V.V.VANNIAPERUMAL COLLEGE FOR WOMEN
(Belonging to Virudhunagar Hindu Nadars)
An Autonomous Institution Affiliated to Madurai Kamaraj University, Madurai
Re-accredited with 'A' Grade (3rd Cycle) by NAAC
VIRUDHUNAGAR – 626 001

BIOINFORMATICS LABORATORY MANUAL

(UNDER DBT STAR COLLEGE SCHEME)
No HRD-11011/163/2020-HRD-DBT
Department of Biotechnology, Ministry of Science and Technology
MHRD, New Delhi

K. Sudha Rameshwari

BIOINFORMATICS LABORATORY MANUAL

K.Sudha Rameshwari

V.V.VANNIAPERUMAL COLLEGE FOR WOMEN

(Belonging to Virudhunagar Hindu Nadars)

An Autonomous Institution Affiliated to Madurai Kamaraj University

Re-accredited with 'A' Grade (3rd cycle) by NAAC

VIRUDHUNAGAR – 626 001 (TAMIL NADU)

DBT STAR COLLEGE SCHEME

Department of Biotechnology, Ministry of Science and Technology

Government of India, New Delhi

Chairman & Principal : Dr.(Tmty.) S.M. MEENA RANI, M.Sc.,M.Phil.,P.G.D.C.A.,Ph.,
Contact number: 9498088703
e-mail:smmeenarani@gmail.com

Coordinator & Member Secretary : Dr.(Tmty.) M.TAMILSELVI, M.Sc.,M.Phil.,Ph.D.,
Contact Number: 9894883106
e-mail:tamilasc.selvi08@gmail.com

FOREWORD

This Lab Manual on "**BIOINFORMATICS**" is prepared in accordance with the updated syllabus under DBT Star College Scheme sponsored by the Department of Biotechnology, Ministry of Science and Technology, MHRD, New Delhi to fulfill the needs of students.

This manual could enable the students to retrieve protein, nucleic acid sequence using different tools and to interpret the results. The protocols included in this manual elaborate the step by step procedure with URL link and illustrations.

We thank the **Department of Biotechnology, Ministry of Science and Technology, MHRD, New Delhi** for providing a good opportunity under Star College Scheme (No.HRD11011/163/2020-HRD-DBT Dt.24.8.2020).

We hope this manual will definitely meet out the student's needs to perform the experiments that enhance their research skills in the drug development.

Member Secretary/Coordinator **Chairman/Principal**

FOREWORD

Bioinformatics is an evergreen, emerging interdisciplinary field in life science. It provides a medium for the exchange of information in the fields of computational molecular biology and the post-genome era, with an emphasis on the documentation of large data sets and databases that allow the progress of biomedical research in a significant manner.

This lab manual presents a collection of twenty practical exercises, aimed at providing standard protocols to access nucleic acid and protein sequence databases, perform sequence alignments, predict secondary structure, and visualize proteins. This manual introduces the theory and provides a systematic procedure to facilitate students to carry out the practical exercises in an easy manner. Overall understanding NCBI resources, accessing biological sequences from GenBank, performing sequence alignments, and using protein visualization tools will become easy if the students and researchers use this manual effectively. I assure that this will become a handy tool and motivation factor for the basic analysis of sequences.

When the readers finish practicing the exercises in this manual, they will possess the knowledge to use the various online tools available for processing the biological data and will have a platform to develop their skills in handling bioinformatics techniques in the future.

P. Sugapriya Menaga
Reviewer
Assistant Professor of Biotechnology,
ANJAC, Sivakasi.

PREFACE

The field of bioinformatics, which applies both information technology and biological science make to prediction about biological processes using bioinformatics tools. Moreover bioinformatics is an essential part of biological research and has applications in different practical fields.

This bioinformatics manual helps the beginners to learn step-by-step procedure that will make the learners to follow the experiments easily. This manual covers the curriculum for the Biotechnology, Microbiology and Biochemistry students at the undergraduate level.

I would like to express my heartfelt gratitude to the reviewer, Mrs.P.Sugapriya Menaga, Assistant Professor, Department of Biotechnology, ANJAC, Sivakasi for her useful suggestions to improve the quality of this manual.

K.Sudha Rameshwari
Assistant Professor,
Department of Biochemistry,
V.V.Vanniaperumal College for Women,
Virudhunagar-626001,
Tamilandu, India.

BIOINFORMATICS LABORATORYMANUAL

LIST OF PRACTICALS

INTRODUCTION TO BIOINFORMATICS

Bioinformatics is a newly emerging scientific discipline for the computational analysis and storage of biological data. Bioinformatics is the field in which biology, computer science and information technology merges into a single discipline for managing and analyzing biological data using advanced computing techniques.

OBJECTIVES

- To organize data, to access existing information and to submit new entries as they are produced.
- To develop tools and resources that aid in the analysis of data
- To analyze the data and interpret the results in a biologically meaningful manner

SCOPE OF BIOINFORMATICS

Biological and medical labs use methods that produce extremely large data sets, which cannot be analyzed for instance sequencing human genomes. Thus, modern biological and medical research and development cannot be carried out without bioinformatics. In addition, bioinformatics plays an important role in biomedical research. Research in the area of genetic diseases and medical genomics is rapidly developing and the future of

personalized medicine depends on the application of bioinformatics approaches.

This course focuses on employing existing bioinformatics resources - mainly web-based programs and databases - to access the wealth of data to answer questions relevant to the average biologist, and is highly hands-on. Different types of career opportunities are available for the students of bioinformatics like Scientific Curator, Gene Analyst, Protein Analyst, Phylogenitist, Research Scientist / Associate, Data Base programmer, Bioinformatics software developer, Computational biologist, Network Administrator / Analyst, Structural Analyst, Molecular Modeler, Biostatistician, Biomechanics, Database programmer, Cheminformatician, Pharmacogenetician, Pharmacogenomics etc.

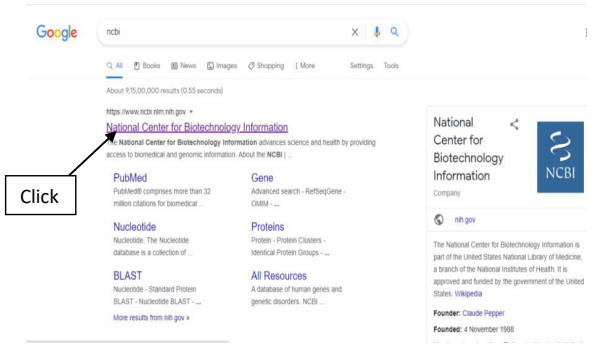


Figure 1: Open the web browser

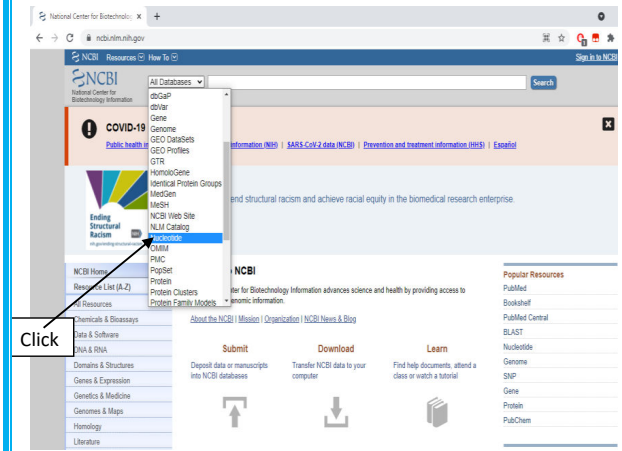


Figure 2: NCBI Home page

RETRIEVAL OF NUCLEOTIDE SEQUENCE FROM GENBANK

AIM

To retrieve the gene from Genbank and to save the sequence in FASTA format

INTRODUCTION

Entries of nucleotide sequence are stored in Genbank nucleotide database. Each entry contains complete information about the particular gene. Genbank website is accessible for anyone who is interested in getting the entry for research.

PROCEDURE

- Type NCBI in the web browser and click search, click National Center for Biotechnology Information, it directs to the URL : https://www.ncbi.nlm.nih.gov/ (Figure 1) OR
- Type the URL https://www.ncbi.nlm.nih.gov/ directly in the space for address in the address bar and press the enter key.
- NCBI homepage will appear.
- Click the All Databases drop –up menu and drag the bar and select nucleotide (Figure 2)
- Search list will be displayed, click the suitable accession number or any gene of interest (figure 4)

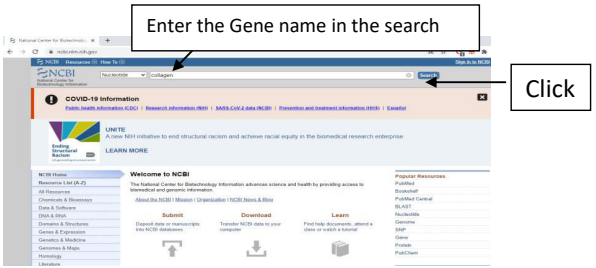


Figure 3: Home page of NCBI the gene enter in the search area

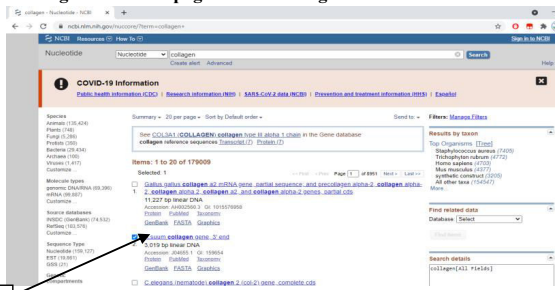


Figure 4: search list page

Figure 5: Genebank entry format

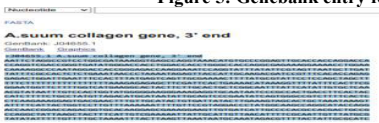


Figure 6: FASTA sequence in new window

- Click the gene of interest/accession number (Figure 4)
- Type Collagen in the search area and click (Figure 3)
- A new window will appear and shows the entry of the collagen gene in detail (Figure 5)
- Click the FASTA, FASTA sequence appears in the new window (Figure 6)
- Copy the FASTA sequence and paste it in note pad
- Save it for further investigation.

RESULT

The collagen gene was retrieved and saved as FASTA format in notepad.

OUTCOME

Students acquire the knowledge to retrieve any gene from Genbank. From the retrieved sequence flat file, details of sequences, submitter's details, biological significance, and the scientific name and taxonomy of the organism is understood. A feature table shows characteristics that indicate coding regions, transcription units, mutation locations, etc. Retrieved gene sequence is useful for gene analysis. It can be compared with other sequences to determine which gene is mutant. It enlightens the diagnosis of hereditary disease.

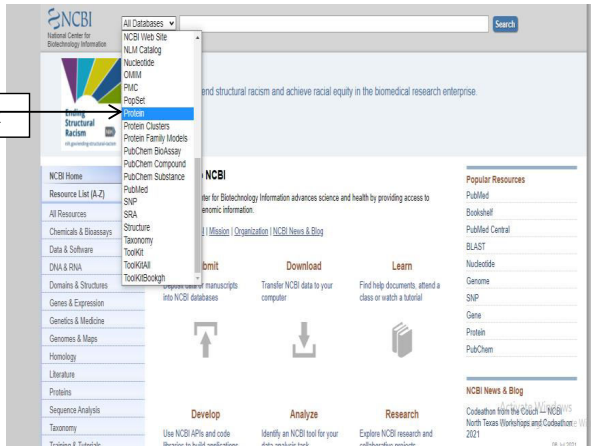


Figure 7: Home page of NCBI, Selection of Protein

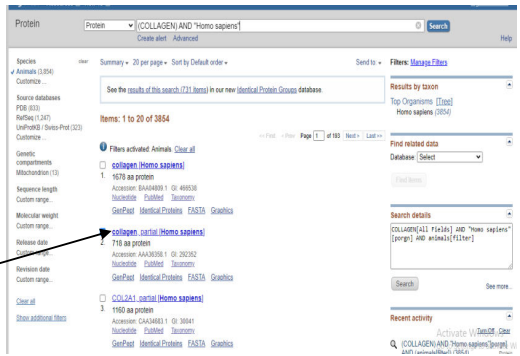


Figure 8: Search list page of protein

RETRIEVAL OF PROTEIN SEQUENCE FROM GENBANK

AIM

To retrieve the protein from Genbank and to save the sequence in FASTA format

INTRODUCTION

Entries of protein sequence are stored in Genbank database. Each entry contains complete information about a particular protein. Genbank website is accessible for anyone who is interested in getting the entry for research.

PROCEDURE

- Type NCBI in the web browser and click search, it shows National Center for Biotechnology Information, click National Center for Biotechnology Information, it directs to the url : https://www.ncbi.nlm.nih.gov/ or home page of NCBI (Figure 7)

  Or
- Type the url: https://www.ncbi.nlm.nih.gov/ web/Genbank/ directly in the address bar and press the enter key
- NCBI homepage will appear
- Click the All Databases drop –up menu, and drag the bar to select protein (Figure 8)

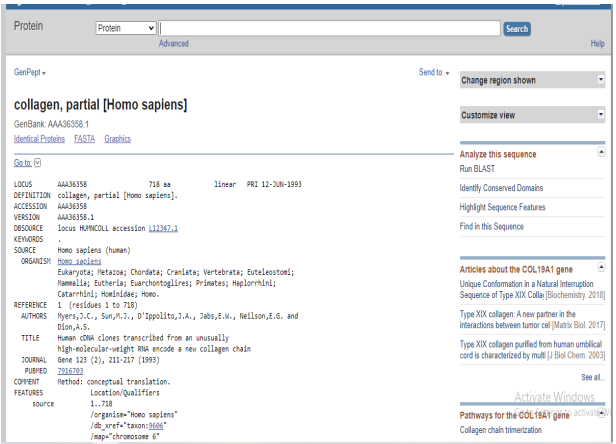


Figure 9: Collgen gene Genbank entry page

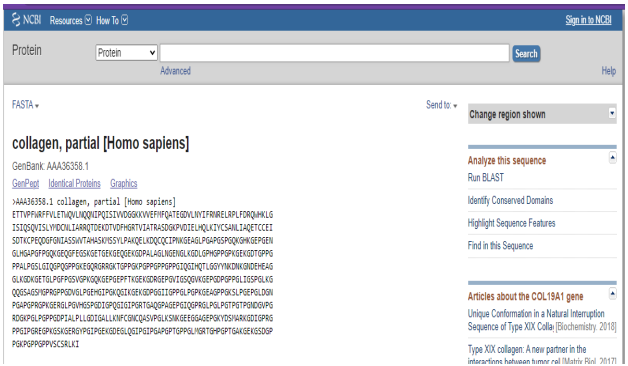


Figure 10: Fasta format Page

- Type collagen in the search area and press enter key or click search
- Search list will be displayed (Figure 8)
- Select the suitable accession number that describes collagen
- Click the accession number of collagen
- A new window will appear and shows the entry of the collagen gene (Figure 9)
- In display format, in place of summary choose “FASTA”
- Only sequences will be shown (Figure 10)
- Copy the sequence (Figure 11) and paste it on note pad (Figure 12) and save it for the further investigation

RESULT

The Protein sequence was retrieved and saved as FASTA format in notepad.

