



LSHG-CT-2006-036814

ProDaC

Proteomics Data Collection

Instrument: Coordination Action

Thematic Priority: Life science, genomics and biotechnology for health

W6.D1.: Documentation of export criteria for confirmation of hypothetical proteins, definition of splice variants, tissue specificity, and protein modifications, to Ensembl, UniProt/TrEMBL, and UniProt/SwissProt

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

Although the public availability of proteomics data in standardized formats is a worthwhile goal in itself, once data are accumulated and readily accessible they can be put to other uses as well. Apart from meta-analyses to study the influences of experimental context [1] and the ability of data to support specific scientific queries [2], an obvious way to reuse these data is to provide feedback loops. Indeed, a long standing grievance in the proteomics community has been the time-lability of proteins in sequence databases. This unfortunate effect is discussed in detail in a comparative paper [3] where the HUPO Plasma Proteome Project (PPP) [4] and a platelet dataset [5] were moved forward in time to enable comparison with the Human Brain Proteome Project (BPP) [6] results. Briefly, this study showed clearly that changes in the open reading frame prediction algorithms used to create Ensembl and RefSeq XP led to a sudden disappearance of proteins from the integrative IPI sequence database [7] (see Figure 1), which made automated retrieval of many proteins (appr. 10% of the total protein count in the HUPO PPP) impossible. In contrast, the platelet dataset, which was obtained after this downsizing event, did not show this problem to the same extent – here only 3% of the proteins were lost when automatically converting them to the HUPO BPP version.

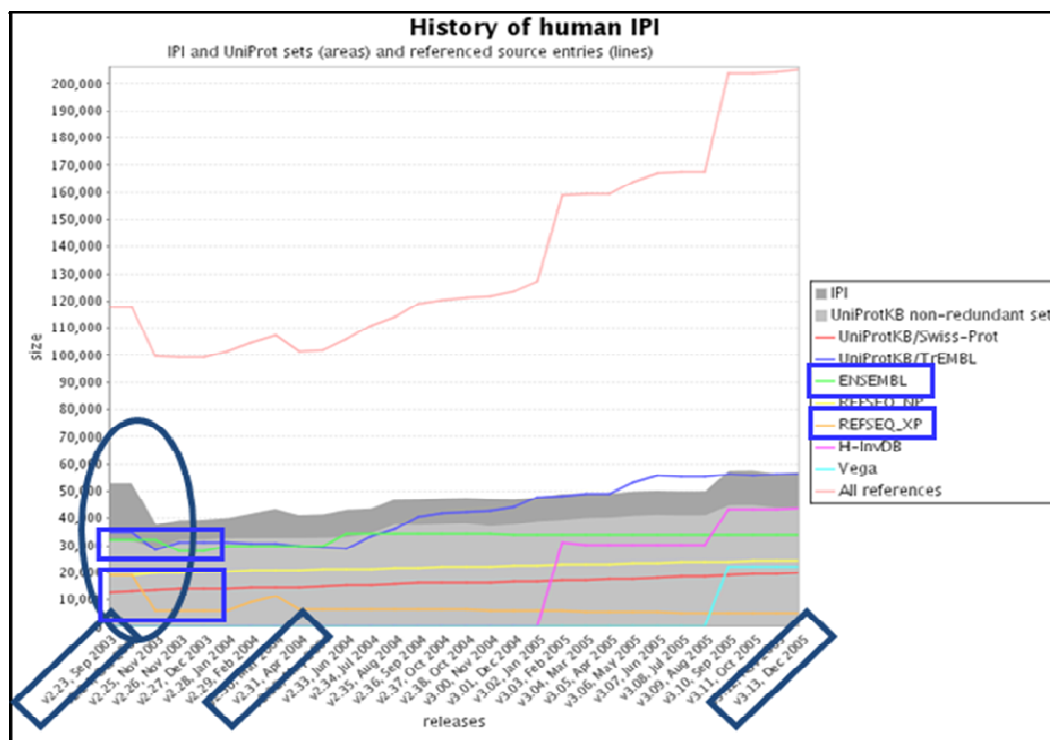


Figure 1: The IPI composition history at the time of conversion of the HUPO PPP and the platelet dataset into the IPI version used for the HUPO BPP project. The plot shows the subsidiary databases on which IPI is built, and the total number of IPI entries. The dark blue boxes over the X-axis indicate the dates of origin for the HUPO PPP, platelets, and HUPO BPP, respectively (from left to right). The dark blue oval highlights the downsizing event in IPI (dark grey area), with the bright blue boxes identifying the origins of this event (Ensembl and RefSeq XP), both on the actual plot, and in the legend.

