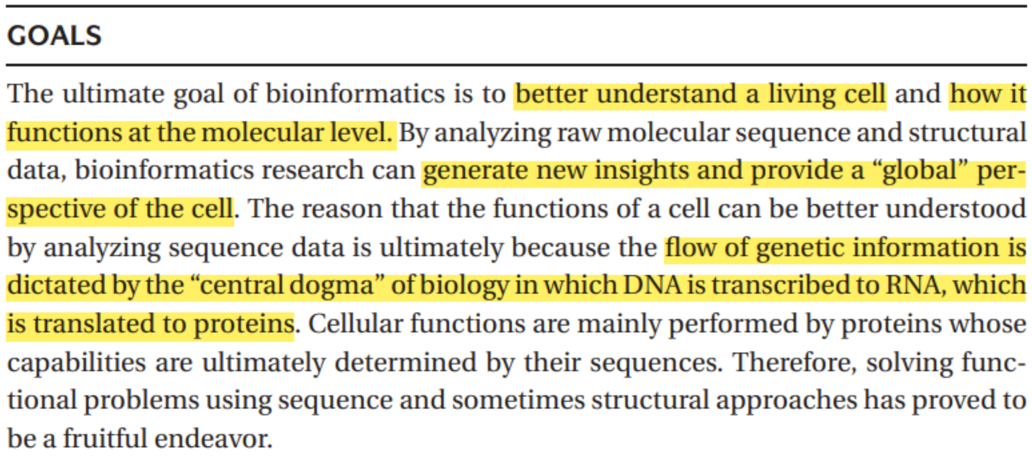
PAPER 1 – UNIT 4
INTRODUCTION TO BIOINFORMATICS AND SEQUENCE ANALYSIS

BIOINFORMATICS

- An interdisciplinary research area at the interface between computer science and biological science.
- Involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA and proteins.
- Limited to sequence, structure and functional analysis of genes and genomes and their corresponding products.

Goals of Bioinformatics –



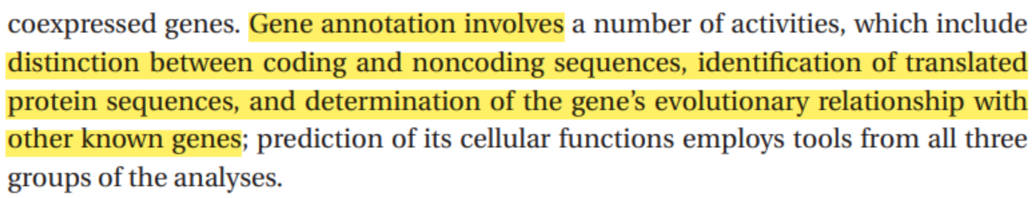
GOALS

The ultimate goal of bioinformatics is to better understand a living cell and how it functions at the molecular level. By analyzing raw molecular sequence and structural data, bioinformatics research can generate new insights and provide a "global" perspective of the cell. The reason that the functions of a cell can be better understood by analyzing sequence data is ultimately because the flow of genetic information is dictated by the "central dogma" of biology in which DNA is transcribed to RNA, which is translated to proteins. Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and sometimes structural approaches has proved to be a fruitful endeavor.

Scope of Bioinformatics –

1. **Development of computational tools and databases –**
   - ✓ Writing software for sequence, structural and functional analysis
   - ✓ Construction and curating of biological databases

2. **Application of these tools and databases in generating biological knowledge to better understand living systems –**
   - **(a) Molecular sequence analysis** – sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, genome assembly and comparison.
   - **(b) Molecular structural analysis** – protein and nucleic acid structure analysis, comparison, classification and prediction.
   - **(c) Molecular functional analysis** – gene expression profiling, protein – protein interaction prediction, protein subcellular localization prediction, metabolic pathway reconstruction and simulation.



coexpressed genes. Gene annotation involves a number of activities, which include distinction between coding and noncoding sequences, identification of translated protein sequences, and determination of the gene's evolutionary relationship with other known genes; prediction of its cellular functions employs tools from all three groups of the analyses.

