POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Basics of Bioinformatics - Introduction

Lecture 1. Adam Kuzdraliński, 2026/2027

National Cancer Institute definition:

“The use of computing tools to manage and analyze genomic and molecular biological data.”

Sources: NCI: Glossary Terms

Bioinformatics: the EMBL-EBI definition

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The European Bioinformatics Institute defines bioinformatics as follows:

“Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, and statistics to analyse and interpret large-scale biological data.”

In this course, we will use biological databases, analyse sequences and explore other applications.

Sources: EMBL-EBI: What is bioinformatics?

Biology, chemistry and biochemistry

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Biology studies organisms and the processes of life.

Chemistry describes substances, their structure and their transformations.

Biochemistry studies the substances and chemical reactions involved in the functioning of organisms.

Molecular biology explains life processes at the level of molecules, including DNA, RNA and proteins.

Sources: OpenStax 1.1 OpenStax 6.1

Atoms and molecules

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Atoms are components of matter, and an element is a particular type of atom.
A molecule consists of atoms held together by chemical bonds.
One water molecule, H_2O, contains two hydrogen atoms and one oxygen atom.
DNA and proteins also consist of atoms, joined into much larger structures.

Level of description	**Example**
Atom	one oxygen atom: O
Small molecule	water: H_2O, 3 atoms in total
Large molecule	DNA or a protein, many connected atoms

Sources: OpenStax 2.1

A nucleotide is also made of atoms

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A nucleotide is a building block of DNA and RNA.

It contains a sugar, a phosphate group and a nitrogenous base.

Here, “sugar” means a type of chemical compound, rather than table sugar.

The chemical structure is extra background. For the exercise, you need to recognise the nucleotide symbols.

Part of a nucleotide	**Role in a simplified description**
Sugar and phosphate group	help form the backbone of the strand
Nitrogenous base	distinguishes the units represented by letters
Example of a free nucleotide	dAMP: $C_{10}H_{14}N_5O_6P$, 36 atoms

Sources: NHGRI: Nucleotide PubChem: dAMP OpenStax 3.5

Polymers and macromolecules

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A polymer consists of many building blocks joined together.

DNA and RNA are chains of nucleotides, while proteins consist of amino acids.

We call these very large molecules macromolecules.

We also know polymers from plastics, although their structures and properties can be very different.

Molecule	**Units in the chain**
DNA	DNA nucleotides
RNA	RNA nucleotides
Protein	amino acids

Sources: OpenStax 3.1 OpenStax 3.3

The cell and its organelles

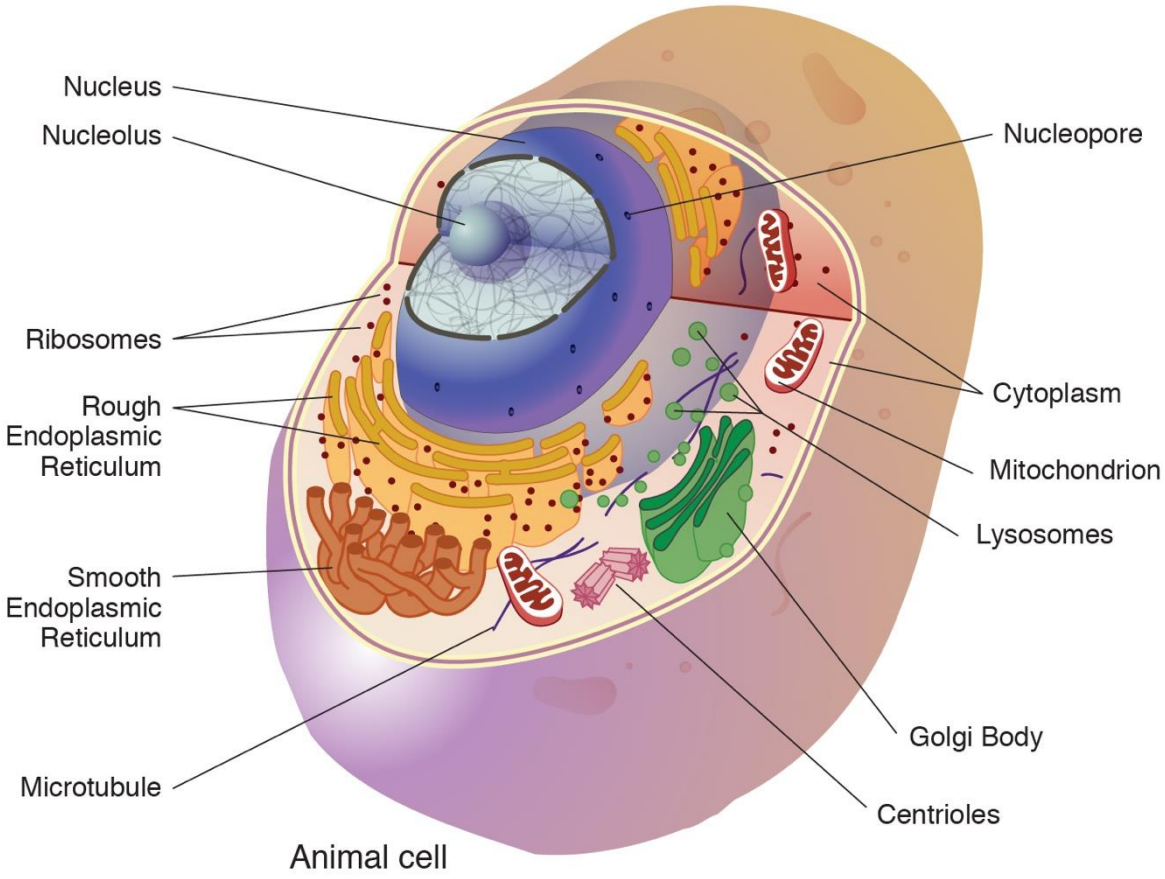
POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The cell is the basic structural and functional unit of organisms.

