Semester – III

18BTP312 GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI 4H – 2C

Instruction Hours / week: L: 0 T: 0 P: 4 **Marks:** Internal: **40** External: **60** Total: **100**
End Semester Exam: 3 Hours

1. Exploring of primary databases (Proteins and Nucleic acids) and sequence retrieval
2. Physicochemical and structural analyses of primary sequences (Proteins and Nucleic acids)
3. Sequence similarity searches and pairwise alignments
4. Multiple sequence alignments and phylogenetic analysis
5. Comparative modeling using online and standalone tools
6. Molecular visualization tools: RasMol, JMol and PyMol
7. Structural analysis and verification tools
8. Molecular dockings of biological macromolecules

Suggested Readings

1. Baxevanis, A.D. & Ouellette, B.F. (2001). *Bioinformatics – A practical guide to the analyze of genes and proteins* (2nd ed.). Wiley-Blackwell Publishers, New York, United States.
2. Leach, A.R. & Gillet, V.J. (2009). *An Introduction to Chemoinformatics*. Springer Publishers, New York, United States.
3. Ibrahim, K.S., Gurusubramanian, G., Zothansanga, Yadav, R.P., Kumar, N.S., Pandian, S.K., Borah, P., & Mohan, S. (2017). *Bioinformatics - A Student's Companion*. Springer Publishers, New York, United States.

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

1. Exploring of primary databases (Proteins and Nucleic acids) and sequence retrieval

Aim: To browse various primary databases and to retrieve sequences of proteins and nucleic acids from their corresponding databases

Procedure

The **European Molecular Biology Laboratory (EMBL)** is a molecular biology research institution supported by 22 member states, four prospect and two associate member states. EMBL was created in 1974 and is an intergovernmental organization funded by public research money from its member states. Research at EMBL is conducted by approximately 85 independent groups covering the spectrum of molecular biology. The Laboratory operates from five sites: the main laboratory in Heidelberg, and outstations in Hinxton (the European Bioinformatics Institute (EBI), in England), Grenoble (France), Hamburg (Germany), Monterotondo (near Rome) and Barcelona (Spain). EMBL groups and laboratories perform basic research in molecular biology and molecular medicine as well as training for scientists, students and visitors. The organization aids in the development of services, new instruments and methods, and technology in its member states. Israel is the only Asian state that has full membership. In March 2010, the EMBL Advanced Training Centre (ATC) was inaugurated on the main campus in Heidelberg. Shaped in the form of a double helix, it hosts conferences and provides training. EMBL also runs an active Science and Society Programme which offers activities and events on current questions in life science research for the general public and the scientific community.

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).

Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see BLAST info for more information about the numerous BLAST databases.

Search, link, and download sequences programatically using NCBI e-utilities.

The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank.

UniGene is an NCBI database of the transcriptome and thus, despite the name, not primarily a database for genes. Each entry is a set of transcripts that appear to stem from the same transcription locus (i.e. gene or expressed pseudogene). Information on protein similarities, gene expression, cDNA clones, and genomic location is included with each entry.

The Protein Information Resource (PIR) is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies. The PIR continues to offer world leading resources to assist with proteomic and genomic data integration and the propagation and standardization of protein annotation.

NCBI DATABASE

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services.

Gene database has been implemented at NCBI to characterize and organize the information about genes. It serves as a major node in the nexus of genomic map, expression, sequence, protein function, structure and homology data. A unique GeneID is assigned to each gene record that can be followed through revision cycles. Protein database is an important protein resource at NCBI. It maintains the text record for individual protein sequences, derived from many different resources such as NCBI Reference Sequence (RefSeq) project, GenbBank,

PDB and UniProtKB/SWISS-Prot. Protein records are present in different formats including FASTA and XML and are linked to other NCBI resources. Protein provides the relevant data to the users such as genes, DNA/RNA sequences, biological pathways, expression and variation data and literature

- The amino acid sequence of proteins, nucleic acids and SNP would be downloaded in the FASTA format from the website **https://www.ncbi.nlm.nih.gov/**
- Once the FASTA file is downloaded, it can be opened in a word pad file and stored in txt format.

SWISS - PROT DATABASE

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recent developments of the database include format and content enhancements, cross-references to additional databases, new documentation files and improvements to TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL Nucleotide Sequence Database, except the CDSs already included in SWISS-PROT. SWISS-PROT is available at: http://www.expasy.ch/sprot/ and http://www.ebi.ac.uk/swissprot/

FASTA FORMAT

FASTA format is a text based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single letter codes. FASTA format is more useful for BLAST searches and alignments of sequences. First line begins with '>' symbol. No separate designations for protein and nucleic acid sequences. The code is entered followed by the comments on the same line. Delimit the comments with a '|' symbol.

Example:

>gi|20072668|gb|AAH27258.1|

MAHAGRTGYDNREIVMKYIHYKLSQRGYEWDAGDVGAAPPGAAPAPGIFSSQPGHTPH
PAASRDPVARTSPLQTPAAPGAAAGPALSPVPPVVHLTLRQAGDDFSRRYRRDFAEMSS
QLHLTPFTARGRFATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSPLVDNIAL
WMTEYLNRHLHTWIQDNGGWDAFVELYGPSMRPLFDFSWLSLKTLLSLALVGACITLG
AYLGHK

For instance, retrieve amino acid and nucleotide sequences of the following proteins and write the accession IDs of the sequences.

Myoglobin from *Homo sapiens*

Lysozyme from *Homo sapiens*

Cardiotoxins from snake venom

Neurotoxins from snake venom

Conclusions:

2. Physicochemical and structural analyses of primary sequences (Proteins and Nucleic acids)

Aim: To analyze various physicochemical and structural properties of proteins and nucleic acids from their primary structures using an array of computational tools.

Procedure

Choose the following tools for the data analysis.

- Compute pI
- ProtParam
- ProtScale (hydrophobicity index)
- Peptide Cutter (Chymotrypsin, pepsin, trypsin, Formic acid, CNBr)
- CFSSP
- GOR
- APSSP

The Hydropathy values of the proteins can be calculated using various algorithms as shown below herein.

- Hphob. OMH / Sweet et al.
- Hphob. / Hopp & Woods
- Hphob. / Kyte & Doolittle
- Hphob. / Manavalan et al.
- Hphob. / Abraham & Leo

The parameter setting can be manipulated as per the following features.

ProtScale allows you to compute and represent the profile produced by any amino acid scale on a selected protein. An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist

which are based on different chemical and physical properties of the amino acids. This program provides 50 predefined scales entered from the literature. You can set several parameters that control the computation of a scale profile, such as the window size, the weight variation model, the window edge relative weight value, and scale normalization.

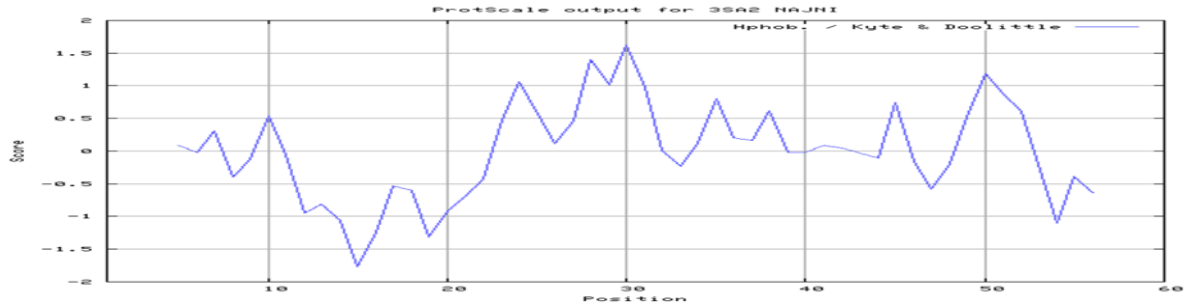
Window size

The window size is the length of the interval to use for the profile computation. When computing the score for a given residue i, the amino acids in an interval of the chosen length, centered around residue i, are considered. In other words, for a window size n, we use the i - (n-1)/2 neighboring residues on each side of residue i to compute the score for residue i. The score for residue i is the sum of the scale values for these amino acids, optionally weighted according to their position in the window.

Relative weight of the window edges

The central amino acid of the window always has a weight of 100%. By default, the amino acids at the remaining window positions have the same weight, but you can make the residue at the center of the window have a larger weight than the others by setting the weight value for the residues at the beginning and end of the interval to a value between 0 and 100%. The decrease in weight between the weight of the center and that of the edges will either be linear or exponential depending on the setting of the weight variation model option. Weight variation modelIn the following example, the window size is 7, and the window edge relative weight value is 10%. Linear weight variation model - This option divides the weight into equally spaced intervals between 100% and the window edge relative weight (here: 10%).

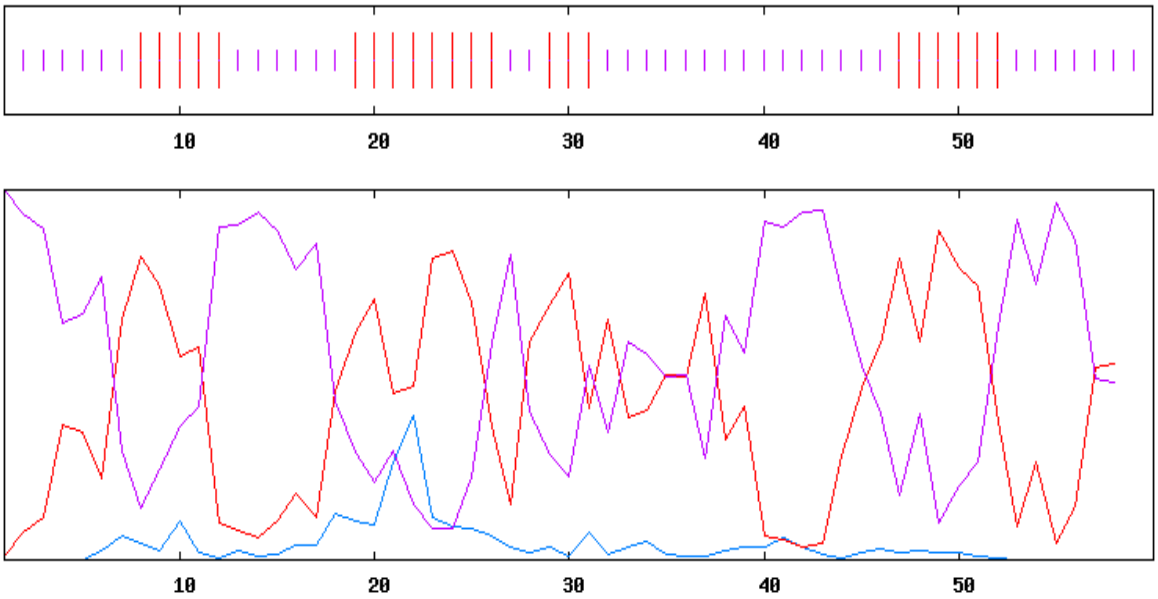
Typical output for a give protein sequence is as shown below.



CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

The secondary structures of the proteins would be estimated by means of various computational tools. Some of them are statistical-based tools and others are knowledge-based tools. Moreover, a few of them use more than one prediction methods and deliver consensus data. In addition, the tool such as AGADIR is a prediction algorithm based on the helix/coil transition theory. The Agadir predicts the helical behavior of monomeric peptides. It only considers short range interactions. Conditions such as pH, temperature and ionic strength are used in the calculation. Modifications of the termini are also allowed. In other words, the Agadir is not a program to predict secondary structure of proteins. Outputs of each of the tools mentioned above are unique and reliability of the tools are also different from each other. The output of GOR tool is shown below herein as reference.

```
      10        20        30        40        50        60
      |         |         |         |         |         |
LKCNKLVPLFYKTCPAGKNLCYKMFMVSNLTVPVKRGCIDVCPKNSALVKYVCCNTD
RCN
ccccccccceeeeeccccccceeeeeeeeccceeeccccccccccccccceeeeeeccccccc
Sequence length :  60
GOR4:
  Alpha helix     (Hh) :     0 is   0.00%
  3(10) helix     (Gg) :     0 is   0.00%
  Pi helix        (Ii) :     0 is   0.00%
  Beta bridge     (Bb) :     0 is   0.00%
  Extended strand (Ee) :    22 is  36.67%
  Beta turn       (Tt) :     0 is   0.00%
  Bend region     (Ss) :     0 is   0.00%
  Random coil     (Cc) :    38 is  63.33%
  Ambiguous states (?) :     0 is   0.00%
  Other states         :     0 is   0.00%
```



Detection of Open Reading Frames in the nucleotide sequences using ORF finder

Fetch the nucleotide sequence with accession ID from EMBL database and save the file in FASTA format.

The following gene – finding programs would be employed and the results would be comprehensively analyzed.

GENSCAN

Web server: http://genes.mit.edu/GENSCAN.html

GeneID

Web server: http://genome.crg.es/geneid.html

GRAIL

Web server: http://pbil.univ-lyon1.fr/members/duret/cours/insa2004/exercise4/pgrail.html

GrailEXP is a software package that predicts exons, genes, promoters, polyas, CpG islands, EST similarities, and repetitive elements within DNA sequence. GrailEXP is used by the Computational Biosciences Section at Oak Ridge National Laboratory to annotate the entire known portion of the human genome (including both finished and draft data).

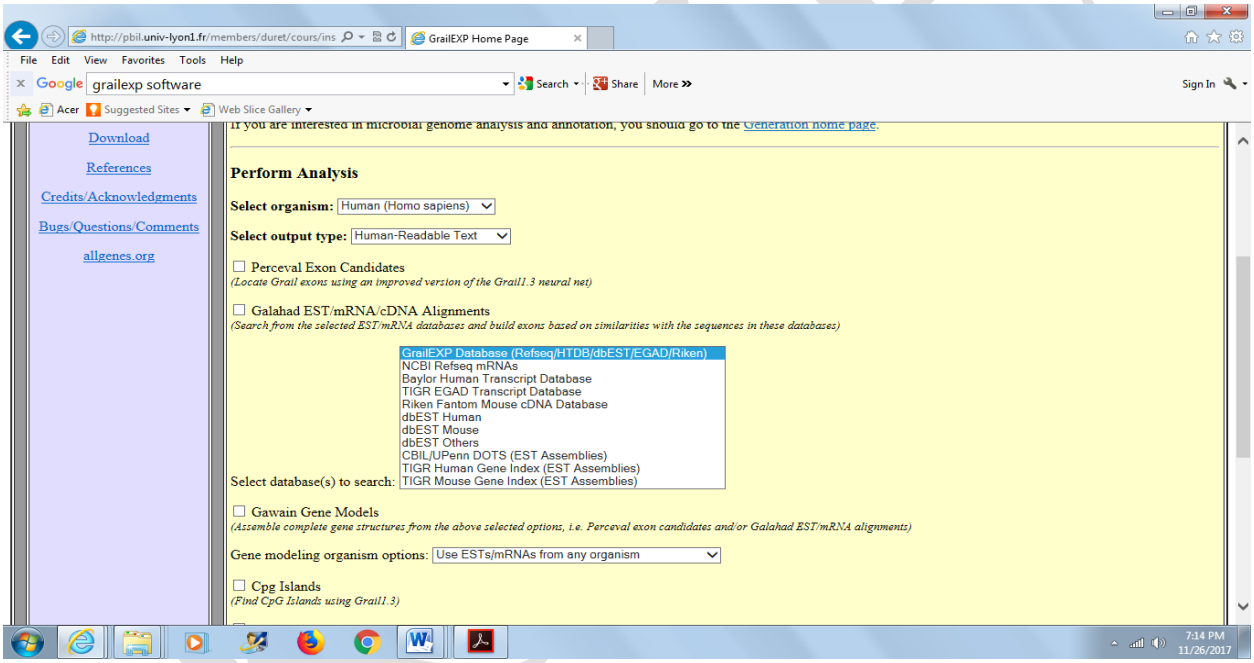
GeneMark

Web server: http://exon.gatech.edu/Genemark/genemarks.cgi

ORF Finder

Web server: http://www.bioinformatics.org/sms2/orf_find.html

ORF Finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF Finder to search newly sequenced DNA for potential protein encoding segments. ORF Finder supports the entire IUPAC alphabet and several genetic codes.



Conclusions:

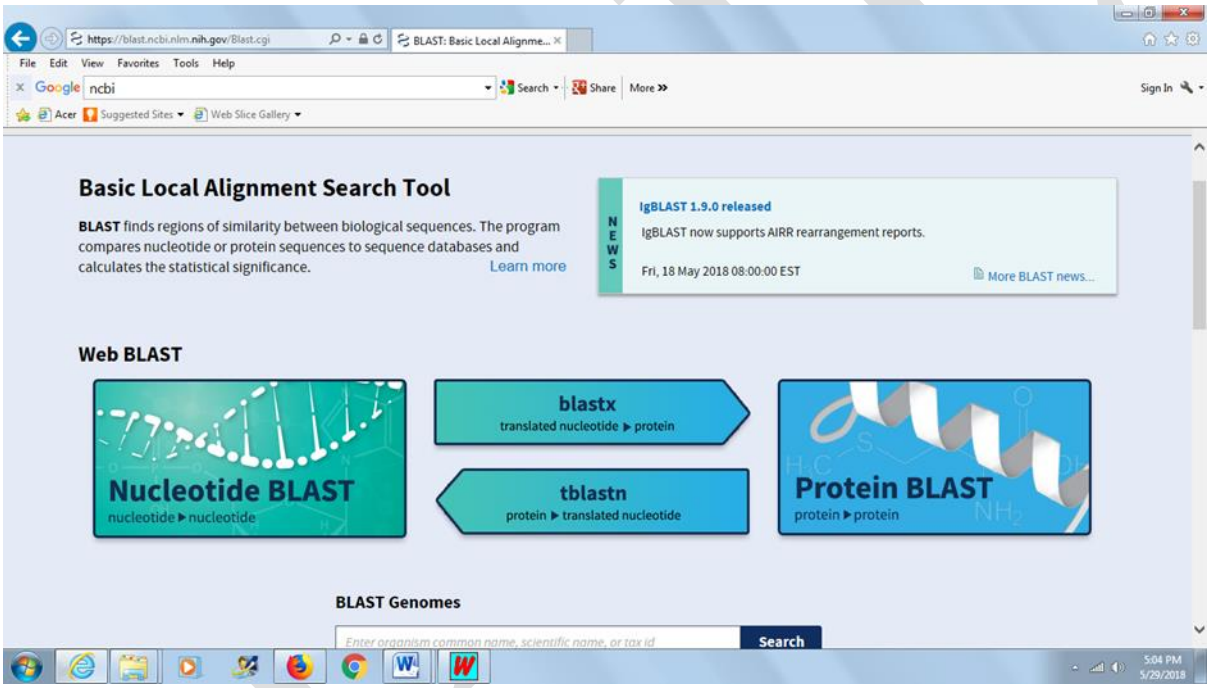
CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

3. Sequence similarity searches and pairwise alignments

Aim: To retrieve similar sequences for a given protein/nucleotide primary structure using BLAST and rationalizing the outcomes

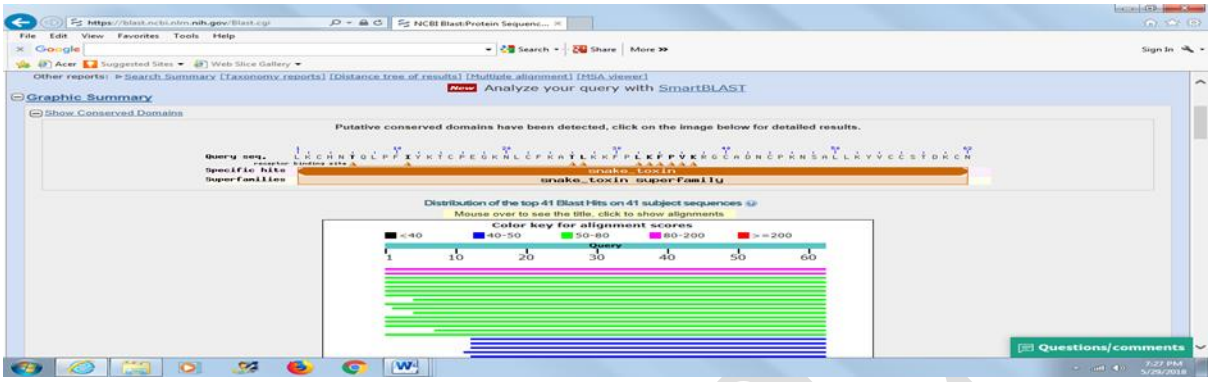
Procedure

Retrieve a few numbers of protein/nucleotide sequences from NCBI/UniProt databases (https://www.ncbi.nlm.nih.gov/) and save the retrieved sequences in FASTA format and accession IDs for all the sequences as well. The 'sequence similarity search' can be executed using various BLAST algorithms.

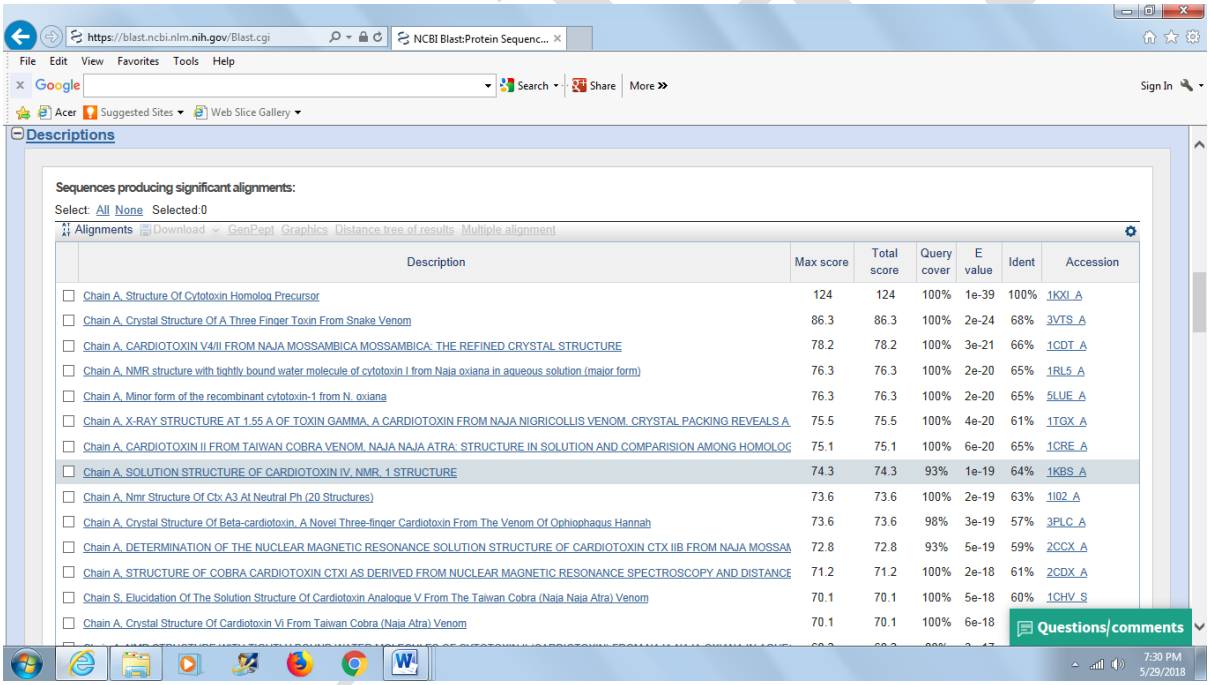


Using protein blast, similar sequences can be obtained either from primary or from structural databases. In general, for distant related and close related sequences, PSI-BLAST, BLASTP algorithms are preferred, respectively. For instance, a cardiotoxin (1CVO) was subjected to BLASTP similarity search and the outcomes were as shown herein below.

