Automatic annotation in UniProtKB using UniRule, and Complete Proteomes

Wei Mun Chan

EMBL

Talk outline

- Introduction to UniProt
- UniProtKB annotation and propagation
- Data increase and the need for Automatic Annotation
- Automatic annotation systems in UniProtKB
- UniRule Automatic Annotation System
- Complete Proteomes in UniProtKB

Talk outline

- **Introduction to UniProt**
- UniProtKB annotation and propagation
- Data increase and the need for Automatic Annotation
- Automatic annotation systems in UniProtKB
- UniRule Automatic Annotation System
- Complete proteomes in UniProtKB

UniProt Consortium

- Formed in 2002
- Previously known as “Swiss-Prot” since 1986

- UniProt group at the EBI is led by Claire Odonovan and Maria Jesus Martin, part of the PANDA proteins group led by Rolf Apweiler

- UniProt group at PIR, Georgetown University is led by Cathy Wu

- UniProt group at SIB (Geneva/Lausanne) is led by Ioannis Xenarios and Lydie Bougeleret (heirs to Amos Bairoch, left 2009)

- UniProtKB is UniProt KnowledgeBase, and includes TrEMBL and Swiss-Prot entries

www.uniprot.org

Search | Blast | Align | Retrieve | ID Mapping

Search in: Protein Knowledgebase (UniProtKB) — Query: Search | Advanced Search » | Clear

WELCOME

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.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval
- Database identifier mapping (ID Mapping)

UniProt

NEWS

UniProt release 2011_11 - Nov 16, 2011

Who wants to be a millionaire? The first million HAMAP-annotated entries in UniProtKB/TrEMBL | Cross-references to KO

› Statistics for UniProtKB: Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives

Follow @uniprot 130 followers

SITE TOUR

Learn how to make best use of the tools and data on this site.

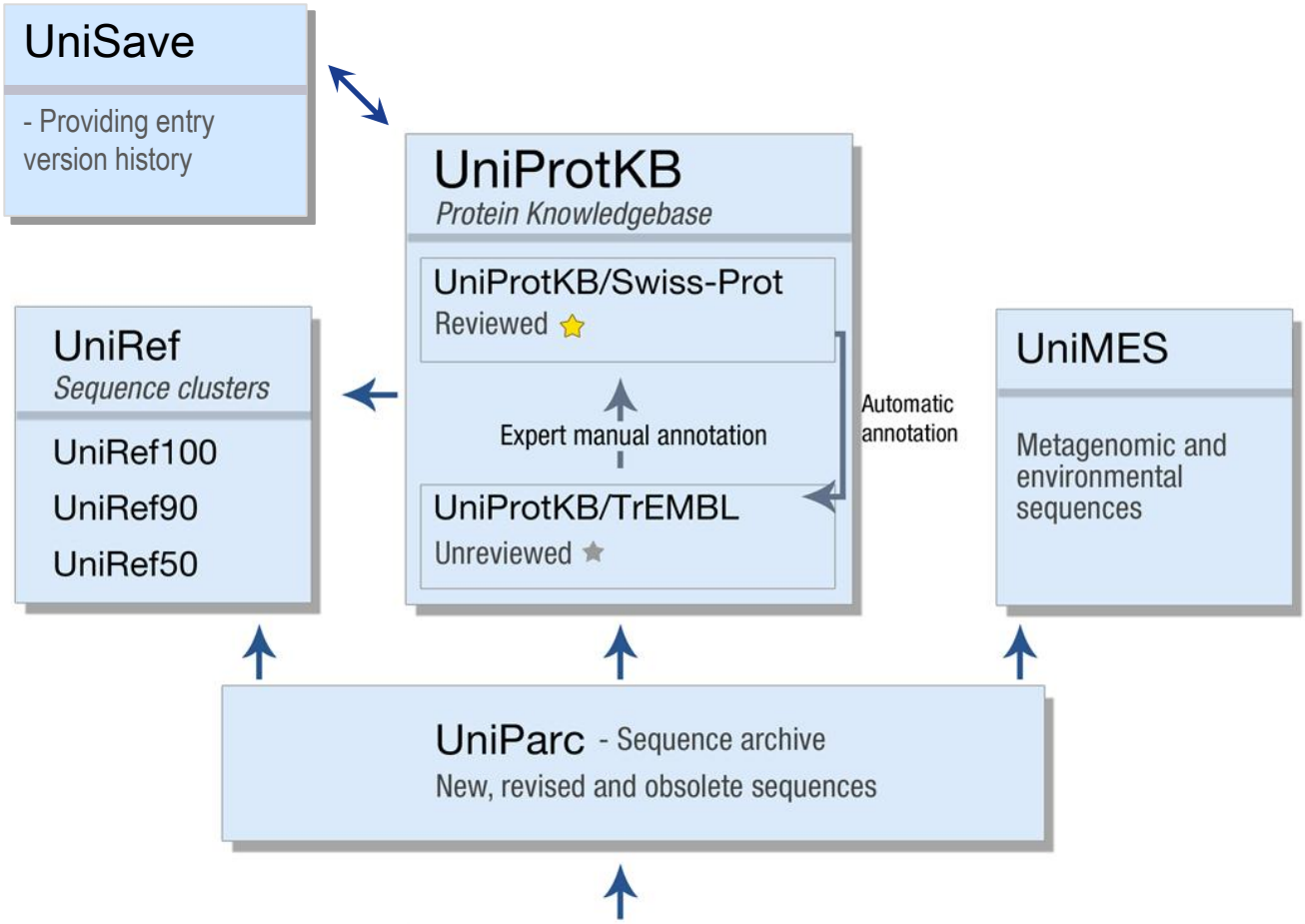
PROTEIN SPOTLIGHT

a missing sense
November 2011

We are reminded regularly of how fragile life is and how easily the subtle balance of our molecular make-up can be shifted and cause devastating effects. Deafness is one. Deafness can be brought about by a number of incidents...

