Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Automatic Annotation in UniProtKB

Alan Bridge, Automation and Enhancements, Swiss-Prot group

alan.bridge@isb-sib.ch

Swiss Institute of Bioinformatics

Fourth International Biocuration Conference,
Tokyo, Japan
October 12th 2010

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Talk outline

- Introduction
- What?

  What we annotate in UniProtKB, and what we propagate
- Why?

  Need for automatic annotation in UniProtKB
- How?

  Automatic annotation systems for UniProtKB
- Current status and future plans

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Talk outline

- **Introduction**
- What?

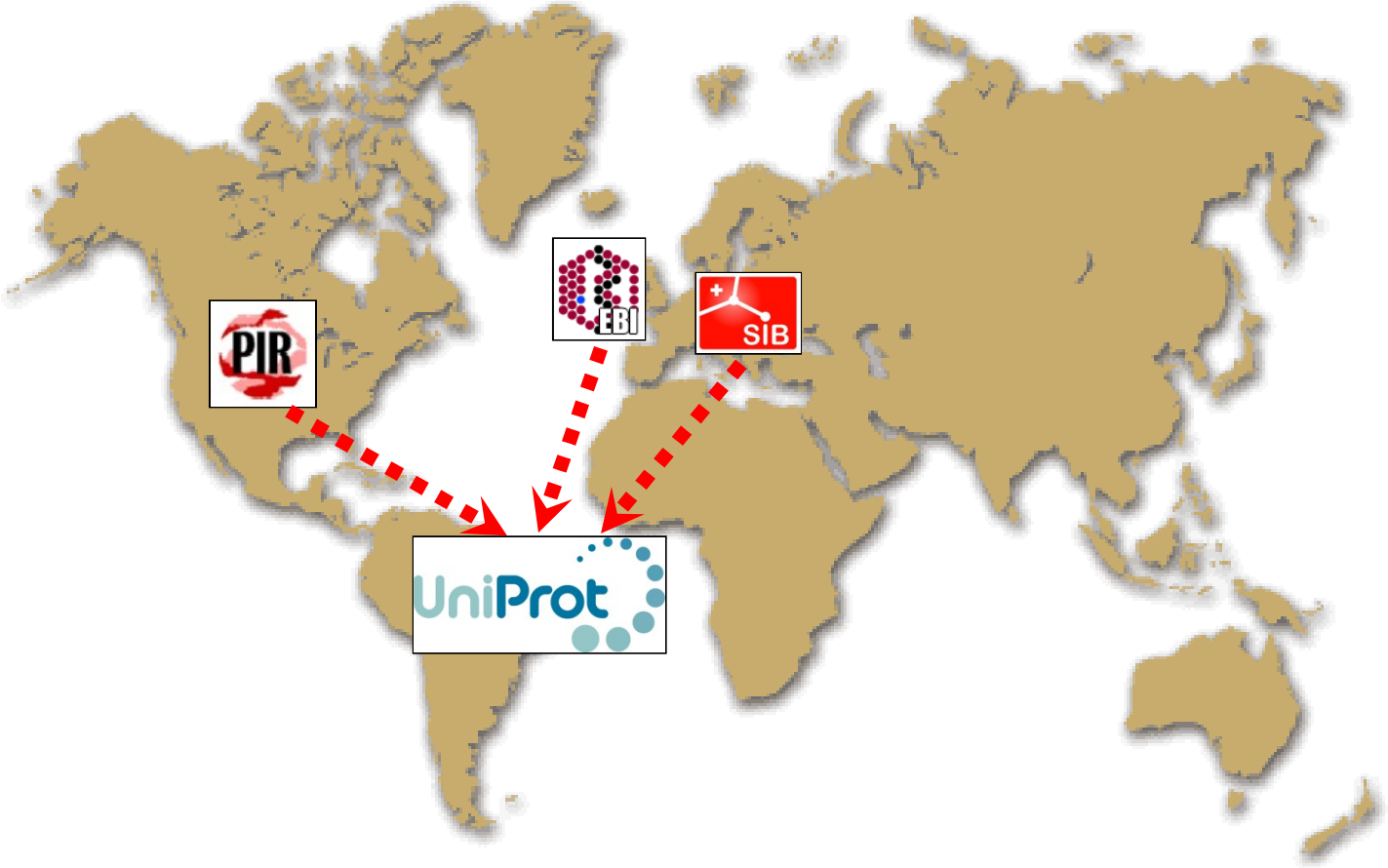
  What we annotate in UniProtKB, and what we propagate
- Why?

  Need for automatic annotation in UniProtKB
- How?

  Automatic annotation systems for UniProtKB
- Current status and future plans

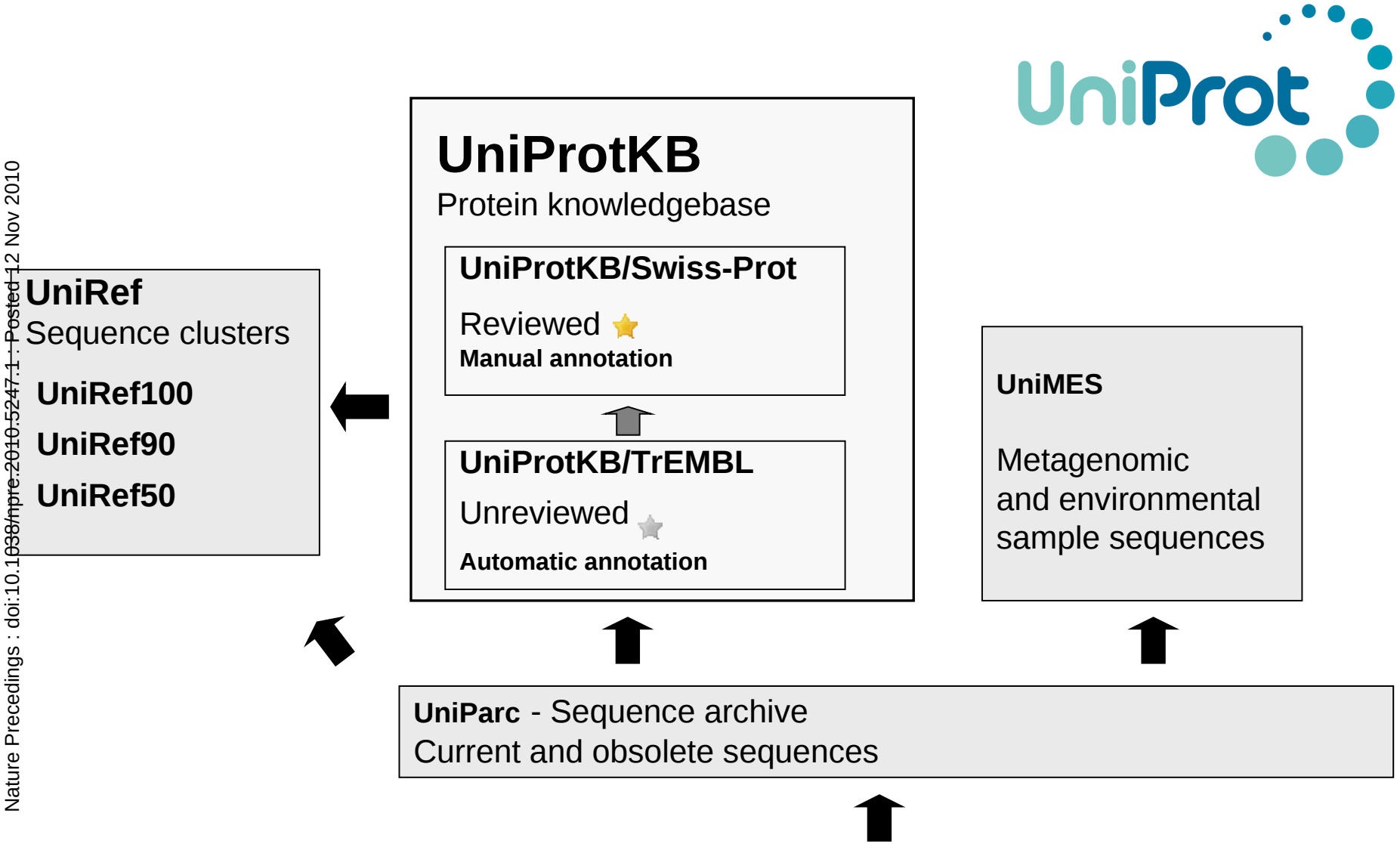
Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

The UniProt consortium



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**.

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Talk outline

- Introduction
- **What?**

  What we annotate in UniProtKB, and what we propagate
- Why?

  Need for automatic annotation in UniProtKB
- How?