➢ Applications of Bioinformatics –

1. **Computational studies of protein – ligand interactions –** provides a rational basis for rapid identification of novel leads for synthetic drugs.
2. **Knowledge of 3D structures of proteins –** allows molecules to be designed that are capable of binding to the receptor site of a target protein with great affinity and specificity. **Advantage –** reduces time and cost necessary to develop drugs with higher potency, fewer side effects and less toxicity than using traditional trial – and – error approach.
3. **Forensics** – results from molecular phylogenetic analysis accepted as evidence in criminal courts.
4. **Personalized and customized medicine** – allows a doctor to quickly sequence a patient's genome and easily detect potential harmful mutations ; and to engage in early diagnosis and effective treatment of diseases.
5. **Agriculture** – plant genome databases + gene expression profile analyses → development of new crop varieties that have higher productivity and more resistance to disease.

➢ Limitations of Bioinformatics –

1. Bioinformatics depends on experimental science to produce raw data for analysis.
2. Bioinformatic predictions are not formal proofs of any concepts. They do not replace the traditional experimental research methods of actually testing hypotheses.
3. Quality of bioinformatics predictions depends on – the quality of data and the sophistication of the algorithms being used.
4. Sequence data from high – throughput analysis often contain errors. If the sequences are wrong or annotations incorrect, results from the downstream analysis are misleading as well.
5. Errors in sequence alignments can affect the outcome of structural or phylogenetic analysis.
6. Depending on computing power available –
   (a) Highly accurate but exhaustive algorithms cannot be used.
   (b) Less accurate but faster algorithms have to be used.

DATABASE

- A computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.
- **Composed of –** computer hardware + software – for data management
- **Chief objective –**
    - **(a) Data retrieval –** to organize data in a set of structured records ("entry") to enable easy retrieval of information ("value").
    - **(b) Knowledge discovery –** identification of connections between pieces of information that were not known when the information was first entered.

- **File formats –**

File formats	**Description**
Flat file format	✓ Long text file that contains many entries separated by a **delimiter** (a special character such as a vertical bar |). ✓ Within each entry, various fields are separated by tabs or commas. Each field contains raw values. ✓ Does not contain any hidden instructions for computers to search for specific information / create reports based on certain fields from each record. ✓ **DRAWBACK –** (a) Inefficient process (b) Not manageable for retrieving information from complex data sets (c) Memory – intensive nature of operation
GenBank file format	GenBank – relational database Search output for sequence files – flat files Each field has a unique identifier for easy indexing by computer software – helps in designing effective search strategies Contain 3 sections – **1. Header –** describes the origin of the sequence, identification of the organism, and unique identifiers associated with the record. **(a) LOCUS –** contains a unique database identifier for a sequence location in the database (not a chromosome locus) – followed by – o Sequence length o Molecule type (eg – DNA, RNA) o 3 – letter code for GenBank division (total 17 divisions ; eg – PRI = primate sequences, BCT = bacterial sequences o Date of record – when the record was made public – different from date of submission **(b) DEFINTION –** provides the summary information for the sequence record including – o Name of the sequence o Name and taxonomy of the source organism if known o Whether the sequence is complete or partial

	(c) ACCESSION NUMBER – unique number assigned to a piece of DNA when it was first submitted to GenBank and is permanently associated with that sequence. o 2 formats – 2 letters + 5 digits / 1 letter + 6 digits o New accession number for a nucleotide sequence that has been translated into a protein sequence **(d) VERSION NUMBER + GENE INDEX (gi) NUMBER** – to identify the current version of the sequence o If the sequence annotation is revised at a later date – **(a)** Accession number remains the same **(b)** Version number and gi number incremented A translated protein sequence also has a different gi number from the DNA sequence it is derived from. **(e) ORGANISM** – source of the organism with the scientific name of the species + taxonomic classification + tissue type (sometimes) **(f) REFERENCE** – publication citation related to the sequence entry – author + title information of published work / tentative title of unpublished work **(g) JOURNAL** – citation information linked to PubMed record + date of sequence submission **2. Features** – includes annotation information about the gene and gene product, as well as regions of biological significance reported in the sequence, with identifiers and qualifiers. **(a) SOURCE** – length of the sequence + scientific name of the organism + taxonomy identification number + clone source + tissue type + cell line **(b) GENE** – nucleotide coding sequence + name **(c) CDS (for DNA entries)** – boundaries of the sequence that can be translated into amino acids → for eukaryotic DNA : location of exons and translated proteins sequences **3. Sequence entry** – flat file (format of sequence display can be changed) **(a) ORIGIN** – start of the sequence **(b) BASE COUNT REPORT** – only for DNA entries – numbers of A, G, C and T in the sequence **(c) // - 2 forward slashes** – end of sequence
FASTA file format	✓ Contains plain sequence information ✓ Readable by many bioinformatics analysis programs. ✓ Has a single definition line that begins with a right angle bracket (>) followed by a sequence name ✓ Optional information – gi number or comments – separated from the sequence name by a "|" symbol – ignored by sequence analysis programs ✓ Plain sequence – in standard one – letter symbols → starts in the second line ✓ Each line of sequence data is limited to 60 – 80 characters in width. ✓ Drawback – Much annotation information is lost.

