# *PRIDE Converter* 2
# GUI User Manual

## Table of Contents

# 1. Introduction

The PRIDE Converter 2 tool suite is a composed of 4 independent applications:

- The ***PRIDE Converter 2*** application will convert MS search result files containing identification and spectra into PRIDE XML.
- The ***PRIDE mzTab Generator*** will produce skeleton mzTab files from MS search results files. These skeleton files require either manual or scripted editing to add quantitation and/or gel information.
- The ***PRIDE XML Filter*** will remove identifications or spectra from PRIDE XML files based on a series of configurable filters.
- The ***PRIDE XML Merger*** will combine several PRIDE XML files into a single one.

All tools have both a Graphical User Interface (GUI) and a Command-Line Interface (CLI). The GUIs have been designed to provide a rich, user-friendly interface while the CLIs have been developed mainly for tool and pipeline developers to be able to integrate the PRIDE Converter 2 tools in their own software to provide an efficient way to generate PRIDE XML from their own resources.

# 2. *PRIDE Converter 2* Requirements

## 2.1. System Requirements

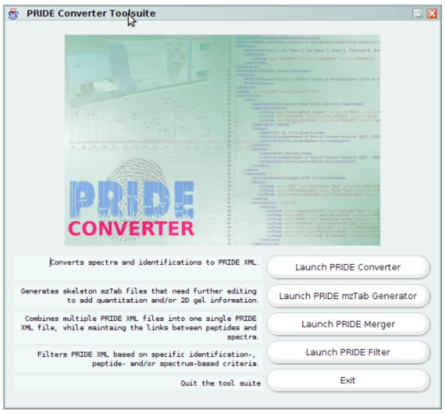
- Java: JRE 1.5 +
- CPU: 1 gigahertz (GHz) or faster 32-bit or 64-bit processor.
- Memory: 1 gigabyte (GB) RAM.
- Hard Disk: 55 MB available for installation, more if doing file conversions.
- Platform: Tested in Mac OS X, Linux, and Windows (XP, Vista, 7).

## 2.2. Additional Requirements

PRIDE Converter 2 requires Internet access for connection to the Ontology Lookup Service (OLS) web service and automatic access to PubMed records already published.

# 3. *PRIDE Converter 2* Tool Suite GUI User Guide

To start the GUI, users must either double-click on the *PRIDE Converter 2* jar file or invoke the command-line without any arguments. The first screen that the user sees will be the tool selection screen. From there, the user can select which tool from the tool suite to launch (Figure 2).



Converts spectra and identifications to PRIDE XML.

Generates skeleton mzTab files that need further editing to add quantitation and/or 2D gel information.

Combines multiple PRIDE XML files into one single PRIDE XML file, while maintaing the links between peptides and spectra.

Filters PRIDE XML based on specific identification-, peptide- and/or spectrum-based criteria.

Quit the tool suite

**Figure 2** – PRIDE Converter Tool Suite tool selection form

Each tool shares a common design framework to ensure a common and consistent look and feel across all applications (see Figure 3).

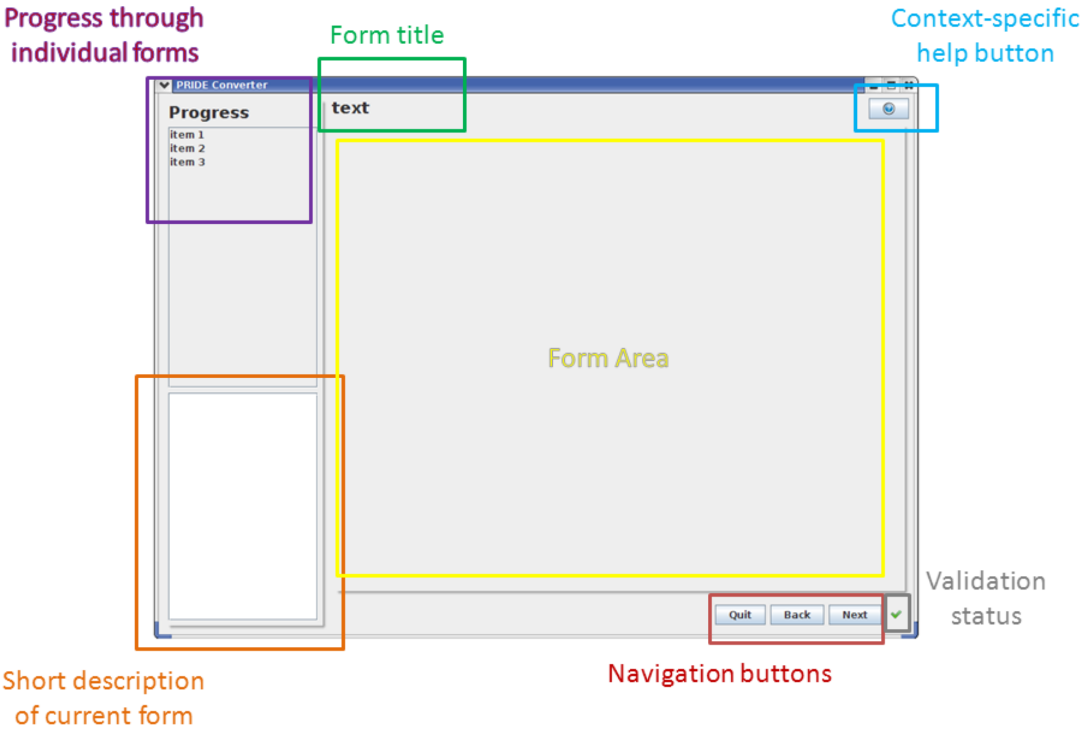


**Figure 3.** The Common GUI Navigation Panel.

Tools are made up of a series of forms, that can be navigated using the Next/Back/Quit buttons at the bottom right corner of the nagivation frame. Each form will provide information that is used to update the GUI: a form title, a short description of its purpose and required fields, a context-specific help page and a mechanism to validate the form content to ensure that all required fields are filled in and contain correctly-formatted values.

The result of this validation are shown next to the navigation buttons (✓ if all fields are OK, ⚠ on warnings, ✖ on errors). If errors are detected, the users will be shown an error dialog with more information on the validation rule that generated the error. Users will not be able to proceed with the conversion process until all errors are corrected.

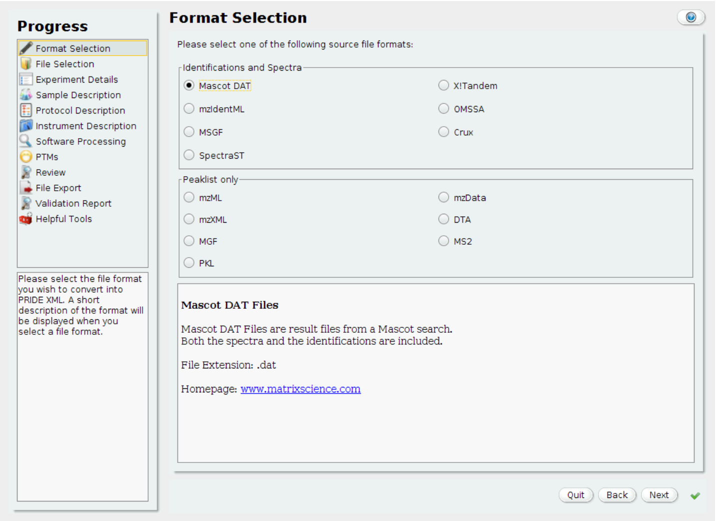
PRIDE Converter 2 forms maintain a standard functionality scheme. Any field marked with a red asterisk (*) is mandatory and the user cannot proceed to the next form until all mandatory fields are filled. The ✚ button can be used to add entries. When entries are selected from a table, the ✎ button can be used to edit those entries and the 🗑 button can be used to delete them.

## 3.1. PRIDE Converter 2

When launched in GUI mode, *PRIDE Converter 2* guides the user through a 12-step process to convert their search engine output files into well-annotated PRIDE XML files.

### 3.1.1. Format Selection

Input format selection is the first choice that the user must make. When selecting a format, the user will be given information on the exact file type(s) required, as well as links to the vendor and/or format specification documentation (Figure 4). Please note that Mascot conversion is not available on the Mac OS platform, due to limitations imposed by the Mascot API itself. Please refer to Table 1 for the supported formats in *PRIDE Converter 2*.



Please select the file format you wish to convert into PRIDE XML. A short description of the format will be displayed when you select a file format.

Format Selection

Please select one of the following source file formats:

Identifications and Spectra

● Mascot DAT ○ X!Tandem
○ mzIdentML ○ OMSSA
○ MSGF ○ Crux
○ SpectraST

Peaklist only

○ mzML ○ mzData
○ mzXML ○ DTA
○ MGF ○ MS2
○ PKL

Mascot DAT Files

Mascot DAT Files are result files from a Mascot search.
Both the spectra and the identifications are included.

File Extension: .dat

Homepage: www.matrixscience.com

**Figure 4** – *PRIDE Converter 2* Format Selection form

**Table 1.** Supported Formats in *PRIDE Converter 2*.

| Format Name | File Type | Data Content |
|---|---|---|
| Mascot | .dat | Spectra and Identifications |
| mzIdentML | .xml | Spectra and Identifications |
| X!Tandem | .xml | Spectra and Identifications |
| OMSSA | .csv | Spectra and Identifications |
| SpectraST | .txt | Spectra and Identifications |
| CRUX | .txt | Spectra and Identifications |
| MSGF | .txt | Spectra and Identifications |
| Proteome Discoverer | .msf | Spectra and Identifications |
| mzML | .xml | Spectra Only |
| DTA | .dta | Spectra Only |
| MGF | .mgf | Spectra Only |
| mzData | .xml | Spectra Only |
| mzXML | .xml | Spectra Only |
| PKL | .pkl | Spectra Only |

### 3.1.2. File Selection

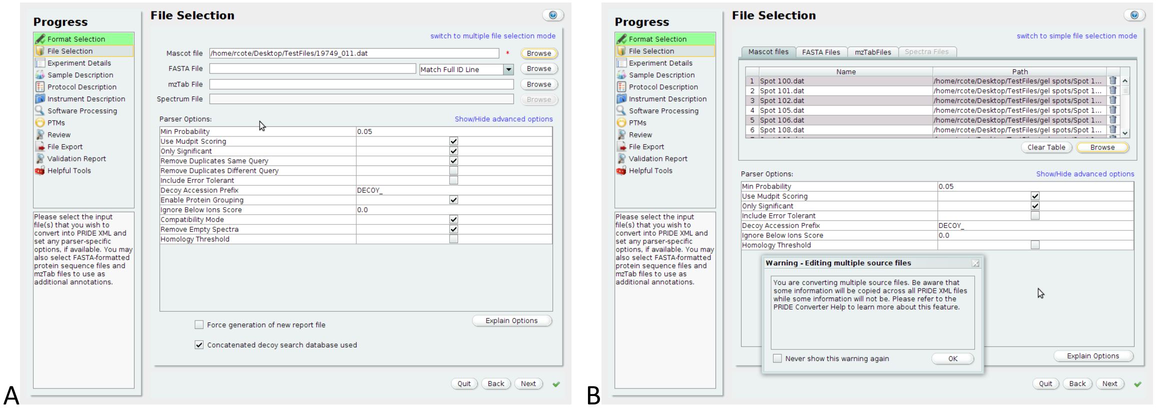
The next form is for file selection and setting *DAO* options, if applicable (Figure 5). Each format-specific *DAO* can have one or several custom options that can be set through the GUI (for example, the Mascot *DAO*, as shown in Figure 5). Sensible default options are always provided and only the basic required options are shown by default. Power users have the option to show all available options (if applicable) and those options will be stored in the report file such that the user can always review how the conversion process was configured. Tool tips will provide a description of the options and sensible/default values. Currently, only the peaklist-only *DAO*s do not have any options.

