UniProt

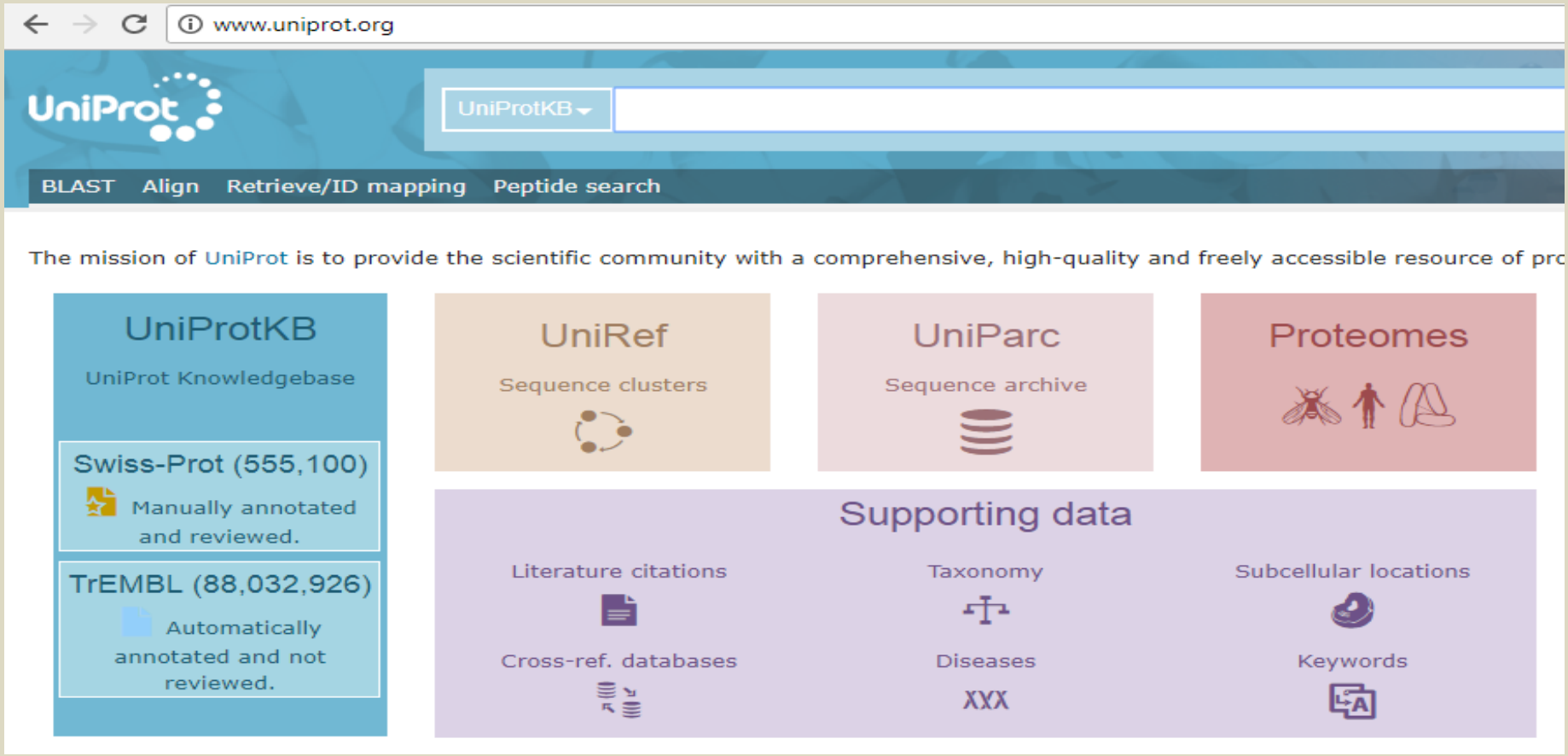
Unit-2

Paper- Bioinformatics (DSE-1)

