

SERVERS

- NCBI (National Centre for Biotechnology Information)
<http://www.ncbi.nlm.nih.gov>
- SIB (Swiss Institute of Bioinformatics)
<https://www.sib.swiss/>
- EBI-EMBL (European Bioinformatics Institute)
<http://www.ebi.ac.uk>

SIB (Swiss Institute of Bioinformatics)



The SIB central administration is hosted in the Génopode Building of the University of Lausanne (Switzerland)

<https://www.sib.swiss/>

The screenshot shows the SIB website homepage. At the top left is the SIB logo and the text "Swiss Institute of Bioinformatics". The top navigation bar includes links for "Services & resources", "Competence centres", "Personalized health", "Research", "Training", "Bioinformatics for all", and "About". On the left side, there's a sidebar with "Empowering life sciences" text and links for "ExPASy", "Technology Transfer", and "Embedded bioinformaticians". The main content area features a large image of molecular structures. A red arrow highlights the "ExPASy" link in the sidebar.

The screenshot shows the ExPasy homepage with the SIB logo and the text "ExPasy Bioinformatics Resource Portal". A search bar at the top contains "Query all databases". To the right, the Swiss Institute of Bioinformatics logo and links to Home, About, and Contact are visible. On the left, a sidebar menu includes "Visual Guidance", "Categories" (selected), "proteomics", "genomics", "structural bioinformatics", "systems biology", "phylogeny/evolution", "population genetics", "transcriptomics", "biophysics", "imaging", "IT infrastructure", and "drug design". Below this is a "Resources A..Z" link. At the bottom left is a "Links/Documentation" link next to a question mark icon. In the center, a "Featuring today" box highlights the "Bgee" database, described as a tool to retrieve and compare gene expression patterns between animal species, with a "details" link. A blue callout bubble on the right lists several resources: UniProtKB, SWISS-MODEL (Homology Modelling), T-Coffee (Multiple Sequence Alignment), LALIGN (pairwise alignment), and etc.

Resources :

- UniProtKB
- SWISS-MODEL
(Homology Modelling)
- T-Coffee
(Multiple Sequence Alignment)
- LALIGN
(pairwise alignment)
- etc

Branch of the EMBL, started in 1980 in Heidelberg.
Since 1995, at the Wellcome Trust Genome Campus at Hinxton



Mission (from the EBI website):

- To provide freely available **data and bioinformatics services** to all facets of the scientific community.
- To contribute to the advancement of biology through basic investigator-driven **research in bioinformatics**.
- To provide advanced bioinformatics **training to scientists** at all levels, from PhD students to independent investigators.
- To help disseminate cutting-edge technologies to **industry**.

European Bioinformatics Institu... (GB) https://www.ebi.ac.uk 80% ... Search EMBL-EBI Services Research Training About us EMBL-EBI

Open data

OMI.-E) (sraras)atF .rom Iif@ Wciez~e M=p.rOmen [ss

The home for big data in biology

More about EMBL-EBI's impact in our annual report ► Data from 2016

Our unique Search service helps you explore dozens of biological data resources.

More about EBI Search ►

All Find a gene, protein or chemical

Example searches: blast keratin bf11

Find a tool ► Deposit data ►

Find a tool for your data analysis.

Share your scientific data with the world.

We are EMBL-EBI

The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact. ►

Training

Access a wealth of world-leading training in bioinformatics and scientific service provision, regardless of your career stage or sector ►

Data resources

Explore our open data resources to enrich your research. Browse data, perform analyses or share your own results. ►

Industry

Explore our knowledge-exchange Industry Programme and take part in translational partnerships and projects ►

Research

Find out about our research groups, postdoctoral schemes and PhD Programme ►

ELIXIR

We support, as an ELIXIR node, the coordination of biological data provision throughout Europe ►

Databases at EBI

UniProt

A comprehensive resource for protein sequence and functional annotation.

Ensembl

Genome browser, API and database, providing access to reference genome annotation.

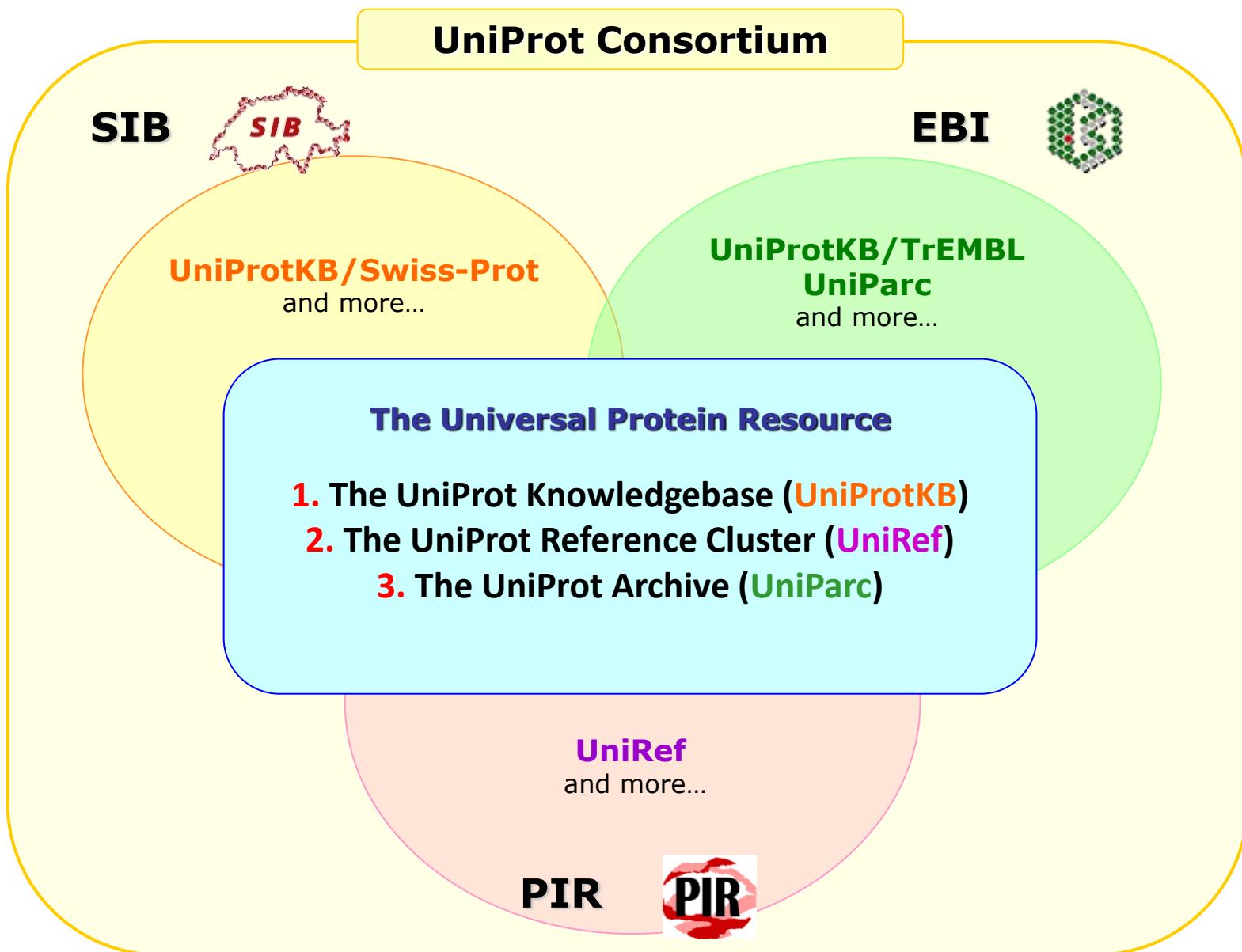
PDBe

The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB).

Who produces UniProt ?



The UniProt Consortium



UniProt: the Universal Protein knowledgebase

Rolf Apweiler*, Amos Bairoch¹, Cathy H. Wu², Winona C. Barker³, Brigitte Boeckmann¹, Serenella Ferro¹, Elisabeth Gasteiger¹, Hongzhan Huang², Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale², Claire O'Donovan, Nicole Redaschi¹ and Lai-Su L. Yeh³

The EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, ¹Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, ²Department of Biochemistry and Molecular Biology and ³National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Box 571414, Washington, DC 20057-1414, USA

Received August 25, 2003; Revised and Accepted October 27, 2003

ABSTRACT

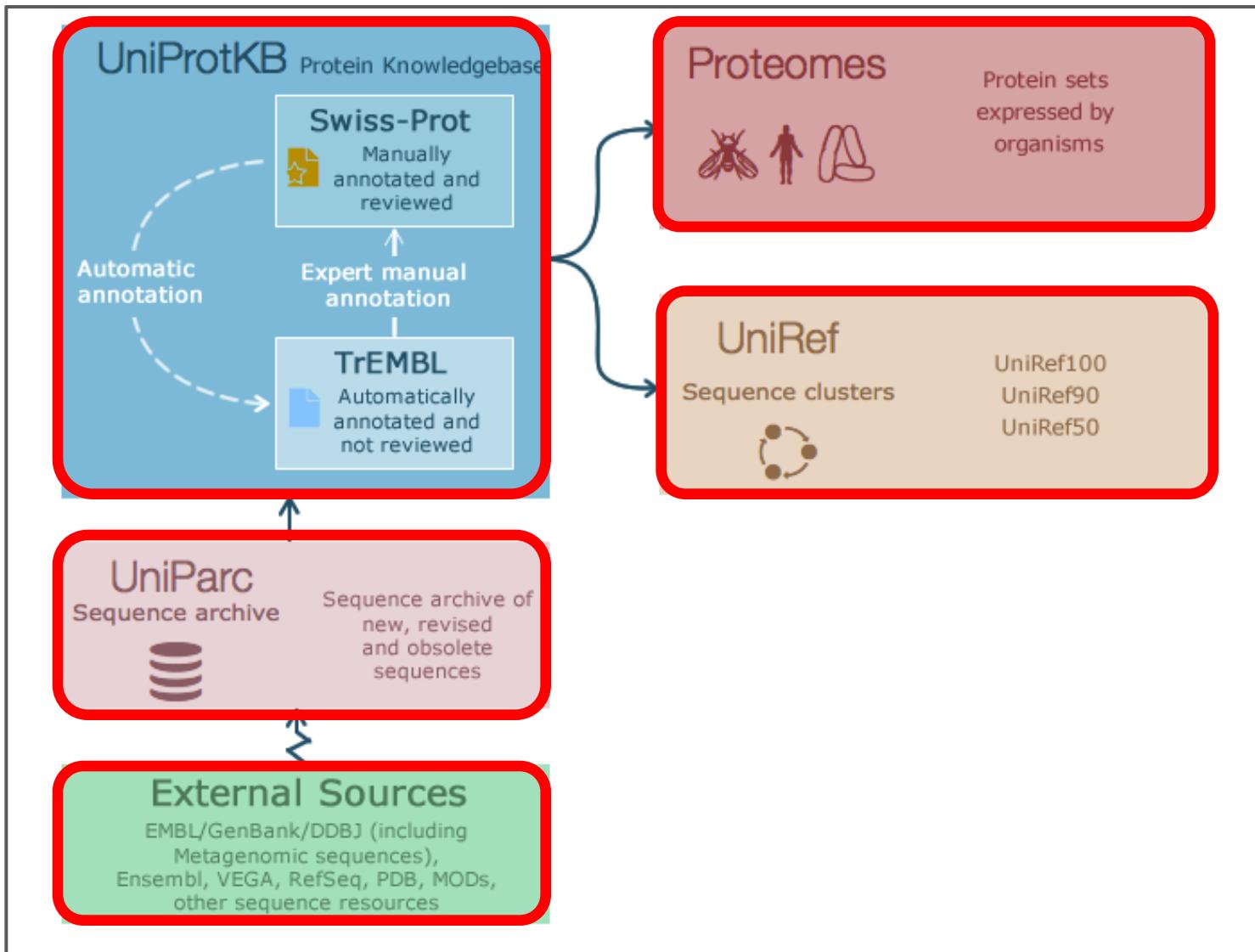
To provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information, the Swiss-Prot, TrEMBL and PIR protein database activities have united to form the Universal Protein Knowledgebase (UniProt) consortium. Our mission is to provide a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and query interfaces. The central database will have two sections, corresponding to the familiar Swiss-Prot (fully manually curated entries) and TrEMBL (enriched with automated classification, annotation and extensive cross-references). For convenient sequence searches, UniProt also provides several

