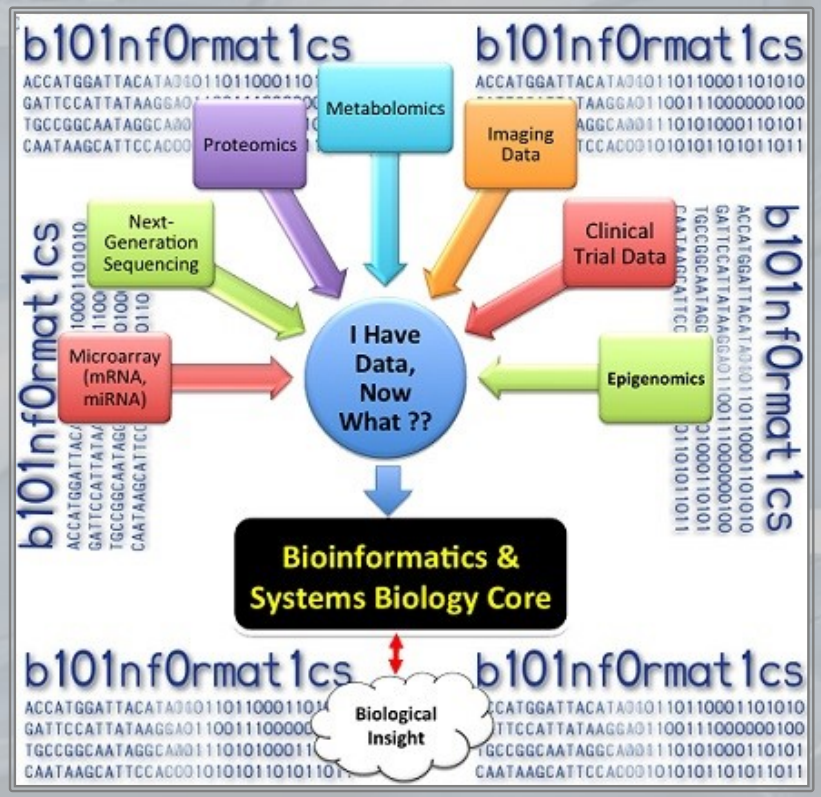
Molecular Biology and Biochemistry

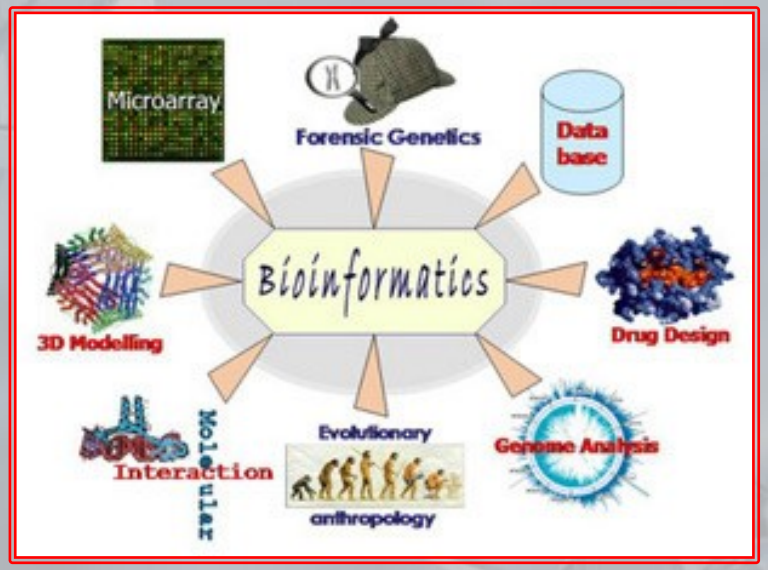


"Science cannot solve the ultimate mystery of Nature. And that is because, in the last analysis, we ourselves are a part of the mystery that we are trying to solve." (M. Planck)

Table of contents

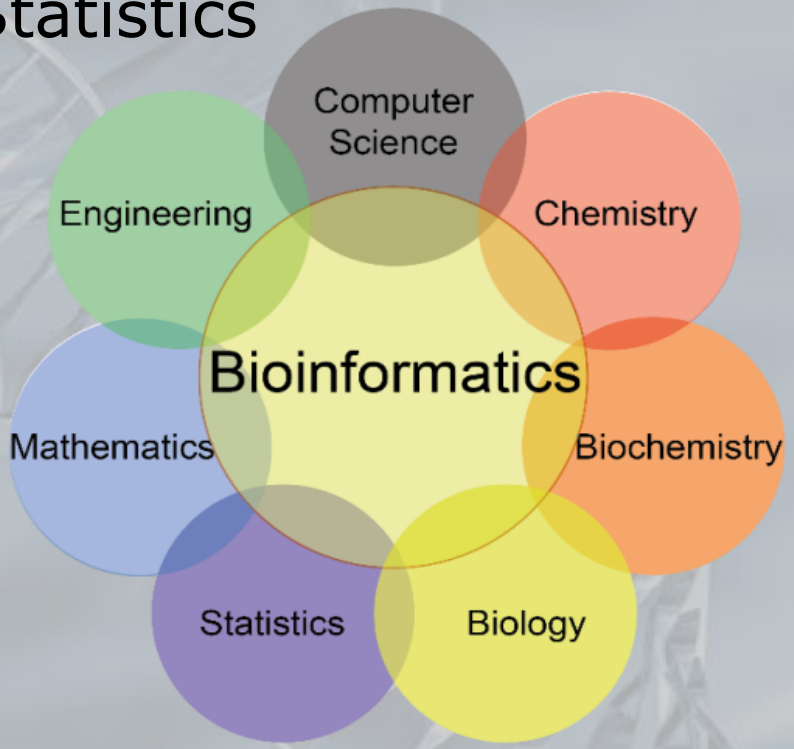
Introduction – 1

- The main feature of living organisms is their ability of saving, managing, and transmitting information
- Bioinformatics aims at:
  - determining the biologically relevant information
  - deciphering how it is used in order to control the internal chemistry of living organisms
- Computers aid to collect, analyze, and interpret bio-logical information at the molecular level



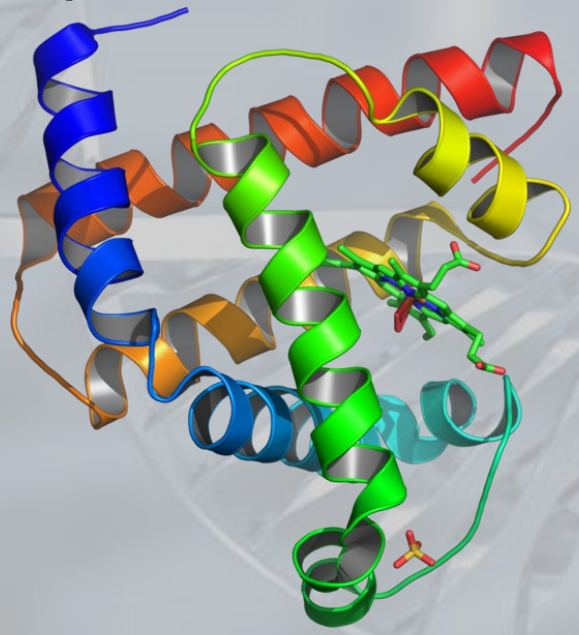
Introduction – 2

- Bioinformatics is a discipline that lies at the crossroad of Biology, Chemistry, Computer Science, Mathematics, and Statistics
- In fact, it is characterized by the application of mathematical, statistical, and computational tools for biological, biochemical and biophysical data processing



Introduction – 3

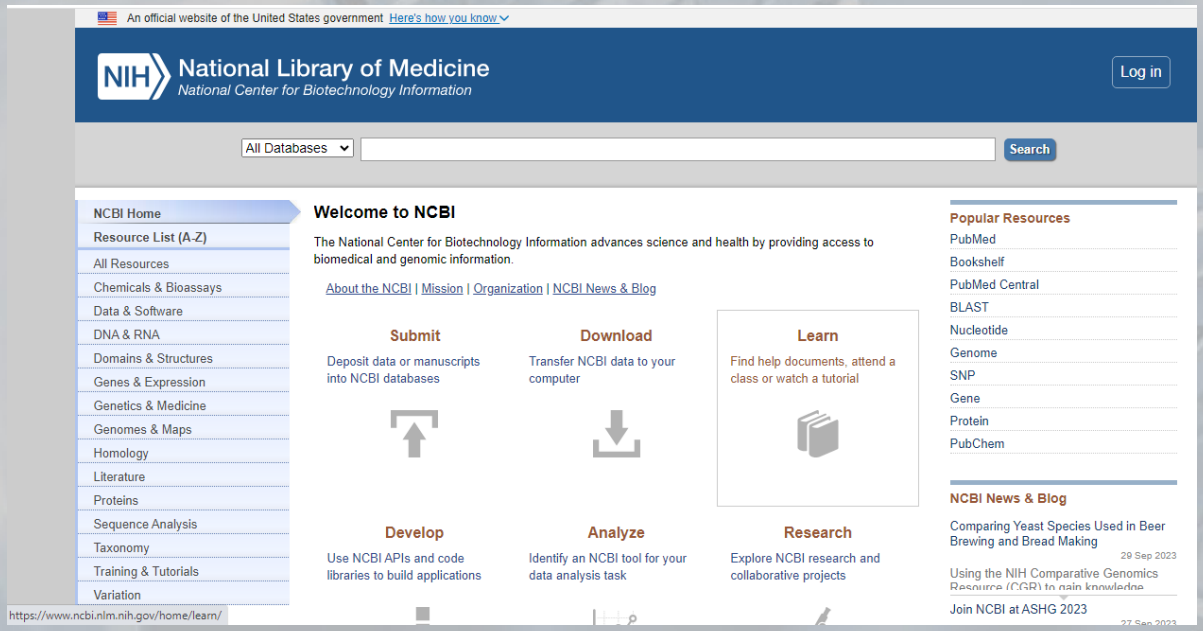
- The main topic in Bioinformatics is the analysis of DNA but, with the wide diffusion of more economic and effective techniques to acquire protein data, also proteins have become largely available for bioinformatics analyses



Introduction – 4

- Bioinformatics comprises two main subtopics:

1) Development and implementation of tools to store, manage, and analyze biological data
   - Store data related to Genomics, Transcriptomics, Proteomics, Metabolomics, etc., and related literature
   - Database building techniques, suitable for biomedical research, and *ad hoc* query tools



Introduction – 5

2) Information extraction via data analysis and inter-pretation; development of new algorithms to explain/verify relationships among a huge amount of information
   - Data mining, to generate knowledge from data
   - Statistical and artificial intelligence methods, for the analysis of biological data and the identification/ automatic extraction of significant information
   - Indeed, data mining is the probing of available datasets in order to identify patterns and anomalies
   - Machine learning is the process of automatic learning from heterogeneous data in a way that mimics the human learning process
   - Together, they enable both past data characterization and future data prediction

Introduction – 6

- Recent technological advances allow for high throughput profiling of biological systems in a cost–efficient manner
- These data are being extensively used to decipher the mechanisms of biological systems at the most basic level
- The low cost of data generation has brought us into the big data era
- The availability of big data (both in dimension and amount) provides unprecedented opportunities, but also raises new challenges, for Bioinformatics

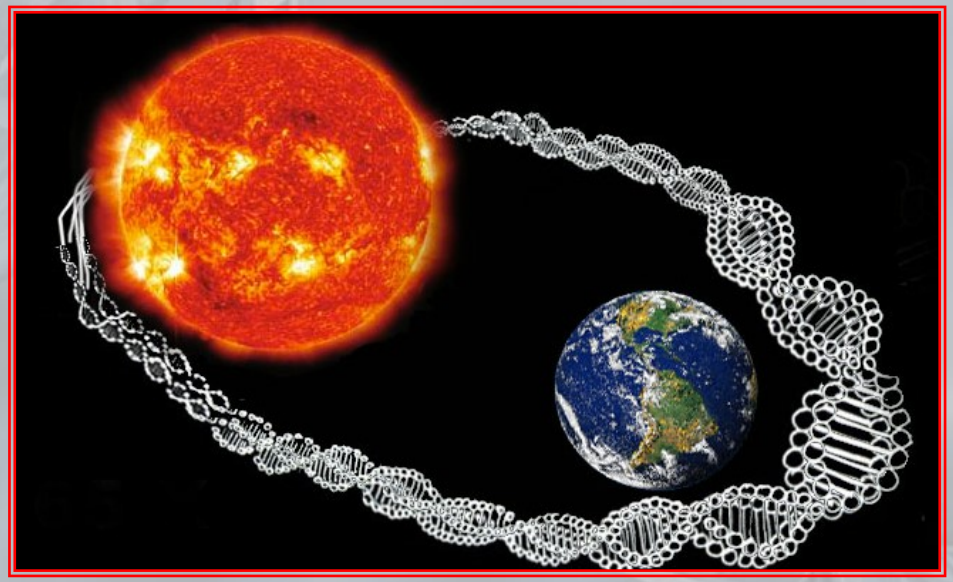
But how big are the data? – 1

- Total size of human genome
  - Each cell carries 3.2 billion base pairs (aploid genome): a code that can be written in 500 books, with 500 pages each
  - Length of DNA in adult man calculated as

(length of 1 bp)×(number of bps per cell)×(number of cells in the body)

$= (0.34\times10^{-9}\ \text{m})(6.4\times10^{9})(10^{13}) \sim 2.2\times10^{13}\ \text{m}$

- That is the equivalent of nearly 70 trips from the earth to the sun and back

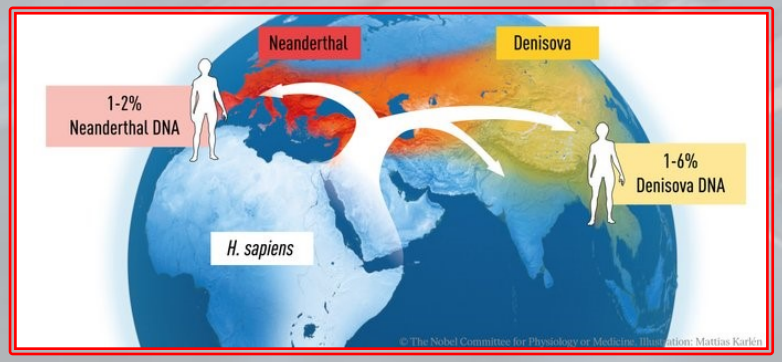


But how big are the data? – 2

- How is it possible?
  - The DNA in our cells is packaged into 46 chromosomes in the nucleus; it is a helical molecule supercoiled using enzymes so that it takes up less space
  - Actually, our DNA is arranged as a coil of coils of coils of coils of coils! This allows the approx 3 billion base pairs in each cell to fit into a space just 6 microns across
  - Stretching the DNA in one cell all the way out, it would be about 2 meters long and all the DNA in all our cells put together would be about twice the diameter of the Solar System

And which is the ultimate goal?

- The goal of Bioinformatics is bivalent
- Actually, it is aimed at providing scientists with a means…
  - …to understand the basis of biological diversity and to trace the evolutionary history of the life on the Earth, which is written in our molecules
  - …to explain normal biological processes, to highlight malfunctions which lead to diseases, and to define approaches that can improve drug discovery and design

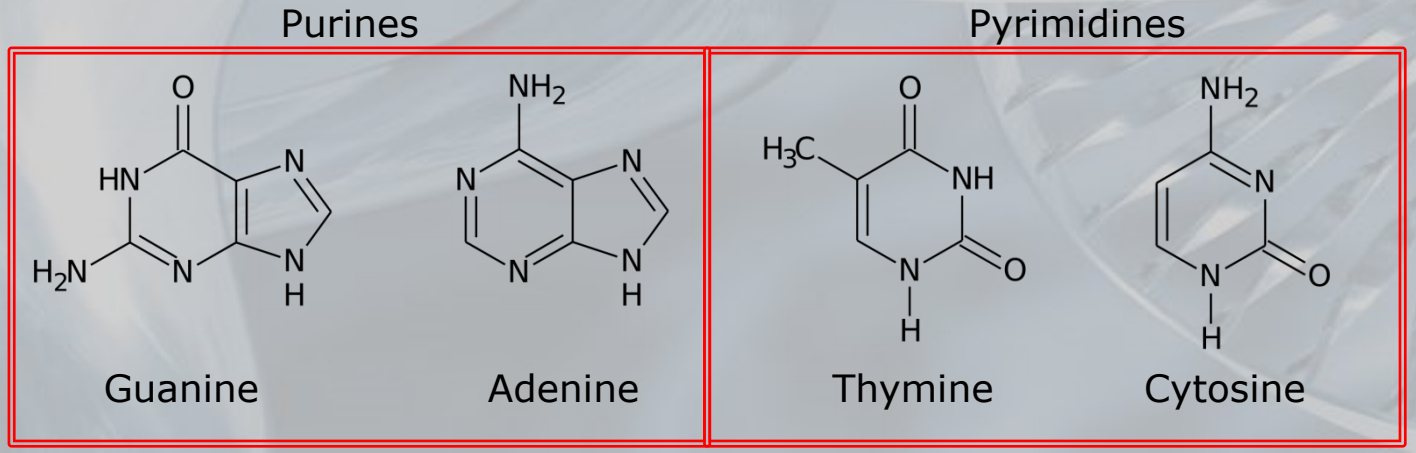


The genetic material – 1

- The genetic material is the DNA (deoxyribonucleic acid); DNA molecules are made up of a few kinds of atoms — carbon, hydrogen, nitrogen, oxygen and phosphorus
- It is actually the information contained in the DNA that allows the organization of inanimate molecules in living cells and organisms, capable of regulating their internal chemical composition and their growth and reproduction
- It is also the DNA that gives us the inheritance of our ancestors' physical traits, through the transmission of genes

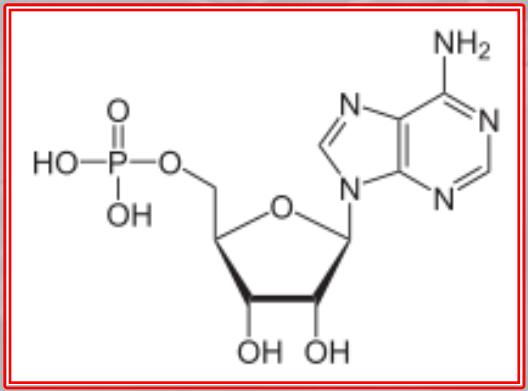
The genetic material – 2

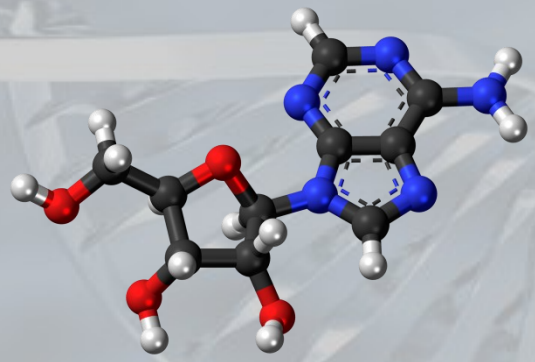
- Genes contain the information, in the form of specific nucleotide sequences, which constitute the DNA molecules
- DNA molecules use only four nucleobases, guanine, adenine, thymine and cytosine (**G**, **A**, **T**, **C**), which are attached to a phosphate group (PO_4) and to a deoxyribose sugar ($C_5H_{10}O_4$), to form a nucleotide



The genetic material – 3

- All the information collected in the genes depends on the order in which the four nucleotides are organized along the DNA molecule
- Complicated genes may be composed by hundreds of nucleotides
- The genetic code, "which describes" an organism, known as its genome, is conserved in millions/billions of nucleotides



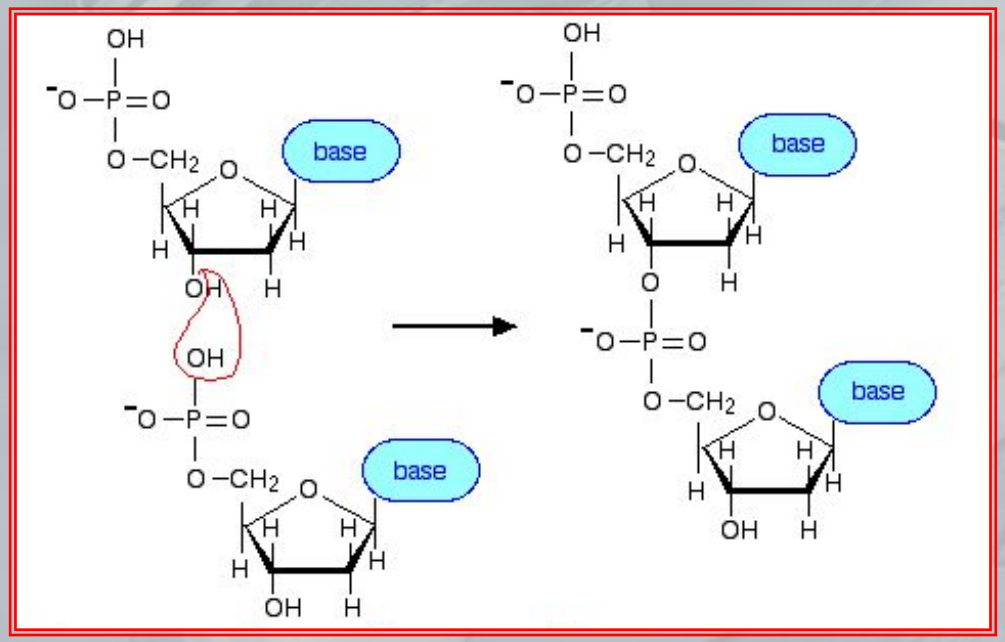


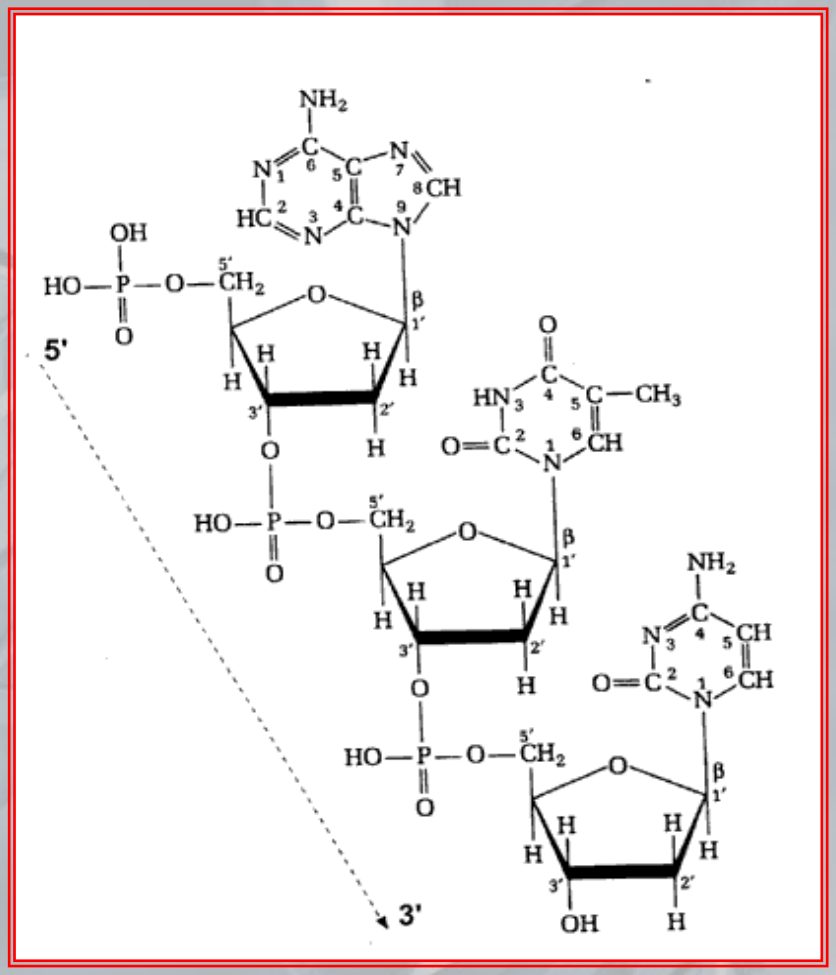
Chemical structure of the Adenosine 5' monophosphate: each nucleotide is composed of three parts, a phosphate group, a central deoxyribose sugar, and one out of the four nucleobases

The genetic material – 4

- Nucleotide strings can be joined together to form long polynucleotide chains and, on a larger scale, a chromosome
- The union of two nucleotides occurs through the formation of a phosphodiester bond, which connects the phosphate group of a nucleotide and the deoxyribose sugar of another nucleotide
- Ester bonds involve connections mediated by oxygen atoms
  - Phosphodiester bonds occur when a phosphorus atom of a phosphate group binds two molecules via two ester bonds

The genetic material – 5

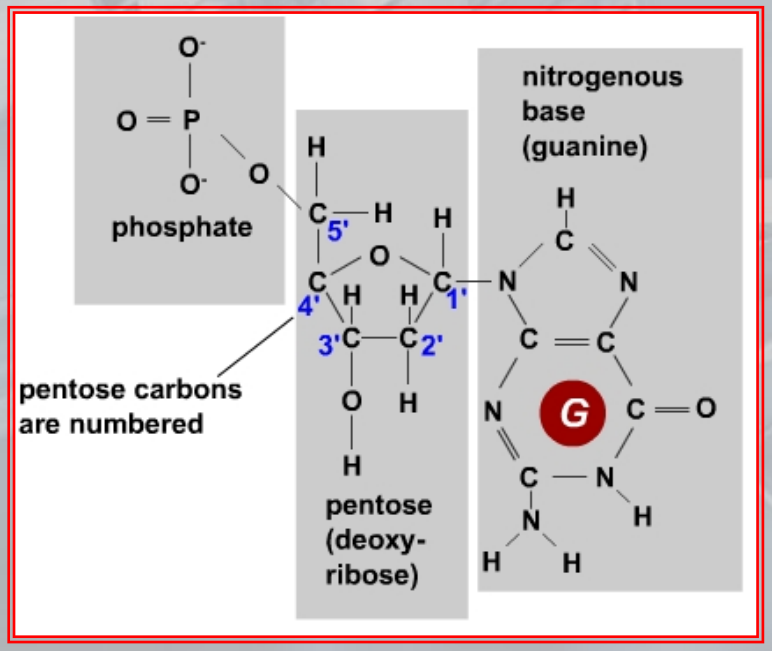




DNA nucleotides are joined together by covalent bonds between the hydroxyl groups (_OH) in 5’ and 3’: the alternation of phosphate residues and pentoses constitutes the skeleton of nucleic acids

The genetic material – 6

- All living organisms form phosphodiester bonds exactly in the same way
- To all the five carbons in the deoxyribose sugar – also called a pentose – an order number (from 1’ to 5’) is assigned



The genetic material – 7

- Phosphate groups of each free nucleotide are always linked to the 5′ carbon of the sugar
  - Phosphate groups constitute a bridge between the 5′ carbon in the deoxyribose of a new attaching nucleotide and the 3′ carbon of a pre–existing polynucleotide chain
  - A nucleotide string always starts with a free 5′ carbon, whereas, at the other end, there is a free 3′ carbon
- The orientation of the DNA molecule is crucial for deciphering the information content of the cell

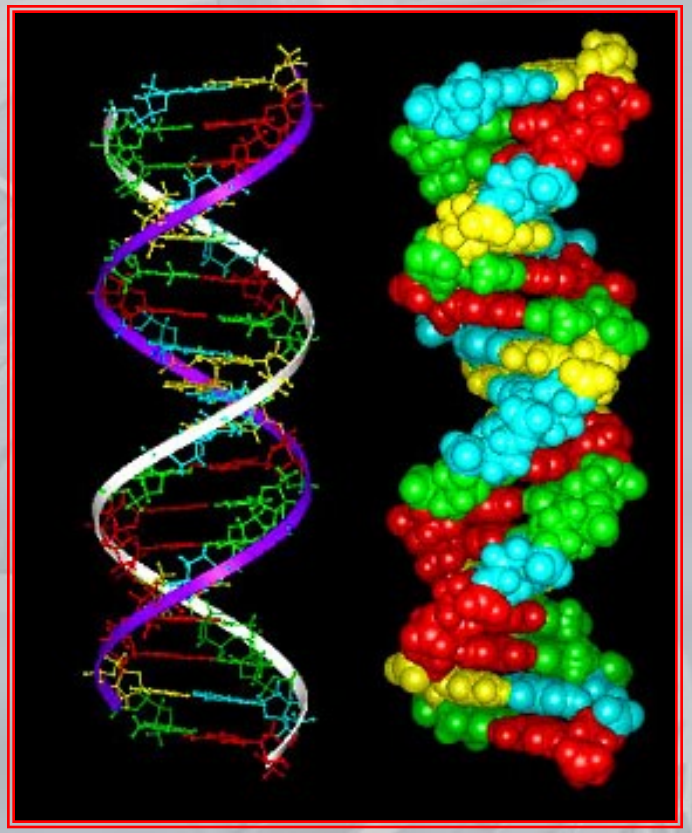
The genetic material – 8

- A common theme in all biological systems, and at all levels, is the idea that the structure and the function are intimately correlated
- The discovery of J. Watson and F. Crick (1953) that the DNA, inside the cells, does not exist as single molecules, but rather as strands of molecules entwined together, provided an invaluable clue on how DNA could act as genetic material