  Automatic annotation systems for UniProtKB
- Current status and future plans

What? UniProtKB annotation content

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

What? UniProtKB annotation content

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

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	Defects in KYNU may be a cause of hydroxykynureninuria [MIM:236800]; also known as kynureninase deficiency. Hydroxykynureninuria is characterized by urinary excretion of large amounts of kynurenine, 3-hydroxykynurenine and xanthurenic acid. This suggests a block in the conversion of kynurenine and 3-hydroxykynurenine to anthranilate and 3-hydroxyanthranilate, respectively, which leads to impaired niacin biosynthesis. Hydroxykynureninuria can be associated with psychomotor retardation and non-progressive encephalopathy. Ref.8
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:

What? UniProtKB annotation content

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Ontologies

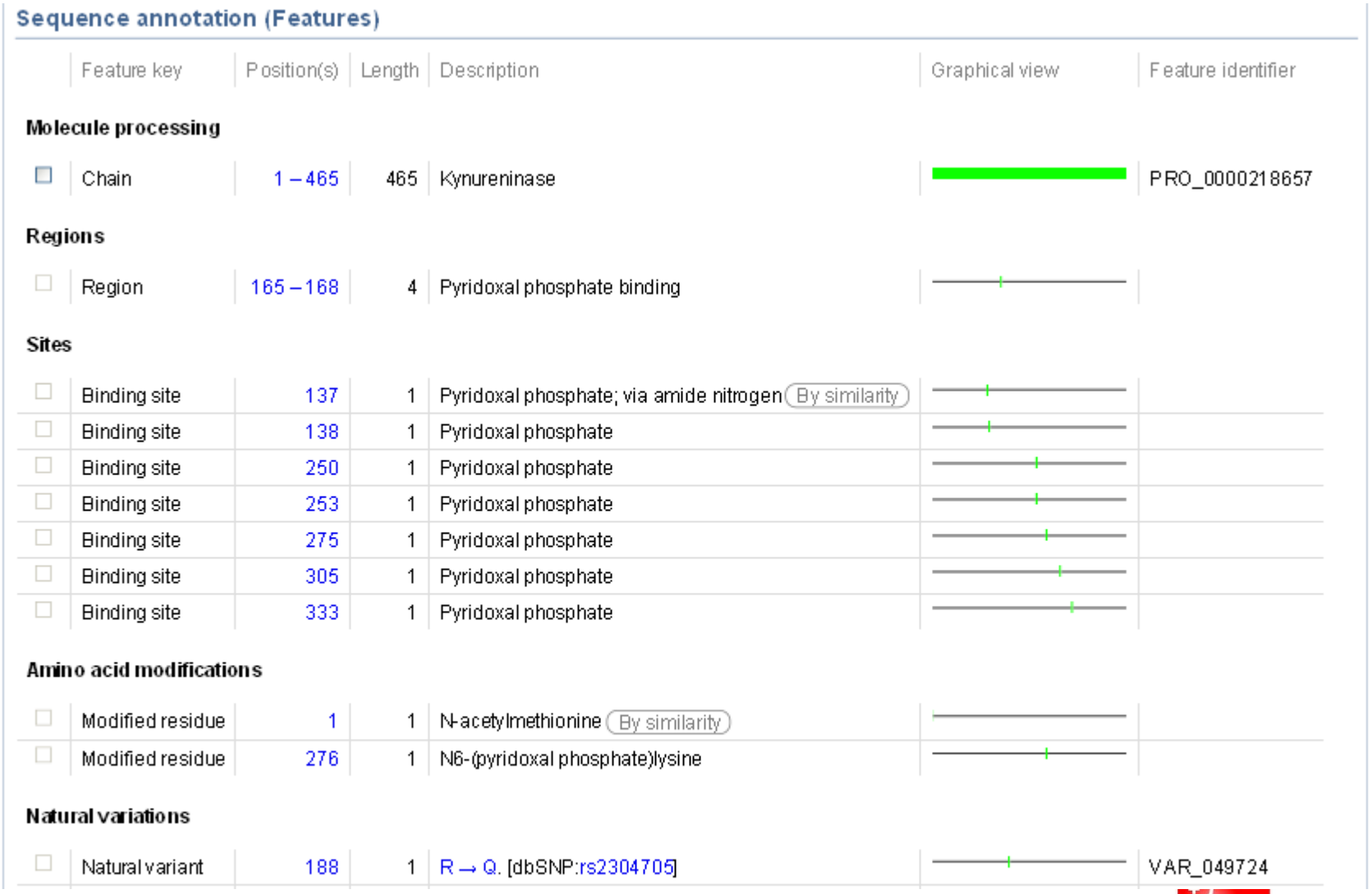
Keywords

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

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

What? UniProtKB annotation content



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

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

What? Annotation propagation in UniProtKB

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

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

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Talk outline

- Introduction
- What?

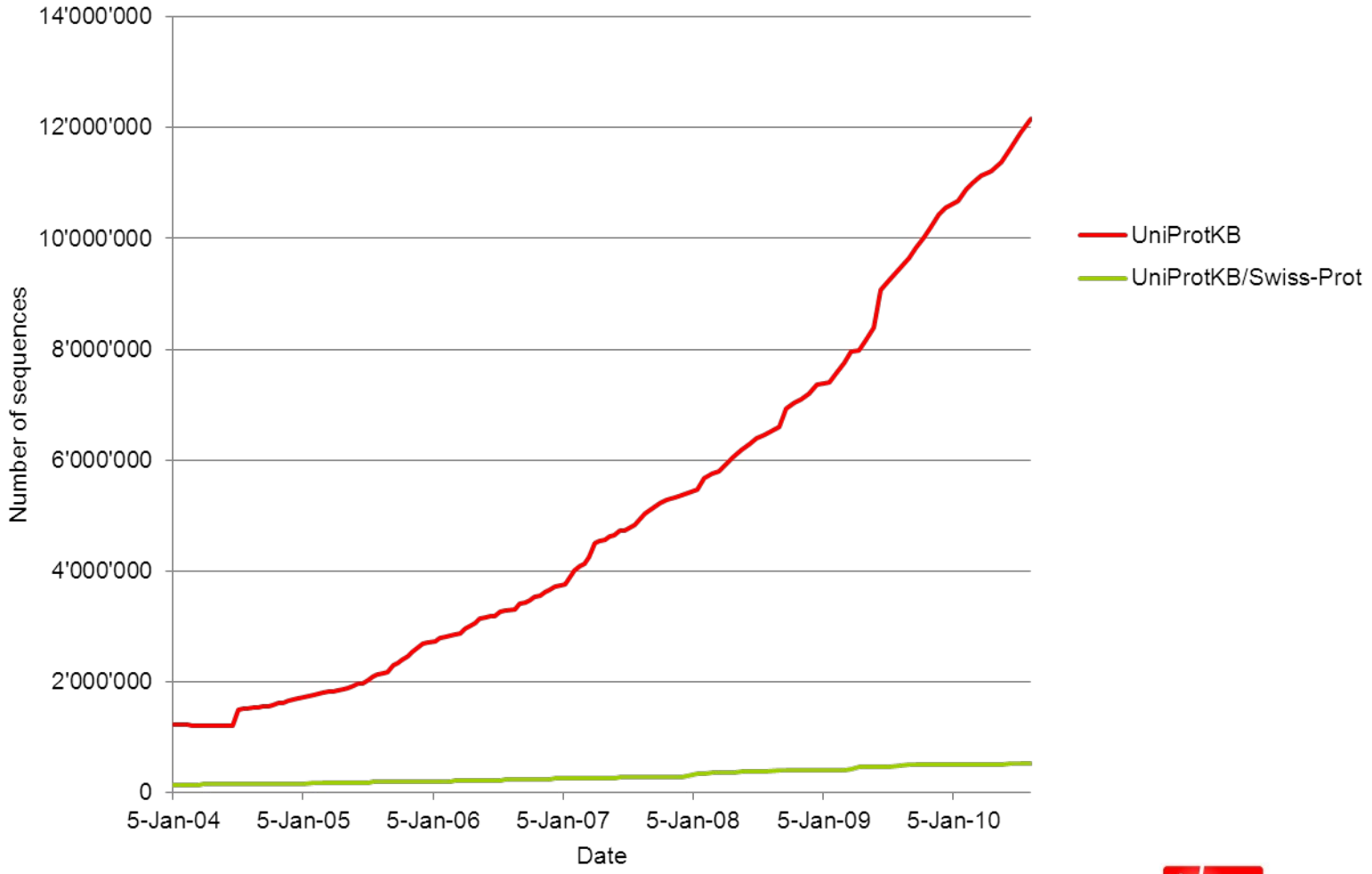
  What we annotate in UniProtKB, and what we propagate
- **Why?**

  Need for automatic annotation in UniProtKB
- How?

  Automatic annotation systems for UniProtKB
- Current status and future plans

Why? Data increase in UniProtKB

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Talk outline

- Introduction
- What?

  What we annotate in UniProtKB, and what we propagate
- Why?

