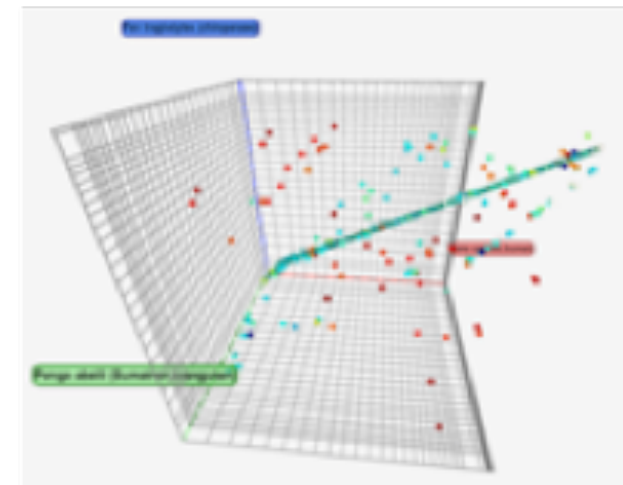
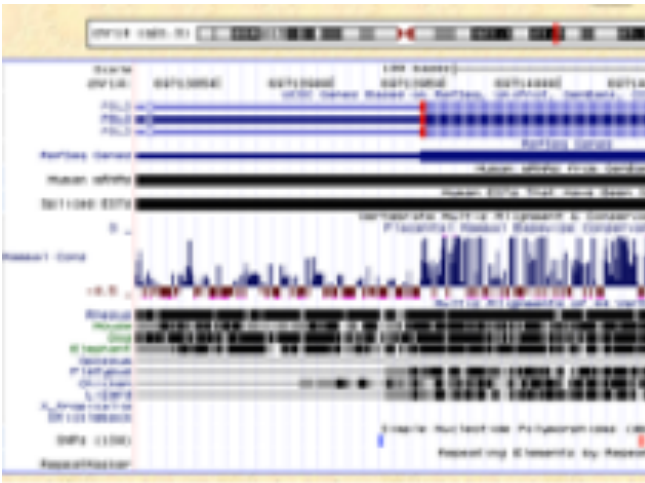


# Computational Genomics

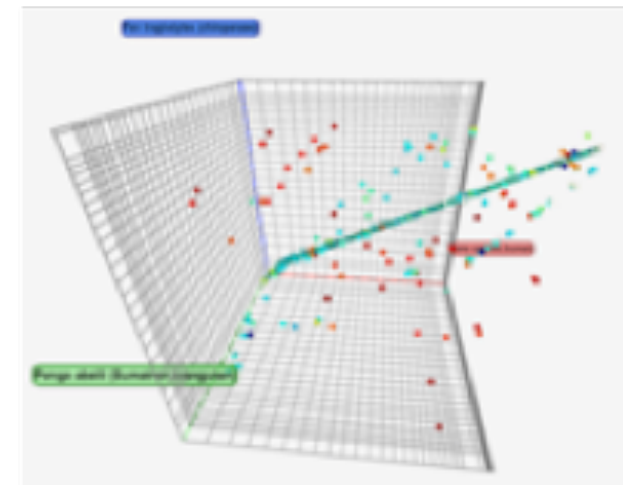
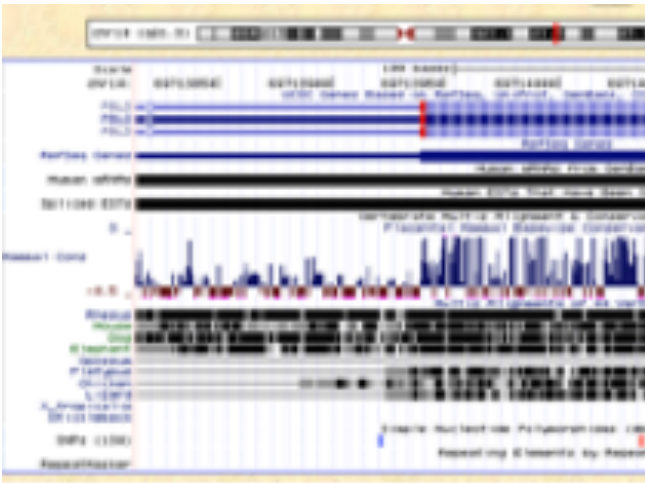
## Introduction To Databases Finding: Genes, and Genomes Proteins and Proteomes



# Computational Genomics

## Introduction To Databases

### Protein Databases



# Protein Databases

## Uniprot

**Proteins**  
UniProt Knowledgebase

**Species**  
Proteomes

**Protein Clusters**  
UniRef

**Sequence Archive**  
UniParc

**Supporting Data**

[Diseases](#) | [Keywords](#) | [Literature Citations](#) | [Taxonomy](#) | [Subcellular locations](#) | [UniRule automatic annotation](#) | [Cross-referenced databases](#) | [ARBA automatic annotation](#)

**Constructive futility**  


Who has not seen mould spread in the corner of a bathroom or on the edges of wallpaper in a damp house? Or winced at the green growth on the edges of a jam jar or on yoghurt left in the open air too long? Moulds are also sure to flourish on...

