

Architecture of large projects in bioinformatics (ADP)

Lecture 02

Łukasz P. Kozłowski

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**

Perl → *BioPerl*

Perl → *BioPerl*

Php → *BioPHP*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Rust → *Rust-Bio*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Rust → *Rust-Bio*

C++ → *Bio++*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Rust → *Rust-Bio*

C++ → *Bio++*

Julia → *BioJulia*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Rust → *Rust-Bio*

C++ → *Bio++*

Julia → *BioJulia*

JavaScript → *BioJS*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Rust → *Rust-Bio*

C++ → *Bio++*

Julia → *BioJulia*

JavaScript → *BioJS*

Python → *BioPython*

Perl → *BioPerl*

Php → *BioPHP*

Java → *BioJava*

R → *Bioconductor*

Rust → *Rust-Bio*

C++ → *Bio++*

Julia → *BioJulia*

JavaScript → *BioJS*

Python → *BioPython*, but also *Cogent3*, *bioconda*

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

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



Python → BioPython, but also Cogent3, bioconda

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

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





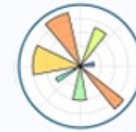
pandas
pandas



SciPy



NumPy



Matplotlib

django



Keras



Tensor Flow



PyTorch

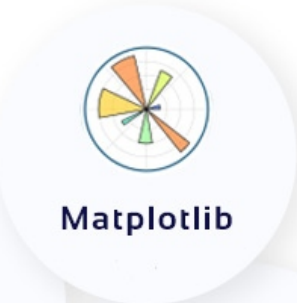


Scikit-learn

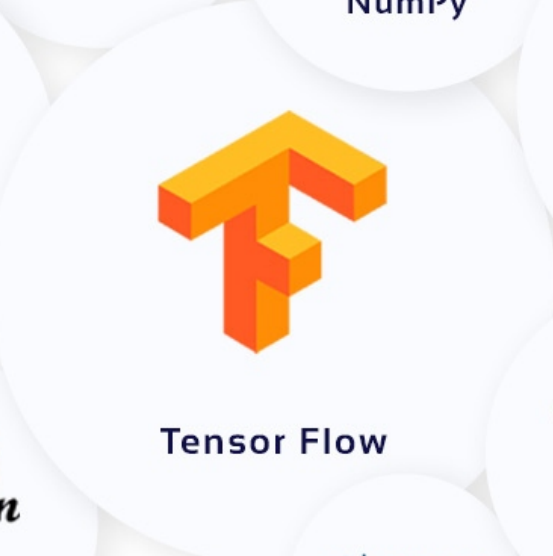


Orange3

theano
theano

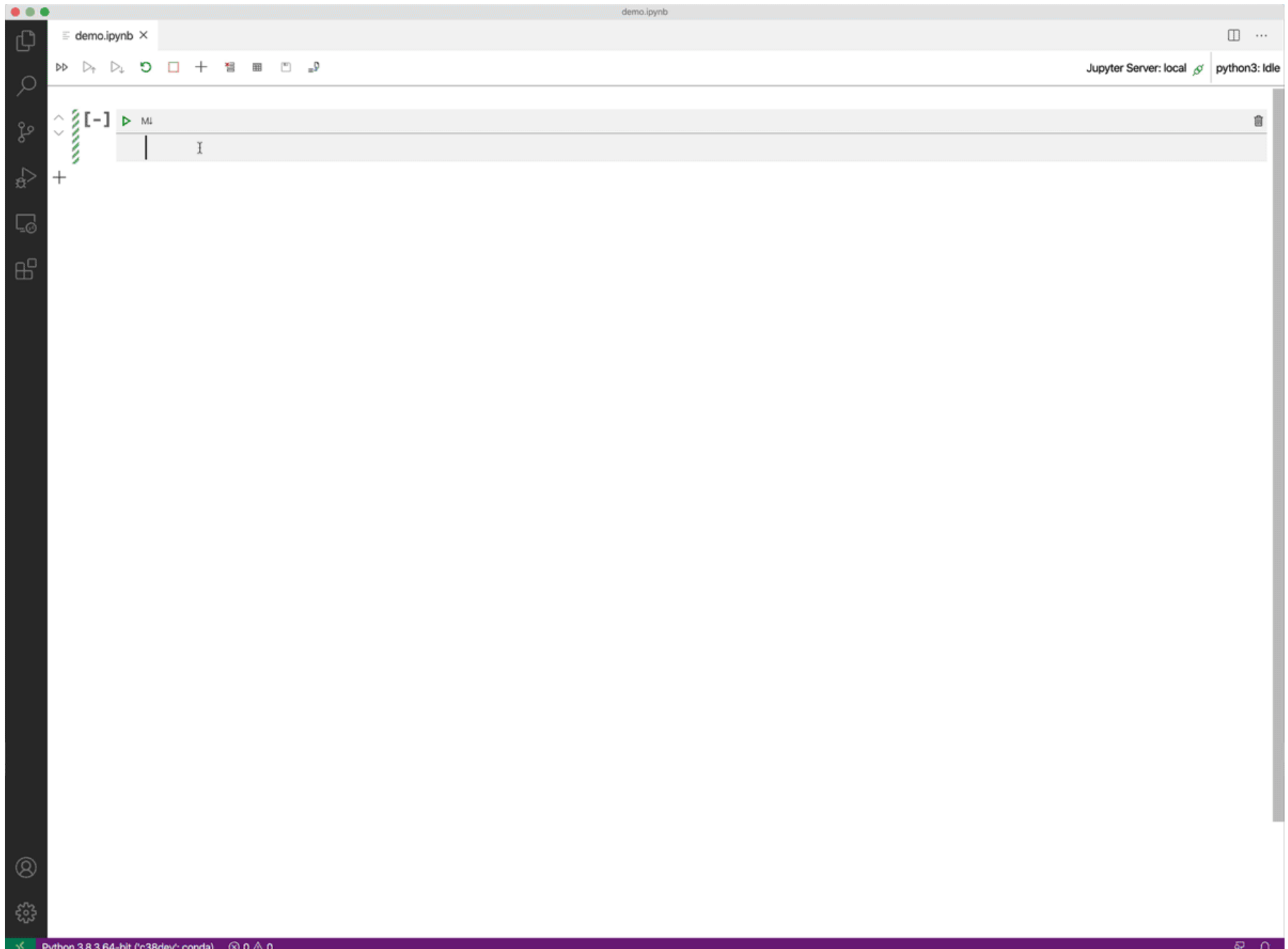


django





Cogent3 (as an alternative to biopython)



**Most important
bioinformatics databases**



> 230M

<https://www.uniprot.org>

A screenshot of the UniProt search interface. The background is a dark blue gradient. At the top center, the text "Find your protein" is written in white. Below this is a search bar with a dropdown menu on the left showing "UniProtKB" and a "Search" button on the right. To the right of the search bar are the options "Advanced" and "List". Below the search bar, there are examples of search terms: "Insulin, APP, Human, P05067, organism_id:9606". At the bottom of the screenshot, there is a white text box with the UniProt mission statement: "UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt".

A card for "Proteins UniProt Knowledgebase". The background is blue with a pattern of white dots. The text "Proteins" is in white, and "UniProt Knowledgebase" is in a smaller white font below it. At the bottom, there are two categories: "Reviewed (Swiss-Prot) 569,213" and "Unreviewed (TrEMBL) 245,871,679".

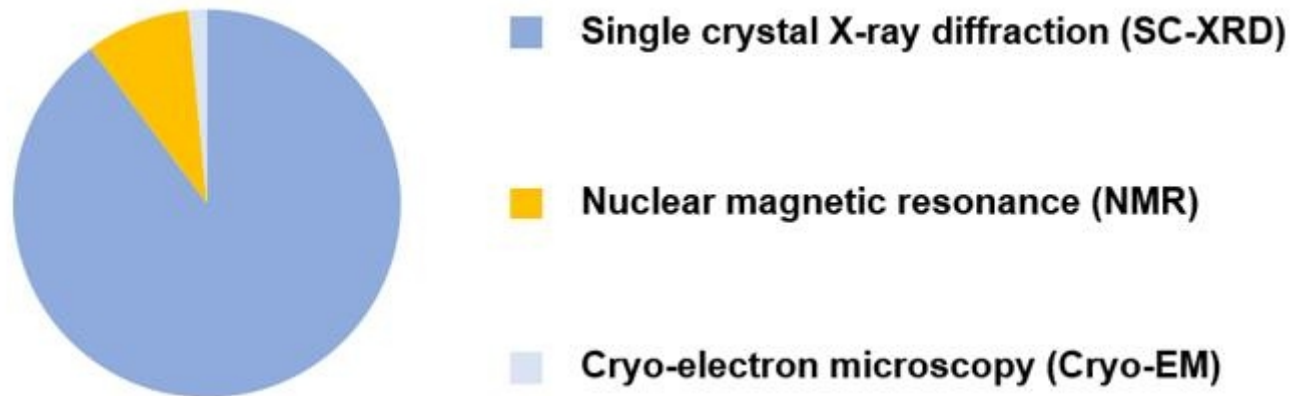
A card for "Species Proteomes". The background is red with a silhouette of a human figure and various animals. The text "Species" is in white, and "Proteomes" is in a smaller white font below it. The description reads: "Protein sets for species with sequenced genomes from across the tree of life".

