

# Computational Molecular Biology and Bioinformatics

## A Little Bit of Biology

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit  
Indian Statistical Institute, Kolkata

July, 2025

# Imagine yourself having such eyes

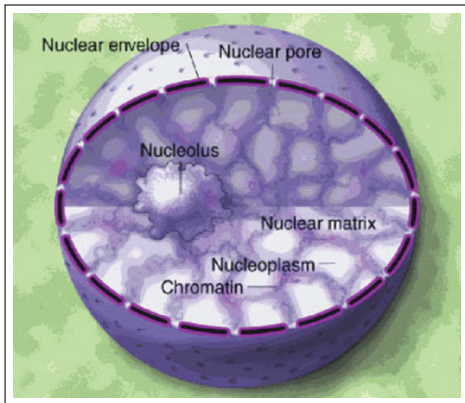


- 1 Introduction to molecular biology
- 2 Scope of research
  - Computational biology vs bioinformatics
  - Sequence analysis
  - Expression analysis
  - Genetic analysis
  - Epigenetic analysis
  - System-level analysis
  - Pathway analysis
- 3 Bioinformatics databases
- 4 Suggested reading





# Nucleus to chromosome



**Zooming from nucleus into the chromosome**



# DNA to nucleotide

...	T	T	A	C	C	G	T	C	G	G	C	A	...
...	A	A	T	G	G	C	A	G	C	C	G	T	...

**Note:** There are 3 billion base pairs in each cell to fit into a space of approximately 6  $\mu\text{m}$  across the human cells.

# This is your haystack!!!

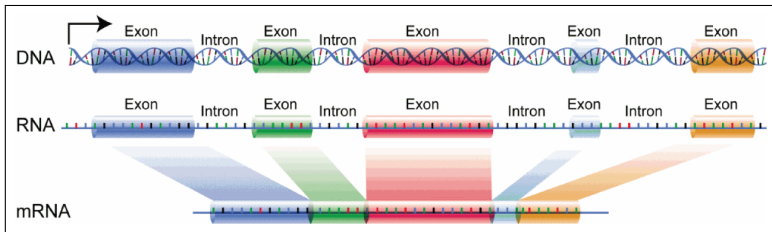
... AATTTTCGGATTGCCAATTCGGAAAATTTTCGGATTGCCAA  
TTCGGTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
AAATTGCCAATTTTTCGAATTCGGAAAATTTTCGGATTGCCAA  
TTCGGTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
AAAATTTTCGGATTGCCAATTCGGAAAATTTTCGGATTGCCAA  
TTCGGCGGATTATTGCCAATTCCTCAAAGCCAATTCGGAAAAC  
CAAATTTTCGGATTGCCAATTCGGAAAATTTTCGGATTGCCAA  
TTCGGTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
TTCAATTTTCGGATTGCCAATTCGGAAAATTTTCGGATTGCCAA  
GGGAATTTTCGGATTGCCAATTCGGAAAATTTTCGGATTGCCAA  
ATCGGTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
GAAAATTTTCGGATTGCCAATTCGGAAAATTTTCGGATTGCCAA  
ATCGGTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAA ...

# This is your haystack!!!

... AATTTTCGGATTGCCAATTCGGAAAAATTTTCGGATTGCCAA  
TTCGGTTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
AAATTGCCAATTTTGGCAATTCGGAAAAATTTTCGGATTGCCAA  
TTCGGTTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
AAAATTTTCGGATTGCCAATTCGGAAAAATTTTCGGATTGCCAA  
TTCGGCGGATTATTGCCAATTCTTCAAAGCCAATTCGGAAAAC  
CAAATTTTCGGATTGCCAATTCGGAAAAATTTTCGGATTGCCAA  
TTCGGTTTC *Blue eyes???* CGGAAAAGCCAATTCGGAAAAC  
TTCAATTTTCGGATTGCCAATTCGGAAAAATTTTCGGATTGCCAA  
GGGAATTTTCGGATTGCCAATTCGGAAAAATTTTCGGATTGCCAA  
ATCGGTTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAAAAC  
GAAAATTTTCGGATTGCCAATTCGGAAAAATTTTCGGATTGCCAA  
ATCGGTTTTTCGGATTGCCAATTCGGAAAAGCCAATTCGGAA ...