EMBL-EBI PIR SIB

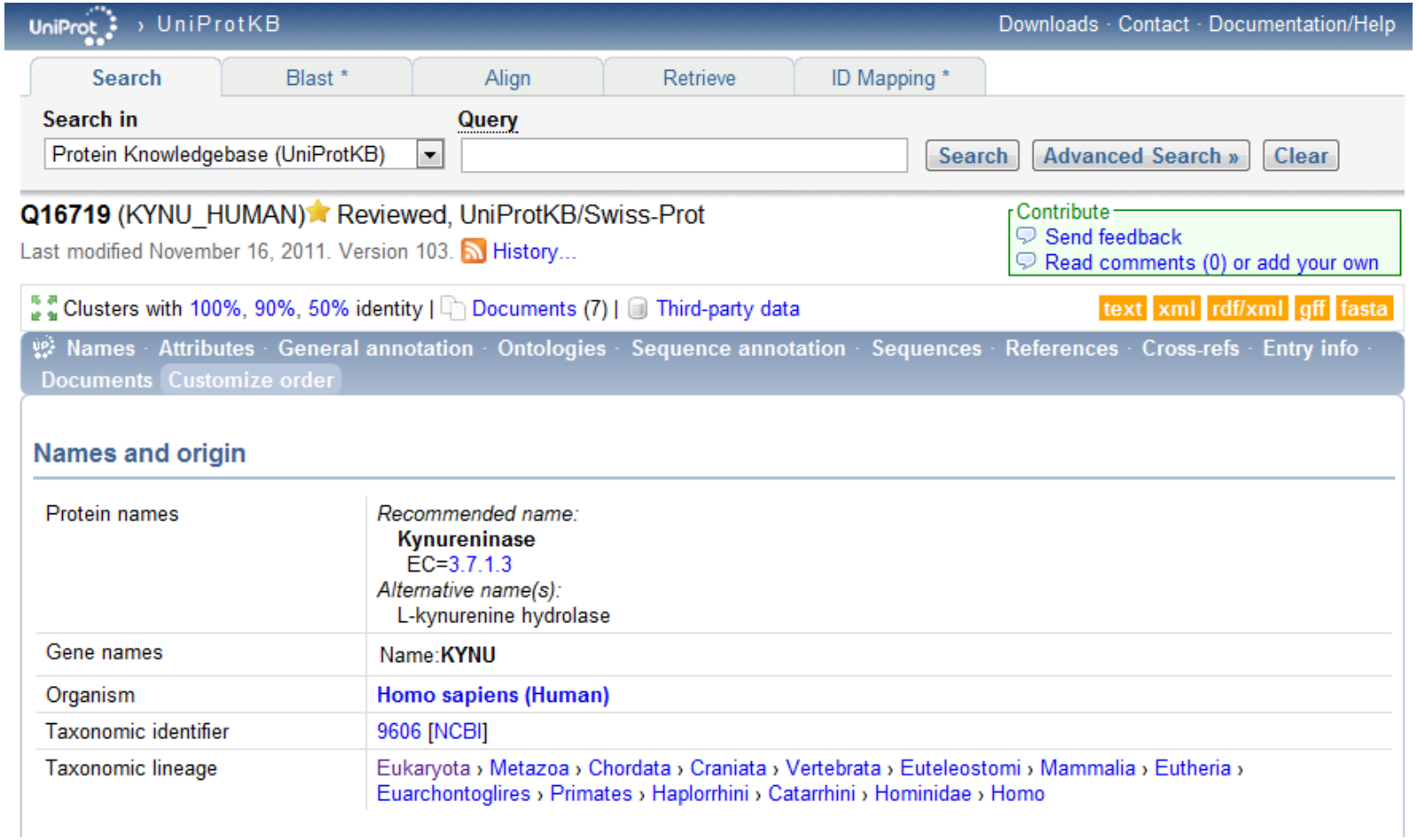
UniProt databases



Talk outline

- Introduction to UniProt
- **UniProtKB annotation and propagation**
- Data increase and the need for Automatic Annotation
- Automatic annotation systems in UniProtKB
- UniRule Automatic Annotation System
- Complete Proteomes in UniProtKB