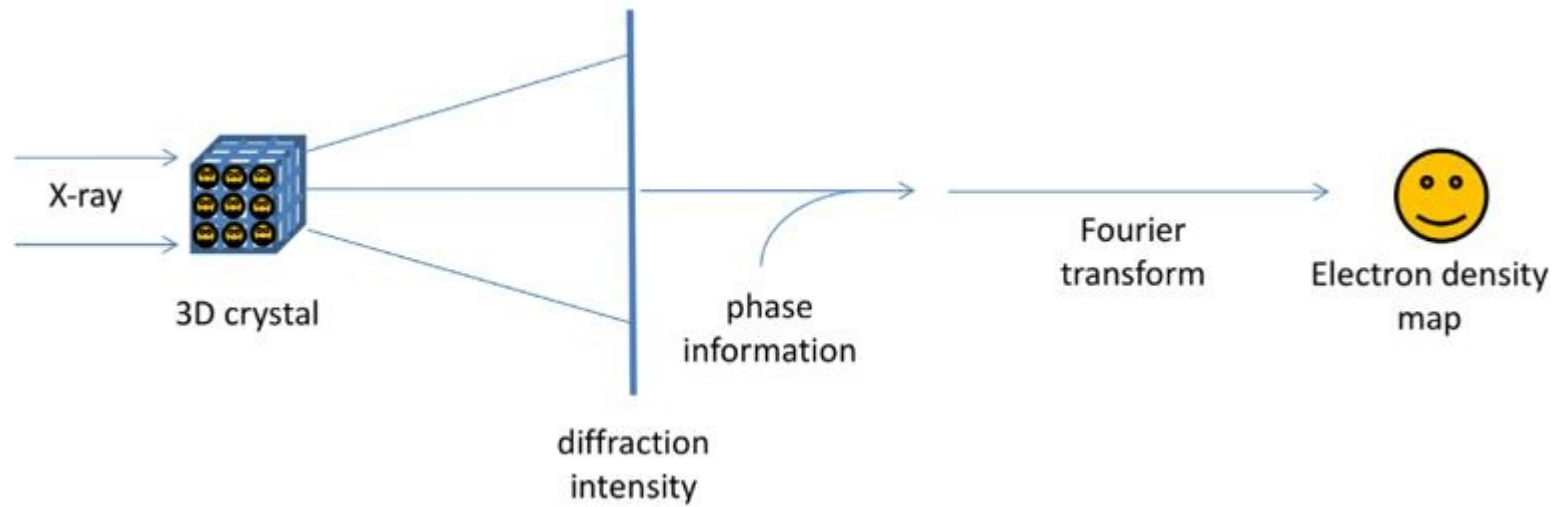
A card for "Protein Clusters UniRef". The background is orange with a pattern of white circles. The text "Protein Clusters" is in white, and "UniRef" is in a smaller white font below it. The description reads: "Clusters of protein sequences at 100%, 90% & 50% identity".

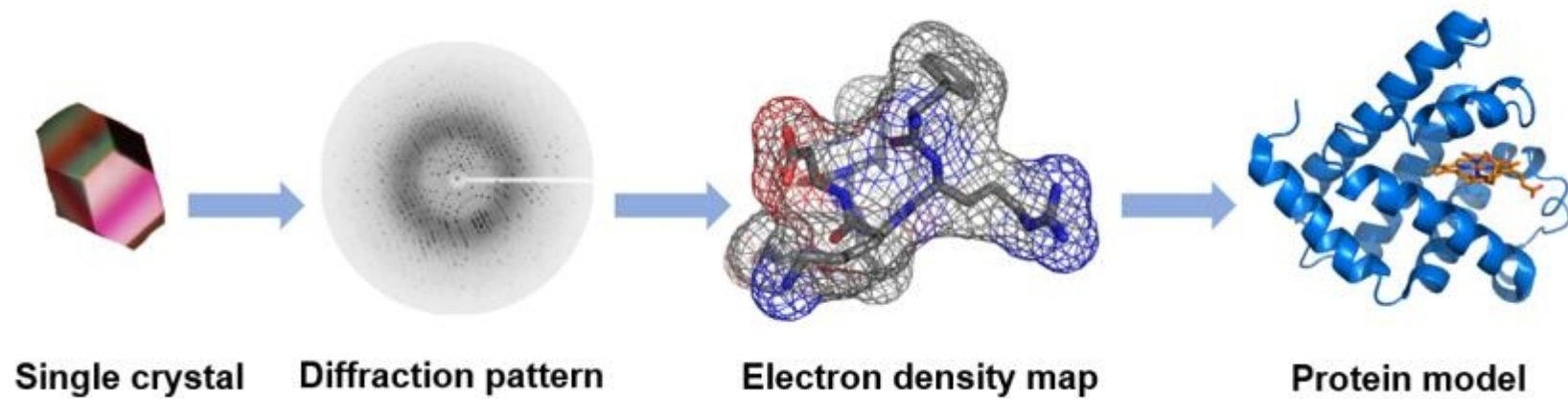
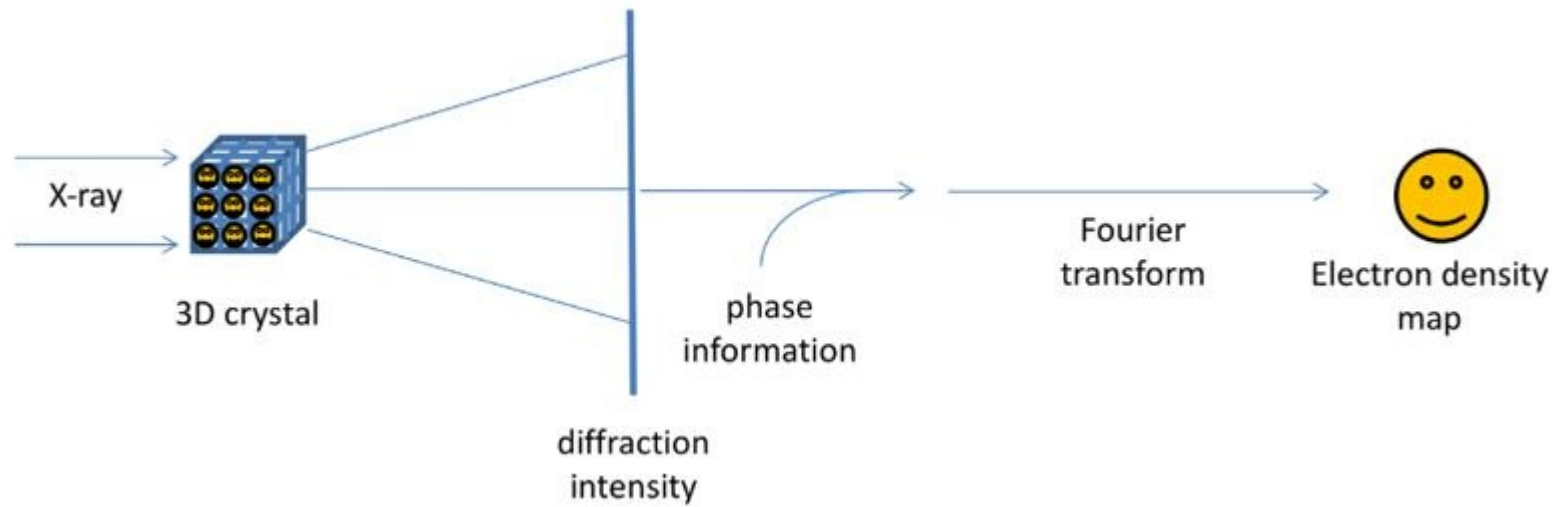
A card for "Sequence Archive UniParc". The background is green with a globe and arrows. The text "Sequence Archive" is in white, and "UniParc" is in a smaller white font below it. The description reads: "Non-redundant archive of publicly available protein sequences seen across different databases".

