16BTU512A BIOINFORMATICS PRACTICAL 4H - 2C

Total hours/week: T:0 P:4 Marks: Internal: 40 External: 60 Total: 100

Practical

1. Sequence information resource
2. Understanding and use of various web resources: EMBL,Genbank, Entrez, Unigene, Protein information resource (PIR)
3. Understanding and using: PDB, Swissprot, TREMBL
4. Using various BLAST and interpretation of results.
5. Retrieval of information from nucleotide databases.
6. Sequence alignment using BLAST.
7. Multiple sequence alignment using Clustal W.

References

1. Ghosh, Z., & Bibekanand M. (2008). *Bioinformatics: Principles and Applications*. Oxford University Press.
2. Pevsner, J. (2009). *Bioinformatics and Functional Genomics* (2nd ed.). Wiley-Blackwell.
3. Campbell, A. M., & Heyer, L.J. (2006). *Discovering Genomics Proteomics and Bioinformatics* (2nd ed.). Benjamin Cummings.

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

1. Sequence information resources

Aim: To explore unique features and applications of various sequence resources

Procedure

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.

UniProt is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. UniProt Archive (UniParc) is a comprehensive and non-redundant database, which contains all the protein sequences from the main, publicly available protein sequence databases.

There are many publicly available protein sequence databases.

FlyBase: the primary repository of genetic and molecular data for the insect family Drosophilidae (FlyBase)

H-Invitational Database (H-Inv)

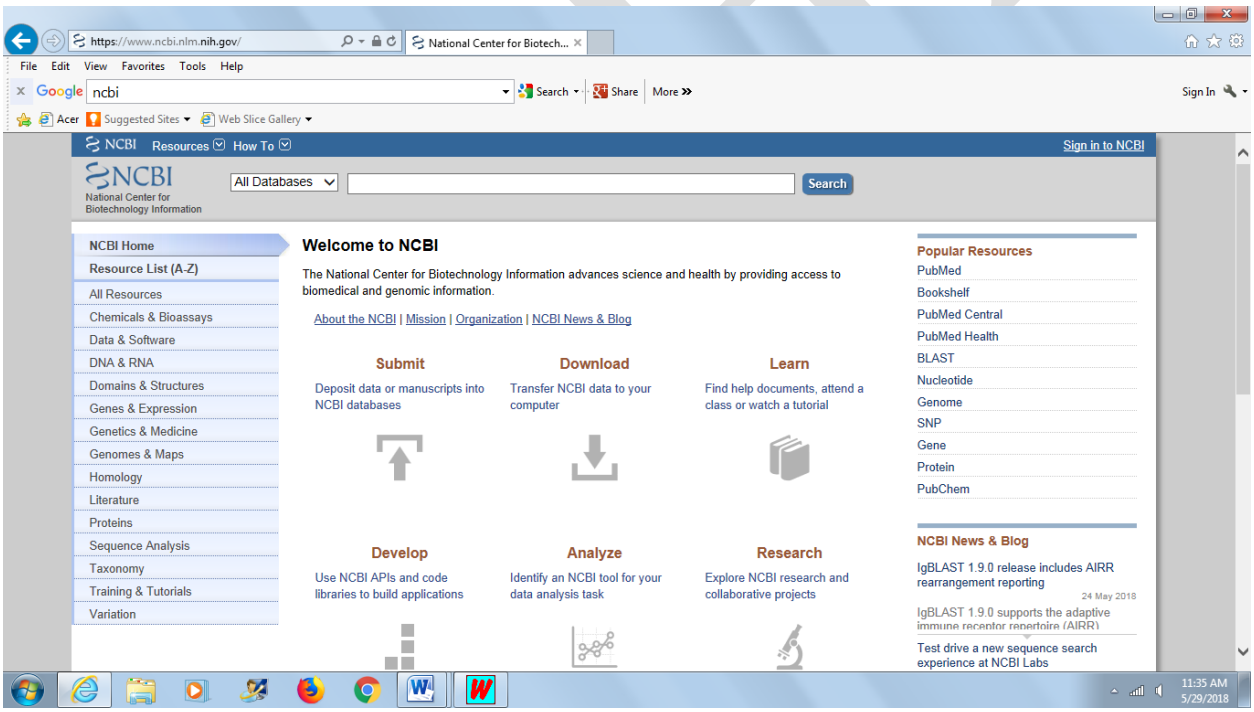
International Protein Index (IPI)

Saccharomyces Genome Database (SGD)

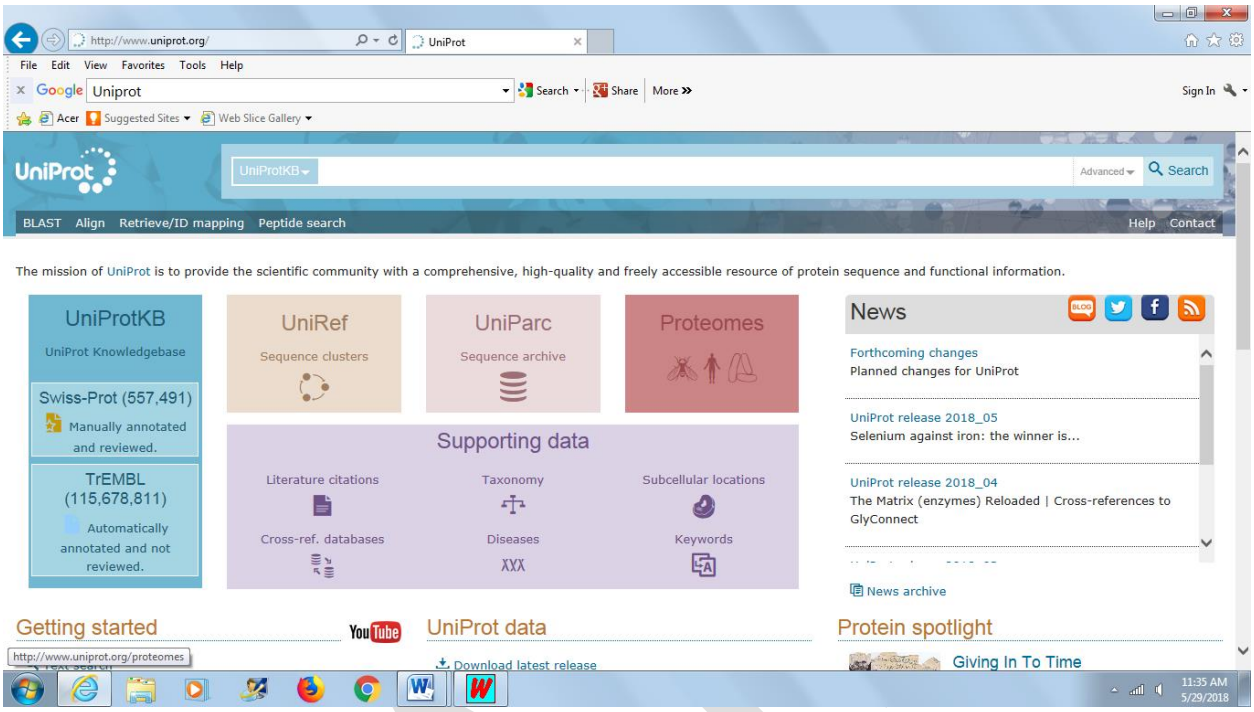
The Arabidopsis Information Resource (TAIR)

Vertebrate and Genome Annotation Database (VEGA)

WormBase



CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A



Information pertaining to primary, secondary and tertiary structures of proteins/nucleic acids can be inferred from the annotation webpages of the corresponding sequences. For instance, annotation data of myoglobin are herein listed out for the quick understanding.