Some MS search result file formats link to spectra that are kept in external files (X!Tandem, mzIdentML and CRUX for example). For those *DAO*s, the user has the possibility to select a file containing the spectra to include in the conversion process. For the *DAO*s where externally-linked spectra are not applicable, the option to select spectra files is disabled.

By default, users will be shown a form that will only provide single-file selection dialogues but users who wish to do multiple file batch conversion can switch to an alternative view, which will allow them to select multiple files of each type (source, mzTab, spectra, etc).

When converting multiple files at once, a warning message will be displayed to the user indicating that some metadata will be copied across all files while some will not. The information that will be common across all files will include: sample annotations, post-translational modifications (PTM), protocol and instrumentation description and any other experiment-level annotations. Software processing, which can include file-specific software settings, score thresholds and false discovery rates will be taken from the individual report files associated with each input file to be converted.

Once the files have been selected, the *PRIDE Converter 2* runs in *prescan* mode to generate the report files, where the annotations that the user will input in the coming steps will be stored prior to actual conversion. Report file generation can take anywhere between seconds and several minutes and is directly proportional to the size and complexity of the input files. If several files are being converted at once, a report file needs to be generated for each input file, which will compound the time required.



Please select the input file(s) that you wish to convert into PRIDE XML and set any parser-specific options, if available. You may also select FASTA-formatted protein sequence files and mzTab files to use as additional annotations.

Mascot file: /home/rcote/Desktop/TestFiles/19749_011.dat *
FASTA File: [Match Full ID Line]
mzTab File:
Spectrum File:

Parser Options:

| | |
|---|---|
| Min Probability | 0.05 |
| Use Mudpit Scoring | ☑ |
| Only Significant | ☑ |
| Remove Duplicates Same Query | ☑ |
| Remove Duplicates Different Query | ☐ |
| Include Error Tolerant | ☐ |
| Decoy Accession Prefix | DECOY_ |
| Enable Protein Grouping | ☑ |
| Ignore Below Ions Score | 0.0 |
| Compatibility Mode | ☑ |
| Remove Empty Spectra | ☑ |
| Homology Threshold | ☐ |

☐ Force generation of new report file

☑ Concatenated decoy search database used

| | Name | Path |
|---|---|---|
| 1 | Spot 100.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 2 | Spot 101.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 3 | Spot 102.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 4 | Spot 105.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 5 | Spot 106.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 6 | Spot 108.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |

Parser Options:

| | |
|---|---|
| Min Probability | 0.05 |
| Use Mudpit Scoring | ☑ |
| Only Significant | ☑ |
| Include Error Tolerant | ☐ |
| Decoy Accession Prefix | DECOY_ |
| Ignore Below Ions Score | 0.0 |
| Homology Threshold | ☐ |

Warning - Editing multiple source files

You are converting multiple source files. Be aware that some information will be copied across all PRIDE XML files while some information will not be. Please refer to the PRIDE Converter Help to learn more about this feature.

☐ Never show this warning again

**Figure 5** – *PRIDE Converter 2* File Selection form. Panel A shows the single-file selection mode while panel B shows the multi-file selection mode. Note that panel A shows the full set of options for the Mascot DAO while panel B only shows the default options.

Users have the possibility to link their search database protein sequences to the PRIDE XML files that they generate. This is particularly helpful in the case where custom-made protein sequence databases are used. The protein sequence database must be exported in FASTA format. From experience, we have observed that the protein accession that is assigned to an identification can differ significantly from the FASTA header, depending on if and how the search engine processes the protein sequence database. To solve this issue, PRIDE Converter 2 has several options on how the FASTA file should be processed during conversion.

### 3.1.2.1. FASTA – First word

In this scenario, the search engine assigns the first word of the FASTA ID line as the identification accession. Given a FASTA ID line that looks like the following:

```
>P30443 Full=HLA class I histocompatibility antigen, A-1 alpha chain;
```

The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
    <Accession>P30443</Accession>
    ...
    <FastaSequenceReference>8</FastaSequenceReference>
</Identification>
```

The *FastaSequenceReference* points to a Sequence element, as illustrated below.

```
<Sequence id="8" accession="P30443">
MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIEQEGPEYWDQETRNMKA
HSQTDRANLGTLRGYYNQSEDGSHTIQIMYGCDVGPDGRFLRGYRQDAYDGKDYIALNEDLRSWTAADMAAQITKRKWEAVHAAEQRRVYLEG
RCVDGLRRYLENGKETLQRTDPPKTHMTHHPISDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQ
RYTCHVQHEGLPKPLTLRWELSSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSDRKGGSYTQAASSDSAQGSDVSLTACKV
</Sequence>
```

### 3.1.2.2. FASTA – Full ID line

In this scenario, the search engine assigns the complete ID line FASTA ID line as the identification accession. This occurs when the search engine does not process or index the source FASTA database, and is especially useful when custom databases or custom identifiers are used.

The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
    <Accession>Custom Fasta Header Protein 1</Accession>
    ...
    <FastaSequenceReference>12</FastaSequenceReference>
</Identification>

    ...

<Sequence id="12" accession="Custom Fasta Header Protein 1">
MGDREQLLQRARLAEQAERYDDMASAMKAVTELNEPLSNEDRNLLSVAYKNVVGARRSSWRVISSIEQKTMADGNEKKLEKVKAYREKIEKEL
ETVCNDVLSLLDKFLIKNCNDFQYESKVFYLKMKGDYYRYLAEVASGEKKNSVVEASEAAYKEAFEISKEQMQPTHPIRLGLALNFSVFYYEI
QNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDQQDEEAGEGN
</Sequence>
```

### 3.1.2.3. FASTA – Match UniProt AC

If the FASTA file used contains ID lines that are generated according to the UniProtKB specification, they should look something like the following:

```
    >sw|P31946|1433B_HUMAN RecName: Full=14-3-3 protein beta/alpha;AltName: Full=Protein
1054;AltName: Full=Protein kinase C inhibitor protein 1;Short=KCIP-1;Contains:RecName:
Full=14-3-3 protein beta/alpha, N-terminally processed;
```

Some search engines will process these FASTA ID lines and only assign the UniProtKB protein accession. The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
    <Accession>P31946</Accession>
    ...
    <FastaSequenceReference>9</FastaSequenceReference>
</Identification>

...

<Sequence id="9" accession="P31946">
MTMDKSELVQKAKLAEQAERYDDMAAAMKAVTEQGHELSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTERNEKKQQMGKEYREKIEAELQ
DICNDVLELLDKYLIPNATQPESKVFYLKMKGDYFRYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNS
PEKACSLAKTAFDEAIAELDTLNEESYKDSTLIMQLLRDNLTLWTSENQGDEGDAGEGEN
</Sequence>
```

### 3.1.2.3. FASTA – Match UniProt ID

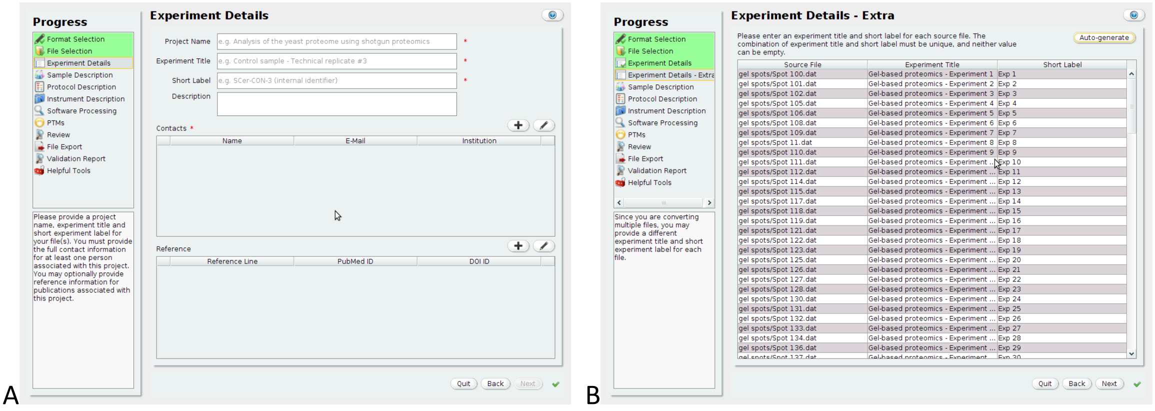
Using the same FASTA ID line as in the preceding section, some search engines will assign the UniProtKB ID. The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
    <Accession>1433B_HUMAN</Accession>
    <CuratedAccession>P31946</CuratedAccession>
    ...
    <FastaSequenceReference>9</FastaSequenceReference>
</Identification>

    ...

<Sequence id="9" accession="P31946">
MTMDKSELVQKAKLAEQAERYDDMAAAMKAVTEQGHELSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTERNEKKQQMGKEYREKIEAELQ
DICNDVLELLDKYLIPNATQPESKVFYLKMKGDYFRYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNS
PEKACSLAKTAFDEAIAELDTLNEESYKDSTLIMQLLRDNLTLWTSENQGDEGDAGEGEN
</Sequence>
```

Note, in this last case, the presence of a *CuratedAccession* element where the correct UniProtKB accession has been linked to the UniProtKB identifier. Users are strongly urged to use UniProt accessions instead of identifiers, as entry identifiers are notoriously changeable, making data comparisons and downstream analysis more difficult and cumbersome.

### 3.1.3. Experimental Details

The first annotation form deals with experimental and project details (Figure 6). Users must provide a project name, an experiment title and a short label to describe the file being converted. In the case where multiple files are being converted at once, the user needs to provide a project name in this form but will have the opportunity to provide individual experiment titles and short labels in a special form which is only displayed for this purpose. Users need to provide contact information for at least one person – usually the data submitter or the primary investigator. A full name, valid e-mail address and institute contact details are required for each contact. Contacts can be saved as templates to avoid the necessity to re-enter the information at each run of the application. Users also have the option to provide a reference that supports the data being submitted to PRIDE. References can either be provided as free text or use an automatic PubMed ID lookup to retrieve the full citation, if indexed and available online. An Internet connection is required for this.