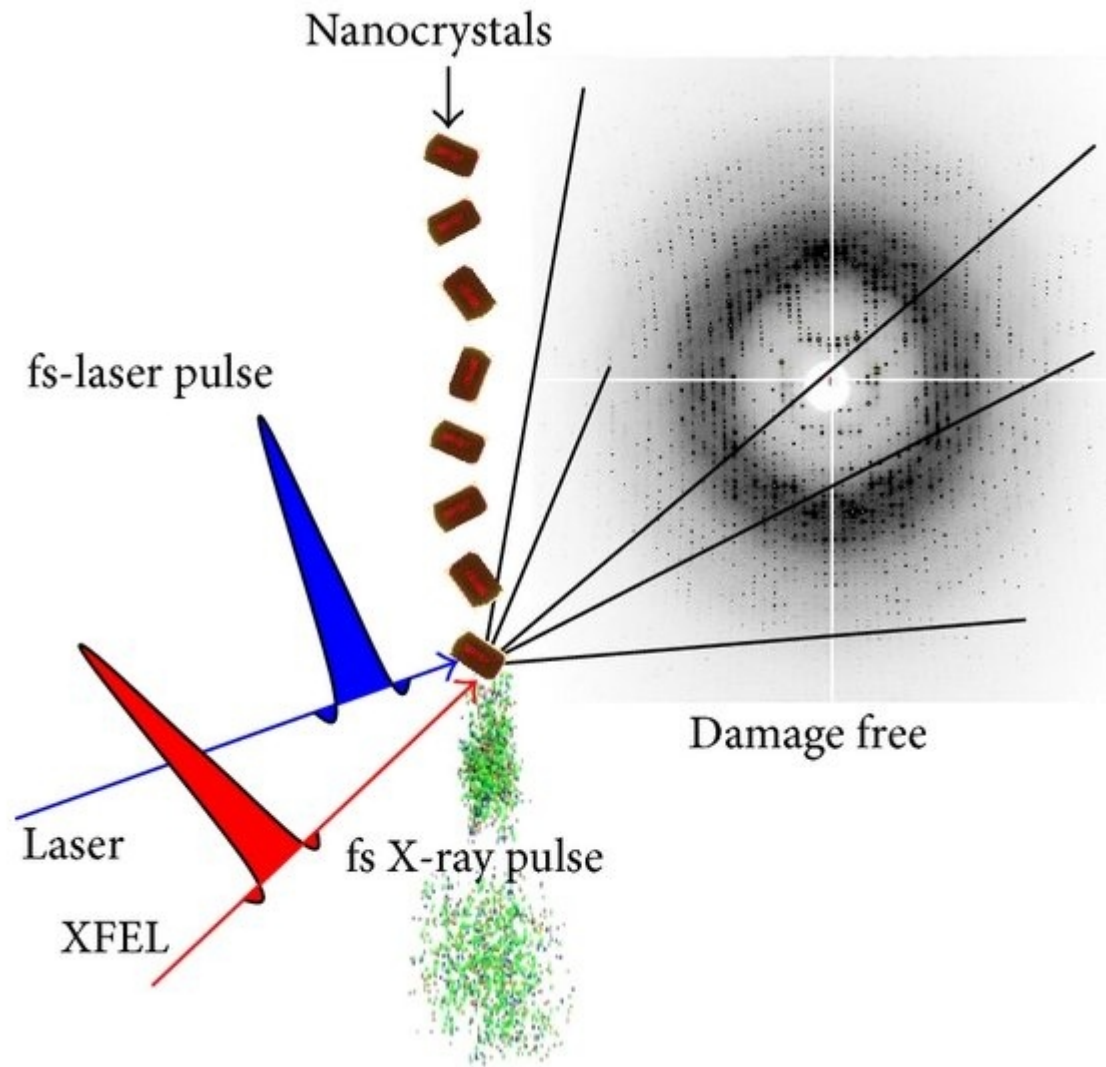


RCSB PDB PROTEIN DATA BANK

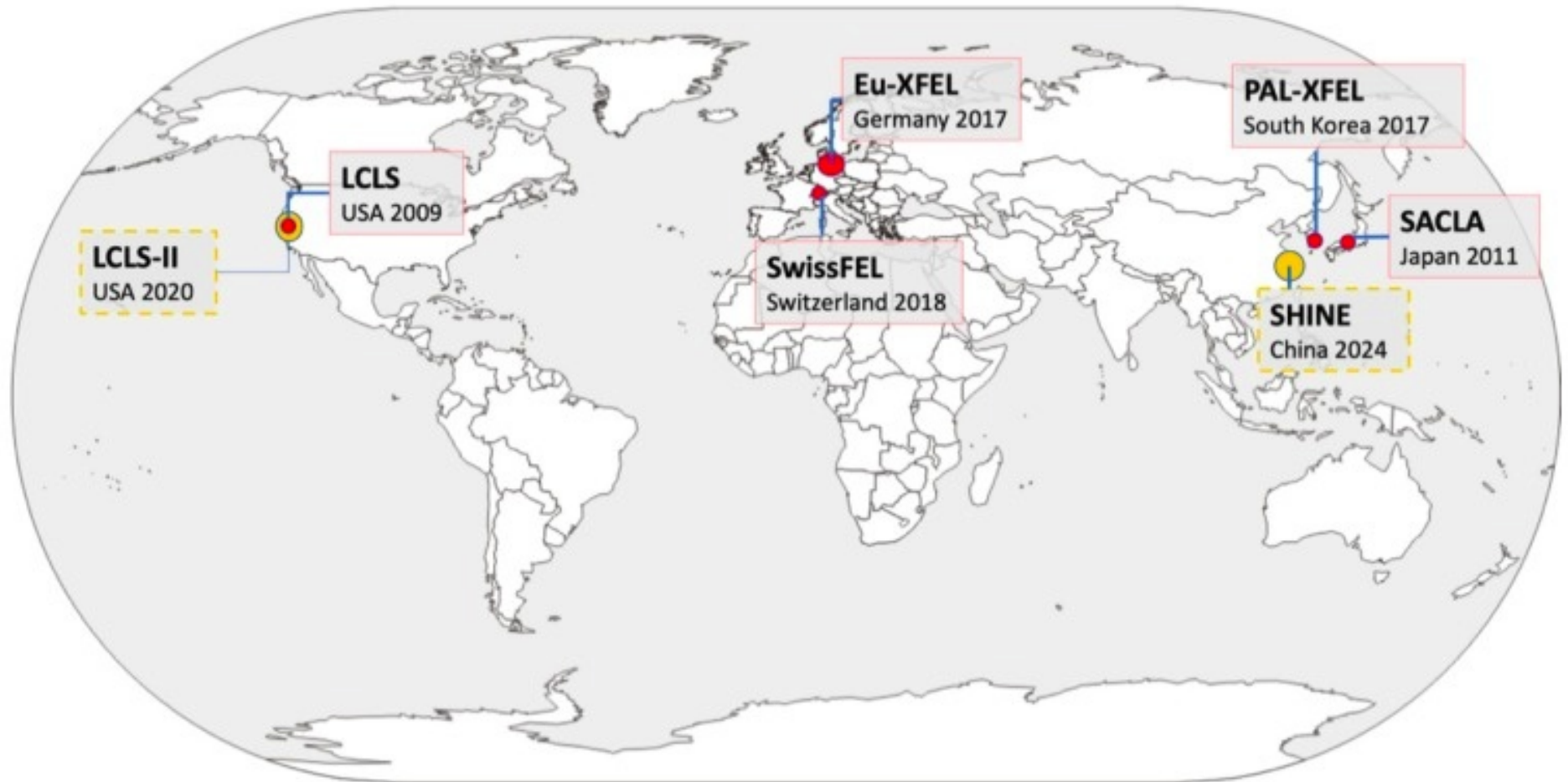


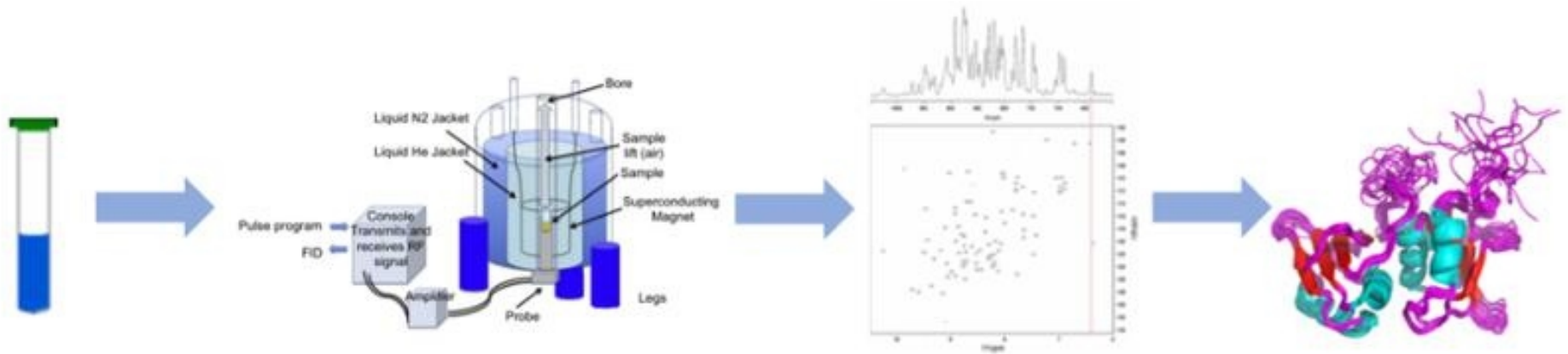






For more watch: <https://www.youtube.com/watch?v=-VMDytbTbNw>



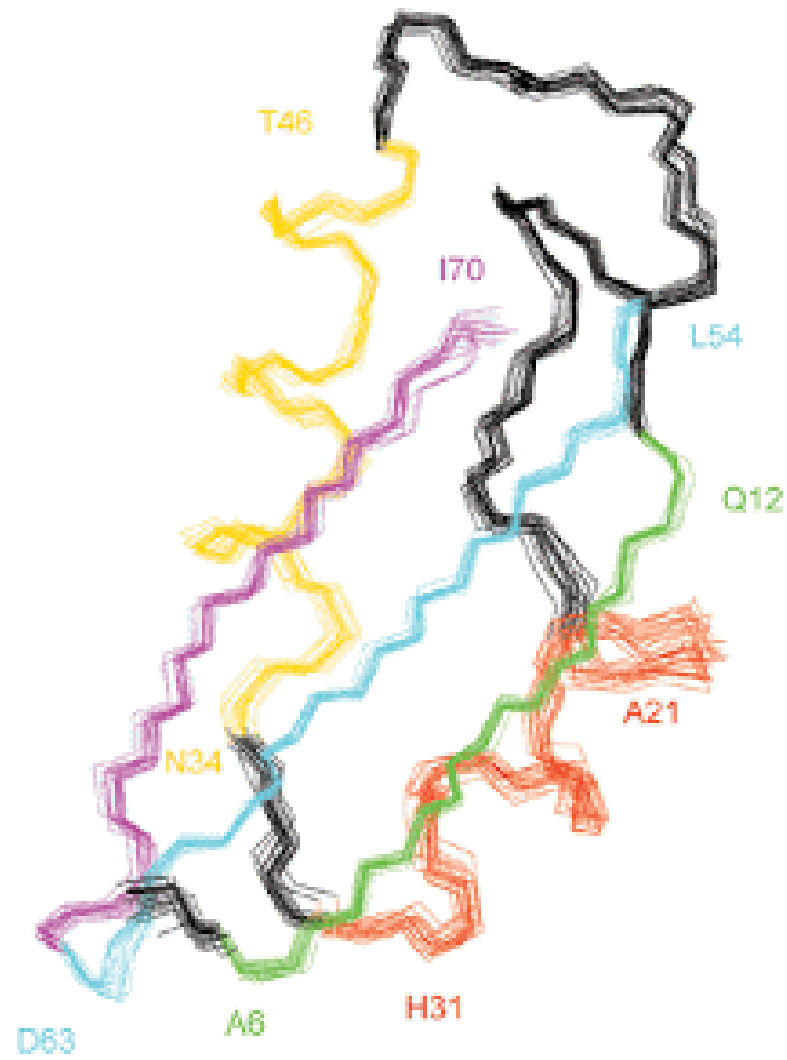


Sample preparation