**Analysis Tools**  


BLAST  
Align  
Search with Lists  
Map IDs  
Search Peptides

**Latest News** [View archive](#)  
Forthcoming changes  
Planned changes for UniProt  
UniProt release 2021\_04  
ZTGC: bacteriophages reinvent the DNA alphabet  
UniProt release 2021\_03  
The importance of being disordered | MobiDB-lite predictions for intrinsically disordered regions [...]  
UniProt release 2021\_02

**Analysis Tools**

**BLAST**  
Search with a sequence to find homologs through pairwise sequence alignment

**Align**  
Align two or more protein sequences with Clustal Omega to find conserved regions

**Search with Lists**  
Map IDs  
Find proteins with lists of UniProt IDs or convert from/to other database IDs

**Search Peptides**  
Search with a peptide sequence to find all UniProt proteins that contain exact matches

**Live webinar**  
Open | 03 Feb 2022 | Online  
Automated annotation in UniProt  
UniProt is a high quality, comprehensive protein resource in which the core activity is the expert review...



- [Tutorial & videos](#)
- [Past webinars](#)
- [Online courses](#)

**Need help?**  
Find answers through our help center or get in touch  
[Attend training](#)  
European Bioinformatics Institute (EBI)  
Protein Information Resource (PIR)  
Swiss Institute of Bioinformatics (SIB)  
[Help center](#) [Contact us](#)  
 

**UniProt data**

**FTP Download**  
  
Download UniProt release data

**Technical Documentation**  
  
Manuals, schemas and ontology descriptions

**Programmatic Access**  
  
Query UniProt data using APIs providing REST, SPARQL and Java services

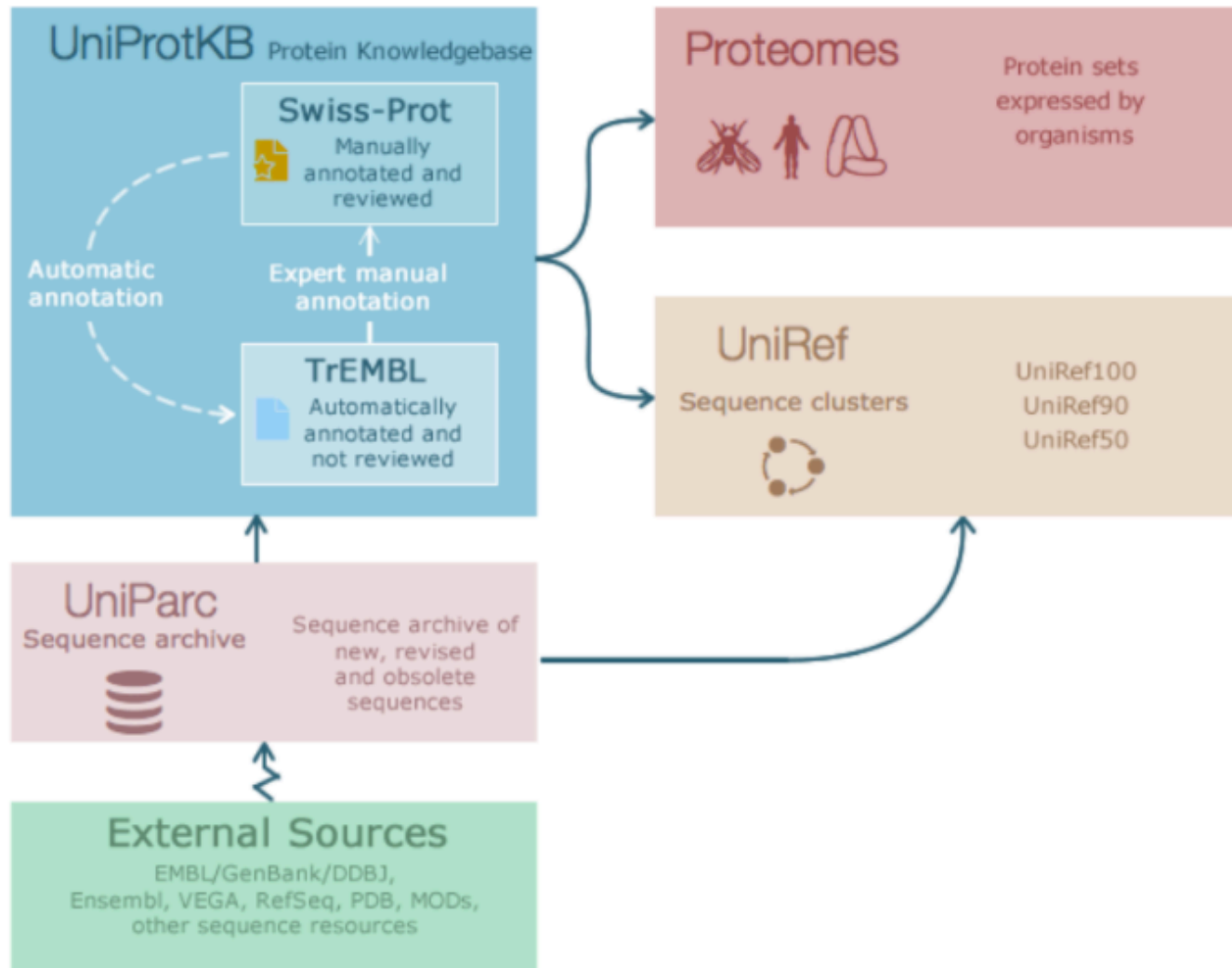
**Submit Data**  
  
Submit your sequences, publications and annotation updates

# Protein Databases

## Uniprot

### What is UniProt?

- The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and functional data (Figure 1).



