Machine learning methods for bioinformatics

INFO-F-528

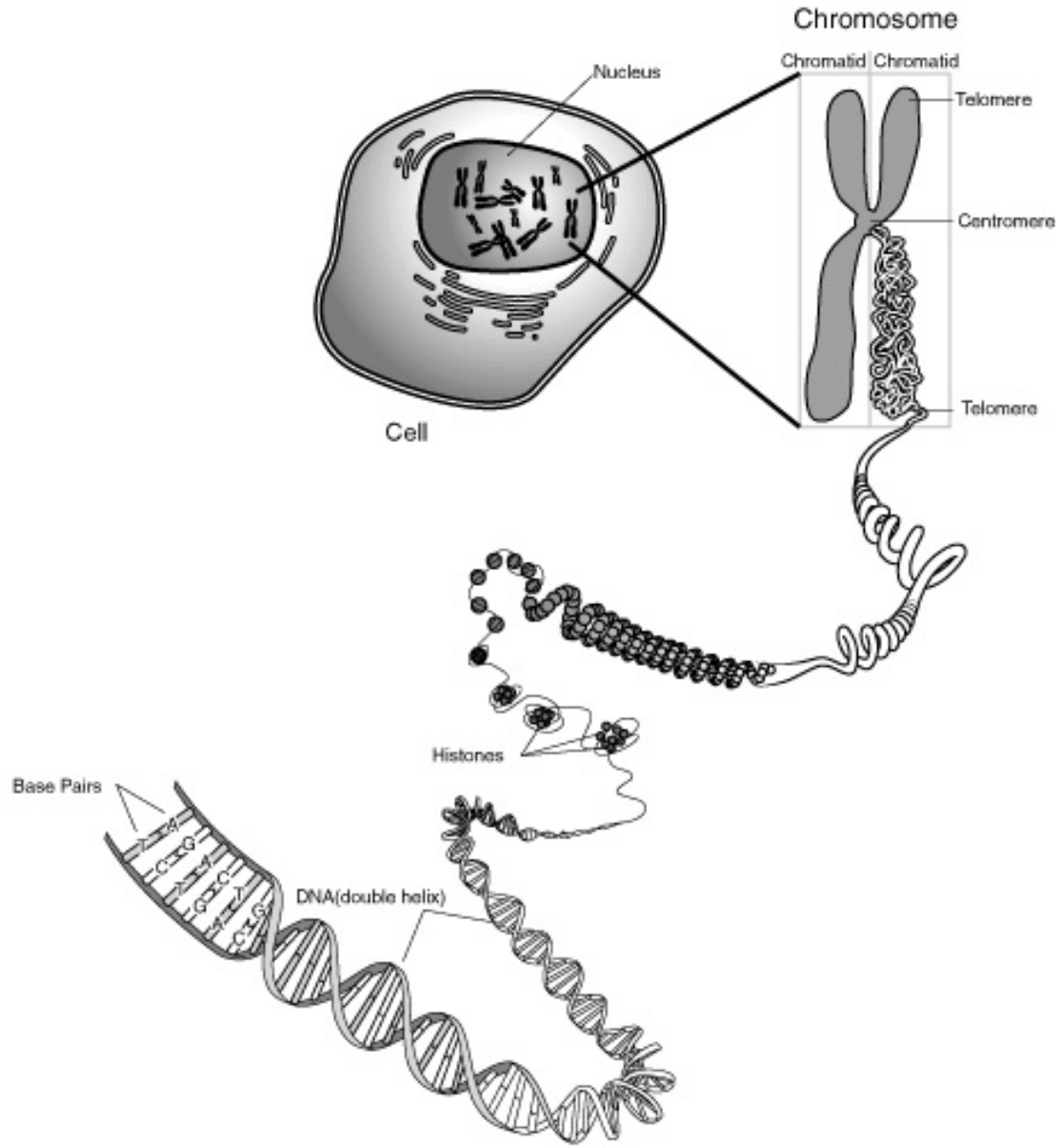
Gianluca Bontempi

ULB Machine Learning Group, Département d'Informatique
Boulevard de Triomphe - CP 212
http://mlg.ulb.ac.be

