



LSHG-CT-2006-036814

ProDaC

Proteomics Data Collection

Instrument: Coordination Action

Thematic Priority: Life science, genomics and biotechnology for health

W6.D2.: Monthly updated PRIDE export files for confirmation of hypothetical proteins, definition of splice variants, tissue specificity, and protein modifications, available publicly as part of the PRIDE distribution.

Due date of deliverable: 31.03.2009

Actual submission date: 31.03.2009

Start date of project: 1st of October 2006

Duration: 30 months

Organisation name of lead contractor for this deliverable

European Bioinformatics Institute

Draft]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Background

The re-use of publicly available proteomics data to extend the annotation and coverage of sequence databases is the primary example of feedback loops that make the accumulated data in repositories work for the community at large (see Figure 1). Additional background on the overall concept of data exploitation can be found in the Background section of the report for W6.D1.

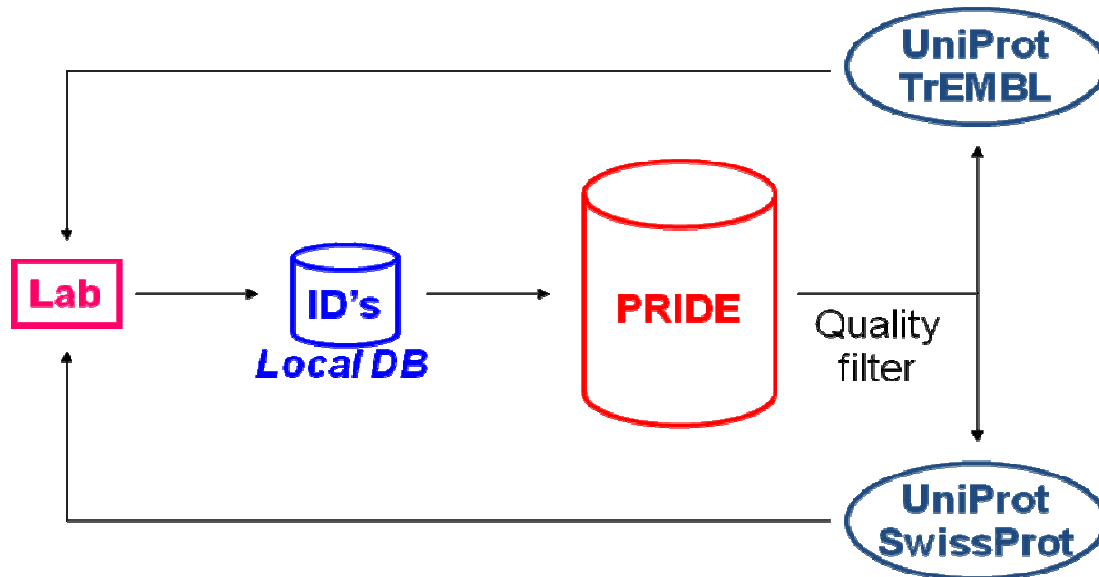


Figure 1: Schematic diagram of a feedback loop involving the reuse of (quality filtered) proteomics data in the extension of annotation and coverage in sequence databases (here the Swiss-Prot and Trembl parts of UniProt). This feedback cycle ultimately presents the proteomics community with more valuable resources that are based on their collective results as captured in public data repositories.

Description

The inclusion of proteomics data from public repositories into sequence database production pipelines hinges on two main issues: the availability of the data in a format that can be easily accommodated in the respective production pipeline, and pre-filtering of these data to ensure that only reliable results are used in these pipelines (see also W6.D1). This particular workpackage is concerned with the former issue, and is aimed at providing the actual export files that can be incorporated in the production pipeline of the sequence databases.

PRIDE currently exports two different such files at regular intervals, one aimed at the inclusion of novel proteins in UniProtKB/Trembl (Trembl), and the other to provide cross-references from UniProtKB/Swiss-Prot (Swiss-Prot) and Trembl to the evidence stored in PRIDE.

The first of these, provided to enhance the Trembl database, is based on a multistage workflow. First of all, every protein identification in PRIDE is regularly mapped across databases using the PICR tool [1]. A specialized analysis pipeline tool then looks up all protein identifications in PRIDE that have no PICR-derived mapping in Trembl or Swiss-Prot, and collects various pieces of information about these potential proteins of interest. This additional data consists of the corresponding UniParc accession number, the NEWT taxonomy ID for that accession number, the total number of peptides contributing to the identification of this protein (across all experiments in PRIDE!), the number of unique peptides for this protein, and finally, the full list of sequences for these unique peptides. The actual comma-separated output file is shown as rendered by Microsoft Excel in Figure 2.

