Proteome Science

Research

Open Access

An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ™

Jennifer A Siepen†[1], Neil Swainston†[2], Andrew R Jones[1,3], Sarah R Hart[4], Henning Hermjakob[5], Philip Jones[5] and Simon J Hubbard*[1]

Address: [1]Faculty of Life Sciences, University of Manchester, M13 9PT, UK, [2]Manchester Interdisciplinary Biocentre, University of Manchester, UK, [3]School of Computer Science, Faculty of Engineering and Physical Sciences, University of Manchester, UK, [4]MBCMS, School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, UK and [5]EMBL Outstation EBI, Wellcome Trust Genome Campus, Hinxton, Cambs, UK

Email: Jennifer A Siepen - jennifer.siepen@manchester.ac.uk; Neil Swainston - neil.swainston@manchester.ac.uk; Andrew R Jones - ajones@cs.man.ac.uk; Sarah R Hart - sarah.hart@manchester.ac.uk; Henning Hermjakob - hhe@ebi.ac.uk; Philip Jones - pjones@ebi.ac.uk; Simon J Hubbard* - simon.hubbard@manchester.ac.uk

* Corresponding author †Equal contributors

Published: 1 February 2007

Proteome Science 2007, **5**:4 doi:10.1186/1477-5956-5-4

This article is available from: http://www.proteomesci.com/content/5/1/4

Received: 20 December 2006
Accepted: 1 February 2007

Abstract

Background: Proteomics continues to play a critical role in post-genomic science as continued advances in mass spectrometry and analytical chemistry support the separation and identification of increasing numbers of peptides and proteins from their characteristic mass spectra. In order to facilitate the sharing of this data, various standard formats have been, and continue to be, developed. Still not fully mature however, these are not yet able to cope with the increasing number of quantitative proteomic technologies that are being developed.

Results: We propose an extension to the PRIDE and mzData XML schema to accommodate the concept of multiple samples per experiment, and in addition, capture the intensities of the iTRAQTM reporter ions in the entry. A simple Java-client has been developed to capture and convert the raw data from common spectral file formats, which also uses a third-party open source tool for the generation of iTRAQTM reported intensities from Mascot output, into a valid PRIDE XML entry.

Conclusion: We describe an extension to the PRIDE and mzData schemas to enable the capture of quantitative data. Currently this is limited to iTRAQTM data but is readily extensible for other quantitative proteomic technologies. Furthermore, a software tool has been developed which enables conversion from various mass spectrum file formats and corresponding Mascot peptide identifications to PRIDE formatted XML. The tool represents a simple approach to preparing quantitative and qualitative data for submission to repositories such as PRIDE, which is necessary to facilitate data deposition and sharing in public domain database. The software is freely available from http://www.mcisb.org/software/PrideWizard.

Background

Proteomics continues to play a critical role in post-genomic science as continued advances in mass spectrometry and analytical chemistry support the separation and identification of increasing numbers of peptides and proteins from their characteristic mass spectra. A desirable trait for such a functional genomics technique is the ability to produce data on a genome-wide basis and, importantly, to be able to do this in a quantitative manner. In proteomics this means being able to quantify the protein changes in different conditions, be they temporal, pathogenic or environmental. Proteomics is beginning to address both these issues; wider genome coverage and quantitation of the proteins present. The latter has been driven by the continued development of techniques for the relative and absolute quantification of protein levels [1-6]. Equally, superior instrumentation and analytical approaches have improved the coverage of genomes, so that genome-wide quantitative proteomics is becoming a reality. This is epitomised by a recent quantitative study acquiring data for the majority of the yeast proteome [7], where the majority of proteins had peptide identifications with available quantitative data obtained using stable isotope labelling in cell culture (SILAC).

Clearly, these types of experiments will become more widespread and detailed. This presents several challenges to the proteomics community and the bioinformatics teams in particular, since it is desirable that this data is captured and stored in appropriate databases in consistent formats, to support data sharing and comparison. Although there are a growing number of data standards [8-12] and databases [13-18] for the storage of proteomic data, at present there is no formal model for quantitative proteomic data that has been fully developed. The Proteome Standards Initiative (PSI) and leading proteomics groups have helped drive the development of several standards for the mass spectral data itself, namely mzXML [8] and mzData [19], and these two are expected to soon merge. These support a comprehensive data model for the storage of proteomic-related mass spectral data, ranging from basic details about the sample, through instrument details and data processing steps, to the actual spectral lists of mass-to-charge values and intensities. This provides a relatively simple yet extensible format for any type of peptide or protein spectra, allowing users to support parent/precursor ion concepts and sophisticated MS^n experiments. Both formats utilise base64 encoding to represent the floating point mass-to-charge (m/z) and ion intensity pairs which form the core of the spectral information. Although this supports the capture of any protein, peptide or fragment ion MS spectra, quantitative data is not explicitly represented in the model. Furthermore, it is not clear how to link the spectra to rich descriptions of the experimental sample, or mixture of samples, within these formats. Indeed, the work to bring this together into a considered whole for proteomics and indeed, in a wider functional genomics context, is well underway, with standards in development for identifications (analysisXML), gels (GelML etc), general sample processing (spML) and functional genomics experiments more generally (FuGE, [11]). Even though the standards development community has not finished this process, database developers in proteomics have already provided solutions for many of these issues in the growing range of proteomics databases now available. These include PeptideAtlas [14], Open Proteomics Database (OPD) [13], Global Proteome Machine (GPM) [15], Pedro [17], PepSeeker [16], and the PRoteomics IDEntifications database (PRIDE) [18,20] as well as others. PeptideAtlas, GPM and PRIDE in particular already contain extensive collections of many millions of peptide identifications. PRIDE, for example, has integrated the mzData data standard into its own PRIDE XML format, which allows users to provide a rich description of their experiment and uses a range of well-supported ontologies to populate the model for a range of meta-data including taxonomy, instrument type, etc. The other databases are also able to capture a similar range of data.

