University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Uniwersytet Warszawski

Architecture of large projects in bioinformatics (ADP)

Lecture 02

Łukasz P. Kozłowski

lukaskoz@mimuw.edu.pl

Warsaw, 2024

1. Data formats in bioinformatics

2. Popular software libraries (BioPerl, BioPython)

3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)

4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

1. Data formats in bioinformatics

2. Popular software libraries (BioPerl, BioPython)

3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)

4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

The essay: mini-review about specific bioinformatics topic

Exemplary subjects
Review about available software for:
– Structural biology (proteins, RNA, drugs)
– Phylogenetics
– NGS
– Chemoinformatics
– Data warehouse in bioinformatics (e.g. Biomart)
– Genomics (e.g. chip-seq)
– Machine learning (clustering, classification, deep learning, etc.)
– Image processing from microscopes/scanners etc.
– own suggestions ...?

You had 1 week to decide/find the subject.

Please send your proposition to ***lukaskoz@mimuw.edu.pl*** *with the email subject* ***ADP24_essay_Surname_Name***

Popular software libraries

Perl → BioPerl

Popular software libraries

Perl → BioPerl

Php → BioPHP

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Julia → BioJulia

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Julia → BioJulia

JavaScript → BioJS

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Julia → BioJulia

JavaScript → BioJS

Python → BioPython

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Julia → BioJulia

JavaScript → BioJS

Python → BioPython, but also Cogent3, bioconda

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Julia → BioJulia

JavaScript → BioJS

Python → BioPython, but also Cogent3, bioconda

Never restrict yourself only to Bio* libraries

Popular software libraries

Perl → BioPerl

Php → BioPHP

Java → BioJava

R →Bioconductor

Rust → Rust-Bio

C++ → Bio++

Julia → BioJulia

JavaScript → BioJS

Python → BioPython, but also Cogent3, bioconda

Never restrict yourself
only to Bio* libraries

Popular software libraries

Python → BioPython, but also Cogent3, bioconda

Frequently solving bioinformatic problem also means that you will use some custom, small libraries (often with multiple bugs)

Popular software libraries

Python → BioPython, but also pyCogent, bioconda

Frequently solving bioinformatic problem also means that you will use some custom, small libraries (often with multiple bugs)

For statistics and machine learning:

LIBSVM

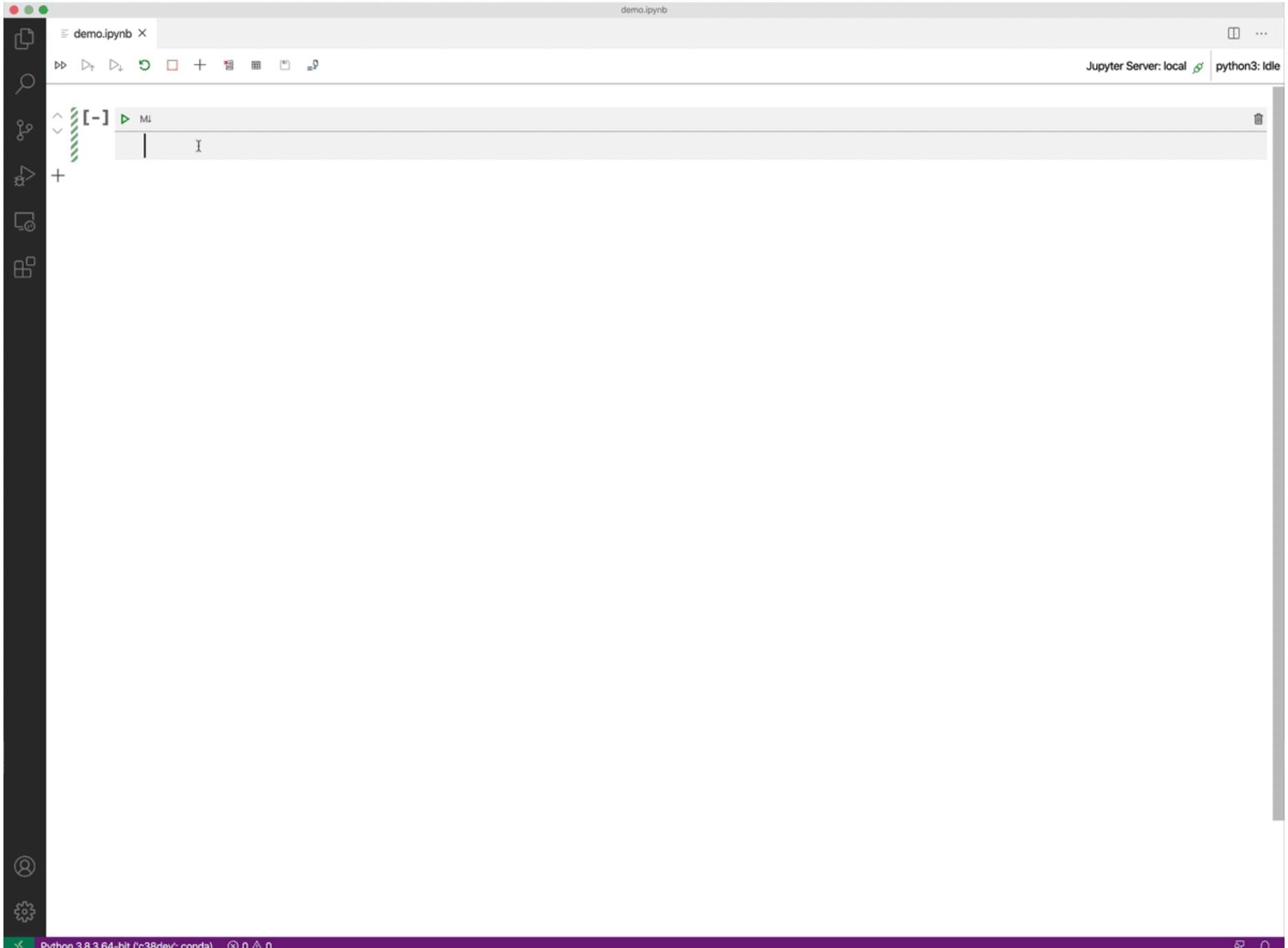
Popular software libraries

Popular software libraries

Popular software libraries

Cogent3 (as an alternative to biopython)