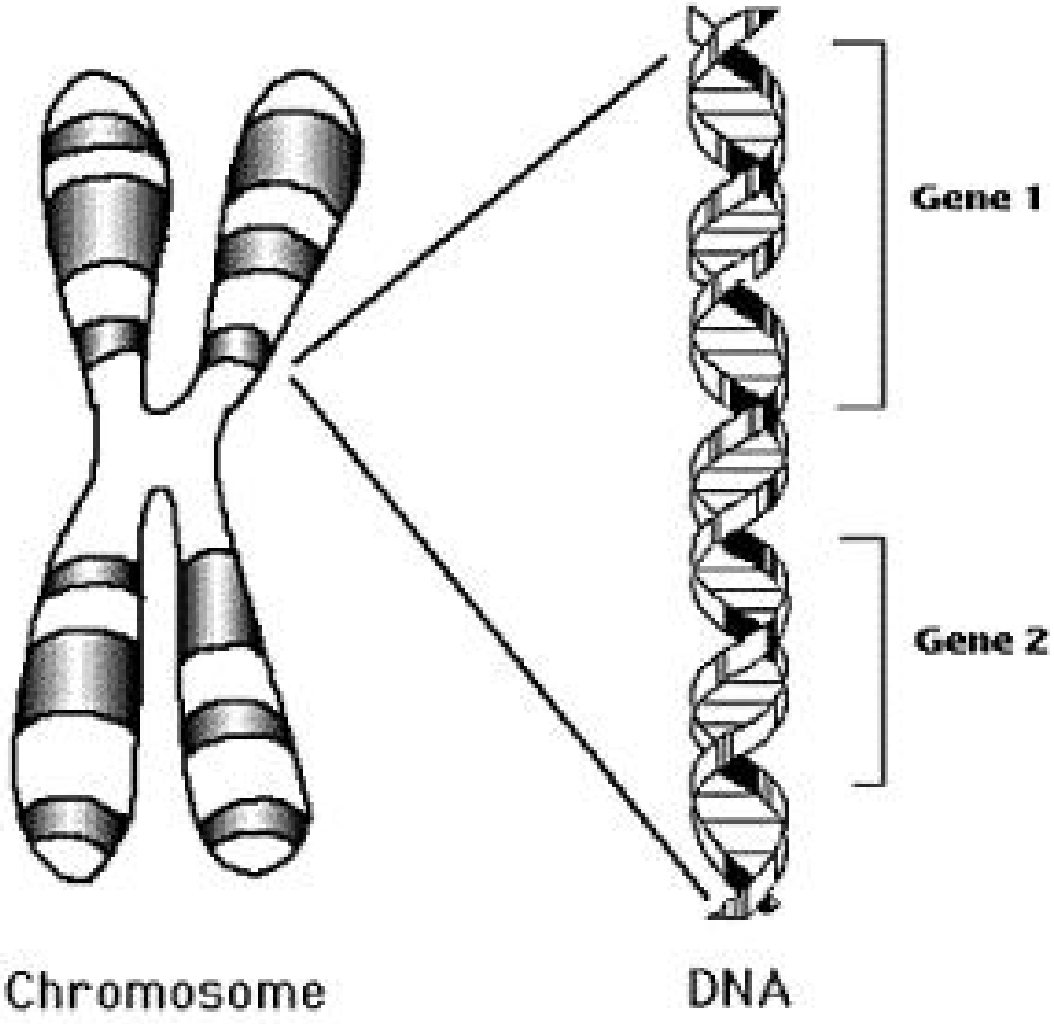
Some basic notions

- *Genes* are the carriers of the genetic information stored as a code of four letters.They normally reside on a stretch of DNA that codes for a type of protein that has a function in the organism.
- They are physically embodied within complex DNA macromolecules that lie within structures called *chromosomes* which occur in every living cell.
- In humans there are forty-six chromosomes. All but two of these (the sex chromosomes) occur in pairs of "homologous" chromosomes.
- The total content of the DNA molecules within the chromosomes is called the *genome* of an organism. Within an organism, each cell contains a complete copy of the genome. The human genome contains about three billion base pairs and about 35,000 genes.