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312



The sequences would be further analyzed on the basis of query coverage, total score, percentage of identities and E-values. The score values and other statistical parameters for a few hits of the query sequences are depicted in the following illustration.

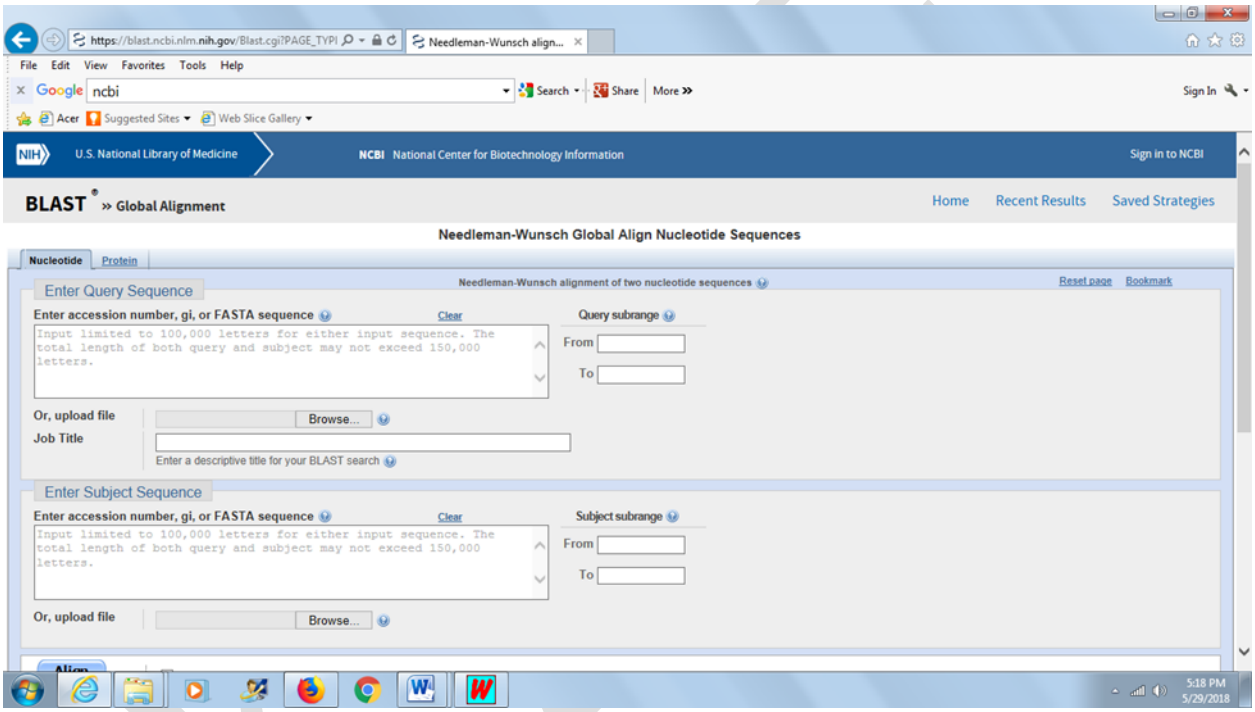


Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Structure Of Cytotoxin Homolog Precursor	124	124	100%	1e-39	100%	1KXI_A
Chain A, Crystal Structure Of A Three Finger Toxin From Snake Venom	86.3	86.3	100%	2e-24	68%	3VTS_A
Chain A, CARDIOTOXIN V4/II FROM NAJA MOSSAMBICA MOSSAMBICA: THE REFINED CRYSTAL STRUCTURE	78.2	78.2	100%	3e-21	66%	1CDT_A
Chain A, NMR structure with tightly bound water molecule of cytotoxin I from Naja oxiana in aqueous solution (major form)	76.3	76.3	100%	2e-20	65%	1RL5_A
Chain A, Minor form of the recombinant cytotoxin-1 from N. oxiana	76.3	76.3	100%	2e-20	65%	5LUE_A
Chain A, X-RAY STRUCTURE AT 1.55 A OF TOXIN GAMMA, A CARDIOTOXIN FROM NAJA NIGRICOLLIS VENOM. CRYSTAL PACKING REVEALS A	75.5	75.5	100%	4e-20	61%	1TGX_A
Chain A, CARDIOTOXIN II FROM TAIWAN COBRA VENOM, NAJA NAJA ATRA: STRUCTURE IN SOLUTION AND COMPARISION AMONG HOMOLOG	75.1	75.1	100%	6e-20	65%	1CRE_A
Chain A, SOLUTION STRUCTURE OF CARDIOTOXIN IV, NMR, 1 STRUCTURE	74.3	74.3	93%	1e-19	64%	1KBS_A
Chain A, Nmr Structure Of Ctx A3 At Neutral Ph (20 Structures)	73.6	73.6	100%	2e-19	63%	1II2_A
Chain A, Crystal Structure Of Beta-cardiotoxin, A Novel Three-finger Cardiotoxin From The Venom Of Ophiophagus Hannah	73.6	73.6	98%	3e-19	57%	3PLC_A
Chain A, DETERMINATION OF THE NUCLEAR MAGNETIC RESONANCE SOLUTION STRUCTURE OF CARDIOTOXIN CTX IIB FROM NAJA MOSSA	72.8	72.8	93%	5e-19	59%	2CCX_A
Chain A, STRUCTURE OF COBRA CARDIOTOXIN CTXI AS DERIVED FROM NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY AND DISTANCE	71.2	71.2	100%	2e-18	61%	2CDX_A
Chain S, Elucidation Of The Solution Structure Of Cardiotoxin Analogue V From The Taiwan Cobra (Naja Naja Atra) Venom	70.1	70.1	100%	5e-18	60%	1CHV_S
Chain A, Crystal Structure Of Cardiotoxin Vi From Taiwan Cobra (Naja Atra) Venom	70.1	70.1	100%	6e-18		

Similar types of data analyzes can be carried out for nucleotide sequences and as well for other BLAST algorithms. Try the BLAST searches using the sequences of proteins and nucleic acids you are interested in and analyze the data outputs.

Pairwise alignments

Retrieve the two protein sequences under interest either from primary or structural databases and save them in 'Fasta' format. Either the sequences or accession IDs of the sequences can be pasted in the space provided under categories of 'Query sequence' and 'Subject sequence' of the algorithm. Upon selecting appropriate alignment algorithm and gap penalty options, the program can be run for aligning the sequences.



For example, two cardiotoxins bearing PDB IDs of 1CRF and 1CVO were aligned and the results are shown below herein.

Sequence ID: Query_206921Length: 62Number of Matches: 1
Related Information
Range 1: 1 to 62Graphics Next Match Previous Match First Match

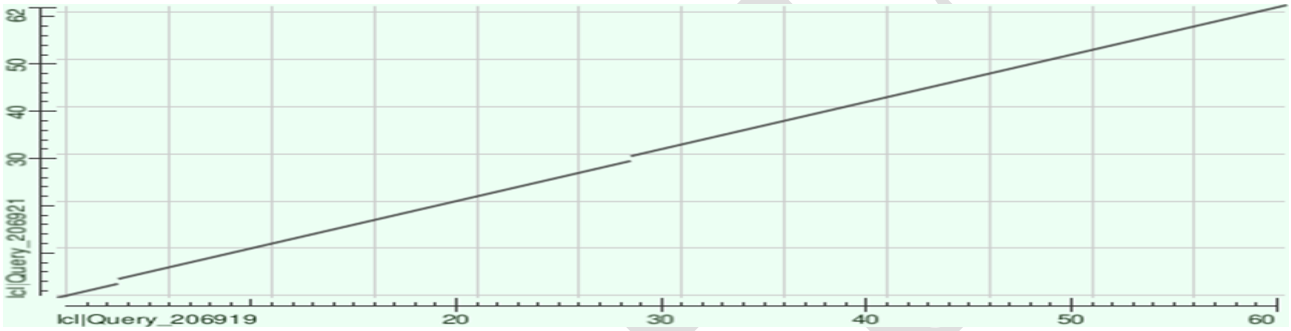
Alignment statistics for match #1

NW Score	Identities	Positives	Gaps	Frame
213	40/62(65%)	46/62(74%)	2/62(3%)	

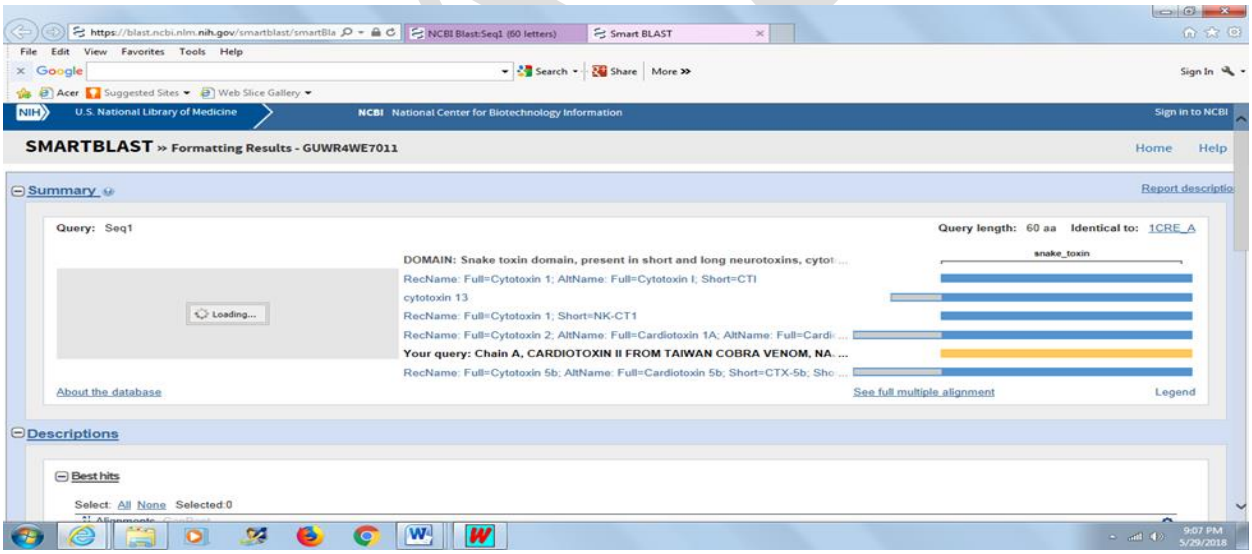
Features:

```
Query  1    LKC-NKLVPLFYKTCPAGKNLCYKMFMVS-NLTVPVKRGCIDVCPKNSALVKYVCCNTDR  58
            LKC N  +P  YKTCP GKNLC+K  +    L  PVKRGC D CPKNSAL+KYVCC+TD+
Sbjct  1    LKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRGCADNCPKNSALLKYVCCSTDK  60

Query  59   CN  60
             CN
Sbjct  61   CN  62
```

The dot plot for the sequences is depicted in the figure shown herein.