B.Sc. (H) Microbiology V Sem

UniProt

It provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information

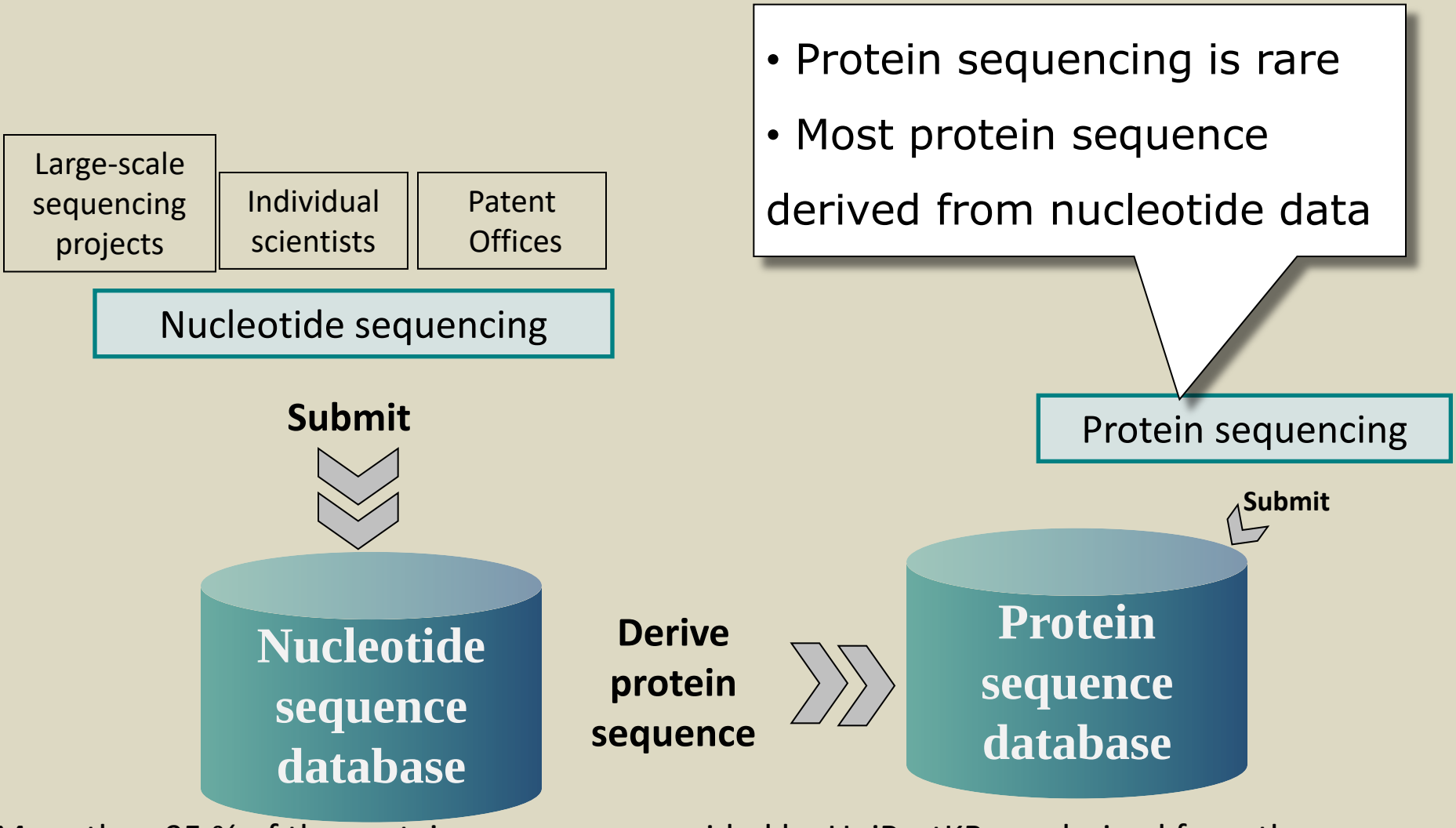


UniProt - **Universal protein resource**

Swiss-Prot & TrEMBL PIR-PSD

- Merger of these three databases, since 2002
- Funded mainly by NIH (US) to be the highest quality, most thoroughly annotated protein sequence database
- UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR).
- Collaboration through different tasks such as database curation, software development and support.

