Protein Databases

Outline

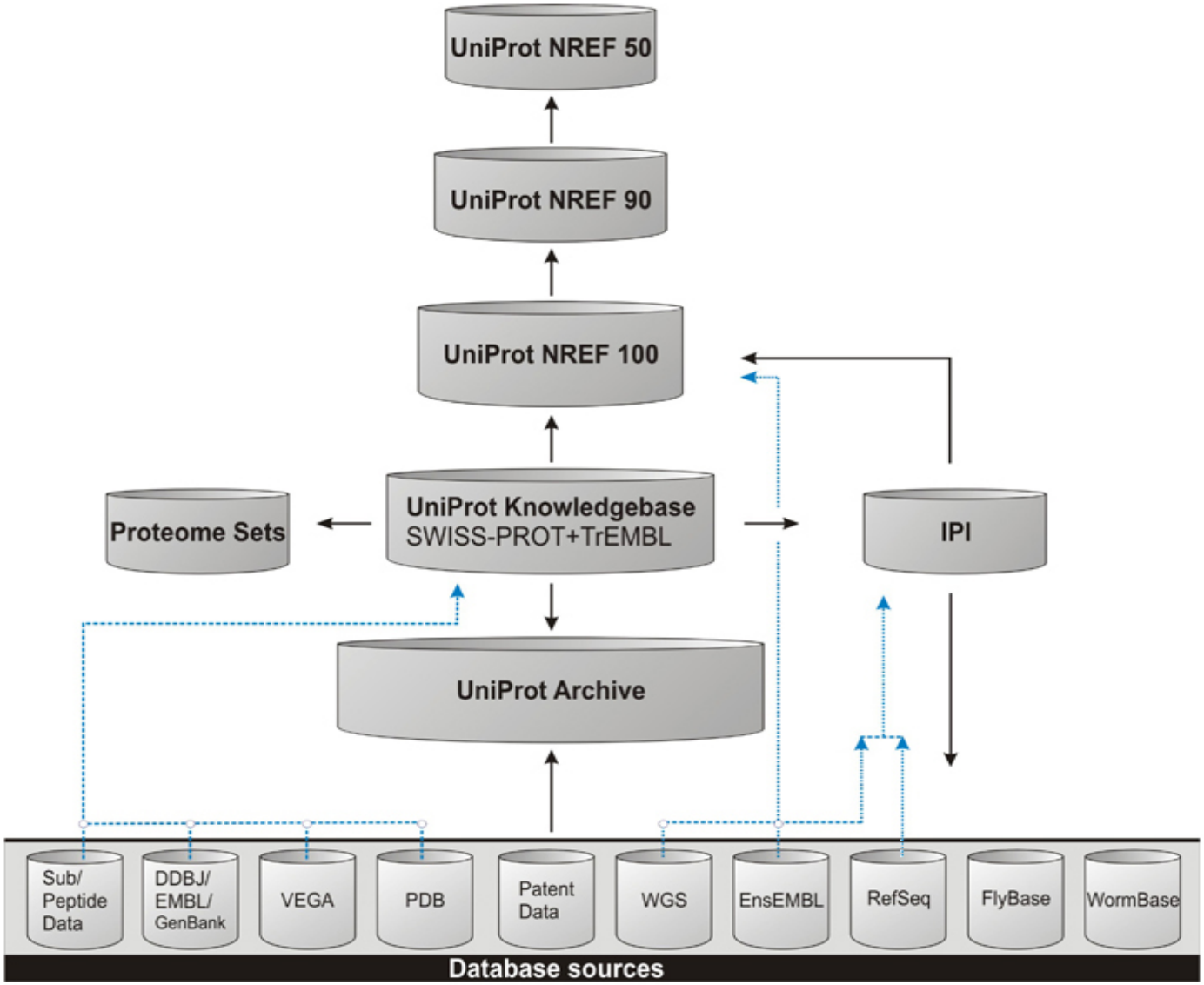
- UniProt
- Expectations and problems in using Proteomics data

UniProt

- Collaboration between EBI, SIB and PIR
- Funded mainly by NIH
- Based on the original work on PIR, Swiss-Prot and TrEMBL

UniProt Goals

- High level of annotation
- Minimal redundancy
- High level of integration with other databases
- Complete and up-to-date


UniProt NREF 50
UniProt NREF 90
UniProt NREF 100
Proteome Sets
UniProt Knowledgebase
SWISS-PROT+TrEMBL
IPI
UniProt Archive
Sub/ Peptide Data
DDBJ/ EMBL/ GenBank
VEGA
PDB
Patent Data
WGS
EnsEMBL
RefSeq
FlyBase
WormBase
Database sources

UniProt Non-Redundancy Concepts

- UniProt Archive (UniParc):
  - All sequences that are 100% identical over their entire length are merged into a single entry, regardless of species. UniParc represents each protein sequence once and only once, assigning it a unique Identifier. UniParc cross-references the accession numbers of the source databases.
- UniProt Knowledgebase:
  - Aims to describe in a single record all protein products derived from a certain gene (or genes if the translation from different genes in a genome leads to indistinguishable proteins) from a certain species.
    - Proteome sets and IPI
- UniProt Nref (UniRef):
  - Merges sequences automatically across different

