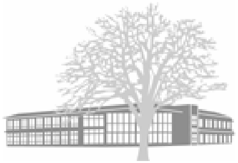
EMBL – EBI
European Bioinformatics Institute

UniProt -
The Universal Protein Resource

Claire O'Donovan



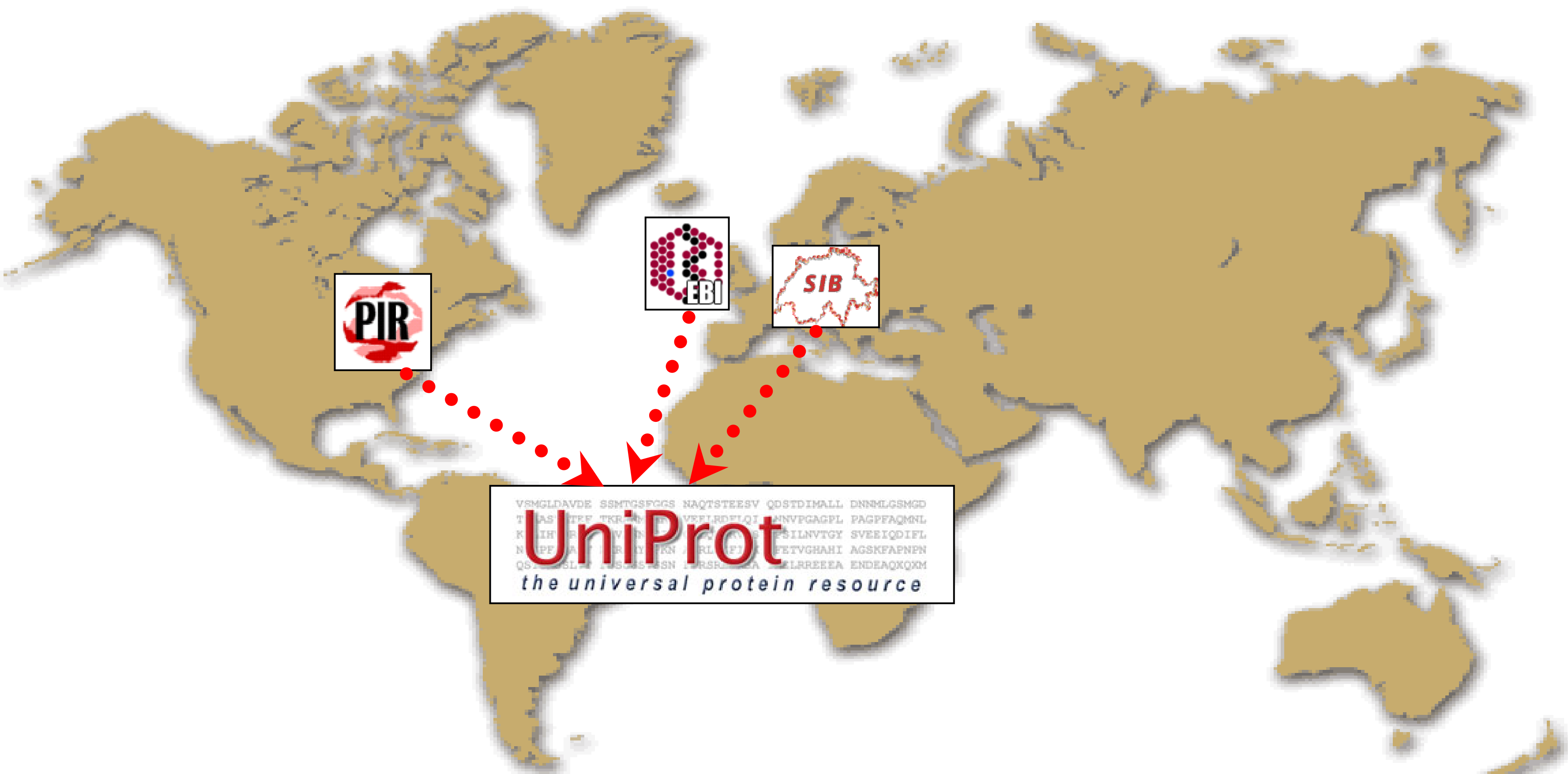
Pre-UniProt

- Swiss-Prot: created in July 1986; since 1987, a collaboration of the SIB and the EMBL/EBI;
- TrEMBL: created at the EBI in November 1996 as a computer-annotated protein sequence database supplementing Swiss-Prot. It was introduced to deal with the increased data flow from genome projects.

