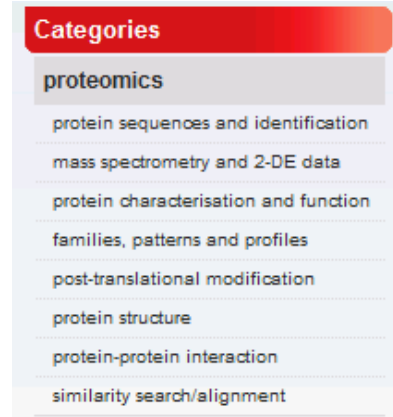


Protein Research & beyond: ExPASy - the new SIB Bioinformatics Resource Portal.

This workshop will introduce you to ExPASy, the new SIB Bioinformatics Resource Portal, which provides access to scientific databases and software tools in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. Different protein sequence formats (Flat file, FASTA, Prosite) will be used to search protein databases and protein analysis tools will be discussed.



Categories
proteomics
protein sequences and identification
mass spectrometry and 2-DE data
protein characterisation and function
families, patterns and profiles
post-translational modification
protein structure
protein-protein interaction
similarity search/alignment

UniProt

UniProt Consortium

- European Bioinformatics Institute, Hinxton UK (EBI)
- Swiss Institute of Bioinformatics, Genf CH (SIB)
- Protein Information Resource Washington DC USA (PIR)

Goal: Minimal redundancy, high quality, integration to other databases, free access

The UniProt databases consist of three database layers:

- The **UniProt Archive (UniParc)** provides a stable, comprehensive sequence collection without redundant sequences by storing the complete body of publicly available protein sequence data.
- The **UniProt Knowledgebase (UniProtKB)** is the central database of protein sequences with accurate, consistent, and rich sequence and functional annotation.
- The **UniProt Reference Clusters (UniRef)** databases provide non-redundant reference clusters based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions. (UniRef 100, 90, 50 - 40% (bzw. 65%) less entries)

UniProt KB

The UniProtKB provides the central database of protein sequences with accurate, consistent, rich sequence and functional annotation.

The UniProt Knowledgebase consists of two sections:

- Swiss-Prot - a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis (standard and preliminary)
- TrEMBL - a section with computationally analyzed records that await full manual annotation.

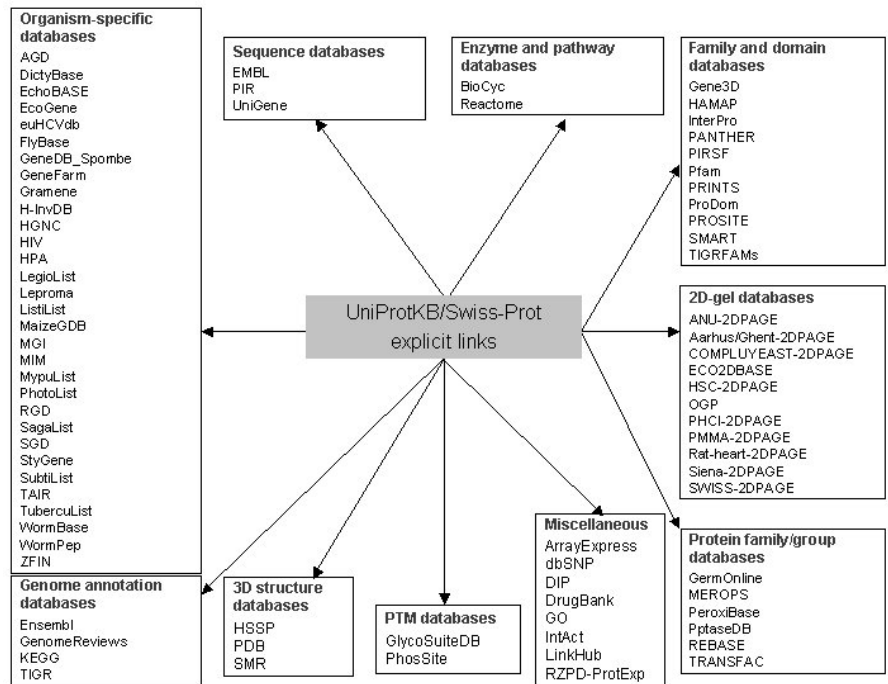
1. In Swiss-Prot, two classes of data can be distinguished: the core data and the annotation.