UniProt Concepts of Complete and Up-to-date

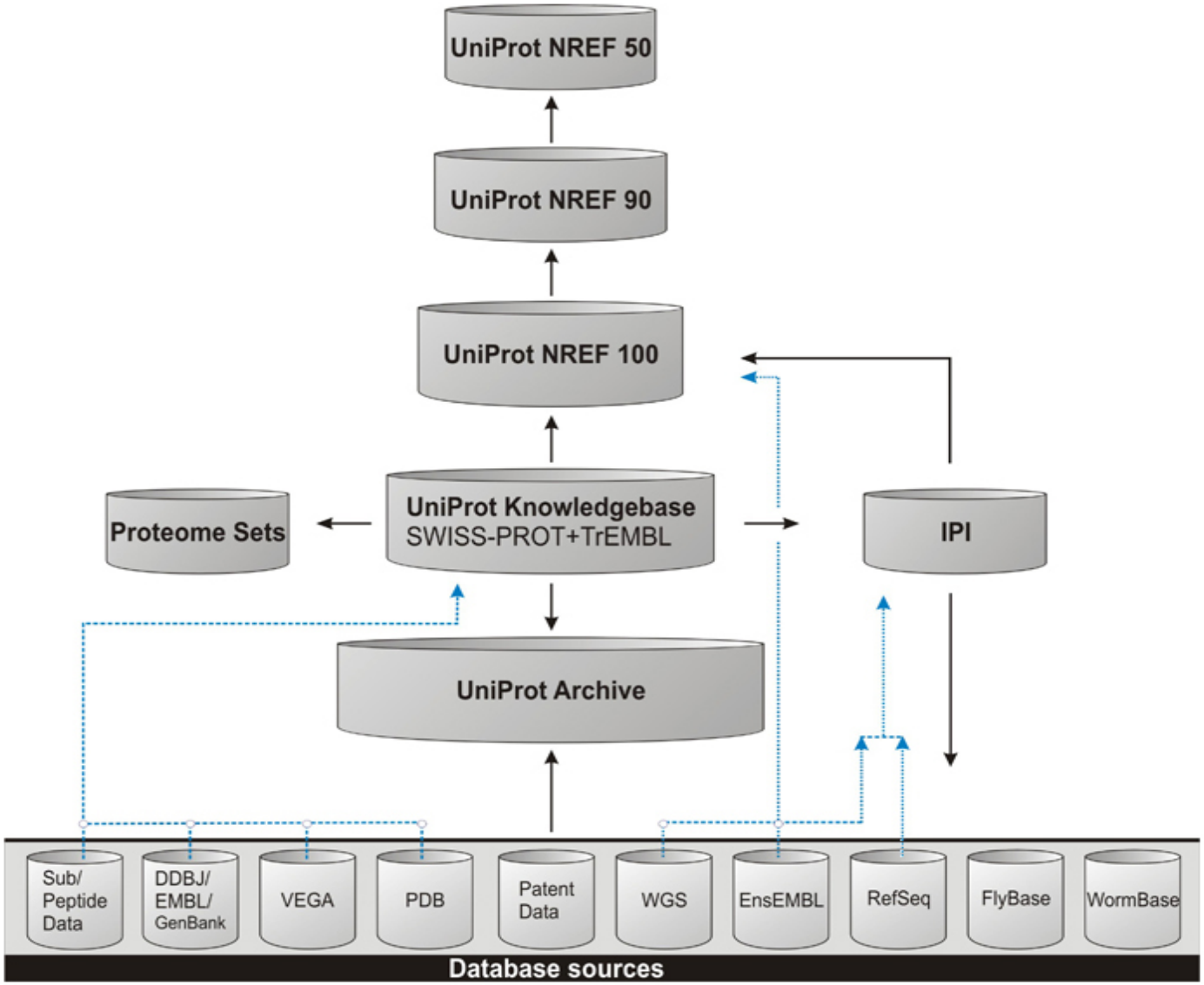
- UniProt Archive (UniParc):
  - All publically available protein sequences, updated every 2 weeks (12/04, Rel 3.4: 4,775,042 entries)
- UniProt Knowledgebase:
  - All suitable stable protein sequences, updated every 2 weeks (12/04, Rel 3.4: 1,707,421 entries)
- UniProt Nref (UniRef):
  - All protein sequences in the Knowledgebase and in UniParc useful for sequence similarity searches, updated every 2 weeks (12/04, Rel 3.4: 2,774,459 UniRef100, 1,762,793 UniRef90, 837,961430 UniRef50)

UniProt Concepts of Integration with other Databases

- UniProt Archive (UniParc):
  - Linked back to source records
- UniProt Knowledgebase:
  - Linked to >60 other databases
- UniProt Nref (UniRef):
  - UniRef clusters link back to Knowledgebase and UniParc records in the cluster

UniProt Annotation Concepts

- UniProt Archive (UniParc):
  - No annotation
- UniProt Knowledgebase:
  - Annotated
- UniProt Nref (UniRef):
  - No annotation, just description line of Knowledgebase or UniParc master entry in the cluster for use in FASTA files


UniProt NREF 50
UniProt NREF 90
UniProt NREF 100
Proteome Sets
UniProt Knowledgebase
SWISS-PROT+TrEMBL
IPI
UniProt Archive
Sub/ Peptide Data
DDBJ/ EMBL/ GenBank
VEGA
PDB
Patent Data
WGS
EnsEMBL
RefSeq
FlyBase
WormBase
Database sources

UniParc 3.4. December 2004

- 4,775,042 unique sequences from 11,095,078 source records (incl. 1,992,408 dead source records)
- Source databases are DDBJ/EMBL/GenBank, UniProt/Swiss-Prot, UniProt/TrEMBL, PIR-PSD, Ensembl, International Protein Index (IPI), PDB, RefSeq, FlyBase, WormBase, H-Inv, TROME, European Patent Office, United States Patent and Trademark Office and Japan Patent Office

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)

