Genomics and gene recognition

“Είμαστε αυτό που κάνουμε επανειλημμένα. Ως εκ τούτου, η Αριστεία, δεν είναι πράξη αλλά συνήθεια”
“We are what we repeatedly do. Excellence, then, is not an act, but a habit.” (*Aristotèlēs*)

Table of contents

- Genomics
- The genome of prokaryotes
- The structure of prokaryotic genes
- GC content in prokaryotic genes
- Density of prokaryotic genomes
- The genome of eukaryotes
- Open reading frames
- GC content in eukaryotic genes
- Gene expression
- Transposition
- Repeated elements
- Density of eukaryotic genes

Introduction – 1

- The enormous development of biomolecular investigation techniques makes it possible to acquire genetic information at a rate unimaginable until very recent years (e.g. DNA sequences, transcription profiles, protein structures, etc.)
- There has been a drastic change in perspective and horizons in biological research

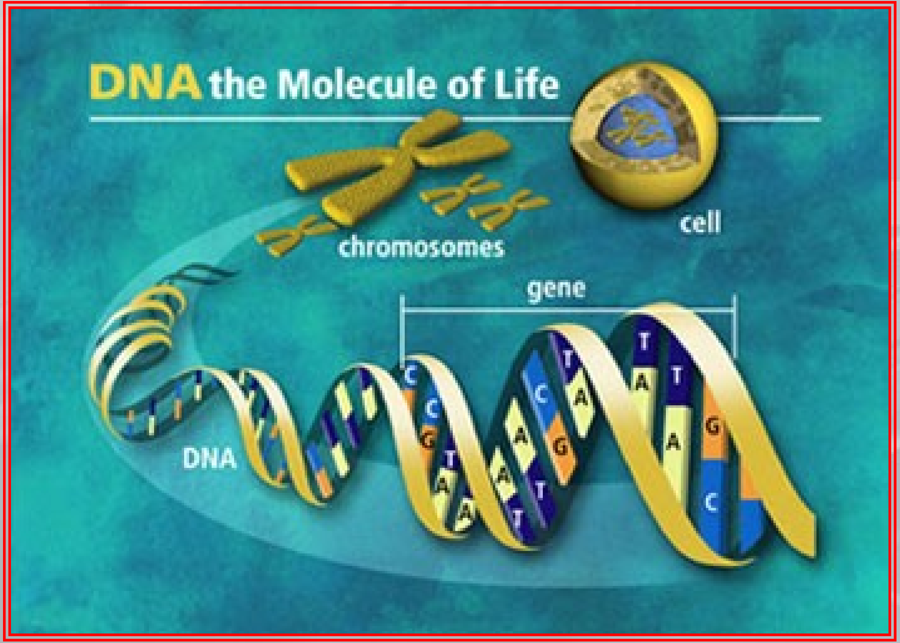
From GENE to GENOME

The set of elementary units of a given system (e.g.: Proteome, Exome, etc.)

Genome: 1920 Hans Winkler, Botanist
Genomics: 1979 Victor McKusic, Geneticist

Introduction – 2

- The analogy between written text and genome information is evident
  - Alphabetic characters correspond to nucleotides
  - Sentences correspond to genes
  - Volumes that make up an encyclopedia are comparable to chromosomes



Introduction – 3

- However, deciphering the information content of a genome is much more difficult than determining the meaning of a text, though written in an unfamiliar language
  - Difficulty in establishing the start/end point of each “sentence” and in fully understanding its meaning
  - In eukaryotes, the genome is mixed with a surprising amount of “junk” DNA, with no information content (at the best of our knowledge…)

Introduction – 4

- However, like any other system for information storage, the genome contains signals that allow the cell to determine the beginning and the end of a gene and when/how much it should be expressed
  - A sense must be attached to the “disconcerting” organization of **A**, **T**, **C**, and **G**, which is a typical raw genomic sequence
- Finally, it is worth noting that the development of new tools for finding genes did turn our attention to before unsuspected biological mechanisms, responsible for the gene expression regulation

Genomics

- Structural genomics: It deals with the study of the genome structure, with the identification of genes and of their expression products, with the analysis of promoter and regulatory elements
- Functional genomics: It deals with the study of the function of genes, their interactions (metabolic pathways), the mechanisms that regulate their expression in healthy cells and how their malfunction induces a pathological state
- The research on structural and functional genomics, which form together comparative genomics, derives a great benefit from the correlative analysis of genomes and of their expression products
  - Comparisons among "homologous" entities help in interpreting genetic information

Comparative genomics

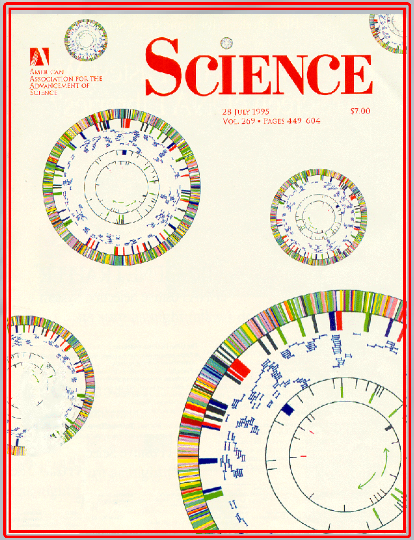
Nothing in Biology makes sense except in the light of evolution

Theodosius Grigorevich Dobzhansky
(1900 - 1975)

- Conservation allows us to observe the effects of evolution
- What it is stored or preserved during evolution is very likely to have a precise biological function
- Conservation can be realized at the sequence level (nucleotide or protein), at the structure level, at the expression level, etc.
- Therefore, we can assign the same function to genes (or to other biological entities) that are similar and conserved during the evolution process

The genomic era

- 1995: The year of publication of the first prokaryotic genome (*Haemophilus influenzae*) marks the beginning of genomics
- Since that date, many other genomes have been sequenced, both of prokaryotic and eukaryotic organisms
- Currently (10.11.2024), there are 517466 sequenced genomes (5366 archaea, 437610 bacteria, 22124 viruses, 52366 eukaryotes), 37914 of which were complete and published



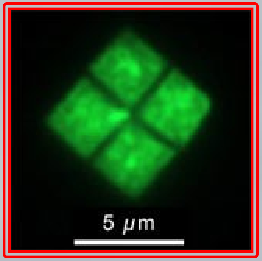
(Source: GOLD, *Genomes On Line Databases* https://gold.jgi.doe.gov/)

The genome of prokaryotes – 1

- Prokaryotes, from the Greek words *Pro* “before, in front” and *Karyon* “core, nucleus”, are microscopic single–cell or colonial organisms (micron size), living in a variety of environments (soil, water, other bodies)
- Although more than 10000 prokaryotic species are known today, it is estimated that their number is possibly 10^6 to 10^7
- The definition of “species” in the case of bacteria is somewhat arbitrary and normally relies on a series of biochemical, molecular (e.g., 16s rRNA), and morphological peculiarities

The genome of prokaryotes – 2

- Molecular classification subdivides prokaryotes into two domains: bacteria and archaea that, with eukaryotes, form the three main branches of the tree of life
- Archaea and bacteria are generally similar in size and shape, although a few archaea actually show unusual patterns (such as the flat and square–shaped cells of *Haloquadratum walsbyi*)



- Despite this visual similarity to bacteria, archaea possess genes and several metabolic pathways that are more closely related to those of eukaryotes, notably the enzymes involved in transcription and translation processes

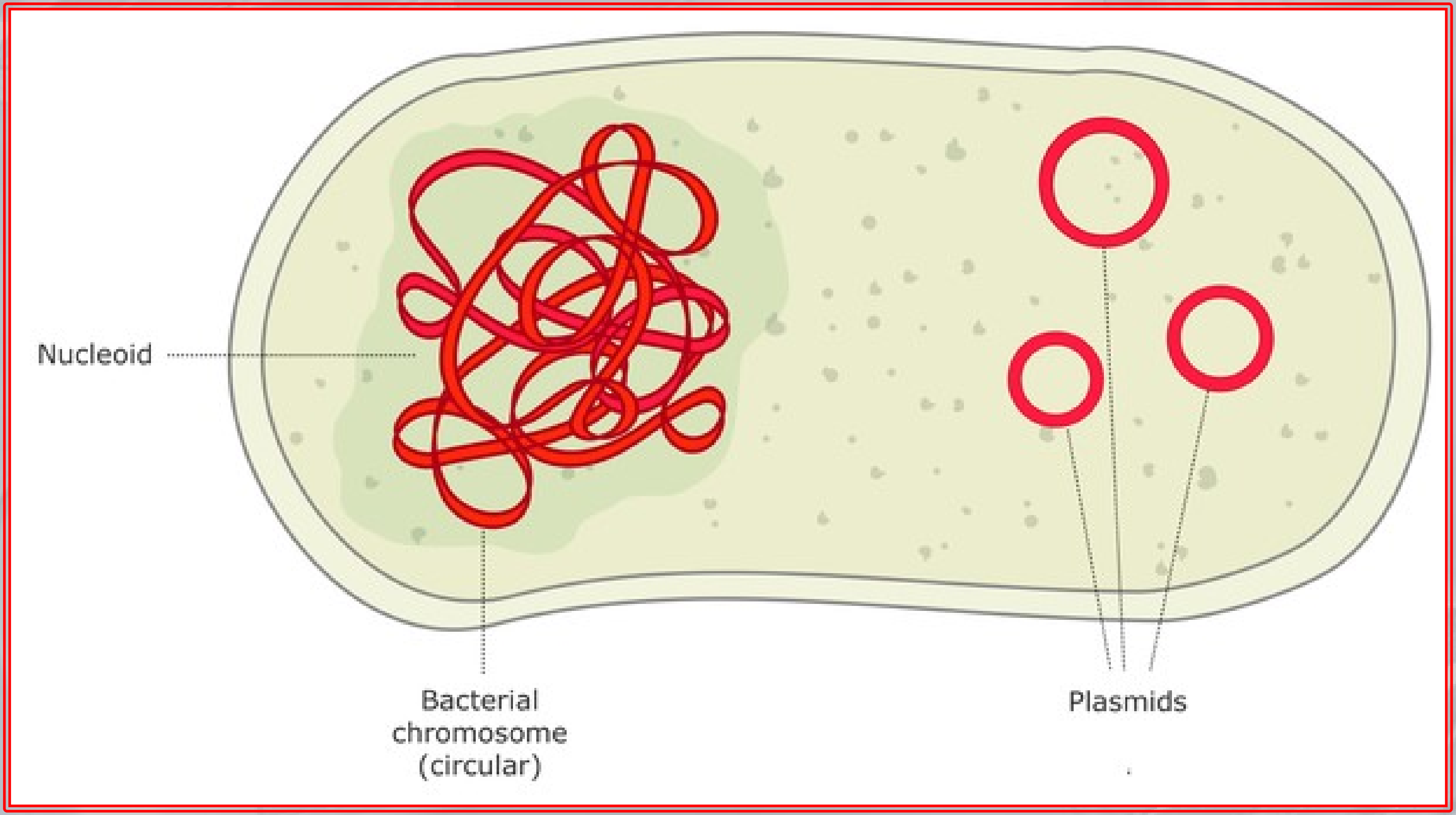
The genome of prokaryotes – 3

- The ability to respond to external stimuli is the central feature of the concept of living organisms
- Being prokaryotes the simplest forms of life, they represent an excellent case–study to determine the molecular basis of such responses
- Actually, in a prokaryotic perspective, appropriate reactions to external stimuli invariably involve alterations in the gene expression levels
- The ability to analyze the whole bacterial genome provides a particularly relevant aid to understand the minimal requirements for life

The genome of prokaryotes – 4

- A great deal of information contained in the prokaryotic genome is dedicated to maintaining the basic infrastructure of the cell, such as its ability in:
  - building and replicating the DNA (no more than 32 genes)
  - synthesizing proteins (between 100 and 150 genes)
  - obtaining and storing energy (at least 30 genes)
- Prokaryotic genomes are generally constituted by a single circular chromosome
- In many species, small circular extrachromosomal DNAs are also present, coding for additional genes, used to better fit the external environment

The genome of prokaryotes – 5



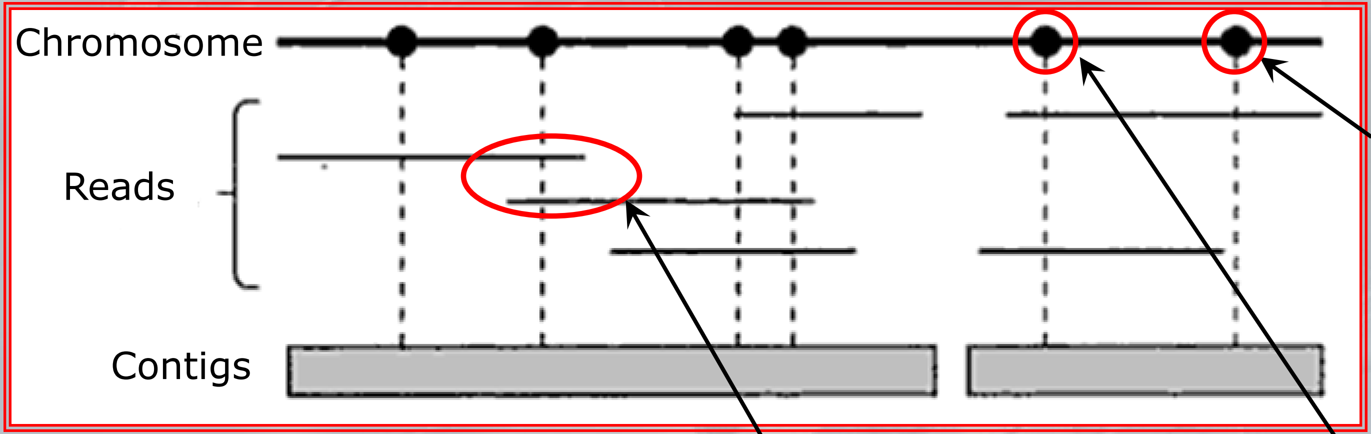
The genome of prokaryotes – 6

- In particular:
  - Some very simple prokaryotes, such as *Haemophilus influenzae* (the first to be completely sequenced), have a genome just a little longer than the minimum (between 256 and 300 genes)
  - More complex prokaryotes use their additional information content to efficiently take advantage of the wide range of resources that can be found in the outdoor environment

The genome of prokaryotes – 7

- DNA sequencing techniques are essentially unchanged from the ‘80s, apart from the recent NGS approach
- The key difference between Sanger and NGS is the sequencing volume, as NGS is highly parallel and produces millions of fragments per run
- Anyway, contiguous data fragments, called reads, longer than 1000 nucleotides are rarely provided
- With a single circular chromosome of 4.6 million nucleotides…
  - The *Escherichia coli* genome requires a minimum of 4600 reactions to be completely sequenced
  - Set of contiguos (partially overlapped) reads are then assembled into contigs, used to guide DNA sequencing and assembly
  - Contigs are continuous sequences of nucleotides longer than that obtained from a single sequencing reaction

The genome of prokaryotes – 8



Sequence-tagged sites (STSs) are short (200/500 base pairs) DNA sequences that occur only once in the genome and whose location and base sequence are known

An overlapping zone which may help in reconstructing the whole sequence

Expressed sequence tags (ESTs) are short final subsequence of cDNA sequences

- Furthermore, what has become the standard approach to genomic sequencing usually starts with a random assortment of subclones of the genome of interest
  - There are no guarantees that any portion of the genome is represented at least once, if we do not accept also the presence of replicated regions

The genome of prokaryotes – 9

- From the statistical point of view:
  - the probability to cover a particular nucleotide in a genome of 4.6 million bases with a single clone, 1000 base pairs long, is equal to 1000/4600000 ≈ 2.174×10^{-4}
  - vice versa, the probability that a specific region is not covered is 4599000/4600000 ≈ 0.9998
- Assuming that, in a given library, a large enough sample of subclones was present, a 95% coverage is obtained, having sequenced N clones, with N such that

$$(4599000/4600000)^N = 0.05$$

  - It is necessary to have more than 20 million nucleotides (20000 subclones, approximately four genome–equivalents) to obtain a 95% probability that each sequence is represented at least once

The genome of prokaryotes – 10

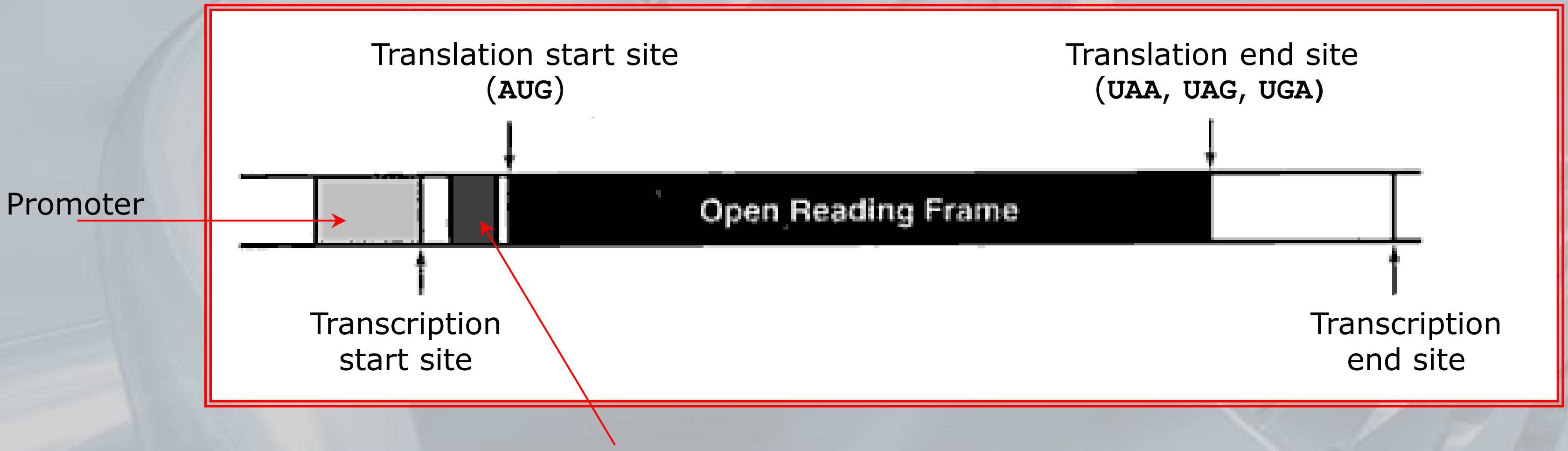
- In spite of logistic difficulties associated with both experimental and computational aspects of DNA sequencing, more than 517K procaryotic genomes have been sequenced to date
- … and since 2000, such sequencing could be achieved within a few weeks/days – now hours
    - In 2001, The U.S. Center for Disease Control and Prevention was able to compare the complete genome of anthrax species, used in bio–terroristic attacks via mail, in less than a month, although such species differed only for four nucleotides (out of more than 5 millions) from those used in US military laboratories

Structure of prokaryotic genes – 1

- Prokaryotic genomes have a very high gene density: on average, the protein–coding genes occupy 85–88% of the genome
- In addition, prokaryotic genes are not interrupted by introns and are sometimes organized in transcriptional polycistronic units (leading information related to several genes), called operons
- The high plasticity of prokaryotic genomes is reflected by the fact that the order of genes along the genome is poorly conserved among different species and taxonomic groups
- Therefore, groups of contiguous genes contained in a single operon in a certain genome can be dispersed in another

Structure of prokaryotic genes – 2

- The structure of prokaryotic genes is normally quite simple
  - Just as we rely on punctuation to decipher the information contained in a written text, proteins, responsible for the gene expression, search for a recurring set of signals associated with each gene



Operator: a DNA segment to which a transcription factor binds to regulate the gene expression

Structure of prokaryotic genes – 3

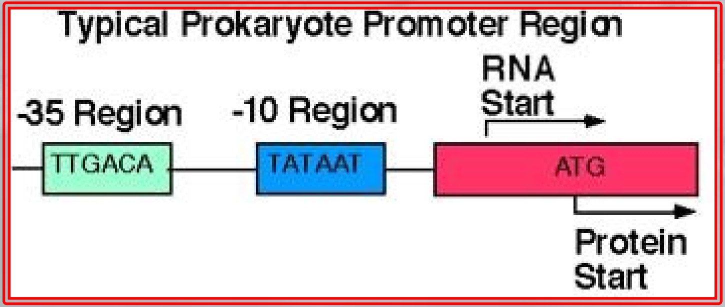
- These genomic punctuation marks and their some-times subtle changes, allow to
  - distinguish between genes that must be expressed
  - identify the beginning and the end of the regions that must be copied into RNA
  - demarcate the beginning and the end of the RNA regions that ribosomes must translate into proteins
- Such signals are represented by short strings of nucleotides, which constitute only a small fraction of the hundreds/thousands of nucleotides necessary to encode the amino acid sequence of a protein

Promoter elements – 1

- The process of gene expression starts with the transcription – the production of an RNA copy of a gene realized by the RNA polymerase
- The prokaryotic RNA polymerases are actually assemblies of a set of different protein subunits, each of which plays a distinct and important role in the overall functioning of the enzyme
- The activities of all the prokaryotic RNA polymerases depend on four different types of protein subunits
  - β', which has the ability to bind the template DNA
  - β, that binds a nucleotide to another
  - α, that holds together all the subunits
  - σ, which is able to recognize the specific nucleotide sequence of the promoter

Promoter elements – 2

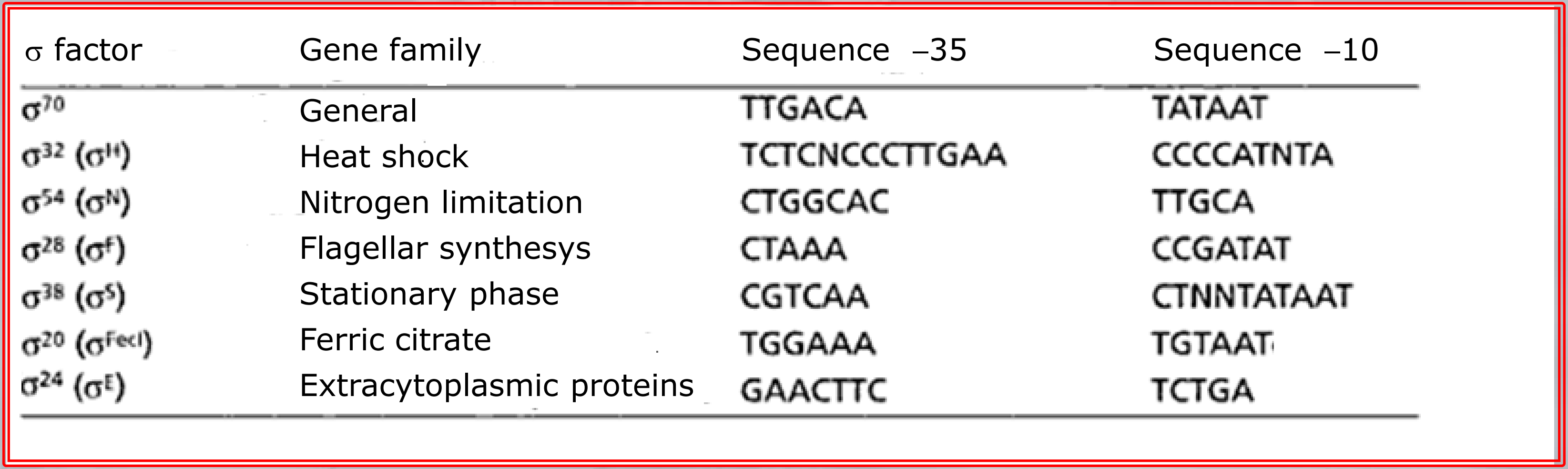
- The subunits β', β and α are well preserved from the evolutionary point of view and are often very similar from one bacterial species to another
- Instead, the subunits σ, responsible for the recognition of the promoter, tend to be less conserved and several variants have been detected in different cell types
- The ability to form RNA polymerases with significantly different σ subunits is the factor responsible for the possibility given to the cell to activate or deactivate the expression of whole sets of genes



Promoter elements – 3

- Example 1

  E.coli has seven different σ factors:



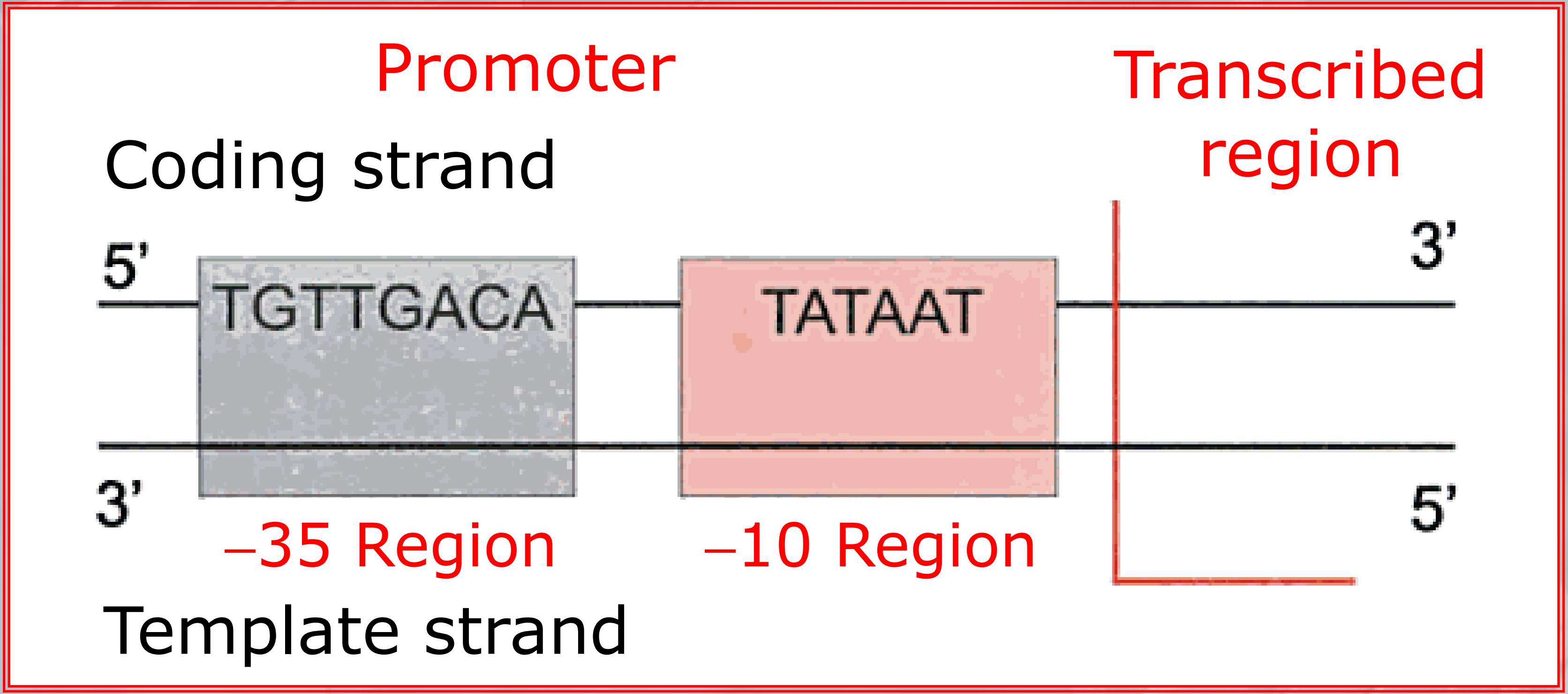
σ factor	Gene family	Sequence –35	Sequence –10
σ^{70}	General	TTGACA	TATAAT
σ^{32} (σ^{H})	Heat shock	TCTCNCCCTTGAA	CCCCATNTA
σ^{54} (σ^{N})	Nitrogen limitation	CTGGCAC	TTGCA
σ^{28} (σ^{F})	Flagellar synthesys	CTAAA	CCGATAT
σ^{38} (σ^{S})	Stationary phase	CGTCAA	CTNNTATAAT
σ^{20} (σ^{FecI})	Ferric citrate	TGGAAA	TGTAAT
σ^{24} (σ^{E})	Extracytoplasmic proteins	GAACTTC	TCTGA