OUTCOME

Students master the technique of sequence retrieval which is the basic step in bioinformatics. From the retrieved sequence flat file, details of sequences, submitter’s details, biological significance, and the scientific name and taxonomy of the organism are understood. Sequence retrieval is essential for the analysis of primary, secondary and tertiary structure of any protein sequence.

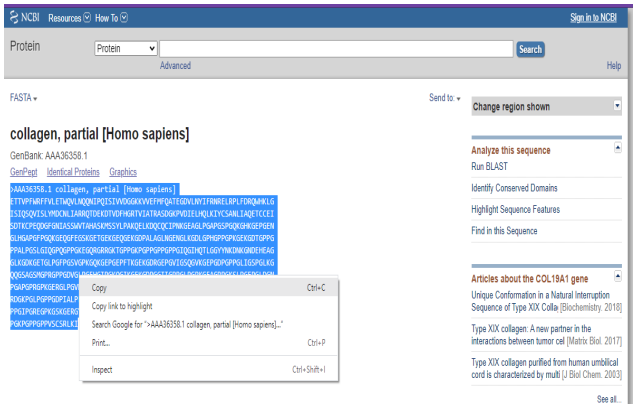


Figure 11: Copy the Fasta Sequence

Figure 12: Paste the Fasta Sequence in the note pad

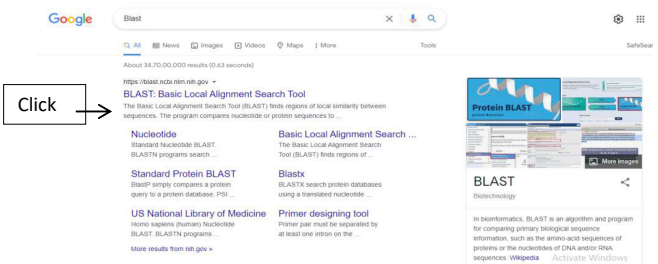


Figure 13: Search list of BLAST

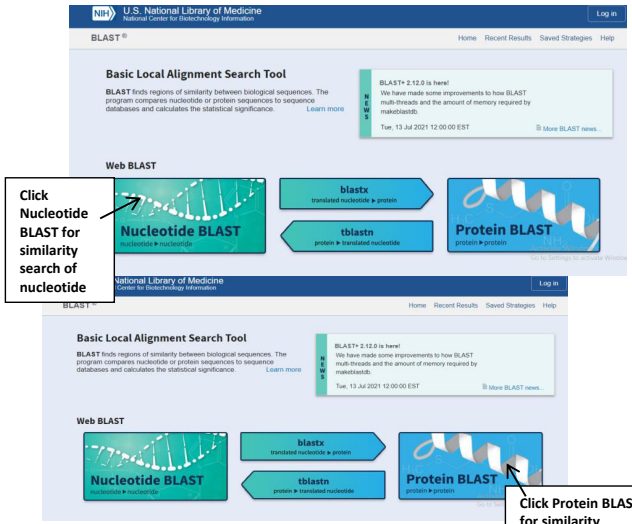


Figure 14: BLAST home page & Select Protein blast

SIMILARITY SEQUENCE SEARCH USING BLASTN

AIM

To find the similarity of sequence for the given nucleotide or protein sequence

INTRODUCTION

BLAST is **B**asic **L**ocal **A**lignment **S**earch **T**ool. It is a technique for finding homology and similarity. It is a tool for searching of sequences that are similar to one another in databases. By matching the novel sequence with previously defined genes, it compares novel gene sequences with nucleotide databases. This tool focuses on identifying areas of sequence similarity. It will provide information on the structure and function of the novel sequence. Instead of using the best alignments, it looks for areas of sequence similarity. It produces ungapped alignments. It reports multiple local alignments between the query and database. It is based on an explicit statistical theory.

PROCEDURE

- Type BLAST in the web browser and click search
- Click the first link, BLAST (Figure 13)
- Go to Blast Home page (Figure 14)
- Click nucleotide blast BLASTN (Figure 15)

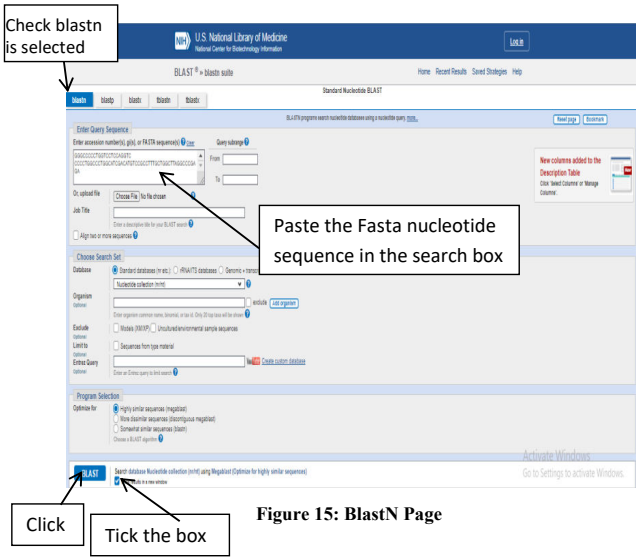


Figure 15: BlastN Page

Depending on the type of sequence, the programs in BLAST differs. They are

- **Blastp** compares an aminoacid query sequence against a protein sequence database
- **Blastn** compares an nucleotide query sequence against a nucleotide sequence database
- **Blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **tblastx** compares the six frame translation of a nucleotide query sequence against the six frame translation of a nucleotide sequence database.

- Paste the sequence (FASTA format) in the given box of BLASTN or paste the accession number (**Figure 15)**
- Tick the box "results in new window" in the last line of BLASTN Home page (**Figure 15)**
- Click BLAST (**Figure 15)**
- In the new window, format request display is shown, wait for few seconds, it leads to result page

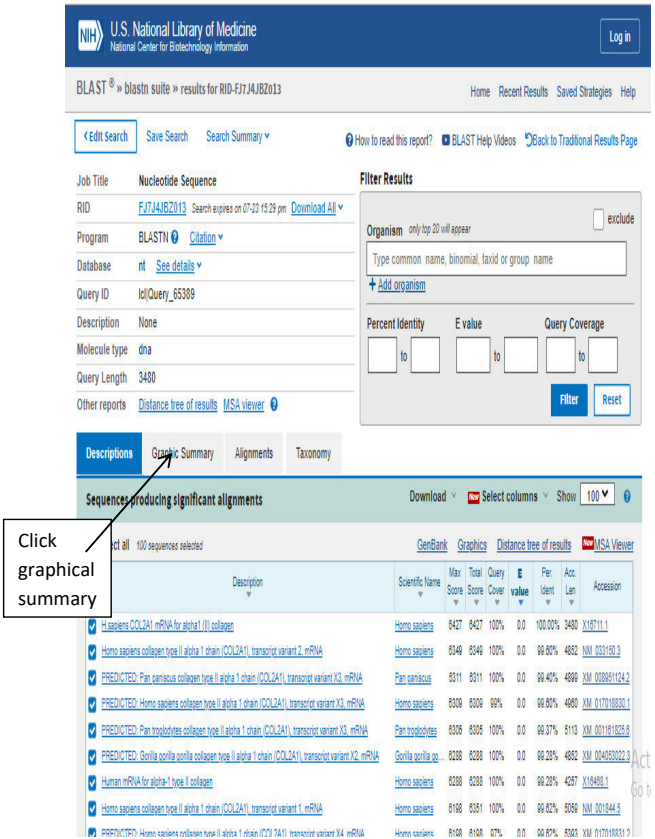


Figure 16: Result window of BLASTN

- In the result page, the blast results for sequence similarity search appear as graphically and as text (descriptions, alignment)
- Click each text and find the similarity between the two sequences
- Click graphic summary (**Figure 16**)
- View the graphical representation of sequence alignment
- Lines in pink, red colors represent the sequence in a set of score values.
- Length of the line indicates the length of the local alignment with the query sequences. (Figure 17)

RESULT

The most significant and similar sequences fetched by the blast are Homo sapiens collagen type II alpha 1 chain (COL2A1), transcript variant 2, mRNA,

OUTCOME

By learning BLASTN, students obtain the skill to compare the new gene sequence with the nucleotide database by aligning the novel sequence with previously characterized gene/protein. From the comparison, students gather the functional and evolutionary clues about the structure and function of the novel sequence.

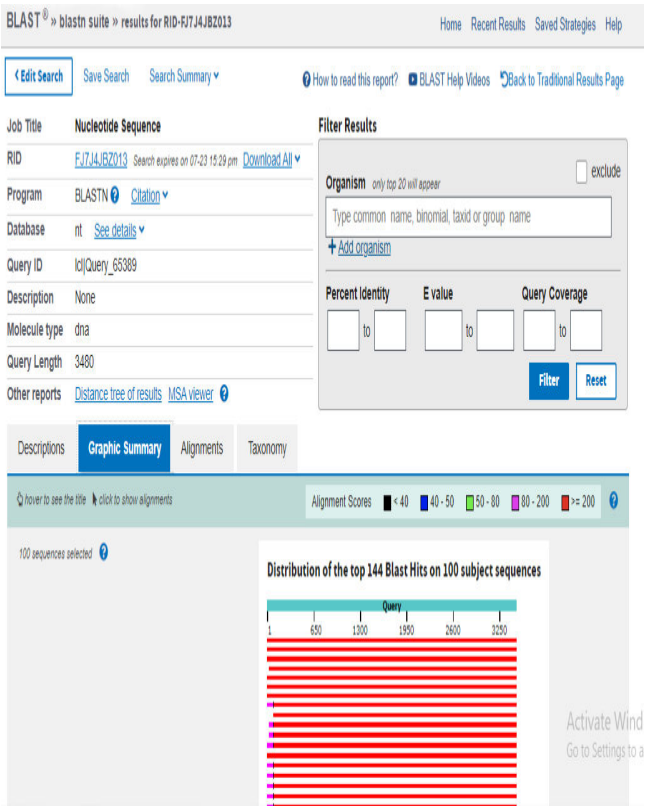


Figure 17: Result page of Graphical summary

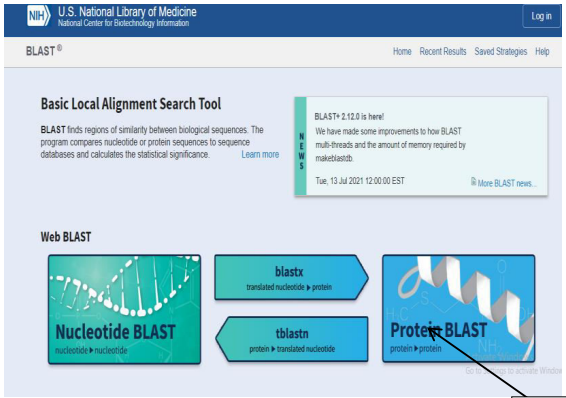


Figure 18: Select Protein blast

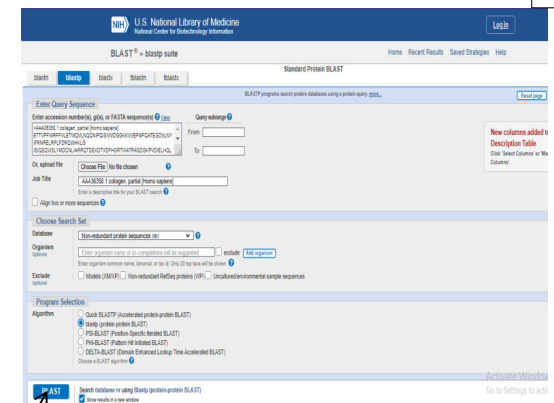


Figure 19: BlastP Page

SIMILARITY SEQUENCE SEARCH USING BLASTP

AIM

To find the similarity of sequence for the given protein sequence

INTRODUCTION

BLAST is **B**asic **L**ocal **A**lignment **S**earch **T**ool. It is a technique for finding homology and similarity of protein sequences. It is made for the Windows platform and is employed to carry out DNA or protein similarity searches. The source is NCBI. It is a tool for searching for sequences that are similar to one another in databases. This tool focuses on identifying areas of sequence similarity. It is used to compare a novel protein sequence against protein database by aligning the novel sequence with previously characterized protein. The emphasis of this tool is to find regions of sequence similarity. It will yield functional and evolutionary clues about the structure and function of the novel sequence. It finds out patches of sequences similarity rather than best alignments. It produces un gapped alignments. It reports multiple local alignments between the query and database. It is based on an explicit statistical theory.

