicates a higher confidence in the prediction of a helical structure at the corresponding position in the protein sequence.

CONCLUSION:

Secondary structure was predicted for query 'Rhodopsin' (UniProt ID: Q8WTQ7) using second generation tool that is PHD.

REFERENCES

1. Jin, X., Guo, L., Jiang, Q., Wu, N., & Yao, S. (2022). Prediction of protein secondary structure based on an improved channel attention and multiscale convolution module. Frontiers in Bioengineering and Biotechnology, 10. https://www.frontiersin.org/articles/10.3389/fbioe.2022.901018
2. Secondary structure predictions. (n.d.). Retrieved February 13, 2024, from https://archive.bioinfo.se/kurser/distanskurs_1998/text/secstr.html
3. S., G., & E.r., V. (2023). Protein secondary structure prediction using cascaded feature learning model. Applied Soft Computing, 140, 110242. https://doi.org/10.1016/j.asoc.2023.110242
4. NPS@ help: Help on PHD tool. (n.d.). Retrieved February 13, 2024, from https://npsa-prabi.ibcp.fr/NPSAHLP/npsahlp_secpredphd.html
5. Rhodopsin structure and function—Proteopedia, life in 3d. (n.d.). Retrieved February 13, 2024, from https://proteopedia.org/wiki/index.php/Rhodopsin_Structure_and_Function

DATE: 16/01/2024

WEBLEM 10
INTRODUCTION TO TERTIARY STRUCTURE PREDICTIONS

Protein tertiary structure often provides a basis for understanding its function. Experimental approaches for protein structure determination, such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) techniques are typically expensive and time-consuming. The increase of the structures in Protein Data Bank (PDB) cannot keep up with the increase of proteins characterized in high-throughput genome sequencing Compared to experimental approaches, computational methods, i.e., to predict the native structure of a protein from its amino acid sequence, are much cheaper and faster. As significant progress has been made over the past two decades, computational methods are becoming more and more importantfor studying protein structures in recent years. In contrast to sequencing techniques, experimental methods to determine protein structures are time consuming and limited in theirapproach. 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.

There are three computational approaches to protein three-dimensional structural modelling and prediction. They are homology modelling, threading, and ab initio prediction. The first two are knowledge-based methods; they predict protein structures based on knowledge of existing protein structural information in databases. Homology modelling builds an atomic model based on an experimentally determined structure that is closely related at the sequence level. Threading identifies proteins that are structurally similar, with or without detectable sequence similarities. The ab initio approach is simulation based and predicts structures based on physicochemical principles governing protein folding without the use of structural templates.

1. Homology modelling

Homology modelling is one of the computational structure prediction methods that are used to determine protein 3D structure from its amino acid sequence. It is considered to be the most accurate of the computational structure prediction methods. It consists of multiple steps that are straightforward and easy to apply. There are many tools and servers that are used for homology modelling. There is no single modelling program or server which is superior in every aspect to others. Since the functionality of the model depends on the quality of the generated protein 3D structure, maximizing the quality of homology modelling is crucial. Homology modelling has many applications in the drug discovery process. Since drugsinteract with receptors that consist mainly of proteins, protein 3D structure determination, and thus homology modelling is important in drug discovery. Accordingly, there has been the clarification of protein interactions using 3D structures of proteins that are built with homology modelling. This contributes to the identification of novel drug candidates. Homology modelling plays an important role in making drug discovery faster, easier, cheaper, and more practical. As new modelling methods and combinations are introduced, the scope of its applications widens.

There are usually four steps in homology-based protein structure prediction methods:

1. Identify one or more suitable structural templates from the known protein structure databases
2. Align the target sequence to the structural template
3. Build the backbone from the alignment, including the loop region and any region that issignificantly different from the template
4. Place the side-chains. The first two steps, identification of structural templates and alignment of the target sequence onto the parent structures, are usually related.

Modeller is used for homology or comparative modelling of protein three-dimensional structures. The user provides an alignment of a sequence to be l with known related structures and modeller automatically calculates a model containing all non-hydrogen atoms. Modeller models 3D structure of proteins. It is built in FORTRAN. Modeller is most frequently used for homology or comparative protein structure modelling. Modeller models protein 3D structure keeping in the constraints of spatial restraints. Swiss Model Server is a fullyautomated protein structure Homology Modelling server. Swiss Model however does not accept the sequences for homology modelling when similarity is less than 25%. ESyPred3Dis a new automated homology modelling program. The method gets benefit of the increased alignment performances of a new alignment strategy using neural networks. Alignments are obtained by combining, weighting and screening the results of several multiple alignment programs.

Sequence comparison methods determine sequence similarity by aligning the sequences optimally. The aligned residuals of the structure templates are used to construct the structural model in the second step. The quality of the sequence comparison thus not only determines whether a suitable structural template can be found but also the quality of the alignment between the target sequence and the parent structure, which in turn determines the accuracy of the structural model. Of critical importance is the ability for the sequence comparison to detect remote homologues and to correctly align the target sequence to and parent structure. In the following I discuss the various sequence comparison methods in relation to homology modelling and their range of applicability, accuracy and shortcomings.

One indication of the accuracy of comparative modelling is the sequence identity between the target and the template. It is believed that if two protein sequences have 50% or higher sequence identity, then the RMSD of the align able potion between the two structures will normally be less than 1. In the so-called “twilight zone”, with sequence identity between 20%~30%, 95% of the sequences with this level of identity have different structures though. When a structure template can indeed be found within the known protein structure databases in such cases, the backbone RMSD can be expected to be no better than 2. Structurally similar proteins can have low sequence identities in the 8~10% range and can still be identified with sensitive profile-profile based comparison, but the RMSD can be as large as 3~6. The error largely comes from the misalignment from sequence comparison. At such low sequence identity, comparison method that can detect the remote homology as well as align the sequences close to the optimal from structure alignment will be desirable.

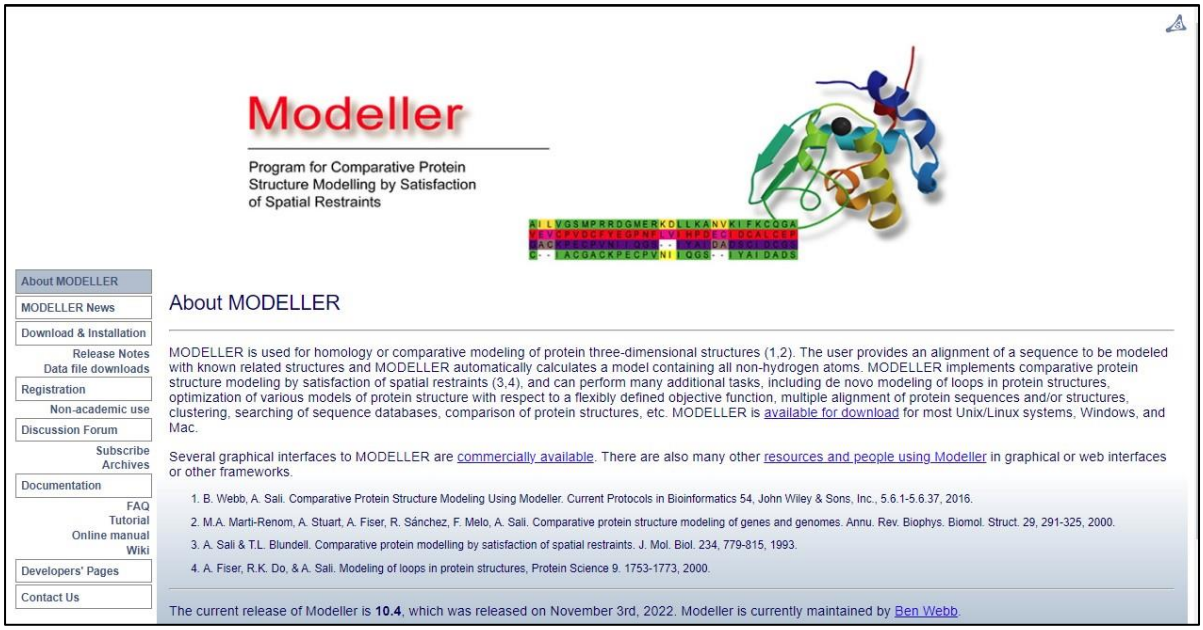


Fig 1: Homepage of Modeller tool

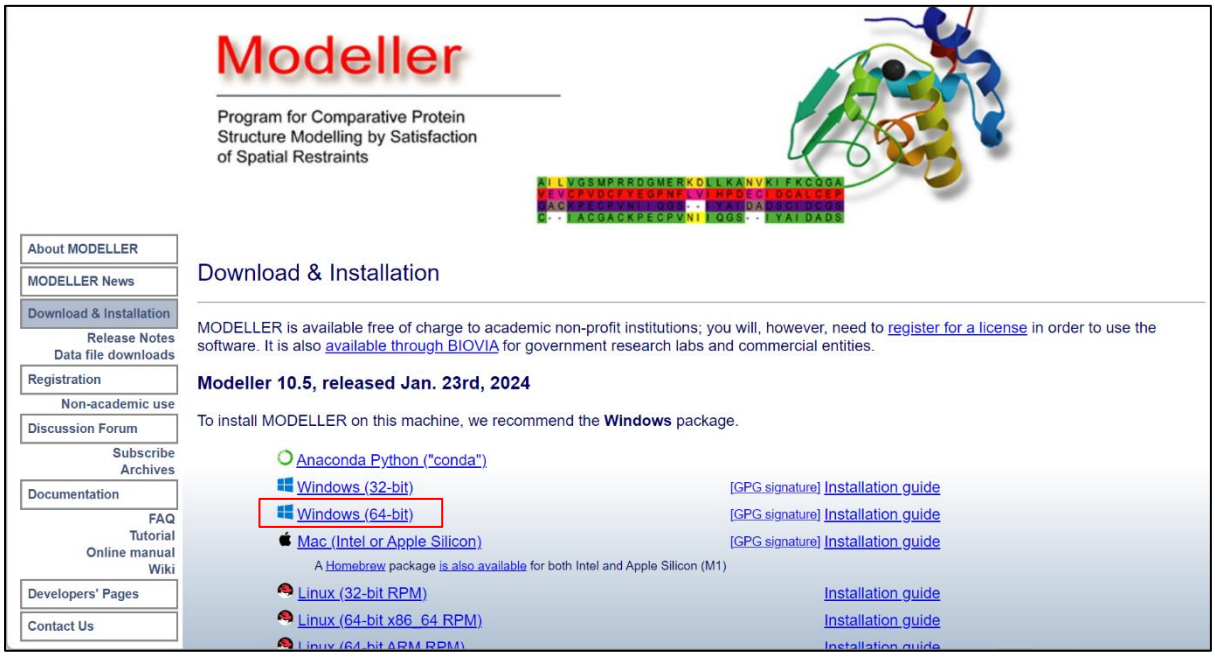


Fig 2: Select Windows 64 bits 10.4 ModellerVersion

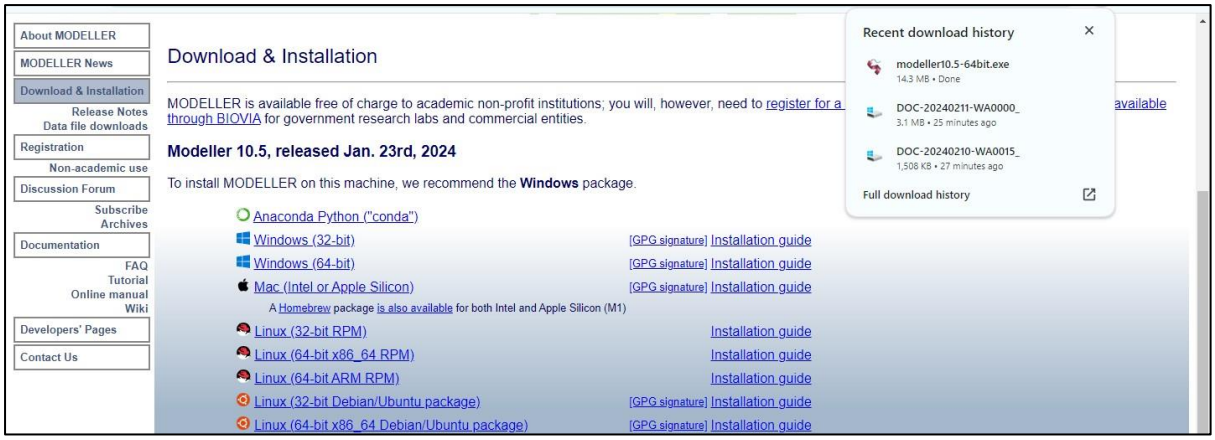


Fig 3: Install and open the folder

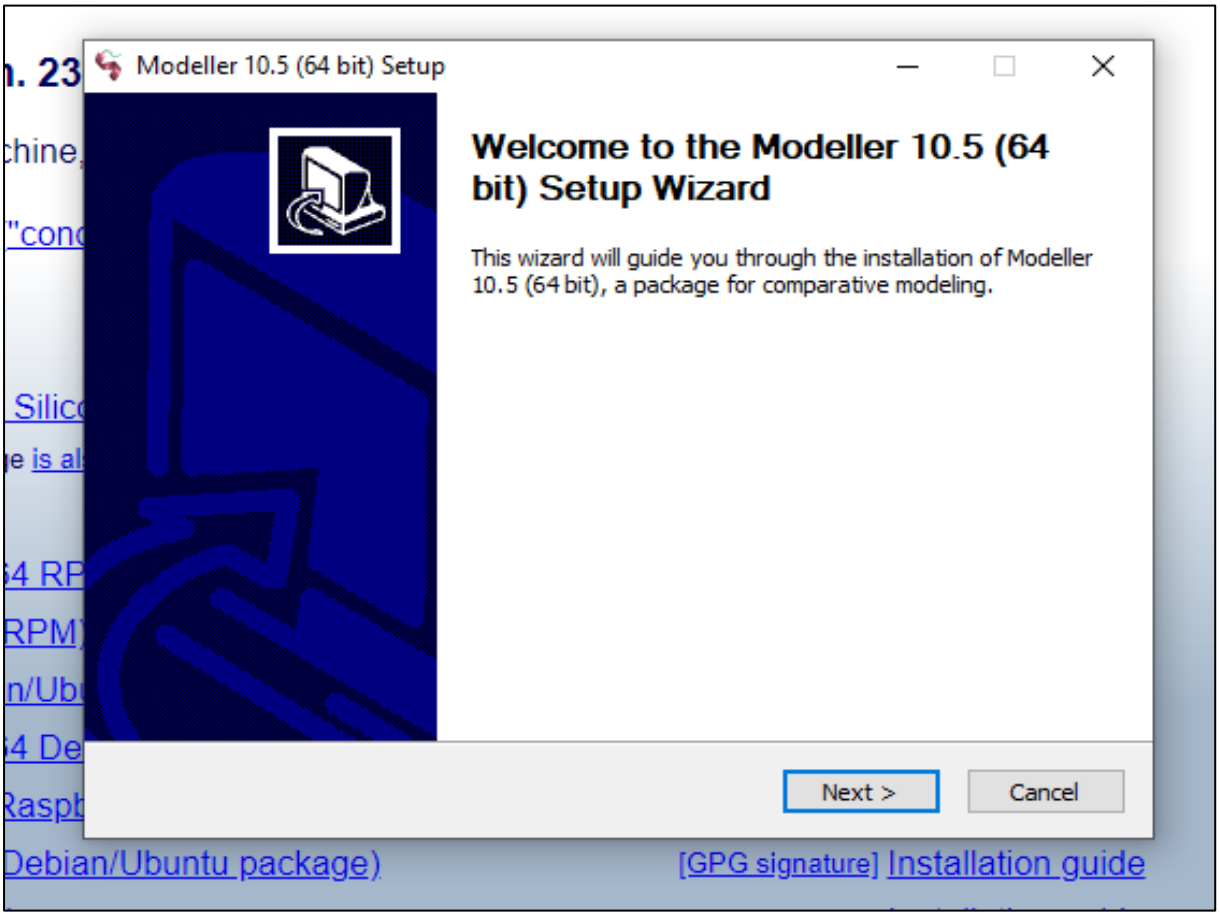


Fig 4: Click on 'Next'

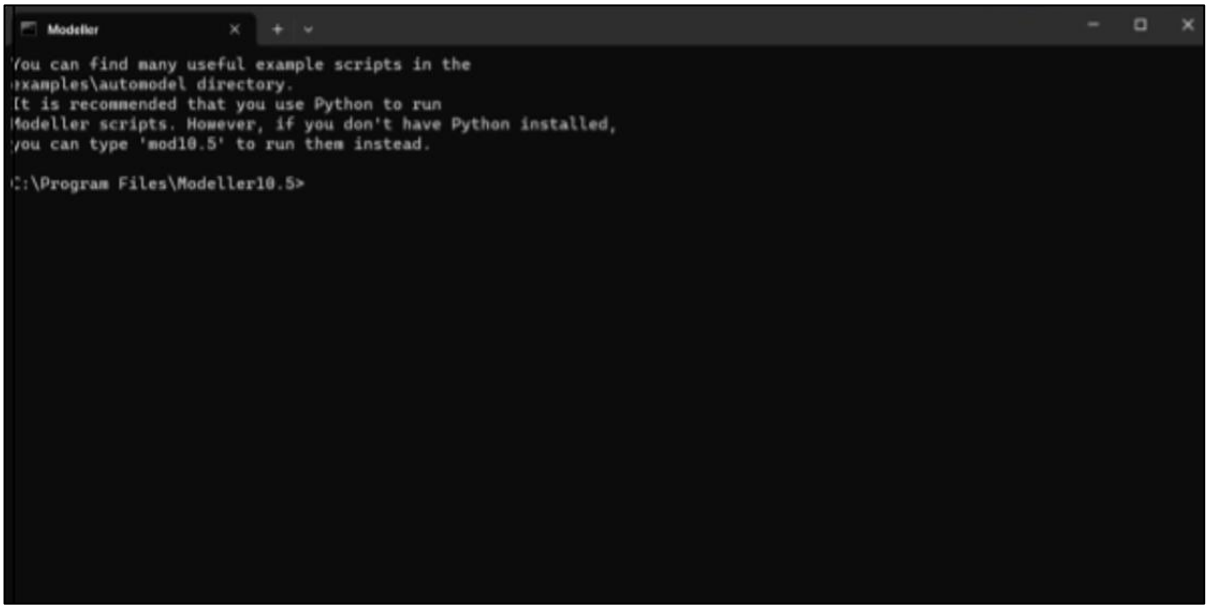


Fig 5: Modeller 10.4 Homepage

2. Threading Method

Protein threading or fold recognition refers to a class of computational methods for predicting the structure of a protein from amino acid sequence. The basic idea is that the target sequence (the protein sequence for which the structure is being predicted) is threaded through the backbone structures of a collection of template proteins (known as the fold library) and a "goodness of fit" score calculated for each sequence-structure alignment. This goodness of fit is often derived in terms of an empirical energy function, based on statistics derived from known protein structures, but many other scoring functions have been proposed and triedover the years. The most useful scoring functions include both pairwise terms (interactions between pairs of amino acids) and solvation terms. Threading methods share some of the characteristics of both comparative modelling methods (the sequence alignment aspect) and ab initio prediction methods (predicting structure based on identifying low-energy conformations of the target protein).

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure prediction and structure-based function annotation. It first identifies structural templates from the PDB by multiple threading approach LOMETS, with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through protein function database BioLiP. I-TASSER (as 'Zhang-Server' or 'UM-TBM') was ranked as the No 1 server for protein structure prediction in recent communitywide CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, CASP13, CASP14,and CASP1 5 experiments. It was also ranked the best for function prediction in CASP9. The server is in active development with the goal to provide the most accurate protein structure and function predictions using state-of-the-art algorithms. The server is only for non-commercial use. Please report problems and questions at I-TASSER message board and our developers will study and answer the questions accordingly. Threading or fold recognition involves similar steps as comparative modelling. The difference

is in the fold identification step. First of all, a structure library needs to be defined. The library can include whole chains, domains, or even conserved protein cores. Once the library is defined, the target sequence will be fitted to each library entry and a energy functionis used to evaluate the fit between the target sequence and the library entries to determine the best possible templates.

Depending on the algorithms to align the target sequence with the folds and the energy functions to determine the best fits, the threading methods can roughly be divided into four classes:

1. The earliest threading methods used the environment of each residue in the structure as the energy function and dynamical programming to evaluate the fit and the alignment
2. Instead of using overly simplified residual environment as the energy function, statistically derived pair wise interaction potentials between residue pairs or atom pairs canbe used to evaluate the best possible fits between the target sequence and library folds. In this method, for efficient optimal alignment between the target sequence and the folds, the potential for residual i is obtained by summing over all the pair wise potentials involving *i*, and then "double dynamical programming" method can be used.
3. The third kind of methods does not use any explicit energy function at all. Instead, secondary structures and accessibility of each residue are predicted first and the target sequence and library folds are encoded into strings for the purpose of sequence-structure alignment.
4. Finally, sequence similarity and threading can be combined for fold recognition. For large-scale genome wise protein structure prediction, sequence similarity can be first used forthe initial alignments and the alignments can be evaluated by threading methods.

Worth mentioning is the threading program PROSPECT, which performed best in its category in the CASP4 competition. What is unique to PROSPECT is that it is designed to find the globally optimal sequence-structure alignment for the given form of energy function. The divide-and-conquer algorithm is used to speed up the calculation by explicitly avoiding the conformation search space that is shown not to contain the optimal alignment. In several cases that have sequence identity as low as 17%, perfect sequence-structure alignment is still achieved for the align able potions between the target and template structures. Even in cases that no fold templates exist for the target sequence, important features of the structure are still recognized through threading the target sequence to the structures.

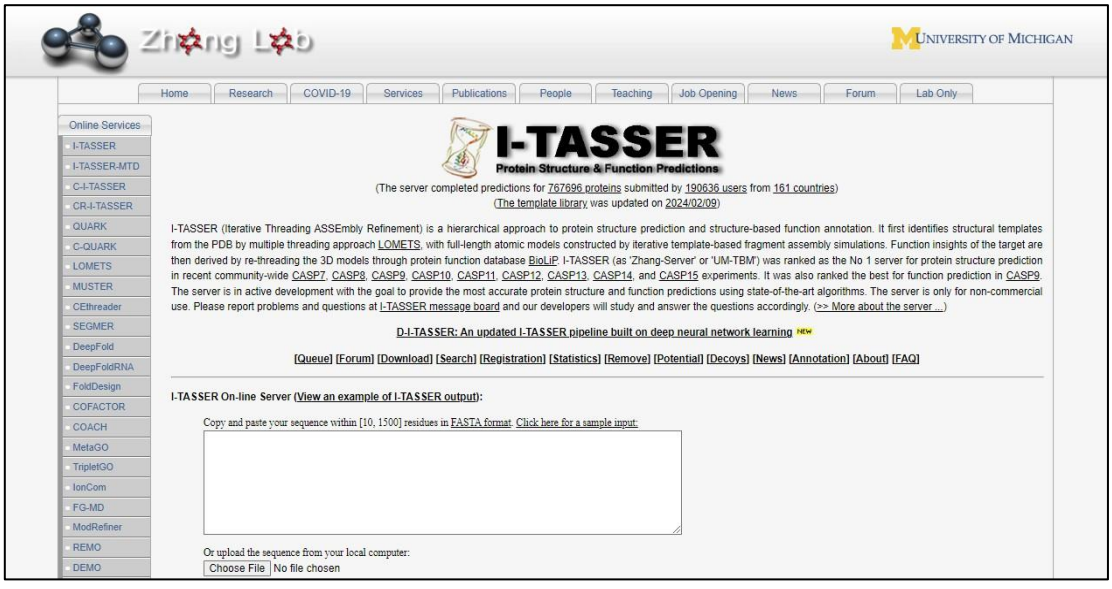


Fig 6: Homepage of I-TASSER server

3. Ab Initio methods:

When no suitable structure templates can be found, Ab Initio methods can be used to predict the protein structure from the sequence information only. Common to all Ab Initio methods are:

1. Suitably defined protein representation and corresponding protein conformation space in that representation.
2. Energy functions compatible with the protein representation
3. Efficient and reliable algorithms to search the conformational space to minimize the energy function

The conformations that minimize the energy function are taken to be the structures that the protein is likely to adopt at native conditions. The folding of the protein sequence is ultimately dictated by the physical forces acting on the atoms of the protein and thus the mostaccurate way of formulating the protein folding or structure prediction problem is in terms of all-atom model subject to the physical forces. Unfortunately, the complexity of such a representation makes the solution simply impossible with today's computational capacity. For practical reasons, most Ab Initio prediction methods use reduced representations of theprotein to limit the conformational space to manageable size and use empirical energyfunctions that capture the most important interactions that drive the folding of the protein sequence toward the native structures. Currently, many Ab Initio methods can predict large contiguous segments of the protein to accuracy within 6_ of RMSD and there are several reviews that highlight the success and failure of the current Ab Initio methods. The ROSETTA Ab Initio method performed better than the other Ab Initio methods in the recent CASP4 meeting and there are extensive literature covering this method so we concentrate on a brief discussion of method used in ROSETTA. The ROSETTA method also illustrates many features and techniques that are common to the majority of the Ab Initio methods based on reduced representation of the protein and empirical potentials.

The ROSETTA method, like many others, uses a reduced representation of the protein as short segments. This representation can be attributed to the observation by Go (Go, 1983) thatlocal segments of the protein sequence have statistically important preferences for specific local structures and that the tertiary structure has to be consistent with this preference. In ROSETTA the protein is represented by short sequence segments and the local structures theycan adopt are assumed to be those found in all the known protein structures. The energy function is defined as the Bayesian probability of structure/sequence matches and this forms the basis of the Monte Carlo sampling of the reduced protein conformational space. The non-local potential, which drives the protein toward compact folded structure, includes terms that favor paired strands and buried hydrophobic residuals. The solvation effect can also be incorporated in the energy function.

Fig 7: Homepage of trRosetta server

REFERENCES:

1. Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol. 1969;42:65–86.
2. 2. Wuthrich K. The way to NMR structures of proteins. Nature Structural Biology. 2001;8:923–925.
3. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. Chem Biol Drug Des. 2019 Jan;93(1):12-20. doi: 10.1111/cbdd.13388. Epub 2018 Oct 8. PMID: 30187647.
4. Altschul, S. F., Gish W., Miller W., Myers E. W., Lipman D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403.
5. Altschul, S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389.

DATE: 16/01/2024

WEBLEM 10 (A)
MODELLER
(URL: https://salilab.org/modeller/tutorial/basic.html)

AIM:

To predict tertiary structure for query "Insulin" (UniProt ID: P06213) using homology-based method – Modeller.

INTRODUCTION:

Homology modeling is one of the computational structure prediction methods that are used to determine protein 3D structure from its amino acid sequence. It is considered to be the most accurate of the computational structure prediction methods. It consists of multiple steps that are straightforward and easy to apply. There are many tools and servers that are used for homology modeling. There is no single modeling program or server which is superior in every aspect to others. Since the functionality of the model depends on the quality of the generated protein 3D structure, maximizing the quality of homology modeling is crucial. Homology modeling has many applications in the drug discovery process. Since drugs interact with receptors that consist mainly of proteins, protein 3D structure determination, and thus homology modeling is important in drug discovery. Accordingly, there has been the clarification of protein interactions using 3D structures of proteins that are built with homology modeling. This contributes to the identification of novel drug candidates. Homology modeling plays an important role in making drug discovery faster, easier, cheaper, and more practical. As new modeling methods and combinations are introduced, the scope of its applications widens.

MODELLER is a computer program for comparative protein structure modeling. In the simplest case, the input is an alignment of a sequence to be modeled with the template structure(s), the atomic coordinates of the template(s), and a simple script file. MODELLER then automatically calculates a model containing all non-hydrogen atoms, without any user intervention and within minutes on a desktop computer. Apart from model building, MODELLER can perform auxiliary tasks such as fold assignment, alignment of two protein sequences or their profiles, multiple alignment of protein sequences and/or structures, clustering of sequences and/or structures, and ab initio modeling of loops in protein structures.

MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints that include: (i) homology-derived restraints on the distances and dihedral angles in the target sequence, extracted from its alignment with the template structures, (ii) stereo chemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics force field, (iii) statistical preferences for dihedral angles and non-bonded interatomic distances, obtained from a representative set of known protein structures, and (iv) optional manually curated restraints, such as those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy,

image reconstruction from electron microscopy, site directed mutagenesis, and intuition. The spatial restraints, expressed as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing. This model building procedure is similar to structure determination by NMR spectroscopy.

Insulin:

Insulin is a vital hormone that regulates blood sugar levels by allowing glucose to enter cells for energy production. It is produced by beta cells in the pancreas and plays a crucial role in glucose storage and production. Insulin was first isolated in 1921 by Canadian scientists Frederick G. Banting and Charles H. Best, leading to life-saving treatments for diabetes. In diabetes, either the body does not produce enough insulin (Type 1) or becomes resistant to its effects (Type 2). Insulin resistance can lead to high blood sugar levels and various health complications. Different types of insulin, including fast, intermediate, and long-acting insulins, are used based on individual needs to manage blood sugar effectively.

Insulin is a composed of two peptide chains, an A chain and a B chain, linked together by disulfide bonds. The A chain consists of 21 amino acids, while the B chain has 30 amino acids. Within the A chain, there is an additional disulfide bond. Insulin molecules have a tendency to form dimers in solution and can associate into hexamers in the presence of zinc ions. The amino acid sequence of insulin is highly conserved among species, with minor variations. Despite these variations, insulin from different species can be biologically active across species. The structure of insulin allows it to regulate blood glucose levels by promoting glucose storage and inhibiting glucose production and release by the liver.

METHODOLOGY:

1. Open Modeller, go to the tutorial section:
   - Basic Modeling
2. Download the library files within Modeller.
3. Create a folder and transfer the library file into it.

STEP 1: Searching for structures related to query structure

- Copy the script generated and paste it into a notepad.
- Open Uniprot database, search for the query ‘Insulin’ download the FASTA sequence, and paste the sequence into another notepad.
- Save the sequence file with a (.ali) extension and place it into the folder.

STEP 2: Selecting a template

- Copy the script for the sequence, paste it into a new notepad, adjust the file name, and save it as script1.py.
- In the Modeller terminal, paste the path and execute the script1.py file to obtain templates in the build profile.

STEP 3: Aligning query structure with the template

- Copy the accession ID from script1 and use it to download all corresponding PDB files from the PDB homepage.
- Transfer all four PDB files into the folder.

STEP 4: Model building

- Copy the compare.py script into a notepad, modify it to select only the top 5 templates, save the file as script2.py, and execute it to obtain a family file containing matrix, NMR, and crystallography values.
- Observe the X-ray values and NMR values.
- Align the QSEQ with the template file using the script in a notepad; save it as script3.py and run.
- Prepare for model building by copying the script into a notepad save it as script4.py and execute it in the terminal to produce models This will create five finished models. GA341 and DOPE scores are also observed.

STEP 5: Model evaluation

- Proceed to model validation on the specified website by uploading and checking the results.

OBSERVATIONS:

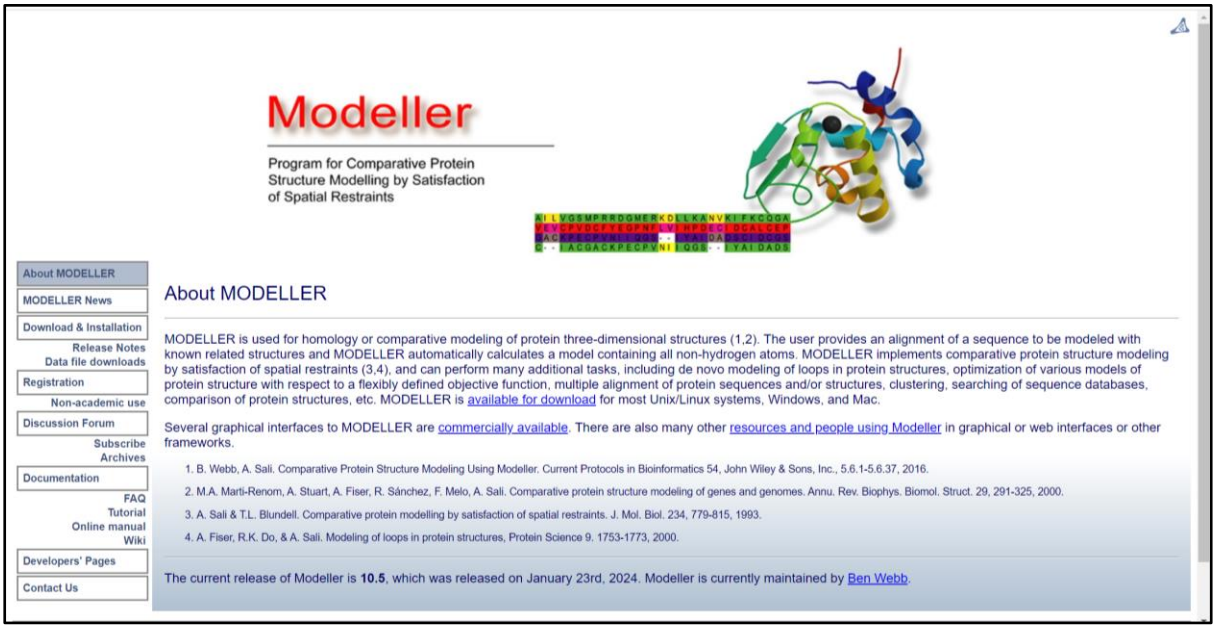


Fig 1: Homepage of Modeller

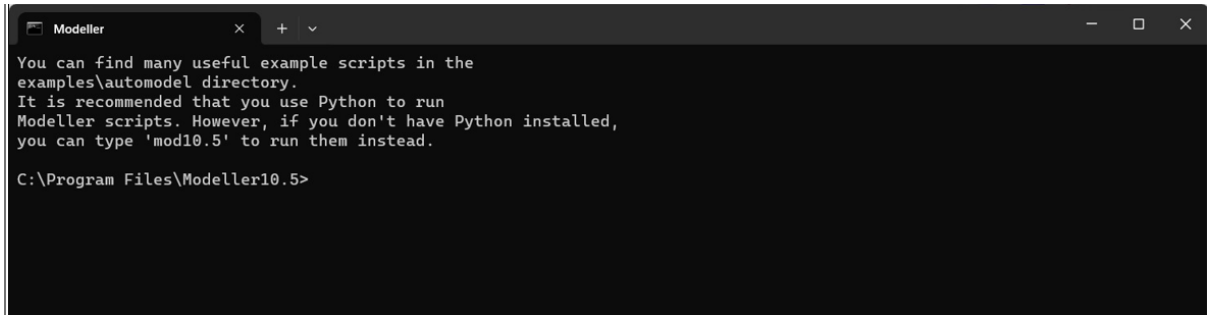


Fig 2: Modeller10.4 Command Line

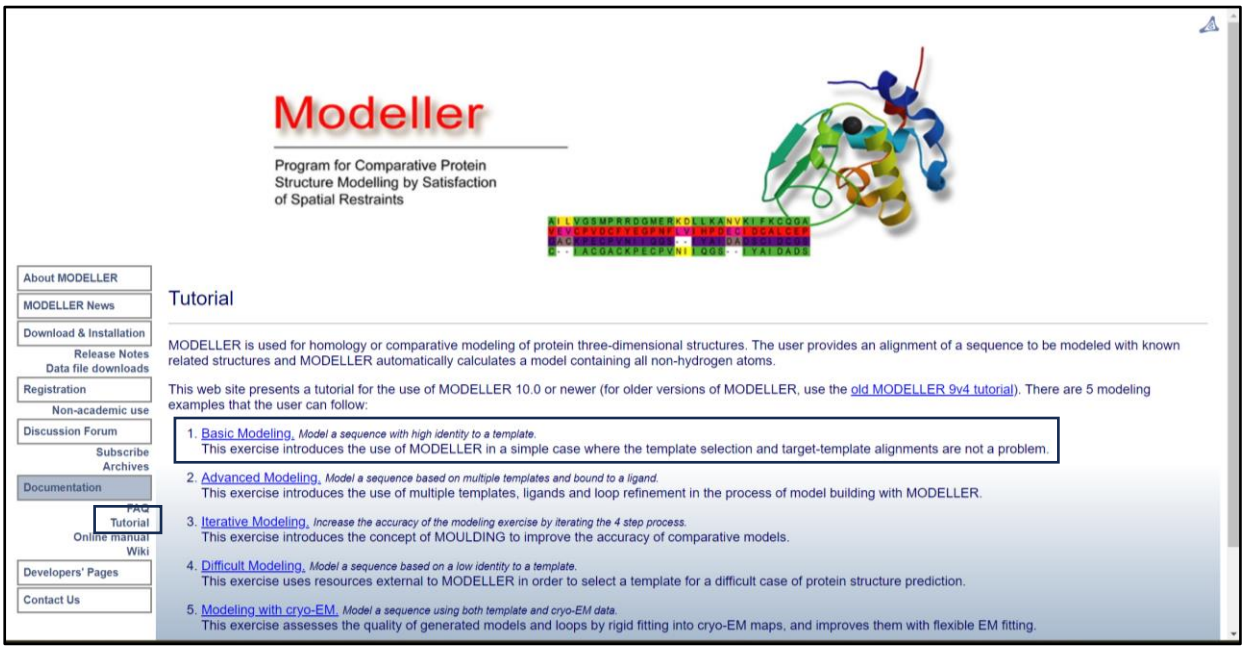


Fig 3: Tutorial page of Modeller

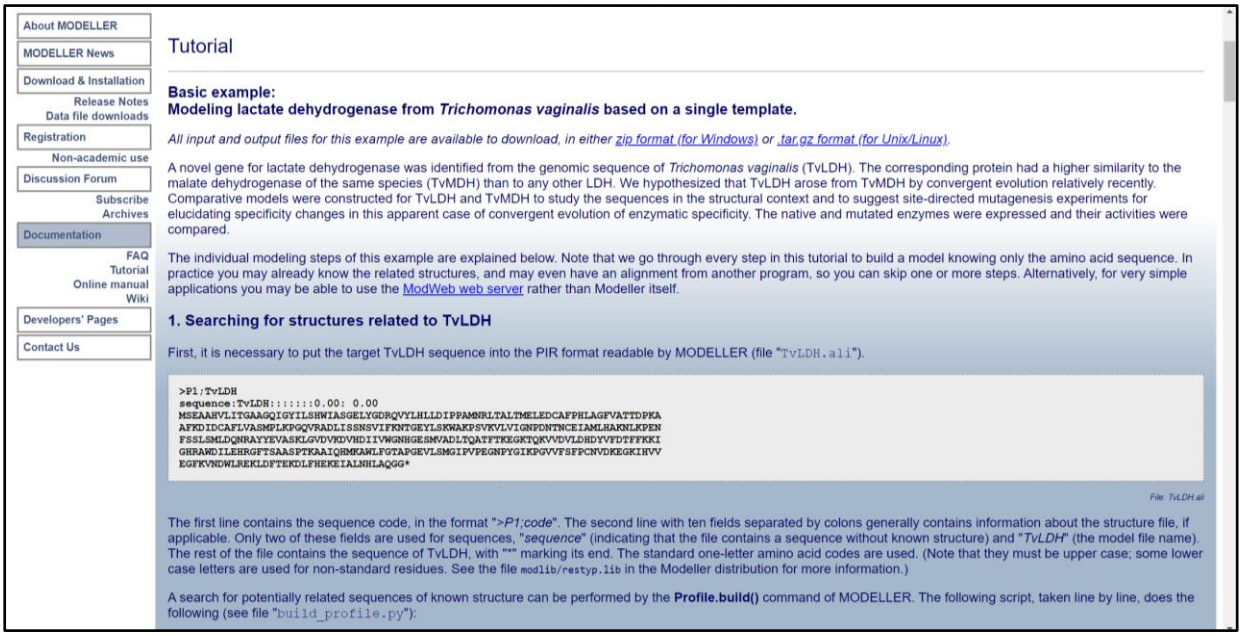


Fig 4: Basic Modelling Homepage

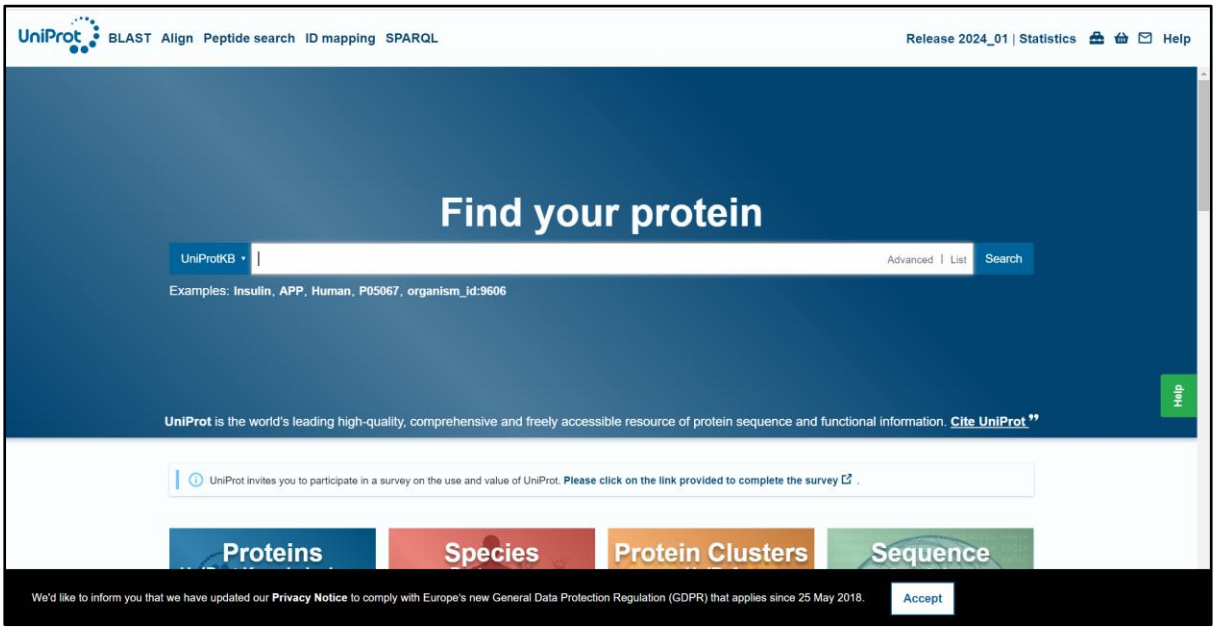


Fig 5: Homepage of UniProt Database

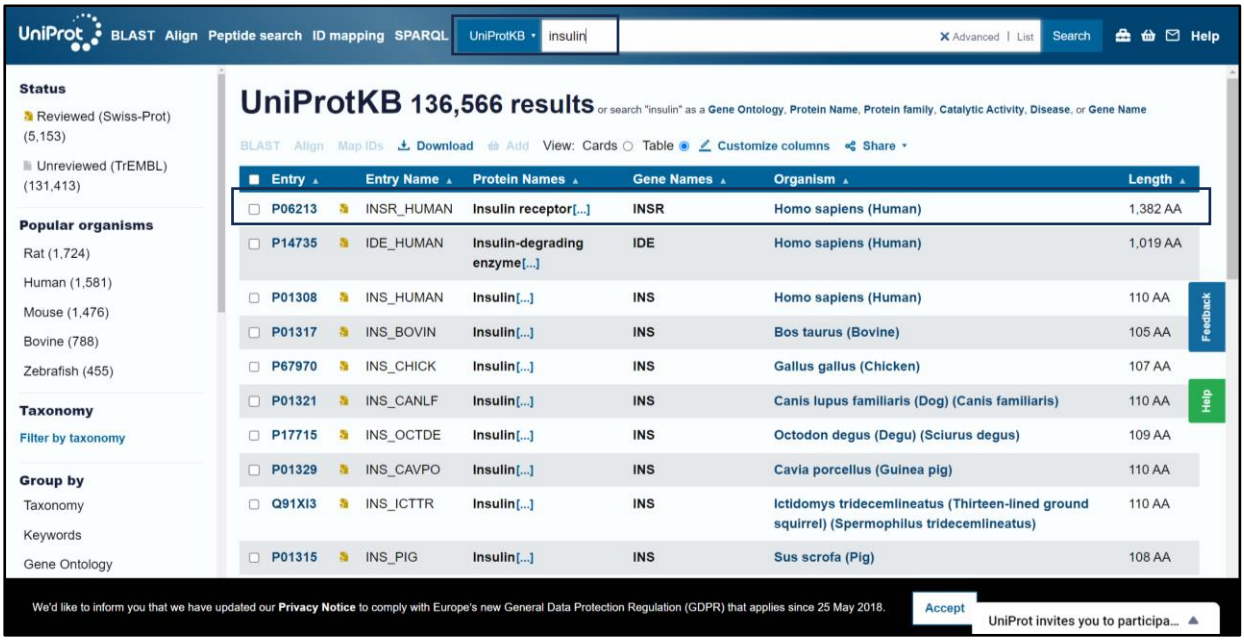


Fig 6: Searching for the query 'Insulin' and selecting the query with UniProt ID: P06213

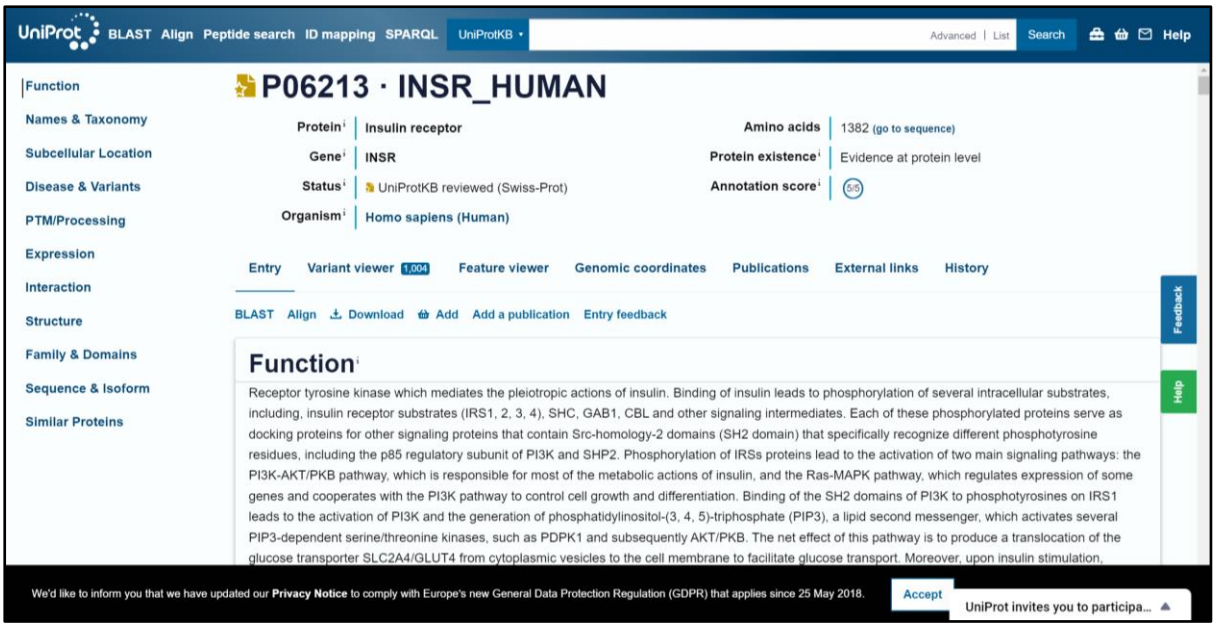


Fig 7: Information for the query 'Insulin' (UniProt ID: P06213)

```
>sp|P06213|INSR_HUMAN Insulin receptor OS=Homo sapiens OX=9606 GN=INSR PE=1 SV=4
MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
QILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFPNLTVIRGSRLFFNYAL
VIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRILDSVEDNYIVLNKDDNE
ECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTICKSHGCTAEGLCCHSECL
GNCSQPDDPTKCVACRNFYLDGRCVETCPPPYYHFQDWRCVNFSFCQDLHHKCKNSRRQG
CHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKVCHLLEGEKTIDSVTSAQELRGC
TVINGSLIINIRGGNNLAAELEANLGLIEEISGYLKIRRSYALVSLSFFRKLRLIRGETL
EIGNYSFYALDNQNLRQLWDWSKHNLTITQGKLFFHYNPKLCLSEIHKMEEVSGTKGRQE
RNDIALKTNGDQASCENELLKFSYIRTSFDKILLRWEPYWPPDFRDLLGFMLFYKEAPYQ
NVTEFDGQDACGSNSWTVVDIDPPLRSNDPKSQNHPGWLMRGLKPWTQYAIFVKTLVTFS
DERRTYGAKSDIIYVQTDATNPSVPLDPISVSNSSSQIILKWKPPSDPNGNITHYLVFWE
RQAEDSELFELDYCLKGLKLPSRTWSPPFESEDSQKHNQSEYEDSAGECCSCPKTDSQIL
KELEESSFRKTFEDYLHNVVFVPRKTSSGTGAEDPRPSRKRRSLGDVGNVTVAVPTVAAF
PNTSSTSVPTSPEEHRPFEKVVNKESLVISGLRHFTGYRIELQACNQDTPEERCSVAAYV
SARTMPEAKADDIVGPVTHEIFENNVVHLMWQEPKEPNGLIVLYEVSYRRYGDEELHLCV
SRKHFALERGCRLRGLSPGNYSVRIRATSLAGNGSWTEPTYFYVTDYLDVPSNIAKIIIG
PLIFVFLFSVVIGSIYLFLRKRQPDGPLGPLYASSNPEYLSASDVFPCSVYVPDEWEVSR
EKITLLRELGQGSFGMVYEGNARDIIKGEAETRVAVKTVNESASLRERIEFLNEASVMKG
FTCHHVVRLLGVVSKGQPTLVVMELMAHGDLKSYLRSLRPEAENNPGRPPPTLQEMIQMA
AEIADGMAYLNAKKFVHRDLAARNCMVAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPV
RWMAPESLKDGVFTTSSDMWSFGVVLWEITSLAEQPYQGLSNEQVLKFVMDGGYLDQPDN
CPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEME
FEDMENVPLDRSSHCQREEAGGRDGGSSLGFKRSYEEHIPYTHMNGGKKNGRILTLPRSN
PS
```

Fig 8: Downloading the FASTA (Canonical) sequence for the query 'Insulin' (UniProt ID: P06213)

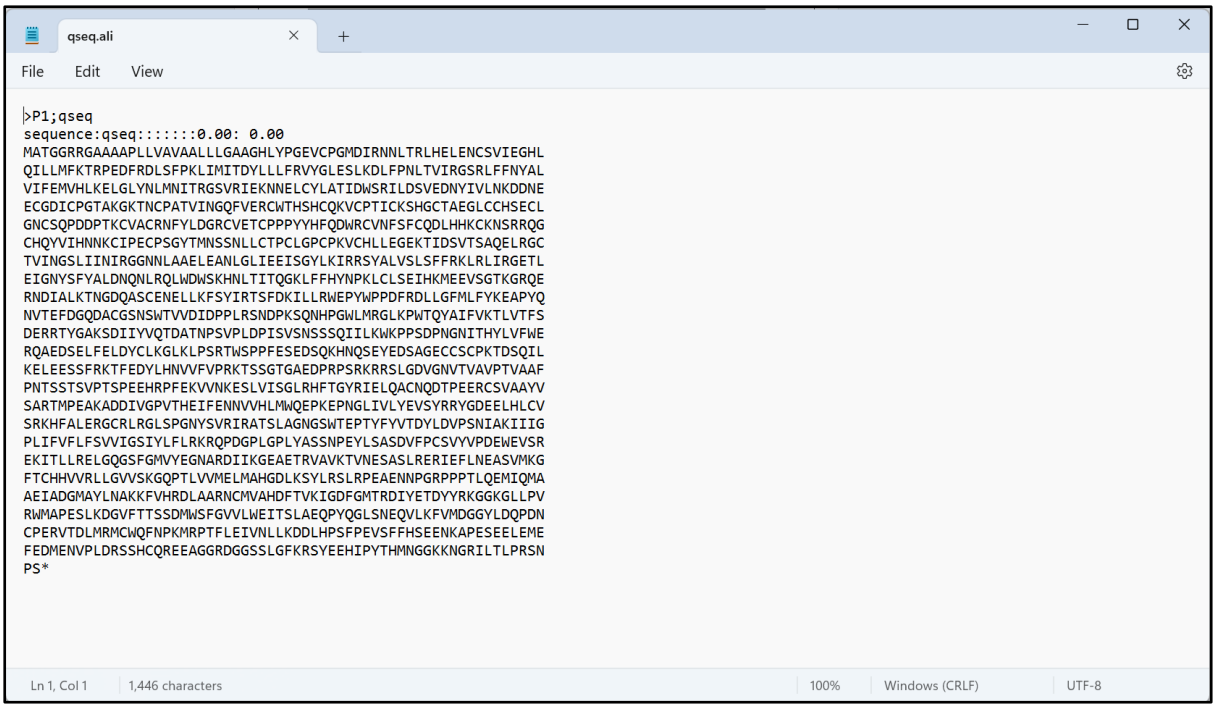


Fig 9: Script for the query sequence (qseq) 'Insulin' (UniProt ID: P06213)

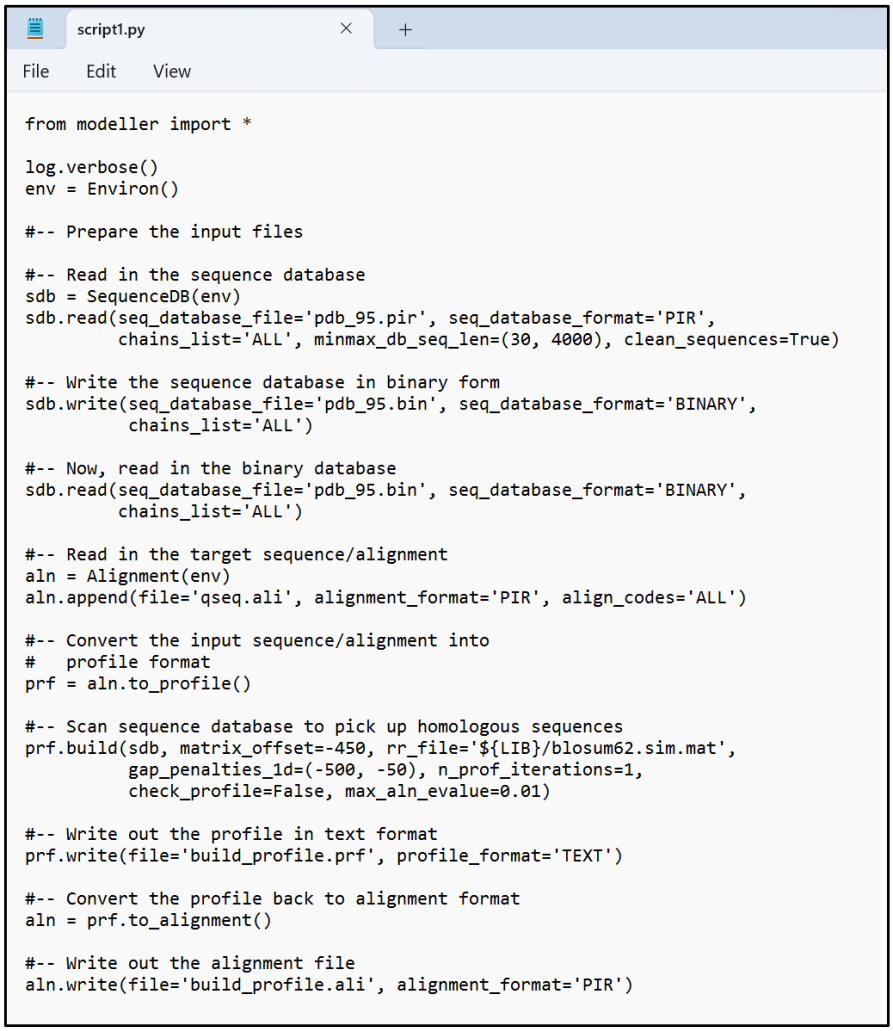


Fig 10: Script file 'script1.py'

