ProCon – Proteomics Conversion Tool

1 Introduction

With ProCon you can convert proteomics identification and quantification results into HUPO standard formats [1], which can be used to upload your results into public repositories.

2 Installation

Install Java SE (JRE): ProCon was tested with JRE 8, Update 66, 64 Bit, but should work with all former JRE 7 and newer JRE 8 versions, e.g. from the actual download version here: http://www.oracle.com/technetwork/java/javase/downloads/index.html.

! Installation of a new ProCon release: !

If you have already used ProCon and install a new release be sure to “save” any changes made to .config files (e.g. added modifications, institute address or server information).

Then unzip **ProCon_dist**<*version*>**.zip** into an arbitrary directory, e.g. **C:\ProCon**. This is the so-called working directory, from which you start ProCon later (see below).

Verify (and potentially change) the configuration files (see next section and configuration subsections in “Sequest .out Import ...” and “ProteinScape® 1.3 Import ...” sections).

The ProteomeDiscoverer version now supports both an access to the SQLite database (.msf file) via JDBC and via a native interface (sqlite4java) written in C++. It's recommended to use the JDBC version, which uses the driver from https://bitbucket.org/xerial/sqlite-jdbc.

If one wants to use the native interface instead, one first must install the operating system specific files for sqlite4java from the website http://code.google.com/p/sqlite4java/ .

For Windows® operating systems copy the DLL's sqlite4java-win32-x64.dll resp. sqlite4java-win32-x86.dll into the folder C:\Windows\SysWOW64 resp. C:\Windows\System32. Also sqlite3.dll available from http://www.sqlite.org/ should be copied into the proper system directory (C:\Windows\System32). (Note that at moment only a sqlite3.dll for 32 Bit is available). Then in the batch file start_ProCon.bat set the JAVA_HOME variable to point to your Java Runtime Environment (JRE).

For Mac OS X® resp. Linux operating systems one should install the files *.dylib resp. *.so from the sqlite4java distribution into the respective system directories.

For starting ProCon you should first set the path to your Java runtime installation in the batch file start_ProCon_from_GUI.bat and then right click on this batch file and run it.

```
@echo on
set JAVA_HOME="C:\Program Files\Java\jre1.8.0_66\bin\java.exe"
%JAVA_HOME% -jar ProCon.jar
pause
```

3 General Configuration

File ./config/ProCon.properties

Specify the details of the person(s) responsible for the mass spectrometry of the sample measured and the data set (probably you) (i.e. “MassSpecContactName=”, “MassSpecInstitution=”, “MassSpecEMailPhoneFax=”, “DataSetContactName=”, “DataSetInstitution=”, “DataSetEMailPhoneFax=”; fill-in directly after the “=”). These details are for example used in the “ProteinScape® 1.3 to PRIDE XML” conversion.

File ./config/log4j.properties:

Lines **log4j.appender.ProCon_file.File** and **log4j.appender.pride_core_file.File** should contain \\ as path separator for Windows and / for Unix-based operating systems.

4 Start ProCon

Start the ProCon.jar by double-clicking it or by giving the command **java –jar ProCon.jar** (ensure, that it starts in the working directory, where you unzipped it; more exactly the directory with subfolders **/config, /instruments, /XML_files**, ...). ProCon now runs with both the Java 7 and Java 8 runtime.
The example batch file start_ProCon.bat shows how one can set Java Runtime options.

Any firewall message or any question of a protection tool about “an application start” should be confirmed / allowed. In case ProCon does not start check the contents of the log file in the log folder: .\log\ProCon.txt.

More detailed information about errors can be found in the log files in the **./log** folder (**ProCon.log** for ProCon, ProteinScape® and Sequest classes and **pride.log** for pride classes).

There are four workflows you can follow in the current version:

1) Import of Sequest .out files and export to mzIdentML [2]
2) ProteinScape® 1.3 import and PRIDE XML export
3) Proteome Discoverer® 1.1, 1.2 and 1.3 to mzIdentML 1.1 conversion,
4) ProteinScape® 2.1 to mzIdentML 1.1 conversion.
5) Spectral counts to mzQuantML [3] conversion.