  Need for automatic annotation in UniProtKB
- **How?**

  Automatic annotation systems for UniProtKB
- Current status and future plans

How? Automatic annotation in UniProtKB/TrEMBL

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

- 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 UniProtKB release

How? Automatic annotation in UniProtKB

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

System		Rule creation	Trigger	Annotations	Scope
SAAS (Spearmint)		automatic	InterPro	comments, KW	all taxa
UniRules	RuleBase	manual	InterPro	protein names, comments, KW	all taxa
	PIR name / site rules	manual	PIRSF*	protein names, comments, features, KW	all taxa
	HAMAP	manual	HAMAP*	protein names, comments, features, KW, GO	Archaea Bacteria plastids

* Signatures produced by UniProt consortium members and submitted to InterPro

How? Automatic annotation in UniProtKB

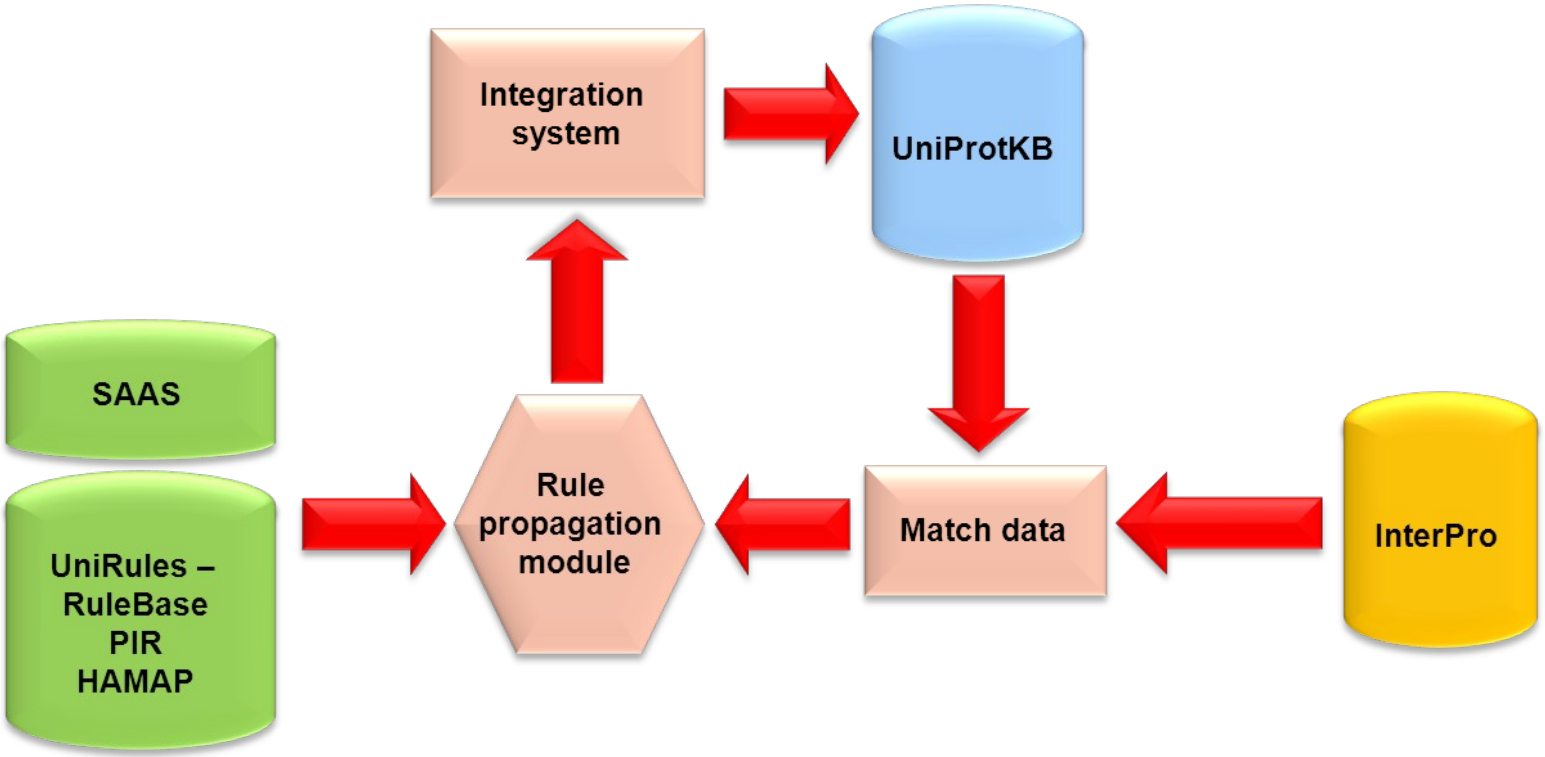
Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

System	Rule creation	Trigger	Annotations	Scope
SAAS	automatic	InterPro	comments, KW	all taxa
UniRules	manual	InterPro*	protein names, comments, features, KW, GO terms	all taxa

* Flexibility to create custom signatures for InterPro as required

How? Automatic annotation in UniProtKB

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



How? SAAS – C4.5 algorithm

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Proteins	Attributes			Annotations
	Prosite PS00847	Pfam PF01493	Mammalia	FAD
Q9ZN27	-	yes	-	yes
Q9JI00	-	-	yes	-
Q9ZL14	yes	-	-	-
Q9NYQ3	-	-	yes	-
Q9UJM8	-	-	yes	-
Q9NYQ2	-	-	yes	-
Q43155	-	yes	-	yes
Q9T0P4	-	yes	-	yes
Q9WU19	-	-	yes	-

- At each node of the tree, C4.5 chooses one attribute of the entries that most effectively splits them into subsets enriched in the desired annotation.
- The split criterion used by C4.5 is based on Shannon entropy.

How? SAAS – C4.5 algorithm

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

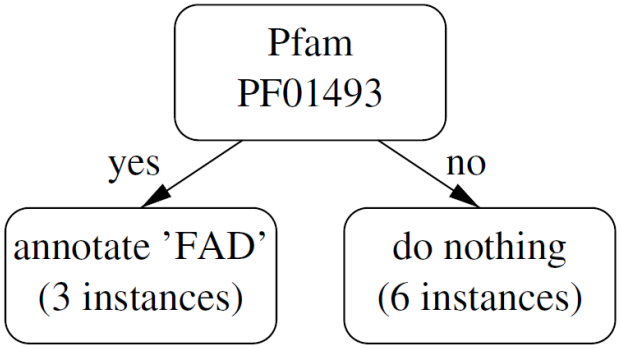
Proteins	Attributes			Annotations
	Prosite PS00847	Pfam PF01493	Mammalia	FAD
Q9ZN27	-	yes	-	yes
Q9JI00	-	-	yes	-
Q9ZL14	yes	-	-	-
Q9NYQ3	-	-	yes	-
Q9UJM8	-	-	yes	-
Q9NYQ2	-	-	yes	-
Q43155	-	yes	-	yes
Q9T0P4	-	yes	-	yes
Q9WU19	-	-	yes	-

- At each node of the tree, C4.5 chooses one attribute of the entries that most effectively splits them into subsets enriched in the desired annotation.
- The split criterion used by C4.5 is based on Shannon entropy.

How? SAAS – C4.5 algorithm

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Proteins	Attributes			Annotations
	Prosite PS00847	Pfam PF01493	Mammalia	FAD
Q9ZN27	-	yes	-	yes
Q9JI00	-	-	yes	-
Q9ZL14	yes	-	-	-
Q9NYQ3	-	-	yes	-
Q9UJM8	-	-	yes	-
Q9NYQ2	-	-	yes	-
Q43155	-	yes	-	yes
Q9T0P4	-	yes	-	yes
Q9WU19	-	-	yes	-



- At each node of the tree, C4.5 chooses one attribute of the entries that most effectively splits them into subsets enriched in the desired annotation.
- The split criterion used by C4.5 is based on Shannon entropy.