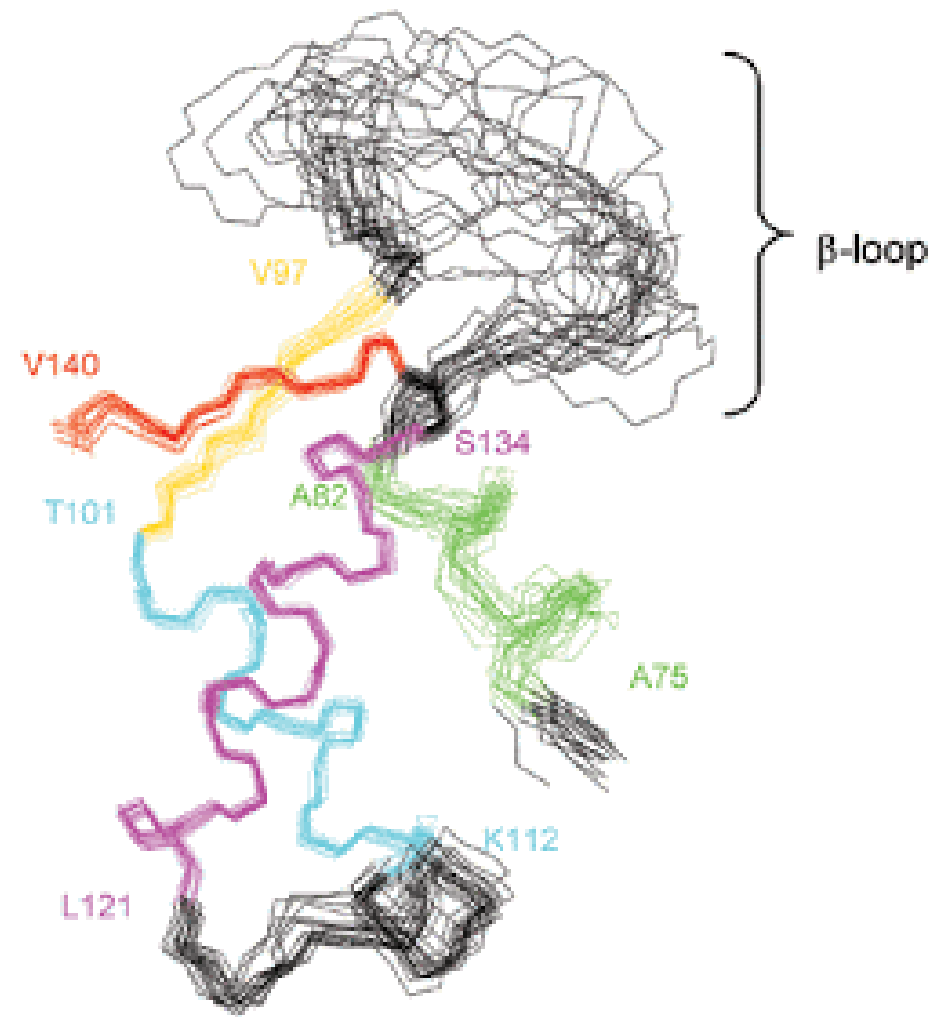
Data acquisition

Spectral processing

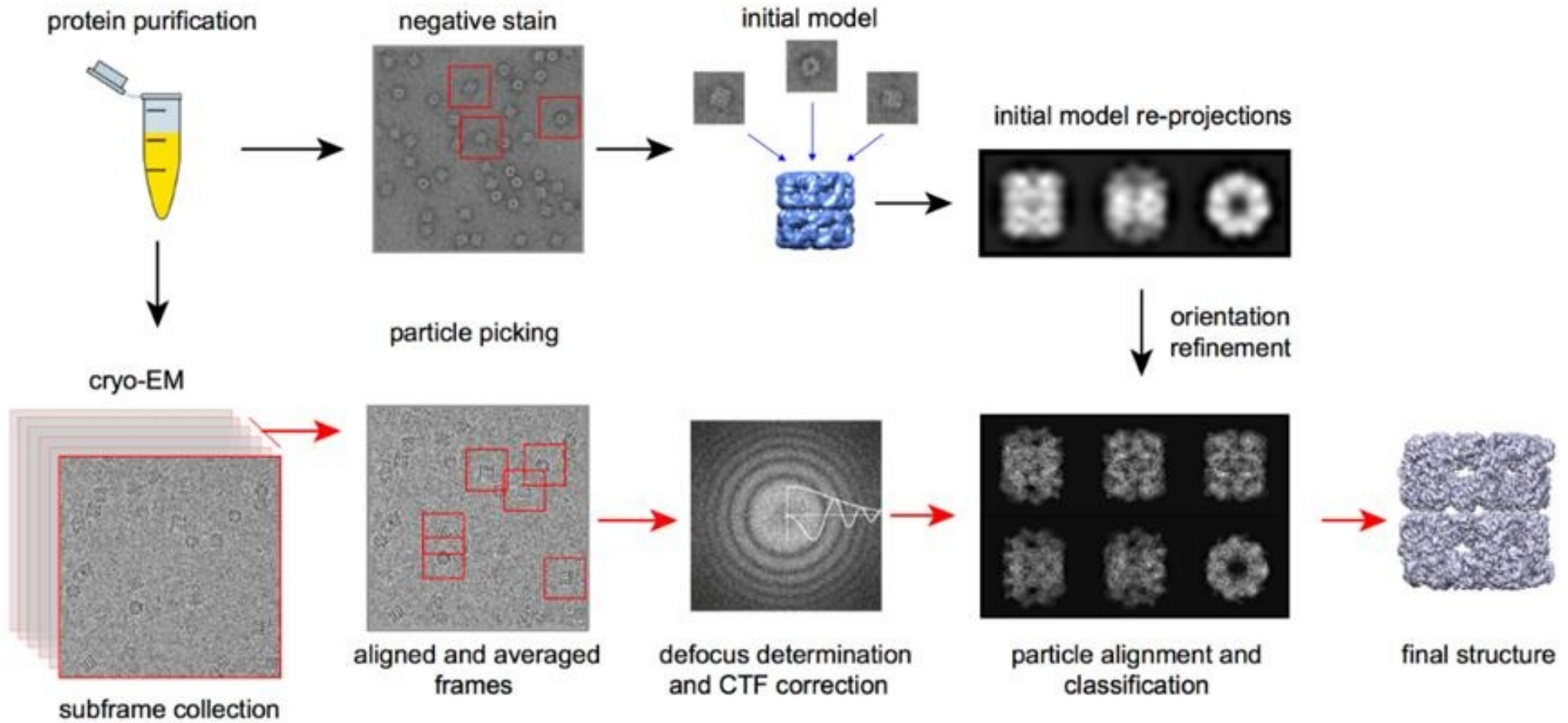
Structural analysis



L11 N-Terminus (5-70)

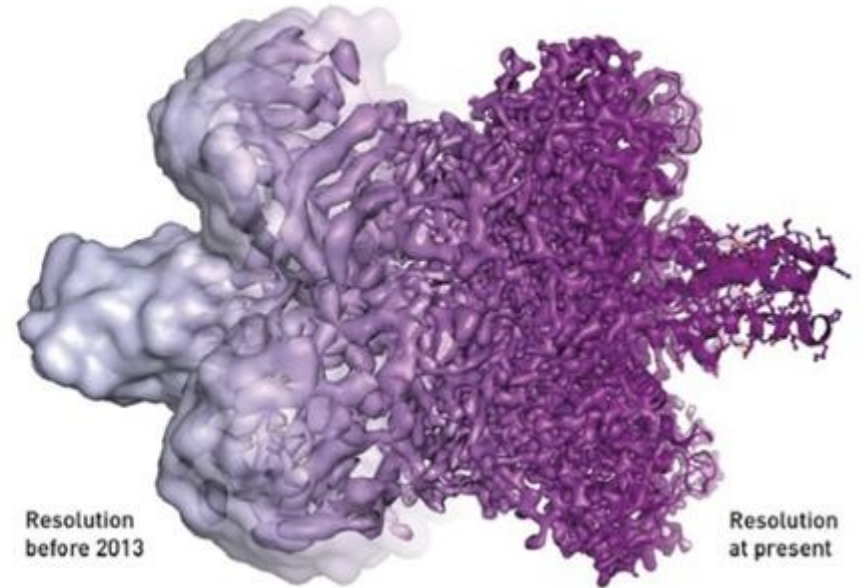
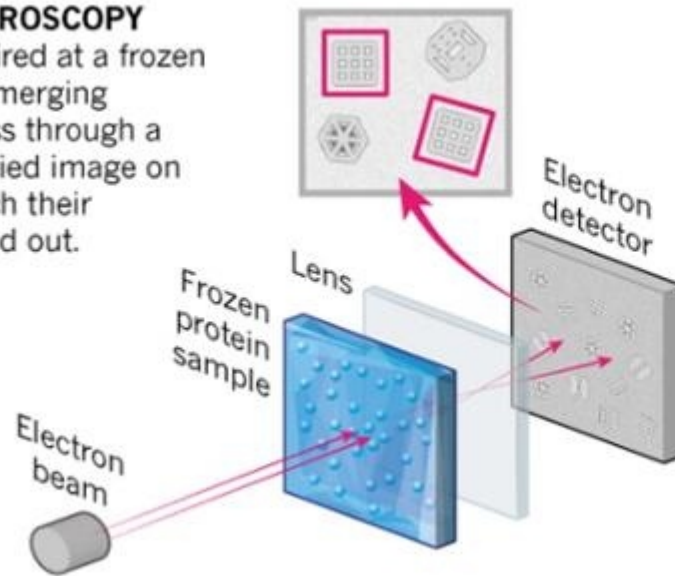


L11 C-Terminus (75-140)