Source of protein sequence data



More than 95 % of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS)

Protein sequence is mainly derived data

submit

DNA sequence

ACGCTCGTACGCATCGTCACTACTAGCTACGACGACGACACGCTACTACTCGACGATTCT

may not have direct evidence

Predicted start

Predicted splice sites

ranscribe

Predicted stop

Derived mRNA sequence

AUGCGUAGUGAUGAAUGCUGCUGUGCGAUGAGCUGC

translate

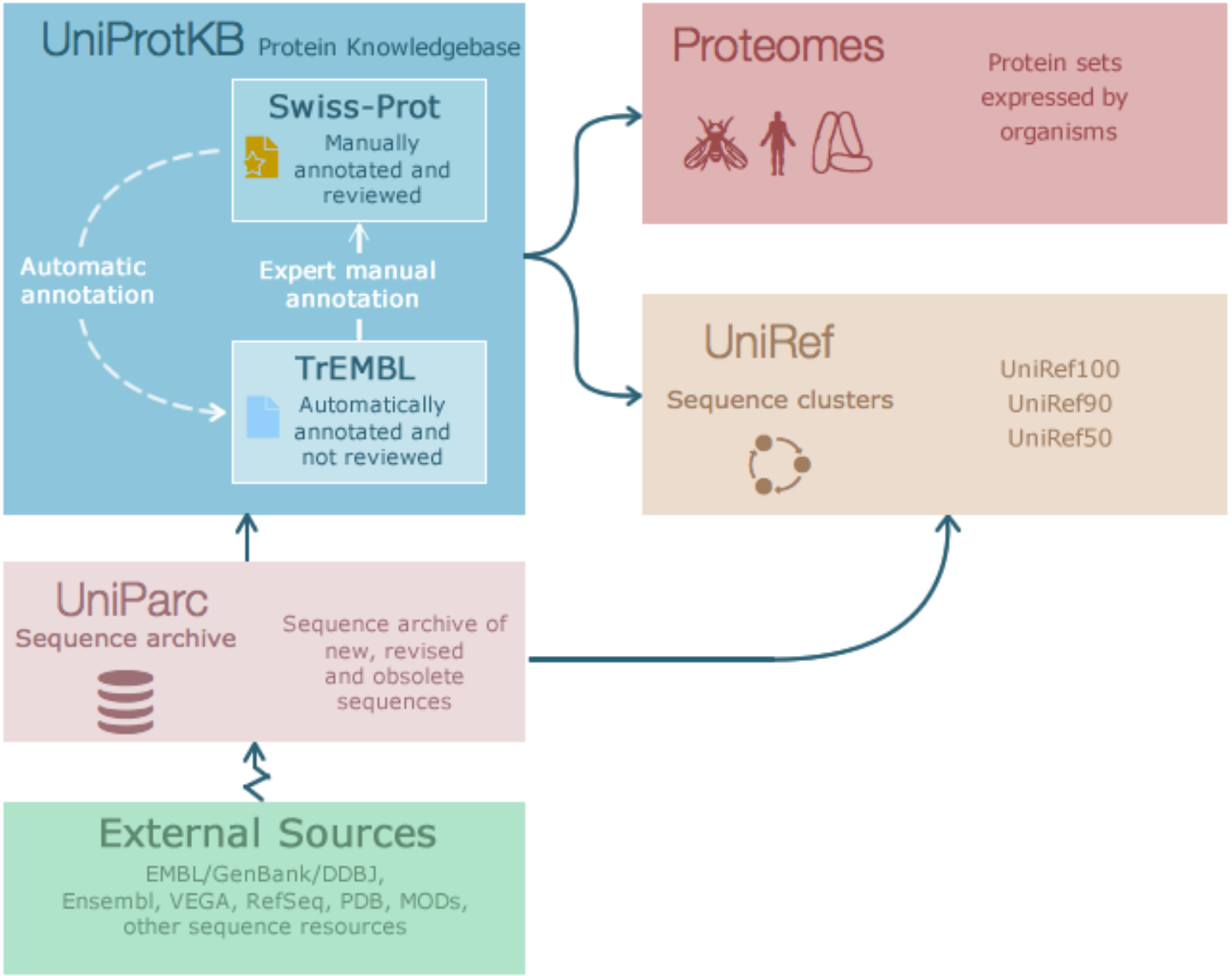
Derived protein sequence

MRSNECCCAMSC

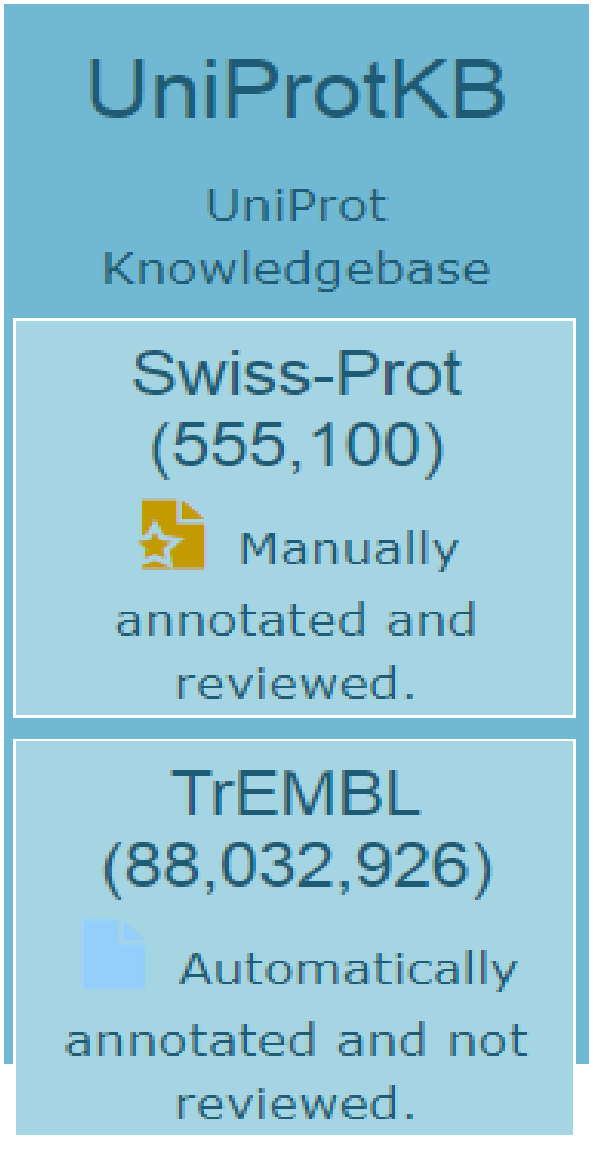
UniProt

- The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data.
- The UniProt databases are:

1. UniProt Knowledgebase (UniProtKB),
2. UniProt Reference Clusters (UniRef),
3. UniProt Archive (UniParc).


UniProtKB
Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
Proteomes
Protein sets expressed by organisms
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ,
Ensembl, VEGA, RefSeq, PDB, MODs,
other sequence resources

UniProtKB


UniProtKB
UniProt Knowledgebase
Swiss-Prot (555,100)
Manually annotated and reviewed.
TrEMBL (88,032,926)
Automatically annotated and not reviewed.

