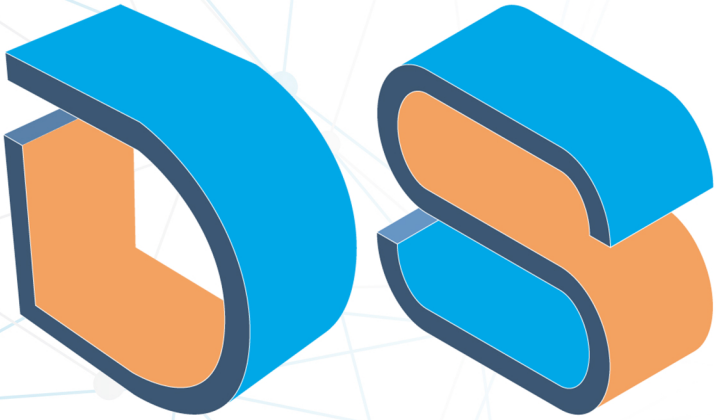


POLIMI
DATA SCIENTISTS

Bioinformatics and Computational Biology

Course Notes

Edited by:
Théo Saulus

Credits

The following notes have been written by Théo Saulus in collaboration with the Polimi Data Scientists student association by combining Prof. Marco Masseroli lectures, the slides of the course and personal notes.

They are meant as a support for the students following the course and they should not be considered as a replacement for the professor's lectures and materials. This document has not been reviewed by the professor and should therefore be used carefully.

Bioinformatics and Computational Biology

Lecture notes

- Marco Masseroli (marco.masseroli@polimi.it)
  - Floor 1, DEIB building 20, Leonardo, office 055
  - tel.: 02 - 2399 3553
  - http://www.deib.polimi.it/ [people/alphabetic list/...]
    http://www.bioinformatics.deib.polimi.it/masseroli/

- Silvia Cascianelli (silvia.cascianelli@polimi.it)

Table of content

I. Introduction (14 Sept.)

I.A Course organisation

The course aims to illustrate how **computer science principles**, **technologies**, **methods** and **instruments** can be profitably used for the computational analysis, information content increment and interpretation of biological data produced by genome sequencing, gene expression measurements, proteomics and cellular metabolic flow quantifications

It will be highlighted as the application to biological data of the engineering themes of **data bases**, **information theory**, **data** and **text mining** and others can contribute to increasing biomedical knowledge and improving health care

The course main objective is to give a systematic view of this **multidisciplinary** sector and to provide students with the base knowledge and instruments required to tackle various issues in computational biology and to take advantage of the opportunities offered by the recent bioinformatics development

Prerequisite:

- None; biological and biochemical concepts required to understand motivations and aims of the bioinformatics computational methodologies presented during the course will be introduced in the first part of the course

Syllabus:

Seminar lectures and practices in informatics room on the following topics compose the course; in case, at the end of the course, an external technical visit to an experimental research laboratory, or a seminar lecture by a field expert, will take place.

- *Introduction* (2 hours): definitions, methodologies and motivations

- *Genetic and molecular biology concepts* (8 hours): organisms, cells, biological molecules and their structure, duplication and expression of genetic information, protein synthesis, structure of genes and transcripts, hints of protein structure, genome, transcriptome, proteome, hints of hereditary pathologies

- *Techniques of biomolecular sequence analysis* (6 hours): importance of biological sequence comparison, local or global alignment of two biomolecular sequences, sequence similarity search

- *Technologies for gene expression measurement and analysis* (2+4 hours): biotechnologies for gene expression measurement, computational methods for gene expression data analysis, data mining of gene expression data

- *Biological network analysis* (2 hours): main characteristics of a biological network, mining and visualization of complex network features, computational methods for gene network extraction and analysis

- *Bio-terminologies, bio-ontologies and methodologies for their analysis* (2+2 hours): functional and phenotypic annotations of genes and proteins, controlled vocabularies for genomic and proteomic annotation, Open Biomedical Ontologies: the Gene Ontology and other bio-ontologies, enrichment and similarity analysis of annotations

- *Genomic and proteomic databanks* (2 hours): databank types and access methodologies, main databanks and their relations, provided data and formats, search methods in databanks, integration and update of data and information

- *Examples of available bioinformatics tools* (20 hours): main software tools available as Web applications, Web services and freeware and open source programs

www.bioinformatics.deib.polimi.it/masseroli/BCB

- **Lessons**: Tuesday, 11,15-14,15 online (Cisco Webex)
- **Practices**:
  - Monday, 17,15-19,15 room 4.0.1
  - Or (rarely) Tuesday, 17,15-19,15 room 7.1.3
- Detailed schedule:
  http://www.bioinformatics.deib.polimi.it/masseroli/BCB/
  Schedule_BCB_2021_2022.pdf

- **Written test**:
  - Some questions about any of the course subjects (given in lessons or practices), to be answered in free text
    More details in "Examination procedure" and "Example of written test questions" on course web site:
    http://www.bioinformatics.deib.polimi.it/masseroli/BCB/
    exams/BCBexamProcedure.pdf
    http://www.bioinformatics.deib.polimi.it/masseroli/BCB/
    exams/BCB_written_test_example_questions.pdf

I.B Definitions

I.B.1 Bioinformatics and Computational Biology

What is Bioinformatics?

The definition of bioinformatics is not universally agreed upon. We define it as the creation and development of advanced information and computational technologies for problems in biology, most commonly molecular biology

Definition of Institut Pasteur:

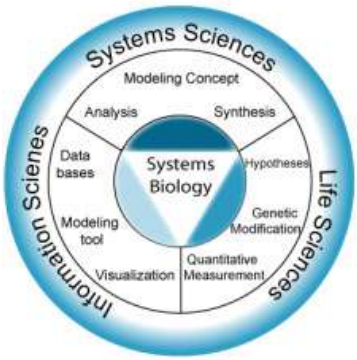
- Bioinformatics derives knowledge from computer analysis of biological data
- Research in bioinfomatics includes method development for storage, retrieval, and analysis of the data
- Bioinformatics uses techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics and linguistics

The NIH (National Institutes of Health, US Dept. of Health and Human Services) Biomedical Information Science and Technology Initiative Consortium definitions:

- *Bioinformatics*: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organize, archive, analyse, or visualize such data
- *Computational Biology*: The development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems

I.B.2 Systems Biology

Systems biology is the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behaviour of that system.



As the objective is a *model of all the interactions in a system*, the experimental techniques that most suit systems biology are those that are system wide. Therefore, *high-throughput techniques are used to collect quantitative data* for the construction and validation of models.

I.B.3 Biomedical Informatics

The increasing convergence of biology, medicine and genetics provides the opportunity to leverage synergies between the three areas for the benefit of health. Biology and genetics are emerging as information sciences while medicine is increasingly adopting information systems and informatics approaches to support healthcare delivery.

Biomedical Informatics is viewed as the discipline that aims to bring together the domains of bioinformatics and medical/health informatics to further the discovery of novel diagnostic and therapeutic methods.

I.B.4 Translational Bioinformatics

AMIA (American Medical Informatics Association) refers to translational bioinformatics as the development of storage, analytic, and interpretative methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health

Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is newly found knowledge from these integrative efforts

http://www.amia.org/applications-informatics/translational-bioinformatics/

I.C Methodologies

I.C.1 Information technology role

From the definitions, it is clear the important role of the Information and Communication Technologies (ICT) in Bioinformatics and Computational Biology to:

Management (store, integrate, query, search, retrieve, …) and analysis of data and information (knowledge)

- Development of models and algorithms
- Implementation of instruments and services (infrastructures)
- Creation of visualization tools
- Computational and systemic approach
- …

Bioinformatics and Computational Biology are not simply another applicative domain of the ICT, but they are disciplines born thanks to the ICT development

They are also disciplines where the relevant ICT contribution can appear only if based on a background (at least minimum) of biological knowledge, since their goal is not only the correct and efficient execution of ICT methods, but the answer to biological questions

In such disciplines computer science shows all its relevance in contributing to the progress of the life science and health care knowledge

I.C.2 Interdisciplinarity

Bioinformatics and Computational Biology are collaborative, interdisciplinary, and globalized activities:

- "World", not national, disciplines
- English is the common language
- Relevant results only in work groups with multiple expertise (informatics, engineering, physical, chemical, biological, medical, social, …)

Need of:

- Using a simple language, clear also to people with different expertise (during the exam, always define the word that could be badly interpreted, and **acronyms**)
- Acquiring base knowledge of other disciplines, to be able to productively collaborate with other expertise people

I.D Motivations

I.D.1 Human Genome Project

Modern Bioinformatics and Computational Biology were born with the DNA sequencing projects

- Sequencing of human DNA was first proposed in 1984
- The Human Genome Project (HGP) started in 1990 as part of an international collaboration
- In June 2000 the public International Human Genome Sequencing Consortium and the private company Celera Genomics announced the completion of the first draft of the whole human DNA sequence
- First draft sequence completed in October 2000 and published in February 2001
- The primary goal of the HGP is to provide a complete, high-quality sequence of human genomic DNA to the research community as a freely, publicly available resource
- Additional goals include developing efficient technologies for gathering information leading to the collection, interpretation, and informed use of that sequence
- Interesting movies on Human Genome Project: *http://www.molecularlab.it/interactive/progetto%20genoma/index.asp*

Other specific HGP **goals** are:

1. *DNA sequencing technology*
   - After first DNA sequencing more DNA sequence required
   - Many two-fold improvements vastly improved cost effectiveness and throughput
   - "$100.000 Genome":
     × Raw measure of human genetic variation
   - "$10.000 Genome":
     × Sequencing of tumor genome collections
     × SNP and disease-associated mutations
     × Signs of natural selection within a population
   - "$1.000 Genome": Personal genome
2. *Human DNA sequence variation:*
   - determine and map common (and less common) variants
   - make the information available
   - develop algorithms for using this information
3. *Comparative genomics:*
   - for interpreting human genome sequence
     × functions of conserved sequences
     × support experiments in model systems
4. *Functional analysis* of genes, coding regions, proteins, and other functional elements of the genome on a high throughput, genome-wide basis:
   - collection of data using these technologies to the extent that resources allow
5. *Genome informatics:*
   - data analysis methods: sequence analysis, gene mapping, complex trait mapping, genetic variation, functional analysis
   - development of database tools
   - development and maintenance of databases of genomic and genetic data
6. *Training and career development:*
   - develop a cadre of new kinds of scientific specialists who can be creative at the interface of biology and other disciplines, such as computer science, engineering, mathematics, physics, chemistry, and the social sciences
7. *Ethical, Legal and Social Implications (ELSI)* of completion of the first human DNA sequence and of human genetic variation:
   - how to integrate this information into clinical, nonclinical, and research settings
   - interaction of this information with philosophical, theological, and ethical perspectives
   - examine how the understanding and use of genetic information are affected by socio-economic factors and concepts of race and ethnicity

After the publishing of the complete sequence of the human genome in 2001 start the "post-genomic era":

- The complete human DNA sequence is known
- Much must still be understood:
  × Which are **all** the genes in the DNA and where they start and end?
  × Which are **all** the DNA components?
  × How they **interact** to each other in a "systemic view"?
  × Which is their **function**; how it is altered in pathologies; how to correct such alterations?

I.D.2 Molecular medicine

Molecular medicine is a broad field, where physical, chemical, biological, and medical techniques are used to describe molecular structures and mechanisms, identify fundamental molecular and genetic errors of disease, and to develop molecular interventions to correct them

The molecular medicine perspective emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs

Therefore, molecular medicine is interconnected with and develops into **pharmacogenomics**, which in turn has molecular medicine as its application focus

I.D.3 Pharmacogenomic

Pharmacogenomics is the branch of pharmacology that deals with the influence of genetic variation on drug response in patients by correlating gene expression or single nucleotide polymorphisms with a drug's efficacy or toxicity. By doing so, pharmacogenomics aims to develop rational means to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects

Such approaches promise the advent of "**personalized medicine**", in which drugs and drug combinations are optimised for each individual's unique genetic makeup

Pharmacogenomics is the whole genome application of **pharmacogenetics**, which examines single gene interactions with drugs (*http://www.phgfoundation.org/tutorials/pharmacogenomics/*)

II. Concepts of Genetics and Molecular Biology

II.A Main concepts

II.A.1 Definitions

Genetics is the science of heredity of characters in living beings, that is the process through which specific traits are passed on to next generations. This heredity is controlled by genes

Biology is the science of structures, functions, and life's conditions of living beings

Molecular biology is a discipline of biology that studies living beings at the level of molecular mechanisms that are the base of their physiology; in particular it studies interactions between macromolecules (proteins and nucleic acids DNA and RNA)

II.A.2 Features of living beings

Living beings have ***7*** common traits that make them different from inorganic matter:

- **Complexity in structure**: organisms have complex structures (at macroscopic, cellular and intracellular level)
- **Organization**: organism structures are highly ordered and functional, both at macroscopic (animals and plants) and microscopic level (cellular and intracellular level)
- **Use of energy**: organisms can get energy from the environment (solar energy (plants), chemical energy from nourishment) in order to produce work (mechanical, chemical, electrical, …) to build and maintain their structure
- **Reproduction**: organisms can reproduce themselves by generating new living beings that unalterably continue basilar characteristics of the specie
- **Development**: in superior organisms, one cell can differentiate in many cells, characterized for each type of tissue
  - × Different types of cells create tridimensional ordered structures and proliferate in a controlled way
  - × Growth is made possible by nutrients incoming from external environment, that are transformed and absorbed
- **Reaction to stimuli**: living beings react to environmental stimuli to get the maximum advantage with the minimum expense of energy. E.g. motion can be the reaction: to the view of a predator, to the light direction (plants), to a climate change, …
- **Evolution**: according to the majority of biologist, living beings can evolve, that means they can transform themselves through changes in time

II.A.3 Scientific method

Natural sciences (physics, chemistry, biology, medicine, …) owe their actual development to **the scientific method**, summarized in the following 5 steps:

- **Definition of the problem**: careful observation of an aspect of reality (e.g. natural phenomena, living organisms) suggests to the scientist a question to solve
- **Study of literature**: before starting his personal research, the scientist analyzes scientific papers other researchers published about the topic
- **Formulation of the hypothesis**: depending on literature data and personal observation, the scientist formulates a hypothesis, as possible answer to the question
- **Conduct of experiments**: to verify the hypothesis (its correspondence to reality) the researcher carry out experiments that mimic in a simplified form and in controlled conditions the real phenomena

Experiments must be **reproducible.** Biologists examine the transformation or the behaviour of living beings; usually it is not highly reproducible due to the interaction of many factors. To reach better reproducibility:

- × Sample of organisms sufficiently substantial
- × Only one factor (the one examined) between the influent ones must be varied in the experiment
- × A control group must be compared and contrasted to the experimental one: the control sample is identical to the experimental one unless for the factor studied

In some cases (e.g. diseases' incidence) information is obtained by the observation of extended group of people, reviewed by statistical analysis

If experiments are not feasible, onerous (costs, time) or numerous, an **"in silico" analysis** can be carried out, to find answers or to select experiments that can give them

- **Formulation of law**: if the hypothesis is validated by the data, it is assumed as *law*; if not, another hypothesis is required

The law consists of a proposition that state the order, or the neatness found in the phenomena (generalization). ***Quantitative*** laws expressed with mathematical equations that allow **predictions** on phenomena not observed yet. Whole sets of laws can be explained with general principles: **theories**.

Nowadays in biology the most important unifying idea is the evolution theory. Laws and theories are **never fixed**! They must be modified or substituted if contrasting facts are observed.

II.A.4 The cell

The biological unit of all living beings is the ***cell***.

Organisms:

- unicellular (made up of only one cell)
- multi-cellular (made up of a variable number of cells); on average, human being have about 37.2 trillions cells

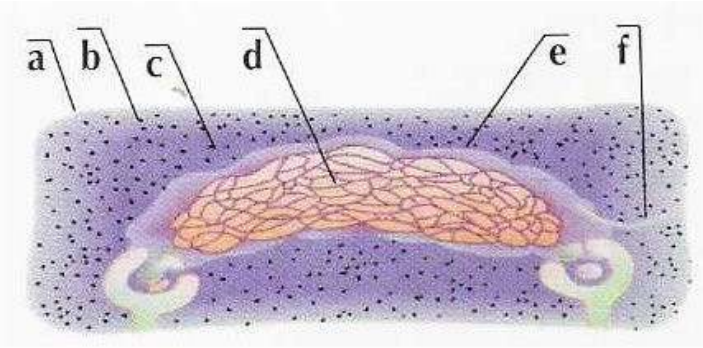
Depending on their structure the cells can be divided into:

- ***prokaryotes***
- ***eukaryotes***

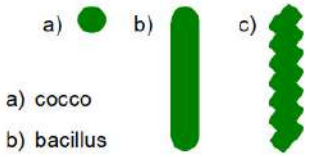
Prokaryote cells: have a nucleus not clearly separated from the rest of cellular matter

Schematized structure of a prokaryote cell (bacterium):

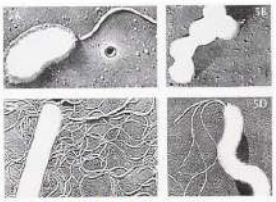
a) Cell wall
b) Plasma membrane
c) Cytoplasm
d) Chromosome
e) Ribosome
f) Flagellum



- **Prokaryotes**:
  - Are unicellular organisms, the most simple and old
  - They includes bacteria, the most numerous living beings on Earth, with different shapes: a) spherical, b) stick-shape, c) spiral-shape



a) cocco
b) bacillus
c) spirillum



Bacteria: A) vibrio; B) staphylococci; C) bacillus; D) spirillum with flagella

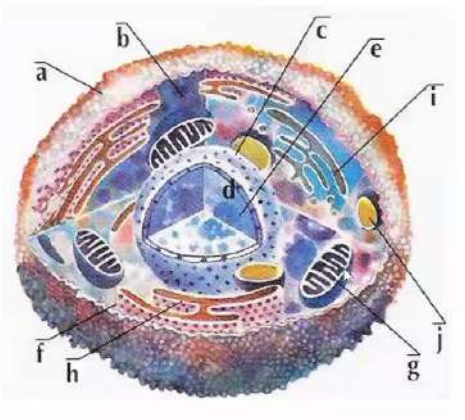
Eukaryote cells: have well-defined structure

They are enclosed by a plasma membrane, that contains:

- Undifferentiated cellular matter (cytoplasm)
- A well-defined nucleus, enclosed by a nuclear envelope

In cytoplasm, there are various organelles enclosed by membrane, that executes accurate vital function for the cell:

- mitochondria
- endoplasmic reticulum
- Golgi apparatus
- …



Schematized structure of a typical eukaryote cell:

a) Plasma membrane
b) Cytoplasm
c) Nuclear envelope
d) Nucleus
e) Chromosomes
f) Ribosome
g) Mitochondria
h) Endoplasmic reticulum
i) Golgi apparatus
j) Lysosome

II.A.5 Building bricks

All the cells are constituted by 4 main types of big biological molecules (macromolecules):

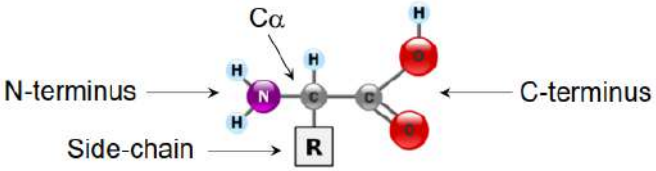
- **proteins**
- **polysaccharides**
- **lipids**
- **nucleic acids**

Proteins: polymers constituted by chains of amino acids (monomers); there are 20 different amino acids:

▪ Alanine	Ala (A)	▪ Methionine	Met (M)
▪ Cysteine	Cys (C)	▪ Asparagine	Asn (N)
▪ Aspartic acid	Asp (D)	▪ Proline	Pro (P)
▪ Glutamic acid	Glu (E)	▪ Glutamine	Gln (Q)
▪ Phenylalanine	Phe (F)	▪ Arginine	Arg (R)
▪ Glycine	Gly (G)	▪ Serine	Ser (S)
▪ Histidine	His (H)	▪ Threonine	Thr (T)
▪ Isoleucine	Ile (I)	▪ Valine	Val (V)
▪ Lysine	Lys (K)	▪ Tryptophane	Trp (W)
▪ Leucine	Leu (L)	▪ Tyrosine	Tyr (Y)

Amino acids are molecules constituted by:

- 1 central atom of carbon (C), linked with:
  - 1 atom of hydrogen (H)
  - 1 amino group ($-NH_2$)
  - 1 carboxylate group (-COOH)
  - 1 side-chain (R), that varies for each amino acid (Ala, Cys, Asp, …)



Amino acids (proteins) can have different functions:

- *structural* (they constitute the cell's frame)
- *enzymatic* (catalyze (allow) the majority of cellular reactions)
- *hormonal* (execute regulative functions)
- *immunity* (defense from other entities, mediated by the Immunoglobulines)

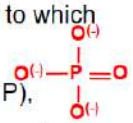
Polysaccharides: polymers constituted by various simple sugars (such as glucose); they include:

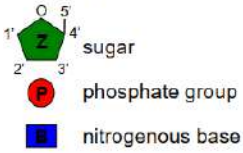
- *glycogen* (the most important energetic reservoir in animals for instant utilization)
- *amid* and *cellulose* (the most important energetic reservoir in plants)

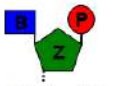
Lipids or **fats**: main constituents of cellular membrane, they represent energetic reservoir for the cell

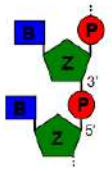
Nucleic acids: the most big molecules in cells; they are constituted by long sequences of monomers, the nucleotides, formed by:

- 1 molecule of sugar with 5 atoms of carbon (C) to which are linked:
  - 1 phosphate group (with 1 phosphoric acid - P),
  - 1 molecule containing nitrogen (N) (nitrogenous base, or base (B))



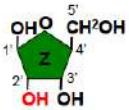


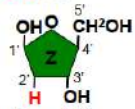

Nucleotide (monomer)



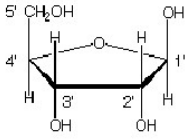
There are 2 types of nucleic acids:

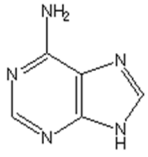
- **Ribonucleic acids**, or **RNA**, constituted by nucleotides with ribose as sugar
- **Deoxyribonucleic acid**, or **DNA**, constituted by nucleotides with deoxyribose as sugar
- DNA contains bases (nucleotides): **adenine** (A), **cytosine** (C), **guanine** (G), **thymine** (T)

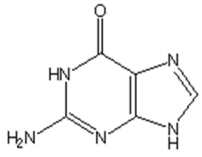


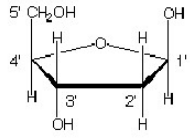


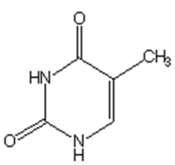
RNAs contain **uracil** (U) instead of thymine

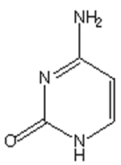

Ribose

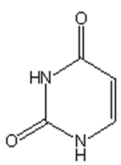

Adenine (DNA/RNA)


Guanine (DNA/RNA)


Deoxyribose


Thymine (DNA)


Cytosine (DNA/RNA)


Uracil (RNA)

DNA contains all the genetic information that is necessary for the life of the host organism. DNA gets organized in structures called ***chromosomes***

- In *prokaryotes* there is only 1 chromosome
- In *eukaryotes* there are more than one chromosome (there are n chromosomes, with n specific to each species; human beings have n = 23 chromosomes)

A cell can be *haploid, diploid, triploid, tetraploid*, … if it contains n, 2n, 3n, 4n, … chromosomes; most cells are diploid.

II.A.6 Cellular reproduction

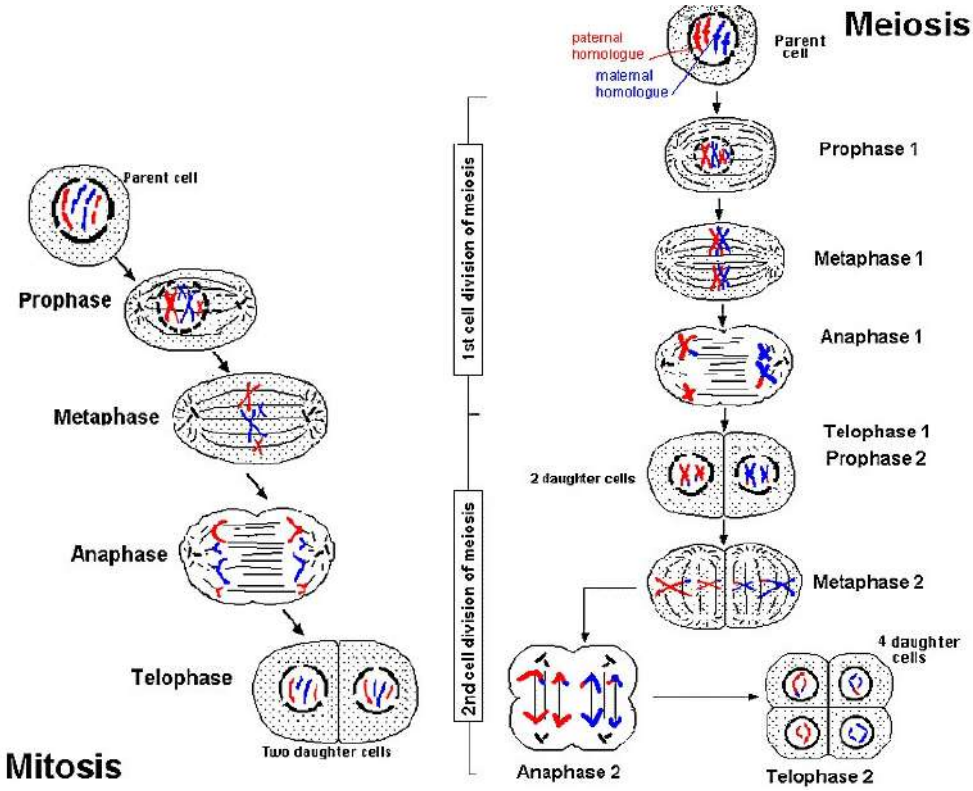
Cells reproduce themselves through cellular division (that is regulated by own genetic information). Time between two cellular divisions varies depending on the type of cell.

- In prokaryotes (e.g. bacteria) reproduction consists in simple binary fission (1, 2, 4, 8, …) until nutrient lasts
- In eukaryotes cells can reproduce themselves in 2 ways:
  - × **asexually** (like, but more complex that prokaryote binary fission)
  - × **sexually**, specific to the superior multi-cellular organisms

In ***sexual reproduction***, the new organism originates from the union (fecundation) of:

- one *female* sexual cell (germinal cell, or gamete), called egg cell or ovum
- one *male* sexual cell (germinal cell, or gamete), called spermatozoon

Sexual cells are **haploid** and originate through a process called **meiosis** (from ancient Greek 'meiosis' = reduction). In meiosis the **number of chromosomes reduces from 2n to n:** from 1 diploid cell, through 1 chromosomes' duplication and 2 subsequent nucleus' divisions, 4 haploid cells are originated.



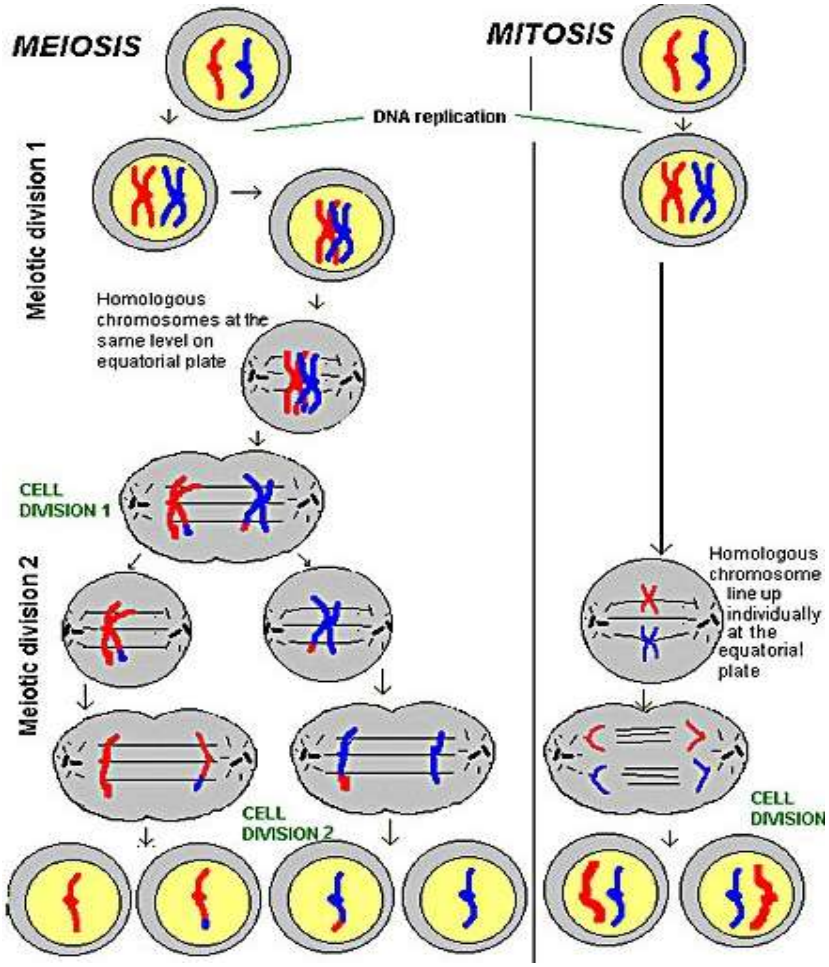
In ***asexual reproduction***, 2 cells originate from the division of an initial single cell, through a process called **mitosis**. In mitosis, the number of chromosomes doubles (from 2n to 4n) before the division, which produces 2 cells with 2n chromosomes each.

From fertilized ovum, the **zygote**, a new living being develops. The zygote, diploid (2n), contains the cellular programs of the father (n) and the mother (n). From zygote all the cells of the new living being are originated:

- ***sexual* cells**, haploid, through meiosis
- ***somatic* cells** (from ancient Greek 'soma' = body), structural and diploid, through **mitosis**

Somatic cells can *differentiate* and *specialize* in determinate functions (e.g. liver cells, blood cells, …)

Comparison between meiosis and mitosis:



DNA replication: divisions, daughter cells, final chromosomes in each cell

Some interesting videos:

Mitosis:

- *http://www.youtube.com/watch?v=VlN7K1-9QB0*
- *http://www.youtube.com/watch?v=3kpR5RSJ7SA*
- *http://www.5min.com/Video/Learn-about-the-Mitosis-Cell-Cycle-117557481*
- Real: *http://www.youtube.com/watch?v=xvOll8rRQSg*
- *http://www.youtube.com/watch?v=m73i1Zk8EA0*
- *http://www.youtube.com/watch?v=DD3IQknCEdc*
- *http://www.youtube.com/watch?v=NVfqzSKa_Bg*

Meiosis: *http://www.youtube.com/watch?v=D1_-mQS_FZ0*

Mitosis vs. meiosis: *http://biologyinmotion.com/cell_division/index.html* (Interactive)

II.B Mendelian genetics

II.B.1 Origins

Genetics as a science was born in the early 1900, when Hugo De Vries (Nederland), Karl Correns (Germany) and Erick Tschermak (Austria) confirmed Gregor Mendel's ***inheritance transmission laws*** (1865):

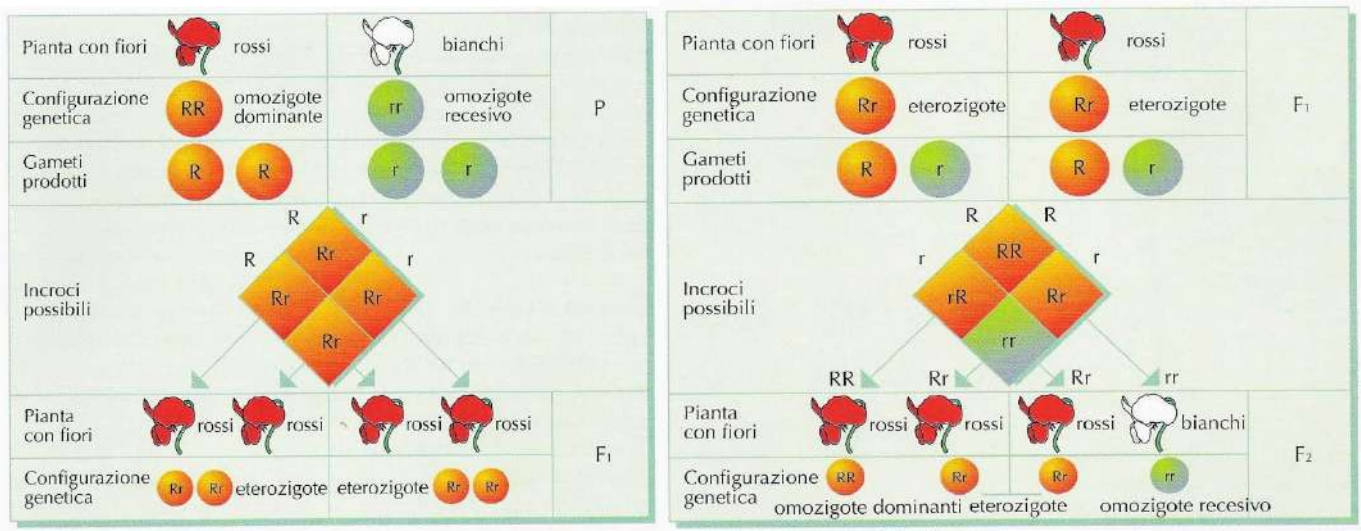
- He started from pure lines of sweet pea plants
- He took in account alternative characteristics (e.g. smooth or rough shape seeds, green or yellow color seeds, …)
- He crossed 1 plant with a characteristic (or trait) with another plant with the alternative trait
- He grouped the offsprings into classes that differed for the characteristic examined (qualitative analysis)
- He counted individuals of each class (quantitative analysis)
- He inferred that inheritance is based on "material units" identifiable by "discrete factors" (separated factors), brought individually by germinal cells. They combine in pairs during fecundation and split up again in subsequent generation, during formation of new gametes

II.B.2 Monohybrid cross

Firstly, Mendel crossed plants (parental individuals, P) that differ for only one *alternative discrete character* (**monohybrid cross**), e.g. green or yellow colour seeds

He obtained all equal hybrids from the first generation (F1) (**uniformity of the first generation**), with only one alternative trait of P, called **dominant trait** (marked with a capital letter)

He crossed the F1 individuals, obtaining F2 offsprings (second generation of P) that show the other alternative discrete character (hidden in F1), called **recessive trait** (marked with a lower letter)



Punnett's table

Mendel hypothesized that the obtained results were caused by inheritance factors

Nowadays:

- This factors are named **genes** and their alternative forms **alleles**
- An hybrid, with both dominant (R) and recessive (r) alleles, is indicated with Rr and named **heterozygote**
- An individual with 2 alleles of only 1 type is called **homozygote** (RR, dominant homozygote, or rr, recessive homozygote)
- The genetic configuration of an individual, that is its allelic composition, is called **genotype**
- The appearance of an individual, that is the way the genotype manifests itself, is called **phenotype**

First Mendel's law, or *law of segregation*:
If:

- each individual produces *equal quantity* of gametes with one or the other allele
- during fecundation, gametes combine *randomly*

Then, in F2 quantitative ratio between:

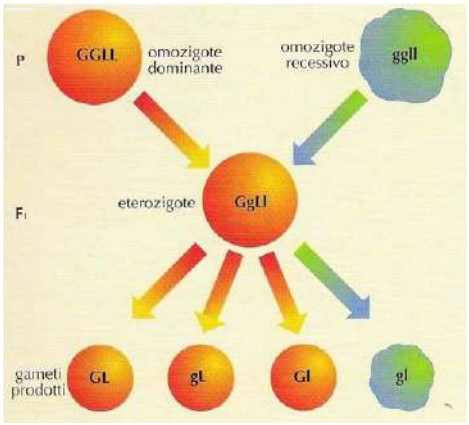
- **dominant** and **recessive** traits (phenotypes) is always **3:1**
- dominant and recessive allelic compositions is always **1RR : 2Rr : 1rr** (see Punnett's table)

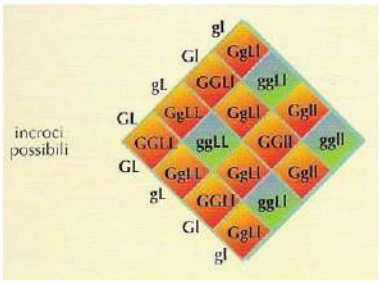
II.B.3 Dihybrid cross

Secondly, Mendel crossed individuals that differ for *2 discrete alternative characters* (**dihybrid cross**), to evaluate if pairs of traits are inherited **separately**

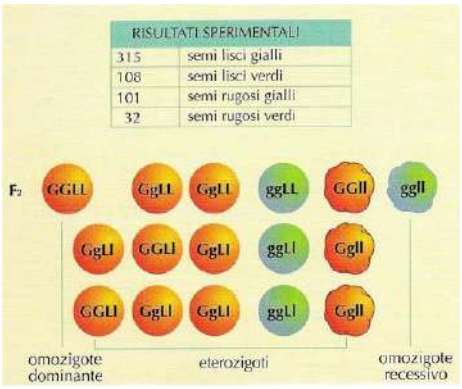
By crossing **dominant homozygote** individuals (e.g. yellow smooth seeds – YYSS) with **recessive homozygote** ones (e.g. green rough seed – yyss), he obtained:

- F1 always showed heterozygote individuals YySs (yellow smooth seeds)
- F2 showed 4 classes of individuals (see Punnett's table):
  - × YYSS, YYss, yySS, yyss
  - × Constant ratio 9 : 3 : 3 : 1





Punnett's table



Results showed that pairs of characters segregate (distribute in gametes) separately one from another

Second Mendel's law, or *law of trait independent segregation*: segregation of traits in dihybrid cross comes from simple mathematical combination of two independent segregations: (3 : 1) * (3 : 1) = **9 : 3 : 3 : 1**

II.B.4 Genes and chromosomes

Where are located the **genes**? In 1902 Walter Sutton (USA) and Theodor Boveri (Germany) state that genes are *located on* ***chromosomes***

This statement conflicts with the second Mendel law:

- the number of traits in an organism is a lot higher than the number of chromosomes, so each chromosome must contain more than one gene
- genes (or rather their alleles) are not always inherited independently (second law), but there is association (linkage) between genes or groups of genes

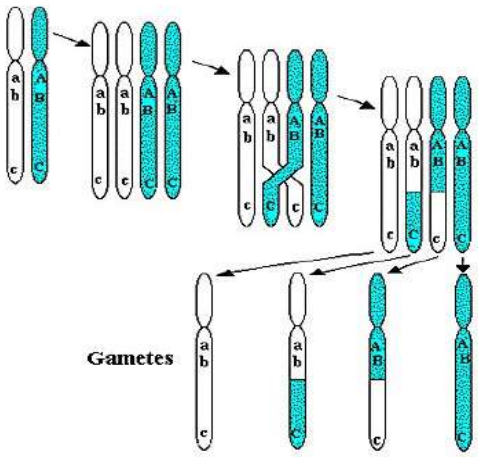
II.B.5 Genes association

In 1910 Thomas H. Morgan (USA), researching on *Drosophila* (fruit flies), demonstrated the association among different genes (alleles) of a chromosome exists, but it is **not total**

During meiosis, *homolog chromosomes* (one with male origin, one with female one) can **exchange** some genetic material (**crossing-over**), generating *recombined chromosomes* and making genes (alleles) on recombined parts *not* to be associated anymore

The point of the chromosome where the exchange and the recombination take place is called ***chiasm***

Video: *http://www.youtube.com/watch?v=op7Z1Px8oO4*



Crossing-over and recombination during meiosis

The more one gene is far from another, the more it is easy that the two genes are separated by crossing-over

For a gene pair, the percentage of recombined chromosomes in offsprings is used as a measure of the relative distance between the two genes (1 centimorgan = 1% of recombination)

The percentage of recombination between two genes is proportional to their relative distance; chiasm prevents other crossing-over processes nearby (interference)

II.B.6 Sexual characters

Inheritance of genetic characters has particular importance for genes located on sexual chromosomes

In eukaryotes, sex is genetically determined by a pair of nonhomolog chromosomes (X and Y)

Generally, one sex has two identical chromosomes (XX, homogametic sex) and the other two different chromosomes (XY, digametic sex)

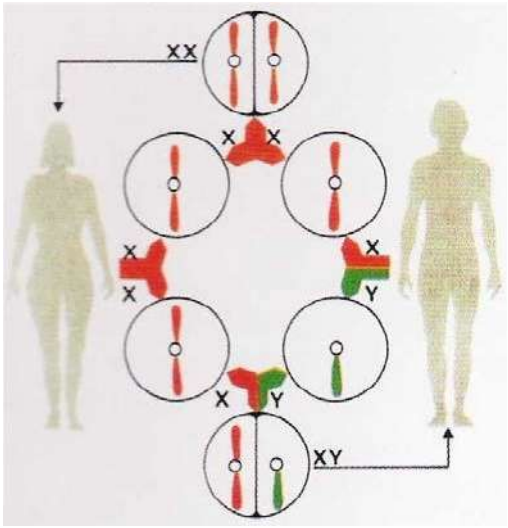
In humans and in many other organisms:

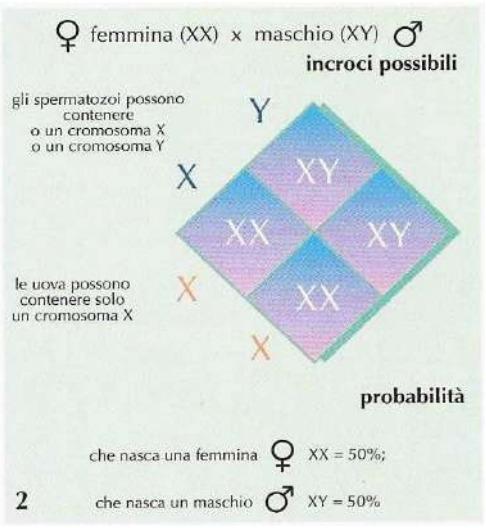
- XX female sex
- XY male sex
- Traits (genes) of Y chromosome are only in male individuals

In humankind:

- Ovum (from female individual) has 1 X chromosome
- Half of spermatozoa (from male individual) has 1 X chromosome, the other half 1 Y chromosome
- If the ovum is fecundated by a spermatozoon with a X chromosome, the new individual is a female, otherwise
- (if it has a Y chromosome) he is a male
- Sex development is a complex process controlled by many genes; genotype (XX or XY) determines sex only from the physical point of view

In humans, genes located on chromosomes X and Y are associated with sex





There are also traits (not on sexual chromosomes):

- *influenced* by sex (e.g. baldness, more common in men)
- or *limited* by sex (e.g. genes for milk production, present but never expressed in men)

Notice that somatic female cells (XX) do not make double quantity (compared with male cells XY) of genes' products of X chromosome: 1 X chromosome is genetically inactivated in female individuals

On X chromosome, around 200 genes have been found that determine various traits (e.g. sight, blood coagulation, nervous system, smell, …)

Transmission of characters linked to X or Y chromosome is different depending on the cross type; this fact is relevant in some pathologies linked to genetic alterations on X or Y chromosome:

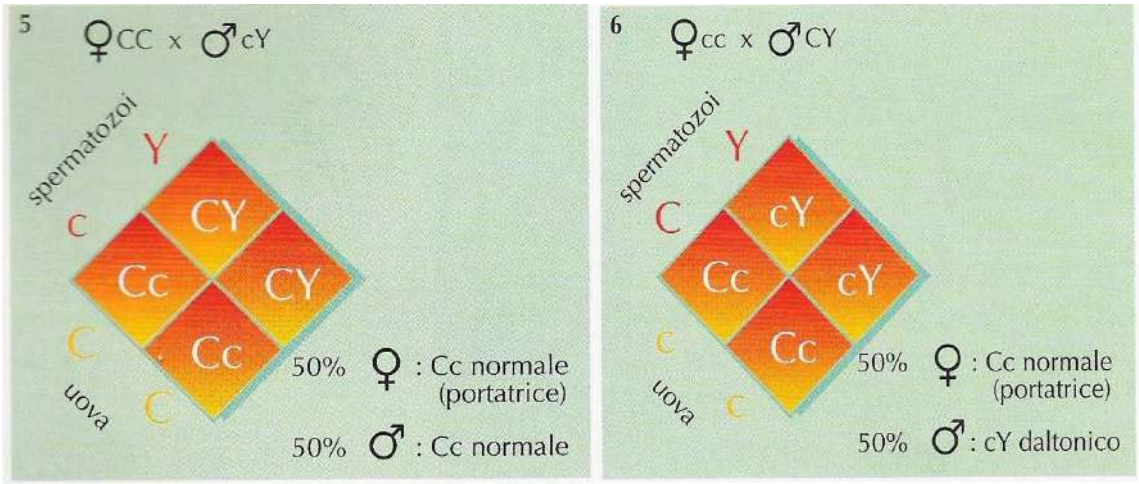
- Color blindness (due to the absence of one or more pigments in retina's cells; incidence 8% men, 0.0064% women)
- Hemophilia (due to lack of proteins necessary to blood coagulation)
- Turner's syndrome (due to lack of one X chromosome in female sex; it brings short height and sterility; incidence 1:5000 women)
- Klinefelter's syndrome (due to one extra X chromosome in XXY individuals; it brings male aspect, small testicles and developed breast, tall height and mental deficiency; incidence 1:500-2000 men)
- Aneuploidy: anomaly in the number of chromosomes (due to errors during cellular division producing gametes)

Transmission of sexual characters (e.g. color blindness)

Normal allele (C) that gives correct colors' perception is dominant on the mutated one (c), they are both on X chromosome

A cross between color-blind father (cY) and normal mother (CC) gives: F1: normal son (unique normal X chromosome: CY), or normal daughter only in phenotype (symptom-free carrier, heterozygote Cc)

A cross between color-blind mother (cc) and normal father (CY) gives: F1: normal daughter only in phenotype (symptom-free carrier, heterozygote cC), or color-blind son (unique X altered chromosome: cY)



In a population, if there are $k\%$ male individuals with altered phenotype of X chromosome, there are basically $k\% * k\% = (k\%)\hat{}2$ female individuals with altered phenotype (e.g. $K = 2\%$ in males, $2\% * 2\% = 0.04\%$ in females)

II.B.7 Genes' interaction

Genes do not act independently; they can interact, generating unpredicted phenotypes or segregations that differ from the expected ones according to the Mendel laws.

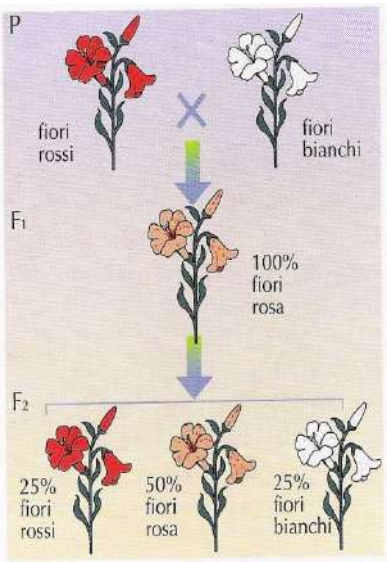
Examples:

- Incomplete dominance
- Co-dominance
- Polygenic inheritance

Not always all F1 individuals have the same dominant phenotype

In some cases, heterozygote individuals have a trait that is intermediate between the parental ones (e.g. by crossing "four o'clock" red flowers' plants (RR) with white flowers' ones (rr), 100% of F1 plants with pink flowers are obtained)

By crossing F1 individuals, in F2: 25% RR, 25% rr and 50% Rr with intermediate phenotype (e.g. pink flowers) (**incomplete dominance**)



In some cases, the heterozygote shows phenotypes of both parents: no phenotype is dominant on the others (**co-dominance**), e.g. AB blood group in humans:

- genotype I^AI^A or I^Ai^A or I^Ai^B generates A group
- genotype I^BI^B or I^Bi^B or I^Bi^A generates B group
- genotype i^Ai^A or i^Bi^B or i^Ai^B or i^Bi^A generates 0 group
- genotype I^AI^B or I^BI^A generate AB group

GRUPPO SANGUIGNO (donatore)	ANTIGENI PRESENTI SUI GLOBULI ROSSI	PUÒ PRODURRE ANTICORPI	GRUPPO SANGUIGNO (ricevente)			
			A	B	A B	O
A	A	anti–B	○	—	○	—
B	B	anti–A	—	○	○	—
A B	A, B		—	—	○	—
O	nessuno	anti–A anti–B	○	○	○	○

○: nessuna immunizzazione —: reazione di immunizzazione

0 group: universal donor / AB group: universal receiver

There are many traits that vary in a continuous way (e.g. height, skin color, …) and not in a discrete alternative way (as Mendel's traits)

These traits are called quantitative (because depend on quantity), or **polygenic** because determined by the expression of more than one gene

Each genotype determines a different phenotype, with slightly different phenotypes and continuous quantitative traits (**polygenic inheritance**)

II.B.8 Gene-environment interaction

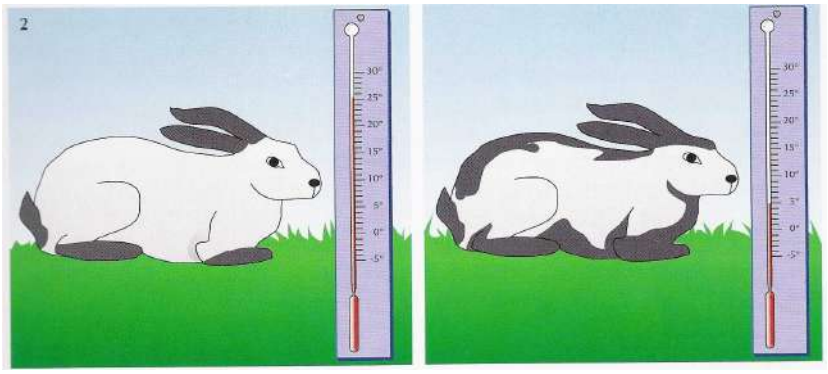
A phenotype is determined not only by genes' expression, but also by the **environment** where the organism lives in. Environment can act on the same genotype, producing different phenotypes (e.g. skin pigmentation depends on genotype, but also on exposure to UV rays)

Genes determines **potentialities** of some traits' realization; these potentialities are **influenced** by the interaction between genes of the same organism and interaction between the organism and its environment

$$\textbf{\textit{phenotype}} = \textbf{\textit{genotype}} + \textbf{\textit{environment}}$$

Individuals with the same genotype can show different phenotype:

- in bees' populations, a female, with the same chromosome complement of others, can become *queen* bee only if fed with royal jelly, otherwise it becomes worker bee
- *homozygote twins* that live in different places
- *Himalayan rabbits*, that have white fur (unless black extremities) if living at 25° and more black fur if living at less than 10° (due to an enzyme that contributes to black pigment, which is active only at low temperatures, e.g. at extremities in 25° environment)



II.B.9 Genes of populations

Mendel's laws point out the principles of inheritance, having as object of study the single individual; are they valid for entire populations?

It is necessary to move the focus from the single individual to Mendelian population, from genotype of single components of population to genetic pool of examined population. ***Populations' genetics*** evaluates, through statistical analysis, frequency of presence of different genes in a population (**genetic variability** of a population)

Contrary to what Mendel laws claim, recessive traits do not disappear in a population with the passing of time. G. Hardy e W. Weimberg in 1908 determined that, in a balanced population, **frequencies of genes and genotypes tend to remain constant in generations.**

Let's see the demonstration:

Given the distribution of paired alleles A and a in a population, we want to know their relative frequencies in it. Each member of the population can have genotypes AA, Aa, or aa.

Examining phenotypes of the population, we can have the percentage of individuals with A (or a) phenotype. We can **not** have the frequency of allele A, because it is present in dominant homozygotes (AA) and in heterozygotes Aa and aA (the same for allele a)

Since in the total population the frequency of the sum of alleles is 100%, if frequency of A is p and of a is q: $p + q = 100\% = 1; \quad p = 1 - q; \quad q = 1 - p$

If alleles A and a are *equally shared* between male and female populations (frequency of A is p and of a is q), in subsequent generation for the first Mendel's law population is characterized by **AA + 2Aa + aa** ratio, but due to **conditional** probability:

- individuals AA have frequency $p * p = p\hat{}2$
- individuals aa have frequency $q * q = q\hat{}2$
- heterozygote individuals, coming from independent events Aa and aA, have frequency $(p * q)$ and $(q * p)$, or rather $Aa + aA = pq + qp = 2pq = 2Aa$

So, the sum of alleles is 100%, i.e. $AA + 2Aa + aa = (A + a)\hat{}2 = p\hat{}2 + 2pq + q\hat{}2 = (\boldsymbol{p + q})\hat{}\mathbf{2} = \mathbf{1}$ (**Hardy-Weimberg's law**)

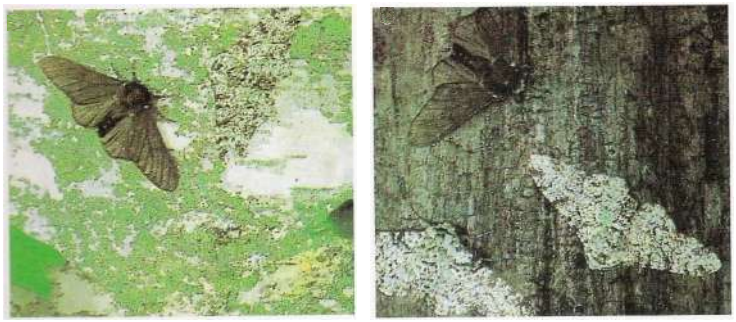
If population is sufficiently *big* and the mating is *random*, allelic frequencies remains constant among generations, in F1:

- Since frequency of AA is p^2, of Aa is pq, of aA is qp and of aa is q^2,
  - frequency of A is $f\_A = p\hat{}2 + pq/2 + qp/2 = p\hat{}2 + pq$
  - frequency of a is $f\_a = q\hat{}2 + pq/2 + qp/2 = q\hat{}2 + pq$
- Since $p = 1 - q$ and $q = 1 - p$,
  - frequency of A is $f\_A = p\hat{}2 + pq = p\hat{}2 + p(1-p) = p$
  - frequency of a is $f\_a = q\hat{}2 + pq = q\hat{}2 + q(1-q) = q$
- that means frequencies of A and a remains the same among generations; thus, this population is said **balanced** (with respect to the two alleles)

Hardy-Weimberg's law is **theoretical**. It is true only if there are no factors that tend to make allelic (genetic) frequencies change. Generally, this is not verified: evolution causes a continuous process of changing in the genetic constitution of a population.

Evolutionary factors that make genetic frequencies vary are:
- Mutation
- Natural (or artificial) selection
- Migrations
- Chance…

Natural selection (i.e. differential reproduction of individuals more suitable for a given environment) is the main evolutionary factor. The more an individual has a suitable genotype, the more it has the possibility to survive, to reproduce and to have an offspring. The measure of reproductive ability of an individual (that is statistic suitability of an individual for an environment, depending on his/her genotype) is called **fitness**



Specimen of peppered moth (Biston betularia) with two phenotypes, one light and one dark, upon birch trunks, healthy or polluted

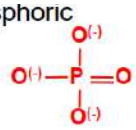
Meiosis, mitosis, and evolution: synthetic visual summary *http://www.youtube.com/watch?v=uH4UUv7Cr4A*

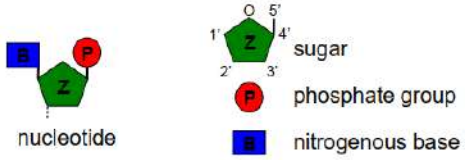
II.C Molecular genetics (21 Sept.)

II.C.1 DNA and its structure

DNA is the biggest macromolecule in the cell: it is a polymer of 4 different types of monomers, nucleotides, made of:

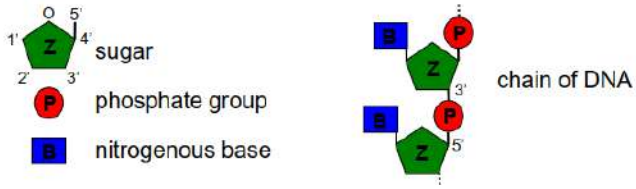
- 1 molecule of sugar (**deoxyribose**, from which the name DeoxyriboNucleic Acid – DNA) with 5 atoms of carbon that bond with:
  - 1 phosphate group (from a molecule of phosphoric acid – P)
  - 1 molecule containing nitrogen (N) (1 of the 4 nitrogenous bases A, C, G, T)

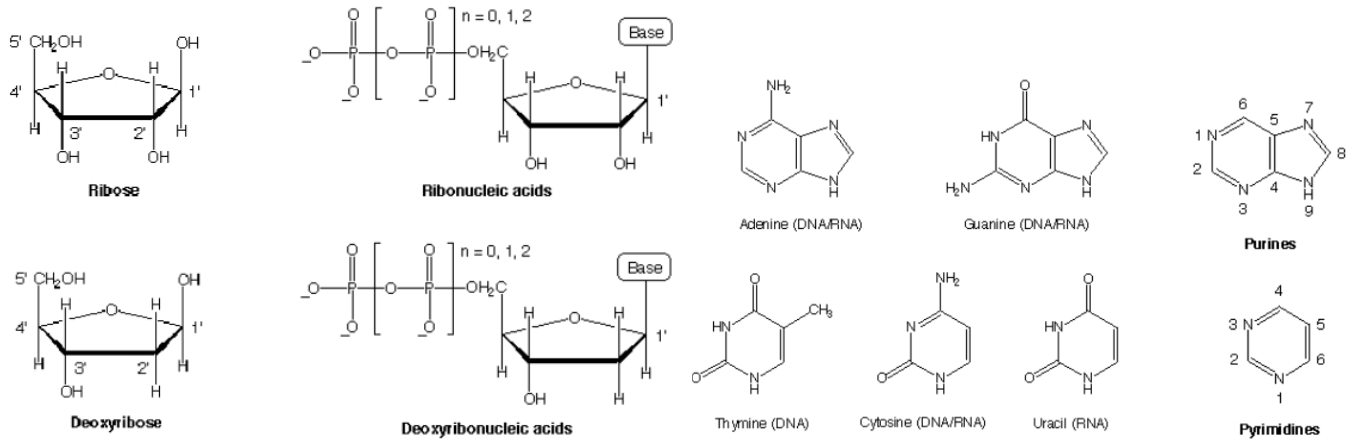


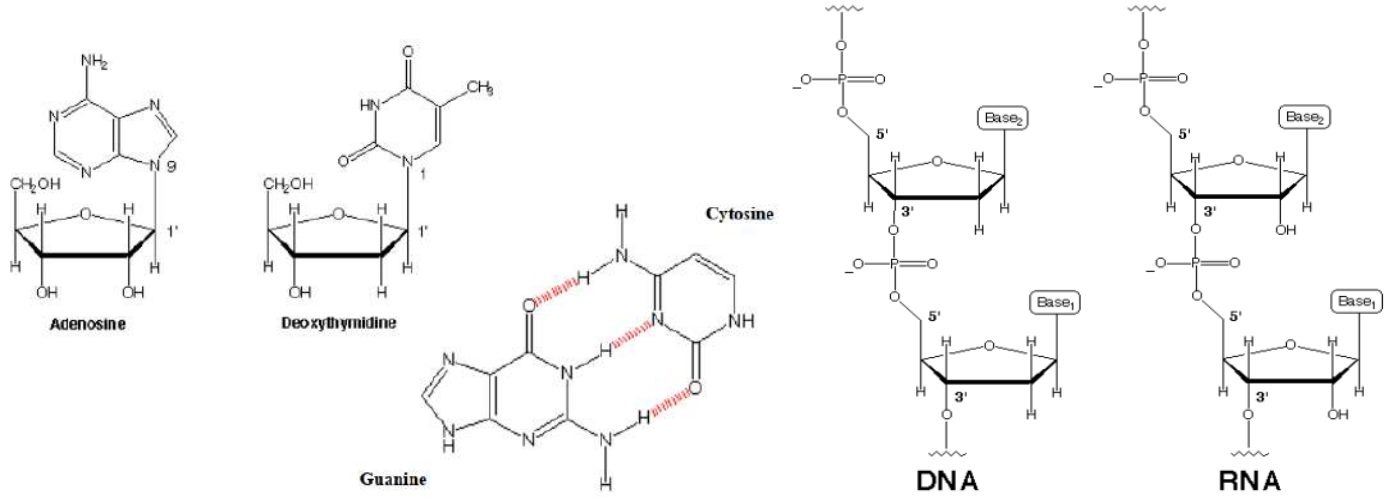


DNA chain: long sequence of nucleotides linked by the bond between the phosphoric acid of a nucleotide and the sugar of the subsequent nucleotide (sugar-phosphoric acid-sugar "bridge")

This *bond* is called **3'-5' phosphodiester bond**, where 3' and 5' is the ordinal number of the atom of carbon of the sugar joining the bond







The chemical structure is mentioned but will not be the main consideration of the course, so it is not compulsory to know the exactly by heart

In 1953, in Cambridge UK, James Watson (USA) and Francis Crick (UK) defined the **exact spatial structure** of DNA, considering two types of experimental results:

1. DNA bases' *composition*

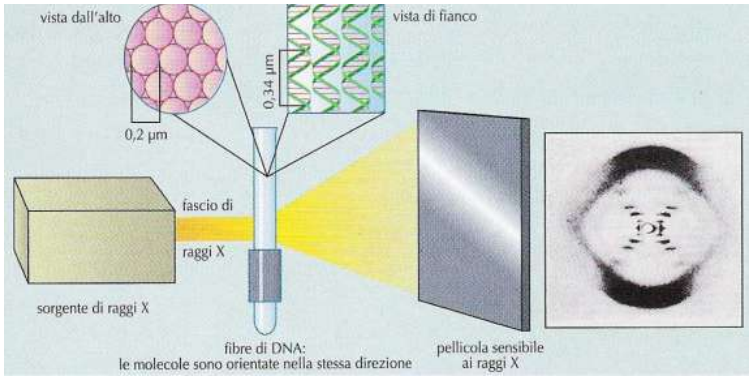
In 1945-1950 E. Chargaff (US) found that in the DNA of each organism (Chargaff's rules):

- Quantity of Adenine (A) = quantity of Thymine (T)
- Quantity of Cytosine (C) = quantity of Guanine (G)

or: $\frac{A}{T} = \frac{C}{G} = 1$ even if $\frac{A+T}{C+G}$ [CG percentage] varies in different organisms

2. *Diffraction* spectra from X rays of pure DNA fibers' crystals

In the same years, Rosalind Franklin (UK), with M. Wilkins, obtained the first photographs of diffraction spectra from X rays of pure DNA fibers' crystals, that showed a **helix structur**e with 2 **characteristic periodicities**, at 0.034 µm and at 0.0034 µm on the main molecular axis



Watson and Crick combined information from Chargaff and Franklin, building tridimensional models of DNA

Model that mostly fitted experimental data was ***double helix***

If temperature, acidity (pH) and humidity are the ones characteristic of living cells, DNA spontaneously arranges itself in a structure with the following characteristics:

1. *Double helix* of two polynucleotide chains that roll up together in a *right-handed coil*, relative to the main molecular axis

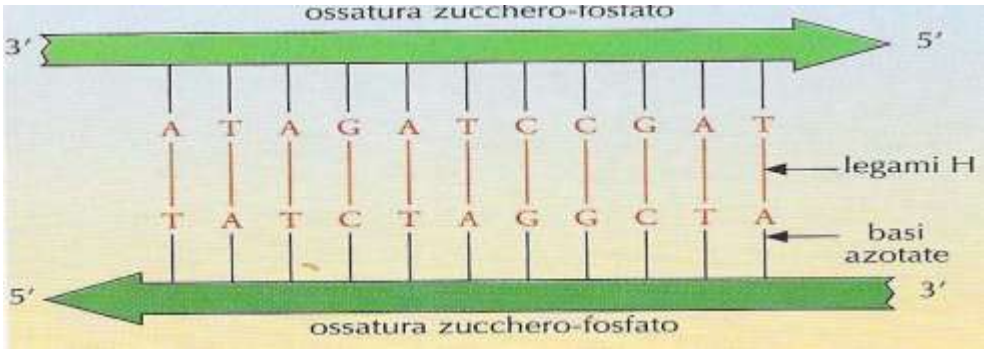
2. Nucleotide bases are arranged in the *internal* part of the helix, perpendicularly to the main molecular axis; backbone sugar-phosphate is in the *external* part

3. Nucleotide bases interact by weak hydrogen bonds (H):

- A and T bond with 2 H bonds
- C and G bond with 3 H bonds

  (A-T and C-G are called pairs of complementary bases)

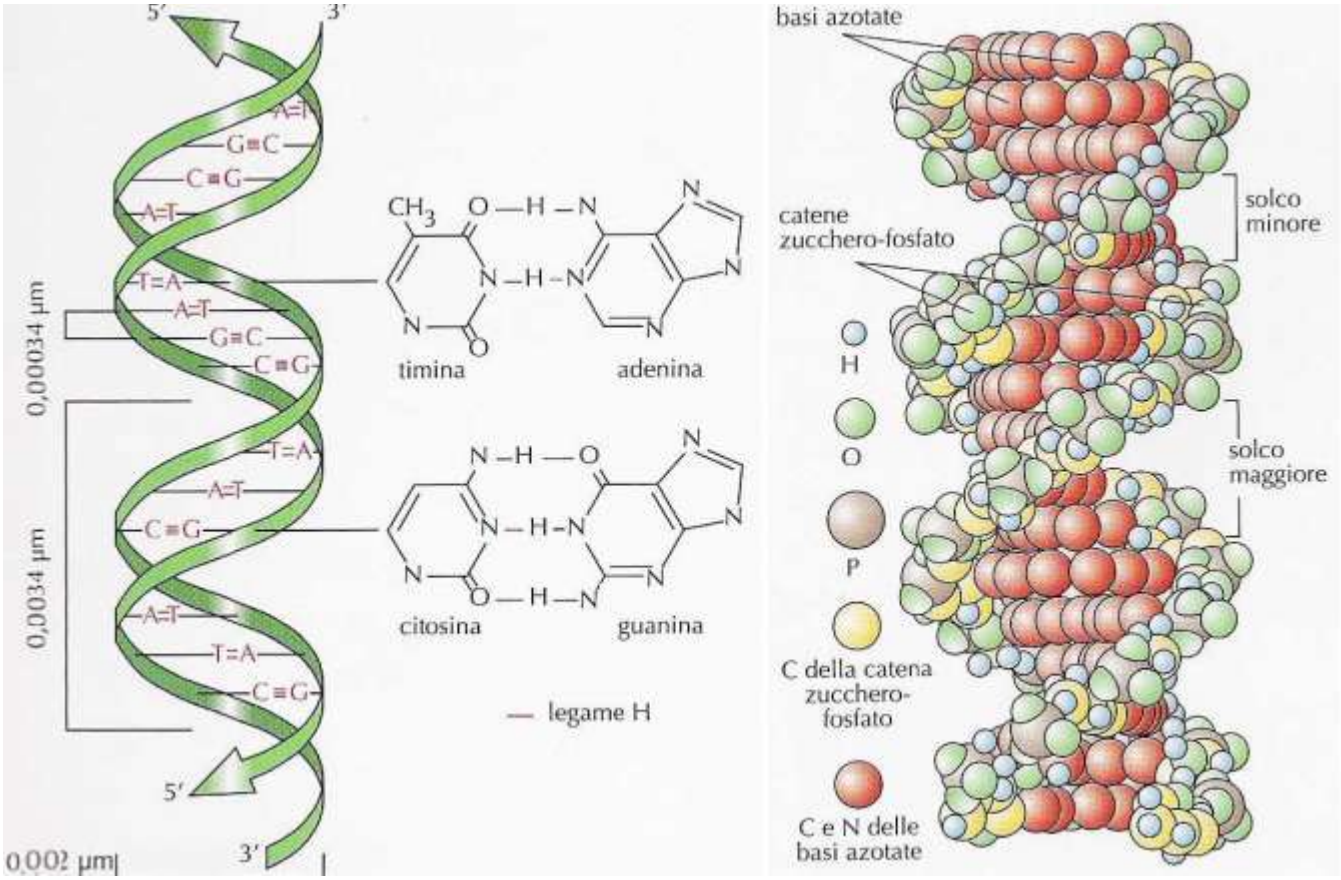
4. H bonds also strengthen structure of the helix



5. There are 10 pairs of bases (10 bp, base pair) for each helix turn; pairs are spaced of 0.00034 µm, so pitch of the helix is 0.0034 µm (as shown by X rays spectra); double helix has a diameter of 0.002 µm

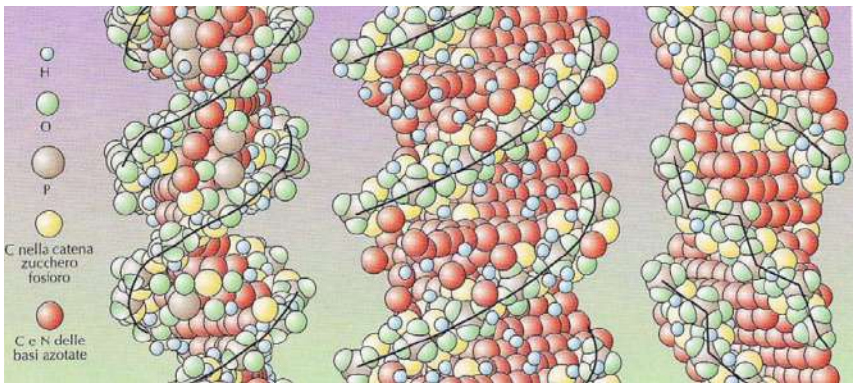
6. The two nucleotide chains are anti-parallel, that means they have opposite directions (one from 3' to 5', the other from 5' to 3')

7. Double helix has 2 grooves on the surface, one bigger than the other; in these grooves, protein interactions occur for DNA replication and genetic transcription



Double helix of Watson and Crick is called B-shape of DNA and it is characteristic of living cells (with much water). When humidity is low, DNA arranges itself in a more compact and wide structure (A-shape)

There are also other types of DNA structure, e.g. Z-shape, with zigzag course of backbone sugar-phosphate, thinner structure and left-handed coiling



DNA molecules are very long: in each human cell there is more than 1 m of DNA. They are thin and flexible, so they run up together and occupy a very small volume

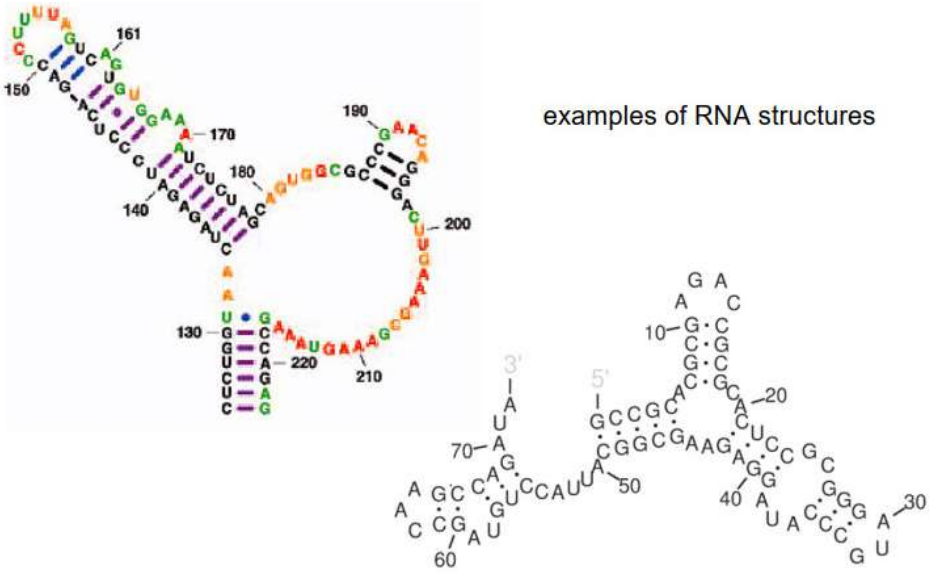
Phenomenon of bending is called packaging of DNA (video: *http://www.youtube.com/watch?v=OStI5pniHPA)*

II.C.2 RNA and its structure

Into the cells, RNA plays a leading role in the synthesis of proteins. From the structural point of view, RNA compared to DNA:

- Comprises ribose as sugar (instead of deoxyribose)
- Has nitrogenous base U (instead of T), that matches with A (as T)
- Always exists as single chain, where the bases interact in complex tridimensional structures, often not wellknown yet

In some living beings without DNA, (e.g. RNA virus), RNA plays a leading role in reproduction process of the individual (e.g. tobacco's mosaic virus)



In cells, there are different types of RNA, with different roles (many of them are not completely understood):

- messenger RNA (mRNA)
- ribosomal RNA (rRNA)
- transfer RNA (tRNA), about 75-95 nucleotides
- small nuclear RNA (snRNA)
- small interfering RNAs (siRNA), 20-25 nucleotides
- micro RNA (miRNA), 21-23 nucleotides
- long non-coding RNAs (long ncRNAs), >200 nucleotides
- antisense RNAs
- …

mRNAs, rRNAs and tRNAs play leading different roles in ***protein synthesis***, others types in its ***regulation***

II.C.3 Genome

Genome: genetic material of an organism

Generally, it indicates DNA contained in each cell (characteristic of species and organized in n chromosomes). Often this term refers also to RNA and proteins (that come from DNA)

Dimension and *organization* of genome vary depending on species. Bacterial cells (**prokaryotes**) have **haploid** genome (n). The majority of **eukaryote** cells have **diploid** genome (2n), with different number of chromosomes depending on species

II.C.4 Genome of viruses

Viruses are the simplest life forms

They are *not independent* cells, but cellular parasites: they can reproduce themselves only inside another cell (by using its enzymatic systems)

Virus genome:

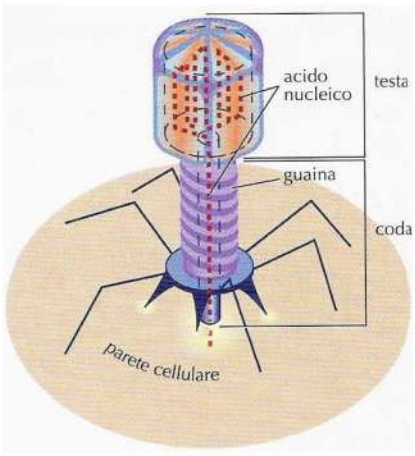
- Composed by *1 molecule* of nucleic acid (DNA or RNA)
- Enclosed in a protein shell (*capsid*) with *different shapes* (icosahedral, helical, or filamentous, head-tail)

There are many viruses, divided into 3 classes:

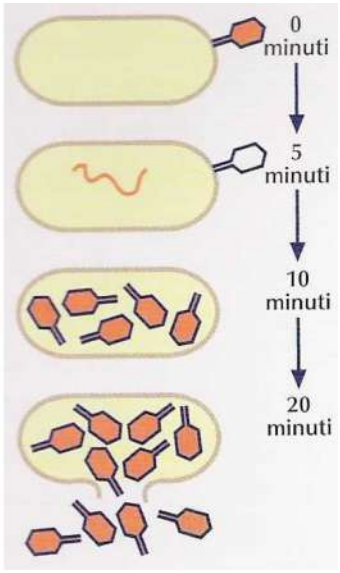
- Viruses of bacteria, or bacteriophages or phages
- Viruses of plants
- Viruses of animals

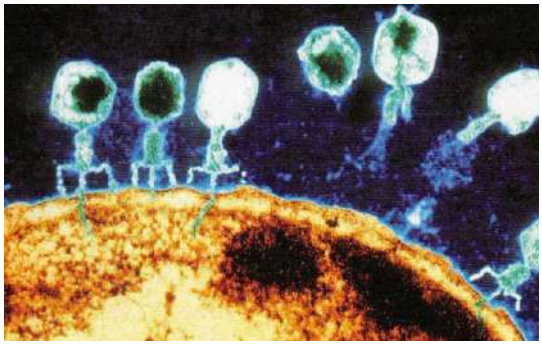
Bacteriophages: capsid with icosahedral head, containing genetic material (DNA or RNA), connected to a hollow cylinder (*tail*) to which filamentous structures (*spikes*) are linked, which allow the hanging of the virus on the bacterial cell's wall

When hanged, virus injects its genetic material, from the head through the tail, inside the cell, where it reproduces itself (video: *http://www.youtube.com/watch?v=gU8XeqI7yts*)



Structure of a bacteriophage and phases of its reproduction

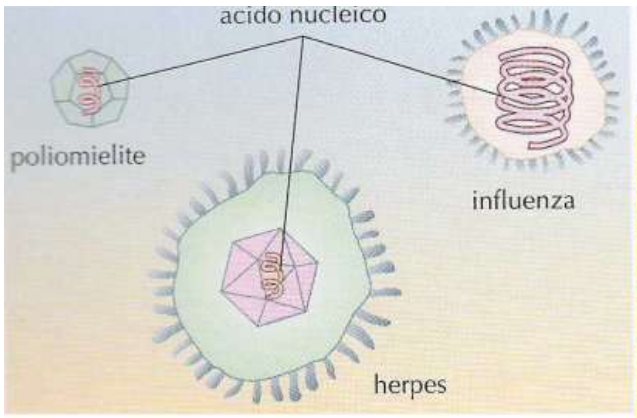


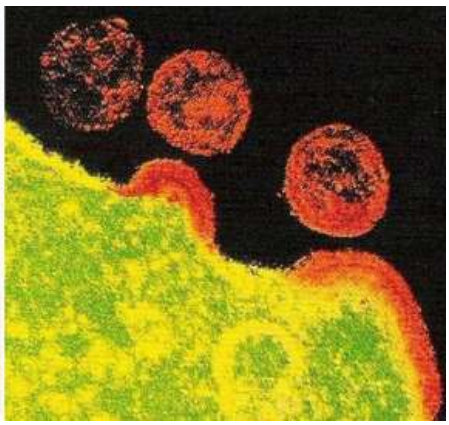


Phages attacking an Escherichia coli bacteria seen by transmission electron microscopy (TEM) in false color

Viruses of eukaryote cells:

- More numerous, both for animals and plants (for plants less important than macro-parasites)
- Capsid mainly icosahedral or filamentous (sometimes coated with lipid membrane coming from infected cell after the leakage of viral particles)
- Genetic material (DNA or RNA) is variable in structure (single or double helix, linear or circular, segmented, or complete)
- The majority of plants' virus are RNA virus

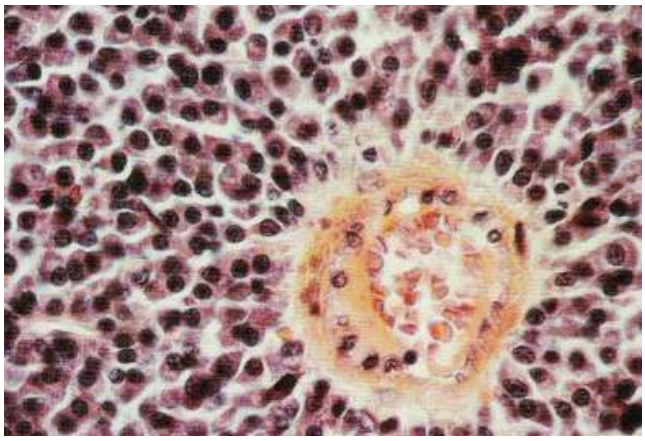




Exocytosis of a new virus from an infected cell. HIV virus (Human Immunodeficiency Virus) (red) exits from an infected lymphocyte: TEM photograph in false colour

Many viruses can include their genome in the host cell, determining dramatic changes (morphological and physiological) in the life of the infected cell (can cause cancer)

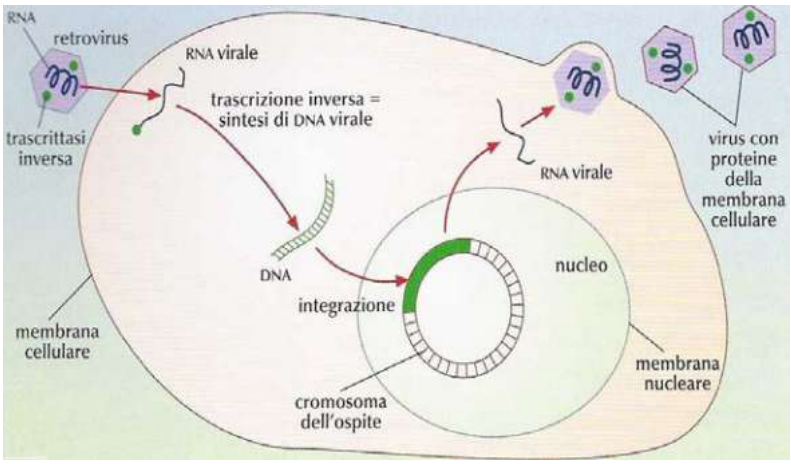
Cellular transformation: phenomena coming from integration of virus in host cell's DNA



Optical microphotograph of bone marrow tumor (myeloma): tumoral cells (purple) replaced the majority of healthy tissue, cells remaining (pink) are dying

Retrovirus (e.g. HIV [Human Immunodeficiency Virus])

- Contains 1 or 2 molecules of RNA
- Thanks to the enzyme reverse transcriptase, firstly RNA is changed in single chain DNA (intermediate form), secondly transformed in double helix by another enzyme active in cell nucleus (DNA polymerase); in this form, genome of virus can integrate in the host cell's genome and reproduce itself



II.C.5 Bacterial genome

Bacterial cells (prokaryotes) *do not have a defined nucleus*, but a compact structure (**nucleoid**) made of 1 molecule of DNA (usually circular) of about 1 mm (bacterial cells is about 1 µm!!)

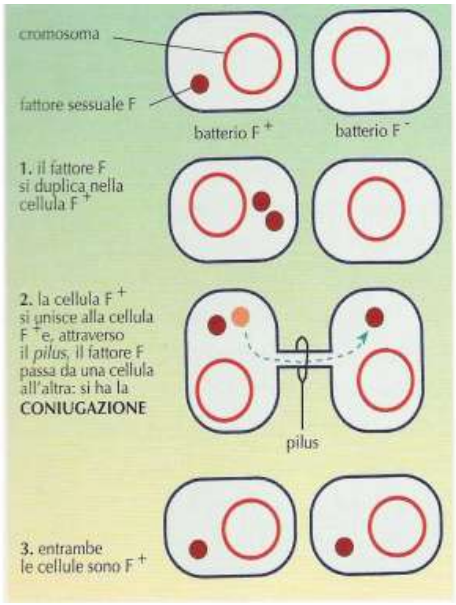
- Exact structure of bacterial nucleoid is not clear, but it seems to have packaging with numerous folds (super-coiling)
- Along bacterial DNA there are 2'000-2'500 genes in **continuous sequence** (without interruptions)

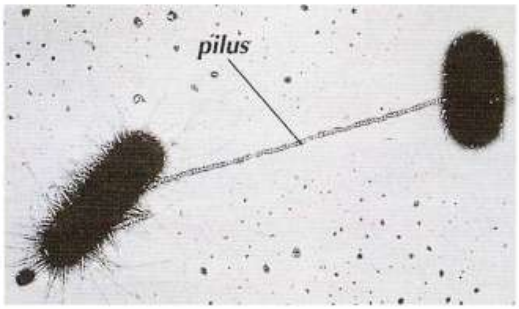
In many bacteria, in addition to proper genome, there are also small circular molecules of DNA: ***plasmids***, which:

- Give to the cell *advantageous* characteristics
- Can *integrate* in cellular genome and detach themselves bringing a variable genetic mix with them

There are various types of plasmids, among which:

- ***R plasmid*** (R = resistance): determines cell's resistance to cations of heavy metals, or to many antibiotics -> difficulties in treatment of bacterial infections. If there is an R plasmid, in environments rich in antibiotic, resistance-genes are activated (not activated in environment poor in antibiotic to save energy)
- ***Degradation plasmids***: allow the bacterium to metabolize stable chemical compounds (oil residuals, pesticides, …) -> used for polluted areas' recovery
- ***Fertility factor*** (or F factor): cells containing it are called male (F+), others female (F-)
    - Through sex pilus (cylindrical structure in cellular wall), F+ cells can transfer to F- cells a copy of F plasmid (conjugation), changing F- in F+
    - Conjugation allows "horizontal transfer" of genetic material





II.C.6 Genome of eukaryotes

Nucleus in **eukaryote** cells is more complex than in prokaryotes. It contains different *linear molecules* of **DNA**, each of which is contained in a **chromosome**.

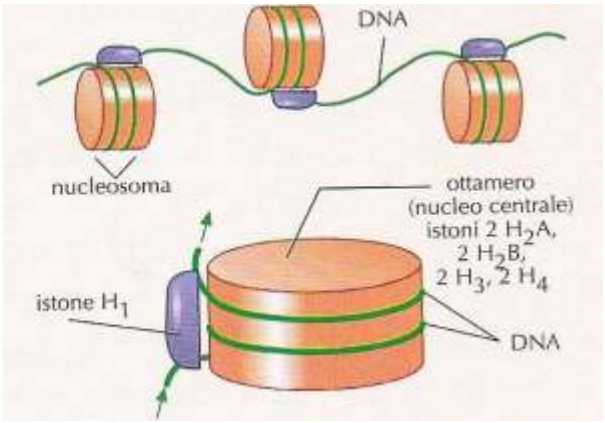
Number of chromosomes is *not proportional* to dimension of genome:

- beer yeast (*fungus saccharomyces cerevisiae*) has genome of about 20.000 Kb (kilobases) subdivided into 16 chromosomes
- human DNA has 3.000.000 Kb and is subdivided into 23 chromosomes.