The sequence alignment can also be performed by using SmartBLAST and the results are as shown below herein.



Conclusions:

4. Multiple sequence alignments and phylogenetic analysis

Aim: To examine multiple sequences alignment of protein sequences using Clustal Omega and ClustalW

Procedure

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. This tool can align up to 4000 sequences or a maximum file size of 4 MB.

The following 11 sequences were first retrieved from PDB database and the data were stored in 'Fasta' format.

>1CRF *Naja naja atra*
LKCNKLVPLFYKTCPAGKNLCYKMFMVSNLTVPVKRGCIDVCPKNSALVKYVCCNTDRCN

>1CRE *Naja naja atra*
LKCNKLVPLFYKTCPAGKNLCYKMFMVSNLTVPVKRGCIDVCPKNSALVKYVCCNTDRCN

>1CHV *Naja naja atra*
LKCNKLVPLFYKTCPAGKNLCYKMFMVSNKMVPVKRGCIDVCPKSSLLVKYVCCNTDRCN

>P60301.1
MKTLLLTLVVVTIVCLDLGYTLKCNKLVPLFYKTCPAGKNLCYKMFMVATPKVPVKRGCIDVCPKSSLLV KYVCCNTDRCN

>P80245.2
MKTLLLTLVVVTIVCLDLGYTLKCNQLIPPFYKTCAAGKNLCYKMFMVAAPKVPVKRGCIDVCPKSSLLV KYVCCNTDRCN

>P60304.1
MKTLLLTLVVVTIVCLDLGYTLKCNKLIPIASKTCPAGKNLCYKMFMMSDLTIPVKRGCIDVCPKNSLLV KYVCCNTDRCN

>1CXO
LKCNQLIPPFWKTCPKGKNLCYKMTMRAAPMVPVKRGCIDVCPKSSLLIKYMCCNTDKCN

>1CXN
LKCNQLIPPFWKTCPKGKNLCYKMTMRAAPMVPVKRGCIDVCPKSSLLIKYMCCNTDKCN

>P01442.2
MKTLLLTLVVVTIVCLDLGYTLKCNKLVPLFYKTCPAGKNLCYKMFMVSNLTVPVKRGCIDVCPKNSALV KYVCCNTDRCN

>1CVO
LKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRGCADNCPKNSALLKYVCCSTDKCN

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

>P62375.1
MKTLLLTMVVVTIVCLDLGYTLKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRGCADNCPKNSA
LLKYVCCSTDKCN

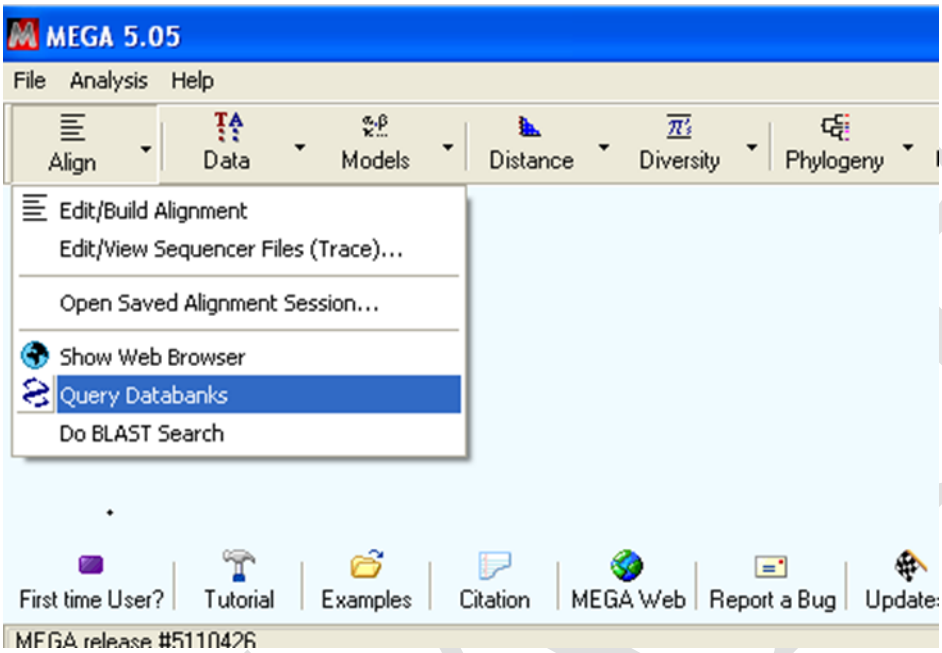
All the 11 sequences were pasted one-by-one in the 'sequence space' provided in the Clustal Omega program. After setting appropriate alignment algorithm, matrix and output format, the program was executed and the outcomes are shown below herein.

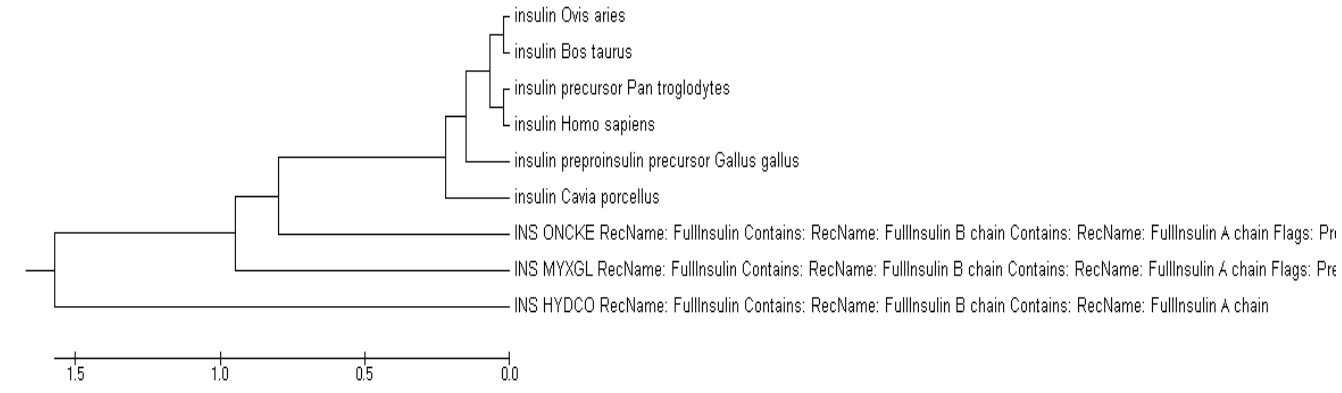
CLUSTAL O(1.2.4) multiple sequence alignment

```
1CVO            ----------------------
LKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRG   39
P62375.1
MKTLLLTMVVVTIVCLDLGYTLKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRG   60
1CXO            ----------------------LKCNQ-LIPPFWKTCPKGKNLCYKMTMRA-
APMVPVKRG       37
1CXN            ----------------------LKCNQ-LIPPFWKTCPKGKNLCYKMTMRA-
APMVPVKRG       37
P60301.1        MKTLLLTLVVVTIVCLDLGYTLKCNK-LVPLFYKTCPAGKNLCYKMFMVA-
TPKVPVKRG       58
P80245.2        MKTLLLTLVVVTIVCLDLGYTLKCNQ-LIPPFYKTCAAGKNLCYKMFMVA-
APKVPVKRG       58
P60304.1        MKTLLLTLVVVTIVCLDLGYTLKCNK-LIPIASKTCPAGKNLCYKMFMMS-
DLTIPVKRG       58
1CRF            ----------------------LKCNK-LVPLFYKTCPAGKNLCYKMFMVS-
NLTVPVKRG       37
1CRE            ----------------------LKCNK-LVPLFYKTCPAGKNLCYKMFMVS-
NLTVPVKRG       37
P01442.2        MKTLLLTLVVVTIVCLDLGYTLKCNK-LVPLFYKTCPAGKNLCYKMFMVS-
NLTVPVKRG       58
1CHV            ----------------------LKCNK-LVPLFYKTCPAGKNLCYKMFMVS-
NKMVPVKRG       37
                                     ***::  :*   ***  *****:*  :
.*****
1CVO            CADNCPKNSALLKYVCCSTDKCN     62
P62375.1        CADNCPKNSALLKYVCCSTDKCN     83
1CXO            CIDVCPKSSLLIKYMCCNTDKCN     60
1CXN            CIDVCPKSSLLIKYMCCNTDKCN     60
P60301.1        CIDVCPKSSLLVKYVCCNTDRCN     81
P80245.2        CIDVCPKSSLLVKYVCCNTDRCN     81
P60304.1        CIDVCPKNSLLVKYVCCNTDRCN     81
1CRF            CIDVCPKNSALVKYVCCNTDRCN     60
1CRE            CIDVCPKNSALVKYVCCNTDRCN     60
P01442.2        CIDVCPKNSALVKYVCCNTDRCN     81
1CHV            CIDVCPKSSLLVKYVCCNTDRCN     60
                * * ***.* *:**:**.**:**
```

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

Phylogenetic tree can be constructed for a set of protein sequences (as well nucleotide sequences) using MEGA computational tool. Snapshots of various steps involved in the procedure are depicted herein below.





Conclusions:

5. Comparative modeling using online and standalone tools

Aim: To predict three-dimensional structures of a given query sequence.

Procedure

The computational techniques that are generally being used to predict 3D structures of protein have been divided into three classes: i) Homology modelling, which is also popularly known as comparative modelling ii) Modelling through Threading and iii) *Ab initio* modelling. In the 'Homology modelling', the unknown protein structure (to be modelled) in any organism can be modelled/predicted from the homologous structure (Template) that is available from the PDB. The sequences of both model and the template need to be aligned further to carry out the homology modelling.

The modeling can be carried out using SWISS-MODEL (a webserver) and as well using MODELLER, a standalone tool. Procedure for generating 3D structures of proteins by using MODELLER tools has been systematically describe below herein.

Target sequence

Obtain FASTA sequence of Cardiotoxin 1 or cytotoxin 1 from *Naja naja*.
UniProt accession No. P01447

Template identification (using BLAST)

The template obtained from BLAST (1CHV) can be viewed using SPDBV/PyMol/RasMol and its sequence should be stored in FASTA format.

File preparations:

The sequence of the protein with unknown 3D structure is the "**target sequence**".

A 3D **template** is chosen by virtue of having the highest sequence identity with the target sequence. A published atomic coordinate "PDB" file from the Protein Data Bank.

An **alignment** between the target sequence and the template sequence is depicted herein.