The UniProt timeline

- Awarded to EBI, SIB, and PIR by NIH
- Run time 9/02-8/05
- ~16 million USD intended to replace Swiss-Prot license fees and previous PIR funding

UniProt Consortium



UniProt Consortium activities

UniParc
The UniProt Archive

UniRef
Non-redundant Reference
100% > 90% > 50%

UniProt
the universal protein knowledgebase

The three-layered approach

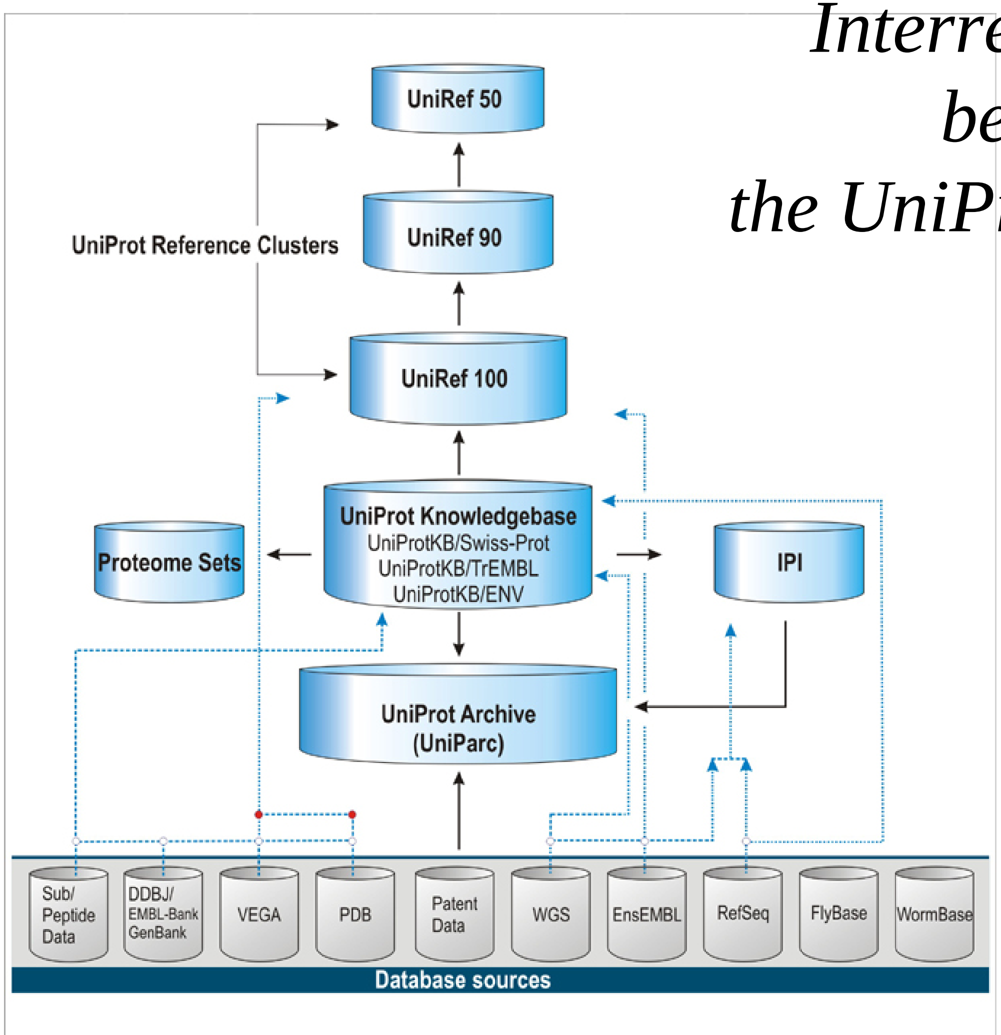
- The UniProt Archive (UniParc)
  - ✓ UniProtKB + all other protein sequences publicly available
  - ✓ Completeness
- The UniProt Reference Clusters (UniRef)
  - ✓ Non-redundant views of UniProtKB + selected UniParc sets
  - ✓ Speed
- The UniProt Knowledgebase (UniProtKB)
  - ✓ Central database of annotated protein sequences and functional information
  - ✓ UniProtKB/Swiss-Prot + UniProtKB/TrEMBL

EMBL – EBI
European Bioinformatics Institute

UniProt Universal Protein Resource

The three layer approach

Interrelationship between the UniProt Databases



UniProt Archive

- UniParc is a non-redundant archive of protein sequences from the public databases
- It contains only protein sequences (no annotations)
- It provides cross-references to the source databases

UniProt Archive: Principles

- UniParc is non-redundant
  - Each unique protein sequence is stored only once and is assigned a unique stable UniParc identifier (e.g UPI0000000356)
- UniParc provides cross-references to the original source: active or retired
- UniParc provides sequence versions.