**Figure 1.** An overview of the databases that comprise UniProt.

# Protein Databases

## Uniprot

### The UniProt databases

- There are three UniProt databases:

The UniProt Knowledgebase (UniProtKB)

The UniProt Reference Clusters (UniRef)

The UniProt Archive (UniParc)

- UniProtKB

UniProtKB is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. It consists of two sections:

- Reviewed (Swiss-Prot) – contains manually annotated records
- Unreviewed (TrEMBL) – contains computationally analysed records

A subset of UniProtKB entries also form the Proteomes dataset. This consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

- UniRef

UniRef provides clustered sets of sequences from UniProtKB (including additional isoforms contained within UniProtKB/Swiss-Prot records) and selected UniParc records. UniRef offers complete coverage of the sequence space at three resolutions:

- UniRef100 database combines identical sequences from any organism into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records.
- UniRef90 is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to the longest sequence (the seed sequence) of the cluster.
- UniRef50 is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to the longest sequence in the cluster.

- UniParc

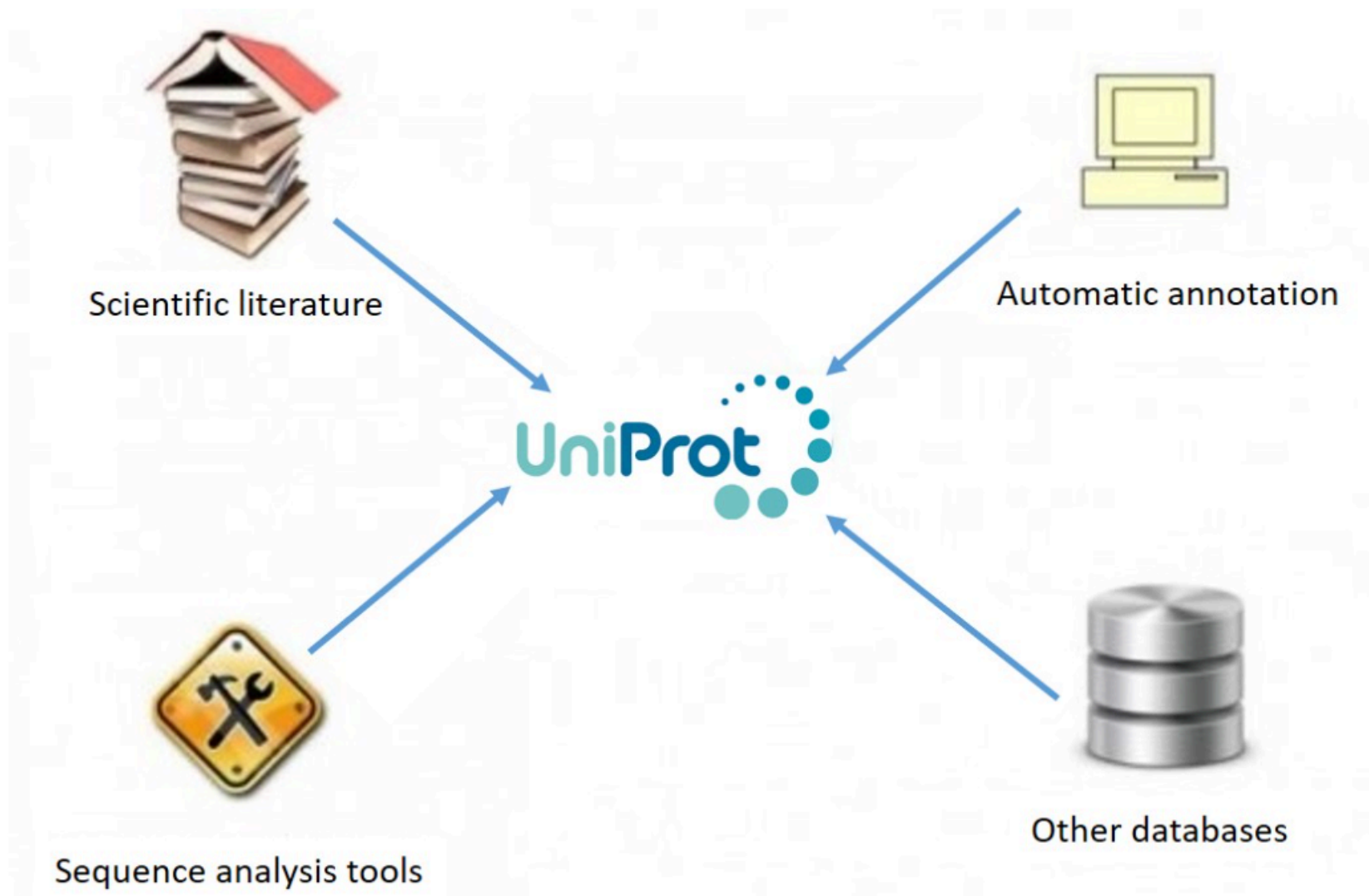
UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. UniParc stores each unique sequence only once, giving it a stable and unique identifier (UPI).