UniProtKB annotation



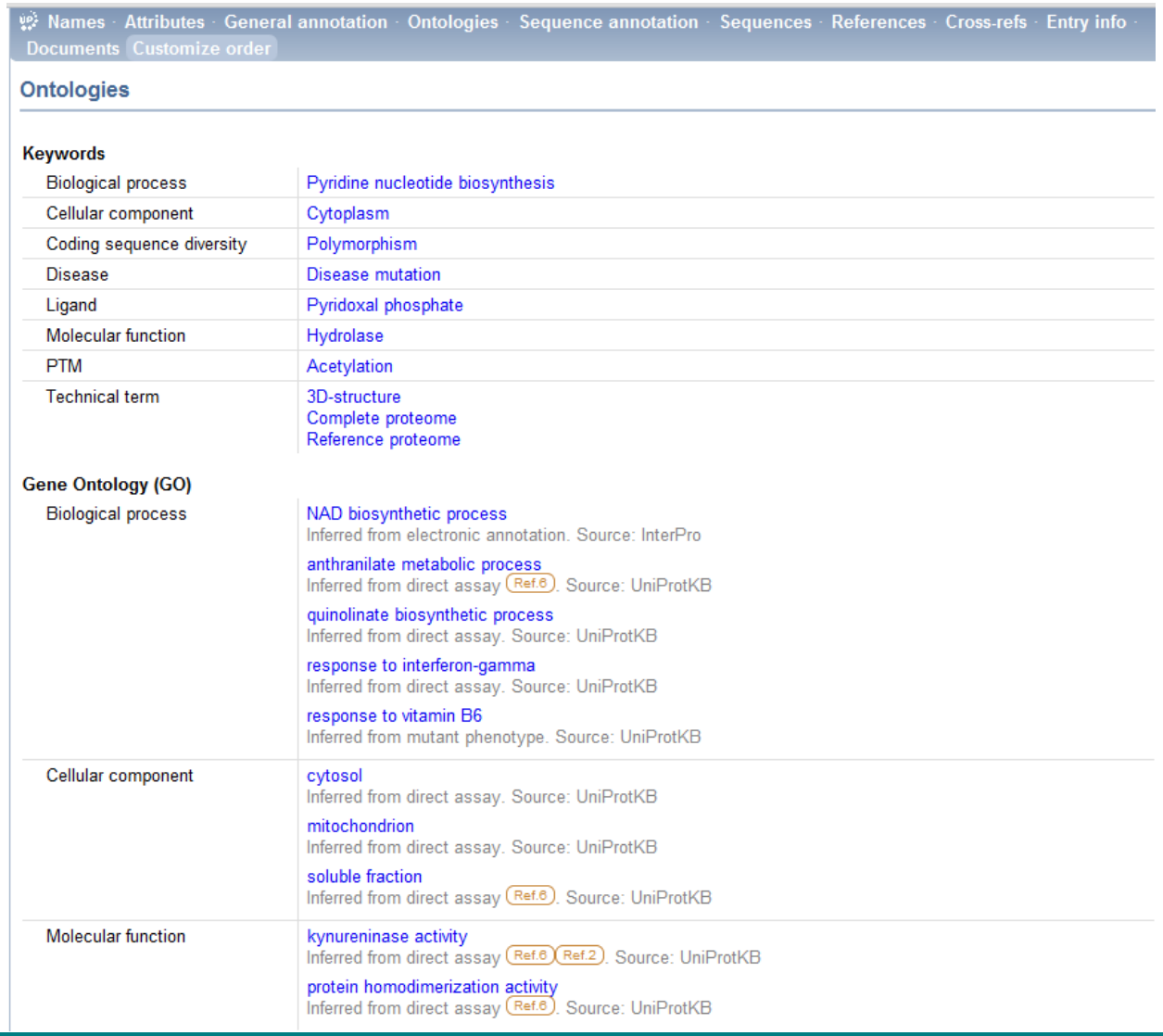
UniProtKB annotation

Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents · Customize order

General annotation (Comments)

Function	Catalyzes the cleavage of L-kynurenine (L-Kyn) and L-3-hydroxykynurenine (L-3OHKyn) into anthranilic acid (AA) and 3-hydroxyanthranilic acid (3-OHAA), respectively. Has a preference for the L-3-hydroxy form. Also has cysteine-conjugate-beta-lyase activity.
Catalytic activity	L-kynurenine + H_2O = anthranilate + L-alanine. Ref.1 Ref.2 Ref.6 Ref.7 L-3-hydroxykynurenine + H_2O = 3-hydroxyanthranilate + L-alanine. Ref.1 Ref.2 Ref.6 Ref.7
Cofactor	Pyridoxal phosphate.
Enzyme regulation	Inhibited by o-methoxybenzoylalanine (OMBA). Ref.1 Ref.2
Pathway	Amino-acid degradation; L-kynurenine degradation; L-alanine and anthranilate from L-kynurenine: step 1/1. Cofactor biosynthesis; NAD(+) biosynthesis; quinolinate from L-kynurenine: step 2/3.
Subunit structure	Homodimer. Ref.7
Subcellular location	Cytoplasm Ref.1.
Tissue specificity	Expressed in all tissues tested (heart, brain placenta, lung, liver, skeletal muscle, kidney and pancreas). Highest levels found in placenta, liver and lung. Expressed in all brain regions. Ref.1 Ref.2
Induction	Increased levels in several cerebral and systemic inflammatory conditions. Ref.1 Ref.2
Involvement in disease	Note=Xanthurenic aciduria manifesting as massive urinary excretion of large amounts of kynurenine, 3-hydroxykynurenine and xanthurenic acid has been observed in an individual carrying a homozygous missense change in KYNU (Ref.8). The urinary pattern in the patient suggests kynureninase deficiency and a block in the conversion of kynurenine and 3-hydroxykynurenine to anthranilate and 3-hydroxyanthranilate, respectively.
Sequence similarities	Belongs to the kynureninase family.
Caution	It has been reported that this enzyme possesses no measurable activity against L-kynurenine and is subject to inhibition by both L-kynurenine and D-kynurenine at pH 7.9 (Ref.6).
Biophysicochemical properties	Kinetic parameters: K_M=493 µM for L-kynurenine (at pH 7.0) Ref.1 Ref.2 Ref.6 Ref.7 K_M=28.3 µM for DL-3-hydroxykynurenine (at pH 7.0) K_M=3.0 µM for DL-3-hydroxykynurenine (at pH 7.9) pH dependence: Optimum pH is 8.25 with DL-3-hydroxykynurenine as substrate.
Mass spectrometry	Molecular mass is 52400 Da from positions 1 - 465. Determined by MALDI. The reported mass is given to only three significant figures. Ref.6

UniProtKB annotation



Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents Customize order

Ontologies

Keywords

Biological process	Pyridine nucleotide biosynthesis
Cellular component	Cytoplasm
Coding sequence diversity	Polymorphism
Disease	Disease mutation
Ligand	Pyridoxal phosphate
Molecular function	Hydrolase
PTM	Acetylation
Technical term	3D-structure Complete proteome Reference proteome

Gene Ontology (GO)

