

Theoretical Physics Methods for Computational Biology.

M. Caselle

Dip di Fisica Teorica, Univ. di Torino

Berlin, 06/04/2006

Plan of the lectures

1. Introduction: Biological background
 - Genome organization.
 - Gene expression and regulation.
2. A survey of most recent results in genome biology
 - The Phantom project.
 - miRNA and post-transcriptional regulation
3. Theoretical questions in genome biology
 - DNA correlators.
 - Statistical mechanics analysis of DNA-T.F. binding.
 - Graph theory and gene networks.
4. A few topical examples:
 - Identification of T.F. binding sites.
 - Graph theory approach to fragile sites characterization.

First lecture: Biological background

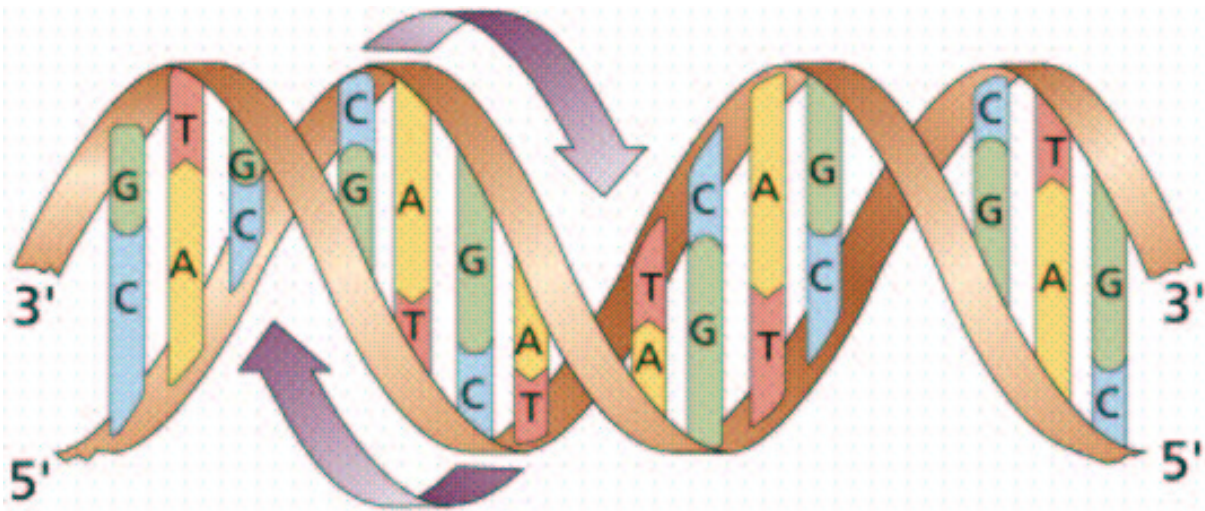
- Genome organization.
- Gene expression: from DNA sequences to proteins.
- Gene regulation.
- RNA versus DNA.
- References and Databases.

Introduction

- Genetic information is encoded in the DNA chain as a long sequence of four types of basis: **A**=Adenine, **C**=Cytosine, **G**=Guanine, **T**=Thymine. (in RNA **U**=Uracil replaces Thymine)

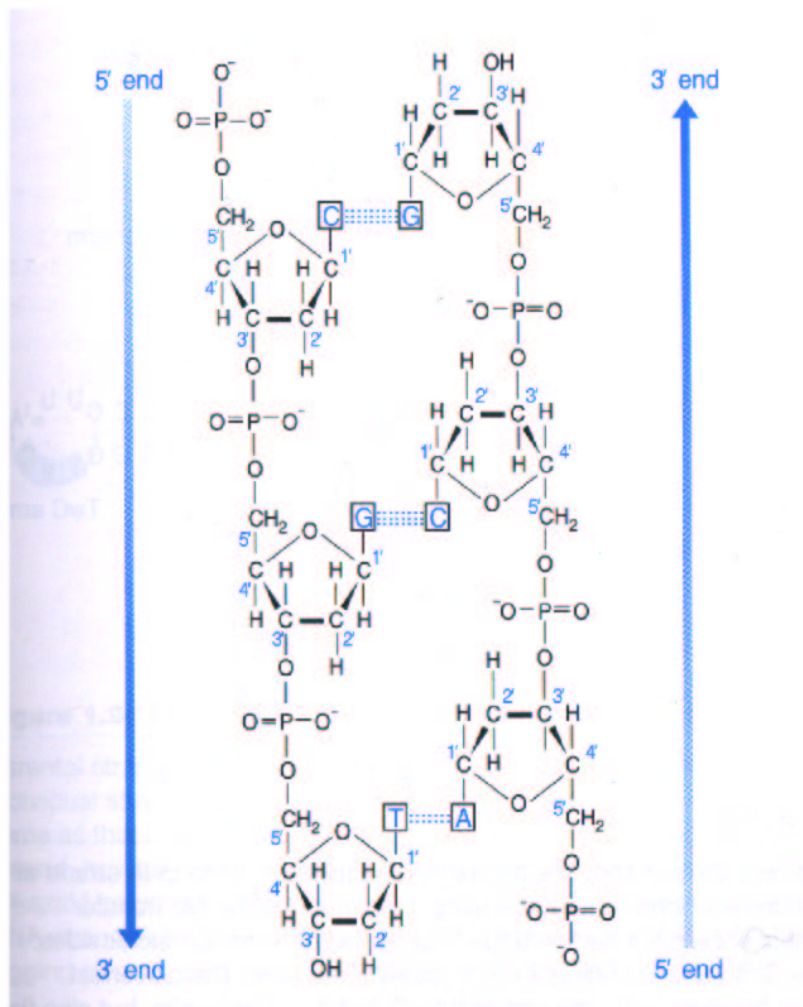
Base pairs bond the two strands of the **double helix** together

Pairing rule: $C \equiv G, A \equiv T$



- The “beginning” of a strand of a DNA molecule is defined as **5'**, the “end” is defined as **3'**. (5' and 3' refer to the position of the bases relative to the sugar molecule in the DNA backbone).

The two strands in a double helix run in opposite directions.



- An organism's total DNA content is known as its **Genome**.

The human genome consists of 22 pairs of **autosomal chromosomes** and two **sex chromosomes** X and Y.

- **Genes** are sequences of base pairs that encode informations for proteins and some RNA molecules (such as ribosomal RNA and Transfer RNA). They can range in size from less than 100 bp (base pairs) to several millions of bp.

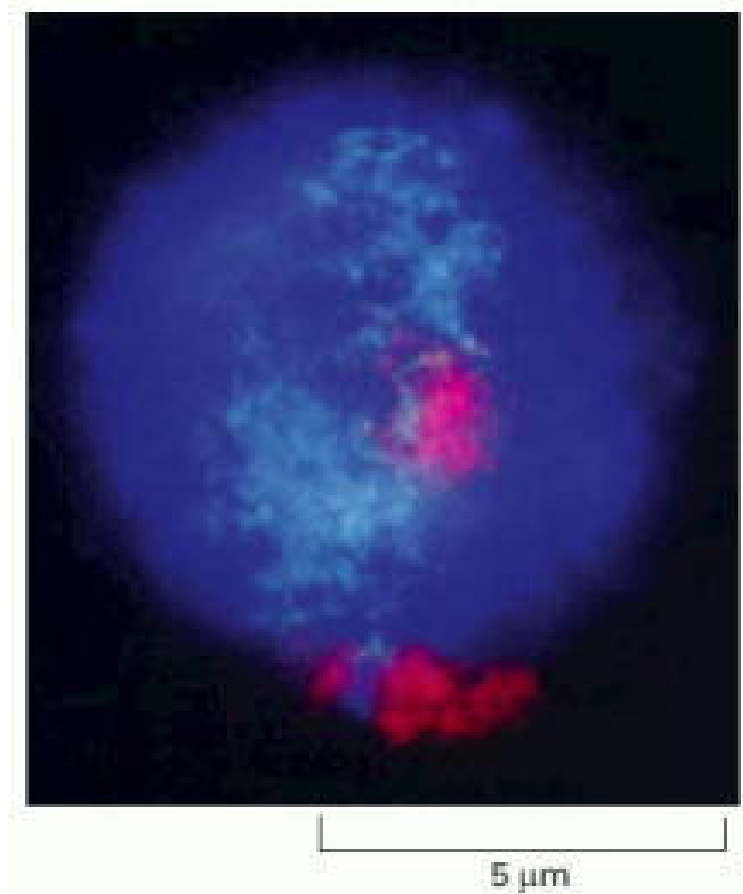
Partial list of sequenced (and publicly available) organisms

Organism	bp (10^6)	Genes
<i>S. cerevisiae</i>	12.1	6,000
<i>C. elegans</i>	97	19,000
<i>D. melanogaster</i>	135.6	13,000
<i>A. thaliana</i>	100	25,000
<i>H. sapiens</i>	3000	30,000

