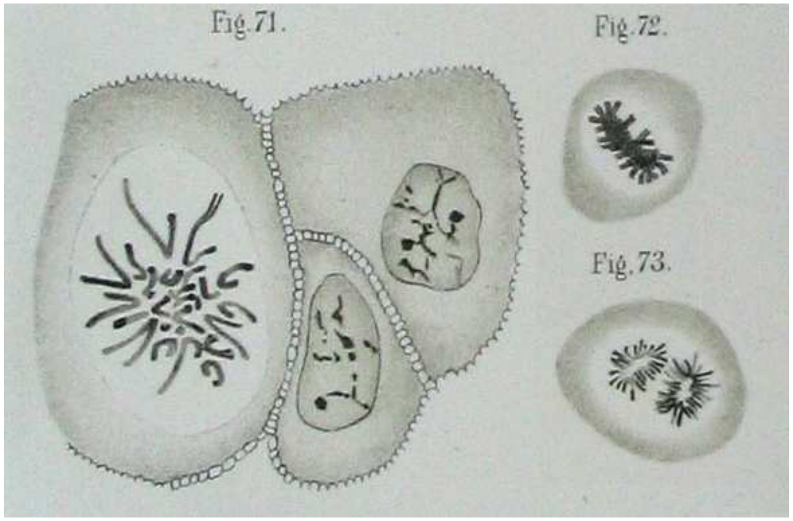
Brief Introduction to Biology

Broňa Brejová

Sept. 23, 2021



Walther Flemming, 1881

Principal players

Deoxyribonucleic acid (DNA)

carrier of genetic information passed from one generation to the next.
Long string of nucleotides from $\{A, C, G, T\}$
(adenine, cytosine, guanine, thymine).
Information stored in symbolic, digital form.

Ribonucleic acid (RNA)