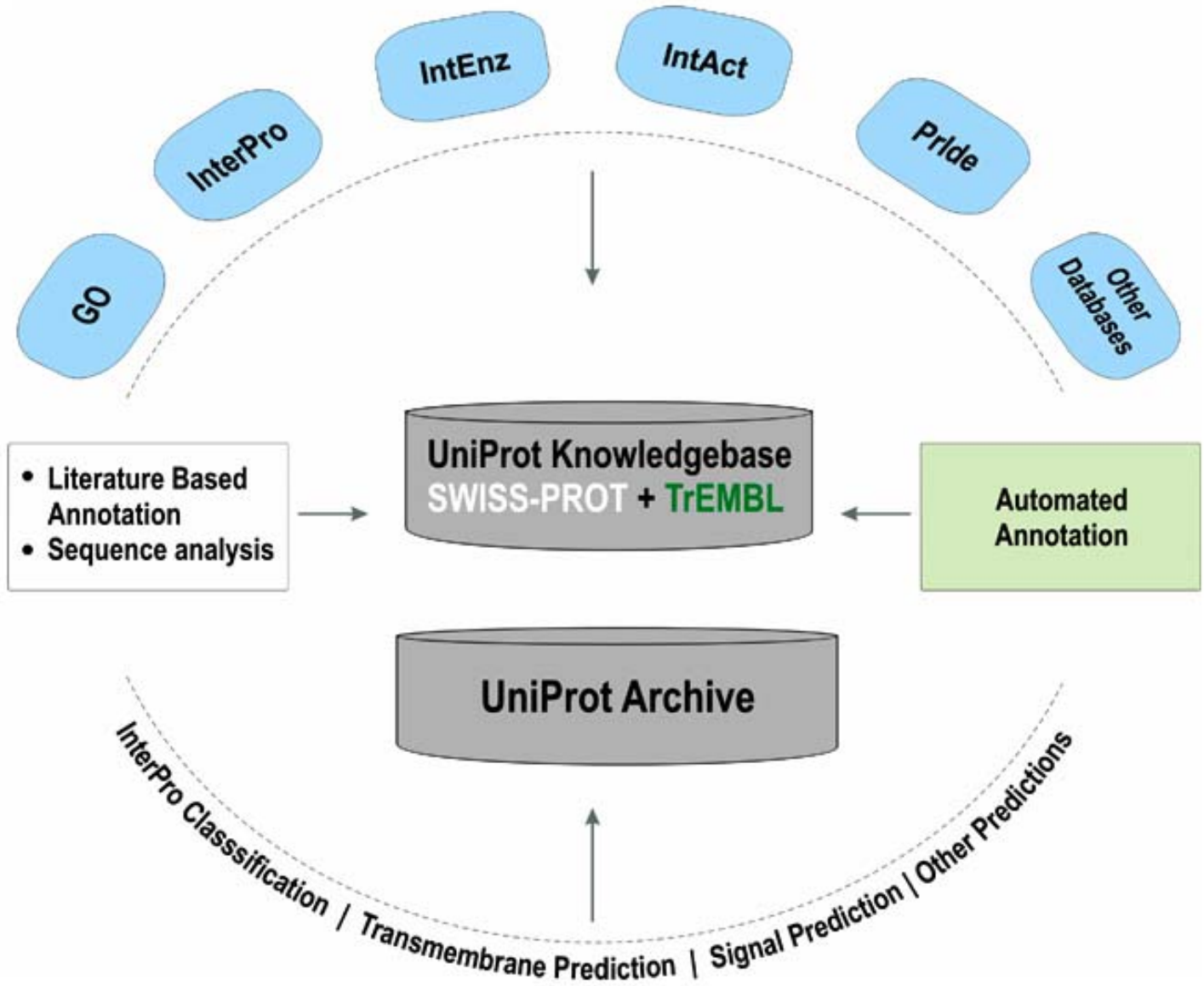
UPI0000000356 | UPI000000060F | UPI0000000C37 | UPI0000000DEF | UPI000004EF93 | UPI000000D91A | UPI000003060C

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	-

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	-
IPI	IPI00016786.1	1	Y	14-MAR-2003	02-JUN-2004	-
IPI	IPI00113849.1	1	Y	14-MAR-2003	02-JUN-2004	-
JPO	BD509573	1	Y	26-MAR-2003	11-JUN-2003	-
JPO	BD513687	1	Y	26-MAR-2003	11-JUN-2003	-
JPO	BD517764	1	Y	26-MAR-2003	11-JUN-2003	-
PDB	1GRN_A	1	Y	27-MAR-2003	09-JUN-2004	-
PDB	2NGR_A	1	Y	27-MAR-2003	09-JUN-2004	-
PIR	A39265	1	Y	11-APR-2003	11-MAY-2004	-
PIR	S57563	1	Y	11-APR-2003	11-MAY-2004	-
PIR archive	A39265	1	Y	31-MAR-2003	04-APR-2003	-
PIR archive	S57563	1	Y	31-MAR-2003	04-APR-2003	-
RefSeq release	NP_001782.1	1	Y	18-JAN-2004	10-JUN-2004	-
RefSeq release	NP_033991.1	1	Y	18-JAN-2004	10-JUN-2004	-
Swiss-Prot	P25763	1	N	29-MAR-2003	-	18-MAY-2003
Swiss-Prot varsplic	P21181-4	1	N	27-MAY-2003	-	19-MAR-2004
Swiss-Prot varsplic	P60766-2	1	Y	02-APR-2004	11-JUN-2004	-
Swiss-Prot varsplic	P60952-2	1	Y	02-APR-2004	11-JUN-2004	-
Swiss-Prot varsplic	P60953-2	1	Y	02-APR-2004	11-JUN-2004	-
TROME_HS	NT_004610_67_13	1	Y	30-DEC-2003	16-APR-2004	-
TROME_HS	NT_004610_67_14	1	Y	30-DEC-2003	16-APR-2004	-
TROME_HS	NT_004610_67_5	2	Y	30-DEC-2003	16-APR-2004	-
TROME_HS	NT_004610_68_12	1	N	11-NOV-2003	-	02-DEC-2003
TROME_HS	NT_004610_68_13	1	N	11-NOV-2003	-	02-DEC-2003
TROME_HS	NT_004610_68_4	1	N	11-NOV-2003	-	02-DEC-2003
TROME_MM	NT_039266_190_0	1	N	11-NOV-2003	-	08-JAN-2004
TROME_MM	NT_039267_143_0	1	Y	14-APR-2004	16-APR-2004	-
TrEMBLnew	AAH02711	1	N	29-MAR-2003	-	15-NOV-2003
TrEMBLnew	AAH03682	1	N	29-MAR-2003	-	15-NOV-2003
TrEMBLnew	AAH18266	1	N	29-MAR-2003	-	15-NOV-2003
TrEMBLnew	AAH60535	1	Y	01-NOV-2003	11-JUN-2004	-
TrEMBLnew	BAB22563	1	Y	29-MAR-2003	11-JUN-2004	-
TrEMBLnew	BAC35825	1	Y	29-MAR-2003	11-JUN-2004	-
TrEMBLnew	CAB57326	1	Y	29-MAR-2003	11-JUN-2004	-
UPO	AAE17065.1	1	Y	26-MAR-2003	11-JUN-2003	-
UPO	AAO95966.1	1	Y	11-JUN-2003	11-JUN-2003	-


IntEnz
IntAct
InterPro
Prlde
GO
Other Databases
UniProt Knowledgebase
SWISS-PROT + TrEMBL
• Literature Based Annotation
• Sequence analysis
Automated Annotation
UniProt Archive
InterPro Classsification / Transmembrane Prediction | Signal Prediction | Other Predictions