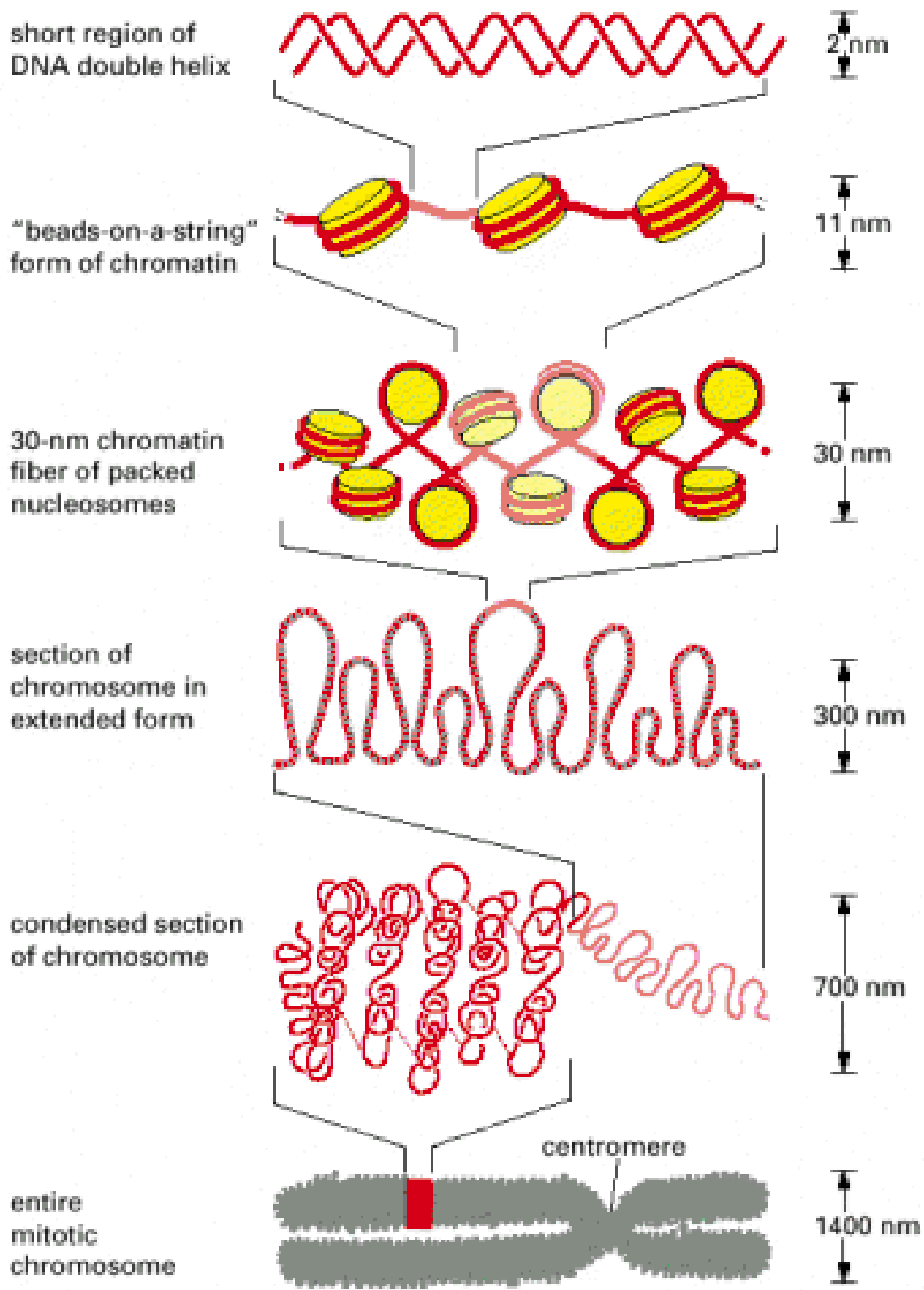
Chromosome organization

- All chromosomes adopt a highly condensed conformation during **mitosis** (see figures below). When they are specially stained, these mitotic chromosomes have a **banding structure** that allows each individual chromosome to be recognized unambiguously. These bands contain millions of DNA nucleotide pairs, and they reflect a poorly-understood coarse heterogeneity of chromosome structure.
- Instead chromosomes are generally decondensed during **interphase** (i.e. in the part of the cell cycle in which most of the genes must be expressed), so that their structure is difficult to visualize directly.

- Although considerably less condensed than mitotic chromosomes, **interphase chromosomes occupy discrete territories in the cell nucleus**; that is, they are not extensively intertwined. The fluorescent light micrograph shows that the two copies of human chromosome 18 (red) and chromosome 19 (turquoise) occupy discrete territories of the nucleus.

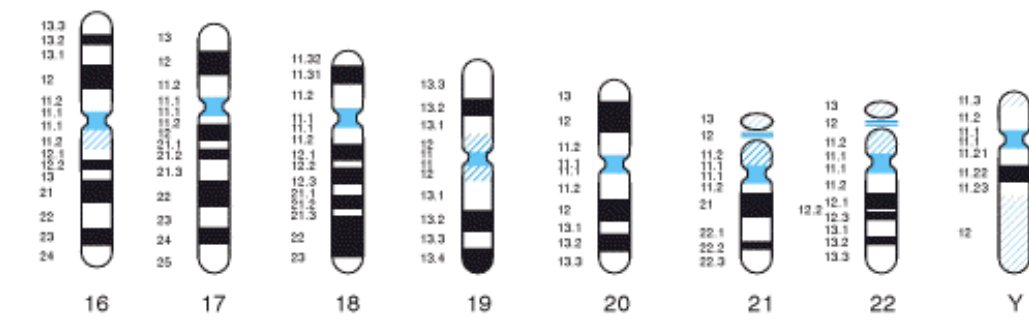
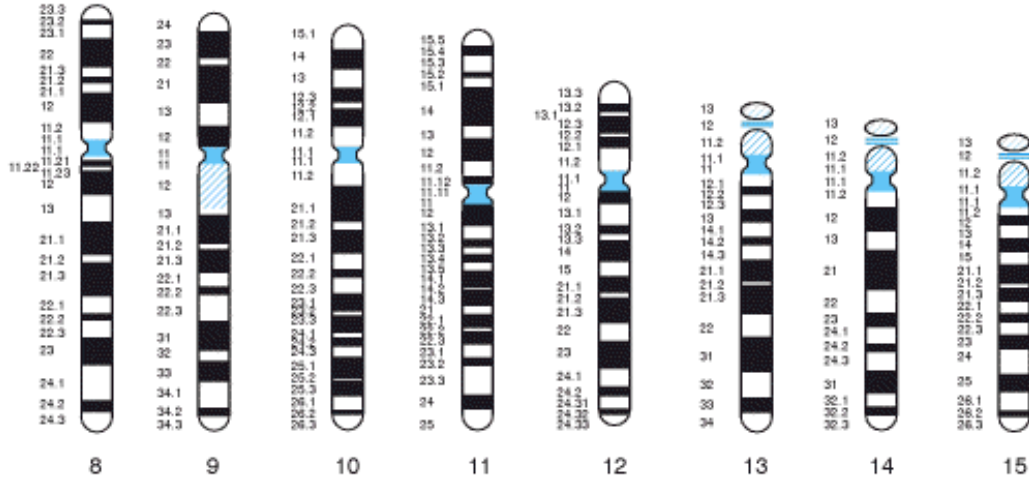
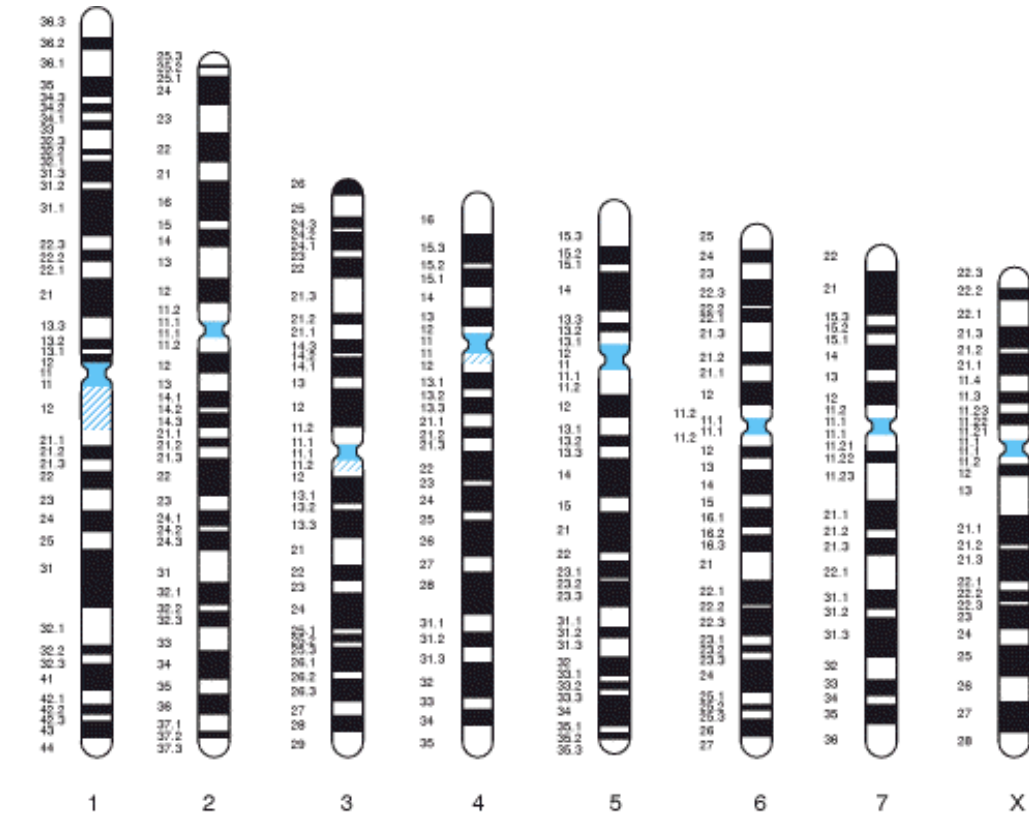


Interphase chromosomes are organized in **Heterochromatic and Euchromatic regions**






NET RESULT: EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 10,000-FOLD SHORTER THAN ITS EXTENDED LENGTH





Key:

-  Centromere
-  rDNA
-  Noncentromeric heterochromatin

Heterochromatin versus Euchromatin

- **Euchromatin** makes up **most of interphase chromosomes** and probably corresponds to looped domains of 30-nm fibers. However, euchromatin is interrupted by stretches of **heterochromatin**, in which 30-nm fibers are subjected to **additional levels of packing** that usually render it **resistant to gene expression**. Heterochromatin is commonly found around **centromeres and near telomeres**, but it is also present at other positions on chromosomes.
- DNA packaged in heterochromatin often consists of **large tandem arrays of short, repeated sequences** that do not code for protein. This turns out to be a **major problem for the sequencing projects**.

- In contrast, **euchromatic DNA is rich in genes** and other single-copy DNA sequences. Although this correlation is not absolute (some arrays of repeated sequences exist in euchromatin and some genes are present in heterochromatin), this trend suggests that some types of repeated DNA may be a signal for heterochromatin formation.
- In the following we shall concentrate on the **euchromatic part of the genome**.

Genome Structure

- The density of protein-coding and RNA-coding sequences becomes lower and lower as the complexity of the organism increases. It is rather high in Prokaryotes, low in *S. Cerevisiae*, very low in the human genome: most of DNA in the human genome is not coding ($\sim 99\%$)
- The biological role of non-coding part of DNA is poorly understood. The common lore is that it should be involved in the regulation of gene expression
- A typical human gene has a complex internal structure. It consists of a set of coding sequences (called **Exons**) interrupted by non-coding sequences called **Introns**. At the beginning and at the end of the gene there are two untranslated regions: **(5'UTR)** and **(3'UTR)** which are important for controlling functions and activities of the genes.