At Manchester, local proteomics groups are active in quantitative proteomics, developing both novel methodology [5,6] and using existing technology to explore quantitative protein levels. In particular, the iTRAQTM technology [4] is widely used by many groups worldwide, since it offers several advantages, including the ability to multiplex several samples in one single experiment, quantifying several samples in one experiment via a series of reporter ions which are fragmented from an isobaric tag attached to free peptide amines. Thus, researchers can quantify the relative levels of several samples, averaging over data from several peptides, using a labelling technique applicable to all peptides, and not relying on cell culture or similar using stable isotope labelling. This ingenious technique presents informatics with a novel modelling challenge, since such a concept cannot be directly modelled in the existing mzData schema, which considers the sample itself to be a single entity to which all spectra in the experiment are related.

To address this problem, we have conducted a case study to further develop the PRIDE and mzData XML schema to accommodate the concept of multiple samples per experiment, and in addition capture the intensities of the iTRAQTM reporter ions in the entry. The model extensions are completely compatible with both the PRIDE and mzData schema, utilising controlled vocabulary terms which are added to the respective ontologies. Furthermore, we have developed a simple Java-client (the "Pride Wizard") to capture and convert the raw data from common spectral file formats, which also uses a third-party open source tool for the generation of iTRAQTM reported intensities from Mascot output. Together, this allows the user to capture large, high-throughput ITRAQTM-based studies, without extensive repetitive manual data entry of individual peptide identifications, and delivered in a valid PSI-consistent data format (PRIDE XML) for submission to the PRIDE repository. The underlying model and Java-client are readily extended to other quantitation techniques. Finally, the Java-client also allows users to directly capture non-quantitative large scale proteomics data, providing the opportunity to convert Mascot-based spectral searches into mzData with associated peptide identifications. We believe this tool will allow proteomics groups to rapidly capture their datasets for submission to a PSI-sanctioned repository and provides a step change in the ease of complex proteomic data available for analysis and sharing for the community in general.

Results

Data capture pipeline

The capture of the mass spectrometry data, associated protein and peptide identifications and quantitative values for multiple samples has been integrated into a single client application, shown in overview in Figure 1. In this example, the mass spectrometry data is represented in Mascot's .mgf format.

As input, the user provides the Pride Wizard with one or more processed mass spectrum files (in either .mgf, mzXML, .pkl or mzData format) and associated Mascot dat files containing protein and peptide identifications. In addition, a number of experimental meta data values are required.

The Pride Wizard can be run in qualitative and quantitative iTRAQTM mode. In the case of the latter, the user specifies a number of samples involved in the experiment, and assigns one or more iTRAQTM labels to each of these samples. Ontology terms can also be assigned to the samples, as the Pride Wizard acts as a client to the EBI Ontology Lookup Service [21] (see Figure 2). A correction factor file must be submitted, in which the isotopic purity of each of the iTRAQTM reagents used in the experiment are specified.

Finally, the user specifies the location of the PRIDE XML file that will be generated upon successful completion of the wizard.

The conversion of this data into valid PRIDE XML takes a number of steps (shown in Figure 1). mzData is required in the final PRIDE XML document, so if necessary, submitted mass spectrum files are converted to mzData using a module provided by ProteomeCommons.org [22]. The ProteomeCommons module is again used to perform conversion of mzData to .mgf files, which are required by the identifications parser module.

The identifications parser takes the form of a Perl script which parses an .mgf mass spectrum file and a Mascot 'dat' file to generate a PRIDE XML fragment containing protein and peptide identifications. In the case where the user has provided iTRAQTM labelled sample data the peptide identification results from the Mascot 'dat' file are merged with iTRAQTM intensities and ratios from i-Tracker [23].

Only the top three ranking identified peptides are reported in the PRIDE XML and the identified peptides are grouped according to the protein accession for the first matching protein for each of the identifications. Where a post-translational modification is assigned by Mascot (fixed or variable) then the name of the modification is matched to the UNIMOD database [24]. If the name of the modification cannot be matched to UNIMOD then the name of the peptide modification is represented as a userParam (see Methods for a description of the schema constructs used).

The i-Tracker software returns the relative ratios of each of the iTRAQTM reporter ions from an .mgf formatted file, a set of correction factors and a user-defined threshold. If the maximum ion peak intensity for any reporter ion peak area is equal to or less than the user-entered threshold a flag of "UT" for "Under Threshold" is reported in the PRIDE XML file.