PROCEDURE

- Go to NCBI Blast page
- Chose the Blast program for proteins BLASTP (Figure 18)

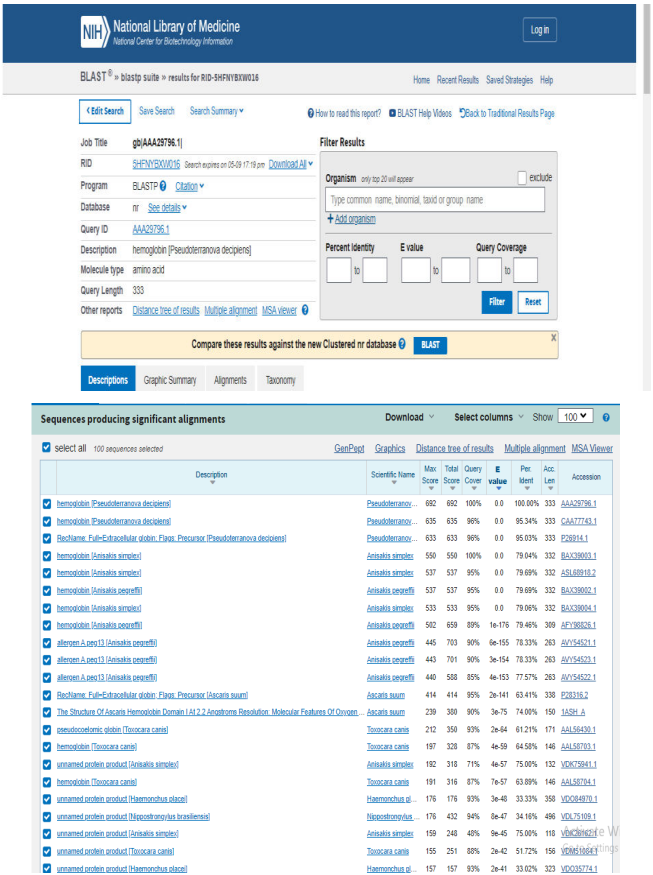


Figure 19a: Result page of Blastp

- Copy the sequence from the notepad. Avoid the first FASTA line. Paste the protein sequence in the search window of BLASTP. (Figure 19)
- Tick the box "results in new window", in the last line of blast Home page
- Click BLAST
- In the new window, format request display is shown, wait for few seconds, it leads to result page (Figure 19a)
- In the result page, the blast results for sequence similarity search appear as graphically and as text
- Click each text and find the similarity between sequence
- Click graphic summary
- View the graphical representation of sequence alignment (Figure 19b)
- Lines in pink, red colors represent the sequence in a set of score values.
- Length of the line indicates the length of the local alignment with the query sequences.

RESULT

The most significant and similar sequences fetched by the blastp are **hemoglobin [*Pseudoterranovadecipiens*],**CAA77743.1

Figure 19b: Graphical summary page

OUTCOME

By learning BLASTP, students gain the knowledge to compare the new protein sequence with the protein database by aligning the novel sequence with previously characterized protein. From the comparison, students can obtain functional and evolutionary clues about the structure and function of the novel sequence.

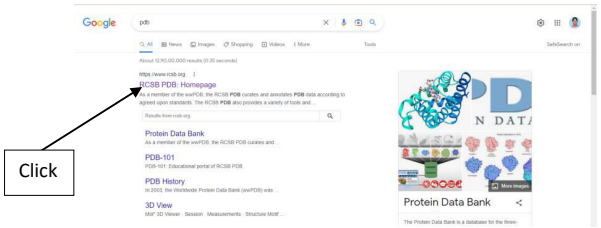


Figure 20 : Web browser page

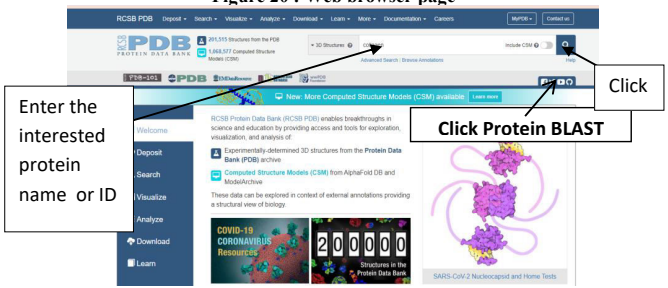


Figure 21: PDB Home page

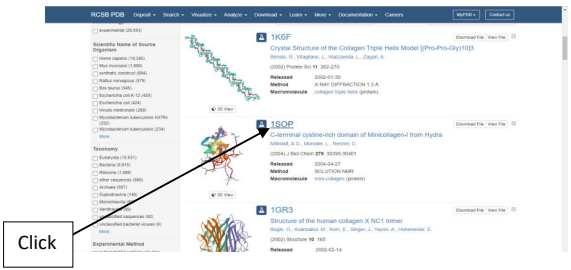


Figure 22: Result summary page

ACCESSING STRUCTURAL DATABASE AND DOWNLOADING THE PROTEIN STRUCTURE

AIM

To access the PDB (Protein Data Bank) structural database and to download the protein structure.

INTRODUCTION

The three dimensional structure of biomolecules plays an important role in the functions and maintenance of the structural features of an organism. These structures were deciphered by research scientists and deposited in the databases, specifically designed for structural submission, for worldwide use. Research Collaborator for Structural Bioinformatics (RCSB), manages of the PDB. It provides free resources to assist the fields of biology and bioinformatics. It provides detailed information about sequence, atomic coordinates, structure factors, and crystallization conditions etc.

PROCEDURE

- Type PDB in the web browser search field (Figure 20)
- Click the RCSB page (Figure 20)
- Go to the WWW.rcsb.org (Figure 21)
- Enter the protein name of interest in the search box
- Click Go

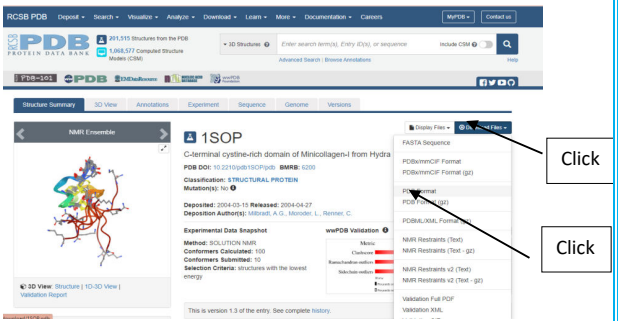


Figure 23: Download the protein structure in PDB format

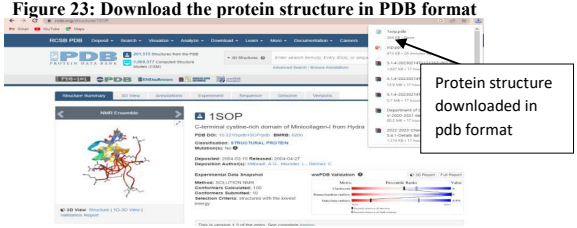


Figure 24: Protein structure was downloaded

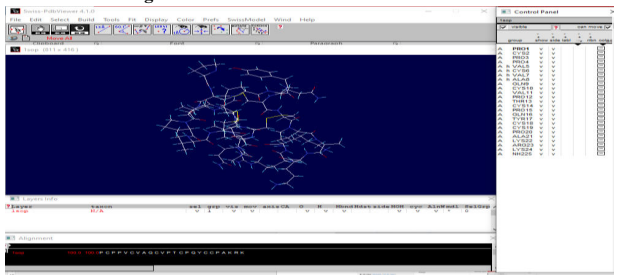


Figure 25: Downloaded protein structure is visualized in Molecular visualization tool (Rasmol and Swisspdbv)

- Choose the protein name of interest (Figure 22)
- Click the ID
- Click the download file (Figure 23)
- Click PDB format (Figure 23)
- Download file as PDB format (Figure 24)
- Save file
- View the protein structure in the molecular visualization tool like RASmol, Swisspdbv (Figure 25)

RESULT

The protein structure was downloaded and their descriptions were observed.

OUTCOME

Students get trained to retrieve any protein structure from structural protein database and can understand the characteristics of the protein structure.

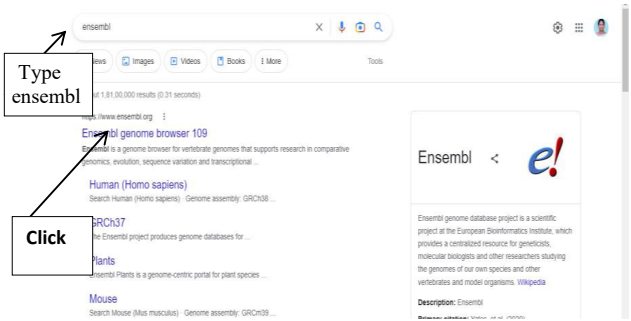


Figure 26: Google search page

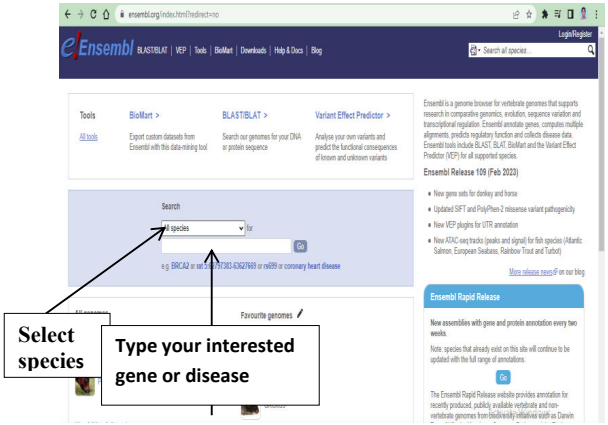


Figure 27: Ensembl home page

WORKING WITH ENSEMBL

AIM

To retrieve the vertebrate genomic information

INTRODUCTION

Ensembl is one of the several well-known genome browsers for the retrieval of genomic information. It is considered to be the universal information source for the human genome. Data available in Ensembl include genes, SNPs, repeats and homologies. Genes may either be known experimentally or deducted from the sequence; because the experimental support for annotation of the human genome is so variable. It presents the evidence for identification of every gene. Extensive linking to other databases containing related information such as OMIM or expression databases is also available. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

PROCEDURE

- Type Ensembl in the web browser search field
- Click Ensembl genome browser 109 (Figure 26)
- Go to the home page, First choose the species of interest and type gene of interest or disease name (Figure 27 & 28) and click go

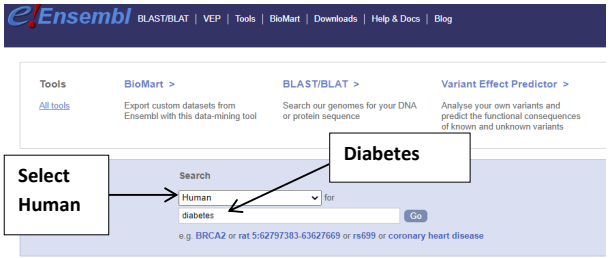


Figure 28: Ensembl home page

Figure 29: List of diabetes gene

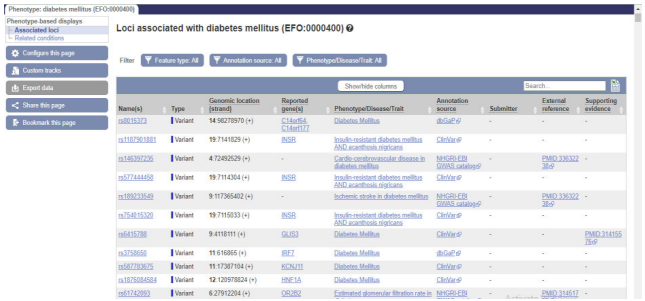


Figure 30: Details of diabetes were listed

- List of genes were displayed (Figure 29)
- Select one and click
- Details of diabetes were shown (Figure 30)
- Click hyperlink one by one; the details of variant, genomic location, reported genes, phenotype/genotype trait, annotation source, submitter, and external reference were known. (Figure 30)

RESULT

The diabetic gene annotations were retrieved from the Ensembl genome browser.

OUTCOME

Students can retrieve any vertebrate gene and its details from the Ensembl genome browser.

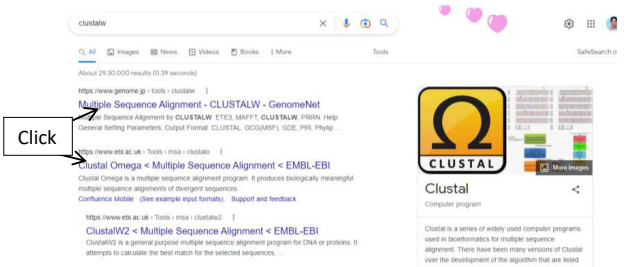


Figure 31: Google Search

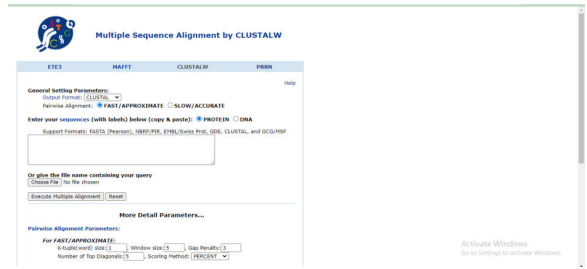


Figure 32: ClustalW Home page

Figure 33: sequences in the notepad

MULTIPLE SEQUENCE ALIGNMENT

AIM

To study the closely related genes or proteins

INTRODUCTION

ClustalW tool is used for aligning multiple nucleotide or protein sequences. It uses progressive alignment methods, which align the most similar sequences first and work their way down to the least similar sequences until a global alignment is created. A multiple sequence alignment tools called Clustal Omega creates alignments between three or more sequences by using HMM profile-profile algorithms and seeded guide trees. Multiple sequence alignment is a tool used to study closely related genes or proteins in order to find the evolutionary relationships between genes and to identify shared patterns among functionally or structurally related genes.