The primary mission of the consortium is to support biological research by maintaining a high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community. UniProt will build upon the solid foundations laid by the consortium members over many years.

The UniProt databases consist of three database layers:

- (i) The UniProt Archive (UniParc) provides a stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data.
- (ii) The UniProt Knowledgebase (UniProt) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation.
- (iii) The UniProt NREF databases (UniRef) provide non-redundant data collections based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions.

Sources and flow of data for UniProt databases



UniProt

<http://www.uniprot.org/>

search bar

advanced

The screenshot shows the UniProt homepage with several key features highlighted:

- UniProtKB search bar:** A red box and arrow point to the search bar at the top right of the header.
- Advanced search:** A red box and arrow point to the "Advanced" dropdown menu next to the search bar.
- UniProtKB section:** A large red box highlights the UniProtKB section, which contains:
 - Swiss-Prot (546,790):** Manually annotated and reviewed.
 - TrEMBL (86,536,393):** Automatically annotated and not reviewed.
- UniRef, UniParc, Proteomes sections:** A red box highlights these three sequence-related sections.
- Supporting data section:** A red box highlights this section, which includes:
 - Literature citations
 - Taxonomy
 - Subcellular locations
 - Cross-ref. databases
 - Diseases
 - Keywords
- News section:** A red box highlights the news feed, showing recent releases:
 - K like Koagulation | Change of the cross-reference ArrayExpress to Expression/Atlas
 - UniProt release 2014_10**
 - Small is beautiful (and useful) | Evidences in the UniProtKB flat file format
 - UniProt release 2014_09**
 - Ubiquitin caught at its own game | New
 - News archive

date of last release

UniProt data

BLAST Align Retrieve/ID mapping

Help Contact

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.

UniProtKB
UniProt Knowledgebase

Swiss-Prot (549,646)
Manually annotated and reviewed.

TrEMBL (52,783,601)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations	Taxonomy	Subcellular locations
Cross-ref. databases	Diseases	Keywords

Getting started

Text search

Our basic text search allows you to search all the resources available

BLAST

Find regions of similarity between your sequences

Sequence alignments

Align two or more protein sequences using the Clustal



UniProt data

[Download latest release](#)
Get the UniProt data

[Statistics](#)

View Swiss-Prot and TrEMBL statistics

[How to cite us](#)

The UniProt Consortium

News



Forthcoming changes

Planned changes for UniProt

UniProt release 2015_10

The smell of the sea in UniProtKB | Cross-references to WBParaSite | Removal of the cross-references to CYGD | UniParc cross-reference t...

UniProt release 2015_09

Life (and death) in 2D | 27 new species in variation files

News archive

Protein spotlight

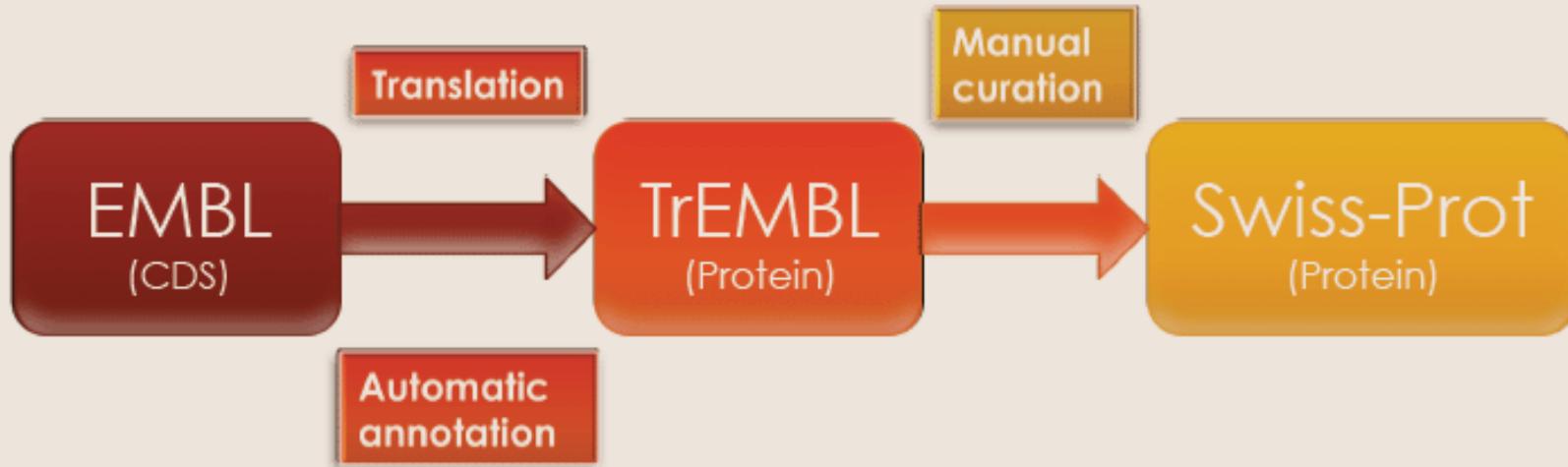


Approaching Happiness

August 2015

We all know what happiness is. At least we know what it feels like to be happy. But the moment you begin to define it, things become complex. And trying to measure a feeling as ungraspable as happiness seems as far-fetched as weighing a poem. Yet understanding what

Data Flow from EMBL DNA sequences to Swiss-Prot



Information Added to a UniProt/Swiss-Prot entry

[1] "The quaking gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins." Eberle T.A., Chen Q., Justice M.J., Ariz K. *Nat. Genet.* 12:260-265 (1995) [PubMed: 8589715] [Abstract]

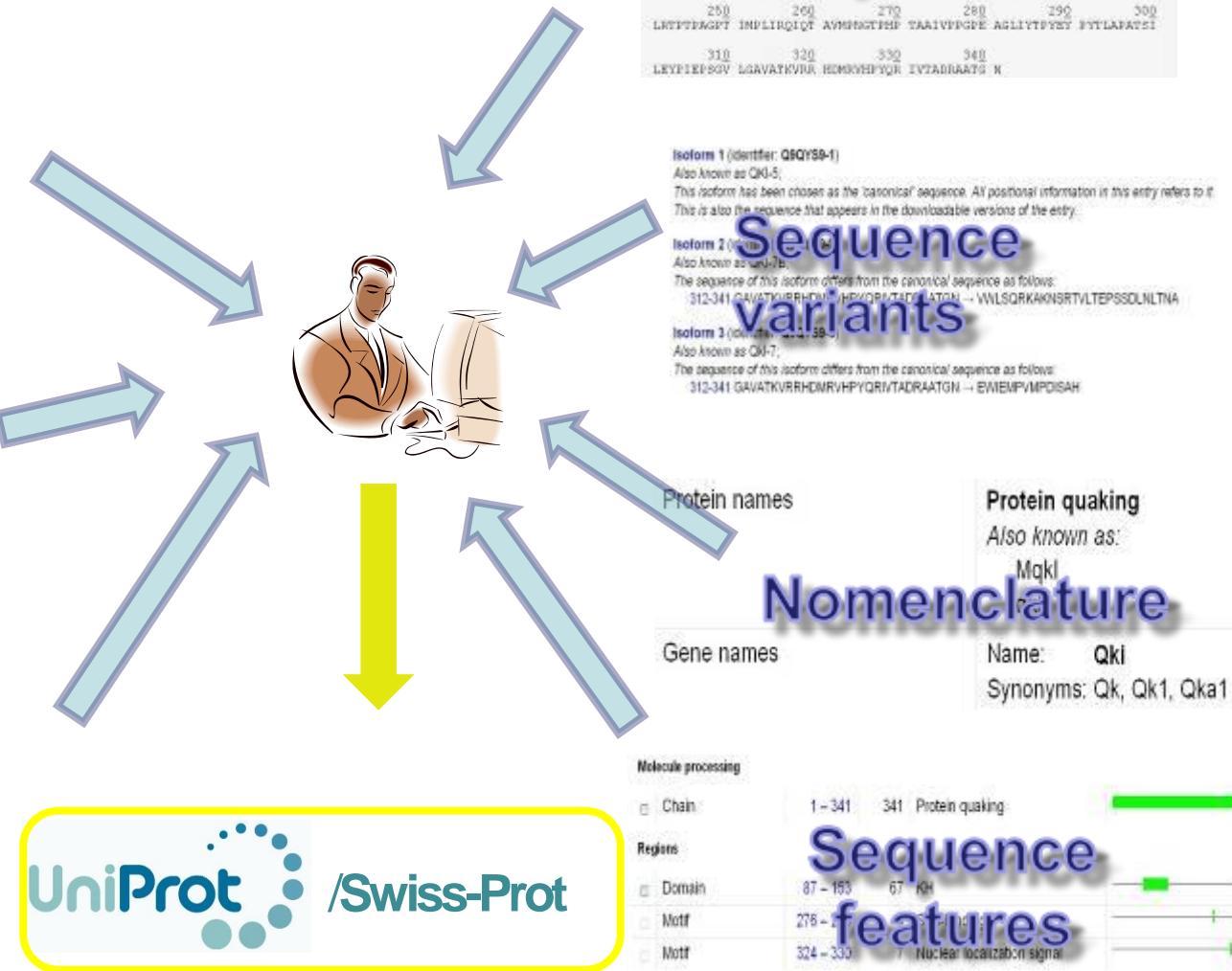
Def for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS 1; INVOLVEMENT IN ORX. TISSUE SPECIFICITY. MUTAGENESIS)

