Bioinformatics-Introduction

Manu Madhavan

Lecture 1

Outline

- Course Introduction
- What is Bioinformatics?
- What we learn in Bioinformatics?
- Applications of ML in Bioinformatics
- Molecular Biology-primer

Course Details

- Code: 15CSE355
- Title: Bioinformatics
- L-T-P-C: 3 0 0 3
- Faculty: Dr. Manu Madhavan
- Slot: B

Course Outcome

1. To understand the fundamental concepts of databases, interfaces,approaches and types of network topology
2. Apply suitable data mining techniques to classify and structure the data
3. Analyze and visualize the structure of the pattern towards pattern discovery
4. Apply machine learning to formulate sequence alignment
5. Analyze the network model to collaborate and communicate

Syllabus

Unit-1

Introduction: The Central Dogma – Killer Application – Parallel Universes – Watson's Definition –Top-Down vs Bottom-Up Approach – Information Flow – Conversance – Communications. Database and Networks: Definition – Data Management – Data Life Cycle – Database Technology – Interfaces – Implementation – Networks: Communication Models – Transmission Technology – Protocols – Bandwidth – Topology – Contents – Security – Ownership –Implementation.

Unit-2

Search Engines and Data Visualization: Search Process – Technologies – Searching and Information Theory – Computational Methods – Knowledge Management – Sequence Visualizations – Structure Visualizations – User Interfaces – Animation vs Simulation. Statistics, Data Mining and Pattern Matching: Statistical Concepts – Micro Arrays – Imperfect Data – Basics – Quantifying – Randomness- Data Analysis – Tools Selection – Alignment – Clustering – Classification – Data Mining Methods – Technology – Infrastructure Pattern Recognition – Discovery.

Unit-3

Machine Learning – Text Mining – Pattern Matching Fundamentals – Dot Matrix Analysis – Substitution Matrix – Dynamic Programming – Word Method – Bayesian Method – Multiple Sequence Alignment Tools. Modelling Simulation and Collaboration: Drug Discovery Fundamentals –Protein Structure – System Biology Tools – Collaboration and Communication – Standards – Issues –Case Study

Evaluation Pattern

	Assessment	Component	Max Marks	Weightage	Total
Internal	Mid-term Exam	Mid-Term	50	30	30
	Quiz-1	CA	10	3.5	20
	Quiz-2	CA	10	3.5	
	Quiz-3	CA	10	3	
	Assignment-1	CA	20	5	
	Assignment-2	CA	20	5	
External	End-sem Exam	End-sem	100	50	50
	Total			100	100

Assignments

Assignment-1

- Programming Assignment on sequence analysis
- Perl/Python/R
- Date of release: 21/12/2021
- Date of submission: 28/12/2021

Assignment-2

- Recent Applications of Machine Learning in Bioinformatics
- Group-wise (3/4 members per group)
- Conduct a survey (at-least 4 papers) and submit a comprehensive report
- Date of release: 11/12/2021
- Group details: 13/12/2021
- Date of submission: 04/11/2022

Text Books and Materials

Tools:

1. KBase, https://www.kbase.us/
2. Biopython
3. BLAST

Text Books

1. [Bergeron, 2003]Bergeron B, “Bioinformatics Computing”, Prentice Hall, 2003.

Reference Materials

1. [Smith 2001] Affward T K and Smith D J P, “Introduction to Bio Informatics”, Pearson Education, 2001.
2. [Baldi, 2003]Baldi P and Brunak S, “Bio Informatics - The Machine Learning Approach”, Second Edition, First East West Press, 2003.
3. [Krane 2002] Dan E. Krane, Michael L. Raymer, Fundamental Concepts of Bioinformatics (Krane), Benjamin/Cummings, 2002.

Online Materials

1. Canadian Bioinformatics Workshops: https://bioinformatics.ca/workshops/
2. Bioinformatics(1)- Coursera: https://www.coursera.org/learn/dna-analysis?specialization=bioinformatics
3. Bioinformatics: Algorithms and Applications: NPTEL Course by Prof. Michel Gromiha, IIT: https://nptel.ac.in/courses/102/106/102106065/
4. Bioinformatics Lectures by Prof. Barry Grant, https://bioboot.github.io/bioinf525_w16/module1/

Biology meets Computers

Did Pasteur and Babbage ever meet?

We do not know if they ever met, but had they met, they possibly would not have talked to each other !
Anyway, what do they have in common to talk, other than the weather?
What is there in common between the gear wheels that were turning away in an attempt to crunch numbers and the microbes playing mysterious role in fermenting alcohol?

- Biology and Computers are becoming close cousins which are mutually respecting, helping and influencing each other
- The flood of data from Biology, mainly in the form of DNA, RNA and Protein sequences, is putting heavy demand on computers and computational scientists.

Bioinformatics

- Bioinformatics = Bio + Informatics (coined by Paulien Hogeweg and Ben Hesper)
- Computational Biology = Computer + Biology
- What Biology?
  - DNA
  - RNA
  - Protein
  - Genome, Proteome,...

Bioinformatics

- Computational Biology/Bioinformatics is the application of computer sciences and allied technologies to answer the questions of Biologists, about the mysteries of life.
- Computational solutions to problems involving data emerging from within cells of living beings
- What is involved in Bioinformatics?

  - Analysing DNA sequence data to locate genes √
  - Analysing RNA sequence data to predict their structure√
  - Analysing protein sequence data to predict their location inside cell √
  - Developing medicinal plant data base ×
  - Analysing gene expression images √
  - Using computers to identify finger prints ×
  - Using computers in process control in bio-technology industries ×
  - Identifying new Drug Molecules √
  - Using computers to analyse ECG signals ×

Bioinformatics

- As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret the biological data.
- Bioinformatics has been used for *in silico* analyses of biological queries using mathematical and statistical techniques.

Bioinformatics-Goals

- Development and implementation of computer programs that enable efficient access to, management and use of, various types of information.
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets.
- For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

Molecular Biology- Essentials

- **Molecular biology** is the branch of biology that studies the molecular basis of biological activity.
- Cell is the basic structural, functional, and biological unit of all living organisms
- Three important molecules of life: Deoxyribo Nucleic Acid (DNA), Ribo Nucleic Acid (RNA) and Protein

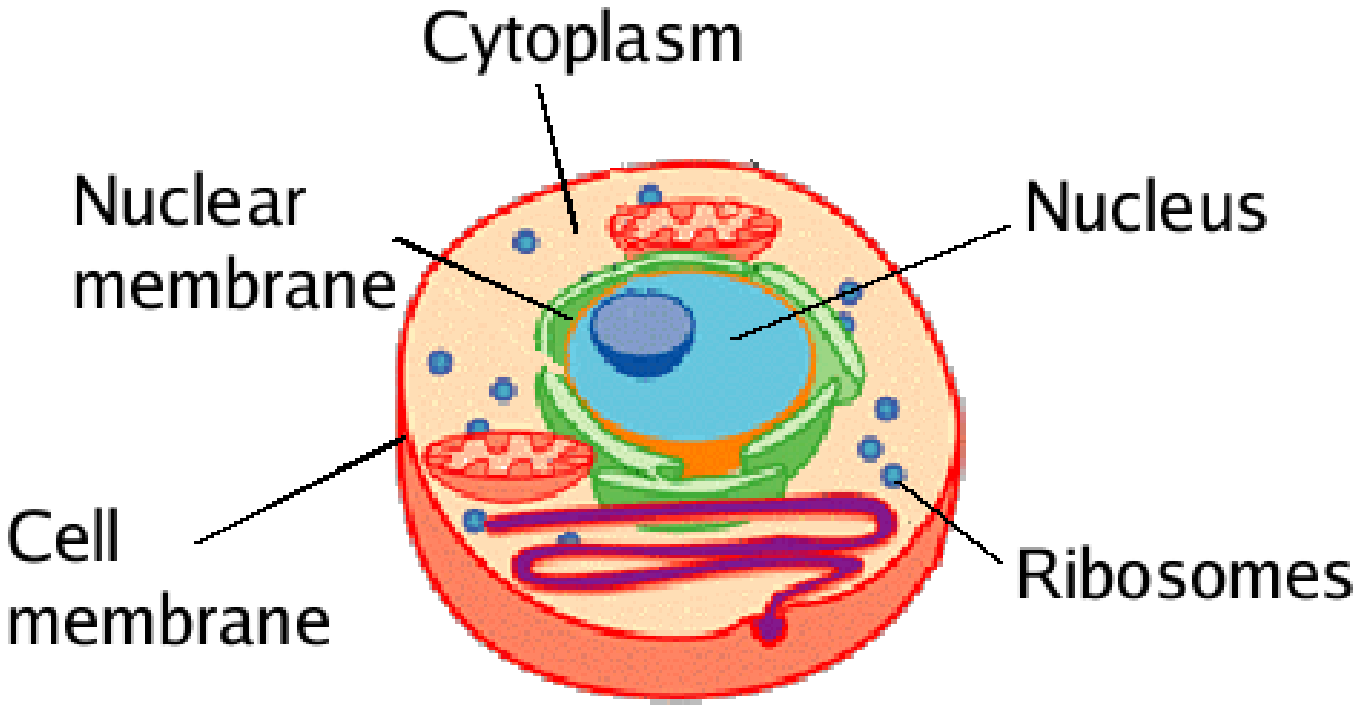


Figura: Basic structure of a cell[1]

Molecular Biology- Essentials

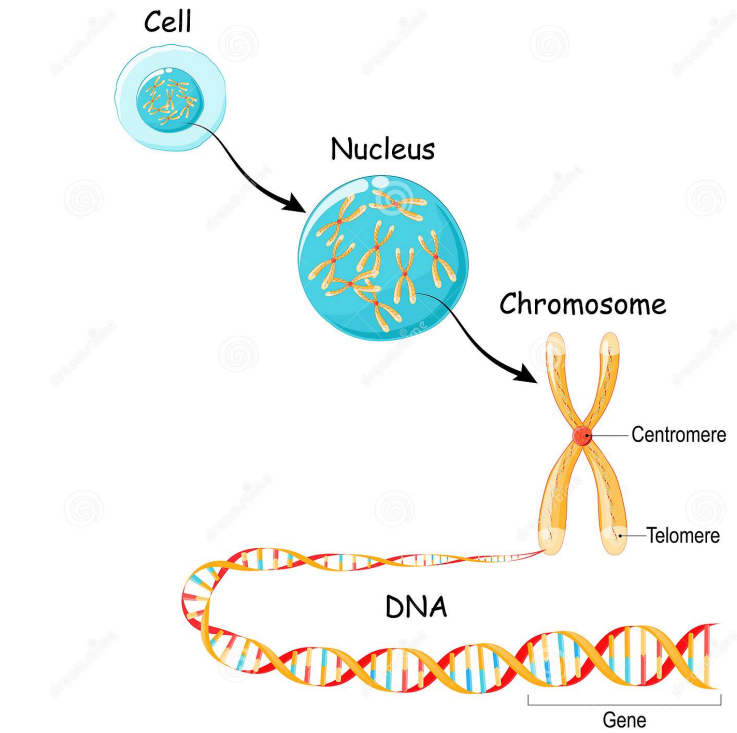
- Cells fall into two categories: *Eukaryotic, Prokaryotic*
- *Eukaryote* is a developed organism like a human being or a tree.
- *Prokaryotes* are lower forms of life like bacteria.

Figura: Eukaryote

Figura: Prokaryote

Molecular Biology- Essentials

- Cell have central core called nucleus
- Nucleus is the store house of important molecule called DNA
- DNA are packaged in units called **Chromosomes**
- DNA is the genetic *blue-print*



Three Key Molecules of Life

- **DNA** is a double stranded molecule, stores genetic information. It is composed of nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T)
- **RNA** is single stranded, decipher genetic information in DNA. It is composed of A, C, G and Uracil (U)

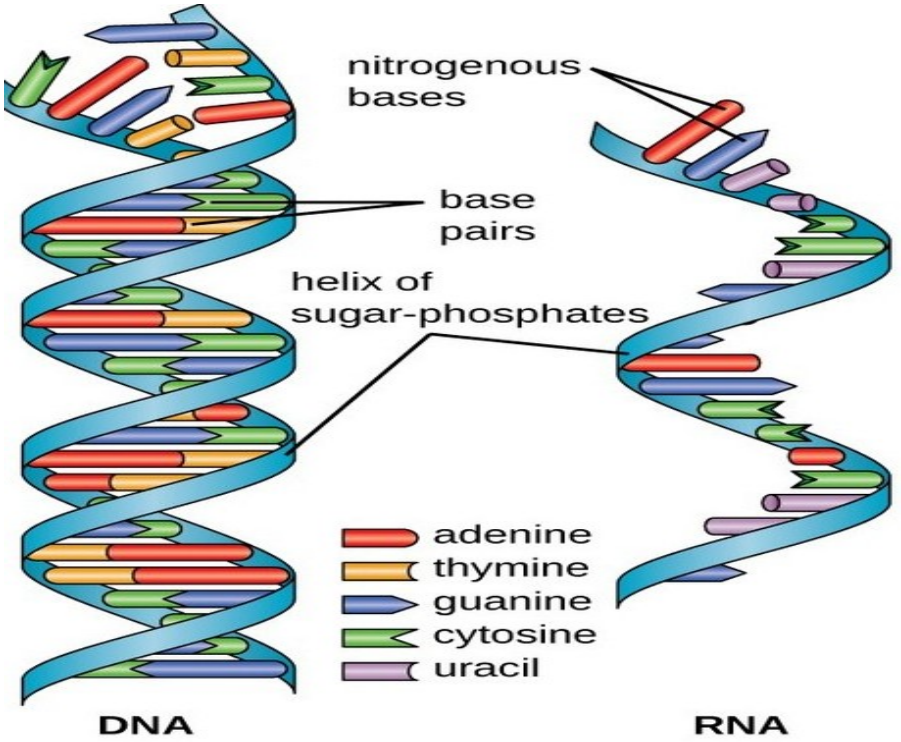


Figura: [a]

[a] Image source: https://courses.lumenlearning.c

Three Key Molecules of Life

- Sequence of three RNA nucleotides are called **codons** that corresponds to an amino acid
- **Proteins** are long chain of amino acids
- Proteins are essential for growth and metabolism of living cell

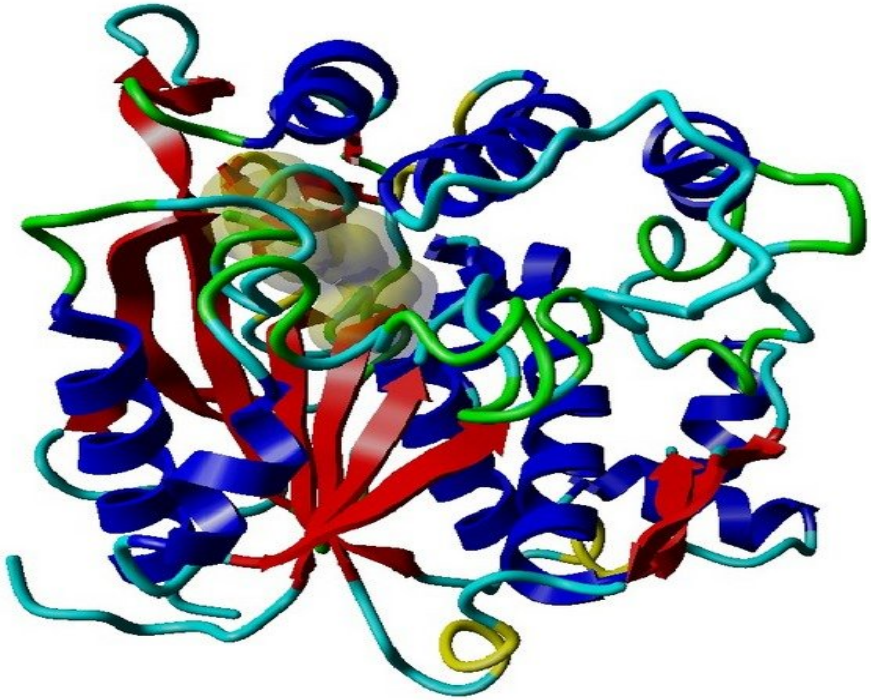


Figura: Protein structure[a]

[a]Image source: https://mirakind.org/

Biological Sequence

(a) DNA Data (4 letter strings)

(b) RNA Data (4 letter strings)

(c) Protein Data (20 letter strings)

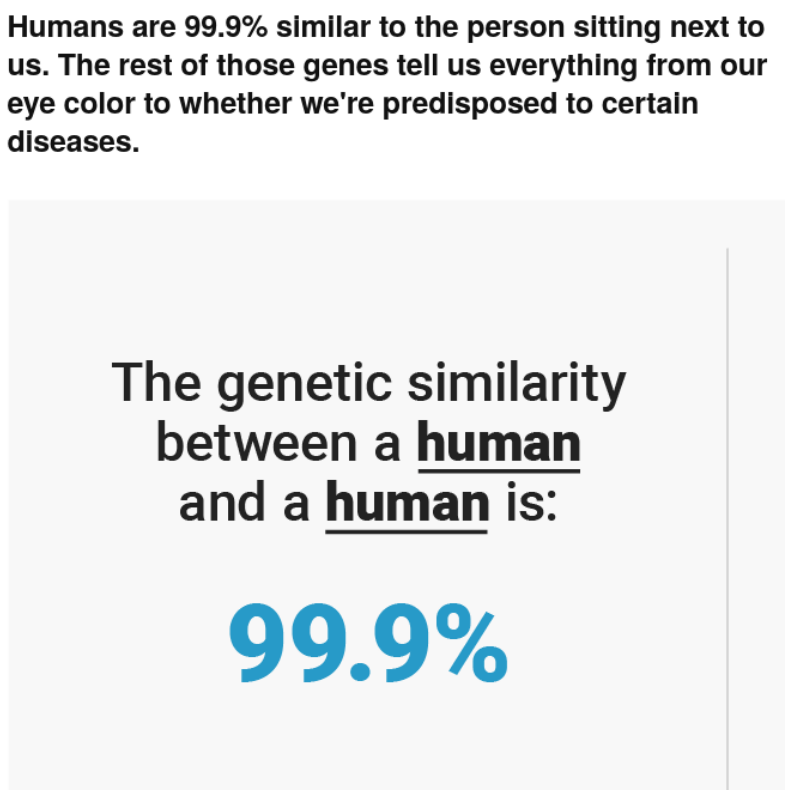
Gene,Genome

- A **gene** is the basic physical and functional unit of heredity, made up of DNA.
- Every person has two copies of each gene, one inherited from each parent.
- Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people
- **Alleles** are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each persons

Figura: Only 0.01% difference in genome between human and ape

Gene,Genome

- Some genes act as instructions to make molecules called proteins (codes for proteins)
- Many genes not code for proteins (non-coding)
- Genome is the sum total of organism's DNA
- Human Genome Project (HGP) estimated that humans have between 20,000 and 25,000 genes.



99% of genes are same in humans

Think!

- What cause some difference in individuals?
- Why some individuals are prone to some diseases? -Can we identify some genetic relations?
- Can we prescribe personalized treatment based on their genetic differences?

Reading Assignment

Jacques Cohen, *Introduction to Bioinformatics for Computer Scientist, ACM Computing Surveys, Vol. 36, No. 2, June 2004, pp. 122*

Reading Assignment

Central Dogma of Molecular Biology