# DNA-RNA-mRNA

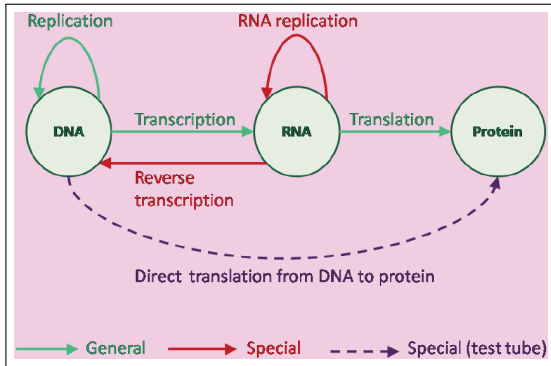


## The formation of DNA-RNA-mRNA

**Note:** The nucleotides A, T, C and G constitutes the DNA, whereas in RNA T is replaced by U.

# Central dogma of molecular biology

The central dogma of molecular biology comprises transcription followed by translation.



# Computational biology vs bioinformatics

Computational biology (case  $\rightarrow$  model) is the study of biology systems using computational models techniques. The goal is to learn new biology, gain knowledge about living systems.

Bioinformatics (data  $\rightarrow$  information) is the creation of tools like statistical methods, algorithms, databases, etc. that solves problems. The goal is to build useful tools that successfully work on the biological data.

**Note:** Computational biology is 'science' whereas bioinformatics is 'engineering'.

# Coverage areas

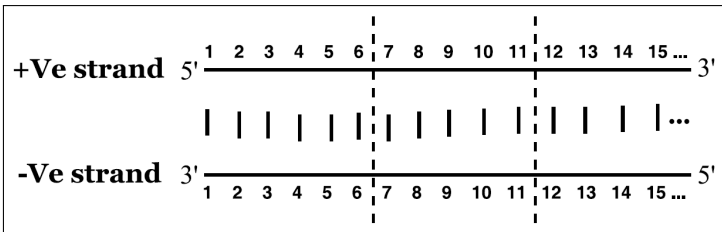
“Biology is the only science in which multiplication is the same thing as division”

– Jupiter Scientific.

Molecular biology, which serves as a link between the genetics and biochemistry, covers the following areas:

- Sequence analysis
- Expression analysis
- Genetic analysis
- Epigenetic analysis
- System-level analysis
- Pathway analysis

# Sequence annotation



## The format of sequence data

The sequence (or any subsequence) is always retrieved from the 5' to the 3' end, irrespective of the strand.

**Note:** 'A and T' and 'G and C' are complementary to each other in both these strands side by side.

# Sequence annotation

## Database 1: Genomic sequence

> ...

AATTCCGCGA ...

5'	1	2	3	4	5	6	7	8	9	10	...	3'
+Ve	A	A	T	T	C	C	G	C	G	A	...	
-Ve	T	T	A	A	G	G	C	T	C	T	...	
3'	1	2	3	4	5	6	7	8	9	10	...	5'

## Database 2: Gene list

Gene Chrom Strand Start End

Gene 1 chr1 + 2 6

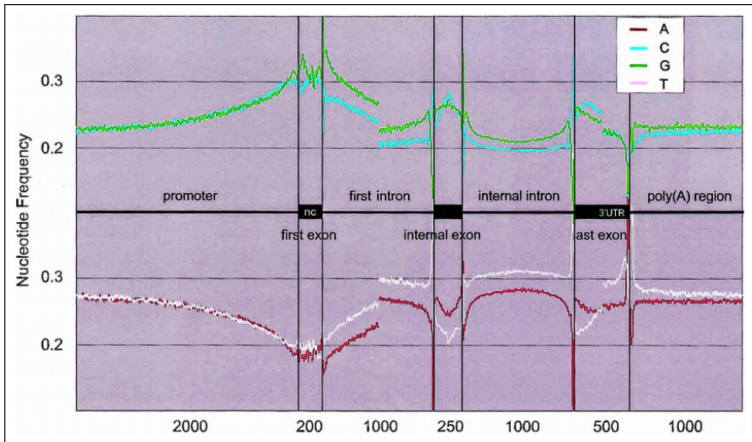
Gene 2 chr1 - 5 8

...