Popular software libraries



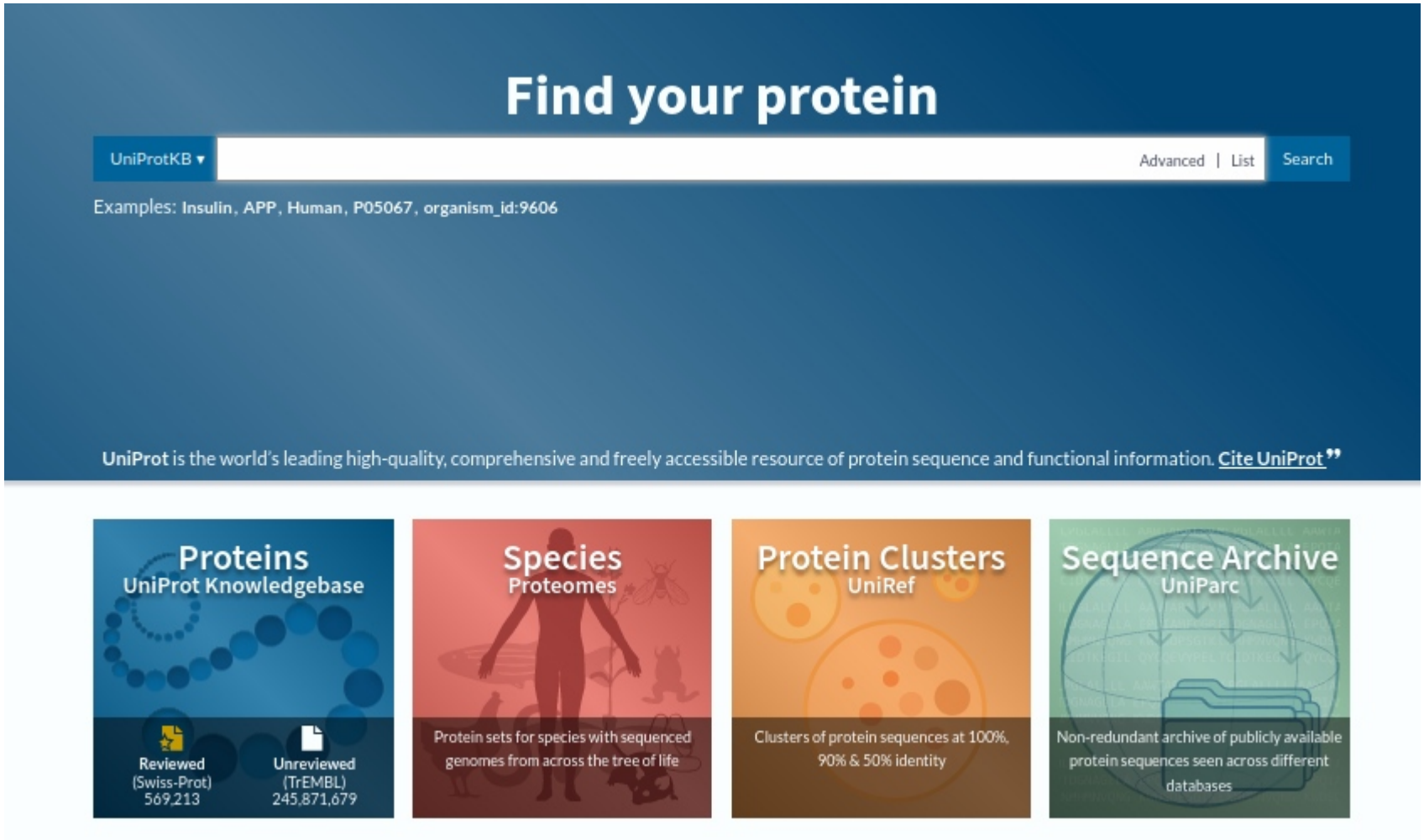
Most important bioinformatics databases

Most important bioinformatics databases

UniProt

> 230M

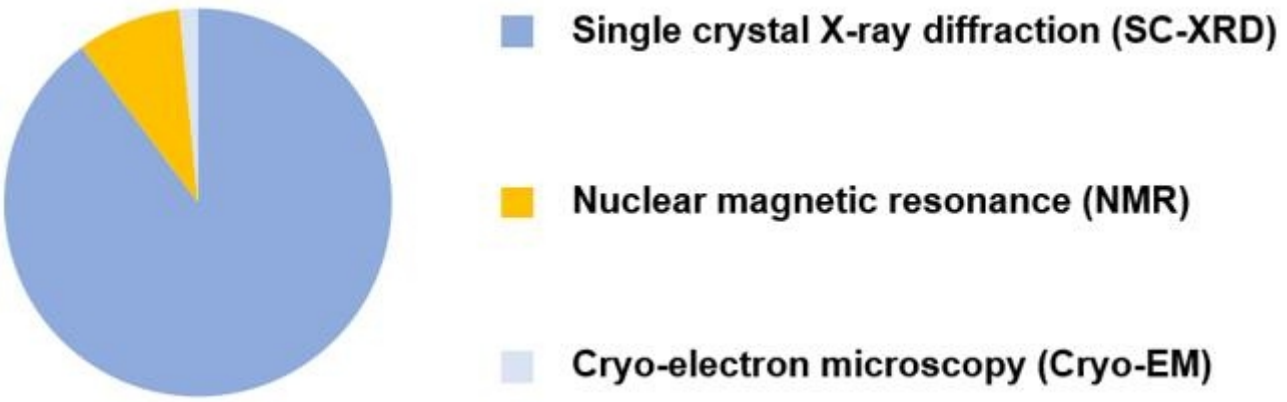
https://www.uniprot.org



Most important bioinformatics databases

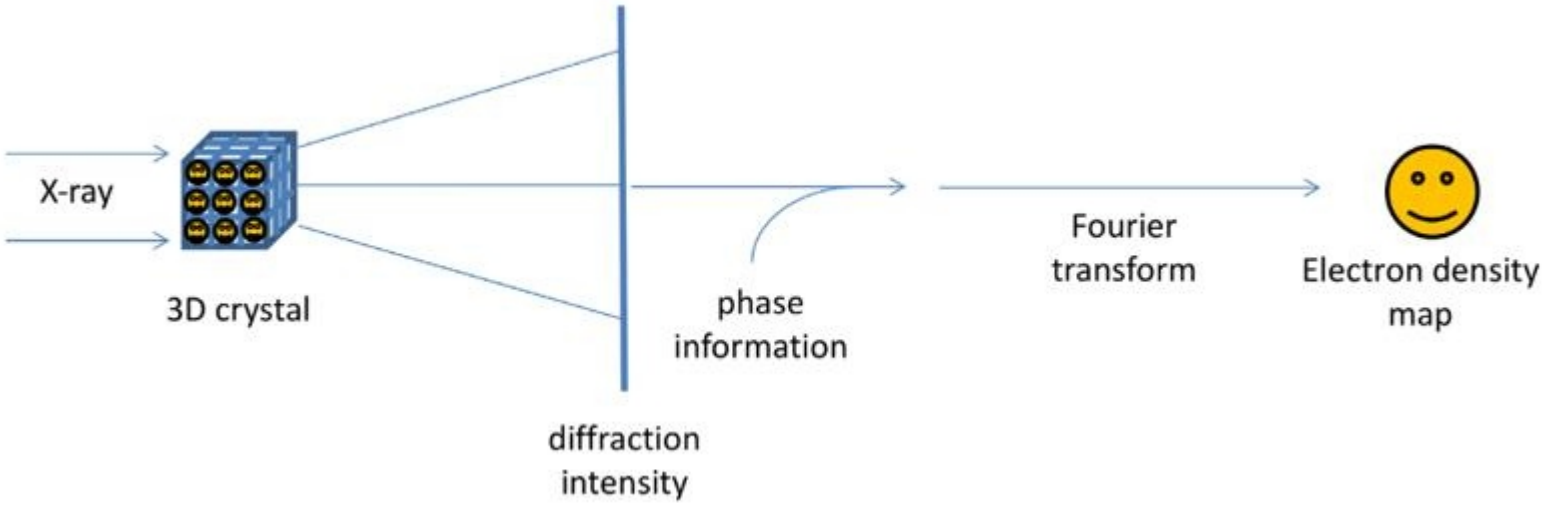
Most important bioinformatics databases

RCSB PDB
PROTEIN DATA BANK



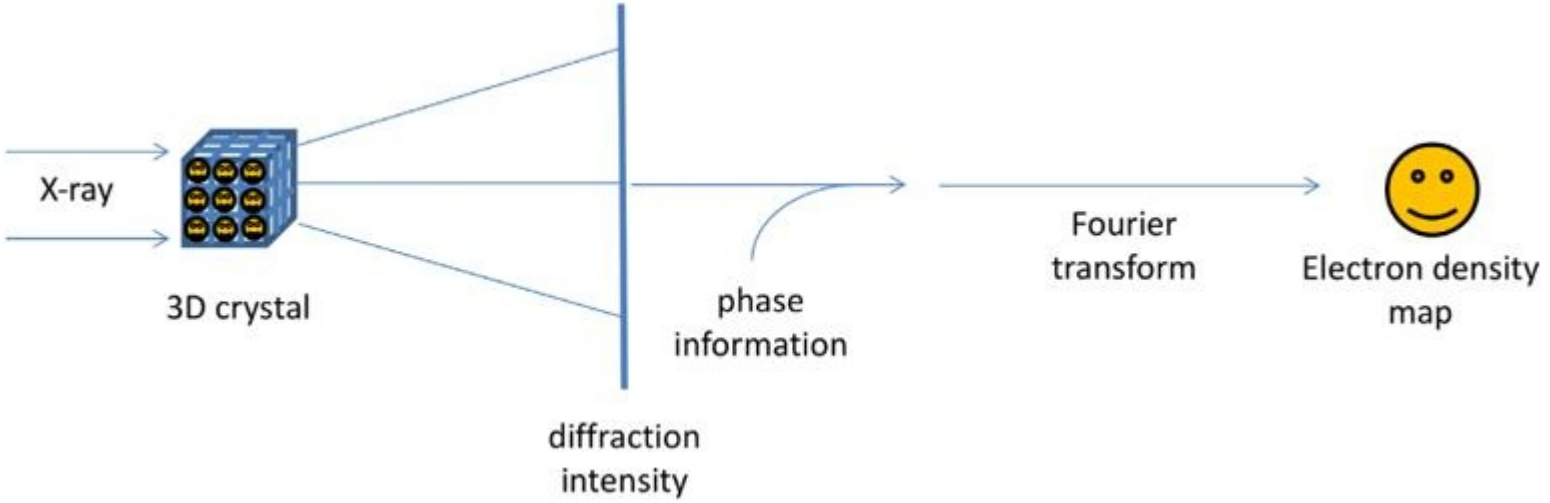
X-ray diffraction

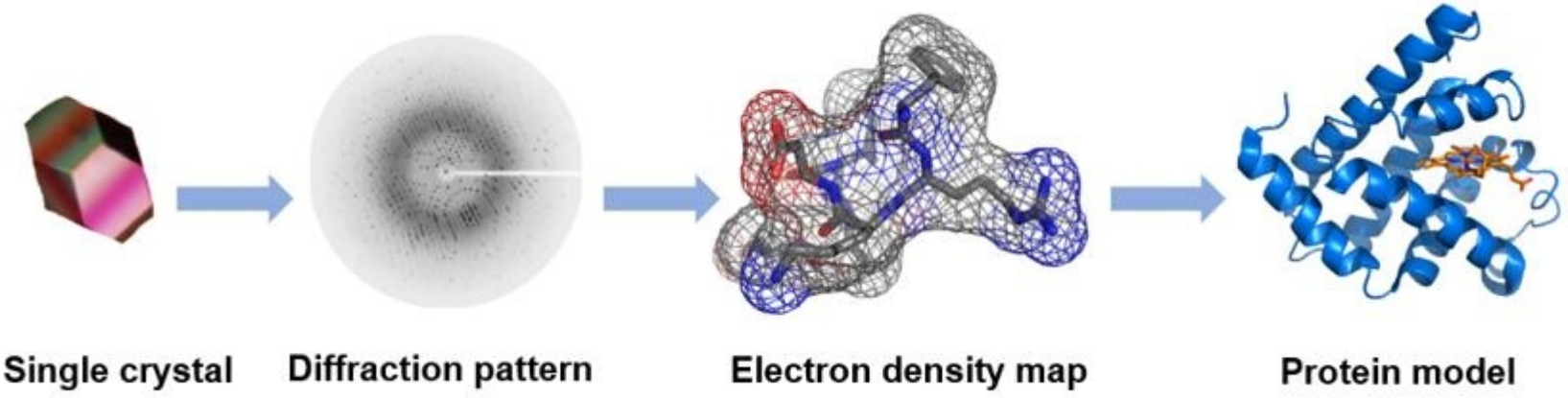
ADP



https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

X-ray diffraction

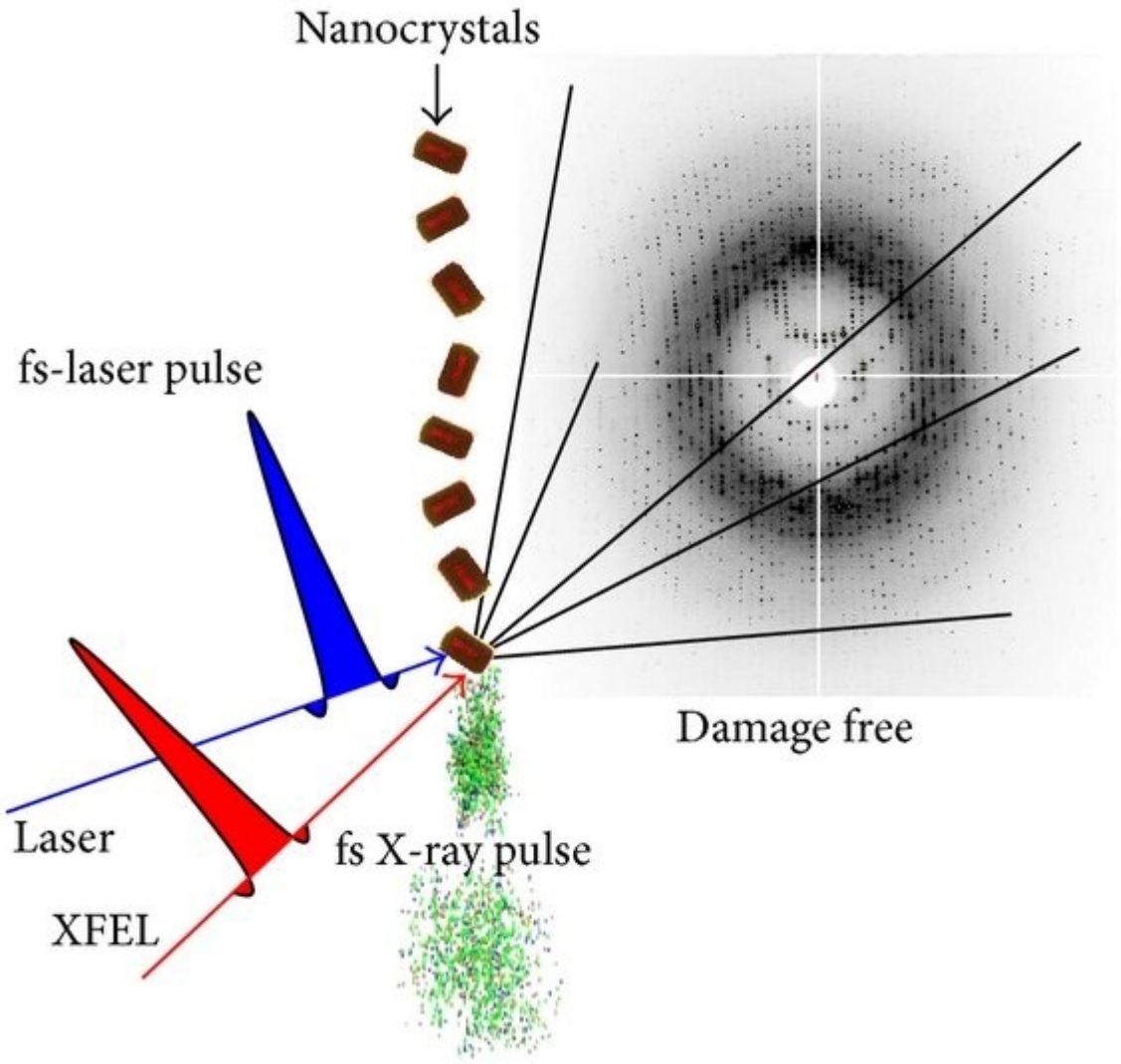




https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

XFEL - X-ray free electron lasers

ADP

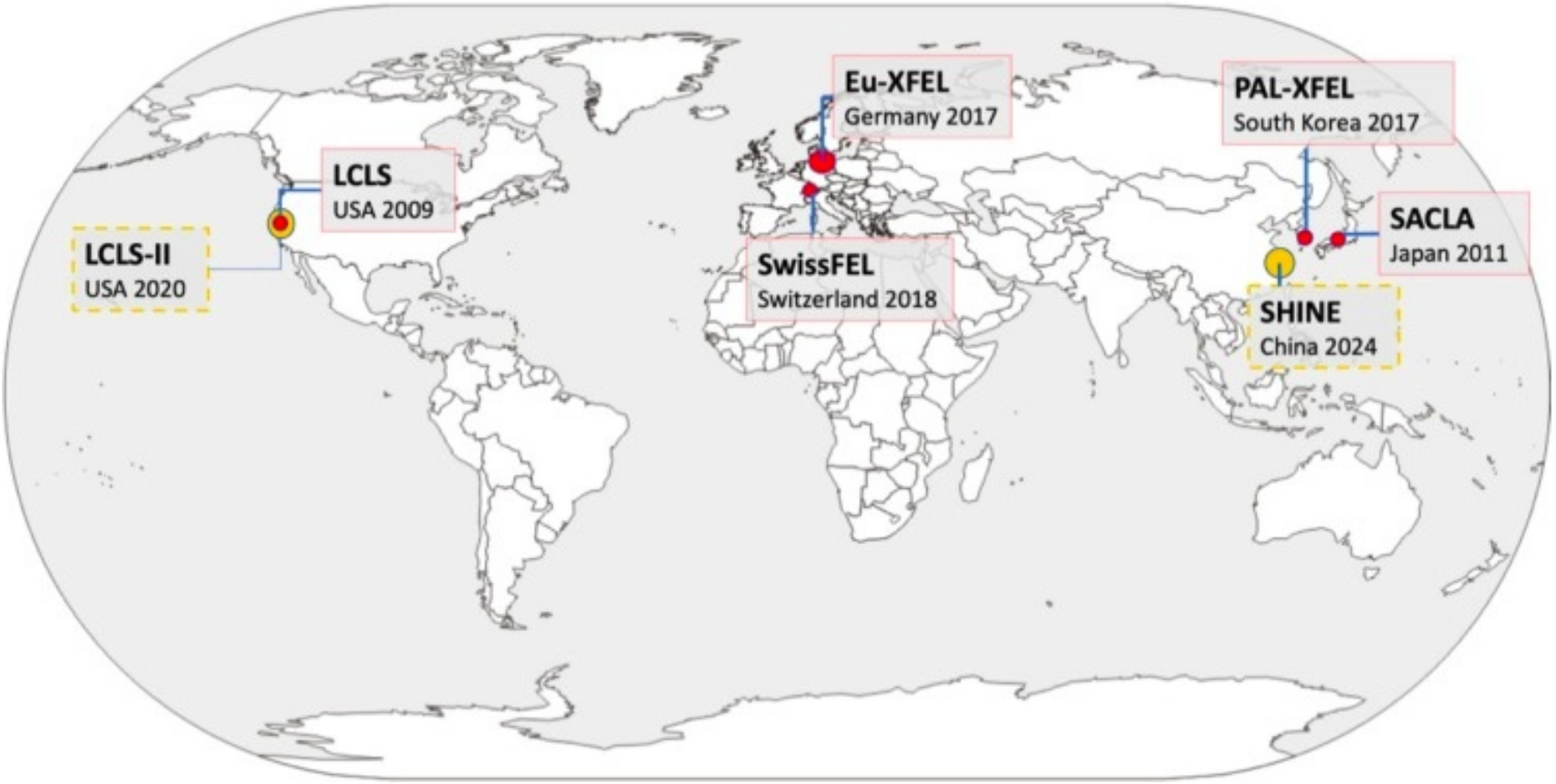


For more watch: https://www.youtube.com/watch?v=-VMDytbTbNw

https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

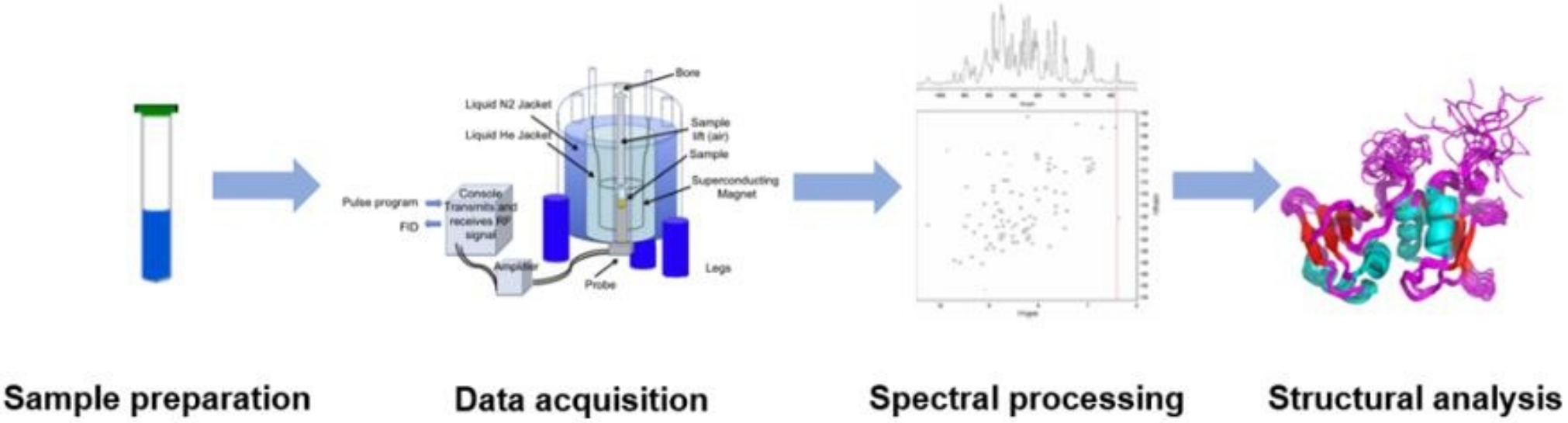
XFEL - X-ray free electron lasers

ADP



Nuclear Magnetic Resonance (NMR)

