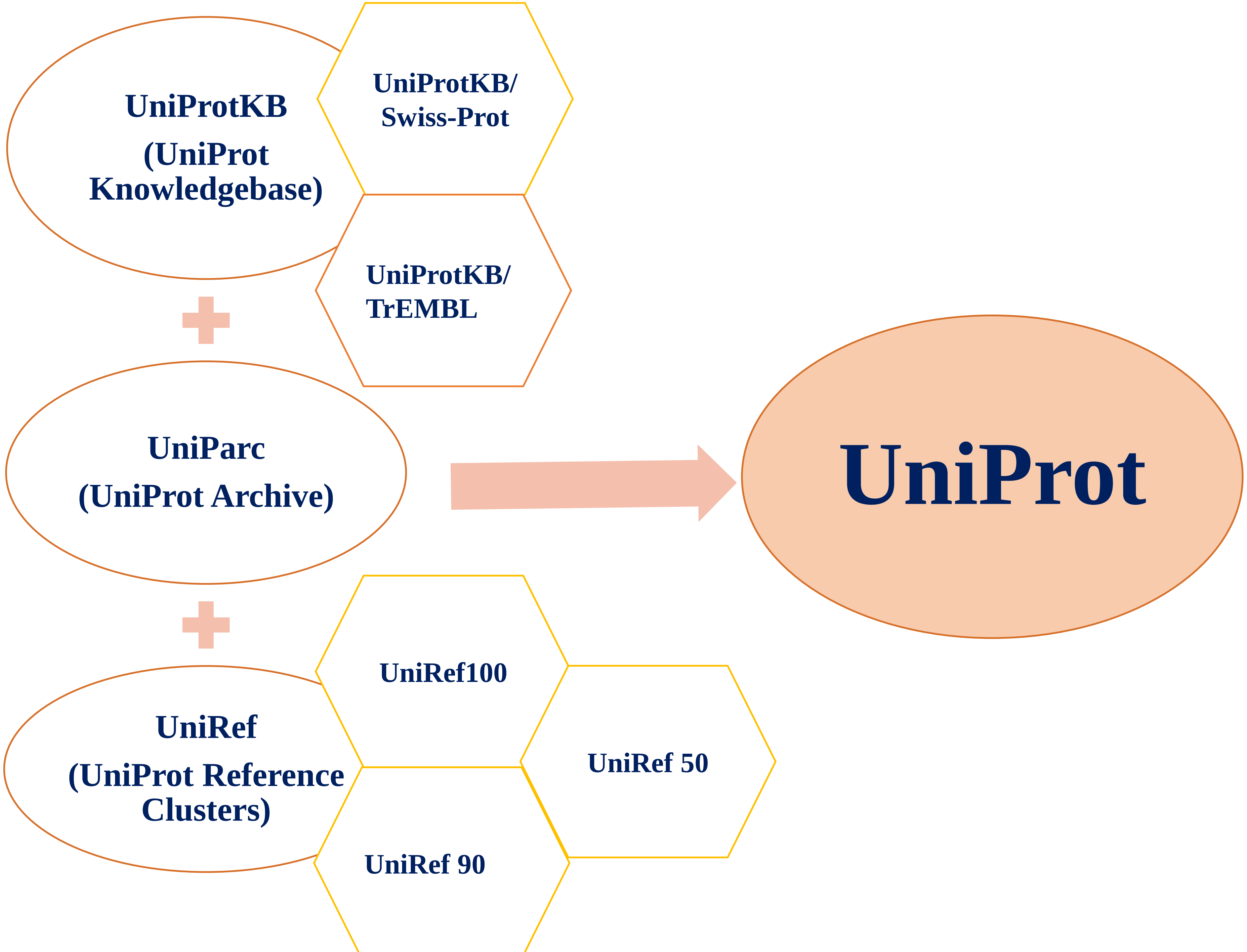
UniPROT

(Universal Protein Resource: An organize and annotated universe of protein sequences http://www.uniprot.org/.)