# Protein Databases

## Uniprot

### Where does the data come from?

- UniProt provides both sequence data and associated functional information, derived from a range of sources (Figure 2).



**Figure 2** UniProt data is derived from a number of different sources.



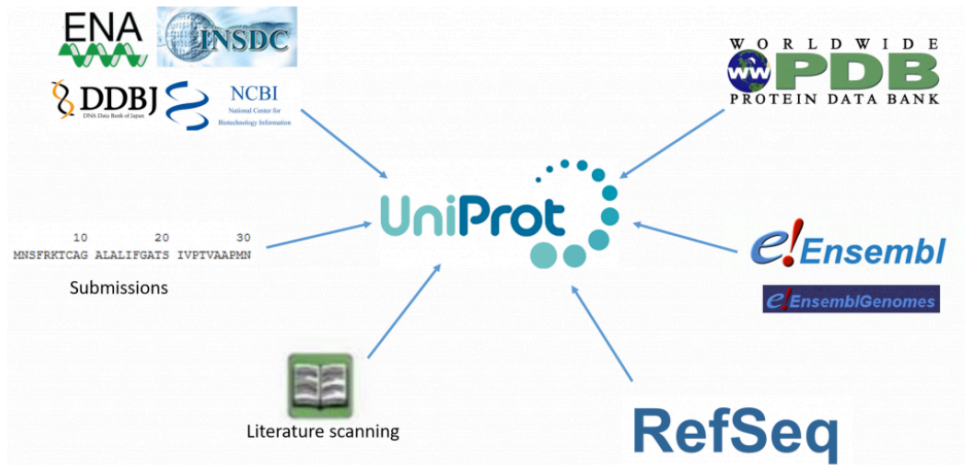
# Protein Databases

## Uniprot Sequence data

- UniProtKB sequences

More than 95% of the protein sequences provided by UniProtKB come from the translations of coding sequences (CDS) submitted to the ENA/GenBank/DDBJ nucleotide sequence resources of the International Nucleotide Sequence Database Collaboration (INSDC).

These CDS are either generated by gene prediction programs or are experimentally proven. The translated CDS sequences are automatically transferred to the TrEMBL section of UniProtKB. The TrEMBL records can be selected for further manual annotation and then integrated into the UniProtKB/Swiss-Prot section.
- In addition to translated CDS, UniProtKB protein sequences may come from:
  - The PDB database
  - Sequences experimentally obtained by direct protein sequencing and submitted to UniProt
  - Sequences scanned from the literature
  - Sequences derived from gene prediction but which have not been submitted to ENA/GenBank/DDBJ. These are imported from resources such as Ensembl and RefSeq
- Importing and combining sequences from a range of sources means that UniProt provides a complete collection of protein sequences and contributes to consistency of protein sets across various sequence resources (Figure 3).



- UniParc sequences

UniParc is designed to capture all publicly available protein sequence data and contains all the protein sequences from the main publicly available protein sequence databases. A complete list of the source databases is available on the UniProt website.
- UniRef sequences

UniRef provides clustered sets of sequences from UniProtKB and selected UniParc records.

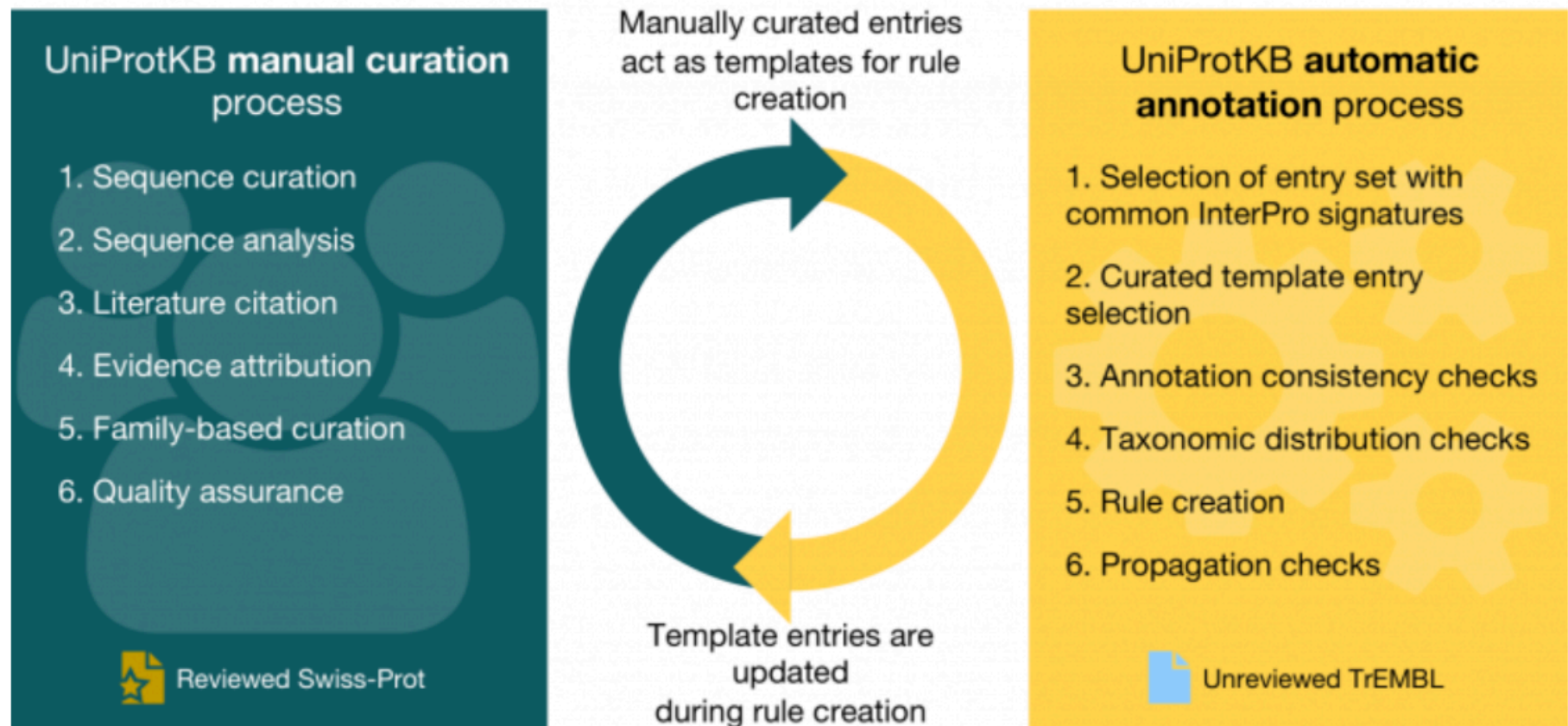
**Figure 3** UniProt imports sequences from a range of sources to ensure that you have access to a complete collection of protein sequences.