How? SAAS – validation

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

- 4-fold cross validation is performed on UniProtKB/Swiss-Prot – true and false positives are measured on this training/test set
- Performance on UniProtKB/TrEMBL is estimated using a standard Bernoulli-process based method, assuming a normal distribution:

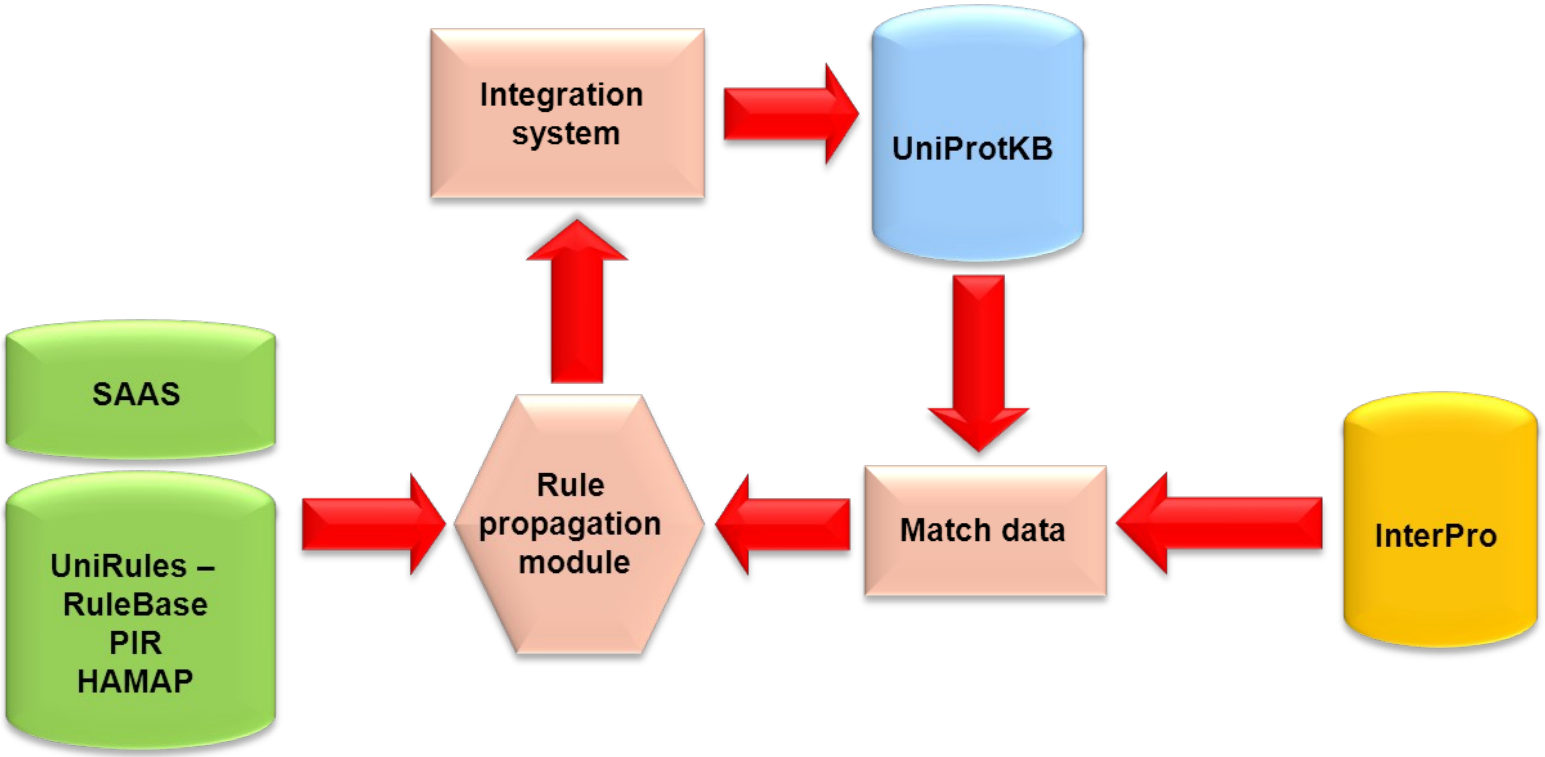
$$p = \frac{\left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right)}{(1 + z^2/N)}$$

f - observed success rate
N - number of predictions
p - estimated success rate
z - confidence limit

- Only SAAS rules with 99% or greater estimated confidence generate annotation. Rules are produced and validated each release.

How? Automatic annotation in UniProtKB

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

How - UniRules

- Rules of varying complexity: annotation varies from simple KW attribution to complete annotation as for UniProtKB/Swiss-Prot
- Curation:
  - From SAAS rules as input
  - From UniProtKB/Swiss-Prot annotation and InterPro match data, taxonomy information – continuously reported to curators
  - From literature based curation of characterized families - with the possibility to create new signatures for specific functional groups
- Validation: rules are continuously monitored – validation on UniProtKB/Swiss-Prot – 97% confidence

How? UniRule – rule complexity (HAMAP)

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Propagated annotation

Identifier, protein and gene names

Identifier IDI

Protein name RecName: Full=Isopentenyl-diphosphate Delta-isomerase;
Short=IPP isomerase;
EC=5.3.3.2;

Comments

FUNCTION: Catalyzes the 1,3-allylic rearrangement of the homoallylic substrate isopentenyl (IPP) to its highly electroph
allylic isomer, dimethylallyl diphosphate (DMAPP) (By similarity).
CATALYTIC ACTIVITY: Isopentenyl diphosphate = dimethylallyl diphosphate.

case <FTGroup:1>
COFACTOR: Binds 1 manganese ion per subunit (By similarity).
end case

case <Property:PHOTOSYN>
PATHWAY: Porphyrin biosynthesis; chlorophyll biosynthesis.
end case

Features

From: IDI_ECOLI (Q46822)

Key	From	To	Description	Condition	FTGroup
ACT_SITE	67	67	*By similarity*	C	
ACT_SITE	116	116	*By similarity*	E	
METAL	25	25	Manganese *(By similarity)*	H	1
METAL	32	32	Manganese *(By similarity)*	H	1
METAL	69	69	Manganese *(By similarity)*	H	1

How? UniRule – rule complexity (HAMAP)

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Propagated annotation

Identifier, protein and gene names

Identifier IDI

Protein name RecName: Full=Isopentenyl-diphosphate Delta-isomerase;
Short=IPP isomerase;
EC=5.3.3.2;

Comments

FUNCTION: Catalyzes the 1,3-allylic rearrangement of the homoallylic substrate isopentenyl (IPP) to its highly electrophi allylic isomer, dimethylallyl diphosphate (DMAPP) (By similarity).
CATALYTIC ACTIVITY: Isopentenyl diphosphate = dimethylallyl diphosphate.

case <FTGroup:1>
COFACTOR: Binds 1 manganese ion per subunit (By similarity).
end case

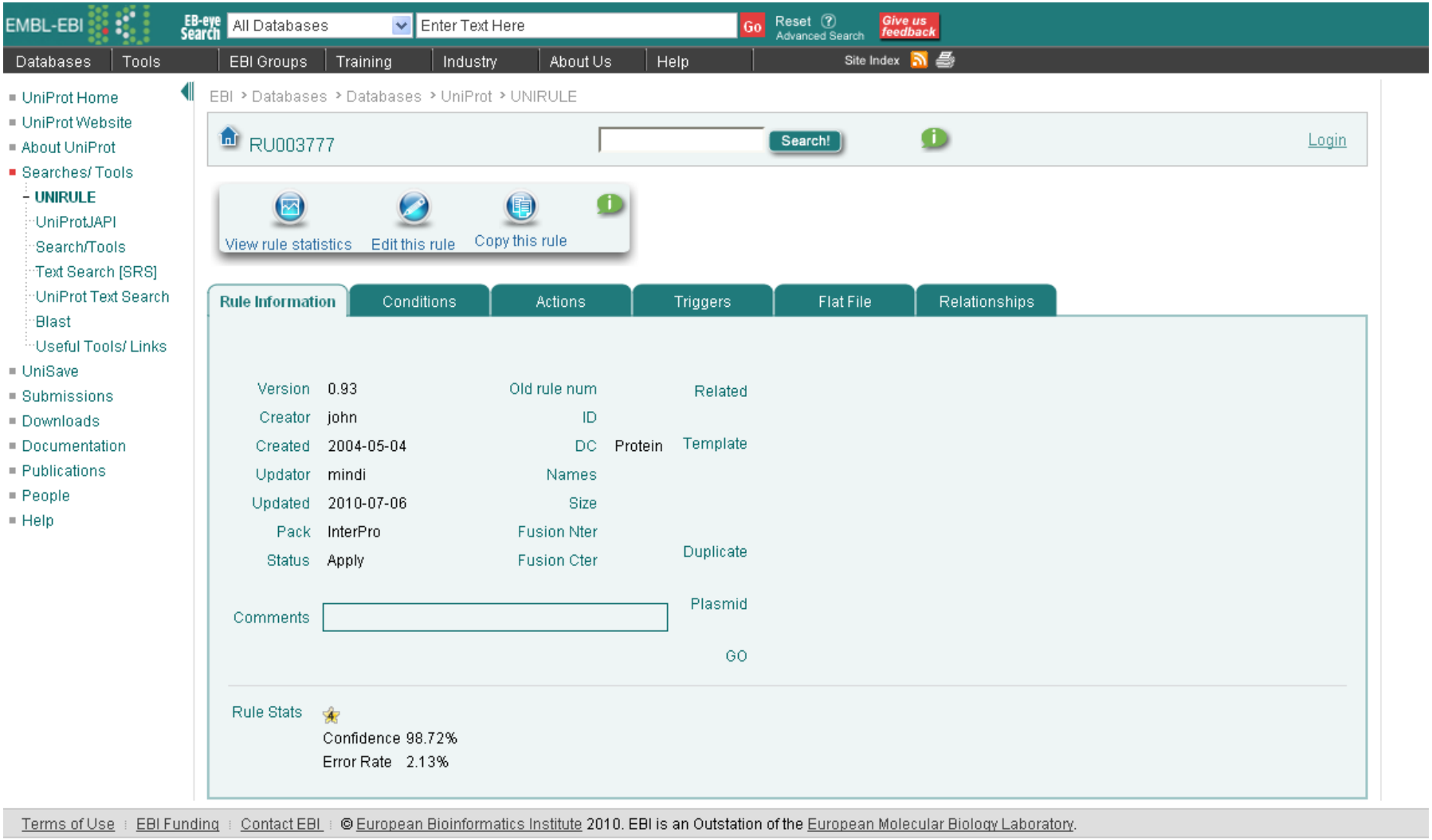
case <Property:PHOTOSYN>
PATHWAY: Porphyrin biosynthesis; chlorophyll biosynthesis.
end case

