

UNIT 1

Bioinformatics

- General
 - ✓ Bioinformatics is the science of managing, mining and interpreting information from biological sequences and structures. In this area of science, biology, computer science and information technology, all the three merge into a single discipline.
 - ✓ During the last few years, bioinformatics has been overwhelmed with increasing floods of data, both in terms of volume and in terms of new databases and new types of data. Basic types of data that can be analyzed in bioinformatics are (a) Raw Deoxyribonucleic Acid sequences (b) Protein sequence (c) Macromolecular structure (d) Genome (e) Gene Expression.

Molecular Biology

- Cells are fundamental working units of every living system. All instructions needed to direct their activities are contained within the chemical DNA (Deoxyribonucleic Acid). DNA from all organisms is made up of the same chemical and physical components. The DNA sequence is the particular side-by-side arrangement of bases along the DNA strand. (E.g.: ATTCCGGA) This order spells out the extra instructions required to create a particular organism with its own unique traits. The genome is an organism's complete set of DNA.
- However only fragments of genome are responsible for the functioning of the cell. These fragments, called genes, are the basic physical and functional units of heredity. Genes are made up of a contiguous set of codons, each of which specifies an amino acid. (Three consecutive nucleotide bases in a DNA sequence constitute a 'codon'; for example, 'AGT' and 'ACG' are two consecutive codons in the DNA fragment AGTACGT.

- Of the 64 possible different arrangements of the four nucleotides (A, T, G, C) in sets of three, three (UAA, UAG, UGA) functionally act as periods to translating ribosome in that they cause the translation to stop.
- These three codons are therefore termed as 'stop codons'. Similarly, one codon of the genetic code, namely ATG, is reserved as start codon, though GTG, CTG, TTG are also rarely observed.) Genes translate into proteins and these proteins perform most life functions and even make up the majority of cellular structures.

Gene Prediction

- The gene identification problem is the problem of interpreting nucleotide sequences by computer, in order to provide tentative annotation on the location, structure, and functional class of protein-coding genes.
- In computational biology, **gene prediction** or **gene finding** refers to the process of identifying the regions of genomic DNA that encode genes. This includes protein-coding genes as well as genes, but may also include prediction of other functional elements such as regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced.

- Gene prediction is one of the key steps in genome annotation, following sequence assembly, the filtering of non-coding regions and repeat masking.
- Gene prediction is closely related to the so-called 'target search problem' investigating how DNA-binding proteins (transcription factors) locate specific binding sites within the genome.

- Many aspects of structural gene prediction are based on current understanding of underlying biochemical processes in the cell such as gene transcription, translation, protein–protein interactions and regulation processes, which are subject of active research in the various omics fields such as transcriptomics, proteomics, metabolomics, and more generally structural and functional genomics.