# Protein Databases

## Uniprot

### Functional information

Functional information is found in the UniProt Knowledgebase. UniProt attempts to attach as much functional information as possible to each protein sequence to provide users with an overview of the available information for a given protein. This information is added manually by the UniProt biocurators who are all trained biologists or added automatically through various annotation systems which have been developed within the group (Figure 4).



**Figure 4** Relationship between manual curation and automatic annotation in the UniProt Knowledgebase (UniProtKB).



# **Protein Databases**

## **Uniprot**

### **Functional information**

#### **Manual curation**

- Manual curation consists of a critical review of experimental and predicted data for each protein as well as manual verification of each protein sequence.
- Curation methods applied include:
  - manual extraction and structuring of information from the literature
  - manual verification of results from computational analyses
  - mining and integration of large-scale data sets
  - continuous updating as new information becomes available

#### **Automatic annotation**

- UniProt has developed two prediction systems to automatically annotate UniProtKB/TrEMBL in a scalable manner with a high degree of accuracy:
  - UniRule is a collection of manually curated annotation rules which define annotations that can be propagated based on specific conditions
  - The Statistical Automatic Annotation System (SAAS) is an automatic decision-tree based rule-generating system

# Protein Databases

## Uniprot

### Data evidence

#### Indicating data origin

The information in a UniProt Knowledgebase (UniProtKB) record comes from a range of different sources. To make it easy to tell where the data have come from, the origin of each piece of information presented in an entry is provided. UniProt makes use of a subset of evidence codes from the Evidence Code Ontology (ECO) to indicate data origin. These ECO codes are shown directly in the text version (also known as flat file version) of the entries. On the UniProt website, they are transformed into user-friendly, easy to understand labels and evidences that are used in manual assertions are colored gold while those that are used in automatic assertions are colored blue.

1. Experimental data If a piece of information has been experimentally shown in a paper, this will be indicated with the details of the paper used (Figure 5).

**FUNCTION**

**Function<sup>1</sup>**

Together with wdr-48, binds to and stimulates the activity of the deubiquitinating enzyme usp-46, leading to deubiquitination and stabilization of the glr-1 glutamate receptor. 1 Publication

Manual assertion based on experiment in:

"The WD40-repeat proteins WDR-20 and WDR-48 bind and activate the deubiquitinating enzyme USP-46 to promote the abundance of the glutamate receptor GLR-1 in the ventral nerve cord of *Caenorhabditis elegans*."

Dahlberg C.L., Juo P.  
J. Biol. Chem. 289:3444-3456(2014) [PubMed] [Europe PMC]  
[Abstract]

Cited for: FUNCTION, INTERACTION WITH USP-46, TISSUE SPECIFICITY, DISRUPTION PHENOTYPE.

Manual assertions are coloured gold

**Figure 5** Function section of UniProtKB entry D9N129 (*Caenorhabditis elegans* wdr-20) showing the paper from which the data have been extracted.

# Protein Databases

## Uniprot

### Data evidence


#### Indicating data origin

The information in a UniProt Knowledgebase (UniProtKB) record comes from a range of different sources. To make it easy to tell where the data have come from, the origin of each piece of information presented in an entry is provided. UniProt makes use of a subset of evidence codes from the Evidence Code Ontology (ECO) to indicate data origin. These ECO codes are shown directly in the text version (also known as flat file version) of the entries. On the UniProt website, they are transformed into user-friendly, easy to understand labels and evidences that are used in manual assertions are colored gold while those that are used in automatic assertions are colored blue.