Your Query Result Sets (Page - 1)[Data Set Manager]

(cro n (cro n..	crossref.pdb ..	(cro n (cro n..	crossref.hinv..	(cro n (cro n..	>>>
8 entries	18412 entries	43 entries	35908 entries	121 entries	

Entry UPI0000000356 New Query | Download Protein | Bookmark Protein (Ctrl+D)

UPI0000000356 | UPI000000060F | UPI0000000C37 | UPI0000000DEF | UPI000004EF93 | UPI000000D91A | UPI000003060C

Viewers: XML | ExPASy | SRS | PIR

UPI		UPI0000000356
Sequence		MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNYAVTVMIGGEPYTLGLFDTAG QEDYDRLRPLSYPQTDVFLVCFSVVSPSSFENVKEKWVPEITHHCPKTPFLLVGTQIDLR DDPSTIEKLAKNKQKPITPETAEKLARDLKAVKYVECSALTQKGLKNVFDEAILAALEPP EPKKSRRCVLL
	Length	191
	CRC64	51A437E22A4D8FFF

References	DataBase	Accession	Version	Active	Created	Last Update	Deleted
	EMBL	AAA37410.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	AAA52592.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	AAC00028.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	AAH02711.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	AAH03682.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	AAH18266.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	AAH60535.1	1	Y	30-OCT-2003	10-JUN-2004	-
	EMBL	AAM21110.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	BAB22563.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	BAC35825.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	CAA90215.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	CAB52602.1	1	N	12-MAR-2003	-	12-MAR-2003
	EMBL	CAB52602.1	1	Y	19-JUN-2003	10-JUN-2004	-
	EMBL	CAB57326.1	1	Y	12-MAR-2003	10-JUN-2004	-
	EMBL	CAE93985.1	1	Y	04-JAN-2004	10-JUN-2004	-
	EPO	AX305419.1	1	Y	26-MAR-2003	11-JUN-2003	-
	Ensembl_HUMAN	ENSP00000251252	2	N	04-JUL-2003	-	04-JUL-2003
	Ensembl_HUMAN	ENSP00000314435	1	N	01-APR-2003	-	03-JUN-2003
	Ensembl_HUMAN	ENSP00000314458	1	N	01-APR-2003	-	03-JUN-2003
	Ensembl_HUMAN	ENSP00000337669	1	Y	13-FEB-2004	07-JUN-2004	-
	Ensembl_MOUSE	ENSMUSP00000030417	1	N	04-MAR-2003	-	08-NOV-2003
	Ensembl_MOUSE	ENSMUSP00000054634	1	Y	09-MAY-2003	07-JUN-2004	-
	Ensembl_RAT	ENSRNOP00000030928	1	Y	13-FEB-2004	07-JUN-2004	-
	H_INV	HIT000031119.1	1	Y	13-MAY-2004	08-JUN-2004	-
	H_INV	HIT000031693.1	1	Y	13-MAY-2004	08-JUN-2004	-
	H_INV	HIT000038320.1	1	N	13-MAY-2004	-	28-MAY-2004
	H_INV	HIT000038320.2	1	Y	08-JUN-2004	08-JUN-2004	-

UniProt Reference Clusters Principles

- It provides non-redundant reference data collections
- It allows faster and more informative sequence similarity searches
- It includes the UniProtKB and some data from UniParc
- It merges across different species

UniProt Reference Clusters Principles

- UniRef100
  - It merges identical sequences and subfragments
- UniRef90
  - Size reduction of 40%
- UniRef50
  - Size reduction of 65%

UniProt

the universal protein knowledgebase

UniProtKB/Swiss-Prot

- Non-redundant
- High level of integration
- High level of manual curation
- Contains 241,242 entries

UniProtKB/TrEMBL

- Translations of CDS in EMBL/GenBank/DDBJ
- Automatic annotation
- Contains 3,313,265 entries

