Computational Molecular Biology and Bioinformatics

Introduction to Computational Biology and Bioinformatics

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

July, 2024

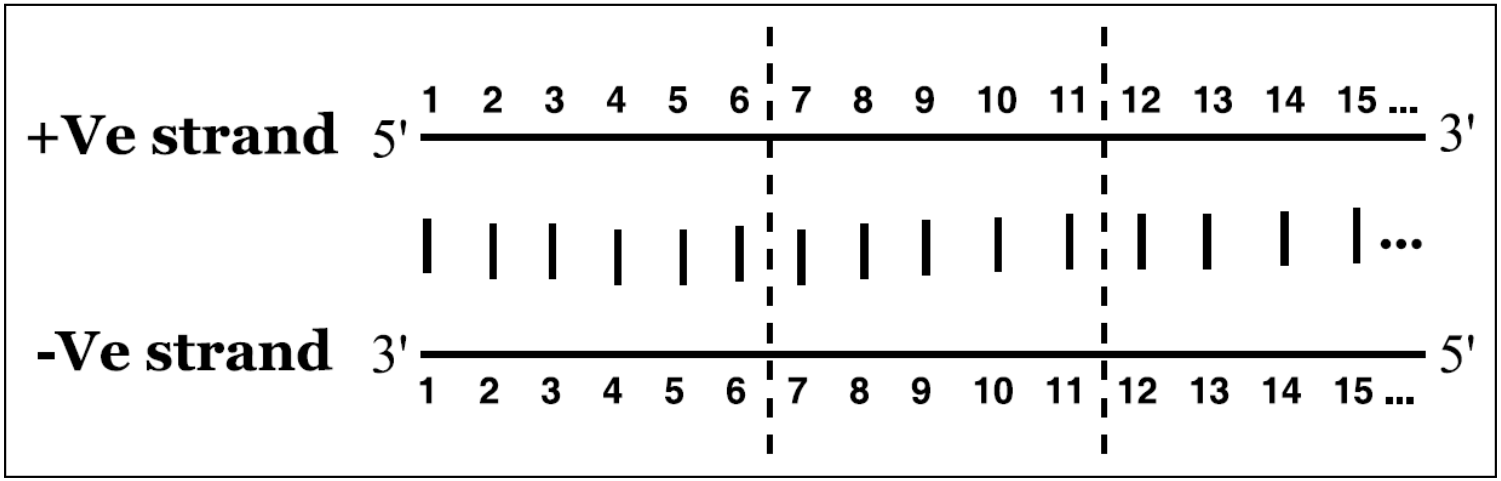
Computational biology vs bioinformatics

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'.

