UniProt complete proteomes

Eleanor Stanley

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

UniProt consortium

- Formed in 2002 (10th anniversary next year)
- Previously known as “Swiss-Prot” since 1986

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

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

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

- UniProtKB is UniProt Knowledgebase, TrEMBL and Swiss-Prot entries

UniProt complete proteomes

- Longstanding project, 3086 proteomes that are spread over the entire taxonomic range,
  - 49% bacteria
  - 42% viruses
  - 5.5% eukaryota
  - 3.5% archaea
- 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

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 are:
  - INSDC
  - Ensembl
- With a view to extending this to include:
  - Ensembl Genomes
  - RefSeq

Proteome products

- UniProt complete proteomes for ~3000 species that include:
  - INSDC defined species
  - Ensembl species
- Proteome sets for 12 IPI species available from FTP
- UniProt reference proteomes (including QfO defined species)
- QfO proteome sets
- GO annotation proteome sets
  - Files of GO annotations for each proteome

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 and more.

NEWS

UniProt release 2011_09 - Sep 21, 2011

Reference proteomes in UniProt

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL

› Forthcoming changes

› News archives

Follow @uniprot · 49 followers

SITE TOUR

INSDC proteomes

- Data captured for each proteome:
  - INSDC accessions for all components that constitute a complete genome
  - Taxonomy identification of the strain/species sequenced
  - Genome project references relevant to a chromosome or entire genome, addition of any subsequent genome reannotation papers
  - Addition of regular expression to recognise genome locus names

- Standardisation of UniProt entries within a proteome:
  - Replace genome submission ref with publication
  - Maintenance of genome locus names
  - Addition of Complete proteome keyword and Reference proteome keyword (where appropriate)

Ensembl import

- Current species list for 2011_09:
  - Chicken, cow, dog, horse, human, macaque, marmoset, mouse, opossum, pig, platypus, rat, sea squirt, xenopus, zebrafish
  - Will include further species in future UniProt releases, for example Anole lizard, gibbon, little flying bat, panda, rabbit and turkey in 2011_10

- Ensembl UniProt 1:1 mapping
  - An ENSP is mapped to a UniProtKB entry where the translation sequences are 100% identical (identity and sequence length)
  - Those ENSPs that are missing in UP generate new UniProtKB/TrEMBL entries

- All UniProtKB/TrEMBL entries will:
  - have Ensembl DR line(s)
  - be enriched with automatic annotation procedures

Ensembl import

Gene: GSTZ1 (ENSG00000100577)

Description glutathione transferase zeta 1 [Source:HGNC Symbol;Acc:4643]

Location Chromosome 14: 77,787,230-77,797,939 forward strand.

Transcripts ⊟ There are 5 transcripts in this gene

Show/hide columns Filter:

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
GSTZ1-201	ENST00000216465	1334	ENSP00000216465	216	Protein coding	CCDS9858
GSTZ1-202	ENST00000349555	1208	ENSP00000314404	174	Protein coding	CCDS9859
GSTZ1-203	ENST00000361389	1054	ENSP00000354959	161	Protein coding	CCDS9860
GSTZ1-204	ENST00000393734	1424	ENSP00000377335	161	Protein coding	CCDS9860
GSTZ1-205	ENST00000393741	859	ENSP00000377342	188	Protein coding	-

← MAAI_HUMAN Isoform 1
← AGNED0_HUMAN
≻ MAAI_HUMAN Isoform 2
← Create new TrEMBL record