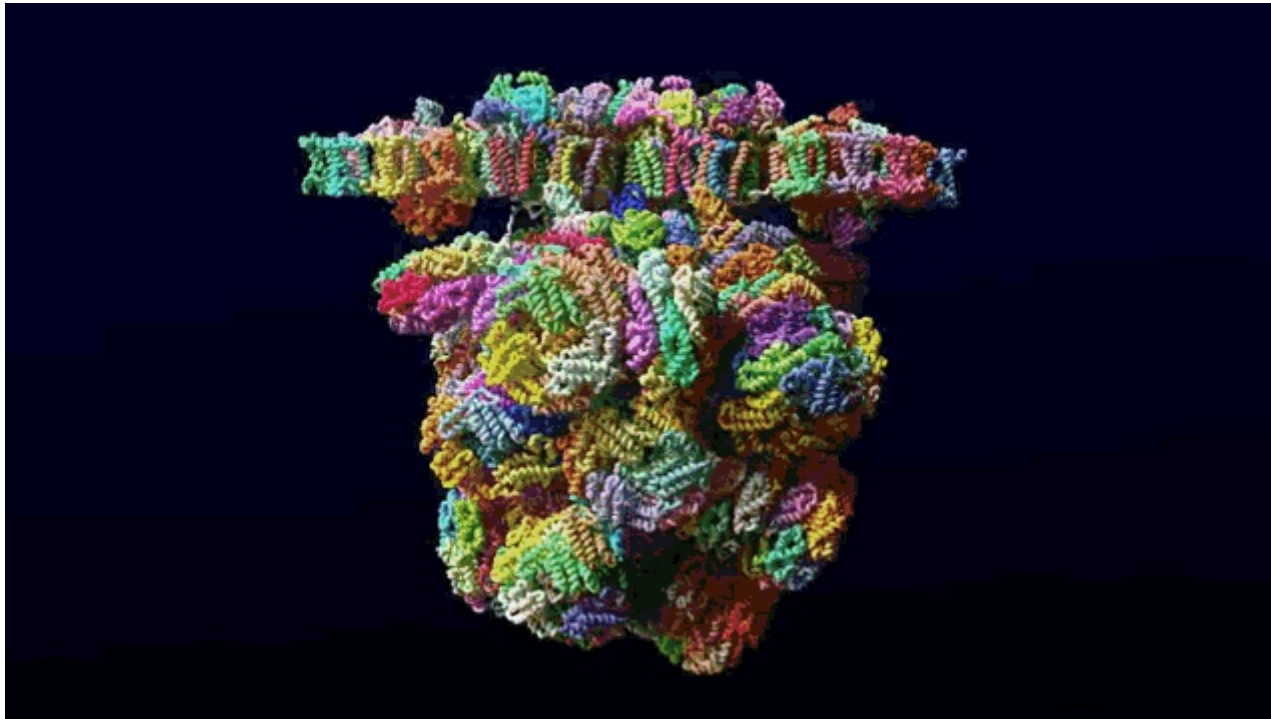
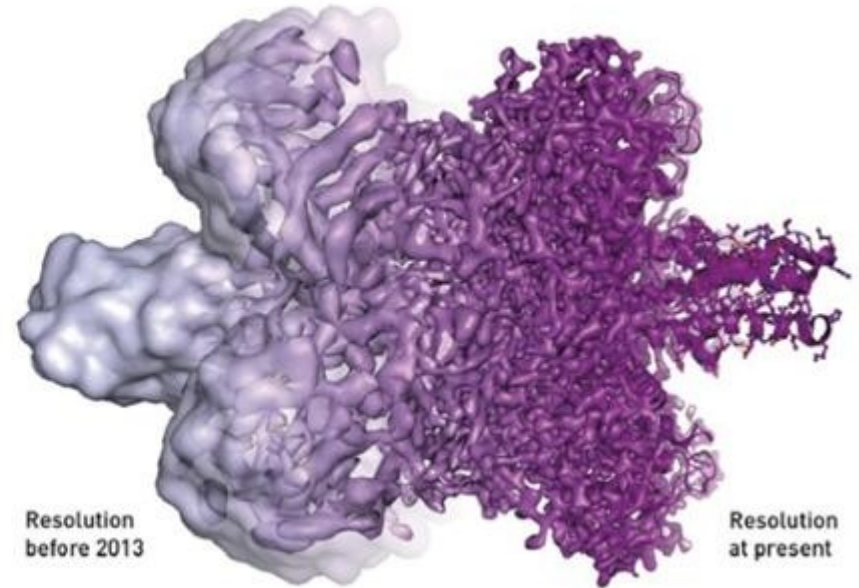
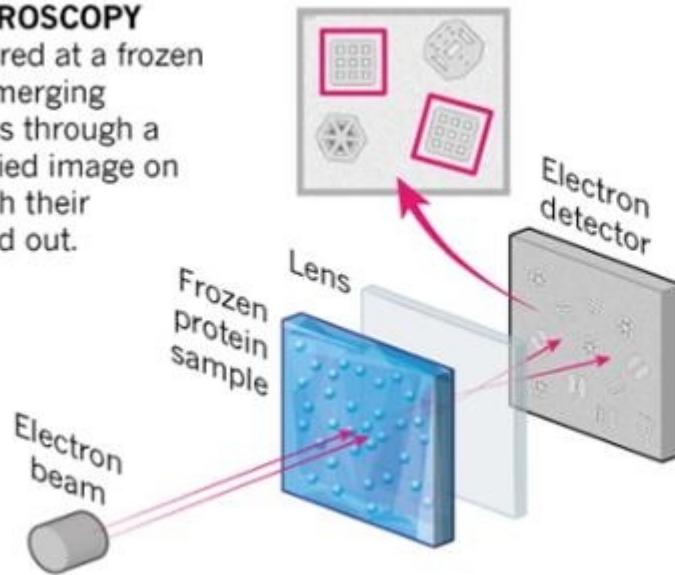
CRYO-ELECTRON MICROSCOPY

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.



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.rcsb.org/>

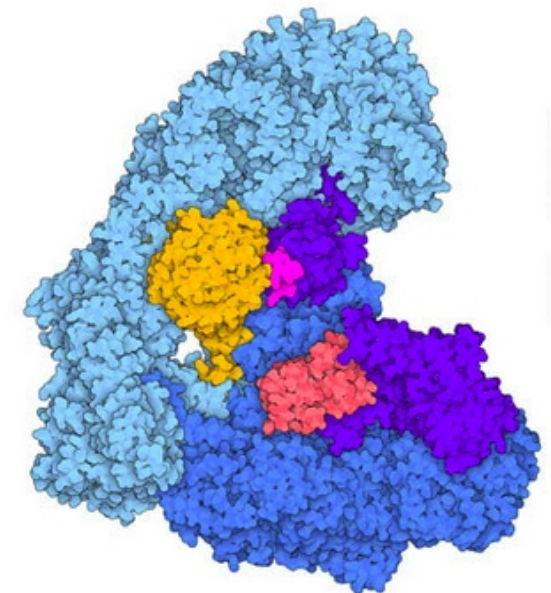
A screenshot of the RCSB PDB website homepage. The top navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, About, Documentation, and Careers, along with 'MyPDB' and 'Contact us' buttons. A left sidebar contains navigation icons for Welcome, Deposit, Search, Visualize, Analyze, Download, and Learn. The main content area features a 'Welcome' message, a list of data types (Experimentally-determined 3D structures and Computed Structure Models), and a description of the data's use. Below this are two promotional banners: one for COVID-19 coronavirus resources and another showing a counter for 20,000 structures in the PDB. On the right, a large 3D molecular model of the Anaphase-Promoting Complex / Cyclosome is shown, with different subunits colored in light blue, yellow, purple, and red.

Welcome

science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.



Anaphase-Promoting Complex / Cyclosome

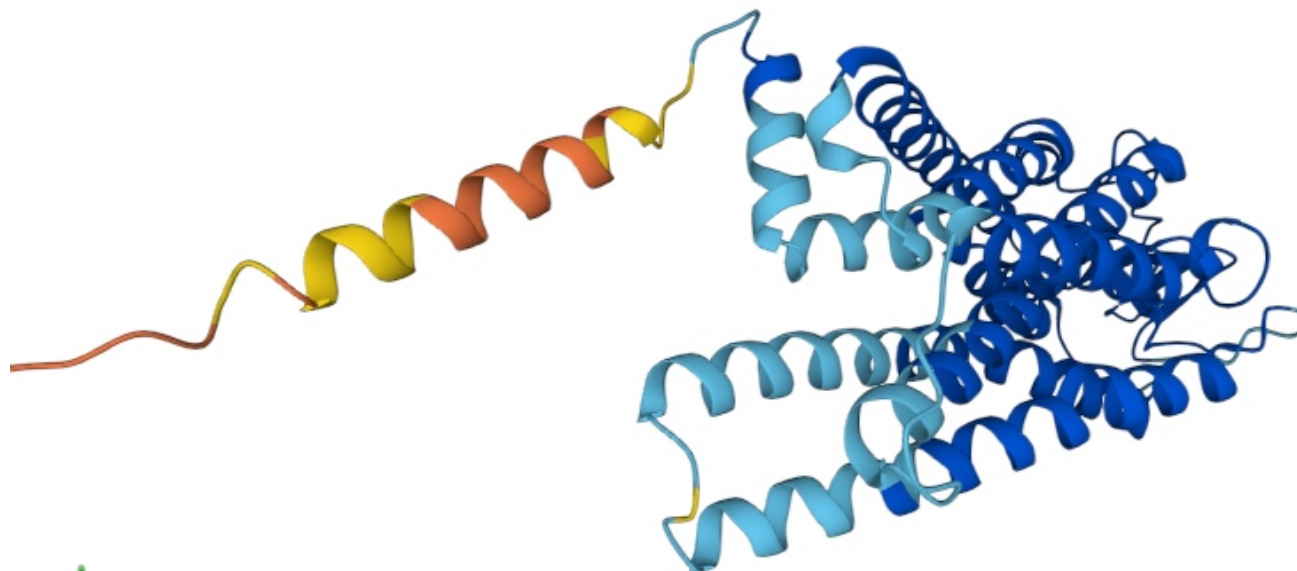


214M
25TB



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

```
MNGTEGPAFYVPMNATGVVRSPEYYPQYYLVAPWAYGLLAAYMFFLIITGFPVNFLTLYVTIEHKKLRTPLNYILLNLAIADLFMVFGGFITTTMYTSLHGYPVFGRLGCNLEGGFFATLGGEMGLW
SLVVLAIERWMVVKPVSNFRFGENHAIMGVAFTWVMACSCAVPPLVGWSRYIPEGMQSCGVDYYTTRPGVNNESFVIYMFIVHFFIPLIVIFFCYGRLVCTVKEAAAQQQESETTQRAEREVTR
MVIIMVIAFLICWLPYAGVAWYIFTHQGSEFGPVFMTLPFAFFAKTSAVYNPCYICMNKQFRHCMITTLCCGKNPFEEEEEGASTTASKTEASSVSSSSVSPA
```





Meta AI

617M sequences & 15TB data

A screenshot of the ESM Atlas website. The browser address bar shows the URL: <https://esmatlas.com/explore/detail/MGY000954702682>. The main content area displays a protein sequence for Chain 1: A. The sequence is: `MLFLQNDVWDAHYIPLTVERQPFLIGFQESMDAGIPQRQPVVHIDLDPKVVSSSQGQAVFLEHGGESPLLERINSVLLTIHQGNEMNQ`
`FSKLLIGLDLVEPSTMEFSLINGEKHTLTGLHIINQERLSKLSGNAETLHQHGHLSIYMMMLASMPNFRKLIDRKNAILKSEIDAV`. Below the sequence is a 3D ribbon diagram of the protein structure, colored by local prediction confidence (pLDDT). The structure is primarily blue, indicating high confidence (pLDDT > 0.9), with a small yellow region indicating low confidence (pLDDT < 0.5). To the right of the structure, there is a "Download" section with two buttons: "PDB file" and "Sequence". Below the buttons, a text box explains: "The predicted structure is colored by local prediction confidence (pLDDT) per amino acid location. Blue indicates confident predictions (pLDDT > 0.9), while red indicates low confidence (pLDDT < 0.5)." At the bottom right, there is a "Share" button with icons for Twitter, Facebook, and a link icon.

