How UniProt serves the proteomics community and makes use of the proteomics data.

Benoît Bely[1], Emanuele Alpi[1], Alan Wilter Sousa da Silva[1], Guoying Qi[1], Maria Jesus Martin[1] and the UniProt consortium[1, 2, 3, 4]

[1] European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2] SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, [3] Protein Information Resource, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA and [4] Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven Street North West, Suite 1200, Washington, DC 20007, USA

[bbely, ealpi, awilter, gqi, martin]@ebi.ac.uk

Abstract

In order to better serve the proteomics community, UniProt provides proteins sets (proteomes) for all complete genomes publicly available. These proteomes can be retrieved via the UniProt FTP and website. In order to complete protein annotation using proteomics data, UniProt has developed a pipeline which maps publicly available lists of experimentally identified peptides to proteins sequences in UniProtKB taking into account peptide unicity. This will be used to set evidence for the existence of a protein at the protein level (PE=1) and will allow users to identify proteins with experimentally proved sequence information.

1 UniProt proteomes replaces IPI.

IPI (International Protein Index) (Kersey, *et al.*, 2004) was an integrated database for proteomics identification. Launched in 2001, IPI was a great resource and its success came from the possibility of providing complete non-redundant data sets by clustering proteins from UniProtKB/SwissProt, UniProtKB/TrEMBL, Ensembl and RefSeq databases. UniProtKB (Apweiler, et al., 2004) contains protein sequences from the CDS translations of te INSDC nucleotide databases, Ensembl, PDB and RefSeq (The UniProt Consortium, 2009) and therefore, IPI was discontinued in September 2011 (Griss, *et al.*, 2011).

UniprotKB has now removed cross-reference links in the 2014_01 release, as the data sets are no longer comparable. However, in 2014 there are articles in PubMed still using direct data from outdated IPI.

There are two ways to retrieve complete and up-to-date proteome sets from UniProt. First, by direct download from the proteome directory of the UniProt FTP[i] for *A. thaliana, B. taurus, C. elegans, C. l. familiaris, D. melanogaster, D. rerio, G. gallus, H. sapiens, M. musculus, R. norvegicus, S. cerevisiae S288c and S. scrofa*. Secondly, by individual proteome downloads from the UniProt web site. (http://www.uniprot.org/proteomes).

[i] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/

2 UniProt proteomes.

2.1 What are the Reference Proteomes?

In the UniProt 2014_09 release, there are 2,290 Reference Proteomes out of a total of 5,548 Proteomes. We define a *Proteome* as the set of proteins encoded by a completely sequenced genome (The UniProt Consortium, 2012); other proteins that can't be mapped to the genome and for which there is no experimental data available, are left outside the proteome. For example in UniProtKB 2014_09 release *homo sapiens* entries are 138,560 when the human proteome contains 20,328 conical proteins and 64,567 UniProtKB/SwissProt and UniProtKB/TrEMBL isoforms.

Reference Proteomes are subset of proteomes (manually and algorithmically) selected. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. Proteins in general and proteomes in particular are not heavenly spread across the taxonomy tree. Generally it is easier to sequence, assemble and analyze, viruses and bacteria genomes than eukaryotic genomes. In UniProt 2014_09 there are 3,545(64%) bacteria, 1,386(25%) viruses, 435(8%) eukaryota and 182(3%) archaea genomes. By selecting the representative proteomes of some highly redundant genomes, UniProtKB is making Reference Proteomes for 40%(1,410) of bacterial proteomes, 17%(382) of viral proteomes, 89%(387) of eucaryotic proteomes and 61%(111) archeal proteomes. For example, amongst the 76 proteomes of *E. coli*, the 38 of *M. tuberculosis* and the 51 of *S. aureus* respective strains K12, NCTC 8325 and ATCC 25618 / H37Rv have been selected to become Reference Proteomes for those species.

2.2 How do I retrieve my favorite proteomes?

Since a new UniProt website was launched in the 2014_08 release of September 2014, a new Proteome portal is available for retrieving proteome data sets. This new proteome portal allows searching for species name and NCBI taxonomy identifier. This leads to a specific proteome page for the selected species containing description, genomic assembly and protein information. Individual downloads for species and chromosomes are available. Notably this can be performed on all proteomes regardless if they are reference ones or not.