By utilizing the data stored in PRIDE as confirmatory evidence for the presence of hypothetical proteins, this protein-level disappearing act can be halted, and valuable genome annotation can thus be obtained from the publicly available proteomics data. This overall strategy can also be applied to provide data for splice variant elucidation, protein tissue specific expression, and protein post-translational modifications.

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 (see also W6.D2), and pre-filtering of these data to ensure that only reliable results are used in these pipelines. This particular workpackage is concerned with this latter issue, and is aimed at providing the quality control criteria that will be applied to the data prior to its inclusion in the different database production pipelines.

The first thing to note here is that this is in fact a three-parameter problem: (i) the experimental context of the data (2D-gel based data might not be evaluated in the same way as gel-free data), (ii) the purpose of the data (confirmation of a post-translational modification is likely to require different criteria than the confirmation of a novel open reading frame), and (iii) the database that will consume the information (UniProtKB/Swiss-Prot will require more stringent criteria than UniProtKB/Trembl). It is therefore necessary to be able to evaluate a single dataset using various criteria depending on its purpose and destination, and to evaluate different datasets using varying criteria even though their purpose and destination will be the same. Obviously, data evaluation should be automated as much as possible (PRIDE already contains more than 10 million identified peptides, and is growing rapidly). In order to accommodate all of the above requirements, a platform-independent, readily configurable semi-automatic framework has been developed in collaboration with Partner 4 – VIB. This application, called Peptizer [8], is a highly configurable and flexible rule-based expert system that is mainly driven by hot-pluggable ‘agents’. A publicly accessible Google Group has also been set up to freely exchange these agents and agent configurations (a predefined set of agents that work in unison to produce a *profile* of an identification), thus implicitly providing a public dissemination mechanism for the implementation of the quality filters.

Although the quality criteria reported here provide the core framework of the evaluation procedure at the time of writing, it is likely that these criteria will evolve. First of all, the current criteria err on the side of caution and are extremely stringent, and it is likely that they will be relaxed in the future for certain purposes or destinations. Second, the field of mass spectrometry based proteomics is in constant evolution, and it is important that our quality control criteria can adapt to changing approaches in the field. Third, refinements specific to certain datasets may be carried out to accommodate highly specialized protocols that fall outside the scope of the ‘default’ criteria.

Quality control criteria

1. For confirmation of hypothetical proteins and definition of splice variants

The quality control takes place on three levels. First of these is the verification of the reported peptide sequence with regards to the evidence in the fragmentation spectrum. In this approach, a robust and generic mapping algorithm analyzes the spectrum for evidence of the amino acid sequence in the form of fragment ions and immonium ions after applying an adaptive, spectrum-specific noise filter. A Peptizer agent panel is then used to verify the quality of the peptide sequence assignment, and the report generated from these agents is captured and stored. Only peptide identifications that pass the Peptizer panel filter will be allowed to proceed to level two.

The second level involves the alteration of the reported peptide sequence to reflect the uncertainty that it may contain. Briefly, only those amino acid residues covered by flanking (consecutive) fragment ions are considered reliable. Whenever a gap of two or more amino acids is found in the fragment ions retrieved from the spectrum, the corresponding sequence segment is replaced by the mass delta between the closest flanking fragment ions. This sequence is called a gapped sequence.

At the third level, the gapped sequence is mapped to the proteome. If there is only one possible place of origin for this sequence, the gapped sequence is accepted as evidence for the corresponding protein. In the case of splice variants, the gapped peptide obviously has to be unique to an individual splice variant.