Gene summary *help*

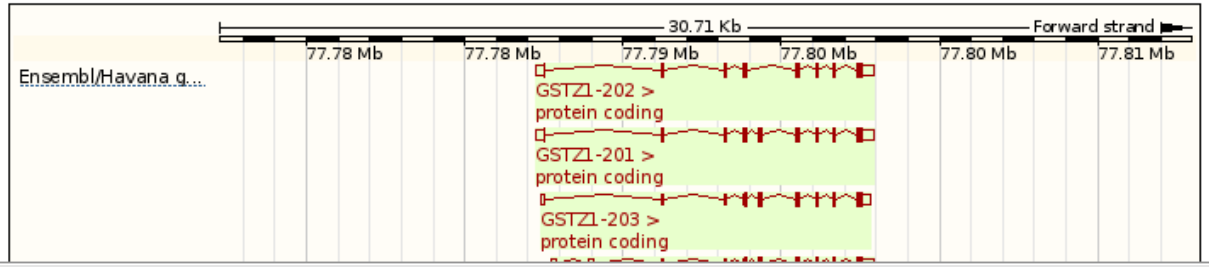
Name GSTZ1 (HGNC Symbol)

Synonyms GSTZ1-1, MAAI, MAI [To view all Ensembl genes linked to the name click here.]

CCDS This gene is a member of the Human CCDS set: CCDS9858, CCDS9859, CCDS9860

Gene type Known protein coding

Prediction Method Transcripts were annotated by the Ensembl genebuild.



IPI

- http://www.ebi.ac.uk/IPI
- Launched in 2001 to cover the gaps in gene predictions between different databases
- An integrated database that clusters protein sequences from different databases (UniProt, Ensembl, Refseq, H-inv, Vega, TAIR) to provide non-redundant complete data sets for human, mouse, rat, zebrafish, arabidopsis, chicken and cow.
- One of the major uses is as a reference database for mass spectrometric (MS) identification of peptides (by sequence similarity searching).

- 27th September 2011 is the final IPI release, now it will exist as archive only

UniProt IPI

- Original 7 species:
  - Arabidopsis thaliana, Bos taurus, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus, Rattus norvegicus
- Two new species from Ensembl:
  - Canis familiaris, Sus scrofa
- User request (EBI search and others) species:
  - Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae

- From 2011_07 UniProt is making IPI-style sets for 12 species

- Sets comprise all the UniProtKB/Swiss-Prot manual annotated protein sequences supplemented by protein sequences in UniProtKB/TrEMBL from high quality predictions cross-referenced or imported from Ensembl

UniProt Reference proteomes

- UniProtKB will define a subset of the Complete proteomes as being 'Reference proteomes'.
- Criteria for the species/strain/isolate choosen is that they are:
  - a model organism
  - species defined standard for a particular user community (eg QfO)
  - provide broad coverage of the tree of life
- All UniProtKB entries for these chosen species will have a new KW: Reference proteome, in addition to the KW Complete proteome

UniProt Reference proteomes

- Current list is 506 reference proteome species
- First available in release 2011_09
- Set will be continuously reviewed as new proteomes of interest become available and as existing taxonomic classifications are revised

Experimental annotation in UniProt proteome entries

- The reference or complete proteome per species/strain/isolate will contain all experimental data and literature for that species/strain/isolate
- Information from other strains will be propagated by similarity when appropriate

QfO Reference Proteomes

- 66 proteomes including
  - 12 GO Reference Genome project
    - Arabidopsis thaliana, Caenorhabditis elegans, Candida albicans, Danio rerio, Dictyostelium discoideum, Drosophila melanogaster, Escherichia coli, Gallus gallus, Homo sapiens, Mus musculus, Plasmodium falciparum, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe
  - 45 UniProt only species
  - 9 Ensembl only species

- Provide a proteome dataset of 1 gene 1 gene product
- Made once a year, so current set is based on UniProt 2011_04 and Ensembl Release 61
- Generated as a standardized set for ortholog prediction algorithms and allow benchmarking

QfO Reference proteome pipeline

- For each proteome we:
  - Identify all the protein coding genes on the genome
  - For each gene identify all translations available
  - Retrieve the longest translation for each gene

- Take into account biological complications:
  - Within genes: alternative transcripts can translate identical sequences
  - Between genes: duplicated genes can have 100% identical translation sequence (e.g. Histone Family)

  - These need to be correctly reflected in the fasta file of unique translations and in the gene/gene product mappings in the gp2protein file.