➢ Types of Databases –

Relational Databases	**CRITERIA**	**Object – Oriented Databases**
Set of tables; Each table = **“relation”** – made up of: **(a) Columns** – represent individual fields; indexed according to a common feature called “**attribute”** for cross – referencing in other tables **(b) Rows –** represent values in the fields of records	ORGANIZATION OF DATA	Objects = a unit that combines data and mathematical routines that act on data; Objects linked by a set of pointers ↓ Hence defining predetermined relationships between the objects.
SQL (Structured Query Language)	CREATED USING	**C++**
1. Easy cross – referencing. **2.** Specific information found more easily than flat file formats. **3.** Information combined in one report by selecting linked data items from different tables. **4.** After creation of the original database, a new category can be easily added without modifying all existing tables.	ADVANTAGE	**1.** More flexible. **2.** Data can be structured based on hierarchical relationships. **3.** Simplified programming tasks for multimedia data with complex relationships.
1. Tables used do not describe complex hierarchical relationships between data items.	DRAWBACK	**1.** Lacks the rigorous mathematical foundation of the relational databases. **2.** Risk of misrepresenting some relationships between objects.
Hence **Object – Relational DBMS** created		

➢ BIOLOGICAL DATABASES

✓ Based on their contents – three categories –

1. **Primary databases –** contain original biological data – archives of raw sequence or structural data submitted by the scientific community.

   Eg – GenBank and Protein Data Bank (PDB)

2. **Secondary databases –** contain computationally processed or manually curated information, based on original information from primary databases.

   Eg - Translated protein sequence databases containing functional annotation, SWISS-Prot and Protein Information Resources (PIR).

3. **Specialized databases –** cater to a particular research interest.

   Eg - Flybase, HIV sequence database, and Ribosomal Database Project (specialize in a particular organism or a particular type of data).

1. PRIMARY DATABASES

- Store raw nucleic acid sequence data produced and submitted by researchers worldwide
- 3 major public sequence databases – **GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA Data Bank of Japan (DDBJ)** – closely collaborate and exchange new data daily – together constitute the **International Nucleotide Sequence Database Collaboration (INSDC).**
- Access to the same nucleotide sequence data with a slightly different kind of format to represent the data.
- Data – contributed directly by authors with a minimal level of annotation.
- **PDB –**
  - ✓ Centralized database for three-dimensional structures of biological macromolecules
  - ✓ Archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by X – ray crystallography and NMR
  - ✓ Uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates.
  - ✓ Web interface of PDB – provides viewing tools for simple image manipulation.

2. SECONDARY DATABASES

- contain computationally processed sequence information derived from the primary databases
- **May be either –**

  (a) simple archives of translated sequence data from identified open reading frames in DNA

  (b) provide additional annotation and information related to higher levels of information regarding structure and functions

- **UniProt Database = SWISS – PROT + TrEMBL + PIR**
  - ✓ **SWISS – PROT** – provides detailed sequence annotation that includes structure, function, and protein family assignment
  - ✓ **TrEMBL –** translated nucleic acid sequences stored in the EMBL database

- **Protein annotations include –**
  - ✓ Function, domain structure, catalytic sites, cofactor binding, posttranslational modification, metabolic pathway information, disease association, and similarity with other sequences, cross – referencing links to other online resources of interest

- ✓ Obtained from scientific literature ; entered by database curators.

- o **Advanatages –**
  - **(a)** Very low redundancy
  - **(b)** High level of integration
  - **(c)** Cross – references
  - **(d)** High quality of annotation

- o **Other secondary databases – Pfam and BLOCKS –** contain aligned protein sequence information as well as derived motifs and patterns – used for classification of protein families and inference of protein functions.

3. **SPECIALIZED DATABASES**
   - o Serve a specific research community / focus on a particular organism
   - o Content – sequences or other types of information.
   - o May have unique organizations and additional annotations associated with the sequences.
   - o Examples include –
     - **(a)** taxonomic specific genome databases – Flybase, WormBase, AceDB, and TAIR
     - **(b)** gene expression databases – GenBank EST database and Microarray Gene Expression Database

DISADVANTAGES OF BIOLOGICAL DATABASES

1. **Overreliance –** sequence information and related annotations – without understanding the reliability of the information

2. **High levels of redundancy –** in primary sequence databases – tremendous duplication of information due to –
   - **(a)** repeated submission of identical / overlapping sequences by the same or different authors
   - **(b)** revision of annotations
   - **(c)** dumping of expressed sequence tags (EST) data
   - **(d)** poor database management that fails to detect the redundancy.
     Remedy – NCBI RefSeq (non – redundant database)

3. **Occasional false / incomplete annotations of genes**

4. **Sequencing errors –** frameshifts, cloning vector contaminations

5. **Erroneous annotations –**
   - **(a)** same gene sequence found under different names resulting in multiple entries and confusion about the data.
   - **(b)** unrelated genes bearing the same name
   - **(c)** genuine disagreement between researchers in the field
   - **(d)** imprudent assignment of protein functions by sequence submitters.
   - **(e)** Omissions or mistakes in typing

Sequence Alignment

- ✓ Method to compare 2 or more sequences (DNA / protein) to identify characters that are identical or similar in the sequences.

- ✓ **Why use Sequence comparison ?**

  Biological basis – many genes + proteins → members of families which have a similar biological function / share common evolutionary origin.

  1. To define evolutionary relationships
  2. To identify conserved patterns
  3. To find similar domains that imply similar functions – when dealing with a sequence of unknown function

- ✓ **Why Protein Alignments are MORE INFORMARIVE than DNA Alignments ?**
  1. Changes in DNA sequence (at 3rd position of codon) – does not change the a.a.
  2. Many a.a. share related biophysical properties.
  3. Important relationships between relate but mismatched a.a. – accounted for using scoring systems. DNA sequences = less informative in this regard.
  4. Identification of homologs by protein sequence comparisons – common ancestors over 1 BYA ; DNA sequence comparisons – only 600 MYA.

- ✓ **Why choose DNA / Nucleotide sequences for study ?**
  1. Confirming identity of a DNA sequence in a DB search
  2. Search for polymorphisms
  3. Analyzing identity of a cloned cDNA fragment

➢ IMPORTANT TERMS

1. **Homologs** – 2 genes or proteins are said to be homologous if they –
   - Share a common evolutionary history
   - Similarity attributed to descent from a common ancestor
   - **Qualitative inference –** No degrees of homology ; either homologous or not.
   - Homologous proteins – always share a significantly related 3 – dimensional structure.
   - Eg – myoglobin and β – globin

2. **Orthologs**
   - Homologous sequences in different species that – arose from a common ancestral gene – during **speciation.**
   - May / may not be responsible for a similar biological function
   - Eg – human myoglobin gene and a rat gene

3. **Paralogs**
   - Homologous sequences within a single species that – arose by **gene duplication mechanism.**
   - Have distinct but related functions
   - Eg – human α – 1 globin and α – 2 globin ; human α – 1 globin and β – globin.

4. **Identity**
   - Extent to which 2 a.a. / nucleotide sequences are invariant.
   - Quantitative measure of **relatedness of sequences**

5. **Similarity**
   - Share similar biochemical properties but not identical ; **Conservative Substitutions**
   - Structurally or functionally related
   - Quantitative measure of **relatedness of sequences**
   - Eg – basic a.a. (K, R, H)
   - **% Similarity = sum of both identical and similar matches.**

Pairwise Sequence Alignment

- ✓ Process of aligning 2 sequences to achieve maximal levels of identity (in case of amino acid alignments – maximal levels of conservation).
- ✓ **Heuristic algorithm** – makes approximations of the best solution without exhaustively considering every possible outcome.
- ✓ **Identical / similar characters** → placed in same column.
- ✓ **Non – identical characters** → placed in either in same column as a **mismatch** / opposite a **gap.** Placed in such a way so as to bring as many identical or similar characters as possible together.
- ✓ **Gaps –**
  - **Caused by –** insertions and deletions (residues added / removed)
  - **Represented by –** dashes (-)
  - **Occurrence –** at the middle / ends of the proteins
  - **2 gap penalties in scoring scheme –**
    - (a) creating a gap
    - (b) gap extension penalty – additional residue that a gap extends

- ✓ **Purpose of PSA –**
  1. To assess the degree of similarity + possibility of homology between 2 molecules.
  2. To identify mutations occurred during evolution + have caused divergence of the sequences