Dr. Kuldeep Sharma
Assistant Professor
Department of Botany
M.L.S. University, Udaipur
Rajasthan, India

- UniProt is a protein sequence database that was formed through the merger of three separate protein databases:
  - ✓ Swiss-Prot database
  - ✓ TrEMBL (Translated EMBL Nucleotide Sequence Data Library) database
  - ✓ PIR-PSD (Protein Information Resource Protein Sequence Database)
- The resource builds upon the solid foundations laid by the three UniProt Consortium members,
  - the European Bioinformatics Institute (EBI),
  - the Swiss Institute of Bioinformatics (SIB)
  - the Georgetown University’s Protein Information Resource - Protein Sequence Database (PIR-PSD).
- UniProt facilitates scientific discovery by organizing biological knowledge and enabling researchers to rapidly comprehend complex areas of biology.
- Plays an ever more important role by providing a central resource on protein sequences and functional annotation for biologists and for scientists active in functional proteomics and genomics research.

- The broad, long-term objective of UniProt can be summarized as the creation and maintenance of stable, comprehensive and high-quality protein databases, coupled with efficient and unencumbered access mechanisms, to enable rich protein information retrieval and scientific querying across multiple databases containing complementary information.
- The core activities in UniProt include
  - sequence archiving
  - manual curation of protein sequences assisted by automated annotation
  - development of a user-friendly UniProt website
  - and interaction with other protein-related databases for expanded cross-references
- Its accession numbers are a primary mechanism for accurate and sustainable tagging of proteins in informatics applications.
- UniProt is updated every four weeks.
- UniProt comprises three database components, each of which addresses a key need in protein bioinformatics.



UniProtKB (UniProt Knowledgebase)

UniProtKB/ Swiss-Prot

UniProtKB/ TrEMBL

UniParc (UniProt Archive)

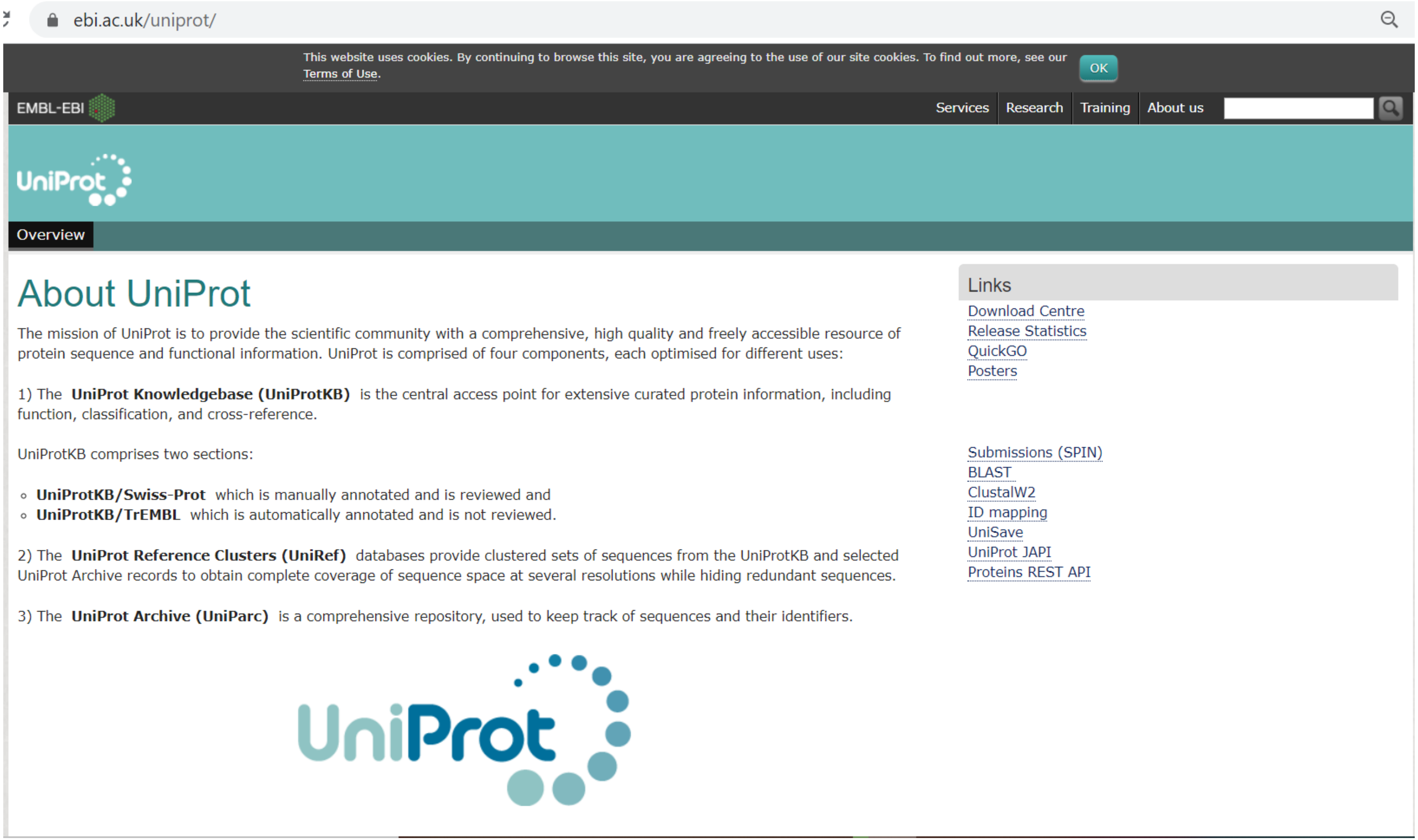
UniRef (UniProt Reference Clusters)