Fixing the scale

It is important to have a grasp on the length scales of the problem. Let us see an example

Example: E. Coli

If E. coli were the size of this room then

DNA: one single molecule,

Size: $\sim 1\text{cm}$ thick, 5×10^6 bp ~ 17 Km long.

Ribosomes: $\# > 10^4$, Size: $\sim 10\text{cm}$ ball

Proteins: $\# \sim 10^6$, Size: $\sim 2\text{cm}$

Under fast growing conditions E.coli

divides in ~ 30 minutes

(Error rate)/replication $< 10^{-6}$

A few general remarks

- Eukaryotes versus Prokaryotes
- DNA versus RNA
- DNA content versus gene content
- The role of Evolution in shaping the genome.
- Phylogenetic trees versus taxonomic trees
- The "RNA world"
-

Gene expression

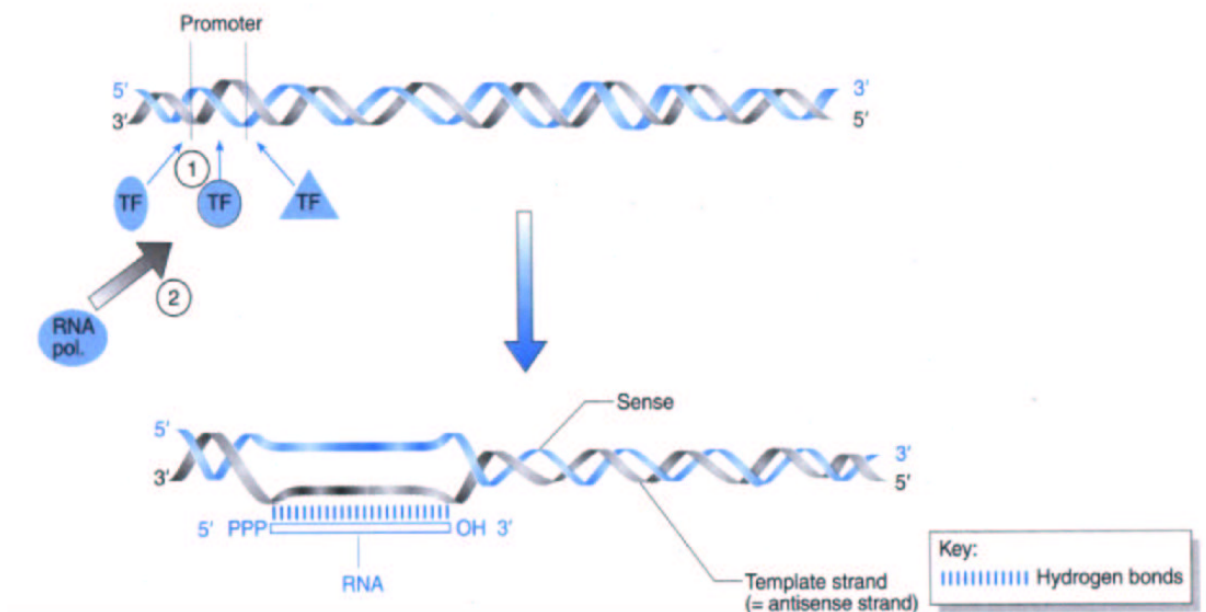
"The Central Dogma" of Molecular Biology

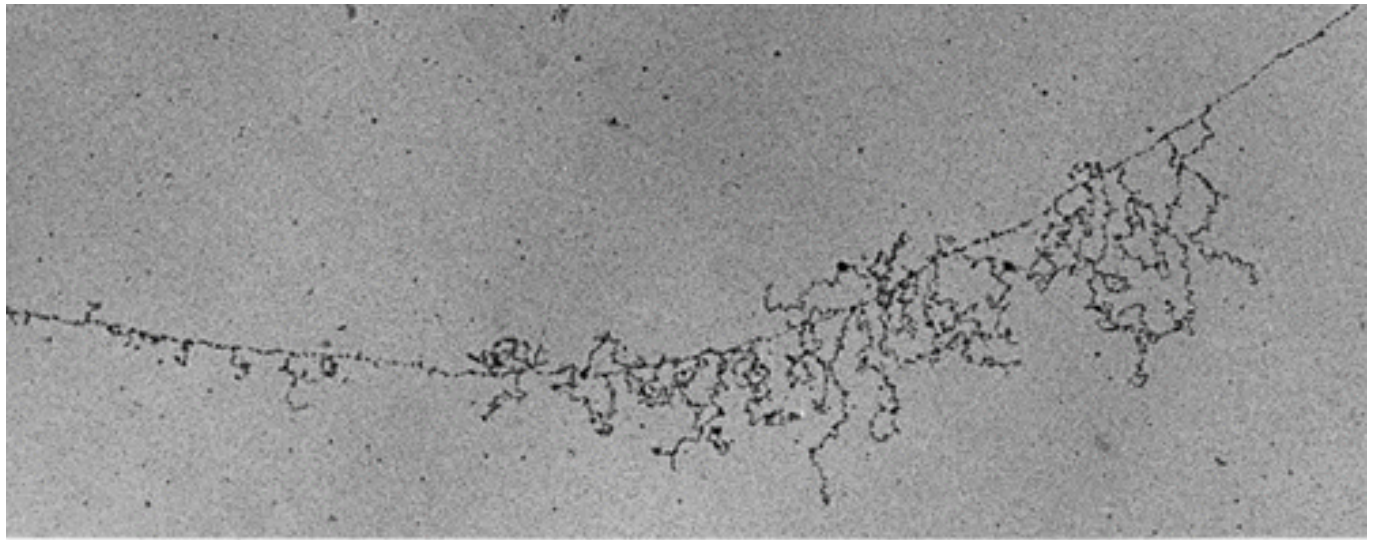
- Gene expression in Eukaryotes involves three main steps:
 - Transcription (from DNA to mRNA)
 - mRNA maturation (splicing)
 - Translation (from mRNA to Proteins)

Transcription

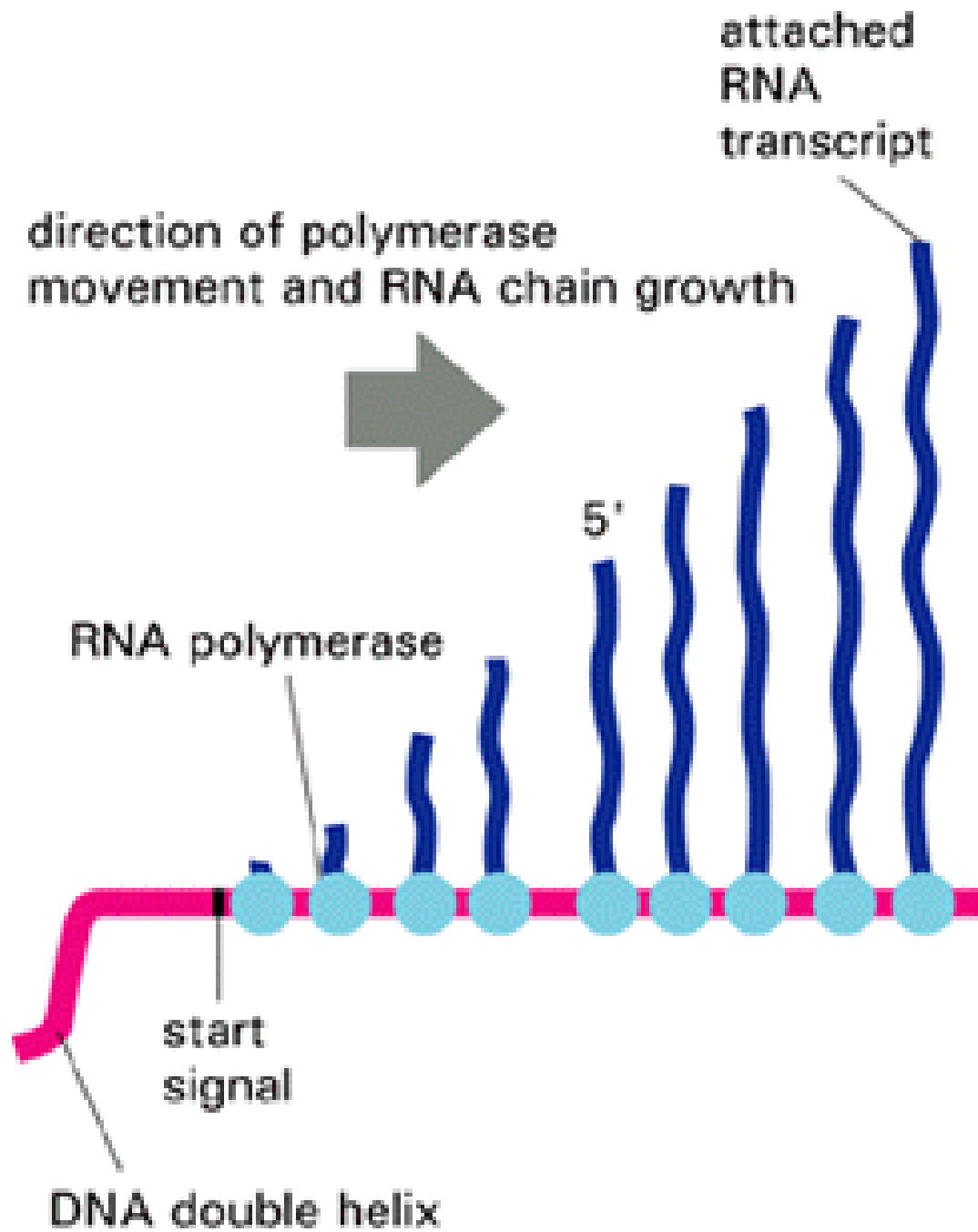
During transcription genetic information in DNA is copied into messenger RNA (mRNA)

- Transcription begins at the start of the gene in 5' (the promoter region) and continues until the end of the gene in 3'.
- The mRNA sequence is complementary to the DNA template strand it transcribes (except uracile bases that replace the thymine ones)