Cell's nucleus

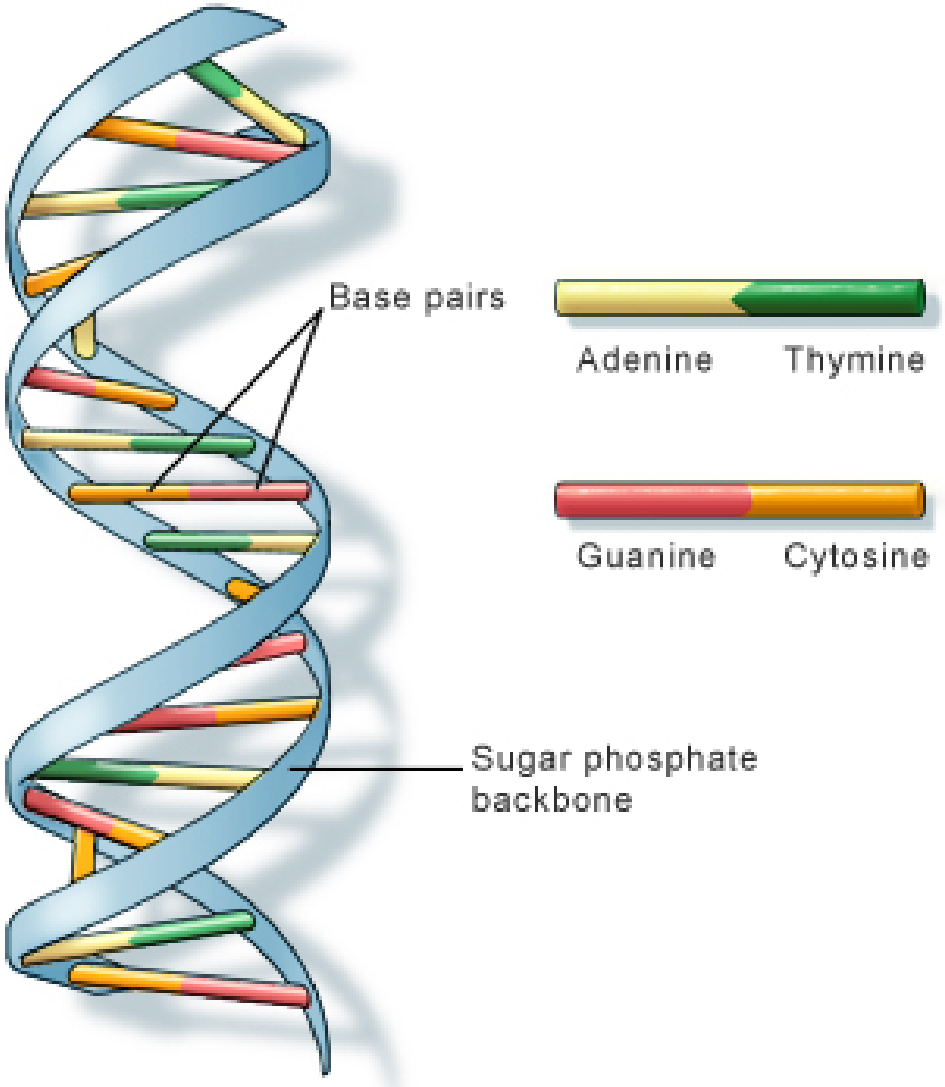


Chromosomes



Genes

The DNA structure

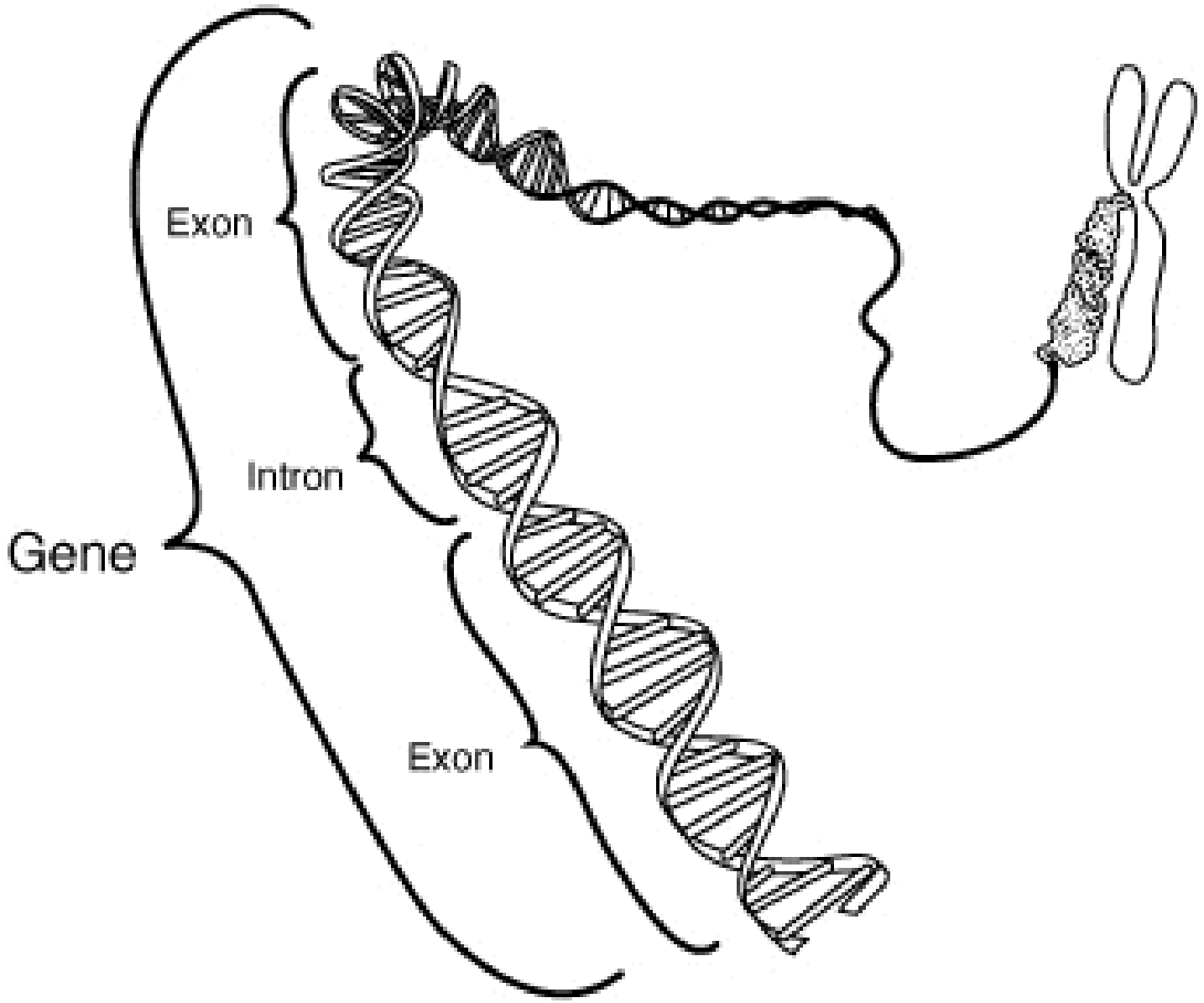


U.S. National Library of Medicine

DNA

- Watson and Crick showed that a DNA molecule is a double helix consisting of two strands.
- Each helix is a chain of bases, chemical units of four types: A, C, T, and G.
- Each base on one strand is joined by a hydrogen bond to a complementary base on the other strand, where A is complementary to T, and C is complementary to G.
- Thus the two strands contain the same information.
- Genetic information is encoded digitally, as strings over the fourletter alphabet A, C, T, G, much as information is encoded digitally in computers as strings of zeros and ones.