Protein Myoglobin
Gene MB
Organism *Homo sapiens (Human)*
Status Reviewed-Annotation score:

Function

Serves as a reserve supply of oxygen and facilitates the movement of oxygen within muscles.

Sites

Feature key	Position(s)	DescriptionActions	Graphical view	Length
Metal binding[i]	65	Iron (heme distal ligand)		1
Metal binding[i]	94	Iron (heme proximal ligand)		1

GO - Molecular function

View the complete GO annotation on QuickGO
GO - Biological process[i]
Keywords

Molecular function	Muscle protein
Biological process	Oxygen transport, Transport
Ligand	Heme, Iron, Metal-binding

Enzyme and pathway databases

Reactome[i]	R-HSA-8981607 Intracellular oxygen transport

Protein family/group databases

TCDB[i]	1.A.107.1.3 the pore-forming globin (globin) family

Names & Taxonomy

Protein names	Recommended name: **Myoglobin**
Gene names	Name: **MB**
Organism	**Homo sapiens (Human)**
Taxonomic identifier	9606 [NCBI]

Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo
Proteomes	UP000005640 Component: Chromosome 22

Organism-specific databases

EuPathDB	HostDB:ENSG00000198125.12
HGNC	HGNC:6915 MB
MIM	160000 gene
neXtProt	NX_P02144

Subcellular location[i]

Extracellular region or secreted Cytosol Plasma membrane Cytoskeleton Lysosome Endosome Peroxisome ER Golgi apparatus Nucleus Mitochondrion Manual annotation Automatic computational assertionGraphics by Christian Stolte; Source: COMPARTMENTS

- GO - Cellular component
- Cytosol
  - cytosol Source: Reactome
- Extracellular region or secreted
  - extracellular exosome Source: UniProtKB

View the complete GO annotation on QuickGO

Pathology & Biotech

Organism-specific databases

DisGeNET	4151
OpenTargets	ENSG00000198125
PharmGKB	PA30658

Chemistry databases

ChEMBL	CHEMBL2406892
DrugBank	DB02671 1-Methylimidazole DB03385 4-Methylimidazole DB02379 Beta-D-Glucose DB02073 Biliverdine Ix Alpha

	DB03399 Ethyl Isocyanide DB04337 Methyl Isocyanide DB02396 Methylethylamine DB01826 N-Butyl Isocyanide DB04050 N-Propyl Isocyanide DB02646 Nitrosoethane DB01710 Porphyrin Fe(III) DB02528 Tetrazolyl Histidine

Polymorphism and mutation databases

BioMuta	MB
DMDM	127661

Results:

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

2. Understanding and use of various web resources: EMBL, GenBank, Entrez, Unigene and Protein information resources

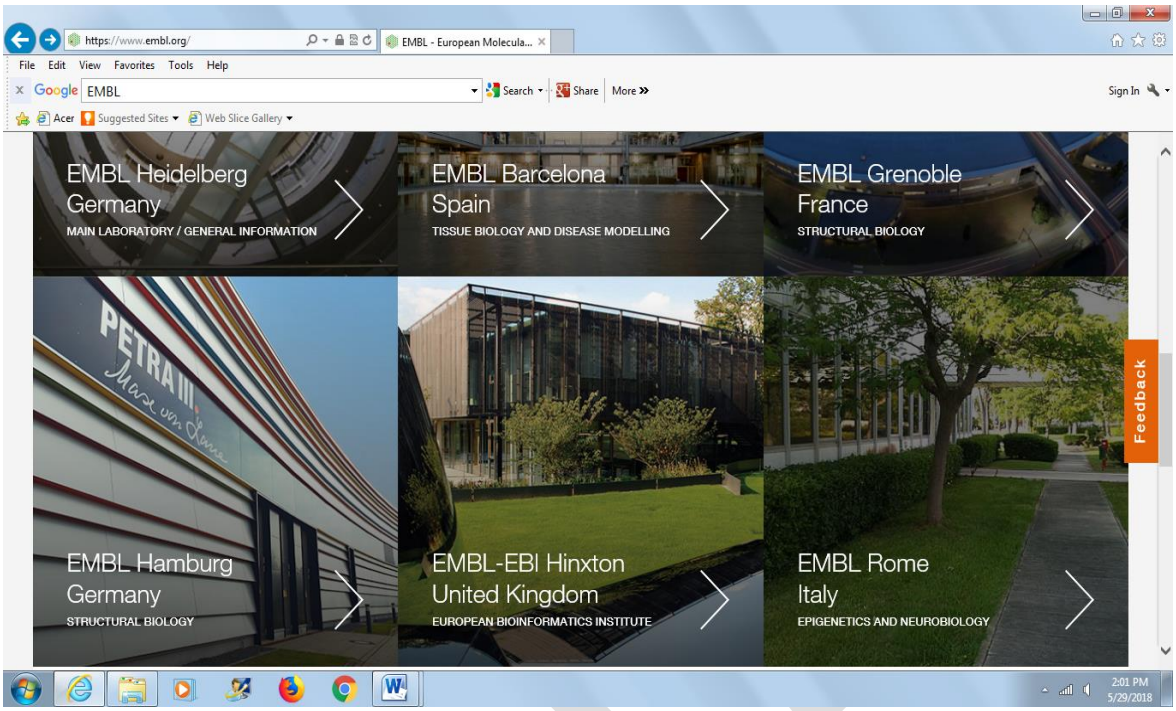
Aim: To understand and exploit the biological data available in various web resources.

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 organisation 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.

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A



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.

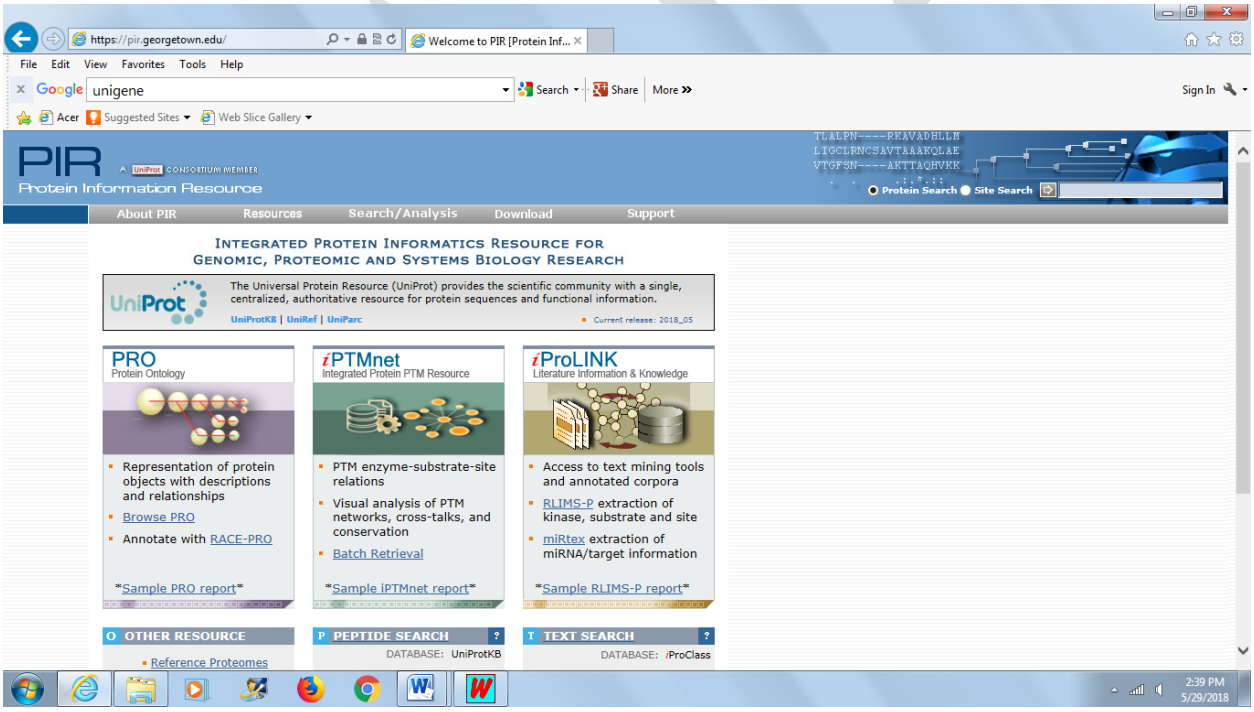
The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank.