GO annotation proteome files

- Species-specific proteomes sets of GO annotations
- For the species to be included, more than 25% of the UniProt entries within the proteome must have GO annotation.
- For those model organism species which are provided with an ‘IPI-style’ proteome set, additional filtering has been carried out to remove redundant electronic annotations where multiple automatic annotation methods have predicted the same term or where one method has applied a less granular GO term.
- http://www.ebi.ac.uk/GOA/proteomes.html

Important limitation of a GO proteome file

- UniProt does not use gp2protein files for cross referencing a MOD to UniProt entries, instead the presence/absence of a MOD cross reference in a UniProt entry relies on separate mapping data provided by the MOD
- If the mapping file lacks up to date UniProt accessions then the linking between MOD and UniProt will be compromised
- Similarly, species specific GO annotation files will lack manual annotations if the supplied gp2protein maps from a MOD id to a UP accession not in the complete proteome
- However these annotations will be available from the GOA UniProt annotation file that contains all UniProt entries

Data and documentation

- UniProt Complete proteomes release 2011_09, Sept 21st 2011
  - http://www.uniprot.org/taxonomy/complete-proteomes
  - http://www.uniprot.org/faq/15
- For more information on Reference proteomes
  - http://www.uniprot.org/taxonomy/complete-proteomes
  - http://www.uniprot.org/faq/47
  - http://www.uniprot.org/news/2011/09/21/release
- How to retrieve proteome sets:
  - http://www.uniprot.org/faq/38

Data and documentation

- Ensembl import proteomes
  - http://www.uniprot.org/news/2011/05/03/release
- Proteome data sets for IPI species
  - http://www.uniprot.org/news/2011/06/28/release
- QfO proteomes from UniProt 2011_04 and Ensembl Release 61
  - http://www.ebi.ac.uk/reference_proteomes/
- QfO consortium page
  - http://questfororthologs.org/

Useful Resources

- Tony Sawford, in the UniProt-GOA team, assists groups in maintaining an up-to-date gp2protein file. However the below UniProtKB files might be of use:
  - gp_information file, containing metadata for valid primary UniProt accessions
  - ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gp_information.goa_uniprot.gz
  - List of secondary accession numbers
  - ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/sec_ac.txt
  - List of deleted accession numbers UniProtKB/Swiss-Prot
  - ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/delac_sp.txt
  - List of deleted accession numbers UniProtKB/TrEMBL
  - ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/delac_tr.txt

Future tasks

- Introduce an Ensembl Genome import pipeline
- Evaluate the RefSeq database
- Provide an additional set of 1 sequence per gene for all species (EBI search)
  - GO annotation files will be created using these protein sets as soon as they become available.
- Improve the Complete proteomes web interface
- Improve website functionality to include a tool to gather all translations for a gene of interest

Acknowledgements

- UniProt consortium for Complete proteomes
  - Benoit Bely
  - Jie Luo
  - Boris Bursteinas
  - Eleanor Stanley
  - Maria Martin
- GO consortium
- QfO consortium
- Ensembl

Additional information

- How often sequence records are reviewed?
  - UniProtKB/Swiss-Prot: time devoted to a review of the literature depends on the priority of the species/protein family but there are some triggers ...
- How often sequence records are automatically updated?
  - UniProtKB/TrEMBL: With an update of the INSDC protein_id, or Ensembl ENSP
- What prompts a review?
  - underlying nucleotide entry changing
  - new papers identified by curators
  - updates from collaborators
  - updates from users
- How and when are fragments and isoforms merged into a single record?
  - When creating a Swiss-Prot entry, all available sequences/isoforms are merged.
  - we don’t annotate fragments as the only sequence source except for those sequences which have biochemical characterization such as submissions.
  - new isoforms are looked for and merged if thought to add value to the entry.