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

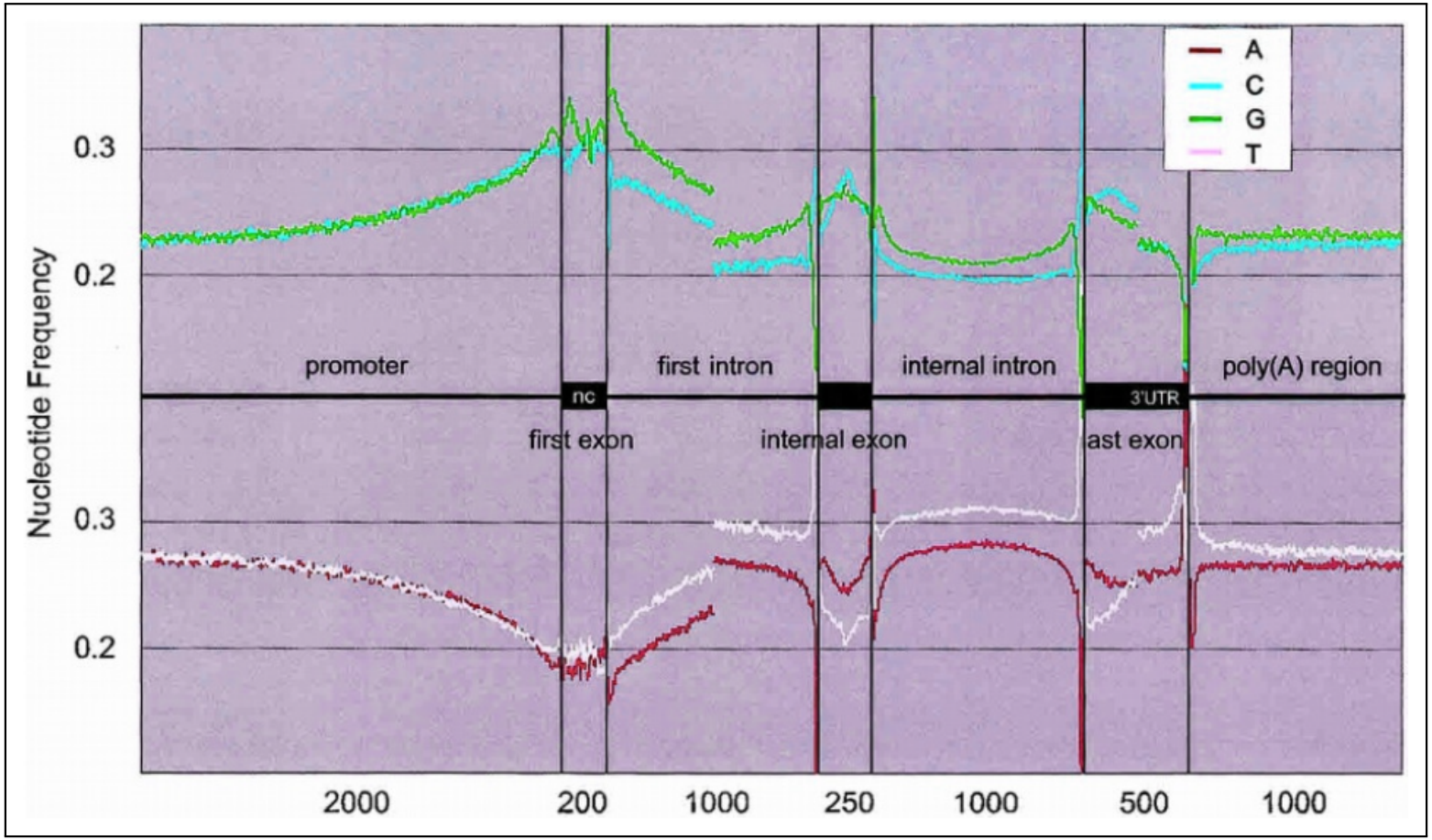
Gene 2 chr1 − 5 8

...

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):