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, and more. The system is produced by the National Center for Biotechnology Information (NCBI) and is available via the Internet

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.



Results:

3. Understanding and use of PDB, Swissprot and TrEMBL

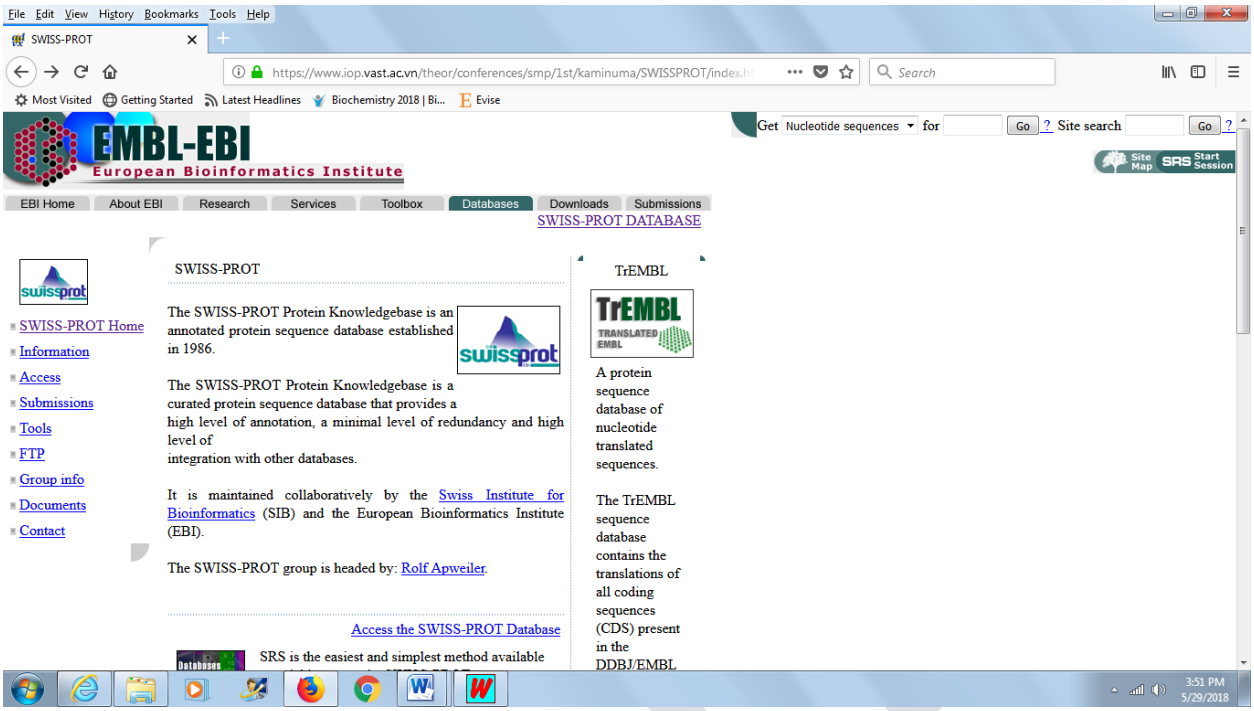
Aim: To analyze various sequence and structural information provided in the PDB, Swiss-Prot and TREMBL databases.

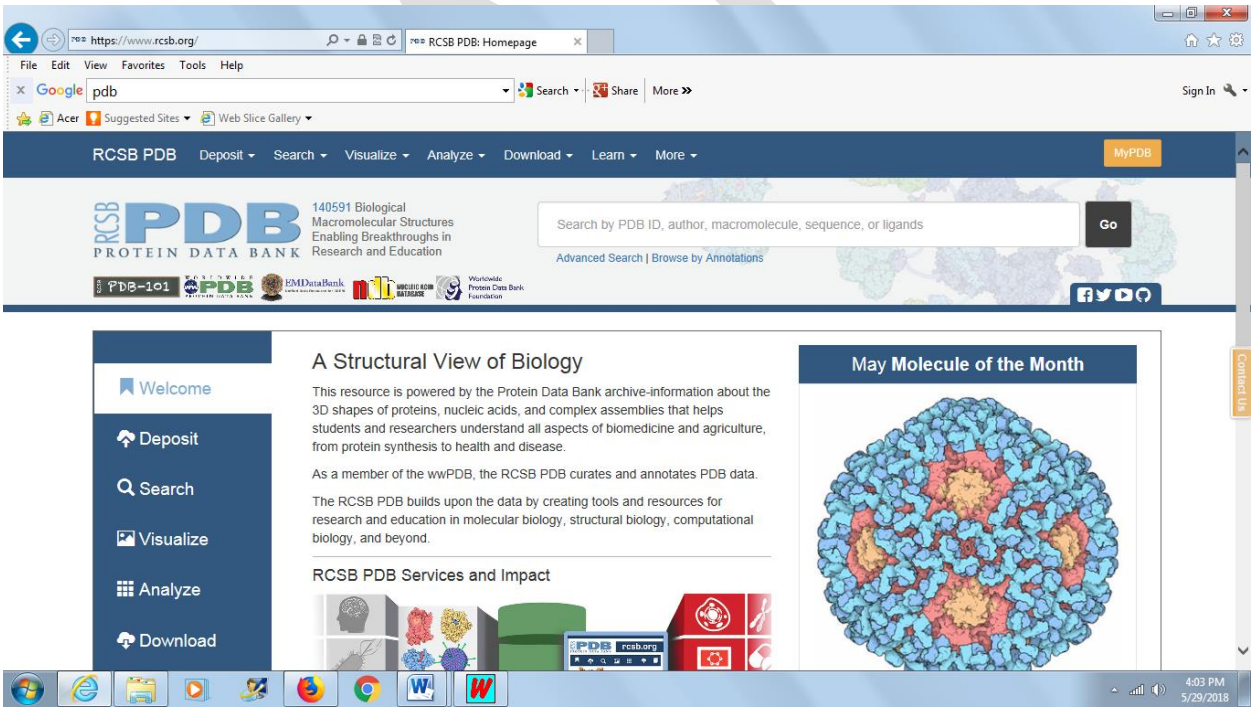
Procedure

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/

The **Protein Data Bank** (**PDB**) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. The PDB is a key resource in areas of structural biology, such as structural genomics.