#### 2. Data copied from an experimentally characterized protein

For information which has been transferred from a related experimentally characterized protein, the accession number of the characterized protein is provided (Figure 6).

Amino acid modifications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Modified residue <sup>i</sup>	196 – 196	1	N6-carboxyllysine			
			By similarity			
Modified residue <sup>i</sup>	609 – 609	1	Manual assertion inferred from sequence similarity to <sup>i</sup>			
			UniProtKB:P11498 (PYC_HUMAN)			

**Figure 6** UniProtKB entry D3DJ41 (*Hydrogenobacter thermophilus* cfiA) showing the accession number of the entry from which the modified residue has been transferred.

# Protein Databases

## Uniprot

### Data evidence

#### Indicating data origin

The information in a UniProt Knowledgebase (UniProtKB) record comes from a range of different sources. To make it easy to tell where the data have come from, the origin of each piece of information presented in an entry is provided. UniProt makes use of a subset of evidence codes from the Evidence Code Ontology (ECO) to indicate data origin. These ECO codes are shown directly in the text version (also known as flat file version) of the entries. On the UniProt website, they are transformed into user-friendly, easy to understand labels and evidences that are used in manual assertions are colored gold while those that are used in automatic assertions are colored blue.

#### 3. Imported data

If information has been imported from another database, the database name and identifier of the entry from which the information has been imported are provided (Figure 7).

Gene names <sup>i</sup>	Name: <b>Acacb</b> Imported
	Synonyms: Ac
Organism <sup>i</sup>	Mus musculus
Taxonomic Identifier <sup>i</sup>	10090 [NCBI]

Manual assertion inferred from database entries<sup>i</sup>  
MGI:2140940

**Figure 7** UniProtKB entry E9Q4Z2 (mouse Acacb) showing that the gene name has been imported from the Mouse Genome Informatics (MGI) resource.

# Protein Databases

## Uniprot

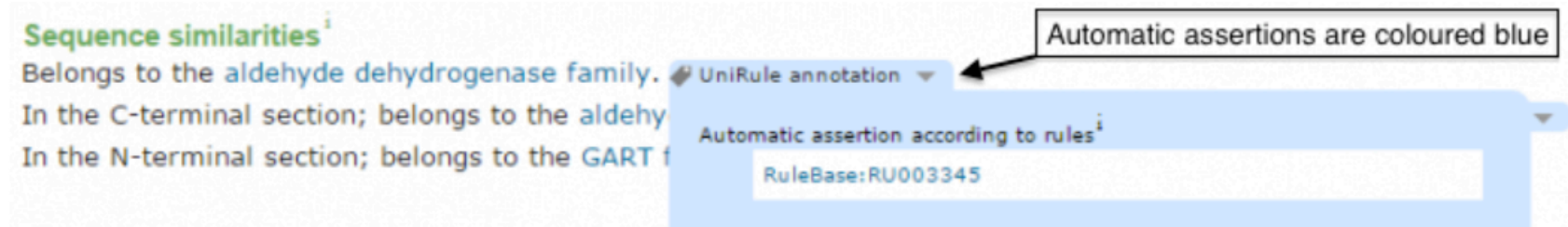
### Data evidence

#### Indicating data origin

The information in a UniProt Knowledgebase (UniProtKB) record comes from a range of different sources. To make it easy to tell where the data have come from, the origin of each piece of information presented in an entry is provided. UniProt makes use of a subset of evidence codes from the Evidence Code Ontology (ECO) to indicate data origin. These ECO codes are shown directly in the text version (also known as flat file version) of the entries. On the UniProt website, they are transformed into user-friendly, easy to understand labels and evidences that are used in manual assertions are colored gold while those that are used in automatic assertions are colored blue.

#### 4. Predicted data

Information which has been predicted by the UniProtKB automatic annotation system or by the sequence analysis programs that are used during the manual curation process are linked to their original source (Figure 8).



**Figure 8** UniProtKB entry Q9VIC9 (*Drosophila melanogaster* CG8665) showing that the protein family information has been predicted by automatic annotation.