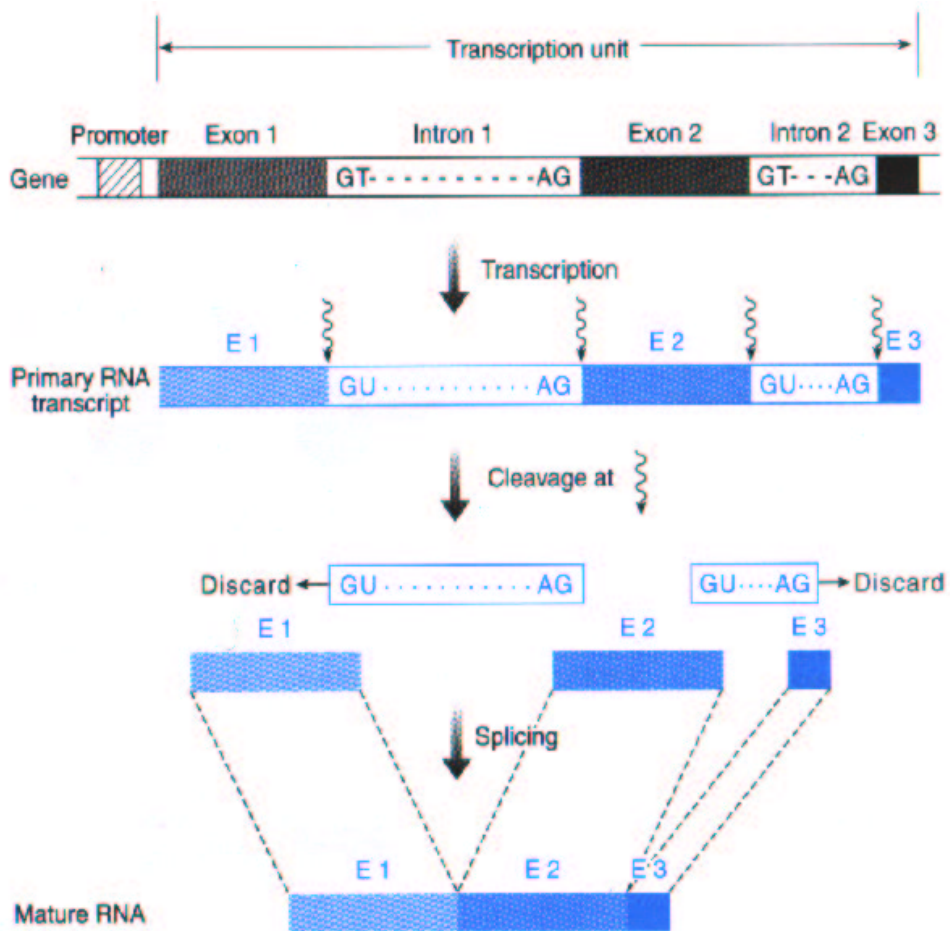
1 μm

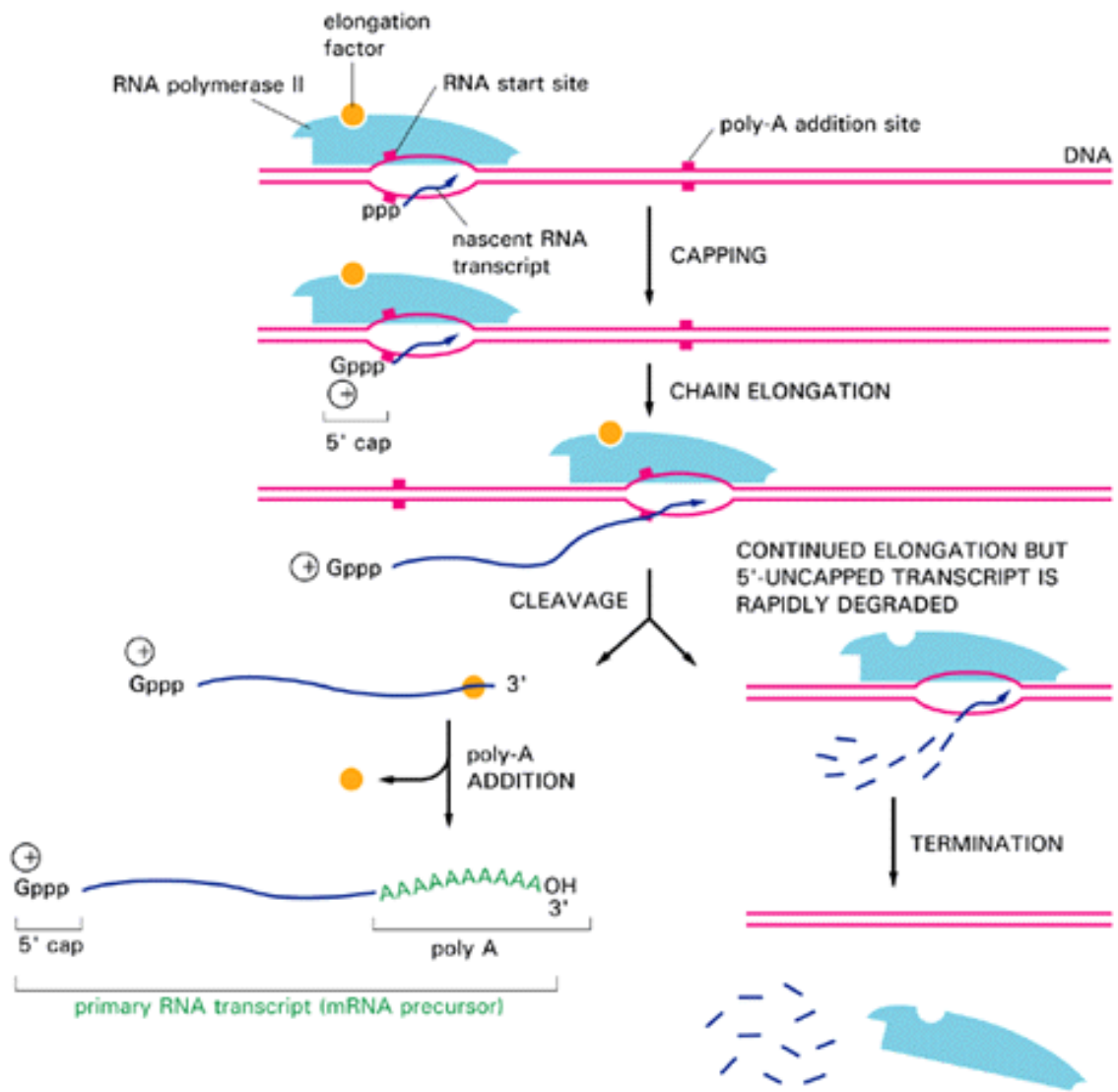


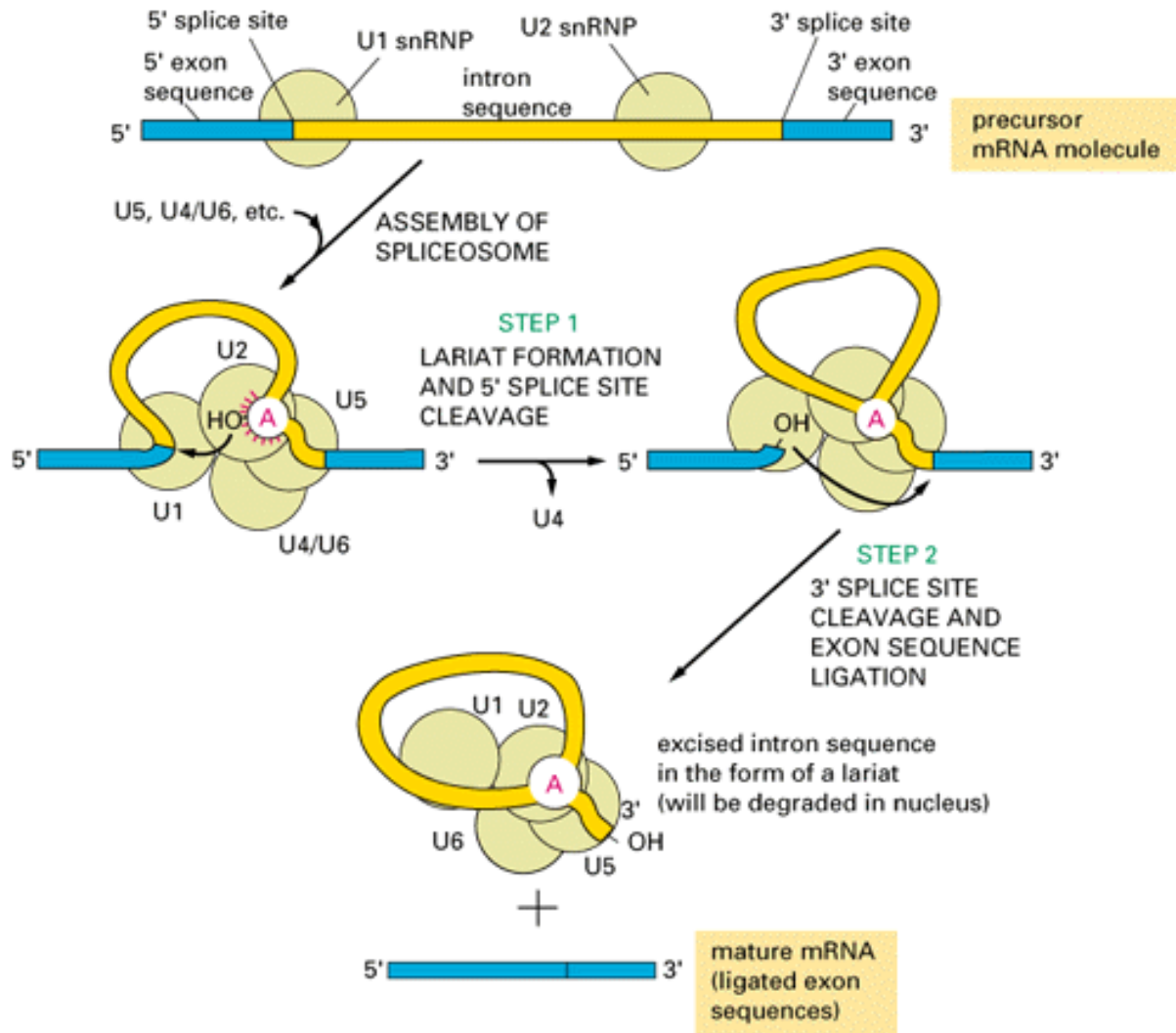
mRNA maturation

The mRNA molecule is processed by

- **Splicing**: remove the introns
- **Capping**: stabilize the 5' end
- **Polyadenylation**: stabilize the 3' end