<https://esmatlas.com/>

EMBL-EBI | MGnify

MGnify

Submit, analyse, discover and compare microbiome data

Example searches: Tara oceans, MGYS00000410, Human Gut

Overview [Submit data](#) [Text search](#) [Sequence search](#) [Browse data](#) [API](#) [About](#) [Help](#) [Login](#)



Human
(150056)



Digestive system
(99928)



Aquatic
(50857)



Marine
(37471)



Digestive system
(33358)



Plants
(28666)



Soil
(24969)



Skin
(10861)



Wastewater
(4261)



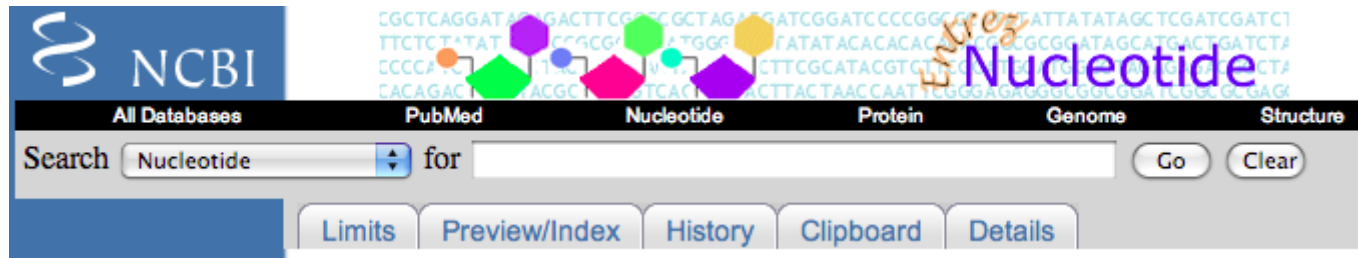
Food production
(2805)

> **2.4B sequences**

[View all biomes](#)

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

GenBank



GenPept

Search Protein for lambda bacteriophage Cro

Display FASTA Show: 20 Send to Text

Items 1-20 of 55 Page 1 of 3 Next

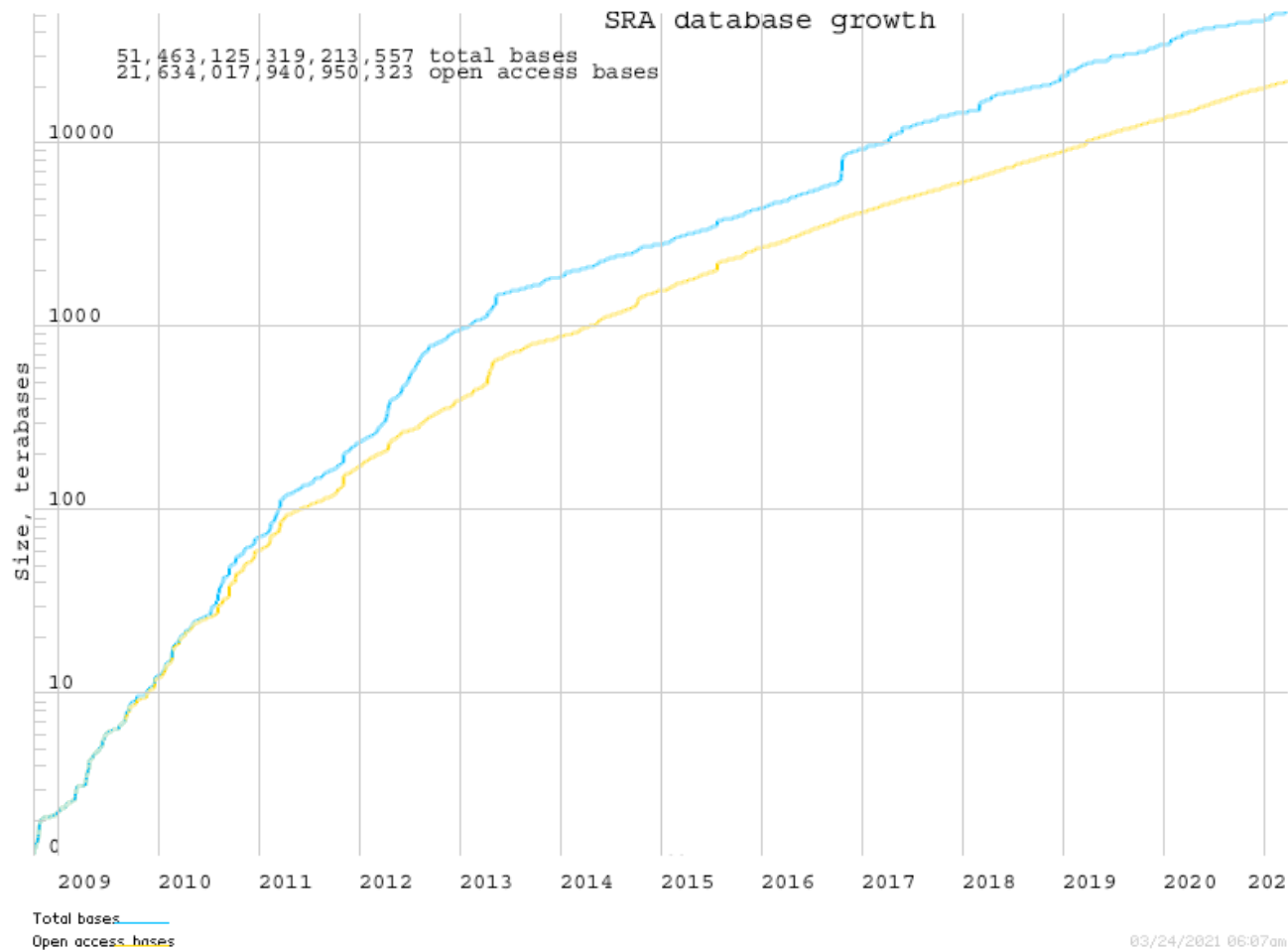
- 1: [NP_050114](#) [BLink](#), [Links](#)
Tec protein [Lactobacillus bacteriophage phi adh]
gi|9633006|ref|NP_050114.1|[9633006]
- 2: [NP_112054](#) [BLink](#), [Links](#)
cro [Bacteriophage HK620]
gi|13559844|ref|NP_112054.1|[13559844]

Entrez Protein Help | FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve sequences

Check sequence revision history

Sequence Read Archive (SRA)



Sequence Read Archive (SRA)

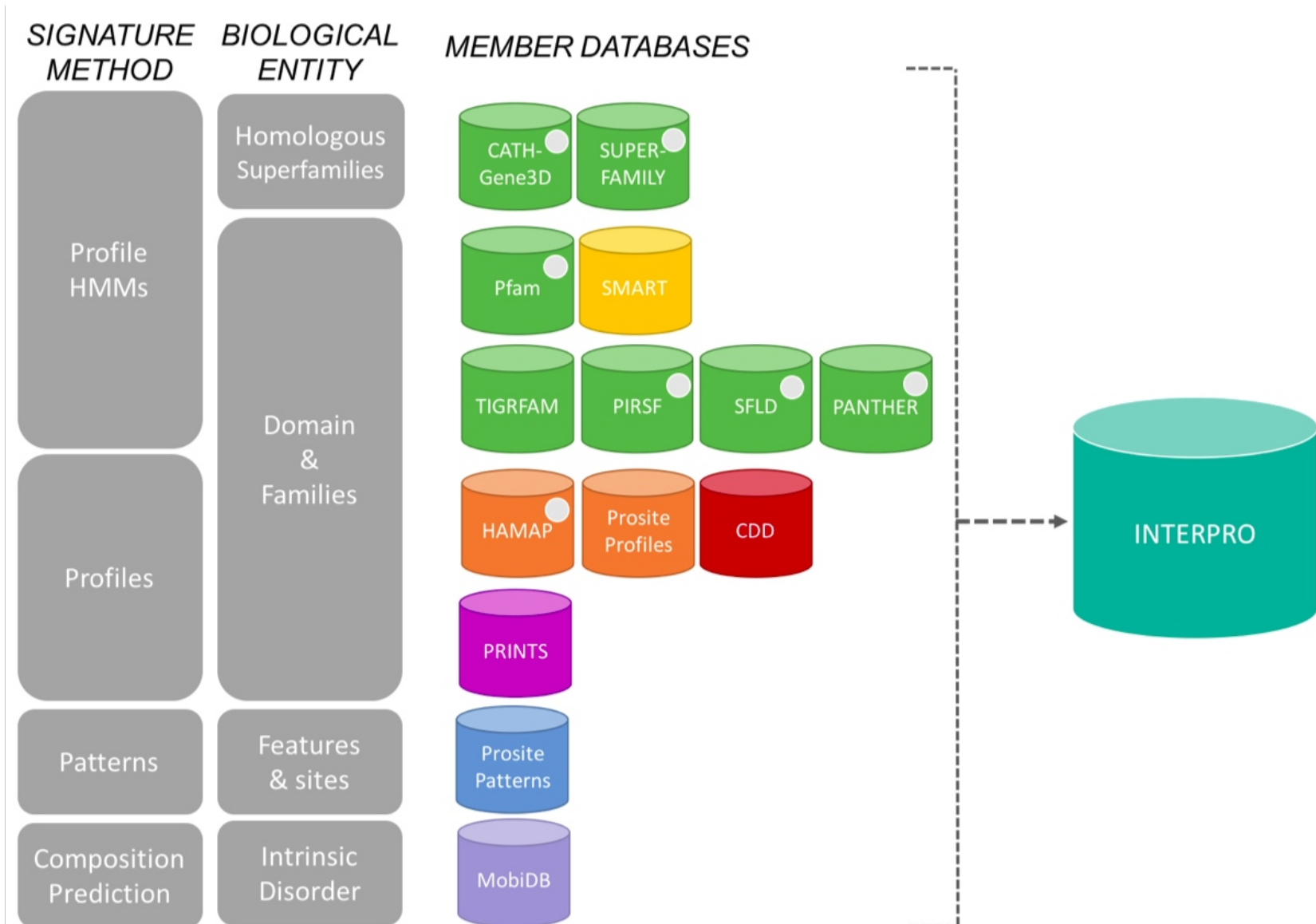
Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores

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





Family: AceK (PF06315)

4 architectures

787 sequences

0 interactions

696 species

10 structures

Summary

Domain organisation

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Structures

Jump to...

Go

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

There are 777 sequences with the following architecture: AceK

[K4KM92_SIMAS](#) [Simidiua agarivorans (strain DSM 21679 / JCM 13881 / BCRC 17597 / SA1)] Isocitrate dehydrogenase kinase/phosphatase {ECO:0000256|HAMAP-Rule:MF_00747} (578 residues)



[Show](#) all sequences with this architecture.