# Protein Databases

## Uniprot

### Why do we need UniProt?

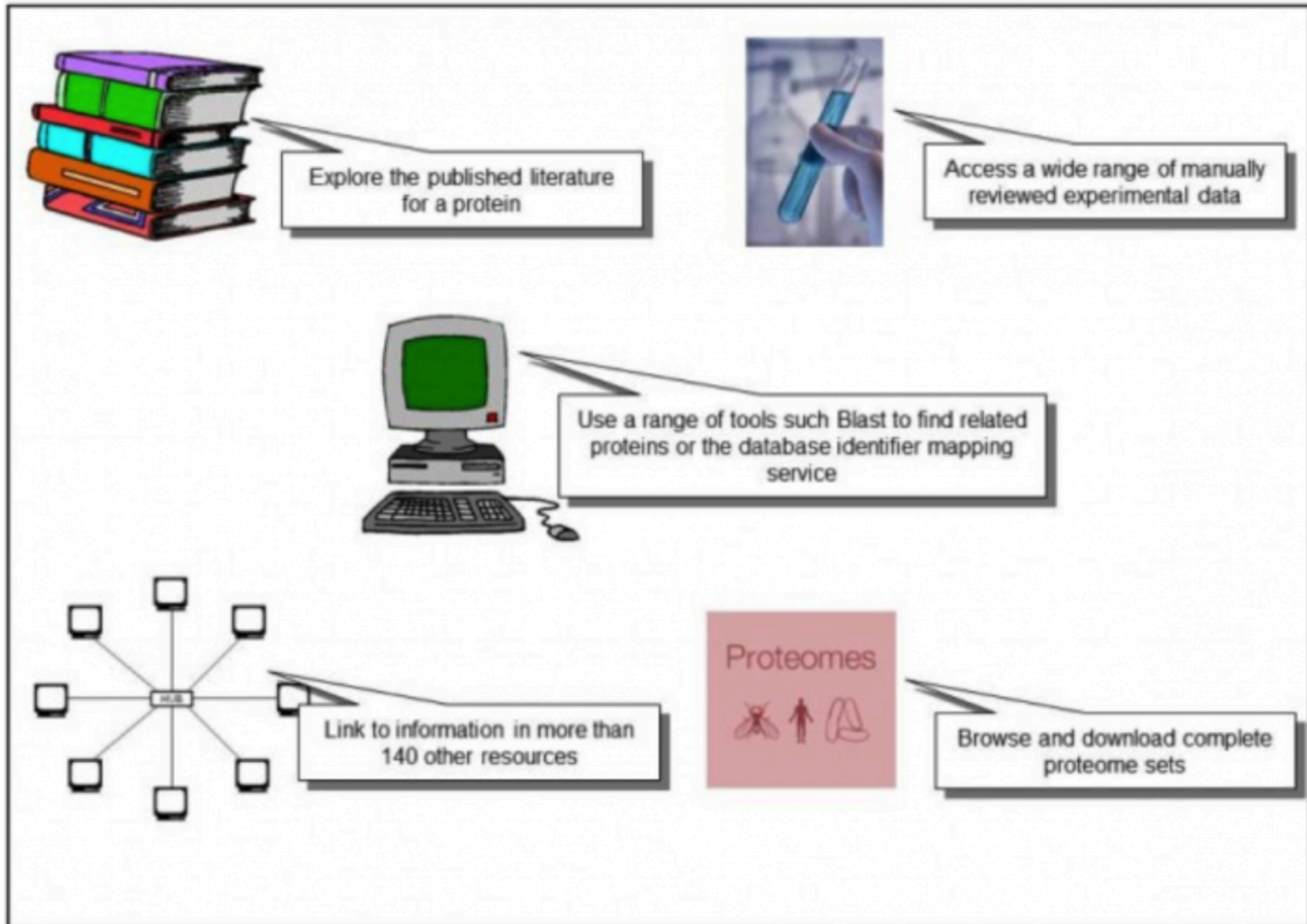
- Understanding protein function is critical to research in many areas of science such as biology, medicine and biotechnology. As the number of completely sequenced genomes continues to increase, huge efforts are being made in the research community to understand as much as possible about the proteins encoded by these genomes. This work is generating large amounts of data which are spread across multiple locations including scientific literature and many biological databases.
- Keeping up with all of this information is a daunting task for most researchers. UniProt helps with this in the following ways:
  - It provides an up-to-date, comprehensive body of protein information at a single site
  - It aids scientific discovery by collecting, interpreting and organizing this information so that it is easy to access and use
  - It saves researchers countless hours of work in monitoring and collecting this information themselves
  - It provides tools to help with protein sequence analysis
  - It provides links to related information in more than 150 other biological databases to help you access additional information in more specialized collections

# Protein Databases

## Uniprot

### When to use UniProt

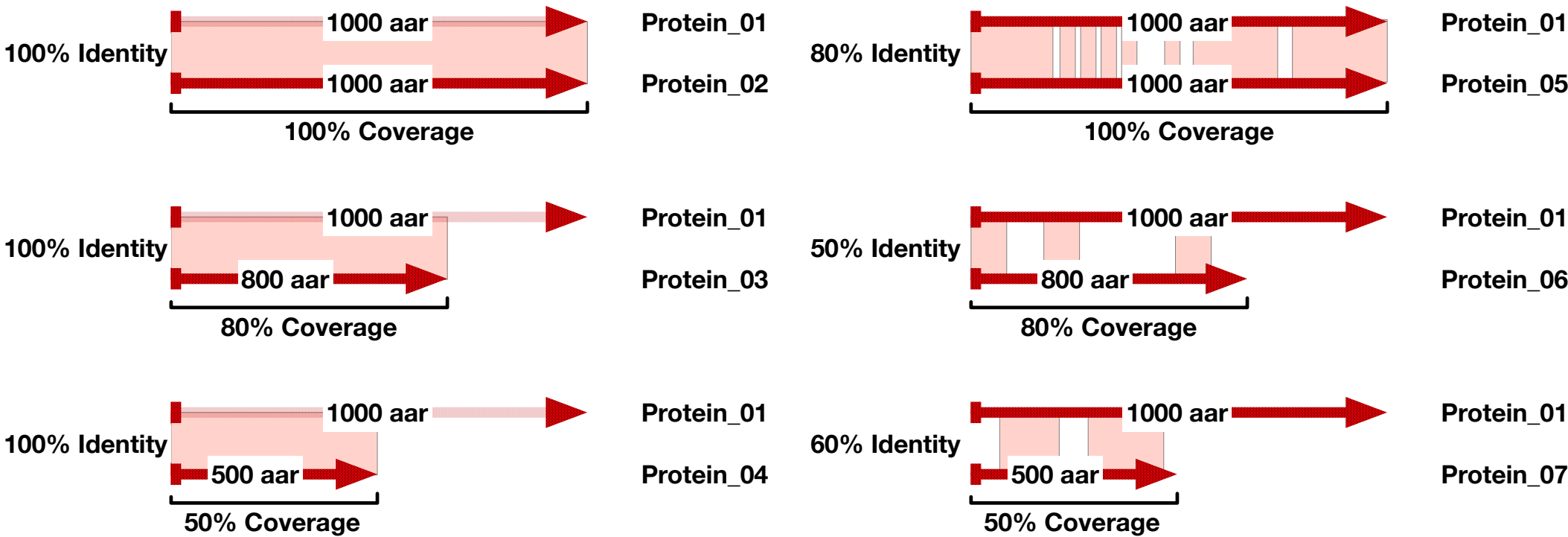
- You can perform many different tasks using UniProt including the following (Figure 9):