Biological process	NAD biosynthetic process Inferred from electronic annotation. Source: InterPro anthranilate metabolic process Inferred from direct assay Ref.6. Source: UniProtKB quinolinate biosynthetic process Inferred from direct assay. Source: UniProtKB response to interferon-gamma Inferred from direct assay. Source: UniProtKB response to vitamin B6 Inferred from mutant phenotype. Source: UniProtKB
Cellular component	cytosol Inferred from direct assay. Source: UniProtKB mitochondrion Inferred from direct assay. Source: UniProtKB soluble fraction Inferred from direct assay Ref.6. Source: UniProtKB
Molecular function	kynureninase activity Inferred from direct assay Ref.6 Ref.2. Source: UniProtKB protein homodimerization activity Inferred from direct assay Ref.6. Source: UniProtKB

UniProtKB annotation

Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents Customize order

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
☐	Chain	1 – 465	465	Kynureninase		PRO_0000218657
Regions						
☐	Region	165 – 168	4	Pyridoxal phosphate binding		
Sites						
☐	Binding site	137	1	Pyridoxal phosphate; via amide nitrogen (By similarity)		
☐	Binding site	138	1	Pyridoxal phosphate		
☐	Binding site	250	1	Pyridoxal phosphate		
☐	Binding site	253	1	Pyridoxal phosphate		
☐	Binding site	275	1	Pyridoxal phosphate		
☐	Binding site	305	1	Pyridoxal phosphate		
☐	Binding site	333	1	Pyridoxal phosphate		
Amino acid modifications						
☐	Modified residue	1	1	N-acetylmethionine (By similarity)		
☐	Modified residue	276	1	N6-(pyridoxal phosphate)lysine		
Natural variations						
☐	Natural variant	188	1	R → Q. [dbSNP:rs2304705]		VAR_049724
☐	Natural variant	198	1	T → A Detected in a boy with xanthurenic aciduria. (Ref.8)		VAR_054401
☐	Natural variant	412	1	K → E. [dbSNP:rs9013] (Ref.3)		VAR_022092

Secondary structure

1 … 465

Helix Strand Turn

Details...

Propagation of annotation in UniProtKB

General annotation

Annotation	Propagated
RecName	Yes
AltName	Yes
Function	Yes
Catalytic activity	Yes
Pathway	Yes
Subunit	Yes
Subcellular location	Yes
Disease	No
Disruption phenotype	No
Polymorphism	No
Alternative products	No

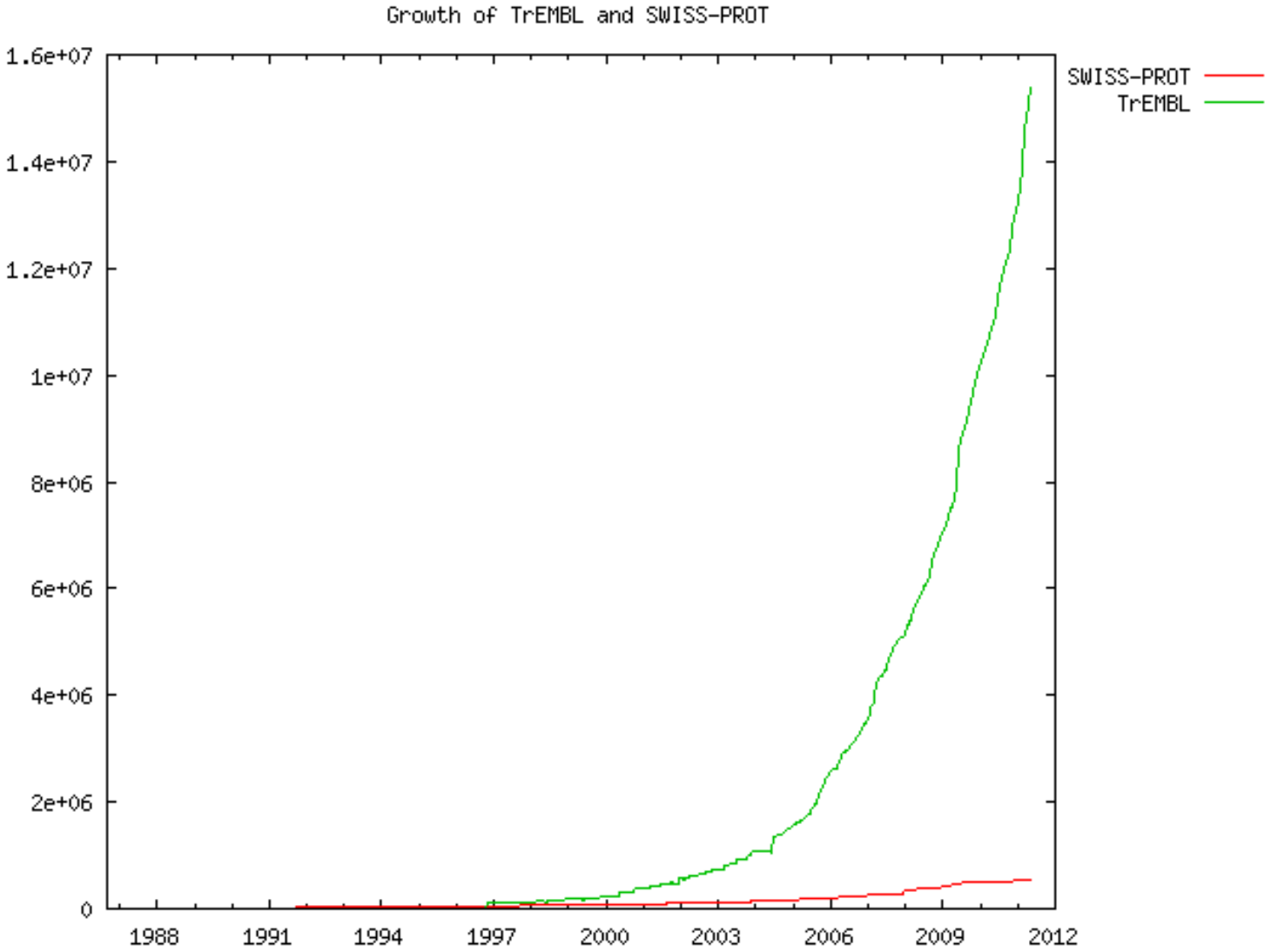
Feature annotation

Annotation	Propagated
KW	Yes
GO	Yes
Regions of interest	Yes
Active site	Yes
Ligand-binding	Yes
Processing	Yes
PTMs	Yes
Ambiguities	No
Conflicts	No
Natural variants	No
Isoforms	No