General information about the UniProt/Swiss-Prot entry	
Entry name	**CD42_MOUSE**
Primary accession number	**P60766**
Secondary accession numbers	P21181 P25763
Entered in Swiss-Prot	Release 18, 01-MAY-1991
Sequence was last modified	Release 18, 01-MAY-1991
Annotations were last modified	Release 44, 15-JUN-2004
Protein description	
Protein name	**Cell division control protein 42 homolog**
Synonyms	G25K GTP-binding protein
Origin of the protein	
Gene	CDC42
From	Mus musculus (Mouse)[TaxID:10090]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
References	
[1]	SEQUENCE FROM N.A. (ISOFORM 2). MEDLINE=93218723; PubMed=8464478; [NCBI, ExPASy, EBI, Israel, Japan] Miki T., Smith C.L., Long J.E., Eva A., Fleming T.P.; "Oncogene ect2 is related to regulators of small GTP-binding proteins."; Nature 362:462-465(1993).
[2]	SEQUENCE FROM N.A. (ISOFORM 2). TISSUE=Liver; MEDLINE=97368185; PubMed=9224952; [NCBI, ExPASy, EBI, Israel, Japan] Gong T.W., Shin J.J., Burmeister M., Lomax M.I.; "Complete cDNAs for CDC42 from chicken cochlea and mouse liver."; Biochim. Biophys. Acta 1352:282-292(1997).
[3]	SEQUENCE FROM N.A. (ISOFORM 1). STRAIN=C57BL/6; TISSUE=Brain; MEDLINE=97124841; PubMed=8954774; [NCBI, ExPASy, EBI, Israel, Japan] Marks P.W., Kwiatkowski D.J.; "Genomic organization and chromosomal location of murine Cdc42."; Genomics 38:13-18(1996).
[4]	INTERACTION WITH CDC42EP4. MEDLINE=99421943; PubMed=10490598; [NCBI, ExPASy, EBI, Israel, Japan] Joberty G., Perlungher R.R., Macara I.G.; "The Borgs, a new family of Cdc42 and TC10 GTPase interacting proteins.";

[4]	INTERACTION WITH CDC42EP4. MEDLINE=99421943; PubMed=10490598; [NCBI, ExPASy, EBI, Israel, Japan] Joberty G., Perlungher R.R., Macara I.G.; "The Borgs, a new family of Cdc42 and TC10 GTPase-interacting proteins."; Mol. Cell. Biol. 19:6585-6597(1999).
[5]	SUBUNIT OF A COMPLEX CONTAINING PARD6B; PARD3 AND PRKCZ, AND MUTAGENESIS OF THR-17. DOI=10.1038/35019573; MEDLINE=20394296; PubMed=10934474; [NCBI, ExPASy, EBI, Israel, Japan] Joberty G., Petersen C., Gao L., Macara I.G.; "The cell-polarity protein Par6 links Par3 and atypical protein kinase C to Cdc42."; Nat. Cell Biol. 2:531-539(2000).

Comments

FUNCTION	Plasma membrane-associated small GTPase which cycles between an active GTP-bound and an inactive GDP-bound state. In active state binds to a variety of effector proteins to regulate cellular responses. Involved in epithelial cell polarization processes. Causes the formation of thin, actin-rich surface projections called filopodia.
ENZYME REGULATION	Regulated by guanine nucleotide exchange factors (GEFs) which promote the exchange of bound GDP for free GTP, GTPase activating proteins (GAPs) which increase the GTP hydrolysis activity, and GDP dissociation inhibitors which inhibit the dissociation of the nucleotide from the GTPase.
SUBUNIT	Interacts with Zizimin1/DOCK9 which activates it by exchanging GDP for GTP. Interacts with PARD6A, PARD6B and PARD6G in a GTP-dependent manner. Part of a complex with PARD3, PARD6A or PARD6B and PRKCI or PRKCZ. Interacts with CDC42EP4.
ALTERNATIVE PRODUCTS	Alternative splicing;2 named isoforms [Display all isoform sequences in Fasta format] Name=1; Synonyms=Brain; IsoformId=P60766-1, P21181-1; This is the isoform sequence displayed in this entry. Name=2; Synonyms=Placental; IsoformId=P60766-2, P21181-4; Sequence=VSP_009844, VSP_009845;
SIMILARITY	Belongs to the small GTPase superfamily. Rho family. CDC42 subfamily.

Copyright

Cross-references

EMBL	L11318; AAA37410.1; -. [EMBL/ GenBank/ DDBJ] [CoDingSequence] U37720; AAC00028.1; -.[EMBL/ GenBank/ DDBJ] [CoDingSequence] L78075; AAB40051.1; -. [EMBL/ GenBank/ DDBJ] [CoDingSequence]

	U37720; AAC00028.1; -.[EMBL/ GenBank/ DDBJ] [CoDingSequence] L78075; AAB40051.1; -.[EMBL/ GenBank/ DDBJ] [CoDingSequence]
MGD	MGI:106211; Cdc42.
GO	

Cellular component	filopodium	GO:0030175	non-traceable author statement
Molecular function	Rho small monomeric GTPase activity	GO:0003931	non-traceable author statement
Biological process	actin filament organization	GO:0007015	non-traceable author statement
[QuickGO]			

InterPro	IPR003578; GTPase_Rho. IPR001806; Ras_trnsfrmng. IPR005225; Small_GTP. Graphical view of the domain strcuture
Additional cross-references	
GeneLynx	CDC42; Mus musculus.
SOURCE	CDC42; Mus musculus.
Ensembl	P60766; Mus musculus.[Entry/Contig].
Prodom	[Domain structure/ List of seq. sharing at least 1 domain].
HOVERGEN	[Familiy/Alignment/Tree].
BLOCKS	P60766.
ProtoNet	P60766.
ProtoMap	P60766.
PRESAGE	P60766.
DIP	P60766.
ModBase	P60766.
SMR	P60766; 34B44F9225EC106B.
SWISS-2DPAGE	Get region on 2D PAGE.
UniRef	View cluster of proteins with at least 50%/90% identity.
Keywords	
	GTP-binding
	Lipoprotein
	Prenylation
	Alternative splicing
	Methylation

	Alternative splicing
	Methylation

Features

Type	From	To	Length	Description	Feature ID
CHAIN	1	188	188	Cell division control protein 42 homolog.	
PROPEP	189	191	3	Removed in mature form.	
NP_BIND	10	17	8	GTP (BY SIMILARITY).	
DOMAIN	32	40	9	Effector region (POTENTIAL).	
NP_BIND	57	61	5	GTP (BY SIMILARITY).	
NP_BIND	115	118	4	GTP (BY SIMILARITY).	
MOD_RES	188	188		Cysteine methyl ester (in mature form) (BY SIMILARITY).	
LIPID	188	188		S-geranylgeranyl cysteine (BY SIMILARITY).	
VARSPLIC	163	163		R -> K (in isoform 2)	VSP_009844
VARSPLIC	182	191		TQPKR KCCIF -> PKKSR RCVLL (in isoform 2)	VSP_009845
MUTAGEN	17	17		T->N: Constitutively inactivate. Abolishes interaction with PARD6 proteins.	