UniRef100

UniRef 90

UniRef 50

UniProt



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

About UniProt

The mission of UniProt is to provide the scientific community with a comprehensive, high quality and freely accessible resource of protein sequence and functional information. UniProt is comprised of four components, each optimised for different uses:

1) The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference.

UniProtKB comprises two sections:

- **UniProtKB/Swiss-Prot** which is manually annotated and is reviewed and
- **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed.

2) The **UniProt Reference Clusters (UniRef)** databases provide clustered sets of sequences from the UniProtKB and selected UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences.

3) The **UniProt Archive (UniParc)** is a comprehensive repository, used to keep track of sequences and their identifiers.

Links

Download Centre
Release Statistics
QuickGO
Posters

Submissions (SPIN)
BLAST
ClustalW2
ID mapping
UniSave
UniProt JAPI
Proteins REST API
]

- The UniProt Knowledgebase (UniProtKB):
  - UniProtKB—a richly annotated protein sequence database with extensive cross references is the centerpiece to UniProt database.
  - The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.
  - The UniProtKB Proteomes portal (https://www.uniprot.org/proteomes/) provides access to proteomes for over 84 thousand (84 387, release 2018 07) species with completely sequenced genomes.
  - The majority of these proteomes are based on the translation of genome sequence submissions to the INSDC source databases—ENA, GenBank and the DDBJ (2).
  - To ensure comprehensiveness, complementary pipelines have been developed to supplement these with genomes sequenced and/or annotated.

- The UniProt Knowledgebase (UniProtKB):
  - The UniProtKB provides an integrated and uniform presentation of protein sequences with extensive annotation and cross-references; annotations are included as:
    - protein name and function
    - taxonomy,
    - Enzyme specific information (catalytic activity, cofactors, metabolic pathway, regulatory mechanisms)
    - domains and sites
    - Posttranslational modifications
    - subcellular locations
    - Tissue specific or developmentally specific expression & interactions
    - splice isoforms
    - polymorphisms
    - diseases and
    - sequence conflicts.

- Literature citations provide evidence for experimental data.
- Entries also connect to various external data collections such as:
  - the underlying DNA sequence entries
  - protein structure databases
  - protein domain and family databases
  - species and function-specific data collections.
- The UniProtKB contains two sections:
  - UniProtKB/ Swiss-Prot: This database contains records with full manual annotation or computer-assisted, manually-verified annotation performed by biologists and based on published literature and sequence analysis.
  - UniProtKB/TrEMBL: This database contains records with computationally generated annotation and large-scale functional characterization.
  - UniProtKB/TrEMBL is a computer-annotated protein sequence database complementing the UniProtKB/Swiss-Prot Protein Knowledgebase. UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProtKB/Swiss-Prot. The database is enriched with automated classification and annotation.
  - Swiss-Prot and TrEMBL continue as two separate sections of the UniProt database.