References

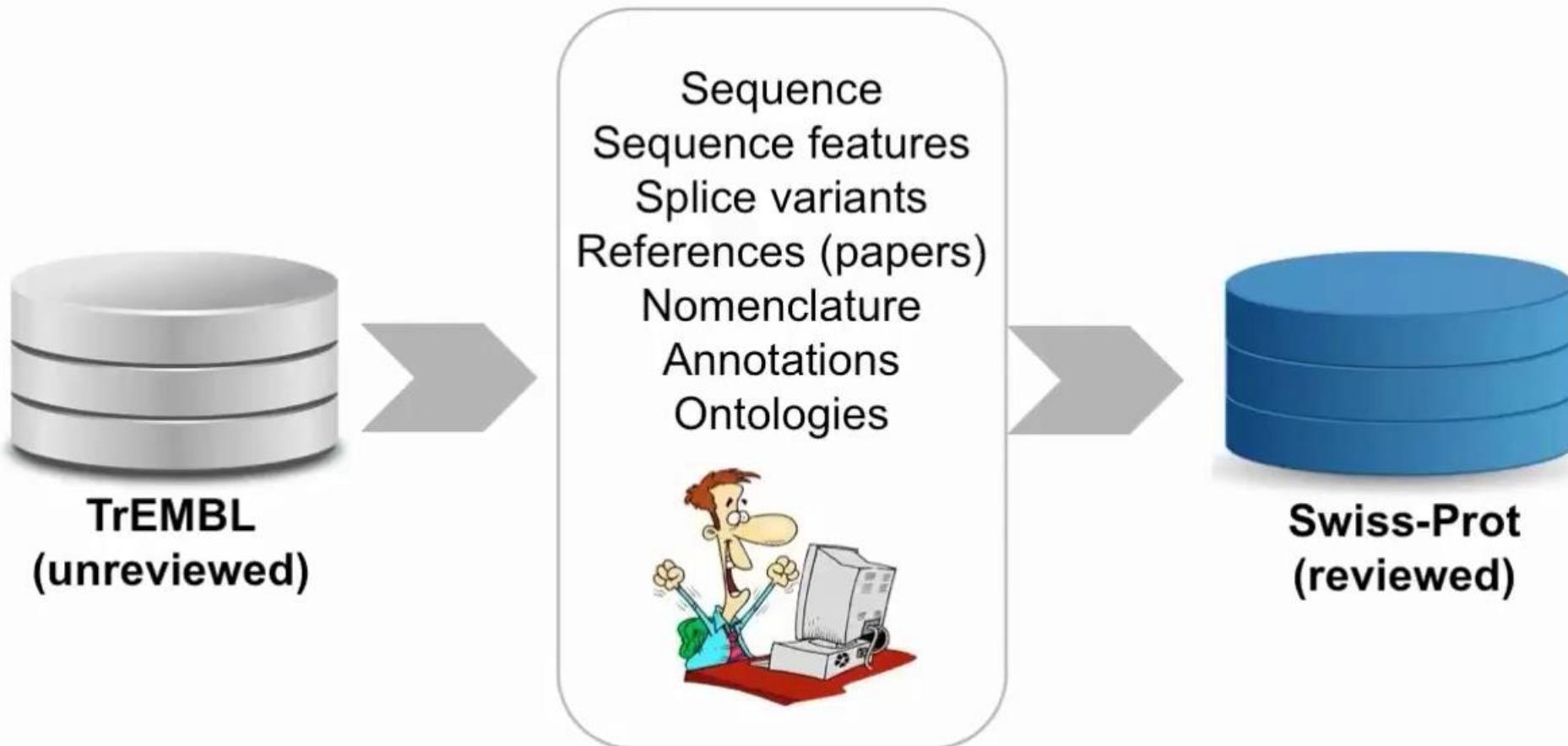
[2] "Genomic org." Kondo T., Furuta T., Matsuoka K., Eberle T.A., Smith M., Wu J., Ariz K., Yamamoto K., Abe K. *Mamm. Genome* 12:662-669 (1999) [PubMed: 10384037] [Abstract]

Def for: NUCLEOTIDE SEQUENCE [gDNA / mRNA] (ISOFORMS 2; 3; 4 AND T1; ALTERNATIVE SPlicing (ISOFORMS 1))

Strain: C57BL



Manual Curation of UniProtKB Entries



Sequence Curation

Discrepancies between sequence reports are identified, and the underlying causes of the sequence differences such as:

- alternative splicing
- natural variations
- frameshifts
- incorrect initiation sites
- incorrect exon boundaries
- unidentified conflicts
- erroneous gene prediction

are documented.

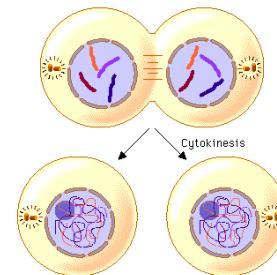
Comparison with homologous sequences is also used to identify additional sequence errors and their causes.

These steps ensure that the sequence described for each protein in UniProtKB/Swiss-Prot is as complete and correct as possible.

Gene Ontology (GO)

1. Biological Process

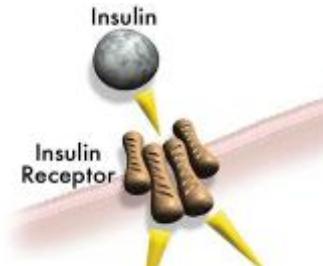
A commonly recognized series of events



- cell division
- Mitosis
- organelle fission

2. Molecular Function

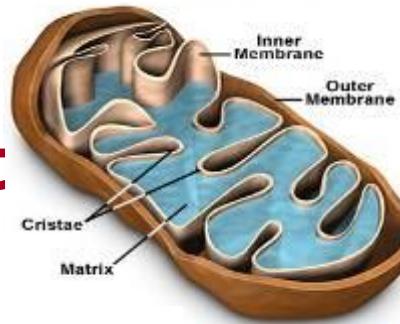
An elemental activity or task or job



- protein kinase activity
- insulin binding
- insulin receptor activity

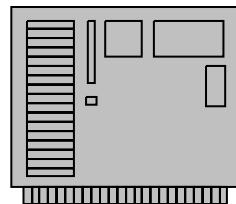
3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix
- mitochondrial membrane

The Two Sides of UniProtKB



UniProtKB/TrEMBL



Redundant
(1 entry per translated ENA entry),
computationally analyzed,
automatically annotated.

Unreviewed

*e.g. Accession Q9N0H9
Entry Name Q9N0H9_EQUAS*



UniProtKB/Swiss-Prot



Non-redundant
(1 entry per protein),
curator-evaluated analysis,
high-quality manual annotation.

Reviewed

*e.g. Accession P38398
Entry Name BRCA1_HUMAN*

UniProt (1) – Entry Name

http://www.uniprot.org/help/entry_name

Find out which ones of the following 5 entries are unreviewed:

- 1) Q84J55 14331_ORYSJ
- 2) Q0DV28 ARK1_ORYSJ
- 3) Q5Z8U4 Q5Z8U4_ORYSJ
- 4) P00986 VKT2_NAJNI
- 5) Q8WA21 Q8WA21_NAJNI

Uniprot KnowledgeBase

Introduction (video)

<https://www.youtube.com/watch?v=ado1r8IDm3U>

Reading a SwissProt or trEMBL Entry

- General information about the Entry
(primary AC, Entry Name, History, annotation score.)
- Name and Origin of the Protein
(Protein Name, Synonymous, Gene Name, Source, Taxonomy, EC Number.)
- Function
(Protein Length, Binding Sites, GO terms, Keywords, EC Number, Subunits.)
- Subcell. location, Pathology, PTM, Expression, Interactions.
- Structure, Family and Domains, Sequence.
- Cross-References, Publications, Entry Information, Similar Proteins
(Sequence databases, Structure databases, link to UniRef.)

UniProt (2) – downloads

- A. Download, as an uncompressed Excel table, all entries from *Saccharomyces cerevisiae* (Baker's yeast) :

- 1) which are reviewed;
- 2) whose function is “DNA binding”;

showing for each entry the content of the following fields:

- 1) Accession Number;
- 2) Entry Name;
- 3) Protein Name;
- 4) Gene Name;
- 5) Length;
- 6) Evidence for existence;
- 7) GO terms for “biological process”, “molecular function”
“cellular component”;
- 8) DNA binding.

- B. Map the UniProtKB IDs of all found entries to their EMBL/GenBank/DDBJ IDs.

UniProt (3) – Feature viewer

Go to the entry with AC number **P06858**.

- 1) Switch the «Display» from «entry» to «Feature viewer».**
- 2) In «Domains & sites», click on the dot representing the active site residue 183.**
- 3) Using the ruler on top of the view, zoom in to that residue, with a window of about 50 residues. In the «Variants» section, find out which variations are known for that residue. Are they pathogenic? Which are the cross-references for each variant?**
- 4) Find the same information with the «Display» switched back to «entry».**

UniProt (4) – From diseases to protein variants

⇒ **Questionnaire on UniProt, #1**

Evidence tags

What do evidences look like?