A gene



Proteins

- Proteins are the workhorses of cells. They act as structural elements, catalyze chemical reactions, regulate cellular activities, and are responsible for cellular structure, producing energy, communication between cells.
- A protein is a linear chain of chemical units called *amino acids*, of which there are twenty common types.
- The function of a protein is determined by the three-dimensional structure into which it folds.
- Amino acids codes:
  $A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V$.

The molecular biology dogma

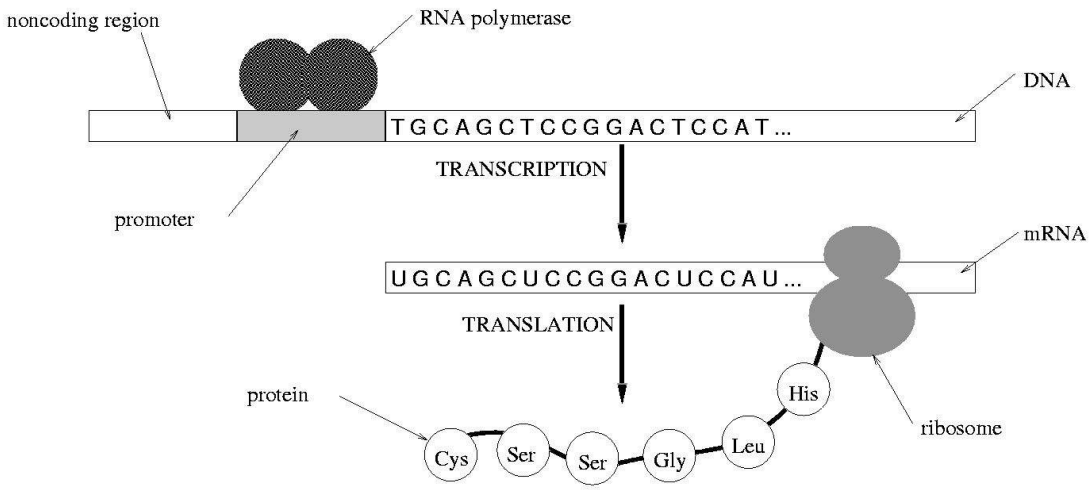


Figure 1: **The process of gene expression.** DNA and RNA are composed of linear chains of nucleotides. *Transcription* involves synthesizing messenger RNA (mRNA) using DNA as a template. The enzyme *RNA polymerase* is the molecule that transcribes DNA into RNA. A site where RNA polymerase binds to DNA to begin transcription is called a *promoter*. Messenger RNA is *translated* to protein by a molecule called a ribosome. Proteins are linear chains of amino acids; each amino acid is encoded by a string of three consecutive nucleotides. *Noncoding regions* are stretches of DNA that do not encode proteins.

From Genes to Proteins

- The fundamental dogma of molecular biology is that DNA codes for RNA and RNA codes for protein.
- Thus the production of a protein is a two-stage process, with RNA playing a key role in both stages.
- An RNA molecule is a single-stranded chain of chemical bases of four types: A, U, C, and G.
- In the first stage, called transcription, a gene within the chromosomal DNA is copied base-by-base into mRNA (messenger RNA) according to the correspondence $A \rightarrow U$, $C \rightarrow G$, $T \rightarrow A$, $G \rightarrow C$.
- The resulting mRNA transcript of the gene is then transported within the cell to a molecular machine called a *ribosome* which has the function of translating the RNA into protein.

From Genes to Proteins

- Translation takes place according to the genetic code, which maps successive triplets of RNA bases to amino acids.
- With minor exceptions, this many-to-one function from the sixty-four triplets of bases to the twenty amino acids is the same in all organisms on Earth.
- One of the main problems in science is the protein folding problem of predicting the three-dimensional structure of a protein from its linear sequence of amino acids.
- This problem is far from being solved, although progress has been made by a variety of methods.

What triggers genes ?