UniProtKB

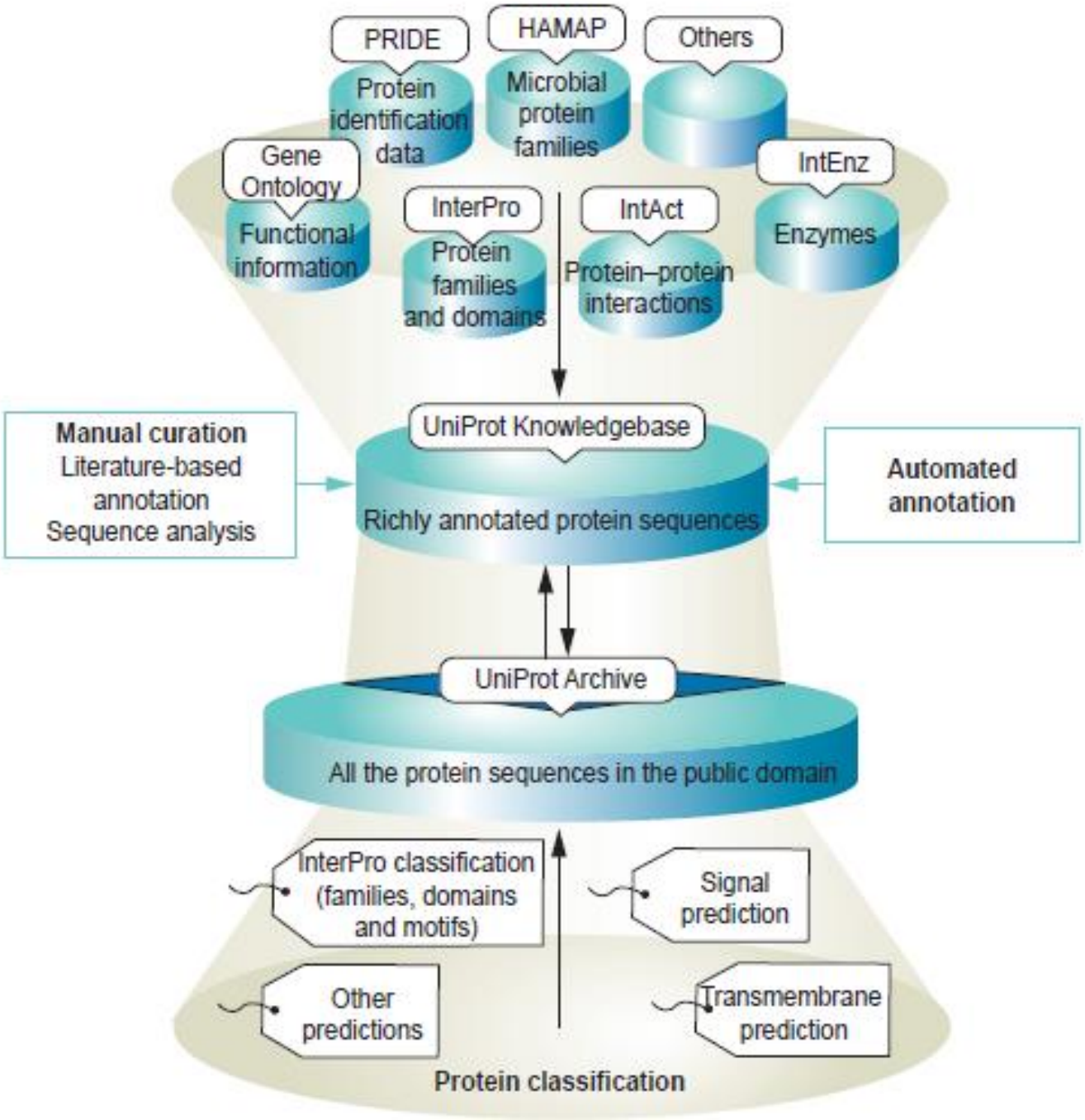
- **Reviewed (Swiss-Prot) - Manually annotated**
  Records with information extracted from literature and curator-evaluated computational analysis and and scientific conclusions.
- Is a high quality non-redundant protein sequence database
- **Unreviewed (TrEMBL) - Computationally analyzed**
  Records that await full manual annotation.
- contains protein sequences associated with computationally generated annotation and large-scale functional characterization