Atom file (.atm, .pdb)

Each atom file is named 'temp.atm' and the temp file must be used as that protein's identifier throughout the modelling.

```
ATOM     33  HB2 LYS S   2      13.260  57.886  21.203  1.00  1.14           H
ATOM     34  HB3 LYS S   2      12.876  58.854  22.560  1.00  1.21           H
ATOM     35  HG2 LYS S   2      13.675  60.309  20.754  1.00  1.31           H
ATOM     36  HG3 LYS S   2      14.781  60.457  22.057  1.00  1.25           H
ATOM     37  HD2 LYS S   2      16.506  59.364  20.776  1.00  1.02           H
ATOM     38  HD3 LYS S   2      15.250  58.436  19.744  1.00  1.07           H
ATOM     39  HE2 LYS S   2      14.775  60.414  18.508  1.00  1.88           H
ATOM     40  HE3 LYS S   2      15.831  61.374  19.535  1.00  1.49           H
ATOM     41  HZ1 LYS S   2      16.833  59.150  17.766  1.00  2.47           H
ATOM     42  HZ2 LYS S   2      16.791  60.889  17.604  1.00  2.43           H
ATOM     43  HZ3 LYS S   2      17.702  60.223  18.796  1.00  2.41           H
```

Alignment file (.ali)

The format for the alignment file is related to the PIR database, this is the preferred format for comparative modelling. For the aligned regions, MODELLER tries to derive a 3D model for the target sequence that is as close to one or the other of the template structures as possible while also satisfying stereo-chemical restraints (e.g. bond lengths, angles, non-bonded atom contacts).

```
>P1;temp
structureX:temp:1    :S:60  :S:ferredoxin:Azotobacter vinelandii: 1.90: 0.19
LKCNKLVPLFYKTCPAGKNLCYKMFMVSNKMVPVKRGCIDVCPKSSLLVK
YVCCNTDRCN*

>P1;target
sequence:target:1    : :60   : ferredoxin:Peptococcus aerogenes: 2.00:-1.00
LKCNKLIPLAYKTCPAGKNLCYKMYMVSNKTVPVKRGCIDVCPKNSLVLK
YECCNTDRCN*
```

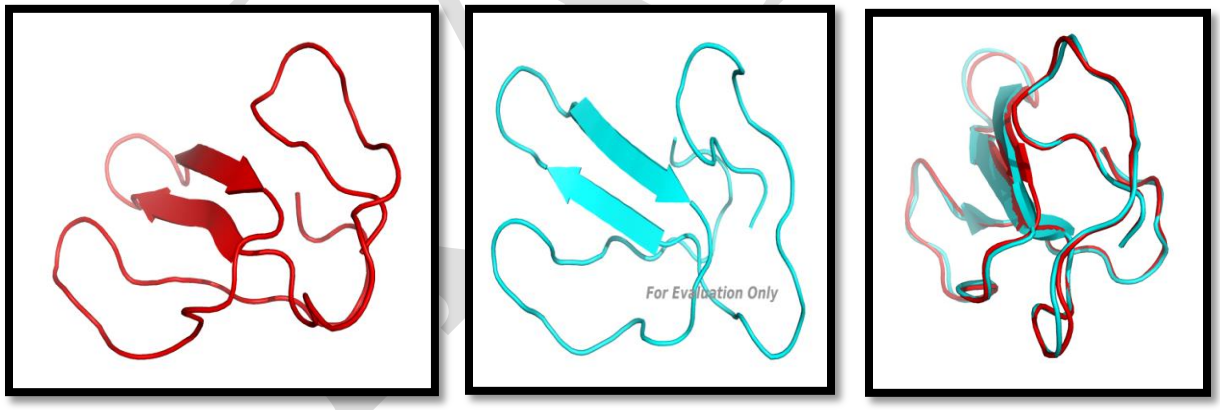
Script file (.py)

MODELLER is a command-line only tool, and has no graphical user interface. Instead, script file containing MODELLER commands should be provided. This file is a Python script. A

script file is given as script.py to produce one model of sequence from the known structure of template.

```
# Homology modeling by the automodel class

from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()                    # request verbose output
env = environ()                   # create a new MODELLER environment to build this model in

                                 # directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']

a = automodel(env,
              alnfile  = 'alignment.ali',      # alignment filename
              knowns   = 'temp',               # codes of the templates
              sequence = 'target')             # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                 # (determines how many models to calculate)
a.make()                            # do the actual homology modeling
```

The model structures obtained can be viewed by using PyMol as shown herein.



Try homology modelling of the protein sequences you are interested in (using single template approach as described in the manual).

Conclusions:

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

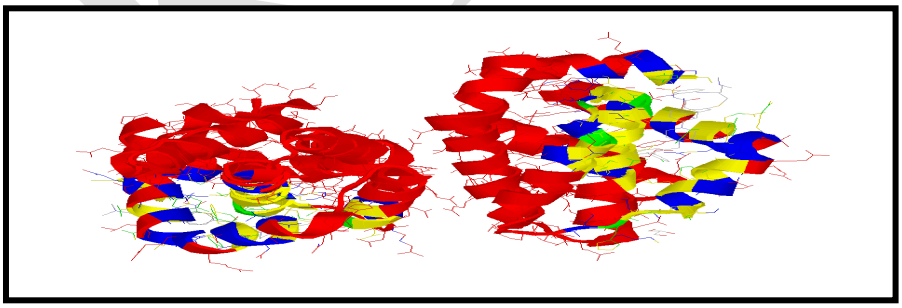
6. Molecular visualization tools: RasMol, JMol and PyMol

Aim: To learn in order to produce various representations of protein structures with high quality resolutions

Procedure

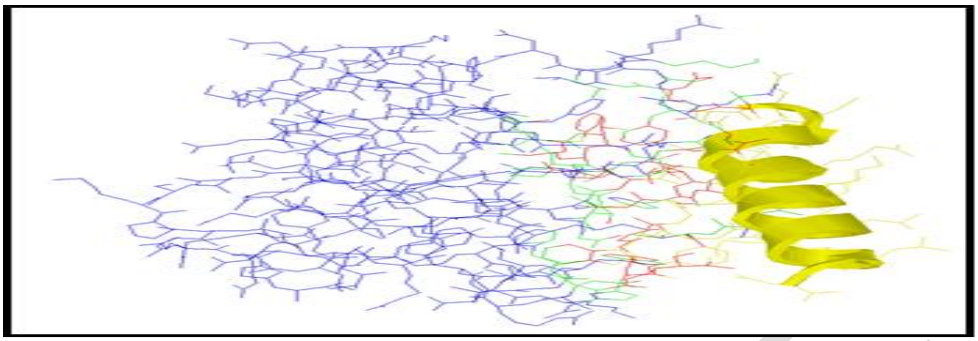
RasMol is a computer program written for molecular graphics visualization intended primarily for use in depiction and exploration of biological macromolecule structures such as those found in the Protein Data Bank (PDB). RasMol includes a language for selecting certain protein chains, changing colours, etc.

SYNTAX	DEFINITION
select all	Selects everything
Colour colour name	Changes the colour of the selected atoms
background colour	Colours the background in the specified colour
select hydrophobic	Selects all hydrophobic residues
select polar	Selects all the polar residues
select alpha	Selects all alpha- carbons
select helix	Selects all residues in alpha- helices



2XYJ protein-ligand dimer viewed in cartoon with for both the chains. The atoms within and 4.0 A° (yellow) , 6.0 A°(blue) and 8.0 A°(green) are represented in different colors.

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

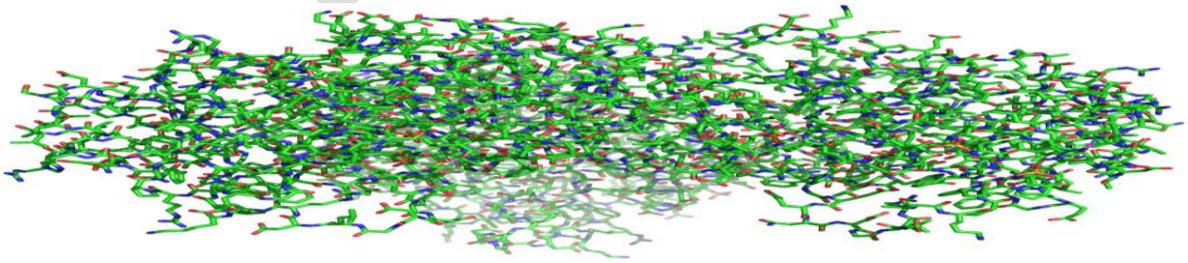


2NLA complex viewed with the protein region as lines and the peptide as a ribbon. The residues within 4.0A°(yellow), 6.0A°(green) and 8.0 A°(red) are shown in different colours.

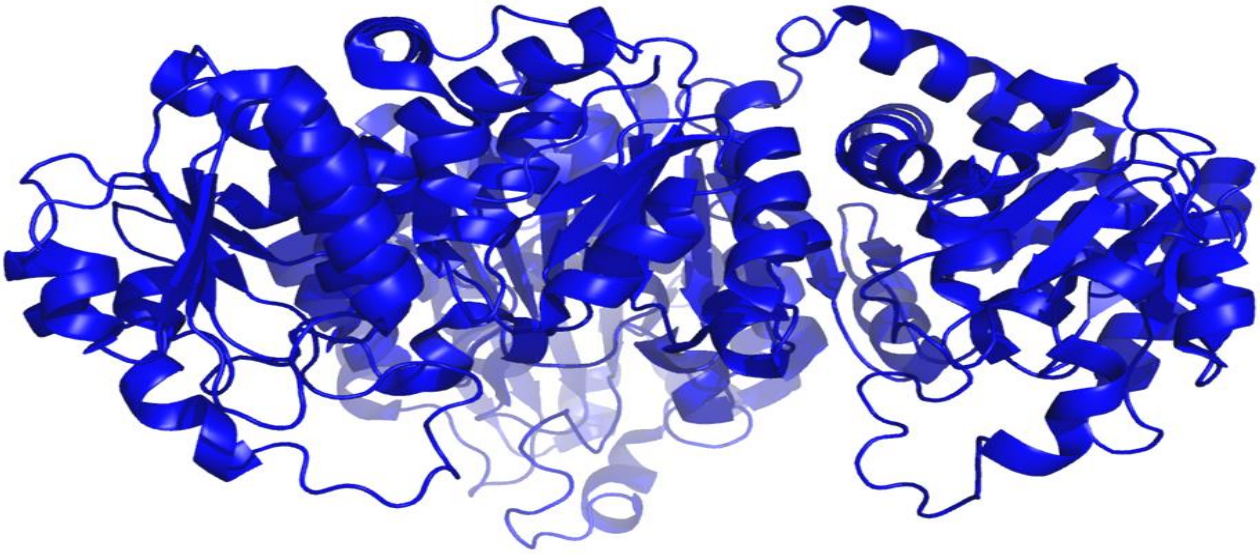
PyMOL is a molecular viewer, render tool and 3D molecular editor intended for visualization of 3B chemical structures including atomic resolution X-ray crystal structures of proteins, nucleic acids, etc., As a stereochemistry viewer, PyMOL illustrates the 3D stereochemical relationships of organic chemistry. PDB is a Protein Data Bank ie., it is a repository of protein 3D structures and nucleic acid 3D structural data (large biological molecules). RMSD is the Root Mean Square Deviation. It tells how an atom deviates from the original position (mean) while superimposing.

Download the PDB files 2I6U from http://www.rcsb.org. The PDB file contains the HEADER, Title, Experiment data, Author, Journal, Remark, SEQRES, HELIX, LINK, SITE, ATOM coordinates, HETATM information.