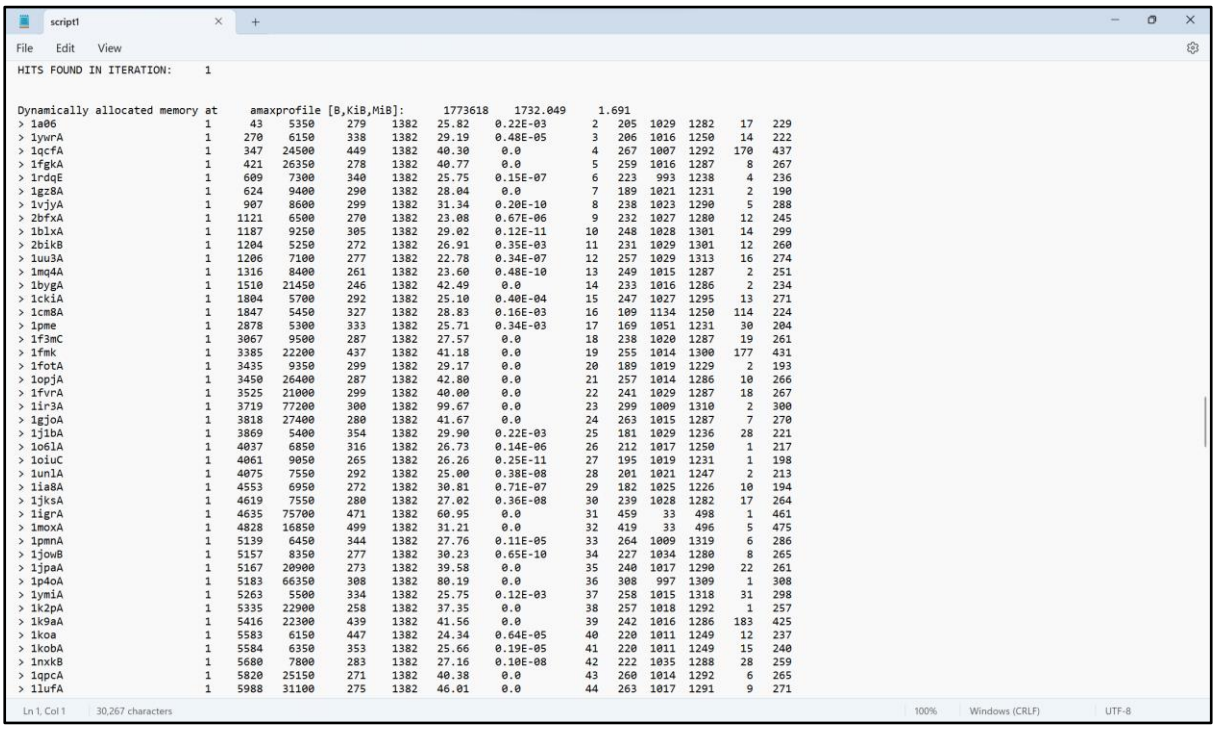
```
You can find many useful example scripts in the
examples\automodel directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.4' to run them instead.

C:\Program Files\Modeller10.4>cd bin

C:\Program Files\Modeller10.4\bin>cd hseq

C:\Program Files\Modeller10.4\bin\hseq>mod10.4 scriptt1.py
'import site' failed; use -v for traceback
```

Fig 11: Executing script1.py in Modeller Terminal



script1

File Edit View

```
HITS FOUND IN ITERATION:     1


Dynamically allocated memory at     amaxprofile [B,KiB,MiB]:      1773618    1732.049     1.691
> 1a06                      1     43    5350     279    1382   25.82    0.22E-03       2   205  1029  1282    17   229
> 1ywrA                     1    270    6150     338    1382   29.19    0.48E-05       3   206  1016  1250    14   222
> 1qcfA                     1    347   24500     449    1382   40.30     0.0           4   267  1007  1292   170   437
> 1fgkA                     1    421   26350     278    1382   40.77     0.0           5   259  1016  1287     8   267
> 1rdqE                     1    609    7300     340    1382   25.75    0.15E-07       6   223   993  1238     4   236
> 1gz8A                     1    624    9400     290    1382   28.04     0.0           7   189  1021  1231     2   190
> 1vjyA                     1    907    8600     299    1382   31.34    0.20E-10       8   238  1023  1290     5   288
> 2bfxA                     1   1121    6500     270    1382   23.08    0.67E-06       9   232  1027  1280    12   245
> 1blxA                     1   1187    9250     305    1382   29.02    0.12E-11      10   248  1028  1301    14   299
> 2bikB                     1   1204    5250     272    1382   26.91    0.35E-03      11   231  1029  1301    12   260
> 1uu3A                     1   1206    7100     277    1382   22.78    0.34E-07      12   257  1029  1313    16   274
> 1mq4A                     1   1316    8400     261    1382   23.60    0.48E-10      13   249  1015  1287     2   251
> 1bygA                     1   1510   21450     246    1382   42.49     0.0          14   233  1016  1286     2   234
> 1ckiA                     1   1804    5700     292    1382   25.10    0.40E-04      15   247  1027  1295    13   271
> 1cm8A                     1   1847    5450     327    1382   28.83    0.16E-03      16   109  1134  1250   114   224
> 1pme                      1   2878    5300     333    1382   25.71    0.34E-03      17   169  1051  1231    30   204
> 1f3mC                     1   3067    9500     287    1382   27.57     0.0          18   238  1020  1287    19   261
> 1fmk                      1   3385   22200     437    1382   41.18     0.0          19   255  1014  1300   177   431
> 1fotA                     1   3435    9350     299    1382   29.17     0.0          20   189  1019  1229     2   193
> 1opjA                     1   3450   26400     287    1382   42.80     0.0          21   257  1014  1286    10   266
> 1fvrA                     1   3525   21000     299    1382   40.00     0.0          22   241  1029  1287    18   267
> 1ir3A                     1   3719   77200     300    1382   99.67     0.0          23   299  1009  1310     2   300
> 1gjoA                     1   3818   27400     280    1382   41.67     0.0          24   263  1015  1287     7   270
> 1j1bA                     1   3869    5400     354    1382   29.90    0.22E-03      25   181  1029  1236    28   221
> 1o6lA                     1   4037    6850     316    1382   26.73    0.14E-06      26   212  1017  1250     1   217
> 1oiuC                     1   4061    9050     265    1382   26.26    0.25E-11      27   195  1019  1231     1   198
> 1unlA                     1   4075    7550     292    1382   25.00    0.38E-08      28   201  1021  1247     2   213
> 1ia8A                     1   4553    6950     272    1382   30.81    0.71E-07      29   182  1025  1226    10   194
> 1jksA                     1   4619    7550     280    1382   27.02    0.36E-08      30   239  1028  1282    17   264
> 1igrA                     1   4635   75700     471    1382   60.95     0.0          31   459    33   498     1   461
> 1moxA                     1   4828   16850     499    1382   31.21     0.0          32   419    33   496     5   475
> 1pmnA                     1   5139    6450     344    1382   27.76    0.11E-05      33   264  1009  1319     6   286
> 1jowB                     1   5157    8350     277    1382   30.23    0.65E-10      34   227  1034  1280     8   265
> 1jpaA                     1   5167   20900     273    1382   39.58     0.0          35   240  1017  1290    22   261
> 1p4oA                     1   5183   66350     308    1382   80.19     0.0          36   308   997  1309     1   308
> 1ymiA                     1   5263    5500     334    1382   25.75    0.12E-03      37   258  1015  1318    31   298
> 1k2pA                     1   5335   22900     258    1382   37.35     0.0          38   257  1018  1292     1   257
> 1k9aA                     1   5416   22300     439    1382   41.56     0.0          39   242  1016  1286   183   425
> 1koa                      1   5583    6150     447    1382   24.34    0.64E-05      40   220  1011  1249    12   237
> 1kobA                     1   5584    6350     353    1382   25.66    0.19E-05      41   220  1011  1249    15   240
> 1nxkB                     1   5680    7800     283    1382   27.16    0.10E-08      42   222  1035  1288    28   259
> 1qpcA                     1   5820   25150     271    1382   40.38     0.0          43   260  1014  1292     6   265
> 1lufA                     1   5988   31100     275    1382   46.01     0.0          44   263  1017  1291     9   271
```

Ln 1, Col 1 30,267 characters 100% Windows (CRLF) UTF-8

Fig 12: Interpretation for script1

script2.py

File Edit View

```
from modeller import *

env = Environ()
aln = Alignment(env)
for (pdb, chain) in (('1qcf', 'A'), ('1fgk', 'A'), ('1gz8', 'A'),
                     ('1byg', 'A'), ('1f3m', 'C')):
    m = Model(env, file=pdb, model_segment=('FIRST:'+chain, 'LAST:'+chain))
    aln.append_model(m, atom_files=pdb, align_codes=pdb+chain)
aln.malign()
aln.malign3d()
aln.compare_structures()
aln.id_table(matrix_file='family.mat')
env.dendrogram(matrix_file='family.mat', cluster_cut=-1.0)
```

Fig 13: Script file 'script2.py' ('compare.py')

```
C:\Program Files\Modeller10.4\bin\hseq>mod10.4 scriptt2.py
'import site' failed; use -v for traceback
```

Fig 14: Executing script2.py in Modeller Terminal

script2

File Edit View

```
Sequence identity comparison (ID_TABLE):

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

         1qcfA @21fgkA @21gz8A @11bygA @21f3mC @2
1qcfA @2      449      88      62      94      41
1fgkA @2       32     278      47      92      48
1gz8A @1       21      17     290      44      59
1bygA @2       38      37      18     246      49
1f3mC @2       14      17      21      20     287


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

                                                          .--- 1qcfA @2.0    62.0000
                                                          |
                                                  .------------ 1bygA @2.4    65.5000
                                                  |
        .-------------------------------------------------------- 1fgkA @2.0    82.3750
        |
        |       .---------------------------------------------- 1gz8A @1.3    79.0000
        |       |
      .-------------------------------------------------------- 1f3mC @2.3

      +----+----+----+----+----+----+----+----+----+----+----+----+
    83.1900   79.5225   75.8550   72.1875   68.5200   64.8525   61.1850
         81.3563   77.6888   74.0213   70.3538   66.6863   63.0188


Total CPU time [seconds]                          :       0.61
```

Ln 2360, Col 7 | 1,23,275 characters | 100% | Windows (CRLF) | UTF-8

Fig 15: Interpretation for script2

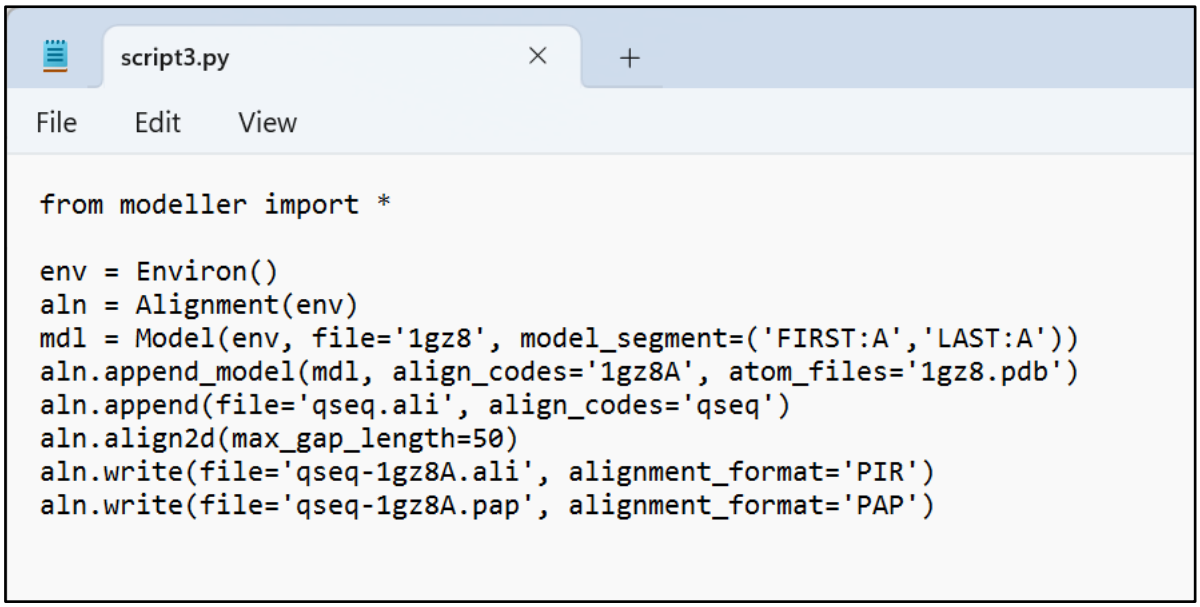


Fig 16: Script file 'script3.py' for aligning the 'qseq' with the template file (1gz8.pdb)

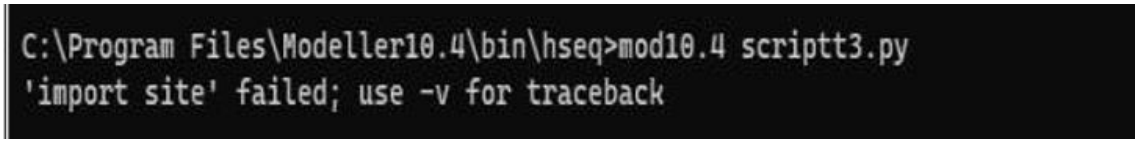


Fig 17: Executing script3.py in Modeller Terminal

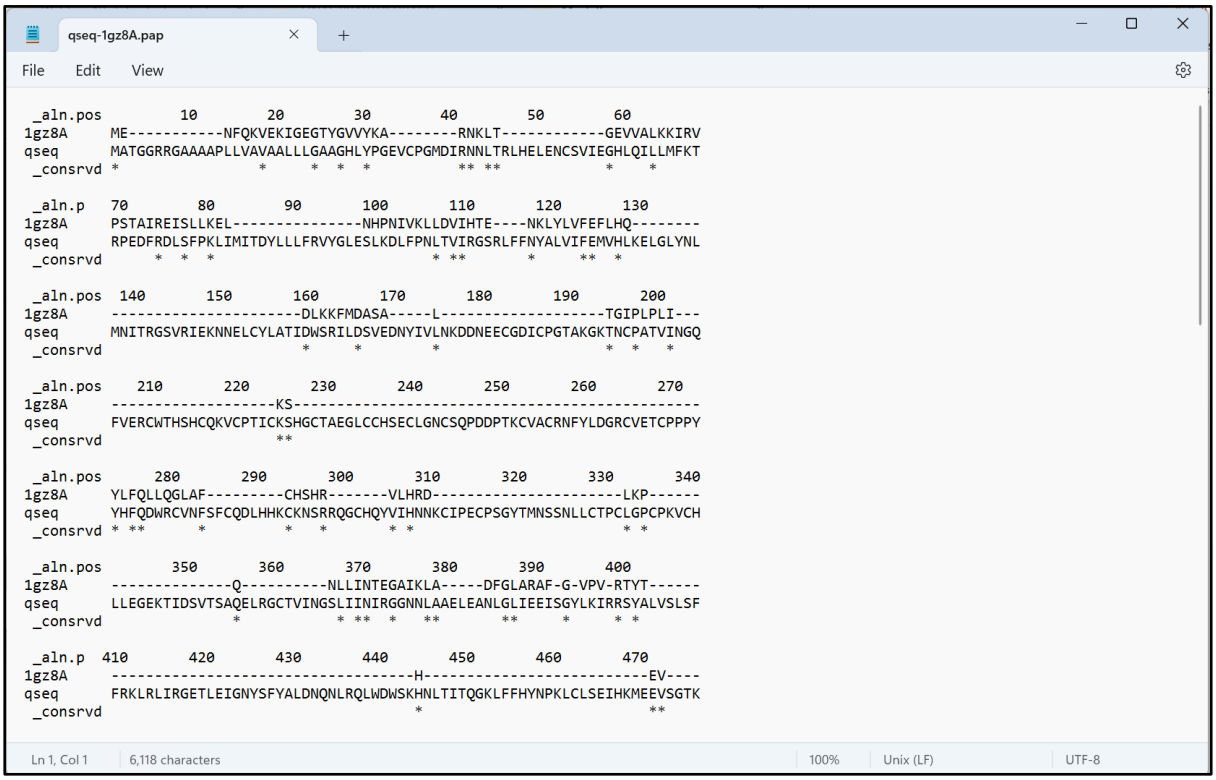


Fig 18: Interpretation for the alignment of the 'qseq' with the template file (1gz8.pdb) in the form of Modeller .pap file format

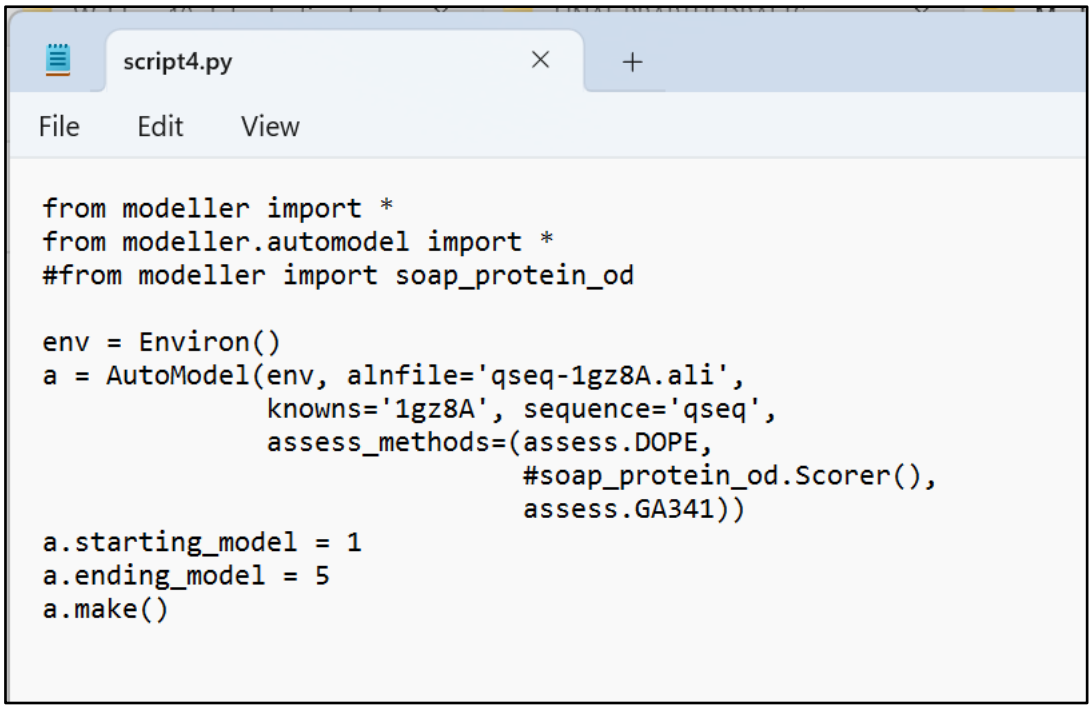


Fig 19: Script file 'script4.py' for generating five models of the aligned sequences

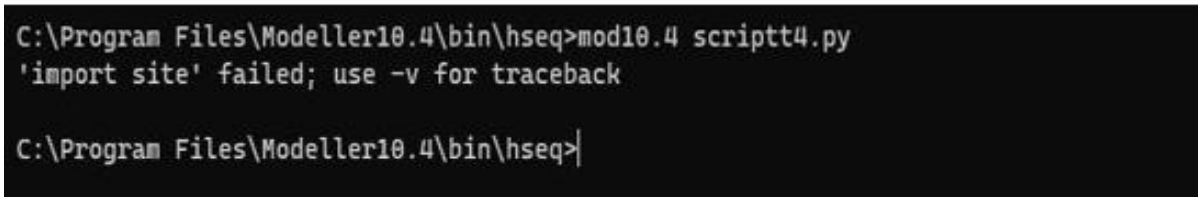


Fig 20: Executing script4.py in Modeller Terminal

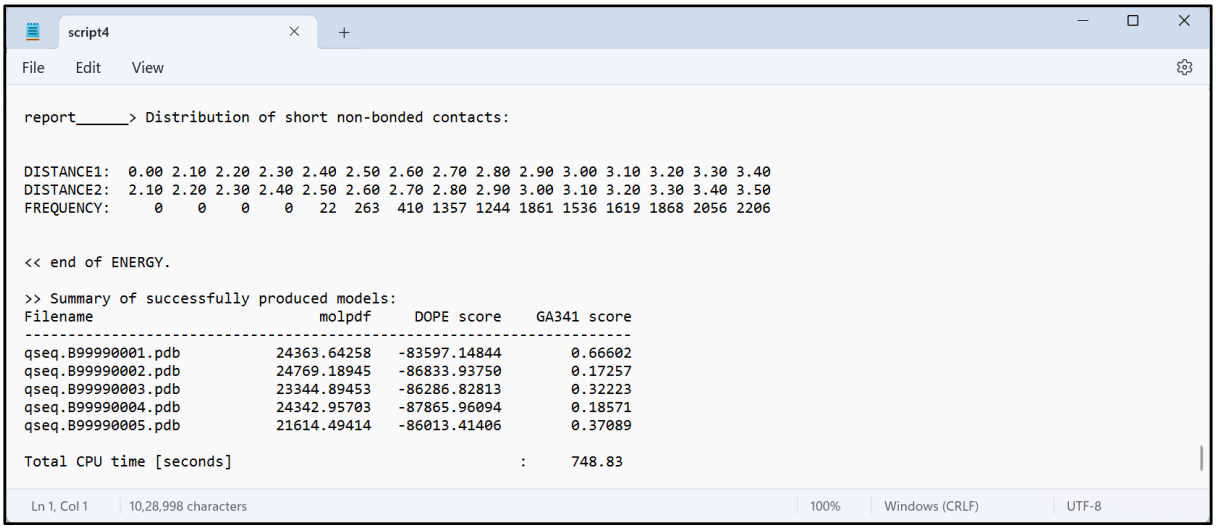


Fig 21: Interpretation for script4 and display of DOPE scores and GA341 scores

RESULTS:

Using the Homology based method (Modeller), tertiary structure was predicted for the query 'Insulin' (UniProt ID: P06213). The templates '1qcfA', '1bygA', '1fgkA', '1qz8A', '1f3mC' were obtained by running the script1.py. After which the PDB files were downloaded from the PDB database. Further, the templates were compared to find the best match after running script2.py based on the values of X-Ray, Crystallography and NMR values obtained. Based on the lowest X-ray value of 1.3 and the NMR value of 79.0000, the template '1gz8A' was found to be the best match. The template '1gz8A' was assigned with the target sequence 'Insulin' (UniProt ID: P06213) and the PAP file format was viewed to observe the conserved region. After aligning the target sequence and the template, script4.py was executed. Based on the results obtained, the final model 'qseq.B99990001.pdb' was selected based on the highest GA341 score of 0.66602 and DOPE score of -83597.14844.

CONCLUSION:

Homology modelling - Modeller program was used to predict tertiary structure of protein query sequence 'Insulin' (UniProt ID: P06213) The model 'qseq.B99990001.pdb' was found to be the best fit for the query which may be further confirmed through validation of the predicted tertiary structure.

REFERENCES:

1. Frank, M., & Schloissnig, S. (2010). Bioinformatics and molecular modeling in glycobiology. Cellular and Molecular Life Sciences, 67(16), 2749–2772.
2. https://doi.org/10.1007/s00018-010-0352-4
3. Berman, H. M. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242. https://doi.org/10.1093/nar/28.1.235
4. Venselaar, H., Beek, T. a. H. T., Kuipers, R., Hekkelman, M. L., & Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics, 11(1). https://doi.org/10.1186/1471-2105-11-548
5. Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of Insulin in Health and Disease: An Update. International journal of molecular sciences, 22(12), 6403. https://doi.org/10.3390/ijms22126403
6. Weiss M, Steiner DF, Philipson LH. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. [Updated 2014 Feb 1]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279029/

DATE: 31/01/2024

WEBLEM 10(B)
I-TASSER
(URL: https://zhanggroup.org/I-TASSER/)

AIM:

To predict protein three-dimensional conformations by exploring I-TASSER (Iterative Threading ASSEmbly Refinement) server for query 'Thrombin' (UniProt ID: TR139059).

INTRODUCTION:

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.

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure prediction and structure-based function annotation. It first identifies structural templates from the PDB by multiple threading approach LOMETS, with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through protein function database BioLiP. I-TASSER (as 'Zhang-Server' or 'UM-TBM') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, CASP13, CASP14, and CASP15 experiments. It was also ranked the best for function prediction in CASP9. The server is in active development with the goal to provide the most accurate protein structure and function predictions using state-of-the-art algorithms. The server is only for non-commercial use. Please report problems and questions at I-TASSER message board and our developers will study and answer the questions accordingly.

Prediction of 3-dimensional protein structures from amino acid sequences represents one of the most important problems in computational structural biology. The community-wide Critical Assessment of Structure Prediction (CASP) experiments have been designed to obtain an objective assessment of the state-of-the-art of the field, where I-TASSER was ranked as the best method in the server section of the recent 7th CASP experiment. Our laboratory has since then received numerous requests about the public availability of the I-TASSER algorithm and the usage of the I-TASSER predictions.

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.

TM-score is a recently proposed scale for measuring the structural similarity between two structures (see Zhang and Skolnick, Scoring function for automated assessment of protein structure template quality, Proteins, 2004 57: 702-710). The purpose of proposing TM-score is to solve the problem of RMSD which is sensitive to the local error. Because RMSD is an average distance of all residue pairs in two structures, a local error (e.g. a misorientation of the tail) will arise a big RMSD value although the global topology is correct. In TM-score, however, the small distance is weighted stronger than the big distance which makes the score insensitive to the local modelling error. A TM-score >0.5 indicates a model of correct topology

and a TM-score<0.17 means a random similarity. These cutoff does not depend on the protein length.

Thrombin:

Thrombin is a Na+-activated, allosteric serine protease that plays opposing functional roles in blood coagulation. Binding of Na+ is the major driving force behind the procoagulant, prothrombotic and signalling functions of the enzyme, but is dispensable for cleavage of the anticoagulant protein C. The anticoagulant function of thrombin is under the allosteric control of the cofactor thrombomodulin. Much has been learned on the mechanism of Na+ binding and recognition of natural substrates by thrombin. Recent structural advances have shed light on the remarkable molecular plasticity of this enzyme and the molecular underpinnings of thrombin allostery mediated by binding to exosite I and the Na+ site. This review summarized our current understanding of the molecular basis of thrombin function and allosteric regulation. The basic information emerging from recent structural, mutagenesis and kinetic investigation of this important enzyme is that thrombin exists in three forms, E*, E and E:Na+, that interconvert under the influence of ligand binding to distinct domains. The transition between the Na+-free slow from E and the Na+-bound fast form E:Na+ involves the structure of the enzyme as a whole, and so does the interconversion between the two Na+-free forms E* and E. E* is most likely an inactive form of thrombin, unable to interact with Na+ and substrate. The complexity of thrombin function and regulation has gained this enzyme pre-eminence as the prototypic allosteric serine protease. Thrombin is now looked upon as a model system for the quantitative analysis of biologically important enzymes.

METHODOLOGY:

1. Access the website:

   Navigate to the I-TASSER website: https://zhanggroup.org/forum/index.php?f=3

2. Prepare your protein sequence:
   - You can either paste your protein sequence directly into the "Protein Sequence" box or upload a FASTA file containing the sequence.
   - Ensure your sequence is in single-letter amino acid code format.
   - If you have multiple sequences, you can predict them individually or use the "Batch Predict" option for simultaneous prediction of up to 10 sequences.

3. Choose your prediction options:
   - Template-based modeling: This is the default option and uses known protein structures (templates) to build a model for your protein.
   - Ab initio modeling: This option attempts to build a model for your protein without relying on templates, suitable for novel or highly divergent proteins.
   - Hybrid modeling: Combines template-based and ab initio approaches for potentially better accuracy.

4. Advanced settings (optional):
   For experienced users, I-TASSER offers various advanced settings like:
   - Confidence level: Choose the desired level of accuracy and computational cost.
   - SP3 refinement: Improves side-chain placements for higher model quality.
   - Ligand binding site prediction: Predicts potential ligand binding sites on the protein.

5. Submit your prediction:
   - Click the "Predict" button to initiate the prediction process. Depending on the chosen options and sequence complexity, the prediction can take minutes to hours.

6. Analyse the results:
   Once completed, you'll receive a web page summarizing the prediction results. This includes:
   - Predicted 3D model: View the predicted protein structure in various interactive formats like PyMOL.
   - Model confidence score: Assess the reliability of the prediction.
   - Secondary structure prediction: Analyse the predicted alpha helices, beta sheets, and coils in the protein.
   - Function prediction: View potential functional annotations based on the predicted structure.
   - Download options: Download the model files, figures, and other data in various formats.

OBSERVATIONS:

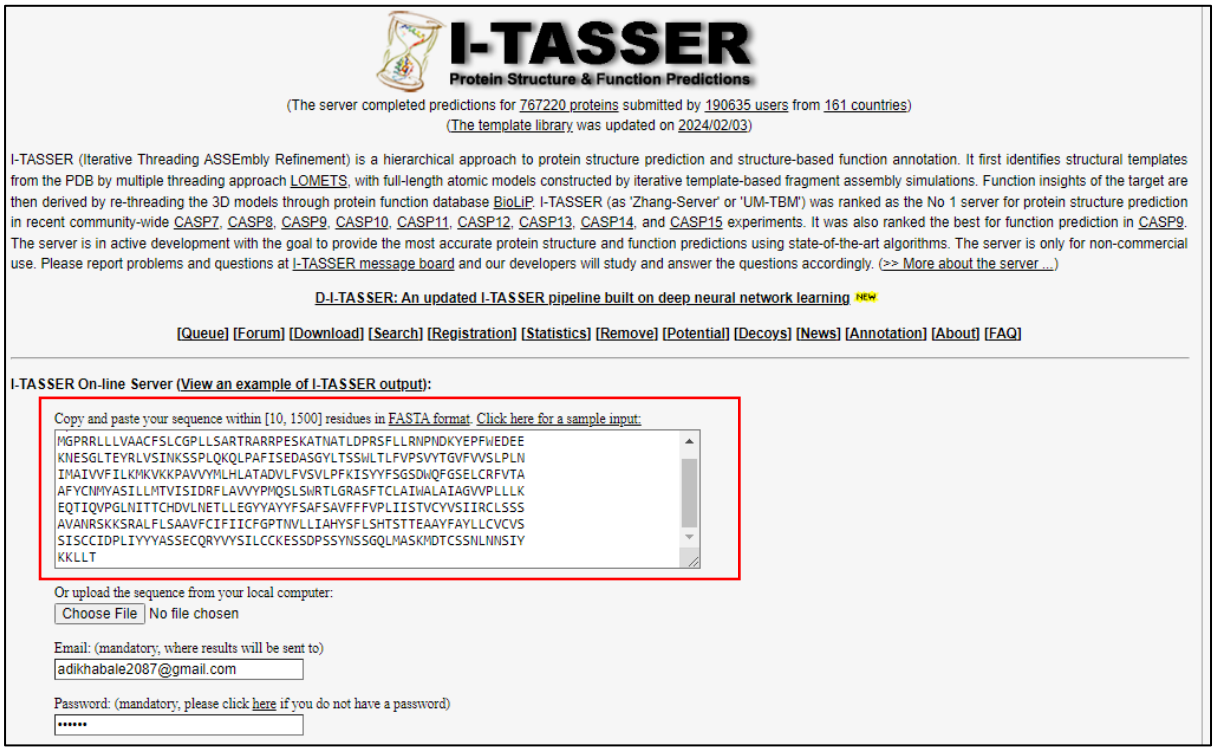


Fig 1: Homepage of I-TASSER server

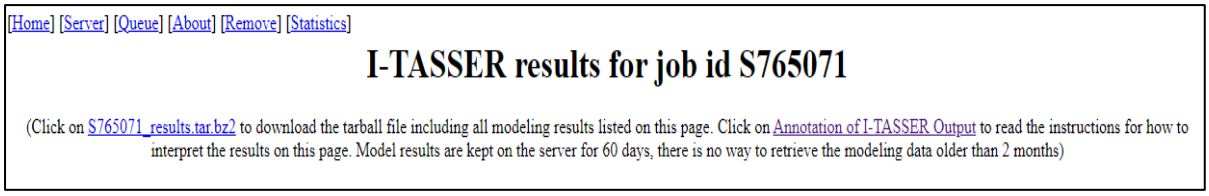


Fig 2: The tarball of I-TASSER modelling results. Tarball allows to download the result

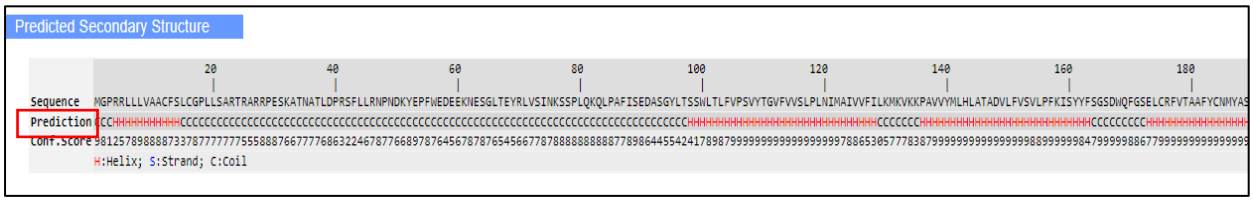


Fig 3: The sequence-based prediction of secondary structure by PSSpred tool. Higher score means more confident prediction of secondary structure

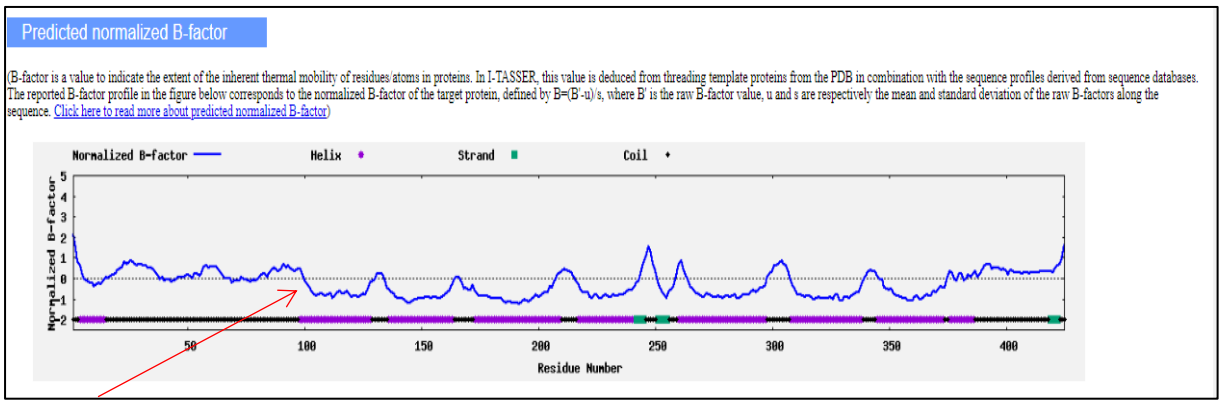


Fig 4: The predicted normalized B-factor by ResQ method indicates Low B-factor (negative or close to zero) This indicates a rigid region in the predicted structure. These residues are likely part of well-defined secondary structures

Fig 5: The top 10 template-query alignments generated by LOMETS server. If Norm. Z-score is greater than 1 it means good alignment. The higher, the better. This can be used to judge whether the protein is an easy or a hard target

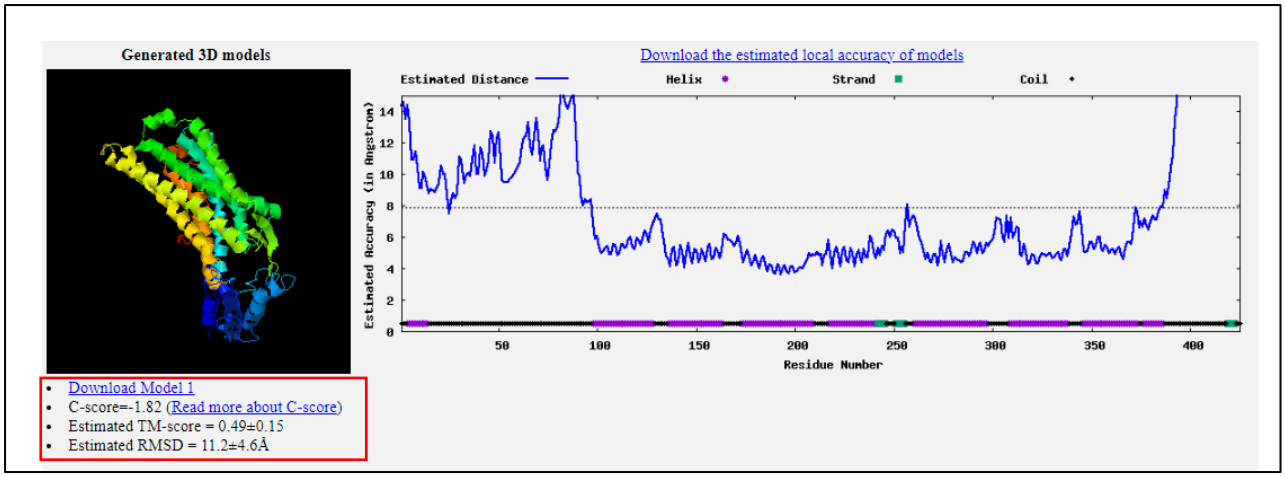


Fig 6: The predicted 3D model and the estimated global and local accuracy by I-TASSER. C-score is in [-5 to 2] and C-score > -1.5 indicates a model of correct global topology

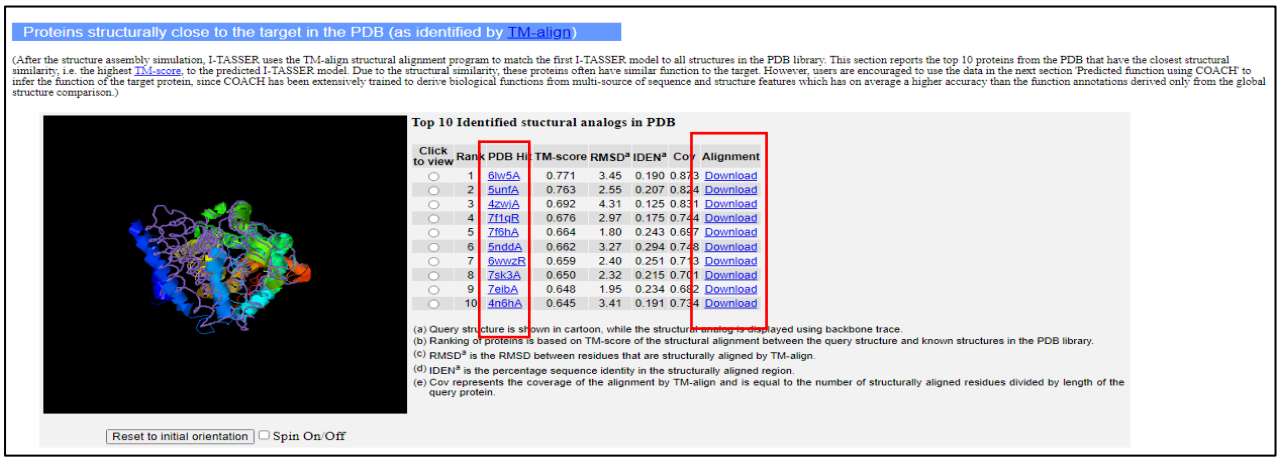


Fig 7: The structure alignment between the first I-TASSER model and the top 10 most similar structure templates in PDB database. Here we can visualize structures and download it in PDB format

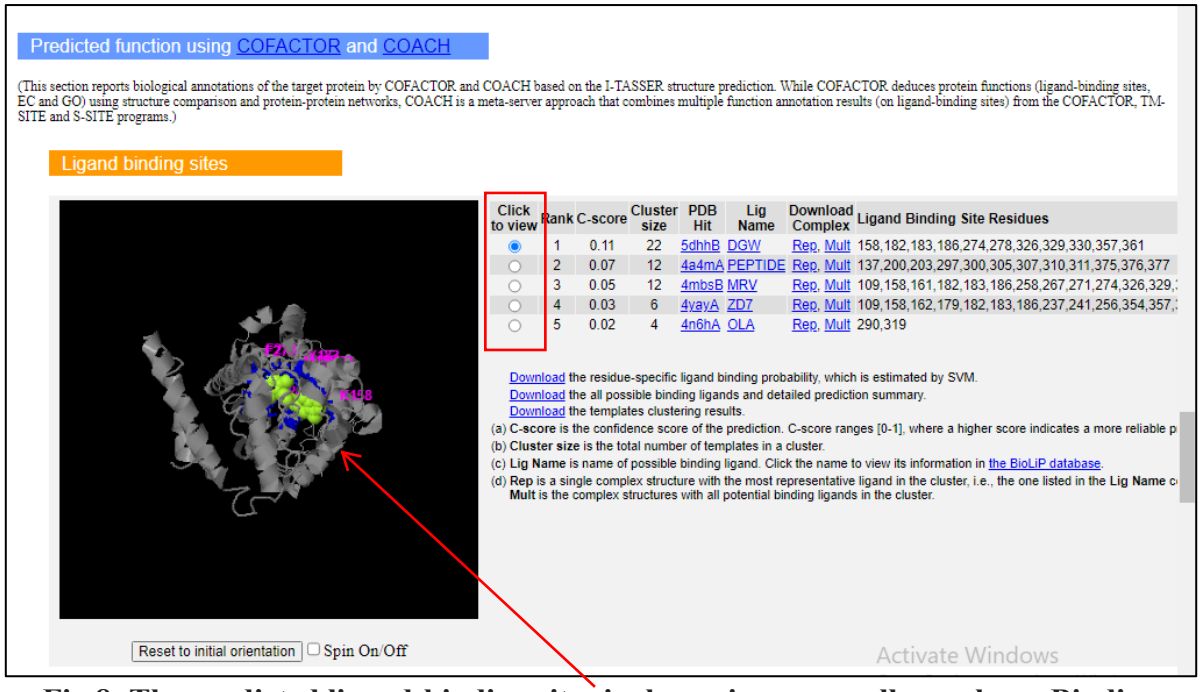


Fig 8: The predicted ligand-binding sites is shown in green-yellow sphere. Binding residues are shown in blue ball & sticks. By clicking radio button we can visualize different ligand-binding sites

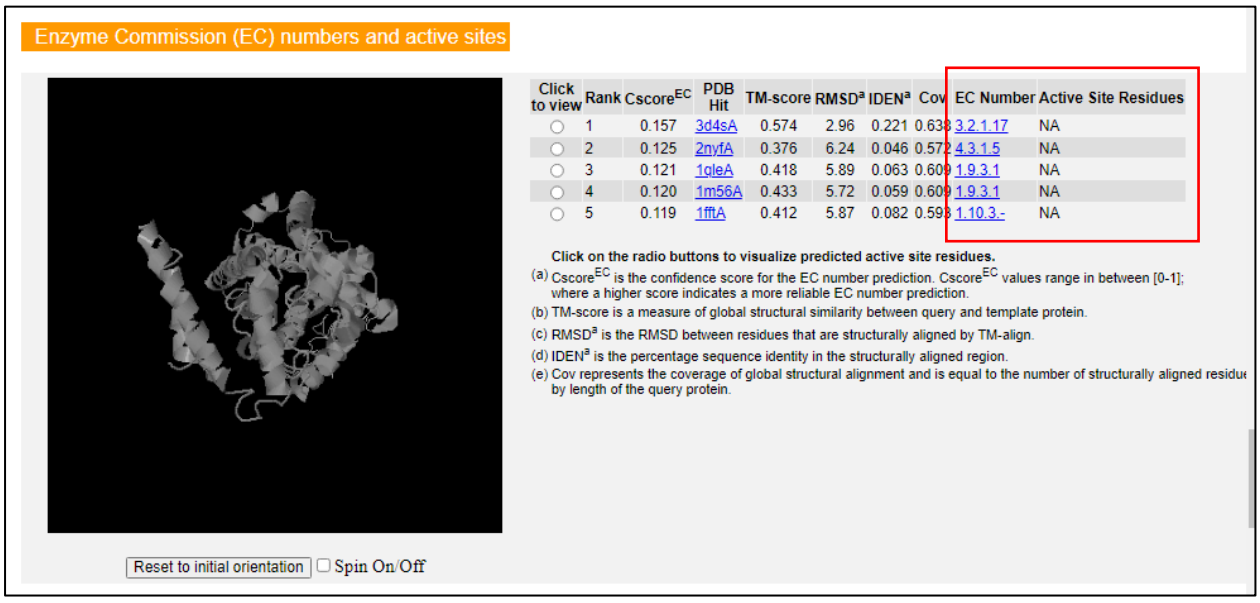
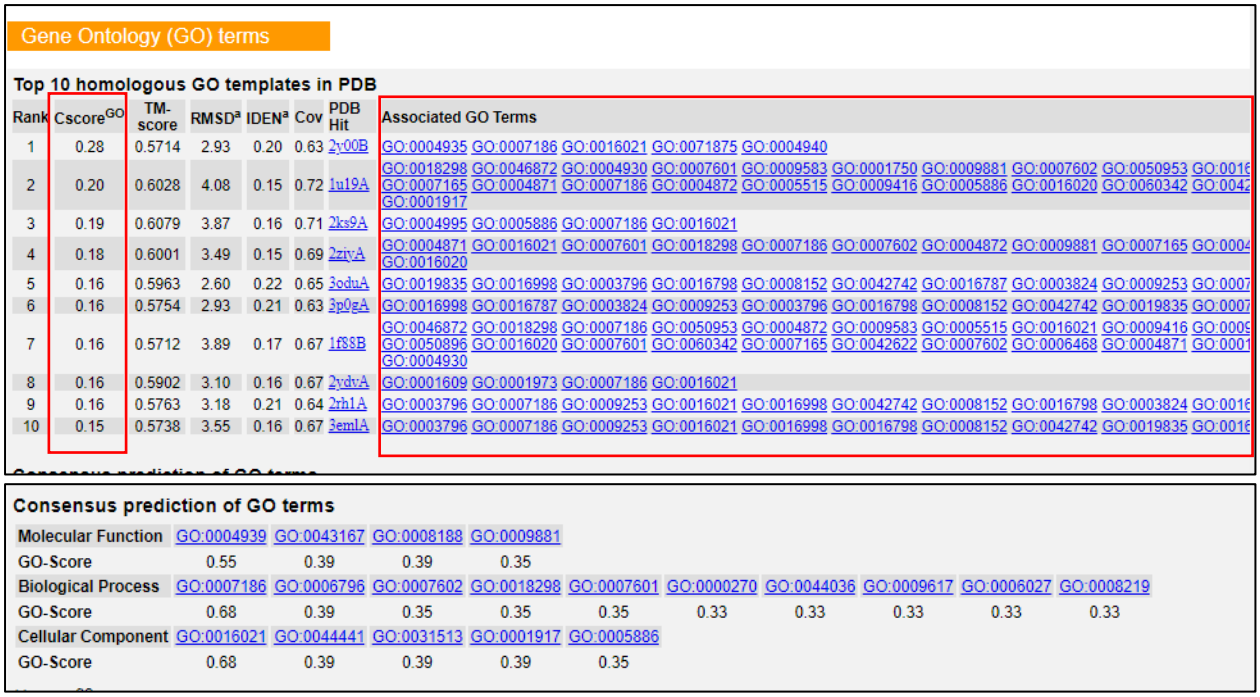


Fig 9: The predicted enzyme commission numbers and active sites results which is predicted catalytic site residues of query protein. By clicking radio button, we can visualize different active sites



Gene Ontology (GO) terms

Top 10 homologous GO templates in PDB

Rank	Cscore^{GO}	TM-score	RMSD^{a}	IDEN^{a}	Cov	PDB Hit	Associated GO Terms
1	0.28	0.5714	2.93	0.20	0.63	2y00B	GO:0004935 GO:0007186 GO:0016021 GO:0071875 GO:0004940
2	0.20	0.6028	4.08	0.15	0.72	1u19A	GO:0018298 GO:0046872 GO:0004930 GO:0007601 GO:0009583 GO:0001750 GO:0009881 GO:0007602 GO:0050953 GO:0016 GO:0007165 GO:0004871 GO:0007186 GO:0004872 GO:0005515 GO:0009416 GO:0005886 GO:0016020 GO:0060342 GO:0042 GO:0001917
3	0.19	0.6079	3.87	0.16	0.71	2ks9A	GO:0004995 GO:0005886 GO:0007186 GO:0016021
4	0.18	0.6001	3.49	0.15	0.69	2zjvA	GO:0004871 GO:0016021 GO:0007601 GO:0018298 GO:0007186 GO:0007602 GO:0004872 GO:0009881 GO:0007165 GO:0004 GO:0016020
5	0.16	0.5963	2.60	0.22	0.65	3oduA	GO:0019835 GO:0016998 GO:0003796 GO:0016798 GO:0008152 GO:0042742 GO:0016787 GO:0003824 GO:0009253 GO:0007
6	0.16	0.5754	2.93	0.21	0.63	3p0gA	GO:0016998 GO:0016787 GO:0003824 GO:0009253 GO:0003796 GO:0016798 GO:0008152 GO:0042742 GO:0019835 GO:0007
7	0.16	0.5712	3.89	0.17	0.67	1f88B	GO:0046872 GO:0018298 GO:0007186 GO:0050953 GO:0004872 GO:0009583 GO:0005515 GO:0016021 GO:0009416 GO:0009 GO:0050896 GO:0016020 GO:0007601 GO:0060342 GO:0007165 GO:0042622 GO:0007602 GO:0006468 GO:0004871 GO:0001 GO:0004930
8	0.16	0.5902	3.10	0.16	0.67	2ydvA	GO:0001609 GO:0001973 GO:0007186 GO:0016021
9	0.16	0.5763	3.18	0.21	0.64	2rh1A	GO:0003796 GO:0007186 GO:0009253 GO:0016021 GO:0016998 GO:0042742 GO:0008152 GO:0016798 GO:0003824 GO:0016
10	0.15	0.5738	3.55	0.16	0.67	3emlA	GO:0003796 GO:0007186 GO:0009253 GO:0016021 GO:0016998 GO:0016798 GO:0008152 GO:0042742 GO:0019835 GO:0016

Consensus prediction of GO terms

Molecular Function	GO:0004939	GO:0043167	GO:0008188	GO:0009881						
GO-Score	0.55	0.39	0.39	0.35						
Biological Process	GO:0007186	GO:0006796	GO:0007602	GO:0018298	GO:0007601	GO:0000270	GO:0044036	GO:0009617	GO:0006027	GO:0008219
GO-Score	0.68	0.39	0.35	0.35	0.35	0.33	0.33	0.33	0.33	0.33
Cellular Component	GO:0016021	GO:0044441	GO:0031513	GO:0001917	GO:0005886					
GO-Score	0.68	0.39	0.39	0.39	0.35					

Fig 10: The top 10 GO templates in PDB and consensus prediction of GO terms in the three function categories (1) Molecular function (2) Biological process (3) Cellular component of Gene Ontology. C-score measures the similarity between the query and template ranges between 0 to1 and the higher the better. Associated GO terms show the experimental gene ontology terms of the templates

RESULTS:

The 3D structure for query thrombin (UniProt ID: TR139059) was designed using I-TASSER (Iterative Threading ASSEmbly Refinement) tool. The predicted model shows C-score **-1.82** which indicates a model of correct global topology whereas the estimated TM score is **0.49 ± 0.15.** which is a standard metric used in protein structure analysis to measure the similarity between two protein structure which ranges from 0 to 1, with 1 indicating identical structures and lower values signifying less similarity. I-TASSER doesn't directly predict TM-score, but it estimates it based on the C-score & Estimated RMSD score is **11.2 ± 4.6A°** and it represents the average distance between corresponding atoms in the two structures, measured in Angstroms (A°) and structure is accepted for further studies. Lower RMSD value indicates closer match between predicted and actual structure. The structure needs further validation to carry ahead in drug designing studies.

CONCLUSION:

The 3D structure for query 'thrombin' (UniProt ID: TR139059) was designed using the I-TASSER tool.

REFERENCES:

1. Wei Zheng, Chengxin Zhang, Yang Li, Robin Pearce, Eric W. Bell, Yang Zhang. Folding non-homology proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Reports Methods, 1: 100014 (2021).
2. Chengxin Zhang, Peter L. Freddolino, and Yang Zhang. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Research, 45: W291-299 (2017)
3. Jianyi Yang, Yang Zhang. I-TASSER server: new development for protein structure and function predictions, Nucleic Acids Research, 43: W174-W181, 2015.

DATE: 31/01/2024

WEBLEM 10(C)
trROSETTA SERVER
(URL: https://yanglab.qd.sdu.edu.cn/trRosetta/)

AIM:

To predict protein three-dimensional conformations using the *Ab initio* method by exploring trRosetta server for query 'thrombin' (UniProt ID: TR139059).

INTRODUCTION:

The goal of protein structure prediction is to determine the spatial location of every atom in a protein from its primary sequence. Depending on whether reliable structural templates are available in the PDB, protein structure prediction methods have been divided into template-based modelling (TBM) and template-free (FM) approaches, the latter of which is also called *ab initio* modelling. For many years, TBM has been the most reliable method for modelling protein structures; however, its accuracy is essentially determined by the availability of close homologous templates and the quality of the query-template alignments. Conversely, *ab initio* methods are designed to use advanced energy functions and sampling techniques to improve the folding performance for proteins that lack homologous templates in the PDB. However, due to the inaccuracy in force field design and the limitations of conformational search engines, the performance of the physics-based FM methods for non-homologous targets has remained significantly worse than that of the TBM methods for targets with readily identifiable homologous templates.

Rosetta is a template-free method developed by the David Baker Lab, which assembles full-length structure based on fragments of 3–9 residues from PDB structures. Similar to template-based methods, these fragments are selected on the basis of local sequence similarity and the similarity between the known and predicted secondary structure. The assembly simulation is then conducted by Monte Carlo simulated annealing search strategy. QUARK is another excellent fragment-assembly method developed by the Yang Zhang Lab. The structural fragments used by QUARK range from 1 to 20 residues, with the assembly simulation conducted by the replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field. There are many other methods which are also based on fragment assembly, such as SCRATCH, PROFESY, FRAGFOLD, and so on. The key difference between these methods and the template-based methods lies in that they do not rely on any global structural template and require no homology or structure similarity between the target protein and the proteins where the fragments come from. Therefore, it can be more capable for template-free methods to model target of new folds. However, it is still a great challenge for template-free methods to model proteins with length > 150 residues because of the huge computing demand and low accuracy of force field. Contact map prediction based on co-evolution approach recently demonstrated promise in breaking up such length limit of *ab initio* structure folding.

Thrombin:

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. These actions of thrombin on cells have been implicated in tissue repair processes and in the pathogenesis of inflammatory and fibroproliferative disorders such as pulmonary fibrosis and atherosclerosis. Thrombin mediates its cellular effects by proteolytically activating cell surface receptors. Presently, two such receptors have been described and their roles in regulation of these functions are currently being investigated. The discovery of multiple thrombin receptors creates the possibility of selective receptor blockade of specific thrombin mediated events. New drugs with these actions should add to our current repertoire of thrombin inhibitors used to treat thrombotic diseases.

METHODOLOGY:

1. Open homepage of trRosetta.
2. Paste a protein sequence copied from UniProt in FASTA format of query 'thrombin' (UniProt ID: TR139059).
3. Select the input type i.e., query sequence of query 'thrombin' (UniProt ID: TR139059).
4. Submit and analyse the result.

OBSERVATIONS:

Fig 1: Homepage of trRosetta server

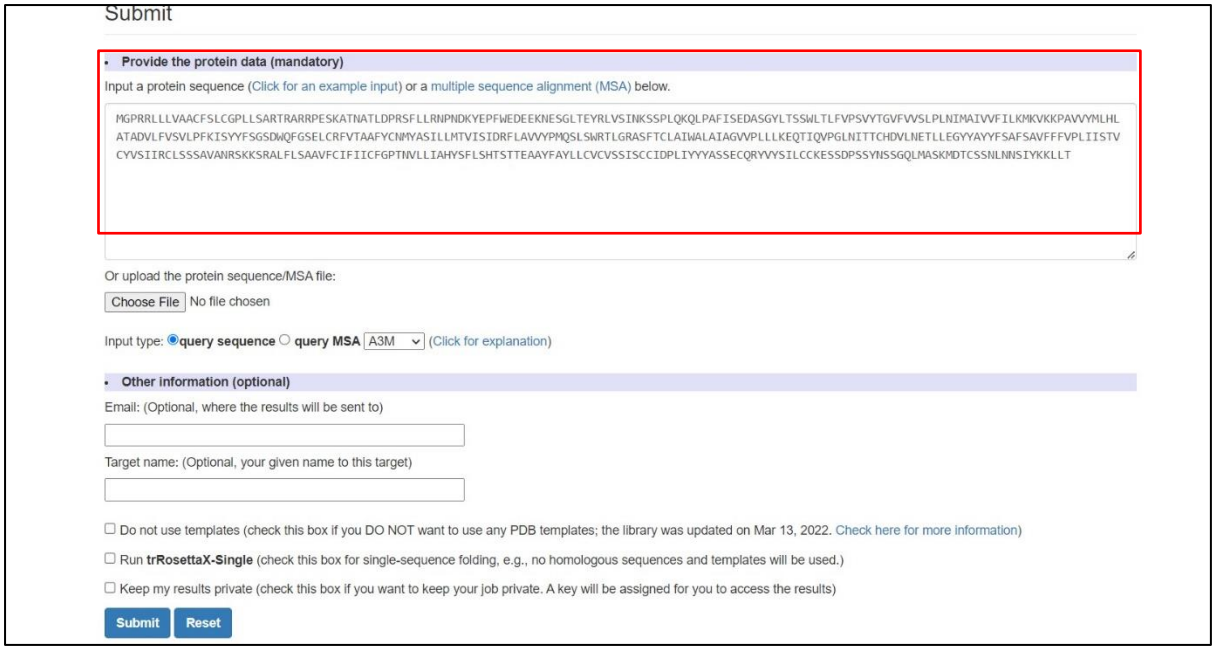


Fig 2: FASTA sequence is pasted for query 'thrombin' (UniProt ID: TR139059).