The cell membrane separates its interior from the surroundings.

The nucleus and mitochondria are examples of organelles, specialised structures within a cell.

Ribosomes take part in protein production.



Sources: MedlinePlus: cell NHGRI: Cell

Cells with and without a nucleus

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Typical eukaryotic cells have a nucleus. Animal, plant and fungal cells belong to this group.

Bacteria and archaea have prokaryotic cells, which lack a nucleus.

Both groups have genetic material and ribosomes.

Viruses are not cells and need host cells to make copies of themselves.

Sources: OpenStax 4.2 OpenStax 4.3 OpenStax 21.1

Species and scientific names

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A species is a basic unit in the classification of organisms.

Under the biological species concept, its members belong to populations that can interbreed and produce fertile offspring.

This definition has limits, especially for organisms that reproduce asexually.

The scientific name of humans is *Homo sapiens*. *Aequorea victoria* is the jellyfish in which green fluorescent protein, GFP, was discovered.

Sources: OpenStax 18.2 ICZN: species names UniProt: GFP P42212 Nobel Prize 2008: GFP

Evolution and sequence similarity

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Evolution involves changes in the inherited traits of populations over generations.

Common ancestry helps explain similarities in the DNA and proteins of different organisms.

Natural selection and random processes affect the frequencies of variants.

Sequence comparison can provide information about relatedness, but requires careful interpretation.

Sources: OpenStax 18.1 OpenStax 19.1

DNA and its four letters

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

DNA stores information in the order of its nucleotides.

The four basic letters in a DNA sequence are A, T, C and G.

They stand for adenine, thymine, cytosine and guanine, the bases in the nucleotides.

At this stage, you need to recognise the letters. You do not need to memorise the full names.

Example DNA sequence	Length
ATGCCGTA	8 nucleotides
AGTCCGTA	8 nucleotides, a different order

Sources: MedlinePlus: dna NCBI: FASTA

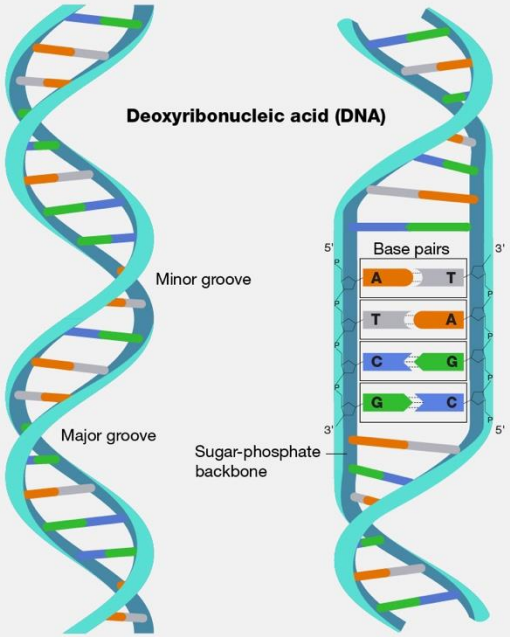
The double helix and complementarity

The two DNA strands usually wind around a common axis, forming a double helix.

A double helix resembles a twisted ladder.

Complementarity means specific base pairing: A with T, and C with G.

If we know one strand, we can work out the sequence of the strand paired with it.



Sources: NHGRI: Double Helix NHGRI: Base Pair

The 5′ and 3′ directions

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Each DNA strand has two chemically different ends, labelled 5′ and 3′.

We read these labels as “five prime” and “three prime”.

Paired strands run in opposite directions. We usually write sequences from 5′ to 3′.

The reverse complement combines base pairing with a change in the direction in which we read the sequence.

What we show	Sequence
Original strand	5′ ATGC 3′
Strand paired below it	3′ TACG 5′
Second strand in standard notation	5′ GCAT 3′

Sources: OpenStax 14.2

DNA replication

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Replication is the process of copying DNA before a cell divides.