  - When *E.coli* has to express the genes involved in the response to a drastic rise in temperature, the RNA polymerases containing σ^{32} seek and find the genes with σ^{32} promoters
  - Approximately 70% of the *E.coli* genes that need to be always expressed during the normal development of the organism are transcribed by RNA polymerases containing σ^{70}

Promoter elements – 4

- The accuracy with which the RNA polymerase recognizes a gene promoter is directly related to how easily the process of transcription begins
- The sequences placed at –35 and –10 (w.r.t. the transcription starting site) recognized by a particular σ factor are called consensus sequences, and represent the set of nucleotides most commonly identified in equivalent positions of several genes transcribed by the RNA polymerases containing the same σ factor
  - The greater the similarity of the sequences placed at –35 and –10 with the consensus sequences, the greater the likelihood that the RNA polymerases actually start the gene transcription from that promoter

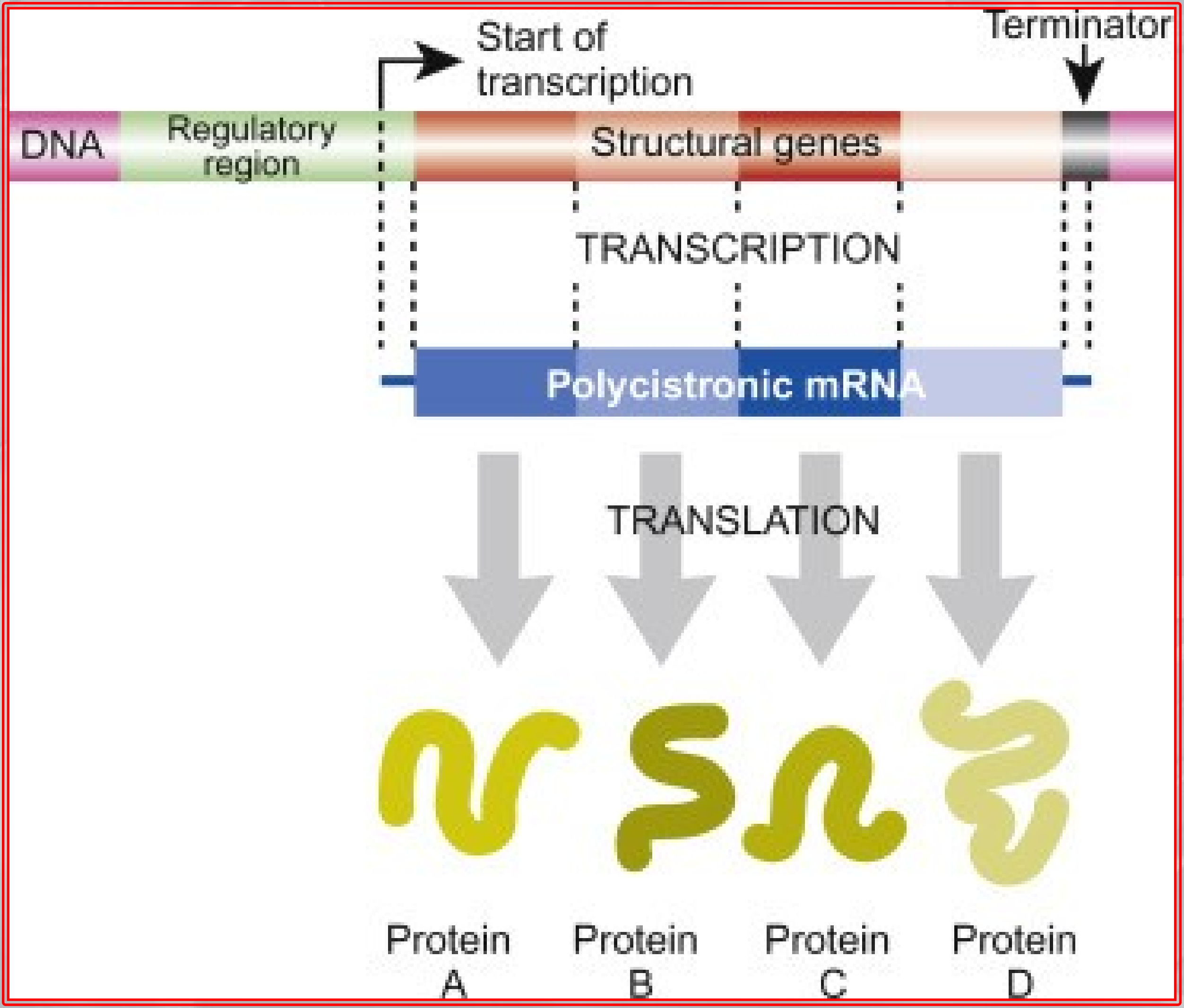
Promoter elements – 5



Promoter elements – 6

- The protein products of many genes are useful only when used in conjunction with the protein products of other genes
- It is very common to have a single, shared, promoter for the expression of genes with related functions in prokaryotic genomes, and that such pool of genes is rearranged in an operon
  - This constitutes a simple and elegant way to ensure that, when a gene is transcribed, all other genes with similar/related roles are also transcribed

Promoter elements – 7



Promoter elements – 8

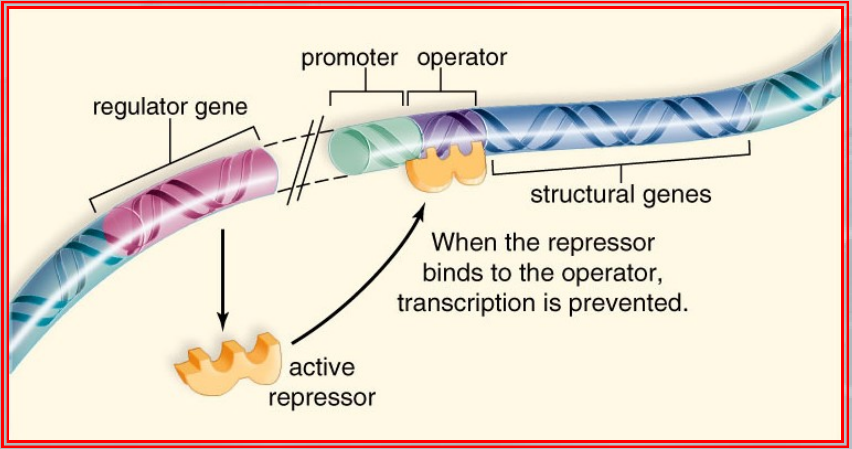
- Example 2
  - The lactose operon is a set of three genes (coding for *beta–galactosidase*, *lactose permease* and *lactose transacetylase*) involved in the metabolism of the lactose sugar in bacterial cells
  - The operon transcription gives rise to the synthesis of a single, very long, RNA molecule, called polycistronic RNA, which contains the coding information needed by ribosomes to synthesize the three proteins
- Regulatory proteins can facilitate the expression of some bacterial genes in response to specific environmental factors, with a much finer adjustment than that achievable using different σ factors

Promoter elements – 9

- Example 2 (cont.)
  - *E.coli* is a bacterium capable of using as a carbon source both glucose and lactose
  - The best suited sugar to its metabolism is glucose, so that, if the bacterium grows in a substrate that presents both sugars, it first uses glucose and, only after, lactose
  - However, if the bacterium grows in an environment in which only lactose is present, it immediately expresses those enzymes needed to metabolize such sugar
  - *E.coli* possesses, therefore, a control mechanism that allows the expression of some genes only when it is needed, and prevents the production of enzymes and proteins that are not strictly necessary

Promoter elements – 10

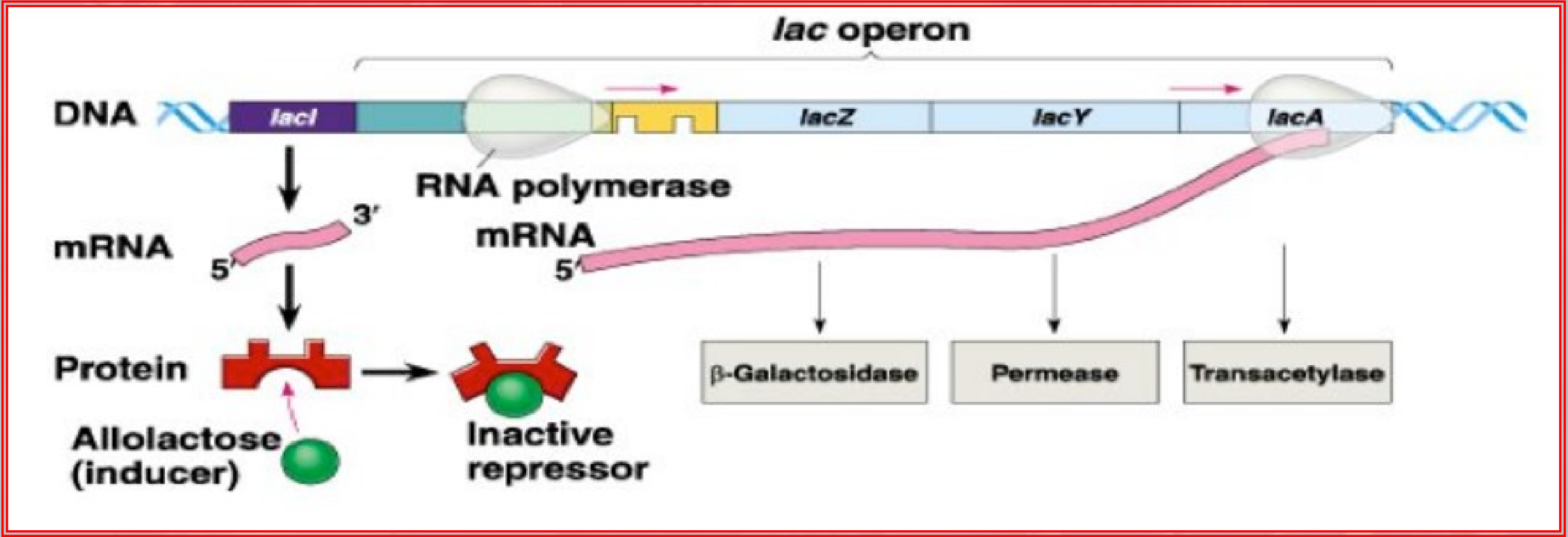
- Example 2 (cont.)
  - The responsiveness to the lactose levels is mediated through a negative regulator, called lactose repressor protein (pLacI), that, when bound to the DNA (in the area of the operator) prevents the polymerase to transcribe the operon
    - When, in the environment, lactose is present, the derivated compound called allolactose binds to the repressor protein, so as to prevent its link with the DNA, making possible the transcription of the operon
    - Even in presence of lactose, the transcription of the operon is poor until glucose is present, since it remains the most easily usable sugar for *E.coli*


promoter
operator
regulator gene
structural genes
When the repressor binds to the operator, transcription is prevented.
active repressor

Promoter elements – 11

- Example 2
  - Instead, with a scarce glucose concentration, the cyclic AMP (cAMP), a molecule that in all the organisms acts as a signal of energy shortage, is produced within the cell
  - The cAMP, binding to CRP (a receptor protein, which acts as a positive regulator), makes it able to bind to DNA (upstream w.r.t the promoter), greatly stimulating the transcription of the operon
  - In summary:
    - In the presence of glucose and lactose, both the repressor and the CRP are inactive ⇨ there is a reduced transcription
    - In the presence of glucose but not lactose, the repressor is active and the CRP is inactive ⇨ there is no transcription
    - In the absence of glucose and lactose, both the repressor and the CRP are active ⇨ there is no transcription
    - In the presence of lactose and absence of glucose, the repressor is inactive and the CRP is active ⇨ the operon is expressed to the maximum level

Promoter elements – 12



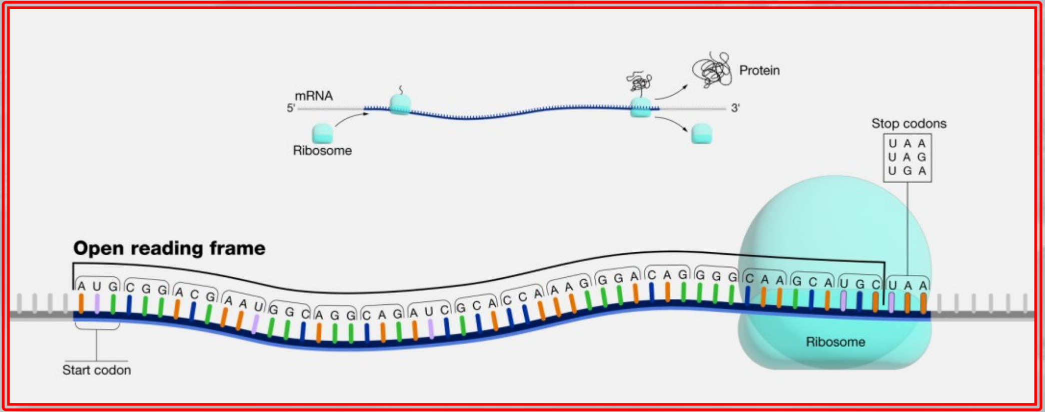
Lac operon

Promoter elements – 13

- Bioinformatics tools, such as pattern matching techniques, can be applied in this context, to detect promoter sequences (placed in position –35 and –10) recognized by the RNA polymerase
  - The penalty score for each nucleotide mismatch within a sequence of the putative promoter allows different operons to be classified according to the greater or less probability of being expressed at high levels in the absence of positive regulators
  - Conversely, many regulatory proteins (such as CRP) were discovered by noting that a particular string of nucleotides, different from the sequences in –35 and –10, was associated with more than one operon promoter

Open reading frames – 1

- Ribosomes translate the triplets of an RNA copy of a gene in the specific amino acid sequence of a protein
- Among all the 64 possible arrangements, three of these codons (**`UAA`**, **`UAG`** and **`UGA`**) functionally act as a full stop at the end of a sentence, causing the termination of the translation phase
- Since stop codons, in uninformative sequences, approximately appear 1 out of 21 positions (3/64), a sequence formed by 30 or more codons that does not include a stop codon, an open reading frame or an ORF, most likely corresponds to the coding sequence of a prokaryotic gene

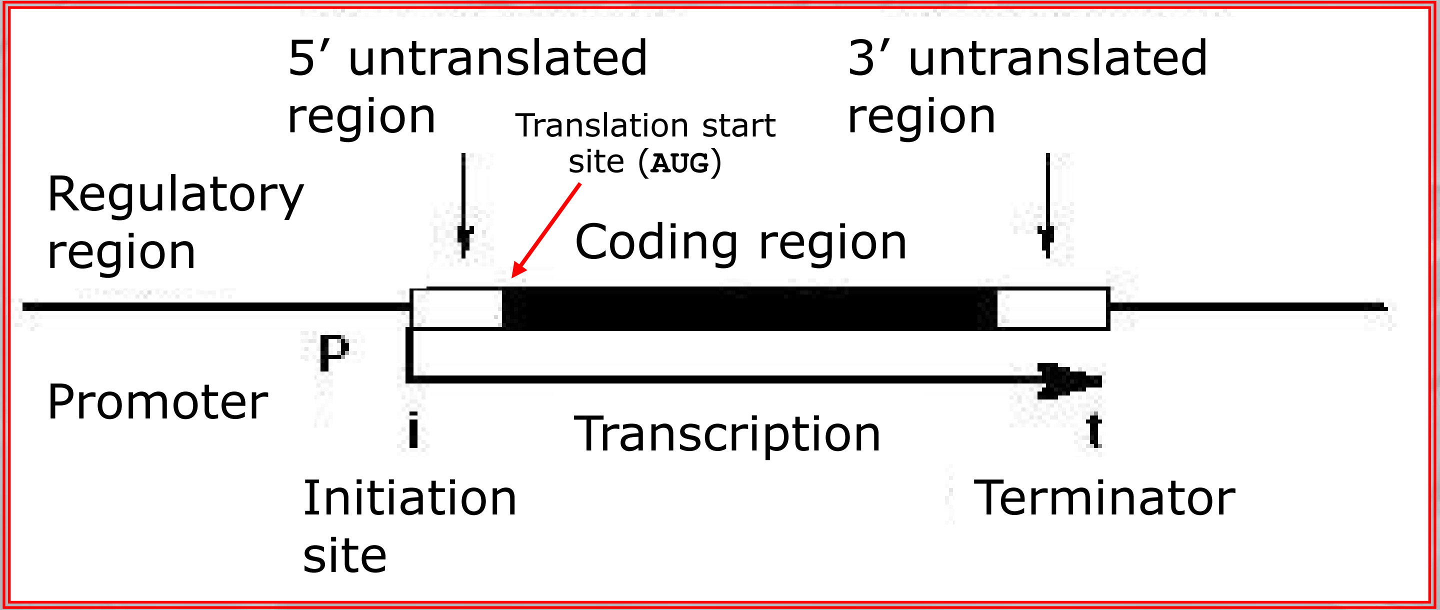

mRNA
5'
3'
Ribosome
Protein
Stop codons
U A A
U A G
U G A
Open reading frame
A U G C G G A C G A A U G G C A G G C A G A U C G C A C C A A A G G G A C A G G G G C A A G C A U G C U A A
Start codon
Ribosome

Open reading frames – 2

- Statistically, if all the codons were present with the same frequency within a random DNA sequence, the probability that a sequence of length N does not contain a stop codon is $(61/64)^N$
- A confidence of 95% on the significance of an ORF is equivalent to the 5% probability of a random success, $(61/64)^N=0.05$
  - $N = 60$
  - Many algorithms for gene mapping in prokaryotic organisms decree the significance of an ORF just according to its length
- Many prokaryotic proteins are formed by more than 60 amino acids
  - Example: In *E.coli*, the average length of a coding region is 316.8 codons, whereas less than 1.8% of the genes are shorter than 60 codons

Open reading frames – 3

- So as three codons are intended to be stop codons, a particular triplet is usually employed as a start codon
- In particular, **AUG** is used both to codify methionine, and to mark the point, along the RNA molecule, where the translation starts
  - **AUG** is the first codon for 83% of *E.coli* genes, while **UUG** and **GUG** are the start codons for the remaining 17%

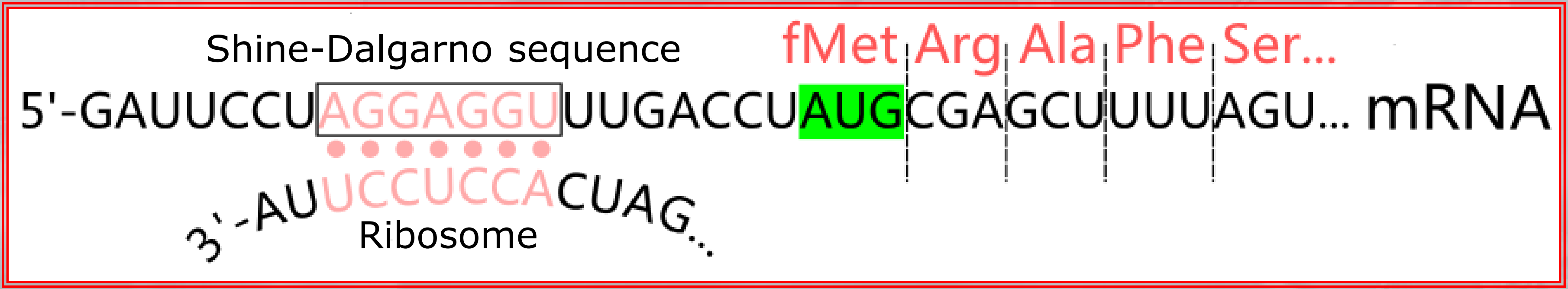


Open reading frames – 4

- If a promoter sequence cannot be found upstream of the start codon of an ORF (and after the end of the previous ORF), it is assumed that the two genes are part of a single operon, the expression of which is controlled by a promoter further upstream
- Another feature of prokaryotic genes, related to their translation, is the presence of a set of sequences around which ribosomes are assembled, located at the 5′ end of each ORF, immediately downstream of the start site of transcription and just upstream of the translation start codon
  - The docking sites of the ribosomes, called Shine–Dalgarno sequences, are purine–rich and almost invariably include the nucleotide sequence 5′–**AGGAGGU**–3′

Open reading frames – 5

- Point mutations in the Shine–Dalgarno sequence of a gene may prevent the translation of an mRNA
- In some bacterial mRNA, where there are very few nucleotides between successive ORFs, the translation of adjacent coding regions in a polycistronic mRNA are linked together because ribosomes gain access to the start codon of the next ORF when they have just completed the translation of the current ORF
- Nevertheless, usually, each start codon is characterized by its own Shine–Dalgarno sequence



Conceptual translation – 1

- During the ‘60s and the ‘70s it was much easier to determine the amino acid sequence of a protein rather than the nucleotide sequence of its encoding gene
- The recent and rapid evolution of methods for DNA sequencing has, however, led to the current situation where the vast majority of protein sequences is derived from their nucleotide sequences
- The process of conceptual translation of a gene sequence into the corresponding amino acid sequence is, in fact, an easily automatable process

Conceptual translation – 2

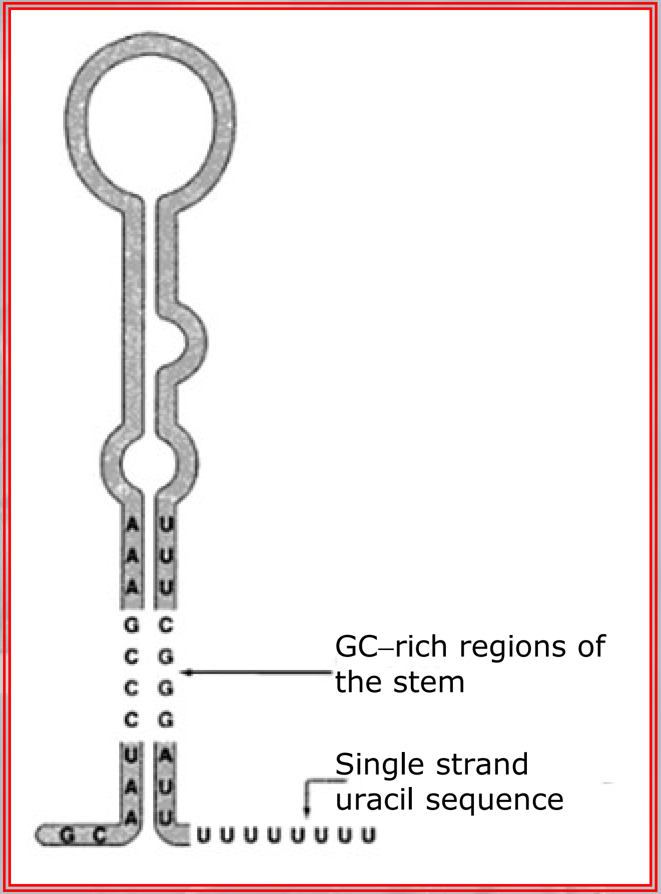
- Amino acid sequences can then be studied to predict their structural trends, such as the propensity to form α–helices or β–sheets
- However, the prediction of the protein structure based on the amino acid sequence (primary structure analysis) rarely produces more than an estimate of the protein function
  - The comparison with amino acid sequences of proteins from different, better characterized, organisms, as well as the promoter sequence and the genomic context of the encoding gene, often provide much more reliable clues on the role of a protein

Terminator sequences – 1

- As the RNA polymerase starts the transcription from easily recognizable sites, placed immediately down-stream of the promoters, so the great majority of prokaryotic genes (over 90%) also contains specific signals for the termination of transcription, called intrinsic terminators
  1) Nucleotide sequences that include an inverted repeated sequence
     - Example: 5′–**CGGAUGC|GCAUCCG**–3′
  2) …immediately followed by a sequence composed by (about) six uracils
- In the intrinsic terminators, each inverted repeated sequence is from 7 to 20 nucleotides long and is rich in **G** and **C**

Terminator sequences – 2

- Although RNA molecules are usually described as single–strand, they can actually adopt secondary structures, due to the formation of intra–molecular base pairs within the inverted repeats
- The stability of the RNA secondary structure is directly connected to the length of the inverted repeats (often imperfect) and to the number of **C/G** and **A/U** inside these repetitions