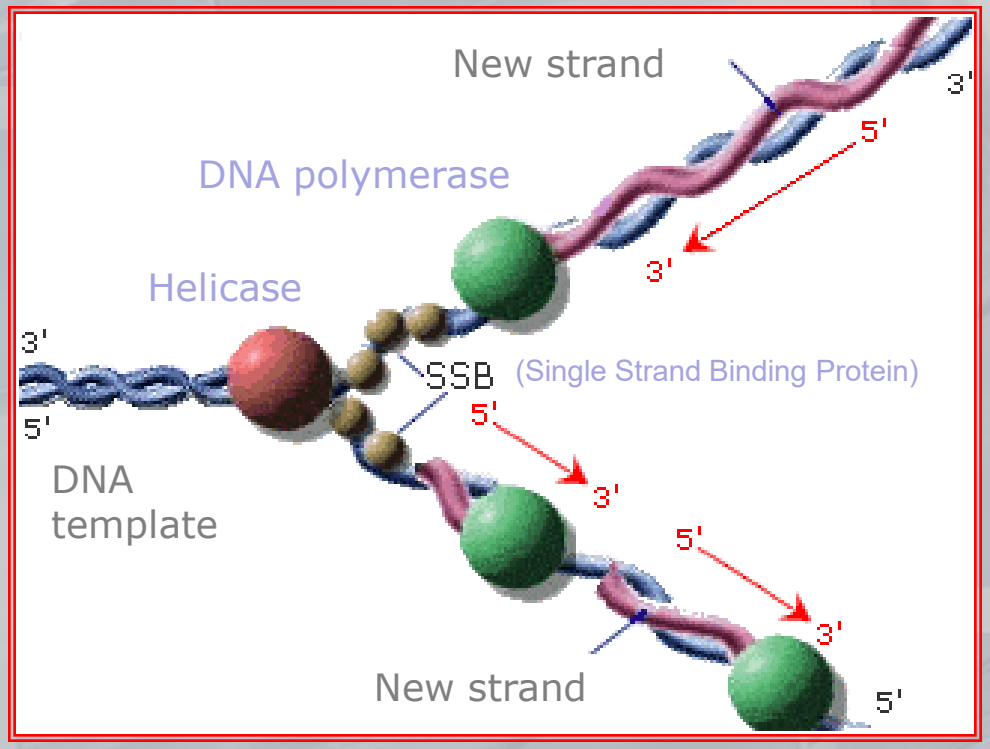
J. Watson e F. Crick (Nobel Prize for Medicine with M. Wilkinson, 1962)

The genetic material – 9

- Therefore, the DNA is made up of two strands, and the information contained in a single strand is redundant with respect to the information contained in the other
- The DNA can be replicated and faithfully transmitted from generations to generations by separating the two strands and using each strand as a template for the synthesis of its companion



The genetic material – 10



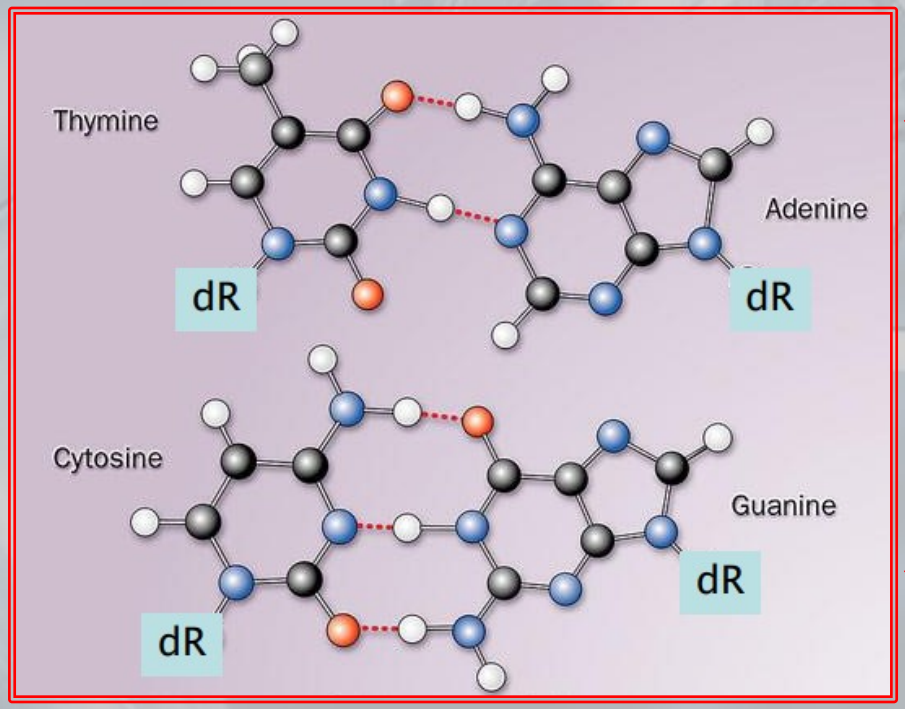
During the DNA replication, the double strand is separated by means of the helicase enzyme; each DNA strand serves as a template for the synthesis of a new strand, produced thanks to the DNA polymerase

The genetic material – 11

- More precisely: the information contained in the two DNA strands is complementary
  - For each **G** on one strand, there is a **C** on the complementary strand and vice versa
  - For each **A** on one strand, there is a **T** on the complementary strand and vice versa
  - The interaction between **G**↔**C** and **A**↔**T** is specific and stable
    - In the space between the two DNA strands (~11Å), guanine, with its double–ring structure, is too large to pair with the double ring of an adenine or another guanine
    - Thymine, with its single–ring structure, is too small to interact with another nucleobase with a single ring (cytosine or thymine)

The genetic material – 12

- However, the space between the two DNA strands does not constitute a barrier for the interaction between **G** and **T**, or between **A** and **C** but, instead, it is their chemical nature that makes them incompatible



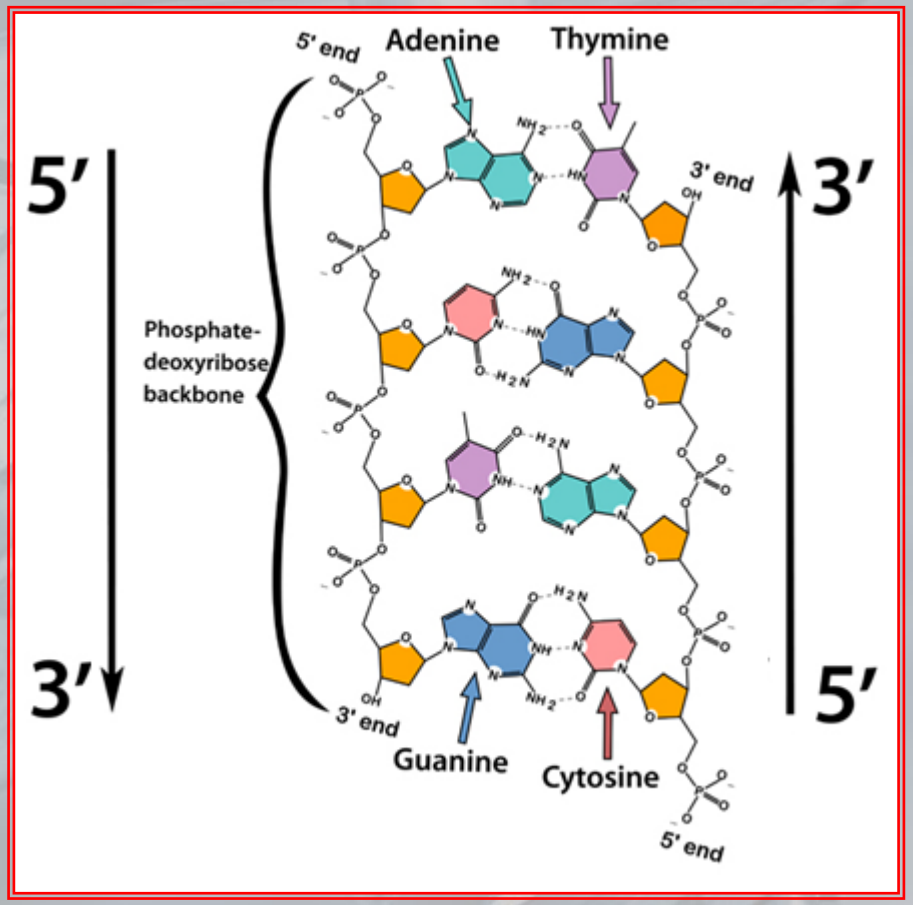
Two hydrogen bonds

The chemical interaction between the two different types of nucleobases is so stable and energetically favorable to be responsible for the pairing of the two complementary strands

Three hydrogen bonds

The genetic material – 13

- The two strands of the DNA molecule do not have the same direction (5′→3′)
- They are anti–parallel, with the 5′ termination of one strand paired with the 3′ termination of the other, and vice versa
- ⇨ The DNA strands are *reverse–complements*



The genetic material – 14

- Example

  If the nucleotide sequence of one strand is 5′–**GTATCC**–3′, the complementary strand sequence will be 3′–**CATAGG**–5′ (or, by convention, 5′–**GGATAC**–3′)

- Characteristic sequences placed in position 5′ or 3′ with respect to a given reference point (f.i., the starting point of a gene) are commonly defined upstream and downstream, respectively

The central dogma of molecular biology – 1

- Although the specific nucleotide sequence of a DNA molecule contains fundamental information, actually the enzymes act as catalysts, continually changing and adapting the chemical environment of the cells
- Catalysts are molecules that allow specific chemical reactions to proceed faster: they do not fray, or alter, and can therefore be used repeatedly to catalyze (or, in other words, to accelerate) the same reaction

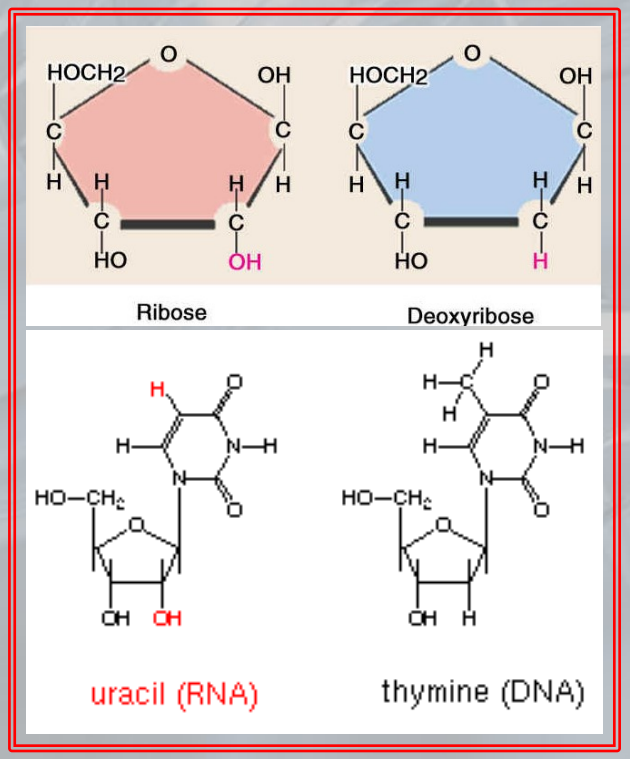
The central dogma of molecular biology – 2

- Instructions needed to describe the enzymatic catalysts — but also non–enzymatic proteins — produced by the cell are contained in genes
- The process of extracting information from genes, for the construction of proteins, is shared by all living organisms
- In detail: the information preserved in the DNA is used to synthesize a transient single–strand polynucleotide molecule, called RNA (ribonucleic acid) which, in turn, is employed for synthesizing proteins

The central dogma of molecular biology – 3

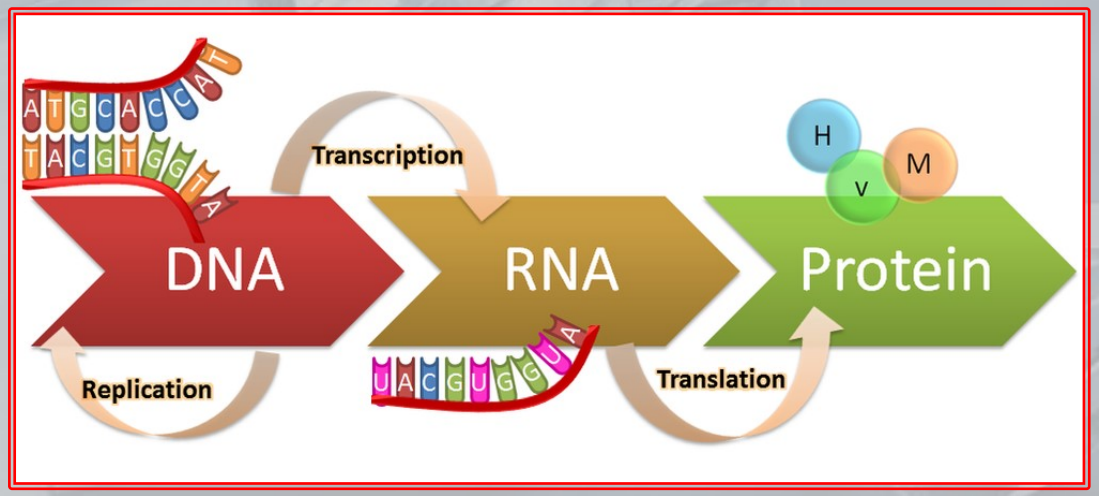
- The process of copying a gene into an RNA is called transcription and is realized through the enzymatic activity of an RNA polymerase
  - A one–to–one correspondence is established between the nucleotides **G**, **A**, **T**, **C** of the DNA and **G**, **A**, **U** (uracil), **C** of the RNA

RNA vs DNA

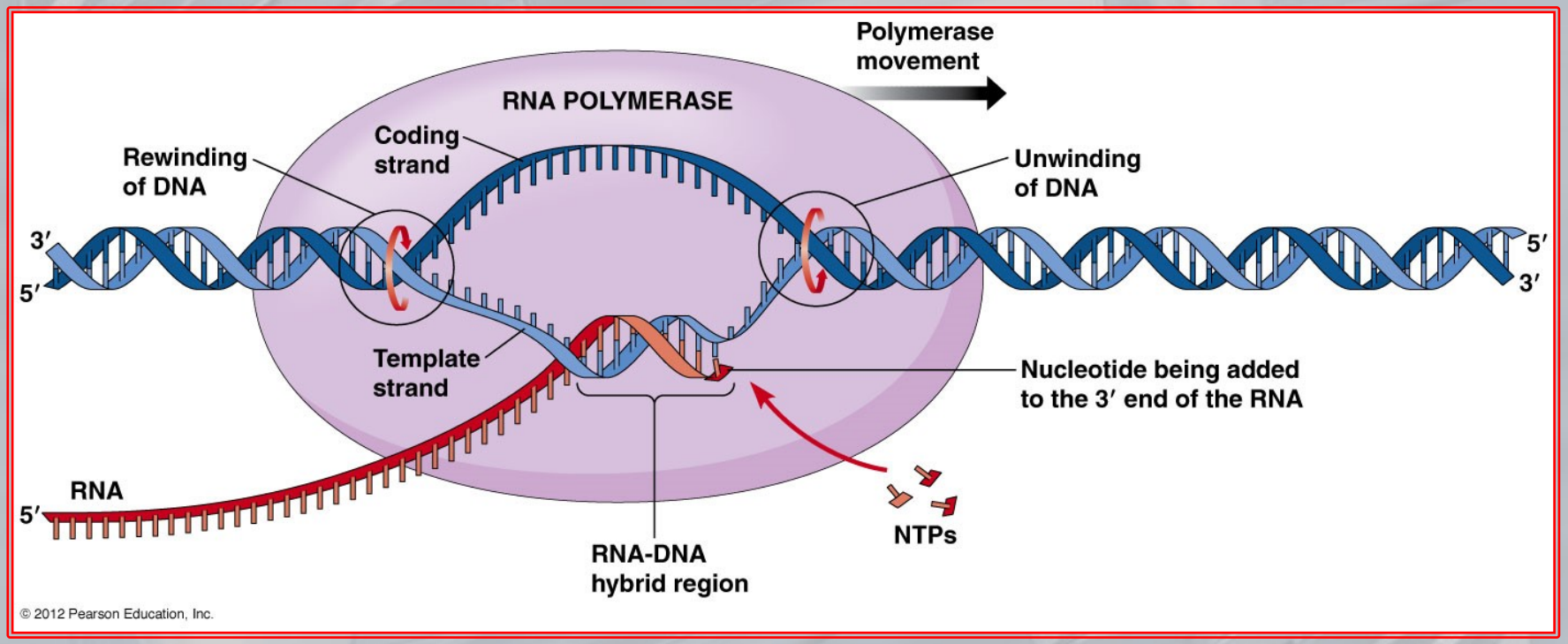


The central dogma of molecular biology – 4

- The conversion process from the RNA nucleotide sequence to the amino acid sequence, that constitutes the protein, is called translation; it is carried out by an ensemble of proteins and RNA, called ribosomes



The central dogma of molecular biology – 5

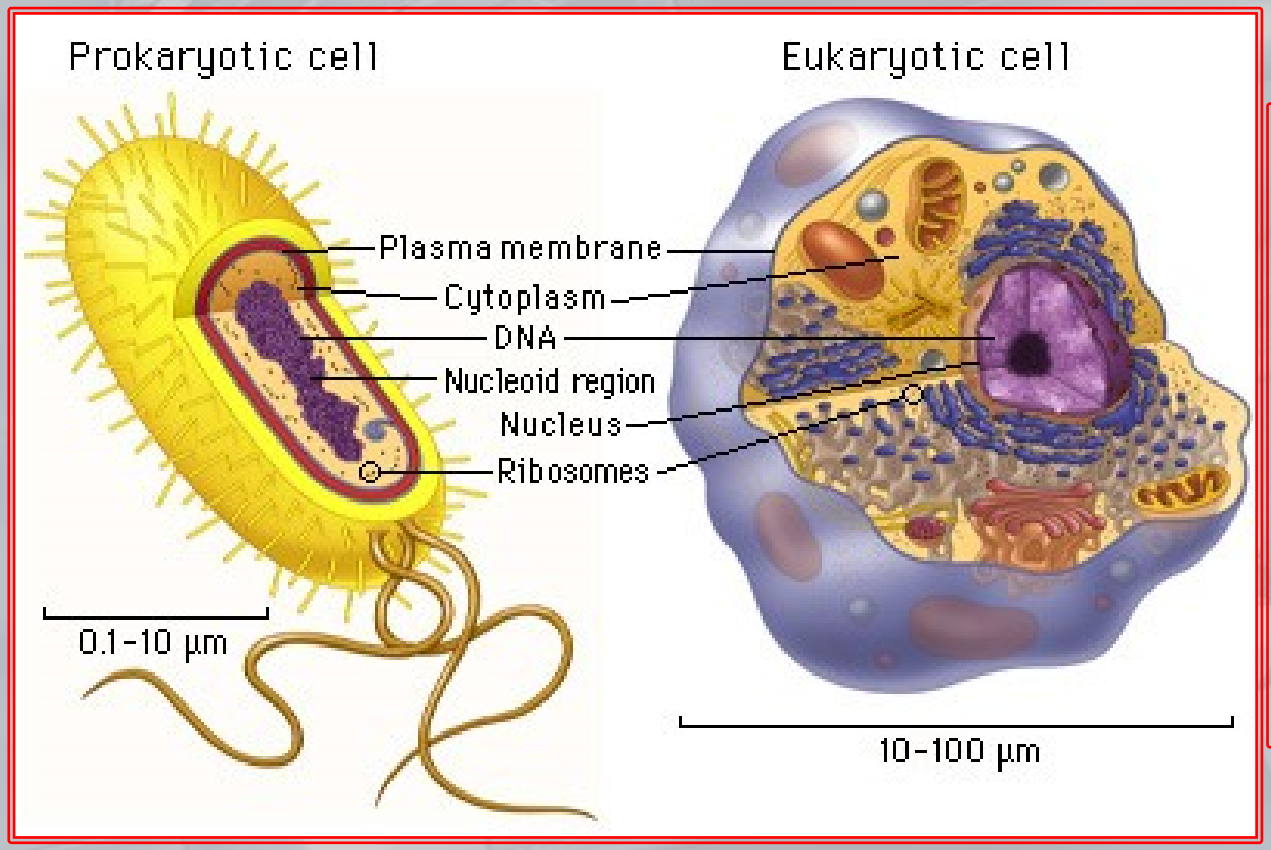


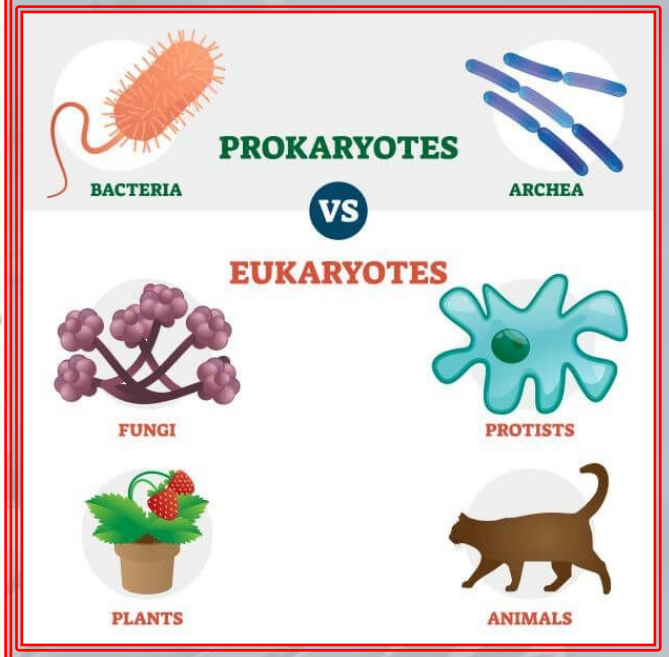
Transcription from DNA to RNA via RNA polymerase

Gene structure – 1

- All the cells interpret the genetic instructions in the same way, thanks to the presence of specific signals used as punctuation marks between genes
- The DNA code was developed at the very beginning of the history of life on the Earth and has undergone few changes over the course of millions of years
- Both prokaryotic organisms (bacteria) and eukaryotes (complex organisms like yeasts, plants, animals and humans) use the same alphabet of nucleotides and almost the same format and methods to store and use the genetic information

Gene structure – 2





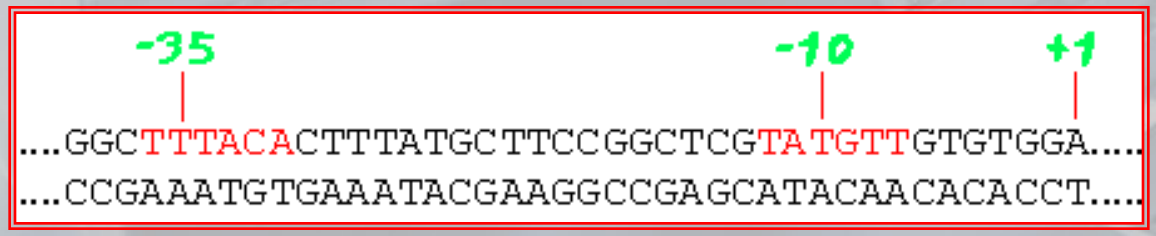
Structure of a prokaryotic cell (left) and of a eukaryotic animal cell (right)

Gene structure – 3

- The gene expression is the process during which the information stored in the DNA is used to construct an RNA molecule that, in turn, encodes for a protein
  - Significant energy wasted by the cell
- Organisms that express unnecessary proteins compete unfavorably for their subsistence, compared to organisms that regulate their gene expression more effectively
- RNA polymerases are responsible for the activation of the gene expression through the synthesis of RNA copies of the gene

Gene structure – 4

- The RNA polymerase must be able to:
  - reliably distinguish the beginning of the gene
  - determine which genes encode for needfull proteins
- The RNA polymerase cannot search for a particular nucleotide, because each nucleotide is randomly located in any part of the DNA
  - The prokaryotic RNA polymerase examines a DNA sequence searching a specific sequence of 13 nucleotides
    - 1 nucleotide serves as the start site for transcription
    - 6 nucleotides are located 10 nucleotides upstream of the start site
    - 6 nucleotides are located 35 nucleotides upstream of the start site
  - Promoter sequence of the prokaryotic gene



Gene structure – 5

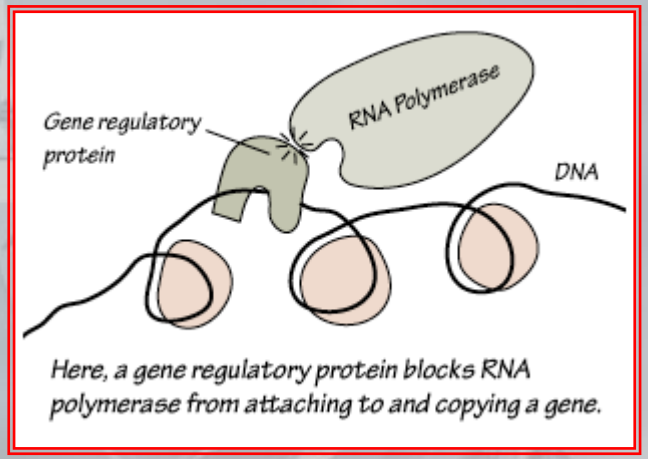
- Because prokaryotic genomes are long few million nucleotides, promoter sequences, that can be randomly found only once every 70 million nucleotides (i.e. $\sim(1/4)^{13}$), allow RNA polymerase to identify the beginning of a gene in a statistically reliable way
- The genome of eukaryotes is several orders of magnitude longer: the RNA polymerase must recognize longer and more complex promoter sequences in order to locate the beginning of a gene

Gene structure – 6

- In the early '60s, F. Jacob and J. Monod – two French biochemists – were the first to obtain experimental evidence on how cells distinguish between genes that should or should not be transcribed
- Their work on the regulation of prokaryotic genes (Nobel 1965, with A. Lwoff) revealed that the expression of the structural genes — coding for proteins involved in cell structure and metabolism — is controlled by specific regulatory genes
  - Proteins encoded by the regulatory genes are able to bind to cellular DNA only near the promoter sequence of structural genes whose expression they control, and only if it is required by the interaction with the external environment

Gene structure – 7

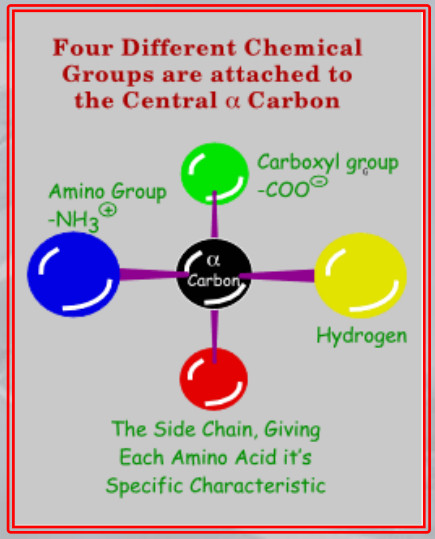
- When the chemical bond established between DNA and regulatory proteins facilitates the initiation of the transcription phase, by the RNA polymerase, it occurs a positive regulation; instead, there is a negative regulation when such bond prevents the transcription
  - The structural genes in prokaryotes are activated/deactivated by one or two regulatory proteins
  - Eukaryotes use an ensemble of seven (or more) proteins to realize the gene expression regulation



The genetic code – 1

- While nucleotides are the basic units used by the cell to maintain the information (DNA) and to build molecules capable to transfer it (RNA), amino acids, instead, are the basic bricks to build up proteins
- The protein function is strongly dependent on the order in which the amino acids are translated and bonded by ribosomes