Project Name: e.g. Analysis of the yeast proteome using shotgun proteomics *

Experiment Title: e.g. Control sample - Technical replicate #3 *

Short Label: e.g. SCer-CON-3 (internal identifier) *

Description:

Please provide a project name, experiment title and short experiment label for your file(s). You must provide the full contact information for at least one person associated with this project. You may optionally provide reference information for publications associated with this project.

Please enter an experiment title and short label for each source file. The combination of experiment title and short label must be unique, and neither value can be empty.

| Source File | Experiment Title | Short Label |
|---|---|---|
| gel spots/Spot 100.dat | Gel-based proteomics - Experiment 1 | Exp 1 |
| gel spots/Spot 101.dat | Gel-based proteomics - Experiment 2 | Exp 2 |
| gel spots/Spot 102.dat | Gel-based proteomics - Experiment 3 | Exp 3 |
| gel spots/Spot 105.dat | Gel-based proteomics - Experiment 4 | Exp 4 |
| gel spots/Spot 106.dat | Gel-based proteomics - Experiment 5 | Exp 5 |
| gel spots/Spot 108.dat | Gel-based proteomics - Experiment 6 | Exp 6 |
| gel spots/Spot 109.dat | Gel-based proteomics - Experiment 7 | Exp 7 |
| gel spots/Spot 11.dat | Gel-based proteomics - Experiment 8 | Exp 8 |
| gel spots/Spot 110.dat | Gel-based proteomics - Experiment 9 | Exp 9 |
| gel spots/Spot 111.dat | Gel-based proteomics - Experiment ... | Exp 10 |
| gel spots/Spot 112.dat | Gel-based proteomics - Experiment ... | Exp 11 |
| gel spots/Spot 114.dat | Gel-based proteomics - Experiment ... | Exp 12 |
| gel spots/Spot 115.dat | Gel-based proteomics - Experiment ... | Exp 13 |
| gel spots/Spot 117.dat | Gel-based proteomics - Experiment ... | Exp 14 |
| gel spots/Spot 118.dat | Gel-based proteomics - Experiment ... | Exp 15 |
| gel spots/Spot 119.dat | Gel-based proteomics - Experiment ... | Exp 16 |
| gel spots/Spot 121.dat | Gel-based proteomics - Experiment ... | Exp 17 |
| gel spots/Spot 122.dat | Gel-based proteomics - Experiment ... | Exp 18 |
| gel spots/Spot 123.dat | Gel-based proteomics - Experiment ... | Exp 19 |
| gel spots/Spot 125.dat | Gel-based proteomics - Experiment ... | Exp 20 |
| gel spots/Spot 126.dat | Gel-based proteomics - Experiment ... | Exp 21 |
| gel spots/Spot 127.dat | Gel-based proteomics - Experiment ... | Exp 22 |
| gel spots/Spot 128.dat | Gel-based proteomics - Experiment ... | Exp 23 |
| gel spots/Spot 130.dat | Gel-based proteomics - Experiment ... | Exp 24 |
| gel spots/Spot 131.dat | Gel-based proteomics - Experiment ... | Exp 25 |
| gel spots/Spot 132.dat | Gel-based proteomics - Experiment ... | Exp 26 |
| gel spots/Spot 133.dat | Gel-based proteomics - Experiment ... | Exp 27 |
| gel spots/Spot 134.dat | Gel-based proteomics - Experiment ... | Exp 28 |
| gel spots/Spot 136.dat | Gel-based proteomics - Experiment ... | Exp 29 |
| gel spots/Spot 137.dat | Gel-based proteomics - Experiment | Exp 30 |

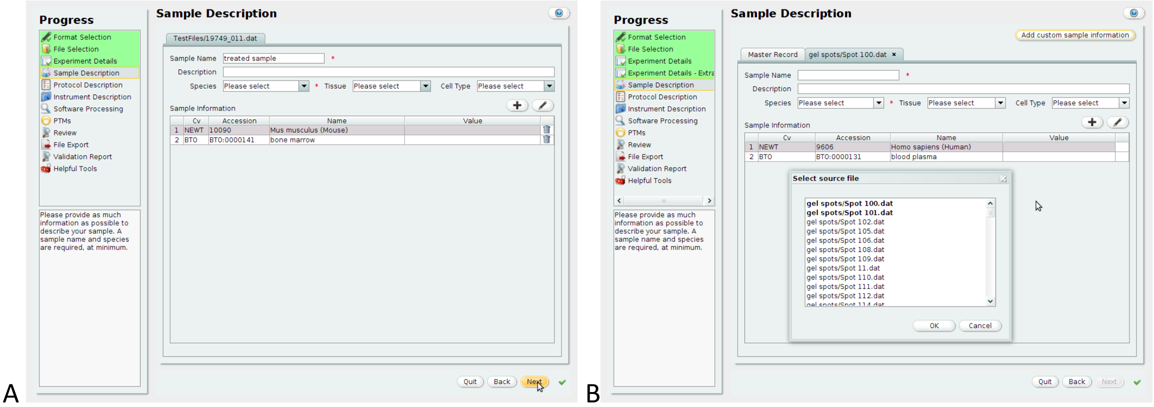
Since you are converting multiple files, you may provide a different experiment title and short experiment label for each file.

**Figure 6** – *PRIDE Converter 2* Experiment Details form. Panel A shows the single-file conversion mode while panel B shows the additional step that is required to provide experiment titles and short labels for each source file to be converted.

If multiple files are being converted, the project name will be identical across all files, but the combination of experiment title and short label must be unique to each file. Users can auto-generate placeholder data, based on the project name, but are encouraged to fill in these fields with meaningful information.

### 3.1.4. Sample Description

The sample description form (Figure 7) allows the user to provide taxonomy, tissue type and cell type annotations using pull-down menus that contain the most commonly used values already in PRIDE. A sample name and species are required. Users still have the possibility to use the comprehensive Ontology Lookup Service (OLS) to look up alternative terms that are not already provided. If multiple files are being converted, the information in the “master report” will be automatically copied over to each file. Users will have the possibility to add file-specific controlled vocabulary (CV) terms by clicking on the “add custom sample information” button and selecting the proper source file. Files with custom annotations will be shown in bold. Any annotation coming from the master report will not be editable or deletable from the individual files' panel, but users will be able to add new terms. Any terms added or deleted from the master report view will always be propagated to all files.



A:

Please provide as much information as possible to describe your sample. A sample name and species are required, at minimum.

Sample Name: treated sample *
Description:
Species: Please select * Tissue: Please select Cell Type: Please select

Sample Information

| | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | NEWT | 10090 | Mus musculus (Mouse) | |
| 2 | BTO | BTO:0000141 | bone marrow | |

B:

Please provide as much information as possible to describe your sample. A sample name and species are required, at minimum.

Sample Name: *
Description:
Species: Please select * Tissue: Please select Cell Type: Please select

Sample Information

| | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | NEWT | 9606 | Homo sapiens (Human) | |
| 2 | BTO | BTO:0000131 | blood plasma | |

Select source file:

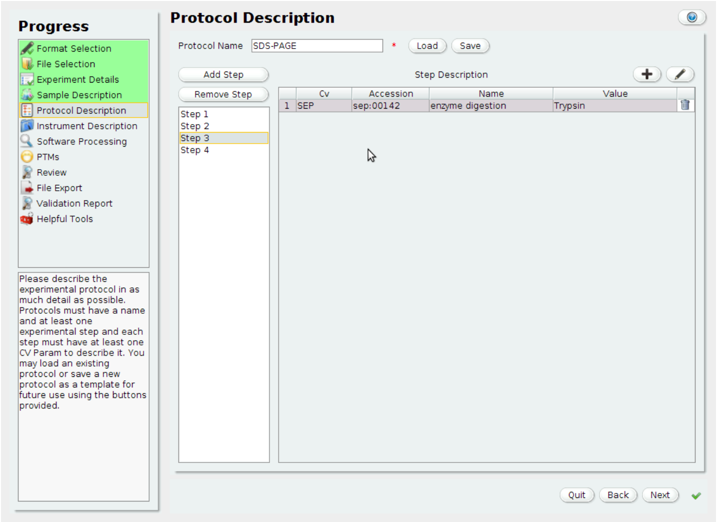
**gel spots/Spot 100.dat**
**gel spots/Spot 101.dat**
gel spots/Spot 102.dat
gel spots/Spot 105.dat
gel spots/Spot 106.dat
gel spots/Spot 108.dat
gel spots/Spot 109.dat
gel spots/Spot 11.dat
gel spots/Spot 110.dat
gel spots/Spot 111.dat
gel spots/Spot 112.dat
gel spots/Spot 114.dat

**Figure 7** – *PRIDE Converter 2* Sample Description form. Panel A shows the single-file conversion mode while panel B shows the how annotation of multiple files. Note that, in panel B, the entries for the selected individual file cannot be deleted, as they were set in the 'Master Record' tab.

## 3.1.5. Protocol Description

In the protocol description form (Figure 8), *PRIDE Converter 2* requires a protocol name to be provided, at minimum. Protocols can have multiple steps, and each step can be described with one or several CV terms. Please look at the templates provided with the tool for examples. The "Load" button can be used to load existing templates. A popup window will be shown, where users can select the desired template from a list. To save a template once all of the information is correctly entered, users simply need to enter a name (in this case, a protocol name) and click on the "Save" button. A file will be saved to the templates folder in the installation directory of *PRIDE Converter 2*. Should there already be a file with the same name, the user will be prompted to either rename the template or allow the existing one to be overwritten.

*PRIDE Converter 2* provides the possibility to save commonly-used annotations such as sample, instrumentation and protocols as templates which can be reused in subsequent conversions. A set of basic templates is provided with the tool suite. Users can load and update templates to better suit their own requirements.



Please describe the experimental protocol in as much detail as possible. Protocols must have a name and at least one experimental step and each step must have at least one CV Param to describe it. You may load an existing protocol or save a new protocol as a template for future use using the buttons provided.

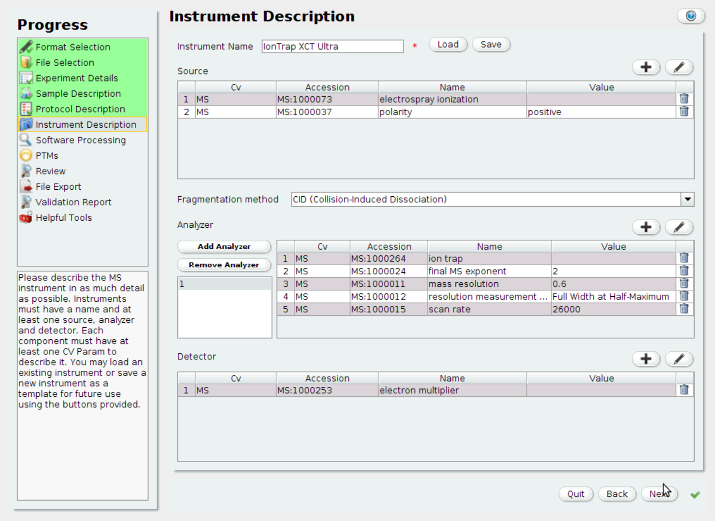
Protocol Name: SDS-PAGE *

| | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | SEP | sep:00142 | enzyme digestion | Trypsin |

**Figure 8** – *PRIDE Converter 2* Protocol Description form.

### 3.1.6. Instrument Description

The instrument description form (Figure 9) is another form that can use templates, as in most cases, the instrumentation available to a user will not change dramatically over time. The instrument name in mandatory. "Source", "Analyzer" and "Detector" elements need to be annotated, preferably with CV terms from the Proteomics Standards Initiative (PSI) Mass Spectrometry (MS) controlled vocabulary (http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo).