Talk outline

- Introduction to UniProt
- UniProtKB annotation and propagation
- **Data increase and the need for Automatic Annotation**
- Automatic annotation systems in UniProtKB
- UniRule Automatic Annotation System
- Complete Proteomes in UniProtKB

Data increase in UniProtKB



Benefits of Automatic Annotation

- Added value for TrEMBL in the face of rapid data growth
  - many species/proteins without published experimental data
- Support for manual curation
  - making manual curation of TrEMBL entries for which there is published data easier
- Correction of misleading annotation in data received from sequencing centres
- Highlighting of patterns
  - knowledge that can be/needs to be propagated across the databases
  - inconsistent annotation e.g. of a protein family

Talk outline

- Introduction to UniProt
- UniProtKB annotation and propagation
- Data increase and the need for Automatic Annotation
- **Automatic annotation systems in UniProtKB**
- UniRule Automatic Annotation System
- Complete Proteomes in UniProtKB

Automatic Annotation in UniProtKB/TrEMBL

We have implemented Automatic Annotation systems based on **annotation rules**

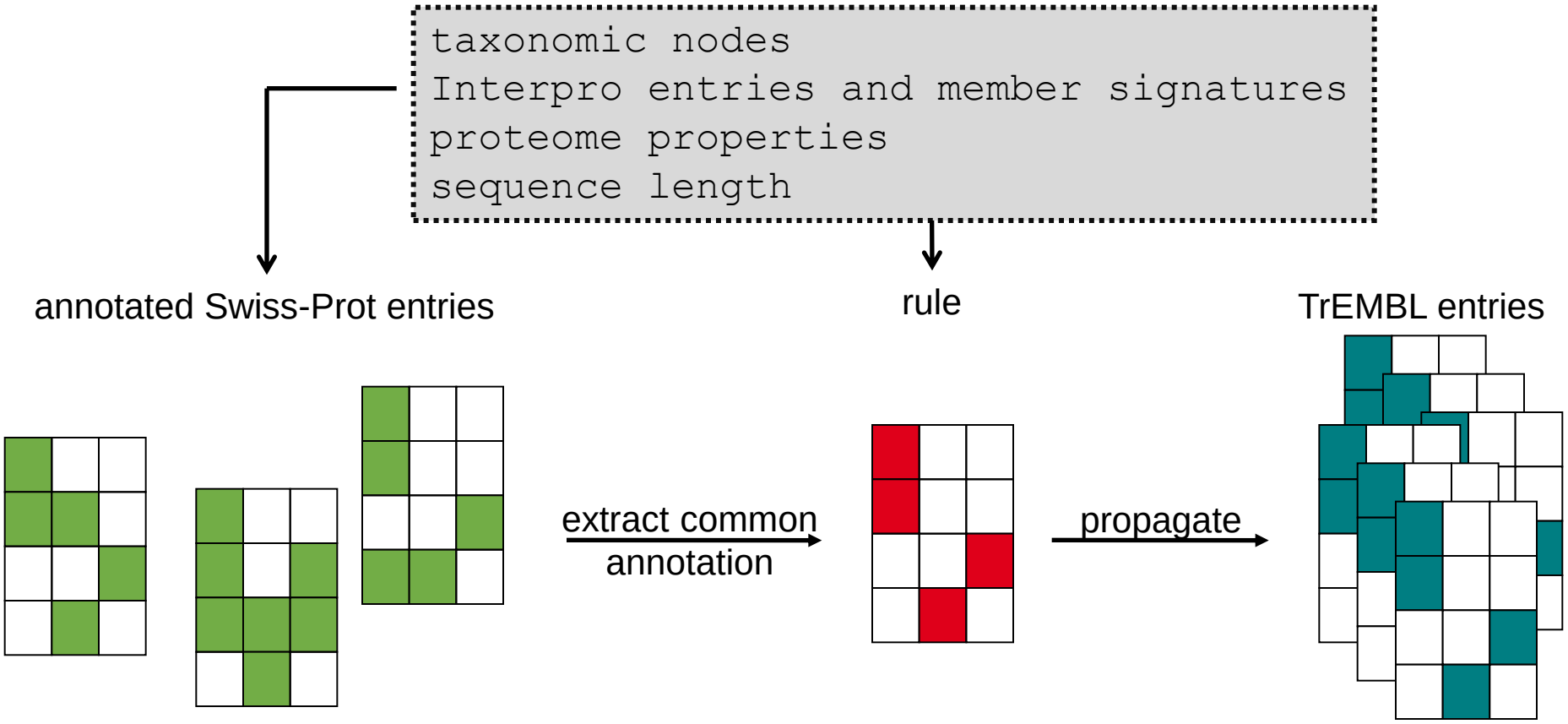
- Rules are linked to specific signatures - **InterPro**
- Annotation rules have
  - annotations
  - conditions
- Rules are tested and validated against UniProtKB/Swiss-Prot
- Rules and annotations are updated each UniProt release

Automatic Annotation Systems in UniProtKB

System	Rule creation	Trigger	Annotations	Scope
SAAS	automatic	taxonomy InterPro	comments, KW	all taxa
UniRule (Rulebase/HAMAP/ PIRNR/PIRSR)	**manual**	taxonomy InterPro* proteome property sequence length	protein names, comments, features, KW, GO terms	all taxa