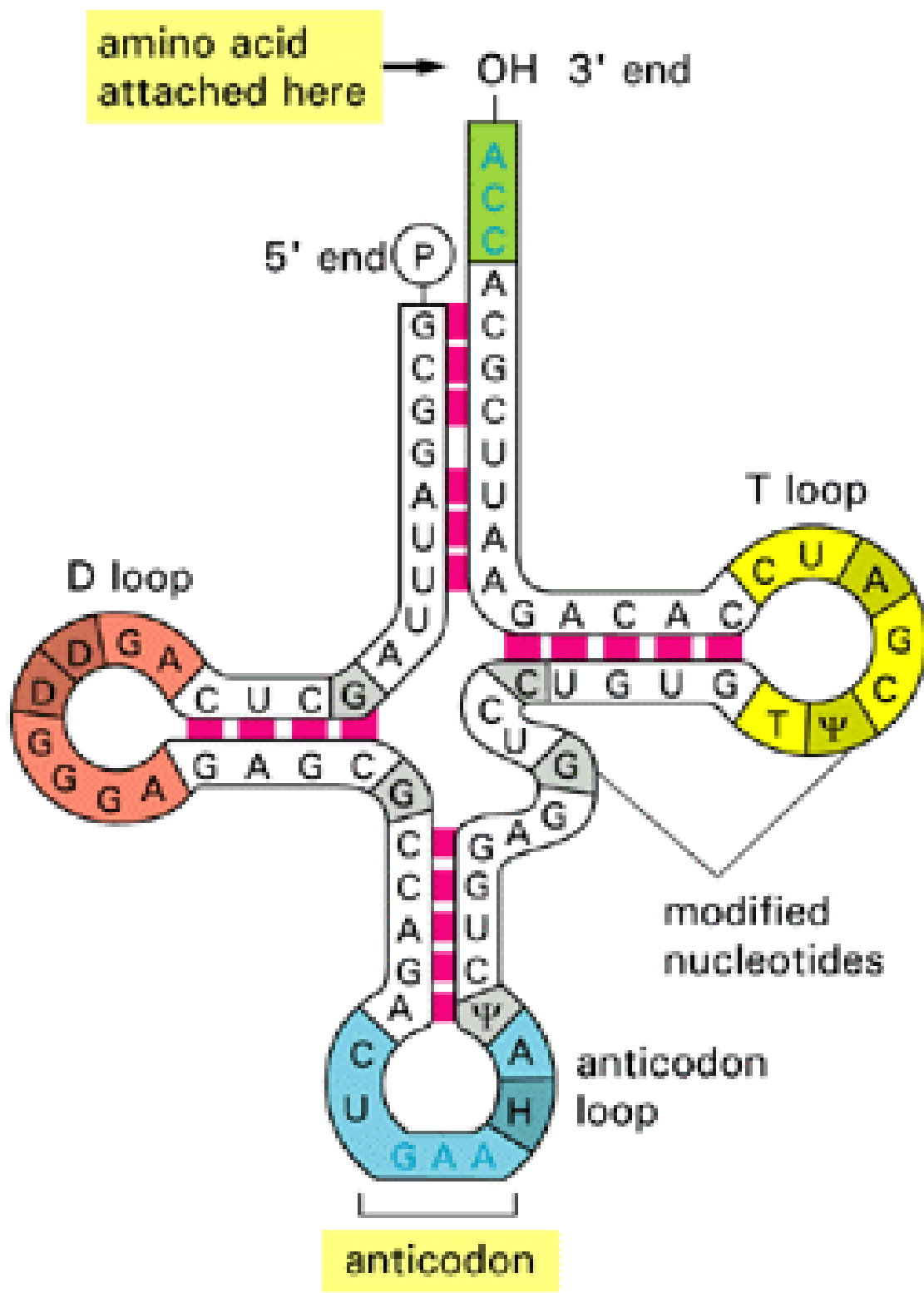


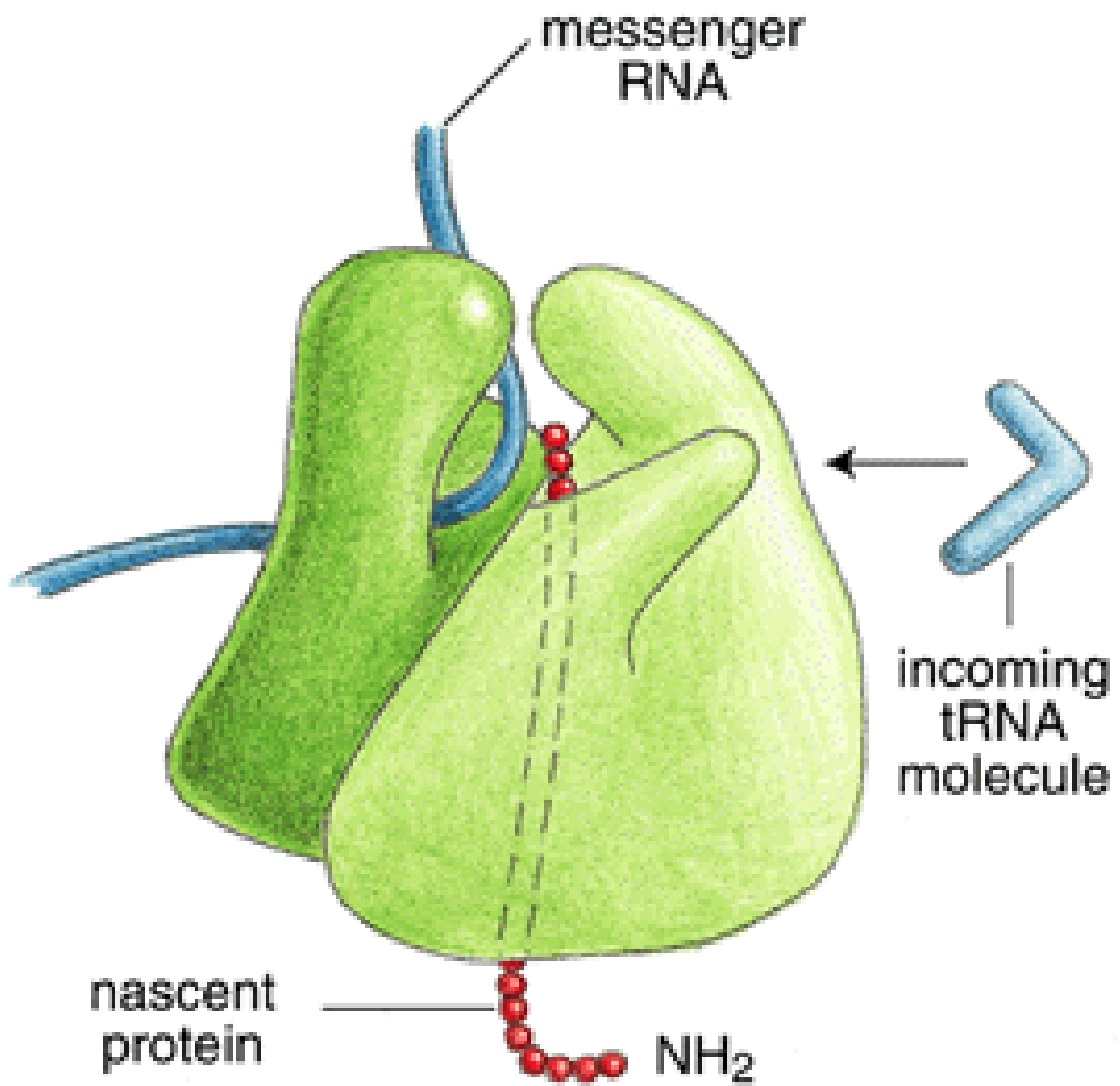
Translation

mRNA is used as a template to synthesize a protein.

- Translation takes place outside the nucleus, in the cytoplasm.
- Proteins are made of amino acids (20 of them)
- Three nucleotides (a codon) specify an amino acid. Since there are only 20 amino acids and $4^3 = 64$ codons, several codons specify the same amino acid. **The genetic code is degenerate.** There are also Start and Stop codons.

		Second Position of Codon				
		T	C	A	G	
First Position	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
		TTA Leu [L]	TCA Ser [S]	TAA Ter [end]	TGA Ter [end]	A
		TTG Leu [L]	TCG Ser [S]	TAG Ter [end]	TGG Trp [W]	G
	C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
		CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
	A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
		ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
		ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G





However....

However, new results from the sequencing projects challenge this too simplified picture: (the "Central Dogma" picture....)

- there are too few genes: ~ 25.000 in human
- half of the genome is made of **repeated sequences**
- non-coding regions (the so called junk DNA) are **impressively conserved** among different species (also very distant ones!)

The "Central Dogma": one gene \rightarrow one protein is **wrong!**

A few tentative answers:

- **Alternative splicing:** one gene → several proteins. This was thought to be an exception. It turns out to be the rule.
- **Retrotransposition:** genetic information flows not only from DNA to RNA but also from RNA to DNA
- **Non coding genes:** the number of "genes" which are transcribed but not translated is much larger than expected, above all in higher eukaryotes.