Please describe the MS instrument in as much detail as possible. Instruments must have a name and at least one source, analyzer and detector. Each component must have at least one CV Param to describe it. You may load an existing instrument or save a new instrument as a template for future use using the buttons provided.

Instrument Description

Instrument Name: IonTrap XCT Ultra *

Source

|  | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | MS | MS:1000073 | electrospray ionization |  |
| 2 | MS | MS:1000037 | polarity | positive |

Fragmentation method: CID (Collision-Induced Dissociation)

Analyzer

1

|  | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | MS | MS:1000264 | ion trap |  |
| 2 | MS | MS:1000024 | final MS exponent | 2 |
| 3 | MS | MS:1000011 | mass resolution | 0.6 |
| 4 | MS | MS:1000012 | resolution measurement ... | Full Width at Half-Maximum |
| 5 | MS | MS:1000015 | scan rate | 26000 |

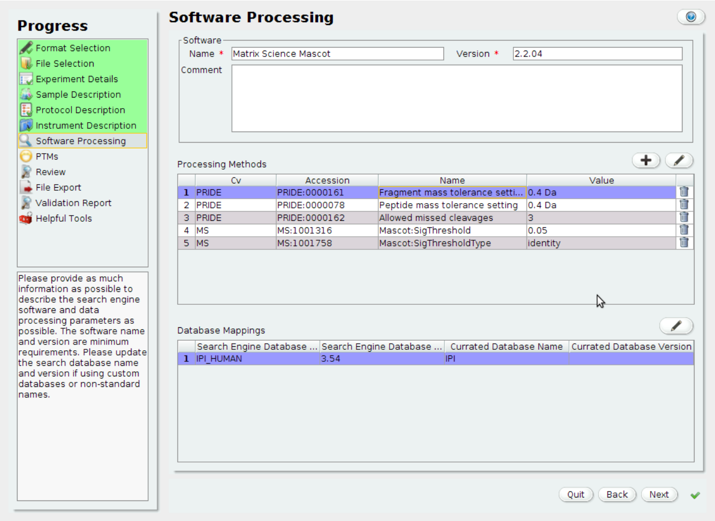
Detector

|  | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | MS | MS:1000253 | electron multiplier |  |

**Figure 9** – *PRIDE Converter 2* Instrument Description form.

### 3.1.7. Software Processing

The software processing form (Figure 10) report information on the software settings and the search database used. The "software" and "processing method" information will more often than not be reported directly by the DAO and should not be modified. **Users should be aware that If converting multiple files**, these two sections will be taken directly from the source file and any additions made to the GUI will not be carried across all converted files.



Please provide as much information as possible to describe the search engine software and data processing parameters as possible. The software name and version are minimum requirements. Please update the search database name and version if using custom databases or non-standard names.

Software Processing

Software

Name * Matrix Science Mascot

Version * 2.2.04

Comment

Processing Methods

| | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | PRIDE | PRIDE:0000161 | Fragment mass tolerance setti... | 0.4 Da |
| 2 | PRIDE | PRIDE:0000078 | Peptide mass tolerance setting | 0.4 Da |
| 3 | PRIDE | PRIDE:0000162 | Allowed missed cleavages | 3 |
| 4 | MS | MS:1001316 | Mascot:SigThreshold | 0.05 |
| 5 | MS | MS:1001758 | Mascot:SigThresholdType | identity |

Database Mappings

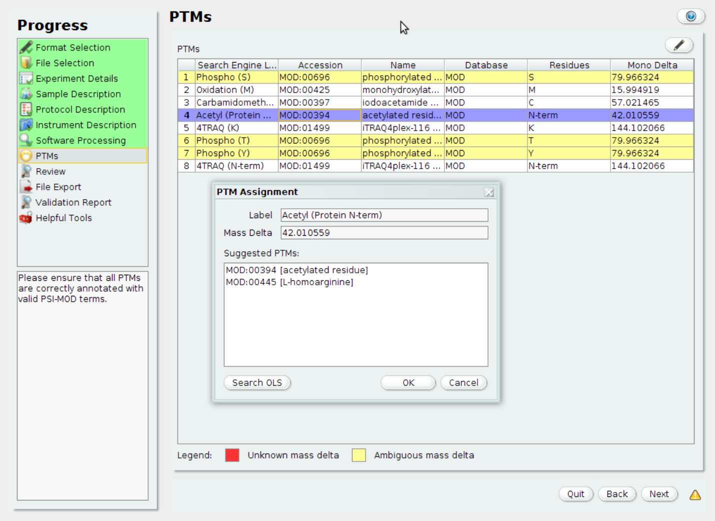
| | Search Engine Database ... | Search Engine Database ... | Currated Database Name | Currated Database Version |
|---|---|---|---|---|
| 1 | IPI_HUMAN | 3.54 | IPI | |

**Figure 10** – *PRIDE Converter 2* Software Processing form.

The "search database" mappings will help standardize the names of the sources of the protein sequences that have been used to generate the search databases. The most common sources of protein sequences have been included (UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, International Protein Index (IPI), Ensembl, RefSeq, NCBI nr database) and the users are encouraged to provide either a release number or the date the search database was built. Users can also manually enter a custom database name, if the supplied options are not appropriate. The "curated database name" is mandatory, but the "curated database version", while desired, is not. A warning will be generated if the database version is omitted however.

## 3.1.8. PTMs

The next step is the review and/or correction of PTM assignments (Figure 11). Please refer to the section entitled "PRIDE Converter Automatic PTM Assignment" for more information on how PTMs are assigned and how they can be manually corrected or annotated.



Please ensure that all PTMs are correctly annotated with valid PSI-MOD terms.

PTMs

PTMs

| | Search Engine L... | Accession | Name | Database | Residues | Mono Delta |
|---|---|---|---|---|---|---|
| 1 | Phospho (S) | MOD:00696 | phosphorylated ... | MOD | S | 79.966324 |
| 2 | Oxidation (M) | MOD:00425 | monohydroxylat... | MOD | M | 15.994919 |
| 3 | Carbamidometh... | MOD:00397 | iodoacetamide ... | MOD | C | 57.021465 |
| 4 | Acetyl (Protein ... | MOD:00394 | acetylated resid... | MOD | N-term | 42.010559 |
| 5 | 4TRAQ (K) | MOD:01499 | iTRAQ4plex-116 ... | MOD | K | 144.102066 |
| 6 | Phospho (T) | MOD:00696 | phosphorylated ... | MOD | T | 79.966324 |
| 7 | Phospho (Y) | MOD:00696 | phosphorylated ... | MOD | Y | 79.966324 |
| 8 | 4TRAQ (N-term) | MOD:01499 | iTRAQ4plex-116 ... | MOD | N-term | 144.102066 |

PTM Assignment

Label Acetyl (Protein N-term)

Mass Delta 42.010559

Suggested PTMs:

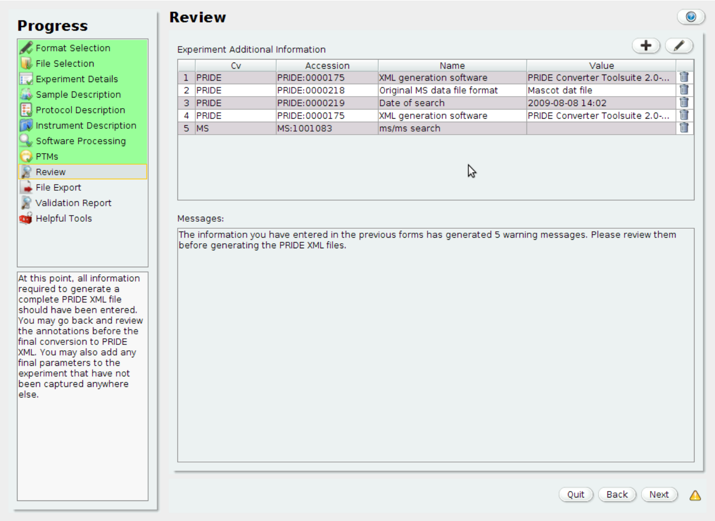
MOD:00394 [acetylated residue]
MOD:00445 [L-homoarginine]

Legend: 🟥 Unknown mass delta 🟨 Ambiguous mass delta

**Figure 11** – *PRIDE Converter 2* PTM form.

## 3.1.9. Review

Once all of the required annotations are provided, the GUI displays a quick review screen before proceeding to copy the metadata across all report files (Figure 12). Any warnings generated by the automatic form validation that occurs at each step will be displayed here. If multiple files are being converted at once, the total number of input files will also be shown. Users can add additional experiment-level CV terms that would reflect the experiment as a whole but are not appropriate in any of the previous sections. The DAOs and *PRIDE Converter 2* itself will automatically add params for ‘XML generation software’ and ‘Original MS data file format’.



At this point, all information required to generate a complete PRIDE XML file should have been entered. You may go back and review the annotations before the final conversion to PRIDE XML. You may also add any final parameters to the experiment that have not been captured anywhere else.

Experiment Additional Information

| | Cv | Accession | Name | Value |
|---|---|---|---|---|
| 1 | PRIDE | PRIDE:0000175 | XML generation software | PRIDE Converter Toolsuite 2.0-... |
| 2 | PRIDE | PRIDE:0000218 | Original MS data file format | Mascot dat file |
| 3 | PRIDE | PRIDE:0000219 | Date of search | 2009-08-08 14:02 |
| 4 | PRIDE | PRIDE:0000175 | XML generation software | PRIDE Converter Toolsuite 2.0-... |
| 5 | MS | MS:1001083 | ms/ms search | |

Messages:

The information you have entered in the previous forms has generated 5 warning messages. Please review them before generating the PRIDE XML files.