The iTRAQTM intensities are reported using the iTRAQTM reagent 114 label (see Methods). The actual ratios for each of the iTRAQTM reporter ions, calculated by i-Tracker, are represented as userParams, where, for every peptide identification, we represent the iTRAQTM reporter ion ratios as:

```
<userParam value="1" name="114_114"/>

<userParam value="1.597" name="114_115"/>

...

<userParam value="1.233" name="117_116"/>

<userParam value="1" name="117_117"/>
```

The final step involves merging together each of the mzData files with the generated PRIDE XML fragments to generate a single PRIDE XML document representing the entire experiment. This document is then saved to the user-specified location.

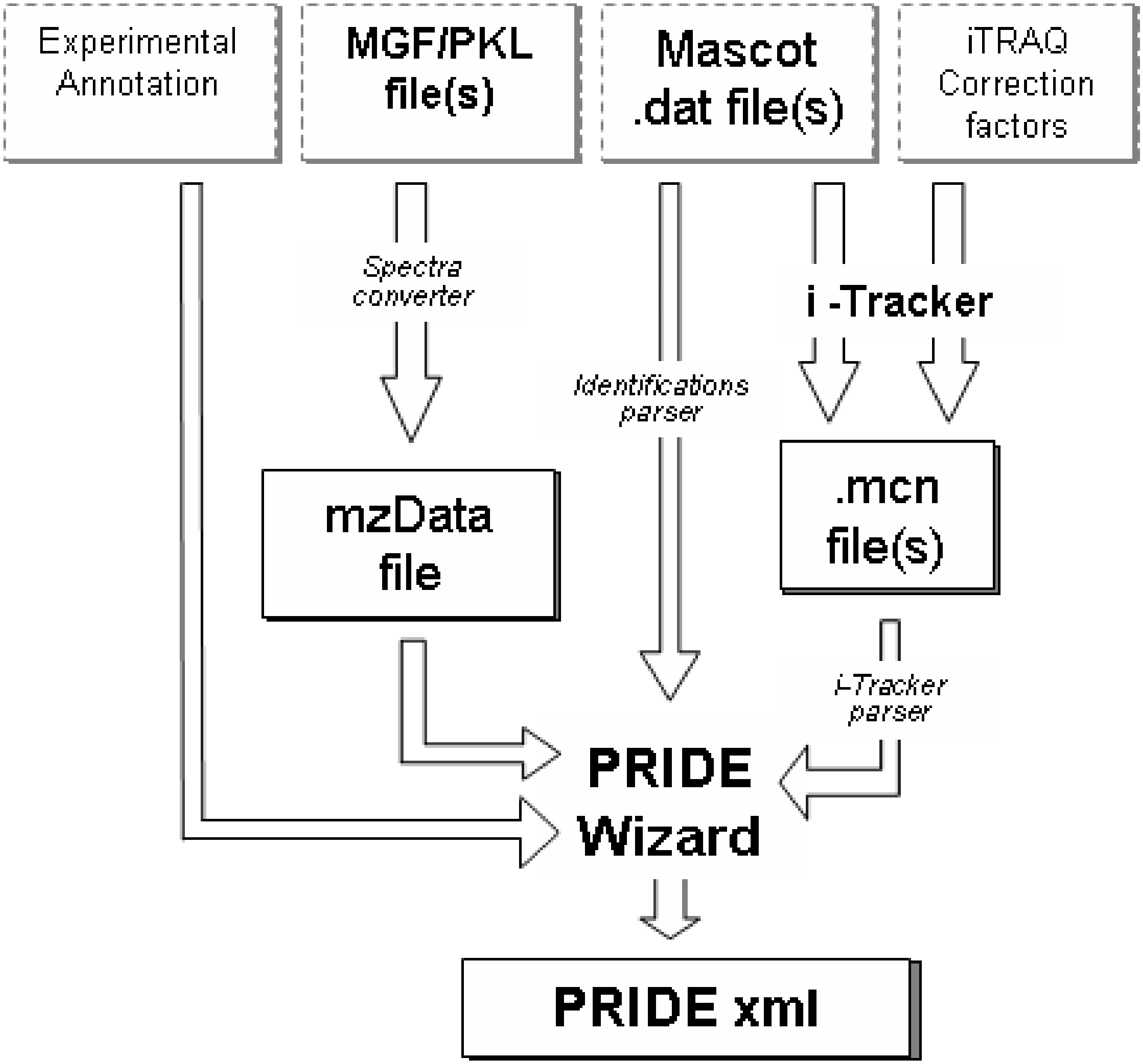


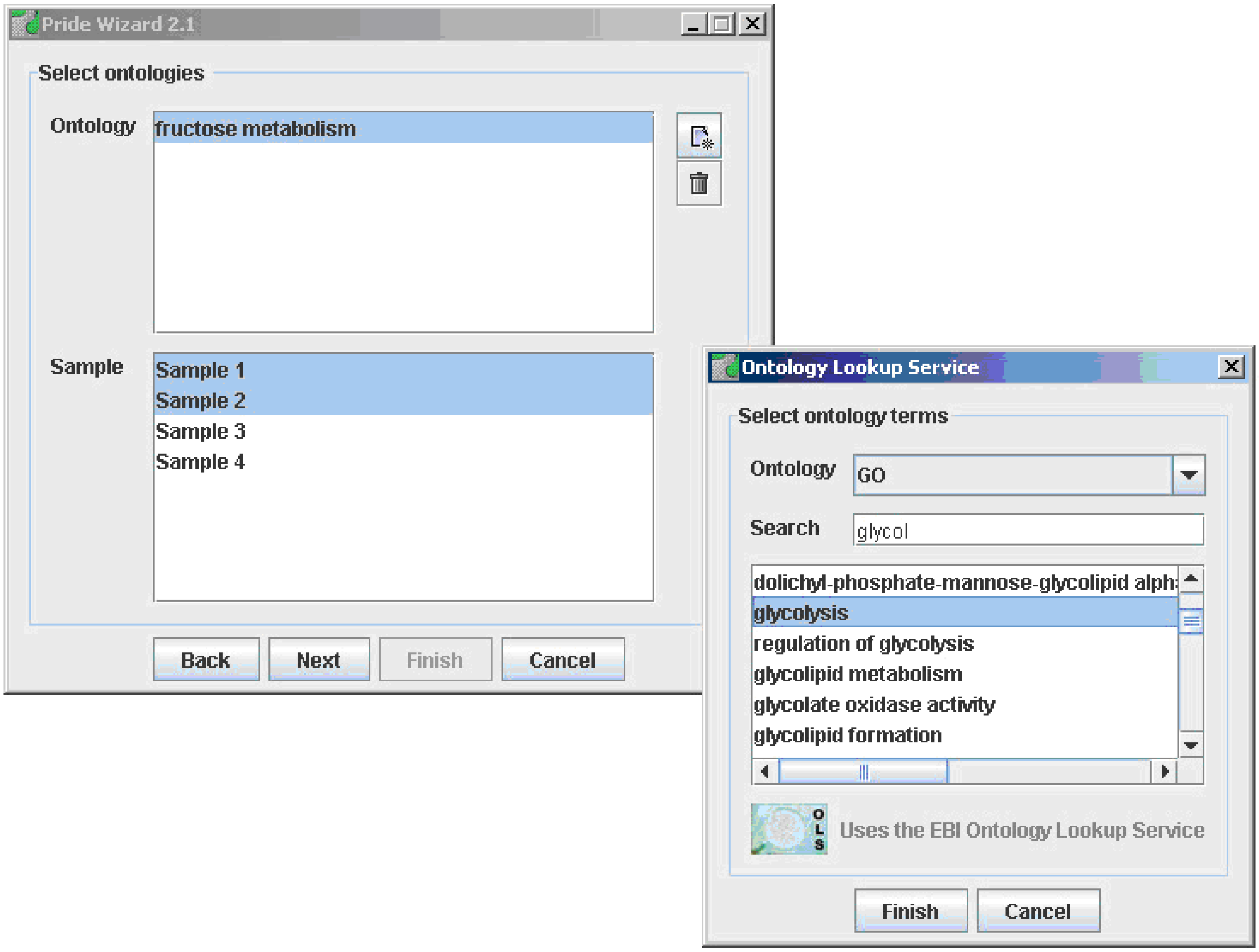
Figure 1
The data processing pipeline for the PRIDE wizard. Grey boxes represent the files/text that are required as input and the black boxes those files created by the PRIDE wizard.

Test data

The software tool was tested on selection of exemplar quantitative data from a number of different collaborating laboratories and successfully created valid PRIDE XML files. The samples included iTRAQTM-based analyses from multiple species, using several instrument types. Full details of the experiments are contained in the methods. The performance of the software was estimated; the wizard takes approximately 4.3 minutes to run on 2314 .mgf formatted mass spectra with 3581 corresponding peptide identifications on a single laptop.

Discussion

We have described a use of controlled vocabulary terms to represent quantitative proteomics data within the PRIDE data format and a software tool to capture and produce the correct file format. Several data standards are currently under development by the Proteomics Standards Initia-