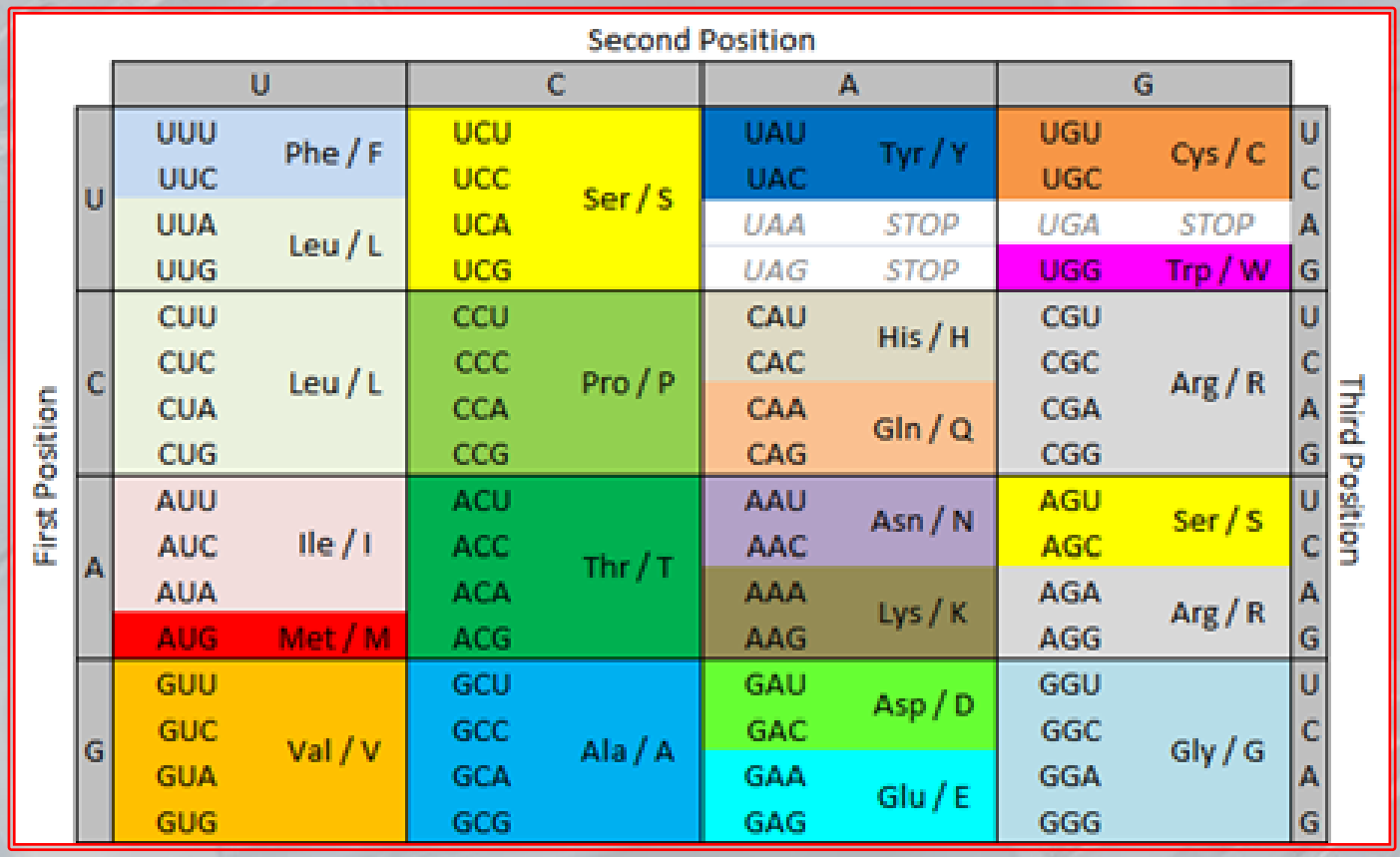
Chemical structure of an amino acid: the amino group, the α–carbon, and the carboxyl group (besides the hydrogen) are identical for all the amino acids, which differ instead with respect to the R–group (side chain)



The genetic code – 2

- However, even if in constructing DNA and RNA only four nucleotides are used, for the protein synthesis, 20 amino acids are needed
- Therefore, there cannot exist a one–to–one relation between nucleotides and amino acids, for which they encode, nor can there be a correspondence between pairs of nucleotides (at most $4^2=16$ different combinations) and amino acids
- Ribosomes must use a code based on triplets
  - With only three exceptions, each group of three nucleotides, called a codon, in an RNA copy of the coding portion of a gene, corresponds to a specific amino acid
  - The three codons that do not tell the ribosomes to insert an amino acid are called stop codons

The genetic code – 3

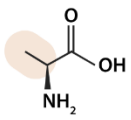


First Position	Second Position: U		Second Position: C		Second Position: A		Second Position: G		Third Position
U	UUU	Phe / F	UCU	Ser / S	UAU	Tyr / Y	UGU	Cys / C	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu / L	UCA		*UAA*	*STOP*	*UGA*	*STOP*	A
	UUG		UCG		*UAG*	*STOP*	UGG	Trp / W	G
C	CUU	Leu / L	CCU	Pro / P	CAU	His / H	CGU	Arg / R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln / Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile / I	ACU	Thr / T	AAU	Asn / N	AGU	Ser / S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys / K	AGA	Arg / R	A
	AUG	Met / M	ACG		AAG		AGG		G
G	GUU	Val / V	GCU	Ala / A	GAU	Asp / D	GGU	Gly / G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu / E	GGA		A
	GUG		GCG		GAG		GGG		G

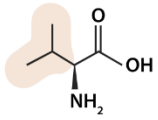
For a limited number of bacterial genes, **UGA** encodes a twenty–first amino acid, selenocysteine; a twenty–second amino acid, pyrrolysine, is encoded by **UAG** in some bacterial and eukaryotic species

The genetic code – 4

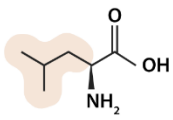
Non-polar side chains, uncharged, hydrophobic



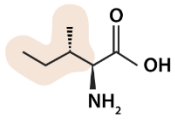
Alanine (Ala, A)
MW: 89,09
pI: 6,01
C3H7N1O2



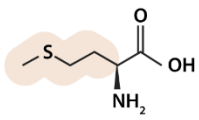
Valine (Val, V)
MW: 117,15
pI: 6,00
C5H11N1O2



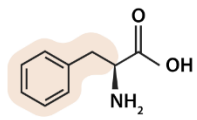
Leucin (Leu, L)
MW: 131,17
pI: 6,01
C6H13N1O2



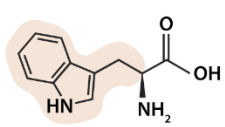
Isoleucine (Ile, I)
MW: 131,17
pI: 6,05
C6H13N1O2



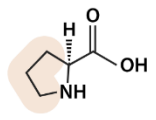
Methionine (Met, M)
MW: 149,21
pI: 5,74
C5H11N1O2S1



Phenylalanine (Phe, F)
MW: 165,19
pI: 5,49
C9H11N1O2



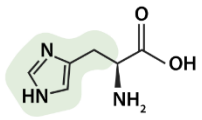
Tryptophan (Trp, W)
MW: 204,23
pI: 5,89
C11H12N2O2



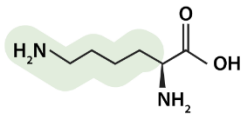
Proline (Pro, P)
MW: 115,13
pI: 6,30
C5H9N1O2

Electrically charged side chains

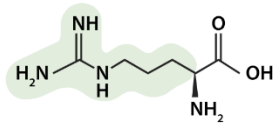
Basic



Histidine (His, H)
MW: 155,16
pI: 7,60
C6H9N3O2

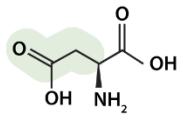


Lysine (Lys, K)
MW: 146,19
pI: 9,60
C6H14N2O2

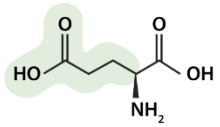


Arginine (Arg, R)
MW: 174,20
pI: 10,76
C6H14N4O2

Acidic

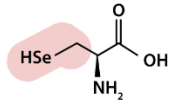


Aspartaic Acid (Asp, D)
MW: 133,1
pI: 2,85
C4H7N1O4



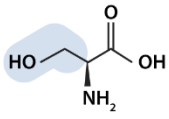
Glutamic Acid (Glu, E)
MW: 147,13
pI: 3,15
C5H9N1O4

Special amino acids

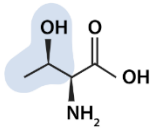


Selenocysteine (Sec, U)
MW: 168,07
pI: 3,9
C3H7N1O2Se

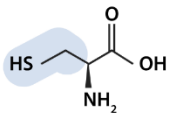
Polar side chains, uncharged



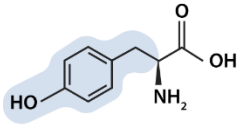
Serine (Ser, S)
MW: 105,09
pI: 5,68
C3H7N1O3



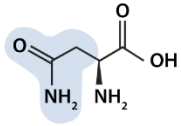
Threonine (Thr, T)
MW: 119,12
pI: 5,60
C4H9N1O3



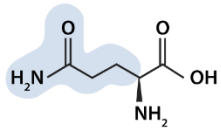
Cysteine (Cys, C)
MW: 121,16
pI: 5,05
C3H7N1O2S1



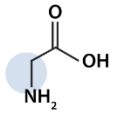
Tyrosine (Tyr, Y)
MW: 181,19
pI: 5,64
C9H11N1O3



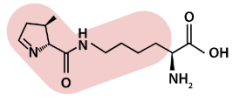
Asparagine (Asn, N)
MW: 132,12
pI: 5,41
C4H8N2O3



Glutamine (Gln, Q)
MW: 146,15
pI: 5,65
C5H10N2O3

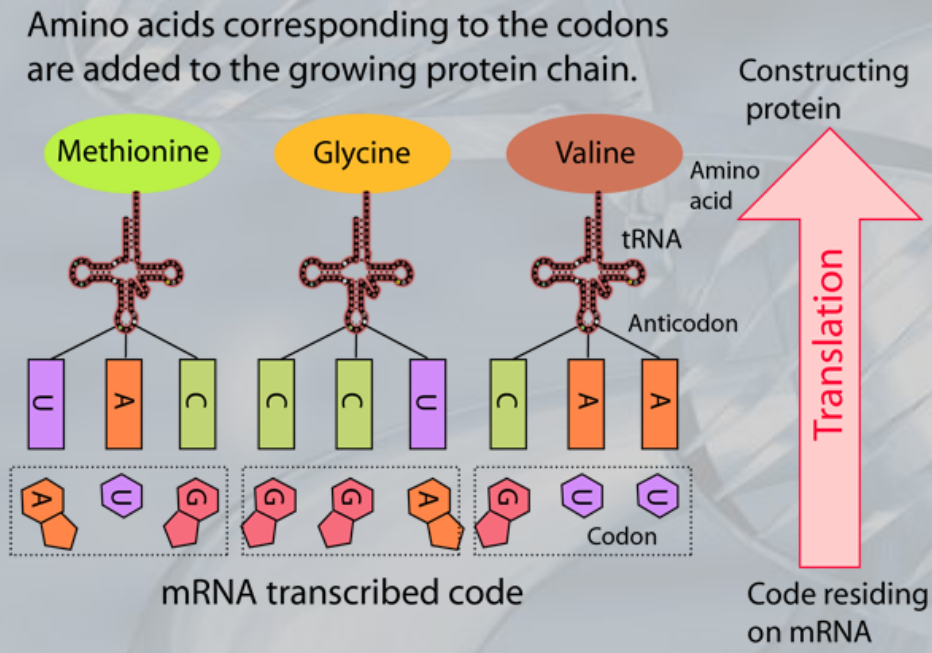


Glycine (Gly, G)
MW: 75,07
pI: 6,06
C2H5N1O2

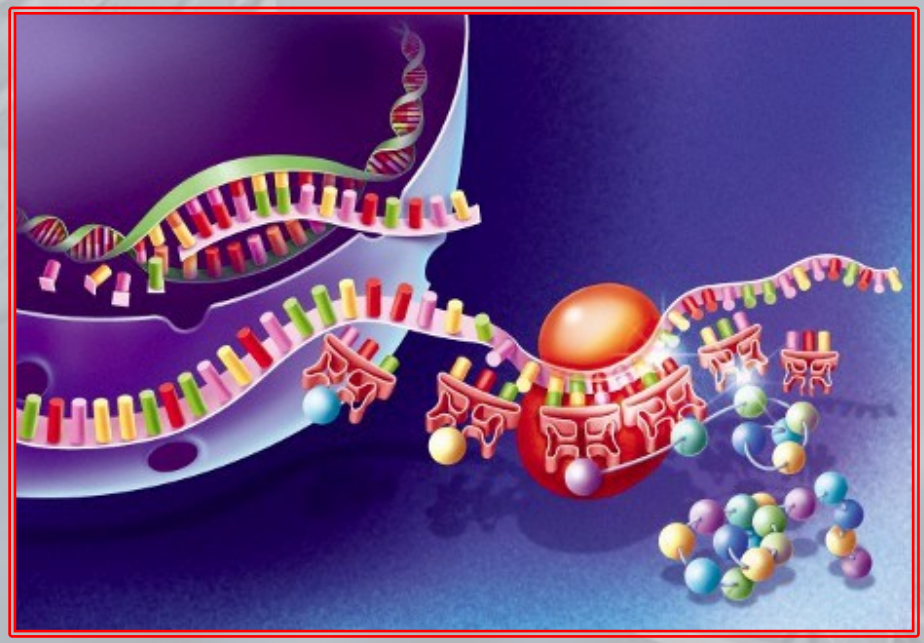


Pyrrolysine (Pyl, O)
MW: 255,31
pI:
C12H21N3O3

The genetic code – 5



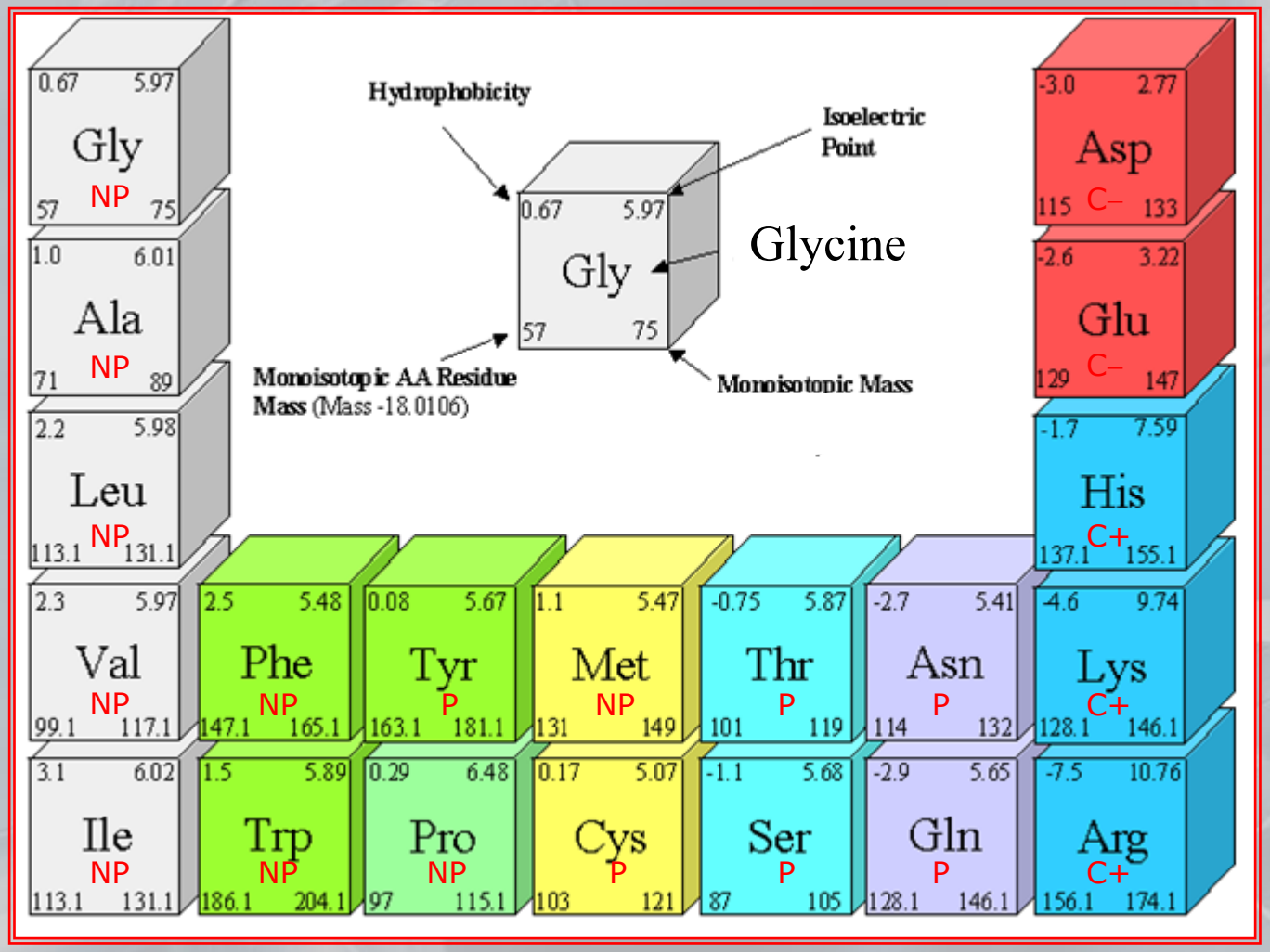
The translation process



The genetic code – 6

- Amino acids are classified into four different categories
  - Nonpolar R–groups, hydrophobic: glycine, alanine, valine, leucine, isoleucine, methionine, phenyla-lanine, tryptophan, proline
  - Polar R–groups, hydrophilic: serine, threonine, cysteine, tyrosine, asparagine, glutamine
  - Acid (negatively charged): aspartic acid, glutamic acid
  - Basic (positively charged): lysine, arginine, histidine

The genetic code – 7

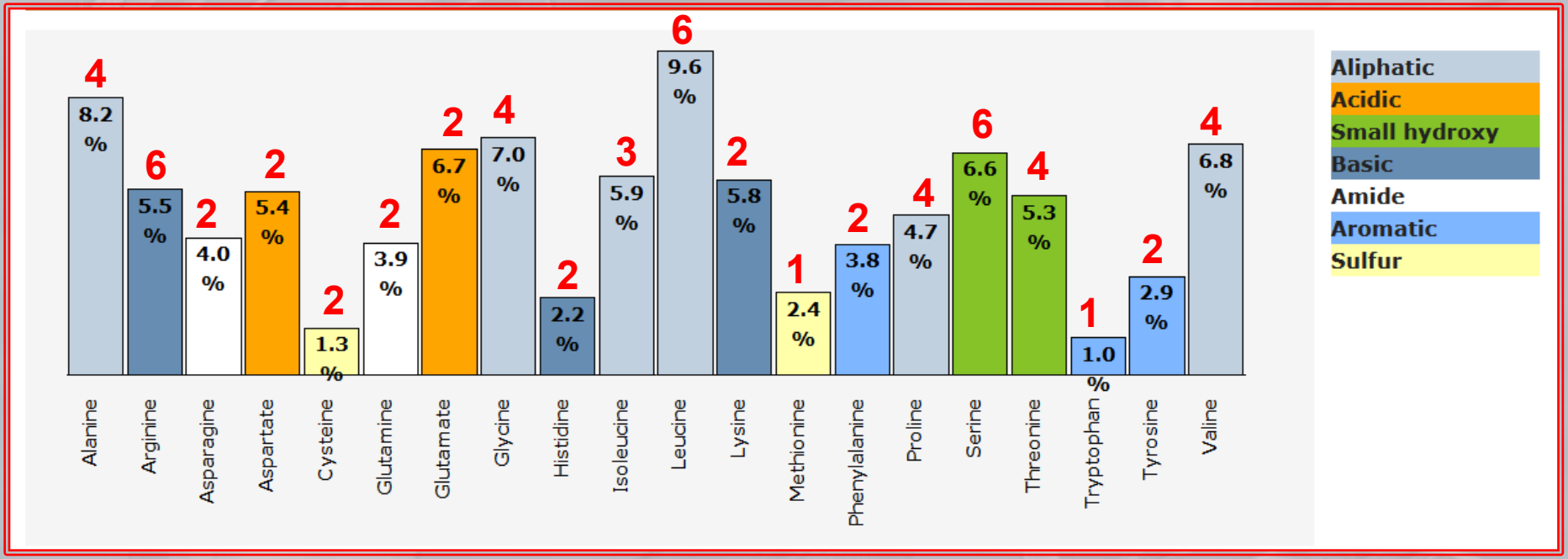


- Hydrophobic
- Aromatic
- Imino acid
- Sulfur
- Subject to O–glycosilation
- Hydrophilic
- Positively charged
- Negatively charged

The genetic code – 8

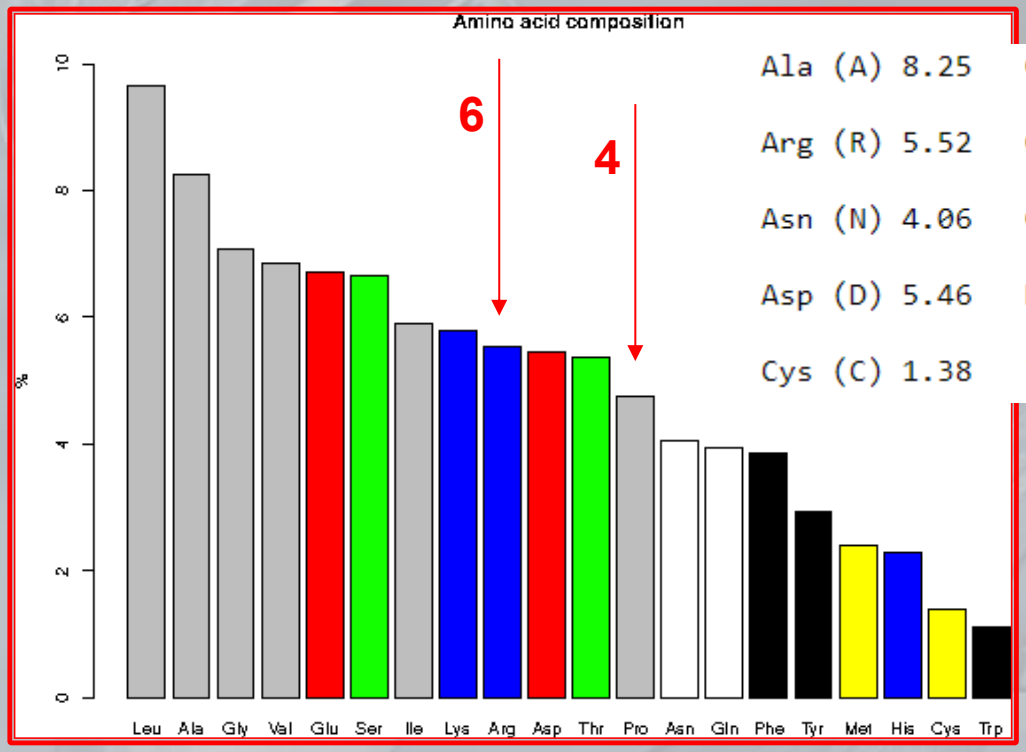
- 18 out of 20 amino acids are encoded by more than one codon – six at most, for leucine, serine, and arginine: this redundancy of the genetic code is called degeneracy
  - A single change in a codon is usually insufficient to cause the encoding of an amino acid of a different class (especially if it happens in the third position)
  - That is, during the replication/transcription of the DNA, mistakes may occur that do not affect the amino acid composition of the protein
  - The genetic code is very robust and minimize the consequences of possible errors in the nucleotide sequence, avoiding devastating impacts on the encoded protein function

The genetic code – 9



Amino acid composition (Release 2023_04 of UniProtKB/Swiss-Prot, containing 570,157 sequence entries https://web.expasy.org/docs/relnotes/relstat.html)

The genetic code – 10



Ala (A) 8.25	Gln (Q) 3.93	Leu (L) 9.64	Ser (S) 6.65
Arg (R) 5.52	Glu (E) 6.71	Lys (K) 5.80	Thr (T) 5.36
Asn (N) 4.06	Gly (G) 7.07	Met (M) 2.41	Trp (W) 1.10
Asp (D) 5.46	His (H) 2.27	Phe (F) 3.86	Tyr (Y) 2.92
Cys (C) 1.38	Ile (I) 5.91	Pro (P) 4.74	Val (V) 6.85

Amino acid composition (2024_04, containing 571,864 sequence entries)

Arg/R	CGU, CGC, CGA, CGG, AGA, AGG

Notice that: Arginine is the amino acid at the lowest concentration among those specified by 6 codons
...but it is very important, because it acts as a plug for proteins that need to bind to DNA

Pro/P	CCU, CCC, CCA, CCG

Notice that: Proline is the amino acid at the lowest concentration among those specified by 4 codons
...but it is very important for the formation of secondary structures in proteins

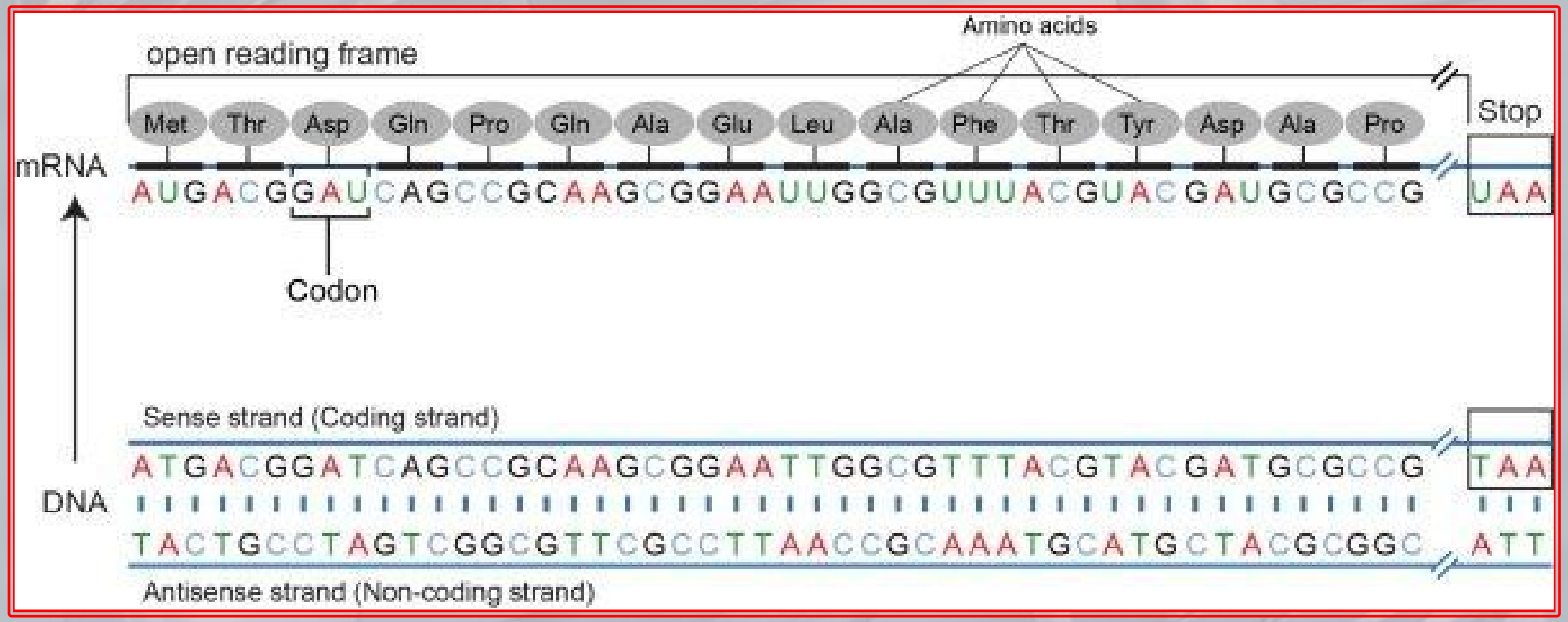
The genetic code – 11

- The translation is realized by the ribosomes from a start site, located on the RNA copy of a gene, and proceeds until the first stop codon is encountered
- The start codon is the triplet **AUG** (which also encodes for methionine), both in eukaryotes and in prokaryotes
- The translation is accurate only when the ribosomes examine the codons contained inside an open reading frame (delimited by a start and a stop codon, respectively)
  - The alteration of the reading frame (except for a multiple of 3) changes each amino acid situated downstream with respect to the alteration itself, and usually causes a truncated, unfunctional, version of the protein

The genetic code – 12

- Most of the genes encodes for proteins composed by hundreds of amino acids
- In a randomly generated sequence, stop codons will appear approximately every 20 triplets (3 codons out of 64), whereas open reading frames, representing genes, usually are very long sequences not containing stop codons
  - Open Reading Frames (ORFs) are a distinctive feature of many genes in both prokaryotes and eukaryotes (w.r.t. mRNA)