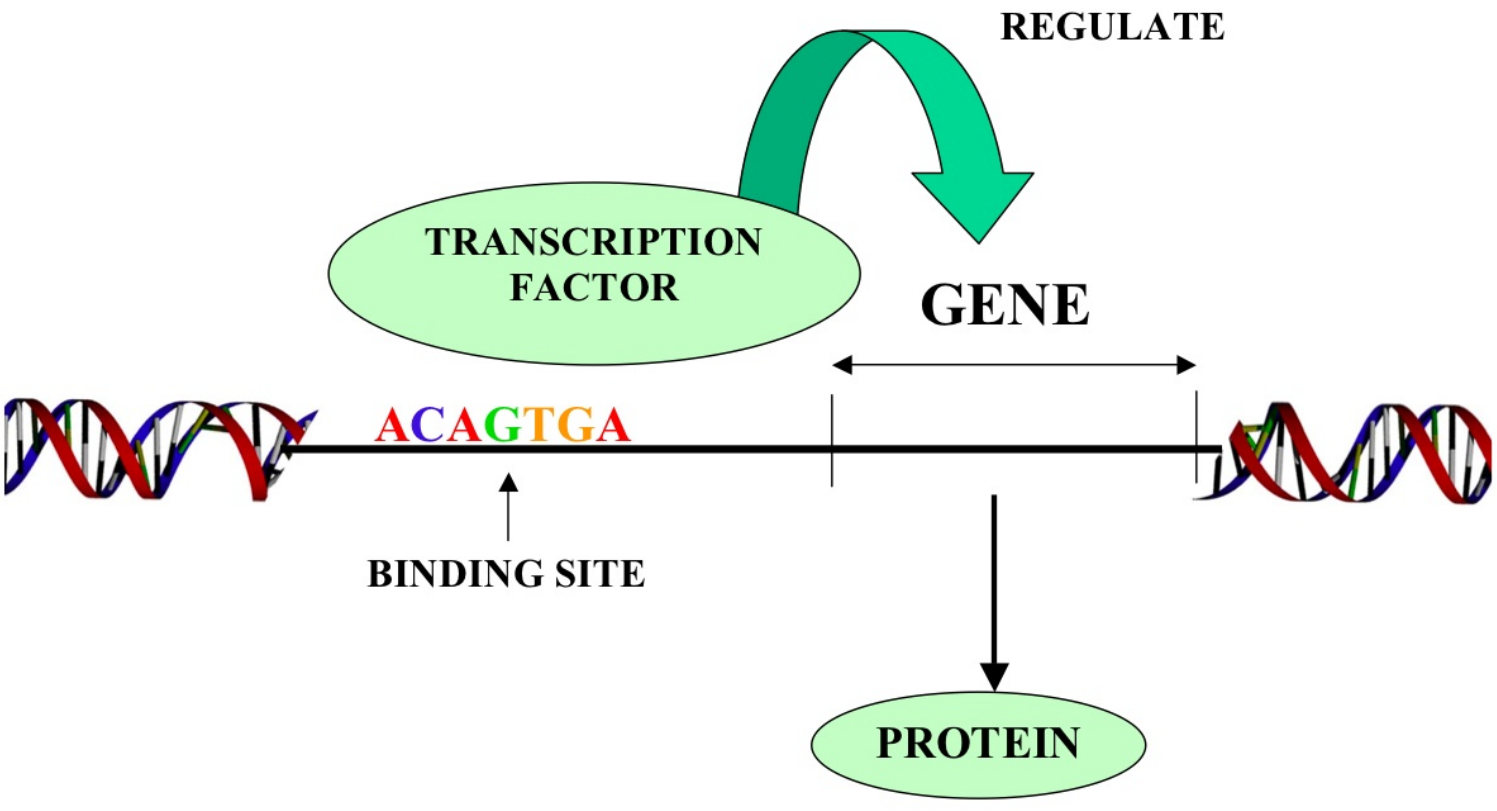
- All the cells within a living organism (with the exception of the sperm and egg cells) contain nearly identical copies of the entire genome of the organism.
- Thus every cell has the information needed to produce any protein that the organism can produce.
- Nevertheless, cells have remarkably distinct properties (e.g. difference between human eye cells, hair cells and liver cells) and differ radically in the proteins that they actually produce.
- For example, there are more than 200 different human cell types, and most proteins are produced in only a subset of these cell types.
- Moreover, any given cell produces different proteins at different stages within its cycle of operation, and its protein production is influenced by its internal environment and by the signals impinging upon it from other cells.

Regulation of Gene Expression

- The expression of a gene within a cell (as measured by the abundance and level of activity of the proteins it produces) is *regulated* by the environment of the cell.
- Transcription of a gene is typically regulated by proteins called *transcription factors* that bind to the DNA near the gene and enhance or inhibit the copying of the gene into RNA. These transcription factors are coded for by other genes - so **genes are often controlled by the activities of other genes**.
- Similarly, translation can be regulated by proteins that bind to the ribosome.
- Certain post-translational processes, such as the chemical modification of the protein or the transport of protein to a particular compartment in the cell can also be regulated so as to affect the activity of the protein.

TF binding

Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.



Regulation of Gene Expression(II)

- Thus gene expression can be viewed as a complex network of interactions involving genes, proteins, and RNA, as well as other factors such as temperature and the presence or absence of nutrients and drugs within the cell.
- whereas it remains difficult to measure the abundances of a cell's proteins, the DNA microarray makes it possible to quickly and efficiently measure the relative representation of each mRNA species in the total cellular mRNA population.

From genetics to genomics

- Genetics is the study of single genes in isolation. Genomics is the study of all the genes in the genome and the interactions among them and their environment(s).
- In studying human disease, for example, genomics examines all the genetic information to determine biological markers predisposing an individual to disease, whereas genetics uses the information from one or two genes to explain a disease state.
- Many diseases due to single gene defects have been identified. Now, geneticists want to tackle multifactorial diseases caused by the complex interactions between multiple genes and the environment.
- Genomics and genetics blend into one another so it is difficult to draw a decisive dividing line between them. Both disciplines are studying DNA to unravel nature's mysteries. But it does help to understand the genetic concepts before you tackle the more complex genomic ideas.