ADP

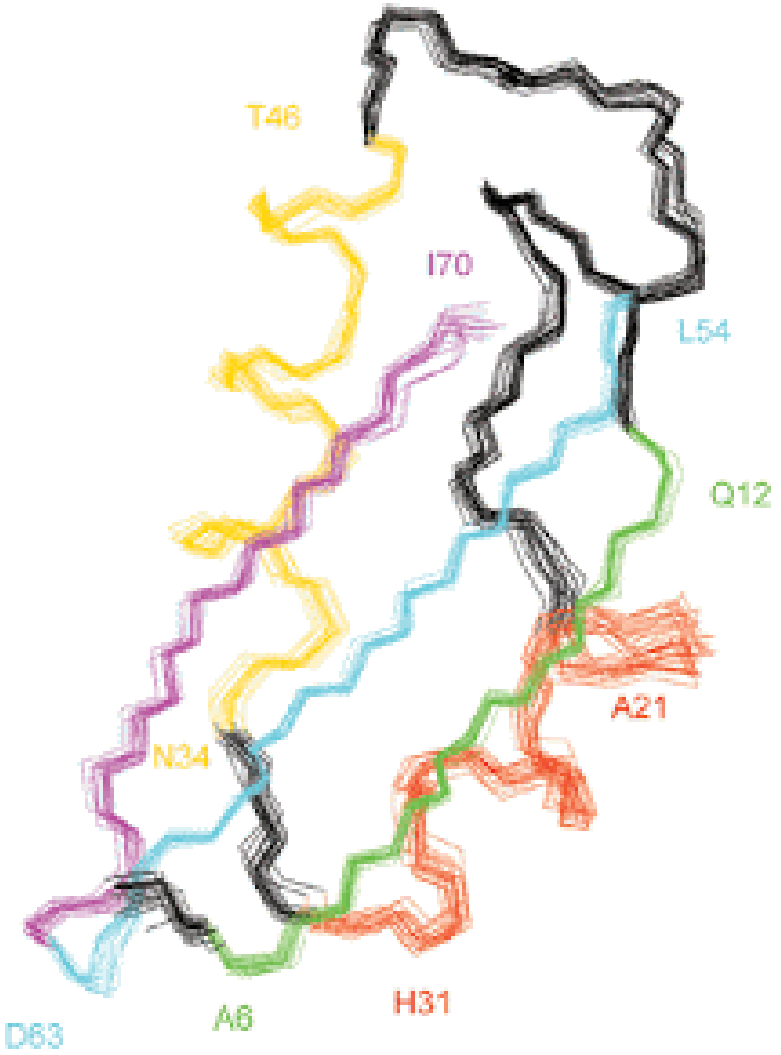


Sample preparation | Data acquisition | Spectral processing | Structural analysis

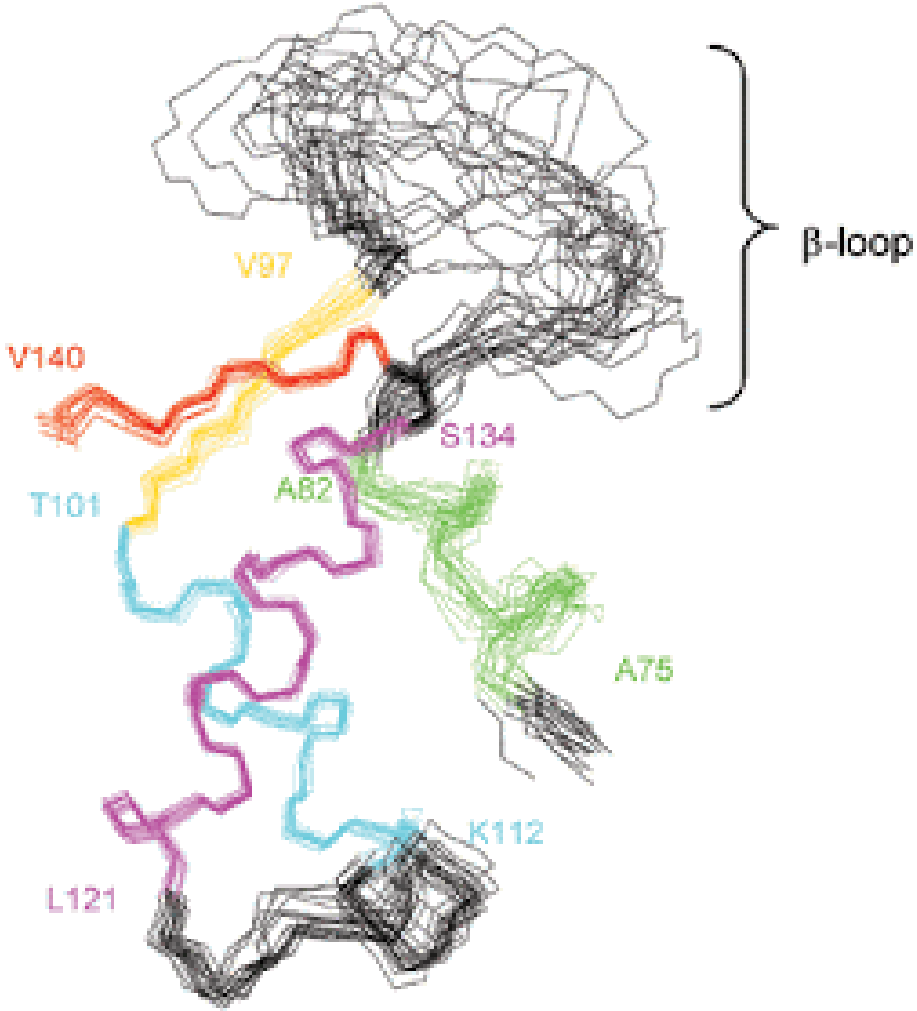
https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

Nuclear Magnetic Resonance (NMR)

ADP

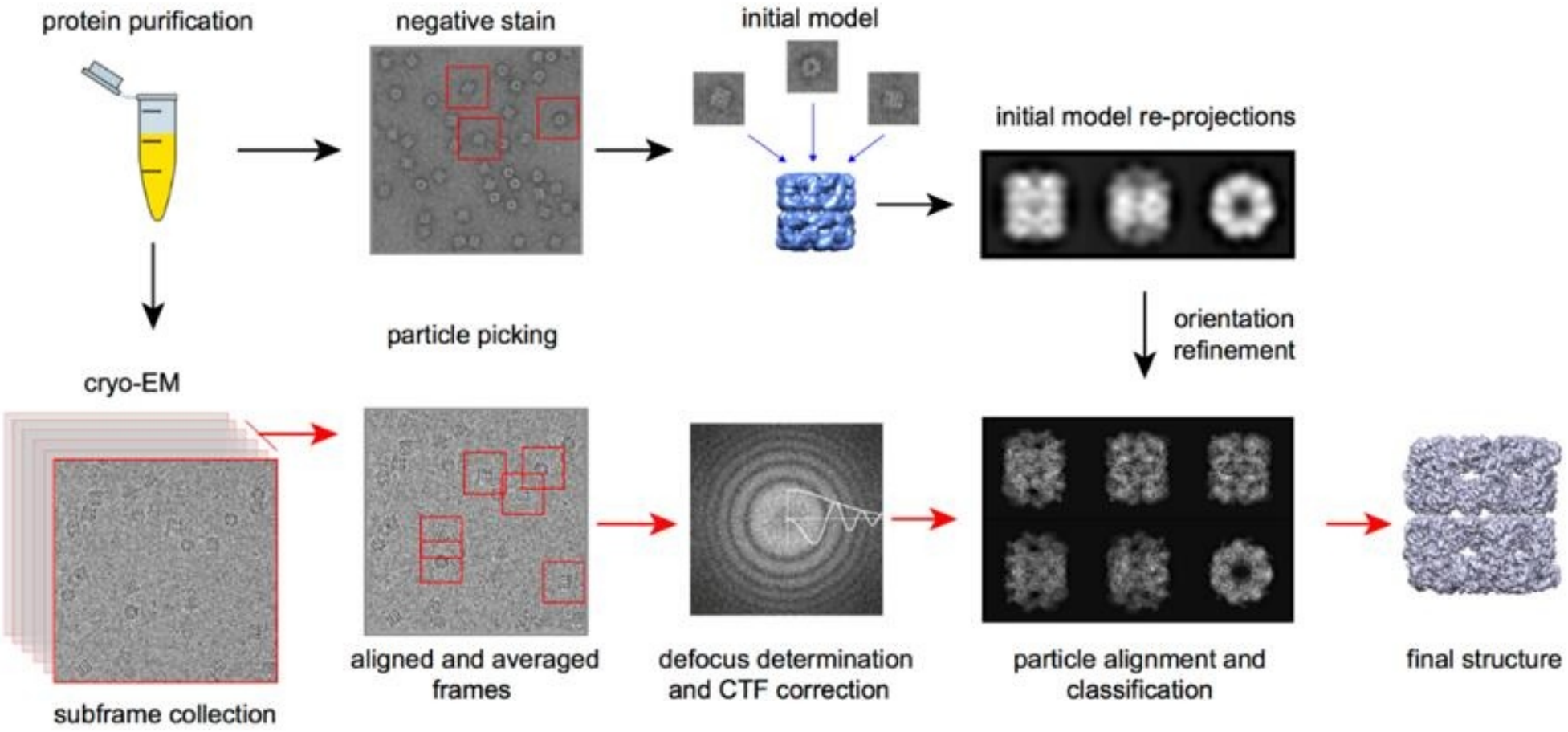


L11 N-Terminus (5-70)



L11 C-Terminus (75-140)

http://schwalbe.org.chemie.uni-frankfurt.de/node/684

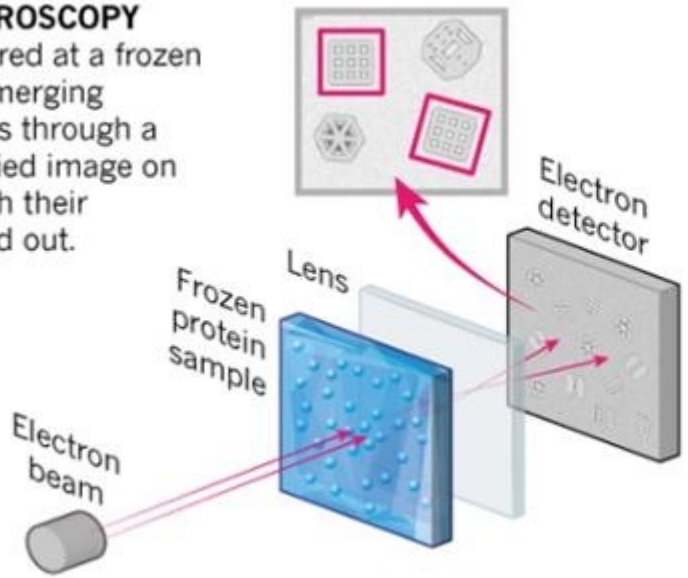


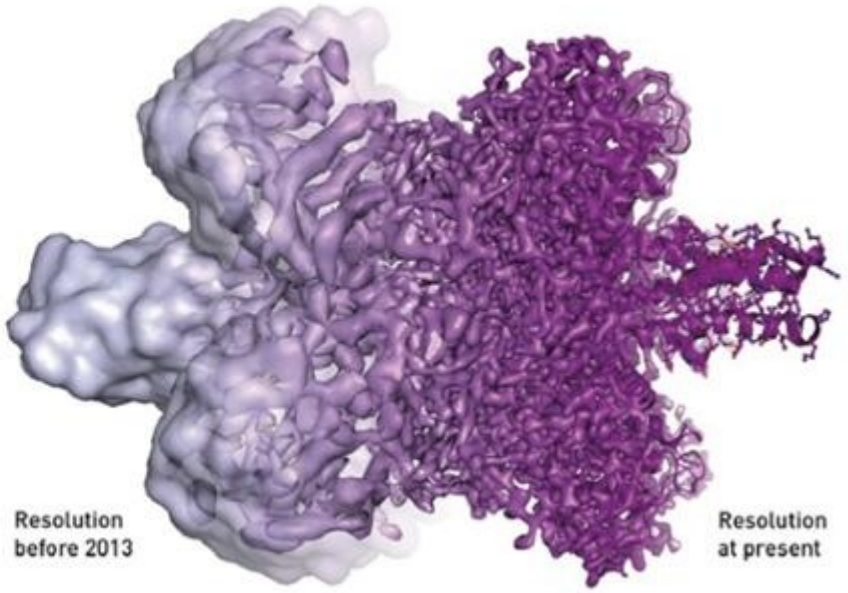
https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

Cryo-EM

ADP

CRYO-ELECTRON MICROSCOPY
A beam of electron is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, from which their structure can be worked out.



