MODY UNIVERSITY
OF SCIENCE AND TECHNOLOGY
LAKSHMANGARH · RAJASTHAN
A Leading Women's University

MODY UNIVERSITY
OF SCIENCE AND TECHNOLOGY
(Established u/s 2(f) of the UGC Act, 1956)
Lakshmangarh (Sikar), Rajasthan 332 311, India
Ph. No. : (01573) 225001-12 · Fax : (01573) 225041
email : registrar@modyuniversity.ac.in · www.modyuniversity.ac.in

Annexure_3.4.7

Criterion III – Research, Innovations and Extension

Key Indicators - 3.4 Research Publications and Awards

Metric No. 3.4.7_ E-content is developed by teachers

NPTEL

swayam

CERTIFICATE

OF APPRECIATION

This Certificate is presented to

Vikrant Sharma

for translating the files of the course

Engineering Mechanics (112106286)

offered by **IIT Madras**

in **Hindi** *language. Hours translated:* 5

Date : 6/10/2020

Andrew Thangaraj
Coordinator, NPTEL

28/12/2024

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

NPTEL

swayam

CERTIFICATE

OF APPRECIATION

This Certificate is presented to

Satyajit Anand

for translating the files of the course

Microwave Integrated Circuits (117101119)

offered by **IIT Madras**

in **Hindi** *language. Hours translated:* **2.75**

Date : **19/10/2020**

Andrew Thangaraj
Coordinator, NPTEL

28/12/2021

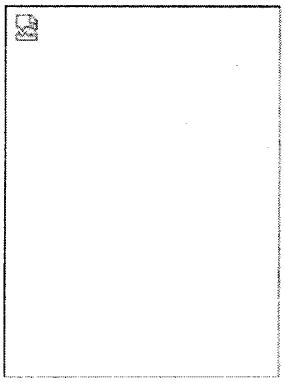
Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Inviting Expression of Interest for Developing Non-Technology PG MOOCs for SWAYAM

University Grants Commission

Applicant's Name: Naveen Dhingra

Course Coordinator Details:



Name: Naveen Dhingra	
E-mail: naveendhingra.slas@modyuniversity.ac.in	**Phone No:** 9752521903
Nature of Employment: Permanent	**Designation:** Assistant Professor
Working Address: Department of Bioscience, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan ,City: Sikar , State: Rajasthan	

Teaching Experience (in year): 6	**Research Experience (in year):** 6

Experience in developing e-Content/MOOCs: Yes
No. of Hours of developing e-Content/MOOCs: 12
Details of Project of developing e-Content/MOOCs: I have developed course on Plant Tissue culture and Bioinformatics MOOCs in previous organization.

Institution Details:

Affiliated Institute: Mody Institute of Technology & Science, Sikar

Course Details:

28/12/2021

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Course Name: Applications of Computers in Chemistry		
Subject Name: Chemistry		
Medium of Instruction: UG	**Category (Applicant):** Private Institute	**Nature of Course:** Core
Type of Course: Credit	**Course Credit:** 1	**Course Duration(week):** 2
Course Objective: 1. To train manpower suited to the needs of Bioinformatics Industry and Research centers. 2. To impart advanced training in Bioinformatics to develop entrepreneurship skills.		
Course Curriculum: https://swayam.inflibnet.ac.in/asset/files_d/chemistry/12.pdf (https://swayam.inflibnet.ac.in/asset/files_d/chemistry/12.pdf)		
Pre-requisites: 1. Higher School Biology 2. Higher School Chemistry 3. Higher School Mathematics		
Introductory Video URL: https://youtu.be/boqpd7Zw-64		

Evaluation Methodology:

Assignment: 10%	**Practical:** 20%	**Final Exam:** 60%	**Others:** 10%

Team (Experts) Details:

Name	Email	Phone No	Affiliation	Designation	Subject Expert
Naveen Dhingra	naveendhingra.slas@modyuniversity.ac.in				

28/12/2021

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Rajeev Singh	rkshing.slass@modyuniversity.ac.in	7297809386	Department of Journalism and Mass Communication, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan, India	Assistant Professor	Electronic Media/Audio and Video Production

28/10/2021

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Sooraj Jangir	soorajjangir.en@modyuniversity.ac.in	8078651292	Department of Journalism and Mass Communication, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan, India	Digital Marketing Executive	Graphic Designer/Audio and Video Production
Dr. Anukriti	anukriti.slas@modyuniversity.ac.in	7060120833	Department of Bioscience, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan, India	Assistant Professor	Bioinformatics

28/12/20

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Dr. Naveen Dhingra	naveendhingra.slas@modyuniversity.ac.in	9752521903	Department of Bioscience, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan, India	Assistant Professor	Drug Designing, Bioinformatics
Dr. Naveen Dhingra	naveendhingra.slas@modyuniversity.ac.in	9752521903	Department of Bioscience, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan, India	Assistant Professor	Drug Designing, Bioinformatics

28/10/2021

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Dr. Anukriti	anukriti.slas@modyuniversity.ac.in	706012083	Department of Bioscience, School of Liberal Arts and Science, Mody University of Science and Technology, Laxmangarh (Sikar)-332311, Rajasthan, India	Assistant Professor	Bioinformatics

Course Schedule:

Week No	Module to be covered	Nature of work	Remarks
1	Introduction of Bioinformatics, Introduction to Database , Sequence Databases, Structure Databases	Video, Text, Discussion forum, Assessment, Assignment, Quiz,	
2	Structure Databases, Sequence Alignment, Phylogenetics Analysis, Molecular Docking	Video, Text, Discussion forum, Assessment, Assignment, Quiz,	

28/12/2021

Registrar
Mody University of Science and Technology
Lakshmangarh-332311 Dist.-Sikar (Rajasthan)

Mody University of Science and Technology

Department of Biosciences

MOOCs Module

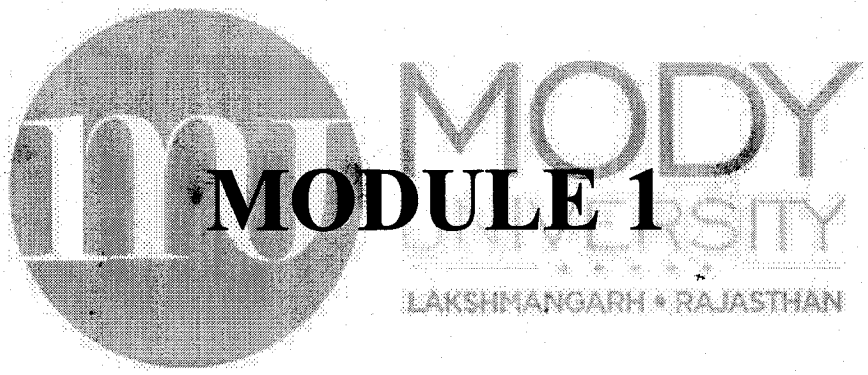
Basic Bioinformatics

Dr. Naveen Dhingra, Dr. Anukriti

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

MODULE 1



28/12/2021

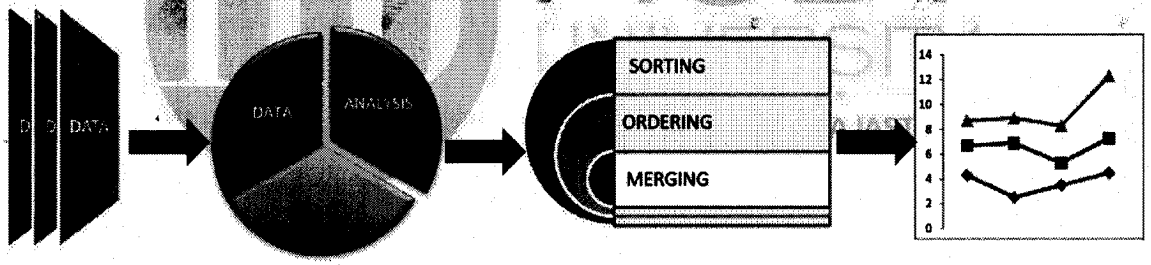
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

1: Introduction to Bioinformatics

1.1: What Is Bioinformatics?

Bioinformatics is an interdisciplinary field harnessing fields like computer sciences, mathematics, physics and biology. It is convergence of biotechnology and information technology. It is a stream of science which arose for the utilitarian purpose of introducing order into the massive data sets produced by the help of new and advanced technologies of molecular biology.

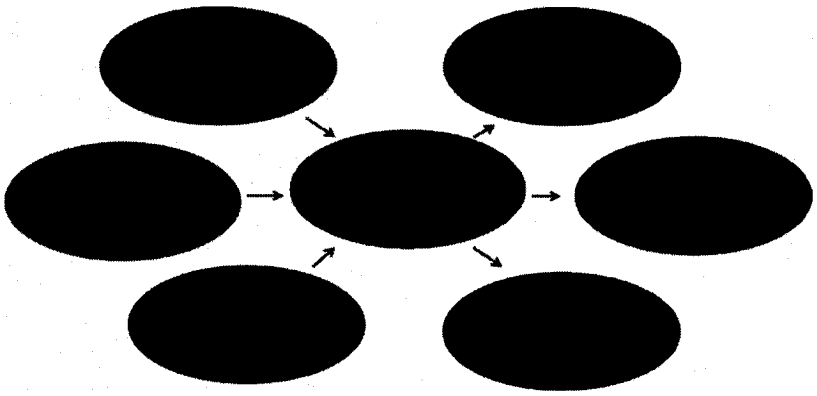
Higgs and Attwood gave two definitions of bioinformatics. "Bioinformatics is the development of computational methods for studying the structure, function, and evolution of genes, proteins and whole genomes" and "bioinformatics is the development of methods for the management and analysis of biological information arising from genomics and high throughput experiments". It deals with collection, organization, data analysis, manipulation, presentation and sharing of biological data.



It requires advanced knowledge of computer science, mathematics and statistical methods to understand various biological functions at the molecular level. It covers specialized areas of biology like functional and structural genomics, DNA microarray, comparative genomics etc. It is a cutting-edge branch of biotechnology that paves its way to novel drug discovery and personalized medical treatments.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



1.2: Origin of Bioinformatics

1960s had witnessed the first efforts towards bioinformatics. Paulien Hogeweg and Ben Hesper first coined the term bioinformatics in the beginning of 1970s, defining it as "the study of informatic processes in biotic systems". Margaret Dayhoff in 1965 developed first protein sequence database called Atlas of Protein Sequence and Structure.

The Brookhaven National Laboratory in early 1970s, established the Protein Data Bank (PDB) for archiving 3D protein structures. Needleman and Wunsch developed the first sequence alignment algorithm in 1970s. Chou and Fasman in 1974 developed the first protein structure prediction algorithm.

In 1980s, GENBANK got established and William Pearson developed a fast database searching algorithm FASTA and Stephen Altschul and coworkers developed BLAST. A major developmental boost to bioinformatics came in late 1980s by the start of Human Genome Project.

1.3: Why We Should Study Bioinformatics?

- For better understanding of living cells at molecular levels.

  All the living cells are composed of various molecules that function in perfect coordination with each other for the proper functioning of the cell. With the help of

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

bioinformatics, we can try to understand a cell as a system that is composed of various molecules working together to perform any function within the cell.

- To get better understanding of evolution.

  Phylogenetic analyses are the processes that are used for the reconstruction of evolutionary relationship among a group of related molecules or organisms. This helps to predict certain features of molecules whose functions are not known, to track the flow of the genetic information, and also to determine genetic relatedness. With the help of all these, one can try to understand the process of evolution.

- For gene predictions and genome annotation.

  Gene prediction, also known as gene finding basically means locating the genes along a genome by identifying the regions of genomic DNA that encode genes using various tools offered by bioinformatics.

  Genome annotation process helps in the identification of functional elements along the sequence of a genome. It is a process of analyzing and interpreting the raw DNA sequences so that significant biological information can be extracted out from it with the use of certain tools and softwares.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

- To study the impact of change in climate on various organisms.

  With the help of bioinformatics, we can try to predict the changes in the climatic conditions and when we combine bioinformatics with genomics, we can try to predict the probable changes in genetic makeup of various organisms with the change in the climatic conditions.

- For the development of genome libraries and databases.

  Bioinformatics plays an important role in managing and analyzing the huge biological data. It helps in the development of databases and libraries from which biological information can be extracted out easily.

- For the development of molecular and personalized medicines and vaccines.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

There are thousands of diseases that are hereditary in nature. Every individual also responds differently towards the environmental changes. Thus, trying to pin point the exact gene or molecule that is involved in the occurrence of any disease can help in designing molecular and personalized medicines. This can be achieved with the help of bioinformatics.

1.4: Uses of Bioinformatics

- Analysis of gene variation and expression.
- Analysis and prediction of gene function.
- Designing primers to predicting the function of gene products.
- Analysis and prediction of protein structure and function.
- Prediction and detection of gene regulation networks.
- Simulation of environments for whole cell modeling.
- With the help of sophisticated computational tools, mining the genomic information.
- Knowledge based drug designing.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

1.5: Scope of Bioinformatics

- Development of computational tools and databases.
- Development of algorithms.
- Data analysis and interpretation.
- In structural analysis of proteins and nucleic acids.
- Managing big data.
- Drug discovery and vaccine designing.
- In functional analysis of proteins.

23/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References:

1) Ardeshir Bayat, Science, medicine, and the future- Bioinformatics, BMJ. 2002 Apr 27; 324(7344): 1018–1022.

2) Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press

3) Lesk, A. M. (2002). Introduction to bioinformatics. Oxford: Oxford University Press.

4) Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer

5) Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol 7(3): e1002021. doi:10.1371/journal.pcbi.100202

6) Wang, Z.; Chen, Y.; Li, Y. (2004) A Brief Review of Computational Gene Prediction Methods; Genomics Proteomics Bioinformatics. 2004 Nov; 2(4): 216–221.

LAKSHMANGARH • RAJASTHAN

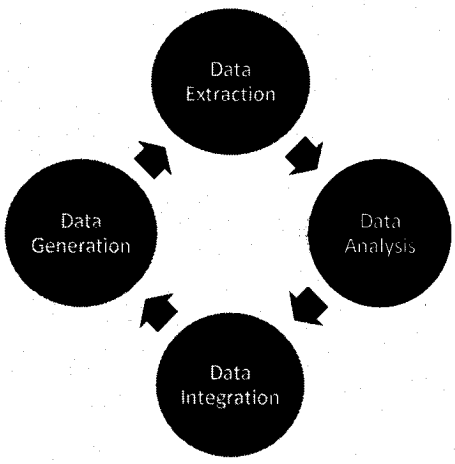
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2: Introduction to databases

With the Technological advances in Biomedical Research like High -throughput Techniques and Lower-Cost DNA sequencing Technologies huge amount of Heterogeneous Data is getting generated daily. To manage this heterogeneous data we need better systems or tools. We also need better systems so that we can efficiently extract our data without distorting the information. For this purpose databases have been developed that can store the biological data in huge volumes and in a proper well defined manner for easy and efficient extraction of information.

Database helps in Extraction of data, its analysis, integration and generation of information or new data in a Fast and Efficient Manner.



2.1: What Is a Database?

Databases are collection of organized information that can be accessed managed and retrieved easily. These are composed of computer hardware and software for the management of data. With the help of a database, the information can be accessed, managed and retrieved easily and efficiently via varieties of search criteria. Sophisticated computational methodologies are being developed daily which help in easy and efficient extraction and analysis of biological data. Databases have evolved dramatically since their inception in the early 1960's. A simple database may contain records in the form of normal files

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2.2: What Are Biological Databases?

A biological database is a large collection of organized biological data (DNA, RNA, gene etc.) which can easily be searched with the help of keywords. The data present in the biological databases can be easily retrieved for the public use. Database organizes the data in the form of Records and each record is known as an Entry

The data present in the biological databases is updated periodically and is cross referenced.

2.3: Purpose of Database:

A database serves various purposes like:

- Data Retrieval.
- Knowledge discovery.
  - Finding connections or relations between different pieces of data/information that were not known at the time that information was fed into the database.

In Biological Databases different types of information is stored like

- The Functions of Genes.
- Structure of Proteins.
- Localization of Proteins and Genes (both cellular and chromosomal)
- Clinical effects of mutations in Genes.
- Similarities of Biological Sequences and Structures.

The biological databases store and organize in a structured and searchable manner for humans and systems. They not only store and organize data but also share data with different databases. They also provide web application programming interfaces (APIs) for computers to exchange and integrate data from various database resources in an Automated Manner.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2.4: Types of Databases

We can classify databases on various criteria like the level of curations or the type of data they store.

- According to level of data curation or source, biological databases can be roughly divided into two categories:
  - Primary Databases
  - Secondary Databases

2.4.1: Primary Databases

Primary databases collect data directly from the Authors. They are archives of raw sequence or structural data submitted by the scientific community. Results from the experiments conducted are submitted directly into the database by researchers. The data in the primary database becomes permanent once accession number has been allotted. These databases are archival in nature. These three public databases (GENBANK, EMBL and DDBJ) closely collaborate and exchange new data daily and constitute the International Nucleotide Sequence Database Collaboration.

For the three-dimensional structures of biological macromolecules, there is only one centralized database, the Protein Data Bank. This database archives atomic coordinates of both Proteins and Nucleic Acids, determined by x-ray crystallography and NMR. They contain the amino acid sequence of proteins, their conformation, structure, and features such as active sites.

2.4.2: Secondary Databases

Secondary Databases contains computationally processed or manually curated information, based on original information from Primary Databases. They are also known as curated databases or knowledgebases. SWISS-Prot and Protein Information Resources (PIR) are examples of Secondary Databases. SWISS Prot provides detailed sequence annotation that

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

includes Structure, Function, and Protein Family Assignment. The Sequence Data are mainly derived from the Database of Translated Nucleic Acid Sequences stored in the EMBL Database. Protein Information Resources (PIR) is an integrated public database which supports genomic, proteomic and systems biology research and scientific studies. It helps in *the identification and interpretation of protein sequence information.*

2.5: Classification of Biological Database according to the scope of Data Coverage.

- Biological database is divided in two parts:
  - Comprehensive Database
  - Specialized Database

2.5.1: Comprehensive Database

Comprehensive Database covers different types of data from numerous species. For example GenBank, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) contain Nucleotide Sequences from various Organisms. Uniprot is a comprehensive database that provides high quality information on the protein sequence and also has information related to the functions performed by the proteins. This is also a freely accessible database.

2.5.2: Specialized Database

Specialized Databases contain specific types of data or data from specific organisms. WormBase is a database that provides accurate, current, accessible information concerning the genetics, genomics and biology of *C. elegans* and other related Nematodes. FlyBase is another database that contains information of the genes and genomes of *Drosophila.* HIV sequence database and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

GenBank EST database and Microarray Gene Expression Database at the European Bioinformatics Institute (EBI) are examples of some available gene expression databases.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

MODULE 2

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2.6: Biological Databases can also be classified on the Basis of Nature of Data they contain

2.6.1: Sequence database

A Sequence database is a biological database that has a huge collection of nucleic acid sequences, amino acid sequences, or other polymer sequences. These databases can be nucleic acid sequence databases and protein sequence databases.

2.6.1.1: Nucleic acid sequence databases

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Nucleic acid sequences are stored in three primary sequence databases GENBANK, EMBL and DDBJ. The International Nucleotide Sequence Database Collaboration (INSDC) is an initiative that operates between DDBJ, EMBL-EBI and NCBI. It collects and provides the information on the nucleotide sequences along with the metadata. INSDC contains the data raw reads, through alignments and assemblies to functional annotation, containing contextual information related to samples and experimental configurations.

All the three databases (GENBANK, EMBL and DDBJ) are synchronized daily and data on nucleotide sequences contributed by research teams is shared between the collaborating organizations.

2.6.1.1.1: RefSeq: NCBI Reference Sequence Database

This is a collection which provides a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses.

23/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

The main features of the RefSeq collection:

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators

2.6.1.1.2: GENBANK

It is a comprehensive database that contains publicly available nucleotide sequences for more than 3,00,000 organisms. The data is obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. It is an annotated collection of all publicly available nucleotide sequences. GenBank is accessible through the NCBI Entrez retrieval system.

Genbank integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed.

2.6.1.1.3: European Molecular Biology Laboratory (EMBL)

The EMBL Nucleotide Sequence Database is Europe's primary nucleotide sequence resource. In collaboration with DDBJ and GenBank the database is produced, maintained and distributed at the European Bioinformatics Institute. EMBL Database entries are grouped into divisions. The EMBL Database currently consists of 17 divisions with each entry belonging to exactly one division. Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages.

2.6.1.1.4: DDBJ (DNA Data Bank of Japan)

28/12/2014
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

The DNA Data Bank of Japan is a public database of nucleotide sequences established at the National Institute of Genetics (NIG). The DDBJ Center also services Japanese Genotype-phenotype Archive (JGA), with the National Bioscience Database Center. The JGA is a permanent archiving service for genotype and phenotype data of human individuals. There are various services that are provided by DDBJ like AGD which is a controlled-access database for sharing individual-level genotype and phenotype information among specific researchers. ARSA is a service provided by DDBJ for the retrieval of the information using accession number and keywords. BLAST, Bioproject, CRISPRDirect, CLUSTAL W, DFAST etc. are some of the other services provided by DDBJ.

2.6.1.2.Protein Sequence databases are

- PIR,
- SWISS-PROT,
- TrEMBL,
- PROSITE and
- PFAM

2.6.1.2.1: Prosite

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. PROSITE is complemented by collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids.

2.6.2: Structural database

Structural databases help in producing, solving, refining and publishing the structure of a new material. One can consult these databases before synthesizing new compounds. PDB, SCOP and CATH are structure databases.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2.6.2.1: The Protein Data Bank (PDB)

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. It contains information related with the 3D structure of proteins, nucleic acids and complex assenblies. RCSB PDB, PDBe, PDBj, BMRB and EMDB are the members of PDB.

2.6.2.2: Structural Classification of Proteins (SCOP)

It provides a detailed and comprehensive description of the relationships of known protein structures. SCOP provides its data with the help of Representation State Transfer (REST) protocol. The classification is on hierarchical levels:

- Family,
- Super family and
- Fold

2.6.3: Functional database

A functional database is one which provides information related to the physiological roles of various biological entities like genes, enzymes etc. KEGG, Reactome and BioCyc are some of the examples of Functional Databases. BRENDA is an enzyme activity Database. It is Comprehensive Enzyme Information system. ExPASY-ENZYME is repository of information relative to the nomenclature of enzymes.

2.6.3.1: KEGG

It is a database that helps in understanding the high level functions and utilities of cells, organisms etc. It provides the computer based representation of biological systems which contain genomic and chemical information. It contains information of diseases and drugs and their inpact on normal biological systems.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2.6.3.2: BRENDA

BRENDA is an enzyme database. It has collection of the enzyme functional data. It is maintained and developed at the Institute of Biochemistry and Bioinformatics at the Technical University of Braunschweig, Germany. Information related to the functional properties of enzymes is taken directly from the works published by the scientists.

2.7: HOW TO RETRIEVE INFORMATION FROM DATABASES

To perform complex queries in a database often requires the use of Boolean operators like AND, OR, and NOT to indicate relationships between the keywords used in a search. With the help of various keywords and the combination of keywords, we can search the biological databases easily.

2.8: LIMITATIONS AND PITFALLS

- Redundancy is a major problem Affecting Primary Databases
  - Some problems are because of Human Errors.
  - Often, the same gene sequence is found under different names resulting in multiple entries and confusion about the data.
- Sequencing Error
  - Most errors in Nucleotide Sequences are caused by Sequencing Errors.

2.9: ADVANTAGES OF BIOLOGICAL DATABASES

- Easy access of Data/information
- Provide Information to answer various biological questions
- Open access to all users
- It allows indexing of all the data
- It Connects the experimental work or data with programming skills

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- With the help of databases, same information can be accessed by multiple users at the same time.
- Different types of databases, helps users narrow down their search criteria and helps in quick retrieval of information.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN
28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References

- https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/
- Lesk, A. M. (2002). Introduction to bioinformatics. (2nd ed.). Oxford: Oxford University Press.
- Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer.
- Xiong, J. (2006). Essential Bioinformatics. (1st ed.).Cambridge University Press.
- Benson DA et al. Genbank; Nucleic Acids Res. 2010 Jan; 38(Database issue). doi: 10.1093/nar/gkp1024
- https://www.ncbi.nlm.nih.gov/genbank/
- http://pd.chem.ucl.ac.uk/pdnn/pubs/databs1.htm
- https://www.wwpdb.org/

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3: SEQUENCE DATABASES

Biological databases provide updated high quality data.

Sequence databases are biological databases that store large collection of nucleotide sequences and protein sequences digitally.

Sequence databases can be divided into nucleic acid sequence databases and protein sequence databases

The Nucleic acid sequence database is a collection of sequences from several sources.

These sources can be:

- GenBank
- RefSeq
- TPA
- PDB

3.1: Uses of these databases:

- discovery of novel genes
- identification of homologous genes
- chromosomal localization of genes
- analysis of alternative splicing
- detection of polymorphisms

GenBank, EMBL and DDBJ are the most comprehensive publically available nucleotide sequence databases.

3.2: GenBank

Genbank is a publically available comprehensive nucleotide sequence database which also supports bibliographic and biological annotation.

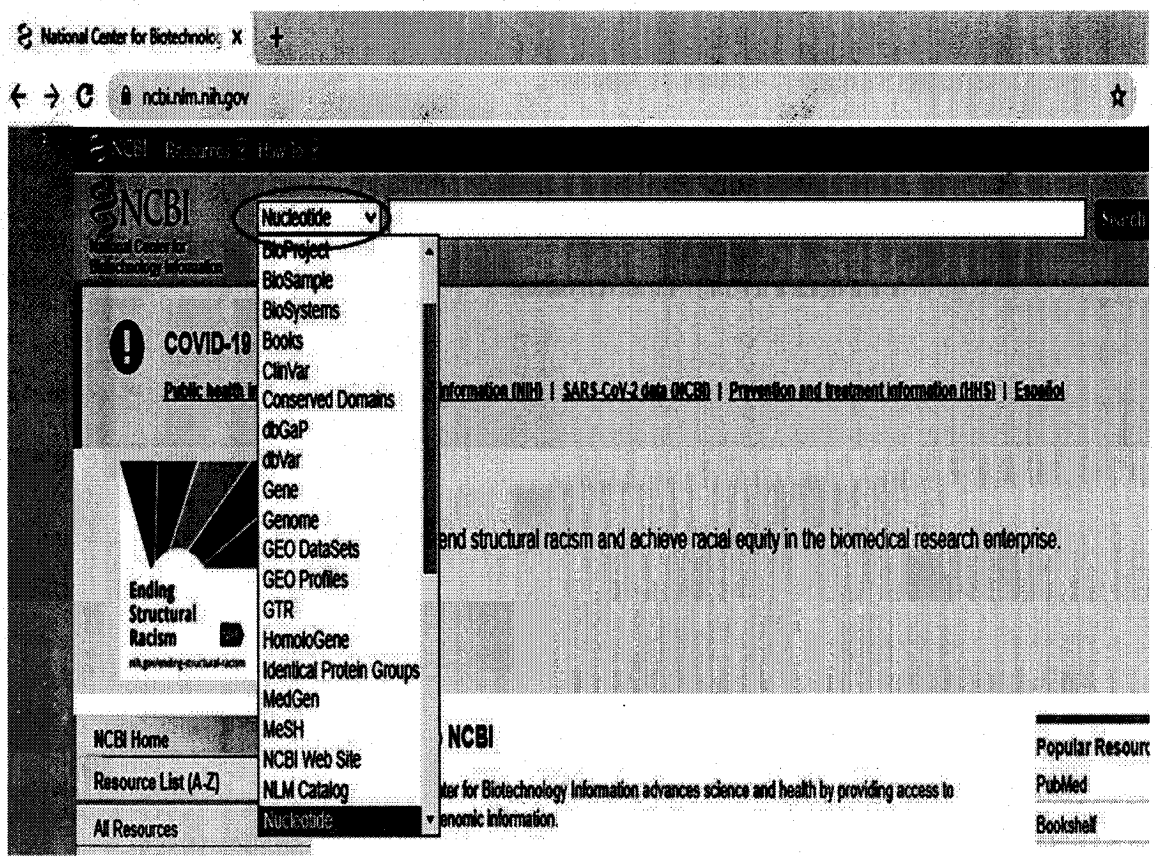
28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

It is built and distributed by the National Center for Biotechnology Information (NCBI).

GenBank, EMBL Data Library, and (DDBJ) comprise the International Nucleotide Sequence Databases, in which information is exchanged daily to ensure a uniform and comprehensive collection of sequence information.

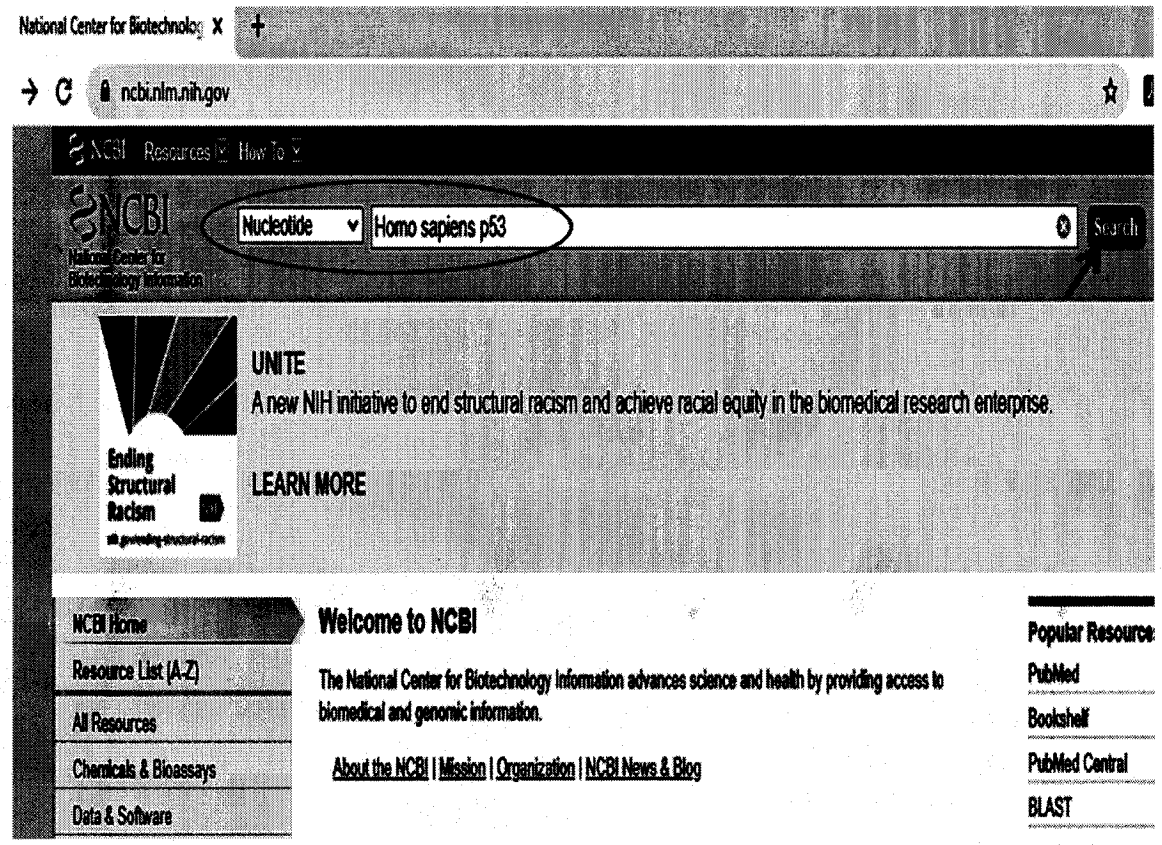
3.2.1: How to use it

- There are several ways to search and retrieve data from GenBank.
- To search any nucleotide sequence-
- Go to the search, select nucleotide from the dropdown, enter the name of any gene whose sequence you want to know and click search



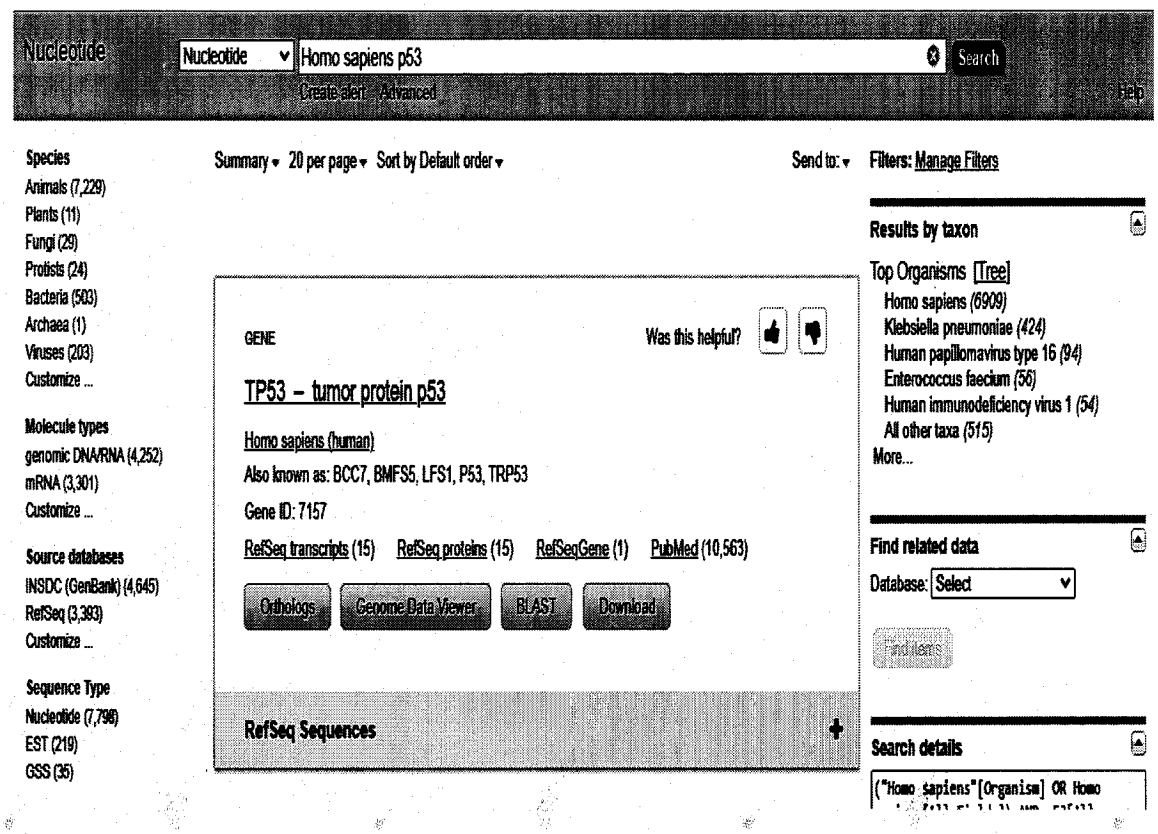
28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



28/12/2021

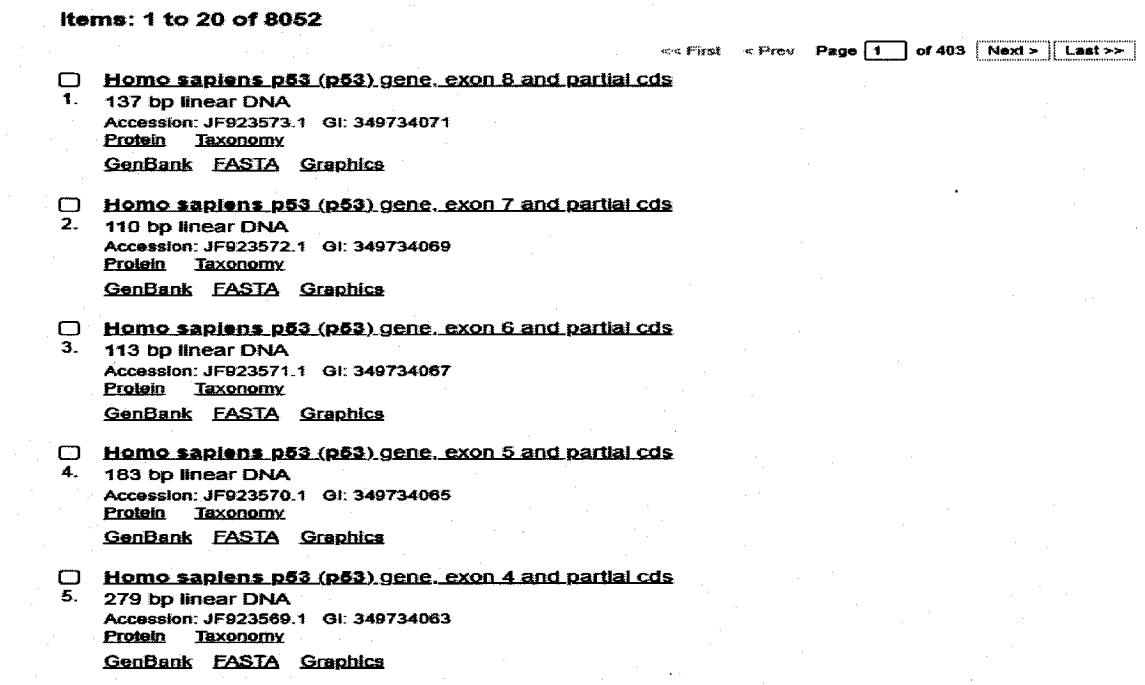
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



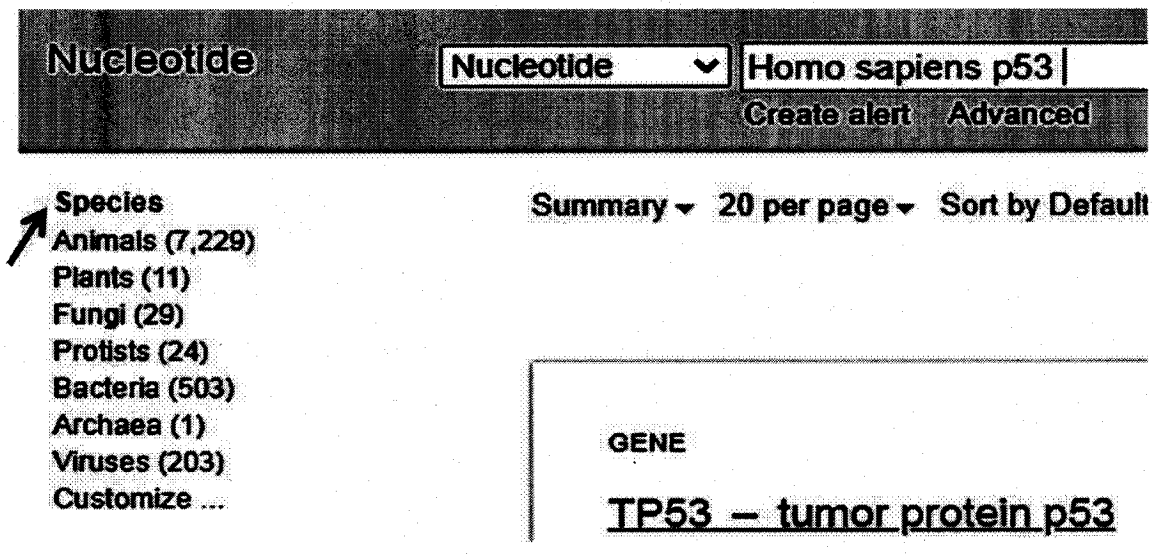
28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

On top of the results, we have total number of hits obtained against our query



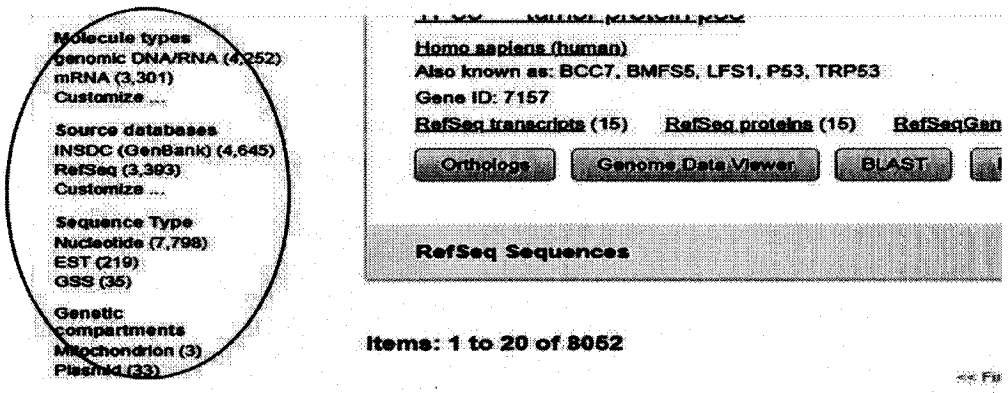
On left, there is a list of species depicting no.of records against various organisms



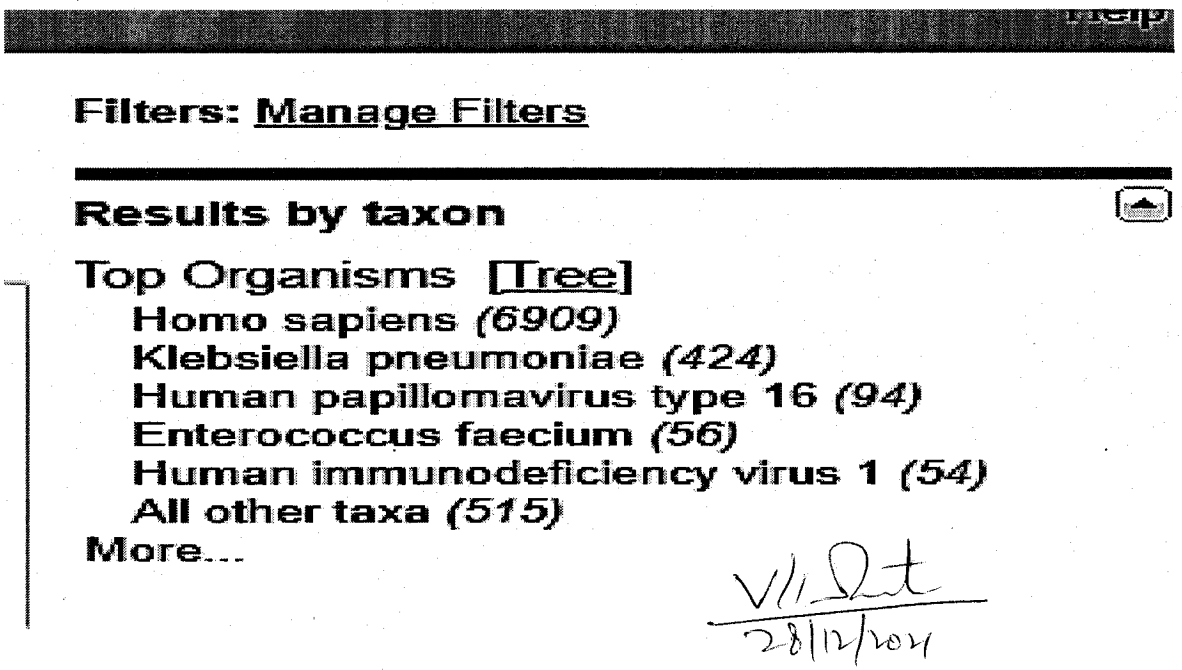
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

We also get information about the source type, types of molecules presented in our results and also the type of sequences for example: nucleotide or GSS or EST.



On the right hand, we have results by Taxon from where we can further refine our search criteria.



28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

In the results that we have obtained, we can get some information on top based on which we can further make our decision on the hit to select as our final result. On top we have information like the length of the sequence in base pairs, the type of sequence i.e whether the sequence is of DNA or RNA, the name of the organism etc.

1. **Homo sapiens p53 (p53) gene, exon 8 and partial cds**
   137 bp linear DNA
   Accession: JF923573.1 GI: 349734071
   Protein Taxonomy
   GenBank FASTA Graphics

2. **Homo sapiens p53 (p53) gene, exon 7 and partial cds**
   110 bp linear DNA
   Accession: JF923572.1 GI: 349734069
   Protein Taxonomy
   GenBank FASTA Graphics

3. **Homo sapiens p53 (p53) gene, exon 6 and partial cds**
   113 bp linear DNA
   Accession: JF923571.1 GI: 349734067
   Protein Taxonomy
   GenBank FASTA Graphics

4. **Homo sapiens p53 (p53) gene, exon 5 and partial cds**
   183 bp linear DNA
   Accession: JF923570.1 GI: 349734065
   Protein Taxonomy
   GenBank FASTA Graphics

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

GenBank FASTA Graphics

☐ **Homo sapiens p53 gene for P53, exon 5, partial cds**
9. **184 bp linear DNA**
Accession: AB118156.1 GI: 42821409
Protein Taxonomy
GenBank FASTA Graphics

☐ **Homo sapiens mRNA for P53, complete cds**
10. **2,451 bp linear mRNA**
Accession: AB082923.1 GI: 23491728
Protein PubMed Taxonomy
GenBank FASTA Graphics

☐ **Homo sapiens p53 gene, exon 7 and partial cds**
11. **110 bp linear DNA**
Accession: DQ075362.1 GI: 70611297
Protein Taxonomy
GenBank FASTA Graphics

Once we have chosen the result of our interest, we land on the main page of our result.

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1
FASTA Graphics

Go to:

```
LOCUS       JF923571                 113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-MAY-2011) Department of Medicine, Institute of
            Medical Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh
            221005, India
```

28/12/16

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

An accession no that is a unique identifier for all the entries that are made in GenBank.

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

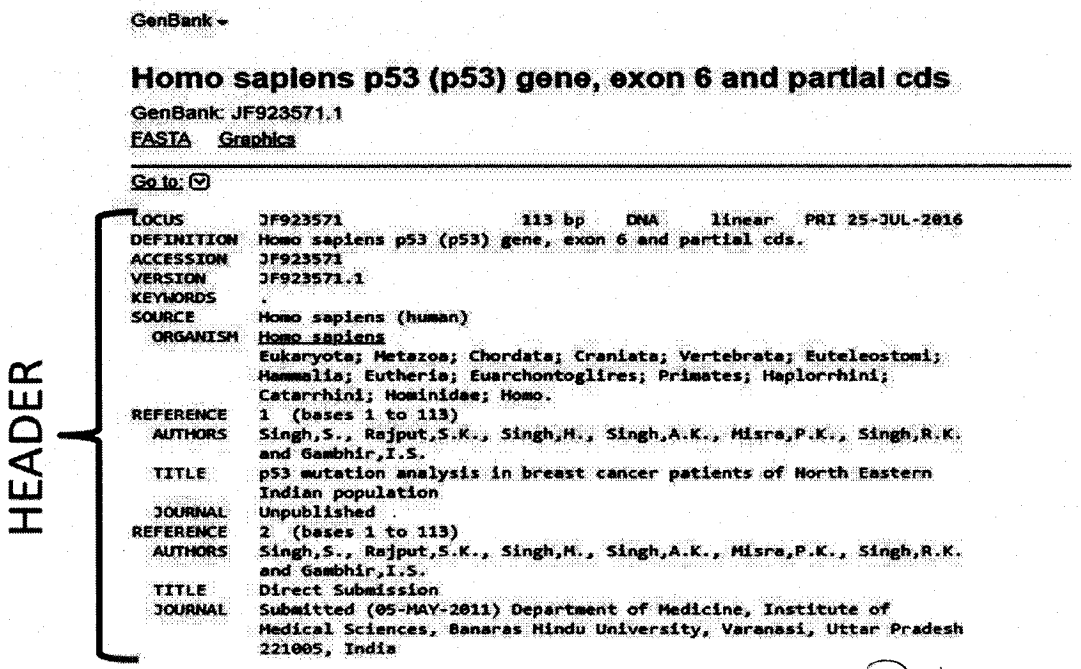
GenBank: JF923571.1

FASTA Graphics

Go to:

```
LOCUS       JF923571                113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
```

The header section



28/12/2017

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

The features section

```
FEATURES             Location/Qualifiers
     source          1..113
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /isolation_source="healthy case in northeastern Indian
                     population"
                     /db_xref="taxon:9606"
     gene            <1..>113
                     /gene="p53"
     mRNA            <1..>113
                     /gene="p53"
                     /product="p53"
     CDS             <1..>113
                     /gene="p53"
                     /codon_start=3
                     /product="p53"
                     /protein_id="AEQ16484.1"
                     /translation="LAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPE"
     exon            1..113
                     /gene="p53"
                     /number=6
```

Origin section

```
ORIGIN
        1 gtctggcccc tcctcagcat cttatccgag tggaaggaaa tttgcgtgtg gagtatttgg
       61 atgacagaaa cacttttcga catagtgtgg tggtgcccta tgagccgcct gag
//
```

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

To retrieve the sequence, click on FASTA at the top of the page

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1

FASTA Graphics

Go to:

```
LOCUS       JF923571                113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
```

FASTA formats start with a more than symbol. The sequences is in continues form without any gap between two base pairs.

FASTA

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1

GenBank Graphics

```
>JF923571.1 Homo sapiens p53 (p53) gene, exon 6 and partial cds
GTCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAA
CACTTTTCGACATAGTGTGGTGGTGCCCTATGAGCCGCCTGAG
```

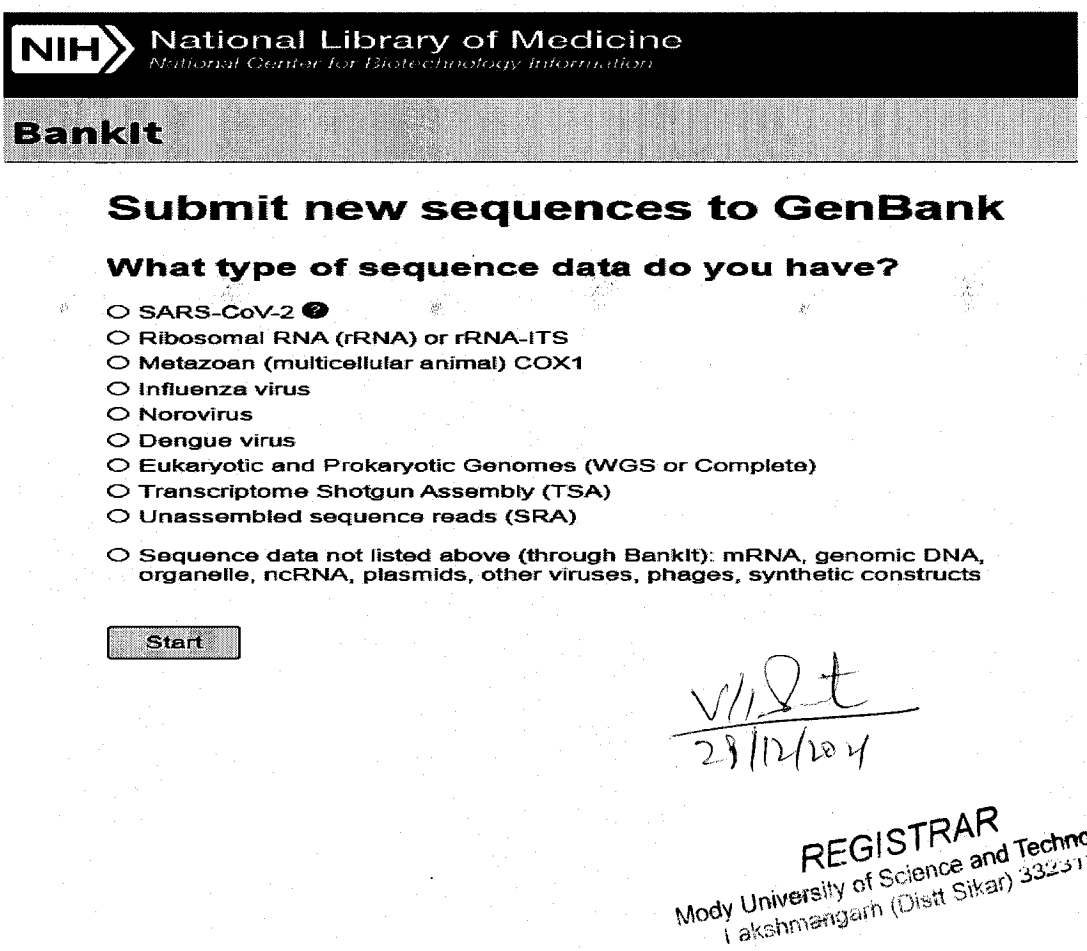
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3.2.2: Submitting data into GenBank

- **Web-based submission tools**
  - **BankIt**
  - **Submission Portal**
  - **tbl2asn**

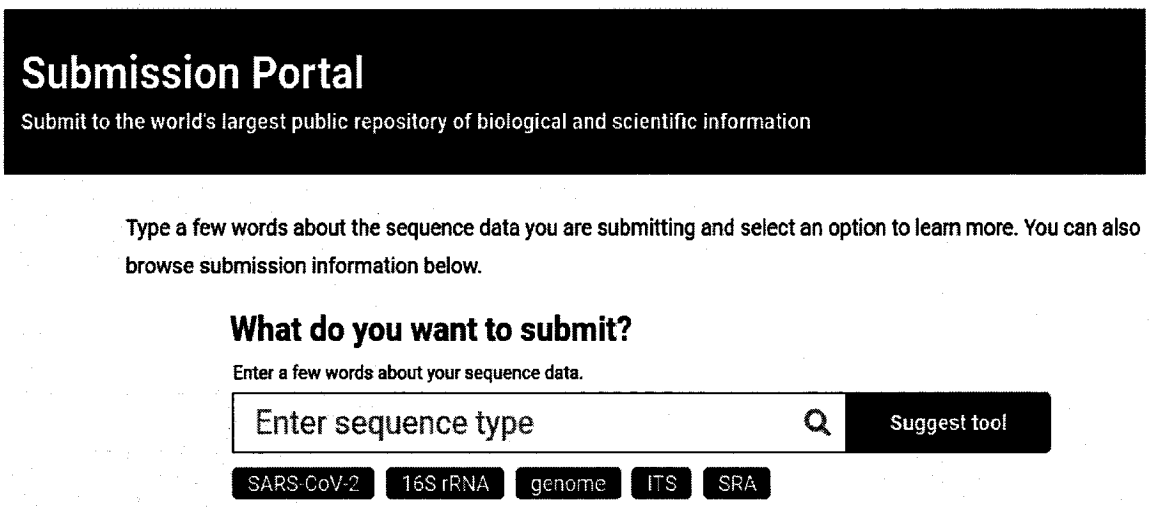
3.2.2.1: BankIt



29/12/24

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3.2.2.2: Submission Portal



- Supports the submissions of
  - ribosomal RNA (rRNA),
  - metazoan mitochondrial COX1,
  - Influenza,
  - Norovirus,
  - Dengue or
  - SARS-CoV-2 sequences
- One can also submit Genome and Transcriptome Assemblies through the Genomes and TSA portals, respectively.
- Submission preparation tools are required for uploading via the Submission Portal.

We can also email to gb-sub@ncbi.nlm.nih.gov

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

3.2.2.3: tbl2asn

- a command-line program that automatically creates the sequence of records for submission to the GenBank.
- used primarily for submission of complete genomes and large batches of sequences.

Genome Workbench

- offers a rich set of integrated tools to study and analyze the genetic data.
- The allows us to prepare submissions of single eukaryotic and prokaryotic genomes.
- One can also use Genome Workbench for editing and visualizing files that are created by tbl2asn.

3.3: The European Molecular Biology Laboratory

- It is the primary nucleotide sequence resource maintained by the European Bioinformatics Institute (EBI), situated in the United Kingdom.
- In this The sequences of DNA or RNA are submitted directly from individual researchers.
- The EMBL Database in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Data Bank of Japan (NIG, Mishima) has been processing nucleotide sequence data reported world-wide since 1982.

3.3.1: How to use it

- Access to the sequence data is provided via ftp and several WWW interfaces.
- The Sequence Retrieval System or SRS allows the databases to be searched using, for example, sequence, annotations, keywords, author names.

Format

Sequence entry starts with an identifier line ("ID")

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").


```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```


28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3.4: DNA Data Bank of Japan

DDBJ initiated the DNA data bank activities in 1986 at the National Institute of Genetics (NIG).

It collects DNA sequence data mainly from Japanese and researchers from all over the world.

3.4.1: How to submit the data

- For submission, It offers interactive web interface as well as mass submission system.
- SAKURA is a Web-based tool of DDBJ which is used for nucleotide sequence submission and annotation.

3.5: Protein sequence databases

- The Protein database is a collection of sequences from several sources
  - translations from annotated coding regions in GenBank, RefSeq and TPA
  - records from SwissProt, PIR, PRF, and PDB
  - Protein sequences are the fundamental determinants of biological structure and function.

3.5.1: GenPept

- It is a part of the National Center of Biotechnology Information (NCBI)
- Its entries are derived from translations of the sequences present in Genbank, DDBJ and EMBL.
- The entries lack additional annotation and the database does not contain proteins derived from amino acid sequencing.
- The format is same as that of Genbank having header, features and origin sections.
- The format ends with double slashes.
-

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3.5.2: Refseq

- RefSeq stands for Reference sequence which is a collection developed by NCBI.
- It provides a non-redundant collection of reference protein sequences.
- It has sequences for some limited organisms which includes human, mouse, rat, honeybee, zebrafish, cow, sea urchin, and several important plant species and also some 1100 viruses and 150 bacteria

3.5.3: Protein information resource

- It is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies.
- PIR was established in 1984 by (NBRF) the National Biomedical Research Foundation for the identification and interpretation of protein sequence information.
- It assists with proteomic and genomic data integration and the propagation and standardization of protein annotation.
- It is the oldest universal curated protein sequence database

LAKSHMANGARH • RAJASTHAN

3.5.4: Swiss Prot

- It is the leading universal curated protein sequence database.
- The database is non-redundant and is highly integrated with other databases.
- Literature-based curation is used to extract experimental data. This experimental knowledge is supplemented by manually confirmed results from various sequence analysis programs.

3.5.5: Translated EMBL Nucleotide Sequence Data Library

- It is a computer annotated protein sequence database.
- TrEMBL translates the nucleotide sequences in the DDBJ/EMBL/GenBank nucleotide sequence database.
- It was introduced to make sequences available quickly.
- TrEMBL has two main sections; SP-TrEMBL and REM-TrEMBL

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries which should be incorporated into SWISS-PROT.
- REM-TrEMBL (REMaining TrEMBL) contains the entries that are not to be included in SWISS-PROT.
- Most REM-TrEMBL entries are immunoglobulins and T-cell receptors.

3.5.6: Uniprot

- It stands for Universal Protein Resource
- It is a publically available protein sequence database.
- UniProt is a collaboration between the EMBL-EBI, the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).
- It is a comprehensive resource for protein sequence and annotation data.

It provides information on

- function,
- sub-cellular location,
- expression,
- interaction,
- structure,
- protein family,
- domain and
- sequence.

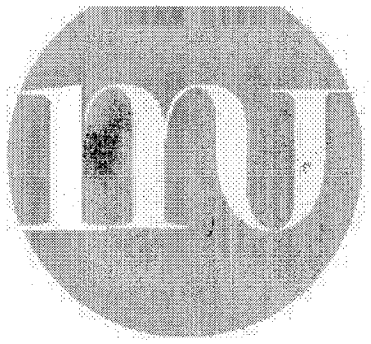
Apart from this, one can also find various published papers in support of the information provided.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Important links

- https://www.ncbi.nlm.nih.gov/genbank/
- https://www.ncbi.nlm.nih.gov/WebSub/
- https://submit.ncbi.nlm.nih.gov/
- https://www.ncbi.nlm.nih.gov/protein/
- https://www.ebi.ac.uk/
- https://www.ddbj.nig.ac.jp/index-e.html
- https://www.ncbi.nlm.nih.gov/refseq/
- https://www.uniprot.org/



MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

4. STRUCTURE DATABASES

Structure databases or databanks contain atomic coordinates, bibliographic citations, and other information related to structural studies. Structural databases are the essential tools for all crystallographic works. They are used in the process of producing, solving, refining and publishing the structure of a new material. Before deciding to synthesize a new compound the database could be used to check how many compounds with a particular chemical composition have been reported.

After synthesizing and indexing the unit cell of a material the database can be searched to see if a material with the same or a similar unit cell is already known. If a material is found in the database with a similar unit cell to the new material, then its structure may be close enough (i.e. same symmetry and similar unit cell contents) to be used as the starting model for the refinement of the new material. To verify the results of a structure refinement the database can be consulted to find structures that have comparable bond distances, bond angles or coordination environments to the new structure.

4.1: Protein Data Bank (PDB)

- Protein Data Bank archive is a single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.
- The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.
- Currently, Most of the structures are determined by X-ray diffraction, but about 10% of structures are determined by protein NMR.
- The .pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity.
- Its fixed-column width format is limited to 80 columns, which was based on the width of the computer punch cards that were previously used to exchange the coordinates.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

MODY UNIVERSITY

The format has a Header section along with the title and author records. This provide information related to the researchers who defined the structure.

Remark can contain free-form annotation, but they also accommodate standardized information.

Atom describe the coordinates of the atoms that are part of the protein.

Hetatm describe coordinates of hetero-atoms, that is those atoms which are not part of the protein molecule.

Ter indicates the end of a chain of residues.

28/12/2017

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332311

4.2: RCSB PDB

- RCSB PDB stands for Research Colaboratory for Structural Bioinformatics PDB.
- This resource is supported by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies
- As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.
- 178229 biological macromolecular structures enabling breakthroughs in research and education.

4.3: SCOP

- SCOP stands for Structural Classification of Proteins.
- Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin.
- It aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- includes 68,816 non-redundant domains representing 772,354 protein structures

UNIVERSITY
LAKSHMANGARH • RAJASTHAN

The main levels of the classification are:

- *Family-* It groups closely related proteins with a clear evidence for their evolutionary origin.
- In most cases, their relationship is detectable with current sequence comparison methods, e.g. BLAST, PSI- BLAST, HMMER.
- *Super-family-* It brings together more distantly related protein domains.
- Their similarity is frequently limited to common structural features that along with a conserved architecture of active or binding sites, or similar modes of oligomerization suggest a probable evolutionary ancestry.
- *Fold* groups super-families on the basis of the global structural features shared by the majority of their members. These features are the composition of the secondary structures in the domain core, their architecture and topology.
- Fold is an attribute of a super-family but the constituent families of some super-families that have evolved distinct structural features can belong to a different fold.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- *IUPR* (Intrinsically Unstructured Protein Region) organizes super-families of proteins or protein regions that do not adopt globular folded structure.
- *Classes* bring together folds and IUPRs with different secondary structural content.
- These include all-alpha and all-beta proteins, containing predominantly alpha-helices and beta-strands, respectively, and 'mixed' alpha and beta classes (a/b) and (a+b) with respectively alternating and segregated alpha-helices and beta-strands, and the fifth class of small proteins with little or no secondary structures.

4.3.1: ORGANISATION AND FACILITIES OF SCOP

SCOP is organized as a tree structure.

- Entering at the top of the hierarchy the user can navigate through the levels of Class, Fold, Super-family, Family and Species to the leaves of the tree which are structural domains of individual PDB entries.
- An alternative hierarchy of Folds, Super-families and Families by the date of solution of the first representative structure is also provided.

We can search in SCOP

- With the help amino acid sequence,

  The Sequence similarity search facility allows any sequence of interest to be searched against databases of protein sequences classified in SCOP using the algorithms BLAST, FASTA or SSEARCH.

- From a keyword

  The keyword search facility returns a list of SCOP pages containing the word entered or combinations of words separated by a series of boolean operators.

- From a PDB identifier

22/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

The PDB entry viewer links PDB entries to various graphical views, external databases and SCOP itself.

4.3.2: OTHER USES OF SCOP

- Non-redundant sequence databases are available from SCOP and it helps in the evaluation of sequence alignment methods.
- Assignment of protein structures to sequences using the intermediate sequence library PDB-ISL.
- Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL.

4.4: CATH

- CATH database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains.
- The CATH database provides hierarchical classification of protein domains based on their folding patterns.
- Domains are obtained from protein structures deposited in the Protein Data Bank and both domain identification and subsequent classification use manual as well as automated procedures.
- The CATH database is a classification of protein domains, based not only on **sequence information**, but also on **structural and functional properties**.
- CATH contains **114,215 domains**.
- It is classified in a hierarchical scheme with four main levels called **class (C), architecture (A), topology (T)** and **homologous** super-family **(H)**.
- From these words they have derived the name CATH.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

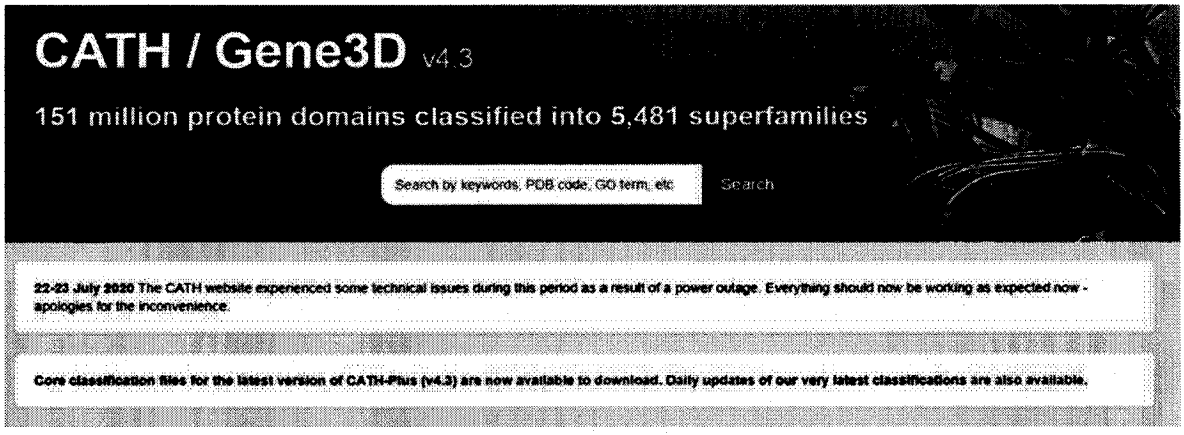
- At the **C-level**, domains are grouped according to their secondary structure content into four categories: mainly alpha, mainly beta, mixed alpha-beta; and the last category contains domains with only few secondary structures.
- The **A-level** groups domains according to the general orientations of their secondary structures.
- At the **T-level**, the connectivity (ie the order) of the secondary structures is taken into account.
- The grouping of domains at the **H-level** is based on a combination of both sequence similarity.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

4.4.1: Searching in CATH



MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

IMPORTANT LINK

- **https://scop.mrc-lmb.cam.ac.uk/**
- **https://www.cathdb.info/**
- **https://www.rcsb.org/**
- **https://www.wwpdb.org/**

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

5: STRUCTURE DATABASES (Part 2)

5.1: PFAM

- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
- Pfam, version 27.0, contains 14 831 Pfam-A families. Of these families, 4563 have been classified into 515 clans.
- Pfam also generates higher-level groupings of related entries, known as clans.
- A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.
- Profile HMMs turn a multiple sequence alignment into a position-specific scoring system, which can be used to align sequences and search databases for remotely homologous sequences.
- Profile HMMs have a formal probabilistic basis and uses consistent theory behind gap and insertion scores.
- Profile hidden Markov models (HMMs) have several advantages.
- HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency.
- The HMMER package, is a set of programs that allow us to create and manipulate profile HMMs and databases of profile HMMs, perform sensitive searches of sequence and profile HMM databases and create multiple sequence alignments efficiently

5.1.1: FORMAT of PFAM

- **Plain Pfam format**
- **Flatfile**

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