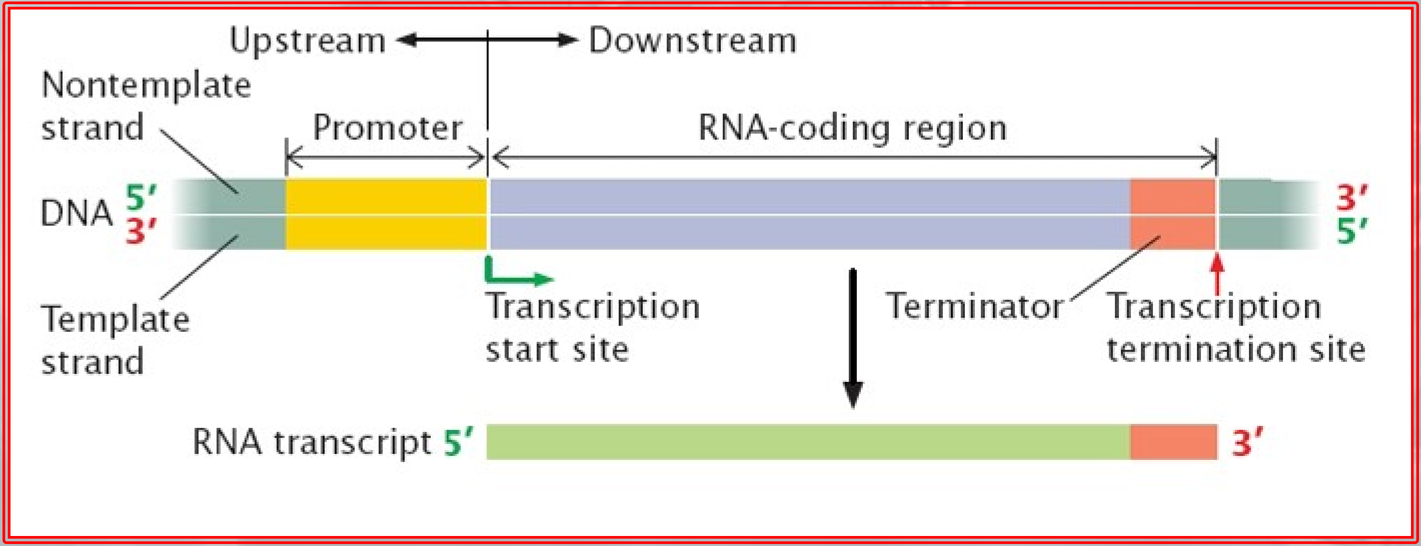
"Hairpin" RNA structure

Terminator sequences – 3

- It has been experimentally proved that the formation of a secondary structure in an RNA molecule, during its transcription, causes a break of the RNA polymerase of approximately one minute
  - The prokaryotic RNA polymerases normally incorporate one hundred nucleotides per second!
- If the RNA polymerase pause occurs during the synthesis of a sequence of uracils within the new RNA molecule, the unusually weak coupling of bases that occurs between the RNA uracils and the template DNA adenines causes the two polynucleotides to dissociate which, indeed, stops the transcription

Terminator sequences – 4

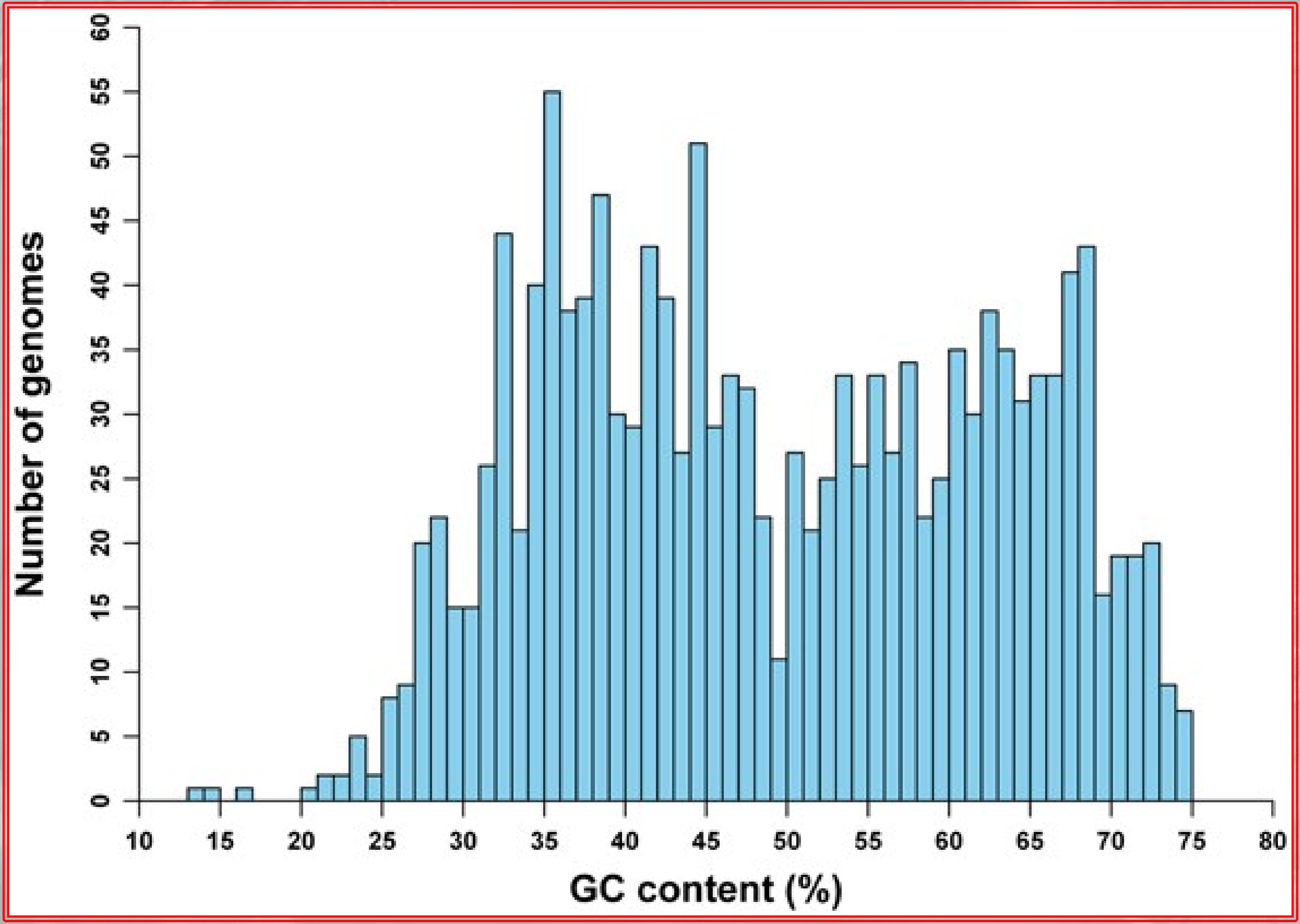
- In other words… while the standard process of transcription by RNA polymerases allows them to transcribe such adenine sequences within the template DNA, in conjunction with a break of the synthesis caused by the RNA secondary structure, the instability of the base coupling uracil/adenine leads to stop the transcription



GC content in prokaryotic genomes – 1

- The coupling rules between bases require that, in a double–strand DNA, each **G** corresponds to a complementary **C**, each **A** to a complementary **T**, but the only physical constraint with regard to the fraction of nucleotides **G**/**C** as opposed to that of **A**/**T** is that they sum up to 100%
- The abundance of nucleotides **G** and **C** with respect to **A** and **T** has long been recognized as a distinctive attribute of bacterial genomes
- The measurement of the **GC** content in prokaryotic genomes is very variable, ranging from 25% to 75%
- It was also noted that the base composition is not uniform along the genome

GC content in prokaryotic genomes – 2



GC content in prokaryotic genomes – 3

- The GC content of each bacterial species seems to be independently modeled by a tendency to mutations in its DNA polymerase and by the mechanisms of DNA repair acting over extended periods of time
  - The relative ratio between G/C and A/T remains almost constant in any bacterial genome
- Having available the complete sequence of an increasing number of prokaryotic genomes, the analysis of their GC content revealed that most of the bacterial evolution takes place on a large scale through the acquisition of genes from other organisms, through a process called horizontal gene transfer

GC content in prokaryotic genomes – 4

- Given that the bacterial species have a significantly variable GC content, the genes that were most recently acquired by horizontal gene transfer often have a GC content very different from that originally possessed by the genome
- Moreover, the differences in the GC content lead to somewhat different preferences in the use of codons and, consequently, in the use of amino acids, between the genes recently acquired and those historically present within the genome
    - Many bacterial genomes are “patchwork” of regions with different GC content, which reflects the evolutionary history of bacteria based on their environmental and pathogenic characteristics

Horizontal gene transfer – 1

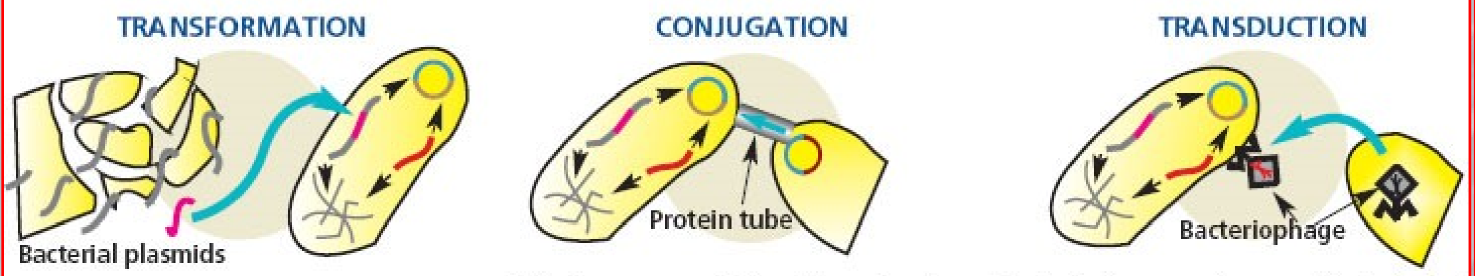
- Entire genes, a set of genes, or even whole chromosomes can be transferred from one organism to another
- Unlike many eukaryotes, prokaryotes do not sexually reproduce
- However, there are mechanisms that allow genetic exchange also in prokaryotes, both based on gene transfer and recombination; these mechanisms are fundamentally horizontal gene transfer (HGT), because the genes are transferred from donors to recipients, rather than vertically from a mother to a daughter cell

Horizontal gene transfer – 2

- Actually, HGT phenomena mostly occur in bacteria, but there are also many eukaryotes that display characteristics known to be associated to horizontal gene transfer (even if HGT is commonly considered rare in eukaryotes and with a very limited evolutionary significance)
- Horizontal gene transfer is made possible in large part by the existence of mobile genetic elements, such as plasmids (extrachromosomal genetic material), transposons ("jumping genes"), and viruses infecting bacteria (bacteriophages)
- These elements are transferred between organisms through different mechanisms, which in prokaryotes include transformation, conjugation, and transduction

Horizontal gene transfer – 3

- In transformation, prokaryotes take up free fragments of DNA, often in the form of plasmids, found in their environment
- In conjugation, genetic material is exchanged during a temporary union between two cells, which may entail the transfer of a plasmid or a transposon
- In transduction, DNA is transmitted from one cell to another via a bacteriophage



TRANSFORMATION

Bacteria acquire new genetic material from the environment

CONJUGATION

It is the process of direct transfer of DNA from one bacterium to another through a protein tube

TRANSDUCTION

Bacteriophage, viruses that infect bacteria pick up genetic material in the process and pass it on to other bacteria

Horizontal gene transfer – 4

- *Streptococcus pneumoniae*, the bacterium that causes pneumonia, won the cover of Science thirteen years ago (January 2011)
- According to a study conducted by the Wellcome Trust Sanger Institute, the strain resistant to antibiotics PNEM1 has undergone many alterations in the genetic code, from the '70s, that have allowed him to resist drugs and vaccines
- By sequencing approximately 240 samples taken in different parts of the world, the researchers were able to reconstruct the evolutionary history of this strain and found that 75% of the genome of PNEM1 has been affected by events of horizontal transfer

Prokaryotic gene density – 1

- The density of prokaryotic genes is very high
- Chromosomes of bacteria and archea completely sequenced indicate that from 85% to 88% of the nucleotides are associated with coding regions
  - Example: *E.coli* contains a total of 4288 genes, with coding sequences which are long, on average, 950 base pairs and separated, on average, from 118 bases
- In addition, prokaryotic genes are not interrupted by introns and are organized in polycistronic transcriptional units (operons)

Prokaryotic gene density – 2

- The number of genes and the genome size – which, in the case of prokaryotes, are “linearly” correlated – reflect the bacterium style of life
  - The specialized parasites have about 500–600 genes, while the generalist bacteria have a much greater number of genes, typically between 4000 and 5000
  - The archea have a number of genes ranging from 1700 to 2900
- A rapid reproduction phase is important for the evolutionary success of bacteria
  - Maximize the coding efficiency of the chromosomes to minimize the time of DNA replication during cell division

Prokaryotic gene density – 3

- Finding a gene in a prokaryotic genome is just a simple task
  - Simple promoter sequences (a small number of factors that support RNA polymerase in the recognition of the promoter sequences placed in –35 and –10)
  - Transcription termination signals simply recognizable (inverted repeats followed by a sequence of uracils)
  - Possible comparison with the nucleotide or amino acid sequences of other well known organisms
- High probability that any randomly chosen nucleotide is associated with the coding sequence or with the promoter of an important gene
- The genome of prokaryotes contains *no wasted space*

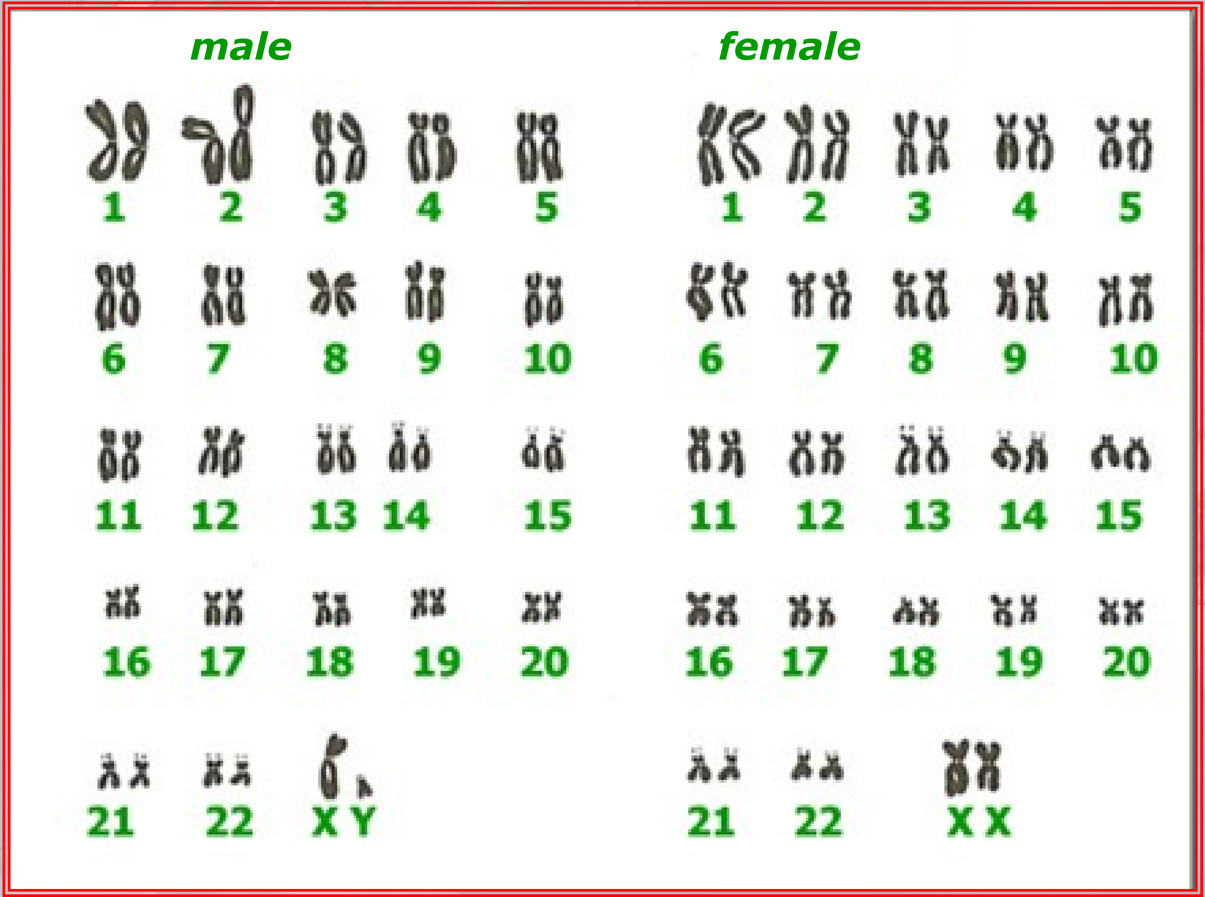
The genome of eukaryotes – 1

- Eukaryotic organisms are much more complex than prokaryotes:
  - The interior compartments surrounded by membranes allow them to maintain a variety of chemically distinct environments within the same cell
  - In contrast to prokaryotes, almost all eukaryotes live as multicellular organisms, and each cell type is usually characterized by a distinctive gene expression pattern, even though every cell of an organism has the same genome
  - Few constraints on the genome size ⇨ eukaryotes contain long sequences of “junk” DNA, which, at the best of our knowledge, look superfluous
  - The eukaryotic genome and the gene expression apparatus, devoted to its interpretation, are much more complex and flexible compared to that of prokaryotes

The genome of eukaryotes – 2

- Completely sequencing and annotating an eukaryotic genome is a difficult undertaking:
  - In contrast to prokaryotes, characterized by a single copy circular chromosome, the nucleus of eukaryotic cells usually contains two copies for each of the (many) linear chromosomes
  - Example
    - Most human cells (apart from egg and sperm cells) have two copies of 22 chromosomes (the autosomal chromosomes) and two sexual chromosomes (two Xs in females, one X and one Y in males)
    - The shortest human chromosome owns 55 million base pairs (55Mbps) whereas the longest one is composed by 250 million base pairs (250Mbps)
    - The total genome length is 3200Mbps

The genome of eukaryotes – 3



The genome of eukaryotes – 4

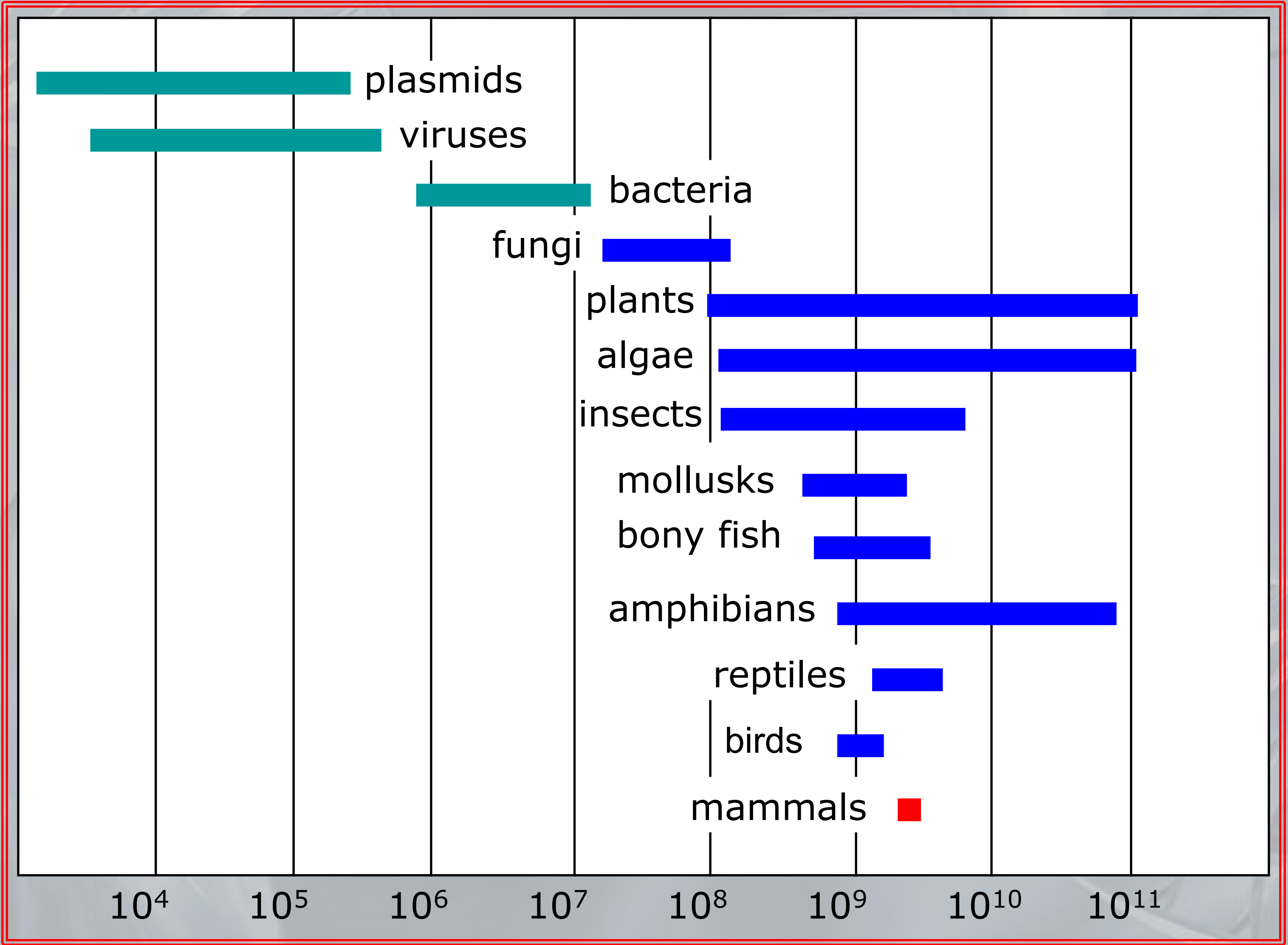
- All the eukaryotic genomes are several magnitude orders longer than those of prokaryotic organisms
- It is worth noting that the total content of DNA in eukaryotes, and therefore the size of the genome, is “weakly” related to the complexity of the organisms (e.g., the human genome is larger than that of insects, which is, in turn, larger than that of fungi)
- However, there are several exceptions: for example, the genome of *X.laevis* is as large as that of mammals; other amphibians have a genome approximately 50 times longer than the human genome; between the plants, the *Zea Mays* genome (5000 Mbps) is larger than that of humans

The genome of eukaryotes – 5

Xenopus laevis: Pipidae family aquatic frog, endemic in Southern Africa

Zea mays: Herbaceous annual plant belonging to the Poaceae family (common maize)

The genome of eukaryotes – 6



The genome of eukaryotes – 7

- In general, for a given taxonomic group, the minimum size of the genome is approximately proportional to the complexity of the organisms belonging to that group
- From a different point of view, the number of cell types present in each organism may constitute a reliable index of its complexity
  - In humans, it is estimated that there are about 400 different types of cells

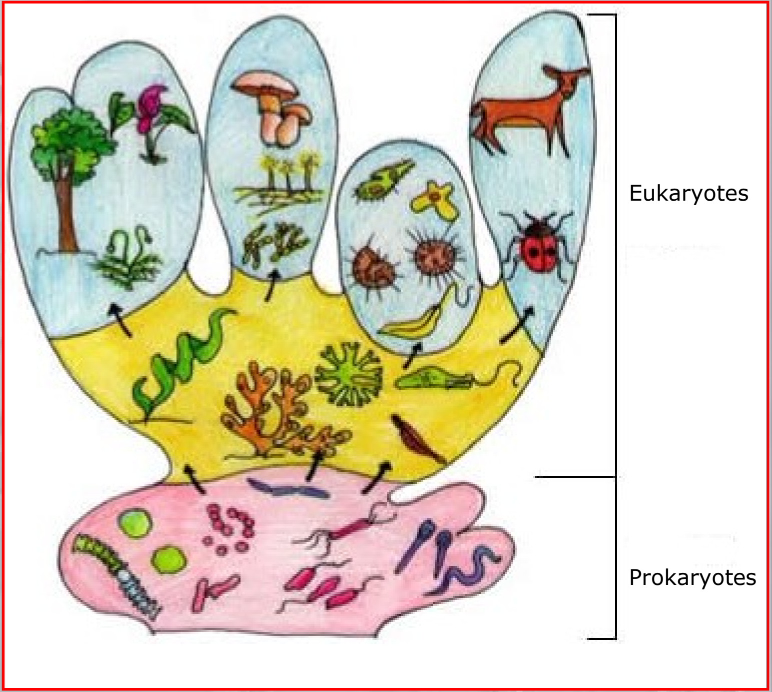
The genome of eukaryotes – 8

- Instead a direct correlation between the size of the genome and the number of chromosomes, or between the number of chromosomes and the complexity of an organism, does not exist
- Finally, a comparison between prokaryotes and eukaryotes with respect to the estimated number of genes is very complicated, because of the difficulty in the prediction of eukaryotic genes, starting from the simple analysis of DNA sequences

The genome of eukaryotes – 9

Prokaryotic organism	Genome length (Mbp)	Number of genes
Mycoplasma genitalium	0.58	470
Helycobacter pylori	1.66	1590
Haemophilus influenzae	1.83	1727
Bacillus subtilis	4.21	4100
Escherichia coli	4.60	4288

Eukaryotic organism	Genome length (Mbp)	Number of genes
Saccharomyces cerevisiae (yeast)	13.5	6241
Caernohabditis elegans (worm)	100	18424
Arabidopsis thaliana (thale cress)	130	25000
Drosophila melanogaster (fruit fly)	180	13601
Danio rerio (zebrafish)	1700	n.d.
Homo sapiens (man)	3200	45000