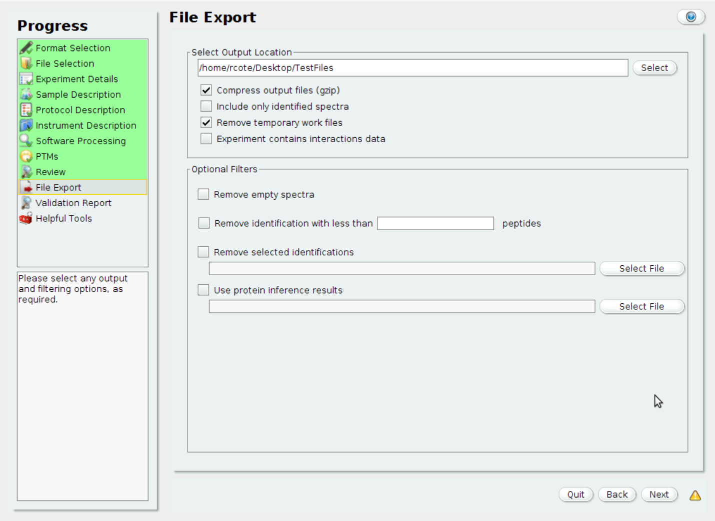
**Figure 12** – *PRIDE Converter 2* Review form.

Updating the report file (if performing a single conversion) or copying the data across all report files (if converting multiple files) can take anywhere between seconds and several minutes and is directly proportional to the number of files that need to be updated.

### 3.1.10. File Export

The next step is the generation of the PRIDE XML files (Figure 13). Alternatively, the graphical conversion process can be halted at this point, as all of the report files are now complete and validated and the conversion process can be scripted and batched using the Command Line Interface (CLI). This is generally only practical for users who need to convert a significantly large number of files and who have access to a cluster of computers where the conversions can be parallelized. In most cases, a single desktop machine with sufficient memory and disk space will be more than sufficient.

*PRIDE Converter 2* can generate compressed files using the gzip algorithm to save on disk space. This is enabled by default, as is the option to remove any temporary work files, such as report files, once the application exits. *PRIDE Converter 2* will always include all spectra in the generated PRIDE XML file unless the “include only identified” spectra option is selected. If the experiment contains interaction data that would benefit from being submitted to the IntAct interaction database (http://www.ebi.ac.uk/intact) at the EBI, users are encouraged to select the relevant option in the export dialog. This will add a CV term in the generated XML file that will be recognized by the submission process and the person whose contact details have been captured will be contacted by an IntAct curator for further discussion.



Please select any output and filtering options, as required.

File Export

Select Output Location

/home/rcote/Desktop/TestFiles Select

☑ Compress output files (gzip)

□ Include only identified spectra

☑ Remove temporary work files

□ Experiment contains interactions data

Optional Filters

□ Remove empty spectra

□ Remove identification with less than ______ peptides

□ Remove selected identifications Select File

□ Use protein inference results Select File

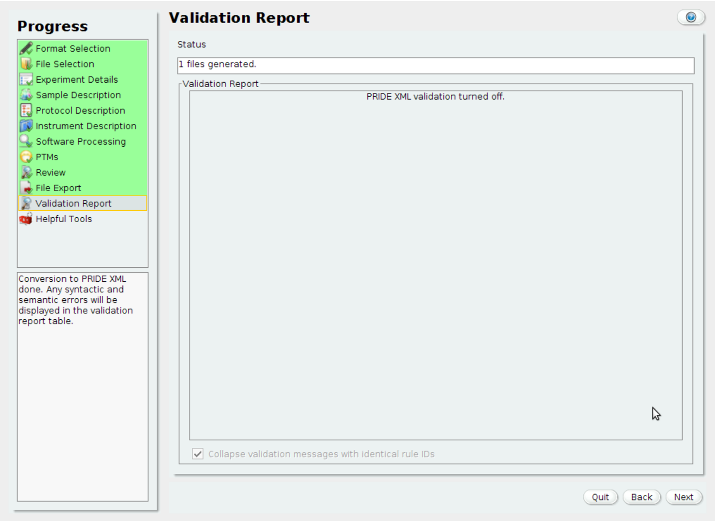
**Figure 13** – *PRIDE Converter 2* File Export form.

This panel also sets the various filtering options, which are discussed in greater detail in the section entitled “PRIDE XML Filter”.

Generating the PRIDE XML files can take anywhere between minutes and hours. This is proportional to the size, number and complexity of the source files. Users should ensure that they have sufficient disk space. If filtering is enabled, the full PRIDE XML file is first written to disk, then written again in its filtered form after it has been processed. The report file and the intermediate PRIDE XML file are only deleted when *PRIDE Converter 2* exits if the “remove temporary files” option is selected.

## 3.1.11. Validation Report

Once the PRIDE XML files are written, a status report is shown, indicating how many PRIDE XML files have been generated (Figure 14).



Conversion to PRIDE XML done. Any syntactic and semantic errors will be displayed in the validation report table.

Validation Report

Status

1 files generated.

Validation Report

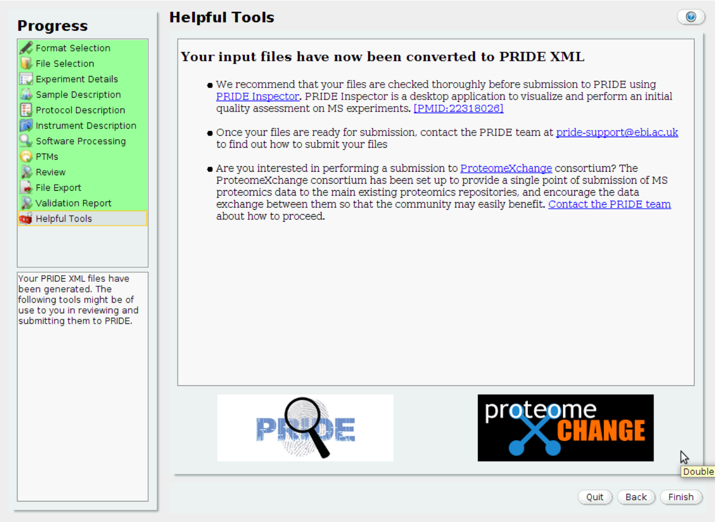
PRIDE XML validation turned off.

☑ Collapse validation messages with identical rule IDs

**Figure 14** – *PRIDE Converter 2* Validation Report form.

### 3.1.12. Helpful Tools

The final screen (Figure 15) of the GUI invites users to review the generated PRIDE XML files using the PRIDE Inspector tool (http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector) and to submit their data *via* the ProteomeXchange (PX) consortium (http://www.proteomexchange.org).



Your PRIDE XML files have been generated. The following tools might be of use to you in reviewing and submitting them to PRIDE.

**Your input files have now been converted to PRIDE XML**

- We recommend that your files are checked thoroughly before submission to PRIDE using PRIDE Inspector. PRIDE Inspector is a desktop application to visualize and perform an initial quality assessment on MS experiments. [PMID:22318026]
- Once your files are ready for submission, contact the PRIDE team at pride-support@ebi.ac.uk to find out how to submit your files
- Are you interested in performing a submission to ProteomeXchange consortium? The ProteomeXchange consortium has been set up to provide a single point of submission of MS proteomics data to the main existing proteomics repositories, and encourage the data exchange between them so that the community may easily benefit. Contact the PRIDE team about how to proceed.

**Figure 15** – *PRIDE Converter 2* Helpful Tools form.

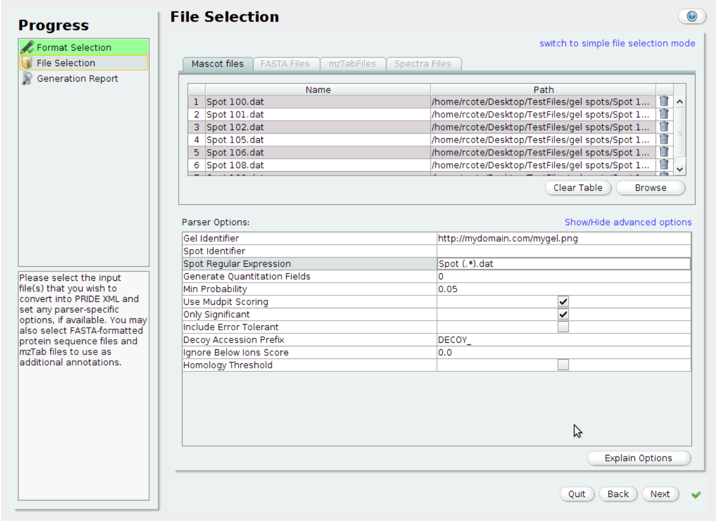
## 3.2. PRIDE mzTab Generator

### 3.2.1. Format Selection

The input format selection form for the PRIDE mzTab Generator is identical to the one used in PRIDE Converter 2 (Figure 4), with the exception that peaklist-only formats are disabled and cannot be selected.

### 3.2.2. File Selection

The look and usability of the file selection form for the PRIDE mzTab Generator is almost identical to that described in the "File Selection" section of the PRIDE Converter 2 GUI user guide, with the addition of 4 mzTab-specific configuration options (Figure 16).



Please select the input file(s) that you wish to convert into PRIDE XML and set any parser-specific options, if available. You may also select FASTA-formatted protein sequence files and mzTab files to use as additional annotations.

| | Name | Path |
|---|---|---|
| 1 | Spot 100.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 2 | Spot 101.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 3 | Spot 102.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 4 | Spot 105.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 5 | Spot 106.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 6 | Spot 108.dat | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |

Parser Options:

| | |
|---|---|
| Gel Identifier | http://mydomain.com/mygel.png |
| Spot Identifier | |
| Spot Regular Expression | Spot (.*).dat |
| Generate Quantitation Fields | 0 |
| Min Probability | 0.05 |
| Use Mudpit Scoring | ☑ |
| Only Significant | ☑ |
| Include Error Tolerant | □ |
| Decoy Accession Prefix | DECOY_ |
| Ignore Below Ions Score | 0.0 |
| Homology Threshold | □ |

**Figure 16** – PRIDE mzTab Generator File Selection form.

The DAO options, if any, will be identical to those present in the PRIDE Converter 2 and will be stored in the mzTab file. In order to avoid unpredictable DAO behaviour, users **must** use the same settings when generating the mzTab files and when using the PRIDE Converter 2. As the options are stored in both the report file and the mzTab file, validity checks will be performed to ensure that both sets of options are identical. Any differences will cause an exception and the Conversion process will be blocked until all options match.

### *3.2.2.1. Adding Gel-based information*

The “gel identifier” option is used to set the gel identifier to be used for identifications in the generated mzTab file. This identifier should correspond to the gel’s “name” in the publication (f.e. “Gel A”, “Gel B” *etc*.). It should enable reviewers and readers to easily link the submitted data to the data presented in the corresponding manuscript.

The “gel spot identifier” option sets the gel spot identifier to be used for identifications in the generated mzTab file. This represents the gel spot’s description used in the manuscript – similar to the “gel identifier” for the gel itself. Expected values would be “Spot 1”, “Spot 2”, *etc.* Together with the “gel identifier” it is therefore possible for the reviewer and reader to unambiguously link findings presented in the manuscript to specific identifications reported in the PRIDE XML file. This option is ignored if the “gel spot regex” option is set. The “gel spot regex” option is used to extract the gel spot identifier based on the source file's name. The value of this option must correspond to a valid Java regular expression. The first matching group in the pattern will be used as a spot identifier.