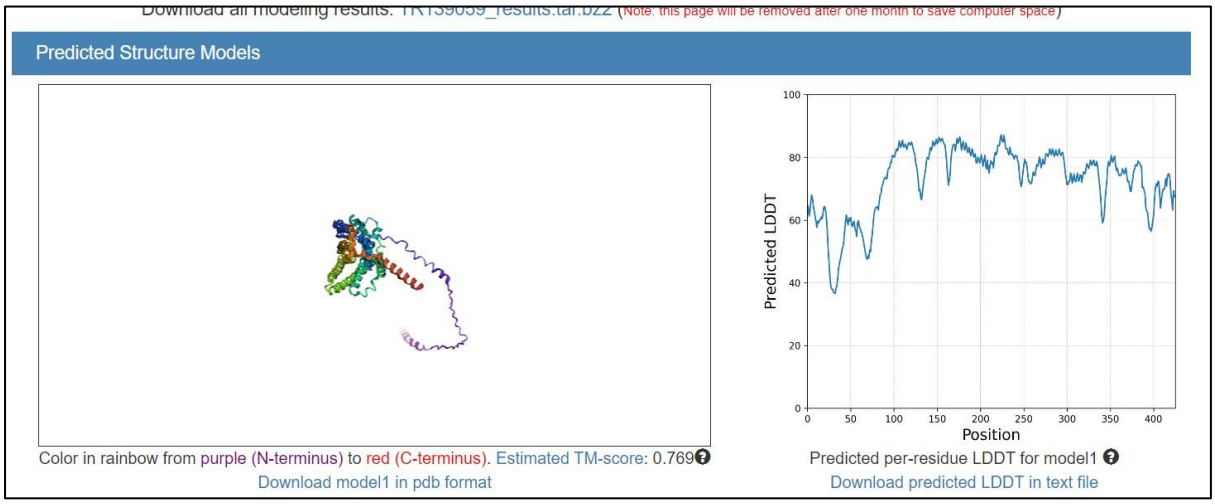


Fig 3: Results shown for predicted structure models

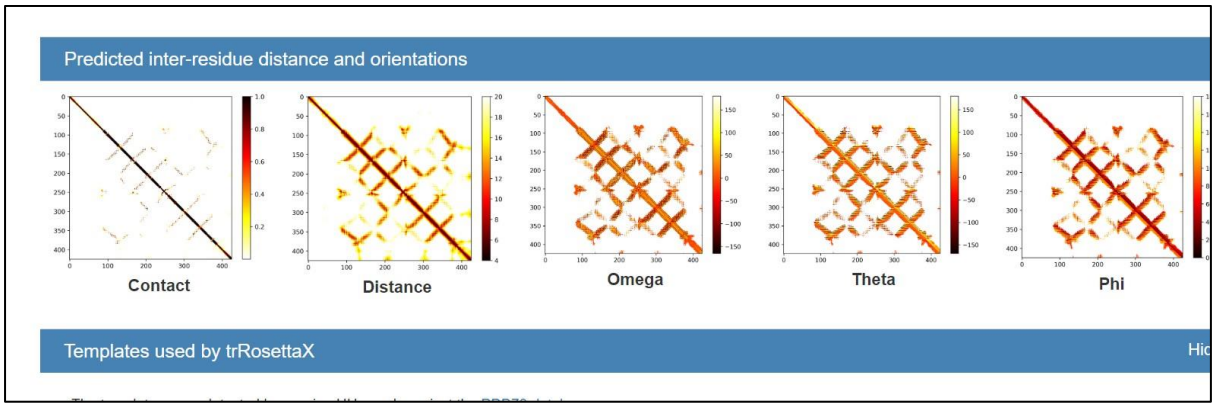


Fig 4: Results shown for predicted inter-residue distance orientations

RESULTS:

Using the ab initio method by exploring trRosetta tool for the query 'thrombin' (UniProt ID: TR139059). Following templates were found:

Template	Confidence	Coverage	Identity	E-value	Z-score
3VW7_A	100.0	66.4	65.9	2.6E-46	9.759
5T1A_A	100.0	66.1	17.9	4.3E-46	9.716
7F8U_A	100.0	64.2	18.8	6.3E-46	9.684
7F8Y_A	100.0	64.2	18.8	6.3E-46	9.684
6OSA_R	100.0	69.6	22.7	8.1E-46	9.663

The TM score was found to be 0.769 i.e., higher than 0.5 TM score usually indicates a model with correctly predicted topology.
Therefore, the query sequence is analogous to our template which means it can be used for the further studies.

CONCLUSION:

The protein prediction of three-dimensional conformations was studied using the Ab initio method by exploring trRosetta tool for query 'thrombin' (UniProt ID: TR139059).

REFERENCES:

1. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008;18(3):342–8. Epub 2008/04/26. doi: 10.1016/j.sbi.2008.02.004; PubMed Central PMCID: PMC2680823.
2. Dunbrack R, editor Template-based modeling assessment in CASP11. 11th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction; 2014; Riviera Maya, Mexico.
3. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Evaluation of free modeling targets in CASP11 and ROLL. Proteins. 2016;84 Suppl 1:51–66. doi: 10.1002/prot.24973.
4. Jones DT, Proteins Struct. Funct. Bioinform 5, 127 (2001).
5. Goldsack NR, Chambers RC, Dabbagh K, Laurent GJ. Thrombin. Int J Biochem Cell Biol. 1998 Jun;30(6):641-6. doi: 10.1016/s1357-2725(98)00011-9. PMID: 9695019.

DATE:09/02/2024

WEBLEM 11
SAVES SERVER
(URL: https://saves.mbi.ucla.edu/)

INTRODUCTION:

SAVES has been in operation since 2004 and is currently operating version 6. It is an interactive validation server for 5 programs commonly used in protein structure validation. Some of these tools were initially invented in the early 1990s in FORTRAN code and some have been upgraded to be in C++ code. Some of them have source code available. Validation is a process that determines whether the output of a software program conforms to user's expectations of the intended function. Protein structure prediction is a primary challenge in structural biology and is essential for gaining better insights into biological function. An understanding of three-dimensional structures is very crucial for rational drug design. Despite having strong methods like X-ray crystallography and NMR for protein 3D structure determination, the time and cost involved restrict their implementation to selective proteins. This led to the emergence of reliable and efficient computational methods to validate protein tertiary structures. The following are the 5 programs used for the prediction and validation of data.

1. **ERRAT:** ERRAT is a program for verifying protein structures determined by crystallography. ERRAT is a novel method that can detect incorrect regions of protein structures according to errors leading to random distributions of atoms, which can be distinguished from correct distributions 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). Generally speaking, the method is sensitive to smaller errors than 3-D Profile analysis but is more forgiving than Procheck. ERRAT is used for the backbone statistical analysis of protein model. It is a protein model evaluation tool that is sensitive to smaller errors than 3-D profile analysis. Overall score >= 80 is considered as more than a accurately predicted model.

2. **VERIFY 3-D:** This is a tool utilized to evaluate the correctness of a protein model by its 3D profile. 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. The interpretation of plots is done on the basis of green dots that represent a raw score (score for each amino acid residue) and blue dots representing (score for overall residues).

3. **PROVE:** PROVE (Protein volume evaluation) analyzes the packing in protein models by evaluating the regularity of the atom volume, defined by the atom's radius and the planes separating it from other atoms.

4. **PROCHECK:** PROCHECK is used to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures. Unusual regions highlighted by PROCHECK are not necessarily errors as such, but may be unusual features for which there is a reasonable explanation (eg. distortions due to ligand-binding in the protein's active site). Nevertheless, they are regions that should be checked carefully.

5. **WHATCHECK:** This is an efficient tool built upon C++ programming which evaluates the predicted models of proteins by checking stereochemical parameters of amino acids residues. WHATCHECK is derived from a protein verification tool from the WHATIF program, which does the extensive checking of many stereochemical parameters of the residues in the model. The different colors show the description of amino acid like the green color depicts that the amino acid is highly reliable validated, the yellow colors shows that the molecule is not much reliable and the red color indicates that the molecule is not validated.

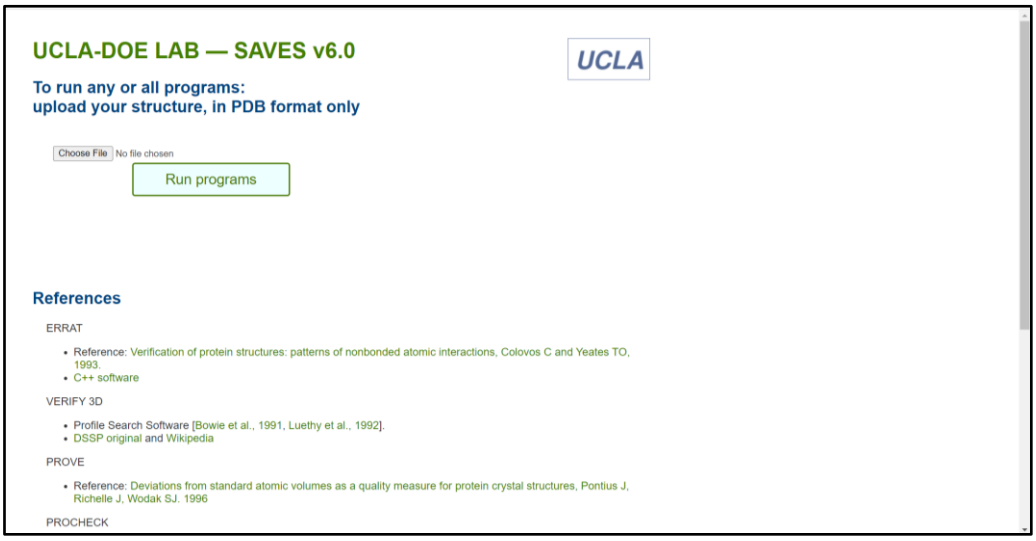


Fig 1: Homepage of SAVES server

REFERENCES:

1. Laskowski R A, Rullmannn J A, MacArthur M W, Kaptein R, Thornton J M (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR, 8, 477-486. [PubMed id: 9008363]
2. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 1993 Sep;2(9):1511-9. doi: 10.1002/pro.5560020916. PMID: 8401235; PMCID: PMC2142462.
3. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 1991 Jul 12;253(5016):164-70. doi: 10.1126/science.1853201. PMID: 1853201.
4. Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol. 1996 Nov 22;264(1):121-36. doi: 10.1006/jmbi.1996.0628. PMID: 8950272.

DATE: 09/02/2024

WEBLEM 11(A)
SAVES SERVER
(URL: https://saves.mbi.ucla.edu/)

AIM:

To validate structure design from Modeller, I-TASSER and T-ROSSETTA using SAVES server.

INTRODUCTION:

Validation is a process that determines whether the output of a software program conforms to user's expectations of the intended function. Protein structure prediction is a primary challenge in structural biology and is essential for gaining better insights into biological function. An understanding of three-dimensional structures is very crucial for rational drug design. Despite having strong methods like X-ray crystallography and NMR for protein 3D structure determination, the time and cost involved restrict their implementation to selective proteins. This led to the emergence of reliable and efficient computational methods to validate protein tertiary structures. The validation has three aspects:

1. checking on the validity of the thousands to millions of measurements of the structure
2. checking how consistent the atomic model is with those experimental data
3. checking consistency of the model with known physical and chemical properties.

The following are the tools used for the different algorithms:

Sr. No.	Algorithm	Tool Used
1.	Homology – based	Modeller (qseq.B99990001.pdb)
2.	Threading	I-TASSER(Model1.pdb)
3.	Ab - initio	T-ROSSETTA (2y00.pdb)

METHODOLOGY:

1. Open SAVES server.
2. Validate the structure from the above mentioned programs using SAVES server.

OBSERVATIONS:

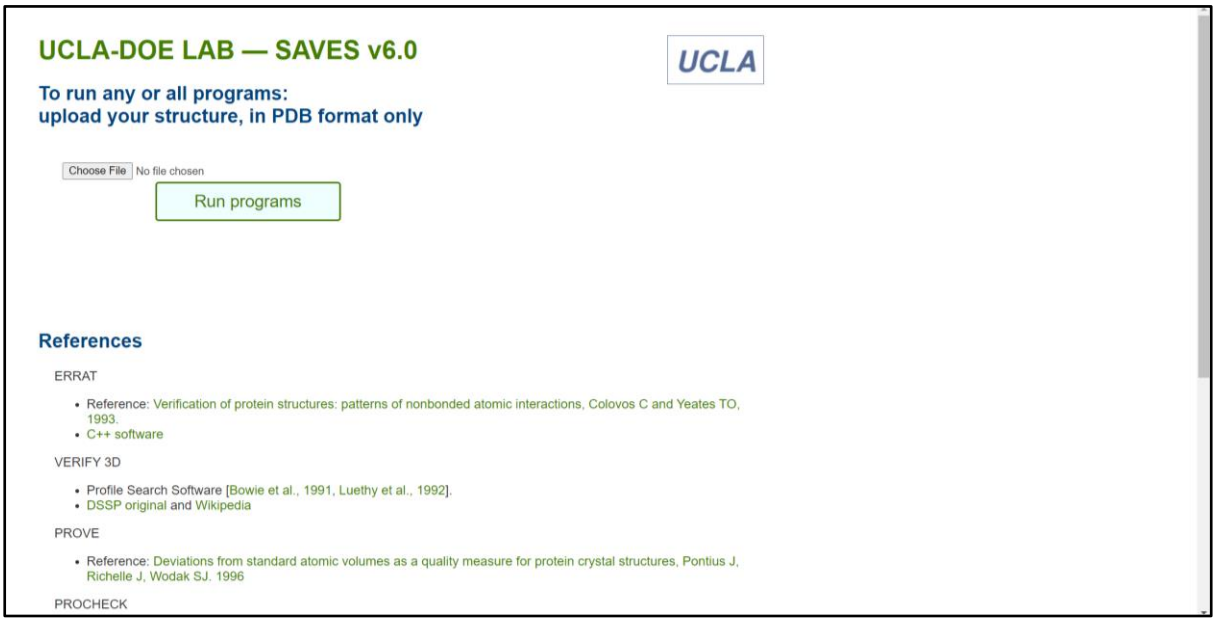


Fig 1: Homepage of SAVES server

Modeller: Homology method

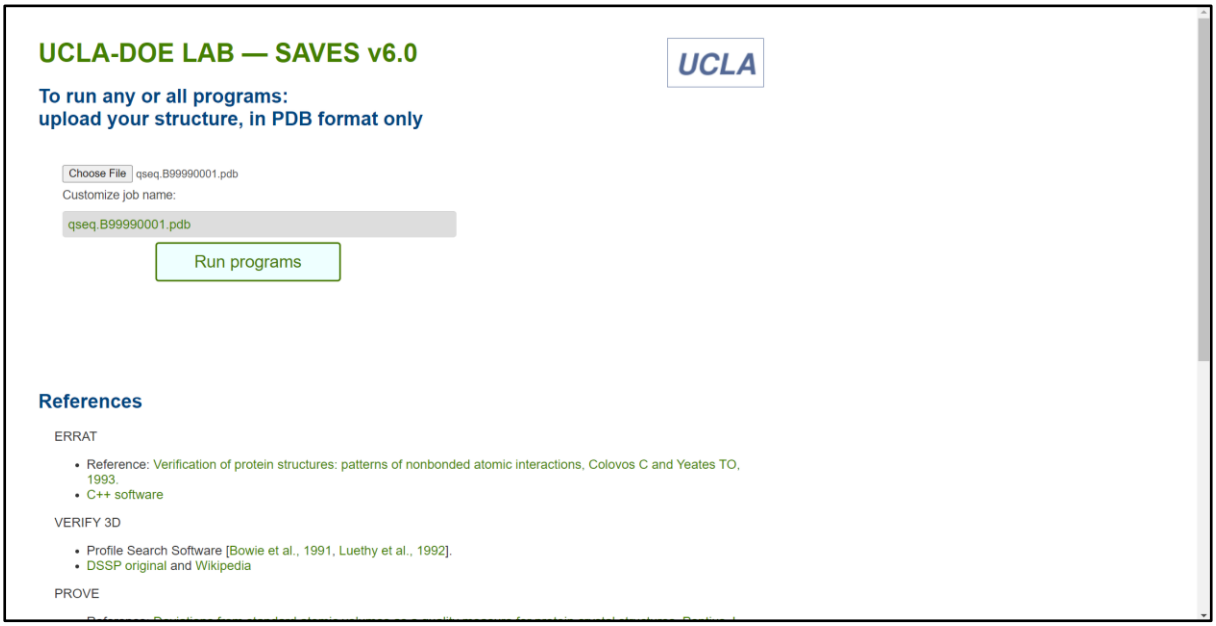


Fig 1.1: Structure model 'qseq.B99990001.pdb' obtained from Modeller tool was uploaded in SAVES server

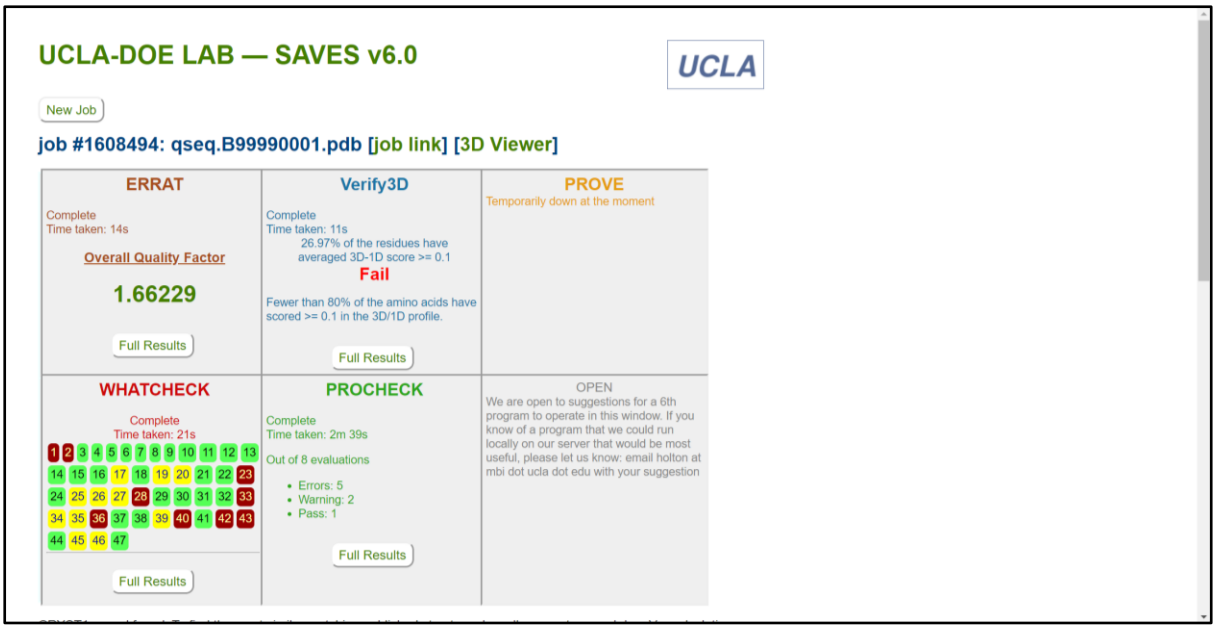


Fig 1.2: Results obtained after running the program

Fig 1.3: Results of ERRAT Tool

Fig 1.4: Results of VERIFY 3D Tool

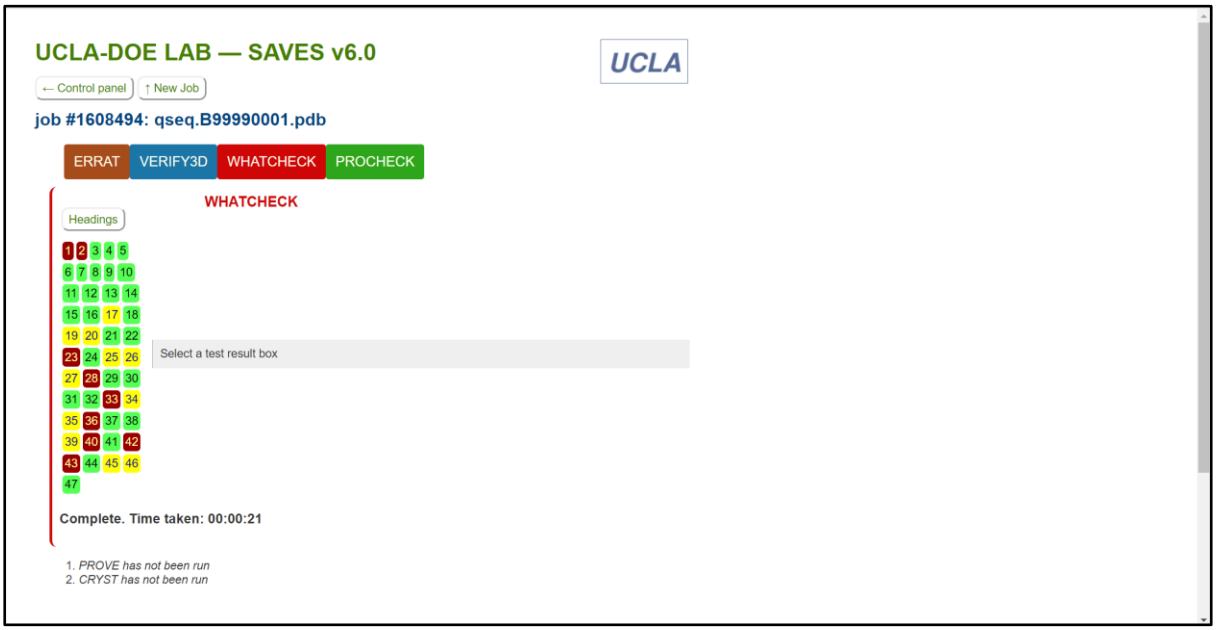


Fig 1.5: Results of WHATCHECK Tool

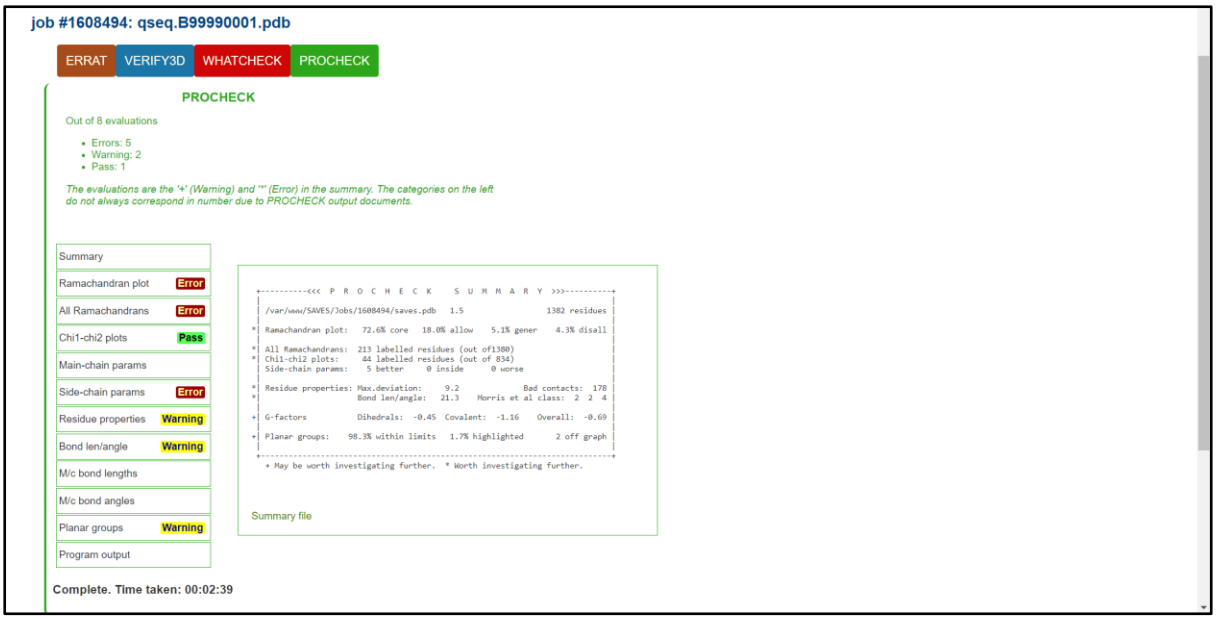


Fig 1.6: Results of PROCHECK Tool

I-TASSER - Threading Method (Iterative Threading Assembly Refinement)

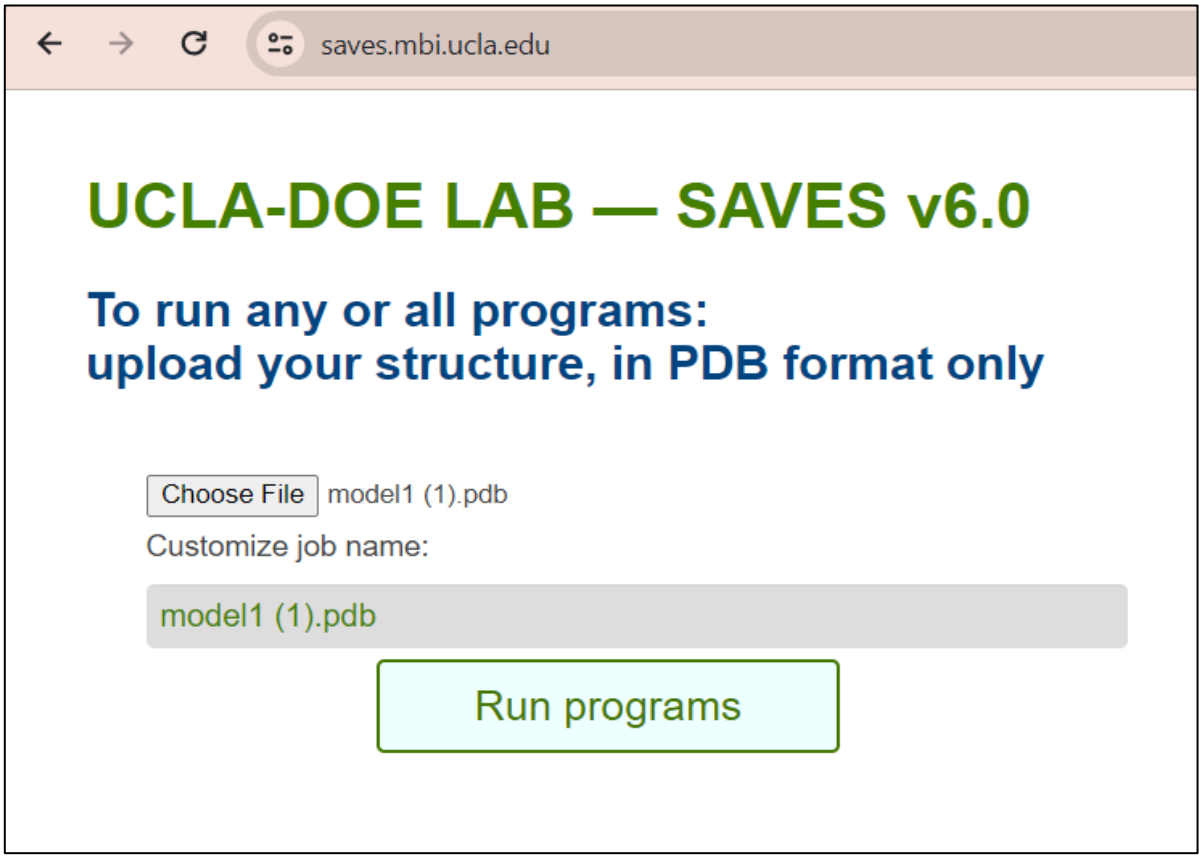


Fig 2: Structure model 'Model1.pdb' obtained from I-TASSER tool was uploaded in SAVES server

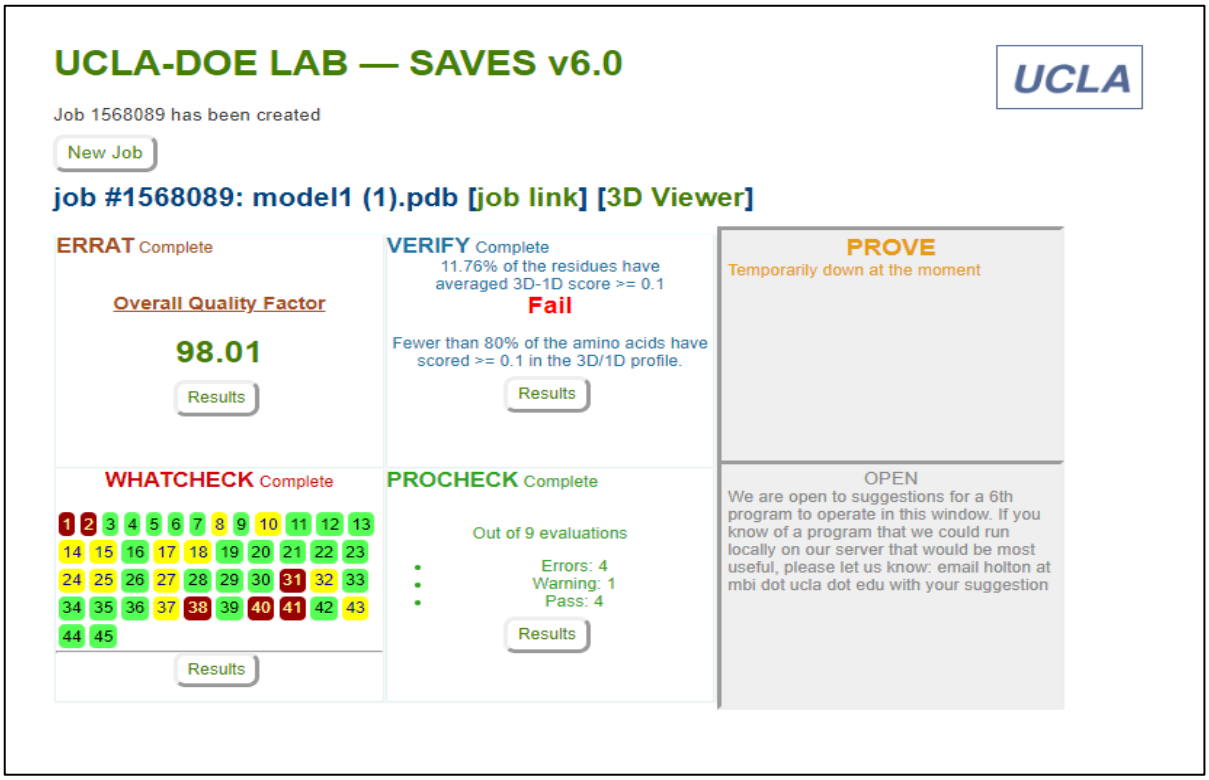


Fig 2.1: Results obtained after running the program

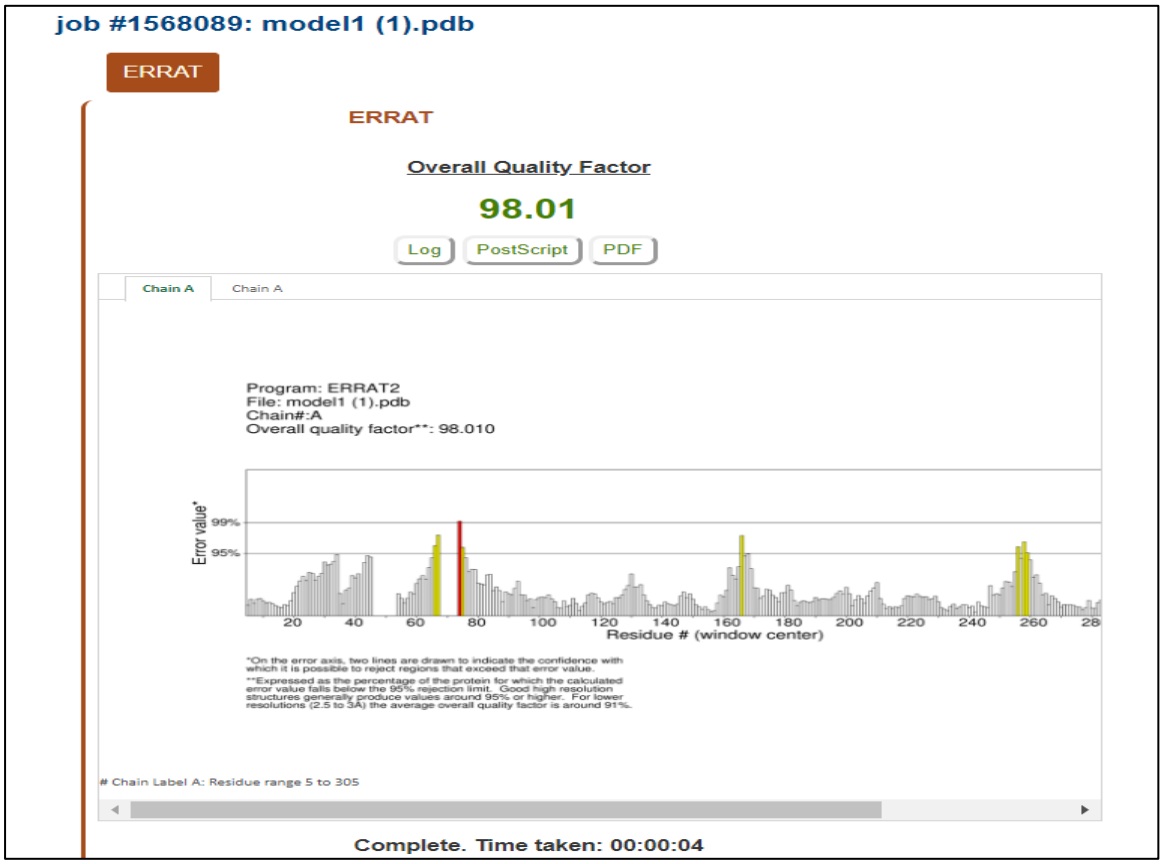


Fig 2.2: Result of ERRAT Tool

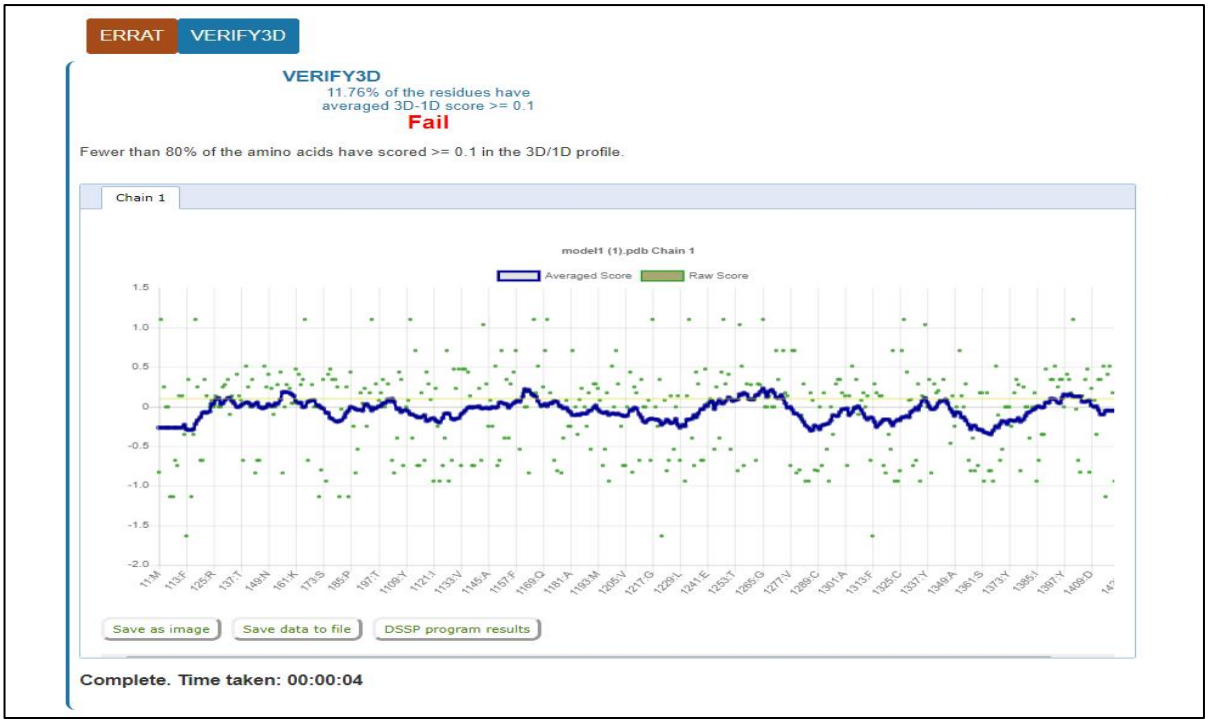


Fig 2.3: Result of VERIFY3D Tool

Fig 2.4: Result of WHATCHECK Tool

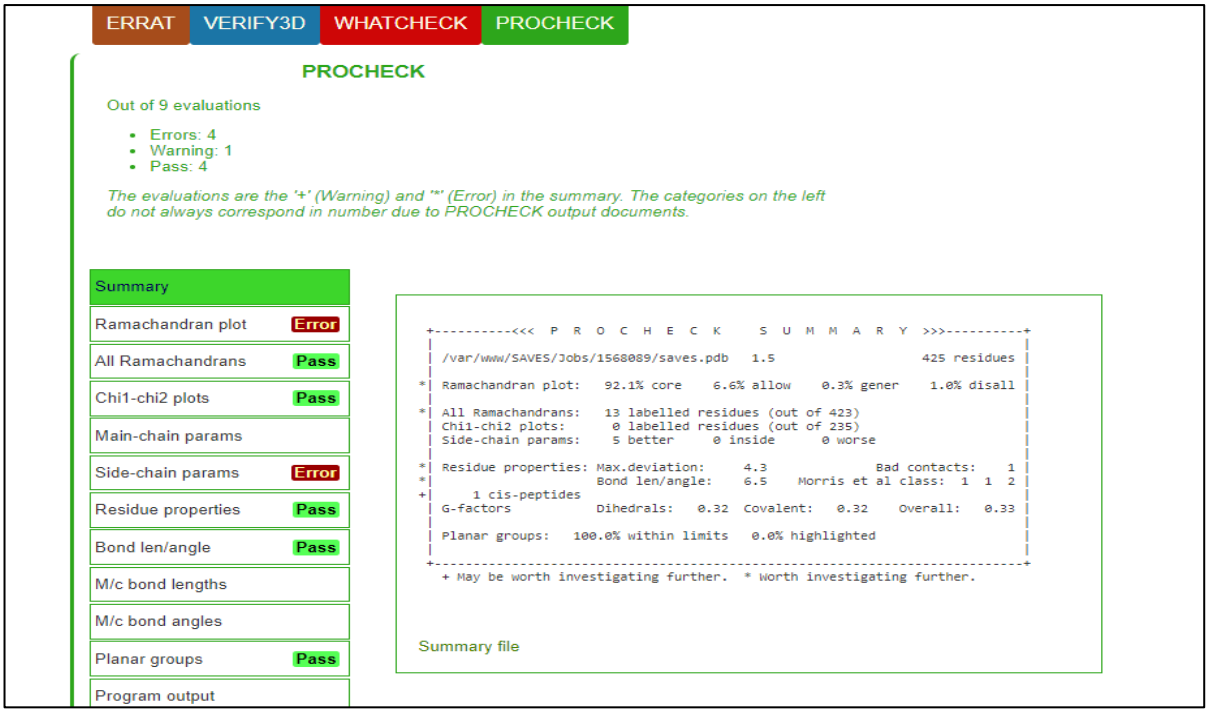


Fig 2.5: Result of PROCHECK Tool

T-ROSSETTA- Ab Initio methods

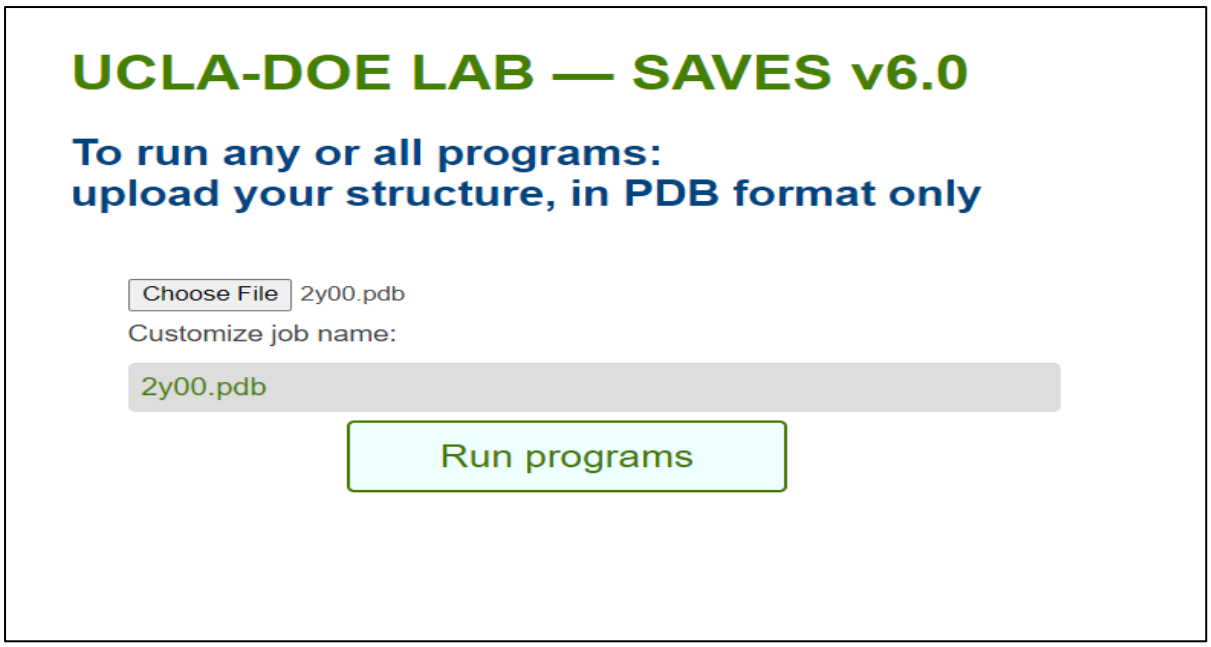


Fig 3: Structure model '2y00.pdb' obtained from T-ROSETTA tool was uploaded in SAVES server

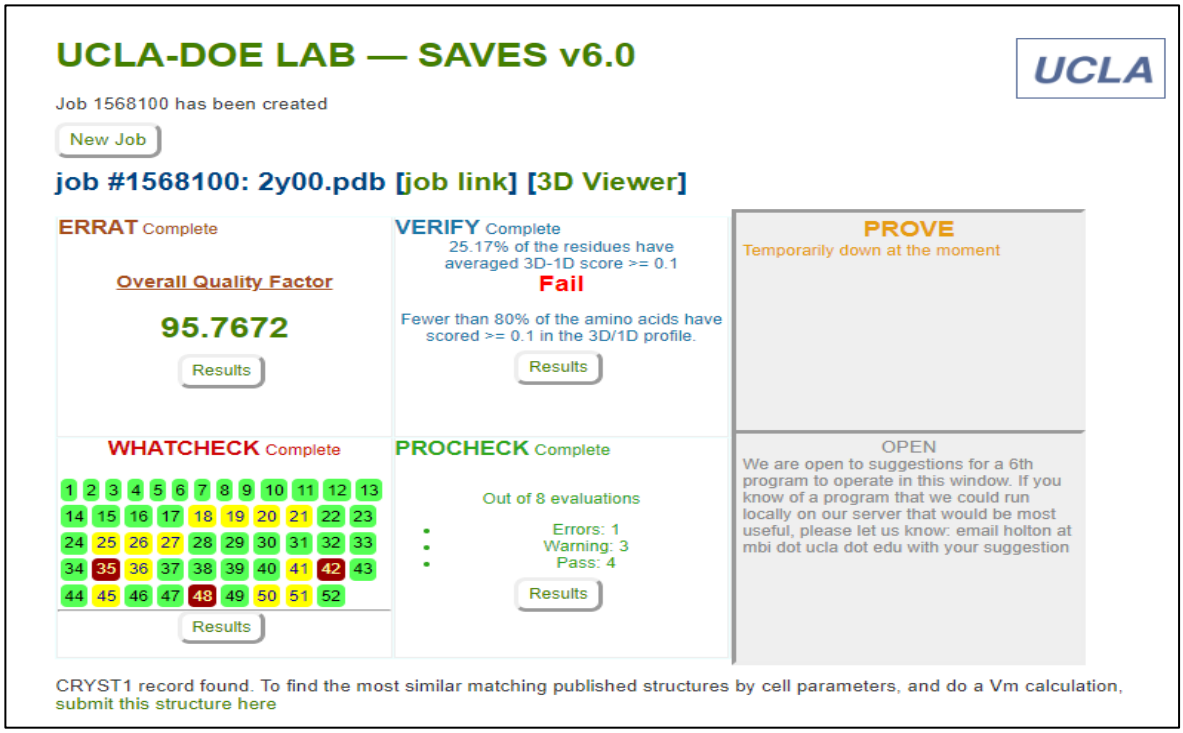


Fig 3.1: Results obtained after running the program

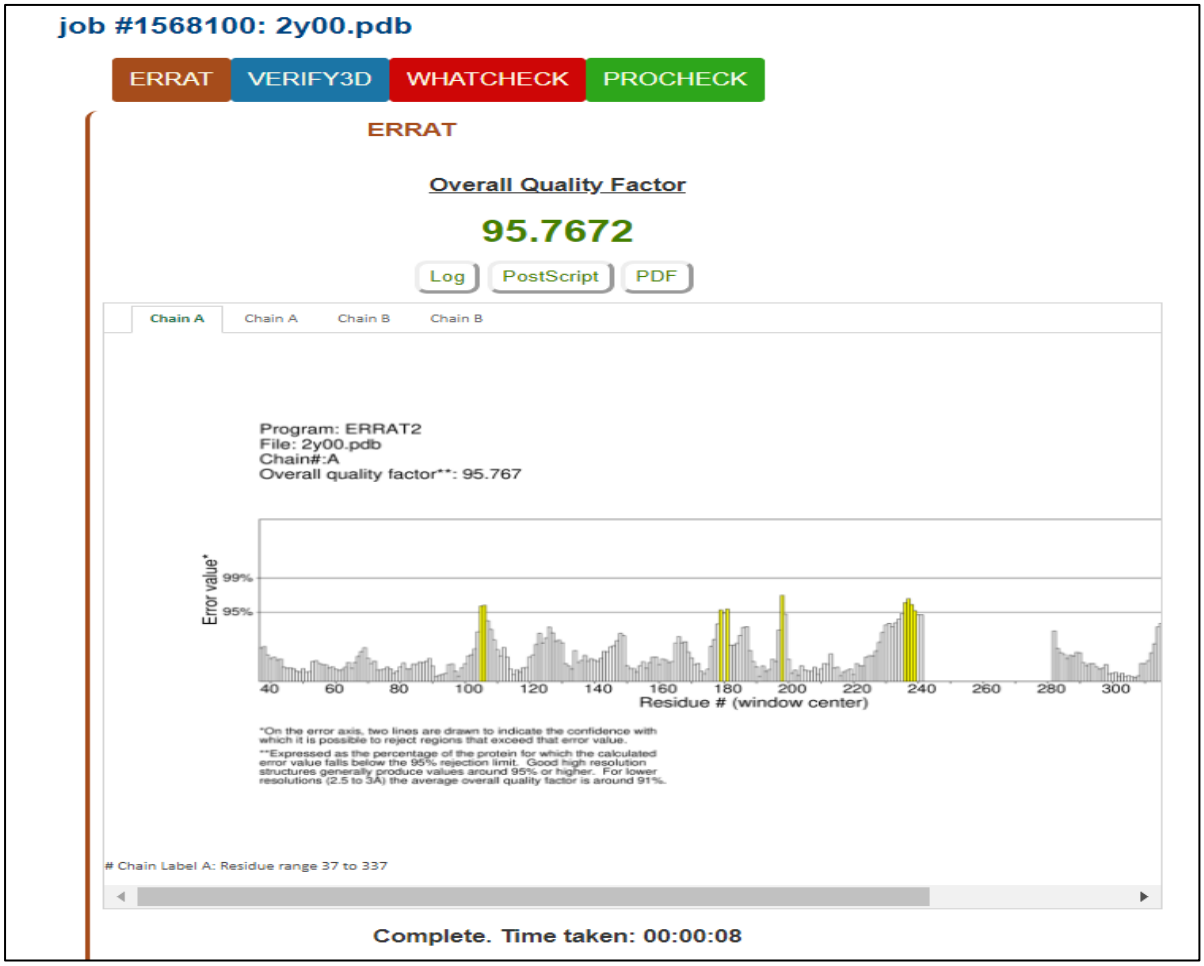


Fig 3.2: Result of ERRAT Tool

Fig3.3: Result of VERIFY3D Tool

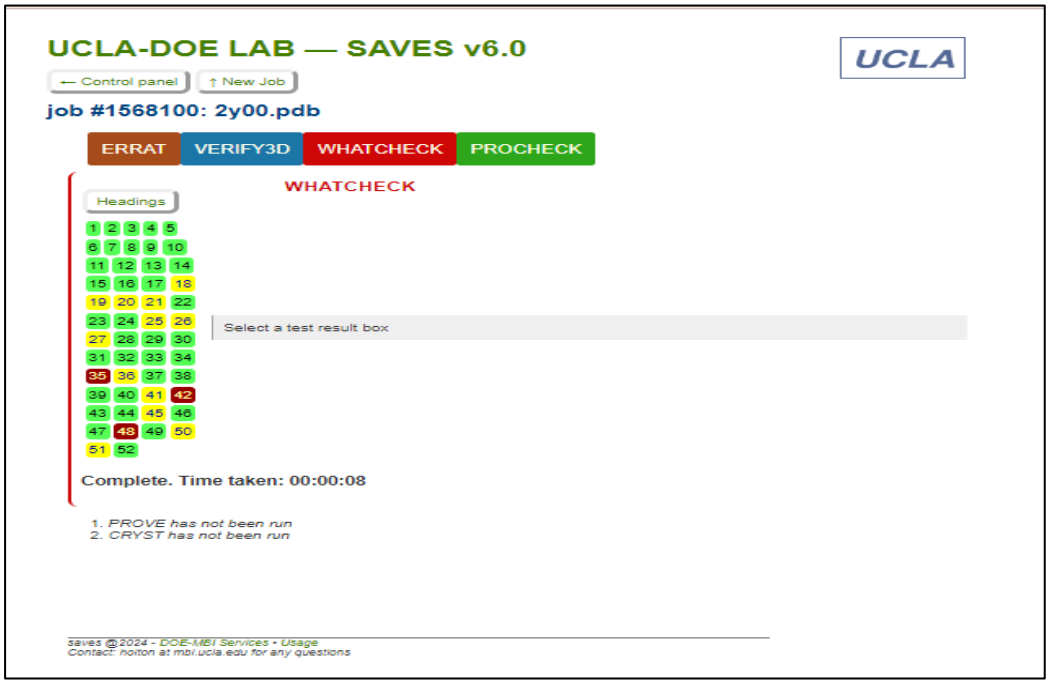


Fig 3.4:Result of WHATCHECK Tool

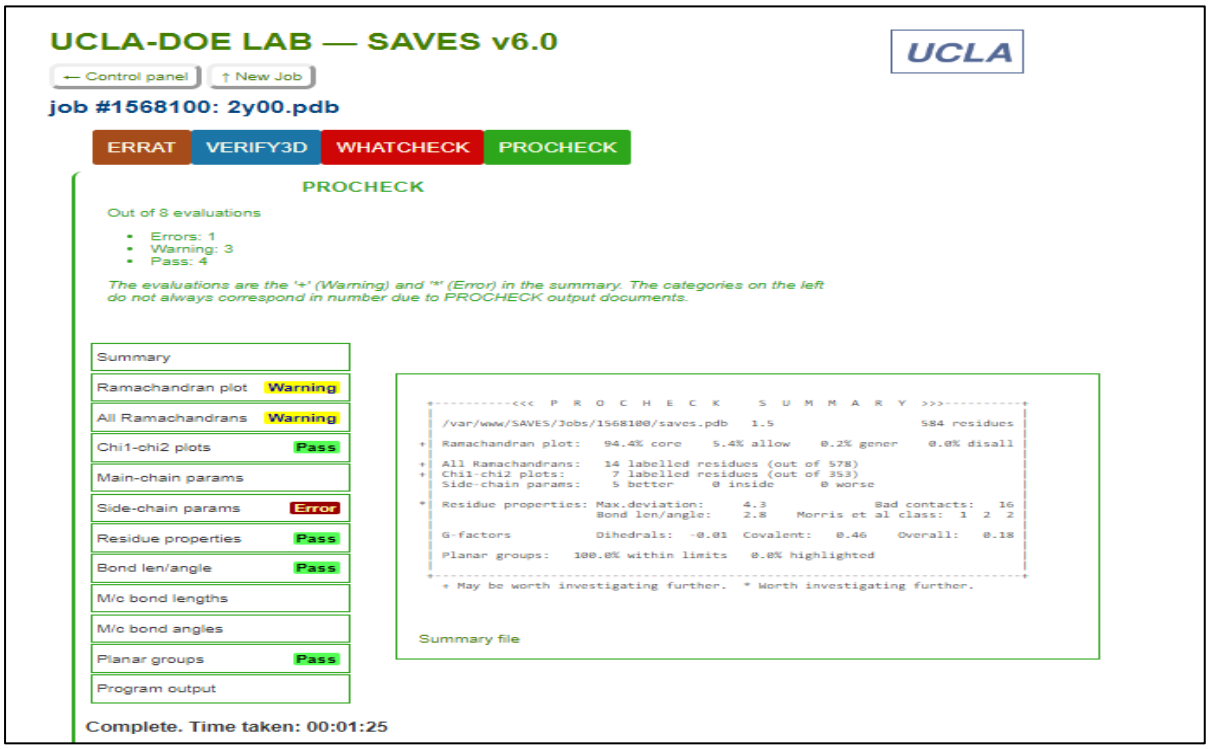


Fig 3.5: Result of PROCHECK Tool

RESULTS:

Saves tool was explored to validate tertiary protein structure. Saves is an Umbrella tool which holds various programs that validate model through four sub programs which include ERRAT, VERIFY-3D, WHATCHECK and PROCHECK. The models were obtained from various tool such as MODELLER, I-TASSER and T-ROSSETTA. Appropriate models were selected for validation using saves on the basis of various parameters. Three different modeling tools were utilized to generate models, but none produced appropriate results. Each tool evaluated the models in various ways. In the case of Modeller, the generated model achieved an ERRAT score of 1.66229 and was evaluated through eight evaluations in PROCHECK, identifying 5 errors. Although the Chi-Chi plot passed, errors were detected in the Ramachandran Plot which looks for residues in favored regions and outliers, and various errors were also detected by the WHATCHECK program. Similarly, I-TASSER generated a model with an ERRAT score of 98.01. PROCHECK identified four errors out of nine evaluations, while Chi1-Chi2 plot and bond len/angles both were passed. However, the Ramachandran Plot showed errors. T-ROSSETTA, on the other hand, generated a model with an ERRAT score of 95.7672. Only one error was found out of eight PROCHECK evaluations, with several criteria passing, but errors were detected in the Ramachandran Plot. Therefore, all three models were failed generated by different modelling tools, it seems that the protein structure is not homologous to the template which was validated by SAVES.

CONCLUSION:

The SAVES subprograms such as ERRAT, VERIFY3D, WHATCHECK and PROCHECK were used to validate tertiary protein structures. These tools helped to check the quality of models generated by various methods like Modeller, I-TASSER and T- ROSSETTA.

REFERENCES:

1. Chen, J., Xu, B., Yang, M., & Luo, X. (2015). ProTSAV: A protein tertiary structure analysis and validation server. Database (Oxford), 2015, bav045. https://pubmed.ncbi.nlm.nih.gov/26478257/
2. Wei Zheng, Chengxin Zhang, Yang Li, Robin Pearce, Eric W. Bell, Yang Zhang. Folding non-homology proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Reports Methods, 1: 100014 (2021).
3. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008;18(3):342–8. Epub 2008/04/26. doi: 10.1016/j.sbi.2008.02.004 ; PubMed Central PMCID: PMC2680823.
4. Frank, M., & Schloissnig, S. (2010). Bioinformatics and molecular modeling in glycobiology. Cellular and Molecular Life Sciences, 67(16), 2749–277 https://doi.org/10.1007/s00018-010-0352-4

DATE: 12/02/2024

WEBLEM 12
INTRODUCTION TO VISUALIZATION OF 3D PROTEIN STRUCTURE

Protein structure visualization tools are software applications that allow users to view, manipulate, and analyse the three-dimensional (3D) structures of proteins and other biomolecules. These tools are essential for understanding the molecular architecture, function, and interactions of proteins, as well as for designing new drugs, vaccines, and biotechnologies. 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 colours. Further manipulations can include changing the conformation of a structure by protein modelling 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.

1. 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.
2. 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.
3. 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.
4. 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.

Before computer visualization software was developed, molecular structures were presented by physical models of metal wires, rods and spheres. With the development of computer hardware and software technology and computer graphics programs were developed to visualizing and manipulating 3D structures. The computer graphics help to analyse and compare protein structure to gain the functions of protein.

Molecular visualization helps the scientists to bioengineer the protein molecules. User-friendly graphic interface makes this area of Bioinformatics a full filled, scientific thrill to the bio-scientists.

1. RasMol
2. Chimera

1. RasMol:

RasMol is a molecular graphics program intended for the visualisation of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images. The program has been developed at the University of Edinburgh's Biocomputing Research Unit and the Biomolecular Structure Department at Glaxo Research and Development, Greenford, UK. RasMol reads in molecular co-ordinate files in a number of formats and interactively displays the molecule on the screen in a variety of colour schemes and representations.