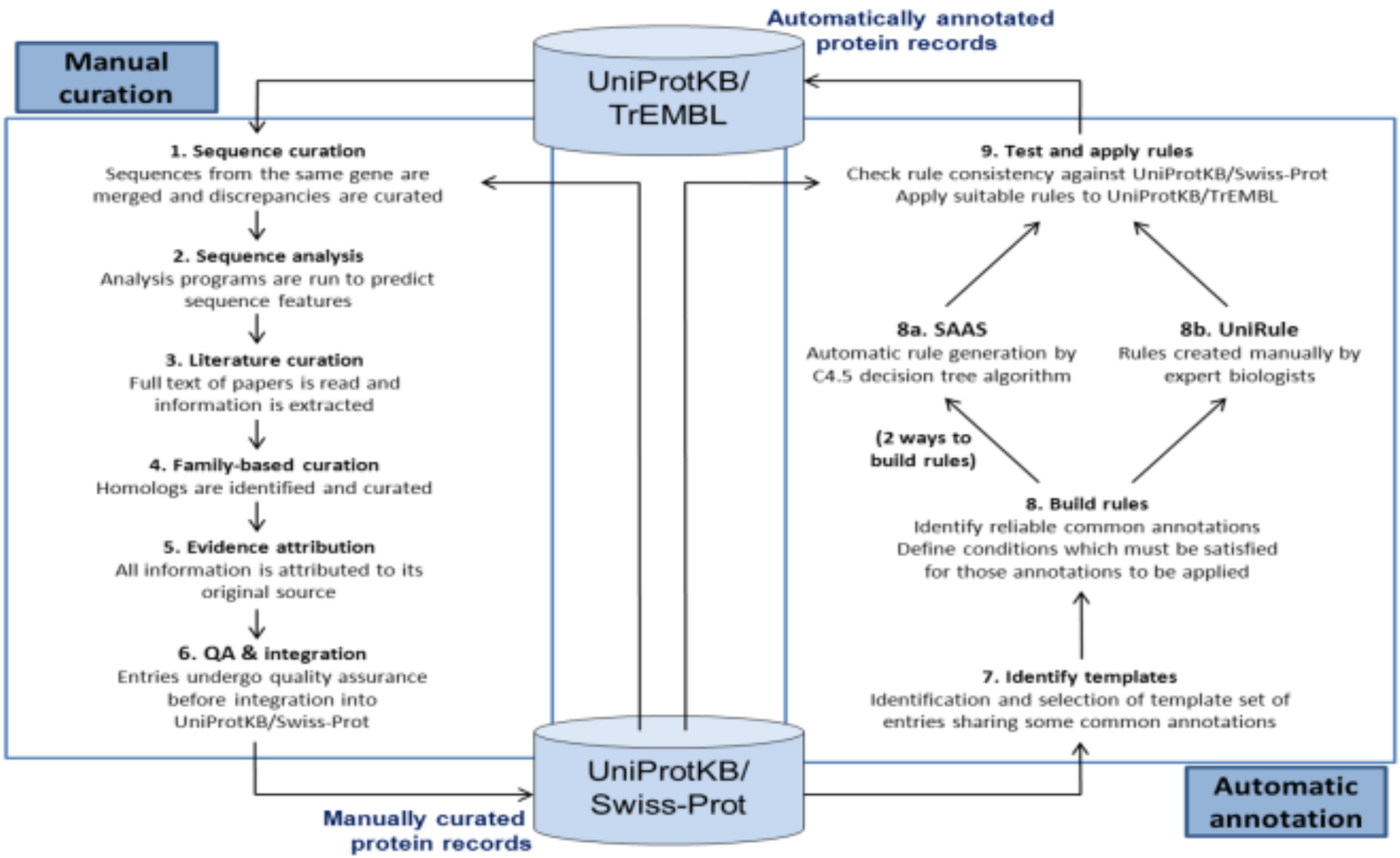
UniProtKB and its functions

- The UniProt Knowledgebase is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.
- It captures the core data mandatory for each UniProtKB entry mainly,
  - the amino acid sequence,
  - protein name or description
  - taxonomic data
  - citation information
- It also adds annotation information as much as possible.

UniProtKB and its functions

- Annotation includes:
- widely accepted biological ontologies, classifications
- cross-references
- It also gives clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.


PRIDE
Protein identification data
HAMAP
Microbial protein families
Others
Gene Ontology
Functional information
InterPro
Protein families and domains
IntAct
Protein–protein interactions
IntEnz
Enzymes
Manual curation
Literature-based annotation
Sequence analysis
UniProt Knowledgebase
Richly annotated protein sequences
Automated annotation
UniProt Archive
All the protein sequences in the public domain
InterPro classification (families, domains and motifs)
Signal prediction
Other predictions
Transmembrane prediction
Protein classification


Automatically annotated protein records
Manual curation
UniProtKB/ TrEMBL
1. Sequence curation
Sequences from the same gene are merged and discrepancies are curated
2. Sequence analysis
Analysis programs are run to predict sequence features
3. Literature curation
Full text of papers is read and information is extracted
4. Family-based curation
Homologs are identified and curated
5. Evidence attribution
All information is attributed to its original source
6. QA & integration
Entries undergo quality assurance before integration into UniProtKB/Swiss-Prot
9. Test and apply rules
Check rule consistency against UniProtKB/Swiss-Prot
Apply suitable rules to UniProtKB/TrEMBL
8a. SAAS
Automatic rule generation by C4.5 decision tree algorithm
8b. UniRule
Rules created manually by expert biologists
(2 ways to build rules)
8. Build rules
Identify reliable common annotations
Define conditions which must be satisfied for those annotations to be applied
7. Identify templates
Identification and selection of template set of entries sharing some common annotations
UniProtKB/ Swiss-Prot
Manually curated protein records
Automatic annotation

UniProt automatic annotation

- UniProt has developed two complementary approaches to automatically annotate protein sequences with a high degree of accuracy.