UniProtKB/TrEMBL

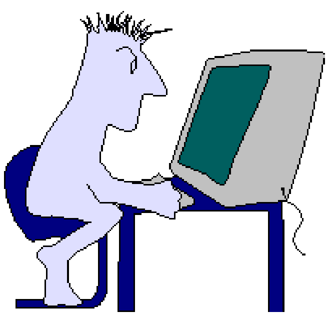
- Automatically generated in a biweekly cycle from the data present in EMBL/GenBank/DDBJ and some other sources such as TAIR/SGD
- Exclusions: pseudogenes, synthetic, immunoglobulins, patents, small sequences <8
- /product, /gene, /locus_tag
- RefSeq and Ensembl

UniProtKB/TrEMBL

- Proteome annotation
- Cross-references to other databases
- Addition of relevant publications (eg PDB)
- Redundancy
- Automatic annotation
- Future plans for manual annotation eg human proteome project

Literature

Other databases



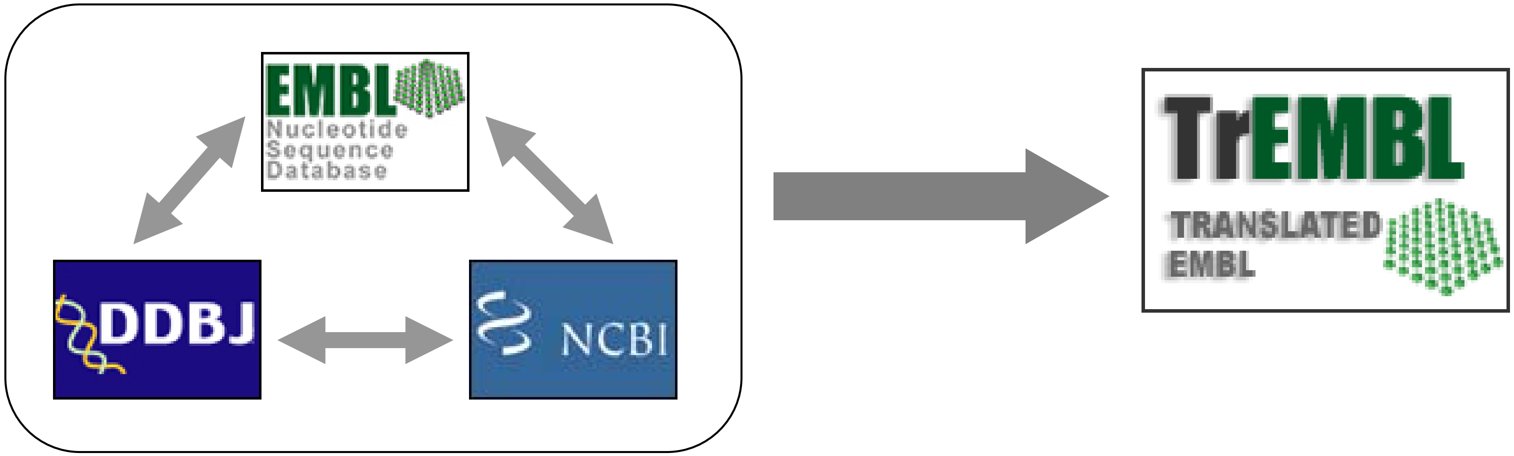
Analysis tools

External expertise

EMBL – EBI
European Bioinformatics Institute

UniProt
Universal Protein Resource

Capturing the correct sequence



- Archive collections
- Each sequence report stored in its own entry

- Merging at 100% identity
- Still some redundancy

Sequence similarity searches

- Identify potential merge candidates
- Identify similar already curated entries

Sequence comparison

- Sequence alignments
- Identification of sequence differences
- Helps in identifying underlying causes

Causes of sequence differences

- Polymorphisms, disease variants
- Splice variants
- Sequencing errors
- Incorrect predictions

Literature curation

- 1741 different journals cited in Swiss-Prot
- Total of 383,401 references
- Average of 2 references per entry