RasMol is a program for molecular graphics visualisation originally developed by Roger Sayle. This site is provided for the convenience of users of RasMol and developers of open-source versions of RasMol. The site itself is provided courtesy of Bernstein + Sons. Maintenance of RasMol, much of the development, and integration of modifications provided by the community is done at the ARCiB project at RIT. RasMol runs on wide range of architectures and operating systems including Microsoft Windows, Apple Macintosh, UNIX and VMS systems. UNIX and VMS versions require an 8-, 24- or 32-bit colour X Windows display (X11R4 or later). The X Windows version of RasMol provides optional support for a hardware dials box and accelerated shared memory communication (via the XInput and MIT-SHM extensions) if available on the current X Server. The program reads in a molecule coordinate file and interactively displays the molecule on the screen in a variety of colour schemes and molecule representations. Currently available representations include depth-cued wireframes, 'Dreiding' sticks, spacefilling (CPK) spheres, ball and stick, solid and strand biomolecular ribbons, atom labels and dot surfaces. The X Windows version of RasMol provides optional support for a hardware dials box and accelerated shared memory communication (via the XInput and MIT-SHM extensions) if available on the current X Serve.

The program reads in molecular coordinate files and interactively displays the molecule on the screen in a variety of representations and colour schemes. Supported input file formats include Protein Data Bank (PDB), Tripos Associates' Alchemy and Sybyl Mol2 formats, Molecular Design Limited's (MDL) Mol file format, Minnesota Supercomputer Center's (MSC) XYZ (XMol) format, CHARMm format, CIF format and mmCIF format files. If connectivity information is not contained in the file this is calculated automatically. The loaded molecule can be shown as wireframe bonds, cylinder 'Dreiding' stick bonds, alpha-carbon trace, space-filling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands), hydrogen bonding and dot surface representations. Atoms may also be labelled with arbitrary text strings. Alternate conformers and multiple NMR models may be specially coloured and identified in atom labels. Different parts of the molecule may be represented and coloured independently of the rest of the molecule or displayed in several representations simultaneously. The displayed molecule may be rotated, translated, zoomed and z-clipped (slabbed) interactively using either the mouse, the scroll bars, the command line or an attached dial box. RasMol can read a prepared list of commands from a 'script' file (or via inter-process communication) to allow a given image or viewpoint to be restored quickly. RasMol can also create a script file containing the commands required to regenerate the current image. Finally, the rendered image may be written out in a variety of formats including either raster or vector PostScript, GIF, PPM, BMP, PICT, Sun rasterfile or as a MolScript input script or Kinemage.

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.

STEPS FOR RASMOL INSTALLATION:

Download the latest Version

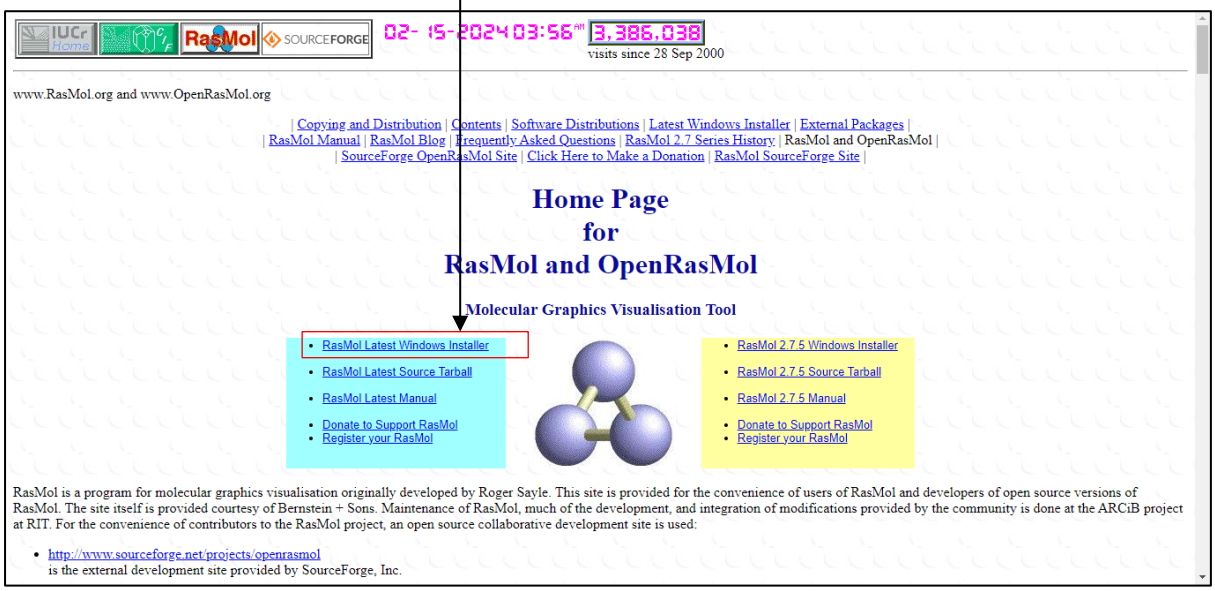


Fig 1: Homepage of RasMol tool

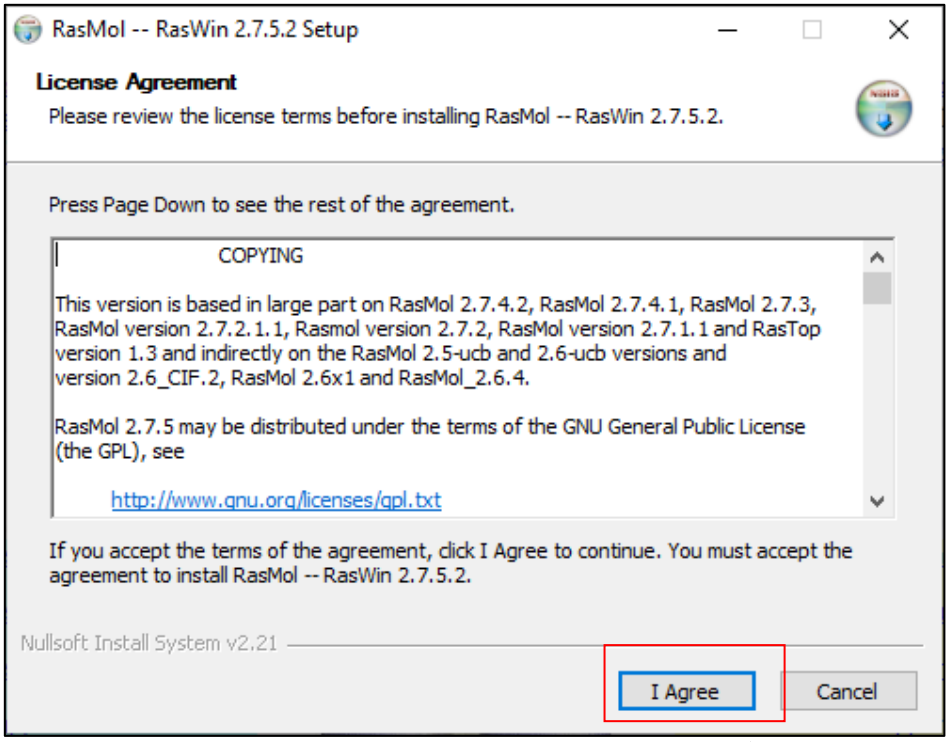


Fig 2: Accept License agreement

Click on I Agree

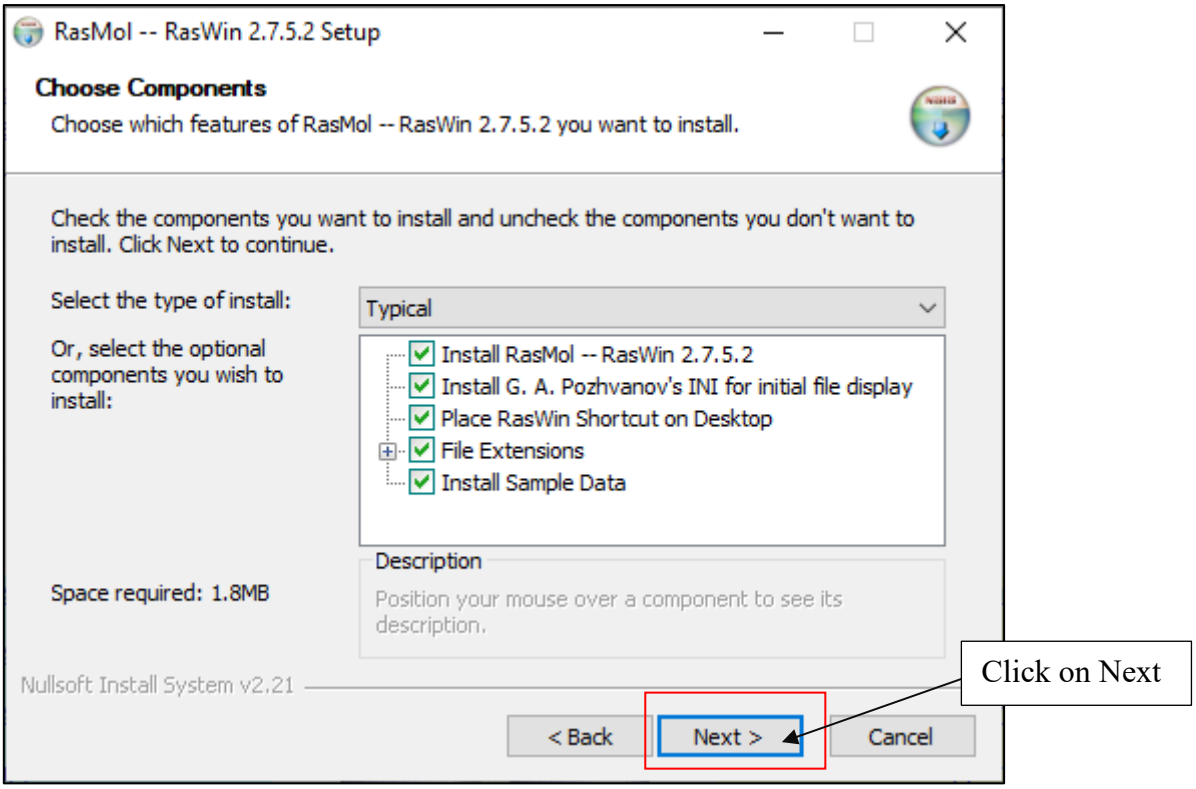


Fig 3: Choose the components as per required

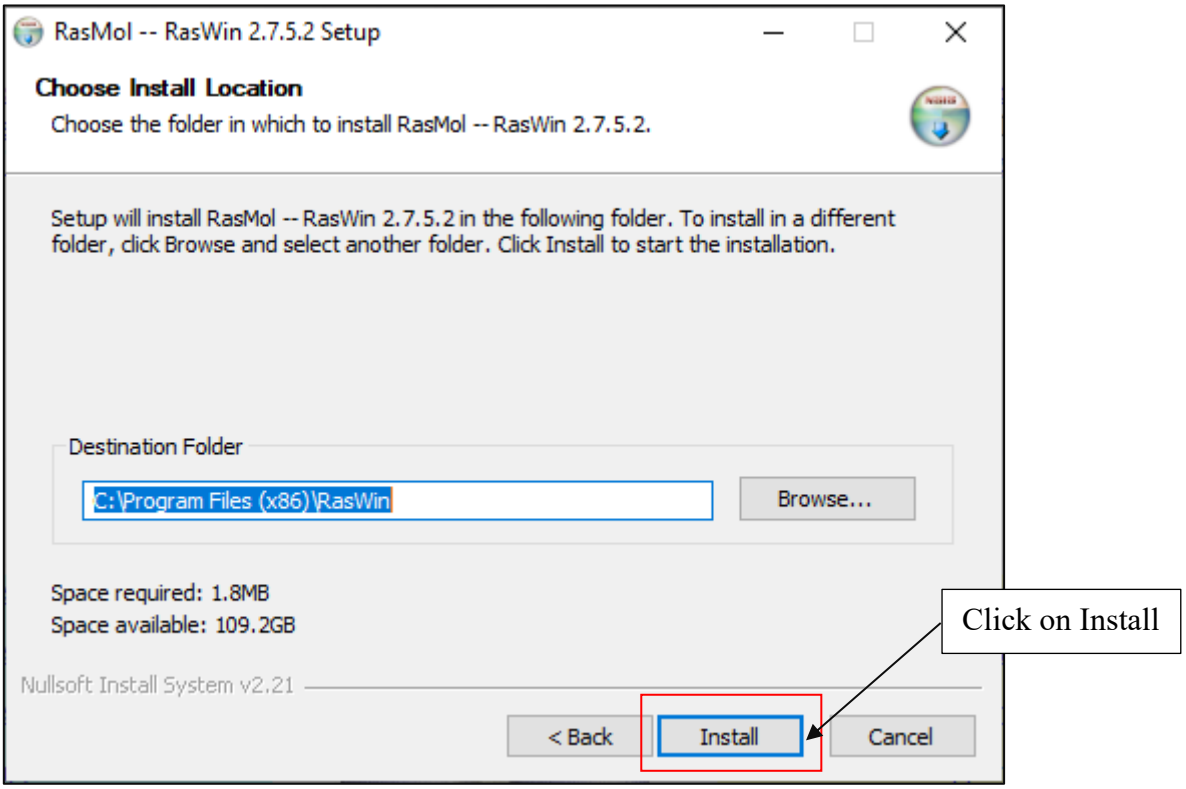


Fig 4: Click on Install to download the RasWin application

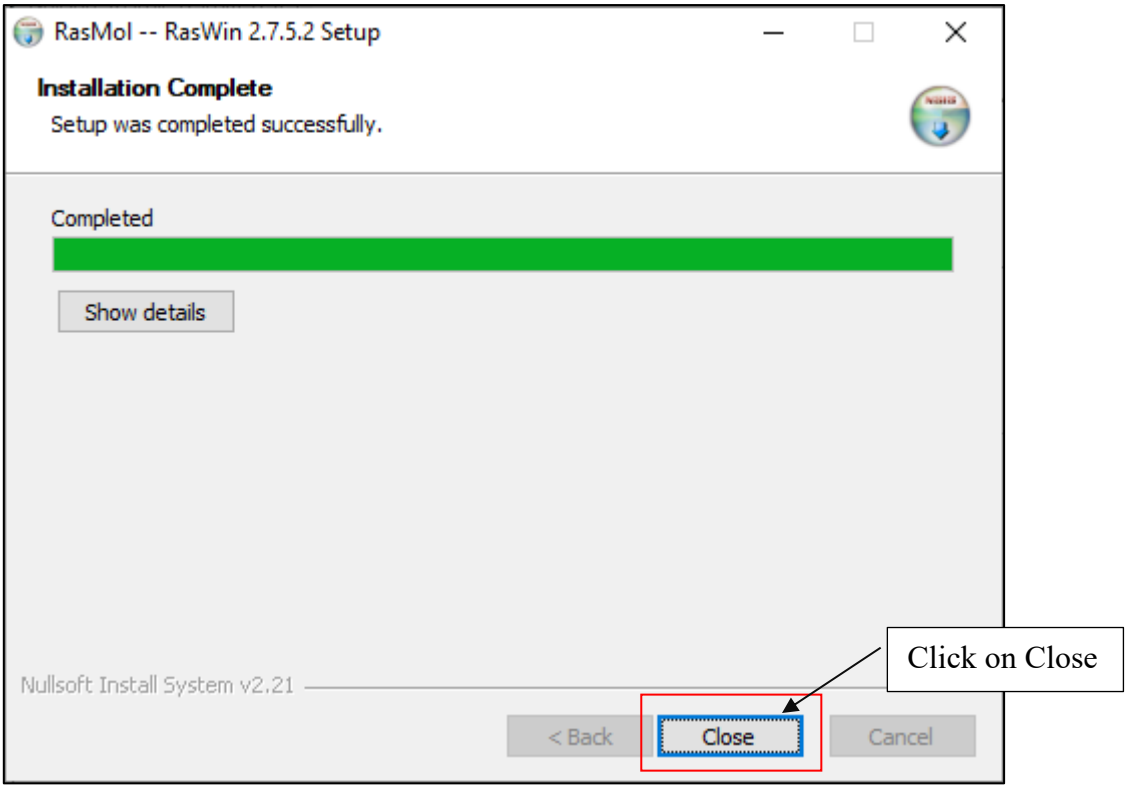


Fig 5: Installation Completed

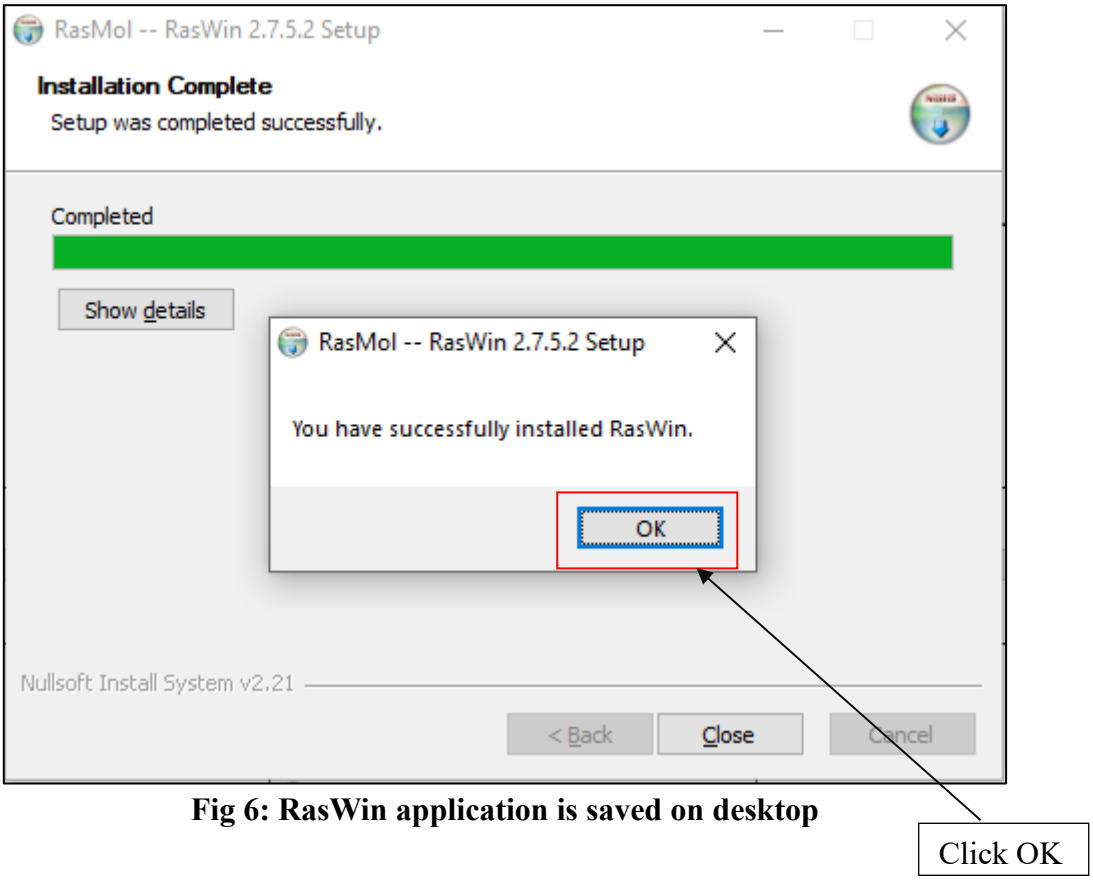


Fig 6: RasWin application is saved on desktop

Fig 7: Open RasWin application from desktop

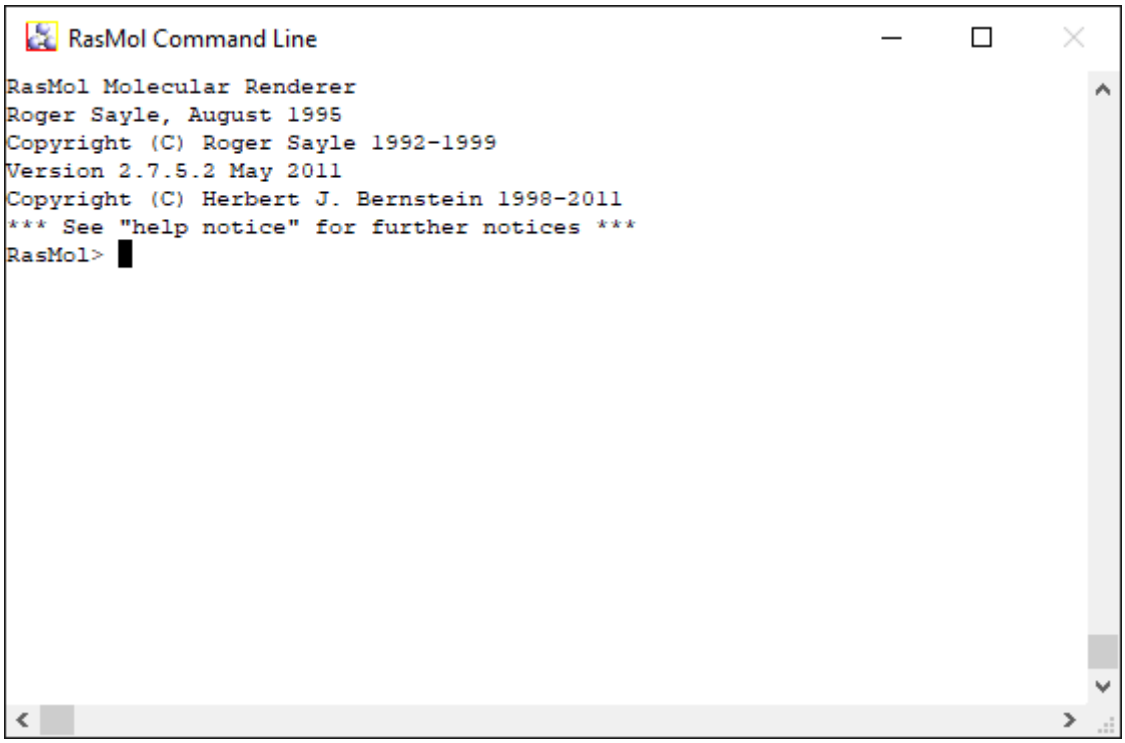


Fig 8: RasMol Command Line Tab

2. UCSF Chimera:

UCSF Chimera is a powerful visualization tool remarkably present in the computational chemistry and structural biology communities. Built on a C++ core wrapped under a Python 2.7 environment, one could expect to easily import UCSF Chimera's arsenal of resources in custom scripts or software projects. Nonetheless, this is not readily possible if the script is not executed within UCSF Chimera due to the isolation of the platform.

UCSF Chimera is a program for the interactive visualization and analysis of molecular structures and related data, including density maps, trajectories, and sequence alignments. It is available free of charge for non-commercial use.

UCSF Chimera is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. High-quality images and animations can be generated.

The molecular graphics program UCSF Chimera includes a suite of tools for interactive analyses of sequences and structures. Structures automatically associate with sequences in imported alignments, allowing many kinds of crosstalk. A novel method is provided to superimpose structures in the absence of a pre-existing sequence alignment. The method uses both sequence and secondary structure, and can match even structures with very low sequence identity. Another tool constructs structure-based sequence alignments from superpositions of two or more proteins. Chimera is designed to be extensible, and mechanisms for incorporating user-specific data without Chimera code development are also provided.

The tools described here apply to many problems involving comparison and analysis of protein structures and their sequences. Chimera includes complete documentation and is intended for use by a wide range of scientists, not just those in the computational disciplines. UCSF Chimera is free for non-commercial use and is available for Microsoft Windows, Apple Mac OS X, Linux, and other platforms from http://www.cgl.ucsf.edu/chimera.

Chime (www.mdlchime.com/chime/) is a plug-in for web browsers; it is not a standalone program and has to be invoked in a web browser. The program is also derived from RasMol and allows interactive display of graphics of protein structures inside a web browser.

STEPS FOR UCSF CHIMERA INSTALLATION:

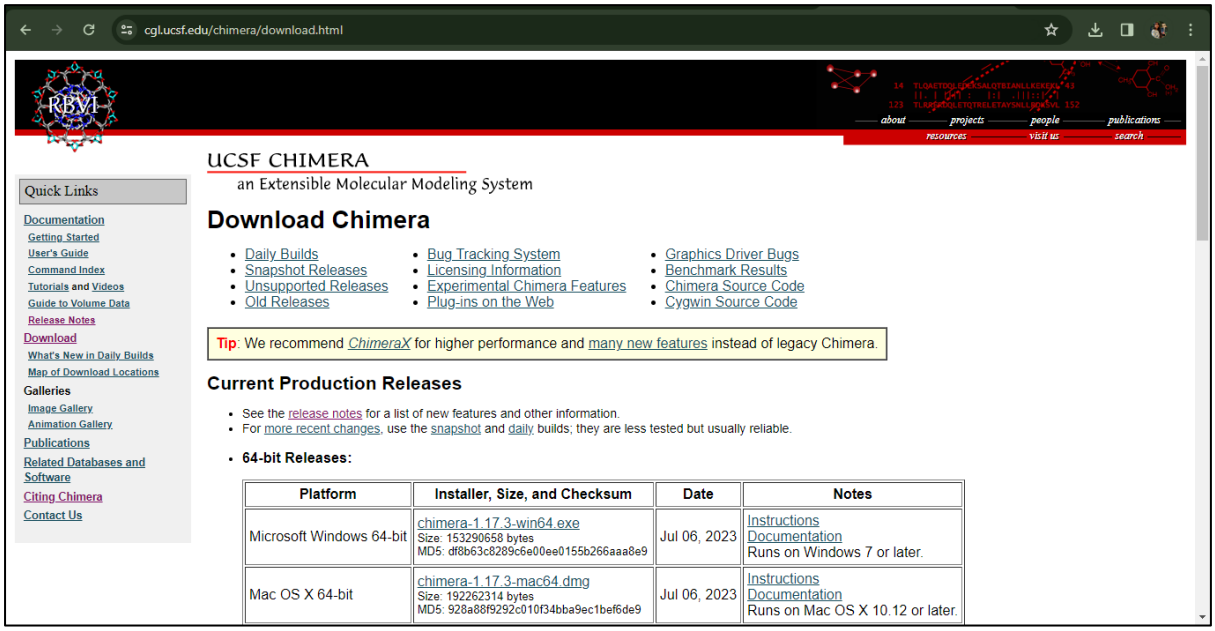


Fig 1: Homepage of UCSF Chimera.

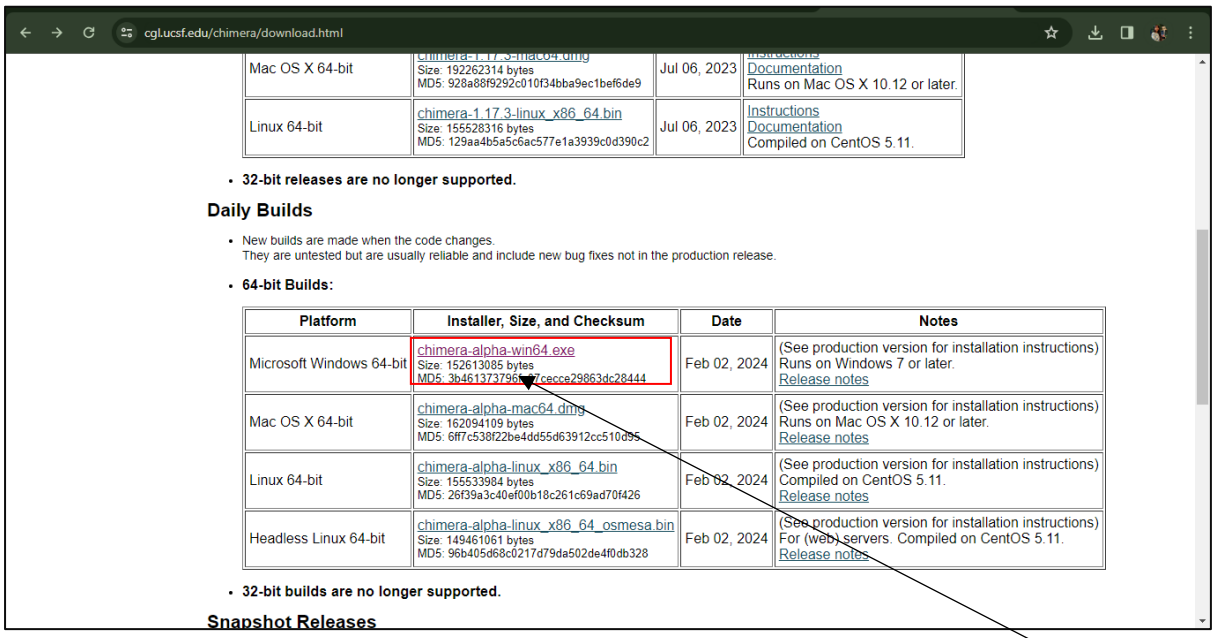


Fig 2: Download Microsoft windows 64-bit

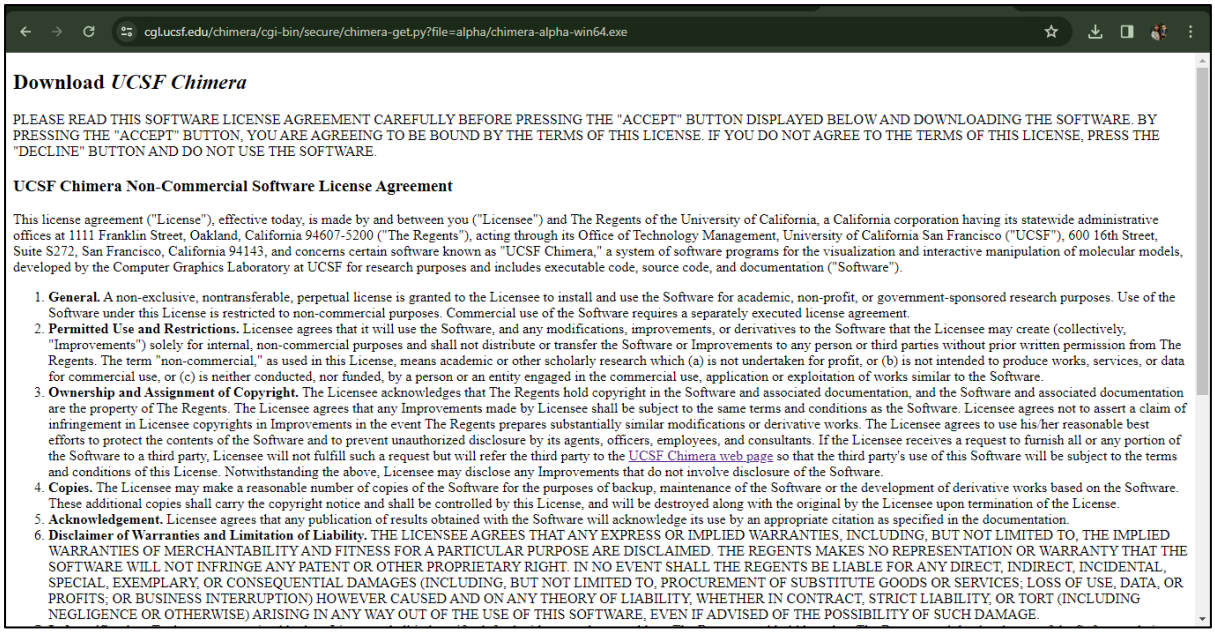


Fig 3.1: Download UCSF Chimera

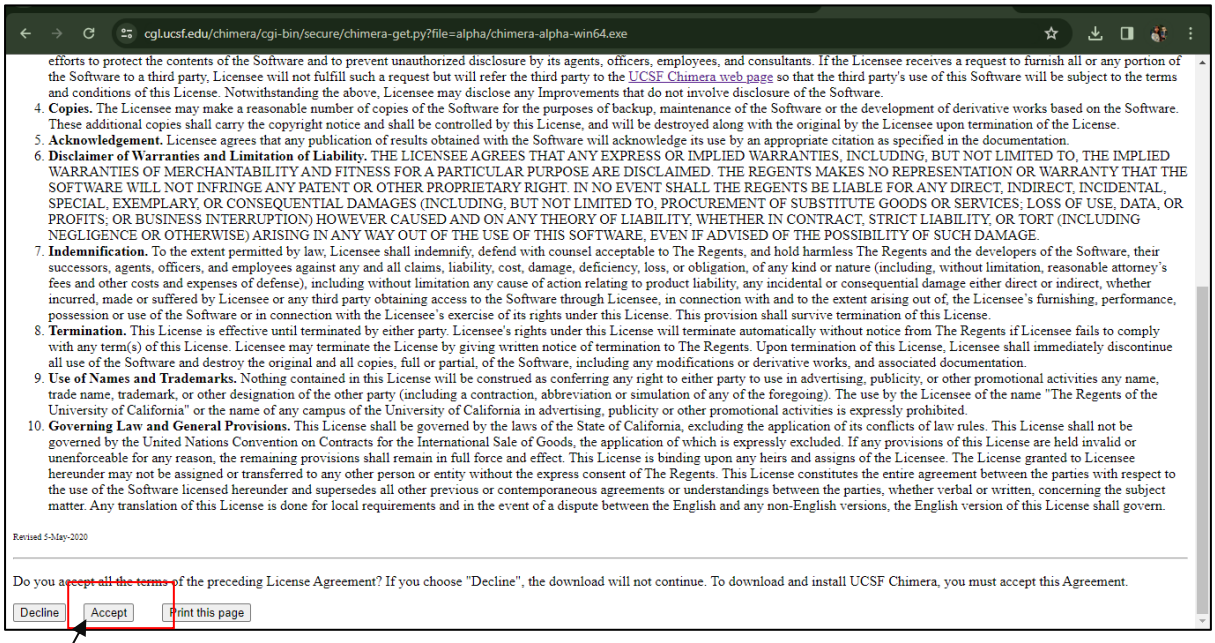


Fig 3.2: Accept terms and conditions to download UCSF Chimera

Click on Accept

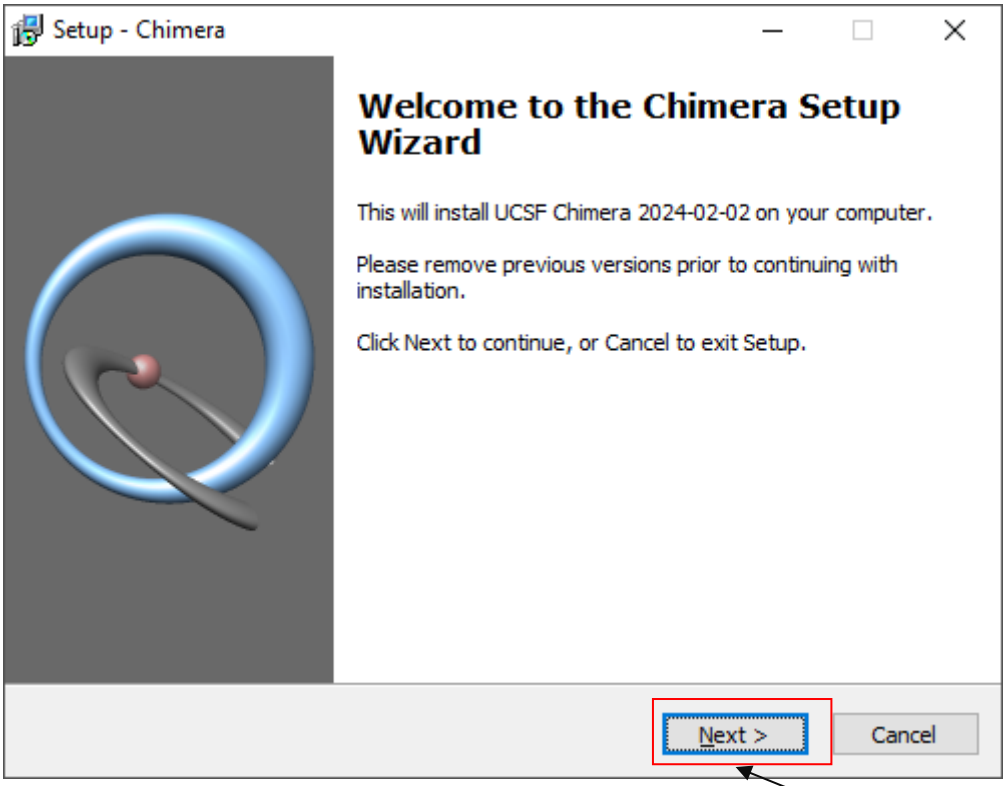


Fig 4: Setup to install UCSF Chimera

Click on Next

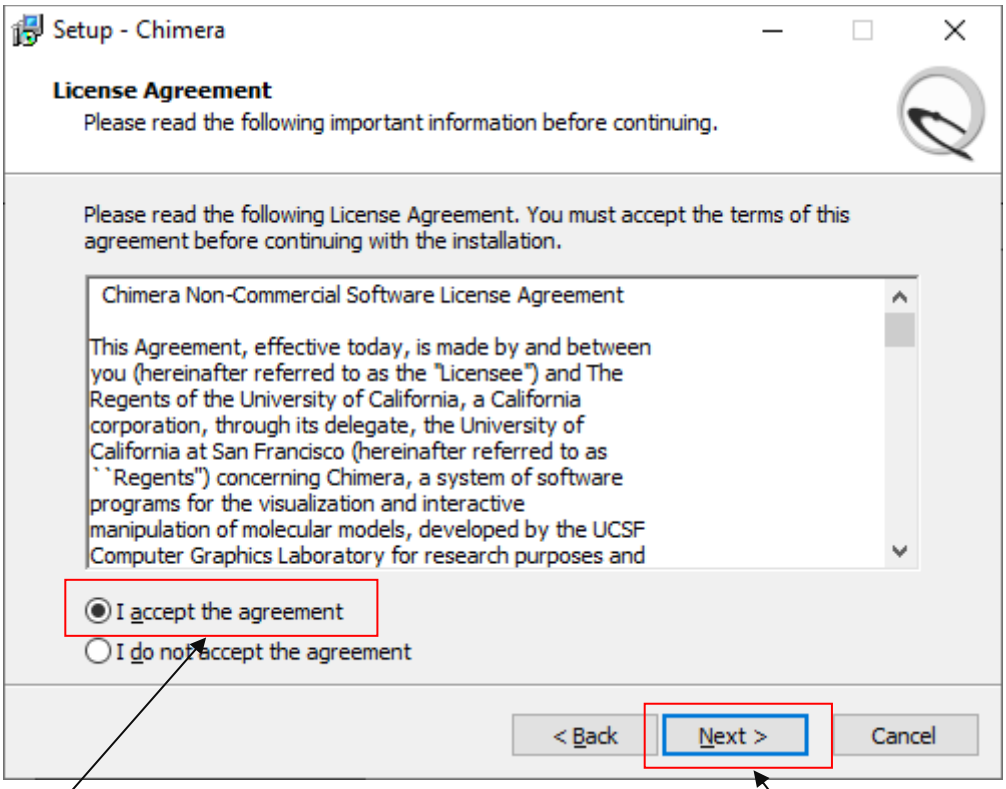


Fig 5: Accept License agreement.

Select accept the agreement

Click on Next

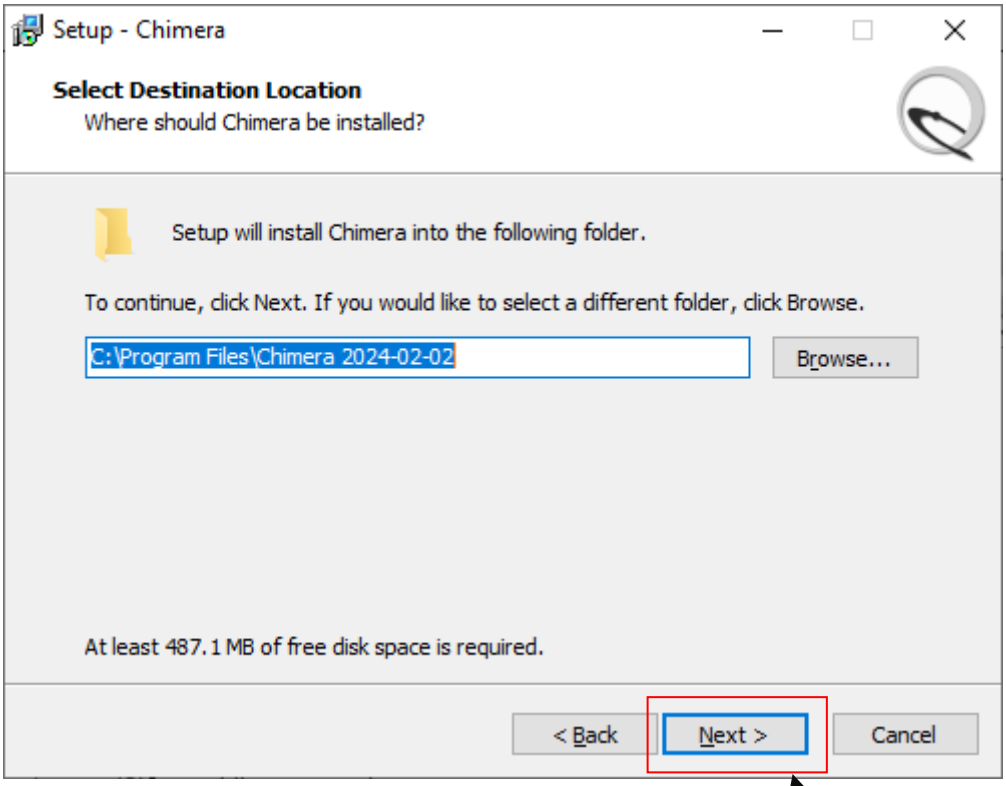


Fig 6: Select Destination location

Click on Next

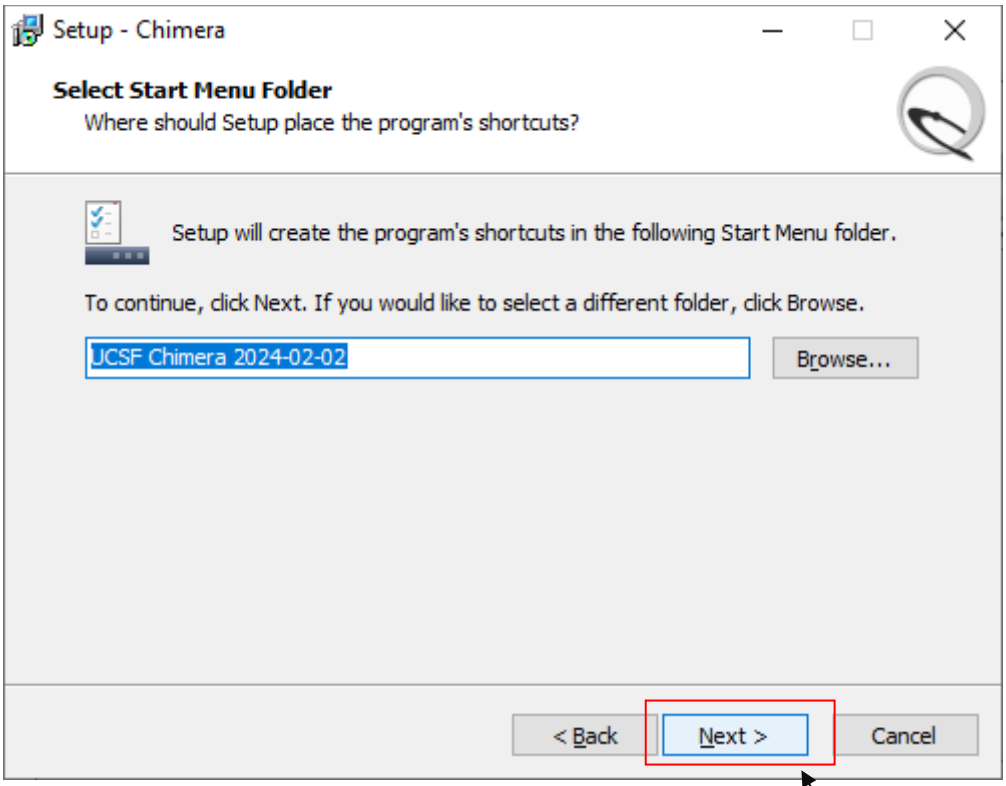


Fig 7: Select start menu folder

Click on Next

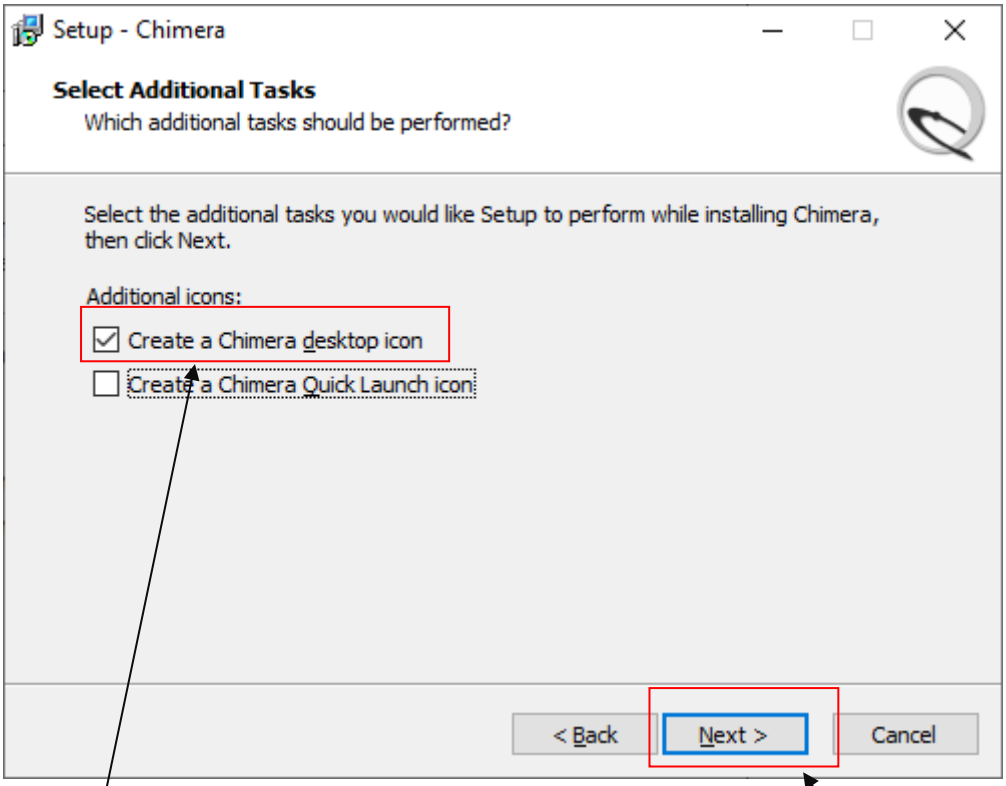


Fig 8: Select additional task

Select on desktop icon

Click on Next

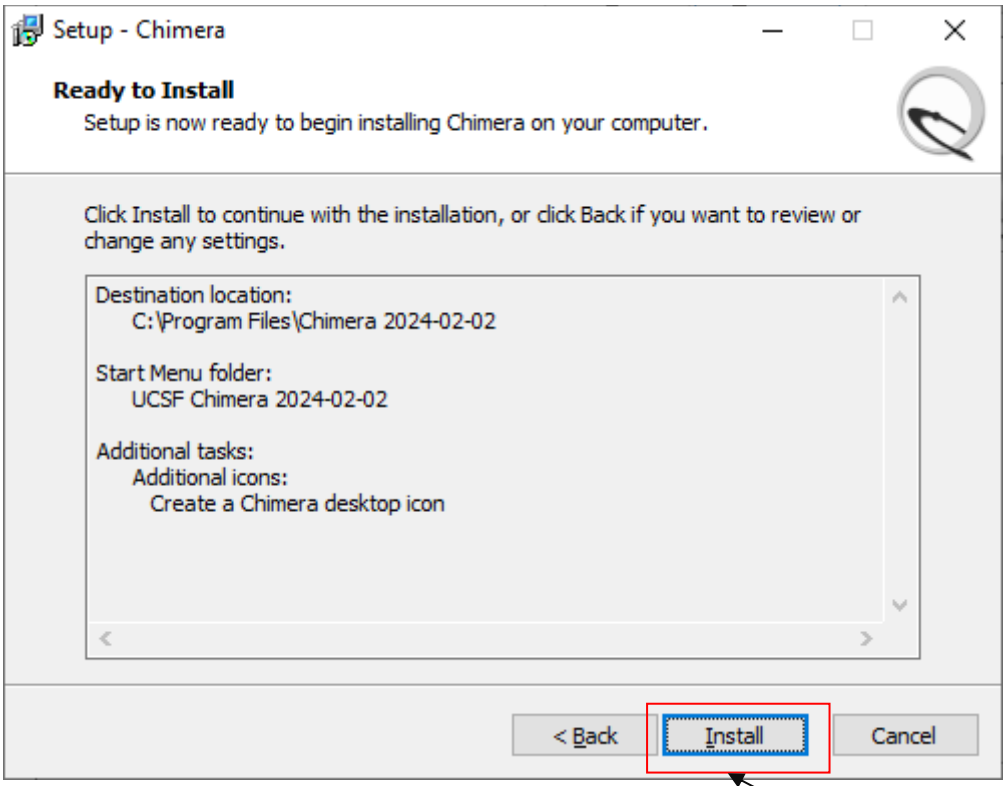


Fig 9: Install the UCSF Chimera

Click on Install

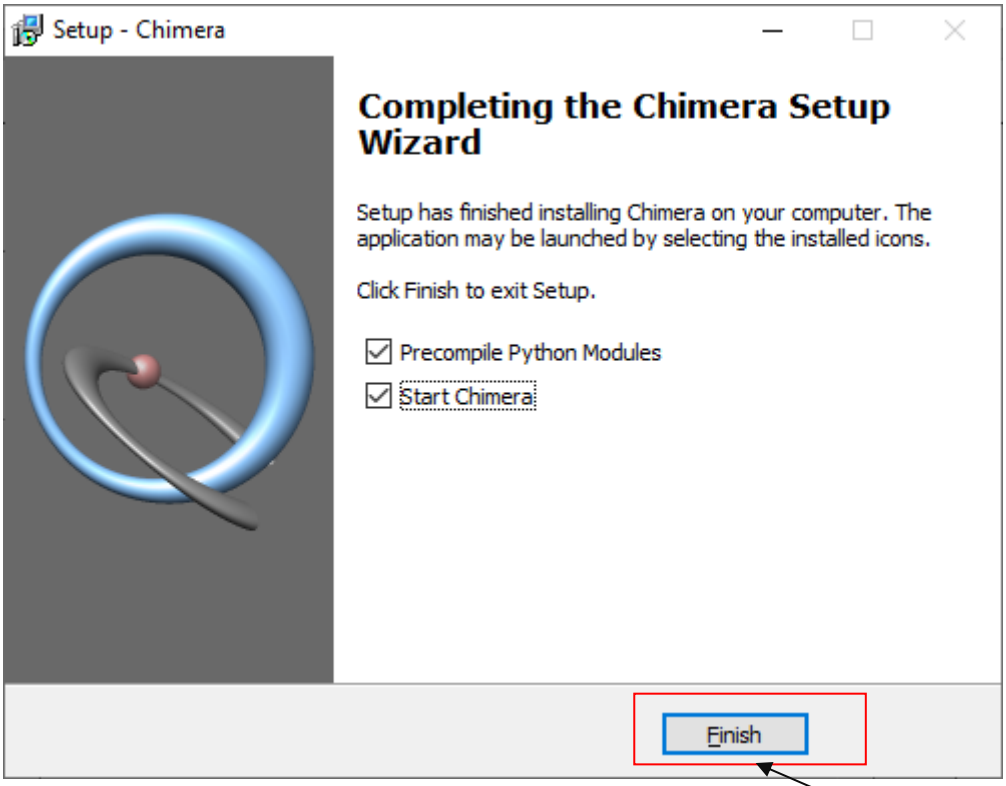


Fig 10: Finish Chimera installing setup

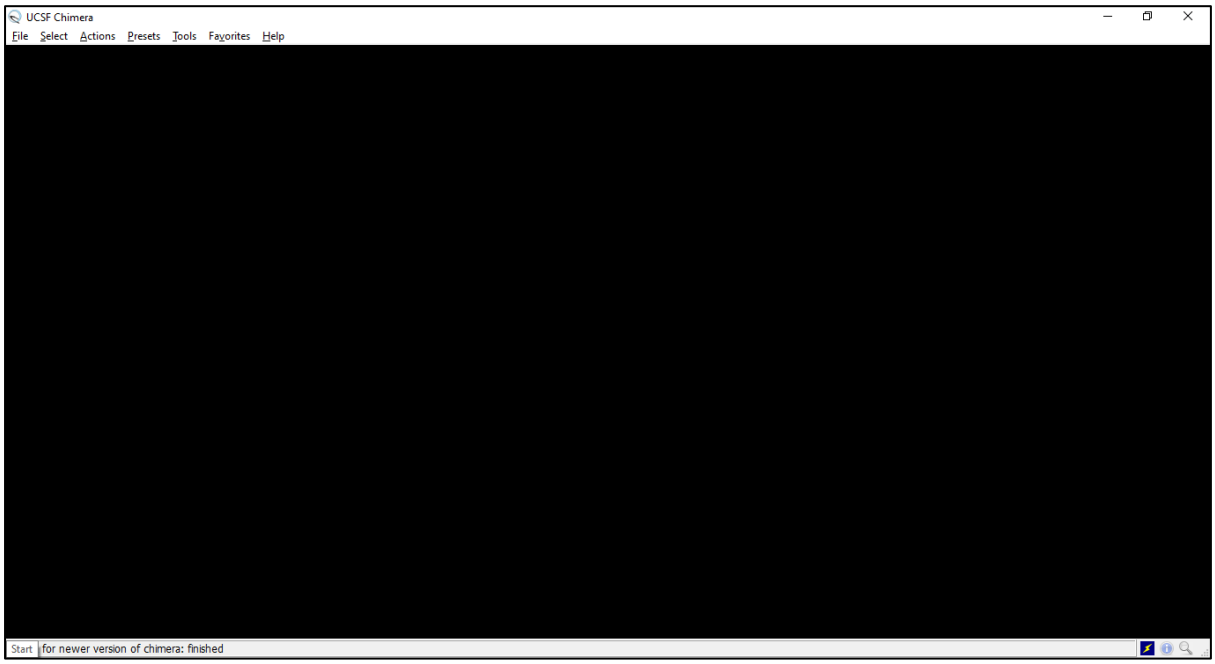


Fig 11: Open Chimera application from desktop

REFERENCES:

1. Rodríguez-Guerra Pedregal J, Maréchal JD.Bioinformatics. 2018 May 15;34(10):1784-1785. doi: 10.1093/bioinformatics/bty021.PMID: 29340616
2. Sheth, Vrunda, "Visualization of protein 3D structures in reduced representation with simultaneous display of intra and intermolecular interactions" (2009). Thesis. Rochester Institute of Technology.
3. UCSF Chimera--a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. J Comput Chem. 2004 Oct;25(13):1605-12.
4. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem 2004, 25(13):1605–1612. 10.1002/jcc.20084
5. Meng, E.C., Pettersen, E.F., Couch, G.S. et al. Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics 7, 339 (2006). https://doi.org/10.1186/1471-2105-7-339.
6. Tate, J. 2003. “Molecular visualization.” In Structural Bioinformatics, edited by P. E. Bourne and H. Weissig, 135–58. Hoboken, NJ: Wiley-Liss.

DATE: 12/02/2024

WEBLEM 12 (A)
RASMOL TOOL

AIM:

To visualize 3D structure of protein 'Insulin' (PDB ID: 1GZ8) using RasMol tool.

INTRODUCTION:

In the field of molecular visualization, RasMol is undoubtedly a venerable piece of software that, since its creation, has had a tremendous influence on the scientific community. RasMol, one of the first applications to give researchers an interactive tool for viewing molecular structures on computers, was created by Roger Sayle in the early 1990s. Scientists, teachers, and students all favored it because of its use and accessibility.

The capacity of RasMol to read and display molecular structures contained in the Protein Data Bank (PDB) format is one of its primary features. The atomic coordinates and associated data of biological macromolecules, including proteins, nucleic acids, and carbohydrates, may be represented uniformly using the PDB format. Users may work with and examine these molecules' three-dimensional structures in real-time by importing PDB files into RasMol.

Users may rotate, translate, and zoom in on molecular structures using the RasMol user interface, which offers a dynamic and user-friendly approach to examine intricate biomolecular assemblies. It also provides a range of rendering choices, including wireframe, ball-and-stick, and space-filling models, to depict molecules in various forms. RasMol's adaptability makes it appropriate for a variety of uses, such as structural biology research, molecular modeling, and molecular teaching.

Additionally, RasMol's source code has been made publicly available, which has encouraged the creation of several modifications and derivative applications throughout time. These include, among others, RasTop, PyMOL, and Jmol; each builds on the base that RasMol provided and offers a unique set of features and capabilities.

All things considered, RasMol's reputation as a trailblazing molecular visualization tool lives on today, continuing to be essential to the investigation and comprehension of molecular structures and their activities.

Insulin:

Insulin is a vital hormone that regulates blood sugar levels by allowing glucose to enter cells for energy production. It is produced by beta cells in the pancreas and plays a crucial role in glucose storage and production. Insulin was first isolated in 1921 by Canadian scientists Frederick G. Banting and Charles H. Best, leading to life-saving treatments for diabetes. In diabetes, either the body does not produce enough insulin (Type 1) or becomes resistant to its effects (Type 2). Insulin resistance can lead to high blood sugar levels and various health complications. Different types of insulin, including fast, intermediate, and long-acting insulins, are used based on individual needs to manage blood sugar effectively.