Chromosomes are constituted by **chromatin** (50% of DNA, 50% of proteins and a little part of RNA). *Proteins* more strictly linked to DNA are **histones** (H1, H2A, H2B, H3, H4)

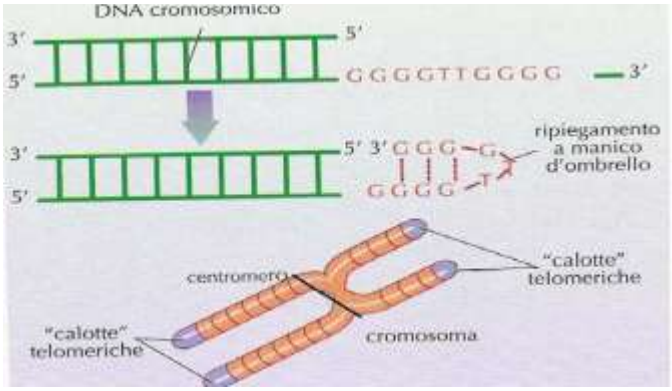
Chromatin has structure similar to a necklace: central filament with spherical particles (**nucleosomes**) of around 10 µm-diameter. Each nucleosome is about 50-70 bp (linker) from the subsequent one.

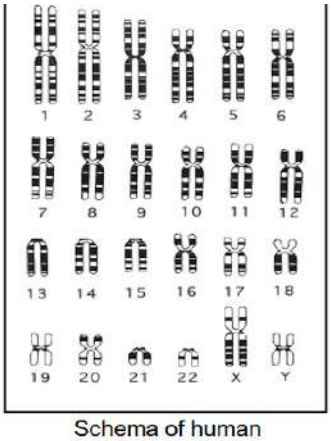
DNA rolling on nucleosome partially explains how DNA is packaged in chromosomes



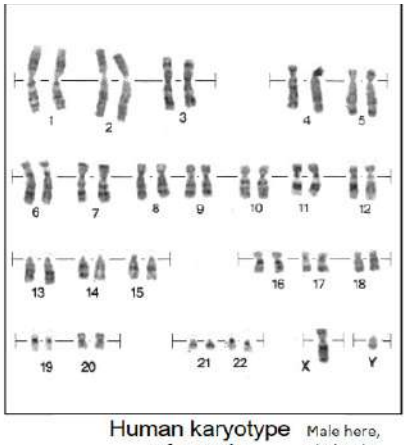
During a particular phase of cellular division (***metaphase***) chromosomes assume the *X shape*, with 2 linear elements (**chromatids**) connected in one point called ***centromere***, that divides each chromatid in 2 *p arms* (*short* arm) and 2 *q arms* (*long* arm)

Relative *position* of centromere, *length* of chromatids and *dimension* of chromosomes clinically identify different chromosomes (that constitutes ***karyotype*** of the organism). At the extremities of chromatids there are ***telomeres*** to stabilise the end part (from, Ancient Greek, "thin tip")

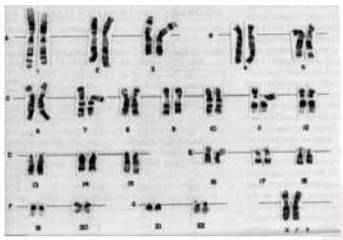


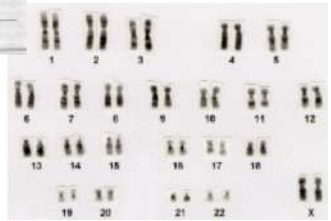


Schema of human chromosomes



Human karyotype of a male Male here, obviously


Human male karyotype


Human female karyotype

With artificial staining, areas of chromosomes rich in *A and T bases become dark-colored*, areas rich in G and C bases remains pale, generating a striped arrangement (bands)

Each band has specific **nomenclature** (e.g. 6p21.3) that indicates its position, like on a map, specifying:

- chromosome
- arm
- region
- band
- sub-band

E.g., 6p21.3 indicates:

- 6: number of chromosome (chromosome 6)

- p: shortest arm of chromosome
- 2: group of bands (region) visible on the arm starting from centromere
- 1: band inside the group, counting from centromere to telomere
- 3: sub-band, a thin band visible inside the thicker one, counted from centromere

That means, position 6p21.3, which indicates third sub-band in the first visible band on the second group localized on the short arm of chromosome 6 (*http://homepages.uel.ac.uk/V.K.Sieber/human.htm*)

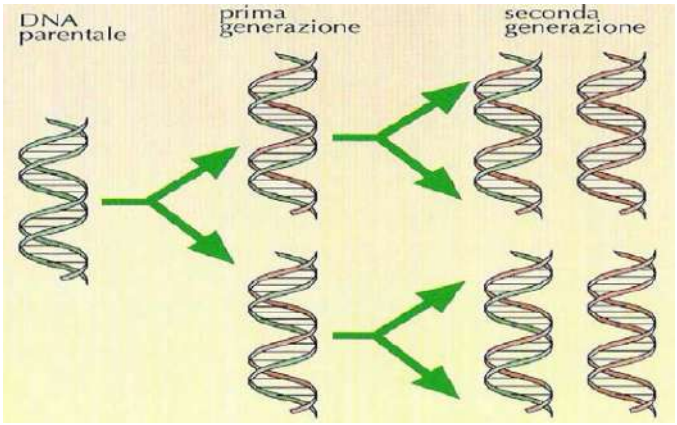
Not all genetic material of eukaryotes is in the nucleus. A small fraction of circular DNA is in *cellular organelles*: ***chloroplasts*** (in vegetal cells, delegated to photosynthesis) and ***mitochondria*** (produce chemical energy for the cell)

Extra-nuclear genes of chloroplasts and mitochondria, being in cytoplasm, are transmitted to the offspring just by the ovum (that has more cytoplasm than the spermatozoon), leading to a maternal inheritance, not mendelian

II.C.7 Duplication of genetic information

To guarantee *transmission of genetic information* towards offspring, before cellular duplication, DNA is copied (**duplication**, or **replication of DNA**)

Same process in eukaryotes and prokaryotes. DNA molecule replicates according to a *semi-conservative* model: each daughter molecule has 1 strand of DNA of the mother molecule and 1 of new synthesis

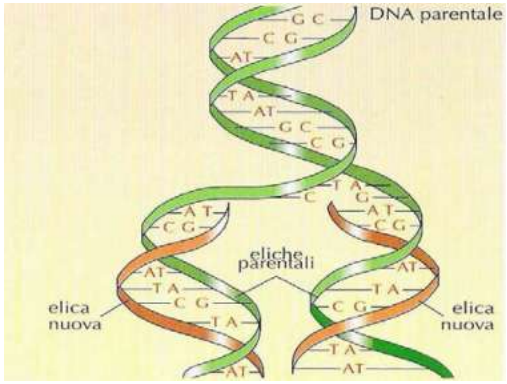


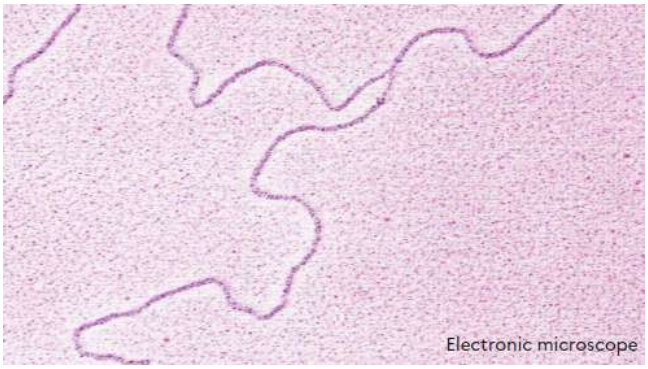
Model of ***semi-conservative duplication*** comprises:

- *"Zip" opening* of the double helix
- *Exposition of single bases* on the two strands that act as *mold*
- *Pairing of bases* between free nucleotides in the cell and the *complementary* ones in the mold (hydrogen bond)
- Nucleotides of paired bases bond to create *new strand*

Finally, **2 double helixes**, both consisting in:

- 1 parental helix
- 1 new synthesized helix





Area where double helix opens, and synthesis starts is called ***replication fork*** (for its shape)

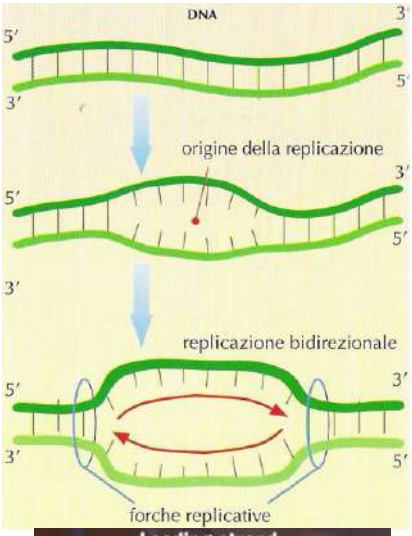
Not concurrent duplication along all the DNA molecule, but just in a specific position at a time, starting from the *origin* of duplication, following determined sequence:

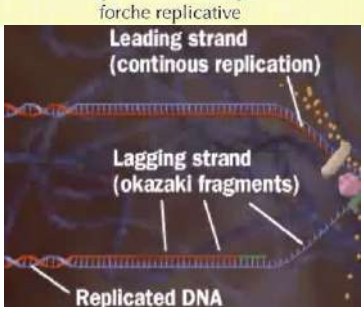
1. Localized *opening* of the double helix by specific enzymes
2. Copying through pairing of bases and polymerization of nucleotides thanks to **DNA polymerase** (III and I) enzymes, from 5' to 3' direction on both DNA strands

   On 5' to 3' strand (leading strand) continuous copying

   On 3' to 5' strand (lagging strand) copying from 5' to 3'

   must occur in restricted areas (***okasaki fragments***), after linked together by **DNA ligase** enzyme
3. Re-closing of double helix by many specific enzymes



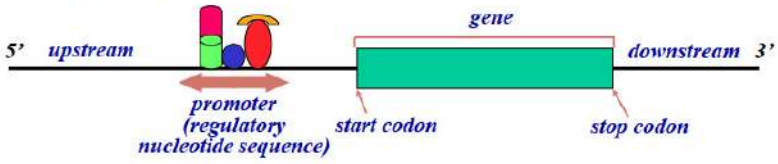


Video DNA replication: *http://www.youtube.com/watch?v=teV62zrm2P0*

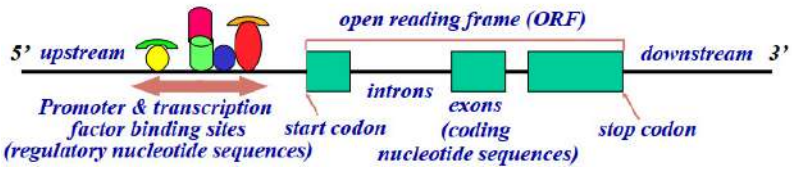
II.C.8 Gene structure

Sequence of 3 bases (triplet) = ***codon***

- In **prokaryotes**:



- In **eukaryotes**:



Human genome consists of about *3 billion bases* (3.000 Mb), but only 22.000-25.000 genes *encoding* proteins. Coding area comprises about 90 Mb (only 3% of genome)

More than 50% of genome is constituted by repeated sequences:

- Tandem repeats (10-15% DNA, from 1 to millions of bp): repetitions are adjacent (e.g. ATTCG*ATTCG*ATTCG)
- Interspersed repeats (35-40% DNA, from 100 to 10'000 bp): repetitions are scattered along DNA

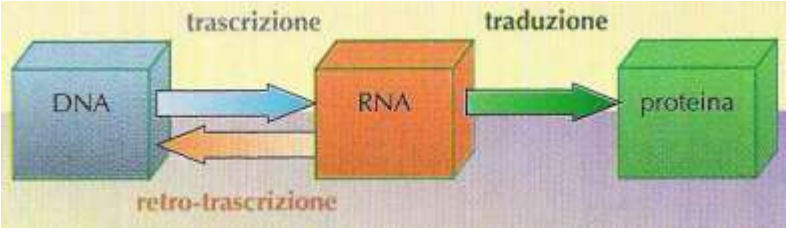
Many sequences are repeated differently in different individuals

II.C.9 Expression of genetic information

Term ***genetic expression*** [Crick 1958] indicates the biological process that transfers genetic information from DNA to RNA to protein (**central dogma of Molecular Biology**)

Information is transferred one-way:

- From DNA to RNA: ***transcription***
- From transcribed RNA to protein: ***translation***



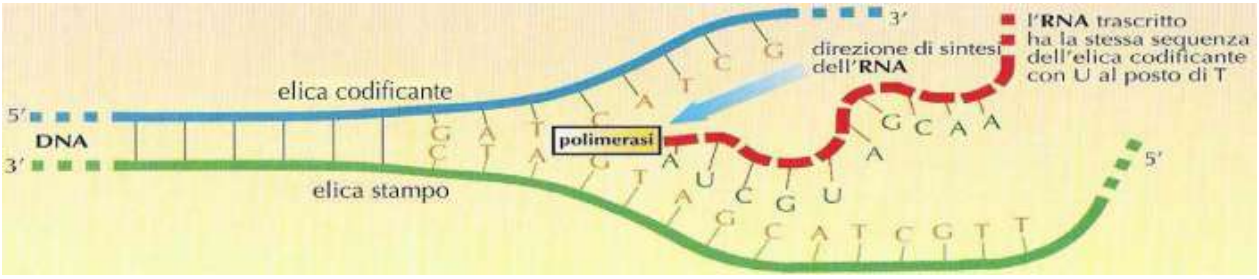
Always verified even if retrovirus can transfer information from RNA to DNA

II.C.10 Transcription

In ***transcription***, DNA information leads the synthesis of RNA molecule from 5' to 3'. Only *1* helix of DNA is used to synthesize one RNA (***mold helix***)

Synthesized RNA molecule has the same sequence of the stretch of the other DNA strand (coding helix), but with base U (uracil) instead of base T (thymine)

Mold and coding helixes are not the same for all types of RNA
Synthesis is catalyzed by enzyme RNA polymerase



RNA polymerase is *different* in prokaryotes and eukaryotes

a) Prokaryotes

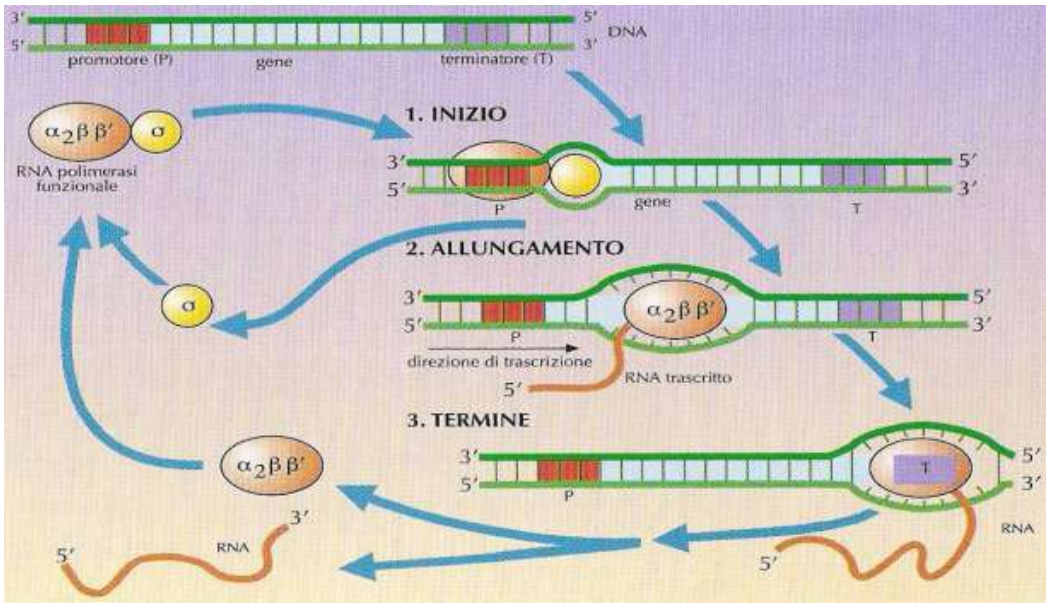
In *prokaryotes* RNA polymerase:

- Constituted by 5 subunit (2 α = α2; 1 β; 1 β'; and 1 σ)
- Transcribes all classes of cell's RNA (mRNA, rRNA, tRNA, …)

In **prokaryotes**, transcription comprises 3 different phases:

1. *Starting* transcription:
   - RNA polymerase enzyme bonds (with its σ subunit) with gene's promoter
   - subunit σ detaches and transcription starts
2. *Polymerization* of polynucleotide RNA (**elongation**):
   - Enzyme continues synthesis of RNA up to stop zone of gene (termination)

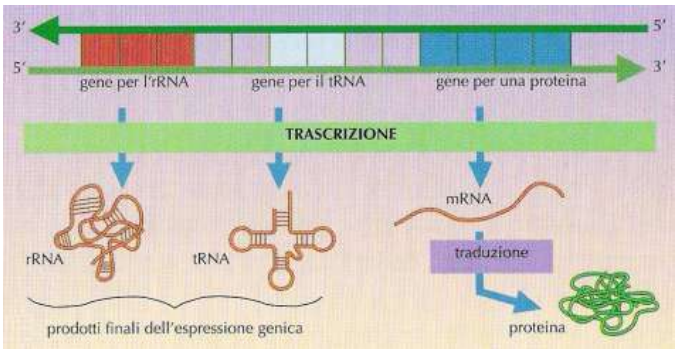
3. *Detaching* of synthesized RNA and end of transcription:
   - enzyme detaches from DNA and RNA remains free (video: *http://www.youtube.com/watch?v=toCc6ZU8D9A*)



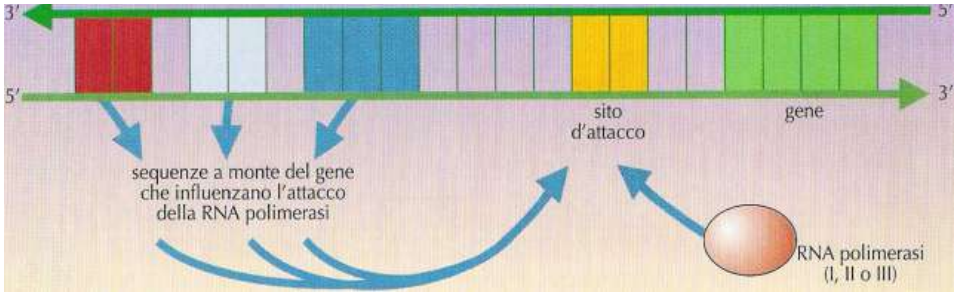
b) Eukaryotes

In *eukaryotes*, transcription is a more complex process:

- **3 different types** of RNA polymerase (RNA polymerase I, RNA polymerase II, RNA polymerase III). Each type recognizes and transcribes a different genes' set (genes for mRNA, genes for rRNA, genes for tRNA, …)

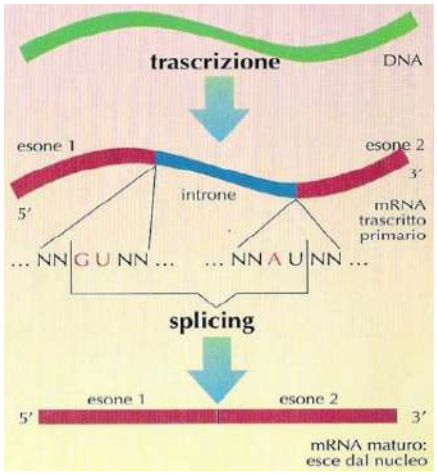


- Many other *proteins* are necessary (**transcription factors**), specific for each polymerase, that lead specific RNA polymerase linkage and gene's transcription. They bond in particular areas of DNA (sites for transcriptional factors' bond) when they have favourable tridimensional access



Video: *http://www.youtube.com/watch?v=WsofH466lqk*

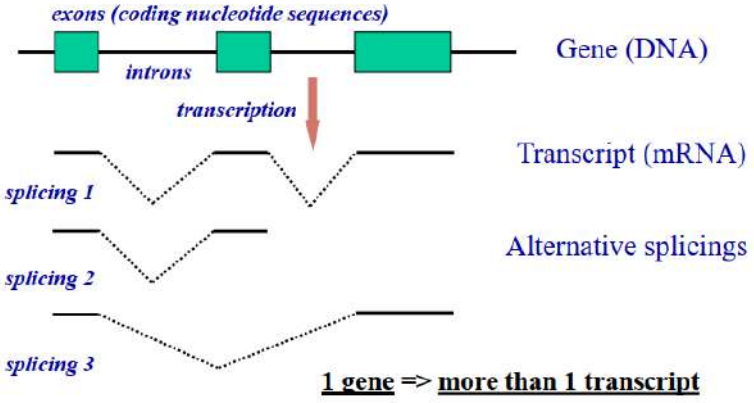
RNA polymerase synthesizes RNA in **continuous** way (cannot skip a part of mold chain of DNA)



In eukaryotes, *genes* (composed of introns and exons) are **fully transcribed**, producing a primary transcription (pre-RNA)

Complex molecular processes (***splicing***), characteristic of eukaryotes, **remove intron** sequences from pre-RNA, producing **mature-RNA**, that migrates through nuclear envelope in cytoplasm, where it is translated (video: *http://www.youtube.com/watch?v=FVuAwBGw_pQ*)

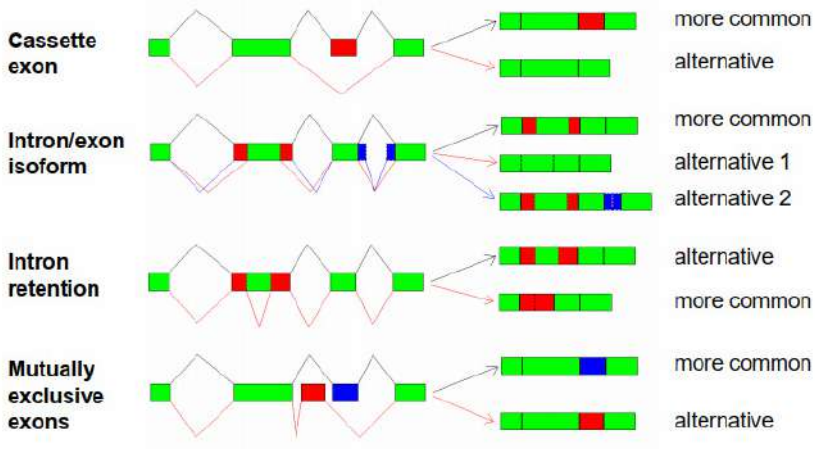
Different types of splicing can be observed (***alternative splicing***)



In a gene with alternative splicing, the *majority* of exons is *always* included in final mRNA

4 different types of ***alternative splicing*** can occur, each of them generates different final transcripts:

- **cassette exons**: full exons transcribed only in some cases
- **isoforms of introns/exons**: boundaries of introns and/or exons can be different, with clipping/extension
- **introns retention**: introns can be contained in final transcript
- **mutually exclusive exons**: different exons can be included in different final transcripts



Splicing is a not well-known process. In some cases, *molecular complexes* comprising small RNA molecules and proteins (*SNRNP* – Small Nuclear RiboNucleoProtein) cut:

- firstly in 5' end of intron by dinucleotide GU
- secondly in 3' end by dinucleotide AG

Finally, exons (become adjacent due to removal of intron) link together

There are some introns in not-mRNA genes (e.g. for rRNA) that follow GU-AG rule without using SNRNP; remove autonomously introns (**auto-splicing**); these RNA are called ***riboenzyme***

After splicing, the process of transformation of premRNA to mature-RNA is completed stabilizing mRNA (constituted by only exons) by adding:

- at initial extremity a "hat" of **7-MethylGuanine**
- at final extremity a "tail" of Adenines (**poly-A tail**)

In this phase complex phenomena can occur, that bring *degradation of mRNA* by preventing subsequent translation and gene's expression.

Mature-mRNA exits from nucleus, through nuclear envelope, and move into cytoplasm, where its translation occurs

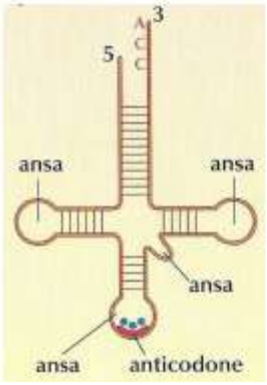
II.C.11 Different types of RNA

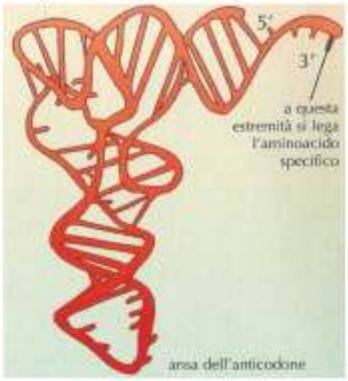
Ribosomal RNA (rRNA):

- *Structural* and *functional* components of ribosomes, big cellular organelles with ovoid shape (in eukaryotes and prokaryotes) where the *synthesis of proteins* occurs

Transfer RNA (tRNA):

- Small molecules of nucleic acid (75-95 nucleotides)
- *Function*: transport of amino acids (used in proteins' synthesis) to molecules of RNA bonded to ribosomes
- tRNA has "cloverleaf" structure, with stretches paired in double helix alternated to stretches without bases' pairing (loops)
    - Acceptor stem, (with extremity 3') where a particular amino acid can link
    - Anticodon area, consisting of a sequence of 3 bases complementary to codon of mRNA to translate (during translation, these parts get paired)





Messenger RNA (mRNA):

- mRNA produced during *transcription*, decides amino acids sequences of codified protein
- Molecules intermediate *between genes and proteins*
- In prokaryotes, mRNA are *translated* immediately after transcription
- In eukaryotes, are subjected to numerous modifications, among which **splicing** *before translation*

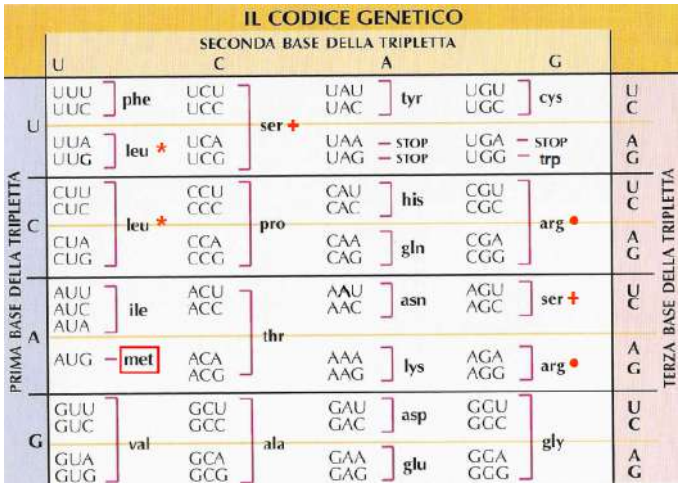
rRNA, tRNA and mRNA produced in transcription participate in translation

II.C.12 Genetic code

Genetic code: group of rules defining how the information of *nucleotides' sequence* in mRNA (4 bases A, G, C, U) is translated in *amino acids' sequence* of the codified protein (20 amino acids)

It is "universal": it is valid for almost all cells (not for mitochondrial genes of some organisms)

Discovered in 1964 with a test that demonstrates ribosome bond to mRNA only if mRNA molecule has at least 3 bases. Given there is 4 nucleotide bases, it means there are $4^3 = 64$ possible triplets (codons)



Features of genetic code:

- Each amino acid is codified by a triplet of bases
- Triplets are "read" (cfr. open reading frame - ORF) one after the other, *without any interruption*
- Each triplet can codify only one of **20 amino acids**
  Almost all amino acids codified by more than 1 triplet

Some "instrumental" triplets:

- **AUG** codifies Met (unique codon) and indicates beginning of message (**"start" triplet**)
- **UAA, UAG, UGA** (**"stop" triplets**) do not codify any amino acid, but the end of message (gene)

Genetic code based on triplets

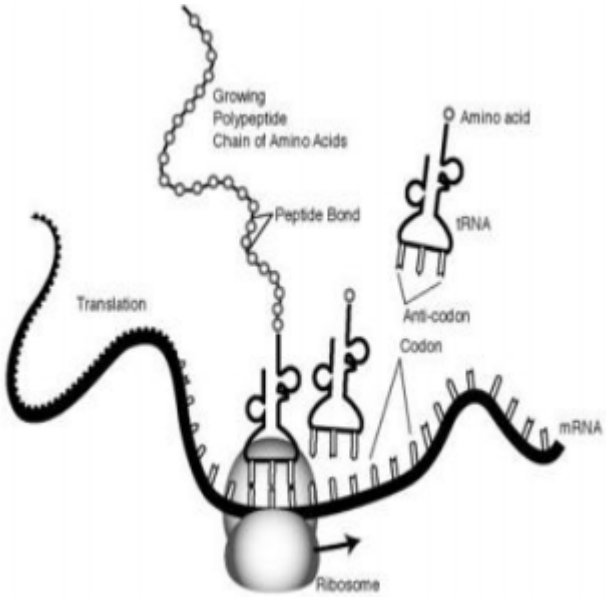
Splicing and mostly alternative splicing can alter reading of triplet inside **open reading frame** (ORF). In same space, different concurrent encoding can occur (encoding compression)

Some alternative transcripts are *tissue-specific*, that means they are expressed only in one specific type of cell (e.g. muscular, or nervous, or …)

Mechanisms of genetic code and alternative splicing allow encoding and production of many proteins with different functions from the ***same DNA*** (main cause of error of estimating 100K genes instead of 25-30K)

II.C.13 Translation

Translation is a *complex* process involving many cellular components, among which ribosomes (rRNA), mRNA, tRNA



tRNAs (adapters) are *junctions* between nucleotides of mRNA and amino acids of protein:

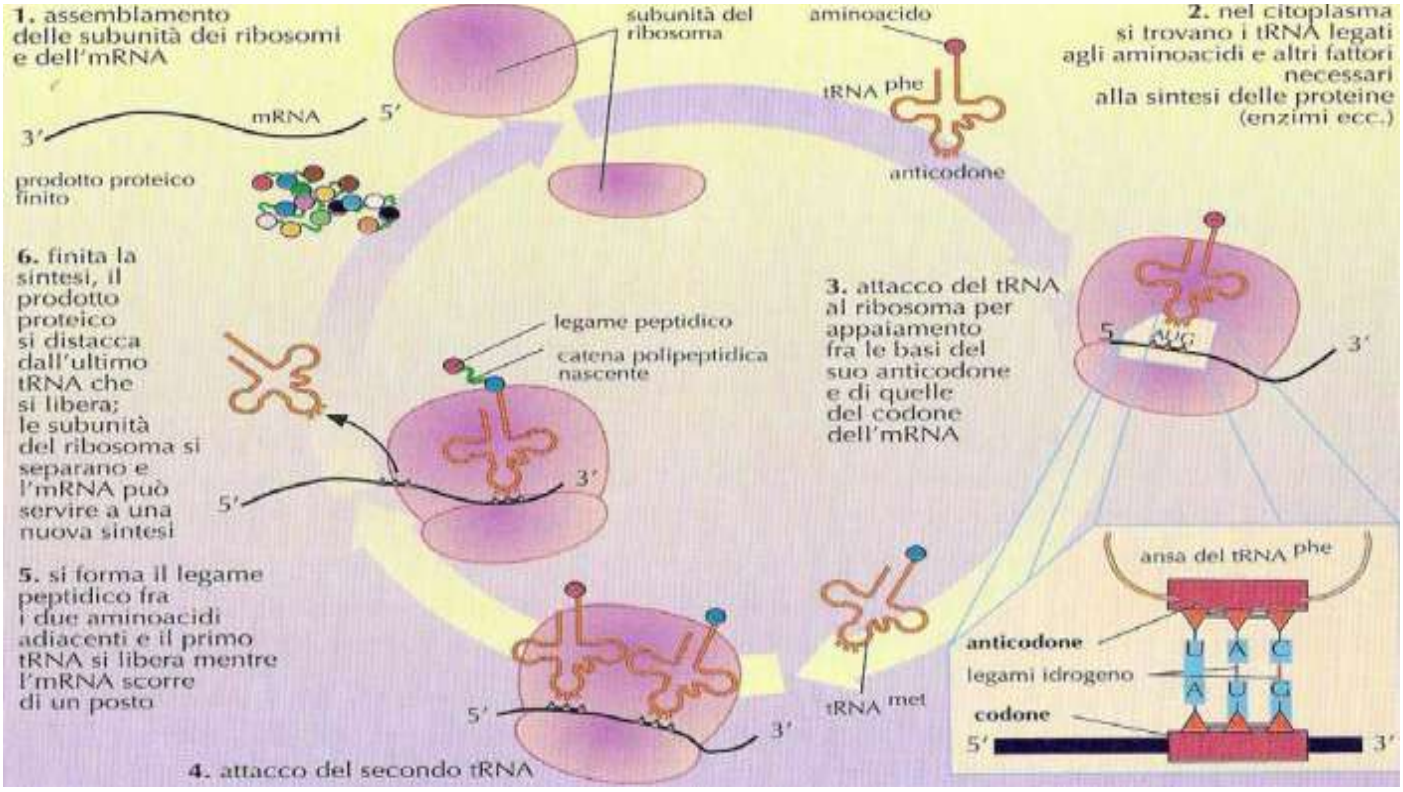
- Anticodon area bonds to codon (e.g. UUU) on mRNA that codifies a specific amino acid (Phe)
- In extremity 3' of tRNA, only that amino acid (Phe) bonds specifically and with covalent bond

Same translation in prokaryotes and eukaryotes; it has 3 phases (*http://www.youtube.com/watch?v=5bLEDd-PSTQ*)

1. Start:
   - Ribosome bonds to mRNA by starting triplet (AUG)
   - Identification of mRNA's AUG triplet by complementary specific tRNA triplet (anticodon)
   - Bond of tRNA that brings amino acid corresponding to AUG triplet (Met)
2. Synthesis:
   - Process goes on
   - Ribosome moves along mRNA
   - Only 1 triplet available at a time for bonding to specific tRNA

- Amino acids brought by tRNA are near
- When ribosome moves, a peptide bond is created between last amino acid transported by tRNA and previously transported one (last of forming peptide)
- Protein chain extends due to ribosome moving

3. <u>End:</u>
   - When ribosome reaches a stop triplet (UAA, UAG, UGA):
   - Detaches from mRNA
   - Sets protein chain free

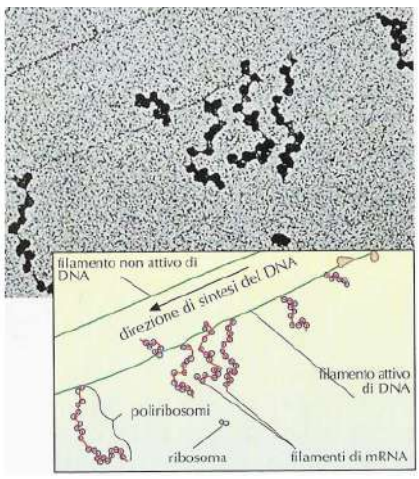


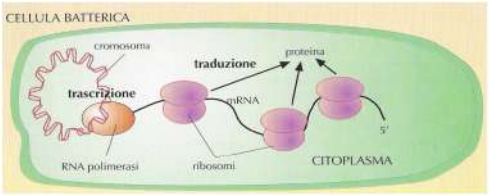
Each ribosome builds only 1 protein at a time

In bacteria (<u>prokaryotes</u>), requiring synthesis of many copies of the same protein in short time (some minutes):

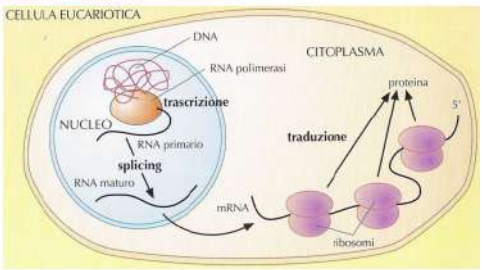
- More than one ribosome (***polyribosomes***) translate *concurrently* the same mRNA (moving in the same synthesis direction one after the other)
- Ribosomes can start translation of mRNA before its synthesis is completed (no splicing phenomena and no membrane-separated nucleus)
- In bacteria transcription and translation are paired

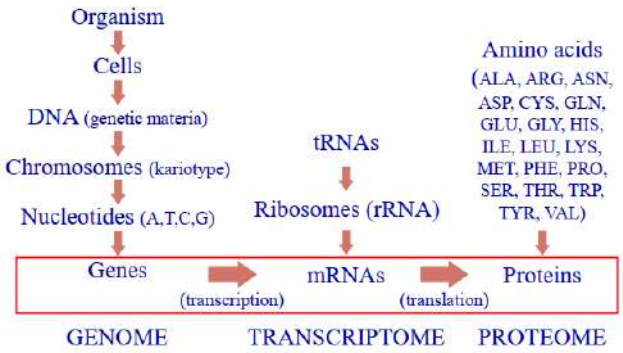
Polyribosomes and pairing of transcription and translation



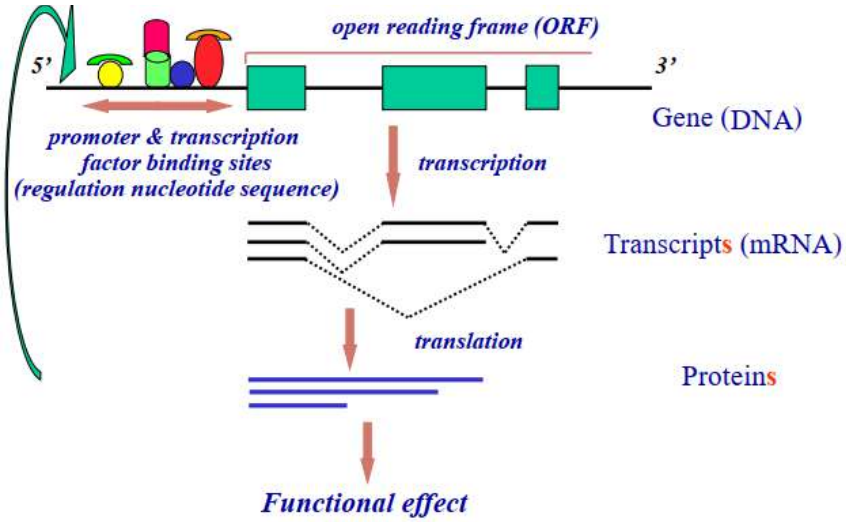


Synthesis of proteins





Summarizing video: *http://www.youtube.com/watch?v=4PKjF7OumYo*



II.C.14 Control of genetic expression

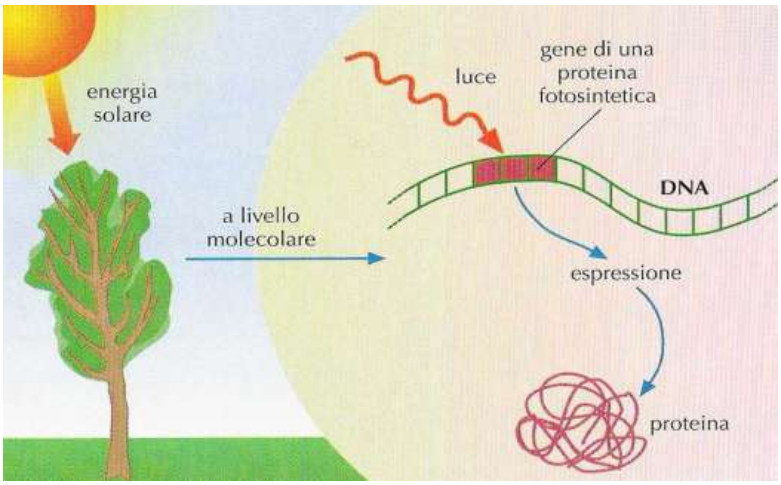
Genes of a cell codify biological information

Not all genes are always necessary for the life of a cell

Only ***constituent genes*** (codifying enzymes of basal metabolism, necessary for the life of the cell) are always expressed; other genes expressed when necessary

Expression of genes is **controlled** by cellular needs: environment conditions and functions to execute; e.g.:

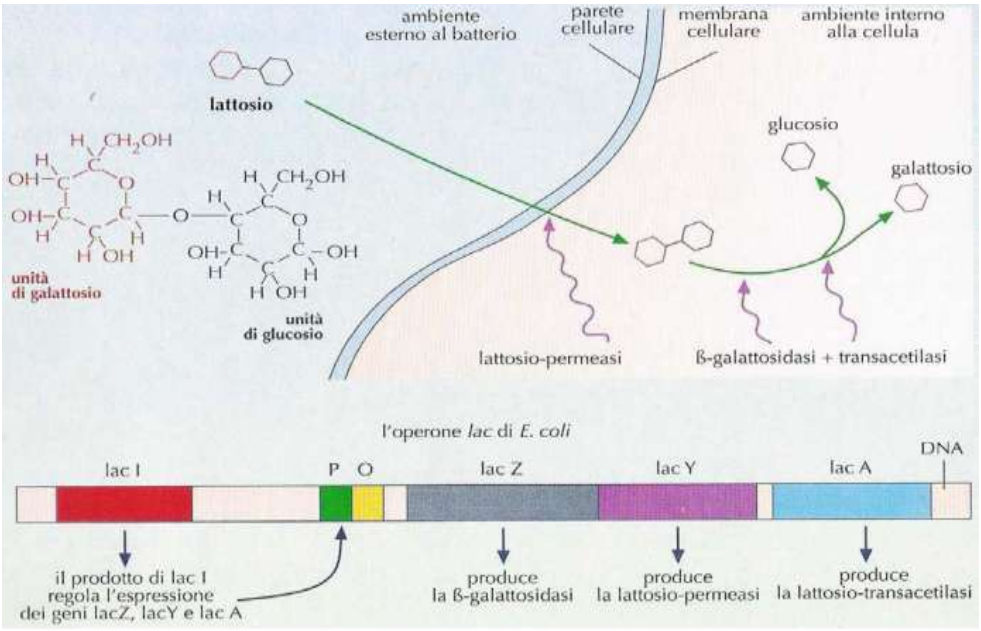
- Bacterium Escherichia coli (E. coli), living in human intestine and taking energy from various sugars, if glucose is available (easily usable), adjusts its genes by producing only enzymes for glucose, "putting out" genes that codify enzymes for other types of sugars
- In plant cells, genes of photosynthesis activated by sunlight



In multi-cellular organisms:

- *Environment* of a cell is the organism itself: single cells answer to stimuli (substances, e.g. **hormones**) produced by other cells of the organism
- In addition, there is a mechanism called "***differential regulation***" that allows one cell to divide in many specialized cells (all with the same DNA)
    - In *humans*, ~250 types of cells with different morphology and function (e.g. lymphocytes, myocytes, osteocytes, …)
    - Variety genetically established very early during growth of zygote (not "reversible"); only stem cells can differentiate in specialized cells

Genetic regulation in bacteria: Knowledge from François Jacob and Jaques Monod research (1960-64, France) on use of lactose in Escherichia coli (bacteria); Nobel prize in 1965 for proposed **model of regulation**



Lactose is a disaccharide (sugar of 2 monomers, glucose and galactose) that can be utilized when divided into the 2 components inside the cell

Splitting of lactose is realized by 3 enzymes codified by 3 genes:

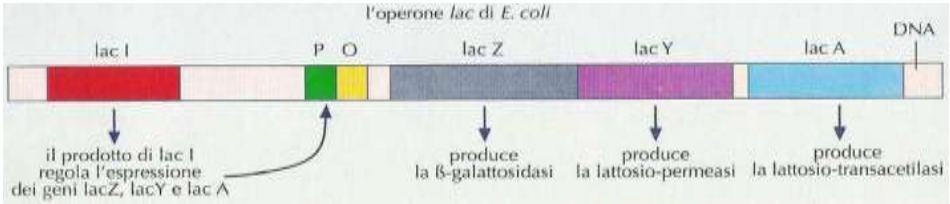
$$lacZ \rightarrow \ \beta - galactosidase$$
$$lacY \ \rightarrow \ lactose - permease$$
$$lacA \ \rightarrow \ lactose - transacetylase$$

In default of lactose, in the cell ~5 molecules of each enzyme. If lactose is the only source of energy, synthesis of enzymes is rapidly stimulated (***inducible enzymes***): in short time ~5.000 molecules

Genes *lacZ*, *lacY* and *lacA* that determine structure of 3 enzymes (**structural genes**) are consecutive on bacterial chromosome and transcribed in the same mRNA

Before the 3 genes there is gene *lacI* that ***regulates*** (down) them: its elimination brings continuous synthesis of 3 the enzymes. *lacI* codifies protein (***repressor***) that bonds to an area on chromosome called ***operator*** (o), between ***promoter*** (p) of 3 genes and the first of them (lacZ)

The whole of p, o, *lacZ*, *lacY* and *lacA* is called ***lac operon*** (operon = group of genes under common control)



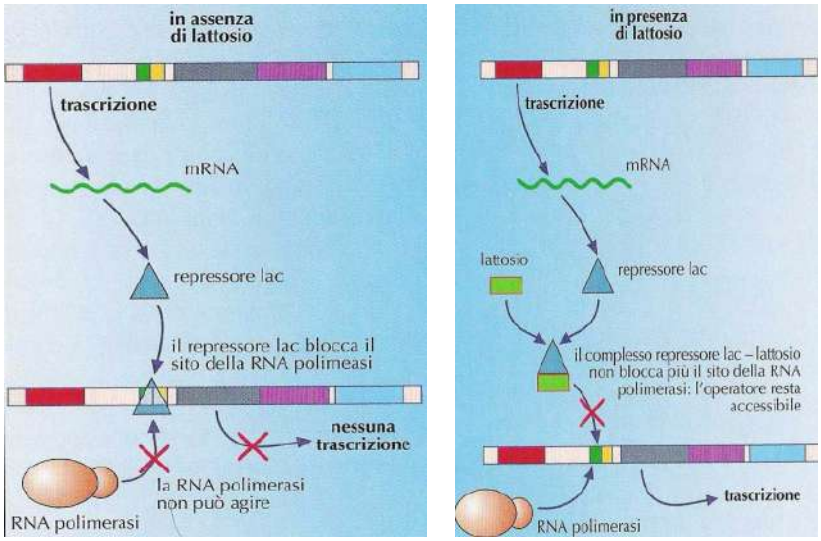
Mechanism of regulation of lactose operon:

- In default of lactose, repressor bonded to operator prevents RNA polymerase transcription of 3 structural genes
- If lactose is present, it bonds to repressor, changes its 3D conformation preventing its bond to operator. *Repressor detaches* from DNA allowing transcription of operon genes (lacZ, lacY and lacA) and synthesis of the 3 enzymes for lactose splitting
- When lactose is totally consumed, repressor bonds again to operator and synthesis of the 3 enzymes stops

Video (lac operon):

http://www.youtube.com/watch?v=oBwtxdI1zvk&NR=1

http://www.youtube.com/watch?v=aEtuaEe0C-I



Genetic regulation in superior organisms

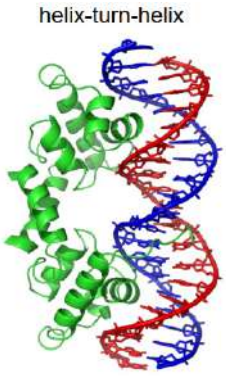
- Main mechanisms are *similar* to bacterial ones, but regulation is more complex
- Genetic expression regulated by proteins (***transcription factors***) that bond DNA sites before gene (***Transcription Factor Binding Sites***) and can allow or stop bond of RNA polymerase to promoter of gene
- Example of complexity in regulation is regulation of the protein (***metallothionein***) that protects cells from toxic effect of metals free in the environment (e.g. cadmium)
    - Small quantities of metallothionein are always present in the cell
    - Increasing of its synthesis if heavy metals are present

Regulation of metallothionein

- Gene of metallothionein is transcribed by RNA polymerase II
- Many traits of DNA before gene are involved in its expression:
    - Binding site of polymerase
    - Sequences (***enhancers***) that act allowing genetic transcription. Probably control tissue-specific expression of gene
- In addition to zones with "continuing" influence, other zones act to answer to specific stimuli (from inside or outside the cell)
- For metallothionein such zones (***elements of response to metals***) modulate transcription based on metals' concentration
- With high metals' concentration, these DNA sites are occupied by regulatory proteins (transcription factors) that activate RNA polymerase II: gene is "on" and much metallothionein is synthesized

- When metallothionein reduces metals' concentration: regulatory proteins detach from DNA, gene is "off"

Transcription factors have leading role in regulation: they have structure (or part of it) that let them enter in DNA grooves and interact with nucleotide bases (DNA binding proteins). Their more common structures are helix-turn-helix and zinc-finger



Zinc-finger video:
http://www.youtube.com/watch?v=GRL_rdB30GY

II.D Molecular genetics II (28 Sept.)

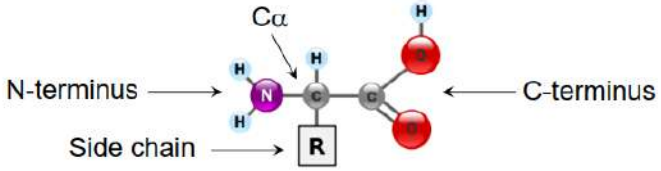
II.D.1 Proteins

Proteins: macro-polymers constituted by linking of amino acids (from 3 to various hundreds); there are *20* amino acids:

▪ Alanine	Ala	(A)	▪ Methionine	Met	(M)
▪ Cysteine	Cys	(C)	▪ Asparagine	Asn	(N)
▪ Aspartic acid	Asp	(D)	▪ Proline	Pro	(P)
▪ Glutamic acid	Glu	(E)	▪ Glutamine	Gln	(Q)
▪ Phenylalanine	Phe	(F)	▪ Arginine	Arg	(R)
▪ Glycine	Gly	(G)	▪ Serine	Ser	(S)
▪ Histidine	His	(H)	▪ Threonine	Thr	(T)
▪ Isoleucine	Ile	(I)	▪ Valine	Val	(V)
▪ Lysine	Lys	(K)	▪ Tryptophane	Trp	(W)
▪ Leucine	Leu	(L)	▪ Tyrosine	Tyr	(Y)

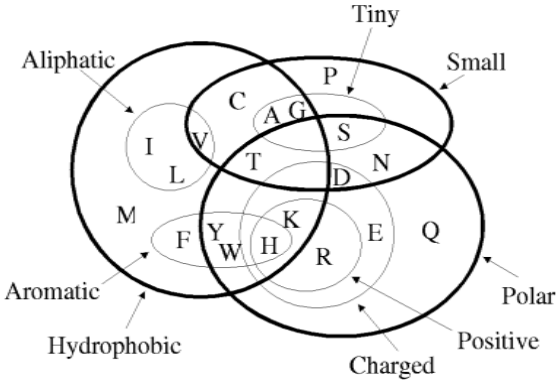
Amino acids are molecules containing 1 central atom of carbon, bonded with:

- 1 atom of *hydrogen* (H)
- 1 *amine* group (-NH2)
- 1 *carboxylic acid* group (-COOH)
- 1 *side chain* (R), that varies depending on the amino acid (Ala, Cys, Asp, …)



Side chain of each amino acid determines chemical **properties**, making it:

- hydrophobic
- polar
- acid
- …



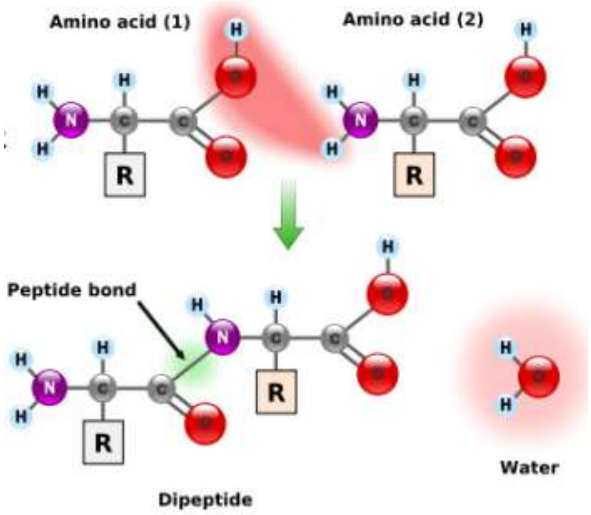
Venn's diagram of amino acids' properties (http://www.russelllab.org/aas/aas.html)

Peptide (from Ancient Greek "small digestible") is a short polymer constituted by the *linkage* of *amino acids* bonded with peptide bonds

Peptide bond: bond between N-terminus of an amino-acid and C-terminus of another one. It is *planar* and *rigid*

(Poly)peptides have 1 free N-terminus (at the beginning) and 1 free C-terminus (at the end)

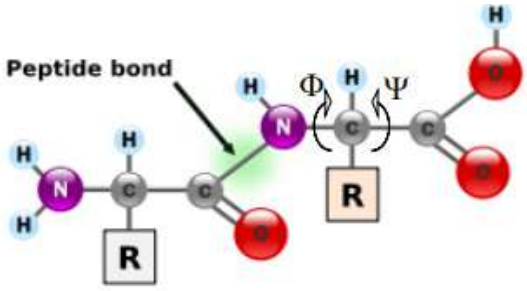
Video: http://www.youtube.com/watch?v=va0DNJId_CM



Proteins are ***polypeptide chains***. Polypeptide contains from 3 to various hundreds of amino acids

Polypeptide chain is long sequence of rigid and planar peptide groups

- 3D conformation of polypeptide is determined by torsion angles around $C\alpha - N$ (Φ) and $C\alpha - C$ (Ψ) bonds of each amino acid
- Only some values of Φ and Ψ are possible, depending on side chain of amino acid



Proteins have different **functions**, ultimate for all organisms:

- **energetic**
- **immune**
- **structural** of support (constitute backbone of cell)
- of **transport** (of oxygen, metals, lipids)
- of **identification** of genetic identity
- **enzymatic** (catalyse, that means allow the majority of cellular reactions)
- **hormonal** (lead regulative functions and transmit signals within the organism)
- **contractile**

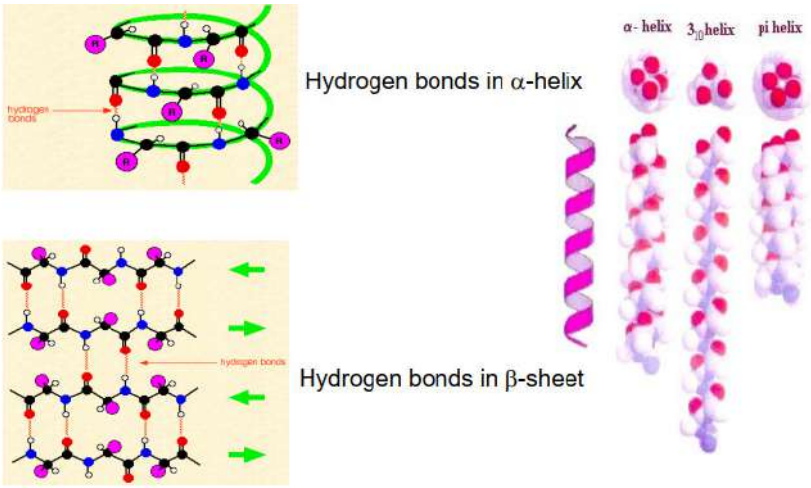
II.D.2 Structure of proteins

Function executed by protein depends on **properties** of protein, determined by:

- Properties of components (amino acids)
- Mainly of 3D structure adopted by the protein

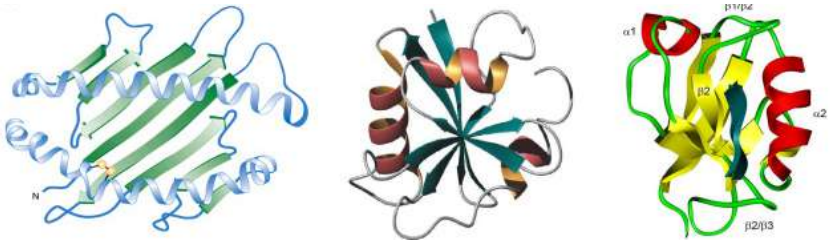
Structure of protein is structured in 4 related levels:

- ***Primary structure***: *sequence* of amino acids. E.g., AFYYWVTNMACDHIRSSWAA
- ***Secondary structure***: *local 3D conformation* with regular and repetitive bonding of polypeptide chain in substructures with well-defined and fixed geometric structures:
  - spiral (**α-helix**): pitch 0.54 nm; 3.6 residuals per turn; hydrogen bonds inside polypeptide chain
  - plane (**β-strand**, or **β-sheet**): in parallel or anti-parallel shape depending on direction of polypeptide chain, stabilized by hydrogen bonds between adjoining parts of the chain
  - **loops**: linkages between α-helix and β-strand (often they can rapidly change their direction); averagely in protein 40% loops and 60% α-helix and β-strand



Video of a α-helix: http://www.youtube.com/watch?v=eUS6CEn4GSA
Video of a β-helix: http://www.youtube.com/watch?v=wM2LWCTWlrE

- ***Tertiary structure***: *3D arrangement* of secondary structure elements in the environment
  - Stabilized by *hydrogen* bonds, *disulphide* bonds, *Van der Waals* forces and *hydrophobic* interactions
  - Assumed shape is the one with *lower free energy*
  - Mainly *globular* or *fibrous*
  - Defines ***properties*** and ***function*** of protein



v · d · e	Chemical bonds	[hide]
"Strong"	Covalent bonds & Antibonding	**Sigma bonds**: 3c-2e (bent bond) · 3c-4e (Hydrogen bond, Dihydrogen bond, Agostic interaction) · 4c-2e **Pi bonds**: π backbonding · Conjugation · Hyperconjugation · Aromaticity · Metal aromaticity **Delta bond**: Quadruple bond · Quintuple bond · Sextuple bond Coordinate covalent bond · Hapticity
	Ionic bonds	Cation-pi interaction · Salt bridge
	Metallic bonds	Metal aromaticity
"Weak"	Hydrogen bond	Dihydrogen bond · Dihydrogen complex · Low-barrier hydrogen bond · Symmetric hydrogen bond · Hydrophile
	Other noncovalent	**van der Waals force** · Mechanical bond · Halogen bond · Aurophilicity · Intercalation · Stacking · Entropic force · Chemical polarity
other	Disulfide bond · Peptide bond · Phosphodiester bond	
Note: the weakest strong bonds are not necessarily stronger than the strongest weak bonds		

Types of chemical bonds and their strength

Protein domain: part of sequence and protein structure that can *exist, work, and evolve independently* of the remainder of protein chain. Each domain has *stable 3D structure* and folds always in the same way, independently of environment

It has length of about 25 to 500 amino acids (among the shortest ones is the zinc finger, ordinary in DNA binding proteins, stabilized by ionic or disulfide bonds, *http://www.youtube.com/watch?v=GRL_rdB30GY*)

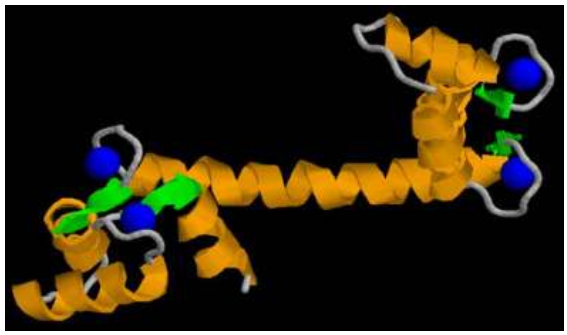
It can be present in proteins *evolutionary-related.* Often it corresponds to a ***functional unit***, e.g. EF hand domain of calcium bond. Many proteins comprise *more than one domain*

The protein Calmodulin with Calcium binding domain

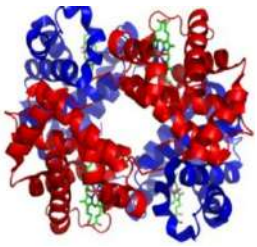
- calcium: blue
- alpha-Helices: orange
- beta-sheets: green

Clearly visible the helix-turn(sheet)-helix structure of four Ca^{2+}-binding sites

Picture from protein 1CLL in Protein Data Base (PDB) [*http://www.rcsb.org/pdb/*]

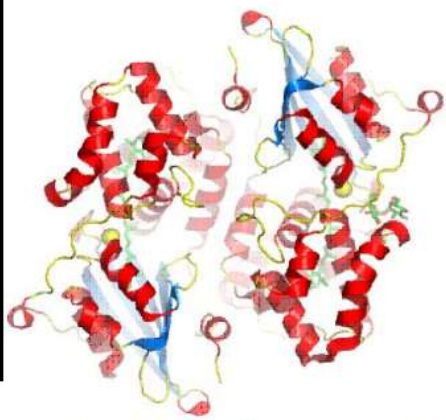


- ***Quaternary structure***: *spatial organization* of multi-subunit complexes (two or more polypeptides with defined tertiary structure, linked in specific way by noncovalent weak bonds, such as hydrogen bonds, Van der Waals forces)
  - phosphorylase (4 sub-units)
  - *hemoglobin*, has the function to carry oxygen and iron in blood (2 sub-units)



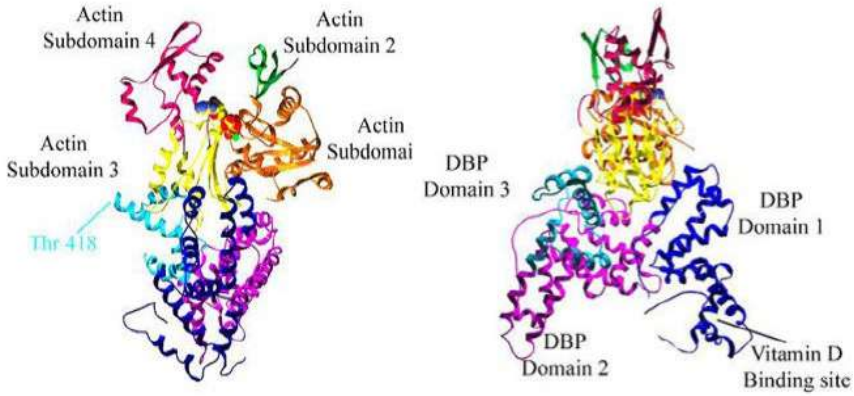
Sub-units α (red) and β (blue) of human hemoglobin: in green, iron-containing heme groups

LSm Hfq protein (LSm is a family of RNA-binding proteins) with hexamer torus



Orange carotenoid protein from *Arthrospira Maxima*

Ribbon representations of two orthogonal views of the DBP-actin complex



Primary structure highly determines *tertiary* one. From amino acids' sequence it is possible to obtain "*tertiary structure's predictions*" (i.e., folding) with specific software (there is also experimental game http://fold.it/portal/)

3D conformation is vital for biological activity of protein. ***Denatured*** protein (it has lost its tertiary structure, even if it maintains its primary one); does not execute its function, unless tertiary structure is restored

Two proteins are called ***isoform*** if they differ for little details, due to *alternative splicing* or to *polymorphisms* (SNPs)

II.D.3 Genetic mutations

During **duplication of DNA** it is possible to have variation in the sequence of nucleotide bases (***mutations***) that are transmitted to *offspring* (***mutants***)

- Generally *rare* (1:10K–1M individuals)
- Mainly *spontaneous* and accidental in all organisms
- *Can* generate individuals with new characteristics

Biological phenomenon vital for evolution (increases genetic variability of populations)

Can also be pathogenic:

- Direct cause of abnormal phenotype
- Increased susceptibility to a pathology

Mutations *can be lethal* for single individual:

- *All* organisms have various cellular mechanisms to fix possible damages to DNA
- Low % of codifying DNA on total DNA decreases probability of mutation in codifying areas (potentially lethal)

Non-lethal mutations are *transmitted* to offspring, introducing between individuals of a species many little and big differences in DNA sequence

- Are called ***polymorphisms*** [nucleotide, or of DNA] (or allelic variants) when frequency in population is greater than **0.01%**
- Some genes (e.g. HLA) are very polymorphic, and alleles of different individuals can have very different sequence

In an individual, the *majority of mutations is inherited*, since percentage of new mutations is low

Genome of an individual is **unique**. Above all, *tandem repetitions* of stretches of DNA sequence. Genome of eukaryotes contains many *stretches* (~50%) of short repeated sequences

In some points of DNA, *number* of these repeats is highly *variable among individuals* and within the chromosome pairs of an individual (heterozygosity)

To evaluate variation *number* in some DNA points allows identification of an individual with good reliability (*forensic identification*)

Single Nucleotide Polymorphism, or SNP ("snip"), is the variation of *1 single nucleotide* in an individual's DNA sequence (e.g. AAGGTTA → ATGGTTA)

Averagely, frequency of SNPs in human DNA > 1%. In humans, SNPs mainly out of codifying region (only 3-5% of human DNA codifies proteins)

SNPs in codifying regions have *more* chance of altering biological functionality. Change in sequence of bases can determine different encoding of amino acids, with possible different biological meaning

- CCU → Pro and CCC → Pro
- AAG → Lys and GAG → Glu

Video: http://www.youtube.com/watch?v=kp0esidDr-c

SNPs (groups of) are likely to be good **biological markers**

Problem: very big number of SNPs (*variables*). It would be necessary a huge number of cases (individuals, that means observations) for a *statistically relevant* association

Often near SNPs appear *together* in different individuals. ***Haplotypes***: *group* of SNPs always present together

Many common human pathologies are caused not by a single genetic variation, but by complex interactions among genes, environment, and lifestyles (**genetic predisposition** and **non-genetic factors**)

II.D.4 Genetic susceptibility

Various types of *epidemiological studies* are used for evaluating ***genetic susceptibility*** to a pathology

- *Linkage* and gene-pathology association (http://www.phgfoundation.org/tutorials/variantsDisease/)
- "*Twin*" and "*adoption*" (http://www.phgfoundation.org/tutorials/twinAdoption/)

It is useful to define the ***penetrance*** of a pathology, both at individual and population levels (http://www.phgfoundation.org/tutorials/penetrance/)

For each study and genetic test realized, it is important to check analytic and *clinical* ***validity***, clinical ***utility***, ethical, legal and *social* ***implications*** (http://www.phgfoundation.org/tutorials/acce/)

Studies of linkage and of **association**: they have the goal of identifying differences in SNPs patterns between a group of healthy subjects and a group of pathological individuals, pointing out which pattern is more probably associated with the gene responsible for a pathology (**class discovery**)

This *pattern* can then be utilized for genetic screening of the pathology (**class prediction**)

Genetic factors influence the answer to pharmacological therapy; analysis of SNPs can help understanding why

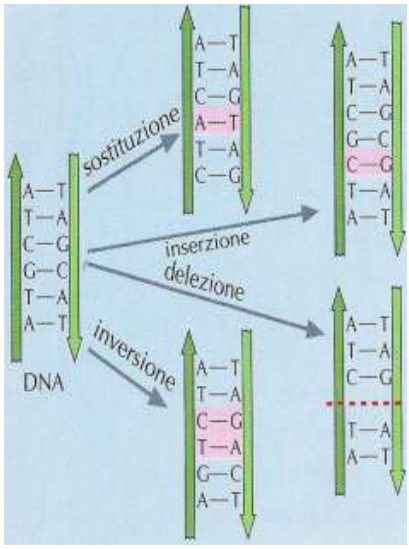
Genetic tests can indicate if a drug can have good action on an individual, or if it possible to have bad reaction (***personalized medicine***)

II.D.5 Types of genetic mutations

3 classes of mutations (depending on type of change):

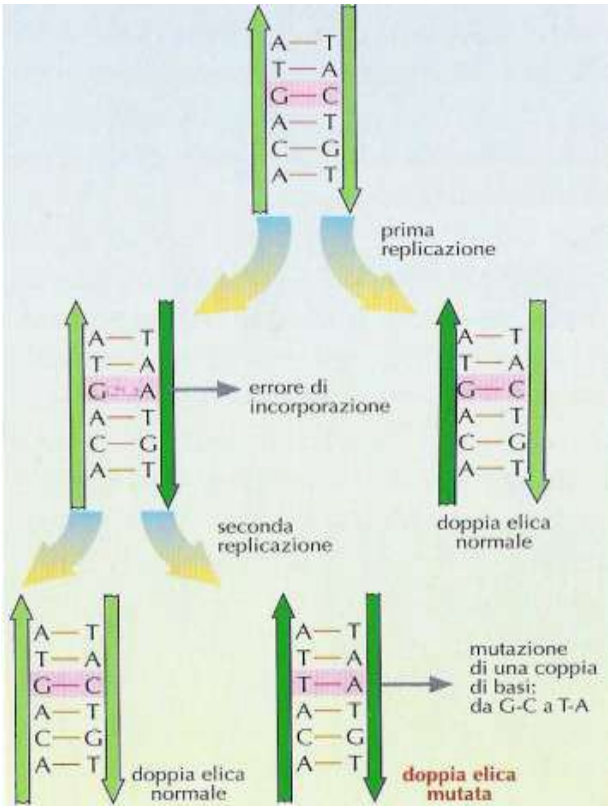
- ***Genetic***: alter *gene's structure*; also change in a single pair of bases (SNP)
- ***Chromosomal***: alter *chromosomal structure*
- ***Genomic***: alter *number* of *chromosomes*, characteristic of the species including mutant

a) Genetic mutations



Genetic mutations can derive from different alterations:

- **Substitution** of a base with another in the nucleotide sequence
- **Insertion** of *one* base in nucleotide sequence
- **Deletion** of *one* base in nucleotide sequence
- **Inversion** of *short nucleotide sequence*, due to excision of a stretch of double helix followed by re-insertion in opposite direction



Substitution for pairing error

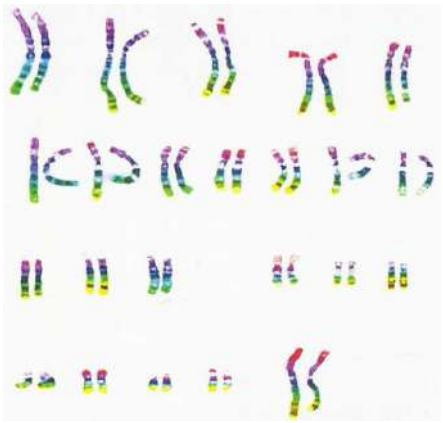
Mutation can be *effective* or not on phenotype:

- ***neutral*** (or ***silent***) if it brings a new codon codifying the *same* amino acid
- ***missense*** if it brings a new triplet codifying a different amino-acid

Different amino acid in produced protein can have *effect* (more or less strong, or none) on *phenotype*, depending on the role of the amino acid in the *structure/function* of the protein

If the amino acid has *important* role, but protein not essential for life, cell survives; otherwise mutation is **lethal**. There are also useful mutations for the cell (e.g. mutant bacteria resistant to antibiotics)

b) Chromosomal mutations

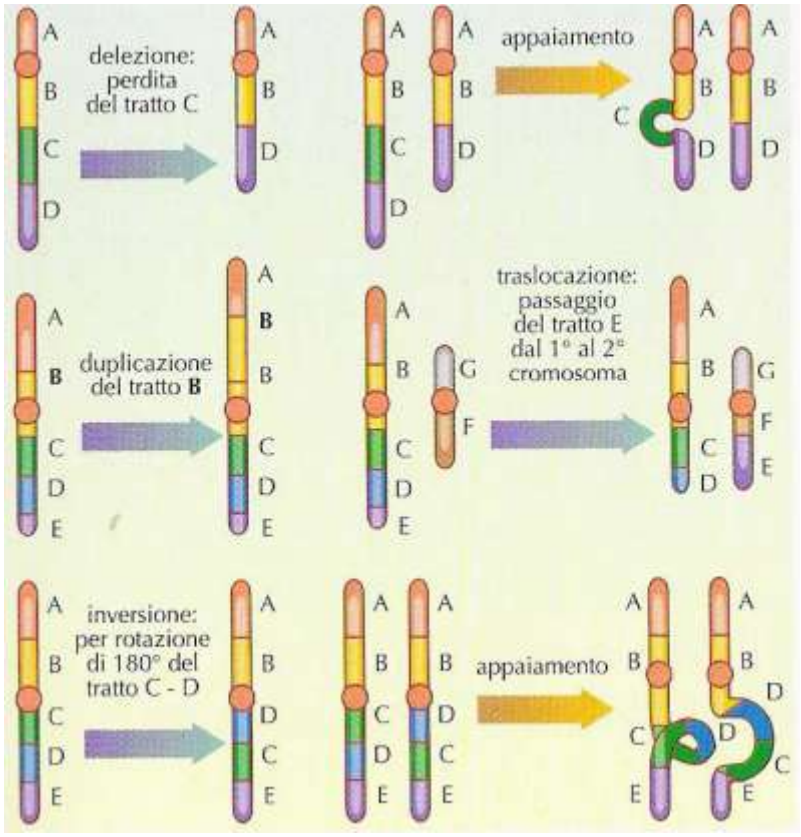


Chromosomal mutations: changes in chromosomal structure compared to normal karyotype

Easily detectable with optical microscope in dividing cell. Coloured with specific technique, assume "band" coloration

Main types of *chromosomal anomalies*:

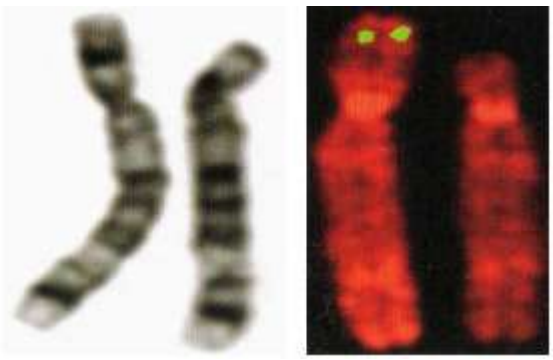
- **Deletion**
- **Duplication**
- **Inversion**
- **Translocation**



Deletion: *loss* of a stretch of chromosome of variable length. Pairing of 1 normal to 1 incomplete chromosome brings formation of *twisted structure*.

Phenotypical consequences depend on genes lost

- In *homozygotes*, deletions often *lethal*
- In *heterozygotes*, effects can be (partly) *balanced* by normal genes on *homolog* chromosome



Syndrome of the "cat cry" (affected babies cry as cats): deletion on one chromosome 5

Duplication: *doubling* of a stretch of chromosome (generally less harmful than deletions)

- In **tandem** when segment is repeated in the same direction
- **Inverse** when duplication has opposite direction

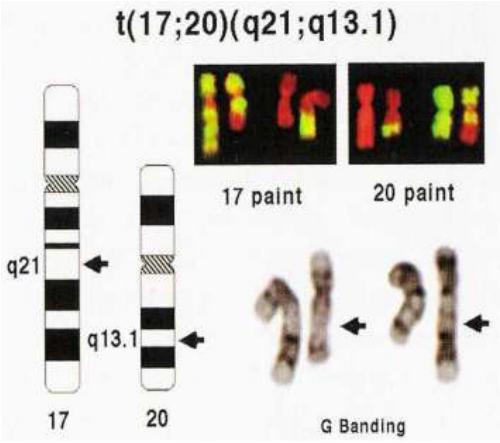
Inversion: change of order of two or more genes in a chromosome. It is caused by 2 breaking, 1 rotation of 180° of excised DNA stretch and 1 reunion

When two homolog chromosomes (one of them with duplication) pair, twisted or ring structures are created

Translocation: alteration of structure due to detachment of a stretch of chromosome and attachment to non-homolog chromosome (video: http://www.youtube.com/watch?v=eUZYACO236c)

- **simple**
- **mutual** when involves 2 chromosomes

In human, involved in tumour onset



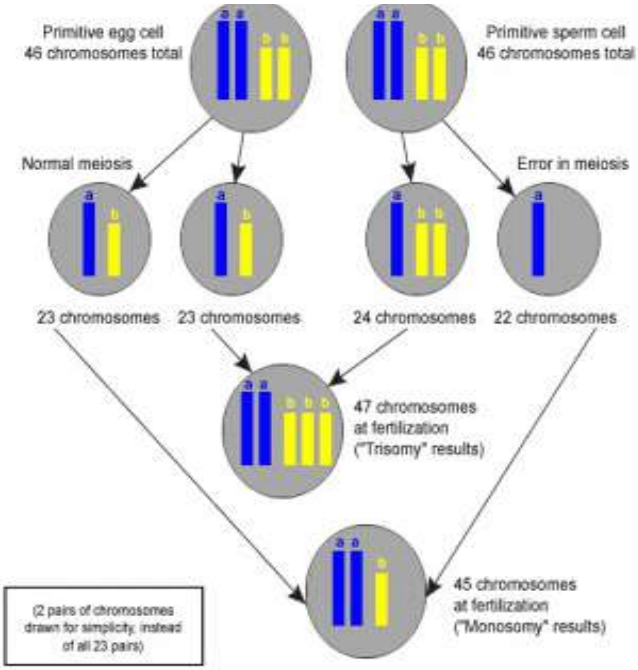
c) Genomic mutations

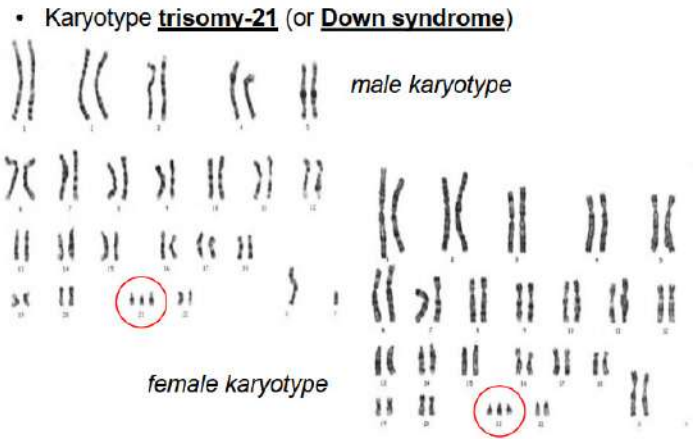
Genomic mutations concern *total number* of chromosomes in each cell of an individual

- **Polyploidy** (or euploidy) if number of chromosomes is multiple of haploid arrangement (3n, 4n, …)
  - In animals more dangerous
  - In plants more common
- **Aneuploidy** if loss or addition of 1 or few chromosomes
  - In animals: strong alterations in phenotype
  - In plants more common
  - E.g. Down syndrome, or aneuploidies of sexual chromosomes (e.g. *Turner syndrome* and *Klinefelter syndrome*)
    - **Turner's syndrome** (due to *lack of one X chromosome* in female sex; it brings short height and sterility; incidence 1:5000 women)
    - **Klinefelter's syndrome** (due to *one extra X chromosome* in XXY individuals; it brings male aspect, small testicles and developed breast, tall height, and backwardness; incidence 1:500-2000 men)

From errors in *meiotic* process, e.g., failed disjunction in pair of homolog chromosomes

- 1 gamete has *pair* of chromosomes
- 1 gamete without chromosomes
- From matching with normal gametes, ***trisomic*** (3 chromosomes) and ***monosomic*** (1 odd chromosome) individuals are obtained. E.g., trisomy-21 (or Down syndrome)

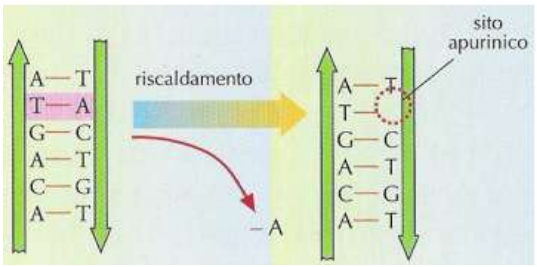


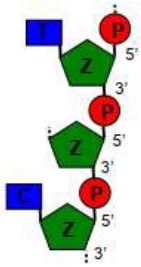


II.D.6 Mutagens

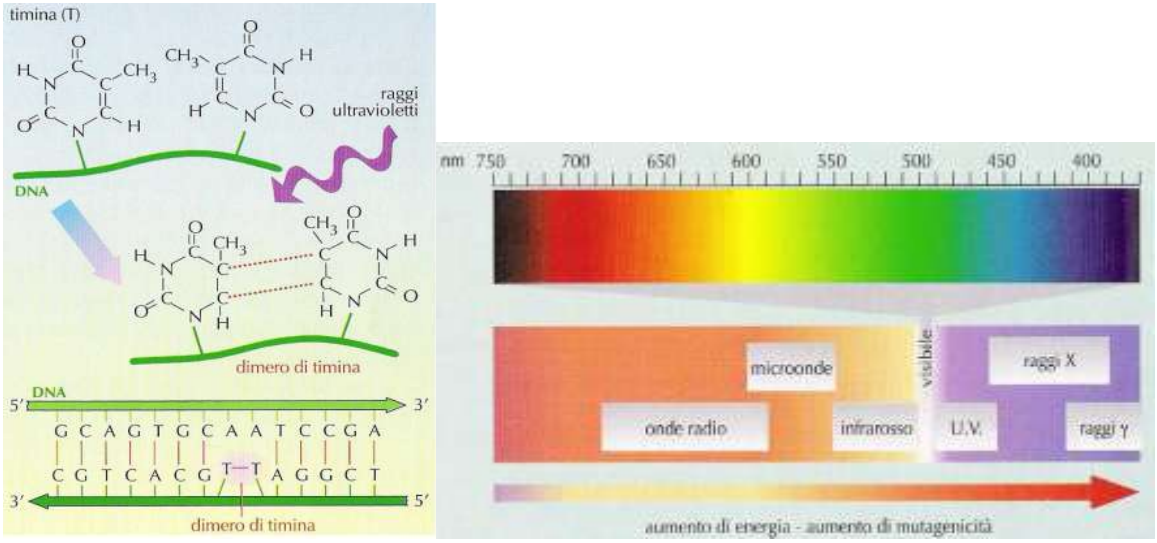
Frequency of mutations can increase (up to 1:100-1000 individuals) if organism is exposed to substances and radiations (**mutagens**) that interact with DNA and can induce changes in nucleotide sequence

- **Physical**: *radiations* with different wavelength
  - *Heat*: break bond between A or G base and sugar, with loss of the base (apurinic site)



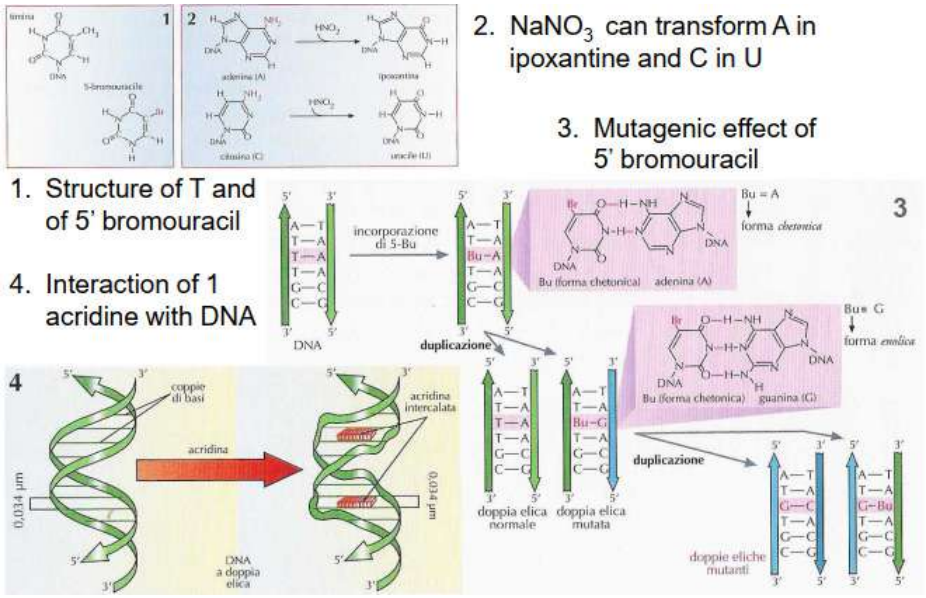


  - *High energy radiations* (wavelength < 30 µm)
    - ***Ionizing radiations*** (cosmic rays, γ rays (emitted by radium (RA)), and X rays): penetrate cellular tissues and *ionize* ($\sim$ charge + or -) molecules
      - Can provoke *breaking of DNA*, loss of bases, bonding between helixes (*cross-link*), mainly in dividing cells
      - Used for tumoral treatment (cells with accelerated reproductive activity)
    - ***Ultraviolet radiations*** (minor energy, but wavelength $\sim$ 26 µm absorbed by DNA bases)
      - Can provoke formation of *dimers* of T or C (if 2 T nearby -> dimer T-T, that prevents pairing of A on the other helix, with distortion of DNA molecule)



Ultraviolet radiations can form dimers T-T, that prevent pairing of A on the other helix, with distortion of DNA molecule

- **Chemical** (e.g. iprite, constituent toxic gas)
  - *Analogs of nucleotide bases*: can *substitute* normal bases during replication (e.g., 5-bromouracil, brings SNPs)
  - *Reactive of nucleic acids*: chemically react with DNA bases altering them. They are the most numerous, some in products of wide use, e.g. nitrite in antioxidants in food (NaNO3 brings SNPs)
  - *Intercalating agents of DNA bases*: synthesized substances (e.g. acridine, used in colorants) that inserts between bases. They bring genetic mutation of triplets reading (frame shift)



- **Environmental mutagenesis**: presence in the environment of many mutagenic substances (e.g. tar, benzene, heavy metals, …) capable of *inducing tumours*. A survey of World Health Organization assessed about 70-80% of tumours is determined by environmental factors

II.D.7 Fixing DNA damages

All living beings have various cellular **mechanisms** for **fixing DNA damages**, among which:

- ***Photoreactivation of dimers T-T***: Due to enzyme *DNA photolyase*, present in all cells, that cuts bond T-T if activated by light

- ***Repair towards damage's excision***: complex process that requires *many enzymes*:
  - *Detection* of damage
  - *Cutting* of DNA before and after the damage
  - *Removal* of damaged stretch of single helix (**excision**)
  - *Polymerization* of missing stretch (**DNA polymerase**) and *welding* of extremities (**DNA ligase**)

Repair towards excision is the most important mechanism for fixing common DNA damages (video: http://www.youtube.com/watch?v=CcTayxEblio)

In mankind lack or reduction of one or more involved enzymes is associated with *inherited pathology* (Xeroderma pigmentosum) that brings formation of *skin tumours* due to ultraviolet radiations present in *solar rays.*

II.D.8 Genome

Genome: entire genetic material of an organism

- It is *identical* in each *cell* of the same individual
- It is for *99% the same* in all individuals of the *same species*
- Constituted by all *possible nucleotide sequences* of a species (or rather by genomes of all individuals of the species)
- For extension, the term indicates also all products of genetic material (RNAs, proteins, …)

In *bioinformatics*, **genomic data/information**: whole of available data and information, related to *genetic material* of an organism (and/or to its products)

II.D.9 Transcriptome and proteome

For analogy with the definition of genome:

Transcriptome: *whole* of all possible *transcripts* (mRNA sequences) of an organism. In *bioinformatics*, data/information of transcriptome: whole of available data and information, related to all possible transcripts of an organism

Proteome: *whole* of all possible *proteins* (amino acids sequences) of an organism, deriving from different transcripts. In bioinformatics, data/information of proteome: whole of available data and information, related to all possible proteins of an organism

Also, other *–ome*: metabolome, …

II.D.10 Studied organisms

On October 18, 2016 **complete genomic sequences** of more than *9'700* species are known, including 4'026 viruses, 2'010 phages, 3'316 bacteria, 202 archea, 179 eukaryotes

Main studied genomes (http://www.ebi.ac.uk/genomes/):

- Human [~3'500 Mb (~750 MB)]
- Mouse [~2'600 Mb] , Rat [~2'600 Mb]
- Zebrafish (Danio rerio) [~1'300 Mb]
- Fruit fly (Drosophila melanogaster) [~120 Mb]
- Thale cress (Arabidopsis thaliana) [~115 Mb]
- Escherichia coli [~4 Mb], Yeast [~12 Mb]
- Pea [~4'800 Mb], Maize [~5'000 Mb], Wheat [~17'000 Mb]

Complexity of an organism is not related to *dimension* of its genome.