5 Workflow 1: Sequest.out / comet.out Import and mzIdentML export

Functionality has been added for import of a Sequest out folder (one search engine run) and export of this Sequest data set to mzIdentML. Use the tab **Sequest.out / comet.out to mzIdentML** for export.

Sequest import was tested with Bioworks Sequest (version 3.2).

Comet import was tested with Comet (version 2015.01 rev. 1)

Because this implementation is important to establish the mzIdentML standard, please report all errors and suggestions to the ProCon developers specified on http://www.medizinisches-proteom-center.de/software.

5.1 Sequest-specific Configuration

The SEQUEST converter converts the .out files from an arbitrary folder into the mzIdentML format. Besides the .out (and .dta files) the folder must contain the following 2 (for comet.out) resp. 4 (for SEQUEST.out) files:

- **Header.txt** only for SEQUEST.out; not required for comet.out conversions
- **sequest.log** only for SEQUEST.out; not required for comet.out conversions
- **sequest.params** resp. **comet.params**
- **<folder-name>.log**, e.g. if your folder is named 'test_folder' then the file should be named 'test_folder.log'

If you don't find this information in your SEQUEST folder containing the .out files, then you have to create the content of these files by your own using any text editor according to the following description:

a) Header.txt: (only for SEQUEST.out; not required for comet.out conversions)

Sample

LastName

e.g.

LastName:Joppich

Sample:PMXPWE080620_38

b) sequest.log: (only for SEQUEST.out; not required for comet.out conversions)

TurboSEQUEST - xxxxxxxxxxxxxx ... // (xxxxxxxxxxxxxx = SEQUEST version)

e.g.
TurboSEQUEST - PVM Master v.27 (rev. 12), (c) 1998-2007

c) sequest.params: (in case of SEQUEST.out conversion)
diff_search_options
term_diff_search_options
database_name
first_database_name
second_database_name
mass_type_parent
mass_type_fragment
max_num_internal_cleavage_sites
peptide_mass_tolerance
peptide_mass_units
fragment_ion_tolerance
enzyme_info

e.g.
diff_search_options = 15.9949 M 57.0 C 0.000 X 0.000 X 0.000 X 0.000 X
term_diff_search_options = 0.0000 0.0000
database_name = D:/Database/StdCry.fasta
first_database_name = D:/Database/StdCry.fasta
second_database_name = mass_type_parent = 0 // 0=average masses, 1=monoisotopic masses
mass_type_fragment = 1 // 0=average masses, 1=monoisotopic masses
max_num_internal_cleavage_sites = 5 // maximum value is 5
peptide_mass_tolerance = 1.5000
peptide_mass_units = 0 // 0=amu, 1=mmu, 2=ppm
fragment_ion_tolerance = 1.5000 // width in amu of bins for fragment ions
enzyme_info = Trypsin 1 1 KR –

c) comet.params: (in case of comet.out conversion)
e.g.
…
search_enzyme_number = 1 # choose from list at end of this params file

num_enzyme_termini = 2 # valid values are 1 (semi-digested), 2 (fully digested, default), 8 N-term, 9 C-term

…

fragment_bin_tol = 1.0005 # binning to use on fragment ions

…

output_outfiles = 1 # 0=no, 1=yes write .out files

…

d) <folder-name>.log:

Sequest queued xxxxxxxxxxxxxxxxxxxxxxxxx ...

// (xxxxxxxxxxxxxxxxxxxxxxxxx = activity date in the format "EEE MMM dd kk:mm:ss: yyyy")

e.g.

Sequest queued Tue Jun 24 11:32:27 2008 StdCry_nr.fasta Trypsin 15.99491 M 57.0000 C 0.0000 X 0.0000 X 0.0000 X 0.0000 X mods 0.0000 0.0000 cj

4.1.1 File ./config/SEQUEST_massvalues.txt

This config file contains the mass values Sequest uses. Be sure to use the mass value file of your Sequest installation! (Server path e.g. **C:\Inetpub\etc\config**.)