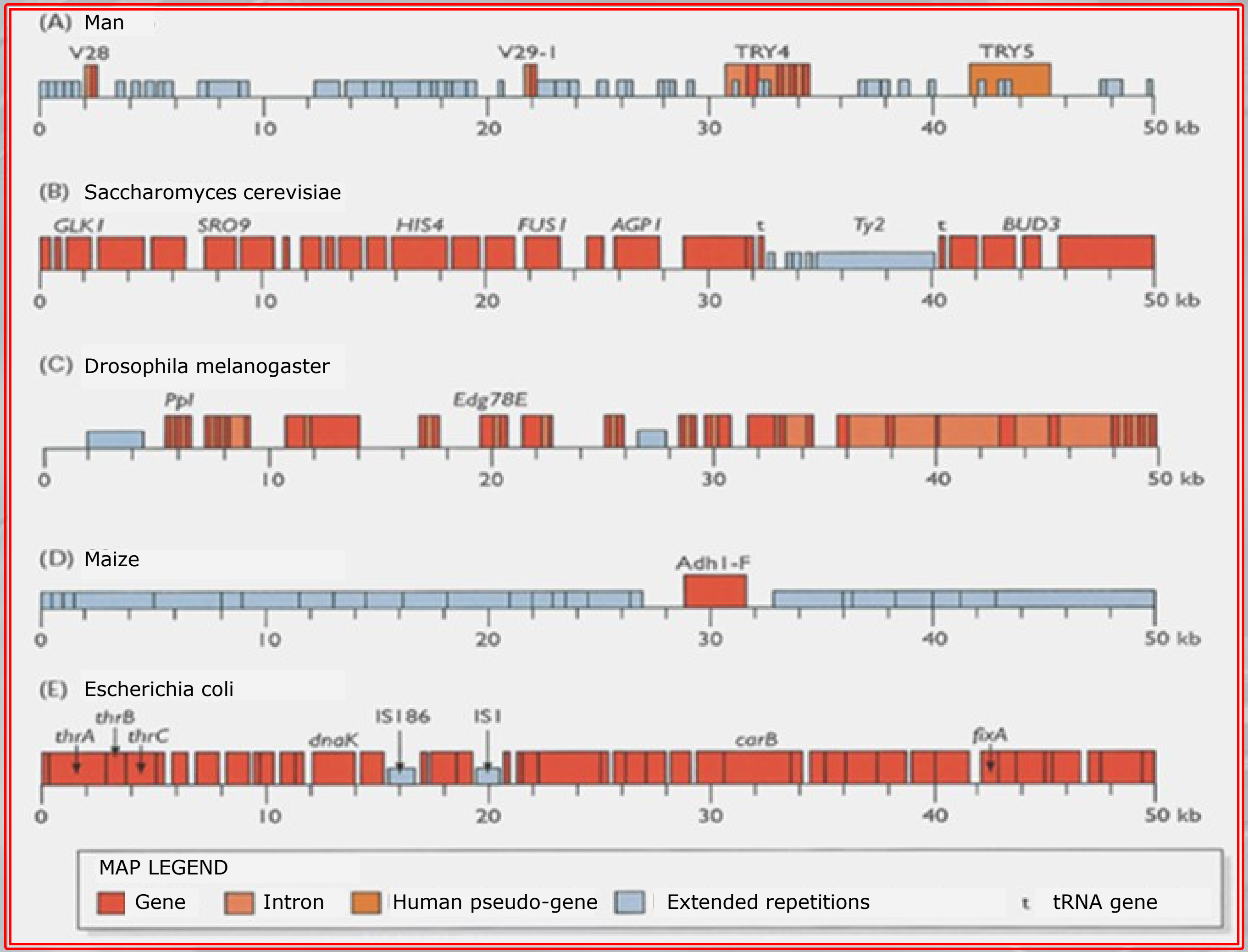
Eukaryotic gene structure – 1

- By definition, among the most difficult search problems, there is the classic “find a needle in a haystack”
- This old analogy is far from being sufficient to give an idea of the complexity of finding eukaryotic genes inside the huge amount of sequence data
  - Actually, finding a needle of 2 grams inside 6000 kilos of straw is thousand times easier than finding a gene in the eukaryotic genome, even assuming that this gene is as different from the rest of the DNA as an iron needle is from a piece of straw

Eukaryotic gene structure – 2

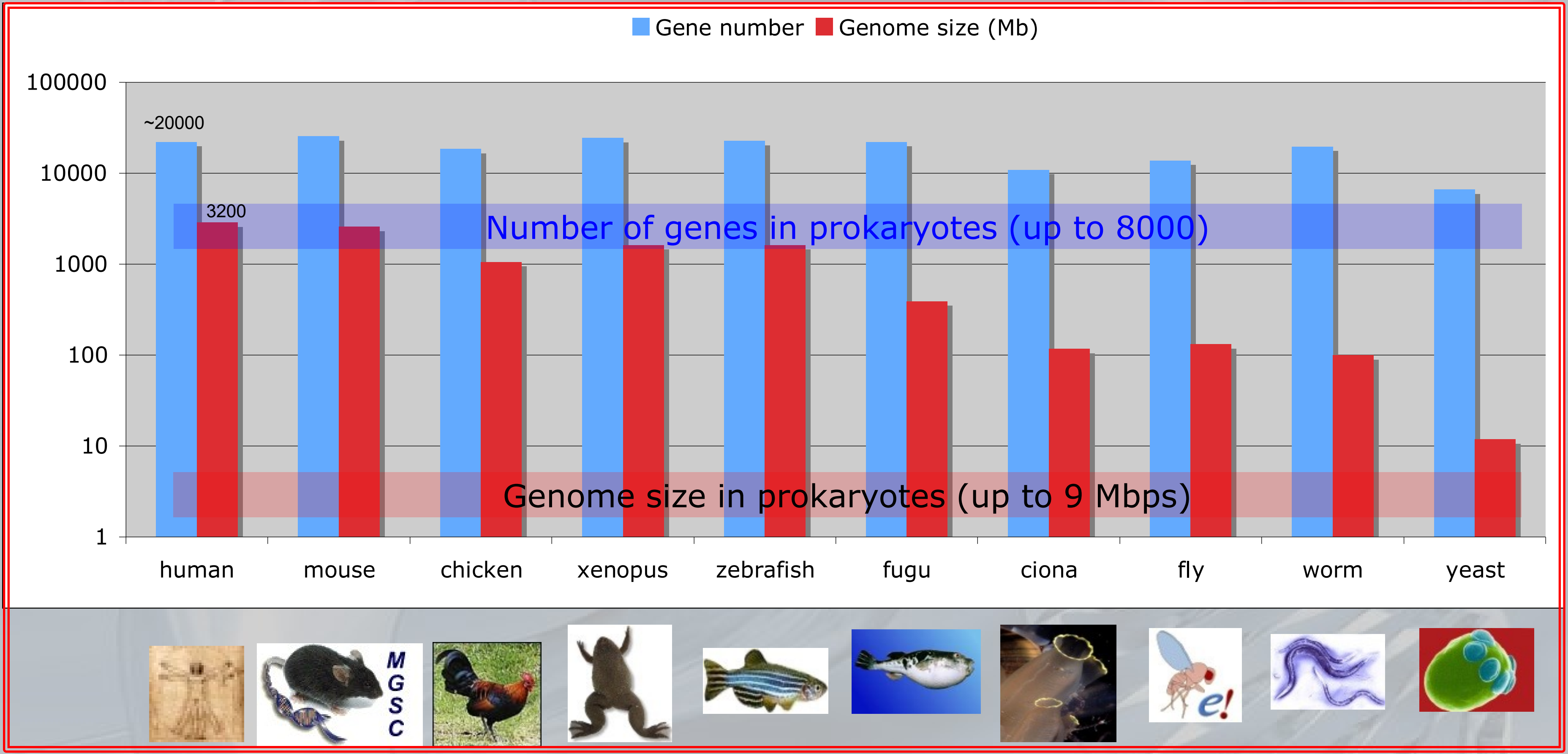
- In fact,…
  - Eukaryotic genomes have a very low gene density: on average, the protein–coding genes occupy only 2–4% of the entire genome
  - The peculiarities of prokaryotic ORFs, with their statistically significant lengths, are not found in eukaryotic genes, due to the abundant presence of introns (which in mammals can reach sizes around 20–30 Kbps) and repeated elements
  - Eukaryotic promoters, like their prokaryotic counterparts, contain, in their sequence, some preserved features, that can be used as reference points in search algorithms
    - However, such sequences tend to be much more dispersed and positioned at a great distance from the transcription start site

Eukaryotic gene structure – 3



Comparison among human, yeast, fruit fly, maize, and *E.coli* genomes

Eukaryotic gene structure – 4



Number of genes in prokaryotes (up to 8000)

Genome size in prokaryotes (up to 9 Mbps)

The absence of correlation between the number of genes and the genome size in eukaryotes

Eukaryotic gene structure – 5

- The problem of recognizing eukaryotic genes in sequence data is therefore a great challenge, which promises to remain such for some future decades
- So far, the best attempts to solve the problem are based on the use of pattern recognition techniques (such as neural networks and Generalized Hidden Markov Models) and on dynamic programming
  - Free software is available on the Web — such as GrailEXP or GenScan (http://argonaute.mit.edu/GENSCAN.html) — which, however, shows only acceptable performance

Eukaryotic gene structure – 6

- All the algorithms for the recognition of genes scan the DNA sequence to search particular nucleotide strings, having *ad hoc* orientations and relative positions
- Any feature, in itself, could be detected at random, but the simultaneous presence of more “markers”, such as possible promoters, sequences that indicate the vicinity of introns and exons, and a putative ORF with codons not uniformly distributed, increases the probability that a given region corresponds to a gene

Promoter’s elements – 1

- All the information needed by a liver cell are also present in muscle or brain cells
- The gene expression regulation is the only mechanism by which their differences are taken into account and, as in the case of prokaryotes, the transcription start point is fundamental for efficiently regulating the gene expression
  - Eukaryotes follow complex strategies in order to adjust the transcription phase
  - Unlike prokaryotes, which have a single RNA polymerase, constituted by few protein subunits, all eukaryotic organisms use three different types of RNA polymerase, consisting of a minimum of 8 up to 12 proteins

Promoter’s elements – 2

- Each eukaryotic RNA polymerase recognizes a different set of (core) promoters and it is used to transcribe different types of genes

RNA polimerase	Promoter position	Promoter complexity	Transcribed genes
RNA polimerase I	From –45 to +20	Simple	Ribosomal RNA
RNA polimerase II	Upstream w.r.t. –25	Very complex	Protein–coding genes
RNA polimerase III	From +50 to +100	Simple	tRNA (transfer RNA) and other small RNAs

Promoter’s elements – 3

- RNA polymerases I and III construct RNA molecules that are functionally important (and must be maintained at constant levels) in all eukaryotic cells and in every moment
- RNA polymerase II is responsible solely for the transcription of eukaryotic genes that encode for proteins
- The variety of promoter sequences recognized by RNA polymerase II reflects the complexity of the distinction between genes that should or should not be expressed at a given time and for a given cell type

Promoter’s elements – 4

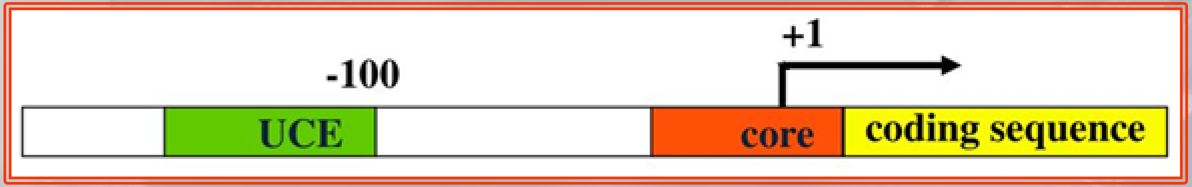
- As in prokaryotes, also in eukaryotes, the term promoter is used to describe all the sequences that are important for the initiation of the gene transcription
- Unlike prokaryotic operons, where multiple genes share a single promoter, in eukaryotes, each gene has its own promoter
- Many promoters, recognized by RNA polymerase II, contain a set of sequences, known as basal or core promoter, around which an initiation set of RNA polymerase II is concentrated, and from which transcription begins

Promoter’s elements – 5

- The promoters of most genes transcribed by RNA polymerase II also include several upstream promoter elements, to which some proteins, different from RNA polymerase II, bind in a specific manner
- Considering the number of genes and the different types of cells present in eukaryotes, it has been estimated a minimum of five upstream promoter elements required to uniquely identify a particular gene, and ensure that it is expressed in an appropriate manner
- If the proteins, that recognize the upstream promoter elements, do not bind correctly, the transcription process can become inefficient

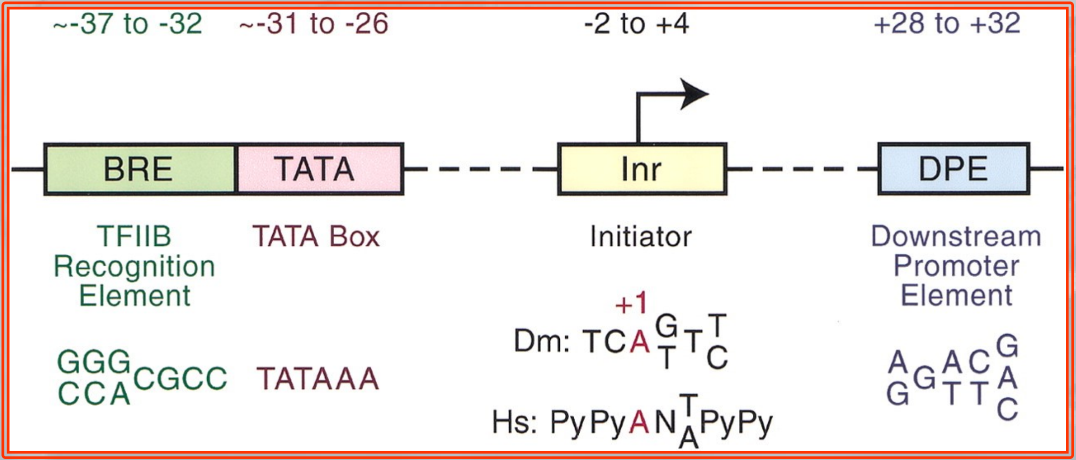
Promoter's elements – 6

- In detail… Each of the three RNA polymerases (RNAPs) recognizes different eukaryotic promoter sequences; in fact, it is just the difference between the promoters that defines which genes will be transcribed and which polymerase will be implicated
- In particular, in vertebrates:
  - The RNAP I promoters are constituted by a core promoter which, with respect to the point of the transcription initiation, can be found between nucleotides –45 and +20, and by a control element, about 100 bases upstream (*upstream control element*)



Promoter’s elements – 7

- The RNAP II promoters are variable and can extend for some kilobases upstream (and most rarely downstream) w.r.t. the start site of transcription
  - The core promoter consists of two segments: the region located at –25, called the TATA box (consensus sequence 5′–**TATAWAW**–3′, **W**=**A**/**T**), and the initiator sequence, Inr (consensus 5′–**YYCARR**–3′, **Y**=**C**/**T**, **R**=**A**/**G**) in position +1 (transcription starting site)
  - The nucleotide in +1 is, almost always, an **A**, very conserved in the Inr sequence

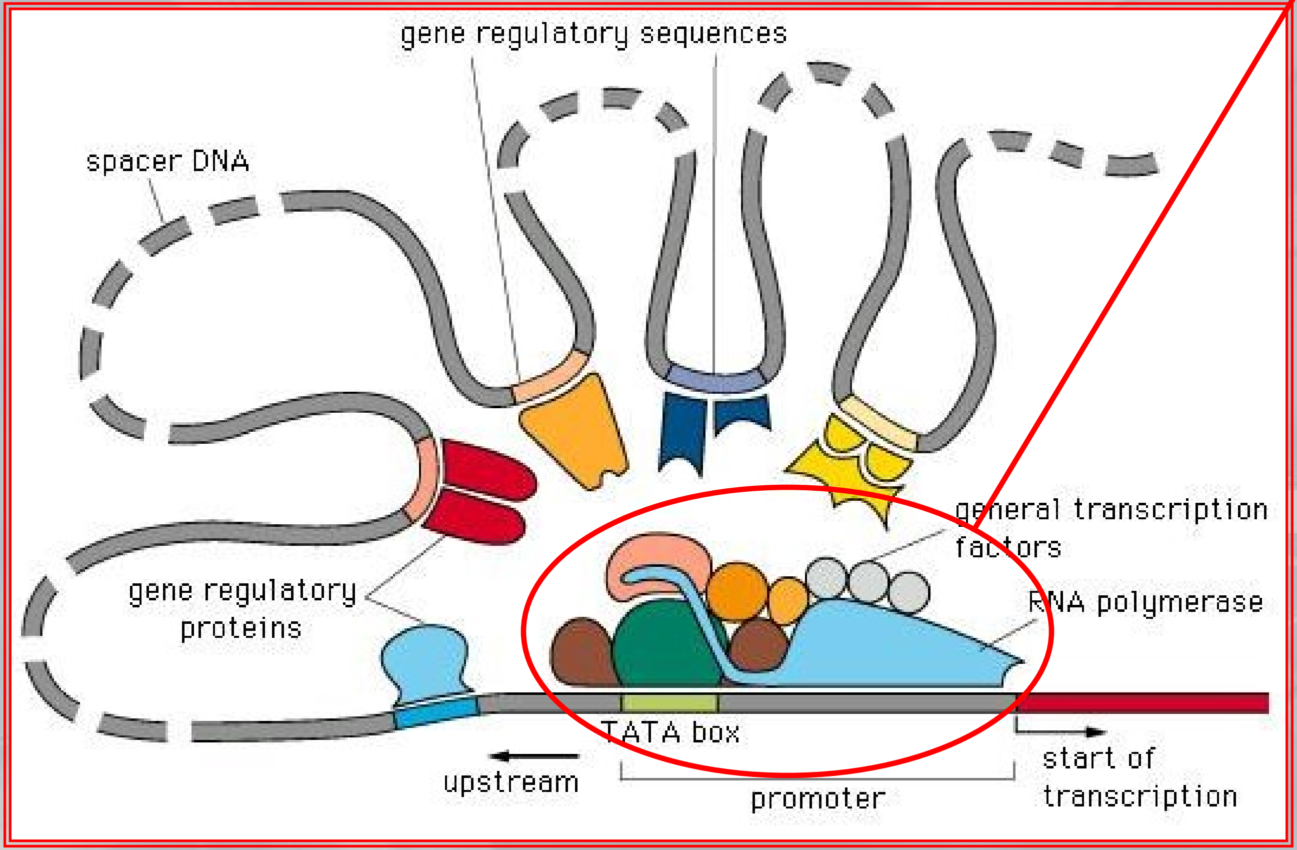


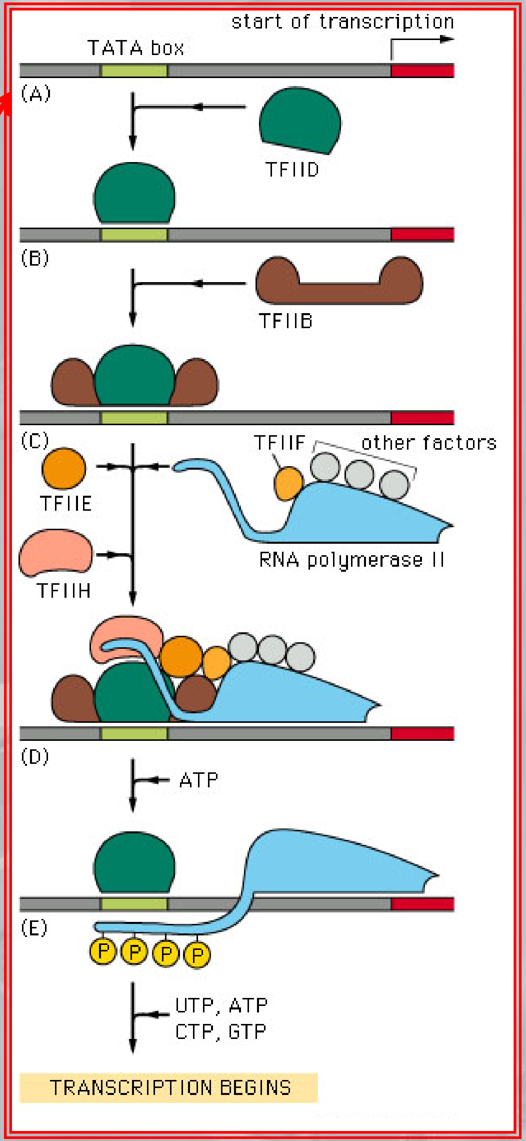
Promoter’s elements – 8

- Actually, the RNAP II does not directly recognize the core promoter, which is first attached by the basal transcription factors — composed by a TATA–binding protein (TBP) and by, at least, 12 TBP–associated factors (TAF)
- The basal transcription factors bind the core promoter sequences, preparing the chemical environment in which the catalytic unit of RNAP II can work
- In addition to the core promoter, the genes recognized by RNAP II have different upstream promoter elements recognized by external transcription factors

Promoter’s elements – 9

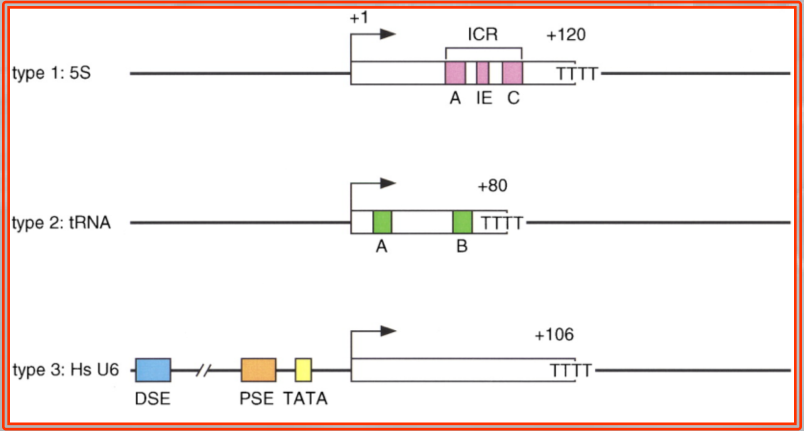
Basal transcription factors binding to the core promoter sequences, preparing the chemical environment for RNAP II activation





Promoter's elements – 10

- The RNAP III promoters are variable and belong to at least three categories, two of which contain fundamental sequences localized within their promoted genes
  - These sequences typically extend for about 50–100 bases and comprise (at least) two conserved regions separated by variable regions
  - The other category of class III promoters is very similar to the promoters of the RNA polymerase II, having the TATA box and a series of additional upstream promoter elements



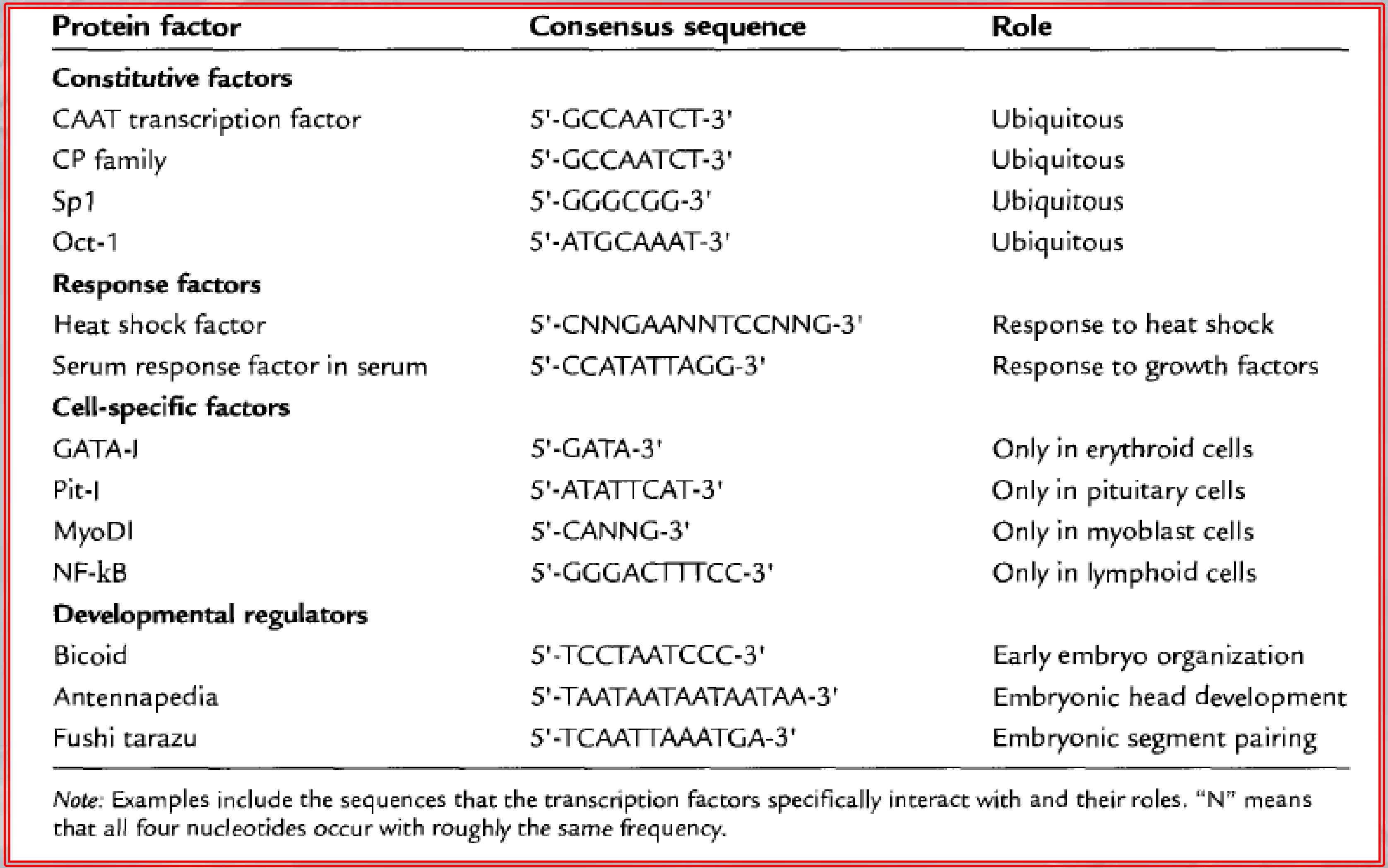
Binding sites of regulatory proteins – 1

- The transcription initiation in eukaryotes is very different from that of prokaryotic organisms
- In bacteria, RNA polymerases have a high affinity for their promoters and the negative regulation, realized by proteins that prevent the gene expression at inappropriate times (such as that made by pLacI), assumes a particular importance
- In eukaryotes, RNA polymerases II and III do not assemble around their promoters efficiently, and the speed of transcription is very low, regardless of how well a promoter corresponds to the expected consensus sequence
- The presence of additional proteins that act as positive regulators is fundamental

Binding sites of regulatory proteins – 2

- Some positive regulators are essentially constitutive, i.e. they operate on many different genes and do not seem to respond to external signals
- Other proteins instead are aimed at regulating the expression of a limited number of genes and respond to environmental signals (they are sometimes called transcription factors)
- Most of regulatory proteins are able to bind only to specific DNA sequences

Binding sites of regulatory proteins – 3



Protein factor	Consensus sequence	Role
Constitutive factors		
CAAT transcription factor	5'-GCCAATCT-3'	Ubiquitous
CP family	5'-GCCAATCT-3'	Ubiquitous
Sp1	5'-GGGCGG-3'	Ubiquitous
Oct-1	5'-ATGCAAAT-3'	Ubiquitous
Response factors		
Heat shock factor	5'-CNNGAANNTCCNNG-3'	Response to heat shock
Serum response factor in serum	5'-CCATATTAGG-3'	Response to growth factors
Cell-specific factors		
GATA-I	5'-GATA-3'	Only in erythroid cells
Pit-I	5'-ATATTCAT-3'	Only in pituitary cells
MyoDI	5'-CANNG-3'	Only in myoblast cells
NF-kB	5'-GGGACTTTCC-3'	Only in lymphoid cells
Developmental regulators		
Bicoid	5'-TCCTAATCCC-3'	Early embryo organization
Antennapedia	5'-TAATAATAATAATAA-3'	Embryonic head development
Fushi tarazu	5'-TCAATTAAATGA-3'	Embryonic segment pairing

Note: Examples include the sequences that the transcription factors specifically interact with and their roles. “N” means that all four nucleotides occur with roughly the same frequency.