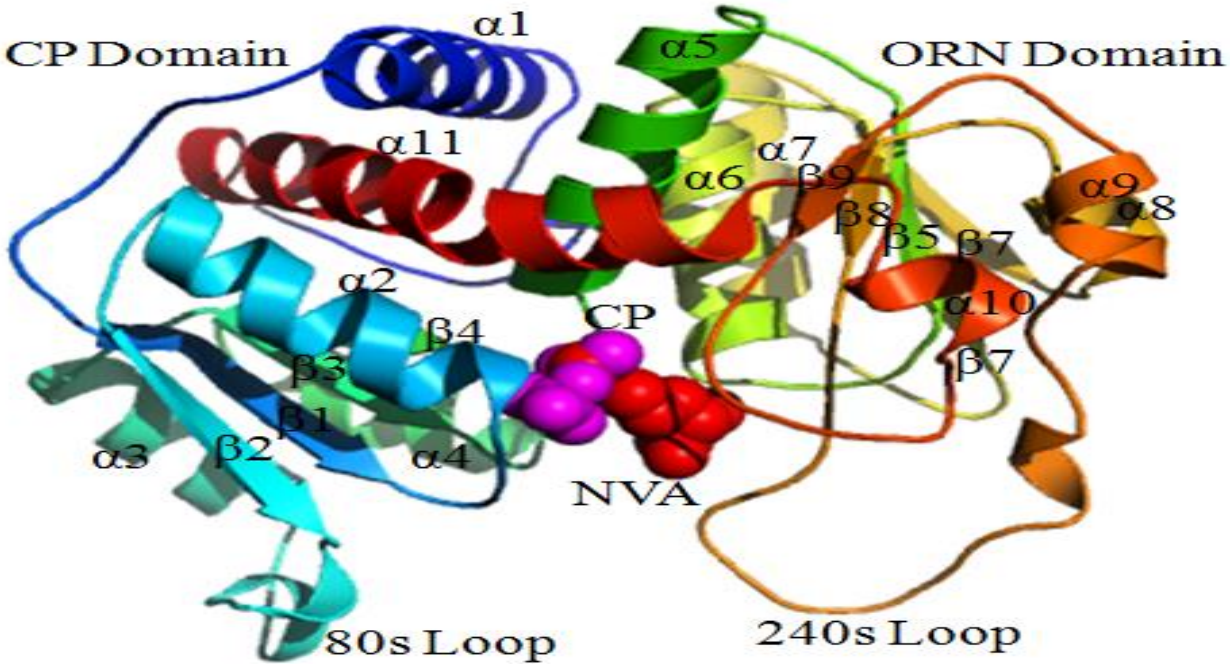
The PDB file 2I6U was downloaded and its resolution was found to be 2.20Å. Various representation (Stick, ribbon, cartoon, surface representations etc.) of the PDB file 2I6U were produced. The advantage of using stick representation is location of oxygen, nitrogen and carbon atoms present in the residues can be known easily.



CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

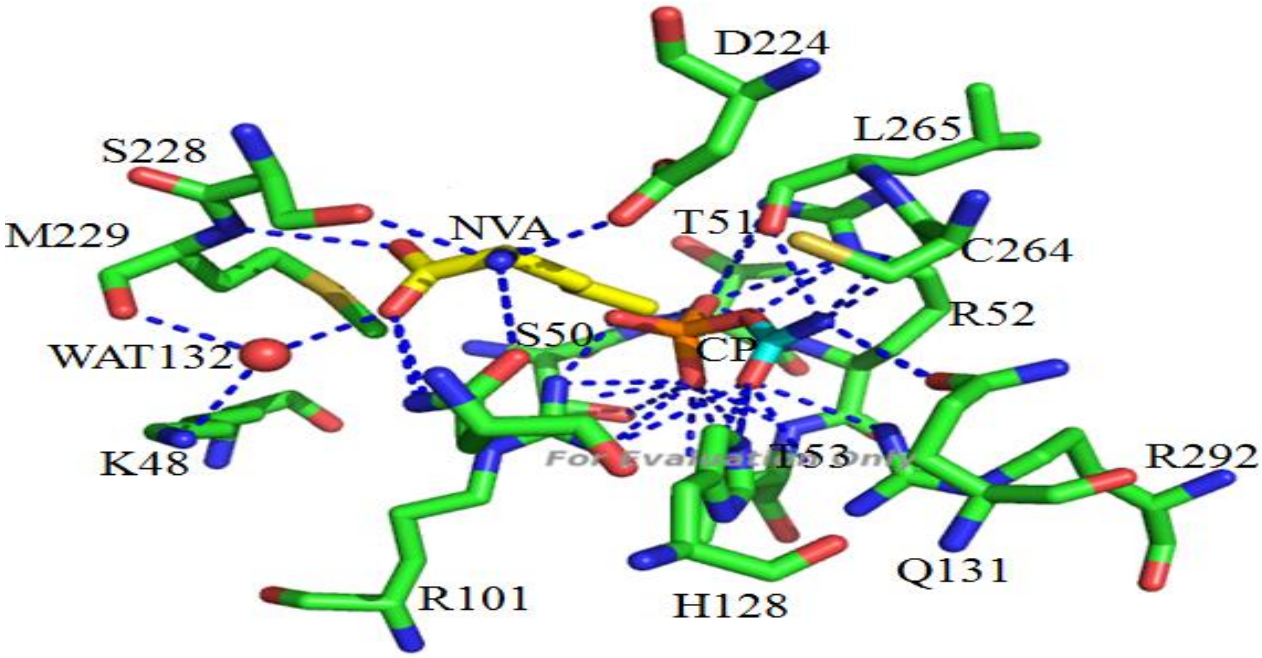


Cartoon representation of the PDB file 2I6U.

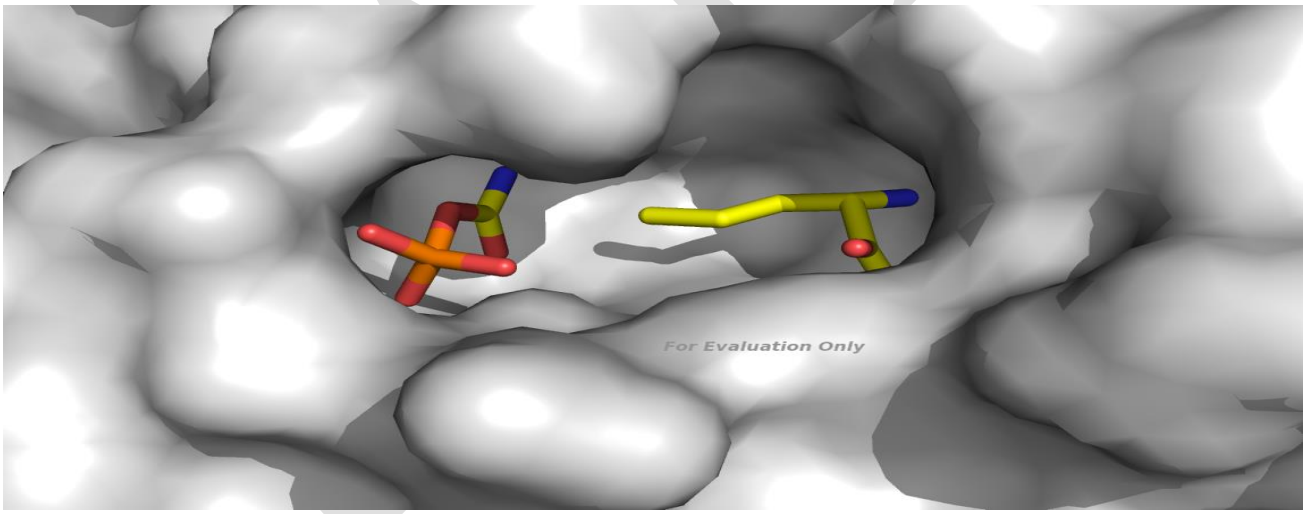


Cartoon representation of a monomer of *Mtb* Ornithine Carbamoyltransferase

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312



Stereo views of hydrogen bonding interactions.



Cartoon representation of active site of the 2I6U with ligands

Similarly, the above exercises can also performed using JMol and other molecular visualizing tools you are interested in.

Conclusions:

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

7. Structural analysis and verification tools

Aim: To assess the quality of 3D structures of proteins determined by experimental and as well by computational methods

Procedure

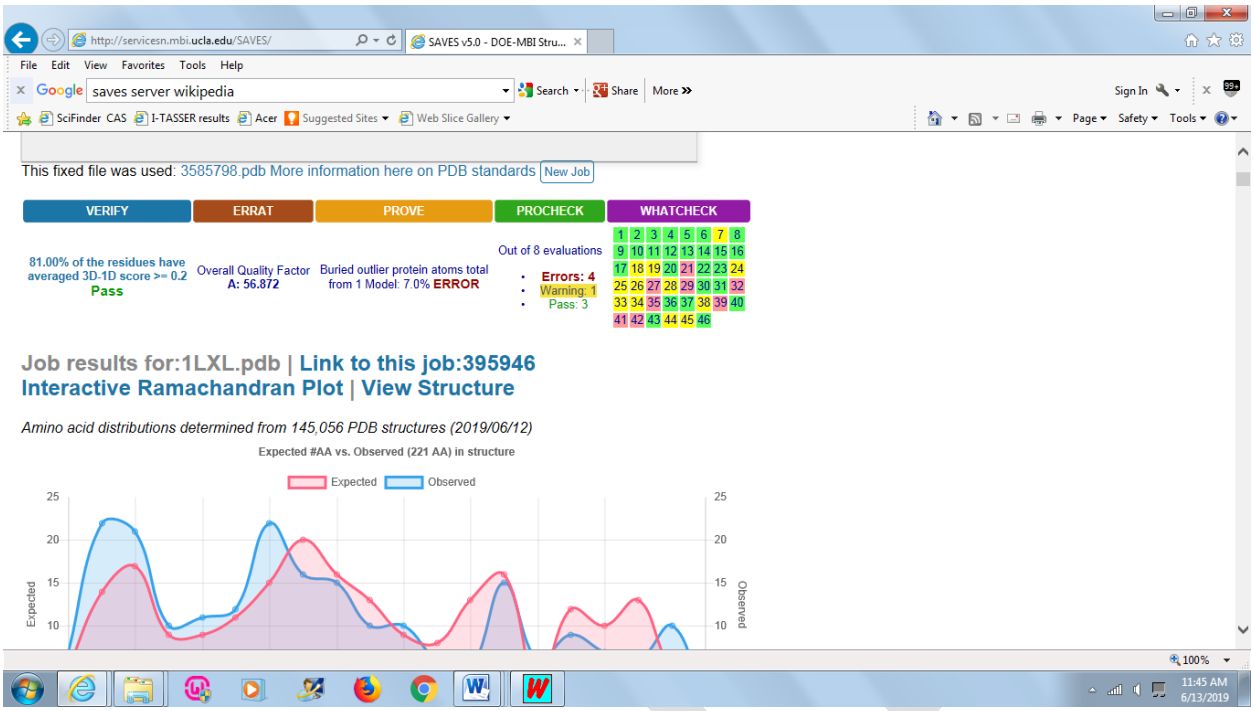
Macromolecular structure validation is the process of evaluating reliability for three-dimensional atomic models of large biological molecules such as proteins and nucleic acids. These models, which provide 3D coordinates for each atom in the molecule (see example in the image), come from structural biology experiments such as x-ray crystallography or nuclear magnetic resonance (NMR). The validation has three aspects: 1) checking on the validity of the thousands to millions of measurements in the experiment; 2) checking how consistent the atomic model is with those experimental data; and 3) checking consistency of the model with known physical and chemical properties.