2. For tissue-specific protein expression

Tissue specificity is established by searching the PRIDE database with those proteins that pass the criteria outlined in (1) and analyzing the annotated tissue of origin for all the samples that are reported to contain this protein. It is important to note that PRIDE uses the BRENDA Tissue Ontology (BTO) for the purpose of annotating the tissue origin of a sample. Also note that the result of this analysis will be a list of all tissues in PRIDE in which this protein was found, as well as a list of tissues in which it was not found. Therefore, the tissue specificity might not be absolute (for instance, if 'smooth muscle' is a tissue for which there is no data available in PRIDE, it is impossible to definitively claim that a protein is not expressed in this tissue). This is however similar to the current tissue specificity annotations in UniProtKB/Swiss-Prot, since here too only the available data can be taken into account.

3. For post-translational protein modifications

Proteins that pass the filter outlined in (1) can be analyzed for the occurrence of post-translational modifications, but only peptides that pass the level 1 filter outlined in (1) will be considered for this protein. The evidence criterion then is that the modification of interest has to be found on a residue identified by two flanking fragment ions (i.e., not in a gapped region), or has to provide a clearly identifiable neutral loss signature along with a unique specificity for a unique residue in that sequence.

Future work

The PRIDE group at EBI will commence work on the implementation of the abovementioned quality criteria, ensuring in the process that both the software used (Peptizer) as well as the actual agent panels are publicly available. Furthermore, the reports obtained from the agent panels for each identification will also be made available to users of the PRIDE database, to ensure complete transparency of the entire pipeline. This implementation of the quality control criteria falls outside the scope of the ProDaC project, but will build on the groundwork performed within ProDaC with regards to the definition of the criteria and the creation of the quality-control software.

References

- [1] Klie S, Martens L, Vizcaíno JA, Côté R, Jones P, Apweiler R, Hinneburg A & Hermjakob H. Analyzing large-scale proteomics projects with latent semantic indexing. *J. Proteome Res.* (2008) **7**: pp. 182-191.
- [2] Mueller M, Vizcaíno JA, Jones P, Côté R, Thorneycroft D, Apweiler R, Hermjakob H & Martens L. Analysis of the experimental detection of central nervous system related genes in human brain and cerebrospinal fluid datasets *Proteomics* (2008) **8**: pp. 1138-1148.
- [3] Martens L, Muller M, Stephan C, Hamacher M, Reidegeld KA, Meyer HE, Bluggel M, Vandekerckhove J, Gevaert K & Apweiler R. A comparison of the HUPO Brain Proteome Project pilot with other proteomics studies *Proteomics* (2006) **6**: pp. 5076-5086.
- [4] Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H, Paik YK, Yoo JS, Ping P,

- Pounds J, Adkins J, Qian X, Wang R, Wasinger V, Wu CY, Zhao X, Zeng R, Archakov A, Tsugita A, Beer I, Pandey A, Pisano M, Andrews P, Tammen H, Speicher DW & Hanash SM. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database *Proteomics* (2005) **5**: pp. 3226-3245.
- [5] Martens L, Van Damme P, Van Damme J, Staes A, Timmerman E, Ghesquière B, Thomas GR, Vandekerckhove J & Gevaert K. The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics* (2005) **5**: pp. 3193-3204.
- [6] Hamacher M, Apweiler R, Arnold G, Becker A, Bluggel M, Carrette O, Colvis C, Dunn MJ, Frohlich T, Fountoulakis M, van Hall A, Herberg F, Ji J, Kretzschmar H, Lewczuk P, Lubec G, Marcus K, Martens L, Palacios Bustamante N, Park YM, Pennington SR, Robben J, Stuhler K, Reidegeld KA, Riederer P, Rossier J, Sanchez JC, Schrader M, Stephan C, Tagle D, Thiele H, Wang J, Wiltfang J, Yoo JS, Zhang C, Klose J & Meyer HE. HUPO Brain Proteome Project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy *Proteomics* (2006) **6**: pp. 4890-4898.
- [7] Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E & Apweiler R. The International Protein Index: an integrated database for proteomics experiments *Proteomics* (2004) **4**: pp. 1985-1988.
- [8] Helsens K, Timmerman E, Vandekerckhove J, Gevaert K & Martens L. Peptizer, a tool for assessing false positive Peptide identifications and manually validating selected results *Mol. Cell Proteomics* (2008) **7**: pp. 2364-2372.