* Flexibility to create custom signatures and submitted to InterPro as required

Principle of an Annotation Rule Creation



TrEMBL entries remain in TrEMBL, but offer more (predicted) annotation

SAAS – Statistically Automatic Annotation System

- Automatically generated annotation rule system to supplement the labour intensive UniRule system
- Employs a C4.5 decision-tree algorithm to find the most concise rule

Talk outline

- Introduction to UniProt
- UniProtKB annotation and propagation
- Data increase and the need for Automatic Annotation
- Automatic annotation systems in UniProtKB
- **UniRule Automatic Annotation System**
- Complete Proteomes in UniProtKB

UniRule Automatic Annotation System

- Manually created/curated rules of varying complexity: annotation varies from simple Keyword attribution to complete annotation

- Sources for rule creation

  - automatically generated SAAS rules as input

  - literature based curation of characterised families – as a potential source for creating new signatures for a specific functional group

  - also …

UniRule - conditions used to created a rule

Conditions (can be positive or negative)

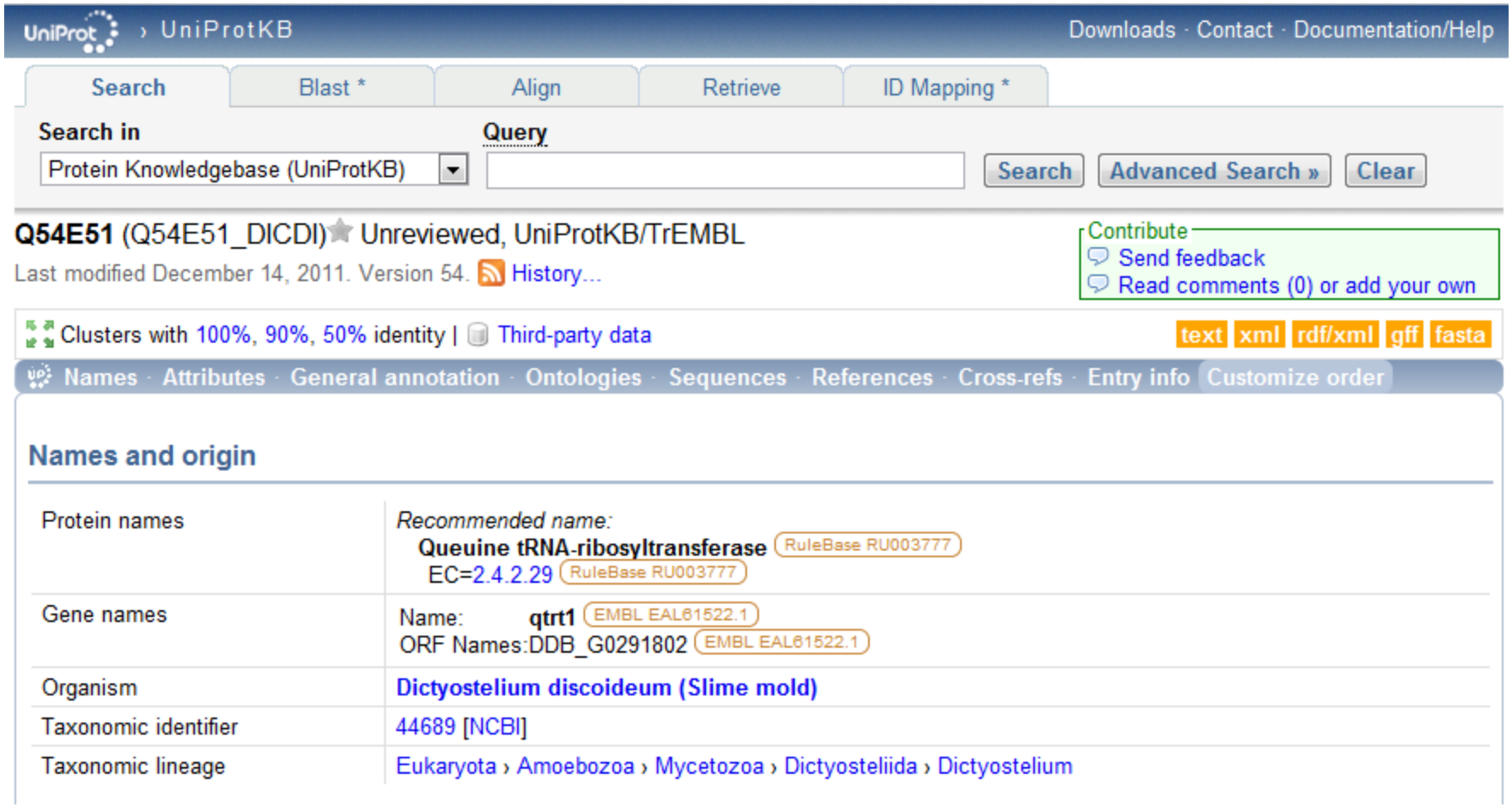
- Taxonomy
- InterPro entries and member signatures
- Subcellular location e.g. organelles
- Proteome properties e.g. photosynthetic
- Sequence length

UniRule – UniProtKB annotations defined in a rule

Annotations

- Description lines
  - Protein names
  - EC numbers
- Gene names
- General annotation (comments)
- UniProtKB Keywords
- GO terms

UniRule – output with evidence attribution



UniRule – output with evidence attribution

General annotation (Comments)

Function	Exchanges the guanine residue with 7-aminomethyl-7-deazaguanine in tRNAs with GU_N anticodons (tRNA-Asp, -Asn, -His and -Tyr). After this exchange, a cyclopentendiol moiety is attached to the 7-aminomethyl group of 7-deazaguanine, resulting in the hypermodified nucleoside queuosine (Q) (7-(((4,5-cis-dihydroxy-2-cyclopenten-1-yl)amino)methyl)-7-deazaguanosine) By similarity. RuleBase RU003777
Catalytic activity	[tRNA]-guanine + queuine = [tRNA]-queuine + guanine. RuleBase RU003777
Cofactor	Binds 1 zinc ion per subunit By similarity. RuleBase RU003777
Sequence similarities	Belongs to the queuine tRNA-ribosyltransferase family. RuleBase RU003777