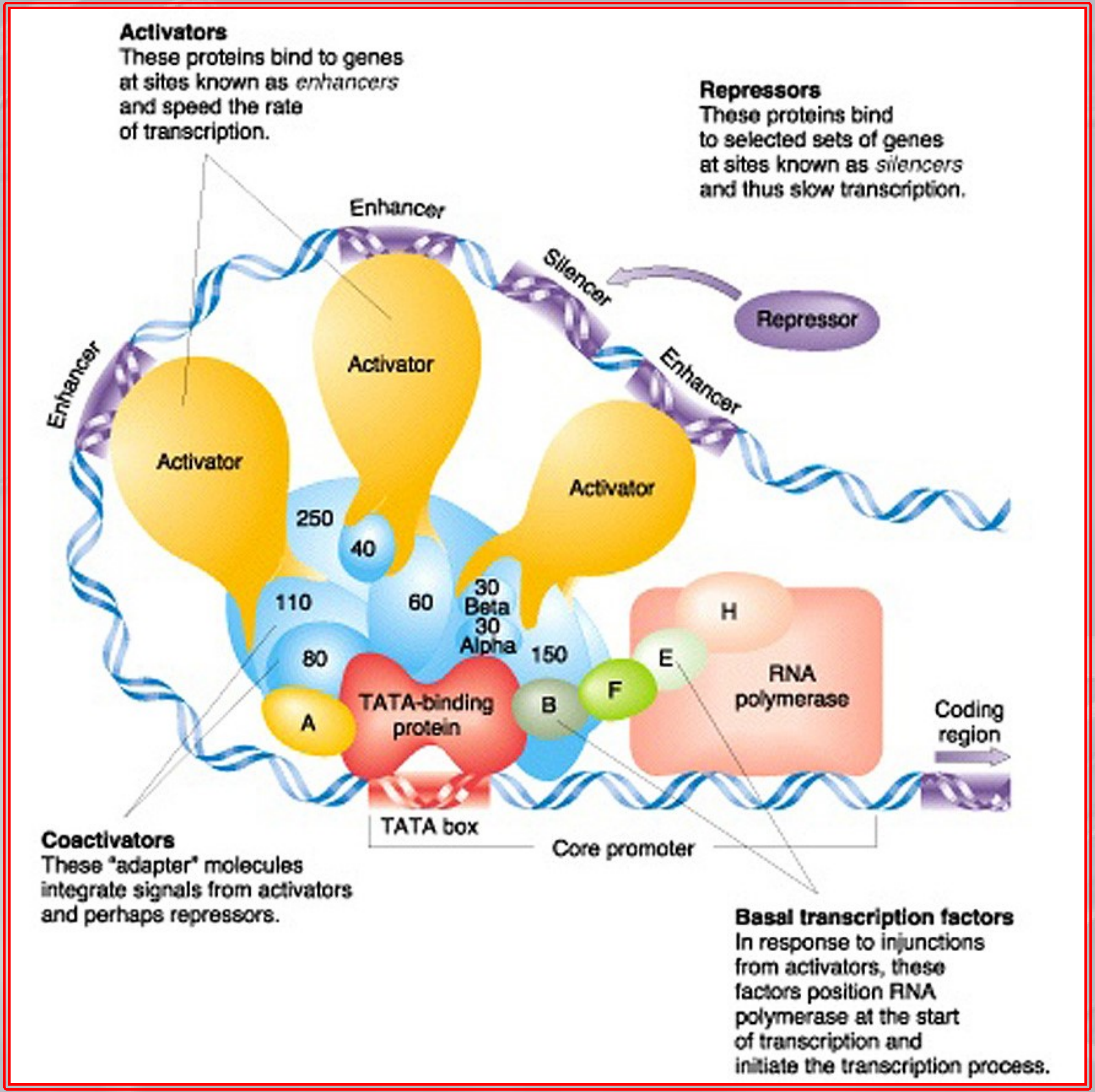
Binding sites of regulatory proteins – 4

- Examples
  - The transcription factor CAAT and the family of CP proteins recognize consensus sequences relatively close to the transcription initiation sites, such as the CAAT box, located at the position –80, in most eukaryotic genes
  - CAAT and CP are constitutive factors, that is they are not related to the expression of specific genes

Binding sites of regulatory proteins – 5

- Examples (cont.)
  - The Sp1 transcription factor binds to the so–called enhancers ("amps"), short DNA regions that increase the transcription levels of the genes for both the orientations and over a wide range with respect to the start site (from –500 to +500)
  - The eukaryotic enhancers work also at several tens of thousands nucleotides upstream of the start site of transcription, and perform their function by bending the DNA into a specific shape that brings the transcription factors in contact, to form structures called enhanceosoms

Binding sites of regulatory proteins – 6



Activators
These proteins bind to genes at sites known as *enhancers* and speed the rate of transcription.

Repressors
These proteins bind to selected sets of genes at sites known as *silencers* and thus slow transcription.

Coactivators
These "adapter" molecules integrate signals from activators and perhaps repressors.

Basal transcription factors
In response to injunctions from activators, these factors position RNA polymerase at the start of transcription and initiate the transcription process.

Binding sites of regulatory proteins – 7

- Examples (cont.)

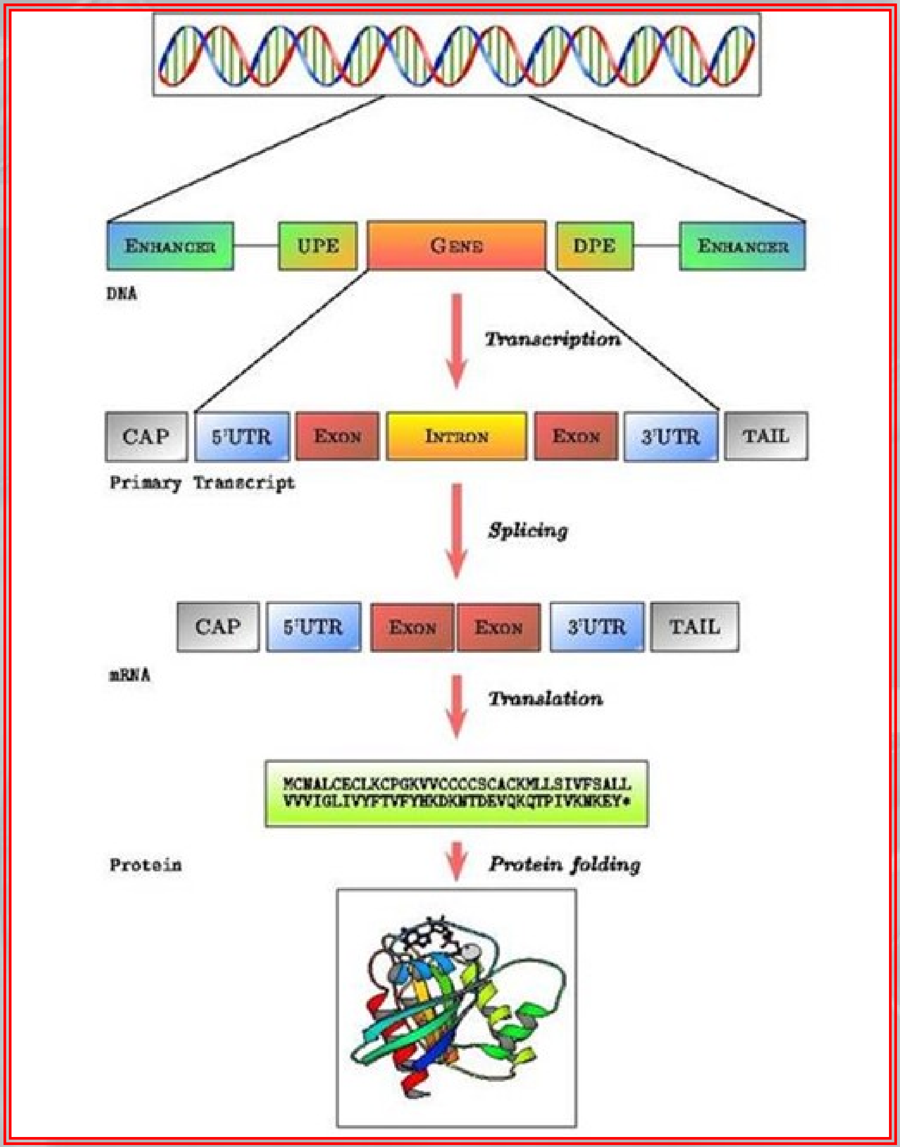
  Several regulatory proteins are activated only in special circumstances and help to mediate the cell response to external environmental stimuli, such as exposure to heat, or allow genes to be expressed only in specific tissues or in particular life stages

  - Heat shock factor, response to a sudden increase in temperature
  - GATA–I, present only in erythroid cells (precursors of red blood cells)
  - Antennapedia, controls the embryonic development of the head

Open reading frames – 1

- The nuclear membrane of eukaryotic cells is a physical barrier that separates the processes of transcription and translation
  - In prokaryotes, this barrier is not present, and the process of translation by the ribosomes starts as soon as the RNA polymerase has started to produce an RNA copy of a coding region
- Eukaryotes benefit from the delay of the translation phase – needed for transporting the RNA out of the nucleus – to change significantly the primary transcript (produced by the RNA polymerase II)

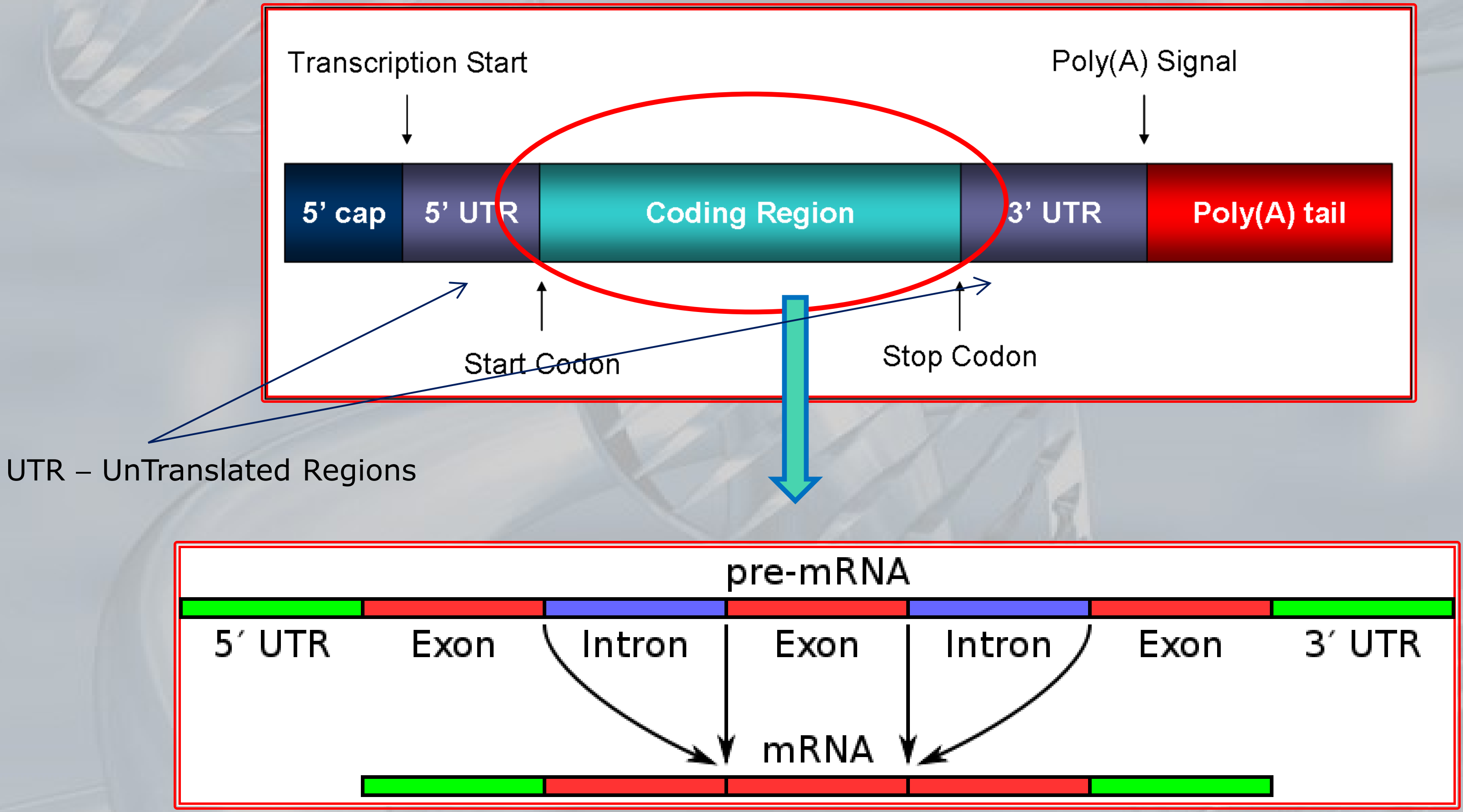
Open reading frames – 2



Open reading frames – 3

- Known as the primary transcript or as hnRNA (for heterogeneous nuclear RNA), before being translated, the transcript of the RNA polymerase II undergoes several changes
  - Capping (addition of a “hood”): It consists of a set of chemical alterations (including methylation) that happen at the 5′ end of all hnRNAs
  - Splicing (exon junction): It provides the total and precise removal of (sometimes very long) segments inside the hnRNA
  - Polyadenylation (to transform hnRNA into mRNA, usable by ribosomes): It is the process of replacing the 3′ end of a hnRNA with a sequence of about 250 adenines, that are not present in the nucleotide sequence of the gene

Open reading frames – 4



Open reading frames – 5

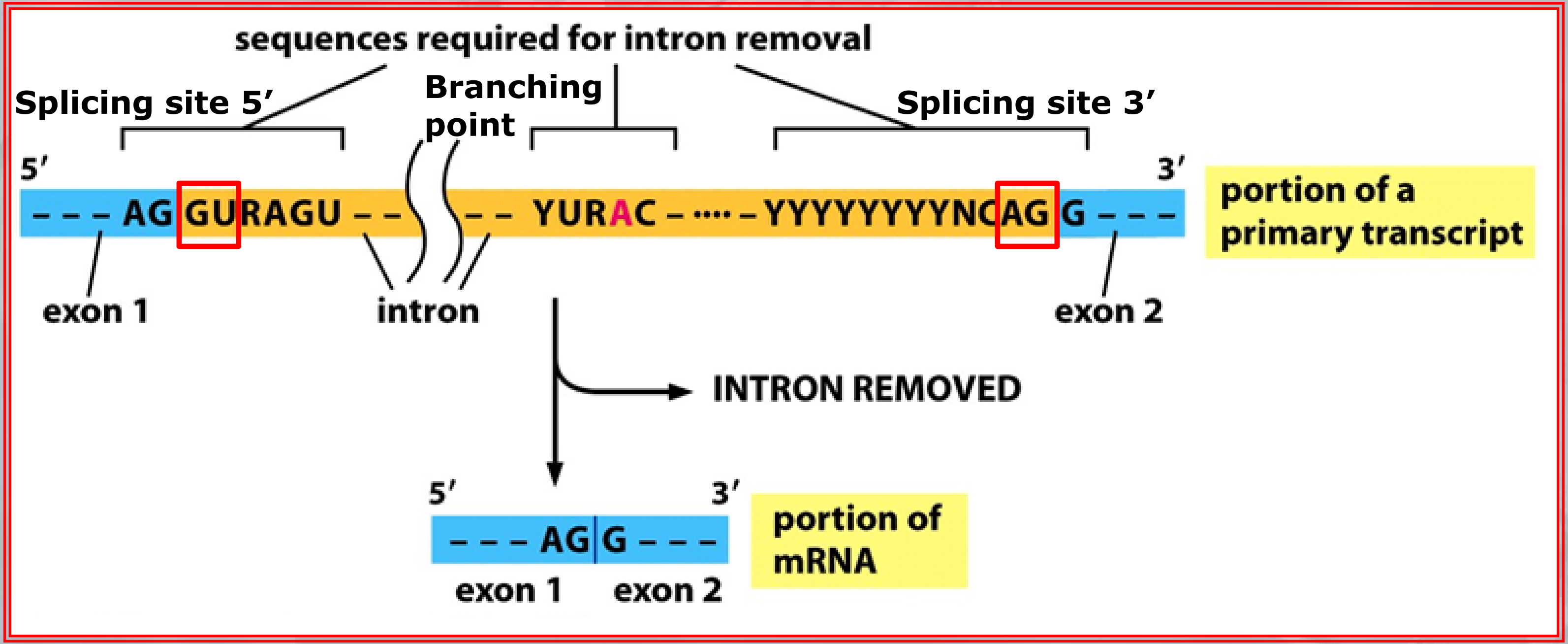
- Each of the three possible modifications may occur differently in different types of cells
- In particular, splicing differentiations allow eukaryotic organisms to meet the demand of tissue–specific gene expression, without paying a high price in terms of genomic complexity
  - Considerable difficulties for gene recognition algorithms due the actual inability of automatically modeling the splicing process

Introns and exons – 1

- The genetic code was experimentally deciphered – in the sense that the correspondence between triplets of nucleotides and amino acids was known – long before the nucleotide sequence of genes was determined
- Therefore, it was really surprising when, in 1977, the first eukaryotic genomic sequence was obtained and it was discovered that many genes contain intervening sequences, called introns, interrupting the coding regions, which were recombined into the mature RNA
- Since then, in eukaryotic cells, at least eight different types of introns have been identified, although only one of these, which follows the rule **GU–AG**, is mainly associated with eukaryotic genes that encode for proteins (those transcribed by RNAP II)

Introns and exons – 2

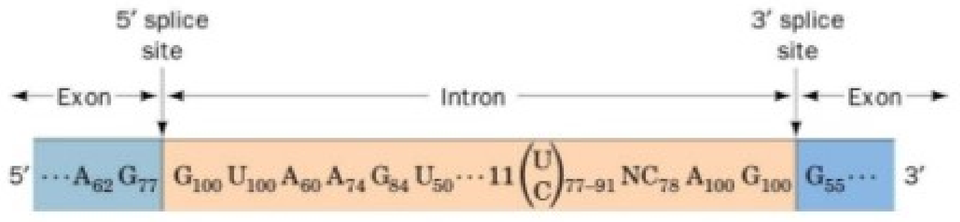
- The **GU–AG** rule takes its name from the fact that the first pair of nucleotides, located at the 5′ end of the RNA sequence of the introns of this type, is always 5′–**GU**–3′, while the last two nucleotides, at the 3′ end, are always 5′–**AG**–3′



Introns and exons – 3

- Some additional nucleotides associated with the splicing junctions located in 5′ and 3′ (~6), as well as an internal "branching point", located from 18 to 40 base pairs upstream of the splicing junction 3′, are all the "markers" for the splicing apparatus
  - Most of the sequences to be examined to realize the splicing process lies within the intron, not involving the information content of the sequences coding for the flanking exons, which will be reconnected to form the messenger RNA

The numbers written in the subscript of each consensus nucleotide indicate the frequency of its presence in known vertebrate introns


5′ splice site
3′ splice site
Exon
Intron
Exon
5′ $\cdots A_{62} G_{77}$ $G_{100} U_{100} A_{60} A_{74} G_{84} U_{50} \cdots 11 \binom{U}{C}_{77-91} NC_{78} A_{100} G_{100}$ $G_{55} \cdots$ 3′

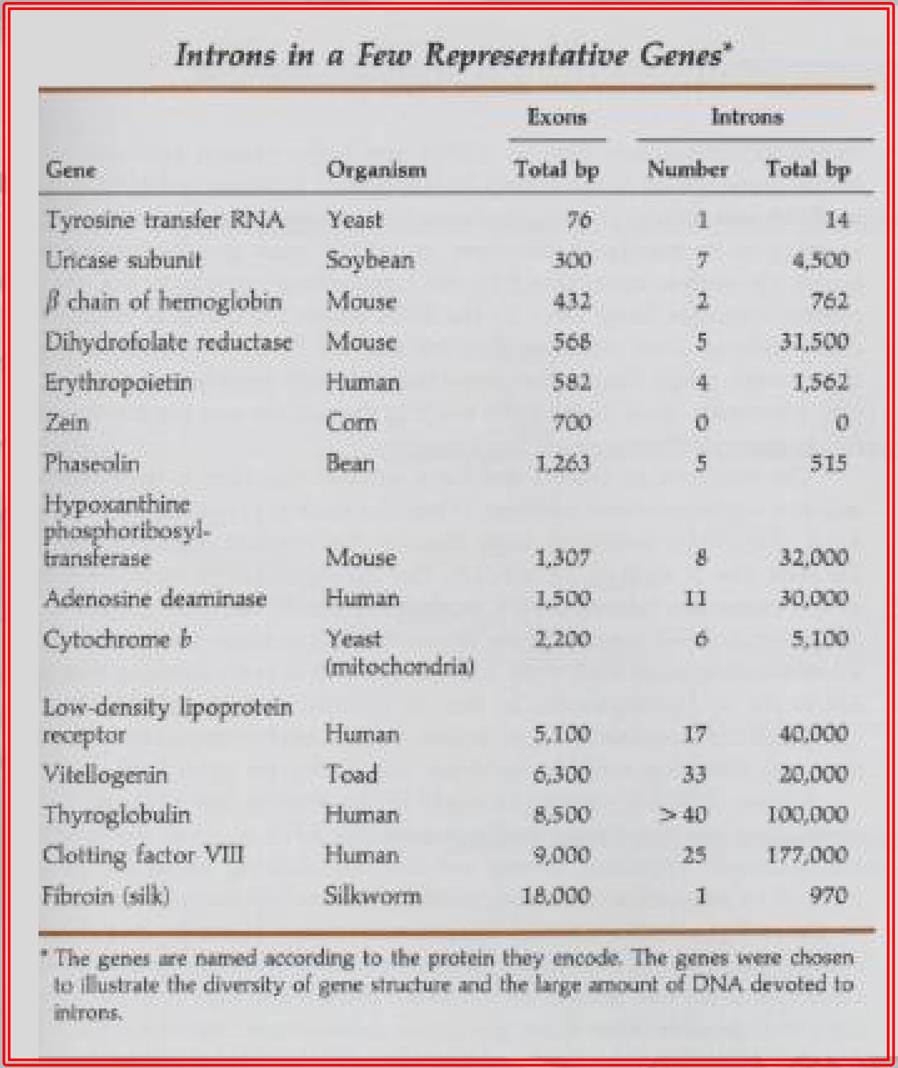
The consensus sequence at the exon–intron junctions of vertebrate pre-mRNAs

Introns and exons – 4

- Introns usually have a minimum length of about 60 base pairs (necessary to maintain the splicing signals), but there are no predetermined upper limits to their length
  - Example: Human introns can be long tens of thousands nucleotides
- Similarly, the average exon length is 450 bps, but there are very short (less than 100 bps) and very long (over 2000 bps) exons
- The intron distribution appears not to be governed by rigid rules, even if they are not common in the simpler eukaryotes
  - Example: Within the 6000 genes of the yeast genome there are only 239 introns

Introns and exons – 5

- Conversely, introns are widespread in the genes of vertebrates, and ~95% of human genes contains at least one intron (while, sometimes, more than 100 introns may be contained in a unique gene)
- Exons constitute about 2% of our genome, whereas introns represent almost 25%



*Introns in a Few Representative Genes**

		Exons	Introns	
Gene	Organism	Total bp	Number	Total bp
Tyrosine transfer RNA	Yeast	76	1	14
Uricase subunit	Soybean	300	7	4,500
β chain of hemoglobin	Mouse	432	2	762
Dihydrofolate reductase	Mouse	568	5	31,500
Erythropoietin	Human	582	4	1,562
Zein	Corn	700	0	0
Phaseolin	Bean	1,263	5	515
Hypoxanthine phosphoribosyl-transferase	Mouse	1,307	8	32,000
Adenosine deaminase	Human	1,500	11	30,000
Cytochrome *b*	Yeast (mitochondria)	2,200	6	5,100
Low-density lipoprotein receptor	Human	5,100	17	40,000
Vitellogenin	Toad	6,300	33	20,000
Thyroglobulin	Human	8,500	>40	100,000
Clotting factor VIII	Human	9,000	25	177,000
Fibroin (silk)	Silkworm	18,000	1	970

* The genes are named according to the protein they encode. The genes were chosen to illustrate the diversity of gene structure and the large amount of DNA devoted to introns.