The genetic code – 13

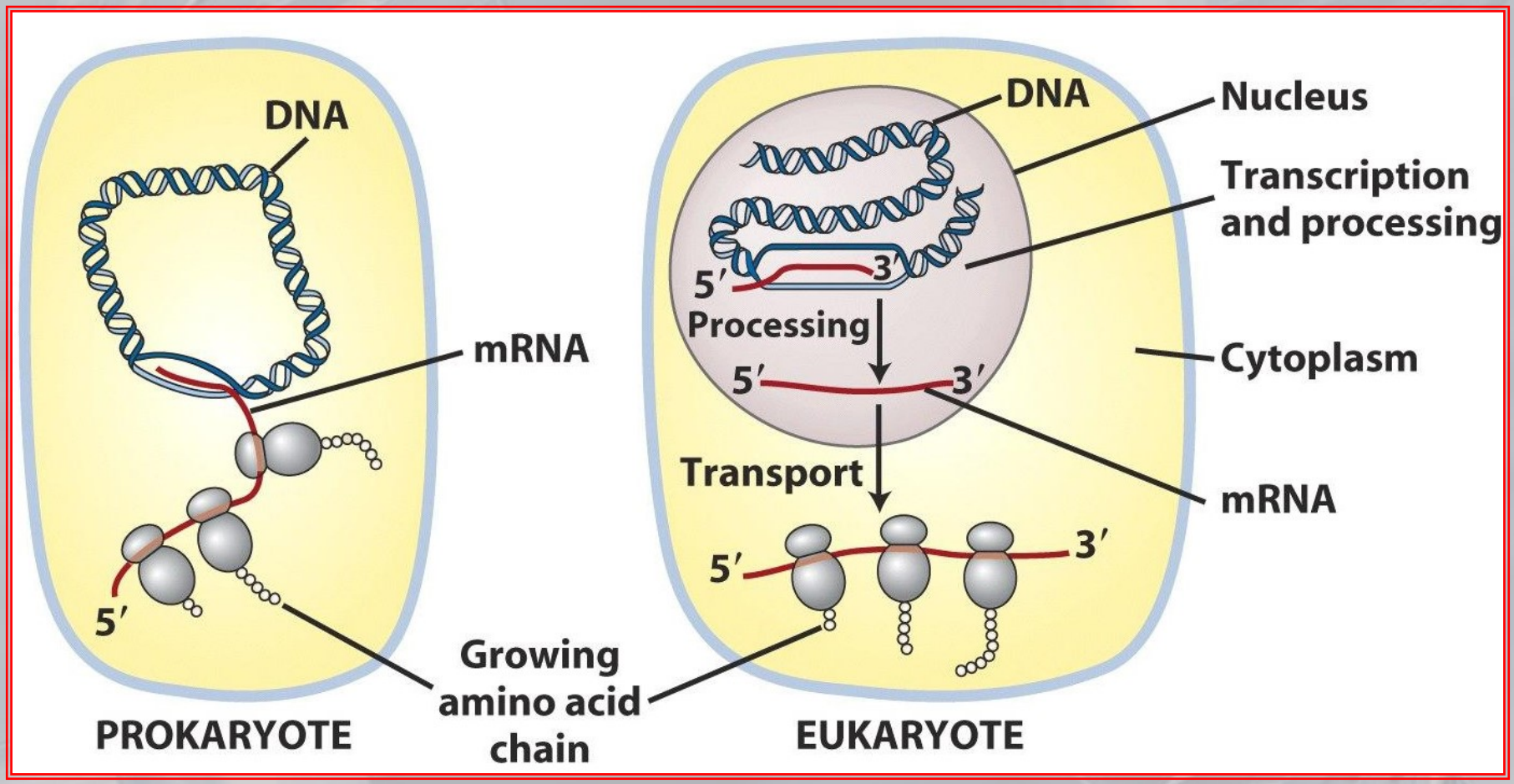


An open reading frame is a portion of an RNA molecule that, when translated into amino acids, contains no stop codons; a long ORF is likely to represent a gene (95% of confidence for a length ≥ 60)

Introns and exons – 1

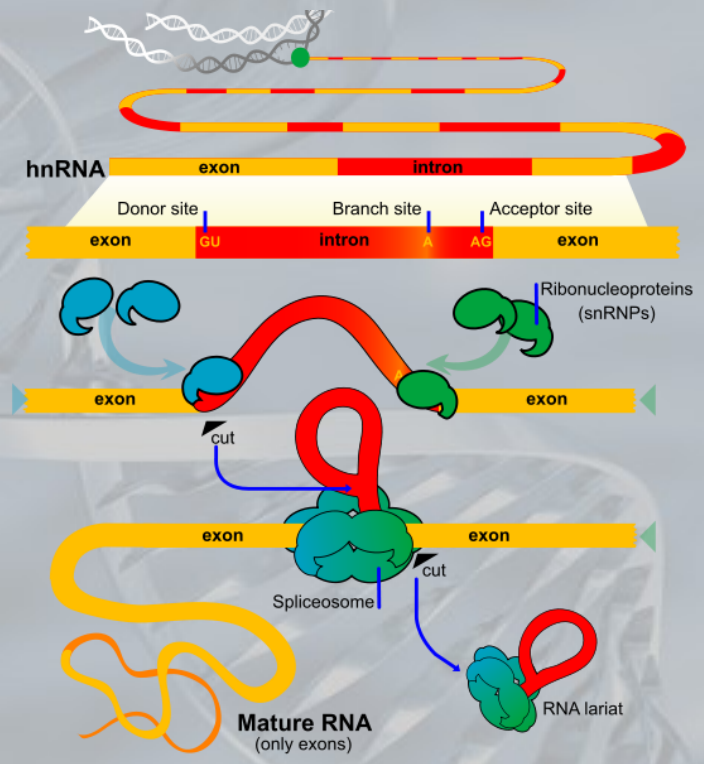
- The messenger RNA (mRNA) copies of prokaryotic genes correspond perfectly to the DNA sequences of the genome (with the exception of uracil, which is used in place of thymine), and the steps of transcription and translation are partially overlapped
- In eukaryotes, the two phases of gene expression are physically separated by the nuclear membrane: the transcription occurs in the nucleus, whereas the translation starts only after that the mRNA has been transported into the cytoplasm
  - RNA molecules transcribed by the eukaryotic polymerase may be significantly modified before that ribosomes initiate their translation

Introns and exons – 2

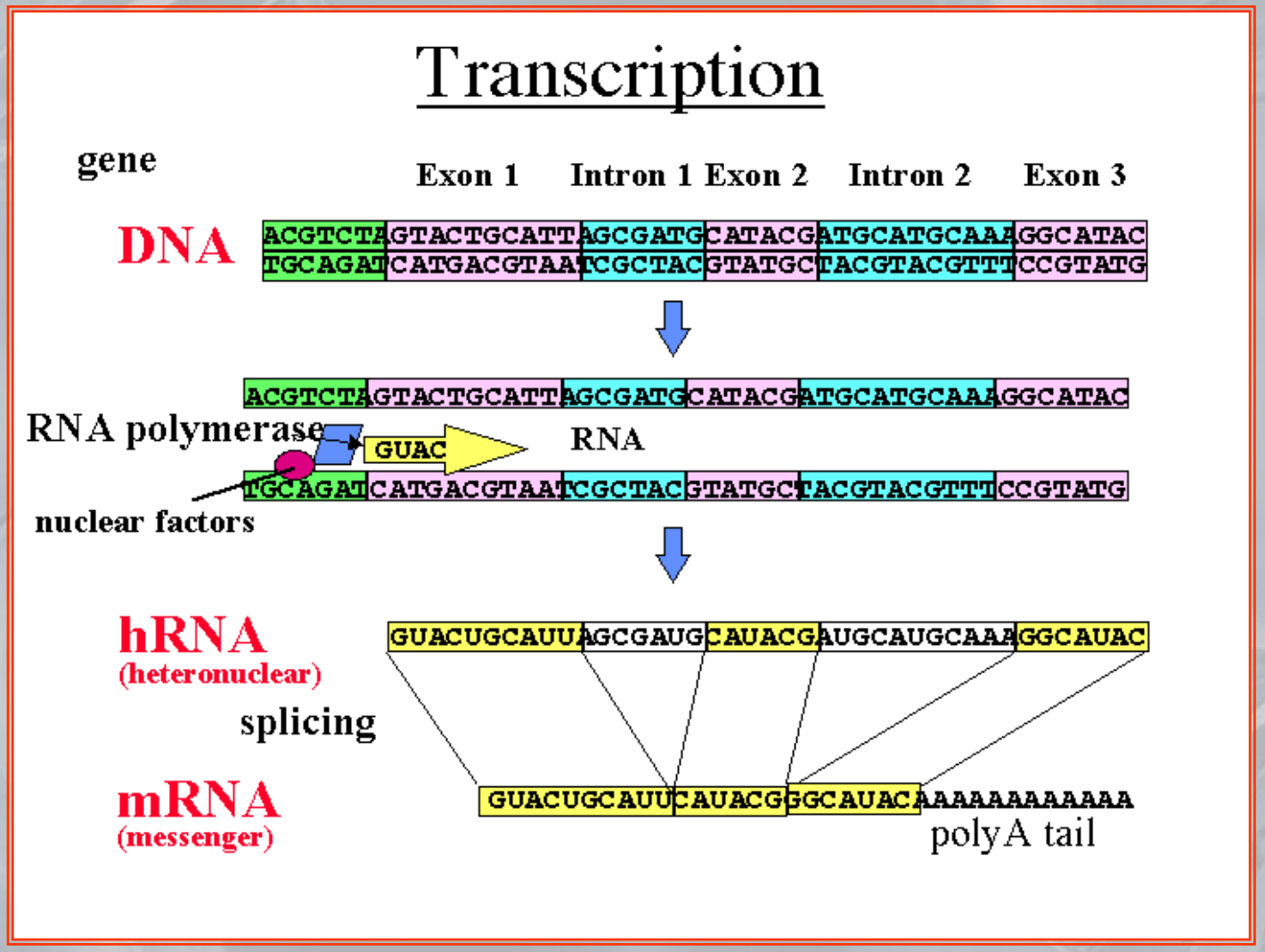


Introns and exons – 3

- The most striking change that affects the primary transcript of eukaryotic genes is splicing, which involves the cutting of introns and the successive reconnection of flanking exons
- Most eukaryotic genes contain a very high number of introns
- Example: the gene associated with the human *cystic fibrosis disease* (cystic fibrosis trans-membrane conductance regu-lator, CFTR) has 24 introns and contains more than a million base pairs, while the mRNA translated by the ribosomes is made up of about a thousand nucleotides



Introns and exons – 4



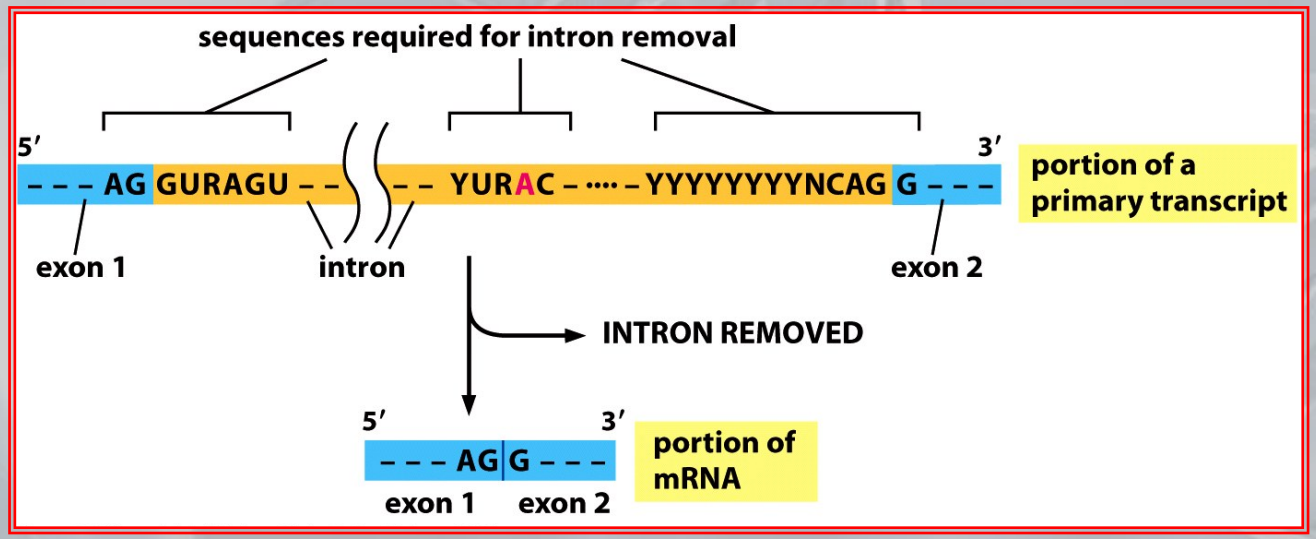
Intronic sequences degrade rapidly after their removal from the hnRNA

Introns and exons – 5

- The majority of eukaryotic introns follows the **GT–AG** rule, according to which introns normally begin with the dinucleotide **GT** and end with the dinucleotide **AG**
- Anyway, pairs of nucleotides are statistically too frequent to be considered as a sufficient signal for the recognition of introns by the splicesome (the ensemble of proteins devoted to splicing)
  - Additional nucleotides, at the endpoints 5’ and 3’ (~6) and inside the intron, are examined, which can differ for different cells

Introns and exons – 6

- A failure in the correct splicing of introns from the primary transcript of eukaryotic RNA can:
  - introduce a shift of the reading frame
  - generate the transcription of a premature stop codon
  - Inoperability of the translated protein

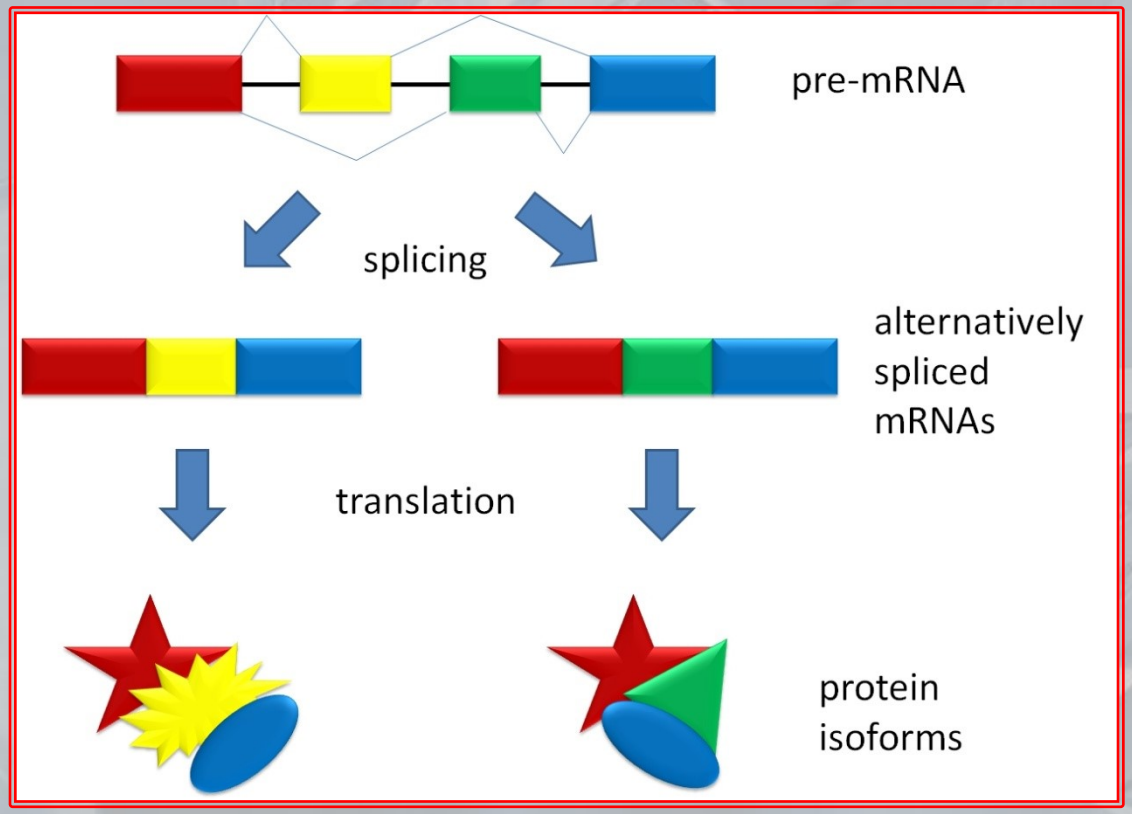


Introns and exons – 7

- *Alternative splicing*: increasing of the protein bio-diversity caused by modifications of the splicesome and by the presence of accessory proteins respons-ible for the recognition of the intron/exon neigh-borhood
  - A single gene encodes for multiple proteins (for instance in different tissues)
  - It is estimated that almost 95% of the approximately 20,000 multiexonic protein–encoding genes in the human genome undergo alternative splicing and that, on average, a given gene gives rise to four/five alternatively spliced variants – encoding a total of 90–100,000 proteins which differ in their sequence and therefore, in their activities

Introns and exons – 8

- Exons from the same gene are joined in different combinations, leading to different, but related, mRNA transcripts: translation of different protein isoforms with distinct structures and functions — all from a single gene



Protein isoforms

The protein function – 1

- The functions of proteins are incredibly diversified
  - Structural proteins, such as collagen, provide support to the bones and connective tissues
  - Enzymes act as biological catalysts, like pepsin, which regulates the food metabolism
  - Proteins are also responsible for the transport of atoms and small molecules through the body (hemoglobin), for signals and intercellular communications (insulin), for the absorption of photons by the visual apparatus (rhodopsin), for the activation of muscles (actin and myosin), etc.

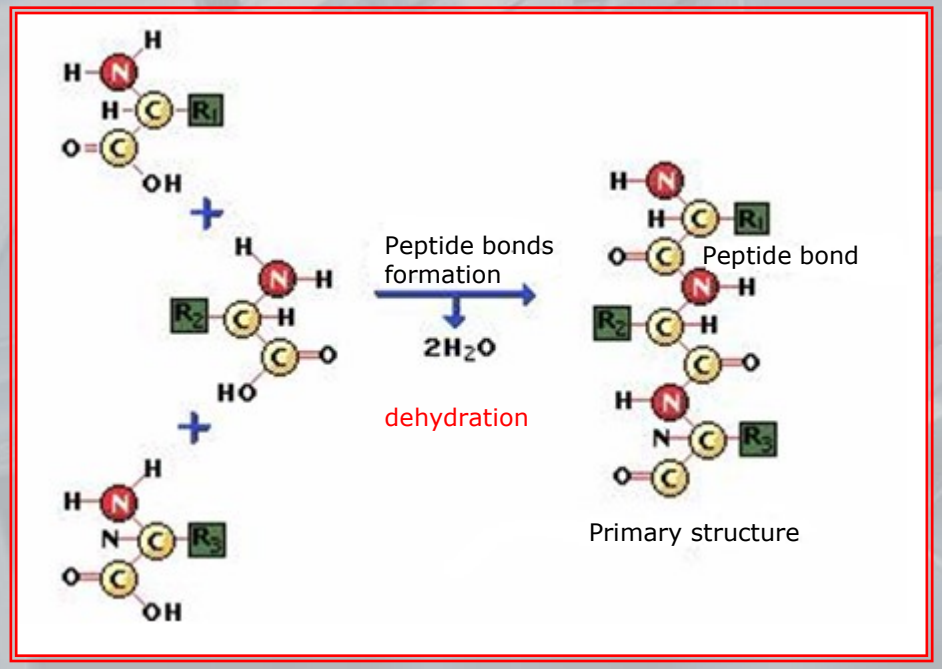
The protein function – 2

The primary structure – 1

- Following the instructions contained in the mRNA, ribosomes translate a linear polymer (chain) of amino acids
- The 20 amino acids share a similar chemical structure and differ only according to the R–group
  - The structural region common to all amino acids is called the main chain or the backbone, while the different R–groups constitute side chains
- The linear, ordered, chain in which the amino acids are assembled in a protein constitutes its primary structure

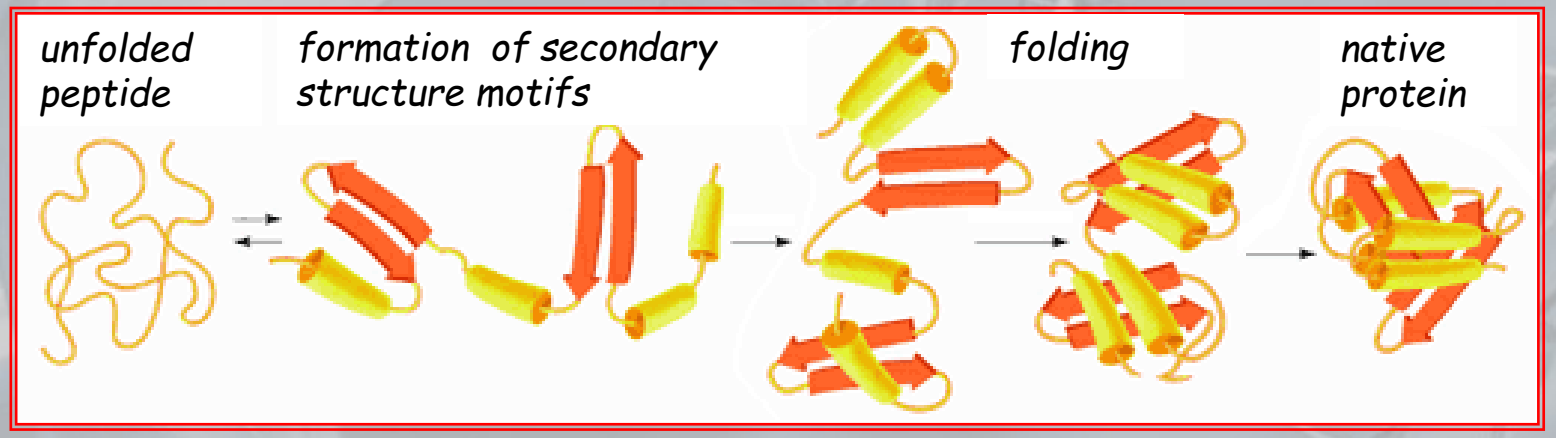
The primary structure – 2

- The protein chain has a direction
  - At one end of the chain there is a free amino group (NH_2), while, at the other termination, there is a carboxyl group (COOH)
  - The chain goes from the amino–terminal (which is the first to be synthesized) to the carboxy–terminal



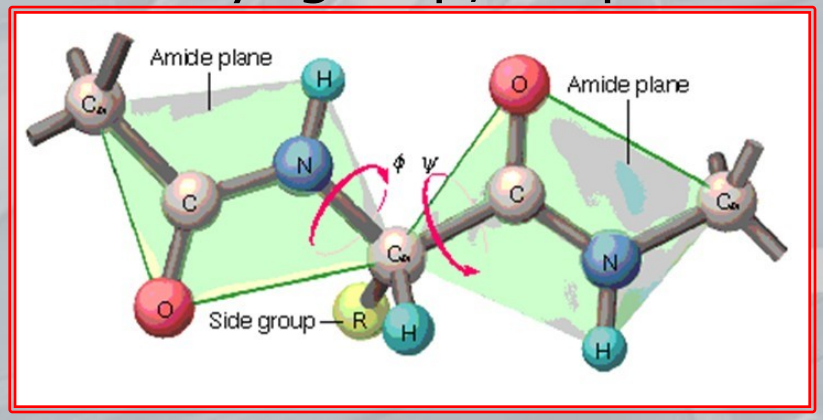
The primary structure – 3

- After the translation phase, the protein collapses, i.e., it bends and shapes itself in a complex globular 3D conformation, assuming the native structure, which, however, depends on the arrangement of the amino acids in the primary chain
  - The native structure defines the protein functionality



The primary structure – 4

- The chemical structure of the protein backbone forces its tridimensional arrangement which is, locally, basically planar
- The only movable segments in the backbone are the bonds of the α–carbon with the nitrogen and with the carbon of the carbonyl group, respectively

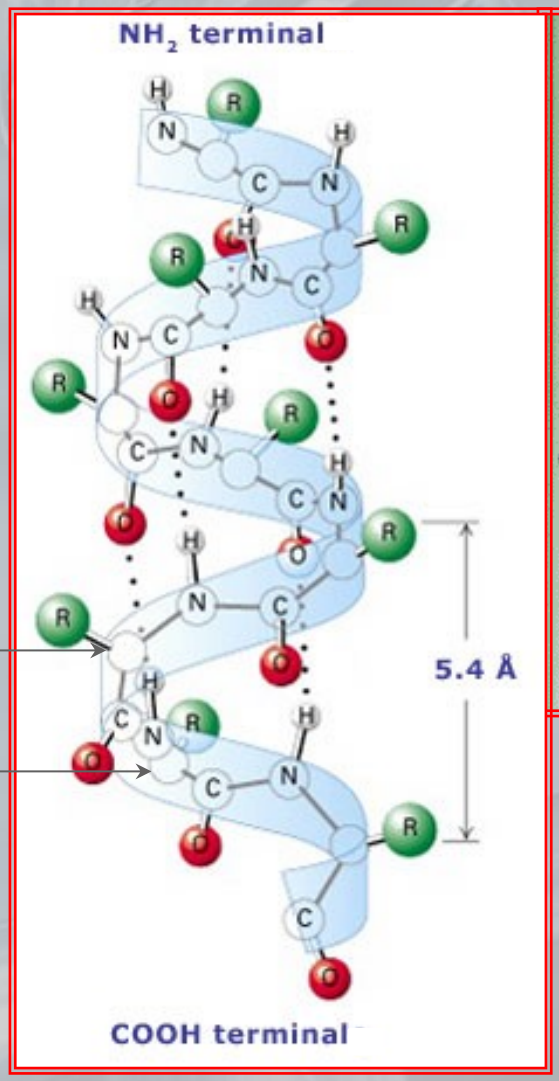


- These two links allow a circular rotation, giving rise to the dihedral angles (ϕ and ψ)
- The protein folding exclusively derives from these rotations

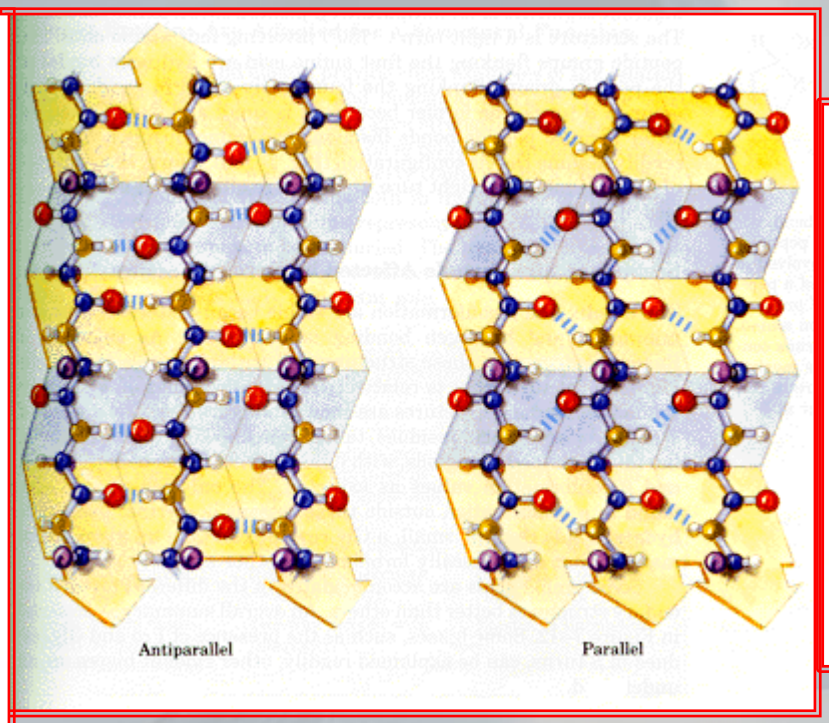
The secondary structure – 1

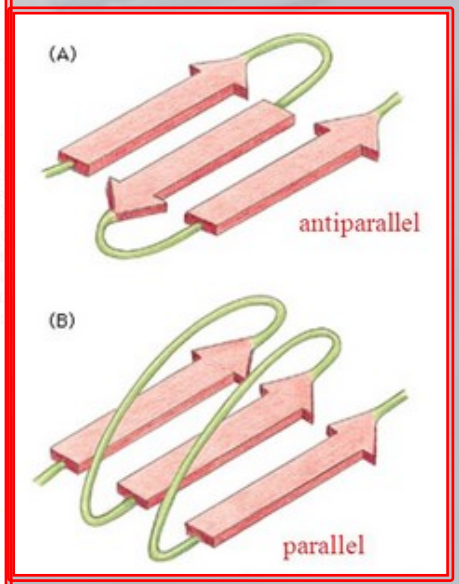
- Examining the structure of known proteins reveals that there are a small number of common local motifs
- These structures, formed by regular patterns of intramolecular hydrogen bonds, are found in almost all proteins
- The position and the direction of these regular motifs define the secondary structure of the protein
- The most common structures are α–helices (which have a helical shape similar to a spring, $\phi=\psi=-60°$, 3.6 amino acids in each turn) and β–sheets ($\phi=-135°$, $\psi=135°$)