4.1.2 File ./config/SEQUESTMOD.obo

Sequest uses only masses for modifications. In the Sequest .obo file a mapping between these masses (added to an amino acid) have to be mapped to UNIMOD modifications (used in mzIdentML).

If you used a “modification mass / amino acid” not specified, an error occurs during export. The edit the .obo file and add this new combination in the following form:

Example:

Oxidation (here with mass 15.9949) of Methionine.

```
[Term]
id: SEQMOD:00002
name: M+15.9949
def: "Oxidation of Methionine" [UNIMOD:UNIMOD\:35]
```

```
is_a: SEQMOD:00001 ! Modification
```

Be aware, that in Sequest fixed and variable modifications are specified separately and can therefore have different masses (e.g. different number of decimals).

4.1.3 File ./config/unimod.obo

This is just the unimod.obo file from http://www.unimod.org/obo/unimod.obo. The unimod.obo file coming with ProCon should be sufficient for most cases. Overwrite with the latest version (date stamp inside the file) to be up-to-date.

4.1.4 File ./config/Sequest.properties

The mzIdentML output contains a globally unique Sequest server URI to specify the location of some files (e.g. the search database file) This URI is not necessarily a browsable web address!

In the **Sequest.properties** file specify the URL part and the name of your Sequest server.

Example:

```
URISequestServerURL=www.medizinisches-proteom-center.de
URISequestServerName=sequestmaster
```

This will lead to the following URI for the search database in the mzIdentML file:

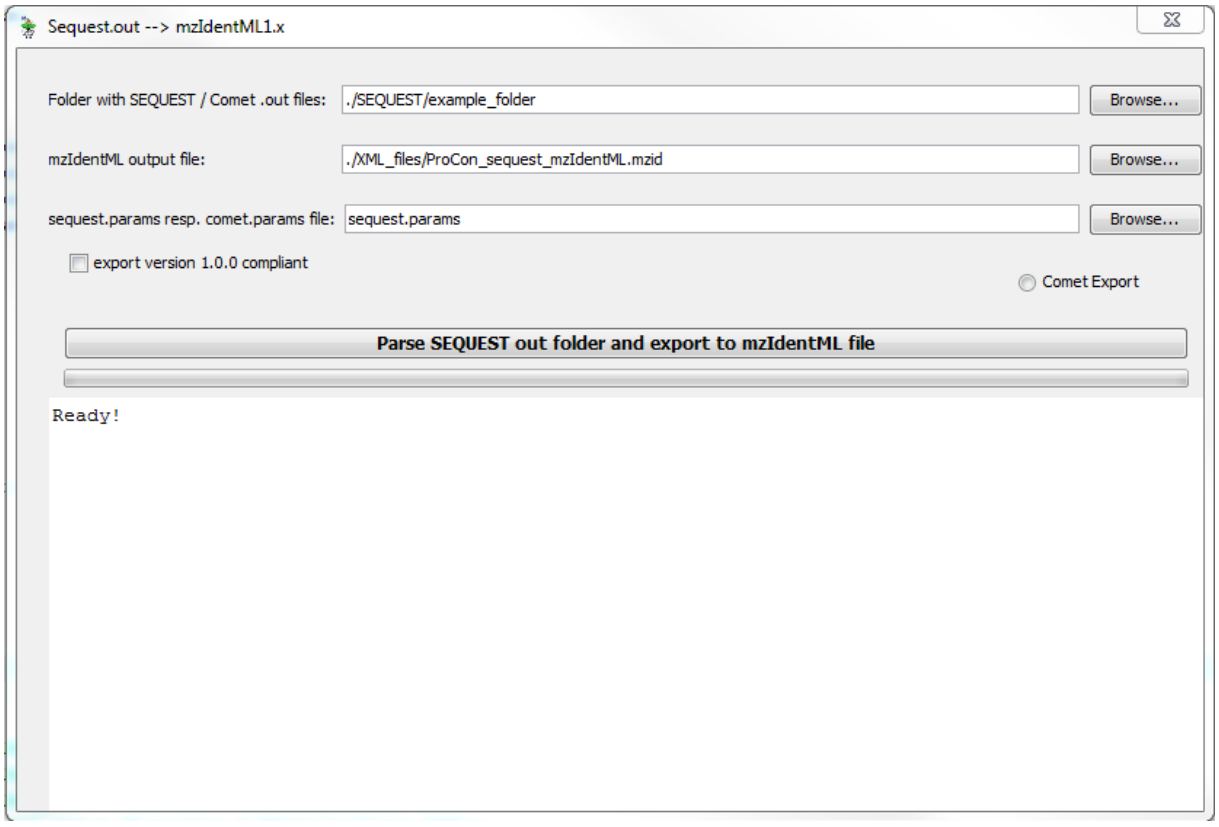
```
file://www.medizinisches-proteom-
center.de/sequestmaster/work/Datenbank/StdCry_nr.fasta
```