Each separated strand can serve as a template for a new strand.

Complementarity guides nucleotide pairing, while enzymes carry out and control the process.

The laboratory method PCR makes many copies of a selected DNA fragment.

Sources: NHGRI: DNA Replication NHGRI: Polymerase Chain Reaction

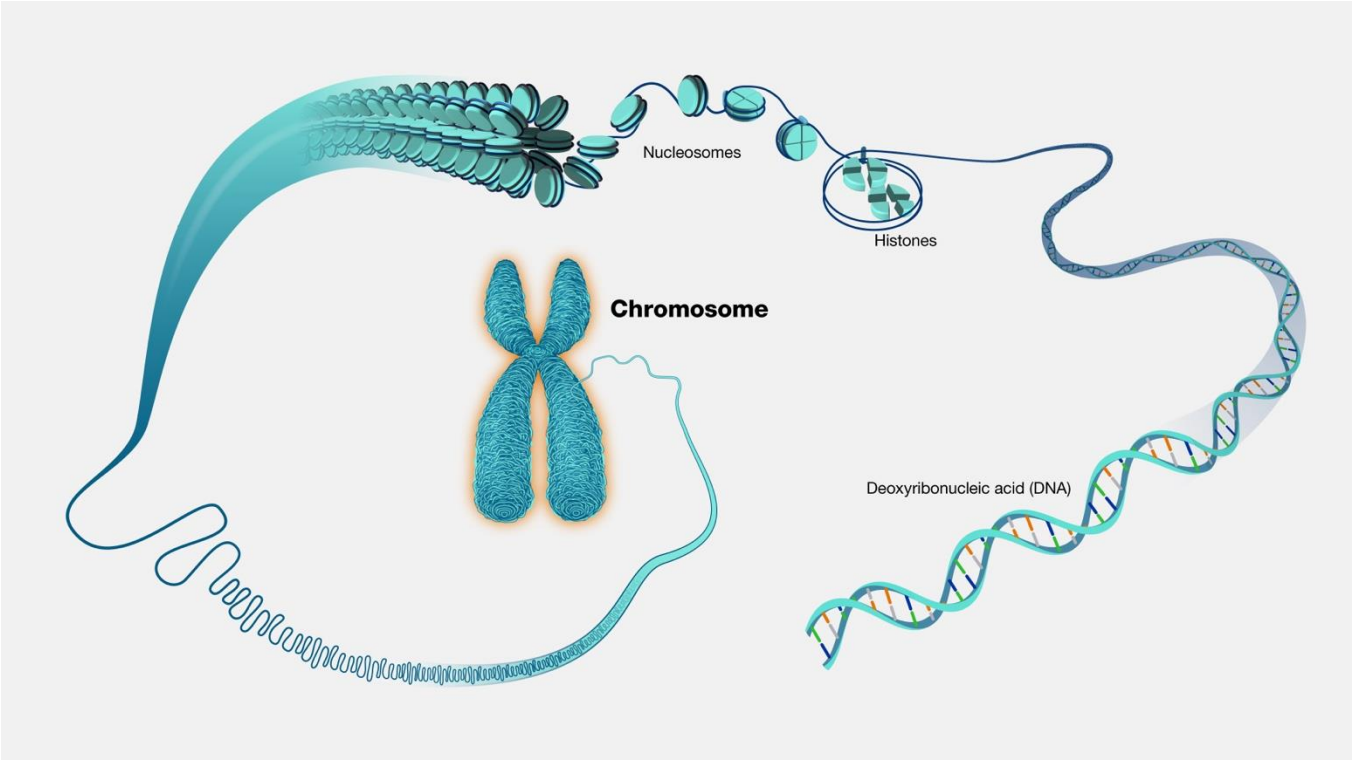
Chromosomes and the genome

A human nuclear chromosome consists of DNA associated with proteins.

The genome includes all genetic material, including regions that do not code for proteins.

A typical human somatic cell has 46 nuclear chromosomes, arranged in 23 pairs.

Chromosomes also exist between cell divisions, although their packaging looks different then.



Sources: NHGRI: Chromosome NHGRI: Genome MedlinePlus: howmanychromosomes

Gene, allele and *locus*

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A gene is a region of DNA used to produce a functional RNA, which may then be used to make a protein.

An allele is one version of a sequence at a particular place in the genome.

A *locus* is a location in the genome, such as the position of a gene on a chromosome.

One gene can influence several traits, and one trait can depend on many genes.

Sources: MedlinePlus: gene NHGRI: Allele NHGRI: Locus MedlinePlus: genelocation

RNA and messenger RNA

RNA is a chain of nucleotides represented by the letters A, U, C and G.

RNA uses uracil, represented by U, in place of the thymine found in DNA.

mRNA, or messenger RNA, carries the information needed to build a protein chain.

Other types of RNA form part of ribosomes or help deliver amino acids.