Submitted accession	UniParc accession	taxonomy	Total number of peptides in	Number of unique peptides i	Unique peptide sequences
ENSP00000261364	UPI00015E0706	9606	1	1	1 CGYCGRAFAGATLNNHIR
ENSP00000323659	UPI0000161FAE	9606	1	1	1 DMMPSTRFDDLMANIPLPEYTR
IPI0003052.2	UPI00001FCB69	9606	1	1	1 KTTDPCQLQR
IPI00220697.4	UPI00001D3EFA	9606	1	1	1 VASTLTEEGGGGGGGGSVAPK
IPI00026958.1	UPI0000125608	9606	13	13	13 AGLLPSGPRPGYAAIQALLSSR,ALEIPGEELPGVCSAR,FGVAPDHPE
ENSP00000329896	UPI000022DC3D	9606	1	1	1 AARGSVPLQPPLPPAALGAYSGGAGPIR
IPI00321375	UPI00005ABB84	10090	1	1	1 FNDTEVLQR
ENSP00000323755	UPI00015E06DB	9606	1	1	1 EGAQAATORVER
IPI00153158	UPI00001BD2E0	10090	1	1	1 VPFSPGPAPPPHMGELDQER
IPI00783040.1	UPI000006F374	9606	1	1	1 SSIVWRFVEDSFDPNINPTIGASFMTK
IPI00746299.1	UPI0000061DF6	9606	7	1	1 HNLACSEERMAYLSYERAK
XP_421854	UPI0000448734	9031	2	2	2 SPLLGGSPQPVPVPTHK,SPLLGGSPQPVPVPTHKDK
IPI00175005	UPI0000160AD8	9606	1	1	1 ECLPLIFIR
IPI00032311	UPI000012E251	9606	3	3	3 GLQYAAQEGLLALQSELLR,LAEGFPLPLLK,VQLYDLGLQIHK
22094121	UPI000007104B	9606	1	1	1 STLIPILHQKAKR
IPI00135244	UPI00001E3489	10090	2	2	2 AQSPSPK,KNATVKTNK
IPI00007127	UPI000006CFDE	9606	1	1	1 LSSEETNKQQR

Figure 2: PRIDE output file for the inclusion of novel proteins in Trembl, as rendered by Microsoft Excel.

The second export file is much simpler in format, and allows Swiss-Prot and Trembl to link out to mass spectrometry based protein evidence stored in PRIDE (see Figure 3). This file is also created by a custom workflow, which relies on the PRIDE PICR mappings to obtain a list of UniProt proteins for which evidence is available in PRIDE. The primary UniProt accession number for each of these proteins is written to an export file (which is a simple text file, with one primary UniProt accession number per line). Additionally, PRIDE presents a specific URL for direct linking to all data it holds for a specific protein, based on the accession number for that protein (http://www.ebi.ac.uk/pride/searchSummary.do?queryTypeSelected=identification%20accession%20number&identificationAccessionNumber=##accession_number##).

All the software code for these workflows is also included in the freely available PRIDE source code.

Cross-references	
Sequence databases	
EMBL	X05491 Genomic DNA. Translation: CAA29043.1 . AE008777 Genomic DNA. Translation: AAL20660.1 .
PIR	QREBOF . D29333.
RefSeq	NP_460701.1 .
3D structure databases	
HSSP	HSSP built from PDB template 1B0U based on UniProtKB P02915 .
ModBase	Search...
Proteomic databases	
PRIDE	P08007 .
Genome annotation databases	
GeneID	1253261 .
GenomeReviews	Gene locus STM1742 in contig AE006468_GR .
KEGG	stm:STM1742 .
NMPDR	fig 99287.1.peg.1687 .

Figure 3: PRIDE cross-reference from UniProtKB, here from the Swiss-Prot protein **P08007** (OPPF_SALTY), Oligopeptide transport ATP-binding protein oppf.

Future work

With the implementation of the quality criteria defined in W6.D1, the exports from PRIDE will grow in number and complexity, as ever more detailed annotations can be made available

based on proteomics evidence. The work performed in the ProDaC project has however already proven crucial for the future integration of this more detailed data in the production pipeline of the sequence databases, since the groundbreaking work of integrating proteomics data and sequence databases has already been accomplished.

References

- [1] Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R & Hermjakob H. The Protein Identifier Cross-Reference (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* (2007) **8**: p. 401.