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A





For instance, the following is the structural information provided in the PDB file of a cardiotoxin from *Naja naja atra*. The PDB ID of the protein is 2CRT.

```
HEADER    CARDIOTOXIN                             12-MAR-94   2CRT
TITLE     CARDIOTOXIN III FROM TAIWAN COBRA (NAJA NAJA ATRA)
TITLE    2 DETERMINATION OF STRUCTURE IN SOLUTION AND COMPARISON WITH
TITLE    3 SHORT NEUROTOXINS
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CARDIOTOXIN III;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: NAJA ATRA;
SOURCE   3 ORGANISM_COMMON: CHINESE COBRA;
SOURCE   4 ORGANISM_TAXID: 8656
KEYWDS    CARDIOTOXIN
EXPDTA    SOLUTION NMR
AUTHOR    R.BHASKARAN,C.C.HUANG,K.D.CHANG,C.YU
REVDAT   3   24-FEB-09 2CRT    1       VERSN
REVDAT   2   01-APR-03 2CRT    1       JRNL
REVDAT   1   01-NOV-94 2CRT    0
JRNL        AUTH   R.BHASKARAN,C.C.HUANG,D.K.CHANG,C.YU
JRNL        TITL   CARDIOTOXIN III FROM THE TAIWAN COBRA (NAJA NAJA
JRNL        TITL 2 ATRA). DETERMINATION OF STRUCTURE IN SOLUTION AND
JRNL        TITL 3 COMPARISON WITH SHORT NEUROTOXINS.
JRNL        REF    J.MOL.BIOL.                   V. 235  1291 1994
JRNL        REFN                   ISSN 0022-2836
JRNL        PMID   8308891
JRNL        DOI    10.1006/JMBI.1994.1082
REMARK   1
REMARK   2
REMARK   2 RESOLUTION. NOT APPLICABLE.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : X-PLOR
REMARK   3   AUTHORS     : BRUNGER
REMARK   3
REMARK   3  OTHER REFINEMENT REMARKS: NULL
REMARK   4
REMARK   4 2CRT COMPLIES WITH FORMAT V. 3.15, 01-DEC-08
REMARK 100
REMARK 100 THIS ENTRY HAS BEEN PROCESSED BY BNL.
REMARK 210
REMARK 210 EXPERIMENTAL DETAILS
REMARK 210  EXPERIMENT TYPE                : NMR
REMARK 210  TEMPERATURE           (KELVIN) : NULL
```

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

The following is atomic coordinates of the protein. It provides information pertaining to atom types, atom nomenclature, atom number, atom position and atom occupancy.

```
ATOM      1  N   LEU A   1     12.825   6.867   0.006  1.00  1.76
N
ATOM      2  CA  LEU A   1      12.032   6.006   0.919  1.00  0.90
C
ATOM      3  C   LEU A   1      11.849   4.631   0.268  1.00  0.73
C
ATOM      4  O   LEU A   1      12.036   4.471  -0.924  1.00  0.90
O
ATOM      5  CB  LEU A   1      10.658   6.639   1.165  1.00  1.16
C
ATOM      6  CG  LEU A   1      10.813   8.094   1.633  1.00  1.61
C
ATOM      7  CD1 LEU A   1       9.429   8.662   1.957  1.00  2.23
C
ATOM      8  CD2 LEU A   1      11.676   8.150   2.900  1.00  2.57
C
ATOM      9  H1  LEU A   1      13.292   6.273  -0.707  1.00  2.15
H
ATOM     10  H2  LEU A   1      12.194   7.547  -0.470  1.00  2.36
H
ATOM     11  H3  LEU A   1      13.547   7.380   0.551  1.00  2.21
H
```

Results:

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

4. Using various BLAST and interpretation of results

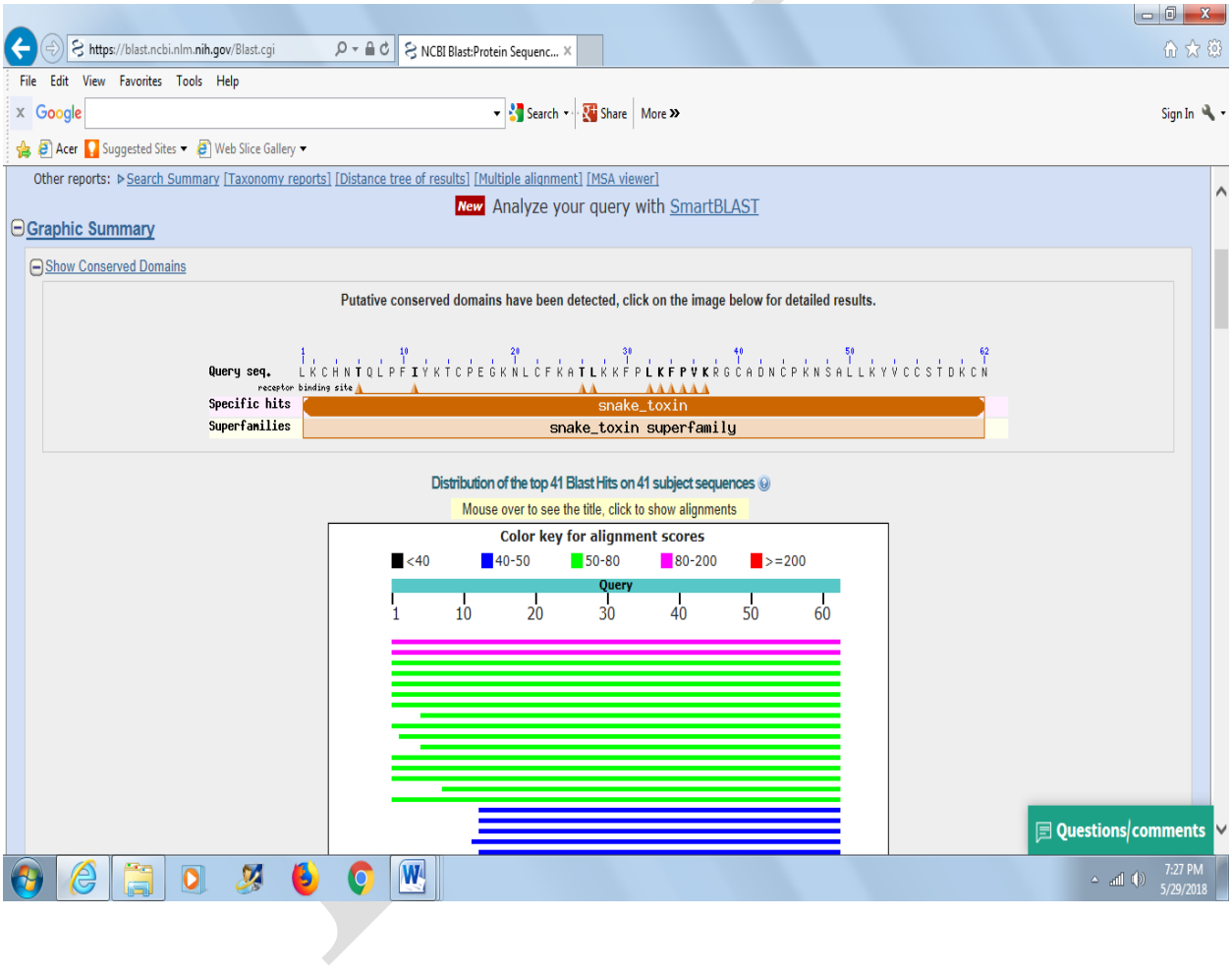
Aim: To retrieve similar sequences for a given protein/nucleotide primary structure using BLAST and rationalizing the outcomes

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.

