UniProt

Nociones Básicas de Bioinformática y Genómica

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2015-05-04

UniProt

http://www.uniprot.org/

Protein **sequence** and **functional information**.

- Curated protein sequences
- Consistent/standard nomenclature (including synonyms)
- Cross-referencing (80 $<$ external databases)
- Identification of splice variants
- Identification of variants at amino acid level
- Annotation of proteins and their sequences
- Usage of controlled vocabularies
- Allows for data submission, updates, corrections ...

UniProt Consortium

- European Bioinformatics Institute: EBI
- Swiss Institute of Bioinformatics: SIB
- Protein Information Resource at Georgetown University: PIR

Internal Organization

- UniParc *(UniProt Archive)*: protein sequences from the main, publicly available protein sequence databases.
- UniRef *(UniProt Reference Clusters)*: homology-reduced database; similar sequences merged into clusters.
- UniProtKB *(UniProt Knowledge Base)*: protein sequences with annotation and references.

UniParc

Protein sequence archive

- Stores protein **sequences** from publicly available databases.
- Non-redundant: identical sequences are merged.
- Each sequence is given a stable and unique identifier (UPI)

Just protein sequences. No annotation.

UniRef

UniProt reference clusters.

Three databases:

- UniRef100: clusters of 100% sequence identity proteins. Identical sequences or sequence fragments (form any organism) are combined into a single entry.
- UniRef90: clusters of 90% sequence identity proteins
- UniRef50: clusters of 50% sequence identity proteins

Just protein sequences. No annotation.

UniProtKB

Protein Knowledge Base

UniProtKB/Swiss-Prot

Manually annotated and reviewed entries:

- information extracted from scientific literature
- and curated repositories

UniProtKB/TrEMBL

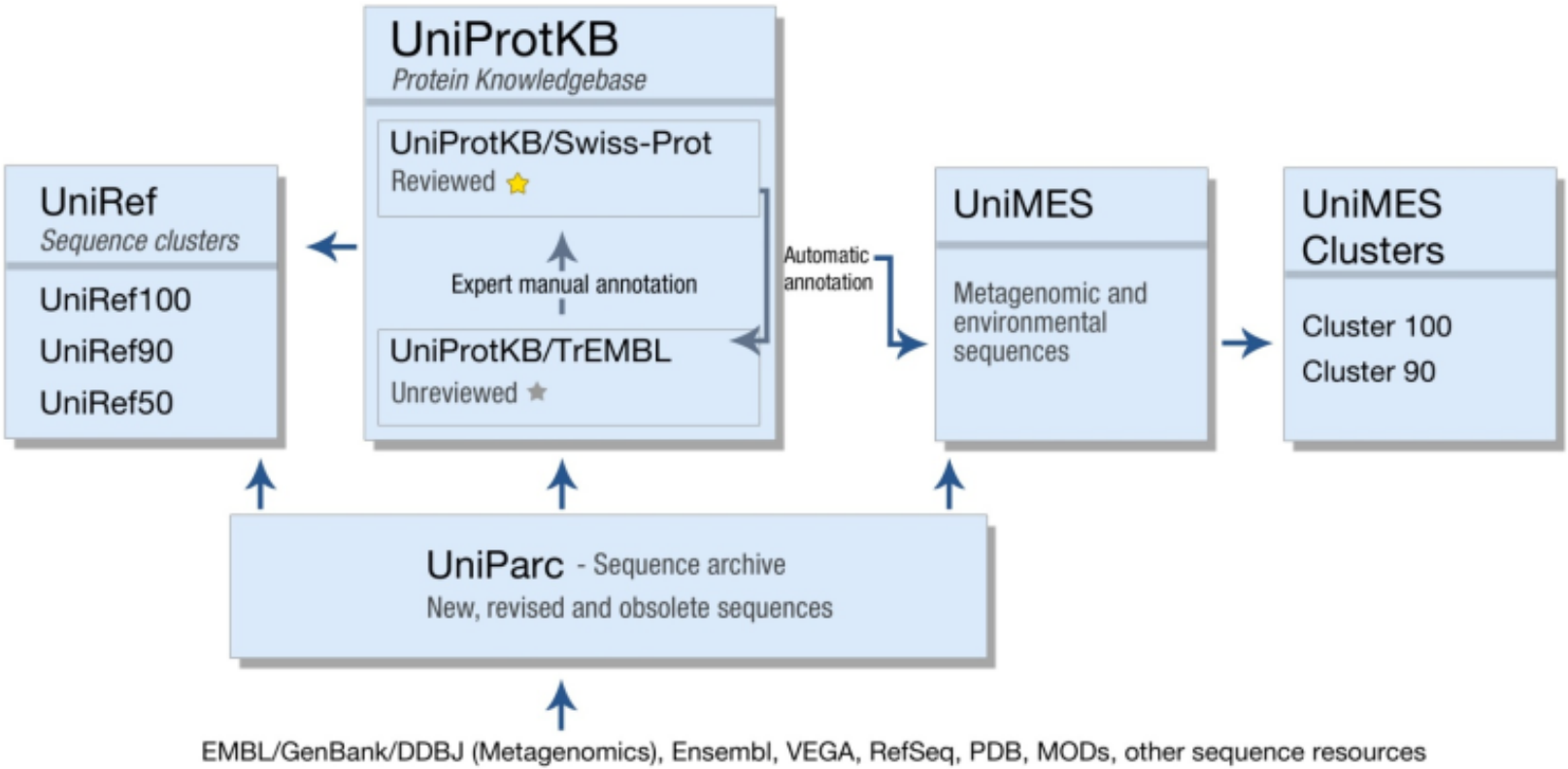
Automatically annotated entries:

- translations of annotated coding sequences (GenBank, EMBL-Bank, DDBJ, ...)
- sequences from PDB
- translations of predicted genes (Ensembl, RefSeq ...)