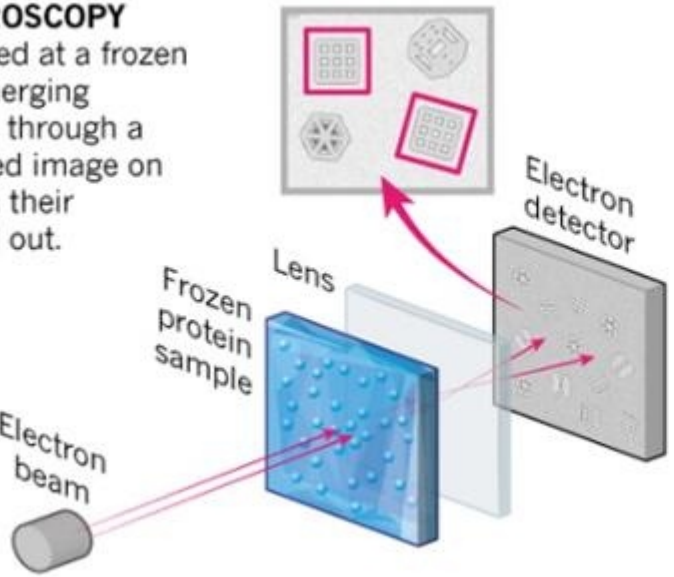


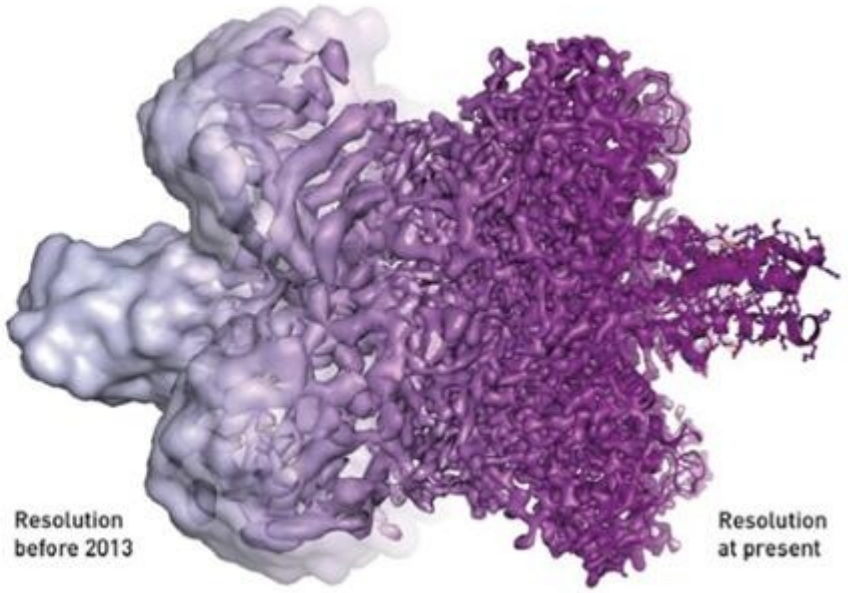
https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

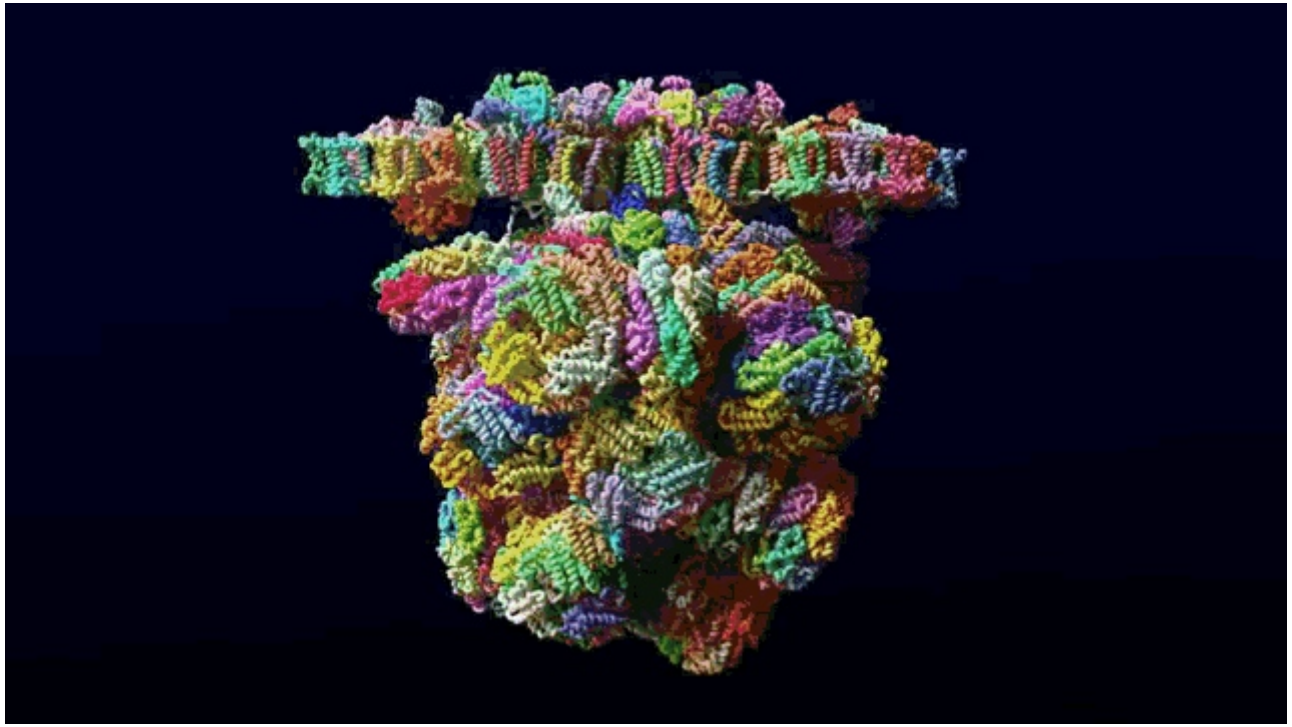
Cryo-EM

ADP

CRYO-ELECTRON MICROSCOPY
A beam of electron is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, from which their structure can be worked out.







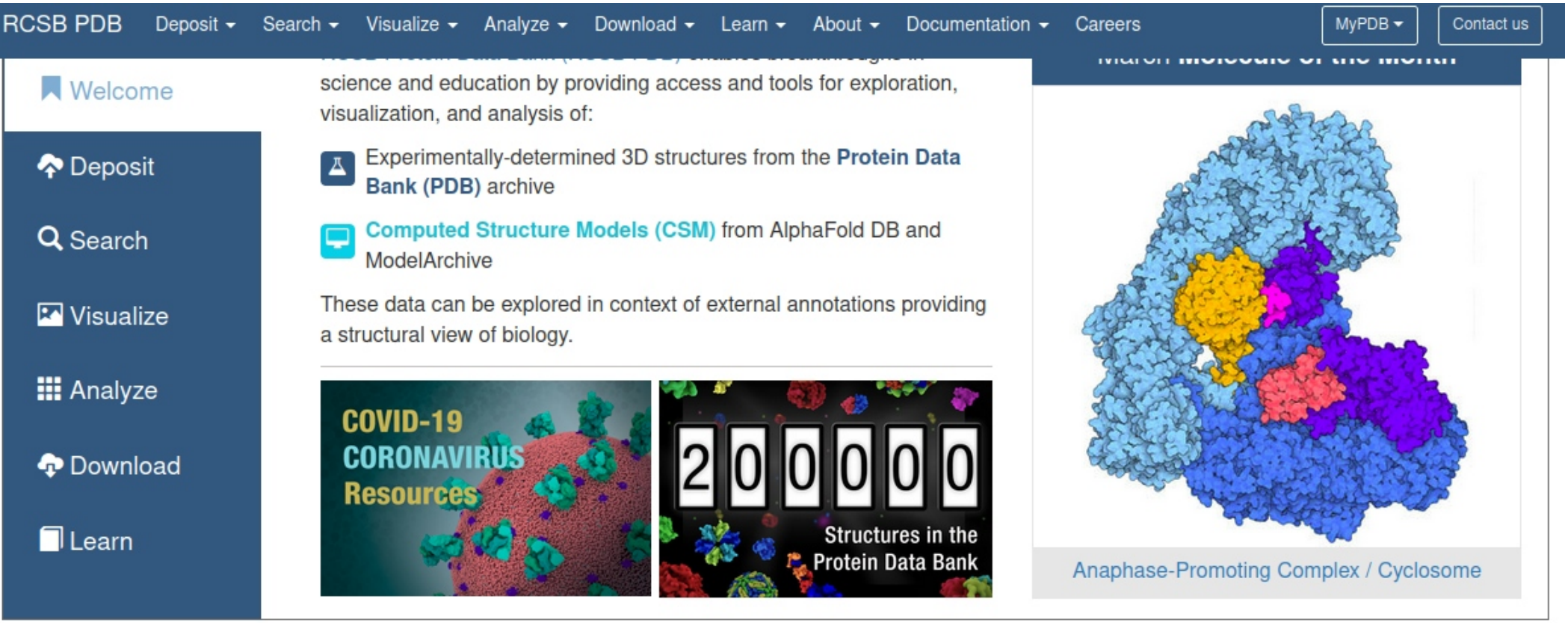
https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

Most important bioinformatics databases

ADP

RCSB PDB
PROTEIN DATA BANK

https://www.rcsb.org/



Most important bioinformatics databases

ALPHA FOLD

PROTEIN STRUCTURE

DATABASE

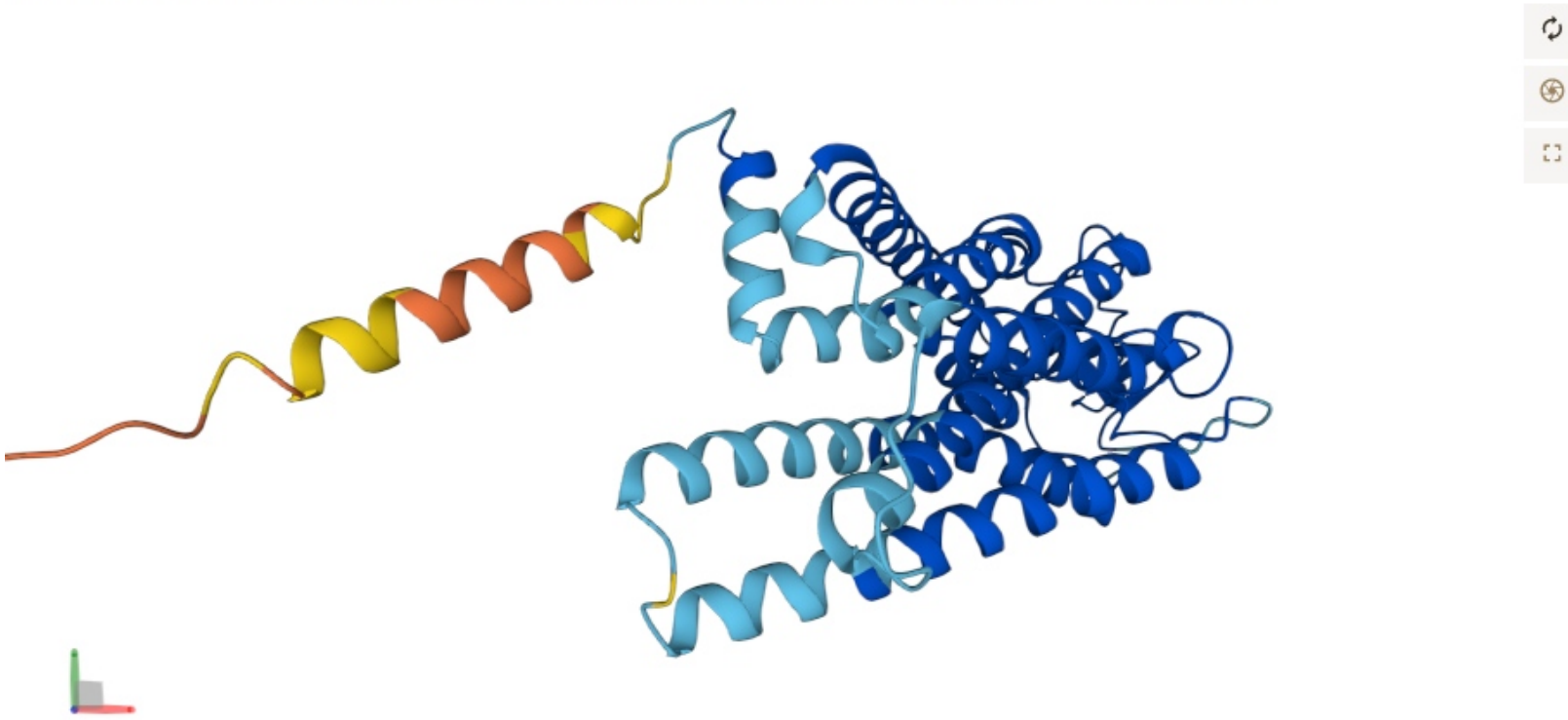
DeepMind

214M

25TB

Sequence of AF-P35359-F1 Chain 1: Rhodopsin A

```
MNGTEGPAFYVPMSNATGVVRSPYEYPQYYLVAPWAYGLLAAYMFFLIITGFPVNFLTLYVTIEHKKLRTPLNYILLNLAIADLFMVFGGFTTTMYTSLHGYFVFGRLGCNLEGFFATLGGEMGLW
SLVVLAIERWMVVCKPVSNFRFGENHAIMGVAFTWVMACSCAVPPLVGWSRYIPEGMQCSCGVDYYTRTPGVNNESFVIYMFIVHFFIPLIVIFFCYGRLVCTVKEAAAQQQESETTQRAEREVTR
MVIIMVIAFLICWLPYAGVAWYIFTHQGSEFGPVFMTLPAFFAKTSAVYNPCIYICMNKQFRHCMITTLCCGKNPFEEEEGASTTASKTEASSVSSSSVSPA
```