II.D.11 Evolutionary biology

Evolutionary biology is a sub-field of biology regarding the origin of species from a *common ancestor*, as well as their changes, multiplications, and diversifications over time

- In evolutionary biology, **homology** (from Greek: "to agree"): *similarity* between characters due to *descent* from a *common ancestor*. ***Similarity*** = an *observable* property that comes with a *significance* measure
- Before Darwin, homology was defined by Linneus only *morphologically* (based on anatomical structures)
- Darwin explained homology as the result of descent with *modification* from a common ancestor
- Modern genetics: homology is in *DNA sequences*

Homology among proteins and DNA is often concluded based on ***sequence similarity***, especially in bioinformatics. If two or more genes have highly similar DNA sequences, it is likely that they are homologous

Sequence similarity may arise from *different ancestors*:

- Short sequences may be similar *by chance*, and sequences may be similar because both bind to a *particular protein*, such as a transcription factor
- Such sequences are **similar**, but not homologous

Sequence regions that are **homologous** are also called **conserved**
Sequence homology may indicate common function

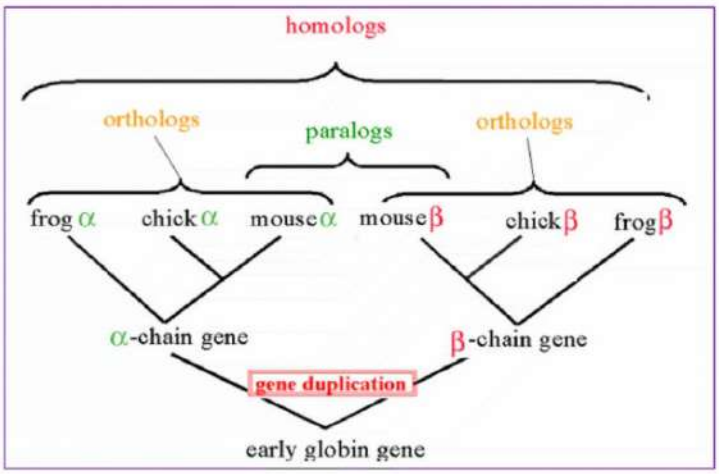
Homologous sequences are said ***orthologous*** if they were separated by a ***speciation event***: evolutionary process in which a species diverges into *two separate species*, the divergent copies of a single gene in the resulting species are said to be orthologous. ***Orthologs***, or orthologous genes, are genes in *different species* that are similar to each other because they originated from a *common ancestor*

Homologous sequences are said ***paralogous*** if they were separated by a ***gene duplication event***: if a gene in an organism is duplicated to occupy *two different positions* in the same genome, then the two copies are said paralogous

- A set of sequences that are paralogous are called ***paralogs*** of each other
- Paralogs typically have the *same or similar function*, but sometimes do not: due to lack of the original selective pressure (keeping the functions required by the species) upon one copy of the duplicated gene, this copy is free to mutate and acquire *new functions*

Homologous sequences can be divided into two groups:

- **Orthologs**: genes that share the same ancestral gene and perform the **same** biological function in different species, but have diverged in sequence makeup due to selective evolution
- **Paralogs**: genes within the *same genome* that share an ancestral gene and perform **diverse** biological functions



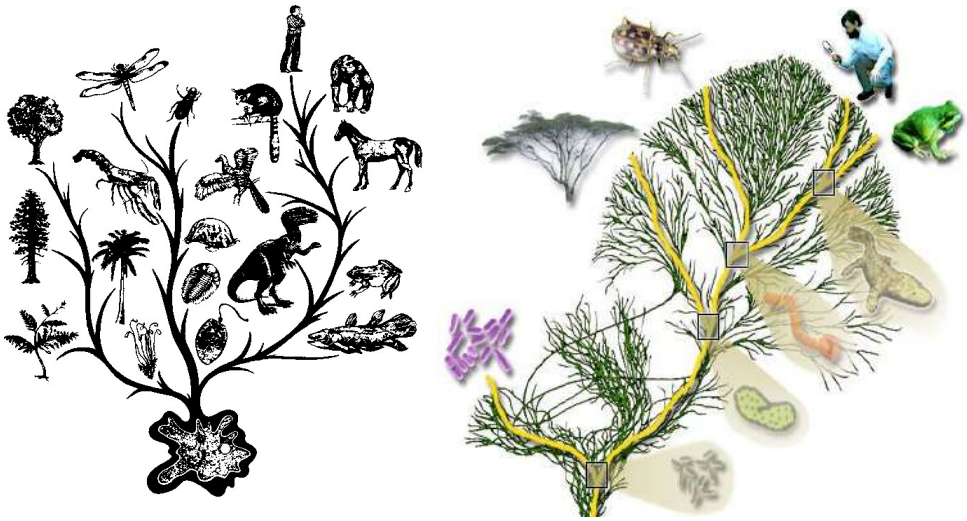
Phylogenesis or ***phylogenetics*** (from ancient Greek "origin of species"): study of life's evolution

- It is a fundamental instrument that *reconstructs* relations of evolutionary kindship of groups of organisms in any systematic level
- **Taxonomy**: *classification* of *organisms* depending on *similarities*. Highly influenced by phylogenetics, even if it remains logically and methodologically separated

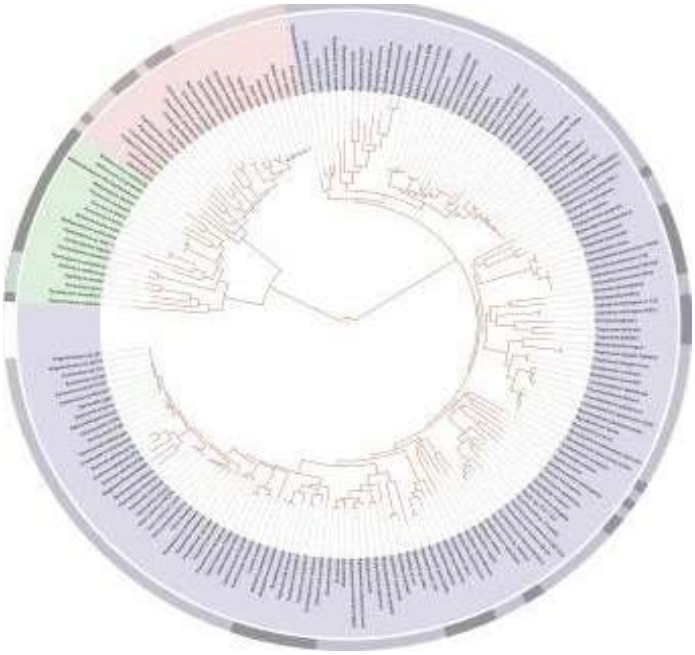
Computational phylogenetics: concerns the compilation of *phylogenetic trees* and the study of anatomic, biochemical, genetic, and paleontological data used for their construction

Phylogenetic trees: diagram that shows *relation of common descent* of taxonomic groups of organisms

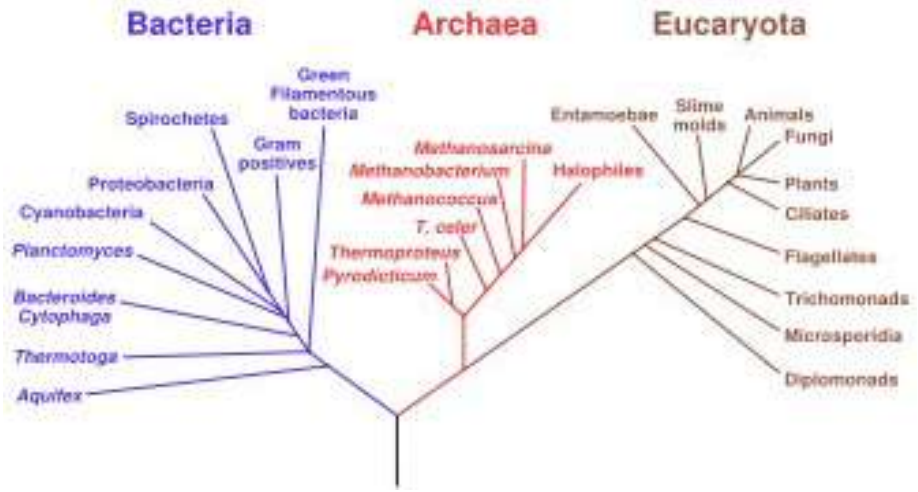
- Evolutionary vision, development of life forms through speciation, from a common ancestor (root of tree), along *different lines*, to the *present species* (leaves of tree)
- Each node (or bifurcation) represents the most recent common ancestor of subjects in subsequent nodes
- Length of ramifications can be, or not, related to *time* or *genetic changes* between to subsequent nodes



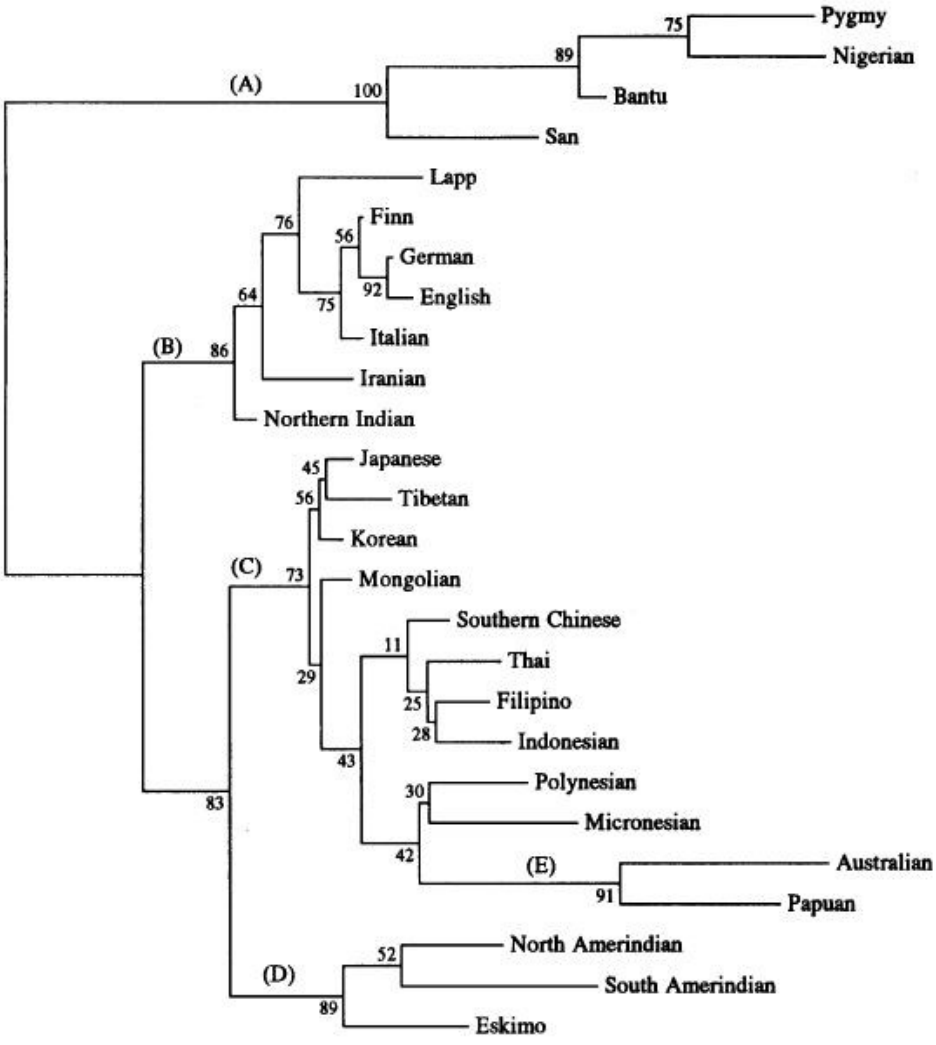
Phylogenetic trees or trees of life



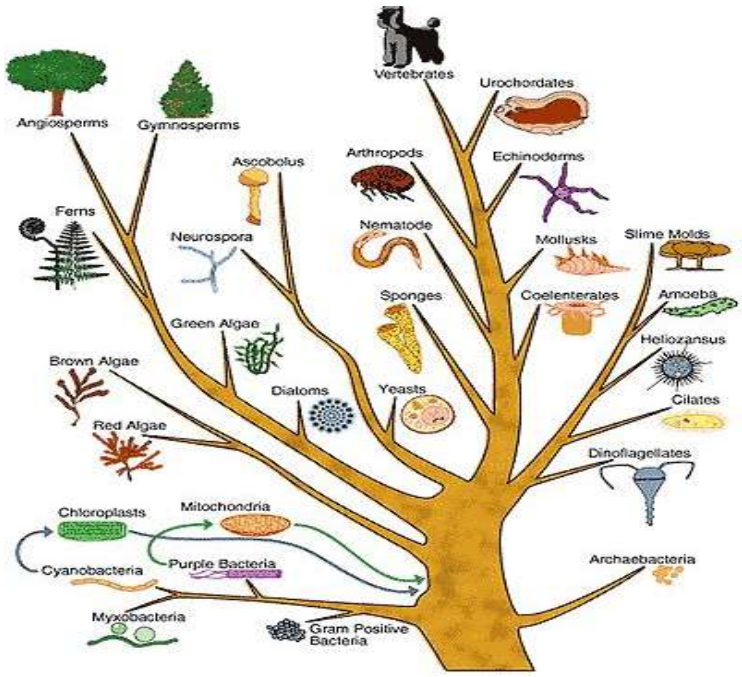
High resolution phylogenetic tree, based on analysis of completely sequenced genomes. Image generated with iTOL: Interactive Tree Of Life (http://itol.embl.de/), online viewer of phylogenetic trees



Phylogenetic tree of living organisms based on RNA analysis



Phylogenetic tree of 26 human populations; main are: Africans (A), Caucasians (B), Asians (C), Amerindians (D) and Australopapuans (E)



Phylogenetic tree of family of dogs

Phylogenetic trees are built on the base of a high number of genetic sequences; they are built using *computational methods*. Various techniques exist for building trees using different methods for *sequences' alignment* (e.g. ClustalW), referring or not to an evolutionary model

Many techniques used to *identify the best tree* on the basis of data, with high computational *complexity NP* (Nondeterministic Polynomial-time). *Heuristic researches* and optimization methods are used, combined with scoring functions of trees, to identify a tree acceptably fitting given data

Phylogenetic trees based on *genomic analysis* are important to evolutionary analysis, but have some limits:

- Often do **not** represent exact evolutionary history of a gene or organism
- Are based on data disturbed by different factors, easily confusing analysis based on phylogenetic principles:
  - Genetic *horizontal* transfer
  - *Hybridization* between different species, very far in a tree before hybridization
  - *Converging* evolution
  - *Conservation* of genetic *sequences*

III. Techniques of biomolecular sequence analysis (5, 12 Oct.)

III.A Motivations

III.A.1 Importance of sequence comparison

Given two or more *biomolecular sequences* we would like to:

- Measure the *degree of similarity*
- Determine the *correspondence* between elements of distinct sequences
- Observe the *patterns* of *conservation* and *variability*
- Deduce the *evolutionary relationship*

To *compare* amino acids or nucleic acids in two or more sequences in *"corresponding" position* it is necessary to allocate these correspondences, i.e. to *align* the sequences. **Sequences alignment** is the most important problem together with searching for a specific sequence in a database

Biomolecular sequence comparison allows:

- **Prediction** of structure
  - If 2 amino acid sequences have *20-30%* of the same residues aligned, their 3D structure can be very *similar*
  - Shape and function follow the structure: similarity of *sequence* may result in similarity of *function*
- **Identifying** known or preserved patterns
- **Inferring** function: preserved positions may represent important functional sites
- **Phylogenetic** analysis: assessing similarities to infer evolutionary information

III.A.2 Homology versus similarities

Homology (*orthology* and *paralogy*): conclusion that two or more genes (proteins) derive from a *common ancestor* and thus from a common basic structure. Generally, sequences with *more than 100 items*:

- *Amino acidic sequences* are homologous if they share **25%** of the amino acids aligned
- *Nucleotide sequences* are homologous if they share **70%** of the nucleotides aligned

The difference in percentage is due to the difference in "alphabet": there are 20 amino acids, vs. 4 nucleotides.

Similarity: is an observable quantity that can be expressed as a percentage. There is only the ***degree*** *of similarity*, not of homology. Genes (and proteins) are homologous or not homologous

Homology search: given a sequence (query), search among *known* sequences is carried out. Homologous sequences are used to *interpret* the new sequence

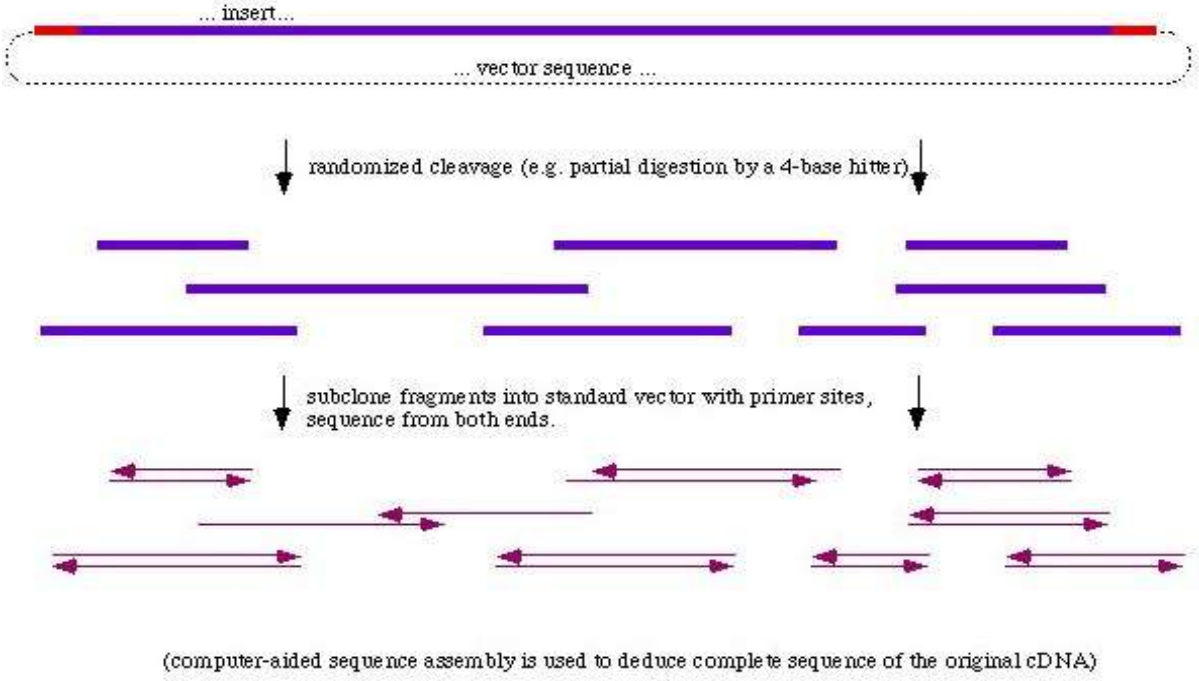
Motive search:

- ***Motive***: Set of *characters* (nucleic acids or amino acids) not necessarily contiguous in a sequence
- Adequate *vocabularies* are used, knowing their *grammars*
- It is needed to *know* (and thus to derive a priori) the relationships between *motives and functions*
- Often it is not feasible in practice, homology search is the preferred approach

Similarity searches:

- To identify *unknown* sequences
- To find other members of *families of genes or proteins*
- To find proteins that are *related*
- To identify, in proteins and nucleic acids, *regions* that are conserved through evolution (i.e. the most biologically important regions)

- Find regions that *overlap* during the assembly as a result of sequencing
  - *Sequencing* long molecule (DNA) is **not** possible
  - Take many copies of the molecule, broke them randomly (shot gun) into subparts and sequence their subparts
  - Reassemble sequences obtained by subparts



(computer-aided sequence assembly is used to deduce complete sequence of the original cDNA)

Reassembling DNA molecule according to overlapping regions between its sequenced subparts (http://www.phgfoundation.org/tutorials/dna/5.html or http://www.youtube.com/watch?v=oYpllbI0qF8) after *fragmentation* (shot gun) of its several copies

III.A.3 Sequence alignment

In order to *compare* nucleotide or amino acid sequences it is necessary to **align** the sequences

Alignment issue scenario:

- Depending on *number* of involved *sequences*
  - Alignment of *two* sequences
  - Alignment of *multiple* sequences (HP, cf. complements)
- Depending on *type* of *alignment*
  - *Global* alignment: it aligns sequences along their *whole length*
  - *Local* alignment: it defines the *longest* subsequence that gives the maximum similarity

III.B Alignment of two sequences

III.B.1 Dot matrix (dot plot)

Technique of *visual inspection* (Gibbs and McIntyre, 1970).

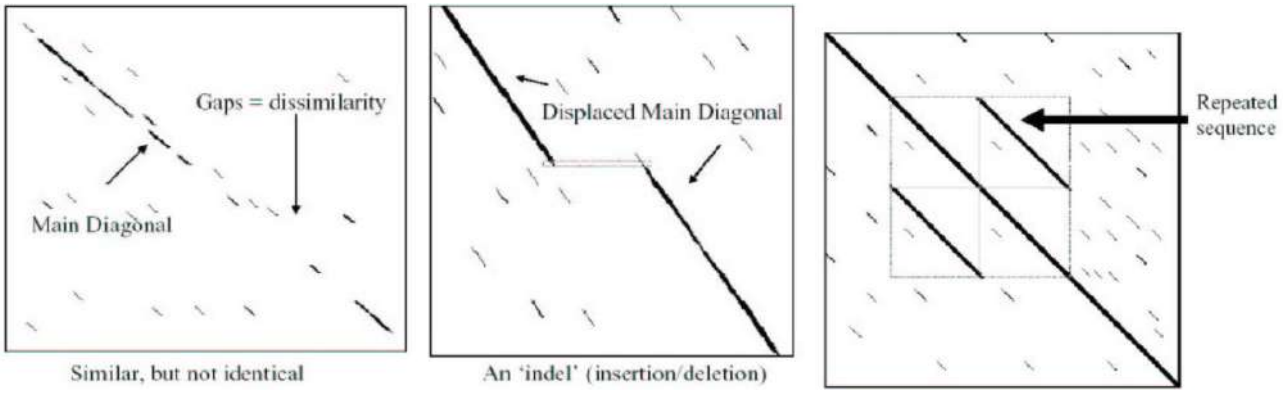
Intuitive *representation* of the comparison between two sequences

Each **point** in the matrix represents a **pair** of **identical characters** in the two sequences

Diagonal lines correspond to similarity regions

Sequence 1

Sequence 2	A	G	G	T	C
A	X				
C					X
C					X
G		X	X		
T				X	
C					X



Use of dot matrixes / Repeated sequences

Sequence 1

Complementary sequence	G	A	A	T	T	C
C						X
T				X	X	
T				X	X	
A		X	X			
A		X	X			
G	X					

Palindrome sequences

Filtering of *background* ***noise***

- *Strong* background noise, especially with nucleic acid sequences
- Filters can be used to improve *readability* of the graph
  - In particular, we seek patterns with a minimum number of correct alignments (***stringency***) in a defined window (***window size***)
  - E.g. stringency 7 with window size 11 means keeping the points that are within a window of 11 elements in which there are at least 7 exact matches
- Generally, you choose a *window* of about the *same size of the motive* you want to highlight

Dot matrix is the *first step* to compare two sequences

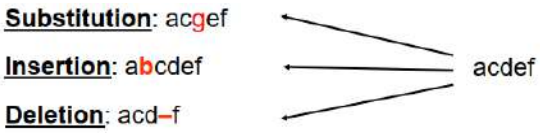
- Pros
  - All possible matches between two sequences are found
  - You can find *repeated* sequences, *direct* and *inverse*
  - Useful for *quick* visual inspection
- Cons
  - Visual inspection
  - Method *not fully automated*
  - *Image compression* for long sequences
- Practice
  - To compare *DNA*: large windows and high stringency
  - To compare *proteins*: small windows and not necessarily high stringency

III.B.2 Pairwise alignment

Besides visual alignment:

- How can we *quantitatively estimate* the degree of *similarity* (or difference) between two sequences?
- The *optimal alignment* of two sequences determines their similarity
- In general, how can we quantitatively estimate the *goodness of the alignment* between any two strings of characters?

First of all, we need to cope with the problem of *alignment of any couples* of strings of characters; we need to consider the following events:



Insertion and *deletion*:

- They are the opposite one of the other
- They are globally defined as ***indels***
- They imply ***gaps***

Given two strings acbcdb and cadbd, a possible alignment:

a c – – b c d b
– c a d b – d –

The special character "–" represents insertion of a *space* indicating a deletion in a sequence, or an insertion in the other sequence (*indels operations*)

To *discriminate* between a good and a bad alignment it is necessary to use a ***scoring*** system that manages *indels and substitution* events, and that assigns to each pair of characters in the obtained alignment (***pairwise alignment***) a value which depends on its content

The total score is the sum of the values of each pair

This scoring system can be used to estimate the *degree of correlation* between strings, and also to describe distance (or ***similarity***) between strings, giving a score:

- Lower (higher) to pairs of *identical* characters
- Higher (lower) to pairs of *different* characters, or to gaps and optimizing the alignment by minimizing (maximizing) the score function

Two sequences with high similarity are quite close, two sequences with low similarity are very distant

Example of scoring system (for DNA):

- Identical characters (match): +1
- Different characters (mismatch): -1
- Indel (gap): -1

Example of scoring system (for similarity):

a	c	–	c	b	c	d	b
a	c	a	d	b	–	d	–

- If we assign a score of +2 to each perfect match and a score of -1 to each mismatch or indel, the similarity between the two sequences considering the alignment is:

$$S = 4 \times 2 + 4 \times (-1) = 4$$

Example of scoring system (for distance):

a	c	–	c	b	c	d	b
a	c	a	d	b	–	d	–

- If we assign a score of 0 in the case of matches, of 1 in the case of substitution of characters and 2 in the case of alignment with a space (insertion or deletion), the distance between the two sequences of characters considering the alignment is:

$$d = 4 \times 0 + 1 \times 1 + 3 \times 2 = 7$$

To calculate distance between two strings it can be used:

- **Hamming distance**: defined between two strings of equal length as the *number of mismatches*
- **Levenshtein distance** (or editing distance): minimum *number of operations* (insertions, deletions, substitutions) to transform a string in the other

E.g.: to transform the string acatga in agctta you must:
- Insert one g
- Substitute one a with one t
- Deleting one g

a	-	c	a	t	g	a
a	g	c	t	t	-	a

- The editing distance between the two strings is: 3

Examples of non-unique alignments

	a c b a d c b a a d	
1)	a c b a – d – c b a a d	Editing distance = 2
2)	a c b – a d – c b a a d	Editing distance = 2
	a b c d b b b b c d a a	
1)	a b c d b b b – b c d – a a	Editing distance = 4
2)	a b c d b b b – b c d a – a	Editing distance = 4

III.B.3 Formal definitions

Pairwise alignment: Formal definitions

Given a character sequence S:

- The symbol $|S|$ indicates the length of S
- $S[i]$ indicates the i-th character of S
- Example: if S = acbcdb, $|S| = 6$ and $S[3]$ = b

Given two sequences S and T:

- Alignment associates to S and T the sequences S' and T', which can contain space symbols "-", so that:
  - $|S'| = |T'|$
  - Deleting the spaces form S' and T' we obtain S and T

Alignment score of individual character or space pairs is denoted as $\sigma(x, y)$

Scoring function of a pair of sequences is given by:

$$\sum_{i=1}^{l} \sigma(S'[i], T'[i])$$

with $l = |S'| = |T'|$

Optimal alignment of S and T is the one that *maximizes the similarity* between sequences S' and T', or *minimizes their distance*

We can create a *scoring function ad hoc* for each problem. Example: scoring function to compare amino acids has to consider the chemical-physical similarities and differences between amino acids

III.B.4 Substitution matrices

Biologically, *substitution* of nucleic acids or amino acids should be considered all with the *same weight*, or are there *more important substitutions* (less likely) and other *less significant* (e.g., purine-purine, pyrimidine-pyrimidine, hydrophobic-hydrophobic, ...)?

- Equivalent *amino acid classes* can be derived on the basis of their:
  - Chemical and physical property
  - Frequencies of substitution calculated on the protein sequences known to be homologous

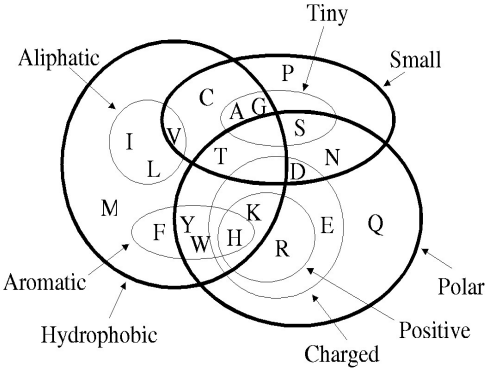
In *evaluating an alignment* of bio-sequences, one wonders if the alignment is random or biologically *significant*, and if so, *how much* it is significant

Different *weights* must be assigned to the editing, through substitution matrices

Substitution matrices assign a *numerical value* to each possible pair of characters (nucleotides or amino acids), which represents the probability of a nucleotide or amino acid to become another in a certain evolutionary time

- *Nucleotides*: simple scoring schemes are sufficient
- *Amino acids*: their chemical differences should be considered

Since a scoring function associates a numerical value to each pair of characters, these matrices can be used as scoring functions for alignment of nucleic acids or proteins



Different chemical features of amino acids

Symmetric substitution matrices (by necessity more than by choice)

There are various substitution matrices; main types are:

- ***PAM matrices*** (Percent/Point Accepted Mutations)
  - Margaret Oakley Dayhoff (1978), for amino acids
  - States (1991), for nucleotides
- ***BLOSUM matrices*** (BLOcks SUbstitution Matrices)
  - Henikoff and Henikoff (1992), for amino acids

a) PAM matrices

PAM matrices (Percent/Point Accepted Mutations): developed in the late 70s looking for mutations in closely correlated superfamilies of amino acid sequences. Accepted: mutations accepted by evolution

It was noticed that the *substitutions* that occur between closely related sequences are *not random*

It was concluded that *certain amino acid substitutions occur more easily than others*, probably because these substitutions do not significantly alter the structure and function of a protein. Homologous proteins do not necessarily need to have the same amino acids in each position

For the *construction* of PAM matrices homogeneous blocks of aligned sequences are considered

Phylogenetic trees are created (to model the evolution) and changes in proteins adjacent in the tree are *counted*

To avoid the problem of *multiple substitutions*, very similar sequences are chosen; to determine PAM matrix:

- Consider substitutions in 71 groups of protein sequences similar to at least 85%
- Counted 1572 changes, or "accepted" mutations

For each amino acid (j), count all N_{jk} changes (quantity of changes) in another amino acid (k)

- Example: How many phenylalanines (F) stay unchanged and how many change in one of the other 19 amino acids?
- A *symmetric* matrix n (n = 20) is derived

Normalize by dividing by the total number of changes $\sum_{m=1}^{n} A_{jm}$, obtaining the frequency *substitution matrix* A (normalized) in the considered blocks of sequences (if there are more blocks, counts are summed before normalizing)

$$A_{jk} = \frac{N_{jk}}{\sum_{m=1}^{n} A_{jm}}$$

PAM matrix is derived by evaluation of A_{jk} substitutions in 71 groups of protein sequences similar to at least 85%

Counted 1572 changes, or "accepted" mutations e.g.:

$A_{F,A}$: 0.0002	$A_{F,L}$: 0.0013
$A_{F,R}$: 0.0001	$A_{F,K}$: 0.0000
$A_{F,N}$: 0.0001	$A_{F,M}$: 0.0001
$A_{F,D}$: 0.0000	$A_{F,F}$: 0.9946
$A_{F,C}$: 0.0000	$A_{F,P}$: 0.0001
$A_{F,Q}$: 0.0000	$A_{F,S}$: 0.0003
$A_{F,E}$: 0.0000	$A_{F,T}$: 0.0001
$A_{F,G}$: 0.0001	$A_{F,W}$: 0.0001
$A_{F,H}$: 0.0002	$A_{F,Y}$: 0.0021
$A_{F,I}$: 0.0007	$A_{F,V}$: 0.0001

K \ j	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Substitution matrix (quantity of changes) Nj,k for PAM1

A *probabilistic Markov model* is then constructed to model the substitutions

PAM matrix (P), of *transition probability* of each amino acid into another amino acid, is defined as the matrix that in each step allows *preservation of 99%* of the sequence

Calculated from substitution matrix A as follows:

- $P_{jk} = c * A_{jk}$ *for* $k \neq j$ and $P_{jj} = 1 - \sum_{k=1}^{n} P_{jk}$
- c chosen in order that the *portion of expected changes* by the model in a step is equal to ***1%,*** calculated on the initial distribution (p_J), observed in the initial blocks
- The condition on c is obtained, by imposing:

$$\sum_{k} \sum_{j \neq k} P_{jk} p_j = c * \sum_{k} \sum_{j \neq k} A_{jk} p_j = 0.01$$

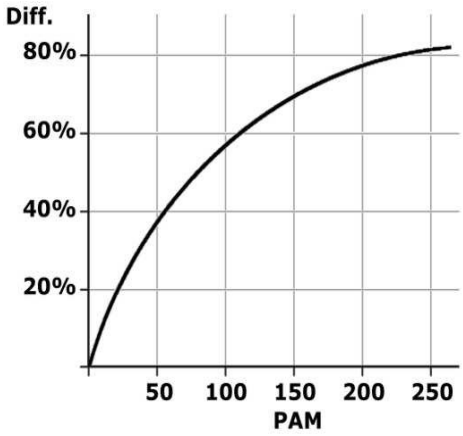
PAM matrix contains the log odd (*logit*) probability (p) of transition of each amino acid into another amino acid:

$$(p) = \log(odd(p)) \quad ; \quad odd(p) = \frac{p}{1-p}$$

In practice:

- If $PAM_{i,j} > 0$, likely transition of i in j
- If $PAM_{i,j} = 0$, random transition of i in j
- If $PAM_{i,j} < 0$, unlikely transition of i in j

PAM matrix expresses *probability of change* "in a step" (1% of sequence change)



To obtain higher percentages, multiply the matrix by itself

- PAM2 = PAM1*PAM1 = $(PAM1)^2$
- PAM10 = $(PAM1)^{10}$

PAM250 matrix is the *most used*

- It accepts a change of *250%*
- The amino acid sequences maintain at this level *20% of similarity*

Example: what is inserted in the matrix of scores when using PAM250 as substitution matrix?

- Calculate PAM250 = $(PAM1)^{250}$
- Convert PAM probability (p) in $logit(p) = \log(\frac{p}{1-p})$:
  - $PAM250(F \to Y) = 0.15$
  - Divide by the frequency of changes into F (0.04), with respect to all observed changes, then do the logarithm:

$$\log_{10}\left(\frac{0.15}{0.04}\right) = 0.57$$

  - Do likewise for $Y \to F$: $\log_{10}(\frac{0.2}{0.03}) = 0.83$
- Calculate the *score* for a change F, Y as $10 * \frac{0.83+0.57}{2} = 7$ (by convention, multiplied by 10 to get *integer values*)

ORIGINAL AMINO ACID

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	Tot.	Freq.
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9	157	8
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2	86	4
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3	80	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3	91	4
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2	83	4
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3	77	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3	94	4
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7	152	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2	72	3
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9	70	3
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13	164	7
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5	153	7
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2	30	1
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3	93	4
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4	96	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6	131	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6	109	5
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0	62	2
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2	77	3
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17	191	9
Tot.	99	100	99	99	98	100	98	99	102	88	106	107	95	104	100	100	97	172	104	101	2068	1

Substitution matrix (in %) for PAM250 (= evolutionary distance)

log odds matrix for PAM250

C	12																	
S	0	2																
T	-2	1	3															
P	-3	1	0	6														
A	-2	1	1	1	2													
G	-3	1	0	-1	1	5												
N	-4	1	0	-1	0	0	2											
D	-5	0	0	-1	0	1	2	4										
E	-5	0	0	-1	0	0	1	3	4									
Q	-5	-1	-1	0	0	-1	1	2	2	4								
H	-3	-1	-1	0	-1	-2	2	1	1	3	6							
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6						
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5					
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6				
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5			
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6		
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	
F	-4	-3	-3	-5	-4	-5	-3	-6	-5	-5	-2	-4	-5	0	1	2	-1	9
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

To measure the *distance* between *amino acid sequences* the "PAM units" can be used. Two sequences S1 and S2 have *1 PAM unit distance* if S1 can be transformed into S2 with an *average of 1 point mutation every 100* amino acids

Each PAM matrix is used to *compare sequences* that are a number of PAM units distant. E.g. the PAM120 matrix can be used to compare sequences whose distance is 120 PAM units

In a sequence the same position may *change several times* and then go back to the original character. Two sequences that are 1 PAM distant may differ by less than 1% (their similarity is at least 99%)

b) BLOSUM matrices

BLOSUM matrices (Blocks Substitution Matrices) of substitution of amino acids, similar to PAM (Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. PNAS 1992;89(22):10915-10919.)

- PAM based on *global* alignments between sequences
- BLOSUM based on alignments of **blocks of segments** of amino acid sequences closely related

A *block* is a highly *conserved* region without *gaps*

- To align sequences, the *matrix of weights* must be known. We use "unitary" matrix (match = 1, mismatch = 0)
- *Iterative* techniques have also been used

The principle is based on the *likelihood of substitutions*, based on available data

For each pair of amino acids x and y, calculate the ratio of the likelihood (e_{xy}) that x and y are aligned by chance:

- Calculated based on frequency of occurrences of x and y in the block, and portion of times that x and y are in the same column (i.e., had a likely evolutionary substitution)
  - $e_{xy} = \frac{p_x p_y}{p_{xy}}$ if $x = y$
  - $e_{xy} = \frac{2p_x p_y}{p_{xy}}$ if $x \neq y$
  - p_x and p_y probability of x and y in block; pxy probability of finding x and y paired in the same column, calculated as

$$p_{xy} = \frac{n_{xy}}{\sum_{x \neq y} n_{xy}}$$

Then calculate $-2 \log_2(e_{xy})$ and round to the nearest integer result to obtain the BLOSUM matrix

That is, obtained blocks of aligned sequences are used to count the occurrences of each pair of aligned amino acids (in the same column)

Example on only *one block* and a simplified alphabet of 3 characters (A, B, C):

sequence1: B A B A
sequence2: A A A C
sequence3: A A C C
sequence4: A A B A
sequence5: A A C C
sequence6: A A B C

Observed values:
24 characters: A: 14; B: 4; C: 6
Frequencies: A: 14/24; B: 4/24; C: 6/24
(= p_x, or p_y)
Aligned pairs: $4 * \binom{6}{2} = 60$ (= $\Sigma\, n_{xy}$)
($p_{xy} = n_{xy} / \Sigma\, n_{xy}$ =) AA: 26/60 BB: 3/60
AB: 8/60 BC: 6/60
AC: 10/60 CC: 7/60

Observed proportions are compared with *expected* proportions

Aligned Pairs	*Expected Proportions (EP)*	*Observed Proportions (OP)*	$-2\log_2\left(\frac{EP}{OP}\right)$
AA	$p_x p_y = \frac{14}{24} * \frac{14}{24} = \frac{196}{576}$	$p_{xy} = \frac{26}{60}$	0.70
AB	$2p_x p_y = 2 * \frac{14}{24} * \frac{4}{24} = \frac{112}{576}$	$p_{xy} = \frac{8}{60}$	-1.09
AC	$2p_x p_y = 2 * \frac{14}{24} * \frac{6}{24} = \frac{168}{576}$	$p_{xy} = \frac{10}{60}$	-1.61
BB	$p_x p_y = \frac{4}{24} * \frac{4}{24} = \frac{16}{576}$	$p_{xy} = \frac{3}{60}$	1.70
BC	$2p_x p_y = 2 * \frac{4}{24} * \frac{6}{24} = \frac{48}{576}$	$p_{xy} = \frac{6}{60}$	0.53
CC	$p_x p_y = \frac{6}{24} * \frac{6}{24} = \frac{36}{576}$	$p_{xy} = \frac{7}{60}$	1.80

Rounded to nearest integer: matrix BLOSUM

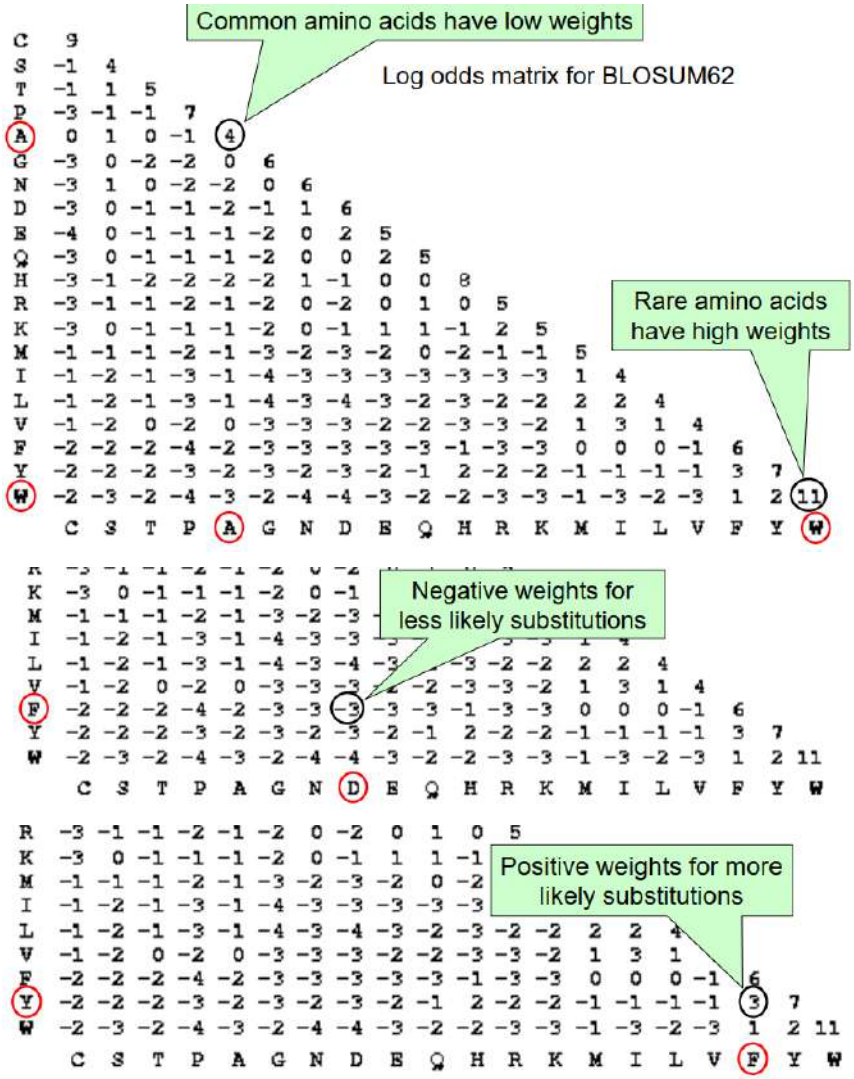
	A	B	C
A	1	-1	-2
B	-1	2	1
C	2	1	2

Entry (i, j) of the matrix is *proportional to the frequency of the substitution* of the amino acid A_i with the amino acid A_j. Values calculated based on the substitutions in a set of 2000 conserved patterns (blocks)

To avoid that very similar sequences in a block polarize the estimation, *clusters* are created in the block. Each sequence in a block weighs in *inverse proportion to the number of its occurrences* in the cluster

BLOSUM-n means that the sequences used in each cluster are similar to *at least n%*. To find relationships between sequences close in time by evolutionary point of view, a large n is used.

BLOSUM62 is the standard



c) Differences between PAM and BLOSUM matrices

PAM based on *explicit evolutionary model* (i.e. replacements are counted on branches of phylogenetic tree)

BLOSUM based on *implicit evolutionary model*

PAM based on mutations observed in *global alignment*, which includes both highly *conserved* regions and highly *mutated* ones

BLOSUM based only on *highly conserved regions* in series of alignments without gaps

Substitutions are counted differently: unlike PAM matrices, BLOSUM procedure uses *clusters* of sequences whose mutations are counted not in the same way

High numbers in names of PAM matrices indicate *high evolutionary distance*

High numbers in names of BLOSUM matrices show *high sequence similarity* (i.e. low evolutionary distance)

Example:

- PAM150 used for sequences more distant than those for which PAM100 is used
- BLOSUM62 used for sequences closer to BLOSUM50

Current research: present innovative approaches include the incorporation of information on the sequence secondary structure in substitution matrices

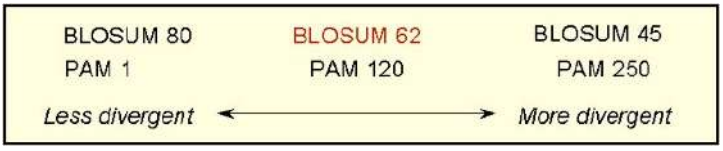
III.B.5 PAM vs BLOSUM matrices

PAM matrices:

- Based on *global* alignments of closely related proteins
- The PAM1 is the matrix calculated from comparisons of sequences with *no more than 1% divergence*
- Other PAM matrices are derived from PAM1
- The number with the matrix (e.g. PAM40, PAM100) refers to the *evolutionary distance*; greater numbers mean greater distances

BLOSUM matrices:

- The number after "BLOSUM" refers to the *minimum percentage identity* of the *blocks* used to construct the matrix; hence greater numbers mean smaller distances
- BLOSUM 62 is a matrix calculated from comparisons of sequences with *no less than 62% similarity*
- The BLOSUM series of matrices generally perform better than PAM matrices for *local* similarity searches



BLOSUM with high numbers and PAM with low numbers are both designed for comparisons of *closely* related sequences.

BLOSUM with low numbers and PAM with high numbers are designed for comparisons of *distantly* related proteins.

III.B.6 Gaps and gap penalty

Two biomolecular sequences can differ not only by the substitution of a residue with another, but also for **insertion or deletion of residues**

To align sequences is therefore often necessary to *introduce spaces* "–" in one or both sequences, also to lead them to the same length. A sequence of contiguous spaces is defined gap

A criterion to include gaps is required

- Inserting a gap lowers the alignment score
- Having to maximize alignment score, gaps are inserted only when strictly necessary

Example of opening a gap:

```
Seq. 1: a c t c a a ...
Seq. 2: t c a t c a ...
Seq. 2: - t c a t c a ...
```

Example of extending a gap:

```
Seq. 1: a c t c a a ...
Seq. 2: - t c a t c a ...
Seq. 2: - - t c a t c a ...
```

How much should *"weigh"* the introduction of a "gap"? Does deletion of n *consecutive* bases has equal weight of n *independent* deletions of 1 single base?

- More *consecutive* gaps are more likely, given that they can be due to the same mutation (of more elements)
- *Individual* gaps are due to different mutations

Usually, we distinguish between gap ***opening*** (go) and gap ***extension*** (ge), penalizing more *beginning* than an extension (of length l):

$$g_o > g_e$$

- Penalty example: $g = g_o + g_e * l$ or $g = g_o + g_e * (l-1)$
- Usually insertions and deletions of several residues at a time, rather than scattered insertions or deletions

Example (parameters recommended by ClustalW):

- For DNA: match = 1, mismatch = 0, $g_o = 10$, $g_e = 0.1 * l$ for insertions/deletions of length l
- For amino acids: BLOSUM62, $g_o = 11$, $g_e = 1 * l$

III.B.7 Computational techniques

To computationally solve the alignment problem, i.e. to find the *optimal alignment*, means:

- To maximize the number of *identical symbols aligned* (meaning in the same position)
- To minimize *insertion and deletion* events and their *length*
- To minimize the number of *different symbols aligned*

In order to find the optimal alignment, we can build all the alignments and then select the best one. Obvious method, but prohibitive (or impossible) due to very long length.

Computational cost to compare all the possible alignments requires time proportional to the product of the lengths of the two sequences (without considering gaps)

- If the two sequences are long about n the problem becomes n^2
- Also including the possibility of gaps, the problem becomes *exponential*
- If n is the number of elements of sequence A and m is the number of elements of sequence B, about nm comparisons should be done

Dynamic programming techniques allow to obtain the optimum solutions with time proportional to n^2, where n is the longest sequence length

- A dynamic programming algorithm finds the best solution by *dividing the original problem* in *subproblems* simpler to solve
- The solution of each subproblem is based on the solutions of the already solved subproblems

Used in the resolution of *optimization* problems. In this case, we must maximize the alignment score

Let's see the *grid-based method*:

Consider the following dynamic programming algorithm:

- Given two sequences S and T, we compare the first character of S with the first character of T considering the scores: $\sigma(S[1], T[1]), \sigma(S[1], -), \sigma(-, T[1])$

One can ask if it is better to align:

- The first character of S with the first character of T
- Or the first character of S with a gap
- Or the first character of T with a gap

You must choose the action that is associated with the *highest* score and *iterate* the process

We use a matrix $n * m$ $[(n+1) * (m+1)]$, with $|S| = n, |T| = m$

We put the first sequence along the top edge of the matrix and the second sequence along the left edge, leaving *1 row and 1 column free* to consider the possible *insertions* of gaps

We fill the matrix row by row:

			0	1	2	3	4	n=5
				f	i	r	s	t
0								
1	s							
2	e							
3	c							
4	o							
5	n							
m=6	d							

The value of each entry is computed with the formula:

$$V(i,j) = \max \begin{pmatrix} V(i-1, j-1) + \sigma(S[i], T[j]), \\ V(i-1, j) + \sigma(S[i], -), \\ V(i, j-1) + \sigma(-, T[j]) \end{pmatrix}$$

With initialization: $V(0,0) = 0$

Base case: $V(i,0) = \sum_{(k=0)}^{i} \sigma(S[k], -)$, and $V(0,j) = \sum_{k=0}^{j} \sigma(-, T[k])$

At each step you choose the best among the scores that would be obtained:

- By aligning character i of S with character j of T
- Or by aligning character i of S with a gap
- Or by aligning character j of T with a gap

Formal expression: alignment equation Given 2 sequences x and y to be aligned:

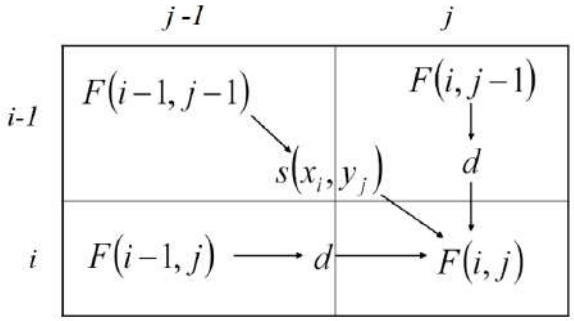
$F(V(i,j)) = F(i,j)$ is the matrix of the alignment procedure with:

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \end{cases}$$

- $s(x_i, y_j)$ the substitution matrix
- d the linear penalty for a gap

That is, on the matrix:



The value in the position (i,j) is the score of the best alignment of the first i positions of the sequence along the top edge of the matrix respect to the first j positions of the sequence along the left edge

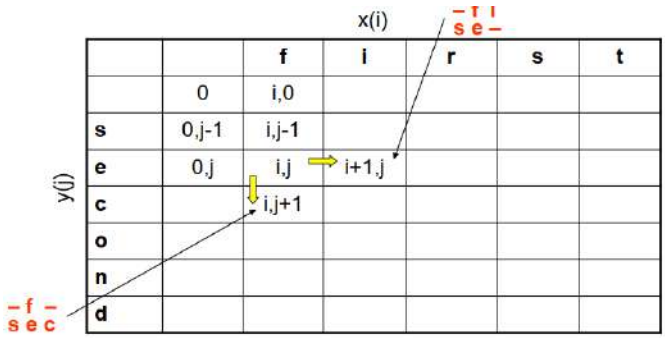
Initialization:

y(j) \ x(i)		f	i	r	s	t
	0					
s						
e						
c						
o						
n						
d						

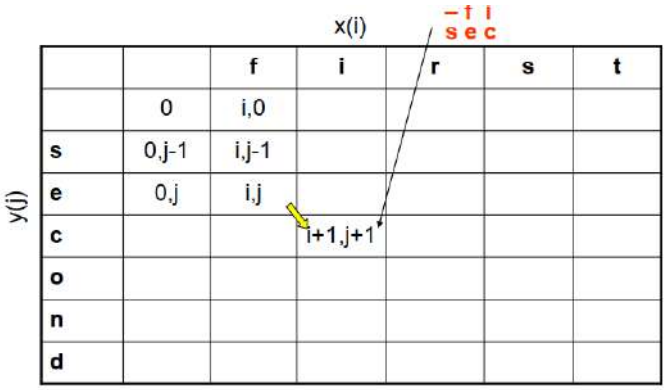
y(j) \ x(i)		f	i	r	s	t
	0	i,0				
s	0,j-1	i,j-1				
e	0,j	i,j				
c						
o						
n						
d						

Moving *horizontally* in the matrix is equivalent to introduce a gap into the sequence along the *left edge*

Moving *vertically* in the matrix is equivalent to introduce a gap into the sequence along the *top edge*



To move diagonally in the matrix is equivalent to align the corresponding characters in the two sequences

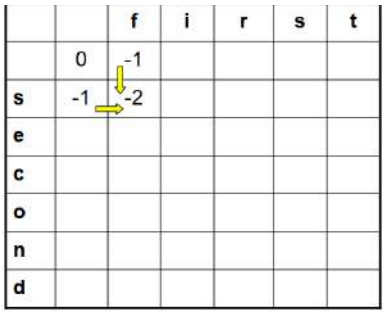


Example:

If:

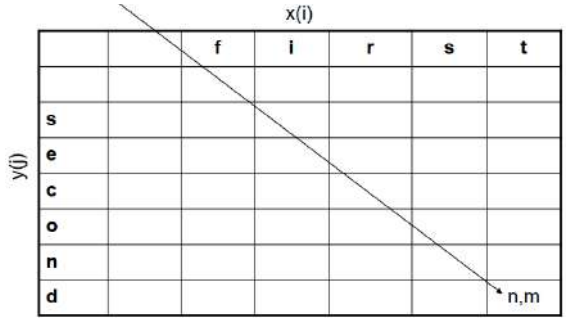
$$\sigma(f,s) = -3$$

$$\text{gap} \begin{cases} \sigma(f,-) = -1 \\ \sigma(-,s) = -1 \end{cases}$$

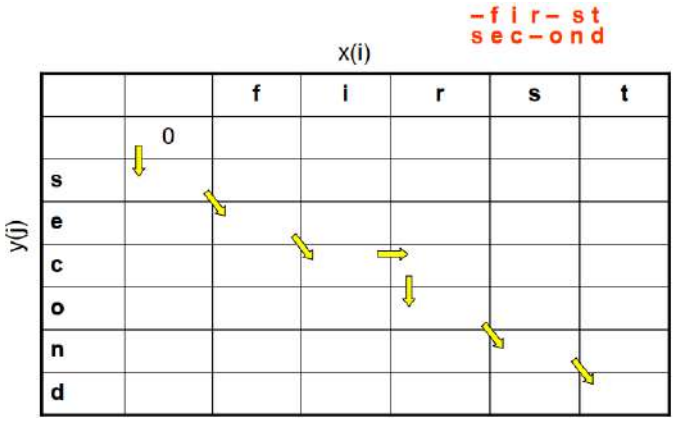


We have: $V(1,1) = \max(0-3, \quad -1-1, \quad -(1-1)) \; = \; -2$

But what will be the optimal alignment and its score? For sequences with *equal or similar length*, the (n,m) cell gives the score of the optimal alignment



Having taken into account at each step of the move done to complete the value of a cell, starting from the (n,m) cell through ***traceback*** *go backward in order to find the optimal alignment*, for example:



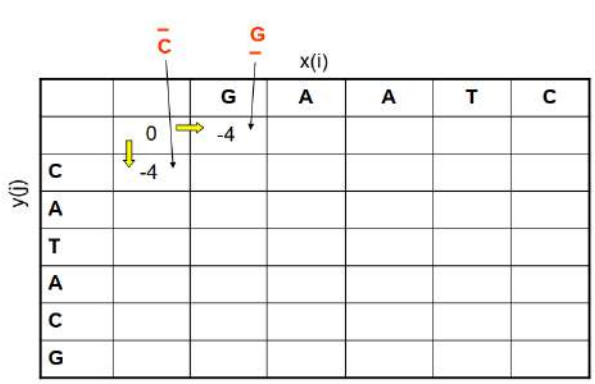
A trivial example of *nucleotide alignment* problem:

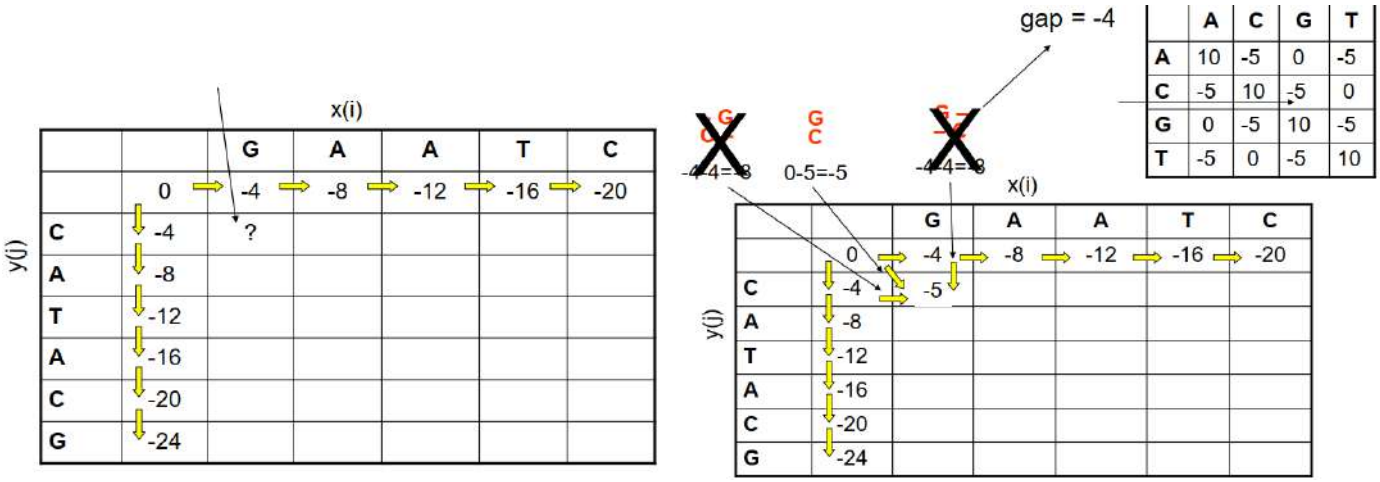
- Find the best pairwise alignment between the sequences GAATC and CATACG
  - Use for the *gap a linear penalty* of -4
  - Use the following *substitution matrix*:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

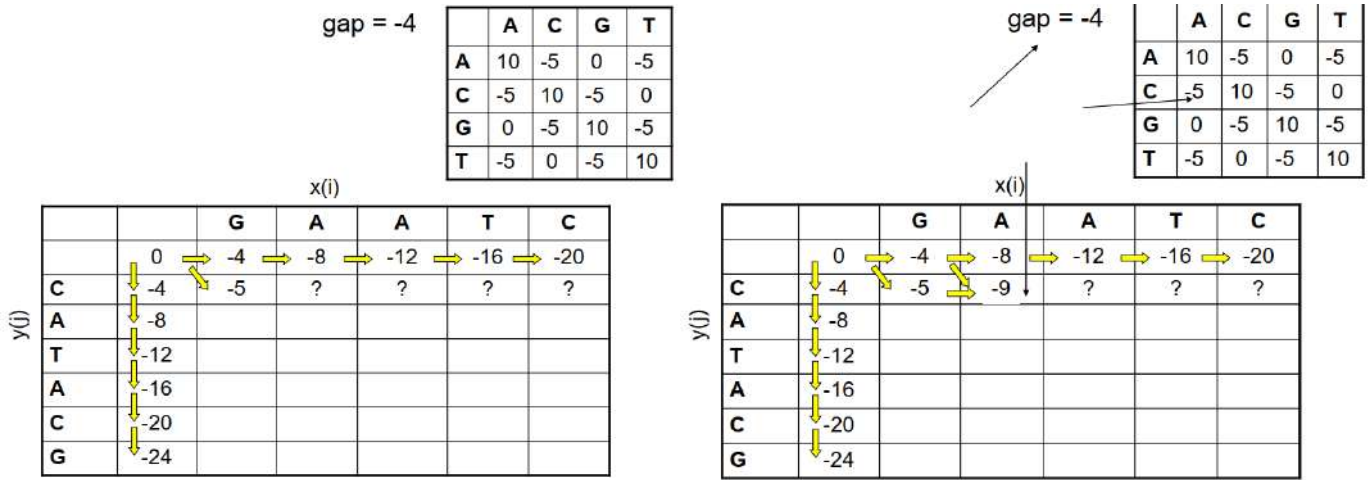
(A and G are Purine, T and C are Piramidine)

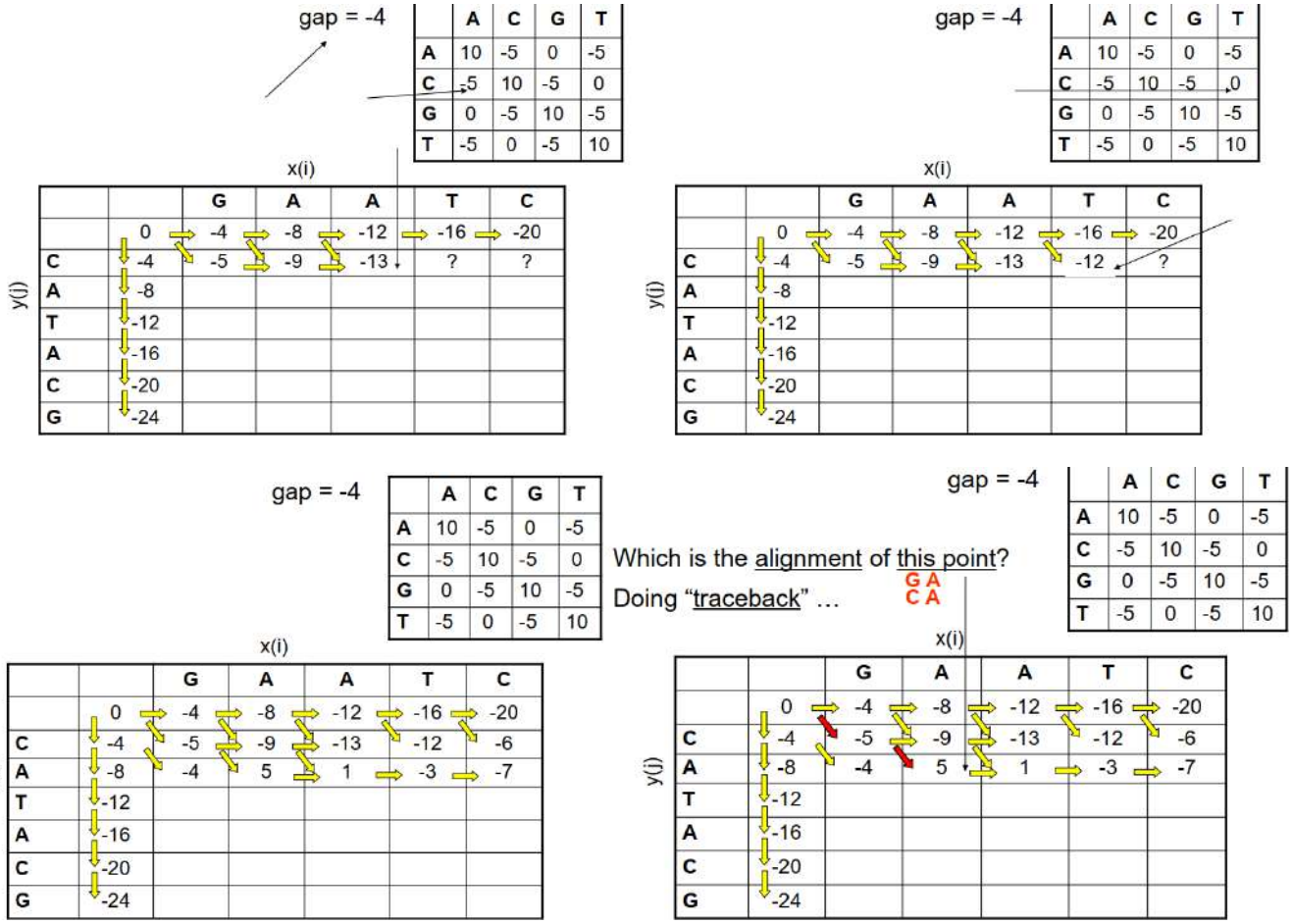
Initialization and *gap introduction* (-4):



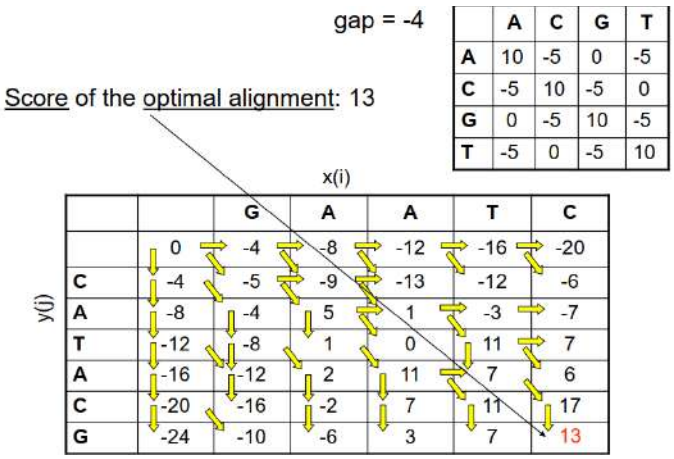


Warning: when selecting the max, beware of *negative numbers*!





After successive computations,



Starting from the last cell (n, m) and doing "**traceback**" we can find optimal alignments:

```
GA-ATC-   GAAT-C-   GAAT-C-   GAAT-C-
CATA-CG   CA-TACG   C-ATACG   -CATACG
```

All alignments with score = 13

Multiple solutions: when we find an alignment of 2 sequences, this may not be the only best alignment

With techniques of *dynamic programming*, the computational complexity of optimal alignment of two sequences of characters of length n and m is:

- $O(n + m)$ for *initialization*
- $O(n \cdot m)$ for *calculation* of other elements of the matrix
- $O(n + m)$ for *traceback*

So, if $n \cong m$: the total computational complexity is $O(n^2)$

III.B.8 Global alignment

Actually, used algorithms are based on *maximizing scores* (similarity) and not on penalties to minimize. Scores usually derived from substitution matrices and calculated based on statistics (like PAM and BLOSUM)

Global alignment: two sequences of comparable length are aligned to find the best score of similarity between the entire sequences, which aligns two sequences over their entire length

- Algorithm of *Needleman-Wunsch* (1970)
- Application example: comparison of genomic DNA and cDNA (coding)

Needleman-Wunsch algorithm:

- Matrix of scores is constructed as in previously described algorithm (based on grids)
- Rule for calculating the scores:

$$S(i,j) = \max \begin{Bmatrix} S(i-1,j-1) + s(ai,bj), \\ S(i-1,j) - Wx, \\ S(i,j-1) - Wy \end{Bmatrix}$$

  where $s(a_i, b_j)$: score assigned to the *match/mismatch*, Wx: *row gap* penalty, Wy: *column gap* penalty
- *Scoring matrix s*: must contain only positive values (the BLOSUM or PAM may be used if normalized)

Optimal global alignment calculated from the cell with the best score (which will be in the *last row* **or** *last column* of the score matrix) and *reconstructed backwards* (traceback)

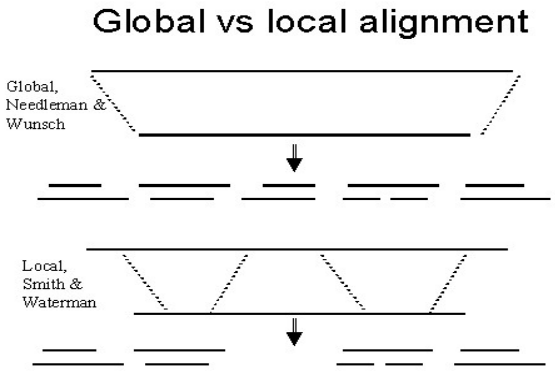
Needleman-Wunsch algorithm implementation: Needle - EMBOSS Pairwise Alignment (European Bioinformatics Institute): http://www.ebi.ac.uk/Tools/psa/emboss_needle/

III.B.9 Local alignment

Local alignment: two sequences are aligned to find the best score of similarity between *subsequences*; alignment on the full length of both sequences is not required

- *Smith-Waterman algorithm* (1981), Needleman-Wunsch algorithm variant that gives local alignment of two sequences
- Useful when comparing sequences that do **not** have high similarity over their entire length, but still contain regions with high similarity
- Application example: search for ***short motives*** shared between long sequences, or for common subunits in proteins

Local alignment algorithm provides the n alignments with maximum score of the subsequences of two sequences



Smith-Waterman algorithm: search for *local subsequences*

- Implies that the alignment *outer regions* should not influence, positively or negatively
- A negative score upstream the subsequence suggests *ignoring* the sequence upstream and to start a *new* alignment
- When the total score value is *negative*, the score is *set to zero* and the alignment is ended

- Rule for calculating the scores:

$$H(i,j) = \max \begin{Bmatrix} H(i-1,j-1) + s(ai,bj), \\ H(i-1,j) - Wx, \\ H(i,j-1) - Wy, \\ 0 \end{Bmatrix}$$

Scoring matrix s: it must also contain negative values (BLOSUM or PAM are generally used)

Initialization:

- $H(0,0) = 0$
- $H(i,0) = 0$
- $H(0,j) = 0$

Local alignment are identified by looking for scores *above certain* ***threshold*** and backwards reconstructed (traceback)

Traceback:

- Started from $H(i,j)$ with highest value above the threshold
- Ended when a $H(i,j) = 0$ is met

End of *optimal local alignment* of sequences not only in a cell of the last row or last column of the score matrix

A simple example of a local alignment problem: find the best local pairwise alignments with *score* ≥ 20 between the GAATC and CATACG sequences.

Use a linear *penalization* of -6 for a *gap*

Use the following substitution matrix:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

(A and G are Purine, T and C are Piramidine)

Score threshold ≥ 20 gap = -6

Initialization:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

x(i)

y(j)		G	A	A	T	C
	0	0	0	0	0	0
C	0					
A	0					
T	0					
A	0					
C	0					
G	0					

Score threshold ≥ 20 gap = -6

The cell with the highest value gives the optimal alignment score: 24

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

x(i)

y(j)		G	A	A	T	C
	0	0	0	0	0	0
C	0	0	0	0	0	10
A	0	0	10	10	4	4
T	0	0	4	5	20	14
A	0	0	10	14	14	15
C	0	0	4	8	14	24
G	0	10	4	4	8	18

After computation, we obtain the matrix above

Starting from the cell with the highest value and doing the "traceback", we find:

A T – C
A T A C

Another local alignment with over-threshold score (20) is:

A T
A T

Implementations of Smith-Waterman algorithm:

SSEARCH - Protein Similarity Search (European Bioinformatics Institute): http://www.ebi.ac.uk/Tools/sss/fasta/

Previously (now retired):

- MPsrch (Edinburgh University): http://www.ebi.ac.uk/MPsrch/ [not available anymore]
- Scanps2.3 (Geff Barton - European Bioinformatics Institute): http://www.ebi.ac.uk/scanps/ [not available anymore]

Popular algorithms have *quadratic complexity* (in the length of the sequences) in *time* and *space*. Variant of **Myers and Miller** allows to produce alignments of pairs of sequences in *quadratic time and linear space*

III.B.10 Global vs. Local

It is important to underline the behavioural *difference between local and global alignment*:

- pairwise *global* alignment highlights any **overall** similarity between two sequences
- pairwise *local* alignment highlights any **local** similarity between two sequences
    - Two sequences can be very *different globally* (in their entirety), but they can still have very *similar regions*
    - From such local similarity, it is often possible to formulate *interesting hypotheses* about the presence of certain motifs and therefore on the function of the analyzed molecules

It is also important to highlight *differences between algorithms*:

- ***Needleman-Wunsch*** algorithm:
    - *Global* alignments
    - Requires the *alignment score* of a couple of nucleotides or amino acids to be ≥ 0
    - Does *not* necessarily require penalized gap
    - Alignment score cannot decrease between two cells of an alignment path
    - Best alignment score is in a cell of the *last row or last column* of the alignment matrix
- ***Smith-Waterman*** algorithm:
    - *Local* alignments
    - Alignment score of nucleotide or amino acid pairs can be *positive or negative*
    - *Gap penalty* required to function effectively
    - Alignment score may increase, decrease or remain the same between two cells of an alignment path
    - Best alignment score can be found in *any cell* in the alignment matrix

III.B.11 Significance

It is important to *automatically evaluate* the **significance** of the alignments found. How much are the highlighted similarities significant and how much are they *by chance*?

For any pairwise alignment, the used measures are:

- ***Z***: measure of how the found *correspondence* is *different* from the random one (better high Z)
- ***p***: probability that the *alignment* found is *not better* than the random one (better low p)
- ***E***: number of sequences in a database of random sequences with equal length of the query sequence that have the score of the alignment with the query sequence greater than or equal to that of the found sequence

a) Score *Z*

Generate a *large number of permutations* of one of the two sequences, *align* each permutation with the other sequence and store the scores obtained

It is estimated

$$Z = \frac{original\ score\ -\ mean\ of\ scores\ obtained\ by\ permutations}{standard\ deviation\ of\ scores\ obtained\ by\ permutations}$$

$Z \geq 5$ score that suggests **significance** of the alignment found between the two sequences

b) Probability *p*

Probability of finding, by chance, a score equal to or greater than some value S is:

$$p = 1 - e^{-k \cdot m \cdot n \cdot \exp(-\lambda \cdot S)}$$

With :

m: length of query sequence

n: length of sequence found in the queried database

k and λ: parameters dependent on the substitution matrix used and the queried database

Guide generally used to read the probability:

$p < 10^{-100}$	exact correspondence
$10^{-100} \leq p < 10^{-50}$	almost equal sequence (e.g. alleles or with SNP)
$10^{-50} \leq p < 10^{-10}$	sequences closely related, homology is certain or almost
$10^{-5} \leq p < 10^{-1}$	sequences of species distantly related
$10^{-1} \leq p$	probably not significant correspondence

c) Value *E*

Expected number of sequences in the database of random sequences with equal length of the query sequence, that have the score of the alignment with the query sequence greater than or equal to that of the found sequence

Guide generally used to read E:

$0 \leq \mathrm{E} < 10^{-100}$	identical sequences (or almost)
$10^{-100} \leq E < 10^{-10}$	sequences usually homologous
$\mathrm{E} < 10^{-5}$	good alignment
$1 \leq E \leq 10$	often related sequences

Note that for $E < 10^{-3}$ we have that: $E \cong p$

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

III.B.12 Database research

Classic programs that *search for sequences* in databases are ***FASTA*** and ***BLAST***. The heuristic principle that these programs use is the *search for "words"* in databases

A word is a *short series of characters* in the sequences of amino acids or nucleic acids. Normally, these words are indicated with the term k-tuple (k = number of characters)

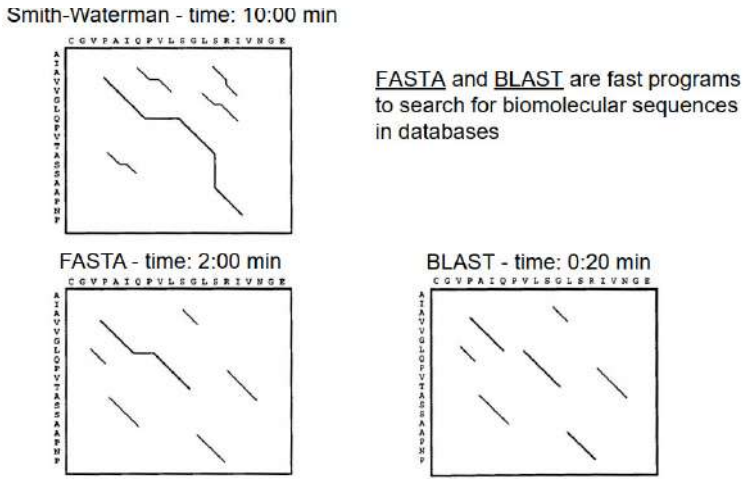
The foundational hypothesis of methods based on search of words is that ***related sequences share many words***

Sensitivity: is the ability to identify sequences related, although evolutionary distant. Increasing the sensitivity, increases the *number of matches observed*, but *decreases the speed of research*

Specificity: is the ability to *avoid false positives* (i.e., sequences not related, but with a high value of similarity).

Long words lead to increased specificity and reduced sensitivity: two evolutionary distant sequences do not share anymore long words, but shorter words

In searching within databases, a compromise between sensitivity and specificity must always be sought



III.B.13 FASTA

FASTA (Lipman and Pearson, 1985), i.e. FAST-All, http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

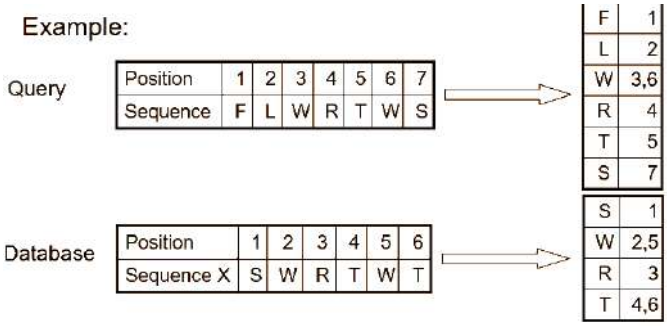
- In the classic version, it is an heuristic program that can search for global homology of sequences
- Two variants, LFASTA and PLFASTA, that can search for local homology of sequences

FASTA is **specific** but not quite sensitive

FASTA uses *4 phases* to perform the search

a) Phase 1a: *k*-tuple

Initially, create a **positional table** containing *all the positions for each amino acid* (or nucleotide) in the query sequence and in each sequence in the database



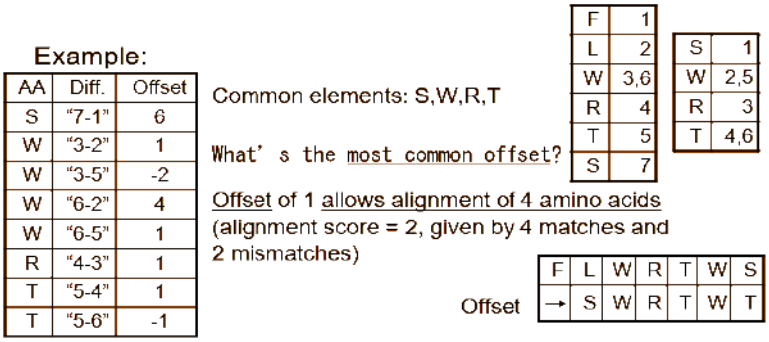
The *positional table* can be built considering the position of the amino acids (nucleotides) taken individually (*k*-tuple = 1) or in pair (*k*-tuple = 2)

In the case of nucleotide sequences, *k*-tuple is 4 or 6

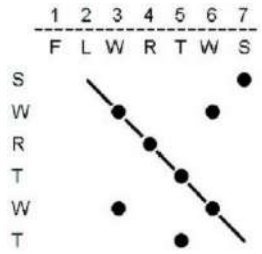
Number of operations is proportional to the length of the query sequence

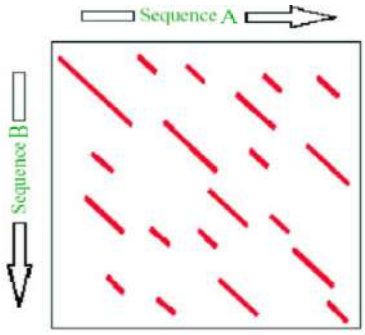
b) Phase 1b: offset calculation

Calculate the **difference (offset) of positional values** of each amino acid (nucleotide) between the query sequence and the sequences in the database



The correspondent dot matrix :

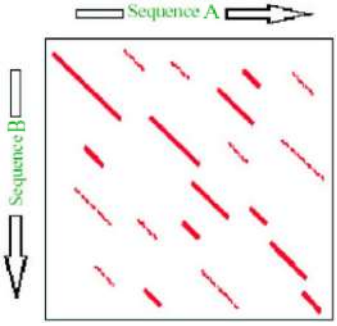




As we proceed, we make *growing the regions* (traits of diagonals with match), interrupting them and starting a new one if the alignment score becomes negative

The *10 best regions of similarity are selected* for the subsequent analysis, regardless if they belong to the same or different diagonal

c) Phase 2: evaluation of substitutions between nucleotides (amino acids)

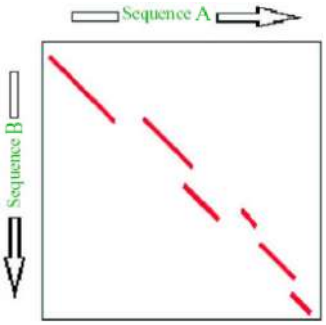


The best 10 regions selected in the phase 1 are **evaluated** through the *score matrices* (e.g. PAM250)

The subregions that contain the bases (amino acids) that *maximize the region score* are identified. These regions are called ***initial regions*** and their score is called initial score, or **INIT1**

The aim is finding the initial region with the best score, to be used to create a rank of the sequences in the database, in order to define which of them are the *most similar* to the query sequence

d) Phase 3: joining of the initial regions

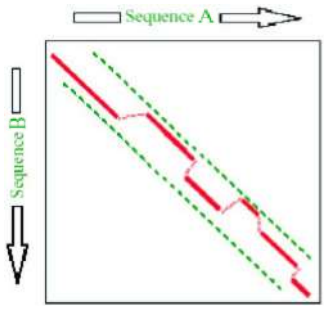


FASTA evaluates if it is possible to **join together** different regions of similarity. The constraints to create the join are:

- Excluding any *areas of overlap* between regions
- Score above a "*threshold*"
- Introduction of a scoring *penalty for each gap* introduced to join 2 regions

The regions are joined if the cost to be paid is less than what is gained through the reunion. The similarity score obtained by joining regions is denoted with **INITN**

e) Phase 4: optimization alignment



Sequences with *higher similarity* are aligned to the query sequence using the procedure based on a *modified Smith-Waterman algorithm* (then partial local alignment). This allows obtaining an **optimized score** (OPT)

Originally built only for a restricted range (20 amino acids / nucleotides). Newer versions do it throughout the whole matrix

Score evaluation: obtained the final scores (OPT), FASTA estimates the *statistical significance* of the results as follows:

- It generates a statistically *significant number* of *random combinations* of sequences with the *same length* and composition of amino acids (or bases) of query sequence
- For each of them, it runs a FASTA alignment against a subset of the database sequences
- It computes score **mean** (M) and **standard deviation** (SD), assuming the values are **normally distributed**
- It compares obtained OPT values with the mean value of the distribution of the scores of the random sequences
- $Z - score = \frac{OPT_{alignment} - Mean_{random}}{SD_{random}}$, measures how much the OPT value deviates from the mean of the scores of the random sequences

III.B.14 BLAST

BLAST (Basic Local Alignment Search Tool): http://blast.ncbi.nlm.nih.gov/Blast.cgi http://www.ebi.ac.uk/Tools/blast2/

- It searches for *best local alignment* between a query sequence and the sequences in a database.
- Developed and supported by NCBI (US National Center for Biotechnology Information) (1990)

Features:

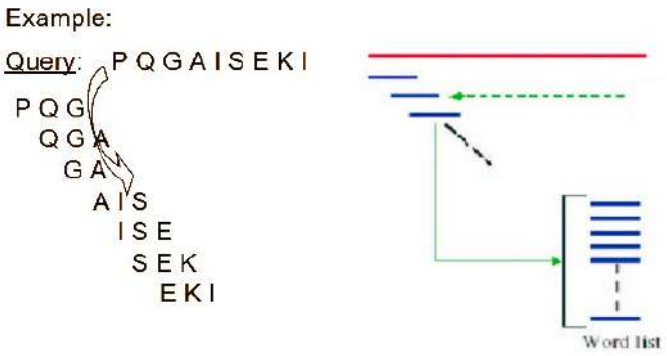
- *Local* alignments
- Alignments with *gaps*
- Heuristic
- *Rapid*

Features of the algorithm:

- While FASTA searches all possible words of the same length, BLAST limits the search to *the most significant words* using a ***"preventive" filter***
- To calculate the score, in the case of proteins it uses the BLOSUM62 matrix
- BLAST fixes the length of the word to:
  - 3 (previously it was 4) for proteins
  - 11 for nucleotides
- BLAST follows several "phases"

a) Phase 1: Generation of words

It **generates a list of words** of length **W** (3 for proteins, 11 for nucleotides) from the query sequence



For each word identified in step 1, a *list of all the associated words* is generated

- Using BLOSUM62 substitution scores (PAM250 for nucleotides), similarity of all possible words is evaluated

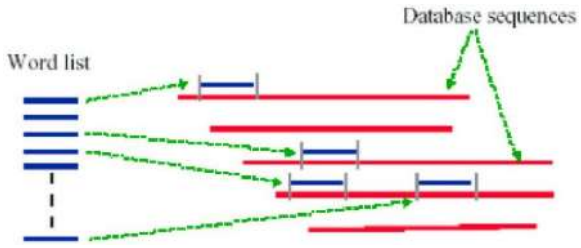
- Example for the word PQG
  - Score BLOSUM62 of:
    - P-P, Q-Q, G-G = 7 + 5 + 6 = 18
    - P-P, Q-Q, G-E = 7 + 5 + -2 = 10
    - P-P, Q-R, G-G = 7 + 1 + 6 = 14
    - …
- A score is assigned to each $20^3 = 8'000$ words that can be found in the database (similar calculations for nucleotides: $4^{11} = 4'194'304$)

A ***threshold T*** is used to limit the number of analogous words (to be subsequently used to search in the database).