Ontology: fructose metabolism

Sample: Sample 1, Sample 2, Sample 3, Sample 4

Ontology: GO

Search: glycol

dolichyl-phosphate-mannose-glycolipid alph
glycolysis
regulation of glycolysis
glycolipid metabolism
glycolate oxidase activity
glycolipid formation

Figure 2
Ontology term selection in the PRIDE Wizard.

tive which will be adopted by PRIDE, allowing a complete proteomics pipeline to be represented. This includes detailed descriptions of protein or peptide separations and labelling (in spML), the mass spectrometry data (mzData) and the protein identifications and quantitative values (analysisXML). However, it is unlikely that spML and analysisXML will be stable and implemented by PRIDE until late 2007 or early 2008. Therefore, the format extension proposed here represents a suitable interim solution for storing quantitative data, and we encourage other laboratories to adopt the conventions. This will allow quantitative data to be represented now in a "pseudo-standard" format and will enable other groups to download such data from PRIDE and perform re-analysis.

In addition to this functionality for iTRAQTM based data, we believe the tool is readily extensible for other quantitative proteomic technologies in a similar fashion, by extending the model and making minor adaptations to the associated Perl and Java code in the Pride Wizard. To this end, we have made the source code available [25]. As data capture needs for both SILAC [1] and QconCAT [5,6] methodologies are underway in our laboratories we expect to provide specific solutions for these approaches in early 2007.