- manual assertions:

1 Publication By similarity UniRule annotation Imported

- automatic assertions:

UniRule annotation Imported

TrEMBL – Automatic Annotation

E3PCB7 - E3PCB7_ECOH1

Basket ▾

Protein Submitted name: **Arabinose operon regulatory protein**
Gene **ETEC_0063**
Organism *Escherichia coli* O78:H11 (strain H10407 / ETEC)
Status Unreviewed - ○○○○○ - Protein predictedⁱ

Display

None

BLAST Align Format Add to basket History

Comment (?) Feedback Help video

- FUNCTION
- NAMES & TAXONOMY
- SUBCELLULAR LOCATION
- PATHOLOGY & BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCE
- CROSS-REFERENCES
- PUBLICATIONS

Functionⁱ

GO - Molecular functionⁱ

- ▶ sequence-specific DNA binding
- ▶ sequence-specific DNA binding, transcription factor activity

GO - Biological processⁱ

- ▶ transcription, DNA-templated

Complete GO annotation...

Keywords - Molecular functionⁱ

Activator

Keywords - Biological processⁱ

Transcription, Transcription regulation

Keywords - Ligandⁱ

DNA-binding

Enzymeⁱ

Source: InterPro

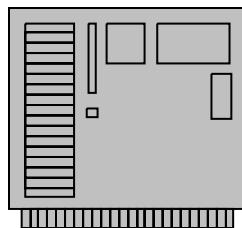
Source: UniProtKB-KW

UniRule annotation SAAS annotation

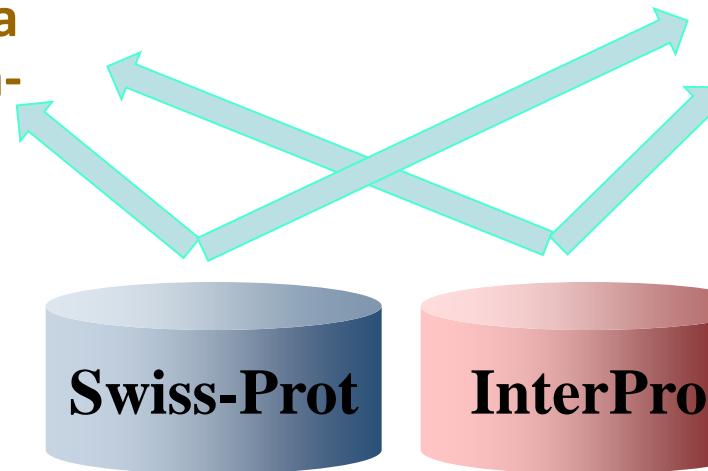
UniRule annotation SAAS annotation

Automatic Annotation

UniProtKB employs two prediction programs which are referred to as **UniRule** and **SAAS**.



SAAS, Statistical Automatic Annotation System, generates a new set of decision-trees with every UniProtKB release using data-mining.



UniRule maintains a set of manually established annotation rules.

InterPro is an EBI database that collates domain signatures and defines protein family membership.

UniProt (5) – example of UniRule

- 1) Go to the entry **Q02PG5**. The recommended name for this protein is «Glyceraldehyde-3-phosphate dehydrogenase-like protein». Which is the code of the UniRule rule according to which this annotation has been issued?
- 2) Which are the common conditions required to issue such annotation?
- 3) How many rules are there in UniRule?

UniProt (6) – example of SAAS

- 1) Go to the entry **Q87644**. The catalytic activity has been annotated with «ATP + H₂O = ADP + phosphate». Which is the code of the SAAS rule according to which this annotation has been issued?
- 2) Which are the common conditions required to issue such annotation?
- 3) How many proteins are annotated with this rule in the whole database?

HAMAP: Automatic Annotation Programs

Entry information¹

Entry name ¹	CYOB_ECOLI
Accession ¹	Primary (citable) accession number: POAB18 Secondary accession number(s): P18401, Q2MBZ5
Entry history ¹	Integrated into UniProtKB/Swiss-Prot: Last sequence update: October 25, 2005 Last modified: October 25, 2017
	This is version 106 of the entry and version 1 of the sequence. See complete history .

Entry status¹

Annotation program

Pokaryotic Protein Annotation Program

Prokaryotic protein annotation project

Basket 11 ▾

Details

References

Contributors

Links

Statistics

Contribute

Submissions and updates

Headlines

From mouth to g...



High-quality Automated and Manual Annotation of Proteins

Home | Browse | HAMAP-Scan | Proteomes | Documentation | Downloads | About

Search HAMAP

Search

HAMAP is a system for the classification and annotation of protein sequences. It consists of a collection of manually curated family profiles for protein classification, and associated, manually created [annotation rules](#) that specify annotations that apply to family members. HAMAP is used to annotate protein records in UniProtKB via [UniProt's automatic annotation pipeline](#). We also provide an interface to scan user sequences against HAMAP family profiles [\[More...\]](#).

HAMAP

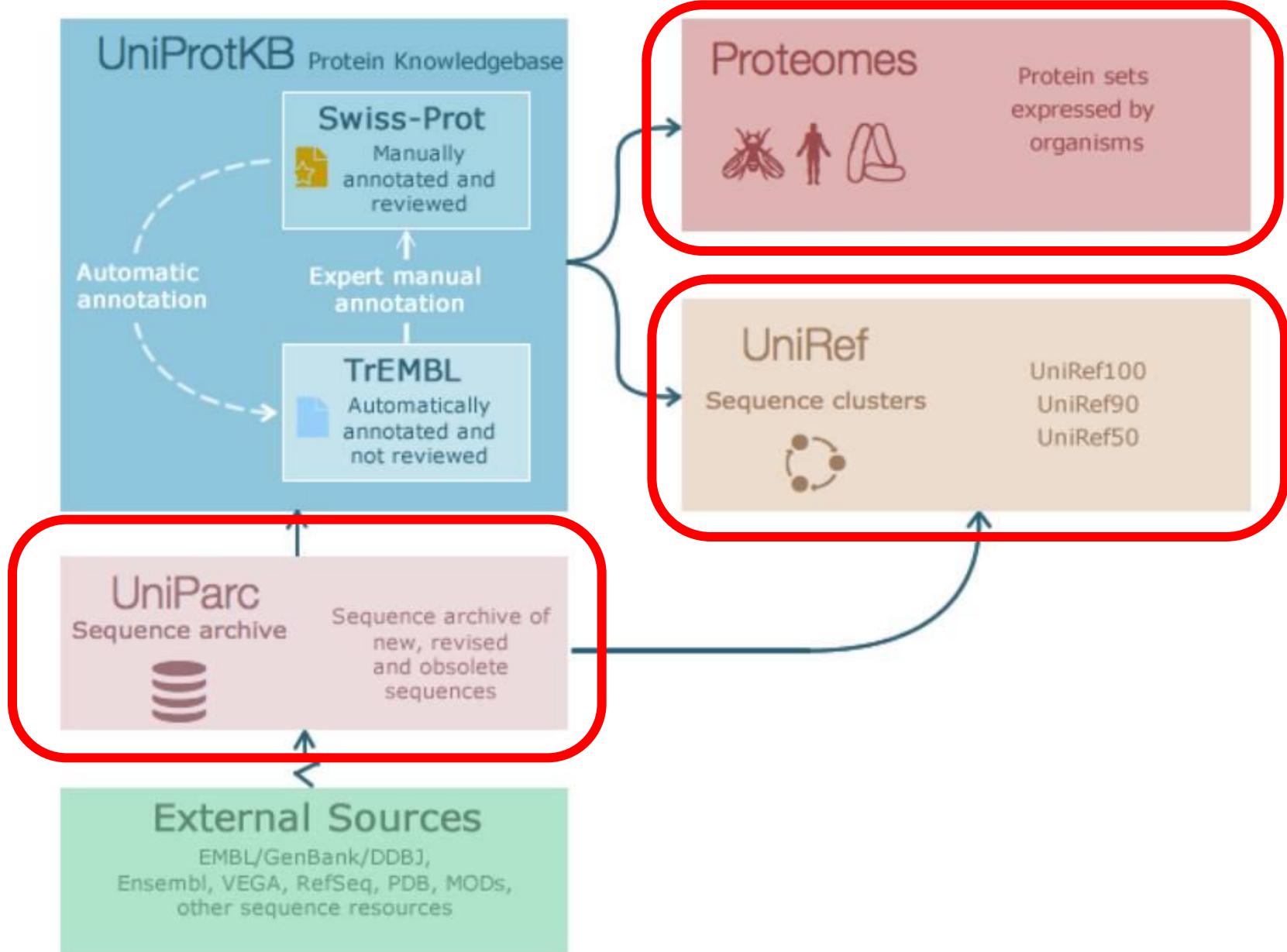
HAMAP is based on a collection of expert curated protein family profiles, which are used to determine family membership of protein sequences, and annotation rules.

HAMAP rules permit the annotation of protein sequences to the same level of detail and quality as manually curated UniProtKB/Swiss-Prot records, annotating protein and gene names, function, catalytic activity, cofactors, subcellular location, protein–protein interactions, etc.

HAMAP rules also specify the conditions under which these annotations may be applied, such as a requirement for key functional residues (identified by structural or other experimental studies). Such conditions can reduce the incidence of erroneous annotation, particularly in large, functionally diverse families— errors that tend to persist in public sequence databases.

About UniProt

<http://www.uniprot.org/help/about>



UniParc

The screenshot shows the UniProt website homepage. At the top, there is a navigation bar with the UniProt logo, a dropdown menu for 'UniProtKB', an 'Advanced' search bar, and a search icon. Below the navigation bar, there are links for BLAST, Align, Retrieve/ID Mapping, Help, and Contact.

Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still [go back to the old site](#).

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.

UniProtKB

- Swiss-Prot** (546,790)
Manually annotated and reviewed.
- TrEMBL** (86,536,393)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations	Taxonomy	Subcellular locations
Cross-ref. databases	Diseases	Keywords

News

- K like Koagulation | Change of the cross-reference ArrayExpress to ExpressionAtlas
[UniProt release 2014_10](#)
- Small is beautiful (and useful) | Evidences in the UniProtKB flat file format
[UniProt release 2014_09](#)
- Ubiquitin caught at its own game | New
[News archive](#)

UniParc

Display None

Sequences (5)ⁱ

Sequence statusⁱ: Complete.
Sequence processingⁱ: The displayed sequence is further processed into a mature form.
This entry describes 5 isoformsⁱ produced by **alternative splicing**. [Align](#)

Note: Additional isoforms seem to exist.