The secondary structure – 2



α–helix



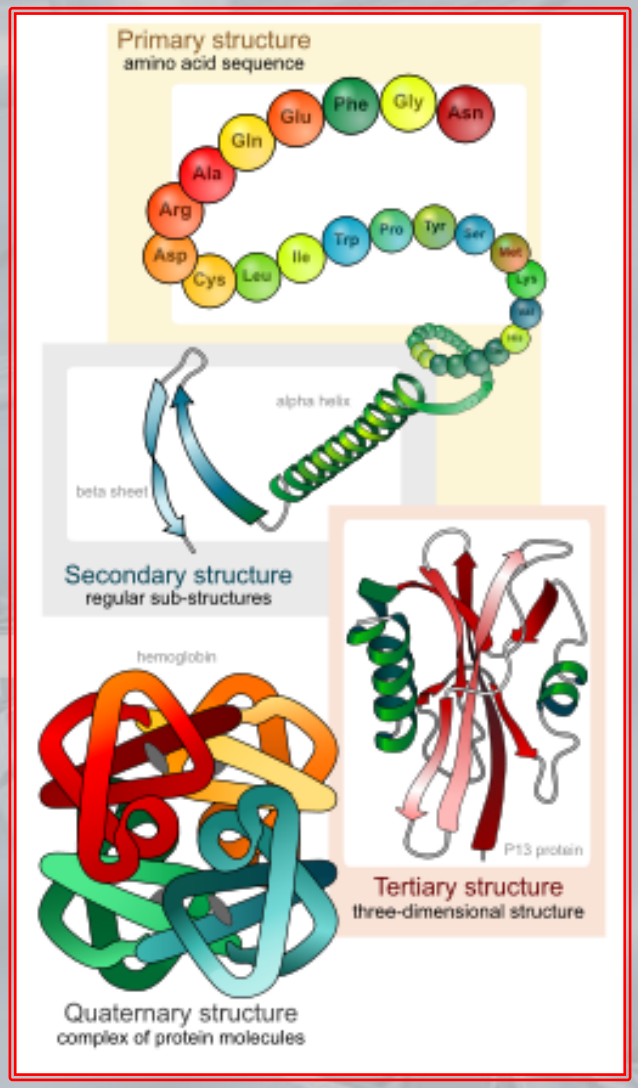


β–sheets

In the antiparallel β–sheets, adjacent filaments were oriented in the opposite direction, while in the case of parallel β–sheets, they are oriented in the same direction; to make this possible, parallel β–sheets are often composed of non–contiguous amino acids in the primary structure

Tertiary and quaternary structures

- Those backbone areas showing a secondary structure are packaged together and combined with other less structured regions, to form a 3D agglomerate, called the tertiary, native structure of the protein
- Sometimes, an active enzyme is composed of two or more proteins, which are made up together in a single large ensemble, to form the quaternary structure of the enzyme

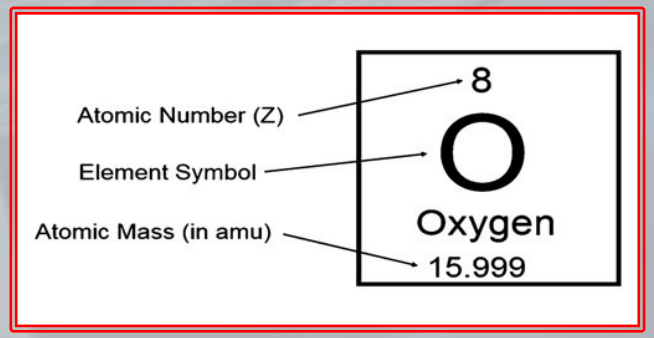


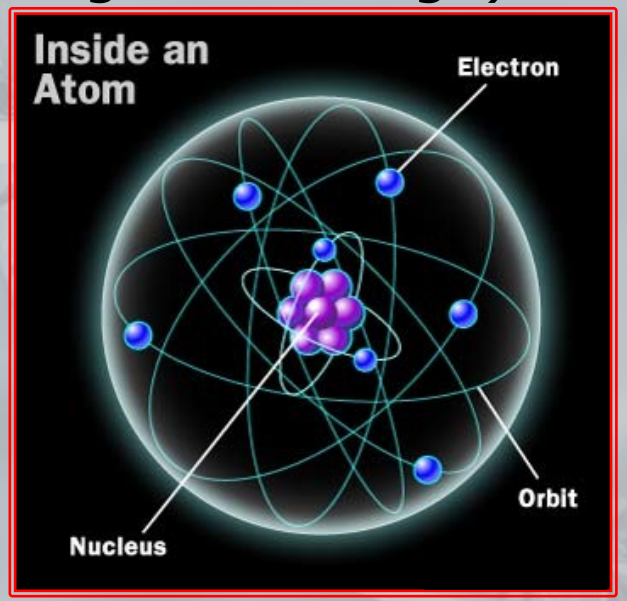
The nature of chemical bonds

- Much of what we consider to be essential for life can be traced back to a set of chemical reactions – occurring inside the cell – and to the features of the enzymes that control their speed
- By definition, an element is a pure chemical substance, that cannot be further reduced through chemical reactions
  - Elements are atoms (from the Greek word *àtomos*: *indivisible*, with *à* [alpha privative] and *tomé*, *division*), indivisible entities composed by small subatomic particles, that can be separated only through physical reactions

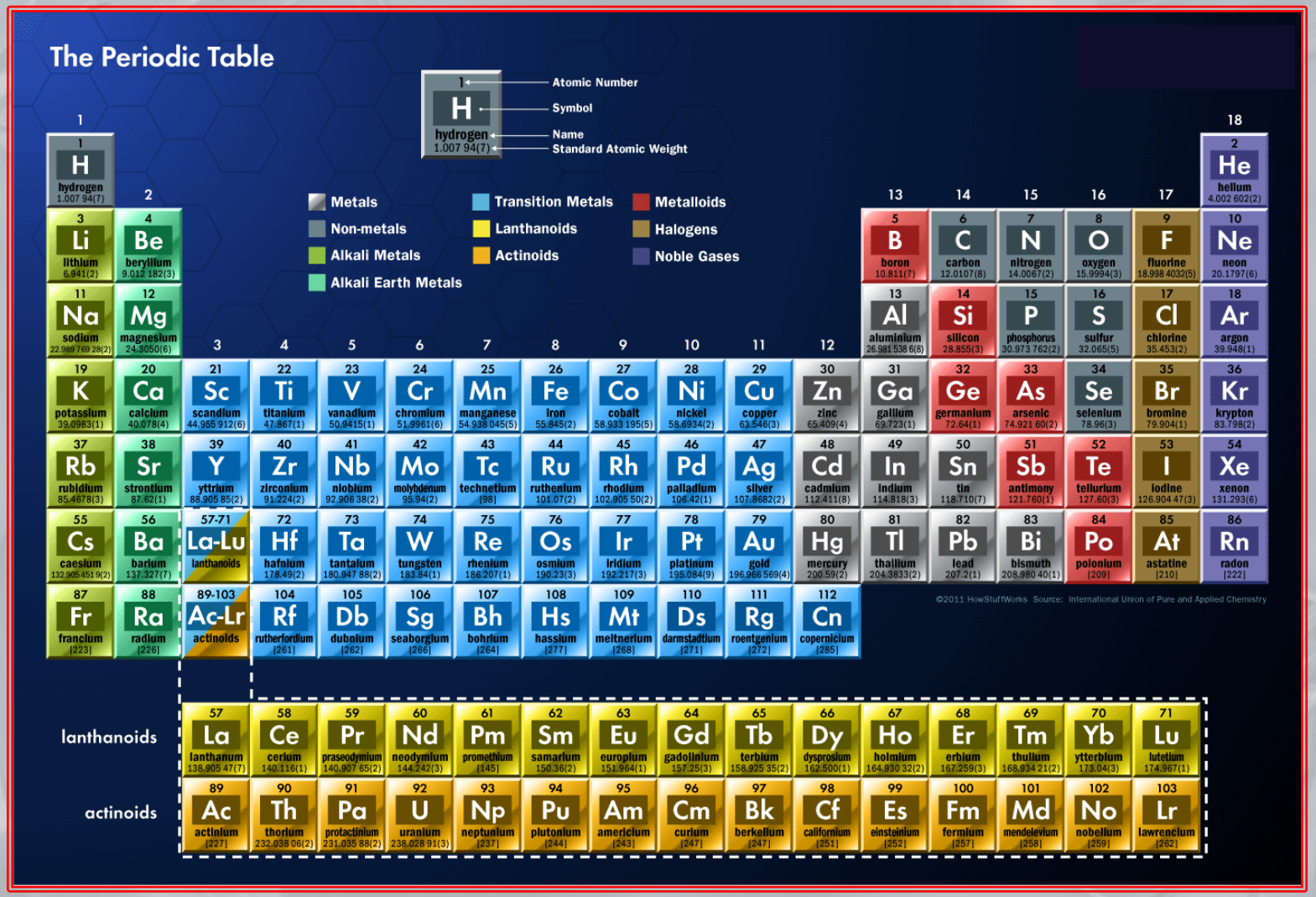
Atom's anatomy – 1

- There are hundreds of subatomic particles, but only three are stable and particularly important for the chemistry of living organisms:
  - neutrons (1.7×10^{-24} grams, with no charge)
  - protons (1.7×10^{-24} grams, with positive charge)
  - electrons (8.5×10^{-28} grams, with negative charge)
- The number of protons in the nucleus determines the element type and represents its atomic number



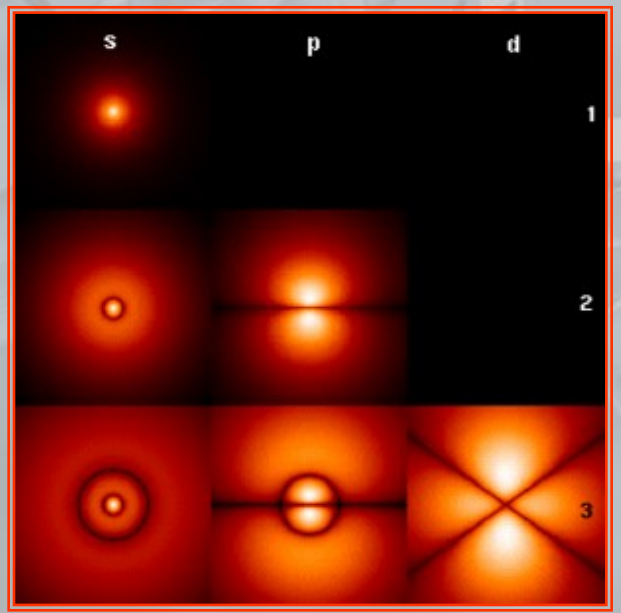


The Periodic Table of Elements



Atom's anatomy – 2

- For each proton, an electron exists which rotates on an orbital relatively far from the nucleus, at a speed close to the light speed
  - Large empty spaces inside the atom
  - Electrons do not have predictable orbits: the orbital is the three–dimensional space in which the electron passes 90% of its time



Atom's anatomy – 3

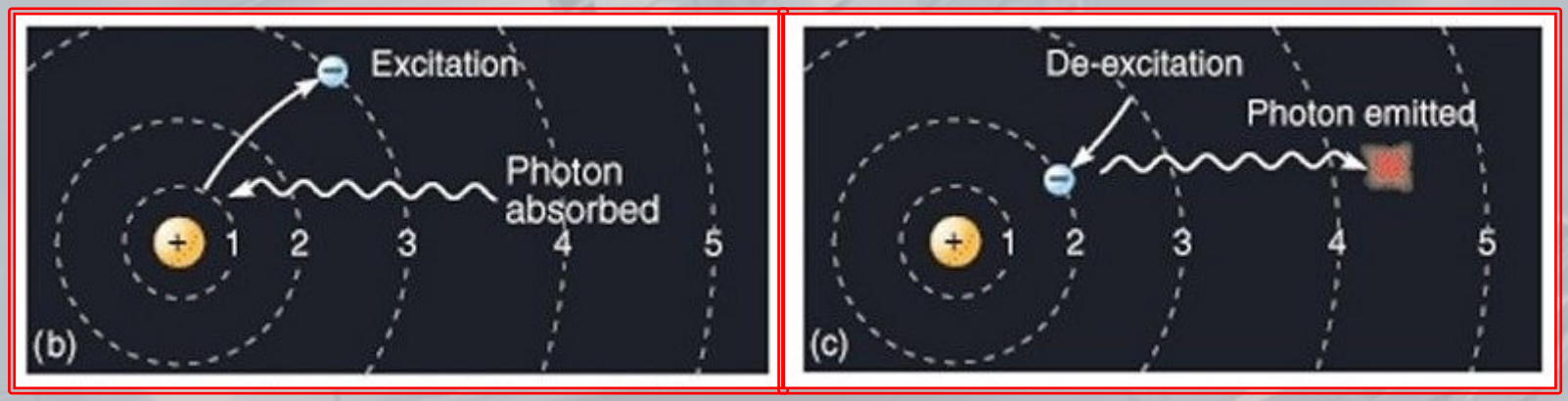
- At the atomic level, the energy is divided into "packets" of light, or photons: the more the electron is far from the nucleus, the greater is its potential energy:

$$U = -kZe^2/r$$

where $k = 8.98 \times 10^9\ \mathrm{N\,m^2\,C^{-2}}$ is the Coulomb's constant, Ze is the positive charge due to the protons (Z atomic number), and r is the distance between the electron and the nucleus

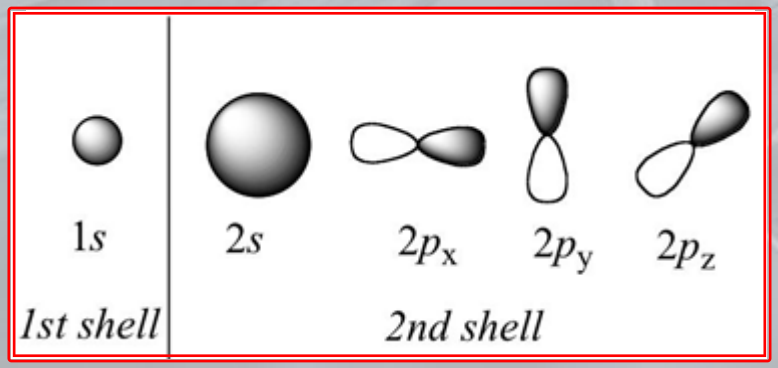
Atom's anatomy – 4

- Electrons emit light when they move from a far orbital to an orbital closer to the nucleus
  - The amount of energy required is a quantum
  - Actually, electrons realize a quantum jump



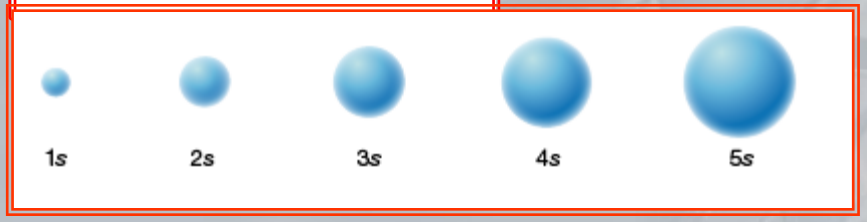
Valence – 1

- In the ground state of a multielectronic atom, electrons occupy the orbitals to minimize the total energy of the atom
- However, due to their negative charges, electrons repel each other and, therefore, only two can share an orbital
  - Electrons having the lowest energy stay in the (spherical) orbital closest to the nucleus, indicated by 1s
  - The second energy level is constituted by four orbitals: a spherical 2s orbital and three "handlebar" 2p orbitals
  - …

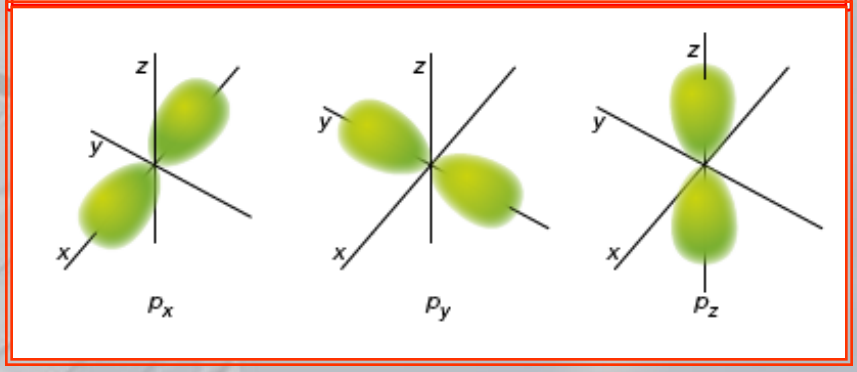


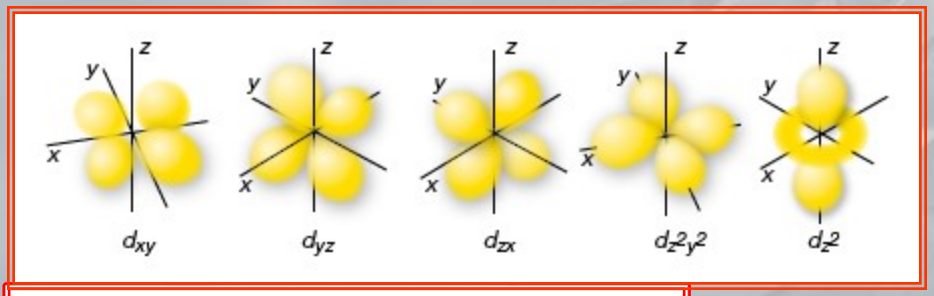
Valence – 2

For each energy level, there is only one s orbital (which can contain 1 or 2 electrons)



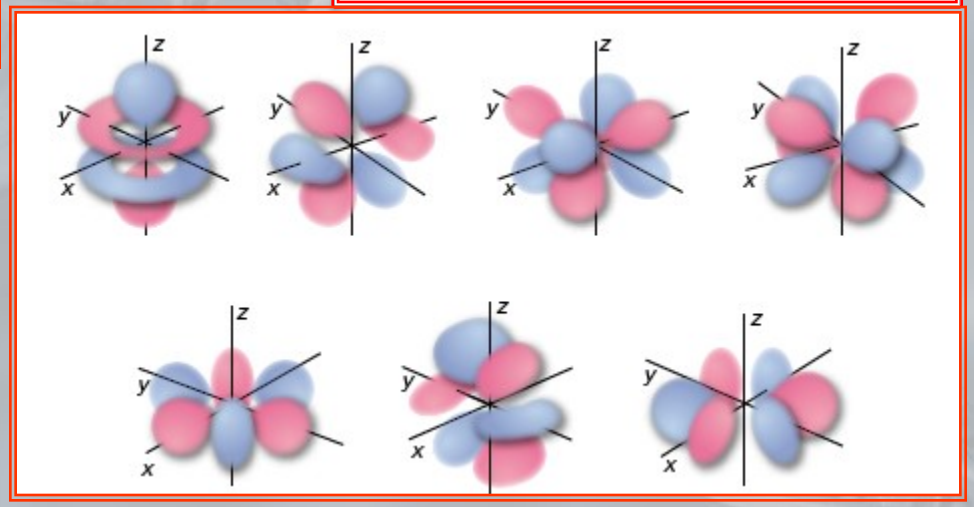
Starting from the 2nd energy level, each level has 3 different p orbitals (i.e., having different spatial orientations), which may contain up to a maximum of 6 electrons (2 for each orbital)





The d orbitals, which are present in the 3rd, 4th, 5th and 6th level, are 5 in total and can contain up to a maximum of 10 electrons

The f orbitals, which are present in the 4th and in the 5th levels, are 7 and can accommodate up to 14 electrons

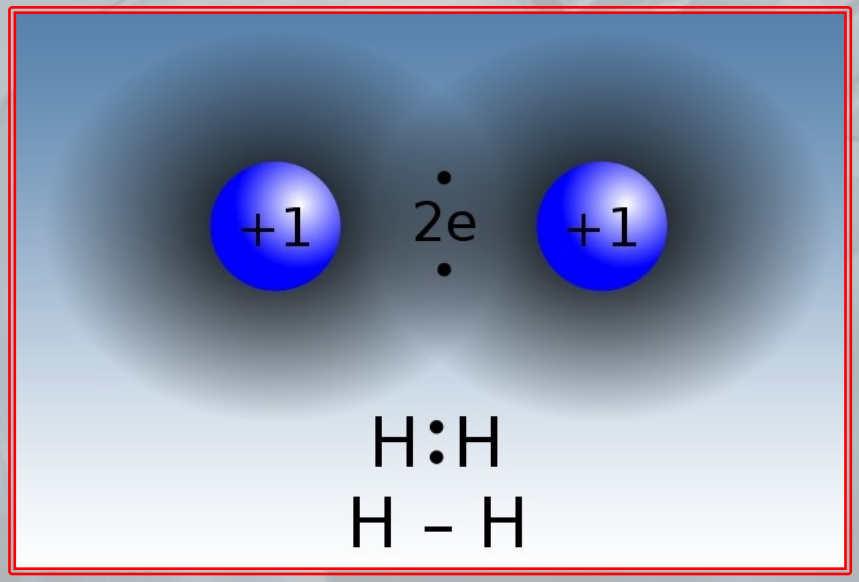


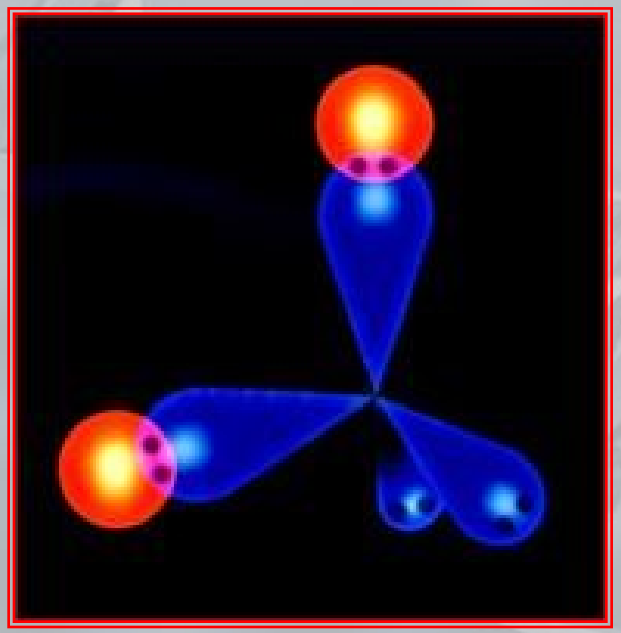
Valence – 3

- The chemical properties of an atom depend on its outermost electron shell
  - Since atoms are mainly composed of empty space, nuclei never meet during normal chemical reactions
  - However, while the number of protons in the nucleus does not change during chemical reactions, the relative position of the electrons, and sometimes even their number, is variable
- Even if the charge balancing principle represents a fundamental rule of Nature, there is also, however, a tendency to maintain the outermost atomic orbitals completely full or completely empty
  - Octet rule: In the formation of bonds, each atom tends, by giving, acquiring or sharing electrons, to achieve the electron configuration of the noble gases, corresponding to the presence of 8 electrons in the more external s and p orbitals

Valence – 4

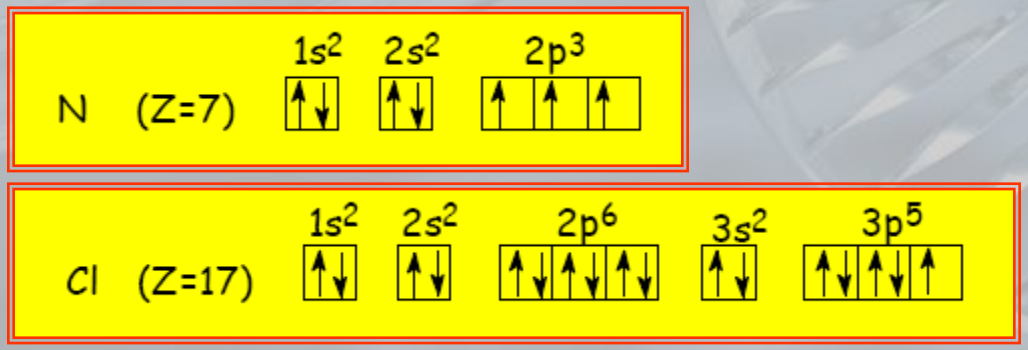
- These two potentially conflicting priorities are guaranteed by enabling the orbitals of an atom to overlap with those of other atoms
- The electrons' sharing, which derives from the orbitals' overlay, is the basis of the covalent bond





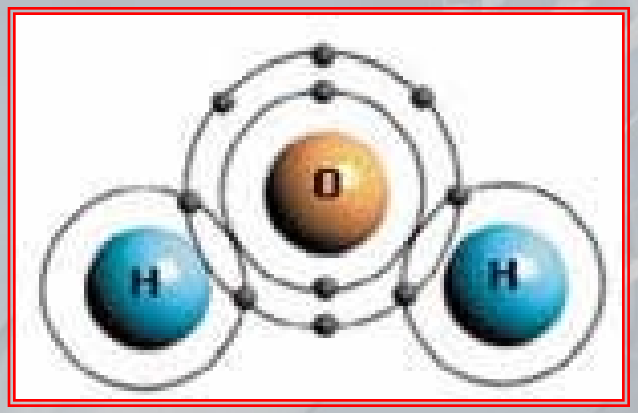
Valence – 5

- The elements, like helium ($_2$He), which do not present unpaired electrons in the more external orbitals, are not chemically reactive and cannot be covalently bonded to other atoms
- Conversely, atoms with a similar valence have similar chemical properties ($_{14}$**Si**, $_6$**C**)
  - Actually, the number of unpaired electrons in the more external s and p orbitals of an atom, its valence, measures its ability to bond ($_1$**H**=1, $_8$O=2, $_7$**N**=3, $_6$**C**=4)

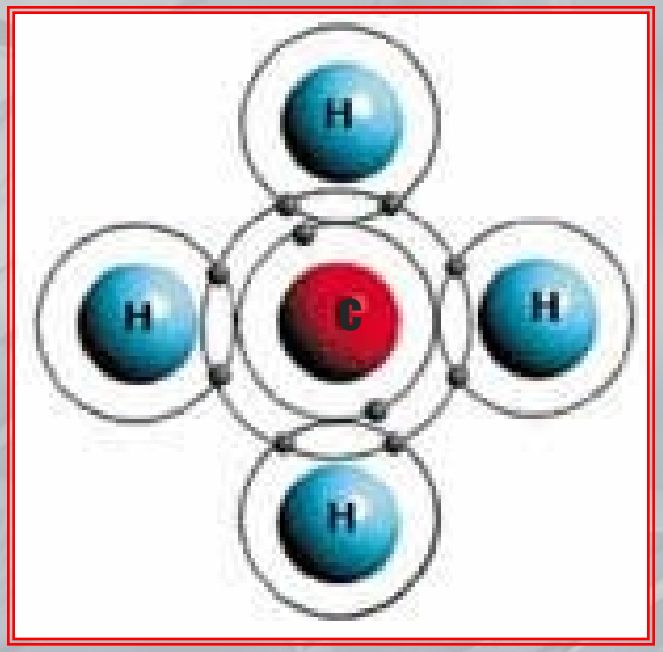


Valence – 6

- The shape and the size of a chemical compound depend on the valence of the atoms with which it is formed



(a)



(b)

Two examples of covalent bonds: Water (a) and Methane (b); in the water molecule, covalent bonds exist between the two hydrogen atoms and the oxygen atom; in the molecule of methane, four hydrogen atoms form covalent bonds with a single carbon atom

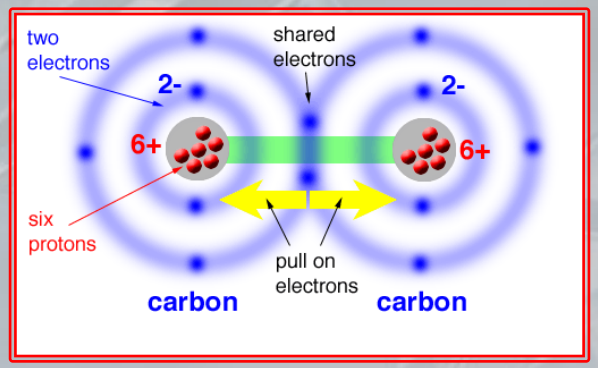
Electronegativity – 1

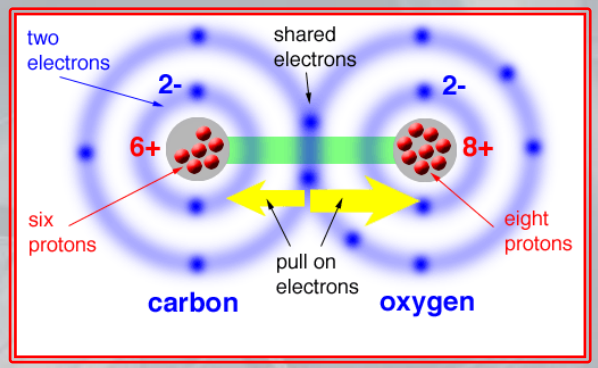
- The electronegativity of an atom is a measure of its "affinity for electrons"
- It is represented by the percentage of electrons that the atom contains in its outer orbitals
- For example, both $_{1}\mathbf{H}$ and $_{6}\mathbf{C}$ have a half full outer shell
  - They have the same electronegativity
  - In covalent bonds between hydrogen and carbon, atoms "participate" in the same way