Although the tool was designed ostensibly solely for quantitative data capture, it clearly is able to capture large volumes of identification data and deliver this automatically in PRIDE XML format. We anticipate this will be extremely useful to many groups with high-throughput data sets they wish to capture without tedious manual input. In order to capture the associated experiment, instrument and sample data that can be associated with a PRIDE entry we recommend the PedroDC data capture tool developed at Manchester [26]. Since the PRIDE XML delivered by our pipeline validates against the PRIDE schema, the data capture tool allows further flexibility to load the PRIDE XML and make suitable additions and edits. Alternatively, the PRIDE team have developed a spreadsheet-based approach linked directly to the Ontology Lookup Service at the EBI which provides an efficient means of entering the higher level data into a PRIDE entry. We anticipate that all of the above will be useful to different user groups, and that a suite of different approaches are probably necessary in proteomics, as any enhancement of data capture capabilities which facilitates data deposition and sharing in public domain repositories is to be welcomed.

Methods

Data capture overview

To generate iTRAQTM quantitative data requires several key component data types which must be integrated. An overview of these data types and the associated analysis tools are shown in Figure 1. A typical iTRAQTM experiment involves the analysis of several samples in a single MS run where peptides are identified in a standard fashion using a search tool such as Mascot [27]. The spectral data are typically delivered to Mascot using Matrix Science's Mascot Generic Format (.mgf), although the tool can cope with a variety of vendor specific formats, as well as mzData. The peptide identifications themselves are contained in Mascot's .dat output file. Finally, to generate the quantitative data for each peptide, users can employ ABI's ProQuant software, or if they prefer, third-party open source tools such as i-Tracker [23]. The latter uses a correction file supplied by ABI to adjust the reporter ion intensities for each identified peptide. The Pride Wizard we have developed integrates these data into a single PRIDE XML file. The model extensions are detailed in the following section.

Modelling quantitative data in PRIDE XML

The mzData schema lies at the heart of a PRIDE XML entry. PRIDE's model is deliberately "light touch" whilst data standards mature, and is readily extensible via inbuilt controlled vocabulary (CV) terms. However, mzData contains only a single sample description object which is also used by PRIDE to capture sample information.

Controlled vocabularies are frequently employed in data formats to provide a consistent extension mechanism allowing a format to capture unanticipated data types [12]. PRIDE files can be annotated with CV and user-defined terms to describe details of the experimental protocol employed, the sample analysed, the instrument used and protein or peptide identifications. We have made use of CV and user-defined terms in PRIDE to support multiplexed sample descriptions and the corresponding quantitative data for each sample (Figure 3)

A CV term in PRIDE has a name (the term itself), a unique accession from the source CV, a label to identify the CV source and optionally a value that can be completed by the user. An example is as follows, where the term *Homo sapiens* (from the NEWT taxonomy) [28] is used to describe the sample:

```
<sampleDescription>

<cvParam name="Homo sapiens" accession="NEWT:9606" cvLabel="NEWT"/>

</sampleDescription>
```

A further example, where a user-entered value (2000 for "Mass Resolution") has been included:

```
<analyzer>

<cvParam accession="PSI:1000011" name="Mass Resolution" value="2000" cvLabel="PSI"/>

</analyzer>
```

In the extension of PRIDE, we have utilised the userParam facility and the CV representation to capture the names of multiple samples within a single file. The userParam term supports the inclusion of a placeholder for the name of each sub-sample (SUBSAMPLE_1 is used in the example below). This is easily supplemented by the use of additional CV terms in the standard way to add taxonomic or further related information pertaining to the sub-sample. Finally, we have created a further list of CV terms named after the 4 standard iTRAQTM report ions to link the sub-samples to specific reagents called iTRAQTM reagent 114, iTRAQTM reagent 115 etc. The user completes the value attribute of cvParam with the name of each sample they wish to record in the file.