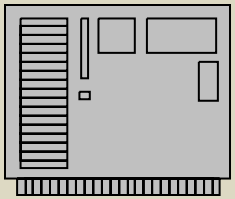
1. **UniRule** is a collection of manually curated annotation rules which define annotations that can be propagated based on specific conditions
2. **Statistical Automatic Annotation System (SAAS)** is an automatic decision-tree based rule-generating system.

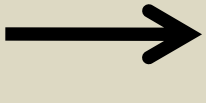
- The central components of these approaches are rules based on InterPro classification and the manually curated data in UniProtKB/Swiss-Prot.
- Predictions of sequence features such as Signal, Transmembrane and Coil regions are generated using software from external providers

- UniProt uses InterPro to **classify sequences at superfamily, family and subfamily levels** and to predict the occurrence of functional domains and important sites. InterPro integrates predictive models of protein function, so-called **'signatures'**, from a number of member databases.
- InterPro matches are automatically annotated to UniProtKB entries as database cross-references with every InterPro release.
- In UniProtKB/TrEMBL entries, domains from the InterPro member databases **PROSITE, SMART or Pfam** are predicted and annotated automatically, and their evidence/source labels indicate "InterPro annotation".

Automatic annotation

UniProtKB uses 2 prediction programs:





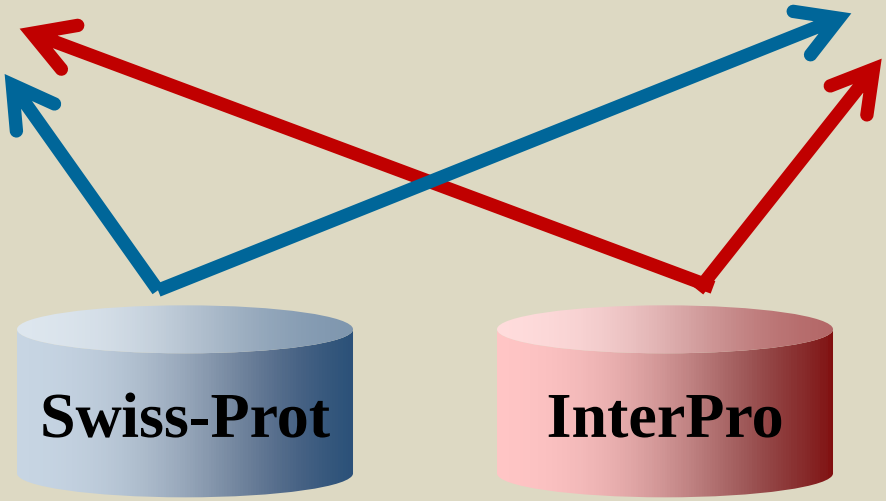
SAAS:

generates a set of decision trees using data mining.