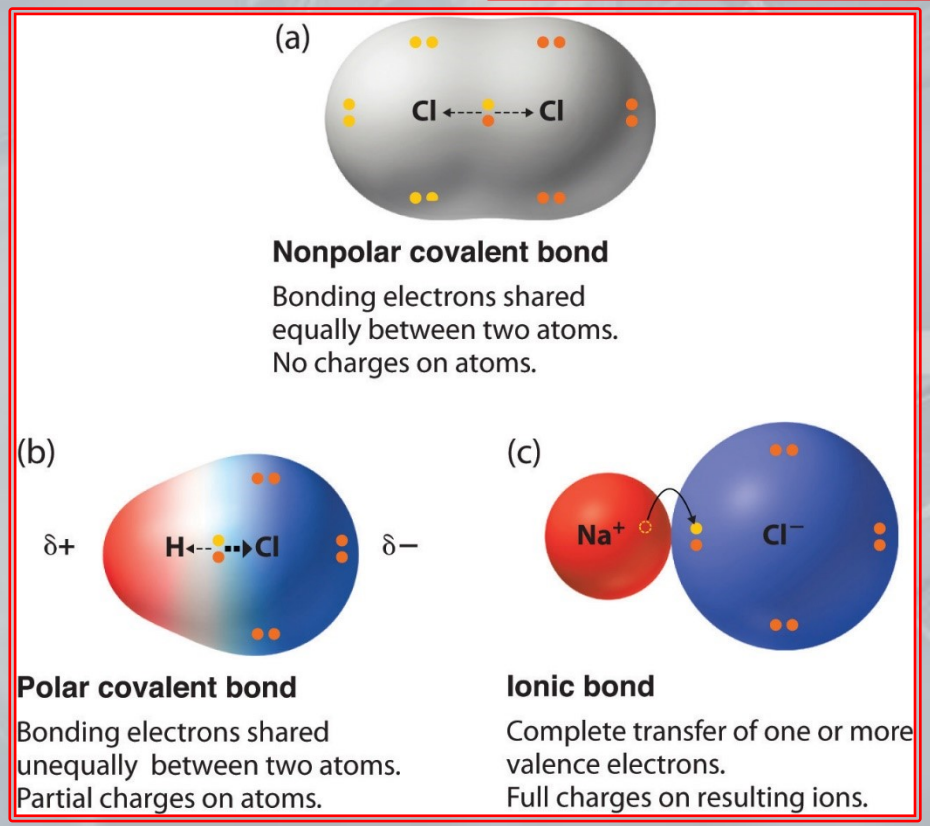
Electronegativity – 2

- The situation is very different for the bonds with oxygen; since $_8$O must gain two or miss six electrons, it is much more electronegative with respect to hydrogen and carbon
- The electrons involved in the covalent bonds of the water, for example, *tend to spend more time in the vicinity of the oxygen atom*
- These polar bonds, thereby, cause a small charge separation, which makes the oxygen of the H_2O molecule slightly negative, while the hydrogens are slightly positive

Electronegativity – 3



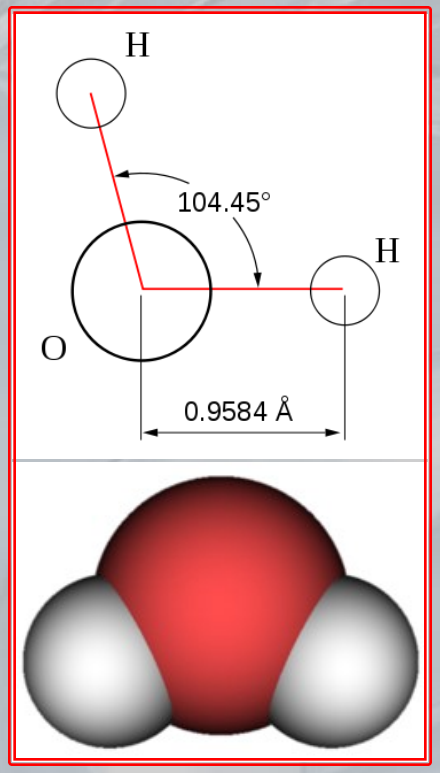




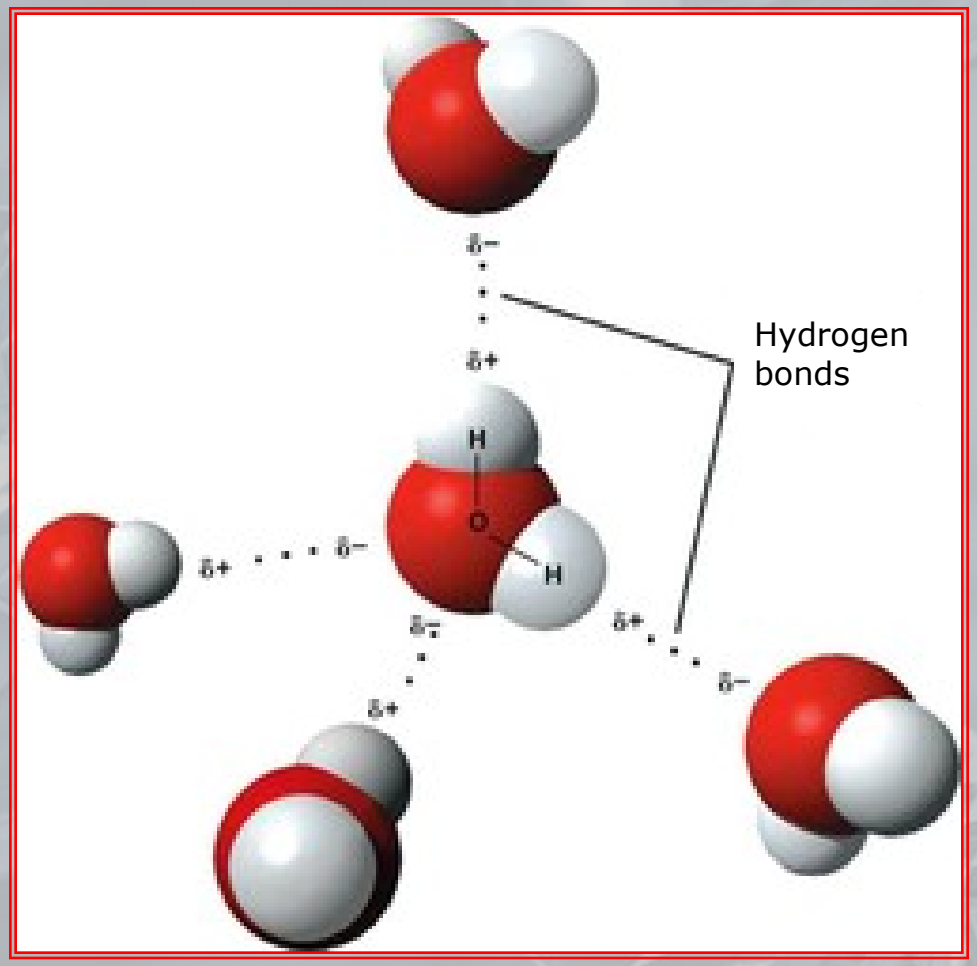
Electronegativity – 4

- The slight separation of charges arising from the polar bonds generates a particular interaction between molecules, called hydrogen bond
  - Each water molecule is weakly associated with a network of other water molecules, since the small positive charges of hydrogens generate affinity with the small negative charges of the neighboring atoms of oxygen
- Much less energy is required to break a hydrogen bond with respect to a covalent bond since, in the first case, there is no electrons' sharing between atoms

Hydrogen bonds



(a)

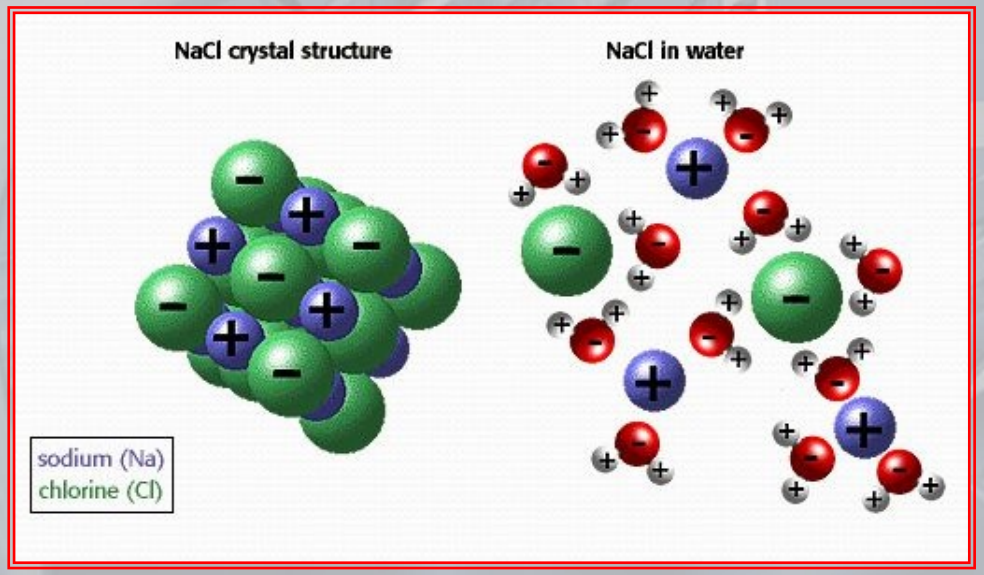


(b)

The water molecule (a) and the interaction between molecules via hydrogen bonds (b)

Hydrophilicity and hydrophobicity – 1

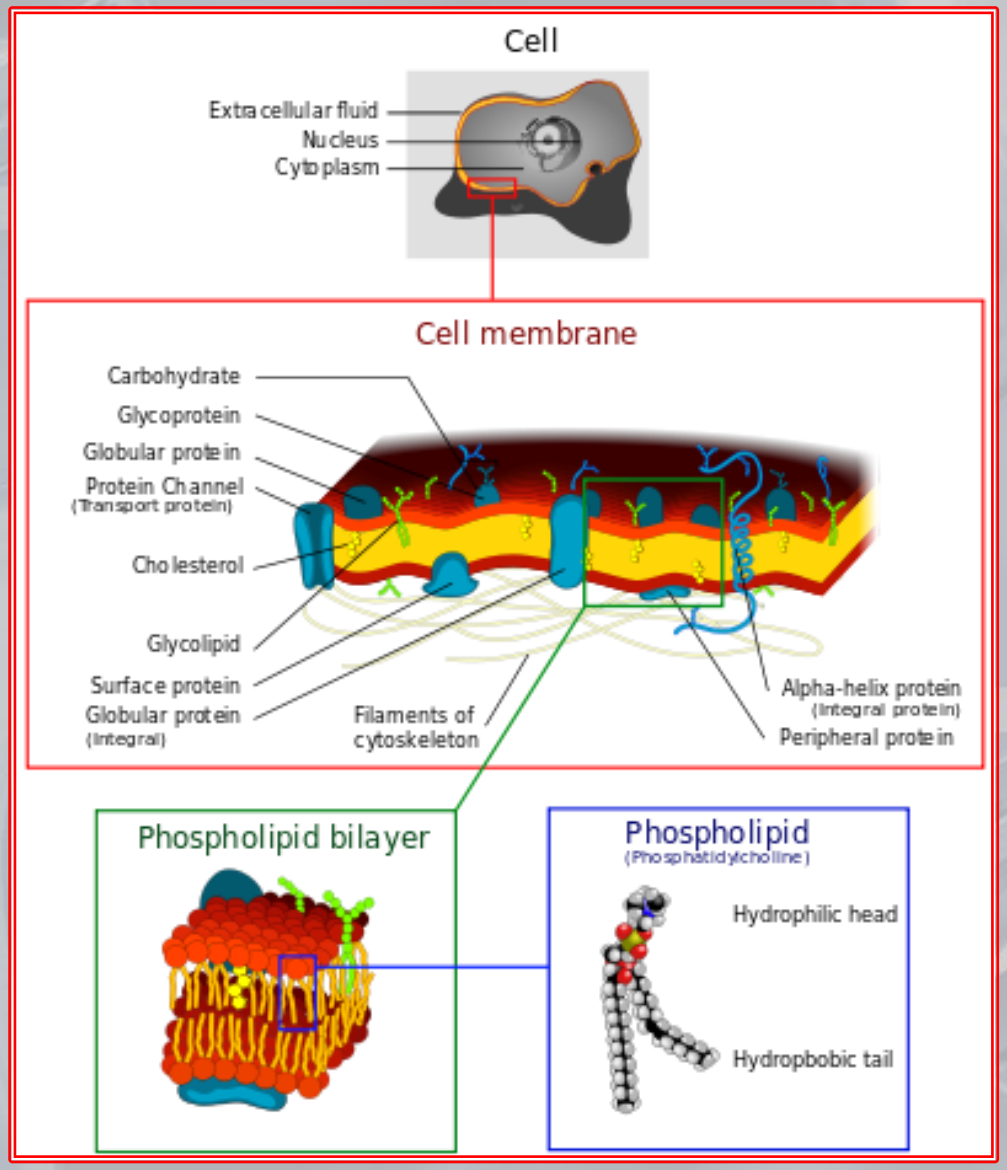
- Chemical compounds can be divided into two categories, those that interact and those that do not interact with water
- Hydrophilic compounds are made up of molecules with polar bonds, capable of forming hydrogen bonds with water
  - They dissolve easily in aqueous solutions



Hydrophilicity and hydrophobicity – 2

- Those chemical compounds that have only non–polar covalent bonded atoms are called hydro-phobic, and do not interact with water molecules
  - Their physical presence causes the water molecules to interact with each other, whereas non–polar molecules form micelles
  - Therefore, molecules such as lipids (with carbon–carbon and carbon–hydrogen bonds) are ex-cluded from aqueous solutions (and forced to associate with each other as in the double lipid layer which forms the cell membrane)

Hydrophilicity and hydrophobicity – 3



Molecular Biology tools – 1

- The interpretation of the information contained in the prokaryotic DNA is easier than the equivalent task on the complex eukaryotic genome
- Moreover, in eukaryotes, the problem of identifying the information that encodes a protein is further complicated by the fact that…
  - …DNA sequences of eukaryotic genes may not be represented by long ORFs, because of the presence of introns
  - …an intron in a particular cell type can be an exon in another

Molecular Biology tools – 2

- However... the problem of deciphering the information content collected in the genome is not insurmountable if we become aware of the "rules" used by the cells for this purpose
- Bioinformatics aims at recognizing/extracting rules from *patterns*, which can be observed in the enormous amount of available data
- Instead, the number of tools commonly used by molecular biologists for:
  - generating raw data needed for such analysis...
  - …checking the biological significance of the extracted rules

  looks surprisingly small

Molecular Biology tools – 3

- Six main laboratory techniques define the entire discipline of Molecular Biology:
  - Restriction enzymes
  - Gel electrophoresis
  - Blotting and hybridization, microarray
  - Cloning
  - Polymerase chain reaction (PCR)
  - DNA sequencing and NGS

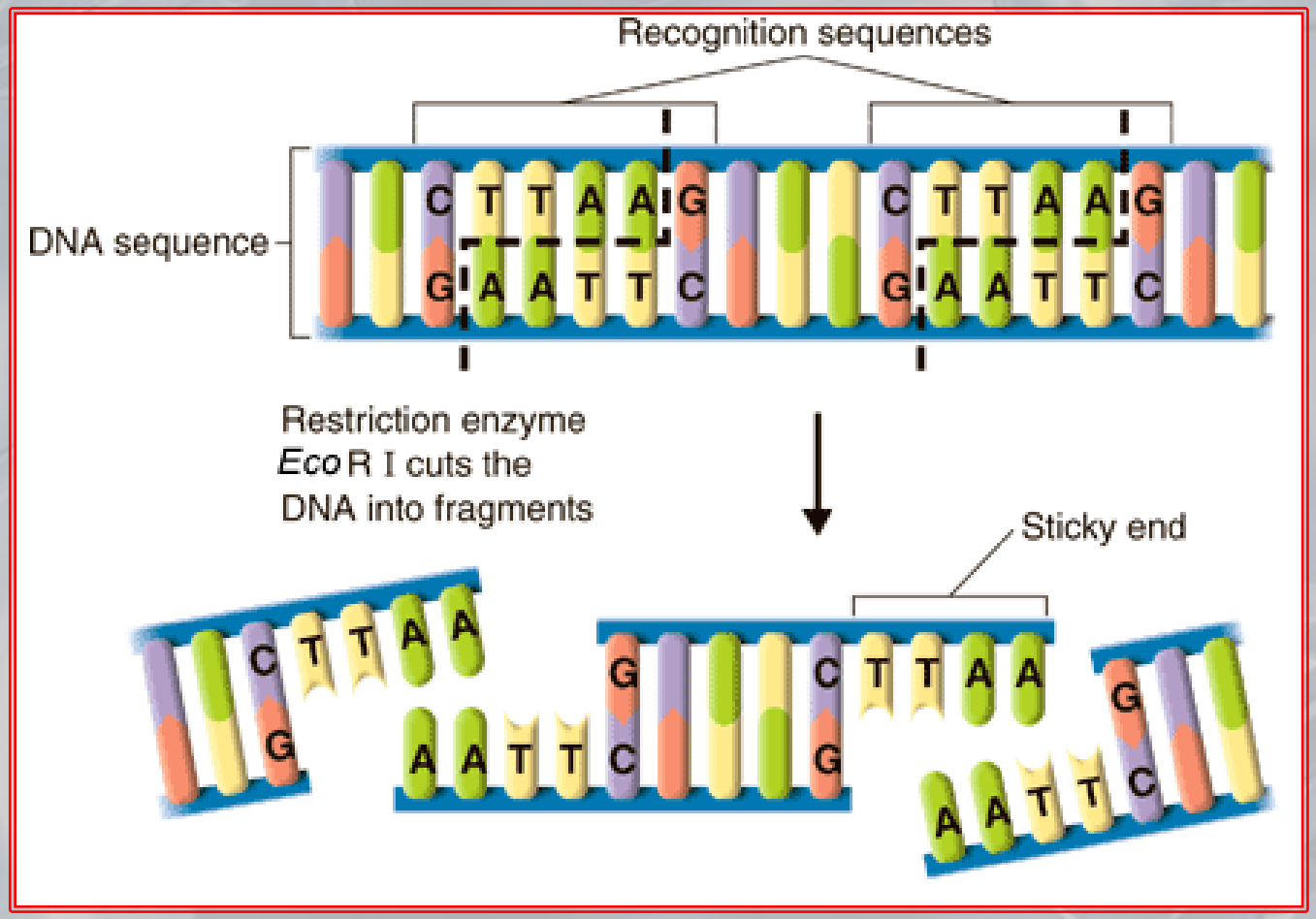
Restriction enzymes – 1

- The original work of Wilkinson, Watson and Crick (1953), for which they were awarded with the Nobel Prize in 1962, explained how DNA can serve as genetic material
- Nevertheless, twenty years passed before that H. Smith *et al.* discovered restriction enzymes, which, since then, have allowed DNA molecules to be manipulated, in order to decipher their information content
- During their study on the causes that led some bacterial cells to improve their defense against viral infections, in fact, they observed how bacteria produce enzymes capable of breaking DNA molecules in correspondence of particular nucleotide strings

Restriction enzymes – 2

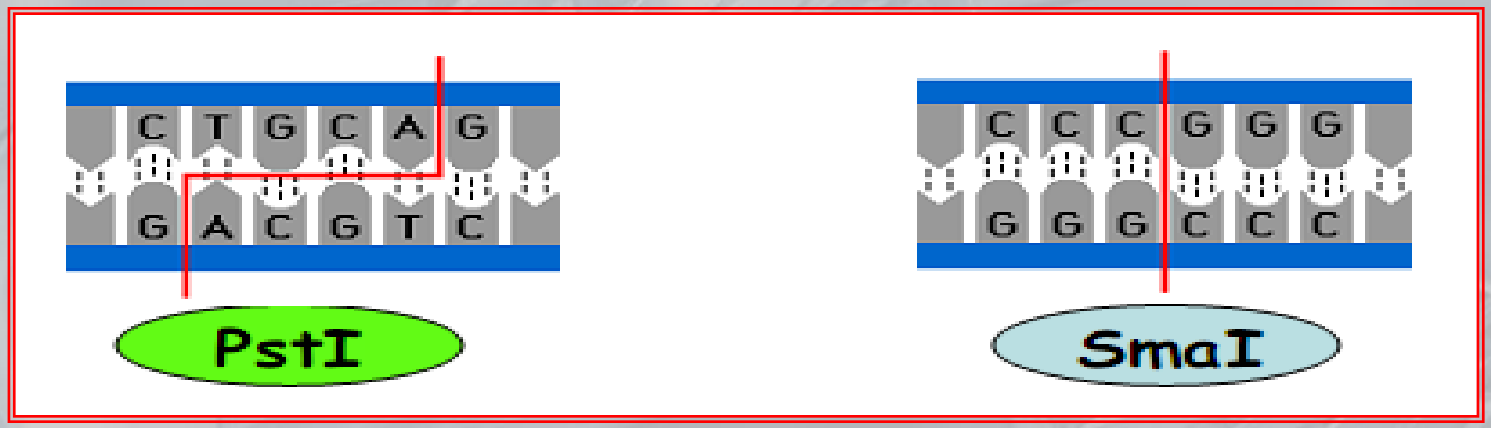
- Restriction enzymes can be isolated from bacterial cells and used as "scissors", which allow biologists to cut and paste the DNA molecules
  - Usually, the double–strand DNA molecules are cleaved in order to leave a short sequence of single–strand DNA at the end of each fragment
  - Such sequence is, by its nature, complementary to any other single–strand sequence produced by the same enzyme
  - The two sequences have, therefore, cohesive or "sticky" terminations, able to hold together the two fragments until another special enzyme, called ligase, rejoins them permanently, reconstructing the phosphodiester bonds cut by the restriction enzyme

Restriction enzymes – 3



Restriction enzymes – 4

- Restriction enzymes can also give rise to flat terminations, which bind to DNA molecules with the same endings

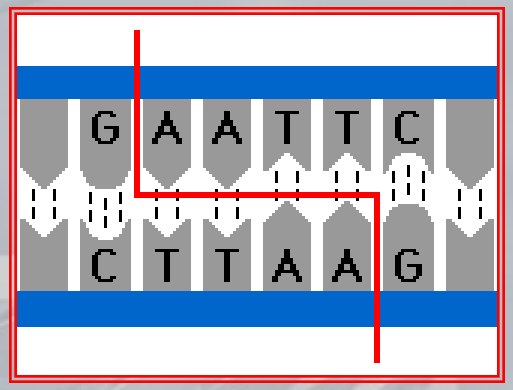


Restriction enzymes – 5

- Example

*Eco*RI (*Escherichia Coli* – bacterium that lives in the lower intestine of warm–blooded animals, Restriction, I)

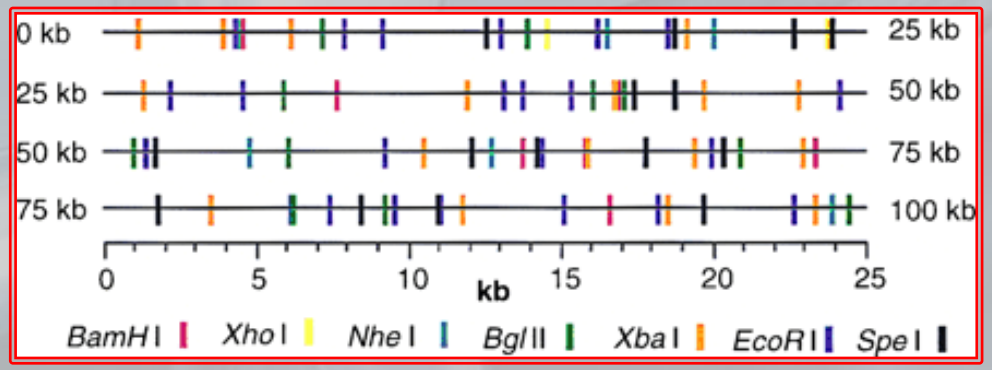
- *Eco*RI cleaves the DNA molecule between the nucleotides **G** and **A** whenever it encounters them in the sequence 5′–**GAATTC**–3′ (which occurs, on average, every $(1/4)^6$=1/4096 base pairs)
- The nucleotide string recognized by *Eco*RI represents its restriction site
- *Eco*RI causes the cohesive termination 5′–**AATT**–3′, which is complementary to 3′–**TTAA**–5′ (5′–**AATT**–3′, by convention)



Restriction enzymes – 6

- Therefore…
  - …cutting a DNA molecule, determining the order of the cuts (obtained with different enzymes) and the number of fragments can help in understanding the specific organization of the molecule (i.e., its sequence)

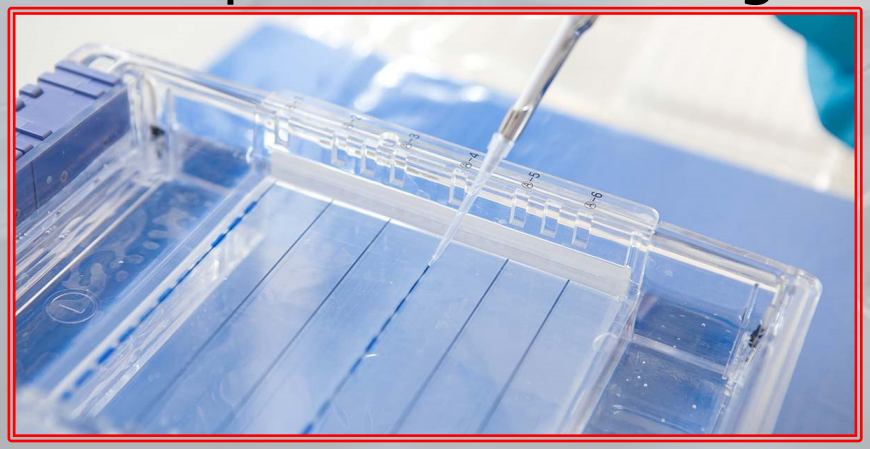
    ⇨ restriction map: a map showing the positions of restriction cut sites within a particular DNA sequence
  - With the restriction enzymes, individual genes can be isolated and experimentally manipulated



Restriction map of BAC 360E4, constructed by using 7 different restriction enzymes; the smallest mapped fragment is 0.5 kb (total insert size: 100 kb)

Gel electrophoresis – 1

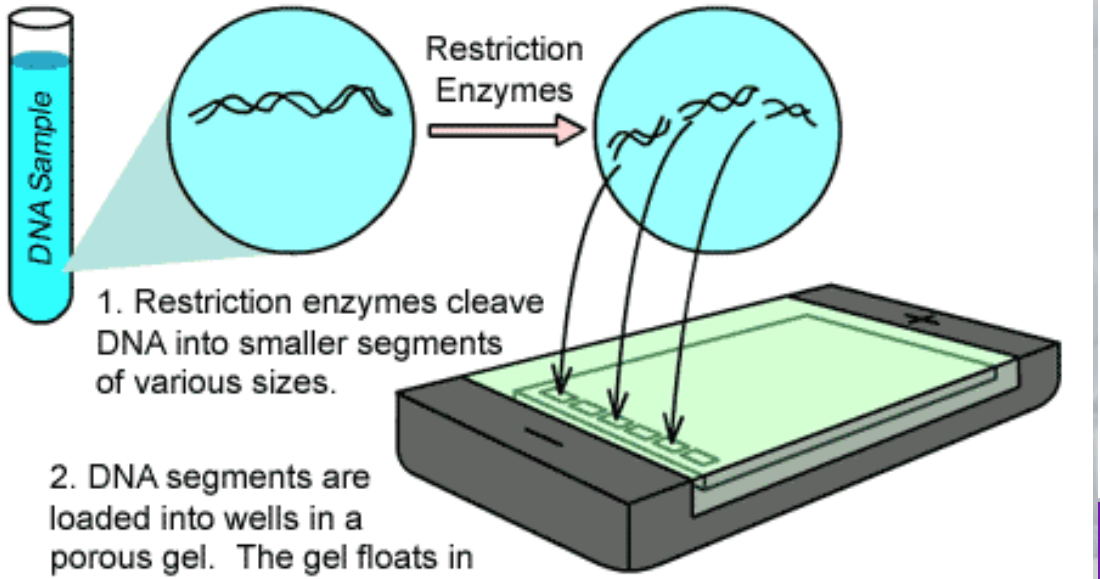
- For genomes composed of millions (as the genome of *E.coli*) or billions of base pairs (such as the human genome), the full DNA splitting by restriction enzymes can provide hundreds of thousands of fragments
- The gel electrophoresis (on *agarose* or on a *poly-acrylamide* support) is a common technique used to analyze and separate these fragments