```
YHI1_MYCLE/3-102      TIFTKIINRELPGRFVYEDDDVVAFLTIEPMTQGHTLVVP..CAEIDQMQMVDP...AI....FGRVIAVSQLIGKGVCRAFNAERAGVII.....AGFEVPHLHIHVFPTHSL
YHIT_MYCHR/7-111      ELFLKIIKREEPATILYEDDKVIAFLDKYAHTKGHFLVVP..KNYSRNLFSISD...ED....LSYLIVKAREFALQEIKKLGATGFKLLINNEPDAEQSIFHTHVHIIPYYKK
YHIT_HELPY/2-104      MVFEKIIQGEIPCSKILENERFLSFYDINPKAKVHALVIP..KQSIQDFNGITP...EL....MAQMTSFIFEVVEKL..GIKEKGYKLLTNVGKNAGQEVMHLHFHILSGDKH
YHIT_AZOBR/9-115      MVFARILRGEIPCKKVLETEHALAFHDINPQAPTHILVIP..KGAYVDMDDFSA...RATEAEIAGLFRAVGEVARGA..GAAEPGYRILSNCGEDANQEVPHLHIHVFAGRRL
PKIA_DICDI/68-172     TIFAKIISGAIPSKKFYDDEYCIAIEDINPQAPVHLLVIP..KLAVGGLSDVAMVDLEKYKESMGHIMSKIHHIASLK....GADSYRLVINEGVLGQQSVRWLHIHIPWW...
ZB14_MAIZE/17-123     TIFDKIIKKEIPSTWVYEDEKVLAFRDINPQAPTHILIIPKVKDGLTGLAKAEE...RHIEI.LGYLLYVAKVVAKQE..GL.EDGYRVVINDGPSGCQSVYHIHMHLLGGRQM
Y961_HAEIN/5-110      TIFSKIIRKEIPANIVYQDELVTAFRDISPQAKTHILIIP..NKVIPTVNDVTE...QDEVA.LGRLFSVAAKLAKEE..GVAEDGYRLIVNCNKHGGQEVFHLHMHLVGGEPL
YCFF_ECOLI/5-110      TIFSKIIRREIPSDIVYQDDLVTAFRDISPQAPTHILIIP..NILIPTVNDVSA...EHEQA.LGRMITVAAKIAEQE..GIAEDGYRLIMNTNRHGGQEVYHIHMHLLGGRPL
YHIT_CAEEL/21-125     TLFGKIIRKEIPAKIIFEDDEALAFHDVSPQAPIHFLVIP..KRRIDMLENAVD...SDAAL.IGKLMVTASKVAKQL..GM.ANGYRVVVNNGKDGAQSVFHLHLHVLGGRQL
HINT1_RAT/16-120      TIFGKIIRKEIPAKIIFEDDRCLAFHDISPQAPTHFLVIP..KKYISQISAAED...DDESL.LGHLMIVGKKCAADL..GLK.KGYRMVVNEGSDGGQSVYHMHLHVLGGRQM
YHIT_SYNY3/5-109      TIFSKIIRREIPAAIVYEDDLCLAFKDVNPQAPVHVLLIP..KKPLPQLSAATP...EDHAL.LGHLLLKAKEVAADL..GI.GDQFRLVINNGAEVGQTVFHLHLHILGGRPF
YHIT_SYNP7/5-109      TIFGKIIRREIPADIVYEDDLCLAFRDVAPQAPVHILVIP..KQPIANLLEATA...EHQAL.LGHLLLTVKAIAAQE..GL.TEGYRTVINTGPAGGQTVYHLHIHLLGGRSL
Y866_METJA/2-105      CIFCKIINGEIPAKVVYEDEHVLAFLDINPRNKGHTLVVP..KKHYERFDEMPD...DE....LCNFIKGVKKTVEVL.KKLGFDGYNIVMNNGRVAGQEVNHVHFHIIPRYEG
```

5.2: Nucleic acid structure databases

The 3D structure of nucleic acid can be well understood by investigating the different types of nucleic acid functions such as nucleic acid–protein interactions, functional RNAs, etc.

Nucleic Acid Database (NDB)

- The Nucleic Acid Database (NDB) was the first database for nucleic acid structures.
- It was designed in the early 1990s.
- It contains the contents of primary data.
- It also contains some derived geometric data, classification of structure and motifs, standards for describing nucleic acid features,
- NBA also has tools and software for analysis of nucleic acids.
- The structures available in the NDB are annotated with various information that are unique to the nucleic acids
- The structures of NBA are almost not available from the original entry of PDB.
- The structures of NBA are analyzed by using computational methods of structural geometries.
- They are then classified by searching the sequence, structure levels, and secondary structure by incorporating the PDB entries to the NDB.

NDB Entry Format

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- The main content of NDB is the primary structure information that is related to the nucleic acids that are acquired from the PDB and also from classification and derived data
- The primary data present in the NDB is obtained from PDB.
- It contains experimental files, identifiers, structural descriptions, citations, crystal data, coordinate information, and various details related to the crystallization, data collection, and structure refinement
- The structures available in the NDB database possess their own personal summary page that is mostly appropriate to that particular structure.
- There are about four sections of data in the structure summary page, including
    - primary structure information,
    - downloads,
    - derived structural data, and
    - images.

Home page of NDB where one can search for the 3D structural information of DNA and RNA along with some advanced searches.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

NUCLEIC ACID DATABASE

A Portal for Three-dimensional Structural Information about Nucleic Acids

As of 26-May-2021 number of released structures: 11337

Search DNA | Search RNA | Advanced Search

Enter an NDB ID or PDB ID

Search for released structures

Welcome to the NDB

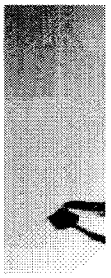
The NDB contains information about experimentally-determined nucleic acids and complex assemblies.
Use the NDB to perform searches based on annotations relating to sequence, structure and function, and to download, analyze, and learn about nucleic acids.

Search Structures

Search DNA
Search DNA and its complexes

Search RNA
Search for RNA structures in the NDB archive or in the Non-Redundant list

Advanced Search
Search for structures based on structural features, chemical features, binding modes, citation and experimental information



28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Featured Tools in NBA

- **RNA 3D Motif Atlas**
- **Non-redundant Lists**
- **RNA Base Triple Atlas**
- **WebFR3D**
- **R3D Align**

Tools available in NBA

Nucleic Acid tools:

- **DNA**
  - DNAProDB
  - Assignment of DNA conformers
- **RNA**
  - RNA Basepair catalog
  - RNA 3D motif Atlas
  - Representative RNA 3D Structure sets
  - R3D Align
  - WebFR3D
  - RNA viewer

FRASS

28/12/2021

Nucleic Acid Softwares

- FR3D
- RNAView
- RNAMLView
- 3DNA
- FreeHelix98
- Hel_param
- The UNAFold Web server

REGISTRAR
Maharishi University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References

- http://ndbserver.rutgers.edu/
- http://pfam.xfam.org/
- Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, Alex Bateman, Pfam: The protein families database in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D412–D419, https://doi.org/10.1093/nar/gkaa913

UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

6. SEQUENCE ALIGNMENT (PART 1)

What we understand by Sequence Alignment?

- Arranging DNA, RNA or Protein sequences in order to find the regions of similarity.
- This help to derive functional, structural and evolutionary relationships between them.
- We can also make prediction about the 3D structures of proteins.
- If two sequences from two different organisms are similar then we can say there is a possibility that both sequences are Homologous i.e they share some common ancestry.

Types of Sequence Alignments

Sequence Alignment can be classified on the basis of:

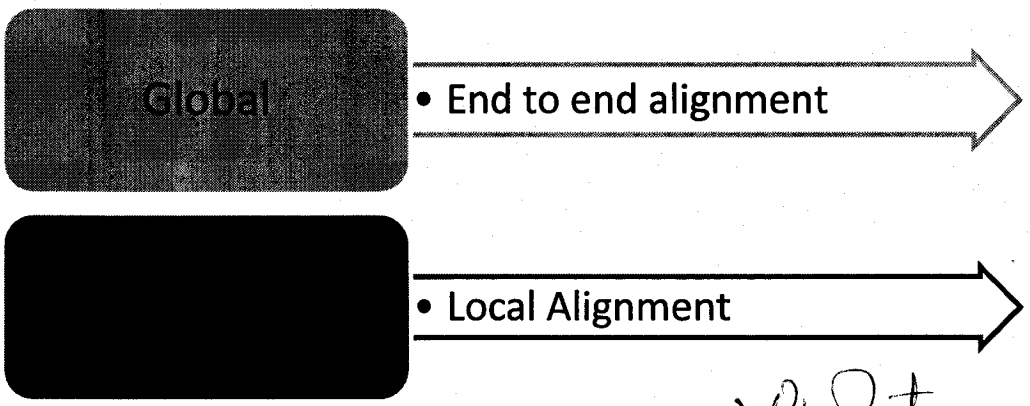
1. Sequence Length-

   How much portion of the sequences under study are being covered.

2. Number of sequence-

   If we are aligning two sequences or we are trying to align multiple sequences

Based on sequence Length



MODY UNIVERSITY

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Global Sequence Alignment

- In this alignment sequences are aligned along their entire length to include as many matching characters as possible.
- We usually do global alignment, when we want to compare two homologous genes or the genes that perform the same function in two different species
- Or to compare two proteins with similar function

Example of Global Sequence Alignment

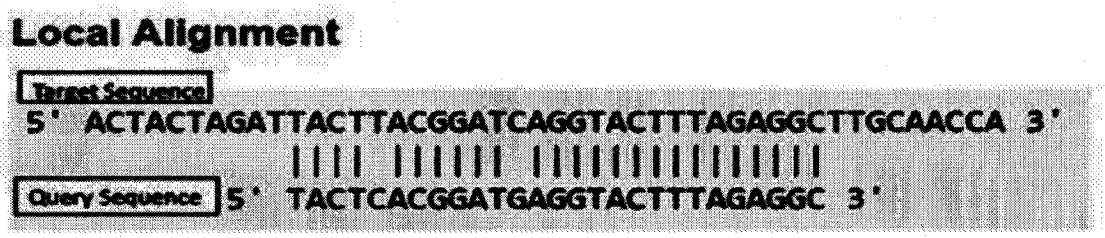
Local Sequence Alignment

- The sequences are aligned to find a region of higher density or strong similarity.
- This helps in finding conserved domains or motifs in two protein sequences or some conserved patterns in two DNA sequences.
- The target sequence and a query sequence and our query sequence is not covering the target sequence in its entire length.

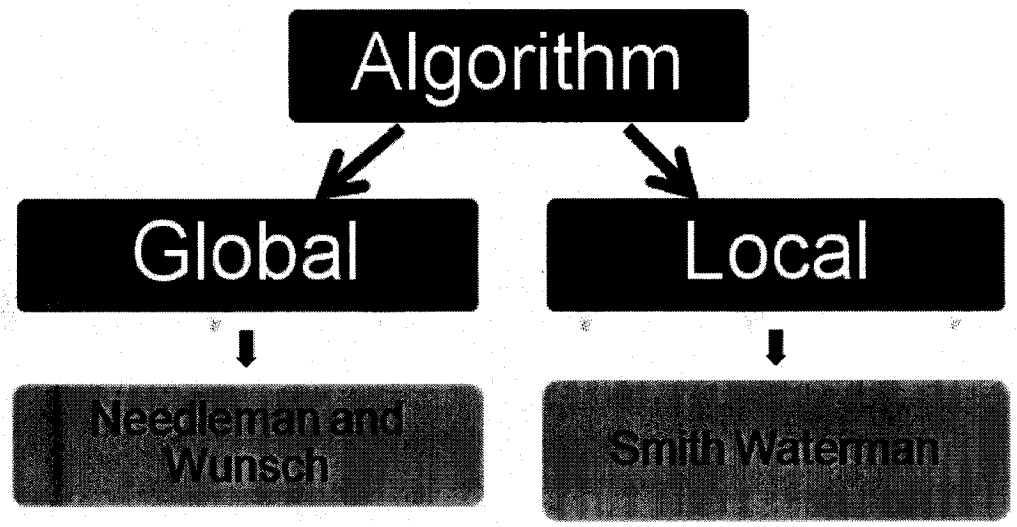
28/12/2014

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

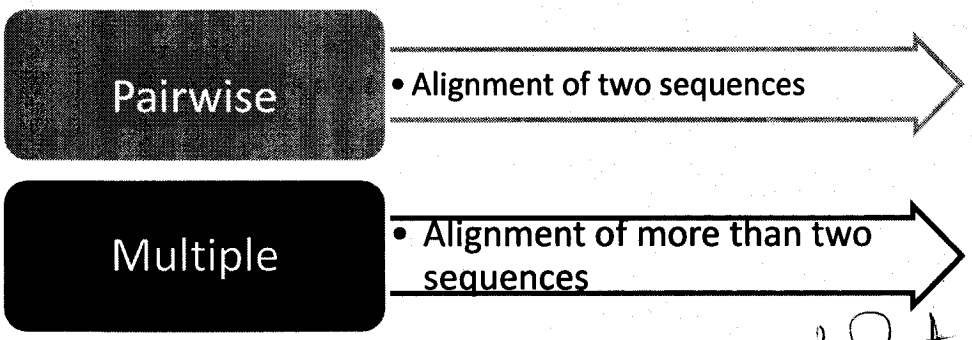
Example of Local Sequence Alignment



Algorithms for Sequence Alignments



Based on Number of Sequence



28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Pairwise Sequence Alignment

The methods that are used for Pairwise alignment of nucleotide and protein sequences are:

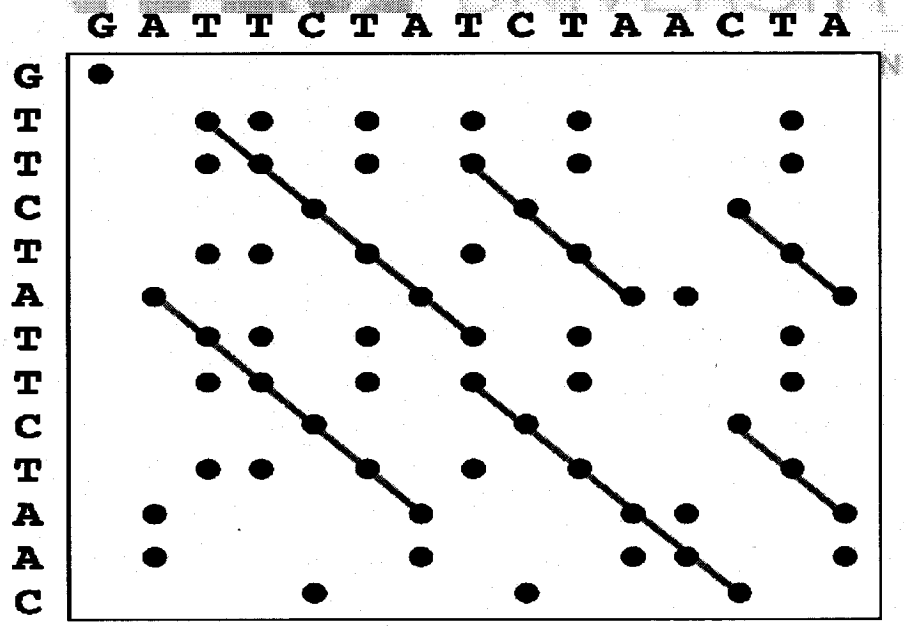
1) Dot Plot method

2) Dynamic Programming method

3) Heuristic Method

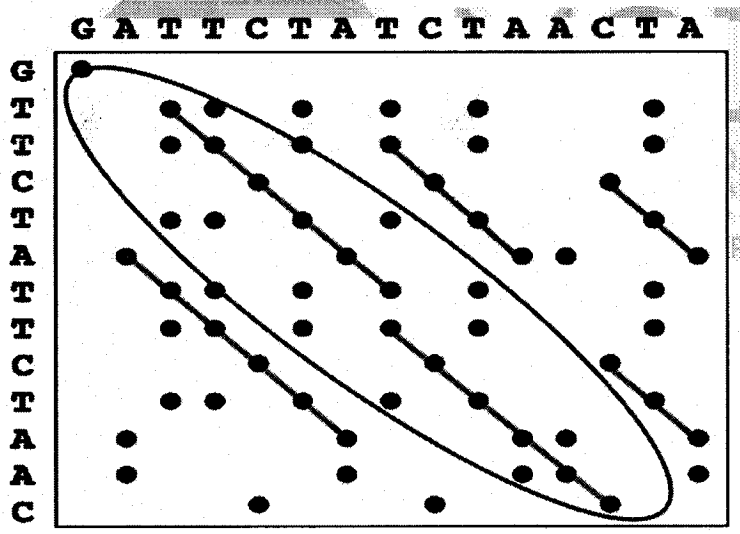
Dot Plot

- It is graphical method used for aligning two sequences for the identification of the region of maximum similarity and dissimilarity, depicted by the presence or absence of DOTS.
- Also known as the dot matrix method.
- The comparison is done by scanning each and every residue of one sequence for similarity with all residues in the other sequence.



MODY UNIVERSITY
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

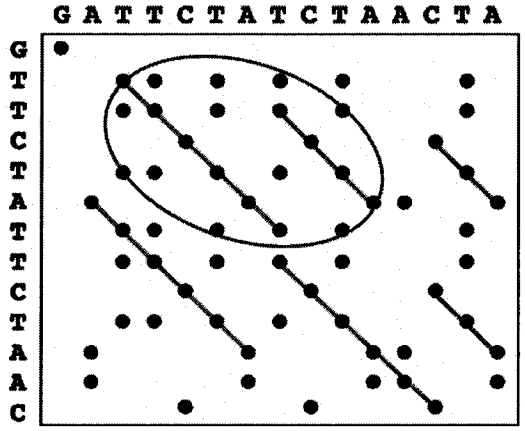
- We pick one residue from our query sequence and compare it with the residues on the X axis.
- We plot a dot where ever we find the same residue along the X axis.
- When the two sequences have substantial regions of similarity, many dots line up to form continuous diagonal lines, which reveal the sequence alignment as shown in example.
- If there are interruptions in the middle of a diagonal line, they indicate insertions or deletions in the sequences.
- Based on the patterns of lines in the matrix, we can say the following:
- If we obtain a Principal diagonal line, it will show identical sequence.



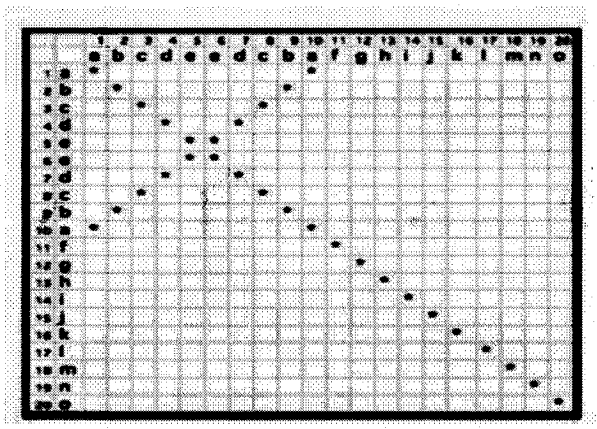
28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- If we have Multiple diagonal lines, they indicate repetition in the sequences.



- **Reverse diagonal (perpendicular to diagonal) indicate INVERSIONs**



MODY UNIVERSITY
HIMANGARH • RAJASTHAN

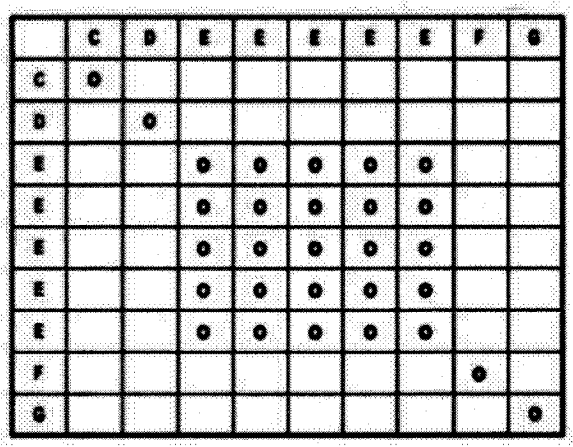
- **Reverse diagonal crossing diagonal (X) indicate PALINDROMES or the sequences are palindromic.**

	A	T	G	C	G	T	A
A	*						*
T		*				*	
G			*		*		
C				*			
G			*		*		
T		*				*	
A	*						*

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- **If we obtain lines that Form of a box as shown in this example, they indicate the Low complexity region**



Limitations of Dot Matrix method

- It lacks statistical rigor in assessing the quality of the alignment.
- It is difficult for the method to scale up to multiple alignment.
- For longer sequence, memory required for the graphical representation is very high. So long sequence can not be aligned.
- Aligning larger sequences will take longer time

Applications of Dot Matrix method

- Shows the all possible alignment between two nucleic acid and amino acid sequences
- All kind of local and global alignment can be traced
- Help to recognize large region of similarity.
- An excellent approach for finding sequence transposition.
- To find the non sequential alignment.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

SOFTWARES

- Dotmatcher : https://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher
- Dottup : https://www.ebi.ac.uk/Tools/seqstats/emboss_dottup/
- MatrixPlot : http://www.cbs.dtu.dk/services/MatrixPlot/
- Dotter : https://sonnhammer.sbc.su.se/Dotter.html
- Dothelix

Dynamic Programming

- This method basically breaks down a large problem into small sub-problems and then uses the solution of the sub-problems to compute the solution of the larger problem.
- Used for global as well as for local sequence alignments using Needleman-Wunsch and Smith-Waterman algorithms respectively.

Heuristic Method

When a single sequence is to be compared against the whole database, we use heuristic methods like BLAST and FASTA.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References:

- Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press
- Lesk, A. M. (2002). Introduction to bioinformatics. Oxford: Oxford University Press.
- Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/24

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

7: SEQUENCE ALIGNMENT (PART 2)

Dynamic programming methods

- Match: this means that there is no change since their divergence. The residue of the query sequence is the same as the residue in the target sequence.
- Once we get a match, we assign some value to it, which we call as a match score
- Mismatch: this means that there is a substitution since their divergence. The residue of the query sequence is the different from the residue in the target sequence.
- It is quite common that when comparing sequences we will find that at some positions in some of the sequences there will be some extra residues which we call insertions, or missing residues which we call deletions.
- A gap in one of the sequences simply means that one or more residues have been deleted from the sequence, or we could also say that there is an insertion of residues in the second sequence.
- When aligning sequences, introducing gaps in the sequences can allow an alignment algorithm to match more terms in comparison to a gap-less alignment can.
- to get a useful alignment, minimizing gaps in an alignment is important.
- Too many gaps can lead to a meaningless alignment
- Gap penalties are used to adjust alignment scores based on the number and length of gaps.
- It can be changed each time an alignment is run. This will affect the number of gaps, their length and position in the sequence alignment.

DYNAMIC PROGRAMMING METHOD

- A method that determines optimal alignment by matching two sequences for all possible pairs of characters between them.
- It is a more quantitative method. In this a dot matrix is converted into a scoring matrix to account for matches and mismatches between sequences.
- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/24

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

STEPS:

- **Initialization:**

The first step in the global alignment dynamic programming approach is to create a matrix with M + 1 columns and N + 1 rows where M and N correspond to the size of the sequences to be aligned.

M is no.of residues in target sequence

N means no.of residues in query sequence.

Matrix filling(scoring):

We fill the matrix with highest possible scores. For the calculation of the scores, we need the match score, mismatch score and the gap penalty. We have three calculations based on which we decide the highest score to be filled in the matrix.

These calculations are:

$$M_{i,j} = Maximum\ [M_{i-1,j-1} + S_{i,j},\ M_{i,j-1} + W,\ M_{i-1,j} + W]$$

Traceback and aligning :

Move from last corner and follow arrow.

Now lets consider the two given sequences.

ACTGATTCA

ACGCATCA

The length of the sequence 1 and sequence 2 are 9 and 8 respectively.

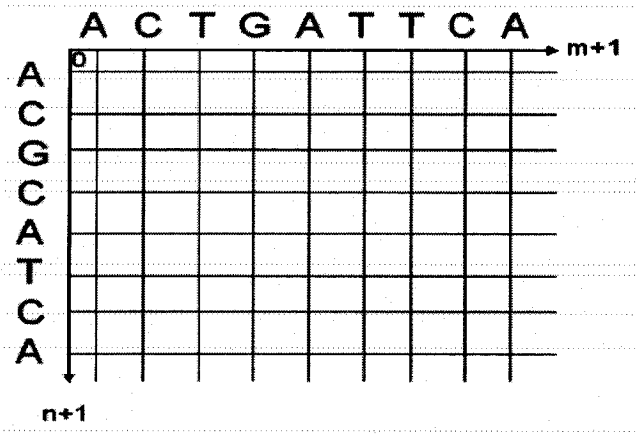
For match..we will give a score of 2, for mismatch -3 and gap penalty is -2

This is the M+1 and N+1 matrix

Then the scoring schema has to be introduced which can be user defined with specific scores.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



Initialization Step

Let's assumes that there is gap penalty.

First row and first column of the matrix can be initially filled with 0.

If the gap score is assumed, the gap score can be added to the previous cell of the row or 0^{th} row and to the 0^{th} column as shown here.

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

28/11/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Matrix Fill Step

The second and crucial step of the algorithm is matrix filling starting from the upper left hand corner of the matrix.

From the assumed values, add the match or mismatch (assumed) score to the diagonal value.

$$M_{1,1} = Max\ [M_{0,0} + S_{1,1},\ M_{1,0} + W,\ M_{0,1} + W]$$

•

Lets generalize this equation

M(I,J)= M(i-1,j-1)+S(I,j)

M(i,j-1)+W

M(I-1, j)+W

M(I,j)- coordinates of the cell

S(I,j)- score corresponding the cell of given coordinate (whether it's a match or a mismatch)

W- gap penalty

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

We compute all the values and choose the highest score

In this example, the value of M(i-1,j-1) ie M(0,0) is 0

To this add the score of match or mismatch

We will check whether the jth residue of our query sequence is matching with the ith residue of the target sequence.

If there is a match we will add the score of match.

In this example, yes there is a match, so we will add the score of match which is 2

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

		A
	0	-2
A	-2	2

In second equation we will look at the score in the M(I,j-1) cell

Here the value at M(1,0) is -2

And the gap panelty is -2, so the value will be -4

In equation 3 value at M(0,1) is again -2 and the gap penalty is also -2 so the value comes out to be -4

Out of all these 3 values 2 is the highest value and hence the score at M(1,1) will be 2

This is how the final matrix will appear.

Step 2: Assign scores

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

28/11/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

Trace back Step

The final step in the algorithm is the trace back for the best alignment.

We will start our trace back from the cell with the highest score.

In this example, we have 8 at the right most bottom corner of the matrix.

We will check for the next highest value adjacent to this cell.

We have values of 4,6 and 4 in the adjacent cells

Out of these 6 is the highest value which is in the diagonally adjacent cell.

After 6 we will again look for the highest value in the same manner.

We have 4 which is also in the diagonally adjacent cell.

Once again lets check for the highest value.

Now here we have 6 which is the value of the horizontal left cell.

MODY UNIVERSITY

Step 3: Trace back
The optimal path is traced beginning from the lower right-hand corner

		A	C	T	G	A	T	T	C	A
	(0)	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	(2)	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	(4)	(2)	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	(4)	2	0	-2	-4	-6
C	-8	-4	0	-1	(2)	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	(4)	2	0	-2	2
T	-12	-8	-4	0	-2	2	(6)	(4)	2	0
C	-14	-10	-6	-2	-4	0	4	2	(6)	4
A	-16	-12	-8	-4	-5	-2	2	1	4	(8)

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Aligning the sequences

The diagonal lines indicate alignment

```
CA
| |
CA
```

If we are moving horizontally left, like here from 4 to 6 we will insert a gap in the sequence on the Y axis

```
?TCA
 | |
?-CA
```

And if we are moving vertically up, we will insert a gap in the sequence at the X axis or our target sequence

```
-ATTCA
 || ||
CAT-CA
```

So we have alligned A with A, C with C, then we have a gap in our query sequence so this is T of target sequence and a gap in query sequence. Then T is aligning with T, A with A and then this time we have a gap in our target sequence against C in the query sequence.

Then G is aligning with G and again we have a gap in our query sequence against T of the target sequence. Then C is aligning with C and A is in alignment with A.

Result:

This analysis yielded the following alignment:

```
ACTG-ATTCA
||   || ||
AC-GCAT-CA
```

The alignment score is equal to the value in the lower right-hand corner of the matrix (8).

- **Softwares which are used for pairwise sequence alignment are**
- **SIM, (http://bioinformatics.iastate.edu/aat/align/align.html)**
- **SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html)**
- **LALIGN (www.ch.embnet.org/software/LALIGN form.html)**

24/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8: MULTIPLE SEQUENCE ALIGNMENT

A Multiple Sequence Alignment is an alignment of more than two sequences.

Multiple alignments are computationally much more difficult than pair-wise alignments.

There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments can

Multiple sequence alignment (MSA) methods refer to a series of algorithmic solution for the alignment of evolutionarily related sequences, while taking into account evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions.

These methods can be applied to DNA, RNA or protein sequences.

Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment.

Multiple sequence alignment is also an essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.

Multiple sequence alignment also has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

MSA is indeed an important modeling tool whose development has required addressing a complex combination of computational and biological problems.

8.1: Commonly used algorithms

Given a set of biological sequences (RNA, proteins, DNA), the purpose of a MSA method is to align the sequences in a way that will either reflect their evolutionary, functional or structural relationship Usually the multiple sequence algorithms assume that the sequences are similar in all its length and they behave like global alignment algorithms.

They also assume that there are not many long insertions and deletions. Thus the algorithms will work for some sequences, but not for others.

there are exhaustive and heuristic approaches used in multiple sequence alignment.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

exhaustive approaches are not feasible in most cases, heuristic programs are commonly used

8.2: Progressive Alignment Method

Progressive alignment depends on the stepwise assembly of multiple alignment and is heuristic in nature.

It first conducts pairwise alignments for each possible pair of sequences using the Needleman–Wunsch global alignment method and records these similarity scores from the pairwise comparisons.

The scores can either be percent identity or similarity scores

Both scores correlate with the evolutionary distances between sequences.

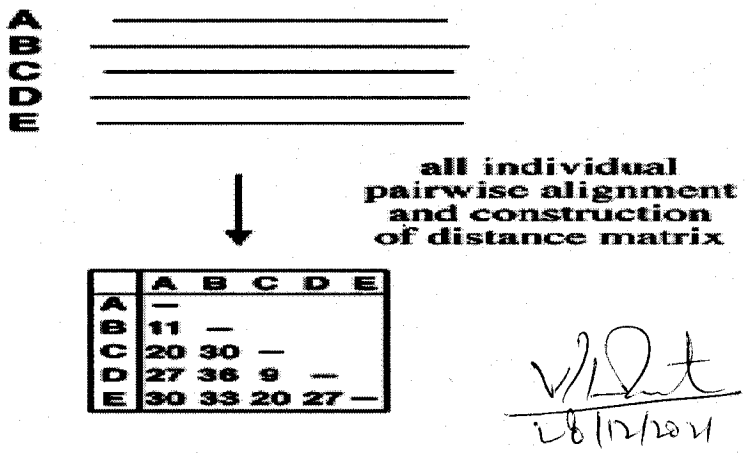
The scores are then converted into evolutionary distances to generate a distance matrix for all the sequences involved.

A simple phylogenetic analysis is then performed based on the distance matrix to group sequences based on pairwise distance scores. As a result, a phylogenetic tree is generated.

The tree reflects evolutionary proximity among all the sequences.

the resulting tree is an approximate tree and does not have the rigor of a formally constructed phylogenetic tree.

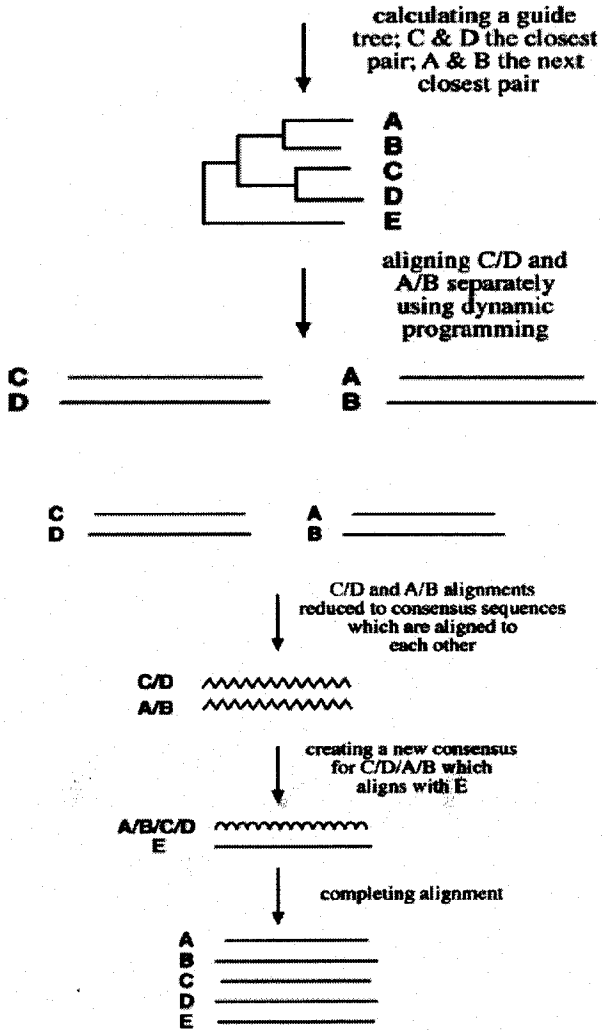
we have 5 different sequences A, B, C, D and E



	A	B	C	D	E
A	—				
B	11	—			
C	20	30	—		
D	27	36	9	—	
E	30	33	20	27	—

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332311



8.3: Softwares for MSA

Clustal (www.ebi.ac.uk/clustalw/) is a progressive multiple alignment program available either as a stand-alone or on-line program.

CLUSTAL 2 has two versions ClustalW and ClustalX. The W version provides a simple text-based interface and the X version provides amore user-friendly graphical interface.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8.3.1: CLUSTAL W

In Clustal W we can select the type of the output format

Multiple Sequence Alignment by CLUSTALW

ETE3 | MAFFT | CLUSTALW | PRRN

Help

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: ◉ FAST/APPROXIMATE ○ SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): ◉ PROTEIN ○ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Choose File | No file chosen

Execute Multiple Alignment | Reset

28/12/2014

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1, Window size: 5, Gap Penalty: 3

Number of Top Diagonals: 5, Scoring Method: PERCENT

For SLOW/ACCURATE:

Gap Open Penalty: 10.0, Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN)

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

Multiple Alignment Parameters:

Gap Open Penalty: 10, Gap Extension Penalty: 0.05

Weight Transition: ○ **YES (Value:** 0.5 **),** ◉ **NO**

Hydrophilic Residues for Proteins: GPSNDQERK

Hydrophilic Gaps: ◉ **YES** ○ **NO**

Select Weight Matrix: BLOSUM (for PROTEIN)

Type additional options (delimited by whitespaces) below:

(**-options** for help)

Execute Multiple Alignment | Reset

The type of the pairwise alignment that we need ie FAST and approximate or slow and accurate.

We then choose the type of sequence we want to enter id DNA or protein

CLUSTAL W also provides some options in parameters that users can change as per their requirements like the scores for gap opening and gap extension, the type of matrix to be used.

In case of fast and accurate pairwise alignment, the user can change the range of K-tuple, window size, gap panelty etc.

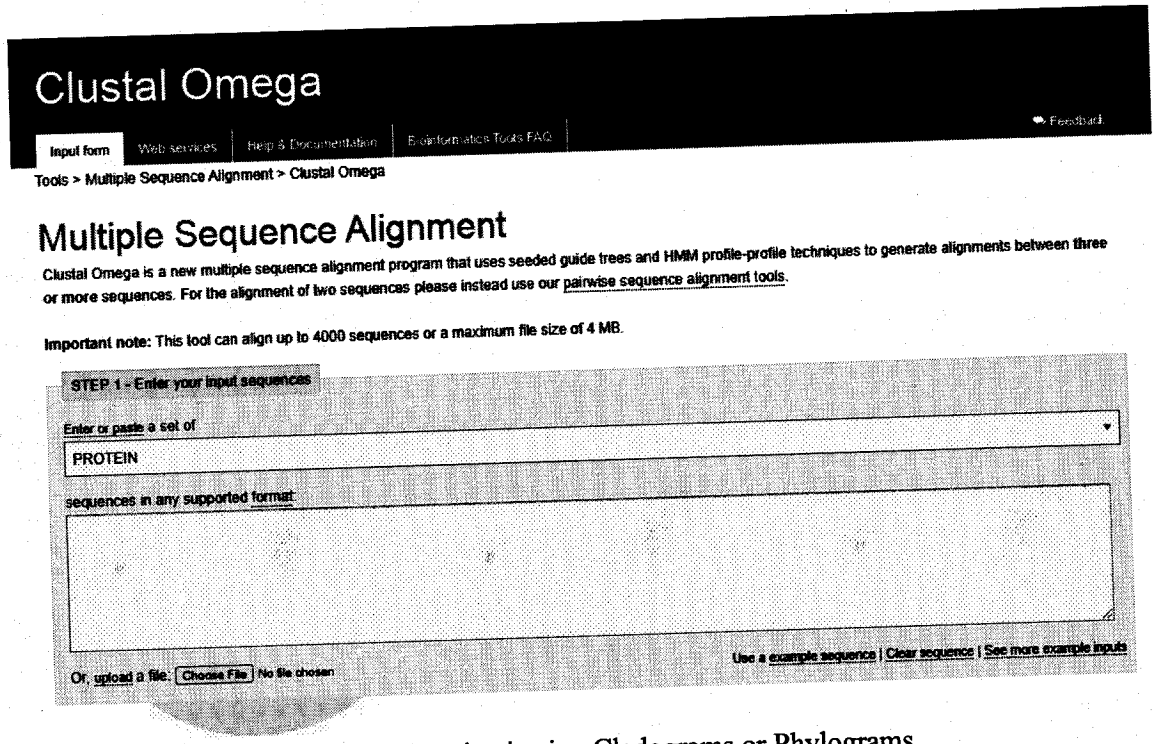
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8.3.2: CLUSTAL Omega

Clustal Omega is a multiple sequence alignment program that uses seeded guide trees and HMM techniques to generate alignments between three or more sequences.

It produces biologically meaningful multiple sequence alignments of divergent sequences.



Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

In the home page we have Sequence Input Window

Select the type of sequences to be aligned ie Protein sequences, DNA sequences or RNA sequences

Three or more sequences to be aligned can be entered directly into this box

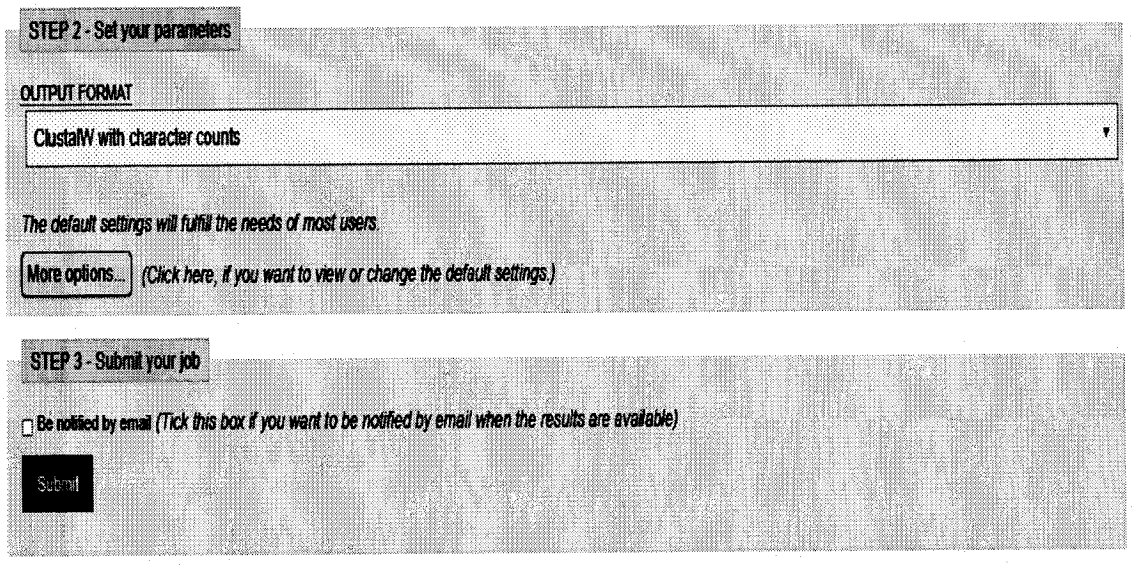
Sequences can be in GCG, FASTA, EMBL (Nucleotide only), GenBank, PIR/NBRF, PHYLIP or UniProtKB/Swiss-Prot (Protein only) format.

One can also upload the sequence files.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- **Step 2 - Set Your Parameters**



There are various parameter options for the users to obtain the alignment most suited for their requirements.

One can also choose the format of the output file generated.

And the last step is submission

8.3.3: EMBOSS Cons

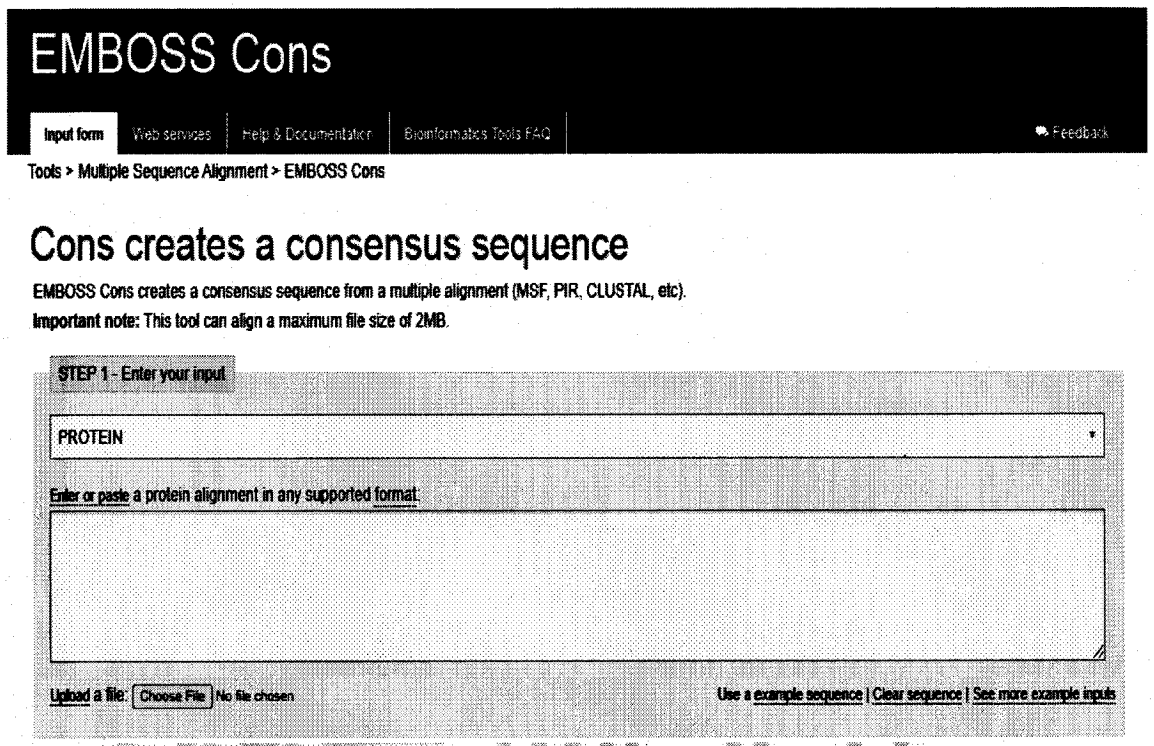
EMBOSS Cons creates a consensus sequence from a multiple alignment

This tool can align a maximum file size of 2MB.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

V/1 Dut
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

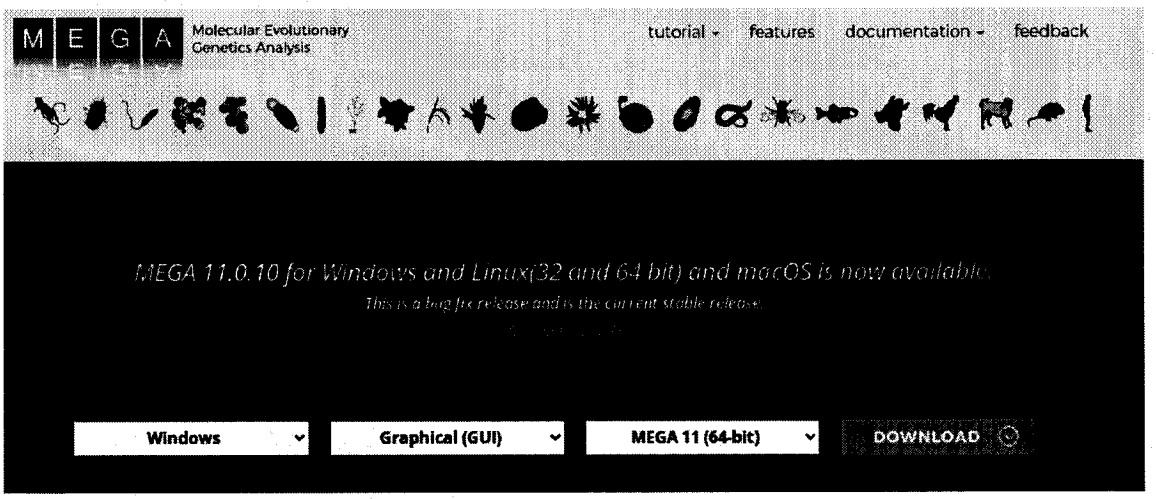
8.3.4: MEGA

It stands of molecular evolutionary genetics analysis

MEGA can be used with either a graphical user interface or a command-line interface which is useful for batch or scripted execution.

MEGA supports sequence alignment using both the ClustalW and MUSCLE programs.

22/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



Alignment (or refinement) is done in the Analysis Explorer

Once the sequences are in the Alignment Explorer (AE), we select Alignment either *ClustalW* or Muscle.

Set the alignment parameters to the values you wish or we can go with the default values. Click Compute/OK

Depending on the length and number of sequences you may see a progress bar while the alignment is running.

Once the alignment is done, we can export them to MEGA or Fasta format for analysis.

8.3.5: MAFFT

It is a MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments

Multiple alignment program for amino acid or nucleotide sequences

MAFFT is a multiple sequence alignment program for unix-like operating systems.

It offers a range of multiple alignment methods, L-INS-i, FFT-NS-2, *etc.*

28/12/24

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

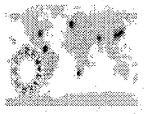
MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Download version
- Mac OS X
- Windows
- Linux
- Source

Online version
- **Alignment**
- mafft --add
- Merge
- Phylogeny
- Rough tree

Merits / limitations
Algorithms
Tips
Benchmarks
Feedback



To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

The Q-INS-i option (for RNA) did not work because of a bug in version 7.489, Oct.23. Fixed in version 7.490, Oct.30, 2021.

A compatibility problem of MSAViewer with Chrome 92 and Firefox 90 has been fixed, Aug. 5, 2021.

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

Multiple sequence alignment and NJ / UPGMA phylogeny

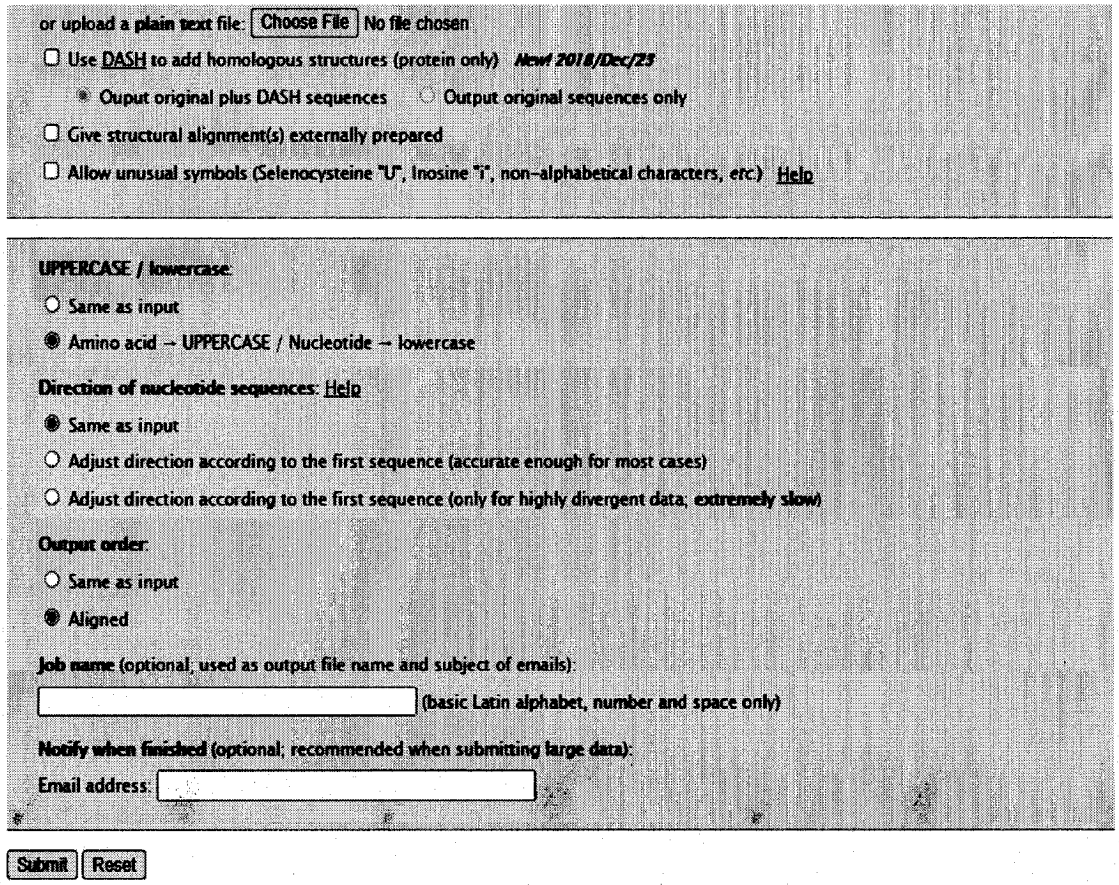
Input

Paste protein or DNA sequences in fasta format. Example

LAKSHMANGARH • RAJASTHAN

29/12/24

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



LAKSHMANGARH • RAJASTHAN

Merits

- (Accuracy): L-INS-i is one of the most accurate multiple sequence alignment methods currently available. L-INS-i is in particular suitable to align 10-100 protein sequences
- (Scalability): FFT-NS-2 and other progressive methods used in MAFFT can align many long DNA/protein sequences
- The scoring system was designed to allow large gaps.

8.3.6: MUSCLE

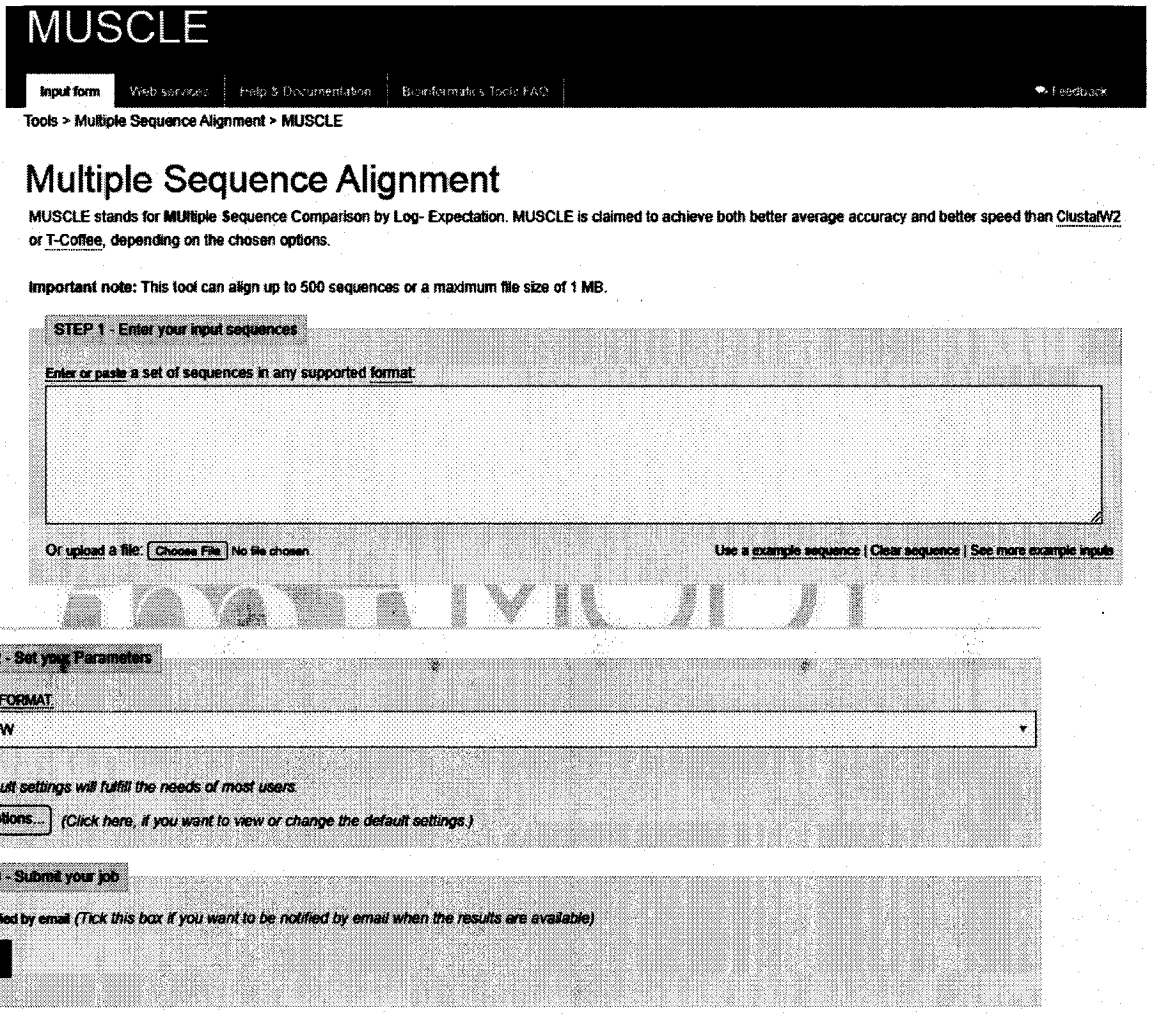
MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation.

MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee,

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt. Sikar) 332311

This tool can align up to 500 sequences or a maximum file size of 1 MB in a second.

MUSCLE enables high-throughput applications to achieve average accuracy



8.3.7: KALIGN

Very fast MSA tool that concentrates on local regions..

It is a fast and robust alignment method, especially well suited for the increasingly important task of aligning large numbers of sequences.

Suitable for large alignments

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Kalign

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback

Tools > Multiple Sequence Alignment > Kalign

Multiple Sequence Alignment

A fast and accurate multiple sequence alignment algorithm.

Important note: This tool can align up to 2000 sequences or a maximum file size of 2 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your Parameters

OUTPUT FORMAT

ClustalW

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332311

Important links

https://mafft.cbrc.jp/alignment/server/

https://www.ebi.ac.uk/Tools/msa/emboss_cons/

https://www.ebi.ac.uk/Tools/msa/clustalo/

https://www.genome.jp/tools-bin/clustalw

https://www.megasoftware.net/

https://www.ebi.ac.uk/Tools/msa/kalign/

https://www.ebi.ac.uk/Tools/msa/muscle/

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

9: Basic local alignment search tool (BLAST)

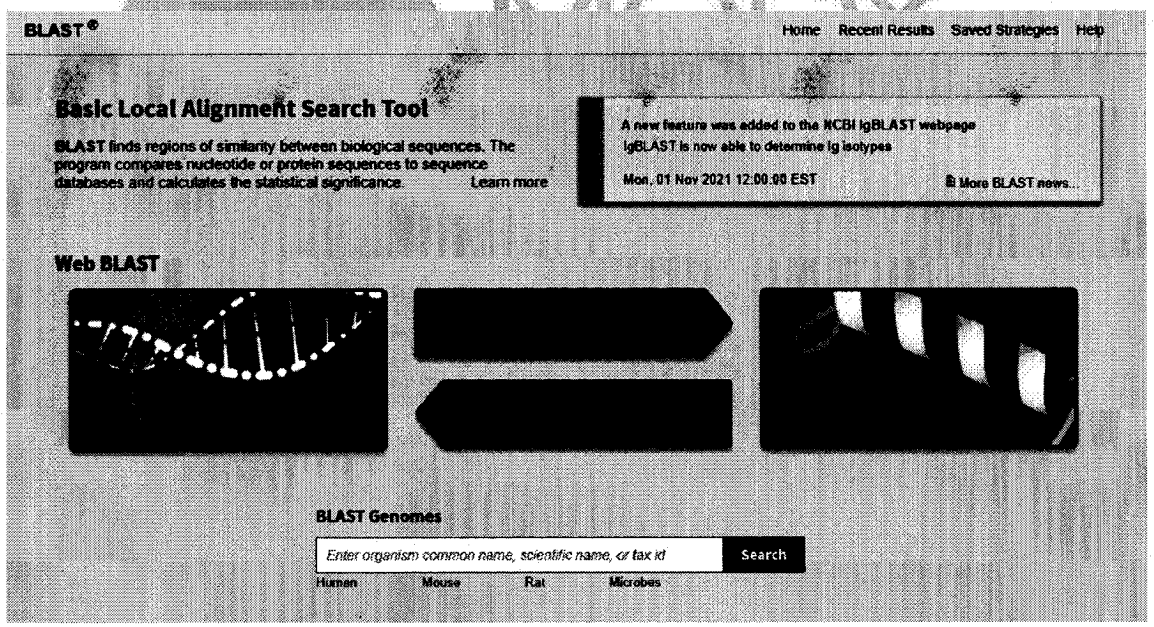
BLAST stands for BASIC LOCAL ALIGNMENT SEARCH TOOL.

It helps in finding the similarities between biological sequences. It is a complete package which helps the researchers in various aspects for comparing and identifying DNA, RNA and protein sequences. BLAST suite offers multiple services like:

Web BLAST offers platforms for searching the protein and nucleotide sequences against a protein or nucleotide database.

BLAST also provides **command line applications** to perform searches locally.

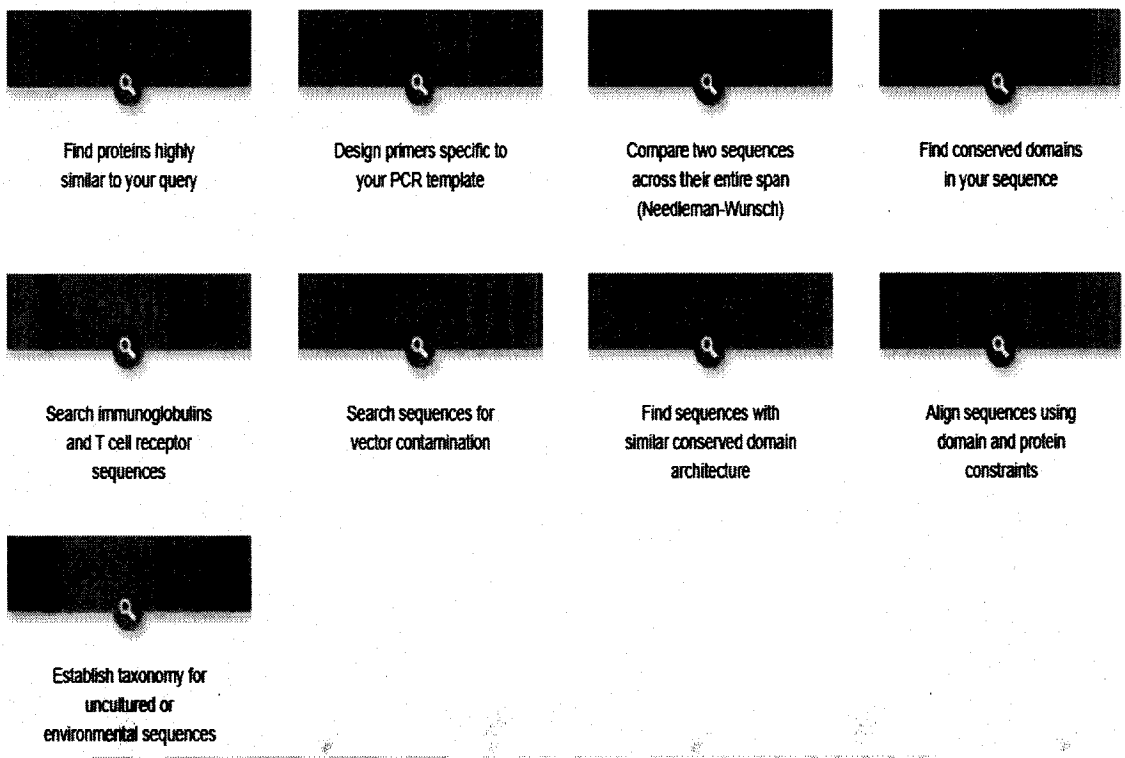
In BLAST suite, one can also find specialized services like SMART BLAST, PRIMER BLAST, GLOBAL ALIGN, CD SEARCH, Ig BLAST, VECSCREEN, CDART, MULTIPLE ALIGN and MOLE BLAST.



8/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Specialized searches



UNIVERSITY
LAKSHMANGARH • RAJASTHAN

SMART BLAST

It is a new tool of NCBI that is used for sequence analysis tasks like finding a candidate protein name for a sequence, locating regions of high sequence conservation, or identifying regions covered by database sequences but missing from the query.

To perform all such tasks, it takes much less time as compared to the protein blast.

PRIMER BLAST

It helps in designing the primers that are specific for the PCR template of our choice.

GLOBAL ALIGN

This helps in aligning two sequences globally i.e end to end alignment is done.

REGISTRAR
Mody University of Science and Technolo
Lakshmangarh (Distt Sikar) 332311

This is done using Needleman-Wunsch algorithm which we have already discussed in our previous lecture.

CD SEARCH

CD search is Conserved Domain Search service

It is a web-based tool for the detection of structural and functional domains in protein sequences.

Ig BLAST

Is A tool for immunoglobulin (IG) and T cell receptor (TR) Variable domain sequences

With this tool, we can view the matches to the germline V, D and J genes, we can find details at rearrangement junctions, we can also find the delineation of immunoglobulin Variable domain framework regions and complementarity determining regions.

VECSCREEN

VecScreen is a system that quickly finds segments of a nucleic acid sequence that may be of vector origin.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

It is recommended to researchers as its helps in identifying and removing any segments of vector origin before they analyze or submit sequences

CDART

Finds sequences with similar conserved domain architecture

28/12/2024

MULTIPLE ALIGN

It is a constraint based multiple alignment tool.

It finds a collection of pairwise constraints derived from conserved domain database, protein motif database, and sequence similarity, using RPS-BLAST, BLASTP, and PHI-BLAST

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

MOLE BLAST

MOLE-BLAST is an experimental tool that helps taxonomists find closest database neighbors of submitted query sequences

It computes a multiple sequence alignment (MSA) between the query sequences along with their top BLAST database hits, and generates a phylogenetic tree.

The BLAST suite also offers downloadable link for BLAST through which one can get a standalone version of BLAST along with the databases.

WEB BLAST

WEB BLAST is the most commonly used section of BLAST which offers multiple versions to compare all combinations of nucleotide or protein queries against a nucleotide or protein database.

The versions are namely nucleotide BLAST or blastn, Protein BLAST or blastp, translated BLASTs blastx, tblastn and tblastx

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

Nucleotide BLAST or blastn

BLASTN program searches nucleotide databases using a nucleotide query.

It is a general purpose nucleotide search and alignment program that can be used to align tRNA or rRNA sequences as well as mRNA or genomic DNA sequences containing a mix of coding and noncoding regions.

Protein BLAST or blastp

28/12/2021

Protein-to-protein sequence searches are performed using blastp or protein blast.

The default wordsize for a blastp search is three.

During a "blastp" search, low-complexity regions of the query sequence are filtered to reduce the construction of spurious alignments and enhance search speed.

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Translated BLAST

Translated BLAST searches use a genetic code to translate either the "query," database "subjects," or both, into protein sequences, which are then aligned as in "blastp.

There are three varieties of translated BLAST search;

- "tblastn,"
- "blastx," and
- "tblastx."

Blastx

In this type of translated BLAST, a nucleotide sequence query is translated in six reading frames, and the resulting six-protein sequences are compared, in turn, to those in a protein sequence database.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

Tblastn

In this a protein sequence query is compared to the six-frame translations of the sequences in a nucleotide database.

As we are searching translated sequences in nucleotide database, we call it T blast N

Tblastx

In this BOTH, the "query" and database "subject" nucleotide sequences are translated in six reading frames, after which 36 protein "blastp" comparisons are made.

Six-frame translations for query sequence and six-frame translations for subject nucleotide sequence

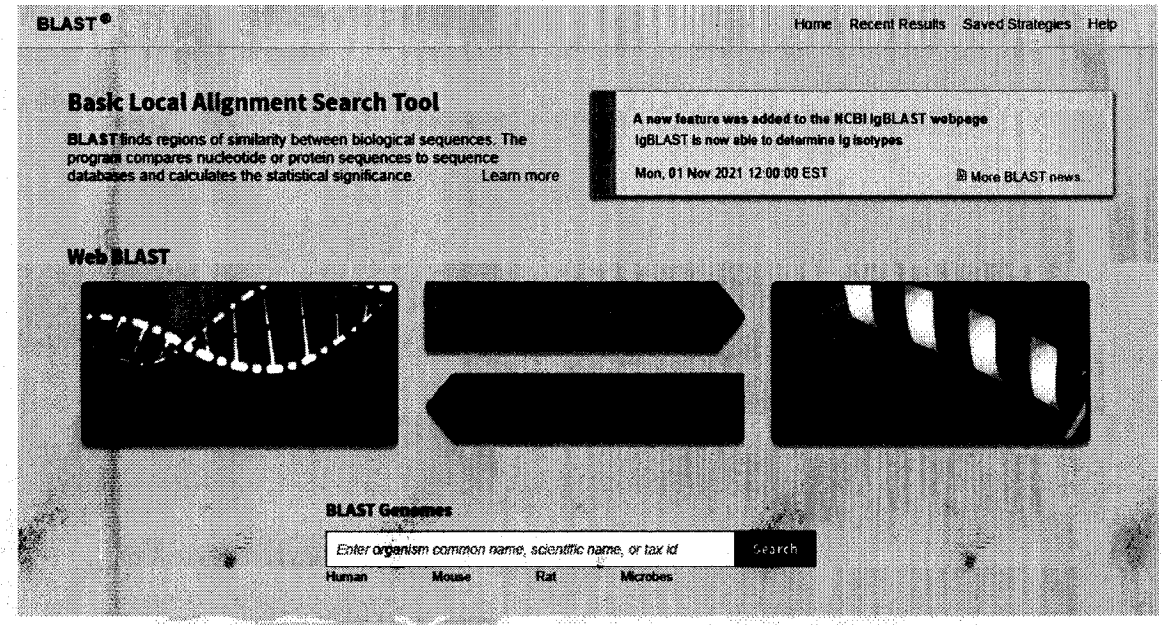
28/12/2021

In web blast, we also find BLAST GENOMES which is a graphical tool for simplifying BLAST searches against complete and unfinished genome sequences.

REGISTRAR
Mody University of Science and Technolo
Lakshmangarh (Distt Sikar) 332311

This tool is used to compare the query sequence against a virtual database of DNA sequences or protein sequences from a selected group of organisms with finished or unfinished genomes.

9.1: BLAST software



Recent Results which lists a set of links to recently completed search results.

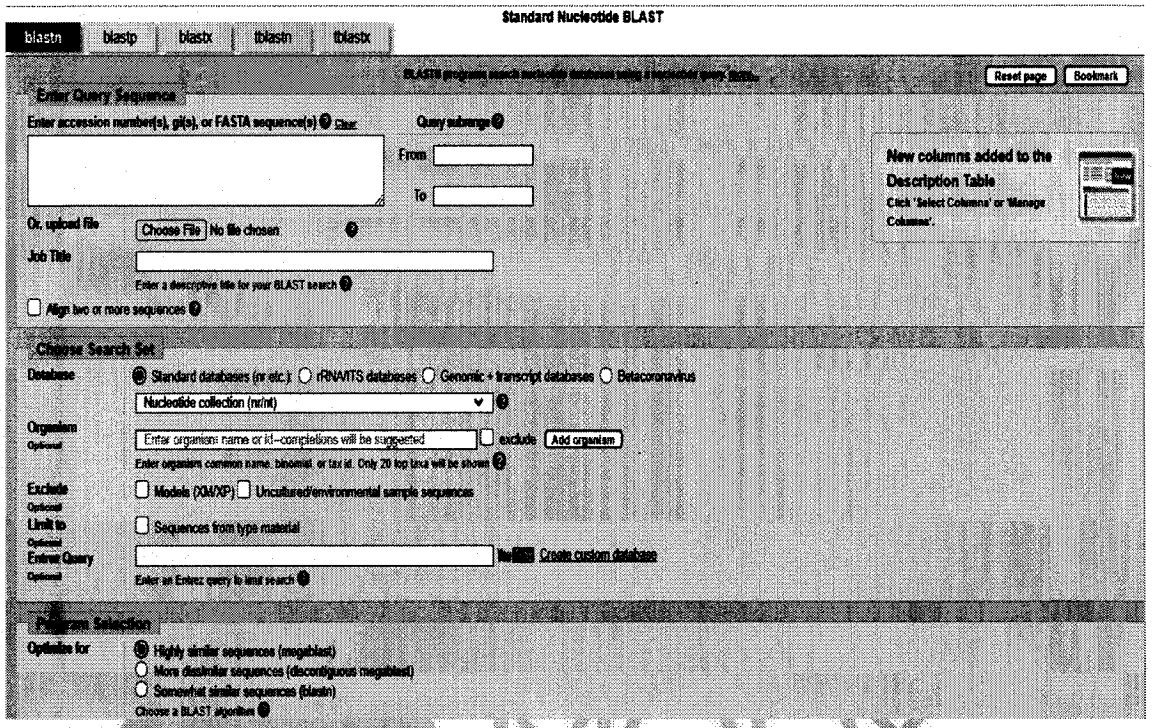
We have "**saved strategies**" which displays a list of search strategies which we have saved earlier. On home page, we can find links to various types of services provided in the BLAST suite.

9.1.1: Nucleotide BLAST

On the homepage of blast, click on nucleotide blast

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



Here on the first section, we provide the query sequence.

We can provide query sequence, either by entering the accession number(s) or by providing the gi number(s), or by providing the FASTA sequence(s).

We can also upload a file from our local disk. The file may contain a single sequence or a list of sequences and the data can be either a list of database accession numbers, NCBI gi numbers, or sequences in FASTA format.

Now here Enter a descriptive title for your BLAST search

We can also align two or more sequences. For this we can Enter one or more queries in the top text box and one or more subject sequences in the lower text box and click on the BLAST button at the bottom.

Now the next segment is to **Choose Search Set**

From here we can choose the database against which we want to search our query sequence.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

We can search against standard databases or rRNA/ITS databases or genomic+ transcript databases and now also against beta corona virus genbank .

Then we can choose the list of organisms that we want to exclude from our search.

This is optional. Like wise we have few more options available to restrict our search like we can exclude the model sequences or uncultured experimental sequences.

Now next we can choose the type of the program we want to run. Whether we want to run megablast or discontinuous megablast or blastn. By choosing the program, we are selecting the algorithm we want to run for our similarity search. Before hitting the BLAST button lets have a look at the additional parameters. We can leave these parameters with their default values.

So here, in parameter section, we can choose the maximum no.of target sequences, the default value of which is 100

Short queries. If the entered sequence is a short query sequence then the algorithm will automatically adjust word size and other parameters to improve results for short queries.

Expected threshold, the default value of which is 0.05. This setting specifies the statistical significance threshold for reporting matches against database sequences

In the same way we have other parameters like word size, match mismatch score, gap cost and we have certain filters and masks.

Now once we have adjusted all the parameters and settings as per our requirement, we will hit the BLAST button

Now we are moving towards the results page. BLAST takes few seconds to generate the results page

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

20/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Format Request Status

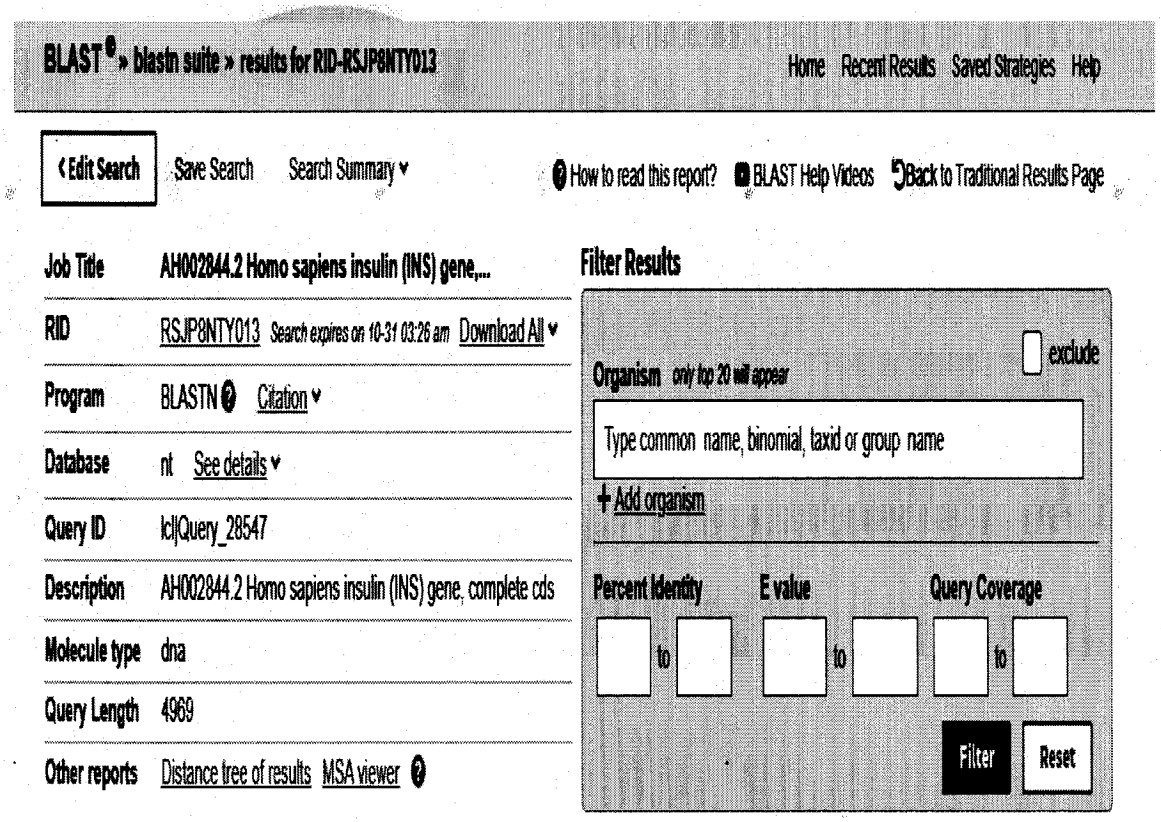
Formatting options]

Job Title: AH002844.2 Homo sapiens insulin (INS) gene,...

Request ID	RSJP8NTY013
Status	Searching
Submitted at	Fri Oct 29 15:26:00 2021
Current time	Fri Oct 29 15:26:15 2021
Time since submission	00:00:14

This page will be automatically updated in 2 seconds

The results page of BLAST



Job Title shows the first query's sequence id by default

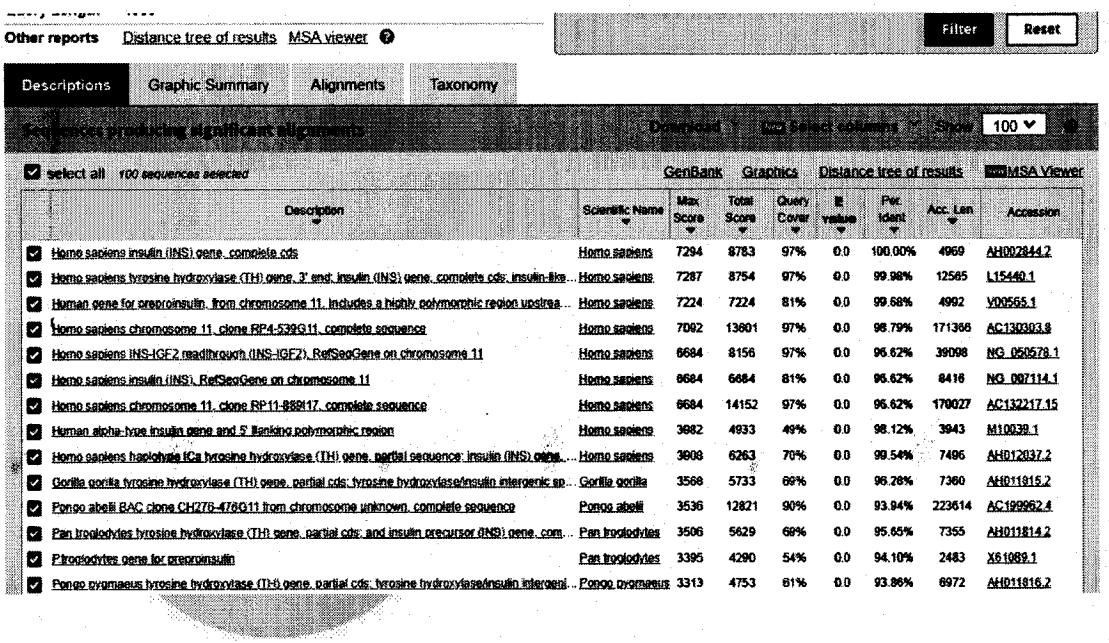
RID shows the assigned unique identifier to this search for sharing and problem reporting

29/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Program tells the search conducted. Database names the target database searched. Query ID is the query's sequence id for which the result is shown. Description is the title of the query sequence from its FASTA defline. Molecule type shows the type of the sequence. Query length is the length of the current query. From here we can also refine our results. The Descriptions tab contains a summary table of hits found by BLAST.

Default sorting is by E value column



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens insulin (INS) gene, complete cds	Homo sapiens	7294	8783	97%	0.0	100.00%	4969	AH002844.2
Homo sapiens tyrosine hydroxylase (TH) gene, 3' end, insulin (INS) gene, complete cds; insulin-like...	Homo sapiens	7287	8754	97%	0.0	99.98%	12565	L15440.1
Human gene for preproinsulin, from chromosome 11, includes a highly polymorphic region upstrea...	Homo sapiens	7224	7224	81%	0.0	99.68%	4992	V00565.1
Homo sapiens chromosome 11, clone RP4-539G11, complete sequence	Homo sapiens	7092	13601	97%	0.0	98.79%	171366	AC130303.8
Homo sapiens INS-IGF2 readthrough (INS-IGF2), RefSeqGene on chromosome 11	Homo sapiens	6684	8156	97%	0.0	96.62%	39098	NG_050578.1
Homo sapiens insulin (INS), RefSeqGene on chromosome 11	Homo sapiens	6684	6684	81%	0.0	96.62%	8416	NG_007114.1
Homo sapiens chromosome 11, clone RP11-888I17, complete sequence	Homo sapiens	6684	14152	97%	0.0	96.62%	178027	AC132217.15
Human alpha-type insulin gene and 5' flanking polymorphic region	Homo sapiens	3982	4933	49%	0.0	98.12%	3943	M10039.1
Homo sapiens haplotype ICa tyrosine hydroxylase (TH) gene, partial sequence; insulin (INS) gene,...	Homo sapiens	3908	6263	70%	0.0	99.54%	7496	AH012037.2
Gorilla gorilla tyrosine hydroxylase (TH) gene, partial cds; tyrosine hydroxylase/insulin intergenic sp...	Gorilla gorilla	3568	5733	69%	0.0	96.28%	7360	AH011915.2
Pongo abelii BAC clone CH276-476G11 from chromosome unknown, complete sequence	Pongo abelii	3536	12821	90%	0.0	93.94%	223614	AC199962.4
Pan troglodytes tyrosine hydroxylase (TH) gene, partial cds; and insulin precursor (INS) gene, com...	Pan troglodytes	3506	5629	69%	0.0	95.65%	7355	AH011814.2
Ptroglodytes gene for preproinsulin	Pan troglodytes	3395	4290	54%	0.0	94.10%	2483	X61069.1
Pongo pygmaeus tyrosine hydroxylase (TH) gene, partial cds; tyrosine hydroxylase/insulin intergeni...	Pongo pygmaeus	3313	4753	61%	0.0	93.86%	6972	AH011816.2

Graphical summary:

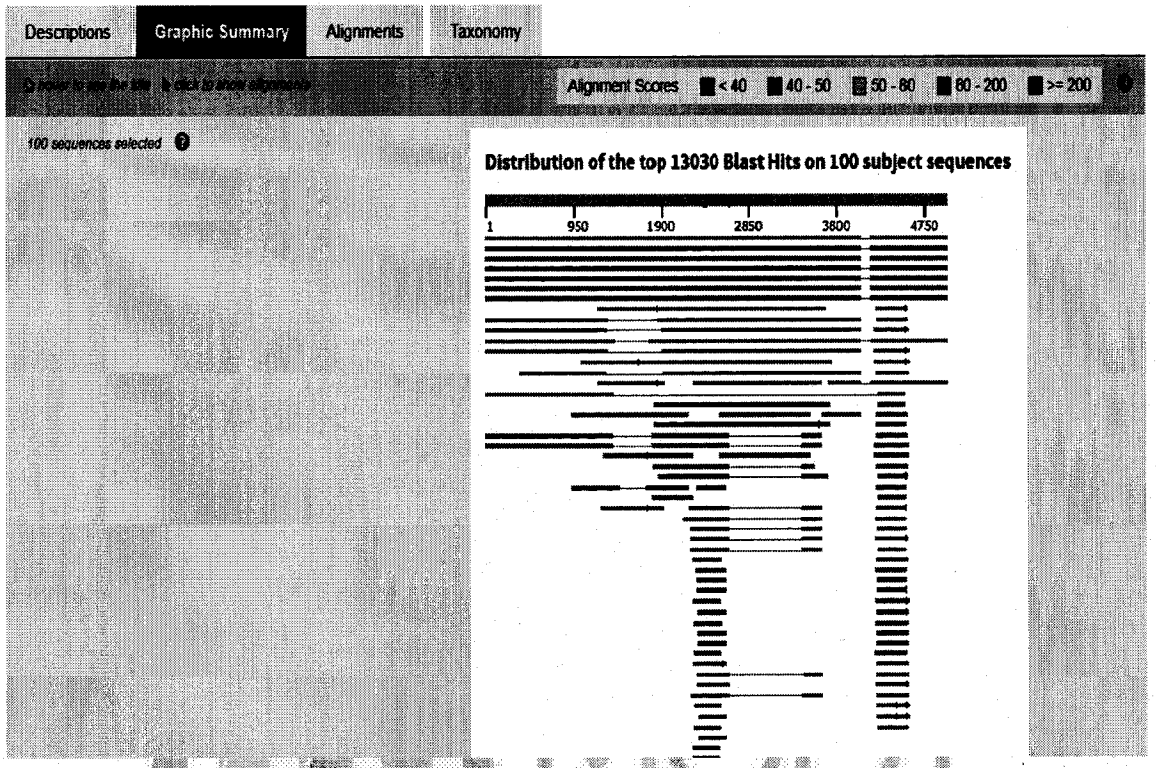
Here you can see various lines in different colours. These colours are on the basis of the alignment scores.

red represents the highest alignment scores

Generally, the higher the alignment score, the more significant the hit.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



The Alignments section contains the detailed pairwise alignments between query and database sequences.

It makes alternative displays of the alignment results readily accessible through a simple pull-down menu.

The sequence alignments show us how well our query sequence matches with the subject sequence in the database.

Each alignment block begins with a summary that includes the Max score and Expect value (the statistical significance of the alignment), sequence identity (number of identical bases between the query and the subject sequence), the number of gaps in the alignment, and the orientation of the query sequence relative to the subject sequence

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Descriptions | Graphic Summary | Alignments | Taxonomy

Alignment view: Pairwise | CDS feature | Restore defaults

100 sequences selected

Download | GenBank Graphics | Sort by: E value

Homo sapiens insulin (INS) gene, complete cds

Sequence ID: AH002844.2 **Length:** 4969 **Number of Matches:** 2

Range 1: 1 to 4044 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
7294 bits(8088)	0.0	4044/4044(100%)	0/4044(0%)	Plus/Plus