Features

From: IDI_ECOLI (Q46822)

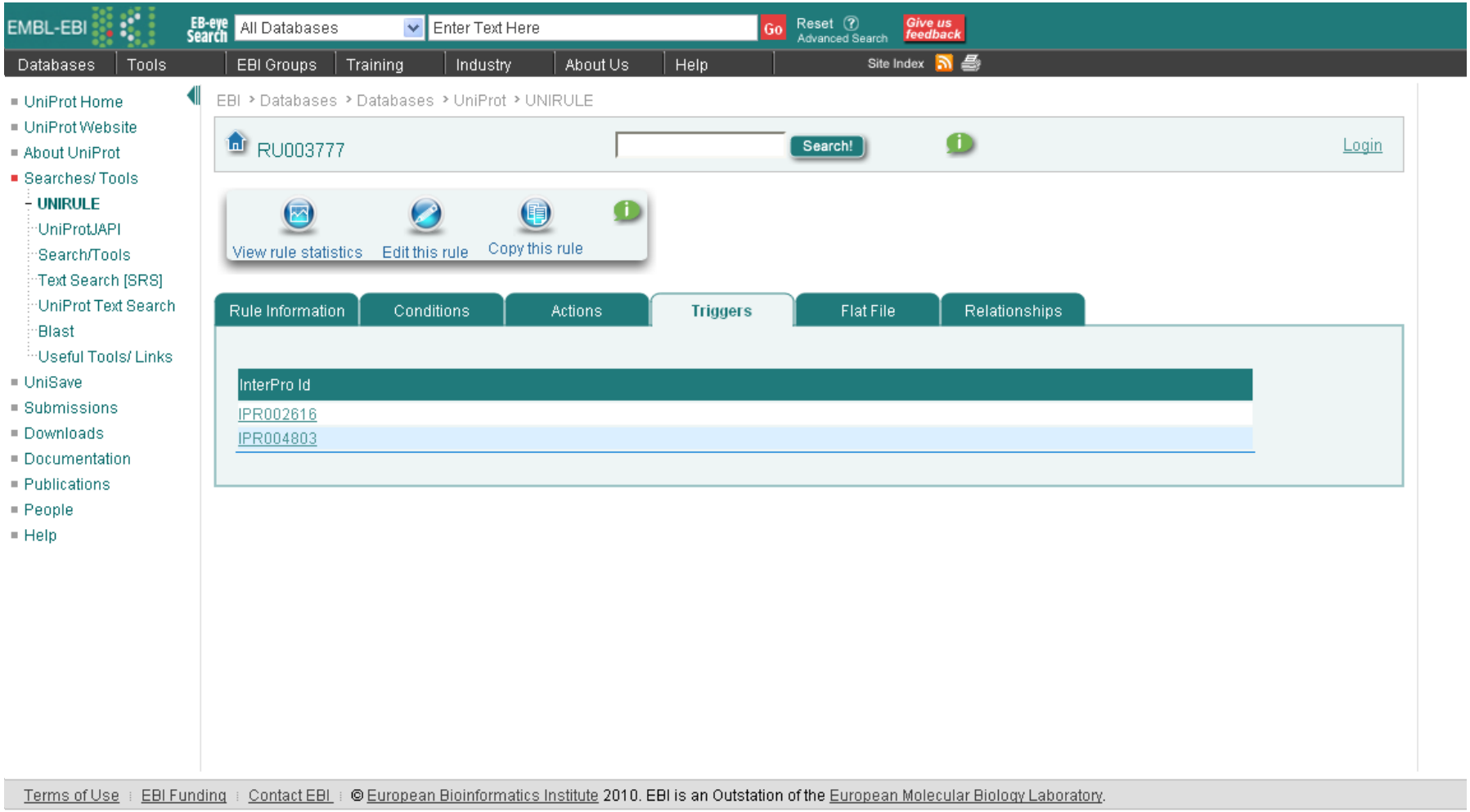
Key	From	To	Description	Condition	FTGroup
ACT_SITE	67	67	*By similarity*	C	
ACT_SITE	116	116	*By similarity*	E	
METAL	25	25	Manganese *(By similarity)*	H	1
METAL	32	32	Manganese *(By similarity)*	H	1
METAL	69	69	Manganese *(By similarity)*	H	1