Introns and exons – 6

- Apart from the splicing signal sequences, the length of introns and their nucleotide composition appear to be subject to weak selective constraints
- On the contrary, the position of the introns within the genes appears to be conserved from an evolutionary point of view, in the sense that they often occupy identical positions in homologous genes

Alternative splicing – 1

- The primary transcripts of RNA polymerase II, the hnRNAs, before being translocated into the cytoplasm, where they are translated, undergo a series of changes, the most notably of which is the removal of introns (via the splicing process)
- For some messengers, splicing can take place in alternative ways
- The alternative splicing can generate, from a single gene, different mature transcripts and, therefore, distinct protein isoforms

Alternative splicing – 2

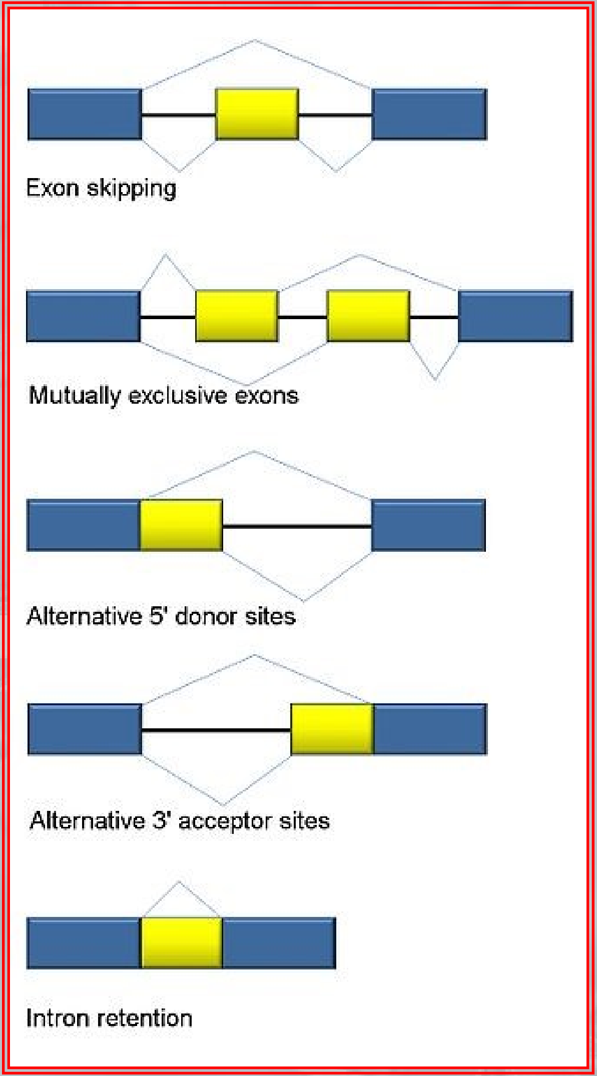
- All the splicing junctions at 5’, as well as those in 3’, appear indeed functionally equivalent for the splicing apparatus
- Furthermore, in normal circumstances, splicing occurs only between sites 5’ and 3’ of the same intron
- In fact, it seems that some eukaryotic genes are transcribed into a single mRNA, i.e., introns and exons are recognized in the same way in all the cell types

Alternative splicing – 3

- However, it was estimated that about 90% of human genes gives rise to more than one type of mRNA
  - In an extreme case, it was found that a single human gene has generated up to 64 different mRNAs by the same primary transcript…
  - ...while it was established that human genes encode, on average, for four/five different proteins
- The alternative splicing is, therefore, a versatile mechanism for the gene expression regulation at the post–transcriptional level
- The alternative splicing (partially) explains why, in the most complex forms of life, a linear relationship between the number of genes and the complexity of the organism does not exist

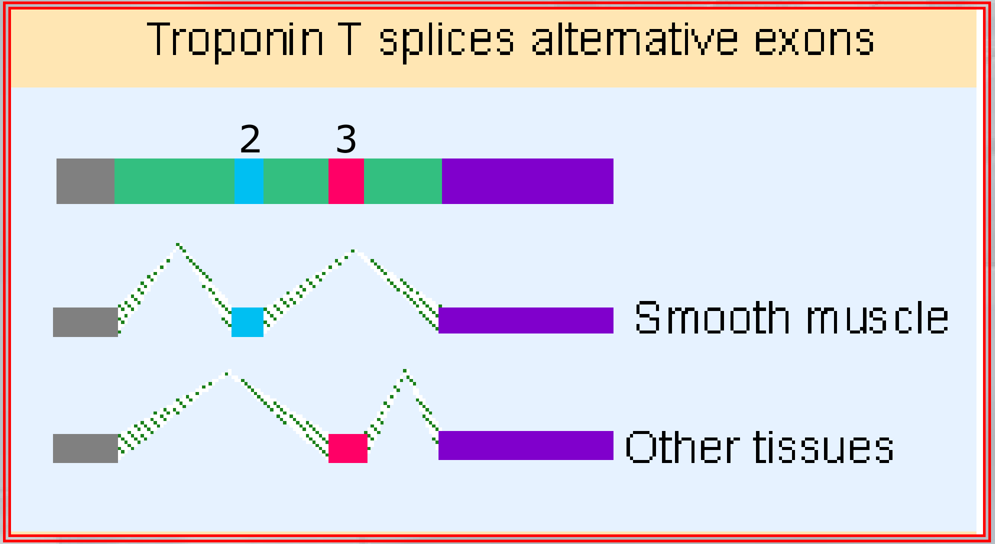
Alternative splicing – 4

- There are five different modes in which alternative splicing occurs
  - Exon skipping: In this case an exon can be eliminated from the primary transcript (very common in mammals)
  - Mutually exclusive exons: Only one, out of two exons, is maintained in the mature mRNA
  - Alternative cutting site in 5’: An alternative cutting site is used at 5’, changing the 3’ termination of the upstream exon
  - Alternative cutting site in 3’: An alternative cutting site is used at 3’, changing the 5’ termination of the downstream exon
  - Intron retention: The cutting sites of an intron may not be recognized; in this case, the intron is not deleted from the mRNA transcript



Alternative splicing – 5

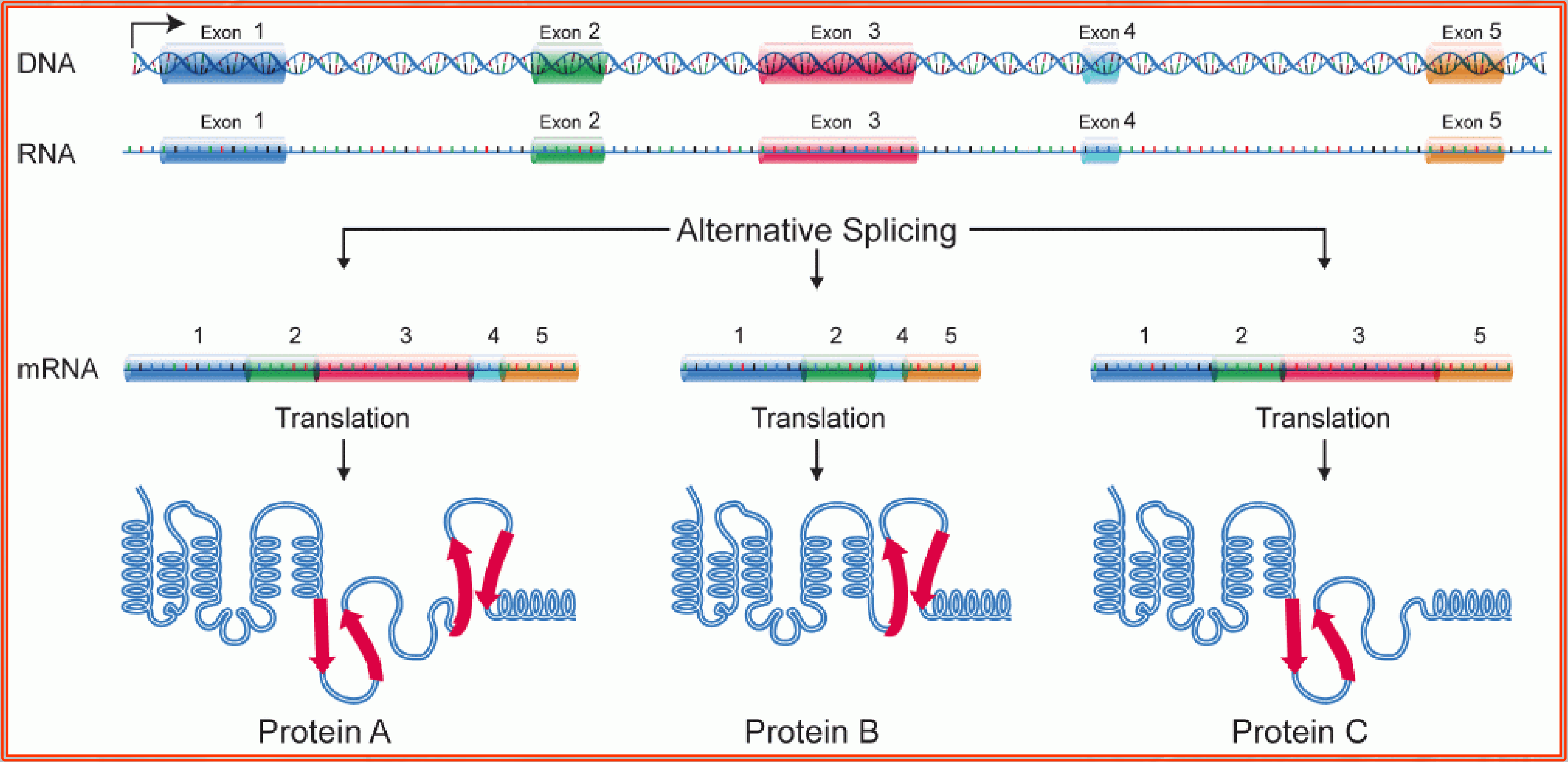
- Example: Exons 2 and 3 of the troponin T gene are mutually exclusive
  - Exon 2 is used in the smooth muscle
  - Exon 3 is used in all other tissues



The smooth muscle cells possess a protein which binds repeated sequences present on both sides of the exon 3 of the hnRNA and, apparently, masks the splice junctions useful to recognize the exon and include it in the mRNA

Troponin T is an integral protein that contribute to the contraction of skeletal and heart muscles; it is expressed in skeletal and cardiac myocites

Alternative splicing – 6



Alternative splicing – 7

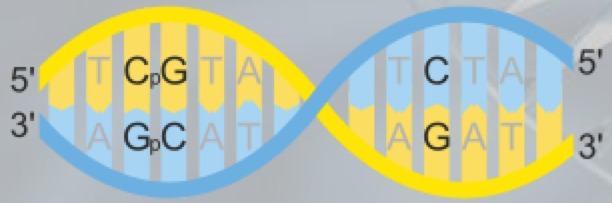
- In recent years, the importance of understanding the alternative splicing mechanism has increased, based on the discovery that at least 15% of genetic diseases is caused by aberrant splicing events, often induced by mutations that alter the efficiency with which a certain exon is recognized and mounted on the mature messenger RNA
- In addition, it has become increasingly clear that the deregulation of the alternative splicing in some genes is accompanied by the appearance of a tumor phenotype and, in some cases, by the tumor’s ability to form metastases
- The recent isolation of proteins and factors involved in the splicing reaction opens the possibility of giving a description, up to now missing, of the deregulation that occurs in tumors

GC content in eukaryotic genomes

- The total GC content of the genome does not have the same variability among eukaryotic species, so as in prokaryotes
- However, it seems to play a very significant role in gene recognition algorithms, because:
  - eukaryotic ORFs are much more difficult to recognize
  - the large–scale variation of GC content within eukaryotic genomes is the basis for useful correlations between genes and upstream promoter sequences, for the choice of codons, the length of genes and their density

CpG islands – 1

- One of the oldest bioinformatic analysis carried out on DNA data was the statistical evaluation of the frequency of all possible pairs of nucleotides in generic sequences extracted from the human genome
- It was observed that the CG dinucleotide – often called CpG to highlight the phosphodiester bond that connects the two nucleotides – appears with a frequency equal to 20% of what it should be detected if each dinucleotide (on the single–strand DNA) should appear with an equal probability
- No other dinucleotide presents a so unusual over/under–distribution



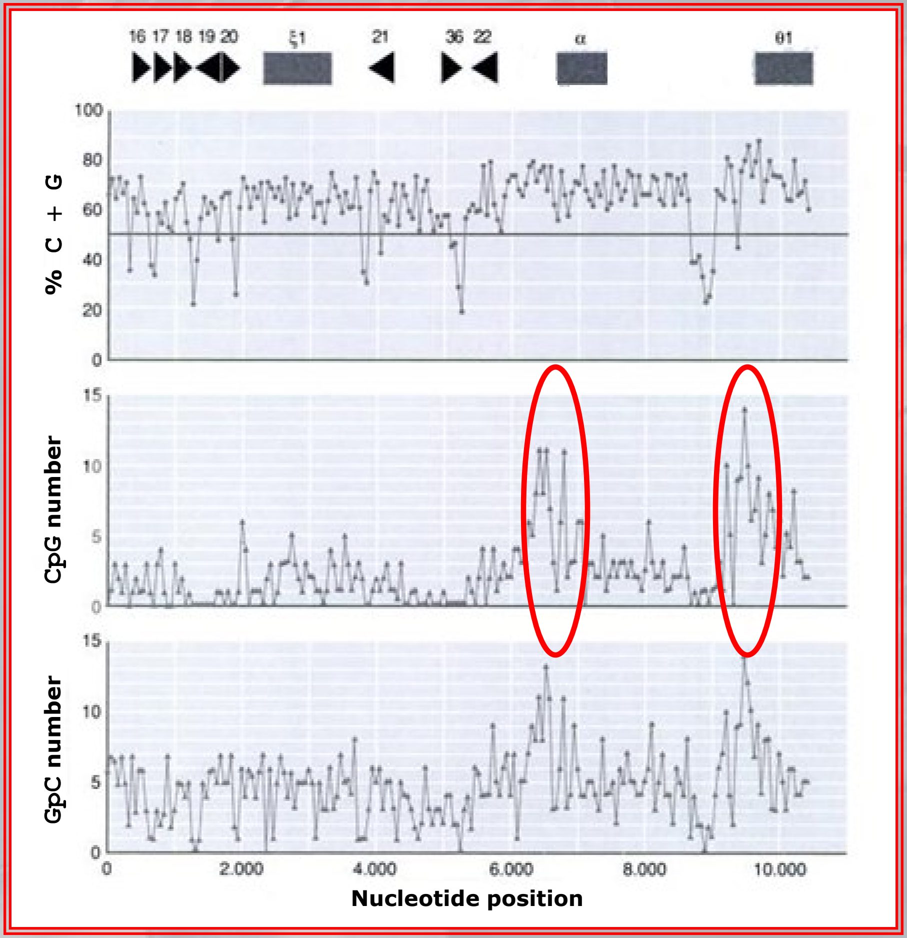
CpG islands – 2

- An interesting exception to the general scarcity of **CpG** was detected in sequences 1–2Kb long, posed at the 5′ termination of many human genes
- The so–called CpG islands are typically found in a position that ranges from approximately –1500 to +500, and have a density of **CpG** similar to that which would be expected if the dinucleotides were uniformly distributed
- Many individual **CpG** islands are involved in the binding sites of known transcriptional enhancer sequences (e.g., in that of the ubiquitous constitutive factor Sp1)

CpG islands – 3

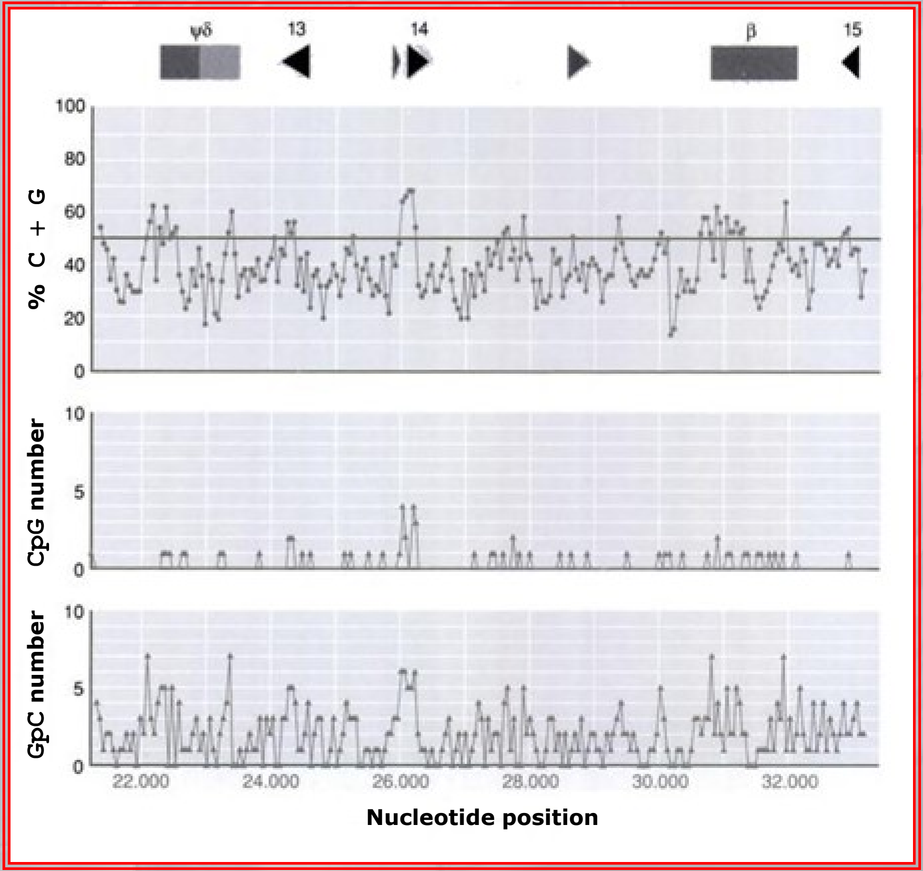
- The analysis of the complete human genome indicates that there would be approximately 45,000 islands and that about half of them are associated with *house-keeping* genes, expressed at a constant level in all the tissues and throughout the life of the organism
- Many of the remaining CpG islands appear to be associated with the promoters of tissue–specific genes (such as the human α–globin), although less than 40% of the known tissue–specific genes exhibits these islands (for instance, the human β–globin does not have a CpG island in its vicinity)
- Instead, CpG islands are found very rarely in non–coding regions, or in genes that have accumulated inactivating mutations

CpG islands – 4



- The set of α–globin genes lie in a tissue–specific portion of the human genome with a high GC content
- Gray rectangles indicate genes, the (numbered) black arrows describe repeated sequences (junk DNA, C repetitions)
- A CpG island is associated with the 5′ termination of both globin genes (α and θ_1)
- The number of appearances of the dinucleotide 5′–GC–3′ in a window of 200 bps is generally higher than that of CpG, which is very changeable due to methylation

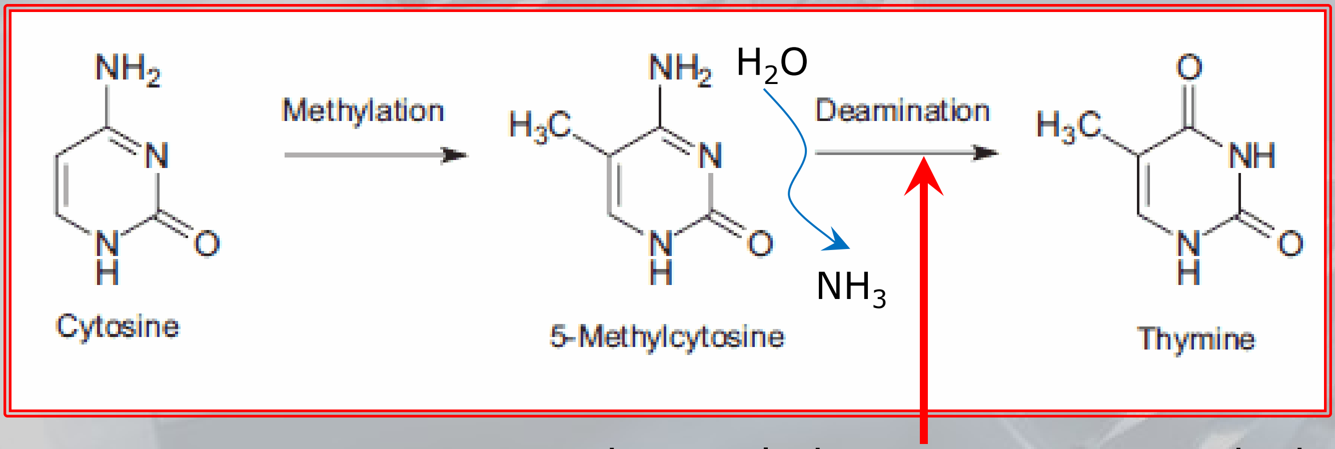
CpG islands – 5



- The set of β–globins lie in a tissue–specific portion of the human genome with a poor GC content
- In this case, a CpG island is not present in the promoter of the β–globin gene

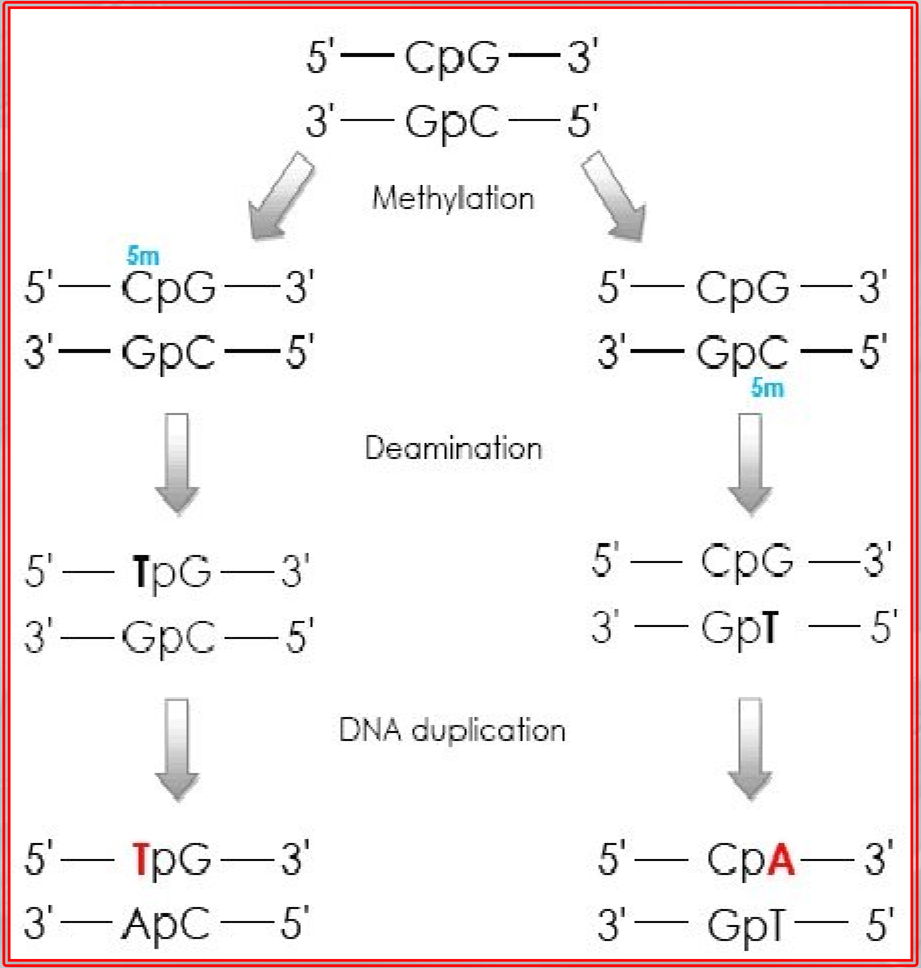
CpG islands – 6

- CpG islands are also intimately associated with a significant chemical modification of the DNA of many eukaryotes, called methylation
- A specific enzyme, DNA methylase, attacks the methyl group CH_3 (negative) to the cytosine, but only when it is present in dinucleotides 5′–CG–3′



A common chemical damage to DNA, which converts methylcytosine into thymine

CpG islands – 7



CpG ⇨ TpG/CpA

CpG islands – 8

- Methylation itself seems to be responsible for the rarity of CpG in the whole genome, because methylated cytosines appear particularly prone to mutations (in particular, TpG and CpA)
- High levels of DNA methylation in a certain region are associated with low levels of histone acetylation and vice versa
  - Histones are proteins that bundle the DNA, and that are found only in eukaryotes
- The degree of histone acetylation (addition of an acetyl group, $COCH_3$, to the N–terminal of a lysine) regulates the gene expression
- Low levels of DNA methylation and high levels of histone acetylation are strongly correlated with high levels of gene expression

CpG islands – 9

- In the human γ–globin gene, for example, the presence of six methyl groups in the region between –200 and +90 effectively suppresses the transcription
- The removal of the three methyl groups present upstream of the transcription start site or of the three methyl groups localized downstream, however, does not allow the initiation of transcription
- Nevertheless, the total removal of the six methyl groups enables the operation of the promoter
- Although there are exceptions to this rule, transcription seems to require that the promoter region should be free from methyls
    - Housekeeping genes have unmethylated CpG islands, whereas the CpG islands of tissue–specific genes are unmethylated only in the tissues in which the adjacent gene is actually expressed

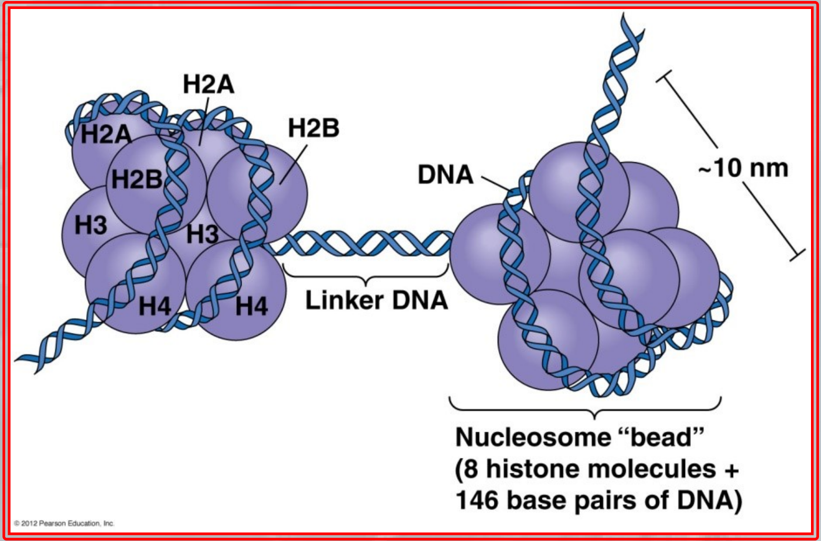
CpG islands – 10

- The methylation patterns differ significantly from one type of cell to another and, for instance, the γ–globin gene is generally free from methyl groups only in erythroid cells (i.e., cells which will develop into red blood cells)
- While the mere presence of **CpG** islands indicates the proximity of an eukaryotic gene, patterns of DNA methylation are sometimes difficult to be experimentally determined and are infrequently reported in the context of genomic sequence data (in the annotations)

CpG islands – 11

- Histones are well preserved eukaryotic proteins, with a very high positive charge, which gives them a strong affinity with the negatively charged DNA molecules
- The mixture, in an approximately equal amount in terms of mass, of DNA and histones (closely associated to it), present within eukaryotic nuclei, is called chromatin
  - Chromatin is the “form” in which the nucleic acids are found in the nucleus of an eukaryotic cell