For each sequence entry the core data consists of:

- sequence data
- citation information (bibliographical references)
- taxonomic data (the biological source of the protein).

The annotation/description consists of the following items:

- Function(s) of the protein
- Posttranslational modification(s): carbohydrates, phosphorylation
- Domains and sites: calcium- or ATP-binding sites, zinc fingers
- Secondary structure: alpha helix, beta sheet
- Quaternary structure: homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associations
- Sequence conflicts, variants, etc.



2. TrEMBL is the computer-annotated section of the UniProt Knowledgebase.

It contains translations of all coding regions in the DDBJ/EMBL/GenBank nucleotide databases, and protein sequences extracted from the literature or submitted to UniProtKB, which are not yet integrated into Swiss-Prot.

Steps to improve the data quality

- Automatic annotation
- Redundancy removal
- Evidence attribution

Example: multidrug resistance

PROSITE

consists of documentation entries describing protein domains, families and functional sites as well as associated **patterns** and **profiles** to identify them.

- **Browse by documentation entry**

Pattern

,... a short (not more than four or five residues long) conserved sequence which is part of a region known to be important or which include biologically significant residue(s). We call the pattern(s) created at this stage the 'core' pattern(s)“.

These biologically significant regions or residues are generally:

- Enzyme catalytic sites.
- Prosthetic group attachment sites (heme, pyridoxal-phosphate, biotin, etc).
- Amino acids involved in binding a metal ion.
- Cysteines involved in disulfide bonds.
- Regions involved in binding a molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) or another protein.

PROSITE examples in PROSITE Format:

[AC]-x-V-x(4)-{ED} = [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

<A-x-[ST](2)-x(0,1)-V = N-terminal (<) Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

<{C}*> = all sequences without Cysteine

IIRIFHLRNI = all sequences with the exact motive

Pattern syntax

1. The standard IUPAC one-letter codes for the amino acids are used in PROSITE.
2. The symbol `x' is used for a position where any amino acid is accepted.
3. Ambiguities are indicated by listing the acceptable amino acids for a given position, between square brackets `[]' . For example: [ALT] stands for Ala or Leu or Thr.
4. Ambiguities are also indicated by listing between a pair of curly brackets `{ }' the amino acids that are not accepted at a given position. For example: {AM} stands for any amino acid except Ala and Met.
5. Each element in a pattern is separated from its neighbor by a `-' .
6. Repetition of an element of the pattern can be indicated by following that element with a numerical value or, if it is a gap ('x'), by a numerical range between parentheses.

Examples:

x(3) corresponds to x-x-x

x(2,4) corresponds to x-x or x-x-x or x-x-x-x

A(3) corresponds to A-A-A

Note: You can only use a range with 'x', i.e. A(2,4) is not a valid pattern element.

7. When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a `<' symbol or respectively ends with a `>' symbol. In some rare cases (e.g. PS00267 or PS00539), '>' can also occur inside square brackets for the C-terminal element. 'F-[GSTV]-P-R-L-[G>]' means that either 'F-[GSTV]-P-R-L-G' or 'F-[GSTV]-P-R-L->' are considered.

Profile

„A profile or weight matrix (the two terms are used synonymously here) is a table of position-specific amino acid weights and gap costs. These numbers (also referred to as scores) are used to calculate a similarity score for any alignment between a profile and a sequence, or parts of a profile and a sequence. An alignment with a similarity score higher than or equal to a given cut-off value constitutes a motif occurrence.

Unlike patterns, profiles are usually not confined to small regions with high sequence similarity. Rather they attempt to characterize a protein family or domain over its entire length. Profiles are supposed to be more sensitive and more robust than patterns

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206 D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207 G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208 V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209 I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210 S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211 S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212 C	-4	-7	-6	-7	1															
213 N	-2	0	2	-1																
214 G	-2	-3	-3	-4																
215 D	-5	-5	-2	9	-7	-4	-1	-5	-5	-7	-7	-4	-7	-7	-5	-4	-4	-8	-7	-7
216 S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217 G	-5	-6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218 G	-3																			
219 P	-2																			
220 L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221 N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222 C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223 Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4
224 A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

because they provide discriminatory weights not only for the residues already found at a given position of a motif but also for those not yet found.