PROCEDURE

- Type multiple sequence alignment tool in the web browser search field (Figure 31)
- Click Clustal Omega < Multiple Sequence Alignment < EMBL-EBI Or
- Click Multiple Sequence Alignment-CLUSTALW – Genome Net ((Figure 31)

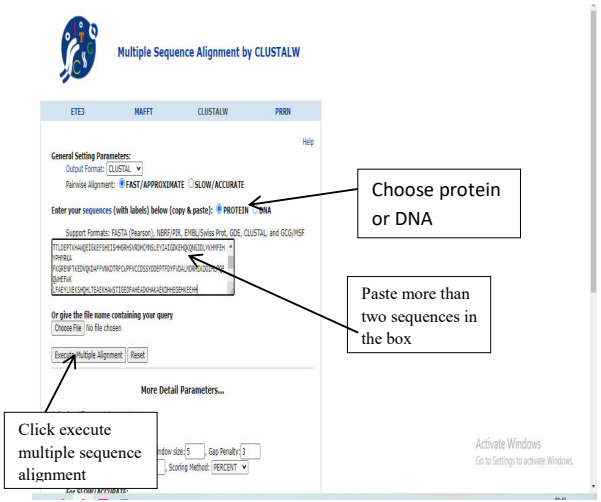


Figure 34: Paste the sequences in the box

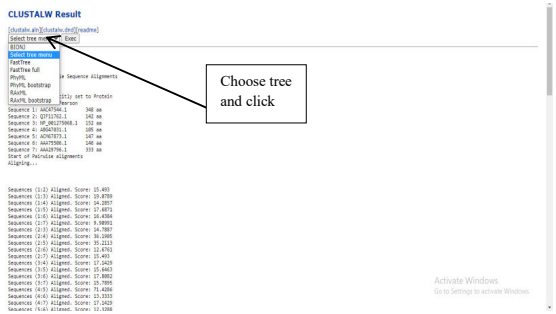


Figure 35: Result page

- Go to home page (Figure 32)
- Copy the sequences from the notepad (Figure 33)
- Paste the more than two sequences in input sequences box (Figure 34)
- Choose the sequence as protein or DNA (Figure 34)
- Click submit or execute multiple sequence alignment (Figure 34)
- Result page is shown (Figure 35)
- Choose fast tree and click
- Phylogenetic tree was displayed (Figure 36)

RESULT

- The close relationship for this NP_001275068.1 (any one Id which sequences are pasted in input sequence box) is NP_001104966.1.
- The far relationship for this NP_001275068.1 (any one Id which sequences are pasted in input sequence box) is AAA29796.1.

OUTCOME

Using this multiple sequence tool, students attain the ability to infer the evolutionary relationships between the sequences under study.

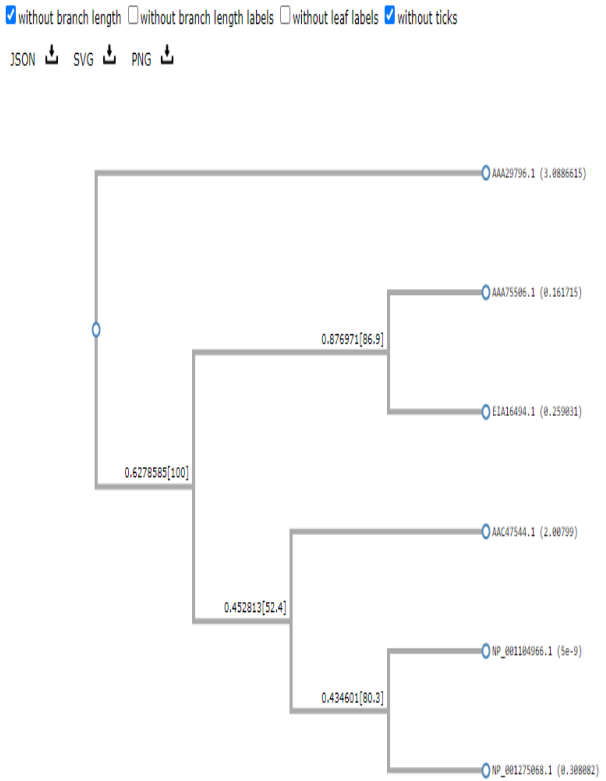


Figure 36: Phylogenetic tree

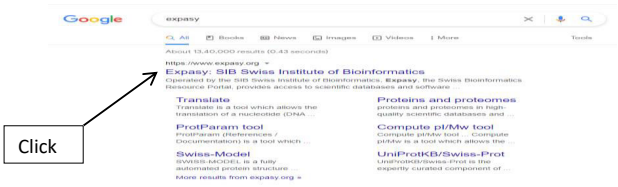


Figure 37: search bar

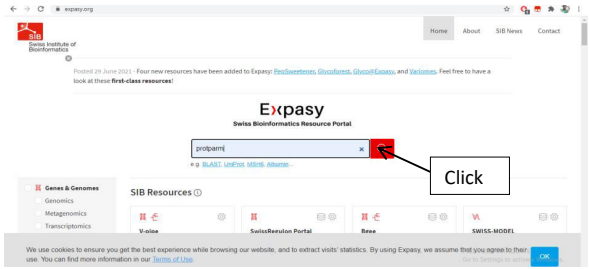


Figure 38: Home page of ExPASy

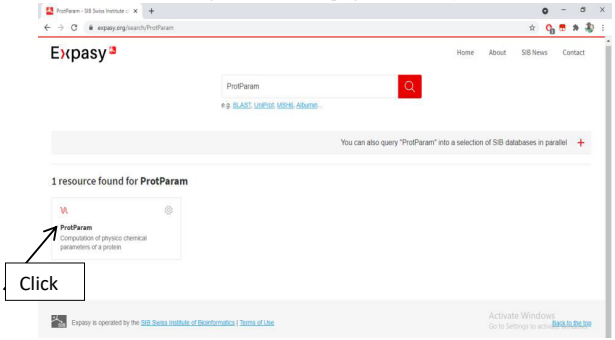


Figure 39: Search Results Page

PREDICTING PHYSIOCHEMICAL PROPERTIES OF PROTEIN SEQUENCE

AIM

To predict the physiochemical properties of protein sequence using ExPASy resources.

INTRODUCTION

With the help of the tool ProtParam, users may compute a number of physical and chemical parameters for a specific protein that is contained in Swiss-Prot or TrEMBL, as well as for a user-entered protein sequence. Molecular weight, theoretical pI (Isoelectric point), amino acid composition, atomic composition, extinction coefficient, assumed half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY) are among the calculated characteristics.

PROCEDURE

- Type ExPASy in the web browser search box (Figure 37)
- Click ExPASy
- Go to ExPASy page, enter specific proteomic tool in search
- box (Figure 38) Or
- Enter the url: https://web.ExPASy.org/ProtParam/
- For predicting physiochemical properties, type ProtParam and enter

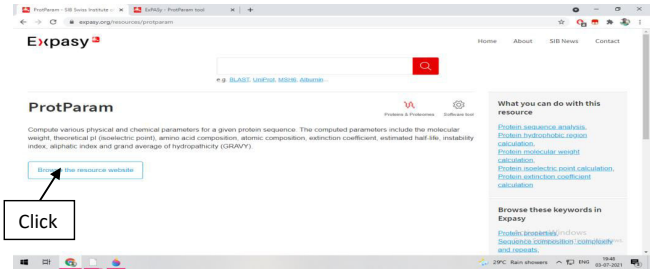


Figure 40: Resource page

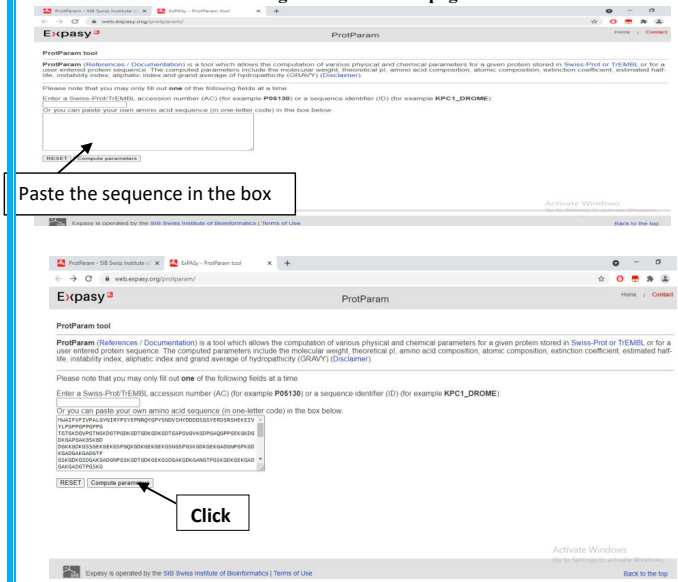


Figure 41: sequence pasted in the ProtParam tool box

- Select protoparm tool (Figure 39)
- Click "browse the resource website", protoparm tool page is opened in new window (Figure 40)
- Paste the protein sequence / accession number in the given box (Figure 41)
- Click compute parameters (Figure 41)

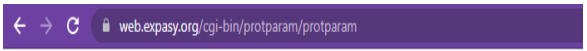
RESULT

Using ProtParam tool, the physicochemical properties of protein sequence were predicted as follows (Figure 42):

- Number of aminoacids:718
- Molecular weight: 72106.72
- Theoretical pI**:** 7.31
- Total number of negatively charged residues (Asp + Glu): 77
- Total number of positively charged residues (Arg + Lys): 77

OUTCOME

Students learnt the art of analysing the physical and chemical properties for a specific protein.



Atomic composition:

Carbon	C	3150
Hydrogen	H	4974
Nitrogen	N	926
Oxygen	O	986
Sulfur	S	16

Formula: $C_{3150}H_{4974}N_{926}O_{986}S_{16}$
Total number of atoms: 10052

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 34545
Abs 0.1% (=1 g/l) 0.479, assuming all pairs of Cys residues form cystines

Ext. coefficient 33920
Abs 0.1% (=1 g/l) 0.470, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is E (Glu).

The estimated half-life is: 1 hours (mammalian reticulocytes, in vitro).
30 min (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to be 24.32
This classifies the protein as stable.

Aliphatic index: 55.43

Grand average of hydropathicity (GRAVY): -0.777

Figure 42: Result page of Protoparam tool

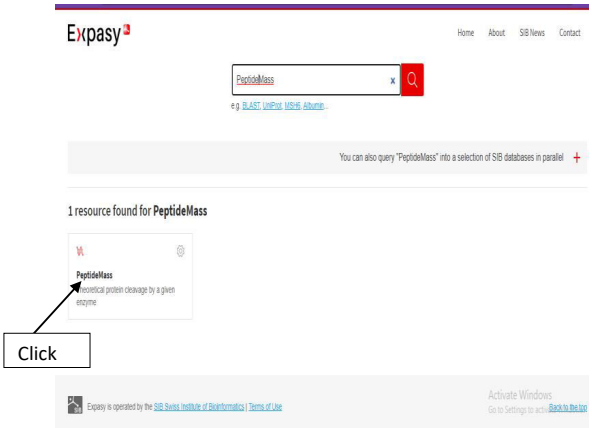


Figure 43: ExPASy Proteomic tool search site

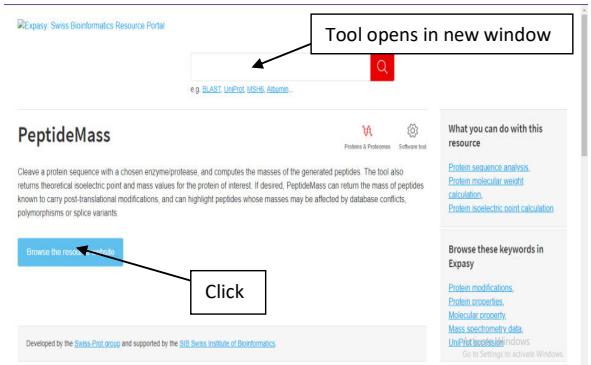


Figure 44 : Peptide Mass resources site

PREDICTING PEPTIDE MASS OF PROTEIN SEQUENCE

AIM

To predict the peptide mass of protein sequence using ExPASy resource

INTRODUCTION

In the analytical method of protein identification known as peptide mass fingerprinting (PMF), the unknown protein of interest is first broken up into smaller peptides whose absolute masses may be precisely determined with a mass spectrometer like the MALDI-TOF or ESI-TOF.

PROCEDURE

- Type ExPASy in the web browser search field
- Click ExPASy
- Go to ExPASy page, Enter Peptide Mass in the search field (Figure 43) Or
- Type url : https://web.ExPASy.org/peptide_mass/
- Click peptide mass tool
- It leads to peptide mass resource page
- Click the “Browse the resource website”, Peptide mass tools site opens in new window (Figure 44)
- Paste the protein sequence of interest

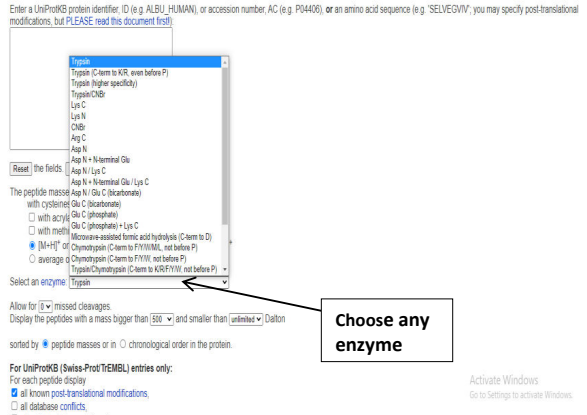


Figure 45: Peptide Mass tool site showing different enzymes for selecting to cleave protein sequence

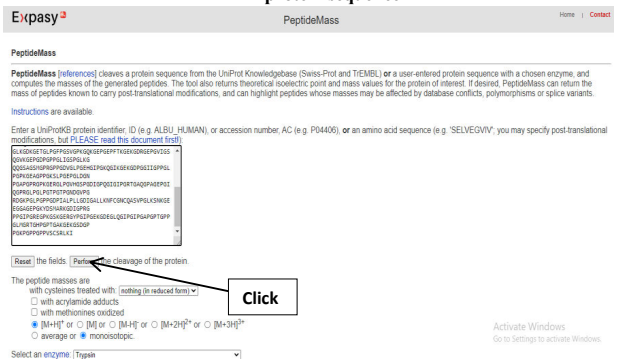


Figure 46: PeptideMass tool site with pasted sequence and click to perform

- Move the cursor to choose which type of enzyme to cleave the interested protein sequence (Figure 45)
- Select "trypsin" from the drop-down box under "Select enzyme: in the tool," if we want other enzyme use drop menu and select the desired enzyme (Figure 45)
- Click perform (Figure 46)
- Result page shown (Figure 47)

RESULT

- The high molecular weight peptide mass sequence is TGPPGKPGPPGPPGPPGPIQGIHQTLGGYYNK, position 309-339and its molecular mass is 3036.5689
- The low molecular weight peptide mass sequence is QELK, position 174-177 and its molecular weight is 517.2980.