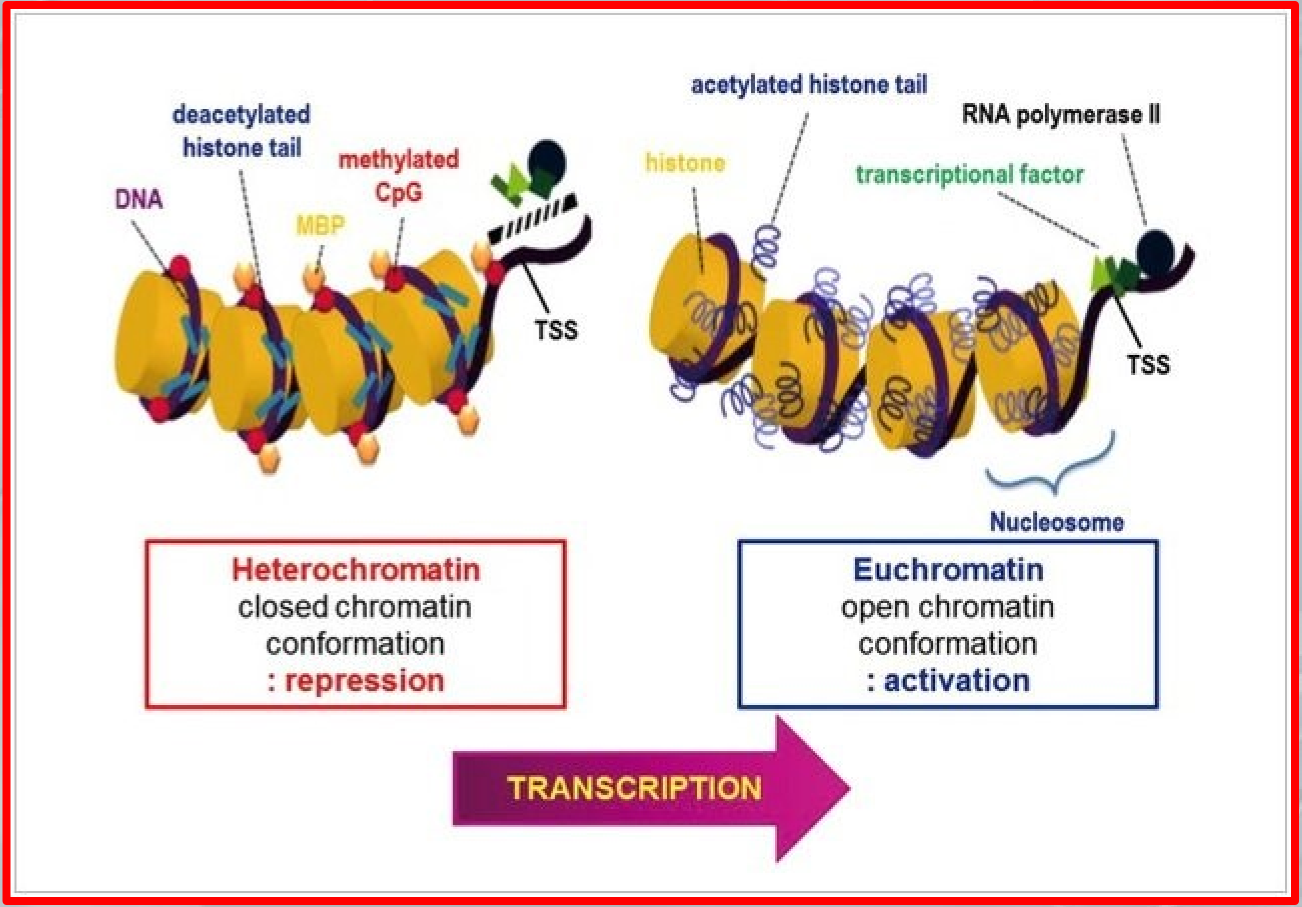
CpG islands – 12



CpG islands – 13

- The transcriptionally active regions are, generally, areas where the histone positive charge is reduced through the addition of acetyl groups
- The resulting lower affinity of these histones to the negatively charged DNA causes the chromatin to be less tightly packed and, therefore, makes the DNA strands more accessible for the RNA polymerase
    - These open chromatin areas are known as euchromatin, in contrast to the transcriptionally inactive and densely packed chromatin, called heterochromatin
    - The information stored into the heterochromatin is not lost, but it is less likely to be used in gene expression

CpG islands – 14

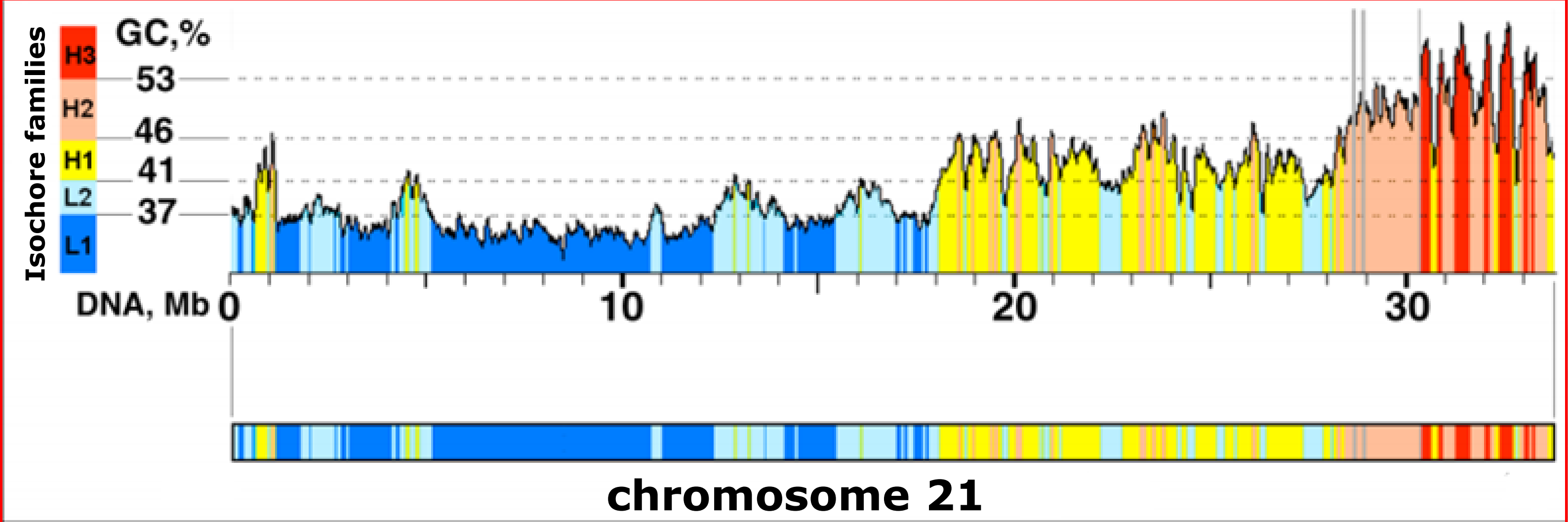


Isochores – 1

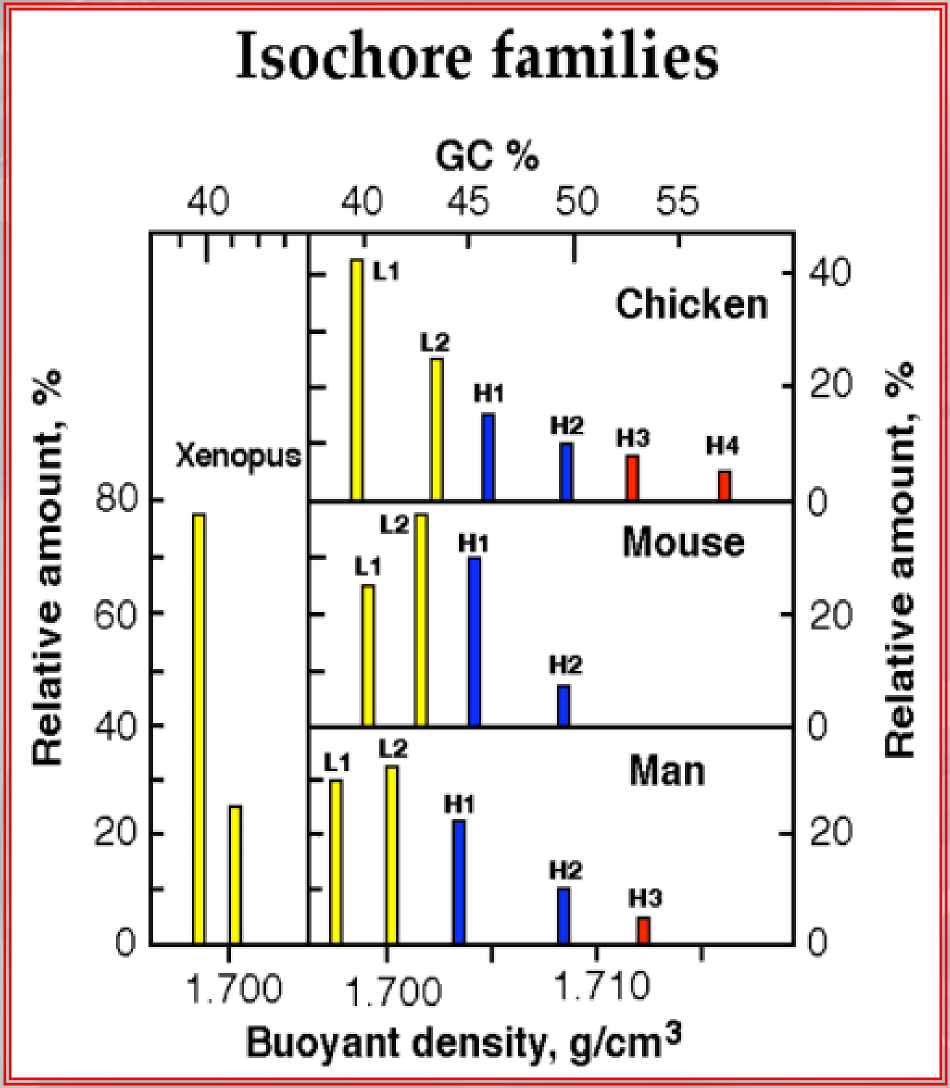
- The vertebrate genome can be considered as a mosaic of isochores, i.e. of large DNA segments having a homogeneous nucleotide composition
- Nucleotide composition refers to the frequency with which guanine and cytosine are present in a specific isochore (i.e., the molar ratio of G+C is considered)
- The very definition of isochores as "long regions with homogeneous nucleotide composition" implies two important concepts:
  - The isochore genomic sequences go beyond the one million base pairs
  - The isochore GC content is relatively uniform from its start to its end (variations <1%), although, in general, it is significantly different also for contiguous isochores

Isochores – 2

- The experiments performed on human chromosomes suggest that our genome is a mosaic of five different isochore classes:
  - Two isochores are poor in GC content (L1 and L2, with an average GC content of 39% and 42%, respectively)
  - Three isochores are relatively rich in GC (H1, H2 and H3, with an average GC content of, respectively, 46%, 49% and 54%)



Isochores – 3


Isochore families
GC %
40
40
45
50
55
Xenopus
Chicken
L1
L2
H1
H2
H3
H4
Mouse
L1
L2
H1
H2
Man
L1
L2
H1
H2
H3
Relative amount, %
80
60
40
20
0
40
20
0
20
0
20
0
Relative amount, %
1.700
1.700
1.710
Buoyant density, g/cm3

Isochores – 4

- The H isochores of humans and other eukaryotes are particularly rich in genes and represent an excellent starting point for genomic sequencing and gene searching
  - Example: The isochore with the maximum **GC** content, H3, has a density of genes at least 20 times higher than that of the isochore L1, rich of **AT**
- Perhaps even more interesting is the fact that the genes found in the **GC**–rich isochores are very different from those coming from low density isochores
  - Although the human H3 isochore represents a relatively small fraction of our genome (3–5%), it contains almost 80% of our housekeeping genes
  - In contrast, isochores L1 and L2 (which together comprise about 66% of the human genome) contain about 85% of our tissue–specific genes

Isochores – 5

- The diversification of isochore families is associated with several other important features of the eukaryotic genome
  1) The methylation pattern and the chromatin structure – GC–rich isochores tend to have low levels of methylation of their CpG and to be stored as transcriptionally active euchromatin
  2) The way to regulate gene expression – the GC–rich regions tend to have elements of the promoter sequence closest to the transcription start site
  3) Introns and gene length – the GC–rich regions tend to have shorter introns and genes
  4) The relative abundance of repeated long and short sequences – short sequences predominate in GC–rich isochores, long ones in GC–poor isochores
  5) The relative frequency of the amino acids used to build proteins – genes contained in GC–rich isochores tend to use amino acids that correspond to codons rich in G and C

Preferences in the use of codons – 1

- It was experimentally proved that each organism prefers to use the same codon, out of a set of equivalent triplets, to code for a certain amino acid
- Examples:
  - Along the entire yeast genome, arginine is represented by the codon **AGA** in 48% of the cases, although it can be translated by five other functionally equivalent codons (**CGT**, **CGC**, **CGA**, **CGG** and **AGG**), which, compared to the first one, are used with lower frequencies (approx 10% for each codon)
  - The fruit fly shows a similar preference in the use of codons for arginine, but in this organism, the preferred codon is **CGC** (33% compared to a rate of 13% for the other equivalent codons)

Preferences in the use of codons – 2

- The biological basis to explain these preferences are related to the need of avoiding codons that are similar to stop codons, as well as to ensure efficient translation by choosing codons that correspond to tRNA particularly abundant in the organism
- Regardless of the reasons for such preferences, the choice of certain codons over others is significantly different among eukaryotic species
  - Exons generally reflect these preferences, but this is not true for any, randomly chosen, string of codons

Gene expression – 1

- The term gene expression is defined as the series of events that, after the activation of the gene transcription, leads to the production of the corresponding protein
- The regulation of these processes is very precise and its complexity increases going up the evolutionary ladder
- Studying the regulation of the gene expression means to ascertain in what tissues a gene is expressed (and at which levels), under what conditions, and what is the effect of this event

Gene expression – 2

- All the cells of a given organism share the same genomic kit
- The tissue–specific gene expression determines the morpho–functional phenotype of both the cell and the tissue
- In any differentiated cell and in each particular development phase of an organism only a subset of genes is active
- All the problems encountered in the recognition of eukaryotic genes lead to consider that one of the most correct checks to confirm a gene prediction is the experimental demonstration that a given living cell actually transcribes that region in an RNA molecule

Gene expression – 3

- Some characteristics of the DNA sequences useful for the gene recognition algorithms are:
  - Known promoter elements (e.g., **TATA** and **CAAT** box)
  - **CpG** islands
  - Splicing signals associated with the introns
  - Open reading frames using particular codons
  - Similarity with "expressed sequence tags" or with known genes from other organisms
- Even if only the nucleotide sequence of certain RNA transcripts is known for an organism, such information can be used to facilitate the recognition of genes, for example using pair alignments

Gene expression – 4

- It is important to remember that the ability of an organism to alter its gene expression pattern in response to environmental changes is a central feature of the concept of living beings
- Gene expression regulation
  - It can happen on each of the phases that characterize the passage of genetic information from DNA to proteins
  - In complex eukaryotes, the gene expression regulation primarily takes place via the transcription control
  - Main types of regulation
    - Epigenetic control (methylation, acetylation)
    - Transcriptional control (chromatin structure)
    - Post–transcriptional control (maturation, transport, translation and stability of mRNA)

Gene expression – 5

- Short–term gene expression regulation

  Genes are rapidly activated or repressed in response to changes in environmental or physiological conditions in a cell or in the whole organism

- Long–term gene expression regulation

  Related to genes involved in the development and in the differentiation of cells within an organism

- Methods for the large–scale gene expression study
  - Systematic sequencing of ESTs from cDNA libraries
  - SAGE (*Serial Analysis of Gene Expression*)
  - cDNA microarray

cDNA and ESTs – 1

- cDNAs, short form for “complementary DNAs”, provide the most convenient way to isolate and manipulate portions of the eukaryotic genome transcribed by RNA polymerase II
- The complementary DNA is a double–strand DNA synthesized from a sample of mature messenger RNA
- In order to produce the cDNA, two helices are synthesized in two steps: the first helix is produced using the mRNA as a template, while the second is obtained starting from the first produced helix

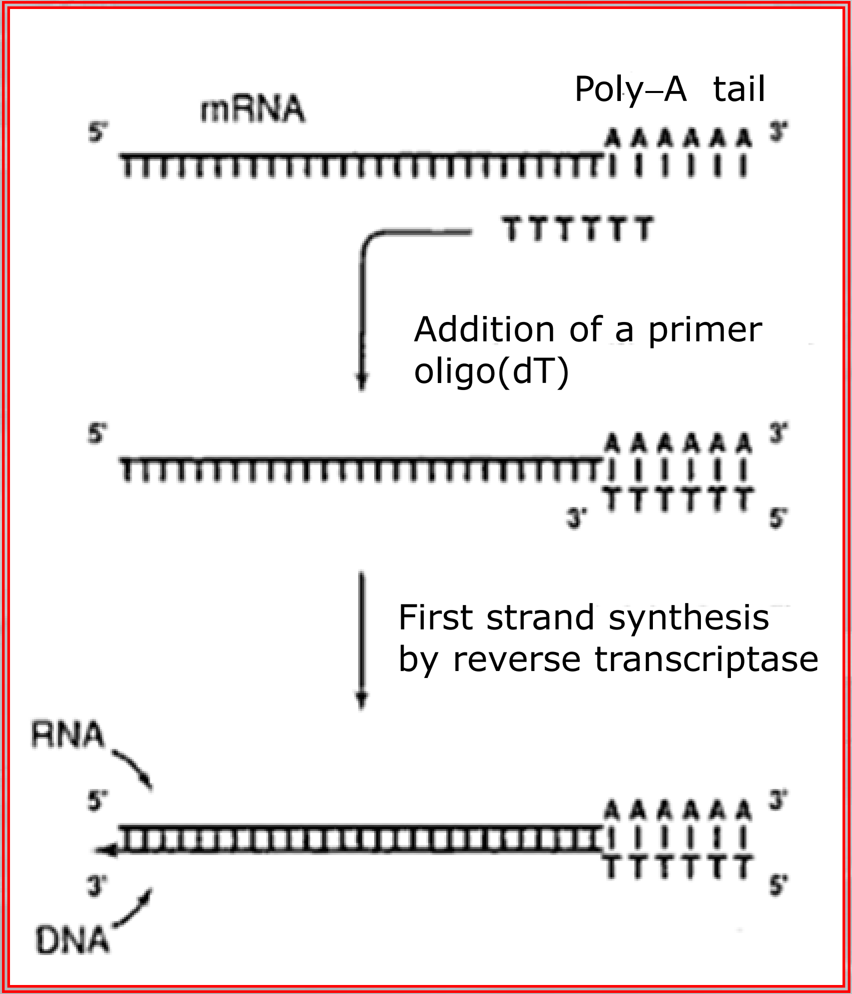
cDNA and ESTs – 2

- Template helix synthesis
  - For the synthesis of the template helix, complementary to the mRNA sequence, the reverse transcriptase enzyme is used
  - This enzyme operates on a single mRNA strand, generating its complementary DNA, based on the coupling of the RNA nitrogenous bases (**A**, **U**, **G**, **C**) with the complementary DNA bases (**T**, **A**, **C**, **G**)

cDNA and ESTs – 3

- Synthesis procedure
  - The eukaryotic cell transcribes the DNA into the RNA (pre–mRNA or hnRNA)
  - The same cell processes the pre–mRNA filaments by eliminating the introns, also adding a cap to 5′ and poly–**A** tail to 3′
  - The mature mRNA filaments are extracted from the cell cytoplasm
  - The mRNA is put in a contact solution with an oligo-nucleotide primer of poly–**T**, which hybridizes with the mRNA poly–**A** tail
  - The reverse transcriptase recognizes the primer and starts the production of the cDNA, based on the presence of deoxynucleotides required for its elongation (without the primer, the enzyme does not work)

cDNA and ESTs – 4

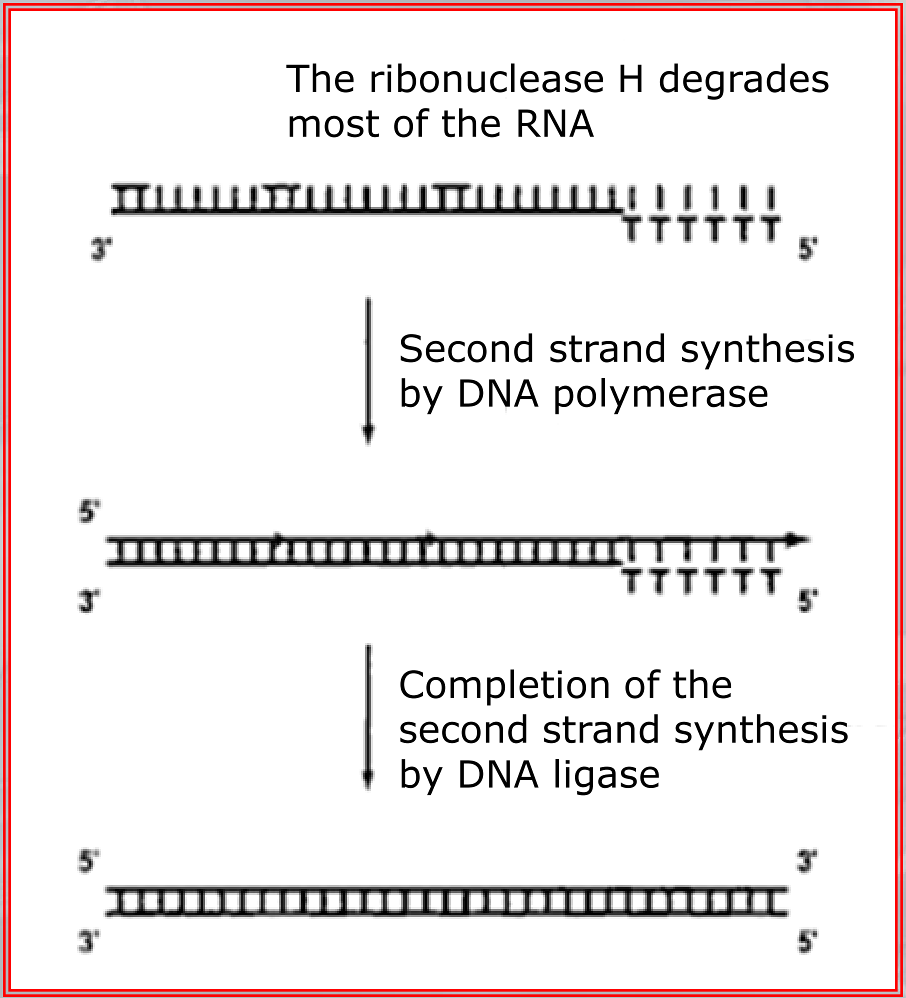

Poly–A tail
mRNA
5'
AAAAAA
3'
TTTTTT
Addition of a primer oligo(dT)
5'
AAAAAA
3'
3'
TTTTTT
5'
First strand synthesis by reverse transcriptase
RNA
5'
AAAAAA
3'
3'
TTTTTT
5'
DNA

cDNA and ESTs – 5

- Coding helix synthesis
  - The coding helix synthesis takes place in the same way as in DNA replication, and uses three enzymes: DNA polymerase, Ribonuclease H, and DNA ligase
- Synthesis procedure
  - Ribonuclease H recognizes the RNA–DNA dimers and degrades the mRNA, leaving only some short fragments
  - The short RNA fragments serve as primers for DNA polymerases, that copy the complementary helix
  - The exonuclease activity of the enzyme causes the degradation of the RNA primers and their replacement with DNAs
  - The DNA ligase joins all the fragments, generating the complete helix

Many DNA polymerases, such as the bacterial DNA polymerase I, possess exonuclease activity, generally in the 3′→5′ direction, and thanks to this property, correct pairing errors that can occur during DNA replication

cDNA and ESTs – 6



The ribonuclease H degrades most of the RNA

Second strand synthesis by DNA polymerase

Completion of the second strand synthesis by DNA ligase

cDNA and ESTs – 7

- Being obtained by reverse transcription of mRNA, which has already undergone the process of splicing, the cDNA does not present non–coding intronic sequences
- Typically, the cDNAs are fragmented and cloned: collections of filaments are obtained, corresponding to fragments of expressed genes, forming cDNA libraries
- A cDNA library, which is prepared from the mRNA contained in the cells of some particular tissue, may be considered as a snapshot that reproduces the composition of the mRNA population present in the tissue at a particular development phase of the organism and for certain physiological conditions

cDNA and ESTs – 8

- Or, in other words… cDNA libraries in which the clones to be sequenced are chosen randomly can be used to describe, both qualitatively and quantitatively, the population of the mRNAs
  - Approximately 50% of the mass of mRNA is found exclusively in specific tissues
  - Mammals have approx 10000 housekeeping (constitutive) genes, always expressed in all the cells
- cDNAs are also used for the production of probes employed in microarray hybridization experiments
- Moreover, partial sequences of cDNAs are used as ESTs (Expressed Sequence Tags), useful in the assembly of contigs, for gene mapping and recognition

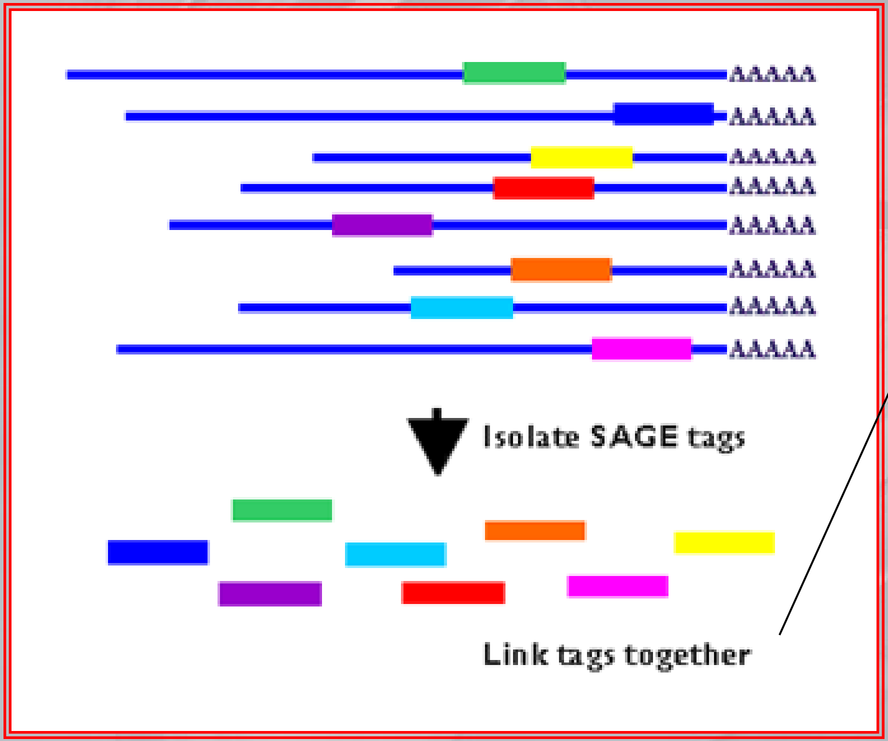
Serial analysis of gene expression – 1

- SAGE (*Serial Analysis of Gene Expression*) is an experimental method designed to use the advantages of large–scale sequencing with the aim to obtain gene expression quantitative information (*Velculescu et al.*, 1995, *Zhang et al.*, 1997)
- It consists in sequencing short oligonucleotides, which act as sequence labels ("tags")
- SAGE allows the estimation of the expression level of each gene, through the measurement of the number of times that the "tag" representing such gene appears in a large enough sample of "tags", sequenced starting from the messenger of the analyzed tissue