Sequence of Gene1: ATTCC

Sequence of Gene2: GCGG

# Distribution of nucleotides in DNA



The frequencies of nucleotides throughout the genome

# Sequence analysis

Various types of features are interesting in sequence analysis:

- $k$ -mers: The frequency counts of subsequences of length  $k$ .
- Palindromes: The palindromic subsequences.
- CpG islands: The regions consisting of high density of CG and G+C contents.
- Special patterns: Special subsequence patterns.

**Note:** Several important genomic regions (e.g., promoter, gene body, etc.) have their unique features.

# CpG islands

CpG islands are identified based on the following criteria:

- It has a length greater than 200 bp.
- The GC content is 50% or greater.
- Ratio between the observed and expected CpG in the segment is greater than 0.6.

The ratio between the observed and expected CpG ( $Obs/Exp$  CpG) is calculated as follows (Gardiner-Garden et al., JMB, 1987):

$$Obs/Exp \text{ CpG} = N * \frac{\#CG}{\#C * \#G},$$

where  $N$  denotes the length of sequence.

# FASTA and FASTQ formats

Notably, the alphabet size of DNA sequences (e.g., genes) is 4, whereas it is 20 for the amino acid sequences (e.g., proteins).

```
>gi|31563518|ref|NP_852610.1| microtubule-associated proteins 1A/1B light chain 3A isoform b [Homo sapiens]
MKMRFFSSPCGKAADVDPADRCKEVQQIRDQHPKIPVIERKGEKQLPVLDTKFLVPDHNMSSELVKI
IRRRLLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDFLYMVIYASQETFGF
```