- Threshold T is usually chosen to reduce to 50 (the 50 most analogous) the number of words to be used
- Example for the word PQG with a threshold 13
  - P Q G = 18
  - ~~P Q E = 10~~
  - P R G = 14
  - …
- The procedure is repeated *for all the words* with 3 amino acids (11 nucleotides)
- In total, the words to be used to search in the database will be (for proteins): $50 * (query\ length - 2)$

b) Phase 2: Find the words in the database

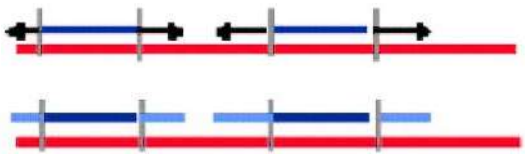
The search (exact) of the best analogous words in the sequences of the database is performed



c) Phase 3: Hit's extension

When searched *analogous words* are found in database's sequences, they identify regions of possible local alignment (without gap) between the query sequence and the sequences found in the database

The algorithm tries to *extend aligned regions*, without allowing gaps, and until extended alignment score does not decrease



In this way, we get alignments with higher score than the original one, called HSP (High-scoring Segmented Pairs)

Extension occurs considering, one at a time, the two bases (residues) immediately next to the current alignment. If score increases, or decreases within a threshold **X**, the two bases are included, otherwise extension stops

A HSP score is considered relevant (and selected) if exceeds a threshold value **S**. At the end, it is generated the *best alignment*, according to Smith-Waterman algorithm, of query sequence only with the sequences of the database selected (by the HSP)

W, T, X, S are parameters of BLAST algorithm. Variant (*two-hit method*): as 90% of the time is due to extension and alignment, try to extend alignment only when there are two independent hits on the same diagonal less far away than a threshold **A**

Parameter choice:

- Usually **T** is set *automatically* (e.g. **W** = 3, BLOSUM62, **T** = 13; or with variant: **T** = 11 and **A** = 40)
- **S** is set through E (expected value of sequences that, in a database of random sequences with the same length of the query sequence, would have score of alignment with the query sequence equal or higher than **S**)
- Increasing **S** decreases E; default $E = 10$

E-value: alternative interpretation

- E is equal to the *number of sequences that we would expect* to find if the database contained random sequences
- The factors that influence the value E are:
  - *Number of sequences* in the database
  - *Length* of query sequence
- It is more probable to find at random a match in a *big* database
- It is more probable to find at random a match (of local alignment) in a long sequence compared to a short one

p-value

- The *probability of getting at random a value* equal or higher than the obtained HSP is: $p = 1 - e^{-E}$
- p goes from 0 to 1
- When $E < 0.01$, p-value and E-value are very similar
- A value of p-value of 0.03 means that there is 3% chance of getting at random a score equal or higher than HSP (but we prefer to look at the E-value than the p-value)

Filters: BLAST has filters to skip regions with repetitions or low complexity (not applied to FASTA):

- SEG: filters low *protein* complexity regions
- DUST: filters low *DNA* complexity regions
- XNU: filters regions containing *protein tandem repetitions*

```
  1 MAAKIFCLIMXXXXXXXXXXXXIFPQCSQAPIASLLPPYLSPAMSSVCENPILLPYRIQQ 60
  1 MAAKIFCLIMLLGLSASAATASIFPQCSQAPIASLLPPYLSPAMSSVCENPILLPYRIQQ 60

 61 AIAAGIXXXXXXXXXXXXXXXXXXXXXXXXXXNIRXXXXXXXXXXXXXXYSQQQQFLPFN 120
 61 AIAAGILPLSPLFLQQSSALLQQLPLVHLLAQNIRAQQLQQLVLANLAAYSQQQQFLPFN 120

121 QXXXXXXXXXXXXXXXXXPFSQLAAAYPRQFLPFNQLAALNSHAYVXXXXXXPFSQLAAVS 180
121 QLAALNSAAYLQQQQLLPFSQLAAAYPRQFLPFNQLAALNSHAYVQQQQLLPFSQLAAVS 180

181 PAAFLTQQQLLPFYLHTAPNVGTXXXXXXXXXXXXXXXXTNPAAFYQQPIIGGALF 235
181 PAAFLTQQQLLPFYLHTAPNVGTLLQLQQLLPFDQLALTNPAAFYQQPIIGGALF 235
```

Masked regions (XXX) in an amino acid query sequence

BLAST	*FASTA*
"Standard" tool, used in practice	
More sensitive in **protein** research	More sensitive, in particular for **nucleotide** sequences
Local alignment	**Global** alignment (although BLAST for global alignments exist)
Fast (on-line)	**Slow** (e-mail answer)
First analysis	**Secondary** analysis (or replaced with a BLAST with a more precise matrix)

BLAST and FASTA are both **heuristic**

- They use a strategy that is expected to find *most of matches*, but sacrifices sensitivity to gain velocity
  - They could not find existing matches
  - They *do not grant* to find the best alignment between two sequences

Implementation examples:

- FASTA: http://www.ebi.ac.uk/Tools/fasta/ http://fasta.bioch.virginia.edu/
- BLAST: http://www.ebi.ac.uk/Tools/blast/ http://www.expasy.org/tools/blast
- You can download a *local version*: BLASTALL (http://blast.ncbi.nlm.nih.gov/): it allows to align amino acid (or nucleotide) sequences in a FASTA format with a database defined by the user, giving results also in tables (tab separated values)
- It is also available a collection of *pre-computed* BLAST results for each protein sequence in the Entrez Protein database: http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?mode=query

Other implementations: to *speed up algorithm execution* there are implementation:

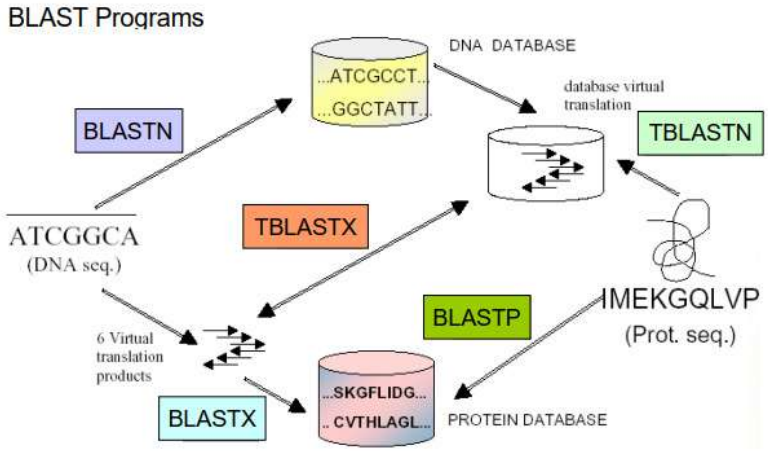
- *Parallelized*
- On dedicated *hardware*
- On *FPGA* - Field-Programmable Gate Array (http://www.timelogic.com/catalog/752/biocomputing-platforms)

Comparisons in databases

- Which type of sequence do we need to compare: DNA, protein, or DNA as a protein?
  - If the query sequence is a sequence of *amino acids* or *nucleic acids* that code for protein, it is better that we do the **research at the protein level.** So, you can point out more distant homologies than with nucleic acid sequences
- Search at *protein-level* helps to identifying **genes in evolutionary relationships**, whereas search in *DNA* helps identifying identical regions

BLAST programs /!

- **BLASTP**: searches a *protein* sequence in a database of proteins
- **BLASTN**: searches a *nucleotide* sequence in a database of *nucleotide sequences*
- **BLASTX**: searches a *nucleotide* sequence translated in ***all six possible reading frames*** in a database of proteins (because the translation is based on codons, triplets of nucleotides [=3 reading frames], that are potentially read one way or the other: $3 * 2 = 6$)
- **TBLASTN**: searches a *protein* sequence in a database of nucleotide sequences that are automatically translated into *all six possible reading frames*
- **TBLASTX**: searches *translations in all six possible reading frames* of a sequence of *nucleotides* in a database of nucleotide sequences *dynamically translated*



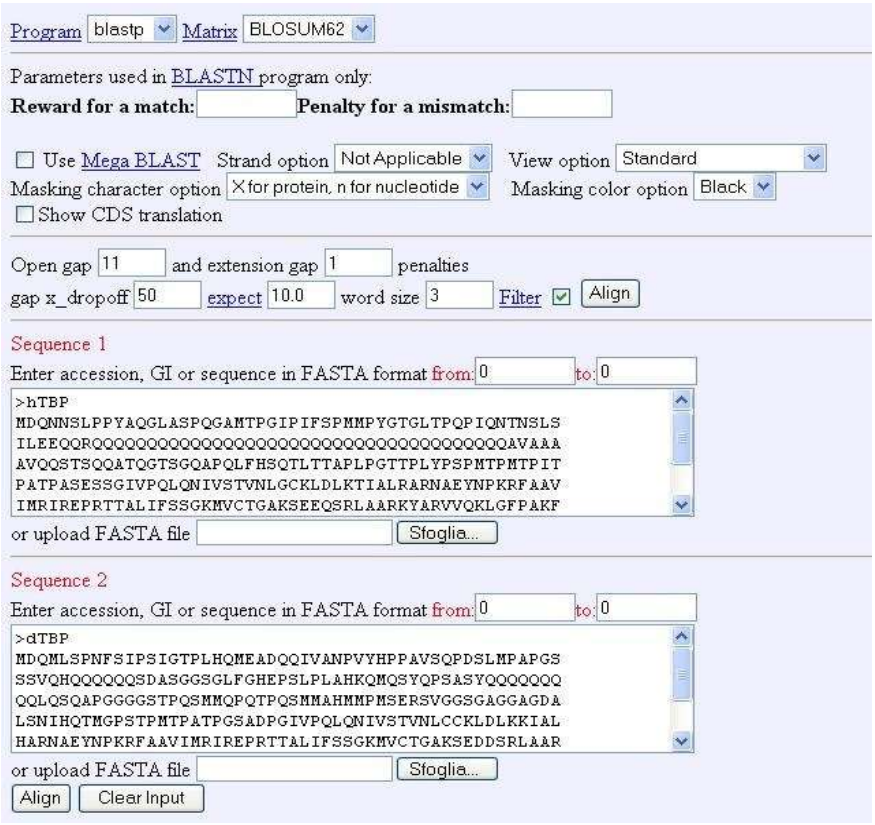
BLASTX vs. TBLASTX: the protein database contains only proteins that have been observed so far, whereas if I compare all possible translation I will find all possible matches, but some of these matches may not be possible in nature, and be thus irrelevant. BLASTX generates false negatives, while TBLASTX generates false positives.

Other BLAST tools:

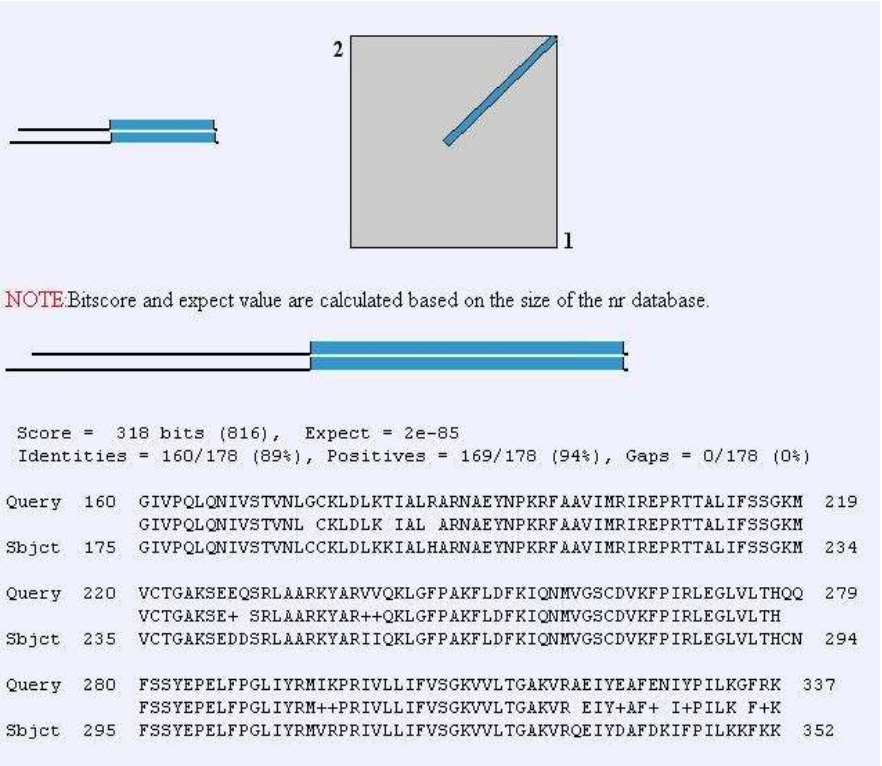
- *MegaBLAST*: it is an optimized program to align nucleotide sequences that differ slightly and therefore they could originate from sequencing errors. It may become 10 times faster than other programs depending on word size and it manages efficiently big sequences
- *PSI-BLAST* (Position Specific Iterated BLAST): designed for analysis of protein sequences, it *increases the sensitivity* of the algorithm using iterative procedure
- *BEAUTY* (BLAST-Enhanced Alignment UTilitY): interface which adds information to the output of BLAST with text and graphics. http://searchlauncher.bcm.tmc.edu/seq-search/protein-search.html

Blast2Seq:

- http://blast.ncbi.nlm.nih.gov/bl2seq/wblast2.cgi
- http://blast.ncbi.nlm.nih. gov/docs/align_seqs.pdf



Example with input the sequences of protein (transcription factor) TBP (TATA binding protein) of human and Drosophila



In the Blast2Seq output the aligned segments are highlighted

In the example were aligned the C-terminal parts of the two sequences, with identity of 89% (the famous "saddle domain" by which TBP interacts with DNA, which is highly conserved with respect to the remaining of the sequence)

III.B.15 Motif search

Motifs are regular *combinations* of *secondary protein structures* associated with particular functions. The presence of the same motif in different proteins may indicate a *similar function*

The **search for protein motifs** within genomic sequences can:

- Study the *diffusion* of specific motifs in different genomes
- Identify *new genes structurally similar to known genes*

Distribution of motifs in different genomes

- ***S.M.A.R.T.***: Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de/)

- In S.M.A.R.T. the most important *protein motifs* are *noted and described* together with their localization and distribution in *different genomes*
- The study of the distribution of motifs in a genome and in different genomes allows us to *understand the evolution of the motif and its functional importance* in cellular system (useful for genomes that are not fully characterised)

Identification of new genes structurally similar to known genes

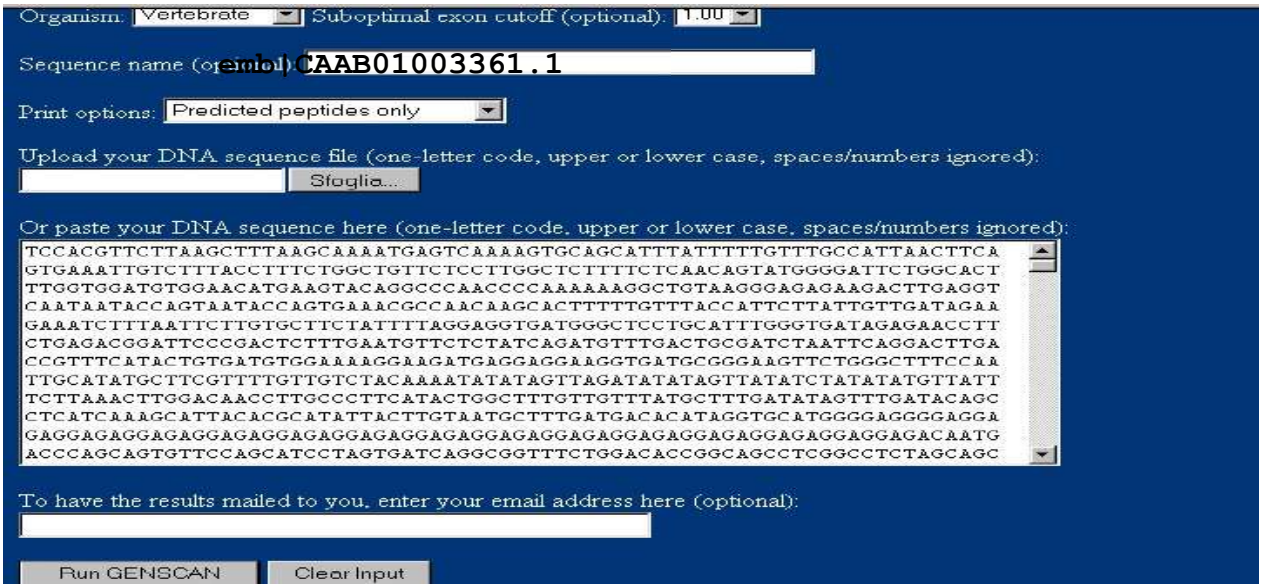
- Protein motifs are the ideal probes for *genome screening* in search for *unknown genes*
- The **motif-probe** can be used for:
  - Search *new members* of a gene family
  - Search for the presence of a *known gene* within an organism in which the gene has *not yet been characterized* (ortholog sequence search)

Example:

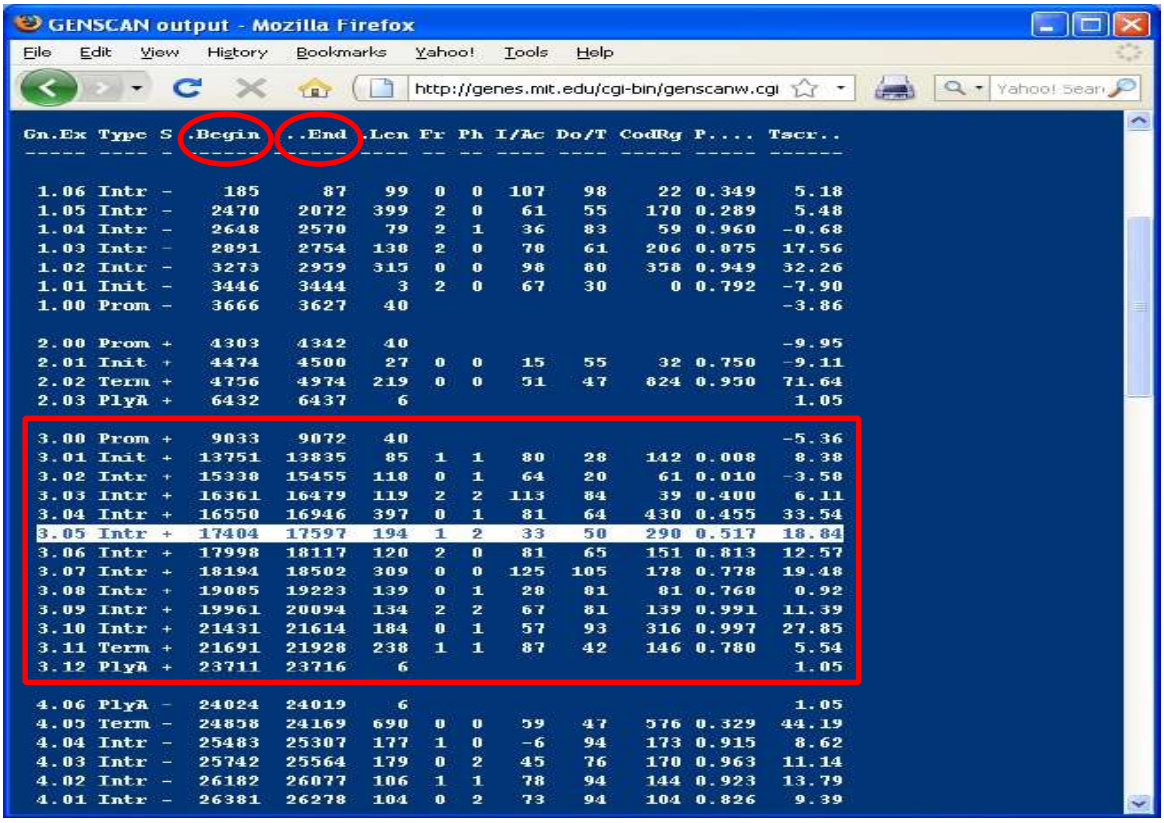
- Search *zinc finger* motifs (a motif that binds DNA, a very conserved one) in Fugu Rubripes genome (puffer fish):
  - Use as *probe* the zinc finger motif of an organism similar to Fugu
  - "Blast" this probe *against the Fugu genomic database* (NCBI BLAST – TBLASTN)
  - Use *GENSCAN* to predict position of genes and their *intron-exon* in genomic sequence given by **TBLASTN**. If it exists, the result to consider is the one that covers the whole region with the motif identified by TBLASTN
  - Validate the result of GENSCAN with **BLASTP** on Fugu Rubripes or other similar organisms

```
>emb|CAAB01003361.1| Fugu rubripes whole genome shotgun assembly SCAFFOLD_3361, whole genome
            shotgun sequence
          Length = 26503

 Score = 66.6 bits (161), Expect = 9e-12
 Identities = 26/27 (96%), Positives = 27/27 (100%)
 Frame = +2

Query: 1     CAVCNDYASGYHYGVWSCEGCKAFFKR 27
              CAVC+DYASGYHYGVWSCEGCKAFFKR
Sbjct: 17507 CAVCHDYASGYHYGVWSCEGCKAFFKR 17587


>emb|CAAB01001203.1| Fugu rubripes whole genome shotgun assembly SCAFFOLD_1203, whole genome
            shotgun sequence
          Length = 73025

 Score = 66.6 bits (161), Expect = 9e-12
 Identities = 26/27 (96%), Positives = 27/27 (100%)
 Frame = -3

Query: 1     CAVCNDYASGYHYGVWSCEGCKAFFKR 27
              CAVC+DYASGYHYGVWSCEGCKAFFKR
Sbjct: 71601 CAVCHDYASGYHYGVWSCEGCKAFFKR 71521
```

TBLASTN result of zinc finger motif against genomic database of Fugu Rubripes



Submit the genome sequence identified by TBLASTN to GENSCAN (http://genes.mit.edu/GENSCAN.html)



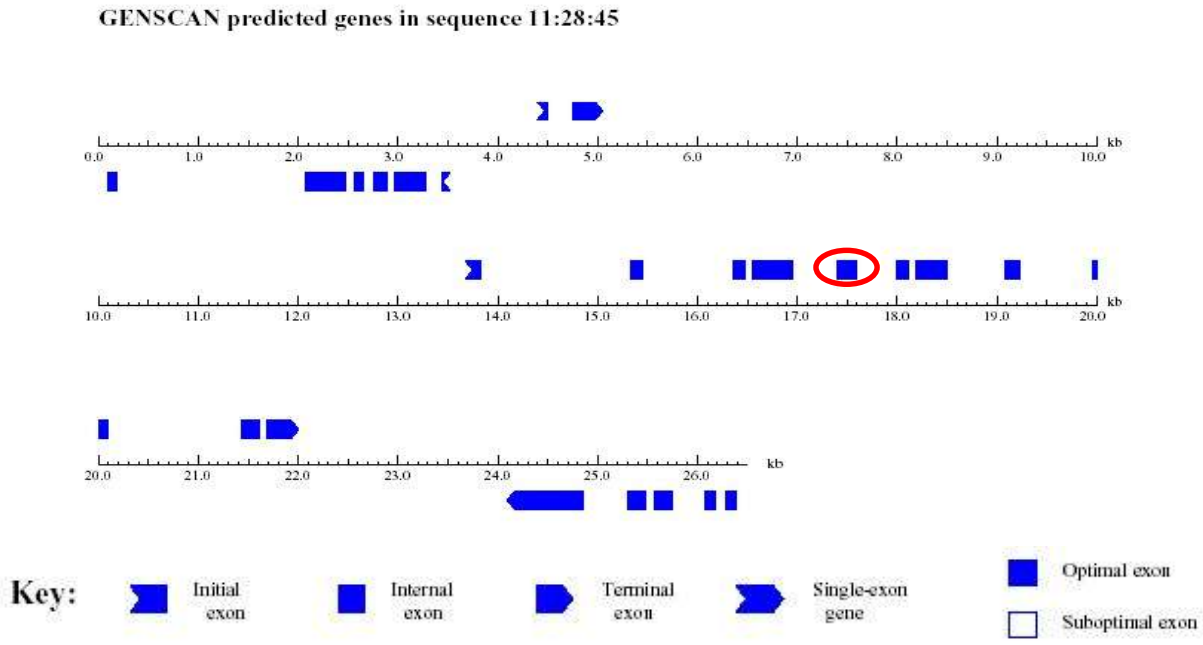
Gn.Ex	Type	S	.Begin	..End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.06	Intr	-	185	87	99	0	0	107	98	22	0.349	5.18
1.05	Intr	-	2470	2072	399	2	0	61	55	170	0.289	5.48
1.04	Intr	-	2648	2570	79	2	1	36	83	59	0.960	-0.68
1.03	Intr	-	2891	2754	138	2	0	78	61	206	0.875	17.56
1.02	Intr	-	3273	2959	315	0	0	98	80	358	0.949	32.26
1.01	Init	-	3446	3444	3	2	0	67	30	0	0.792	-7.90
1.00	Prom	-	3666	3627	40							-3.86
2.00	Prom	+	4303	4342	40							-9.95
2.01	Init	+	4474	4500	27	0	0	15	55	32	0.750	-9.11
2.02	Term	+	4756	4974	219	0	0	51	47	824	0.950	71.64
2.03	PlyA	+	6432	6437	6							1.05
3.00	Prom	+	9033	9072	40							-5.36
3.01	Init	+	13751	13835	85	1	1	80	28	142	0.008	8.38
3.02	Intr	+	15338	15455	118	0	1	64	20	61	0.010	-3.58
3.03	Intr	+	16361	16479	119	2	2	113	84	39	0.400	6.11
3.04	Intr	+	16550	16946	397	0	1	81	64	430	0.455	33.54
3.05	Intr	+	17404	17597	194	1	2	33	50	290	0.517	18.84
3.06	Intr	+	17998	18117	120	2	0	81	65	151	0.813	12.57
3.07	Intr	+	18194	18502	309	0	0	125	105	178	0.778	19.48
3.08	Intr	+	19085	19223	139	0	1	28	81	81	0.760	0.92
3.09	Intr	+	19961	20094	134	2	2	67	81	139	0.991	11.39
3.10	Intr	+	21431	21614	184	0	1	57	93	316	0.997	27.85
3.11	Term	+	21691	21928	238	1	1	87	42	146	0.780	5.54
3.12	PlyA	+	23711	23716	6							1.05
4.06	PlyA	-	24024	24019	6							1.05
4.05	Term	-	24858	24169	690	0	0	59	47	576	0.329	44.19
4.04	Intr	-	25483	25307	177	1	0	-6	94	173	0.915	8.62
4.03	Intr	-	25742	25564	179	0	2	45	76	170	0.963	11.14
4.02	Intr	-	26182	26077	106	1	1	78	94	144	0.923	13.79
4.01	Intr	-	26381	26278	104	0	2	73	94	104	0.826	9.39

Select the gene identified in the region where TBLASTN found the zinc-finger motif

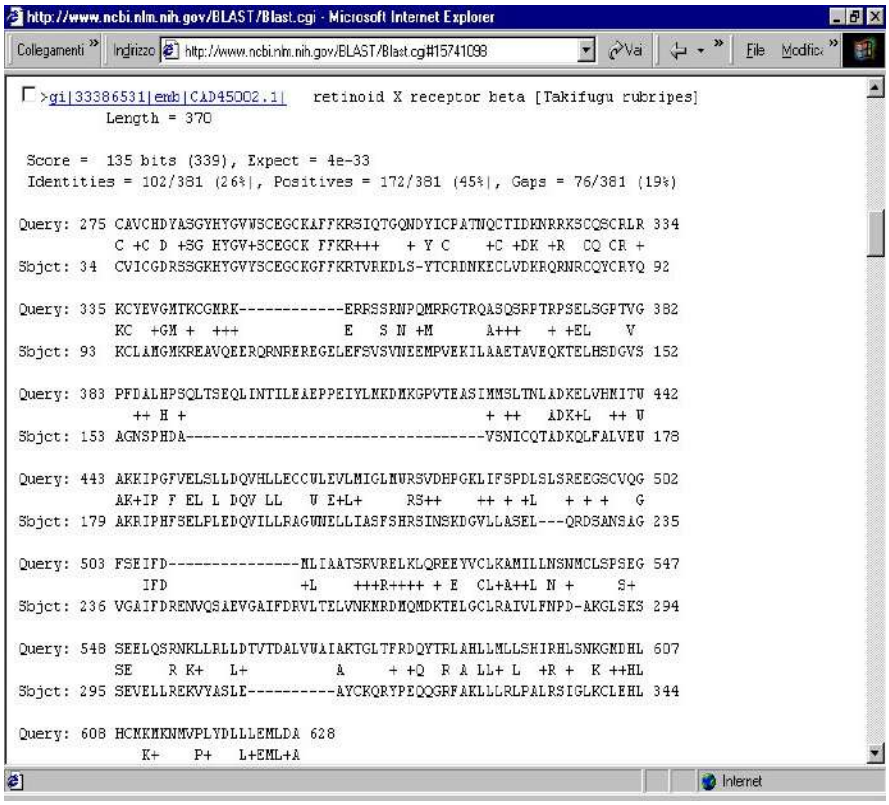
Type:

- Init: Initial exon
- Intr: Internal exon
- Term: Terminal exon
- Prom: Promoter
- PlyA: poly-A signal

There is an internal exon of the candidate that perfectly matches at coordinate 17404-17597, which contains the 17507-17587 previously identified by TBLASTN. So this exon probably encodes the protein sequence



The protein predicted according to the computed codifying sequence is submitted to a BLAST with *protein databases* (BLASTP) of the analyzed organism (Fugu) and of analogous organisms to determine if it is an *unknown* gene sequence. In this case, seek *orthologous* sequences in other organisms



Protein identified by BLASTP in Fugu protein databaseshows low homology with protein calculated by GENSCAN

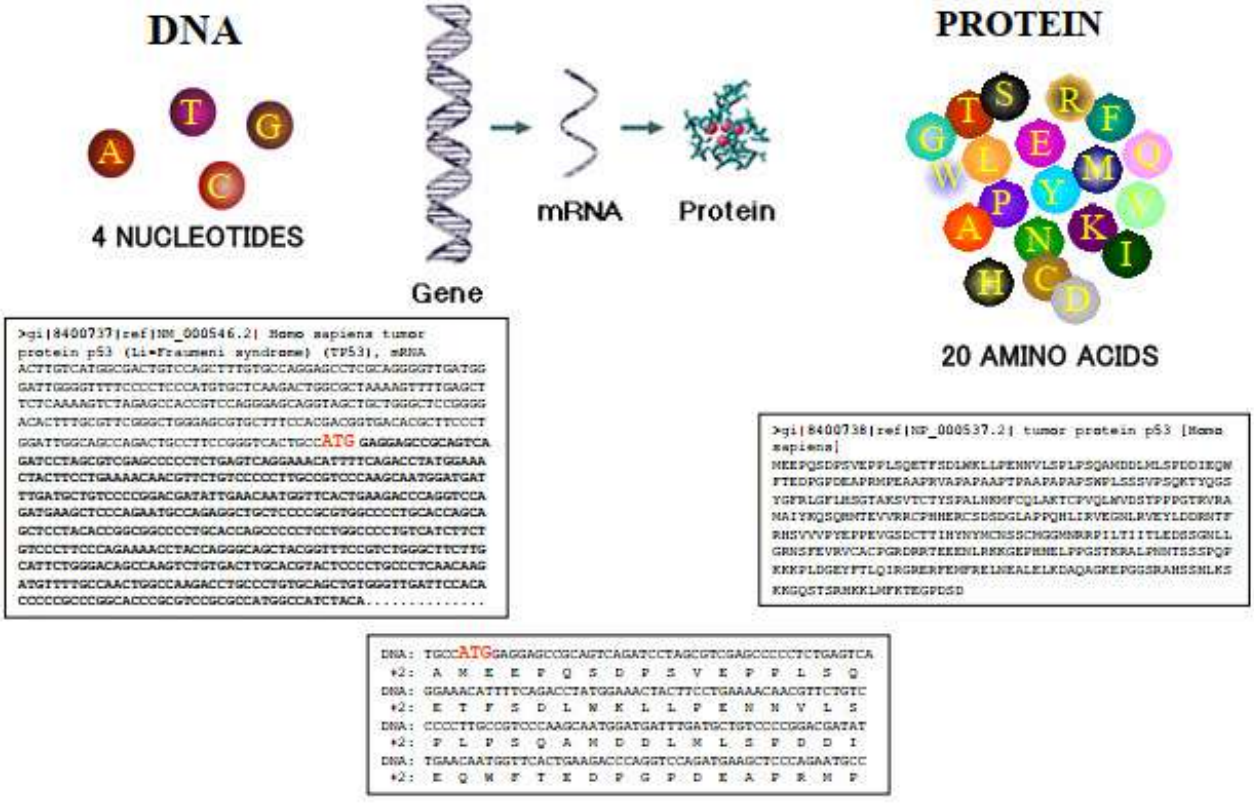
Is it a not yet identified gene (with its protein still unknown)?

```
>gi|16151550|emb|CAC93849.1|  estrogen receptor beta2 protein [Danio rerio]
          Length = 553

 Score =  692 bits (1785), Expect = 0.0
 Identities = 357/540 (66%), Positives = 413/540 (75%), Gaps = 15/540 (2%)

Query: 116 VASTPHKNQPLLQLQKVDSSRLGARVVSPILGASLET----SQPICIPSPYTDLNHDFSG 171
           ++  P  + PLLQLQ+VDS R+G  ++SPI  +S  +    + PICIPSPYTDL HDFS
Sbjct: 1   MSEYPEGDSPLLQLQEVDSGRVGGHILSPIFNSSSPSLPVENHPICIPSPYTDLGHDFST 60

Query: 172 IPFYGPTIFGYASPAISDRASIHRSMSPSLFWXXXXXXXXXXXXXXSQPRPQHGQPIQSP 231
           +PFY P + GY++  +SD +S+ +S+SP+LFW               Q R Q
Sbjct: 61  LPFYSPALLGYSTSPLSDCSSVRQSLSPTLFWPPHSHVSSLTLQ--QQSRLQQNHATSGT 118

Query: 232 WAELSPLDRDKDSKLSVGKSTRRRSQESEEAVVSSGGKADLHYCAVCHDYASGYHYGVWS 291
           W E +P D  ++      K   +R  ++EE  VS  GKAD+HYCAVC DYASGYHYGVWS
Sbjct: 119 WTEHTPHDHVEEEN---SKPLVKRVADTEETSVSLRGKADMHYCAVCSDYASGYHYGVWS 175

Query: 292 CEGCKAFFKRSIQTGQNDYICPATNQCTIDKNRRKSCQSCRLRKCYEVGMTKCGMRKERR 351
           CEGCKAFFKRSIQ G NDYICPATNQCTIDKNRRKSCQ+CRLRKCYEVGM KCG+R++R
Sbjct: 176 CEGCKAFFKRSIQ-GHNDYICPATNQCTIDKNRRKSCQACRLRKCYEVGMMKCGLRRDRS 234

Query: 352 S--SRNPQMRRGTRQASQSRPTRPSEL---SGPTVGPFDALHPSQLTSEQLINTILEAEP 406
           S   R  Q +R  R + + R T P      S P     + +    L+ E+LI+ I+EAEP
Sbjct: 235 SYQQRGAQQKRLVRFSGRMRMTGPRSQEIKSIPRPLSGNEVVRISLSPEELISRIMEAEP 294

Query: 407 PEIYLMKDMKGPVTEASIMMSLTNLADKELVHMITWAKKIPGFVELSLLDQVHLLECCWL 466
           PEIYLMKDMK P TEA++MMSLTNLADKELVHMI+WAKKIPGFVELSL DQVHLLECCWL
Sbjct: 295 PEIYLMKDMKKPFTEANVMMSLTNLADKELVHMISWAKKIPGFVELSLFDQVHLLECCWL 354

Query: 467 EVLMIGLMWRSVDHPGKLIFSPDLSLSREEGSCVQGFSEIFDMLIAATSRVRELKLQREE 526
           EVLM+GLMWRSV+HPGKLIFSPDLSLSR+E SCVQG  EIFDML+AATSR RELKLQREE
Sbjct: 355 EVLMLGLMWRSVNHPGKLIFSPDLSLSRDESSCVQGLVEIFDMLLAATSRFRELKLQREE 414

Query: 527 YVCLKAMILLNSNMCLSPSEGSEELQSRNKLLRLLDTVTDALVWAIAKTGLTFRDQYTRL 586
           YVCLKAMILLNSNMCL  SEG E+LQSR+KLL LLD+VTDALVWAI+KTGL+F+ + TRL
Sbjct: 415 YVCLKAMILLNSNMCLGSSEGGEDLQSRSKLLCLLDSVTDALVWAISKTGLSFQQRSTRL 474

Query: 587 AHLLMLLSHIRHLSNKGMDHLHCMKMKNMVPLYDLLLEMLDAHIMHSSRLSHRPPQQDLA 646
           AHLLMLLSHIRH+SNKGMDHLHCMKMK MVPLYDLLLEMLDAHIMHSSRLSH  P+   A
Sbjct: 475 AHLLMLLSHIRHVSNKGMDHLHCMKMKKMVPLYDLLLEMLDAHIMHSSRLSHSGPRAPAA 534
```

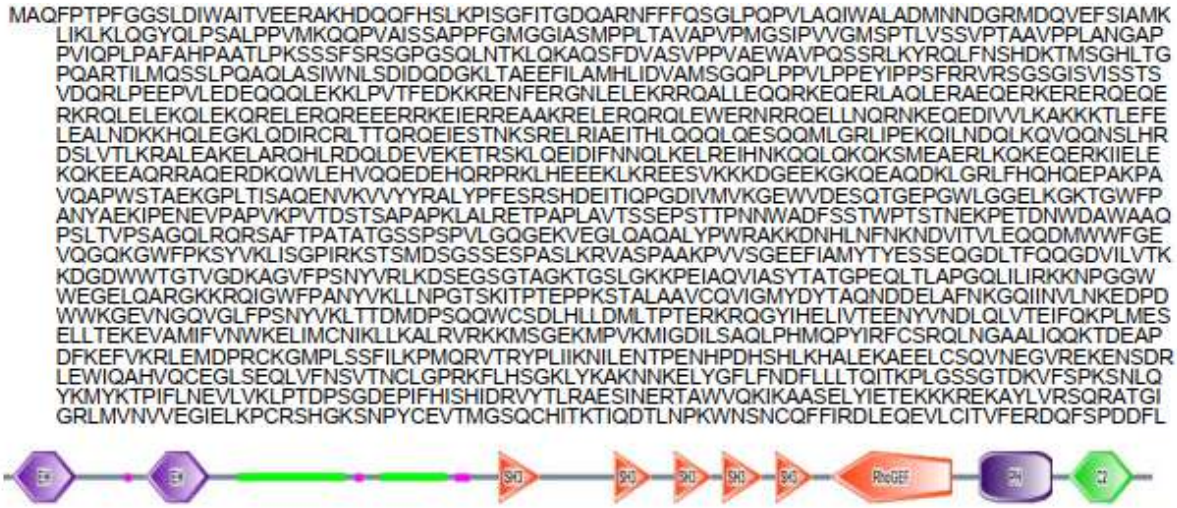
Protein identified by BLASTP in protein database of Danio rerio (another fish) shows high homology with the protein calculated by GENSCAN

A *new coding gene* in Fugu, homologous to a Danio rerio gene, has been **identified**!

III.B.16 Quick sum up



A simple sequence of letters *does not say much*...



...using, however, one of the many *tools* available on the Internet for the analysis of protein sequences, it is possible to know if the protein contains *conserved domains*, which are present in *other proteins*, whose function is well *know*

Bibliography:

- Ewens WJ, *Grant GR. Statistical Methods in Bioinformatics*, Springer. 2001.
- Principles and Methods of Sequence Analysis: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=sef& part=A166

III.C Multiple alignment of protein sequences [add.]

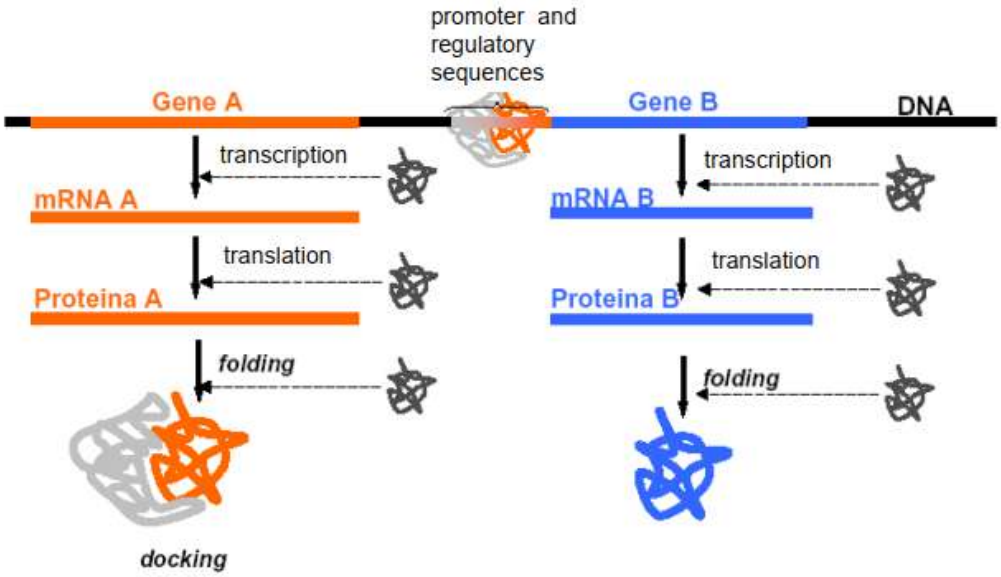
See slides

IV. Technologies for measurement and analysis of gene expression

IV.A Measurement of genetic expression (19 Oct.)

IV.A.1 Introduction

From genes to proteins: the genetic information is encoded in the DNA



Cell *genes* code for a "*pool*" of biological information

Genetic expression: conversion of coded information in a gene; for coding genes, first in messenger RNA and then in protein

Not every gene is always necessary for the cell life

- Only *constitutive genes* are <u>always</u> expressed
- Other genes are expressed only when they are <u>necessary</u>

Gene expression is *regulated* by the cell necessity: *environment conditions* and functions necessary to be performed (e.g. genes for lactose synthesis)

In multi-cellular organisms:

- The ***environment*** of a cell is the *organism itself*
- Starting from the same cell, the *"differential gene regulation"* mechanism causes the creation of *different specialized cells* (each one with the same DNA)

The *genetic* ***information*** is the same in every ***somatic*** *cell* of an organism (unless external factors, i.e. mutagens, change it). It specifies the nature of all proteins in the organism

The *genetic* ***expression***, and so the protein synthesis, is different depending on the ***cell type*** and the answer from the ***environment*** (the state of the cell)

The *transcriptome* is the complete *set of gene transcripts* and of their levels of expression, in a particular type of cells or tissue, in well-defined conditions

Only about *20% of the transcriptome is expressed by a cell*

N.B. The *transcript levels* do <u>not</u> necessarily translate in *protein synthesis or activity* (some transcripts are not translated; some translated proteins "do not function"). However, the quantification of the transcriptome in certain types of cell is a good approximation evaluation of the functional activity in a cell

Knowing *genome* and genes is <u>not</u> *sufficient* to understand how a gene, a cell, an organism works

To understand *biological organisms* in their entirety (and complexity) it is necessary to study:

- gene *expression* and *regulation*
- synthesized protein *functionality*
- quantitative *occurrences* of metabolites
- effects of *gene defects* on organism phenotype

Systems biology: study of *interactions between components* of a biological system and how such interactions induce *functions* and *behaviour* of the system

For *functional analysis of genomes*, modern methods exist:

- Transcriptomics
- Proteomics
- Metabolomics

Usually these methods use ***high-throughput procedures*** requiring relevant activities of data managing and analysis. The objective is to *identify* components of the system (i.e. transcripts, proteins, metabolites) and their interactions and functions

These approaches of **genotyping** (determine the genotype and its components of an individual/organism) must be correlated with, and completed by, *phenotypic analysis* of model organisms and *cells in vitro*

IV.A.2 Gene expression analysis techniques

After *sequencing* (knowledge of the sequence) and structural *annotation* (knowledge of components: genes, regulatory elements, ...) of genome, **transcriptome analysis** is a very important field of the functional genomic science

How to *measure* the gene expression?

Methods to measure the expression level of **a single gene at a time**: *RT-PCR* (Reverse Transcriptase Polymerase Chain Reaction)

Main analysis techniques of the **whole transcriptome** are:

- *cDNA microarray*
- *Oligonucleotide microarray*
- *SAGE* (Serial Analysis of Gene Expression) [Additional mat.]

SAGE is the only technique that can provide an exact quantification of the transcript produced by a gene at a certain time, while the other techniques (especially the microarray ones) are relative quantification between multiple genes in a cell.

1980: RNA analysis of *one or few* genes at a time:

- Northern blotting
- quantitative PCR (Q-RT-PCR or real-time PCR)

1995 - ...: RNA analysis of the *whole genome*

- Molecular biology techniques
- Micro / Nano technologies
- Computer science: high density (potentially measures of cell whole genome), big data

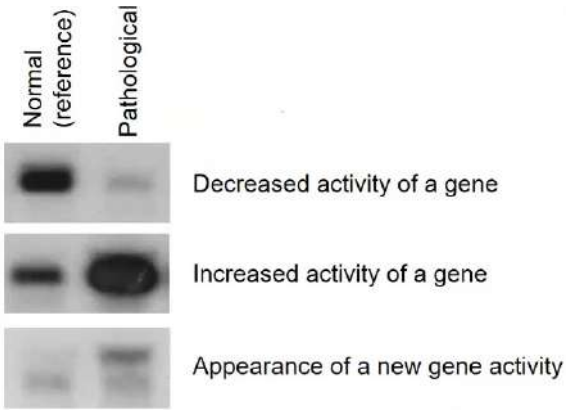
Two main technologies of *DNA microarrays*:

- cDNA spotted arrays (Schena et al., 1995)
- Oligonucleotide arrays (Lockhart et al., 1996)

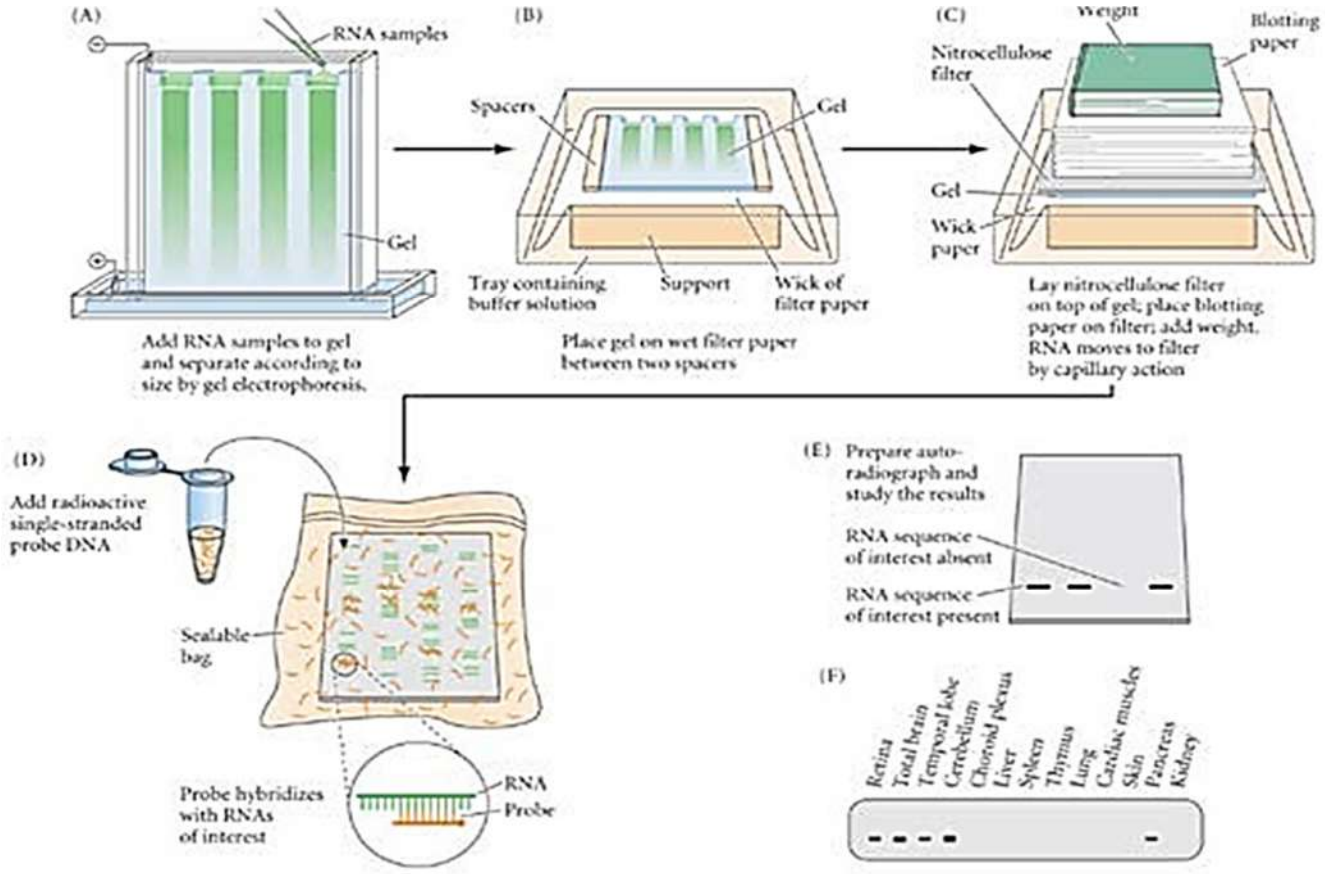
IV.A.3 Northern blot – Single transcript analysis

Northern blot: *laboratory technique* to study genetic expression, by finding the RNA (or isolated mRNA) in a sample

- *Gene expression:* quantifying the level of abundance of a transcript in a single sample
- *Gene regulation*: behaviour of the transcript in comparison test-control



Example of a northern blot



Northern blot: Transcript analysis (mRNA) (video: http://www.youtube.com/watch?v=KfHZFyADnNg)

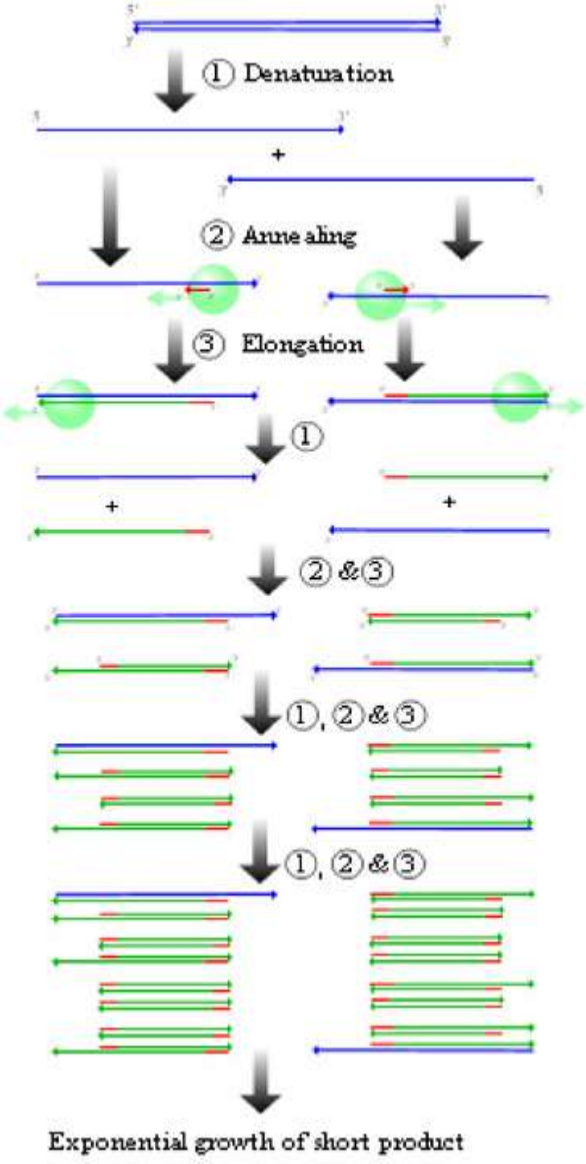
Southern blot: DNA analysis (video: http://www.youtube.com/watch?v=ftkdAbV_5gE)

Western blot: Protein analysis

IV.A.4 RT-PCR – Analysis of a single transcript

The ***polymerase chain reaction*** (PCR) is a *laboratory technique* exploiting *DNA replication* to **amplify** a single or a few couples of a specific *sequence of DNA*, up to ~10 kb long, but also up to 40 kb long, by synthesizing billions of couples (http://www.phgfoundation.org/tutorials/dna/4.html)

PCR is based on ***thermal cycles*** *of heating and cooling* of a solution where the replication reaction of the DNA occurs (Good video: http://www.youtube.com/watch?v=_YgXcJ4nkQ)



The ***reverse transcription polymerase chain reaction*** (**RT-PCR**) is a variation of the PCR, in which an RNA helix firstly is *reverse-transcribed* in its *complementary DNA* (**cDNA**), by using the enzyme *reverse transcriptase*, and the resulting cDNA is amplified by using *traditional PCR*, or real-time PCR, made in a thermal cycler for automatic time and temperature control

RT-PCR must not be confused with *real-time* polymerase chain reaction, or *quantitative* PCR (Q-PCR, or qRT-PCR)

Thermal cycler for PCR

IV.A.5 Cloning with plasmids

Another laboratory technique to replicate pieces of DNA uses ***plasmids*** (extranuclear DNA) of bacterial (e.g. E. coli) as *vectors to clone* DNA sequences

DNA fragments to be cloned (exogenous to the bacterium used as a carrier) are inserted in the DNA sequence of the plasmid, using ***restriction enzymes*** to cut the DNA of the plasmid and the ***DNA ligase enzyme*** to bind to the plasmid DNA the DNA fragments to be cloned, creating, in this way, recombinant plasmid (video: http://www.youtube.com/watch?v=acKWdNj936o and http://www.youtube.com/watch?v=x2jUMG2E-ic)

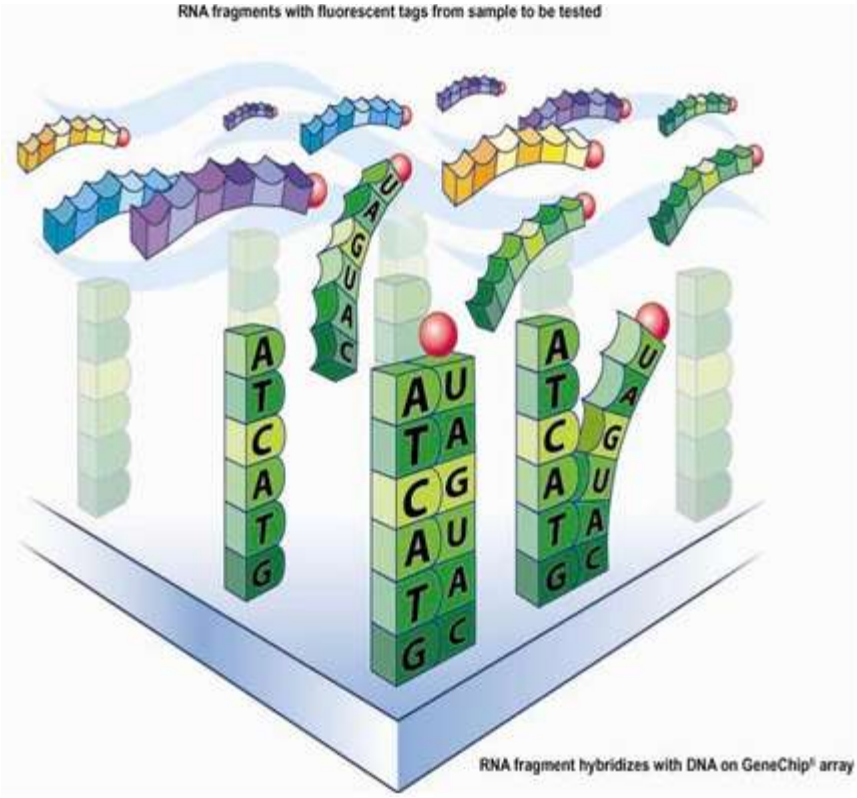
IV.A.6 DNA microarrays

Microarrays: orderly and miniaturized arrangements of *fragments of DNA* with known sequences on *solid support*

Every position contains a *fragment of DNA specific probe*, *complementary* to the transcribed sequence

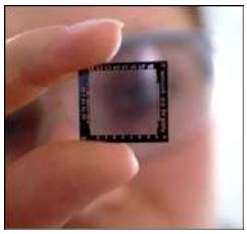
When the **probe fragment** is placed in presence of the *complementary fragment*, the test (marked with *fluorochrome*), they will tend to pair with strong interaction because of complementarity

During **scan**, the amount of *fluorescent signal* arising from a specific position in the array is directly *proportional* to the amount of the correspondent transcript in the biological sample used



http://www.phgfoundation.org/tutorials/dna/6.html

Applications: not only measurement of genetic expression:

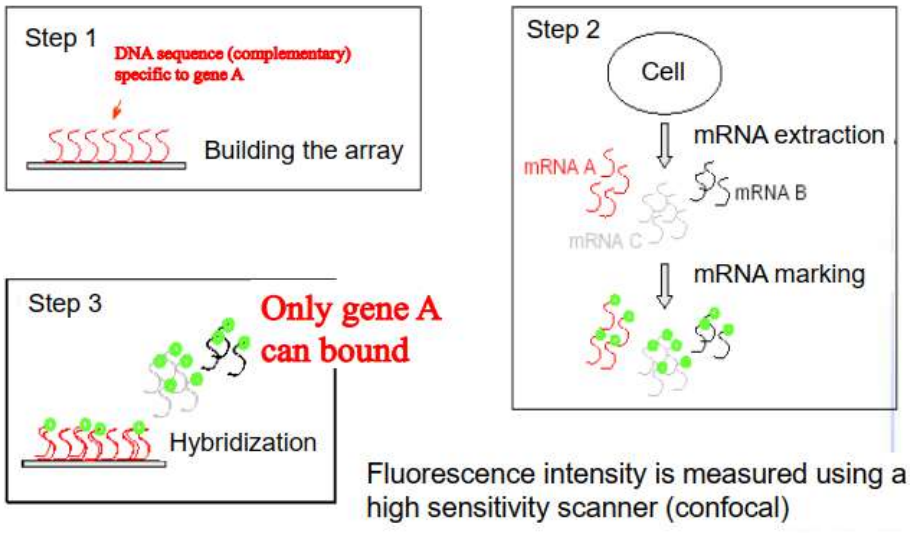


- Measurement of *abundance* of a genetic *transcript*
- Characterization of a gene sequence (i.e. exons / introns)
- Characterization of *alteration* of the *number of copies* of a given gene or DNA sequence (i.e. due to chromosomal mutations of duplication)
- Characterization of *DNA-proteins interactions*
- …

DNA microarrays: principle and main steps:

1. *Building* of the microarray (solid holder where thousands of sequences of different genes, called probes, are firmly placed, in well-known locations)
2. *RNA full extraction* from the cells to be examined (test)
   - *Retro-transcription* to cDNA (coding DNA), if needed
   - *Amplification* and *marking*
3. *Hybridization* (of the test) to the microarray – the
4. *Evaluation* of gene activity

Since they allow to determine the *profile of expression* of the cell in a given state, it is also said that microarrays allow **expression profiling**



A technology that changed the way to study molecular biology

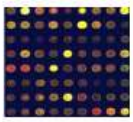
Traditional methods: a gene/some genes are observed in one experiment: a global vision is missing. They are used in *hypothesis-guided* research

Microarray technology: *thousands of genes* on an array to study their functions simultaneously. Generating hypothesis research, data driven

The two most used technologies: **cDNA microarray** and **oligonucleotides chip**. Both measure genetic expression levels, through mRNA abundance

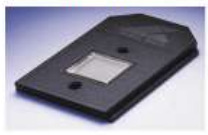
cDNA microarrays

Detail:

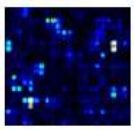


- ~25'000 clones in 5 cm x 2 cm slide
- Fluorescent marking
- 2 experimental conditions for each array slide

Oligonucleotide arrays
(oligo chips)



Detail:

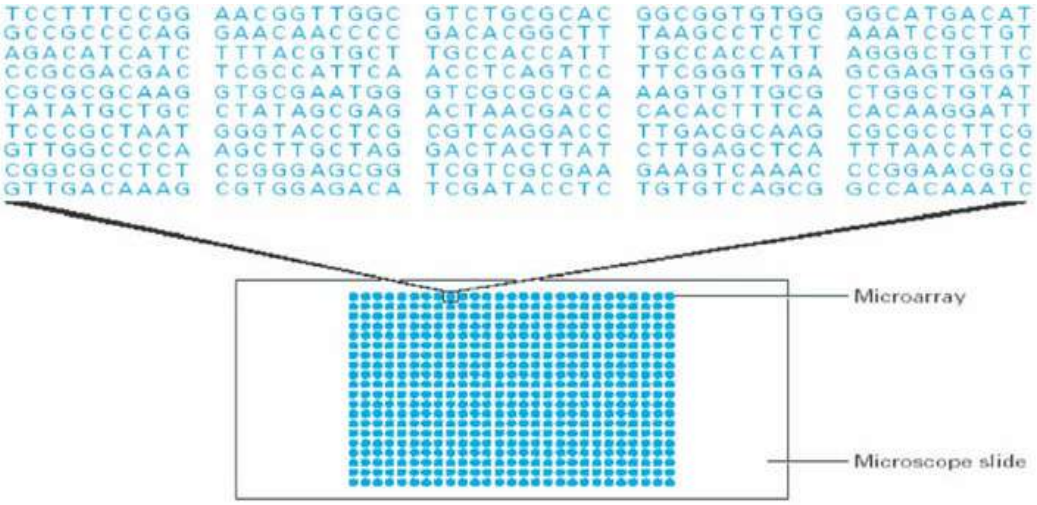


- ~60'000 genes in 1.28 cm x 1.28 cm
- Fluorescent marking
- 1 experimental condition for each array

IV.A.6.1 cDNA microarrays (spotted)

Building of the *cDNA microarrays*: *full* sections of *ESTs* (Expressed Sequence Tags, short sub-sequences of a transcribed cDNA sequence)

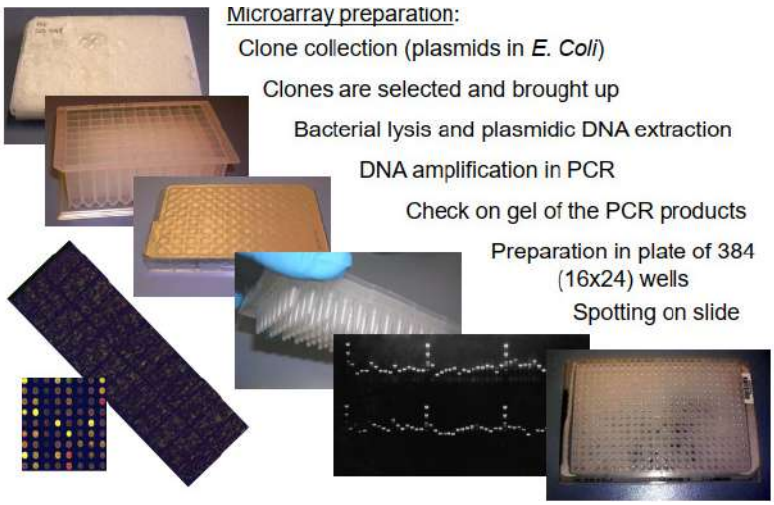
For every gene, a lot of copies of EST (500-5000 bp) are arranged. They are obtained using cDNA libraries (boosted with RT-PCR) on a *spot* of a *slide* (10-50 spots per mm2). Each EST should be specific (unigenic set)



Sample preparation: *two mRNA samples* are prepared, retro-transcribed into cDNA and made *fluorescent* with *different colours* (markers: Cy3, green and Cy5, red)

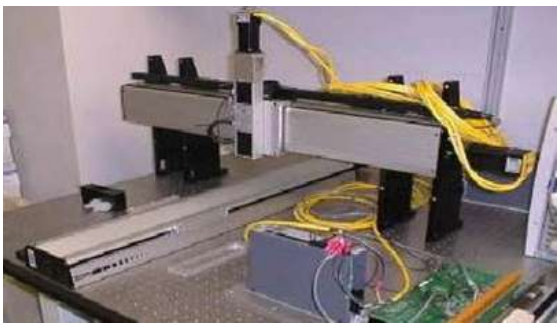
Hybridization: gene transcripts expressed in sample, prepared and marked are hybridized with their *complementary sequence* on the microarray

Measure of the *gene expression*: the *fluorescent measure* in every spot gives a measure of which genes are expressed in each of the two samples. Example: fluorescent Cy5 measure is done using a 633 nm helium-neon laser (HeNe) and the emission is at 680 nm



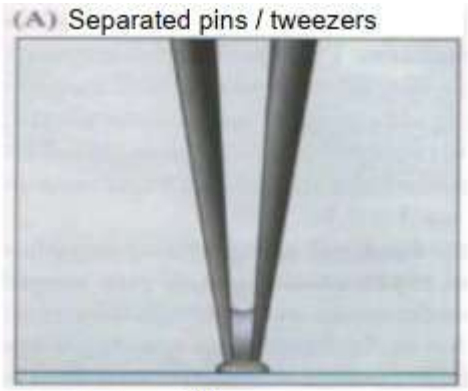
"Technical" details:

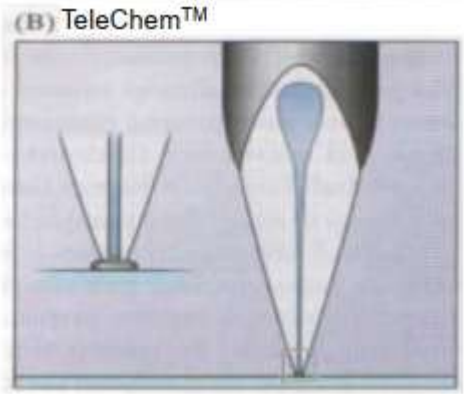
- There can be up to 15'000 elements per slide (usually about 5'000)
- Slide dimensions are typically 2.5 x 2.5 cm (longer slides can be used, but require samples with a larger amount of marked cDNA)
- Spot distance is usually 120-250 µm
- They are printed by robots with heads containing from 4 to 32 pins at a distance of about 1 cm
- **Pins** with *different shapes* exist



Video: http://www.youtube.com/watch?v=Pjr1Oyc0KrY

Several printing pins exist:



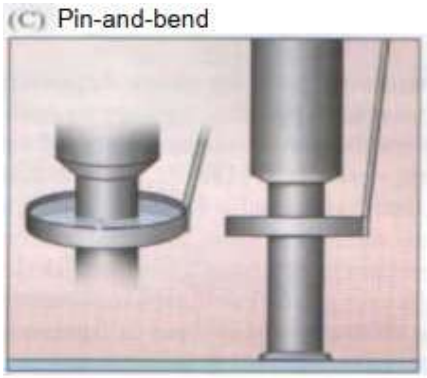


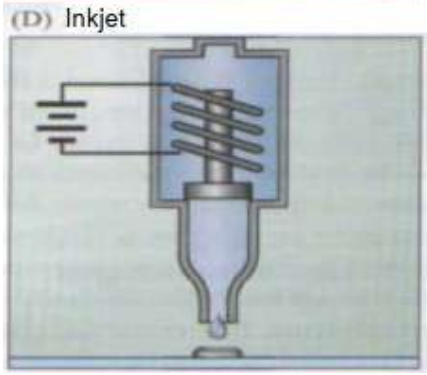
A. *Tweezers* models, or with separated pins, transfer a very small amount (nanoliters) of DNA solution to the array, by capillarity, when the pin touches the solid surface

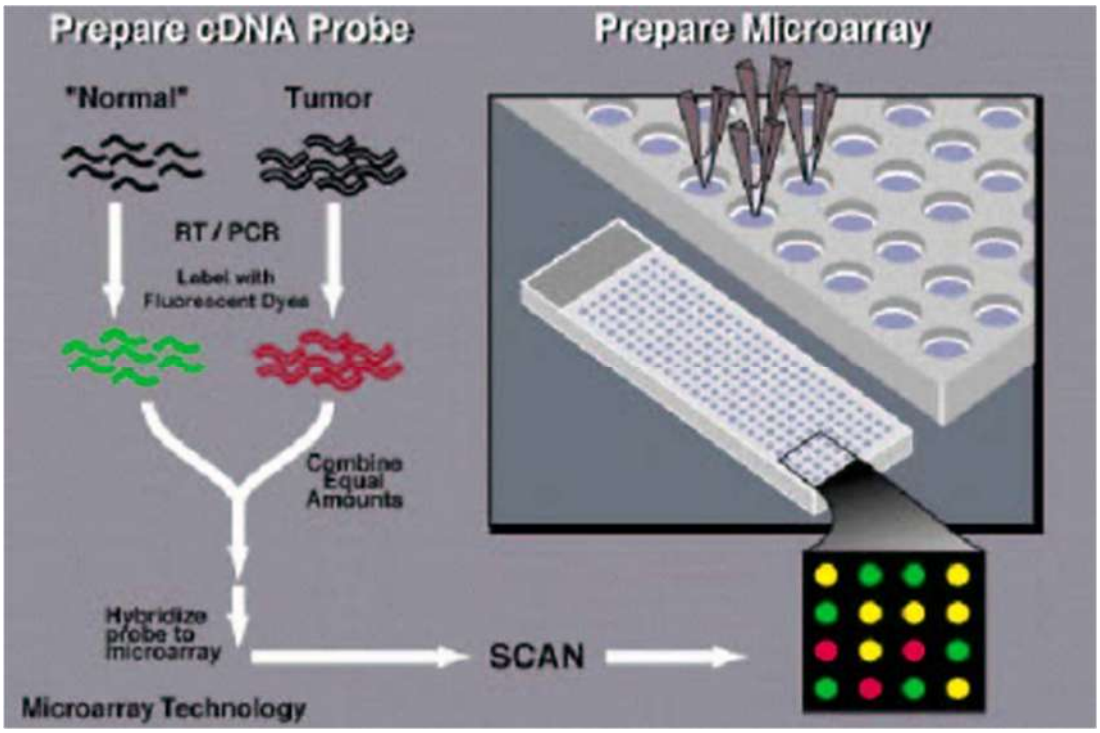
B. Pins and *TeleChemTM* tips apply small drops when the pin touches the solid surface (substrate)

C. *Pin-and-bend* models, collect DNA in a small bend and afterword a pin transfers the solution on a slide, keeping an uniform density

D. *Inkjet models* (e.g. STMicroelectronics) spray even smaller amount (picoliters) of pressured liquid drops







Sample preparation, hybridization, and measure

Even in one category of printing pin, every pin is different!

