Biological Databases and Tools

1. Introduction

Biological databases are structured repositories that store, organize, and provide access to biological data such as DNA sequences, protein structures, gene expression, metabolic pathways, and molecular interactions. These databases are essential for bioinformatics research, enabling data retrieval, analysis, and visualization.

With the advancement of high-throughput sequencing and computational biology, various bioinformatics tools have been developed to analyze biological data. These tools assist in tasks such as sequence alignment, structure prediction, and molecular modeling.

2. Types of Biological Databases

Biological databases can be categorized based on their content and purpose:

A. Based on Data Type

1. **Nucleotide Sequence Databases** – Store DNA and RNA sequences.
    - **Examples:** GenBank, ENA (European Nucleotide Archive), DDBJ (DNA Data Bank of Japan)
2. **Protein Sequence Databases** – Contain amino acid sequences and functional annotations.
    - **Examples:** UniProt (Universal Protein Resource), Swiss-Prot (Manually Curated Protein Database), PIR (Protein Information Resource)
3. **Genomic Databases** – Store whole genome sequences.
    - **Examples:** Ensembl Genome Browser, UCSC Genome Browser, NCBI (National Center for Biotechnology Information) Genome
4. **Protein Structure Databases** – Store 3D structures of proteins.
    - **Examples:** PDB – Protein Data Bank, SCOP – Structural Classification of Proteins, CATH – Class, Architecture, Topology, Homology
5. **Pathway and Interaction Databases** – Contain biochemical pathways and protein-protein interactions.
    - **Examples:** KEGG – Kyoto Encyclopedia of Genes and Genomes, Reactome – Reactome Pathway Database, BioGRID – Biological General Repository for Interaction Datasets
6. **Expression Databases** – Store gene expression profiles.
    - **Examples:** GEO – Gene Expression Omnibus, ArrayExpress – ArrayExpress Archive of Functional Genomics Data: A repository for **high-throughput gene expression data** (microarrays & sequencing)

7. **Variant and Mutation Databases** – Store genetic variants and disease-associated mutations.
    - **Examples:** dbSNP – Database of Single Nucleotide Polymorphisms, ClinVar – Clinical Variation Database, HGMD – Human Gene Mutation Database

B. Based on Data Processing

1. **Primary Databases**
    - Store raw experimental data submitted by researchers.
    - **Examples:** GenBank, UniProt, PDB
2. **Secondary Databases**
    - Contain curated, processed, and analyzed data derived from primary databases.
    - **Examples:** Swiss-Prot (annotated proteins), CATH (protein structures), KEGG (pathways)

3. Key Biological Databases

A. PDB (Protein Data Bank)

- Stores 3D structures of biomolecules, including proteins, DNA, and RNA.
- Contains structural information derived from X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy.
- **File Formats:**
    - **PDB format (.pdb):** Standard text format for 3D molecular structures.
    - **mmCIF (.cif):** Modern format with more detailed metadata.

B. NCBI (National Center for Biotechnology Information)

- Provides access to multiple biological databases, including:
    - **GenBank:** Nucleotide sequences.
    - **PubMed:** Biomedical literature.
    - **GEO:** Gene expression data.
    - **dbSNP:** Genetic variations.
- **File Formats:**
    - **FASTA (.fasta):** Simple text format for nucleotide/protein sequences.
    - **GenBank (.gb, .gbk):** Rich annotation format for sequences.

4. Sequence Retrieval and Manipulation

A. Sequence Retrieval

- **Tools:**
    - **NCBI Entrez:** Search and retrieve sequences from GenBank.
    - **UniProt:** Fetch protein sequences.

- **EBI-ENA:** Access nucleotide sequences from the European Nucleotide Archive.

- **Methods:**
  - Search by gene name, accession number, or keyword.
  - Use BLAST to find similar sequences.

B. Sequence Manipulation

- **Tools:**
  - **FASTA Tools:** Convert sequences to different formats.
  - **SeqTK:** Command-line tool for sequence processing.
  - **EMBOSS (European Molecular Biology Open Software Suite):** Provides various bioinformatics tools for sequence handling.

5. Primer Design

- **Primers** are short sequences used in PCR (Polymerase Chain Reaction) to amplify DNA.
- **Tools for Primer Design:**
  - **Primer3:** Widely used for designing PCR primers.
  - **OligoAnalyzer (IDT):** Checks primer properties (melting temperature, GC content, secondary structures).
  - **NCBI Primer-BLAST:** Designs primers while checking specificity against genomic databases.

6. Restriction Mapping

- **Restriction mapping** identifies restriction enzyme cut sites in a DNA sequence.
- **Tools:**
  - **NEB Cutter:** Finds restriction sites and predicts digestion patterns.
  - **REBASE:** A comprehensive database of restriction enzymes.
  - **EMBOSS Restriction Mapper:** Analyzes restriction sites in DNA sequences.

7. ORF (Open Reading Frame) Finding

- **ORF Finding** identifies regions in a DNA sequence that can potentially code for proteins.
- **Tools:**
  - **ORF Finder (NCBI):** Identifies all possible ORFs in a given sequence.
  - **EMBOSS getorf:** Extracts ORFs from a DNA sequence.
  - **GeneMark:** Predicts coding regions in prokaryotic and eukaryotic genomes.