Serial analysis of gene expression – 2

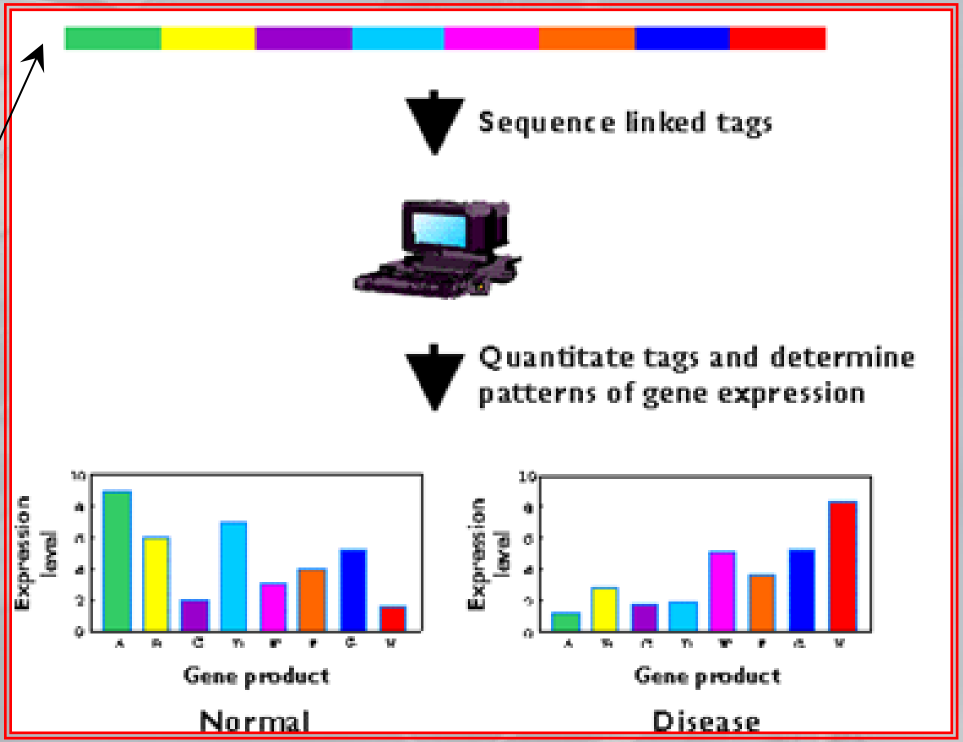
- SAGE is based on three main principles:
  - The cDNAs produced from a cell are segmented into small fragments (originally from 10 to 14 nucleotides, now also a bit longer) obtained with the use of restriction enzymes
  - The “tags” can be joined together in series, to form long DNA molecules, which are cloned and sequenced in an automated way
  - The number of times in which a single “tag” is observed allows to quantify the abundance of that particular messenger, identified in the population of mRNAs and, indirectly, the expression level of the corresponding gene

Serial analysis of gene expression – 3



cDNA synthesis and restriction enzyme cleavage to generate SAGE tags

Linkage of tags (concatenation) prior to sequencing



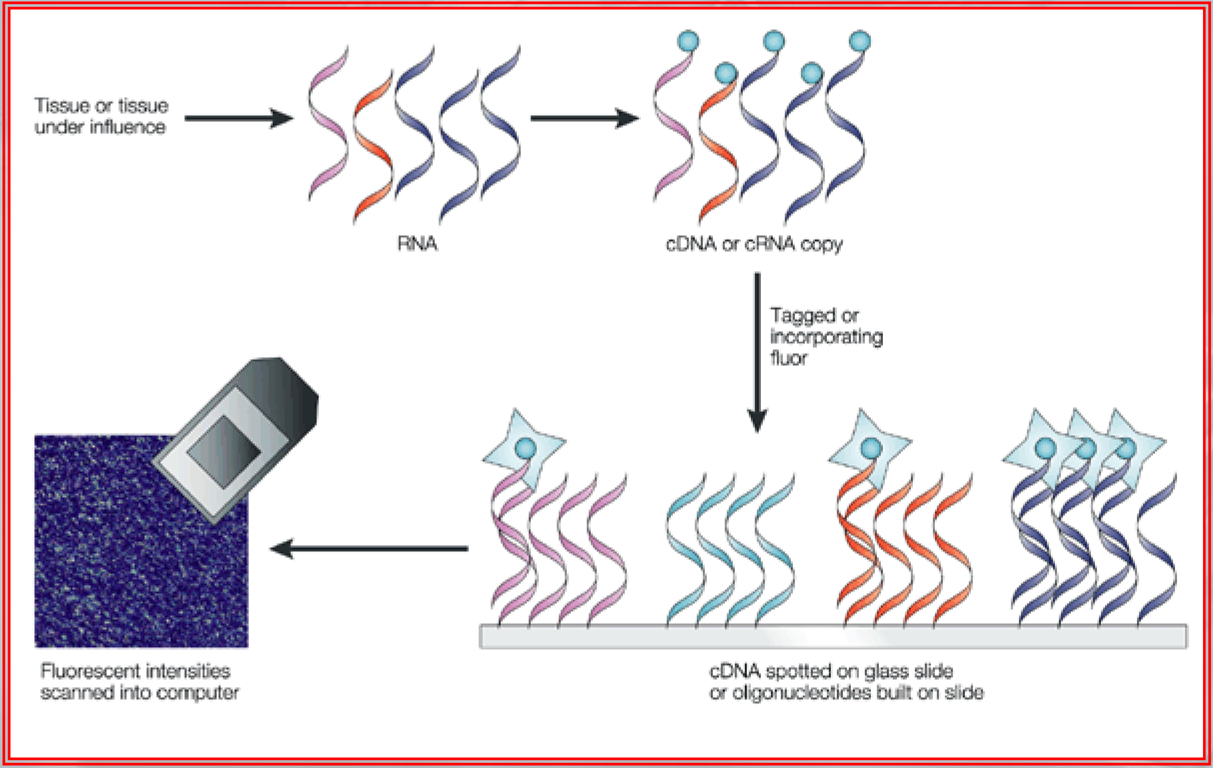
Comparison of gene expression data between normal and diseased tissues

Microarrays – 1

- Microarrays are small glass or silicon slides upon the surface of which short cDNA fragments (in multiple copies) are fixed – usually from hundreds of thousands to many millions of probes
- In other words… a microarray is a set of short Expressed Sequence Tags (ESTs) made from a cDNA library of a set of known (or partially known) gene loci
- Using a conventional hybridization process, the level of expression of genes is measured
- Microarrays are read using laser–based fluorescence scanners
- The process is "high throughput"

Microarrays – 2

- The results are typically displayed as a grid in which each “dot” represents a particular gene and the relative level of expression is indicated by colors or gray scales


Tissue or tissue under influence
RNA
cDNA or cRNA copy
Tagged or incorporating fluor
Fluorescent intensities scanned into computer
cDNA spotted on glass slide or oligonucleotides built on slide

Pharmacogenomics – 1

- The gene expression profile, or the transcriptional profile, has been applied to a wide variety of biological problems, such as metabolic pathways mapping, tissue identification, and medical diagnosis
- Recently, the gene expression patterns have been used to distinguish between two types of lymphoma, that often are not diagnosed correctly: the diffuse large B–cell and the follicular lymphoma
- The microarray technique, with probes for 6817 different human genes, indicates that, between the two types of cancer, there are significant differences in the expression level of 30 genes

Pharmacogenomics – 2

- Taken together, the patterns of expression of the 30 distinctive genes have allowed the correct classification of 71 out of 77 tumors (91%), with a substantial increase in performance with respect to the previously used cytological indicators
- Improvements in the diagnostic field may be of decisive importance, especially when the drug treatments are significantly different in different cases
- Medical applications of gene expression profiles are not restricted, however, to diagnostics

Pharmacogenomics – 3

- For 58 patients with diffuse large B–cell lymphoma, changes in gene expression patterns, in response to specific treatments, were evaluated
- Prediction techniques, based on supervised learning, have been applied to the obtained data, allowing the binary classification of patients with very different five–year survival rates (70% vs. 12%) with a high degree of reliability
- The implications are clear: the quicker it is possible to determine that a patient does not respond adequately to a certain treatment, the greater the probability of being able to intervene in time to modify the drug, in order to produce a positive evolution of the disease

Pharmacogenomics – 4

- Possibility of a massive development of targeted treatments to individual problems
- The relatively new field of pharmacogenomics (closely related to that of precision medicine) aims to maximize the effectiveness of treatments while minimizing unwanted side effects, using information about the genetic makeup of individuals and how their gene expression patterns change in response to various therapies

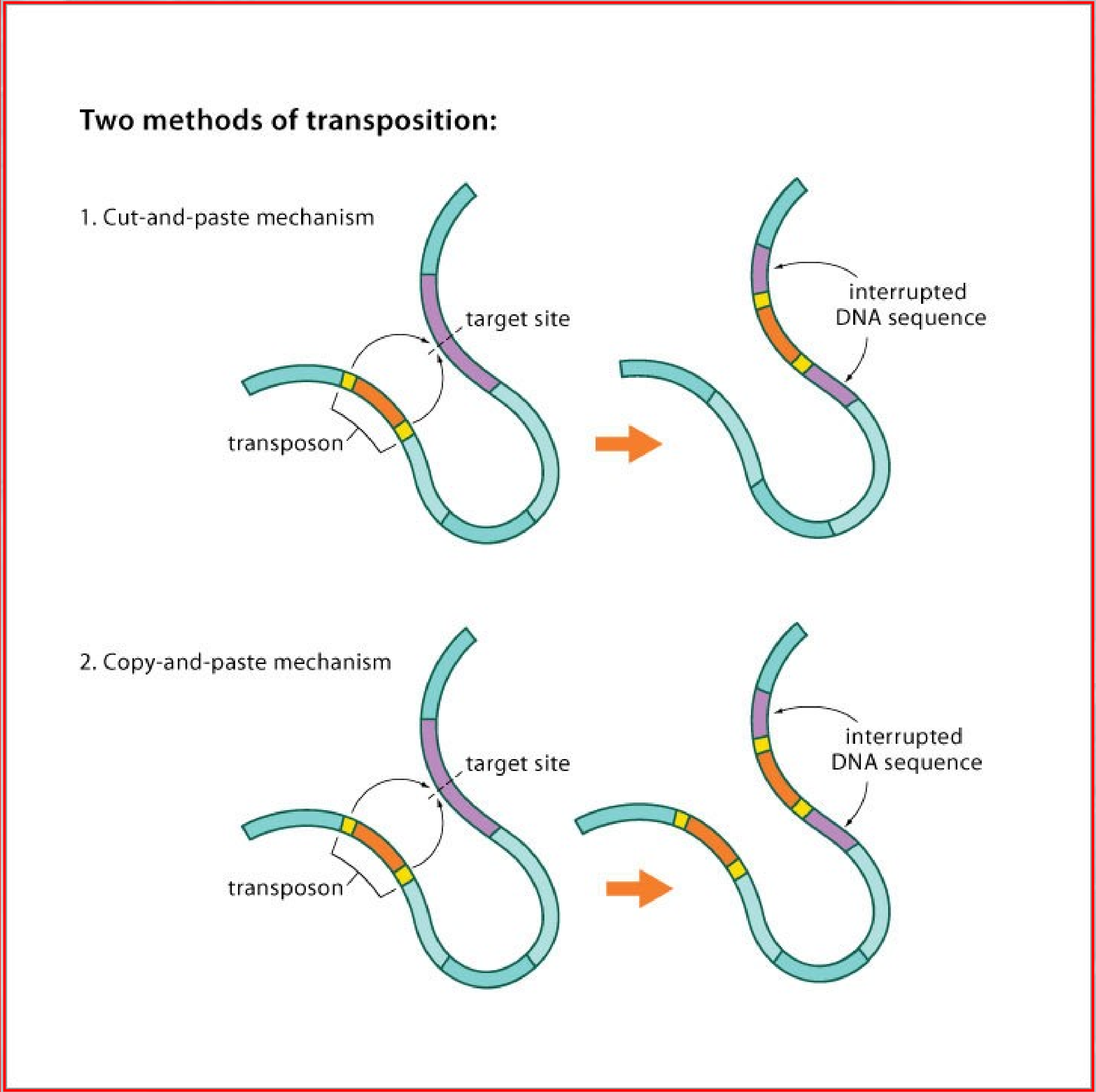
Transposition – 1

- Transposons are “transposable” genetic elements present in the chromosomes, able to move from one position to another within the genome
- Prokaryotic genomes are extremely simplified in terms of their information content
- However, the transposonic DNA, which is often present in multiple copies and is quite superfluous to its host, is also an important component in the anatomy of the bacterial genome
- Example: A single *E.coli* genome may actually contain 20 different insertion sequences (ISs)

Transposition – 2

- Most of the sequence of an IS is dedicated to one or two genes that encode for an enzyme, called transposase, which catalyzes the IS transposition within the genome in a conservative (the number of copies does not change) or in a replicative (the number of copies increases) way
- Different bacterial transposons are:
  - Composite transposons – pairs of IS elements that facilitate transposition and, sometimes, horizontal gene transfer
  - Tn3 transposons – always transposed in a replicative manner
  - Transposable bacteriophages – viruses replicated as part of their normal infection cycle

Transposition – 3



Two methods of transposition:

1. Cut-and-paste mechanism

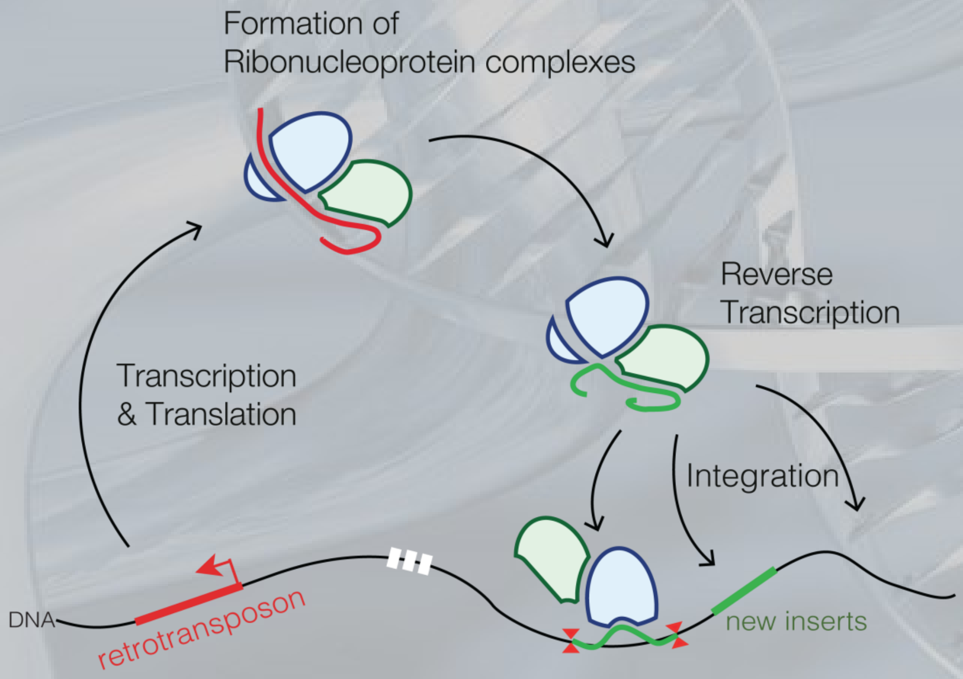
2. Copy-and-paste mechanism

Transposition – 4

- Prokaryotic transposons are often randomly distributed in the genome, and their presence is usually sufficiently variable to allow a reliable distinction among descendants of the same species
- Even the eukaryotic DNAs, with their abundance of non–coding data, contain transposonic DNA, although a recent estimate suggests that there are no more than 23000 transposons in the human genome, belonging to 11 different families
- An important eukaryotic transposon is the *mariner*, 1250 bps long, originally found in fruit flies, but detected, since then, in many eukaryotes, including man
  - It may be used for both natural and engineered horizontal gene transfer in eukaryotic organisms

Transposition – 5

- However, in eukaryotes, those transposons, called retrotrasposons, that propagate through intermediate RNAs, are actually much more common



Repeated elements – 1

- DNA transposons present in multiple copies (and often constituted by repeated sequences) within eukaryotic or prokaryotic genomes are qualified as “repetitive DNA”
- Uncommon in prokaryotes, the repeated elements which do not propagate themselves through the transposase action, constitute, instead, a very large portion of the genome of many eukaryotes
  - Tandem repeated DNA (5’–**CACACACA**–3’)
    - Satellite DNA
    - Mini/micro–satellites
  - Interspersed repeats throughout the genome

Repeated elements – 2

- The satellite DNA (representing a large part of constitutive heterochromatin) takes its name from the fact that its very simple sequence (from 5 to 200 bps), with an abnormal composition of nucleotides, originates DNA fragments with an unusual density of certain bases with respect to the other genomic sequences
  - Although some satellite DNA is shed in the eukaryotic genome, most of it is located in the centromeres (the decentralized "bottleneck" of chromosomes) and in the telomeres (their final parts), suggesting that it must play a structural role and, given its location, must protect the chromosome from degradation or shortening that would cause the loss of encoding genes
- Minisatellites form clusters which are up to 20,000 base long, containing many copies of sequences, no longer than 25 bps, arranged in tandem

ATTCGATTCGATTCGATTCG

Repeated elements – 3

- Microsatellites form clusters of short repeated sequences (typically consisting of four nucleotides, at most) covering about 150 bases in total
  - They are rather regularly distributed within the eukaryotic genome
  - Example
    - In humans, microsatellites with **CA** repeats are approximately present once every 10,000 bps and represent 0.5% of the entire genome
    - Individual nucleotides repetitions (for example, **A**) constitute another 0.3% of the human genome

Repeated elements – 4

- Surprisingly, the DNA polymerase can “lose the thread” during the replication of these simple sequences and, often, gives rise to longer or shorter versions of the sequence so that tandem repetitive DNA can stretch over several generations
    - The high level of variability in the microsatellite length from one individual to another has made them excellent genetic markers (for geneticists, for the forensic use, for maternity/paternity tests)

Repeated elements – 5

- We define retrotransposons those DNA fragments that independently transcribe themselves into an intermediate RNA and that are consequently able (via the intervention of inverse transcriptase) to produce replicated copies in different positions within the genome

  ⇨ interspersed repeats
- Inverse transcriptase does not generally belong to the set of genes of normal cells: It is acquired by infectious (single strand) retroviruses (similar to the human AIDS)
- Retrotransposons are particularly abundant in plants, where they constitute a substantial fraction of the entire genome, and in mammals, including humans

Repeated elements – 6

- LINEs (*Long Interspersed Nuclear Elements*) are long (more than 5000 base pairs) interspersed DNA sequences
  - They code for two genes, which have a reverse transcriptase and integrase activity, allowing the copy and the transposition of the same genes and of other non–coding sequences (such as SINEs)
  - Since LINEs transpose themselves by replication, they are able to increase the size of a genome
  - Example: *L1* repetitions are very common LINEs in the human genome; they are approximately 6100 bps long and are present in roughly 3500 copies dispersed throughout the genome
  - The human genome contains more than 900,000 LINEs, which constitute about 21% of the entire genome

Repeated elements – 7

- SINEs (*Short Interspersed Nuclear Elements*) are short (less than 500 base pairs) DNA sequences
  - SINEs are rarely transcribed, and do not encode reverse transcriptase; therefore, they need proteins, encoded by other sequences (such as LINEs), for being transposed
  - Example: The most common SINEs in mammalian genomes (although different in different species) belong to the family of *Alu sequences* and can be identified by the fact that they are capable of binding the enzyme Alu I (hence the name); the elements of this family have an average length of 258 bps and are present in the human genome with over a million copies
  - SINEs constitute about 11% of the total genetic heritage

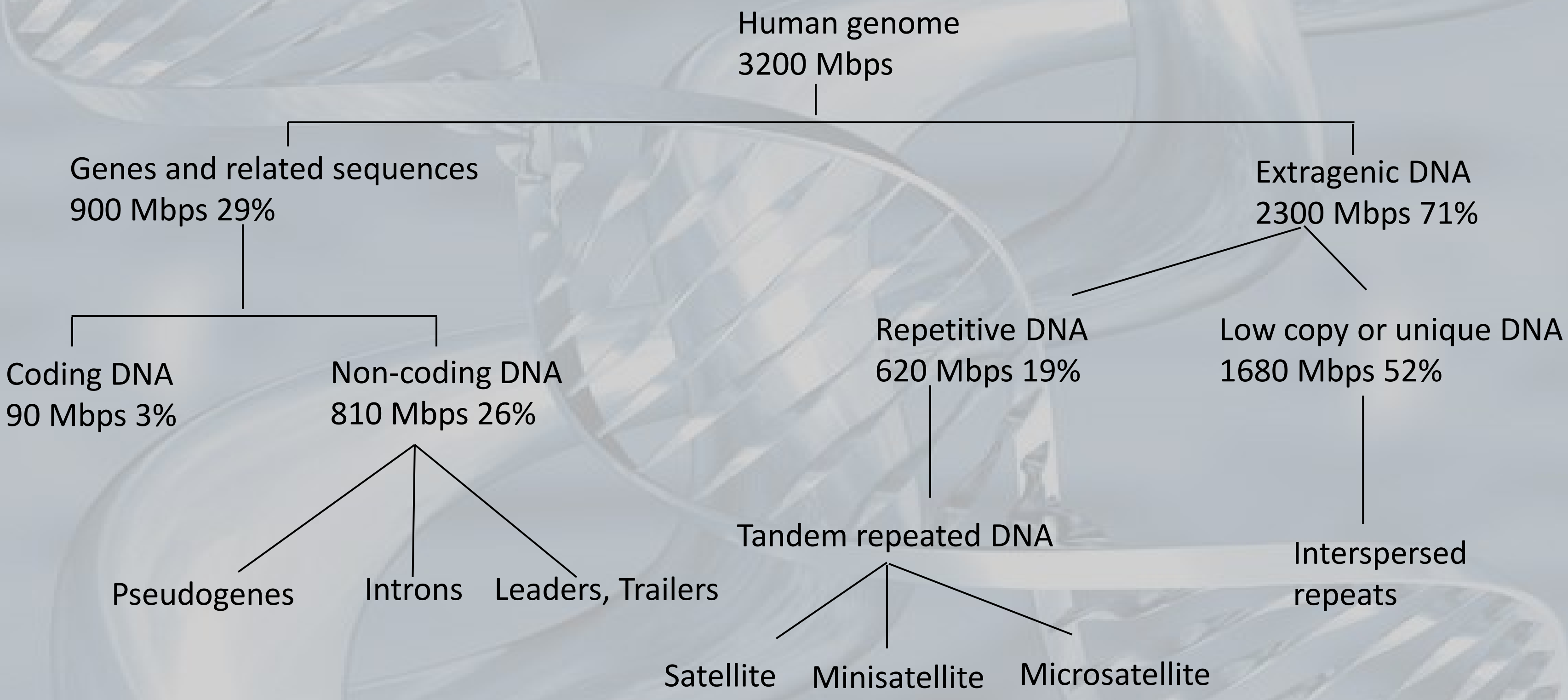
Repeated elements – 8

- Although usually classified as junk DNA, recent research has suggested that both LINEs and SINEs may have had an important role in the evolution of genomes, so as significant effects at the structural and transcriptional level
- However, in a bioinformatics perspective, many algorithms for genome analysis “camouflage” known repeated sequences, because their information content, usable in the detection of genes or for sequence comparisons, is negligible

Eukaryotic gene density – 1

- The C–value paradox made it clear that much of the eukaryotic genome is unnecessary many decades before that molecular biologists have provided the complete nucleotide sequence of several genomes
- The human genome project has largely confirmed the hypothesis underlying that paradox:
  - Out of the ∼3200Mbps of the human genome, not more than 90Mbps (less than 3%) correspond to coding sequences and, approximately, 820Mbps (26%) correspond to sequences associated with them (introns, promoters, pseudogenes)
  - The remaining 2300Mbps are divided into two kinds of “junk DNA” (subject to any selective constraint): low copied or unique sequences (1680Mbps, 52%) and tandem repeated DNA (620Mbps, 19%)

Eukaryotic gene density – 2



Human genome
3200 Mbps

Genes and related sequences
900 Mbps 29%

Extragenic DNA
2300 Mbps 71%

Coding DNA
90 Mbps 3%

Non-coding DNA
810 Mbps 26%

Repetitive DNA
620 Mbps 19%

Low copy or unique DNA
1680 Mbps 52%

Pseudogenes

Introns

Leaders, Trailers

Tandem repeated DNA

Interspersed
repeats

Satellite

Minisatellite

Microsatellite

Eukaryotic gene density – 3

- Genes are far from each other, even in those regions of complex eukariots that are particularly rich of coding information, as the H3 isochore of the human genome
  - The average distance between human genes is around 65,000 base pairs, approximately equal to 10% of the genome size of a simple prokaryotic organism
- Moreover:
  - Mutational analyses have revealed that many genes encode proteins that perform multiple functions
  - Many genes are present in multiple, redundant copies
  - Simple eukaryotes tend to have a higher density of genes compared to more complex organisms, such as vertebrates

Concluding… – 1

- The gene recognition in prokaryotes is a relatively simple task and can be based on the search of statistically significant, long open reading frames
- Moreover, the prokaryotic genome is characterized by a very high density of information content and, normally, it is quite simple to be analyzed
- Conversely, the eukaryotic genome, with its low density with respect to its prodigious dimensions, represents an open challenge for any automatic gene recognition technique

Concluding… – 2

- The recognition software must in fact take into account a wide variety of different characteristics:
  - Preferential use of codons within ORFs
  - Presence of CpG islands located upstream with respect to genes
  - Splicing junctions and branching sites internal to the introns which have good correspondence with the relative consensus sequences
- Unfortunately, the rules associated with the “standard markers” are confused with a lot of common exceptions, and often vary greatly from one organism to another and even from one genomic context or cell type to another

Concluding… – 3

- Nowdays, the best algorithms for the recognition of genes show good performance, although significantly affected by high rates of false positives and negatives
- The recent increase in data availability (both in quantity and variety) for the training and evaluation of such algorithms, however, suggests a significant performance improvement for years to come