In 2009, UniProt joined the Quest for Orthologs consortium (QfO) (Dessimoz, *et al.*, 2012) in order to provide standard protein sets for tool benchmarking; UniProt has been providing protein data sets for Reference Proteomes (http://www.ebi.ac.uk/reference_proteomes) for all kingdoms, excluding viruses. For each Reference Proteomes, UniProt provides five files (1) a fasta file containing canonical protein sequences; (2) a fasta file with the additional isoforms of a given gene; (3) a file containing the mapping between UniProtKB accessions and the corresponding gene; (4) a fasta file with the underlining canonical DNA sequences for each gene in a proteome and (5) a file containing mappings between UniProtKB accessions and many external databases and their identifiers.

In early 2015, UniProt will provide these data sets via the UniProtKB FTP site.

3 UniProt proteomics analysis pipeline

3.1 The methodology

UniProt has set up a pipeline to look into filtered repository peptides that are uniquely mapped to a single or group of UniProtKB sequences. One of the key points is how to evaluate peptide unicity. Among the ways for doing that, sequence-centric uniqueness versus gene-centric uniqueness (see below) were considered (Alpi, et al., In Press). The process consists in 3 steps (cf.: Figure 1): (1) build a UniProtKB species specific, non-redundant and filtered list of *in-silico* digested unique peptides; (2) retrieve and filter the corresponding MS-proteomics repository species-specific list of all experimentally identified peptides; (3) build the mapping between list (1) and list (2) depending on how peptide unicity was evaluated in step (1)(see below).

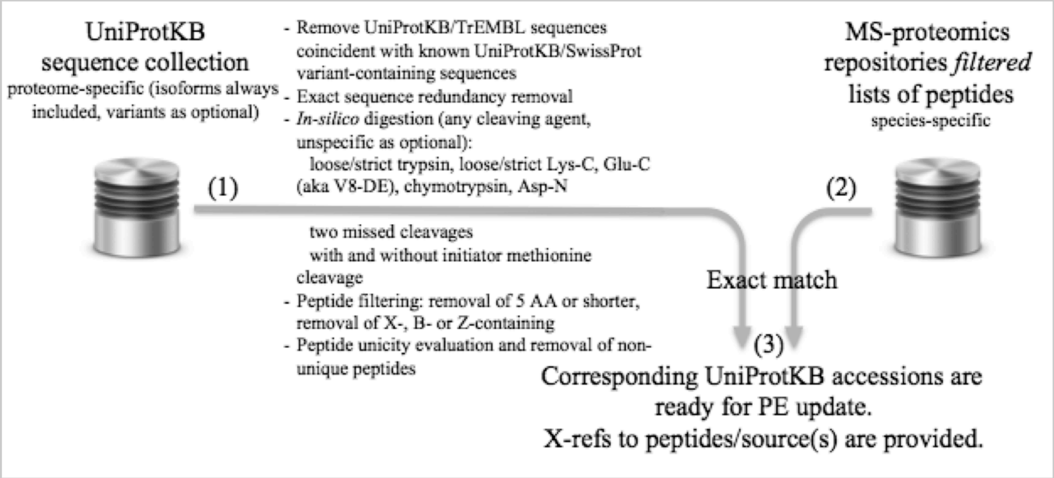
UniProtKB
sequence collection
proteome-specific (isoforms always included, variants as optional)
- Remove UniProtKB/TrEMBL sequences coincident with known UniProtKB/SwissProt variant-containing sequences
- Exact sequence redundancy removal
- *In-silico* digestion (any cleaving agent, unspecific as optional):
loose/strict trypsin, loose/strict Lys-C, Glu-C (aka V8-DE), chymotrypsin, Asp-N
two missed cleavages
with and without initiator methionine cleavage
- Peptide filtering: removal of 5 AA or shorter, removal of X-, B- or Z-containing
- Peptide unicity evaluation and removal of non-unique peptides
MS-proteomics
repositories *filtered*
lists of peptides
species-specific
(1)
(2)
Exact match
(3)
Corresponding UniProtKB accessions are ready for PE update.
X-refs to peptides/source(s) are provided.