Proteins and nucleic acids are the workhorses of biology, providing the necessary chemical reactions, structural organization, growth, mobility, reproduction, and environmental sensitivity. Essential to their biological functions are the detailed 3D structures of the molecules and the changes in those structures. To understand and control those functions, we need accurate knowledge about the models that represent those structures, including their many strong points and their occasional weaknesses.

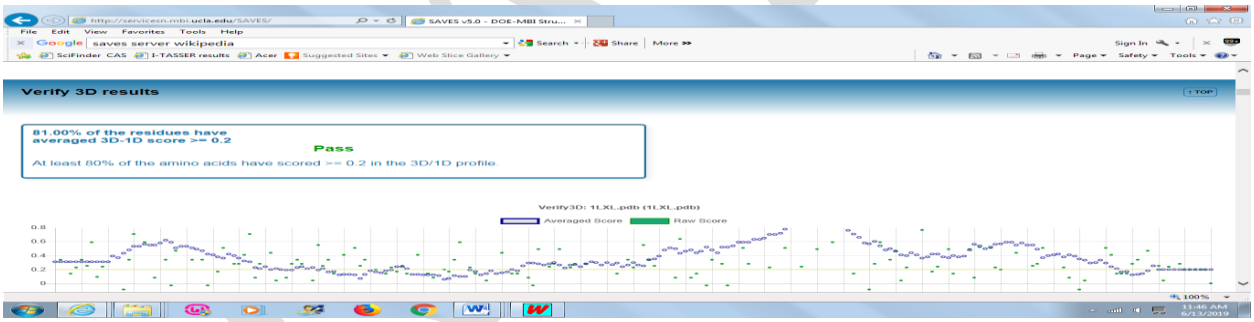
The structural quality of 3D structures of proteins can be assessed from many facets using the following webserver: http://servicesn.mbi.ucla.edu/SAVES/. The server provides following computations methods to achieve the above said tasks: Verify3D, ERRAT, PROVE, PROCHECK and WHATCHECK.

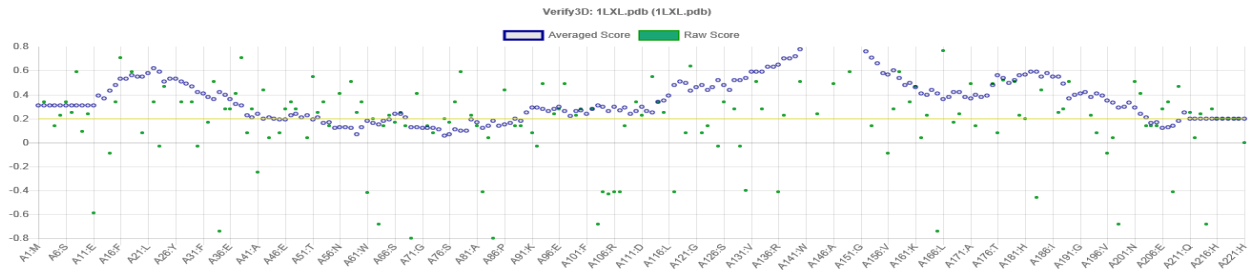
For instance, the PDB file 1LXL is used to assess its structural quality using the server and the results have been depicted herein below.

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312



The Verify 3D results are shown herein.



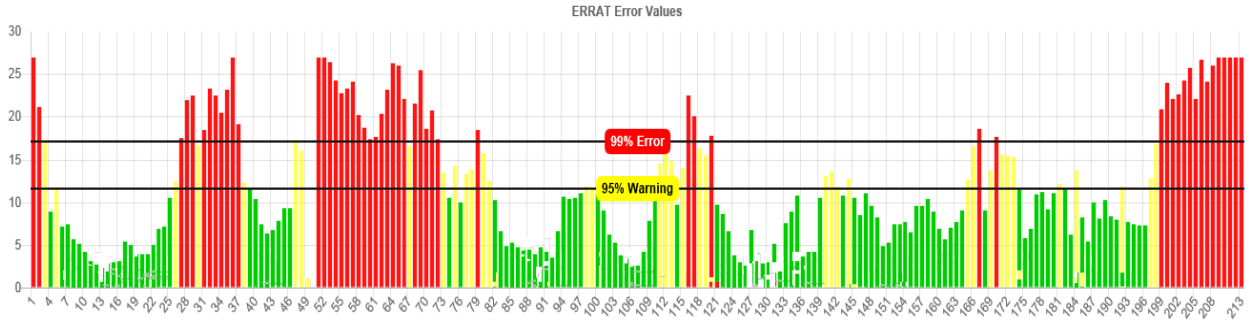


The ERRAT results are shown herein.

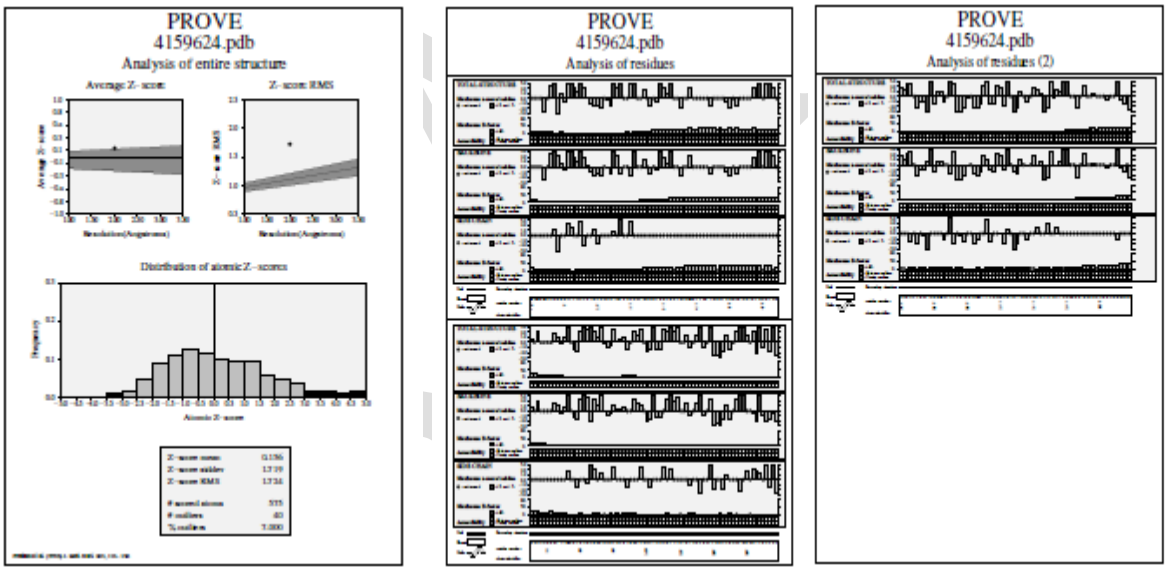
ERRAT results ↑ TOP

Input: 1LXL.pdb (1LXL.pdb

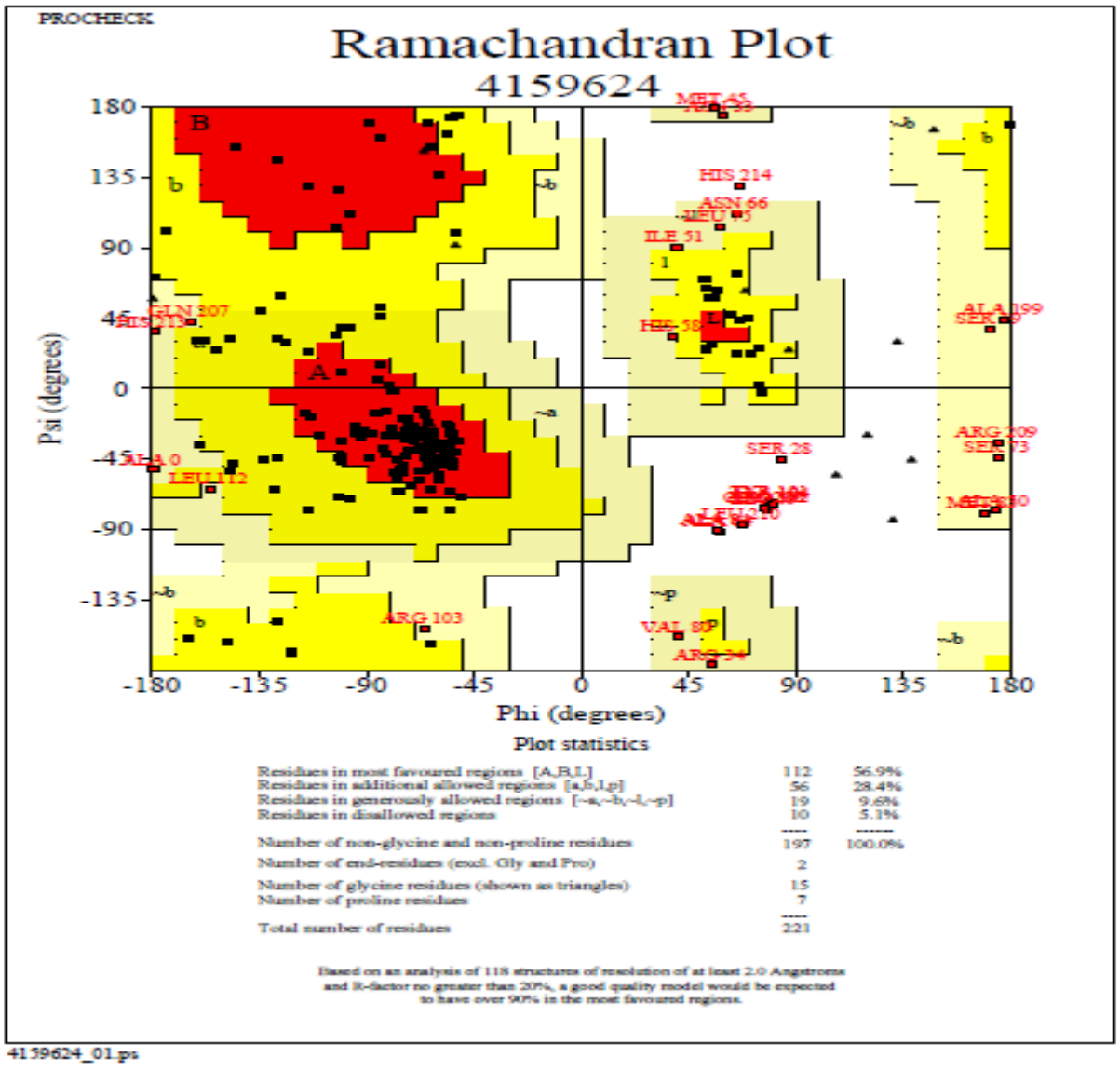
Quality Factor: A: 56.872 | PDF | PostScript | Log



The PROVE results for the protein structure.



The 'Ramachandran Plot' analyses of the protein structure.



The above results should be systematically analyzed and on the basis of the outputs, the overall structural quality of the protein structures would be rationalized.

Conclusions:

CLASS : II M. Sc., BT
COURSE NAME : GENOMICS, PROTEOMICS AND BIOINFORMATICS PRACTICAL VI
BATCH : 2018 – 2020
COURSE CODE : 18BTP312

8. Molecular dockings of biological macromolecules

Aim: To generate docking complexes of two proteins using a few molecular docking tools and to analyze the results in a comprehensive manner.