There are 4 sequences with the following architecture: AceK x 2

[A0A257D8B1_9PSED](#) [Pseudomonas sp. PGPPP3] Bifunctional isocitrate dehydrogenase kinase/phosphatase {ECO:0000313|EMBL:OYT97460.1} (Fragment) (401 residues)



[Show](#) all sequences with this architecture.

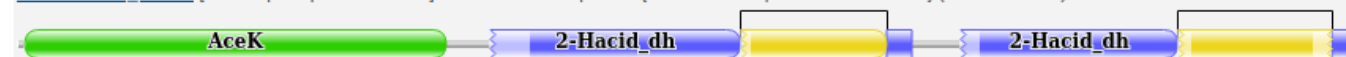
There is 1 sequence with the following architecture: Pro_CA, AceK

[E9HWS8_DAPPU](#) [Daphnia pulex (Water flea)] Carbonic anhydrase {ECO:0000256|ARBA:ARBA00012925} (408 residues)

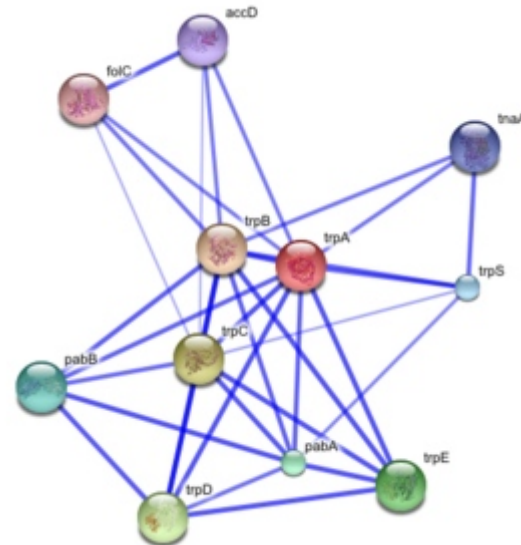


There is 1 sequence with the following architecture: AceK, 2-Hacid_dh, 2-Hacid_dh_C, 2-Hacid_dh, 2-Hacid_dh_C

[A0A4U1AIU3_9DELT](#) [Desulfopila sp. IMCC35006] Uncharacterized protein {ECO:0000313|EMBL:TKB25201.1} (1703 residues)



Showing all 4 architectures.



Reactome – biological pathways

The screenshot displays the Reactome website interface. At the top, the Reactome logo is on the left, and navigation options for 'Homo sapiens' are in the center. On the right, there are buttons for 'Citation', 'Analysis', 'Tour', and 'Layout'. Below the navigation bar is a search bar with the placeholder text 'Search for a term, e.g. pten ...'. The main area features a large, complex network diagram of biological pathways, with various nodes and connections. A sidebar on the left lists an 'Event Hierarchy' with categories such as Autophagy, Cell Cycle, Cell-Cell communication, Cellular responses to external stimuli, Chromatin organization, Circadian Clock, Developmental Biology, Digestion and absorption, Disease, DNA Repair, DNA Replication, Extracellular matrix organization, Gene expression (Transcription), Hemostasis, Immune System, Metabolism, Metabolism of proteins, Metabolism of RNA, Muscle contraction, Neuronal System, Organelle biogenesis and maintenance, Programmed Cell Death, and Protein localization. At the bottom, there is a navigation bar with tabs for 'Description', 'Molecules', 'Structures', 'Expression', 'Analysis', and 'Downloads'. A text box with a clipboard icon explains that the interface displays details when an item is selected in the Pathway Browser, including input and output molecules, summary, references, and supporting evidence.

reactome 3.7 75

Pathways for: Homo sapiens

Citation: Analysis: Tour: Layout:

Search for a term, e.g. pten ...

Event Hierarchy:

- Autophagy
- Cell Cycle
- Cell-Cell communication
- Cellular responses to external stimuli
- Chromatin organization
- Circadian Clock
- Developmental Biology
- Digestion and absorption
- Disease
- DNA Repair
- DNA Replication
- Extracellular matrix organization
- Gene expression (Transcription)
- Hemostasis
- Immune System
- Metabolism
- Metabolism of proteins
- Metabolism of RNA
- Muscle contraction
- Neuronal System
- Organelle biogenesis and maintenance
- Programmed Cell Death
- Protein localization

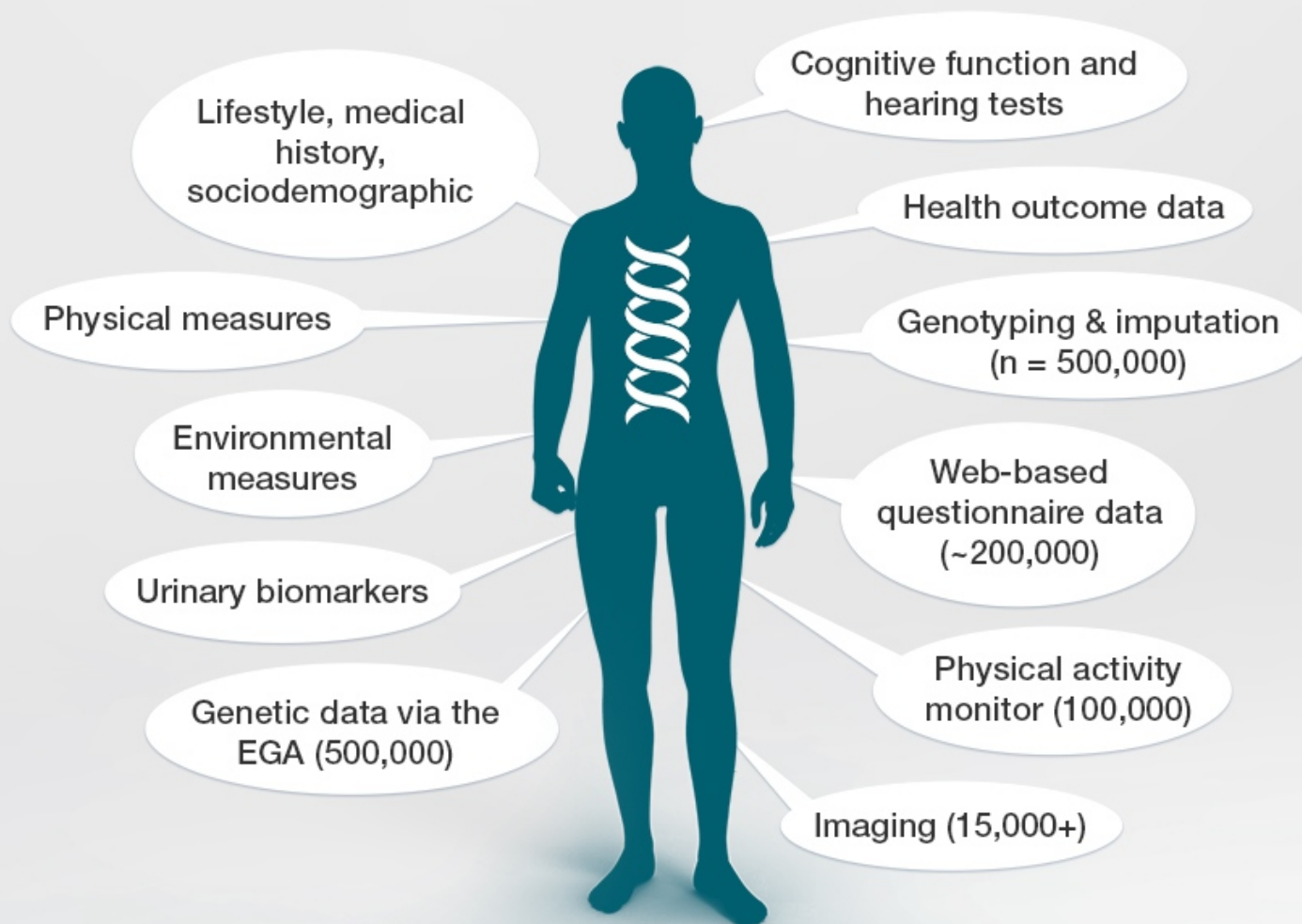
Description Molecules Structures Expression Analysis Downloads

Displays details when you select an item in the Pathway Browser. For example, when a reaction is selected, shows details including the input and output molecules, summary and references containing supporting evidence. When relevant, shows details of the catalyst, regulators, preceding and following events.

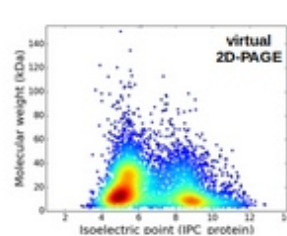


Enabling scientific discoveries that improve human health

Data on UK Biobank participants

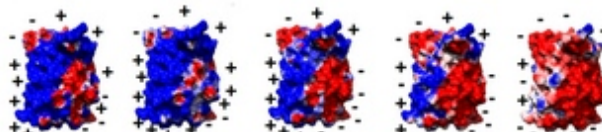


<https://www.ukbiobank.ac.uk/>

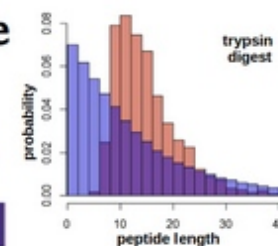


Proteome-pI 2.0: Proteome Isoelectric Point Database

protein
protonated
positively charged



protein
deprotonated
negatively charged



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

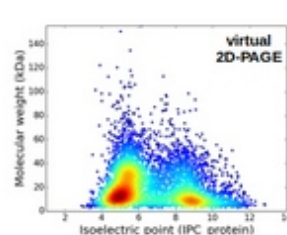
Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

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

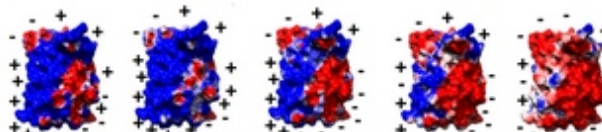
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

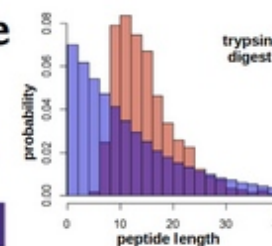


Proteome-pI 2.0: Proteome Isoelectric Point Database

protein
protonated
positively charged



protein
deprotonated
negatively charged



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

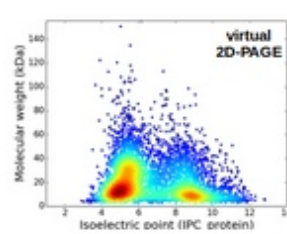
Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