Generally, they are quite *expensive*, due to the **setup** of their preparation

- cDNA clones must be created using *EST amplification* with PCR (for each spot 10 ng of material are needed)
- Clones with a length of 1-2 kb are used

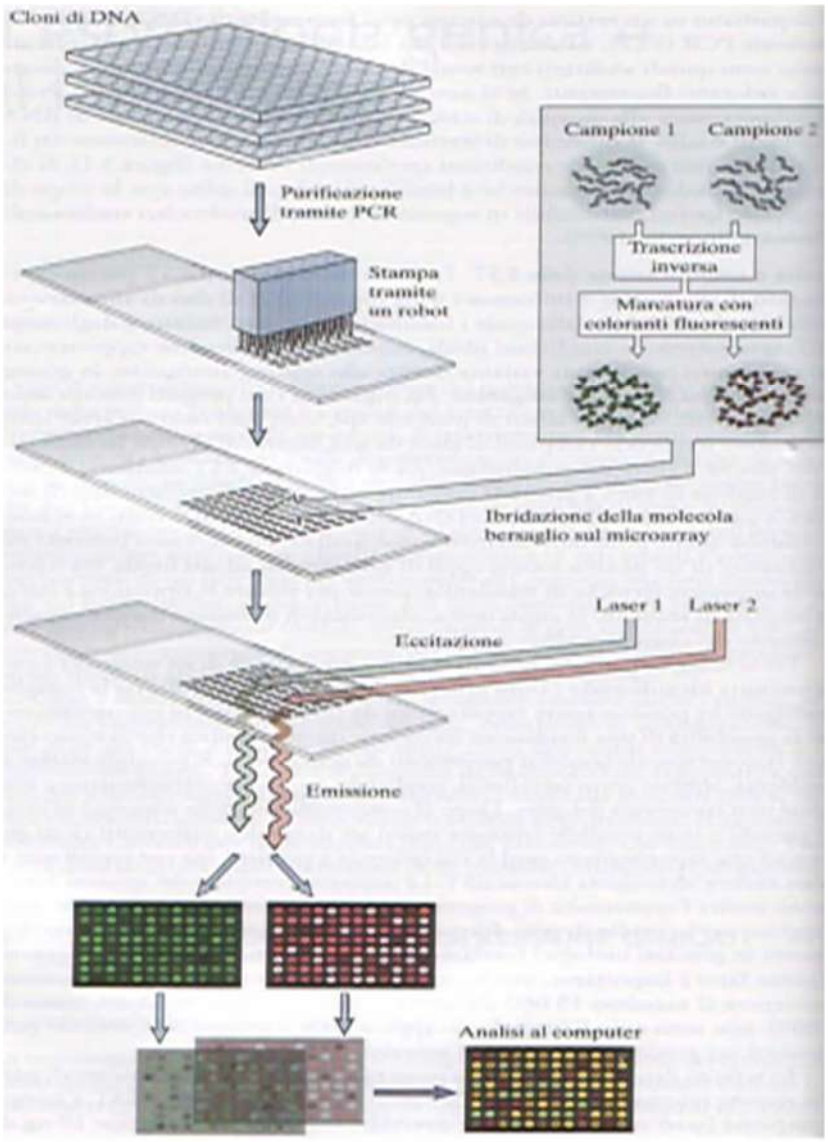
The solutions containing the amplified clones can be used to produce up to 1,000 slides

It takes *2 days* to produce *100 microarrays* with *5,000 genes*

Finally, it should produce *labelled cDNA* from the biological samples, using the reverse transcriptase in presence of fluorescent or radioactive nucleotides

Principle of the analysis with cDNA microarrays:

- After that fragment of EST arranged in 96 or 384 well plates are deposited at high density on a microscope slide, make an hybridization on the array with two different types of cDNA, labelled with different fluorescent dyes, and derived from independent samples of mRNA
- After washing, a laser scans the slide and calculates the ratio of fluorescence induced in the two samples for each EST: this value indicates the relative amount of transcript for each EST in the samples
- Video: http://www.youtube.com/watch?v=ffOgVQekKnk



Video: http://www.youtube.com/watch?v=VNsThMNjKhM

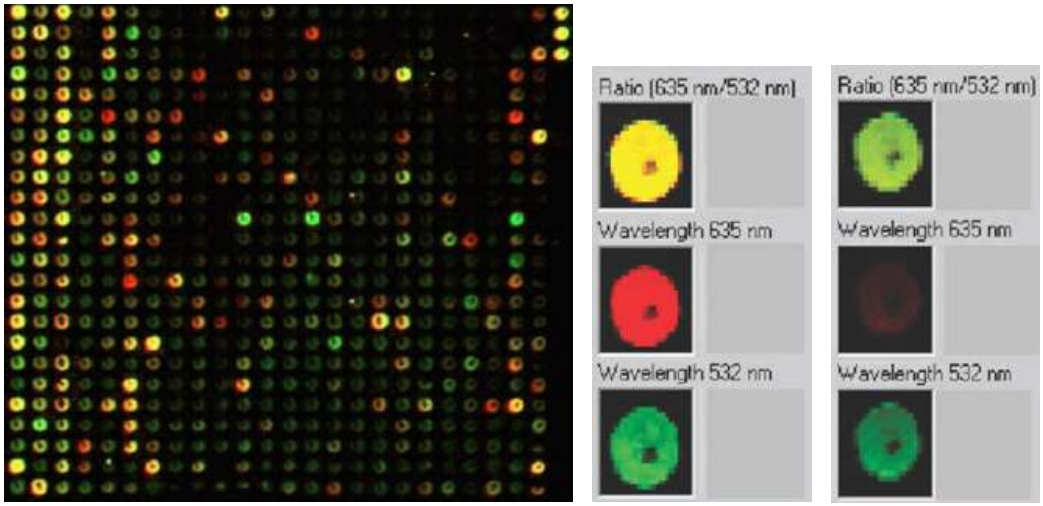
Images of cDNA microarray:

Red: expressed gene, e.g. *in tests but not in controls*

Green: expressed gene, e.g. *in controls but not in tests*

Yellow: expressed gene in *both samples*

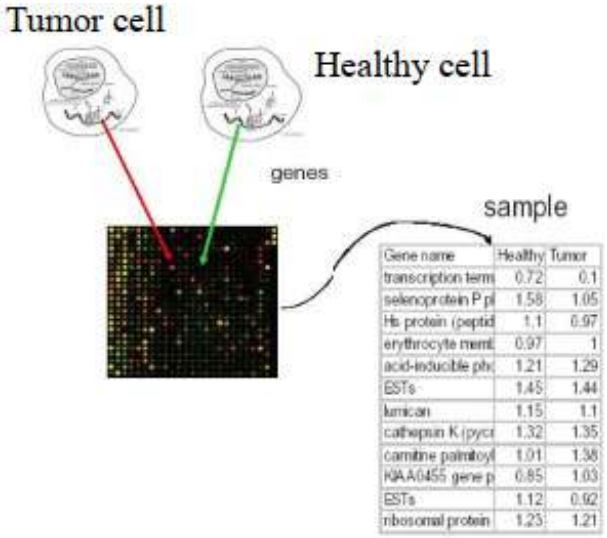
Gray: *not expressed* gene in any samples



Most of the gene are not active in a certain time. In order to quantify, we have to distinguish the background from the pixels

From images to data:

- *Grid alignment*: each probe must be localized in the array image
- *Segmentation*: identification of pixels belonging to each spot
- *Intensity extraction*: evaluation of a numeric value representing the expression level (mean, median, ...)
- *Background correction*: the background intensity is calculated and subtracted from the spot intensity value
- *Spot quality*: parameters are calculated (e.g. circularity, uniformity, size, ...) to evaluate the quality of an experiment Pros and cons of "spotted" technology:



Pros and cons of "spotted" technology

- **Competitive hybridization**: analysis of mRNA from cells in two conditions
  - *Pro*: *relative* measures (often expressed as log2)
  - *Handicap*: the definition of the *reference*, *colorimetric* problems, possible differences in the amounts of the two mRNA
- *Difficult to compare results* from different arrays: the intensity depends on the amount of probes deposited
- It takes a lot of mRNAs to prepare the target (50-200 µg)

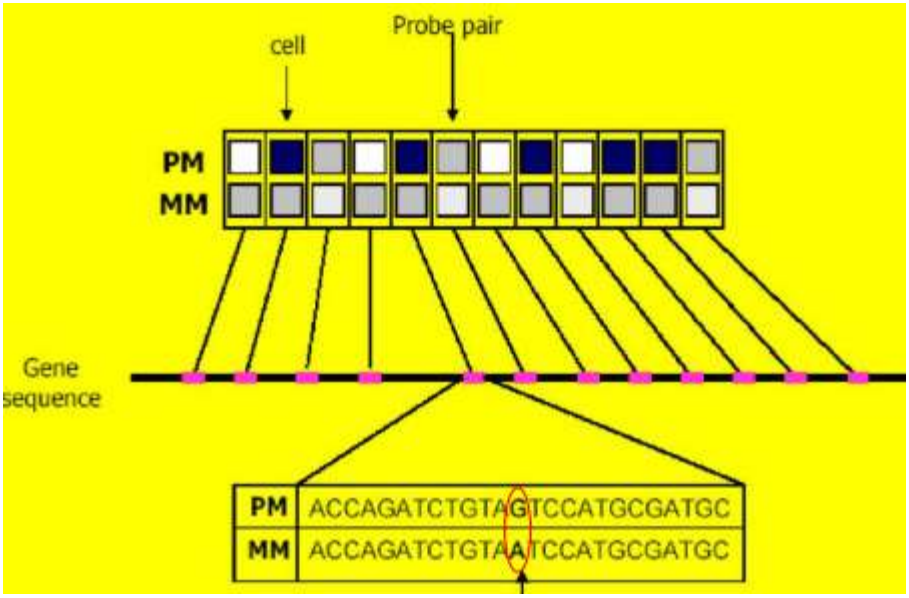
IV.A.6.2 Oligonucleotide microarrays

Oligonucleotide microarrays: Affymetrix genechips (it was the first company to provide them on the market) and others

- In place of the ESTs, there are *oligonucleotides long 20-80 bases*, designed to represent ORFs
- Composition of each *set of sequences* of oligonucleotides:
  - Perfect match (PM): a sequence that *could hybridize*
  - Mismatch (MM): a sequence that *should not hybridize*, because the *central* base is inverted

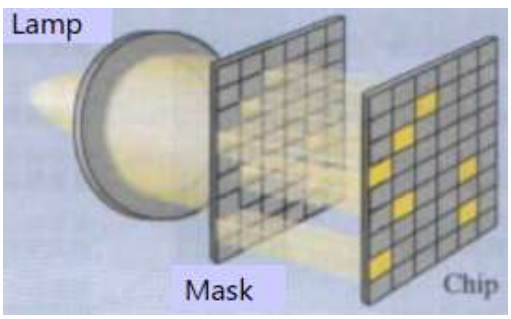
    PM ATGAGCTGATGCGATGCCATGAGAG
    MM ATGAGCTGATGCCATGCCATGAGAG
- For each probe with sequence of PM there is on chip another probe with sequence of MM

Each gene is represented as a *set of 10-20 oligonucleotides* (e.g. 25 bp long in Affymetrix chips), corresponding to some positions of the represented gene, each with PM and MM

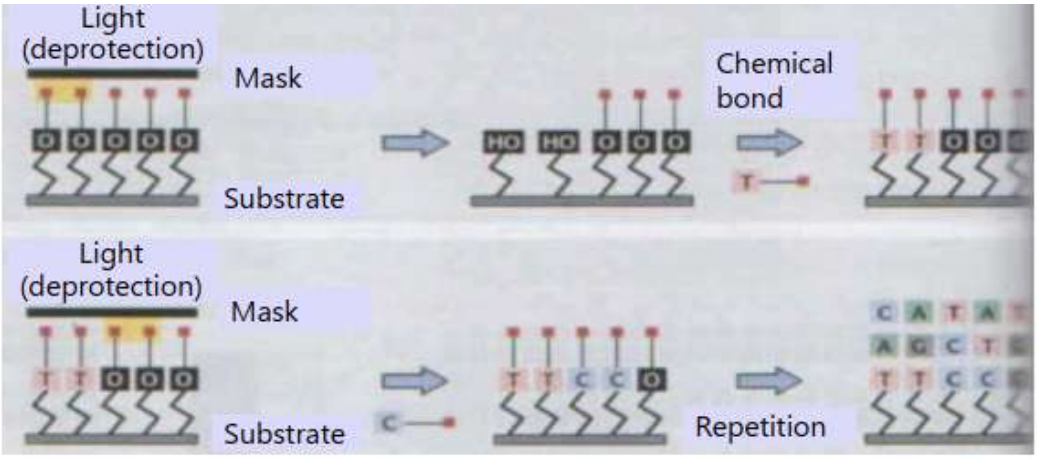


Construction of arrays of oligonucleotides (performed by specialized companies selling such arrays)

Oligonucleotides are *synthesized in situ* on the silicon chip by lithography: using a *flash of light* and a *mask*, which allows the light to hit only the required points on the surface of the chip (process similar to the production of computers' CPUs)

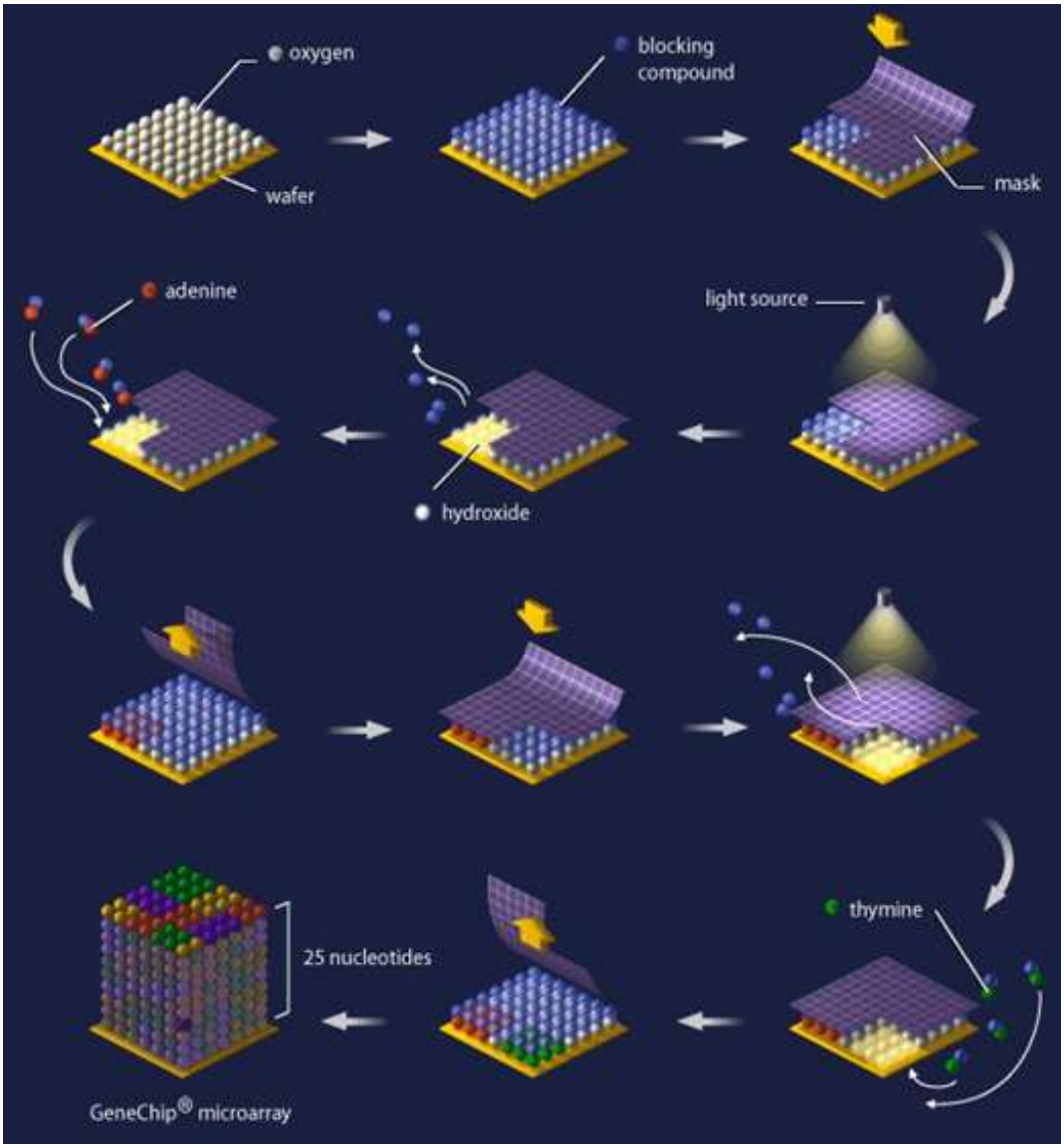


In each step, the flash of light "*unprotects*" (i.e. frees) the oligonucleotides in the *desired point* of the chip; then "protected" nucleotides of one of the four possible types (A, C, G, or T) are added, such that *only one nucleotide* is added to the desired chains



On site oligonucleotide synthesis

- On a *substrate of silicon*, oligonucleotides are synthesized through addition cycles of a specific nucleotide in specific positions; "blocked" nucleotides are deprotected through exposition to light
- Only nucleotides localized in correspondence of holes on the photolithographic mask are accessible to adding the next one

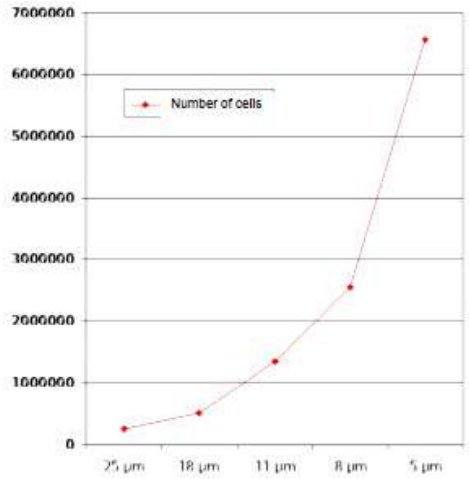


Advantages:

- *Verified* oligo sequences
- *Predetermined* oligo length

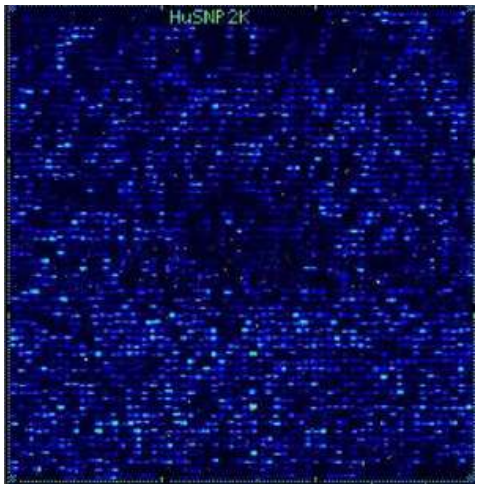
Video: http://www.youtube.com/watch?v=MuN54ecfHPw

Creation of *photolithographic masks* with constantly *increasing resolution* allows to synthesize on the same surface an increasing number of individual cells:

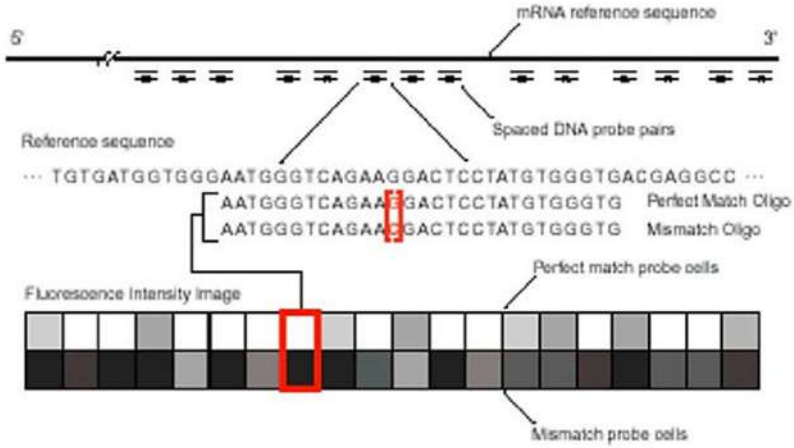


Cell dimensions (µm)	Number of cells synthesizable on "standard" array 12.8 mm^2
25	262'144
18	505'679
11	1'354'000
8	2'560'000
5	6'553'600
2	…
1	…

Each chip group of cells measures the **expression level of a gene sequence**

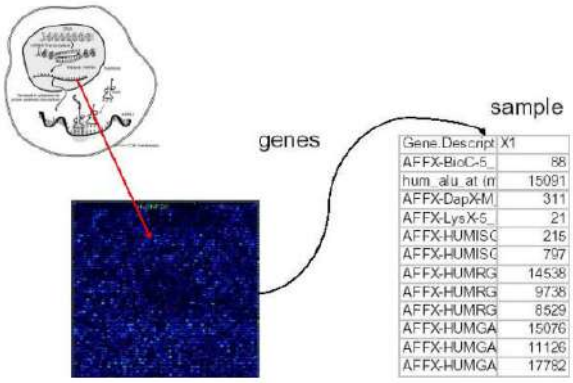


Scanned microarray (with confocal laser)



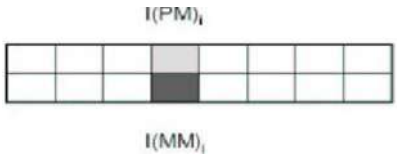
Gene expression level quantified by the intensity (I) of the chip cell in the scanned image

Measured level = $avg[I(PM) - I(MM)]$



From images to data: same steps as cDNA microarrays

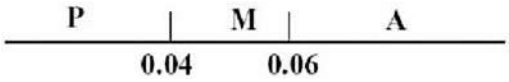
Detection scoring



- Calculate the following score:

$$Ri = \frac{I(PM)_i - I(MM)_i}{I(PM)_i + I(MM)_i}$$

- It determines the probe ability to **identify the target**
- Detection p-value: performed hypothesis test that the score *differs significantly from a close to zero threshold* (evidence of a non-random hybridization is evaluated)
  The Wilcoxon signed ranked test is used
- A ***detection call*** describes if the *hybridization* of a probe set took place (P, presence), or not (A, absence), or it was only marginally (M)
  It is assigned on the basis of p-value
- Suggested values:
  - *Presence*: p-value < 0.04
  - *Marginal*: $0.04 \leq$ p-value ≤ 0.06
  - *Absence*: otherwise



Remark:

- Procedure very *fast and effective*, but it is very *expensive* to make the masks, so this technique is carried out by specialized companies and used only for *model organisms*
- It also requires to appropriately *design* the combination of oligonucleotide sequences that *discriminate* among the various ORFs
- Note: as **sample** (test) a marked and amplified mRNA is used instead of cDNA (as in cDNA microarrays)

IV.A.6.3 cDNAs vs oligonucleotides

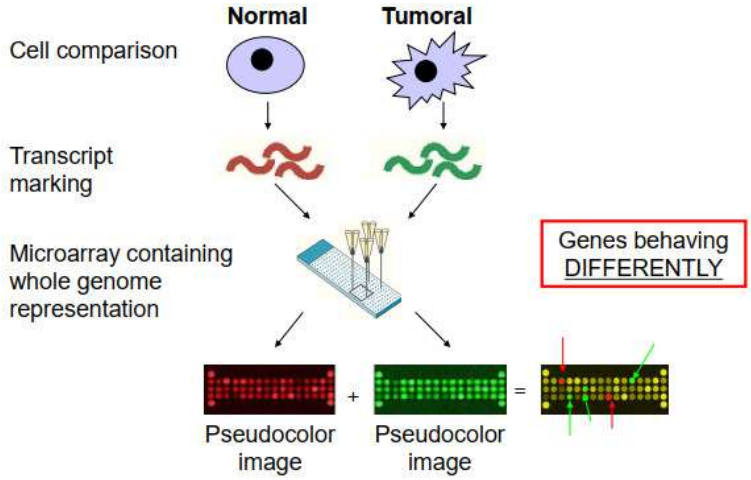
cDNA microarrays (for specific tasks, used a lot in clinical settings):

- They can be *applied to any organism* without the need to have sequenced the complete genome
- Overall are *cheaper* (but expensive setup)
- They are more *flexible* and rely on hybridization between many bases and not a few
  In this way they overcome some problems associated with polymorphisms
- Currently, there are also some other solutions offered by Agilent (http://www.home.agilent.com/) that use longer oligonucleotides (with 60-80 bases) and ink-jet deposition

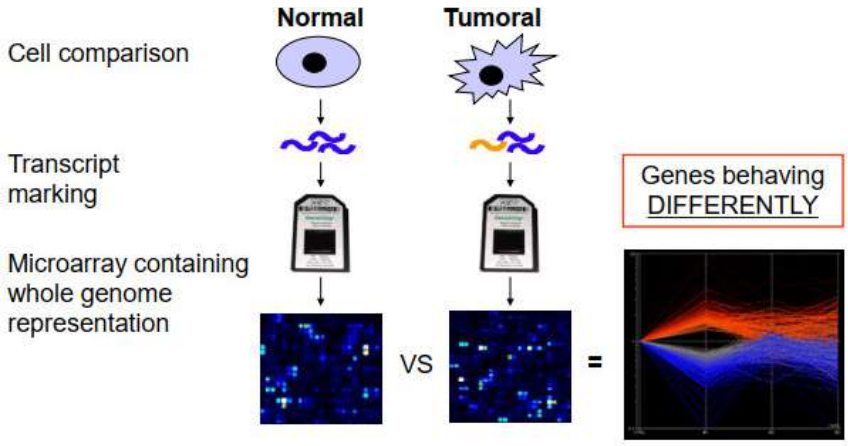
Chips of oligonucleotides (for an initial analysis [transcriptome for example]):

- They can contain a *higher amount of genes*, even *predicted* genes, that have not been inserted already in cDNA libraries
- Can be used only for *sequenced organisms*
- They can be used even by who cannot build a slide
- They have *less variability* between one chip and another
  It is easier (even if with difficulties) to *compare data* generated by different research groups

Two colour frameworks 2 biological samples on the same array

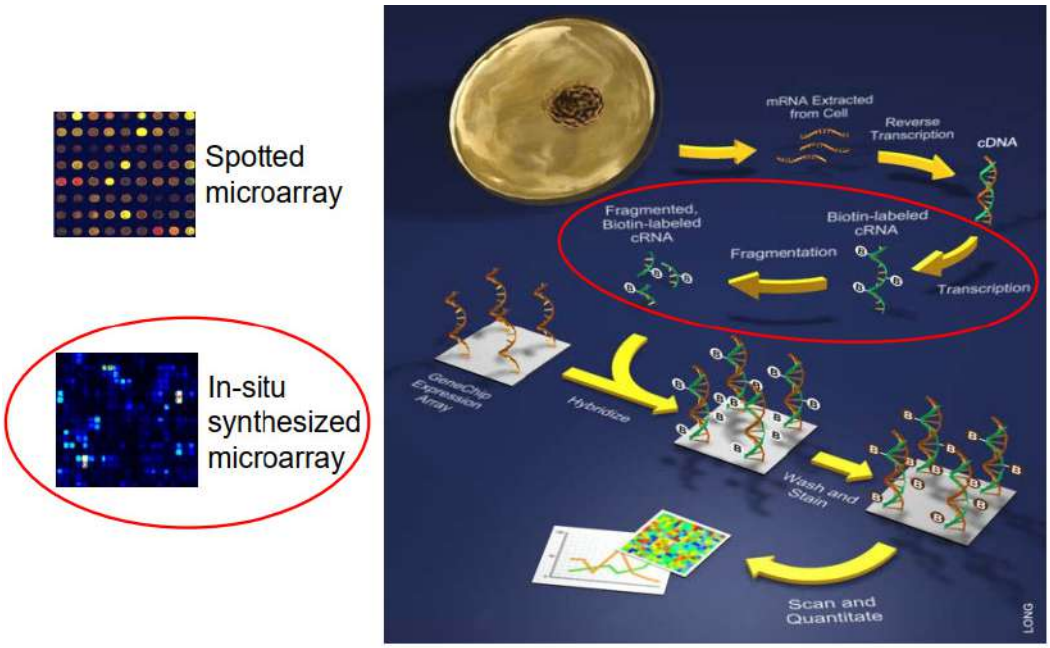


Single colour framework: 1 biological sample on 1 array



Expression variation (y axe) of several genes (lines) over-[orange] and under-[blue] expressed in subsequent temporal instants (x axe)

IV.A.7 Summary of a microarray experiment



Expression levels of tens of thousands of transcripts are measured in a *single* experiment. For this reason, it is called *genomic analysis*

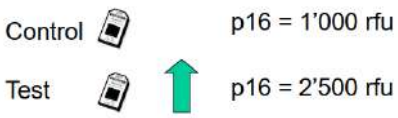
ID Transcribed **Expression levels (rfu) [relative fluorescent units]** ***p*-value**

Analysis Name	Probe Set Name	Stat Pairs	Stat Pairs Used	Signal	Detection	Detection p-value
A2_R-372	7332_at	16	16	2220.6	P	0.000219
A2_R-372	7333_at	16	16	727.7	A	0.204022
A2_R-372	7334_at	16	16	958.0	A	0.060419
A2_R-372	7335_at	16	16	161.4	A	0.060419
A2_R-372	7336_at	16	16	286.7	M	0.054470
A2_R-372	7337_at	16	16	146.4	P	0.021866
A2_R-372	7338_at	16	16	42.8	A	0.378184
A2_R-372	7339_at	16	16	446.6	P	0.021866
A2_R-372	7340_at	16	16	155.5	A	0.189687
A2_R-372	7341_at	16	16	561.7	P	0.008689
A2_R-372	7342_at	16	16	1418.4	P	0.003067
A2_R-372	7343_at	16	16	3.9	A	0.975289
A2_R-372	7344_i_at	3	3	3.3	A	0.937500
A2_R-372	7345_s_at	16	16	56.1	A	0.520620
A2_R-372	7346_at	16	16	44.1	A	0.641310
A2_R-372	7347_at	16	16	963.1	P	0.000219
A2_R-372	7348_at	16	16	20212.9	P	0.000219
A2_R-372	7349_at	16	16	4713.3	P	0.000266
A2_R-372	7350_at	16	16	939.4	P	0.000388
A2_R-372	7351_at	16	16	507.8	P	0.008689
A2_R-372	7306_at	16	16	418.4	P	0.000562
A2_R-372	7307_at	16	16	465.3	P	0.000266
A2_R-372	7308_at	16	16	340.5	P	0.000218

Array reading results

Search of *regulated transcripts* (in test-control comparison)

- Comparative analysis allows to compare, for each represented transcript, the *expression level* of one condition to another. Directly comparing the *expression level* of the same *transcript* [probeset]
- In this way, it is possible to *identify* and to *quantify accurately alterations at transcriptional level* between two samples



rfu : relative fluorescent unit

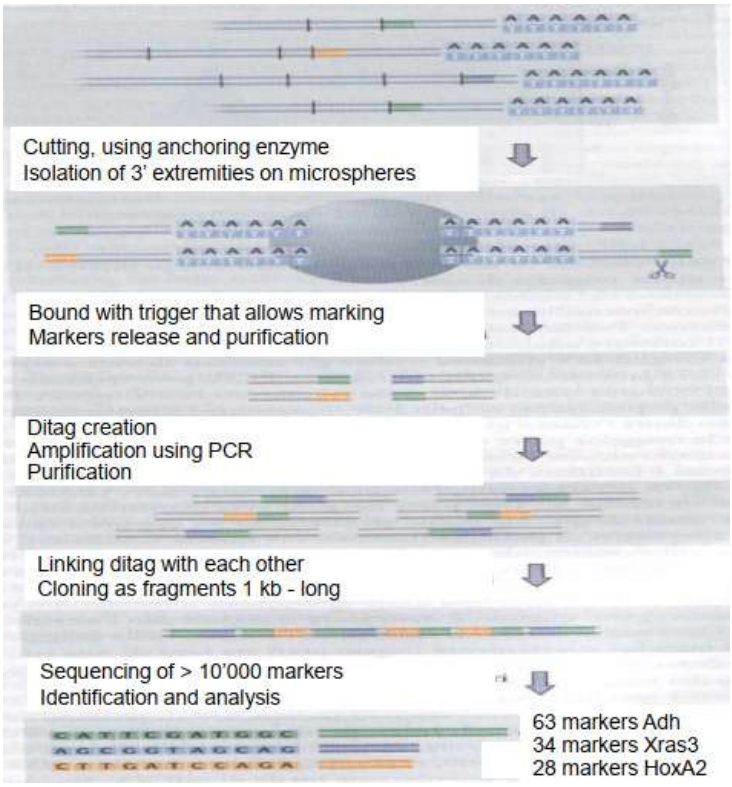
IV.A.8 Serial Analysis of Gene Expression (SAGE) [add.]

SAGE analysis:

- Method used to determine absolute abundance (or concentration) of every transcript expressed in a cell population, by using automatic sequencing in series of specific markers for each gene, produced by connected molecular techniques
- Principles upon which SAGE is based:
  - A short sequence of 10-14 bps contains sufficient information to univocally identify a transcript. These sequences can be attached to each other in order to create a longer sequence
  - The number of times a short sequence occurs measures the expression level of the correspondent transcript
- Therefore, unique markers, 15 bps long, are sequenced in series. In every sequencing reaction, 50 markers are obtained
- Usually, two markers for each gene are used, so 50'000 markers are needed for the whole human genome
  - Complex and expensive procedure (5'000 Euro for each sample). Unfeasible to repeat experiments several times
  - It provides what can be considered the **exact measure** of the **transcript number**

SAGE analysis steps:

- Isolate mRNA from biological sample to be analyzed
- mRNA to cDNA transcription
- Cut of appropriate cDNA sequences with restriction enzymes, in order to get short sequences
- Attach an adapter to create a "di-tag"
- Connect di-tags between each other
- Amplify, using a bacterial vector, long di-tag chains
- Sequence amplified chains
- Recognize, using software, short sequences, count them and associate them with the related transcript



IV.A.9 Microarray and Gene Expression Data (MGED)

Standardization of microarray data and annotations (comments here are true for microarray, but also for any technology that produces the same type of data)

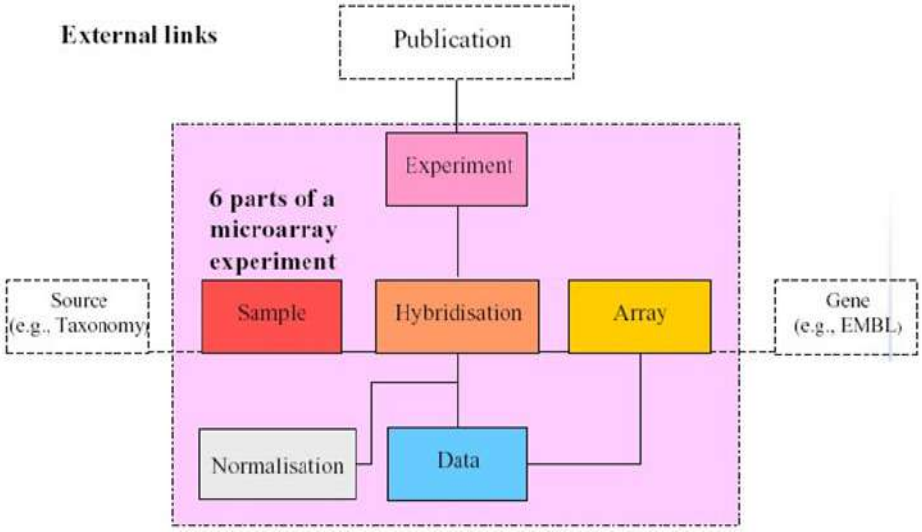
- MGED team (Microarray and Gene Expression Data): http://www.mged.org/
- Team's goal is to simplify:
  - Use of *standards* for DNA microarray *experiment annotation* and *data representation*
  - Introduction of *experimental tests* and data/result *normalization* methods
- Several international centers are involved (TIGR, Affymetrix, Stanford, Sanger, Agilent, Rosetta, etc.). Coordinated by European Bioinformatics Institute (EBI)

Glossary MGED (http://www.mged.org/)

- *MIAME* (Minimum Information About a Microarray Experiment): standard for *experiment* annotation
- *MAGE-OM* (MicroArray Gene Expression - Object Model): *model* of the data generated by microarrays
- *ArrayExpress* (http://www.ebi.ac.uk/microarray-as/ae/): *database* based on MAGE-OM
- *MAGE-ML* (MicroArray Gene Expression – Markup Language): *markup language* to share, among databases, the experiments and their data
- *Expression Profiler* (http://www.ebi.ac.uk/expressionprofiler/): tool for the analysis of microarray data that directly uses ArrayExpress

MIAME General principles [Brazma et al., Nature Genetics, 2001] – The standard to define the information to be collected

- Information **acquired** should be enough to *interpret the results* and to *replicate* and integrate experiments
- Information **structure** should allow queries and *automatic analysis* on data

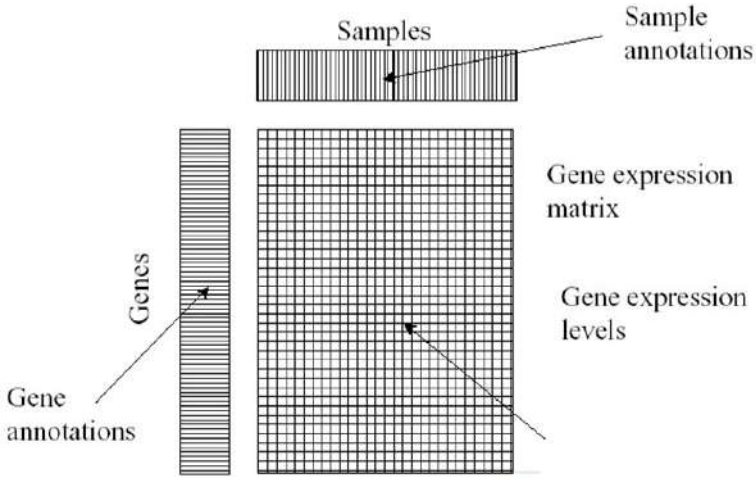


ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/)

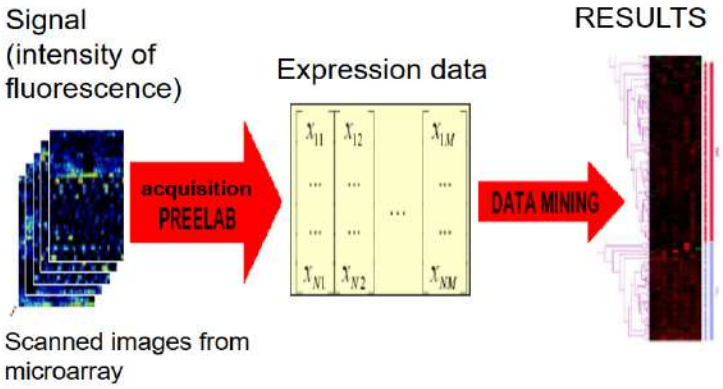
- Database with microarray experimental data, where information are described in a standard way
  It complies with MIAME
- Web interface:
  - *Queries*: genes, conditions, experiments, arrays, samples
  - *Browsing*: gene or experiment views

Database content in the context of data analysis

- *Samples* – Annotations
- *Genes* – Annotations
- *Genetic expressions* – Expression levels



IV.A.10 Analysis of expression data



IV.A.10.1 Data acquisition and signal pre-processing

Acquisition and **pre-processing** of the signal (fluorescent intensity) consists of:

- Image *analysis*
- Image/data *normalization*
- Data *transformation*

Software: GenePix, MAS5, ...

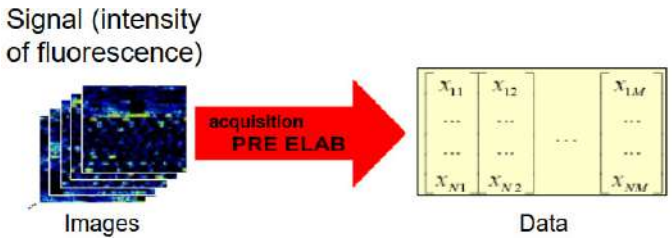
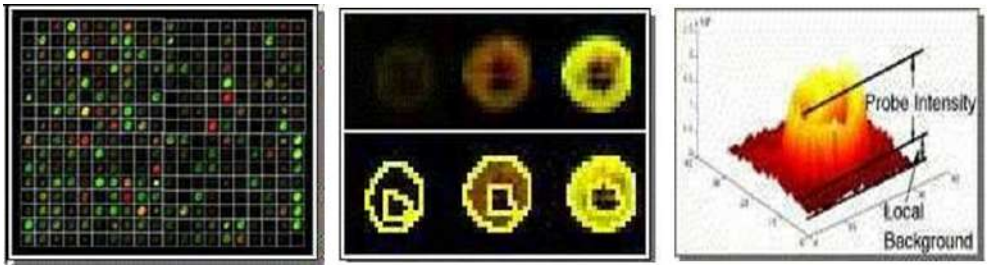


Image analysis*:*

- Identify *probe positions* related to every gene (gridding)
- *Distinguish* pixels related to foreground and background (segmentation)
- *Intensity extraction* for various intensities of the image



Images/data normalization:

- Identify and remove **systematic error**
  - *Different concentration of probes* and different hybridization efficiency lead to different brightness in different array measures
    - For experiments with *several arrays* and oligonucleotides, it is possible, for instance, to scale measures such that average intensities are all the same
  - *Normalization based on a set of genes* (housekeeping genes) whose expression must be invariant in different experimental conditions
- *Software*: dChip (http://www.dchip.org/)

Data transformation:

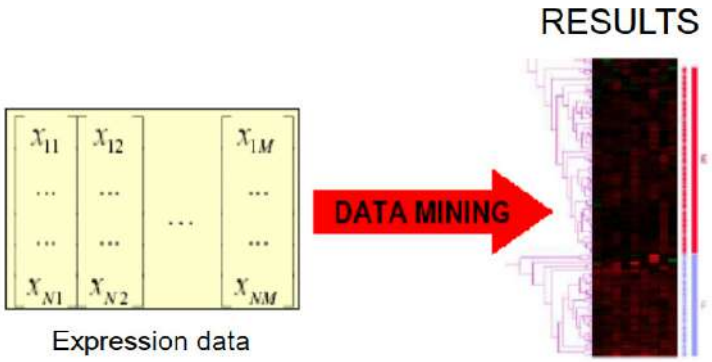
- *Logarithmic* transformation
- *Outliers'* detection
- *Missing values* management

IV.A.10.2 Data mining

Selection of differently expressed genes

Clustering, class discovery: *Unsupervised*

Classification, class prediction: *Supervised*



IV.A.10.3 Microarray data analysis issues

Data set *dimensions*

Different *supports*

Different *technologies* on different platforms: Oligonucleotides / spotted cDNA

External database references are *not stable*: identification codes of sequences on microarrays can change in different database versions

Array and sample *annotations*: they are often incomplete/not sufficient, written using a non-standard terminology

IV.A.10.4 Microarray data analysis tools

Expression Profiler (http://www.ebi.ac.uk/expressionprofiler/), completely integrated with DB ArrayExpress

Bioconductor (http://www.bioconductor.org/)

- Data analysis tool
- It is the result of an open-source software project

Specialized tools

- Public / open source
- Commercial

IV.A.11 References

Some references:

- *Interactive example of a microarray experiment*: http://www.bio.davidson.edu/courses/genomics/chip/chip.html
- *PCR interactive animation*: http://www.dnalc.org/ddnalc/resources/pcr.html
- *DNA hybridization*, *sequencing*, *PCR*, *microarray*: http://www.phgfoundation.org/tutorials/dna/
- "Our" tools:
  - GAAS: http://www.bioinformatics.deib.polimi.it/GAAS/
  - Microgen: http://www.bioinformatics.deib.polimi.it/Microgen/
- Orange - Open-source data visualization and analysis tool: http://orange.biolab.si/

IV.B DNA Microarray data analysis (26 Oct.)

IV.B.1 DNA Microarrays

IV.B.1.1 Microarrays

Microarrays: microscope slides or chips that contain ordered series of probes

- DNA → DNA microarray
- RNA → RNA microarray
- Protein → Protein microarray
- Tissue → Tissue microarray

Here, we focus on DNA microarrays used to determine expression *levels of genes* (***expression profiling***). Goal: study the effect of treatments, diseases, developmental stages, etc. on gene expression. DNA microarrays can also be used to analyse *gene sequence* in a sample (SNP analysis, minisequencing)

Spotted microarrays, where probes are

- Small *fragments of PCR* products that correspond to mRNA
- *cDNA* (complementary DNA, synthesized from a mRNA template [helix opened during the hybridisation step])

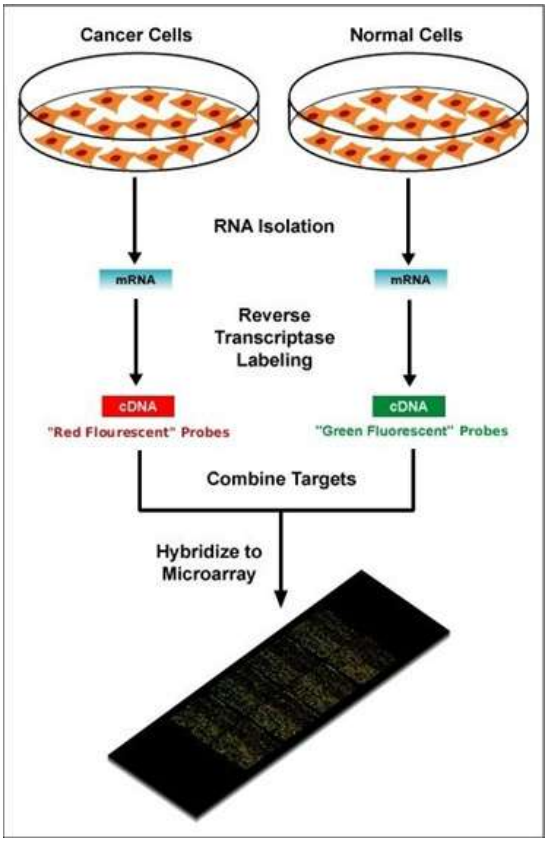
Oligonucleotide microarrays, where probes are short sequences designed to match parts of the sequence of known or predicted open reading frames

IV.B.1.2 Two channels (cDNA microarrays)

Typically hybridized with cDNA from *two samples to be* ***compared*** (e.g., diseased tissue vs. healthy tissue) and labelled with two different dyes

- Each gene is represented by *one partial cDNA clone*
- *Heat variation cycles* are used to break cDNA double strain bonds and allow hybridization
- *Dye relative intensities* are used to identify up-regulated and down-regulated genes

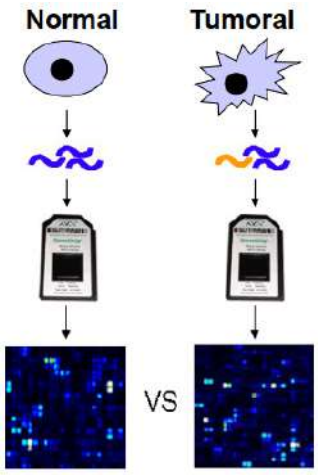
Two samples on the same microarray



IV.B.1.3 Single channel (oligonucleotide microarrays)

Arrays are designed to give estimations of the **absolute levels** of gene expression. Comparison of two conditions requires two separate single-dye hybridizations on two single channel microarrays

- Each gene is represented as a probe set of 10-25 oligonucleotide pairs (each probe)
- Most popular single-channel system: Affymetrix GeneChip™



IV.B.2 Normalisation

Goal: obtaining quantitative information from the images we obtained through the processes mentioned before and estimate the quality of the measurement. We aim to compare the measurements in two conditions.

Normalization: process of removing *systematic variations* that affect measured gene expression levels in microarray experiments

Hypotheses:

- Measured intensities for each arrayed gene represent *gene expression levels* (major hypothesis)
  - There is no *saturation* of any probe
  - The amount of photochrom measured is proportional to the *quantity of mRNA produced*
  - Biologically relevant *patterns of expression* are typically identified by comparing measured expression levels between different states on a *gene-by-gene basis*
- We assume that, for each biological sample we assay, we have a *high-quality measurement* of the intensity of hybridization (the quality of the spot we measure is good)
- We do not take into account:
  - Particular microarray *platform* used
  - *Type of measurement* reported (e.g. mean, median or integrated intensity, or average difference for Affymetrix GeneChips™)
  - Thus, we suppose that there is no influence on the quantity of expression between the measures or tools used
- We suppose that the following are performed:
  - *Background* correction
  - *Spot-quality assessment* and trimming (outlier elimination, unless of interest for the specific analysis)

IV.B.2.1 Systematic variations

Sources of **systematic variation**:

- *Dye effect*: differences in dye (labelling) efficiencies → intensity varying bias: in a channel the intensity is higher than in the other
- Scanner *malfunction* / uneven functioning
- Uneven hybridization → *spatially varying bias*
- *Printing tips*: slides are printed with more than one pen; if any of these pens works differently from others, the corresponding sub-array could differ → *spatially varying bias*
- …

IV.B.2.2 Expression ratio

Let R (red) and G (green) denote respectively the target and reference samples

The **expression ratio** of the i-th gene on all arrays used is defined as:

$$T_i = \frac{R_i}{G_i}, \qquad i = 1, \dots, N_{gene}$$

The measures R_i and G_i can be made on either a *single two-channel* array, or on *two single channel* arrays

Issue: expression ratios treat up- and down-regulated genes differently (despite having a factor of 10, the expression ratio can be different):

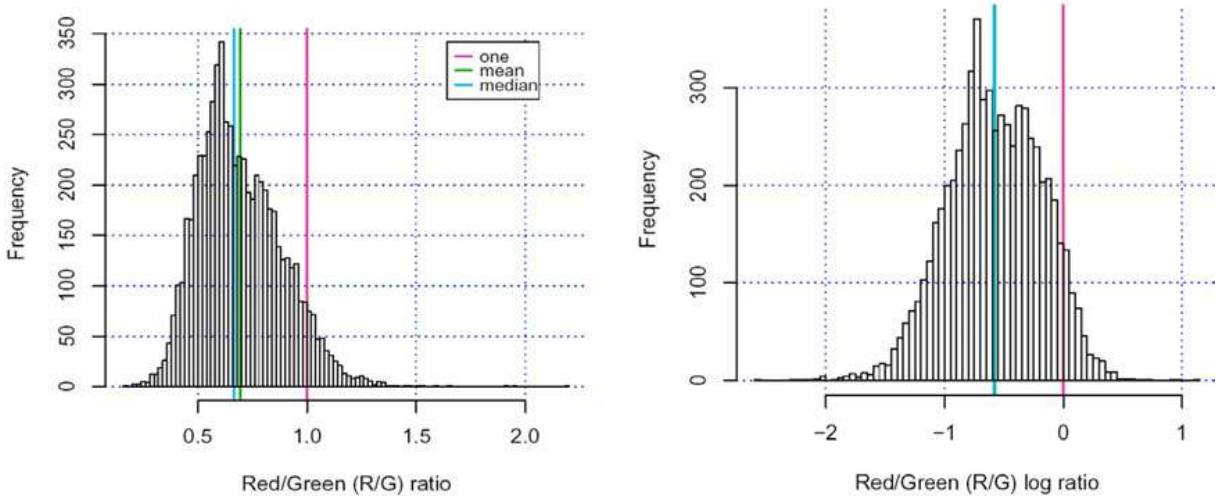
- genes *up-regulated* by a factor of 10 à expression ratio of 10
- genes *down-regulated* by a factor of 10 à expression ratio 0.1

IV.B.2.3 Log expression ratio

To solve the problem of up- and down-regulated genes, we can use $\log_2(ratio)$:

$$\log_2 T_i = \log_2 \left(\frac{R_i}{G_i} \right)$$

The distribution of log2(ratio) is **symmetric** à it *does not favour* up- or down-regulation



Comparison between expression ratio and log expression ratio

IV.B.2.4 Assumptions

Let us assume that:

- We are starting with **equal total quantities of mRNA** for the two samples we are going to compare
- Given that there are millions of individual mRNA molecules in each sample, we will assume that the **average mass of each molecule is approximately the same**, and that, consequently, the *total number of molecules in each sample is also the same*
- The arrayed elements (probes) represent a *total or random sampling of the genes* in the organism (note that in some applications, this might *not* be the case, e.g., for the cDNA microarray built to monitor a set of genes predefined)
    - This point is important because we also assume that the arrayed elements interrogate the two mRNA samples totally or randomly
        - If the arrayed genes are selected to represent only the genes we know will change, then we will likely over- or under-sample the genes in one of the biological samples being compared
        - If the array contains all genes or a large enough assortment of random genes of the considered organism, we do not expect to see such bias

IV.B.2.5 Global normalisation

Given the previous assumptions, we expect to observe the same average intensity on both channels in all arrays:

- Average $R/G\ ratio = 1$
- Average $\log_2(ratio) = 0$

Global normalization is achieved by:

$$\begin{aligned} R'_i &= R_i \\ G'_i &= K_{global} G_i \\ K_{global} &= \frac{\sum_{i=1}^{N_{array}} R_i}{\sum_{i=1}^{N_{array}} G_i} \end{aligned}$$

In terms of log2(ratio), we obtain:

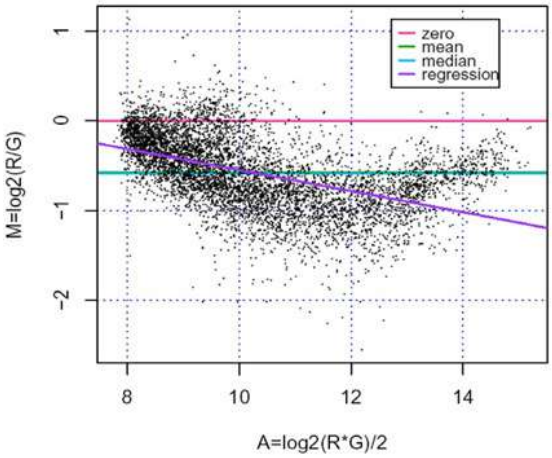
$$\log_2 T'_i = \log_2 T_i - \log_2 K_{global}$$

Possible alternatives: *equate medians*, use a *subset of arrayed genes*, ...

IV.B.2.6 Intensity adaptive normalisation

Several studies have indicated that the $\log_2(ratio)$ values can have a *systematic dependence on intensity*, which most commonly appears as a ***deviation from zero*** *for low-intensity spots*

Such dependency can be studied in **MA (*Minus-Add*) plots**



$$M_i = \log_2\left(\frac{R_i}{G_i}\right) = \log_2 R_i - \log_2 G_i$$

$$A_i = \frac{\log_2(R_i \cdot G_i)}{2} = \frac{\log_2 R_i + \log_2 G_i}{2}$$

The banana shape shows that there is not a linear dependency of the ratio to the specific absolute intensity measured

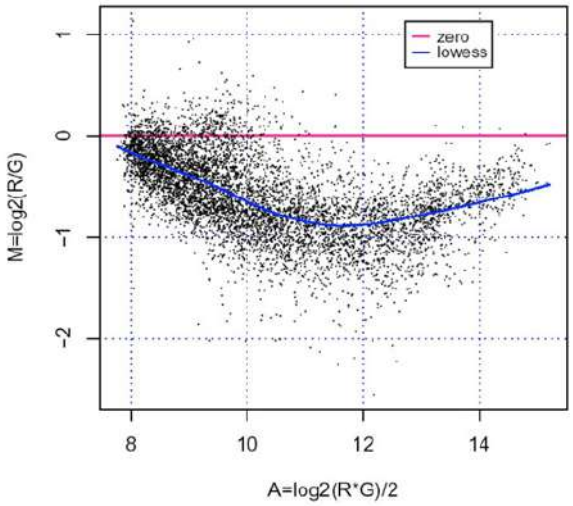
IV.B.2.7 LOWESS normalisation

Global average of M value differs from 0: corrected by global normalization

Local average of M value depends on A (*banana shape curve*)

Simple linear regression is unable to correct for the intensity-dependent bias: corrected by LOWESS normalization

LOWESS: **Locally Weighted Linear Regression**: it consists of calculating, for each A value, the regression line on the basis of a subset of points (M, A) around such A value.

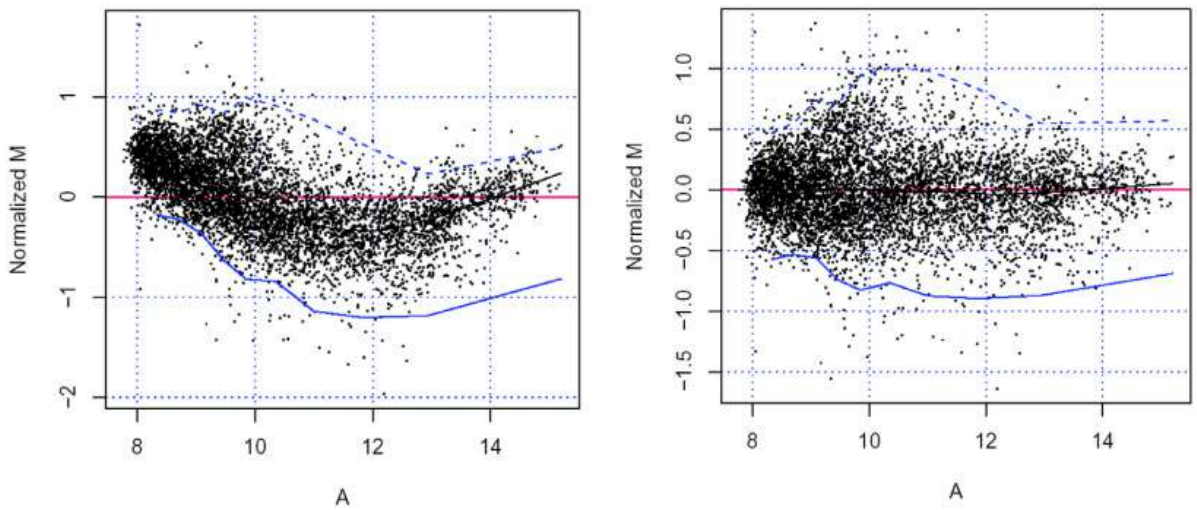


Let $M(A)$ denote the LOWESS fit for a given value of A. LOWESS correction is obtained by:

$$M_i' = \log_2 T_i' = \log_2 T_i - M(A_i) = M_i - M(A_i)$$

This equation can be made equivalent to a *transformation on the intensities*:

$$R_i' = R_i, \qquad G_i' = G_i \cdot 2^{M(A_i)}$$



Mean centering (linear normalisation) vs. LOWESS normalisation

IV.B.2.8 Local normalisation

Most normalization algorithms, including LOWESS, can be applied either *globally* (to the entire data set), or **locally** (to some physical subset of the data)

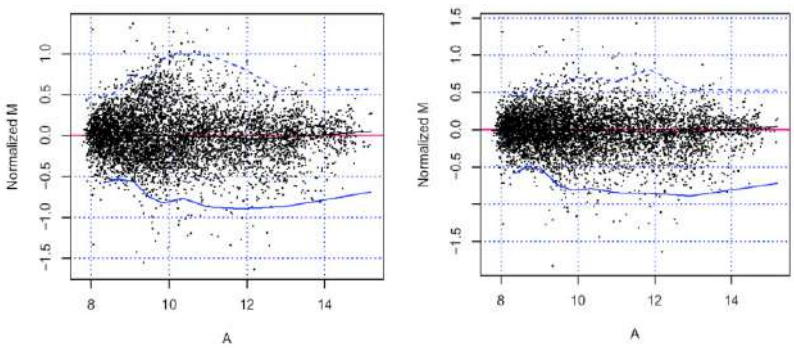
For spotted arrays, local normalization is often applied to each *group of array elements* deposited by a single spotting pen (referred to as a '*pen group*' or 'sub-grid').

Local normalization has the advantage that it can help *correcting for systematic spatial variation* in the array

Let $M_j(A)$ denote the LOWESS fit for the sub-array j:

$$M_i' = \log_2 T_i' = \log_2 T_i - M_j(A_i) = M_i - M_j(A_i)$$

if the spot I belongs to the sub-array j



Global LOWESS vs. Local (print-tip) LOWESS

IV.B.2.9 Variance regularisation

Normalization adjusts the mean of the $\log_2(ratio)$ measurements, but the **variance** of the measured $\log_2(ratio)$ values might differ from an array region to another, or between arrays

An approach to deal with this problem is to adjust the $\log_2(ratio)$ measures so that the variance does not vary.

Let σ_j^2 denote the variance of the normalized $\log_2(ratio)$ values in the j-th sub-array:

$$\sigma_j^2 = \frac{1}{N_j} \sum_{i=1}^{N_j} {M_i'}^2$$

with $M_i' = M_i - M_j(A_i)$

The **scaling factor** for the $\log_2(ratio)$ values in the j-th sub-array is:

$$a_j = \frac{\sigma_j^2}{\left[\prod_{k=1}^{N_{subarray}} \sigma_k^2\right]^{\frac{1}{N_{subarray}}}}$$

Then, all elements within the j-th sub-grid can be *scaled* by dividing their values by the scaling factor value a_j computed for that sub-array:

$$M_i'' = \frac{M_i'}{a_j} = \frac{\log_2 T_i'}{a_j}$$

Since $T_i' = R_i'/G_i'$, the equation $M_i'' = \frac{M_i'}{a_j} = \frac{\log_2 T_i'}{a_j} = \frac{\log_2 \frac{R_i'}{G_i'}}{a_j}$ can be made equivalent to a transformation on the intensities:

$$R_i'' = R_i'^{\frac{1}{a_j}}$$

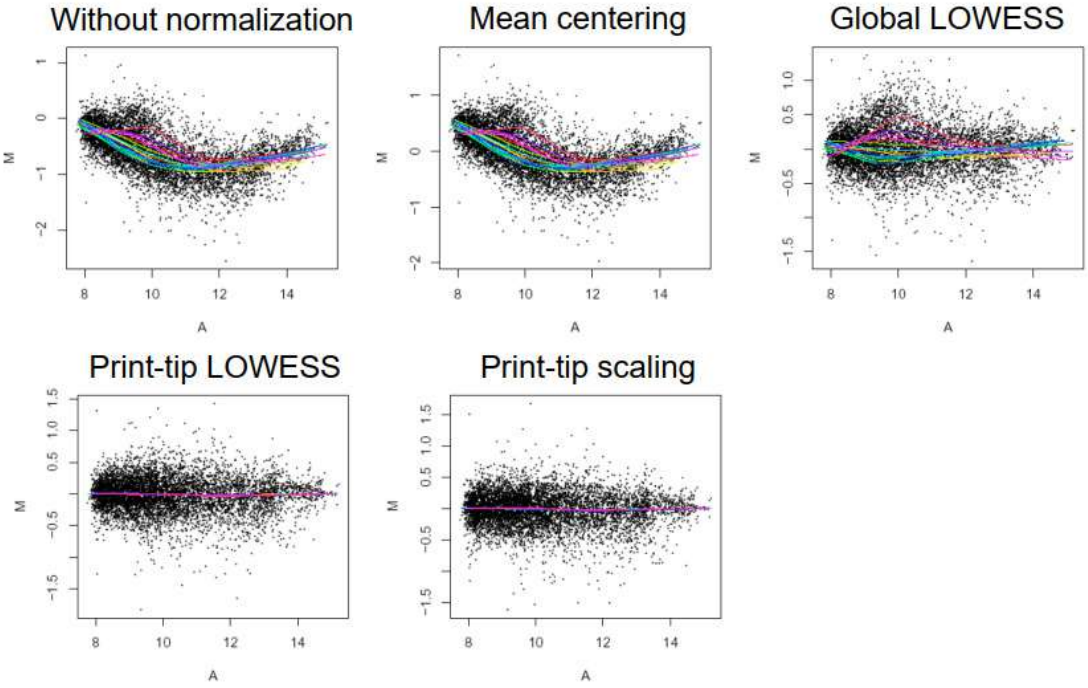
$$G_i'' = G_i'^{\frac{1}{a_j}}$$

if the spot i belongs to the sub-array j

To measure the *sub-array dispersion*, it is also possible to adopt the **MAD** (*Mean Absolute Dispersion*), which is more *robust to outliers* compared to the variance:

$$MAD_j = median_j(|M_i - median_j(M_i)|)$$

IV.B.2.10 Comparison



The colour lines are the mean of the quantity measured

We can see that the Global LOWESS even increase the difference between arrays, contrary to Print-tip methods

IV.B.2.11 Normalisation between array

Replication is essential for *identifying* and *reducing the variation* in any experimental assay

- **Technical replicates** provide information on *natural* and *systematic* variability that occurs in performing an assay
- **Biological replicas** use mRNA obtained from distinct biological sources

The particular approach used for between *array normalization* depends on the chosen experiment design. In the following, we consider two simple experiment design choices:

- *Dye-reversal* analysis (for two-channel cDNA microarrays)
- *Replicate averaging*

IV.B.2.11.1 Dye-reversal analysis

Let us assume to have two samples, A and B. Perform two hybridizations:

- $A \to Red, B \to Green$
- $A \to Green, B \to Red$

$$T'_{1,i} = \frac{R'_{1,i}}{G'_{1,i}} = \frac{A_i}{B_i}, \qquad T'_{2,i} = \frac{R'_{2,i}}{G'_{2,i}} = \frac{B_i}{A_i}$$

As we are making a *comparison* between (ideally) *identical samples*:

$$\log_2(T'_{1,i} \cdot T'_{2,i}) = \log_2\left(\frac{A_i}{B_i} \cdot \frac{B_i}{A_i}\right) = 0$$

- Consistent measurements should have the previous quantity *close to zero*, and can be averaged
- Measurements for which the previous quantity *deviates significantly* from zero (e.g., more than two standard deviations) can be eliminated from further analysis

IV.B.2.11.2 Replicate averaging

Let us assume to have *more than one replicate* for the *same experiment*

$$M'_{k,i} = \log_2 T'_{k,i} = \log_2\left(\frac{R'_{k,i}}{G'_{k,i}}\right)$$

$$A'_{k,i} = \frac{1}{2}\log_2(R'_{k,i} \cdot G'_{k,i})$$

with $k = 1, \dots, N_{replicates}$

The simplest strategy is to ***average*** *the M and A values*:

$$\overline{M_i} = \frac{1}{N_{replicates}} \sum_{k=1}^{N_{replicates}} M'_{k,i}$$

$$\overline{A_i} = \frac{1}{N_{replicates}} \sum_{k=1}^{N_{replicates}} A'_{k,i}$$

that corresponds to taking the *geometric average* of the raw measurements R and G

IV.B.3 Detection of differential expression

Goal: identification of genes that are significantly **differentially expressed** between one or more pairs of samples in the data set (after data normalization)

Simple strategies (**thresholding**):

- *Constant* thresholding:
  - Compare *fold-change* with a constant threshold τ (e.g. $\tau = 2$)
  - If $|M_i| > \tau$, then the i-th gene is differentially expressed
- *Fixed* thresholding:
  - Compute the *standard deviation* σ_M of all M values
  - If $|M_i| > c \cdot \sigma_M$, then the i-th gene is differentially expressed (c is typically 2 or 3, like for τ)
- *Adaptive* thresholding:
  - Compute a *local* standard deviation $\sigma_M^{local}(A_i)$ of M values, as a function of A
  - If $|M_i| > c \cdot \sigma_M^{local}(A_i)$, then the i-th gene is differentially expressed

More sophisticated techniques adopt the framework of hypothesis testing:

- Gene-by-gene differential expression (DE) analysis
- Gene Set DE analysis (GSEA: Gene Set Enrichment Analysis), not discussed here

Gene-by-gene hypothesis testing

- Goals:
  - *Select a statistic* that ranks the genes in order of evidence for differential expression, from strongest to weakest evidence (easier, but more important)
  - Choose a *critical value* for the ranking statistic, above which any value is considered to be significant (harder)
- Let us define the (unbiased) sample *mean* and the sample *standard variance* as:

$$\overline{M_i} = \frac{1}{N_{replicates}} \sum_{k=1}^{N_{replicates}} M'_{k,i}$$

$$s_i^2 = \frac{1}{N_{replicates} - 1} \sum_{k=1}^{N_{replicates}} (M'_{k,i} - \overline{M_i})^2$$

Here we focus on a single gene, where the different measurements are those obtained in different replicates of genes expression experiment

Hypothesis testing

- Formulate *two* ***hypotheses***:
  - Null hypothesis: the i-th gene is NOT differentially expressed
  - Alternative hypothesis: the i-th gene is differentially expressed
- Define the distribution under the *null* hypothesis
  - Assumption: the normalized $\log_2(ratio)$ measurements are ***zero-mean Gaussian*** distributed with unknown variance σ_i^2:

$$M'_{k,i} \in \mathcal{N}(0, \sigma_i^2)$$

- Compute the ***test statistic***:
  - t-statistic: $t_i = \frac{\overline{M_i}}{s_i / \sqrt{N_{replicates}}}$ with s_i the sample standard variance
  - So, it is possible to *rank* genes, by sorting them according to t_i as the evidence of differential expression
- Compute ***p-value*** (see later on)
- Compute ***significance*** (use *multiple testing correction*, if needed) (see later on)

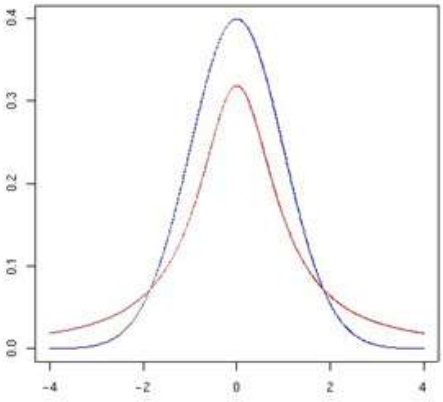
IV.B.3.1 t-static

The t-statistic is used to compare a *sample mean* to a specific value μ_0 (independent one-sample t-statistic),

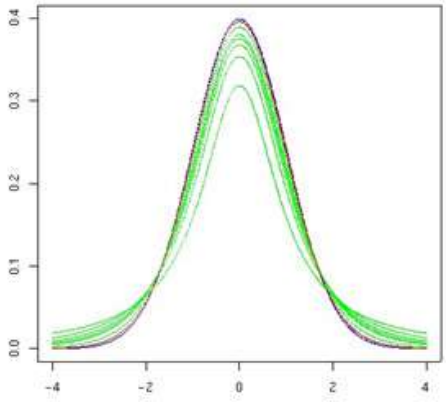
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$

where N is the sample number

If the population is *normally distributed*, under the null hypothesis the t-statistic is distributed as a t-student distribution with $(N-1)$ degrees of freedom (dof)



Normal (blue) vs. t-student (1 dof)



Normal (blue) vs. t-student (1,3,5,10,30 dof)

IV.B.3.2 z-statistic

Note that if the *true* variance σ of the population was known in advance, we would have used the z-statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$$

which is normally distributed $\mathcal{N}(0,1)$

The t-statistic *converges* to the z-statistic when the number of samples N is **large** (in our case, N is the number of replicates, which is generally low: 3, 5, or 7 at most [so definitely not large…])

IV.B.3.3 *p*-value

To compute the ***p-value***, we need to have that:

- The test statistic has been computed
- The distribution of the test statistic under the null hypothesis is defined

The p-value represents, intuitively, the probability of observing, *under the null hypothesis*, a value less likely than that of the test statistic

Let $F(t) = \int_{-\infty}^{t} f(t)dt$ denote the *cumulative density function* of the probability density function under the null hypothesis; the p-value of the i-th gene is given by:

$$p_i = 2 \cdot (1 - F(|t_i|))$$

Note: the factor of 2 is due to the fact that we are using a two-sided test, i.e. we do not distinguish between up- and down-regulated genes

IV.B.3.4 Significance

The p-value is also defined as the *significance level*

- Set a threshold α (often, $\alpha = 0.05$)
- If $p_i \leq \alpha$, **reject** the null hypothesis, the gene i is *significantly* differentially expressed (DE)
- If $p_i > \alpha$ **accept** the null hypothesis, the gene i is *not* significantly DE
- The threshold α controls the false positive rate, i.e.:
  - It sets the *probability* of *discarding the null hypothesis* when it is *true*
  - In our context, the p-value is the probability of declaring a gene to be *significantly DE* when it is *not*

IV.B.3.5 Multiple testing correction

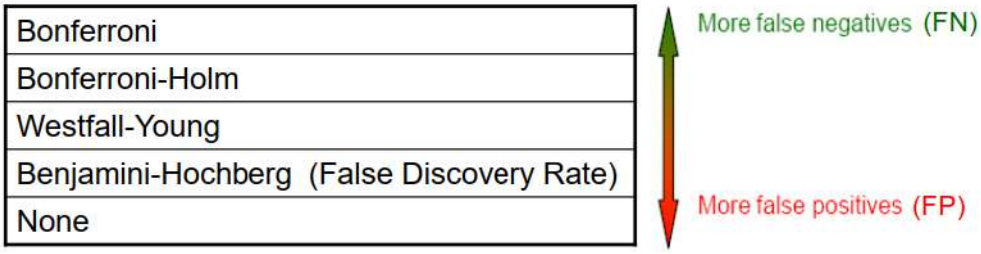
Note: defines the actual α *false positive rate* only when testing the differential expression of one gene at a time. When testing multiple genes simultaneously, as it is usually the case, *multiple test correction* is needed

Why multiple test correction? Example:

- Imagine a box with 20 marbles: 19 blue and 1 red
- What are the odds of randomly *sampling the red marble* by chance? It is 1 out of 20 (i.e., 5% chance)
- Now let's say that you get to sample a single marble (and put it back into the box) multiple times (e.g., 20 times)
- You have a much higher chance to sample the red marble (there is a 64% chance in the latter case):

$$p_{FA} = (1 - (1 - \alpha)^{N_{tests}})$$

Widely adopted multiple testing correction methods:



All these methods define different ways to *correct* (adjust) the p-value of the performed tests, in order to provide a p^{adj}-value that takes into account the *variation of test significance* due to the high number of multiple tests preformed

Bonferroni: (overly conservative: fewer FP, more FN)

$$p_i^{adj} = p_i \cdot N_{tests}$$

All the p-values are adjusted the same way

Bonferroni-Holm:

- Let p_i, $i = 1, \dots, N_{tests}$ be the ranked p-values ($p_i < p_{i+1}$)
- For $i = 1{:}\, N_{tests}$; with $N_{tests} = N_{genes}$
  - If $p_i^{adj} = p_i(N_{tests} - i + 1) \leq \alpha$, then the null hypothesis rejected: gene i is DE
  - Else, the null hypothesis accepted: gene i is not DE
- End

It controls the *probability of one or more* FP (type I errors) among all tests done (i.e., the family wise error rate, *FWER*)

Westfall-Young:

- For $j = 1{:}M$ (e.g., $M = 1000$)
  - Perform random sampling of the N_{genes}
  - Compute the p-value p_i^j for each expression level i
  - Compute the minimum p-value: $p^j = \min_{i=1,\dots,N_{tests}} p_i^j$
- End
- Compute p_i^{adj} as the fraction of p^j-values that are less than p_i (original p-value for expression level i)

$$p_i^{adj} = \frac{N_{p^j < p_i}}{N_{p^j}}$$

Accurate, but *costly* (requires *resampling*)

Benjamini-Hochberg (**False Discovery Rate** – FDR):

- Let $p_i, i = 1, \dots, N_{tests}$ be the ranked p-values ($p_i < p_{i+1}$)
- For $i = 1{:}N_{tests}$; with $N_{tests} = N_{genes}$
  - If $p_i^{adj} = p_i \cdot \frac{N_{tests}}{i} \leq \alpha$, then the null hypothesis is rejected: gene i is DE
  - Else, the null hypothesis is accepted: gene i is not DE
- End

The least conservative: more FP, fewer FN

FDR controls the *expected proportion* of *incorrectly rejected null hypotheses* (type I errors, or FP)

IV.B.3.6 Moderated statistics

Ordinary t-statistic is *not ideal* because a large t-statistic can be driven by an *unrealistically small* value of s

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$

So, genes with small sample variance have a good chance of giving a large t-statistic even if they are not DE

Alternative statistics:

- **B-statistic** is an estimate of the *posterior log-odds* that each gene is DE (i.e., the log of the ratio of the probability of being DE and not being DE)
  Values for B-statistic greater than zero correspond to a greater than 50% chance that the gene is DE
- **Penalized t-statistic** (equivalent to B-statistic in terms of ranking):

$$t_i^p = \frac{\overline{M_i}}{\sqrt{\frac{a + s_i^2}{N_{replicates}}}}$$

  where the penalty a is estimated from the mean and standard deviation of the *sample* variances s_i^2

IV.B.3.7 Example

Given the relative expression levels of two genes i and j in $N = 4$ replicate experiments:

$$\begin{aligned} M_i' &= [-0.4326, \quad -1.6656, \quad 0.1253, \quad 0.2877] \\ M_j' &= [0.8535, \quad 3.1909, \quad 3.1892, \quad 1.9624] \end{aligned}$$

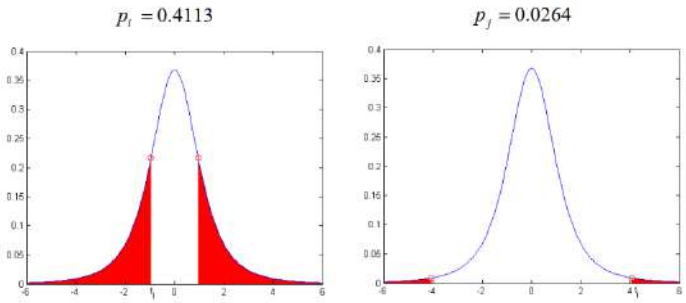
Compute the t-statistic:

$$\begin{aligned} \overline{M_i} &= -0.4213, \quad s_i = 0.8850, \quad t_i = -0.9520 \\ \overline{M_j} &= 2.2990, \quad s_j = 1.1241, \quad t_j = 4.0905 \end{aligned}$$

Compute the p-value (from t-statistic tables, with $N - 1 = 3$ degree of freedom):

$$p_i = 0.4113, \qquad p_j = 0.0264$$

If $\alpha = 0.05$, only gene j would be declared as significantly DE



IV.B.4 Experimental design of transcriptome studies

Experimental design of genetic expression studies:

"Static" experiments: two or more *subject classes* (different phenotypes / treatments)

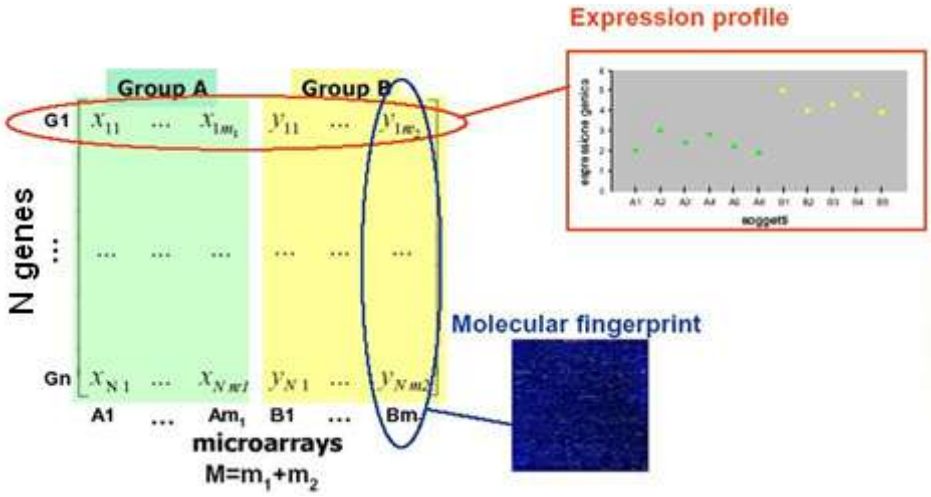
- Example 1: Selection of genes with different expressions (in different subject classes)
- Example 2: Classification (supervised)

"Dynamic" experiments: same subject in *different times*

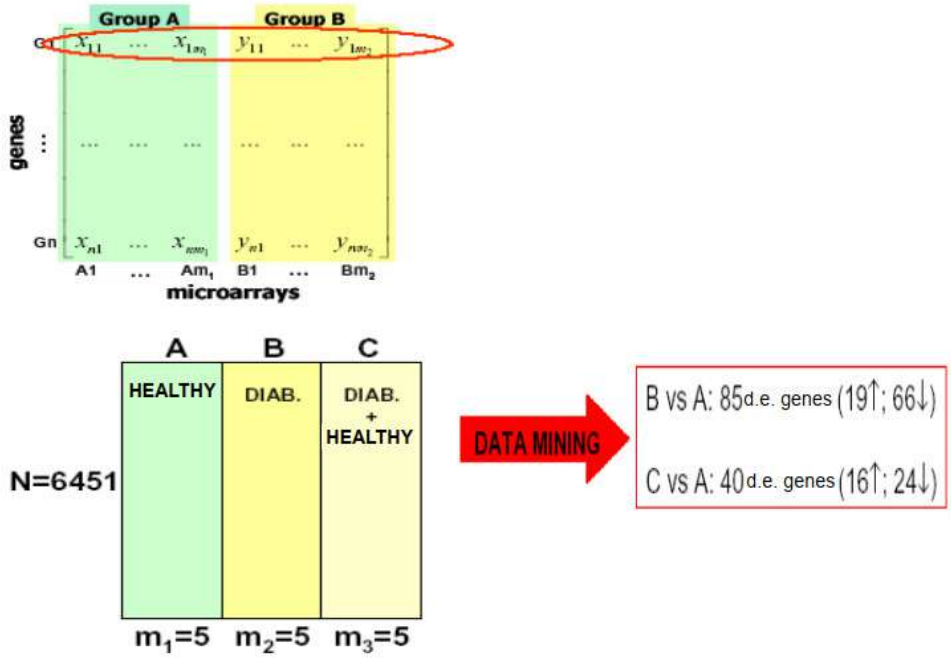
- Temporal series of genetic expressions during a perturbation
- Example 3: Selection of differentially expressed genes in time, clustering

IV.B.4.1 "Static" experiments with microarrays

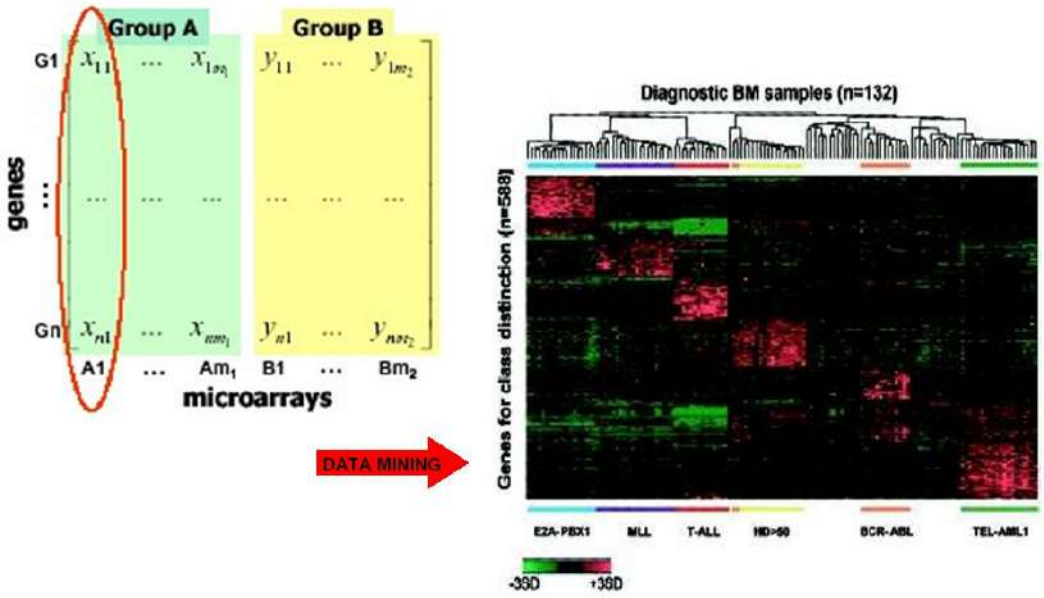
Two or more *subject classes* (different phenotypes / treatments)



Example 1: Selection of genes with *different expressions* (d.e.)



Example 2: *Classification* (supervised)

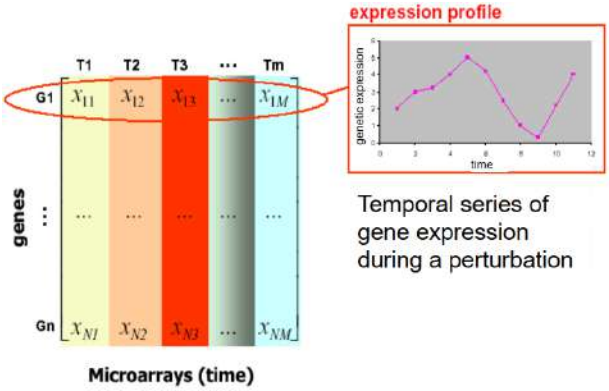


6 groups (types) of childhood acute lymphoblastic leukaemia (ALL)

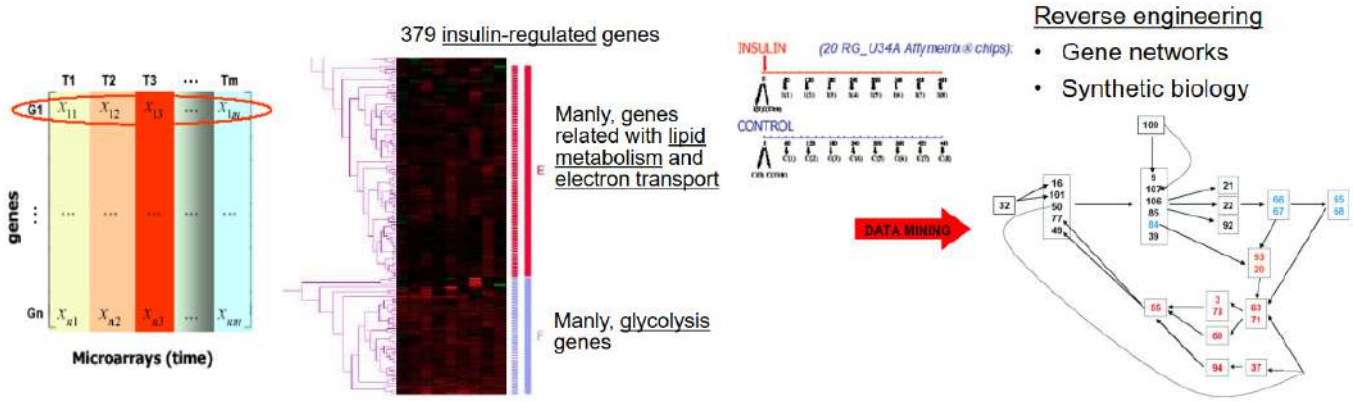
$$M = m_1 + m_2 = 132, \qquad N = 26'000$$

IV.B.4.2 "Dynamic" experiments with microarrays

Same subject at *different times*:



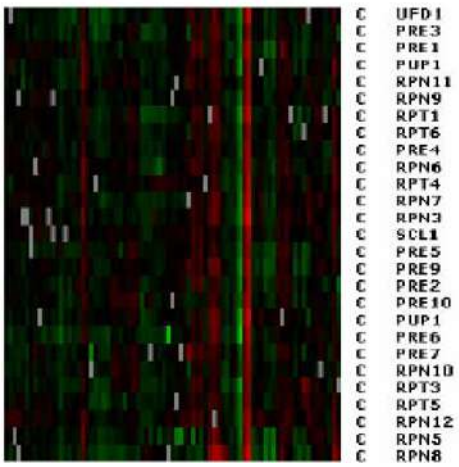
Example 3: Selection of differentially expressed genes in time



Reverse engineering: computations that we perform to understand the biological process that occur at a cellular level:

- The *gene networks analysis* refers to the study of the interactions between the proteins encoded by the gene interact and the DNA on places that regulate the activity of the gene, to build a gene regulatory network, that explain the biological process occurring within a cell (see further lessons).
- *Synthetic biology* refers to the field of engineering that focuses on mimicking the biological process that occur in the cell, to study these processes or to be used in industrial processes (use of some materials, or in the agricultural engineering)

IV.B.5 Machine learning basics (2 Nov.)



Input data: a (typically small) set of microarray experiments (*samples*) with, usually, many *variables* (genes)

For each gene, a **feature vector** is formed by combining its *normalized expression values* in the available samples

- Each row represents the feature vector associated with a gene
- Each column represents an experiment
- Typically, we work in a *small sample* set scenario: much more genes than experiments (usually not good for statistical analysis)

Unsupervised learning (clustering / class discovery)

- Feature vectors are *unlabelled*
- Goal: attach to each gene a *cluster label* by grouping together genes that exhibit *similar expression* behaviour

Supervised learning (classification / class prediction)

- Feature vectors are labelled (e.g. before / after treatment)
- Goal: predict the label of a *new unlabelled* sample
- When the class variable has *continue* instead of discrete values, regression analysis is used instead of classification (not covered in this class)

IV.B.5.1 Definitions

Features: *variables* or attributes of the samples that are *used to cluster* or classify genes

Distance: method used to decide whether two samples are *similar* or not

Model: *how* clustering or classifying, e.g.:

- Hierarchical clustering
- k-means
- k-Nearest Neighbors (k-NN)
- Support Vector Machines (SVM),
- Neural Nets
- …

IV.B.5.2 Distances

All (every!) machine learning tools rely on some *measure of* ***distance*** between samples. You **must be aware** of the distance function being used

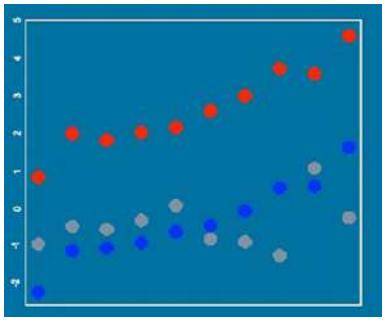
- There are very many *different distances* (e.g., Euclidean, Correlation [used in biological data most of the time], Manhattan, …)
- The *choice of distance* is important and in general substantially affects the outcome
- The choice of distance should be made *carefully*

Metric distances: a distance measure d_{ij} between two vectors, i and j, must obey several rules:

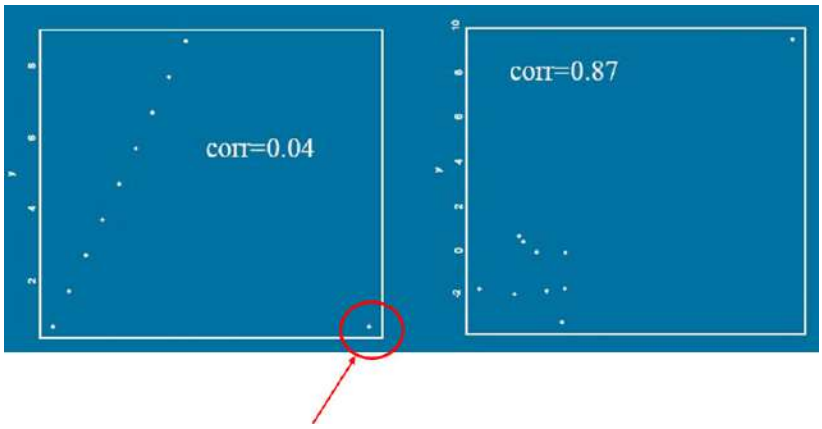
- The distance must be *positive* and *definite*, $d_{ij} \geq 0$ (i.e. it must be zero or positive). An object is zero distance from itself, $d_{ii} = 0$
- The distance must be *symmetric*, $d_{ij} = d_{ji}$ (i.e. the distance from i to j is the same as the distance from j to i)
- *'Triangle' inequality*: when considering three objects, i, j and k, the distance from i to k is always less than or equal to the sum of the distance from i to j, and the distance from j to k (i.e., $d_{ik} \leq d_{ij} + d_{jk}$)

Distance matrix: distances can be represented as *matrices*, where the value in row i and column j is the distance between sample i and sample j (or between genes i and j). These matrices are called distance matrices and they are symmetric

It is not simple to select the distance function. Let's consider the following set of points:



Distance	**Euclidian**	**Correlation**
Expression	$d_{ij}^E = \sqrt{\sum_{k=1}^{N}(x_i^k - x_j^k)^2}$	$d_{ij}^C = \sum_{k=1}^{N} \frac{(x_i^k - \mu_i)(x_j^k - \mu_j)}{\sqrt{\sum_{k=1}^{N}(x_i^k - \mu_i)^2}\sqrt{\sum_{k=1}^{N}(x_j^k - \mu_j)^2}}$
Red-blue distance	9.45	0.006
Red-grey distance	10.26	0.768
Blue-grey distance	3.29	0.7101



Correlation measures linear association and is <u>not resistant</u> (one outlier can ruin it)

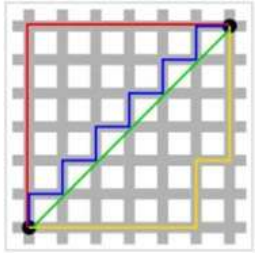
Manhattan distance: the distance between two points as the sum of the (absolute) differences of their coordinates:

$$d_{ij}^M = \sum_{k=1}^{N} |x_i^k - x_j^k|$$

Also known as *rectilinear* distance, L_1 distance or l_1 norm, or *city block* distance, since it is induced by the p-norm distance, or *Minkowski distance*, when $p = 1$:

$$d_{ij}^{Mi} = \left(\sum_{k=1}^{N} |x_i^k - x_j^k|^p\right)^{\frac{1}{p}}$$

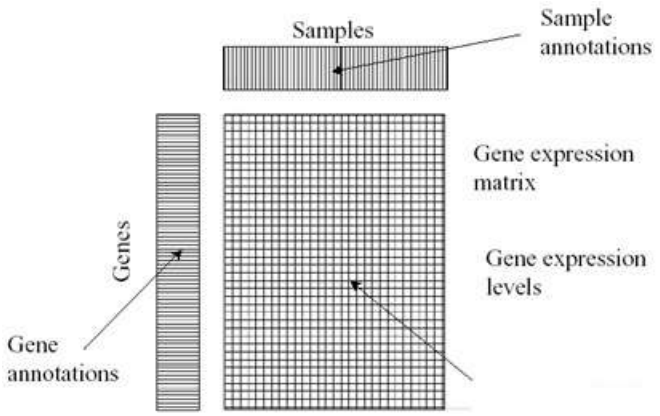
It is called Manhattan distance because it is the distance along the *square path* connecting two points, i.e. the only possible in Manhattan (where it is not possible to go in diagonal through a skyscraper block!)