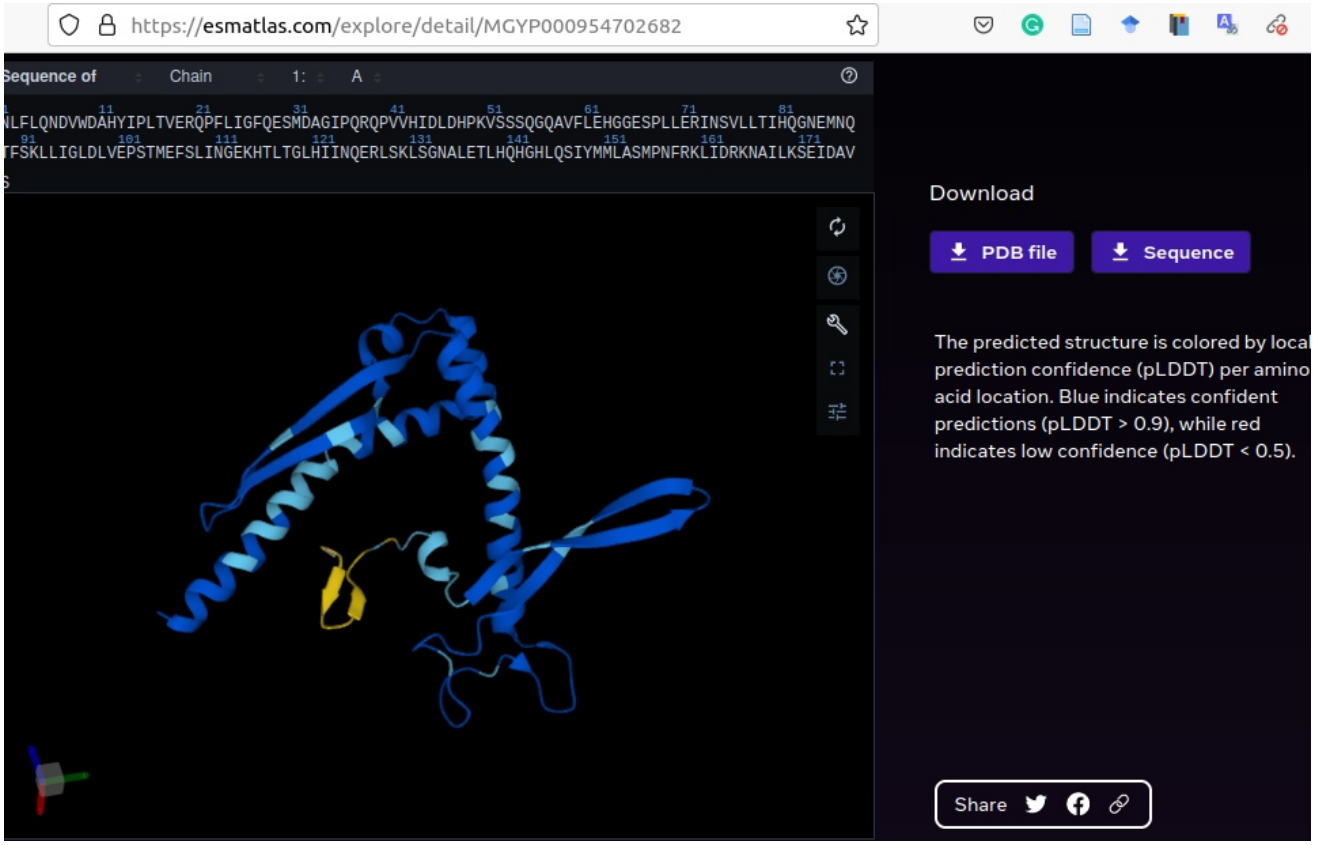
https://alphafold.ebi.ac.uk/entry/P35359