Both of these fields can also be added manually to an existing mzTab file: the “gel spot identifier” is mapped to the “opt_gel_identifier” column and the “gel spot identifier” to the “opt_gel_spotidentifier” column in the mzTab file. Additional gel specific information can be supplied using the following optional columns in the mzTab file:

| | |
|---|---|
| **opt_gel_url** | The gel's URL to use as a link. |
| **opt_ycoord_pixel** | Y coordinates in pixel of the spot in the gel |
| **opt_xcoord_pixel** | X coordinates in pixel of the spot in the gel |
| **opt_mw** | The protein's (observed) molecular weight. |
| **opt_pi** | The protein's (observed) pI. |

Once these columns are present in the supplied mzTab file, PRIDE Converter will automatically add the corresponding values to the PRIDE XML file. For more information on mzTab and on optional columns in mzTab files please refer to the mzTab format specification at http://mztab.googlecode.com.

### *3.2.2.2. Adding quantitative data*

The “generate quant fields” option will add empty placeholder quantitative fields to the generated mzTab file for the number of specified labels. These fields and columns will need to be manually or programmatically filled in with the proper values before the mzTab file is ready to be used by the PRIDE Converter 2. If, for example, a 4-plex iTRAQ approach was used in which 4 different samples were labelled, the “generate quant fields” should be set to “4”.

As a result, the mzTab generator will automatically add the required meta-data fields to the generated mzTab file:

- A placeholder for the used quantification method (“[UNIT_ID]-quantification_method”)
- A placeholder for the used protein quantification unit (“[UNIT_ID]-protein-quantification_unit”)

- A placeholder for the used peptide quantification unit ("[UNIT_ID]-peptide-quantification_unit")

Additionally, so-called "subsamples" will be added to the mzTab file for every label. These subsamples will contain place holders to describe the used label as well as the characteristics of the labelled (biological) sample:

- A placeholder for the used quantification reagent
- A placeholder for the human readable description of the labelled sample (f.e. "healthy control")

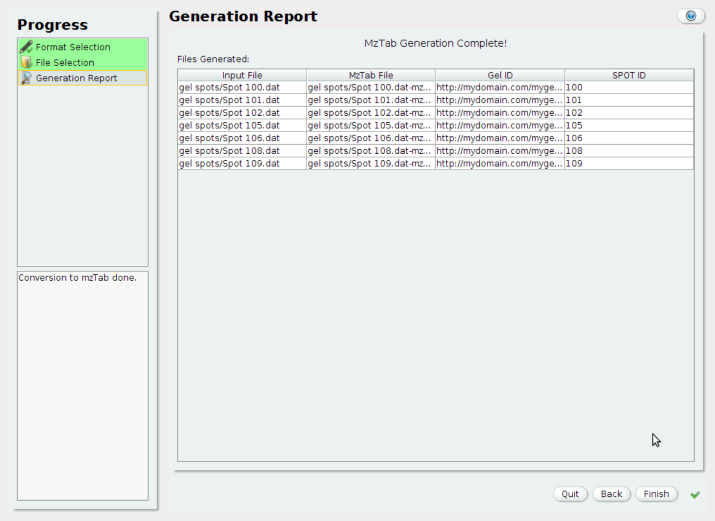
The way these data is represented in the mzTab file is documented in detail in the mzTab format specification (http://mztab.googlecode.com).

At last, the mzTab generator will also add columns to the mzTab file's protein and peptide section to hold the actual measured quantitative values for each protein and peptide identified in the MS run. The user must then manually or preferably automatically insert the actual quantification values into this mzTab file. Once PRIDE Converter is supplied with an mzTab file containing such quantitative information, these data will automatically be reported in the generated PRIDE XML file.

A step-by-step description of how to generate quantitative mzTab file using the PRIDE mzTab Generator can be found in the PRIDE Converter 2 wiki: https://code.google.com/p/pride-converter-2/wiki/QuantitativeMzTabFiles.

### 3.2.3. Generation Report

The final screen of the PRIDE mzTab Generator (Figure 17) will show the path of the newly generated mzTab file corresponding to each input file. If a gel identifier and/or a spot identifier or spot regex have been provided, the relevant information is also included in the report. This is particularly useful when a gel spot regular expression has been used to confirm that the correct spot identifier has been extracted from the input file name.



MzTab Generation Complete!

Files Generated:

| Input File | MzTab File | Gel ID | SPOT ID |
|---|---|---|---|
| gel spots/Spot 100.dat | gel spots/Spot 100.dat-mz... | http://mydomain.com/myge... | 100 |
| gel spots/Spot 101.dat | gel spots/Spot 101.dat-mz... | http://mydomain.com/myge... | 101 |
| gel spots/Spot 102.dat | gel spots/Spot 102.dat-mz... | http://mydomain.com/myge... | 102 |
| gel spots/Spot 105.dat | gel spots/Spot 105.dat-mz... | http://mydomain.com/myge... | 105 |
| gel spots/Spot 106.dat | gel spots/Spot 106.dat-mz... | http://mydomain.com/myge... | 106 |
| gel spots/Spot 108.dat | gel spots/Spot 108.dat-mz... | http://mydomain.com/myge... | 108 |
| gel spots/Spot 109.dat | gel spots/Spot 109.dat-mz... | http://mydomain.com/myge... | 109 |

Conversion to mzTab done.

**Figure 17** – PRIDE mzTab Generator Generation Report form.

## 3.3. PRIDE XML Filter

### 3.3.1. Filter Information

The first screen of the PRIDE XML Filter provides an overview of the role of the tool and informs the user of several design assumptions and restrictions.

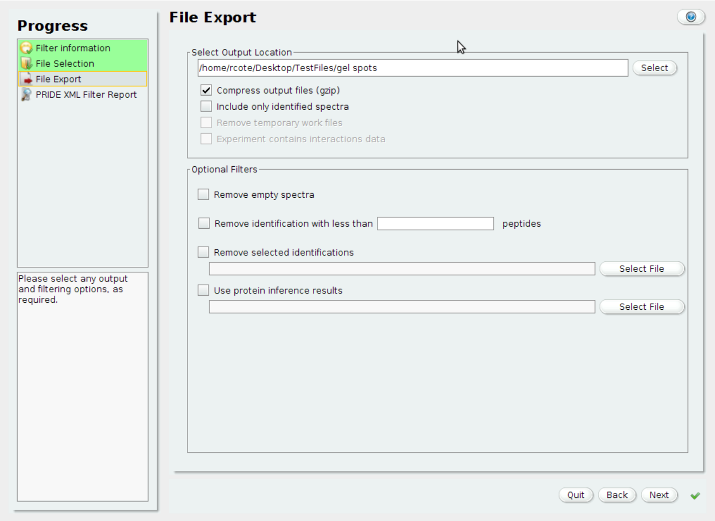
All the PRIDE XML files must be valid according to the PRIDE XML schema. Any number of filters can be selected in one go and all the filters will be applied to each selected input file.

### 3.3.2. File Selection

The look and usability of the file selection form for the PRIDE XML Filter is almost identical to that described in the “File Selection” section of the PRIDE Converter 2 GUI user guide, with the exception that all other file type selectors have been disabled and there are no configuration options to set. The PRIDE XML Filter can either take compressed (gzipped) or uncompressed PRIDE XML files as input, but all files need to be in the same state.

### 3.3.3. File Export

The look and usability of the file export form for the PRIDE XML Filter is almost identical to that described in the “File Export” section of the PRIDE Converter 2 GUI user guide, with the exception that some options have been disabled (Figure 18).



File Export

Select Output Location

/home/rcote/Desktop/TestFiles/gel spots

☑ Compress output files (gzip)
☐ Include only identified spectra
☐ Remove temporary work files
☐ Experiment contains interactions data

Optional Filters

☐ Remove empty spectra
☐ Remove identification with less than ______ peptides
☐ Remove selected identifications
☐ Use protein inference results

Please select any output and filtering options, as required.

**Figure 18** – PRIDE XML Filter File Export form.

The “include only identified spectra” filter is the only filter that does not need to generate an intermediate PRIDE XML file, as this can be done at the DAO level. All other filters will first write a complete PRIDE XML file, as normal, then post-process this file to generate the fully-filtered PRIDE XML file. This will have consequences on the time and disk space required to fully produce the final file.

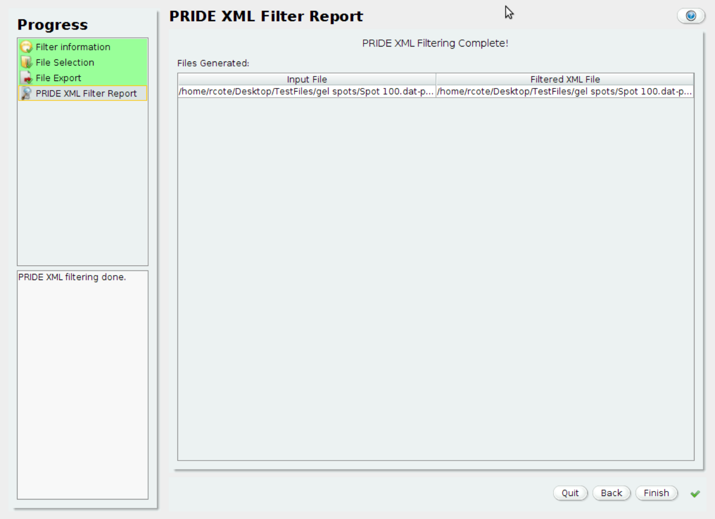
The “remove empty spectra” filter will remove any spectra that do not contain any peaks. The “remove identification with less than X peptides” filter will scan and remove any protein identification from the PRIDE XML file that contains less peptides than the number given as parameter.

The “remove selected identifications” filter requires a list of contaminant protein identifications in the form of a text file with one protein accession per line. It will use this blacklist to remove corresponding identifications from the XML file where the submitted protein accession identically matches one of the blacklisted protein identifiers.

The “use protein inference results” filter also requires a list of protein identifications in the same manner as the previous filter but works in the complete opposite fashion. Only identifications found in the whitelist are retained in the final PRIDE XML file.

### 3.3.4. Filter Report

The final screen of the PRIDE XML Filter (Figure 19) will show the path of the newly generated filtered PRIDE XML file corresponding to each input file.



PRIDE XML Filter Report

PRIDE XML Filtering Complete!

Files Generated:

| Input File | Filtered XML File |
|---|---|
| /home/rcote/Desktop/TestFiles/gel spots/Spot 100.dat-p... | /home/rcote/Desktop/TestFiles/gel spots/Spot 100.dat-p... |

PRIDE XML filtering done.