Isoform Alpha (identifier: P29461-1) [UniParc] [ASTA](#) [Add to Basket](#)

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

Length: 404
Mass (Da): 45,159
Last modified: April 1, 1993 - v1
Checksum:ⁱ ABF33CF33CC71584

BLAST [GO](#)

10	20	30	40	50
MADKVILKEKR	KLFIRSMGEG	TINGLLDELL	QTRVLNKEEM	EKVKRENATV
60	70	80	90	100
MDKTRALIDS	VIPKGAAQACQ	ICITYICEED	SYLAGTLGLS	ADQTSGNYLN
110	120	130	140	150
MQDSQGVLSS	FPAPQAVQDN	PAMPTSSGSE	GNVKLCSLEE	AQRIWKQKSA
160	170	180	190	200
EIYPIMDKSS	RTRLALIICN	EEFDSSIPRT	GAEVDITGMT	MLLQNLGYSV
210	220	230	240	250
DVKKNLTASD	MTTELEAFAH	RPEHKTSDDST	FLVFMSHGIR	EGICGKKHSE
260	270	280	290	300

caution [Highlight All](#) [Match Case](#) 1 of 1 match

UniParc

UniParc is the most comprehensive publicly accessible **non-redundant** protein database.

UniParc records are **not annotated** because annotation is context-dependent: proteins with the same sequence can have different functions depending on species, tissue, developmental stage or other variables.

This context-dependent information is the scope of **UniProtKB**.

UniRef



UniProtKB ▾

Advanced ▾



BLAST Align Retrieve/ID Mapping

Help Contact

Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still [go back to the old site](#).

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.

UniProtKB
Swiss-Prot (546,790) Manually annotated and reviewed.
TrEMBL (86,536,393) Automatically annotated and not reviewed.

 UniRef
Sequence clusters 

UniParc
Sequence archive 

Proteomes 

Supporting data

Literature citations 	Taxonomy 	Subcellular locations 
Cross-ref. databases 	Diseases 	Keywords 

News



K like Koagulation | Change of the cross-reference ArrayExpress to ExpressionAtlas
[UniProt release 2014_10](#)

Small is beautiful (and useful) | Evidences in the UniProtKB flat file format
[UniProt release 2014_09](#)

Ubiquitin caught at its own game | New

 [News archive](#)

UniRef

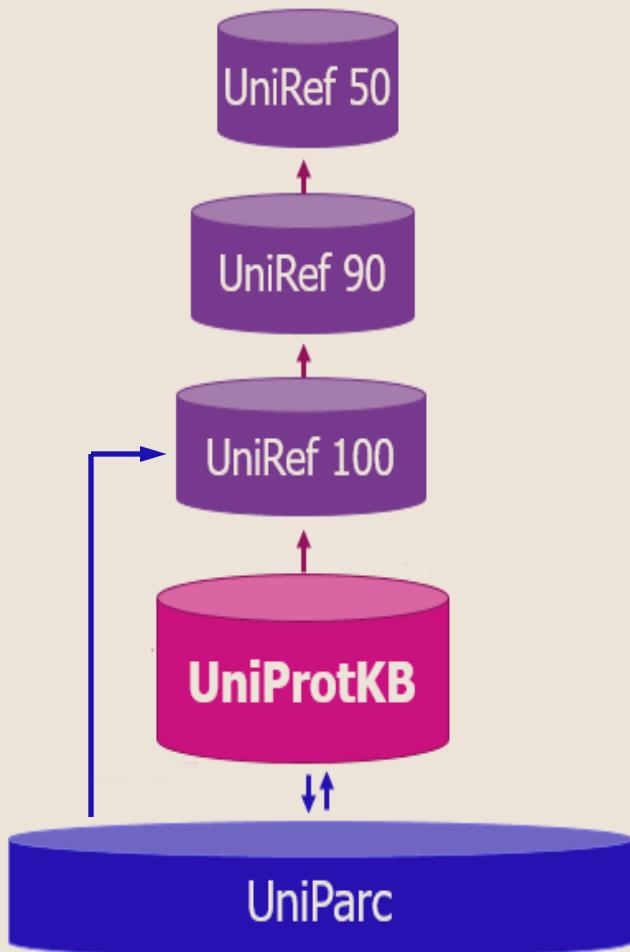
The **UniProt Reference Clusters (UniRef)** provide **clustered sets** of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. This **hides redundant sequences** and obtains complete coverage of the sequence space at three resolutions:

UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.

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 (seed).

UniRef50 is built by clustering **UniRef90** seed sequences that have **at least 50% sequence identity to, and 80% overlap with**, the longest sequence (seed) in the cluster.

UniRef



Database size reduction:

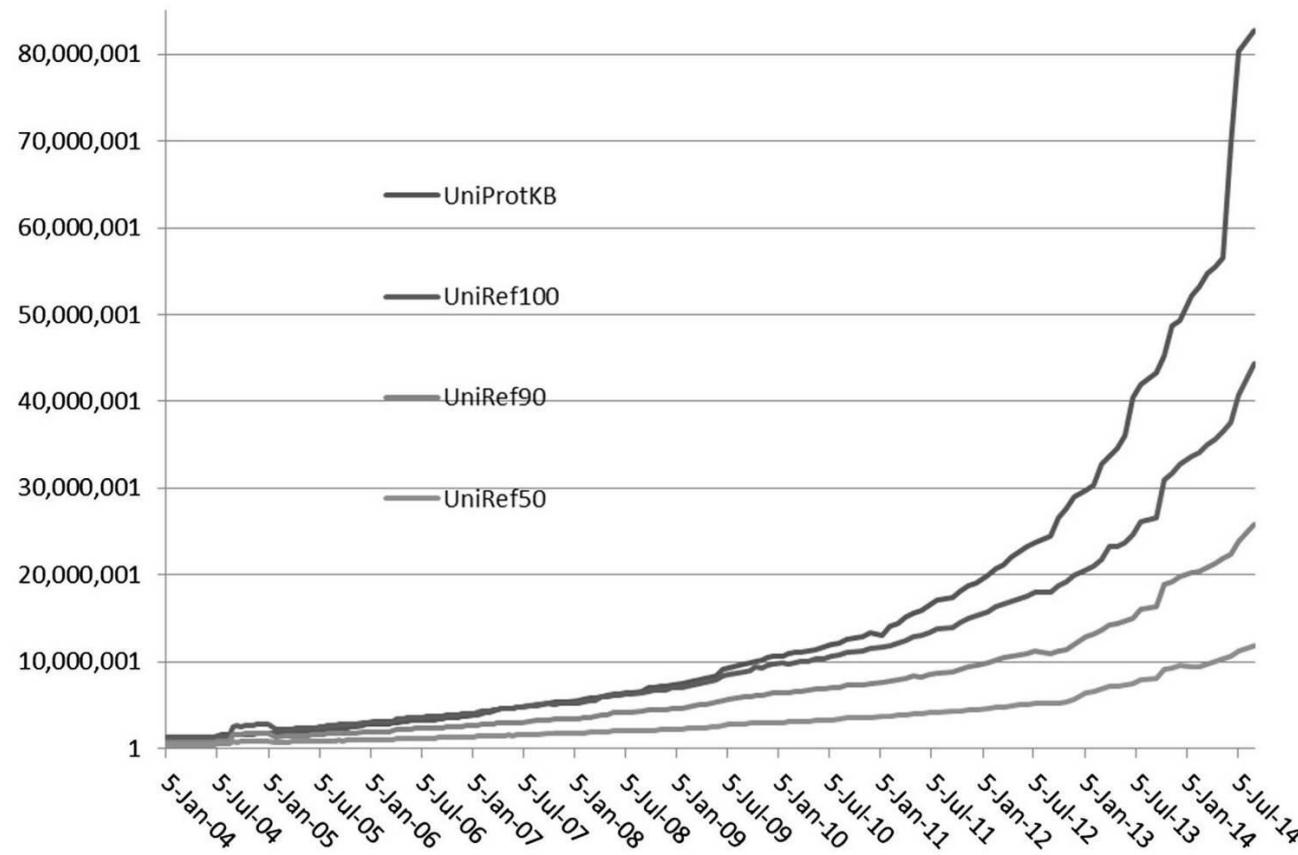
≈ 70% (UniRef90)

≈ 85% (UniRef50)



**significantly faster
sequence searches.**

Growth of UniRef Databases and UniProt Knowledgebase



Baris E. Suzek et al. Bioinformatics 2015;31:926-932

UniRef50-based BLASTP searches are faster (6x), more concise (lists are 7x shorter), and overall more sensitive in the detection of remote similarities.

From <http://www.uniprot.org/help/uniref/>

The **seed sequences** are the longest members of the cluster.

However, the biologically relevant information (**name, function, cross-references**) is often available on other cluster members.

The **biologically relevant representative** for the cluster is chosen based on:

- ✓ **quality of the entry:** manually reviewed entries are preferred
- ✓ **UniProtKB annotation score:** higher are preferred
- ✓ **organism:** entries from reference proteomes and model organisms are preferred
- ✓ **length of the sequence:** longest sequence is preferred

UniProt (8) – Searches

⇒ **Questionnaire on UniProt #2**

UniSave

UniProtKB Advanced Search

BLAST Align Retrieve/ID Mapping Help Contact Basket

P29466 - CASP1_HUMAN

Protein Caspase-1
Gene CASP1
Organism Homo sapiens (Human)
Status Reviewed - Experimental evidence at protein level

Display None

FUNCTION NAMES & TAXONOMY SUBCELLULAR LOCATION PATHOLOGY & BIOTECH PTM / PROCESSING EXPRESSION INTERACTION

BLAST Align Format Add to basket History

Comment (?) Feedback Help video

Function

Thiol protease that cleaves IL-1 beta between an Asp and an Ala, releasing the mature cytokine which is involved in a variety of inflammatory processes. Important for defense against pathogens. Cleaves and activates sterol regulatory element binding proteins (SREBPs). Can also promote apoptosis. 2 Publications

Catalytic activity

Strict requirement for an Asp residue at position P1 and has a preferred cleavage sequence of Tyr-Val-Ala-Asp-|-.
Enzyme regulation

Specifically inhibited by the cowpox virus Crma protein.

Sites

UniProt – BLAST, Align, ID Mapping

The screenshot shows the UniProtKB results page for the query "egfr human". A red box highlights the navigation bar at the top, which includes links for BLAST, Align, and Retrieve/ID mapping.