8. EMBOSS (European Molecular Biology Open Software Suite)

- A collection of bioinformatics tools for sequence analysis, including:
    - **Water/Needle:** Sequence alignment.
    - **Transeq:** Translates nucleotide sequences into proteins.
    - **Patmatmotifs:** Identifies sequence motifs.

9. Molecular Visualization

- **Molecular visualization tools** help researchers analyze protein structures and interactions.
- **Tools:**
    - **PyMOL:** High-quality 3D visualization of molecular structures.
    - **Chimera:** Advanced molecular modeling and visualization.
    - **Jmol:** Web-based interactive molecular viewer.

10. Sequence Analysis

Sequence analysis involves various computational techniques to compare, annotate, and interpret sequences.

A. **Sequence Alignment:**
    - **BLAST (NCBI):** Finds similar sequences using local alignment.
    - **Clustal Omega:** Performs multiple sequence alignments.
    - **MAFFT:** Fast multiple sequence alignment tool.

These are three widely used **sequence alignment tools** in bioinformatics. They help in comparing DNA, RNA, or protein sequences to identify similarities, evolutionary relationships, and functional annotations.

1. BLAST (Basic Local Alignment Search Tool)

- **Developed by:** NCBI (National Center for Biotechnology Information)
- **Purpose:** Finds regions of similarity between biological sequences.
- **Type:** Local sequence alignment.
- **Input:** DNA, RNA, or protein sequences.
- **Output:** Aligned sequences with similarity scores and statistical significance (E-value).

Types of BLAST:

1. **BLASTN** – Compares a nucleotide sequence against a nucleotide database.
2. **BLASTP** – Compares a protein sequence against a protein database.
3. **BLASTX** – Translates a nucleotide sequence into protein and compares it against a protein database.
4. **TBLASTN** – Compares a protein sequence against a translated nucleotide database.

5. **TBLASTX** – Translates both query and database sequences and performs a comparison.

Features:

- **Fast local alignment** (compared to global alignment algorithms like Needleman-Wunsch).
- Returns **E-value** (expectation value) to determine significance.
- Provides **high-scoring segment pairs (HSPs)**, showing the best local alignments.

Applications:

- Identifying **homologous genes** in different species.
- Locating **mutations and variations** in DNA sequences.
- Function prediction for unknown sequences by similarity comparison.

Where to Use It?

- **NCBI BLAST Web Server:** https://blast.ncbi.nlm.nih.gov/
- **Standalone BLAST:** Available for local analysis with command-line options.

2. Clustal Omega

- **Developed by:** European Bioinformatics Institute (EMBL-EBI).
- **Purpose:** Multiple sequence alignment (MSA) of nucleotide or protein sequences.
- **Type:** Global sequence alignment.
- **Input:** DNA or protein sequences in **FASTA** format.
- **Output:** Aligned sequences with conserved regions highlighted.

Features:

- Uses **progressive alignment algorithms** and **hidden Markov models (HMMs)** for better accuracy.
- Supports **large datasets** with thousands of sequences.
- Provides **phylogenetic tree output** based on aligned sequences.

Applications:

- **Phylogenetic analysis** – Studying evolutionary relationships among species.
- **Conserved region detection** – Identifying functionally important motifs in proteins.
- **Primer design** – Aligning sequences for conserved region identification.

Where to Use It?

- **Web Server:** https://www.ebi.ac.uk/Tools/msa/clustalo/
- **Command-line version:** Available for Linux/macOS for large-scale alignments.

3. MAFFT (Multiple Alignment using Fast Fourier Transform)

- **Developed by:** Kazutaka Katoh & colleagues.

- **Purpose:** Multiple sequence alignment (MSA) with speed and accuracy.
- **Type:** Global sequence alignment.
- **Input:** DNA or protein sequences in **FASTA** format.
- **Output:** Aligned sequences with similarity scores.

Features:

- **Fast and scalable** – Can align **millions of sequences** efficiently.
- Uses **FFT (Fast Fourier Transform)** to speed up alignments.
- Provides different alignment strategies based on **sequence length and complexity**.

Comparison with Clustal Omega:

Feature	**Clustal Omega**	**MAFFT**
Speed	Slower for large datasets	Faster
Scalability	Handles thousands of sequences	Handles millions of sequences
Accuracy	High for moderately sized datasets	High for large and complex datasets
Phylogenetics	Yes (guide trees)	Yes (more accurate trees)

Applications:

- Large-scale **genomic and metagenomic analyses**.
- Alignment of **highly divergent sequences**.
- **Structural and functional analysis** of conserved sequence motifs.

Where to Use It?

- **Web Server:** https://mafft.cbrc.jp/alignment/server/
- **Command-line version:** Available for high-performance computing.

Summary

Tool	**Type**	**Key Use**	**Best For**
BLAST	Local alignment	Finds sequence similarity	Identifying homologous sequences
Clustal Omega	Global alignment	Multiple sequence alignment	Evolutionary and functional studies
MAFFT	Global alignment	Fast multiple sequence alignment	Large-scale and complex datasets