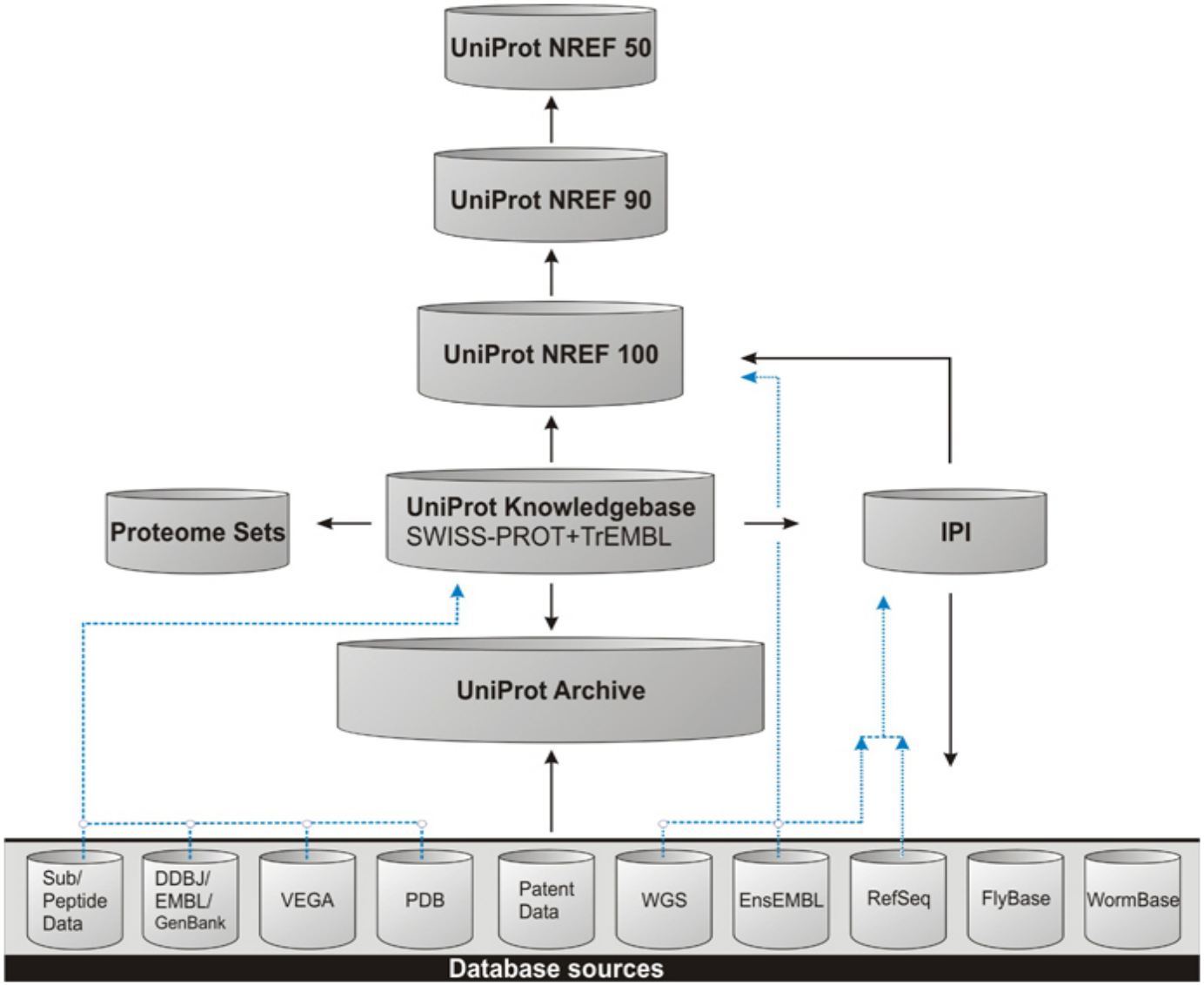
FeatureTableViewer Feature aligner

Sequence information

Length	191 AA
Molecular weight	21310 Da
CRC64	34B44F9225EC106B [This is a checksum on the sequence]

```
         10         20         30         40         50         60
          |          |          |          |          |          |
MQTIKCVVVG DGAVGKTCLL ISYTTNKFPS EYVPTVFDNY AVTVMIGGEP YTLGLFDTAG
         70         80         90        100        110        120
          |          |          |          |          |          |
QEDYDRLRPL SYPQTDVFLV CFSVVSPSSF ENVKEKWVPE ITHHCPKTPF LLVGTQIDLR
        130        140        150        160        170        180
          |          |          |          |          |          |
DDPSTIEKLA KNKQKPITPE TAEKLARDLK AVKYVECSAL TQRGLKNVFD EAILAALEPP
        190
          |
ETQPKRKCCI F
```

UniRef100 | UniRef90 | UniRef50


UniProt NREF 50
UniProt NREF 90
UniProt NREF 100
Proteome Sets
UniProt Knowledgebase
SWISS-PROT+TrEMBL
IPI
UniProt Archive
Sub/ Peptide Data
DDBJ/ EMBL/ GenBank
VEGA
PDB
Patent Data
WGS
EnsEMBL
RefSeq
FlyBase
WormBase
Database sources

Entry UniRef100_P60953

New Query | Download Entry | Bookmark UniRef (Ctrl+D)

Viewers: XML | SRS | ExPASy | PIR

Accession Number		UniRef100_P60953	
Name		Cell division control protein 42 homolog	
UniRef90 ID		UniRef90_P60953	
Sequence		MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNYAVTVMIGGEPYTLGLFDTAG QEDYDRLRPLSYPQTDVFLVCFSVVSPSSFENVKEKWVPEITHHCPKTPFLLVGTQIDLR DDPSTIEKLAKNKQKPITPETAEKLARDLKAVKYVECSALTQRGLKNVFDEAILAALEPP ETQPKRKCCIF	
	Length	191	
	CRC64	34B44F9225EC106B	
Representative	UniProt ID	CD42_HUMAN	
	UniProt Accession Numbers	P60953, P21181, P25763.	
	UniParc ID	UPI0000000CF6	
	Protein Name	Cell division control protein 42 homolog	
	Source Organism	Homo sapiens	
	NCBI Taxonomy ID	9606	
Member	UniProt ID	CD42_MOUSE	
	UniProt Accession Numbers	P60766, P21181, P25763.	
	UniParc ID	UPI0000000CF6	
	Protein Name	Cell division control protein 42 homolog	
	Source Organism	Mus musculus	
	NCBI Taxonomy ID	10090	
	Length	191	
	Overlap Region	1 - 191	
	Member	UniProt ID	CD42_CANFA
		UniProt Accession Numbers	P60952, P21181, P25763.
	UniParc ID	UPI0000000CF6	
	Protein Name	Cell division control protein 42 homolog	
	Source Organism	Canis familiaris	
	NCBI Taxonomy ID	9615	
	Length	191	
	Overlap Region	1 - 191	