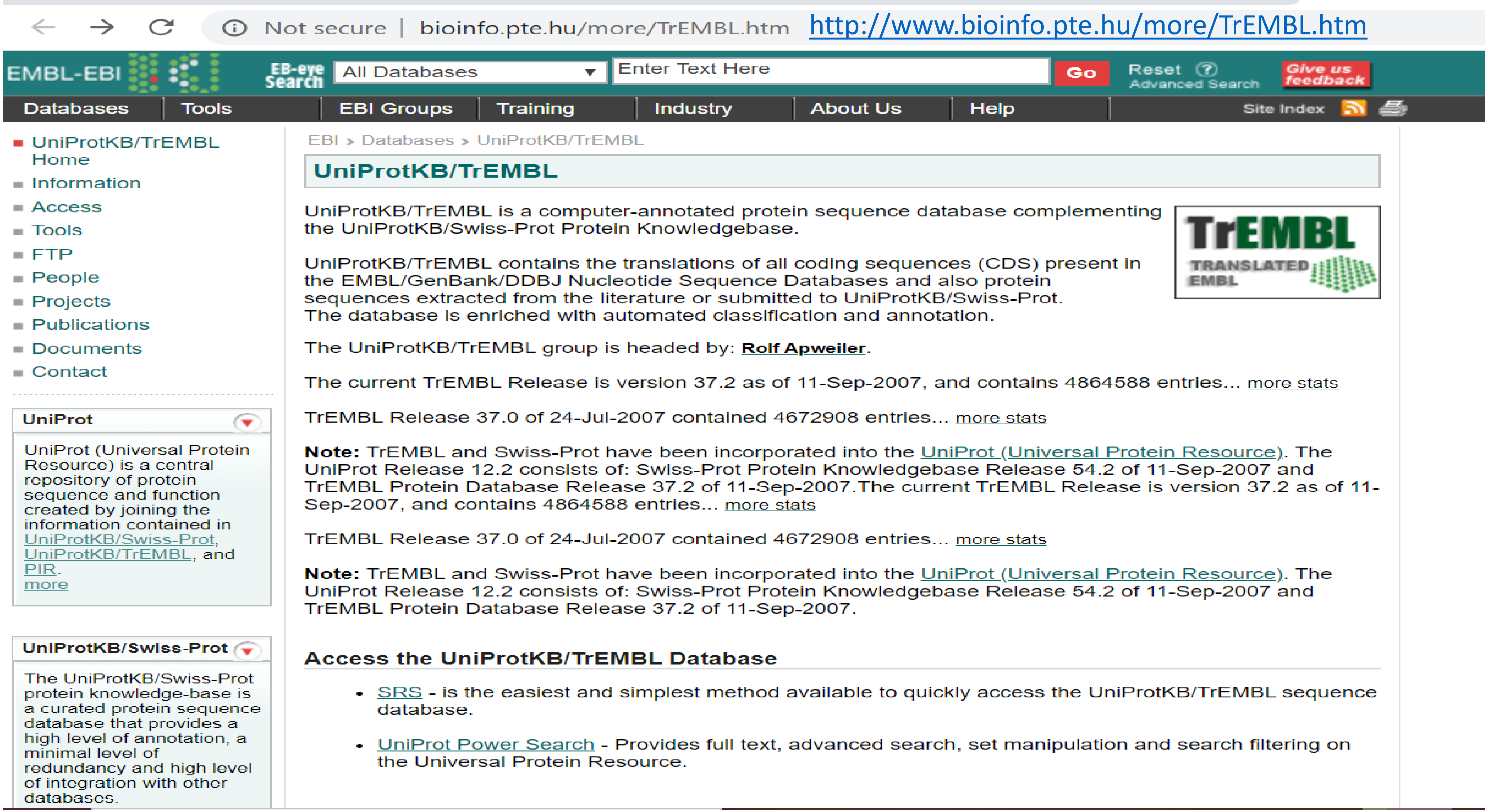
UniProtKB/Swiss-Prot
and
UniProtKB/TrEMBL?