```
Query  1    CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCG  60

Query  61   GCCAGGGTGTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GCCAGGGTGTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  120

Query  121  TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGC  180

Query  181  TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACTCCTACTCATCCACGA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACTCCTACTCATCCACGA  240

Query  241  AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACCTTGT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACCTTGT  300

Query  301  TAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTGAAGGGCTGTGAGCCCCCAGG  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  TAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTGAAGGGCTGTGAGCCCCCAGG  360
```

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Download GenBank Graphics

Pongo pygmaeus insulin gene, partial cds

Sequence ID: AY092024.1 Length: 972 Number of Matches: 1

Range 1: 1 to 972 GenBank Graphics Next Match Previous Matc

Score	Expect	Identities	Gaps	Strand
1481 bits(1642)	0.0	922/989(93%)	24/989(2%)	Plus/Plus

```
Query  2513  GTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTA  2572
             ||||||||| ||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  1     GTGCGGCTCCCACCTGGTGGAAGCTCTCTACCTAGTGTGTGGGGAACGAGGCTTCTTCTA  60

Query  2573  CACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCATTGCTG  2632
              |||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
Sbjct  61    CACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCCTTGCTG  120

Query  2633  CCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTCCCACCCAGCATGGGCAGAAGG  2692
              ||||||| ||||||||||||| |||||||| ||| ||||||||||||||||||||||||
Sbjct  121   CCCCTGGTCGCCCCCAGCCATCCCCTGCTTCTGCCGCTCCCACCCAGCATGGGCAGAAGG  180

Query  2693  GGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAA-----GTT  2747
              |||||||||||||||||||||||||||||||||||||||||||||||||||     |||
Sbjct  181   GGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGTACTTTTTTAAAAAGAAATGAAGTT  240

Query  2748  CTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGA  2807
              |||||||||| |||||||||||||||||||||||||||||| |||||||
Sbjct  241   CTCTTGGTCACATCCTAAAAGTGACCAGTTCCCTGTGGCCCATTCAGAATG---------  291

Query  2808  GGACGGTGTTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCC  2867
                 ||   |||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  292   ----GG----GGCTTCAGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCC  343

Query  2868  ACTCGCCCCTCAAACAAATGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAG  2927
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  344   ACTCGCCCCTCAAACAAATGCCCCACAGCCCATTTCTCCACCCTCATTTGATGACCGCAG  403

Query  2928  ATTCAAGTGTTTTGTTAAGTA-AAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGC  2986
              |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  404   ATTCAAGTGTTTTGTTAAGTACAAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGC  463