Entry UniRef90_P60953

New Query | Download Entry | Bookmark UniRef (Ctrl+D)

Viewers: XML | SRS | ExPASy | PIR

Accession Number		UniRef90_P60953
Name		Cell division control protein 42 homolog related cluster
UniRef100 ID		UniRef100_P60953
Sequence		MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNYAVTVMIGGEPYTLGLFDTAG QEDYDRLRPLSYPQTDVFLVCFSVVSPSSFENVKEKWVPEITHHCPKTPFLLVGTQIDLR DDPSTIEKLAKNKQKPITPETAEKLARDLKAVKYVECSALTQRGLKNVFDEAILAALEPP ETQPKRKCCIF
	Length	191
	CRC64	34B44F9225EC106B
Representative	UniProt ID	CD42_HUMAN
	UniProt Accession Numbers	P60953, P21181, P25763.
	Protein Name	Cell division control protein 42 homolog
	Source Organism	Homo sapiens
	NCBI Taxonomy ID	9606
Member	UniProt ID / Accession Number	Q7PQZ1
	Protein Name	ENSANGP00000023777
	Source Organism	Anopheles gambiae str. PEST
	NCBI Taxonomy ID	180454
	UniRef100 ID	UniRef100_Q7PQZ1
Member	UniProt ID / Accession Number	Q86DH9
	Protein Name	Cdc42
	Source Organism	Aplysia californica
	NCBI Taxonomy ID	6500
	UniRef100 ID	UniRef100_Q86DH9
Member	UniProt ID / Accession Number	Q86RA4
	Protein Name	GTP-binding protein
	Source Organism	Brugia malayi (Filarial nematode worm)
	NCBI Taxonomy ID	6279
	UniRef100 ID	UniRef100_Q86RA4
Member	UniProt ID / Accession Number	Q9U9S6
	Protein Name	CDC42 protein
	Source Organism	Drosophila melanogaster
	NCBI Taxonomy ID	7227
	UniRef100 ID	UniRef100_Q9U9S6
Member	UniProt ID / Accession Number	Q9U9S5
	Protein Name	CDC42 protein

Member	UniProt ID / Accession Number	Q9U9S5
	Protein Name	CDC42 protein
	Source Organism	Drosophila melanogaster
	NCBI Taxonomy ID	7227
	UniRef100 ID	UniRef100_Q9U9S5
Member	UniProt ID / Accession Number	Q9U9S4
	Protein Name	CDC42 protein
	Source Organism	Drosophila melanogaster
	NCBI Taxonomy ID	7227
	UniRef100 ID	UniRef100_Q9U9S4
Member	UniProt ID / Accession Number	Q9U9S3
	Protein Name	CDC42 protein
	Source Organism	Drosophila melanogaster
	NCBI Taxonomy ID	7227
	UniRef100 ID	UniRef100_Q9U9S3
Member	UniProt ID	CD42_CHICK
	UniProt Accession Numbers	Q90694.
	Protein Name	Cell division control protein 42 homolog
	Source Organism	Gallus gallus
	NCBI Taxonomy ID	9031
	UniRef100 ID	UniRef100_Q90694
Member	UniProt ID / Accession Number	P60953-2
	Protein Name	Splice isoform 2 of P60953
	Source Organism	Homo sapiens
	NCBI Taxonomy ID	9606
	UniRef100 ID	UniRef100_P60953-2
Member	UniProt ID	CC42_DROME
	UniProt Accession Numbers	P40793, Q9V465.
	Protein Name	Cdc42 homolog
	Source Organism	Drosophila melanogaster
	NCBI Taxonomy ID	7227
	UniRef100 ID	UniRef100_P40793
Member	UniProt ID	CC42_ANOGA
	UniProt Accession Numbers	Q17031, Q93110.
	Protein Name	CDC42 homolog
	Source Organism	Anopheles gambiae
	NCBI Taxonomy ID	7165
	UniRef100 ID	UniRef100_Q17031

Common problems for UniProt to make use of Proteomics data

- False peptide and protein IDs
- Use of outdated, redundant and incomplete databases, no use of annotated features like known variants and PTMs
- Redundant Identifications
- IDs based on incomplete data
- IDs based on different Organisms
- Blind Trust in DE lines
- Lack of Proteomics data repositories or lack of collaboration between them