TrEMBL

(Translation of the EMBL database)

- ❑ SWISS-PROT, a curated protein sequence data bank, contains not only sequence data but also annotation relevant to a particular sequence.
- ❑ The annotation added to each entry is done by a team of biologists and comes, primarily, from articles in journals reporting the actual sequencing and sometimes characterisation.
- ❑ Review articles and collaboration with external experts also play a role along with the use of secondary databases like PROSITE and Pfam in addition to a variety of feature prediction methods.
- ❑ TrEMBL consists of entries in a SWISS-PROT format that are derived from the translation of all coding sequences in the EMBL nucleotide sequence database, that are not in SWISS-PROT.
- ❑ Unlike SWISS-PROT entries those in TrEMBL are awaiting manual annotation.

http://www.bioinfo.pte.hu/more/TrEMBL.htm



UniProtKB/TrEMBL

UniProtKB/TrEMBL is a computer-annotated protein sequence database complementing the UniProtKB/Swiss-Prot Protein Knowledgebase.

UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProtKB/Swiss-Prot. The database is enriched with automated classification and annotation.

The UniProtKB/TrEMBL group is headed by: **Rolf Apweiler**.

The current TrEMBL Release is version 37.2 as of 11-Sep-2007, and contains 4864588 entries... more stats

TrEMBL Release 37.0 of 24-Jul-2007 contained 4672908 entries... more stats

Note: TrEMBL and Swiss-Prot have been incorporated into the UniProt (Universal Protein Resource). The UniProt Release 12.2 consists of: Swiss-Prot Protein Knowledgebase Release 54.2 of 11-Sep-2007 and TrEMBL Protein Database Release 37.2 of 11-Sep-2007.The current TrEMBL Release is version 37.2 as of 11-Sep-2007, and contains 4864588 entries... more stats

TrEMBL Release 37.0 of 24-Jul-2007 contained 4672908 entries... more stats

Note: TrEMBL and Swiss-Prot have been incorporated into the UniProt (Universal Protein Resource). The UniProt Release 12.2 consists of: Swiss-Prot Protein Knowledgebase Release 54.2 of 11-Sep-2007 and TrEMBL Protein Database Release 37.2 of 11-Sep-2007.

Access the UniProtKB/TrEMBL Database

- SRS - is the easiest and simplest method available to quickly access the UniProtKB/TrEMBL sequence database.
- UniProt Power Search - Provides full text, advanced search, set manipulation and search filtering on the Universal Protein Resource.

UniProt

UniProt (Universal Protein Resource) is a central repository of protein sequence and function created by joining the information contained in UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and PIR.
more

UniProtKB/Swiss-Prot

The UniProtKB/Swiss-Prot protein knowledge-base is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases.

- The Swiss-Prot component consists of manually annotated protein sequence records that have added information, such as binding sites for drugs. The TrEMBL portion consists of computationally analyzed sequence records that are awaiting full manual annotation; following curation, they are transferred to Swiss-Prot.
- TrEMBL is derived from the CDS translations annotated on records in the INSDC databases, with some additional computational merging and adjustment. Given the very high rate of sequencing, and the effort it takes to do manual annotation.
- The Swiss-Prot component of UniProt is generally much smaller than the TrEMBL component. Because Swiss-Prot's manual annotation provides much additional information, NCBI's protein databases provide links to Swiss-Prot records, even if the sequence is the same as one or more INSDC translations.
- Swiss-Prot (created in 1986) is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. UniProtKB/Swiss-Prot is now the reviewed section of the UniProt Knowledgebase.
- The TrEMBL section of UniProtKB was introduced in 1996 in response to the increased dataflow resulting from genome projects. It was already recognized at that time that the traditional time- and labour-intensive manual curation process which is the hallmark of Swiss-Prot could not be broadened to encompass all available protein sequences. UniProtKB/TrEMBL contains high quality computationally analyzed records that are enriched with automatic annotation and classification. These UniProtKB/TrEMBL unreviewed entries are kept separated from the UniProtKB/Swiss-Prot manually reviewed entries so that the high quality data of the latter is not diluted in any way. Automatic processing of the data enables the records to be made available to the public quickly.