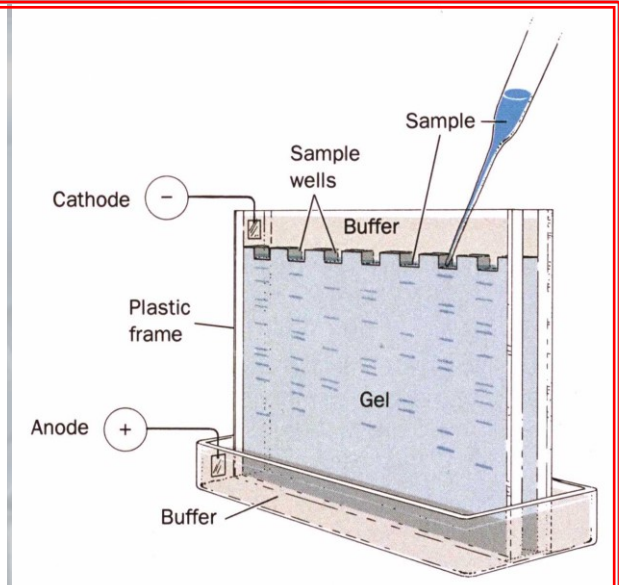


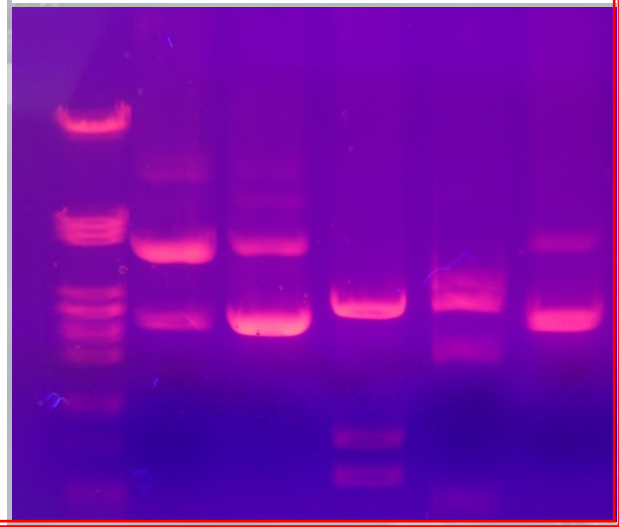
Gel electrophoresis – 2

- The gel electrophoresis exploits the charges present in the molecules of DNA or RNA (the main chain is negatively charged, due to the presence of the phosphate group) to make them migrate through the gel, using an electric field
- The gel acts as a sieve, being constituted by a network of pores, which allows to separate the molecules according to their size: smaller molecules will traverse the gel faster than the greater ones
  - A separation will occur as a function of the speed

Gel electrophoresis – 3

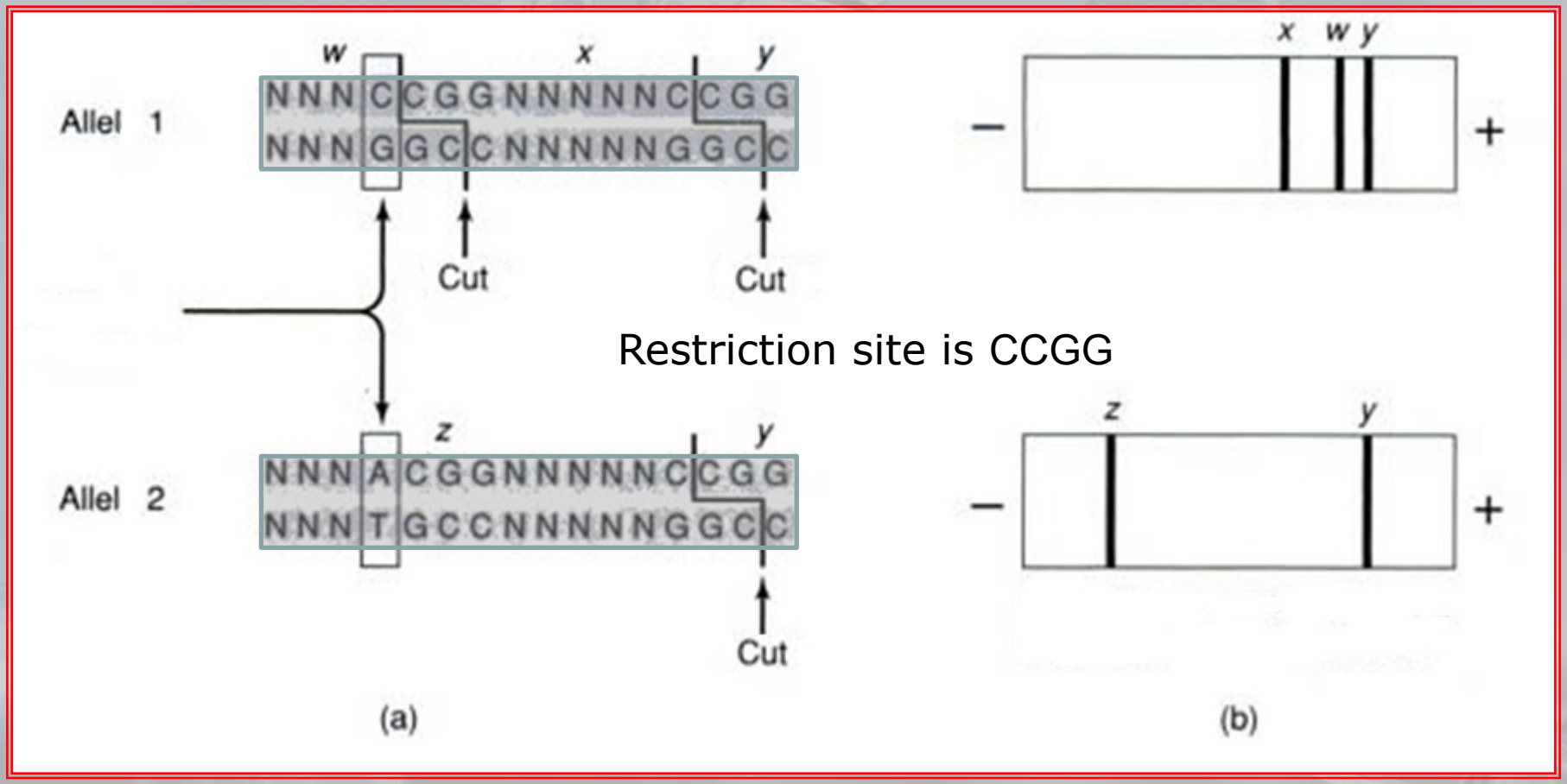






Gel electrophoresis – 4

The gel electrophoresis allows the separation of DNA fragments according to their length; (a) The presence of alterations in the nucleotide sequence causes the restriction enzymes not to recognize a cutting site; (b) Differences in the restriction fragment lengths

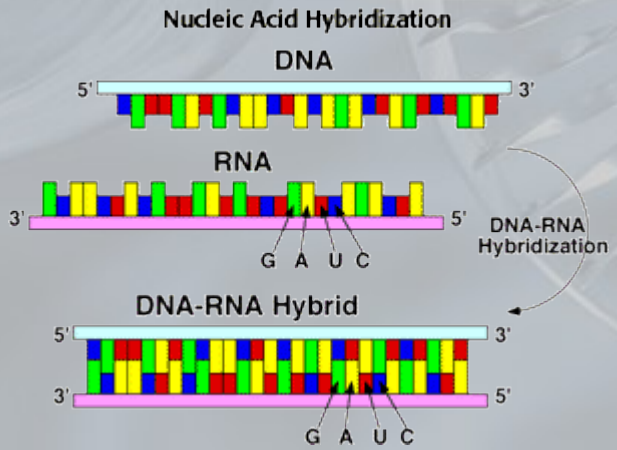


Blotting and hybridization – 1

- Locating a single DNA fragment, that contains a specific gene, among hundreds of thousands of fragments, although divided by size, is an impossible task
- Hybridization is a technique that allows to verify the complementarity between two single–strand nucleic acid molecules according to their sequence homology
- This is possible because those proteins which have a key biological function, such as hemoglobin, insulin, or the growth hormone, have very similar sequences among phylogenetically close organisms

Blotting and hybridization – 2

- The base coupling can occur between two DNAs, two RNAs or a DNA and an RNA molecule
- The reaction principle consists in...
  - ...exposing two preparations of single–strand nucleic acid one to another (one of which is radioactively labeled)
  - and, subsequently, displaying (by autoradiography or staining) and measuring the amount of double–strand material (f.i., by quantifying its incorporated radioactivity)



Blotting and hybridization – 3

- The technique of hybridization on a membrane, or a filter, is the most used to verify the presence of particular sequences (and thus of particular genes) of different origin, coming from diverse species
- During the blotting phase, polynucleotides are transferred from the gel to a solid support (a nitrocellulose or a nylon membrane), using an alkaline transfer method (sodium hydroxide), which produces also DNA denaturation; because of the capillarity phenomenon, the DNA molecules remain attached to the membrane in the position previously occupied in the gel, while the molecules of water are filtered out; a subsequent cooking process, finally fixes the DNA fragments to the membrane

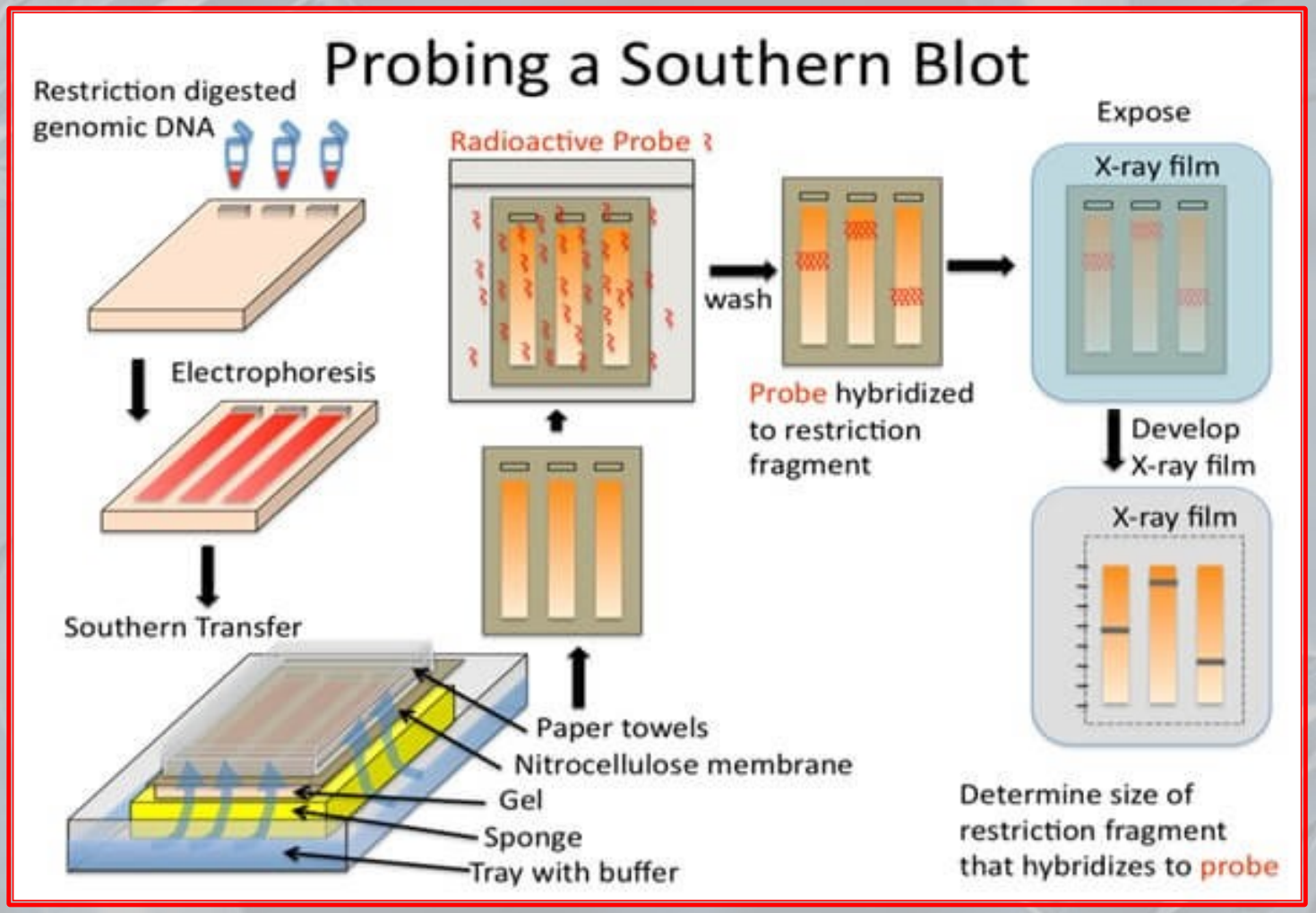
Blotting and hybridization – 4

- For hybridization, molecular probes are employed, i.e. DNA fragments coding for the gene to be analyzed
- The probe is labeled with a radioactive isotope and mixed in solution (with salts and detergents), at a temperature close to 100°, with the membrane on which DNA fragments, that should contain the complementary sequence, are fixed
- The hybridization conditions, i.e. both the salt concentration and the temperature, can be adjusted to allow the coupling of not perfectly complementary sequences

Blotting and hybridization – 5

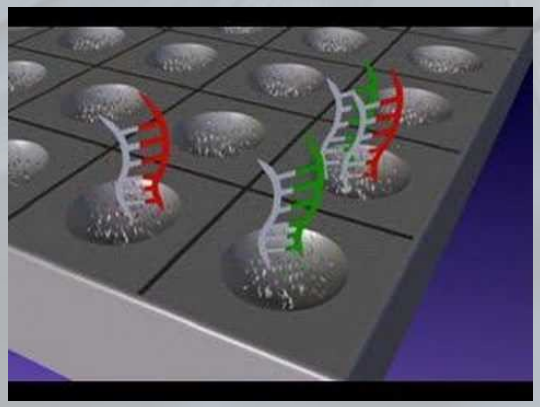
- At the end of the hybridization process, the unbound probe is washed away, whereas the membrane is examined in order to see where the coupling between the probe and the target sequence took place
- The exposure of the filter to a photosensitive plate (or coloring) will allow the detection of a track that will confirm the coupling; on the other hand, if homology does not exist, hybridization will not happen

Blotting and hybridization – 6



Microarrays – 1

- A different hybridization technique is represented by microarrays, or DNA chips, or biochips
- In this case, tens of thousands of nucleotide sequences are fixed one by one in a specific position on the surface of a small silicon chip
- Microarrays exploit a reverse hybridization technique, which consists in fixing all the DNA segments, called *probes*, on a support, marking instead the nucleic acid, called *target*, that should be identified

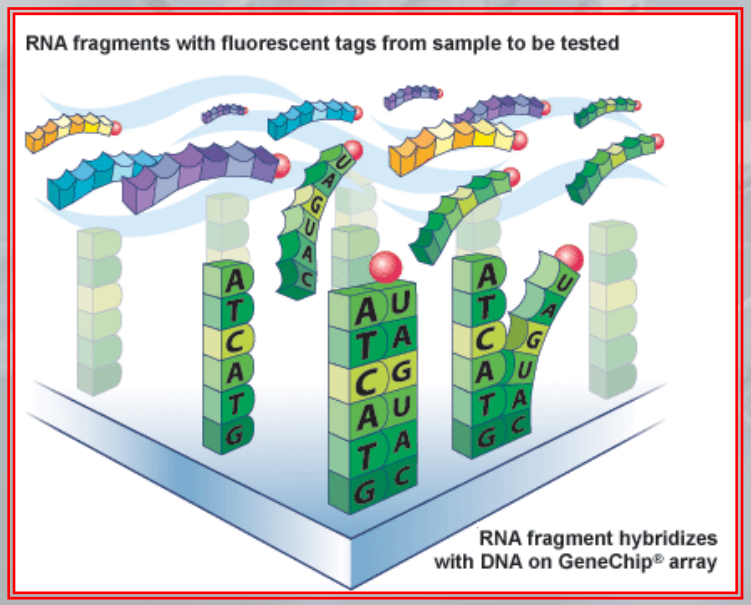


Microarrays – 2

- In other words, microarrays consist of a solid support on which a large number of specific DNA probes (20–25 bps) are neatly arranged to form a regular dot matrix
- Each dot in the matrix is typically less than 200 microns in size and is made up of many copies of the same DNA sequence; it represents the minimum unit of the microarray and is called a feature
- Microarrays are classified by the number of features present on their surface, a sort of measure of their complexity and resolution capacity
- Each feature is made up of several identical copies of probe sequences, which will hybridize with the complementary marked target sequences contained in the samples under examination

Microarrays – 3

- The microarray technique was developed in the late '90s; today, it enables the analysis of gene expression, monitoring at once the RNA (or cDNA) products from thousands of genes
- Significant computational efforts are needed both to build the chip and to interpret the obtained results



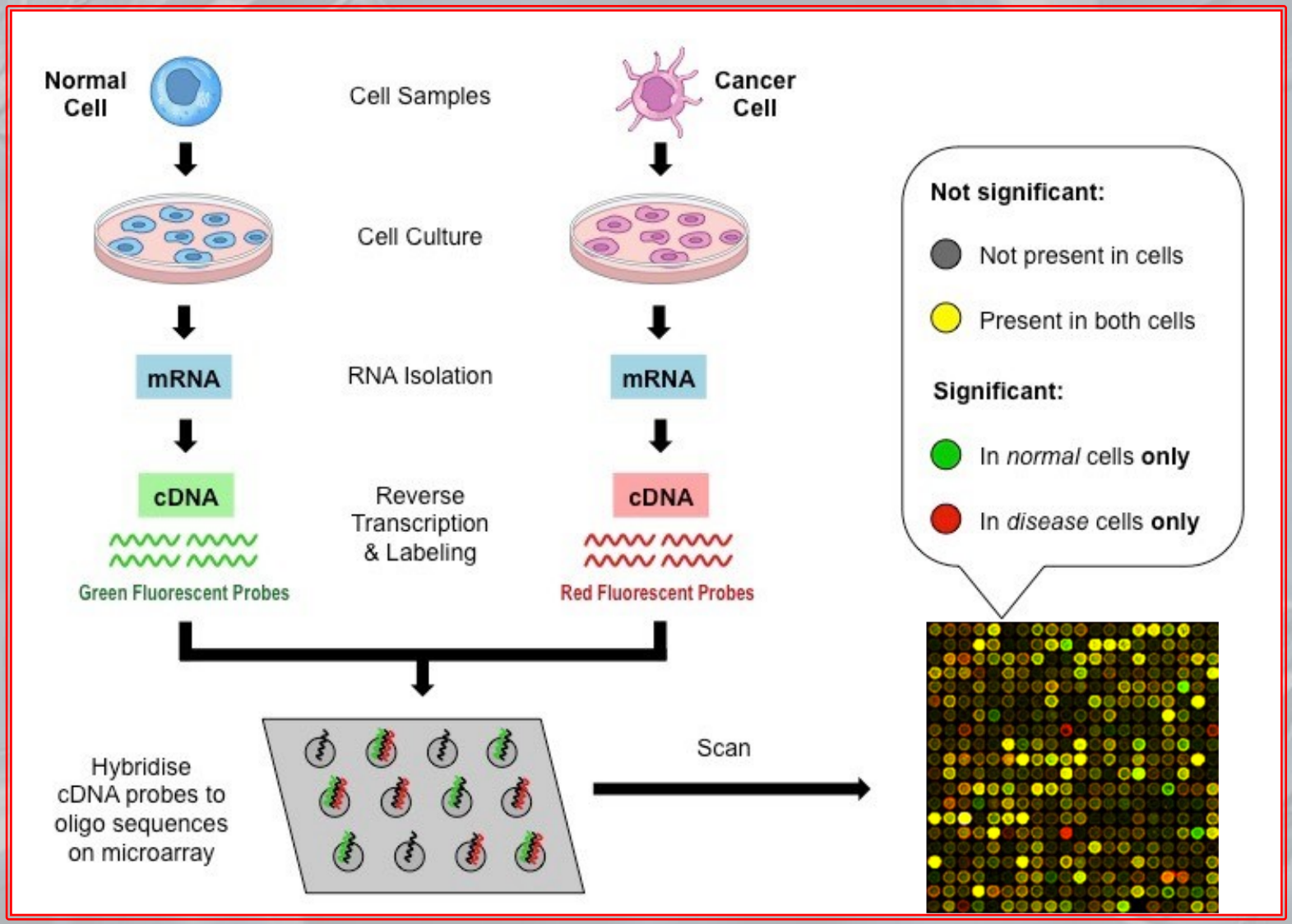
Microarrays – 4

- To perform a microarray analysis, mRNA molecules are typically collected from both an experimental sample (f.i., an individual with a disease like cancer) and a reference sample (a healthy individual)
- The two mRNA samples are then converted into complementary DNA (cDNA) — using the reverse transcriptase —, and each sample is labeled with a fluorescent dye of a different color
- The experimental cDNA sample may be labeled with a red fluorescent dye, whereas the reference cDNA may be labeled with a green fluorescent dye
- The two samples are then mixed together and allowed to bind to the microarray slide

Microarrays – 5

- After hybridization, the microarray is scanned to measure the expression of each gene printed on the slide
  - If the expression of a particular gene is higher in the experimental sample than in the reference sample, then the corresponding spot on the microarray appears red
  - In contrast, if the expression in the experimental sample is lower than in the reference sample, then the spot appears green
  - Finally, if there is equal expression in the two samples, then the spot appears yellow
- The data gathered through microarrays can be used to create gene expression profiles, which show simultaneous changes in the expression of many genes in response to a particular condition or medical treatment

Microarrays – 6



Cloning – 1

- While it is normal for the cells to extract information from a single DNA molecule, a great amount of genetic material must be analyzed using Molecular Biology tools (many millions of molecules, a quantity almost visible with the naked eye)
  - Cell aid must be invoked for the generation of a sufficient amount of specific DNA molecules
- In the genome of a diploid organism (an organism having two sets of chromosomes: usually, one from the mother and one from the father), for example in a human cell, a gene is typically duplicated, "diluted" in the midst of many other DNA sequences

Cloning – 2

- Through molecular cloning, it is possible to isolate a single gene or, more generally, a DNA fragment from a particular genome, to produce many identical copies
- The availability of a gene in a pure form and in large quantities allows its analysis at the molecular level
- DNA extraction or purification aims to separate DNA from cellular components (proteins, lipids, polysaccharides) and from interfering substances

Cloning – 3

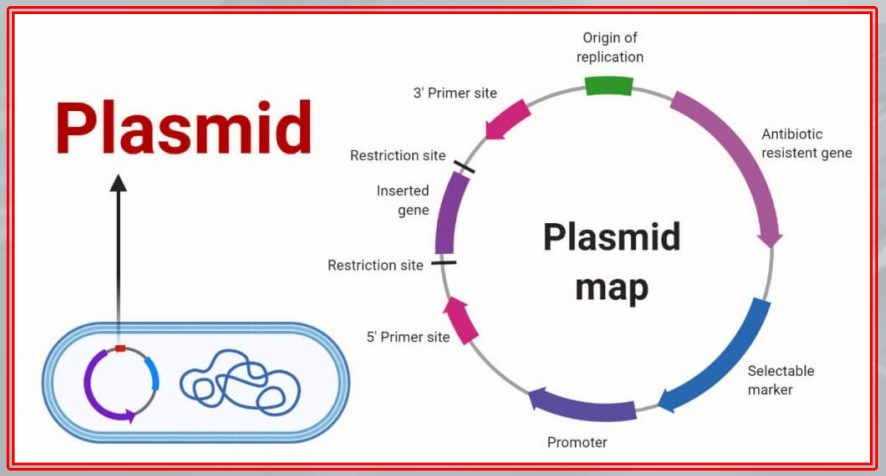
- The cloning procedure involves the insertion of specific DNA fragments within conveyors, called vectors, similar to the chromosomes, which allow the replication and the isolation of genes within living cells
- The first conveyors that were employed were derived from bacterial viruses, or from small extra–chromosomal DNA fragments, called *plasmids*, which are present in prokaryotic cells

Cloning – 4

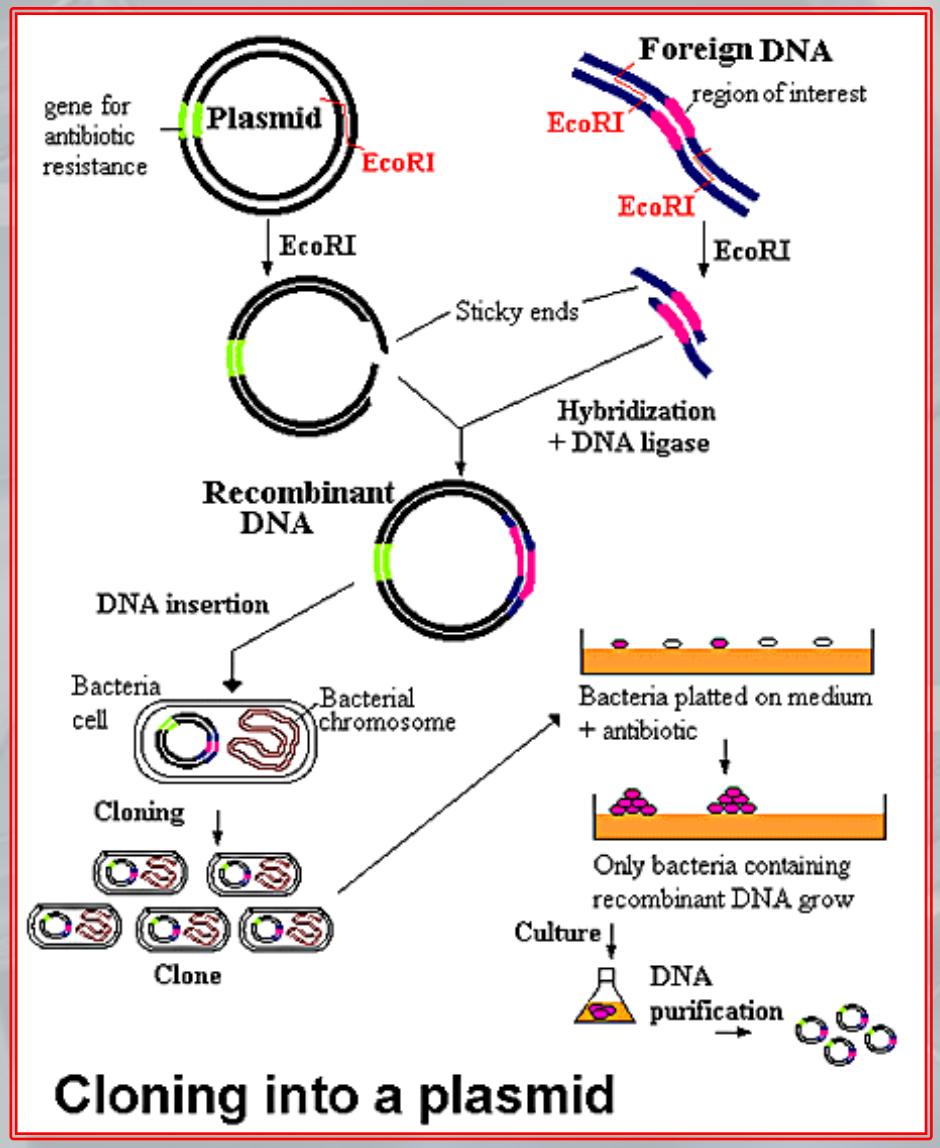
- After obtaining a restriction fragment to clone, its sticky ends can be used to insert and bind it into an *ad hoc* conveyor, previously cut with the same restriction enzyme
- Standard conveyors can be easily processed in a wet lab and can be used to clone DNA segments no longer than 25000 bps
- Recent alternative conveyors, derived from bacterial or yeast chromosomes, can be used to clone longer sequences (100,000 to 1 million bps)

Cloning – 5

- All the vectors must share the following features:
  - They have a zone in which the exogenous DNA can be inserted (called *polylinker*)
  - They contain appropriate sequences that allow their replication within a living cell
  - They contain appropriate sequences that confer to the host cell the ability to detect their presence
  - They must have a size and a shape such as they can be separated from the host cell DNA



Cloning – 6


gene for antibiotic resistance
Plasmid
EcoRI
EcoRI
Foreign DNA
region of interest
EcoRI
EcoRI
EcoRI
Sticky ends
Hybridization + DNA ligase
Recombinant DNA
DNA insertion
Bacteria cell
Bacterial chromosome
Cloning
Clone
Bacteria platted on medium + antibiotic
Only bacteria containing recombinant DNA grow
Culture
DNA purification
Cloning into a plasmid

Cloning – 7

- All the produced identical copies of the fragments, the molecular clones, can be immediately analysed or stored in a genomic library
- A genomic library should ideally contain (at least) one copy for each DNA segment of an organism
  - The number of clones (given by the genome size divided by the average length of the fragments) defines a genomic equivalent
  - Example: *E.coli* has a genome composed by ~4,600,000 base pairs; if it were completely digested by a restriction enzyme such as *EcoRI*, for the construction of a complete genomic library, each DNA fragment should be cloned (i.e. a total of more than 1000 fragments with an average length of 4096 base pairs)