How? UniRule – curator tool



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

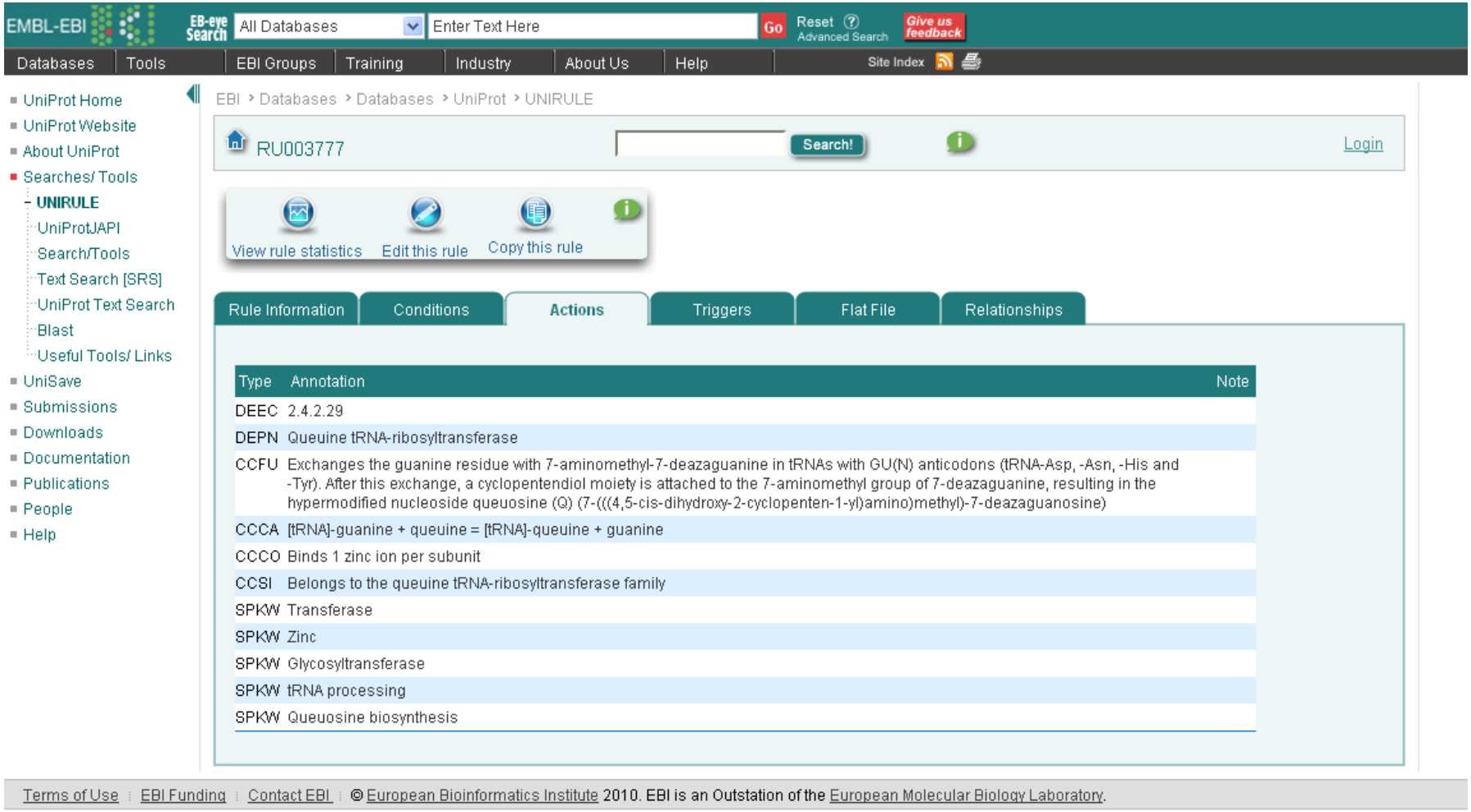
How? UniRule – curator tool



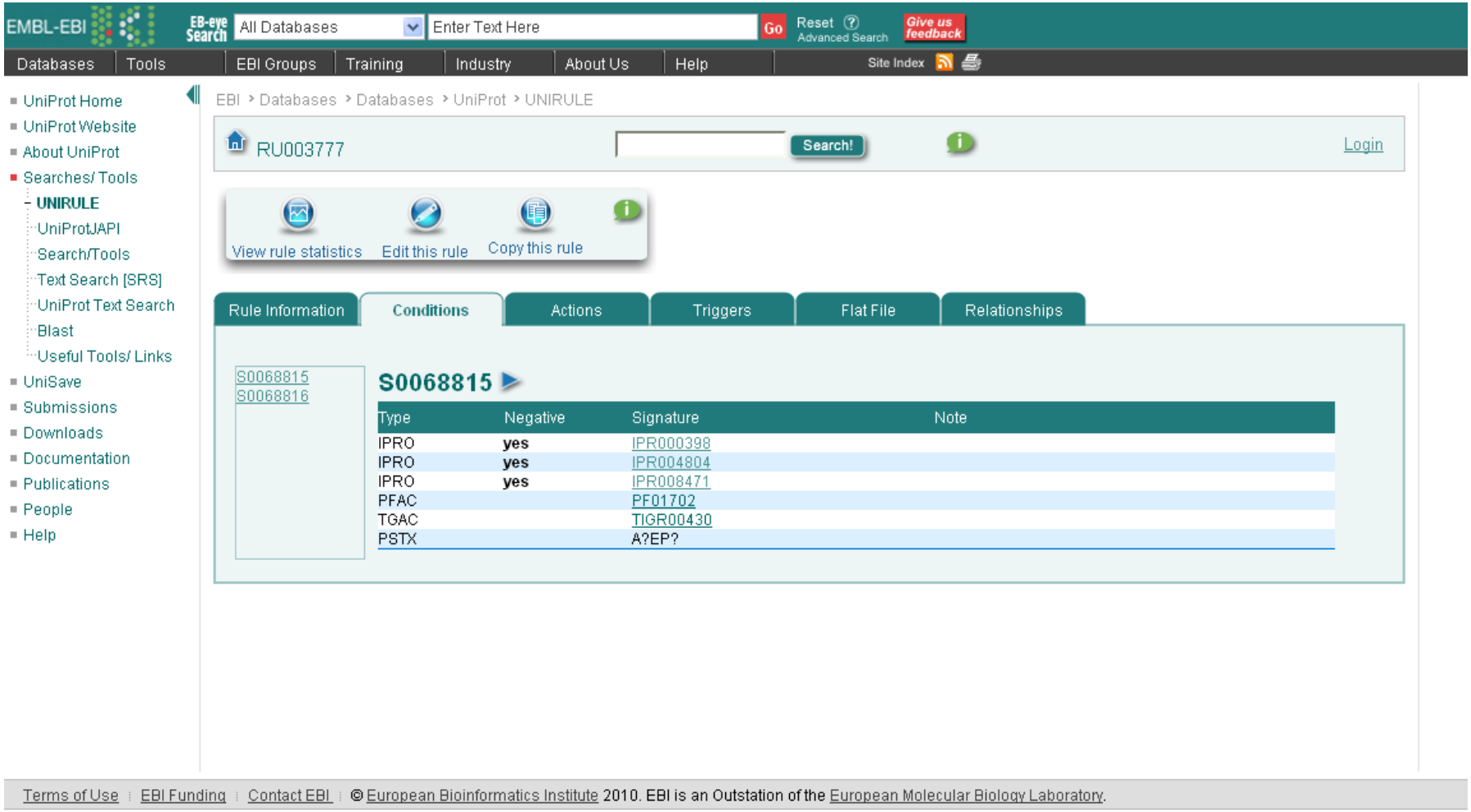
Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