OUTCOME

Students can identify the molecular weight of peptide after cleavage of protein using different enzymes or chemicals.

The selected enzyme is: Trypsin

Maximum number of missed cleavages (MC): 0

All cysteines in reduced form.

Methionines have not been oxidized.

Displaying peptides with a mass bigger than 500 Dalton.

Using monoisotopic masses of the occurring amino acid residues and giving peptide masses as [M+H]+.

The peptide masses from your sequence are:

[Theoretical pI: 7.31 / Mw (average mass): 72106.72 / Mw (monoisotopic mass): 72062.31]

mass	position	#MC	modifications	peptide sequence
3036.5689	309-339	0		TGPPGKPGPPGPPGPPGIQG IHQTLGGYYNK
2972.5880	9-34	0		FFVLETWQVLNQQNIPQISI VVDGGK
2591.4806	562-588	0		DGKPGLPGPPGDPIALPLLG DIGALLK
2398.2183	657-682	0		GDEGLQGIPGIPGAPGPTGP PGLMGR
2349.2196	274-299	0		GDTGPPGPPALPGSLGIQGP QGPPGK
2337.2305	69-89	0		LGISIQSQVISLYMDCNLIA R
2278.1212	36-54	0		VVEFMFQATEGDVLNYIFR
2205.0029	145-165	0		CPEQDGFGNIASSWVTAHAS K
2204.9872	125-144	0		IYCSANLIAQETCCEISDTK
2038.0828	504-525	0		GLPGVHGSPGDIGPQGIGIP GR
1845.8911	696-715	0		GSDGPPGKPGPPGPPVSCSR
1815.9347	542-561	0		GLPGLPGTPGTPGNDGVPGR
1748.8350	205-222	0		GEPGENGLHGAPGFPGQK
1687.8398	480-497	0		SLPGEPGLDGNPGAPGPR
1492.7502	526-541	0		TGAQGPAGEPGIQGPR
1437.6613	589-602	0		NFCGNCQASVPGLK
1428.7845	456-471	0		GDPGGIIGPPGLPGPK
1425.7332	244-258	0		GDPALAGLNGENGLK
1399.7215	357-371	0		GETGLPGFPGSVGPK
1322.6698	187-201	0		GEAGLPGAPGSPGQK
1132.5164	420-431	0		GQQGSAGSMGPR
1128.5796	259-270	0		GDLGPHGPPGPK
1127.6055	390-401	0		GEPGVIGSQGVK
1126.5123	343-353	0		GNDEHEAGGLK
1048.4550	178-186	0		DQCQCIPNK
1045.5789	57-64	0		ELRPLFDR
1035.5258	1-8	0		ETTVPFWR
987.4377	606-616	0		GEEGSAGEPGK
961.4625	375-383	0		GEPGEPTIK
946.4377	96-103	0		DTVDFHGR
938.4214	223-231	0		GEQGFEGSK
922.4741	683-692	0		TGHPGPTGAK
917.4727	648-656	0		GYPGIPGEK
896.4546	166-173	0		MSSYLPAK
750.4257	630-637	0		GPPGIPGR
742.3188	617-622	0		YDSMAR
712.3624	472-479	0		GEAGPPGK
660.4039	104-109	0		TVIATR
647.2995	238-243	0		GEQGEK
620.2886	91-95	0		QTDEK
620.2886	232-237	0		GETGEK
614.3256	624-629	0		GDIGPR
598.3096	65-68	0		QWHK
517.2980	174-177	0		QELK

90.7% of sequence covered (you may modify the input parameters to display also peptides < 500 Da or > 100000000000 Da).

Activate Windows
Go to Settings to activate Windows.

Figure 47: Result page of peptide mass

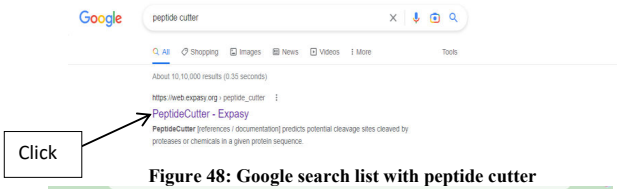


Figure 48: Google search list with peptide cutter

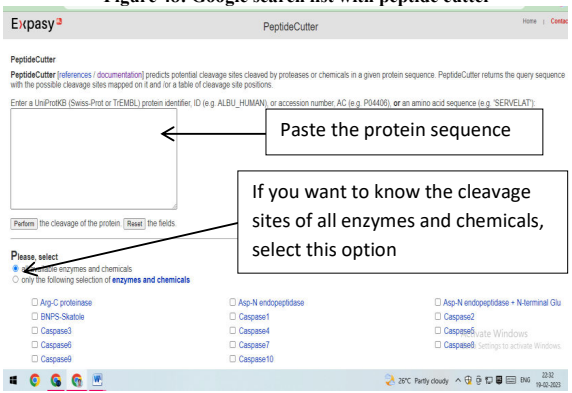


Figure 49: Peptide cutter Home page

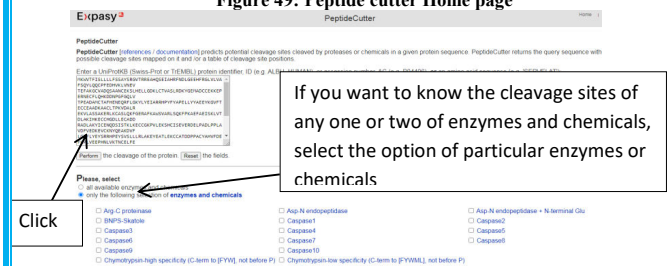


Figure 50: Paste the sequence and click to perform

PREDICTING CLEAVAGE SITE OF PROTEIN SEQUENCE

AIM

To predict the cleavage site of protein sequence using ExPASy resource (Peptide cutter).

INTRODUCTION

Peptide Cutter explores for protease cleavage sites in a protein sequence given by the user, a protein sequence from the SWISS-PROT and/or TrEMBL databases, or both. Users can use a protease, a group of proteases, or the entire list of proteases and compounds. Different forms of output of the results are available: Tables of cleavage sites that are either ordered alphabetically by the names of the enzymes or sequentially by the number of amino acids are available. A map of the cleavage locations is a third possible result. The user can choose the block size to print out the sequence and the cleavage sites that have been mapped onto it.

PROCEDURE

- Type Peptide cutter in the web browser search field and press enter
- Click Peptide cutter tool (Figure 48) Or Type url : https://web.ExPASy.org/peptide_cutter/

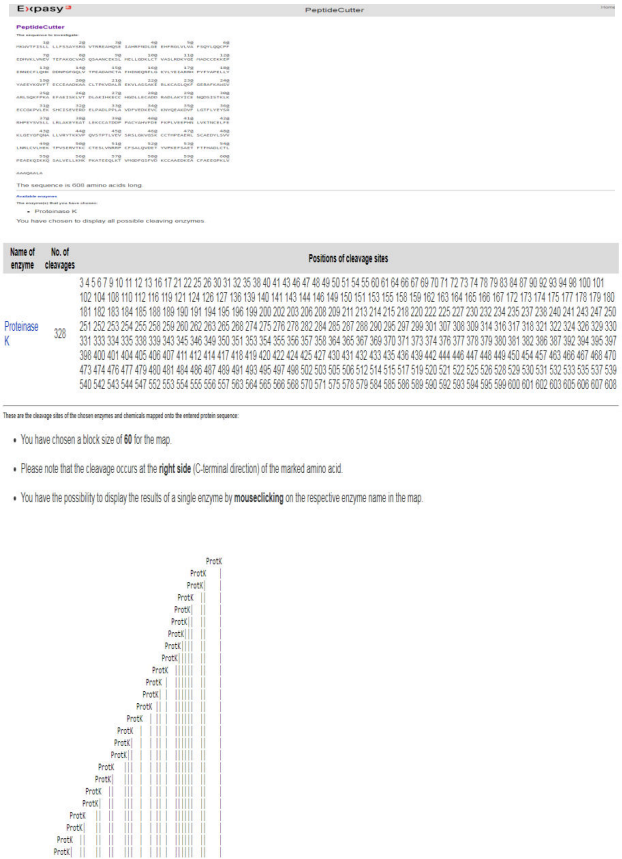


Figure 51: Result page

- Go to the peptide cutter tool home page (Figure 49)
- Paste the protein sequence of interest in the given box
- Move the cursor to choose the enzyme type to cleave the sequence of interest or select all available enzymes and chemicals; Click perform, (Figure 50)
- Result page is displayed (Figure 51)

RESULT

Proteinase K enzyme (Specific enzyme choose) cleaves 328 sites in a given protein sequence.

OUTCOME

Students can predict the potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. Peptide Cutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

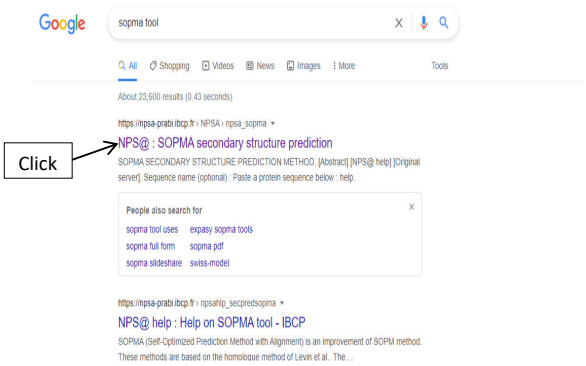


Figure 52: Google search list with secondary structure prediction tool

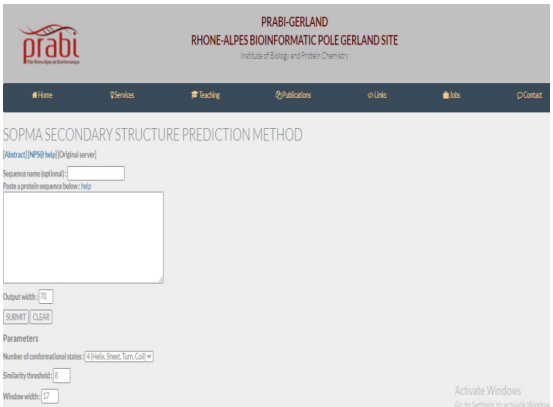


Figure 53: Home page of SOPMA tool

PREDICTING SECONDARY STRUCTURE OF PROTEIN SEQUENCE USING SOPMA TOOL

AIM

To predict the secondary structure of protein sequence using ExPASy resources

INTRODUCTION

Secondary structure prediction is a group of techniques in bioinformatics that aim to calculate the secondary structures of proteins and nucleic acid sequences based on the information from their basic structures (primary structures). Through base pairing and base stacking interactions, it predicts the formation of nucleic acid structures like helixes and stem-loop structures while predicting the formation of protein structures like alpha helices and beta strands for proteins. The Self-Optimized Prediction Method with Alignment (SOPMA) is a tool to predict the secondary structure of a protein. Based on the query (primary sequence of a protein), SOPMA will predict its secondary structure. Protein secondary structure prediction offers insight into the activity, interactions, and functions of proteins as well as serving as an important initial step toward tertiary structure prediction. The polypeptide backbone of the local conformation proteins is referred to as the protein's secondary structure.

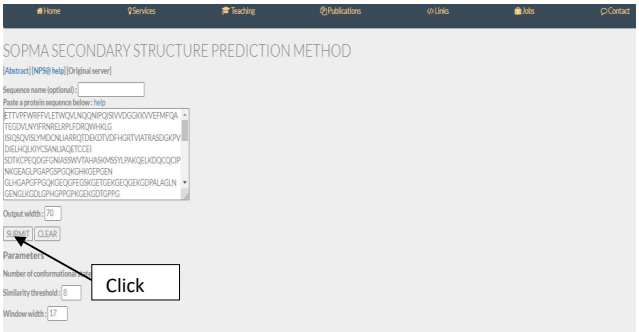


Figure 54: sequence pasted in box

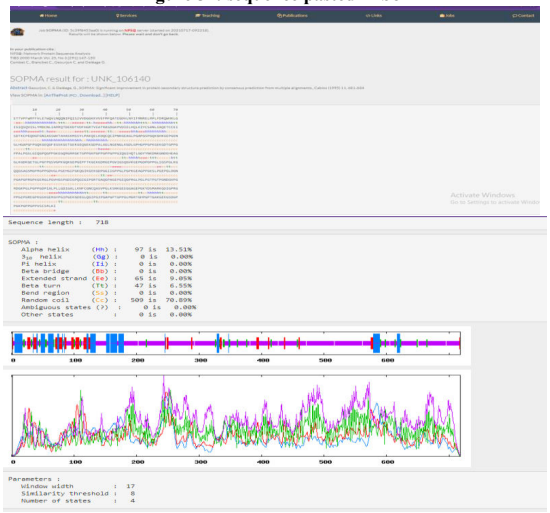


Figure 55: Result page of SOPMA

PROCEDURE

- Type SOPMA tool in the web browser search field and press
- enter
- Click the NPS@SOPMA secondary structure prediction (Figure 52)
- Open the SOPMA secondary structure prediction tool home page (Figure 53)
- Paste a protein sequence in the given box (Figure 54)
- Click submit (Figure 54)
- Result page (Figure 55) is displayed.