Insulin is a composed of two peptide chains, an A chain and a B chain, linked together by disulfide bonds. The A chain consists of 21 amino acids, while the B chain has 30 amino acids. Within the A chain, there is an additional disulfide bond. Insulin molecules have a tendency to form dimers in solution and can associate into hexamers in the presence of zinc ions. The amino acid sequence of insulin is highly conserved among species, with minor variations. Despite these variations, insulin from different species can be biologically active across species. The structure of insulin allows it to regulate blood glucose levels by promoting glucose storage and inhibiting glucose production and release by the liver.

METHODOLOGY:

1. Download RasMol from http://RasMol.org/ and run the executive file.
2. Register RasMol for the first time by double-clicking the “RasWin” icon.
3. Open two windows: 3D visualization window and RasMol command line prompt.
4. Select the PDB file 1GZ8.pdb from PDB database.
5. Display patterns available in the Display menu include “Wireframe” etc.

OBSERVATIONS:

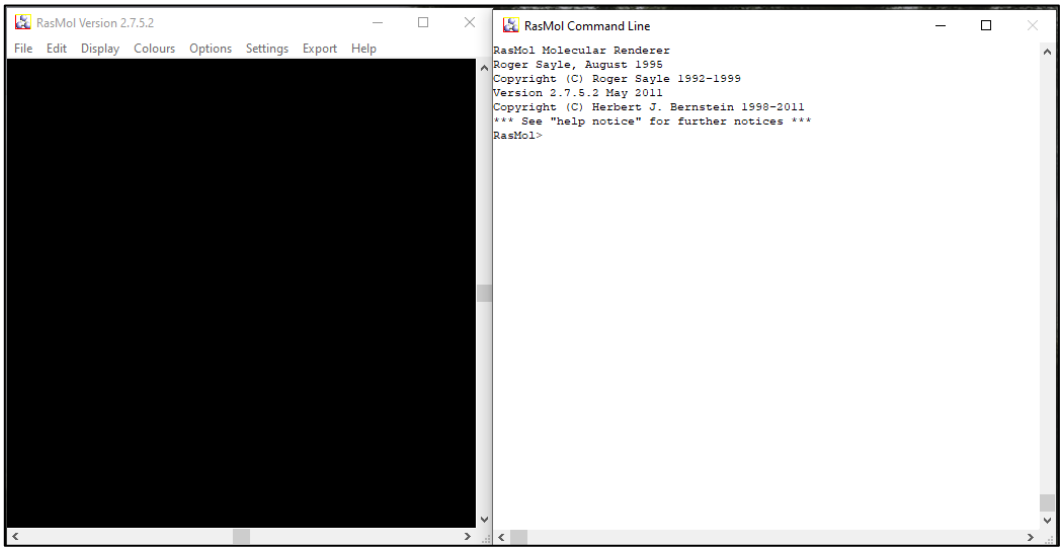


Fig 1: Homepage and Command line of RasMol tool

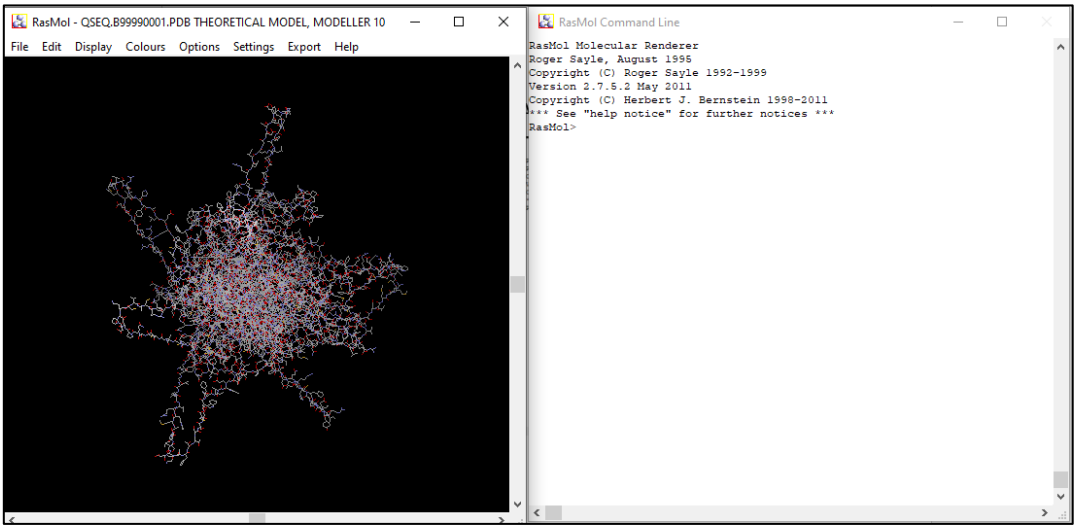


Fig 2: ‘Insulin’ (PDB ID: 1GZ8) protein opened in the RasMol tool

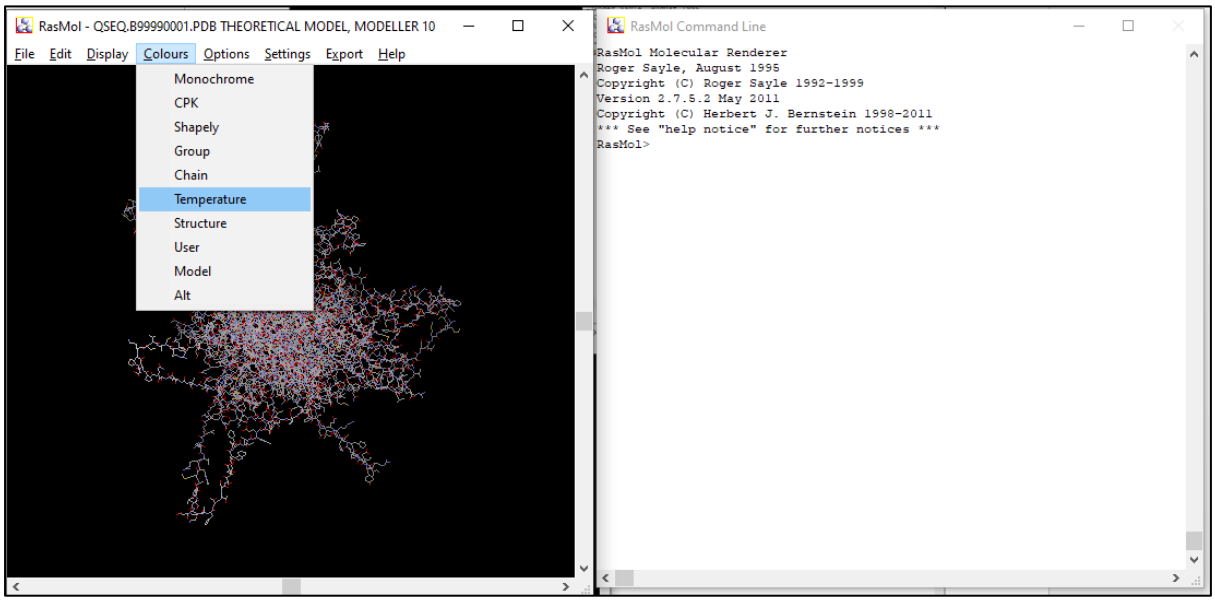


Fig 3: Option selected – Colors: Temperature

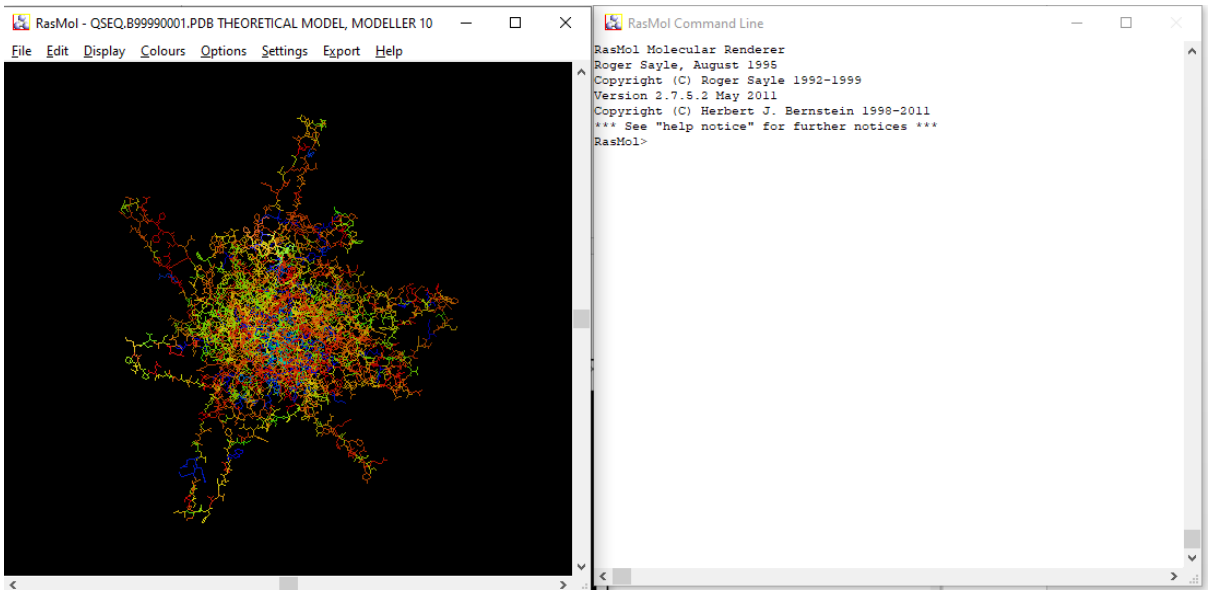


Fig 3.1: Output for Option selected – Colors: Temperature

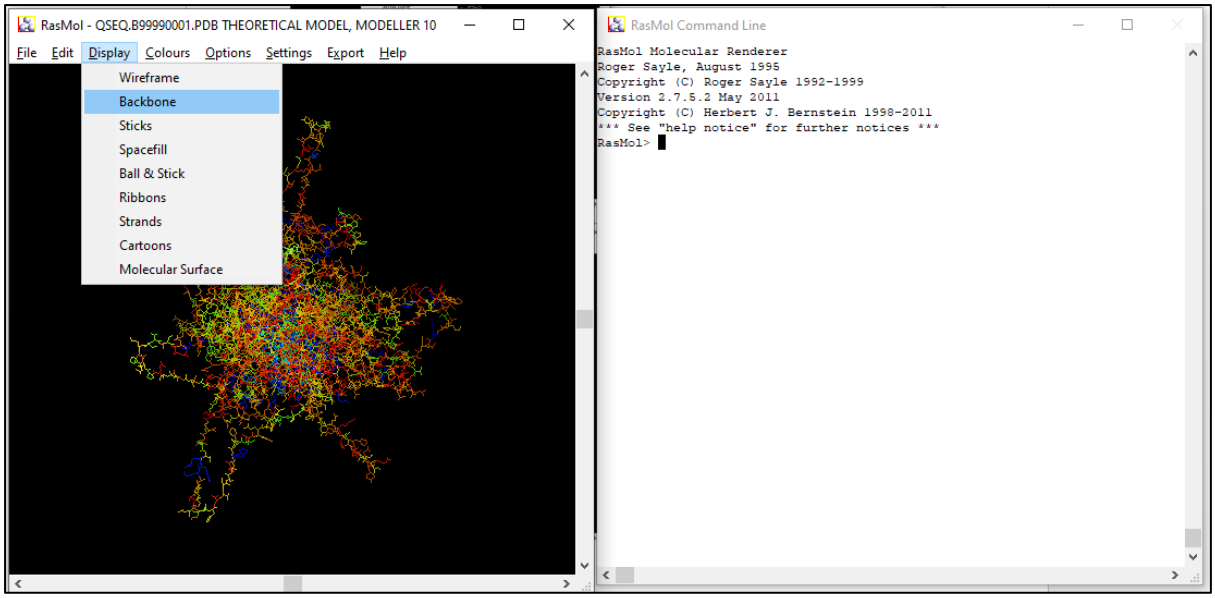


Fig 4: Option selected – Display: Backbone

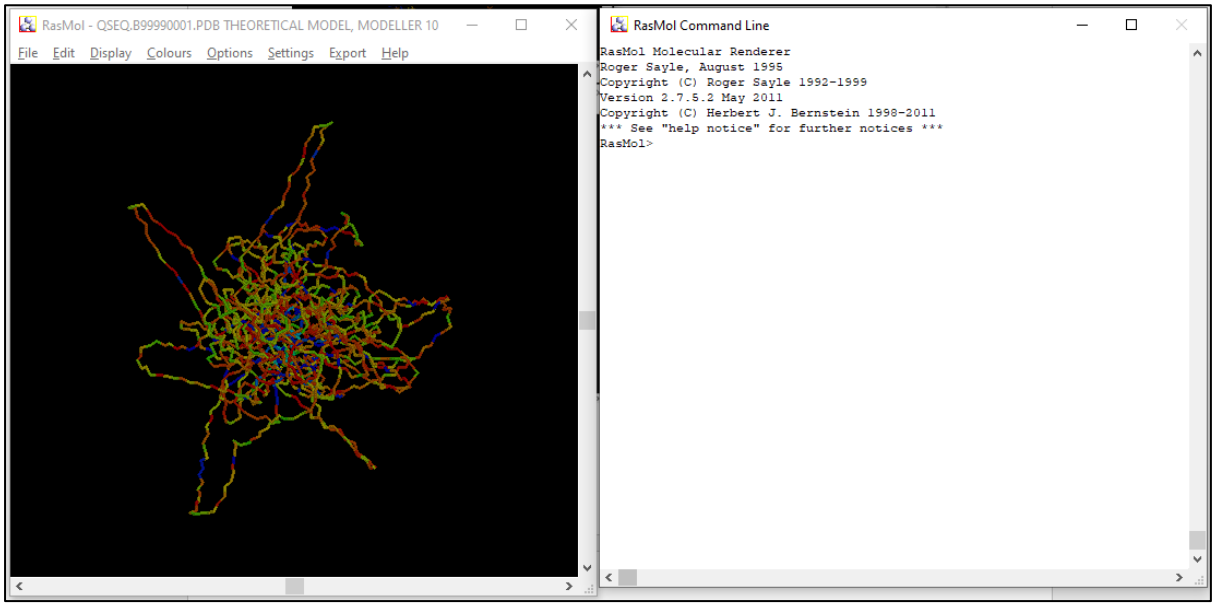


Fig 4.1: Output for Option selected – Display: Backbone

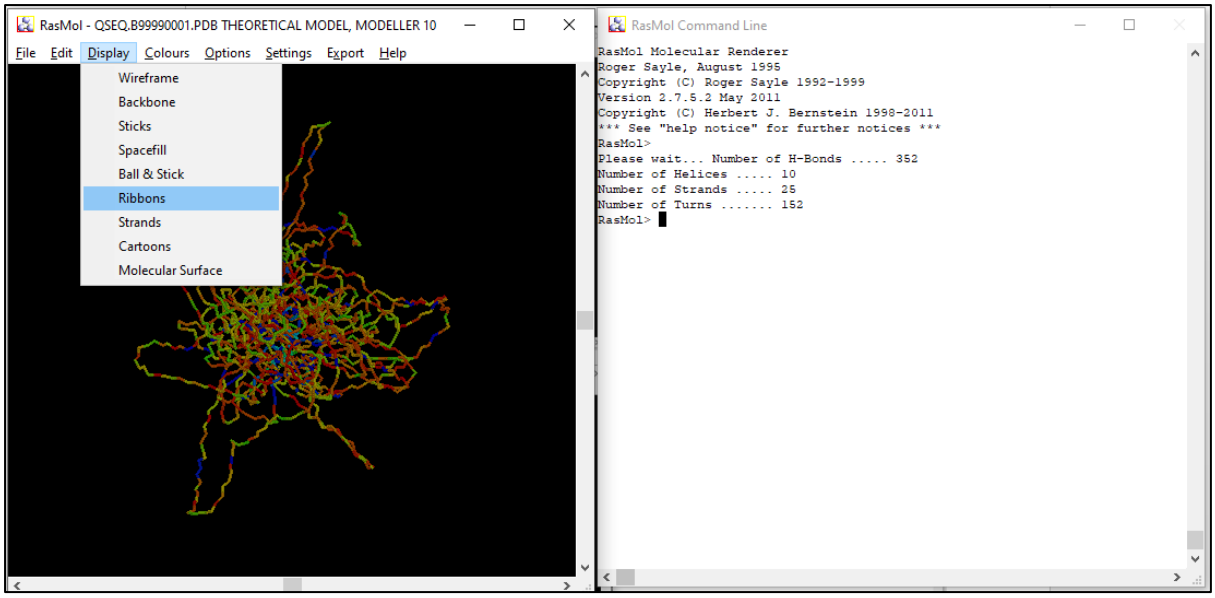


Fig 5: Option selected – Display: Ribbons

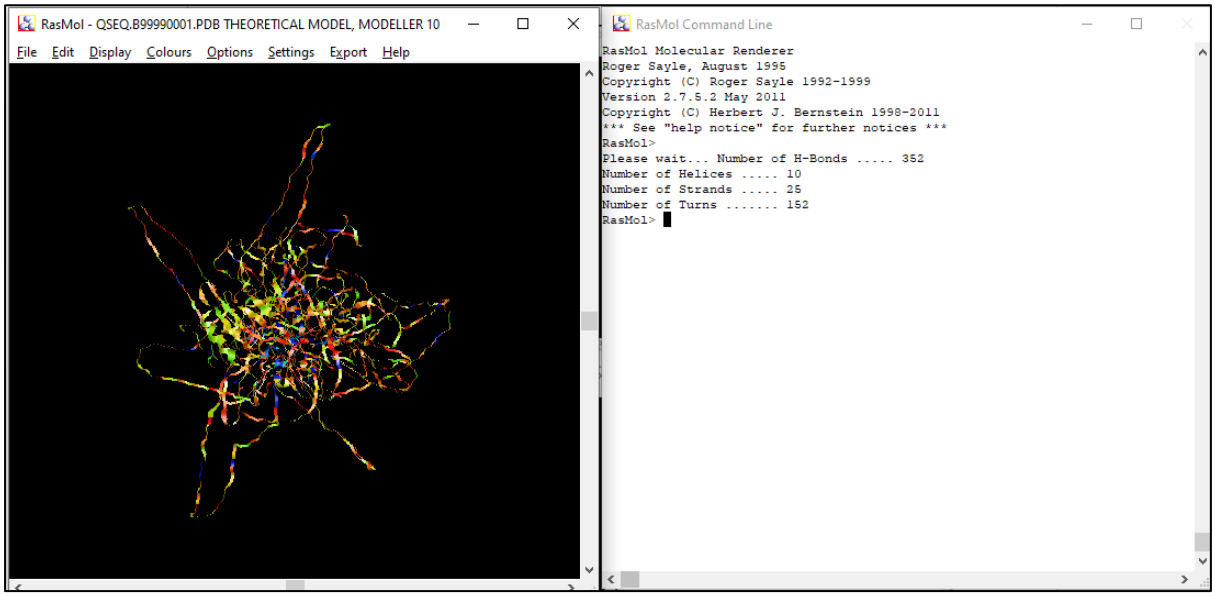


Fig 5.1: Output for Option selected – Display: Ribbons

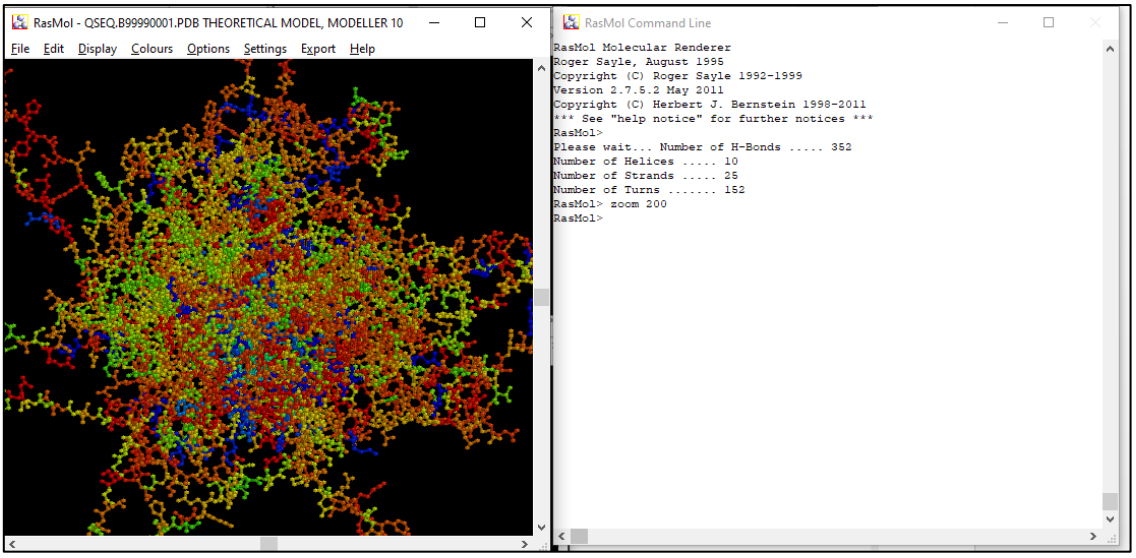


Fig 6: Output for command executed in command line: zoom 200

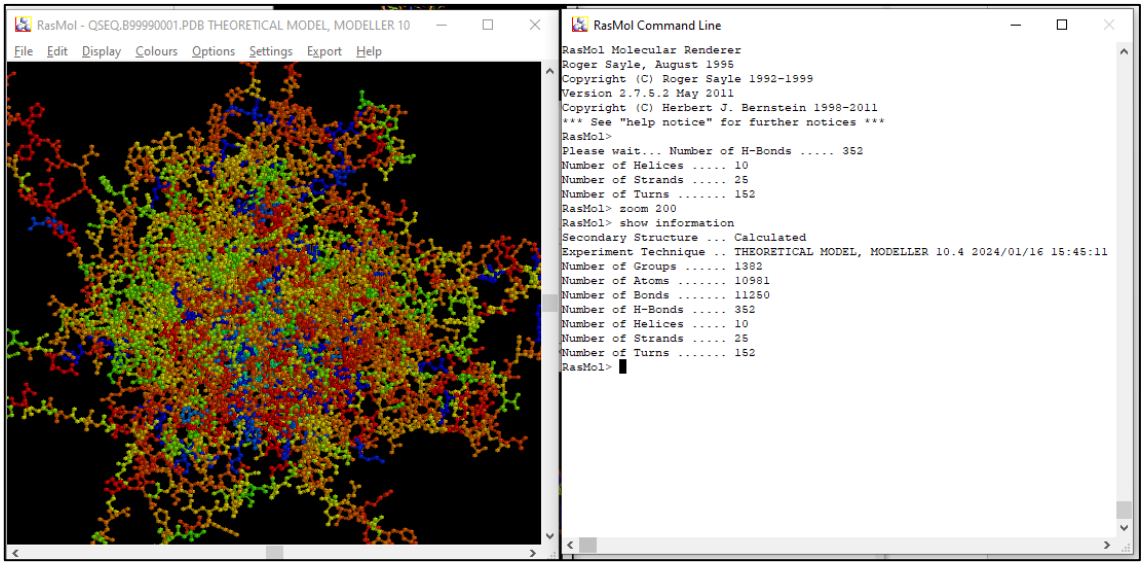


Fig 7: Output for command executed in command line: show information

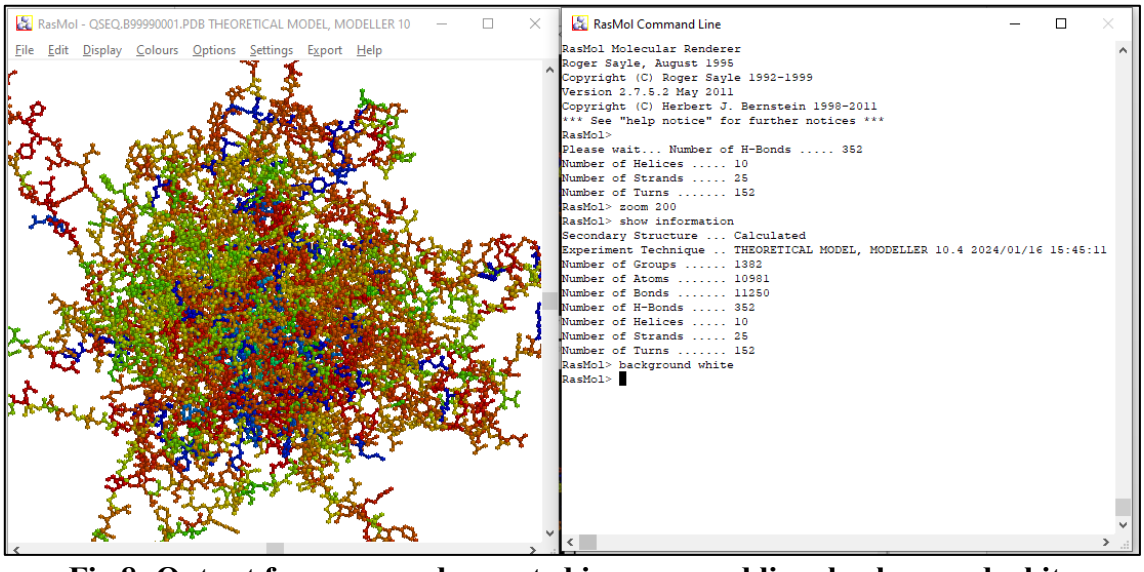


Fig 8: Output for command executed in command line: background white

RESULTS:

RasMol is a comprehensive tool for molecular structures the secondary structural elements are highlighted when the 'Colors' option is changed to 'Temperature' and the 'Display' option is changed to 'Backbone' and 'Ribbons' in the RasMol visualization tool. The structure's temperature and color correlate, revealing information about its thermal characteristics. The representation is further improved by zooming into the structure by a magnitude of 200 and changing the background color to white, to provide the best contrast for the visualization using the command line. Information about the number of groups, atoms, bonds H-bonds, helices, strands and turns were obtained by executing the command: 'show information' in the command line.

CONCLUSION:

RasMol is a crucial molecular visualization tool, providing a user-friendly platform for exploring complex biomolecular structures. Its open-source nature has led to developers expanding its capabilities, making it a cornerstone of computational biology and structural biology research.

REFERENCES:

1. RasMol and OpenRasMol. (n.d.). http://www.openRasMol.org/
2. Mukhopadhyay, C. S. (n.d.). Basic Applied Bioinformatics. O'Reilly Online Learning. https://www.oreilly.com/library/view/basic-applied-bioinformatics/9781119244332/c04.xhtml#:~:text=The%20name%20%E2%80%9CRasMol%E2%80%9D%20is%20an,written%20in%20PDB%20file%20format.
3. RCSB PDB: Homepage. (2010). Rcsb.org. https://www.rcsb.org/
4. Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of Insulin in Health and Disease: An Update. International journal of molecular sciences, 22(12), 6403. https://doi.org/10.3390/ijms22126403
5. Weiss M, Steiner DF, Philipson LH. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. [Updated 2014 Feb 1]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279029/

DATE: 12/02/2024

WEBLEM 12(B)
CHIMERA
(URL: http://www.cgl.ucsf.edu/chimera/)

AIM:

To visualization of 3D structure of protein 'Insulin' (PDB ID: 1GZ8) using Chimera.

INTRODUCTION:

UCSF Chimera is a powerful visualization tool remarkably present in the computational chemistry and structural biology communities. UCSF Chimera is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. High quality images and animations can be generated. Protein structure visualization tools are software applications that allow users to view, manipulate, and analyze the three-dimensional (3D) structures of proteins and other biomolecules. These tools are essential for understanding the molecular architecture, function, and interactions of proteins, as well as for designing new drugs, vaccines, and biotechnologies. Built on a C++ core wrapped under a Python 2.7 environment, one could expect to easily import UCSF Chimera's arsenal of resources in custom scripts or software projects. The Chimera core consists of a C++ layer that handles time critical operations (e.g., graphics rendering) and a Python layer that handles all other functions. Nonetheless, this is not readily possible if the script is not executed within UCSF Chimera due to the isolation of the platform. All significant C data and functions are made accessible to the Python layer. Core capabilities include molecular file input/output, molecular surface generation using the MSMS algorithm, and aspects of graphical display such as wire-frame, ball-and-stick, ribbon, and sphere representations, transparency control, near and far clipping planes, and lenses. Chimera's primary programming language is Python. Importantly, Python is an interpreted, object-oriented programming language that is also easy to learn and very readable. Python is interpreted, it is good for rapid development and debugging. Readability is important for a team development project like Chimera, and an easy-to-learn language enables others to develop extensions without undue effort. Chimera includes the Python-standard IDLE interactive development environment to help diagnose problems during extension development. Chimera is divided into a core and extensions. The core provides basic services and molecular graphics capabilities. All higher level functionality is provided through extensions. This design, with the bulk of Chimera functions provided by extensions, ensures that the extension mechanism is robust enough to handle the needs of outside researchers wanting to extend Chimera in novel ways. The tools described here apply to many problems involving comparison and analysis of protein structures and their sequences. Chimera includes complete documentation and is 8 intended for use by a wide range of scientists, not just those in the computational disciplines. Chimera's Collaborators extension enables researchers at geographically distant locations to share a molecular modeling session in real time. UCSF Chimera is freely available to academic and nonprofit researchers from http://www.cgl.ucsf.edu/chimera/, and is available for

Microsoft Windows, Apple Mac OS X, Linux, and other platforms and can be licensed by commercial institutions for a fee. Extensions to Chimera developed by outside researchers can be redistributed freely.

Insulin:

Insulin is a vital hormone that regulates blood sugar levels by allowing glucose to enter cells for energy production. It is produced by beta cells in the pancreas and plays a crucial role in glucose storage and production. Insulin was first isolated in 1921 by Canadian scientists Frederick G. Banting and Charles H. Best, leading to life-saving treatments for diabetes. In diabetes, either the body does not produce enough insulin (Type 1) or becomes resistant to its effects (Type 2). Insulin resistance can lead to high blood sugar levels and various health complications. Different types of insulin, including fast, intermediate, and long-acting insulins, are used based on individual needs to manage blood sugar effectively.

Insulin is a composed of two peptide chains, an A chain and a B chain, linked together by disulfide bonds. The A chain consists of 21 amino acids, while the B chain has 30 amino acids. Within the A chain, there is an additional disulfide bond. Insulin molecules have a tendency to form dimers in solution and can associate into hexamers in the presence of zinc ions. The amino acid sequence of insulin is highly conserved among species, with minor variations. Despite these variations, insulin from different species can be biologically active across species. The structure of insulin allows it to regulate blood glucose levels by promoting glucose storage and inhibiting glucose production and release by the liver.

METHODOLOGY:

1. Install Chimera 1.17.3 Application on your computer.
2. Open Chimera.
3. Click on File option and Open to upload ‘Insulin’ (PDB ID: 1GZ8) PDB file.
4. Move & interact with displayed molecule.
5. Review “preset” views.

OBSERVATIONS:

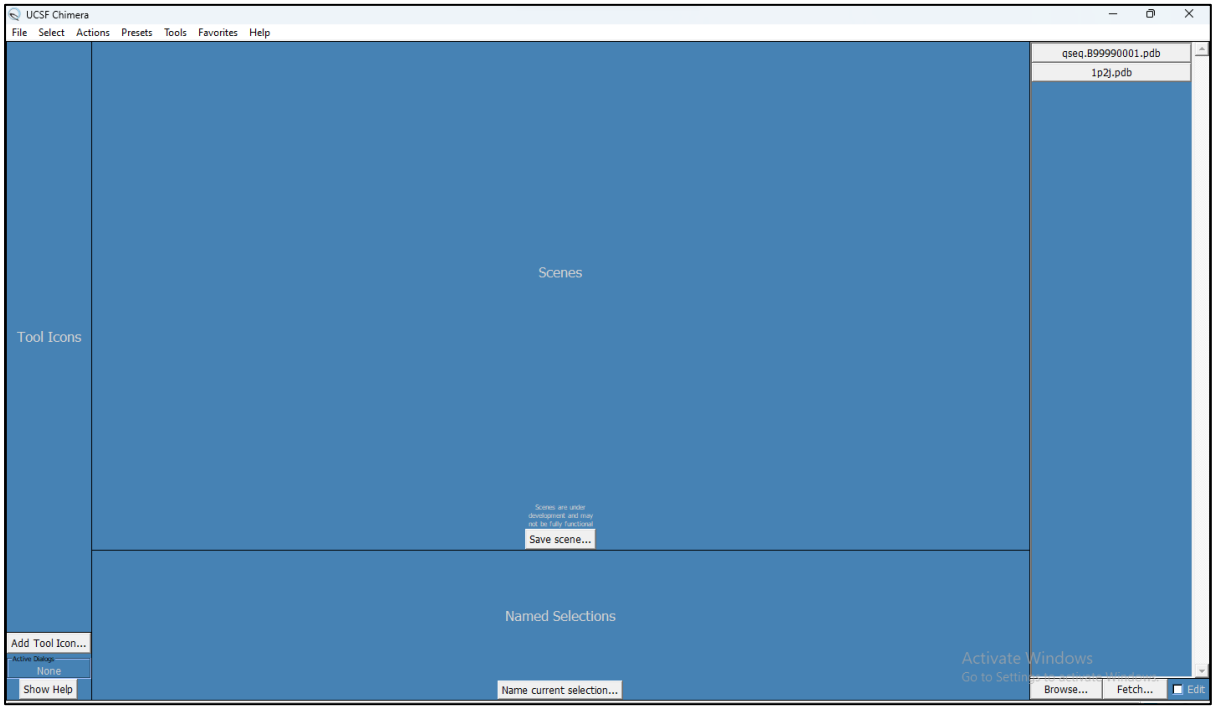


Fig 1: Homepage of Chimera tool

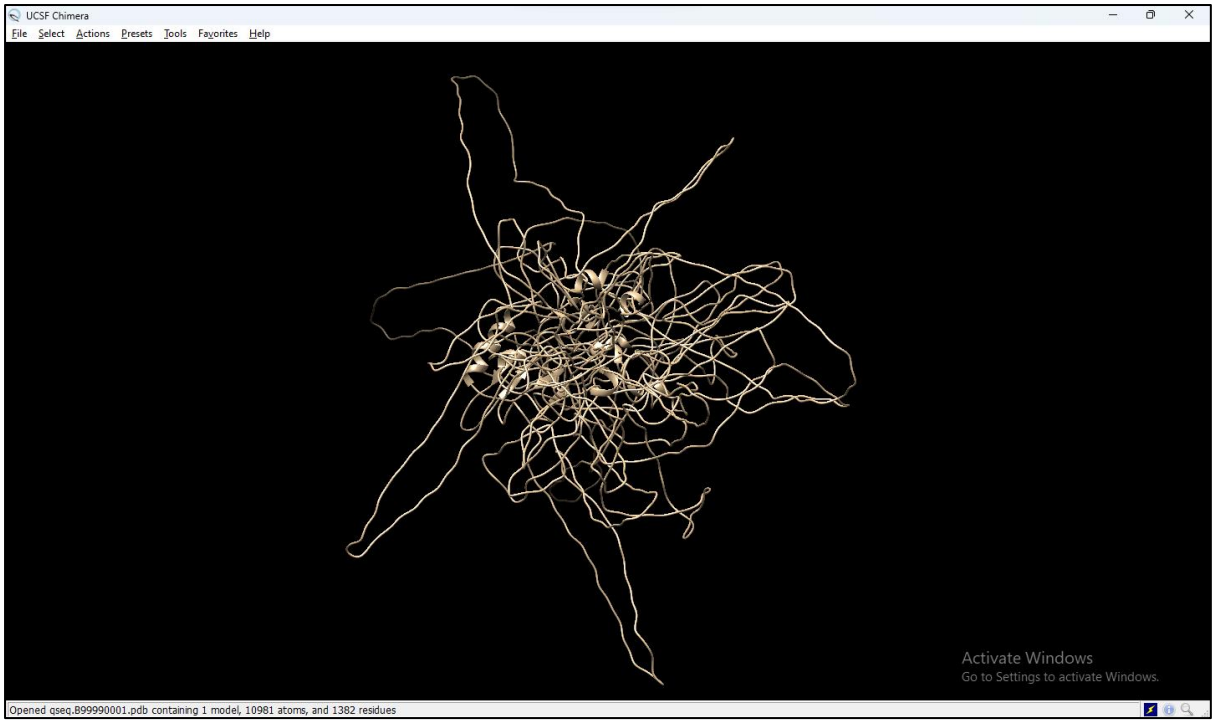


Fig 2: 'Insulin' (PDB ID: 1GZ8) protein opened in the Chimera tool

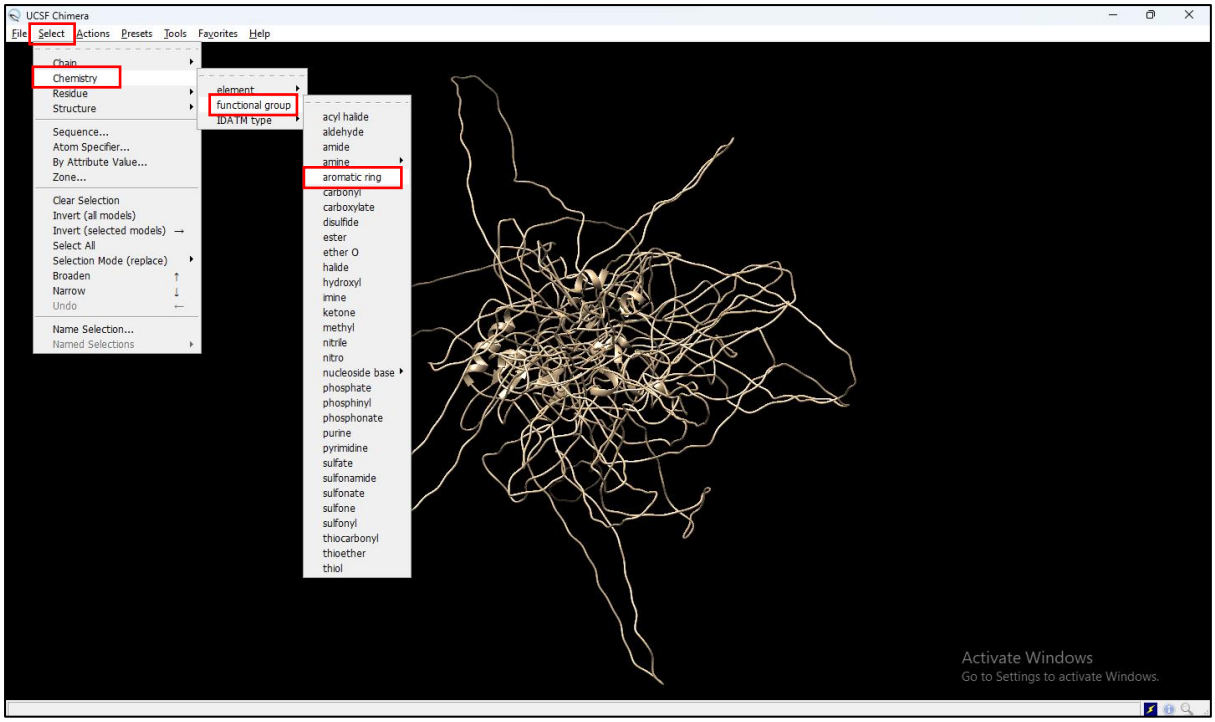


Fig 3: Option selected – 'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'

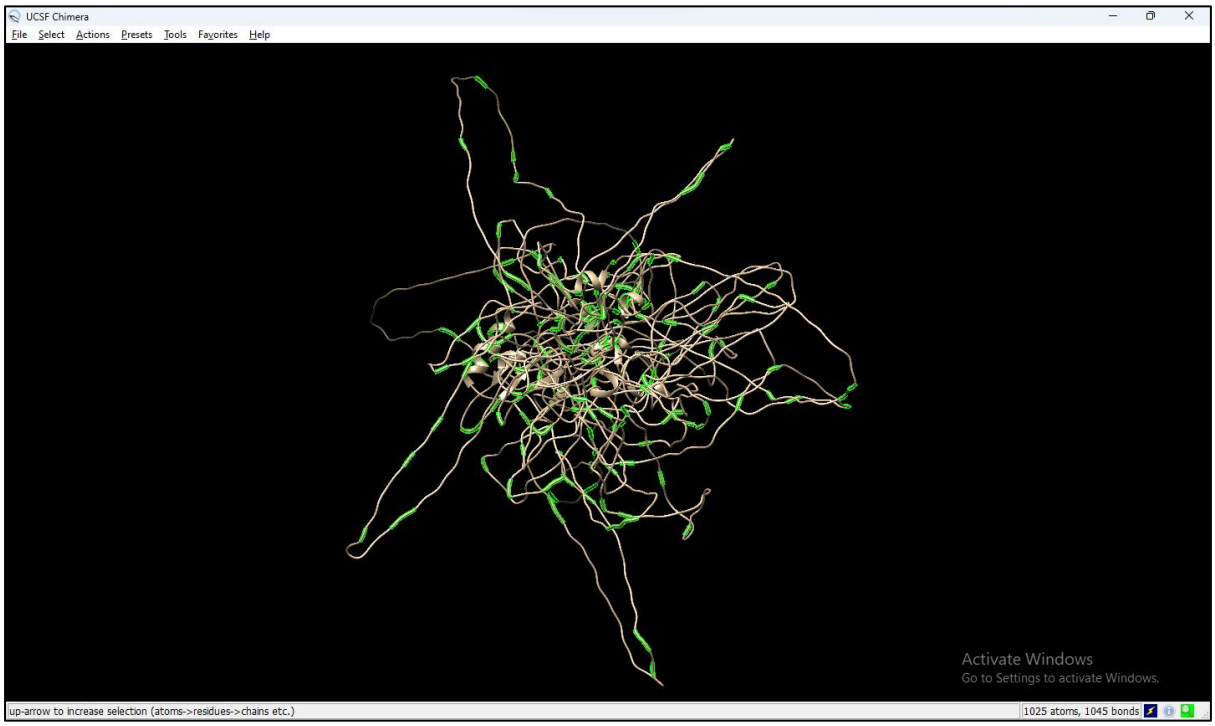


Fig 3.1: Output for Option selected – 'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'

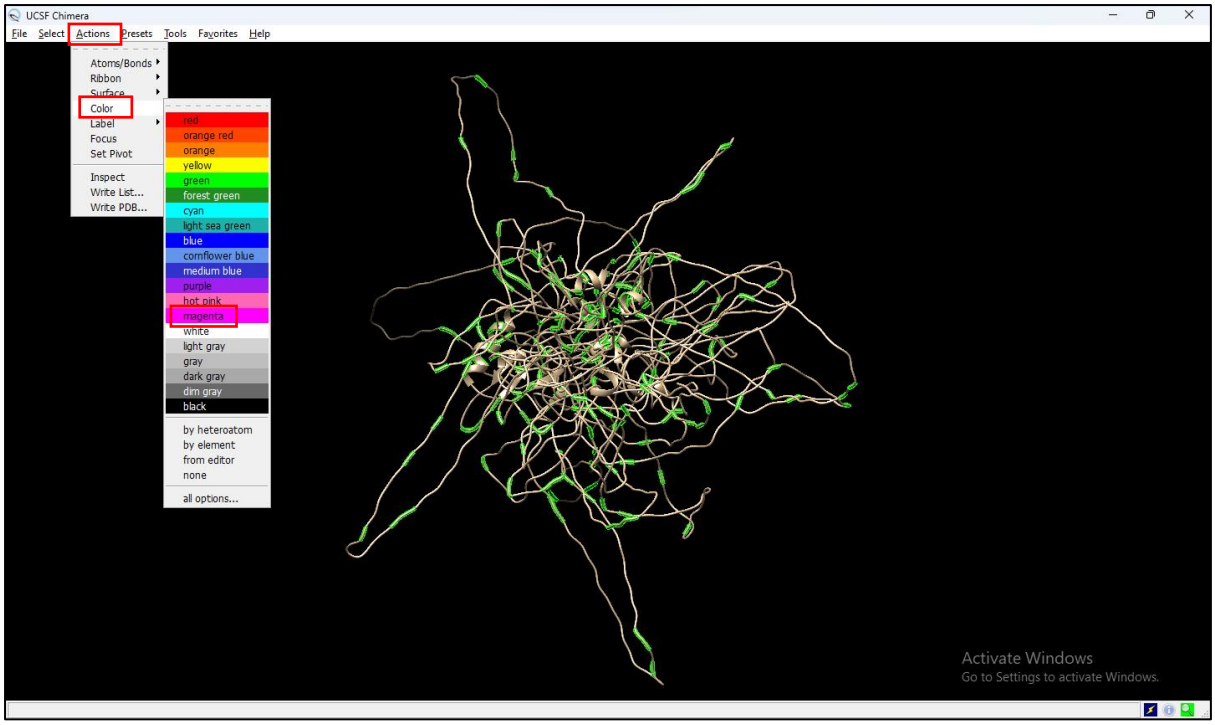


Fig 4: Options selected –

1. **'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'**
2. **'Actions' → 'Color' → 'Magenta'**

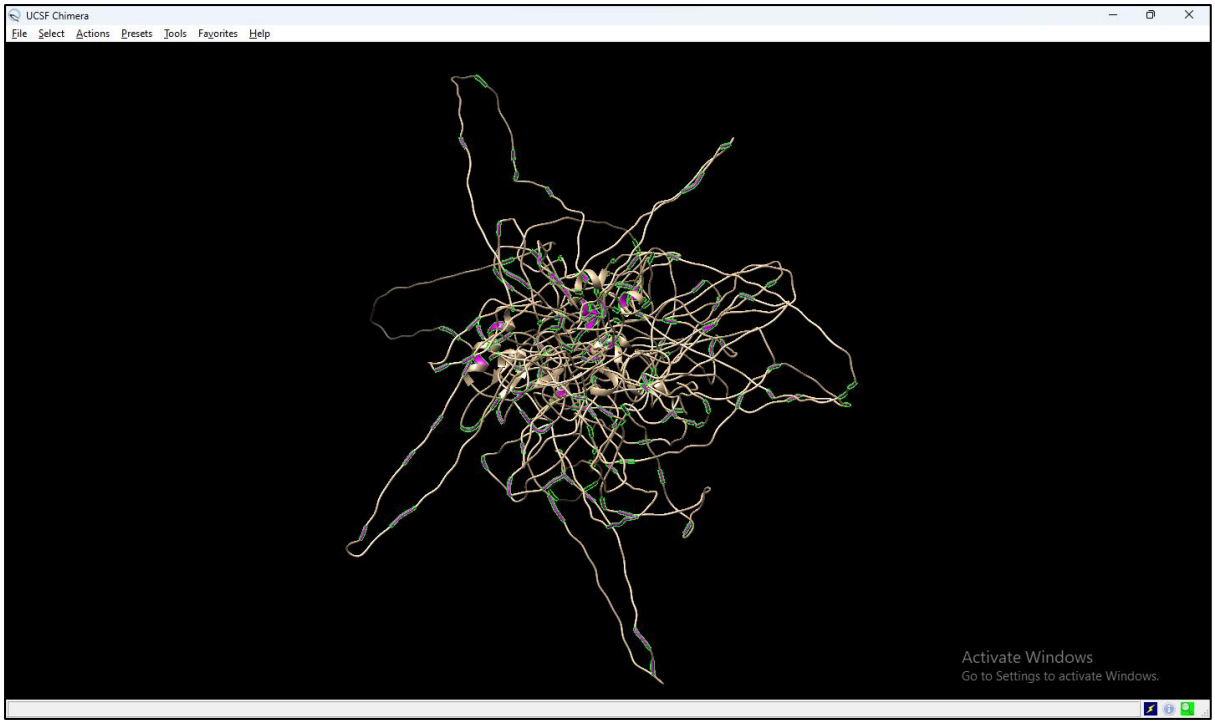


Fig 4.1: Output for the following Options selected –

1. **'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'**
2. **'Actions' → 'Color' → 'Magenta'**

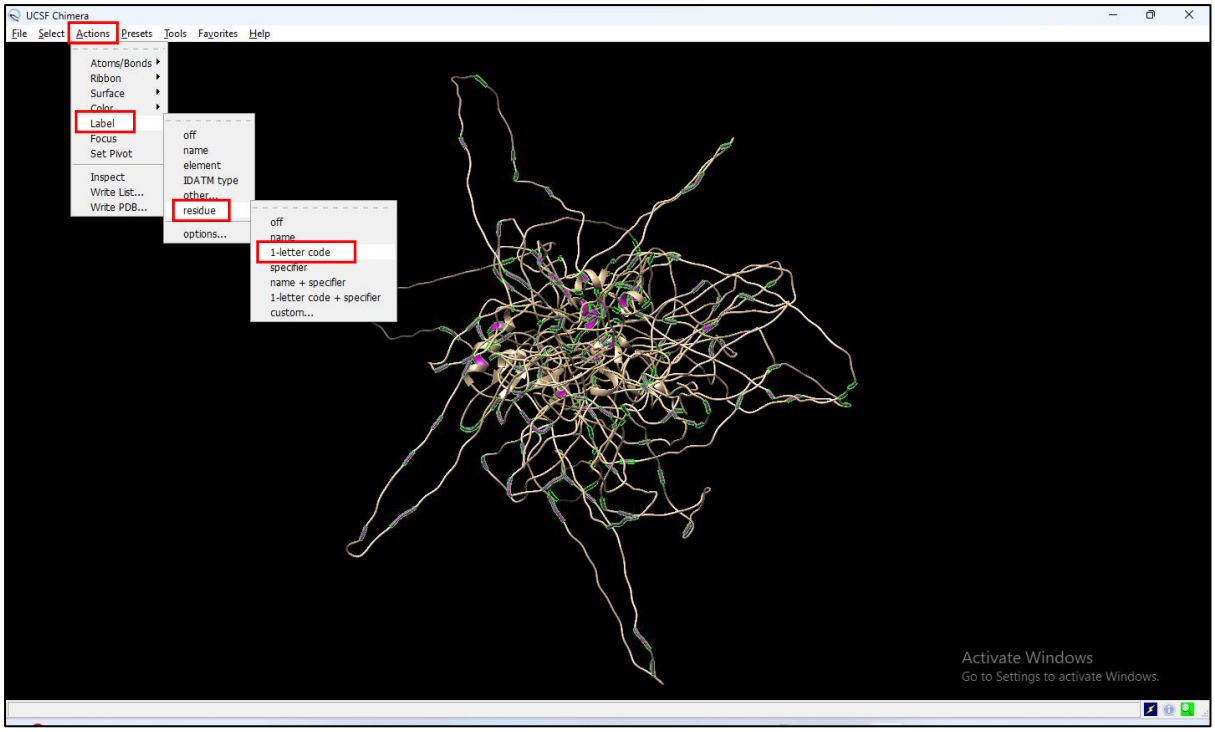


Fig 5: Options selected –

1. **'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'**
2. **'Actions' → 'Label' → 'residue' → '1-letter code'**

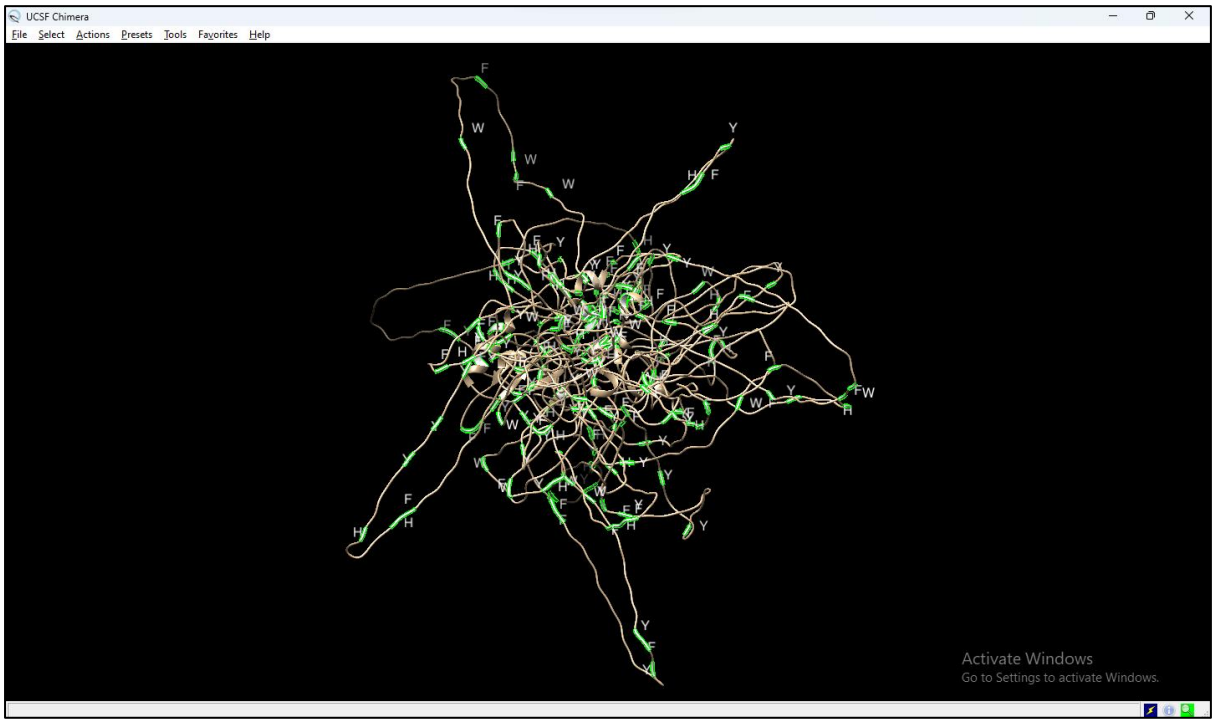


Fig 5.1: Output for the following Options selected –

1. **'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'**
2. **'Actions' → 'Label' → 'residue' → '1-letter code'**

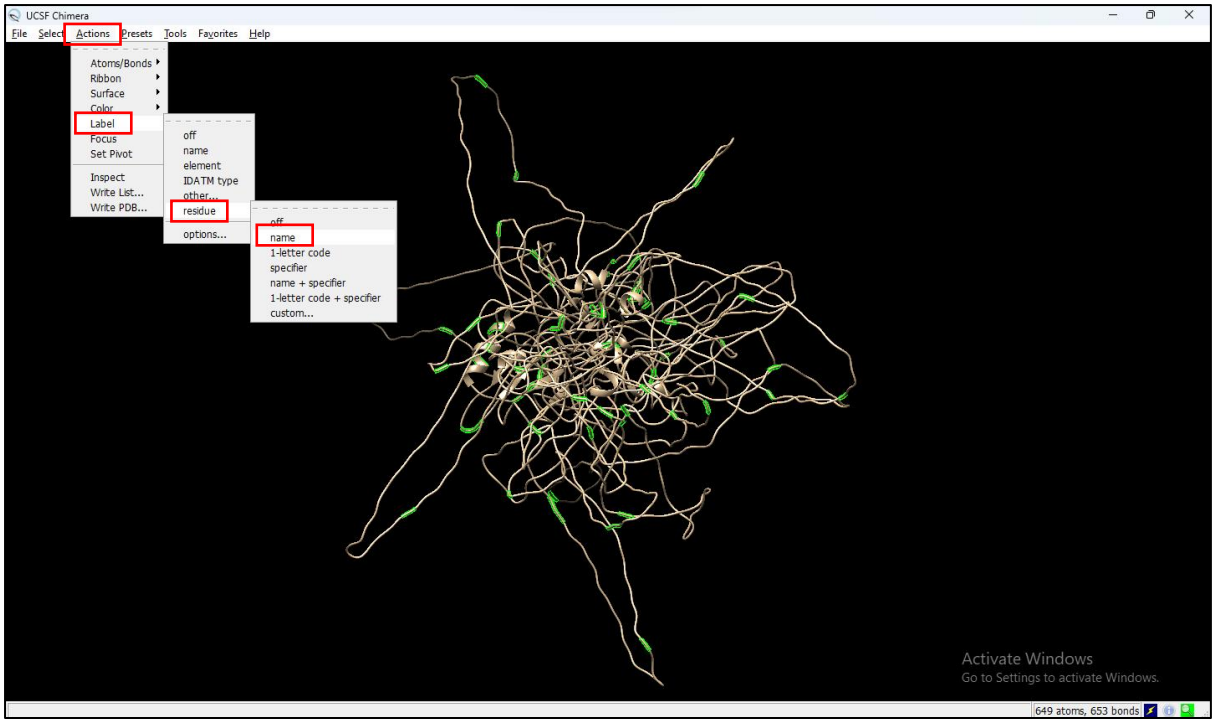


Fig 6: Options selected –

1. **'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'**
2. **'Actions' → 'Label' → 'residue' → 'name'**

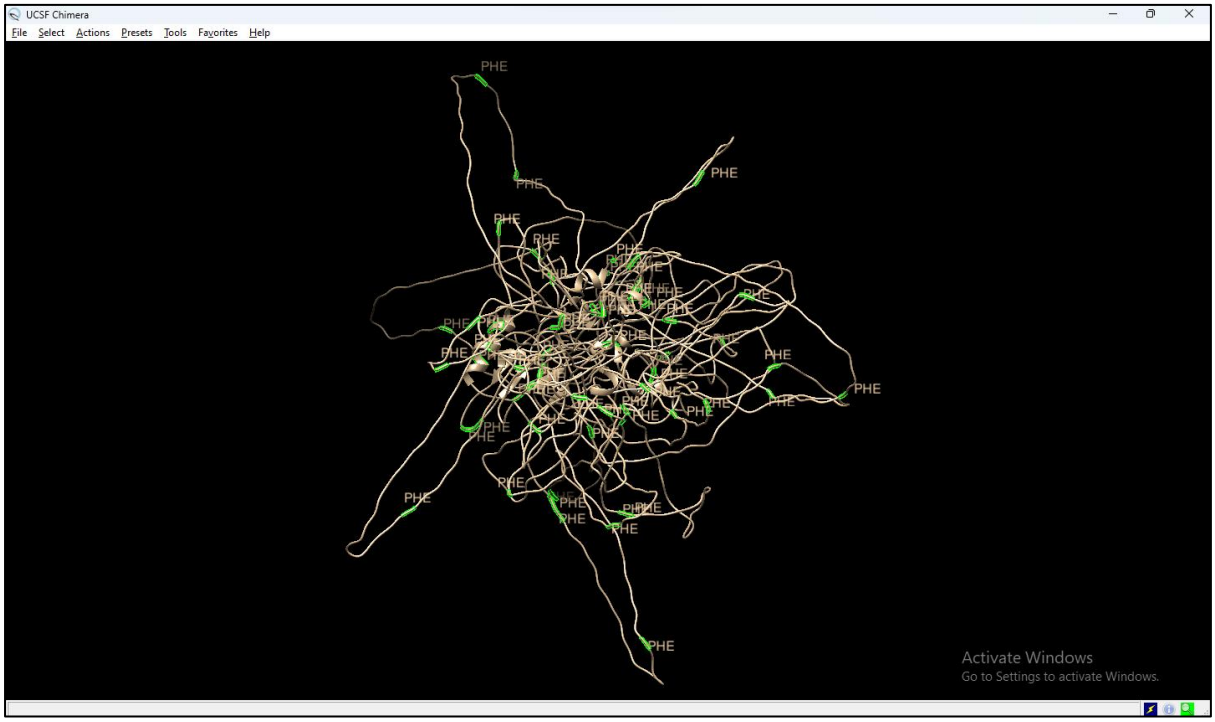


Fig 6.1: Output for the following Options selected –

1. **'Select' → 'Chemistry' → 'functional group' → 'aromatic ring'**
2. **'Actions' → 'Label' → 'residue' → 'name'**

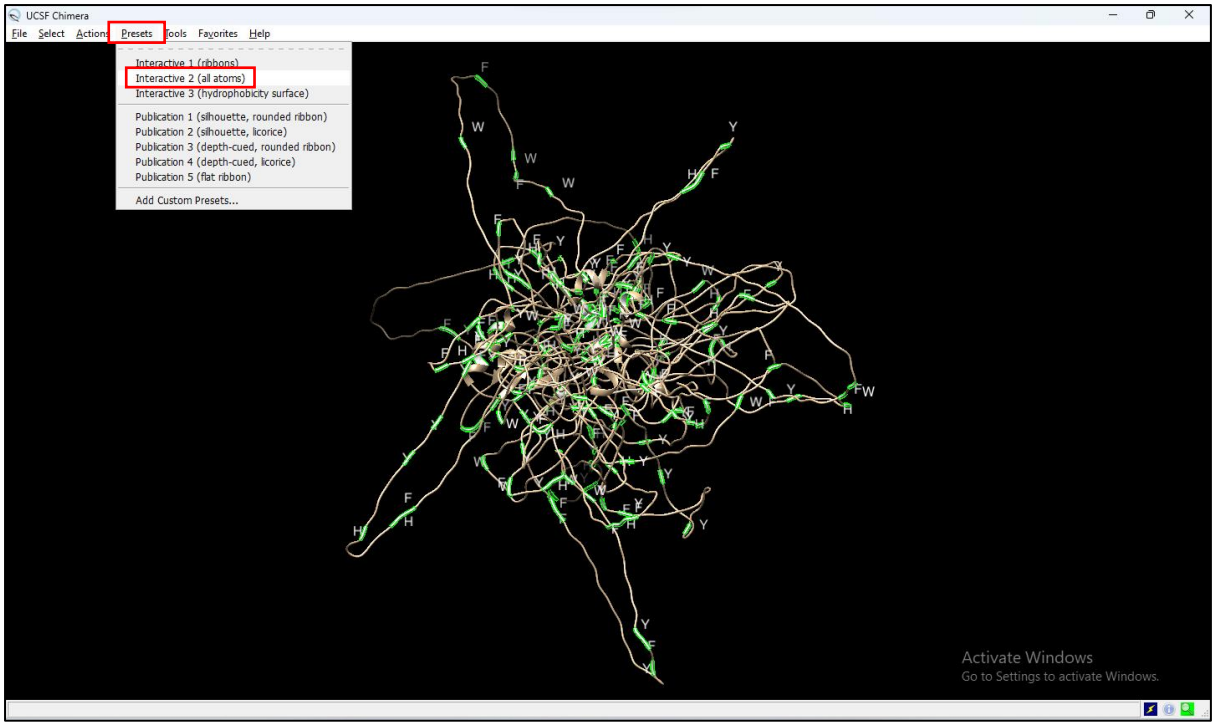


Fig 7: Option selected – 'Presets' → 'Interactive 2 (all atoms)'

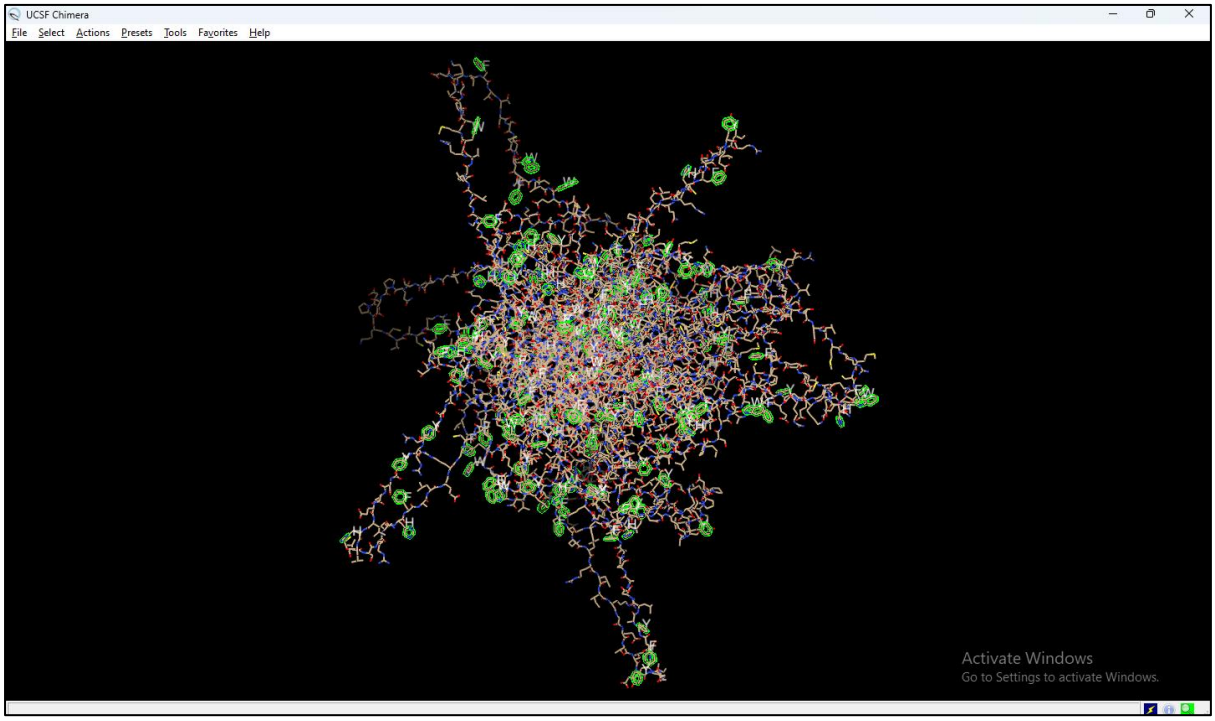


Fig 7.1: Output for Option selected – 'Presets' → 'Interactive 2 (all atoms)'

RESULTS:

The protein 'Insulin' structure with PDB ID: 1GZ8 were analyzed and visualized in various tools of UCSF Chimera. The aromatic rings, 1 – letter codes and the names of the selected functional groups were observed using the 'Select' and 'Actions' options for the visualization of the protein. The structural changes were observed in the 'Interactive 2(all atoms)' using the 'Presets' option. The Chimera tool may further be used to understand the 3D molecular structure, high quality figure and sequence alignment.