Cloning – 8

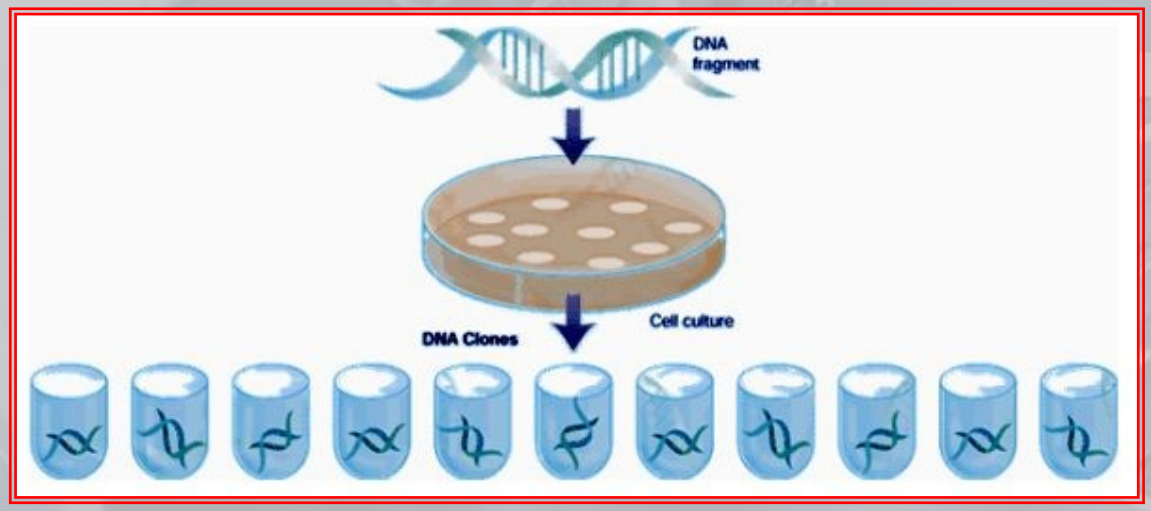
- Unfortunately, the library cannot be built with the simple digestion of the genomic DNA of a single cell, and then with the construction of clones for each fragment
- The process of cloning is not efficient and it is usually necessary to collect DNA from thousands of cells to clone even a single fragment
- In addition, the random nature of the cloning process implies that some fragments are cloned multiple times, while others are not represented in the genomic equivalent

Cloning – 9

- Increasing the number of clones in a genomic library means increasing the likelihood that it will contain at least one copy of each DNA segment
- A genomic library containing four or five genomic equivalents has, on average,
  - four or five copies of each DNA fragment
  - 95% chance of containing at least one copy of each portion of the genome

Cloning – 10

- Practical implications:
  - Vectors allowing the cloning of long DNA fragments are most suitable for the construction of genomic libraries, because they require few clones to achieve the genomic equivalent
  - Cloning the last 5% of a genome is often as difficult as cloning the first 95%



Cloning – 11

- Otherwise: construction of a cDNA library
  - The most interesting portion of the genome is that corresponding to the regions that encode proteins, which are then transcribed into mRNA
  - mRNAs can be separated from the other poly-nucleotides of the cell and then converted, by an enzyme called reverse transcriptase, into comple-mentary DNA (cDNA); after, they can be cloned as a part of a library
  - However, the genomic contents related to, for instance, introns, promoters, regulatory sequences, etc., are irretrievably lost

Polymerase chain reaction – 1

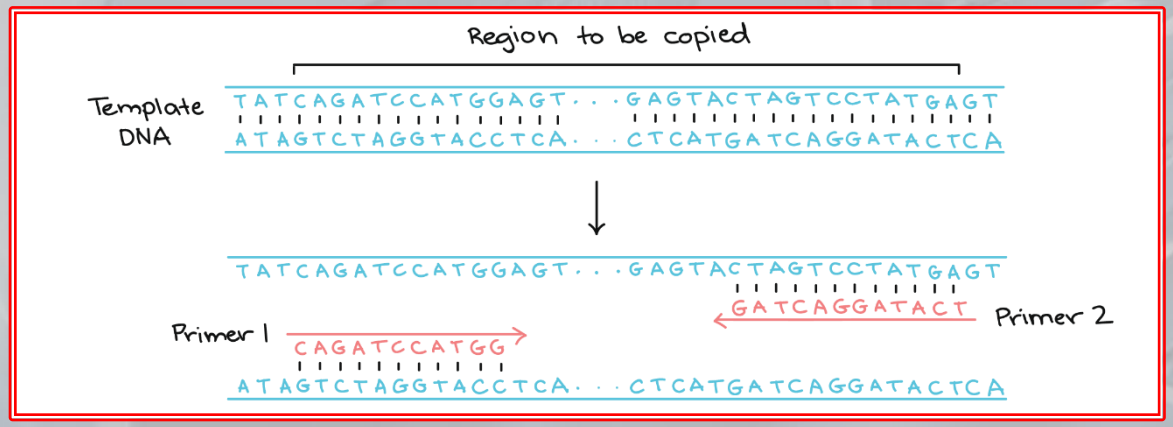
- The Polymerase Chain Reaction (PCR, K. Mullis, 1985, Nobel prize for Chemistry in 1993) has revolutionized the genetic engineering methods, providing a very powerful tool for amplifying, in vitro and very rapidly, specific DNA sequences
- This technique allows to obtain hundreds of thousands of DNA copies for subsequent characterization (length evaluation, sequencing, etc.), without having to resort to the common methods of cloning

Polymerase chain reaction – 2

- PCR exploits the in vitro DNA synthesis, a reaction catalyzed by DNA polymerase
- This enzyme requires a "mold" or a *template*, represented by a DNA strand, to which a *primer* must be paired, that acts as a trigger for the DNA replication
- Primers are necessary because many DNA polymerases cannot initiate the synthesis of a new strand *de novo*, but, instead, can only add nucleotides to a preexisting filament

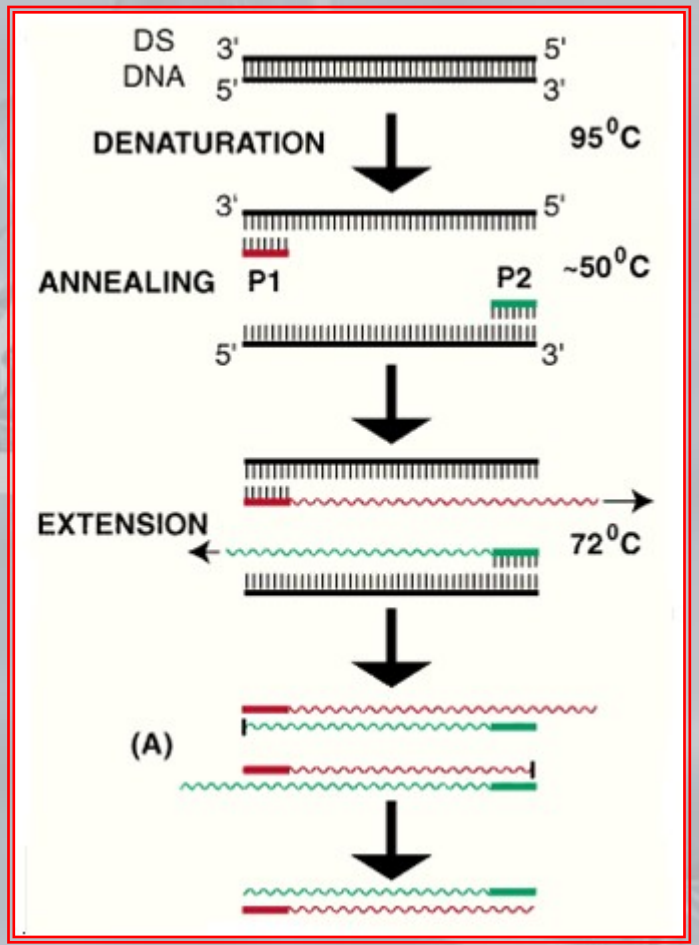
Polymerase chain reaction – 3

- PCR is based on the use of two primers, 18–20 nucleotides long, designed to be exactly complementary to the corresponding sequences flanking the DNA segment to be amplified
- The two primers are directed in opposite directions, but they are convergent, and define the ends of the future amplification product
- The DNA polymerase activity results in the synthesis of a new strand from each primer



Polymerase chain reaction – 4

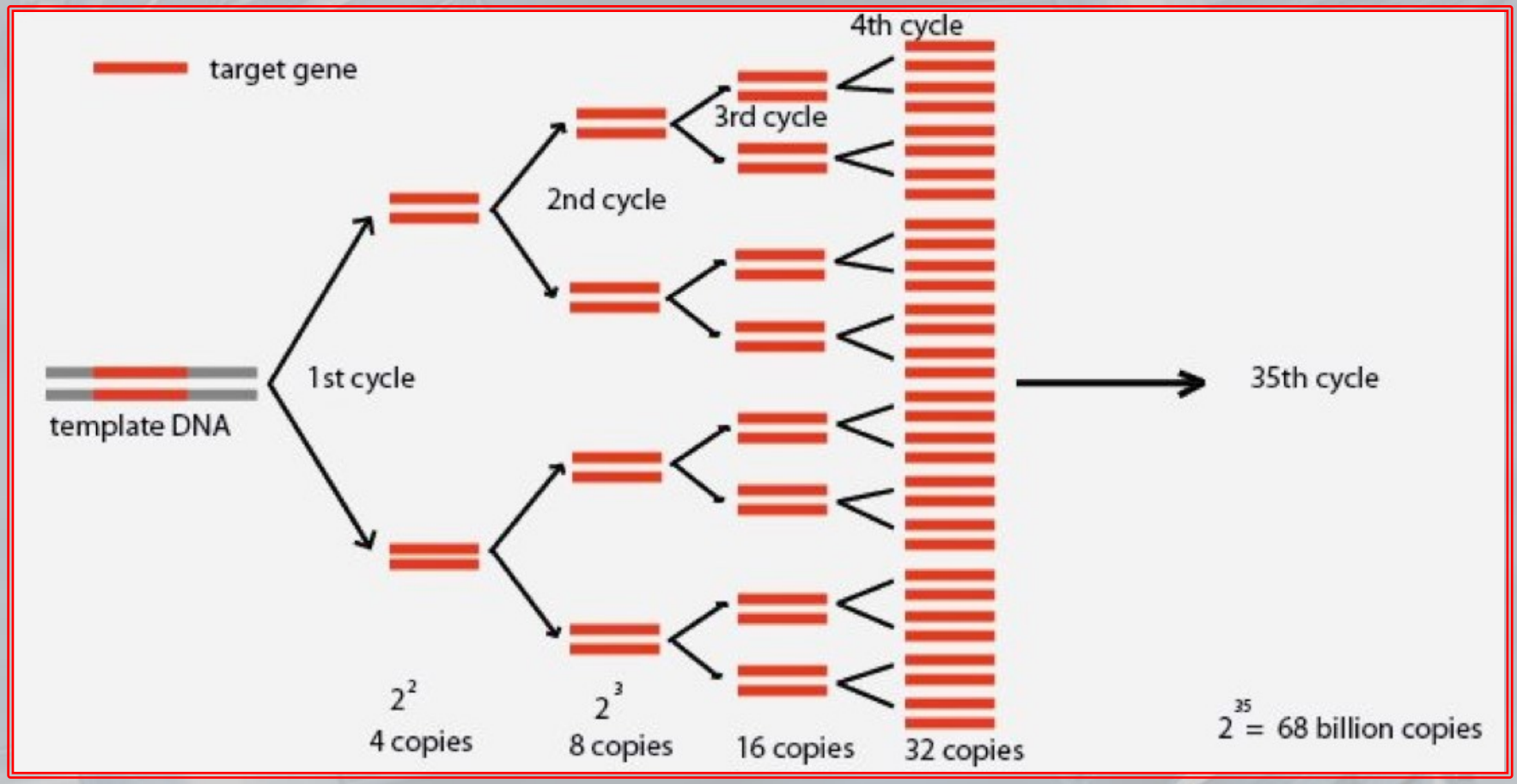
- The reaction is divided into three stages, each carried out at a different temperature
  - denaturation (95°C), to separate the two strands of the template molecule
  - annealing (50–60°C), during which the primers anneal to the denatured strands, determining the point of initiation of the DNA synthesis
  - polymerization (extension, 72°C), to produce two new double–strand DNA molecules (thanks to DNA polymerase)



Polymerase chain reaction – 5

- The cycle of denaturation–annealing–extension is repeated 20–30 times, so as to obtain a large amplification of the DNA included in the annealing region of the two primers
- Since many DNA polymerases, including the human one, cannot resist the temperatures necessary for denaturation, polymerases belonging to thermophilic organisms, which are not inactivated by high temperatures, are used — for example the Taq polymerase coming from the thermophilic bacterium *Thermus aquaticus*
- The first PCR products of "good quality" are formed starting from the third cycle and accumulate with an exponential trend ($N \times 2^{n-2}$, where N is the initial number of molecules, and n represents the number of amplification cycles)

Polymerase chain reaction – 6



Polymerase chain reaction – 7

- The amplified DNA molecules are produced much more quickly (each cycle lasts a few minutes) and effectively than those obtainable from clones
- The major advantage of PCR, however, is that the process can be started with an amount of material (f.i., in the case of specimens from museums, or fossils or forensic) much lower than that usually available (and necessary) in cloning experiments

DNA sequencing – 1

- Sequencing refers to the process of determining the exact primary structure of a biopolymer (a macromolecule), by establishing the order of nucleotides, for a nucleic acid, or amino acids, in the case of proteins
- The molecular characterization of a DNA fragment consists, in fact, in the determination of the order, or sequence, of its nucleotides

DNA sequencing – 2

- All the strategies of DNA sequencing include three steps
  1) Generation of a complete set of subfragments of the region of interest, which differ from one another for a single nucleotide
  2) Fragment marking based on one out of four different labels (*tags*), which depends on the nucleotide located at the end of the fragment
  3) Fragment separation based on size (by gel electrophoresis), to read the sequence after the ordered recognition of all the tags

DNA sequencing – 3

- Method due to A. M. Maxam and W. Gilbert (1977): chemical degradation of DNA
- Method due to F. Sanger (1978): based on chain termination
  - They are both based on the generation of a sequence of single–strand DNA fragments, each of which differ from (is longer than) the previous fragment by only one basis

DNA sequencing – 4

- Maxam and Gilbert method
  - Based on the chemical degradation of DNA fragments, it is now fallen into disuse because it employs chemical compounds harmful to human health and because it is not suited for automated sequencing
  - It differs from the Sanger method since it does not use enzymatic synthesis, but chemical processes that act at the individual nucleotide level
  - It is based on the ability of certain chemical compounds – hydrazine, N_2H_4, formic acid, $HCOOH$, dimethyl sulfate, $(CH_3)_2SO_4$ – of changing the DNA bases in a very specific way, and on the capacity of others, like piperidine, $(CH_2)_5NH$, of catalyzing the rupture of the DNA strand at the modified nucleotides

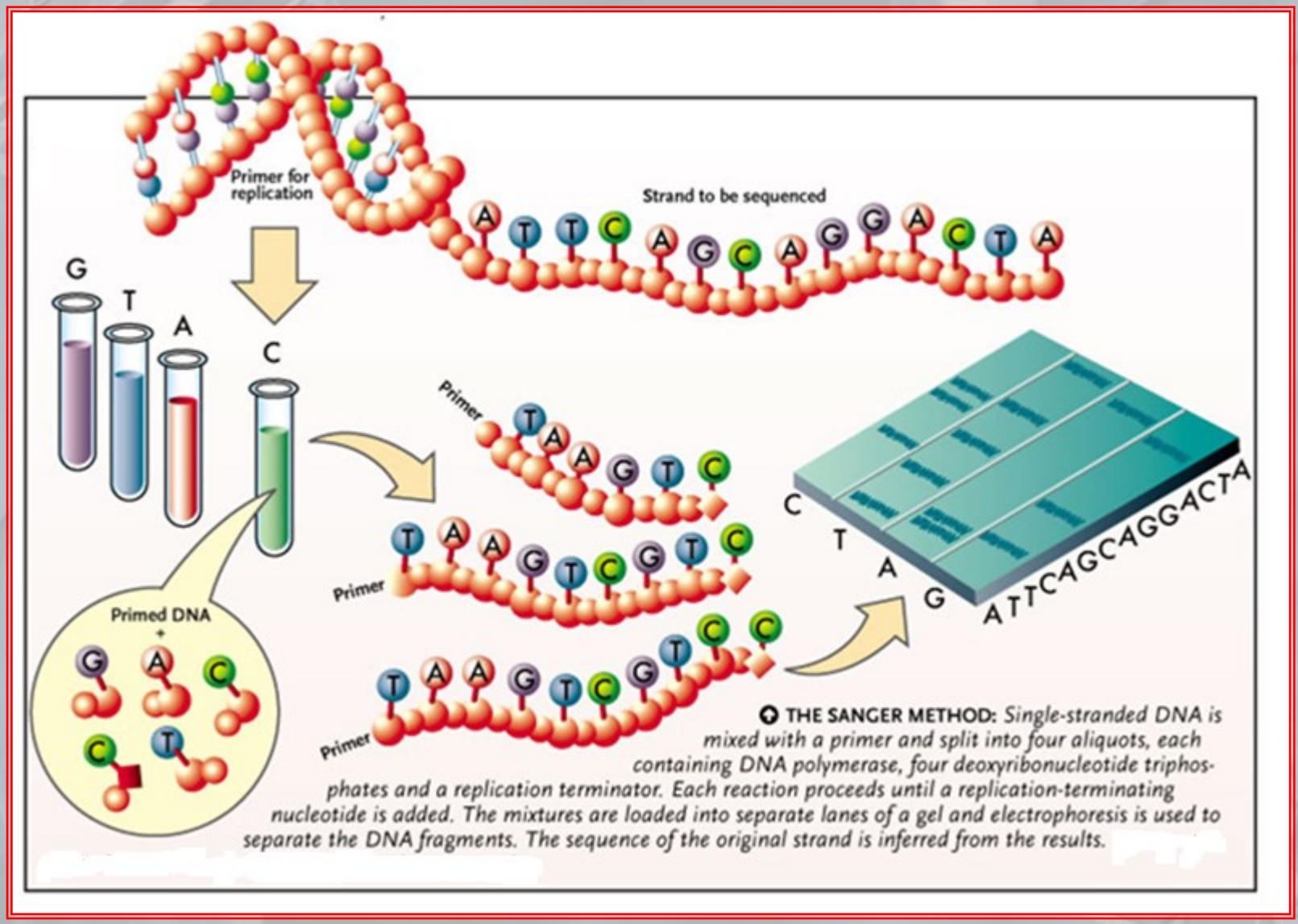
DNA sequencing – 5

- Sanger method
  - It is based on the selective incorporation of chain–terminating di–deoxynucleotides by DNA poly-merase during in vitro DNA replication
  - In other words… the group of fragments needed for sequencing is generated through the incorporation of modified nucleotides into the DNA, that prevent the DNA polymerase to add further bases to the chain
  - Such nucleotides differ from their standard counter-part because of the lack of the hydroxyl group (–**OH**) at the 3’ end, which should attach the next nucle-otide in a normal DNA growing fragment
    - They are equipped with tags (fluorescent dye) to locate them when they are divided by size

DNA sequencing – 6

- The method requires a single–strand DNA template, a DNA primer, the DNA polymerase, and modified nucleotides, which terminate the DNA strand elongation
- The DNA sample is divided into four separate sequencing reactions, containing all four standard nucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase; to each reaction is added only one of the four modified nucleotide types (ddATP, ddGTP, ddCTP, or ddTTP)
- Finally, the resulting DNA fragments are heat de-natured and separated by size using gel electro-phoresis

DNA sequencing – 7



THE SANGER METHOD: *Single-stranded DNA is mixed with a primer and split into four aliquots, each containing DNA polymerase, four deoxyribonucleotide triphosphates and a replication terminator. Each reaction proceeds until a replication-terminating nucleotide is added. The mixtures are loaded into separate lanes of a gel and electrophoresis is used to separate the DNA fragments. The sequence of the original strand is inferred from the results.*

DNA sequencing – 8

- Next Generation Sequencing (NGS)
  - A high–throughput method used to determine a portion of the nucleotide sequence of an individual's genome
  - It utilizes DNA sequencing technologies that are capable of processing multiple DNA sequences in parallel
  - Using NGS, an entire human genome can be sequenced within a single day
  - In contrast, the previous Sanger sequencing technology, used to decipher the human genome, required over a decade to deliver the final draft
  - Different technologies with some common steps – for more details on the Illumina process: https://www.youtube.com/watch?v=ToKUGz_YhC4

The C value paradox – 1

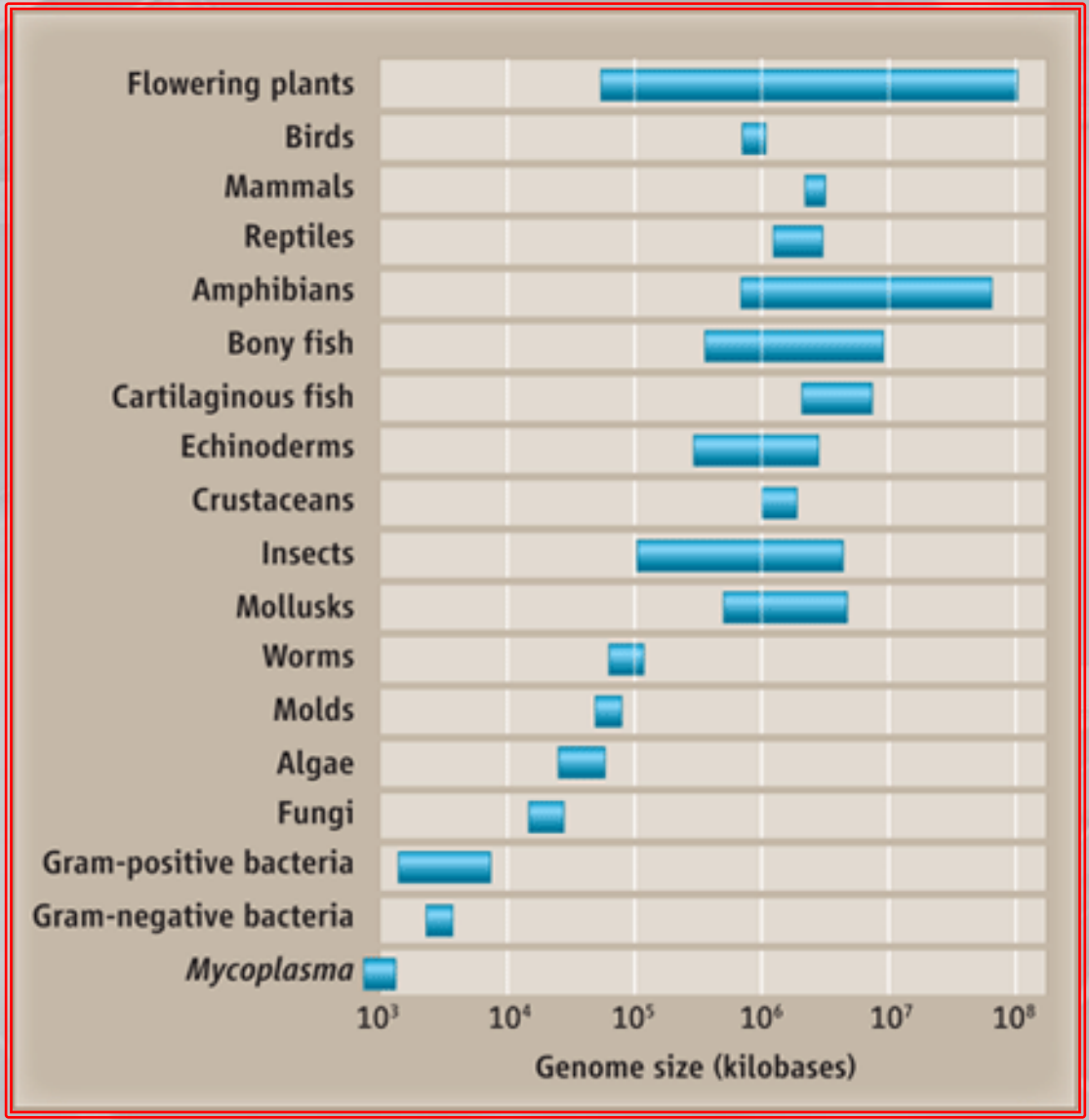
- In 1948, it was discovered that the DNA amount within each cell of the same organism is the same
- This quantity is called the C value, a term coined in 1950 by Henson Swift

  "*I am afraid the letter C stood for nothing more glamorous than 'constant', i.e., the amount of DNA that was characteristic of a particular genotype*" (H. Swift to Michael Bennett)
- It is worth noting that, while the size of the genome of a species is constant, there are strong variations among different species (possibly belonging to the same phylum), but without a correlation with the complexity of the organisms

The C value paradox – 2

- The lack of correlation between the complexity and the size of the genome gives rise to the C–value paradox
- The DNA amount often differs by a factor greater than one hundred among very similar species
- The clear implication – even if difficult to prove – is that a great quantity of DNA, present in some organisms, is in excess and, apparently, does not provide a significant contribution to the complexity of the organism itself

The C value paradox – 3



Reassociation kinetics – 1

- When the complementary strands of the double–strand DNA are separated, or denatured, by means of the heat or by alkali treatments, they can easily rejoin in the double–strand structure if the conditions within the cell return to normality
- The genome content can be understood by the way in which the denatured DNA reforms its structure
  - The more a genome sequence is rare, the greater is the time required for each strand to hybridize with its complementary strand

Reassociation kinetics – 2

- In 1968, Britten and Kohne described the cot equation, which defines the DNA reassociation kinetics in the form:

$$c_t/c_0 = 1/(1+kc_0t)$$

where c_t is the single–strand DNA concentration at time t (c_0 is the initial concentration), and k is a constant (dependent on the temperature, on the size of fragments, and on the complexity of the nucleotide sequence)

Reassociation kinetics – 3

- The *cot equation* establishes an inverse proportionality relation between the remaining fraction of single–strand DNA and $c_0 t$
- From this equation, a specific value $c_0 t_{1/2}$ can be derived for each organism, which is directly proportional to the number of nucleotides belonging to not repeated sequences
- Therefore, the time interval $t_{1/2}$, required by half of the single–strand DNA to rejoin their complementary counterpart (c_t/c_0=0.5), allows the experimental determination of the amount of "unique" information encoded in a given genome

Reassociation kinetics – 4

- Since $c_0 t_{1/2}$ is the product of the concentration and the time required for the renaturation of half of the DNA, a large value of $c_0 t_{1/2}$ implies a slow reaction, and reflects the situation in which there are few copies (or only one copy) of a particular sequence within a given DNA
- The value $c_0 t_{1/2}$ is a measure of the total length of different sequences within the genome and, therefore, it "describes its complexity"

Reassociation kinetics – 5

- However, if the initial denatured DNA concentration is 10 pg, it can "contain" 2000 copies of each sequence of a bacterial genome whose size is 0.005pg, but it will be made up of only two copies of a 5pg eukaryotic genome
- Since the renaturation speed depends on the concentration of the complementary sequences, the initial amount of the eukaryotic genome should be a thousand times greater

Reassociation kinetics – 6

- Therefore, while the C value is not directly related to the complexity of an organism, the value $c_0 t_{1/2}$ represents a significant marker
- The disparity between the two values usually indicates the presence of multiple copies of unnecessary DNA sequences, called *junk DNA*
- Repeated sequences constituting junk DNAs differ much in terms of complexity (from one or two to thousands of nucleotides) and distribution (local groupings or random sparsity) within the genome

Concluding… – 1

- DNA molecules contain the information stored within the cell needed for its life and reproduction
- The ordered chain of the four different nucleotides is transcribed by RNA polymerase into mRNA, which is then translated into proteins by ribosomes
- Twenty different amino acids are used to build proteins, and the order and the specific composition of these "building blocks" play an important role in the construction and in the structural and functional maintenance of proteins

Concluding… – 2

- Molecular biologists have a rather limited number of tools to study DNA
  - Restriction enzymes cut the DNA molecule when they recognize a specific string of nucleotides
  - Gel electrophoresis allows the separation of DNA fragments according to their length and charge
  - Blotting and hybridization ensure the retrieval of specific DNA fragments in a mixture, so as microarrays
  - The cloning technique allows the propagation of specific sequences, for repeated use and analysis
  - PCR provides a rapid amplification and characterization of specific DNA fragments
  - Finally, DNA sequencing determines the order of nucleotides, ensuring the characterization of the DNA molecule

Concluding… – 3

- The reassociation kinetics revealed that the DNA content of a cell, its C value, does not always directly correspond to the information content (the complexity) of an organism
  - In complex organisms, there are large areas of junk (useless) DNA