Filter by:

- Reviewed (352)
Swiss-Prot
- Unreviewed (200)
TrEMBL
- Popular organisms:**
 - Human (338)
 - Zebrafish (80)
 - Mouse (44)
 - Rat (10)
 - Fruit fly (5)
 - Other organisms

Search terms: Filter "egfr" as:

UniProtKB results:

1 to 250 of 552 | Show 250

Entry	Entry name	Protein names	Gene names	Organism	Length
P04412	EGFR_DROME	Epidermal growth factor receptor	Egfr, c-erbB, DER, top, CG10079	Drosophila melanogaster (Fruit fly)	1,426
Q86T13	CLC14_HUMAN	C-type lectin domain family 14 memb...	CLEC14A, C14orf27, EGFR5, UNQ236/PRO269	Homo sapiens (Human)	490
Q96AW1	VOPP1_HUMAN	Vesicular, overexpressed in cancer,...	VOPP1, ECOP, GASP	Homo sapiens (Human)	172
A2VCQ7	A2VCQ7_HUMAN	EGFR protein	EGFR	Homo sapiens (Human)	388
Q8IYD2	KLD8A_HUMAN	Kelch domain-containing protein 8A	KLHDC8A	Homo sapiens (Human)	350
<input checked="" type="checkbox"/> P00533	EGFR_HUMAN	Epidermal growth factor receptor	EGFR, ERBB, ERBB1, HER1	Homo sapiens (Human)	1,210
Q96B97	SH3K1_HUMAN	SH3 domain-containing kinase-bindin...	SH3KBP1, CIN85	Homo sapiens (Human)	665
Q14ZT7	Q14ZT7_HUMAN	EGFR protein	EGFR	Homo sapiens	454

Proteomes Section

The screenshot shows the UniProt website homepage. At the top, there is a navigation bar with links for UniProtKB (selected), Advanced search, Help, and Contact. Below the navigation bar, a yellow banner welcomes users to the new site and provides a link to go back to the old site. A main message states: "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." On the left, there are two main sections: UniProtKB (Swiss-Prot: 546,790 entries, Manually annotated and reviewed; TrEMBL: 86,536,393 entries, Automatically annotated and not reviewed) and UniRef (Sequence clusters). In the center, there are four main categories: UniParc (Sequence archive), Proteomes (highlighted with a red circle), Supporting data (with sub-sections: Literature citations, Taxonomy, Subcellular locations, Cross-ref. databases, Diseases, and Keywords), and News. The Proteomes section features icons for a fly, a person, and a DNA helix.

Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still [go back to the old site](#).

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.

UniProtKB

Swiss-Prot (546,790)
Manually annotated and reviewed.

TrEMBL (86,536,393)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations
Taxonomy
Subcellular locations

Cross-ref. databases
Diseases
Keywords

News

K like Koagulation | Change of the cross-reference ArrayExpress to ExpressionAtlas
[UniProt release 2014_10](#)

Small is beautiful (and useful) | Evidences in the UniProtKB flat file format
[UniProt release 2014_09](#)

Ubiquitin caught at its own game | New

[News archive](#)

Proteomes Section

UniProt Proteomes▼ Advanced Search Basket 4

Proteomes results

What are proteomes?
A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes. [? Help](#)

What are reference proteomes?
Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. [? Help](#)

[Proteomes help video](#) [Other tutorials and videos](#) [Downloads](#)

Filter by: [Download](#) [Columns](#) 1 to 25 of 160,232 ▶ Show 25

Show only non-redundant proteomes? X

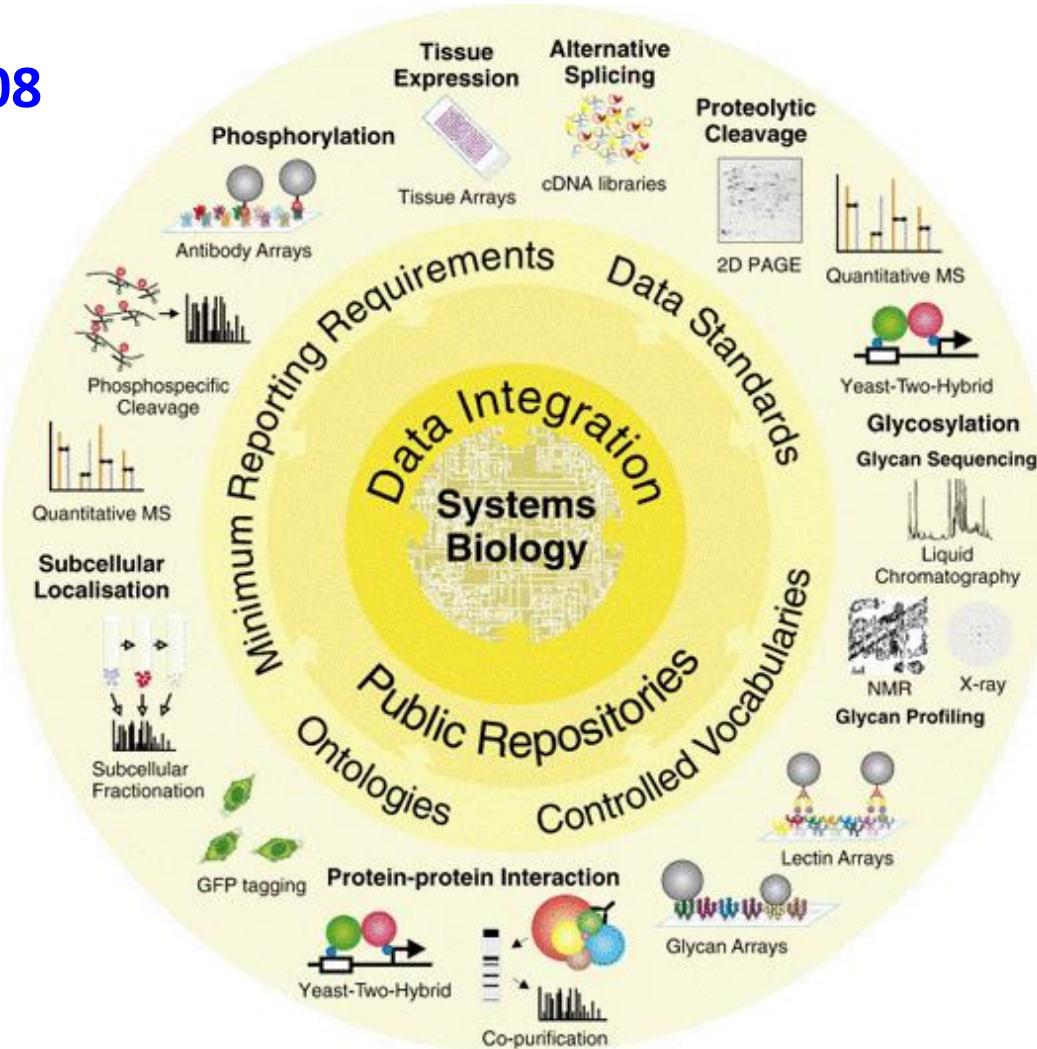
<input type="checkbox"/>	Proteome ID	Organism	Organism ID	Protein count	⋮
<input type="checkbox"/>	UP000000832	R Banna virus (strain Indonesia/JKT-6423/1980) (BAV) (Strain: Indonesia/JKT-6423/1980)	649604	12	
<input type="checkbox"/>	UP000134051	R Budgerigar fledgling disease virus (BFPyV) (Aves polyomavirus 1)	1891747	5	
<input type="checkbox"/>	UP000017836	R Amborella trichopoda	13333	27371	
<input type="checkbox"/>	UP000000352	R Helminthosporium victoriae virus-190S (Hv190SV)	45237	2	
<input type="checkbox"/>	UP000000431	R Chlamydia trachomatis (strain D/UW-3/Cx) (Strain: D/UW-3/Cx)	272561	895	
<input type="checkbox"/>	UP000000429	R Helicobacter pylori (strain ATCC 700392 / 26695) (Campylobacter pylori) (Strain: ATCC 700392 / 26695)	85962	1553	
<input type="checkbox"/>	UP000007110	R Strongylocentrotus purpuratus (Purple sea urchin)	7668	28594	
<input type="checkbox"/>	UP000002006	R Escherichia phage PhiEco32 (Escherichia coli phage phi32)	490103	128	
<input type="checkbox"/>	UP000008912	R Ailuropoda melanoleuca (Giant panda)	9646	21136	
<input type="checkbox"/>	UP000008592	R Lettuce necrotic yellows virus (isolate 318) (LNYV) (Strain: Isolate 318)	928304	6	