Query  2987  TGCCTGCCTCTGGGCGAACACCCCATCACGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAG  3046
              |||||||||||||||||||||||||||||||| |||| ||||||||||||||||||||
Sbjct  464   TGCCTGCCTCTGGGCGAACGCCCCATCACGCCCTGAGGAGGACATGGCCGCCTGCCTGAG  523
```

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	D
root			106	
. Simiiformes	primates		99	
. . Catarrhini	primates		98	
. . . Hominidae	primates		86	
. . . . Homininae	primates		82	
. . . . . Homo sapiens	primates	7294	75	Homo sapiens hits
. . . . . Gorilla gorilla	primates	3568	2	Gorilla gorilla hits
. . . . . Pan troglodytes	primates	3506	4	Pan troglodytes hits
. . . . . Pan paniscus	primates	1324	1	Pan paniscus hits
. . . . Pongo abelii	primates	3536	2	Pongo abelii hits
. . . . Pongo pygmaeus	primates	3313	2	Pongo pygmaeus hits
. . . Chlorocebus aethiops	primates	2338	1	Chlorocebus aethiops hits
. . . Chlorocebus sabaeus	primates	1760	3	Chlorocebus sabaeus hits
. . . Papio anubis	primates	618	1	Papio anubis hits
. . . Theropithecus gelada	primates	614	3	Theropithecus gelada hits
. . . Piliocolobus tephrosceles	primates	589	1	Piliocolobus tephrosceles hits
. . . Rhinopithecus bieti	primates	532	1	Rhinopithecus bieti hits
. . . Macaca mulatta	primates	426	2	Macaca mulatta hits
. . Aotus trivirgatus	primates	1763	1	Aotus trivirgatus hits
. Expression vector pSP20	other sequences	1299	1	Expression vector pSP20 hits
. eukaryotic synthetic construct	other sequences	431	6	eukaryotic synthetic construct hits

28/12/2021

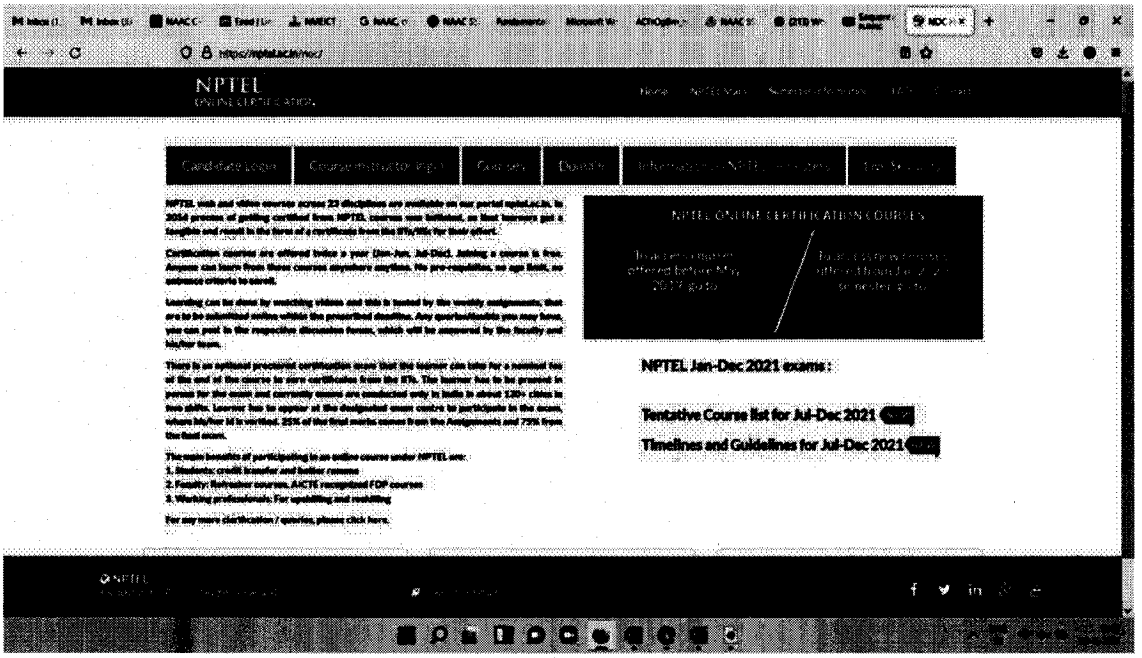
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References

- https://ncbiinsights.ncbi.nlm.nih.gov/2013/03/19/blastdbinfo-api-access-to-a-database-of-blast-databases/
- Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes, FEMS Microbiol Lett . 2002 Nov 5;216(2):133-8. doi: 10.1111/j.1574-6968.2002.tb11426.x.
- https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



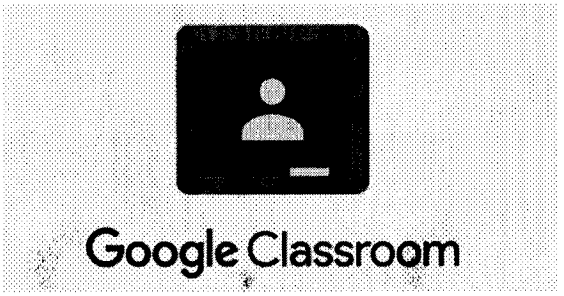
https://drive.google.com/file/d/1gnxwH8pkw3pHpjZKWPWD_3C-TIPoPf2g/view

14/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

SCHOOL OF **ENGINEERING** AND **TECHNOLOGY**

MODY UNIVERSITY
A Leading Women's University



28/12/2021

Student Manual

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Table of Contents

Table of Contents 1

Google Classroom on a Computer 3

1: Making an Account 3

2: Change Your Profile Photo 5

3: Customize your Notifications 7

4: Join a Class 8

Join a Class with a Class Code 8

Join a Class with an Email Invitation 8

5: View Your Class Resource Page 9

6: Comment on an Announcement 10

7: Access a Topic 11

8: Post an Announcement 12

Attach a File From Your Computer 12

Attach a File From Google Drive 12

Attach a Youtube Video 12

Attach a Link 13

9: Complete a PDF Assignment 14

Complete an Assignment Without Attaching a File 15

Attach an Already Existing File 15

Create a New File to Complete Your Submission 16

10: Complete a Video Assignment 18

Complete an Assignment Without Attaching a File 19

Attach an Already Existing File 19

Create a New File to Complete Your Submission 20

11: Take a Quiz 22

Multiple Choice Questions 23

Short/Long Answer Questions 23

Checkbox Questions 23

Dropdown Questions 24

File Upload Question 24

Linear Scale Question 25

Multiple Choice Grid Questions 26

Checkbox Grid Question 27

Date Question 28

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Time Question 28
12: Basic Functions 30

Google Classroom on a Cell Phone **31**
1:Download the Application 31
2: Change Your Profile Photo 32
3: Customize Notifications 34
4: Join a Class 35
Join a Class with a Class Code 35
Join a Class with an Email Invitation 35
5: View Class Resource Page 37
6: Comment on an Announcement 38
7: Access a Topic 39
8: Make an Announcement 40
Attach a File From Your Computer 40
Attach a File From Google Drive 40
Attach a Youtube Video 41
Attach a Link 41
9: Complete a PDF Assignment 42
Complete an Assignment Without Attaching a File 43
Attach an Already Existing File 43
Create a New File to Complete Your Submission 43
10: Complete a Video Assignment 45
Complete an Assignment Without Attaching a File 45
Attach an Already Existing File 46
Create a New File to Complete Your Submission 46
11: Take a Quiz 48
12: Basic Functions 49

28/12/2021

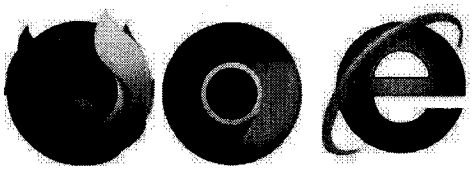
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Google Classroom on a Computer

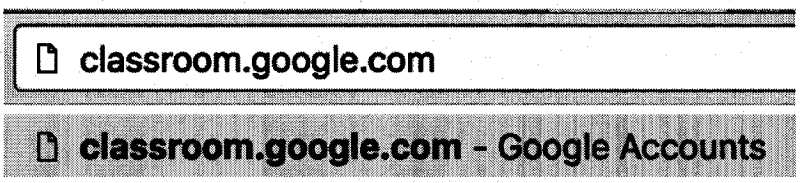
1: Making an Account

If you already have a Google Account, you can skip these instructions.

1. Open up your **internet browser** by clicking on the internet icon. (This could be **Firefox**, **Google Chrome**, **Internet Explorer**, etc., whichever internet explorer you use most often.)



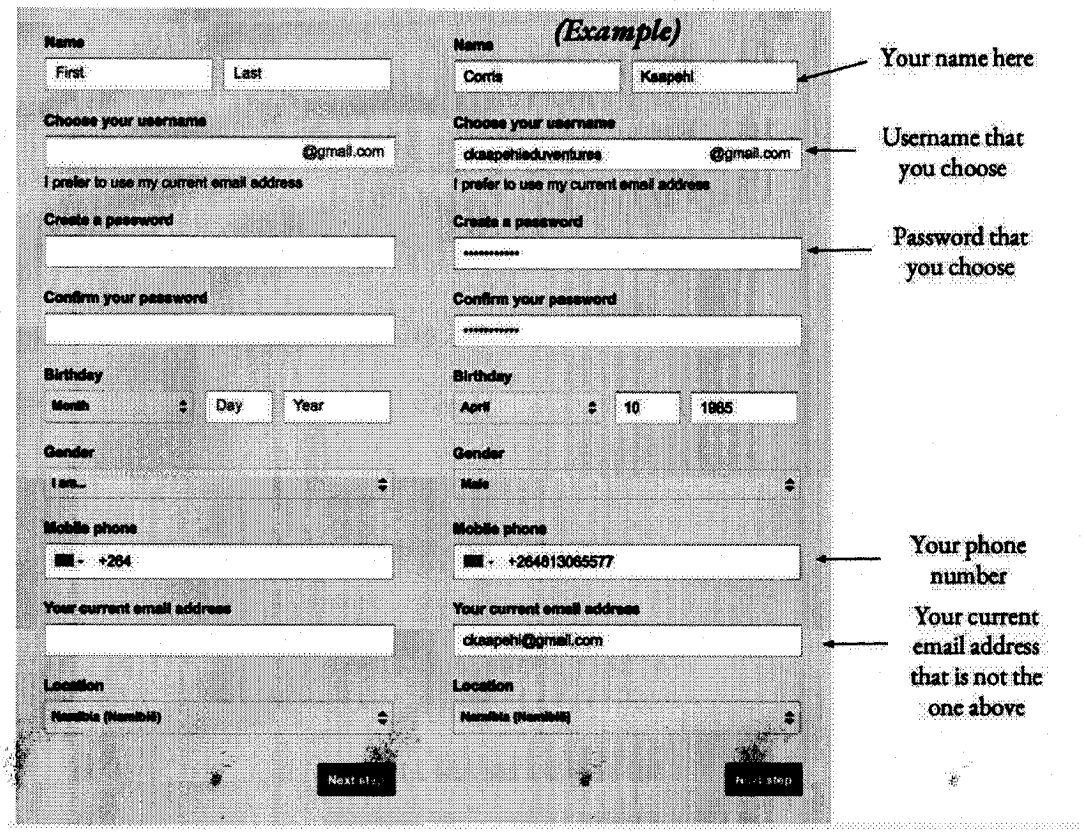
2. In the toolbar of your internet browser, **type in the following URL**: classroom.google.com and **press enter.**



3. Click the SIGN IN button in the middle of the page.
4. Create a Google Account by clicking on the Create account link.
5. On the right side of the page, **fill in the necessary information to create your Google Account**. You will have to choose your own username and password. (***It is recommended to write this information down elsewhere so that you don't forget it.***)

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



6. After you have filled out all of the necessary information, click the **Next step** button.

7. Click the **Continue to Classroom** button in the middle of the page.

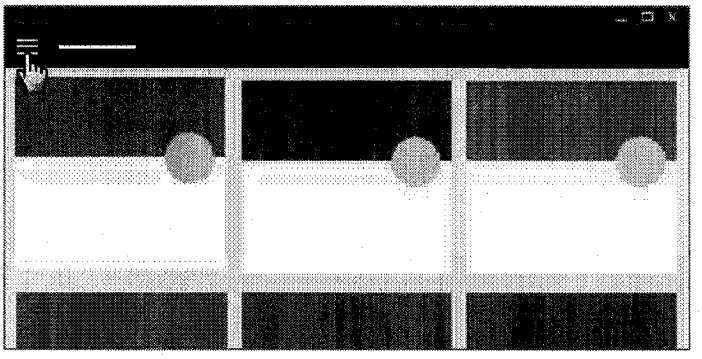
You have now successfully created a Google Classroom account!

28/12/2021

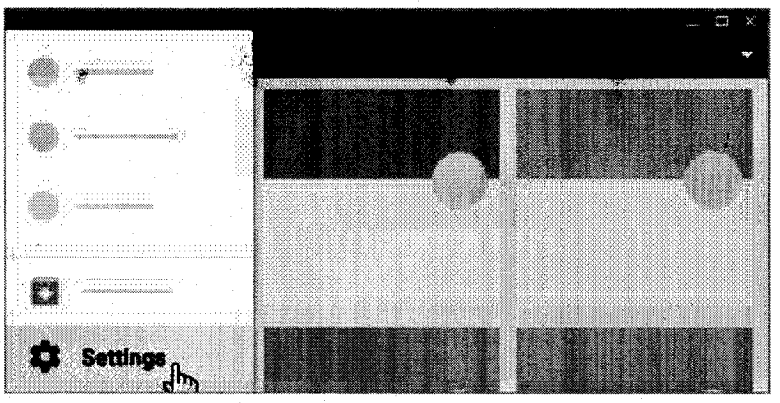
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2: Change Your Profile Photo

1. At the top left, click Menu ☰.



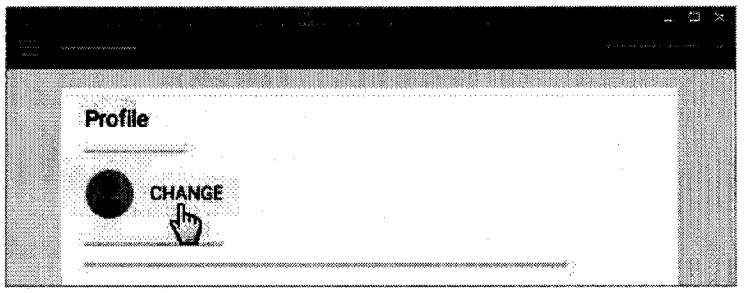
2. Scroll down and click **Settings**.



28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3. Under **Profile picture**, click **Change**.



4. Click **Select a photo from your computer** or drag a photo from your computer.

5. ***(Optional)*** Resize the box over your photo.

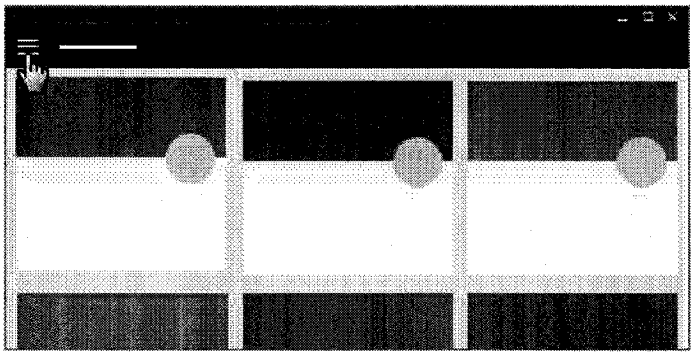
6. Click **Set as profile photo**.

29/12/2021

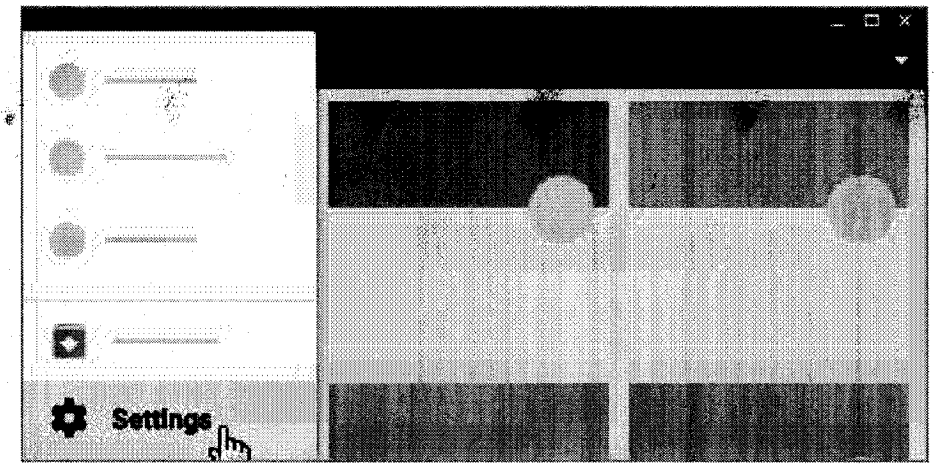
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3: Customize your Notifications

1. At the top left, click Menu ☰.



2. Click **Settings** in the bottom left (you might need to scroll down).



3. Click any notification to turn it on or off.
4. ***(Optional)*** To turn all notifications off, at **Receive email notifications**, click Turn off.

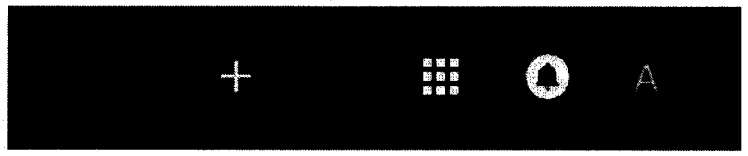
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

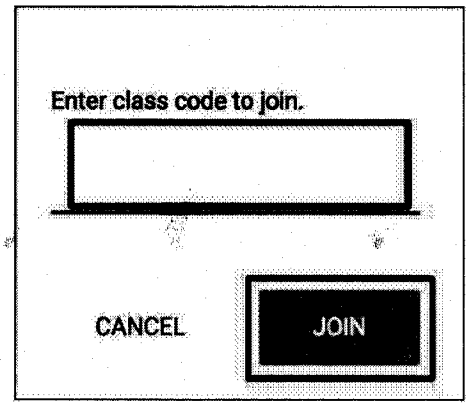
4: Join a Class

Join a Class with a Class Code

1. Go to classroom.google.com. If it is your first login, be sure to select your role as Student
2. Click the + on the top right of the page to Join Class.



3. Enter the Class Code given to you by your teacher, and click Join.



Join a Class with an Email Invitation

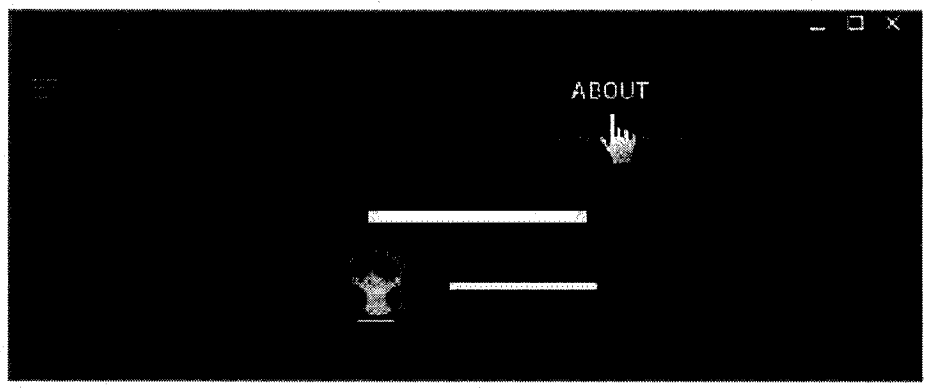
1. Go to classroom.google.com. If it is your first login, be sure to select your role as Student.
2. Click the on the top right of the page to access other Google software.
3. Press the symbol. Your newly created Google Mail page should open on a new tab in your browser.
4. Open the email with the following subject: **Class Invitation: "<CLASS NAME>"**
5. Click the button to accept the class invitation.

28/12/2021

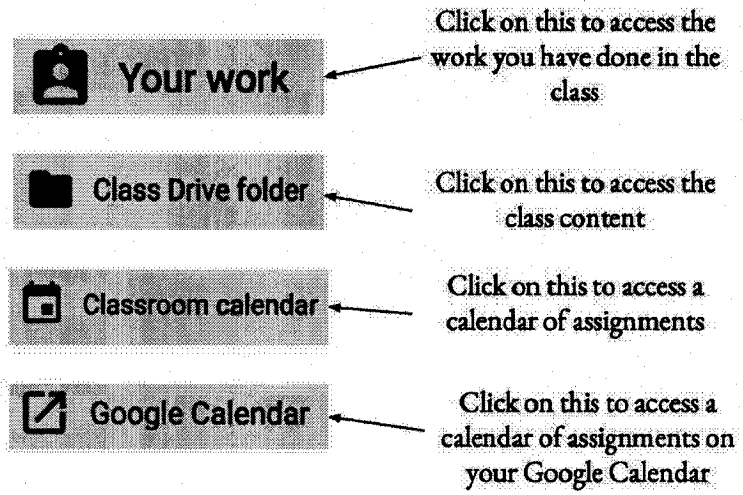
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

5: View Your Class Resource Page

1. Click a class.
2. At the top, click **About**.



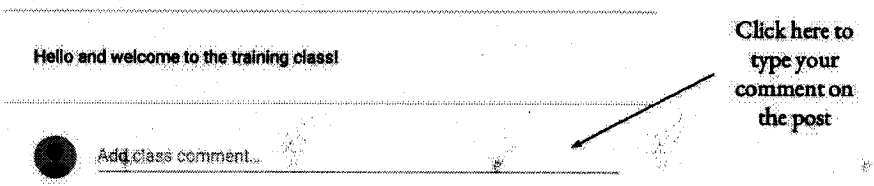
3. Here you can click on a variety of options to open them.



REGISTRAR
Moay University of Science and Techno
Lakshmangarh (Distt Sikar) 332311

6: Comment on an Announcement

1. If you are not already on the Stream page, click on the STREAM button.
2. Your screen may say "Stream was updated" in the top middle. If it does, click on the SHOW button.
3. Near the middle of the screen there will be a post to the class stream. Click on the "Add a class comment" to type your comment on the post.



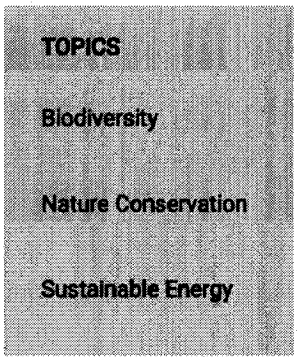
4. After you have typed your comment out, click the POST button to post your comment for the rest of the class to see.

28/12/2021

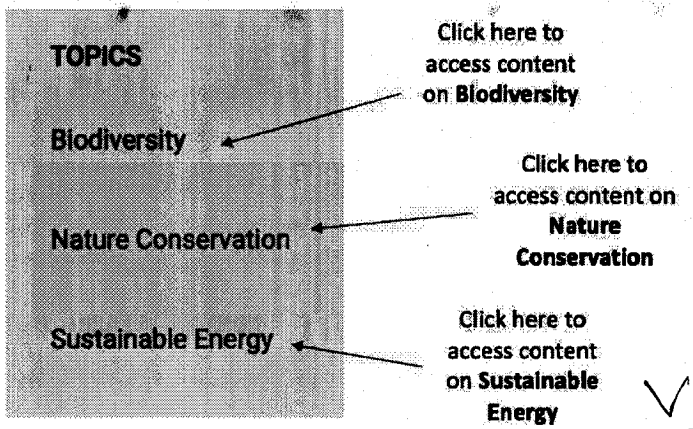
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

7: Access a Topic

1. On the left hand side of your "Stream" page, you will see a TOPICS section, which is where your instructor has categorized your class into topics. A class topic section might look like this:



2. You can access the content in each topic by clicking on a topic name:

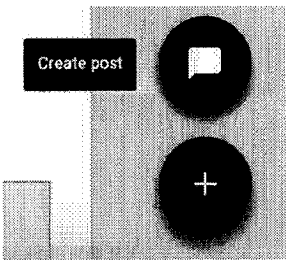


28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8: Post an Announcement

1. Click on a class.
2. At the bottom right of the page, click + and then click **Create post**.



3. In the **Share with your class** box, enter your message.
4. To post your announcement, click **POST**

Attach a File From Your Computer

1. Click on the paper clip icon

2. Click **Select files from your computer**, find the file on your computer, then double click on the file.
3. Click **Upload**

Attach a File From Google Drive

1. Click on the Google Drive icon
2. Select the file, then click **Add**

Attach a Youtube Video

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

1. Click on the Youtube icon
2. Search for the video using the search bar. Once you have found the video, click **Add**

Attach a Link

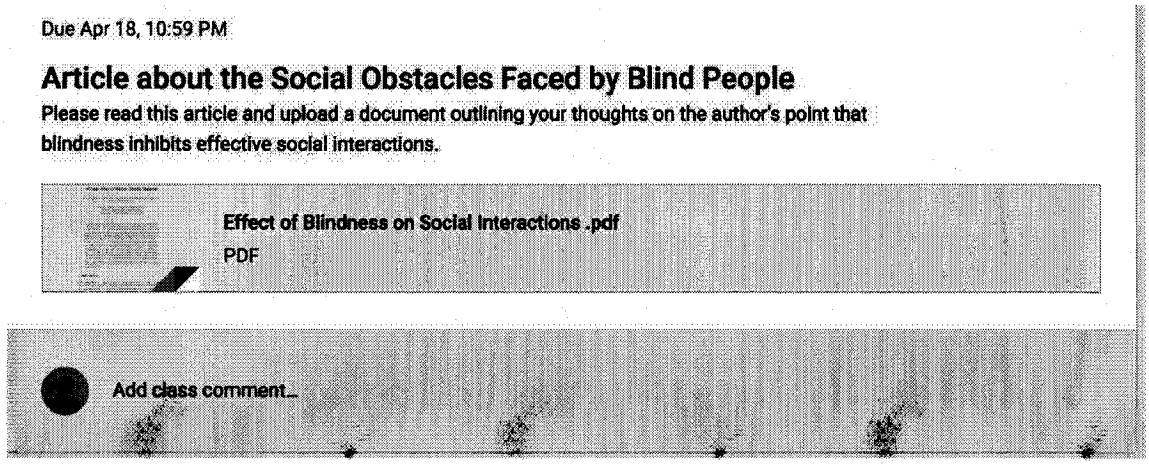
1. Click on the link icon
2. Paste the link in the **Link** box
3. Click **ADD LINK**

29/12/2021

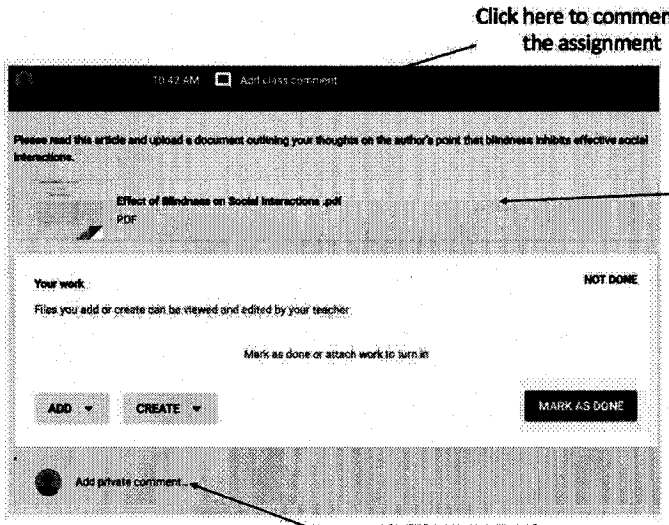
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

9: Complete a PDF Assignment

1. Click a class.
2. Scroll until you find the article assignment, and then click on the title of the assignment. In this case, the title is "Article about the Social Obstacles Faced by Blind People." The assignment can either be under a topic or on the homepage of the class.



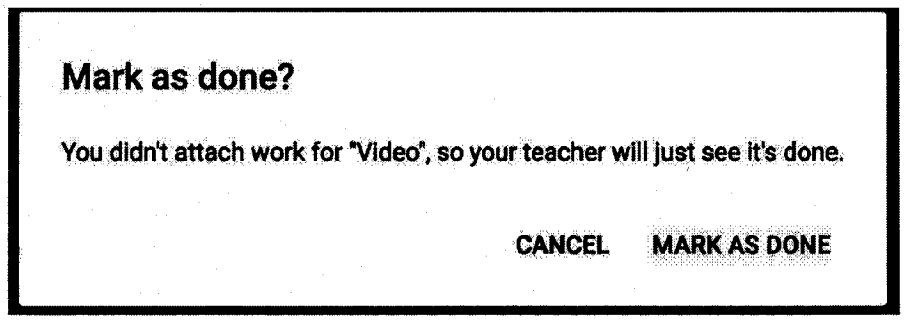
3. This will bring you to a new page.



REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

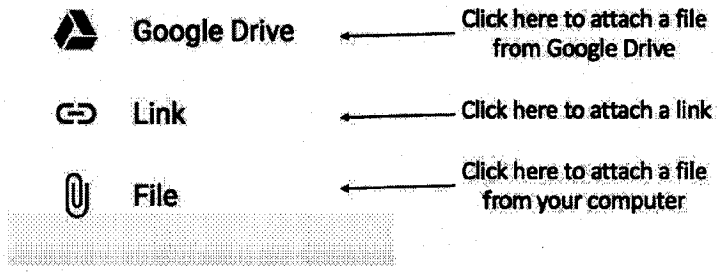
Complete an Assignment Without Attaching a File

1. If your assignment does not ask you to complete any work or attach a file, to mark your assignment as complete click on **MARK AS DONE**.
2. This will open the following pop up. Click **MARK AS DONE** again to mark the assignment as complete.



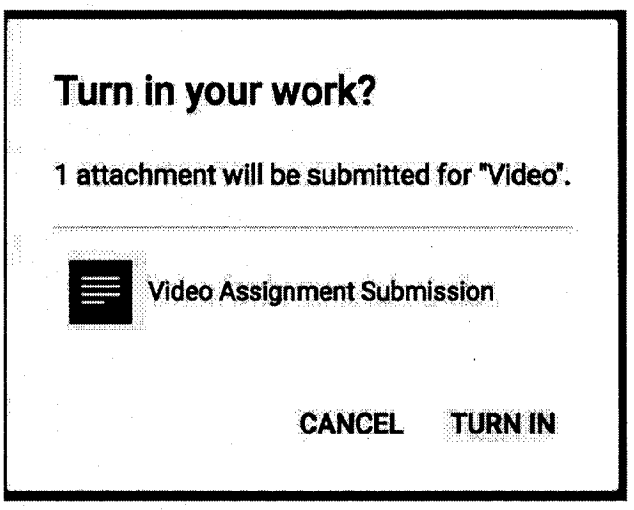
Attach an Already Existing File

1. To attach a file that you have already created, click **ADD**, which will bring up the following drop down menu.



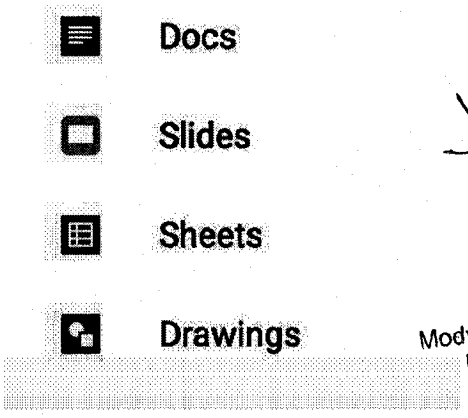
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2. After attaching your submission, click **TURN IN**.
3. This will open the following pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.



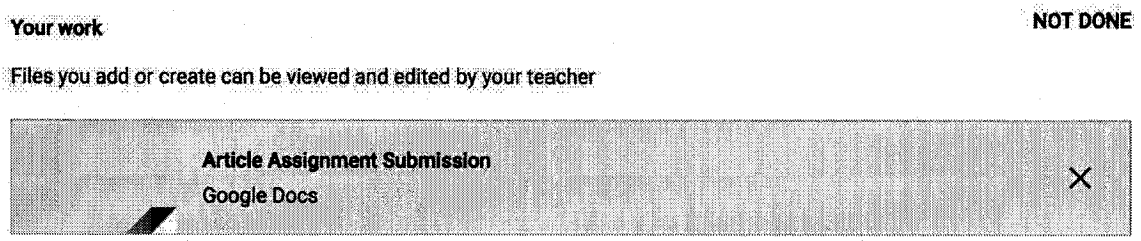
Create a New File to Complete Your Submission

1. To create a file from scratch on Google Drive, click **CREATE**, which will bring up the following drop down menu.

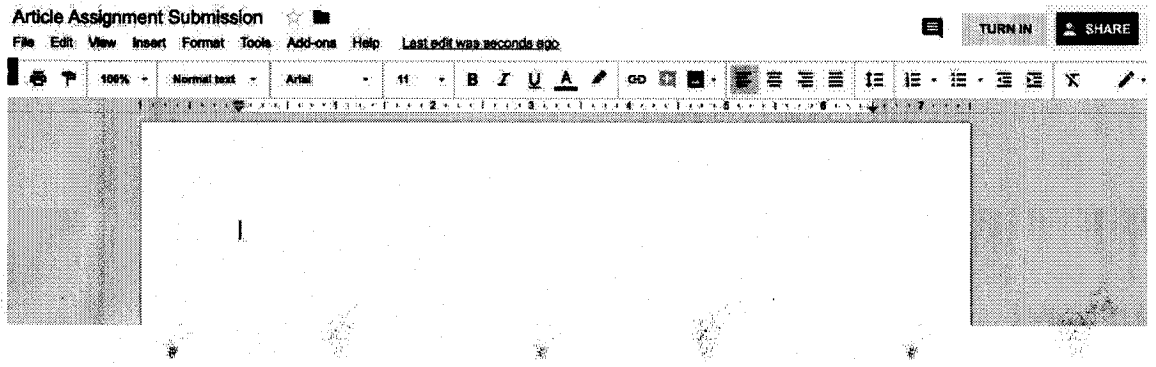


28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

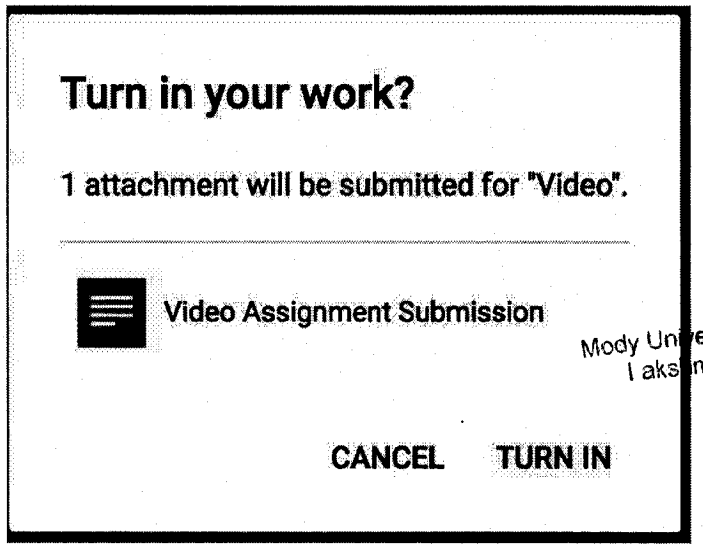
2. Click on one of the options, and this will develop a file for your submission.



3. Click on the file and a new tab will open up for you to edit the file and complete the assignment.



4. Once you have completed the assignment, click **TURN IN** in the top right corner.
5. This will open the following pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.

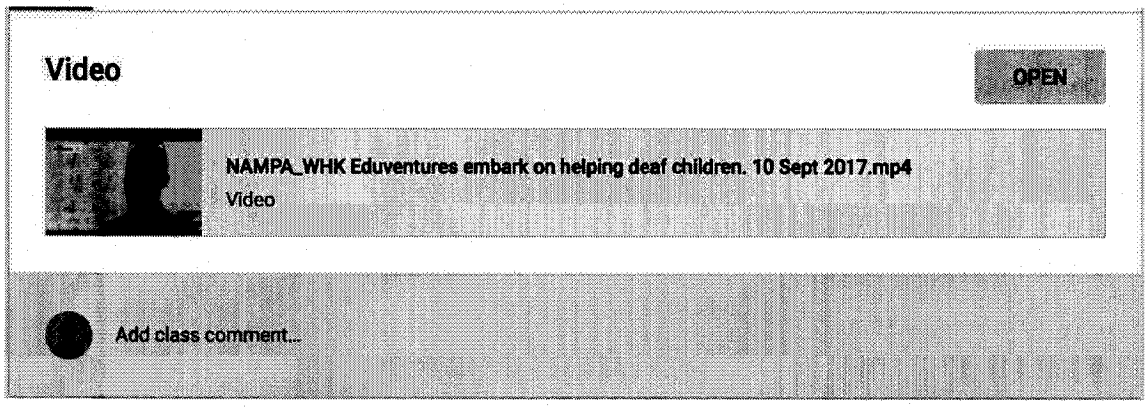


28/12/2024

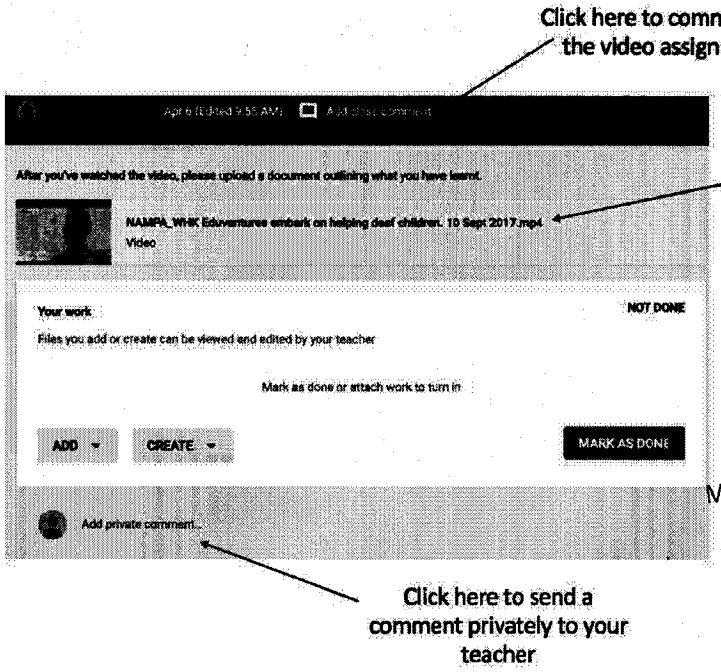
REGISTRAR
Mody University of Science and Technology
Laxmangarh (Distt Sikar) 332311

10: Complete a Video Assignment

1. Click a class.
2. Scroll until you find the video assignment, and then click on the title of the assignment. In this case, the title is "Video." The video can either be under a topic or on the homepage of the class.



3. This will bring you to a new page.

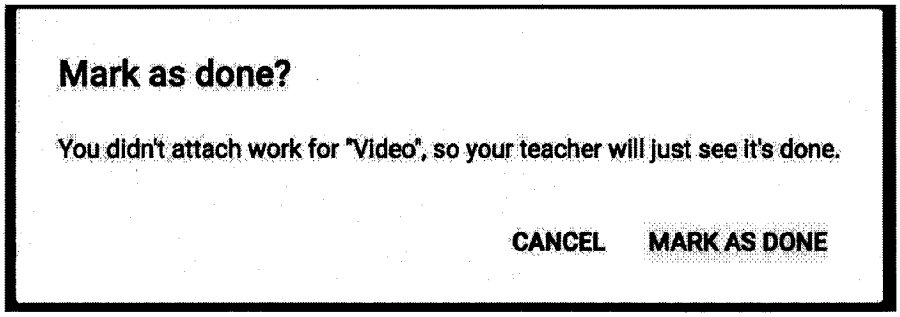


28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

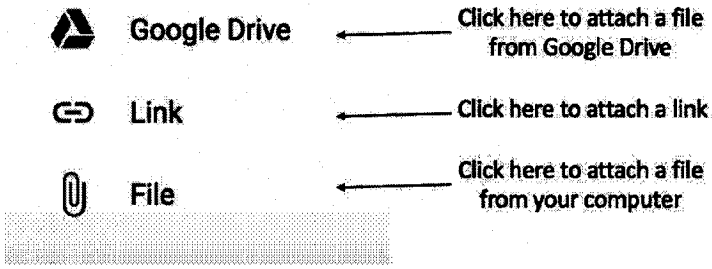
Complete an Assignment Without Attaching a File

1. If your assignment does not ask you to complete any work or attach a file, to mark your assignment as complete click on **MARK AS DONE**.
2. This will open the following pop up. Click **MARK AS DONE** again to mark the assignment as complete.



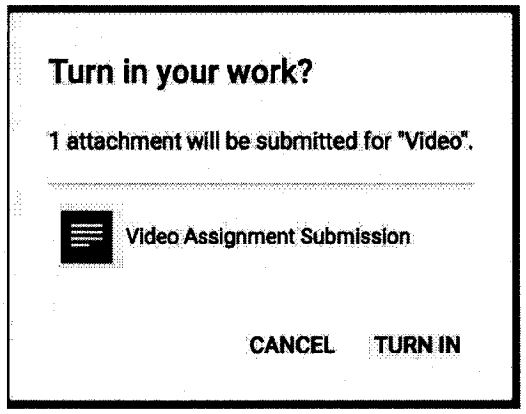
Attach an Already Existing File

1. To attach a file that you have already created, click **ADD**, which will bring up the following drop down menu.



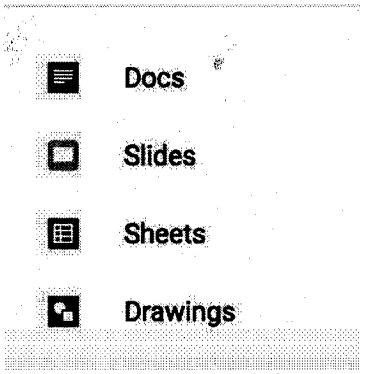
REGISTRAR
Mody University of Science and Technolog
Lakshmangarh (Distt Sikar) 332311

2. After attaching your submission, click **TURN IN**.
3. This will open the following pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.



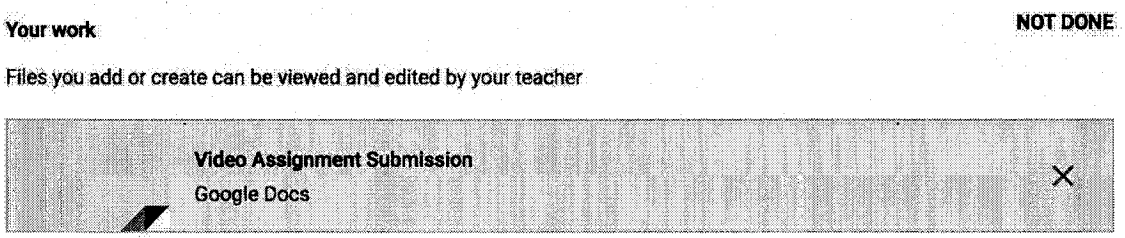
Create a New File to Complete Your Submission

1. To create a file from scratch on Google Drive, click **CREATE**, which will bring up the following drop down menu.

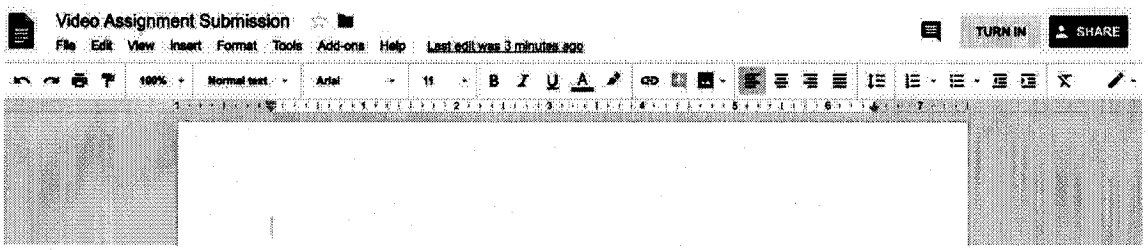


28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2. Click on one of the options, and this will develop a file for your submission.

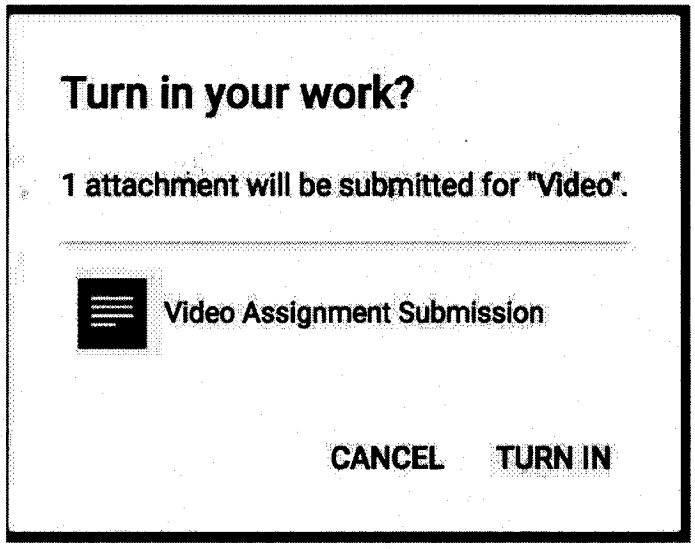


3. Click on the file and a new tab will open up for you to edit the file and complete the assignment.



4. Once you have completed the assignment, click **TURN IN** in the top right corner.

5. This will open the following pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.

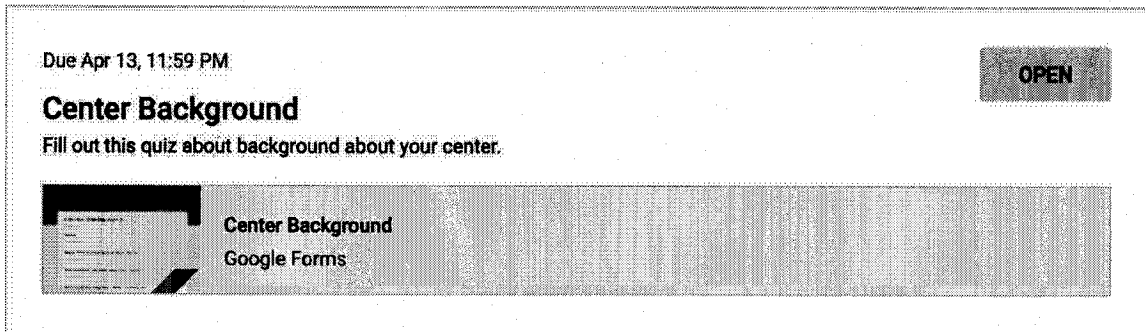


28/12/2024

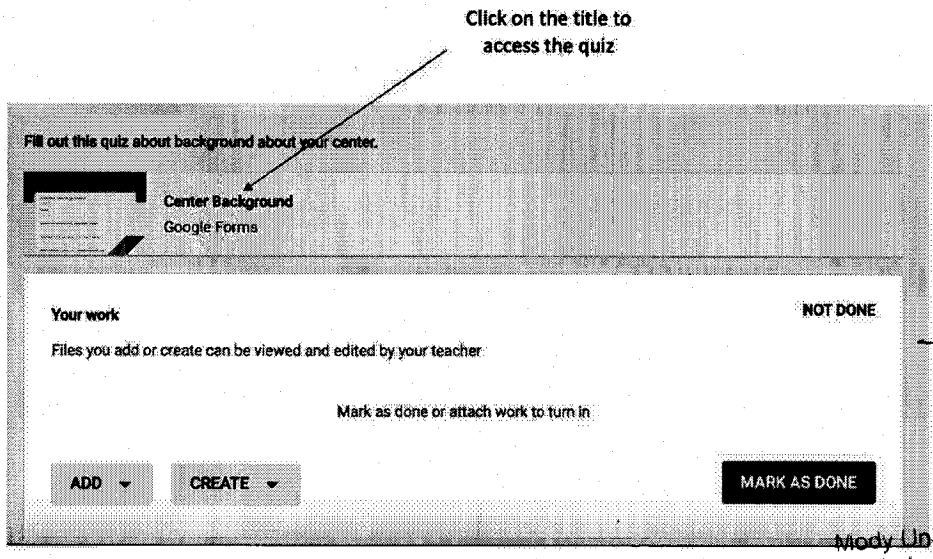
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

11: Take a Quiz

1. If your instructor has posted a quiz, you can access it in the "Stream" page, or in a specific "Topic" page.



2. Click the OPEN button to access the quiz.
3. This will lead you to another page, where you will need to click on the quiz title in the middle of the page to access the quiz.

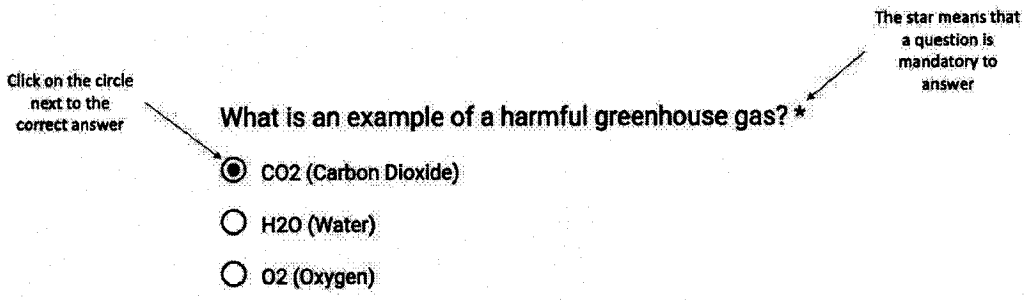


28/12/2021

REGISTRAR
Mody University of Science and Techn
Lakshmangarh (Distt Sikar) 33229

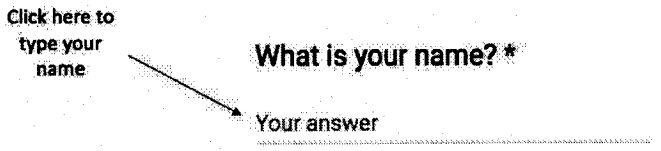
Multiple Choice Questions

1. To answer a multiple choice question, click on one circle which you believe to be the correct answer.

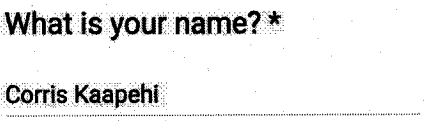


Short/Long Answer Questions

1. To answer a short answer question, click on the text



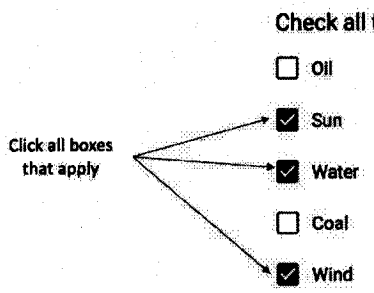
2. Type your name in the box provided. You would follow the same procedure for a question that requires a longer answer.



REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

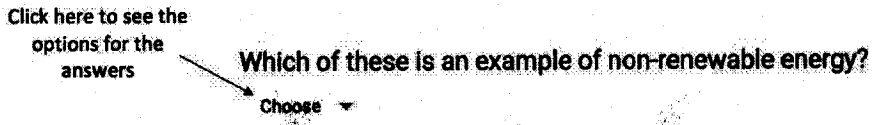
Checkbox Questions

1. To answer questions that require boxes to be checked, click on all of the boxes that apply to the correct answer.

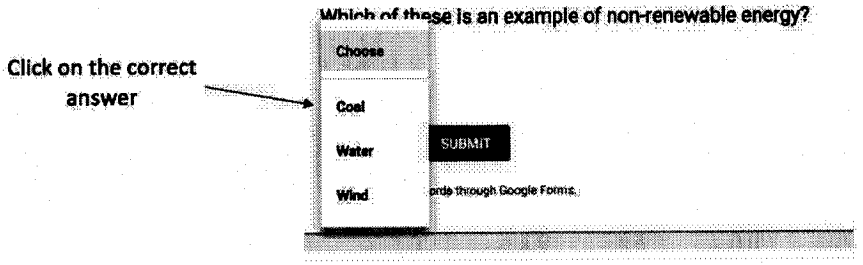


Dropdown Questions

1. To answer a question with dropdown options, click on the "Choose" button.



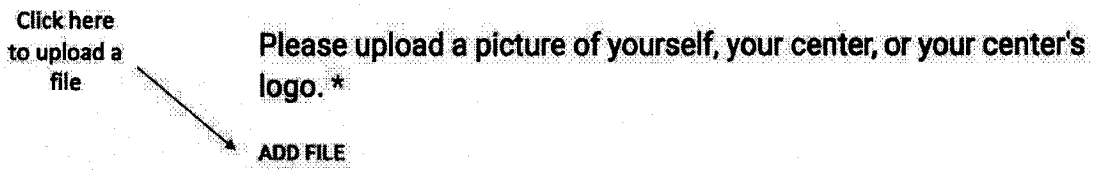
2. Select the correct answer from the dropdown menu.



28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt. Sikar)

File Upload Question

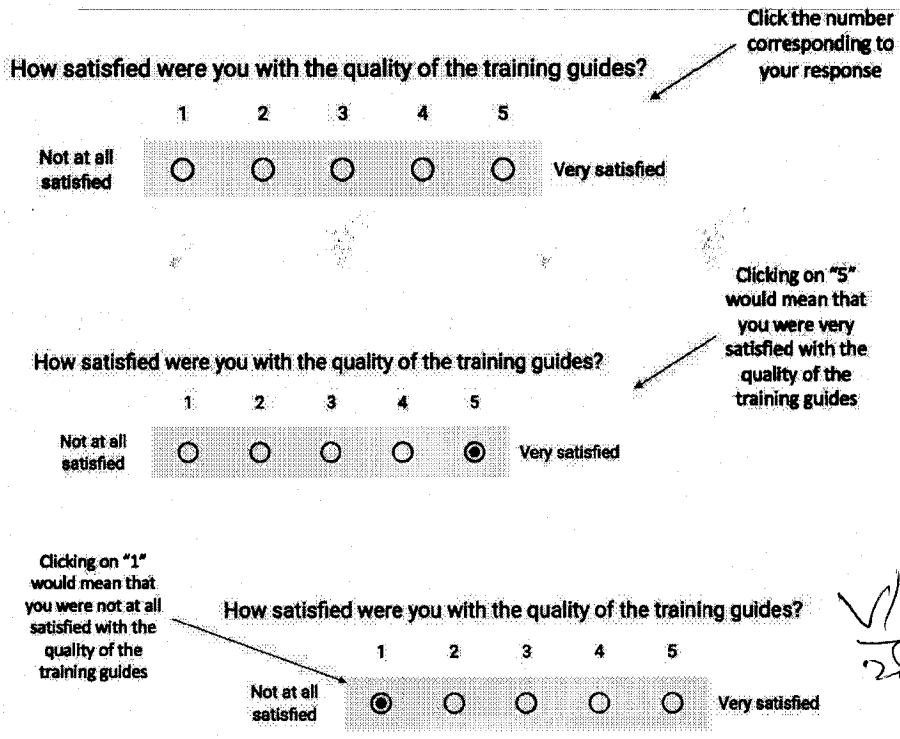
1. To answer a question that requires a file upload, click on the "Add File" button.



2. Select a file from your computer to upload.

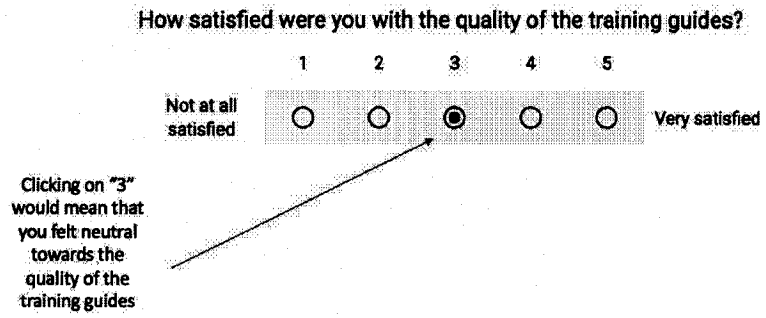
Linear Scale Question

1. To answer a linear scale question, select the number which corresponds best to your selected answer.



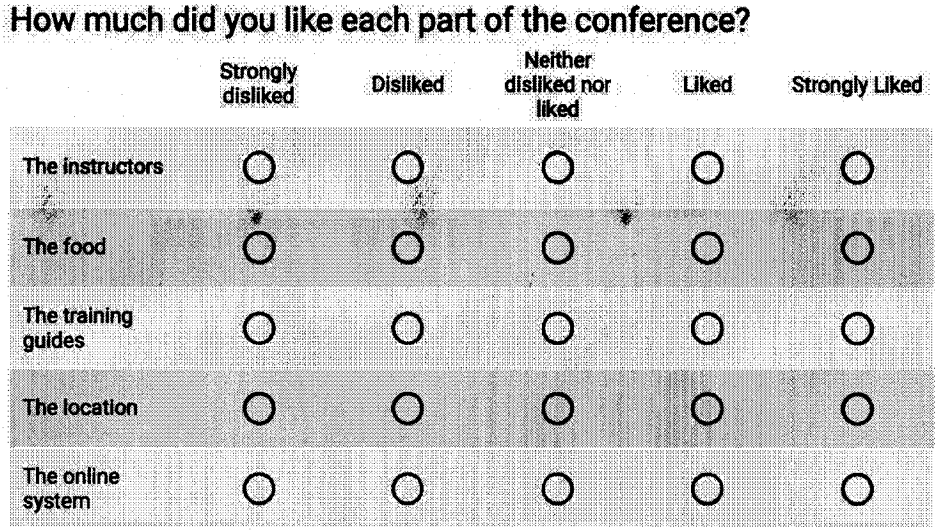
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

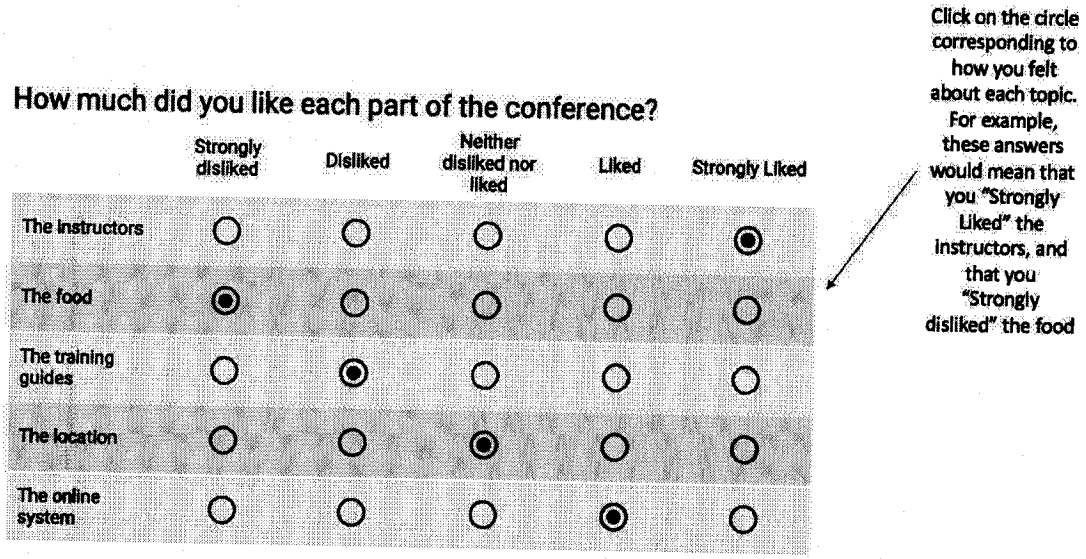


Multiple Choice Grid Questions

1. To answer a multiple choice grid question, select the number that best corresponds to each category, similar to how you answered a linear scale question.

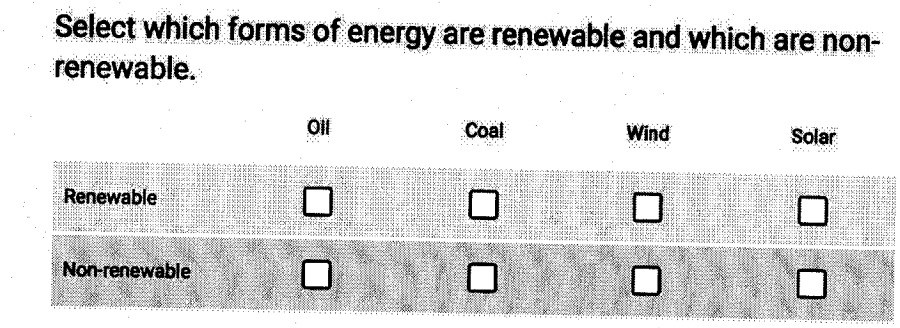


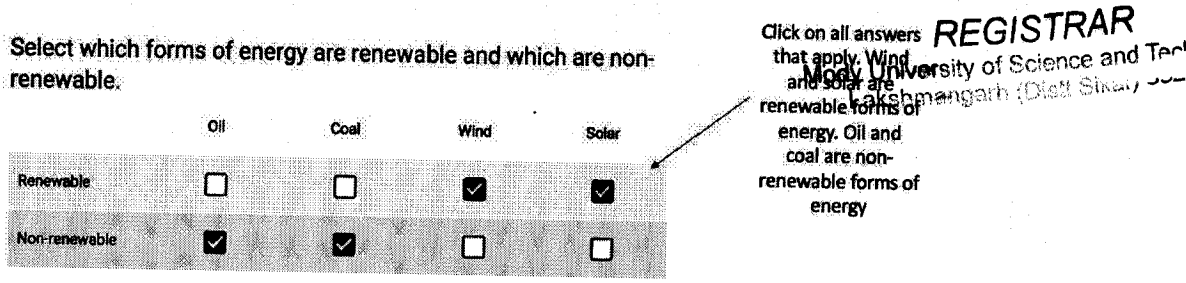
29/12/2021
REGISTRAR
Moody University of Science and Technology
Lakshmangarh (Distt Sikar)



Checkbox Grid Question

1. To answer a checkbox grid question, select the answers on the left that best correspond with the answers on the top. The difference between this type of question and the multiple choice grid question is that you can choose the same answer for one category.

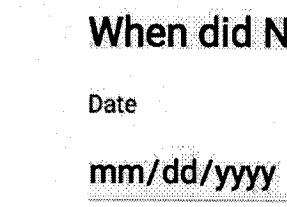


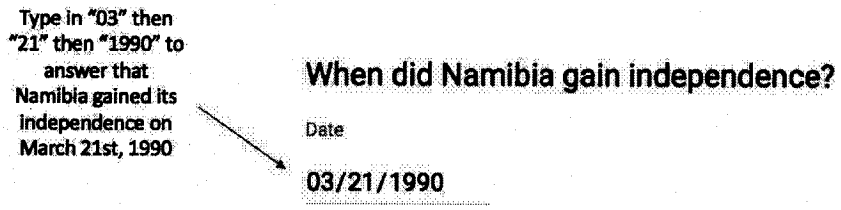


28/12/2024
REGISTRAR
Maharaja Agrasen University of Science and Technology
Lakshmangarh (Distt Sikar)

Date Question

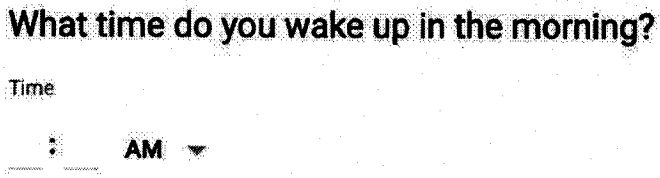
1. To answer a question that requires a date as a response, type in the correct answer in the space provided.



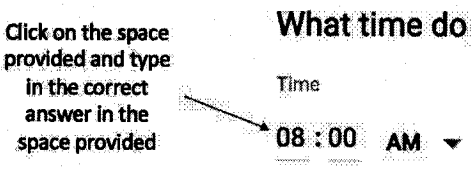


Time Question

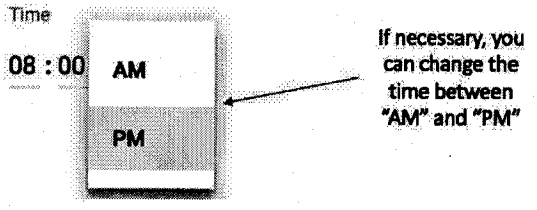
1. To answer a question that requires a time as an answer, type in the correct answer in the space provided.



28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



What time do you wake up in the morning?

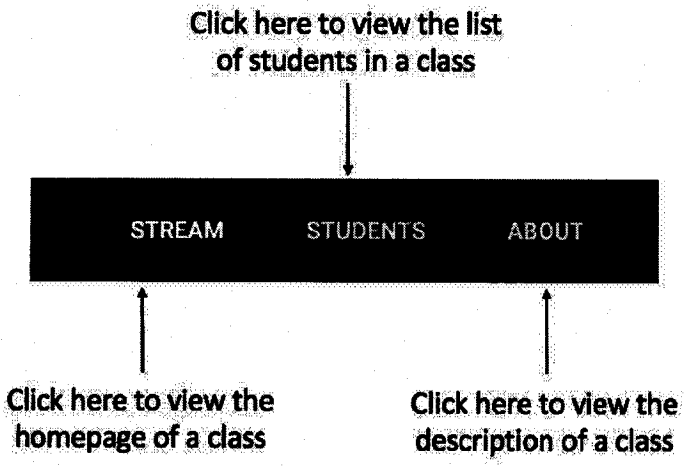


28/12/24

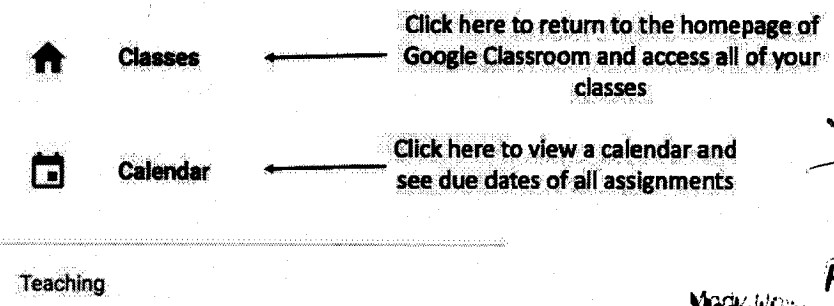
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332311

12: Basic Functions

1. Click on a class
2. At the top of the page, there is the following menu



3. Click this icon in the top left to access the following sidebar



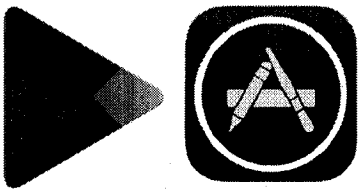
In this sidebar, you can also access all of the classes that you are enrolled in, classes that you are a instructor for, and the settings for your account.

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

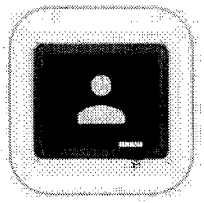
Google Classroom on a Cell Phone

1:Download the Application

1. Open your phone's respective **app store**. This may be the google play store or the apple store.



2. In the search toolbar of your app store, **type in the following search query**: "Google Classroom" and **press enter.**
3. Download the Google classroom application.



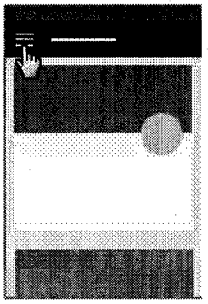
4. Click the **Get Started** button in the middle of the page.
5. If you have an Google Account, skip steps 6 and onwards and simply sign in using your username and password.
6. If you do not have an account, click the **create account** prompt.

29/1/24
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

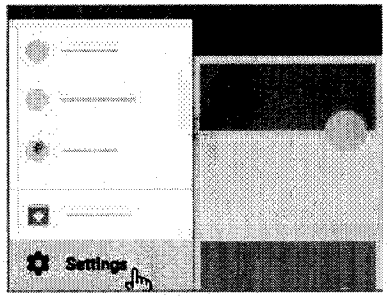
7. Fill in the necessary personal information to create the account.

2: Change Your Profile Photo

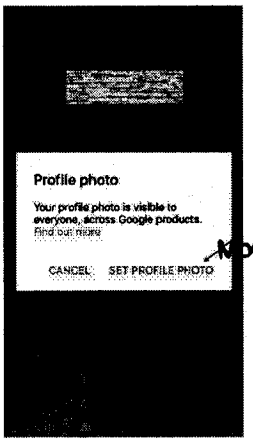
1. At the top, click Menu ☰.



2. Scroll down and click **Settings**.



3. Under **Account Setting**, click **Update Photo** then **Set Profile Photo**.



28/12/2024

REGISTRAR

Kiody University of Science and Technology
Lakshmangarh (Dist Sikar)

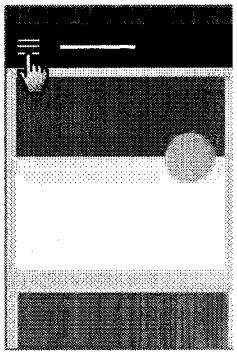
4. Click to **choose from photos** or to **take a photo.**

28/12/2017

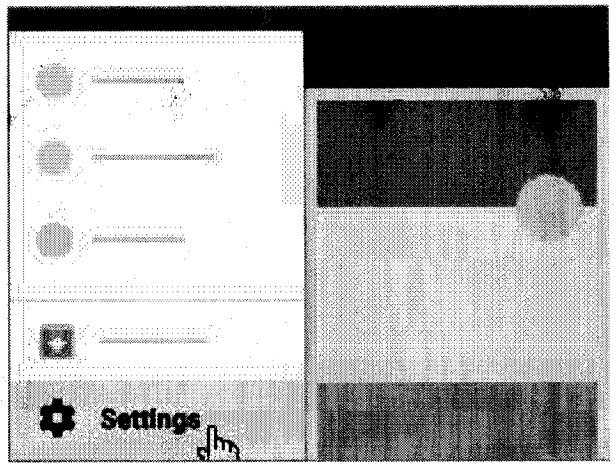
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

3: Customize Notifications

1. At the top, click Menu ☰.



2. Click **Settings** (you might need to scroll down).



3. Go into the **notifications** tab and click any notification to turn it on or off.
4. (Optional) To turn all notifications off, at **Receive email notifications**, click Turn off ⬬. To turn of all notifications off, **Device Notifications**, click Turn off ⬬.

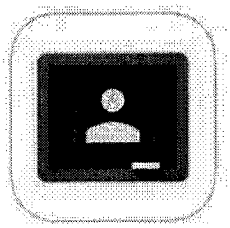
28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

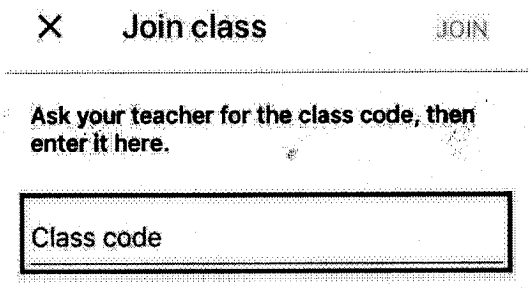
4: Join a Class

Join a Class with a Class Code

1. Open the Google Classroom app.

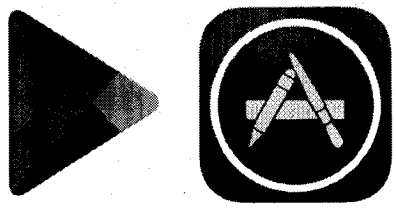


2. Click the on the top right of the page then **Join Class**.
3. Enter the Class Code given to you by your teacher, and click **Join**.



Join a Class with an Email Invitation

1. Open your phone's respective **app store**. This may be the google play store or the apple store.



REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

1. In the search toolbar of your app store, **type in the following search query**: "Google Classroom" and **press enter.**
2. Download the Google classroom application.

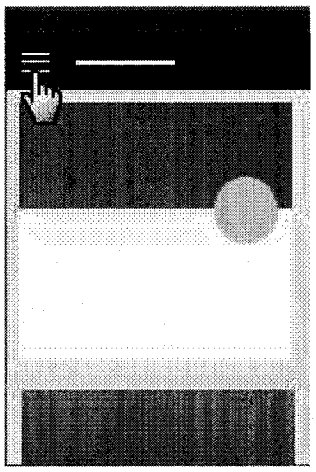
3. Press the symbol. Your newly created Google Mail page should open on a new tab in your browser.
4. Open the email with the following subject: **Class Invitation: "<CLASS NAME>"**
5. Click the JOIN button to accept the class invitation.

28/12/2021

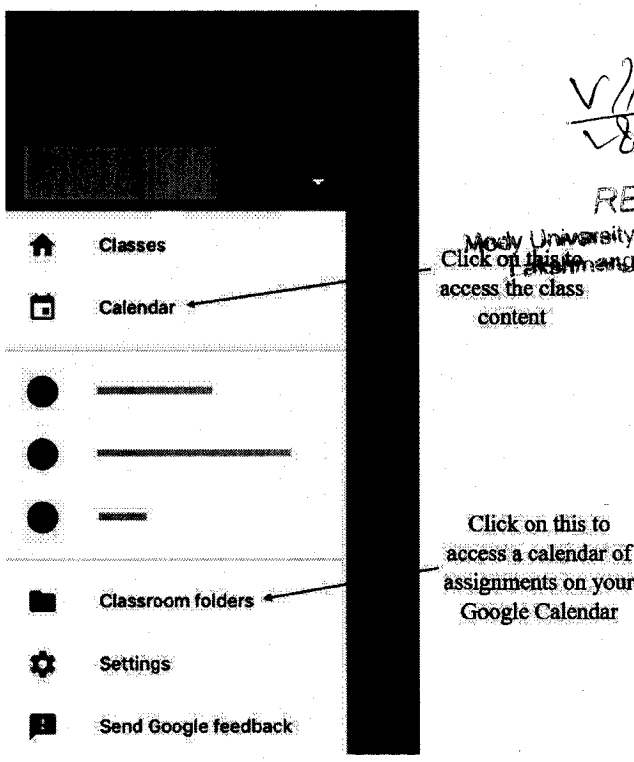
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

5: View Class Resource Page

1. Click a class.
2. At the top, click Menu ☰.



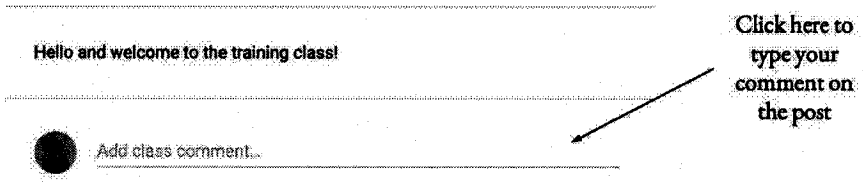
3. Here you can click on a variety of options to open them.



REGISTRAR
Science and Technology
(Deptt. Sikar) 332311

6: Comment on an Announcement

1. If you are not already on the Stream page, click on the STREAM button.
2. Your screen may say "Stream was updated" in the top middle. If it does, click on the SHOW button.
3. Near the middle of the screen there will be a post to the class stream. Click on the "Add a class comment" to type your comment on the post.

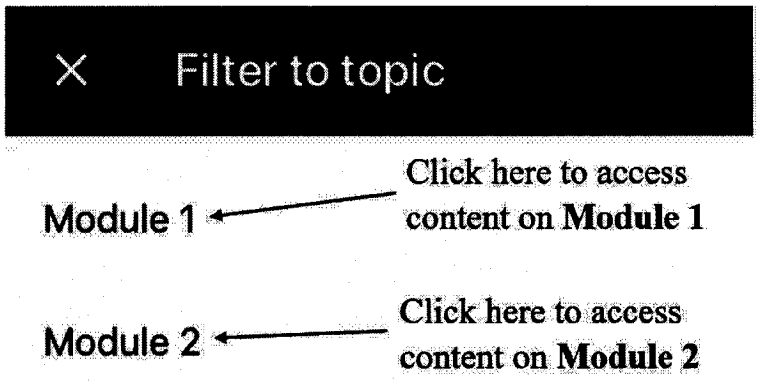


4. After you have typed your comment out, click the POST button to post your comment for the rest of the class to see.

28/7/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

7: Access a Topic

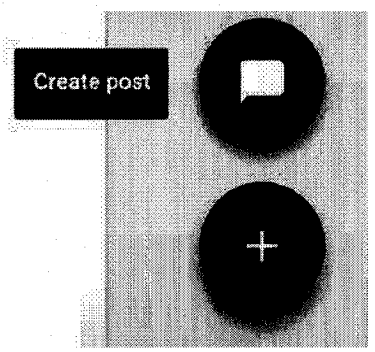
1. On the top right side of your screen, you will see a **topic filter** sign.
2. Click on the topic you would like to access.



28/12/2021
REGISTRAR
Mody University of Science and Technology

8: Make an Announcement

1. Click on a class
2. At the bottom right of the page, click + and then click **Create post**.



3. In the **Share with your class** box, enter your message.
4. To post your announcement, click **POST**

Attach a File From Your Computer

1. Click on the paper clip icon

2. Click **Select files from your computer**, find the file on your computer, then double click on the file.
3. Click **Upload**

28/12/2024

REGISTRAR
Moody University of Science and Technology

Attach a File From Google Drive

1. Click on the Google Drive icon
2. Select the file, then click **Add**

Attach a Youtube Video

1. Click on the Youtube icon
2. Search for the video using the search bar. Once you have found the video, click **Add**

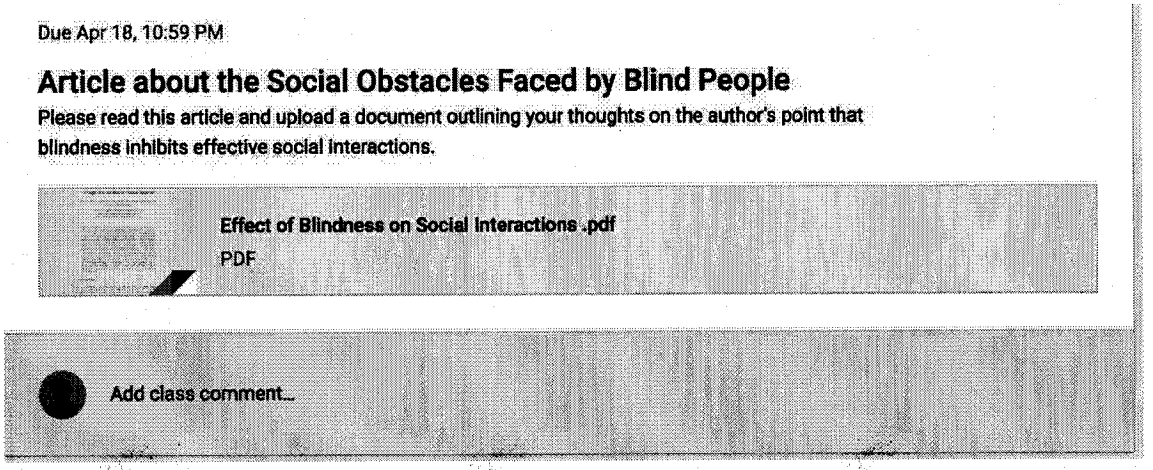
Attach a Link

1. Click on the link icon
2. Paste the link in the **Link** box
3. Click **ADD LINK**

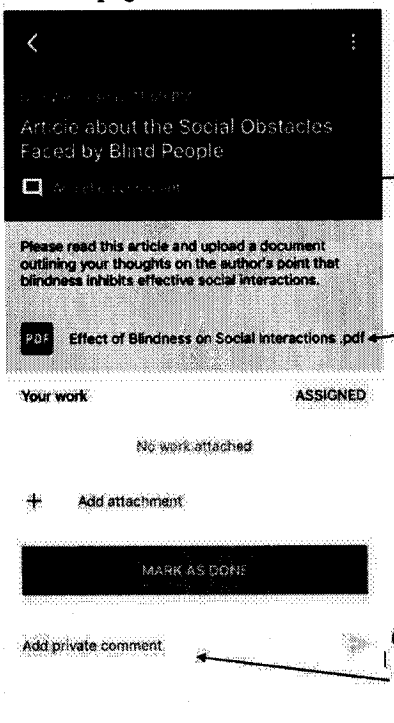
28/12/24
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

9: Complete a PDF Assignment

1. Click a class.
2. Scroll until you find the article assignment, and then click on the title of the assignment. In this case, the title is "Article about the Social Obstacles Faced by Blind People." The assignment can either be under a topic or on the homepage of the class.



3. This will bring you to a new page.



Click here to comment on the assignment

Click here to read the article

Click here to send a comment privately to you teacher

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

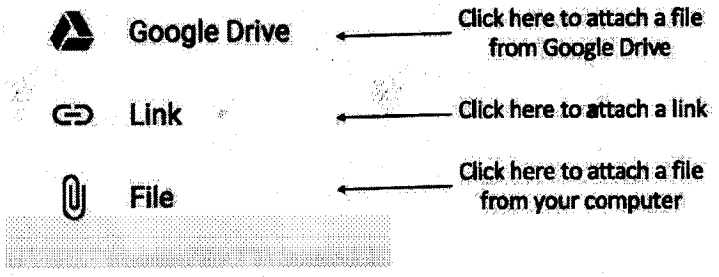
Complete an Assignment Without Attaching a File

1. If your assignment does not ask you to complete any work or attach a file, to mark your assignment as complete click on **MARK AS DONE**.

2. This will open the following pop up. Click **MARK AS DONE** again to mark the assignment as complete.

Attach an Already Existing File

1. To attach a file that you have already created, click **ADD**, which will bring up the following drop down menu.

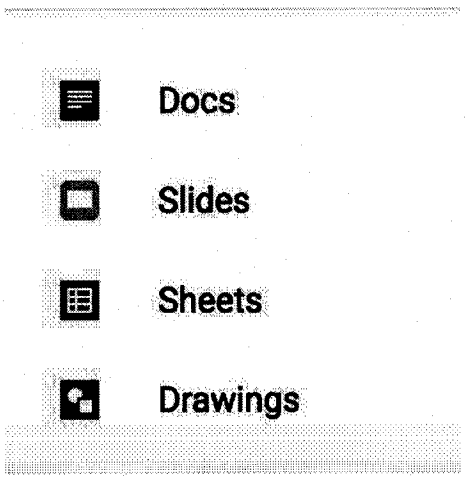


28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

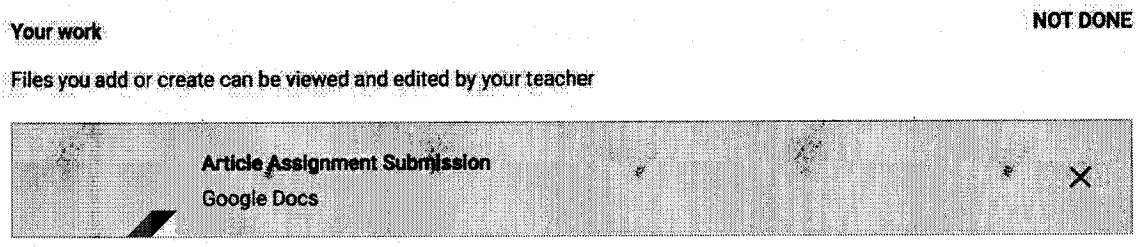
2. After attaching your submission, click **TURN IN**.
3. This will open a pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.

Create a New File to Complete Your Submission

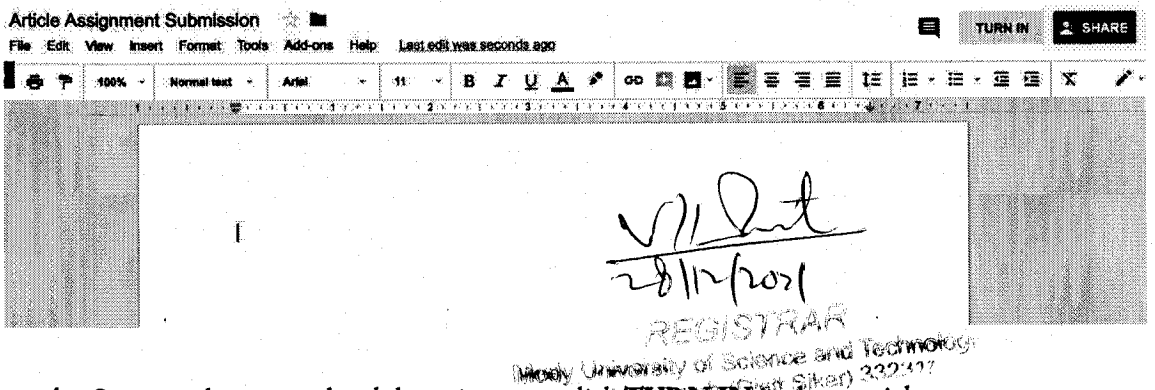
1. To create a file from scratch on Google Drive, click **CREATE**, which will bring up the following drop down menu.



2. Click on one of the options, and this will develop a file for your submission.



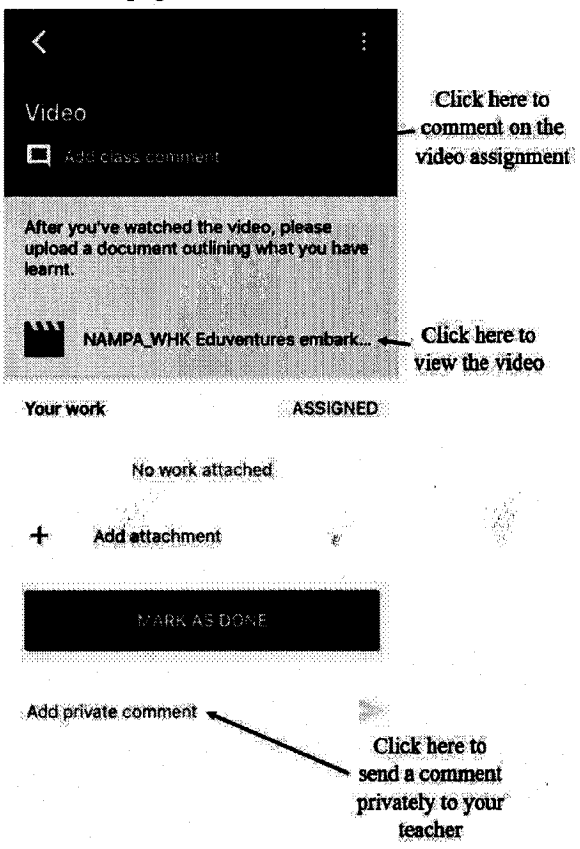
3. Click on the file and a new tab will open up for you to edit the file and complete the assignment.



4. Once you have completed the assignment, click **TURN IN** in the top right corner.
5. This will open a pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.

10: Complete a Video Assignment

1. Click a class.
2. Scroll until you find the video assignment, and then click on the title of the assignment. In this case, the title is "Video." The video can either be under a topic or on the homepage of the class.
3. This will bring you to a new page.



Complete an Assignment Without Attaching a File

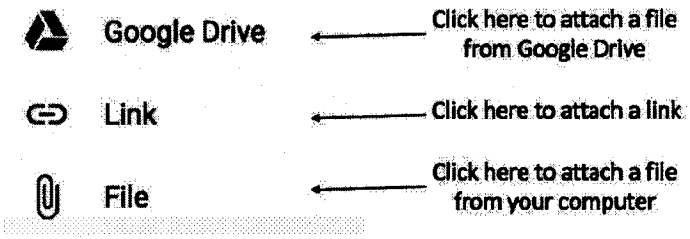
1. If your assignment does not ask you to complete any work or attach a file, to mark your assignment as complete click on **MARK AS DONE**.

2. This will open the following pop up. Click **MARK AS DONE** again to mark the assignment as complete.

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Attach an Already Existing File

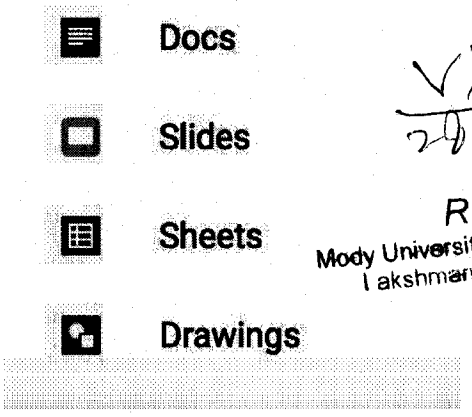
1. To attach a file that you have already created, click **ADD**, which will bring up the following drop down menu.



2. After attaching your submission, click **TURN IN**.
3. This will open the following pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.

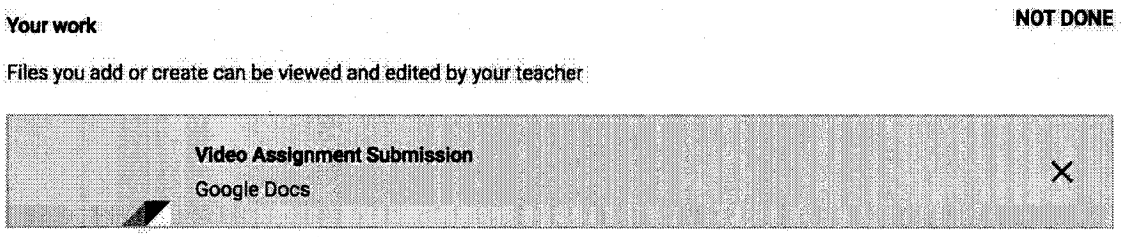
Create a New File to Complete Your Submission

1. To create a file from scratch on Google Drive, click **CREATE**, which will bring up the following drop down menu.

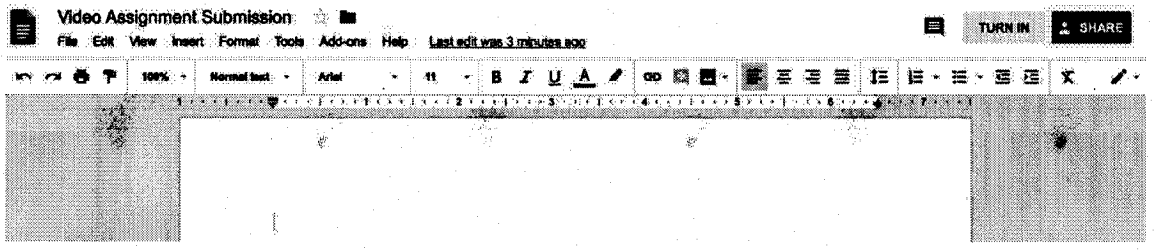


29/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2. Click on one of the options, and this will develop a file for your submission.



3. Click on the file and a new tab will open up for you to edit the file and complete the assignment.



4. Once you have completed the assignment, click **TURN IN** in the top right corner.
5. This will open the following pop up, asking if you would like to submit your work. Click **TURN IN** again to submit your assignment.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (D...n Sikar)

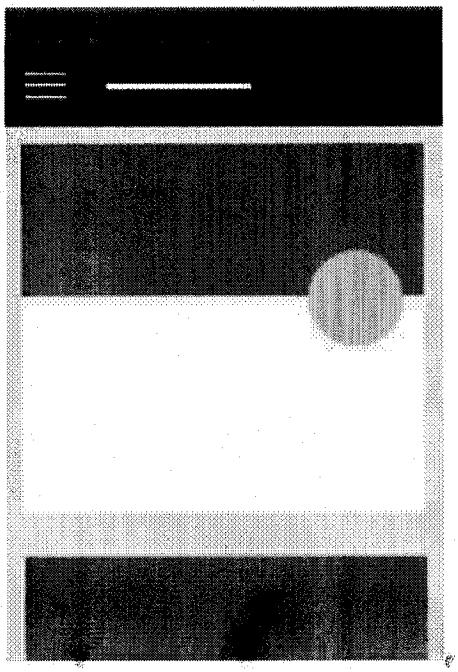
11: Take a Quiz

1. If your instructor has posted a quiz, you can access it in the "Stream" page, or in a specific "Topic" page.
2. Click the button to access the quiz.
3. Click **SUBMIT** once your done with the quiz.

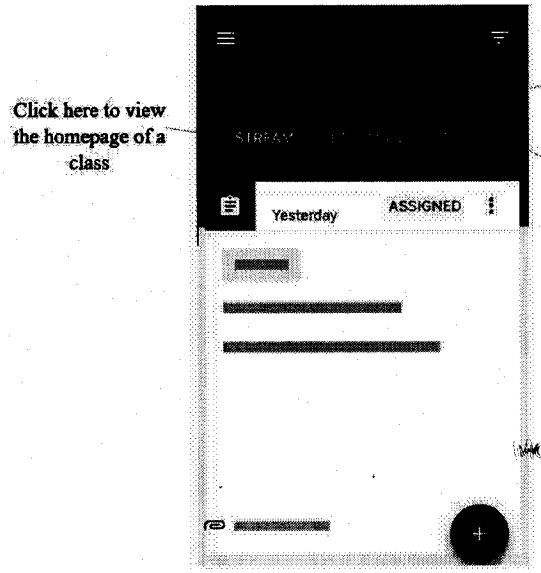
REGISTRAR
Lakshmangarh (Distt Sikar)

12: Basic Functions

1. Click on a class.



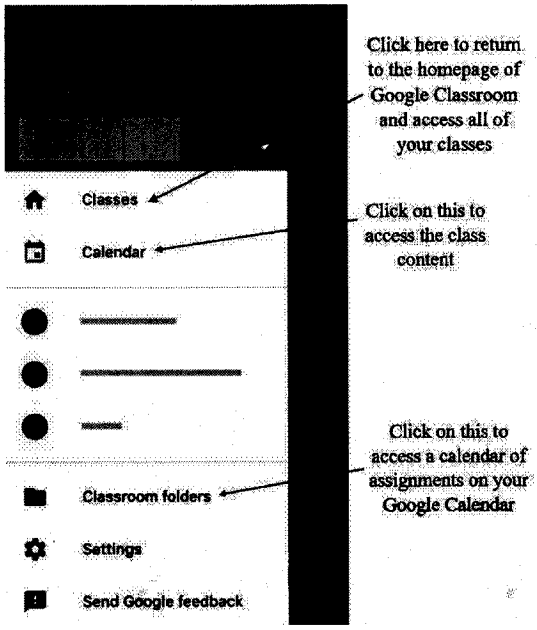
2.



28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) Rajasthan

3. Click this icon in the top left to access the following sidebar

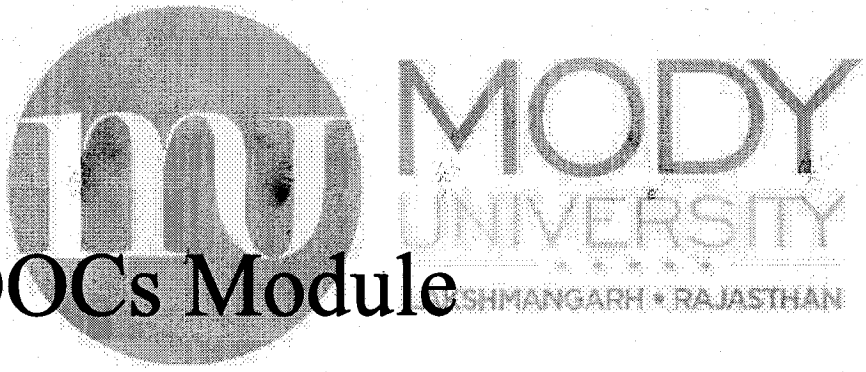


In this sidebar, you can also access all of the classes that you are enrolled in, classes that you are a teacher for, and the settings for your account.

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Mody University of Science and Technology

Department of Biosciences



MOOCs Module

Basic Bioinformatics

Dr. Naveen Dhingra, Dr. Anukriti

29/12/2021

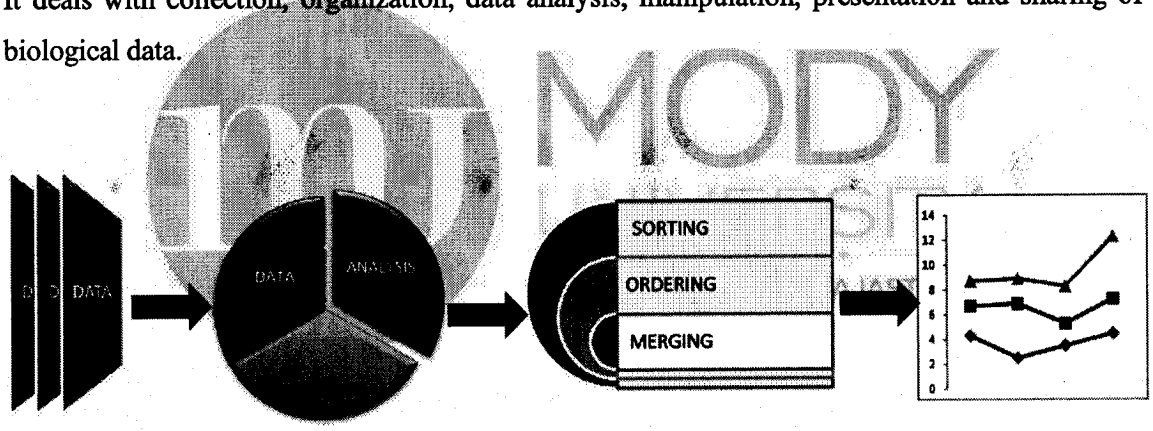
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

1: Introduction to Bioinformatics

1.1: What Is Bioinformatics?

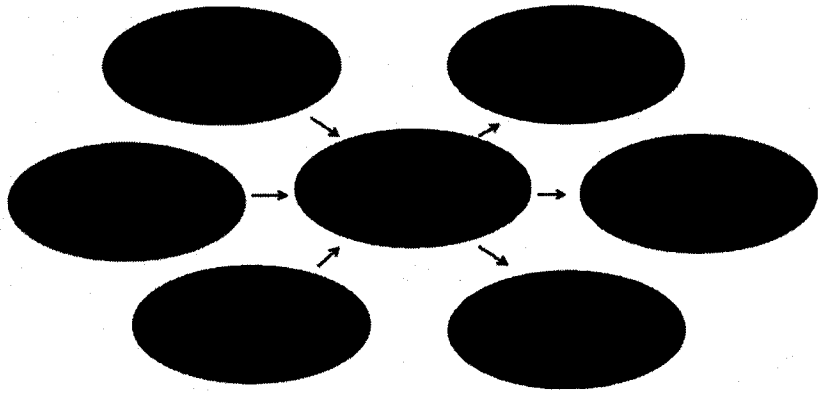
Bioinformatics is an interdisciplinary field harnessing fields like computer sciences, mathematics, physics and biology. It is convergence of biotechnology and information technology. It is a stream of science which arose for the utilitarian purpose of introducing order into the massive data sets produced by the help of new and advanced technologies of molecular biology.

Higgs and Attwood gave two definitions of bioinformatics. "Bioinformatics is the development of computational methods for studying the structure, function, and evolution of genes, proteins and whole genomes" and "bioinformatics is the development of methods for the management and analysis of biological information arising from genomics and high throughput experiments". It deals with collection, organization, data analysis, manipulation, presentation and sharing of biological data.



It requires advanced knowledge of computer science, mathematics and statistical methods to understand various biological functions at the molecular level. It covers specialized areas of biology like functional and structural genomics, DNA microarray, comparative genomics etc. It is a cutting-edge branch of biotechnology that paves its way to novel drug discovery and personalized medical treatments.

28/12/202
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)



1.2: Origin of Bioinformatics

1960s had witnessed the first efforts towards bioinformatics. Paulien Hogeweg and Ben Hesper first coined the term bioinformatics in the beginning of 1970s, defining it as "the study of informatic processes in biotic systems". Margaret Dayhoff in 1965 developed first protein sequence database called Atlas of Protein Sequence and Structure.

The Brookhaven National Laboratory in early 1970s, established the Protein Data Bank (PDB) for archiving 3D protein structures. Needleman and Wunsch developed the first sequence alignment algorithm in 1970s. Chou and Fasman in 1974 developed the first protein structure prediction algorithm.

In 1980s, GENBANK got established and William Pearson developed a fast database searching algorithm FASTA and Stephen Altschul and coworkers developed BLAST. A major developmental boost to bioinformatics came in late 1980s by the start of Human Genome Project.

1.3: Why We Should Study Bioinformatics?

- For better understanding of living cells at molecular levels.

  All the living cells are composed of various molecules that function in perfect coordination with each other for the proper functioning of the cell. With the help of

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

bioinformatics, we can try to understand a cell as a system that is composed of various molecules working together to perform any function within the cell.

- To get better understanding of evolution.

  Phylogenetic analyses are the processes that are used for the reconstruction of evolutionary relationship among a group of related molecules or organisms. This helps to predict certain features of molecules whose functions are not known, to track the flow of the genetic information, and also to determine genetic relatedness. With the help of all these, one can try to understand the process of evolution.

- For gene predictions and genome annotation.

  Gene prediction, also known as gene finding basically means locating the genes along a genome by identifying the regions of genomic DNA that encode genes using various tools offered by bioinformatics.

  Genome annotation process helps in the identification of functional elements along the sequence of a genome. It is a process of analyzing and interpreting the raw DNA sequences so that significant biological information can be extracted out from it with the use of certain tools and softwares.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

- To study the impact of change in climate on various organisms.

  With the help of bioinformatics, we can try to predict the changes in the climatic conditions and when we combine bioinformatics with genomics, we can try to predict the probable changes in genetic makeup of various organisms with the change in the climatic conditions.

- For the development of genome libraries and databases.

  Bioinformatics plays an important role in managing and analyzing the huge biological data. It helps in the development of databases and libraries from which biological information can be extracted out easily.

- For the development of molecular and personalized medicines and vaccines.

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

There are thousands of diseases that are hereditary in nature. Every individual also responds differently towards the environmental changes. Thus, trying to pin point the exact gene or molecule that is involved in the occurrence of any disease can help in designing molecular and personalized medicines. This can be achieved with the help of bioinformatics.

1.4: Uses of Bioinformatics

- Analysis of gene variation and expression.
- Analysis and prediction of gene function.
- Designing primers to predicting the function of gene products.
- Analysis and prediction of protein structure and function.
- Prediction and detection of gene regulation networks.
- Simulation of environments for whole cell modeling.
- With the help of sophisticated computational tools, mining the genomic information.
- Knowledge based drug designing.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

1.5: Scope of Bioinformatics

- Development of computational tools and databases.
- Development of algorithms.
- Data analysis and interpretation.
- In structural analysis of proteins and nucleic acids.
- Managing big data.
- Drug discovery and vaccine designing.
- In functional analysis of proteins.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

References:

1) Ardeshir Bayat, Science, medicine, and the future- Bioinformatics, BMJ. 2002 Apr 27; 324(7344): 1018–1022.

2) Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press

3) Lesk, A. M. (2002). Introduction to bioinformatics. Oxford: Oxford University Press.

4) Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer

5) Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol 7(3): e1002021. doi:10.1371/journal.pcbi.100202

6) Wang, Z.; Chen, Y.; Li, Y. (2004) A Brief Review of Computational Gene Prediction Methods; Genomics Proteomics Bioinformatics. 2004 Nov; 2(4): 216–221.

LAKSHMANGARH • RAJASTHAN

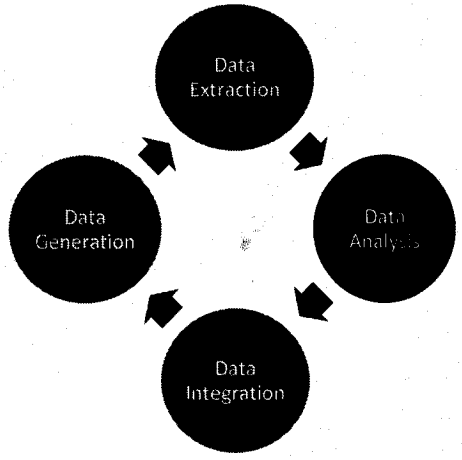
28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2: Introduction to databases

With the Technological advances in Biomedical Research like High -throughput Techniques and Lower-Cost DNA sequencing Technologies huge amount of Heterogeneous Data is getting generated daily. To manage this heterogeneous data we need better systems or tools. We also need better systems so that we can efficiently extract our data without distorting the information. For this purpose databases have been developed that can store the biological data in huge volumes and in a proper well defined manner for easy and efficient extraction of information.

Database helps in Extraction of data, its analysis, integration and generation of information or new data in a Fast and Efficient Manner.



28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2.1: What Is a Database?

Databases are collection of organized information that can be accessed managed and retrieved easily. These are composed of computer hardware and software for the management of data. With the help of a database, the information can be accessed, managed and retrieved easily and efficiently via varieties of search criteria. Sophisticated computational methodologies are being developed daily which help in easy and efficient extraction and analysis of biological data. Databases have evolved dramatically since their inception in the early 1960's. A simple database may contain records in the form of normal files

2.2: What Are Biological Databases?

A biological database is a large collection of organized biological data (DNA, RNA, gene etc.) which can easily be searched with the help of keywords. The data present in the biological databases can be easily retrieved for the public use. Database organizes the data in the form of Records and each record is known as an Entry

The data present in the biological databases is updated periodically and is cross referenced.

2.3: Purpose of Database:

A database serves various purposes like:

- Data Retrieval.
- Knowledge discovery.
  - Finding connections or relations between different pieces of data/information that were not known at the time that information was fed into the database.

In Biological Databases different types of information is stored like

- The Functions of Genes.
- Structure of Proteins.
- Localization of Proteins and Genes (both cellular and chromosomal)
- Clinical effects of mutations in Genes.
- Similarities of Biological Sequences and Structures.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

The biological databases store and organize in a structured and searchable manner for humans and systems. They not only store and organize data but also share data with different databases. They also provide web application programming interfaces (APIs) for computers to exchange and integrate data from various database resources in an Automated Manner.

2.4: Types of Databases

We can classify databases on various criteria like the level of curations or the type of data they store.

- According to level of data curation or source, biological databases can be roughly divided into two categories:
  - Primary Databases
  - Secondary Databases

2.4.1: Primary Databases

Primary databases collect data directly from the Authors. They are archives of raw sequence or structural data submitted by the scientific community. Results from the experiments conducted are submitted directly into the database by researchers. The data in the primary database becomes permanent once accession number has been allotted. These databases are archival in nature. These three public databases (GENBANK, EMBL and DDBJ) closely collaborate and exchange new data daily and constitute the International Nucleotide Sequence Database Collaboration.

For the three-dimensional structures of biological macromolecules, there is only one centralized database, the Protein Data Bank. This database archives atomic coordinates of both Proteins and Nucleic Acids, determined by x-ray crystallography and NMR. They contain the amino acid sequence of proteins, their conformation, structure, and features such as active sites.

28/12/202
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

2.4.2: Secondary Databases

Secondary Databases contains computationally processed or manually curated information, based on original information from Primary Databases. They are also known as curated databases or knowledgebases. SWISS-Prot and Protein Information Resources (PIR) are examples of Secondary Databases. SWISS Prot provides detailed sequence annotation that

includes Structure, Function, and Protein Family Assignment. The Sequence Data are mainly derived from the Database of Translated Nucleic Acid Sequences stored in the EMBL Database. Protein Information Resources (PIR) is an integrated public database which supports genomic, proteomic and systems biology research and scientific studies. It helps in *the identification and interpretation of protein sequence information.*

2.5: Classification of Biological Database according to the scope of Data Coverage.

- Biological database is divided in two parts:
  - Comprehensive Database
  - Specialized Database

2.5.1: Comprehensive Database

Comprehensive Database covers different types of data from numerous species. For example GenBank, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) contain Nucleotide Sequences from various Organisms. Uniprot is a comprehensive database that provides high quality information on the protein sequence and also has information related to the functions performed by the proteins. This is also a freely accessible database.

28/12/21
REGISTRAR
Sikkim University of Science and Technology

2.5.2: Specialized Database

Specialized Databases contain specific types of data or data from specific organisms. WormBase is a database that provides accurate, current, accessible information concerning the genetics, genomics and biology of *C. elegans* and other related Nematodes. FlyBase is another database that contains information of the genes and genomes of *Drosophila.* HIV sequence database and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

GenBank EST database and Microarray Gene Expression Database at the European Bioinformatics Institute (EBI) are examples of some available gene expression databases.

2.6: Biological Databases can also be classified on the Basis of Nature of Data they contain

2.6.1: Sequence database

A Sequence database is a biological database that has a huge collection of nucleic acid sequences, amino acid sequences, or other polymer sequences. These databases can be nucleic acid sequence databases and protein sequence databases.

2.6.1.1: Nucleic acid sequence databases

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Nucleic acid sequences are stored in three primary sequence databases GENBANK, EMBL and DDBJ. The International Nucleotide Sequence Database Collaboration (INSDC) is an initiative that operates between DDBJ, EMBL-EBI and NCBI. It collects and provides the information on the nucleotide sequences along with the metadata. INSDC contains the data raw reads, through alignments and assemblies to functional annotation, containing contextual information related to samples and experimental configurations.

All the three databases (GENBANK, EMBL and DDBJ) are synchronized daily and data on nucleotide sequences contributed by research teams is shared between the collaborating organizations.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

2.6.1.1.1: RefSeq: NCBI Reference Sequence Database

This is a collection which provides a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses.

The main features of the RefSeq collection:

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators

2.6.1.1.2: GENBANK

It is a comprehensive database that contains publicly available nucleotide sequences for more than 3,00,000 organisms. The data is obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. It is an annotated collection of all publicly available nucleotide sequences. GenBank is accessible through the NCBI Entrez retrieval system.

Genbank integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed.

2.6.1.1.3: European Molecular Biology Laboratory (EMBL)

The EMBL Nucleotide Sequence Database is Europe's primary nucleotide sequence resource. In collaboration with DDBJ and GenBank the database is produced, maintained and distributed at the European Bioinformatics Institute. EMBL Database entries are grouped into divisions. The EMBL Database currently consists of 17 divisions with each entry belonging to exactly one division. Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages.

2.6.1.1.4: DDBJ (DNA Data Bank of Japan)

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

The DNA Data Bank of Japan is a public database of nucleotide sequences established at the National Institute of Genetics (NIG). The DDBJ Center also services Japanese Genotype-phenotype Archive (JGA), with the National Bioscience Database Center. The JGA is a permanent archiving service for genotype and phenotype data of human individuals. There are various services that are provided by DDBJ like AGD which is a controlled-access database for sharing individual-level genotype and phenotype information among specific researchers. ARSA is a service provided by DDBJ for the retrieval of the information using accession number and keywords. BLAST, Bioproject, CRISPRDirect, CLUSTAL W, DFAST etc. are some of the other services provided by DDBJ.

2.6.1.2.Protein Sequence databases are

- PIR,
- SWISS-PROT,
- TrEMBL,
- PROSITE and
- PFAM

2.6.1.2.1: Prosite

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. PROSITE is complemented by collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids.

2.6.2: Structural database

Structural databases help in producing, solving, refining and publishing the structure of a new material. One can consult these databases before synthesizing new compounds. PDB, SCOP and CATH are structure databases.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2.6.2.1: The Protein Data Bank (PDB)

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. It contains information related with the 3D structure of proteins, nucleic acids and complex assenblies. RCSB PDB, PDBe, PDBj, BMRB and EMDB are the members of PDB.

2.6.2.2: Structural Classification of Proteins (SCOP)

It provides a detailed and comprehensive description of the relationships of known protein structures. SCOP provides its data with the help of Representation State Transfer (REST) protocol. The classification is on hierarchical levels:

- Family,
- Super family and
- Fold

2.6.3: Functional database

A functional database is one which provides information related to the physiological roles of various biological entities like genes, enzymes etc. KEGG, Reactome and BioCyc are some of the examples of Functional Databases. BRENDA is an enzyme activity Database. It is Comprehensive Enzyme Information system. ExPASY-ENZYME is repository of information relative to the nomenclature of enzymes.

28/12/2021
REGISTRAR
Mody University of Science and Technology

2.6.3.1: KEGG

It is a database that helps in understanding the high level functions and utilities of cells, organisms etc. It provides the computer based representation of biological systems which contain genomic and chemical information. It contains information of diseases and drugs and their inpact on normal biological systems.

2.6.3.2: BRENDA

BRENDA is an enzyme database. It has collection of the enzyme functional data. It is maintained and developed at the Institute of Biochemistry and Bioinformatics at the Technical University of Braunschweig, Germany. Information related to the functional properties of enzymes is taken directly from the works published by the scientists.

2.7: HOW TO RETRIEVE INFORMATION FROM DATABASES

To perform complex queries in a database often requires the use of Boolean operators like AND, OR, and NOT to indicate relationships between the keywords used in a search. With the help of various keywords and the combination of keywords, we can search the biological databases easily.

2.8: LIMITATIONS AND PITFALLS

- Redundancy is a major problem Affecting Primary Databases
  - Some problems are because of Human Errors.
  - Often, the same gene sequence is found under different names resulting in multiple entries and confusion about the data.
- Sequencing Error
  - Most errors in Nucleotide Sequences are caused by Sequencing Errors.

2.9: ADVANTAGES OF BIOLOGICAL DATABASES

- Easy access of Data/information
- Provide Information to answer various biological questions
- Open access to all users
- It allows indexing of all the data
- It Connects the experimental work or data with programming skills

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- With the help of databases, same information can be accessed by multiple users at the same time.
- Different types of databases, helps users narrow down their search criteria and helps in quick retrieval of information.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)

References

- https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/
- Lesk, A. M. (2002). Introduction to bioinformatics. (2nd ed.). Oxford: Oxford University Press.
- Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer.
- Xiong, J. (2006). Essential Bioinformatics. (1st ed.).Cambridge University Press.
- Benson DA et al. Genbank; Nucleic Acids Res. 2010 Jan; 38(Database issue). doi: 10.1093/nar/gkp1024
- https://www.ncbi.nlm.nih.gov/genbank/
- http://pd.chem.ucl.ac.uk/pdnn/pubs/databs1.htm
- https://www.wwpdb.org/

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3: SEQUENCE DATABASES

Biological databases provide updated high quality data.

Sequence databases are biological databases that store large collection of nucleotide sequences and protein sequences digitally.

Sequence databases can be divided into nucleic acid sequence databases and protein sequence databases

The Nucleic acid sequence database is a collection of sequences from several sources.

These sources can be:

- GenBank
- RefSeq
- TPA
- PDB

3.1: Uses of these databases:

- discovery of novel genes
- identification of homologous genes
- chromosomal localization of genes
- analysis of alternative splicing
- detection of polymorphisms

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

GenBank, EMBL and DDBJ are the most comprehensive publically available nucleotide sequence databases.

3.2: GenBank

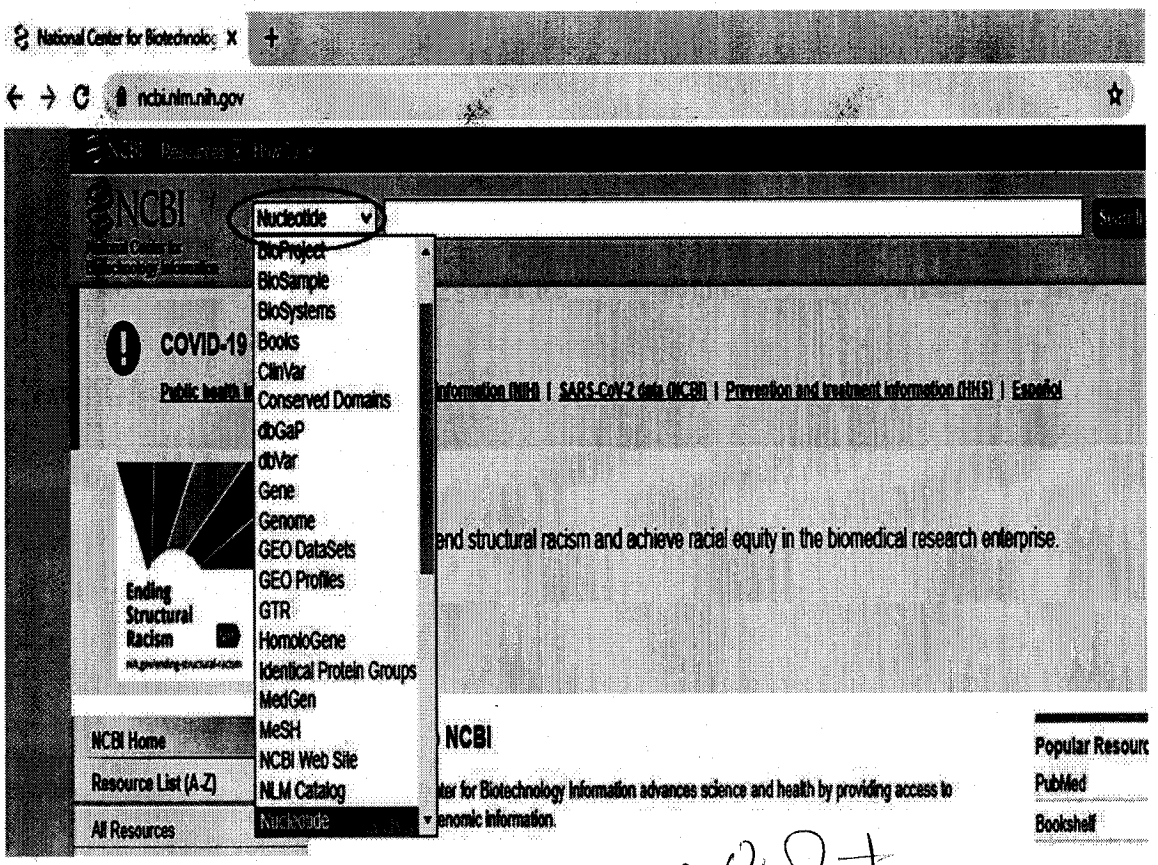
Genbank is a publically available comprehensive nucleotide sequence database which also supports bibliographic and biological annotation.

It is built and distributed by the National Center for Biotechnology Information (NCBI).

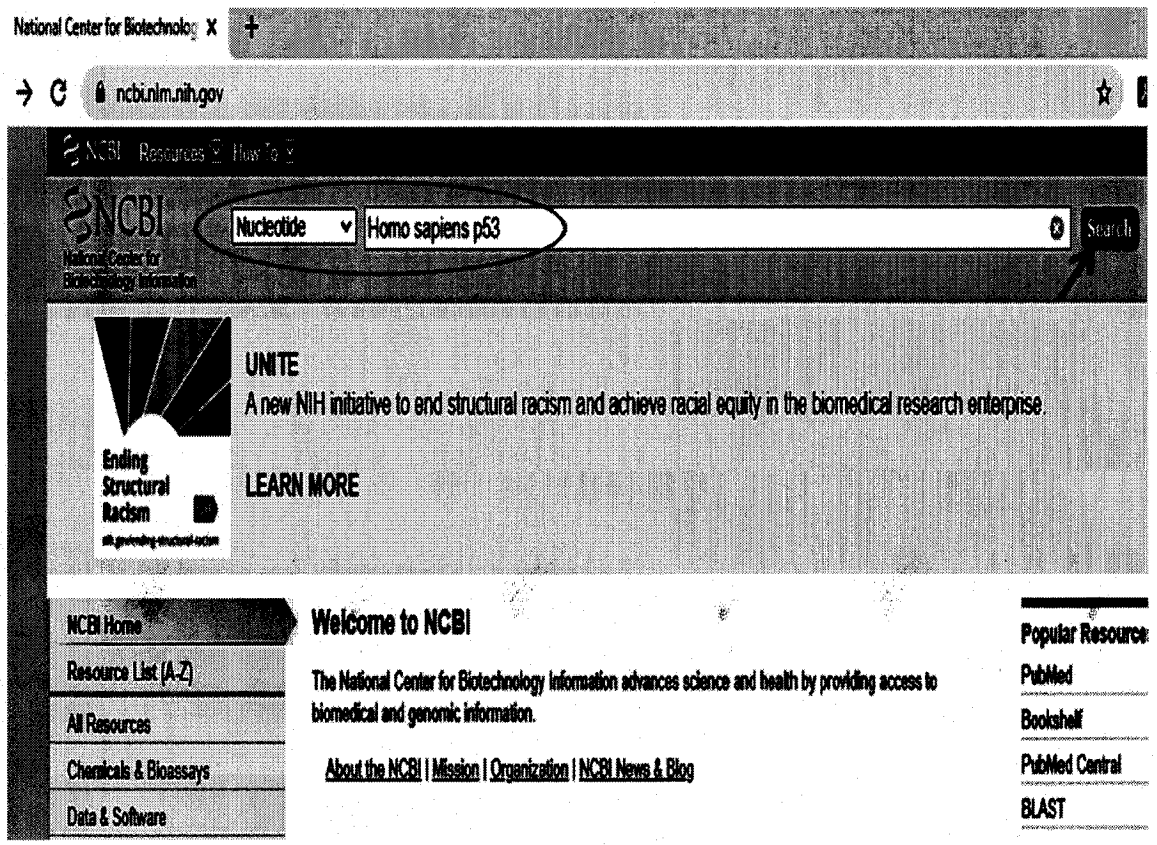
GenBank, EMBL Data Library, and (DDBJ) comprise the International Nucleotide Sequence Databases, in which information is exchanged daily to ensure a uniform and comprehensive collection of sequence information.

3.2.1: How to use it

- There are several ways to search and retrieve data from GenBank.
- To search any nucleotide sequence-
- Go to the search, select nucleotide from the dropdown, enter the name of any gene whose sequence you want to know and click search

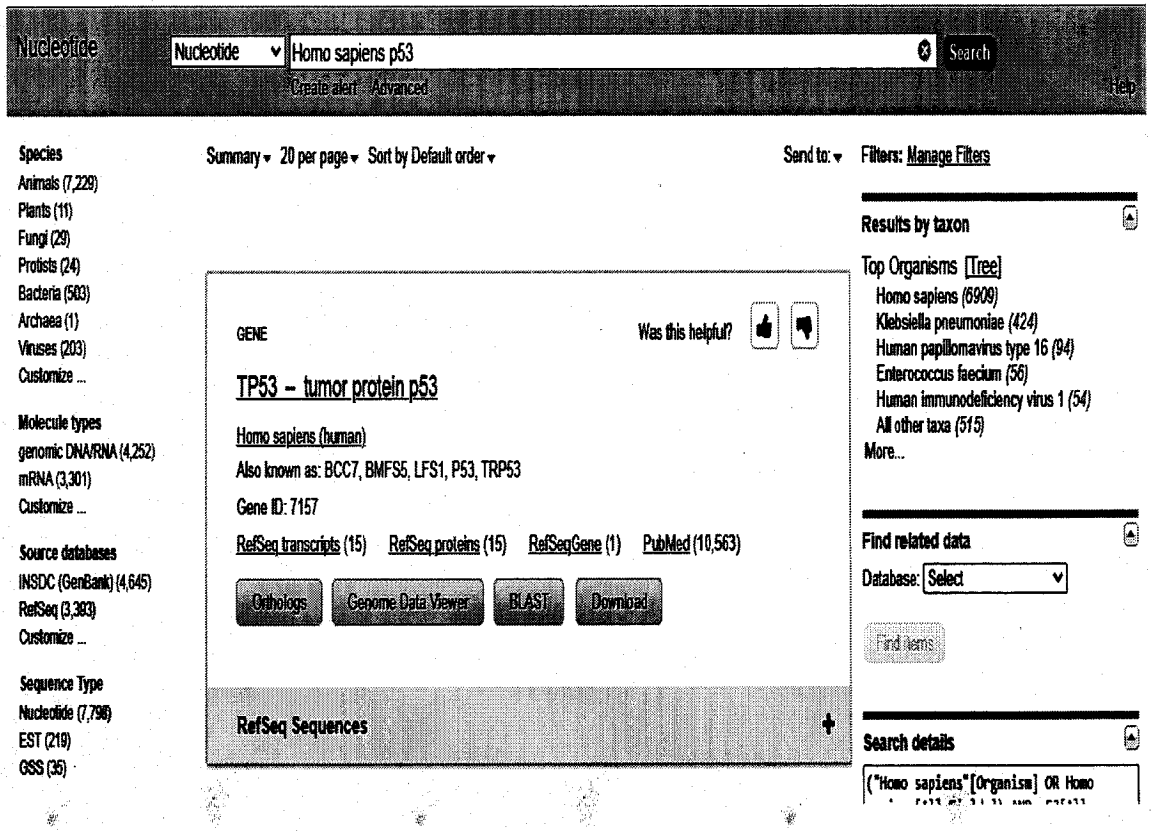


28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



28/12/2021

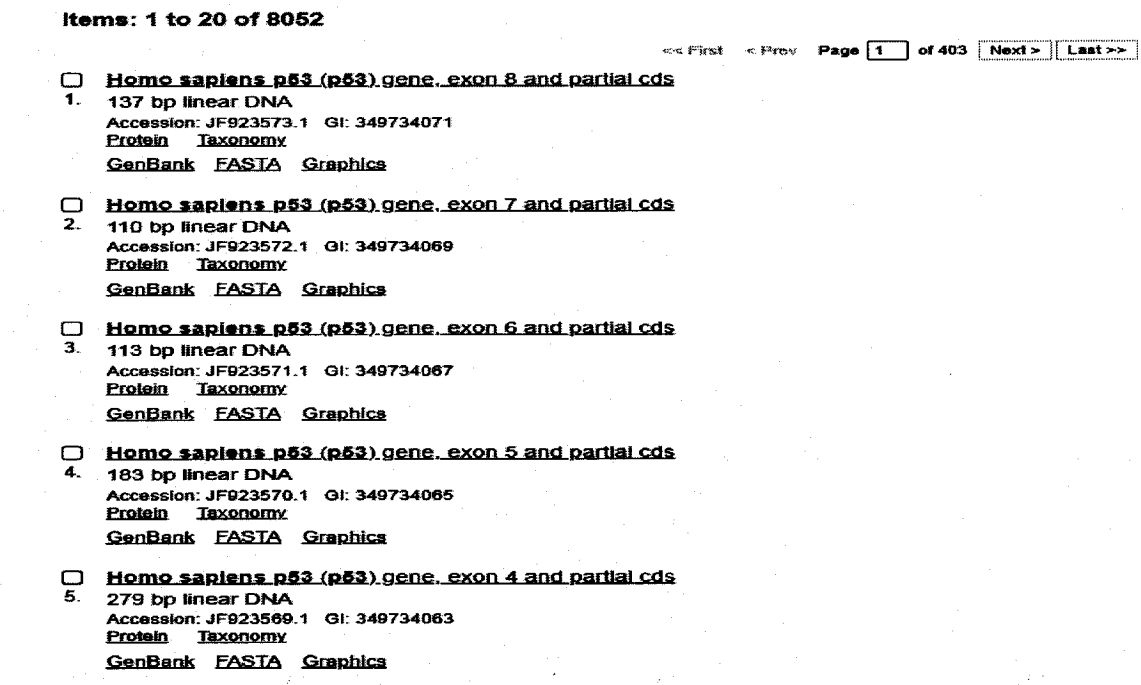
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)



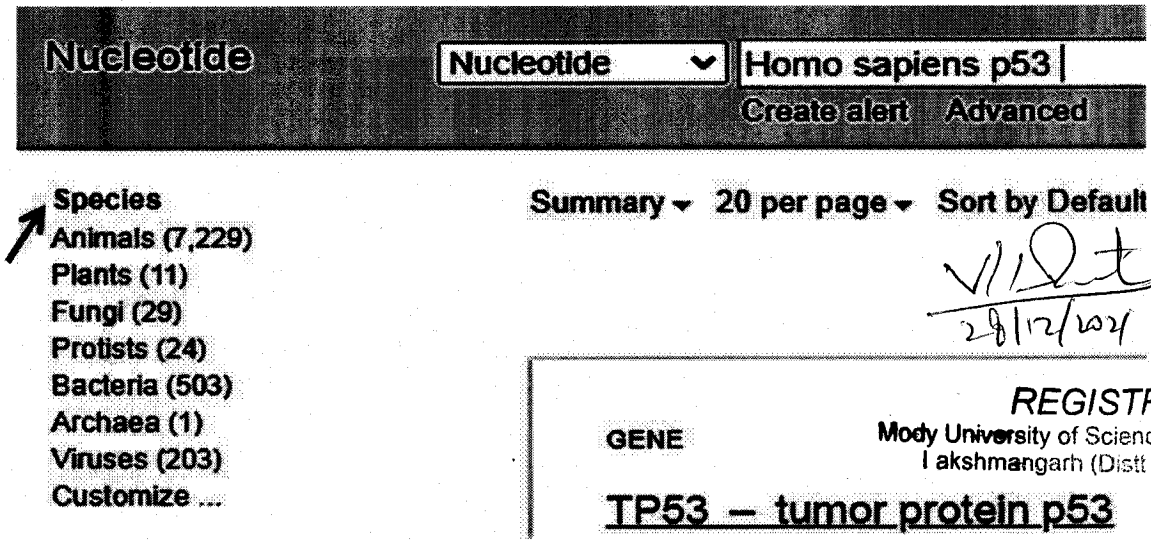
28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

On top of the results, we have total number of hits obtained against our query

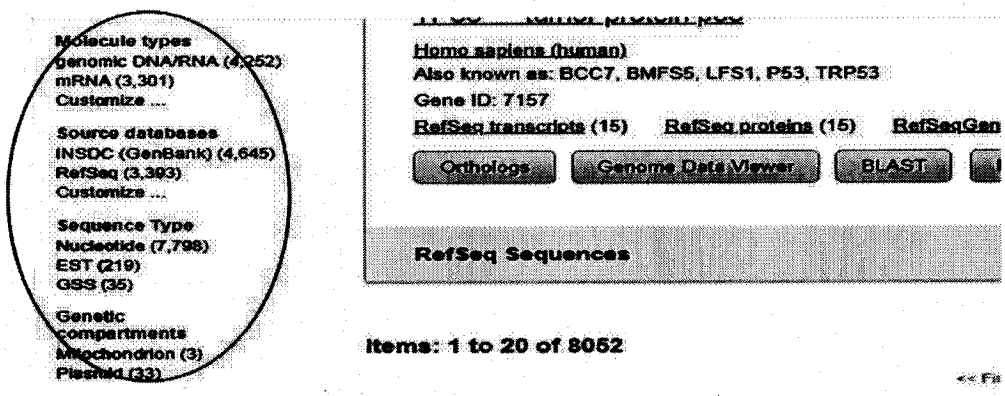


On left, there is a list of species depicting no.of records against various organisms

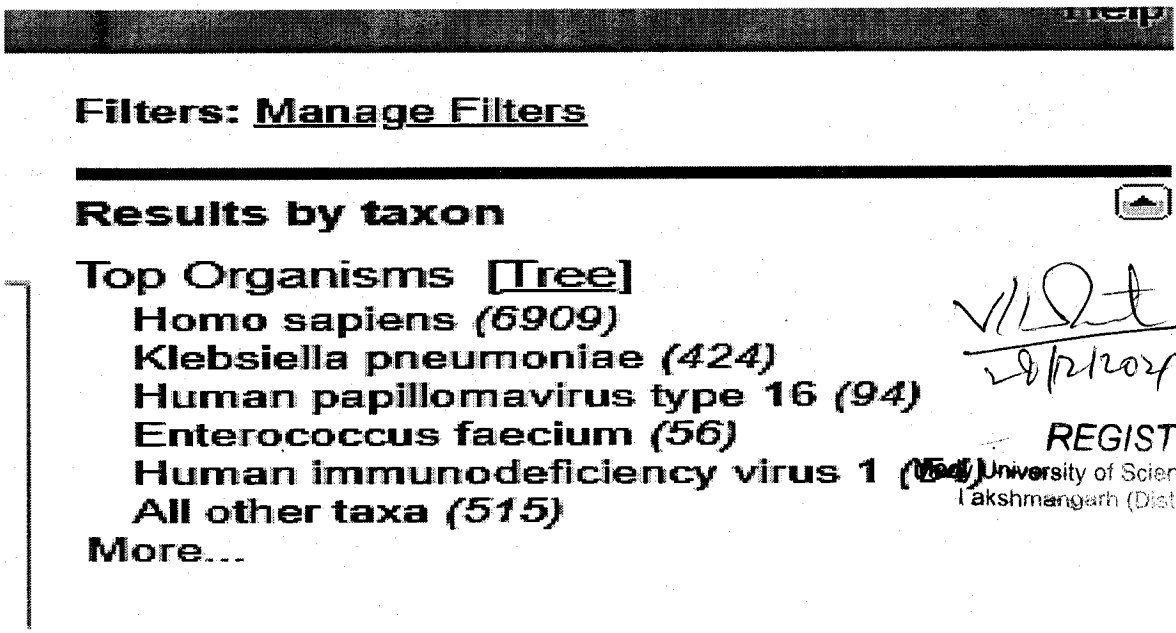


REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

We also get information about the source type, types of molecules presented in our results and also the type of sequences for example: nucleotide or GSS or EST.



On the right hand, we have results by Taxon from where we can further refine our search criteria.



REGISTRAR
University of Science and Technol
Lakshmangarh (Distt Sikar)

In the results that we have obtained, we can get some information on top based on which we can further make our decision on the hit to select as our final result. On top we have information like the length of the sequence in base pairs, the type of sequence i.e whether the sequence is of DNA or RNA, the name of the organism etc.

1. **Homo sapiens p53 (p53) gene, exon 8 and partial cds**
   137 bp linear DNA
   Accession: JF923573.1 GI: 349734071
   Protein Taxonomy
   GenBank FASTA Graphics

2. **Homo sapiens p53 (p53) gene, exon 7 and partial cds**
   110 bp linear DNA
   Accession: JF923572.1 GI: 349734069
   Protein Taxonomy
   GenBank FASTA Graphics

3. **Homo sapiens p53 (p53) gene, exon 6 and partial cds**
   113 bp linear DNA
   Accession: JF923571.1 GI: 349734067
   Protein Taxonomy
   GenBank FASTA Graphics

4. **Homo sapiens p53 (p53) gene, exon 5 and partial cds**
   183 bp linear DNA
   Accession: JF923570.1 GI: 349734065
   Protein Taxonomy
   GenBank FASTA Graphics

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

GenBank FASTA Graphics

9. Homo sapiens p53 gene for P53, exon 5, partial cds
184 bp linear DNA
Accession: AB118156.1 GI: 42821409
Protein Taxonomy
GenBank FASTA Graphics

10. Homo sapiens mRNA for P53, complete cds
2,451 bp linear mRNA
Accession: AB082923.1 GI: 23491728
Protein PubMed Taxonomy
GenBank FASTA Graphics

11. Homo sapiens p53 gene, exon 7 and partial cds
110 bp linear DNA
Accession: DQ075362.1 GI: 70611297
Protein Taxonomy
GenBank FASTA Graphics

Once we have chosen the result of our interest, we land on the main page of our result.

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1
FASTA Graphics

Go to:

```
LOCUS       JF923571                113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-MAY-2011) Department of Medicine, Institute of
            Medical Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh
            221005, India
```

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

An accession no that is a unique identifier for all the entries that are made in GenBank.

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

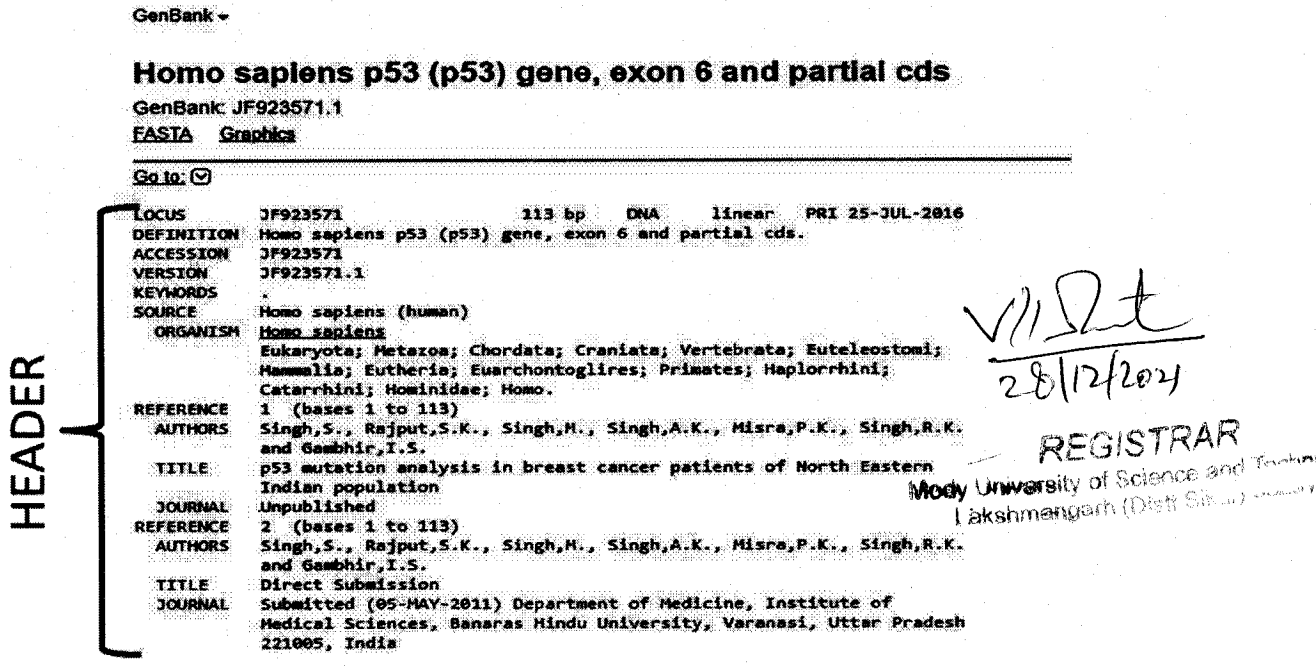
GenBank: JF923571.1

FASTA Graphics

Go to:

```
LOCUS       JF923571                 113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
```

The header section



28/12/2024

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Dist. Sikar)

The features section

```
FEATURES             Location/Qualifiers
     source          1..113
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /isolation_source="healthy case in northeastern Indian
                     population"
                     /db_xref="taxon:9606"
     gene            <1..>113
                     /gene="p53"
     mRNA            <1..>113
                     /gene="p53"
                     /product="p53"
     CDS             <1..>113
                     /gene="p53"
                     /codon_start=3
                     /product="p53"
                     /protein_id="AEQ16484.1"
                     /translation="LAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPE"
     exon            1..113
                     /gene="p53"
                     /number=6
```

Origin section

```
ORIGIN
        1 gtctggcccc tcctcagcat cttatccgag tggaaggaaa tttgcgtgtg gagtatttgg
       61 atgacagaaa cacttttcga catagtgtgg tggtgcccta tgagccgcct gag
//
```

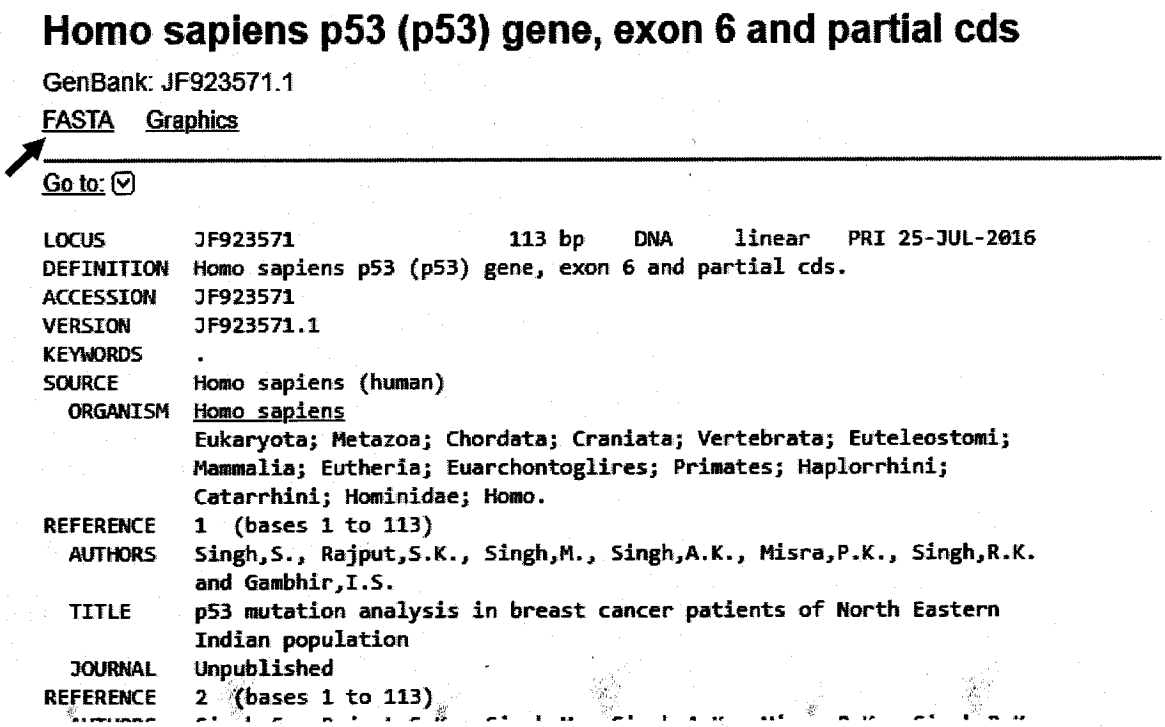
28/12/2024

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Dist Sikar)

To retrieve the sequence, click on FASTA at the top of the page



Homo sapiens p53 (p53) gene, exon 6 and partial cds

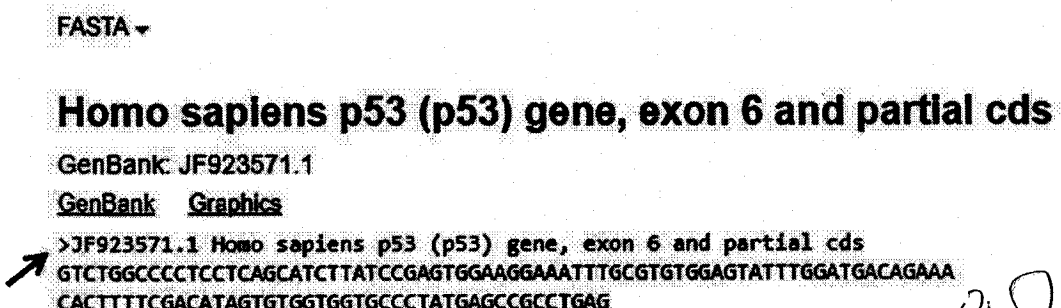
GenBank: JF923571.1

FASTA Graphics

Go to:

```
LOCUS       JF923571                 113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
```

FASTA formats start with a more than symbol. The sequences is in continues form without any gap between two base pairs.



FASTA

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1

GenBank Graphics

```
>JF923571.1 Homo sapiens p53 (p53) gene, exon 6 and partial cds
GTCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAA
CACTTTTCGACATAGTGTGGTGGTGCCCTATGAGCCGCCTGAG
```

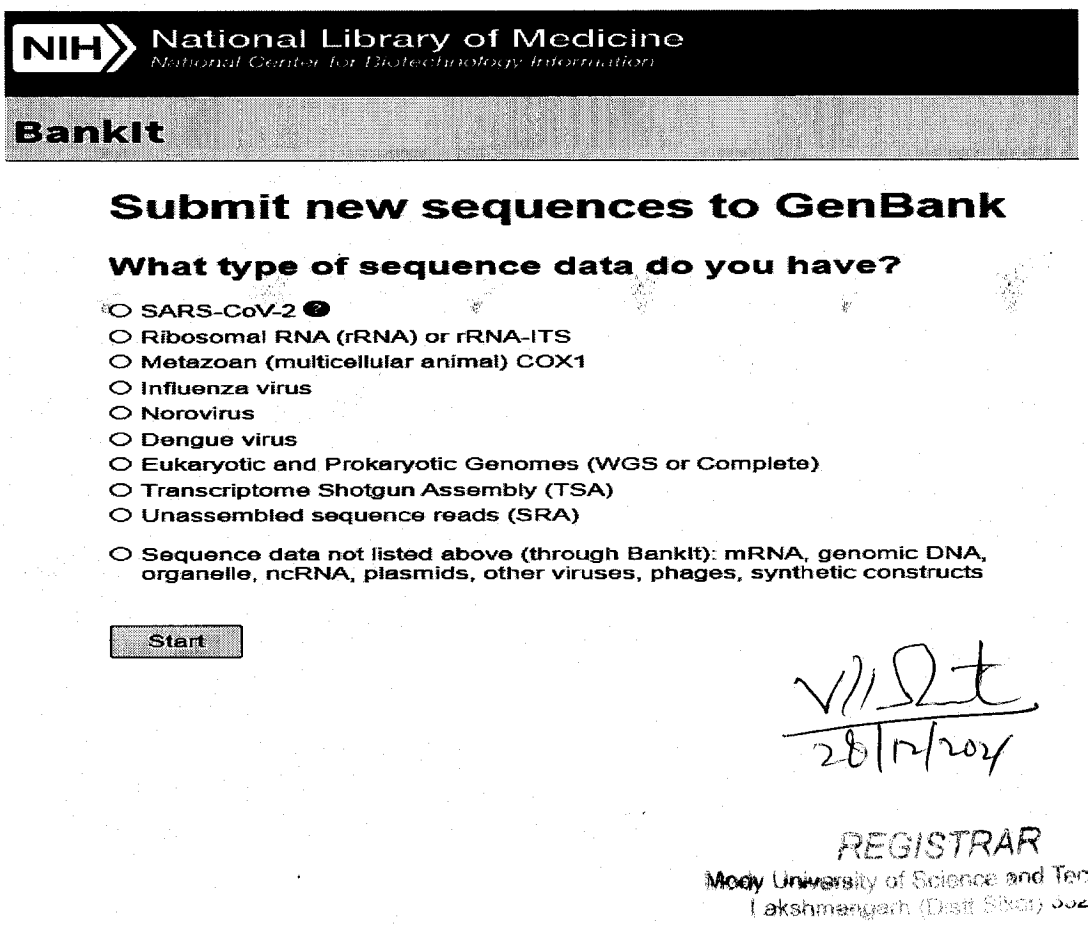
8/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

3.2.2: Submitting data into GenBank

- **Web-based submission tools**
  - **BankIt**
  - **Submission Portal**
  - **tbl2asn**

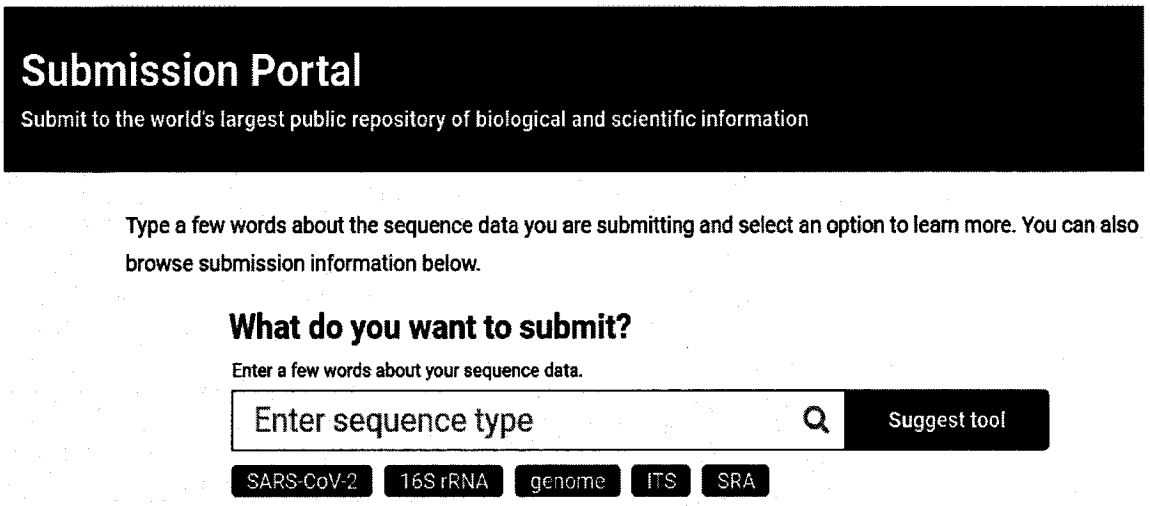
3.2.2.1: BankIt



28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3.2.2.2: Submission Portal



- Supports the submissions of
  - ribosomal RNA (rRNA),
  - metazoan mitochondrial COX1,
  - Influenza,
  - Norovirus,
  - Dengue or
  - SARS-CoV-2 sequences
- One can also submit Genome and Transcriptome Assemblies through the Genomes and TSA portals, respectively.
- Submission preparation tools are required for uploading via the Submission Portal.

We can also email to gb-sub@ncbi.nlm.nih.gov

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

3.2.2.3: tbl2asn

- a command-line program that automatically creates the sequence of records for submission to the GenBank.
- used primarily for submission of complete genomes and large batches of sequences.

Genome Workbench

- offers a rich set of integrated tools to study and analyze the genetic data.
- The allows us to prepare submissions of single eukaryotic and prokaryotic genomes.
- One can also use Genome Workbench for editing and visualizing files that are created by tbl2asn.

3.3: The European Molecular Biology Laboratory

- It is the primary nucleotide sequence resource maintained by the European Bioinformatics Institute (EBI), situated in the United Kingdom.
- In this The sequences of DNA or RNA are submitted directly from individual researchers.
- The EMBL Database in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Data Bank of Japan (NIG, Mishima) has been processing nucleotide sequence data reported world-wide since 1982.

3.3.1: How to use it

- Access to the sequence data is provided via ftp and several WWW interfaces.
- The Sequence Retrieval System or SRS allows the databases to be searched using, for example, sequence, annotations, keywords, author names.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh

Format

Sequence entry starts with an identifier line ("ID")

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").


```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```


REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

3.4: DNA Data Bank of Japan

DDBJ initiated the DNA data bank activities in 1986 at the National Institute of Genetics (NIG).

It collects DNA sequence data mainly from Japanese and researchers from all over the world.

3.4.1: How to submit the data

- For submission, It offers interactive web interface as well as mass submission system.
- SAKURA is a Web-based tool of DDBJ which is used for nucleotide sequence submission and annotation.

3.5: Protein sequence databases

- The Protein database is a collection of sequences from several sources
  - translations from annotated coding regions in GenBank, RefSeq and TPA
  - records from SwissProt, PIR, PRF, and PDB
  - Protein sequences are the fundamental determinants of biological structure and function.

3.5.1: GenPept

- It is a part of the National Center of Biotechnology Information (NCBI)
- Its entries are derived from translations of the sequences present in Genbank, DDBJ and EMBL.
- The entries lack additional annotation and the database does not contain proteins derived from amino acid sequencing.
- The format is same as that of Genbank having header, features and origin sections.
- The format ends with double slashes.
-

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh

3.5.2: Refseq

- RefSeq stands for Reference sequence which is a collection developed by NCBI.
- It provides a non-redundant collection of reference protein sequences.
- It has sequences for some limited organisms which includes human, mouse, rat, honeybee, zebrafish, cow, sea urchin, and several important plant species and also some 1100 viruses and 150 bacteria

3.5.3: Protein information resource

- It is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies.
- PIR was established in 1984 by (NBRF) the National Biomedical Research Foundation for the identification and interpretation of protein sequence information.
- It assists with proteomic and genomic data integration and the propagation and standardization of protein annotation.
- It is the oldest universal curated protein sequence database

LAKSHMANGARH • RAJASTHAN

3.5.4: Swiss Prot

- It is the leading universal curated protein sequence database.
- The database is non-redundant and is highly integrated with other databases.
- Literature-based curation is used to extract experimental data. This experimental knowledge is supplemented by manually confirmed results from various sequence analysis programs.

3.5.5: Translated EMBL Nucleotide Sequence Data Library

- It is a computer annotated protein sequence database.
- TrEMBL translates the nucleotide sequences in the DDBJ/EMBL/GenBank nucleotide sequence database.
- It was introduced to make sequences available quickly.
- TrEMBL has two main sections; SP-TrEMBL and REM-TrEMBL

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries which should be incorporated into SWISS-PROT.
- REM-TrEMBL (REMaining TrEMBL) contains the entries that are not to be included in SWISS-PROT.
- Most REM-TrEMBL entries are immunoglobulins and T-cell receptors.

3.5.6: Uniprot

- It stands for Universal Protein Resource
- It is a publically available protein sequence database.
- UniProt is a collaboration between the EMBL-EBI, the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).
- It is a comprehensive resource for protein sequence and annotation data.

It provides information on

- function,
- sub-cellular location,
- expression,
- interaction,
- structure,
- protein family,
- domain and
- sequence.

28/12/202
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Apart from this, one can also find various published papers in support of the information provided.

Important links

- https://www.ncbi.nlm.nih.gov/genbank/
- https://www.ncbi.nlm.nih.gov/WebSub/
- https://submit.ncbi.nlm.nih.gov/
- https://www.ncbi.nlm.nih.gov/protein/
- https://www.ebi.ac.uk/
- https://www.ddbj.nig.ac.jp/index-e.html
- https://www.ncbi.nlm.nih.gov/refseq/
- https://www.uniprot.org/

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

4. STRUCTURE DATABASES

Structure databases or databanks contain atomic coordinates, bibliographic citations, and other information related to structural studies. Structural databases are the essential tools for all crystallographic works. They are used in the process of producing, solving, refining and publishing the structure of a new material. Before deciding to synthesize a new compound the database could be used to check how many compounds with a particular chemical composition have been reported.

After synthesizing and indexing the unit cell of a material the database can be searched to see if a material with the same or a similar unit cell is already known. If a material is found in the database with a similar unit cell to the new material, then its structure may be close enough (i.e. same symmetry and similar unit cell contents) to be used as the starting model for the refinement of the new material. To verify the results of a structure refinement the database can be consulted to find structures that have comparable bond distances, bond angles or coordination environments to the new structure.

MODY
28/10/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

4.1: Protein Data Bank (PDB)

- Protein Data Bank archive is a single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.
- The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.
- Currently, Most of the structures are determined by X-ray diffraction, but about 10% of structures are determined by protein NMR.
- The .pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity.
- Its fixed-column width format is limited to 80 columns, which was based on the width of the computer punch cards that were previously used to exchange the coordinates.

```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

The format has a Header section along with the title and author records. This provide information related to the researchers who defined the structure.

Remark can contain free-form annotation, but they also accommodate standardized information.

Atom describe the coordinates of the atoms that are part of the protein.

Hetatm describe coordinates of hetero-atoms, that is those atoms which are not part of the protein molecule.

Ter indicates the end of a chain of residues.

REGISTRAR
Lakshmangarh (Distt Sikar)

4.2: RCSB PDB

- RCSB PDB stands for Research Colaboratory for Structural Bioinformatics PDB.
- This resource is supported by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies
- As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.
- 178229 biological macromolecular structures enabling breakthroughs in research and education.

4.3: SCOP

- SCOP stands for Structural Classification of Proteins.
- Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin.
- It aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- includes 68,816 non-redundant domains representing 772,354 protein structures

LNIVERSITY
LAKSHMANGARH • RAJASTHAN

The main levels of the classification are:

- *Family-* It groups closely related proteins with a clear evidence for their evolutionary origin.
- In most cases, their relationship is detectable with current sequence comparison methods, e.g. BLAST, PSI- BLAST, HMMER.
- *Super-family-* It brings together more distantly related protein domains.
- Their similarity is frequently limited to common structural features that along with a conserved architecture of active or binding sites, or similar modes of oligomerization, suggest a probable evolutionary ancestry.
- *Fold* groups super-families on the basis of the global structural features shared by the majority of their members. These features are the composition of the secondary structures in the domain core, their architecture and topology.
- Fold is an attribute of a super-family but the constituent families of some super-families that have evolved distinct structural features can belong to a different fold.

28/11/2024
REGISTRAR
Lakshmangarh (D

- *IUPR* (Intrinsically Unstructured Protein Region) organizes super-families of proteins or protein regions that do not adopt globular folded structure.
- *Classes* bring together folds and IUPRs with different secondary structural content.
- These include all-alpha and all-beta proteins, containing predominantly alpha-helices and beta-strands, respectively, and 'mixed' alpha and beta classes (a/b) and (a+b) with respectively alternating and segregated alpha-helices and beta-strands, and the fifth class of small proteins with little or no secondary structures.

4.3.1: ORGANISATION AND FACILITIES OF SCOP

SCOP is organized as a tree structure.

- Entering at the top of the hierarchy the user can navigate through the levels of Class, Fold, Super-family, Family and Species to the leaves of the tree which are structural domains of individual PDB entries.
- An alternative hierarchy of Folds, Super-families and Families by the date of solution of the first representative structure is also provided.

We can search in SCOP

- With the help amino acid sequence,

  The Sequence similarity search facility allows any sequence of interest to be searched against databases of protein sequences classified in SCOP using the algorithms BLAST, FASTA or SSEARCH.

- From a keyword

  The keyword search facility returns a list of SCOP pages containing the word entered or combinations of words separated by a series of boolean operators.

- From a PDB identifier

REGISTRAR
Lakshmangarh (Distt Sikar)

The PDB entry viewer links PDB entries to various graphical views, external databases and SCOP itself.

4.3.2: OTHER USES OF SCOP

- Non-redundant sequence databases are available from SCOP and it helps in the evaluation of sequence alignment methods.
- Assignment of protein structures to sequences using the intermediate sequence library PDB-ISL.
- Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL.

4.4: CATH

- CATH database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains.
- The CATH database provides hierarchical classification of protein domains based on their folding patterns.
- Domains are obtained from protein structures deposited in the Protein Data Bank and both domain identification and subsequent classification use manual as well as automated procedures.
- The CATH database is a classification of protein domains, based not only on **sequence information**, but also on **structural and functional properties**.
- CATH contains **114,215 domains**.
- It is classified in a hierarchical scheme with four main levels called **class (C)**, **architecture (A)**, **topology (T)** and **homologous** super-family **(H)**.
- From these words they have derived the name CATH.

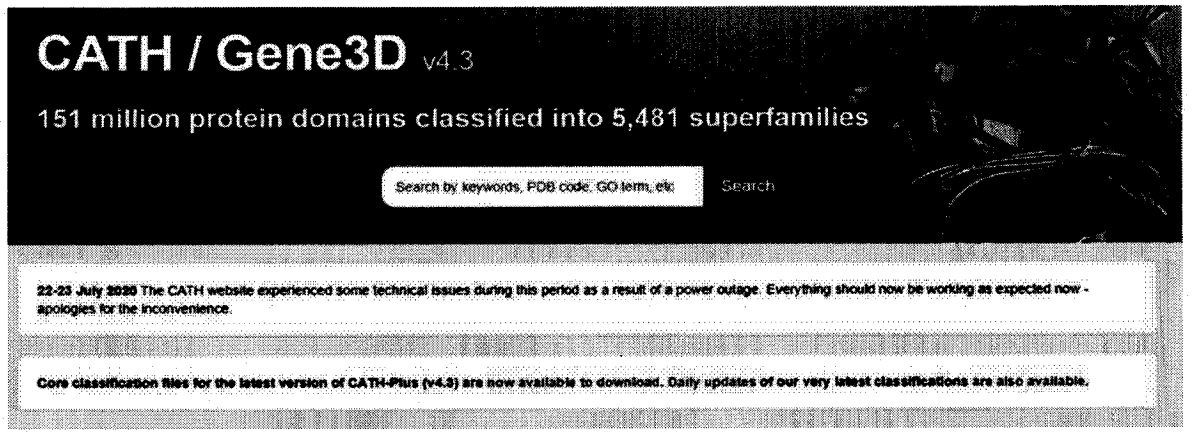
MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN
28/12/2024
REGISTRAR
Moody University of Science and Technology
Lakshmangarh (Distt Sikar)

- At the **C-level**, domains are grouped according to their secondary structure content into four categories: mainly alpha, mainly beta, mixed alpha-beta; and the last category contains domains with only few secondary structures.
- The **A-level** groups domains according to the general orientations of their secondary structures.
- At the **T-level**, the connectivity (ie the order) of the secondary structures is taken into account.
- The grouping of domains at the **H-level** is based on a combination of both sequence similarity.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

4.4.1: Searching in CATH



MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)

IMPORTANT LINK

- **https://scop.mrc-lmb.cam.ac.uk/**
- **https://www.cathdb.info/**
- **https://www.rcsb.org/**
- **https://www.wwpdb.org/**

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

5: STRUCTURE DATABASES (Part 2)

5.1: PFAM

- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
- Pfam, version 27.0, contains 14 831 Pfam-A families. Of these families, 4563 have been classified into 515 clans.
- Pfam also generates higher-level groupings of related entries, known as clans.
- A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.
- Profile HMMs turn a multiple sequence alignment into a position-specific scoring system, which can be used to align sequences and search databases for remotely homologous sequences.
- Profile HMMs have a formal probabilistic basis and uses consistent theory behind gap and insertion scores.
- Profile hidden Markov models (HMMs) have several advantages.
- HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency.
- The HMMER package, is a set of programs that allow us to create and manipulate profile HMMs and databases of profile HMMs, perform sensitive searches of sequence and profile HMM databases and create multiple sequence alignments efficiently

5.1.1: FORMAT of PFAM

- **Plain Pfam format**
- **Flatfile**

20/12/202/
REGISTRAR
Maity University of Science and Technology
Lakshmangarh (Distt Sikar)

```
YHI1_MYCLE/3-102      TIFTKIINRELPGRFVYEDDDVVAFLTIEPMTQGHTLVVP..CAEIDQMQMVDP...AI....FGRVIAVSQLIGKGVCRAFNAERAGVII.....AGFEVPHLHIHVFPTHSL
YHIT_MYCHR/7-111      ELFLKIIKREEPATILYEDDKVIAFLDKYAHTKGHFLVVP..KNYSRNLFSISD...ED....LSYLIVKAREFALQEIKKLGATGFKLLINNEPDAEQSIFHTHVHIIPYYKK
YHIT_HELPY/2-104      NVFEKIIQGEIPCSKILENERFLSFYDINPKAKVHALVIP..KQSIQDFNGITP...EL....MAQMTSFIFEVVEKL..GIKEKGYKLLTNVGKNAGQEVMHLHFHILSGDKH
YHIT_AZOBR/9-115      NVFARILRGEIPCKKVLETEHALAFHDINPQAPTHILVIP..KGAYVDMDDFSA...RATEAEIAGLFRAVGEVARGA..GAAEPGYRILSNCGEDANQEVPHLHIHVFAGRRL
PKIA_DICDI/68-172     TIFAKIISGAIPSKKFYDDEYCIAIEDINPQAPVHLLVIP..KLAVGGLSDVAMVDLEKYKESMGHTMSKIHHIASLK....GADSYRLVINEGVLGQQSVRNLHIHIPNN...
ZB14_MAIZE/17-123     TIFDKIIKKEIPSTVVYEDEKVLAFRDINPQAPTHILIIPKVKDGLTGLAKAEE...RHIEI.LGYLLYVAKVVAKQE..GL.EDGYRVVINDGPSGCQSVVHIHVHLLGGRQM
Y961_HAEIN/5-110      TIFSKIIRKEIPANIVYQDELVTAFRDISPQAKTHILIIP..NKVIPTVNDVTE...QDEVA.LGRLFSVAAKLAKEE..GVAEDGYRLIVNCNKHGGQEVFHLHMHLVGGEPL
YCFF_ECOLI/5-110      TIFSKIIRREIPSDIVYQDDLVTAFRDISPQAPTHILIIP..NILIPTVNDVSA...EHEQA.LGRHITVAAKIAEQE..GIAEDGYRLIMNTNRHGGQEVYHIHMHLLGGRPL
YHIT_CAEEL/21-125     TLFGKIIRKEIPAKIIFEDDEALAFHDVSPQAPIHFLVIP..KRRIDMLENAVD...SDAAL.IGKLMVTASKVAKQL..GM.ANGYRVVVNNGKDGAQSVFHLHLHVLGGRQL
HINT1_RAT/16-120      TIFGKIIRKEIPAKIIFEDDRCLAFHDISPQAPTHFLVIP..KKYISQISAAED...DDESL.LGHLMIVGKKCAADL..GLK.KGYRMVVNEGSDGGQSVYHVHLHVLGGRQM
YHIT_SYNY3/5-109      TIFSKIIRREIPAAIVYEDDLCLAFKDVNPQAPVHVLLIP..KKPLPQLSAATP...EDHAL.LGHLLLKAKEVAADL..GI.GDQFRLVINNGAEVGQTVFHLHLHILGGRPF
YHIT_SYNP7/5-109      TIFGKIIRREIPADIVYEDDLCLAFRDVAPQAPVHILVIP..KQPIANLLEATA...EHQAL.LGHLLLTVKAIAAQE..GL.TEGYRTVINTGPAGGQTVYHLHIHLLGGRSL
Y866_METJA/2-105      CIFCKIINGEIPAKVVYEDEHVLAFLDINPRNKGHTLVVP..KKHYERFDEMPD...DE....LCNFIKGVKKTVEVL.KKLGFDGYNIVMNNGRVAGQEVNHVHFHIIPRYEG
```

5.2: Nucleic acid structure databases

The 3D structure of nucleic acid can be well understood by investigating the different types of nucleic acid functions such as nucleic acid–protein interactions, functional RNAs, etc.

Nucleic Acid Database (NDB)

- The Nucleic Acid Database (NDB) was the first database for nucleic acid structures.
- It was designed in the early 1990s.
- It contains the contents of primary data.
- It also contains some derived geometric data, classification of structure and motifs, standards for describing nucleic acid features,
- NBA also has tools and software for analysis of nucleic acids.
- The structures available in the NDB are annotated with various information that are unique to the nucleic acids
- The structures of NBA are almost not available from the original entry of PDB.
- The structures of NBA are analyzed by using computational methods of structural geometries.
- They are then classified by searching the sequence, structure levels, and secondary structure by incorporating the PDB entries to the NDB.

NDB Entry Format

28/12/2021
REGISTRAR
Mody University of Scie
Lakshmangarh (
Technology

- The main content of NDB is the primary structure information that is related to the nucleic acids that are acquired from the PDB and also from classification and derived data
- The primary data present in the NDB is obtained from PDB.
- It contains experimental files, identifiers, structural descriptions, citations, crystal data, coordinate information, and various details related to the crystallization, data collection, and structure refinement
- The structures available in the NDB database possess their own personal summary page that is mostly appropriate to that particular structure.
- There are about four sections of data in the structure summary page, including
    - primary structure information,
    - downloads,
    - derived structural data, and
    - images.

Home page of NDB where one can search for the 3D structural information of DNA and RNA along with some advanced searches.

28/12/2021

REGISTRAR

Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

NUCLEIC ACID DATABASE

A Portal for Three-dimensional Structural Information about Nucleic Acids

As of 26-May-2021 number of released structures: 11337

Search DNA | Search RNA | Advanced Search

Enter an NDB ID or PDB ID

Search for released structures

Welcome to the NDB

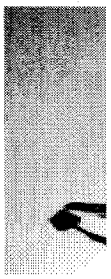
The NDB contains information about experimentally-determined nucleic acids and complex assemblies.
Use the NDB to perform searches based on annotations relating to sequence, structure and function, and to download, analyze, and learn about nucleic acids.

Search Structures

Search DNA
Search DNA and its complexes

Search RNA
Search for RNA structures in the NDB archive or in the Non-Redundant list

Advanced Search
Search for structures based on structural features, chemical features, binding modes, citation and experimental information



26/11/2024

REGISTRAR
Jain Deemed-to-be University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Featured Tools in NBA

- **RNA 3D Motif Atlas**
- **Non-redundant Lists**
- **RNA Base Triple Atlas**
- **WebFR3D**
- **R3D Align**

Tools available in NBA

Nucleic Acid tools:

- **DNA**
  - DNAProDB
  - Assignment of DNA conformers
- **RNA**
  - RNA Basepair catalog
  - RNA 3D motif Atlas
  - Representative RNA 3D Structure sets
  - R3D Align
  - WebFR3D
  - RNA viewer

FRASS

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Nucleic Acid Softwares

- FR3D
- RNAView
- RNAMLView
- 3DNA
- FreeHelix98
- Hel_param
- The UNAFold Web server

References

- http://ndbserver.rutgers.edu/
- http://pfam.xfam.org/
- Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, Alex Bateman, Pfam: The protein families database in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D412–D419, https://doi.org/10.1093/nar/gkaa913

UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

6. SEQUENCE ALIGNMENT (PART 1)

What we understand by Sequence Alignment?

- Arranging DNA, RNA or Protein sequences in order to find the regions of similarity.
- This help to derive functional, structural and evolutionary relationships between them.
- We can also make prediction about the 3D structures of proteins.
- If two sequences from two different organisms are similar then we can say there is a possibility that both sequences are Homologous i.e they share some common ancestry.

Types of Sequence Alignments

Sequence Alignment can be classified on the basis of:

1. Sequence Length-

   How much portion of the sequences under study are being covered.

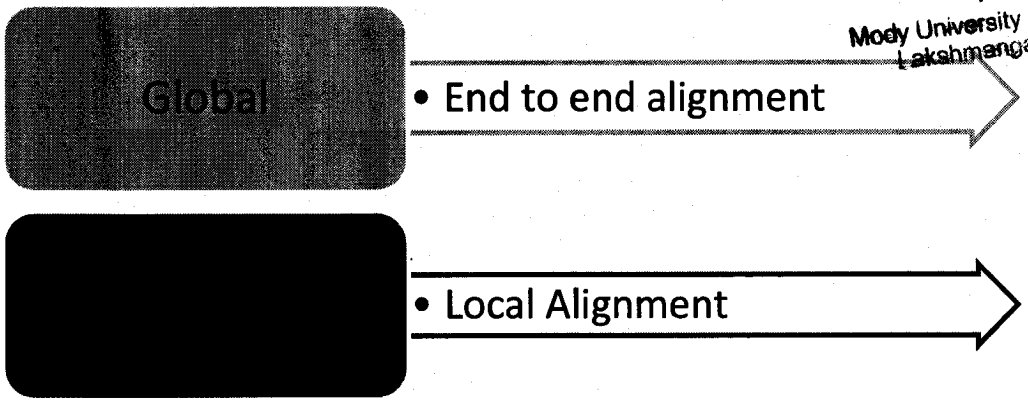
2. Number of sequence-

   If we are aligning two sequences or we are trying to align multiple sequences

MODY UNIVERSITY

Based on sequence Length

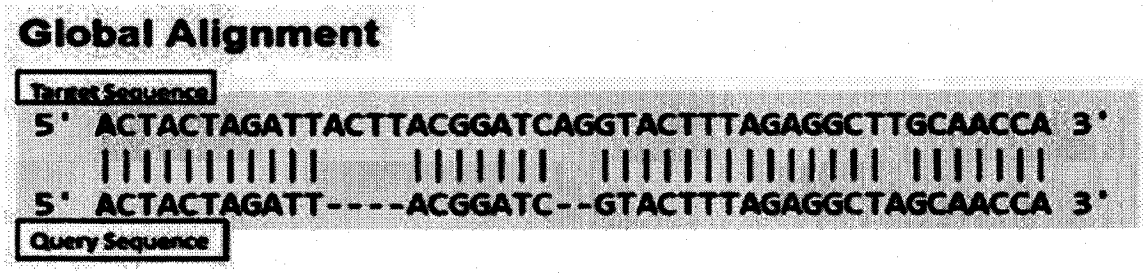
28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



Global Sequence Alignment

- In this alignment sequences are aligned along their entire length to include as many matching characters as possible.
- We usually do global alignment, when we want to compare two homologous genes or the genes that perform the same function in two different species
- Or to compare two proteins with similar function

Example of Global Sequence Alignment

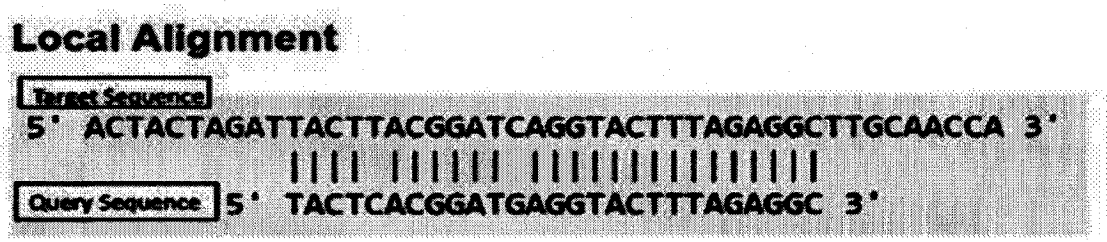


Local Sequence Alignment

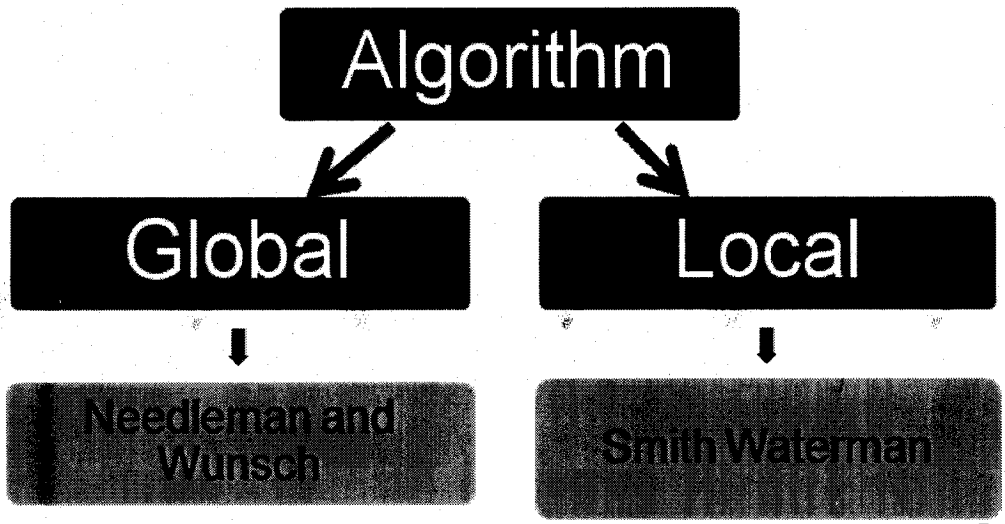
- The sequences are aligned to find a region of higher density or strong similarity.
- This helps in finding conserved domains or motifs in two protein sequences or some conserved patterns in two DNA sequences.
- The target sequence and a query sequence and our query sequence is not covering the target sequence in its entire length.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

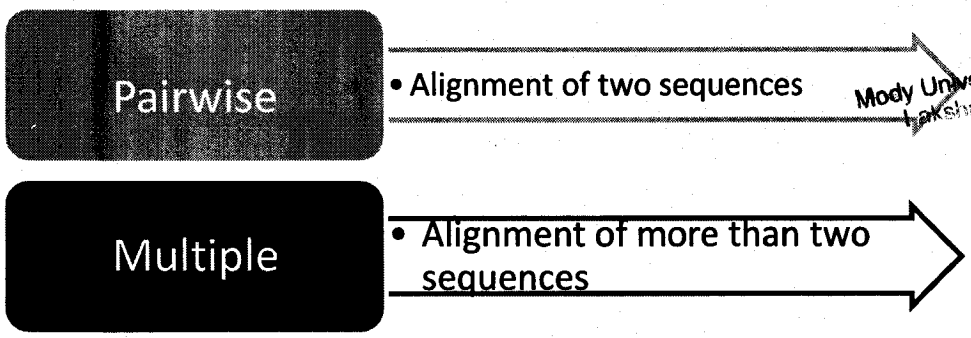
Example of Local Sequence Alignment



Algorithms for Sequence Alignments



Based on Number of Sequence



28/12/2024
REGISTRAR
Mody University of Science and Technology

Pairwise Sequence Alignment

The methods that are used for Pairwise alignment of nucleotide and protein sequences are:

1) Dot Plot method

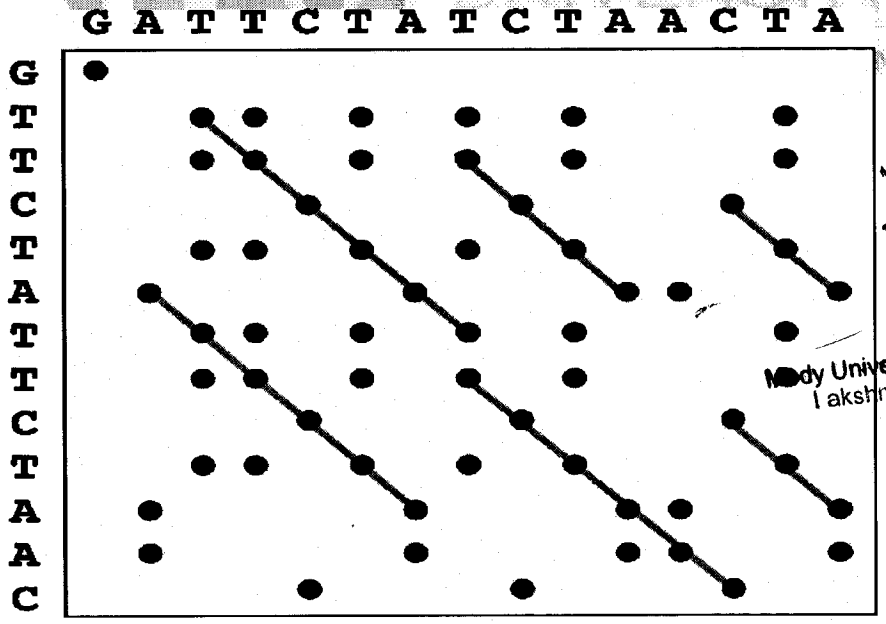
2) Dynamic Programming method

3) Heuristic Method

Dot Plot

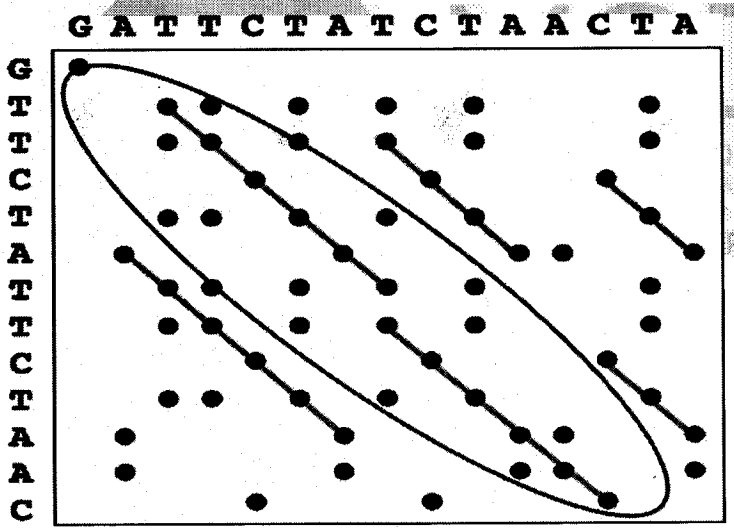
- It is graphical method used for aligning two sequences for the identification of the region of maximum similarity and dissimilarity, depicted by the presence or absence of DOTS.
- Also known as the dot matrix method.
- The comparison is done by scanning each and every residue of one sequence for similarity with all residues in the other sequence.

MODY UNIVERSITY



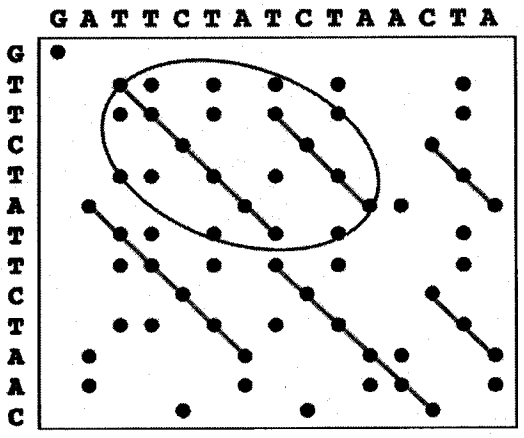
28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- We pick one residue from our query sequence and compare it with the residues on the X axis.
- We plot a dot where ever we find the same residue along the X axis.
- When the two sequences have substantial regions of similarity, many dots line up to form continuous diagonal lines, which reveal the sequence alignment as shown in example.
- If there are interruptions in the middle of a diagonal line, they indicate insertions or deletions in the sequences.
- Based on the patterns of lines in the matrix, we can say the following:
- If we obtain a Principal diagonal line, it will show identical sequence.

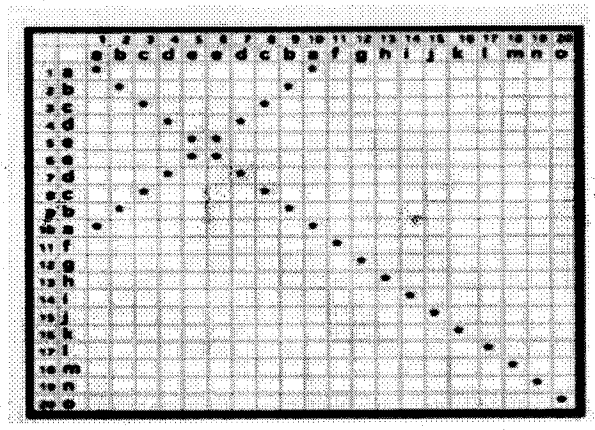


REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- If we have Multiple diagonal lines, they indicate repetition in the sequences.



- **Reverse diagonal (perpendicular to diagonal) indicate INVERSIONs**



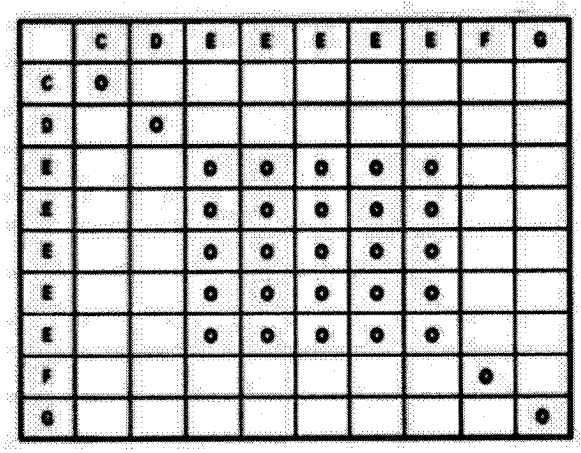
MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

- **Reverse diagonal crossing diagonal (X) indicate PALINDROMES or the sequences are palindromic.**

	A	T	G	C	G	T	A
A	*						*
T		*				*	
G			*		*		
C				*			
G			*		*		
T		*				*	
A	*						*

28/1/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- **If we obtain lines that Form of a box as shown in this example, they indicate the Low complexity region**



Limitations of Dot Matrix method

- It lacks statistical rigor in assessing the quality of the alignment.
- It is difficult for the method to scale up to multiple alignment.
- For longer sequence, memory required for the graphical representation is very high. So long sequence can not be aligned.
- Aligning larger sequences will take longer time

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt. Sikar)

Applications of Dot Matrix method

- Shows the all possible alignment between two nucleic acid and amino acid sequences
- All kind of local and global alignment can be traced
- Help to recognize large region of similarity.
- An excellent approach for finding sequence transposition.
- To find the non sequential alignment.

SOFTWARES

- Dotmatcher : https://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher
- Dottup : https://www.ebi.ac.uk/Tools/seqstats/emboss_dottup/
- MatrixPlot : http://www.cbs.dtu.dk/services/MatrixPlot/
- Dotter : https://sonnhammer.sbc.su.se/Dotter.html
- Dothelix

Dynamic Programming

- This method basically breaks down a large problem into small sub-problems and then uses the solution of the sub-problems to compute the solution of the larger problem.
- Used for global as well as for local sequence alignments using Needleman-Wunsch and Smith-Waterman algorithms respectively.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

Heuristic Method

When a single sequence is to be compared against the whole database, we use heuristic methods like BLAST and FASTA.

28/12/2021

REGISTRAR
Mody University of Science and Technology

References:

- Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press
- Lesk, A. M. (2002). Introduction to bioinformatics. Oxford: Oxford University Press.
- Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

7: SEQUENCE ALIGNMENT (PART 2)

Dynamic programming methods

- Match: this means that there is no change since their divergence. The residue of the query sequence is the same as the residue in the target sequence.
- Once we get a match, we assign some value to it, which we call as a match score
- Mismatch: this means that there is a substitution since their divergence. The residue of the query sequence is the different from the residue in the target sequence.
- It is quite common that when comparing sequences we will find that at some positions in some of the sequences there will be some extra residues which we call insertions, or missing residues which we call deletions.
- A gap in one of the sequences simply means that one or more residues have been deleted from the sequence, or we could also say that there is an insertion of residues in the second sequence.
- When aligning sequences, introducing gaps in the sequences can allow an alignment algorithm to match more terms in comparison to a gap-less alignment can.
- to get a useful alignment, minimizing gaps in an alignment is important.
- Too many gaps can lead to a meaningless alignment
- Gap penalties are used to adjust alignment scores based on the number and length of gaps.
- It can be changed each time an alignment is run. This will affect the number of gaps, their length and position in the sequence alignment.

DYNAMIC PROGRAMMING METHOD

- A method that determines optimal alignment by matching two sequences for all possible pairs of characters between them.
- It is a more quantitative method. In this a dot matrix is converted into a scoring matrix to account for matches and mismatches between sequences.
- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

STEPS:

- **Initialization:**

The first step in the global alignment dynamic programming approach is to create a matrix with M + 1 columns and N + 1 rows where M and N correspond to the size of the sequences to be aligned.

M is no.of residues in target sequence

N means no.of residues in query sequence.

Matrix filling(scoring):

We fill the matrix with highest possible scores. For the calculation of the scores, we need the match score, mismatch score and the gap penalty. We have three calculations based on which we decide the highest score to be filled in the matrix.

These calculations are:

$$M_{i,j} = Maximum\ [M_{i-1,j-1} + S_{i,j},\ M_{i,j-1} + W,\ M_{i-1,j} + W]$$

Traceback and aligning :

Move from last corner and follow arrow.

Now lets consider the two given sequences.

ACTGATTCA

ACGCATCA

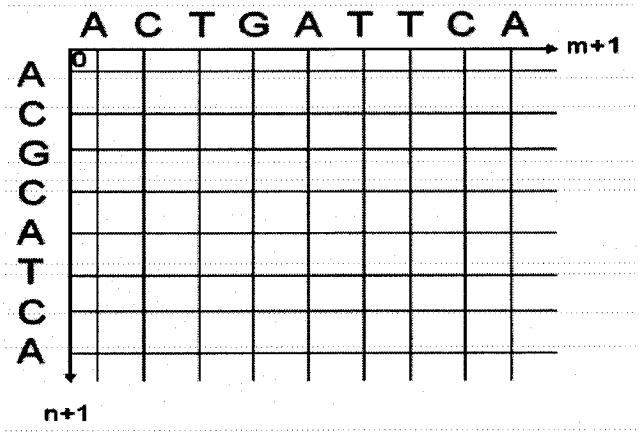
The length of the sequence 1 and sequence 2 are 9 and 8 respectively.

For match..we will give a score of 2, for mismatch -3 and gap penalty is -2

This is the M+1 and N+1 matrix

Then the scoring schema has to be introduced which can be user defined with specific scores.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)



Initialization Step

Let's assumes that there is gap penalty.

First row and first column of the matrix can be initially filled with 0.

If the gap score is assumed, the gap score can be added to the previous cell of the row or 0^{th} row and to the 0^{th} column as shown here.

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Matrix Fill Step

The second and crucial step of the algorithm is matrix filling starting from the upper left hand corner of the matrix.

From the assumed values, add the match or mismatch (assumed) score to the diagonal value.

$$M_{1,1} = Max\ [M_{0,0} + S_{1,1},\ M_{1,0} + W,\ M_{0,1} + W]$$

•

Lets generalize this equation

M(I,J)= M(i-1,j-1)+S(I,j)

M(i,j-1)+W

M(I-1, j)+W

M(I,j)- coordinates of the cell

S(I,j)- score corresponding the cell of given coordinate (whether it's a match or a mismatch)

W- gap penalty

We compute all the values and choose the highest score

In this example, the value of M(i-1,j-1) ie M(0,0) is 0

To this add the score of match or mismatch

We will check whether the jth residue of our query sequence is matching with the ith residue of the target sequence.

If there is a match we will add the score of match.

In this example, yes there is a match, so we will add the score of match which is 2

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar), 332311

		A
	0	-2
A	-2	2

In second equation we will look at the score in the M(I,j-1) cell

Here the value at M(1,0) is -2

And the gap panelty is -2, so the value will be -4

In equation 3 value at M(0,1) is again -2 and the gap penalty is also -2 so the value comes out to be -4

Out of all these 3 values 2 is the highest value and hence the score at M(1,1) will be 2

This is how the final matrix will appear.

Step 2: Assign scores

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Trace back Step

The final step in the algorithm is the trace back for the best alignment.

We will start our trace back from the cell with the highest score.

In this example, we have 8 at the right most bottom corner of the matrix.

We will check for the next highest value adjacent to this cell.

We have values of 4,6 and 4 in the adjacent cells

Out of these 6 is the highest value which is in the diagonally adjacent cell.

After 6 we will again look for the highest value in the same manner.

We have 4 which is also in the diagonally adjacent cell.

Once again lets check for the highest value.

Now here we have 6 which is the value of the horizontal left cell.

MODY UNIVERSITY

Step 3: Trace back

The optimal path is traced beginning from the lower right-hand corner

		A	C	T	G	A	T	T	C	A
	(0)	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	(2)	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	(4)	(2)	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	(4)	2	0	-2	-4	-6
C	-8	-4	0	-1	(2)	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	(4)	2	0	-2	2
T	-12	-8	-4	0	-2	2	(6)	(4)	2	0
C	-14	-10	-6	-2	-4	0	4	2	(6)	4
A	-16	-12	-8	-4	-5	-2	2	1	4	(8)

28/12/2021

REGISTRAR

Mody University of Science and Technology

Aligning the sequences

The diagonal lines indicate alignment

```
CA
| |
CA
```

If we are moving horizontally left, like here from 4 to 6 we will insert a gap in the sequence on the Y axis

```
:TCA
  | |
:-CA
```

And if we are moving vertically up, we will insert a gap in the sequence at the X axis or our target sequence

```
-ATTCA
 || ||
CAT-CA
```

So we have alligned A with A, C with C, then we have a gap in our query sequence so this is T of target sequence and a gap in query sequence. Then T is aligning with T, A with A and then this time we have a gap in our target sequence against C in the query sequence.

Then G is aligning with G and again we have a gap in our query sequence against T of the target sequence. Then C is aligning with C and A is in alignment with A.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

Result:

This analysis yielded the following alignment:

```
ACTG-ATTCA
||   || ||
AC-GCAT-CA
```

The alignment score is equal to the value in the lower right-hand corner of the matrix (8).

28/12/2011
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- **Softwares which are used for pairwise sequence alignment are**
- **SIM, (http://bioinformatics.iastate.edu/aat/align/align.html)**
- **SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html)**
- **LALIGN (www.ch.embnet.org/software/LALIGN form.html)**

8: MULTIPLE SEQUENCE ALIGNMENT

A Multiple Sequence Alignment is an alignment of more than two sequences.

Multiple alignments are computationally much more difficult than pair-wise alignments.

There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments can

Multiple sequence alignment (MSA) methods refer to a series of algorithmic solution for the alignment of evolutionarily related sequences, while taking into account evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions.

These methods can be applied to DNA, RNA or protein sequences.

Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment.

Multiple sequence alignment is also an essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.

Multiple sequence alignment also has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

MSA is indeed an important modeling tool whose development has required addressing a complex combination of computational and biological problems.

8.1: Commonly used algorithms

Given a set of biological sequences (RNA, proteins, DNA), the purpose of a MSA method is to align the sequences in a way that will either reflect their evolutionary, functional or structural relationship Usually the multiple sequence algorithms assume that the sequences are similar in all its length and they behave like global alignment algorithms.

They also assume that there are not many long insertions and deletions. Thus the algorithms will work for some sequences, but not for others.

there are exhaustive and heuristic approaches used in multiple sequence alignment.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

24/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

exhaustive approaches are not feasible in most cases, heuristic programs are commonly used

8.2: Progressive Alignment Method

Progressive alignment depends on the stepwise assembly of multiple alignment and is heuristic in nature.

It first conducts pairwise alignments for each possible pair of sequences using the Needleman–Wunsch global alignment method and records these similarity scores from the pairwise comparisons.

The scores can either be percent identity or similarity scores

Both scores correlate with the evolutionary distances between sequences.

The scores are then converted into evolutionary distances to generate a distance matrix for all the sequences involved.

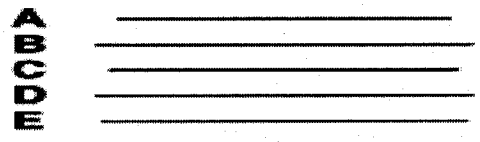
A simple phylogenetic analysis is then performed based on the distance matrix to group sequences based on pairwise distance scores. As a result, a phylogenetic tree is generated.

The tree reflects evolutionary proximity among all the sequences.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

the resulting tree is an approximate tree and does not have the rigor of a formally constructed phylogenetic tree.

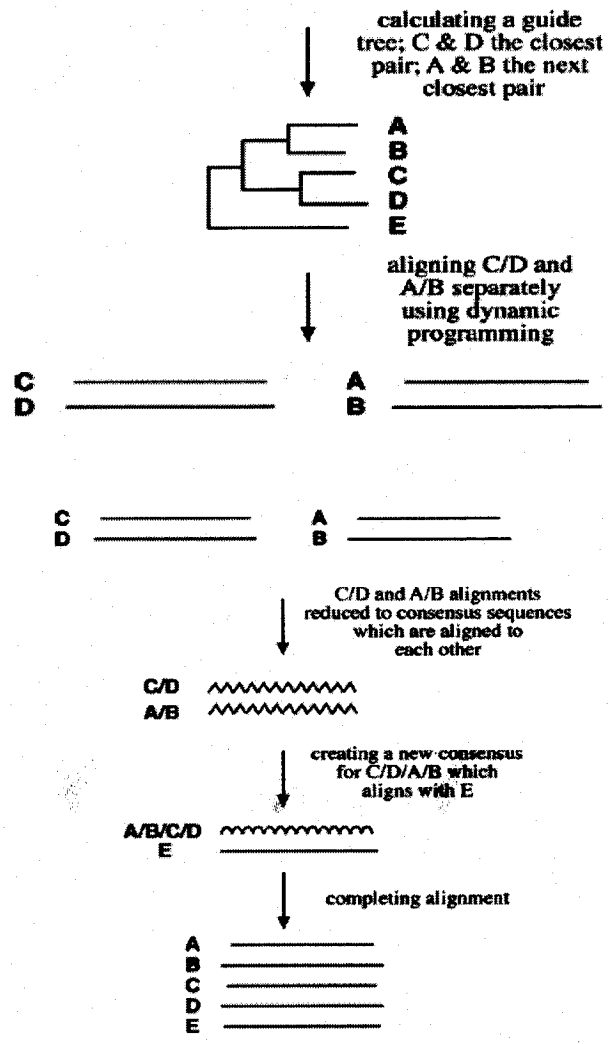
we have 5 different sequences A, B, C, D and E



all individual pairwise alignment and construction of distance matrix

	A	B	C	D	E
A	—				
B	11	—			
C	20	30	—		
D	27	36	9	—	
E	30	33	20	27	—

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh



8.3: Softwares for MSA

Clustal (www.ebi.ac.uk/clustalw/) is a progressive multiple alignment program available either as a stand-alone or on-line program.

CLUSTAL 2 has two versions ClustalW and ClustalX. The W version provides a simple text-based interface and the X version provides amore user-friendly graphical interface.

28/12/04

REGISTRAR
University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8.3.1: CLUSTAL W

In Clustal W we can select the type of the output format

Multiple Sequence Alignment by CLUSTALW

ETE3 | MAFFT | CLUSTALW | PRRN

Help

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: ◉ FAST/APPROXIMATE ○ SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): ◉ PROTEIN ○ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Choose File | No file chosen

Execute Multiple Alignment | Reset

28/12/2021

REGISTRAR

Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1, Window size: 5, Gap Penalty: 3

Number of Top Diagonals: 5, Scoring Method: PERCENT

For SLOW/ACCURATE:

Gap Open Penalty: 10.0, Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN)

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

Multiple Alignment Parameters:

Gap Open Penalty: 10, Gap Extension Penalty: 0.05

Weight Transition: ○ **YES (Value:** 0.5 **),** ◉ **NO**

Hydrophilic Residues for Proteins: GPSNDQERK

Hydrophilic Gaps: ◉ **YES** ○ **NO**

Select Weight Matrix: BLOSUM (for PROTEIN)

Type additional options (delimited by whitespaces) below:

(**-options** for help)

Execute Multiple Alignment | Reset

The type of the pairwise alignment that we need ie FAST and approximate or slow and accurate.

We then choose the type of sequence we want to enter id DNA or protein

CLUSTAL W also provides some options in parameters that users can change as per their requirements like the scores for gap opening and gap extension, the type of matrix to be used.

In case of fast and accurate pairwise alignment, the user can change the range of K-tuple, window size, gap panelty etc.

28/12/2021

REGISTRAR

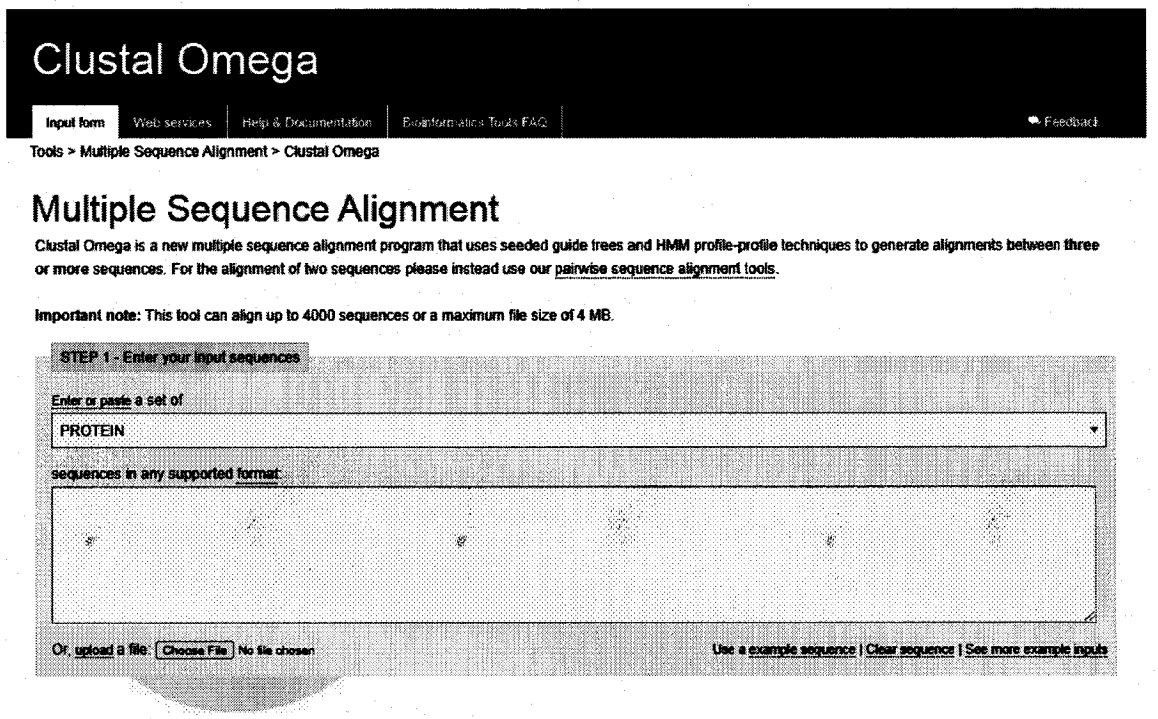
Mody University of Sci...

Lakshmangarh

8.3.2: CLUSTAL Omega

Clustal Omega is a multiple sequence alignment program that uses seeded guide trees and HMM techniques to generate alignments between three or more sequences.

It produces biologically meaningful multiple sequence alignments of divergent sequences.



Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

In the home page we have Sequence Input Window

Select the type of sequences to be aligned ie Protein sequences, DNA sequences or RNA sequences

Three or more sequences to be aligned can be entered directly into this box

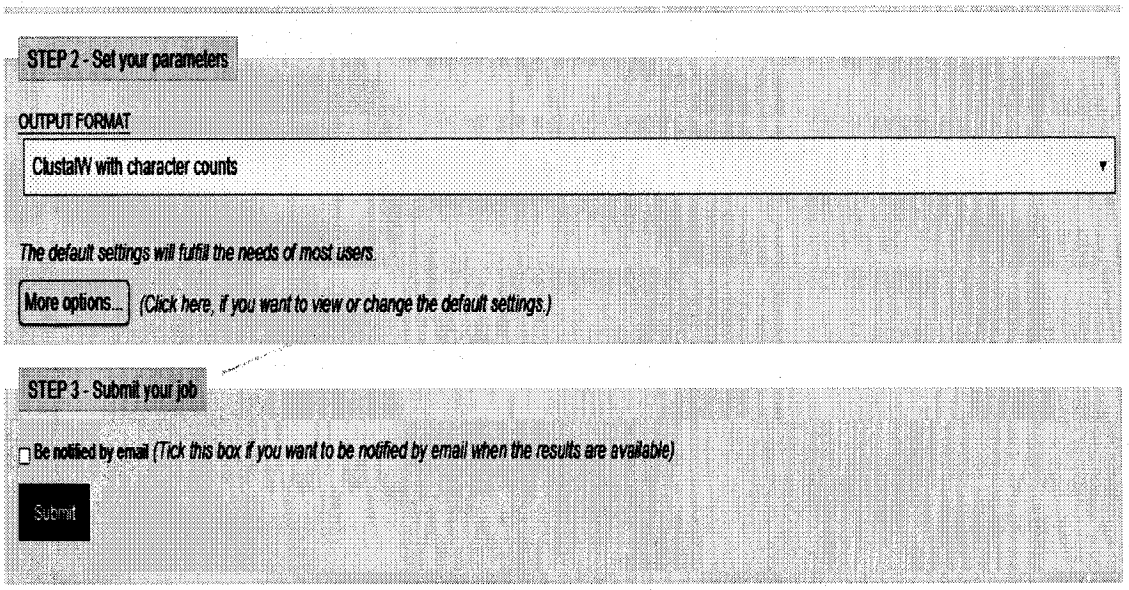
Sequences can be in GCG, FASTA, EMBL (Nucleotide only), GenBank, PIR/NBRF, PHYLIP or UniProtKB/Swiss-Prot (Protein only) format.

One can also upload the sequence files.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- **Step 2 - Set Your Parameters**



There are various parameter options for the users to obtain the alignment most suited for their requirements.

One can also choose the format of the output file generated.

And the last step is submission

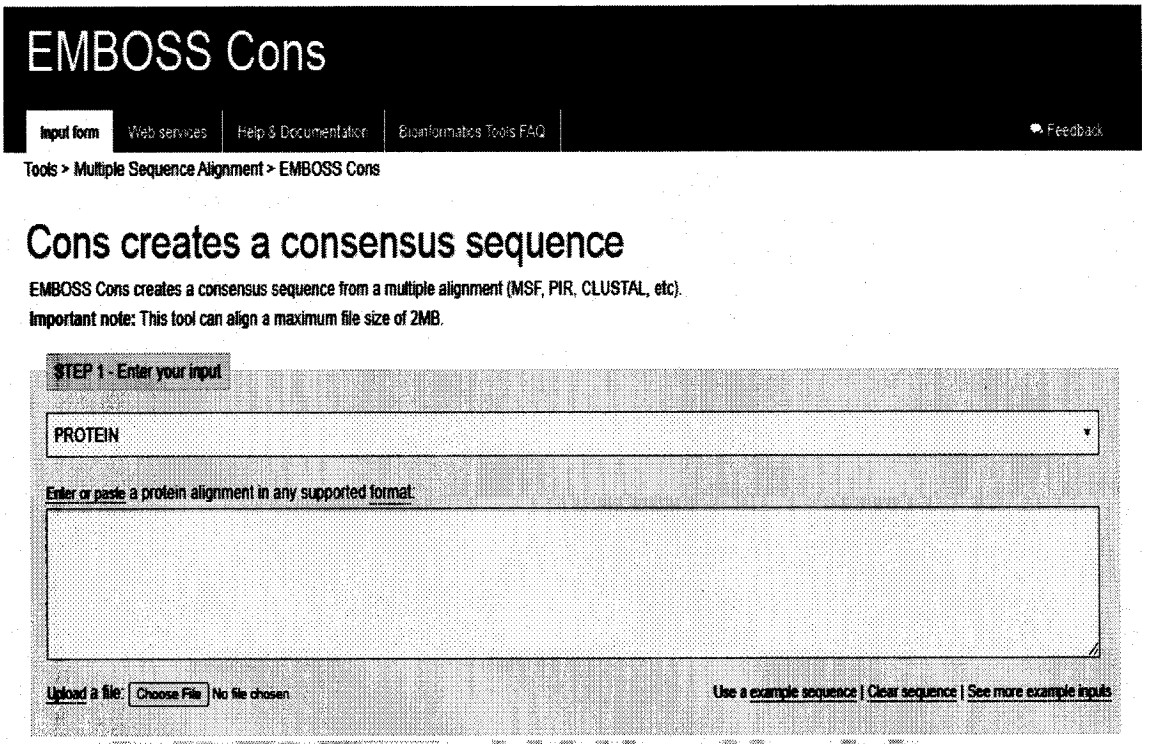
8.3.3: EMBOSS Cons

EMBOSS Cons creates a consensus sequence from a multiple alignment

This tool can align a maximum file size of 2MB.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh



8.3.4: MEGA

It stands of molecular evolutionary genetics analysis

MEGA can be used with either a graphical user interface or a command-line interface which is useful for batch or scripted execution.

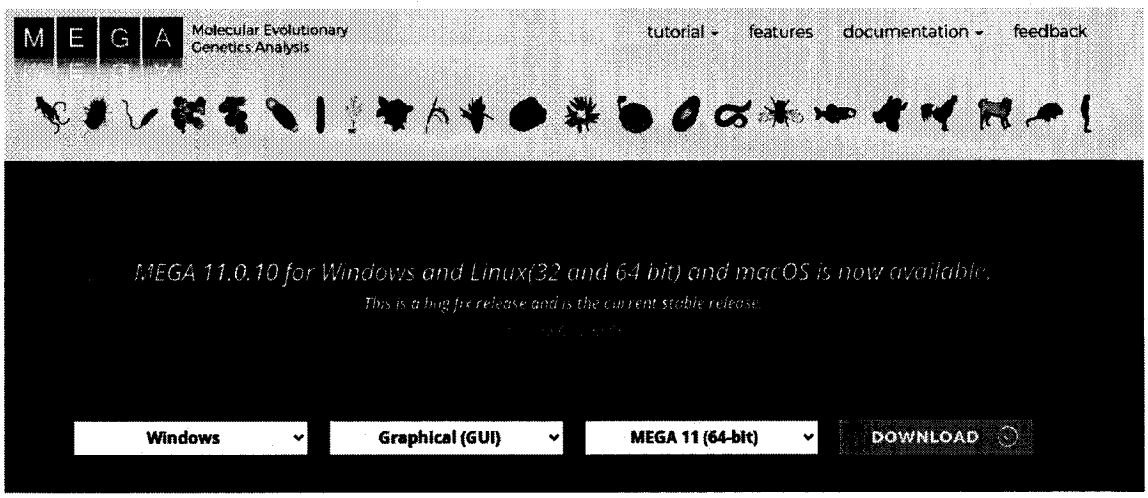
MEGA supports sequence alignment using both the ClustalW and MUSCLE programs.

28/12/2024

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Distt Sikar)



Alignment (or refinement) is done in the Analysis Explorer

Once the sequences are in the Alignment Explorer (AE), we select Alignment either *ClustalW* or Muscle.

Set the alignment parameters to the values you wish or we can go with the default values. Click Compute/OK

Depending on the length and number of sequences you may see a progress bar while the alignment is running.

Once the alignment is done, we can export them to MEGA or Fasta format for analysis.

8.3.5: MAFFT

It is a MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments

Multiple alignment program for amino acid or nucleotide sequences

MAFFT is a multiple sequence alignment program for unix-like operating systems.

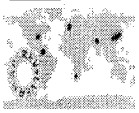
It offers a range of multiple alignment methods, L-INS-i, FFT-NS-2, *etc.*

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences



To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

The Q-INS-i option (for RNA) did not work because of a bug in version 7.489, Oct.23. Fixed in version 7.490, Oct.30, 2021.

A compatibility problem of MSAViewer with Chrome 92 and Firefox 90 has been fixed, Aug. 5, 2021.

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

Multiple sequence alignment and NJ / UPGMA phylogeny

Input

Paste protein or DNA sequences in fasta format. Example

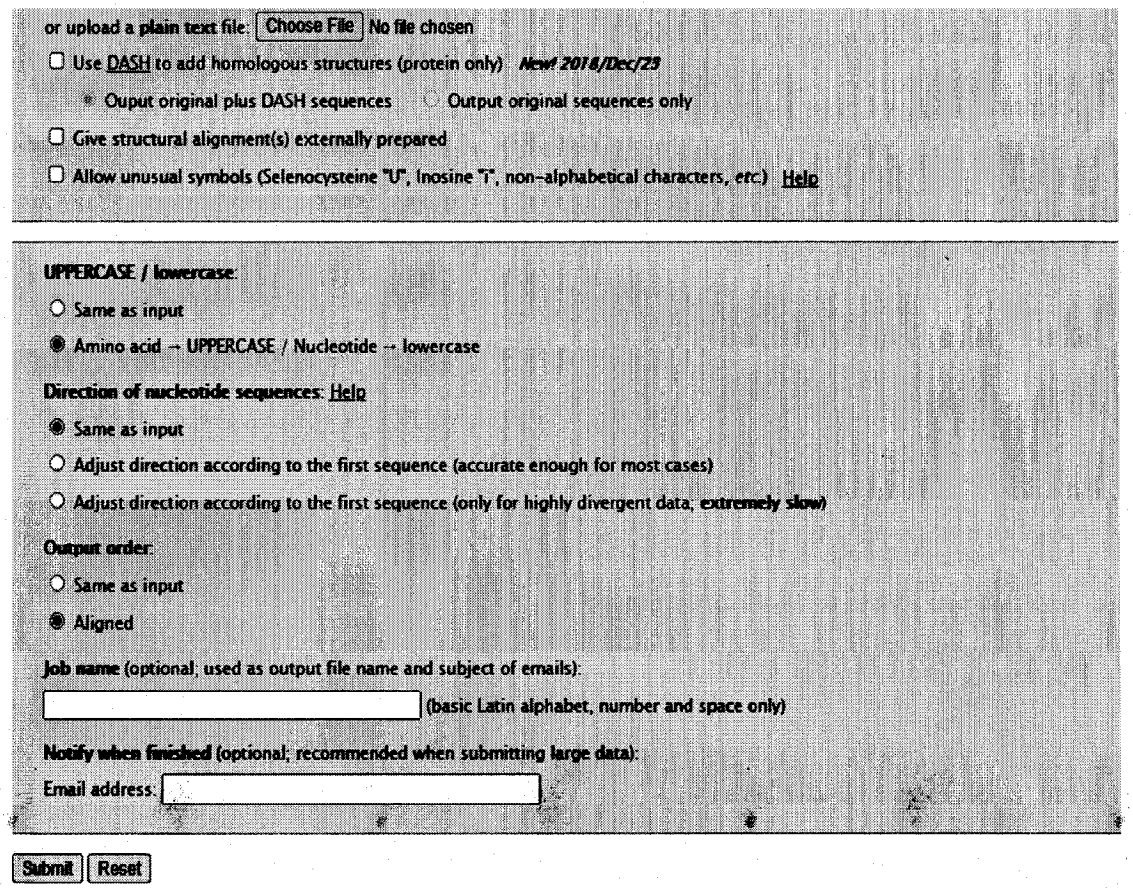
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Distt Sikar) 332311



LAKSHMANGARH • RAJASTHAN

Merits

- (Accuracy): L-INS-i is one of the most accurate multiple sequence alignment methods currently available. L-INS-i is in particular suitable to align 10-100 protein sequences
- (Scalability): FFT-NS-2 and other progressive methods used in MAFFT can align many long DNA/protein sequences
- The scoring system was designed to allow large gaps.

29/12/2024

8.3.6: MUSCLE

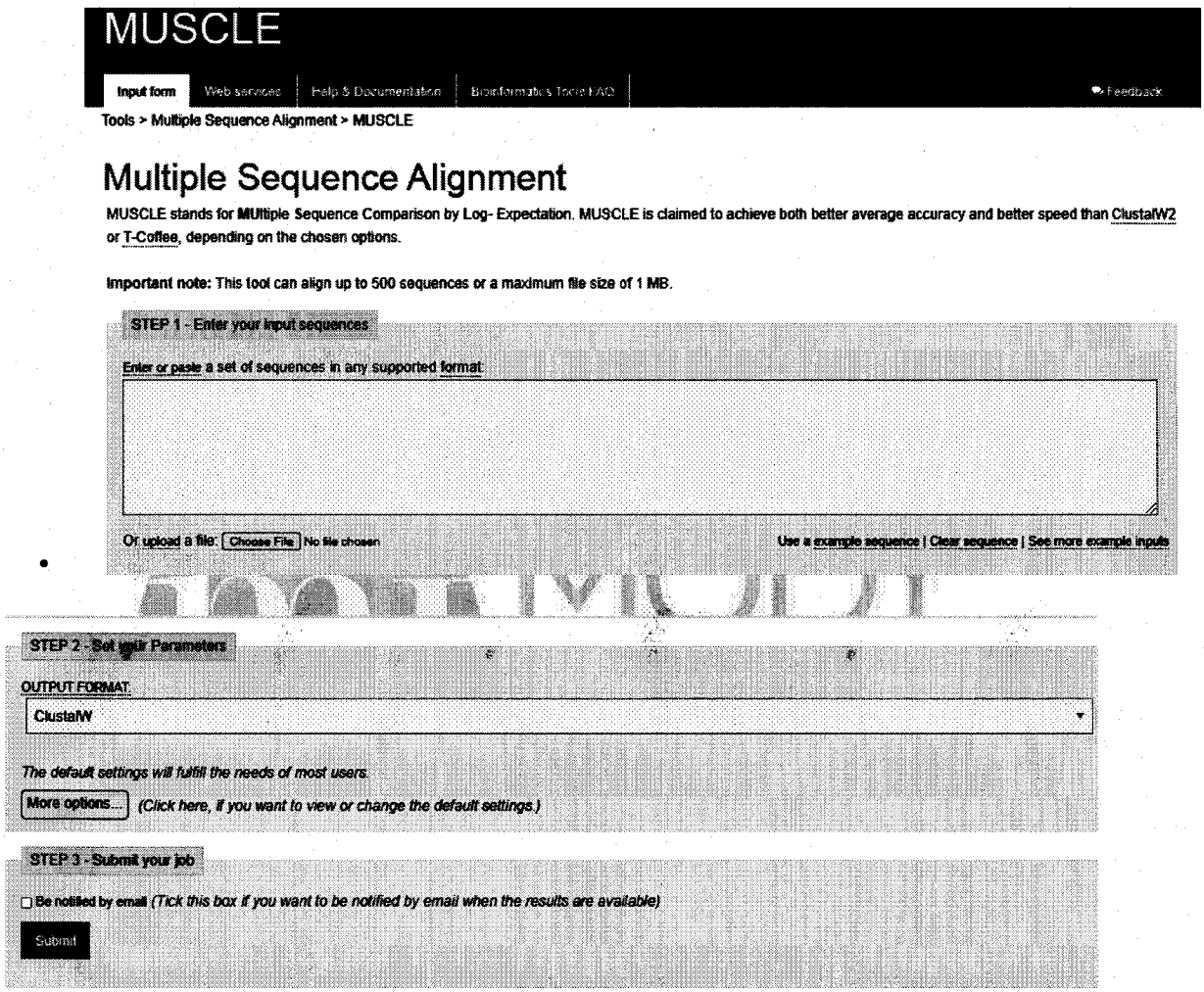
MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation.

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee,

This tool can align up to 500 sequences or a maximum file size of 1 MB in a second.

MUSCLE enables high-throughput applications to achieve average accuracy



28/12/2021

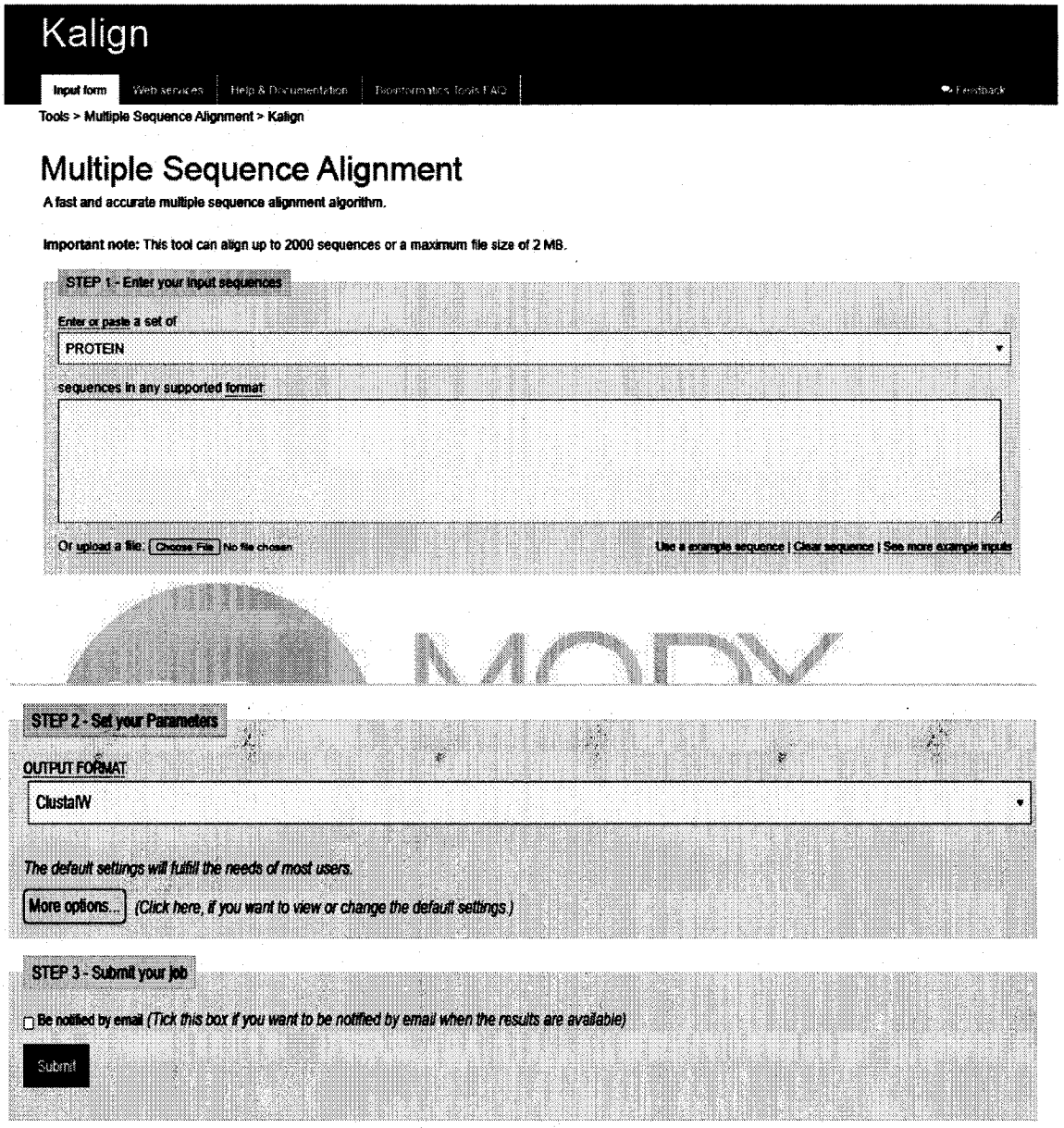
REGISTRAR
Mody University of Science and Technology
(Lakshmangarh) (Rajasthan)

8.3.7: KALIGN

Very fast MSA tool that concentrates on local regions..

It is a fast and robust alignment method, especially well suited for the increasingly important task of aligning large numbers of sequences.

Suitable for large alignments



20/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Important links

https://mafft.cbrc.jp/alignment/server/

https://www.ebi.ac.uk/Tools/msa/emboss_cons/

https://www.ebi.ac.uk/Tools/msa/clustalo/

https://www.genome.jp/tools-bin/clustalw

https://www.megasoftware.net/

https://www.ebi.ac.uk/Tools/msa/kalign/

https://www.ebi.ac.uk/Tools/msa/muscle/

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

9: Basic local alignment search tool (BLAST)

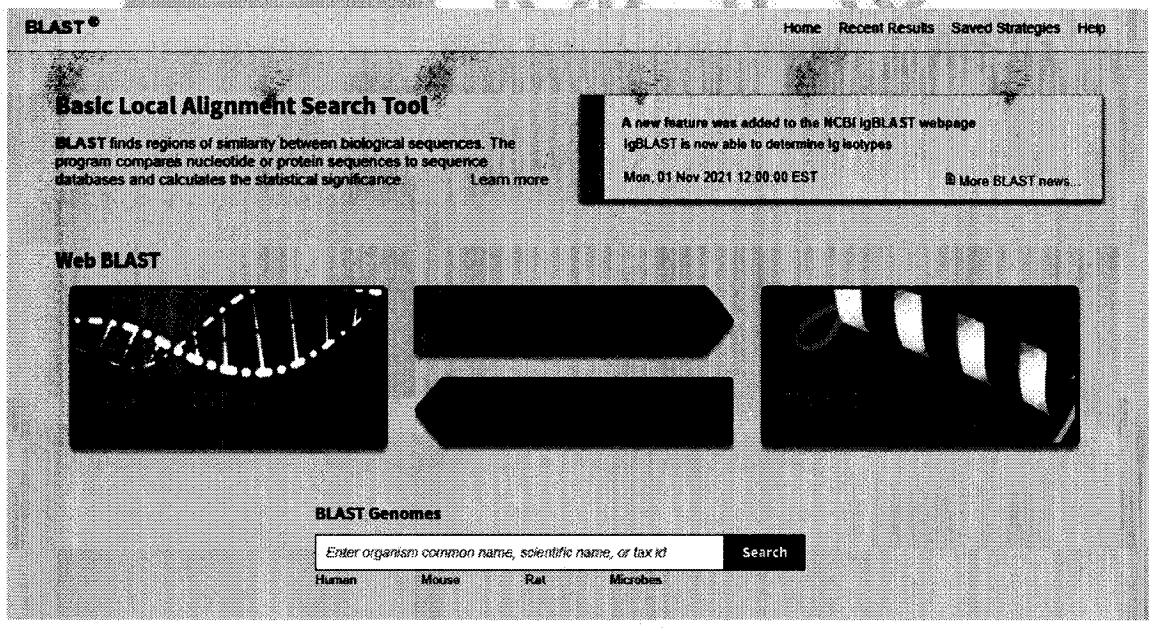
BLAST stands for BASIC LOCAL ALIGNMENT SEARCH TOOL.

It helps in finding the similarities between biological sequences. It is a complete package which helps the researchers in various aspects for comparing and identifying DNA, RNA and protein sequences. BLAST suite offers multiple services like:

Web BLAST offers platforms for searching the protein and nucleotide sequences against a protein or nucleotide database.

BLAST also provides **command line applications** to perform searches locally.

In BLAST suite, one can also find specialized services like SMART BLAST, PRIMER BLAST, GLOBAL ALIGN, CD SEARCH, Ig BLAST, VECSCREEN, CDART, MULTIPLE ALIGN and MOLE BLAST.



28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Specialized searches

Find proteins highly similar to your query

Design primers specific to your PCR template

Compare two sequences across their entire span (Needleman-Wunsch)

Find conserved domains in your sequence

Search immunoglobulins and T cell receptor sequences

Search sequences for vector contamination

Find sequences with similar conserved domain architecture

Align sequences using domain and protein constraints

Establish taxonomy for uncultured or environmental sequences

UNIVERSITY
LAKSHMANGARH • RAJASTHAN

SMART BLAST

It is a new tool of NCBI that is used for sequence analysis tasks like finding a candidate protein name for a sequence, locating regions of high sequence conservation, or identifying regions covered by database sequences but missing from the query.

To perform all such tasks, it takes much less time as compared to the protein blast.

PRIMER BLAST

It helps in designing the primers that are specific for the PCR template of our choice.

GLOBAL ALIGN

This helps in aligning two sequences globally i.e end to end alignment is done.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

This is done using Needleman-Wunsch algorithm which we have already discussed in our previous lecture.

CD SEARCH

CD search is Conserved Domain Search service

It is a web-based tool for the detection of structural and functional domains in protein sequences.

Ig BLAST

Is A tool for immunoglobulin (IG) and T cell receptor (TR) Variable domain sequences

With this tool, we can view the matches to the germline V, D and J genes, we can find details at rearrangement junctions, we can also find the delineation of immunoglobulin Variable domain framework regions and complementarity determining regions.

VECSCREEN

VecScreen is a system that quickly finds segments of a nucleic acid sequence that may be of vector origin.

It is recommended to researchers as its helps in identifying and removing any segments of vector origin before they analyze or submit sequences

CDART

Finds sequences with similar conserved domain architecture

MULTIPLE ALIGN

It is a constraint based multiple alignment tool.

It finds a collection of pairwise constraints derived from conserved domain database, protein motif database, and sequence similarity, using RPS-BLAST, BLASTP, and PHI-BLAST

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

MOLE BLAST

MOLE-BLAST is an experimental tool that helps taxonomists find closest database neighbors of submitted query sequences

It computes a multiple sequence alignment (MSA) between the query sequences along with their top BLAST database hits, and generates a phylogenetic tree.

The BLAST suite also offers downloadable link for BLAST through which one can get a standalone version of BLAST along with the databases.

WEB BLAST

WEB BLAST is the most commonly used section of BLAST which offers multiple versions to compare all combinations of nucleotide or protein queries against a nucleotide or protein database.

The versions are namely nucleotide BLAST or blastn, Protein BLAST or blastp, translated BLASTs blastx, tblastn and tblastx

Nucleotide BLAST or blastn

BLASTN program searches nucleotide databases using a nucleotide query.

It is a general purpose nucleotide search and alignment program that can be used to align tRNA or rRNA sequences as well as mRNA or genomic DNA sequences containing a mix of coding and noncoding regions.

Protein BLAST or blastp

Protein-to-protein sequence searches are performed using blastp or protein blast.

The default wordsize for a blastp search is three.

During a "blastp" search, low-complexity regions of the query sequence are filtered to reduce the construction of spurious alignments and enhance search speed.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Translated BLAST

Translated BLAST searches use a genetic code to translate either the "query," database "subjects," or both, into protein sequences, which are then aligned as in "blastp.

There are three varieties of translated BLAST search;

- "tblastn,"
- "blastx," and
- "tblastx."

Blastx

In this type of translated BLAST, a nucleotide sequence query is translated in six reading frames, and the resulting six-protein sequences are compared, in turn, to those in a protein sequence database.

Tblastn

In this a protein sequence query is compared to the six-frame translations of the sequences in a nucleotide database.

As we are searching translated sequences in nucleotide database, we call it T blast N

Tblastx

In this BOTH, the "query" and database "subject" nucleotide sequences are translated in six reading frames, after which 36 protein "blastp" comparisons are made.

Six-frame translations for query sequence and six-frame translations for subject nucleotide sequence

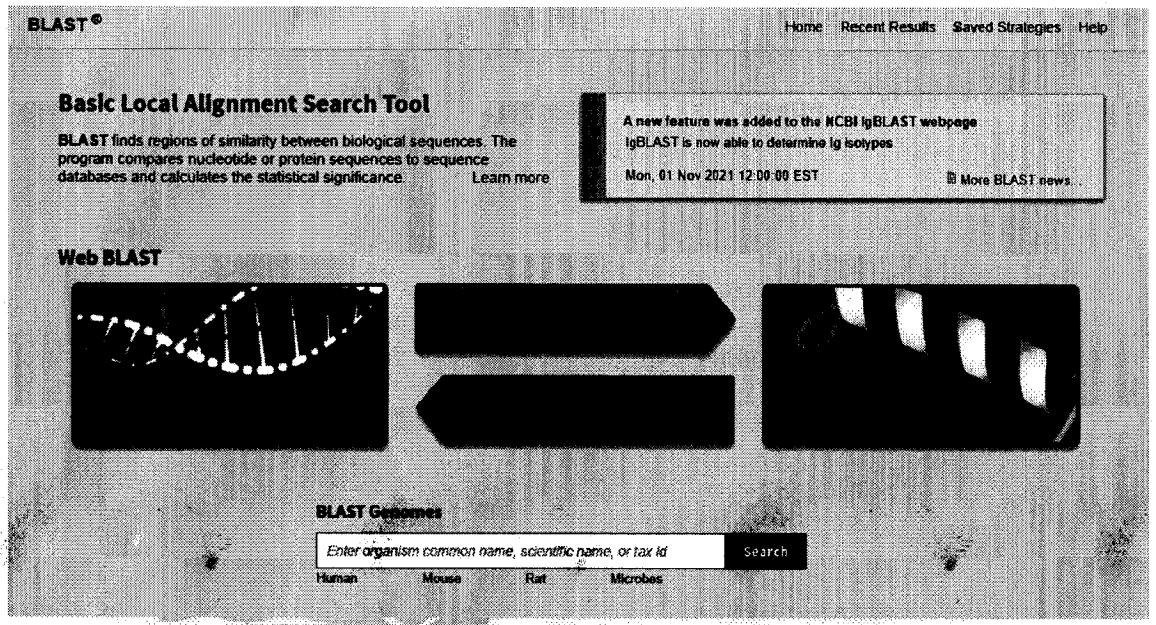
In web blast, we also find BLAST GENOMES which is a graphical tool for simplifying BLAST searches against complete and unfinished genome sequences.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2017
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

This tool is used to compare the query sequence against a virtual database of DNA sequences or protein sequences from a selected group of organisms with finished or unfinished genomes.

9.1: BLAST software



Recent Results which lists a set of links to recently completed search results.

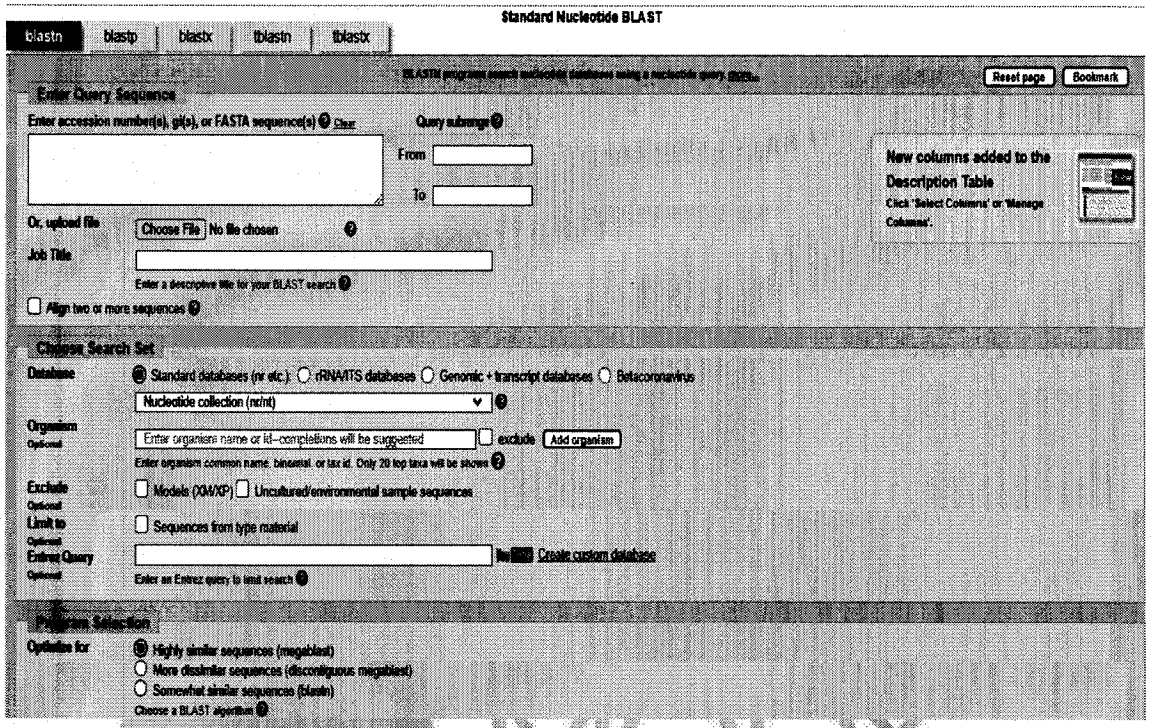
We have "**saved strategies**" which displays a list of search strategies which we have saved earlier. On home page, we can find links to various types of services provided in the BLAST suite.

9.1.1: Nucleotide BLAST

On the homepage of blast, click on nucleotide blast

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



Here on the first section, we provide the query sequence.

We can provide query sequence, either by entering the accession number(s) or by providing the gi number(s), or by providing the FASTA sequence(s).

We can also upload a file from our local disk. The file may contain a single sequence or a list of sequences and the data can be either a list of database accession numbers, NCBI gi numbers, or sequences in FASTA format.

Now here Enter a descriptive title for your BLAST search

We can also align two or more sequences. For this we can Enter one or more queries in the top text box and one or more subject sequences in the lower text box and click on the BLAST button at the bottom.

Now the next segment is to **Choose Search Set**

From here we can choose the database against which we want to search our query sequence.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

20/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

We can search against standard databases or rRNA/ITS databases or genomic+ transcript databases and now also against beta corona virus genbank .

Then we can choose the list of organisms that we want to exclude from our search.

This is optional. Like wise we have few more options available to restrict our search like we can exclude the model sequences or uncultured experimental sequences.

Now next we can choose the type of the program we want to run. Whether we want to run megablast or discontinuous megablast or blastn. By choosing the program, we are selecting the algorithm we want to run for our similarity search. Before hitting the BLAST button lets have a look at the additional parameters. We can leave these parameters with their default values.

So here, in parameter section, we can choose the maximum no.of target sequences, the default value of which is 100

Short queries. If the entered sequence is a short query sequence then the algorithm will automatically adjust word size and other parameters to improve results for short queries.

Expected threshold, the default value of which is 0.05. This setting specifies the statistical significance threshold for reporting matches against database sequences

In the same way we have other parameters like word size, match mismatch score, gap cost and we have certain filters and masks.

Now once we have adjusted all the parameters and settings as per our requirement, we will hit the BLAST button

Now we are moving towards the results page. BLAST takes few seconds to generate the results page

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

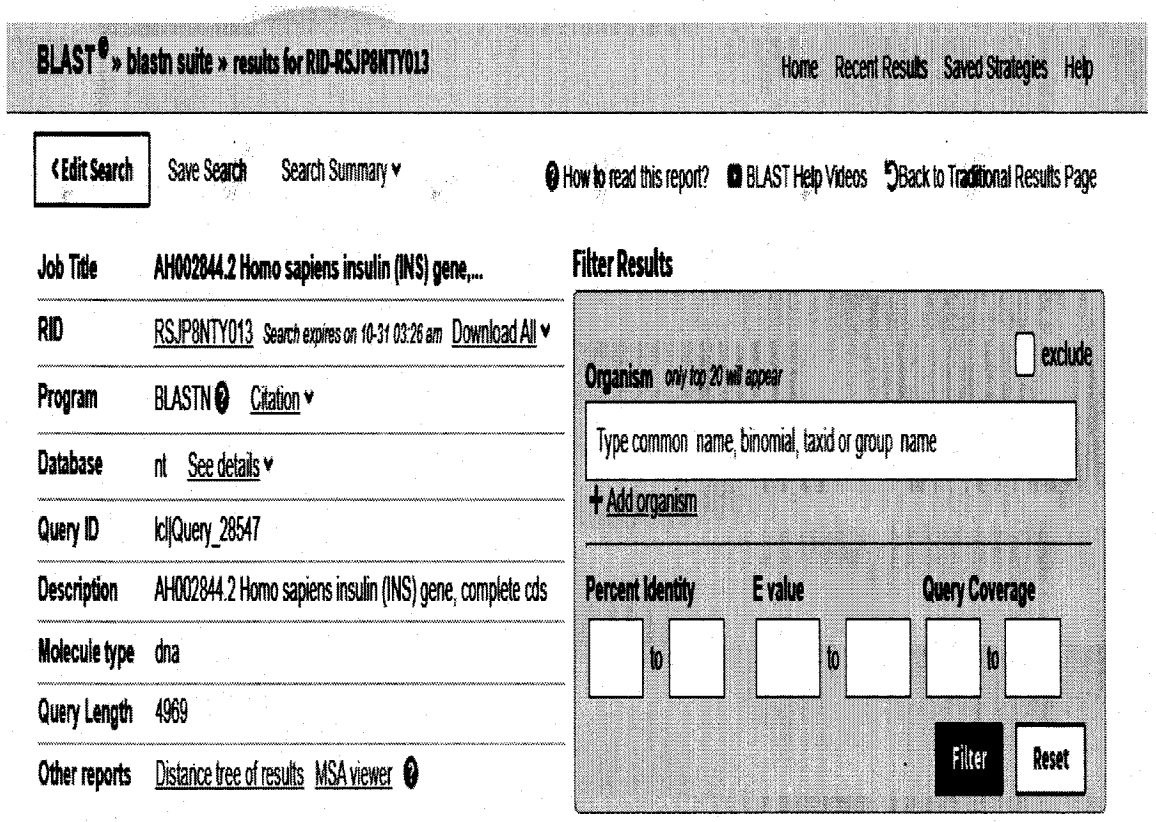
Formatting options]

Job Title: AH002844.2 Homo sapiens insulin (INS) gene,..

Request ID	RSJP8NTY013
Status	Searching
Submitted at	Fri Oct 29 15:26:00 2021
Current time	Fri Oct 29 15:26:15 2021
Time since submission	00:00:14

This page will be automatically updated in 2 seconds

The results page of BLAST



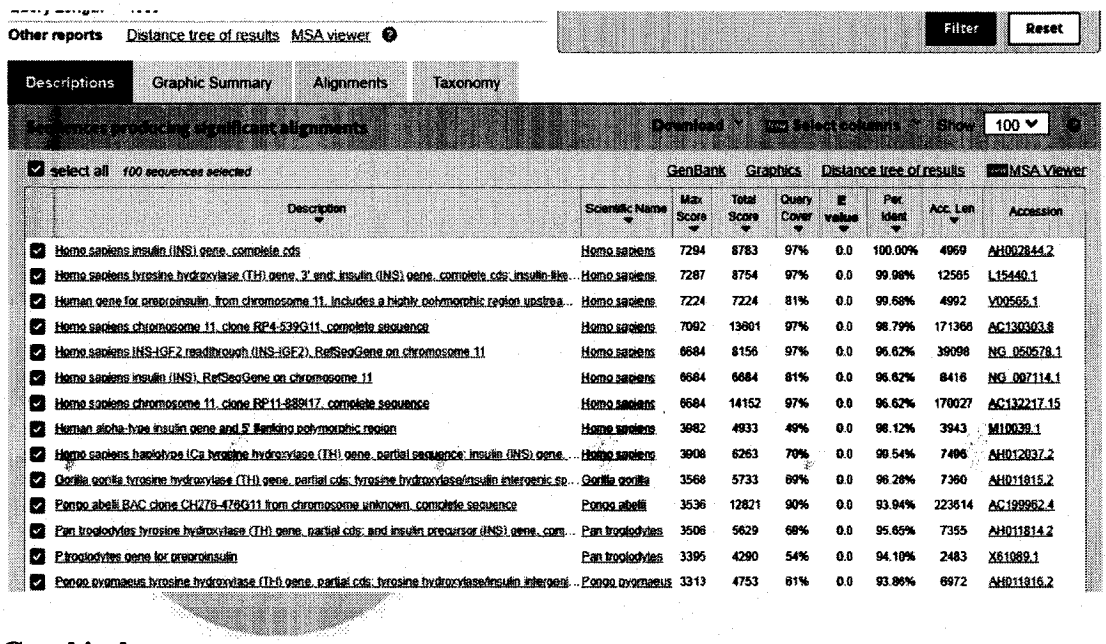
Job Title shows the first query's sequence id by default

RID shows the assigned unique identifier to this search for sharing and problem reporting

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Program tells the search conducted. Database names the target database searched. Query ID is the query's sequence id for which the result is shown. Description is the title of the query sequence from its FASTA defline. Molecule type shows the type of the sequence. Query length is the length of the current query. From here we can also refine our results. The Descriptions tab contains a summary table of hits found by BLAST.

Default sorting is by E value column



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc. Len	Accession
Homo sapiens insulin (INS) gene, complete cds	Homo sapiens	7294	8783	97%	0.0	100.00%	4969	AH002844.2
Homo sapiens tyrosine hydroxylase (TH) gene, 3' end; insulin (INS) gene, complete cds; insulin-like...	Homo sapiens	7287	8754	97%	0.0	99.98%	12565	L15440.1
Human gene for preproinsulin, from chromosome 11, includes a highly polymorphic region upstrea...	Homo sapiens	7224	7224	81%	0.0	99.68%	4992	V00565.1
Homo sapiens chromosome 11, clone RP4-539G11, complete sequence	Homo sapiens	7092	13801	97%	0.0	98.79%	171366	AC130303.8
Homo sapiens INS-IGF2 readthrough (INS-IGF2), RefSeqGene on chromosome 11	Homo sapiens	6684	8156	97%	0.0	96.62%	39098	NG_050578.1
Homo sapiens insulin (INS), RefSeqGene on chromosome 11	Homo sapiens	6684	6684	81%	0.0	96.62%	8416	NG_007114.1
Homo sapiens chromosome 11, clone RP11-889I17, complete sequence	Homo sapiens	6684	14152	97%	0.0	96.62%	170027	AC132217.15
Human alpha-type insulin gene and 5' flanking polymorphic region	Homo sapiens	3982	4933	49%	0.0	98.12%	3943	M10039.1
Homo sapiens haplotype (Ca tyrosine hydroxylase (TH) gene, partial sequence; insulin (INS) gene,...	Homo sapiens	3908	6263	70%	0.0	99.54%	7496	AH012037.2
Gorilla gorilla tyrosine hydroxylase (TH) gene, partial cds; tyrosine hydroxylase/insulin intergenic sp...	Gorilla gorilla	3568	5733	69%	0.0	98.28%	7360	AH011815.2
Pongo abelii BAC clone CH276-476G11 from chromosome unknown, complete sequence	Pongo abelii	3536	12821	90%	0.0	93.94%	223614	AC199962.4
Pan troglodytes tyrosine hydroxylase (TH) gene, partial cds; and insulin precursor (INS) gene, com...	Pan troglodytes	3508	5629	68%	0.0	95.65%	7355	AH011814.2
Ptroglodytes gene for preproinsulin	Pan troglodytes	3395	4290	54%	0.0	94.10%	2483	X61089.1
Pongo pygmaeus tyrosine hydroxylase (TH) gene, partial cds; tyrosine hydroxylase/insulin intergeni...	Pongo pygmaeus	3313	4753	61%	0.0	93.86%	6972	AH011816.2

Graphical summary:

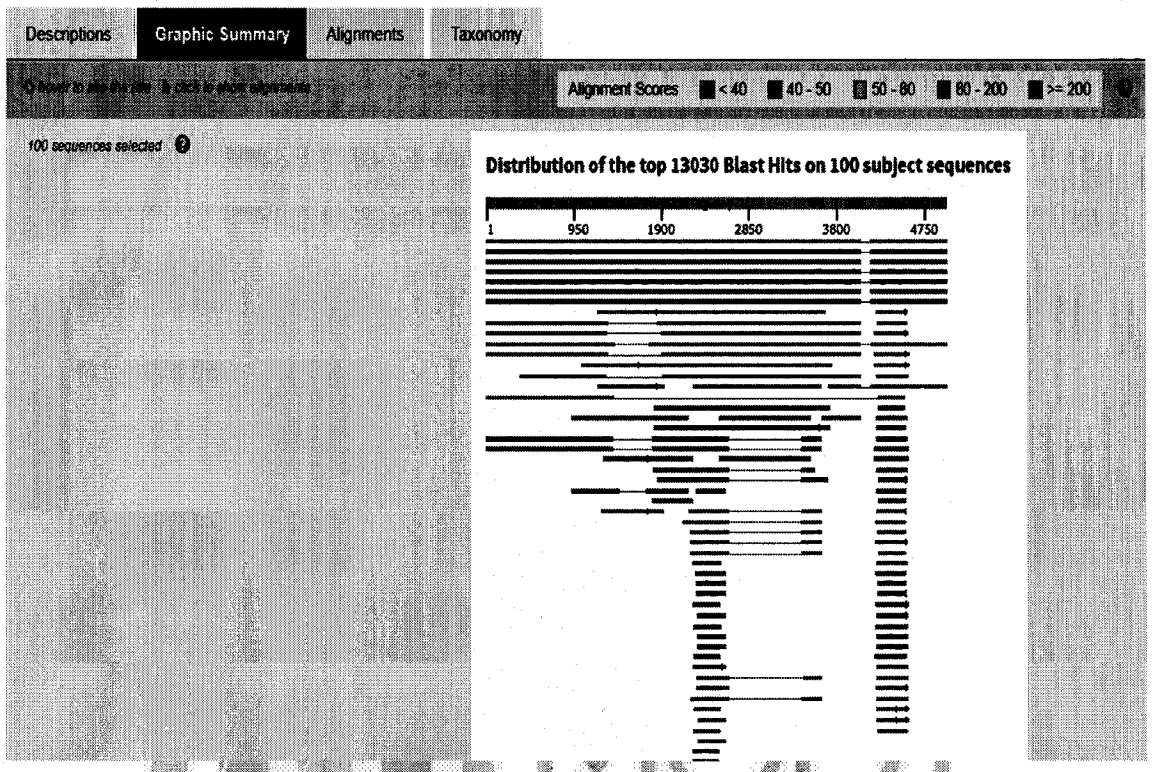
Here you can see various lines in different colours. These colours are on the basis of the alignment scores.

red represents the highest alignment scores

Generally, the higher the alignment score, the more significant the hit.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)



The Alignments section contains the detailed pairwise alignments between query and database sequences.

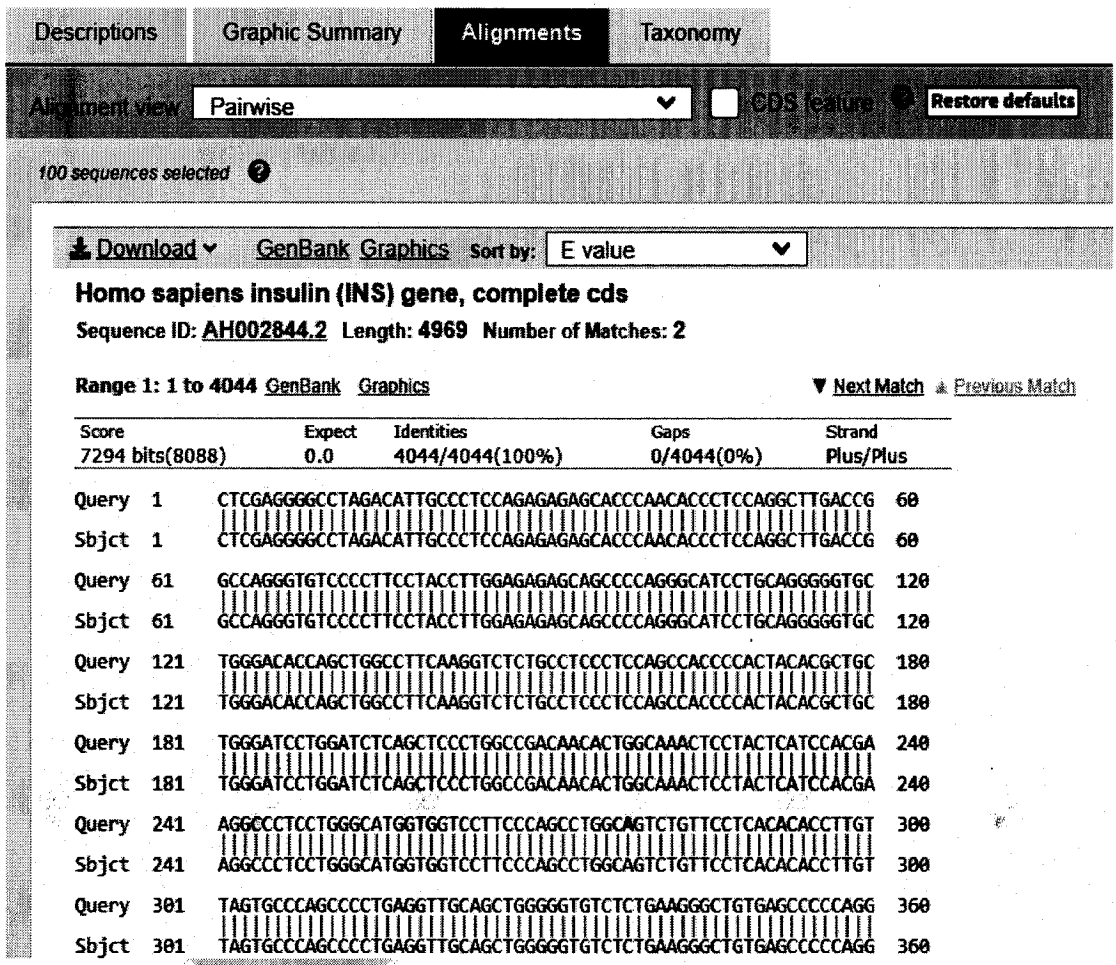
It makes alternative displays of the alignment results readily accessible through a simple pull-down menu.

The sequence alignments show us how well our query sequence matches with the subject sequence in the database.

Each alignment block begins with a summary that includes the Max score and Expect value (the statistical significance of the alignment), sequence identity (number of identical bases between the query and the subject sequence), the number of gaps in the alignment, and the orientation of the query sequence relative to the subject sequence

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

Download ▾ GenBank Graphics

Pongo pygmaeus insulin gene, partial cds

Sequence ID: AY092024.1 **Length:** 972 **Number of Matches:** 1

Range 1: 1 to 972 GenBank Graphics ▼ Next Match ▲ Previous Matc

Score	Expect	Identities	Gaps	Strand
1481 bits(1642)	0.0	922/989(93%)	24/989(2%)	Plus/Plus

```
Query  2513  GTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTA  2572
             ||||||||| ||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GTGCGGCTCCCACCTGGTGGAAGCTCTCTACCTAGTGTGTGGGGAACGAGGCTTCTTCTA  60

Query  2573  CACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCATTGCTG  2632
              ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct  61    CACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCCTTGCTG  120

Query  2633  CCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTCCCACCCAGCATGGGCAGAAGG  2692
              ||||||| |||||||||||| ||||||| ||| ||||||||||||||||||||||||||
Sbjct  121   CCCCTGGTCGCCCCCAGCCATCCCCTGCTTCTGCCGCTCCCACCCAGCATGGGCAGAAGG  180

Query  2693  GGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAA-----GTT  2747
              ||||||||||||||||||||||||||||||||||||||||||||||||||      |||
Sbjct  181   GGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGTACTTTTTTAAAAAGAAATGAAGTT  240

Query  2748  CTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGA  2807
              |||||||||| |||||||||||||| |||||||||||| |||||||
Sbjct  241   CTCTTGGTCACATCCTAAAAGTGACCAGTTCCCTGTGGCCCATTCAGAATG---------  291

Query  2808  GGACGGTGTTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCC  2867
              ||   |||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  292   ----GG----GGCTTCAGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCC  343

Query  2868  ACTCGCCCCTCAAACAAATGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAG  2927
              |||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||
Sbjct  344   ACTCGCCCCTCAAACAAATGCCCCACAGCCCATTTCTCCACCCTCATTTGATGACCGCAG  403

Query  2928  ATTCAAGTGTTTTGTTAAGTA-AAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGC  2986
              |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||
Sbjct  404   ATTCAAGTGTTTTGTTAAGTACAAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGC  463