Procedure

Docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex i.e., how well the particular (ligand) molecule fits into the particular target (how well the drug binds to the receptor). There are many types of docking. They are:

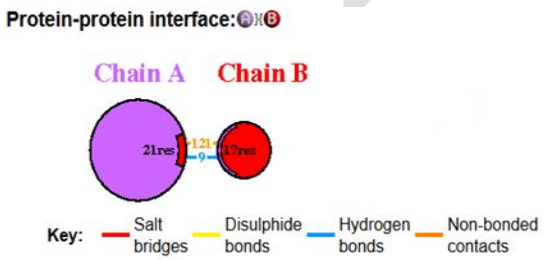
- Protein-protein docking
- Protein-Ligand docking
- Protein-DNA docking

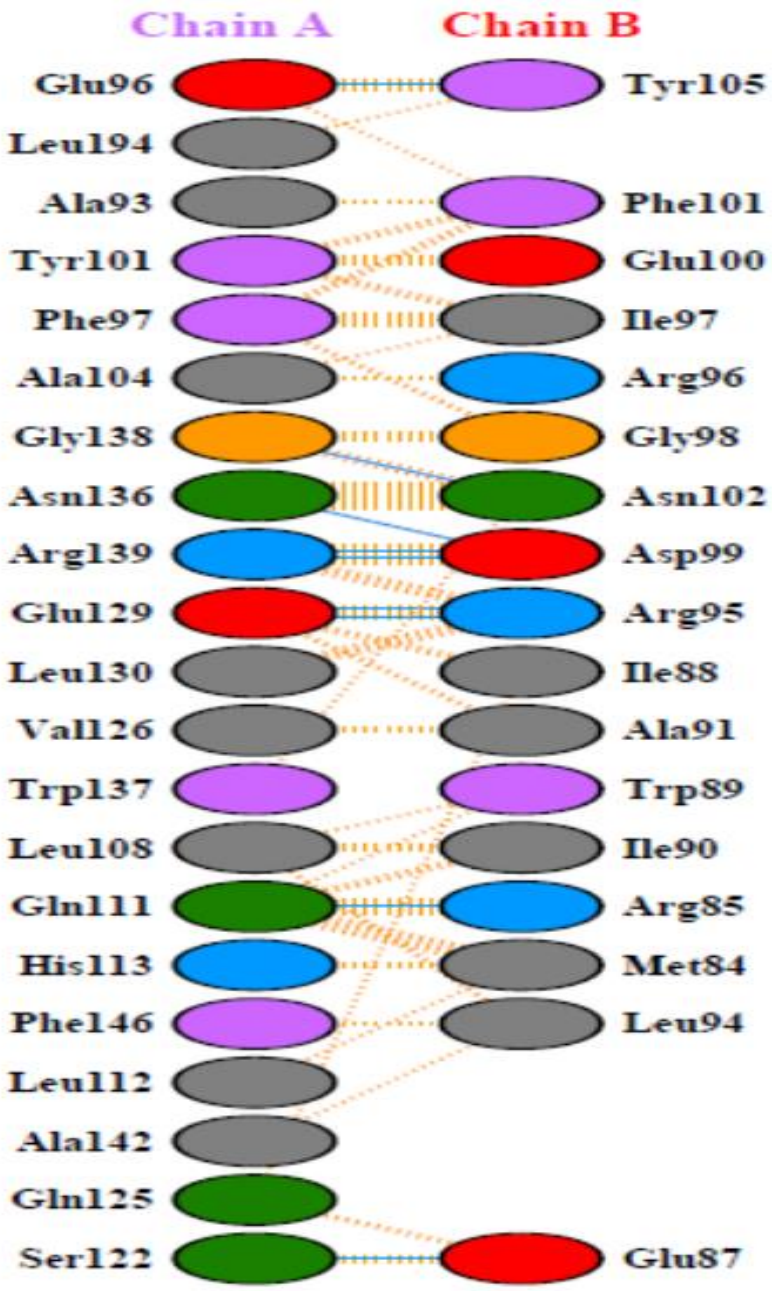
HEX: Hex (http://hex.loria.fr/) is an interactive protein docking and molecular superposition program. Hex understands protein and DNA structures in PDB format, and it can also read small-molecule SDF files.

ZDOCK: ZDOCK (http://zdock.umassmed.edu/) server is an interactive docking prediction of protein-protein complexes and symmetric multimers.

ClusPRO: ClusPro (http://nrc.bu.edu/cluster) represents the first fully automated, web-based program for the computational docking of protein structures.

A protein-peptide complex (PDB ID: 3FDL; BCLXL and BIM complex) was downloaded from RCSB PDB. Binding site information could be gathered from the corresponding references. The two polypeptide chains were saved as follows: Chain A is BCLX and chain B is BIM; presence of water molecules was deleted in both cases.

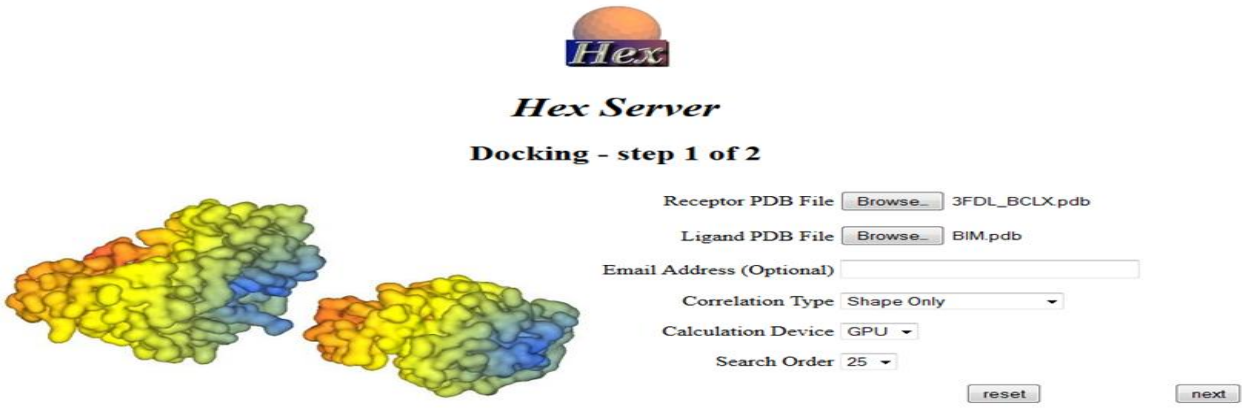




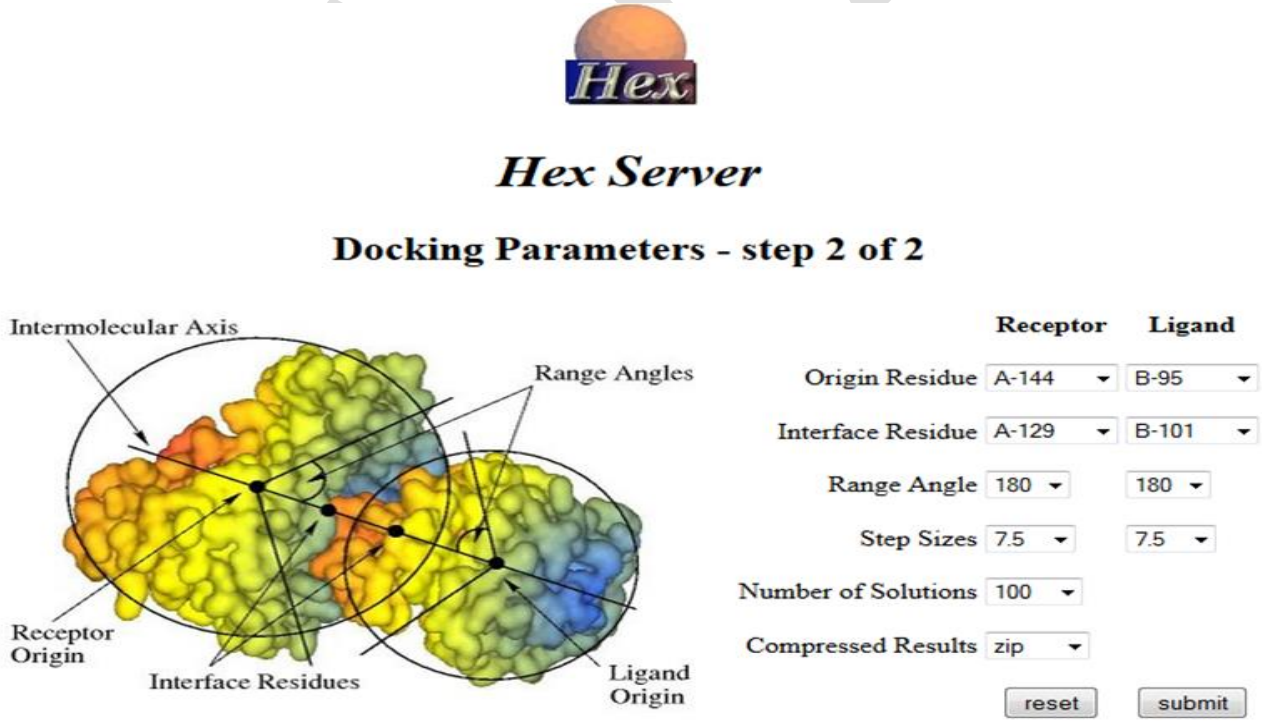
Interfacial residues of BclXL and Bim complex.

Complex generations were carried out using the HEX, Z-DOCK and ClusPro servers.

The results have been displayed one-by-one below herein.

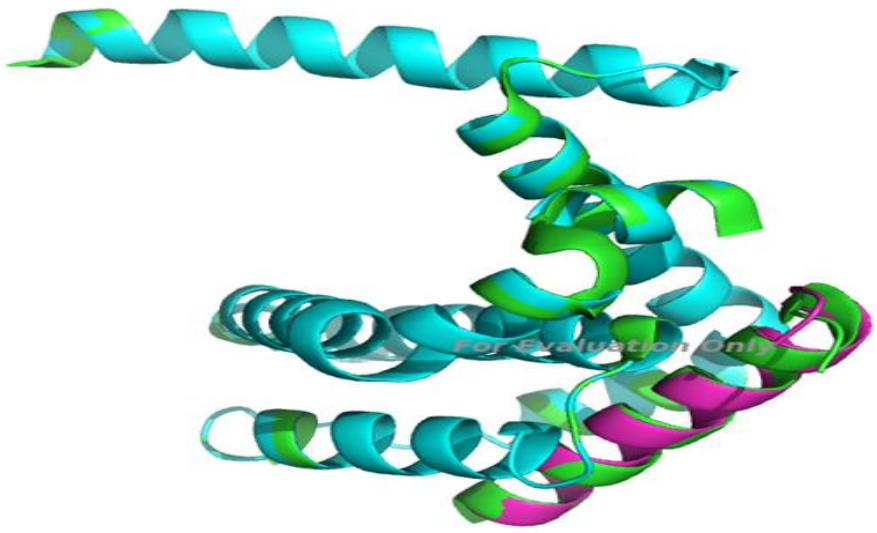


In the Origin residue (present near the centroid), Phe144 and Arg95 were given for receptor and ligand respectively. In the Interface residue (making many interactions), Glu129 and Phe101 were given for receptor and ligand respectively.



ClusPro:

- The receptor and ligand files along with their interfacial residues were given as input.
- 10 models were generated.
- Randomly model 2 was selected and aligned with the reference complex in PyMol. The observed RMSD value for receptor is : 0.0000 Å and for ligand is 1.541 Å.



Superimposition of 3FDL (experimental structure shown in green) and computationally generated structure (receptor shown in cyan) and ligand (shown in pink).

Conclusions: