Introduction to the course

BT302 Bioinformatics (2-0-2-6)

Syllabus

Introduction to biological databases: collection, organization, storage and retrieval of data; Concept of homology and definition of associated terms; Pairwise sequence alignment: dynamic programming algorithm, global (Needleman-Wunsch) and local (Smith-Waterman) alignments; BLAST Scoring matrices (PAM and BLOSUM families), gap penalty, statistical significance of alignment; Multiple sequence alignment: progressive alignment, iterative alignment, Sum-of-pairs method, CLUSTAL W; Pattern recognition in protein and DNA sequences, Hidden Markov Model, Profile construction and searching, PSI-BLAST.

Big data analysis: Introduction to Next-generation sequencing analysis, RNA-seq, CHIP-seq, Introduction to phylogeny: maximum parsimony method, distance method (neighbor-joining), maximum-likelihood method; Gene prediction in prokaryotes and eukaryotes, homology and ab-initio methods; Genome analysis and annotation; comparative genomics, cluster of orthologous groups.

Text Books/References

1. Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison. Biological Sequence Analysis. Cambridge University Press, 1998.

2. Gibas C and Jambeck P, *Developing Bioinformatics Computer Skills*, O'Reilly Media Inc., 2001.

3. Bergeron B, *Bioinformatics Computing: The Complete Practical Guide to Bioinformatics for Life Scientists*, Prentice Hall, 2002.

4. Krane De and Raymer ML: *Fundamental Concepts of Bioinformatics*. Pearson Education, 2002.

5. Kochan, Stephen G; Wood, Patrick. UNIX shell programming (3rd Ed.). SAMS, 2003.

6. **Mount, David W. Bioinformatics: Sequence and Genome Analysis (3rd Ed.). CSH press, 2005**

7. Jean-Michel Claverie and Cedric Notredame, *Bioinformatics for Dummies*. Wiley Publishing Inc., 2006.

8. Attwood TK, Parry-Smith DJ and Phukan S, *Introduction to Bioinformatics*. Pearson Education, 2007.

9. Marketa Zvelebil, Jeremy O. Baum. Understanding Bioinformatics. Garland Science, 2007.

10. **Zhumur Ghosh and Bibekanand Mallick, *Bioinformatics: Principles and Applications*, 2008.**

11. Bourne, Philip E. Structural Bioinformatics. (2nd Ed.). Wiley, 2009.

Course Instructors

Prof. Shankar Prasad Kanaujia

First half: 26th July – 24th September 2023

Prof. B. Anand

Second half: 25th September – 25th November 2023

Tentative distribution of classes

S. No.	Date/ Day	Time	Topic to be covered
1	26.07.2023 (Wednesday)	10.00 – 12:00 PM	Introduction to the course
2	31.07.2023 (Monday)	3:00 – 4:00 PM	Biological databases (BDBs)
3	01.08.2023 (Tuesday)	2:00 – 3:00 PM	BDBs and Concepts of homology and related terms
4	02.08.2023 (Wednesday)	10.00 – 12:00 PM	Pairwise sequence alignment
5	07.08.2023 (Monday)	3:00 – 4:00 PM	,,
6	08.08.2023 (Tuesday)	2:00 – 3:00 PM	,,
7	09.08.2023 (Wednesday)	10.00 – 12:00 PM	Lab 1 (Biological databases)
8	14.08.2023 (Monday)	3:00 – 4:00 PM	Scoring matrices
9	16.08.2023 (Wednesday)	10:00 – 12:00 PM	Lab 2 (Sequence alignment)
10	17.08.2023 (Thursday)	2:00 – 3:00 PM	Scoring matrices (Tuesday time table)
11	21.08.2023 (Monday)	3.00 – 4:00 PM	BLAST
12	22.08.2023 (Tuesday)	2:00 – 3:00 PM	,,
13	23.08.2023 (Wednesday)	10:00 – 12:00 PM	Lab 3 (BLAST and Scoring matrices)

Tentative distribution of classes

S. No.	Date/ Day	Time	Topic to be covered
14	28.08.2023 (Monday)	3.00 – 4:00 PM	Multiple sequence alignment
15	29.08.2023 (Tuesday)	2:00 – 3:00 PM	,,
16	30.08.2023 (Wednesday)	10:00 – 12:00 PM	Lab 4 (MSA)
17	04.09.2023 (Monday)	3.00 – 4:00 PM	Pattern recognition (Motifs)
18	05.09.2023 (Tuesday)	2:00 – 3:00 PM	,,
19	06.09.2023 (Wednesday)	3:00 – 4:00 PM	Quiz (Thursday time table)
20	11.09.2023 (Monday)	3.00 – 4:00 PM	Hidden Markov Model
21	12.09.2023 (Tuesday)	2:00 – 3:00 PM	,,
22	13.09.2023 (Wednesday)	10:00 – 12:00 PM	Lab 5 (Motifs)
23	16.09.2023 (Saturday)	9:00 – 1:00 PM	Lab Viva
24	18.09.2023 to 24.09.2023	XXX	Mid semester examination

Evaluation scheme

Weightage: 50% each (first and second halves)

Marks distribution (1st Half):
Quiz: 1x10 marks = 10 marks
Lab assignments: 4 x 5 marks = 20 marks
Lab viva: 30 marks
Mid semester examination: 40 marks
Total marks: 100 marks

Lecture notes availability

MOODLE

https://www.iitg.ac.in/moodle/login/index.php

(bt302_iitg2023)

Introduction to Bioinformatics

Interview with Dr. Satyajit Mayor (NCBS)

EMBO *reports: In your opinion, what are the upcoming fields in biology that are worth focusing on, maybe also with regards to where Indian science should be headed?*

Mayor: One area with great potential is quantitative biology that could be applied to every area of the life sciences.

EMBO Reports, 14, 2013

Quantitative biology is an umbrella term encompassing the use of mathematical, statistical or computational techniques to study life and living organisms.

The central theme and goal of quantitative biology is the **creation of predictive models based on fundamental principles governing living systems**.

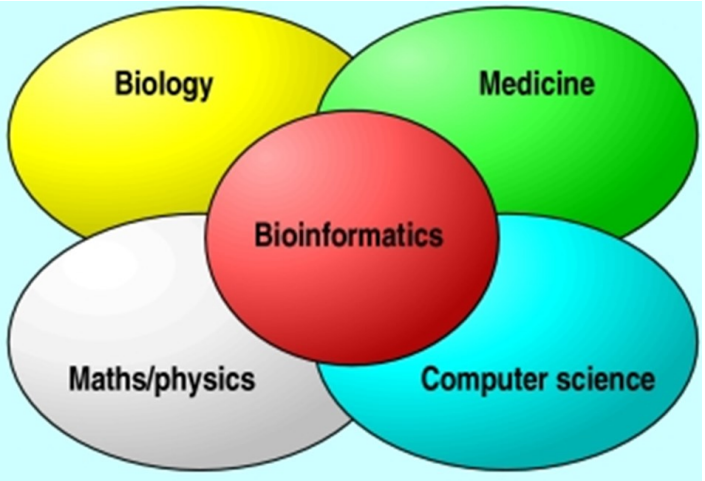
The subfields of biology that employ quantitative approaches include: (1) Mathematical and theoretical biology, (2) **Computational biology**, (3) **Bioinformatics**, (4) Biostatistics, (5) Systems biology, (6) Synthetic biology, (7) Epidemiology, etc.

Bioinformatics

Bioinformatics is the **application of computer technology to the understanding and effective use of biological and biomedical data**.

It is the discipline that stores, analyses and interprets the big data generated by life-science experiments, or collected in a clinical context.

This multidisciplinary field is driven by experts from a variety of backgrounds: biologists, computer scientists, mathematicians, statisticians and physicists.



The journal Nature defines Bioinformatics as
Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.

Bioinformatics vs Computational Biology

Bioinformatics is a multidisciplinary field that combines biological knowledge with computer programming and large sets of big data.

Computational biology is a multidisciplinary field that uses computer science, statistic, and mathematics to help solve problems in biology.

There is a great deal of **overlap between bioinformatics and computational biology**. **But bioinformatics requires more programming and technical knowledge**. **In bioinformatics, scientists can interpret the results of more complex research studies. Some well-known examples of bioinformatics include analysis of genetic and genomic data, chemi-informatic comparison of proteins to assist personalized medicine, prediction of protein function from data sequence and structural information, etc**.

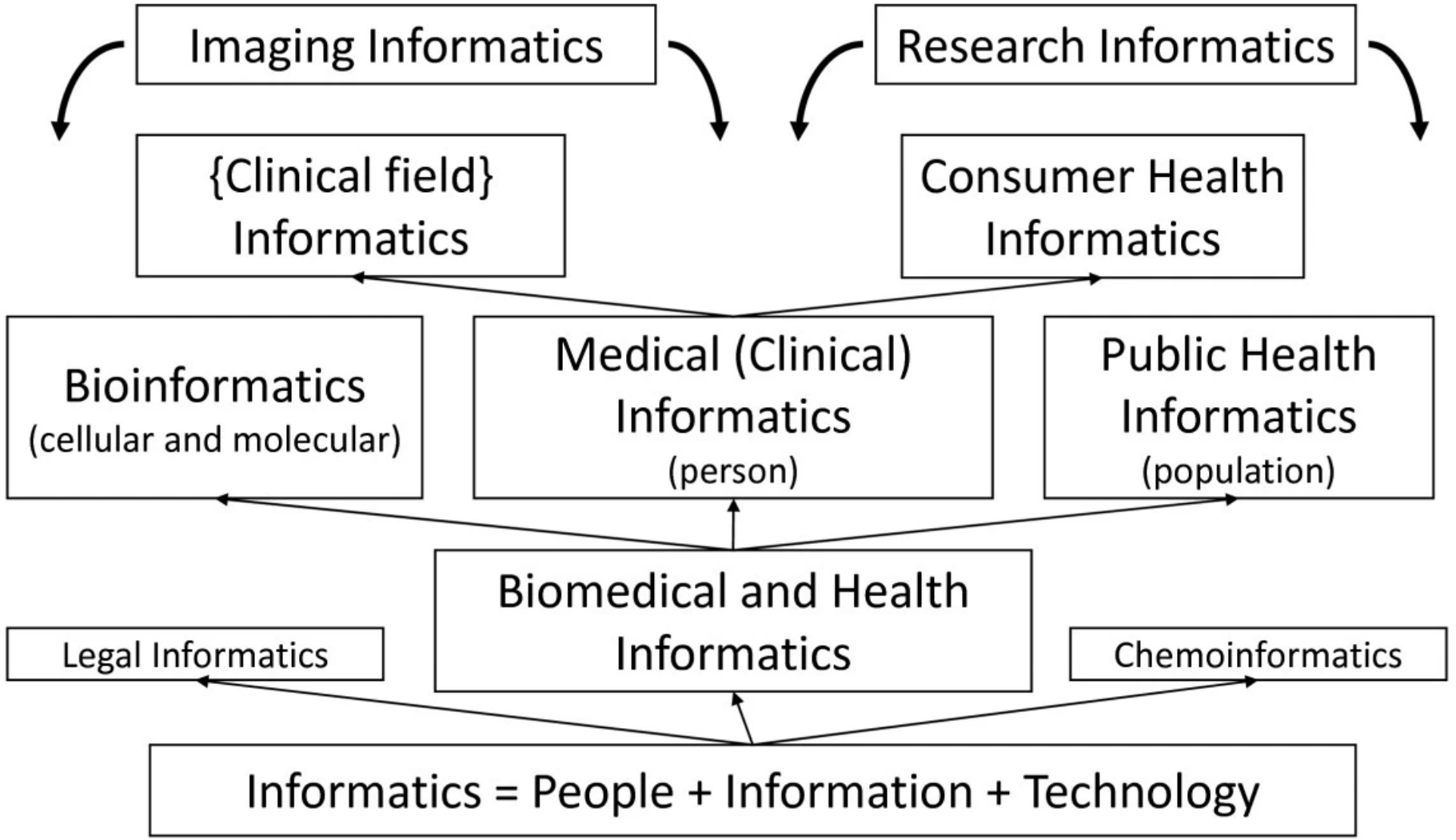
Computational biology uses mathematical and computational approaches to address experimental and theoretical queries in biology. It can also include the development of algorithms, theoretical models, computational simulation, and mathematical models.

Bioinformatics vs Computational Biology

Bioinformatics describe the field that answers the question "***How can I efficiently store, annotate, search and compare information from biological measurements and observations***?"

Computational biology is the science that answers the question "***How can we learn and use models of biological systems constructed from experimental measurements?***"

Major subcategories of the informatics field



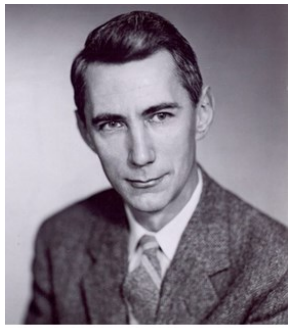
Historical aspects of Bioinformatics

(1890 – 1962)

1930: Sir Ronald A Fisher helped put both Mendelism and Darwinism on a firm mathematical footing, and he is also credited with being the **first to apply a computer to biology**. **Created ANOVA, ML, DoE, etc**.

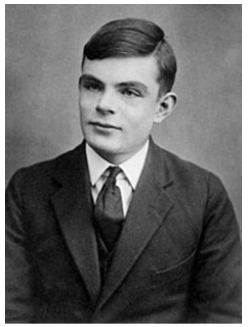
(1902 – 1984)

1939: George G Simpson had co-authored the first book on ***quantitative methods in biology***.



(1916 – 2001)

1940: Claude E Shannon's PhD thesis, entitled "**An Algebra for Theoretical Genetics**", formalized **population genetics**. **Today it would be labeled bioinformatics**. It is intriguing to think that two giants of mathematics (**Claude E Shannon**) and computer science (**Alan M Turing**) may have come so close to committing their careers to biology.



1912 – 1954

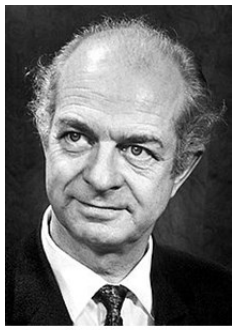
Historical aspects of Bioinformatics

1946: **Erwin Schrodinger**'s ***what is life***, which was stimulated by the work of physicist-turned-biologist Max Delbruck (mentor to James Watson), influenced Francis Crick.

1887 – 1961

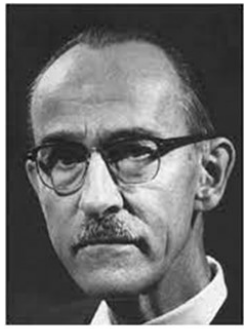
1904 – 1968

1947: **George Gamow** book ***One Two Three … Infinity*** (fly genetics) – framed ***sequence bioinformatics*** and the ***conceptualization of genetic code***.



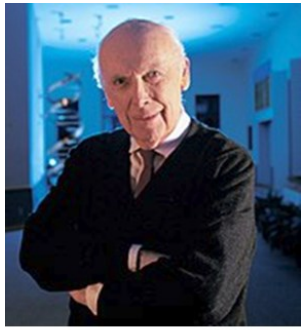
1901 – 1994

1951: **Configurations of polypeptide chains** – **Linus Pauling** and **Robert Corey**, the first milestones concerning the prediction of a protein structure that reported the prediction of α-helices and β-sheets.



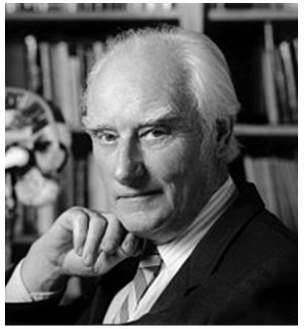
1897 – 1971

Historical aspects of Bioinformatics

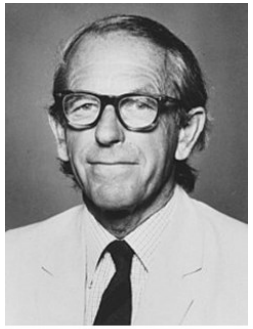


1928 – till date

1953: **Molecular structure of nucleic acids** – paper by **James Watson** and **Francis Crick**.



1916 – 2004



1918 – 2013

1955: **The sequence of the first protein to be analyzed**, **bovine insulin**, is announced by **Frederick Sanger**.

1955: **Robert Ledley (Gamow lab)** went on to promote **computer-based medical diagnosis** and ***protein sequence tools and databases***.

1926 – 2012

Historical aspects of Bioinformatics

1925 – 1983

1965: ***Atlas of Protein Sequence and Structure***, the first ever biological sequence database. 1978: Creation of PAM matrices by Margaret Dayhoff (***The First Bioinformatician***). David J. Lipman, Director NCBI, called her ***the mother and father of bioinformatics***. **Dayhoff later developed the one-letter amino acid code that is still in use today**.

1970: The details of the ***Needleman-Wunsch*** algorithm for sequence comparison are published (JMB).

1981: The ***Smith-Waterman*** algorithm for sequence alignment is published (JMB).

1986: The term "***Genomics***" appeared for the first time to describe the scientific discipline of mapping, sequencing, and analyzing genes by **Thomas Roderick**.

1986: The **SWISS-PROT** database is created by the **Department of Medical Biochemistry of the University of Geneva** and the **European Molecular Biology Laboratory (EMBL)**.

Historical aspects of Bioinformatics

1988: **Hwa A Lim (HAL)** coined the term ***Bioinformatics*** and is credited with establishing and shaping the field of ***bioinformatics***, credits that earn him the title of "***Father of Bioinformatics***." He is CEO of MIRAHI Inc. (Tong Ren Tang, Chinese Largest producer of traditional Chinese Medicine).

1988: The National Center for Biotechnology Information (NCBI). The Human Genome Initiative is started. The FASTA algorithm for sequence comparison is published by **David J Lipman** and **Robert Pearson**.

1990: The **BLAST** program is implemented.

1991: **First paper on bioinformatics from India (University of Pune)**.

2001: The **human genome** (3,000 Mbp) is published.

Historical aspects of Bioinformatics

Bioinformatics Policy of India (BPI – 2004):

The principal aim of the bioinformatics programme was to ensure that India emerged a key international player in the field of bioinformatics; enabling a greater access to information wealth created during the post-genomic era and catalyzing the country's attainment of lead position in medical, agricultural, animal and environmental biotechnology.

India should make a niche in Bioinformatics industry and would work to create bioinformatics industry with US$10 billion by the end of 10^{th} Plan period.

Thank You