Microarrays

- Microarray are assays for quantifying the type and amount of mRNA transcripts present in a collection of cells.
- The number of mRNA molecules derived from the transcription of a gene is an approximate estimate of the activity of the gene.
- RNA is extracted from the specimen and the mRNA is isolated.

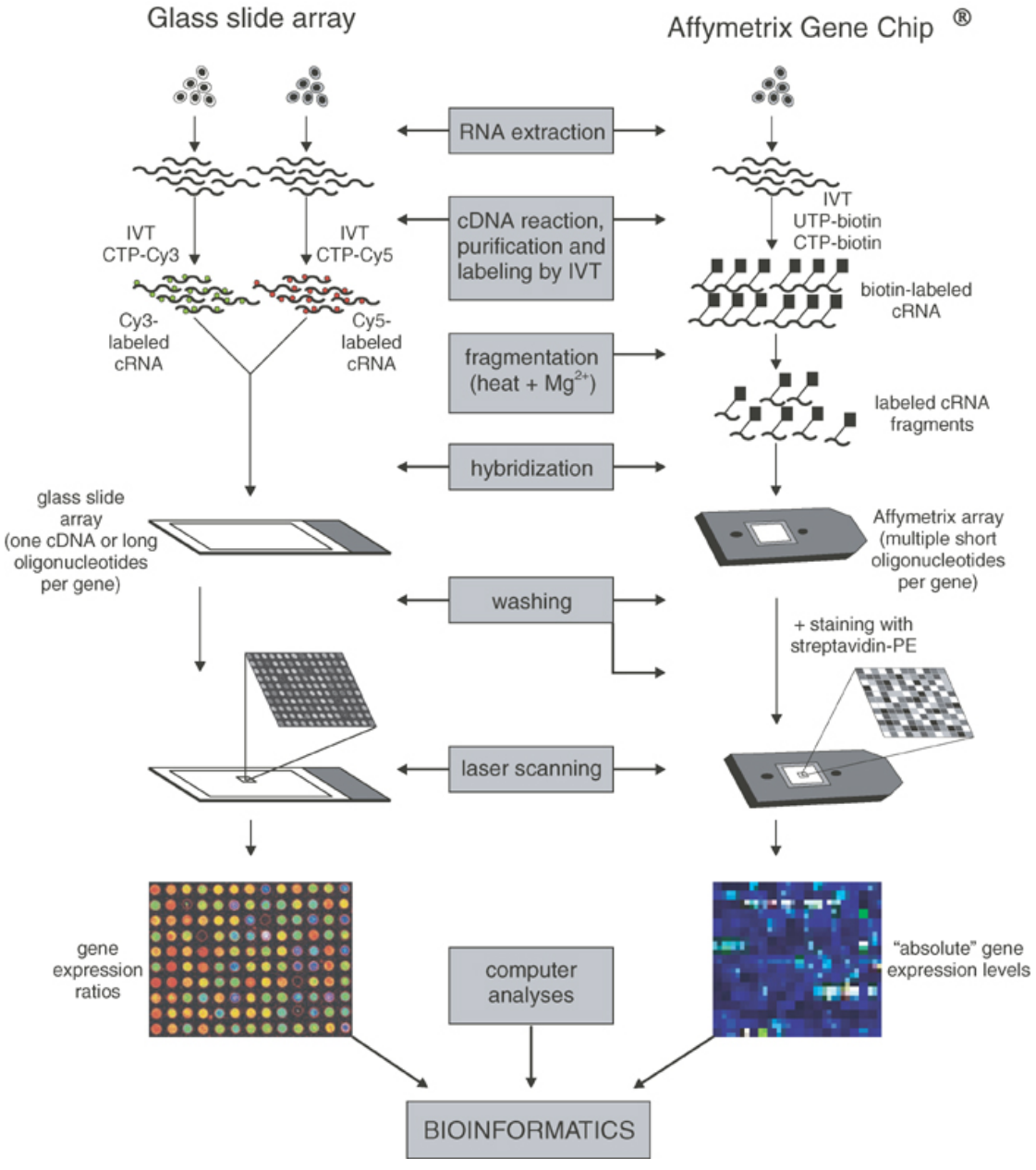
Microarrays (II)

- Microarray have many *single* strands of a gene sequence attached to their surface, known as *probes*.
- The attachment is sometimes achieved by physically spotting them on the array (as in cDNA microarrays) and sometimes by immobilizing them to the quartz wafer surface via hydroxylation.
- The single strands wait for complementary strands to bond (*hybridize*) and stick to the surface of the array.
- Chemically speaking, RNA is similar to a single strand of DNA.
- The purpose of a microarray is to measure for each gene in the genome the amount of message that was broadcast through the RNA.
- Simultaneous quantification of expression allows to evolve beyond the one-gene-at-a-time paradigm.