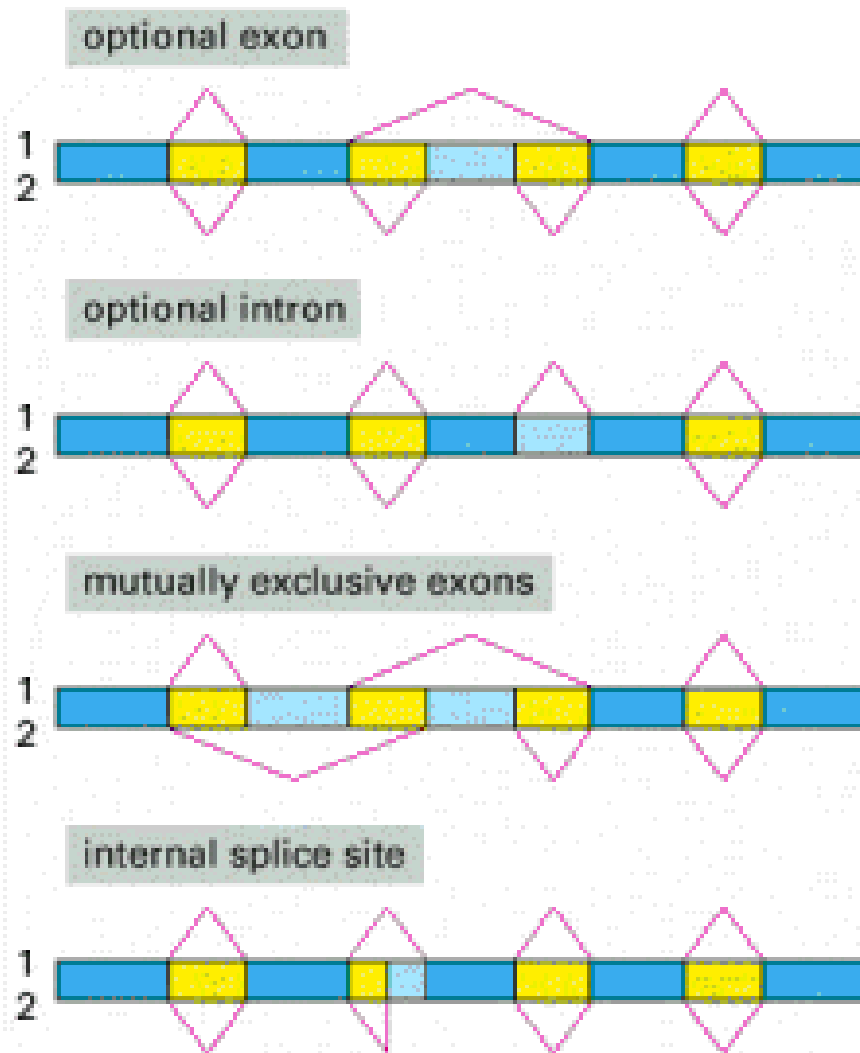
Alternative splicing

In the splicing process some of the exons can be neglected (treated as introns). In this way **from one DNA sequence ("gene") one can obtain several different mature mRNA** and thus several different proteins. This seems to be the rule: almost all the genes may have different alternative transcripts. In some case also a huge number of them.

Alternative splicing is tightly regulated: it often happens that alternative transcripts of the same gene are "alternatively expressed" in different tissues.

In most of the current genome databases the fundamental unit is not any more the gene but the **transcript**.

There are essentially four different type of splicing patterns



In the figure the dark blue boxes mark exon sequences that are retained in both mRNAs. The light blue boxes mark possible exon sequences that are included in only one of the mRNAs. The boxes are joined by red lines to indicate where intron sequences (yellow) are removed

Repeats

A significant fraction of many vertebrate chromosomes is made up of repeated DNA sequences. In human chromosomes, these repeats are mostly mutated and truncated versions of a **retrotransposon** called an **L1 element** (sometimes referred to as a **LINE** or long interspersed nuclear element). Although most copies of the L1 element are immobile, **a few retain the ability to move.**

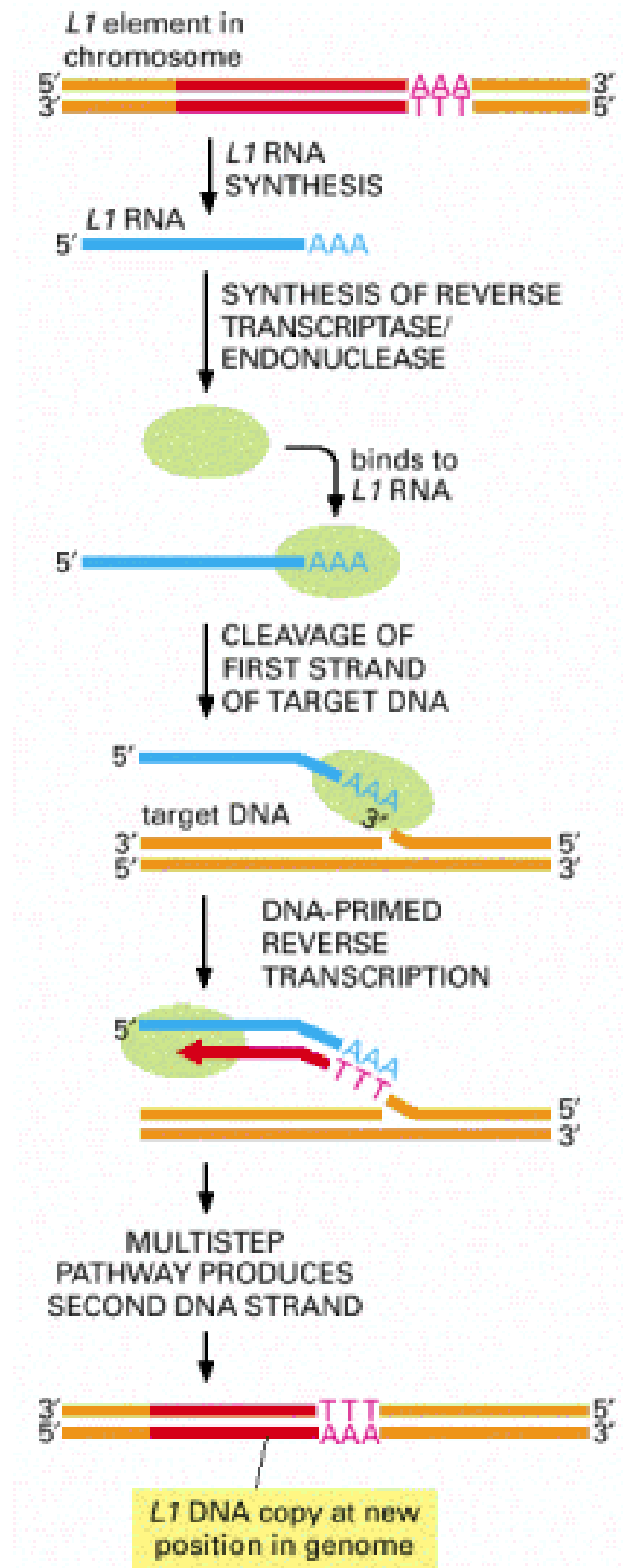
Translocations of the element have been identified, some of which result in human disease; for example, a particular type of hemophilia results from an L1 insertion into the gene encoding a blood clotting factor, Factor VIII.

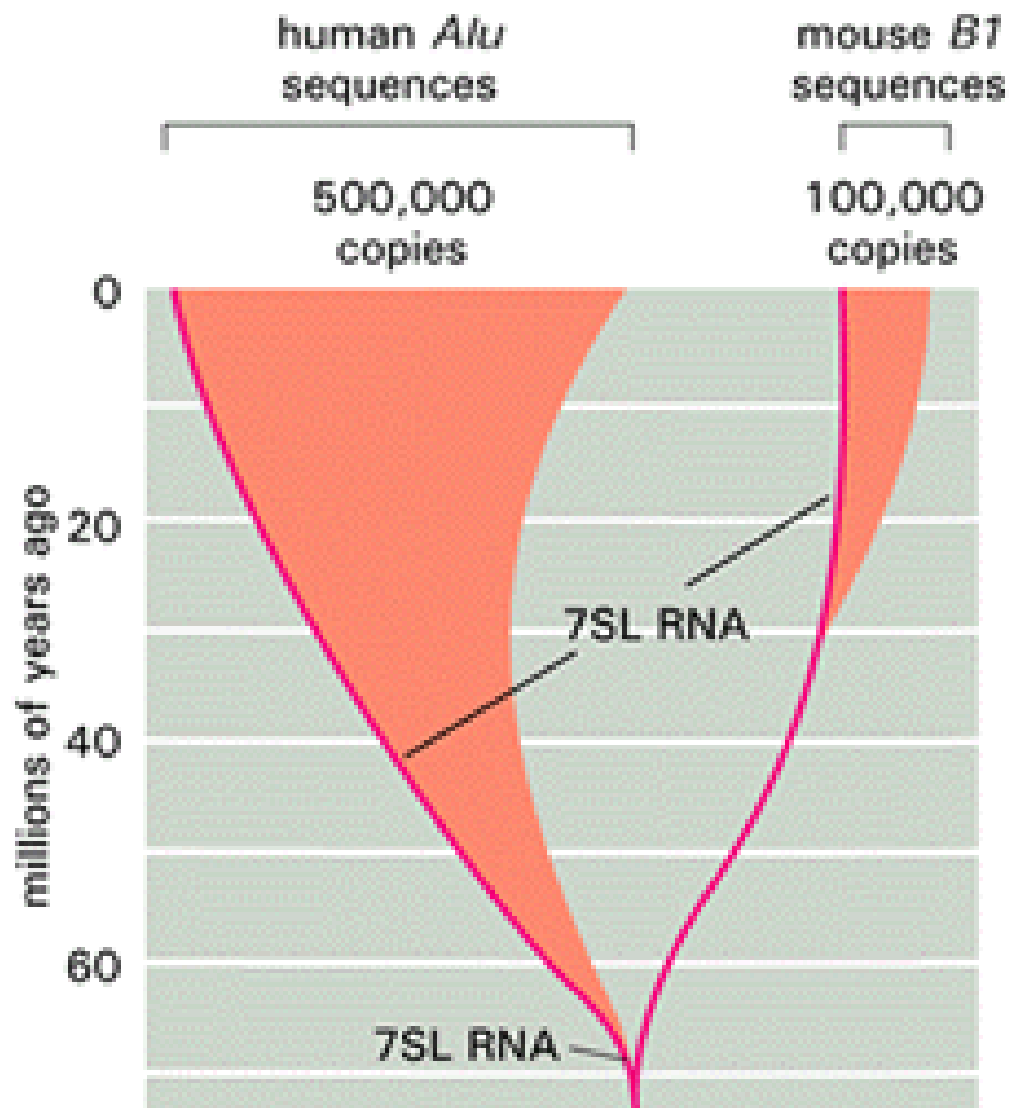
Related mobile elements are found in other mammals and insects, as well as in yeast mitochondria.

These "retrotransposons" move via a distinct mechanism that requires a **complex of an endonuclease and a reverse transcriptase**.

It is thought that other repeated DNAs that fail to encode an endonuclease or a reverse transcriptase in their own nucleotide sequence can multiply in chromosomes by a similar mechanism, using various endonucleases and reverse transcriptases present in the cell, including those encoded by L1 elements. For example, the abundant **Alu element** lacks endonuclease or reverse transcriptase genes, yet it has amplified to become a **major constituent of the human genome** .

The L1 and Alu elements seem to have multiplied in the human genome **relatively recently**. Thus, for example, the mouse contains sequences closely related to L1 and Alu, but their placement in mouse chromosomes is very different from that in human chromosomes





Gene regulation

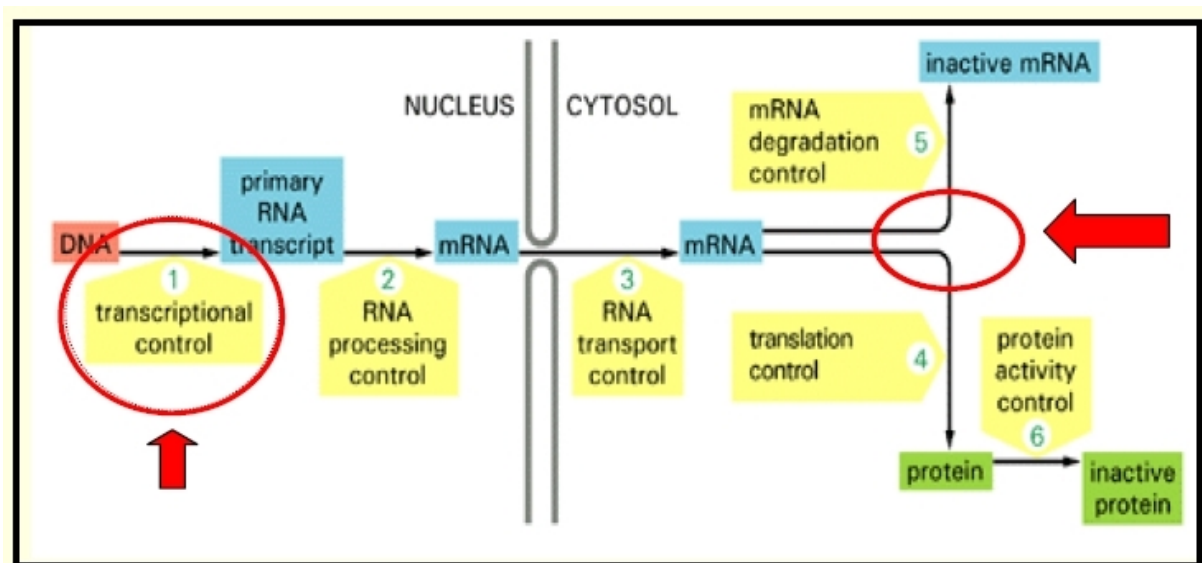
Gene expression is tightly controlled and regulated:

- All cells in the body carry the full set of genes, but only express about 20% of them at any particular time
- Different proteins are expressed in different cells (neurons, muscle cells....) according to the different functions of the cell.

As more and more complete genomes are sequenced it is becoming of crucial importance to understand **how the gene expression is regulated**.

The challenge is now to identify and fully characterize the **network of interactions among genes and their products** in an organism.

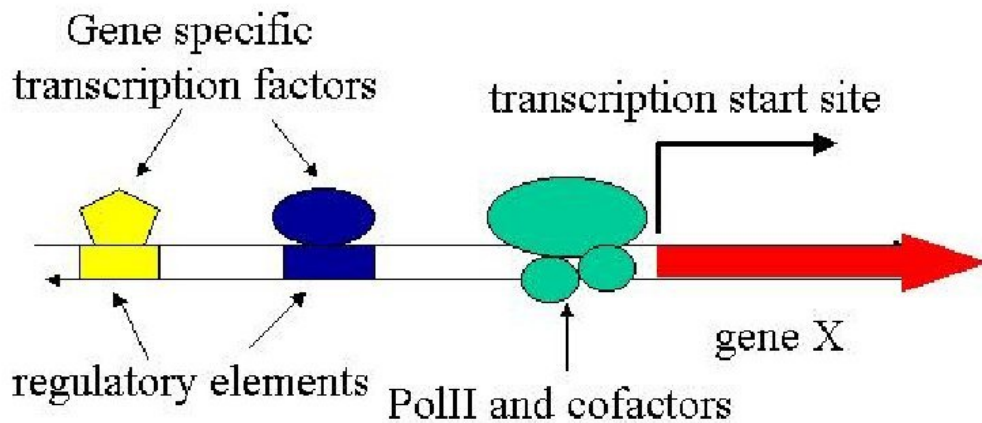
A very important example (but not the only one!) of such interactions is the transcriptional regulation of protein coding genes. This was considered up to a few years ago the main regulatory mechanism in the cell. However it has been recently realized that in higher eukaryotes a very important role is played by **miRNA** mediated **post-transcriptional** regulation.



Transcription factors.

TFs act by binding to specific, often short (5-10 bp) DNA sequences in the upstream noncoding region of genes.

Transcriptional regulation



TFs may have a twofold action on gene transcription.

They can **activate** gene transcription by:

- binding enhancer sequences in the upstream noncoding region,
- recruiting the transcription machinery to the transcription starting site

but they can also **repress** the transcription by interfering with the transcriptional apparatus.

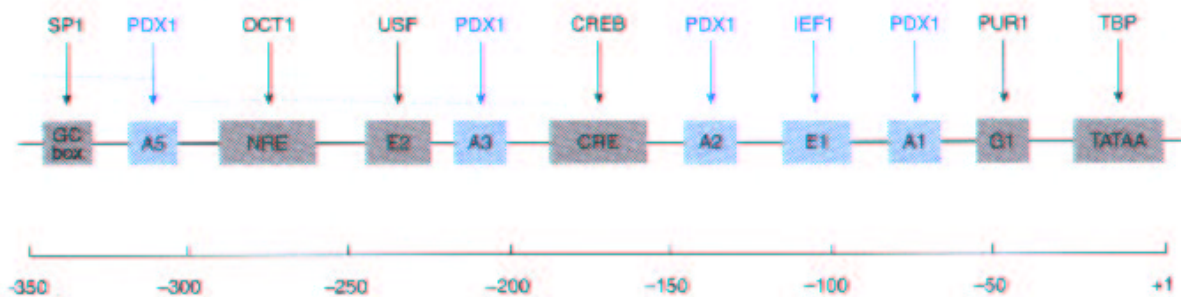
Binding sites in Eukaryotes.

There are two main classes of binding sites

- Promoters

These are localized in the region immediately upstream the coding region (often within 200 bp from the transcription starting point. They can be of two types:

- short sequences like the well known CCAAT-box, TATA-box, GC-box which are **not** tissue specific and are recognized by ubiquitous TFs
- tissue specific sequences which are only recognized by tissue specific TFs

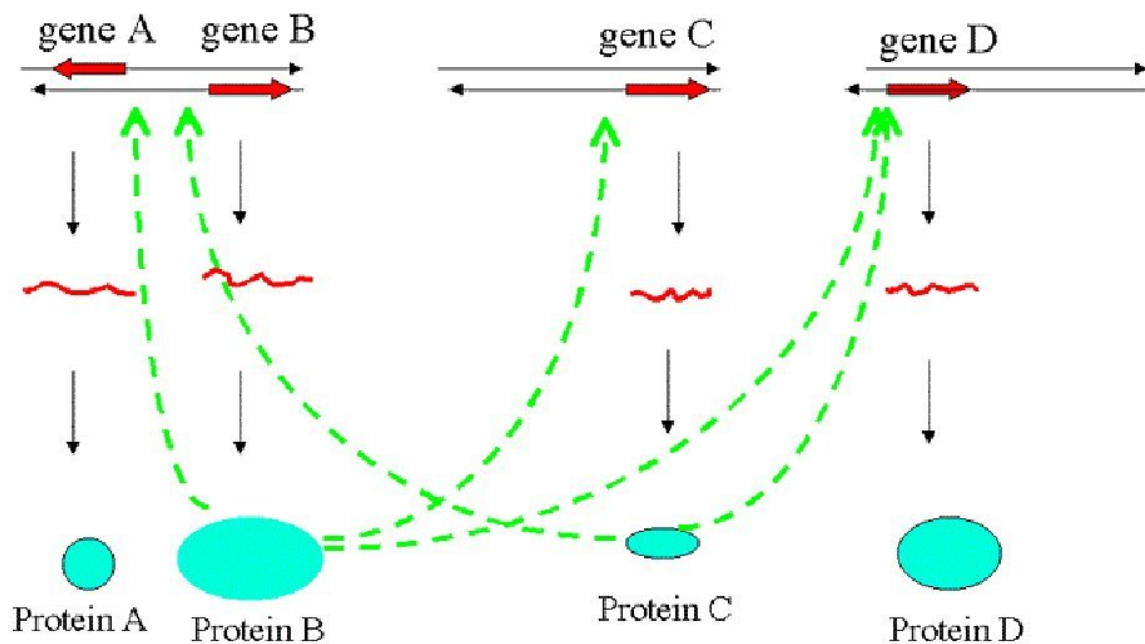


- Enhancers/Silencers

these are regulatory elements which, differently from the promoters, can act in both orientations and (to a large extent) at any distance from the transcription starting point (there are examples of enhancers located even 50-60kb upstream). They enhance (or repress) the expression of the corresponding gene.

Regulatory network

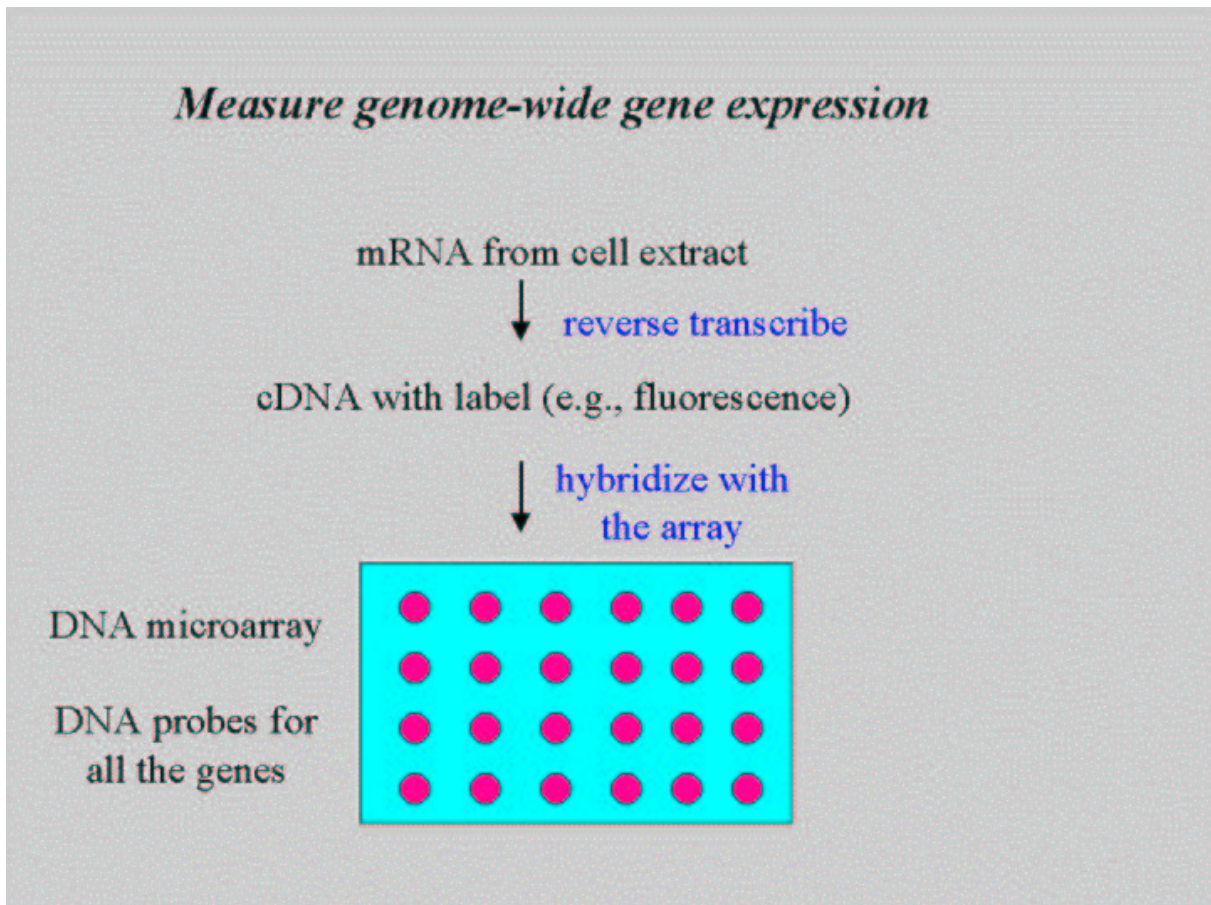
T.F.'s themselves are proteins produced by other genes.



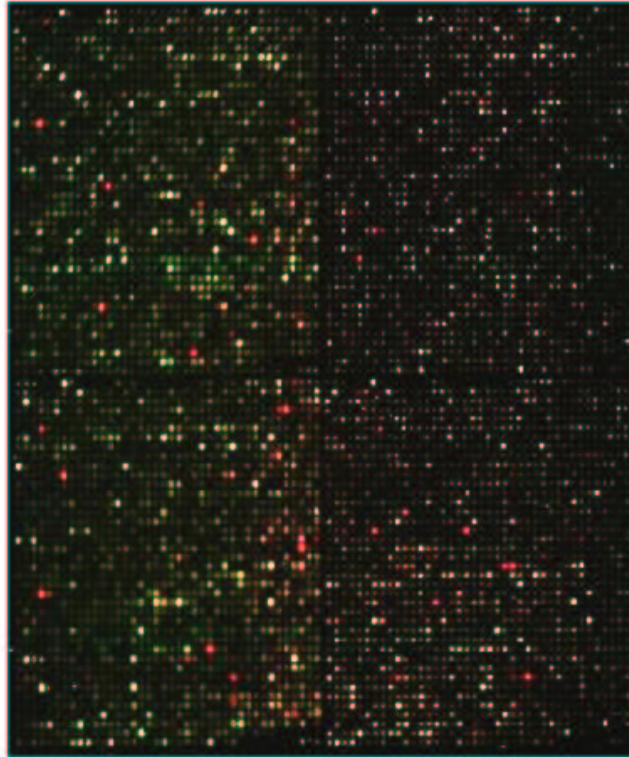
The Genome is a complex network of interactions between genes and their products **This network pattern is ubiquitous in Postgenomic biology**

DNA Microarrays

DNA microarrays can estimate **genome-wide gene expression levels** by measuring the amount of mRNA levels in the cell. Thousands of genes can be simultaneously studied in a single microchip.



The result of the experiment is a slide of this type:



The fluorescence level is proportional to the amount of mRNA produced in the experimental condition under study (usually one studies the ratio with respect to the expression level in some “reference” state of the cell).

Example : Microarray samples in *S. Cerevisiae*

The diauxic shift

DeRisi et al., Science 278 (1997) 680

- a yeast culture is inoculated into a glucose-rich medium
- rapid anaerobic growth fueled by **fermentation**, with production of ethanol, insues
- upon glucose depletion, the yest cells turn to ethanol as a carbon source for aerobic growth (**respiration**)

Expression data from DNA microarrays

- samples of cells are harvested at seven time-points during the diauxic shift
- using DNA microarray techniques **mRNA levels** for all the genes can be measured and compared to their initial values
- therefore the experiment answers the question: **which genes are switched on, and which are switched off**, as the available glucose becomes progressively scarcer?

The **output** of the experiment is, for each gene, the ratio between initial expression level and expression level at each of the seven timepoints during the diauxic shift.

References

- Alberts, B. et al. "Molecular Biology of the Cell", Paperback 1616 pages (fourth edition: March 2002) Publisher: Garland Science
online at: <http://www.ncbi.nlm.nih.gov/books/>
- Durbin R. et al. "Biological sequence analysis" Paperback 356 pages (first edition: 1998) Publisher: Cambridge University Press
- Ewens, J.W. and Grant, G.R. "Statistical methods in bioinformatics" 597 pages (second edition: 2005) Publisher: Springer
- Nature 15 Feb. 2001 "The human genome"
- Nature 5 Dec. 2002 "The mouse genome"

Databases

The existing genomic databases can be divided in two classes:

- **Primary databases:** Complete collection of DNA and RNA sequences with minimal information.
 - EMBL
 - GenBank
 - DDBJ
- **Specialized databases:** Large amount of databases (see the NAR compilation) with subsets of the existing sequences and high level of information

A selection of specialized databases:

- Protein database:

Swissprot <http://www.expasy.ch/swissprot>

- A database of functional annotations:

Gene Ontology <http://www.geneontology.org>

- Transcription factors:

Transfac <http://transfac.gbf.de/TRANSFAC>

- Untranscribed regions:

UTRdb <http://bighost.area.br.cnr.it/BIG/UTRHome>

- Single Nucleotide Polymorphism:

dbSNP <http://www.ncbi.nlm.nih.gov/SNP>

- Metabolic pathways

KEGG <http://www.genome.ad.jp/kegg/kegg2.html>

"Organism oriented" specialized databases

- Human:

GDB <http://www.gdb.org>

- D. melanogaster

FLYBASE <http://www.flybase.org>

- Mouse

MGD <http://www.informatics.jax.org>

- S. cerevisiae

SGD <http://genome-www.stanford.edu/Saccharomyces>

” Reference databases”

- **Ensembl** <http://www.ensembl.org>
- **NCBI** <http://www.ncbi.nlm.nih.gov/>
- **UCSC** <http://genome.ucsc.edu/>

Preprints

Since September 2003 in the ArXive web page you can find (just below hep-th, cond-mat etc..) a preprint archive devoted to “quantitative biology”. The name is: [q-bio](#).

The link is <http://xxx.lanl.gov/archive/q-bio>

Papers

Main source for published papers (contains more than $16 * 10^6$ papers): [PubMed](#)

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>

Congress

- Intelligent Systems for Molecular Biology
[ISMB 2006](#), Fortaleza (Brasil) August 6-10 2006
<http://ismb2006.cbi.cnptia.embrapa.br/>
- European Conference on Computational Biology
[ECCB 2006](#), Eliat (Israel) September 10-13 2006
<http://www.eccb06.org/>

- Research in Computational Biology

RECOMB 2006, Venice (Italy) April 2-5 2006

<http://recomb06.dei.unipd.it/>

Topics:

- Genomics
- Molecular sequence analysis
- Recognition of genes and regulatory elements
- Molecular evolution
- Protein structure
- Structural genomics
- Gene Expression
- Gene Networks
- Drug Design
- Combinatorial libraries
- Computational proteomics
- Structural and functional genomics

[arXiv.org](#) > [q-bio](#)

Search for

([Help](#) | [Advanced search](#))

Quantitative Biology (since 9/03)

- [search](#) q-bio titles/authors or full-text
- [get](#) q-bio/abstract if you know the paper number
- e-Prints are available for years:
[2006](#) [2005](#) [2004](#) [2003](#) [2002](#) [2001](#) [2000](#) [1999](#) [1998](#) [1997](#) [1996](#) [1995](#) [1994](#) [1993](#)
[1992](#)

Subject Classes

- BM - Biomolecules ([new](#), [recent](#), [find](#))
- CB - Cell Behavior ([new](#), [recent](#), [find](#))
- GN - Genomics ([new](#), [recent](#), [find](#))
- MN - Molecular Networks ([new](#), [recent](#), [find](#))
- NC - Neurons and Cognition ([new](#), [recent](#), [find](#))
- OT - Other ([new](#), [recent](#), [find](#))
- PE - Populations and Evolution ([new](#), [recent](#), [find](#))
- QM - Quantitative Methods ([new](#), [recent](#), [find](#))
- SC - Subcellular Processes ([new](#), [recent](#), [find](#))
- TO - Tissues and Organs ([new](#), [recent](#), [find](#))

Additional:

- [new](#) q-bio papers received (most recent mailing)
- [recent](#) q-bio listings
- [current](#) month's q-bio listings
- [lastupdate](#) of daily changes to q-bio database (ftp format)
- some [info](#) for q-bio

Links to: [arXiv](#), [form interface](#), [q-bio](#), [/find](#), [/abs](#), [/0603](#), [help](#)