CONCLUSION:

By exploring molecular graphics program UCSF Chimera which is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. High-quality images and animations can be generated. A novel method is provided to superimpose structures and structural modeling. Chimera provides graphical user interfaces to simplify setting up input data and parameters for the fitting process, evaluating results, and performing cycles of refinement for building models of macromolecular assemblies.

REFERENCES:

1. UCSF Chimera--a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. J Comput Chem. 2004 Oct;25(13):1605-12. https://doi.org/10.1002/jcc.20084
2. UCSF Chimera -I -Introduction. Sgro, J.-Y. (2017). https://static-bcrf.biochem.wisc.edu/tutorials/chimera/chimera_i_introduction.pdf
3. RCSB PDB: Homepage. (2010). Rcsb.org. https://www.rcsb.org/
4. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., & Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics, 7(1), 339. https://doi.org/10.1186/1471-2105-7-339
5. Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of Insulin in Health and Disease: An Update. International journal of molecular sciences, 22(12), 6403. https://doi.org/10.3390/ijms22126403
6. Weiss M, Steiner DF, Philipson LH. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. [Updated 2014 Feb 1]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279029/

DATE: 12/02/2024

WEBLEM 13
BINDING POCKET PREDICTION

INTRODUCTION:

Predicting binding pockets refers to identifying the regions on a protein surface, where other molecules, such as ligands or substrates, bind. This prediction is crucial for understanding protein function and designing drugs. Various computational methods are used, including geometric, energetic, and machine learning approaches, to predict binding pockets based on factors like amino acid composition, structural features, and interactions with solvent molecules. These predictions aid in drug discovery and protein engineering efforts. Binding pockets are essential for understanding protein function, as they are the sites where other molecules, such as small molecules, ions, or other proteins, interact with the protein. These interactions often play a crucial role in biological processes, such as enzyme-substrate binding, protein-protein interactions, and signal transduction pathways. Predicting binding pockets helps in understanding these interactions and can aid in drug design and protein engineering.

Predicting which molecules can bind to a given binding site of a protein with known 3D structure is important to decipher the protein function, and useful in drug design. A classical assumption in structural biology is that proteins with similar 3D structures have related molecular functions, and therefore may bind similar ligands. A cavity on the surface or in the interior of a protein that possesses suitable properties for binding a ligand is usually referred to as a binding pocket. The set of amino acid residues around a binding pocket determines its physicochemical characteristics and, together with its shape and location in a protein, defines its functionality. Residues outside the binding site can also have a long-range effect on the properties of the binding pocket. Cavities with similar functionalities are often conserved across protein families. For example, enzyme active sites are usually concave surfaces that present amino acid residues in a suitable configuration for binding low molecular weight compounds.

Proteins perform their biological functions in biological processes mainly by interacting with other molecules such as other proteins, small molecules, DNAs and RNAs. Usually not all the residues on a protein surface participate in these interactions. Thus, identification of these functional sites is of great importance to understanding the function of a protein and the mechanism of the interactions. In addition, knowledge of these functional sites can be used to guide the mutagenesis experiments. There exist a number of cavities or pockets on protein surface where small molecules bind. Therefore, identification of such cavities is often the starting point in protein–ligand binding site prediction for protein function annotation and structure-based drug design. Proper ligand binding site detection is a prerequisite for protein–ligand docking and high-throughput virtual screening to identify drug candidates in drug discovery processes. Many computational algorithms and tools have been developed in last two decades to identify pocket for protein–ligand binding site prediction.

A. CASTp:

Computed Atlas of Surface Topography of proteins (CASTp) provides an online resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins. These include pockets located on protein surfaces and voids buried in the interior of proteins. The measurement includes the area and volume of pocket or void by solvent accessible surface model (Richards' surface) and by molecular surface model (Connolly's surface), all calculated analytically. CASTp can be used to study surface features and functional regions of proteins. CASTp includes a graphical user interface, flexible interactive

visualization, as well as on-the-fly calculation for user uploaded structures. The CASTp web server aims to provide a comprehensive and detailed quantitative characterization of interior voids and surface pockets of proteins, which are prominent concave regions of proteins that are frequently associated with binding events. CASTp is based on the alpha shape and the pocket algorithm developed in computational geometry.
The CASTp web server aims to provide a comprehensive and detailed quantitative characterization of interior voids and surface pockets of proteins, which are prominent concave regions of proteins that are frequently associated with binding events. CASTp is based on the alpha shape and the pocket algorithm developed in computational geometry. In CASTp, voids are defined as buried unfilled empty space inside proteins after removing all hetero atoms that are inaccessible to water molecules (modeled as a spherical probe of 1.4 Å) from outside. Pockets are defined as concave caverns with constrictions at the opening on the surface regions of proteins. Unlike voids, pockets allow easy access of water probes from the outside.
CASTp identifies all pockets and voids on a protein structure and provides detailed delineation of all atoms participating in their formation. It also measures the volume and area of each pocket and void analytically, using both the solvent accessible surface model (Richards' surface) and molecular surface model (Connolly's surface). In addition, it measures the size of mouth openings of individual pockets, which helps to assess the accessibility of binding sites to various ligands and substrates. CASTp computation has been shown to be useful in a number of biological studies.

B. NetNGlyc 1.0:

N-linked glycosylation, is the attachment of an oligosaccharide, a carbohydrate consisting of several sugar molecules, sometimes also referred to as glycan, to a nitrogen atom (the amide nitrogen of an asparagine (Asn) residue of a protein), in a process called N-glycosylation, studied in biochemistry. The N-glycosylation process includes two principal phases the assembly of a lipid-linked oligosaccharide (LLO) and the transfer of the oligosaccharide to selected asparagine residues of polypeptide chains. In all eukaryotes, N-glycosylation is obligatory for viability. Glycosylation is a ubiquitous modification of newly synthesized proteins in the endoplasmic reticulum (ER). Dependent on the linkage of the oligosaccharide to the amino acid side chain of the protein.

Glycosylation is an important post-translational modification, and is known to influence protein folding, localisation and trafficking, protein solubility, antigenicity, biological activity and half-life, as well as cell-cell interactions. We investigate the spread of known and predicted N-glycosylation sites across functional categories of the human proteome. It functions by modifying appropriate asparagine residues of proteins with oligosaccharide structures, thus influencing their properties and bioactivities. Protein glycosylation, a general posttranslational modification of proteins involved in cell membrane formation, is crucial to dictate proper conformation of many membrane proteins, retain stability on some secreted glycoproteins, and play a role in cell–cell adhesion. The NetNGlyc 1.0 server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons.

C. NetPhos 3.1:

Protein phosphorylation is a major post-translational modification (PTM), occurring when a phosphate group bonds with specific amino acids (such as serine, threonine and tyrosine). Numerous experimental studies have demonstrated that phosphorylation is involved in regulation of a variety of fundamental cellular processes, such as protein-protein interaction,

protein degradation, signal transduction and signalling pathways and also multitude of cellular signalling pathways. Therefore, it is crucial to accurately identify human phosphorylation sites and to further characterise their biological functions. Protein phosphorylation is a mechanism of regulation that is extremely important in most cellular processes such as protein synthesis, cell division, signal transduction, cell growth, development and aging as many enzymes and receptors are activated and deactivated via phosphorylation/dephosphorylation events.

NetPhos3.1 is a online server which predicts and identifies the phosphorylation sites present in the specific proteins. 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. hence, we can say that NetPhos3.1 is a integrated version of NetPhos2.1.0 and 2.0 it is a neural network based server.

Predictions are made for 17 different kinases which include ATM, CKI, CKII, CaM-II, DNAPK, EGFR, GSK3, INSR, PKA, PKB, PKC, PKG, RSK, SRC, cdc2, cdk5 and p38MAPK

REFERENCES:

1. Lakowski R.A., Luscombe,N.M., Swindells,M.B. and Thornton,J.M. (1996) Protein clefts in molecular recognition and function. Protein Sci., 5, 2438–2452.
2. Liang J., Edelsbrunner,H. and Woodward,C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci., 7, 1884–1897.
3. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. 2002:310-22. https://services.healthtech.dtu.dk/services/NetNGlyc-1.0
4. Biswas AK, Noman N, Sikder AR. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics. 2010 May 21; 11:273. Doi: 10.1186/1471-2105-11-273. PMID: 20492656; PMCID: PMC2887807.
5. Xu C, Ng DT. O-mannosylation: The other glycan player of ER quality control. *Semin Cell Dev Biol.* 2015; 41:129–134.

DATE: 12/02/2024

WEBLEM 13(A)
Computed Atlas of Surface Topography of Proteins(CASTp)
(URL: http://sts.bioe.uic.edu/)

AIM:

To predict binding pockets located on protein surface for query 'Insulin' (PDB ID: 1H59) using CASTp Tool.

INTRODUCTION:

A cavity on the surface or in the interior of a protein that possesses suitable properties for binding a ligand is usually referred to as a binding pocket. The set of amino acid residues around a binding pocket determines its physicochemical characteristics and, together with its shape and location in a protein, defines its functionality. Residues outside the binding site can also have a long-range effect on the properties of the binding pocket. Cavities with similar functionalities are often conserved across protein families. For example, enzyme active sites are usually concave surfaces that present amino acid residues in a suitable configuration for binding low molecular weight compounds.

Computed Atlas of Surface Topography of proteins (CASTp) provides an online resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins. These include pockets located on protein surfaces and voids buried in the interior of proteins. The measurement includes the area and volume of pocket or void by solvent accessible surface model (Richards' surface) and by molecular surface model (Connolly's surface), all calculated analytically. CASTp can be used to study surface features and functional regions of proteins. CASTp includes a graphical user interface, flexible interactive visualization, as well as on-the-fly calculation for user uploaded structures. The CASTp web server aims to provide a comprehensive and detailed quantitative characterization of interior voids and surface pockets of proteins, which are prominent concave regions of proteins that are frequently associated with binding events. CASTp is based on the alpha shape and the pocket algorithm developed in computational geometry.

Insulin:

Insulin is a vital hormone that regulates blood sugar levels by allowing glucose to enter cells for energy production. It is produced by beta cells in the pancreas and plays a crucial role in glucose storage and production. Insulin was first isolated in 1921 by Canadian scientists Frederick G. Banting and Charles H. Best, leading to life-saving treatments for diabetes. In diabetes, either the body does not produce enough insulin (Type 1) or becomes resistant to its effects (Type 2). Insulin resistance can lead to high blood sugar levels and various health complications. Different types of insulin, including fast, intermediate, and long-acting insulins, are used based on individual needs to manage blood sugar effectively.

Insulin is a composed of two peptide chains, an A chain and a B chain, linked together by disulfide bonds. The A chain consists of 21 amino acids, while the B chain has 30 amino acids. Within the A chain, there is an additional disulfide bond. Insulin molecules have a tendency to form dimers in solution and can associate into hexamers in the presence of zinc ions. The amino acid sequence of insulin is highly conserved among species, with minor variations. Despite these variations, insulin from different species can be biologically active across species. The structure of insulin allows it to regulate blood glucose levels by promoting glucose storage and inhibiting glucose production and release by the liver.

METHODOLGY:

1. Open Homepage of CASTp tool (Computed Atlas of Surface Topography of Proteins).
2. Paste the PDB ID: 1H59 from the PDB (Protein data bank) database for query 'Insulin' protein.
3. Search with PDB ID and observe the results.

OBSERVATIONS:

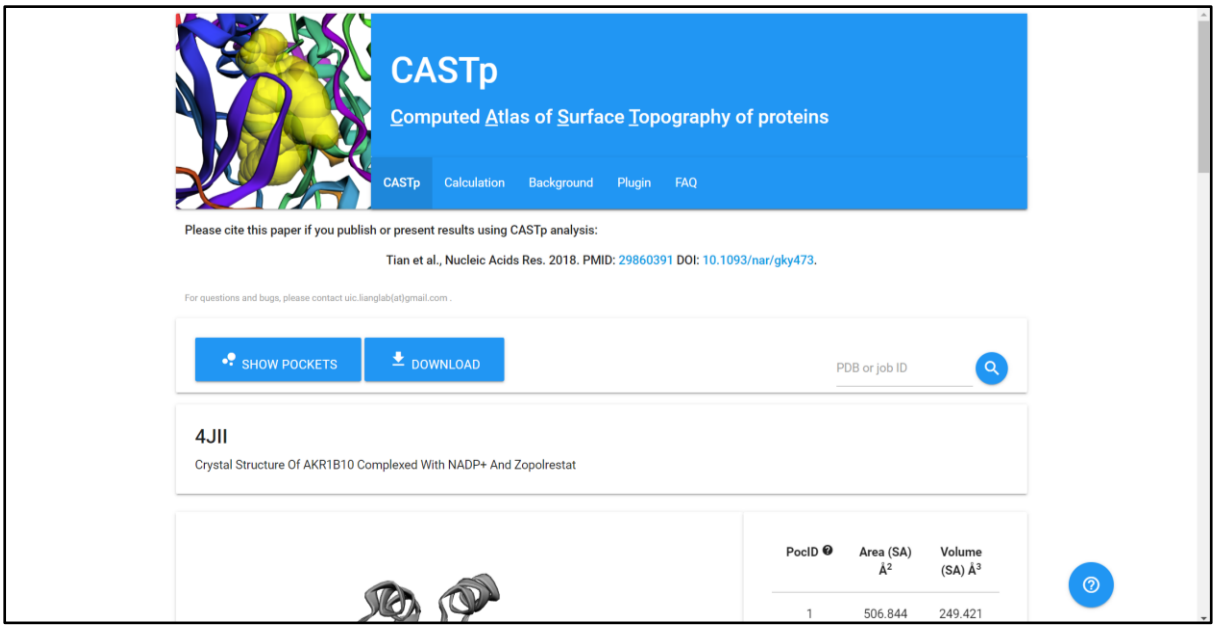


Fig 1: Homepage of CASTp tool

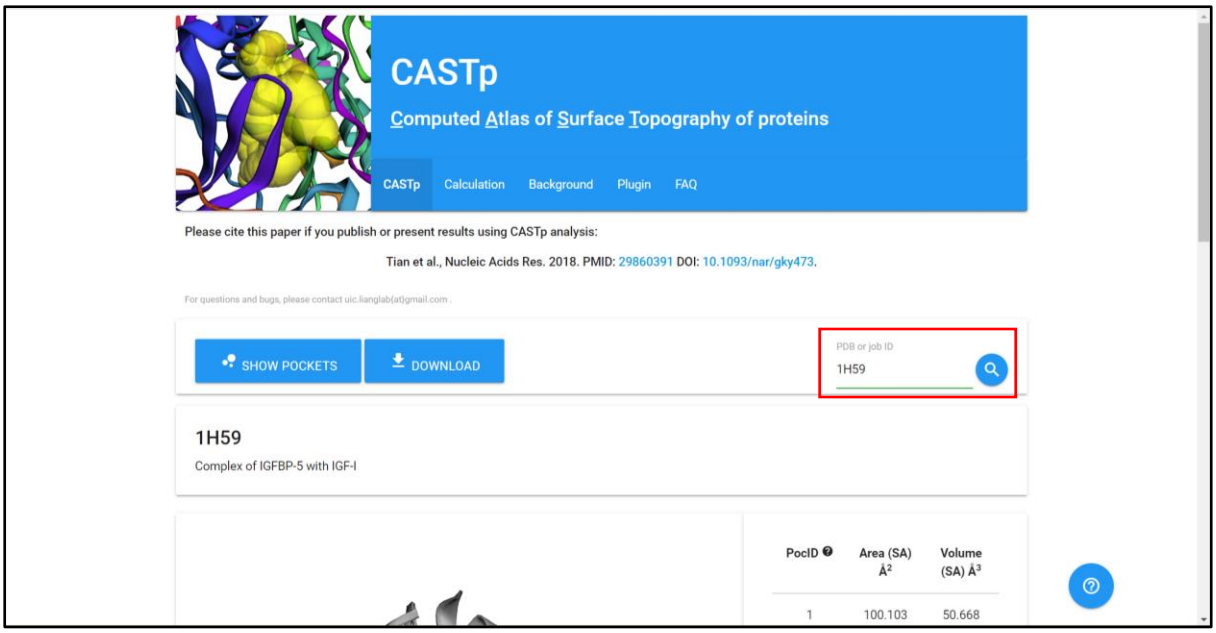


Fig 2: Search for query 'Insulin' (PDB ID: 1H59)

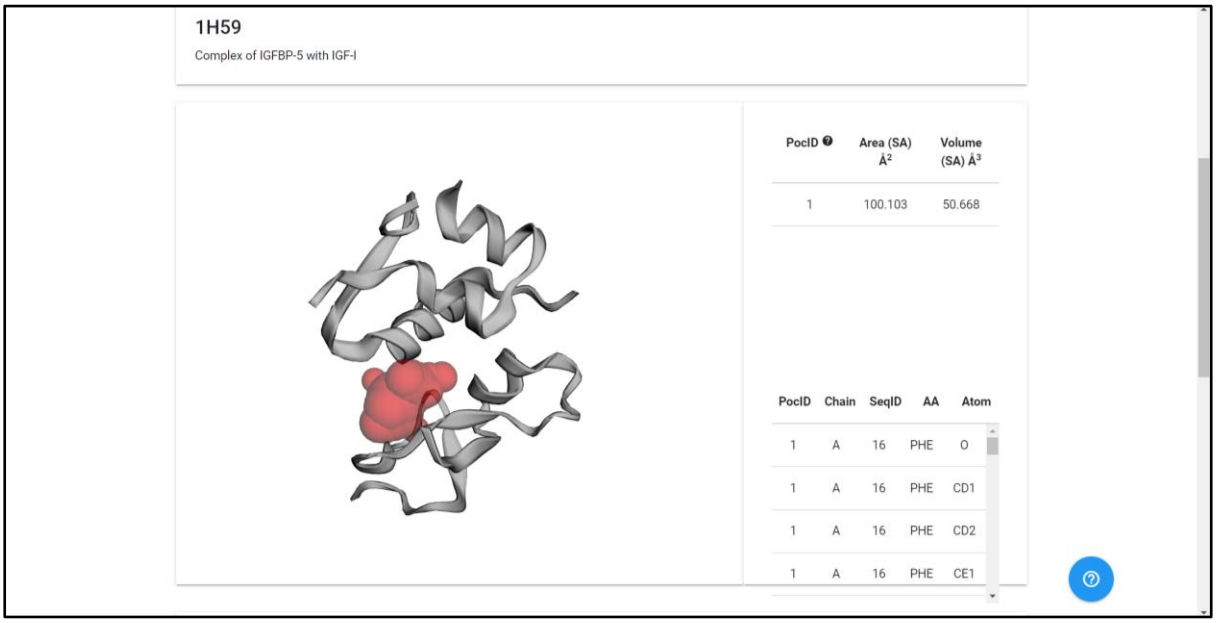


Fig 3: Result page of query 'Insulin' (PDB ID: 1H59)

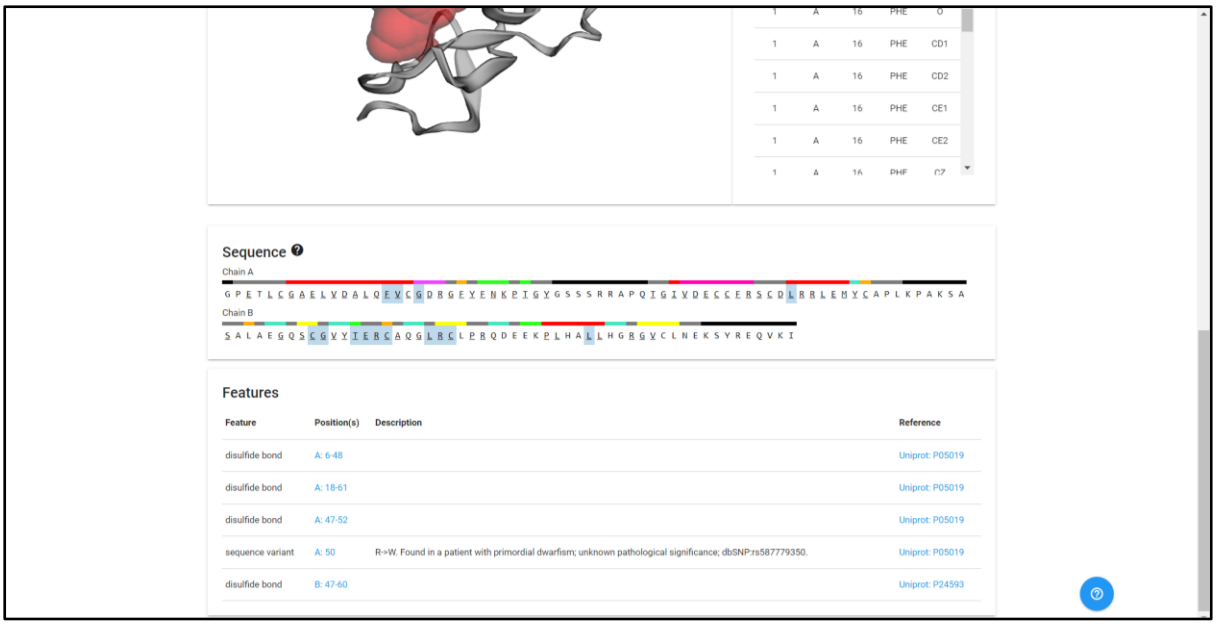


Fig 4: Chains of Sequence with actual amino acid coding

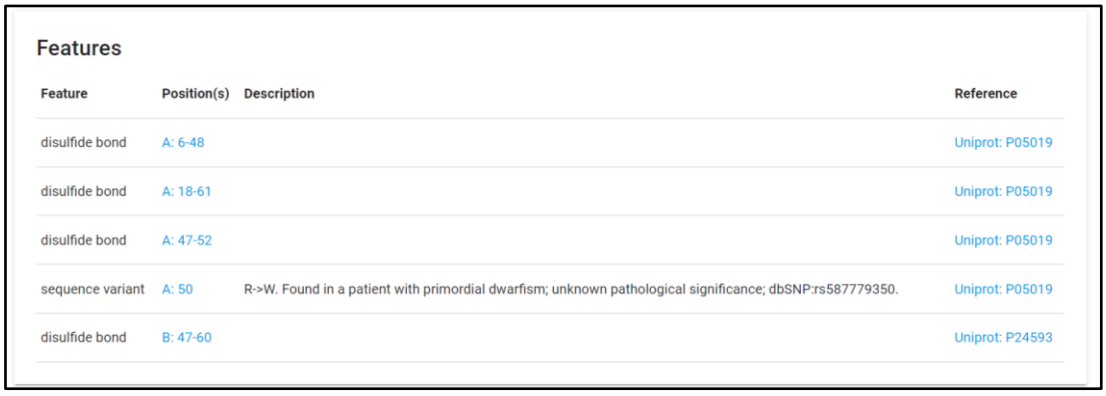


Fig 5: Features, Position, Reference of protein structure 'Insulin' (PDB ID: 1H59)

RESULTS:

By exploring CASTp tool for the query 'Insulin' (PDB ID: 1H59), the binding pocket and sub binding pockets were predicted with different chains amino acids and their atoms. The chains with underlined base pair were coded with actual amino acid on surface topology of protein. The features, position and references were also predicted and studied.

CONCLUSION:

By exploring CASTp tool, binding pockets located on protein surface topography were predicted for query 'Insulin' (PDB ID: 1H59).

REFERENCES:

1. Wei Tian, Chang Chen, Xue Lei, Jieling Zhao, Jie Liang, CASTp 3.0: computed atlas of surface topography of proteins, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W363–W367, https://doi.org/10.1093/nar/gky473
2. Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of Insulin in Health and Disease: An Update. International journal of molecular sciences, 22(12), 6403. https://doi.org/10.3390/ijms22126403.
3. Weiss M, Steiner DF, Philipson LH. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. [Updated 2014 Feb 1]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279029/

DATE: 12/02/2024

WEBLEM 13(B)
N-linked glycosylation sites in human proteins (NetNGlyc – 1.0)
(URL:https://services.healthtech.dtu.dk/services/NetNGlyc-1.0)

AIM:

To predict and study the N-glycosylation sites present in the protein for query 'Insulin' (UniProt ID: P06213) using NetNGlyc – 1.0 tool.

INTRODUCTION:

N-linked glycosylation, is the attachment of an oligosaccharide, a carbohydrate consisting of several sugar molecules, sometimes also referred to as glycan, to a nitrogen atom (the amide nitrogen of an asparagine (Asn) residue of a protein), in a process called N-glycosylation, studied in biochemistry. The N-glycosylation process includes two principal phases the assembly of a lipid-linked oligosaccharide (LLO) and the transfer of the oligosaccharide to selected asparagine residues of polypeptide chains. In all eukaryotes, N-glycosylation is obligatory for viability.

Glycosylation is an important post-translational modification, and is known to influence protein folding, localisation and trafficking, protein solubility, antigenicity, biological activity and half-life, as well as cell-cell interactions. We investigate the spread of known and predicted N-glycosylation sites across functional categories of the human proteome.It functions by modifying appropriate asparagine residues of proteins with oligosaccharide structures, thus influencing their properties and bioactivities. Protein glycosylation, a general posttranslational modification of proteins involved in cell membrane formation, is crucial to dictate proper conformation of many membrane proteins, retain stability on some secreted glycoproteins, and play a role in cell–cell adhesion.

The NetNglyc server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons.

Insulin:

Insulin is a vital hormone that regulates blood sugar levels by allowing glucose to enter cells for energy production. It is produced by beta cells in the pancreas and plays a crucial role in glucose storage and production. Insulin was first isolated in 1921 by Canadian scientists Frederick G. Banting and Charles H. Best, leading to life-saving treatments for diabetes. In diabetes, either the body does not produce enough insulin (Type 1) or becomes resistant to its effects (Type 2). Insulin resistance can lead to high blood sugar levels and various health complications. Different types of insulin, including fast, intermediate, and long-acting insulins, are used based on individual needs to manage blood sugar effectively.

Insulin is a composed of two peptide chains, an A chain and a B chain, linked together by disulfide bonds. The A chain consists of 21 amino acids, while the B chain has 30 amino acids. Within the A chain, there is an additional disulfide bond. Insulin molecules have a tendency to

form dimers in solution and can associate into hexamers in the presence of zinc ions. The amino acid sequence of insulin is highly conserved among species, with minor variations. Despite these variations, insulin from different species can be biologically active across species. The structure of insulin allows it to regulate blood glucose levels by promoting glucose storage and inhibiting glucose production and release by the liver.

METHODOLOGY:

1. Go to UniProt database and search query ‘Insulin’ in search.
2. Download the FASTA(Canonical) sequence of the protein and copy the query ‘Insulin’ (UniProt ID: P06213) protein sequence.
3. Open Homepage Page of N-linked glycosylation sites in human proteins (NetNglyc-1.0 server) and paste the protein sequence.
4. Click on ‘Send file’.
5. Interpret the results displayed.

OBSERVATIONS:

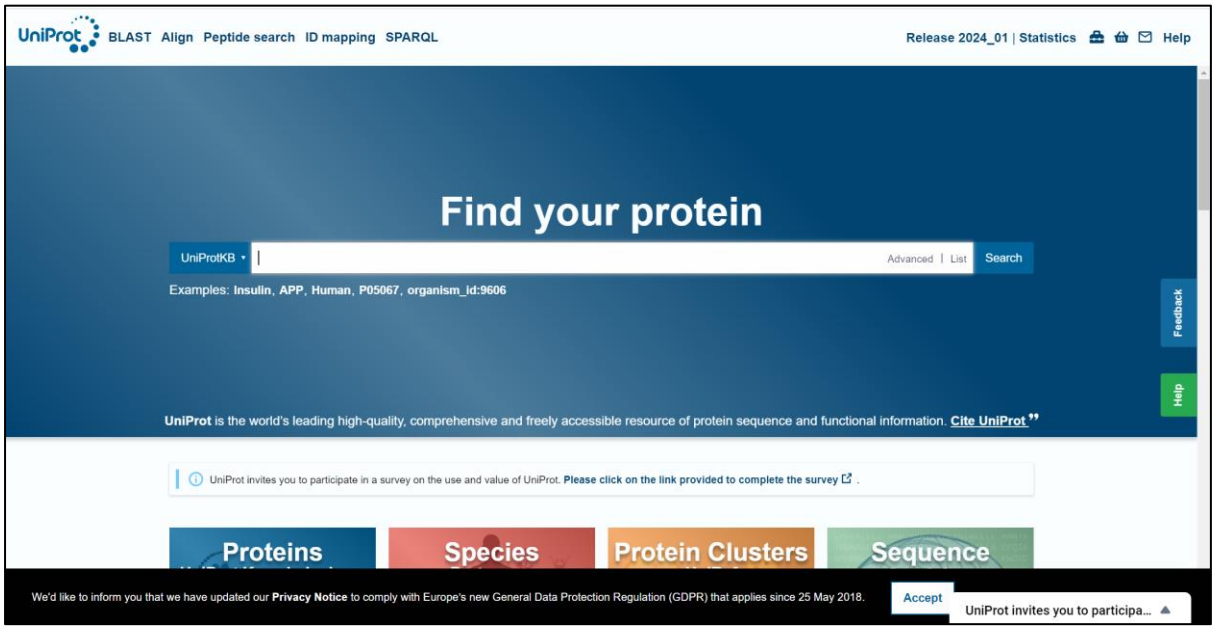


Fig 1: Homepage of UniProt Database

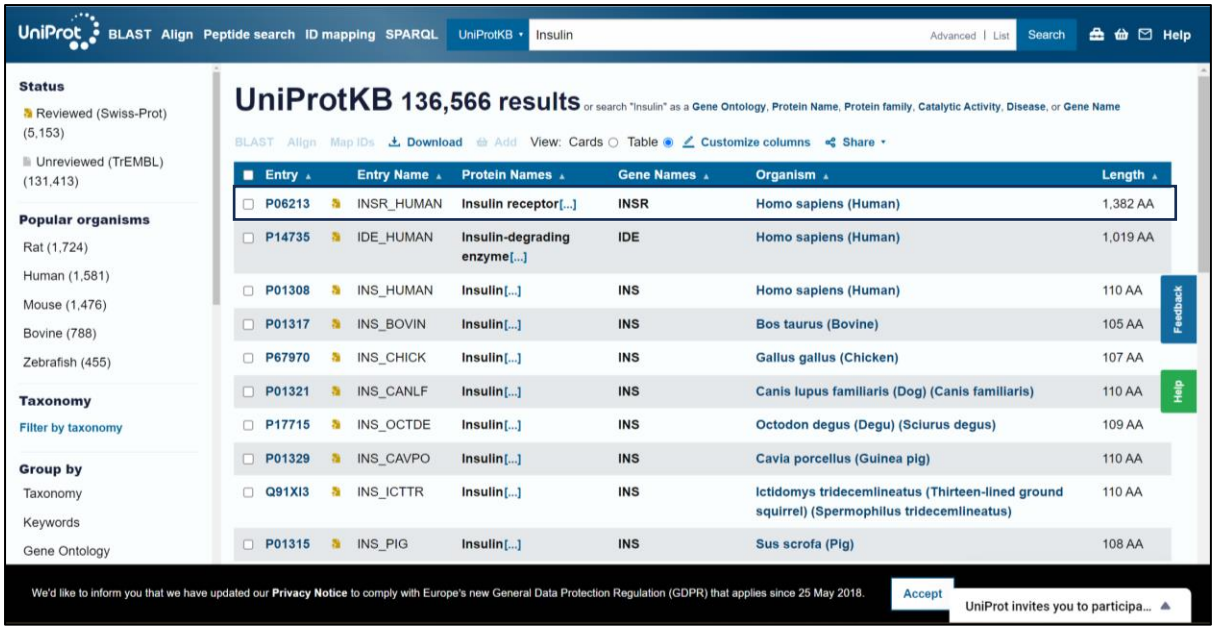


Fig 2: 'Insulin' (UniProt ID: P06213) selected

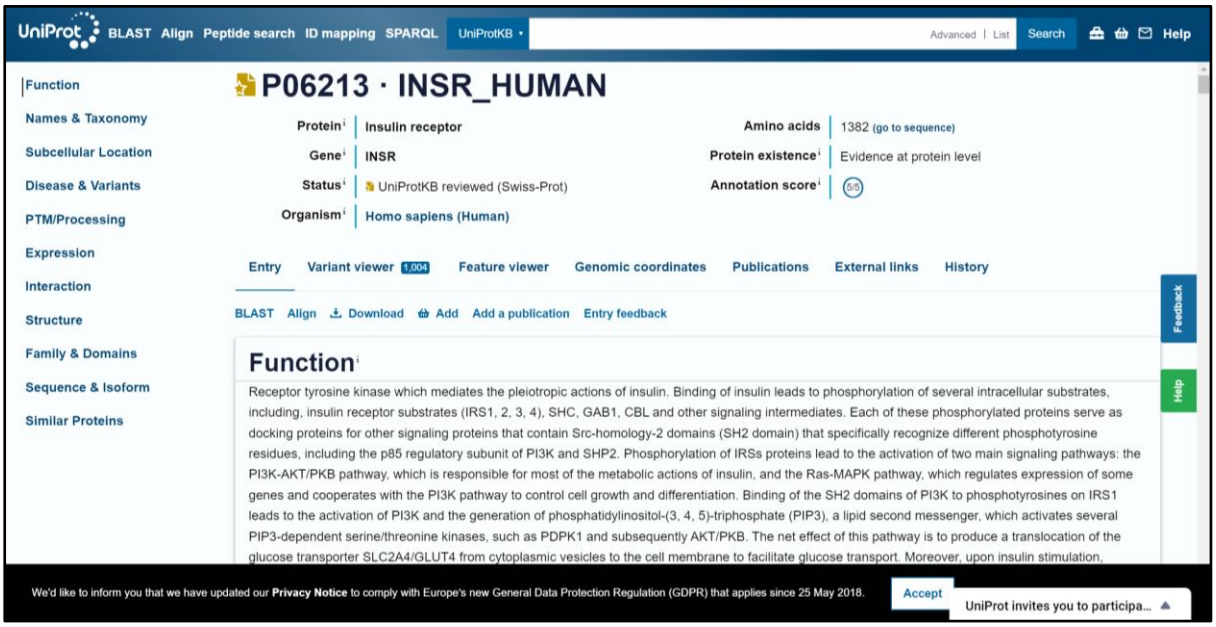


Fig 3: Information for the query 'Insulin' (UniProt ID: P06213)

```
>sp|P06213|INSR_HUMAN Insulin receptor OS=Homo sapiens OX=9606 GN=INSR PE=1 SV=4
MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
QILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFPNLTVIRGSRLFFNYAL
VIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRILDSVEDNYIVLNKDDNE
ECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTICKSHGCTAEGLCCHSECL
GNCSQPDDPTKCVACRNFYLDGRCVETCPPPYYHFQDWRCVNFSFCQDLHHKCKNSRRQG
CHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKVCHLLEGEKTIDSVTSAQELRGC
TVINGSLIINIRGGNNLAAELEANLGLIEEISGYLKIRRSYALVSLSFFRKLRLIRGETL
EIGNYSFYALDNQNLRQLWDWSKHNLTITQGKLFFHYNPKLCLSEIHKMEEVSGTKGRQE
RNDIALKTNGDQASCENELLKFSYIRTSFDKILLRWEPYWPPDFRDLLGFMLFYKEAPYQ
NVTEFDGQDACGSNSWTVVDIDPPLRSNDPKSQNHPGWLMRGLKPWTQYAIFVKTLVTFS
DERRTYGAKSDIIYVQTDATNPSVPLDPISVSNSSSQIILKWKPPSDPNGNITHYLVFWE
RQAEDSELFELDYCLKGLKLPSRTWSPPFESEDSQKHNQSEYEDSAGECCSCPKTDSQIL
KELEESSFRKTFEDYLHNVVFVPRKTSSGTGAEDPRPSRKRRSLGDVGNVTVAVPTVAAF
PNTSSTSVPTSPEEHRPFEKVVNKESLVISGLRHFTGYRIELQACNQDTPEERCSVAAYV
SARTMPEAKADDIVGPVTHEIFENNVVHLMWQEPKEPNGLIVLYEVSYRRYGDEELHLCV
SRKHFALERGCRLRGLSPGNYSVRIRATSLAGNGSWTEPTYFYVTDYLDVPSNIAKIIIG
PLIFVFLFSVVIGSIYLFLRKRQPDGPLGPLYASSNPEYLSASDVFPCSVYVPDEWEVSR
EKITLLRELGQGSFGMVYEGNARDIIKGEAETRVAVKTVNESASLRERIEFLNEASVMKG
FTCHHVVRLLGVVSKGQPTLVVMELMAHGDLKSYLRSLRPEAENNPGRPPPTLQEMIQMA
AEIADGMAYLNAKKFVHRDLAARNCMVAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPV
RWMAPESLKDGVFTTSSDMWSFGVVLWEITSLAEQPYQGLSNEQVLKFVMDGGYLDQPDN
CPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEME
FEDMENVPLDRSSHCQREEAGGRDGGSSLGFKRSYEEHIPYTHMNGGKKNGRILTLPRSN
PS
```

Fig 4: Downloading the FASTA (Canonical) sequence for the query 'Insulin' (UniProt ID: P06213)

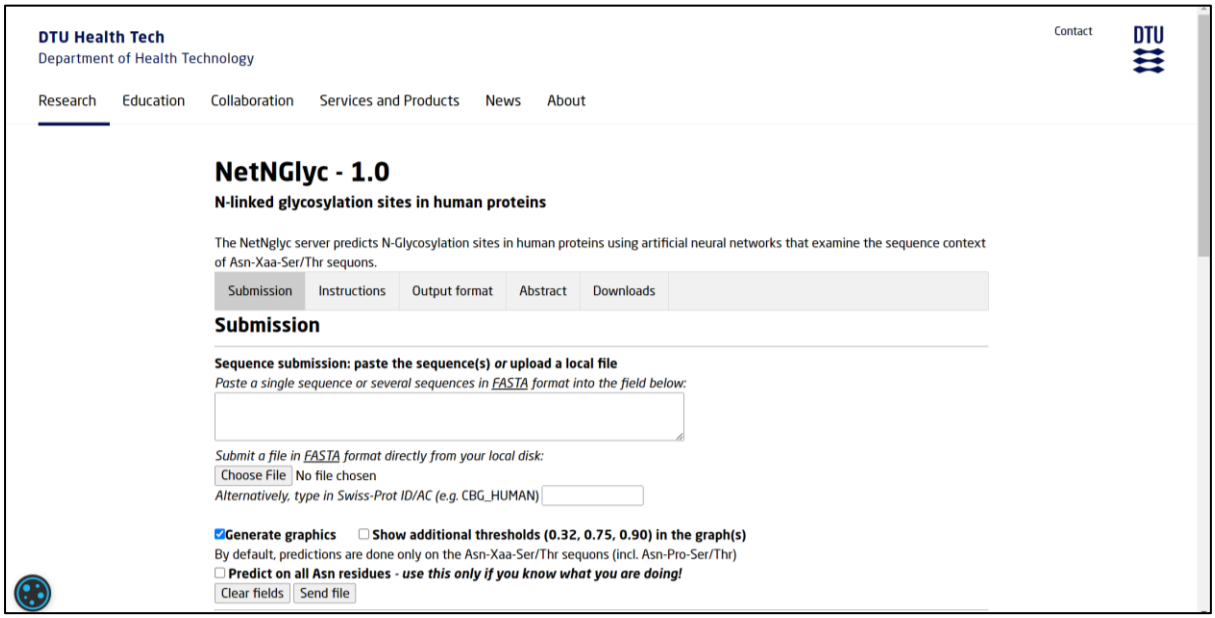


Fig 5: Homepage of NetNGlyc – 1.0 Server

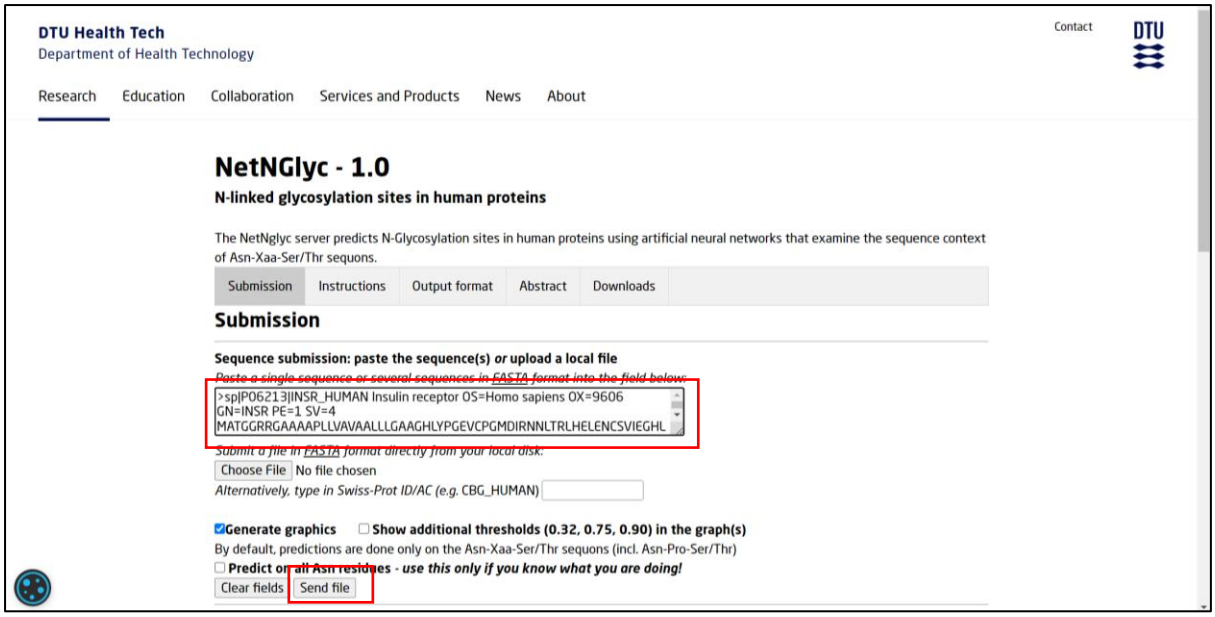


Fig 6: Pasting the protein sequence in the NetNGlyc – 1.0 Server

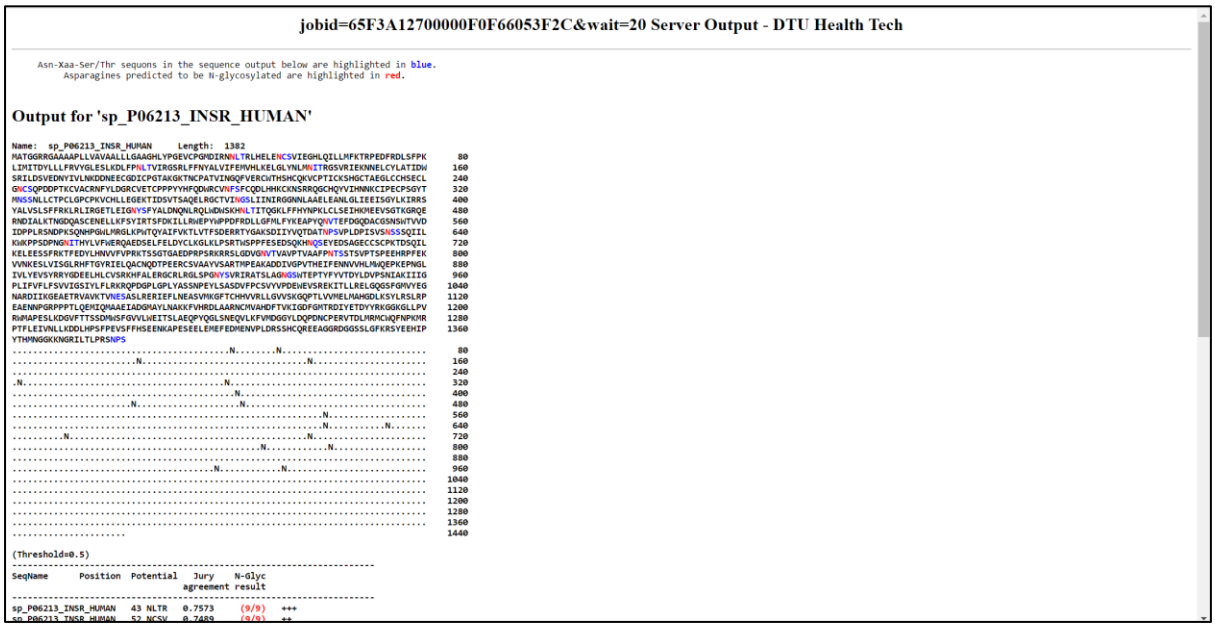
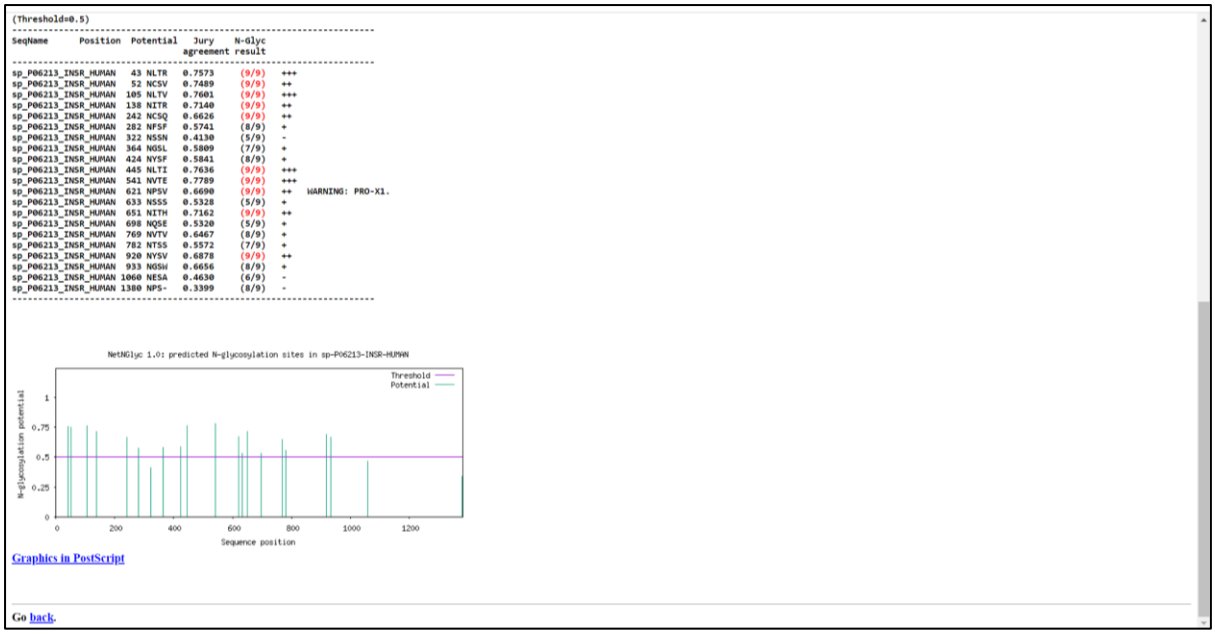


Fig 7: Result page for the protein query sequence 'Insulin' (UniProt ID: P06213)



(Threshold=0.5)

SeqName	Position	Potential	Jury agreement	N-Glyc result	
sp_P06213_INSR_HUMAN	43	NLTR	0.7573	(9/9)	+++
sp_P06213_INSR_HUMAN	52	NCSV	0.7489	(9/9)	++
sp_P06213_INSR_HUMAN	105	NLTV	0.7601	(9/9)	+++
sp_P06213_INSR_HUMAN	138	NITR	0.7140	(9/9)	++
sp_P06213_INSR_HUMAN	242	NCSQ	0.6626	(9/9)	++
sp_P06213_INSR_HUMAN	282	NFSF	0.5741	(8/9)	+
sp_P06213_INSR_HUMAN	322	NSSN	0.4130	(5/9)	-
sp_P06213_INSR_HUMAN	364	NGSL	0.5809	(7/9)	+
sp_P06213_INSR_HUMAN	424	NYSF	0.5841	(8/9)	+
sp_P06213_INSR_HUMAN	445	NLTI	0.7636	(9/9)	+++
sp_P06213_INSR_HUMAN	541	NVTE	0.7789	(9/9)	+++
sp_P06213_INSR_HUMAN	621	NPSV	0.6690	(9/9)	++ WARNING: PRO-X1.
sp_P06213_INSR_HUMAN	633	NSSS	0.5328	(5/9)	+
sp_P06213_INSR_HUMAN	651	NITH	0.7162	(9/9)	++
sp_P06213_INSR_HUMAN	698	NQSE	0.5320	(5/9)	+
sp_P06213_INSR_HUMAN	769	NVTV	0.6467	(8/9)	+
sp_P06213_INSR_HUMAN	782	NTSS	0.5572	(7/9)	+
sp_P06213_INSR_HUMAN	920	NYSV	0.6878	(9/9)	++
sp_P06213_INSR_HUMAN	933	NGSW	0.6656	(8/9)	+
sp_P06213_INSR_HUMAN	1060	NESA	0.4630	(6/9)	-
sp_P06213_INSR_HUMAN	1380	NPS-	0.3399	(8/9)	-

Fig 7.1: Result page for the protein query sequence 'Insulin' (UniProt ID: P06213)

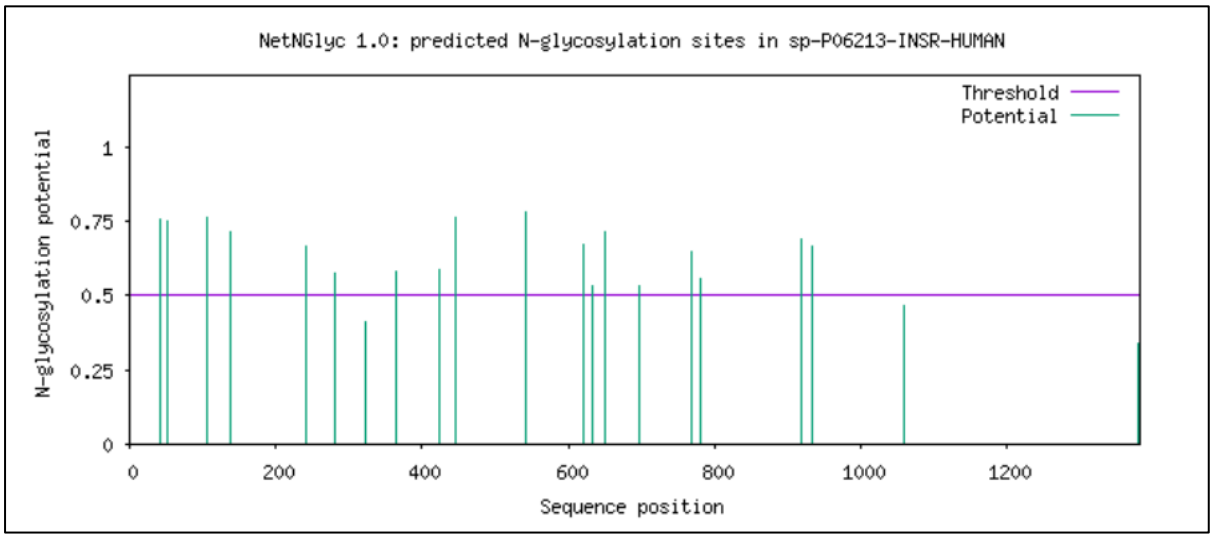


Fig 8: Graphical Representation of N-glycosylation potential.