RESULT

- The secondary structure location present in the protein sequence is predicted.
- The secondary structures present in the given sequence are alpha helix, extended strand, beta turn and Random coil.
- Random coil is the most prominent secondary structure (70.89%) in the given sequence.
- The lowest percentage of secondary structure in the given sequence is Beta turn.

OUTCOME

Students learn how to predict the regions of different forms of secondary structure from the protein sequence.

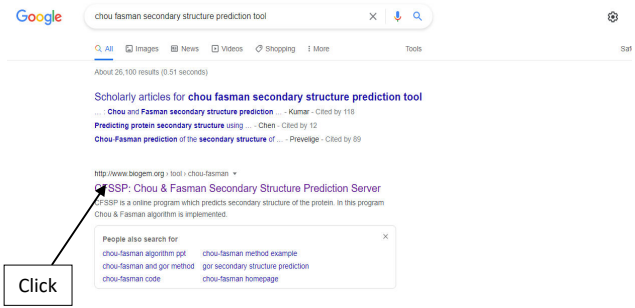


Figure 56: Search list of Secondary structure prediction tool

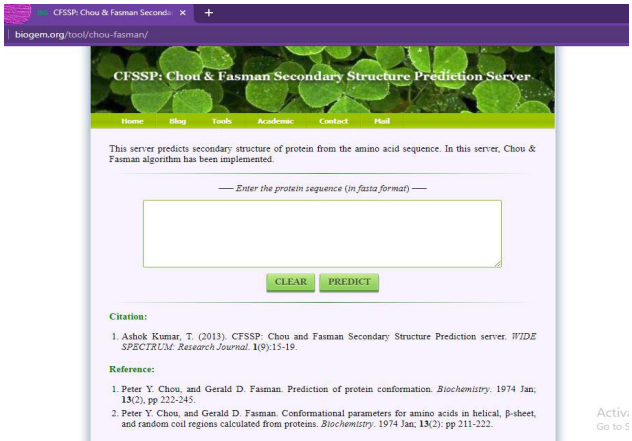


Figure 57: CFSSP Home Page

PREDICTING SECONDARY STRUCTURE OF PROTEIN SEQUENCE USING CFSSP TOOL

AIM

To predict the secondary structure of the given protein sequence through CFSSP tool.

INTRODUCTION

CFSSP (**Chou and Fasman Secondary Structure Prediction** Server) is an online protein secondary structure prediction server. The output predicts regions of secondary structure from the protein sequence such as alpha helix, beta sheet, and turns from the amino acid sequence. The method implemented in CFSSP is Chou and Fasman algorithm, which is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. CFSSP is freely accessible via ExPASy server or directly from BioGem tools at http://www.biogem.org/tool/chou-fasman.

PROCEDURE

- In web browser, type Chou Fasman secondary structure prediction tool, search list will be displayed and click CFFSSP (Figure 56)
- Click CFSSP or http://www.biogem.org › tool › chou-fasman

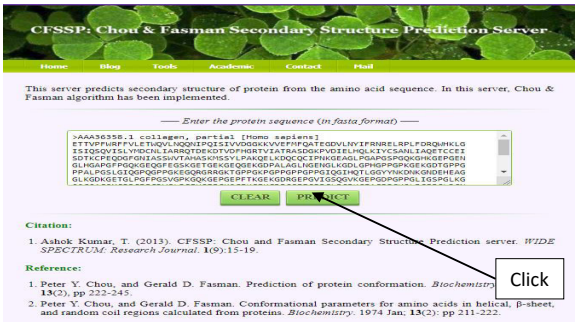


Figure 58: Sequence pasted in the box

Figure 59: Result page

- Open the CFSSP secondary structure prediction tool home page (Figure 57)
- Paste a protein sequence in the given box (Figure 58)
- Click predict (Figure 58)
- Result page (Figure 59) is displayed.

RESULT

- The secondary structures present in the given sequence are alpha helix, beta sheet and betas turn.
- Alpha helix is the most prominent secondary structure (32.9%) in the given sequence.
- The lowest percentage of secondary structure (16.7) in the given sequence is β turn.

OUTCOME

Students understand how to predict the regions of different forms of secondary structure from the protein sequence.

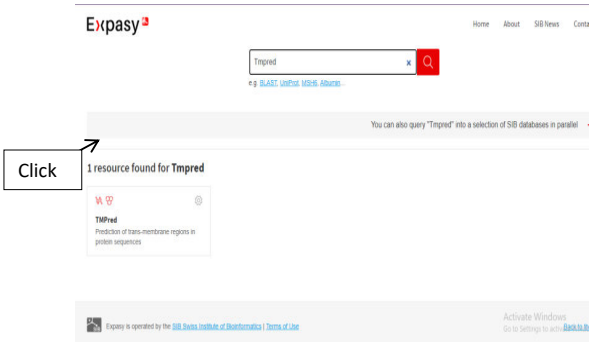


Figure 60: ExPASy home page

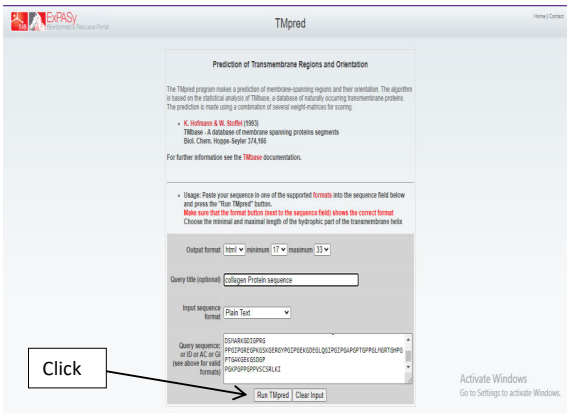


Figure 61: Home page of Mpred

PREDICTING TRANSMEMBRANE REGION OF PROTEIN SEQUENCE

AIM

To predict the transmembrane region in the given protein sequence/ID.

INTRODUCTION

The orientation of membrane-spanning sections is predicted by the TMpred software. The technique is based on a statistical investigation of the transmembrane protein database known as TMbase. A combination of various weight matrices for scoring are used to make the prediction.

PROCEDURE

- Go to ExPASy page, enter TMpred in search box (Figure 60)
- Click TMpred
- Go to the home page of TMpred tool (Figure 61)
- Paste the protein sequence in the search field
- Click Run TMpred (Figure 61)
- Results page is displayed (Figure 62)

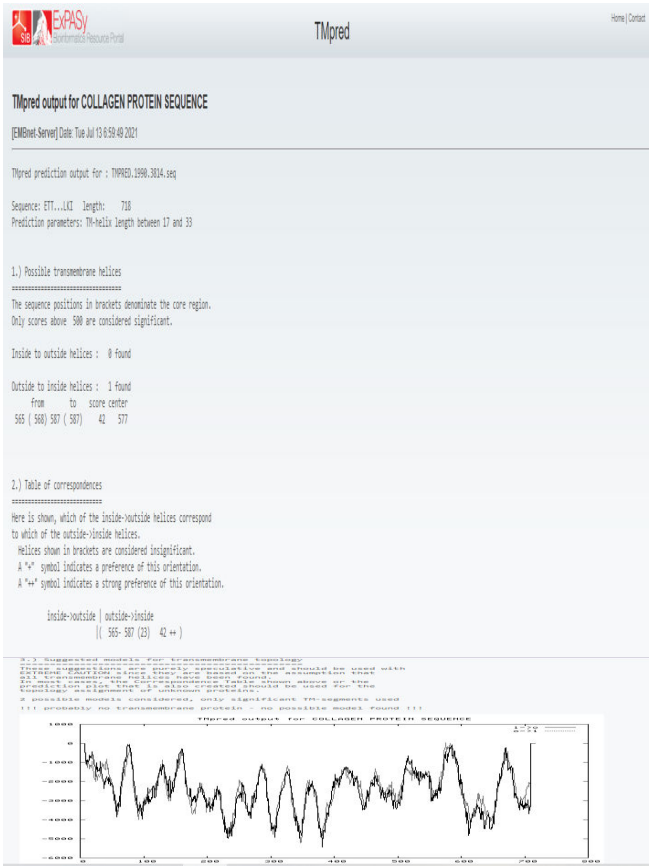


Figure 62: Result page of TMpred

RESULT

- Outside to inside helices : **1 found**

  from to score center

  565 (568) 587 (587) 42 577

- outside->inside

  "++" symbol indicates a strong preference of orientation

  (565- 587 (23) 42 ++)

OUTCOME

Students comprehend the technique to predict the membrane-spanning regions and their orientation. They also understand that the transmembrane proteins act as gateways for transporting specific substances across the membrane.

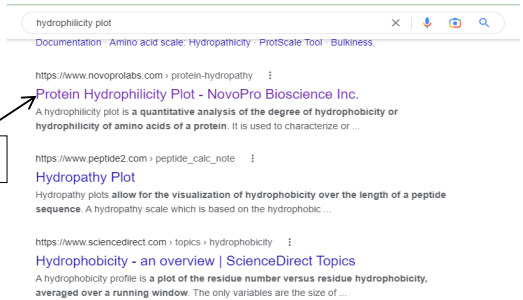


Figure 63: Search list of Hydrophilicity plot

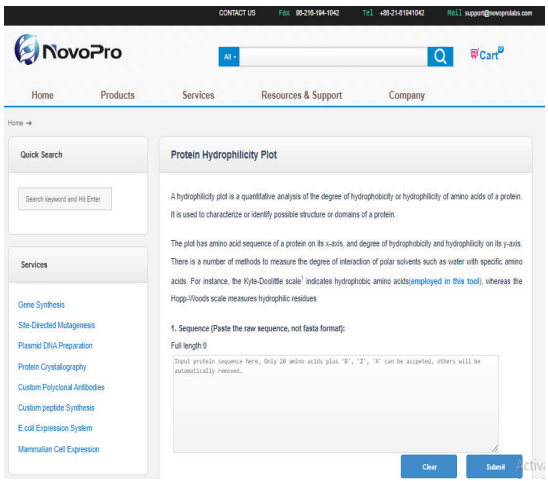


Figure 64: Hydrophilicity home page

PREDICTING HYDROPHILICITYREGION IN THE PROTEIN SEQUENCE

AIM

To predict the hydrophilic region in the given protein sequence

INTRODUCTION

The hydrophobicity or hydrophilicity of the amino acids in a protein is statistically analysed using a hydrophilicity plot. It is used to analyse or identify a protein's potential structure or domains. If the protein fragment is sufficiently hydrophobic to interact with or remain in a membrane can be predicted from the plot. The plot's x-axis represents the amino acid sequence of a protein, while the y-axis represents the degree of hydrophobicity and hydrophilicity. The degree of interaction between certain amino acids and polar solvents like water can be determined using a variety of techniques. For instance, the Hopp-Woods scale assesses hydrophilic residues while the Kyte-Doolittle scale identifies hydrophobic amino acids. Understanding the plot's shape reveals the details of the protein's partial structure. For instance, if a group of roughly 20 amino acids exhibits positive hydrophobicity, then it is possible that these amino acids are a portion of an alpha-helix that spans a lipid bilayer made up of hydrophobic fatty acids.

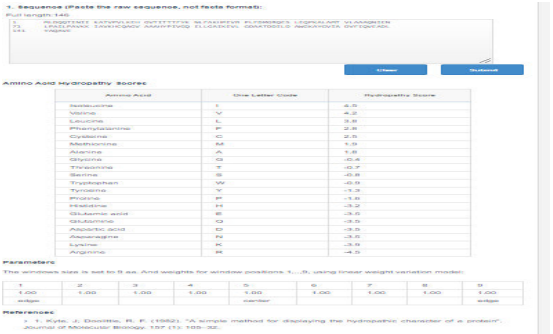


Figure 65: Home page

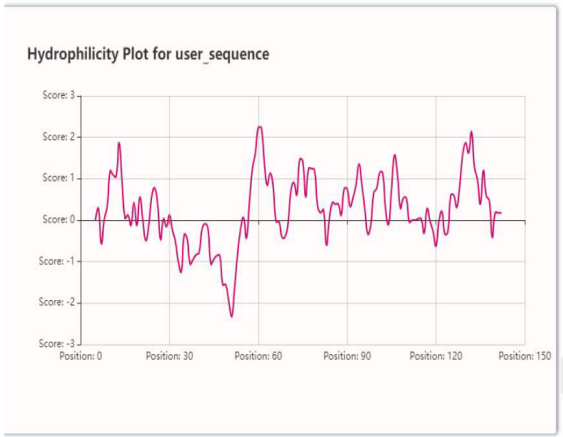


Figure 66: Hydrophilic regions in the protein sequence

On the other hand, amino acids with high hydrophilicity show that these residues are in touch with a solvent, such as water, and are therefore likely to be found on the protein's outer surface. First, a hydrophobicity rating between 4.6 and -4.6 is provided to each amino acid. The highest hydrophobic value is 4.6, and the least hydrophilic score is -4.6.

PROCEDURE

- Type hydrophilicity plot in the web browser and press enter
- Search list will be displayed, Click protein hydrophilic city plot –novoprolabs (Figure 63)
- Go to the home page (Figure 64)
- Paste the protein sequence in the given field
- Click submit (Figure 65)
- Results will be displayed (Figure 66)

RESULT

The hydrophilic residues in given Protein sequence was predicted.

OUTCOME

Students learn to predict whether or not the protein segment has enough hydrophilicity to either interact with or reside in a membrane.