Figure 1 : Proteomics analysis pipeline

Sequence-centric uniqueness means that if an *in-silico* digested peptide is present in more than one sequence it is filtered out from the list in step (1) above.

Gene-centric uniqueness means that if an in-silico digested peptide is present in more than one sequence but all of these sequences belong to the same gene, it is still considered unique and kept in the list from step (1) above. But if the peptide is present in several sequences that belong to different genes then this is filtered out from the list in step (1) above.

3.2 Results for human

Human proteome analysis on UniProtKB 2014_09 release against PeptideAtlas human MS-repository (August 2013 build) shows that out of a total of 84,895 proteins, 8.47% of them (7,187) can be mapped to at least one unique repository peptide and 874 of the 7,187 ones should be promoted to PE=1 (cf. Table 1) since they currently have other PE values. This low percentage is due to the sequence-centric peptide unicity used here, which is the most conservative one.

The analysis with the gene-centric peptide unicity shows that out of a total of 20,328 gene groups, 60.54% of them (12,306) can be mapped to at least one unique repository peptide and 1,412 of the 12,306 ones should be promoted to PE=1 since they currently have other PE values. If we count the sequences constituting the 12,306 gene groups we get the number of 61,940 sequences. They represent 72.96% of the 84,895 human proteome proteins, which are mapped to at least one unique PeptideAtlas human peptide.

The proteomics analysis pipeline will be soon integrated in UniProt release, it will provide new annotations and additionally, the mapping between UniProtKB accessions and peptides coming from this pipeline will be made available on the UniProt FTP site.

	NR seq.	Uni. Pep.	AUP	Having Unique	%Uni
Seq.	84,895	3,176,729	37	73,453	86.52%
Gene	20,328	7,711,001	379	20,252	99.63%

	Rep. pep.	Map P	%Map P	Map gp.	%Map gp.	PE1	Not PE1
Seq.	338,013	65,629	19.42%	7,187	8.47%	6,313	874
Gene	338,013	238,683	70.61%	12,306	60.54%	10,894	1,412

Rows: Seq. are results sequence-centric analisys; Gene are results from gene-centric analisys

Columns: NR seq.: number of non-redundant records in the corresponding dataset; Uni. Pep.: number off unique peptides after *in silico* digestion (five cleaving agents); AUP: the average of unique peptide per sequence/gene; Having Unique: number of sequences/genes having at least one unique peptide; %Uni: the percentage having unique in NR seq.; Rep. pep.: number of unique peptides from MS-repository (PeptideAtlas in this analisys); Map P: number of Uni. Pep. matching Rep. Pep.; %Map P: percentage of Map P in Rep. Pep.; Map gp.: number of sequences/genes having at least one Map P.; %Map gp: percentage of Map pg. in NR seq.; PE1: number of sequences/genes already PE1; Not PE1: number of sequences/genes that could be promoted to PE1.

Table 1 : Human proteome analysis for UniProt 2014_09 release

References

Alpi, E., Griss, J., Sousa da Silva, A. W., Bely, B., Antunes, R., Zellner, H., et al. (In Press). Analysis of the tryptic search space in UniProt databases. *Proteomics* .

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucl. Acids Res. , 32* (1), D115-D119.

Dessimoz, C., Gabaldón, T., Roos, D. S., Sonnhammer, E. L., Herrero, J., & Consortium, T. Q. (2012). Toward community standards in the quest for orthologs. *Bioinformatics , 28* (6), 900-904.

Griss, J., Martín, M. J., O'Donovan, C., Apweiler, R., Hermjakob, H., & Vizcaíno, J. A. (2011). Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. *Proteomics , 11* (22), 4434-38.

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., & Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *Proteomics , 4* (7), 1985-1988.

The UniProt Consortium. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucl. Acids Res. , 40* (1), D71-D75.

The UniProt Consortium. (2009). The Universal Protein Resource (UniProt). *Nucl. Acids Res. , 37* (1), D169-D174.