IPI International Protein Index: Statistics Page

Current composition of Human IPI

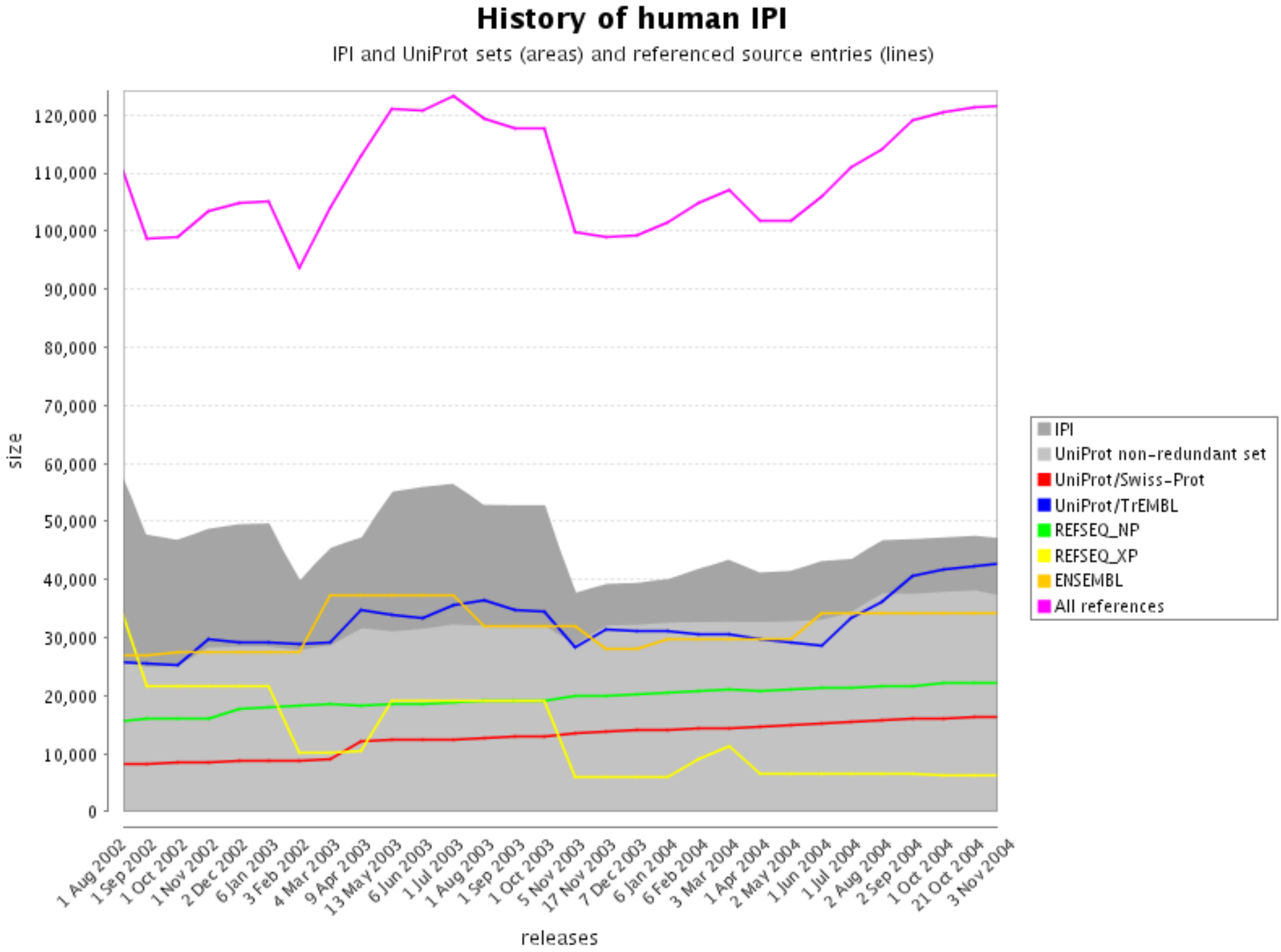
The following table describes the latest version of IPI (Human, 3.00), released on **Wed, 3 Nov 2004** and assembled using the publicly released data available in the source databases on **Wed, 3 Nov 2004**
IPI data sets are released monthly, usually at the start of each calendar month.

Click here to download the IPI for Human.

Click here to get the statistics for Mouse.

Click here to get the statistics for Rat.

Number of IPI entries	
Number of entries in IPI	47094
Total number of entries referenced by IPI	121596
Number of references to UniProt	59132
Number of references to RefSeq	28353
Number of references to ENSEMBL	34111
Composition of IPI entries	
Number of IPI entries pointing only to UniProt	9308
Number of IPI entries pointing only to RefSeq	3910
Number of IPI entries pointing only to UniProt and RefSeq	2055
Number of IPI entries pointing only to ENSEMBL	4547
Number of IPI entries pointing only to UniProt and ENSEMBL	6236
Number of IPI entries pointing only to RefSeq and ENSEMBL	1512
Number of IPI entries pointing only to UniProt and RefSeq and ENSEMBL	19526
Download	FASTA format UniProt format