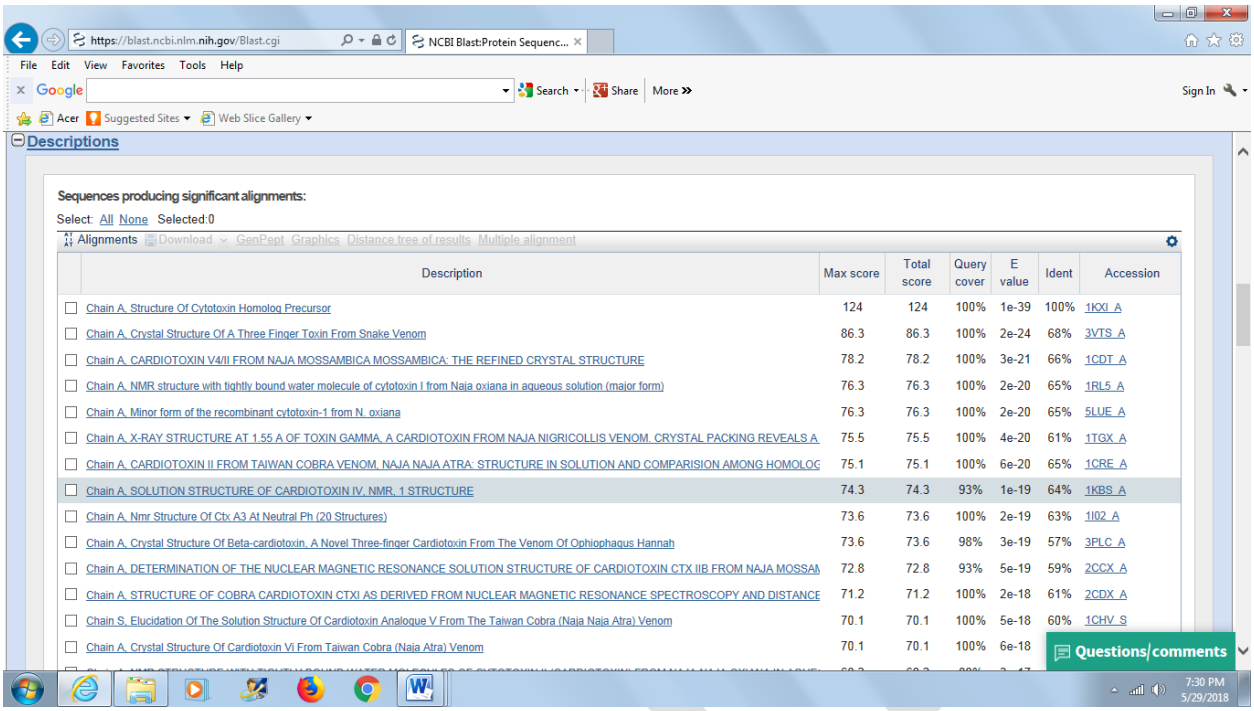
CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

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.



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.

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A



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%	1I02_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 MOSSAM	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.

Results:

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

5. Retrieval of information from nucleotide databases

Aim: To retrieve nucleotide sequences and their annotations from various nucleotide databases

Procedure

The nucleotide sequences under interest can be retrieved from any one of the suitable databases such as Gene, UniGene, EST, Nucleotide and Species-specific Genome repository. Using name of the protein, a general search can be carried out at the beginning and then species-specific sequences can be obtained in a systematic search options. Following are the data of myoglobin (*Homo sapiens*) as taken from the 'Nuclotide (NCBI database)'.

Homo sapiens myoglobin (MB) gene, complete cds

GenBank: AH002877.2

```
LOCUS       AH002877                6889 bp    DNA     linear   PRI 01-AUG-
2016
DEFINITION  Homo sapiens myoglobin (MB) gene, complete cds.
ACCESSION   AH002877 M10090 M14602 M14603
VERSION     AH002877.2
KEYWORDS    Fok family repetitive sequence; direct repeat; myoglobin; repeat
            region; tandem repeat.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 6889)
  AUTHORS   Akaboshi,E.
  TITLE     Cloning of the human myoglobin gene
  JOURNAL   Gene 33 (3), 241-249 (1985)
   PUBMED   2989088
COMMENT     On or before Aug 1, 2016 this sequence version replaced M10090.1,
            M14602.1, M14603.1, AH002877.1.
FEATURES             Location/Qualifiers


     source          1..6889
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="22q11.2-qter"
```

```
     repeat_region   1100..1750
                     /note="Fok family repetitive element"

     prim_transcript 1895..>2552
                     /note="myoglobin mRNA and intron"

     gene            1965..5924
                     /gene="MB"

     CDS             join(1965..2059,4067..4289,5778..5924)
                     /gene="MB"
                     /note="myoglobin"
                     /codon_start=1
                     /protein_id="AAA59595.1"

/translation="MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKF

DKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPV
                     KYLEFISECIIQVLQSKHPGDFGADAEGAMNKALELFRKDMASNYKELGFQG"

     exon            <1965..2059
                     /gene="MB"
                     /note="myoglobin; G00-119-378"
                     /number=1

     intron          2060..>2552
                     /gene="MB"
                     /note="myoglobin intron A"
     gap             2553..2652
                     /estimated_length=unknown

     prim_transcript <2653..>5323
                     /gene="MB"
                     /note="myoglobin mRNA"

     intron          <2653..4066
                     /gene="MB"
                     /note="myoglobin intron A"

     repeat_region   <2653..2873
                     /gene="MB"
                     /note="Alu family repetitive element"

     repeat_region   3246..3254
                     /gene="MB"
                     /note="9 bp direct repeat copy 1"

     repeat_region   3255..3386
                     /gene="MB"
                     /note="33 bp tandem repeat (4 copies)"
```

```
     repeat_region   3387..3395
                     /gene="MB"
                     /note="9 bp direct repeat copy 2"

     exon            4067..4289
                     /gene="MB"
                     /note="G00-119-378"
                     /number=2

     intron          4290..>5323
                     /gene="MB"
                     /note="myoglobin intron B"
     gap             5324..5423
                     /estimated_length=unknown

     prim_transcript <5424..6336
                     /note="myoglobin mRNA"

     intron          <5424..5777
                     /gene="MB"
                     /note="myoglobin intron B"

     exon            5778..>5924
                     /gene="MB"
                     /note="myoglobin; G00-119-378"
                     /number=3
ORIGIN      2 bp upstream of RsaI site.
        1 gtactgtatt ttcattcctc ttagttatct ccctaaaaag actctgagtt ccttgaacac
       61 aggaaggtgt tttatttgat tttgttatcc tcagcatgta gcagtgtctg acacacagta
      121 ggtgctctat cactgtgaga gggatggatg gatgggtgga gttacagatg gatagaagga
      181 tagatggagg gatgggtgga tgatggatgg atagatggat ggaggggtgga tgatgaatgg
      241 agggataatg agtggatgaa tgagggaatg ggtggatgga tggatggagg gatggaggaa
      301 cagatagata gatggaggga tgggtgggtg atggatggat agatggatgg agggagggat
      361 gatgaatgga gggataatga atggatgaat gaggggatgg gtggatggat gaatggaggg
      421 atgatgggtg gatgaatgaa ttgagggatg gatggatgaa cacatggatg gatggataga
      481 tggatagatg gaggaactgg tggattttgg atggatgggt ggatggatag atgaatgaat
      541 gcctggatag acaaagagat gatggataga tgaatagatg aattaaggga tgtcggatag
      601 atggagggat tgatagatgt tggatggatg ggtggtggat ggatagatga gtgaatgcat
      661 ggatagacaa agagatgatg gatggatgaa ttaagggatg acagatggat ggatggatga
      721 gtaactggat ggacaagtgg ataaatggat agatggttga atacctgaat ggattgaagg
      781 aggatgcatg gatgtaagat aaggctaatc atcctccact ctctttcttt gcaaaaccat
      841 ccacccattt actcaataaa catttattca gttcaaactt ggcacaaagc accatgtgag
      901 gcccaagaga tacgtgggtt aataaaacag agctcctgcc ctcctgaaaa ctgcaaagaa
      961 aggggcgtgg cttcctgagt tcaaatccca actctgccag cgactagctg tacatcagtg
     1021 atgtttccct actttctctc aattaaatag ggataatgtc agtacctatc acattgggag
     1081 gtcttgcggg gattaaatga gttaccaaat gccaagtgtt tgggacaggg cctggcaccc
     1141 agcaaagtct cttgtgagtg ctggctgcta ttatcctaat ggagaagatg gcatgaaaac
     1201 caggaaatag gatgcccttt gggaagcaat gcaacaggaa cttacacaaa gaaaggaaag
     1261 gaggaagcaa ttagtggtgt ctcaaaggag tatgtcaaga aaaacttttc agagggaaac
     1321 ctttgagcag ggtcatgaaa acaggagttc tctaagagat tgtggacttg cctgggacca
     1381 cctggctata agcacaaaac catccggttc ctttctgtca cttctggcgg gtgaggggtc
```

CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

```
1441 tctggcaaag gggcagaagg tgcgtgagag gttgcgaatg gccaggactg tcctggggcc
1501 agccggggca cctggtggcc aagcttagaa acatgacagg tcctcttggg agggctgacc
1561 gcagggagcg ttgggtttca ggctgctggc gtcggcttct gtggtgccct ttctgtcggc
1621 tatgagagtc cagacagtgc ccaacctcct ccccttcttt ccacacgcac aaccacccca
1681 ccccctgtgg cctgagctgt cctgcctcgc cacaatggca cctgccctaa aatagcttcc
1741 catgtgaggg ctagagaaag gaaaagatta gaccctccct ggatgagaga gagaaagtga
1801 aggagggcag gggaggggga cagcgagcca ttgagcgatc tttgtcaagc atcccagaag
1861 gtataaaaac gcccttggga ccaggcagcc tcaaacccca gctgttgggg ccaggacacc
1921 cagtgagccc atacttgctc tttttgtctt cttcagactg cgccatgggg ctcagcgacg
1981 gggaatggca gttggtgctg aacgtctggg ggaaggtgga ggctgacatc ccaggccatg
2041 ggcaggaagt cctcatcagg taaaaggaag agattccatt gcccctgcca cccacaccct
2101 aagatcaagg gtgttcagct gcaaggtgga aagtttgcac gtggggtagg tcagttggct
2161 gcattagtta agggtgttag aacggtcact tgctttttct ttgcttttaa gtgtcaggga
2221 ttggactcag gagagggaaa ggagccattt caggctgatg tcagcagctg gaggaagcat
2281 gagaatcaaa cctaggatgc tcagagtcca ccaggaagaa ttttagaatt atagacagtc
2341 agagttaaca agggtcctga gagattttgt acagccacct ctcttacagg atgaggacaa
2401 aaagcgactg agaaggggag gacatttcca gagtcacagc tcattaaatg ctcttaaagt
2461 gtcaaggtta agacatgctc ttcaagggga gacagatctg gttctagact tggctctgcc
2521 actgagccac tgggtgacct ttgggaaggt ac
     [gap 100 bp]
2653              gaattctg agaattgctc aaacccagga ggtggaggtt gcggtgagca
2701 gagattgcac cactgcactc cagcctgggc aacagagcca gattccatct caaaaaagaa
2761 aaacaaaaaa caaaaaagcc atgaactcat tttcaggttg aggagctcag catcctggtt
2821 gtgaaatacc ctcctcataa aaccctggga tggagactac ggggatcagg tgcttccttg
2881 tgacaacttc tgggcatggt ggctcagggc gcaaactgga gtgtggccac aatacatact
2941 gtgtactttt acaaggatgt cacagagcct gggtatcata aaagaggagc ttttcaagga
3001 actgaaacca ttagacagga gagagagccc tgggcagaca gggttgcccg tgccaaacat
3061 ttcagctgtg gcacaaggga aagggtggga gttatgaaac tgttccattt tgggtttagg
3121 tctgggctct gccgctagct agccaagtga ccttggccac ttatctctgt ggtcttccat
3181 gagtaaaagg cggaaactca ctcctaccca gagggcaggt ctgactccct ttaaccagca
3241 cccacctgct cacagcagga aggactgagg tctaaagctg gaggtgggca ggaaggactg
3301 aggtctaaag ctggaggtgg gcaggaagga ccgaggtcta aagctggagg tgggcaggaa
3361 ggaccgaggt ctaaagctgg aggtggctgc tcagagtccc agcagaggcc tctggggcac
3421 ctcactgagt gcctggcagg agtgggtgcc tgtctcaggg ctgggttgag ttgctcccac
3481 caggaccctt cgtcatctgc acagtgaggg gactgggagg ttcagagagt cacagcttgg
3541 gctcaaaaca agcaagaggt ttctgagtgt gaggattgct ctggagtgga atggccctca
3601 caggtaggag tgagcctcct gtagctagag gtatttaagc agctgaagga caatccctgg
3661 gcaggaagct gcagagatgg tcgcagcgtg gactagaact gctgttttgg tcactcagac
3721 ctcattccag cctggcttct ctggacagca cccctgcaat agtgagctgg tgactttacg
3781 cctcagaacc tcggtttcta catctgtaaa atgggaatta tatgacactc actatgtgcc
3841 agacaccctg ttggtacata gcacacacta tctcacttaa tccttcaagt agggacaagt
3901 tatccccatc ccttatatga ggaagctgag gcacagagag gtgaagtgaa tggcccaagg
3961 tcacacagct gggaagacag ggagctaaac ttgaactcta gtctggctgc ccccagacct
4021 cacaccgcac ctcccatgcc gactccagcc ttccctgtgc ccacaggctc tttaagggtc
4081 acccagagac tctggagaag tttgacaagt tcaagcacct gaagtcagag gacgagatga
4141 aggcatctga ggacttaaag aagcatggtg ccactgtgct caccgccctg ggtggcatcc
4201 ttaagaagaa ggggcatcat gaggcagaga ttaagcccct ggcacagtcg catgccacca
4261 agcacaagat ccccgtgaag tacctggagg taggaggcag agcctgggca ggtgggagga
4321 tgcggggaag gcctcgggtg gggcaatggg atctgggttc gagtccaagc tcagccacta
4381 acttgtggga tgacctatgc cactcttctc tgtgccccag gtttctcatt tgtaaagggg
4441 actgccaccc actttgcctt cctcctggga ttgttgagaa tgaacacatt tagcattttt
```

```
     4501 aatttagtat gccaaattca catcttatta ccaaagagga aagggagagg ggatattggg
     4561 tgcaaaattt gcatcctctc catgggtagg taccattatc atatccactt gatagatggg
     4621 gaaactgagg ctcacagagg ttaagcagct tgtccacggt cacaggaggt ggataatggc
     4681 agagccaaga ttcaaacgca ggtctctatt actacagaac cccagcccct aactgctgtg
     4741 ccactgggag tctggtacat gcaggactta tgtggcagga gctcagcaag tggggctcaa
     4801 tttggggtgg gggtgaccag caggctggct ctattggttc cagcatcttc acagatgaag
     4861 agacaggacc tcggtttcca gcacaagcaa ttggtttgaa cctcctgaga tgggttggaa
     4921 agttgggtgg atcagggttg ggggcaggag cctgggcttc aggttgtgtg tctataactg
     4981 gtgggaggag gcgatttggg gagaggaggg agctggggat gaaggaccac agggacaggt
     5041 gcatcccccg agggtagaaa cagcaggaag tctggtgcag ccatgaggat taggatgtgg
     5101 tgatagctac ccgctgggat gggccacagt gagcatttgc tgccatgcct agcacatgca
     5161 tccatcctca aagttgcctc atggccaaaa tgactgcaag agctccagcc agctcttcta
     5221 tattcccaac tggaagcagg agaaagagag gaatgctctc ttttgaggag tttaaggagt
     5281 cccagaaatc tcatccaaca attttattta catctcattg gcc
          [gap 100 bp]
     5424                            gagctct cacctggttt cagtggggtc tacatcctga
     5461 tggagtggag ggggctgtga gtaagagcgt gggctccgga gccggccctc ctgggtccaa
     5521 atgtcccttc cattcaacct ccccctcgcct cagtttctgc atctgtaaat cgagggcagt
     5581 tgtagtatct atctcacagt ggttgtgggg atcaaagggg ttcatccgtg gagatcacac
     5641 agactctcac ctggtgccta gcaagtgctc aatacacggt cctggaataa agagaaggta
     5701 ggaggacaac tgactcccat ctggcccctg gcttgtccca ccctggtgac cattttctct
     5761 cctcaccctc cctgcagttc atctcggaat gcatcatcca ggttctgcag agcaagcatc
     5821 ccggggactt tggtgctgat gccgaggggg ccatgaacaa ggccctggag ctgttccgga
     5881 aggacatggc ctccaactac aaggagctgg gcttccaggg ctaggccccct gccgctccca
     5941 cccccaccca tctgggcccc gggttcaaga gagagcgggg tctgatctcg tgtagccata
     6001 tagagtttgc ttctgagtgt ctgctttgtt tagtagaggt gggcaggagg agctgagggg
     6061 ctggggctgg ggtgttgaag ttggctttgc atgcccagcg atgcgcctcc ctgtgggatg
     6121 tcatcaccct gggaaccggg agtggccctt ggctcactgt gttctgcatg gtttggatct
     6181 gaattaattg tcctttcttc taaatcccaa ccgaacttct tccaacctcc aaactggctg
     6241 taaccccaaa tccaagccat taactacacc tgacagtagc aattgtctga ttaatcactg
     6301 gccccttgaa gacagcagaa tgtccctttg caatgaggag gagatctggg ctgggcgggc
     6361 cagctgggga agcatttgac tatctggaac ttgtgtgtgc ctcctcaggt atggcagtga
     6421 ctcacctggt tttaataaaa caacctgcaa catctcagtt tctgcctggc atttttcatc
     6481 tcctagagta aatgatgccc ccaccagcac cagcatcaag gaagaaatgg gaggaaggca
     6541 gaccctgggc ttgtgtgtgc agagagcctc aggaaagagg agaaggggag gaggaaaggc
     6601 aggagggtga gagggacagg agcccaccct ccctgggcca ccgctcagag gcaggcccag
     6661 tgcagggcat ggggaaatgg aagggacagg cttggcccca gccttgggag caccttctct
     6721 tcgggggagg tgggaggcag cgaacagacc tctgcaatac gaggagagag tgacaggtgc
     6781 gccaggctgt gggaacccag aggagagggg aagccatcat catcatggct gcaatacctt
     6841 cagtaacgtg ggaaggtcac cctgctagta agtggcagag ctgggactc
//
```