Comments

FUNCTION	Converts the abundant, but inactive, zymogen plasminogen to plasmin by hydrolyzing a single Arg-Val bond in plasminogen. By controlling plasmin-mediated proteolysis, it plays an important role in tissue remodeling and degradation, in cell migration and many other physiopathological events.
CATALYTIC ACTIVITY	Specific cleavage of Arg-|-Val bond in plasminogen to form plasmin.
SUBUNIT	Heterodimer of chain A and chain B held by a disulfide bond. Binds to fibrin with high affinity. This interaction leads to an increase in the catalytic efficiency of the enzyme between 100-and 1000-fold, due to an increase in affinity for plasminogen. Similarly, binding to heparin increases the activation of plasminogen. Binding to laminin and fibronectin has also been demonstrated. Binds to mannose receptor and the low-density lipoprotein receptor-related protein (LRP1). These proteins are involved in TPA clearance. Also binds to annexin II and to cytokeratin 8. Yet unidentified interactions on endothelial cells and vascular smooth muscle cells (VSMC) lead to a 100-fold stimulation of plasminogen activation. In addition, binding to VSMC reduces TPA inhibition by PAI-1 by 30-fold. Binds LRP1B; binding is followed by internalization and degradation.
SUBCELLULAR LOCATION	Secreted; extracellular.
ALTERNATIVE PRODUCTS	Alternative splicing;2 named isoforms [Display all isoform sequences in Fasta format] Name: 1 Synonyms: Long IsoformId: P00750-1 Sequence: This is the isoform sequence displayed in this entry. Name: 2 Synonyms: Short IsoformId: P00750-2 Sequence: VSP_005411, VSP_005412 Note: May be produced at very low levels due to a premature stop codon in the mRNA, leading to nonsense-mediated mRNA decay
TISSUE SPECIFICITY	Synthesized in numerous tissues (including tumors) and secreted into most extracellular body fluids, such as plasma, uterine fluid, saliva, gingival crevicular fluid, tears, seminal fluid, milk.
DOMAIN	Both FN1 and one of the kringle domains are required for binding to fibrin.
DOMAIN	Both FN1 and EGF-like domains are important for binding to LRP1.
PTM	N-glycosylation of Asn-152; the bound oligomannosidic glycan is involved in the interaction with the mannose receptor.
PTM	Characterization of O-linked glycan was studied in Bowes melanoma cell line.
DISEASE	Increased activity of TPA causes hyperfibrinolysis, with excessive bleeding as a consequence.
DISEASE	Defective release of TPA causes hypofibrinolysis, leading to thrombosis or embolism.
PHARMACEUTICAL	Available under the names Activase (Genentech) and Retavase (Centocor and Roche) [Retavase is a fragment of TPA that contains kringle 2 and the protease domain; it was also known as BM 06.022]. Used in Acute Myocardial Infarction (AMI), in Acute Ischemic Stroke (AIS) and Pulmonary Embolism (PE) to initiates fibrinolysis.
SIMILARITY	Contains 1 EGF-like domain.
SIMILARITY	Contains 1 fibronectin type-I domain.
SIMILARITY	Contains 2 kringle domains.
SIMILARITY	Contains 1 peptidase S1 domain.
SIMILARITY	Belongs to the peptidase S1 family.

Sequence analysis

- Range of sequence analysis tools used to predict important sequence features
- Use of most appropriate programs
- Development of new predictive methods

Evidence attribution

- System which allows linking of all information in an entry to its original source.
- Allows users:
  - to trace origin of all data
  - to differentiate easily between literature-derived and computational data
  - to assess data reliability

UniProtKB curation group

14 curators

24 curators

2 curators

EBI curation projects

- Submissions
- Journal scanning
- Species-specific curation
  - human, mouse, rat, C.elegans, Drosophila, Xenopus, zebrafish, S.cerevisiae, S.pombe
- Protein family curation
  - kinases, keratins
- UniProtKB-MSD collaboration
- PTM standardisation

Some future curation plans

- Improvements to SPIN
- Extension of evidence attribution system to Swiss-Prot
- New annotation projects
- Community participation
- Further database collaborations