Similar to DNA, thymine T replaced with uracil U

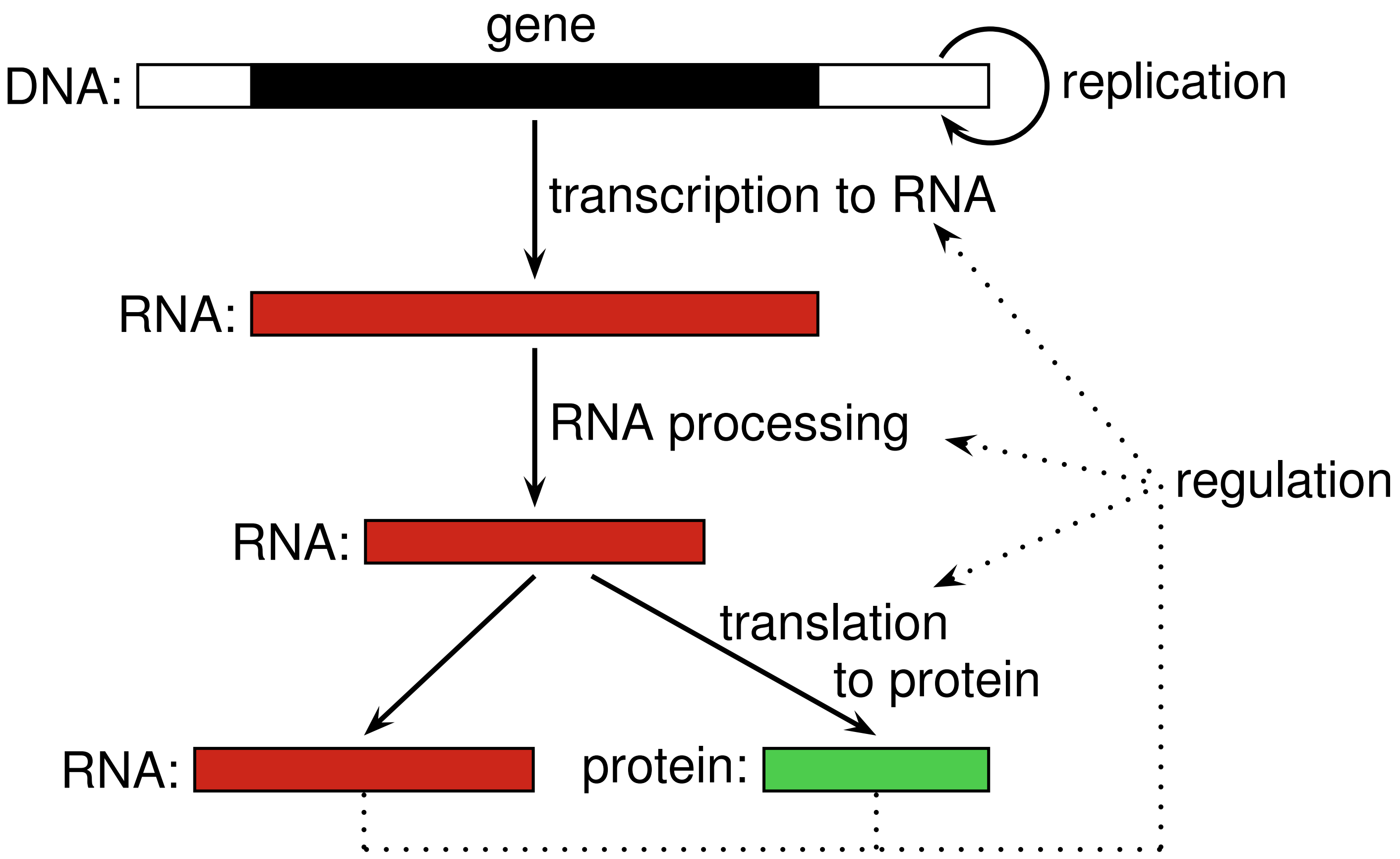
Proteins

catalyze biochemical reactions in the cell (enzymes),
carry signals within/between cells,
also important for cell structure and movement, etc.
String of amino acids (20 different amino acids).

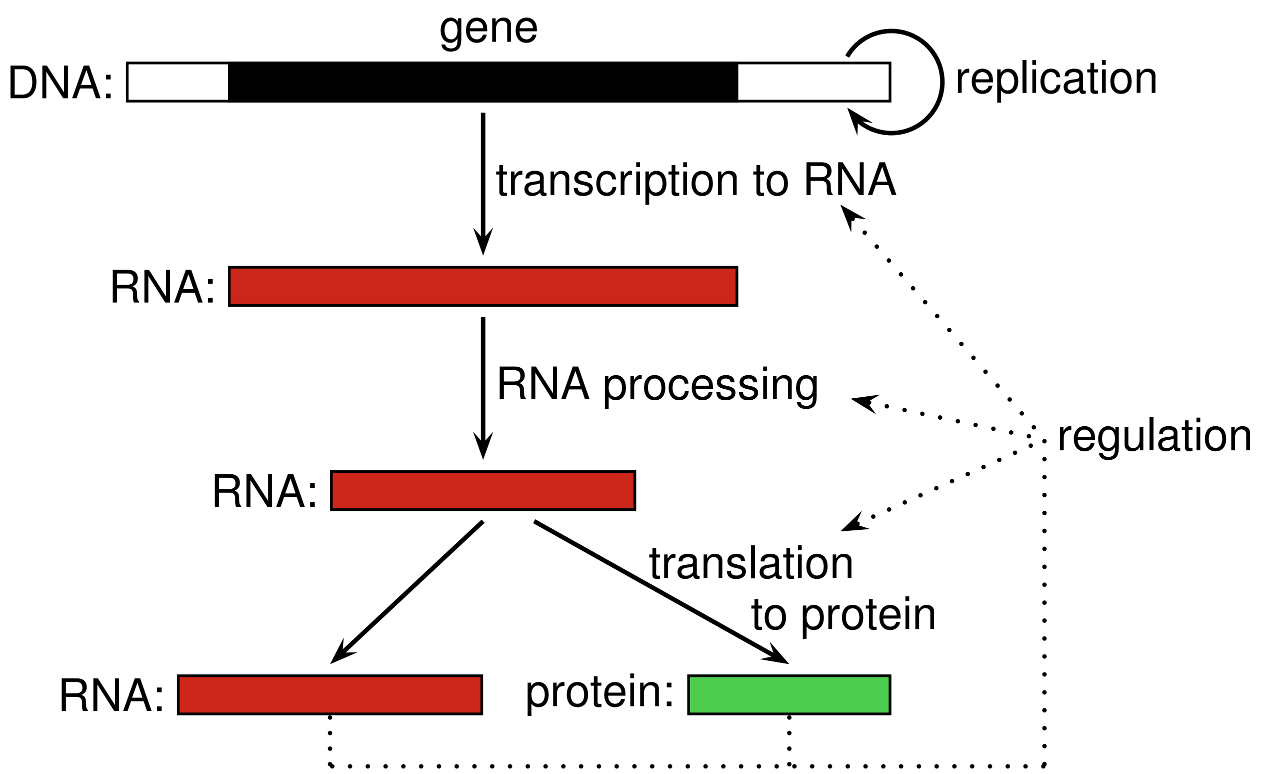
What information is stored in DNA?

Genes: Recipes for synthesis of proteins and functional RNAs.

Regulation of their expression: when and how many molecules to synthesize



Central dogma (Francis Crick 1958,1970)

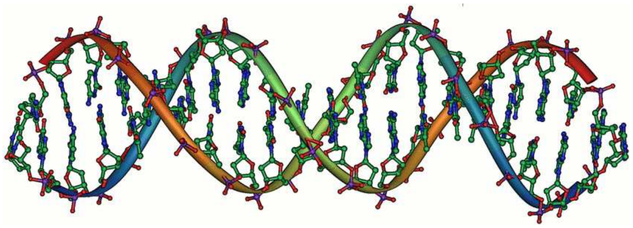


"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid."

DNA, chromosomes

DNA: two complementary strands (pairs A-T, C-G), in opposite orientation (ends are called 5' and 3'). For example ACCATG is complementary with CATGGT.