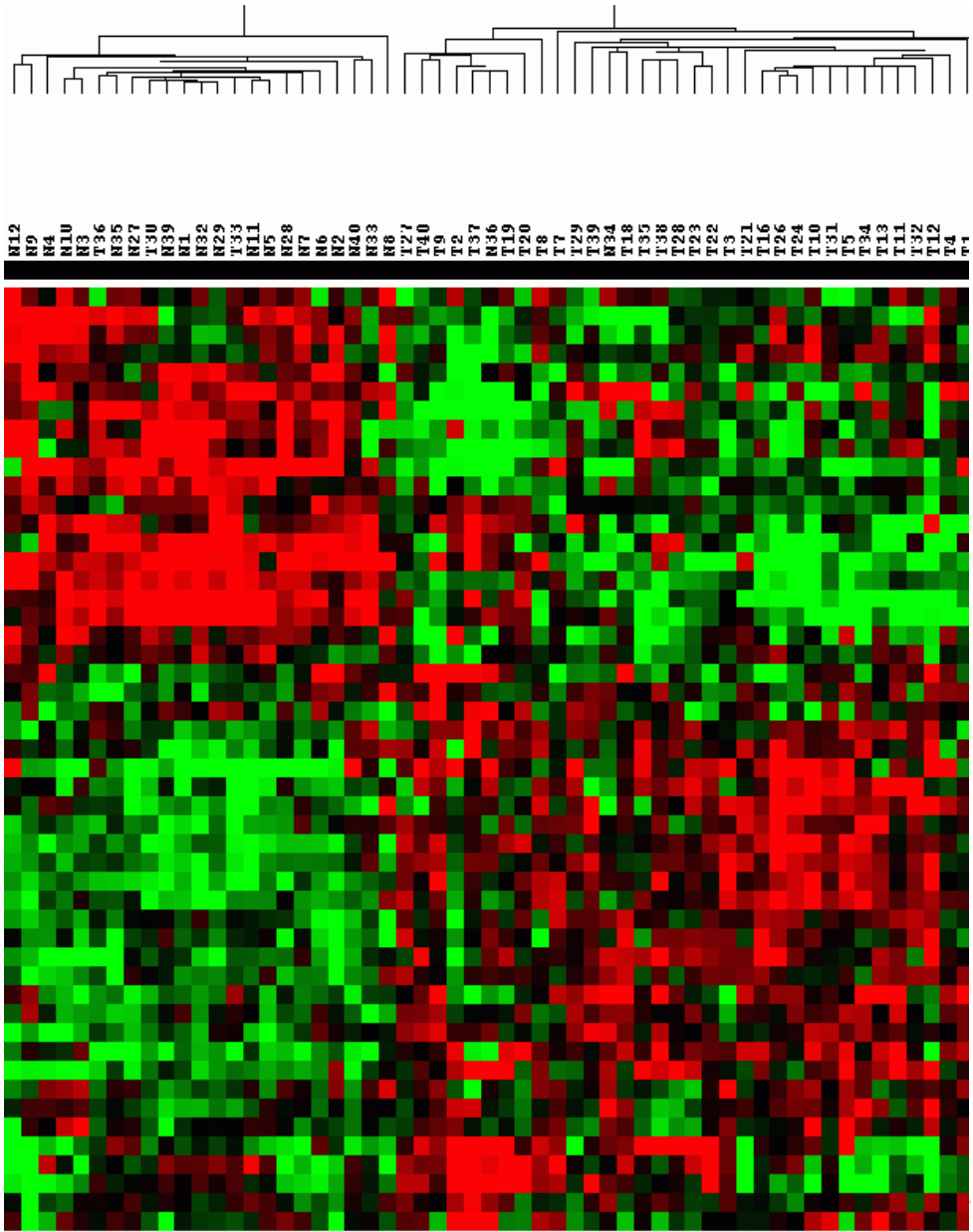
Microarray technology

- There are two major kinds of microarrays: oligonucleotide and cDNA array.
- In an oligo array a single gene is represented by several sequences (e.g. corresponding to exons) or probes. Typically there are about 20 probe pairs (aka probe set) for each gene. The perfect match (pm) probe is designed to match a small subsequence of the gene about 25 bases long. The mismatch probe is a control being identical to pm except with the middle base flipped to its complement.
- In a spotted cDNA microarray one base sequence matching all or part of a gene is printed on a glass slide. Colour-labelled RNA is applied to the microarray and if the RNA finds its complementary strand on the array, it naturally binds and sticks to the array. By measuring the amount of color emitted by the array, one can get a sense of how much RNA was produced for each gene.