IV.B.6 Unsupervised learning

Unsupervised learning is known as **clustering**, or class discovery in some cases

- The basic idea is to determine *how many groups* are in the data, and *which variables* seem to define the groupings
- Clustering algorithms are methods to *divide* a set of n observations into g groups so that *within* group similarities are larger than *between* group similarities
- The *inputs* are typically the *feature vectors* of the data elements (genes), i.e., the rows of the gene-experiment matrix



- The number of groups, g, is generally *unknown* and must be *selected* in some way
- Implicitly both *features* and a *distance* must have been already selected (there are interactions between the distance being used and the clustering method)
- There is *no training sample* (and the groups are unknown before the process begins)
- Unlike classification (supervised learning) there is *no easy way* to use cross-validation

IV.B.6.1 Hierarchical clustering

There are two types of **hierarchical clustering**:

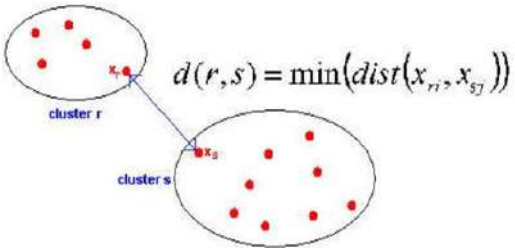
- **Agglomerative**: generates a *hierarchy of clusters* going from n clusters of 1 element each, to 1 cluster of n elements
- **Divisive**: divides the data into g groups using some (re)allocation algorithm (not covered here)

In both types, before it must be defined:

- Distance between *feature* vectors (see previous slides)
- Distance between *groups* of feature vectors (clusters):
  - Single linkage
  - Complete linkage
  - Average linkage

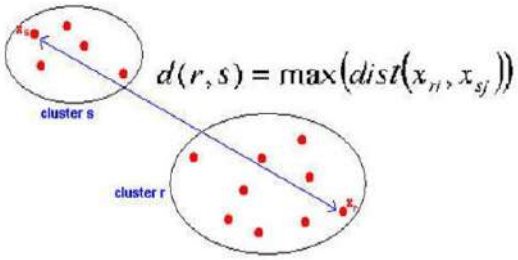
Distance between *groups* of feature vectors:

Single linkage: distance between two clusters is the *smallest distance* between an element of the first cluster and an element of the second cluster:



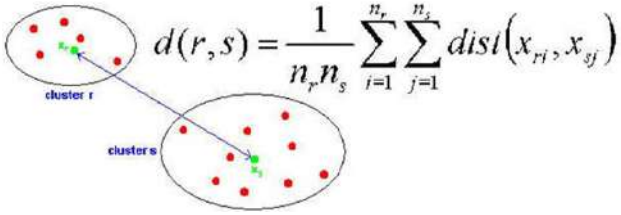
- *Chaining* issue: tends to force clusters together due to single entities being close to each other regardless of the positions of other entities in each cluster

Complete linkage: distance between two clusters is the *maximum distance* between an element of the first cluster and an element of the second cluster



- Not to be used if lot of *noise* is expected in the dataset
- This method also produces *very compact clusters*; useful if entities of the same cluster are expected to be *far apart in multi-dimensional space* (provided there is no noise), i.e., *outliers* have more weight in cluster definition

Average linkage: distance between the two clusters is the *average of all pairwise distances*



- More *computationally expensive* than the other methods
- It is halfway between single and complete linkage. Several variations of this method exist
- The *chaining issue is not observed,* and outliers are not given any special favours in the cluster definition; This makes it the most popular method of the three
- It is also referred to as UPGMA (Unweighted Pair-Group Method using Arithmetic averages)

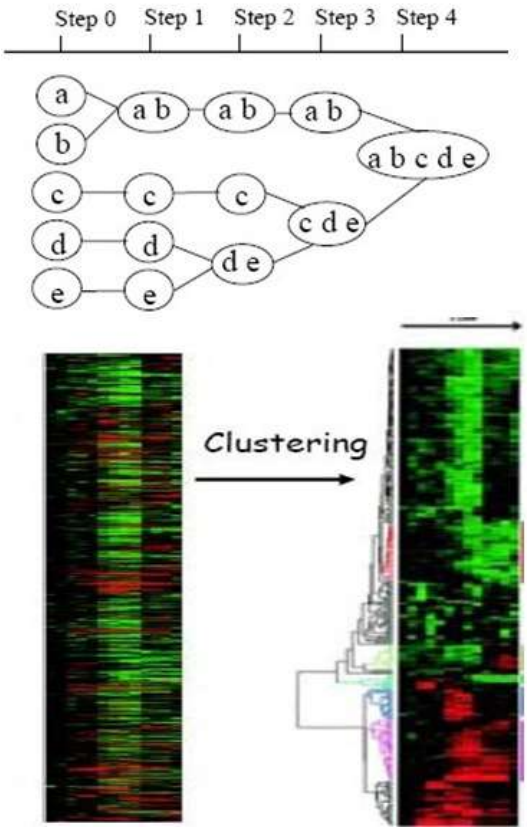
IV.B.6.2 Agglomerative hierarchical learning

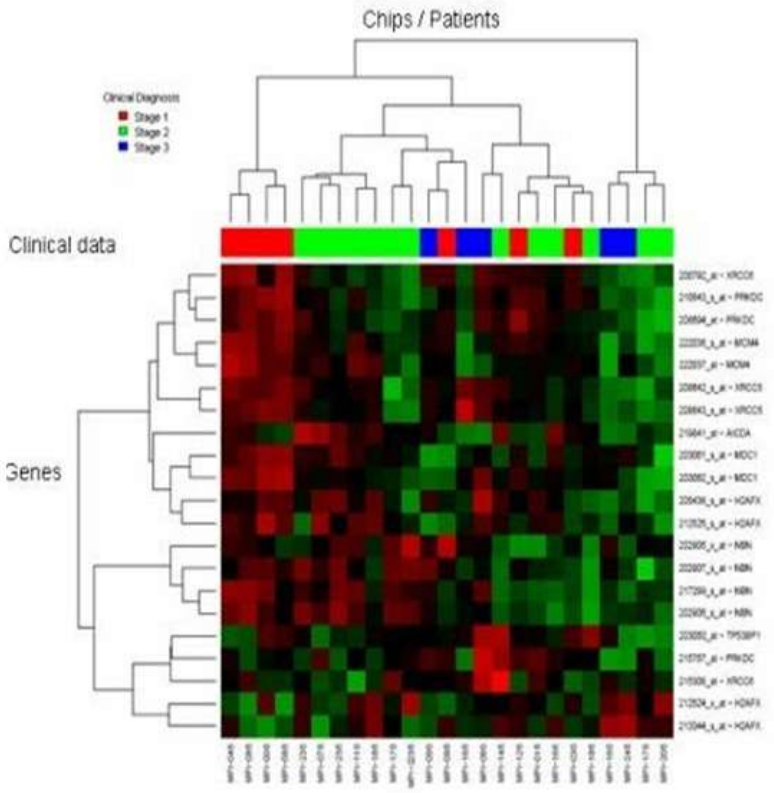
Input: one feature vector for each gene

1. *Initialization*: each cluster consists of a gene
2. Compute the distance between *each pair of clusters*
3. *Merge* the two clusters with the *smallest inter-cluster distance*
4. Go to step 2, until all genes are contained within *one big cluster*

Output: **dendrogram**:

- A tree structure with the *genes at the bottom* (the leaves)
- The *height* of the joins indicates the *distance* between the left branch and the right branch
- A *threshold* on the dendrogram levels defines the clusters





IV.B.6.3 Partitioning methods

Agglomerative clustering partitioning methods:

- *k-means*
- PAM (Partitioning Around Medoids) [used in biology]
- SOM (Self-Organizing Maps) [used in biology]

It must be first *predefined the number of clusters K* (and their centres) after which the algorithm partitions the data iteratively until a solution is found

IV.B.6.4 *k*-means clustering

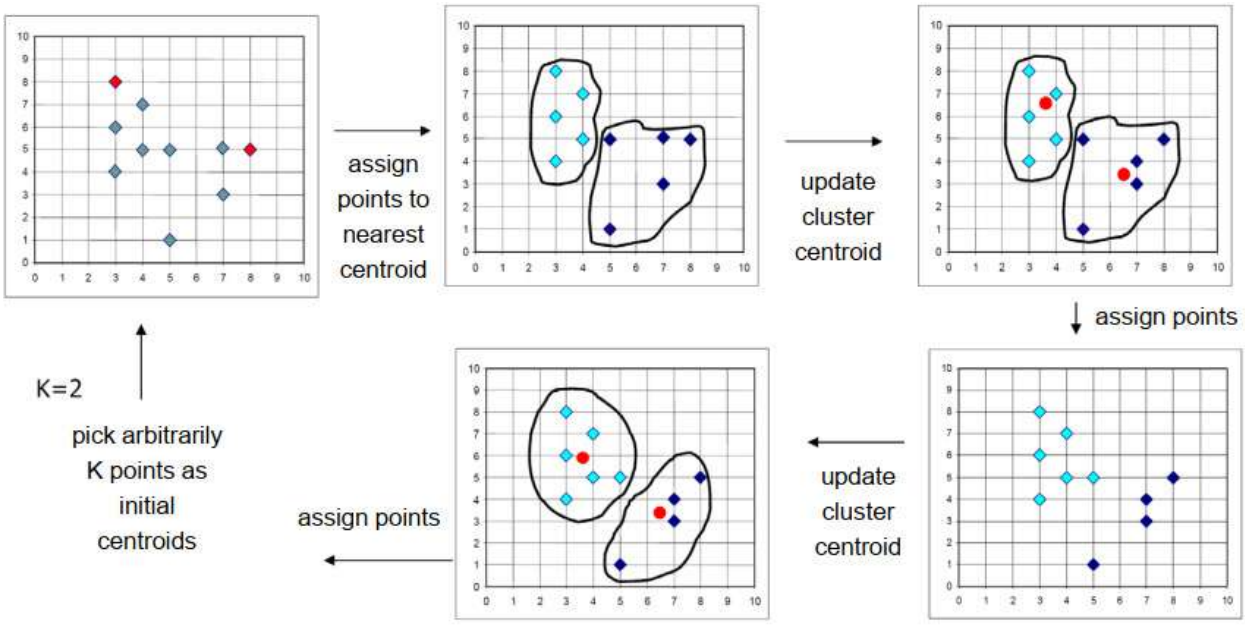
1. Initialisation:
   - Define the *number* of clusters k
   - Designate a cluster *centre* (a vector quantity that is of the same dimensionality of the data) for *each* cluster
2. Assign each data point to the *closest cluster centre*; the data point is now a member of the cluster
3. Calculate the *new cluster centre* (the geometric average of all the members of the cluster)

4. Calculate the *sum* of *within-cluster sum-of-squares* of distances of cluster elements from cluster centroid:
   - If this value has *not significantly changed* over a certain number of *iterations*, exit the algorithm
   - If it has changed, or the change is insignificant but it has not been seen to persist over a certain number of iterations, go back to step 2

A common problem in k-means partitioning: if the initial partitions are not chosen carefully enough the computation has the chance of converging to a ***local minimum***, rather than to the *global minimum* solution.

- The *initialisation* step is therefore *very* important
- To combat this problem, it might be a good idea to run the algorithm *several times* with **different initialisations**
  - If results converge to the *same partition*, it is likely that a global minimum has been reached
  - Drawback: *computationally expensive* and very *time consuming*

Another way to combat it: *dynamically* change the *number of partitions* (clusters) as the iterations progress



There is a growing interest in using **fuzzy logic** in clustering algorithms. Fuzzy logic allows the algorithm to accept the possibility that a *single data point* (i.e., gene) can *belong to more than one cluster*. By defining a *vector of memberships* for each data element (gene)

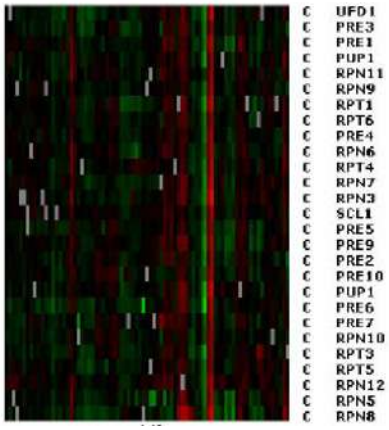
IV.B.6.5 Singular Value Decomposition

The **Singular Value Decomposition** (SVD), sometimes called **Principal Component Analysis** (PCA), can be used to:

- *Reduce the dimensionality* of the data to summarise the most important components whilst simultaneously filtering out noise → A conventional clustering algorithm (e.g., k-means) is applied afterwards on reduced feature vectors
- Perform *clustering* directly

Consider the matrix $A \in \mathbb{R}^{m \times n}$ where:

- m is the number of genes
- n is number of experiments (samples)
- A_{ij} is the expression level of the i-th gene in the j-th sample



The SVD of A is defined $A = U\Sigma V^T$ where:

- U is a $m \times r$ *orthogonal* matrix (i.e. $U^T U = I$; I: identity matrix)
- V is a $n \times r$ *orthogonal* matrix (i.e. $V^T V = I$; I: identity matrix)
- Σ is a $r \times r$ *diagonal* matrix
- r is the *rank* of A (i.e., the number of linearly independent columns)

Linear algebra parenthesis:

- Linear algebra studies *linear transformations*, which are represented by *matrices* acting on *vectors*
- A matrix can act on a vector by changing both its *magnitude* and its *direction*. On certain vectors, the **eigenvectors**, a matrix acts only by multiplying their magnitude by a *factor* (the **eigenvalue** associated with that eigenvector), which is positive if the vector direction is unchanged or negative if the vector direction is reversed
- An **eigenspace** is the set of all eigenvectors that have the same eigenvalue, together with the zero vector
- Eigenvalues, eigenvectors and eigenspaces are *properties* of a matrix. They give important information about the matrix and can be used in matrix *factorization* (decomposition)

The SVD of A is related to the eigenvalue/eigenvector decomposition:

- U is a set of eigenvectors of the matrix AA^T
- V is a set of eigenvectors of the matrix $A^T A$
- The diagonal elements of Σ are (in non-increasing order) the square roots of the eigenvalues of AA^T (or $A^T A$)

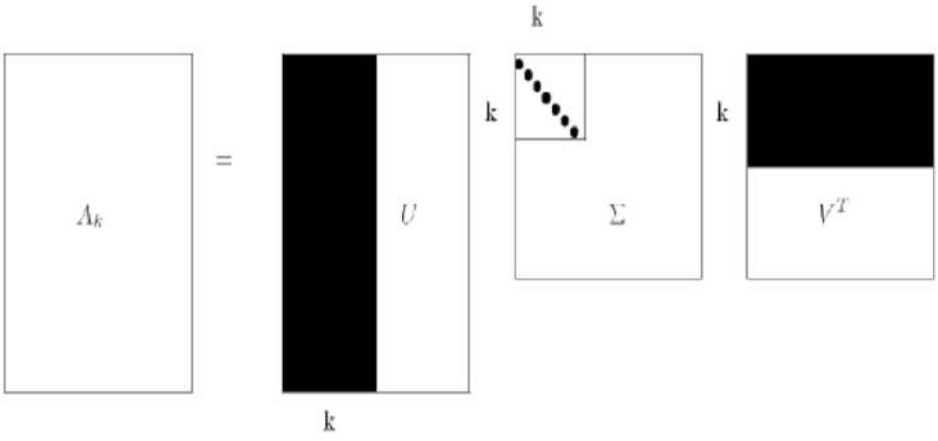
Let V_k denote the *submatrix* obtained with the first $k < r$ columns of V (e.g. the eigenvectors of $A^T A$ associated with the k largest eigenvalues). The new k-dimensional features are obtained by:

$$A_k = AV_k V_k^T \quad \text{or} \quad A_k = U_k \Sigma_k V_k^T$$

with $A_k \in \mathbb{R}^{m \times n}$

The A_k is the same dimension as A, but only takes into account the k first components

$$A = U\Sigma V^T \qquad A_k = U_k \Sigma_k V_k^T$$



Clustering based on SVD; two options:

- Use A_k as input for a clustering algorithm (e.g., hierarchical clustering, k-means, ...)
- Decide the number of clusters k and extract the first k columns of the matrix U:
  - Each column of U represents a cluster
  - Each real-valued entry i of column j of U represents the membership of gene i to cluster j
  - Unlike k-means, genes are associated with *multiple clusters* (with varying membership degrees, $\rightarrow$ fuzzy clustering)

IV.B.7 Supervised learning

Supervised learning is known as *classification* (or *class prediction*)

The basic idea is to be able to *predict the class label* of an input sample (**test sample**) given the prior knowledge of a set of *labelled samples* (**training samples**)

There are several techniques used in supervised learning:

- Linear classifiers
- k-NN (k-Nearest Neighbors)
- SVM (Support Vector Machines)
- ANN (Artificial Neural Networks)
- ...

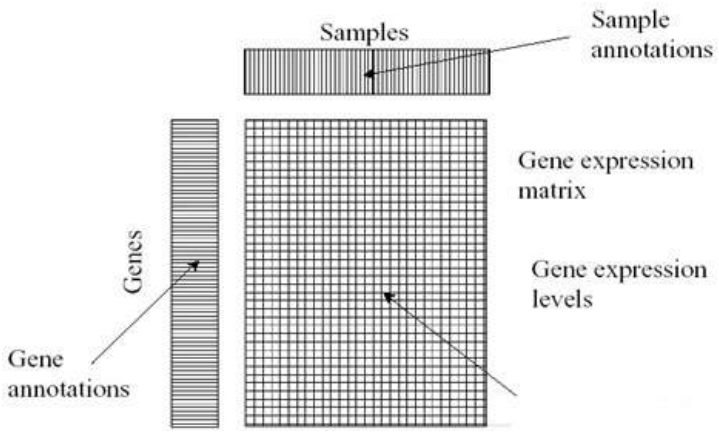
Unlike unsupervised learning, the input feature vectors are typically the columns of the gene-experiment matrix

The **dimensionality** (i.e., the *number of features* (genes) in each vector) is often *huge*, since it depends on the number of genes assayed in a single microarray experiment.

- We may want to reduce de dimensionality using the algorithms mentioned before (e.g., PCA).
- However, we need to consider that the variable obtained after this decomposition are different from the original values, and this does not facilitate the interpretation.
- Instead, we apply some feature selection algorithm that do not combine together the original genes but keep them separate, and only consider the one that have a significant contribution to the sample in the classes they belong to

The **sample size** (i.e., the *number of feature vectors* (microarray experiments) is typically *small*

Many classification algorithms may benefit from pre-processing (e.g. SVD, i.e. PCA) to reduce the dimensionality



IV.B.7.1 *k*-nearest neighbours

***k*-Nearest Neighbours** classification technique:

1. Initialization
   - Define k
   - Define a *distance* metric
   - Consider a set of *labelled training samples*
2. Given a test sample:
   - *Compute the distance* between the *test* sample and all *training* samples
   - Retain the *top k training samples* sorted based on the distance from the test sample (k-Nearest Neighbours)
   - Each neighbour votes for its label: assign to the test sample the label that receives more votes

Characteristics:

- *Non-parametric* (i.e., fitting of sample population to any parametrized distributions is not required)
- *Time-consuming* (distances need to be re-computed for each test sample)

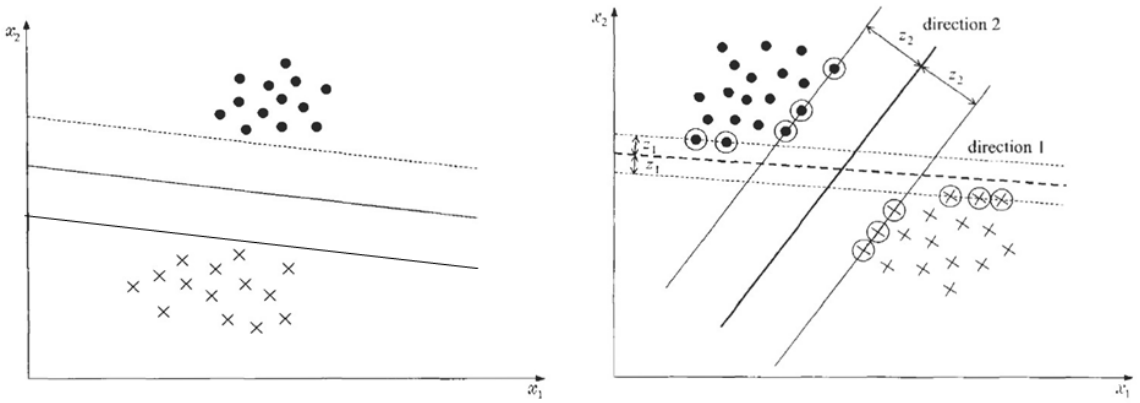
IV.B.7.2 Support Vector Machines

There are two types of **Support Vector Machines** (SVM):

- **Linear SVM**:
  - Work for *linearly* separable samples (two classes)
  - Receive in input the *original* feature vectors
  - Find the *optimal hyperplane* that separates the samples of the two classes
- **Non-linear SVM**:
  - A *non-linear mapping* (kernel function) is applied from the original feature space (called attribute space) to a *higher dimensional target* feature space (that can separate the data better)
  - Linear SVM is applied in the *target* feature space
  - Can often classify non-linearly separable samples

IV.B.7.3 Linear Support Vector Machines

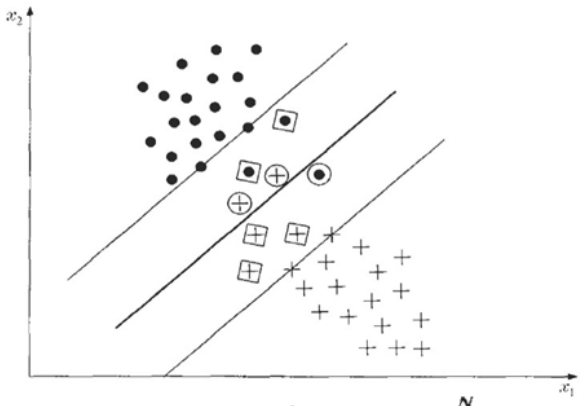
For *linearly separable samples*, the linear SVM finds the *hyper-plane* that *maximize the margin* (i.e., the distance of the decision boundaries (hyper-plane) from each sample)



$$\text{minimize} \quad J(\boldsymbol{w}) \equiv \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1, \quad i = 1, 2, \ldots, N$$

For *non-linearly separable samples*, in linear SVM a *loss factor* is added to the cost function to account for misclassified samples



$$\text{minimize} \quad J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad y_i[\boldsymbol{w}^T\boldsymbol{x}_i + w_0] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N$$
$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N$$

From the original (*primal*) problem:

$$\text{minimize} \quad J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad y_i[\boldsymbol{w}^T\boldsymbol{x}_i + w_0] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N$$
$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N$$

We can write the *dual problem*:

$$\text{Maximise} \sum_{k=1}^{R}\alpha_k - \frac{1}{2}\sum_{k=1}^{R}\sum_{l=1}^{R}\alpha_k\alpha_l Q_{kl}\,, \qquad \text{where } Q_{kl} = y_k y_l(\boldsymbol{x_k}\cdot\boldsymbol{x_l})$$

$$\text{Subject to the constraint: } \forall k, \qquad 0 \leq \alpha_k \leq C, \qquad \sum_{k=1}^{R}\alpha_k y_k = 0$$

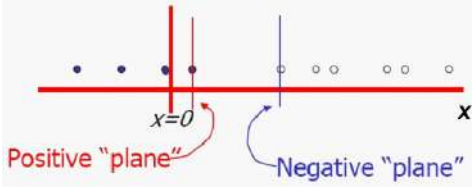
The data vectors enter the problem only in the inner product $x_k x_l$

Why is it useful to solve the dual instead of the primal problem?

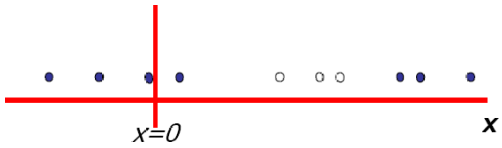
- Dual problem is a **QP** (quadratic programming) [QP is a special type of mathematical optimization problem; it is the problem of optimizing (minimizing or maximizing) a quadratic function of several variables subject to linear constraints on these variables]
  → there are several optimized algorithms to quickly solve QPs
- Using the "***kernel function*** *trick*" (following illustrated), we are able to separate non-linearly separable data

IV.B.7.4 Non-linear Support Vector Machines

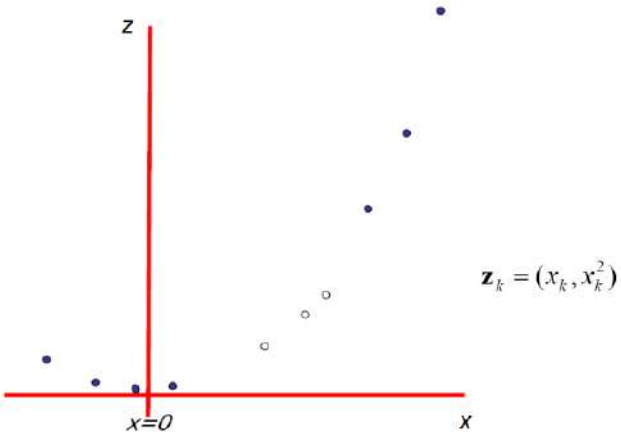
Kernel function trick: Suppose to be in a *1D space*:



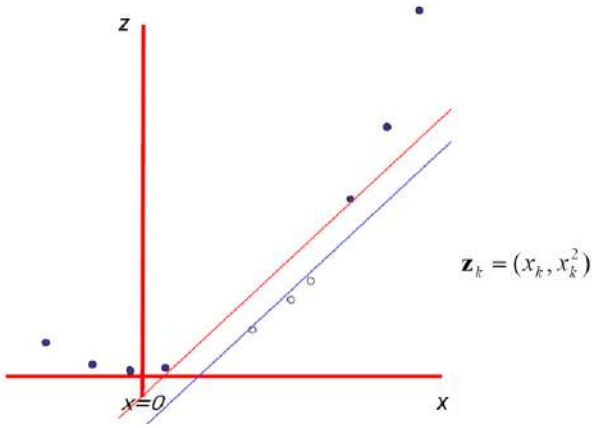
... but data might *not* be linearly separable:



However, there could be a trick ... go *from 1D to 2D*:



Now data is *linearly* separable in the *2D space*:



Common SVM kernel functions (z_k):

- z_k = (polynomial terms of $\boldsymbol{x_k}$ of degree 1 to q) : $z_k[j] = \varphi_j(\boldsymbol{x_k}) = (\boldsymbol{x_k}, \boldsymbol{x_k^2}, \ldots, \boldsymbol{x_k^q})$
- z_k = (radial basis functions of $\boldsymbol{x_k}$) : $z_k[j] = \varphi_j(\boldsymbol{x_k}) = KernelFn\left(\frac{|\boldsymbol{x_k} - \boldsymbol{c_j}|}{KW}\right)$
- z_k = (sigmoid functions of $\boldsymbol{x_k}$)

Modified dual problem with kernel function:

$$\text{Maximise } \sum_{k=1}^{R} \alpha_k - \frac{1}{2} \sum_{k=1}^{R} \sum_{l=1}^{R} \alpha_k \alpha_l Q_{kl}\,, \qquad \text{where } Q_{kl} = y_k y_l \big(\varphi(\boldsymbol{x_k}) \cdot \varphi(\boldsymbol{x_l})\big)$$

$$\text{Subject to the constraint: } \forall k, \qquad 0 \leq \alpha_k \leq C, \qquad \sum_{k=1}^{R} \alpha_k y_k = 0$$

$\frac{R^2}{2}$ *dot products* must be done to get the matrix ready. Each dot product requires $\frac{m^2}{2}$ additions and multiplications where:

- R is the *number of feature vectors* (usually low)
- m is the *dimension* of the target space (usually high)

The whole evaluation seems to cost $\frac{R^2m^2}{4}$ operations ... But, if a proper kernel function φ is used, it can be computed in $\frac{R^2m}{2}$ operations

Example: Polynomial kernel of degree m

$$\mathbf{\Phi(a)} \bullet \mathbf{\Phi(b)} = \begin{pmatrix} 1 \\ \sqrt{2}a_1 \\ \sqrt{2}a_2 \\ \vdots \\ \sqrt{2}a_m \\ a_1^2 \\ a_2^2 \\ \vdots \\ a_m^2 \\ \sqrt{2}a_1a_2 \\ \sqrt{2}a_1a_3 \\ \vdots \\ \sqrt{2}a_1a_m \\ \sqrt{2}a_2a_3 \\ \vdots \\ \sqrt{2}a_1a_m \\ \vdots \\ \sqrt{2}a_{m-1}a_m \end{pmatrix} \bullet \begin{pmatrix} 1 \\ \sqrt{2}b_1 \\ \sqrt{2}b_2 \\ \vdots \\ \sqrt{2}b_m \\ b_1^2 \\ b_2^2 \\ \vdots \\ b_m^2 \\ \sqrt{2}b_1b_2 \\ \sqrt{2}b_1b_3 \\ \vdots \\ \sqrt{2}b_1b_m \\ \sqrt{2}b_2b_3 \\ \vdots \\ \sqrt{2}b_1b_m \\ \vdots \\ \sqrt{2}b_{m-1}b_m \end{pmatrix} \begin{matrix} 1 \\ + \\ \sum_{i=1}^{m} 2a_ib_i \\ + \\ \sum_{i=1}^{m} a_i^2b_i^2 \\ + \\ \sum_{i=1}^{m}\sum_{j=i+1}^{m} 2a_ia_jb_ib_j \end{matrix}$$

Computing the inner product cost $O(m^2)$ … but it is the same as:

$$\mathbf{\Phi(a)} \bullet \mathbf{\Phi(b)} = 1 + 2\sum_{i=1}^{m} a_ib_i + \sum_{i=1}^{m} a_i^2b_i^2 + \sum_{i=1}^{m}\sum_{j=i+1}^{m} 2a_ia_jb_ib_j$$

$$\begin{aligned} &(\mathbf{a}.\mathbf{b}+1)^2 \\ &= (\mathbf{a}.\mathbf{b})^2 + 2\mathbf{a}.\mathbf{b} + 1 \\ &= \left(\sum_{i=1}^{m} a_ib_i\right)^2 + 2\sum_{i=1}^{m} a_ib_i + 1 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} a_ib_ia_jb_j + 2\sum_{i=1}^{m} a_ib_i + 1 \\ &= \sum_{i=1}^{m} (a_ib_i)^2 + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} a_ib_ia_jb_j + 2\sum_{i=1}^{m} a_ib_i + 1 \end{aligned}$$

Cost: $O(m^2)$ *vs.* $O(m)$

IV.B.8 References

Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002 Dec;32 Suppl:496-501.

Quackenbush J. Computational genetics: Computational analysis of microarray data. N*ature Reviews Genetics* 2001 June;2:418-427.

Smyth, GK, Speed TP. Normalization of cDNA microarray data. *Methods* 2003;31:265-273.

Theodoridis S, Koutroumbas K. *Pattern recognition*. Academic Press, San Diego, CA. 2006.

Moore A. *Statistical data mining tutorials*. http://www.autonlab.org/tutorials/

V. Introduction to biological networks

V.A Why networks in biology?

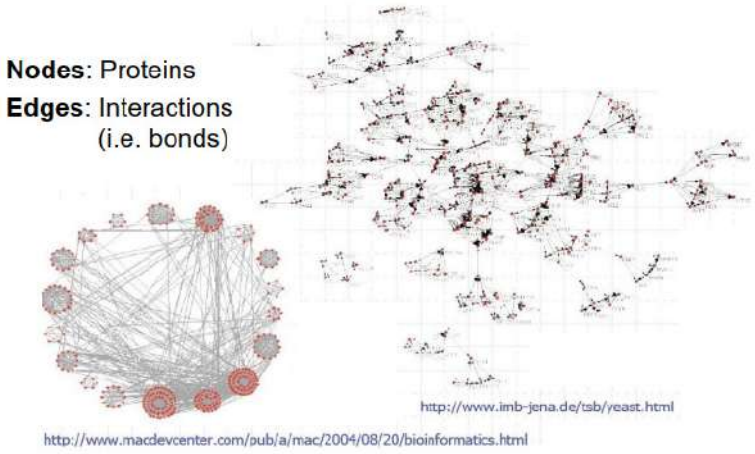
Biology studies complex systems such as **cells**. The properties of complex systems **do not result from single components alone, but mainly from their interactions**.

For example, *transcription* from DNA to RNA is not explained only by the information in a DNA fragment alone, but in relation with *other molecules* (e.g., transcription factors) that interact with DNA to start/stop/regulate the transcription.

Models from complex networks theory provide useful representations of complex biological systems and allow to understand them through their properties.

All *biological processes* can be modelled as networks since they occur thanks to interactions among molecules (system biology):

- *Proteins* interact with each other, or with other molecules (e.g., DNA), generating and regulating several mechanisms (protein networks)
- *Genes* encode for proteins; their interrelated activity determine protein abundance and related processes (gene networks).



V.B Types of biological networks

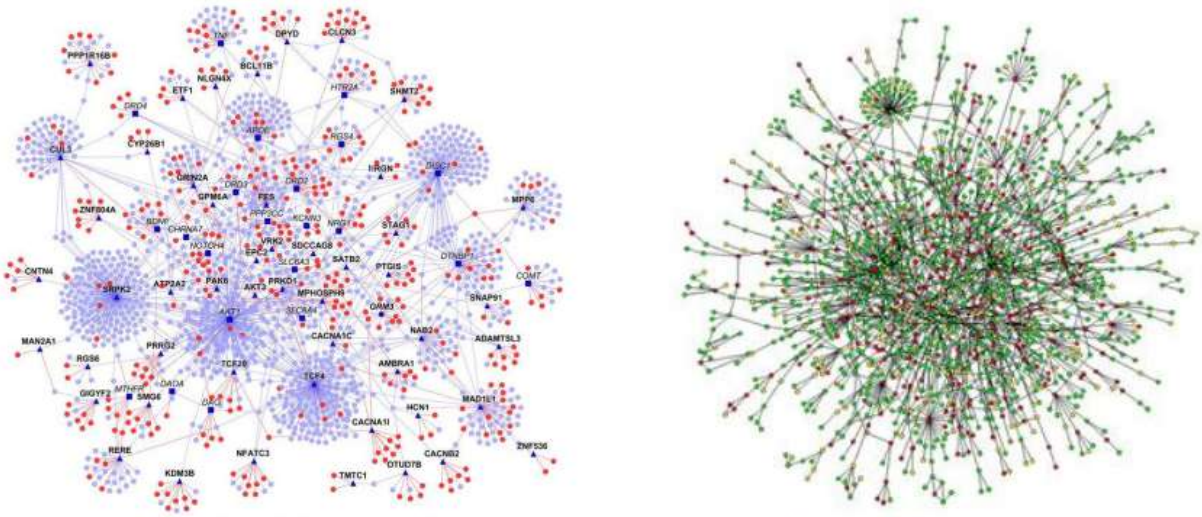
In biology there are different types of networks that can be considered, some of them are:

- **Protein** networks → interactions between proteins
- **Gene regulatory** networks → interactions between molecular regulators within the cell
- **Gene co-expression** networks → associations between variables that measure the abundance of transcripts
- **Metabolic** networks → biochemical reactions in a living cell
- **Signalling** networks → signals between cells or within cells
- **Neuronal** (not neural!) networks → connections between neurons in the brain.

V.B.1 Example of protein networks

Since protein-protein interactions are essential to almost every process in a cell, understanding them is crucial.

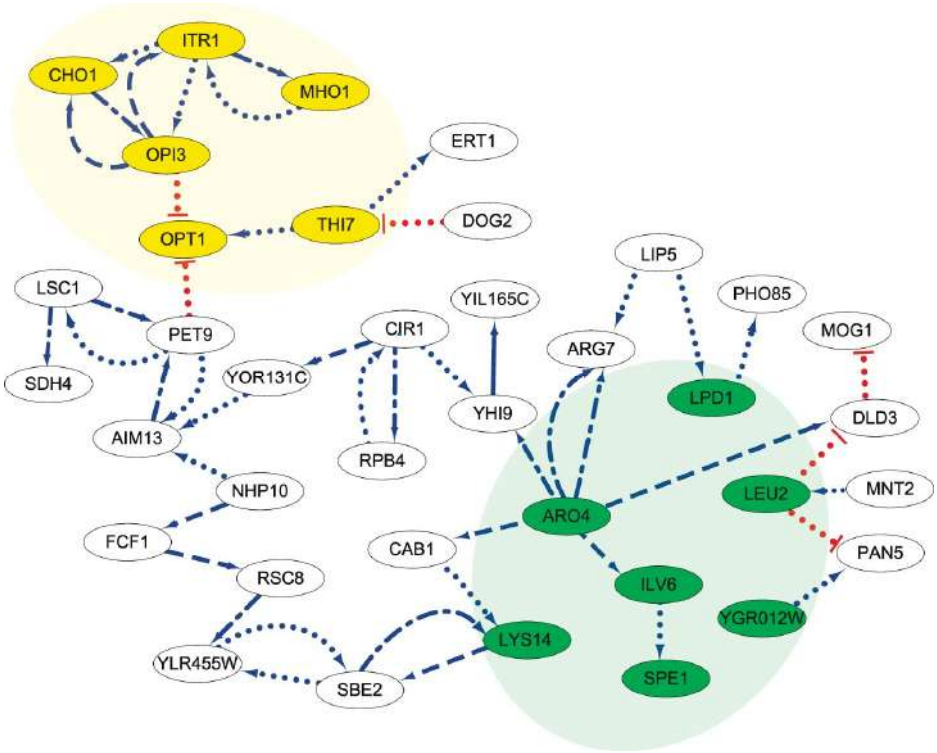
A **protein-protein interaction network** is a mathematical representation of the physical contacts between proteins in the cell.



V.B.2 Example of gene regulatory networks

Gene regulatory networks represent mechanisms between molecular regulators that interact with each other and with other substances in the cell. The molecular regulators can be DNA, RNA, proteins.

Usually, these networks are *directed* (from a source node to a target node). Different kind of edges may exist, to represent different kind of interactions.

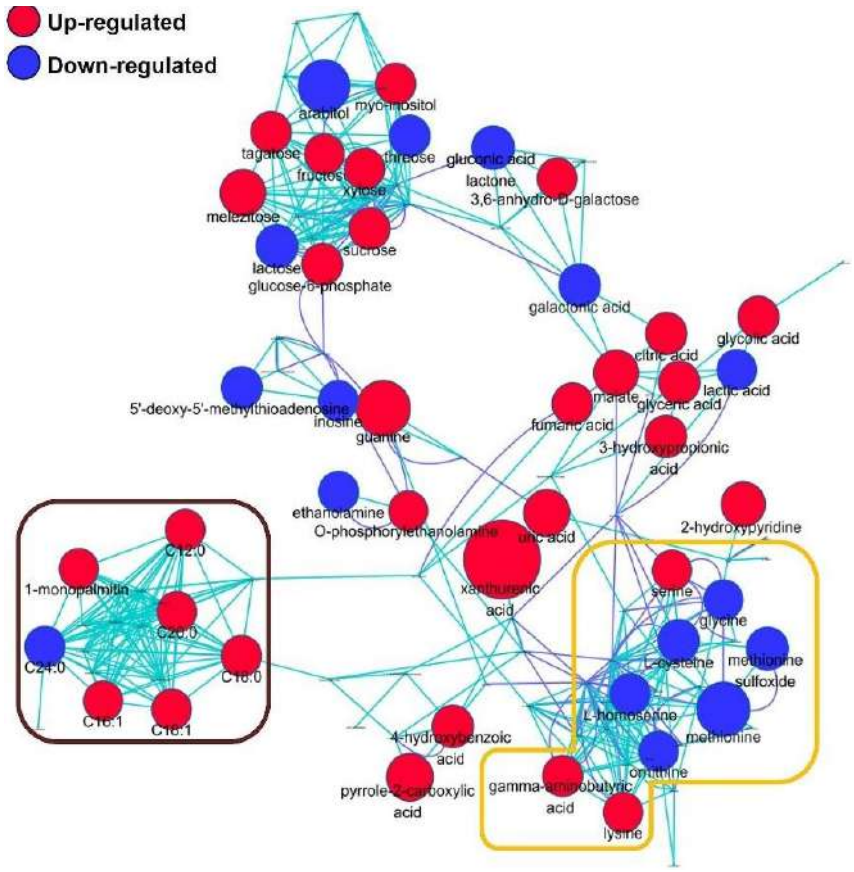


The largest gene regulatory subnetworks in yeast: https://www.nature.com/articles/s41598-018-37667-4#Fig1

V.B.3 Example of metabolic networks

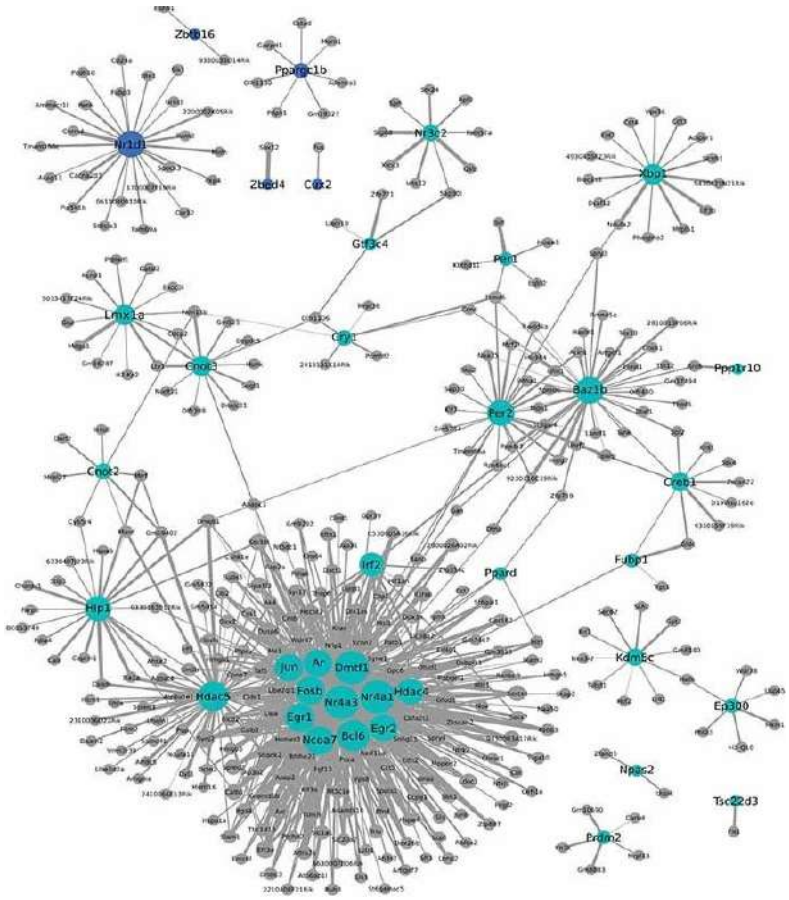
Metabolic networks describe the relationships between small biomolecules and proteins or enzymes that produce biochemical reactions.

Usually in these types of networks, the small biomolecules constitute the nodes, whereas the proteins or enzymes are the edges that allow the reaction.



V.B.4 Example of gene co-expression networks

Gene co-expression networks represent the interconnectedness between genes. The nodes are the *genes*, the edges are the *co-expression relationships* between them.



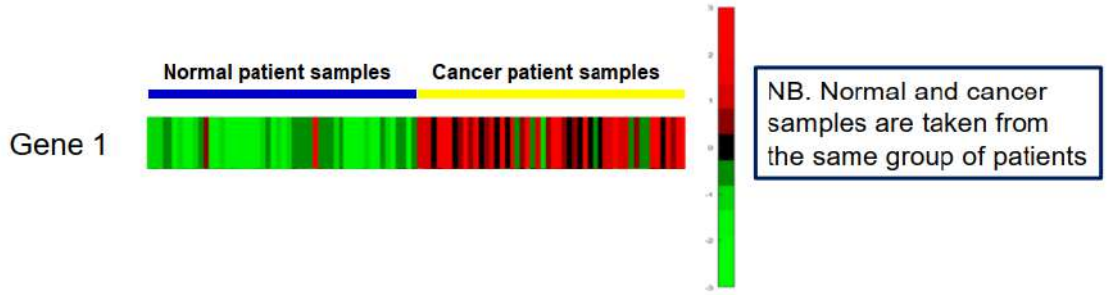
Gene Coexpression Network of transcription factors and genes in postpartum NAC (nucleus accumbens in the brain)

V.C Gene co-expression networks from gene expression data

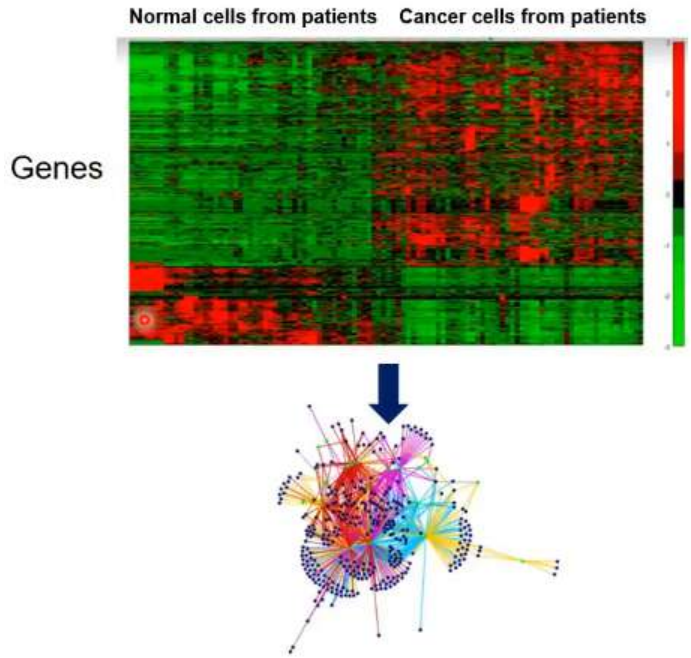
Gene expression data can be seen as scores associated with each gene of each sample according to the gene expression value.

The **profile of expression** values of a gene *across considered samples* can be visualized in a heatmap (as shown below), where columns represent samples and colors represent *gene expressions*.

In the example heatmap below, the goal is to show the effect of a cancer disease; the **fold change** is used to visualize differential expression.



We can compute a **similarity measure** among *gene expression profiles* and **visualize** the results as a **gene network**. We identify genes with similar expression profiles, e.g., involved in the same process, i.e., related to each other.



From the similarity matrices that we obtain by comparing one gene expression with the others, we can obtain the gene co-expression network. The easiest way to do it is computing all the similarities and fix a threshold.

V.D Similarity measures

Most typical similarity measures:

- **Pearson's correlation** (the most used one): measures the correspondence of two vectors (here genes). It has the benefit of being *scalable*, i.e., it can be efficiently computed for *large numbers of genes*. However, it is *sensitive to outliers*, and it assumes that the gene expression data follow a *normal* distribution.
- **Euclidean distance** (for genes that are similar in their expression values and wide expressed): measures the *geometric distance* between two vectors (here genes). It is not appropriate when the absolute expression levels of functionally related genes are highly different. Furthermore, if two genes have consistently low expression levels but are otherwise randomly correlated, they might still appear close in the Euclidean space.

- **Mutual information**: measures how much the information of a gene reduces the uncertainty about the expression levels of another. It can *detect non-linear relationships*; however, sophisticated non-linear relationships may *not be biologically meaningful*. In addition, the distribution of the data is needed for the computation of the mutual information, and it needs many samples for a good estimate.
- **Spearman's rank correlation:** it is the Pearson's correlation calculated for the ranks of gene expression values in a gene expression vector. It is more *robust to outliers*, but it is less sensitive to expression values and with small number of samples it may *detect many false positives*.

V.E Biological complex networks

Biological networks are complex networks. Any descriptive property of complex networks can be applied

A biological complex network can be analysed in multiple different ways through:

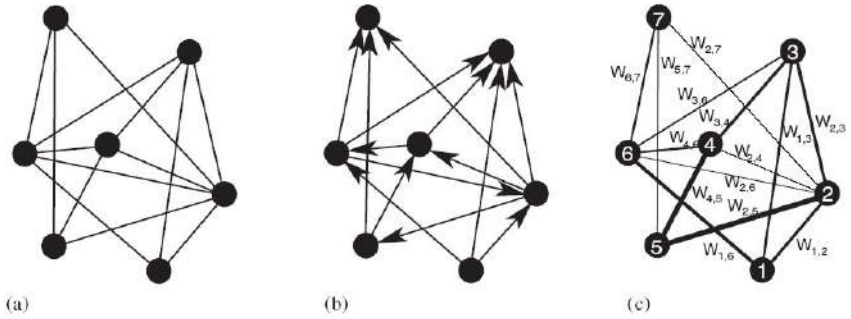
- statistical descriptive measures
- biomolecular annotation analyses

Here, we focus on statistical descriptive measures; biomolecular annotation analyses will be illustrated in the practice on the analysis of controlled biomolecular annotations.

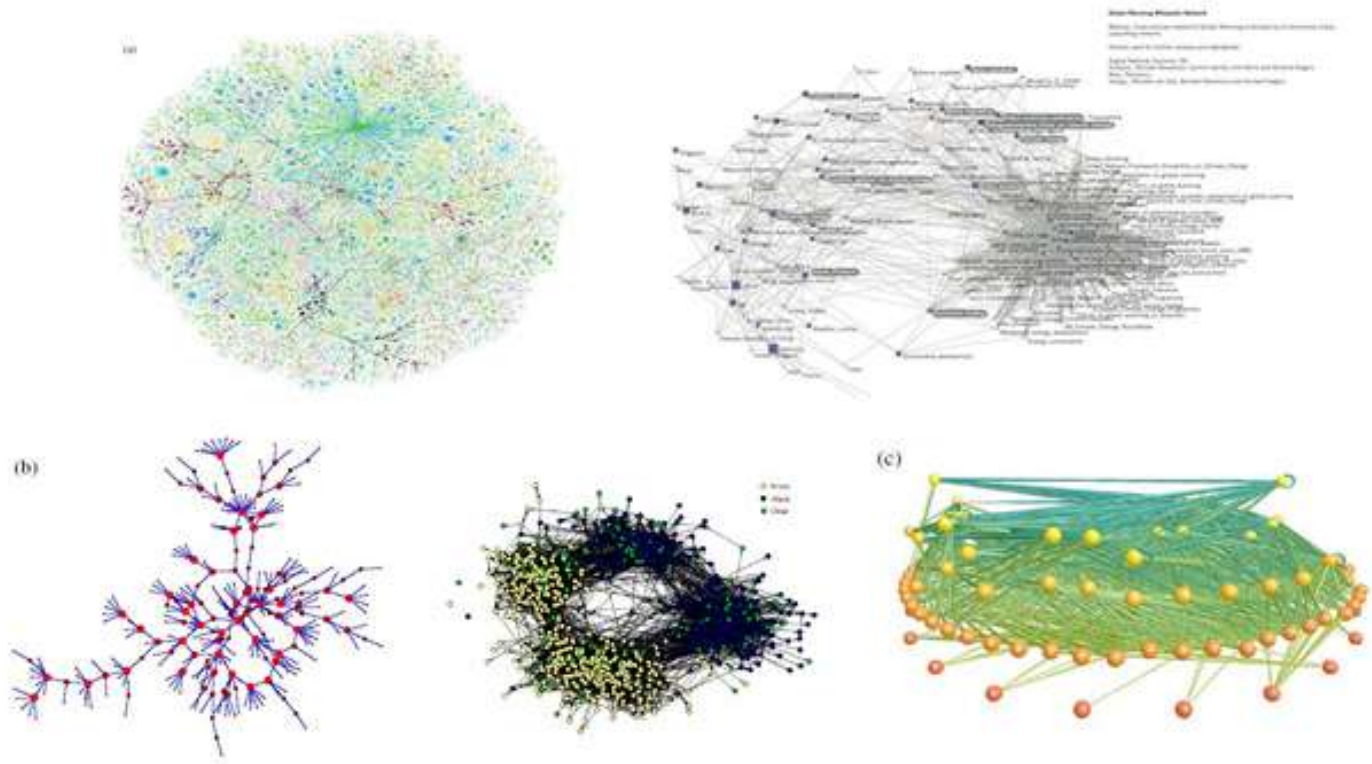
V.E.1 Complex networks

According to Wikipedia: In the context of network theory, a **complex network** is a **graph** (network) with **non-trivial topological features** – features that do not occur in simple networks such as lattices or random graphs but often occur in graphs modelling of real systems. The study of complex networks is a young and active area of scientific research (since 2000) inspired largely by the empirical study of real-world networks such as computer networks, technological networks, brain networks and social networks

What is a network? It is a series of components, systems, subsystems or entities **linked/interacting** to one another. Entities are represented by N nodes (or vertices) and E edges (or links):

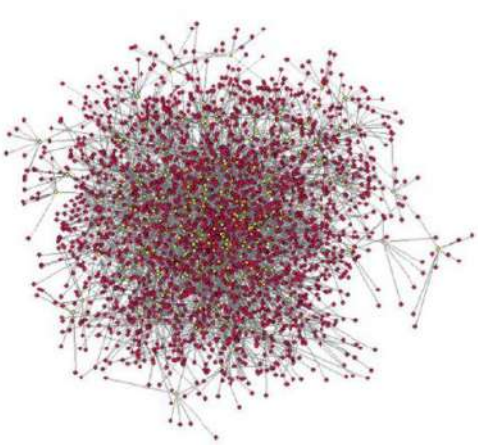


Networks can be **classified** according to edge properties: *undirected* (a,c) or *directed* (b); *weighted* (c) or *unweighted* (binary) (a,b).

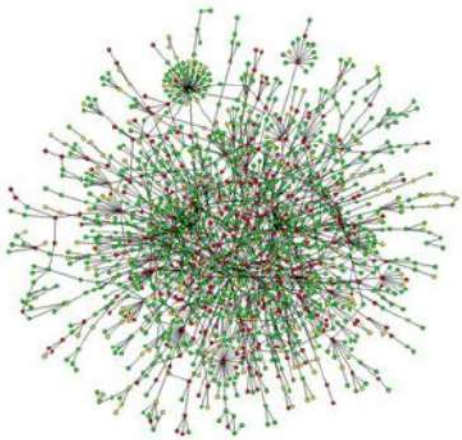


Networks examples

Similar problems are found in different contexts, which leads to common theories, methods and algorithms:



The "directors network" of the Italian companies

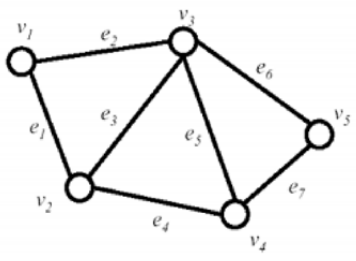


The protein interaction network of yeast

V.E.2 Adjacency matrix

An unweighted network of N elements is completely described by an **adjacency matrix** A of size $N * N$

$$a_{ij} = 1, \text{ if link } i \to j \text{ exists}; \qquad a_{ij} = 0, \text{ otherwise}$$



$A_G =$

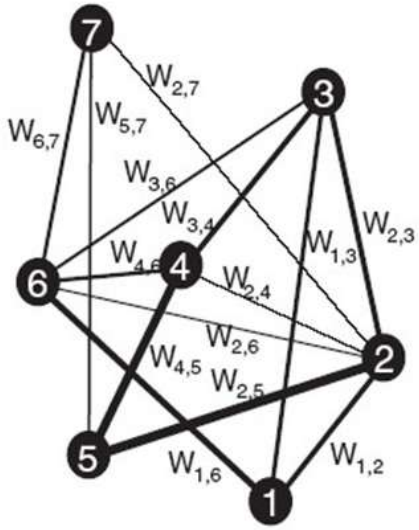
	v_1	v_2	v_3	v_4	v_5
v_1	0	1	1	0	0
v_2	1	0	1	1	0
v_3	1	1	0	1	1
v_4	0	1	1	0	1
v_5	0	0	1	1	0

A is **symmetrical** if the network is *undirected*, **asymmetrical** if the network is *directed*.

Typically, A has a small density: $\rho = \frac{E}{N(N-1)}$ (dir.) or $\rho = \frac{E}{N(N-1)/2}$ (undir.)

A weighed network is described by the $N * N$ weight matrix, $W = [w_{ij}]$:

$$w_{ij} > 0, \text{if link } i \rightarrow j \text{ exists}, \qquad w_{ij} = 0, \text{otherwise}$$



V.E.3 Topological measures of networks (avg., diam., clustering, centrality)

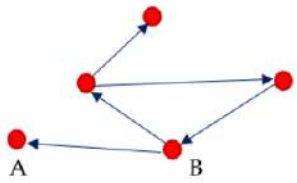
- **Average distance** (in a network, computed between each pair of nodes of the network)

The distance $l(A, B)$ between 2 nodes A and B of a network is the **minimum number of edges** between A and B.

The average distance L of a network is the sum of all distances $l(A, B)$ computed among every pair of nodes of the network (that are linked…), normalized by the total number of node pairs:

$$L = \langle l(A,B) \rangle = \frac{\sum_{A,B} l(A,B)}{N(N-1)/2} \quad \text{(for undirected network)}$$

$$L = \langle l(A,B) \rangle = \frac{\sum_{A,B} l(A,B)}{N(N-1)} \quad \text{(for directed network)}$$



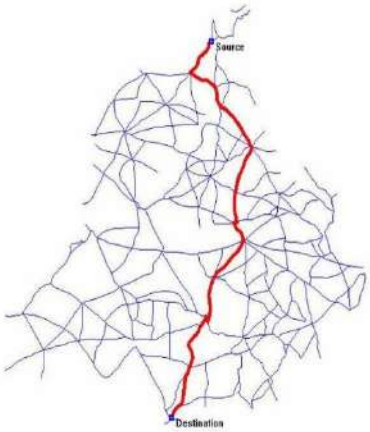
Here, we cannot move from A to B, so the distance is infinite, but the distance from B to A is 1

- **Diameter** (of a network)

The diameter D of a network is the maximum length among all l_{ij}:

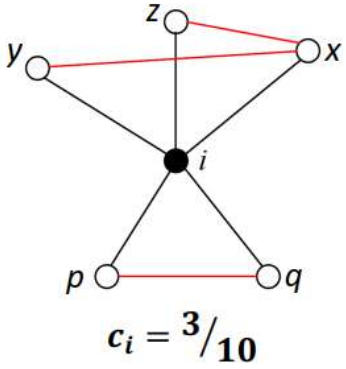
$$D = \max l_{ij}, \qquad 1 \leq L \leq D \leq N-1$$

Where l_{ij} is the distance between node i and node j, L is the average distance, and N is the total number of nodes



- **Clustering coefficient** (of each node of a network and of an entire network)

The clustering coefficient quantifies the "*local link density*" of the network by counting the triangles in the network.



How frequently, if we have the links $j \leftrightarrow i$ and $i \leftrightarrow l$, then we also have the link $j \leftrightarrow l$ (thus, the triangle *j, i, l*)?

The **(local) clustering coefficient** c_i, with $0 \leq c_i \leq 1$, of node i is:

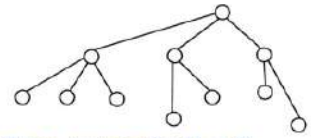
$$c_i = \frac{\#triangles\ connected\ to\ i}{\#triplets\ j,i,l\ centered\ on\ i} = \frac{e_{k_i}}{k_i(k_i-1)/2}$$

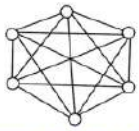
Where $\boldsymbol{k_i}$ is the total number of connections that i has (in this example the black lines) and neighbours of $\boldsymbol{e_{k_i}}$ is the number of links directly connecting the i (in this example the red lines)

Once we decide that we want to compute the clustering coefficient of node i, we check to which other nodes it is linked, and then we can "forget about i" and just check whether the other nodes are linked together

The **(global) clustering coefficient** C of a network is the average of the network c_i:

$$C = < c_i > = \frac{1}{N}\sum_i c_i$$


Tree network: C = 0


Complete network: C = 1

Network	Size	Clustering coefficient	Average path length
Internet, domain level [13]	32711	0.24	3.56
Internet, router level [13]	228298	0.03	9.51
WWW [14]	153127	0.11	3.1
E-mail [15]	56969	0.03	4.95
Software [16]	1376	0.06	6.39
Electronic circuits [17]	329	0.34	3.17
Language [18]	460902	0.437	2.67
Movie actors [5, 7]	225226	0.79	3.65
Math. co-authorship [19]	70975	0.59	9.50
Food web [20, 21]	154	0.15	3.40
Metabolic system [22]	778	–	3.2

This table shows that for the networks that have a global clustering coefficient near 0 (resp. 1), they will behave somewhat like tree (resp. complete) networks

- **Centrality measures**: degree, closeness, betweenness and eigenvector (of each node of a network)

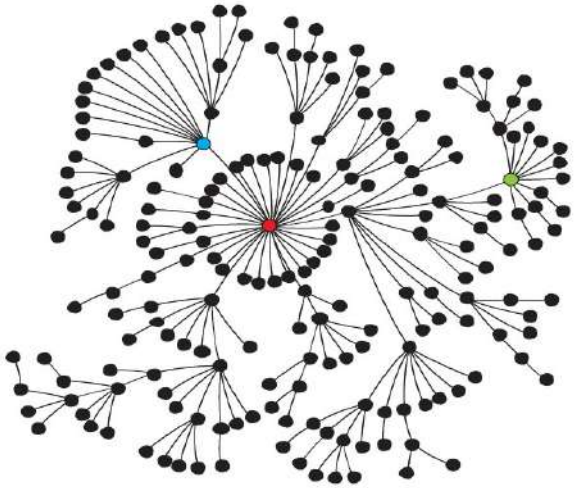
The centrality of a node is a measure of its importance in the network, it is a ranking. All centrality measures should provide more or less similar rankings, because they rely on the same network.

Degree centrality

The importance of a network node can trivially be captured by the **number $\boldsymbol{k_i}$ of its neighbours** (i.e., interactions, communication channels, sources-destinations of information, etc.).

Therefore, the network "**hubs**" are the most central nodes of the network.

- For weighted networks: **strength centrality**
- For directed networks, the degree of a node can be distinguished in **out** and **in degree**.



Closeness centrality

A node of a network is central if, on average, it is **close** (i.e., has short distance) to all the other nodes of the network: it has better access to the network information, more direct influence on other nodes, etc.

The **average distance** (number of nodes to access it) from node i to all the other $n-1$ network nodes is:

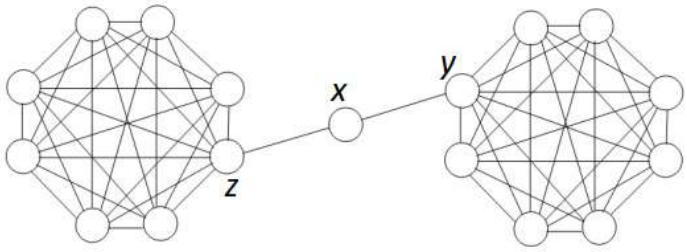
$$l_i = \frac{1}{n-1}\sum_j d_{ij}$$

The **closeness centrality** of a network node is defined as:

$$c_i = \frac{1}{l_i} = \frac{n-1}{\sum_j d_{ij}}$$

- If the network is directed, we must distinguish between **in-** and **out-closeness** of a node of the network
- If the network is weighted, several (non-trivial) generalized definitions of node closeness are available

Betweenness centrality



The **betweenness** b_i of node i of the network is the *fraction of shortest paths* connecting all the pairs of nodes of the network that pass through the node i.

$$b_i = \sum_{j,k} \frac{\#shortest\ paths\ connecting\ j\ and\ k\ via\ i}{\#shortest\ paths\ connecting\ j\ and\ k} = \sum_{j,k} \frac{n_{jk}(i)}{n_{jk}}$$

Eigenvector centrality

The **eigenvector centrality** γ_i of a network node is (proportional to) the *sum of the centralities of the node neighbours* (i.e., a node is important if it relates to many and/or important nodes):

$$\gamma_i = \alpha \sum_j a_{ij}\gamma_j$$

Where α is a constant and a_{ij} are the network adjacency matrix (A) elements. Letting $\gamma = [\gamma_1, \gamma_2, \dots, \gamma_N]^T$ and $\lambda = \frac{1}{\alpha}$, we obtain the eigenvector equation:

$$A\gamma = \lambda\gamma$$

If the network is **connected** (i.e., A is not reducible to block upper triangular form by simultaneous row/column permutations), the eigenvector centralities γ_i are given by the **only solution** with $\lambda > 0, \gamma > 0$ for all nodes i of the network (Frobenius-Perron theorem).

V.E.4 Degree distribution (of a network)

The **degree distribution** $P(k)$ of a network specifies the fraction of network nodes having exactly degree k (i.e., the probability that a randomly selected node has degree k):

$$P(k) = \frac{\#nodes\ with\ degree\ k}{N}, \qquad \sum_k P(k) = 1$$

It is often more practical to consider the **cumulative degree distribution**:

$$\bar{P}(k) = \frac{\#nodes\ with\ degree \geq k}{N} = \sum_{h=k}^{k_{\max}} P(h)\,, \qquad \bar{P}(k_{\min}) = 1$$

The ***r*-moments** of the degree distribution $P(k)$ are:

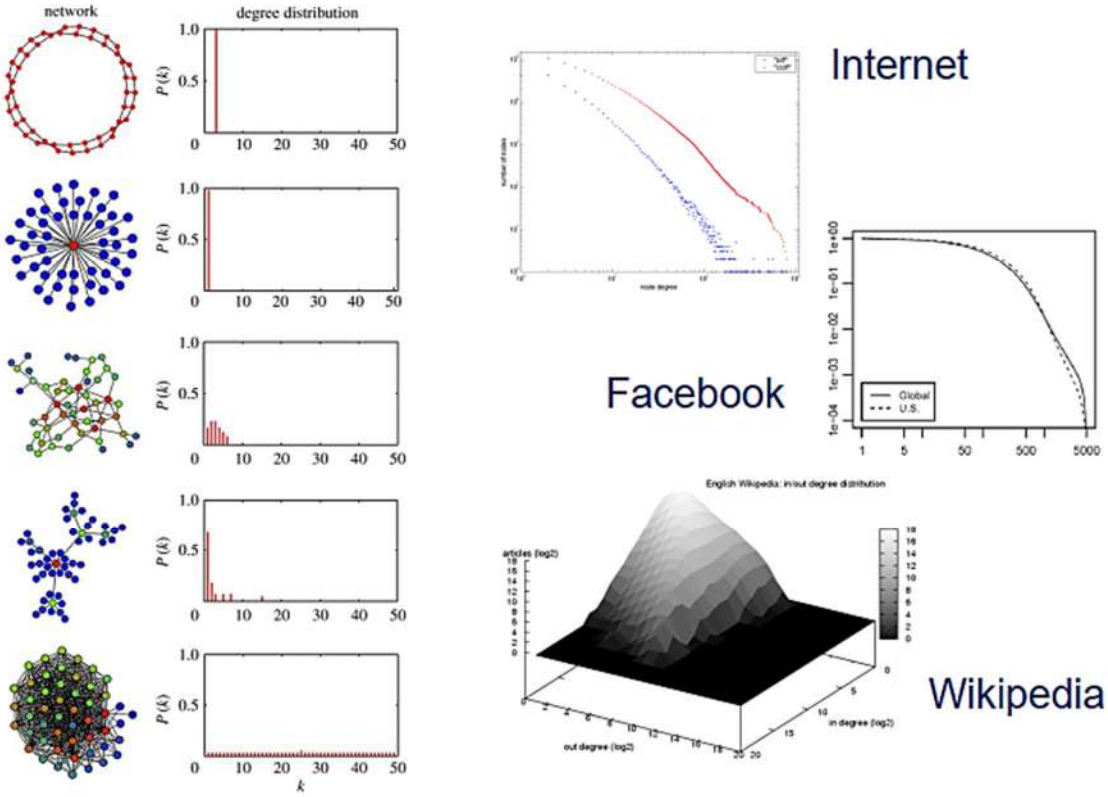
$$< k^r > = \sum_k k^r P(k)\,, \qquad r = 1,2,...$$

The first moment ($r = 1$) is the **average** degree:

$$< k > = \sum_k kP(k) = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N}$$

Where E is the number of edges and N is the number of nodes in the network

Constant degree distribution defines homogeneous networks (all nodes have same degree). Yet, real world networks do not have constant degree distribution:



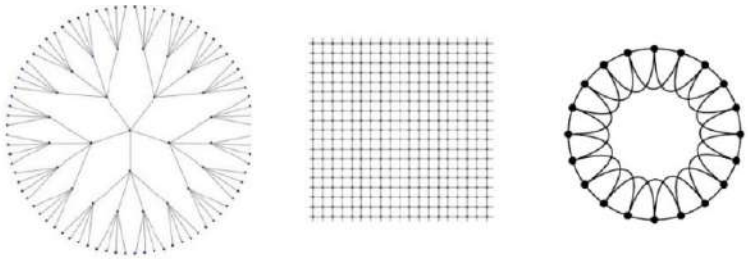
V.F Networks models

Mathematical models of networks have been studied in the past to better understand complex networks in general; some are:

- **Regular** networks (very distant from real complex networks)
- **Random** networks (very distant from real complex networks)
- **Scale-free** networks (real complex networks usually have such features)
- **Small-world** networks (real complex networks usually have such features)

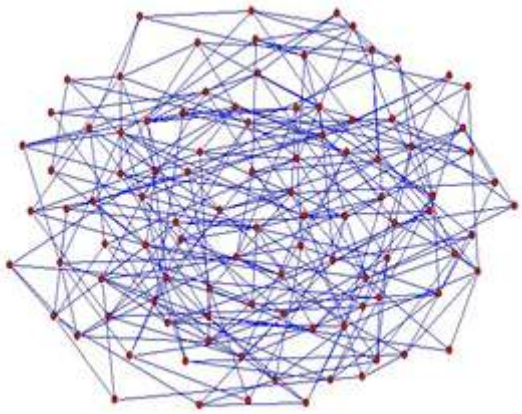
V.F.1 Regular networks

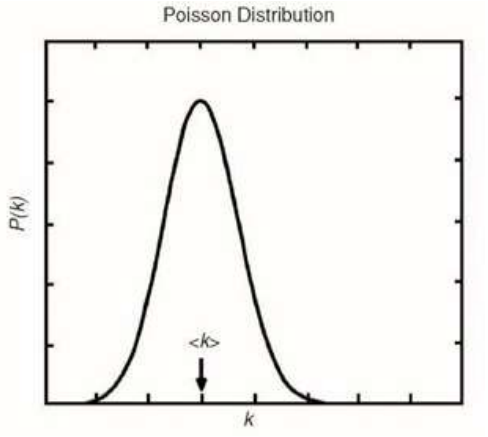
They are not representative of the real-world network



V.F.2 Random networks

Erdös-Rényi model is a mathematical model to generate random graphs. For example, this is a random (Erdös-Rényi) network, with $N = 100$ nodes and $E = 300$ randomly extracted connections between node pairs (hence average degree $< k > = 2 * \frac{300}{100} = 6$).

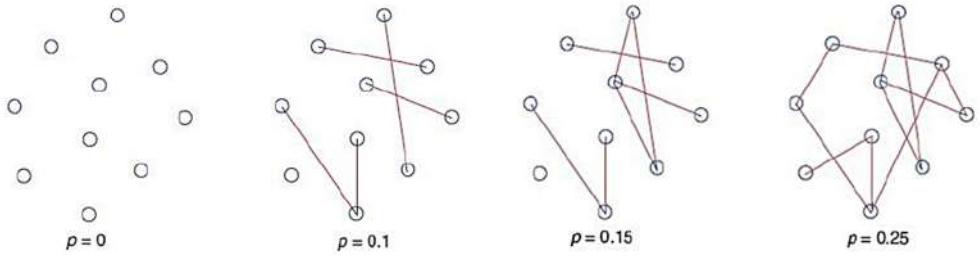




For **large N**, the degree of random networks is Poisson-distributed with $< k > = \frac{2E}{N}$:

- The typical scale of the node degree distribution is $k_i = < k >$ (Poisson distribution centred in $< k >$)
- Node degree distributions have small **fluctuations around $< k >$**
- The network is almost **homogenous**

Erdös-Rényi model of random networks:



Start from a graph with N nodes and no links, and connect each pair of nodes i, j with a given probability p. Pick a pair of nodes at random among the n nodes and add an edge between them if not already present; repeat until exactly E edges have been added. Thus, $p = \frac{2E}{N(N-1)}$

Some properties for $(N \to \infty)$:

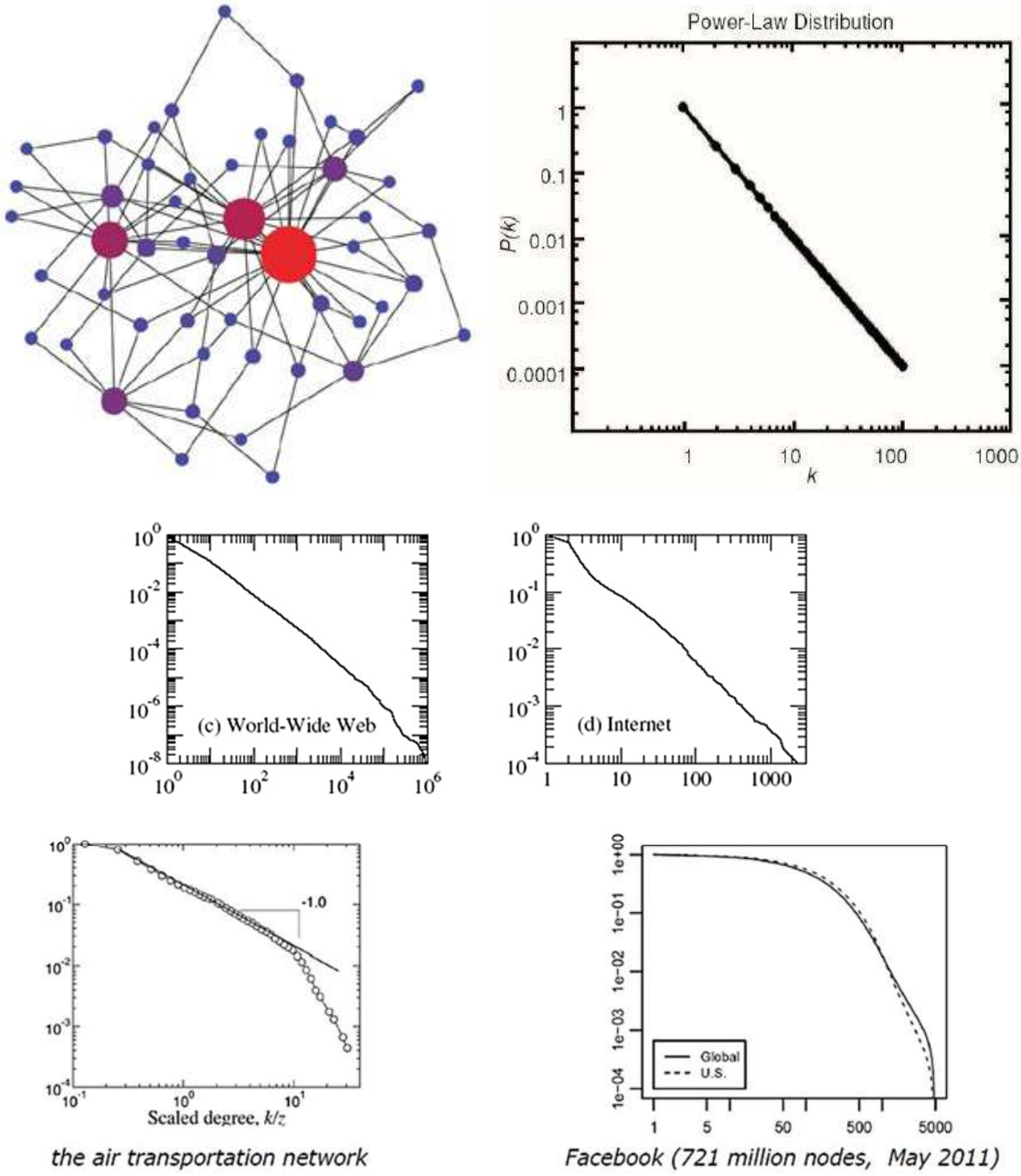
- The degree is Poisson distributed, with $< k > = p(N-1) = \frac{2E}{N}$:
$$P(k) = e^{-<k>} \frac{< k >^k}{k!}$$
- The network has a **giant component** if $< k >$ is *larger than 1* (a giant component is a well-connected part of the graph that contains almost all nodes, that means almost every node is reachable from almost every other).
- The **average distance** L grows slowly with N.
- The **clustering coefficient** C tends to 0 as N grows.

V.F.3 Scale-free networks

This is a scale-free network, obtained by **adding one node at a time**, and connecting it preferentially (i.e., with higher probability) to nodes with *higher degree* (Barabási-Albert algorithm). The network contains few very connected nodes ("hubs") and many scarcely connected nodes.

For large number of nodes N, the degree distribution is a power-law function $P(k) \cong k^{-\alpha}$

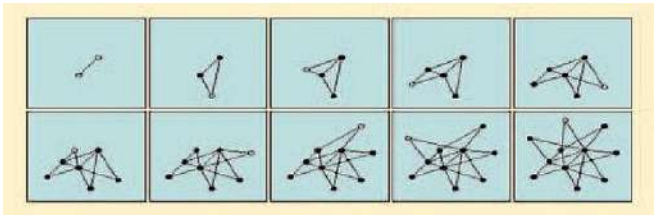
- node degrees have large *fluctuations* around $< k >$: there is no "typical" scale of node degree
- the network is strongly *heterogeneous*.



Examples of real degree distributions

Barabási-Albert algorithm (1999) is inspired by the WWW growth:

- *initialization*: start with m_0 nodes (arbitrarily connected).
- *growth*: at each step, add a new node x with $m \leq m_0$ new links connecting x to m existing nodes.
- *preferential attachment*: attach the new links preferentially (i.e., with higher probability) to nodes with high degree ("rich get richer"): that is, let the probability of connecting a new node x to an existing node i be $\frac{k_i}{\sum k_j}$, where k_i is the degree of node i and $\sum k_j$ the sum made over all pre-existing nodes (i.e. twice the current number of edges in the network)



Then, for $N \to \infty$ and k the node degree:

- The **average degree** tends to $< k >= 2E$, and the degree distribution tends to the power-law $P(k) \cong k^{-3}$
- $< k_2 >$ and thus the **variance** $< \sigma^2 >= < k^2 > - < k >^2$ diverge ($P(k)$ has a "heavy tail").
- The **average distance tends to** $\boldsymbol{L \approx \frac{log(N)}{log(log(N))}}$
- The **clustering coefficient** C vanishes with $c \approx \frac{(logN)^2}{N} \to 0$

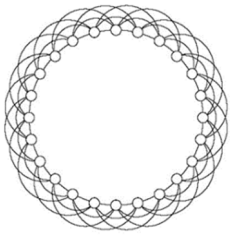
V.F.4 Small-world networks

Small-world networks, according to Watts and Strogatz, are a class of networks that are "**highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs**."

Small-world network are model networks that account for clustering while retaining the *short average path lengths* of the Erdös-Rényi graphs.

They are between a *randomized structure* close to Erdös-Rényi graphs and a *regular ring lattice*. In 1998, Watts and Strogatz demonstrated that adding a few long-distance connections to a regular network yields a dramatic decrease of *average distance* L.

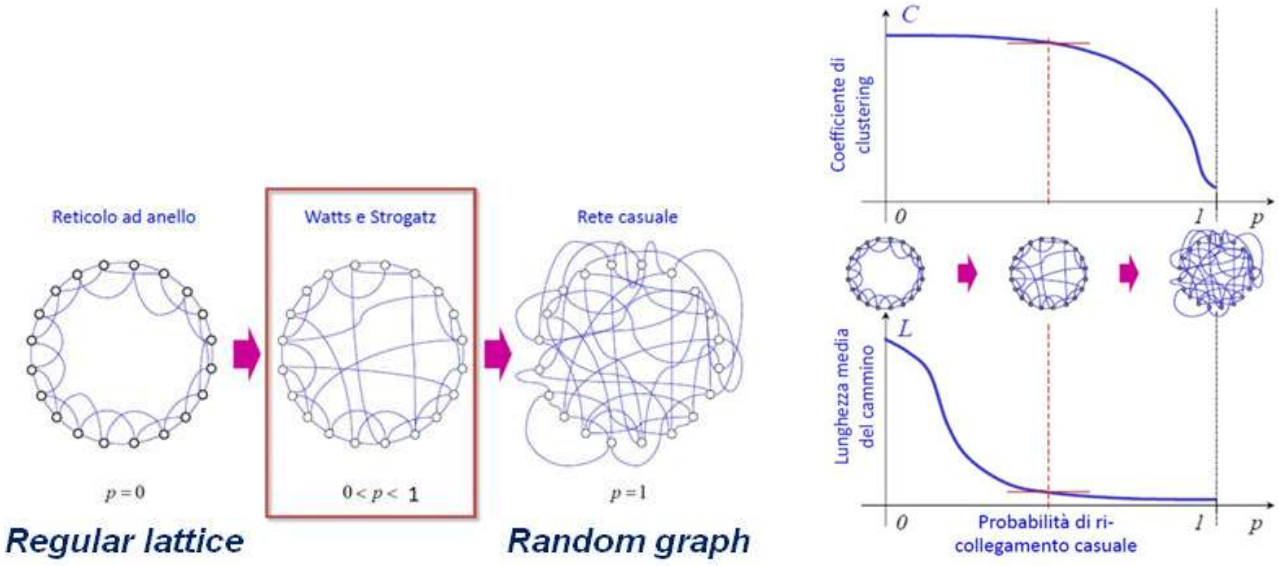
How to build a Small-world network?



1. Start from a **regular "ring" graph** with N nodes, where each node is connected to the m right-neighbours and to the m left-neighbour (i.e., each node has exactly degree $2m$).

The network has large *clustering coefficient* C (typical of "regular" networks): $C = \frac{3m-3}{4m-2}$ and the *average distance* is also large (it grows linearly with N) $L = \frac{N}{4m}$

2. "**Rewiring**": Scan all nodes $i = 1, 2, \dots, N$. Consider all the links $i \leftrightarrow j$ connecting node i to its right neighbor nodes and, with **rewiring probability** p, break the connection to node j and redirect it to a randomly selected node



If p is small, the *local properties* are not significantly modified:

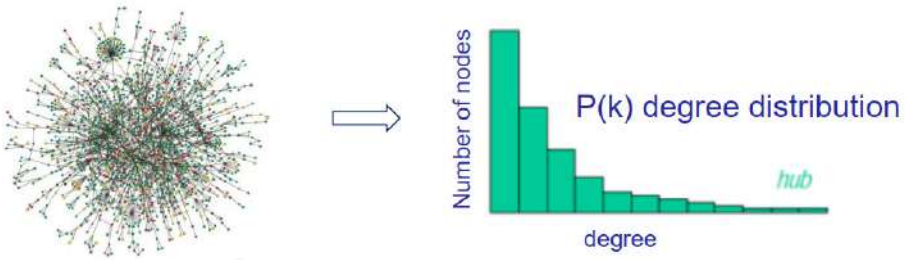
- the *degree distribution* remains concentrated around the average degree (unchanged!) $< k > = 2m$
- the *clustering coefficient* C does not change significantly
- but there is a **dramatic decrease of the average distance** as long-distance connections appear

In a suitable rewiring probability p interval, the network mimics many typical real-world networks, i.e., at the same time:

- The clustering coefficient is *large* (from the regular lattice)
- The average distance is *small* (from the random network)

V.G Biological examples

The majority of biological networks have **scale-free properties** and thus can be modelled as scale-free networks. Here, an example of a **protein-protein interactions** network.



While most proteins participate in only a *few interactions*, a few participate in dozens (hubs). This is a typical feature of scale-free networks. Each hub is associated with functionally different groups of proteins; therefore *clusters of proteins* can be found through the hubs.

Other examples of scale-free organization include **gene regulatory** or **gene co-expression networks**, where nodes are genes and links can be derived from gene expression correlations, e.g., computed on microarray data.

V.G.1 Gene co-expression networks

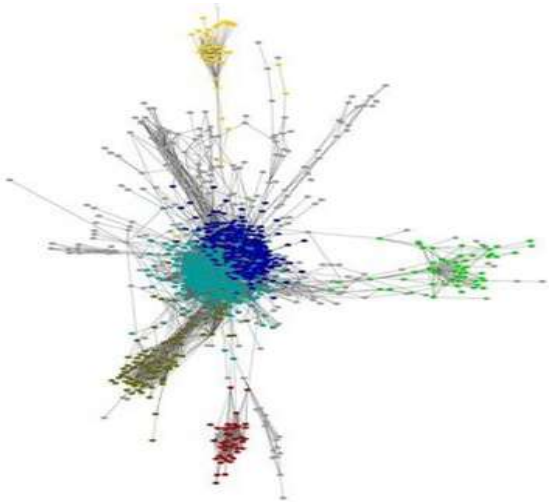
Each gene is estimated on average to interact with four to eight other genes and to be involved in 10 biological functions.

Gene co-expression networks (GCN) represent the *interconnectedness between genes* and connect pairs of genes that are significantly similar/correlated. The nodes are the genes, the edges are the co-expression relationships between them.

Densely connected sub-networks form ***gene modules*** (communities), usually related to biological functions or biologically relevant traits.

The **Guilt By Association** (GBA) paradigm assumes that *strongly co-expressed genes share functionalities.*

Gene networks and particularly **co-expression networks** provide the potential to id*entify hundreds of genes that are associated with complex human diseases* and that could serve as points for therapeutic interventions. This information is important for predicting the functions of new genes and finding genes that play key roles in complex human diseases.



Gene co-expression network analysis of prostate cancer. A smaller edge indicates a higher correlation. Nodes are coloured according to the modules in which they belong

V.G.2 Weighted gene co-expression network analysis

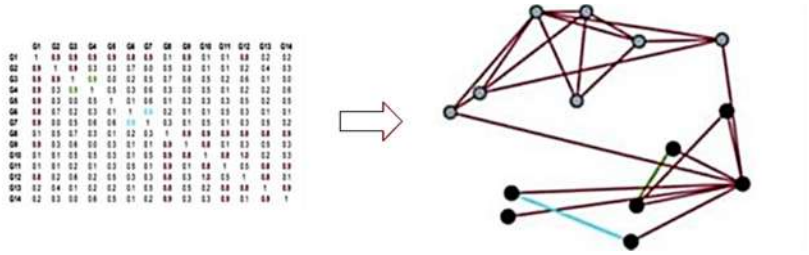
Based on the Guilt By Association (GBA) paradigm, genes with closely functional linkages or involved in similar pathways may have similar expression profiles.