How? UniRule – curator tool

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

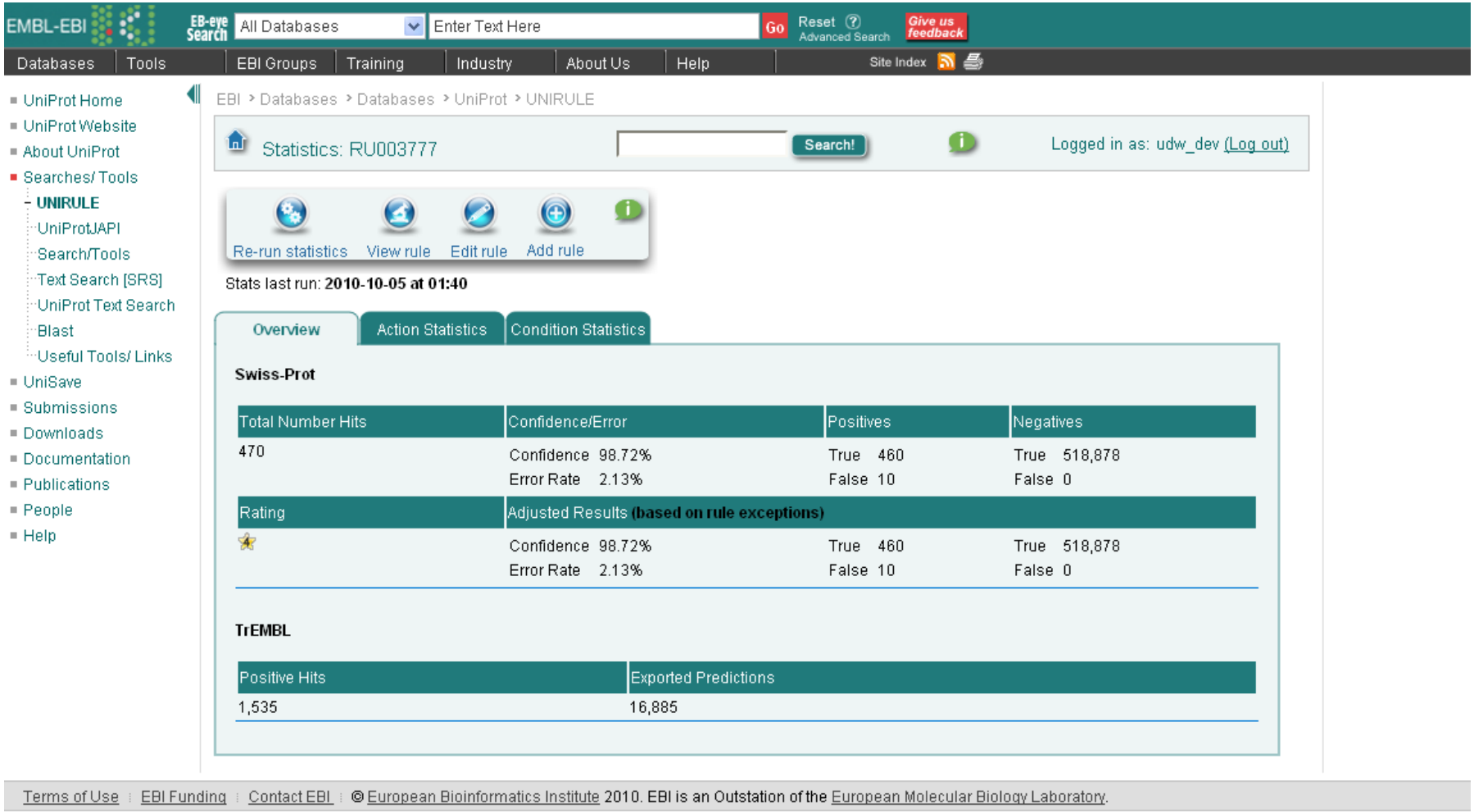


How? UniRule – curator tool



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

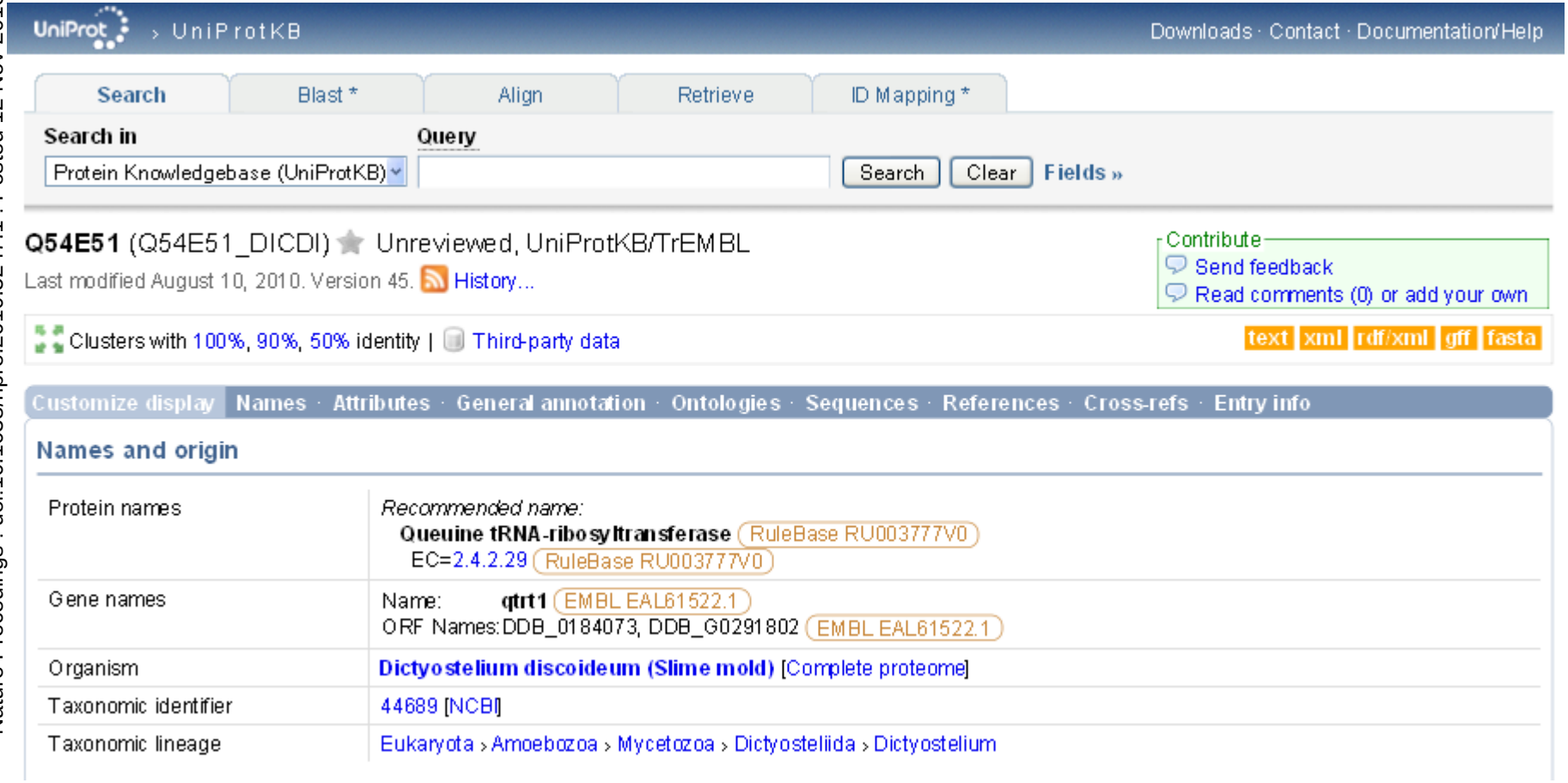
How? UniRule – curator tool



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

How? UniRule – output

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



How? UniRule – output

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

General annotation (Comments)

Function	Exchanges the guanine residue with 7-aminomethyl-7-deazaguanine in tRNAs with GUN 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 RU003777V0
Catalytic activity	[tRNA]-guanine + queuine = [tRNA]-queuine + guanine. RuleBase RU003777V0
Cofactor	Binds 1 zinc ion per subunit By similarity. RuleBase RU003777V0
Sequence similarities	Belongs to the queuine tRNA-ribosyltransferase family. RuleBase RU003777V0
Caution	The sequence shown here is derived from an EMBL/GenBank/DDBJ whole genome shotgun (WGS) entry which is preliminary data.

Ontologies

Keywords

Biological process	Queuosine biosynthesis RuleBase RU003777V0 tRNA processing RuleBase RU003777V0
Ligand	Zinc RuleBase RU003777V0
Molecular function	Glycosyltransferase RuleBase RU003777V0 Transferase
Technical term	Complete proteome

How? UniRule – output

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Unreviewed, UniProtKB/TrEMBL Q54E51 (Q54E51_DICDI)

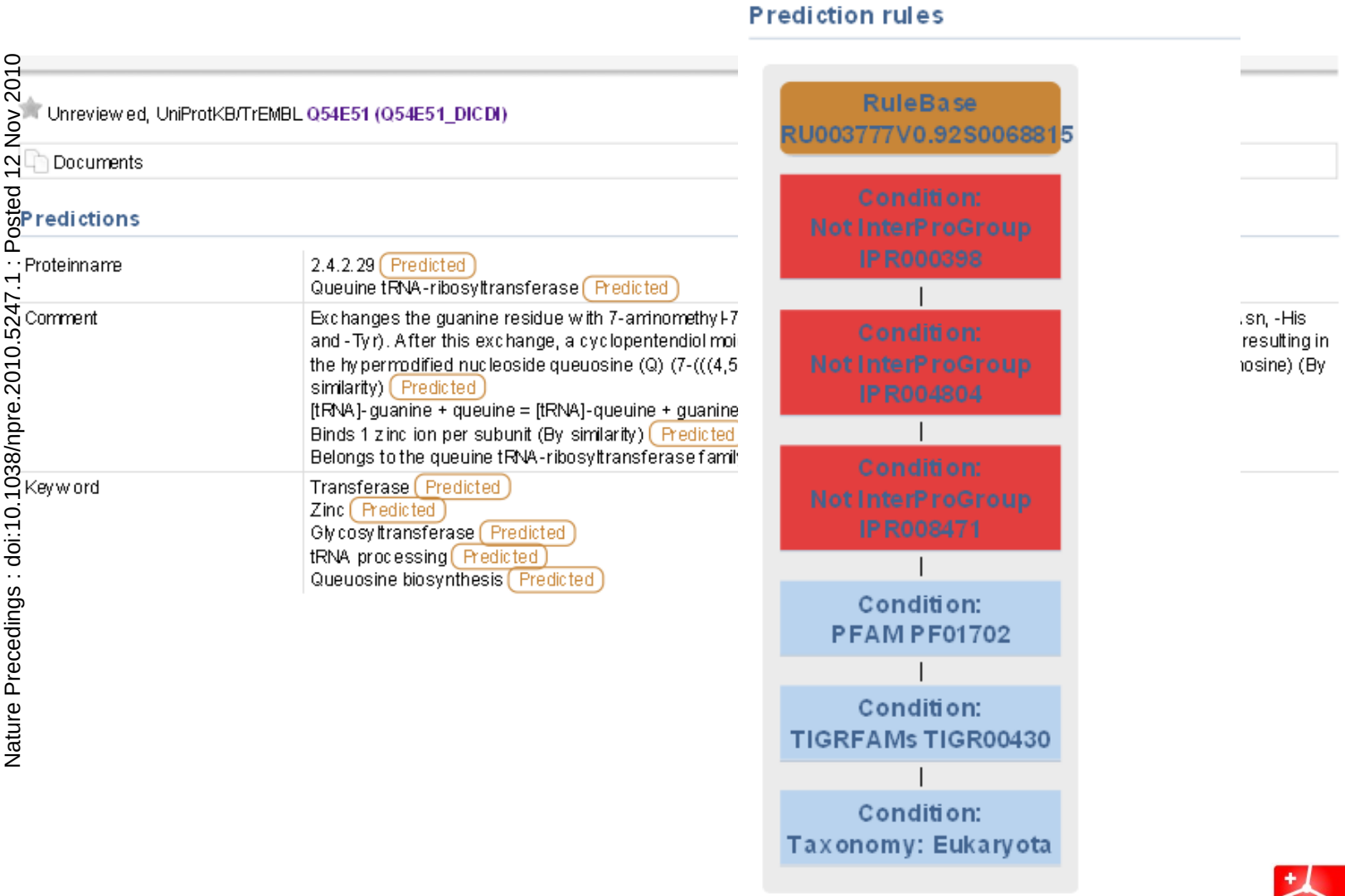
Documents

Predictions

Proteinname	2.4.2.29 Predicted Queuine tRNA-ribosyltransferase 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 Binds 1 zinc ion per subunit (By similarity) Predicted Belongs to the queuine tRNA-ribosyltransferase family Predicted
Keyword	Transferase Predicted Zinc Predicted Glycosyltransferase Predicted tRNA processing Predicted Queuosine biosynthesis Predicted

How? UniRule – output

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

How? Automatic annotation in UniProtKB/TrEMBL - Summary

- Manual annotation cannot keep pace with current or projected rates of growth of UniProtKB -> automatic annotation

- SAAS – automatically generated annotation rules for comments, KWs

- UniRule – manually curated annotation rules
  - annotation varies from simple KWs to full annotation
  - start point can be SAAS rules, InterPro reports, literature-based curation of protein families
  - possibility to create custom signatures -> InterPro

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

- The source of each annotation is indicated - users can access rule logic

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Talk outline

- Introduction
- What?