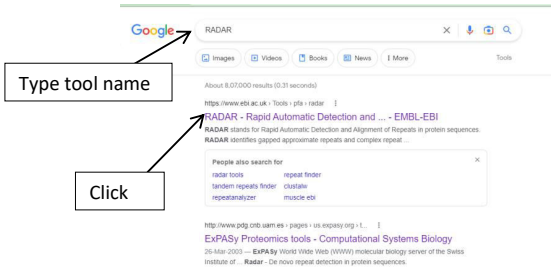


Figure 67: Google search page

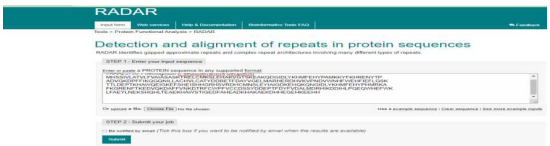


Figure 68: RADAR tool home page

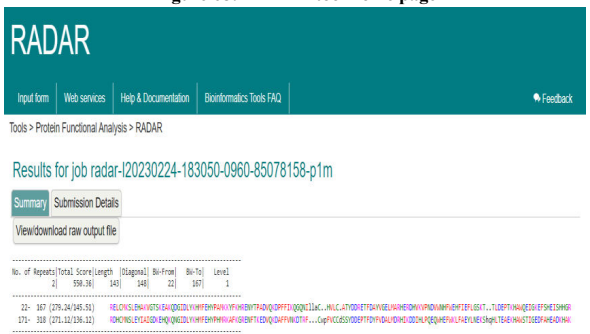


Figure 69: Result Page

DETECTING ALIGNMENT OF REPEATS IN A PROTEIN SEQUENCES

AIM

To detect the alignment of repeats in a given protein sequence.

INTRODUCTION

RADAR stands for Rapid Automatic Detection and Alignment of Repeats in protein sequences. RADAR identifies gapped approximate repeats and complex repeat architectures involving different types of repeats.

PROCEDURE

- Type Radar protein tool in the web browser and press enter
- Click the RADAR tool (ebi.uk)in the search list (Figure 67)
- Go to the RADAR home page (Figure 68)
- Paste the fasta protein sequence in the input box
- Click submit
- Result page was shown (Figure 69)

RESULT

The given sequence contains **2**alignment repeats.

OUTCOME

Students can able to predict the number of alignment repeats in the proteins sequences.

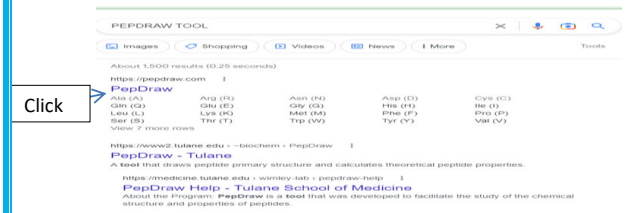


Figure 70: Google search list

Figure 71: Pepdraw home page

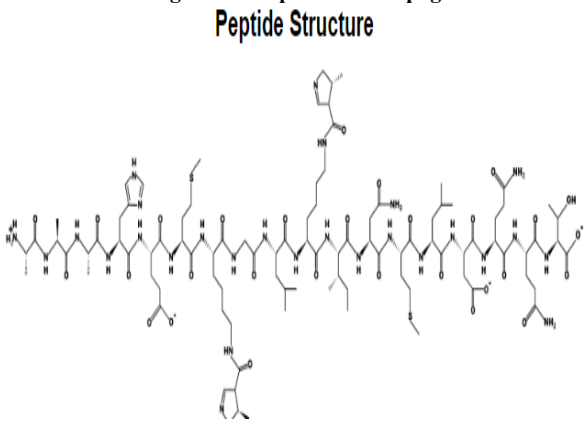


Figure 72: Result page

PREDICTING THE PEPTIDE STRUCTURE FOR THE GIVEN PROTEIN SEQUENCE

AIM

To predict the peptide structure of given protein sequence using Pepdraw tool

INTRODUCTION

Pepdraw tool is used to draw primary peptide and also calculate the physico chemical properties.

PROCEDURE

- Type Pepdraw tool in the web browser and press enter
- Click the Pepdraw in the search list (Figure 70)
- Go to the Pepdraw home page (Figure 71)
- Paste the fasta protein sequence in the input box
- Click draw peptide
- Result page will be displayed (Figure 72)

RESULT

The protein sequence was converted into peptide chain.

OUTCOME

Students acquire the skill to draw the primary chemical structure of an amino acid sequence and to predict chemical properties for any protein sequences.

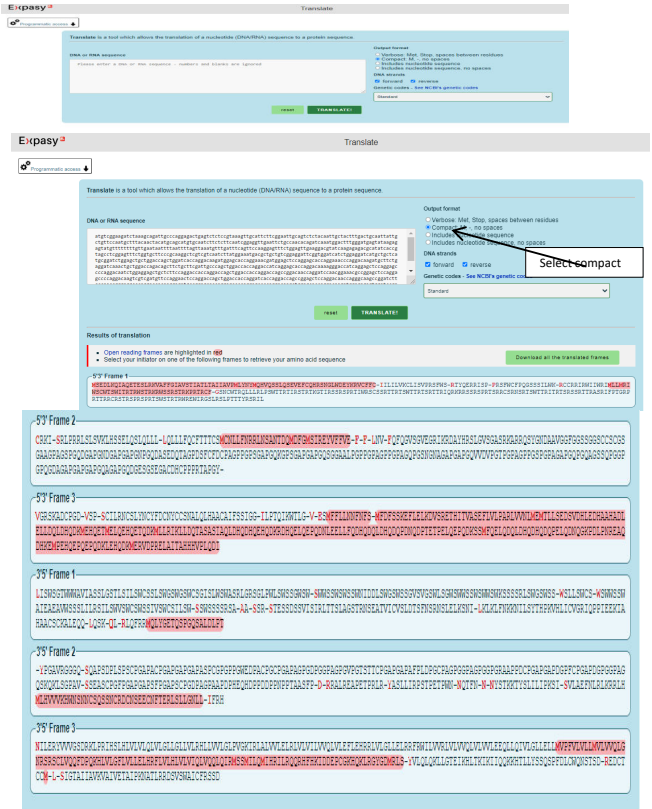


Figure 73: translate Homepage

CONVERSION OF NUCLEOTIDE SEQUENCES INTO PROTEIN SEQUENCES

AIM

To convert the nucleotide sequences into protein sequences and to identify the correct reading frame.

INTRODUCTION

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence into a protein sequence. Translate accepts a DNA sequence and converts it into a protein in the reading frame as specified. Translate supports the entire IUPAC alphabet and several genetic codes. A raw sequence or one or more FASTA sequences is pasted in the text area. Input limit is 200,000,000 characters. Determining is a complex process if a nucleic acid sequence actually codes for a protein. Because, generally it is not known which strand is the coding strand or which is the correct reading frame. Both these questions are resolved by translating both strands in all three reading frames and looking for the one that gives the longest amino acid sequence before a stop codon is encountered. A stop codon is expected to appear on average once for every 20 amino acids when reading a sequence in the incorrect frame. It is possible for an out of frame translation to extend over 100 amino acids before a stop codon is reached.

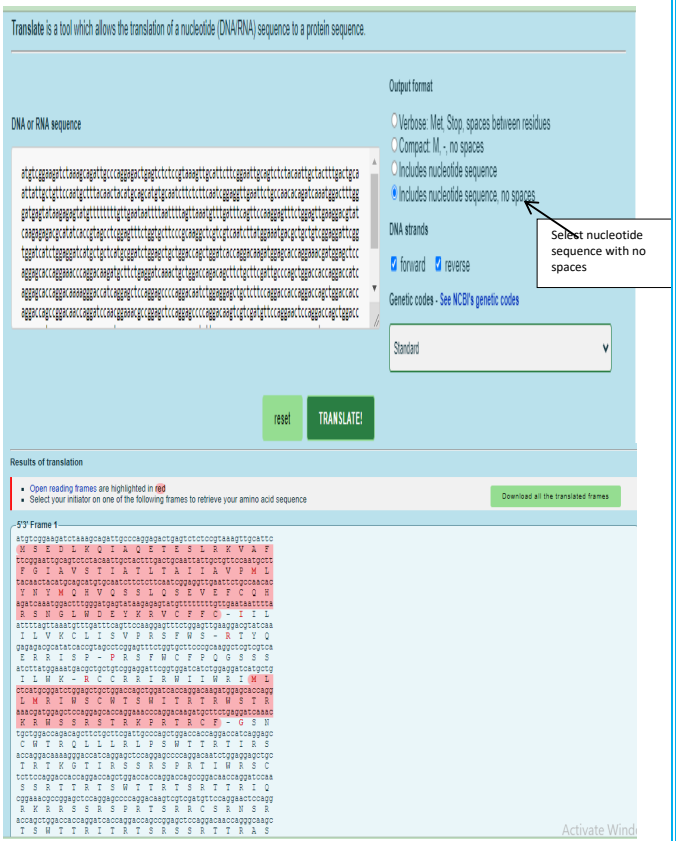


Figure 74: Result page of translate with nucleotide sequence with aminoacid

PROCEDURE

- Select the nucleotide sequence, copy it and then paste it into the translate sequence window in the ExPASy translate tool
- Under Output format, select "Compact" or nucleotide sequence without space
- Click on Translate Sequence
- Result page is displayed (Figure 73 and 74).

RESULT

- Output compact is selected; it gives the amino acid sequence as one letter code with stop codons indicated by a hyphen with different frames.
- Output nucleotide sequence without space is selected; it gives the nucleotide sequence with one letter code aminoacid with different frames.
- Red colour indicates the open reading frame.

OUTCOME

Students gain knowledge to convert any unknown nucleotide sequence into protein sequence and to identify the most correct reading frame. They also acquire knowledge about the exons, pseudogenes, non-coding region of DNA and regulatory functions.

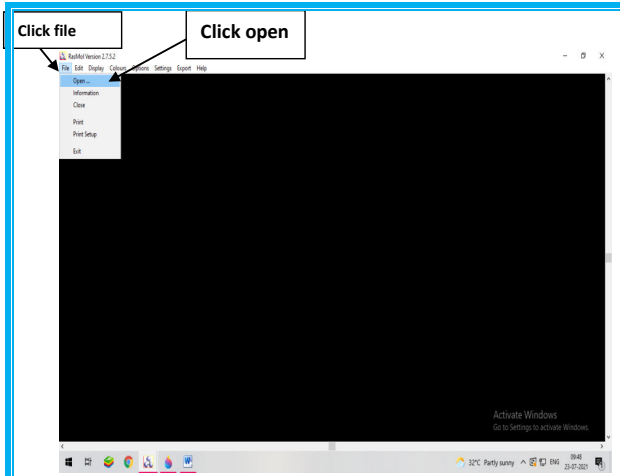


Figure 75 : Open the PDB file in RASMOL tool

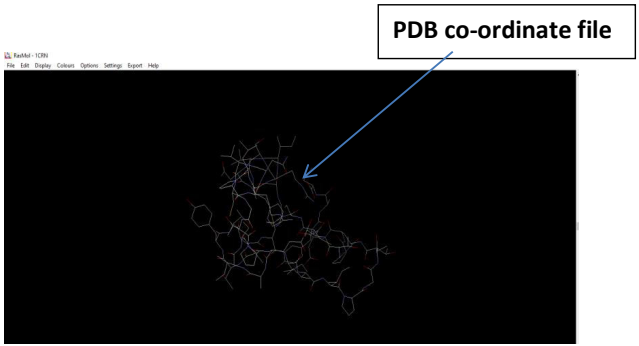


Figure 76: Protein structure open in molecular visualization tab

MOLECULAR VISUALIZATION USING RASMOL

AIM

To visualize the tertiary structure of protein molecule in graphic view and command line

INTRODUCTION

RasMol is free software for molecular visualization created by Roger Sayle. It is a molecular graphics programme intended for the visualization of proteins, nucleic acids and small molecules. The programme aims at display, teaching and generation of publication with quality images. The program reads in a molecule coordinate file and interactively displays the molecular screen in variety of colour schemes and molecular representations.

REQUIREMENT

RASMOL software, PDB molecule

PROCEDURE

- Open a molecular visualization tool (Figure 75)
- From the file menu open a PDB atom co-ordinate file (Figure 76)
- Rotate the molecule

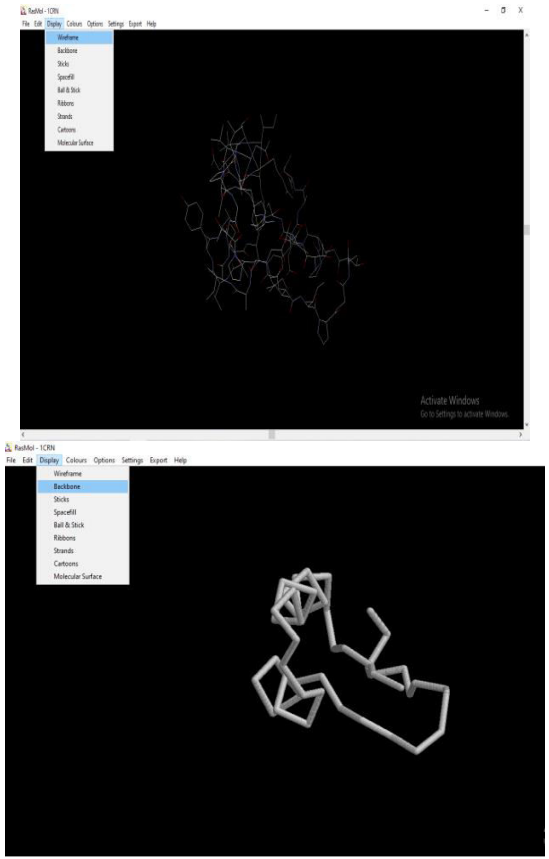


Figure 77:Visualise the different forms of protein structure

- Try various options (Figure 77,78,79)
- Try different commands in command line and visualize the changes in structure (Figure 80, 81)
- Save the required structural view
- Exit the application

COMMANDS

- Select
- Colour
- Zoom on
- Zoom off
- Label on
- Label off
- Spacefill
- Star on
- Background pink
- Stereo on
- Stereo off
- Pick angle
- Pick distance
- Label 250
- Star
- Rotate