**Figure 19** – PRIDE XML Filter Report form.

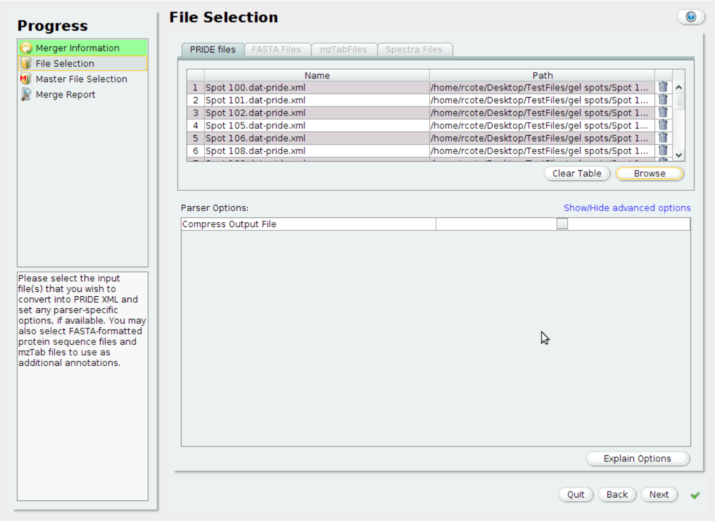
## 3.4. PRIDE XML Merger

### 3.4.1. Merger Information

The first screen of the PRIDE XML Merger provides an overview of the role of the tool and informs the user of several design assumptions and restrictions. All the PRIDE XML files must be valid according to the PRIDE XML schema. A master file will be selected, which will serve as the basis for the final PRIDE XML file in terms of metadata. The identifications and spectra from the other files will be merged into the final output file. Users should note that the spectrum IDs from the source files will not necessarily be kept, but the internal references between peptide-spectrum_refs will be maintained.

### 3.4.2. File Selection

The look and usability of the file selection form for the PRIDE XML Merger is almost identical to that described in the “File Selection” section of the PRIDE Converter 2 GUI user guide, with the exception that all other file type selectors have been disabled and the only DAO option available will be the option to compress the final output file (Figure 20). The PRIDE XML Merger can either take compressed (gzipped) or uncompressed PRIDE XML files as input, but all files need to be in the same state.



|  | Name | Path |
|---|---|---|
| 1 | Spot 100.dat-pride.xml | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 2 | Spot 101.dat-pride.xml | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 3 | Spot 102.dat-pride.xml | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 4 | Spot 105.dat-pride.xml | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 5 | Spot 106.dat-pride.xml | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |
| 6 | Spot 108.dat-pride.xml | /home/rcote/Desktop/TestFiles/gel spots/Spot 1... |

Parser Options:

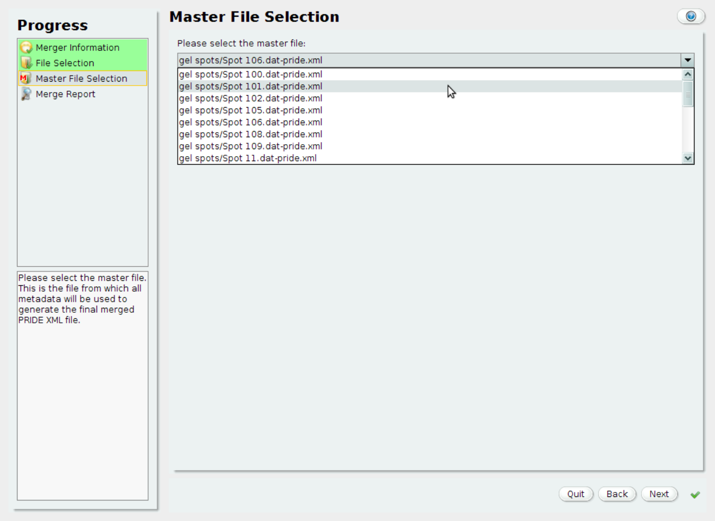
Compress Output File □

Please select the input file(s) that you wish to convert into PRIDE XML and set any parser-specific options, if available. You may also select FASTA-formatted protein sequence files and mzTab files to use as additional annotations.

**Figure 20** – PRIDE XML Merger File Selection form.

## 3.4.3. Master File Selection

A master file must be selected. All metadata annotations will be taken from this file to as a basis to generate the merged PRIDE XML file (Figure 21).



Please select the master file:

gel spots/Spot 106.dat-pride.xml

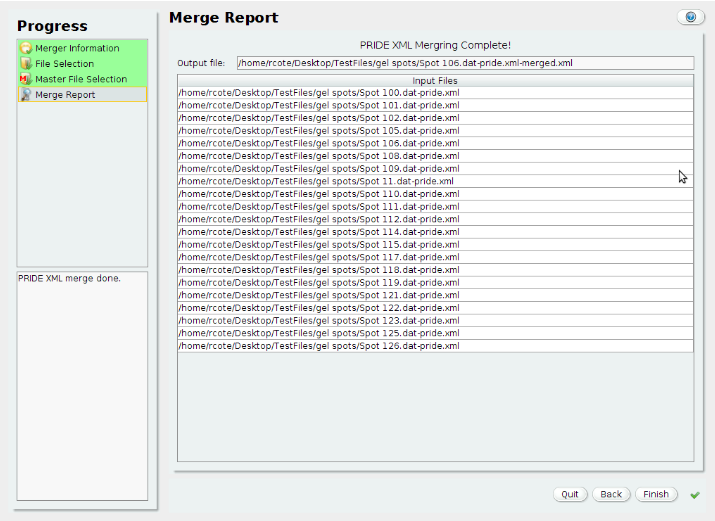
gel spots/Spot 100.dat-pride.xml
gel spots/Spot 101.dat-pride.xml
gel spots/Spot 102.dat-pride.xml
gel spots/Spot 105.dat-pride.xml
gel spots/Spot 106.dat-pride.xml
gel spots/Spot 108.dat-pride.xml
gel spots/Spot 109.dat-pride.xml
gel spots/Spot 11.dat-pride.xml

Please select the master file. This is the file from which all metadata will be used to generate the final merged PRIDE XML file.

**Figure 21** – PRIDE XML Merger Master File Selection form.

## 3.4.4. Merger Report

Once all PRIDE XML files have been merged, the user is shown a report form showing where the output file was written to and all the source PRIDE XML files which were used for the merge (Figure 22).



Merge Report

PRIDE XML Mergring Complete!

Output file: /home/rcote/Desktop/TestFiles/gel spots/Spot 106.dat-pride.xml-merged.xml

| Input Files |
|---|
| /home/rcote/Desktop/TestFiles/gel spots/Spot 100.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 101.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 102.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 105.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 106.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 108.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 109.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 11.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 110.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 111.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 112.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 114.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 115.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 117.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 118.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 119.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 121.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 122.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 123.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 125.dat-pride.xml |
| /home/rcote/Desktop/TestFiles/gel spots/Spot 126.dat-pride.xml |

PRIDE XML merge done.

**Figure 22** – PRIDE XML Merger Report form.

# 4. Troubleshooting

## 4.1. The Application Selector window appears but no tools start when a button is clicked

When *PRIDE Converter 2* starts in the GUI mode, the first screen that is displayed is the application selector (Figure 3). In rare occasions, it is possible that nothing happens when the users click on a button to start a specific application (*Converter*, *Filter*, *Merger*, *mzTab generator*). This is generally a problem caused with improper Java configuration. Please ensure that JAVA_HOME is correctly set in your environment and that the java runtime executable file is in your path. Please refer to the Java documentation on how to properly install and configure the Java Runtime Environment (http://java.com/en/download/help/download_options.xml).

When a button is clicked from the application selector, a new process is launched based on the settings found in the *converter.properties* file that is located in the same directory as the *PRIDE Converter 2* jar file.

```
    ###############################################################
    # Pride Converter GUI Bootstrap configuration
    ###############################################################

    #set the exact path of the java executable if not already in the path
    #do not include the executable itself and include a trailing slash
    #eg java.home=/home/rcote/dev/jdk1.6.0_13/bin/
    #and not #eg java.home=/home/rcote/dev/jdk1.6.0_13/bin/java
    java.home=

    # any JVM argument listed here will be passed verbatim to the JVM
    jvm.args=-Xms128M -Xmx4000M

    # uncomment and set accordingly to configure PROXY settings
    #http.proxyHost=webcache.mydomain.com
    #http.proxyPort=8080
    #http.proxyUser=
    #http.proxyPassword=
    #http.proxySet=true
```

If JAVA_HOME isn't properly set, the new process might not properly start. One way to circumvent this problem is to update the *java.home* setting in the properties file. You need to include the full path to the java executable, including the final slash, but do not include the java application itself.

```
    java.home=/home/rcote/dev/jdk1.6.0_13/bin/              - correct for linux and Mac OS X
    java.home=c:\Program Files\jdk1.6.0_13\bin\             - correct for Windows
```

If this still does not work correctly, try starting *PRIDE Converter 2* from the command line, as useful debugging information is logged to the console window.

```
    rcote@bobble: target$ java -jar pride-converter-2.0-SNAPSHOT.jar

    Reading properties file: ./converter.properties
    Bootstrap command: /home/rcote/dev/jdk1.7.0_03/jre/bin/java -Xms128M -Xmx4000M -cp pride-
converter-2.0-SNAPSHOT.jar
uk.ac.ebi.pride.tools.converter.gui.ConverterApplicationSelector$ConverterLauncher

    Name of the OS: Linux
    Version of the OS: 2.6.32-41-generic
    Architecture of The OS: amd64
    Validator Loaded
```

This information can be communicated with the PRIDE helpdesk to help fix the problem.

## 4.2. A progress window appears but the application seems to hang

If you are converting a large number of files or files that are very big or contain a large number of spectra, the conversion process can take anywhere from minutes to hours. This is normal and the time required will be in direct proportion to the number, size and complexity of the source files. The GUI will always show message dialogs if exceptions occur. If the application still seems to hang but no error message is shown, users should look in the log file that is in the *log* directory where the *PRIDE Converter 2* was installed.

## 4.3. The application runs out of memory

The *PRIDE Converter 2* tool suite was written to be as memory-efficient as possible. However, if converting a large number of big files, it is always possible to run out of memory. If this should happen, please update the *jvm.args* setting in the *converter.properties* file.