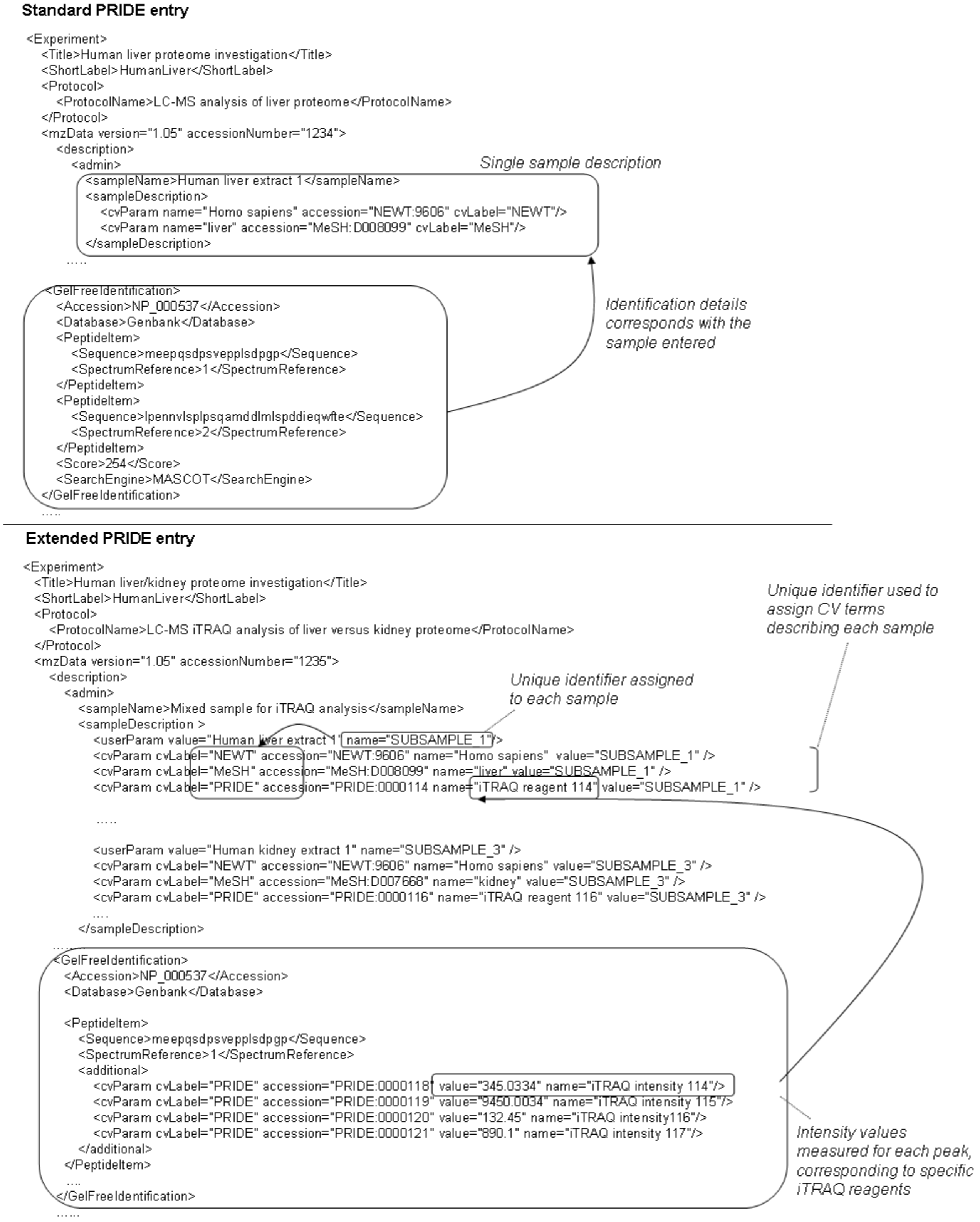
```
<userParam value="Human Liver Extract 1" name="SUBSAMPLE_1"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000114" value="SUBSAMPLE_1" name="iTRAQ reagent 114">
```

The term SUBSAMPLE_1 then serves as a unique identifier for that sample throughout the rest of the file. Where the user wants to add further CV terms to describe the sample, the value attribute is completed with SUBSAMPLE_1.

```
<sampleDescription>

<cvParam name="Homo sapiens" accession="NEWT:9606" cvLabel="NEWT" value="SUBSAMPLE_1"/>
```



Standard PRIDE entry

```
<Experiment>
  <Title>Human liver proteome investigation</Title>
  <ShortLabel>HumanLiver</ShortLabel>
  <Protocol>
    <ProtocolName>LC-MS analysis of liver proteome</ProtocolName>
  </Protocol>
  <mzData version="1.05" accessionNumber="1234">
    <description>
      <admin>
        <sampleName>Human liver extract 1</sampleName>
        <sampleDescription>
          <cvParam name="Homo sapiens" accession="NEWT:9606" cvLabel="NEWT"/>
          <cvParam name="liver" accession="MeSH:D008099" cvLabel="MeSH"/>
        </sampleDescription>
      .....

<GelFreeIdentification>
    <Accession>NP_000537</Accession>
    <Database>Genbank</Database>
    <PeptideItem>
      <Sequence>meepqsdpsvepplsdpgp</Sequence>
      <SpectrumReference>1</SpectrumReference>
    </PeptideItem>
    <PeptideItem>
      <Sequence>lpennvlsplpsqamddlmlspddieqwfte</Sequence>
      <SpectrumReference>2</SpectrumReference>
    </PeptideItem>
    <Score>254</Score>
    <SearchEngine>MASCOT</SearchEngine>
  </GelFreeIdentification>
  .....
```