Ontologies

Keywords

Biological process	Queuosine biosynthesis RuleBase RU003777 tRNA processing RuleBase RU003777
Ligand	Zinc RuleBase RU003777
Molecular function	Glycosyltransferase RuleBase RU003777 Transferase
Technical term	Complete proteome Reference proteome

UniRule – predictions

Predictions

Proteinname	Queuine tRNA-ribosyltransferase Predicted 2.4.2.29 Predicted
Comment	Exchanges the guanine residue with 7-aminomethyl-7-deazaguanine in tRNAs with GU(N) anticodons (tRNA-Asp, -Asn, -His and -Tyr). After this exchange, a cyclopentendiol moiety is attached to the 7-aminomethyl group of 7-deazaguanine, resulting in the hypermodified nucleoside queuosine (Q) (7-(((4,5-cis-dihydroxy-2-cyclopenten-1-yl)amino)methyl)-7-deazaguanosine) (By similarity) Predicted [tRNA]-guanine + queuine = [tRNA]-queuine + guanine Predicted Belongs to the queuine tRNA-ribosyltransferase family Predicted Binds 1 zinc ion per subunit (By similarity) Predicted
Keyword	Queuosine biosynthesis Predicted Transferase Predicted Zinc Predicted Glycosyltransferase Predicted tRNA processing Predicted

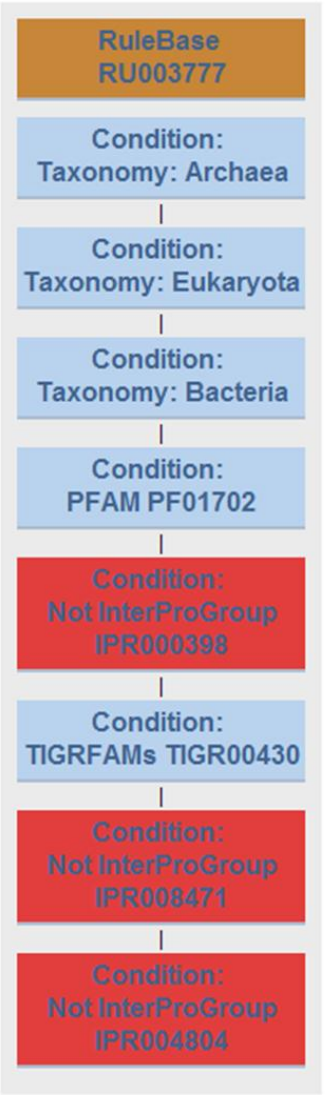
UniRule – prediction rules

Predictions

Proteinname	Queuine tRNA-ribosyltransferase Predicted 2.4.2.29 Predicted
Comment	Exchanges the guanine residue with 7-aminomethyl-7-deazaguanine in -His and -Tyr). After this exchange, a cyclopentendiol moiety is attache deazaguanine, resulting in the hypermodified nucleoside queuosine (Q) yl)amino)methyl)-7-deazaguanosine) (By similarity) Predicted [tRNA]-guanine + queuine = [tRNA]-queuine + guanine Predicted Belongs to the queuine tRNA-ribosyltransferase family Predicted Binds 1 zinc ion per subunit (By similarity) Predicted
Keyword	Queuosine biosynthesis Predicted Transferase Predicted Zinc Predicted Glycosyltransferase Predicted tRNA processing Predicted

RNA-Asp, -Asn,
7-
nten-1-

Prediction rules



Talk outline

- Introduction to UniProt
- UniProtKB annotation and propagation
- Data increase and the need for Automatic Annotation
- Automatic annotation systems in UniProtKB
- UniRule Automatic Annotation System
- **Complete Proteomes in UniProtKB**

How does UniProt define a Complete proteome?

- A complete proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

Status of complete proteomes in UniProt

- Longstanding project, 2902 proteomes that are spread over the entire taxonomic range
  - Archaea
  - Bacteria
  - Eukaryota
  - Viruses
- Capture of “Complete proteome” data is a mixture of automatic and manual procedures
- Aim is to provide a set of UniProtKB entries that define the proteome

Human complete proteome

- First draft of the complete human proteome available in UniProtKB/Swiss-Prot in September 2008
- The first mammalian proteome to be annotated
- Representing approximately 20,000 putative protein-coding genes each represented by one canonical sequence

Other complete proteomes

Human not the only organism to have its proteome annotated

- Sus scrofa (Pig) – 19,576 entries
- Gallus gallus (Chicken) – 21,622 entries
- Mus musculus (Mouse) – 46,656 entries
- Arabidopsis thaliana (Mouse-ear cress) - 32,521 entries

Challenges of proteome data

- How to define a complete genome, what is complete? Does it have a complete set of gene model annotations?
- Track any changes in the genome annotations and the impact on UniProt
- Gather all proteomes available, develop import pipelines to improve species coverage, current sources include:
  - INSDC species
  - Ensembl species
- UniProtKB also define a subset of the Complete proteomes as being 'Reference proteomes'.
  - Complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

Obtaining Proteomes

UniProt — Downloads · Contact · Documentation/Help

Search | Blast | Align | Retrieve | ID Mapping

Search in: Protein Knowledgebase (UniProtKB) — Query — Search | Advanced Search » | Clear

WELCOME

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.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: Swiss-Prot, which is manually annotated and reviewed. TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval
- Database identifier mapping (ID Mapping)

NEWS

UniProt release 2011_11 - Nov 16, 2011

Who wants to be a millionaire? The first million HAMAP-annotated entries in UniProtKB/TrEMBL | Cross-references to KO