Most important bioinformatics databases

ESM Atlas
Metagenomics
Protein Dataset

Meta AI

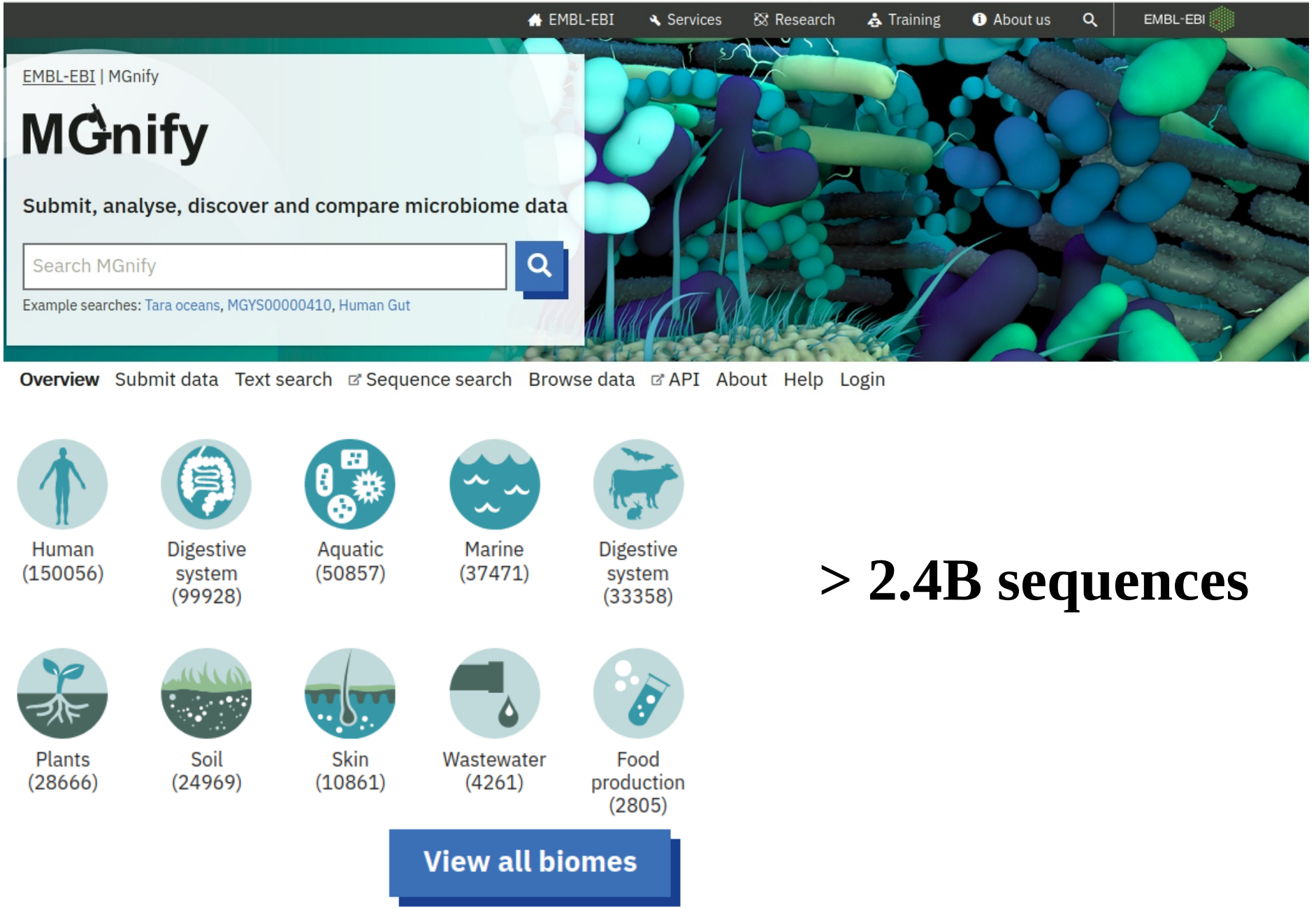
617M sequences & 15TB data



https://esmatlas.com/

Most important bioinformatics databases

ADP

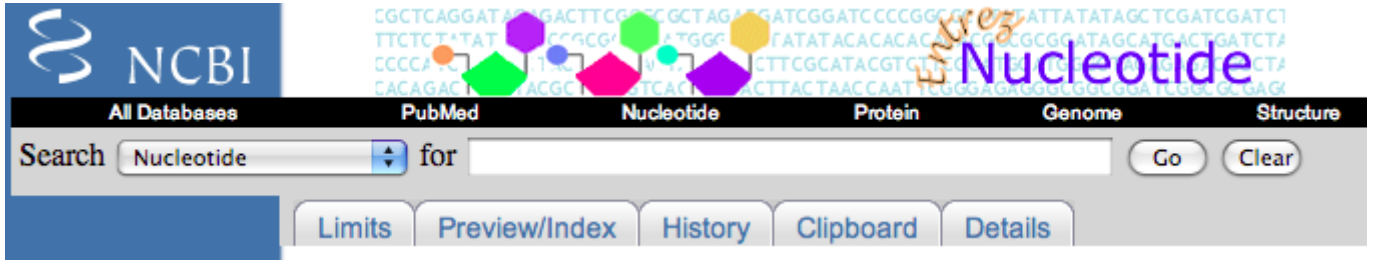


> 2.4B sequences

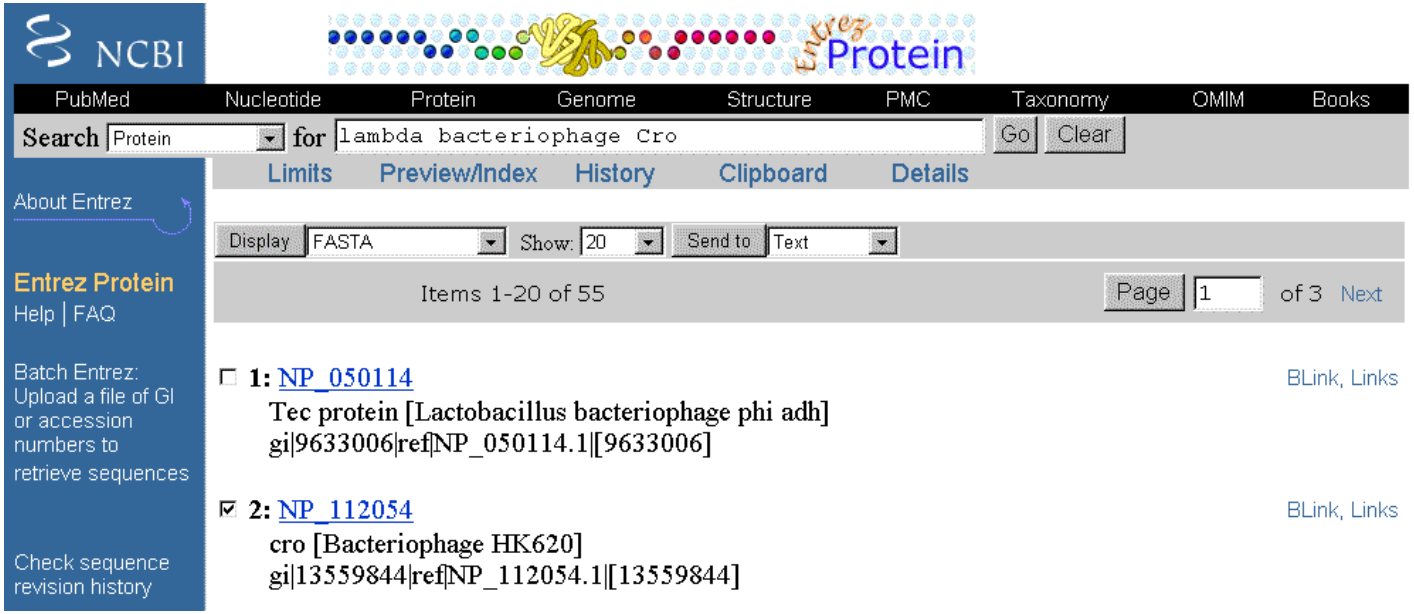
https://www.ebi.ac.uk/metagenomics/

Most important bioinformatics databases

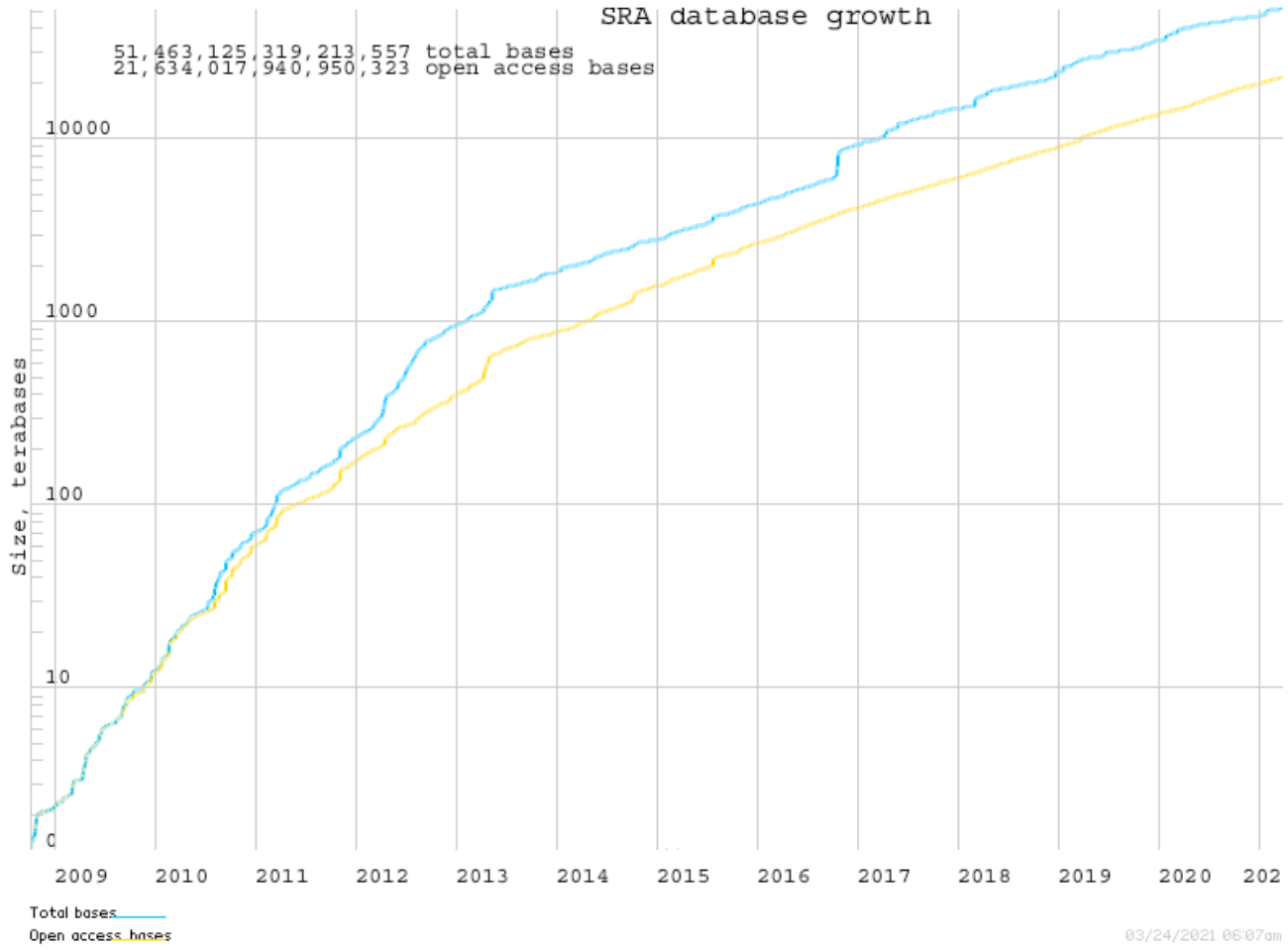
GenBank



GenPept



Sequence Read Archive (SRA)



Sequence Read Archive (SRA)

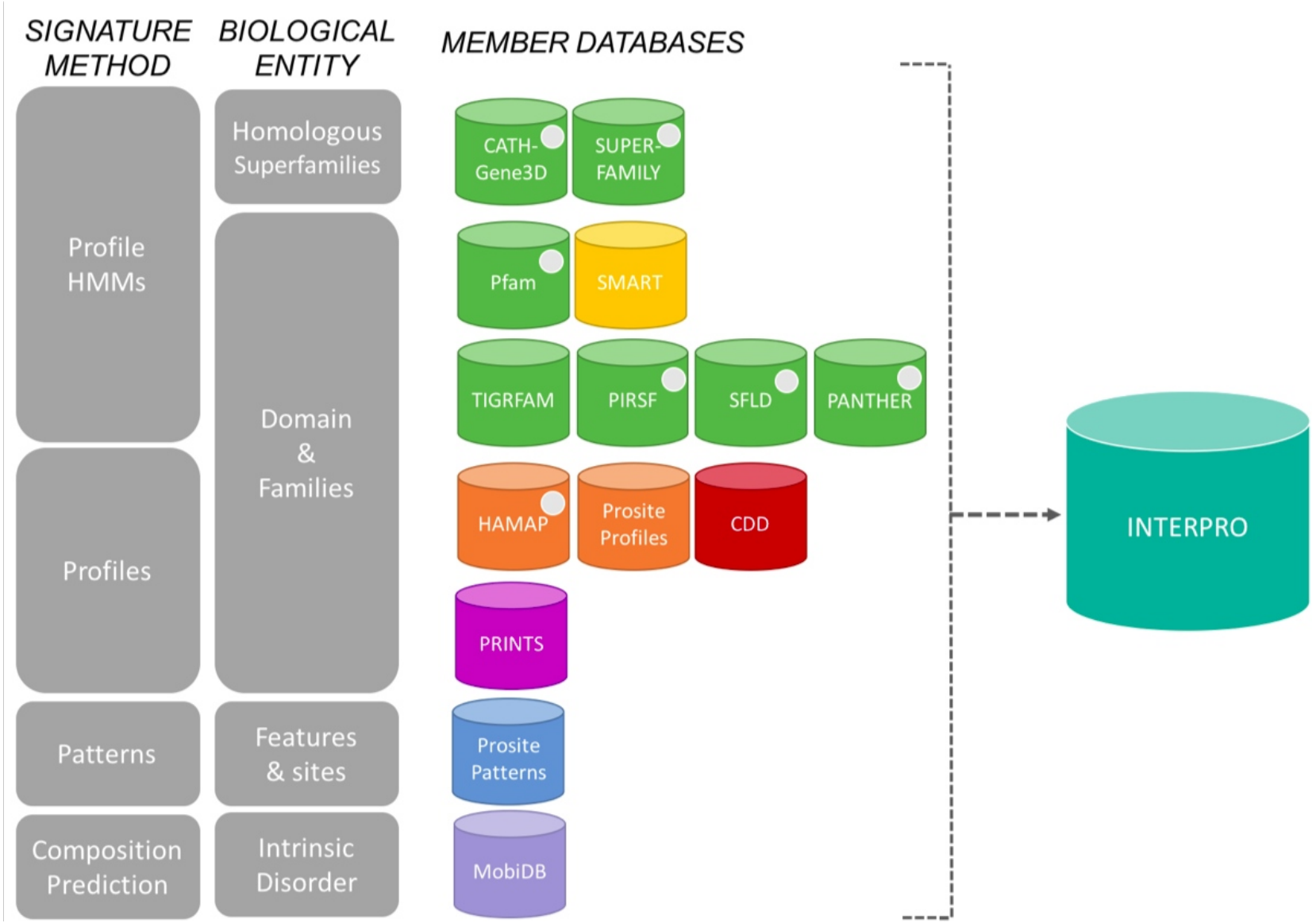
Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores

Sequence Read Archive (SRA)

Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores

SRA = NGS data

InterPro

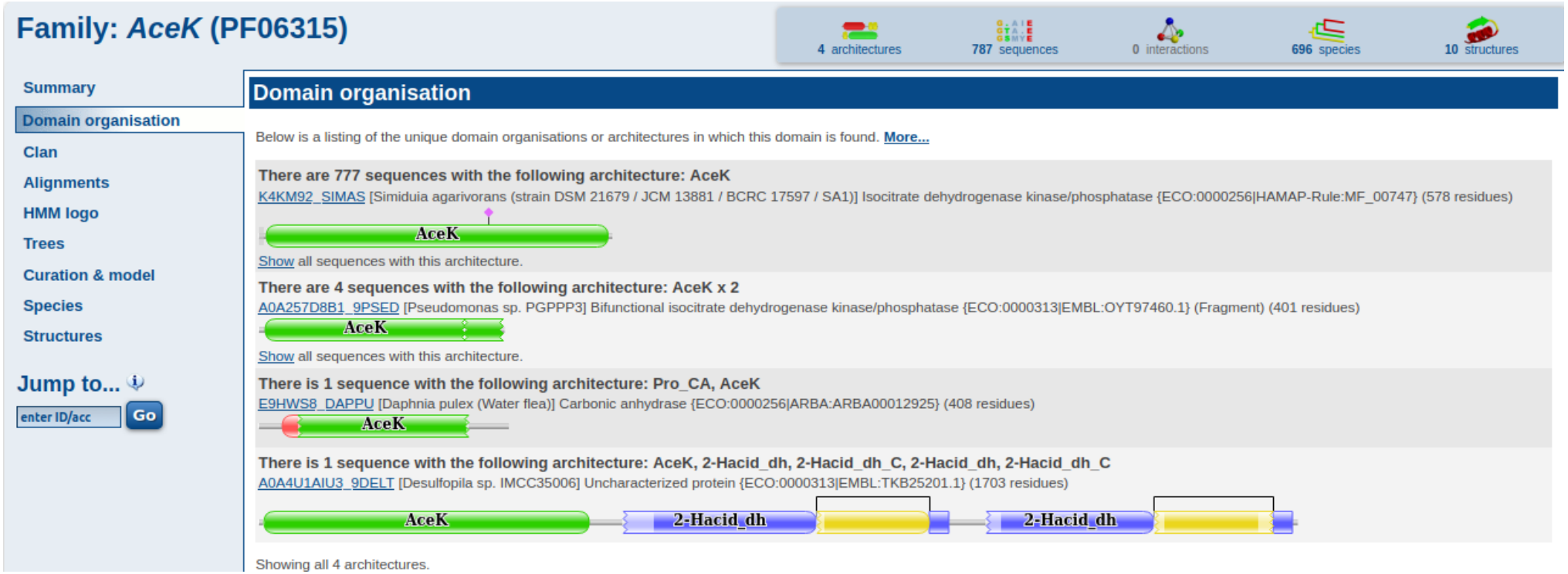

SIGNATURE METHOD
BIOLOGICAL ENTITY
MEMBER DATABASES
Profile HMMs
Profiles
Patterns
Composition Prediction
Homologous Superfamilies
Domain & Families
Features & sites
Intrinsic Disorder
CATH-Gene3D
SUPER-FAMILY
Pfam
SMART
TIGRFAM
PIRSF
SFLD
PANTHER
HAMAP
Prosite Profiles
CDD
PRINTS
Prosite Patterns
MobiDB
INTERPRO