- The Swiss-Prot section of the UniProtKB uses a general purpose naming convention for entry IDs that can be symbolized as X_Y, where X is a mnemonic code of alphanumeric characters representing the protein name and Y is a species identification code of at most five alphanumeric characters representing the biological source of the protein.
- The mnemonic code for the protein name may be elongated up to five characters for UniProtKB/Swiss-Prot entries .
- For TrEMBL entries a biological source indicator to the six-character accession number using the UniProtKB/Swiss-Prot-like format is written.
- Thus, a UniProtKB/Swiss-Prot entry can be distinguished from a UniProtKB/TrEMBL entry by the number of characters preceding the underscore (six for the latter, up to five for the former).
- The IDtracker tool (http://www.expasy.org/cgi-bin/idtracker) is available to trace protein entry identifiers.

- UniProtKB data has also been divided into different taxonomic divisions for:
  - Archaea
  - Bacteria
  - Vertebrates
  - Mammals
  - Humans
  - Plants
  - Rodents
  - Invertebrates,
  - Fungi
  - Viruses and
  - Unclassified (/uniprot/current_release/knowledgebase/taxonomic_divisions/).
- Furthermore, complete proteomes from >200 organisms are also available for download (/uniprot/current_release/knowledgebase/complete_proteomes/).

- Furthermore, complete proteomes from >200 organisms are also available for download (/uniprot/current_release/knowledgebase/complete_proteomes/).
- Besides, a redundancy removal process was introduced in 2015. This process identifies and removes almost identical proteomes of the same species before their inclusion in UniProtKB (https://www.uniprot.org/help/proteome redundancy) and places their sequences in UniParc.
- Currently this process has removed ~38% all complete proteomes (~241 million proteins) from UniProtKB.
- The redundancy reduction both greatly reduced the size of UniProtKB as well as made its growth more scalable.
- Key Characteristics of UniProt versus GenBank and RefSeq:

UniProt	**GenBank and RefSeq**
Produced by SIB, EBI & Georgetown University	Produced by INSDC and NCBI
Protein data only	Protein and nucleotide data
Curated in Swiss-Prot, not in TrEMBL	Curated in RefSeq, not in GenBank
UniProt	GenBank and RefSeq

- (UniParc): The UniProt Archive
  - This database is the main sequence storehouse.
  - It houses all new and revised protein sequences from the various sources (UniProt, GenPept, RefSeq, etc.) to ensure that comprehensive coverage is available at a single site.
  - It is a comprehensive set of all known sequences indexed by their unique sequence checksums and currently contains over 70 million sequences entries.
  - To avoid redundancy, each unique sequence is assigned a unique identifier and is stored only once.
  - The basic information stored with each UniParc entry is the identifier, the sequence, cyclic redundancy check number (CRC64), source database(s) with accession and version numbers, and a time stamp.
  - In addition, each source database accession number is tagged with its status in that database, indicating if the sequence still exists or has been deleted at that source.
  - The archive thus provides a history of protein sequences.

- UniRef: The UniProt Reference Cluster
  - It cluster sequence sets at various levels of sequence identity
  - UniProt also makes available three sets of sequences that have been made non-redundant at various levels of sequences identity regardless of source organisms:
    - UniRef100- UniProt Reference Clusters: 100% identity
    - UniRef90- UniProt Reference Clusters: 90% identity
    - and UniRef50- UniProt Reference Clusters: 50% identity
  - It has cross references to over 150 databases and acts as a central hub to organize protein information.
  - The UniProt Reference Clusters (UniRef) condense sequence information and annotation to facilitate both sequence similarity searches and analyses of the results.
  - Reduction of sequence redundancy speeds sequence similarity searches while rendering such searches more informative.
  - As a result, UniProtKB acts as a central hub connecting biomolecular information archived in more than 100 cross-referenced databases.