Single sample description

Identification details corresponds with the sample entered

Extended PRIDE entry

```
<Experiment>
  <Title>Human liver/kidney proteome investigation</Title>
  <ShortLabel>HumanLiver</ShortLabel>
  <Protocol>
    <ProtocolName>LC-MS iTRAQ analysis of liver versus kidney proteome</ProtocolName>
  </Protocol>
  <mzData version="1.05" accessionNumber="1235">
    <description>
      <admin>
        <sampleName>Mixed sample for iTRAQ analysis</sampleName>
        <sampleDescription >
          <userParam value="Human liver extract 1" name="SUBSAMPLE_1"/>
          <cvParam cvLabel="NEWT" accession="NEWT:9606" name="Homo sapiens"  value="SUBSAMPLE_1" />
          <cvParam cvLabel="MeSH" accession="MeSH:D008099" name="liver" value="SUBSAMPLE_1" />
          <cvParam cvLabel="PRIDE" accession="PRIDE:0000114 name="iTRAQ reagent 114" value="SUBSAMPLE_1" />

          .....

          <userParam value="Human kidney extract 1" name="SUBSAMPLE_3" />
          <cvParam cvLabel="NEWT" accession="NEWT:9606" name="Homo sapiens" value="SUBSAMPLE_3" />
          <cvParam cvLabel="MeSH" accession="MeSH:D007668" name="kidney" value="SUBSAMPLE_3" />
          <cvParam cvLabel="PRIDE" accession="PRIDE:0000116" name="iTRAQ reagent 116" value="SUBSAMPLE_3" />
          ....
        </sampleDescription>
    .......
  <GelFreeIdentification>
    <Accession>NP_000537</Accession>
    <Database>Genbank</Database>

    <PeptideItem>
      <Sequence>meepqsdpsvepplsdpgp</Sequence>
      <SpectrumReference>1</SpectrumReference>
      <additional>
        <cvParam cvLabel="PRIDE" accession="PRIDE:0000118" value="345.0334" name="iTRAQ intensity 114"/>
        <cvParam cvLabel="PRIDE" accession="PRIDE:0000119" value="9450.0034" name="iTRAQ intensity 115"/>
        <cvParam cvLabel="PRIDE" accession="PRIDE:0000120" value="132.45" name="iTRAQ intensity116"/>
        <cvParam cvLabel="PRIDE" accession="PRIDE:0000121" value="890.1" name="iTRAQ intensity 117"/>
      </additional>
    </PeptideItem>
    ....
  </GelFreeIdentification>
  ......
```

Unique identifier used to assign CV terms describing each sample

Unique identifier assigned to each sample

Intensity values measured for each peak, corresponding to specific iTRAQ reagents

Figure 3
Extensions to the PRIDE XML schema.

```
</sampleDescription>
```

In order to enter the actual intensities of the reporter ions, we propose the following convention, adapting the iTRAQTM reagent 114 label further to iTRAQTM intensity 114 as shown below.

```
<cvParam cvLabel="PRIDE" accession="PRIDE:0000118"
value="0.048" name="iTRAQ intensity 114"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000119"
value="0.193" name="iTRAQ intensity 115"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000120"
value="0.204" name="iTRAQ intensity 116"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000121"
value="0.65" name="iTRAQ intensity 117"/>
```

The terms proposed here have been added to the PRIDE CV and have been assigned stable accession numbers. It is apparent from this example that other quantitative data with intensity or ratio values, calculated in a variety of ways, can be represented using similar CV terms.

Test data

The software tool was tested on a number of different data sets from different laboratories. Test set 1 was derived from *Trypanosoma brucei* flagellum samples which were prepared as described previously [29]. Samples were derivatised using iTRAQTM according to the manufacturer's instructions and derivatised peptides from four samples were prepared and analysed online with a QTOF I instrument (Waters, Manchester, upgraded to QTOF II specifications by MS Horizons, Manchester). Data acquisition was performed using MassLynx 3.4, acquiring 3 channels of tandem MS data. Following acquisition, data were processed using ProteinLynx to generate .pkl files.

Test set 2 was derived from soluble extracts from the gram negative plant pathogenic bacterium *Erwinia carotovora* (sp atroseptica SCRI1043) which were prepared as described previously [30]. Three biological replicate samples were labelled with iTRAQTM reagents 114–116 respectively, a fourth sample which consisted of a pool of the three replicates was labelled with the 117 iTRAQTM reagent. Labelling, multidimensional LC and MSMS were carried out as in [31]. The data submitted to the PRIDE wizard was essentially from the combination of running four fractions from strong cation exchange column on LCMSMS (QSTAR, Applied Biosystems).

Test set 3 was derived from primitive hematopoietic cells from mouse bone marrow as described previously [32]. Samples were derivatised using iTRAQTM according to the manufacturer's instructions and derivatised peptides from four samples were prepared and analysed online with a QSTAR XL (Applied Biosystems). Data acquisition was performed using an independent data acquisition protocol as described previously [32].

Availability and requirements

Project name: Pride Wizard

Project homepage: http://www.mcisb.org/software/PrideWizard

Operating system: Windows

Programming language: perl, Java 1.4.2 and above.

Licence: GNU GPL

Authors' contributions

SJH, ARJ, JAS, NS & SRH designed and developed the quantitative data model in the PRIDE XML schema, and jointly proposed the schema extensions with subsequent additional verification from PJ & HH. JAS & NS developed the software tool to capture the data, with testing from SRH. SJH conceived the study, lead the manuscript production with contributions from all authors, who have read and approved the final manuscript.

Declaration of competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

The authors are indebted to Prof Tony Whetton and Dr Richard Unwin at the University of Manchester, Dr Kathryn Lilley at Cambridge University for provision of data sets to evaluate the wizard, and Jayson Falkner of the University of Michigan for assistance with the ProteomeCommons module. This work has been supported by several BBSRC and EPSRC grants to the authors, ISPIDER (BBSB17204) to JAS and SJH, Pedro (BBSB12407) to ARJ, SJH, and the Manchester Centre for Integrative Systems Biology which supported NS.

References

1. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.** *Molecular & Cellular Proteomics* 2002, **1(5):**376-386.
2. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP: **Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(12):**6940-6945.
3. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nature Biotechnology* 1999, **17(10):**994-999.
4. Ross PL, Huang YLN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.** *Molecular & Cellular Proteomics* 2004, **3(12):**1154-1169.