› Statistics for UniProtKB: Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives

Follow @uniprot 127 followers

SITE TOUR

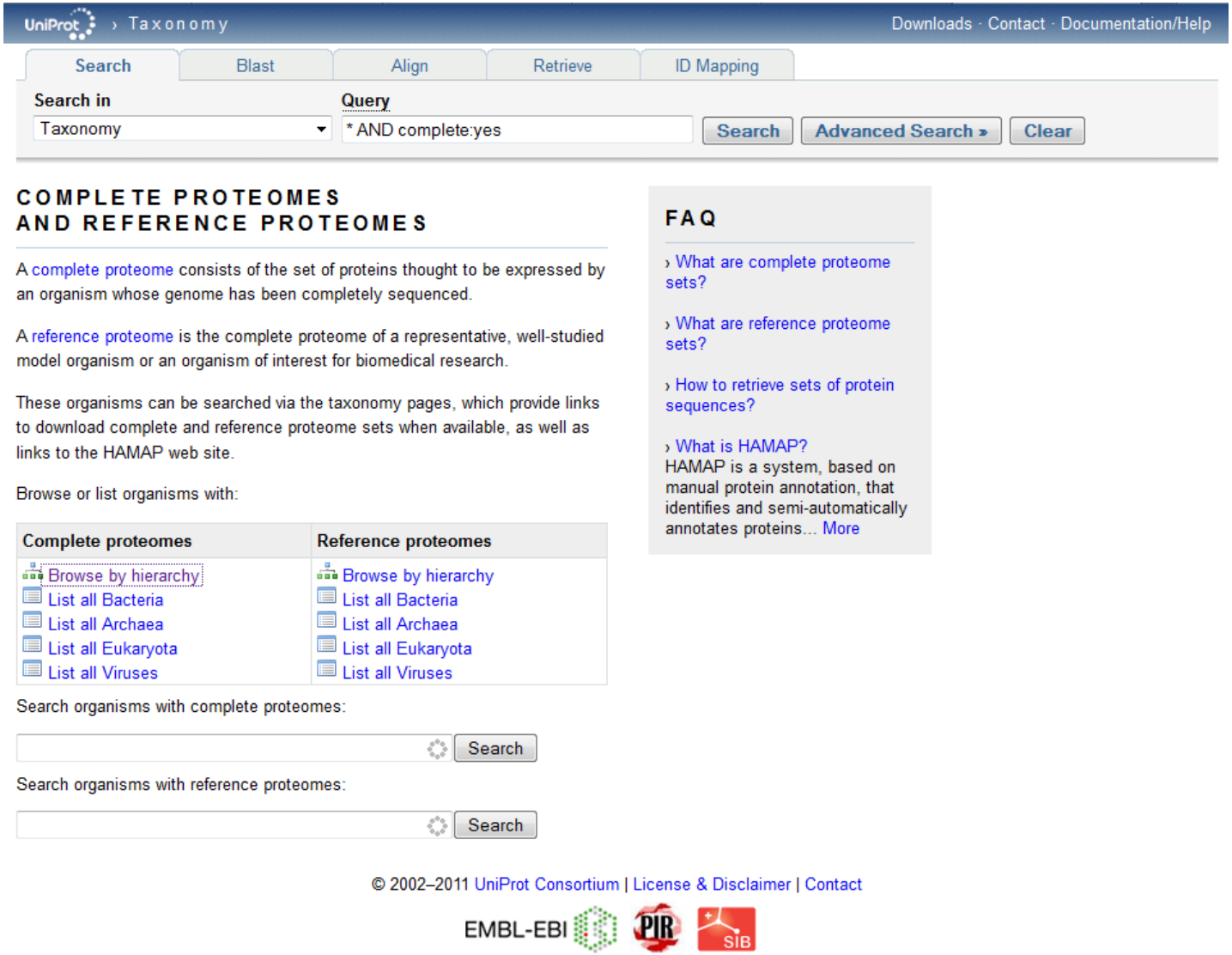
Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT

a missing sense
November 2011

We are reminded regularly of how fragile life is and how easily the subtle balance of our molecular make-up can be shifted and cause devastating effects. Deafness is one. Deafness can be brought about by a number of incidents…

Obtaining Proteomes



Closing remarks

- Manual annotation cannot keep pace with current or future rates of growth of UniProtKB so there is a need for automatic annotation

- UniProtKB currently uses two automatic annotation systems referred to as SAAS and UniRule

- Automatic annotation of TrEMBL is **refreshed** and **validated** using UniProtKB/Swiss-Prot as a reference, each UniProtKB release

Closing remarks

- UniRule – manually annotated rules
  - annotation varies from simple keywords to full annotation
  - starting from SAAS rules, InterPro signatures, literature-based curation of protein families
  - possibility to create custom signatures for InterPro

- Evidence attribution - users to determine the composition of the rule behind predicted annotation

Closing remarks

- Requirements for completed proteomes
  - Completely sequenced genome
  - Good gene prediction models
  - Good quality transcriptome/proteome data
  - Proteins are mapped to genome

Acknowledgements

- UniProt group at the EBI is led by **Claire Odonovan** and **Maria Jesus Martin**, part of the PANDA proteins group led by **Rolf Apweiler**

- UniProt group at PIR, Georgetown University is led by **Cathy Wu**

- UniProt group at SIB (Geneva/Lausanne) is led by **Ioannis Xenarios and Lydie Bougeleret** (heirs to Amos Bairoch, left 2009)

- Thanks also to all **curators**, **developers** and **support staff** at all three sites

Funding

- National Institutes of Health (NIH)
- European Commission (EC)'s SLING
- Swiss Federal Government through the Federal Office of Education and Science
- GEN2PHEN
- MICROME
- National Science Foundation (NSF)