4.1.5 File ./config/mzidAuditCollection_1.1.xml

For export of mzIdentML a Provider contact role (e.g. “researcher”) and the Provider’s organization (e.g. institute within a university) should be specified. The **mzidAuditCollection_1.1.xml** .config file contains this information using a certain structure (similar to FuGE). Please overwrite with your details!

You should not change the sequence of comment and content lines. If you accidently destroy the file, copy over with **mzidAuditCollectionTemplate_1.1.xml**.

If you use the “Sequest .out to mzIdentML_1.0” export (deprecated!), use the **mzidAuditCollection_1.0.xml** file analogously.


Sequest.out --> mzIdentML1.x
Folder with SEQUEST / Comet .out files:
./SEQUEST/example_folder
Browse...
mzIdentML output file:
./XML_files/ProCon_sequest_mzIdentML.mzid
Browse...
sequest.params resp. comet.params file:
sequest.params
Browse...
export version 1.0.0 compliant
Comet Export
Parse SEQUEST out folder and export to mzIdentML file
Ready!

6 Workflow 2: ProteinScape® 1.3 Import and PRIDE XML export

6.1 Prerequisites

Connection to your local ProteinScape® database

ProCon connects to the SQL database of your PS 1.3 installation. The SQL database must be configured to allow connections via TCP/IP (default port 1433).

A connection string is implemented like:

`jdbc:jtds:sqlserver://`<*IP_address_of_server*>`:1433/ProteinScape1_0;`
`user=`<*username*>`;password=`<*password*>

Therefore you must know:

- the **IP address** of the ProteinScape® server (where the SQL database is normally installed, too)
- the **port of the SQL** database (default 1433 for Microsoft SQL Server)
- **no firewall** should prevent the communication between the computer where ProCon runs and the database server
- the **database name** (default:ProteinScape1_0)
- the **SQL user name** (we can use sa because ProCon does not CHANGE anything, but you may create another user having only read permissions)
- the **SQL password** for this database user

Data for First Test

ProCon was tested for 2D gels and 1D-LC, both PMF (MS) or PFF (MS/MS), either protein assembly by search engine or by ProteinExtractor. It was tested with Mascot, Sequest and Phenyx runs.

A best (because simplest) first test would be a LC/MS/MS run performed using one search engine.

Connection to Ontology Lookup Service

During import of ProteinScape® experiments, the taxonomy ID is queried online using the Ontology Lookup Service at the European Bioinformatics Institute (http://www.ebi.ac.uk/ontology-lookup/services/OntologyQuery). That works correctly only,

if ProCon (as Java application) can establish an online connection at runtime. You should configure firewall rules appropriately or answer firewall questions with Yes.

6.2 ProteinScape®-specific Configuration

File ./config/PAG-PS.obo

In this file the mapping from ProteinScape® modifications to PSI-MOD is configured. As it is "name-based" there may exist differences in your ProteinScape® installation. This is most probable for modifications you added yourself (e.g. "Cy3" differs from "Cy3 (C)"!)

You should check your commonly used modifications (names and at least one cross-reference to PSI-MOD) before using ProCon the first time. ProCon looks for "names", so you must be quite exact considering each space and bracket.

Whenever an unknown modification is encountered during import, ProCon aborts and displays an info message to correct the .obo file.