5. Beynon RJ, Doherty MK, Pratt JM, Gaskell SJ: **Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides.** *Nature Methods* 2005, **2(8):**587-589.
6. Pratt JM, Simpson D, Doherty M, Rivers J, Gaskell SJ, Beynon RJ: **Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes.** *Nature Protocols* 2006, **1:**1029-1043.
7. de Godoy LMF, Olsen JV, de Souza GA, Li GQ, Mortensen P, Mann M: **Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system.** *Genome Biology* 2006, **7(6):**.
8. Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu WM, Aebersold R: **A common open representation of mass spectrometry data and its application to proteomics research.** *Nature Biotechnology* 2004, **22(11):**1459-1466.
9. Taylor CF, Hermjakob H, Julian RK, Garavelli JS, Aebersold R, Apweiler R: **The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI).** *Omics-a Journal of Integrative Biology* 2006, **10(2):**145-151.
10. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik R, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li YX, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SGN, Sander C, Bork P, Zhu WM, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios L, Eisenberg D, Steipe B, Hogue C, Apweiler R: **The HUPOPSI's Molecular Interaction format - a community standard for the representation of protein interaction data.** *Nature Biotechnology* 2004, **22(2):**177-183.
11. Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, DeGreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK, Laursen K, Oliver SG, Paton NW, Sarkans U, Sansone S, Stoeckert CJ, Taylor CF, Whetzel PL, White JA, Spellman P, Piazarro A: **The functional genomics experimental model (FuGE): an extensible framework for standards in functional genomics.** *Nature Biotechnology* 2006, **in press.:**.
12. Jones AR, Paton NW: **An analysis of extensible modelling for functional genomics data.** *Bmc Bioinformatics* 2005, **6:**.
13. Prince JT, Carlson MW, Wang R, Lu P, Marcotte EM: **The need for a public proteomics repository.** *Nature Biotechnology* 2004, **22(4):**471-472.
14. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas project.** *Nucleic Acids Research* 2006, **34:**D655-D658.
15. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *Journal of Proteome Research* 2004, **3(6):**1234-1242.
16. McLaughlin T, Siepen JA, Selley J, Lynch JA, Lau KW, Yin HJ, Gaskell SJ, Hubbard SJ: **PepSeeker: a database of proteome peptide identifications for investigating fragmentation patterns.** *Nucleic Acids Research* 2006, **34:**D649-D654.
17. Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S, Stead D, Yin Z, Brown AJP, Hesketh A, Chater K, Hansson L, Mewissen M, Ghazal P, Howard J, Lilley KS, Gaskell SJ, Brass A, Hubbard SJ, Oliver SG, Paton NW: **PEDRo: A database for storing, searching and disseminating experimental proteomics data.** *Bmc Genomics* 2004, **5:**.
18. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: The proteomics identifications database.** *Proteomics* 2005, **5(13):**3537-3545.
19. Orchard S, Hermjakob H, Taylor CF, Potthost F, Jones P, Zhu WM, Julian RK, Apweiler R: **Second Proteomics Standards Initiative Spring Workshop.** *Expert Review of Proteomics* 2005, **2(3):**287-289.
20. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide identifications for the proteomics community.** *Nucleic Acids Research* 2006, **34:**D659-D663.
21. Cote RG, Jones P, Apweiler R, Hermjakob H: **The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries.** *Bmc Bioinformatics* 2006, **7:**.
22. Falkner JA, Ulintz PJ, Andrews PC: **http://www.proteomecommons.org/.** .
23. Shadforth IP, Dunkley TPJ, Lilley KS, Bessant C: **i-Tracker: For quantitative proteomics using iTRAQ (TM).** *Bmc Genomics* 2005, **6:**145.
24. Creasy DM, Cottrell JS: **Unimod: Protein modifications for mass spectrometry.** *Proteomics* 2004, **4(6):**1534-1536.
25. **PrideWizard** [http://www.mcisb.org/software/PrideWizard]
26. **PedroDC Capture Tool** [http://pedrodownload.man.ac.uk]
27. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20(18):**3551-3567.
28. Phan IQH, Pilbout SF, Fleischmann W, Bairoch A: **NEWT, a new taxonomy portal.** *Nucleic Acids Research* 2003, **31(13):**3822-3823.
29. Broadhead R, Dawe HR, Farr H, Griffiths S, Hart SR, Portman N, Shaw MK, Ginger ML, Gaskell SJ, McKean PG, Gull K: **Flagellar motility is required for the viability of the bloodstream trypanosome.** *Nature* 2006, **440(7081):**224-227.
30. Coulthurst SJ, Lilley KS, Salmond GPC: **Genetic and proteomic analysis of the role of luxS in the enteric phytopathogen, Erwinia carotovora.** *Molecular Plant Pathology* 2006, **7(1):**31-45.
31. Dunkley TPJ, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, Griffin JL, Bessant C, Brandizzi F, Hawes C, Watson RB, Dupree P, Lilley KS: **Mapping the Arabidopsis organelle proteome.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103(17):**6518-6523.
32. Unwin RD, Smith DL, Blinco D, Wilson CL, Miller CJ, Evans CA, Jaworska E, Baldwin SA, Barnes K, Pierce A, Spooncer E, Whetton AD: **Quantitative proteomics reveals posttranslational control as a regulatory factor in primary hematopoietic stem cells.** *Blood* 2006, **107(12):**4687-4694.