kinase Highlight All Match Case Whole Words 6 of 10 matches

Annotating the human proteome

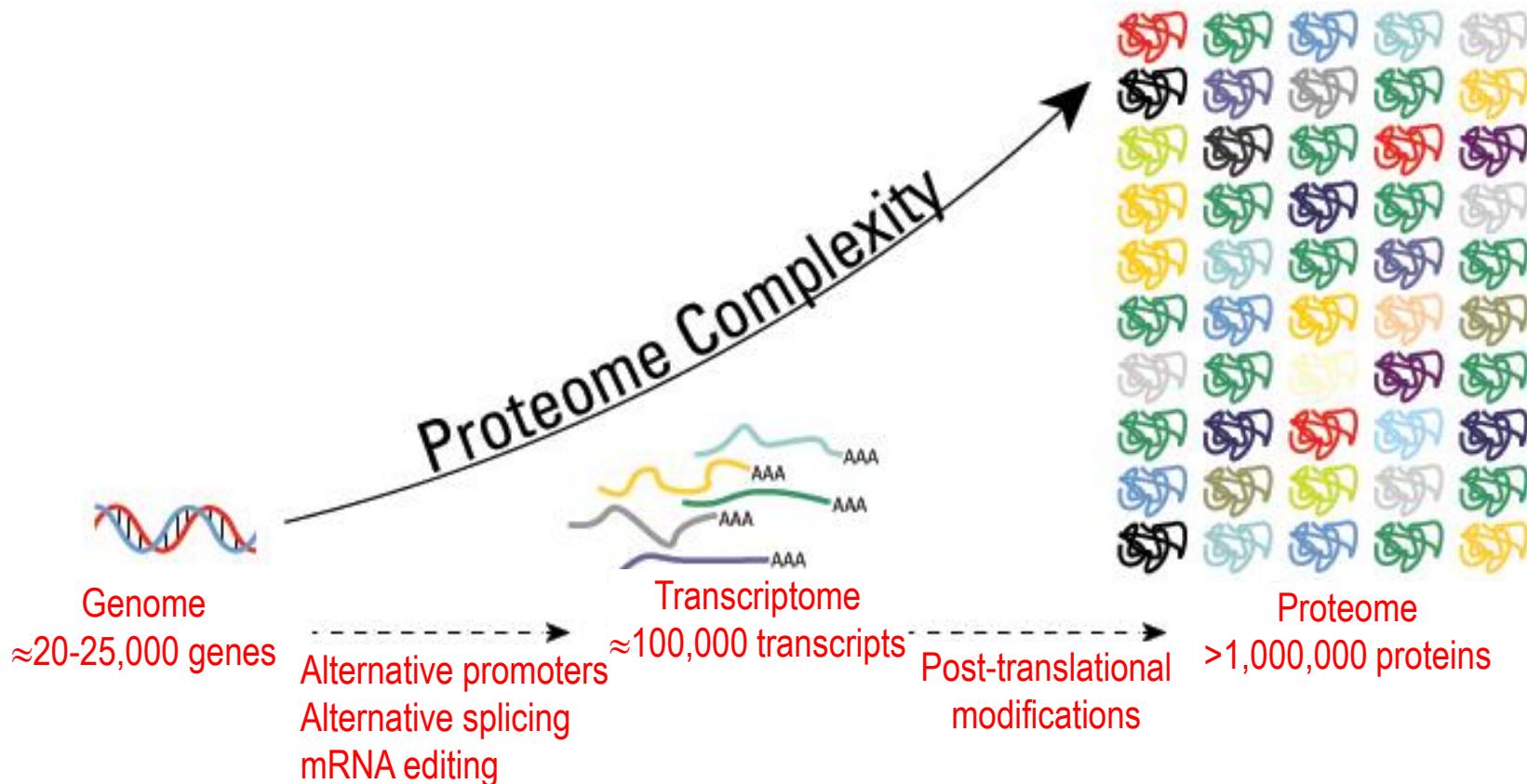
September 2008

First draft of
the complete
human
proteome
available
in UniProtKB/
Swiss-Prot.

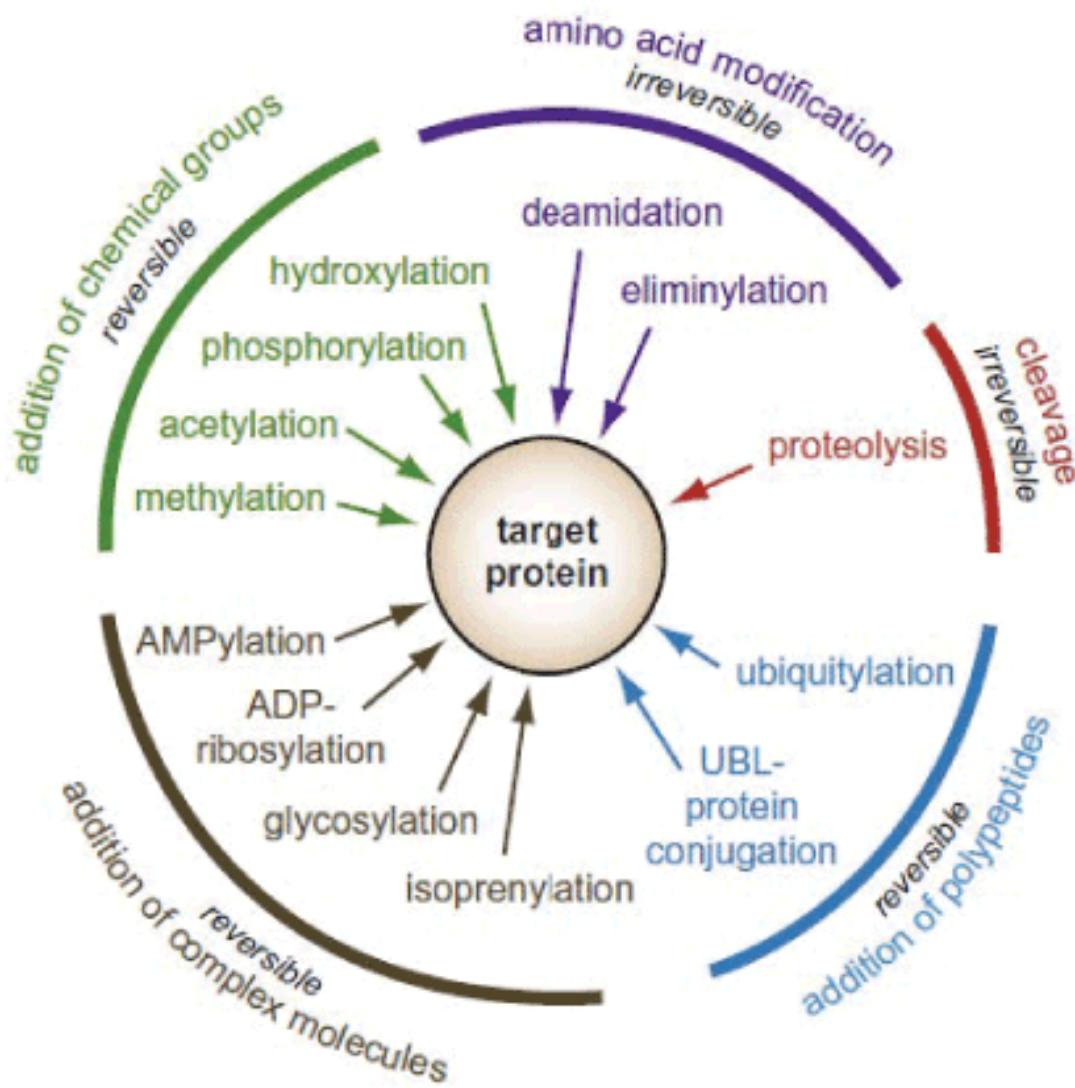


Gene prediction algorithms, plus the existing transcript and protein information, have enabled the identification of most exons in a genome.

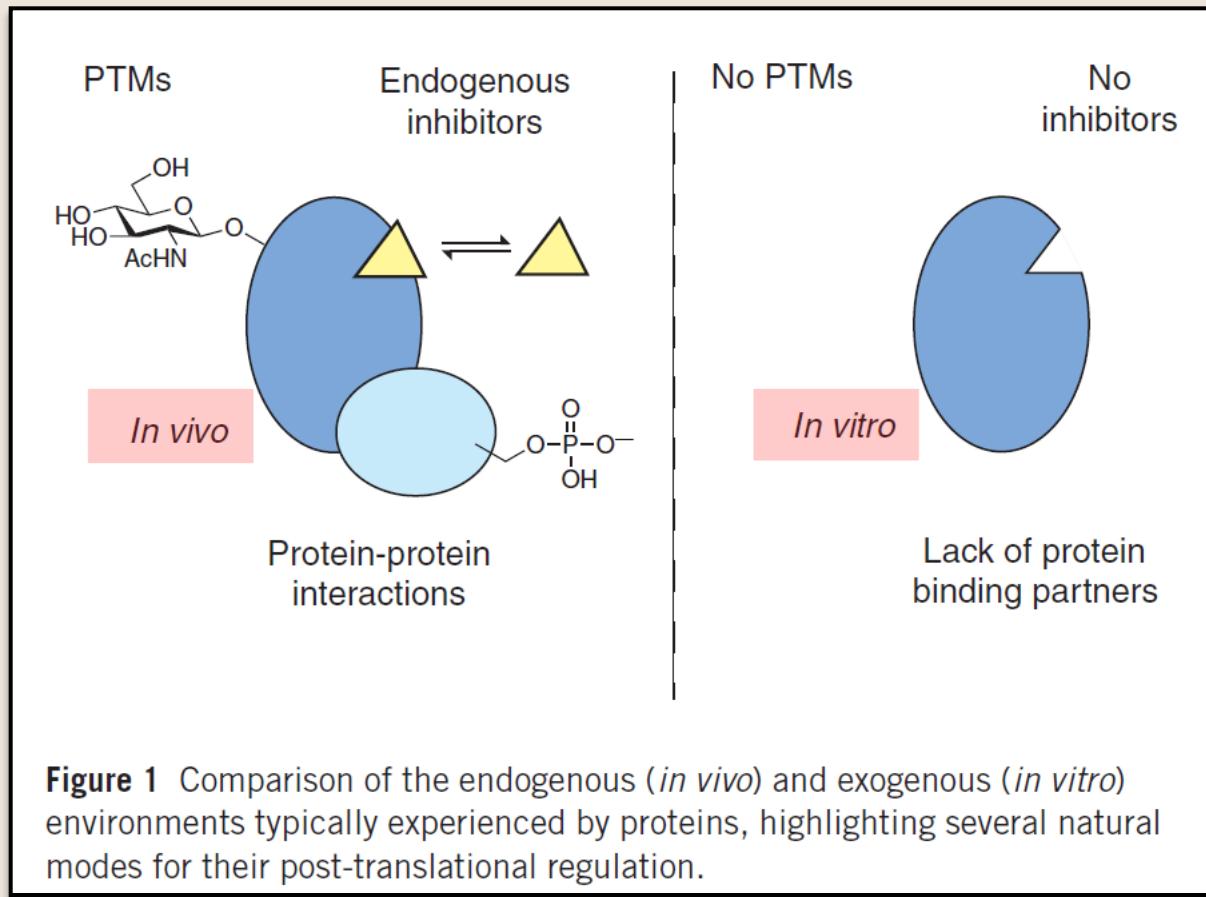
The Human Proteome



Post Translational Modifications (PTMs)



Comparison of *in vivo* and *in vitro* environment



A. Saghatelian & B.F. Cravatt **Assignment of protein function in the postgenomic era**
(2005) *Nature Chemical Biology* 1, 130.

UniProt (7) – PTMs

- A. Which percentage of the whole SwissProt database has PTMs?

- B. Go to the entry with Accession Number: P00533. Find out which PTMs are reported, and which are the evidence for them. Watch them also in the «feature viewer» mode.

Uniprot access from programs

Technical corner

Last modified March 13, 2015

[UniProtKB entry view manual](#)

User manual for the UniProtKB entry view.

[UniProtKB flat file manual](#)

User manual for the UniProtKB flat file format.

[FASTA headers](#)

Description of FASTA headers for UniProtKB (including alternative isoforms), UniRef, UniParc and archived UniProtKB versions.

[UniProtKB query fields](#)

List of supported fields for querying UniProtKB on this web site.

[Browser support](#)

Recommended list of major web browsers for the UniProt website.