File ProteinScape.properties

The property ENDIAN_TYPE allows specification of base64 encoding of mzdata binary arrays ("little" or "big"). Precision is fixed to 64 (double).

Only necessary, if Mascot was used in the analyses, that are to be exported:

"SATParameterType" should be the ParameterType in the SearchAlgorithmTranslations table for instrument mappings (default: 8).

You can check, whether 8 is okay for you, if the SQL query:

```
select distinct AlgorithmName
from SearchAlgorithmTranslations
where ParameterType=8
```

on the ProteinScape® database results in some Mascot instruments like:

```
ESI-FTICR
ESI-QUAD-TOF
ESI-TRAP
ESI-TRAP,ETD-TRAP,ESI-TRAP
MALDI-QTOF
MALDI-TOF-TOF
```

6.3 ProteinScape® Data Generation

In order to obtain concise and complete result files, you should follow some guidelines in ProteinScape® data generation:

- Fill in all fields for to describe project, sample, separation and spot/band (nearly all fields are exported, if not as CVParam [4,5], then as userParam);
    - AVOID empty fields or "default" or "not specified" fields;
    - wrong descriptions go non-validated into the exported XML!
- Use one instrument type for each imported spectrum (or each spectrum package, called "combined spectrum"); otherwise the results cannot be exported into the same <Experiment> element, but have to go into separate data sets.
- Use one SearchMethod for all identification runs of a gel (not only the same name, but really the same method having the same SearchMethodID); otherwise the results cannot be exported into the same <Experiment> element, but have to go into separate data sets.

6.4 Testing / Initializing the Database Connection

A default database connection string is given on the **ProteinScape® Source** tab (which can be anytime restored by clicking the **Reset DB string** button). Provide the correct information for your server as described in section 1 (see above).

Example:

```
jdbc:jtds:sqlserver://134.147.123.124:1433/ProteinScape1_0;user=iu
ser;password=iuser
```

Before you can import ProteinScape® data, you have to click the **Initialize DB Connection:** button. Please be patient, this can take some time! ProCon tries to connect to the ProteinScape® server and database with the specified account information.

If an error occurs, the error/exception text is printed out. Check the connection string and try **Initialize DB Connection:** again.

If no error occurs, you will find three project names of your ProteinScape® server in the text area next to the button. Only if you see these project names, the connection is working!

Otherwise contact the ProCon developers specified on http://www.medizinisches-proteom-center.de/ProCon.

6.5 Converting a Search Event

- Specify a SearchEventID on the **ProteinScape® Source** tab and click the **Import ProteinScape® SearchEvent** button. ProCon imports the proteins marked green, their peptides (with modifications) and the spectra in which those have been identified. During the import ProCon asks you to specify any missing information (see section "Missing Information" below). Please be patient, the import may take some time (scroll-down the Outputs text area for latest progress messages)! Further ProteinScape® data sets can be imported, or the current imports can be cleared.
- Then on the **PRIDE XML** tab click the **Assemble PRIDE XML** button and a PRIDE data set is assembled internally (subsequent imports can be added to this assembly, or the current assembly can be cleared).
- Finally click **Export to PRIDE XML file** and the current data sets currently in the PRIDE assembly are exported to the PRIDE XML file specified in the text field of this tab.

6.6 Converting Gel Data

- On the **ProteinScapeSource** tab there is a ComboBox containing all separations of your server (entries are structured "<project> | <sample> | <gel> (<GelID>)", long names are truncated, GelID is unique!). Specify a separation and click the **Import ProteinScape Separation** button. ProCon then considers all spots, spectra and searches of this separation and exports the proteins marked green, their peptides (with modifications) and the spectra in which those have been identified. Dependent on the selection status of the PMF / PFF check boxes, PMF and/or PFF identifications are exported (leading to 1 or 2 ProteinScape® imports). During the import ProCon asks you to specify any missing information (see section "Missing Information" below).
- PRIDE can only describe one protocol per data set (i.e. per <Experiment> element). If the SearchEvents of the specified gel have been run with different SearchMethods, ProCon asks you to select one. Only SearchEvents done with this SearchMethod are then imported. ATTENTION: In ProteinScape®, if you modify a SearchMethod and run a SearchEvent without saving the method changes, it is not stored but the SearchEvent is named "*origSearchMethod*(modified)" per default. If you anyhow store and then select such a SearchMethod (with the "(modified)" postfix), ProCon will export only last SearchEvent, although there may be more SearchEvents using the same (default) name. RECOMMENDATION: You should optimize a SearchMethod

for your gel, then store it and run all SearchEvents for a gel you want to export with the same stored SearchMethod.