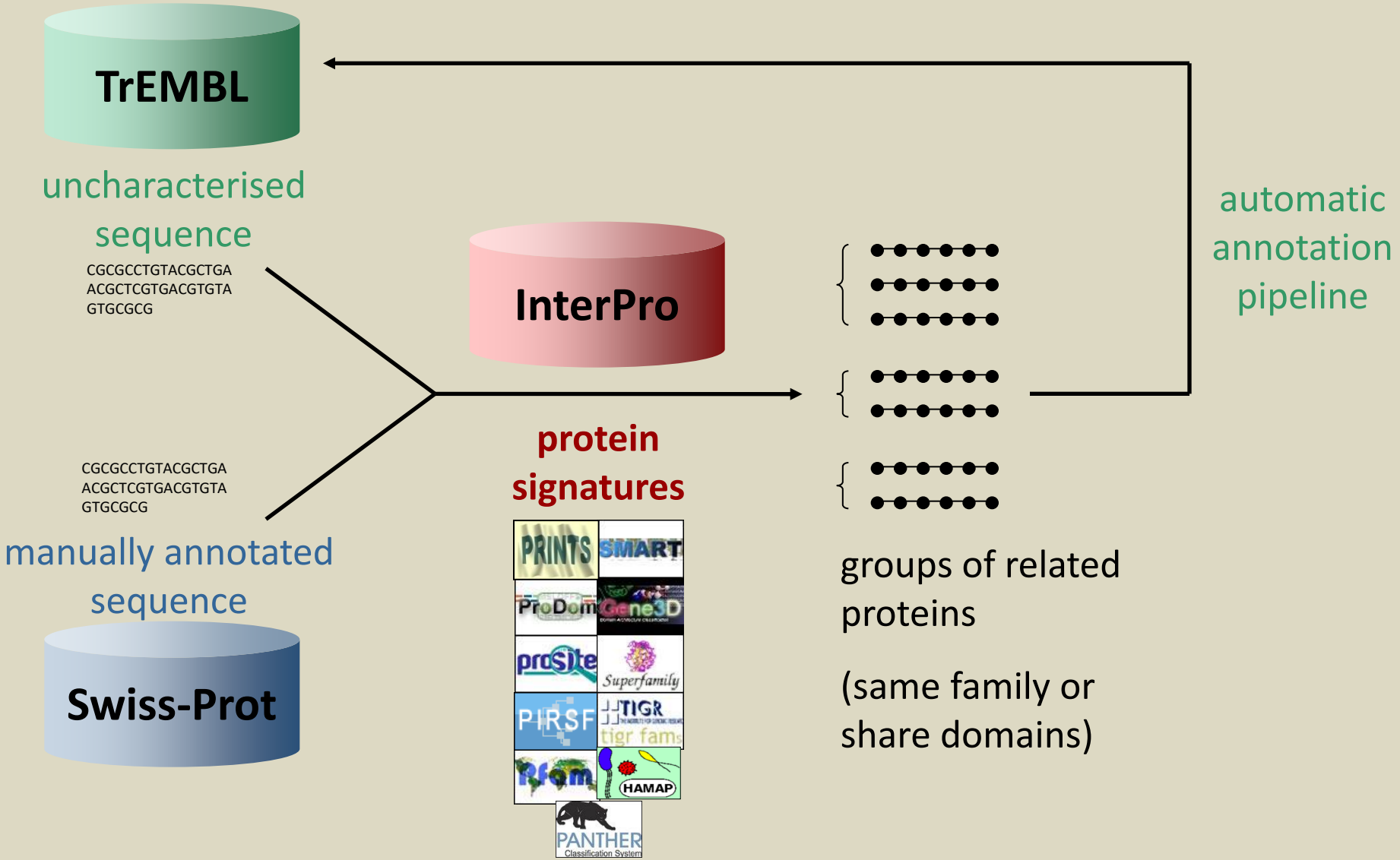
(new set every UniProtKB release)

UniRule:

maintains a set of manual annotation rules.



Automatic annotation - InterPro



UniProt Manual curation

- UniProt provides both manual curation and automatic annotation
- The UniProt manual curation process comprises manual review of results from a range of sequence analysis programs and literature curation of experimental data as well as attribution of all information to its original source.
- Curators also assign GO terms to all manually curated entries.

Manual curation process

- This process consists of 6 major mandatory steps:
  1. Sequence curation
  2. Sequence analysis
  3. Literature curation
  4. Family-based curation
  5. Evidence attribution
  6. Quality assurance and integration of completed entries.
- Curation is performed by expert biologists using a range of tools that have been iteratively developed in close collaboration with curators.

The UniProt Reference Clusters (UniRef)

- It provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. This hides redundant sequences and obtains complete coverage of the sequence space at three resolutions:

1. **UniRef100** combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.
2. **UniRef90** is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to, and 80% overlap with, the longest sequence (a.k.a. seed sequence).
3. **UniRef50** is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to, and 80% overlap with, the longest sequence in the cluster.

UniParc

- A comprehensive & non-redundant database
- Contains most of the publicly available protein sequences in the world.
- Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoids such redundancy by storing each unique sequence only once and giving it a **stable and unique identifier (UPI).**
- A UPI is never removed, changed or reassigned.
- UniParc contains only protein **sequences**. All other information about the protein must be retrieved from the source databases using the database cross-references.

- Predictions of sequence features such as Signal, Transmembrane and Coil regions are generated using the following software from external providers:
- TMHMM
- SignalP
- Phobius
- Coils
- TMHMM and Phobius predictors are used to infer transmembrane regions.