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

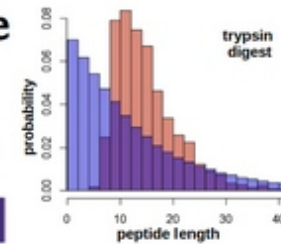
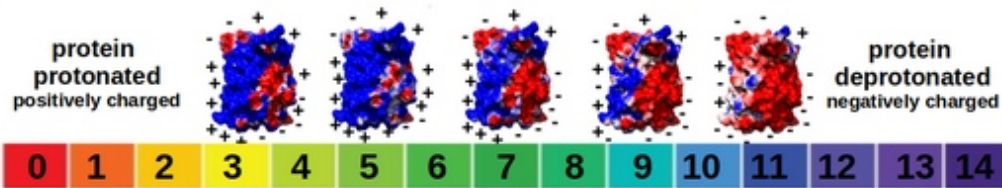
5.38 Billion dissociation constant (pKa) predictions for proteins

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



Proteome-pI 2.0: Proteome Isoelectric Point Database



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

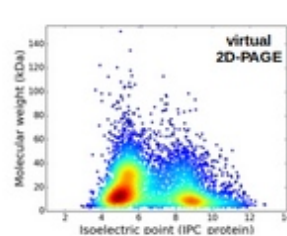
Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

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

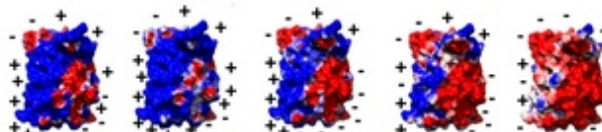
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

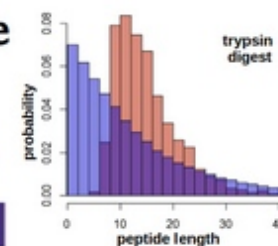


Proteome-pI 2.0: Proteome Isoelectric Point Database

protein
protonated
positively charged



protein
deprotonated
negatively charged



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

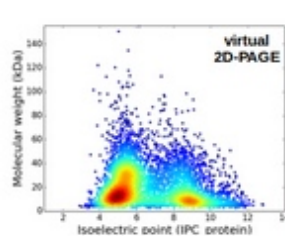
Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

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

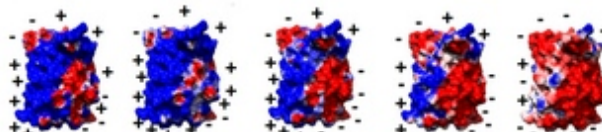
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

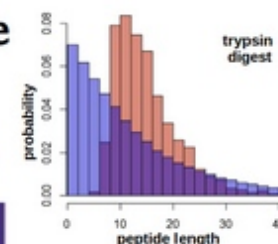


Proteome-pI 2.0: Proteome Isoelectric Point Database

protein
protonated
positively charged



protein
deprotonated
negatively charged



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

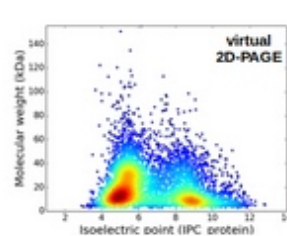
Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

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

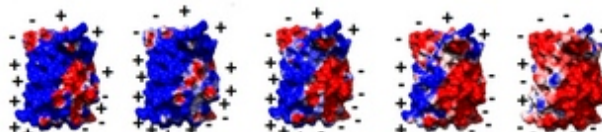
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl

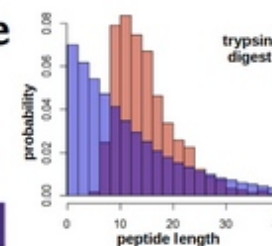


Proteome-pI 2.0: Proteome Isoelectric Point Database

protein
protonated
positively charged



protein
deprotonated
negatively charged



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

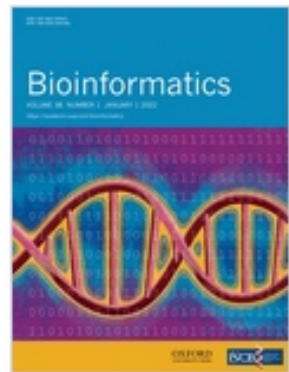
Protease digests (peptides with molecular weight and isoelectric point) can be useful for bottom-up proteomics MS analysis

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



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>

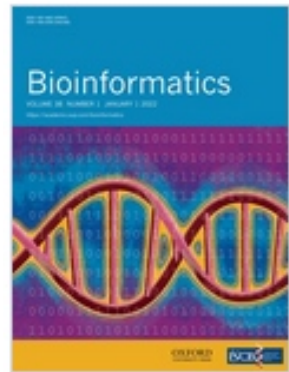
Published: 14 July 2021 **Article history** ▼

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

5.38 Billion dissociation constant (pK_a) 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



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

Denature protein
in 6–8M urea.



Protein resists
digestion due
to tight folding.

 Lys-C

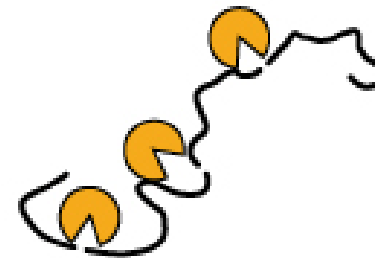
 Trypsin

Digest with Trypsin/Lys-C
Mix for 3–4 hours.



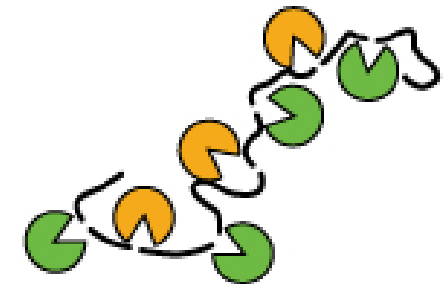
Protein denatures
and is available
for digestion.

Digest with Trypsin/Lys-C
Mix for 3–4 hours.



Lys-C digests protein
into relatively large
fragments. Trypsin is
reversibly inactivated
by urea.

Dilute reaction and
incubate overnight.



Trypsin reactivates
and completes
digestion.

Trypsin cleavage specificity



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

NRRPC**HSHTKE**CE**SAWKNR**RPC**HSHTKK**PC**HSHTKKNR**K**VWKI**PP**FFW**
✂ ✂ ✂ ✂ ✂ ✂ ✂ ✂ ✂

trypsin digest



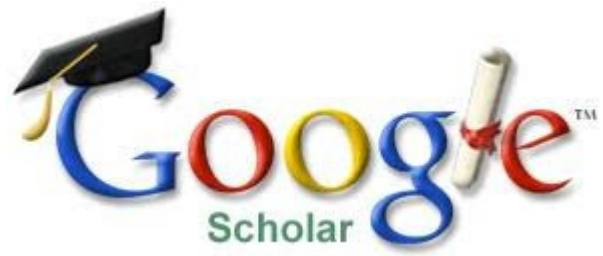
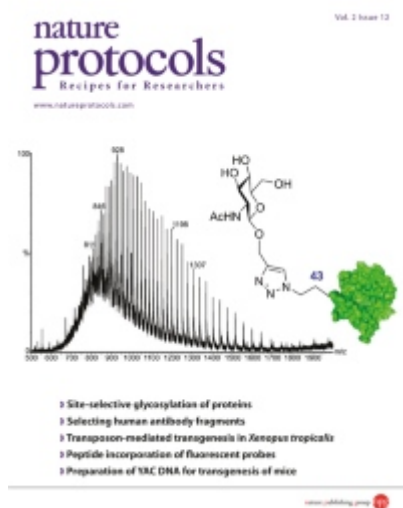
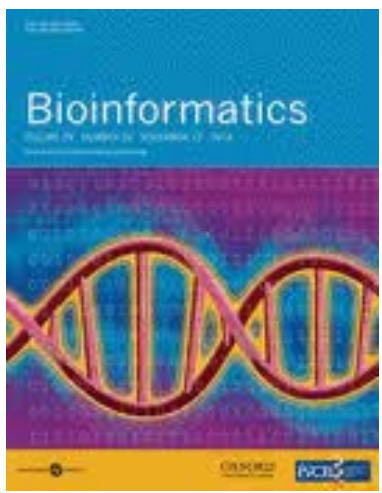
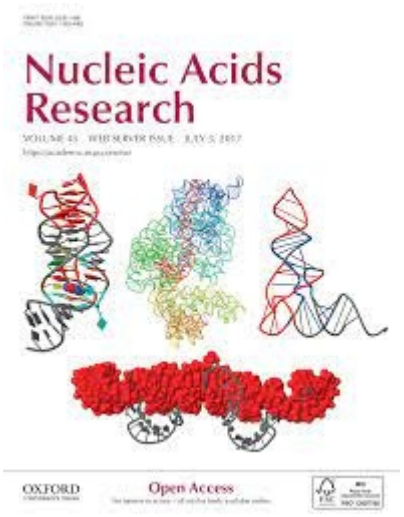
~~NR~~ ECESAWK KPC**HSHTK** ~~NR~~ IPPFFW
RPC**HSHTK** NR**PC**HSHTK ~~HSHTK~~ ~~KVWK~~

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



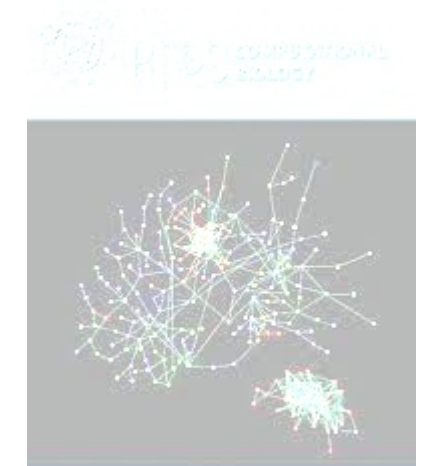
Proteomes *in silico* digested with the five most frequently used proteases (Trypsin, Chymotrypsin, Trypsin+LysC, LysN, ArgC)
In total, **9.58 Billion** peptides

isoelectricpointdb2.mimuw.edu.pl





RESEARCH



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