- Please be patient, the import may take some time (scroll-down the Outputs text area for latest progress messages)!
- Then click the **Assemble PRIDE XML** button and a PRIDE data set is assembled internally (containing 1 or 2 experiments in PRIDE assembly, depending on PMF and/or PFF identifications in the gel).
- Finally click **Export to File** and the current PRIDE data set is exported to the file specified in the text field of this tab.

Subsequently imported ProteinScape® imports are added to the internally assembled PRIDE data set and can be flushed out together using **Export to File**.

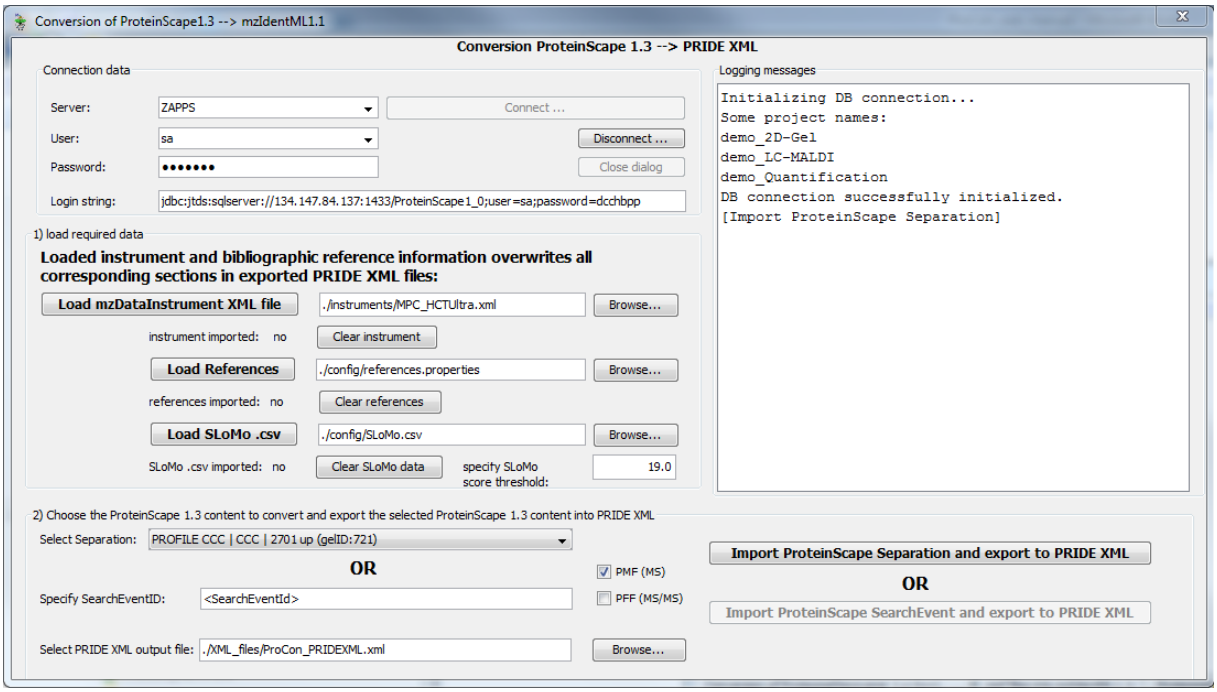
6.7 Instrument information and References

On the **General Source** tab, instrument details and references can be imported. During PRIDE XML export, the respective sections (mzData/instrument and <Reference> elements) are overwritten / filled with the imported information in all data sets of the assembly.