Sources: NHGRI: Ribonucleic Acid RNA NHGRI: Messenger RNA mRNA NHGRI: Transfer RNA tRNA

Transcription: making RNA

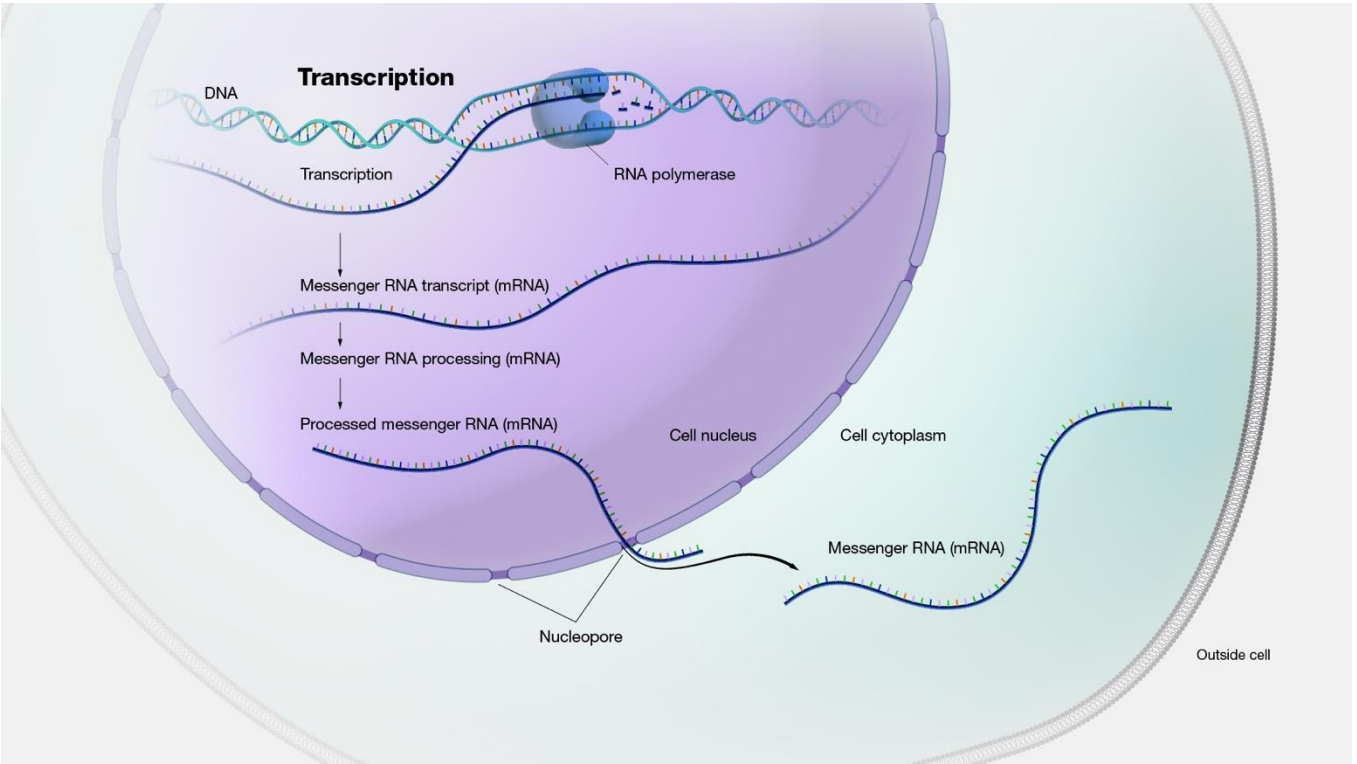
POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

During transcription, an enzyme builds RNA using one DNA strand as a template.

The new molecule consists of RNA nucleotides available in the cell.

In human cells, transcription of nuclear genes takes place in the nucleus.

After processing, mRNA enters the cytoplasm, where ribosomes can read it.



Sources: NHGRI: Transcription MedlinePlus: makingprotein

RNA processing and simplified examples

Many human genes contain introns, which the cell removes from the initial RNA.
During splicing, the cell joins the remaining exons.
For a suitable coding-strand fragment, the RNA sequence has U where the DNA has T.
This is a simplification. It does not describe the whole process of mRNA maturation.

Simplified example, without introns	Sequence
Coding DNA strand	5′ ATG TTC TAA 3′
Template DNA strand	3′ TAC AAG ATT 5′
Corresponding RNA fragment	5′ AUG UUC UAA 3′

Sources: NHGRI: ribonucleic acid fact sheet OpenStax 15.4

Amino acids and proteins

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Amino acids are the units that cells use to build protein chains.
Nucleotides make up DNA and RNA, while amino acids make up proteins.
In a basic description of proteins, we use 20 standard amino acids.
Each has a three-letter abbreviation and a one-letter symbol.

Amino acid	Three-letter code	Symbol
Methionine	Met	M
Phenylalanine	Phe	F
Leucine	Leu	L

Sources: NHGRI: Amino Acids OpenStax 3.4

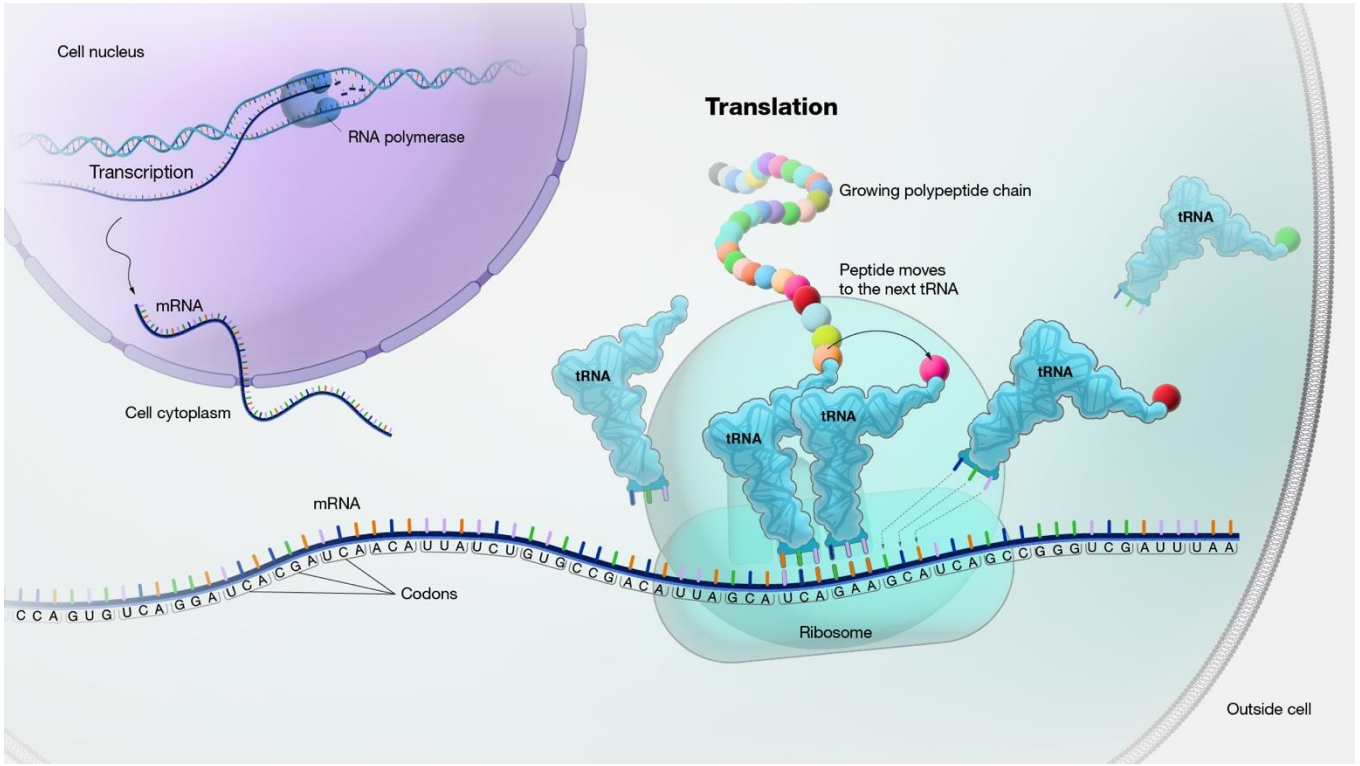
Translation at the ribosome

Translation is the reading of mRNA to build a chain of amino acids.

The ribosome moves along the mRNA and reads successive groups of three nucleotides.

tRNA molecules bring amino acids, which the ribosome joins together.

The RNA nucleotides do not turn into amino acids.



Sources: NHGRI: Translation NHGRI: Ribosome OpenStax 15.5

The genetic code and stop codons

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The genetic code assigns an amino acid or a stop signal to each codon.
A codon is a group of three mRNA nucleotides.
AUG specifies methionine and, in the right context, starts translation.
UAA, UAG and UGA signal the end of amino acid chain assembly.

mRNA codon	Meaning in the standard genetic code
AUG	methionine, and start in the right context
UUU or UUC	phenylalanine
UAA, UAG or UGA	STOP: the end of translation

Sources: NCBI: genetic code NHGRI: Stop Codon NHGRI: Codon

Reading a short mRNA sequence

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

In this simplified example, reading starts at the first AUG.
The sequence AUG UUC UAA specifies methionine, phenylalanine and the end of translation.
The product is a chain of two amino acids, written as MF.
The reading frame determines how we divide the sequence into groups of three.

mRNA	AUG	UUC	UAA
Meaning	methionine	phenylalanine	STOP
Product	M	F	no further amino acid

Sources: NCBI: genetic code NHGRI: Stop Codon NHGRI: Codon

Protein shape and function

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

An amino acid chain takes on a three-dimensional structure that affects its function.

Proteins can speed up reactions, transport substances and receive signals.

An enzyme binds particular molecules and helps transform them.

A change in a protein sequence may affect its structure or function, but does not always do so.

Sources: MedlinePlus: protein NHGRI: Enzyme NHGRI: Protein

Gene expression and types of data

Gene expression produces RNA and, for protein-coding genes, also a protein.

Similar DNA does not mean identical gene activity.

The transcriptome describes RNA, and the proteome describes proteins, in a given system at a given time.

Epigenetic modifications can regulate gene activity without changing the DNA sequence.

Type of data	Example question
DNA, genome	What is the sequence?
RNA, transcriptome	Which RNAs are present, and how much?
Proteins, proteome	Which proteins are present in the sample?

Sources: NHGRI: Gene Expression NHGRI: Transcriptome NHGRI: Epigenomics

Chromosome inheritance

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Gametes are reproductive cells: the egg cell and the sperm cell.

A typical human gamete contains 23 nuclear chromosomes.

The gametes join to form a zygote with 46 chromosomes, one set from each parent.

Meiosis reduces the number of chromosome sets and produces different combinations of inherited DNA.

Stage	Typical number of nuclear chromosomes
Egg cell	23
Sperm cell	23
Zygote after fertilisation	46, arranged in 23 pairs

Sources: NHGRI: Gamete NHGRI: Meiosis NHGRI: homologous recombination MedlinePlus: howmanychromosomes

Dominance and recessiveness

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Consider a fictional gene that affects pigment production.

“A” (capital A) is an allele that allows a working enzyme to form.

“a” (lowercase a) is an allele that does not produce a working enzyme.

One copy of “A” is enough, so AA and Aa have pigment, while aa does not.

Genotype in this model	Working enzyme	Phenotype
AA	yes	pigment present
Aa	yes, one copy is enough	pigment present
aa	no	pigment absent

Sources: OpenStax 12.2 NHGRI: Recessive NHGRI: Heterozygous NHGRI: Homozygous

An Aa × Aa inheritance example

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Each Aa parent can pass on allele A or allele a.

In this simple model, the four possible combinations are AA, Aa, aA and aa.

The probability of aa is 25%. The probability of pigment being present is 75%.

Each fertilisation is a new event. The outcomes for previous children do not have to balance to these proportions.

Gamete / gamete	**A**	**a**
A	AA	Aa
a	aA	aa

Sources: OpenStax 12.2 OpenStax 12.3

Genotype and phenotype

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A genotype describes an organism's genetic variants, such as its alleles at a given *locus*.

A phenotype includes the traits of an organism that we can observe or measure.

Many traits depend on several genes, the environment and development.

A person's appearance alone usually does not allow us to determine their genotype.

Sources: NHGRI: Genotype NHGRI: Phenotype MedlinePlus: height

A DNA variant: a difference between people

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

We compare the same DNA region in two fictional people.

At position five, one person has A and the other has G.

We describe a sequence difference relative to a chosen reference as a DNA variant.

Here we show one copy per person, although humans usually have two copies of each autosomal region.

Position in the fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Person 1: one copy	T	T	A	C	**A**	G	T	C
Person 2: one copy	T	T	A	C	**G**	G	T	C

Sources: NHGRI: Single Nucleotide Polymorphisms NHGRI: Deletion NHGRI: Genetic Marker NIST: STRBase

Deletions, STRs and genetic markers

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

People may also differ in the presence of a DNA fragment or in its number of repeats.

A deletion removes a fragment, while an insertion adds one.

An STR is a short DNA motif, such as AGAT, repeated several times in a row.

A genetic marker is a recognisable location or feature in DNA that we use to track variation in an analysis.

Two versions of the same location	Repeated DNA segment
Version with 4 repeats	AGAT AGAT AGAT AGAT
Version with 5 repeats	AGAT AGAT AGAT AGAT AGAT

Sources: NHGRI: Single Nucleotide Polymorphisms NHGRI: Deletion NHGRI: Genetic Marker NIST: STRBase

Gene symbols and variant identifiers

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

OXTR and *CCR5* are symbols for two different genes.

An identifier starting with rs points to a variant record in NCBI dbSNP.

An rs number is not a gene number, a DNA position or a measure of disease risk.

We check its biological meaning in the record description and scientific studies.

Gene symbol	Variant record identifier	Type of difference
OXTR	rs53576	a single DNA position
CCR5	rs333, known as Δ32	a deletion of 32 base pairs

Sources: NCBI: rs identifiers NCBI Gene: OXTR NCBI Gene: CCR5

Variants, traits and disease

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

OXTR is the oxytocin receptor gene, and rs53576 identifies a studied variant within it.

This variant cannot reliably tell us how empathetic a particular person is.

The *CCR5*-Δ32 deletion, recorded as rs333, changes a protein that some strains of HIV use to enter cells.

Two copies of this deletion greatly reduce susceptibility to these strains, but do not protect against all HIV.

Sources: PubMed 33892530 PubMed 8756719 NCBI: rs333

Bioinformatics in medicine

DNA, RNA and protein analysis can support disease diagnosis and treatment selection.

In cancer care, testing can examine the molecular features of tumour cells.

A screening test indicates risk and may need confirmation by another method.

Clinical interpretation combines laboratory results with information about the patient and the strength of the evidence.

Sources: NCI: biomarker testing MedlinePlus: nipt NHGRI: CRISPR

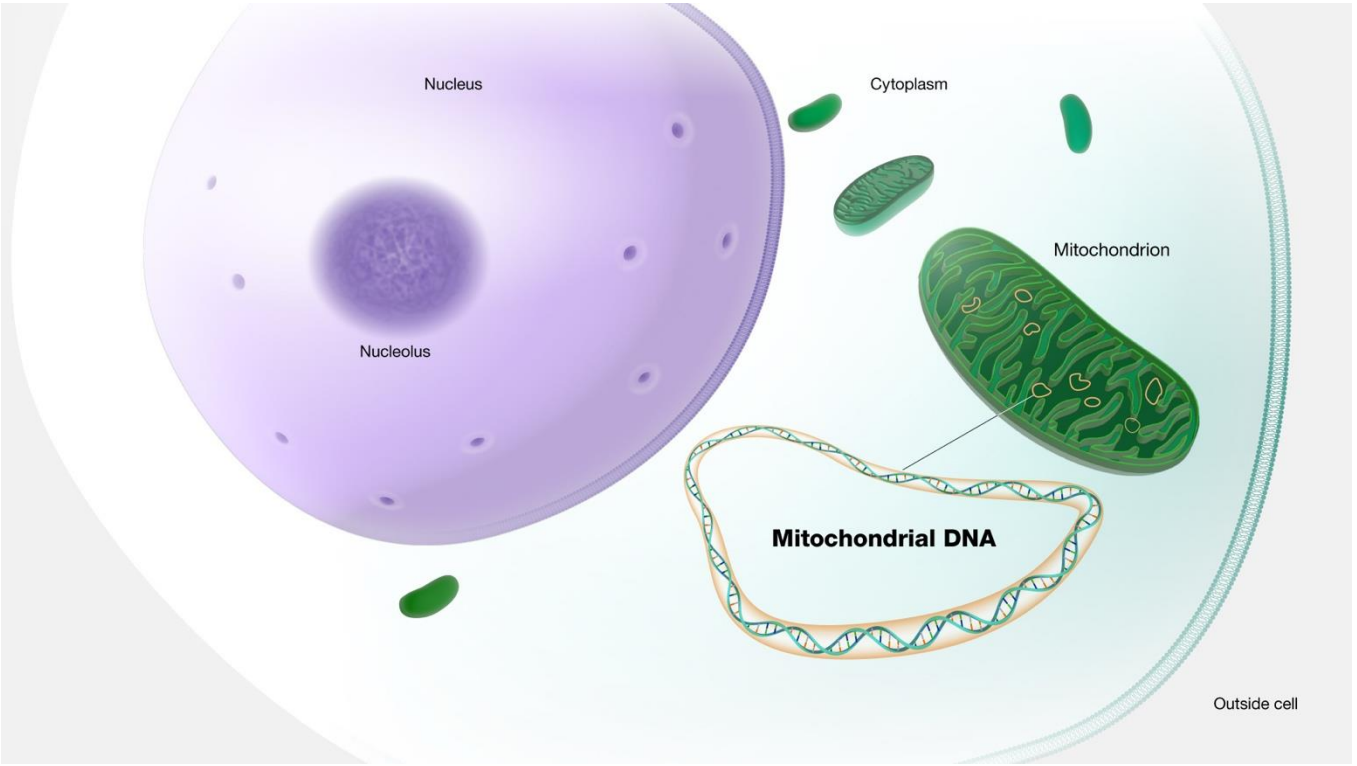
mtDNA and the maternal line

Mitochondria have their own small genome, called mtDNA.

In humans, we inherit mtDNA from our mothers.

Both daughters and sons receive it, but daughters pass it on to the next generation.

mtDNA analysis follows one line of ancestry, rather than a person's whole family history.



Sources: MedlinePlus: mitochondrial dna NHGRI: Mitochondrial DNA Lee et al. 2023

The Y chromosome and autosomal DNA

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Most of the Y chromosome follows a father-to-son line of inheritance, which differs from autosomal inheritance.

Testing this part of Y allows us to follow one paternal line.

Autosomal DNA comes from both parents and contains segments from many ancestral lines.

mtDNA, Y and autosomes answer different questions about ancestry.

Material	Typical inheritance	Information provided
mtDNA	mother to all children	one maternal line
Most of Y	father to son	one paternal line
Autosomes	both parents	many branches of the family

Sources: MedlinePlus: ancestrytesting MedlinePlus: y Mathieson and Scally 2020: What is ancestry?

Haplogroups and frequency maps

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A haplogroup brings together lineages with shared ancestry, identified through inherited DNA variants.

Genetic genealogy includes mtDNA haplogroups and Y chromosome haplogroups.

R1a is an example of a Y haplogroup name.

A haplogroup frequency map does not establish a person's nationality or complete ancestry.

Sources: MedlinePlus: ancestrytesting YHRD

STRs and paternity testing

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

At a typical autosomal marker, a child inherits one allele from each parent.

The numbers in an STR profile describe how many times a short DNA motif repeats.

Comparing many markers helps assess whether results support a proposed family relationship.

A match at one marker does not prove paternity. Real tests use statistics and consider possible mutations.

Person	Alleles at one fictional STR marker
Mother	10 / 12
Child	12 / 14
Candidate A	10 / 14
Candidate B	10 / 11

Sources: NHGRI: DNA Fingerprinting NIST: STRBase

Sequencing and measurement results

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Sequencing uses measured signals to determine the order of nucleotides.

A read represents a segment of the molecule being analysed.

The result may contain errors or uncertain positions marked with N.

Reconstructing a longer sequence requires analysis of many reads and quality control.

Sources: NHGRI: DNA Sequencing EMBL-EBI: sequencing and FASTQ

The FASTA format

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

FASTA is a simple text format for biological sequences.

A line starting with > contains an identifier and an optional description.

The following lines contain the sequence, up to the next header or the end of the file.

The description should help us establish whether we are analysing DNA, RNA or a protein.

```
>example_DNA_1 teaching fragment
ATGCCGTA
>example_DNA_2 teaching fragment
ATGTCGTA
```

Sources: NCBI: FASTA

Databases and reference sequences

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A database record links a sequence to a description, an identifier and information about its origin.

NCBI GenBank stores nucleotide sequences, while UniProt provides information about proteins.

A reference sequence provides a standard point of comparison.

The version number and download date help us reproduce an analysis.

Sources: NCBI: GenBank NCBI: RefSeq UniProt: GFP P42212 UniProt: record status

Sequence alignment

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A sequence alignment arranges DNA or protein sequences so that we can compare corresponding positions.

In this example, position four differs: C in one sequence and T in the other.

Seven of the eight positions match, giving 87.5% identity.

This percentage describes the fragments being compared, rather than the similarity of entire organisms.

Position in the fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Sequence 1	A	T	G	**C**	C	G	T	A
Sequence 2	A	T	G	**T**	C	G	T	A

Sources: NCBI: alignment and scoring NCBI: BLAST Help

BLAST: queries and search results

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

BLAST searches a database for sequences similar to a query sequence.

The query is the sequence we submit. A hit is a database result with a matching region.

blastn compares nucleotide sequences, while blastp compares protein sequences.

A result can suggest an origin or function, but we need to assess that suggestion in context.

Sources: NCBI: BLAST Help UniProt: GFP P42212

BLAST: identity, coverage and E-value

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Identity is the proportion of compared positions that match exactly.

Query coverage is the proportion of our sequence included in the alignment.

The E-value estimates how many alignments with this score or better we would expect by chance in the database search.

Even an excellent match for one sequence may not identify a single species.

Result	Identity	Query coverage
20 of 100 positions aligned	20/20 = 100%	20/100 = 20%
100 of 100 positions aligned	100/100 = 100%	100/100 = 100%

Sources: NCBI: BLAST FAQ NCBI: alignment statistics

Protein structure and AI predictions

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A protein sequence is a starting point for studying its structure and function.

The PDB provides structures determined using experimental methods.

AlphaFold provides structure predictions together with measures of model confidence.

A predicted structure does not automatically replace experimental evidence of function.

Sources: RCSB PDB: methods EMBL-EBI: faq

The microbiome and mixed samples

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The microbiota is a community of microorganisms in a particular environment.

The term microbiome usually also includes their genetic material and environmental context.

A gut, soil or water sample may contain DNA from many organisms.

Metagenomics analyses the genetic material of such a community.

Sources: Berg et al. 2020: microbiome and microbiota NHGRI: Metagenomics

Further work with Python and R

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Python automates work with sequence files and provides access to libraries such as Biopython.
R supports statistical analysis of measurement tables and the creation of plots.
Both languages can perform some of the same tasks.
A script records the steps of an analysis so that we can check and repeat them.

Example task	What the student needs to understand
Counting sequence length	whether we are counting nucleotides or amino acids
G and C content in ATGC	2 of 4 positions, or 50%
Comparing RNA measurements	meaning of the value and the sample it comes from

Sources: Biopython: Tutorial & Cookbook R Project: What is R?

POLSKO-JAPOŃSKA
AKADEMIA TECHNIK
KOMPUTEROWYCH

Thank you for your attention

Basics of Bioinformatics

Lecture 1