Shape of double helix



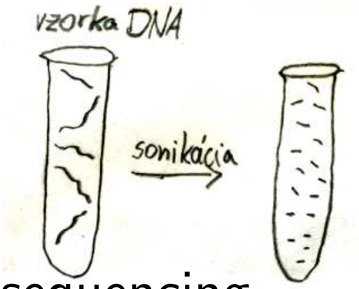
Double stranded structure allows redundancy, repair after damage in one strand. During cell division double stranded DNA unwinds and second strand is synthesized to each original strand (DNA replication).

Chromosome: Complex of double-stranded DNA molecule and support proteins

The human genome has 22 pairs chromosomes plus two sex chrom., together 3GB.

Technology: DNA sequencing

- Technology for determining sequence of nucleotides in chromosomes
- Complex process:
  chromosomes are cut to short pieces,
  each piece is duplicated many times,
  each piece is sequenced separately e.g. by Sanger sequencing
  – uses natural enzymes, e.g. DNA polymerase



Sanger sequencing

Example: sequencing AGCTAGGACT (below drawn right-to-left)

Primer AGT + enzymes + nucleotides

+ modified color-labeled nucleotides

Results of sequencing reaction:

```
TCAGGATCGA
AGTCCTAGC

TCAGGATCGA
AGTCCTA

TCAGGATCGA
AGTCCTAGCT

TCAGGATCGA
AGTCCT

TCAGGATCGA
AGTCC

TCAGGATCGA
AGTCCTAG

TCAGGATCGA
AGTC
```

Order by length on a gel:

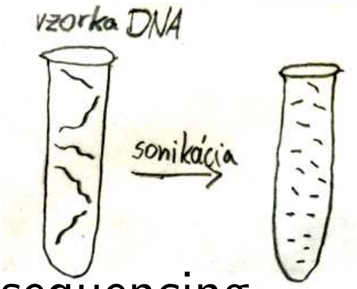
```
AGTCCTAGCT
AGTCCTAGC
AGTCCTAG
AGTCCTA
AGTCCT
AGTCC
AGTC
```

Read color labels to obtain complementary strand: AGTCCTAGCT

Technology: DNA sequencing

- Technology for determining sequence of nucleotides in chromosomes

- Complex process:
  chromosomes are cut to short pieces,
  each piece is duplicated many times,
  each piece is sequenced separately e.g. by Sanger sequencing
  – uses natural enzymes, e.g. DNA polymerase



- **Computational problem:** genome assembly from short pieces.

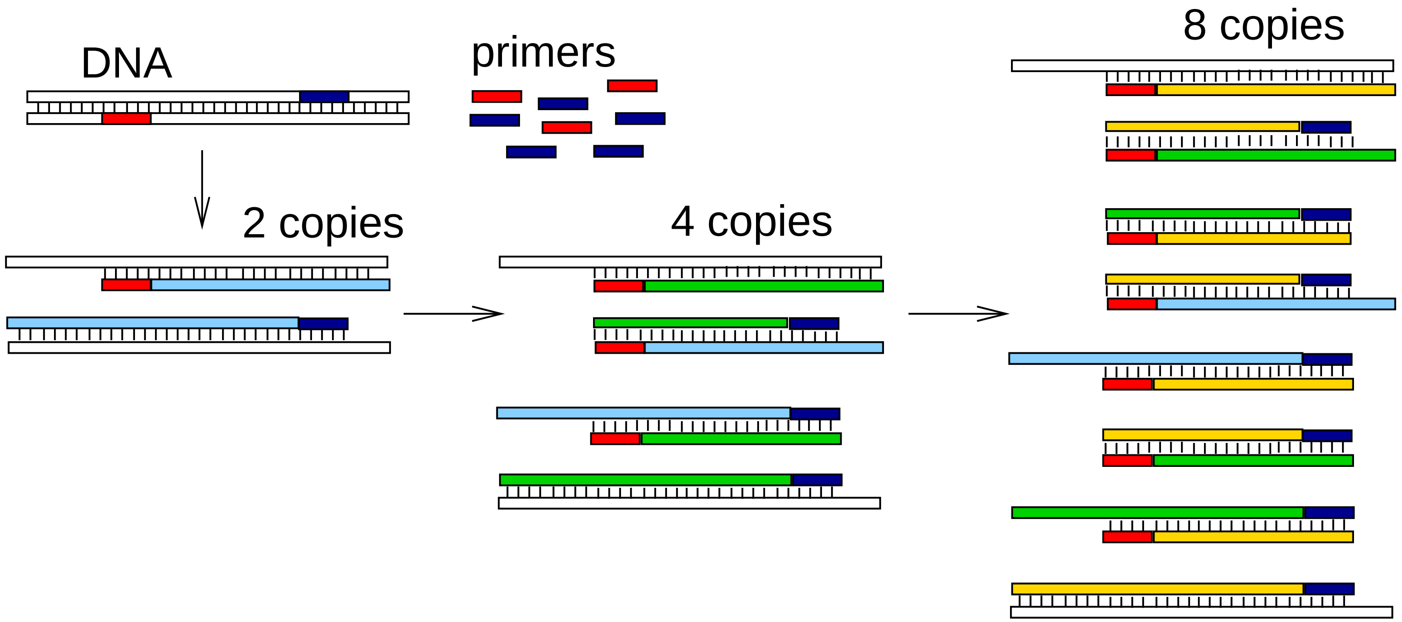
- Genome availability allows
  annotate genes and other functional regions,
  seek similarities and differences between species and organisms.