# Protein Databases

## Uniprot

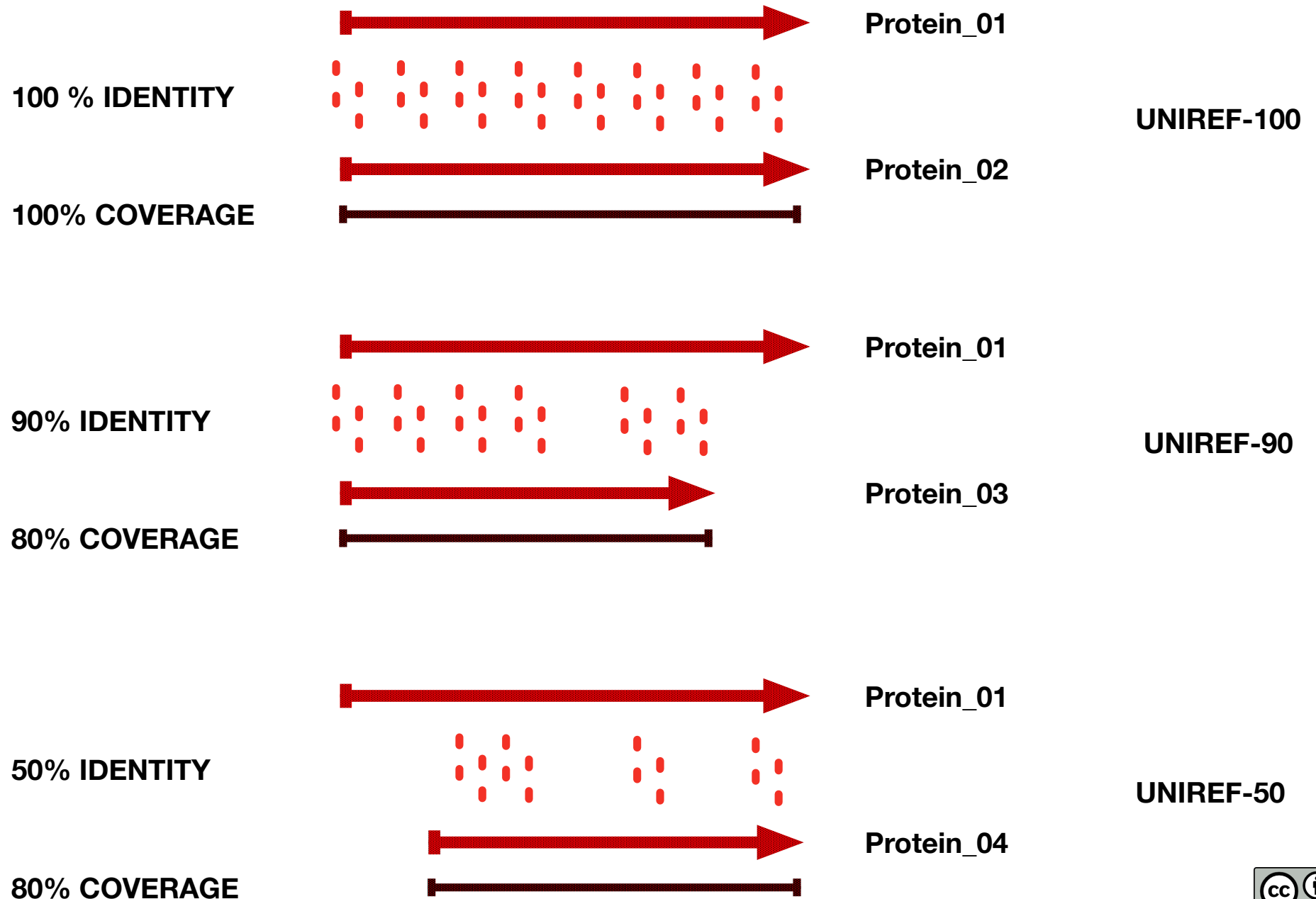
### Identity versus Coverage



# Protein Databases

## Uniprot

### UNIREF Clustering Logic



# Protein Databases

## Uniprot