## The sequence of a protein in FASTA format

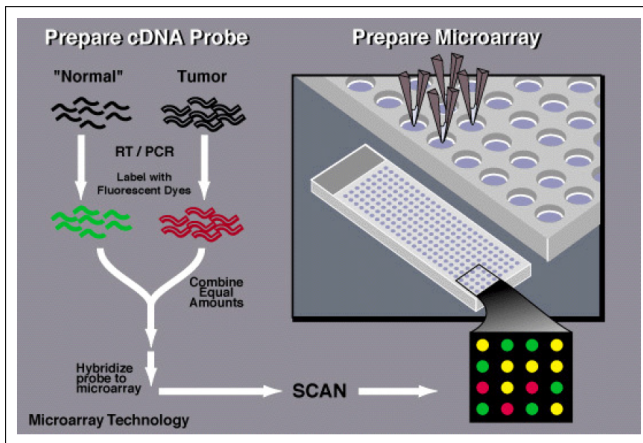
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTCAACTCACAGTTT
+
!'*(((((***+))%%%+))(%%%).1***-+''')**55CCF>>>>>CCCCCCC65
```

## The sequence of a gene in FASTQ format

**Note:** In contrast to FASTA sequences, which includes only the identifier and sequence lines of size 50, a FASTQ sequence additionally takes the quality scores.



# Microarray profiling



## Profiling the cDNA microarrays

# Snapshot of expression data

Gene	ID	t1	t2	t3	t4	t5	t6	t7
G1	...	1.2	1.9	2.4	3.2	1.1	5.7	7.4
G2	...	3.2	3.9	4.4	5.3	3	7.8	9.5
G3	...	1	2.1	3.2	6.2	7.3	8.5	3.7
...	...	...	...	...	...	...	...	...
G1000	...	2.2	3.1	6.3	5.3	8.2	2.5	4.3

## An example of expression dataset

The rows and columns in an expression dataset can be different biomolecules (genes, proteins, microRNAs, etc.) and samples (time points, tissues, patients, etc.), respectively.

Different approaches of analysis – co-expression, differential expression, differential co-expression, co-expression dynamics, etc.

# Genetics

Genetic analysis is the overall process of studying and researching in fields of science that involve genetics and molecular biology.

Different areas of epigenetic analysis are listed below:

- Genomics
- Gene finding
- Phylogenetic analysis
- Genome wide association studies (GWAS)
- Proteomics
- Pharmacogenetics

**Note:** The environment has a major impact over the genome.

# Epigenetics

The term epigenetics refers to heritable changes in gene expression (active versus inactive genes) that does not involve changes to the underlying DNA sequence; a change in phenotype without a change in genotype.

Different areas of epigenetic analysis are listed below:

- DNA methylation
- Histone modification

A methylated sequence (showing methylation with a '\*') might appear as follows

\*   \*

ATCCCGACTGCAT

# System-level analysis

## Molecular networks

- Protein-protein interaction networks
- Metabolic networks
- Regulatory networks - TF-gene networks
- Post-translational networks - Kinase-substrate networks
- RNA networks - TF-miRNA Networks, miRNA-gene networks

## Phenotypic networks

- Co-expression networks
- Genetic networks
- Disease networks

# Pathway

Biological pathway diagrams are used to describe molecular biology processes in a graphical way. A pathway is a set of related reactions in a given context, e.g., glycolysis, Krebs cycle, apoptosis, etc.

The role of bioinformatics on the pathway representations are creating specific requirements for their creation and curation.

Some popular pathway analysis tools are: KEGG, VisANT, etc.

# Bioinformatics databases

- Reference genome – NCBI (<http://www.ncbi.nlm.nih.gov>)
- Genes and proteins – NCBI, EMBL-EBI (<https://www.ebi.ac.uk>)
- MicroRNAs – miRBase (<http://www.mirbase.org>)
- Other RNAs – RNAdb (<http://research.imb.uq.edu.au/RNAdb>)
- Other biomolecules – UCSC Genome Browser (<http://genome.ucsc.edu>)
- GO analysis – Funcassociate 2.0 (<http://llama.mshri.on.ca/funcassociate>)
- Assembly conversion – Galaxy (<http://main.g2.bx.psu.edu>)
- ID conversion – DAVID (<http://david.abcc.ncifcrf.gov/conversion.jsp>)
- Gene enrichment analysis – ShinyGo (<http://bioinformatics.sdstate.edu/go>)

# Resources

## Books:

- 1 C. Setubal and J. Meidanis: Introduction to Computational Molecular Biology, PWS Publishing Company, Boston, 1997.
- 2 P. A. Pevzner: Computational Molecular Biology – An Algorithmic Approach, MIT Press, 2000.
- 3 R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison: Biological Sequence Analysis - Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.
- 4 D. Gusfield: Algorithms on Strings, Trees, and Sequences, Cambridge University Press, USA, 1997.
- 5 H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell: Molecular Cell Biology, W. H. Freeman, USA, 2000.
- 6 C.-I. Branden, J. Tooze: Introduction to Protein Structure, Garland Publishing, 1998.
- 7 A. Kowald, C. Christoph Wierling, E. Klipp, and W. Liebermeister: Systems Biology, Wiley-VCH, 2016.
- 8 B.O. Palsson: Systems Biology – Constraint based Reconstruction and Analysis, Cambridge University Press, 2015.

# Resources

## Journals:

- ① WIREs Computational Molecular Science, Wiley.
- ② Briefings in Bioinformatics, Oxford University Press.
- ③ Bioinformatics, Oxford University Press.
- ④ PLoS Computational Biology, PLoS.
- ⑤ IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE-ACM.

## Conferences:

- ① ISMB/ECCB
- ② RECOMB

# Resources

## Other similar courses:

- 1 Serafim Batzoglou at Stanford – <https://web.stanford.edu/class/cs262>
- 2 Manolis Kellis at MIT – <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-fall-2015>
- 3 Christopher Burge, David Gifford and Ernest Fraenkel at MIT – <https://ocw.mit.edu/courses/biology/7-91j-foundations-of-computational-and-systems-biology-spring-2014>
- 4 Curtis Huttenhower at Harvard – <https://canvas.harvard.edu/courses/71191>

# Evaluation criteria

## ① End-semester Evaluation:

- Written examination (50 marks)

## ② Internal Evaluation:

- Mid-semester examination (30 marks)
- Assignment submission (10 marks)
- Project work (10 marks)

**Web:** <https://www.isical.ac.in/malaybhattacharyya/Courses/CMBB/Fall2025>