

# Introduction to Bioinformatics

## Introduction to Bioinformatics

*An overview of the field, its purpose, core concepts, and typical workflows.*

---

## 1. What Is Bioinformatics?

| Aspect       | Description                                                                                                                                                   |
|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Definition   | The interdisciplinary science that combines biology, computer science, mathematics, and statistics to collect, store, analyze, and interpret biological data. |
| Primary Goal | Transform raw biological data (e.g., DNA sequences, protein structures) into meaningful knowledge about life processes.                                       |
| Key Domains  | Genomics, proteomics, transcriptomics, metabolomics, systems biology, evolutionary biology, and clinical informatics.                                         |

---

## 2. Why Bioinformatics Matters

| Reason                    | Impact                                                                                                            |
|---------------------------|-------------------------------------------------------------------------------------------------------------------|
| Data Explosion            | Modern high-throughput technologies generate terabytes of data daily (e.g., NGS, mass spec).                      |
| Disease Understanding     | Identifies genetic variants linked to disease, informs drug target discovery, and supports personalized medicine. |
| Agriculture & Environment | Improves crop breeding, monitors biodiversity, and tracks pathogen evolution.                                     |
| Fundamental Science       | Reveals evolutionary relationships, gene function, and regulatory networks.                                       |

---

## 3. Core Concepts & Terminology

| Term          | Meaning                                                                 |
|---------------|-------------------------------------------------------------------------|
| Genome        | Complete DNA sequence of an organism.                                   |
| Transcriptome | Full set of RNA transcripts expressed in a cell/tissue at a given time. |
| Proteome      | Entire collection of proteins produced by a cell/organism.              |

# Introduction to Bioinformatics

| Term                       | Meaning                                                                              |
|----------------------------|--------------------------------------------------------------------------------------|
| <b>Sequence Alignment</b>  | Arranging two or more DNA/RNA/protein sequences to identify regions of similarity.   |
| <b>Homology</b>            | Similarity due to shared ancestry; used to infer function.                           |
| <b>Variant (SNP/Indel)</b> | Single nucleotide polymorphism or insertion/deletion relative to a reference genome. |
| <b>Annotation</b>          | Adding functional information (genes, regulatory elements) to raw sequences.         |
| <b>Database</b>            | Structured repository (e.g., GenBank, UniProt, PDB) that stores biological data.     |
| <b>Pipeline</b>            | Automated series of computational steps that process raw data to results.            |

---

## 4. Typical Bioinformatics Workflow

### 1. Experimental Design

- Define hypothesis, select organism, choose technology (e.g., Illumina, PacBio).

### 2. Data Generation

- High-throughput sequencing, microarrays, mass spectrometry, imaging.

### 3. Quality Control (QC)

- Remove low-quality reads, adapters, contaminants. Tools: **FastQC**, **Trimmomatic**.

### 4. Pre-processing

- **Read alignment** (e.g., **BWA**, **STAR**) ? map reads to a reference.
- **Assembly** (de novo or reference-guided) ? build contigs/scaffolds (**SPAdes**, **Canu**).

### 5. Variant Calling / Feature Extraction

- Detect SNPs, indels, structural variants (**GATK**, **FreeBayes**).
- Quantify expression (**HTSeq**, **featureCounts**).

### 6. Annotation & Functional Interpretation

# Introduction to Bioinformatics

- Assign gene names, GO terms, pathways (**ANNOVAR**, **Ensembl VEP**, **KEGG**).

## 7. Statistical Analysis & Visualization

- Differential expression (**DESeq2**, **edgeR**).
- Phylogenetic trees (**MEGA**, **IQ-TREE**).
- Network analysis (**Cytoscape**, **STRING**).

## 8. Validation & Reporting

- Experimental verification (e.g., qPCR, Western blot).