RESULTS:

Netglyc-1.0 server was explored the glycosylation sites present in the protein 'Insulin' (UniProt ID: P06213) were predicted for further studies with the threshold of 0.5. Ten sequences were predicted to be the best results and N-glycosylation potential graph was also observed and studied.

CONCLUSION:

Glycosylation site predicted and studied for the query 'Insulin' (UniProt ID: P06213) by exploring NetNGlyc-1.0 server.

REFERENCES:

1. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to proteinfunction. 2002:310-22. https://services.healthtech.dtu.dk/services/NetNGlyc-1.0
2. R.B. Brown, Jason E. Schaffer, in Encyclopedia of Biological Chemistry (Third Edition),2021.https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/n-linked-glycosylation
3. Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of Insulin in Health and Disease: An Update. International journal of molecular sciences, 22(12), 6403. https://doi.org/10.3390/ijms22126403
4. Weiss M, Steiner DF, Philipson LH. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. [Updated 2014 Feb 1]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279029/

DATE: 09/02/2024

WEBLEM 13(C)
Binding pocket prediction: Netphos3.1
(URL: https://services.healthtech.dtu.dk/services/NetPhos-3.1/)

AIM:

To predict and study the phosphorylation sites present in the protein 'Insulin' (UniProt ID: P06213) using Netphos3.1 program.

INTRODUCTION:

Protein phosphorylation is a major post-translational modification (PTM), occurring when a phosphate group bonds with specific amino acids (such as serine, threonine and tyrosine). Numerous experimental studies have demonstrated that phosphorylation is involved in regulation of a variety of fundamental cellular processes, such as protein-protein interaction, protein degradation, signal transduction and signalling pathways and also multitude of cellular signalling pathways. Therefore, it is crucial to accurately identify human phosphorylation sites and to further characterise their biological functions.

NetPhos3.1 is a online server which predicts and identifies the phosphorylation sites present in the specific proteins. 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. Hence, we can say that NetPhos3.1 is a integrated version of NetPhos2.1.0 and 2.0 it is a neural network based server.

Predictions are made for 17 different kinases which include ATM,CKI, CKII, CaM-II, DNAPK, EGFR, GSK3, INSR, PKA, PKB, PKC, PKG, RSK, SRC, cdc2, cdk5 and p38MAPK.

Insulin:

Insulin is a vital hormone that regulates blood sugar levels by allowing glucose to enter cells for energy production. It is produced by beta cells in the pancreas and plays a crucial role in glucose storage and production. Insulin was first isolated in 1921 by Canadian scientists Frederick G. Banting and Charles H. Best, leading to life-saving treatments for diabetes. In diabetes, either the body does not produce enough insulin (Type 1) or becomes resistant to its effects (Type 2). Insulin resistance can lead to high blood sugar levels and various health complications. Different types of insulin, including fast, intermediate, and long-acting insulins, are used based on individual needs to manage blood sugar effectively.

Insulin is a composed of two peptide chains, an A chain and a B chain, linked together by disulfide bonds. The A chain consists of 21 amino acids, while the B chain has 30 amino acids. Within the A chain, there is an additional disulfide bond. Insulin molecules have a tendency to form dimers in solution and can associate into hexamers in the presence of zinc ions. The amino acid sequence of insulin is highly conserved among species, with minor variations. Despite

these variations, insulin from different species can be biologically active across species. The structure of insulin allows it to regulate blood glucose levels by promoting glucose storage and inhibiting glucose production and release by the liver.

METHODOLOGY:

1. Open UniProt database and enter the protein name ‘Insulin’.
2. Download the FASTA sequence of the protein ‘Insulin’ (UniProt ID: P06213).
3. Open NetPhos3.1 server and paste the protein sequence.
4. Interpret the results displayed.

OBSERVATIONS:

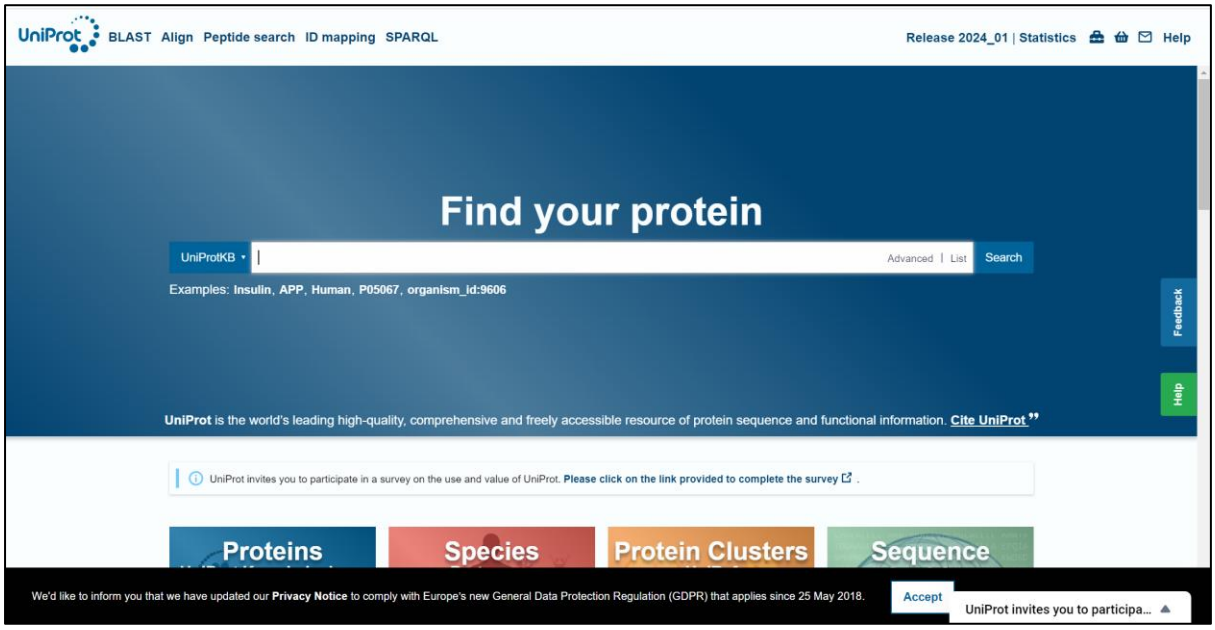


Fig 1: Homepage of UniProt Database

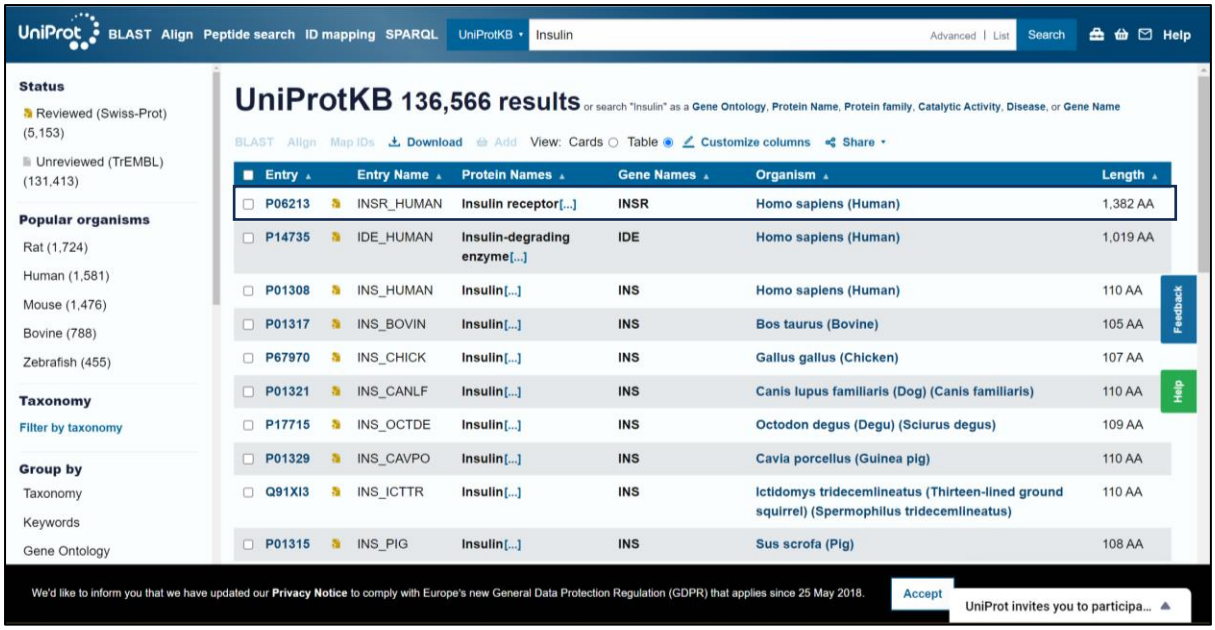


Fig 2: 'Insulin' (UniProt ID: P06213) selected

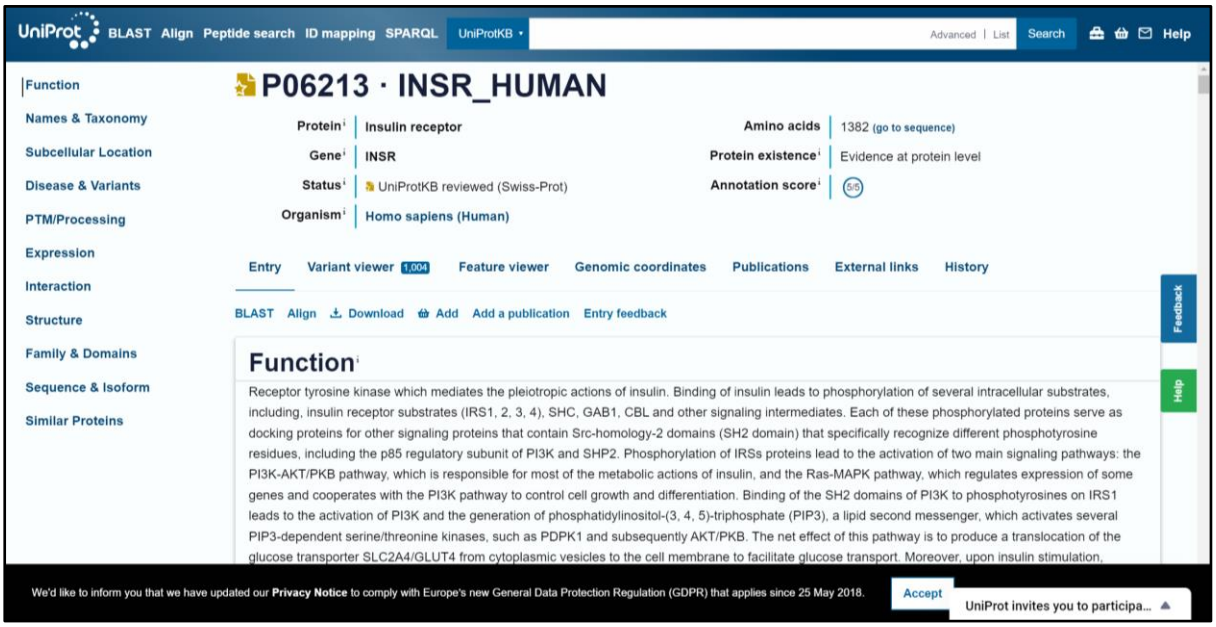


Fig 3: Information for the query 'Insulin' (UniProt ID: P06213)

```
>sp|P06213|INSR_HUMAN Insulin receptor OS=Homo sapiens OX=9606 GN=INSR PE=1 SV=4
MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
QILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFPNLTVIRGSRLFFNYAL
VIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRILDSVEDNYIVLNKDDNE
ECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTICKSHGCTAEGLCCHSECL
GNCSQPDDPTKCVACRNFYLDGRCVETCPPPYYHFQDWRCVNFSFCQDLHHKCKNSRRQG
CHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKVCHLLEGEKTIDSVTSAQELRGC
TVINGSLIINIRGGNNLAAELEANLGLIEEISGYLKIRRSYALVSLSFFRKLRLIRGETL
EIGNYSFYALDNQNLRQLWDWSKHNLTITQGKLFFHYNPKLCLSEIHKMEEVSGTKGRQE
RNDIALKTNGDQASCENELLKFSYIRTSFDKILLRWEPYWPPDFRDLLGFMLFYKEAPYQ
NVTEFDGQDACGSNSWTVVDIDPPLRSNDPKSQNHPGWLMRGLKPWTQYAIFVKTLVTFS
DERRTYGAKSDIIYVQTDATNPSVPLDPISVSNSSSQIILKWKPPSDPNGNITHYLVFWE
RQAEDSELFELDYCLKGLKLPSRTWSPPFESEDSQKHNQSEYEDSAGECCSCPKTDSQIL
KELEESSFRKTFEDYLHNVVFVPRKTSSGTGAEDPRPSRKRRSLGDVGNVTVAVPTVAAF
PNTSSTSVPTSPEEHRPFEKVVNKESLVISGLRHFTGYRIELQACNQDTPEERCSVAAYV
SARTMPEAKADDIVGPVTHEIFENNVVHLMWQEPKEPNGLIVLYEVSYRRYGDEELHLCV
SRKHFALERGCRLRGLSPGNYSVRIRATSLAGNGSWTEPTYFYVTDYLDVPSNIAKIIIG
PLIFVFLFSVVIGSIYLFLRKRQPDGPLGPLYASSNPEYLSASDVFPCSVYVPDEWEVSR
EKITLLRELGQGSFGMVYEGNARDIIKGEAETRVAVKTVNESASLRERIEFLNEASVMKG
FTCHHVVRLLGVVSKGQPTLVVMELMAHGDLKSYLRSLRPEAENNPGRPPPTLQEMIQMA
AEIADGMAYLNAKKFVHRDLAARNCMVAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPV
RWMAPESLKDGVFTTSSDMWSFGVVLWEITSLAEQPYQGLSNEQVLKFVMDGGYLDQPDN
CPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEME
FEDMENVPLDRSSHCQREEAGGRDGGSSLGFKRSYEEHIPYTHMNGGKKNGRILTLPRSN
PS
```

Fig 4: Downloading the FASTA (Canonical) sequence for the query 'Insulin' (UniProt ID: P06213)

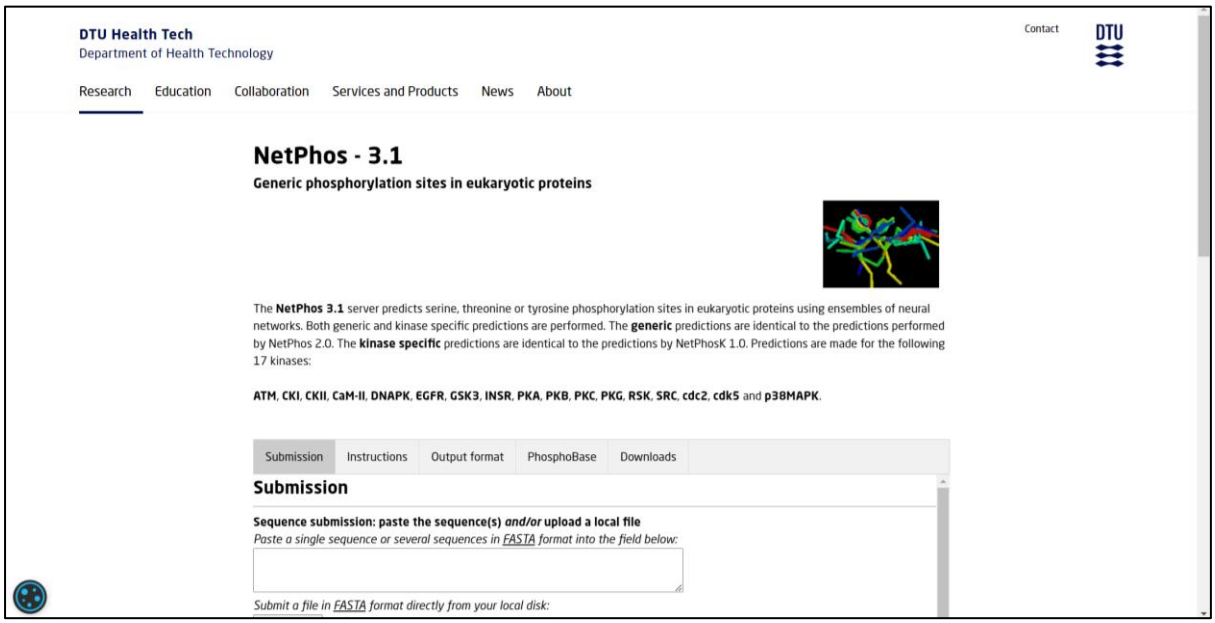


Fig 5: Homepage of NetPhos – 3.1 Server

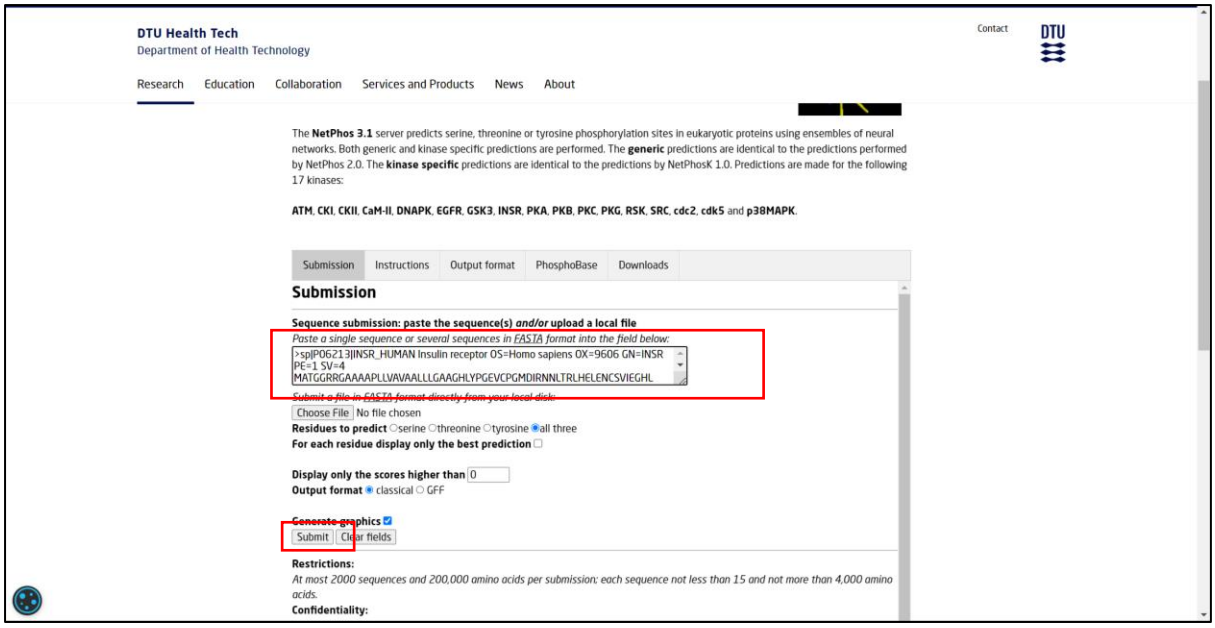
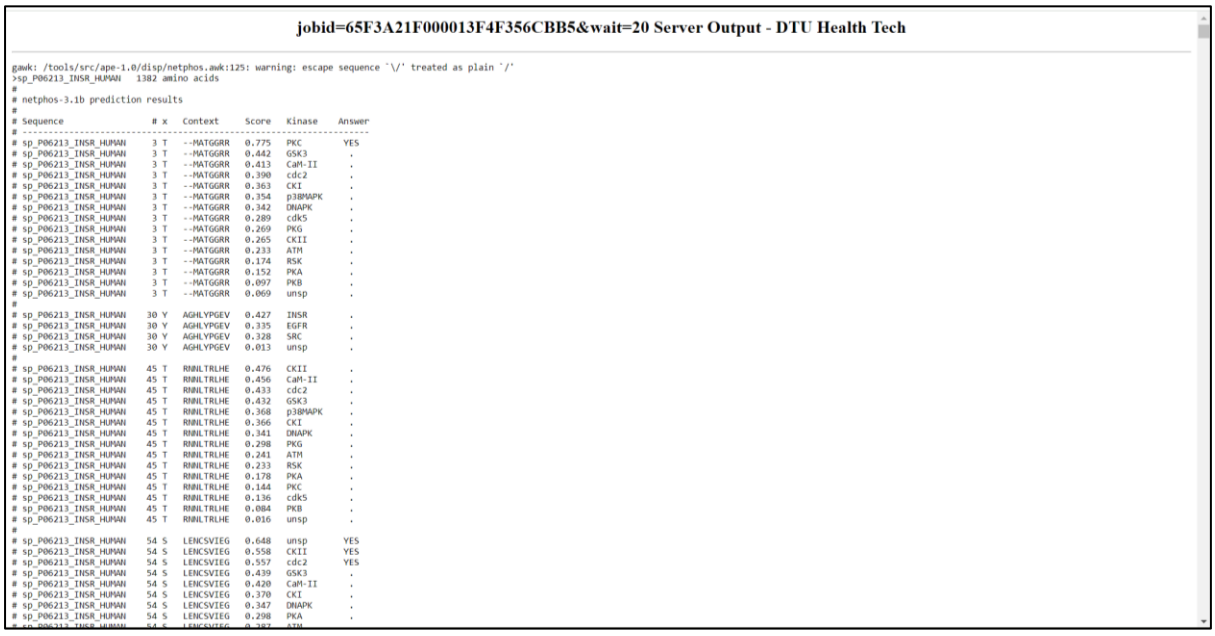


Fig 6: Pasting the protein sequence in the NetPhos – 3.1 Server



jobid=65F3A21F000013F4F356CBB5&wait=20 Server Output - DTU Health Tech

```
gawk: /tools/src/ape-1.0/disp/netphos.awk:125: warning: escape sequence `\/' treated as plain `/'
>sp_P06213_INSR_HUMAN   1382 amino acids
#
# netphos-3.1b prediction results
#
# Sequence                 # x   Context     Score   Kinase    Answer
# -------------------------------------------------------------------
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.775   PKC         YES
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.442   GSK3          .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.413   CaM-II        .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.390   cdc2          .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.363   CKI           .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.354   p38MAPK       .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.342   DNAPK         .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.289   cdk5          .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.269   PKG           .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.265   CKII          .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.233   ATM           .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.174   RSK           .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.152   PKA           .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.097   PKB           .
# sp_P06213_INSR_HUMAN       3 T   --MATGGRR   0.069   unsp          .
#
# sp_P06213_INSR_HUMAN      30 Y   AGHLYPGEV   0.427   INSR          .
# sp_P06213_INSR_HUMAN      30 Y   AGHLYPGEV   0.335   EGFR          .
# sp_P06213_INSR_HUMAN      30 Y   AGHLYPGEV   0.328   SRC           .
# sp_P06213_INSR_HUMAN      30 Y   AGHLYPGEV   0.013   unsp          .
#
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.476   CKII          .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.456   CaM-II        .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.433   cdc2          .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.432   GSK3          .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.368   p38MAPK       .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.366   CKI           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.341   DNAPK         .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.298   PKG           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.241   ATM           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.233   RSK           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.178   PKA           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.144   PKC           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.136   cdk5          .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.084   PKB           .
# sp_P06213_INSR_HUMAN      45 T   RNNLTRLHE   0.016   unsp          .
#
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.648   unsp        YES
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.558   CKII        YES
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.557   cdc2        YES
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.439   GSK3          .
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.420   CaM-II        .
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.370   CKI           .
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.347   DNAPK         .
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.298   PKA           .
# sp_P06213_INSR_HUMAN      54 S   LENCSVIEG   0.287   ATM           .
```

Fig 7: Result page for the protein query sequence 'Insulin' (UniProt ID: P06213)

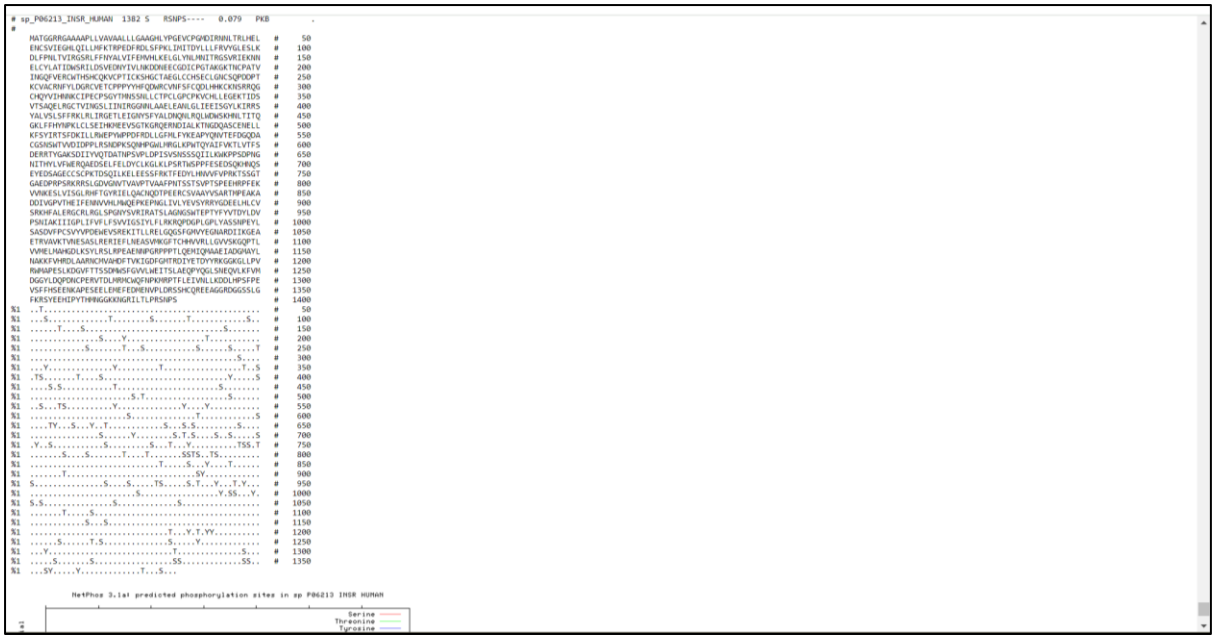


Fig 7.1: Result page for the protein query sequence 'Insulin' (UniProt ID: P06213)

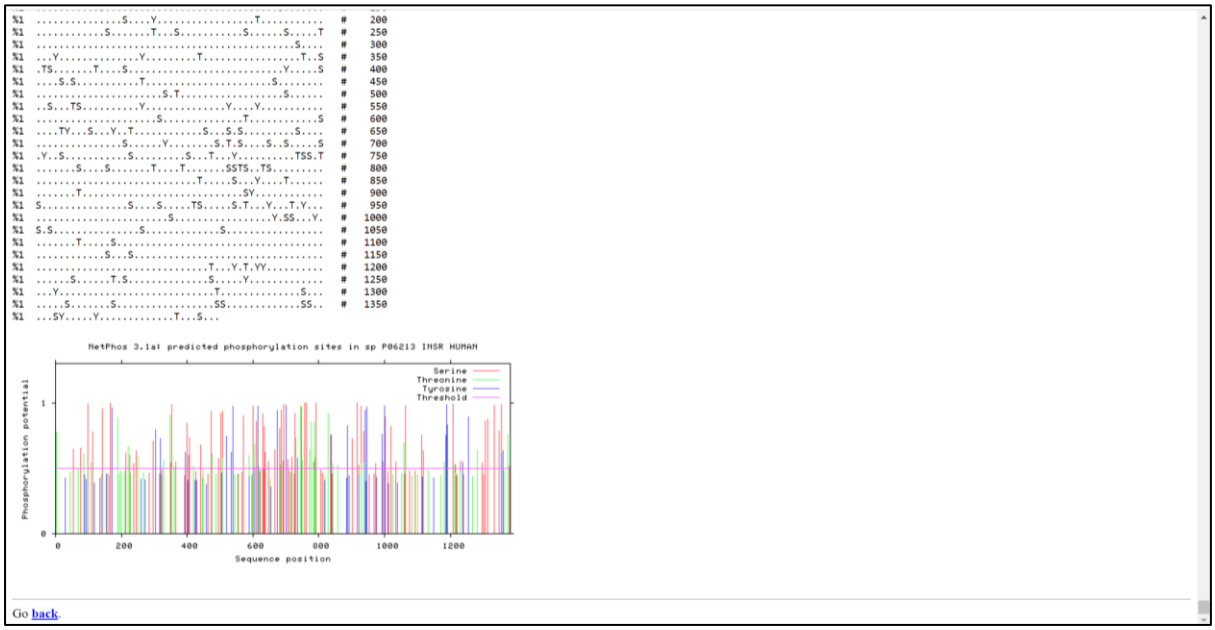


Fig 7.2: Result page for the protein query sequence 'Insulin' (UniProt ID: P06213)

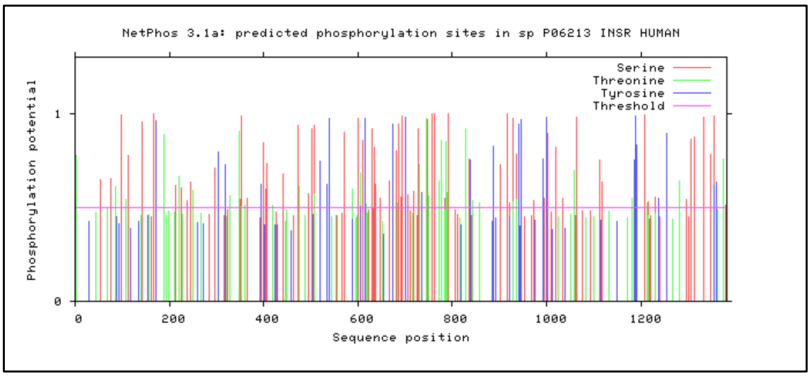


Fig 8: Graphical Representation of predicted phosphorylation sites

RESULTS:

NetPhos-3.1 server was explored to predict the phosphorylation sites present in the protein 'Insulin' (UniProt ID: P06213). The protein FASTA sequence was retrieved from the UniProt database and was submitted in the server. The amino acids which were having maximum score and showed "YES" were considered as phosphorylation site and would probably undergo phosphorylation.

CONCLUSION:

Phosphorylation sites were predicted and studied for the query 'Insulin' (UniProt ID: P06213) by exploring NetPhos-3.1 server.

REFERENCES:

1. Biswas AK, Noman N, Sikder AR. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics. 2010 May 21;11:273. doi: 10.1186/1471-2105-11-273. PMID: 20492656; PMCID: PMC2887807.
2. Fuyi Li, Chen Li, Tatiana T Marquez-Lago, André Leier, Tatsuya Akutsu, Anthony W Purcell, A Ian Smith, Trevor Lithgow, Roger J Daly, Jiangning Song, Kuo-Chen Chou, Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics, Volume 34, Issue 24, December 2018, Pages 4223–4231, https://doi.org/10.1093/bioinformatics/bty522
3. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999 Dec 17;294(5):1351-62. doi: 10.1006/jmbi.1999.3310. PMID: 10600390.
4. Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of Insulin in Health and Disease: An Update. International journal of molecular sciences, 22(12), 6403. https://doi.org/10.3390/ijms22126403
5. Weiss M, Steiner DF, Philipson LH. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. [Updated 2014 Feb 1]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279029/

DATE: 13/02/2024

WEBLEM 14
INTRODUCTION TO STRUCTURAL BLAST- VAST & DALI

INTRODUCTION:

Structural BLAST (Basic Local Alignment Search Tool) is a bioinformatics tool designed for the comparison of protein three-dimensional structures. While traditional BLAST is widely used for comparing sequences, Structural BLAST takes a step further by considering the spatial arrangement of amino acids within proteins. It allows researchers to search a protein structure database to identify proteins with similar folds or overall structural features.

The methodology behind Structural BLAST involves the representation of protein structures as mathematical vectors, capturing information about the position of amino acids in three-dimensional space. By converting complex structural information into a format suitable for computational analysis, Structural BLAST enables the identification of structural similarities between proteins.

Users input a target protein structure, and Structural BLAST compares it against a database of known protein structures. The tool employs algorithms to calculate structural similarities, providing a measure of the degree of resemblance between the query structure and those in the database. This information is valuable in understanding the evolutionary relationships, functional implications, and potential ligand-binding sites of proteins.

Structural BLAST finds applications in various areas of bioinformatics, including protein function prediction, drug discovery, and the study of protein evolution. By extending the principles of sequence alignment to the realm of protein structures, Structural BLAST contributes to our understanding of the relationships between proteins and aids researchers in uncovering the biological significance of structurally conserved regions.

VAST+:

VAST+ builds on the existing VAST database to generate such a report of structure neighbours. 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 neighbours 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.

Structure neighbours as computed by the VAST+ algorithm will be used in the future to provide links to 'similar structures' on Entrez/structure document summaries. Lists of similar structures are then summarized via a new interactive web service, which can also be used independently of the Entrez query and retrieval system and provides tools for sub-setting results, at http://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi. For a query structure specified by the user, the service lists similar structures, should they exist, ranked by the extent of the match. Matches that associate each polymer chain of the query with a corresponding polymer chain of some other structure are considered complete and are indicated with full circles in the search results table; partial matches are indicated with partially filled circles. The default ranking puts matches with the most matched components at the top of the list. Not all queries that have similar structures according to VAST+ are guaranteed to also have complete matches (although monomers usually do). The search results tables provided by the VAST+ web service give a concise summary of the matches and the extent/quality of the similarity. A clickable '+' symbol opens a panel for a selected match that provides more details and functionality.

THE VAST+ WEB SERVICE:

Structure neighbours as computed by the VAST+ algorithm will be used in the future to provide links to 'similar structures' on Entrez/structure document summaries. Lists of similar structures are then summarized via a new interactive web service, which can also be used independently of the Entrez query and retrieval system and provides tools for sub-setting results, at http://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi. For a query structure specified by the user, the service lists similar structures, should they exist, ranked by the extent of the match. Matches that associate each polymer chain of the query with a corresponding polymer chain of some other structure are considered complete and are indicated with full circles in the search results table; partial matches are indicated with partially filled circles. The default ranking puts matches with the most matched components at the top of the list. Not all queries that have similar structures according to VAST+ are guaranteed to also have complete matches (although monomers usually do). The search results tables provided by the VAST+ web service give a concise summary of the matches and the extent/quality of the similarity. A clickable '+' symbol opens a panel for a selected match that provides more details and functionality.

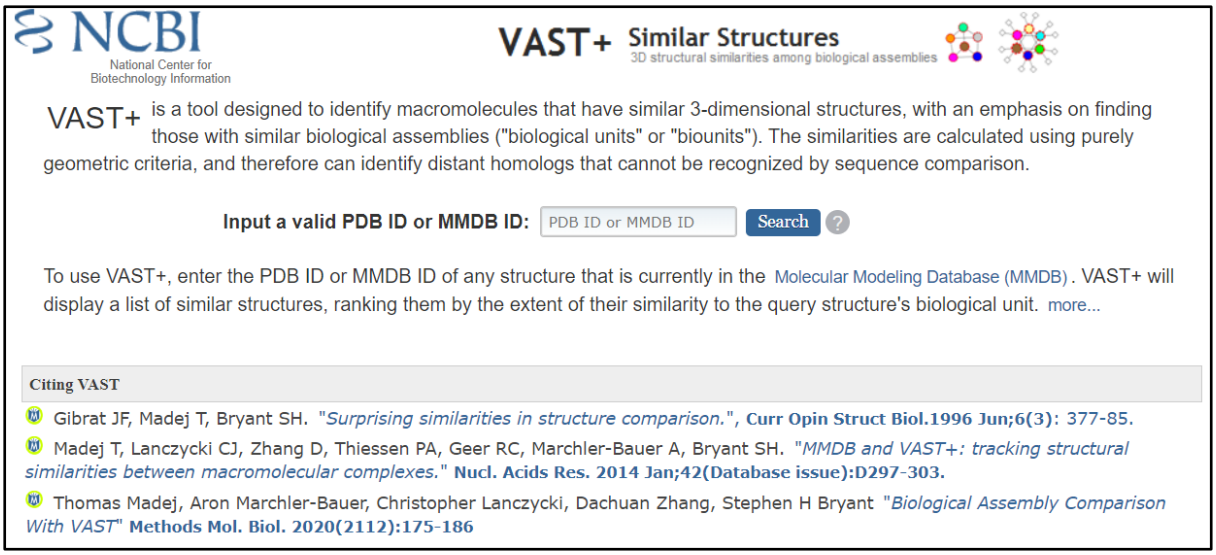


Fig 1: Homepage for VAST+

USING CN3D TO VISUALIZE BIOLOGICAL ASSEMBLY ALIGNMENTS:

The 3D structures of superimposed biological assemblies may be visualized using the 3D viewer Cn3D, which has been re-released as a new version 4.3.1 to support the visualization style. Currently, Cn3D is able to display the structure superposition of the matched biological assemblies and all the protein chains involved, but it can only display one sequence alignment at a time. Therefore, the individual protein match provides separate Cn3D launch points for each matched/aligned protein pair. All of these launch points will result in the same 3D image and rendering, but they will differ in the pair of aligned sequences that are chosen as the content of Cn3D's sequence/alignment viewer window. Figure 3 provides examples of Cn3D visualization sessions. Pairs of matching molecules are rendered in the same color, with unaligned segments rendered in gray. The default rendering settings, as generated and provided by the VAST+ service, can be examined and modified via Cn3D's Style Annotate menu.

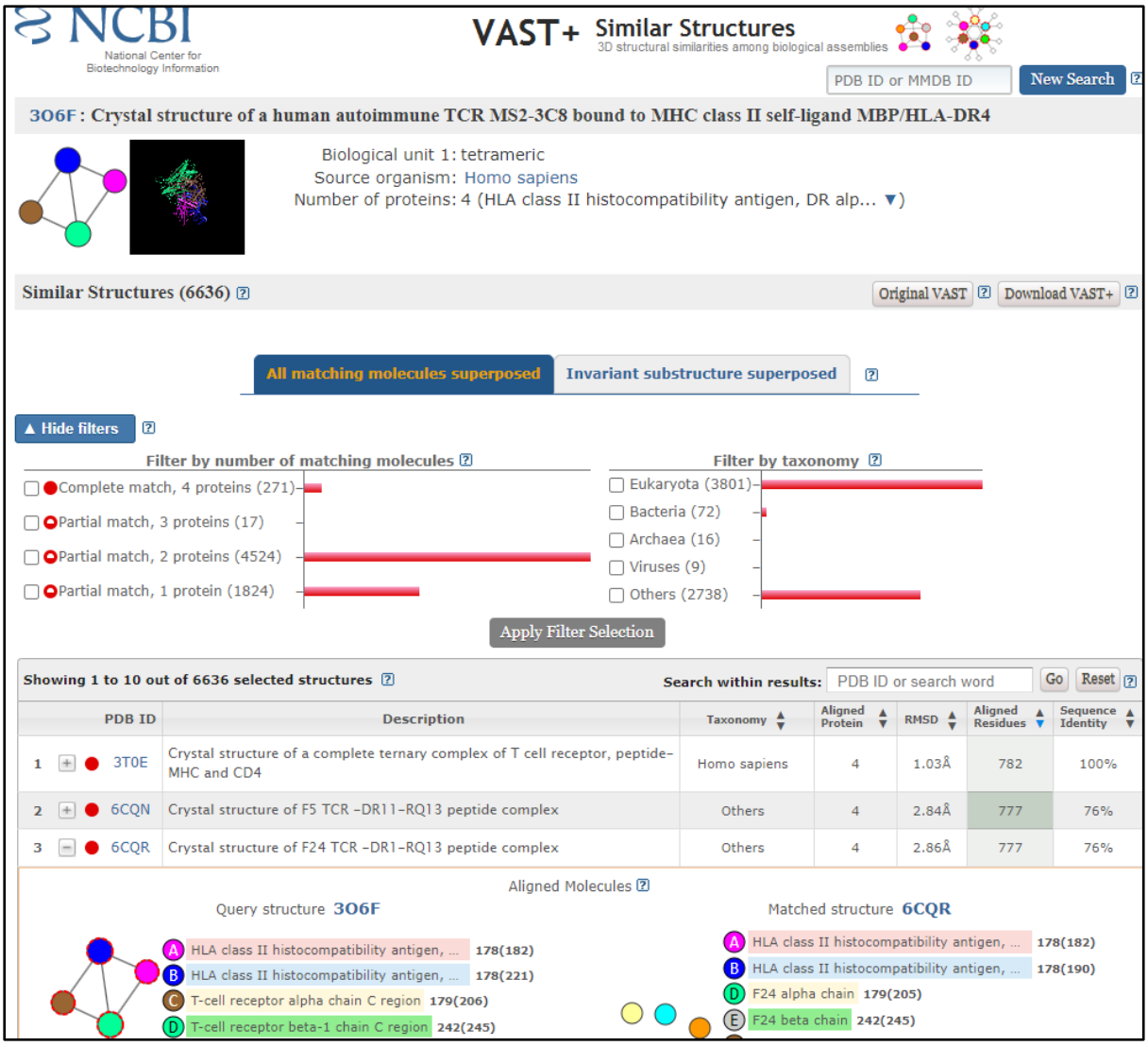


Fig 2: The VAST+ web service generates lists of structures that have 3D similarity to the query

URLs for MMDB and VAST resources:		
MMDB	Database home page	http://www.ncbi.nlm.nih.gov/structure
MMDB FTP	FTP Data distribution	ftp://ftp.ncbi.nih.gov/mmdb/
VAST	Identify structurally similar individual protein molecules	http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
VAST+	Identify structurally similar macromolecular complexes	http://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi
VAST search	search Input the 3D coordinates of a query structure to search for similar structures	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html

Cn3D	Molecular graphics viewer	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
CBLAST	Find 3D structures that are related to a query protein via sequence comparison	http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi

SIMILAR SUBSTRUCTURES: ORIGINAL VAST AND VAST-SEARCH:

MMDB is updated weekly, following PDB's schedule. With each update, computation of new structure neighbors is completed within a few days, and they are available as structure neighbors computed for biological assemblies via the VAST+ service, as well as structure neighbors computed for individual protein chains and domains, via the original VAST service. The latter is accessible on the can be found near the top of the VAST+results page. At this point, the VAST-search service, which accepts 3D structure data uploaded in PDB-format, remains unchanged, and presents similar structures in the Original VAST format.

DALI:

Dali (Distance-matrix Alignment) is a prominent structural bioinformatics tool used for the comparison and alignment of protein structures. Unlike traditional sequence-based alignment tools, Dali focuses on the three-dimensional spatial arrangement of amino acids within proteins, providing valuable insights into structural similarities and evolutionary relationships. Dali operates by converting protein structures into distance matrices, representing the spatial relationships between all pairs of amino acids. These matrices capture the Euclidean distances between the Cα atoms of residues in each protein. The algorithm then aligns the structures by identifying optimal superimpositions that maximize the overlap of structurally equivalent residues.

Dali generates a Z-score to assess the statistical significance of structural similarities, considering the size and complexity of the protein structures being compared. Higher Z-scores indicate more significant structural similarities, providing a measure of the likelihood that the observed structural alignment is not due to random chance.

The output of a Dali search typically includes a ranked list of structurally similar proteins from the Protein Data Bank (PDB) or another specified structural database. The results include information about the alignment, Z-score, root mean square deviation (RMSD), and other parameters, aiding researchers in evaluating the significance and quality of the structural match.

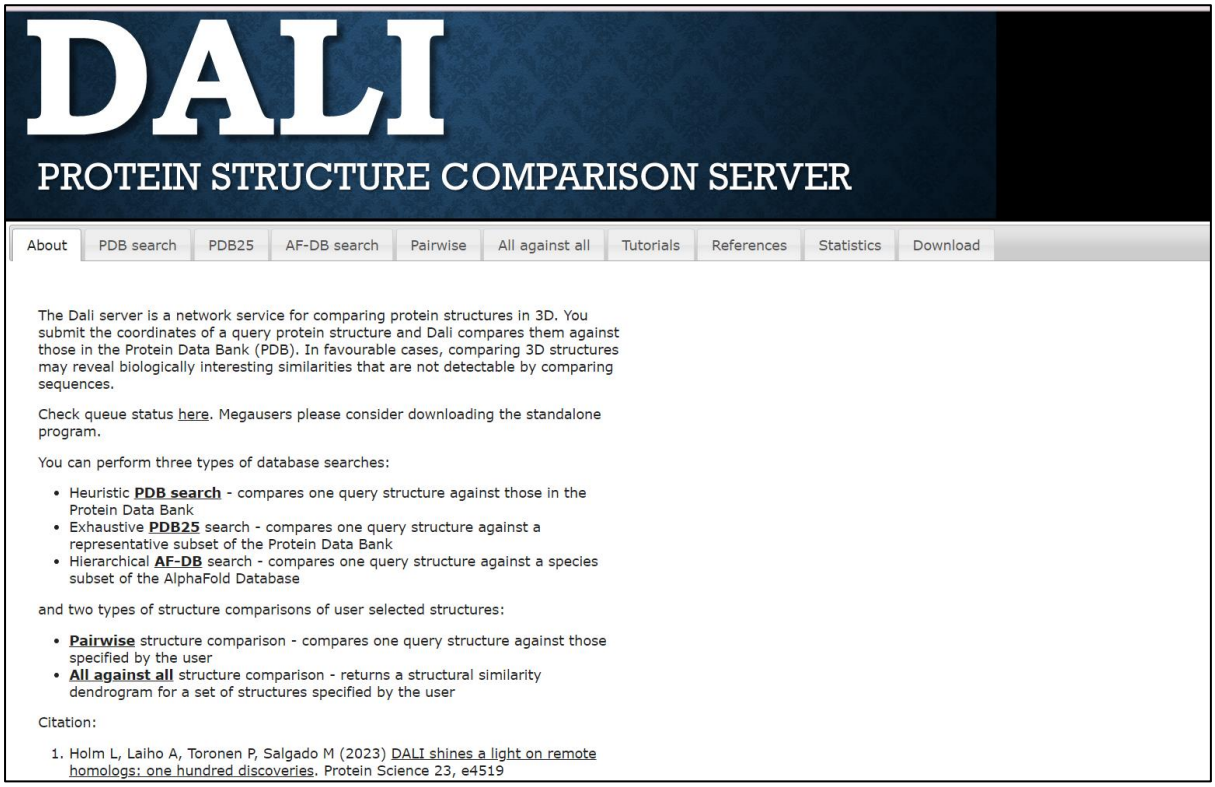


Fig 3: Homepage of Dali Server

REFERENCES:

1. Madej, T., Lanczycki, C. J., Zhang, D., Thiessen, P. A., Geer, R. C., Marchler-Bauer, A., & Bryant, S. H. (2013, December 6). MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Research. https://doi.org/10.1093/nar/gkt1208
2. Holm, L., & Laakso, L. (2016, April 29). Dali server update. Nucleic Acids Research. https://doi.org/10.1093/nar/gkw357
3. Holm, L., & Sander, C. (1998, January 1). Touring protein fold space with Dali/FSSP. Nucleic Acids Research. https://doi.org/10.1093/nar/26.1.316
4. VAST: Vector Alignment Search Tool. (n.d.). https://structure.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
5. Dali server. (n.d.). http://ekhidna2.biocenter.helsinki.fi/dali/

DATE: 13/02/2024

WEBLEM 14(A)
VAST+ (Vector Alignment Search Tool)
(URL: https://structure.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi)

AIM:

To identify similar structure for query 'trypsin' (PDB ID: 1P2J) using VAST+ tool.

INTRODUCTION:

NCBI has maintained the Molecular Modeling Database (MMDB). Since 1996, as a collection of publicly accessible experimentally determined macromolecular structures that have been deposited with the Protein Data Bank (PDB). MMDB serves a variety of VAST, it facilitates searching for macromolecular structure data in NCBI's Entrez query and retrieval system. VAST+ is a tool designed to compare 3-dimensional structures, with an emphasis on finding those with similar macromolecular complexes. The similarities are calculated using purely geometric criteria, without regard to sequence similarity, and therefore can identify distant homologs. VAST+ is built upon the original Vector Alignment Search Tool (VAST), and expands the capabilities of that program by making it possible to now find macromolecular structures that have similarly shaped biological units (also referred to as "biounits"), not just those that share similarly shaped individual protein molecules or fragments.

The similar structures found by the programs are often referred to as "neighbors." VAST neighbors are structures that contain similarly shaped individual protein molecules or 3D domains, and VAST+ neighbors are structures that have similarly shaped biological units. 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.

Structure neighbours are computed by the VAST+ algorithm will be used in the future to provide links to 'similar structures' on Entrez/structure dominant summaries. List of similar structures are then summarized via a new interactive web service, which can also be used independently of the Entrez query and retrieval system and provides tools for sub-setting results.

Matches that associate each polymer chain of the query with a corresponding polymer chain of some other structure are considered complete and are indicated with full circles in the search results table; partial matches are indicated with partially filled circles. The default ranking puts

matches with the most matched components at the top of the list. Not all queries that have similar structures according to VAST+ are guaranteed to also have complete matches (although monomers usually do). The search results tables provided by the VAST+ web service give a concise summary of the matches and the extent/quality of the similarity. A clickable '+' symbol opens a panel for a selected match that provides more details and functionality.

Trypsin:

Trypsin is a pivotal enzyme crucial for protein digestion and absorption in mammals. It is a serine protease found in the small intestine, where it hydrolyzes proteins into peptides and amino acids, facilitating their absorption into the bloodstream. Initially produced as the inactive zymogen trypsinogen by the pancreas, trypsin is activated in the small intestine to perform its digestive functions. Trypsin specifically cleaves peptide chains at the carboxyl side of lysine or arginine amino acids, aiding in the breakdown of proteins into smaller peptides for absorption. Structurally, trypsin consists of a single polypeptide chain comprising 223 amino acids, with an active site that includes key residues like His46 and Ser183. This enzyme's specificity is well-defined, as it hydrolyzes peptide bonds where the carbonyl group is contributed by arginine or lysine residues. Trypsin's mechanism of action involves cleaving the C-terminal of two basic amino acids, utilizing the negative charge associated with aspartate in its active site.

METHODOLOGY:

1. Go to the NCBI VAST Search web page:
2. https://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.
3. Provide the structure you want to search against by entering a PDB ID or MMDB ID.
4. Adjust the search parameters as needed.
5. Click on the "Search" button to initiate the search.
6. The search results will be displayed, showing structures that are similar to the input query. The results will include information such as the PDB ID, description, and the level of structural similarity.
7. Explore the results to identify structures that match your criteria. You can click on individual results to view more detailed information about the structures.
8. Use the visualization tools provided to examine the 3D structures of the matched proteins.