The Weighted Gene Co-expression Network Analysis (**WGCNA**) is a popular systems biology strategy to explore the system-level functionality of a transcriptome not only **constructing gene co-expression networks** but also **detecting gene modules** and **identifying hub genes** within modules.

Also, WGCNA analyses the relationships between gene modules and sample traits to explore the biological mechanisms behind certain traits. WGCNA has been widely applied to identify gene modules associated with clinical annotations in many cancer diseases [see slides for references].

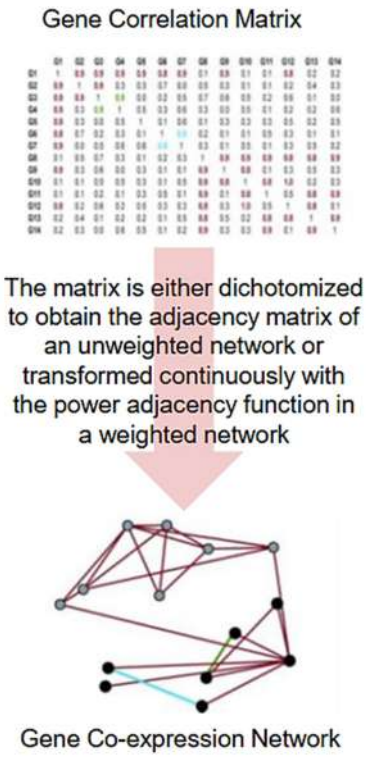
Pipeline

1. **Construct a gene co-expression network** represented mathematically by a matrix, the element of which indicates co-expression similarity between a pair of genes.



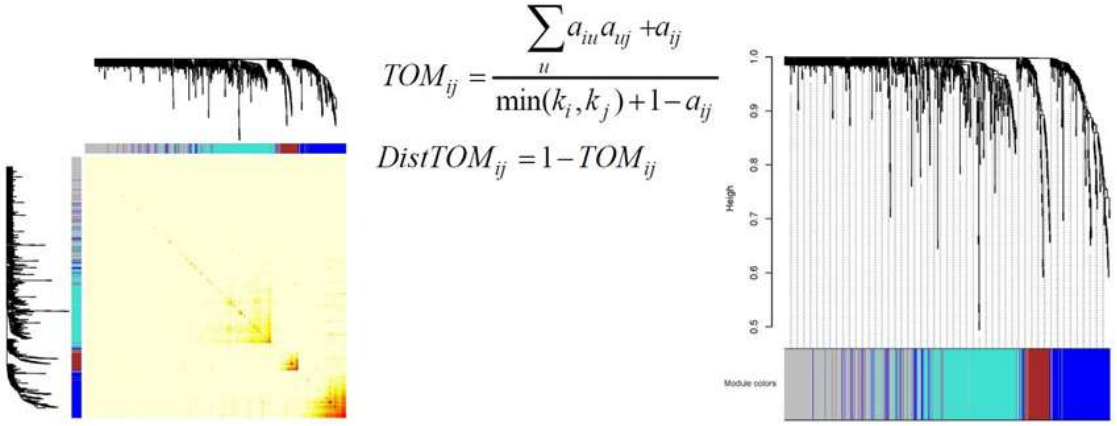
In genomic data analysis, *samples are assumed independent of each* other and WGCNA often uses the absolute value of Pearson correlation to measure the magnitude of co-expression between genes.

Instead of dichotomizing gene co-expression (connected = 1, unconnected = 0), WGCNA uses a *'soft' threshold to determine the weights of the edges* connecting pairs of genes, which has been proven to yield more robust results than unweighted networks. According to the "scale free topology criterion" in Zhang and Horvath (2005), *an appropriate soft threshold makes the resulting co-expression network closer to a scale-free network.*



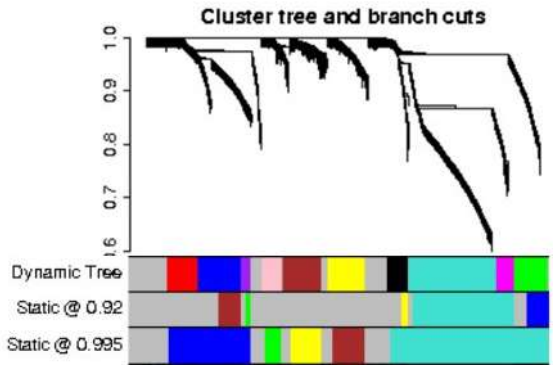
Power adjacency function to obtain a weighted gene co-expression network: $a_{ij} = \left|cor(x_i, x_j)\right|^{\beta}$

2. **Identify modules using hierarchical clustering**: WGCNA uses a *topological overlap matrix* and dissimilarity measure to obtain modules, that can be biologically meaningful in real data analysis.



WGCNA hierarchical clustering works with a *bottom-up approach*: each gene starts in its own cluster and clusters are then joined by merging the two most similar clusters together, iteratively.

A *tree-like structure*, known as a dendrogram, is produced and branches of the hierarchical clustering dendrogram, corresponding to gene modules, can be identified using the dynamic tree cut method.

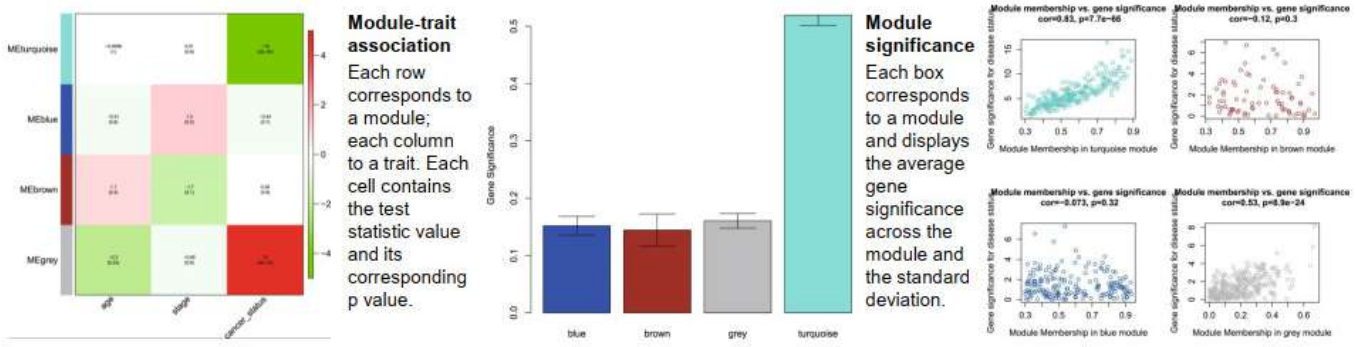


3. **Relate modules to phenotypic and/or clinically relevant traits**:

One can test the association between the *module eigengene* (ME - first principal component of the module) and the trait.

One can also use the *module significance* (MS), which is defined as the average *gene significance* (GS) to a trait of all genes in the module. The GS of a node is the correlation between the node and the trait, while the *module membership* (MM) of a gene i ($MM(i) = cor(x_i, ME)$) measures the importance of the gene within the module.

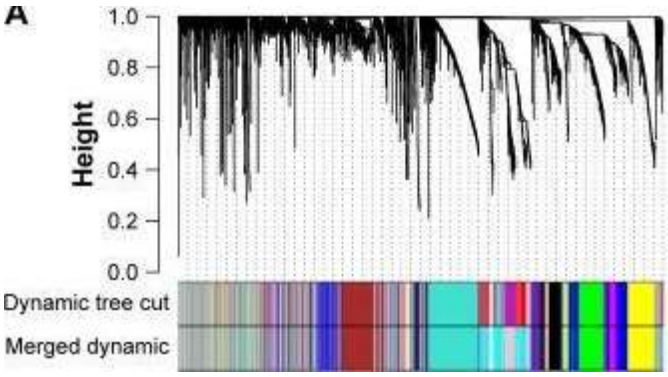
Modules with high trait significance may represent pathways associated with the trait.



4. **Study inter-module relationships and module preservation**:

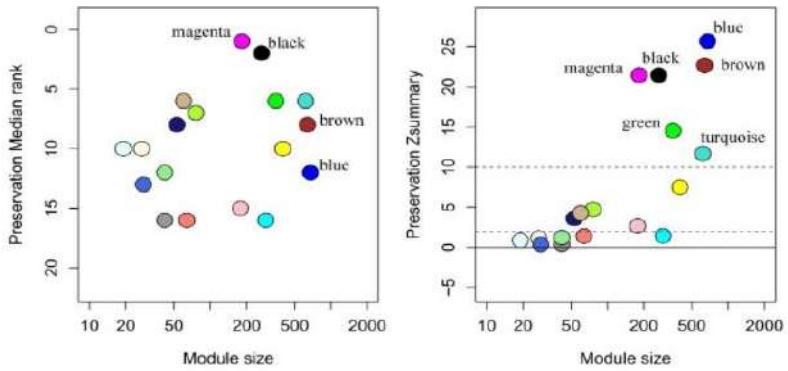
WGCNA uses ME as a *representative profile of a module* and quantifies module similarity by eigengene correlation: studying the relationships of the modules can help to find which modules are *highly related and can be merged*.

Testing module preservation requires to assess whether similar network modules can be *constructed* using other data: studying module preservation can help to find which modules are more interesting and to *check robustness* of module definition



Gene clustering dendrogram

The first colour band indicates the modules detected by dynamic tree cut. The second colour band indicates the modules after merging similar modules (e.g., having the height of ME lower than 0,25)

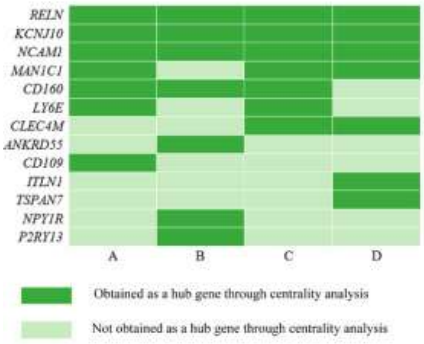


The smaller preservation Median Rank is associated with a stronger preservation. An integrated function of WGCNA package for module preservation was used to calculate the preservation ($n_{Permutations} = 100$) and the Z summary score (Z score).

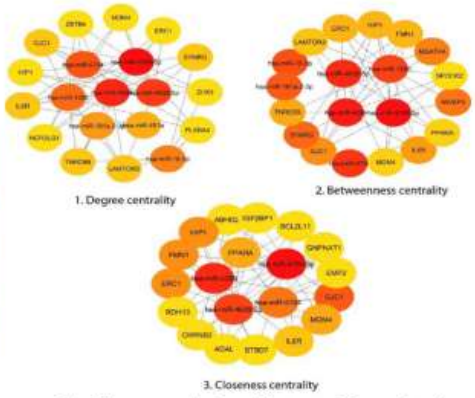
In right graph, the dashed blue and green lines indicate the thresholds $Z = 2$ and $Z = 10$, respectively. The Z score lower than 2 indicates the modules has no preservation; 2 to 10 indicates low to moderate preservation and higher than 10 indicates strong preservation.

5. **Find key drivers in interesting modules**:

Hub genes can be traces through *centrality analysis*. Hub genes can also be corresponding to the nodes having *higher module memberships* and *gene significance*.



Distribution of hub genes identified by several types of centrality. (A) degree centrality; (B) clustering coefficient; (C) betweenness centrality; (D) closeness centrality



Significant module with top 20 ranked genes based on 1. degree (20 nodes - 52 edges); 2. betweenness (20 nodes - 50 edges); and 3. closeness (20 nodes - 49 edges). Red indicates the highest rank, whereas yellow indicates the lowest rank. 5 hub genes were found to be common in all the cases and considered as significant hub nodes of the module

Module	Symbol	MM	GS	Log_2FC
Gmagenta	SPARC	0.95	0.12	1.46
Gmagenta	COL1A2	0.94	0.10	1.61
Gmagenta	COL1A1	0.93	0.08	2.67
Gmagenta	COL5A2	0.93	0.08	1.80
Gmagenta	COL5A1	0.93	0.10	1.60
Gmagenta	BGN	0.91	0.13	2.56
Gmagenta	THBS2	0.90	0.09	2.35
Gmagenta	SULF1	0.90	0.07	1.87
Gmagenta	WISP1	0.90	0.03	2.67
Gmagenta	THY1	0.90	0.10	1.38
Ggreen	TNS1	0.96	0.11	−2.68
Ggreen	LMOD1	0.96	0.10	−3.45
Ggreen	MYLK	0.96	0.09	−3.29
Ggreen	HSPB8	0.95	0.10	−3.34
Ggreen	PDZRN4	0.95	0.09	−4.00
Ggreen	TAGLN	0.95	0.12	−2.83
Ggreen	DACT3	0.95	0.13	−2.41
Ggreen	GNAO1	0.94	0.11	−3.35
Ggreen	TMEM200B	0.94	0.11	−1.64
Ggreen	HAND2	0.94	0.10	−3.82

Hub genes based on MM (module membership), GS (gene significance), reporting also fold change of differential expression (as Log2 ratio)

V.G.3 Open issues in WGCNA

WGCNA only looks at **co-expression across all samples**. As transcriptional regulation is highly context specific, clustering potentially misses local co-expression effects which are present in only a subset of all biological samples.

WGCNA is **unable to assign genes to multiple modules**. The issue of overlap between modules is especially problematic given the increasing evidence that gene regulation is highly combinatorial and that gene products can participate in multiple pathways.

WGCNA **does not account for other kinds of links** like regulatory relationships between genes. As the variation in target gene expression can at least be partly explained by variation in transcription factor expression, including this information could therefore boost module detection.

WGCNA **relies on hierarchical clustering** for module identification, while other interesting community detection techniques could be used

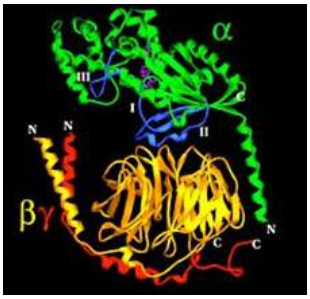
VI. Bio-terminologies and Bio-ontologies

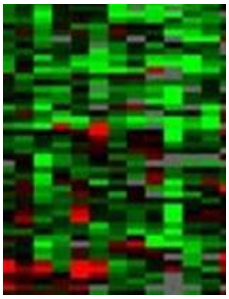
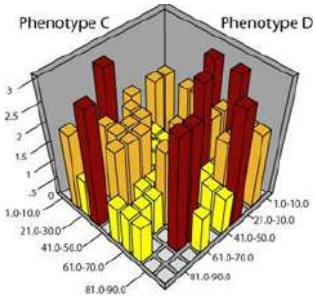
VI.A Introduction

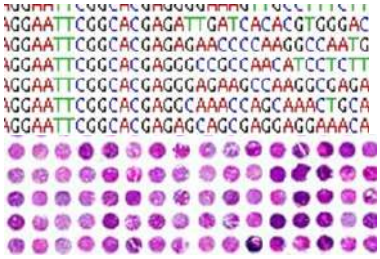
Huge *growth* in online biomedical data sets:

- Genomics (genetic sequences, SNPs)
- Gene expression microarrays
- Proteomics (mass spectrometry, protein arrays)
- Tissue arrays

Need for people and machines to **make sense** of massive data sets

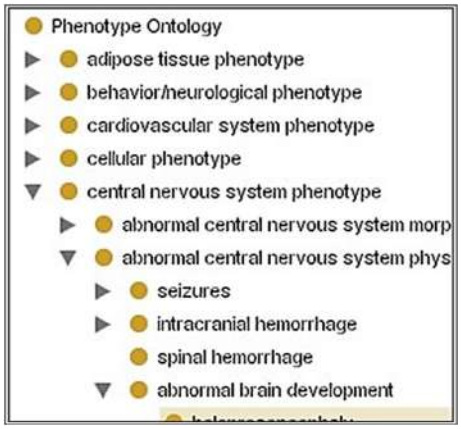




Bio-terminologies and **bio-ontologies** have an important role in e-science

- *Formal* and *explicit declarations* of the entities and relationships
- Built by *humans*, processed by *machines*
- Capabilities:
  - Relate *disparate* data
  - Enable data *interoperability*, data *summarization*, data *mining*



Several bio-terminologies and bio-ontologies are being developed and used to:

- Define, describe, and suitably identify information
- Favour information management and analysis with:
  - *Integration* of sparse and heterogeneous information
  - *Identification* and grouping of "similar" bio-sequences
  - *Statistical analysis* and data mining of controlled annotations to:
    - Highlight most relevant biological features
    - Help unveiling knowledge from data
- Support translational research (to quickly bring in the clinical practice new biomolecular knowledge)

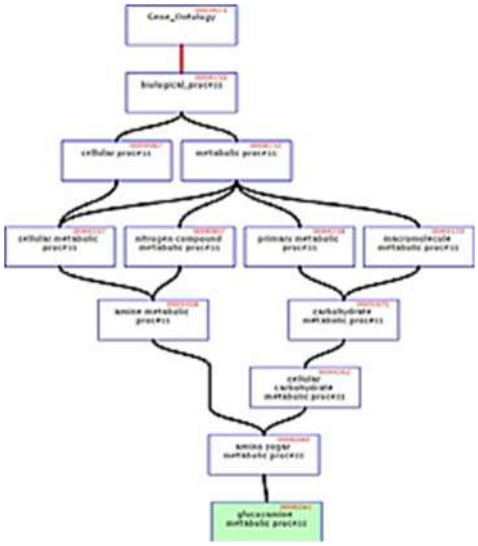
VI.A.1 Bio-terminologies

Collections of terms, *precise* and universally *comprehensible*, that univocally define and identify different concepts

- Useful for knowledge analysis and sharing
- *Controlled*: defined and maintained by groups of experts
- Increasing number, coverage and use in *molecular biology* and *biomedicine*:
  - Biochemical and metabolic pathways (KEGG)
  - Protein families and domains (Pfam)
  - Genetic diseases (OMIM)
  - Biological processes, molecular functions, cellular components (Gene Ontology)

Very useful to *enhance gene lists* with biological *information*

VI.A.2 Semantic networks



Logical structures used to **represent knowledge**, in a specific domain, through a *graph structure* composed of:

- a set of *elements*, the graph nodes, representing the domain concepts
- *relations* among domain concepts, the graph arches, representing the knowledge of the domain

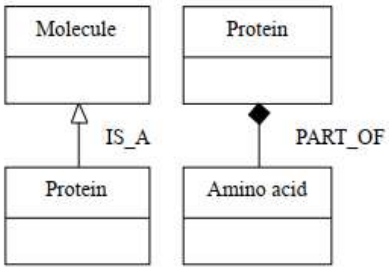
Also a **reasoning tool**: relations can be found between concepts not directly related

It can be implemented in software and automatically processed

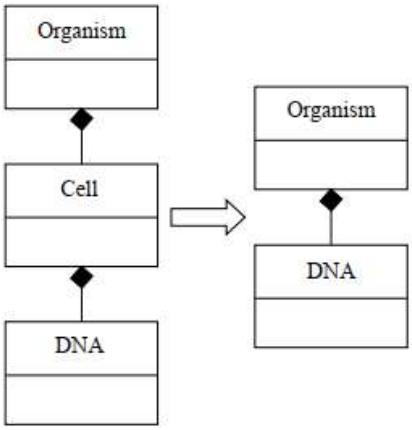
A type of arch for each type of *relation*

Main relations: `IS_A`, `PART_OF`:

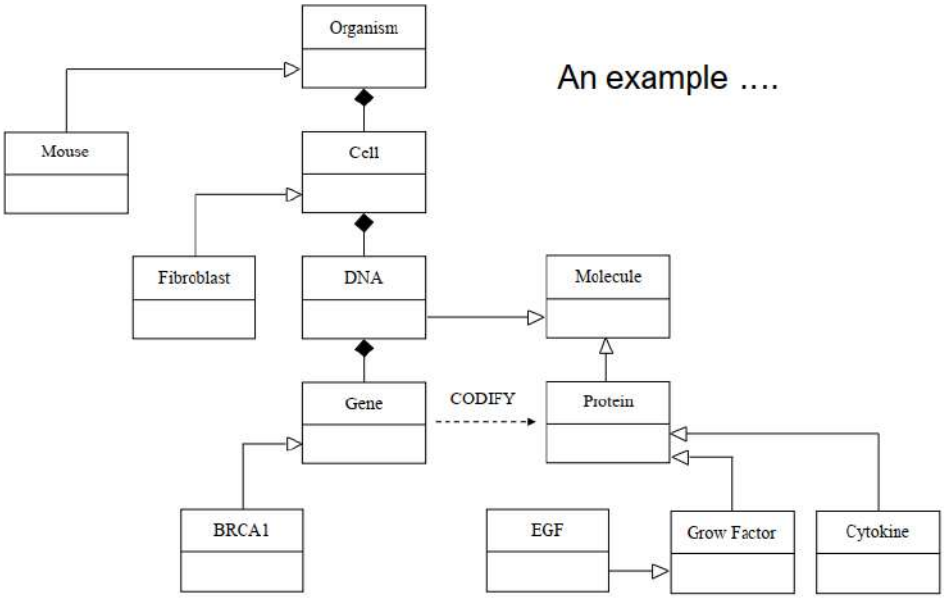
- Model *hierarchical* knowledge with concepts related at different levels of specification
- Their *transitive* property can be used to make useful inferences



`IS_A` allows attribute inheritance among related concepts (which can help to discover new relations):



Other specific relations: `EXPRESSED_BY`, `REGULATED_BY`, `ASSOCIATED_WITH`, …



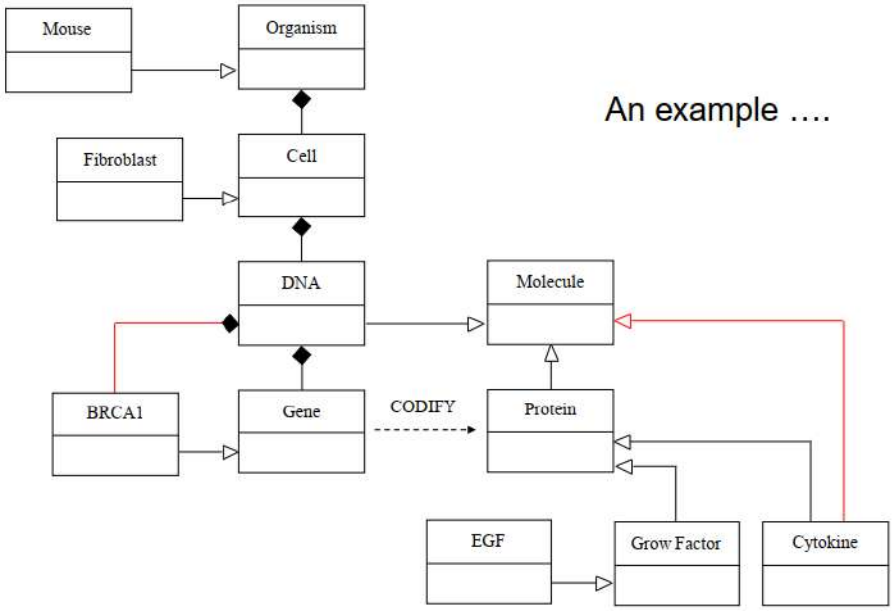
VI.A.3 Bio-ontologies

Several definitions of an **ontology** exist:

- "A specification of a conceptualization" (Gruber, 1993)
- "A partial specification of a conceptualization" (Guarino, 1998)
- "The subject of ontology is the study of the categories of things that exist, or may exist, *in some domain*. The product of such a study, called an ontology, is a structured catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. [...]" (Sowa, 1997)
- "We use the term 'ontology' in what follows to refer to any theory or system that aims to *describe*, *standardize* or provide *rigorous definitions* for *terminologies* used in the *domain*" (Smith 2003)

Ontologies are **semantic structures** used to:

- *Describe* the **knowledge** of a domain in a *textual* and computable form
- *Standardize* and provide rigorous definitions for the terminology used in the domain (bio-molecular/biomedical)
- Composed by a *controlled* (bio-)terminology and a semantic network
- Very useful for *automatic classification* and *inference*



In red: automatically inferred relationships

VI.A.4 Bio-ontology issues

Ontology *development* is **fragmented**:

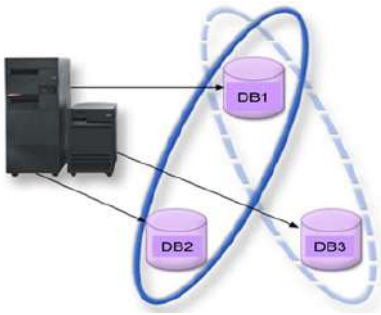
- *Separate* communities of biomedical researchers creating and maintaining ontologies
- *Different model* organism databases using ontologies to annotate experimental data
- Bioinformaticians creating algorithms to analyse these annotations
- These activities are **not unified** and produce often **not matching** ontologies; *unification* could allow:
  - Integration of each other and with other data
  - *Cross-species* analysis

Problems facing *ontology content* **curation**:

- Many different groups/consortia create ontologies; their efforts are *uncoordinated*
- Many different ontologies, *overlapping* content and *variable quality*
- Ontologies are *not interoperable*
- Ontology (and terminology) mapping efforts are *laborious*
  - Constitute barriers to accessing, effectively using and expanding data repositories

Problems facing *experimental data* **annotation** (i.e., controlled description of experimental data features):

- *Growing* amount of data of biomedical resources annotated with ontologies (MGED, GO, BioPAX), but:
  - Current resources confined to using *single ontology* for annotations
  - Difficult to relate *different annotation repositories* to each other
- Data integration efforts are laborious and made difficult by *mapping* difficulties



VI.B National Center of Biomedical Ontology

The US *National Institute of Health* (NIH) has funded the **National Center of Biomedical Ontology** (NCBO) (http://www.bioontology.org/)

- Mission: Advance biomedicine with *tools* and *methodologies* for the *structured* organization of knowledge
- Strategy: Develop, disseminate, and support:
  - *Open-source* ontology development and data annotation tools
  - Resources enabling scientists to *access*, review, and integrate disparate knowledge resources

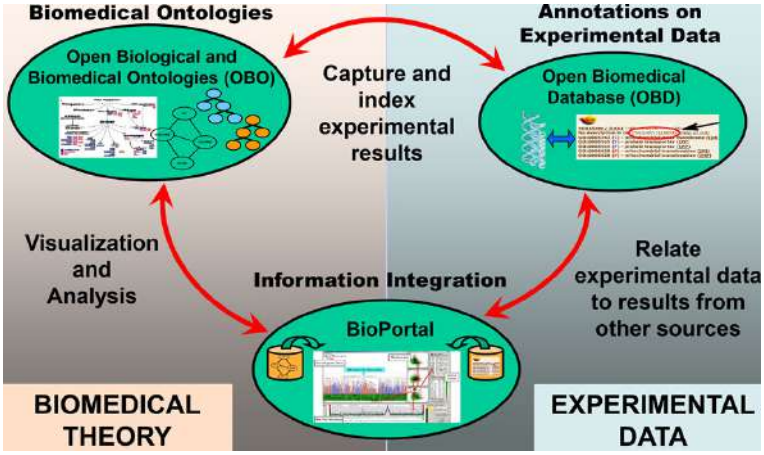
VI.B.1 Resources

Open Biological and Biomedical Ontologies (OBO): (http://www.obofoundry.org/): An integrated virtual library of biomedical ontologies

Open Biomedical Database (OBD): An online repository of OBO annotations on experimental data accessible via BioPortal

BioPortal (http://bioportal.bioontology.org/): A Web-based portal to:

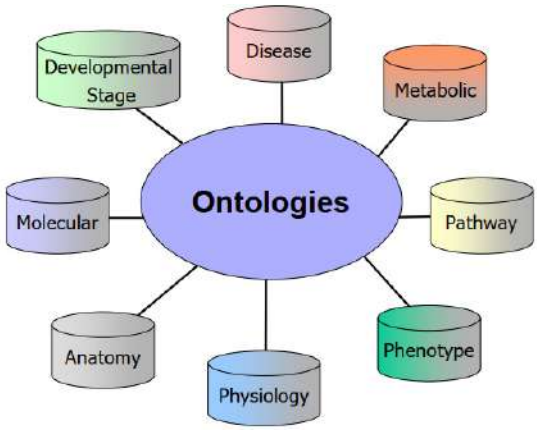
- Allow investigators and intelligent computer programs to *access* and use OBO
- Use OBO to *annotate* experimental data in OBD
- *Visualize* and analyse OBO annotations in OBD



VI.C Open Biological and Biomedical Ontologies

The OBO Foundry (http://www.obofoundry.org/) is an *open*, inclusive and *collaborative* experiment involving developers of science-based ontologies aiming at:

- *Establishing* principles for ontology development
- *Supporting* community members who are developing and publishing ontologies in the biomedical domain
- *Defining* a set of orthogonal, fully interoperable reference ontologies in the biomedical domain by virtue of:
  - Common design philosophy and implementation
  - Sharing of unique identifier space
  - Inclusion of definitions
- Enabling scientists and their instruments to *communicate* with *minimum ambiguity*



Ontology driven interoperability of bio-knowledge databanks

VI.C.1 Documentation

OBO documentation:

- Paper: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html
- The OBO Foundry wiki: http://www.obofoundry.org/wiki/index.php/Main_Page
- Introduction to Biomedical Ontologies: http://bioontology.org/wiki/index.php/Introduction_to_Biomedical_Ontologies
- Presentations:
  - Moving Beyond Ontology Libraries http://virtualgenomics.org/vcgb/talks/RubinVirtualGenomicsConf.ppt
  - US National Center for Biomedical Ontology (cBiO)
    - http://www.xmdr.org/presentations/Natasha-NoycBIO-Oct-2005.ppt
    - http://protege.stanford.edu/conference/2006/submiss ions/slides/2.1_Rubin.pdf

VI.C.2 Related projects [additional material]

The **Microarray Gene Expression Data** (MGED) Society (http://mged.sourceforge.net/), now the **Functional Genomics Data** (FGED) Society (http://www.mged.org/) is an international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments

The **Ontologies for Biomedical Investigations** (OBI) project (http://obi.sourceforge.net/) is developing an integrated ontology for the description of biological and medical experiments and investigations. This ontology will support the consistent annotation of biomedical investigations, regardless of the field of study

The **Human Proteome Organisation** (HUPO) Proteomics Standards Initiative (PSI) (http://www.psidev.info/index.php?q=node/258) defines community standards for data representation in proteomics to facilitate data comparison, exchange, and verification. It is also compiling guidelines for the development of controlled vocabularies

Standards and Ontologies for Functional Genomics (SOFG) (http://www.sofg.org/) is both a meeting and a website; it aims to bring together biologists, bioinformaticians, and computer scientists who are developing and using standards and ontologies with an emphasis on describing high-throughput functional genomics experiments

VI.C.3 OBO ontologies

To be part of OBO, an ontology must be:

- **Open**: accessible to everyone without any constrain
  - Its *origin* must be recognized
  - Subsequent *modifications* must be distributed under different names and identifiers
- Expressed in a **common and shared syntax** (OBO syntax, its extension, or in OWL (Web Ontology Language)), in order to ease use of the same tools and shared implementation of software applications
- **Clearly specified** and with a well-defined content: each ontology must be *orthogonal* to the other OBO ontologies
- **Not overlapping** other OBO ontologies: *partial* overlapping can be allowed to enable combination of ontology terms to form new terms
- **Able to include textual definitions of all terms**: since several biomedical terms can be ambiguous, the concepts they represent must be precisely defined with their meaning within the specific ontology domain they refer (which must be also specified)
- Well **documented**

OBO ontologies tackle several different biological aspects:

- Organism taxonomies
- Anatomies
- Cell types
- Genotypes
- Sequence attributes
- Temporal attributes
- Phenotypes
- Diseases
- …

Mature ontologies undergoing incremental reform:

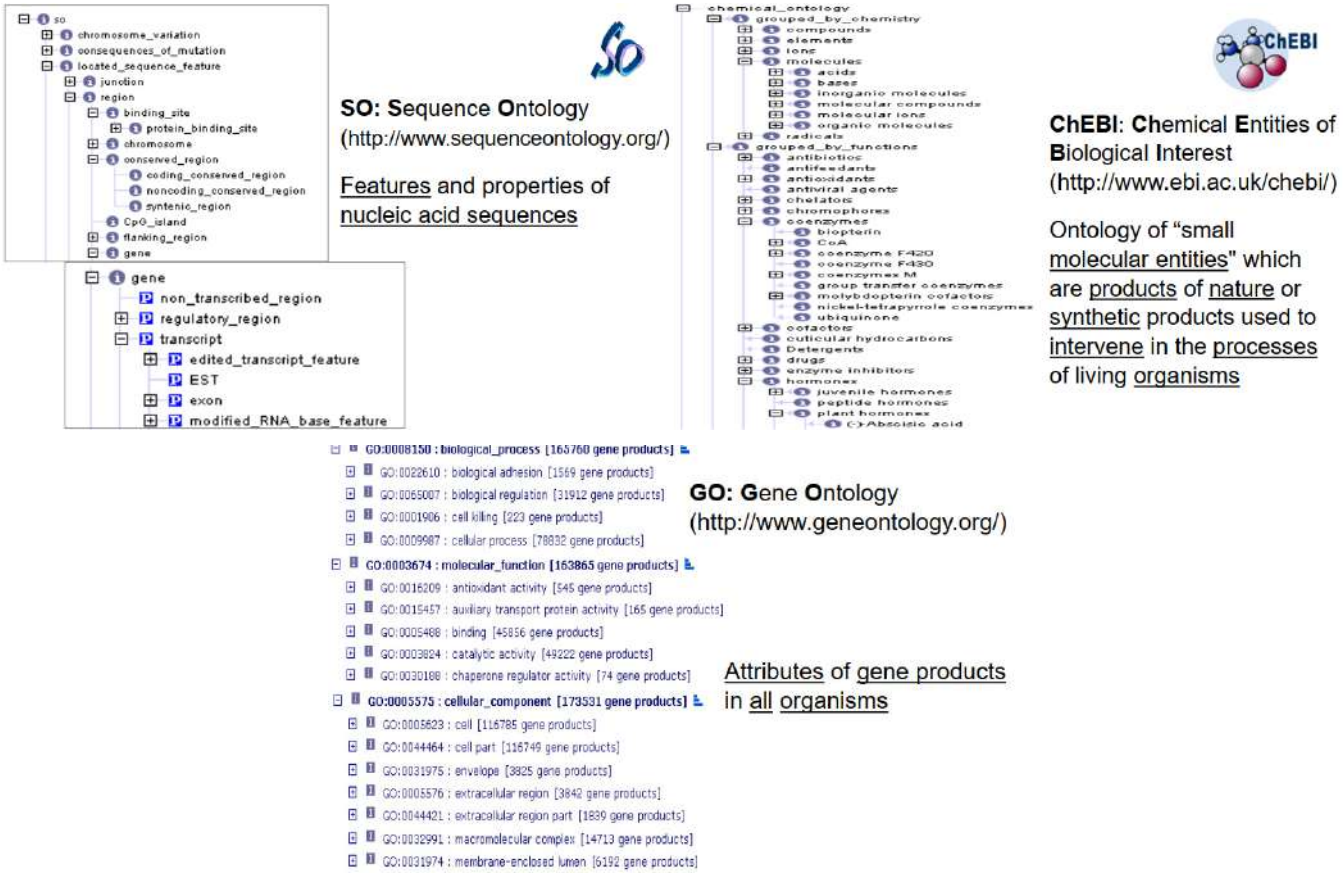
- CL: Cell Ontology (http://obofoundry.org/cgi-bin/detail.cgi?cell)
- **GO**: Gene Ontology (http://www.geneontology.org/)
- **FMA**: Fundational Model of Anatomy (http://fma.biostr.washington.edu/)
- ZAO: Zebrafish Anatomical Ontology (http://zfin.org/zf_info/anatomy/dict/sum.html)

Mature ontologies still in need of thorough *review*:

- **ChEBI**: Chemical Entities of Biological Interest (http://www.ebi.ac.uk/chebi/)
- DO: Disease Ontology (http://diseaseontology.sf.net/)
- PO: Plant Ontology (http://plantontology.org/)
- **SO**: Sequence Ontology (http://www.sequenceontology.org/)

Ontologies for which *early versions* exist:

- **OCI**: Ontology for Clinical Investigations (http://www.bioontology.org/wiki/index.php/CTO:Main_Page)
- **CARO**: Common Anatomy Reference Ontology (http://obofoundry.org/cgi-bin/detail.cgi?caro)
- EO: Environment Ontology (http://www.obofoundry.org/cgibin/detail.cgi?id=envo)
- **OBI**: Ontology for Biomedical Investigations (http://obi.sf.net/)
- **PATO**: Phenotypic Quality Ontology (http://www.obofoundry.org/cgi-bin/detail.cgi)
- PRO: Protein Ontology (http://pir.georgetown.edu/pro)
- RO: Relation Ontology (http://obofoundry.org/ro)
- RnaO: RNA Ontology (http://roc.bgsu.edu/)



VI.D The Gene Ontology

The Gene Ontology (GO) (http://www.geneontology.org/) project is the result of a *collaborative effort* to address the need for *consistent* and *species independent* descriptions of gene and protein features in distinct biomolecular databanks (e.g., gene involved in protein synthesis vs. translation)

Started in 1998 as a collaboration between *three model organism databases* and their curator labs: Drosophila Melanogaster Database (FlyBase - http://flybase.org/), Saccharomyces Genome Database (SGD http://www.yeastgenome.org/), and Mouse Genome Database (MGD - http://www.informatics.jax.org/)

The GO Consortium has now grown to include *several of the major repositories* for plant, animal, and microbial genomes (http://www.geneontology.org/GO.consortiumlist.shtml?all)

VI.D.1 Structure

The Gene Ontology (GO) is the bio-ontology most developed and used to *describe gene and protein features*

Provides **3 controlled terminologies** (terms, or categories, + relationships) to describe biological characteristics of genes and proteins (biological processes, molecular functions, cellular components)

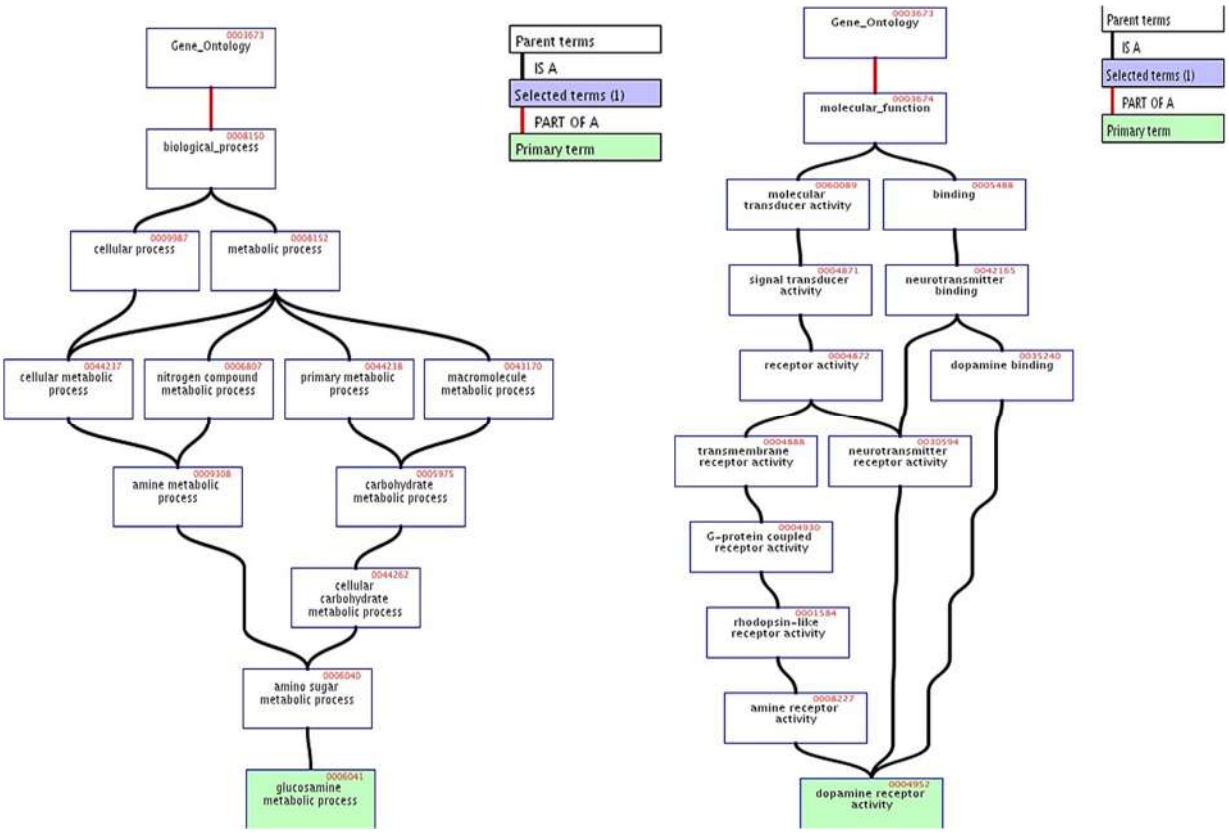
The GO has a **Directed Acyclic Graph** (DAG) structure:

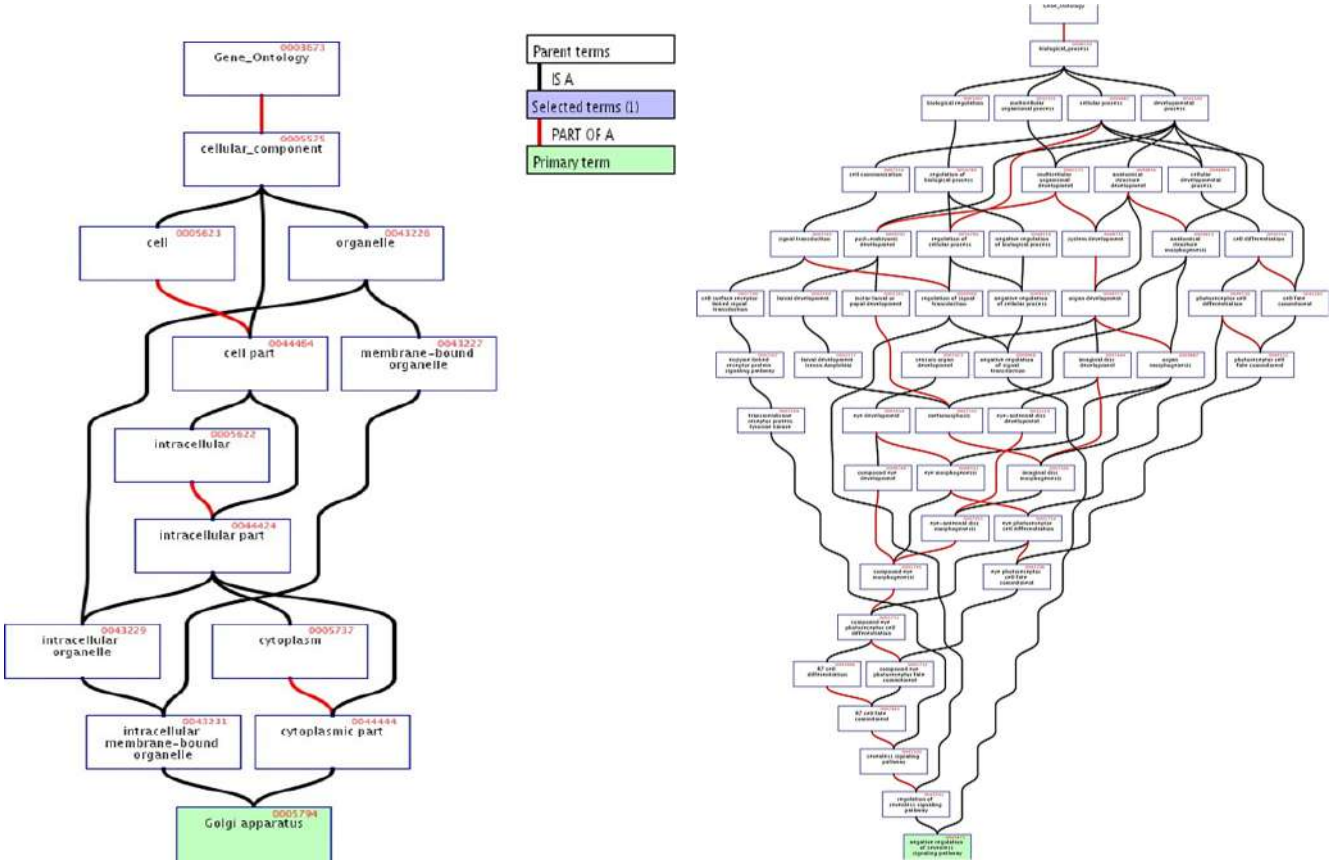
- Like a hierarchical tree (a child specifies the parent), but in a DAG a child node (term) can have *many parent* nodes (terms)
- Nodes are *connected* by oriented arches *without cycles*
- A node can be at *different levels* simultaneously
- Lower levels indicate *generality*, upper levels *specialization* of the represented concept
- Arches represent *relationships* between categories, mainly 2 types of relationships: `IS_A`, `PART_OF`

Each *GO concept* has associated:

- ID (unique and required)
- Name (unique and required)
- Definition (optional, soon required but unique)
- Synonyms (optional, can be more than one)
- Reference databases (optional, can be more than one)
- Relationships ("`IS_A`", "`PART_OF`", or others recently added, e.g., `REGULATES`)

VI.D.2 DAG examples





The concept of node level is not unique

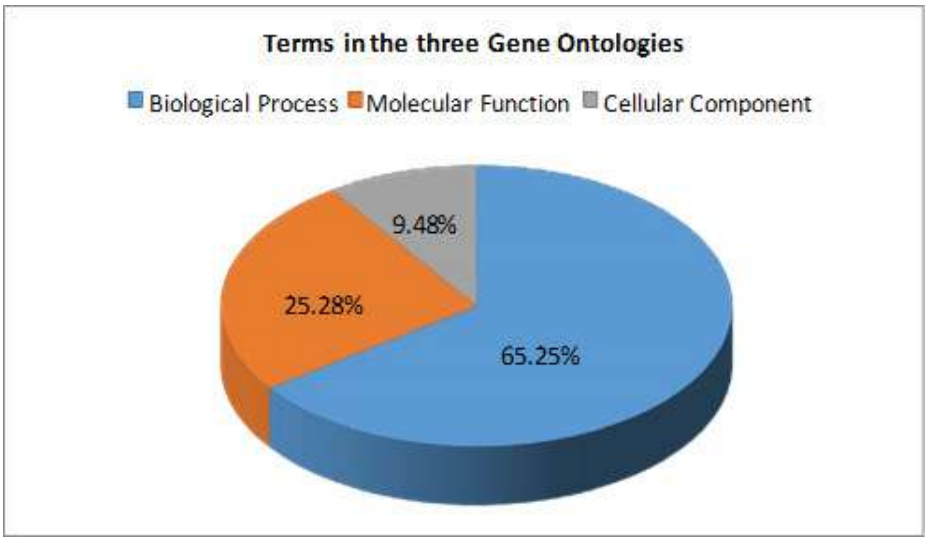
GO DAG structure reflects underlying biological complexity

VI.D.3 Statistics

As of October 31st, 2021, the GO includes (http://geneontology.org/stats.html):

- 43,832 terms, 100.00% with definitions, subdivided in:
  - 28,484 Biological Process (65.25%)
  - 11,166 Molecular Function (25.28%)
  - 4,182 Cellular Component (9.48%)

Additionally, there are 3,394 obsolete terms (7.74%) not included in the above statistics



For each one of the three Ontologies there are many Categories on different levels of the GO hierarchy

N. of Ontology Categories per level

level	Biological Process	Cellular Component	Molecular Function
0	1	1	1
1	22	15	18
2	277	152	176
3	2126	847	634
4	5350	1464	1383
5	8072	1789	3772
6	9185	1799	1685
7	8814	1679	1170
8	7475	1386	793
9	5364	1152	399
10	3421	832	166
11	1991	542	66
12	1067	314	24
13	567	167	38
14	308	36	12
15	147	3	0
16	66	0	0
17	18	0	0

The most developed ontology is the biological processes one

VI.D.4 GO browsers

Some GO browsers have been built to **help visualizing** the complex GO DAG structure:

- EMBL-EBI QuickGO (http://www.ebi.ac.uk/ego/)
- AmiGO (http://www.godatabase.org/cgi-bin/go.cgi)
- Mouse Genome Informatics (MGI) GO Browser http://www.informatics.jax.org/searches/GO_forms.html
- Expression Profiler GO Browser
  - http://www.bioinf.ebc.ee/EP/EP/
  - http://www.ebi.ac.uk/microarray-srv/EP/cgi-bin/ep_ui.pl

All offer *textual search* tools in the GO vocabularies and *graphic visualization* of the DAG of the retrieved terms

VI.D.5 Annotations

Annotation: **association** of a *gene* (or gene product) with a biomedical/biomolecular *concept* (term/concept)

Each gene (or gene product) is associated with *more terms* (has more features)

Each term (category) is assigned to *many genes* (or gene products) (many genes (gene products) have same features)

Annotations can be assigned in different ways:

- *Human* curated (assigned by experts)
- *Computationally* (automatically predicted, with or without human supervision (curation))

Ontological annotations must be assigned by associating genes (or gene products) with the **most specific terms** describing their features

When a gene (or gene product) is *annotated* (associated) to a GO term (i.e., identified as having the feature described by the term), it is **implicitly annotated** also to the parent terms (describing more general features) (true path rule, or *annotation unfolding*)

Note that:

- In an ontology, node (term) properties are *inherited from root to leaves* (i.e., specific concepts have the properties of their more generic concepts)
- Ontological annotations are *unfolded from leaves to root* (i.e., if a gene has a specific property, it has also the more general one)

Example (not biomolecular) of annotation and *annotation unfolding* vs. *term property inheritance*:

- Excerpt of the "car ontology":
  - Car; Sport car; Ferrari (a Ferrari is_a Sport car; a Sport car is_a Car)
- "Having a steering wheel" is a property of the concept identified by the term Car
  - For **property inheritance**, it is also a property of the concept identified by the term Ferrari
- Let X-0344-DS2134 be the chassis number of a Ferrari (annotation of the object with ID X-0344-DS2134 to the ontology term Ferrari)
  - For **annotation unfolding**, the object with chassis number X-0344-DS2134 is a car, in particular a sport car

Evidence (quality) codes exist for *each annotation* (http://www.geneontology.org/GO.evidence.shtml):

- *Automatically* assigned evidence codes:
  - IEA (Inferred from Electronic Annotation): annotation based on automatic computation (lowest quality)
- *Manually* assigned (curated) evidence codes:
  - Experimental: (e.g., EXP: Inferred from Experiment)
  - Computational analysis (e.g., ISA: Inferred from Sequence Alignment)
  - Author statement: (e.g., TAS: Traceable Author Statement)
  - Curatorial statement: (e.g., IC: Inferred by Curator)

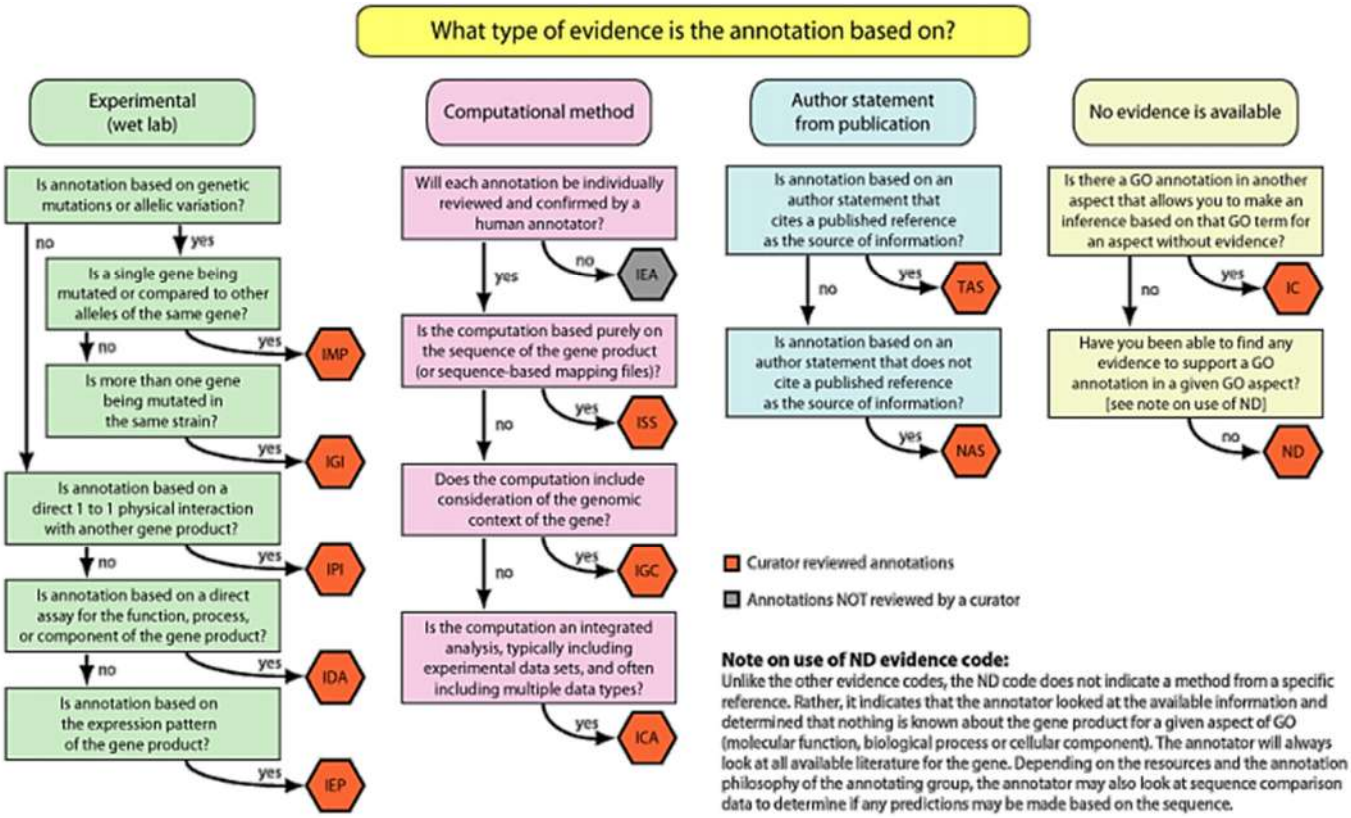
Manually assigned (curated) evidence codes:

- Experimental EXP: (Inferred from Experiment), better to use one of the more specific experimental codes:
  - IMP: (Inferred from Mutant Phenotype)
  - IGI: (Inferred from Genetic Interaction)
  - IPI: (Inferred from Physical Interaction)
  - IDA: (Inferred from Direct Assay)
  - IEP: (Inferred from Expression Pattern)
- Author statement:
  - TAS: (Traceable Author Statement)
  - NAS: (Non-traceable Author Statement)
- Computational analysis:
  - ISS: (Inferred from Sequence or structural Similarity)
    - ISA: (Inferred from Sequence Alignment)
    - ISO: (Inferred from Sequence Orthology)
    - ISM: (Inferred from Sequence Model)
  - IGC: (Inferred from Genomic Context)
  - RCA: (inferred from Reviewed Computational Analysis)
- Curatorial statement:
  - IC: (Inferred by Curator)
  - ND: (No biological Data available)

- Term (evidence) hierarchy:
  - TAS/IDA
  - IMP/IGI/IPI
  - ISS/IEP
  - NAS
  - IEA

GO also provides a "**modifier**" of the annotation, a "**qualifier**" field that can assume values such as: `CONTRIBUTES_TO`, `COLOCALIZES_WITH`, or `null`

But also "`NOT`" and `NOT_CONTRIBUTES_TO`! Be careful to also consider the qualifier field (useful also for ML algorithms)



GO Evidence Code Decision Tree

Each annotation is represented by a *record* including:

- Gene (or gene product) ID
- Gene Ontology ID
- Reference ID(s) (e.g. PubMed ID(s))
- Evidence code(s)
- Evidence modifier (Qualifier)

Protein ID	GO ID	Evidence code	PubMed ID	Qualifier
P05147	GO:0047519	IDA	PMID:2976880	null

or

Protein ID	GO ID	Evidence code	PubMed ID	Qualifier
P98194	GO:0005388	IDA	PMID:16192278	NOT

VI.D.6 Documentation

Gene Ontology documentation available at:

- http://www.geneontology.org/GO.contents.doc.shtml?all
- http://www.geneontology.org/GO.teaching.resources.shtml

Including:

- GO for newbies
- Introduction to the Gene Ontology Project
- Introduction to the Gene Ontology, AmiGO and GO website tutorial
- Introduction to GO and OBO
- Open Biomedical Ontologies
- The Gene Ontology and its insertion into UMLS
- …

VI.E Examples of bio-terminologies

Beside those that are part of the bio-ontologies (such as the Gene Ontology and the OBO ontologies), several *controlled vocabularies* constitute **bio-terminologies** used in some databases to describe:

- Pathways (e.g., in KEGG db)
- Protein families and domains (e.g., in Pfam, InterPro db)
- …

Several of these controlled vocabularies (e.g., KEGG pathways, …) are being enhanced with simple hierarchical (i.e., is-a) relationships, formally constituting an ontology (although with a simple structure)

NB: in bioinformatics, not a lot of bio-terminology exist anymore, because most of them have been transformed in ontologies

VI.F Unified Medical Languages System

Numerous different *controlled vocabularies* exist especially in the **biomedical** domain, besides in the biomolecular one:

- ICD-9, ICD-10
- ICD-9-CM
- CPT, HCPCS
- LOINC
- NDC
- SNOMED, SNOMED III, SNOMED-RT
- READ, SNOMED-CT
- MeSH
- MedDRA
- NCI Thesaurus
- …

Such *medical* and *clinical* terminologies were created in *different times* by different associations for *different purposes*, but:

- Some concepts they define are the same
- Often different controlled terms are used in each terminology to define the same (or similar) concept

For data and knowledge description, integration and interoperability matching are required. Therefore, matching textual descriptions (even controlled) is difficult

The **Unified Medical Language System** (UMLS) was created and is maintained (by the US National Library of Medicine) as a *support for integration* of biomedical textual *annotations* scattered in distinct databases

VI.F.1 Documentation

Resource documentation:

- UMLS: http://www.nlm.nih.gov/research/umls/
- UMLSKS: https://uts.nlm.nih.gov/home.html
- NLP and Lexical Tools: http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html

Tutorials:

- The Unified Medical Language System. What is it and how to use it? - 2004 (http://www.nlm.nih.gov/research/umls/presentations/2004 -medinfo_tut.pdf)
- Unified Medical Language System® (UMLS®) Basics 2007 (http://www.nlm.nih.gov/research/umls/pdf/UMLS_ Basics.pdf): more updated but not so clear; also online version (http://www.nlm.nih.gov/research/umls/user_ education/)
- **The Unified Medical Language System**: What is it and how to use it? - 2007, Bodenreider O. (http://mor.nlm.nih.gov/pubs/pres/20071204-KAISTtutorial.pdf): brief and clear

VI.G The UMLS – KAIST tutorial

VI.G.1 Introduction

Motivation:

- Started in 1986
- National Library of Medicine
- "Long-term R&D project"
- Complementary to IAIMS (Integrated Academic Information Management Systems)

« [...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information. The first is **the variety of ways the same concepts are expressed** in different machine-readable sources and by different people. The second is the **distribution** of useful information among many disparate databases and systems. »

The UMLS **in practice**:

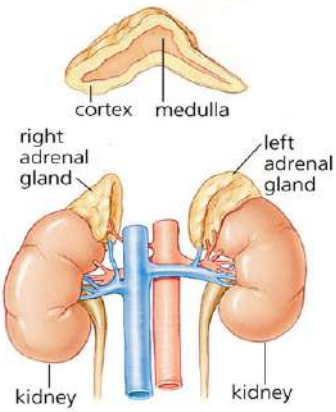
- Database:
  - Series of relational files
- Interfaces:
  - Web interface: Knowledge Source Server (UMLSKS)
  - Application programming interfaces (Java and XML-based)
- Applications
  - lvg (lexical programs)
  - MetamorphoSys (installation and customization)
  - RRF browser (browsing subsets)

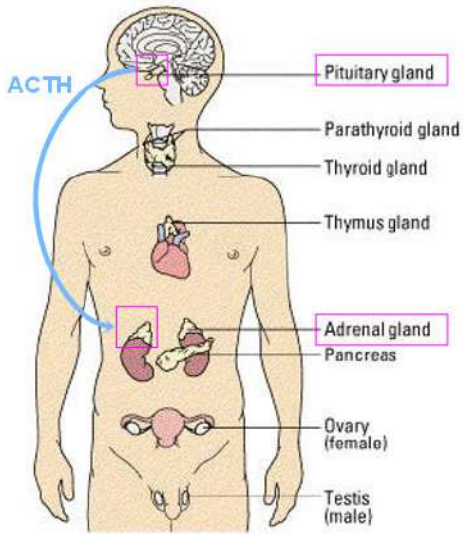
The UMLS is *not* an end-user application!

VI.G.2 What is the UMLS? Overview through an example

Example of Addison's disease

- Addison's disease is a rare *endocrine disorder*
- Addison's disease occurs when the *adrenal glands* do not produce enough of the hormone *cortisol*
- For this reason, the disease is sometimes called *chronic adrenal insufficiency*, or *hypocortisolism*





Clinical variants:

- Primary/Secondary
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland
- Acute / Chronic
- Isolated / Polyendocrine deficiency syndrome

Symptoms:

- Fatigue
- Weakness
- Low blood pressure
- Pigmentation of the skin (exposed and nonexposed parts of the body)
- …

In medical vocabularies:

Synonyms: different terms

- Addisonian syndrome — eponym
- Bronzed disease — symptoms
- Melasma addisonii — symptoms
- Asthenia pigmentosa — symptoms
- Primary adrenal deficiency — clinical variants
- Primary adrenal insufficiency — clinical variants
- Primary adrenocortical insufficiency — clinical variants
- Chronic adrenocortical insufficiency — clinical variants

Contexts: different hierarchies

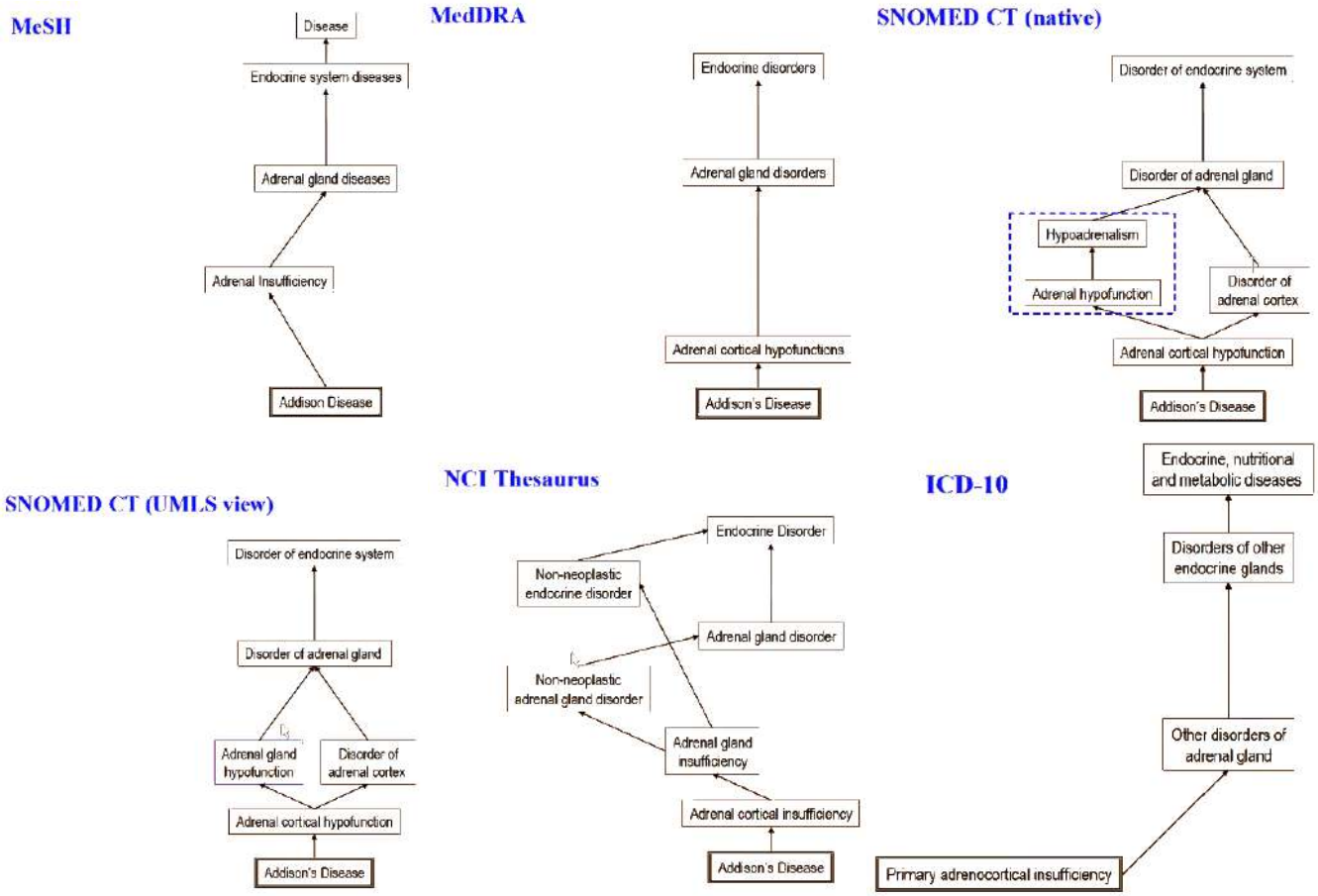
Organise terms:

- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

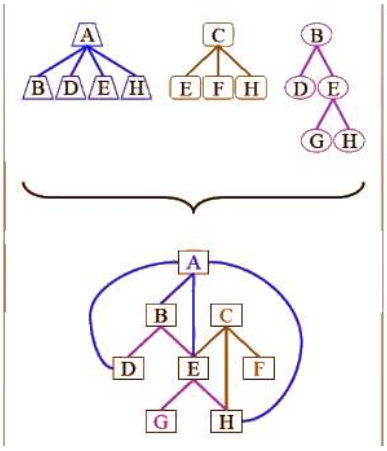
C0001403

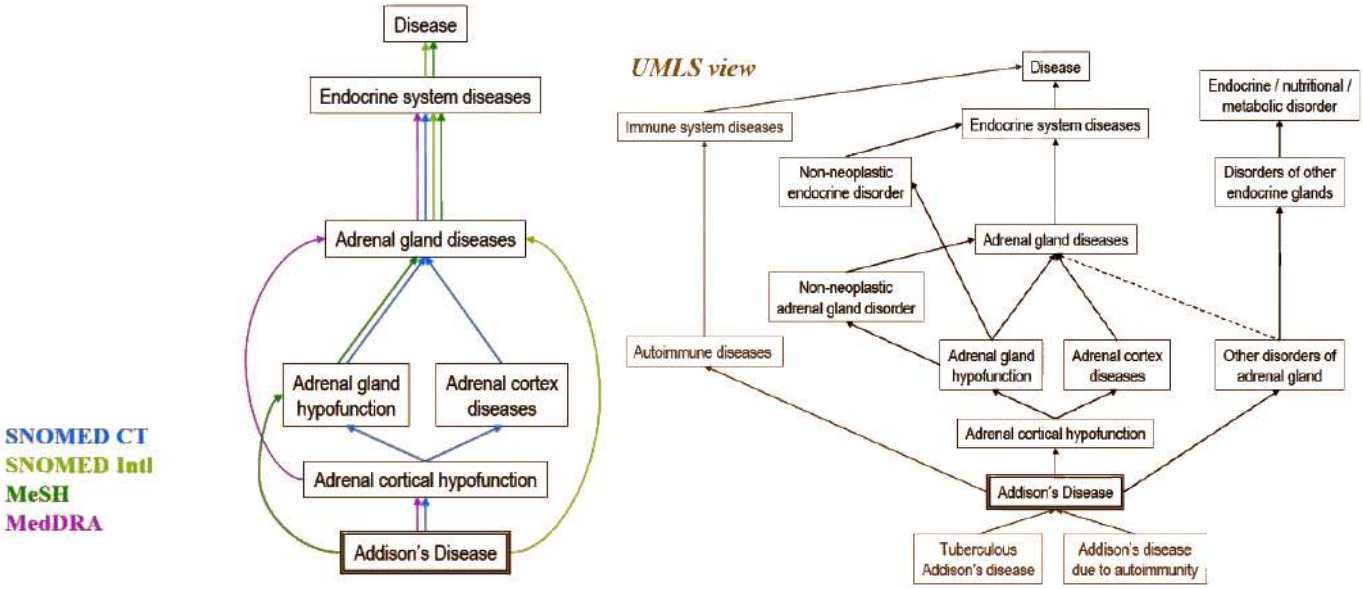
Addison's disease



Organise concepts :

- Inter-concept relationships: hierarchies from the source vocabularies
- Redundancy: multiple paths
- One graph instead of multiple trees (multiple inheritance)



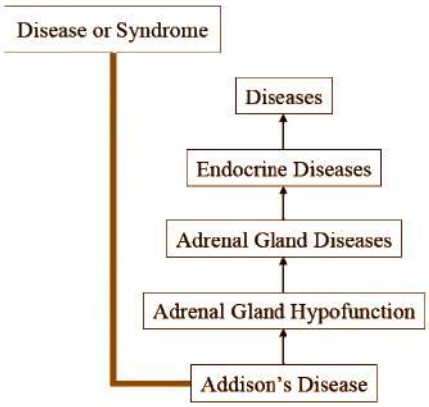


Relate to other concepts:

- Additional hierarchical relationships
  - link to other trees
  - make relationships explicit
- Non-hierarchical relationships
- Co-occurring concepts
- Mapping relationships

Categorise concepts

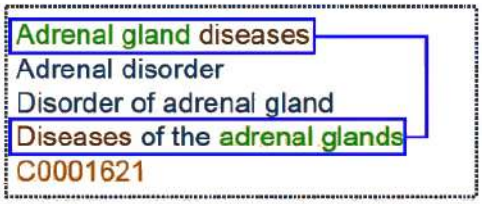
- High-level categories (semantic types)
- Assigned by the Metathesaurus editors
- Independently of the hierarchies in which these concepts are located



How do they do that?

- Lexical knowledge
- Semantic pre-processing
- UMLS editors

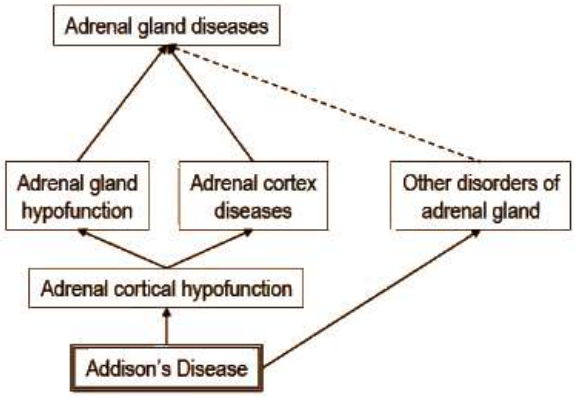
Lexical knowledge:



Semantic pre-processing

- Metadata in the source vocabularies
- Tentative categorization
- Positive (or negative) evidence for tentative synonymy relations based on lexical features

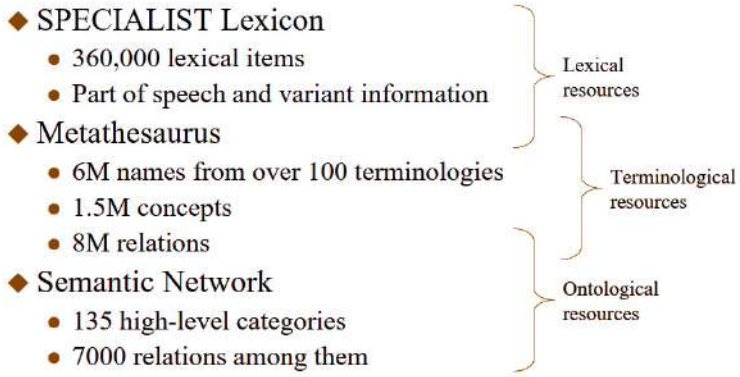
Additional knowledge: UMLS editors



UMLS Summary:

- Synonymous terms clustered into concepts
- Unique identifier
- Finer granularity
- Broader scope
- Additional hierarchical relationships
- Semantic categorization

VI.G.3 UMLS Metathesaurus



Metathesaurus basic organisation:

- Concepts
  - Synonymous terms are clustered into a concept
  - Properties are attached to concepts, e.g.,
    - Unique identifier
    - Definition
- Relations
  - Concepts are related to other concepts
  - Properties are attached to relations, e.g.,
    - Type of relationship
    - Source

Source vocabularies:

- 141 source vocabularies (17 languages)
- Broad coverage of biomedicine
  - 6.1M names
  - 1.5M concepts
  - 8M relations
- Common presentation

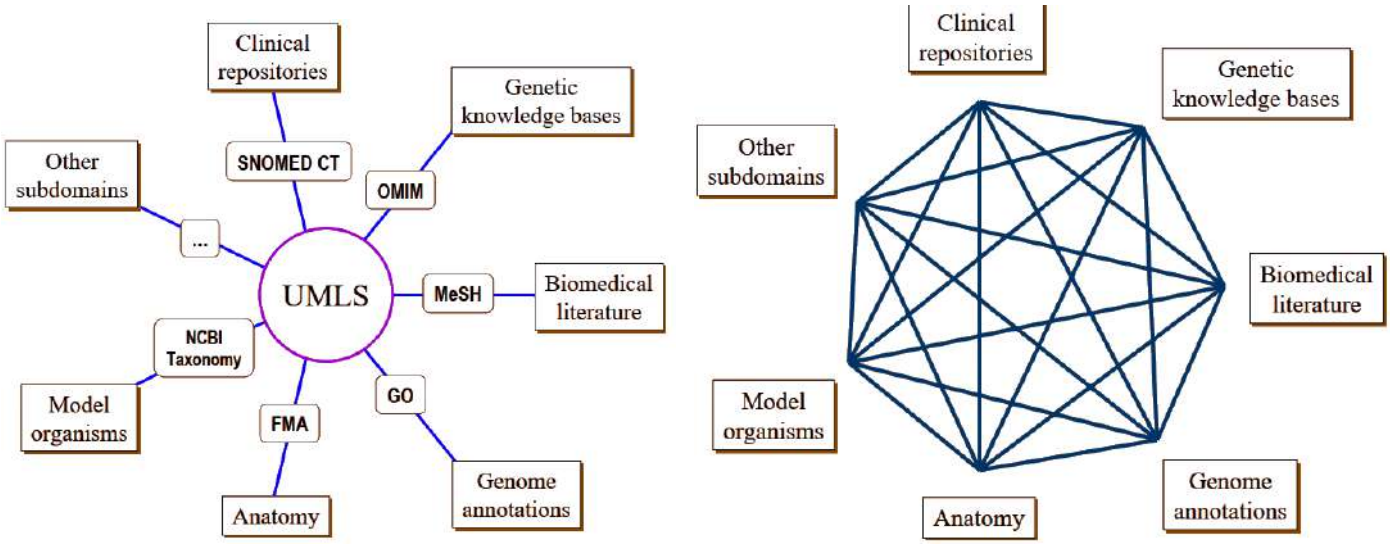
Biomedical terminologies:

- General vocabularies
  - anatomy (UWDA, Neuronames)
  - drugs (RxNorm, First DataBank, Micromedex)
  - medical devices (UMD, SPN)
- Several perspectives
  - clinical terms (SNOMED CT)
  - information sciences (MeSH, CRISP)
  - administrative terminologies (ICD-9-CM, CPT-4)
  - data exchange terminologies (HL7, LOINC)

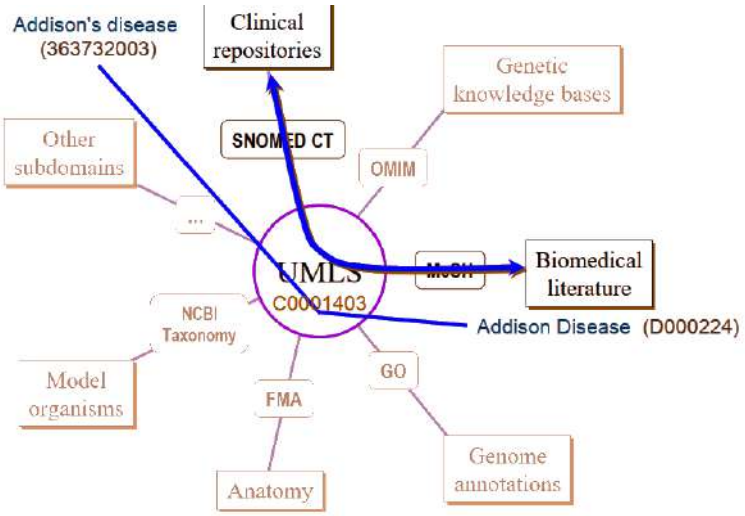
- Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - oncology (PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)
- Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)

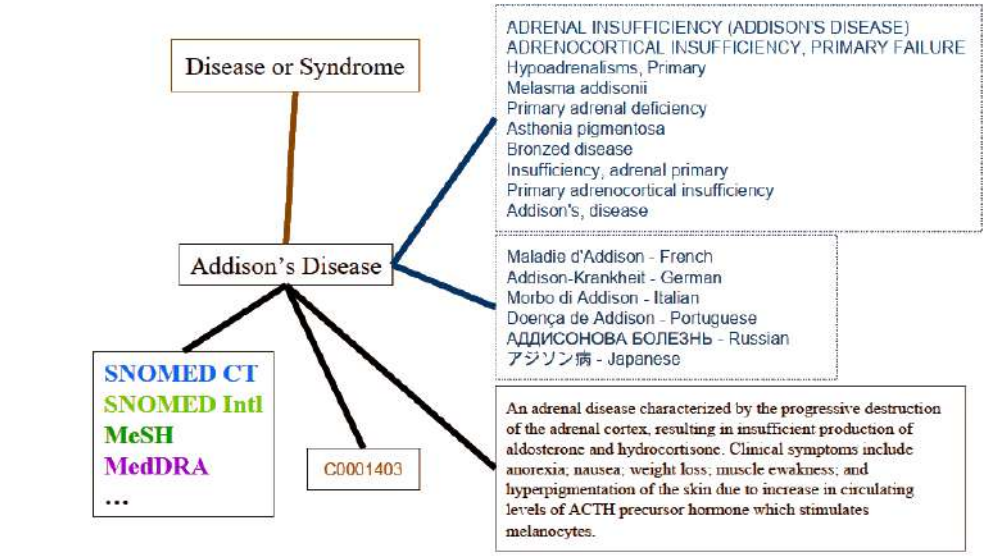
Integrating subdomains:

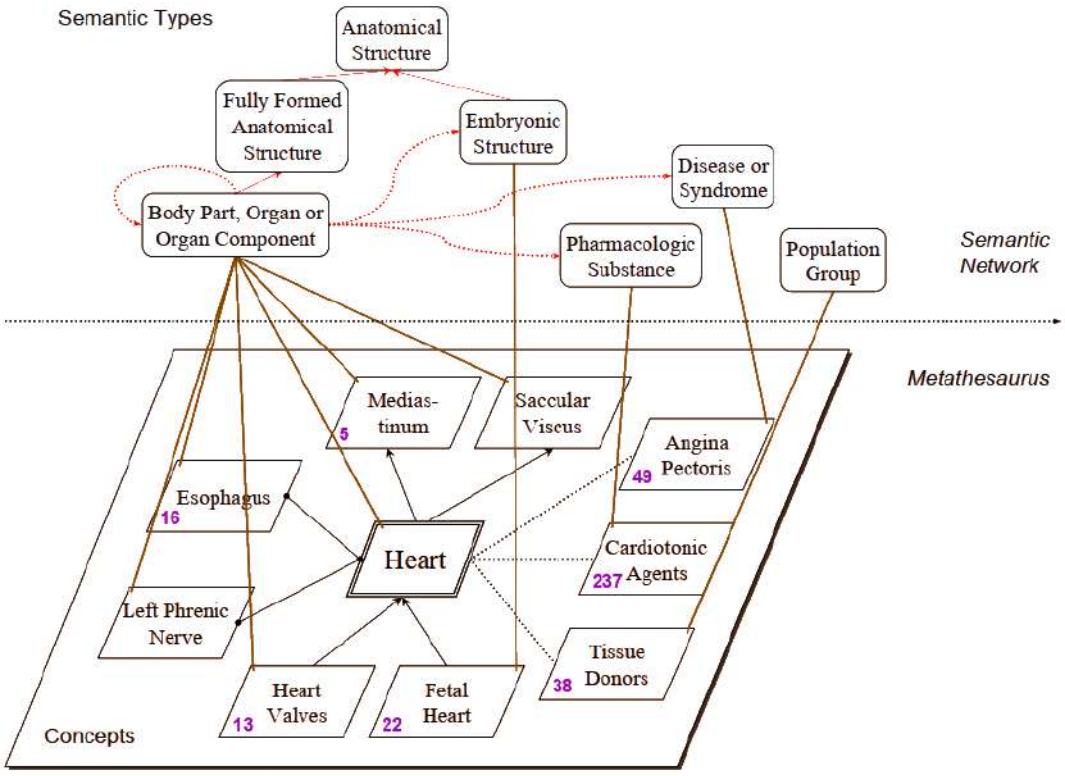


Trans-namespace integration:



Addison's Disease: Concept

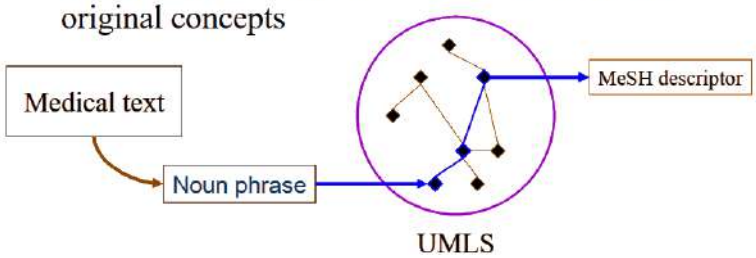




VI.G.4 How to use the UMLS? A UMLS-based algorithm

Indexing initiative:

- For noun phrases extracted from medical texts, map to UMLS concepts
- Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts



Restrict to MeSH:

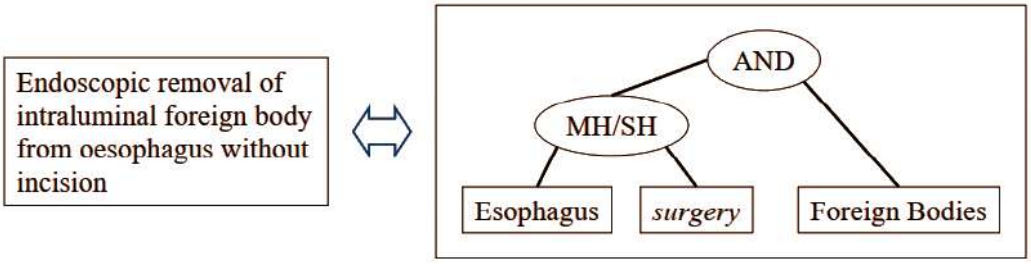
- Based on the principle of semantic locality
- Use different components of the UMLS
- 4 techniques of increasing aggressiveness
  - Use Synonymy MRCONSO
  - Use Associated expressions (ATXs) MRATX + MRREL
  - Explore the Ancestors MRREL + SN
  - Explore the Other related concepts MRREL + SN

Synonymy

- Term mapped to Source concept
- For this concept, is there a synonym term that comes from MeSH? (MRCONSO)

Associated expressions

- If not,
- Is there an associated expression (ATX) that describes this concept using a combination of
- MeSH descriptors? (MRATX/MRMAP + MRREL)



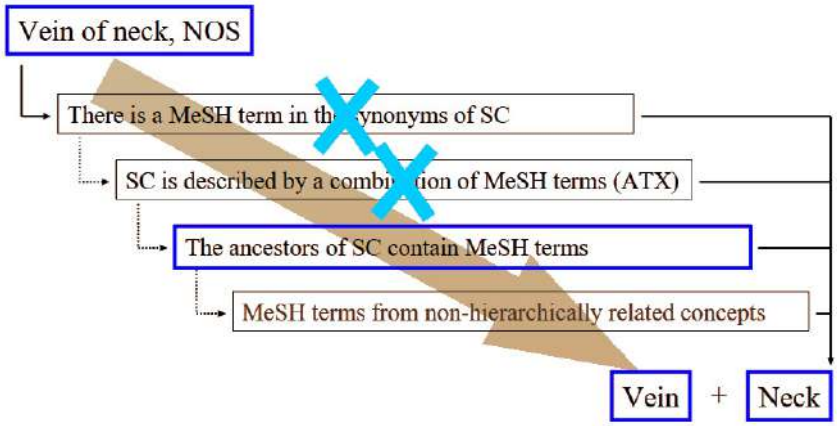
Ancestors

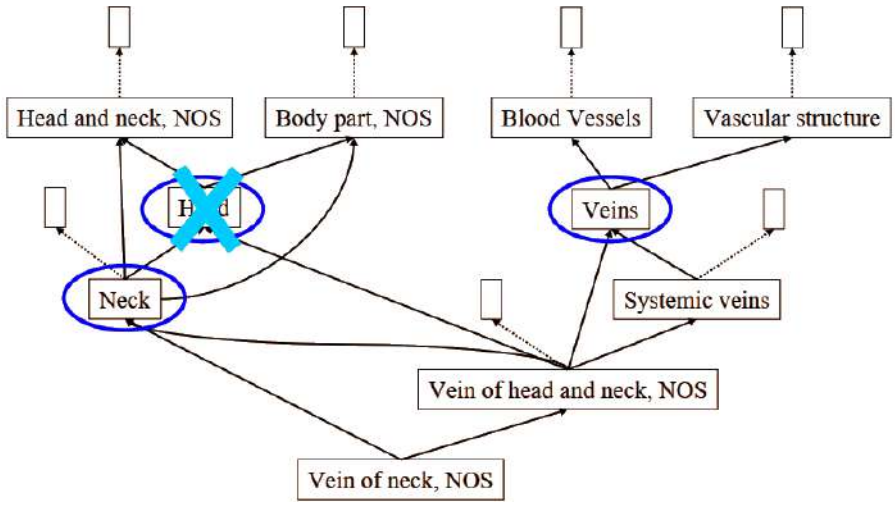
- If not, let us build the graph of the ancestors of this concept
  - using parents and broader concepts (MRREL)
  - all the way to the top
  - excluding ancestors whose semantic types are not compatible with those of the source concept (MRSTY)
- From the graph, select the concepts that come from MeSH (MRCONSO)
- Remove those that are ancestors of another concept coming from MeSH

Other related concepts

- If not, explore the other related concepts (MRREL) whose semantic types are compatible with those of the source concept (MRSTY)
- From those, select the concepts that come from MeSH (MRCONSO)

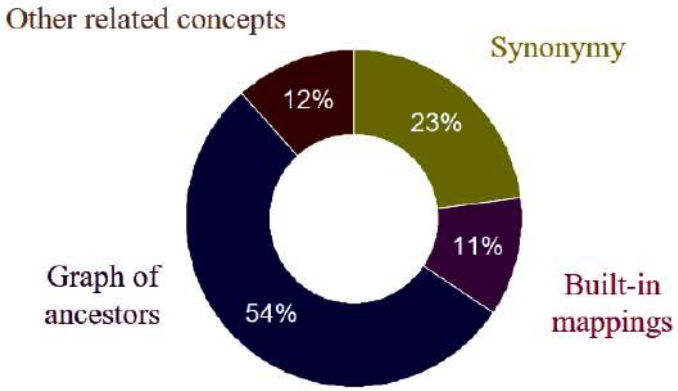
Example





Quantitative results

- 86% of UMLS concepts mapped to MeSH (2007)



Qualitative results

- Qualitative evaluation
  - 1,036 concepts extracted from 200 MEDLINE citations
  - manual review of every mapping or failure
- 61% Relevant
  - Subtotal Gastrectomy → Gastrectomy
  - Encephalopathy, NOS → Brain Diseases
- 28% More or less relevant
  - Vitamin A measurement → Laboratory Procedure
  - Swelling, NOS → Symptoms
- 11% Non relevant

References: UMLS documentation

- UMLS home page http:// www.nlm.nih.gov/research/umls/
- UMLS documentation
  - Formerly know as the “Green Book”
  - Now online documentation
  - http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML

VII. Biomolecular databanks

VII.A Introduction

VII.A.1 Genomic data

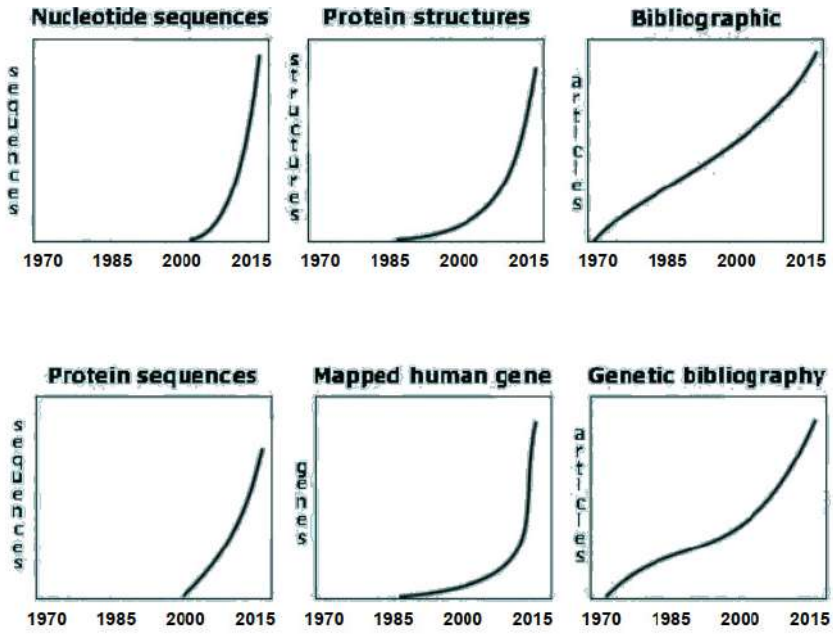
What are **genomic data**?