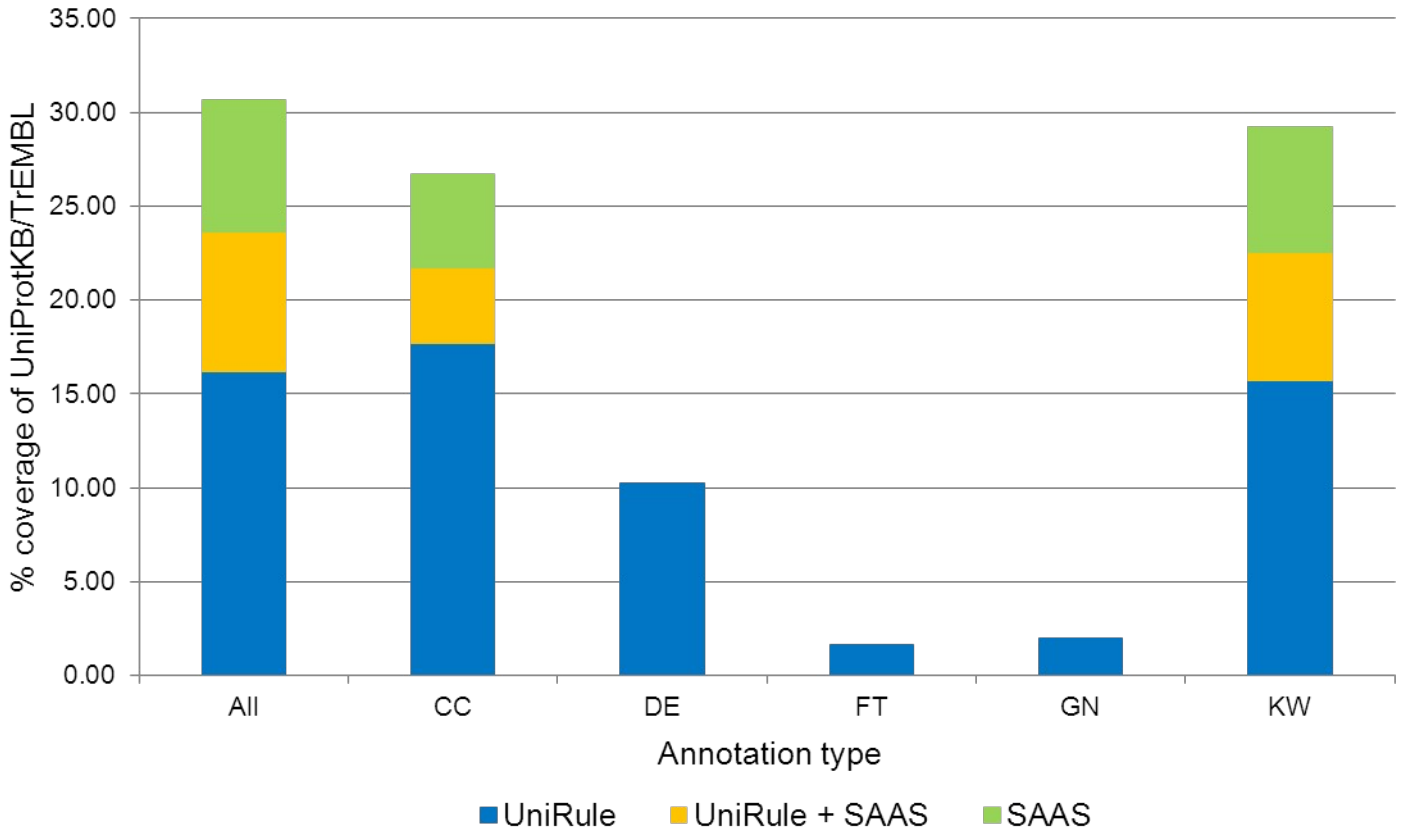
  What we annotate in UniProtKB, and what we propagate
- Why?

  Need for automatic annotation in UniProtKB
- How?

  Automatic annotation systems for UniProtKB
- **Current status and future plans**

Current status - coverage of UniProtKB/TrEMBL

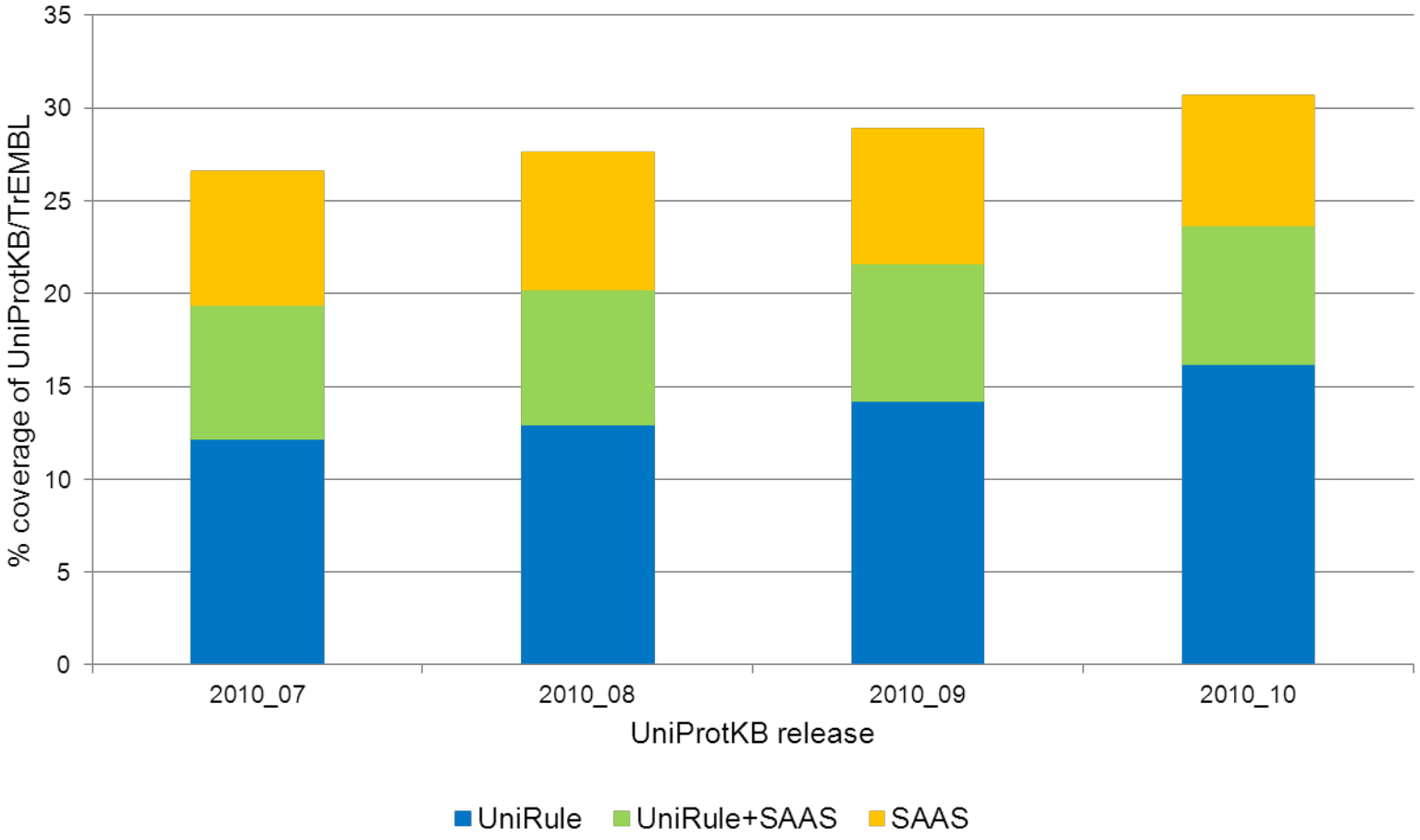
Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



UniProt release 2010_10 included annotations from 7767 SAAS rules, 1814 UniRules

Current status – coverage of UniProtKB/TrEMBL

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010



Future plans

- Complete integration of all existing rules into one unified system
- Improve presentation and access – including prediction server similar to that currently provided for HAMAP rules
- Collaborative exchange of annotations with other groups – to improve coverage and quality

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Acknowledgements

- **Curators and developers at EBI:**
  Ricardo Antunes, Mark Bingley, Alan da Silva, Kati Laiho, Sam Patient, Klemens Pichler, **Diego Poggioli**, Mindi Sehra, Tony Wardell, Maria Jesus Martin, **Claire O'Donovan**

- **Curators and developers at SIB:**
  Andrea Auchincloss, Delphine Baratin, Edouard de Castro, Elisabeth Coudert,

  Thomas Kappler , Guillaiume Keller, Corinne Lachaize, Tania Lima, Karine Michoud, Isabelle Phan, Nicole Redaschi, Catherine Rivoire

- **Curators and developers at PIR:**
  Chuming Chen, Hongzhan Huang, Peter McGarvey, Natalia Petrova, Cecilia Arighi, Winona C. Barker, Raja Mazumder, Darren Natale, C.R. Vinayaka, Lai-Su L. Yeh, Qinghua Wang, Uzoamaka Ugochukwu

- **The rest of the UniProt production team**

- **Prosite team:**
  Nicolas Hulo, Christian Sigrist, Lorenzo Cerutti, Beatrice Cuche

- **InterPro team**

- **National Institutes of Health (NIH), Swiss Federal Government through the Federal Office of Education and Science; European Commission.**

Nature Precedings : doi:10.1038/npre.2010.5247.1 : Posted 12 Nov 2010

Thanks for your attention