Microarrays: 2 platforms



Microarrays



Sources of variations

- Sources of variation in mRNA
  - Differences in experimental conditions (e.g. room temperature, temperature, time of the day)
  - Differences between experimental subjects.
  - Differences between samples from the same subject.
  - Variations in mRNA extraction methods from original sample
- Source of variation in the microarray production
- Source of variation in the hybridization process
- Source of variation in the scanning

Low-level processing

This phase aims to derive from the set of intensities of the probes a single value representing the abundance of the transcript (gene or exon). It is typically composed of three steps:

- background adjustment: it aims to make a correction for background noise and nonspecific binding. There are indications that this step has the largest effect on the preprocessing results.
- normalization: it aims at reducing the nonbiological variation and generally makes the assumption that only a relatively small number of genes is differentially expressed.
- summarization: it combines the probe set values in a single figure

Low-level processing methods

The most common preprocessing methods are

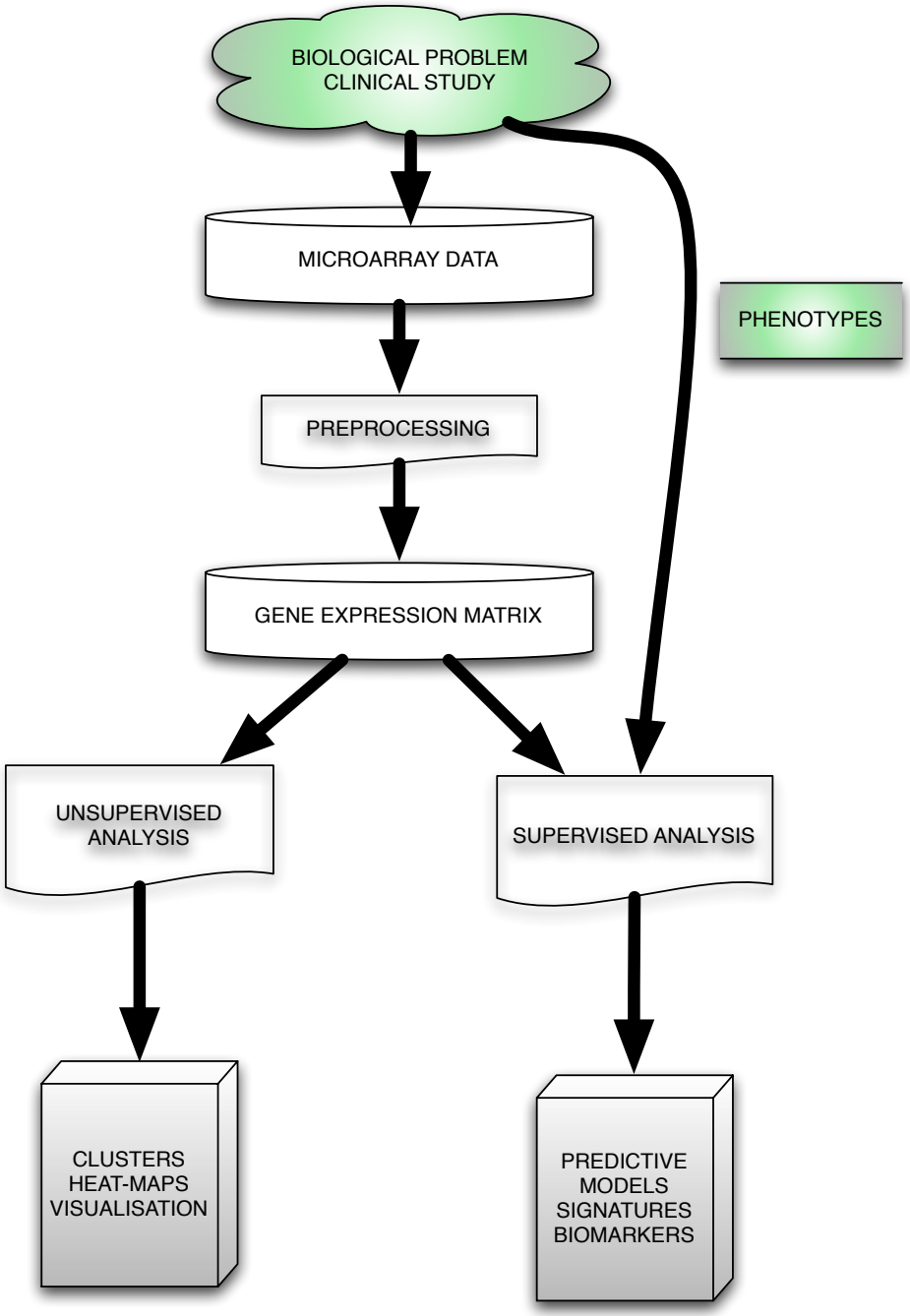
- MAS5
- RMA
- GCRMA

The output of low-level processing is the gene expression matrix.

Gene expression matrix

	1053_at	117_at	1294_at	...	...	91684_at	91703_at	91953_at	Class
Patient 1	93.7	123.4							A
Patient 2	105.2	80.1							C
Patient 3	...								B
Patient 4									A
Patient 5	77.3								A
Patient 6	22.4								C
Patient ...	2.2								
Patient 89	11								A
Patient 90	34								C

The columns represent probesets or genes. Names like 1053_at are Affymetrix probe set IDs, which correspond to genes and whose annotation can be downloaded from affymetrix.com.
Classes represent different phenotypes, (e.g. types of diseases, subtypes of cancer).


BIOLOGICAL PROBLEM
CLINICAL STUDY
MICROARRAY DATA
PHENOTYPES
PREPROCESSING
GENE EXPRESSION MATRIX
UNSUPERVISED
ANALYSIS
SUPERVISED ANALYSIS
CLUSTERS
HEAT-MAPS
VISUALISATION
PREDICTIVE
MODELS
SIGNATURES
BIOMARKERS