Introduction to Genome Biology

Sandrine Dudoit and Robert Gentleman

Bioconductor Short Course

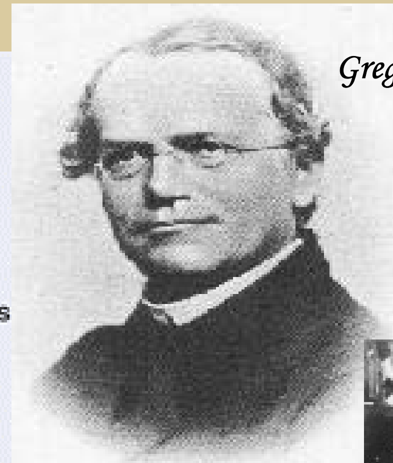
2003

Outline

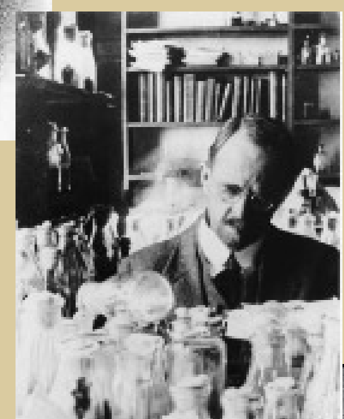
- Cells, chromosomes, and cell division
- DNA structure and replication
- Proteins
- Central dogma: transcription, translation
- Microarrays
- Pathways

A brief history

1865	Genes are particulate factors
1903	Chromosomes are hereditary units
1910	Genes lie on chromosomes
1913	Chromosomes contain linear arrays of genes
1927	Mutations are physical changes in genes
1931	Recombination is caused by crossing over
1944	DNA is the genetic material
1945	A gene codes for a protein
1953	DNA is a double helix
1958	DNA replicates semiconservatively
1961	Genetic code is triplet
1977	DNA can be sequenced
1997	Genomes can be sequenced

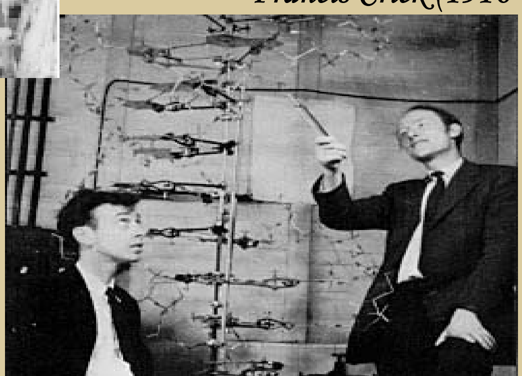


Gregor Mendel (1823-1884)



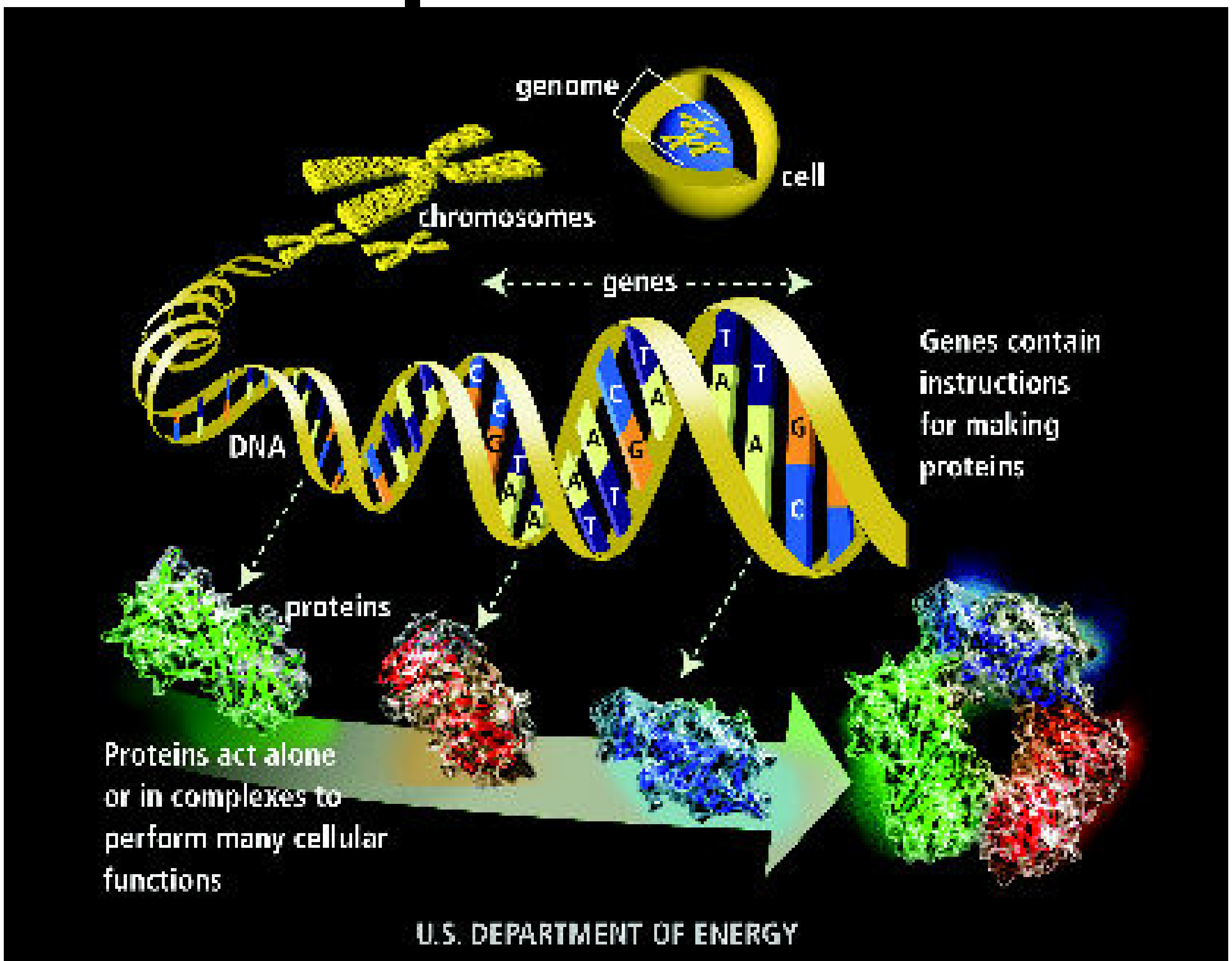
Thomas Hunt Morgan (1866-1945)

Francis Crick (1916-)

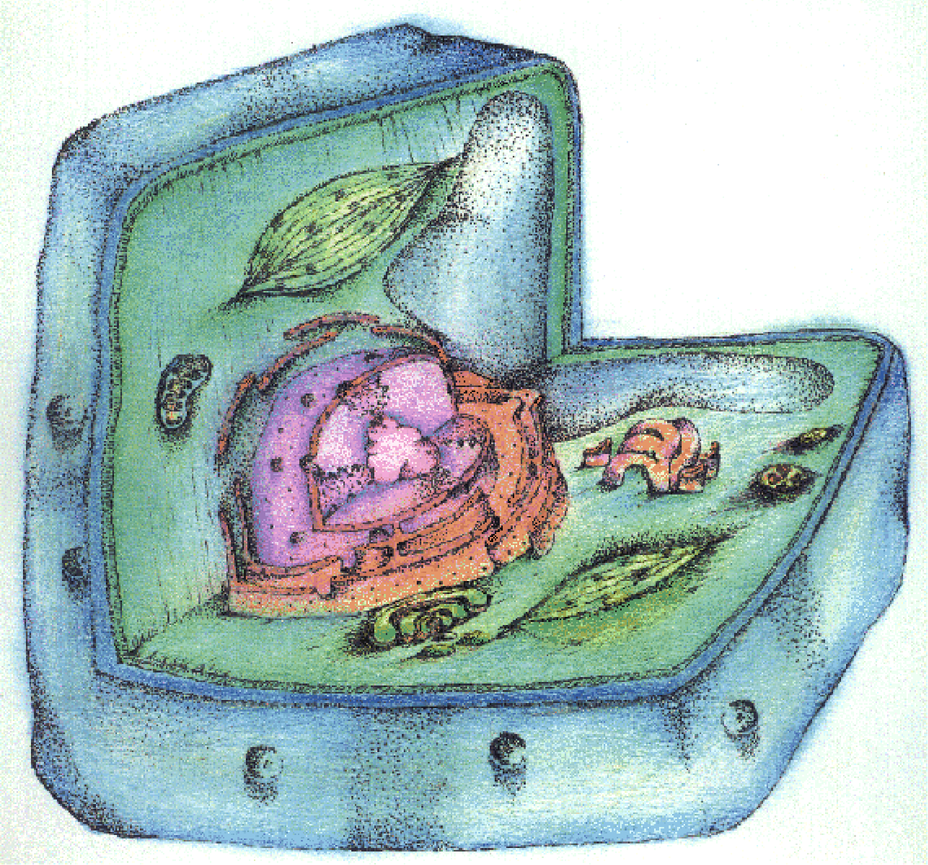


James D. Watson (1928-)

From chromosomes to proteins



Cells



Cells

- **Cells**: the fundamental working units of every living organism.

- **Metazoa**: multicellular organisms. E.g. humans: trillions of cells.
- **Protozoa**: unicellular organisms. E.g. yeast, bacteria.

Cells

- Each cell contains a complete copy of an organism's **genome**, or blueprint for all cellular structures and activities.

- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.

Cell composition

- 90% water.
- Of the remaining molecules, dry weight
  - 50% protein
  - 15% carbohydrate
  - 15% nucleic acid
  - 10% lipid
  - 10% miscellaneous.
- By element: 60% H, 25% O, 12%C, 5%N.

The genome

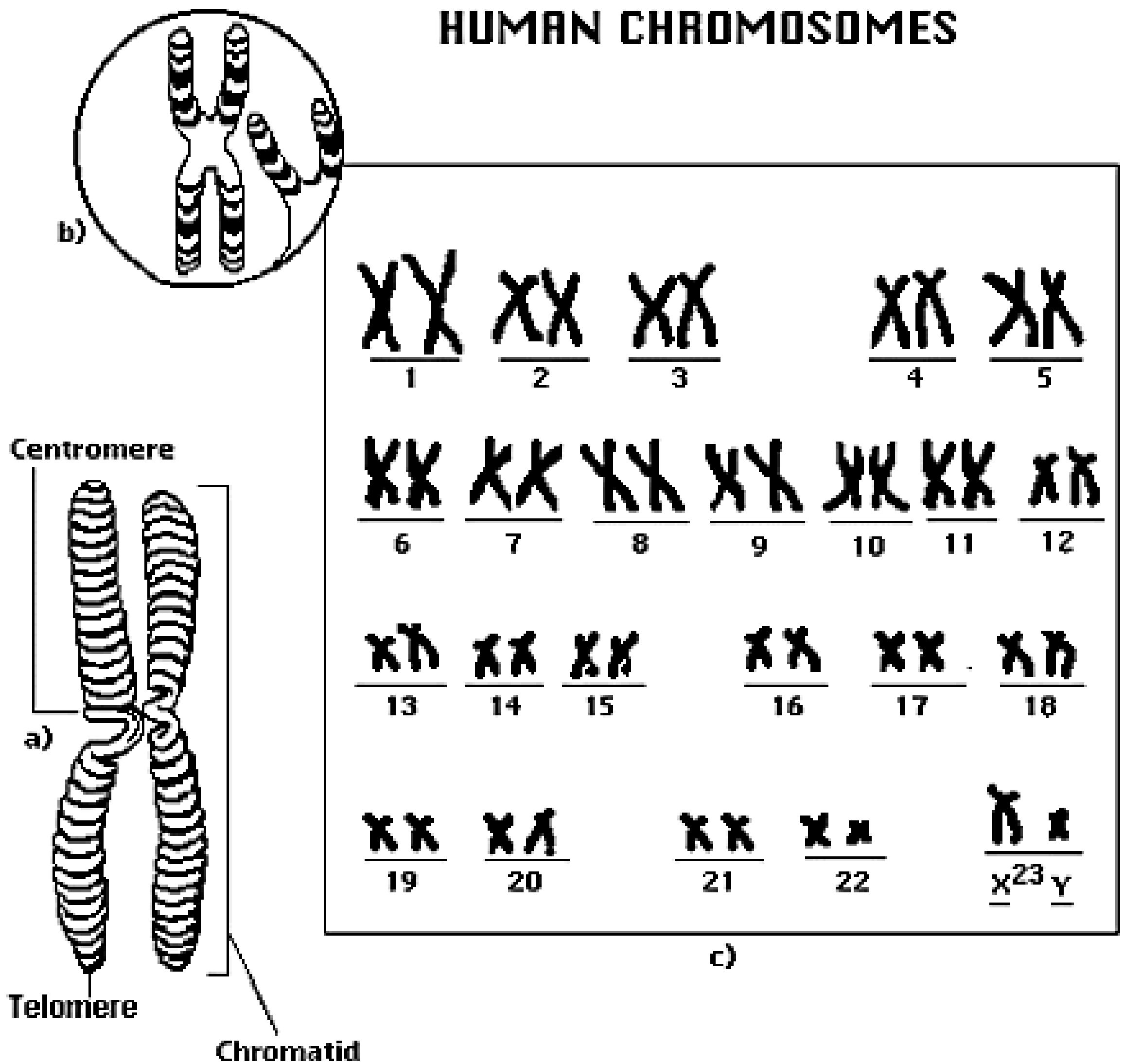
- The genome is distributed along **chromosomes**, which are made of compressed and entwined **DNA**.

- A (protein-coding) **gene** is a segment of chromosomal **DNA** that directs the synthesis of a **protein**.

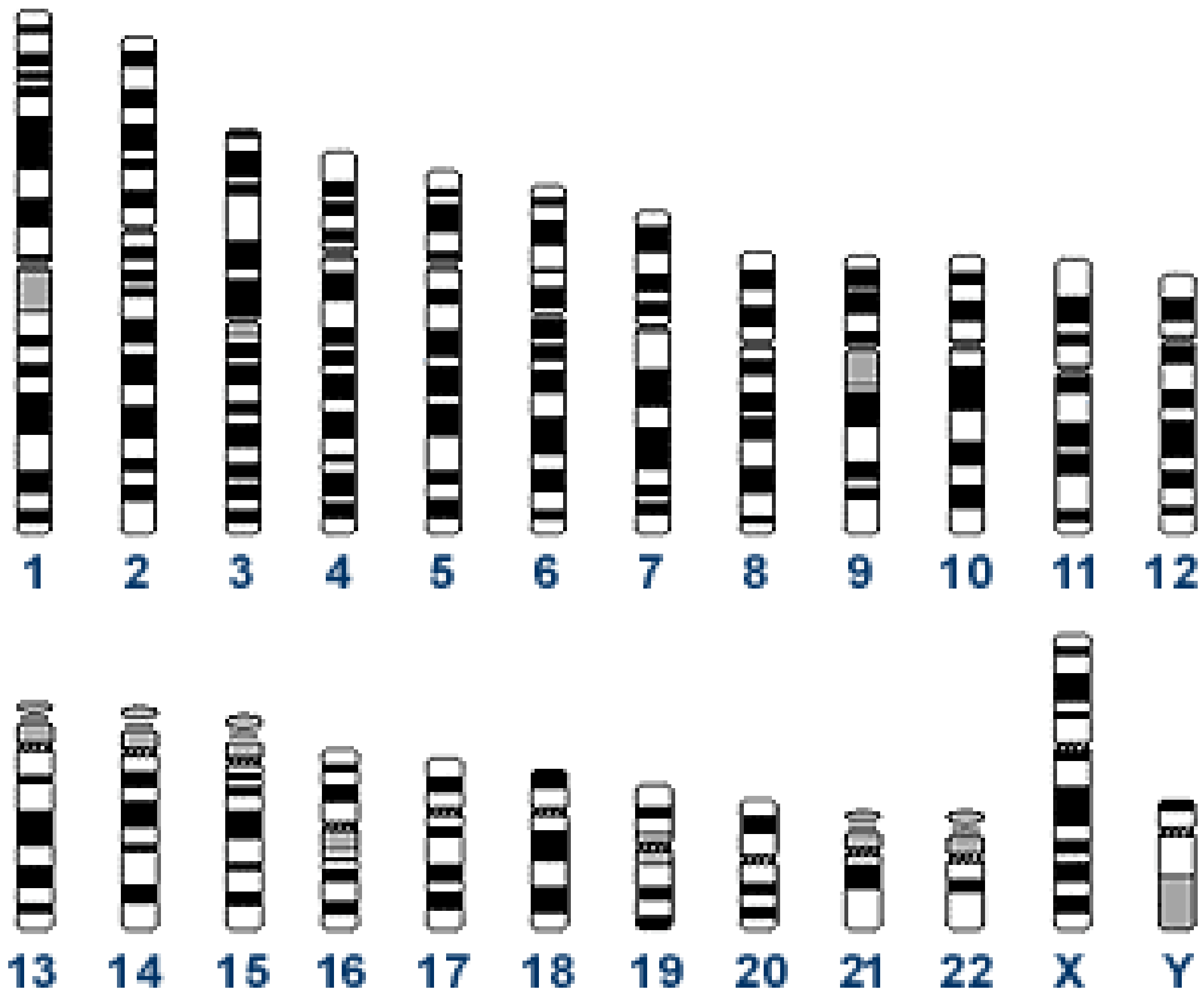
The human genome

- The human genome is distributed along **23 pairs of chromosomes**
  - 22 autosomal pairs;
  - the sex chromosome pair, **XX** for females and **XY** for males.
- In each pair, one chromosome is paternally inherited, the other maternally inherited (cf. meiosis).

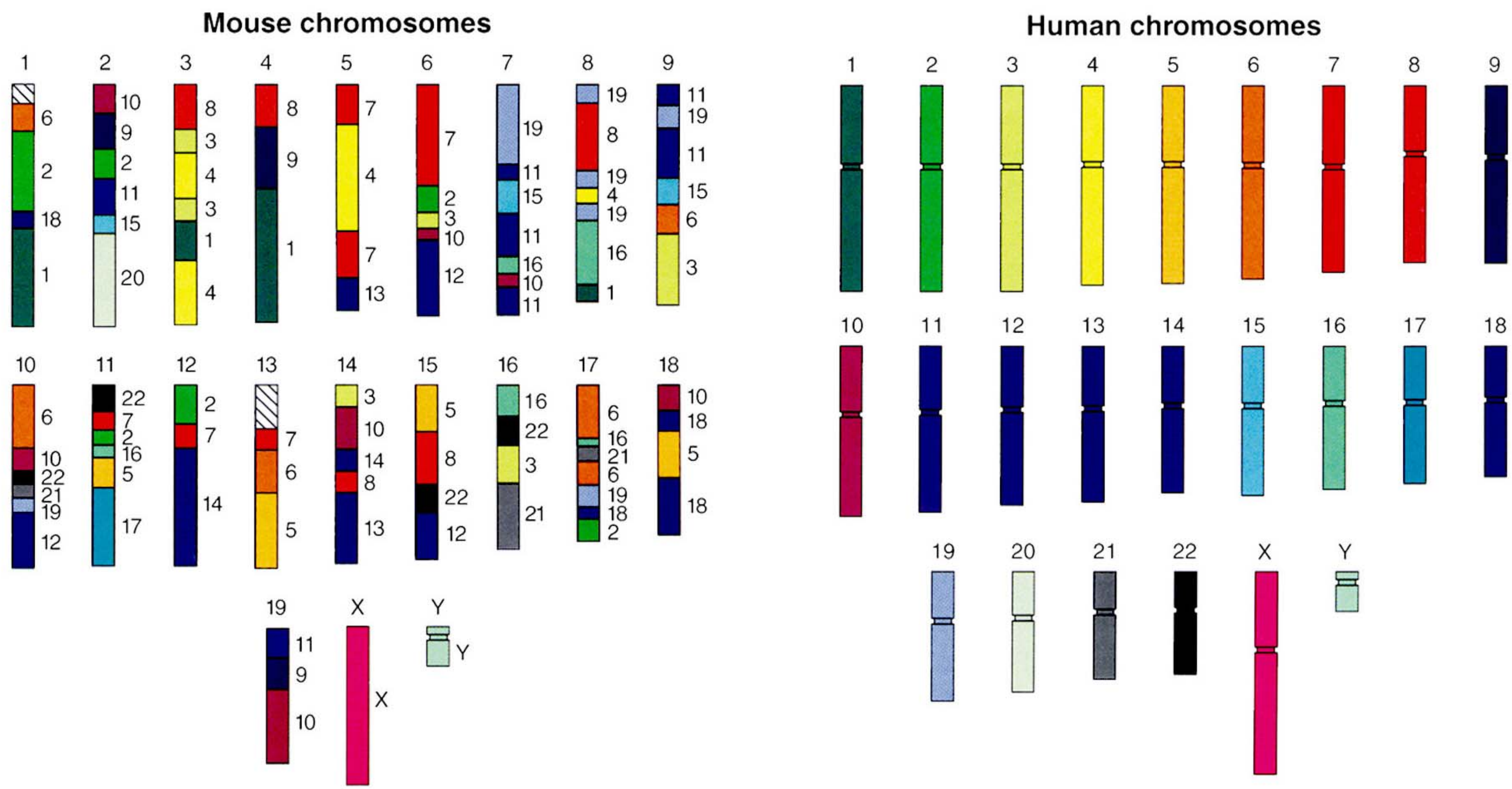
Chromosomes



Chromosome banding patterns



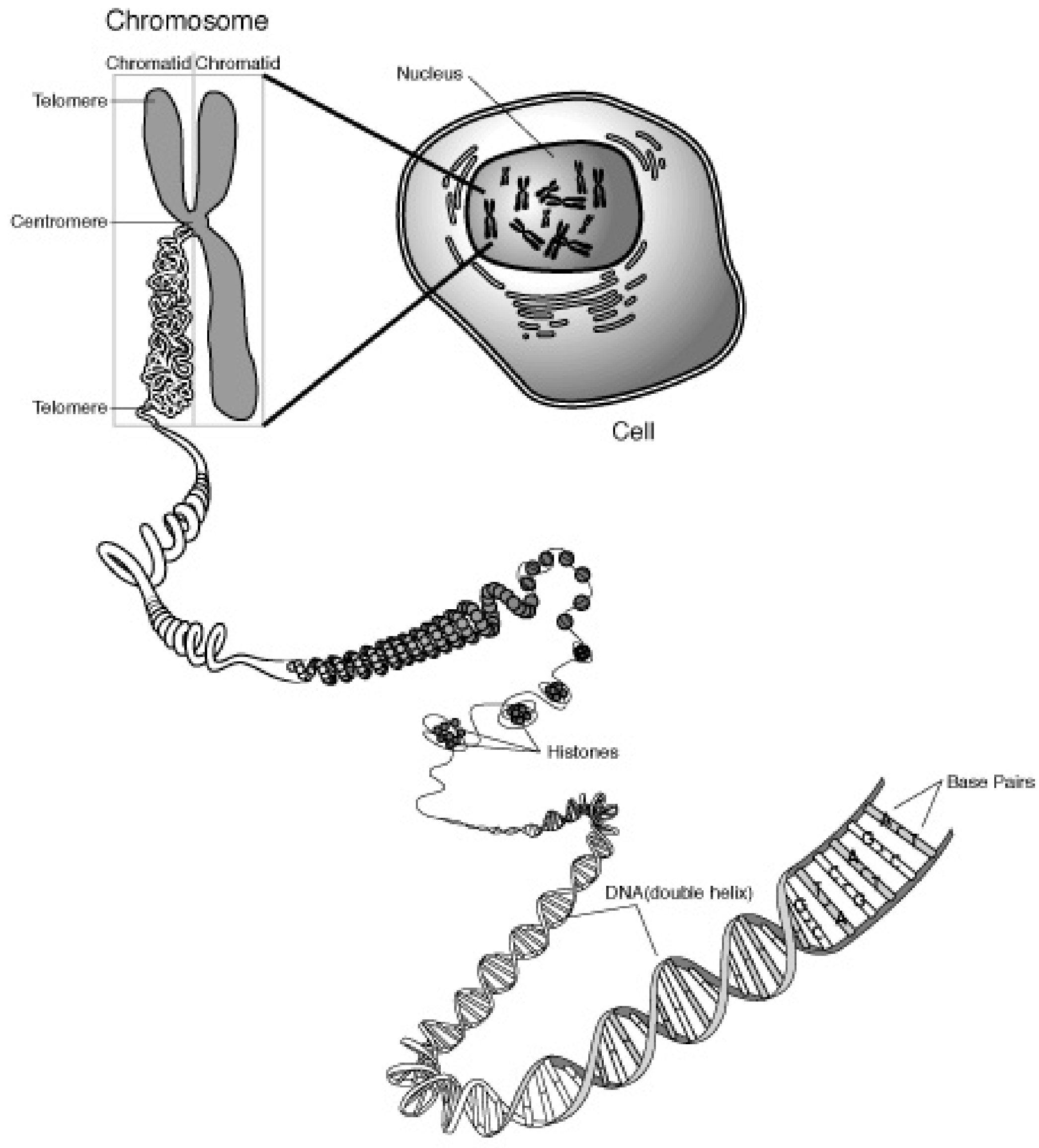
Of mice and men



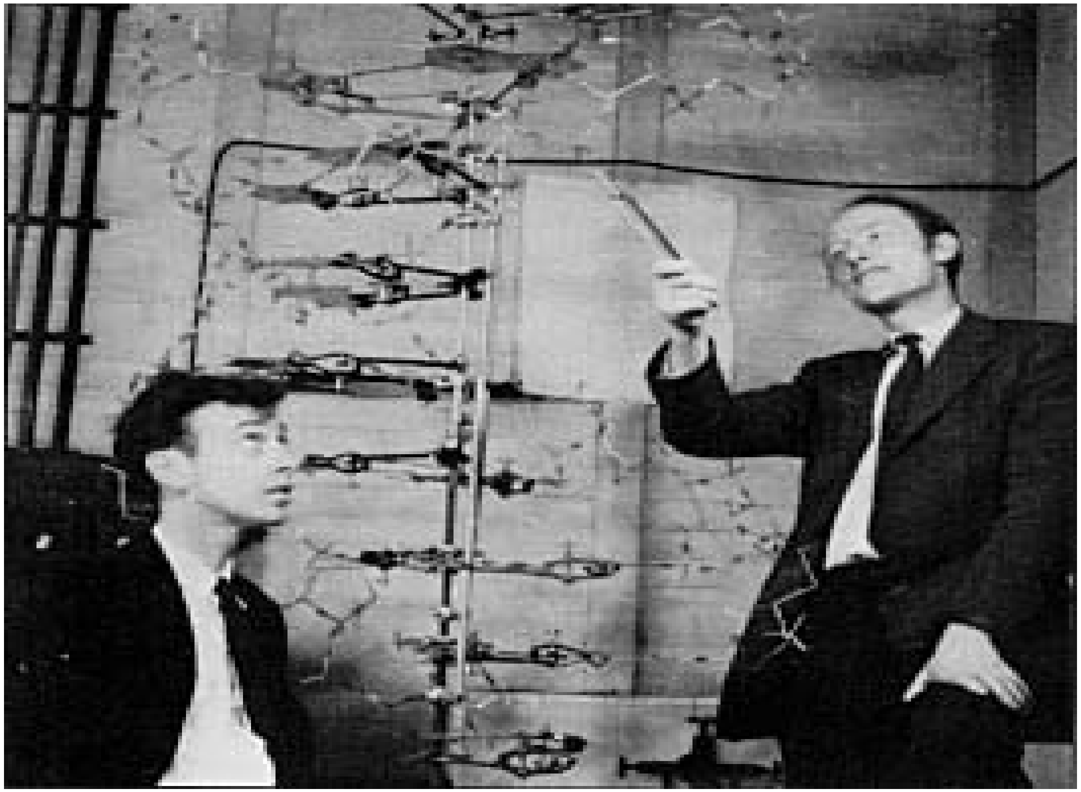
Courtesy Lisa Stubbs
Oak Ridge National Laboratory

YGA 98-075R2

Chromosomes and DNA



DNA structure



"We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest."

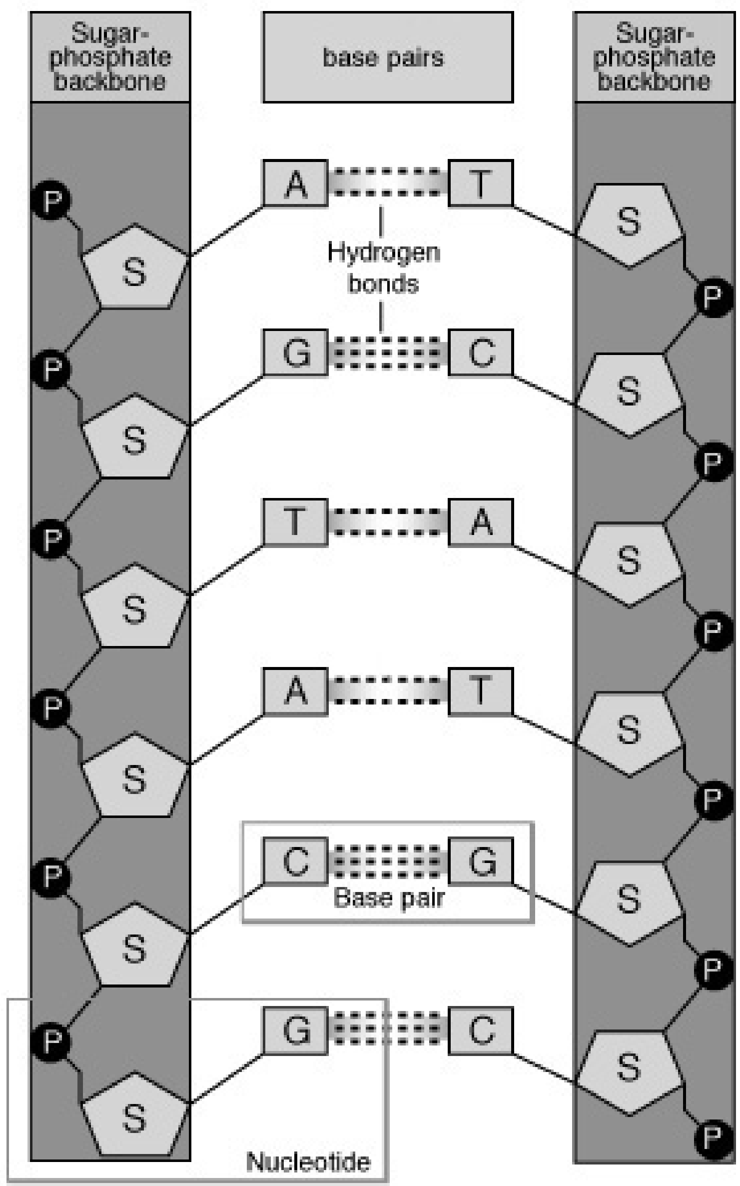
J.D. Watson & F. H. C. Crick. (1953). Molecular structure of Nucleic Acids. *Nature.* **171**: 737-738.

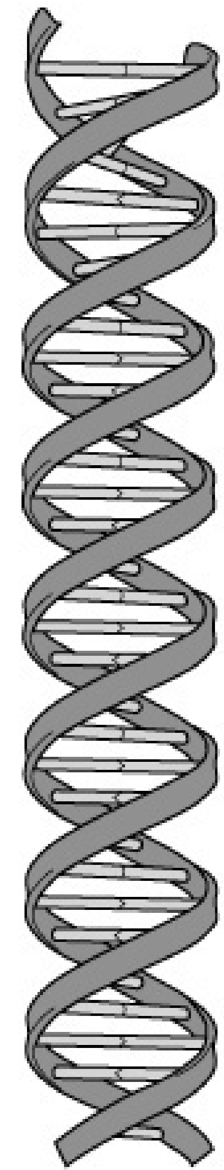
DNA structure

- A **deoxyribonucleic acid** or **DNA** molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
- Each **nucleotide** comprises
  - a phosphate group;
  - a deoxyribose sugar;
  - one of four nitrogen bases:
    - purines: **adenine (A)** and **guanine (G)**,
    - pyrimidines: **cytosine (C)** and **thymine (T)**.

DNA structure

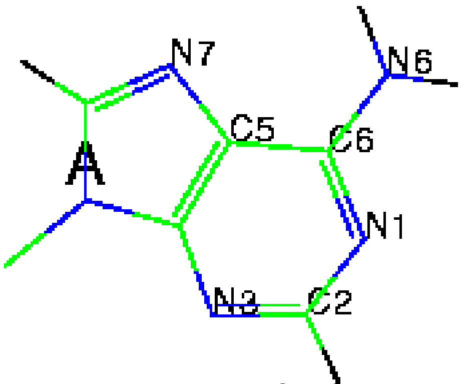
DNA structure



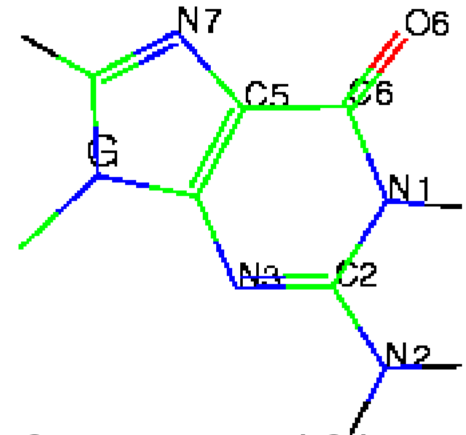


Nucleotide bases

Purines

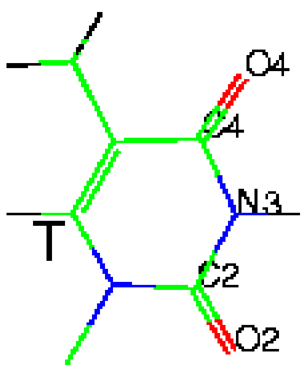


Adenine (A)

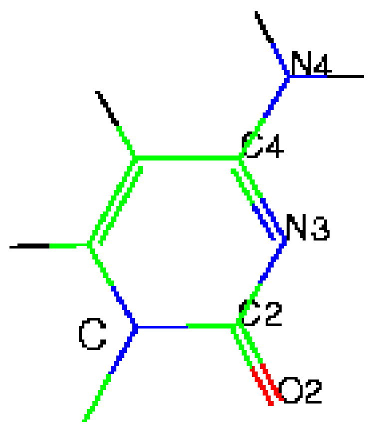


Guanine (G)

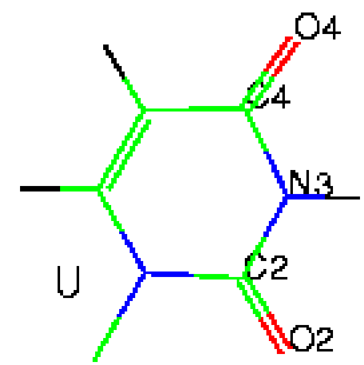
Pyrimidines



Thymine (T)
(DNA)



Cytosine (C)



Uracil (U)
(RNA)

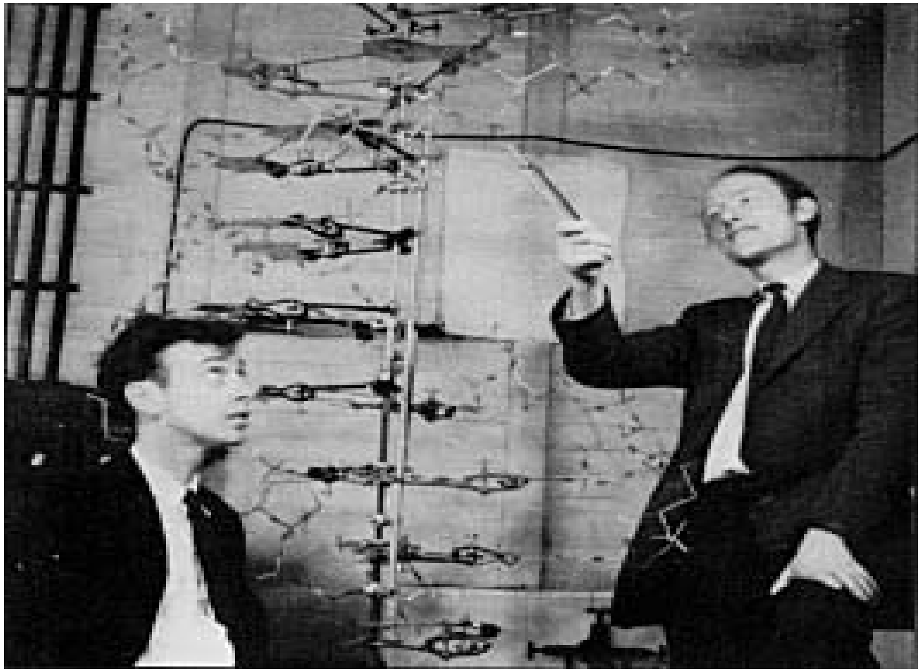
DNA structure

- Polynucleotide chains are **directional** molecules, with slightly different structures marking the two ends of the chains, the so-called **3' end** and **5' end**.
- The 3' and 5' notation refers to the numbering of carbon atoms in the sugar ring.
- The 3' end carries a sugar group and the 5' end carries a phosphate group.
- The two complementary strands of DNA are **antiparallel** (i.e, 5' end to 3' end directions for each strand are opposite)

The human genome in numbers

- 23 pairs of chromosomes;
- 2 meters of DNA;
- 3,000,000,000 bp;
- 35 M (males 27M, females 44M);
- 30,000-40,000 genes.

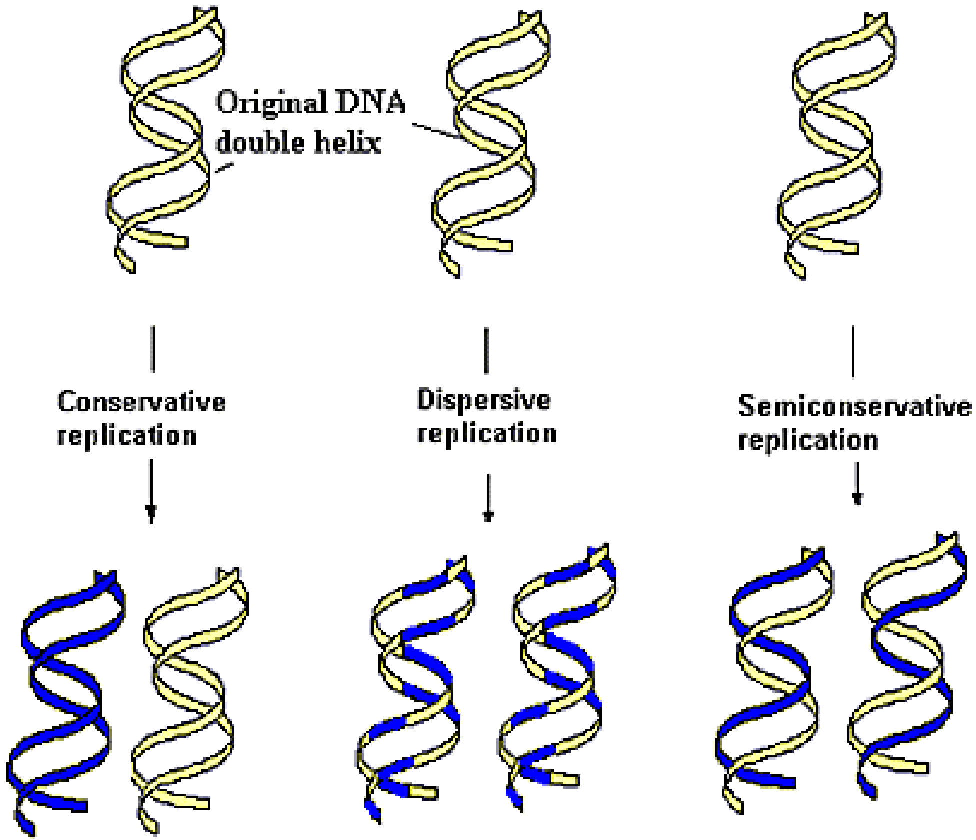
DNA replication



"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

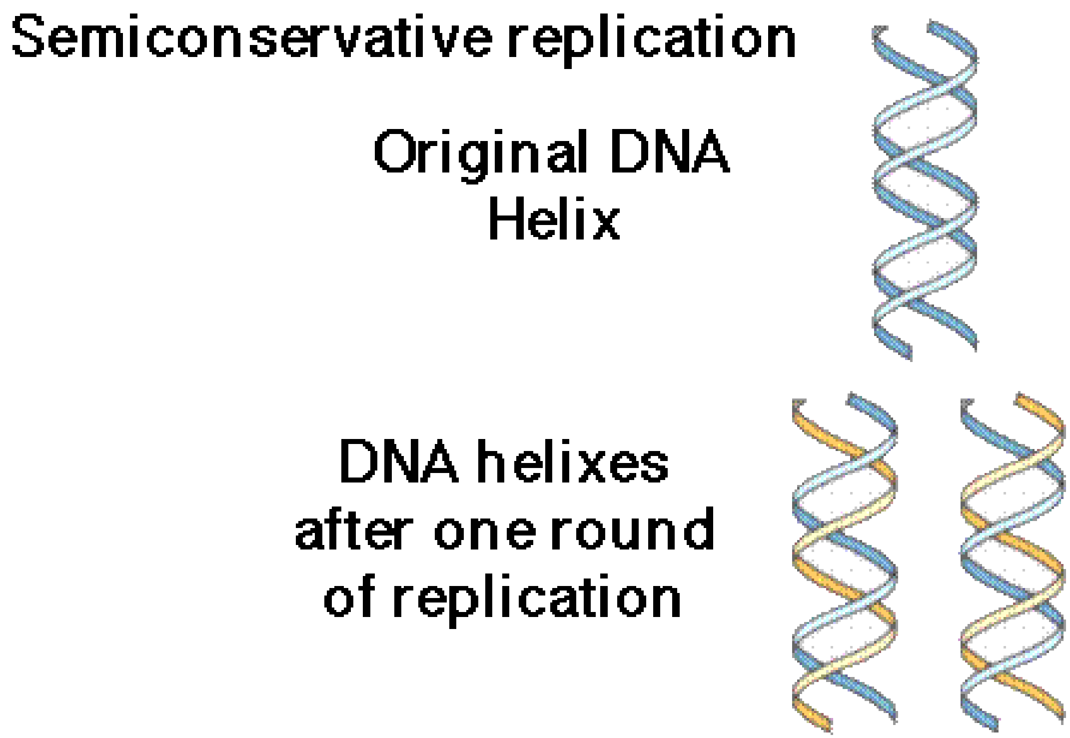
J.D. Watson & F. H. C. Crick. (1953). Molecular structure of Nucleic Acids. *Nature.* **171**: 737-738.

DNA replication



Three possible models

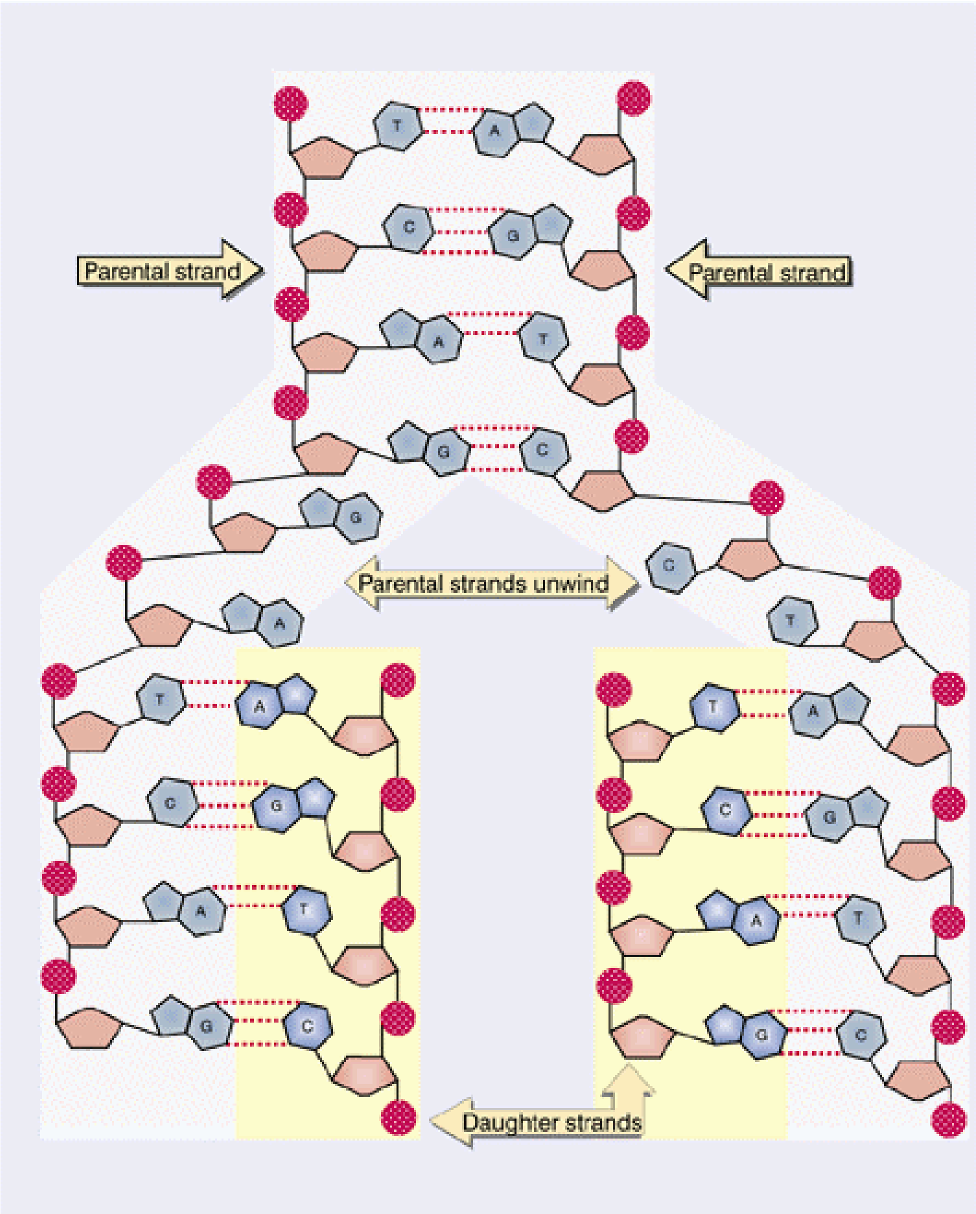
DNA replication



DNA replication

- In the replication of a double-stranded or duplex DNA molecule, **both** parental (i.e. original) DNA strands are copied.
- The parental DNA strand that is copied to form a new strand is called a **template**.
- When copying is finished, the two new duplexes each consist of one of the original strands plus its complementary copy - **semiconservative** replication.

DNA replication

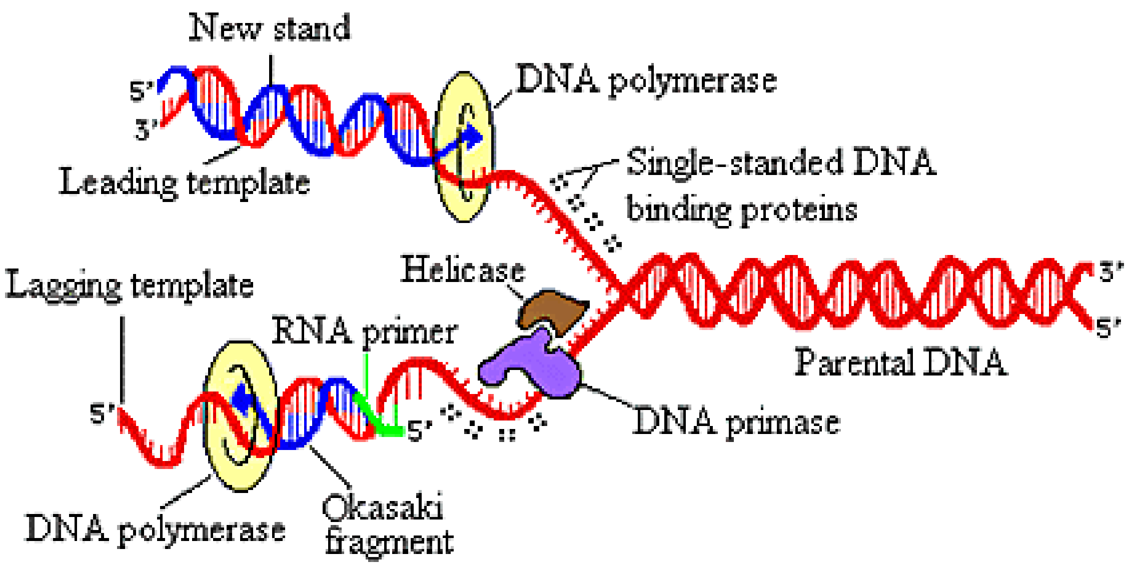


Base pairing provides the mechanism for DNA replication.

DNA replication

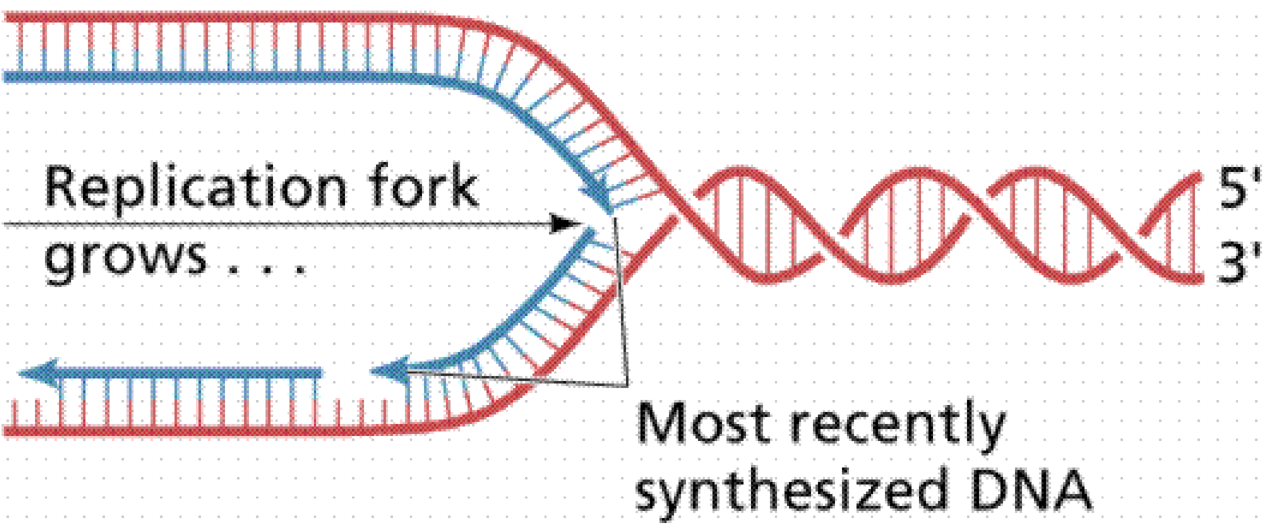
- Many **enzymes** are required to unwind the double helix and to synthesize a new strand of DNA.
- The unwound helix, with each strand being synthesized into a new double helix, is called the **replication fork**.
- DNA synthesis occurs in the **5' → 3'** direction.

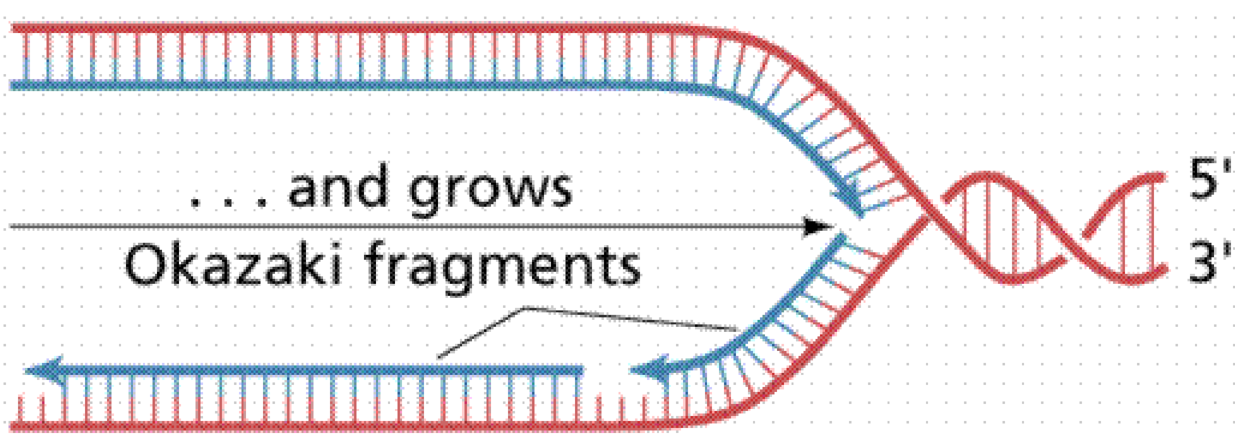
DNA replication



Collaboration of Proteins at the Replication Fork

DNA replication

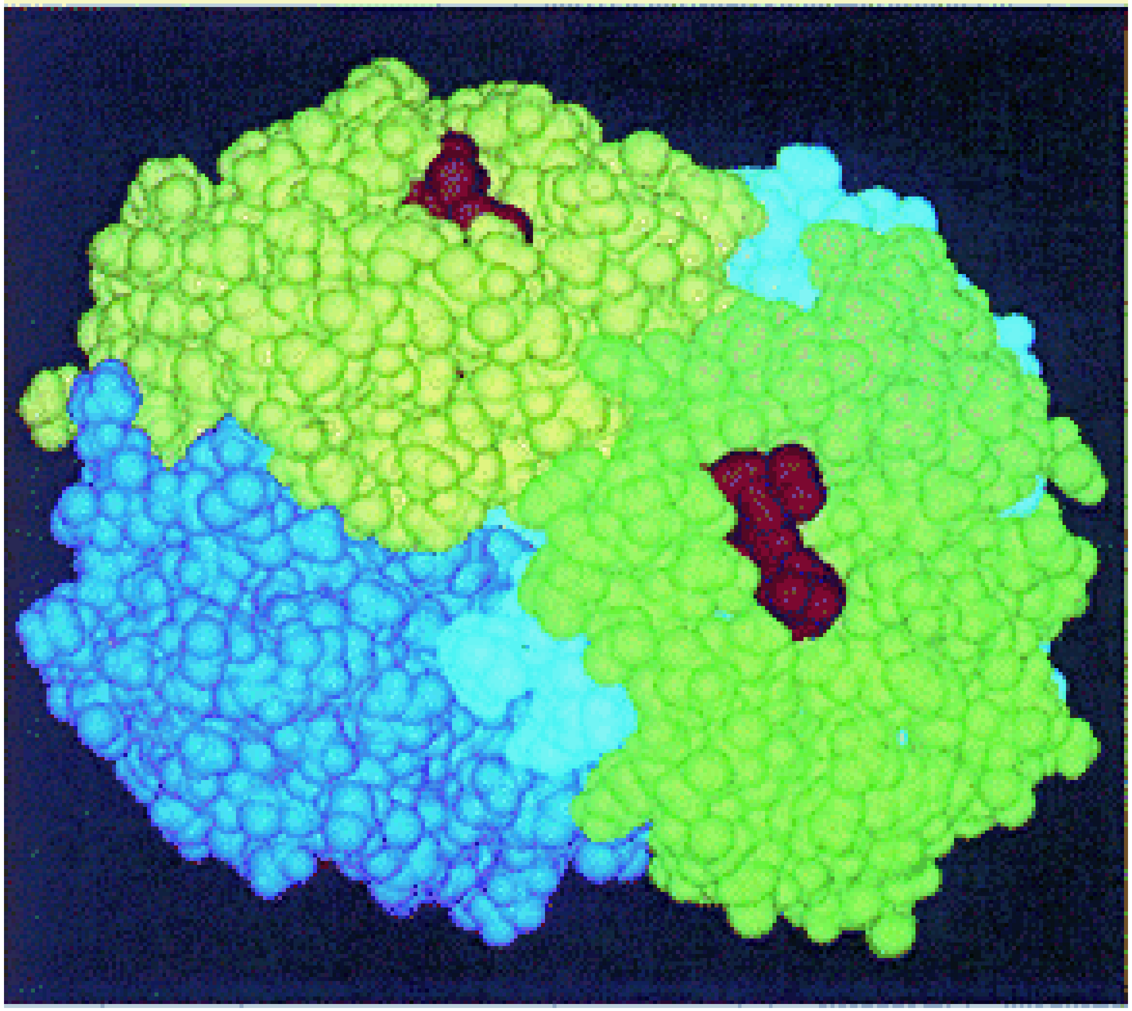




Enzymes in DNA replication

1. **Topoisomerase**: removes supercoils and initiates duplex unwinding.
2. **Helicase**: unwinds duplex.
3. **DNA polymerase**: synthesizes the new DNA strand; also performs proofreading.
4. **Primase**: attaches small RNA primer to single-stranded DNA to act as a substitute 3'OH for DNA polymerase to begin synthesizing from.
5. **Ligase**: catalyzes the formation of phosphodiester bonds.
6. **Single-stranded binding proteins**: maintain the stability of the replication fork.

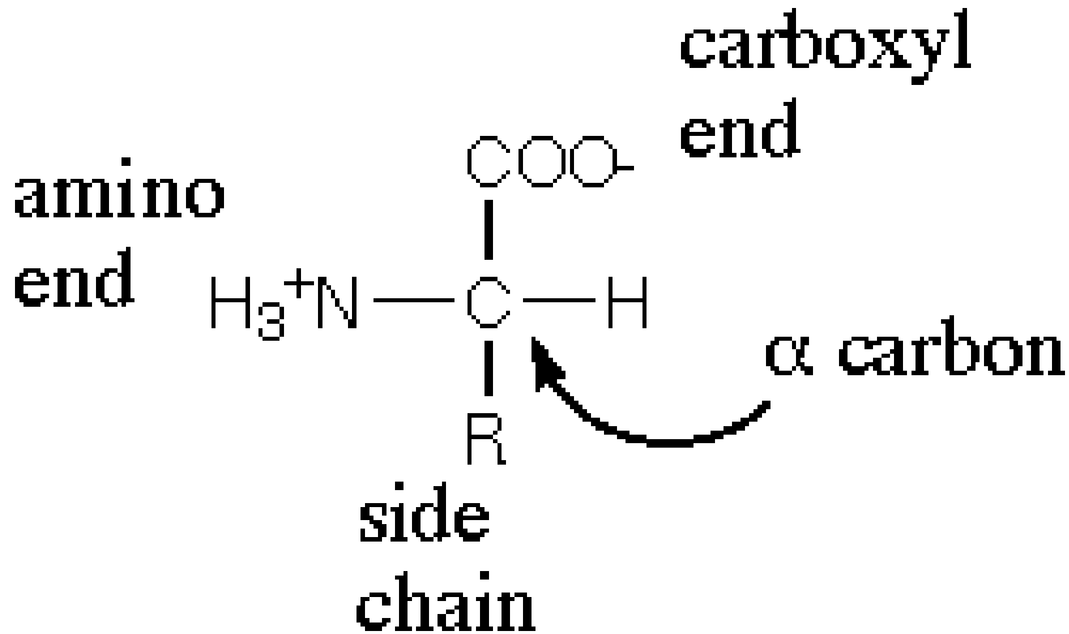
Proteins



Proteins

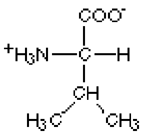
- **Proteins:** large molecules composed of one or more chains of amino acids, **polypeptides**.
- **Amino acids:** class of 20 different organic compounds containing a basic amino group ($-NH_2$) and an acidic carboxyl group ($-COOH$).
- The order of the amino acids is determined by the **base sequence** of nucleotides in the **gene** coding for the protein.
- E.g. hormones, enzymes, antibodies.

Amino acids

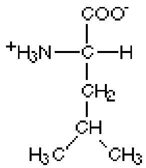


Amino acids

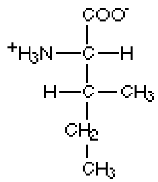
Amino acids with hydrophobic side groups



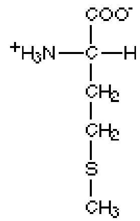
Valine
(val)



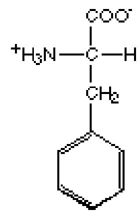
Leucine
(leu)



Isoleucine
(ile)

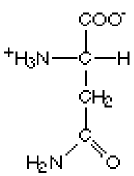


Methionine
(met)

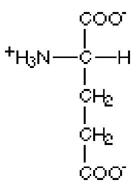


Phenylalanine
(phe)

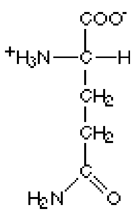
Amino acids with hydrophilic side groups



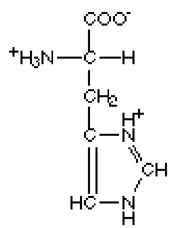
Asparagine
(asn)



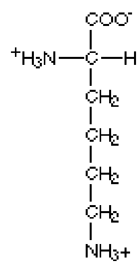
Glutamic acid
(glu)



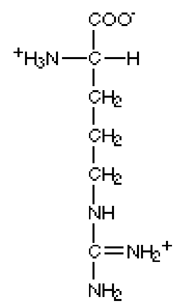
Glutamine
(gln)



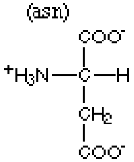
Histidine
(his)



Lysine
(lys)

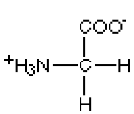


Arginine
(arg)

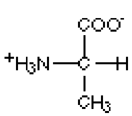


Aspartic acid
(asp)

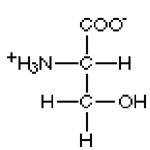
Amino acids that are in between



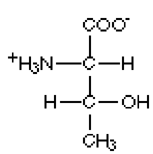
Glycine
(gly)



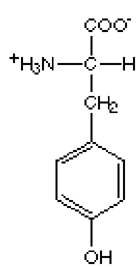
Alanine
(ala)



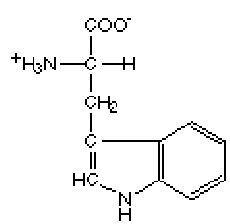
Serine
(ser)



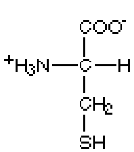
Threonine
(thr)



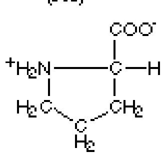
Tyrosine
(tyr)



Tryptophan
(trp)

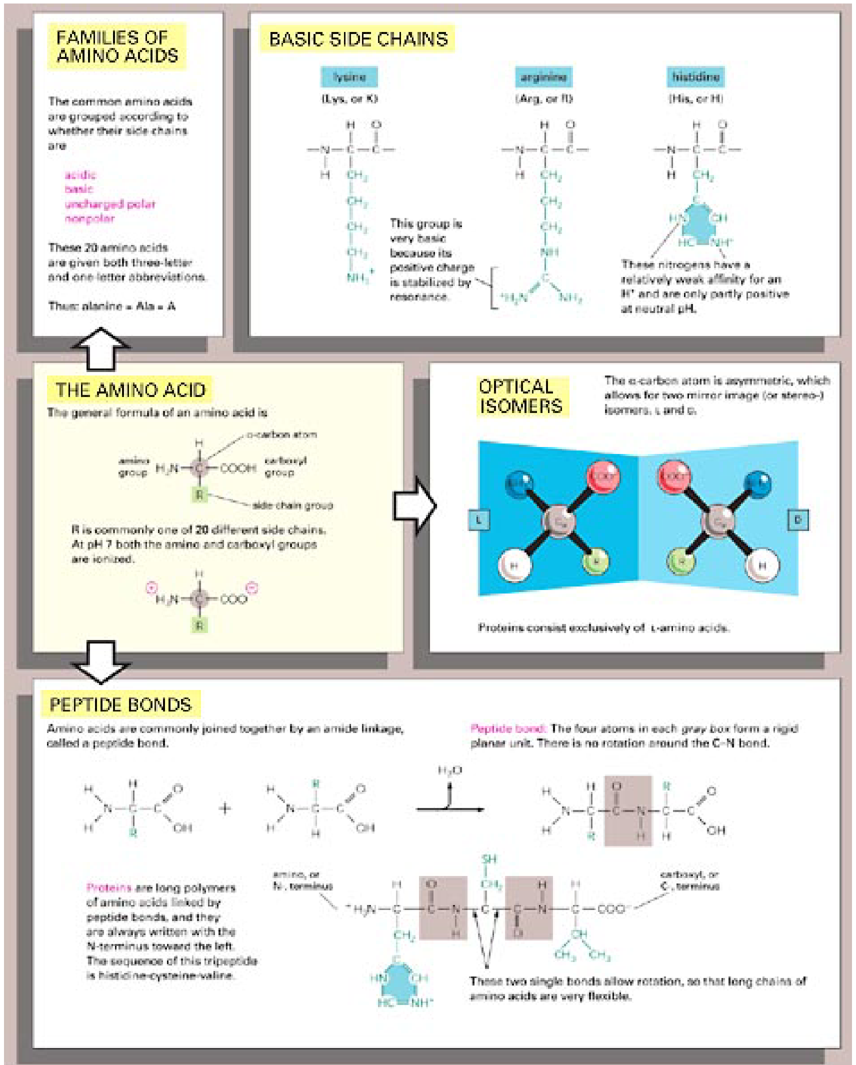


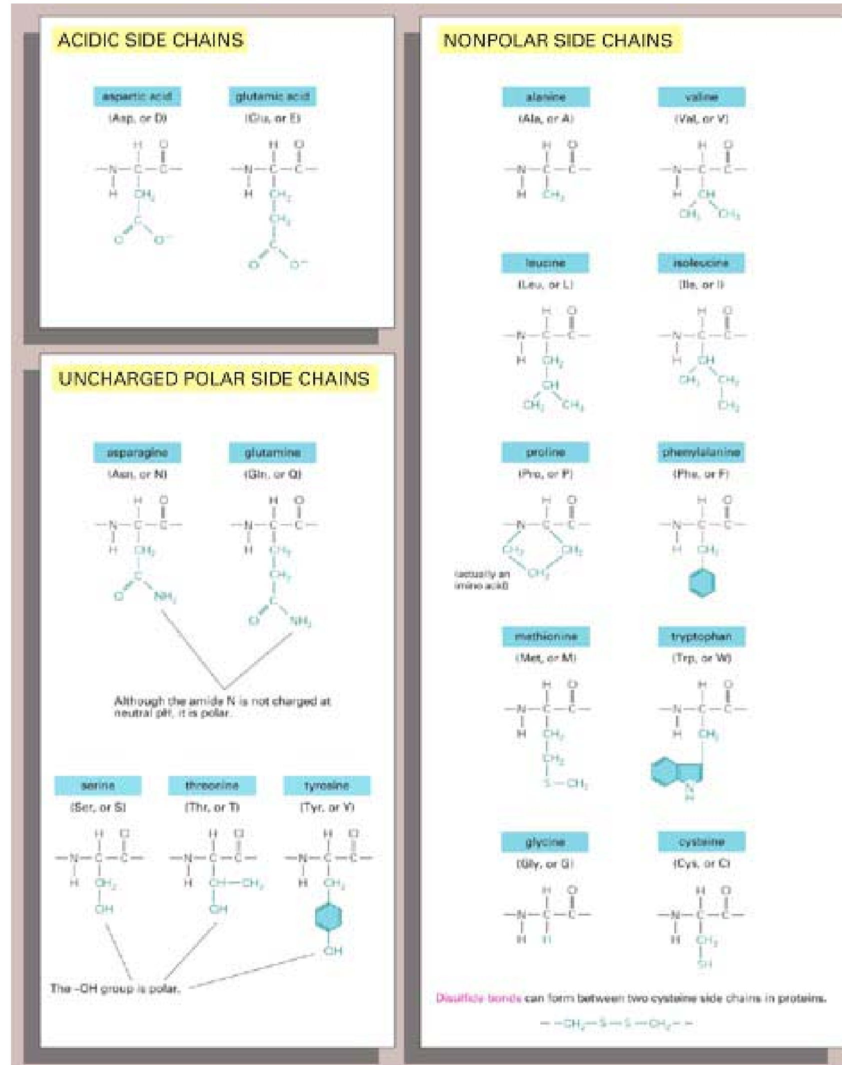
Cysteine
(cys)



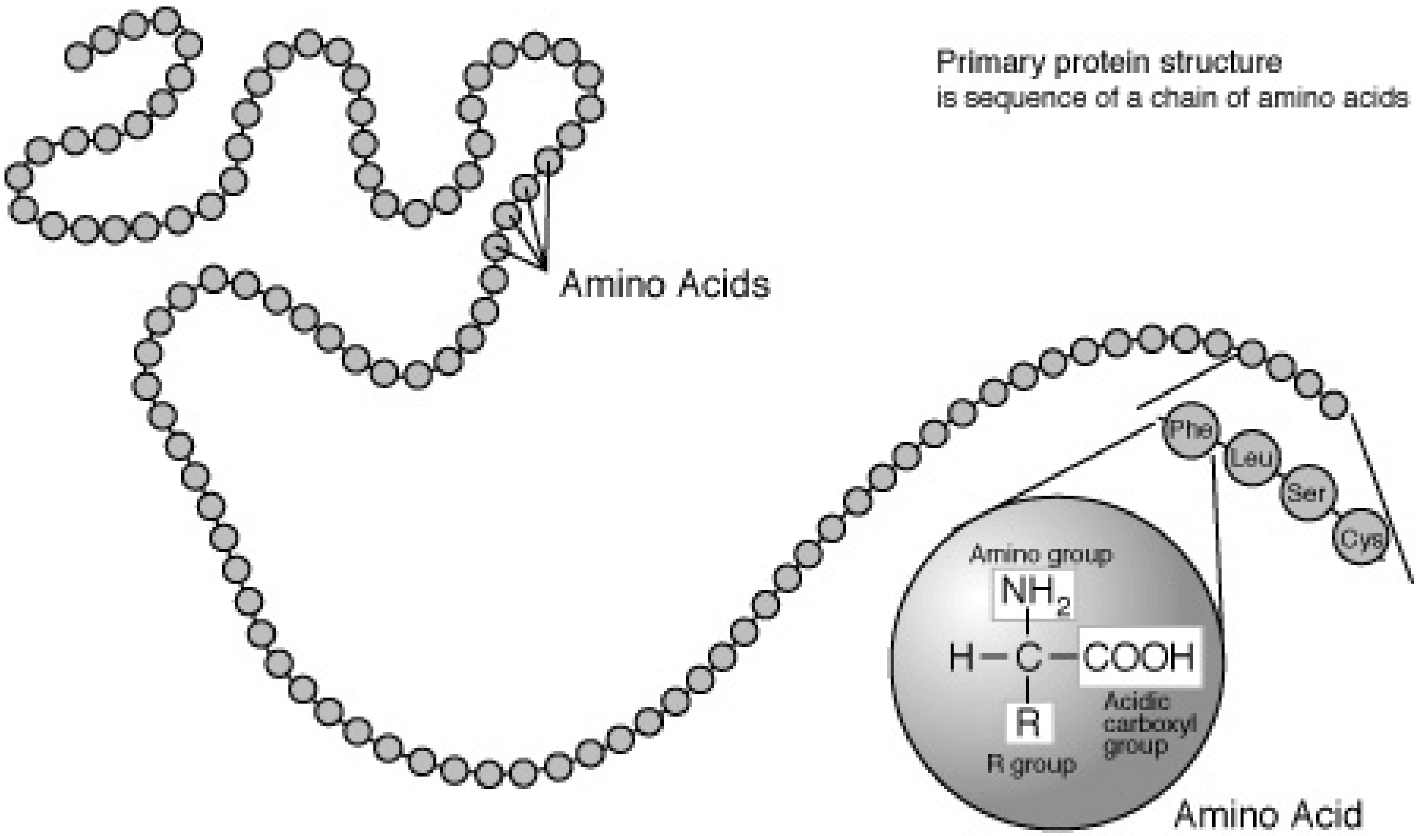
Proline
(pro)

Amino acids

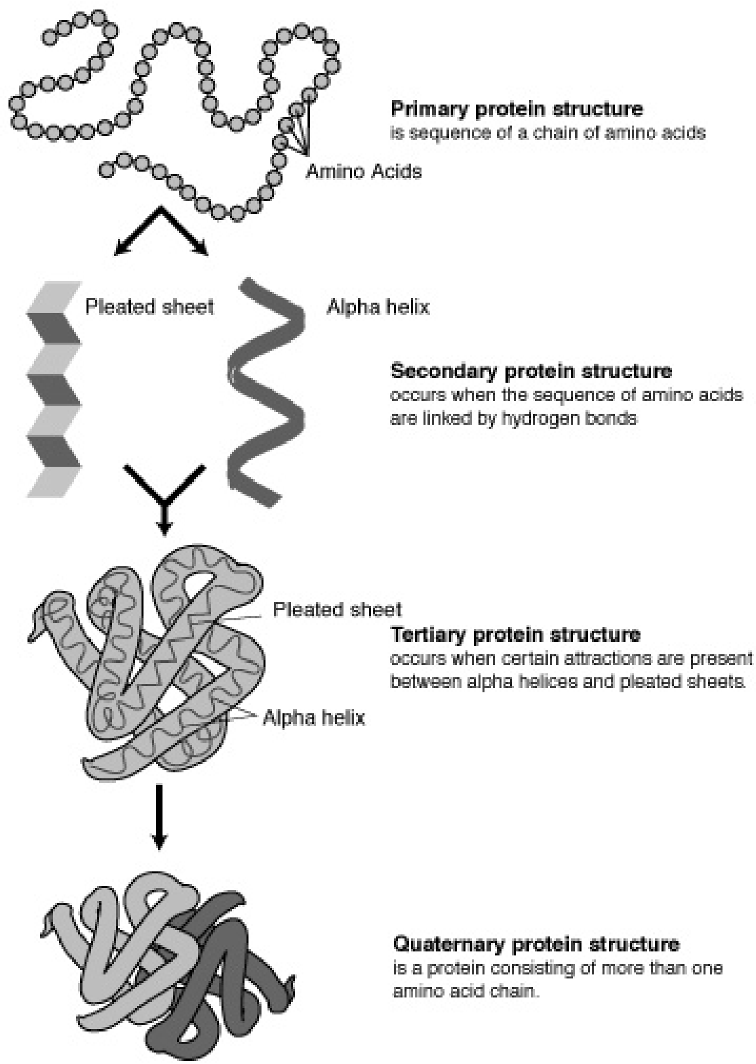

FAMILIES OF AMINO ACIDS
The common amino acids are grouped according to whether their side chains are
acidic
basic
uncharged polar
nonpolar
These 20 amino acids are given both three-letter and one-letter abbreviations.
Thus: alanine = Ala = A
BASIC SIDE CHAINS
lysine
(Lys, or K)
arginine
(Arg, or R)
histidine
(His, or H)
This group is very basic because its positive charge is stabilized by resonance.
These nitrogens have a relatively weak affinity for an H+ and are only partly positive at neutral pH.
THE AMINO ACID
The general formula of an amino acid is
α-carbon atom
amino group
carboxyl group
side chain group
R is commonly one of 20 different side chains. At pH 7 both the amine and carboxyl groups are ionized.
OPTICAL ISOMERS
The α-carbon atom is asymmetric, which allows for two mirror image (or stereo-) isomers, L and D.
Proteins consist exclusively of L-amino acids.
PEPTIDE BONDS
Amino acids are commonly joined together by an amide linkage, called a peptide bond.
Peptide bond: The four atoms in each gray box form a rigid planar unit. There is no rotation around the C–N bond.
Proteins are long polymers of amino acids linked by peptide bonds, and they are always written with the N-terminus toward the left. The sequence of this tripeptide is histidine-cysteine-valine.
amino, or N-terminus
carboxyl, or C-terminus
These two single bonds allow rotation, so that long chains of amino acids are very flexible.


ACIDIC SIDE CHAINS
aspartic acid
(Asp, or D)
glutamic acid
(Glu, or E)
UNCHARGED POLAR SIDE CHAINS
asparagine
(Asn, or N)
glutamine
(Gln, or Q)
Although the amide N is not charged at neutral pH, it is polar.
serine
(Ser, or S)
threonine
(Thr, or T)
tyrosine
(Tyr, or Y)
The –OH group is polar.
NONPOLAR SIDE CHAINS
alanine
(Ala, or A)
valine
(Val, or V)
leucine
(Leu, or L)
isoleucine
(Ile, or I)
proline
(Pro, or P)
(actually an imino acid)
phenylalanine
(Phe, or F)
methionine
(Met, or M)
tryptophan
(Trp, or W)
glycine
(Gly, or G)
cysteine
(Cys, or C)
Disulfide bonds can form between two cysteine side chains in proteins.

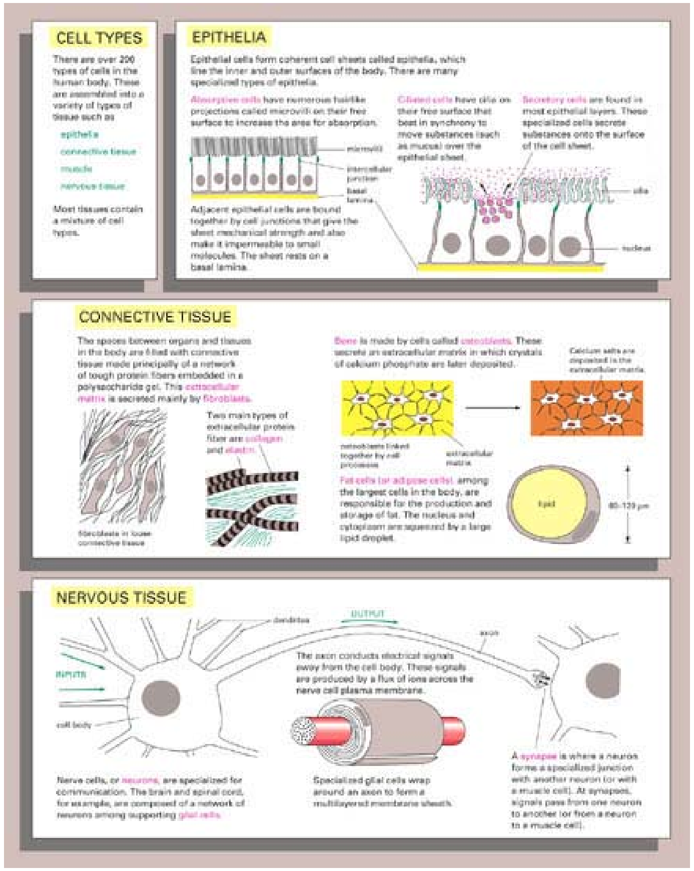
Proteins


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Phe
Leu
Ser
Cys
Amino group
NH2
H–C–COOH
R
Acidic
carboxyl
group
R group
Amino Acid

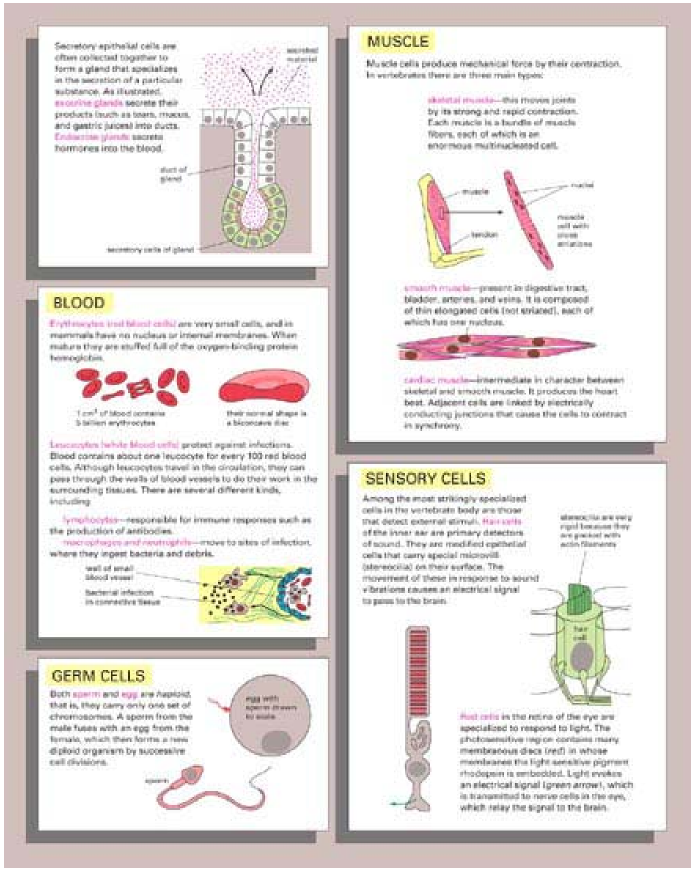
Proteins


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Cell types



CELL TYPES
EPITHELIA
CONNECTIVE TISSUE
NERVOUS TISSUE

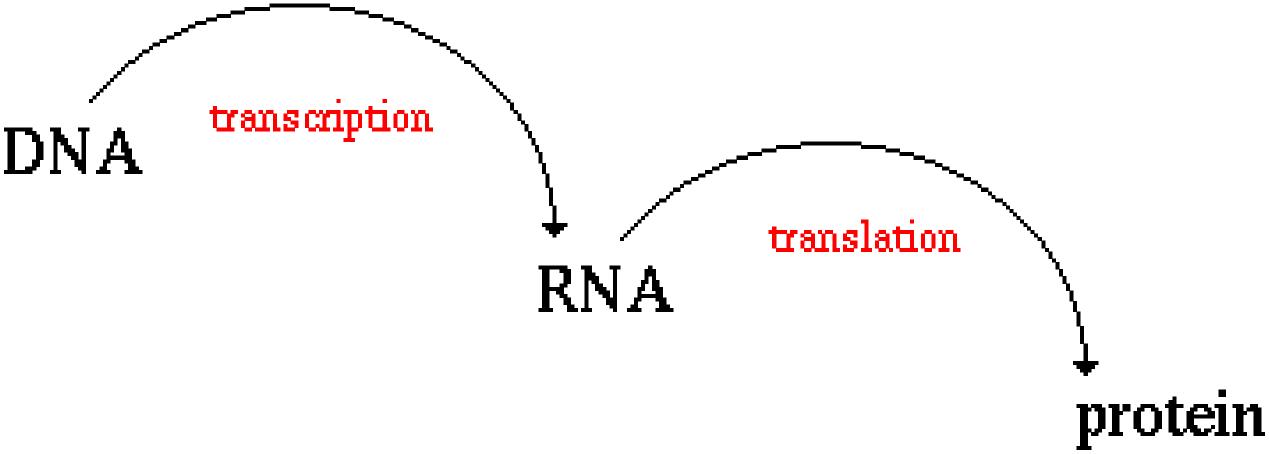


MUSCLE
BLOOD
GERM CELLS
SENSORY CELLS

Differential expression

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states
  E.g. blood, nerve, and skin cells, dividing cells, cancerous cells, etc.
- What makes the cells different?
- **Differential gene expression**, i.e., **when**, **where**, and **how much** each gene is expressed.
- On average, 40% of our genes are expressed at any given time.

Central dogma



Central dogma

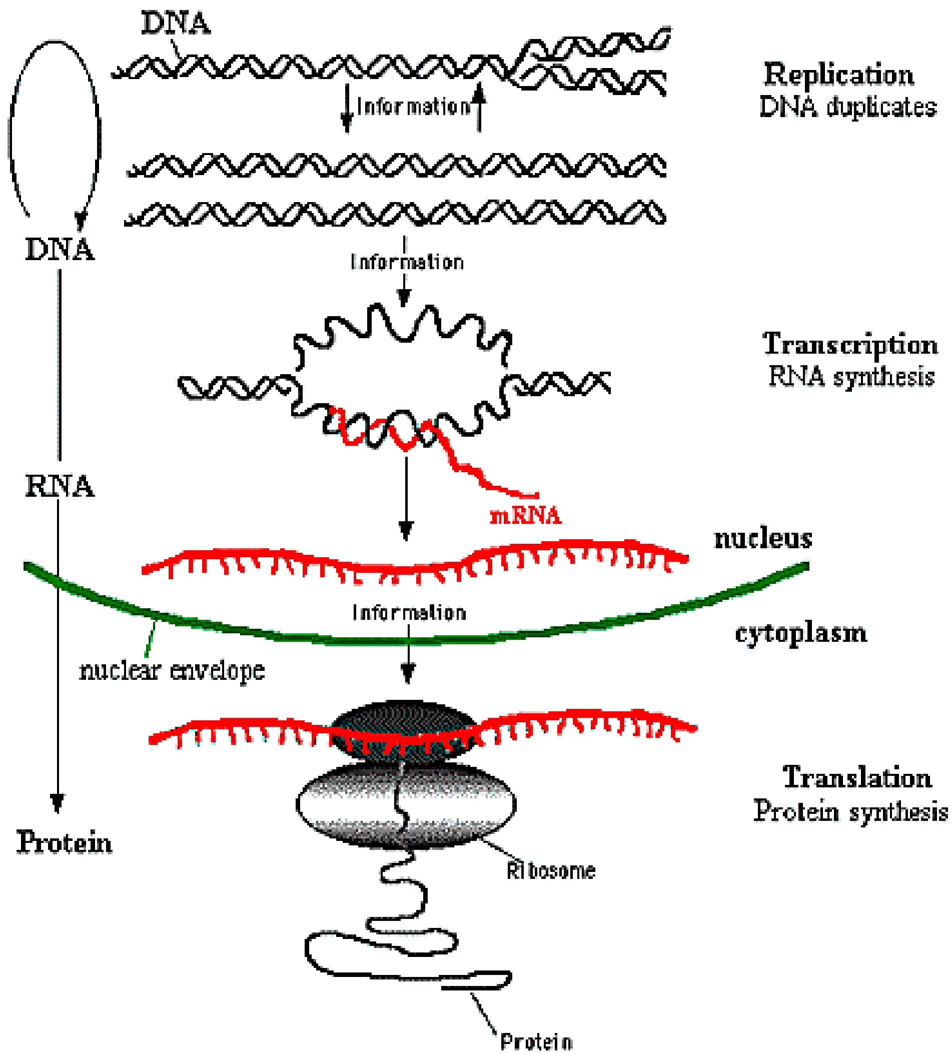
The **expression** of the genetic information stored in the DNA molecule occurs in two stages:

- (i) **transcription**, during which DNA is transcribed into mRNA;
- (ii) **translation**, during which mRNA is translated to produce a protein.

DNA → mRNA → protein

Other important aspects of regulation: methylation, alternative splicing, etc.

Central dogma



The Central Dogma of Molecular Biology

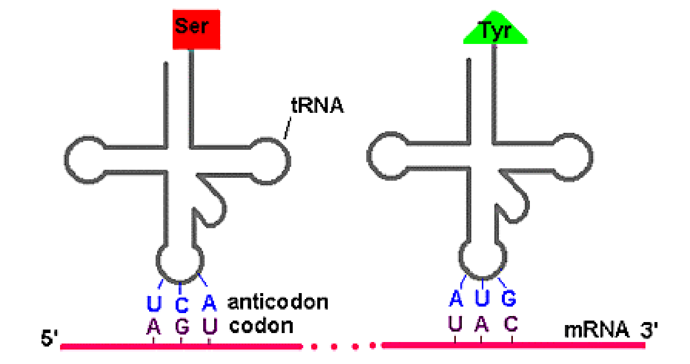
RNA

- A **ribonucleic acid** or **RNA** molecule is a nucleic acid similar to DNA, but
  - single-stranded;
  - ribose sugar rather than deoxyribose sugar;
  - **uracil (U)** replaces thymine (T) as one of the bases.
- RNA plays an important role in protein synthesis and other chemical activities of the cell.
- Several classes of RNA molecules, including **messenger RNA** (**mRNA**), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs.

The genetic code

- **DNA:** sequence of **four** different nucleotides.
- **Proteins:** sequence of **twenty** different amino acids.
- The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the **genetic code**, which relates nucleotide triplets or **codons** to **amino acids**.

The genetic code



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

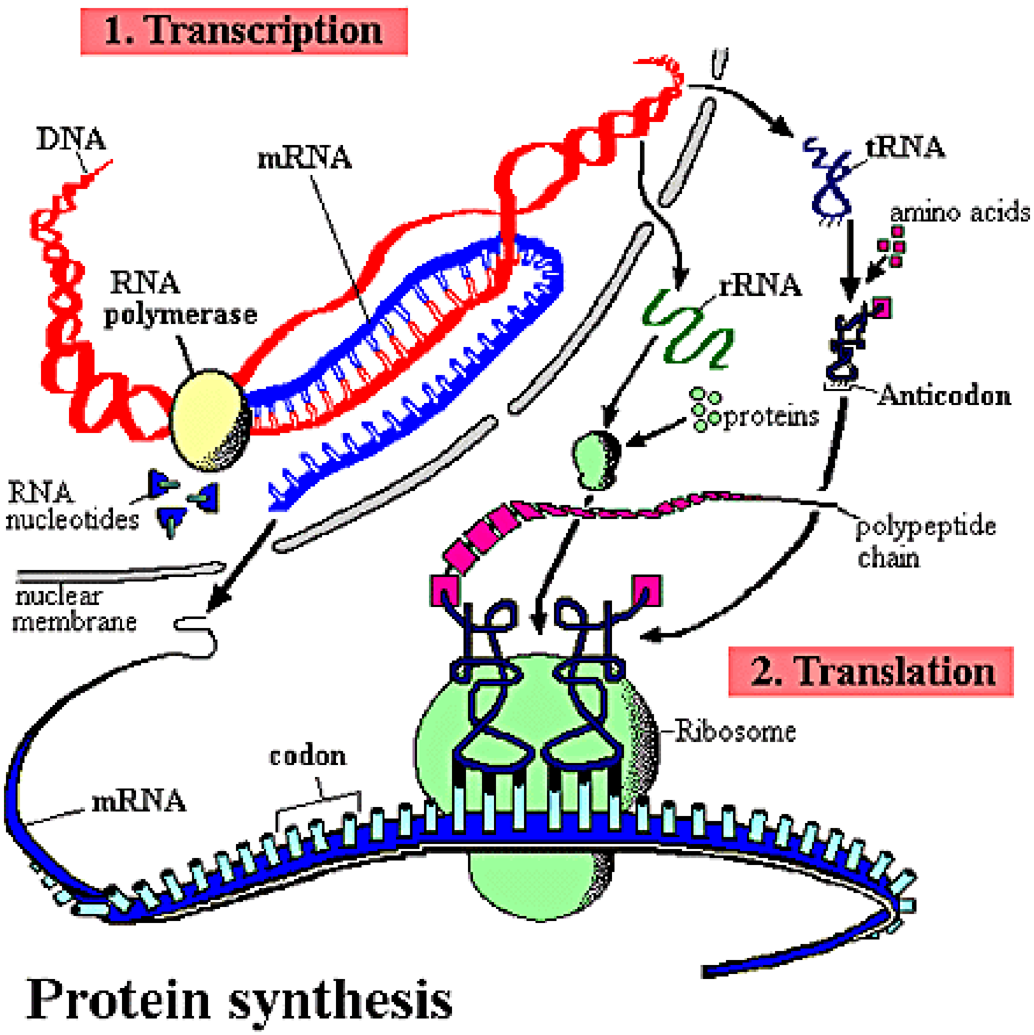
The Genetic Code

Start codon: initiation of translation (AUG, Met).
Stop codons: termination of translation.

Mapping between codons and amino acids is **many-to-one**: 64 codons but only 20 a.a..

Third base in codon is often redundant,
e.g., stop codons.

Protein synthesis



Transcription

- Analogous to DNA replication: several steps and many enzymes.
- **RNA polymerase** synthesizes an RNA strand complementary to one of the two DNA strands.
- The RNA polymerase recruits **rNTPs** (ribonucleotide triphosphate) in the same way that DNA polymerase recruits dNTPs (deoxunucleotide triphospate).
- However, synthesis is **single stranded** and only proceeds in the 5' to 3' direction of mRNA (no Okazaki fragments).

Transcription

- The strand being transcribed is called the **template** or **antisense** strand; it contains **anticodons**.
- The other strand is called the **sense** or **coding** strand; it contains **codons**.
- The RNA strand newly synthesized from and complementary to the template contains the same information as the coding strand.

Transcription

5' ...A T G G C C T G G A C T T C A... 3' Sense strand of DNA
3' ...T A C C G G A C C T G A A G T... 5' Antisense strand of DNA

↓ Transcription of antisense strand

5' ...A U G G C C U G G A C U U C A... 3' mRNA

↓ Translation of mRNA

Met — Ala — Trp — Thr — Ser — Peptide

Transcription

- **Promoter.** Unidirectional sequence upstream of the coding region (i.e., at 5' end on sense strand) that tells the RNA polymerase both **where** to start and on **which strand** to continue synthesis. E.g. TATA box.
- **Terminator.** Regulatory DNA region signaling end of transcription, at 3' end .
- **Transcription factor.** A protein needed to initiate the transcription of a gene, binds either to specific DNA sequences (e.g. promoters) or to other transcription factors.

Transcription

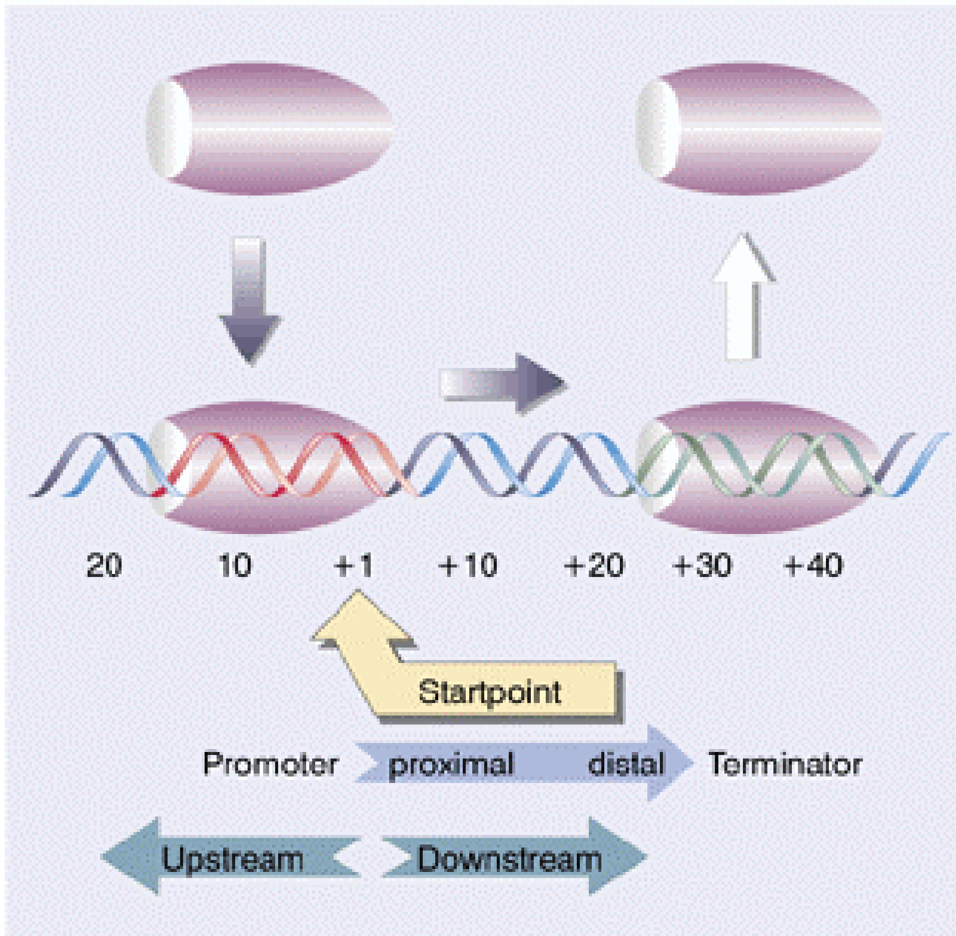
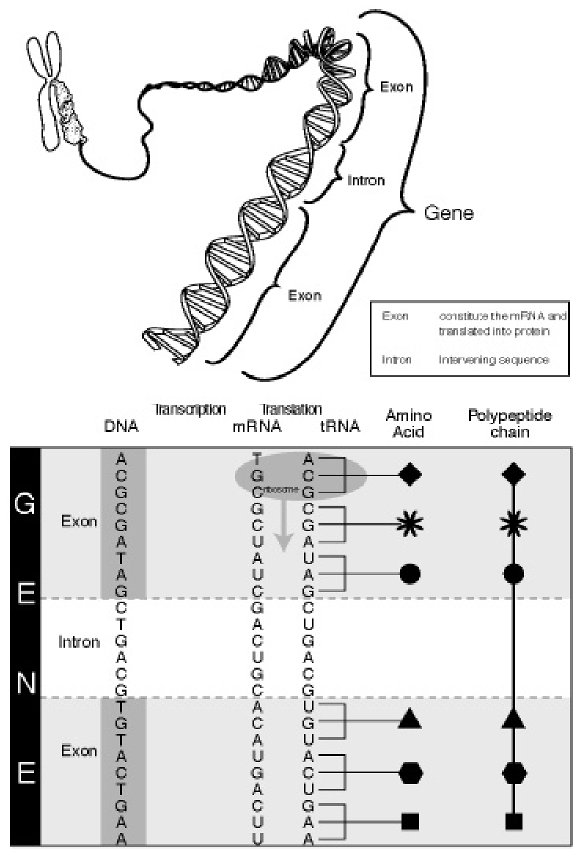


Figure 9.2 *Overview*: a transcription unit is a sequence of DNA transcribed into a single RNA, starting at the promoter and ending at the terminator.

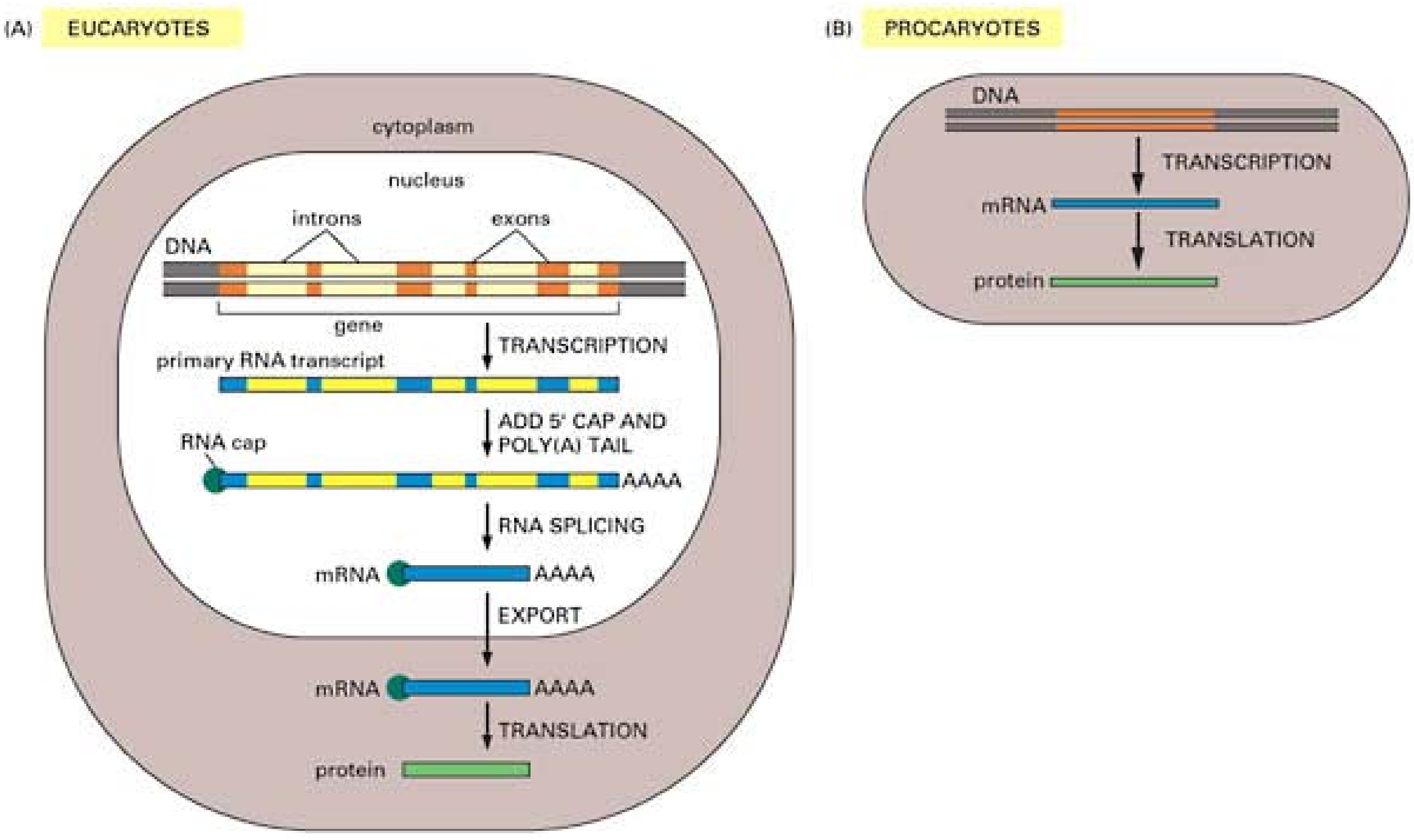
Exons and introns

- Genes comprise only about 2% of the human genome.
- The rest consists of **non-coding** regions
  - chromosomal structural integrity,
  - cell division (e.g. centromere)
  - regulatory regions: regulating when, where, and in what quantity proteins are made .
- The terms **exon** and **intron** refer to coding (translated into a protein) and non-coding DNA, respectively.

Exons and introns



Splicing



(A) EUCARYOTES
cytoplasm
nucleus
introns
exons
DNA
gene
TRANSCRIPTION
primary RNA transcript
ADD 5' CAP AND
POLY(A) TAIL
RNA cap
AAAA
RNA SPLICING
mRNA
AAAA
EXPORT
mRNA
AAAA
TRANSLATION
protein
(B) PROCARYOTES
DNA
TRANSCRIPTION
mRNA
TRANSLATION
protein

Translation

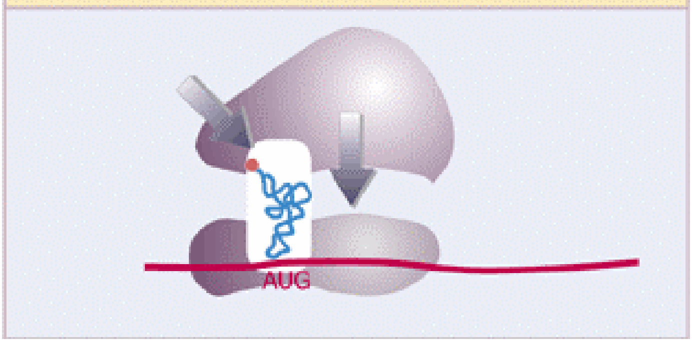
- **Ribosome**:
  - cellular factory responsible for protein synthesis;
  - a large subunit and a small subunit;
  - structural RNA and about 80 different proteins.
- **transfer RNA (tRNA)**:
  - adaptor molecule, between mRNA and protein;
  - specific **anticodon** and **acceptor site**;
  - specific **charger protein**, can only bind to that particular tRNA and attach the correct amino acid to the acceptor site.

Translation

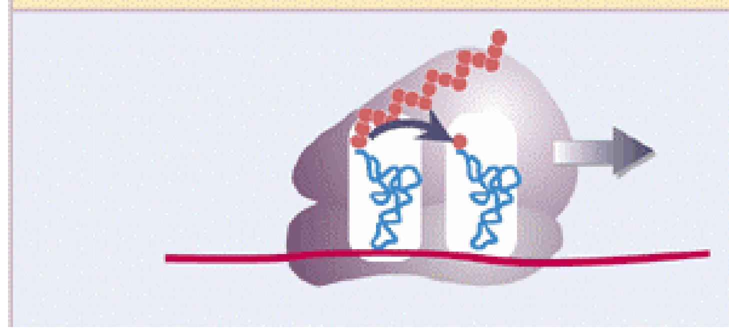
- Initiation
  - **Start codon AUG**, which codes for **methionine, Met**.
  - Not every protein necessarily starts with methionine. Often this first amino acid will be removed in post-translational processing of the protein.
- Termination:
  - **stop codon** (UAA, UAG, UGA) ,
  - ribosome breaks into its large and small subunits, releasing the new protein and the mRNA.

Translation

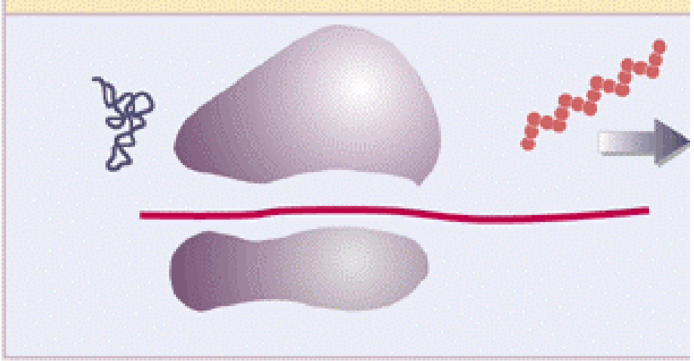
Initiation
30S subunit on mRNA binding site is joined by 50S subunit and aminoacyl-tRNA binds



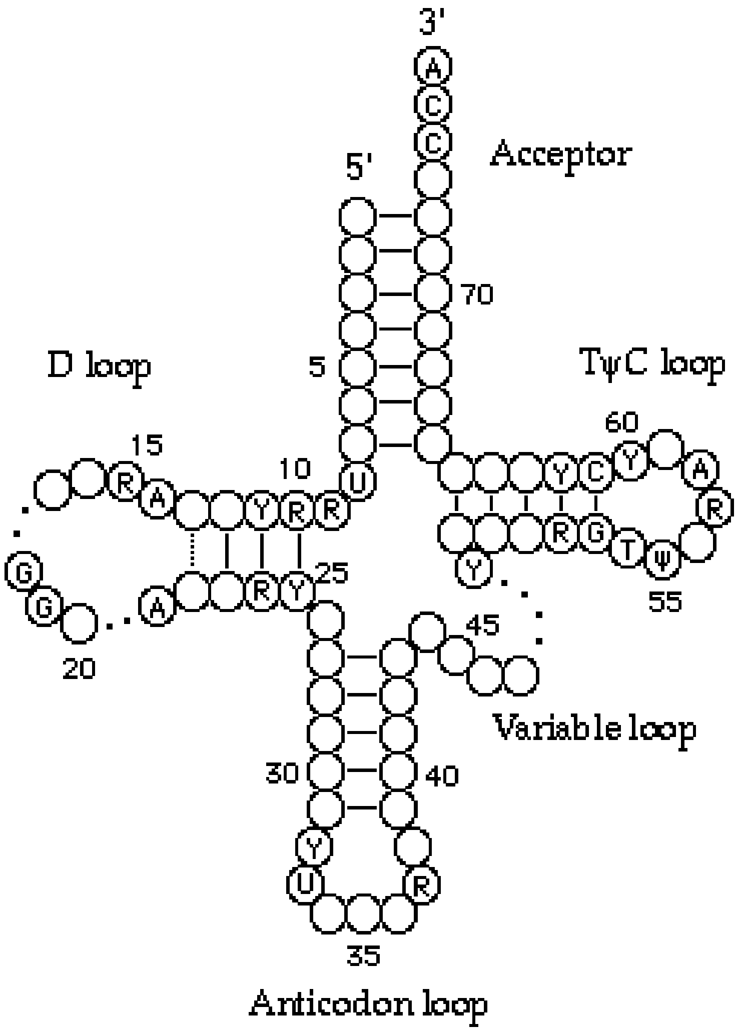
Elongation
Ribosome moves along mRNA and length of protein chain extends by transfer from peptidyl-tRNA to aminoacyl-tRNA



Termination
Polypeptide chain is released from tRNA, and ribosome dissociates from mRNA



tRNA

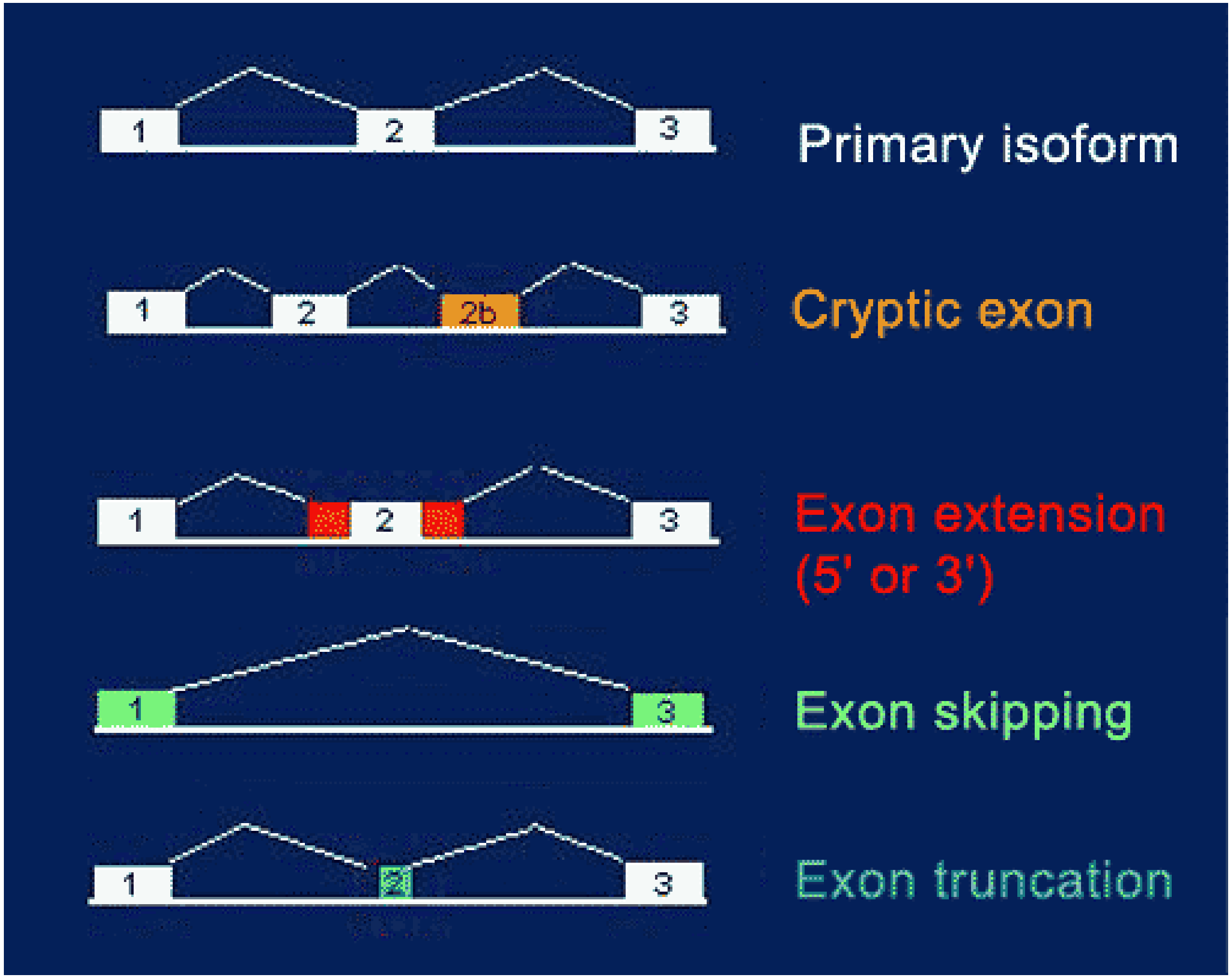


- The tRNA has an **anticodon** on its mRNA-binding end that is complementary to the codon on the mRNA.
- Each tRNA only binds the appropriate amino acid for its anticodon.

Alternative splicing

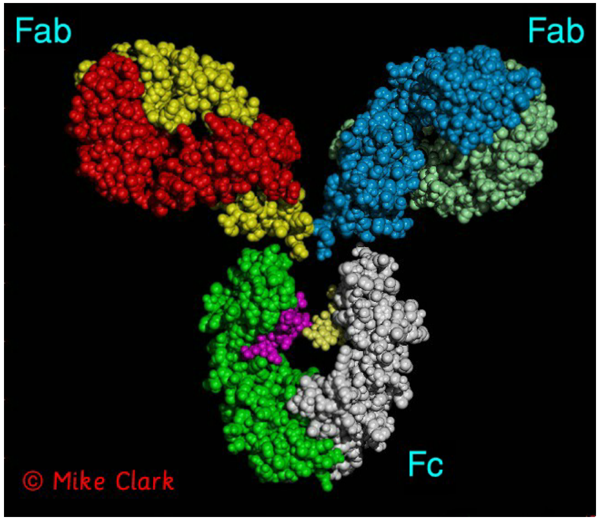
- There are more than 1,000,000 different human antibodies. How is this possible with only ~30,000 genes?
- **Alternative splicing** refers to the different ways of combining a gene's exons. This can produce different forms of a protein for the same gene.
- Alternative pre-mRNA splicing is an important mechanism for regulating gene expression in higher eukaryotes.
- E.g. in humans, it is estimated that approximately 30% of the genes are subject to alternative splicing.

Alternative splicing



Immunoglobulin

- B cells produce antibody molecules called immunoglobulins (Ig) which fall in five broad classes.
- Diversity of Ig molecules
  - DNA sequence: recombination, mutation.
  - mRNA sequence: alternative splicing.
  - Protein structure: post-translational proteolysis, glycosylation.

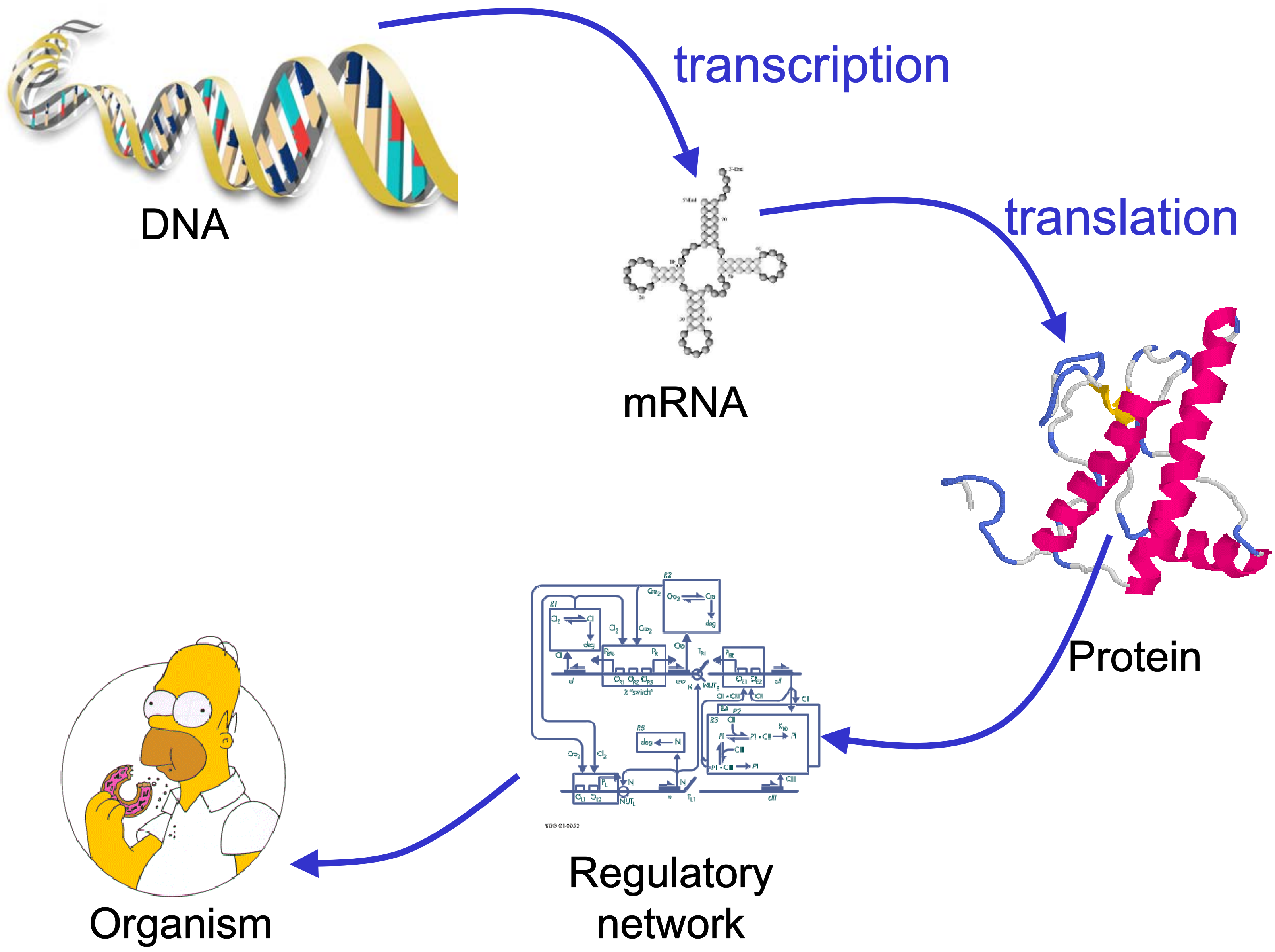


IgG1

Post-translational processing

- Folding.
- Cleavage by a proteolytic (protein-cutting) enzyme.
- Alteration of amino acid residues
  - phosphorylation, e.g. of a tyrosine residue.
  - glycosylation, carbohydrates covalently attached to asparagine residue.
  - methylation, e.g. of arginine.
- Lipid conjugation.

Transcription and translation



Control of Gene Expression

- there is strong evidence that the DNA content of most cells in a multi-cell organism is identical
- different cell types synthesize different sets of proteins at different times

Gene expression

Control of Gene Expression

- there are at least six ways to control protein expression

1. control when and how often a gene is transcribed
2. control how the transcript is spliced
3. select which mRNA's are exported from the nucleus
4. control translation

Controlling Expression

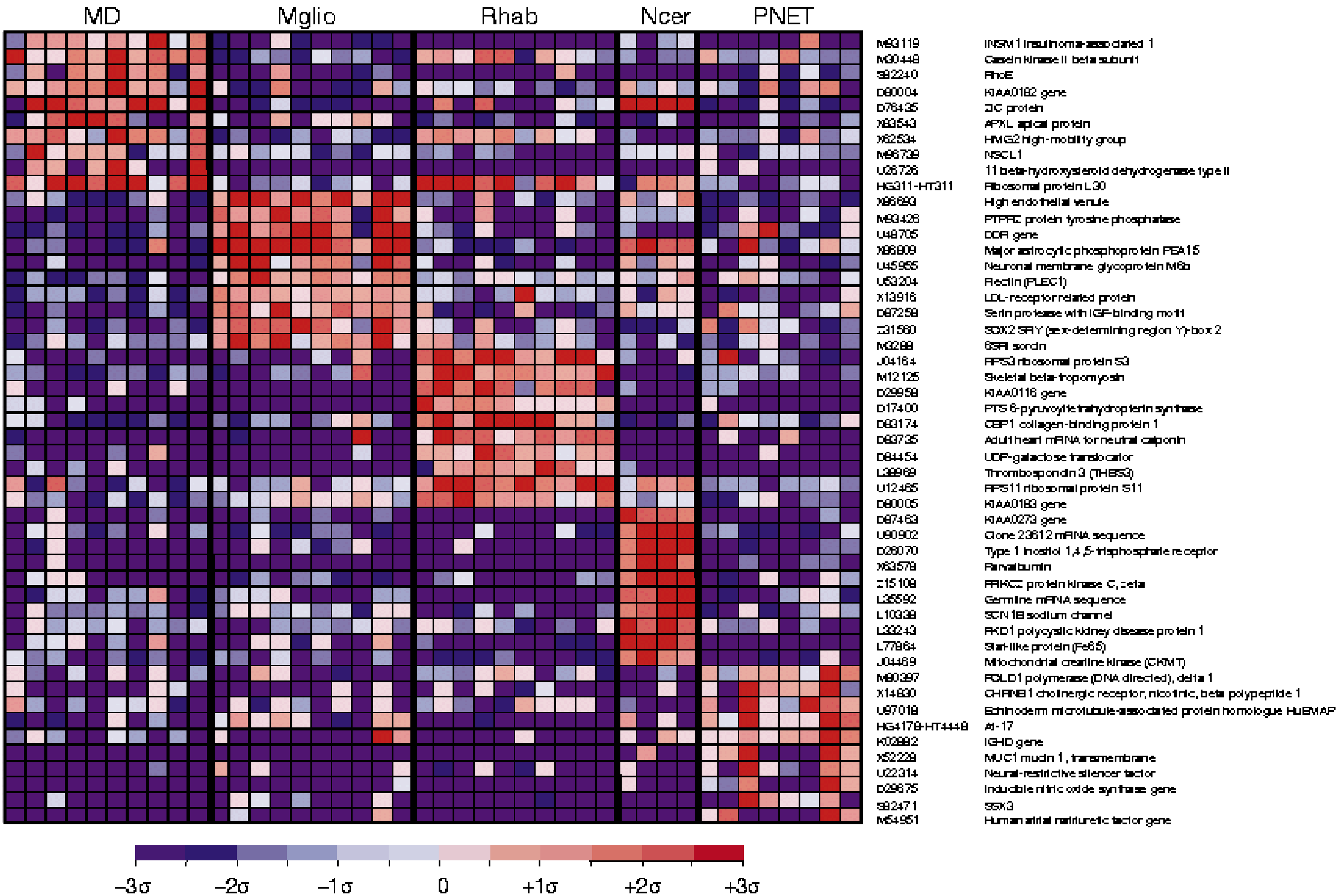
5. selectively destabilize mRNAs in the cytoplasm
6. control protein activity (degradation, inactivate, isolate), post-translational modifications

- for most genes transcriptional control is the most important

Gene Expression

- for many diseases specific patterns of gene expression (mRNA expression) have been associated with the different phenotypes

Class discovery



Different Tumors have different patterns of expression

Fig. from Pomeroy et al. Nature 415 (2002)

DNA Microarrays

- the data obtained from microarray experiments is a measure of the abundance of a nucleic acid
- usually they are used for detecting mRNA levels
- some of the issues mentioned previously can affect the observed abundance of mRNA

Low values of mRNA

- the gene may be deleted
- the gene may be being repressed
- the gene may no longer be enhanced
- the gene may be methylated
- the mRNA may be kept in the nucleus

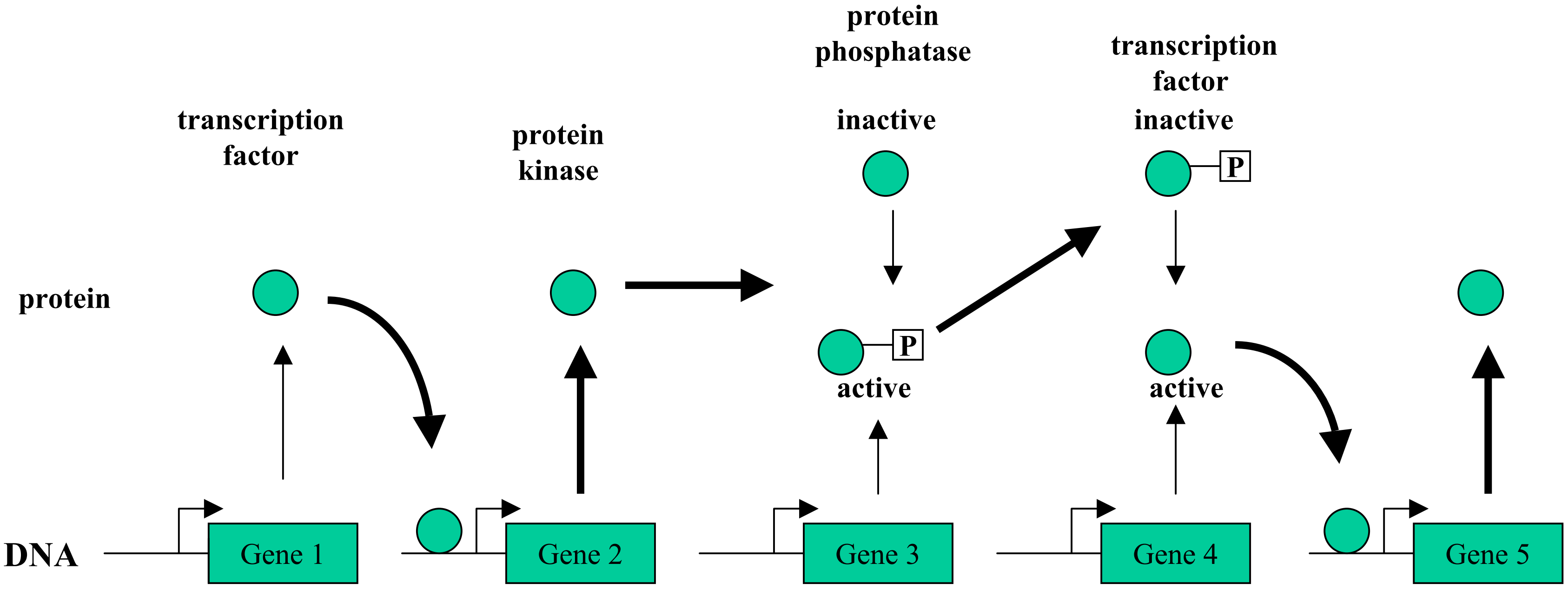
High Levels of mRNA

- the gene may be part of an amplicon
- the gene may no longer be being repressed
- the gene may be being enhanced

Other Issues

- was the right sequence applied to the chip?
- alternative splicing: which one are we measuring?
- cross-hybridization – genes with similar sequences may hybridize

An example of the interactions between some genes (adapted from Wagner 2001)



Downstream Consequences

- many genes fall into the class of genes called transcription factors
- while most genes are transcribed by RNA polymerase II it cannot initiate transcription itself in eukaryotic cells
- transcription factors identify and then bind to specific sites in the DNA
- the TFs then guide and activate RNA polymerase

Downstream Consequences

- TFs tend not to be specific for one gene
- disregulation (or over or under production) of a TF can have large effects on gene expression
- for example ESR1 (estrogen receptor 1) is a transcription factor
- it affects production of cyclin d1 (CCND1)

Downstream Consequences

- CCND1 forms a complex with CDK4 and/or CDK6
- this complex inactivates the repressor function of pRb (retinoblastoma protein) which regulates cell proliferation
- and so on….

Functional genomics

- The various **genome projects** have yielded the complete DNA sequences of many organisms.
  - E.g. human, mouse, yeast, fruitfly, etc.
  - Human: 3 billion base-pairs, 30-40 thousand genes.
- Challenge: **go from sequence to function**, i.e., define the role of each gene and understand how the genome functions as a whole.

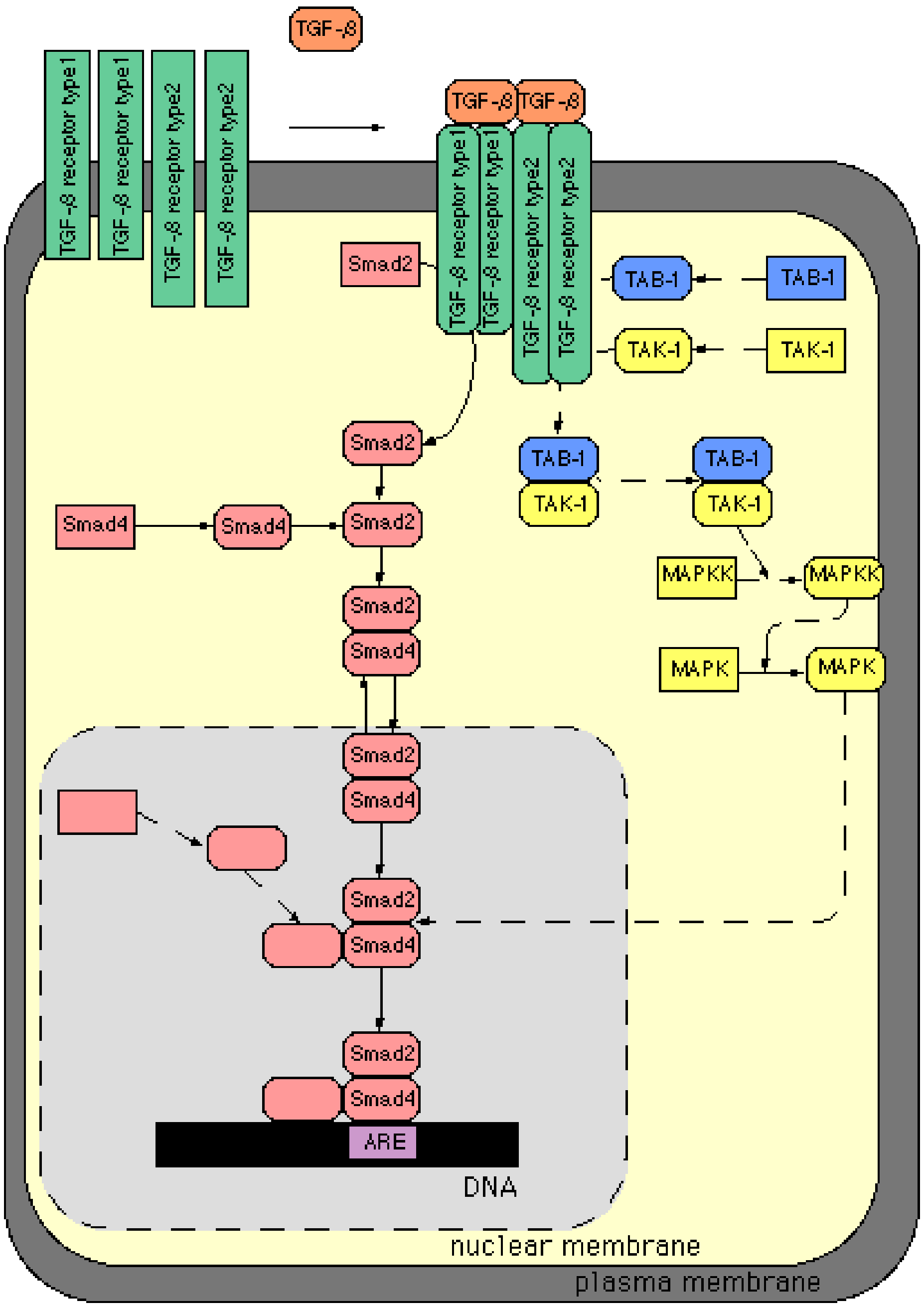
Pathways

- The complete genome sequence doesn’t tell us much about how the organism functions as a biological system.
- We need to study how different gene products interact to produce various components.
- Most important activities are not the result of a single molecule but depend on the **coordinated effects** of multiple molecules.

TGF-β pathway

- **Transforming Growth Factor beta**, **TGF-β**, plays an essential role in the control of development and morphogenesis in multicellular organisms.

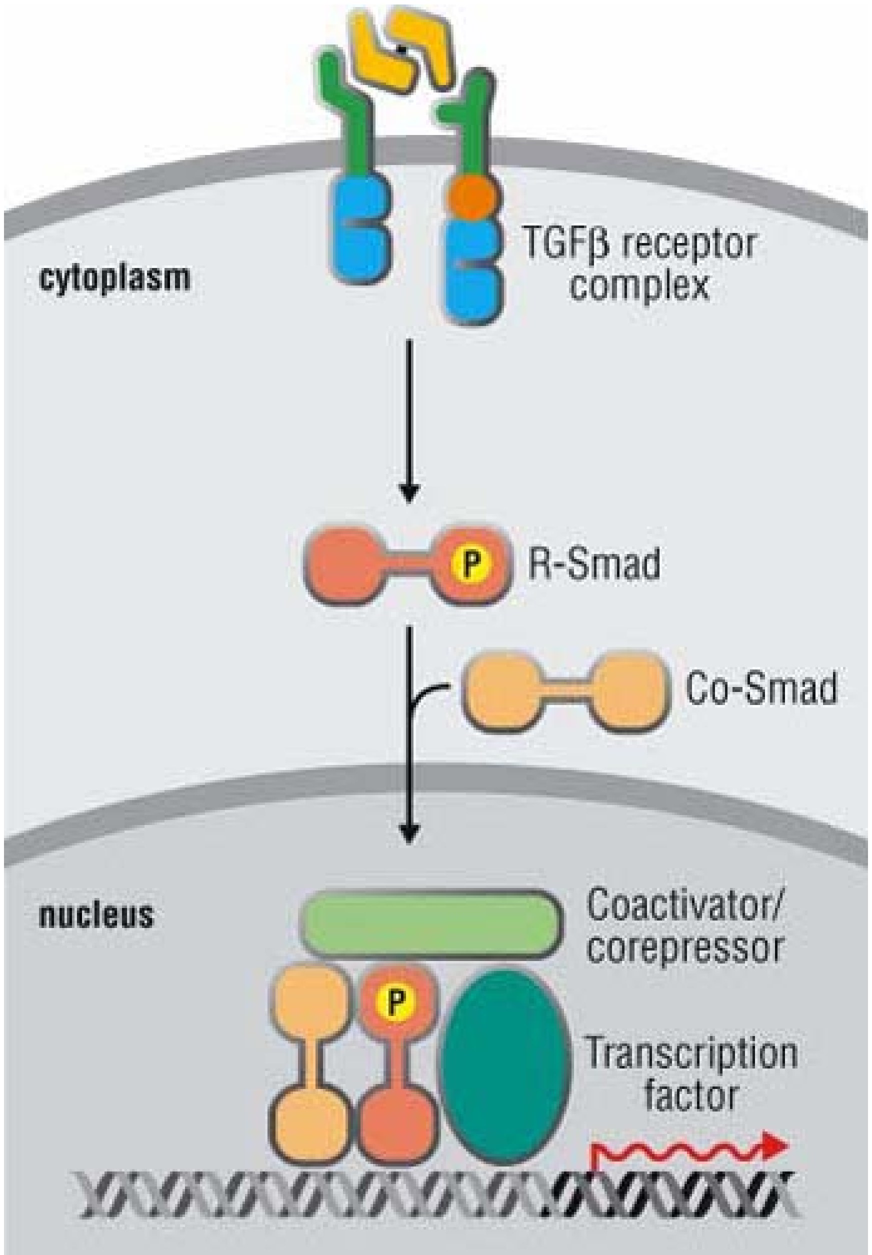
- The basic pathway provides a simple route for signals to pass from the extracellular environment to the nucleus, involving only four types of molecules.



TGF-β pathway

Four types of molecules

- TGF-β
- TGF-β type I receptors
- TGF-β type II receptors
- SMADS, a family of signal transducers and transcriptional activators.



TGF-β pathway

TGF-β pathway

- Extracellular TGF–β ligands transmit their signals to the cell's interior by binding to type II receptors, which form heterodimers with type I receptors.

- The receptors in turn activate the SMAD transcription factors.

TGF-β pathway

- Phosphorylated and receptor-activated SMADs (R-SMADs) form heterodimers with common SMADs (co-SMADs) and translocate to the nucleus.
- In the nucleus, SMADs activate or inhibit the transcription of target genes, in collaboration with other factors.

Pathways

- http://www.grt.kyushu-u.ac.jp/spad/
- There are many open questions regarding the relationship between gene expression levels (e.g. mRNA levels) and pathways.
- It is not clear to what extent microarray gene expression data will be informative.

WWW resources

- **Access Excellence**
  http://www.accessexcellence.com/AB/GG/
- **Genes VII**
  http://www.oup.co.uk/best.textbooks/biochemistry/genesvii/
- **Human Genome Project Education Resources**
  http://www.ornl.gov/hgmis/education/education.html
- **Kimball's Biology Pages**
  http://www.ultranet.com/~jkimball/BiologyPages/
- **MIT Biology Hypertextbook**
  http://esg-www.mit.edu:8001/