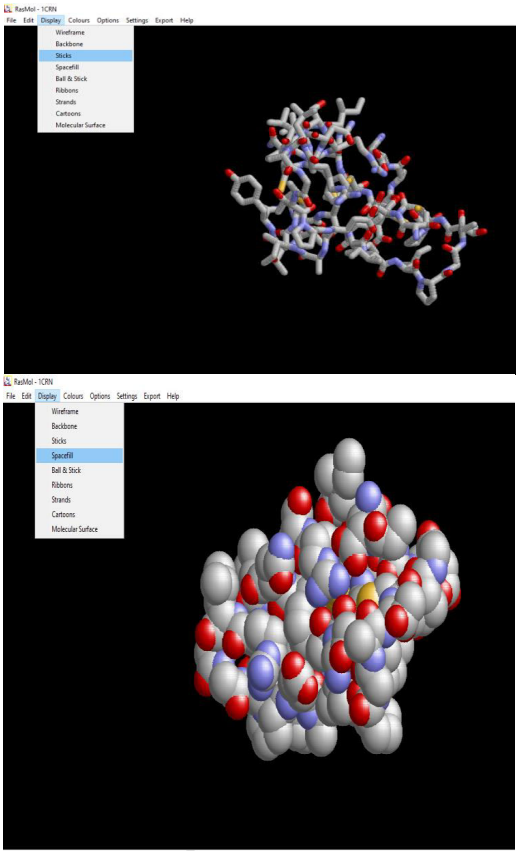


Figure 78: Visualise the different forms of protein structure

- Hbonds
- Wireframe
- Cartoon
- dots
- Quit

Other command reference:

Backbone	Background	Bond	Bulgarian	Cartoon	Centre	Chinese	Clipboard
Colour	ColourMode	Connect	CPK	CPKnew	Defer	Define	Depth
Dots	Echo	English	Execute	Exit	French	HBonds	Help
Italian	Japanese	Label	Load	Map	Molecule	Monitor	NoToggle
Pause	Play	Print	Quit	Record	Refresh	Renumber	Reset
Restrict	Ribbons	Rotate	Save	Script	Select	Set	Show
Slab	Source	Spacefill	Spanish	SSBonds	Star	Stereo	Strands
Structure	Surface	Trace	Translate	UnBond	Wireframe	Write	Zap
Zoom							

RESULT

The tertiary structure of protein molecule is visualized in graphic view and command line

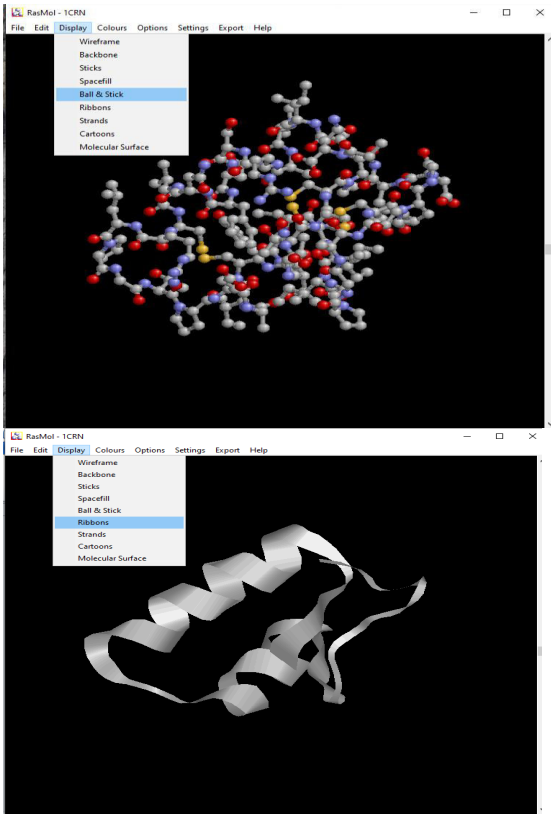


Figure 79: Visualise the different forms of protein structure

OUTCOME

The students are imparted the ability to evaluate and interpret molecular models. Students can interpret the complicated molecule structure, properties, and interactions with the use of molecular visualisation tools. These resources aid their study in the fields of chemistry, pharmacology, biology, and bioinformatics.

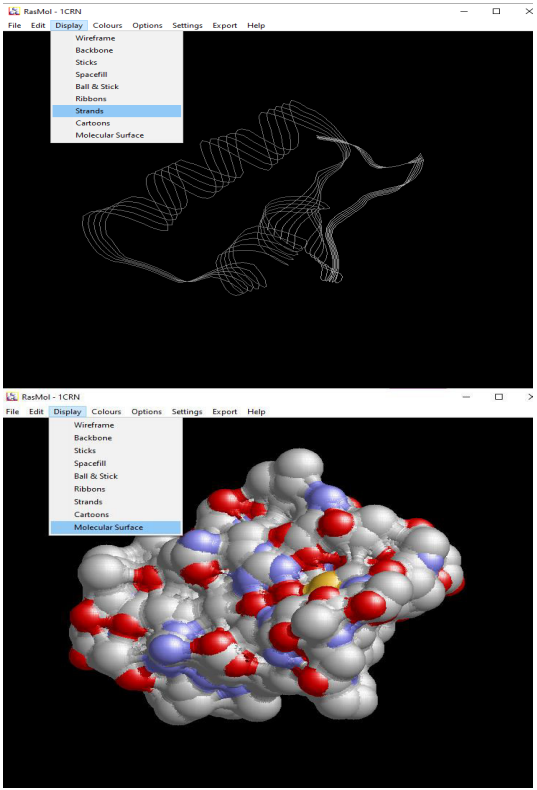


Figure 80: Visualise the different forms of protein structure

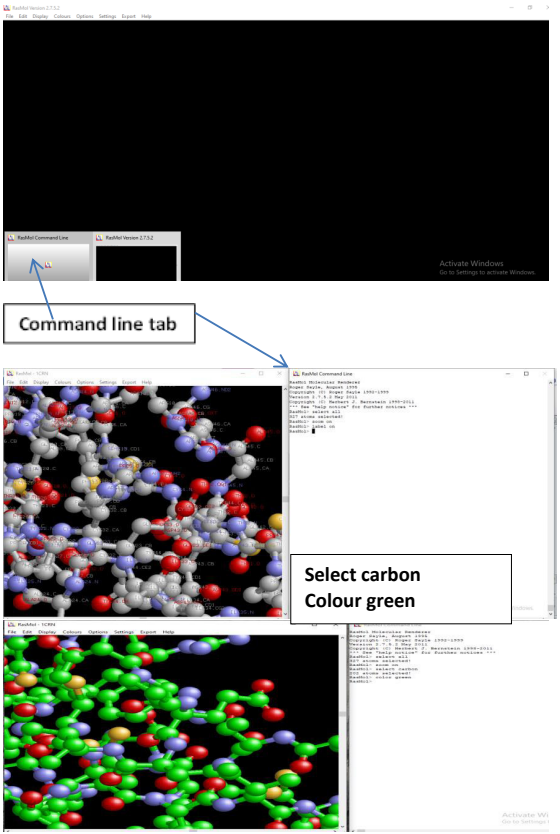


Figure 81: Various commands used in command line and visualize the changes in structure

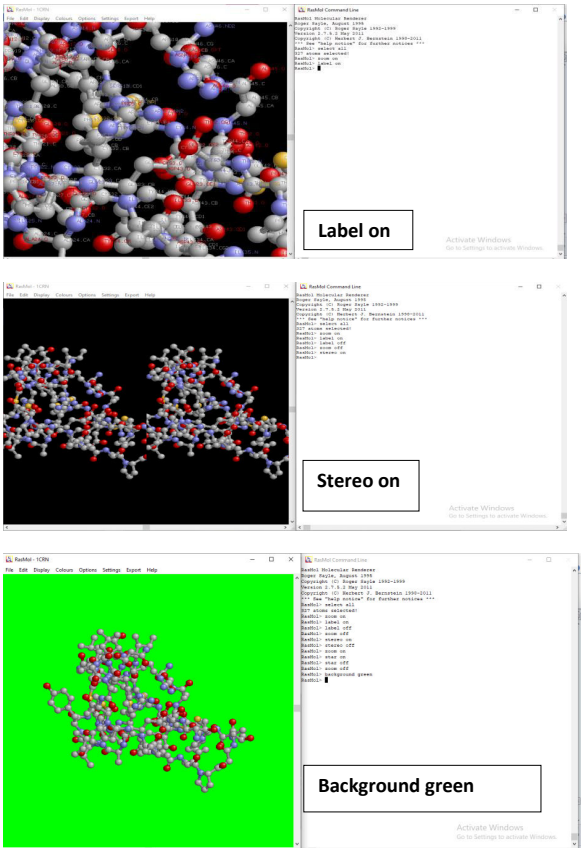


Figure 82: Various commands used in command line and visualize the changes in structure

MEASUREMENT OF BOND LENGTH IN PROTEIN STRUCTURE USING RASMOL

AIM

To measure the bond length of atoms in the protein structure and to visualize it in graphic view and command line.

INTRODUCTION

RasMol is a free software for molecular visualization created by Roger Sayle. It is a molecular graphics programme intended for the visualization of proteins, nucleic acids and small molecules. The program aims at display, teaching and generation of publication with quality images. The program reads in a molecule coordinate file and interactively displays the molecular screen in variety of colour schemes and molecular representations.

PROCEDURE

- Open RasMol and import a file of Pdb atom coordinates
- Use the various menu options and get a composition of the molecule
- Set the display style to “ball and stick” (Preferable, but works with other displaying style as well)
- Use Shift+ mouse down to zoom in on the molecule to see the bonds more clearly.
- Open the command line window
- Type **set picking distance** and press enter key

- Open the display Window and select the two atoms participating in the bond formation by clicking on them successively.
- The command line window displays the bond length.
- Record the results
- Alternatively, to show bond and to measure the bond length between two atoms, type set picking monitor in the command line window.
- Now click on the two atoms again.
- A bond line appears (The bond is removed when a click on the atoms is made more than once.)
- Note the results from the command line window.

RESULT

The distance between two atoms is 90.0Å.

OUTCOME

Students learn how to find out the distance between two atoms using command line and directly picking atoms in structure.

MEASUREMENT OF BOND ANGLE IN PROTEIN STRUCTURE USING RASMOL

AIM

To measure the bond angle between the atoms in the protein structure and to visualize it in graphic view and command line.

INTRODUCTION

RasMol is a free and most popularly used software for molecular visualization created by Roger Sayle. It is a molecular graphics programme intended for the visualization of proteins, nucleic acids and small molecules. The program aims at display, teaching and generation of publication with quality images. The program reads in a molecule coordinate file and interactively displays the molecular screen in variety of colour schemes and molecular representations.

PROCEDURE

- Open RasMol and load a file of Pdb atom coordinates
- Use the various menu options and get a feel of the molecule
- Set the display style to “ball and stick”(Preferable, but works with other displaying style as well)
- Use Shift + mouse down to zoom in on the molecule to see the bonds more clearly.
- Go to command line window.
- Type **set picking angle** and press enter key

- Go to display Window and select the three atoms forming the bond angle by clicking on them successively.
- The command line window displays the bond angle.
- Note the results.

RESULT

The bond angle between atoms is 25.0°

OUTCOME

Students acquire the skills necessary to find out the bond angle between atoms using command line and directly picking atoms in structure.

WEBLINKS:

- https://www.careerindia.com/courses/unique- courses/what-is-bioinformatics-scope- career-opportunities-012034.html
- https://www.ncbi.nlm.nih.gov/
- https://www.ebi.ac.uk/Tools/sss/ncbiblast/
- https://en.wikipedia.org/wiki/Clustal
- https://www.ebi.ac.uk/Tools/msa/clustalo/
- https://www.bing.com/ck/a?!
- https://web.ExPASy.org/
- https://web.ExPASy.org/ProtParam/
- https://web.ExPASy.org/peptide_mass/
- https://web.ExPASy.org/peptide_cutter/
- https://www.academia.edu/3112992/CFSSP_Chou_and_Fasman_Secondary_Structure_Prediction_Server/
- http://www.openrasmol.org/
- https://www.novoprolabs.com/tools/protein-hydropathy
- https://www.ebi.ac.uk/Tools/pfa/radar/
- https://pepdraw.com/
- https://ase.tufts.edu/biology/bioinformatics/exercise3.asp

About the Author

Sudha Rameshwari is a Biochemist and her native is Kingmaker Kamaraj birth town Virudhunagar, Tamilnadu, India. She did her B.Sc. and M.Sc. (Biochemistry) from V.V.Vanniaperumal College for Women, Virudhunagar, Tamilnadu, India. She obtained her M.Phil degree in Life Science (2003) from Manonmaniam Sundaranar University, Tirunelveli. She has completed her Post graduate diploma in Bioinformatics (2005) in Bharathiyar University, Coimbatore. She has completed 24 years' of teaching experience in Biochemistry and her Research interests include Microbiology, Pharmacology, Green nanotechnology and Bioinformatics. She submitted one sequence to GEN BANK on April 2018. She regularly teaches Techniques, Enzymology, Clinical biochemistry, Microbial biochemistry and Bioinformatics. She is well trained in Microbiology and Molecular Biology Techniques. She has published more than 30 research articles in well reputed International journals which are Scopus indexed (8), Web of Science (4) and UGC approved. She is interested in participating and to organize workshops. She guided 2 M.Phil students and 25 PG students. She has filed and published one patent (Indian) Publication. She is a reviewer in 15 reputed International journals. She also published 3 chapters in edited books. She received grants from TNSCST-DBT, autonomy fund and MRP grant from VVVCMB-MRP Scheme sponsored by VVVCollege Manage Board. She has also been awarded as Honorary Doctorate (D.Litt.(hc)) from University of Central America, CV Raman Prize 2022 from Institute of Researchers, Wayanad, Kerala and Bharat Excellence award 2022, Leading educationist of India Award 2022 from Friendship forum , New Delhi, Global Personalities of Asia 2022 from Global brotherhood forum, New Delhi and Outstanding Researcher in Microbial Biochemistry from VIHA, 2018. She is a life member (LM052202) in Institute of Researchers, Waynad and also SAS eminent Fellow Membership (SAS/SEFM/077/2021) in Scholars Academic and Scientific society. She has scored 8.7/10 in VIDWAN expert database and National Researchers network.