Query  2987  TGCCTGCCTCTGGGCGAACACCCCATCACGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAG  3046
              |||||||||||||||||||||||||||||||| |||| ||||||||||||
Sbjct  464   TGCCTGCCTCTGGGCGAACGCCCCATCACGCCCTGAGGAGGACATGGCCGCCTGCCTGAG  523
```

28/12/2024

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Dist Sikar) 332311

Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	D
root			106	
. Simiiformes	primates		99	
. . Catarrhini	primates		98	
. . . Hominidae	primates		86	
. . . . Homininae	primates		82	
. . . . . Homo sapiens	primates	7294	75	Homo sapiens hits
. . . . . Gorilla gorilla	primates	3568	2	Gorilla gorilla hits
. . . . . Pan troglodytes	primates	3506	4	Pan troglodytes hits
. . . . . Pan paniscus	primates	1324	1	Pan paniscus hits
. . . . Pongo abelii	primates	3536	2	Pongo abelii hits
. . . . Pongo pygmaeus	primates	3313	2	Pongo pygmaeus hits
. . . Chlorocebus aethiops	primates	2338	1	Chlorocebus aethiops hits
. . . Chlorocebus sabaeus	primates	1760	3	Chlorocebus sabaeus hits
. . . Papio anubis	primates	618	1	Papio anubis hits
. . . Theropithecus gelada	primates	614	3	Theropithecus gelada hits
. . . Piliocolobus tephrosceles	primates	589	1	Piliocolobus tephrosceles hits
. . . Rhinopithecus bieti	primates	532	1	Rhinopithecus bieti hits
. . . Macaca mulatta	primates	426	2	Macaca mulatta hits
. . Aotus trivirgatus	primates	1763	1	Aotus trivirgatus hits
. Expression vector pSP20	other sequences	1299	1	Expression vector pSP20 hits
. eukaryotic synthetic construct	other sequences	431	6	eukaryotic synthetic construct hits

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

References

- https://ncbiinsights.ncbi.nlm.nih.gov/2013/03/19/blastdbinfo-api-access-to-a-database-of-blast-databases/
- Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes, FEMS Microbiol Lett . 2002 Nov 5;216(2):133-8. doi: 10.1111/j.1574-6968.2002.tb11426.x.
- https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

28/11/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Sikar) 332311

e-PG Paathshaala

E-Content

Basic Bioinformatics

Dr. Anukriti and Dr. Naveen Dhingra

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Contents

i. Objectives

ii. Learning outcome

Chapter 1: Introduction to Bioinformatics

Chapter 2: Introduction to Databases

Chapter 3: Sequence databases

Chapter 4: Structure database

Chapter 5: Structure databases (Part II)

Chapter 6: Sequence alignment (part I)

Chapter 7: Sequence alignment (part II)

Chapter 8: Multiple Sequence Alignment

Chapter 9: Basic Local Alignment Search Tool (BLAST)

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/24

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

i: Objectives

1) To understand about the beginning of Bioinformatics

2) Develop basic understanding of databases

3) Understand about sequence databases

4) Create a basic understanding of structure databases

5) Basic understanding of Pairwise and multiple sequence alignment

6) Hands-on with BLAST software

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

ii: Learning Outcome:

1) A brief introduction about Bioinformatics

2) Be able to learn about various databases like structure databases, sequence databases

3) How to submit data in various databases

4) Algorithms in Pairwise sequence alignment like Dot plot

5) Algorithm for Multiple sequence alignment

6) Basics of BLAST with hands on experiment using BLAST software

28/12/2024

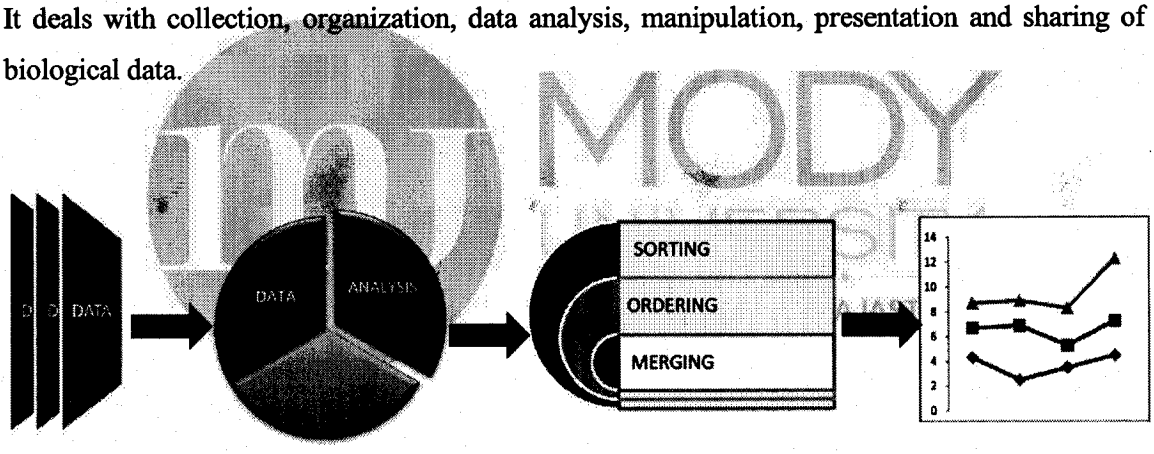
REGISTRAR
Mody University of Science and Technology
Lakshmangarh

1: Introduction to Bioinformatics

1.1: What Is Bioinformatics?

Bioinformatics is an interdisciplinary field harnessing fields like computer sciences, mathematics, physics and biology. It is convergence of biotechnology and information technology. It is a stream of science which arose for the utilitarian purpose of introducing order into the massive data sets produced by the help of new and advanced technologies of molecular biology.

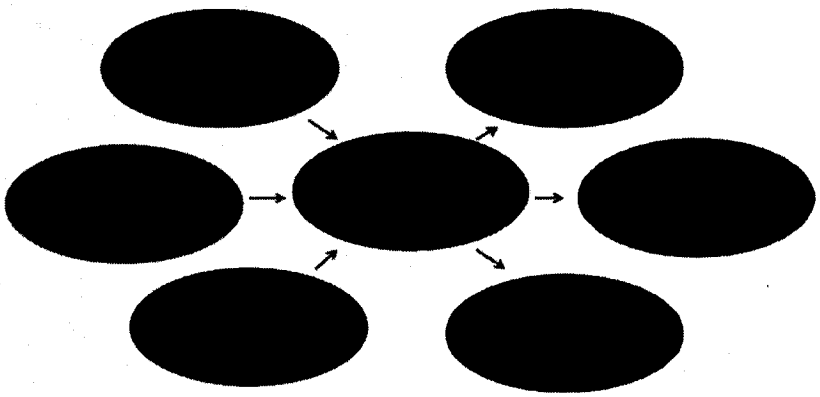
Higgs and Attwood gave two definitions of bioinformatics. "Bioinformatics is the development of computational methods for studying the structure, function, and evolution of genes, proteins and whole genomes" and "bioinformatics is the development of methods for the management and analysis of biological information arising from genomics and high throughput experiments". It deals with collection, organization, data analysis, manipulation, presentation and sharing of biological data.



It requires advanced knowledge of computer science, mathematics and statistical methods to understand various biological functions at the molecular level. It covers specialized areas of biology like functional and structural genomics, DNA microarray, comparative genomics etc. It is a cutting-edge branch of biotechnology that paves its way to novel drug discovery and personalized medical treatments.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



1.2: Origin of Bioinformatics

1960s had witnessed the first efforts towards bioinformatics. Paulien Hogeweg and Ben Hesper first coined the term bioinformatics in the beginning of 1970s, defining it as "the study of informatic processes in biotic systems". Margaret Dayhoff in 1965 developed first protein sequence database called Atlas of Protein Sequence and Structure.

The Brookhaven National Laboratory in early 1970s, established the Protein Data Bank (PDB) for archiving 3D protein structures. Needleman and Wunsch developed the first sequence alignment algorithm in 1970s. Chou and Fasman in 1974 developed the first protein structure prediction algorithm.

In 1980s, GENBANK got established and William Pearson developed a fast database searching algorithm FASTA and Stephen Altschul and coworkers developed BLAST. A major developmental boost to bioinformatics came in late 1980s by the start of Human Genome Project.

1.3: Why We Should Study Bioinformatics?

- For better understanding of living cells at molecular levels.

  All the living cells are composed of various molecules that function in perfect coordination with each other for the proper functioning of the cell. With the help of

20/12/2024
REGISTRAR
Modv University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

bioinformatics, we can try to understand a cell as a system that is composed of various molecules working together to perform any function within the cell.

- To get better understanding of evolution.

  Phylogenetic analyses are the processes that are used for the reconstruction of evolutionary relationship among a group of related molecules or organisms. This helps to predict certain features of molecules whose functions are not known, to track the flow of the genetic information, and also to determine genetic relatedness. With the help of all these, one can try to understand the process of evolution.

- For gene predictions and genome annotation.

  Gene prediction, also known as gene finding basically means locating the genes along a genome by identifying the regions of genomic DNA that encode genes using various tools offered by bioinformatics.

  Genome annotation process helps in the identification of functional elements along the sequence of a genome. It is a process of analyzing and interpreting the raw DNA sequences so that significant biological information can be extracted out from it with the use of certain tools and softwares.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

- To study the impact of change in climate on various organisms.

  With the help of bioinformatics, we can try to predict the changes in the climatic conditions and when we combine bioinformatics with genomics, we can try to predict the probable changes in genetic makeup of various organisms with the change in the climatic conditions.

- For the development of genome libraries and databases.

  Bioinformatics plays an important role in managing and analyzing the huge biological data. It helps in the development of databases and libraries from which biological information can be extracted out easily.

- For the development of molecular and personalized medicines and vaccines.

29/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

There are thousands of diseases that are hereditary in nature. Every individual also responds differently towards the environmental changes. Thus, trying to pin point the exact gene or molecule that is involved in the occurrence of any disease can help in designing molecular and personalized medicines. This can be achieved with the help of bioinformatics.

1.4: Uses of Bioinformatics

- Analysis of gene variation and expression.
- Analysis and prediction of gene function.
- Designing primers to predicting the function of gene products.
- Analysis and prediction of protein structure and function.
- Prediction and detection of gene regulation networks.
- Simulation of environments for whole cell modeling.
- With the help of sophisticated computational tools, mining the genomic information.
- Knowledge based drug designing.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

1.5: Scope of Bioinformatics

- Development of computational tools and databases.
- Development of algorithms.
- Data analysis and interpretation.
- In structural analysis of proteins and nucleic acids.
- Managing big data.
- Drug discovery and vaccine designing.
- In functional analysis of proteins.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References:

1) Ardeshir Bayat, Science, medicine, and the future- Bioinformatics, BMJ. 2002 Apr 27; 324(7344): 1018–1022.

2) Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press

3) Lesk, A. M. (2002). Introduction to bioinformatics. Oxford: Oxford University Press.

4) Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer

5) Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol 7(3): e1002021. doi:10.1371/journal.pcbi.100202

6) Wang, Z.; Chen, Y.; Li, Y. (2004) A Brief Review of Computational Gene Prediction Methods; Genomics Proteomics Bioinformatics. 2004 Nov; 2(4): 216–221.

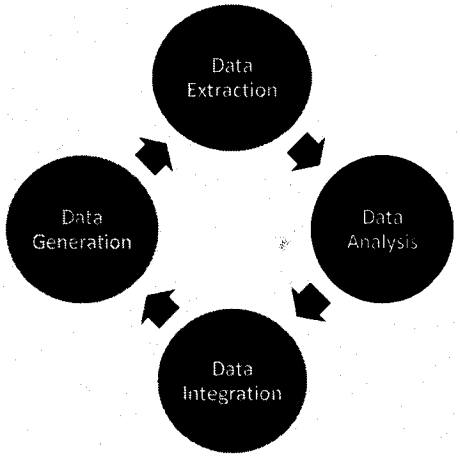
LAKSHMANGARH • RAJASTHAN

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

2: Introduction to databases

With the Technological advances in Biomedical Research like High -throughput Techniques and Lower-Cost DNA sequencing Technologies huge amount of Heterogeneous Data is getting generated daily. To manage this heterogeneous data we need better systems or tools. We also need better systems so that we can efficiently extract our data without distorting the information. For this purpose databases have been developed that can store the biological data in huge volumes and in a proper well defined manner for easy and efficient extraction of information.

Database helps in Extraction of data, its analysis, integration and generation of information or new data in a Fast and Efficient Manner.



2.1: What Is a Database?

Databases are collection of organized information that can be accessed managed and retrieved easily. These are composed of computer hardware and software for the management of data. With the help of a database, the information can be accessed, managed and retrieved easily and efficiently via varieties of search criteria. Sophisticated computational methodologies are being developed daily which help in easy and efficient extraction and analysis of biological data. Databases have evolved dramatically since their inception in the early 1960's. A simple database may contain records in the form of normal files

28/12/2021
REGISTRAR
Madanapalle Institute of Technology & Science
Lakshmanayam (Post), Andhra Pradesh

2.2: What Are Biological Databases?

A biological database is a large collection of organized biological data (DNA, RNA, gene etc.) which can easily be searched with the help of keywords. The data present in the biological databases can be easily retrieved for the public use. Database organizes the data in the form of Records and each record is known as an Entry

The data present in the biological databases is updated periodically and is cross referenced.

2.3: Purpose of Database:

A database serves various purposes like:

- Data Retrieval.
- Knowledge discovery.
  - Finding connections or relations between different pieces of data/information that were not known at the time that information was fed into the database.

In Biological Databases different types of information is stored like

- The Functions of Genes.
- Structure of Proteins.
- Localization of Proteins and Genes (both cellular and chromosomal)
- Clinical effects of mutations in Genes.
- Similarities of Biological Sequences and Structures.

The biological databases store and organize in a structured and searchable manner for humans and systems. They not only store and organize data but also share data with different databases. They also provide web application programming interfaces (APIs) for computers to exchange and integrate data from various database resources in an Automated Manner.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2.4: Types of Databases

We can classify databases on various criteria like the level of curations or the type of data they store.

- According to level of data curation or source, biological databases can be roughly divided into two categories:
  - Primary Databases
  - Secondary Databases

2.4.1: Primary Databases

Primary databases collect data directly from the Authors. They are archives of raw sequence or structural data submitted by the scientific community. Results from the experiments conducted are submitted directly into the database by researchers. The data in the primary database becomes permanent once accession number has been allotted. These databases are archival in nature. These three public databases (GENBANK, EMBL and DDBJ) closely collaborate and exchange new data daily and constitute the International Nucleotide Sequence Database Collaboration.

For the three-dimensional structures of biological macromolecules, there is only one centralized database, the Protein Data Bank. This database archives atomic coordinates of both Proteins and Nucleic Acids, determined by x-ray crystallography and NMR. They contain the amino acid sequence of proteins, their conformation, structure, and features such as active sites.

2.4.2: Secondary Databases

Secondary Databases contains computationally processed or manually curated information, based on original information from Primary Databases. They are also known as curated databases or knowledgebases. SWISS-Prot and Protein Information Resources (PIR) are examples of Secondary Databases. SWISS Prot provides detailed sequence annotation that

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

includes Structure, Function, and Protein Family Assignment. The Sequence Data are mainly derived from the Database of Translated Nucleic Acid Sequences stored in the EMBL Database. Protein Information Resources (PIR) is an integrated public database which supports genomic, proteomic and systems biology research and scientific studies. It helps in *the identification and interpretation of protein sequence information.*

2.5: Classification of Biological Database according to the scope of Data Coverage.

- Biological database is divided in two parts:
  - Comprehensive Database
  - Specialized Database

2.5.1: Comprehensive Database

Comprehensive Database covers different types of data from numerous species. For example GenBank, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) contain Nucleotide Sequences from various Organisms. Uniprot is a comprehensive database that provides high quality information on the protein sequence and also has information related to the functions performed by the proteins. This is also a freely accessible database.

2.5.2: Specialized Database

Specialized Databases contain specific types of data or data from specific organisms. WormBase is a database that provides accurate, current, accessible information concerning the genetics, genomics and biology of *C. elegans* and other related Nematodes. FlyBase is another database that contains information of the genes and genomes of *Drosophila.* HIV sequence database and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

GenBank EST database and Microarray Gene Expression Database at the European Bioinformatics Institute (EBI) are examples of some available gene expression databases.

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

2.6: Biological Databases can also be classified on the Basis of Nature of Data they contain

2.6.1: Sequence database

A Sequence database is a biological database that has a huge collection of nucleic acid sequences, amino acid sequences, or other polymer sequences. These databases can be nucleic acid sequence databases and protein sequence databases.

2.6.1.1: Nucleic acid sequence databases

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Nucleic acid sequences are stored in three primary sequence databases GENBANK, EMBL and DDBJ. The International Nucleotide Sequence Database Collaboration (INSDC) is an initiative that operates between DDBJ, EMBL-EBI and NCBI. It collects and provides the information on the nucleotide sequences along with the metadata. INSDC contains the data raw reads, through alignments and assemblies to functional annotation, containing contextual information related to samples and experimental configurations.

All the three databases (GENBANK, EMBL and DDBJ) are synchronized daily and data on nucleotide sequences contributed by research teams is shared between the collaborating organizations.

2.6.1.1.1: RefSeq: NCBI Reference Sequence Database

This is a collection which provides a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses.

28/12/2024
REGISTRAR
Lakshmangarh (Dist Sikar)

The main features of the RefSeq collection:

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators

2.6.1.1.2: GENBANK

It is a comprehensive database that contains publicly available nucleotide sequences for more than 3,00,000 organisms. The data is obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. It is an annotated collection of all publicly available nucleotide sequences. GenBank is accessible through the NCBI Entrez retrieval system.

Genbank integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed.

2.6.1.1.3: European Molecular Biology Laboratory (EMBL)

The EMBL Nucleotide Sequence Database is Europe's primary nucleotide sequence resource. In collaboration with DDBJ and GenBank the database is produced, maintained and distributed at the European Bioinformatics Institute. EMBL Database entries are grouped into divisions. The EMBL Database currently consists of 17 divisions with each entry belonging to exactly one division. Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages.

2.6.1.1.4: DDBJ (DNA Data Bank of Japan)

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

The DNA Data Bank of Japan is a public database of nucleotide sequences established at the National Institute of Genetics (NIG). The DDBJ Center also services Japanese Genotype-phenotype Archive (JGA), with the National Bioscience Database Center. The JGA is a permanent archiving service for genotype and phenotype data of human individuals. There are various services that are provided by DDBJ like AGD which is a controlled-access database for sharing individual-level genotype and phenotype information among specific researchers. ARSA is a service provided by DDBJ for the retrieval of the information using accession number and keywords. BLAST, Bioproject, CRISPRDirect, CLUSTAL W, DFAST etc. are some of the other services provided by DDBJ.

2.6.1.2.Protein Sequence databases are

- PIR,
- SWISS-PROT,
- TrEMBL,
- PROSITE and
- PFAM

2.6.1.2.1: Prosite

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. PROSITE is complemented by collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids.

2.6.2: Structural database

Structural databases help in producing, solving, refining and publishing the structure of a new material. One can consult these databases before synthesizing new compounds. PDB, SCOP and CATH are structure databases.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)

2.6.2.1: The Protein Data Bank (PDB)

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. It contains information related with the 3D structure of proteins, nucleic acids and complex assenblies. RCSB PDB, PDBe, PDBj, BMRB and EMDB are the members of PDB.

2.6.2.2: Structural Classification of Proteins (SCOP)

It provides a detailed and comprehensive description of the relationships of known protein structures. SCOP provides its data with the help of Representation State Transfer (REST) protocol. The classification is on hierarchical levels:

- Family,
- Super family and
- Fold

2.6.3: Functional database

A functional database is one which provides information related to the physiological roles of various biological entities like genes, enzymes etc. KEGG, Reactome and BioCyc are some of the examples of Functional Databases. BRENDA is an enzyme activity Database. It is Comprehensive Enzyme Information system. ExPASY-ENZYME is repository of information relative to the nomenclature of enzymes.

2.6.3.1: KEGG

It is a database that helps in understanding the high level functions and utilities of cells, organisms etc. It provides the computer based representation of biological systems which contain genomic and chemical information. It contains information of diseases and drugs and their inpact on normal biological systems.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)

2.6.3.2: BRENDA

BRENDA is an enzyme database. It has collection of the enzyme functional data. It is maintained and developed at the Institute of Biochemistry and Bioinformatics at the Technical University of Braunschweig, Germany. Information related to the functional properties of enzymes is taken directly from the works published by the scientists.

2.7: HOW TO RETRIEVE INFORMATION FROM DATABASES

To perform complex queries in a database often requires the use of Boolean operators like AND, OR, and NOT to indicate relationships between the keywords used in a search. With the help of various keywords and the combination of keywords, we can search the biological databases easily.

2.8: LIMITATIONS AND PITFALLS

- Redundancy is a major problem Affecting Primary Databases
  - Some problems are because of Human Errors.
  - Often, the same gene sequence is found under different names resulting in multiple entries and confusion about the data.
- Sequencing Error
  - Most errors in Nucleotide Sequences are caused by Sequencing Errors.

2.9: ADVANTAGES OF BIOLOGICAL DATABASES

- Easy access of Data/information
- Provide Information to answer various biological questions
- Open access to all users
- It allows indexing of all the data
- It Connects the experimental work or data with programming skills

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- With the help of databases, same infc n be accessed by multiple users at the same time.
- Different types of databases, helps us Biochemist down their search criteria and helps in quick retrieval of information.

collection
ormation re
ished by the

ROM DATA

e often requ
hips betwee
bination of

ing Primary
Human Erro
ice is foun
about the da

iences are ca

BASES

biological

MODY UNIVERSITY
NGARH • RAJASTHAN
5/12/2021
REGISTRAR
versity of Science and Technology
mengarh (Dist. Sikar) 332311

data with pr

Mod

References

- https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/
- Lesk, A. M. (2002). Introduction to bioinformatics. (2nd ed.). Oxford: Oxford University Press.
- Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer.
- Xiong, J. (2006). Essential Bioinformatics. (1st ed.).Cambridge University Press.
- Benson DA et al. Genbank; Nucleic Acids Res. 2010 Jan; 38(Database issue). doi: 10.1093/nar/gkp1024
- https://www.ncbi.nlm.nih.gov/genbank/
- http://pd.chem.ucl.ac.uk/pdnn/pubs/databs1.htm
- https://www.wwpdb.org/

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) Rajasthan

3: SEQUENCE DATABASES

Biological databases provide updated high quality data.

Sequence databases are biological databases that store large collection of nucleotide sequences and protein sequences digitally.

Sequence databases can be divided into nucleic acid sequence databases and protein sequence databases

The Nucleic acid sequence database is a collection of sequences from several sources.

These sources can be:

- GenBank
- RefSeq
- TPA
- PDB

3.1: Uses of these databases:

- discovery of novel genes
- identification of homologous genes
- chromosomal localization of genes
- analysis of alternative splicing
- detection of polymorphisms

GenBank, EMBL and DDBJ are the most comprehensive publically available nucleotide sequence databases.

3.2: GenBank

Genbank is a publically available comprehensive nucleotide sequence database which also supports bibliographic and biological annotation.

28/12/2024

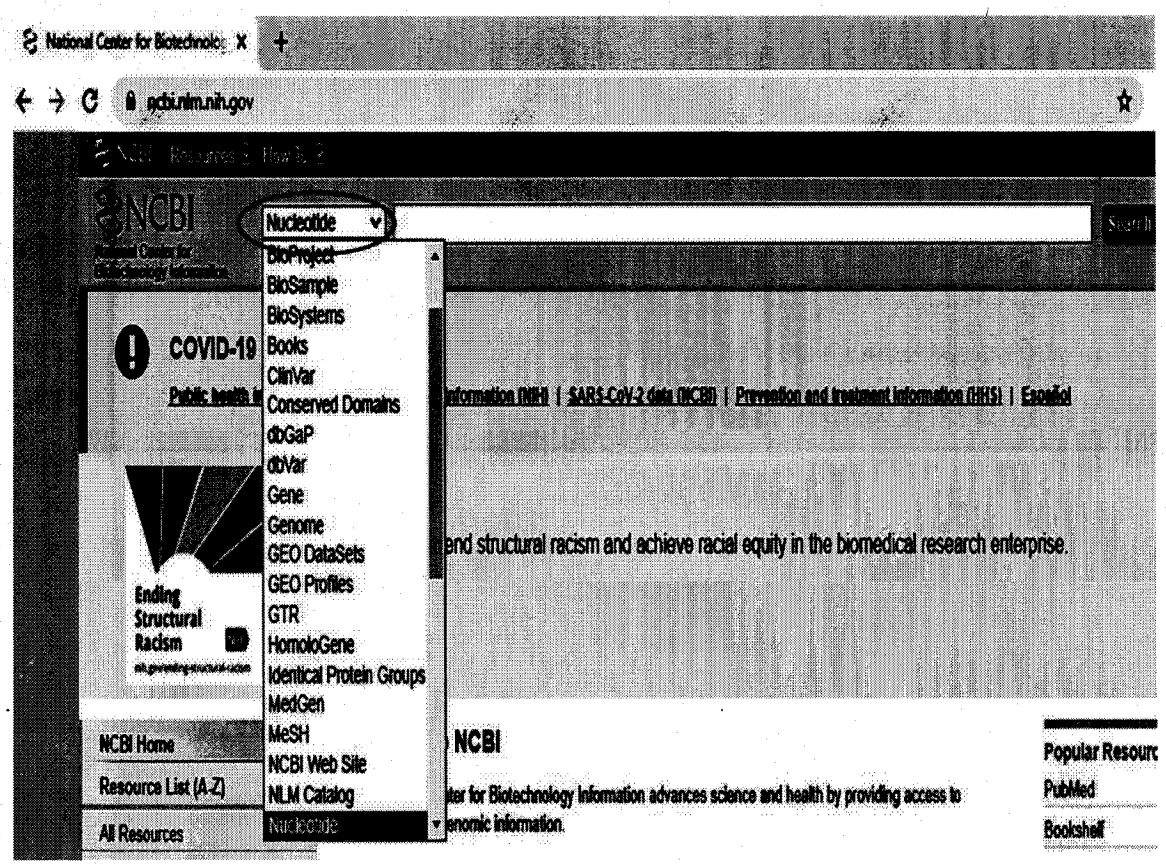
REGISTRAR
Moi University of Science and Technology

It is built and distributed by the National Center for Biotechnology Information (NCBI).

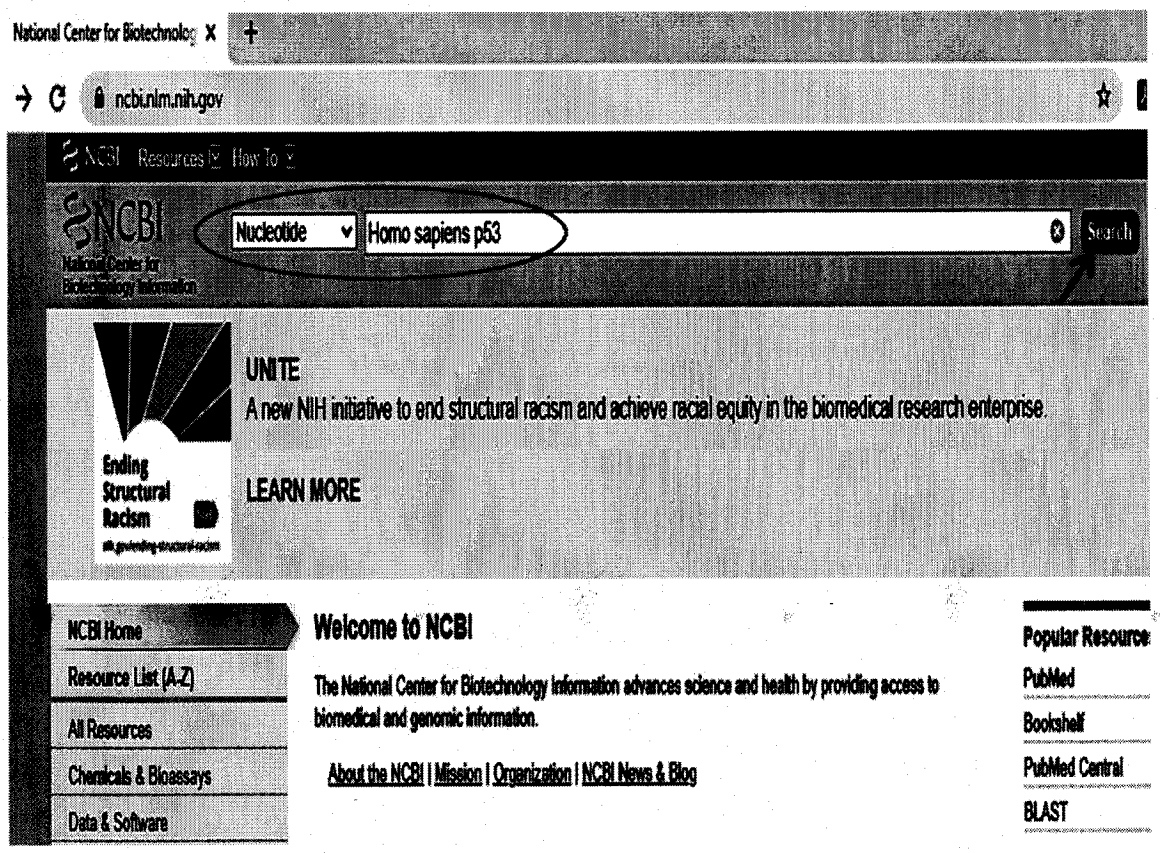
GenBank, EMBL Data Library, and (DDBJ) comprise the International Nucleotide Sequence Databases, in which information is exchanged daily to ensure a uniform and comprehensive collection of sequence information.

3.2.1: How to use it

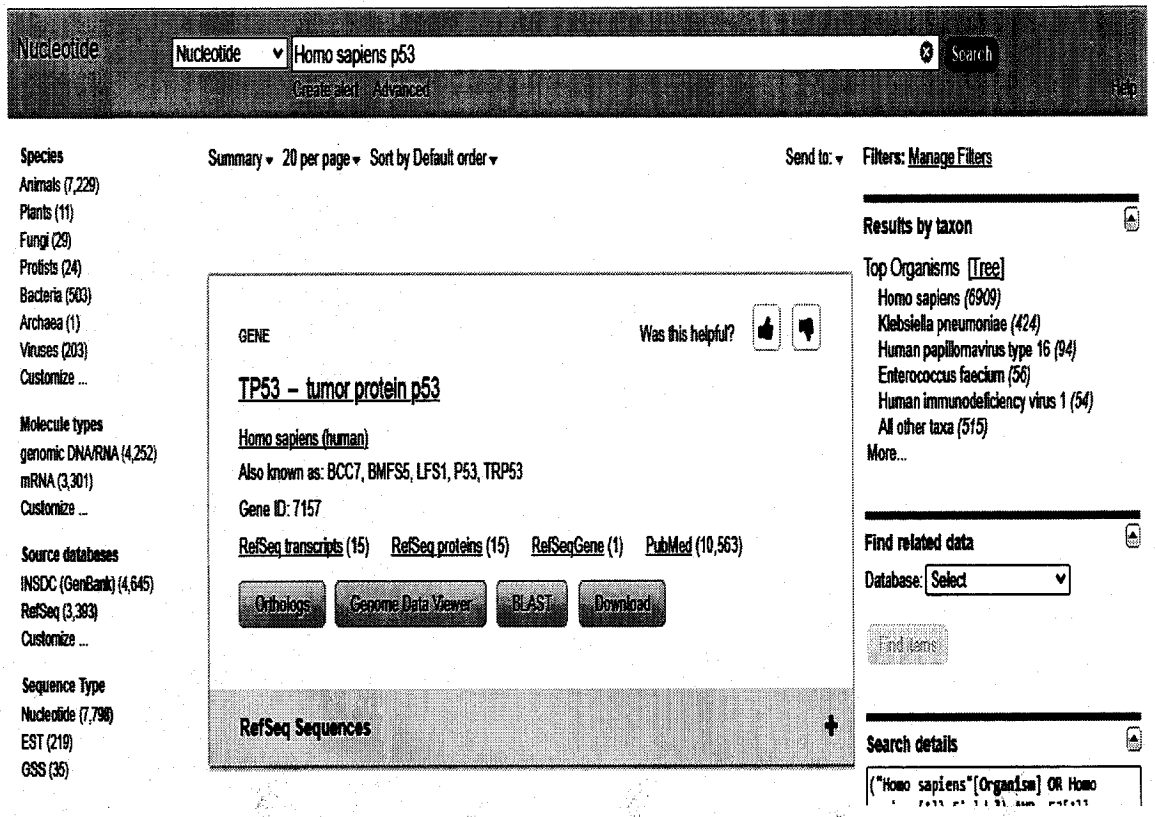
- There are several ways to search and retrieve data from GenBank.
- To search any nucleotide sequence-
- Go to the search, select nucleotide from the dropdown, enter the name of any gene whose sequence you want to know and click search



29/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dear Sikar) 332311



28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

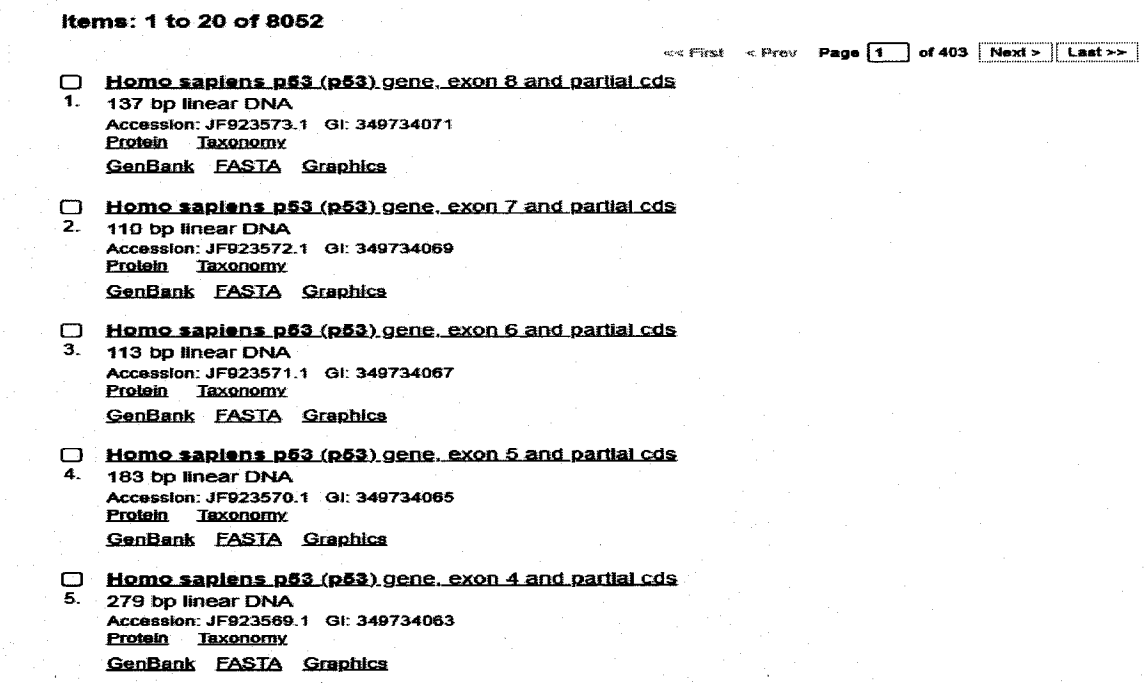


28/12/2024

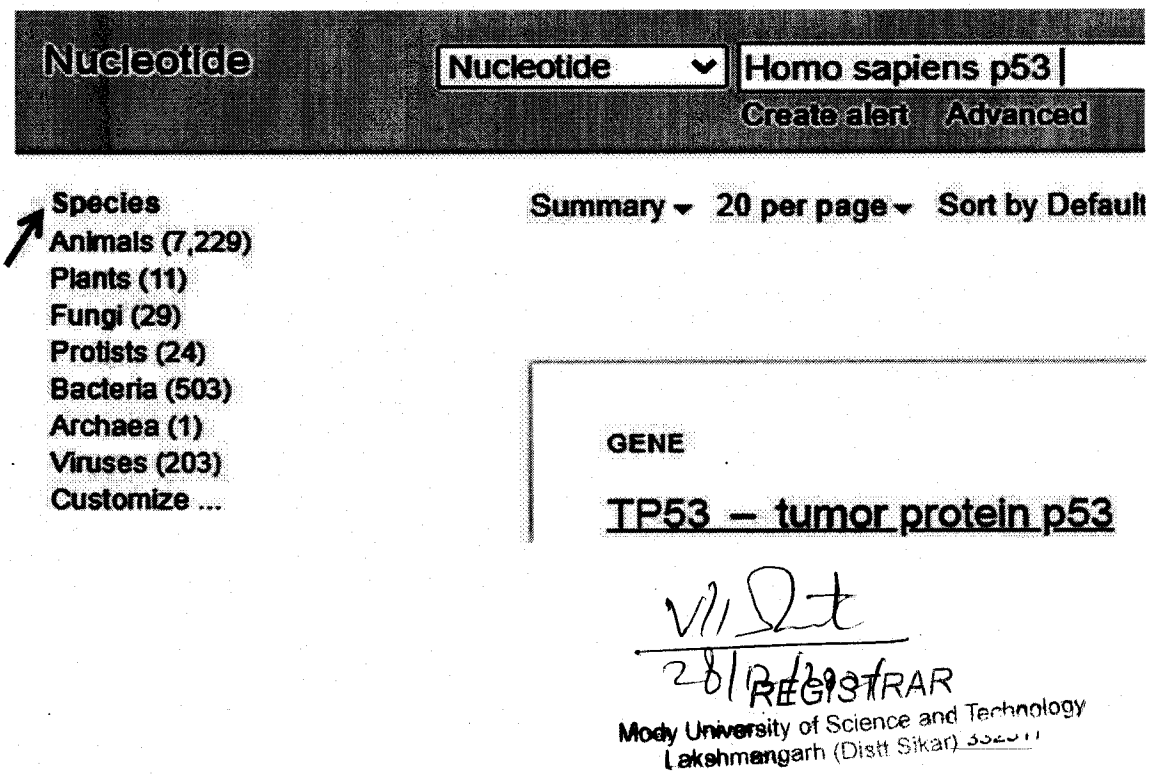
REGISTRAR

Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332311

On top of the results, we have total number of hits obtained against our query

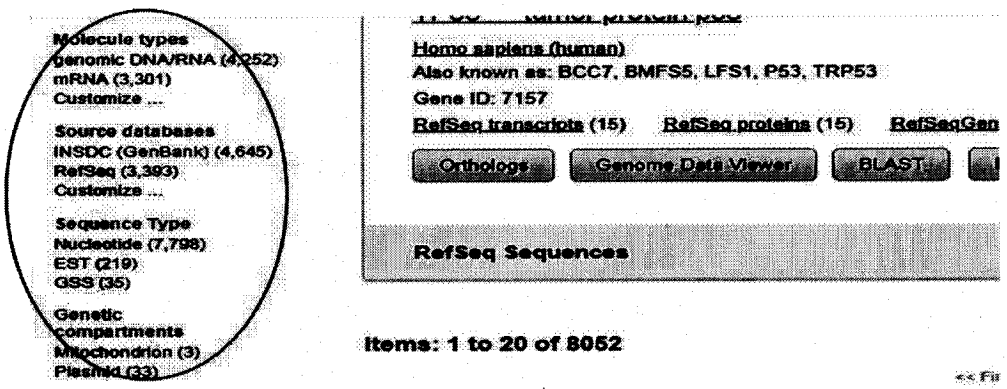


On left, there is a list of species depicting no.of records against various organisms

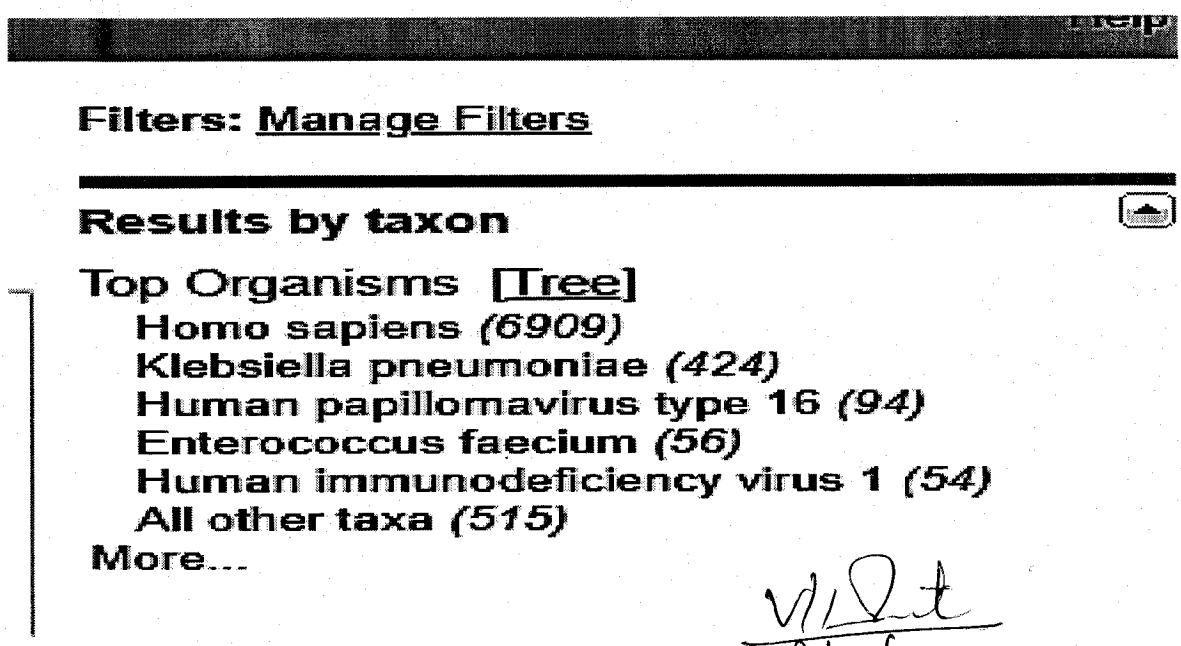


28/12/2017
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

We also get information about the source type, types of molecules presented in our results and also the type of sequences for example: nucleotide or GSS or EST.



On the right hand, we have results by Taxon from where we can further refine our search criteria.



28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

In the results that we have obtained, we can get some information on top based on which we can further make our decision on the hit to select as our final result. On top we have information like the length of the sequence in base pairs, the type of sequence i.e whether the sequence is of DNA or RNA, the name of the organism etc.

1. **Homo sapiens p53 (p53) gene, exon 8 and partial cds**
   137 bp linear DNA
   Accession: JF923573.1 GI: 349734071
   Protein Taxonomy
   GenBank FASTA Graphics

2. **Homo sapiens p53 (p53) gene, exon 7 and partial cds**
   110 bp linear DNA
   Accession: JF923572.1 GI: 349734069
   Protein Taxonomy
   GenBank FASTA Graphics

3. **Homo sapiens p53 (p53) gene, exon 6 and partial cds**
   113 bp linear DNA
   Accession: JF923571.1 GI: 349734067
   Protein Taxonomy
   GenBank FASTA Graphics

4. **Homo sapiens p53 (p53) gene, exon 5 and partial cds**
   183 bp linear DNA
   Accession: JF923570.1 GI: 349734065
   Protein Taxonomy
   GenBank FASTA Graphics

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

GenBank FASTA Graphics

9. **Homo sapiens p53 gene for P53, exon 5, partial cds**
184 bp linear DNA
Accession: AB118156.1 GI: 42821409
Protein Taxonomy
GenBank FASTA Graphics

10. **Homo sapiens mRNA for P53, complete cds**
2,451 bp linear mRNA
Accession: AB082923.1 GI: 23491728
Protein PubMed Taxonomy
GenBank FASTA Graphics

11. **Homo sapiens p53 gene, exon 7 and partial cds**
110 bp linear DNA
Accession: DQ075362.1 GI: 70611297
Protein Taxonomy
GenBank FASTA Graphics

Once we have chosen the result of our interest, we land on the main page of our result.

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1
FASTA Graphics

Go to:

```
LOCUS       JF923571                 113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-MAY-2011) Department of Medicine, Institute of
            Medical Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh
            221005, India
```

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

An accession no that is a unique identifier for all the entries that are made in GenBank.

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1

FASTA Graphics

Go to:

```
LOCUS       JF923571                113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
```

- The header section

GenBank

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1

FASTA Graphics

Go to:

HEADER

```
LOCUS       JF923571                113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-MAY-2011) Department of Medicine, Institute of
            Medical Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh
            221005, India
```

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

The features section

```
FEATURES             Location/Qualifiers
     source          1..113
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /isolation_source="healthy case in northeastern Indian
                     population"
                     /db_xref="taxon:9606"
     gene            <1..>113
                     /gene="p53"
     mRNA            <1..>113
                     /gene="p53"
                     /product="p53"
     CDS             <1..>113
                     /gene="p53"
                     /codon_start=3
                     /product="p53"
                     /protein_id="AEQ16484.1"
                     /translation="LAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPE"
     exon            1..113
                     /gene="p53"
                     /number=6
```

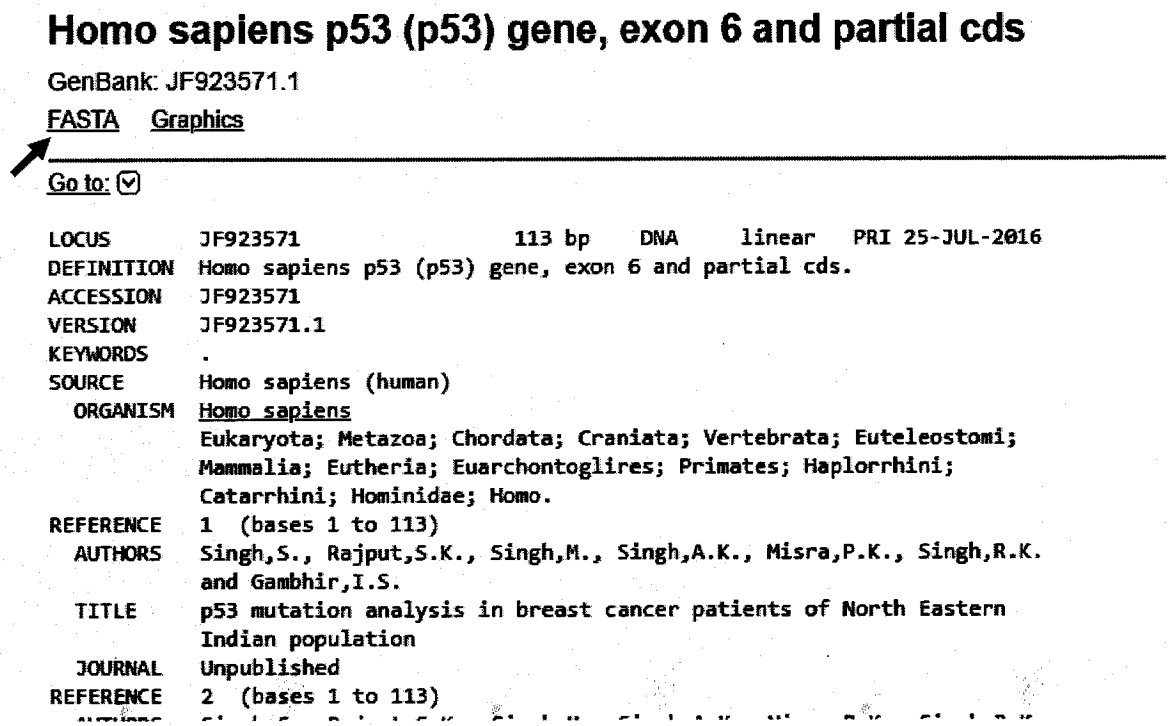
Origin section

```
ORIGIN
        1 gtctggcccc tcctcagcat cttatccgag tggaaggaaa tttgcgtgtg gagtatttgg
       61 atgacagaaa cacttttcga catagtgtgg tggtgcccta tgagccgcct gag
//
```

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

To retrieve the sequence, click on FASTA at the top of the page



Homo sapiens p53 (p53) gene, exon 6 and partial cds

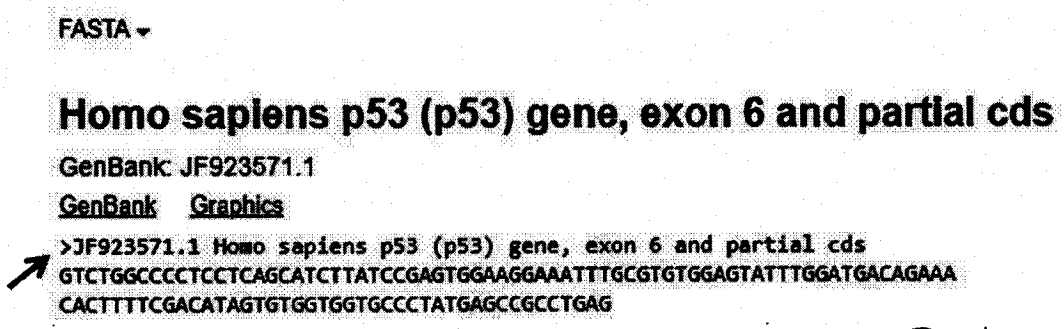
GenBank: JF923571.1

FASTA Graphics

Go to:

```
LOCUS       JF923571                113 bp    DNA     linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens p53 (p53) gene, exon 6 and partial cds.
ACCESSION   JF923571
VERSION     JF923571.1
KEYWORDS    .
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 113)
  AUTHORS   Singh,S., Rajput,S.K., Singh,M., Singh,A.K., Misra,P.K., Singh,R.K.
            and Gambhir,I.S.
  TITLE     p53 mutation analysis in breast cancer patients of North Eastern
            Indian population
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 113)
```

FASTA formats start with a more than symbol. The sequences is in continues form without any gap between two base pairs.



FASTA

Homo sapiens p53 (p53) gene, exon 6 and partial cds

GenBank: JF923571.1

GenBank Graphics

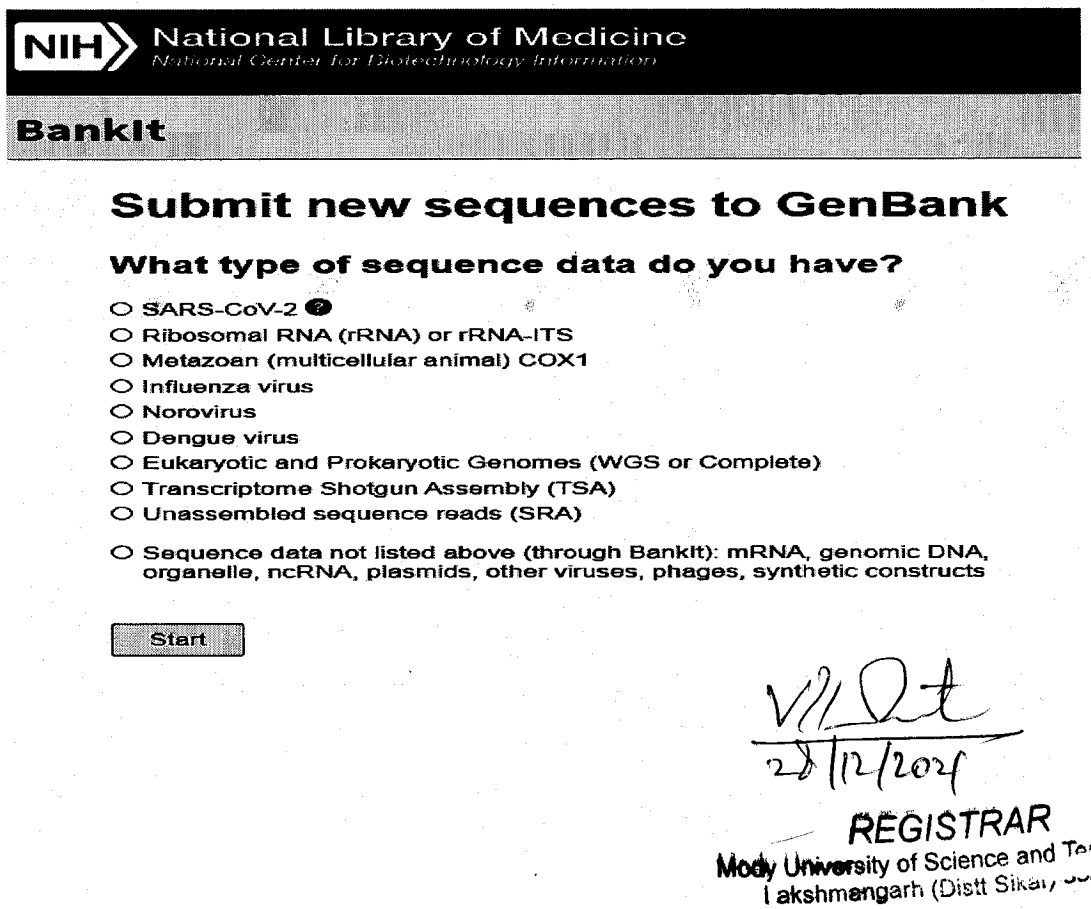
```
>JF923571.1 Homo sapiens p53 (p53) gene, exon 6 and partial cds
GTCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAA
CACTTTTCGACATAGTGTGGTGGTGCCCTATGAGCCGCCTGAG
```

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

3.2.2: Submitting data into GenBank

- **Web-based submission tools**
  - **BankIt**
  - **Submission Portal**
  - **tbl2asn**

3.2.2.1: BankIt



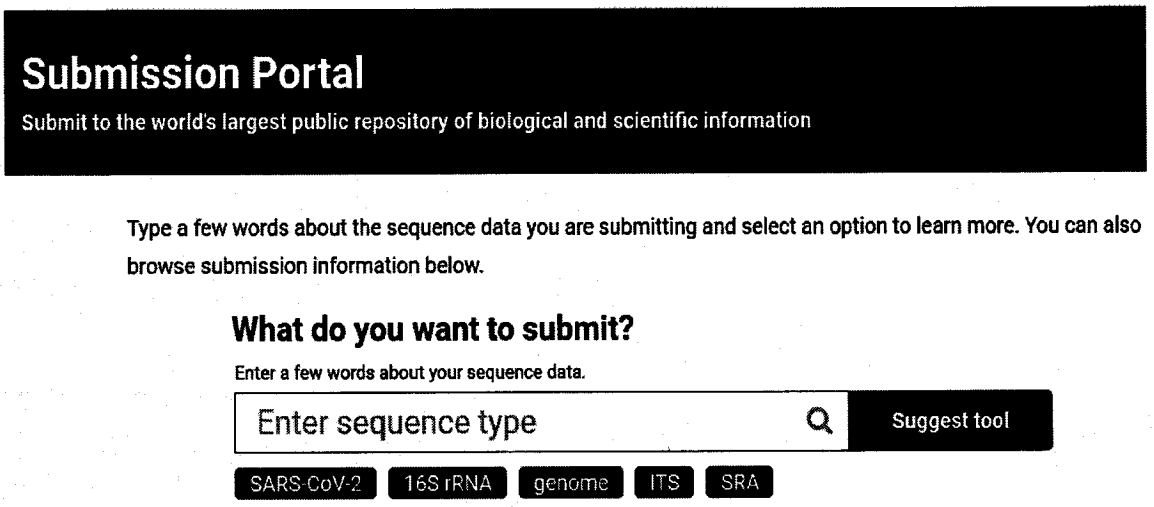
28/12/2024

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Distt Sikar, [illegible])

3.2.2.2: Submission Portal



- Supports the submissions of
  - ribosomal RNA (rRNA),
  - metazoan mitochondrial COX1,
  - Influenza,
  - Norovirus,
  - Dengue or
  - SARS-CoV-2 sequences
- One can also submit Genome and Transcriptome Assemblies through the Genomes and TSA portals, respectively.
- Submission preparation tools are required for uploading via the Submission Portal.

We can also email to gb-sub@ncbi.nlm.nih.gov

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Sikar)

3.2.2.3: tbl2asn

- a command-line program that automatically creates the sequence of records for submission to the GenBank.
- used primarily for submission of complete genomes and large batches of sequences.

Genome Workbench

- offers a rich set of integrated tools to study and analyze the genetic data.
- The allows us to prepare submissions of single eukaryotic and prokaryotic genomes.
- One can also use Genome Workbench for editing and visualizing files that are created by tbl2asn.

3.3: The European Molecular Biology Laboratory

- It is the primary nucleotide sequence resource maintained by the European Bioinformatics Institute (EBI), situated in the United Kingdom.
- In this The sequences of DNA or RNA are submitted directly from individual researchers.
- The EMBL Database in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Data Bank of Japan (NIG, Mishima) has been processing nucleotide sequence data reported world-wide since 1982.

3.3.1: How to use it

- Access to the sequence data is provided via ftp and several WWW interfaces.
- The Sequence Retrieval System or SRS allows the databases to be searched using, for example, sequence, annotations, keywords, author names.

Format

Sequence entry starts with an identifier line ("ID")

28/12/2021

REGISTRAR

Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").


```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```


28/12/2021

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Distt Sikar) 332311

3.4: DNA Data Bank of Japan

DDBJ initiated the DNA data bank activities in 1986 at the National Institute of Genetics (NIG).

It collects DNA sequence data mainly from Japanese and researchers from all over the world.

3.4.1: How to submit the data

- For submission, It offers interactive web interface as well as mass submission system.
- SAKURA is a Web-based tool of DDBJ which is used for nucleotide sequence submission and annotation.

3.5: Protein sequence databases

- The Protein database is a collection of sequences from several sources
  - translations from annotated coding regions in GenBank, RefSeq and TPA
  - records from SwissProt, PIR, PRF, and PDB
  - Protein sequences are the fundamental determinants of biological structure and function.

3.5.1: GenPept

- It is a part of the National Center of Biotechnology Information (NCBI)
- Its entries are derived from translations of the sequences present in Genbank, DDBJ and EMBL.
- The entries lack additional annotation and the database does not contain proteins derived from amino acid sequencing.
- The format is same as that of Genbank having header, features and origin sections.
- The format ends with double slashes.
-

29/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

3.5.2: Refseq

- RefSeq stands for Reference sequence which is a collection developed by NCBI.
- It provides a non-redundant collection of reference protein sequences.
- It has sequences for some limited organisms which includes human, mouse, rat, honeybee, zebrafish, cow, sea urchin, and several important plant species and also some 1100 viruses and 150 bacteria

3.5.3: Protein information resource

- It is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies.
- PIR was established in 1984 by (NBRF) the National Biomedical Research Foundation for the identification and interpretation of protein sequence information.
- It assists with proteomic and genomic data integration and the propagation and standardization of protein annotation.
- It is the oldest universal curated protein sequence database

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

3.5.4: Swiss Prot

- It is the leading universal curated protein sequence database.
- The database is non-redundant and is highly integrated with other databases.
- Literature-based curation is used to extract experimental data. This experimental knowledge is supplemented by manually confirmed results from various sequence analysis programs.

3.5.5: Translated EMBL Nucleotide Sequence Data Library

- It is a computer annotated protein sequence database.
- TrEMBL translates the nucleotide sequences in the DDBJ/EMBL/GenBank nucleotide sequence database.
- It was introduced to make sequences available quickly.
- TrEMBL has two main sections; SP-TrEMBL and REM-TrEMBL

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries which should be incorporated into SWISS-PROT.
- REM-TrEMBL (REMaining TrEMBL) contains the entries that are not to be included in SWISS-PROT.
- Most REM-TrEMBL entries are immunoglobulins and T-cell receptors.

3.5.6: Uniprot

- It stands for Universal Protein Resource
- It is a publically available protein sequence database.
- UniProt is a collaboration between the EMBL-EBI, the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).
- It is a comprehensive resource for protein sequence and annotation data.

It provides information on

- function,
- sub-cellular location,
- expression,
- interaction,
- structure,
- protein family,
- domain and
- sequence.

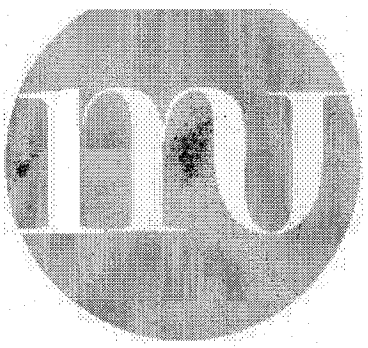
Apart from this, one can also find various published papers in support of the information provided.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Important links

- https://www.ncbi.nlm.nih.gov/genbank/
- https://www.ncbi.nlm.nih.gov/WebSub/
- https://submit.ncbi.nlm.nih.gov/
- https://www.ncbi.nlm.nih.gov/protein/
- https://www.ebi.ac.uk/
- https://www.ddbj.nig.ac.jp/index-e.html
- https://www.ncbi.nlm.nih.gov/refseq/
- https://www.uniprot.org/



MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

4. STRUCTURE DATABASES

Structure databases or databanks contain atomic coordinates, bibliographic citations, and other information related to structural studies. Structural databases are the essential tools for all crystallographic works. They are used in the process of producing, solving, refining and publishing the structure of a new material. Before deciding to synthesize a new compound the database could be used to check how many compounds with a particular chemical composition have been reported.

After synthesizing and indexing the unit cell of a material the database can be searched to see if a material with the same or a similar unit cell is already known. If a material is found in the database with a similar unit cell to the new material, then its structure may be close enough (i.e. same symmetry and similar unit cell contents) to be used as the starting model for the refinement of the new material. To verify the results of a structure refinement the database can be consulted to find structures that have comparable bond distances, bond angles or coordination environments to the new structure.

MODY

4.1: Protein Data Bank (PDB)

- Protein Data Bank archive is a single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.
- The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.
- Currently, Most of the structures are determined by X-ray diffraction, but about 10% of structures are determined by protein NMR.
- The .pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity.
- Its fixed-column width format is limited to 80 columns, which was based on the width of the computer punch cards that were previously used to exchange the coordinates.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

UNIVERSITY

The format has a Header section along with the title and author records. This provide information related to the researchers who defined the structure.

Remark can contain free-form annotation, but they also accommodate standardized information.

Atom describe the coordinates of the atoms that are part of the protein.

Hetatm describe coordinates of hetero-atoms, that is those atoms which are not part of the protein molecule.

Ter indicates the end of a chain of residues.

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

4.2: RCSB PDB

- RCSB PDB stands for Research Colaboratory for Structural Bioinformatics PDB.
- This resource is supported by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies
- As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.
- 178229 biological macromolecular structures enabling breakthroughs in research and education.

4.3: SCOP

- SCOP stands for Structural Classification of Proteins.
- Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin.
- It aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- includes 68,816 non-redundant domains representing 772,354 protein structures

LNIVERSITY
LAKSHMANGARH • RAJASTHAN

The main levels of the classification are:

- *Family-* It groups closely related proteins with a clear evidence for their evolutionary origin.
- In most cases, their relationship is detectable with current sequence comparison methods, e.g. BLAST, PSI- BLAST, HMMER.
- *Super-family-* It brings together more distantly related protein domains.
- Their similarity is frequently limited to common structural features that along with a conserved architecture of active or binding sites, or similar modes of oligomerization suggest a probable evolutionary ancestry.
- *Fold* groups super-families on the basis of the global structural features shared by the majority of their members. These features are the composition of the secondary structures in the domain core, their architecture and topology.
- Fold is an attribute of a super-family but the constituent families of some super-families that have evolved distinct structural features can belong to a different fold.

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)

- *IUPR* (Intrinsically Unstructured Protein Region) organizes super-families of proteins or protein regions that do not adopt globular folded structure.
- *Classes* bring together folds and IUPRs with different secondary structural content.
- These include all-alpha and all-beta proteins, containing predominantly alpha-helices and beta-strands, respectively, and 'mixed' alpha and beta classes (a/b) and (a+b) with respectively alternating and segregated alpha-helices and beta-strands, and the fifth class of small proteins with little or no secondary structures.

4.3.1: ORGANISATION AND FACILITIES OF SCOP

SCOP is organized as a tree structure.

- Entering at the top of the hierarchy the user can navigate through the levels of Class, Fold, Super-family, Family and Species to the leaves of the tree which are structural domains of individual PDB entries.
- An alternative hierarchy of Folds, Super-families and Families by the date of solution of the first representative structure is also provided.

We can search in SCOP

- With the help amino acid sequence,

  The Sequence similarity search facility allows any sequence of interest to be searched against databases of protein sequences classified in SCOP using the algorithms BLAST, FASTA or SSEARCH.

- From a keyword

  The keyword search facility returns a list of SCOP pages containing the word entered or combinations of words separated by a series of boolean operators.

- From a PDB identifier

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

The PDB entry viewer links PDB entries to various graphical views, external databases and SCOP itself.

4.3.2: OTHER USES OF SCOP

- Non-redundant sequence databases are available from SCOP and it helps in the evaluation of sequence alignment methods.
- Assignment of protein structures to sequences using the intermediate sequence library PDB-ISL.
- Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL.

4.4: CATH

- CATH database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains.
- The CATH database provides hierarchical classification of protein domains based on their folding patterns.
- Domains are obtained from protein structures deposited in the Protein Data Bank and both domain identification and subsequent classification use manual as well as automated procedures.
- The CATH database is a classification of protein domains, based not only on **sequence information**, but also on **structural and functional properties**.
- CATH contains **114,215 domains**.
- It is classified in a hierarchical scheme with four main levels called **class (C)**, **architecture (A)**, **topology (T)** and **homologous** super-family **(H)**.
- From these words they have derived the name CATH.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

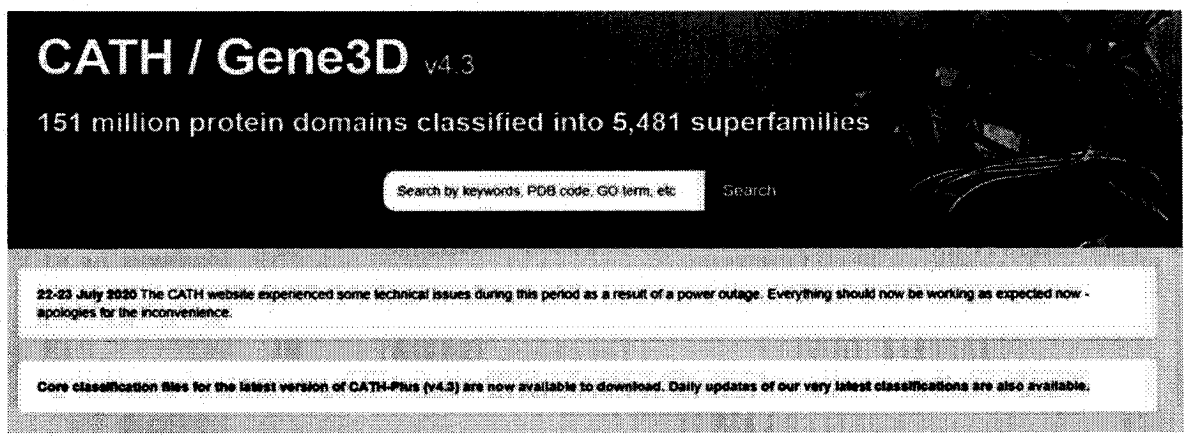
- At the **C-level**, domains are grouped according to their secondary structure content into four categories: mainly alpha, mainly beta, mixed alpha-beta; and the last category contains domains with only few secondary structures.
- The **A-level** groups domains according to the general orientations of their secondary structures.
- At the **T-level**, the connectivity (ie the order) of the secondary structures is taken into account.
- The grouping of domains at the **H-level** is based on a combination of both sequence similarity.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

4.4.1: Searching in CATH



MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh

IMPORTANT LINK

- **https://scop.mrc-lmb.cam.ac.uk/**
- **https://www.cathdb.info/**
- **https://www.rcsb.org/**
- **https://www.wwpdb.org/**

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

29/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

5: STRUCTURE DATABASES (Part 2)

5.1: PFAM

- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
- Pfam, version 27.0, contains 14 831 Pfam-A families. Of these families, 4563 have been classified into 515 clans.
- Pfam also generates higher-level groupings of related entries, known as clans.
- A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.
- Profile HMMs turn a multiple sequence alignment into a position-specific scoring system, which can be used to align sequences and search databases for remotely homologous sequences.
- Profile HMMs have a formal probabilistic basis and uses consistent theory behind gap and insertion scores.
- Profile hidden Markov models (HMMs) have several advantages.
- HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency.
- The HMMER package, is a set of programs that allow us to create and manipulate profile HMMs and databases of profile HMMs, perform sensitive searches of sequence and profile HMM databases and create multiple sequence alignments efficiently

5.1.1: FORMAT of PFAM

- **Plain Pfam format**
- **Flatfile**

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

```
YHI1_MYCLE/3-102     TIFTKIINRELPGRFVYEDDDVVAFLTIEPMTQGHTLVVP..CAEIDQMQMVDP...AI....FGRVIAVSQLIGKGVCRAFNAERAGVII.....AGFEVPHLHIHVFPTHSL
YHIT_MYCHR/7-111     ELFLKIIKREEPATILYEDDKVIAFLDKYAHTKGHFLVVP..KNYSRNLFSISD...ED....LSYLIVKAREFALQEIKKLGATGFKLLINNEPDAEQSIFHTHVHIIPYYKK
YHIT_HELPY/2-104     NVFEKIIQGEIPCSKILENERFLSFYDINPKAKVHALVIP..KQSIQDFNGITP...EL....MAQNTSFIFEVVEKL..GIKEKGYKLLTNVGKNAGQEVMHLHFHILSGDKH
YHIT_AZOBR/9-115     NVFARILRGEIPCKKVLETEHALAFHDINPQAPTHILVIP..KGAYVDMDDFSA...RATEAEIAGLFRAVGEVARGA..GAAEPGYRILSNCGEDANQEVPHLHIHVFAGRRL
PKIA_DICDI/68-172    TIFAKIISGAIPSKKFYDDEYCIAIEDINPQAPVHLLVIP..KLAVGGLSDVAANVDLEKYKESMGHIMSKIHHIASLK....GADSYRLVINEGVLGQQSVRGLHIHIPNW...
ZB14_MAIZE/17-123    TIFDKIIKKEIPSTVVYEDEKVLAFRDINPQAPTHILIIPKVKDGLTGLAKAEE...RHIEI.LGYLLYVAKVVAKQE..GL.EDGYRVVINDGPSGCQSVYHIHVHLLGGRQM
Y961_HAEIN/5-110     TIFSKIIRKEIPANIVYQDELVTAFRDISPQAKTHILIIP..NKVIPTVNDVTE...QDEVA.LGRLFSVAAKLAKEE..GVAEDGYRLIVNCNKHGGQEVFHLHMHLVGGEPL
YCFF_ECOLI/5-110     TIFSKIIRREIPSDIVYQDDLVTAFRDISPQAPTHILIIP..NILIPTVNDVSA...EHEQA.LGRMITVAAKIAEQE..GIAEDGYRLIMNTNRHGGQEVYHIHMHLLGGRPL
YHIT_CAEEL/21-125    TLFGKIIRKEIPAKIIFEDDEALAFHDVSPQAPIHFLVIP..KRRIDMLENAVD...SDAAL.IGKLMVTASKVAKQL..GM.ANGYRVVVNNGKDGAQSVFHLHLHVLGGRQL
HINT1_RAT/16-120     TIFGKIIRKEIPAKIIFEDDRCLAFHDISPQAPTHFLVIP..KKYISQISAAED...DDESL.LGHLMIVGKKCAADL..GLK.KGYRMVVNEGSDGGQSVYHVHLHVLGGRQM
YHIT_SYNY3/5-109     TIFSKIIRREIPAAIVYEDDLCLAFKDVNPQAPVHVLLIP..KKPLPQLSAATP...EDHAL.LGHLLLKAKEVAADL..GI.GDQFRLVINNGAEVGQTVFHLHLHILGGRPF
YHIT_SYNP7/5-109     TIFGKIIRREIPADIVYEDDLCLAFRDVAPQAPVHILVIP..KQPIANLLEATA...EHQAL.LGHLLLTVKAIAAQE..GL.TEGYRTVINTGPAGGQTVYHLHIHLLGGRSL
Y866_METJA/2-105     CIFCKIINGEIPAKVVYEDEHVLAFLDINPRNKGHTLVVP..KKHYERFDEMPD...DE....LCNFIKGVKKTVEVL.KKLGFDGYNIVMNNGRVAGQEVNHVHFHIIPRYEG
```

5.2: Nucleic acid structure databases

The 3D structure of nucleic acid can be well understood by investigating the different types of nucleic acid functions such as nucleic acid–protein interactions, functional RNAs, etc.

Nucleic Acid Database (NDB)

- The Nucleic Acid Database (NDB) was the first database for nucleic acid structures.
- It was designed in the early 1990s.
- It contains the contents of primary data.
- It also contains some derived geometric data, classification of structure and motifs, standards for describing nucleic acid features,
- NBA also has tools and software for analysis of nucleic acids.
- The structures available in the NDB are annotated with various information that are unique to the nucleic acids
- The structures of NBA are almost not available from the original entry of PDB.
- The structures of NBA are analyzed by using computational methods of structural geometries.
- They are then classified by searching the sequence, structure levels, and secondary structure by incorporating the PDB entries to the NDB.

NDB Entry Format

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

- The main content of NDB is the primary structure information that is related to the nucleic acids that are acquired from the PDB and also from classification and derived data
- The primary data present in the NDB is obtained from PDB.
- It contains experimental files, identifiers, structural descriptions, citations, crystal data, coordinate information, and various details related to the crystallization, data collection, and structure refinement
- The structures available in the NDB database possess their own personal summary page that is mostly appropriate to that particular structure.
- There are about four sections of data in the structure summary page, including
    - primary structure information,
    - downloads,
    - derived structural data, and
    - images.

Home page of NDB where one can search for the 3D structural information of DNA and RNA along with some advanced searches.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh

NUCLEIC ACID DATABASE

A Portal for Three-dimensional Structural Information about Nucleic Acids

As of 26-May-2021 number of released structures: 11337

Search DNA | Search RNA | Advanced Search

Enter an NDB ID or PDB ID

Search for released structures

Welcome to the NDB

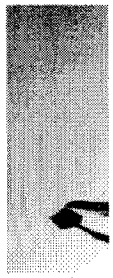
The NDB contains information about experimentally-determined nucleic acids and complex assemblies.
Use the NDB to perform searches based on annotations relating to sequence, structure and function, and to download, analyze, and learn about nucleic acids.

Search Structures

Search DNA
Search DNA and its complexes

Search RNA
Search for RNA structures in the NDB archive or in the Non-Redundant list

Advanced Search
Search for structures based on structural features, chemical features, binding modes, citation and experimental information



28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Featured Tools in NBA

- **RNA 3D Motif Atlas**
- **Non-redundant Lists**
- **RNA Base Triple Atlas**
- **WebFR3D**
- **R3D Align**

Tools available in NBA

Nucleic Acid tools:

- **DNA**
  - DNAProDB
  - Assignment of DNA conformers
- **RNA**
  - RNA Basepair catalog
  - RNA 3D motif Atlas
  - Representative RNA 3D Structure sets
  - R3D Align
  - WebFR3D
  - RNA viewer

FRASS

Nucleic Acid Softwares

- FR3D
- RNAView
- RNAMLView
- 3DNA
- FreeHelix98
- Hel_param
- The UNAFold Web server

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332311

References

- http://ndbserver.rutgers.edu/
- http://pfam.xfam.org/
- Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, Alex Bateman, Pfam: The protein families database in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D412–D419, https://doi.org/10.1093/nar/gkaa913

UNIVERSITY
LAKSHMANGARH • RAJASTHAN

29/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

6. SEQUENCE ALIGNMENT (PART 1)

What we understand by Sequence Alignment?

- Arranging DNA, RNA or Protein sequences in order to find the regions of similarity.
- This help to derive functional, structural and evolutionary relationships between them.
- We can also make prediction about the 3D structures of proteins.
- If two sequences from two different organisms are similar then we can say there is a possibility that both sequences are Homologous i.e they share some common ancestry.

Types of Sequence Alignments

Sequence Alignment can be classified on the basis of:

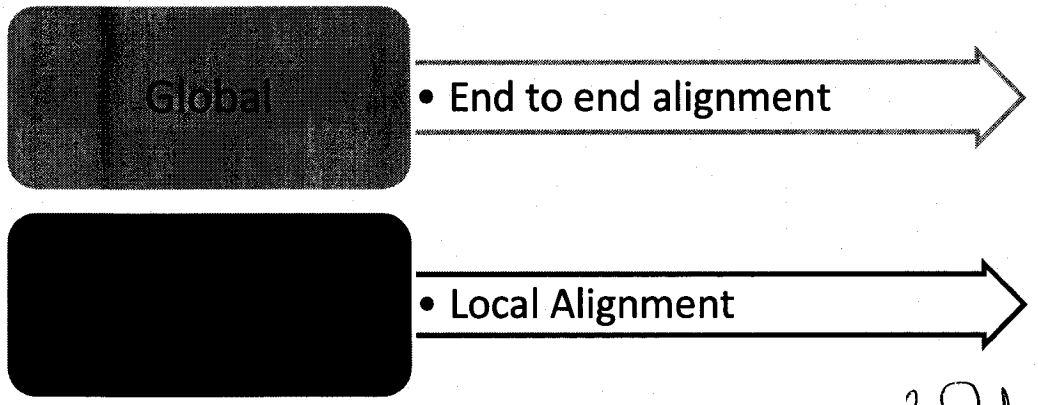
1. Sequence Length-

   How much portion of the sequences under study are being covered.

2. Number of sequence-

   If we are aligning two sequences or we are trying to align multiple sequences

Based on sequence Length

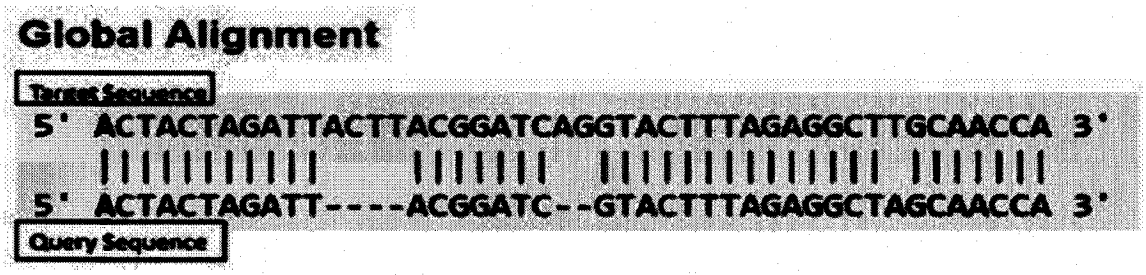


28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Global Sequence Alignment

- In this alignment sequences are aligned along their entire length to include as many matching characters as possible.
- We usually do global alignment, when we want to compare two homologous genes or the genes that perform the same function in two different species
- Or to compare two proteins with similar function

Example of Global Sequence Alignment

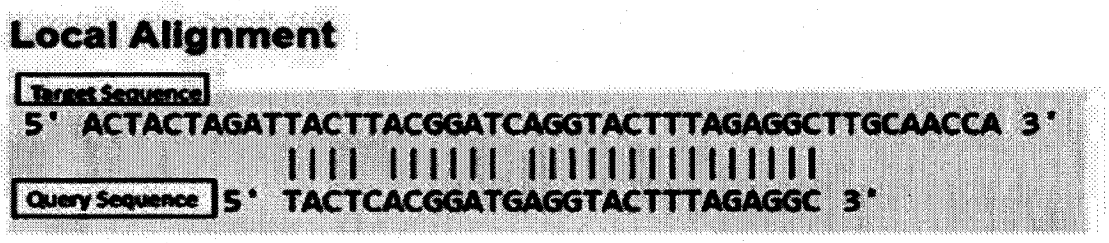


Local Sequence Alignment

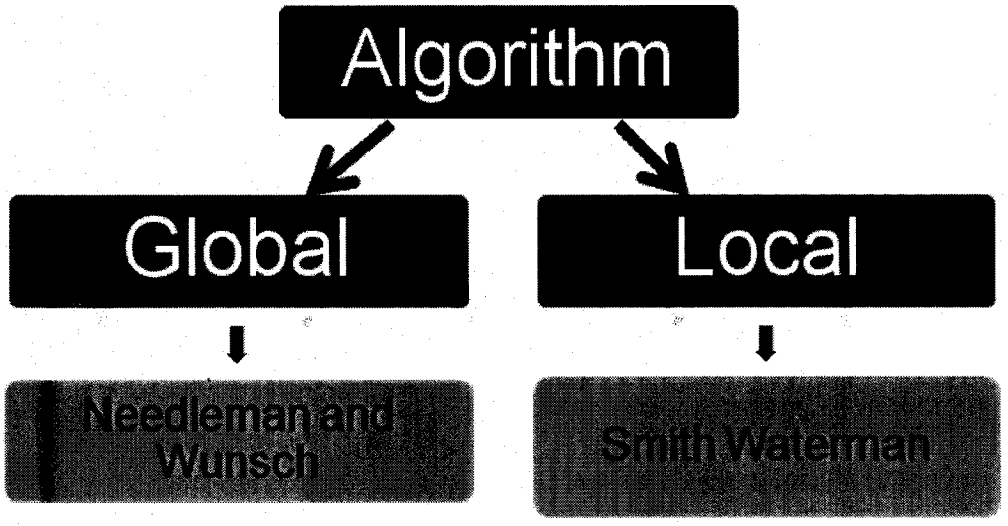
- The sequences are aligned to find a region of higher density or strong similarity.
- This helps in finding conserved domains or motifs in two protein sequences or some conserved patterns in two DNA sequences.
- The target sequence and a query sequence and our query sequence is not covering the target sequence in its entire length.

28/12/2024
REGISTRAR
Lakshmangarh (Distt Sikar) 332...

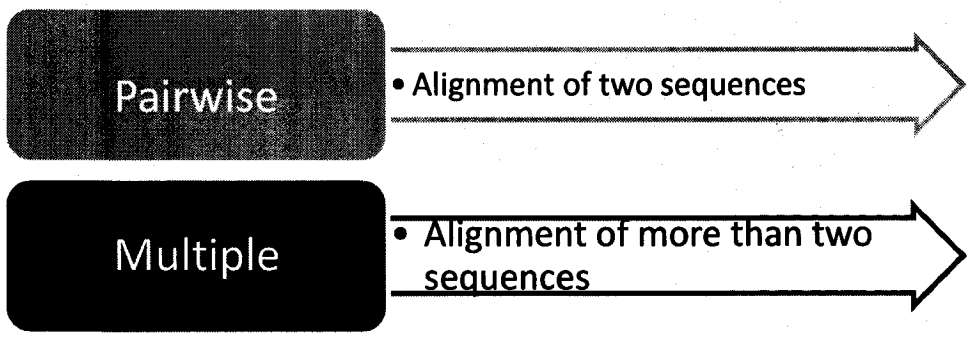
Example of Local Sequence Alignment



Algorithms for Sequence Alignments



Based on Number of Sequence



28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Pairwise Sequence Alignment

The methods that are used for Pairwise alignment of nucleotide and protein sequences are:

1) Dot Plot method

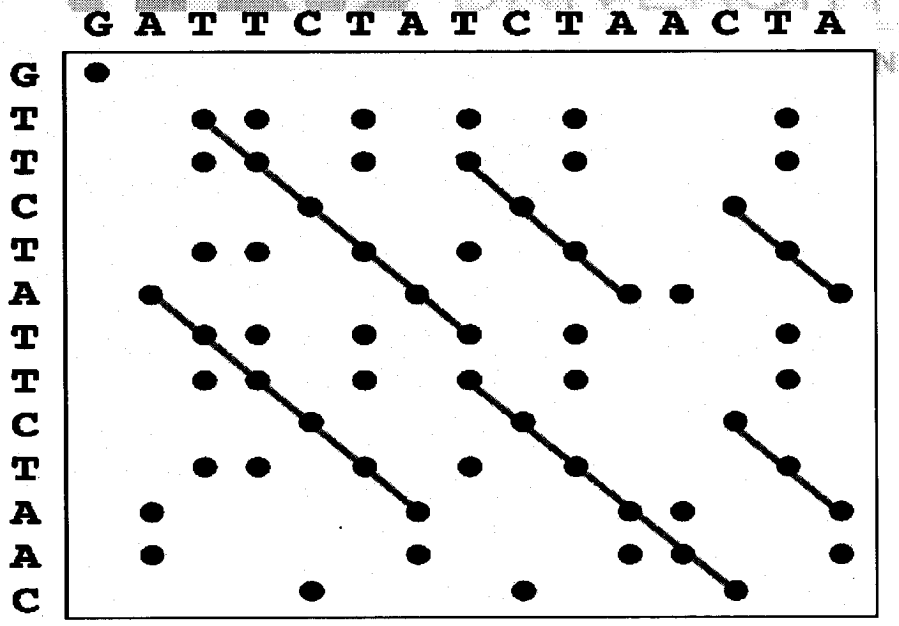
2) Dynamic Programming method

3) Heuristic Method

Dot Plot

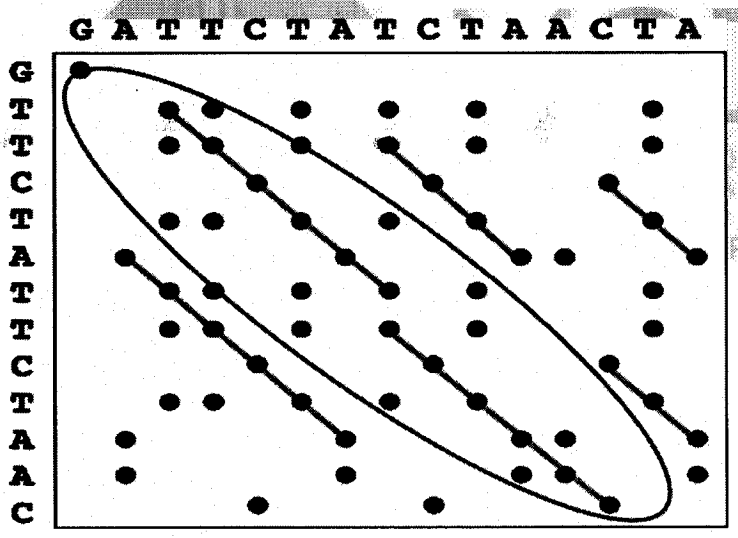
- It is graphical method used for aligning two sequences for the identification of the region of maximum similarity and dissimilarity, depicted by the presence or absence of DOTS.
- Also known as the dot matrix method.
- The comparison is done by scanning each and every residue of one sequence for similarity with all residues in the other sequence.

MODY UNIVERSITY



28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

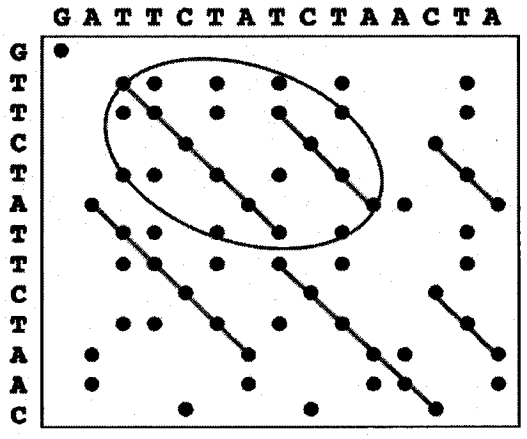
- We pick one residue from our query sequence and compare it with the residues on the X axis.
- We plot a dot where ever we find the same residue along the X axis.
- When the two sequences have substantial regions of similarity, many dots line up to form continuous diagonal lines, which reveal the sequence alignment as shown in example.
- If there are interruptions in the middle of a diagonal line, they indicate insertions or deletions in the sequences.
- Based on the patterns of lines in the matrix, we can say the following:
- If we obtain a Principal diagonal line, it will show identical sequence.



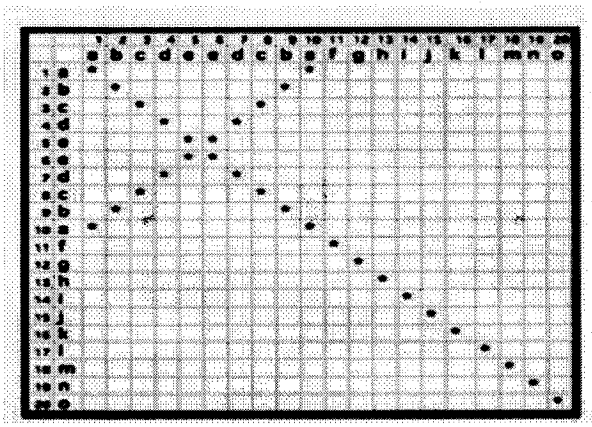
28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar) 332...