UniMES

- Specific repository for **metagenomic** and **environmental** data.
- Now deprecated in favor of the EBI Metagenomics portal
  See `http://www.uniprot.org/help/unimes`

Organization Schema



Non Redundancy

Non redundancy is different for the UniProt databases:

- UniProtKB/TrEMBL: one record for 100% identical full-length sequences in **one species**.
- UniProtKB/Swiss-Prot: one record **per gene** in one species.
- UniParc: one record for 100% identical sequences over the entire length, **regardless of the species**.
- UniRef X: one record for X% identical sequences, including fragments, **regardless of the species**.

See further details here.

UniProtKB

- 542,503 reviewed entries UniProtKB/Swiss-Prot)
- 52,707,211 unreviewed entries (UniProtKB/TrEMBL)

- amino acid sequence
- protein name, description, IDs
- taxonomic data
- functional annotation
- citation information

UniProtKB Record

EXAMPLE: http://www.uniprot.org/uniprot/P01116

Names and origin

- Protein names: Recommended / Alternative
- Gene names: includes synonyms.
- Organism: UniProt internal. May be extinct, EXAMPLE
- Taxonomic identifier: NCBI Taxonomy
- Taxonomic lineage: ... Kingdom, Phylumn, Class, ...

UniProtKB Record

Protein attributes

- Sequence length: number of amino acids
- Sequence status: is the *canonical* sequence complete?
  - Complete
  - Fragment
  - Fragment**s**
- Protein existence: see details HERE
  - Evidence at protein level
  - Evidence at transcript level
  - Inferred from homology
  - Predicted
  - Uncertain

UniProtKB Record

General annotation

The "useful" information about the protein:

- Function
- Structure and domains
- Associate ed metabolic *pathways*
- Polymorphism and variants

See all *subsections* of the protein annotation HERE

UniProtKB Record

Ontologies

- Keywords:
  - controlled vocabulary (internal to UniProt)
  - developed according to the **need and content** of UniProtKB/Swiss-Prot
  - attributed manually
  - provide a summary of the entry
  - 10 categories
  - Can be explored at: `http://www.uniprot.org/keywords/`
- Gene Ontology (GO):
  - controlled vocabulary (external to UniProt)
  - describe gene and gene product attributes **in any organism**
  - developed **independently of any existing** database
  - 3 disjoint categories:
    cellular component / molecular function / biological process

UniProtKB Record

Alternative products

Shows the alternative protein sequences that can be generated from the **same gene** by

- alternative promoter usage
- alternative splicing
- alternative initiation
- ribosomal frameshifting

EXAMPLE

UniProtKB Record

Sequence annotation

Describes regions of interest in the protein sequence.

- Chain: amino acid sequence in the mature protein.
- Alternative sequence: isoforms.
- Modified residue: phosphorylation, methylation, acetylation, amidation ...
- Non-terminal residue: the sequence is incomplete.
- ...

See all *subsections* of the sequence annotation HERE
EXAMPLE

UniProtKB Record

Sequences

- default: the canonical protein sequence.
- upon request: isoforms described in the "sequence annotation" section

References

- Citations

Miscellany

- Binary interactions: curated protein-protein interactions form the IntAct database.

Tools

- Search: search in the UniProt Site. Data + documents
- Blast: your sequence against the tree databases
- Align: aligns amino acid changes in fasta format. Accepts UniProt IDs
- Retrieve: quick query to the UniProt database
  - sequence
  - IDs
  - annotation
- ID Mapping:

Biomart

A Biomart repository for UniProt is available at the EBI
http://www.ebi.ac.uk/uniprot/biomart/martview/
Datasets: - **UNIPROT EBI** - ENSEMBL GENE - INTERPRO EBI

FTP

http://www.uniprot.org/downloads

References

- UniProt Helphttp://www.uniprot.org/help/about
- UniProt Manual http://www.uniprot.org/manual/
- UniProt in Wikipedia:
  http://en.wikipedia.org/wiki/UniProt

- Gene Ontology web site http://www.geneontology.org/
- Gene Ontology in Wikipedia
  http://en.wikipedia.org/wiki/Gene_ontology

- Exercises at Denmark University:
  http://wiki.bio.dtu.dk/teaching/index.php/
  Exercise:_The_protein_database_UniProt

Exercise I

Find out all human proteins related with gene **KRAS**

- How many entries are there in UniProt for the gene?
- Which are the differences among them?
- Do the proteins have the same length?
- Explore the different formats in which you can download from the web for this proteins.

Exercise II

Read this article about Enzyme Codes

- Can you use Enzyme Codes ids in the UniProt web site?
- And in the Biomart UniProt Portal?

Exercise III

Find out to which species belongs the protein:

MAFTSGCNHPSFTLPWRTLLPYLVALHLLQLGSAQLTVVAPSLRVTANVGQDVVLRCHLS
PCKDARSSDIRWIQQRSSRLVHHYRNGVDLGQMEEYKGRTELLRDGLSDGNLDLRITAVT
SSDSGSYSCAVQDGDAYAEAVVNLEVSDPFSQIILYWTVALAVIITLLVGSFVVNVFLHR
KKVAQSRELKRKDAELGNCLEKAAALERKDAELAEQAALSKQRDAMLEKHVLKLEEKTDE
VENWNSVLKKDSEEMGYGFAELKKLAAELEKHSEEMGTRDLKLERLAAKLEHQTKELEKQ
HSQFQRHFQNMYLSAGKQKKMVTKLEEHCEWMVRRNVKL

Exercise IV

Go to the FTP site of UniPot and download the reference proteome for **Escherichia coli**

- How many proteins does it have?