OBSERVATIONS:

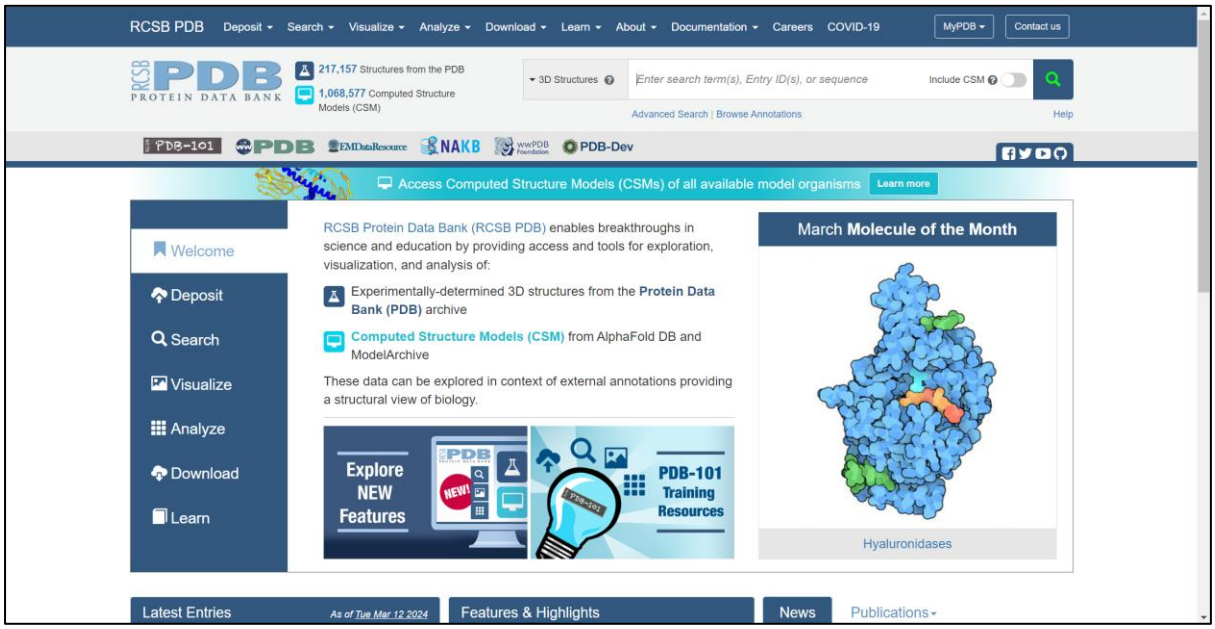


Fig 1: Homepage of PDB Database

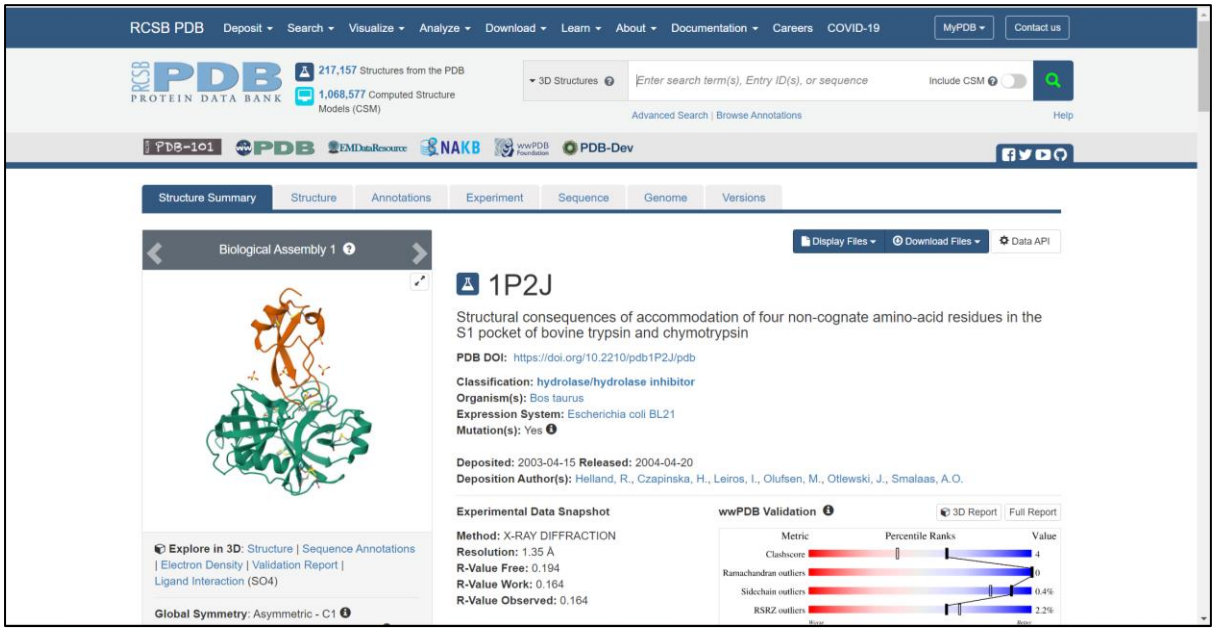


Fig 2: PDB ID: 1P2J obtained from PDB Database for the query 'trypsin'

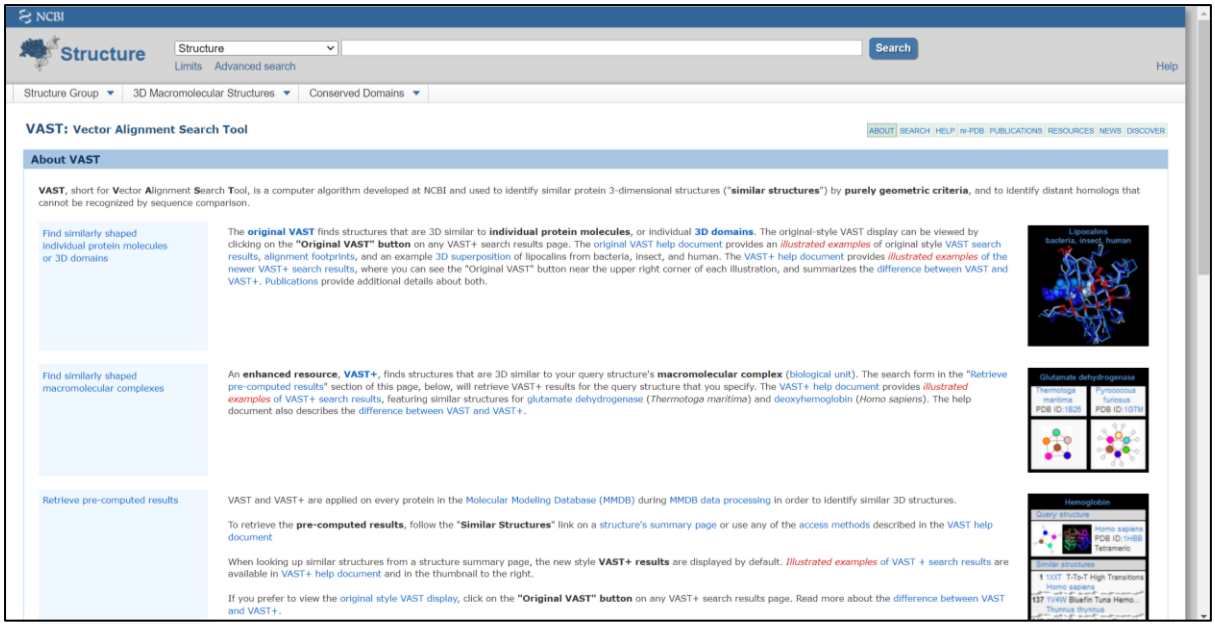


Fig 3: Homepage of VAST+ (Vector Alignment Search Tool)

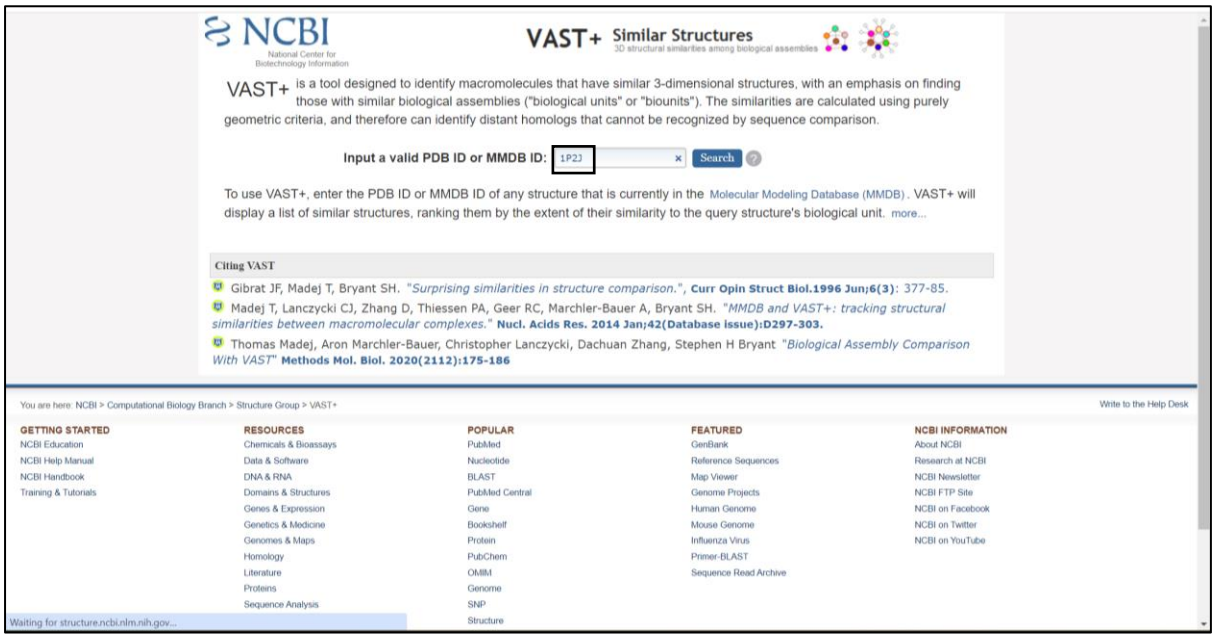


Fig 4: Pasting the PDB ID: 1P2J for the query 'trypsin' in the Search box

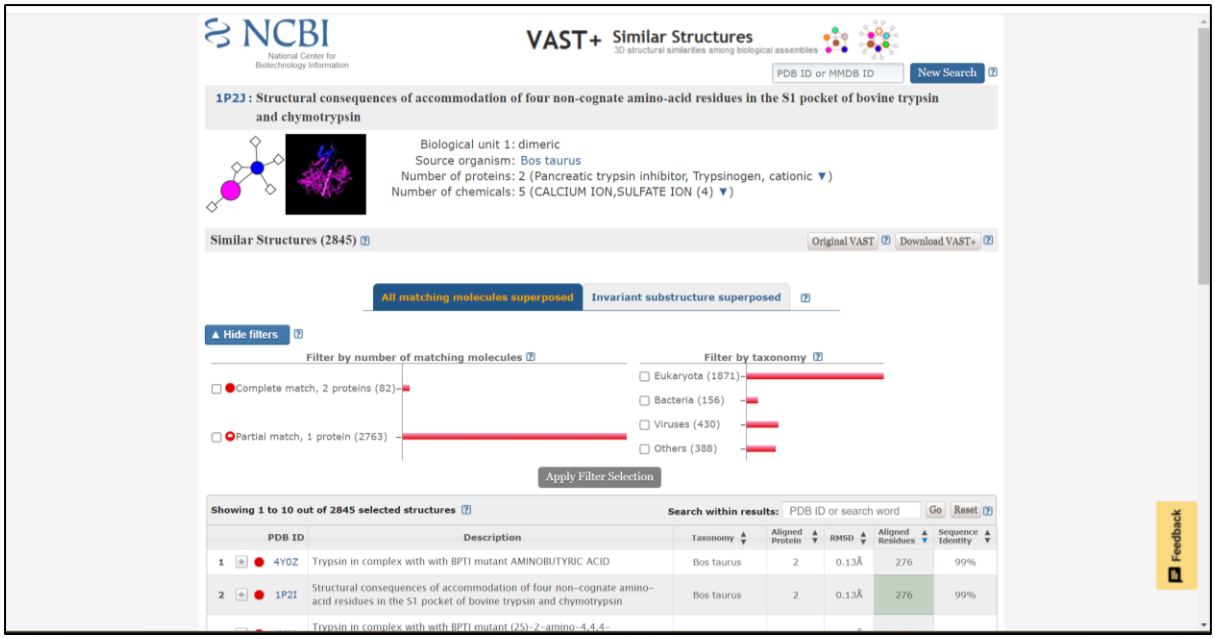


Fig 5: Result page of (2845) Similar Structures and Filter Options (Like Eukaryota, Bacteria, Viruses, Others)

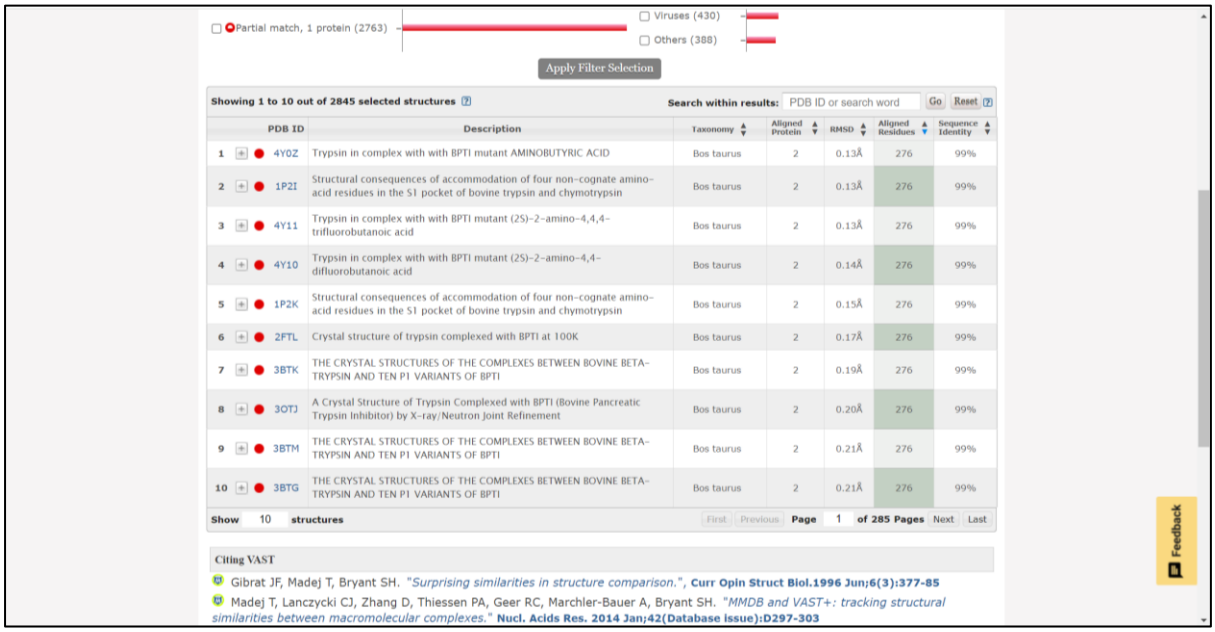


Fig 6: Displaying 10 Similar Structures out of 2845 Hits

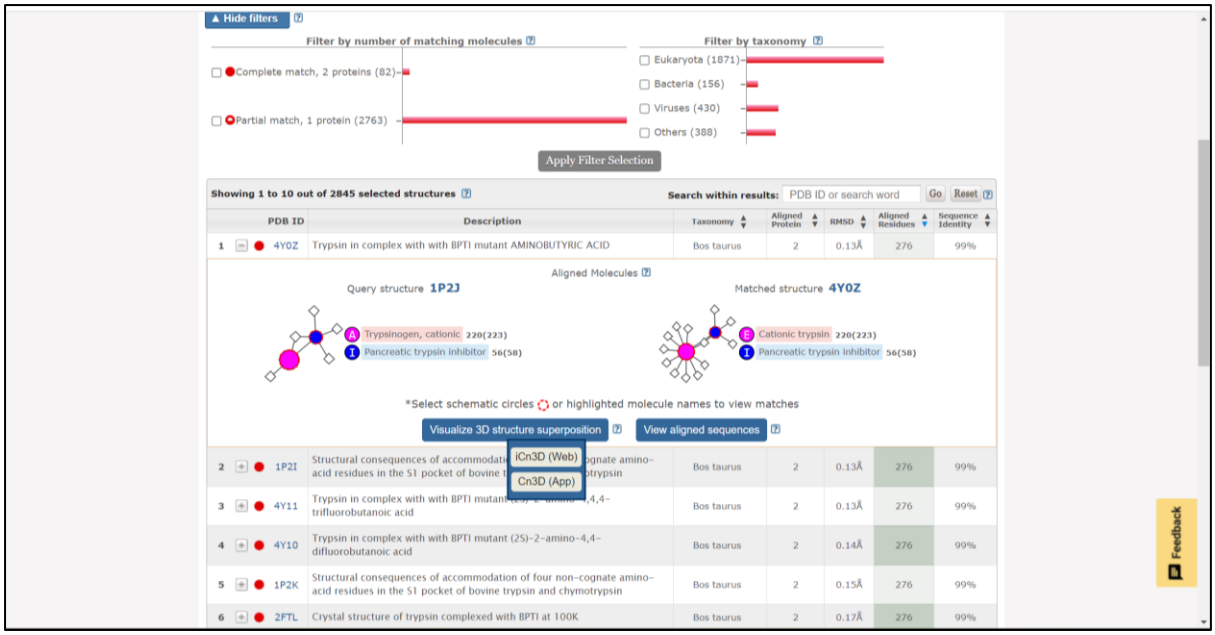


Fig 7: Selecting the 1st hit with the highest sequence similarity of 99%

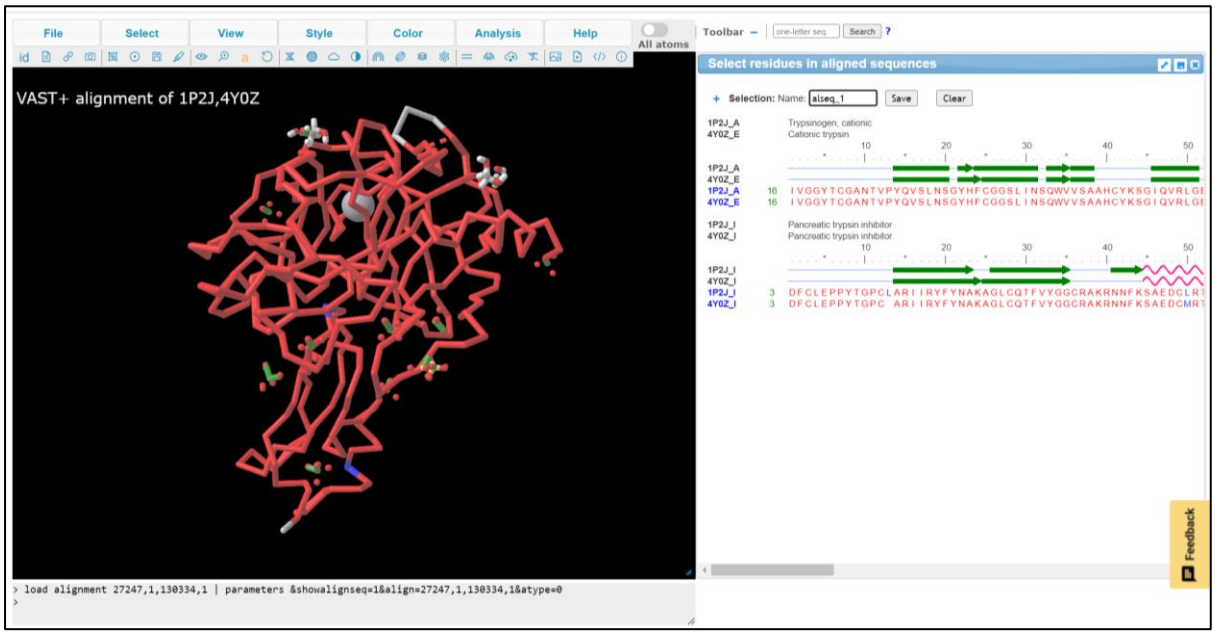


Fig 8: 3D Structure Visualization of the selected structure (PDB ID: 4YOZ) and display of the sequence alignment

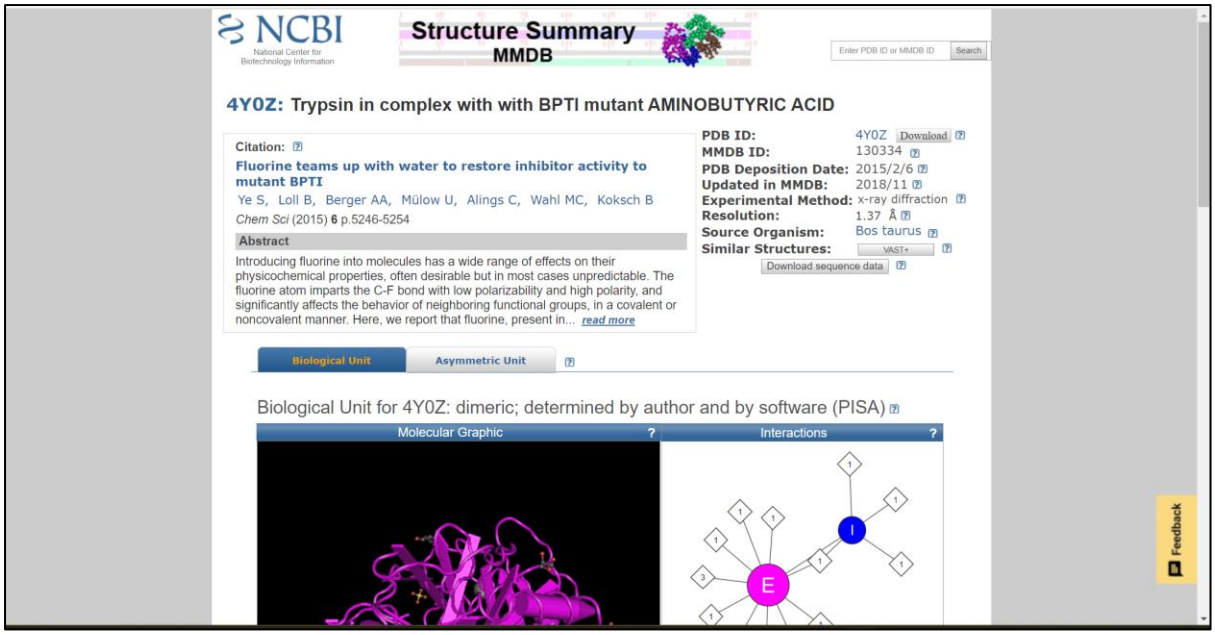


Fig 9: Detailed Description of selected structure (PDB ID: 4YOZ)

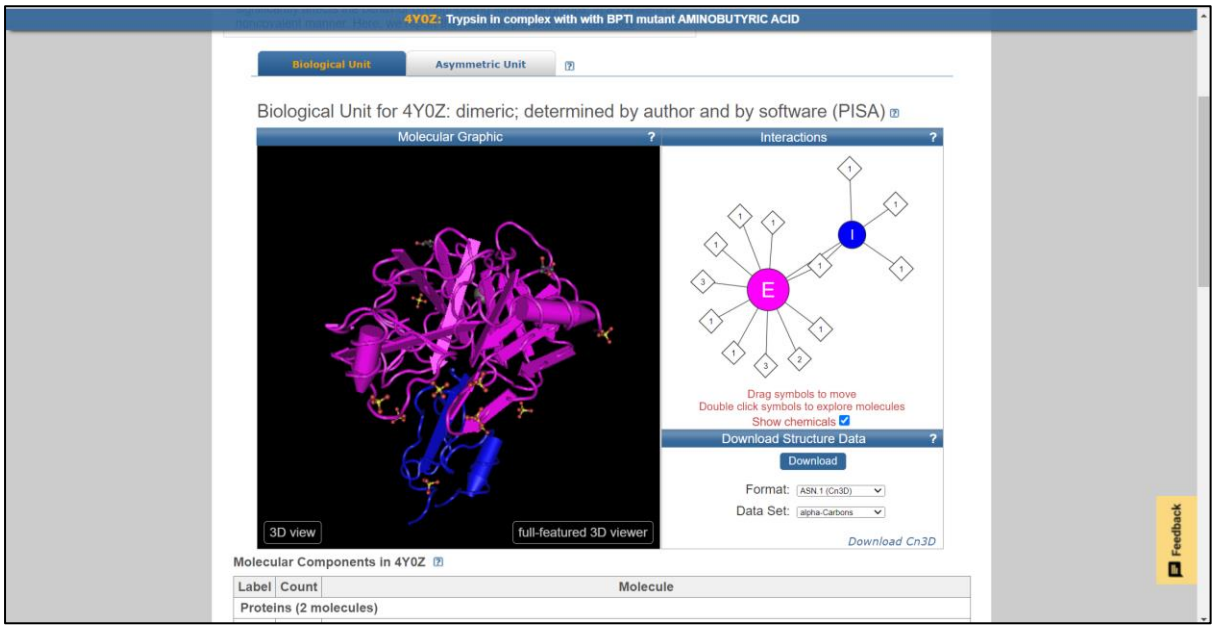


Fig 10: 3D View of selected structure (PDB ID: 4YOZ) and Interactions between the molecules

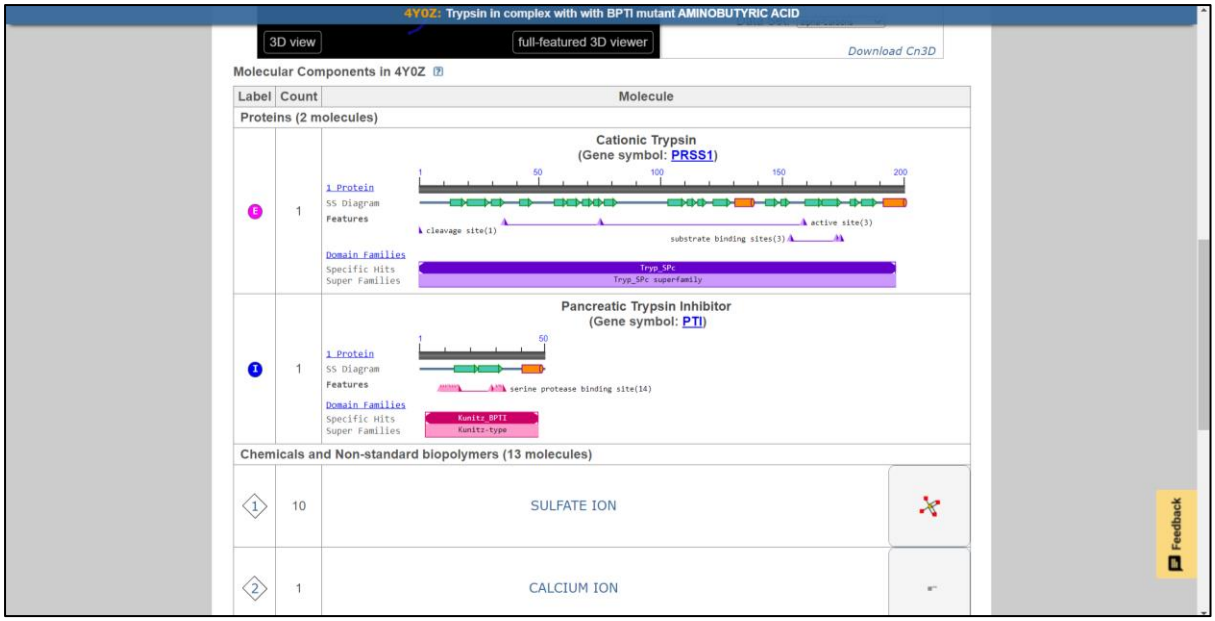


Fig 11: Display of Molecular components of selected structure (PDB ID: 4YOZ)

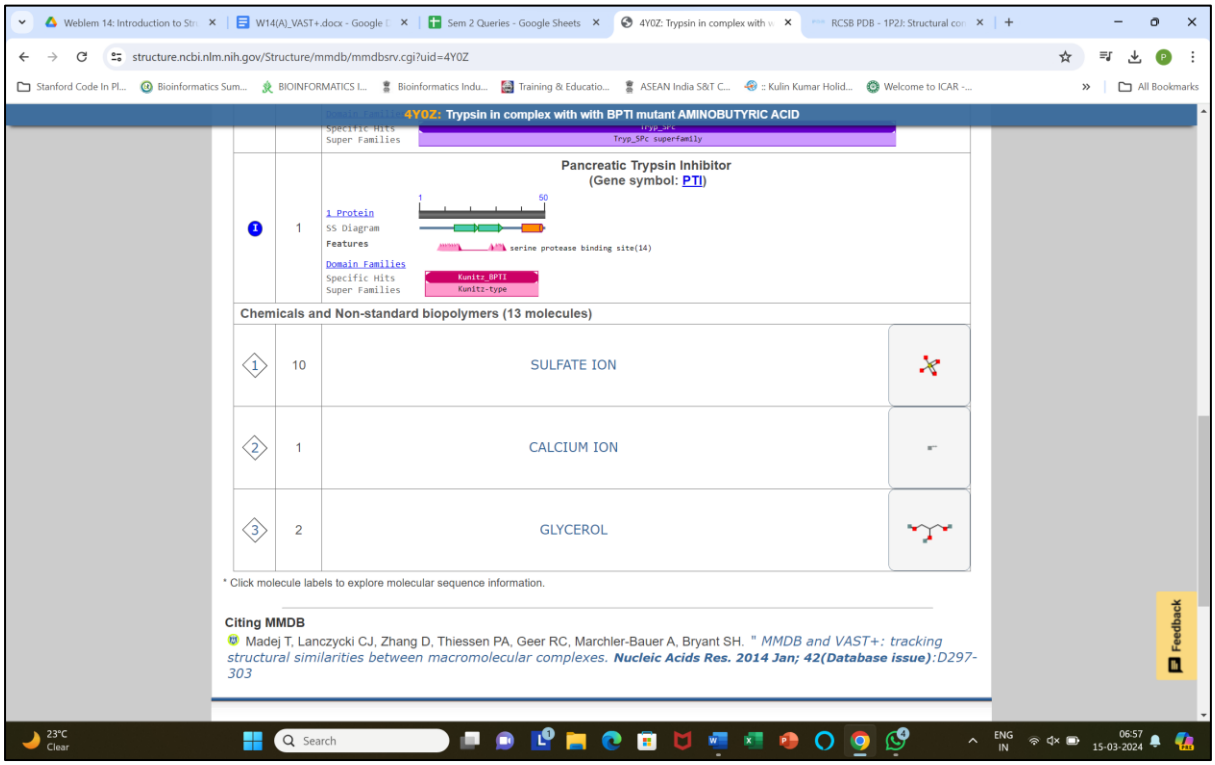


Fig 12: Display of Chemicals and Non – standard biopolymers of selected structure (PDB ID: 4YOZ)

RESULTS:

Based on the query 'trypsin' (PDB ID: 1P2J) in the VAST+ tool, the best match obtained is Trypsin in complex with BPTI mutant AMINOBUTYRIC ACID with 99% of sequence identity which is the highest in the given list. The RMSD score is 0.13A which is the lowest among all entries and the aligned residues is 276. So, we could take the structure of Trypsin in complex with BPTI mutant AMINOBUTYRIC ACID and can map on our query and gained valuable insights of the functions, structure, coding regions, protein folding, molecular interactions, etc.

CONCLUSION:

VAST+ tool was explored and used to find similar structure of query 'trypsin' (PDB ID: 1P2J). VAST is valuable in identifying structural homology among proteins. By aligning 3D vectors, it can reveal similarities in protein structures even when there are variations in sequence. Similar structures often imply similar functions. VAST can aid in predicting the function of a protein by comparing its structure with known structures. Identifying structurally similar proteins can be crucial in drug discovery. The VAST tool has provided structural information about 'trypsin' (PDB ID: 1P2J), including details on their crystal structures, resolutions, and completeness.

REFERENCES:

1. Madej, T., Lanczycki, C. J., Zhang, D., Thiessen, P. A., Geer, R. C., Marchler-Bauer, A., & Bryant, S. H. (2013, December 6). MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Research. https://doi.org/10.1093/nar/gkt1208.
2. Hancock, J. M., & Bishop, M. J. (2004, October 15). VAST (Vector Alignment Search Tool). Dictionary of Bioinformatics and Computational Biology. https://doi.org/10.1002/0471650129.dob0782.
3. VAST: Vector Alignment Search Tool. (n.d.). https://structure.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml.
4. Madej, T., Marchler-Bauer, A., Lanczycki, C. J., Zhang, D., & Bryant, S. H. (2020, January 1). Biological Assembly Comparison with VAST+. Methods in Molecular Biology. https://doi.org/10.1007/978-1-0716-0270-6_13.
5. National Center for Biotechnology Information (2024). PubChem Compound Summary for , Trypsin. Retrieved March 16, 2024 from https://pubchem.ncbi.nlm.nih.gov/compound/Trypsin

DATE: 13/02/2024

WEBLEM 14(B)
DALI SERVER
(URL: http://ekhidna2.biocenter.helsinki.fi/dali/)

AIM:

To perform structure comparison for query 'trypsin' (PDB ID: 1P2J) by using Dali server.

INTRODUCTION:

The Dali server is a network service for comparing protein structure in 3D. In favorable cases, comparing 3D structure may reveal biologically interesting similarities that are not detectable by comparing sequence. The Dali server has been running in various place for over 20 years and is used routinely by crystallographers on newly solved structures. The server performs three types of structure comparisons:

1. Protein Data Bank (PDB) search compares one query structure against those in the PDB and returns a list of similar structures.
2. Pairwise comparison compares one query structure against a list of structures specified by the user.
3. All against all structure comparison returns a structural similarity matrix, a dendrogram and a multidimensional scaling projection of a set of structures specified by the user, structural superimpositions are visualized using the JAVA-free webGL viewers PV. The structural alignment view is enhanced by sequence similarity searches against UniProt. The combine structure sequence alignment information is compressed toa stack of aligned sequence logo. In the stack, each structure is structurally aligned to the query protein and represented by a sequence logo.

Trypsin:

Trypsin is a pivotal enzyme crucial for protein digestion and absorption in mammals. It is a serine protease found in the small intestine, where it hydrolyzes proteins into peptides and amino acids, facilitating their absorption into the bloodstream. Initially produced as the inactive zymogen trypsinogen by the pancreas, trypsin is activated in the small intestine to perform its digestive functions. Trypsin specifically cleaves peptide chains at the carboxyl side of lysine or arginine amino acids, aiding in the breakdown of proteins into smaller peptides for absorption. Structurally, trypsin consists of a single polypeptide chain comprising 223 amino acids, with an active site that includes key residues like His46 and Ser183. This enzyme's specificity is well-defined, as it hydrolyzes peptide bonds where the carbonyl group is contributed by arginine or lysine residues. Trypsin's mechanism of action involves cleaving the C-terminal of two basic amino acids, utilizing the negative charge associated with aspartate in its active site.

METHODOLOGY:

1. Locate the query protein 'trypsin' (PDB ID: 1P2J) by opening the PDB database and searching for it
2. Click the PDB search button after opening Dali server.
3. Click on PDB25.
4. Click the submit button after typing PDB ID into the search field.
5. Select multiple chains to check structure alignments. Select neighbors (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbors is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious.

OBSERVATIONS:

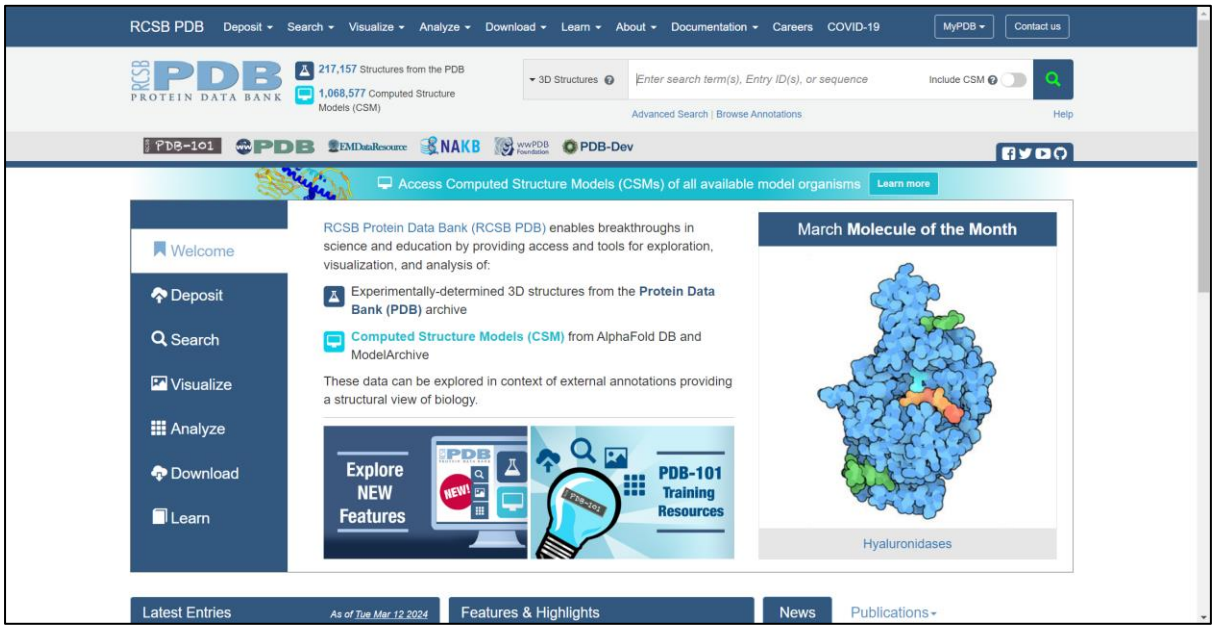


Fig 1: Homepage of PDB Database

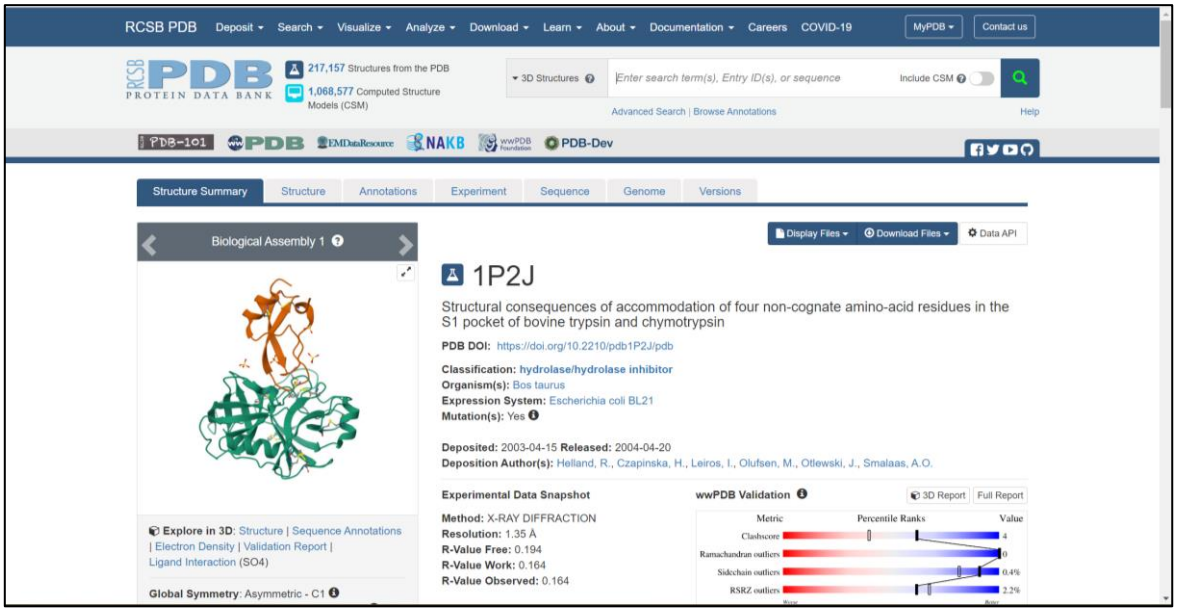


Fig 2: PDB ID: 1P2J obtained from PDB Database for the query 'trypsin'

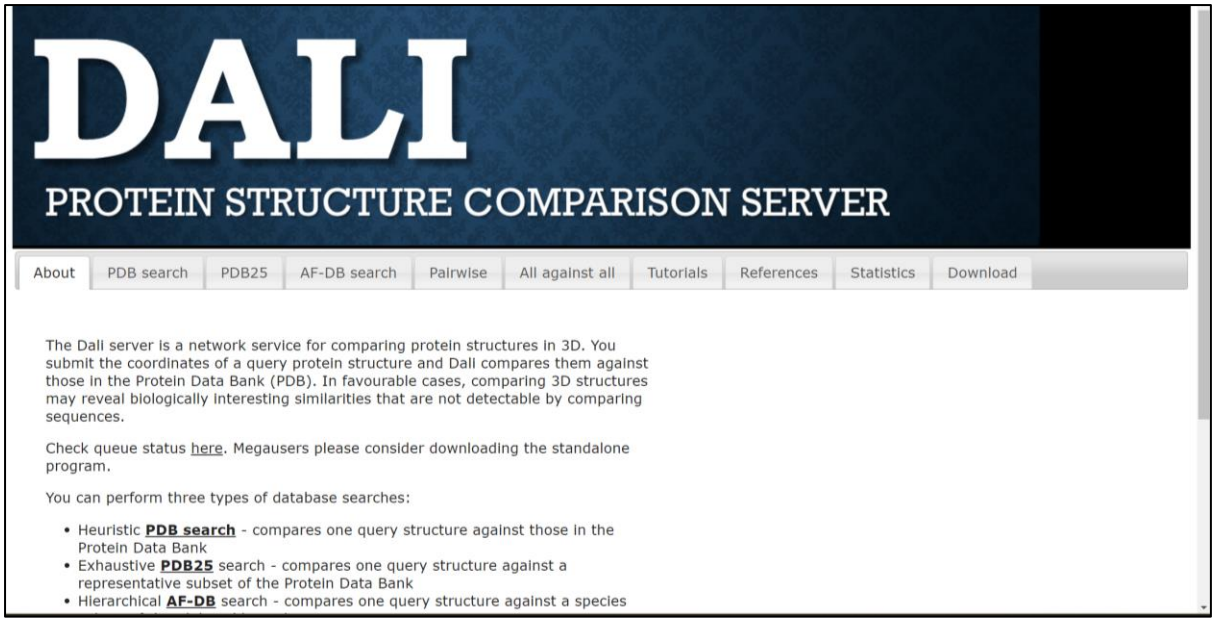


Fig 3: Homepage of DALI Server

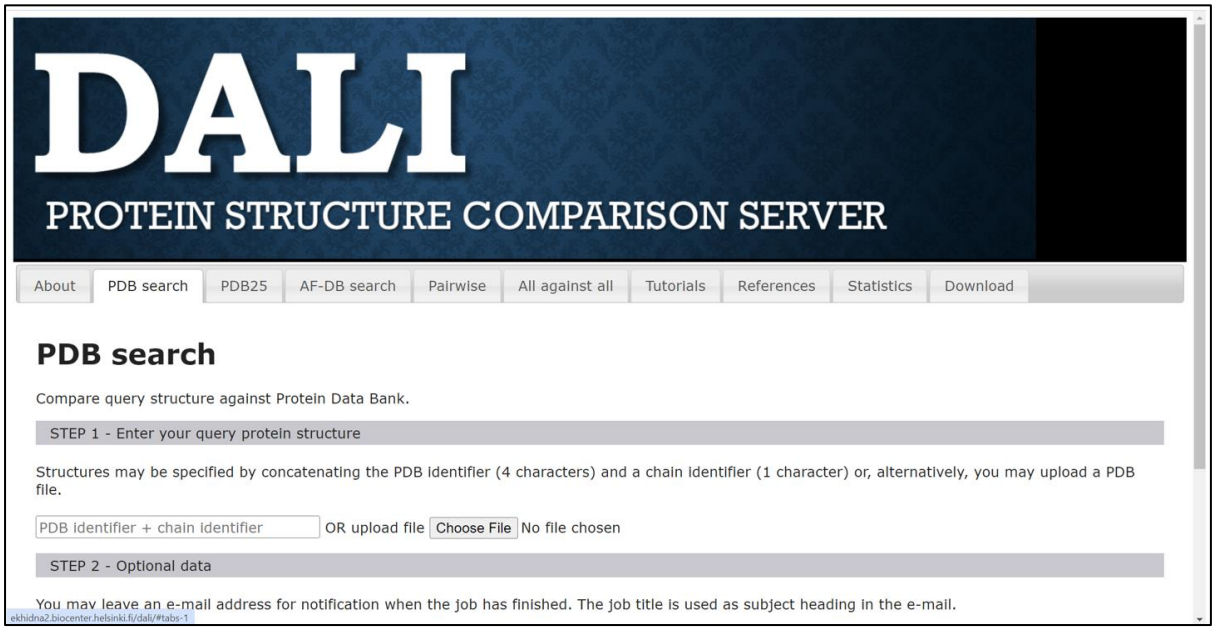


Fig 4: PDB Search page

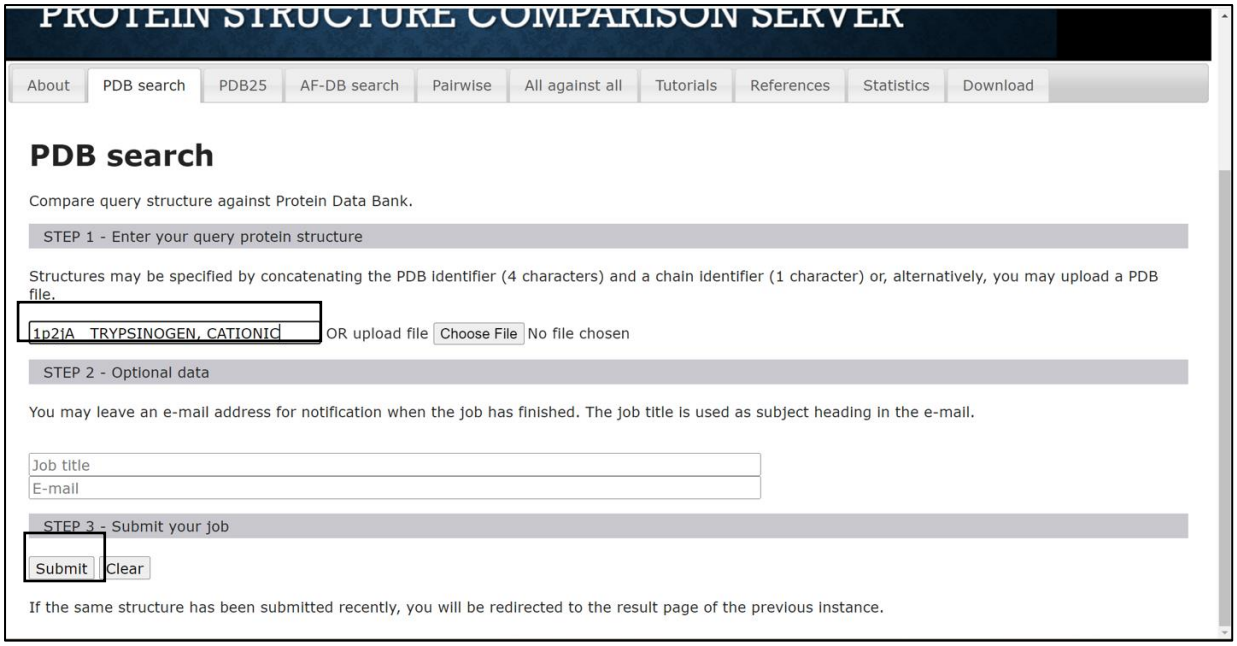


Fig 5: Searching for the PDB ID: 1P2J

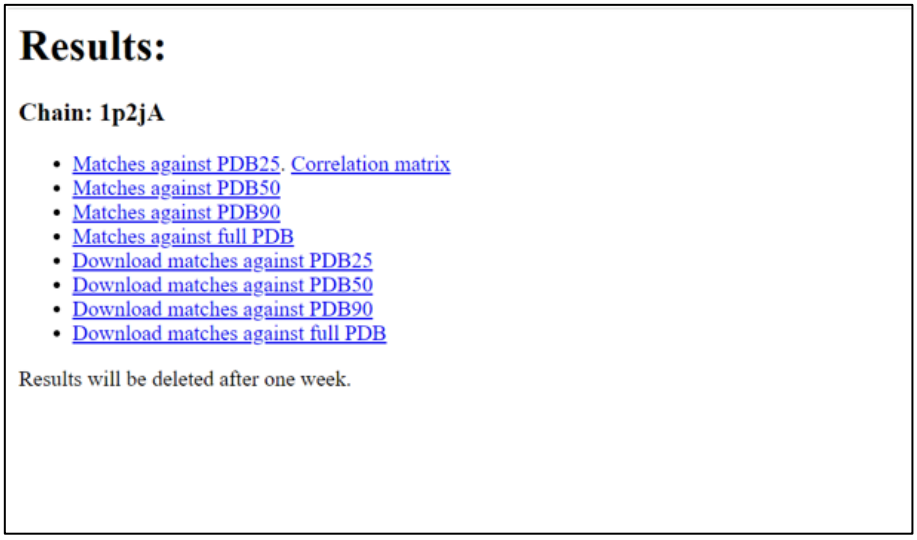


Fig 6: Results obtained for PDB ID: 1P2J

Results: 1p2jA

Query: 1p2jA

MOLECULE: TRYPSINOGEN, CATIONIC;

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:	1z8g-A	34.5	1.2	217	366	43	PDB	MOLECULE: SERINE PROTEASE HEPSIN;
2:	5to3-B	33.9	1.4	218	368	39	PDB	MOLECULE: PROTHROMBIN;
3:	7wqx-A	32.0	1.6	215	501	37	PDB	MOLECULE: ENTEROPEPTIDASE;
4:	3h7t-A	30.6	1.9	213	235	29	PDB	MOLECULE: GROUP 3 ALLERGEN SMIPP-S YVT004A06;
5:	4kkd-B	28.7	1.8	211	405	35	PDB	MOLECULE: MANNAN-BINDING LECTIN SERINE PROTEASE 1;
6:	4a5t-S	28.0	2.1	206	767	40	PDB	MOLECULE: PLASMINOGEN;
7:	5i25-A	25.8	2.3	201	596	34	PDB	MOLECULE: COAGULATION FACTOR XI;
8:	2xxl-B	23.8	2.0	206	354	33	PDB	MOLECULE: GRAM-POSITIVE SPECIFIC SERINE PROTEASE, ISOFORM B
9:	2xrc-B	23.7	1.7	181	485	33	PDB	MOLECULE: HUMAN COMPLEMENT FACTOR I;
10:	3h5c-B	23.3	2.2	196	312	24	PDB	MOLECULE: PROTEIN Z-DEPENDENT PROTEASE INHIBITOR;
11:	6bqm-A	23.2	1.9	200	460	32	PDB	MOLECULE: SERINE PROTEASE VESC;
12:	2b91-A	22.7	2.2	203	372	29	PDB	MOLECULE: PROPHENOLOXIDASE ACTIVATING FACTOR;
13:	4ink-A	21.8	2.0	176	203	13	PDB	MOLECULE: SERINE PROTEASE SPLD;
14:	2xwb-F	21.5	2.4	196	714	23	PDB	MOLECULE: COMPLEMENT C3B BETA CHAIN;
15:	3wy8-A	20.8	2.6	193	219	14	PDB	MOLECULE: SERINE PROTEASE;
16:	7uxg-A	19.5	2.5	189	239	17	PDB	MOLECULE: SERINE PROTEASE;
17:	5c2z-A	19.4	2.5	183	248	17	PDB	MOLECULE: EXFOLIATIVE TOXIN D2;
18:	1agj-A	19.3	2.4	181	242	16	PDB	MOLECULE: EPIDERMOLYTIC TOXIN A;
19:	7u5b-I	18.1	1.4	145	149	51	PDB	MOLECULE: ANTI-KLK5 FAB HEAVY CHAIN;
20:	3ak5-A	17.4	2.4	178	970	19	PDB	MOLECULE: HEMOGLOBIN-BINDING PROTEASE HBP;
21:	4nsy-B	17.2	2.7	183	266	15	PDB	MOLECULE: LYSYL ENDOPEPTIDASE;
22:	2r3y-A	17.1	2.7	170	280	16	PDB	MOLECULE: PROTEASE DEGS;
23:	5il9-A	16.9	2.7	169	465	15	PDB	MOLECULE: PROTEASE DO-LIKE 9;
24:	3k6v-A	16.8	2.6	166	227	21	PDB	MOLECULE: POSSIBLE MEMBRANE-ASSOCIATED SERINE PROTEASE;

Fig 7: List of matches obtained when searched against PDB25

Results: 1p2jA

Query: 1p2jA

MOLECULE: TRYPSINOGEN, CATIONIC;

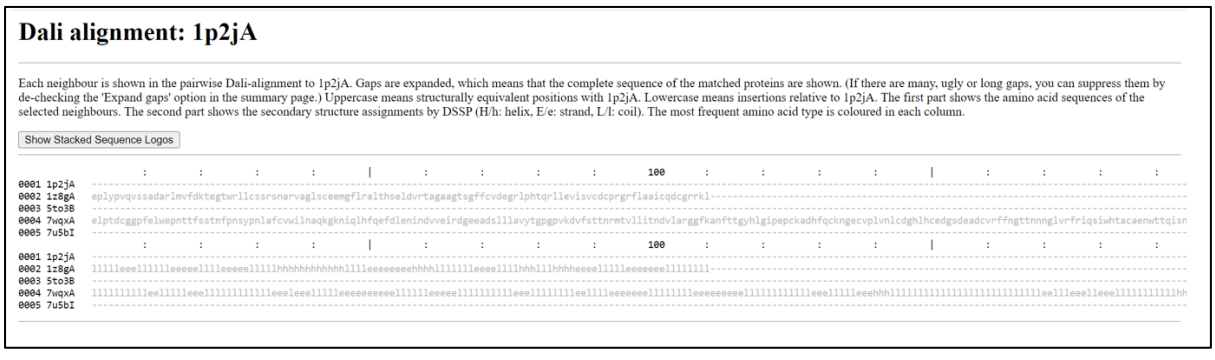
Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format 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:	1z8g-A	34.5	1.2	217	366	43	PDB	MOLECULE: SERINE PROTEASE HEPSIN;
☑ 2:	5to3-B	33.9	1.4	218	368	39	PDB	MOLECULE: PROTHROMBIN;
☑ 3:	7wqx-A	32.0	1.6	215	501	37	PDB	MOLECULE: ENTEROPEPTIDASE;
☐ 4:	3h7t-A	30.6	1.9	213	235	29	PDB	MOLECULE: GROUP 3 ALLERGEN SMIPP-S YVT004A06;
☐ 5:	4kkd-B	28.7	1.8	211	405	35	PDB	MOLECULE: MANNAN-BINDING LECTIN SERINE PROTEASE 1;
☐ 6:	4a5t-S	28.0	2.1	206	767	40	PDB	MOLECULE: PLASMINOGEN;
☐ 7:	5i25-A	25.8	2.3	201	596	34	PDB	MOLECULE: COAGULATION FACTOR XI;
☐ 8:	2xxl-B	23.8	2.0	206	354	33	PDB	MOLECULE: GRAM-POSITIVE SPECIFIC SERINE PROTEASE, ISOFORM B
☐ 9:	2xrc-B	23.7	1.7	181	485	33	PDB	MOLECULE: HUMAN COMPLEMENT FACTOR I;
☐ 10:	3h5c-B	23.3	2.2	196	312	24	PDB	MOLECULE: PROTEIN Z-DEPENDENT PROTEASE INHIBITOR;
☐ 11:	6bqm-A	23.2	1.9	200	460	32	PDB	MOLECULE: SERINE PROTEASE VESC;
☐ 12:	2b91-A	22.7	2.2	203	372	29	PDB	MOLECULE: PROPHENOLOXIDASE ACTIVATING FACTOR;
☐ 13:	4ink-A	21.8	2.0	176	203	13	PDB	MOLECULE: SERINE PROTEASE SPLD;
☐ 14:	2xwb-F	21.5	2.4	196	714	23	PDB	MOLECULE: COMPLEMENT C3B BETA CHAIN;
☐ 15:	3wy8-A	20.8	2.6	193	219	14	PDB	MOLECULE: SERINE PROTEASE;
☐ 16:	7uxg-A	19.5	2.5	189	239	17	PDB	MOLECULE: SERINE PROTEASE;
☐ 17:	5c2z-A	19.4	2.5	183	248	17	PDB	MOLECULE: EXFOLIATIVE TOXIN D2;
☐ 18:	1agj-A	19.3	2.4	181	242	16	PDB	MOLECULE: EPIDERMOLYTIC TOXIN A;
☑ 19:	7u5b-I	18.1	1.4	145	149	51	PDB	MOLECULE: ANTI-KLK5 FAB HEAVY CHAIN;
☐ 20:	3ak5-A	17.4	2.4	178	970	19	PDB	MOLECULE: HEMOGLOBIN-BINDING PROTEASE HBP;
☐ 21:	4nsy-B	17.2	2.7	183	266	15	PDB	MOLECULE: LYSYL ENDOPEPTIDASE;
☐ 22:	2r3y-A	17.1	2.7	170	280	16	PDB	MOLECULE: PROTEASE DEGS;
☐ 23:	5il9-A	16.9	2.7	169	465	15	PDB	MOLECULE: PROTEASE DO-LIKE 9;
☐ 24:	3k6v-A	16.8	2.6	166	227	21	PDB	MOLECULE: POSSIBLE MEMBRANE-ASSOCIATED SERINE PROTEASE;

Fig 8: Selecting matches with higher percent identity for Structural Alignment



Dali alignment: 1p2jA

Each neighbour is shown in the pairwise Dali-alignment to 1p2jA. Gaps are expanded, which means that the complete sequence of the matched proteins are shown. (If there are many, ugly or long gaps, you can suppress them by de-checking the 'Expand gaps' option in the summary page.) Uppercase means structurally equivalent positions with 1p2jA. Lowercase means insertions relative to 1p2jA. The first part shows the amino acid sequences of the selected neighbours. The second part shows the secondary structure assignments by DSSP (H/h: helix, E/e: strand, L/l: coil). The most frequent amino acid type is coloured in each column.

Show Stacked Sequence Logos

```
                 :         :         :         :         |         :         :         :         :       100       :         :         :         :         |         :         :         :         :
0001 1p2jA       ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
0002 1z8gA       eplypvqvssadarlmvfdktegtwrllcssrsnarvaglsceemgflralthseldvrtagaagtsgffcvdegrlphtqrllevisvcdcprgrflaaicqdcgrrkl----------------------------------------------------------------------------
0003 5to3B       ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
0004 7wqxA       elptdcggpfelwepnttfsstnfpnsypnlafcvwilnaqkgkniqlhfqefdlenindvveirdgeeadslllavytgpgpvkdvfsttnrmtvllitndvlarggfkanfttgyhlgipepckadhfqcknqecvplvnlcdghlhcedgsdeadcvrffngttnnnglvrfriqsiwhtacaenwttqisn
0005 7u5bI       ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
                 :         :         :         :         |         :         :         :         :       100       :         :         :         :         |         :         :         :         :
0001 1p2jA       ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
0002 1z8gA       llllleeellllllleeeeellllleeeeellllhhhhhhhhhhhhllllleeeeeeeehhhhlllllllleeeeellllhhhlllhhhhheeeellllleeeeeeellllllll----------------------------------------------------------------------------
0003 5to3B       ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
0004 7wqxA       lllllllllleellllleeellllllllllllleeeleeellllleeeeeeeeeellllleeeeellllllllleeellllllleellleeeeeeellllllleeeeeeeeellllllllllleeellllleeehhhllllllllllllllllllllllllllleellleeellleeellllllllllhh
0005 7u5bI       ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

Fig 9: DALI Alignment of Selected Sequences

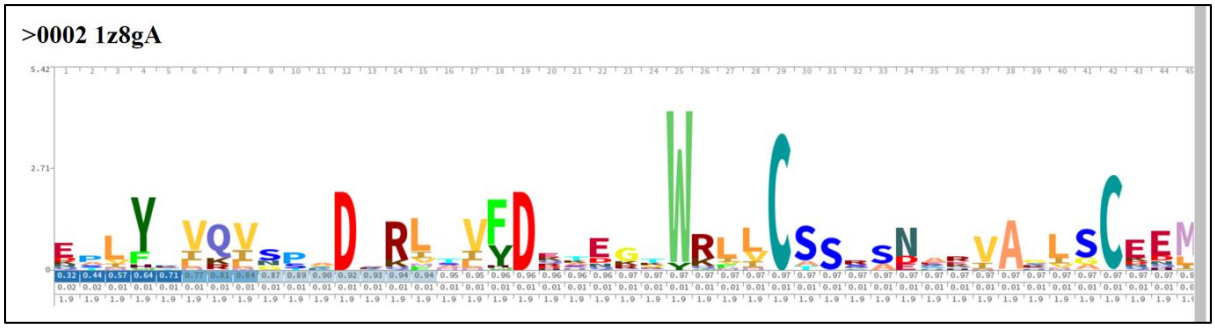


Fig 10: Stacked Sequence Logos

RESULTS:

The Dali server (Distance Alignment Matix Based on Local Structure) is a tool commonly used in bioinformatics and structural biology for protein structure comparison and analysis. The protein query 'trypsin' (PDB ID: 1P2J) was used for protein structure comparison. Matching a structure on PDB25 which shows 25% matching, highest % identity was found to be 43%, which is 1z8g-A: serine protease hepsin with Z score is 34.5. Structure alignment shows number of h which represents Helix.

CONCLUSION:

DALI server was explored and used to perform structural comparison on query 'trypsin' (PDB ID: 1P2J) protein. It provided results based on the best Z-score which helped in further analysis.

REFERENCES:

1. DALI server. (n.d.). Retrieved from http://ekhidna2.biocenter.helsinki.fi/dali/
2. National Center for Biotechnology Information (2024). PubChem Compound Summary for , Trypsin. Retrieved March 16, 2024 from https://pubchem.ncbi.nlm.nih.gov/compound/Trypsin