Similarly, the nuclotide sequences from other databases can also be analyzed in a forthright manner.

Results:

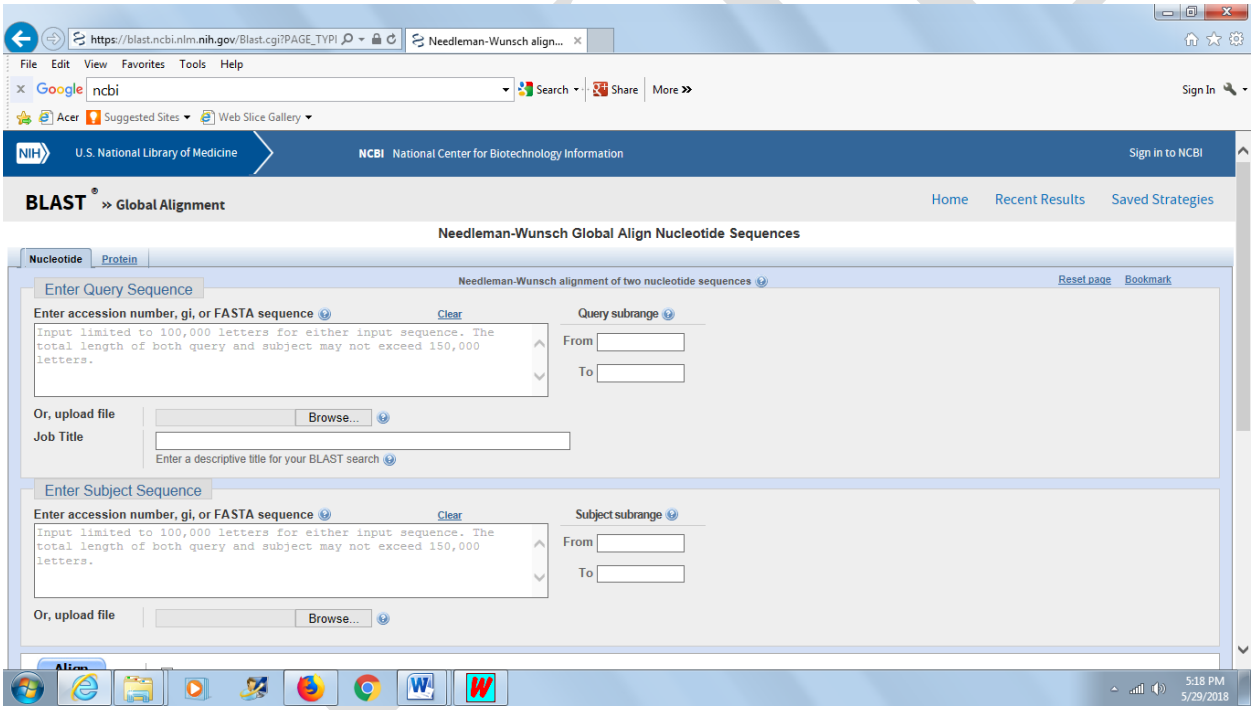
CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

6. Sequence alignment using BLAST

Aim: To examine pair-wise alignments of protein sequences using BLAST

Procedure

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.