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

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]
MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHVNMSELVKI
IRRRLQLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDGFLYMVYASQETFGF
```

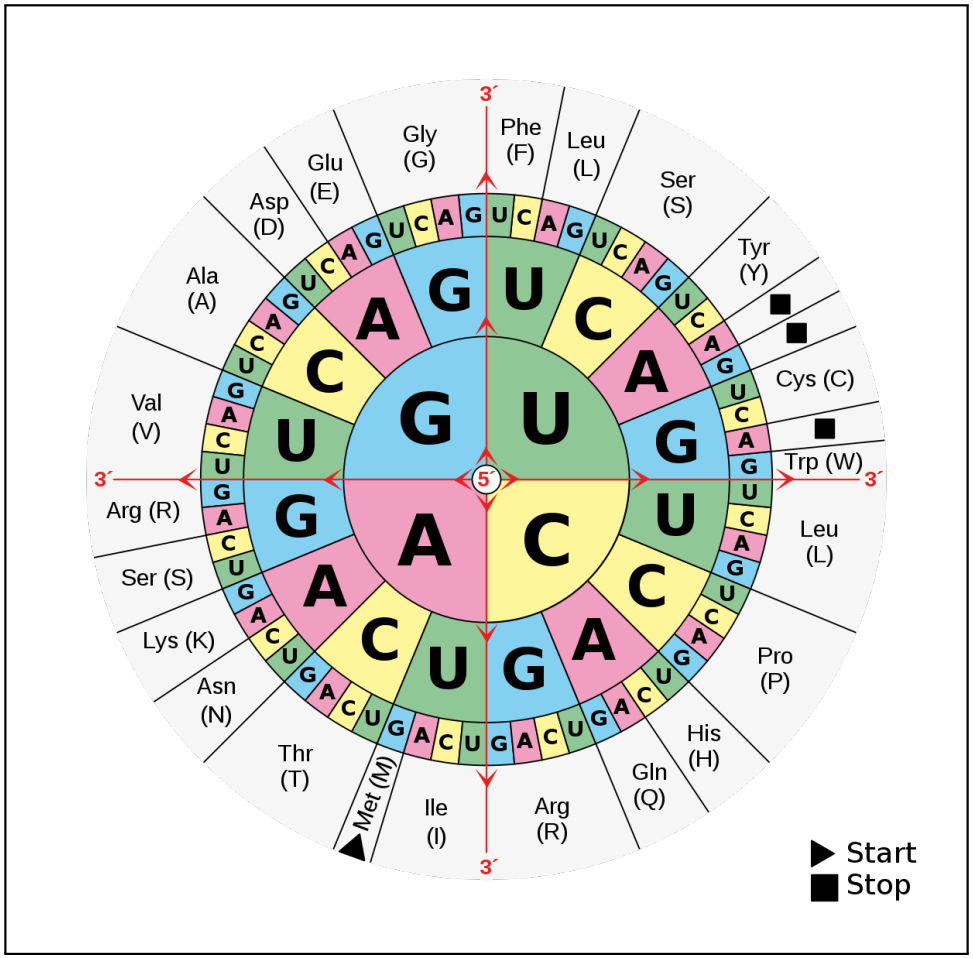
The sequence of a protein in FASTA format

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

The sequence of a gene in FASTQ format

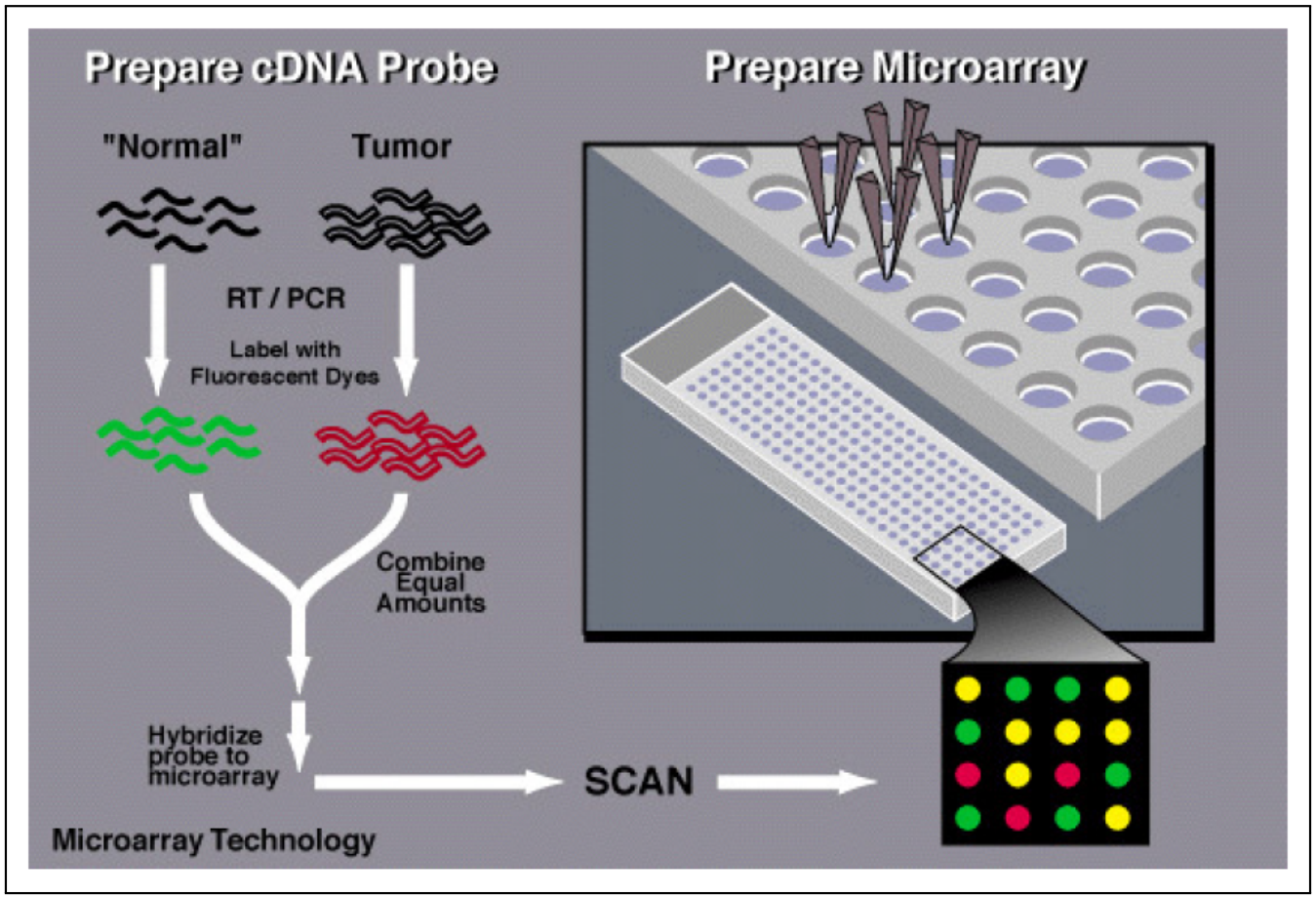
Note: In contrast to FASTA sequences, which includes only the identifier and sequence lines of size 50, a FASTQ sequence additionally takes the quality scores.

The codon table



The conversion from RNA codons to proteins

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

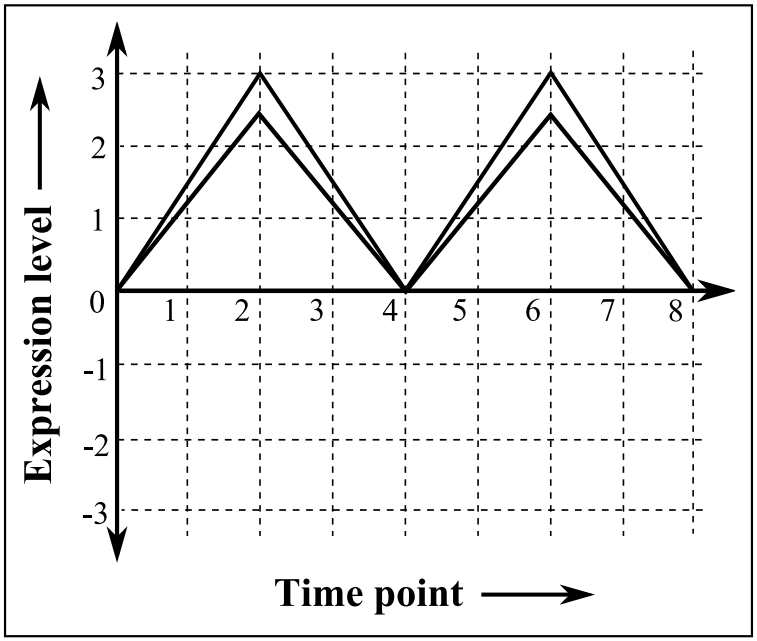
The rows and columns in an expression dataset can be different biomolecules (genes, proteins, microRNAs, etc.) and samples (time points, tissues, patients, etc.), respectively.

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

Co-expression

Definition (Co-expression)

Pairwise similarity pattern (spatial or temporal) of expression vectors.



Co-expression between a pair of expression vectors

Differential expression

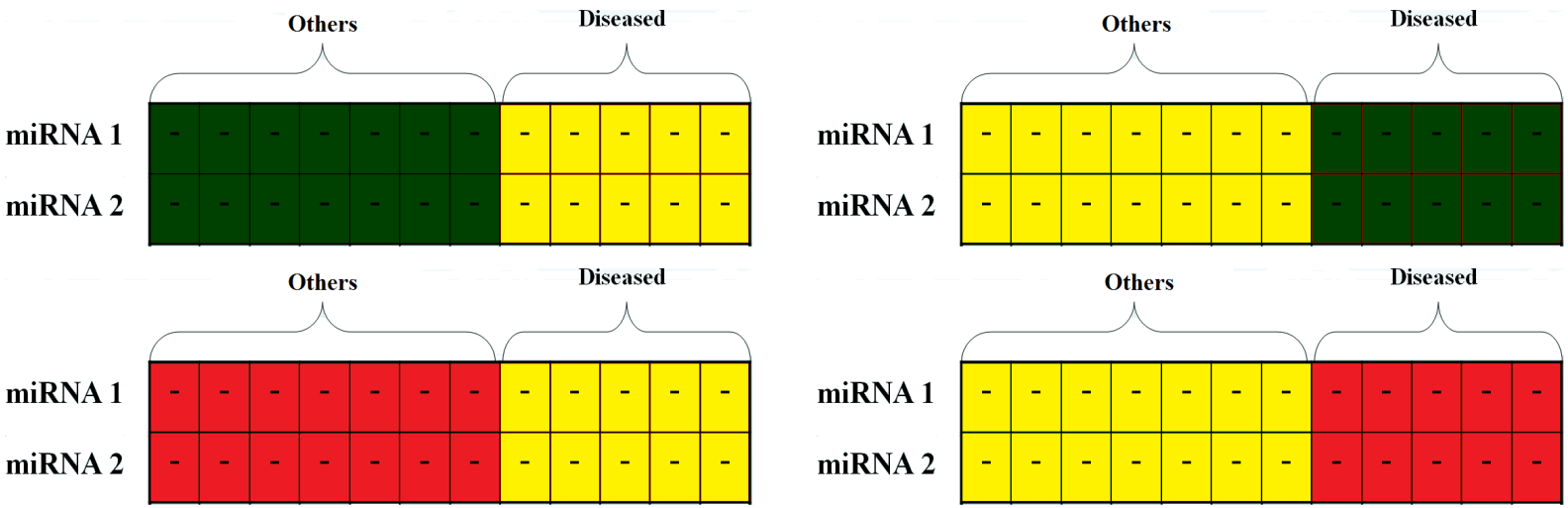
Gene	ID	c1	c2	c3	c4	n1	n2	n3	n4	n5	n6
G1	...	1.2	1.6	1.8	1.1	1	2	1.3	4	2	1.1
G2	...	1.1	1.5	1.3	1.8	2.1	1.1	1.1	1.1	2.3	1.5
G3	...	1.2	1.7	1.8	1.1	2	1.1	2.1	0.8	1.1	1.9
...	...	...	...	...	...	...	...	...	...	...	...
G100	...	2.1	1.6	1.7	1.4	2.3	2.7	2.8	2.9	1.3	2.1

Differential expression denotes varying patterns (spatial or temporal) of expression vectors between different phenotypes.

Differential co-expression

Definition (Differential co-expression)

Pairwise varying dependence (spatial or temporal) between expression vectors in different phenotypes.



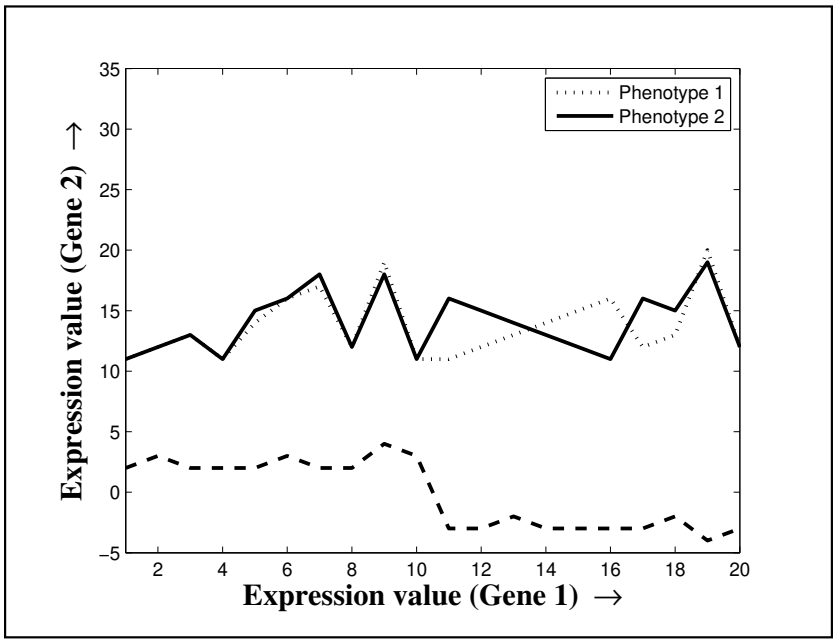
Different types of differential co-expression patterns

It can be mainly of two types – on/off case and gap/subsitution case.

Co-expression dynamics

Definition (Co-expression dynamics)

Pairwise varying dependence (spatial or temporal) between expression vectors based on another expression vector.



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)

Hands-on

1. Download the reference genome of Monkeypox, the virus attributed to Mpox, from NCBI following the steps below:
   i) Open the NCBI portal.
   ii) Select "Assembly" and search for "Monkeypox".
   iii) Download the complete reference genome sequence (RefSeq) in `FASTA` format.
   iv) Find out the frequencies of k-mers (k = 1, 2, 3) in the reference genome.
   v) Identify the proteins that are already characterized in Monkeypox.
   vi) Figure out from where you can download genome and protein sequences, annotation and a data report for all complete genomes of Monkeypox.