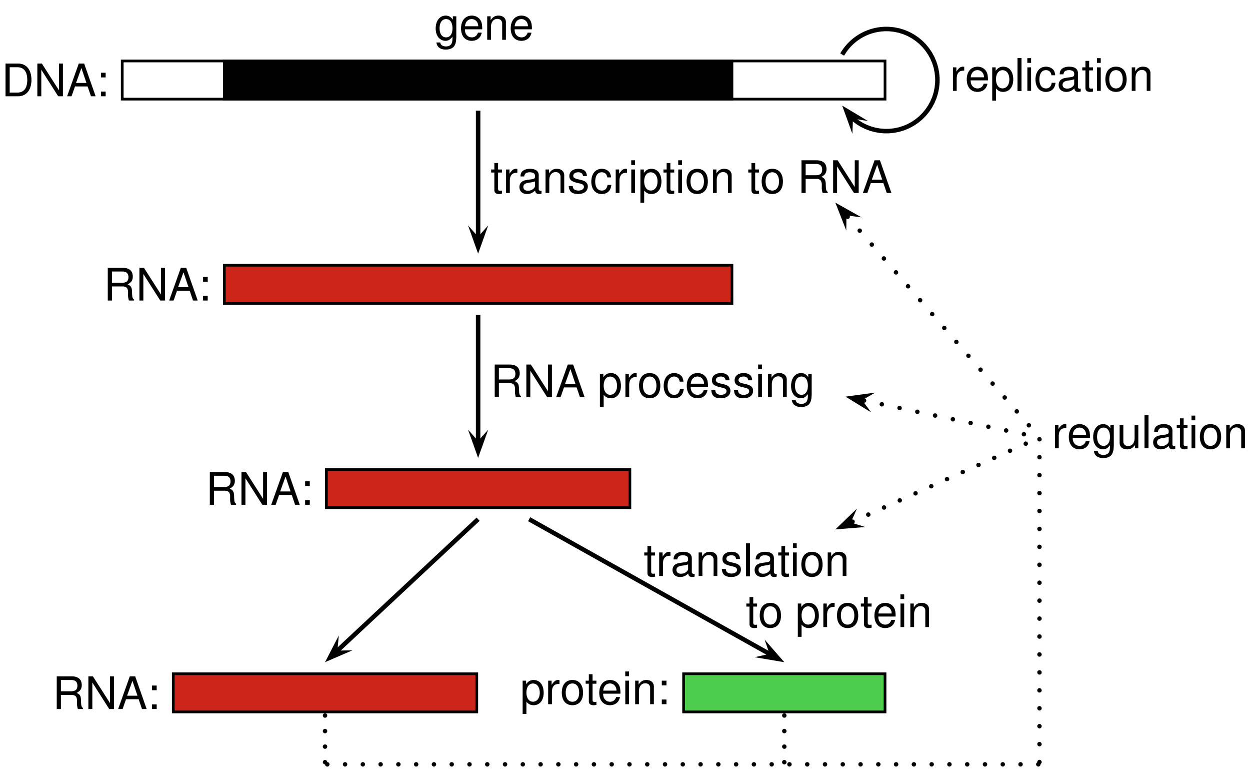
PCR (polymerase chain reaction)

We select two short pieces of DNA (primers)

PCR tests if they are close together in DNA sample
(hundreds/thousands of bases)

If yes, it will make many copies of the region between them

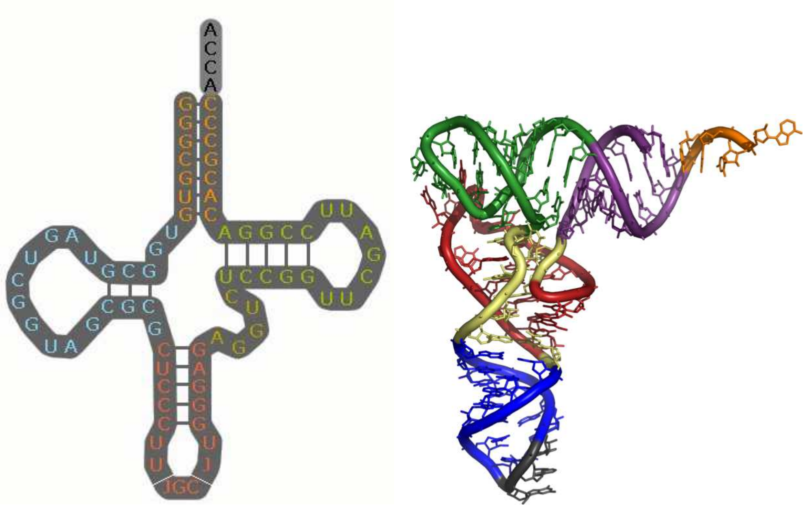



gene
DNA:
replication
transcription to RNA
RNA:
RNA processing
regulation
RNA:
translation
to protein
RNA:
protein:

RNA

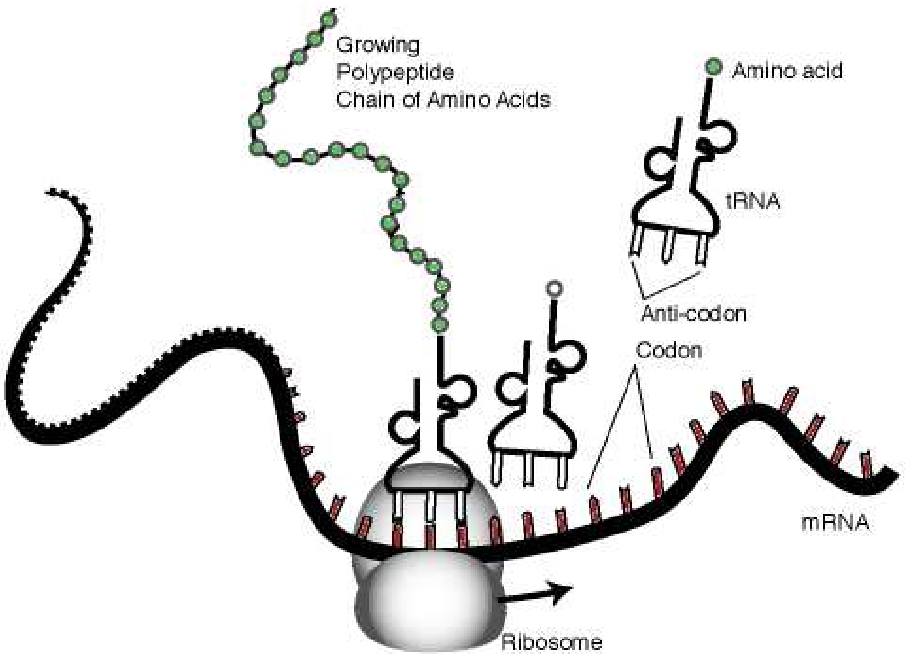
Differences from DNA

- contains ribose instead of deoxyribose
- contains uracil instead of thymine (bases A,C,G,U)
- single-stranded sequences, usually shorter
- complex secondary structure: paired complementary regions

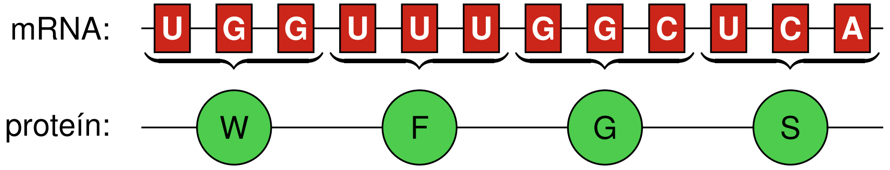


transfer RNA (tRNA), figures from Wikipedia

Translation



Codon (triple of nucleotides) determines 1 amino acid



mRNA: U G G U U U G G C U C A

proteín: W F G S

Genetic code

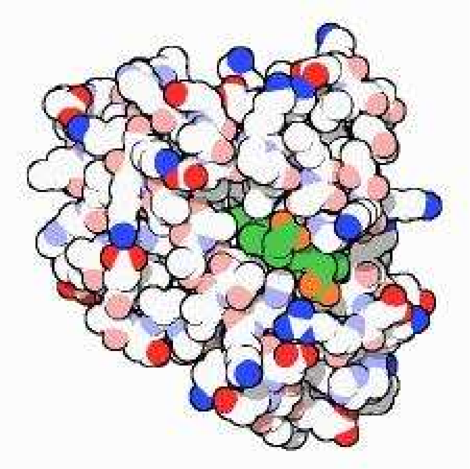
Ala / A	GCT, GCC, GCA, GCG	Leu / L	TTA, TTG, CTT, CTC, CTA, CTG
Arg / R	CGT, CGC, CGA, CGG, AGA, AGG	Lys / K	AAA, AAG
Asn / N	AAT, AAC	Met / M	ATG
Asp / D	GAT, GAC	Phe / F	TTT, TTC
Cys / C	TGT, TGC	Pro / P	CCT, CCC, CCA, CCG
Gln / Q	CAA, CAG	Ser / S	TCT, TCC, TCA, TCG, AGT, AGC
Glu / E	GAA, GAG	Thr /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

Proteins

Strings of 20 different amino acids with different chemical properties:

Amino Acid	Side chain	Its properties
Alanine (A)	-CH3	hydrophobic
Arginine (R)	-(CH2)3NH-C(NH)NH2	basic
Asparagine (N)	-CH2CONH2	hydrophilic
Aspartic acid (D)	-CH2COOH	acidic
Cysteine (C)	-CH2SH	hydrophobic
Glutamic acid (E)	-CH2CH2COOH	acidic
Glutamine (Q)	-CH2CH2CONH2	hydrophilic
Glycine (G)	-H	hydrophilic
Histidine (H)	-CH2-C3H3N2	basic
Isoleucine (I)	-CH(CH3)CH2CH3	hydrophobic
Leucine (L)	-CH2CH(CH3)2	hydrophobic
Lysine (K)	-(CH2)4NH2	basic
Methionine (M)	-CH2CH2SCH3	hydrophobic
Phenylalanine (F)	-CH2C6H5	hydrophobic
Proline (P)	-CH2CH2CH2-	hydrophobic
Serine (S)	-CH2OH	hydrophilic
Threonine (T)	-CH(OH)CH3	hydrophilic
Tryptophan (W)	-CH2C8H6N	hydrophobic
Tyrosine (Y)	-CH2-C6H4OH	hydrophobic
Valine (V)	-CH(CH3)2	hydrophobic

Protein structure

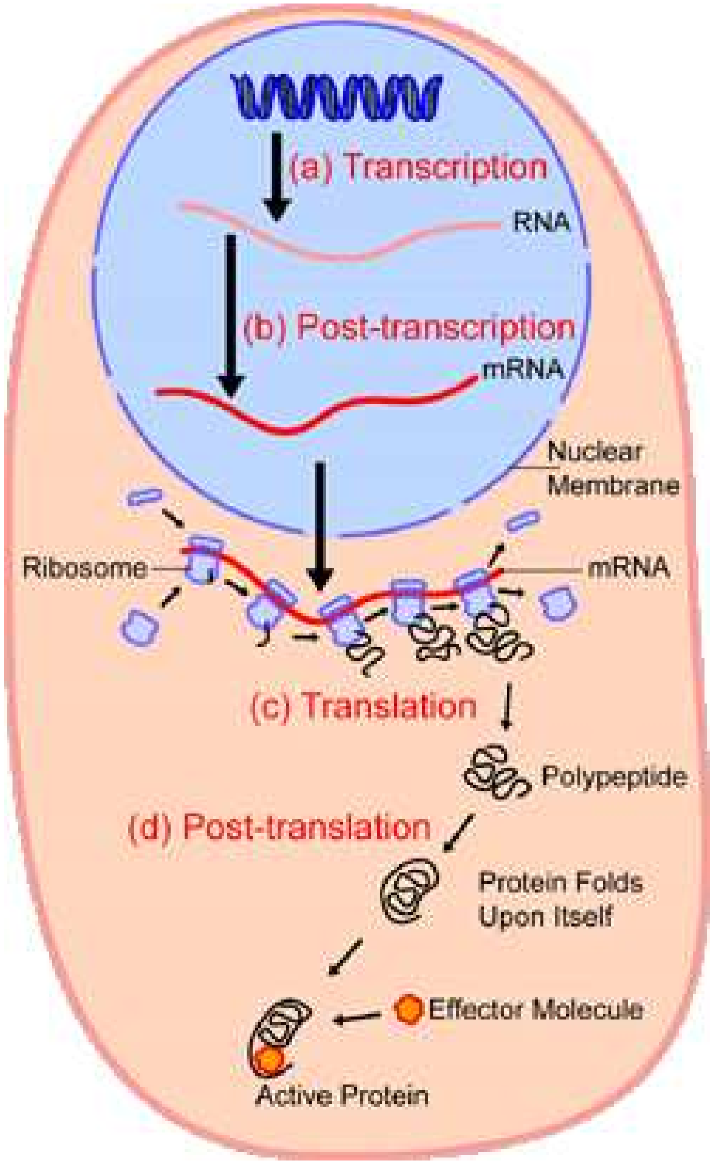


Myoglobin, the first protein with a known structure.

Proteins occur folded in a stable structure,
or move between several conformations.

Hydrophobic amino acids do not interact with water,
usually located inside the structure.

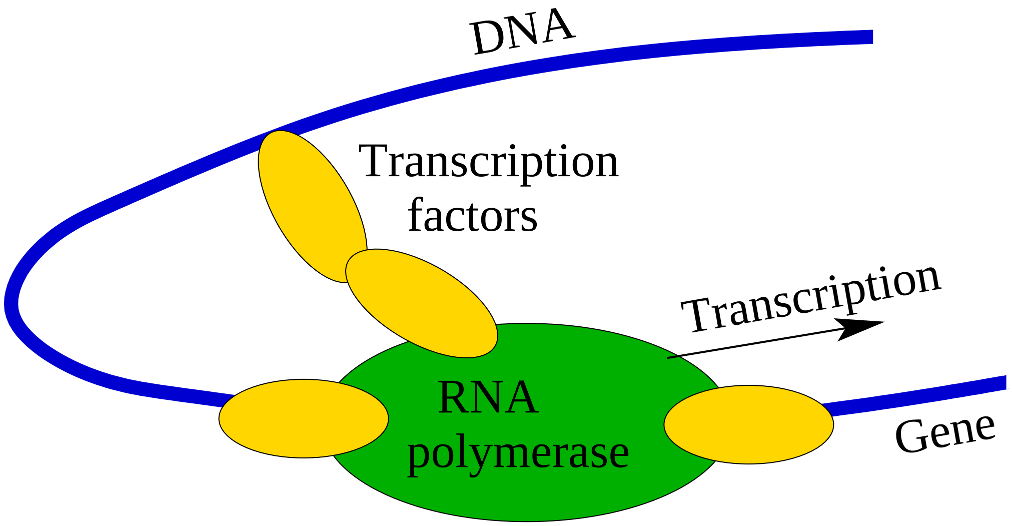
Structure of a protein determines its function.


(a) Transcription
RNA
(b) Post-transcription
mRNA
Nuclear Membrane
Ribosome
mRNA
(c) Translation
Polypeptide
(d) Post-translation
Protein Folds Upon Itself
Effector Molecule
Active Protein

Regulation of gene expression

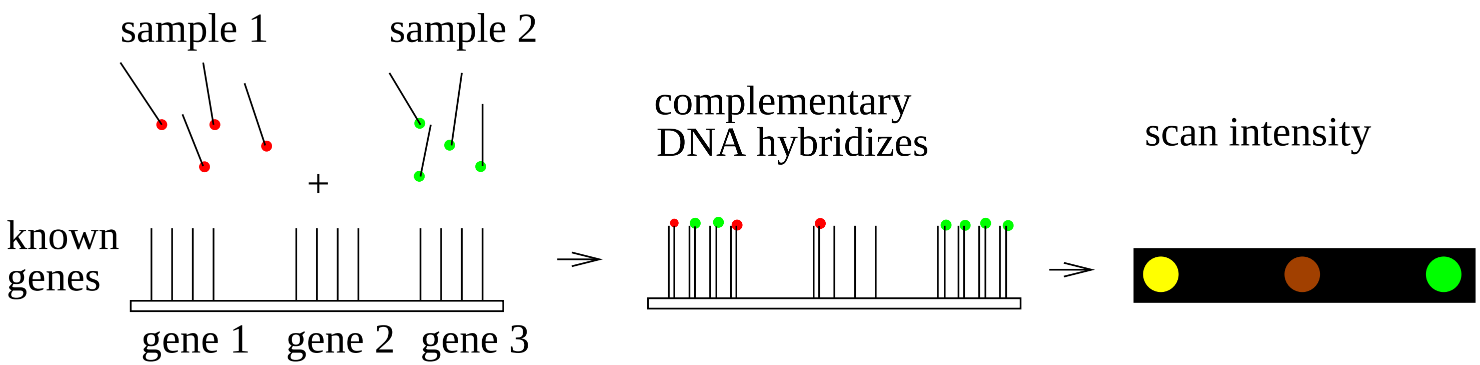
Cells in different tissues of the same organisms share the same genome, yet look and function very differently.
Some proteins are produced only under special circumstances or in variable amounts.

Regulation of transcription initiation by transcription factors:



Computational problem: determine, which factors influence a given gene and where they bind DNA.

Technology: microarray



sample 1

sample 2

+

known genes

gene 1 gene 2 gene 3

complementary DNA hybridizes

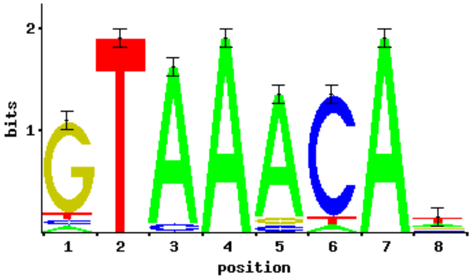
scan intensity

Measurement of the amount of mRNA in the sample for **many genes** simultaneously.

Repeat for different samples, study correlations among genes.

These could be caused by a common regulator (transcription factor).

Computational problem: for several co-regulated genes, find a motif where the common transcription factor could bind **(motif finding)**



Example of microarray data

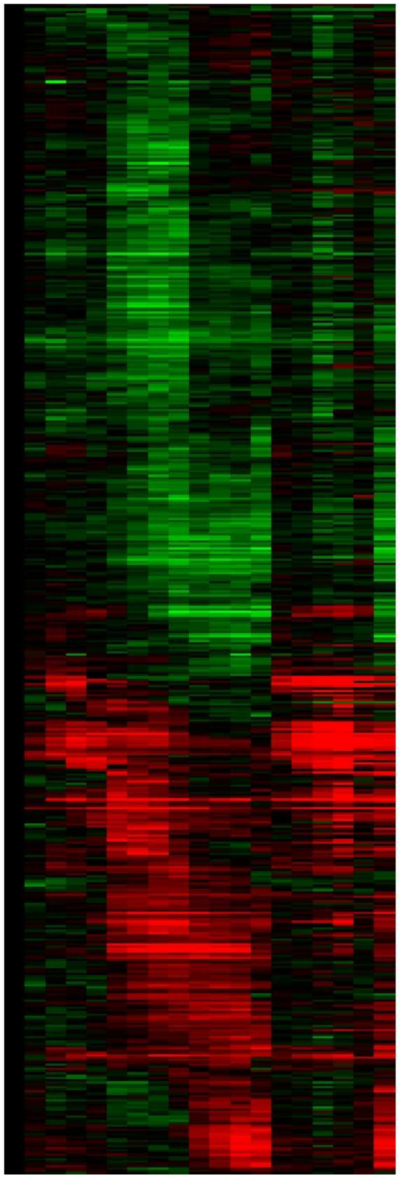
Ratio of gene expression in treated and control sample fg/bg

Red: fg>bg

Green: fg<bg

517 genes

19 experiments



Mutations in DNA

Occasionally DNA is changed, mutated
(for example due to environmental factors, replication errors).

Mutation types:

substitution (one nucleotide changes to another),
insertion (inserts several new nucleotides),
deletion (delets several nucleotides),
large scale changes (e.g. translocations).

Computational problems:

Which sequences evolved from a common ancestor?
(homology search)
Which nucleotides in two related sequences correspond to each other?
(sequence alignment)

Population genetics

Mutations are propagated in a population from parents to offspring. Dangerous mutations are quickly eliminated, advantageous mutations are more likely to spread (natural selection).

Polymorphisms: genetic differences between organisms within species. They cause differences in phenotype, e.g. appearance, genetic diseases. Sequencing several individuals within a species helps to map common polymorphisms.

Computational problem:

Determine polymorphisms linked to a certain disease or other character

Evolution

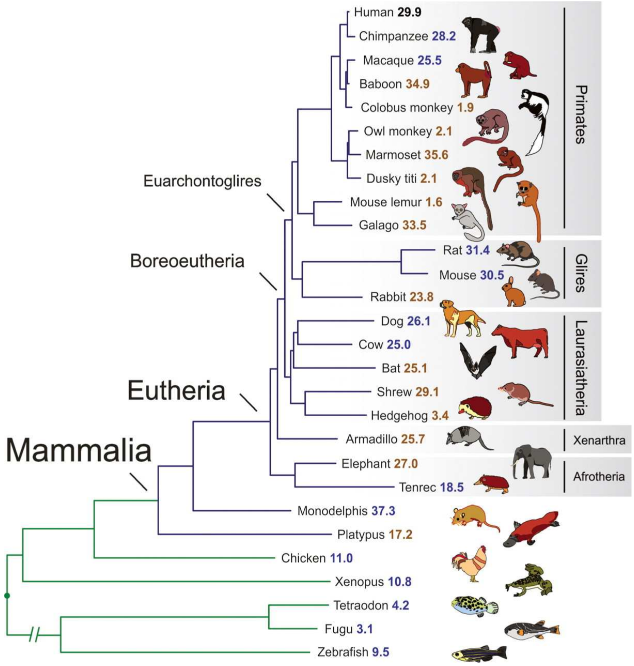
Speciation:

If a population is divided into several isolated subpopulations, genetic material is not exchanged.

Mutations accumulate independently, until mating no longer possible – new separate species.

Computational problem:

Using sequences of current species, reconstruct a **phylogenetic tree** representing their evolutionary history



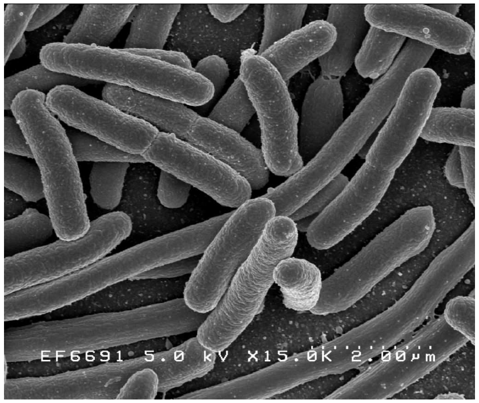
Prokaryotes vs. eukaryotes

Prokaryotes: bacteria and archea, simple unicellular organisms.
DNA directly in the cytoplasm.
Genome in one circular chromosome (plus shorter plasmids),
simple gene structure

Eukaryotes: animals, fungi, plants, some unicellular organisms.
DNA in a nucleus, cell contains also other organelles.
Mitochondria and chloroplasts are engulfed prokaryotes which became part of the eukaryotic cell.
Longer genome in several linear chromosomes.

Model organisms

Species important for biology research, explored more than other related species. General principles applicable to other species as well.



Escherichia coli: bacterium living in digestive tract. Simple to manipulate, cell division every 20 min. Study of basic biological processes: DNA replication, gene expression, etc. Genome 4.6MB, 4000 genes.

Saccharomyces cerevisiae: baker's yeast. Simple eukaryotic organism. Genome with 6000 genes, 13MB. Cell division every 2 hours. Study of processes specific for eukaryotes.

Model organisms

Arabidopsis thaliana: small flowering plant, 6-week reproduction cycle. Model organism for plant research.

Caenorhabditis elegans: small worm, nematode, living in soil. Study of development, cell differentiation.

Drosophila melanogaster: fruit fly. Study of genetics, development genes.

Vertebrates: frog Xenopus laevis (large eggs, easy to manipulates), aquarium fish Danio rerio (translucent embryos), mouse Mus musculus (many laboratory breeds with different properties).

Available data

- DNA sequences: whole genomes or their parts
- Genome annotation: location of genes and other functional elements
- RNA sequences and structures
- Protein sequences, their function and structure
- Measurements of cell state (amount of RNA, protein, etc.)
- . . .

Data obtained by experiments or by computational methods
Often unreliable, noisy (in both cases)

More information

- Zvelebil, Baum: Understanding Bioinformatics, chapter 1
- University textbooks of molecular biology
- English Wikipedia
- Tutorials linked on the course website