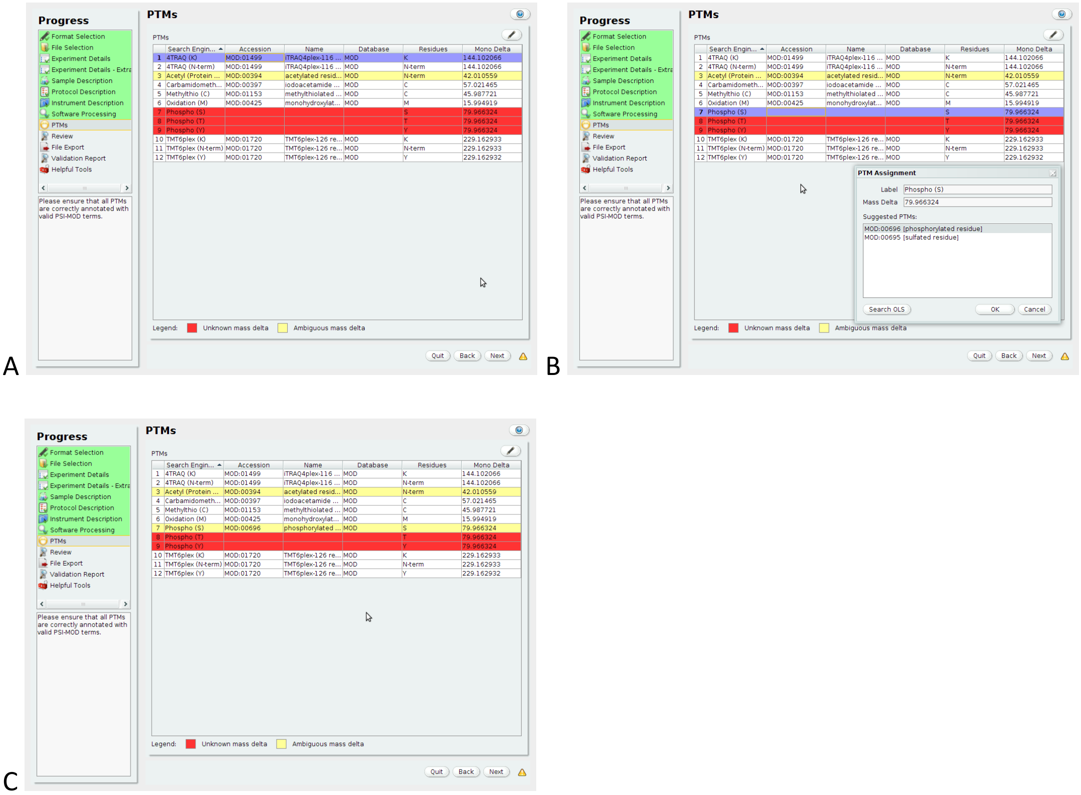
## 4.4. The application cannot use the OLS

The *PRIDE Converter 2* tool suite requires a working internet connection to connect to the OLS web service to perform CV term related queries. If you are on a computer that requires a proxy configuration to access external network resources, you will need to uncomment and update the five *http.proxy* settings in the *converter.properties* file.

# 5. *PRIDE Converter 2* Automatic PTM Assignment

*PRIDE Converter 2* attempts to assign the correct PTM annotation based on a curated list of the most commonly observed PTMs and the mass delta as reported by the search engine. Please refer to Table 2 below for the full list of automatically assigned PTMs. If a unique PTM can be assigned to a mass delta within a 0.1 Da mass tolerance, the annotation is automatically shown to the user (Figure 16).

In cases where multiple PTMs can be assigned to a mass delta with a precision of 0.1 Da, *PRIDE Converter 2* will try to locate a unique PTM assignment to within 0.01 Da. If a unique match is found at the higher precision threshold, it will be assigned but the GUI will report the fact that multiple PTMs have been observed. In the case that multiple PTMs are still found at the higher precision threshold, no mapping is done. The GUI will report the fact that multiple PTMs have been observed within the mass tolerance window by highlighting the conflict in yellow. PTMs that have not been automatically assigned will be highlighted in red  (Figure 16A).



Please ensure that all PTMs are correctly annotated with valid PSI-MOD terms.

PTMs

|  | Search Engin... | Accession | Name | Database | Residues | Mono Delta |
|---|---|---|---|---|---|---|
| 1 | 4TRAQ (K) | MOD:01499 | iTRAQ4plex-116 ... | MOD | K | 144.102066 |
| 2 | 4TRAQ (N-term) | MOD:01499 | iTRAQ4plex-116 ... | MOD | N-term | 144.102066 |
| 3 | Acetyl (Protein ... | MOD:00394 | acetylated resid... | MOD | N-term | 42.010559 |
| 4 | Carbamidometh... | MOD:00397 | iodoacetamide ... | MOD | C | 57.021465 |
| 5 | Methylthio (C) | MOD:01153 | methylthiolated ... | MOD | C | 45.987721 |
| 6 | Oxidation (M) | MOD:00425 | monohydroxylat... | MOD | M | 15.994919 |
| 7 | Phospho (S) |  |  |  | S | 79.966324 |
| 8 | Phospho (T) |  |  |  | T | 79.966324 |
| 9 | Phospho (Y) |  |  |  | Y | 79.966324 |
| 10 | TMT6plex (K) | MOD:01720 | TMT6plex-126 re... | MOD | K | 229.162933 |
| 11 | TMT6plex (N-term) | MOD:01720 | TMT6plex-126 re... | MOD | N-term | 229.162933 |
| 12 | TMT6plex (Y) | MOD:01720 | TMT6plex-126 re... | MOD | Y | 229.162932 |

Legend: 🟥 Unknown mass delta 🟨 Ambiguous mass delta

PTM Assignment

Label: Phospho (S)

Mass Delta: 79.966324

Suggested PTMs:

MOD:00696 [phosphorylated residue]
MOD:00695 [sulfated residue]

Search OLS   OK   Cancel

|  | Search Engin... | Accession | Name | Database | Residues | Mono Delta |
|---|---|---|---|---|---|---|
| 1 | 4TRAQ (K) | MOD:01499 | iTRAQ4plex-116 ... | MOD | K | 144.102066 |
| 2 | 4TRAQ (N-term) | MOD:01499 | iTRAQ4plex-116 ... | MOD | N-term | 144.102066 |
| 3 | Acetyl (Protein ... | MOD:00394 | acetylated resid... | MOD | N-term | 42.010559 |
| 4 | Carbamidometh... | MOD:00397 | iodoacetamide ... | MOD | C | 57.021465 |
| 5 | Methylthio (C) | MOD:01153 | methylthiolated ... | MOD | C | 45.987721 |
| 6 | Oxidation (M) | MOD:00425 | monohydroxylat... | MOD | M | 15.994919 |
| 7 | Phospho (S) | MOD:00696 | phosphorylated ... | MOD | S | 79.966324 |
| 8 | Phospho (T) |  |  |  | T | 79.966324 |
| 9 | Phospho (Y) |  |  |  | Y | 79.966324 |
| 10 | TMT6plex (K) | MOD:01720 | TMT6plex-126 re... | MOD | K | 229.162933 |
| 11 | TMT6plex (N-term) | MOD:01720 | TMT6plex-126 re... | MOD | N-term | 229.162933 |
| 12 | TMT6plex (Y) | MOD:01720 | TMT6plex-126 re... | MOD | Y | 229.162932 |

**Figure 16.** Display of automatically assigned PTMs in the *PRIDE Converter* 2 GUI. Rows with a white background have been unambiguously assigned. Rows in yellow have been assigned, but there is ambiguity in a 0.1 Da window around the reported mass delta. Rows in red have could not be assigned.

By double-clicking on a row highlighted in red, the user will be presented with a window showing the preferred PTMs (those that map to a mass delta with a precision of 0.1 Da), if any. The user has the choice to either select a preferred PTM, if applicable, or to search the PSI-MOD ontology via the OLS (Figure 16B). See Table 2 to see how the different mass deltas are mapped to the PSI-MOD ontology.

**Table 2**. List of PTMs that *PRIDE Converter 2* will attempt to automatically assign, based on the mass delta reported in the PSI-MOD ontology.

| PSI-MOD Accession | Unimod ID | PSI-MOD Description | Mass Delta (Da) |
|---|---|---|---|
| MOD:00394 | 1 | acetylated residue | 42.01057 |
| MOD:00674 | 2 | amidated residue | -0.98402 |
| MOD:01885 | 3 | biotinylated residue | 226.07760 |
| MOD:00696 | 21 | phosphorylated residue | 79.96633 |
| MOD:00397 | 4 | iodoacetamide derivatized residue | 57.02146 |
| MOD:00398 | 5 | carbamoylated residue | 43.00581 |
| MOD:00399 | 6 | iodoacetic acid derivatized residue | 58.00548 |
| MOD:00400 | 7 | deamidated residue | 0.98402 |
| MOD:00403 | 10 | homoserine | -29.99281 |
| MOD:00404 | 11 | homoserine lactone | -48.00337 |
| MOD:00410 | 17 | S-(N-isopropylcarboxamidomethyl)-L-cysteine | 99.06841 |
| MOD:00704 | 23 | dehydrated residue | -18.01057 |
| MOD:00417 | 24 | S-carboxamidoethyl-L-cysteine | 71.03711 |
| MOD:00419 | 26 | (R)-5-oxo-1,4-tetrahydrothiazine-3-carboxylic acid | 39.99492 |
| MOD:00425 | 35 | monohydroxylated residue | 15.99492 |
| MOD:01160 | 28 | deaminated residue | -17.02655 |
| MOD:01813 | 29 | morpholine-2-acetylated residue | 127.06333 |
| MOD:00423 | 30 | monosodium salt | 21.98194 |
| MOD:00424 | 31 | S-pyridylethyl-L-cysteine | 105.05785 |
| MOD:00599 | 34 | monomethylated residue | 14.01565 |
| MOD:01153 | 39 | methylthiolated residue | 45.98772 |
| MOD:00695 | 40 | sulfated residue | 79.95682 |
| MOD:00127 | 42 | N6-lipoyl-L-lysine | 188.03296 |
| MOD:00437 | 44 | farnesylated residue | 204.18780 |
| MOD:00438 | 45 | myristoylated residue | 210.19837 |
| MOD:00128 | 46 | N6-pyridoxal phosphate-L-lysine | 229.01401 |
| MOD:00440 | 47 | palmitoylated residue | 238.22967 |
| MOD:00441 | 48 | geranylgeranylated residue | 272.25040 |
| MOD:00159 | 49 | O-phosphopantetheine-L-serine | 340.08579 |
| MOD:00697 | 50 | flavin modified residue | 783.14149 |
| MOD:00445 | 52 | L-homoarginine | 42.02180 |
| MOD:00493 | 122 | formylated residue | 27.99492 |
| MOD:00405 | 12 | Applied Biosystems original ICAT(TM) d8 modified cysteine | 450.27521 |
| MOD:00406 | 13 | Applied Biosystems original ICAT(TM) d0 modified cysteine | 442.22499 |
| MOD:00480 | 105 | Applied Biosystems cleavable ICAT(TM) light | 227.12699 |
| MOD:00481 | 106 | Applied Biosystems cleavable ICAT(TM) heavy | 236.15719 |
| MOD:00544 | 188 | 6x(13)C labeled residue | 6.02013 |

| | | | |
|---|---|---|---|
| MOD:00546 | 193 | (18)O label at both C-terminal oxygens | 4.00849 |
| MOD:01152 | 299 | carboxylated residue | 43.98983 |
| MOD:00428 | 425 | dihydroxylated residue | 31.98983 |
| MOD:01499 | 214 | iTRAQ4plex-116 reporter+balance reagent acylated residue | 144.10206 |
| MOD:01549 | 730 | iTRAQ8plex-116 reporter+balance reagent acylated residue | 304.20536 |
| MOD:01720 | 737 | TMT6plex-126 reporter+balance reagent acylated residue | 229.16293 |