Seq2
Sequence ID: Query_206921Length: 62Number of Matches: 1
Related Information
Range 1: 1 to 62Graphics Next Match Previous Match First Match

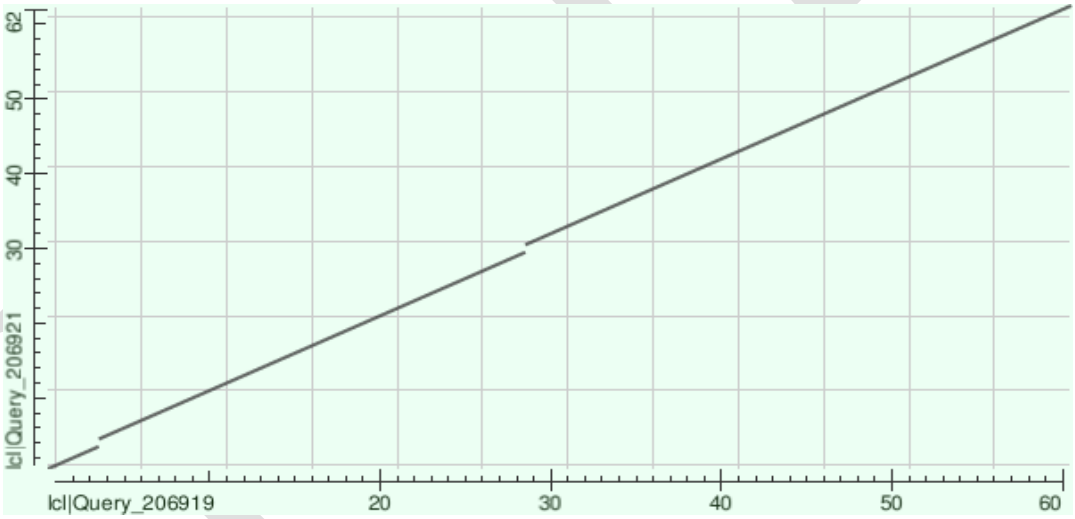
Alignment statistics for match #1

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

Features:

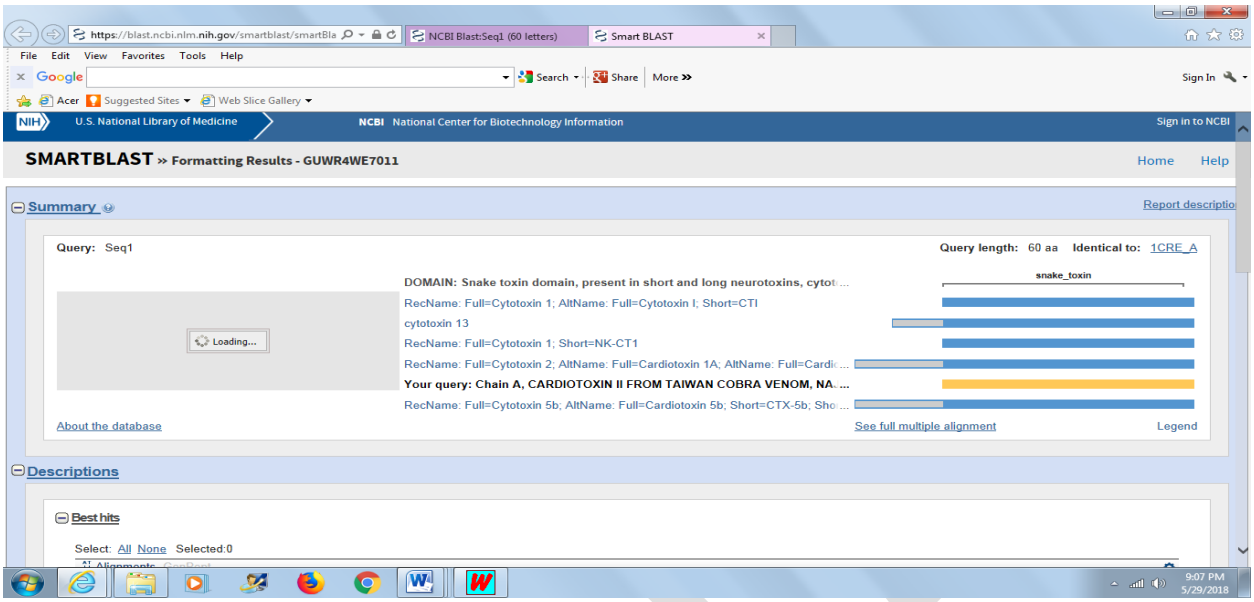
```
Query  1   LKC-NKLVPLFYKTCPAGKNLCYKMFMVS-NLTVPVKRGCIDVCPKNSALVKYVCCNTDR  58
           LKC N  +P  YKTCP GKNLC+K  +    L  PVKRGC D CPKNSAL+KYVCC+TD+
Sbjct  1   LKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRGCADNCPKNSALLKYVCCSTDK  60

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

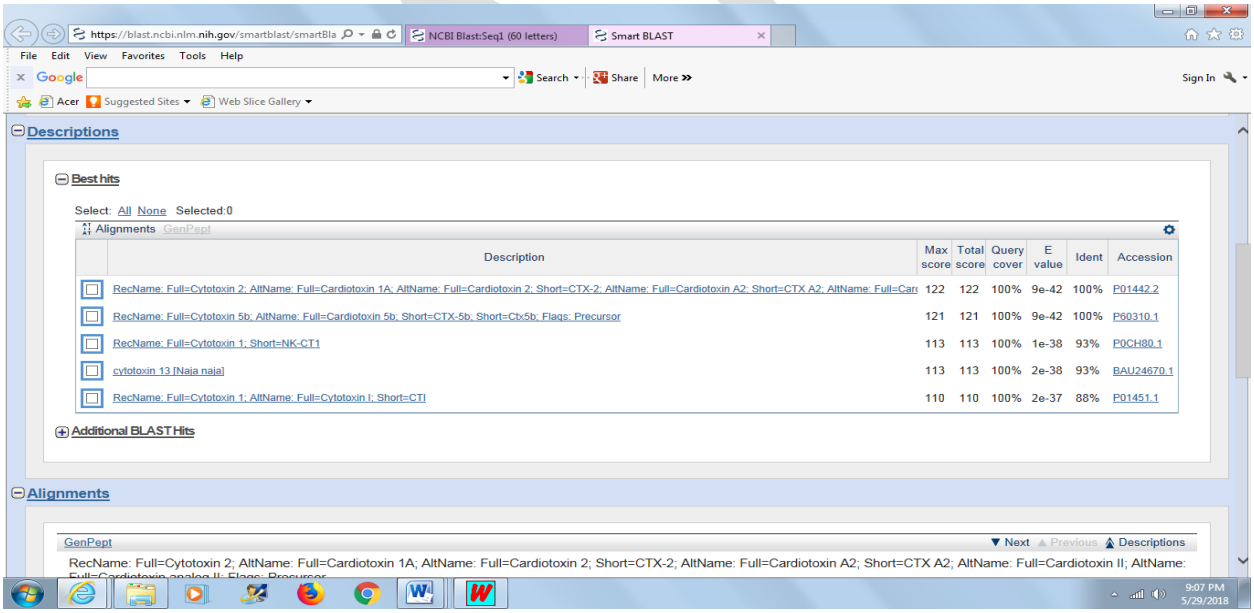
The dot plot for the sequences is depicted in the figure shown herein.



CLASS : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

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





Results:

7. Multiple sequence alignment using ClustalW/Clustal Omega

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

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 : III B. Sc., BT
COURSE NAME : BIOINFORMATICS PRACTICAL
BATCH : 2016 – 2019
COURSE CODE : 16BTU512A

>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
                * * ***.* *:**:**.**:**
```

Results: