POLSKO-JAPOŃSKA
AKADEMIA TECHNIK
KOMPUTEROWYCH

Bioinformatics Databases

Basics of Bioinformatics

Lecture 2

Adam Kuzdraliński

2026/2027

From a question to a dataset

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

First define the question, then choose a database.

Decide what one record represents and which information the analysis needs.

Download a small example, check its contents and only then expand the dataset.

Sources: Ensembl: BioMart

What counts as biological data?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A sequence records the order of nucleotides or amino acids.

A three-dimensional structure describes the arrangement of atoms in a molecule.

A measurement table may show the amount of RNA in samples.

Each type of data needs to be read in a different way.

Sources: UniProt: GFP P42212 | RCSB PDB: APIs | NCBI GEO: data organisation

Records, metadata and annotations

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A record describes a particular object in a database.

Metadata explain details such as the data source, organism and method used to obtain the data.

Biological annotations assign meaning to data, for example by identifying a protein-coding region.

The distinction between metadata and annotations depends on the context.

Sources: NCBI: GenBank record guide | UniProt: evidence

Archives and reference collections

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Archives store data submitted by researchers, such as sequences in GenBank.

Reference collections and knowledge resources organise and describe selected data, as in RefSeq and UniProtKB.

Archives also apply quality checks.

A database's usefulness depends on the question, not on a simple ranking.

Sources: NCBI: GenBank | NCBI: RefSeq | UniProtKB: overview

NCBI and EMBL-EBI are not single databases

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

NCBI and EMBL-EBI provide many databases and tools.

GenBank, NCBI Gene and GEO store different kinds of information.

Links between records help us move from a gene to a protein, structure or publication.

Moving to another database often means moving to a different type of object.

Sources: NCBI Datasets: genes | EMBL-EBI: training

Record identifiers and versions

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

An identifier points to a specific record within a database.

In NC_012920.1, the suffix ".1" is the RefSeq sequence version.

An updated description does not always change the sequence version.

For a project, also record the download date and the database release, if available.

Sources: NCBI: GenBank record guide | NCBI: RefSeq

Genes, transcripts and proteins have different IDs

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A transcript is an RNA molecule produced by copying information from DNA.

One gene can produce different transcripts.

Some transcripts are used to produce proteins, while others have different functions.

Do not assume that one gene equals one transcript equals one protein.

Sources: NCBI: RefSeq | Ensembl: BioMart

Choose a database for the type of data

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

These are starting points, not a list of all available databases.

One main source and its documentation are usually enough for a single project.

Data needed	Starting point
Protein sequences and descriptions	UniProtKB
DNA and RNA sequences	NCBI GenBank / RefSeq
Three-dimensional structures	RCSB PDB
Measurements from samples	NCBI GEO

Sources: UniProtKB: overview | NCBI: RefSeq | RCSB PDB: APIs | NCBI GEO: data organisation

Selecting part of a database: three operations

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Filtering selects records, such as proteins from a particular species.

Field selection limits the columns, for example to IDs and sequence lengths.

Selecting a sequence range retrieves specific positions within one record.

These operations answer different questions and can be combined.

Sources: UniProt: API examples | Ensembl: BioMart | NLM: E-utilities

Species as a search criterion

Scientific names and taxonomy identifiers help specify the organism.

Homo sapiens has the NCBI taxonomy identifier 9606.

In UniProtKB, organism_id:9606 selects records assigned to this organism.

The word “human” anywhere in a description is not an equivalent filter.

Sources: UniProt: search fields | UniProt: API examples

A name, an identifier or a sequence?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A known ID leads to a specific record.

Names and filters are used to search descriptions.

BLAST compares an input sequence with sequences in a selected collection.

Similar descriptions and similar sequences are different kinds of matches.

Sources: UniProt: GFP P42212 | UniProt: search fields | NCBI: BLAST Quick Start

One record: the GFP protein

P42212 describes green fluorescent protein, GFP, from the jellyfish *Aequorea victoria*.

The record contains a sequence of 238 amino acids and information about the protein.

This is the same type of example we encountered when comparing sequences.



Sources: UniProt: GFP P42212

“Reviewed” is not the same as experimental evidence

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

UniProtKB/Swiss-Prot contains manually curated records labelled reviewed.

UniProtKB/TrEMBL contains unreviewed records, annotated mainly automatically.

Reviewed status does not mean that every detail has been confirmed experimentally.

Check the evidence for each annotation separately.

Sources: UniProtKB: overview | UniProt: evidence

Building a UniProtKB query

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

We want human protein records with reviewed status.

AND means that a record must meet both conditions.

First check a few results in the browser.

Record counts may change between database releases.

```
(organism_id:9606) AND (reviewed:true)
```

Sources: UniProt: search fields | UniProt: API examples

Selecting only the columns we need

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

For a length comparison, we need protein IDs, names and lengths.

In the UniProt API, the fields parameter selects the columns.

Check field names in the documentation, as they may differ from labels on the website.

Protein sequence length is expressed as the number of amino acids.

```
fields=accession,protein_name,length
```

Sources: UniProt: return fields

Choose a format for the next step

FASTA stores sequences with headers.

TSV stores a table with tab-separated columns.

JSON preserves the structure of fields, lists and objects.

Changing a file extension does not change its format ☺

Sources: UniProt: API examples

FASTA: a header and a sequence

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The “>” character starts the record header.

The following lines contain the sequence, here using amino acid symbols.

Only the first 60 amino acids of GFP are shown, not the whole protein.

One file can contain many FASTA records.

```
>P42212 GFP | fragment 1-60 | shortened header
MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL
```

Sources: GFP: full FASTA

TSV: a small table we can check

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The IDs and lengths below come from an actual API response.

Three records from the downloaded results page are shown.

The first results are not a random sample and cannot describe all human proteins.

Record ID	Length [amino acids]
A0A0C5B5G6	16
A0A1B0GTW7	788
A0JNW5	1464

Sources: UniProt: tested query

Websites, APIs and downloadable files

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A website helps us inspect records and choose filters.

An API, or application programming interface, lets a program retrieve data.

Prepared bulk files are useful for larger or repeated analyses.

Manual export is a valid way to obtain data if we document the steps.

Sources: UniProt: API examples | NCBI Datasets: genes

What do the query parameters mean?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

The endpoint address identifies the service that handles the query.

query selects records, fields selects columns and format specifies the response format.

size=10 limits the number of records on one results page.

These parameters belong to the UniProt API, not to every database.

```
Endpoint: https://rest.uniprot.org/uniprotkb/search
query: (organism_id:9606) AND (reviewed:true)
fields: accession,protein_name,length
format: tsv     size: 10
```

Sources: UniProt: API examples | UniProt: tested query

Our first download with Python

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

This example requires Python 3 and the requests library.

It downloads one page, containing at most 10 records.

We check the response status and save the data as a UTF-8 file.

```
import requests
url = "https://rest.uniprot.org/uniprotkb/search"
p = {"query": "(organism_id:9606) AND (reviewed:true)",
     "fields": "accession,protein_name,length",
     "format": "tsv", "size": 10}
r = requests.get(url, params=p, timeout=30)
r.raise_for_status()
with open("proteins.tsv", "w", encoding="utf-8") as f:
    f.write(r.text)
```

Sources: UniProt: API examples | Requests: documentation

The first page is not the complete dataset

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Pagination splits the results into pages of records.

The UniProt API provides the next page's URL in the Link header with the relation next.

Keep retrieving pages until there is no next-page link.

Check record counts, unique IDs and the reported total number of results.

Sources: UniProt: pagination | UniProt: API examples

When the server does not respond as expected

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Error 400 may indicate an invalid query, while 404 means the requested resource was not found.

Status 429 indicates a request rate limit.

Retry with a delay and follow the service documentation.

Also check the contents, because status 200 does not guarantee the data we expected.

Sources: Requests: documentation | RCSB PDB: APIs | UniProt: API examples

Could we reproduce this download next month?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Record the database, query, fields, format, date and data version or release.

Keep the downloaded file and a record of later transformations.

A file checksum, such as SHA-256, helps detect changes.

Check the data usage terms and how to cite the source.

Sources: UniProt: licence | GO FAIR: FAIR principles

NCBI Gene: starting with a gene

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

NCBI Gene 351 is the record for the human *APP* gene.

APP codes for amyloid precursor protein.

A gene record connects its location, transcripts, proteins and publications.

A Gene ID does not identify one particular sequence.

Sources: NCBI Gene: APP (351) | NCBI Datasets: genes

GenBank and RefSeq: similar data, different purposes

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

GenBank archives nucleotide sequences submitted by researchers.

RefSeq provides curated reference sequences for DNA, RNA and proteins.

The collections differ in coverage and in how records are created.

State which resource supplied each sequence in your analysis.

Sources: NCBI: GenBank | NCBI: RefSeq | GRC: human genome

A sequence and its GenBank annotations

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

FASTA preserves the sequence and a short header.

The full record also describes the positions and meaning of sequence regions.

CDS identifies the region or regions that code for a protein sequence.

GFF3 stores sequence features in a tabular format with positions.

Sources: NCBI: GenBank record guide | NCBI Datasets: package contents

Downloading a selected DNA fragment

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

NC_012920.1 describes the human mitochondrial DNA reference sequence.

We retrieve positions 1 to 60, inclusive, on the strand selected by strand=1.

The result is 60 nucleotides, not the whole mitochondrial genome.

Document the range, direction and record version.

```
id=NC_012920.1   seq_start=1   seq_stop=60
strand=1   rettype=fasta   retmode=text

GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
```

Sources: NCBI: mtDNA fragment 1-60 | NCBI: EFetch parameters

NCBI Datasets: a ready-made data package

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

NCBI Datasets provides organised packages for genes or genomes.

A package may contain sequences, annotations and reports describing the data.

Before downloading, select the organism, scope and file types you need.

A small project does not require a whole genome and every associated file.

Sources: NCBI Datasets: genes | NCBI Datasets: genomes | NCBI Datasets: package contents

E-utilities: searching and downloading

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

ESearch finds identifiers in a selected NCBI database.

EFetch retrieves data for specified identifiers.

An API key is not required for ordinary use of E-utilities.

Standard limits are up to 3 requests/s without a key and 10 requests/s with a key.

Sources: NLM: E-utilities | NCBI: E-utilities guidelines (PDF)

Ensembl BioMart: filters and attributes

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

BioMart creates selected data tables without writing code.

Dataset selects the collection, filters restrict records and attributes select the output fields.

You can start with your own list of gene identifiers or names.

Check the preview and result count before exporting.

Sources: Ensembl: BioMart | Ensembl: BioMart tutorial

Positions need a reference system

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Describe a region's location together with the reference sequence and its version.

In 1-based inclusive coordinates, positions 101–200 contain 100 positions.

In BED, the same region has boundaries 100 and 200, with the end excluded.

Check the coordinate convention before comparing numbers.

Sources: Ensembl: BED format | UCSC: BED coordinates

Joining databases: check the relationships

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Identifier mapping connects IDs used in different resources.

One ID may have several matches or none at all.

After joining tables, check row counts, missing matches and duplicates.

Do not join records only because their names sound similar.

Sources: UniProt: ID mapping | Ensembl: BioMart

UniRef and UniParc: reducing redundancy

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

UniParc stores unique protein sequences and tracks their presence in source databases.

UniRef groups sequences using defined similarity criteria.

A sequence, a UniProtKB record and a UniRef cluster are different units of analysis.

Changing the unit changes the counts and their interpretation.

Sources: UniProt: UniRef | UniProt: UniParc

PDB: from sequence to structure

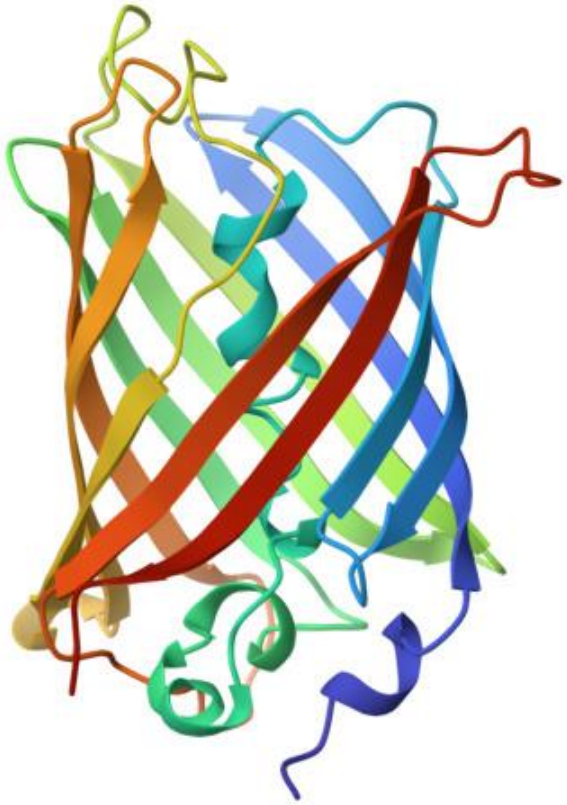
POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

PDB stores models of three-dimensional structures determined experimentally.

Record 1GFL describes a GFP structure.

One record may contain several chains and additional molecules.

The number of structures is not the number of different proteins.



Sources: RCSB PDB: GFP 1GFL | PDB-101: guide

RCSB PDB: search, data and files

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Search API finds identifiers for objects that meet the search conditions.

Data API returns information about selected objects.

PDBx/mmCIF files contain coordinates and structure descriptions.

A project about research methods may only need a metadata table.

Sources: RCSB PDB: APIs | RCSB PDB: structure files | RCSB: 1GFL JSON record

Experimental or predicted structure?

RCSB also provides computed structure models from external resources.

A computer prediction is a different type of result from an experimentally based structure.

Check the method and the quality measures used by each source.

Do not treat all quality scores as a single scale.

Sources: RCSB PDB: APIs | PDB-101: guide

GEO: how much RNA was detected?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Gene expression uses the information in a gene to produce RNA and, for protein-coding genes, protein.

Many experiments measure RNA amounts to study gene activity.

GEO stores such data along with descriptions of the experiments.

The amount of RNA is not a direct measurement of protein abundance.

Sources: NCBI GEO: data organisation

GSE, GSM and GPL: three levels of description

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

GSE identifies a series of related samples, usually within a study.

GSM describes a particular sample and its associated data.

GPL identifies the platform: the technology or measurement system.

We need both numerical data and information about what we are comparing.

Sources: NCBI GEO: data organisation | NCBI GEO: downloads

A measurement matrix needs sample descriptions

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A matrix has genes or measured features in rows and samples in columns.

A second table describes the samples and their groups.

The numbers below are fictional and only illustrate the table layout.

Before analysis, check the units, normalisation and meaning of missing values.

Feature (fictional data)	S1: control	S2: treated
GENE_X	120	240
GENE_Y	80	70

Sources: NCBI GEO: data organisation | NCBI GEO: downloads

Raw or processed data?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Sequencing reads represent sequence fragments, and their length depends on the technology.

Raw reads are available in archives such as SRA and ENA.

GEO may provide processed matrices and supplementary files.

For a first project, choose data you can prepare within the available time.

Sources: NCBI GEO: downloads | ENA: downloads

PubMed: where does the interpretation come from?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

PubMed helps find publications describing data, methods and interpretations.

A bibliographic record is not the same as a full-text article.

Check how the search query was interpreted and follow links to the data.

Publications and database records complement each other but serve different purposes.

Sources: NLM: PubMed Help

Where can we find project instructions?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Start with the database providers' documentation and examples.

Check the date, parameter names and expected response format.

The clickable links below lead to the instructions used in this lecture.

Resource	Useful documentation
UniProt	API queries and pagination
Ensembl	BioMart: selecting and exporting data
NCBI Datasets	Data package tutorials
RCSB PDB	API overview

Sources: UniProt: API examples | Ensembl: BioMart tutorial | NCBI Datasets: tutorials | RCSB PDB: APIs

What does a database record actually prove?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

A missing record does not prove that a biological phenomenon is absent.

Entry counts also reflect research interest and how the database is built.

Missing values, zeros and negative findings are different situations.

Limit conclusions to the dataset you actually analysed.

Sources: UniProt: evidence | PDB-101: guide

Project 1: lengths of recorded protein sequences

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Compare protein sequence length distributions for two species in UniProtKB.

For both species, select reviewed records and the main sequence displayed by UniProt, labelled “canonical”.

Download IDs, organisms and lengths, then describe the counts and distributions.

The result describes the selected records, not every aspect of the species' biology.

Sources: UniProtKB: overview | UniProt: API examples

Project 2: how were protein structures studied?

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

In RCSB PDB, select a time period and explicitly restrict the record types.

Download IDs, release dates and experimental methods.

Compare the counts and proportions of methods across years.

One PDB entry is the unit of analysis, even if its protein has been studied before.

Sources: RCSB PDB: APIs | PDB-101: guide

Project 3: samples from one GEO study

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Choose a well-described study with a processed matrix and metadata.

Check the groups, independent biological replicates and measurement scale.

Start by describing data quality and selected differences between samples.

Statistical inference requires a method suited to the study design.

Sources: NCBI GEO: data organisation | NCBI GEO: downloads

Before we approve a project topic

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

We can name the unit of analysis, data source and record selection criteria.

We have a sample file, understand its format and can read the required fields.

We know how to produce the result and reproduce each step.

Python or R helps us organise, analyse and present the data.

Sources: NCBI Datasets: tutorials | UniProt: API examples

POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

Thank you for your attention

Question → data selection → checks → analysis → interpretation

Basics of Bioinformatics | Lecture 2