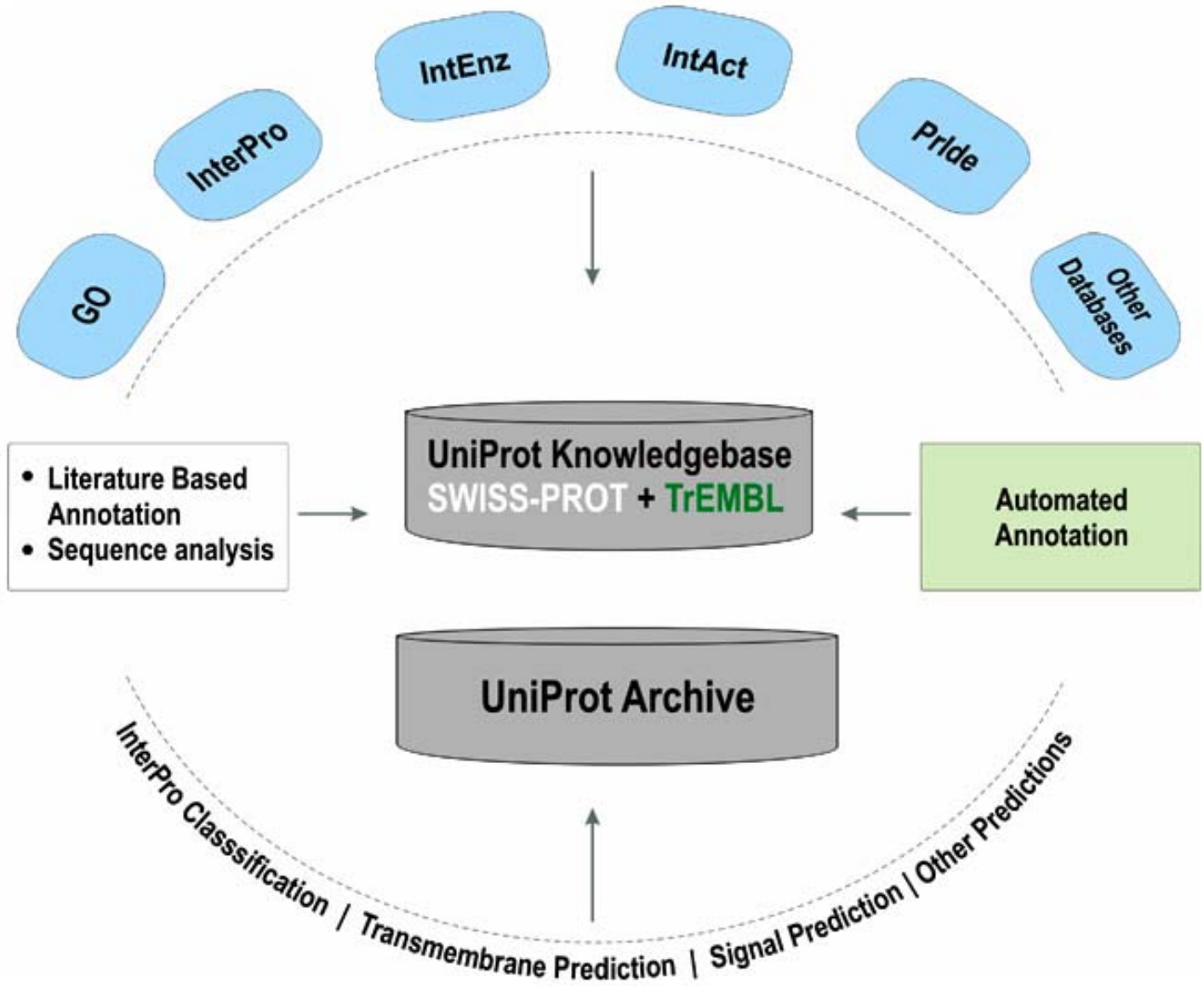

History of human IPI
IPI and UniProt sets (areas) and referenced source entries (lines)
size
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
1 Aug 2002
1 Sep 2002
1 Oct 2002
1 Nov 2002
2 Dec 2002
6 Jan 2003
3 Feb 2002
4 Mar 2003
9 Apr 2003
13 May 2003
6 Jun 2003
1 Jul 2003
1 Aug 2003
1 Sep 2003
1 Oct 2003
5 Nov 2003
17 Nov 2003
7 Dec 2003
6 Jan 2004
6 Feb 2004
3 Mar 2004
1 Apr 2004
2 May 2004
1 Jun 2004
1 Jul 2004
2 Aug 2004
2 Sep 2004
1 Oct 2004
21 Oct 2004
3 Nov 2004
releases
IPI
UniProt non-redundant set
UniProt/Swiss-Prot
UniProt/TrEMBL
REFSEQ_NP
REFSEQ_XP
ENSEMBL
All references

What protein has been identified?

- The ID is ambiguous due to matching against
  - many proteins derived from different transcripts from different genes
  - many proteins derived from different transcripts from one gene
- The ID is unambiguously pointing to a protein derived from a certain transcript of a certain gene, but …

Envisaged use of Proteomics data by UniProt

- Verification of existence of gene products:
  - The ability to define the major splice variants (by tissue) will lead to more accurate structure/function predictions due to specific knowledge of exon/domain structure
  - Avoid false positive protein entries from ab initio gene predictions and spurious ORFs
  - Identify AA-changing SNPs that are validated through *in vivo* conformation at the protein level in primary human tissue
- Temporal and spatial information on proteins
- Protein-Protein Interactions
- Existence and Role of PTMs


IntEnz
IntAct
InterPro
Prlde
GO
Other Databases
UniProt Knowledgebase
SWISS-PROT + TrEMBL
• Literature Based Annotation
• Sequence analysis
Automated Annotation
UniProt Archive
InterPro Classsification / Transmembrane Prediction | Signal Prediction | Other Predictions

Proteomics Standards

- HUPO PSI standards
  - MI, mzData, mzIdent,
  - Without such standards no data harvesting possible and no exchange of data between repositories
- Standards for QA/QC
- Generally already well accepted but need support by journal and funding agencies to speed up implementation and creation of repositories

What we need

- Standards for data exchange
- Standards for QA/QC
- Proteomics repositories (Archives)
- Added-value Proteomics databases
- Data sharing: Mandatory submission of data
  - RO1 vs community resources: at publication vs prepublication (NIH/WT Fort Lauderdale agreement)
- Collaboration on many levels, between
  - databases (like DDBJ/EMBL/GenBank)
  - funding agencies, journals, databases and data providers to enforce data submission and exchange
  - vendors, funding agencies, journals, databases and data providers to create and enforce standards

Funding

- EMBL
- Industry Contributions
- European Commission
- NIH
- MRC
- BBSRC
- HUPO
- IUPHAR

Sequence Database Group

UniProt, InterPro, IntAct, etc, at EBI and SIB

- **Group leaders**: Rolf Apweiler, Amos Bairoch
- **Co-ordinators**:Wolfgang Fleischmann, Henning Hermjakob, Michele Magrane, Maria-Jesus Martin, Nicola Mulder, Claire O'Donovan, Manuela Pruess
- **Annotators/curators**: Yasmin Alam-Faruque, Philippe Aldebert, Nicola Althorpe, Andrea Auchincloss, Kirsty Bates, Marie-Claude Blatter Garin, Brigitte Boeckmann, Silvia Braconi Quintaj, Paul Browne, Evelyn Camon, Wei Mun Chan, Danielle Coral, Elisabeth Coudert, Tania de Oliveria Lima, Kirill Degtyarenko, Sylvie Dethiollaz, Emily Dimmer, Ruth Eberhardt, Marcus Ennis, Ann Estreicher, Livia Famiglietti, Nathalie Farriol-Mathis, Stephanie Federico, Serenella Ferro, Gill Fraser, John Gamble, John Steve Garavelli, Raffaella Gatto, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Janet Harwood, Ursula Hinz, Chantal Hulo, Julius Jacobson, Janet James, Florence Jungo, Vivien Junker, Youla Karavidopoulou, Maria Krestyaninova, Kati Laiho, Vivian Lee, Minna Lehvaslaiho, David Lonsdale, Jennifer McDowall, Michelle McHale, Karine Michoud, Virginie Mittard, Madelaine Moinat, Markiyan Oliynyk, Sandra Orchard, Sandrine Pilbout, Sylvain Poux, Sorogini Reynaud, Catherine Rivoire, Bernd Röchert, Michel Schneider, Christian Sigrist, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Sandra van den Broek, Bob Vaughan, Eleanor Whitfield
- **Programmers:** Daniel Barrell, David Binns, Michael Darsow, Ujjwal Das, Eduardo de Castro, Paula de Matos, Jorge Duarte, Alexander Fedotov, Rodrigo Fernandez, Astrid Fleischmann, Elisabeth Gasteiger, Alain Gateau, Federico Garcia-Diez, Andre Hackmann, Ivan Ivanyi, Eric Jain, Phil Jones, Alexander Kanapin, Samuel Kerrien, Paul Kersey, Asif Kibria, Ernst Kretschmann, Corinne Lachaize, Chris Lewington, Xavier Martin, John Maslen, Peter McLaren, Rupinder Singh Mazara, Lorna Morris, Sugath Mudali, John O'Rourke, Gulam Patel, Sam Patient, Isabelle Phan, Emmanuel Quevillon, Antony Quinn, Astrid Rakow, Muruli Rao, Nicole Redaschi, Siamak Sobhany, Chris Taylor, Nisha Vinod, Daniela Wieser, Allyson Williams, Dan Wu
- **Research staff:** Kristian Axelsen, Pierre-Alain Binz, Nicolas Hulo, Anne-Lise Veuthey
- **Administrative assistance:** Veronique Mangold, Claudia Sapsezian, Margaret Shore-Nye, Veronique Verbegue
- **Students:** Pavel Dobrokhotov, Alexandre Gattiker, various MCF, etc