Example:

Searching PROSITE (online)

1. Scan a protein for PROSITE matches (SWISS-PROT/TrEMBL accession number (AC), sequence ID (ID), PDB ID, or the sequence in FASTA format)
2. Search SWISS-PROT with a PROSITE entry (PROSITE AC, or the Pattern in PROSITE format).

Mass spectrometry and 2-DE Data

Databases	Tools
MIAPEGelDB • MIAPE document edition • [more]	HCD/CID spectra merger • combine HCD and CID MS/MS spectra • [more]
SWISS-2DPAGE • The SWISS-2DPAGE database assembles data on proteins identified on various 2-D PAGE and SDS-PAGE maps. Each SWISS-2DPAGE entry contains textual data on one protein, including mapping procedures, physiological and pathological information, experimental data and bibliographical references. In addition several 2-D PAGE and SDS-PAGE images are provided, showing the experimentally determined location of the protein, as well as a theoretical region where the protein might be found in the gel. [less]	ImageMaster / Melanie • Melanie offers a unique and flexible interface for the comprehensive visualization, exploration and analysis of 2D gel data. It provides powerful and innovative solutions to shorten the path from data acquisition to protein information, both for conventional 2-DE and DIGE (Fluorescence Difference Gel Electrophoresis) gels. [less]
World-2DPAGE Constellation • set of 2DPAGE resources • [more]	Make2D-DB II • package to build web-based proteomics database • [more]
World-2DPAGE Portal • query get-based proteomics databases • [more]	MALDI PepQuant • quantify MALDI peptides • [more]
World-2DPAGE Repository • gel-based proteomics data • [more]	MSight • mass spectrometry imager • [more]
	plcarver • theoretical distributions of peptide pl • [more]

Tools:

PeptideCutter - Predict potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and/or a table of cleavage site positions. *Protein sequence and identification*

PeptideMass - Cleave a protein sequence with a chosen enzyme, and computes the masses of the generated peptides. The tool also returns theoretical isoelectric point and mass values for the protein of interest. If desired, PeptideMass can return the mass of peptides known to carry post-translational modifications, and can highlight peptides whose masses may be affected by database conflicts, polymorphisms or splice variants. *Protein sequence and identification*

ProtParam - Compute various physical and chemical parameters for a given protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). *Protein characterisation and function*

ProtScale - Compute and represent the profile produced by any amino acid scale on a selected protein sequence. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but ProtScale provides more than 50 predefined scales entered from the literature. *Protein characterisation and function*

Biochemical Pathways - Digitized version of the Roche Applied Science "Biochemical Pathways" wall chart. The map, linked to relevant ENZYME database entries, can be browsed online, and keyword searches are available. *Protein characterisation and function*

- <http://web.expasy.org/pathways/>

The **Sequence Manipulation Suite** is a collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.

- <http://bioinformatics.org/sms2/>

Databases:

STRING - Known and Predicted Protein-Protein Interactions The database contains information from numerous sources, including experimental repositories, computational prediction methods and public text collections. STRING is regularly updated and gives a comprehensive view on protein-protein interactions currently available.

- <http://string-db.org/>

neXtProt is an innovative knowledge platform dedicated to human proteins. This resource contains a wealth of high-quality data on all the human proteins that are produced by the 20'000 protein-coding genes found in the human genome. The content of neXtProt is continuously extended so as to provide many more carefully selected data sets and analysis tools.

- <http://www.nextprot.org/db/>

ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided.

- <http://enzyme.expasy.org/>

Related resources

Nucleic Acids Research

- Molecular Biology Database Collection
<http://nar.oxfordjournals.org/content/40/D1.toc>
- Web Server issue
http://nar.oxfordjournals.org/content/39/suppl_2

MPG Bioinformatics Support Service

<http://www.biochem.mpg.de/en/facilities/ivs/SupportTraining/BioInfoSup/index.html>

Wiki on bioinformatics tools developed in the Max Planck Society

<http://www.biinfowiki.mpg.de/>

©N. Gaedeke

Last update: January 18th, 2012