[How can I access resources on this web site programmatically?](#)

FAQ with examples of web access using simple URLs ([REST](#)).

Parsing SwissProt with BioPython

<http://biopython.org/DIST/docs/api/Bio.SwissProt-pysrc.html>

```
#python code
import Bio.SwissProt as sp
example_filename = "./P01892.txt"
with open(example_filename) as handle:
    records = sp.parse(handle)
    for record in records:
        print(record.entry_name)
        print(",".join(record.accessions))
        print(record.keywords)
        print(repr(record.organism))
        print(record.sequence[:20] + "...")
```

1A02_HUMAN

P01892,O19619,P06338,P10313,P30444,P30445,P30446,P30514,Q29680,Q29837,Q29899,
Q95352,Q95380,Q9TPX8,Q9TPX9,Q9TPY0,Q9TQH5,Q9TQI3

['3D-structure', 'Complete proteome', 'Direct protein sequencing', 'Disulfide bond',
'Glycoprotein', 'Host-virus interaction', 'Immunity', 'Membrane', 'MHC I', 'Phosphoprotein',
'Polymorphism', 'Reference proteome', 'Signal', 'Transmembrane', 'Transmembrane helix',
'Ubl conjugation']

'Homo sapiens (Human).'

MAVMAPRTLVLSSGALALT...

Uniprot access from programs

https://www.ebi.ac.uk/training/online/course/accessing-em 80% Search Login/register

Training Train online About Train online Glossary Support and feedback

EMBL-EBI, programmatically: take a REST from manual searches

Introduction to EMBL-EBI resources
Introduction to programmatic access
Europe PMC, programmatically
Ensembl, programmatically
UniProt, programmatically
PDBe, programmatically
ChEMBL, programmatically
Sequence analysis tools, programmatically
EVA, programmatically
Summary
Your feedback
Contributors

UniProt, programmatically



UniProt aims to provide the scientific community with the highest quality, most comprehensive and most thoroughly annotated protein resource. Our mission includes curating detailed information on protein function and sequences including isoforms and disease variants. The UniProt knowledgebase contains millions of sequences, a subset of which have been curated by experts who critically review experimental and predicted data for each protein. The remainder are automatically annotated based on rule systems that rely on the expert curated knowledge. You can learn more about UniProt in [UniProt: Exploring protein sequence and functional information](#) or our [introductory webinar](#).

This webinar took place on 20th June 2017 and will introduce the UniProt REST API. In this webinar we will:

- Introduce the UniProt resources (1:12)
- Search the UniProt Knowledgebase (3:40)
- Convert between different database identifiers (11:43)
- Search with proteomics peptides (16:55)
- Show you where to get help (21:47)

UniProt (8) – Searches

⇒ **Questionnaire on UniProt #3**

Errors in Databases ?

Issue: Automatic Enzymes Functional Annotations in Biological Databases

For over a decade, the majority of sequences found in public databases have been annotated using computational prediction alone, following the principle that proteins with similar sequences and similar structures frequently carry out similar functions.

But protein sequence or structure similarity alone does not necessarily equate to cellular or molecular functional similarity.



Enzyme Function Less Conserved than Anticipated

Burkhard Rost^{1,2*}

¹CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York NY 10032, USA

²Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York

The level of sequence similarity that implies similarity in protein structure is well established. Recently, many groups proposed thresholds for similarity in sequence implying similarity in enzymatic function. All previous results suggest the strong conservation of enzymatic function above levels of 50% pairwise sequence identity. Here, I argue that all groups substantially overestimated the conservation of enzyme function because their data sets were either too biased, or too small. An unbiased analysis suggested that less than 30% of the pair fragments above 50% sequence identity have entirely identical EC numbers. Another surprising finding was that even BLAST *E*-values below 10^{-50} did not suffice to automatically transfer enzyme function without errors. As expected, most mis-

Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies

2009

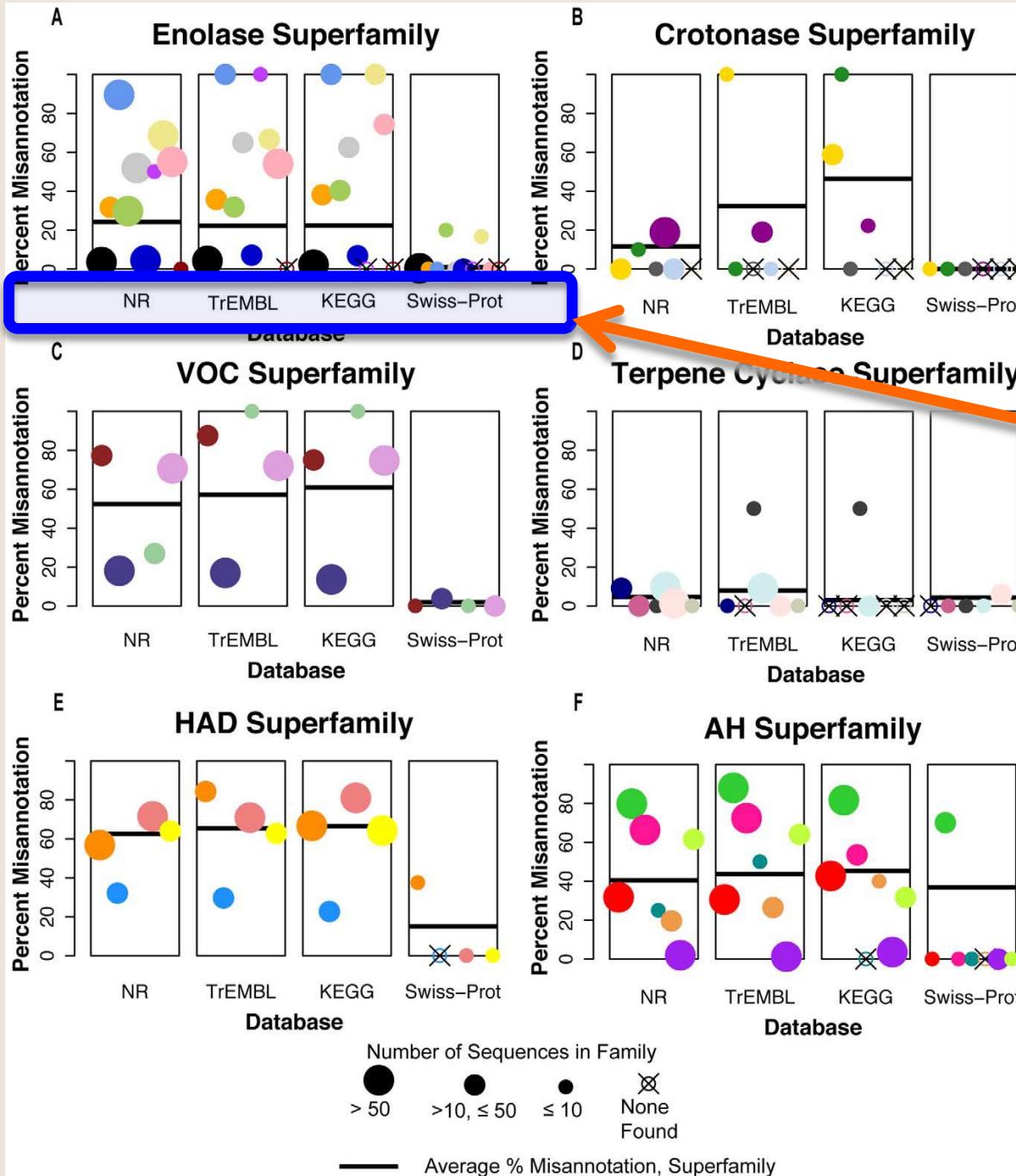
Alexandra M. Schnoes¹, Shoshana D. Brown², Igor Dodevski³, Patricia C. Babbitt^{2,4,5*}

1 Graduate Group in Biophysics, University of California San Francisco, San Francisco, California, United States of America, **2** Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, United States of America, **3** Department of Biochemistry, University of Zürich, Zürich, Switzerland, **4** Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, **5** California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, United States of America

Abstract

Due to the rapid release of new data from genome sequencing projects, the majority of protein sequences in public databases have not been experimentally characterized; rather, sequences are annotated using computational analysis. The level of misannotation and the types of misannotation in large public databases are currently unknown and have not been analyzed in depth. We have investigated the misannotation levels for molecular function in four public protein sequence databases (UniProtKB/Swiss-Prot, GenBank NR, UniProtKB/TrEMBL, and KEGG) for a model set of 37 enzyme families for which extensive experimental information is available. The manually curated database Swiss-Prot shows the lowest annotation error levels (close to 0% for most families); the two other protein sequence databases (GenBank NR and TrEMBL) and the protein sequences in the KEGG pathways database exhibit similar and surprisingly high levels of misannotation that average 5%–63% across the six superfamilies studied. For 10 of the 37 families examined, the level of misannotation in one or more of these databases is >80%. Examination of the NR database over time shows that misannotation has increased from 1993 to 2005. The types of misannotation that were found fall into several categories, most associated with “overprediction” of molecular function. These results suggest that misannotation in enzyme superfamilies containing multiple families that catalyze different reactions is a larger problem than has been recognized. Strategies are suggested for addressing some of the systematic problems contributing to these high levels of misannotation.

Percent misannotation in the tested families and superfamilies



Databases tested:

- GenBank NR (Non-Redundant)
- UniProtKB-TrEMBL
- KEGG (Kyoto Encyclopedia of Genes and Genomes)
- UniProtKB-SwissProt

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies**, PLoS Comput Biol. 2009

What about using conservation of known catalytic residues to improve the annotation of catalytic function?

[EBI Thornton Group CSA](#)

<https://www.ebi.ac.uk/thornton-srv/databases/CSA/>

Catalytic Site Atlas

Enter a PDB code, UniProKB code or EC number in one of the boxes Below to obtain catalytic residue details from the CSA.

Search The CSA

PDB ID	<input type="text"/>	<input type="button" value="SEARCH CSA"/>
UNIPROT ID	<input type="text"/>	<input type="button" value="SEARCH CSA"/>
EC Number	<input type="text"/>	<input type="button" value="SEARCH CSA"/>

A NEW VERSION OF THE CSA UPDATED 14th November 2013

Using the conservation of known catalytic residues has been shown to greatly improve coverage and accuracy in transferring annotations to low-sequence-similarity homologs, at sequence identity thresholds below 40%, though problems still exist.