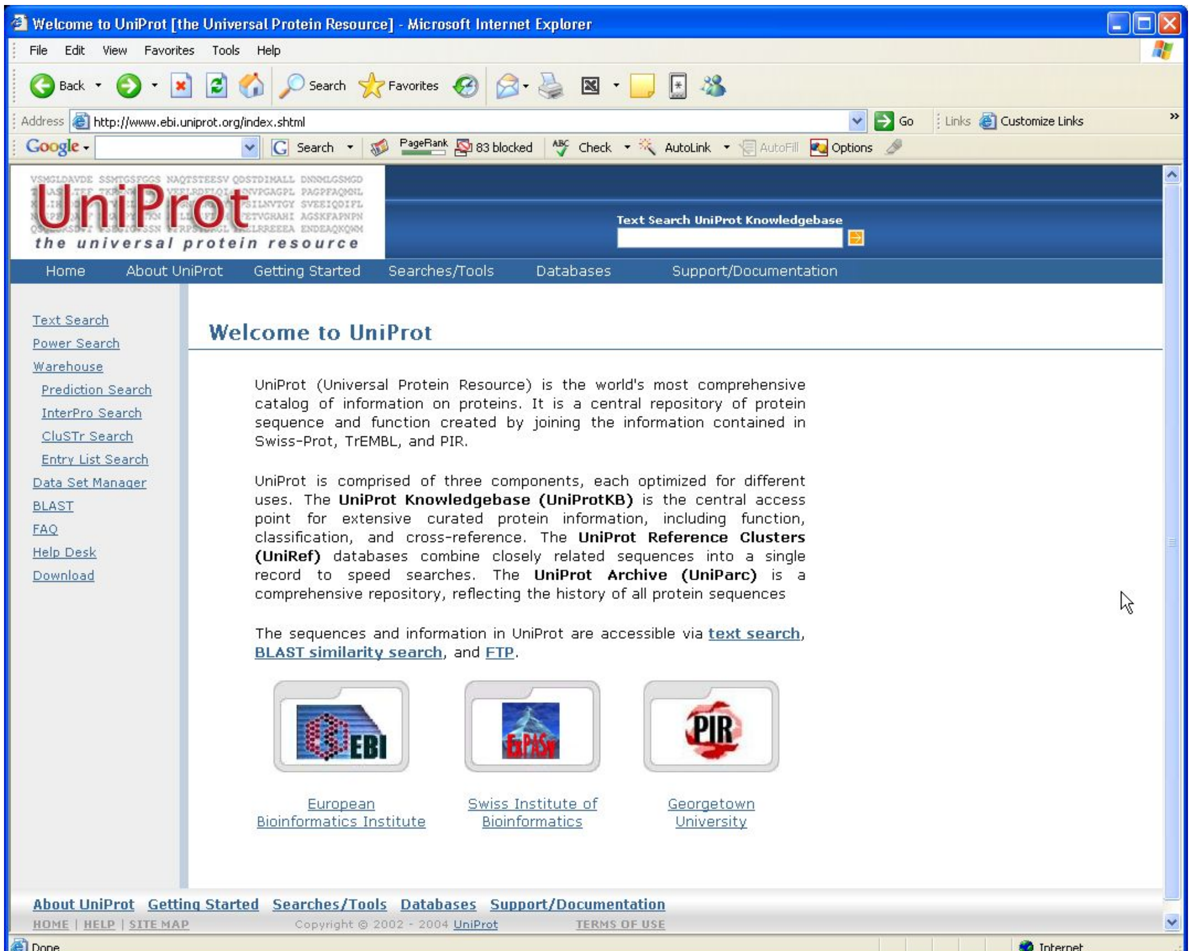
UniProt distribution

- Biweekly distribution
- Website access www.uniprot.org
- FTP access
- DVD of UniProtKB (datalib@ebi.ac.uk)

EMBL – EBI

European Bioinformatics Institute

UniProt Web



The new UniProt grant timeline

- Second Grant awarded to EBI, SIB, and PIR by NIH
- Run time 9/06-8/09

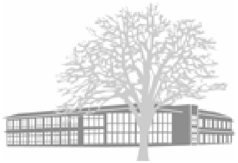
Acknowledgements (1)

Production:
Daniel Barrell
Renato Golin
Alexander Fedetov
Maria Jesus Martin
Patricia Monteiro
Claire O’Donovan
Mark Rijnbeek

UniParc/UniSave:
Quan Lin
Andrey Sitnov
Rasko Leinonen

Proteomes:
Alan Horne
Paul Kersey

AutomaticAnnotation
/Kraken/Website/XML:
Michael Kleen
Ernst Kretschmann
John O’Rourke
Sam Patient
Emilio Salazar
Natalyia Skylar
Dani Wieser



Acknowledgements (2)

- EBI curators:
  - Michele Magrane (Annotation coordinator / Mouse)
  - Yasmin Alam (Keratins)
  - Paul Browne (Journal scan)
  - Wei Mun Chan (Human)
  - Ruth Eberhardt (Submissions)
  - Rebecca Foulger (Xenopus)
  - Gill Fraser (Zebrafish)
  - Gabriella Frigerio (Rat)
  - John Garavelli (PTMs)
  - Jules Jacobsen (Structural data)
  - Kati Laiho (Fungi)
  - Claire O’Donovan (Quality control, data integration)
  - Sandra Orchard (Kinases)
  - Eleanor Whitfield (C.elegans, Drosophila)
- SIB Group
- PIR Group
- Rolf Apweiler