- If we have Multiple diagonal lines, they indicate repetition in the sequences.



- **Reverse diagonal (perpendicular to diagonal) indicate INVERSIONs**



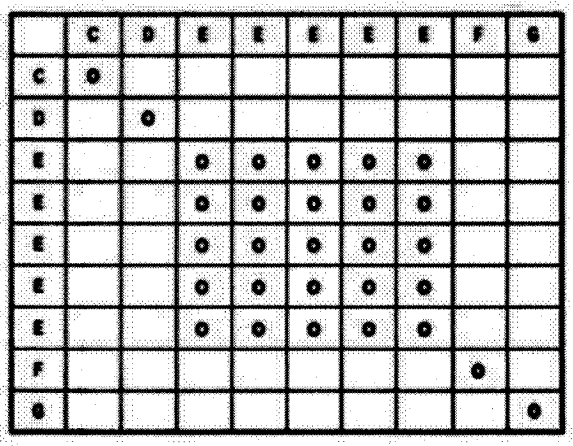
10DY
IVERSITY
HMANGARH • RAJASTHAN

- **Reverse diagonal crossing diagonal (X) indicate PALINDROMES or the sequences are palindromic.**

	A	T	G	C	G	T	A
A	*						*
T		*				*	
G			*		*		
C				*			
G			*		*		
T		*				*	
A	*						*

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 33...

- **If we obtain lines that Form of a box as shown in this example, they indicate the Low complexity region**



Limitations of Dot Matrix method

- It lacks statistical rigor in assessing the quality of the alignment.
- It is difficult for the method to scale up to multiple alignment.
- For longer sequence, memory required for the graphical representation is very high. So long sequence can not be aligned.
- Aligning larger sequences will take longer time

Applications of Dot Matrix method

- Shows the all possible alignment between two nucleic acid and amino acid sequences
- All kind of local and global alignment can be traced
- Help to recognize large region of similarity.
- An excellent approach for finding sequence transposition.
- To find the non sequential alignment.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh Sikar

SOFTWARES

- Dotmatcher : https://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher
- Dottup : https://www.ebi.ac.uk/Tools/seqstats/emboss_dottup/
- MatrixPlot : http://www.cbs.dtu.dk/services/MatrixPlot/
- Dotter : https://sonnhammer.sbc.su.se/Dotter.html
- Dothelix

Dynamic Programming

- This method basically breaks down a large problem into small sub-problems and then uses the solution of the sub-problems to compute the solution of the larger problem.
- Used for global as well as for local sequence alignments using Needleman-Wunsch and Smith-Waterman algorithms respectively.

Heuristic Method

When a single sequence is to be compared against the whole database, we use heuristic methods like BLAST and FASTA.

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

References:

- Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press
- Lesk, A. M. (2002). Introduction to bioinformatics. Oxford: Oxford University Press.
- Polanski, A.; Kimmel, M. (2007) Bioinformatics. Springer

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

7: SEQUENCE ALIGNMENT (PART 2)

Dynamic programming methods

- Match: this means that there is no change since their divergence. The residue of the query sequence is the same as the residue in the target sequence.
- Once we get a match, we assign some value to it, which we call as a match score
- Mismatch: this means that there is a substitution since their divergence. The residue of the query sequence is the different from the residue in the target sequence.
- It is quite common that when comparing sequences we will find that at some positions in some of the sequences there will be some extra residues which we call insertions, or missing residues which we call deletions.
- A gap in one of the sequences simply means that one or more residues have been deleted from the sequence, or we could also say that there is an insertion of residues in the second sequence.
- When aligning sequences, introducing gaps in the sequences can allow an alignment algorithm to match more terms in comparison to a gap-less alignment can.
- to get a useful alignment, minimizing gaps in an alignment is important.
- Too many gaps can lead to a meaningless alignment
- Gap penalties are used to adjust alignment scores based on the number and length of gaps.
- It can be changed each time an alignment is run. This will affect the number of gaps, their length and position in the sequence alignment.

DYNAMIC PROGRAMMING METHOD

- A method that determines optimal alignment by matching two sequences for all possible pairs of characters between them.
- It is a more quantitative method. In this a dot matrix is converted into a scoring matrix to account for matches and mismatches between sequences.
- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

STEPS:

- **Initialization:**

The first step in the global alignment dynamic programming approach is to create a matrix with M + 1 columns and N + 1 rows where M and N correspond to the size of the sequences to be aligned.

M is no.of residues in target sequence

N means no.of residues in query sequence.

Matrix filling(scoring):

We fill the matrix with highest possible scores. For the calculation of the scores, we need the match score, mismatch score and the gap penalty. We have three calculations based on which we decide the highest score to be filled in the matrix.

These calculations are:

$$M_{i,j} = Maximum\left[M_{i-1,j-1} + S_{i,j},\ M_{i,j-1} + W,\ M_{i-1,j} + W\right]$$

Traceback and aligning :

Move from last corner and follow arrow.

Now lets consider the two given sequences.

ACTGATTCA

ACGCATCA

The length of the sequence 1 and sequence 2 are 9 and 8 respectively.

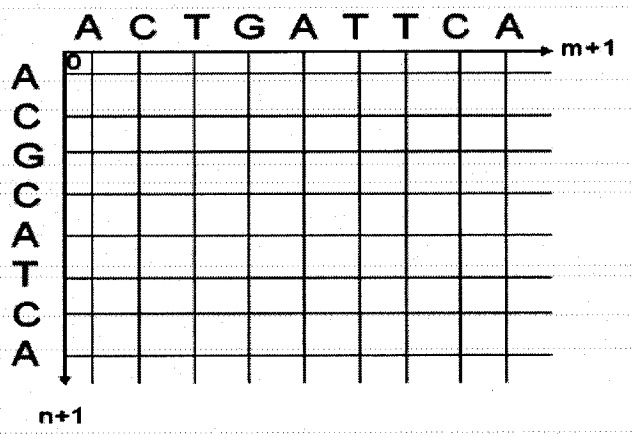
For match..we will give a score of 2, for mismatch -3 and gap penalty is -2

This is the M+1 and N+1 matrix

Then the scoring schema has to be introduced which can be user defined with specific scores.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)



Initialization Step

Let's assumes that there is gap penalty.

First row and first column of the matrix can be initially filled with 0.

If the gap score is assumed, the gap score can be added to the previous cell of the row or 0^{th} row and to the 0^{th} column as shown here.

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

Matrix Fill Step

The second and crucial step of the algorithm is matrix filling starting from the upper left hand corner of the matrix.

From the assumed values, add the match or mismatch (assumed) score to the diagonal value.

$$M_{1,1} = Max\ [M_{0,0} + S_{1,1},\ M_{1,0} + W,\ M_{0,1} + W]$$

•

Lets generalize this equation

M(I,J)= M(i-1,j-1)+S(I,j)

M(i,j-1)+W

M(I-1, j)+W

M(I,j)- coordinates of the cell

S(I,j)- score corresponding the cell of given coordinate (whether it's a match or a mismatch)

W- gap penalty

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

We compute all the values and choose the highest score

In this example, the value of M(i-1,j-1) ie M(0,0) is 0

To this add the score of match or mismatch

We will check whether the jth residue of our query sequence is matching with the ith residue of the target sequence.

If there is a match we will add the score of match.

In this example, yes there is a match, so we will add the score of match which is 2

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh

		A
	0	-2
A	-2	2

In second equation we will look at the score in the M(I,j-1) cell

Here the value at M(1,0) is -2

And the gap panelty is -2, so the value will be -4

In equation 3 value at M(0,1) is again -2 and the gap penalty is also -2 so the value comes out to be -4

Out of all these 3 values 2 is the highest value and hence the score at M(1,1) will be 2

This is how the final matrix will appear.

Step 2: Assign scores

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

28/12/2021

REGISTRAR
Moody University of Science and Technology

Trace back Step

The final step in the algorithm is the trace back for the best alignment.

We will start our trace back from the cell with the highest score.

In this example, we have 8 at the right most bottom corner of the matrix.

We will check for the next highest value adjacent to this cell.

We have values of 4,6 and 4 in the adjacent cells

Out of these 6 is the highest value which is in the diagonally adjacent cell.

After 6 we will again look for the highest value in the same manner.

We have 4 which is also in the diagonally adjacent cell.

Once again lets check for the highest value.

Now here we have 6 which is the value of the horizontal left cell.

MODY

Step 3: Trace back

The optimal path is traced beginning from the lower right-hand corner

		A	C	T	G	A	T	T	C	A
	(0)	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	(2)	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	(4)	(2)	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	(4)	2	0	-2	-4	-6
C	-8	-4	0	-1	(2)	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	(4)	2	0	-2	2
T	-12	-8	-4	0	-2	2	(6)	(4)	2	0
C	-14	-10	-6	-2	-4	0	4	2	(6)	4
A	-16	-12	-8	-4	-5	-2	2	1	4	(8)

28/12/2021

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Distt Sikar) 332...

Aligning the sequences

The diagonal lines indicate alignment

```
CA
||
CA
```

If we are moving horizontally left, like here from 4 to 6 we will insert a gap in the sequence on the Y axis

```
?TCA
  ||
?-CA
```

And if we are moving vertically up, we will insert a gap in the sequence at the X axis or our target sequence

```
-ATTCA
 || ||
CAT-CA
```

So we have alligned A with A, C with C, then we have a gap in our query sequence so this is T of target sequence and a gap in query sequence. Then T is aligning with T, A with A and then this time we have a gap in our target sequence against C in the query sequence.

Then G is aligning with G and again we have a gap in our query sequence against T of the target sequence. Then C is aligning with C and A is in alignment with A.

Result:

This analysis yielded the following alignment:

```
ACTG-ATTCA
||   || ||
AC-GCAT-CA
```

The alignment score is equal to the value in the lower right-hand corner of the matrix (8).

- **Softwares which are used for pairwise sequence alignment are**
- **SIM, (http://bioinformatics.iastate.edu/aat/align/align.html)**
- **SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html)**
- **LALIGN (www.ch.embnet.org/software/LALIGN form.html)**

28/11/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8: MULTIPLE SEQUENCE ALIGNMENT

A Multiple Sequence Alignment is an alignment of more than two sequences.

Multiple alignments are computationally much more difficult than pair-wise alignments.

There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments can

Multiple sequence alignment (MSA) methods refer to a series of algorithmic solution for the alignment of evolutionarily related sequences, while taking into account evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions.

These methods can be applied to DNA, RNA or protein sequences.

Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment.

Multiple sequence alignment is also an essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.

Multiple sequence alignment also has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

MSA is indeed an important modeling tool whose development has required addressing a complex combination of computational and biological problems.

8.1: Commonly used algorithms

Given a set of biological sequences (RNA, proteins, DNA), the purpose of a MSA method is to align the sequences in a way that will either reflect their evolutionary, functional or structural relationship Usually the multiple sequence algorithms assume that the sequences are similar in all its length and they behave like global alignment algorithms.

They also assume that there are not many long insertions and deletions. Thus the algorithms will work for some sequences, but not for others.

there are exhaustive and heuristic approaches used in multiple sequence alignment.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

exhaustive approaches are not feasible in most cases, heuristic programs are commonly used

8.2: Progressive Alignment Method

Progressive alignment depends on the stepwise assembly of multiple alignment and is heuristic in nature.

It first conducts pairwise alignments for each possible pair of sequences using the Needleman–Wunsch global alignment method and records these similarity scores from the pairwise comparisons.

The scores can either be percent identity or similarity scores

Both scores correlate with the evolutionary distances between sequences.

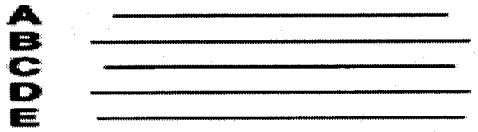
The scores are then converted into evolutionary distances to generate a distance matrix for all the sequences involved.

A simple phylogenetic analysis is then performed based on the distance matrix to group sequences based on pairwise distance scores. As a result, a phylogenetic tree is generated.

The tree reflects evolutionary proximity among all the sequences.

the resulting tree is an approximate tree and does not have the rigor of a formally constructed phylogenetic tree.

we have 5 different sequences A, B, C, D and E

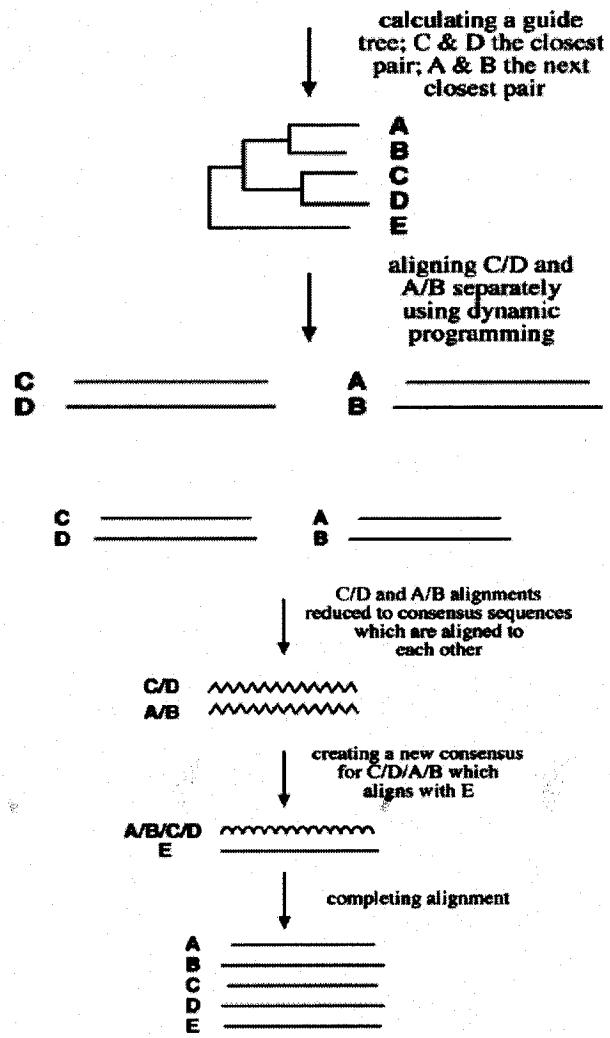


all individual pairwise alignment and construction of distance matrix

	A	B	C	D	E
A	—				
B	11	—			
C	20	30	—		
D	27	36	9	—	
E	30	33	20	27	—

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist Sikar)



8.3: Softwares for MSA

Clustal (www.ebi.ac.uk/clustalw/) is a progressive multiple alignment program available either as a stand-alone or on-line program.

CLUSTAL 2 has two versions ClustalW and ClustalX. The W version provides a simple text-based interface and the X version provides amore user-friendly graphical interface.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

8.3.1: CLUSTAL W

In Clustal W we can select the type of the output format

Multiple Sequence Alignment by CLUSTALW

ETE3 | MAFFT | CLUSTALW | PRRN

Help

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: ◉ FAST/APPROXIMATE ○ SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): ◉ PROTEIN ○ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Choose File No file chosen

Execute Multiple Alignment | Reset

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1, Window size: 5, Gap Penalty: 3

Number of Top Diagonals: 5, Scoring Method: PERCENT

For SLOW/ACCURATE:

Gap Open Penalty: 10.0, Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN)

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

Multiple Alignment Parameters:

Gap Open Penalty: 10, Gap Extension Penalty: 0.05

Weight Transition: ○ **YES (Value:** 0.5 **),** ◉ **NO**

Hydrophilic Residues for Proteins: GPSNDQERK

Hydrophilic Gaps: ◉ **YES** ○ **NO**

Select Weight Matrix: BLOSUM (for PROTEIN)

Type additional options (delimited by whitespaces) below:

(**-options** for help)

Execute Multiple Alignment | Reset

The type of the pairwise alignment that we need ie FAST and approximate or slow and accurate.

We then choose the type of sequence we want to enter id DNA or protein

CLUSTAL W also provides some options in parameters that users can change as per their requirements like the scores for gap opening and gap extension, the type of matrix to be used.

In case of fast and accurate pairwise alignment, the user can change the range of K-tuple, window size, gap panelty etc.

28/12/2024

REGISTRAR

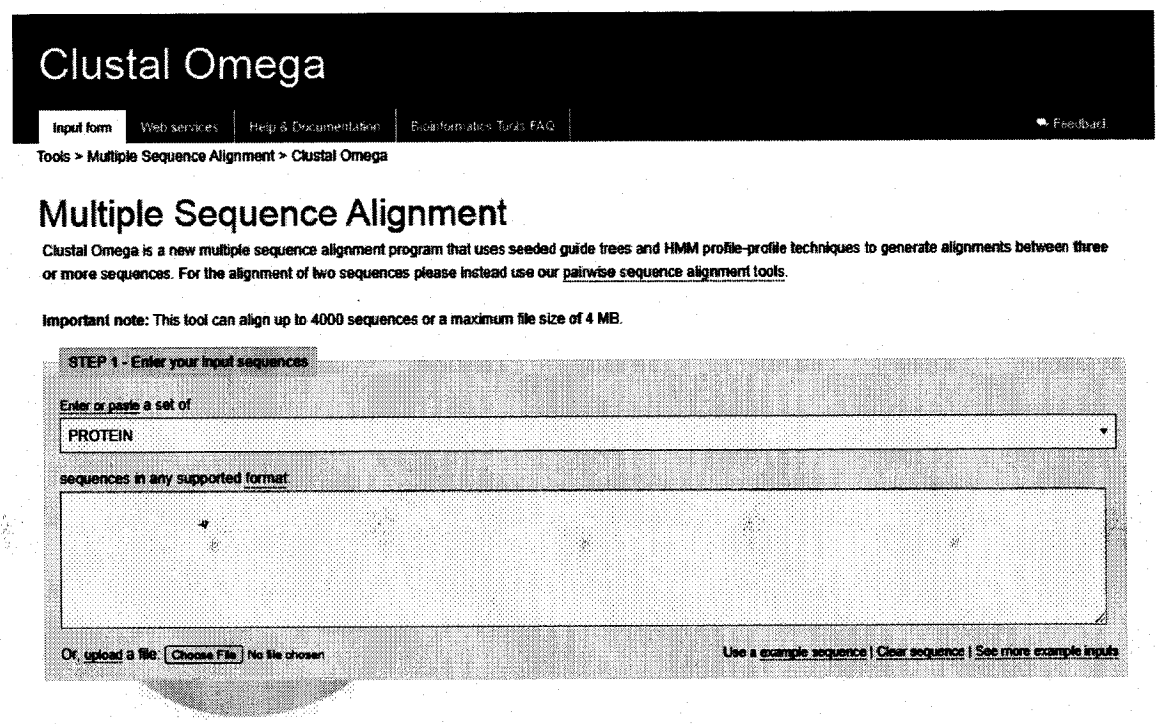
Mody University of Science and Technology

Lakshmangarh (Sikar)

8.3.2: CLUSTAL Omega

Clustal Omega is a multiple sequence alignment program that uses seeded guide trees and HMM techniques to generate alignments between three or more sequences.

It produces biologically meaningful multiple sequence alignments of divergent sequences.



Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

In the home page we have Sequence Input Window

Select the type of sequences to be aligned ie Protein sequences, DNA sequences or RNA sequences

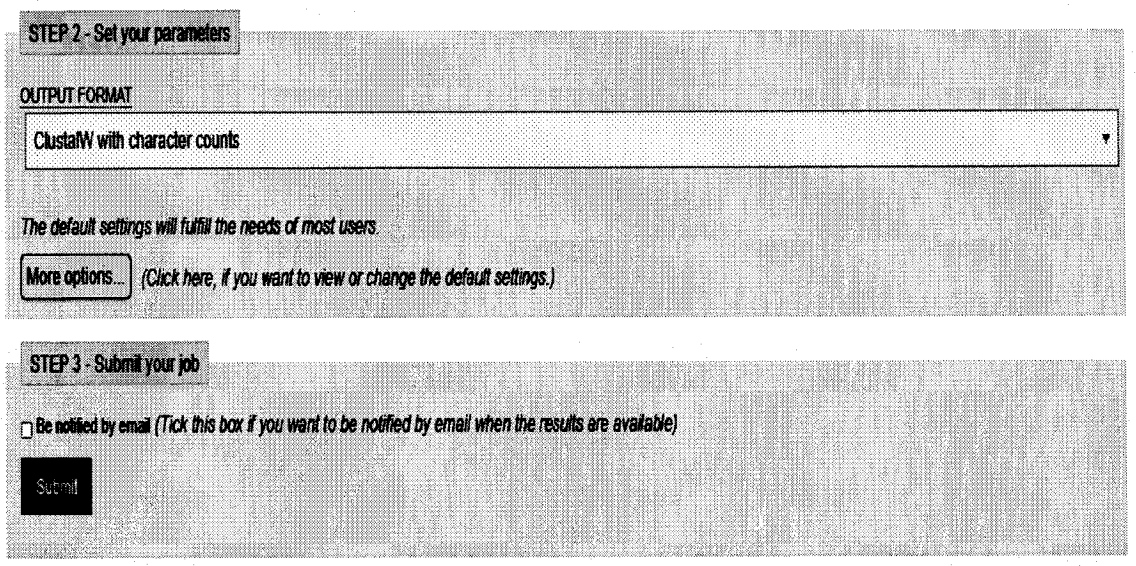
Three or more sequences to be aligned can be entered directly into this box

Sequences can be in GCG, FASTA, EMBL (Nucleotide only), GenBank, PIR/NBRF, PHYLIP or UniProtKB/Swiss-Prot (Protein only) format.

One can also upload the sequence files.

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

- **Step 2 - Set Your Parameters**



There are various parameter options for the users to obtain the alignment most suited for their requirements.

One can also choose the format of the output file generated.

And the last step is submission

8.3.3: EMBOSS Cons

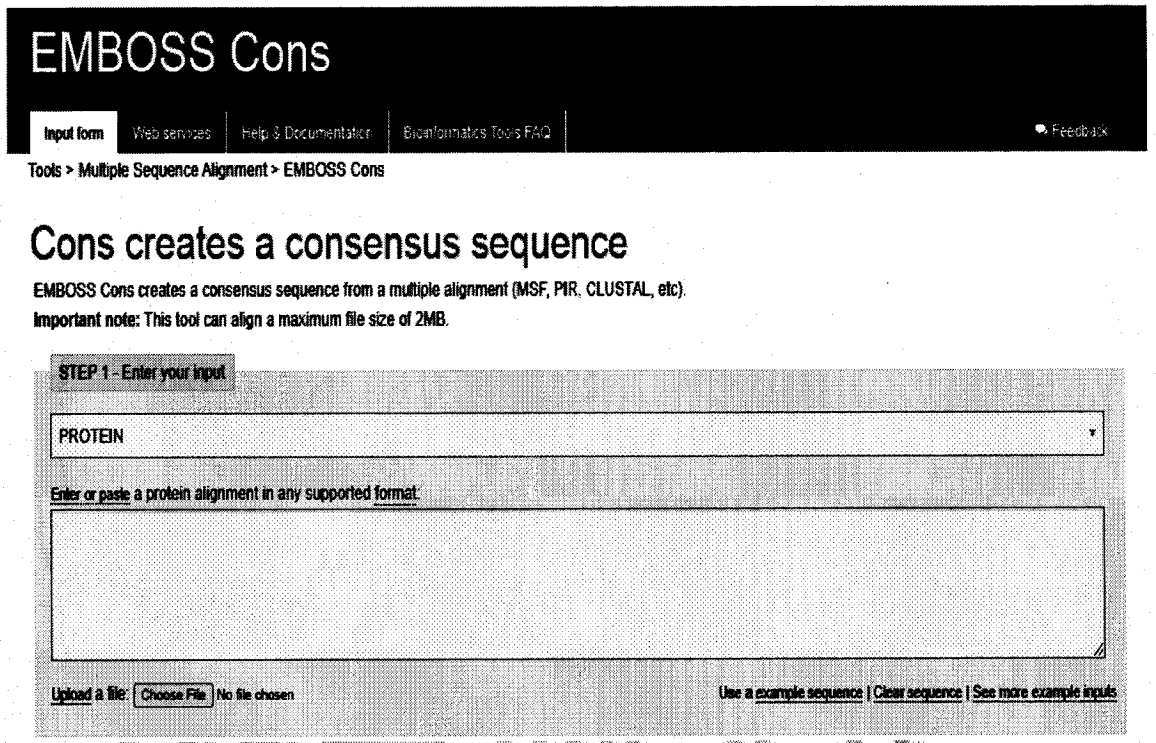
EMBOSS Cons creates a consensus sequence from a multiple alignment

This tool can align a maximum file size of 2MB.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

26/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



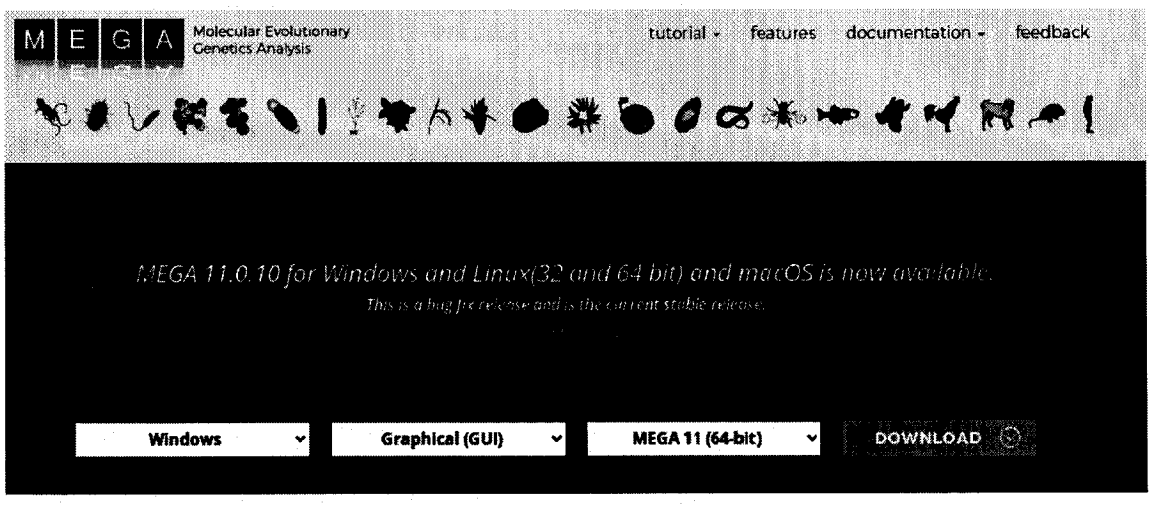
8.3.4: MEGA

It stands of molecular evolutionary genetics analysis

MEGA can be used with either a graphical user interface or a command-line interface which is useful for batch or scripted execution.

MEGA supports sequence alignment using both the ClustalW and MUSCLE programs.

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)



Alignment (or refinement) is done in the Analysis Explorer

Once the sequences are in the Alignment Explorer (AE), we select Alignment either *ClustalW* or Muscle.

Set the alignment parameters to the values you wish or we can go with the default values. Click Compute/OK

Depending on the length and number of sequences you may see a progress bar while the alignment is running.

Once the alignment is done, we can export them to MEGA or Fasta format for analysis.

8.3.5: MAFFT

It is a MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments

Multiple alignment program for amino acid or nucleotide sequences

MAFFT is a multiple sequence alignment program for unix-like operating systems.

It offers a range of multiple alignment methods, L-INS-i, FFT-NS-2, *etc.*

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

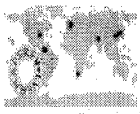
IMD

Download version
- Mac OS X
- Windows
- Linux
- Source

Online version
- Alignment
- mafft --add
- Merge
- Phylogeny
- Rough tree

Merits / limitations
Algorithms
Tips
Benchmarks
Feedback

Follow



To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

The Q-INS-i option (for RNA) did not work because of a bug in version 7.489, Oct.23. Fixed in version 7.490, Oct.30, 2021.

A compatibility problem of MSAViewer with Chrome 92 and Firefox 90 has been fixed, Aug. 5, 2021.

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

Multiple sequence alignment and NJ / UPGMA phylogeny

Input

Paste protein or DNA sequences in fasta format. Example

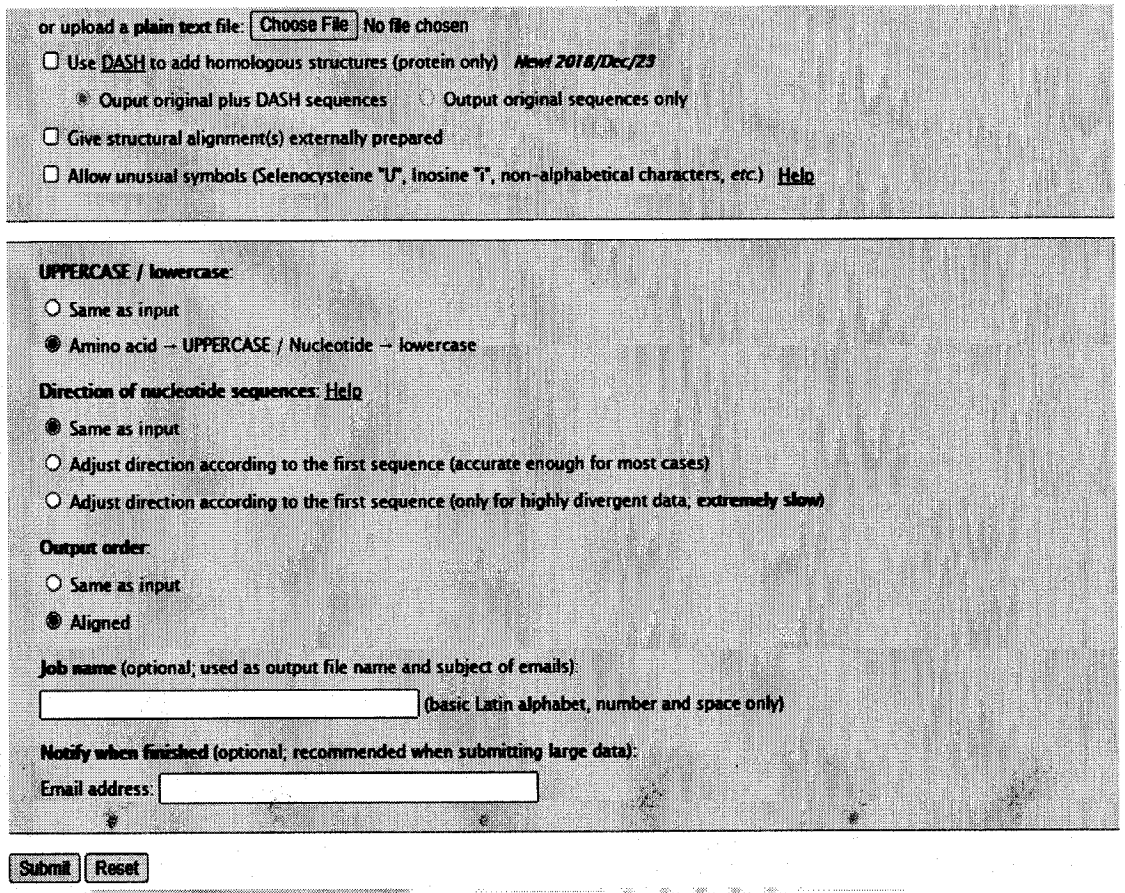
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Sikar) Rajasthan



LAKSHMANGARH • RAJASTHAN

Merits

- (Accuracy): L-INS-i is one of the most accurate multiple sequence alignment methods currently available. L-INS-i is in particular suitable to align 10-100 protein sequences
- (Scalability): FFT-NS-2 and other progressive methods used in MAFFT can align many long DNA/protein sequences
- The scoring system was designed to allow large gaps.

8.3.6: MUSCLE

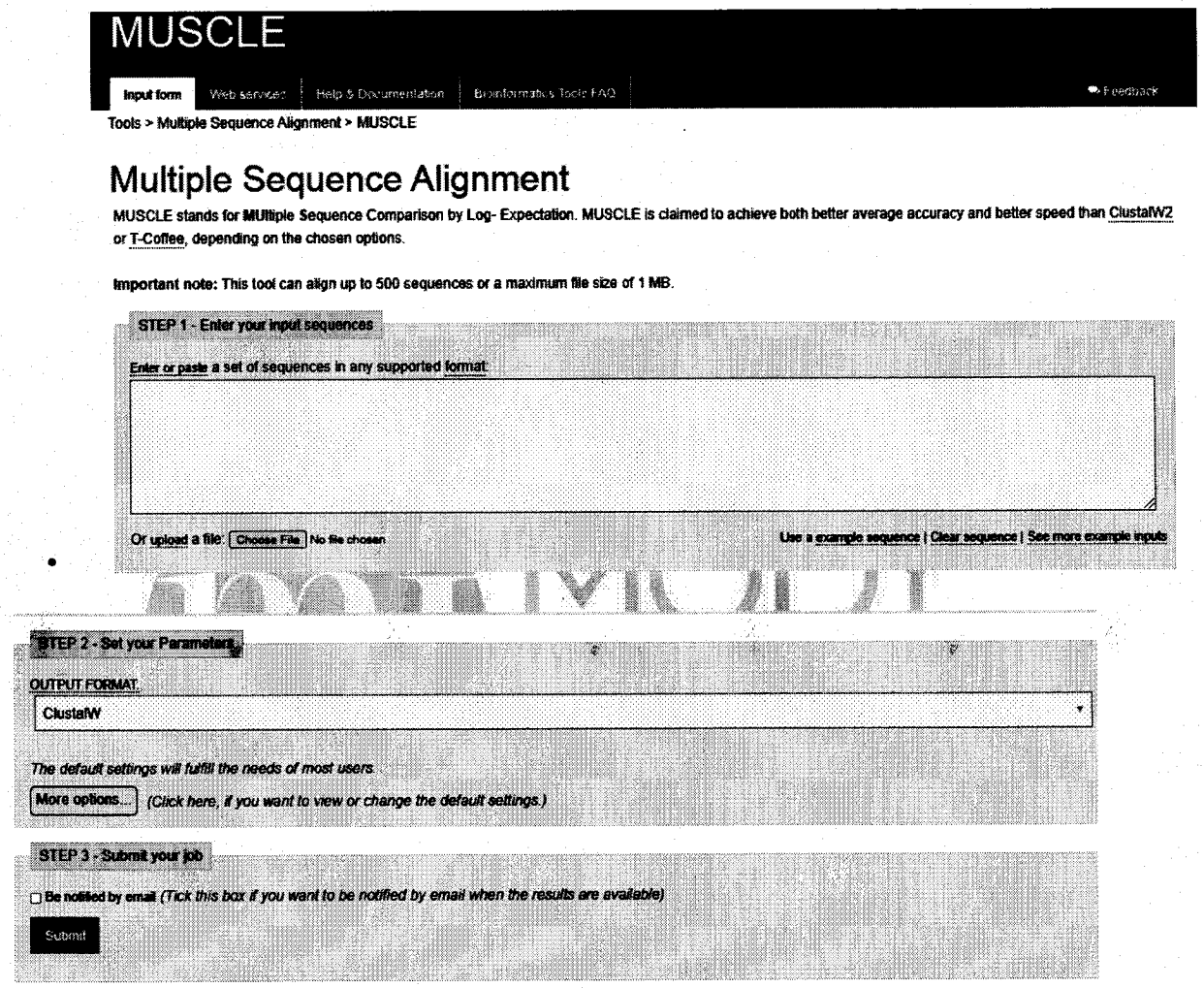
MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation.

MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee,

28/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Sikar Distt.)

This tool can align up to 500 sequences or a maximum file size of 1 MB in a second.

MUSCLE enables high-throughput applications to achieve average accuracy



8.3.7: KALIGN

Very fast MSA tool that concentrates on local regions..

It is a fast and robust alignment method, especially well suited for the increasingly important task of aligning large numbers of sequences.

Suitable for large alignments

28/12/2021
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Kalign

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback

Tools > Multiple Sequence Alignment > Kalign

Multiple Sequence Alignment

A fast and accurate multiple sequence alignment algorithm.

Important note: This tool can align up to 2000 sequences or a maximum file size of 2 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your Parameters

OUTPUT FORMAT

ClustalW

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

28/12/24

REGISTRAR

Mody University of Science and Technology

Lakshmangarh (Distt Sikar)

Important links

https://mafft.cbrc.jp/alignment/server/

https://www.ebi.ac.uk/Tools/msa/emboss_cons/

https://www.ebi.ac.uk/Tools/msa/clustalo/

https://www.genome.jp/tools-bin/clustalw

https://www.megasoftware.net/

https://www.ebi.ac.uk/Tools/msa/kalign/

https://www.ebi.ac.uk/Tools/msa/muscle/

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

9: Basic local alignment search tool (BLAST)

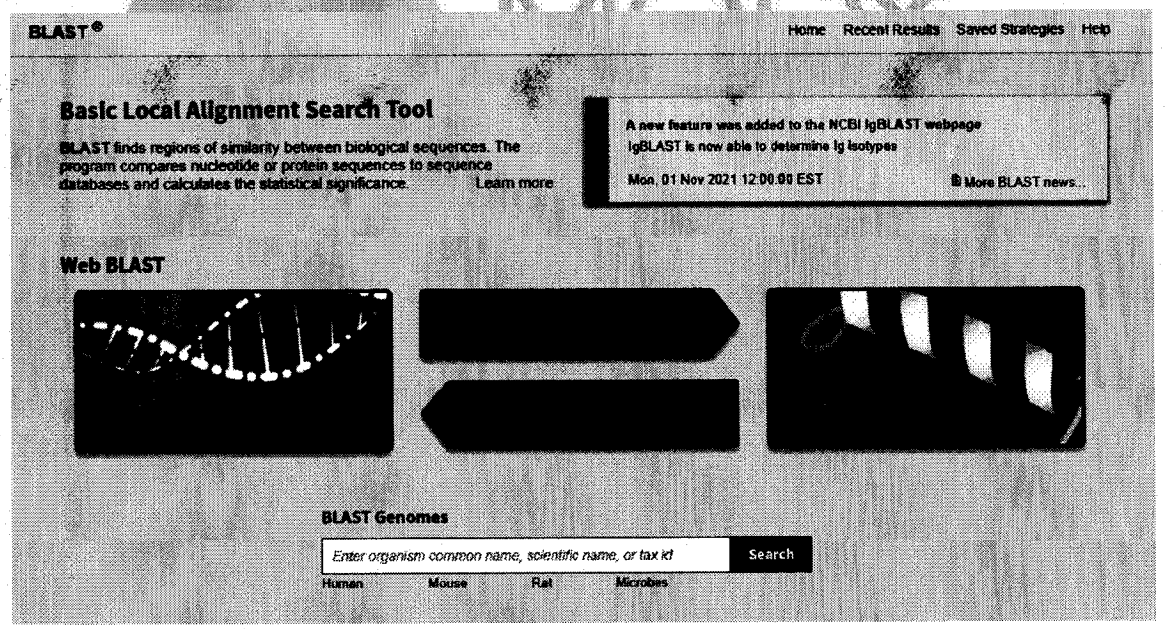
BLAST stands for BASIC LOCAL ALIGNMENT SEARCH TOOL.

It helps in finding the similarities between biological sequences. It is a complete package which helps the researchers in various aspects for comparing and identifying DNA, RNA and protein sequences. BLAST suite offers multiple services like:

Web BLAST offers platforms for searching the protein and nucleotide sequences against a protein or nucleotide database.

BLAST also provides **command line applications** to perform searches locally.

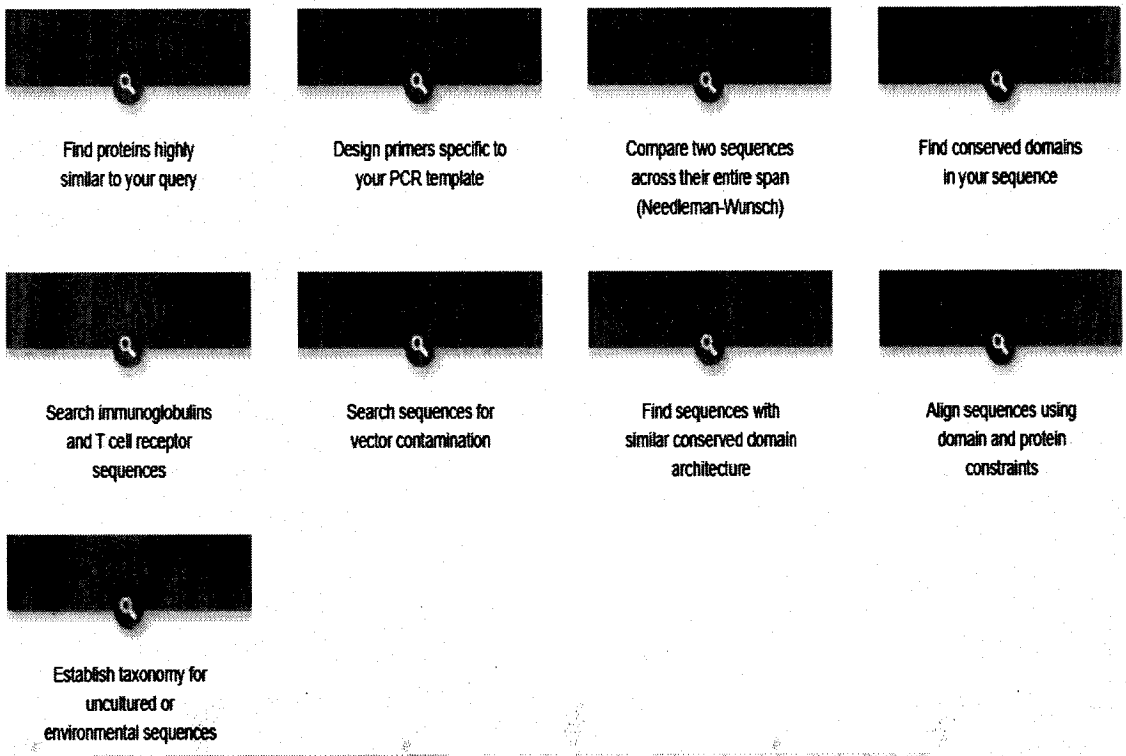
In BLAST suite, one can also find specialized services like SMART BLAST, PRIMER BLAST, GLOBAL ALIGN, CD SEARCH, Ig BLAST, VECSCREEN, CDART, MULTIPLE ALIGN and MOLE BLAST.



28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Specialized searches



UNIVERSITY
LAKSHMANGARH • RAJASTHAN

SMART BLAST

It is a new tool of NCBI that is used for sequence analysis tasks like finding a candidate protein name for a sequence, locating regions of high sequence conservation, or identifying regions covered by database sequences but missing from the query.

To perform all such tasks, it takes much less time as compared to the protein blast.

PRIMER BLAST

It helps in designing the primers that are specific for the PCR template of our choice.

GLOBAL ALIGN

This helps in aligning two sequences globally i.e end to end alignment is done.

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

This is done using Needleman-Wunsch algorithm which we have already discussed in our previous lecture.

CD SEARCH

CD search is Conserved Domain Search service

It is a web-based tool for the detection of structural and functional domains in protein sequences.

Ig BLAST

Is A tool for immunoglobulin (IG) and T cell receptor (TR) Variable domain sequences

With this tool, we can view the matches to the germline V, D and J genes, we can find details at rearrangement junctions, we can also find the delineation of immunoglobulin Variable domain framework regions and complementarity determining regions.

VECSCREEN

VecScreen is a system that quickly finds segments of a nucleic acid sequence that may be of vector origin.

It is recommended to researchers as its helps in identifying and removing any segments of vector origin before they analyze or submit sequences

CDART

Finds sequences with similar conserved domain architecture

MULTIPLE ALIGN

It is a constraint based multiple alignment tool.

It finds a collection of pairwise constraints derived from conserved domain database, protein motif database, and sequence similarity, using RPS-BLAST, BLASTP, and PHI-BLAST

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

22/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

MOLE BLAST

MOLE-BLAST is an experimental tool that helps taxonomists find closest database neighbors of submitted query sequences

It computes a multiple sequence alignment (MSA) between the query sequences along with their top BLAST database hits, and generates a phylogenetic tree.

The BLAST suite also offers downloadable link for BLAST through which one can get a standalone version of BLAST along with the databases.

WEB BLAST

WEB BLAST is the most commonly used section of BLAST which offers multiple versions to compare all combinations of nucleotide or protein queries against a nucleotide or protein database.

The versions are namely nucleotide BLAST or blastn, Protein BLAST or blastp, translated BLASTs blastx, tblastn and tblastx

MODY UNIVERSITY
LAKSHMANGARH • RAJASTHAN

Nucleotide BLAST or blastn

BLASTN program searches nucleotide databases using a nucleotide query.

It is a general purpose nucleotide search and alignment program that can be used to align tRNA or rRNA sequences as well as mRNA or genomic DNA sequences containing a mix of coding and noncoding regions.

Protein BLAST or blastp

Protein-to-protein sequence searches are performed using blastp or protein blast.

The default wordsize for a blastp search is three.

During a "blastp" search, low-complexity regions of the query sequence are filtered to reduce the construction of spurious alignments and enhance search speed.

28/12/2021

REGISTRAR
Moody University of Science and Technology
Lakshmangarh (Distt Sikar)

Translated BLAST

Translated BLAST searches use a genetic code to translate either the "query," database "subjects," or both, into protein sequences, which are then aligned as in "blastp.

There are three varieties of translated BLAST search;

- "tblastn,"
- "blastx," and
- "tblastx."

Blastx

In this type of translated BLAST, a nucleotide sequence query is translated in six reading frames, and the resulting six-protein sequences are compared, in turn, to those in a protein sequence database.

Tblastn

In this a protein sequence query is compared to the six-frame translations of the sequences in a nucleotide database.

As we are searching translated sequences in nucleotide database, we call it T blast N

Tblastx

In this BOTH, the "query" and database "subject" nucleotide sequences are translated in six reading frames, after which 36 protein "blastp" comparisons are made.

Six-frame translations for query sequence and six-frame translations for subject nucleotide sequence

In web blast, we also find BLAST GENOMES which is a graphical tool for simplifying BLAST searches against complete and unfinished genome sequences.

MODY UNIVERSITY LAKSHMANGARH • RAJASTHAN

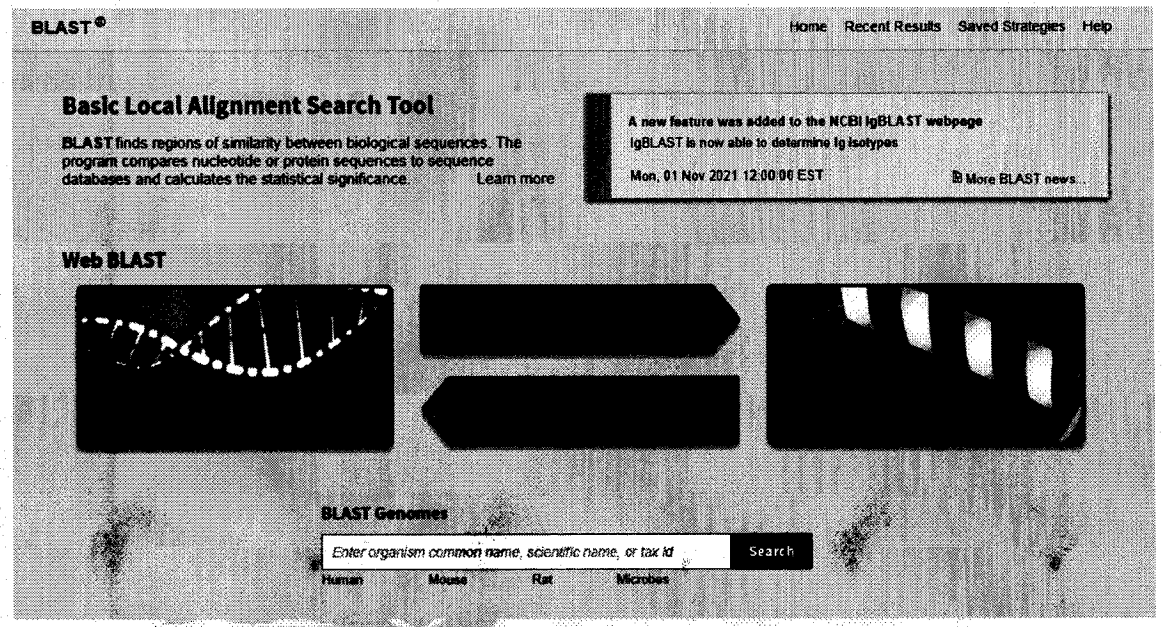
28/12/2021

REGISTRAR

Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

This tool is used to compare the query sequence against a virtual database of DNA sequences or protein sequences from a selected group of organisms with finished or unfinished genomes.

9.1: BLAST software



Recent Results which lists a set of links to recently completed search results.

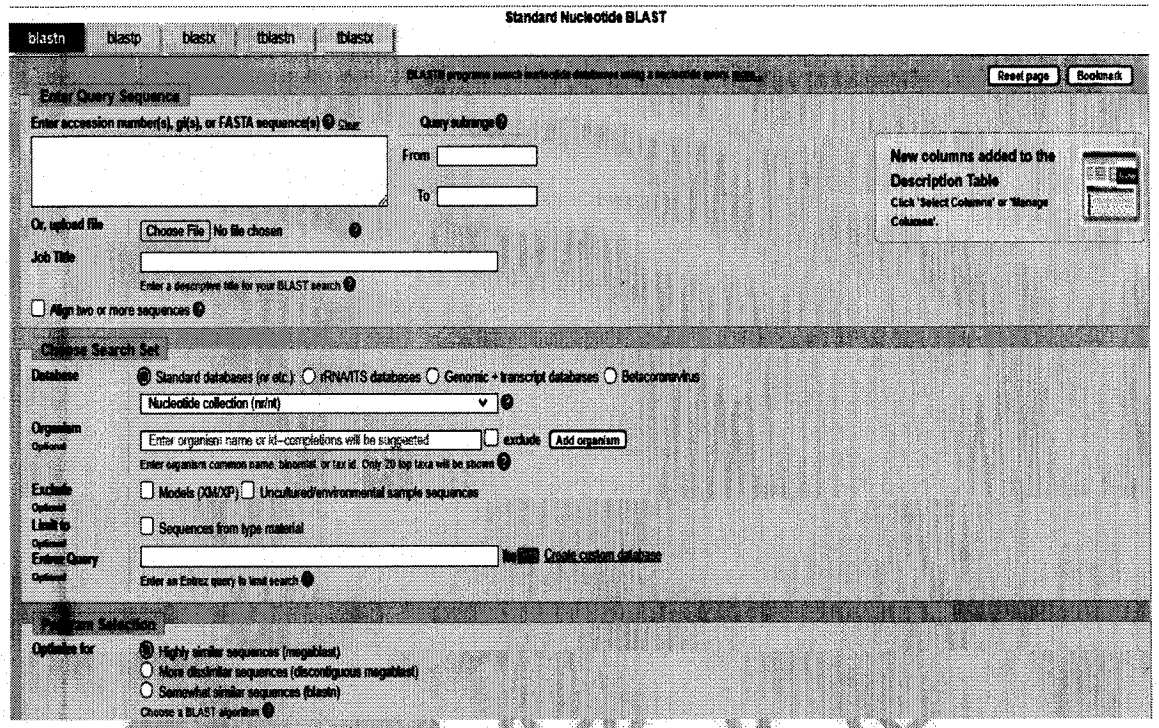
We have "**saved strategies**" which displays a list of search strategies which we have saved earlier. On home page, we can find links to various types of services provided in the BLAST suite.

9.1.1: Nucleotide BLAST

On the homepage of blast, click on nucleotide blast

LAKSHMANGARH • RAJASTHAN

20/12/2024
REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311



Here on the first section, we provide the query sequence.

We can provide query sequence, either by entering the accession number(s) or by providing the gi number(s), or by providing the FASTA sequence(s).

We can also upload a file from our local disk. The file may contain a single sequence or a list of sequences and the data can be either a list of database accession numbers, NCBI gi numbers, or sequences in FASTA format.

Now here Enter a descriptive title for your BLAST search

We can also align two or more sequences. For this we can Enter one or more queries in the top text box and one or more subject sequences in the lower text box and click on the BLAST button at the bottom.

Now the next segment is to **Choose Search Set**

From here we can choose the database against which we want to search our query sequence.

28/12/2021

REGISTRAR

Mody University of Science and Technology
Lakshmangarh (Dist. Sikar)

We can search against standard databases or rRNA/ITS databases or genomic+ transcript databases and now also against beta corona virus genbank .

Then we can choose the list of organisms that we want to exclude from our search.

This is optional. Like wise we have few more options available to restrict our search like we can exclude the model sequences or uncultured experimental sequences.

Now next we can choose the type of the program we want to run. Whether we want to run megablast or discontinuous megablast or blastn. By choosing the program, we are selecting the algorithm we want to run for our similarity search. Before hitting the BLAST button lets have a look at the additional parameters. We can leave these parameters with their default values.

So here, in parameter section, we can choose the maximum no.of target sequences, the default value of which is 100

Short queries, If the entered sequence is a short query sequence then the algorithm will automatically adjust word size and other parameters to improve results for short queries.

Expected threshold, the default value of which is 0.05. This setting specifies the statistical significance threshold for reporting matches against database sequences

In the same way we have other parameters like word size, match mismatch score, gap cost and we have certain filters and masks.

Now once we have adjusted all the parameters and settings as per our requirement, we will hit the BLAST button

Now we are moving towards the results page. BLAST takes few seconds to generate the results page

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Format Request Status

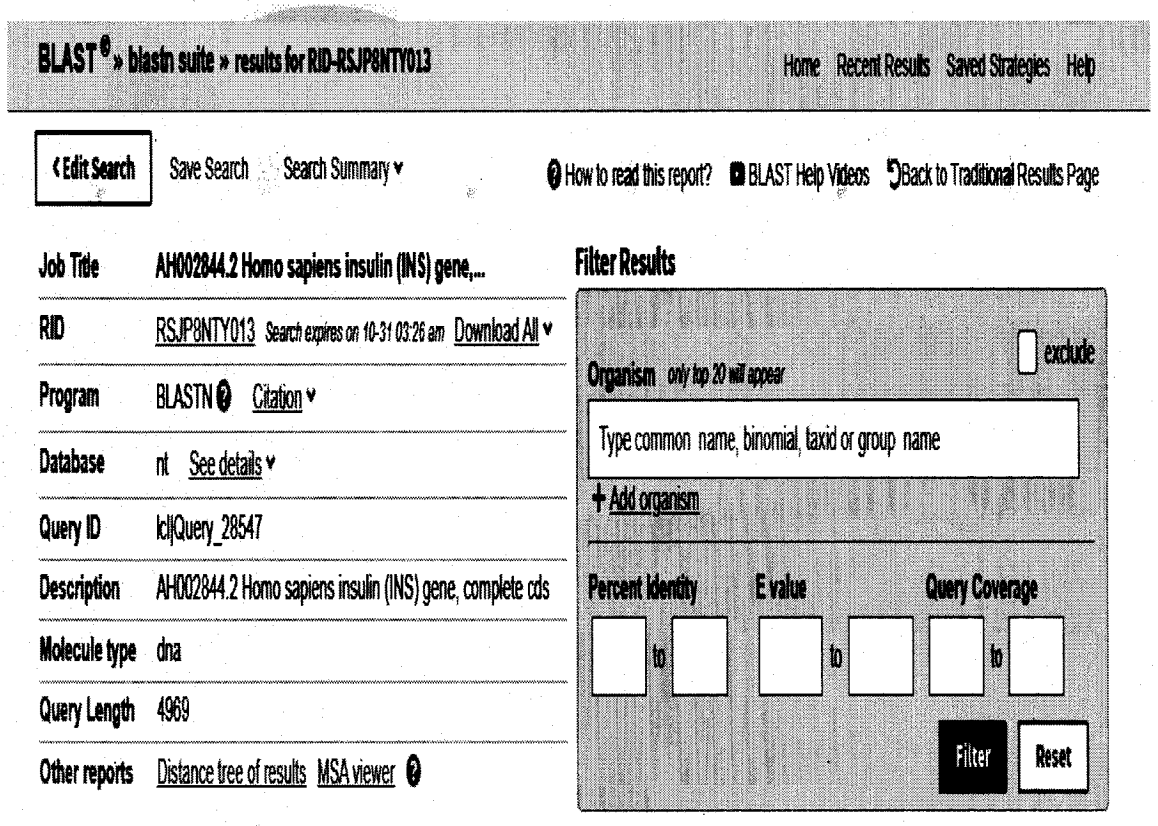
Formatting options

Job Title: AH002844.2 Homo sapiens insulin (INS) gene,...

Request ID	RSJP8NTY013
Status	Searching
Submitted at	Fri Oct 29 15:26:00 2021
Current time	Fri Oct 29 15:26:15 2021
Time since submission	00:00:14

This page will be automatically updated in 2 seconds

The results page of BLAST



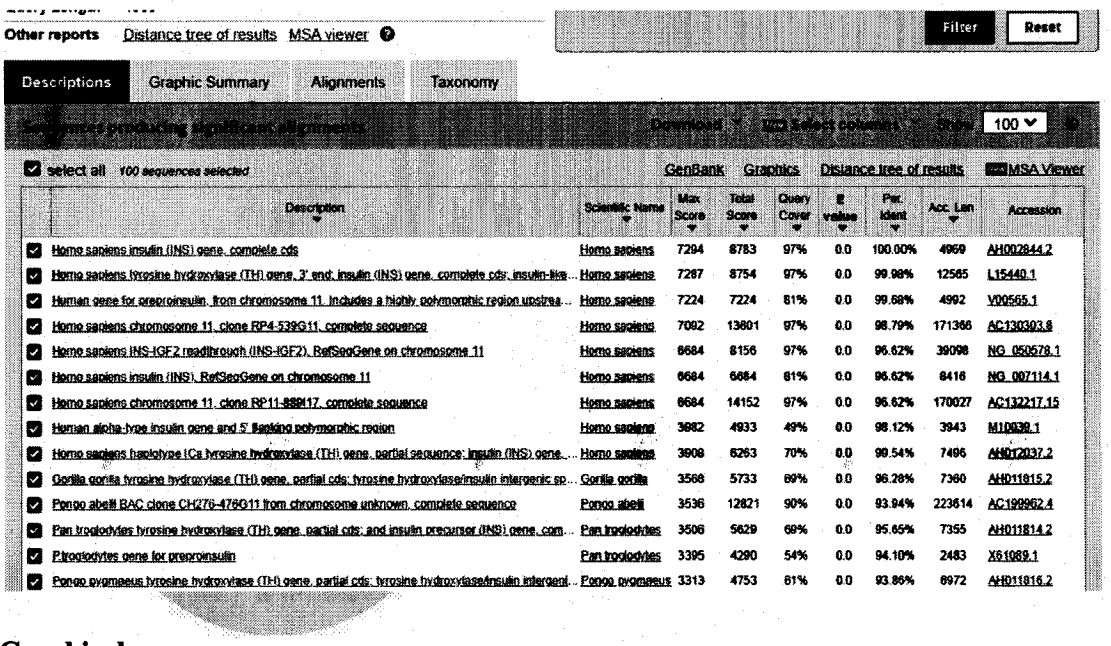
Job Title shows the first query's sequence id by default

RID shows the assigned unique identifier to this search for sharing and problem reporting

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Program tells the search conducted. Database names the target database searched. Query ID is the query's sequence id for which the result is shown. Description is the title of the query sequence from its FASTA defline. Molecule type shows the type of the sequence. Query length is the length of the current query. From here we can also refine our results. The Descriptions tab contains a summary table of hits found by BLAST.

Default sorting is by E value column



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens insulin (INS) gene, complete cds	Homo sapiens	7284	8783	97%	0.0	100.00%	4969	AH002844.2
Homo sapiens tyrosine hydroxylase (TH) gene, 3' end; insulin (INS) gene, complete cds; insulin-like ...	Homo sapiens	7287	8754	97%	0.0	99.98%	12565	L15440.1
Human gene for preproinsulin, from chromosome 11. Includes a highly polymorphic region upstrea ...	Homo sapiens	7224	7224	81%	0.0	99.68%	4992	V00565.1
Homo sapiens chromosome 11, clone RP4-539G11, complete sequence	Homo sapiens	7092	13801	97%	0.0	98.79%	171366	AC130393.8
Homo sapiens INS-IGF2 readthrough (INS-IGF2), RefSeqGene on chromosome 11	Homo sapiens	6684	8156	97%	0.0	96.62%	39098	NG_050578.1
Homo sapiens insulin (INS), RefSeqGene on chromosome 11	Homo sapiens	6684	6684	81%	0.0	96.62%	8416	NG_007114.1
Homo sapiens chromosome 11, clone RP11-889I17, complete sequence	Homo sapiens	6684	14152	97%	0.0	96.62%	170027	AC132217.15
Human alpha-type insulin gene and 5' flanking polymorphic region	Homo sapiens	3982	4933	49%	0.0	98.12%	3943	M10039.1
Homo sapiens haplotype ICa tyrosine hydroxylase (TH) gene, partial sequence; insulin (INS) gene, ...	Homo sapiens	3908	6263	70%	0.0	99.54%	7496	AH012037.2
Gorilla gorilla tyrosine hydroxylase (TH) gene, partial cds; tyrosine hydroxylase/insulin intergenic sp...	Gorilla gorilla	3568	5733	69%	0.0	96.26%	7360	AH011815.2
Pongo abelii BAC clone CH276-476G11 from chromosome unknown, complete sequence	Pongo abelii	3536	12821	90%	0.0	93.94%	223614	AC199962.4
Pan troglodytes tyrosine hydroxylase (TH) gene, partial cds; and insulin precursor (INS) gene, com...	Pan troglodytes	3508	5629	69%	0.0	95.65%	7355	AH011814.2
Ptroglodytes gene for preproinsulin	Pan troglodytes	3395	4290	54%	0.0	94.10%	2483	X61089.1
Pongo pygmaeus tyrosine hydroxylase (TH) gene, partial cds; tyrosine hydroxylase/insulin intergeni...	Pongo pygmaeus	3313	4753	81%	0.0	93.86%	8972	AH011816.2

Graphical summary:

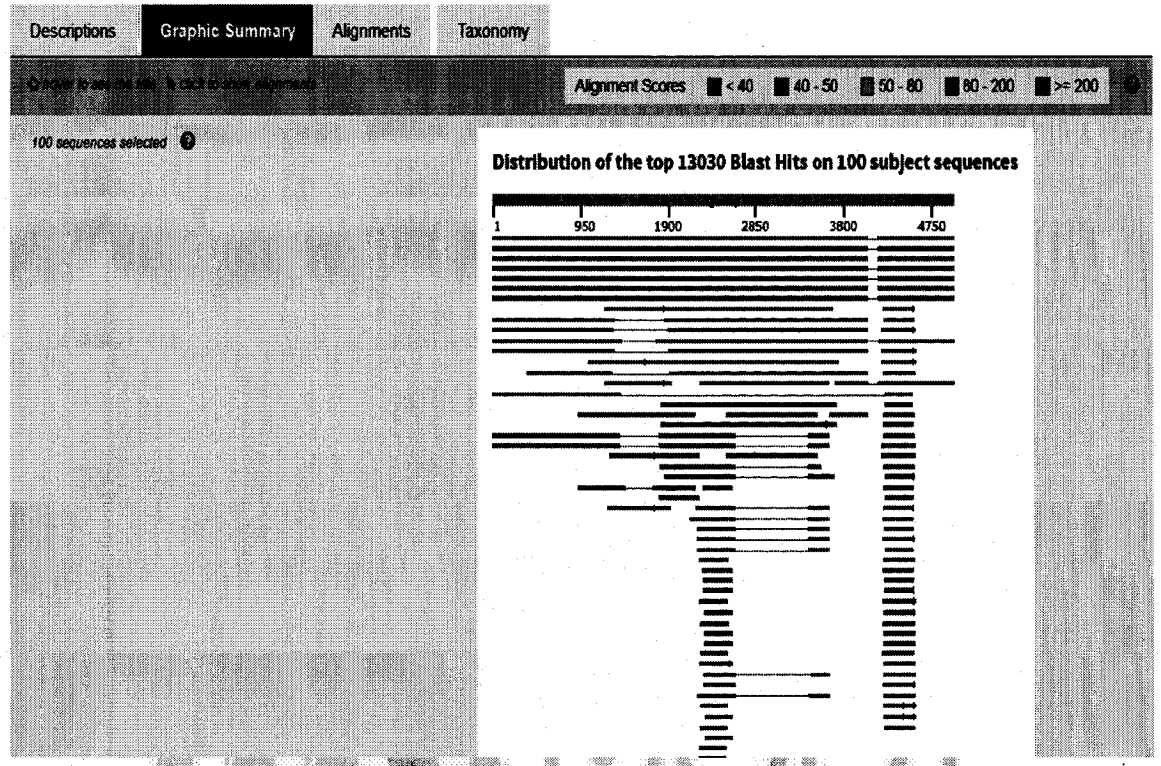
Here you can see various lines in different colours. These colours are on the basis of the alignment scores.

red represents the highest alignment scores

Generally, the higher the alignment score, the more significant the hit.

28/12/2024

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311



The Alignments section contains the detailed pairwise alignments between query and database sequences.

It makes alternative displays of the alignment results readily accessible through a simple pull-down menu.

The sequence alignments show us how well our query sequence matches with the subject sequence in the database.

Each alignment block begins with a summary that includes the Max score and Expect value (the statistical significance of the alignment), sequence identity (number of identical bases between the query and the subject sequence), the number of gaps in the alignment, and the orientation of the query sequence relative to the subject sequence

28/12/2015

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

Descriptions | Graphic Summary | Alignments | Taxonomy

Alignment view: Pairwise | CDS feature | Restore defaults

100 sequences selected

Download | GenBank Graphics | Sort by: E value

Homo sapiens insulin (INS) gene, complete cds

Sequence ID: AH002844.2 **Length:** 4969 **Number of Matches:** 2

Range 1: 1 to 4044 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
7294 bits(8088)	0.0	4044/4044(100%)	0/4044(0%)	Plus/Plus

```
Query  1    CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCG  60

Query  61   GCCAGGGTGTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GCCAGGGTGTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  120

Query  121  TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGC  180

Query  181  TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACTCCTACTCATCCACGA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACTCCTACTCATCCACGA  240

Query  241  AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACCTTGT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACCTTGT  300

Query  301  TAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTGAAGGGCTGTGAGCCCCCAGG  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  TAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTGAAGGGCTGTGAGCCCCCAGG  360
```

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)

Download ▾ GenBank Graphics

Pongo pygmaeus insulin gene, partial cds

Sequence ID: AY092024.1 **Length:** 972 **Number of Matches:** 1

Range 1: 1 to 972 GenBank Graphics ▼ Next Match ▲ Previous Matc

Score	Expect	Identities	Gaps	Strand
1481 bits(1642)	0.0	922/989(93%)	24/989(2%)	Plus/Plus

```
Query  2513  GTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTA  2572
             ||||||||| ||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GTGCGGCTCCCACCTGGTGGAAGCTCTCTACCTAGTGTGTGGGGAACGAGGCTTCTTCTA  60

Query  2573  CACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCATTGCTG  2632
              ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct  61    CACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCCTTGCTG  120

Query  2633  CCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTCCCACCCAGCATGGGCAGAAGG  2692
              ||||||| |||||||||| |||||||| ||| ||||||||||||||||||||||||||
Sbjct  121   CCCCTGGTCGCCCCCAGCCATCCCCTGCTTCTGCCGCTCCCACCCAGCATGGGCAGAAGG  180

Query  2693  GGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAA-----GTT  2747
              ||||||||||||||||||||||||||||||||||||||||||||||||||     |||
Sbjct  181   GGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGTACTTTTTTAAAAAGAAATGAAGTT  240

Query  2748  CTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGA  2807
              |||||||||| |||||||||||||||| |||||||||||| |||||||
Sbjct  241   CTCTTGGTCACATCCTAAAAGTGACCAGTTCCCTGTGGCCCATTCAGAATG---------  291

Query  2808  GGACGGTGTTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCC  2867
              ||    |||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  292   ----GG----GGCTTCAGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCC  343

Query  2868  ACTCGCCCCTCAAACAAATGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAG  2927
              |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||
Sbjct  344   ACTCGCCCCTCAAACAAATGCCCCACAGCCCATTTCTCCACCCTCATTTGATGACCGCAG  403

Query  2928  ATTCAAGTGTTTTGTTAAGTA-AAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGC  2986
              |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||
Sbjct  404   ATTCAAGTGTTTTGTTAAGTACAAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGC  463

Query  2987  TGCCTGCCTCTGGGCGAACACCCCATCACGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAG  3046
              |||||||||||||||||| |||||||||||| ||||||| | |||| ||||||||||
Sbjct  464   TGCCTGCCTCTGGGCGAACGCCCCATCACGCCCTGAGGAGGACATGGCCGCCTGCCTGAG  523
```

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Dist. Sikar) 332311

Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	D
root			106	
. Simiiformes	primates		99	
. . Catarrhini	primates		98	
. . . Hominidae	primates		86	
. . . . Homininae	primates		82	
. . . . . Homo sapiens	primates	7294	75	Homo sapiens hits
. . . . . Gorilla gorilla	primates	3568	2	Gorilla gorilla hits
. . . . . Pan troglodytes	primates	3506	4	Pan troglodytes hits
. . . . . Pan paniscus	primates	1324	1	Pan paniscus hits
. . . . Pongo abelii	primates	3536	2	Pongo abelii hits
. . . . Pongo pygmaeus	primates	3313	2	Pongo pygmaeus hits
. . . Chlorocebus aethiops	primates	2338	1	Chlorocebus aethiops hits
. . . Chlorocebus sabaeus	primates	1760	3	Chlorocebus sabaeus hits
. . . Papio anubis	primates	618	1	Papio anubis hits
. . . Theropithecus gelada	primates	614	3	Theropithecus gelada hits
. . . Piliocolobus tephrosceles	primates	589	1	Piliocolobus tephrosceles hits
. . . Rhinopithecus bieti	primates	532	1	Rhinopithecus bieti hits
. . . Macaca mulatta	primates	426	2	Macaca mulatta hits
. . Aotus trivirgatus	primates	1763	1	Aotus trivirgatus hits
. Expression vector pSP20	other sequences	1299	1	Expression vector pSP20 hits
. eukaryotic synthetic construct	other sequences	431	6	eukaryotic synthetic construct hits

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar) 332311

References

- https://ncbiinsights.ncbi.nlm.nih.gov/2013/03/19/blastdbinfo-api-access-to-a-database-of-blast-databases/
- Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes, FEMS Microbiol Lett . 2002 Nov 5;216(2):133-8. doi: 10.1111/j.1574-6968.2002.tb11426.x.
- https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

28/12/2021

REGISTRAR
Mody University of Science and Technology
Lakshmangarh (Distt Sikar)