- All the *information* (structural and functional) that we have at molecular level on living organisms
- Mainly obtained by means of *molecular biology experiments*

VII.A.2 Biomolecular data production

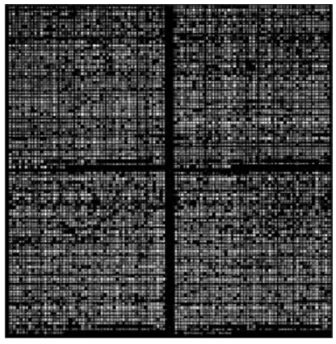
Today many public and private research groups are working in sequencing and *analysing the genomes* of many organisms

New **automatic sequencing** and high-throughput analysis techniques (e.g., microarrays) produce huge amount of data. *Automatic annotations* enable to have **homogeneous genomic** data on which subsequently applying consistent analysis strategies, obtaining comparable results



Different slopes: it can be that part of the information that is available is not of good quality, and does not make it through the literature. On the other hand, just by analysing measures already produced, we can extract many new information.

VII.A.3 Biomolecular data types



Genetic sequences, from raw trace files to base-calls, to protein

Microarrays (gene expression, SNP, …), from pictures to interpretation

Sample annotations

Patient diagnostics:

- Karyotype
- Fluorescent In Situ Hybridization
- Polymorphisms

- **Genetic sequences**

Though the trace files are **large**, the readings take up much less space

- *FASTA*: simple text file format consisting of a header line beginning with a greater than (>) symbol followed by a sequence of one letter base or amino acid codes
- Lowest common denominator between proprietary systems

- The entire genome can be downloaded in FASTA format

```
>TC30326 s1 TC63997 TC16407 TC21735 TC23192 TC30327 TC50687 TC59470
GAGCCTCTGGGTCCCGTCTAGGTACACTTTCTGCATTTCGAGCCCGGGCAGGTGAGGTGCGACAGGTAAATTTAAC
ACAATGGATTTCTCCAAGCTACCCAAAATCCGAGATGAGGATAAAGAAAGTACATTTGGTTATGTGCATGGAGTCTC
AGGGCCTGTGGTTACAGCCTGTGACATGGCGGGCGCTGCCATGTACGAGCTGGTGAGAGTGGGGCACAGCGAGC
TGGTTGGAGAAATTATTCGATTGGAAGGTGACATGGCCACCATTCAGGTGTATGAAGAAACTTCTGGTGTCTCTGTT
GGAGACCCCGTACTCCGCACTGGTAAACCTCTCTCGGTCGAGCTGGGTCCCGGGATTATGGGAGCCATTTTTGATG
GTATACAGAGACCTCTGTCGGATATCAGCAGTCAGACCCAAAGTATCTACATCCCCAGAGGAGTCAATGTGTCTGCT
CTCAGCAGAGATATCAAATGGGAGTTTATACCCAGCAAAAACCTACGGGTTGGTAGTCATATCACTGGTGGAGACAT
TTATGGGATTGTCAATGAGAACTCCCTCATCAAACACAAAATCATGTTGCCCCCACGTAACAGAGGAAGCGTGACTT
ACATCGCGCCGCCTGGGAATTATGATGCATCCGATGTCGTCCTGGAGCTTGAGTTTGAAGGTGTGAAGGAGAAGTT
CAGCATGGTCCAAGTGTGGCCTGTGCGGCAGGT
```

- **Microarrays**

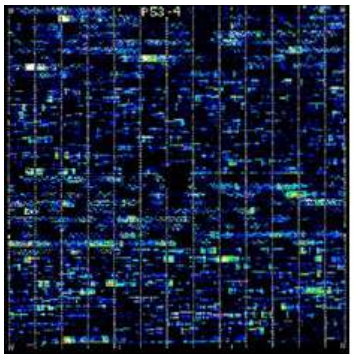
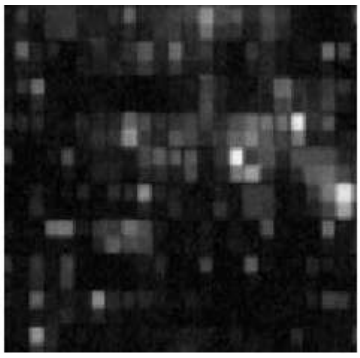
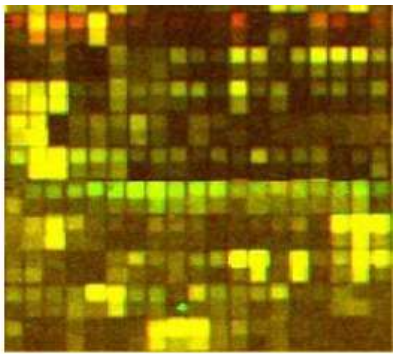
Raw TIFF images from a single microarray take 10-100 MB

File of expression measurements is 0.5-1 MB

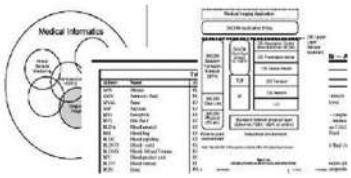
Each experiment of differential gene expression is made of 2-3 *replica* of test and control microarrays: 4-6 microarrays

MIAME: Minimum Information About Microarray Experiment http://www.mged.org/Workgroups/MIAME/miame.html

MGED: Microarray Gene Expression Database http://www.mged.org/

- **Sample annotations**



How to *describe* the *context* of the measured sample?

- The least common denominator
- Equivalent to the medical records problem

- **Specific biomolecular data types**
  - Nucleotide sequences
  - Genomic mapping data
  - Expression profiles (2D-SDS PAGE, DNA chips)
  - Protein sequences
  - 3D Structures of nucleic acids and proteins
  - Transcription and genotyping data
  - Metabolic data
  - Functional and phenotypic annotations
  - Bibliographic information

VII.B Biomolecular databanks

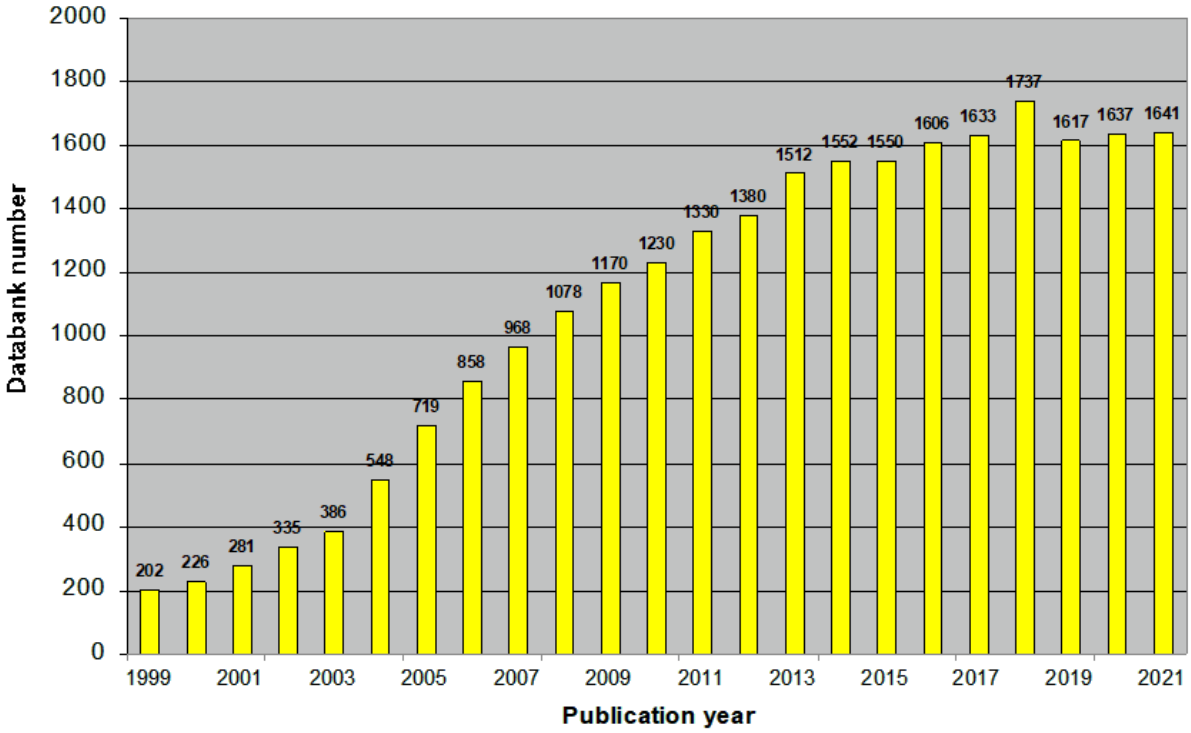
VII.B.1 Growth

Biomolecular data and information are stored in **databanks**, mostly *public* and *freely accessible* through the Internet

Since 1994, every year Nucleic Acids Research publishes an issue dedicated to molecular biology databanks. It includes a list of freely available key databanks, with a brief description and the databank URL

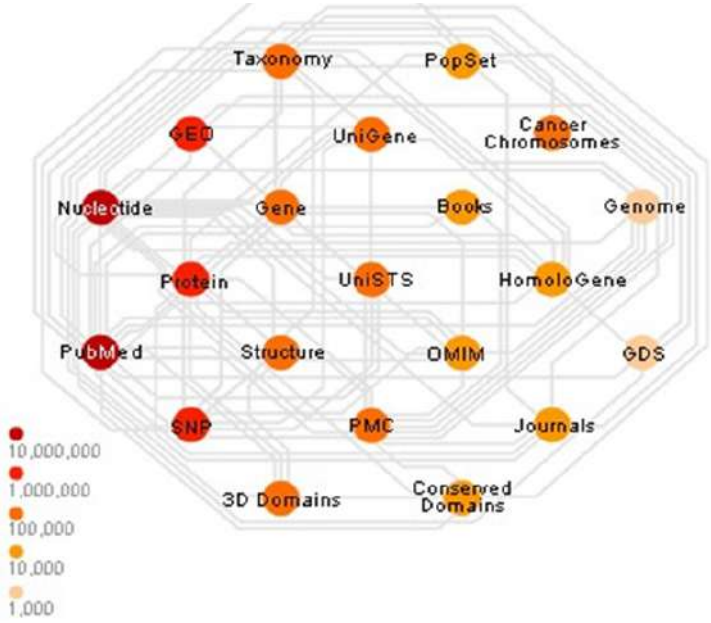
The 2021 update lists 1,641 databanks, with 90 new and 86 eliminated for discontinued URL with respect to the 2020 edition (https://academic.oup.com/nar/issue/49/D1). This number is very high, and despite that some databanks contain replica, most of them are specific to different biomolecular domain.

- Corresponding open access paper is: Rigden DJ, Fernández XM. The 27th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. Nucleic Acids Res. 2021; 49(D1): D1-D9 (https://academic.oup.com/nar/article/49/D1/D1/6059975)

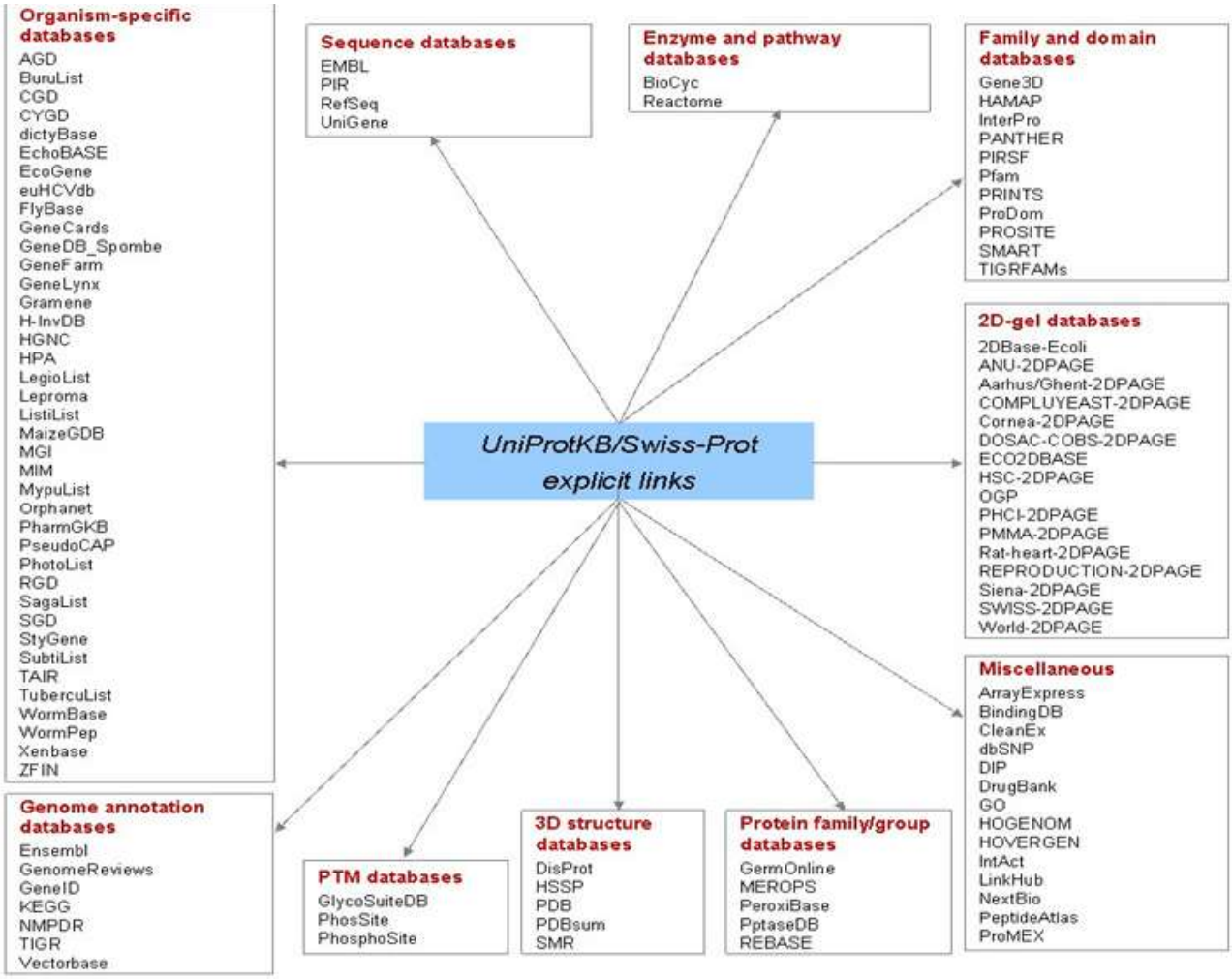


VII.B.2 Interoperability and cross referencing

It is well managed if an institution is responsible for several databanks (e.g., the NCBI below)



Entrez databanks



VII.B.3 Types

Most of biomolecular databanks are *public* and *freely accessible* through internet. They can be subdivided in:

- Primary databanks (DNA, RNA, proteins, …)
- Derivative or specialized databanks (EST, STS, SNP, RNA, genomes, microarray data, protein families and domains, pathways, genetic disorders, ...)

- **Primary**

Databanks of *nucleic* and *amino acid* sequences are defined as **primary** databanks because they contain only *generic* information. This is the *minimal information* to be associate with the sequence in order to identify it from the point of view of specie-function.

Each sequence introduced in a databank with its annotation constitutes an "*entry*" and is identified by an ID or accession number

Two main classes:

- DNA (nucleic acids) databanks, including:
  - EMBL at EBI (Europe - UK) http://www.ebi.ac.uk/embl.html
  - GenBank at NCBI (US) http://www.ncbi.nlm.nih.gov/
  - DDBJ (Japan) http://www.ddbj.nig.ac.jp/
- Protein (amino acids) databanks, including:
  - UniProt (The Universal Protein Resource) http://www.uniprot.org/
    - Swiss-Prot/TrEMBL (high level of annotation) http://www.expasy.org/sprot/
    - PIR (Protein Information Resource) http://pir.georgetown.edu/

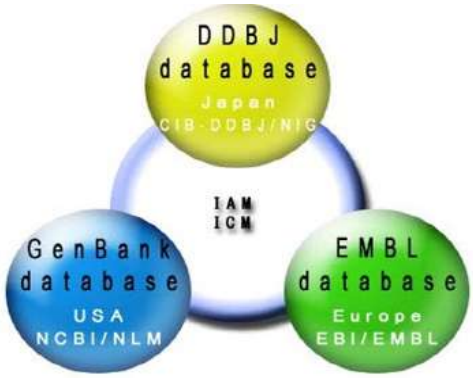
The first databank of nucleic acid sequences, created in 1980, is the *European Molecular Biology Laboratory* (EMBL) Data Library (http://www.ebi.ac.uk/embl/) constituted in the homonym laboratory in Heidelberg in Germany.

In 1982 was created GenBank, the American databank (http://www.ncbi.nlm.nih.gov/Genbank/), with a data format different from the EMBL and developed in parallel with it

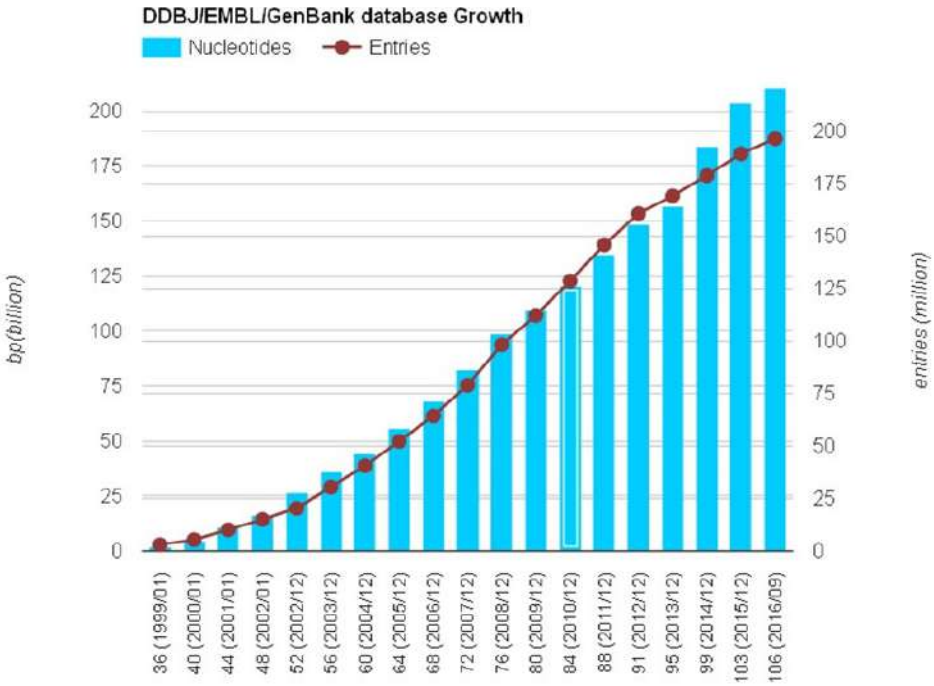
In 1986 was created DDBJ, the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/)

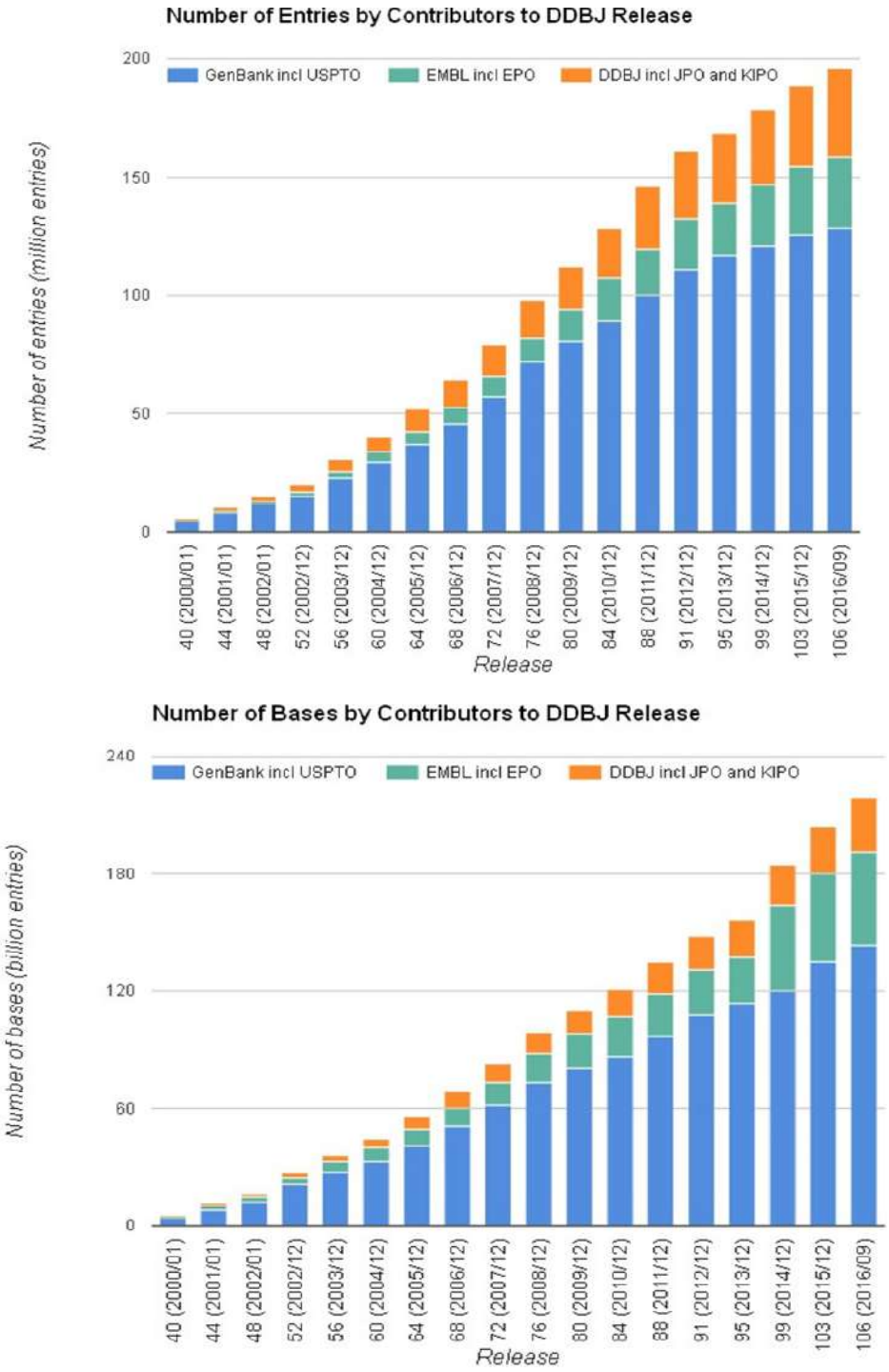
The three major primary databanks joined the **International Nucleotide Sequence Database Collaboration** that promotes the following projects:

- The Taxonomy Project, one of its main goals is using a unified taxonomy in all three databanks
- The Feature Table, identifying a set of information to associate to each sequence and the mechanism of *data exchange*



IAM: International Advisory Meeting - ICM: International Collaborative Meeting





- **Specialized**

The **specialized databanks** collect sets of homogeneous data from the taxonomic and/or functional point of view. These data, available in the primary databanks and/or in the literature, are revised and annotated with *added value information.*

The specialized databanks can be:

- *Human* curated (e.g., Entrez Gene, Swiss-Prot, NCBI RefSeq mRNA)
- *Computationally* derived (e.g., UniGene)
- A combination of *both* (e.g., NCBI Genome Assembly)

Specialized databanks can be classified as:

- A simple **subset of the primary databank** data, homogeneous from the biological point of view, **accurately revised** and enhanced with specific biological information inherent to the considered subset
  - A good example is the PIR Sequence-Structure databank (PIR-NRL3D) (Pattabiraman N. et al., 1990) PIR-NRL3D is a databank of proteins, derived from the Protein Information Resource (PIR) databank, with a known 3D structure and whose atomic coordinates are memorized in the Protein Data Bank (PDB)
- A set of **homologous sequences multi-aligned**, such as:
  - rRNA compilation databank (Neefs JM. et al., 1993) (http://rrna.uia.ac.be/)
  - tRNA compilation databank (Steinberg S. et al., 1993)
- A set of **specific information**, complementary of those in the primary databanks, and specific for a well-defined class of sequences
  - A good example for this class is the Eukaryotic Promoter Databank (EPD) (Cavin Périer R. et al., 1998) (http://www.epd.isb-sib.ch/)

Genomic databanks, representative of the whole set of information derived from mapping and sequencing projects of the Human Genome and of other Genomes selected as Model Organisms. A good example are the Genome Data Base (GDB) (http://gdbwww.gdb.org/), or the Mouse Genome Informatics (http://www.informatics.jax.org/)

Integrative databanks, created to collect information dispersedly stored in other specialized databanks. Good examples are:

- GeneCards (http://bioinformatics.weizmann.ac.il/cards/)
- SOURCE (http://source.stanford.edu/)

VII.B.4 Databank main features to be considered

Scientific community *acknowledgment*

Building procedures, components: curated vs. computationally inferred

Content provided:

- Data:
  - Semantic types, organisms
  - Annotations
  - Cross-references
  - Updating frequency
  - Statistics
- Query and analysis services: query options and response time

Access (Web, FTP, Web service): data format and dimension

VII.B.5 Selected biomolecular databanks

Primary DBs
- EMBL-EBI
- GenBank
- DDBJ

Sequence DBs
- UniGene
- RefSeq
- UCSC
- Ensemble

Genomic DBs
- GDB

Gene DBs
- **Entrez Gene**
- OmoloGene

Protein DBs
- **UniProt**
- Swiss-Prot
- TrEMBL
- PIR

Protein 3D structure DBs
- PDB

Protein domain DBs
- InterPro

Patway DBs
- **KEGG**
- Reactome

Gene Ontology Annot. DBs
- GOA

Disorders DBs
- **OMIM**
- GAD

Mutation DBs
- **dbSNPs**

Microarray DBs
- SMD
- GEO
- **Array Express**

Integrative DBs
- SOURCE
- GeneCards

Literature DBs
- **PubMed**

EMBL: http://www.ebi.ac.uk/embl/
GenBank: http://www.ncbi.nlm.nih.gov/GenBank/
DDJB: http://www.ddbj.nig.ac.jp/
UniGene: http://www.ncbi.nlm.nih.gov/UniGene/
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq/
UCSC: http://genome.ucsc.edu/
GDB: http://www.gdb.org/
Ensemble: http://www.ensembl.org/
Entrez Gene: http://www.ncbi.nih.gov/gene
HomoloGene: http://www.ncbi.nlm.nih.gov/HomoloGene/
UniProt: http://www.pir.uniprot.org/
Swiss-Prot: http://www.expasy.ch/sprot/
TrEMBL: http://www.ebi.ac.uk/trembl/
PIR: http://www-nbrf.georgetown.edu/pirwww/search/
InterPro: http://www.ebi.ac.uk/interpro/

PDB: http://www.rcsb.org/pdb/
KEGG: http://www.genome.ad.jp/kegg/
Reactome: http://www.reactome.com/
GOA: http://www.ebi.ac.uk/GOA/
OMIM: http://www.ncbi.nlm.nih.gov/Omim/
GAD: http://geneticassociationdb.nih.gov/
dbSNPs: http://www.ncbi.nlm.nih.gov/snp
SMD: http://genome-www5.stanford.edu/Microarray/
GEO: http://www.ncbi.nlm.nih.gov/geo/
Array Express: http://www.ebi.ac.uk/arrayexpress/
SOURCE: http://source.stanford.edu/
GeneCards: http://bioinformatics.weizmann.ac.il/cards/
Harvester: http://harvester.embl.de/
PubMed: http://www.ncbi.nlm.nih.gov/pubmed/

Of each *main* databank, you should know **building procedures** (curated vs. computationally inferred) and **content provided** (data types, main organisms, updating frequency))

Example IDs (table of conversion is accessible):

GenBank accession number	UniGene cluster ID	Entrez Gene ID	Swiss-Prot / UniProt accession number / ID	PIR accession	PDB ID
H59260	Hs.1634	993	Q16719	A41648	1C25
H72122	Hs.104925	8507	P30304	A48157	1AH9
H87471	Hs.169139	8942	P09581	I38238	1C04
R43509	Hs.75251	8554	P30291	I53908	2RGF
W96134	Hs.78465	3725	Q14703	JC5517	3EZA
AA039640	Hs.75188	7465	O95644	S10404	4HHB
AA047413	Hs.55606	7571	P28352	S12008	5TMP
AA158990	Hs.80680	9961	P48307	S51342	7ENL
AA399473	Hs.295944	7980	P48431	S55048	9INS
AA447393	Hs.75890	8720	P05412	T04859	13PK

In the case of proteins (right), it is easier to find the correspondence in the table, because UniProt plays a reference role and contains other identifiers

VII.C Issues in effective using the provided data

VII.C.1 Scenario and users' needs

Biomolecular sequence data and annotations describing individual genes and their encoded protein products continue to accumulate in many **different databanks**

Gene and protein databanks are accessible in **different ways**

At present, these ways are mainly *not functional* to efficiently use comprehensively the provided annotations for easily studying lists of genes, e.g., produced by means of high-throughput experiments

VII.C.2 Access types

Access through **Web interface** (HTML):

- Most *common* provided access
- Usually *unstructured* information
- *Heterogeneous* Web interfaces
- Information organized *per single sequence* (gene or protein)
- *Query results* on single biomolecular sequence are mainly returned in *HTML format*
- Requires *time* to comprehensively query *multiple databanks*

Access through **Web service**:

- Available only for a *few* databanks
- Usually designed for *specific queries* regarding a *limited number of IDs*
- Generally, require *informatics skill* and service composition/integration, e.g., through systems like
  - Taverna (http://www.taverna.org.uk/),
  - Galaxy (http://galaxy.psu.edu/),
  - SeCo (http://http://www.searchcomputing.deib.polimi.it/)

Access through **FTP server**:

- Requires having *technological and informatics skilled* human resources for re-implementing locally the databank, which become obsolete soon
- Sometime there are *no explicit relations* among provided data (ASCII flat file format)

Direct access:

- Rarely allowed for *security issues*
- Databank *schemas* are *heterogeneous* and unknown a priori
- *Query languages* differ among databanks
- *Lack of a common vocabulary*, which limits interoperability

Direct HTTP linking to a databank is generally available, if the databank entry identification code/s is/are known

- Each link (URL) returns a *Web page* (usually in HTML format) with all data available in the databank for the considered entry
- Examples of direct links to databanks are:
  - UniGene: http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ACC=XXXX with XXXX the GenBank accession number code for the entry (e.g., M27396)
  - PDB: http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=XXXX with XXXX the four letter identification code for the entry (e.g., http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=2cpk)

VII.C.3 Information extraction requirements

Biomedical researchers need to have in **aggregated form** the *genomic* and *proteomic* data they need for their sets of genes/proteins in order to browse them easily and *perform complex queries* on them to highlight relevant information

Despite efforts to integrate *gene annotations*, relevant gene and protein data are still sparsely stored among **heterogeneous databanks**

The **increasing amount of information** available requires new approaches to integrate, summarize, visualize, and compare the gene and protein annotations in order to make possible *discovering new knowledge*

VII.C.4 Interrogation/search difficulties

The effective use of the **huge amount of data** available in biomolecular databanks presents several difficulties:

- The data are stored in *distinct* databanks
- The databanks:
  - are *heterogeneous* in schema and contents
  - generally can be interrogated only for a *single biomolecular sequence* (i.e. gene or protein) at a time are mostly accessible for interrogation via *Web only*
- The data retrieved as *interrogation*/search results are often available, not structured, in *HTML format* only

VII.C.5 Possible solutions and example tools

For the databanks with access through **FTP server**, solutions to the interrogation difficulties can be:

- Creating *local databases* (i.e., mirrors) associated to the original databanks
- Drawbacks: keeping *updated*, multiple database issues

For the rare databanks with **direct access** (and for those locally mirrored):

- Designing and using *special query languages* to access and query data in *multiple databases* of heterogeneous DBMS; definition and use of *metadata*
- *Automatic mapping* of queries, to answer the need of performing the same query on several databases

For databanks providing access through a **Web interface**, solutions to the interrogation difficulties reside in creating/using tools allowing to:

- *Automatically* wrap and extract *specific data* of interest in HTML pages of different databanks
- Store in *aggregated form* the extracted data
- *Structure* the aggregated data to enable performing subsequent *specific queries* on them

For access through **Web services** (the most useful way from the informatics point of view, to get structured information), solutions regard the development and usage of a *graphical environment* for Web service integration, composition, orchestration and *workflow design* and execution (e.g., similar to Taverna, with better service result integration support and graphical interface)

At "Politecnico di Milano", we developed some tools to effectively use available gene and protein annotations:

- MyWEST: My *Web Extraction Software Tool*, for the automatic extraction of data about several genes from *multiple HTML pages* (http://www.bioinformatics.deib.polimi.it/MyWEST/)
- GFINDer: *Genome Function INtegrated Discoverer* for the statistical functional analysis of different groups of genes (http://www.bioinformatics.deib.polimi.it/GFINDer/)
- GPDW: *Genomic and Proteomic Data Warehouse*, with all the *automatic procedures* for creation and updating of an integrated data warehouse of many genomic and proteomic *annotations* (http://www.bioinformatics.deib.polimi.it/GPKB/). To add a new datatype, just needed to add a module and specify its relation to the previous modules, way easier than the previous one
- Bio-SeCo: *Biomedical Search Computing*, in the context of the Search Computing project (http://http://www.searchcomputing.deib.polimi.it/), focused on *building the answers to complex multi-topic search queries* involving *ranking composition* like "Which genes encode proteins in different organisms with the highest sequence similarity to a given protein and are co-expressed (e.g. over expressed) in the same given tissue?" by interacting with and *integrating* a constellation of available cooperating search services, using *ranking* and *joining* of results as the dominant factors for service composition. Available at http://www.bioinformatics.deib.polimi.it/bio-seco/seco/

VII.D Data integration

Data Integration in the Life Science:

- Several approaches in data integration:
  - Indexed data sources
  - Multi- or Federated databases
  - Data warehousing
  - Mediator based systems
  - LAV and GAV
  - Memory-mapped data structures
- Access tools for:
  - Browsing
  - Querying
  - Visualization
  - Mining

Integration of *Real-World Data*:

- **Quality** of integrated data:
  - Field separator used in the data
  - Missing required field
  - Primary key or data type constraint violation
  - Naming heterogeneity
  - Data format differences
  - Syntactically or semantically inconsistent data
  - Data detail level differences
  - Documentation inconsistency
  - Redundancy
  - Contractual obligations
- Steps to data **cleaning**
  - Elementizing (Parsing)
  - Standardizing
  - Verifying
  - Matching
  - Householding
  - Documenting

VIII. Bio-terminology and bio-ontology analysis

VIII.A Enrichment analysis

VIII.A.1 Motivations

Given a list of genes **found relevant** (e.g., differentially expressed) in a studied *condition*, we like to understand **why** such genes are relevant (e.g. changed significantly their expression) in that condition (w.r.t. the reference one)

Towards this aim, we want to:

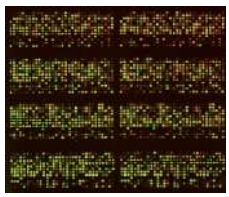
- *know* which are all the known **features** (of a certain type, e.g., functional) of such genes
- *evaluate* which of such features, if any, make a gene having them likely belonging to **such group** of genes

To do so, we can:

- *retrieve* all the gene known **annotations** (direct and unfolded), e.g., from the Gene Ontology
- consider their annotation terms and test which of them, if any, are **significantly more/less annotated** to the found genes w.r.t. all the studied genes

Goal: ***detect*** significant *enrichments* and/or *depletions* of annotation terms (e.g., Gene Ontology (GO) terms) *within a target set of genes* of interest, *with respect to a master set* (target-master scenario)

Example: help biological interpretation of microarray gene expression experiment results



- The target (study) set consists of differentially expressed genes
- The master (population) set consists of all genes that can be detected by the microarray

VIII.A.2 Problem statement

Interpret biologically why those genes have been selected as statistically significantly differentially expressed in our test condition, with respect to the reference condition

Input:

- *Master* set of n_A genes (or gene products)
- *Target* set of n_B genes (or gene products)
- *Controlled vocabulary* term t_i
- *Annotation database*, where each gene (or gene product) is annotated to zero or more terms from the controlled vocabulary

In the case that the annotation are expressed through an ontology, we need to consider all the annotation of those set of genes, not only the direct annotations that are stored in the database, but also the indirect one that can be derived by ontology unfolding starting from the direct annotation. This is because the database only store the direct ontological annotations

Terms

Genes	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

Annotation matrix: one row per gene, one column per annotation term considered

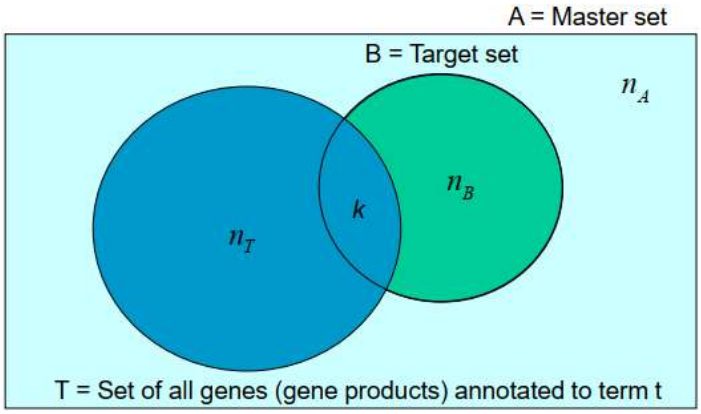
Output:

- Indication of over-representation (*enrichment*) or underrepresentation (*depletion*) of **term t_i** in the target set → p-value (significance level)

No new annotation is generated with the enrichment analysis! [beware during the exam]

For each annotation term t_i:

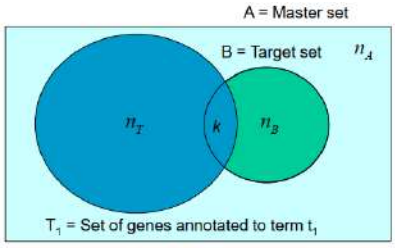
- n_T: number of master set genes (or gene products) annotated to term t_i
- k: number of target set genes (or gene products) annotated to term t_i



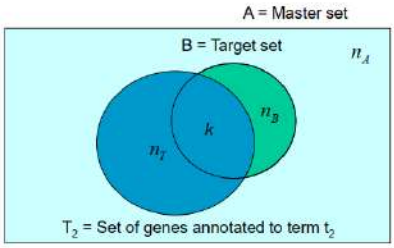
Just to have an idea (it is **not the right testing formula**!), t is enriched in B when $\frac{k}{n_B} \gg \frac{n_B - k}{n_A - n_T}$

Illustrative example:

Term t_1 is not significantly enriched in the target set B
→ High p-value



Term t_2 is significantly enriched in the target set B
→ Low p-value



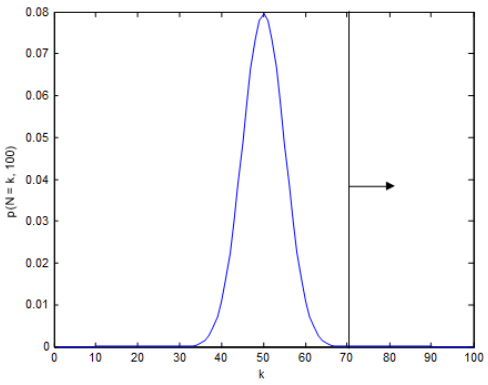
VIII.A.3 Enrichment analysis

Hypothesis testing framework (see reference [1] and [2]):

- Define a *null hypothesis* and the *null distribution*
- Compute the *significance* (p-value) of the observed data
- Compare the p-value to the *significance level* α (e.g., 0.05)

Example: tossing a coin

- Null hypothesis: the coin is fair $q = 0.5$
- Null distribution: binomial
- Observed $k = 70$ "heads" out of 100 experiments, leading to $p < 0.05$ (recall: the p-value is equivalent to the area below the curve)
- Therefore, we reject the null hypothesis

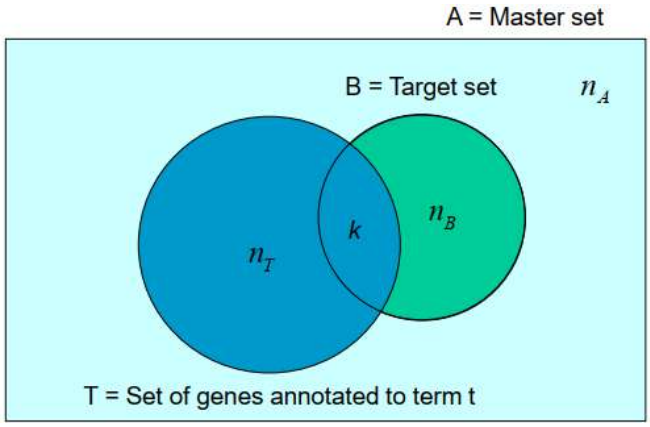


In our case, **null hypothesis**:

- Under the null hypothesis, *belonging* to the target set B is **independent from being annotated** with term t (i.e., having the feature that the term t represents)
- The probability of observing k genes in the target set annotated to the term t is given by the **hypergeometric distribution** (see reference [1] and [2]):

$$\mathbb{P}(N_{B\cap T} = k) = \frac{\binom{n_T}{k}\binom{n_A - n_T}{n_B - k}}{\binom{n_A}{n_B}}, \qquad \text{with } \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

This is the *right* testing formula!



Significance value measured on hypergeometric distribution or through the *Fisher Exact test* (two-sided test, in either case)

- Probability of finding at least k genes annotated to t in the target set B under the *null hypothesis*:

$$p = \sum_{k \geq n_{B\cap T}} \mathbb{P}(N_{B\cap T} = k)$$

If $p > \alpha$: accept the null hypothesis → Belonging to the target set B is *independent* from being annotated with term t (term t not significantly enriched)

If $p \leq \alpha$: reject the null hypothesis → Belonging to the target set B is *dependent* from being annotated with term t (term t significantly enriched), that is belonging to the target set B highlights characteristics represented by the term t

VIII.A.4 Fisher Exact test

Fisher Exact test is a test of significance used in place of χ^2 test in 2×2 **tables**, especially with *small* samples

Gives the probability P of a **contingency table** with proportion of cases on the diagonal with most cases due chance of sampling

		Cases / Class		Row totals
		Class 1	Class 0	
Cases / Group	Group 1	C(1,1)	C(1,0)	C(1,*)
	Group 0	C(0,1)	C(0,0)	C(0,*)
	Column totals	C(*,1)	C(*,0)	Tot

$$P = \frac{C(1,*)! \times C(0,*)! \times C(*,1)! \times C(*,0)!}{C(1,1)! \times C(1,0)! \times C(0,1)! \times C(0,0)! \times Tot}$$

Generally used in *one tailed* tests, but also as a two tailed tests

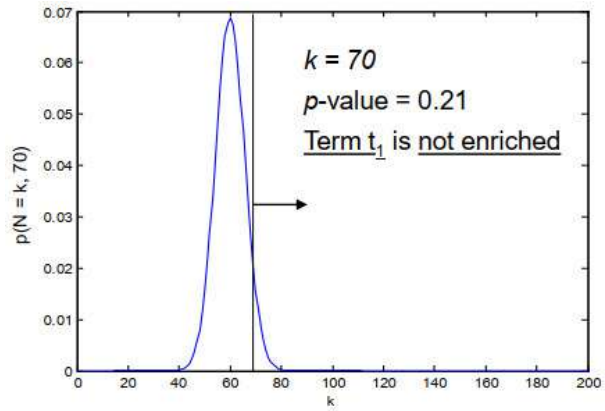
In our application case:

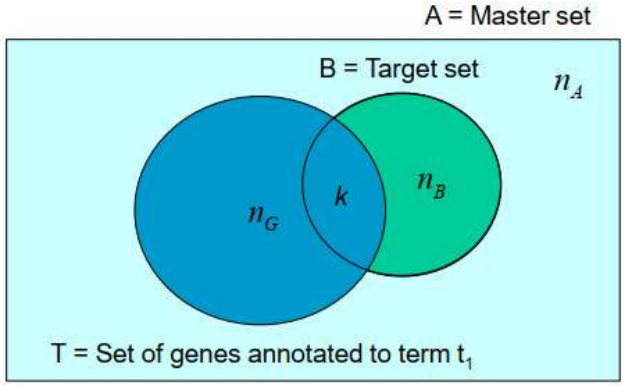
		Genes / Set		Row totals
		Target set	Not target set	
Genes / Term *i*	Annotated	k	$n_T - k$	n_T
	Not annotated	$n_B - k$	$n_A - n_B - n_T + k$	$n_A - n_T$
	Column totals	n_B	$n_A - n_B$	n_A

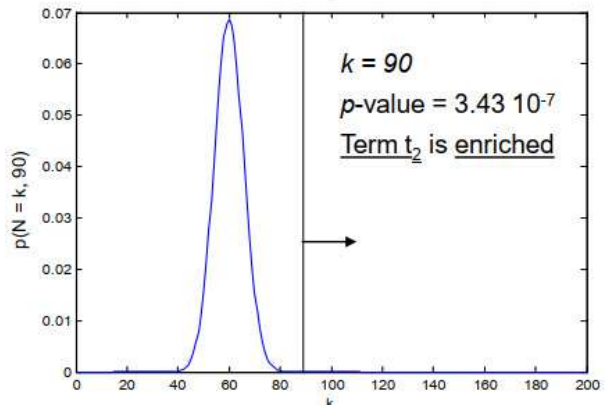
$$\mathbb{P}(N_{B\cap T} = k) = \frac{\binom{n_T}{k}\binom{n_A - n_T}{n_B - k}}{\binom{n_A}{n_B}} = \frac{n_T!\,(n_A - n_T)!\,n_B!\,(n_A - n_B)!}{k!\,(n_T - k)!\,(n_B - k)!\,(n_A - n_B - n_T + k)!\,n_A!} = P$$

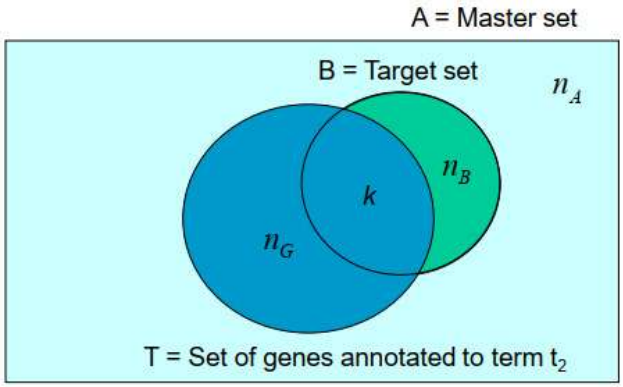
Example:

- $n_A = 1000$
- $n_B = 200$
- $n_T = 300$









VIII.A.5 Biological interpretation

Biological interpretation of gene list (belonging to target set)

- Annotation of genes (gene products) to controlled vocabulary terms means that the ***annotated genes*** (gene products) ***have the features*** described by the controlled ***terms***
- *Terms* statistically significantly **enriched** in a target set of genes **represent** the gene (gene product) **features** that make those genes (gene products) **belonging to the target set**
- If the *target set* has been selected as the genes significantly *differentially expressed* in a given biological *condition*, the significantly enriched terms represent the gene (gene product) *features* that make those genes (gene products) differentially expressed in that biological condition; thus, they *represent* the **significant features** in the given **biological condition**

VIII.A.6 Multiple testing correction

The threshold α controls the **false positive rate**, i.e.:

- It sets the probability of *discarding* the *null* hypothesis when it is *true*
- In our context, the p-value is the probability of declaring that the membership of a gene to the target set significantly depends on the annotation of the gene to the tested term t (i.e., on gene having the feature described by the term t), when such membership is *independent*

α defines the actual *false positive rate* only when testing the enrichment (depletion) of **one annotation term at a time**

When testing **multiple annotation terms simultaneously**, as it is usually the case, multiple testing correction is required to adjust p-values to correct false positives occurrence (see [3]), to not risk to have much more false positive than what we actually want

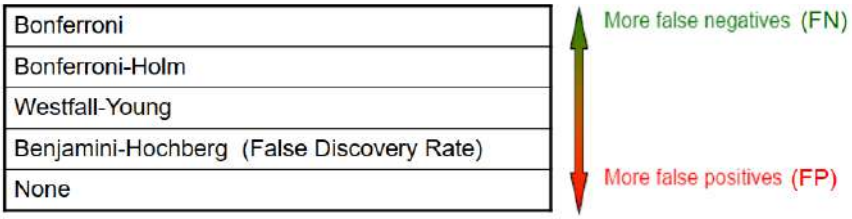
Why multiple test correction?

Example:

- Imagine a box with 20 marbles: 19 blue and 1 red
- What are the odds of randomly sampling the red marble by chance? It is 1 out of 20 (i.e., 5% chance)
- Now let's say that you get to sample a single marble (and put it back into the box) multiple times (e.g., 20 times)
- You have a much higher chance to sample the red marble (there is a 64% chance in the latter case):

$$p_{FA} = (1 - (1 - \alpha)^{N_{tests}})$$

Widely adopted multiple testing correction methods:



All these methods define *different ways to correct* (adjust) the p-value of the performed tests, in order to provide a p^{adj}-value that takes into account the variation of test significance due to the high number of multiple tests preformed

See IV.B.3.5

VIII.A.7 Ontology-based analysis

All previous methods assume that all the multiple tests performed are **independent**, i.e., in our case the *tested annotation terms are independent* (which holds when considering terms part of a terminology, a controlled vocabulary)

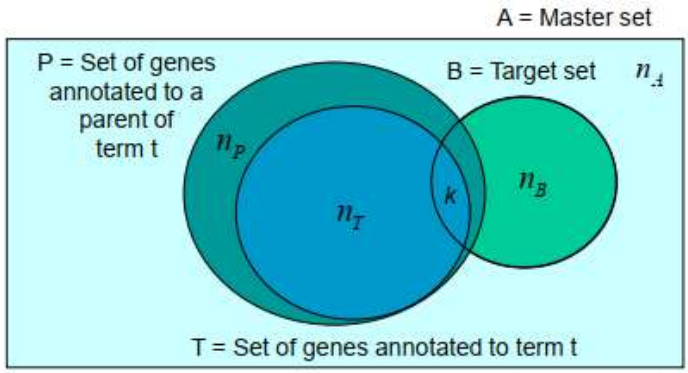
Yet, if ***ontological*** *annotations* are used, *parent-child dependencies* between annotation terms exist. Some methods try to exploit the ontology structure (e.g., the Gene Ontology DAG) to *de-correlate* ontology terms:

Alexa et al., 2006 (see reference [5]):

- Analyse the ontology terms of the annotations **bottom-up** (from more *specific* to more *generic* terms); two methods:
  - *Elim method* - For each level of the ontological hierarchy: if a term is found to be *significantly enriched*, **remove** the *annotations* to its ancestor terms of the genes annotated to it from the target and master set (i.e., do not consider these genes in the analysis of the other ancestor terms)
  - *Weight method* - improves the Elim method; instead of eliminating genes, they are assigned a **soft weight** (less prone to false negative in the enrichment analysis)

Grossman et al. (see reference [6]):

- Goal is to **avoid inheritance problem**: children of enriched terms tend to be also enriched
- The hypergeometric distribution formula is *modified* by considering the set of genes annotated to a parent term of term t and its *intersection* to the target set of genes:



$$P(N_{B\cap T} = k \mid N_{B\cap P} = n_{B\cap P}) = \frac{\binom{n_T}{k}\binom{n_P - n_T}{n_{B\cap P} - k}}{\binom{n_P}{n_{B\cap P}}}$$

VIII.A.8 Basic operations, software and tools

Assuming that ontological annotation unfolding has been performed according to the "true path rule":

1	Count number of genes (or gene products) in the master set that are annotated at least to one term	n_A
2	Count number of genes (or gene products) in the target set that are annotated at least to one term	n_B
3	Count number of genes (or gene products) in the master set annotated to a term t	n_T
4	Count number of genes (or gene products) in the target set annotated to a term t	k
5	Find parents of a term t	
6	Union of lists of genes (or gene products) annotated to a term t and one of its parents	
7	Intersection of gene (or gene product) lists	
8	Find upper induced graph (use 5)	

Enrichment analysis algorithms are provided by several:

- *Standalone* tools, including Bioconductor packages (http://www.bioconductor.org/)
- *Web-based* platforms, such as:
  - DAVID (http://david.abcc.ncifcrf.gov/)
  - GFINDer (http://www.bioinformatics.deib.polimi.it/GFINDer/)
- …

Ontologizer (see reference [6]): provides Eclipse Java project (http://ontologizer.de/) that implements:

- The methods in *Alexa et al.* and in *Grossman et al.*
- Other conventional enrichment analysis tests
- Yet, it is applied only to Gene Ontology annotations

VIII.A.9 References

1. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007; 23(4): 401-407.

2. Günther CC, Langaas M, Lydersen S. Statistical hypothesis testing of association between two lists of genes. Tech. Report. Preprint Statistics, 1/2006, Department of Mathematical Sciences, NTNU. Available from: http://www.math.ntnu.no/preprint/statistics/2006/
3. Multiple Testing Correction, Tech. Report. 2003. Available from:
4. http://envgen.nox.ac.uk/courses/GeneSpring/GS_Mar2006/Multiple%20te sting%20corrections.pdf
5. Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. Bioinformatics 2003; 19(18): 2502-2504.
6. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006; 22(13): 1600-1607.
7. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. Bioinformatics 2007; 23(22): 3024-3031.

VIII.B Functional similarity analysis

VIII.B.1 Motivations

Goal: **computing functional similarity between genes** (or gene products), based on *annotations* describing their *functions*

Traditional *strategies* are based on:

- *Sequence similarity* (sequence homologs), to determine *functional categories* (protein domain families)
- Analysis of *correlation* (co-expression, genes active at the same time) in gene expression, e.g., using microarray experiments

Issues: most **co-functioning genes**:

- Neither is *sequence*-related
- Nor encodes *proteins* in the same protein family, e.g., proteins with the same domains, or in the same pathway
- Can be expressed at different *time* points

Hypothesis: if two genes (or gene products) have *similar functional annotation profiles*, they should be *functionally related.* Therefore, we compute measure of functional similarity based on gene (or gene products) **annotation profiles**

Annotation profiles, expressed through:

- *Controlled vocabularies* (terminologies)
- *Ontologies* (e.g., the Gene Ontology, and many others!)

VIII.B.2 Functional similarity based on controlled vocabularies

Controlled vocabulary schemas mandate the uses of *predefined*, authorized *terms* that have been *preselected*

In their simplest version, there are *no semantic links* between the terms in the controlled vocabulary

Annotation of genes (or gene products): each gene (or gene product) can be annotated to zero or more terms from a controlled vocabulary

	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

Annotation matrix

Example tool: DAVID (The Database for Annotation, Visualization and Integrated Discovery), see reference [6]. Annotations from Gene Ontology, UniProt, KEGG, ...

Use **Kappa (κ) statistical *index*** (http://en.wikipedia.org/wiki/Cohen%27s_kappa): a statistic applied on our contingency tables, to evaluate the *agreement between two observers* (one on every side of the table)

		Gene a: 1	Gene a: 0	Row total
Gene b	1	3 ($C_{1,1}$)	1 ($C_{0,1}$)	4 ($C_{1,\cdot}$)
	0	0 ($C_{0,1}$)	2 ($C_{0,0}$)	2 ($C_{0,\cdot}$)
Column total		3 ($C_{\cdot,1}$)	3 ($C_{\cdot,0}$)	6 (T_{ab})

$$O_{ab} = \frac{C_{1,1} + C_{0,0}}{T_{ab}} = \frac{3+2}{6} = 0.83$$

$$A_{ab} = \frac{C_{\cdot,1} \bullet C_{1,\cdot} + C_{\cdot,0} \bullet C_{0,\cdot}}{T_{ab} \bullet T_{ab}} = \frac{3 \bullet 4 + 3 \bullet 2}{6 \bullet 6} = 0.5$$

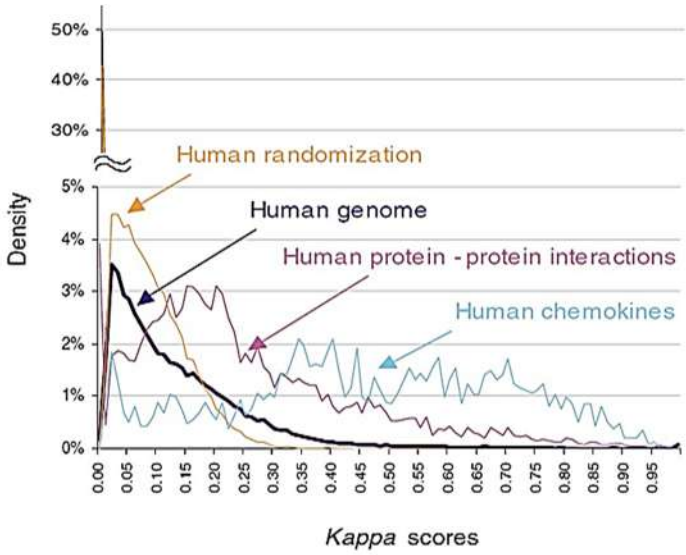
$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}} = \frac{0.83 - 0.5}{1 - 0.5} = 0.66$$

This last value computed K_{ab} ranges from -1 to 1, with:

- $K = 1$: perfect agreement
- $K = 0$: agreement by chance
- $K < 0$: disagreement

Validation of the Kappa index metrics: restricting the analysis to (known) functionally related genes provides higher Kappa scores (see reference [6])

$K \in [0.0, 0.20]$ poor, $[0.20, 0.40]$ fair, $[0.40, 0.60]$ moderate, $[0.60, 0.80]$ good, $[0.80, 1]$ very good agreement



We performed function similarity according to the annotation profile pairwise between pairs of genes belonging to each the category. The distribution of the kappa scores shows many low values, but when we increase the function similarity of the set of genes we analyse (e.g., the chemokines), the kappa score becomes higher and somehow relevant

VIII.B.3 Computing similarity based on annotation profile

Typically, with **ontological annotations**, it is a two-step procedure (more precise evaluation):

- Compute ontological *term-to-term similarity*
- Compute *gene-to-gene* (or gene product-to-gene product) *similarity* based on *annotation profile*

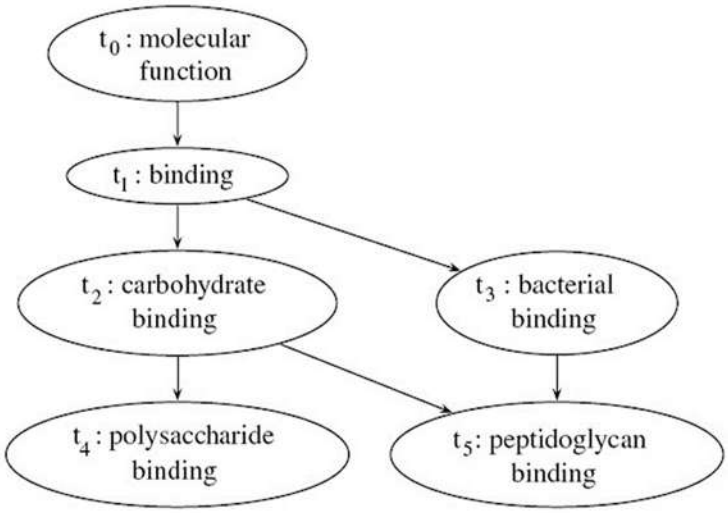
Additional steps (optional):

- Compute *gene* (or gene product) *similarity* based on *multiple ontologies*
- Compute *gene* (or gene product) *clustering* based on *functional similarity*

Step 1 (*term-to-term similarity*) methods:

- Ontology **topology-based methods**:
  - Compute the *distance* (within the ontology) between *two ontological terms* by counting the *number of arches* between them *within* the ontology
  - Shortest or average distance is used for *multiple paths*
  - Issue: assume that nodes and arches are *uniformly* distributed in an ontology (usually not true, because the distribution of nodes and arches within the ontology depend on the development of the knowledge in the domain represented by the ontology, and some parts of the ontology may be very detailed [due to more experiments, etc.] compared to others)
- **Information theoretic methods** (e.g., *Singular Value Decomposition* (SVD), see clustering techniques):
  - Less sensitive to arch density variability

Example of a simple ontology:



A sub-graph of the GO – Molecular function (See reference [7])

VIII.B.3.1 Computing term-to-term similarity

Common method to compute **term-to-term similarity** (5 steps):

1. Compute the *frequency* (probability) of occurrence of a *term* in a *corpus* (e.g., all gene annotations to the ontology terms):

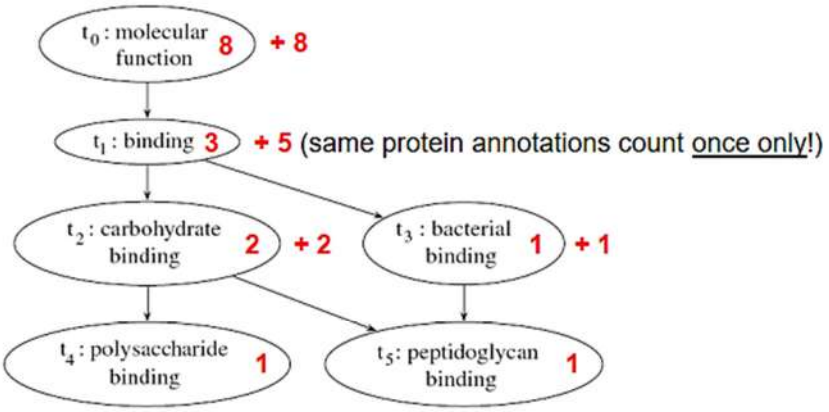
$$Freq(c) = \sum \{occur(c_i) | c \in Ancestors(c_i)\}$$
$$Prob(c) = \frac{Freq(c)}{maxFreq}$$

2. Compute the *information content* (IC) of a term (in a corpus):

$$IC(c) = -\log\big(Prob(c)\big)$$

The *rarer* (specific) is the term, the *lower* is its probability and the *higher* is its IC

The more *common* (generic) is the term, the *higher* is its probability and the *lower* is its IC



GO term	Protein annotations	*Freq*	*Prob*	*IC*
t_0	8	16	1	0
t_1	3	8	0.5	1
t_2	2	4	0.25	2
t_3	1	2	0.125	3
t_4	1	1	0.0625	4
t_5	1	1	0.0625	4

Note that Protein annotations are the direct annotation, Freq is the direct+indirect annotation, and that the same protein annotations count only once in the Freq

3. Find *common ancestors* of two terms:

$$CommonAnc(c_1, c_2) = Ancestors(c_1) \cap Ancestors(c_2)$$

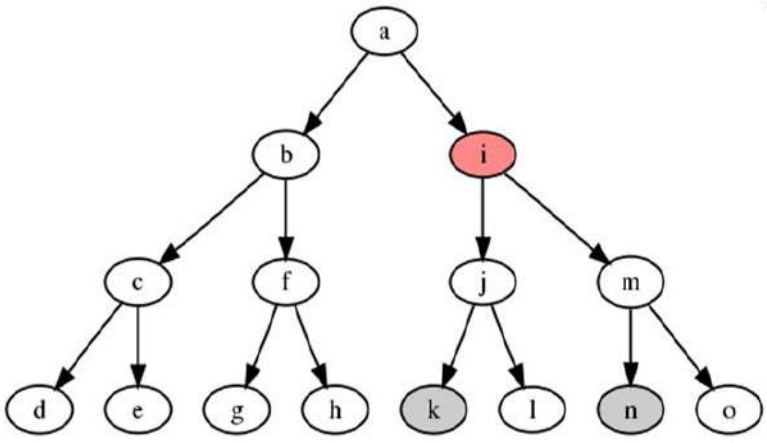
4. Compute *shared information* between two terms:

$$Share(c_1, c_2) = \max\{IC(a) | a \in CommonAnc(c_1, c_2)\}$$

i.e., the information content (IC) of the *Lowest Common Ancestor* (LCA), which is the ancestor term more distant *from the ontology root*

Lowest Common Ancestor (LCA):

- Common ancestors of nodes (terms) k and n are nodes a and i
- The LCA of nodes k and n is node i, the most distant between node a and i from the ontology root



5. Compute *similarity metrics* between two terms:

Resnik:

$$Sim_{Resnik}(c_1, c_2) = Share(c_1, c_2)$$

Jiang:

$$dist_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times Share(c_1, c_2)$$

$$Sim_{JC}(c_1, c_2) = \frac{1}{dist_{JC}(c_1, c_2) + 1}$$

Lin:

$$Sim_{Lin} = \frac{2 \times Share(c_1, c_2)}{IC(c_1) + IC(c_2)}$$

There are *many other metrics* derived from the above (e.g., multiple shared ancestors can be considered)

VIII.B.3.2 Computing gene-to-gene similarity

Step 2 (*gene-to-gene similarity*):

Computing gene-to-gene (gene product-to-gene product) similarity (based on the similarity of their annotation terms):

- Consider *two genes* (gene products), p and q, annotated to N and M terms:

$$GO^p = \{GO_1^p, \dots, GO_N^p\}$$
$$GO^q = \{GO_1^q, \dots, GO_M^q\}$$

- Define the *term-to-term similarity* metrics:

$$s_{ij} = sim(GO_i^p, GO_j^q), \qquad \forall i \in \{1, \dots, N\}, \forall j \in \{1, \dots, M\}$$

- Compute *similarity metrics* between two genes (gene products):
  - Lord (max of annotated term similarity scores):

$$GOscore_{\max}(p, q) = \max(s_{ij}), \qquad \forall i \in \{1, \dots, N\}, \forall j \in \{1, \dots, M\}$$

  - Speer (average of annotated term similarity scores):

$$GOscore_{\text{avg}}(p, q) = \frac{1}{N * M} \sum s_{ij}, \qquad \forall i \in \{1, \dots, N\}, \forall j \in \{1, \dots, M\}$$

  - Schliker (max of averages of max of term similarity scores, more complex and precise):

$$GOscore_{\text{BM}}(p, q) = \max\{rowScore(p, q), columnScore(p, q)\}$$

where *rowScore* is the average of the row maxima and *columnScore* is the average of the column maxima of the scores s between each term annotated to p and each term annotated to q

VIII.B.4 Similarity analysis, basic operations

We assume that unfolding of ontological annotations has been performed

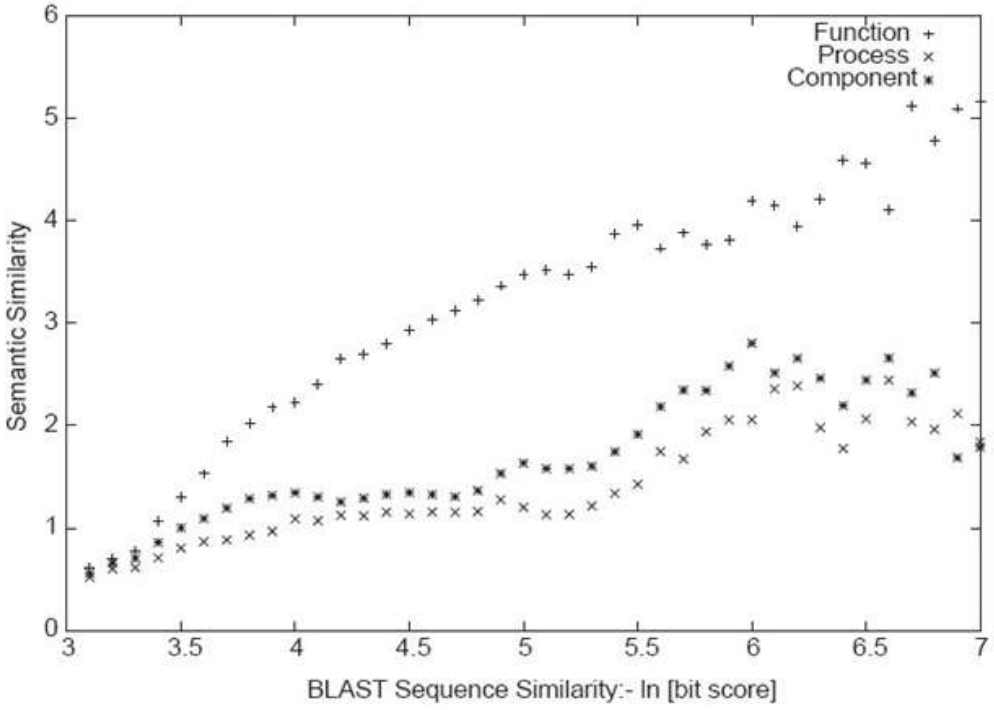
1	Count number of genes (or gene products) annotated to a term t
2	Given two terms t_1 and t_2, find common ancestors
3	Given two terms t_1 and t_2, find the lowest common ancestor
4	Compute/Store/Fetch term-to-term similarity
5	Find the terms annotated to a given gene (or gene product)
6	Compute gene (or gene product) similarity

VIII.B.5 Validating functional similarities metrics

- Using *structural information* (sequence similarity), see references [1] and [2]
- Using *gene expression* data (e.g. microarray experiments), see reference [3]
- Assessing the *functional consistency of clustering*, see references [4], [5] and [6]

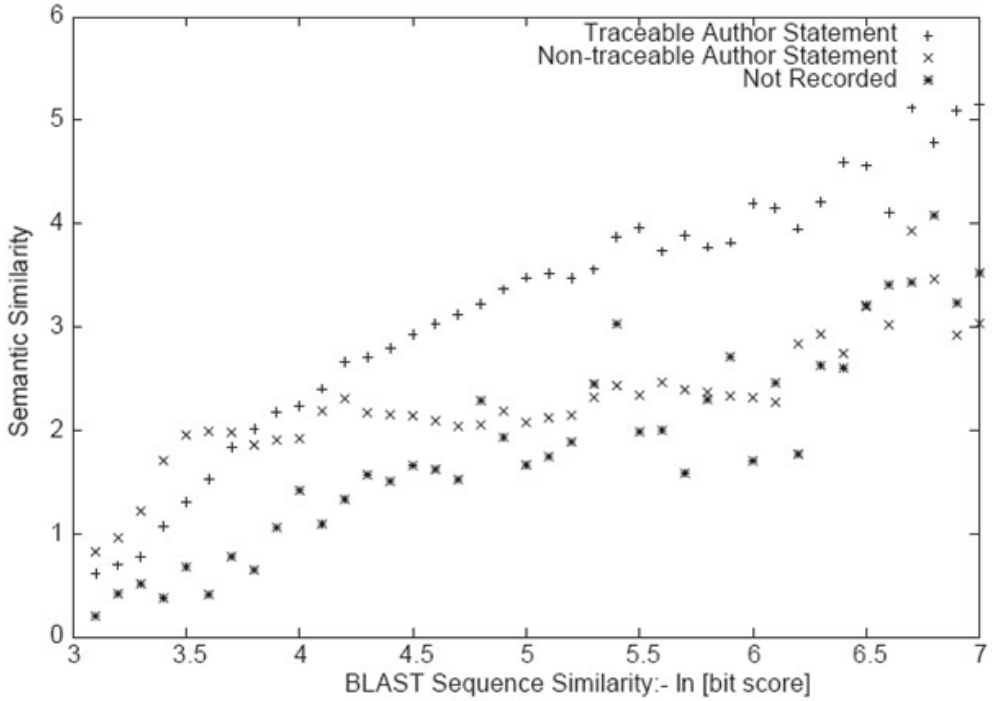
All these approaches have the limitations mentioned in VIII.B.1

- Using **structural information** (sequence similarity):



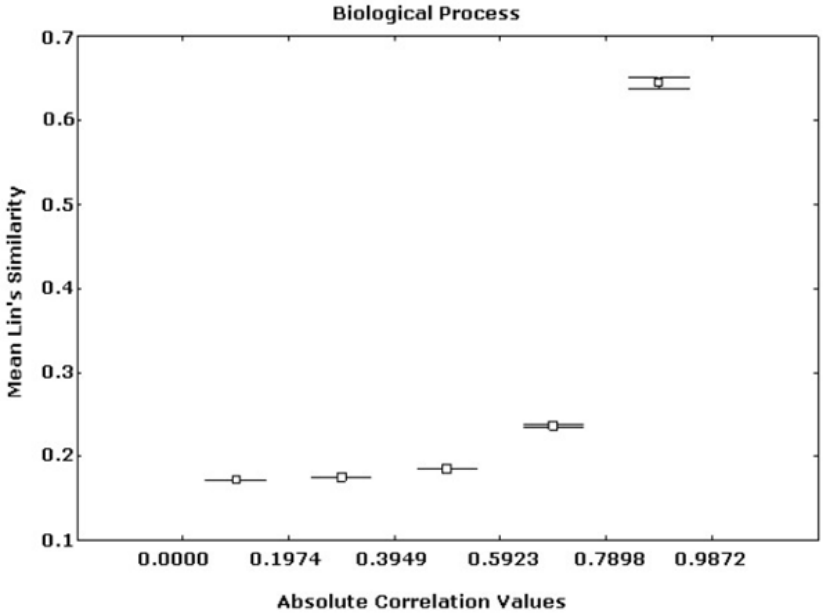
x-axis shows the *structural similarity* between the sequences of a pair of genes (or gene products)

Evidence of annotation:



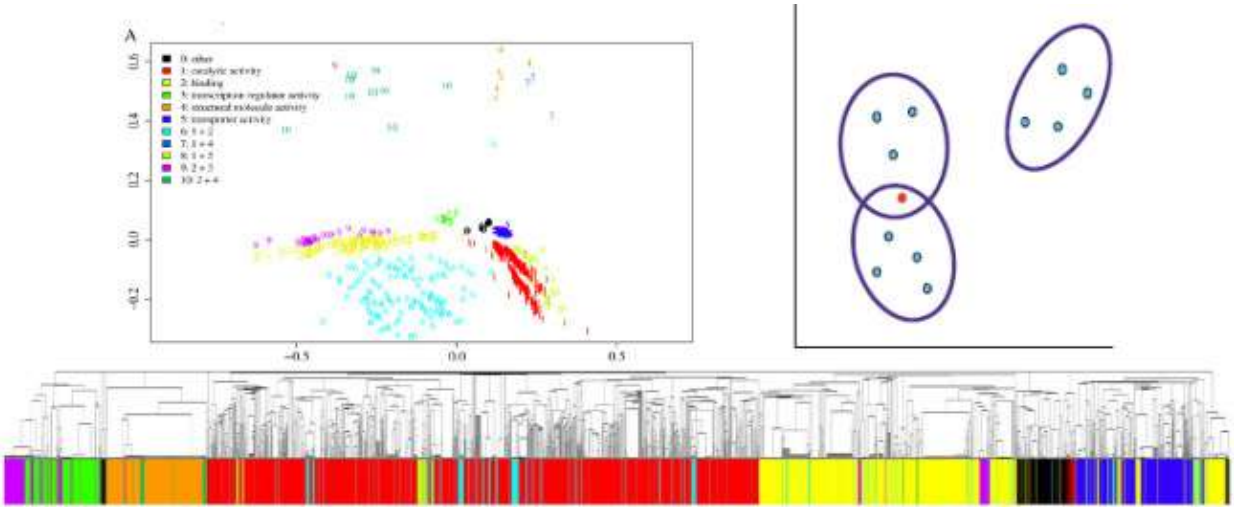
GO annotations tagged as "*traceable author stat*ement" (TAS) are typically more reliable

- Using **gene expression data** (microarray experiments):



x-axis shows the correlation coefficient between *gene expression* data obtained in microarray experiments

- Assessing the **semantic functional consistency of clustering**
  - Similarity measures induce a metric space (not a vector space)
  - Some examples:
    - Multi Dimensional Scaling (MDS)
    - Hierarchical clustering
    - Fuzzy clustering



VIII.B.6 Compute gene similarity based on multiple ontologies

Merge similarity scores obtained using *different ontologies*, e.g., GO Biological Process (BP) and GO Molecular Function (MF) (see reference [8]):

$$funSim(p,q) = \frac{1}{2}\left[\left(\frac{BPscore(p,q)}{\max BPscore}\right)^2 + \left(\frac{MFscore(p,q)}{\max MFscore}\right)^2\right]$$

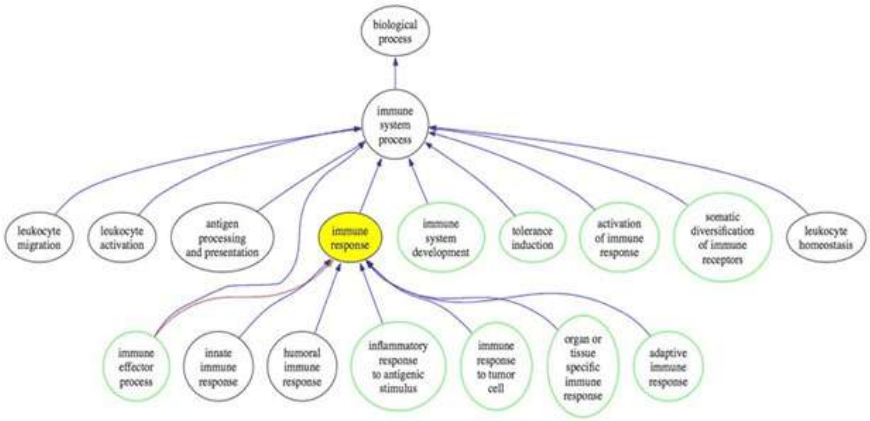
$$rfunSim(p,q) = \sqrt{funSim(p,q)}$$

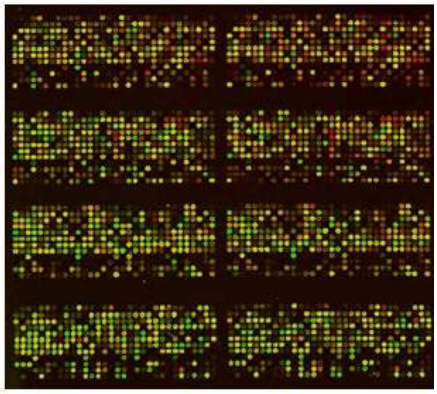
That is, the square root of average of squares of normalized scores. It is very important to normalize the scores, because the ontology can have very different structures (in particular in terms of richness of nodes), and even if the information content methods already take that into account (but not enough)

VIII.B.7 Gene clustering based on functional similarity

Goal: **functional clustering**

- Functional annotation-based clustering (see examples before)
- Co-clustering: *Functional annotations* + microarray *expression data*





VIII.B.8 Conclusion

Most of the literature on *semantic similarity* between genes (or gene products) is based on the *GO*. However, genes (and gene products) can be *annotated* with ontologies (or controlled vocabularies) *other* than the GO

Open issues:

- *Fuse information* provided by multiple ontologies
- Handle ontologies that have a *different structure* than the GO (i.e., which are not directed acyclic graphs)

VIII.B.9 References

1. Lord PW, Stevens RD, Brass A, Goble CA. Semantic similarity measures as tools for exploring the Gene Ontology. Pac Symp Biocomput. 2003: 601-612.
2. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003; 19(10): 1275–1283.
3. Wang H, Azuaje F, Bodenreider O, Dopazo J. Gene expression correlation and Gene Ontology-based similarity: An assessment of quantitative relationships. Proceedings IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2004; 25-31.
4. Schlicker A, Domingues F, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 2006; 7: 302.
5. Kustra R, Zagdanski A. Incorporating Gene Ontology in clustering gene expression data.
6. Proceedings 19th IEEE Symposium on Computer-Based Medical Systems 2006; 555-563.
7. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biology 2007; 8(9): R183.
8. Couto FM, Silva MJ, Coutinho PM. Measuring semantic similarity between Gene Ontology terms. Data & Knowledge Engineering 2007; 61: 137–152.
9. Schlicker A, Albrecht M. FunSimMat: a comprehensive functional similarity database.
10. Nucleic Acids Research 2008; 36(Database issue): D434-439.