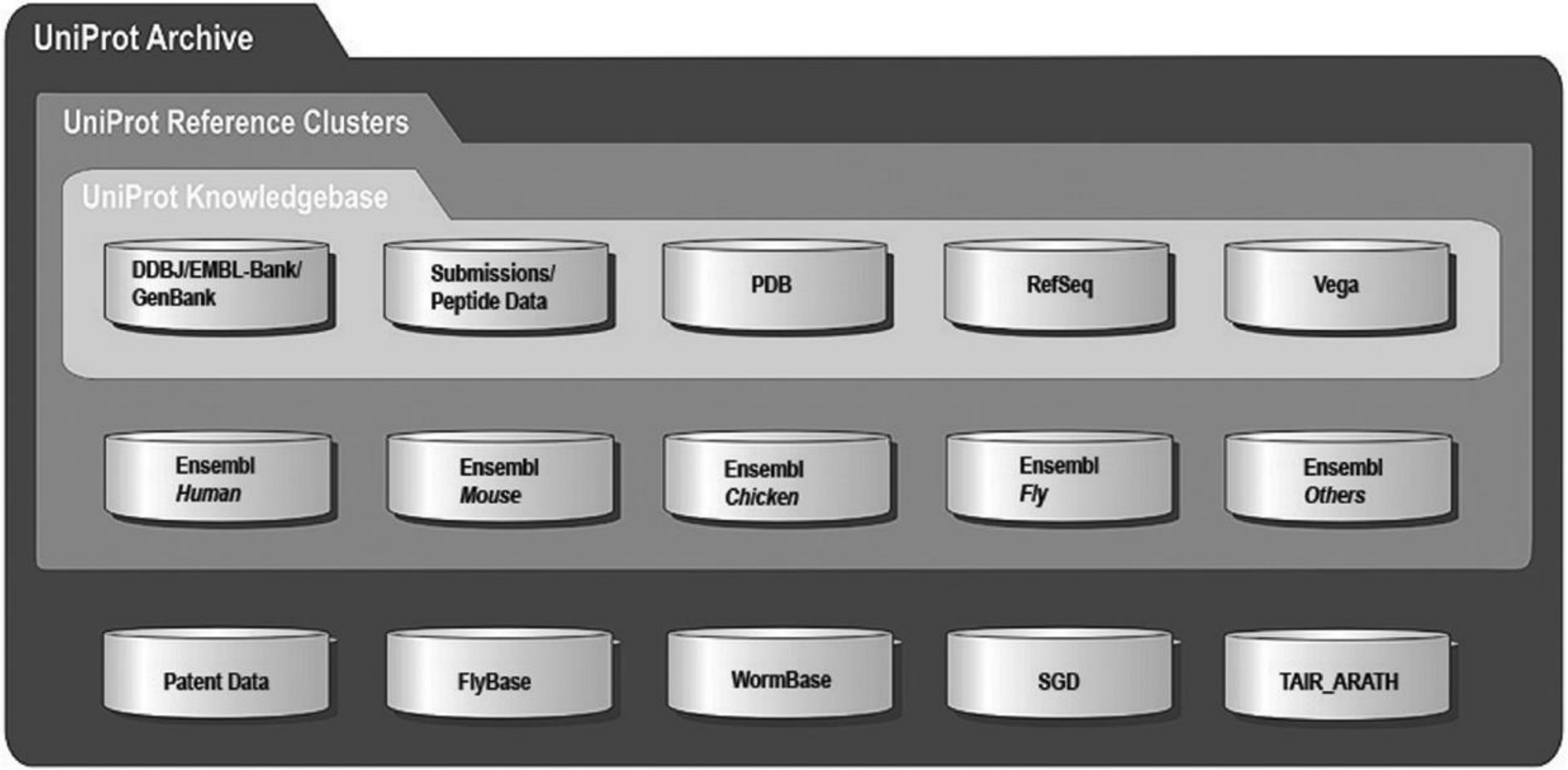


Figure : Overview of the major data sources of the UniProt databases.

Thank You