Most important bioinformatics databases

ADP

Pfam

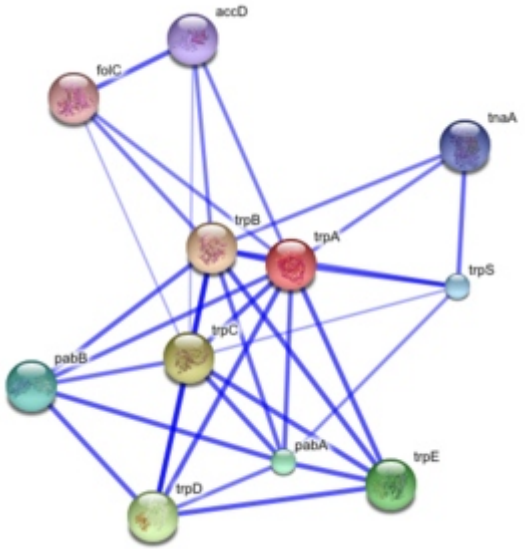
Rfam



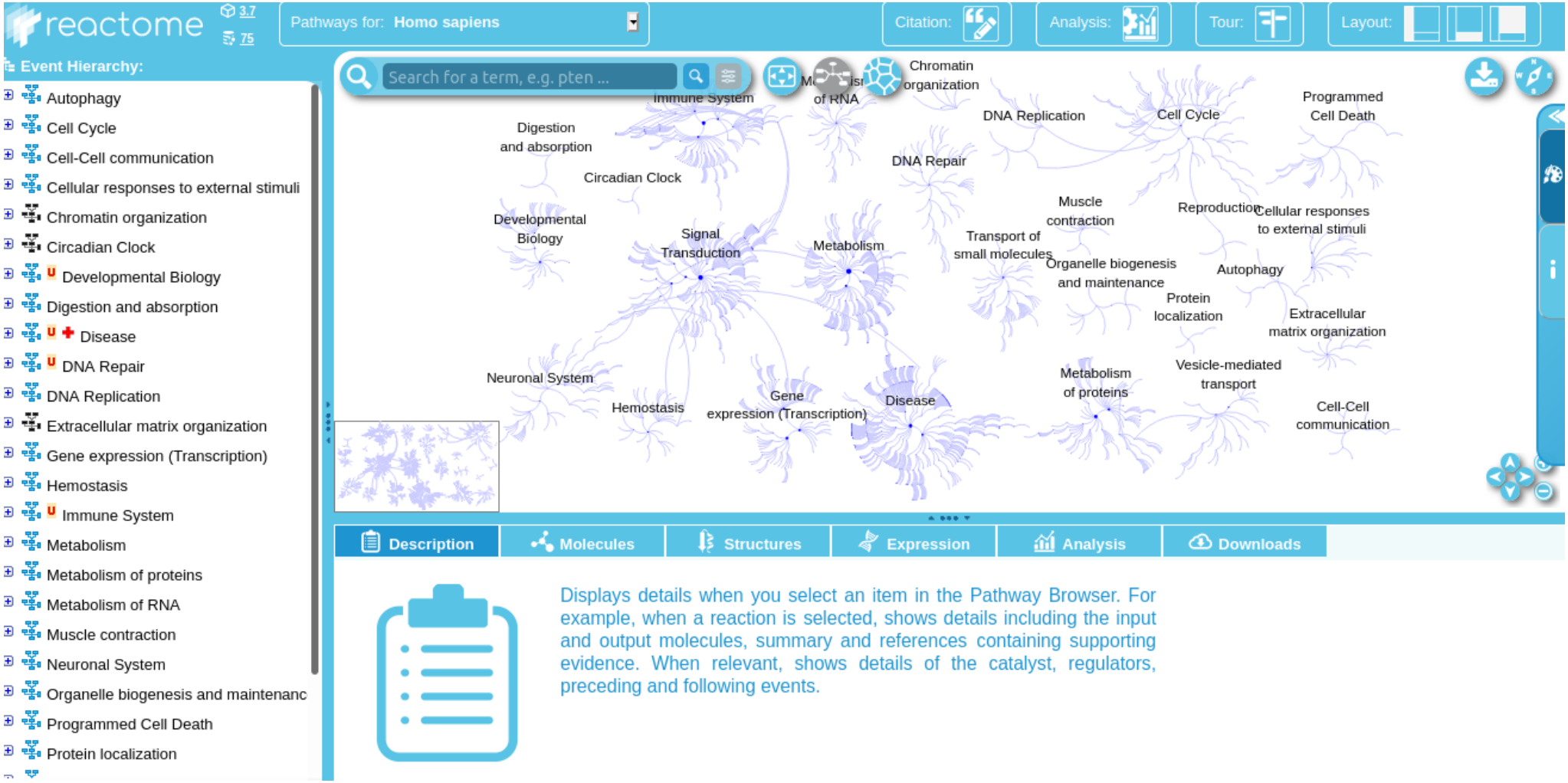
Most important bioinformatics databases

ADP

STRING



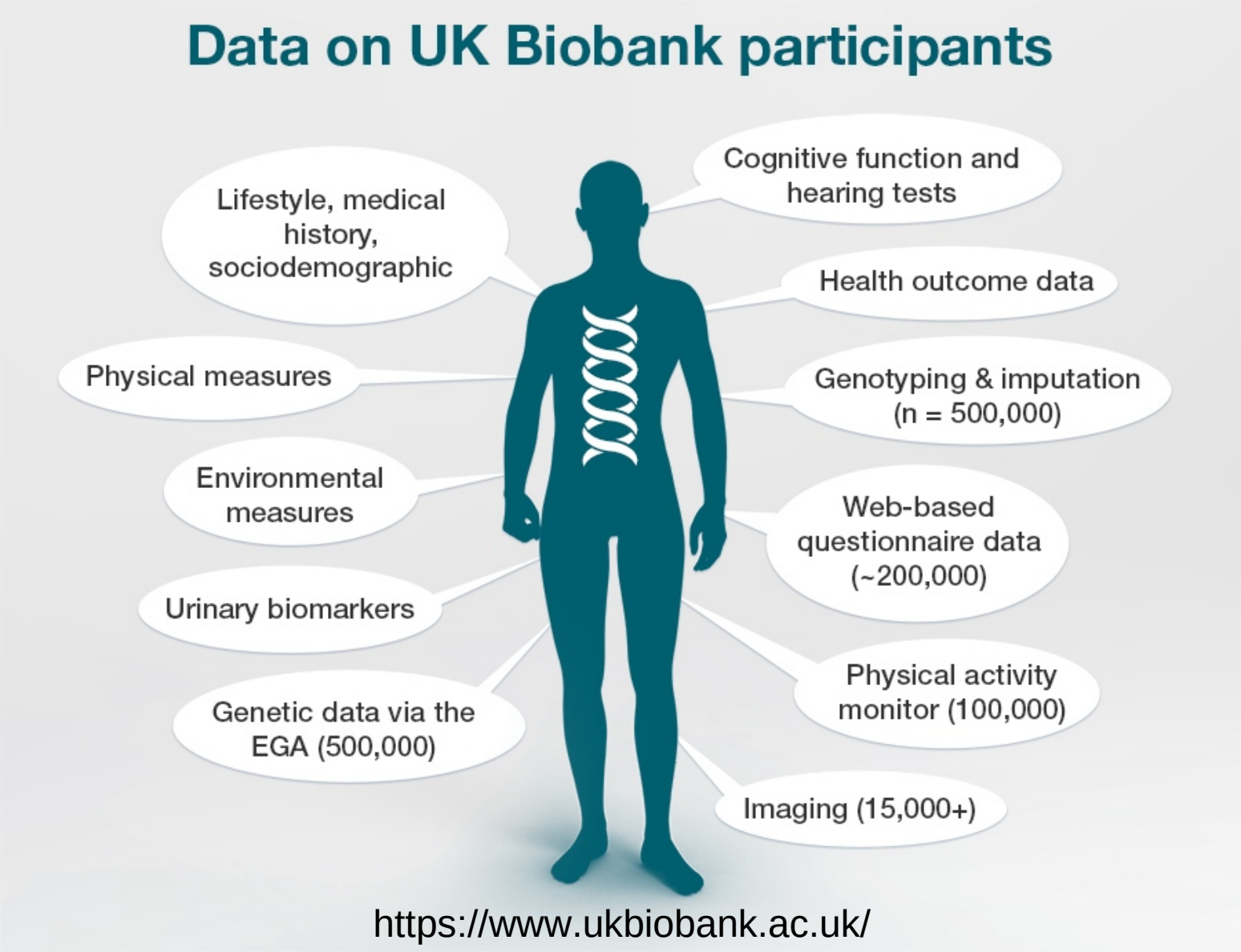
Reactome – biological pathways



Most important bioinformatics databases

biobank uk

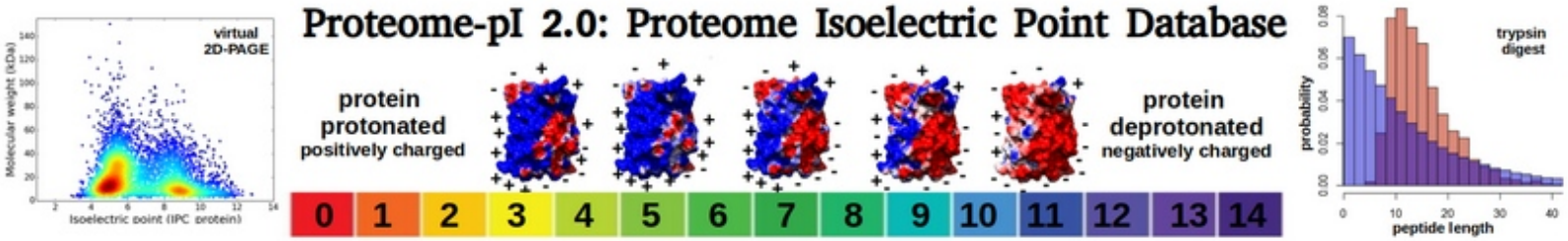
Enabling scientific discoveries that improve human health



https://www.ukbiobank.ac.uk/

Proteome-pI 2.0

NARODOWE CENTRUM NAUKI
Opus



Home | Browse | Search | Statistics | Download | About | Contact

Database of **pre-computed** isoelectric points and molecular weights for proteins and digest peptides from model organism proteomes (20,115 species)

The goals of the database include making statistical comparisons of the various prediction methods (21 algorithms implemented) as well as facilitating the biological investigation of protein isoelectric point space. The isoelectric point, the pH at which a particular molecule carries no net electrical charge, is an important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analysis

Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

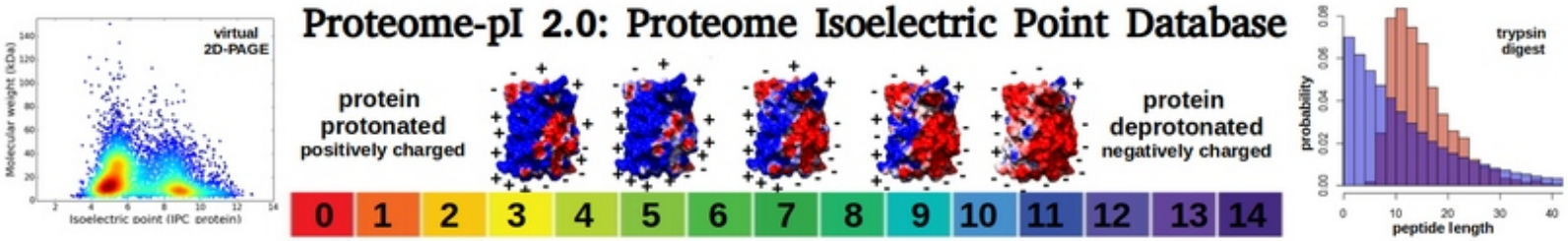
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0

NARODOWE CENTRUM NAUKI
Opus



Home | Browse | Search | Statistics | Download | About | Contact

Database of **pre-computed** isoelectric points and molecular weights for proteins and digest peptides from model organism proteomes (20,115 species)

The goals of the database include making statistical comparisons of the various prediction methods (21 algorithms implemented) as well as facilitating the biological investigation of protein isoelectric point space. The isoelectric point, the pH at which a particular molecule carries no net electrical charge, is an important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analysis

Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

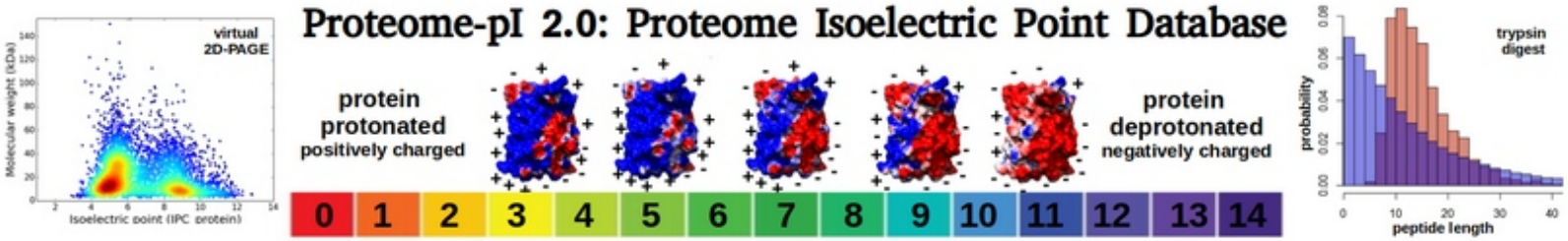
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0

NARODOWE CENTRUM NAUKI
Opus



Home | Browse | Search | Statistics | Download | About | Contact

Database of **pre-computed** isoelectric points and molecular weights for proteins and digest peptides from model organism proteomes (20,115 species)

The goals of the database include making statistical comparisons of the various prediction methods (21 algorithms implemented) as well as facilitating the biological investigation of protein isoelectric point space. The isoelectric point, the pH at which a particular molecule carries no net electrical charge, is an important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analysis

Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

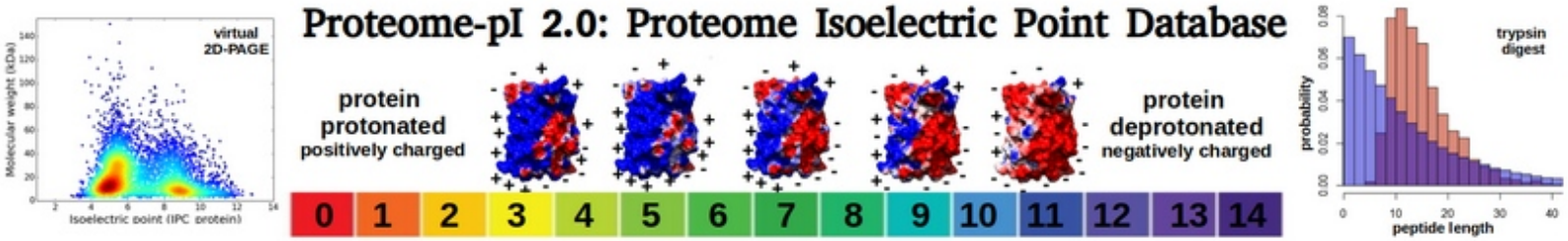
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0

NARODOWE CENTRUM NAUKI
Opus



Home | Browse | Search | Statistics | Download | About | Contact

Database of **pre-computed** isoelectric points and molecular weights for proteins and digest peptides from model organism proteomes (20,115 species)

The goals of the database include making statistical comparisons of the various prediction methods (21 algorithms implemented) as well as facilitating the biological investigation of protein isoelectric point space. The isoelectric point, the pH at which a particular molecule carries no net electrical charge, is an important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analysis

Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

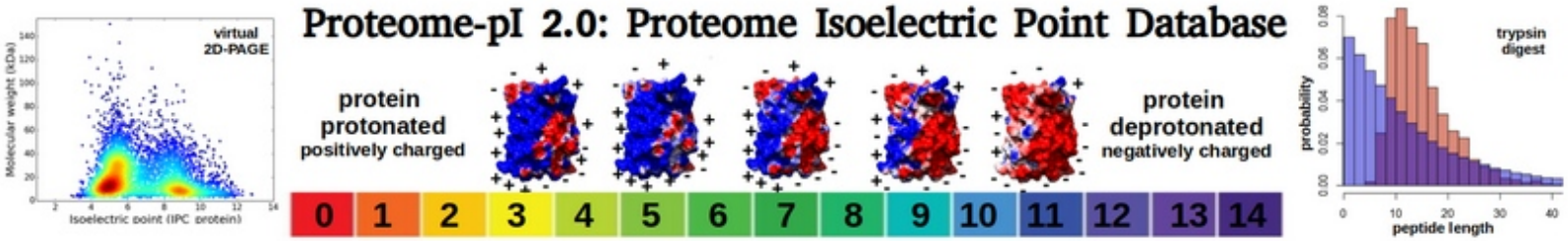
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0

NARODOWE CENTRUM NAUKI
Opus



Home | Browse | Search | Statistics | Download | About | Contact

Database of **pre-computed** isoelectric points and molecular weights for proteins and digest peptides from model organism proteomes (20,115 species)

The goals of the database include making statistical comparisons of the various prediction methods (21 algorithms implemented) as well as facilitating the biological investigation of protein isoelectric point space. The isoelectric point, the pH at which a particular molecule carries no net electrical charge, is an important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analysis

Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

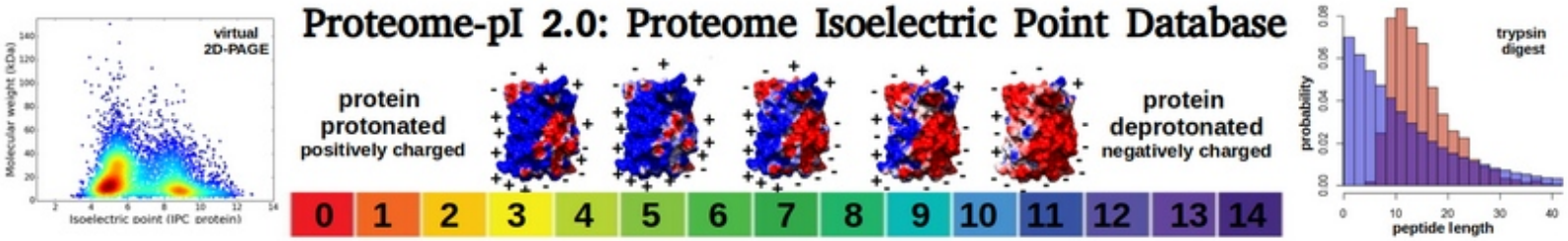
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0

NARODOWE CENTRUM NAUKI
Opus



Home | Browse | Search | Statistics | Download | About | Contact

Database of **pre-computed** isoelectric points and molecular weights for proteins and digest peptides from model organism proteomes (20,115 species)

The goals of the database include making statistical comparisons of the various prediction methods (21 algorithms implemented) as well as facilitating the biological investigation of protein isoelectric point space. The isoelectric point, the pH at which a particular molecule carries no net electrical charge, is an important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analysis

Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

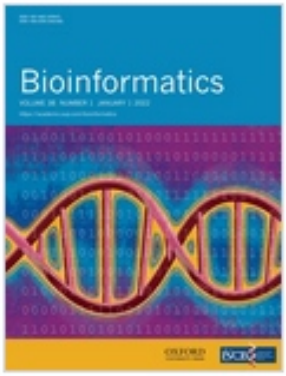
61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540



Volume 38, Issue 1

pKPDB: a protein data bank extension database of pK_a and pI theoretical values

Get access >

Pedro B P S Reis, Djork-Arné Clevert, Miguel Machuqueiro ✉

Bioinformatics, Volume 38, Issue 1, 1 January 2022, Pages 297–298, https://doi.org/10.1093/bioinformatics/btab518

Published: 14 July 2021 **Article history** ▾

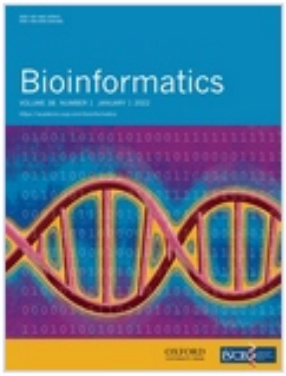
61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

5.38 Billion dissociation constant (pKa) predictions for proteins

Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540



Volume 38, Issue 1

pKPDB: a protein data bank extension database of pK_a and pI theoretical values

Get access >

Pedro B P S Reis, Djork-Arné Clevert, Miguel Machuqueiro ✉

Bioinformatics, Volume 38, Issue 1, 1 January 2022, Pages 297–298, https://doi.org/10.1093/bioinformatics/btab518

Published: 14 July 2021 **Article history** ▾

pKPDB is a database of over 12 M theoretical pKa values calculated over 120k protein structures deposited in the Protein Data Bank

PypKa (structure based method) with ~0.9 RMSD accuracy

61,329,034 protein sequences from **20,115** proteomes with isoelectric point predicted using **21** algorithms

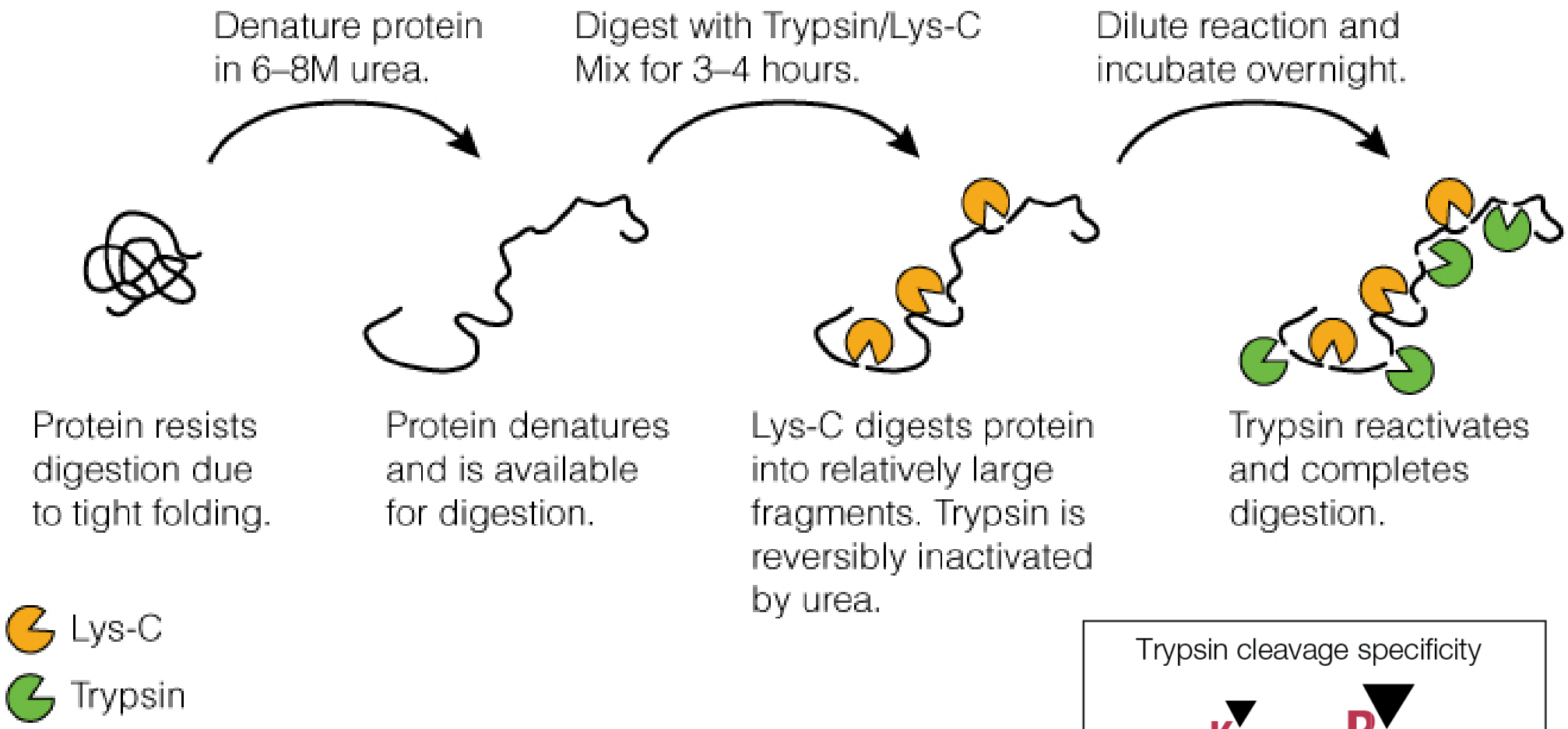
5.38 Billion dissociation constant (pKa) predictions for proteins

Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0

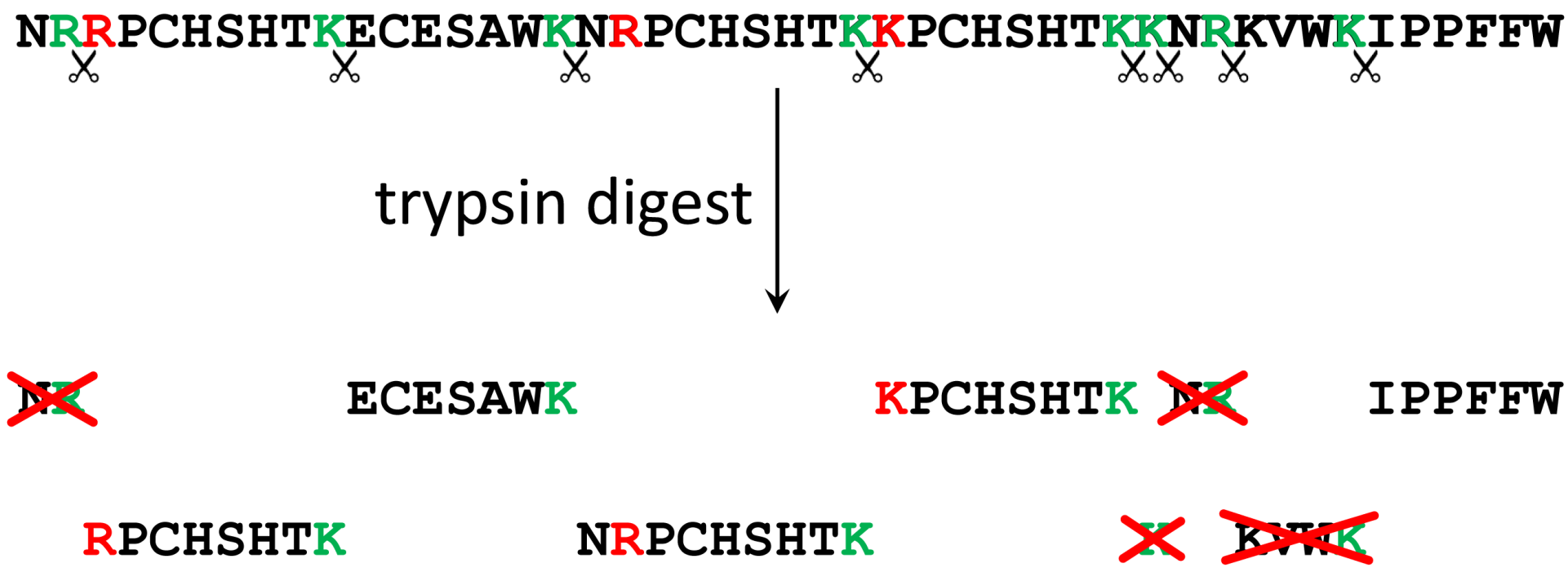


Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

Proteome-pI 2.0



https://doi.org/10.1021/acs.jproteome.8b00716

pl.promega.com

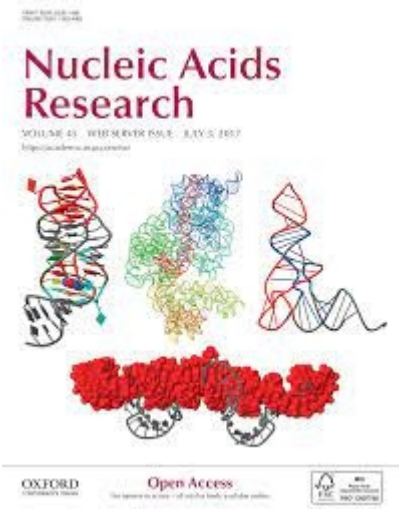
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

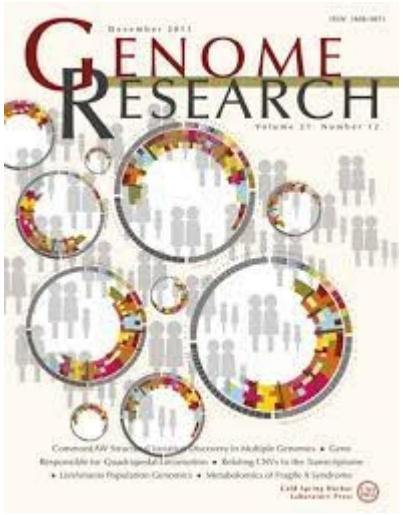
isoelectricpointdb2.mimuw.edu.pl

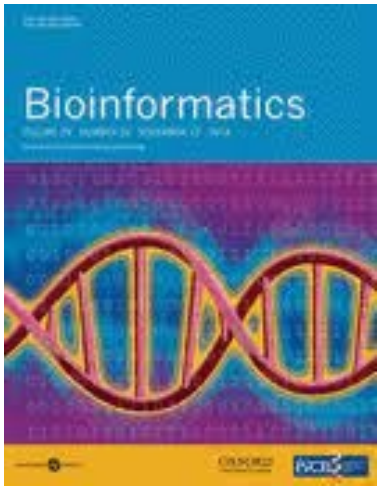
Kozlowski LP. Proteome-pI 2.0: Proteome Isoelectric Point Database Update. NAR 2022 50 (D1), D1535-D1540

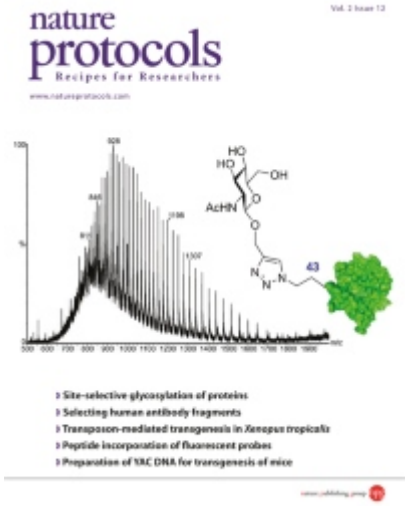
Journals related to bioinformatics resources

ADP







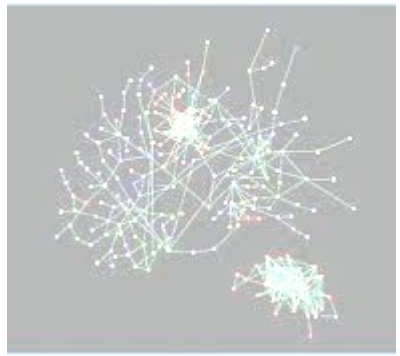


BMC Bioinformatics

Google Scholar

PubMed

Journals related to bioinformatics resources



BMC Bioinformatics

Google

PubMed

BioGRID

http://bio.tools ← tools

https://academic.oup.com/nar/pages/nar_methods_new

https://fairsharing.org ← databases

https://www.oxfordjournals.org/nar/database/c/

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

lukaskoz@mimuw.edu.pl