6.8 Missing Information

Depending on the information or type of results you want to export, ProteinScape® asks you to specify missing information:

- **Dig_before_Sep**: Specify, whether the digestion step was done before the separation step (default in LC protocols) or whether separation was done before digestion (normal in gel protocols).
- **Database version**: Specify the version of the sequence database (e.g. 3.41 for IPI database); this text should not be too long (<10 characters).
- **Protocol Name**: You can specify an "overall" name for the protocol you performed and described in ProteinScape® (ProteinScape® has no possibility for that, but PRIDE has).
- **Instrument Software Name, Version, Comments**: In three dialogs you should specify the details of the instrument software (not search engine!), which was used for spectrum generation.


Conversion of ProteinScape1.3 --> mzIdentML1.1
Conversion ProteinScape 1.3 --> PRIDE XML
Connection data
Server:
ZAPPS
Connect ...
User:
sa
Disconnect ...
Password:
Close dialog
Login string:
jdbc:jtds:sqlserver://134.147.84.137:1433/ProteinScape1_0;user=sa;password=dcchbpp
Logging messages
Initializing DB connection...
Some project names:
demo_2D-Gel
demo_LC-MALDI
demo_Quantification
DB connection successfully initialized.
[Import ProteinScape Separation]
1) load required data
Loaded instrument and bibliographic reference information overwrites all corresponding sections in exported PRIDE XML files:
Load mzDataInstrument XML file
./instruments/MPC_HCTUltra.xml
Browse...
instrument imported: no
Clear instrument
Load References
./config/references.properties
Browse...
references imported: no
Clear references
Load SLoMo .csv
./config/SLoMo.csv
Browse...
SLoMo .csv imported: no
Clear SLoMo data
specify SLoMo score threshold:
19.0
2) Choose the ProteinScape 1.3 content to convert and export the selected ProteinScape 1.3 content into PRIDE XML
Select Separation:
PROFILE CCC | CCC | 2701 up (gelID:721)
OR
PMF (MS)
PFF (MS/MS)
Specify SearchEventID:
<SearchEventId>
Import ProteinScape Separation and export to PRIDE XML
OR
Import ProteinScape SearchEvent and export to PRIDE XML
Select PRIDE XML output file:
./XML_files/ProCon_PRIDEXML.xml
Browse...

7 Workflow 3: Proteome Discoverer® to mzIdentML conversion

The output of the Mascot, SEQUEST, ZCore and MS Amanda [6] search engines of Proteome Discoverer® can be converted into the standard mzIdentML1.1 format. For the conversion of ProteomeDiscoverer® 1.3 and 1.4 results, only the *.msf (Mass Spec Format, Thermo) file is needed. For ProteomeDiscoverer® 1.1 and 1.2 results conversion in addition the *.prot.xml file must be specified: in this case the information (spectra data) missing in the ProtXML output are combined with the data from the *.msf file. This is done by matching the peptide sequences to the proteins in which they are found.

Note that version 2.0 of ProteomeDiscoverer® is not yet supported.

The tab for the ProteomeDiscoverer® output conversion consists of 2 parts:

1. Entering the parameters for the conversion process
2. Starting the conversion process into the mzIdentML1.1 standard format

In the dialog box for entering the conversion parameters (Figure 1) one can choose the **input files** (the *.msf and for ProteomeDiscoverer® 1.1 and 1.2 also the *.prot.xml output files) of the ProteomeDiscoverer® output. After choosing one of them the name of the other one and a name for the *.mzid output file are proposed, but one can also change the proposed file names if needed.

In the panel "**Organization data**" one can enter the name and contact details. If one clicks the checkbox "**Use MPC data**" then these contact fields are filled in with the data of the MPC (Medical Proteome Center) in Bochum.

In the panel "**Conversion parameters**" one can choose if the theoretical m/z values and isoelectric points for the peptide sequences should be calculated.

If the checkbox "**Export the ProteinDetectionList**" is deselected no protein inference information is exported. This can be used if one intends to use one's one protein inference algorithm, as e.g. the Protein Inference Algorithms PIA [7] (http://www.ruhr-uni-bochum.de/mpc/software/PIA/index.html.en).

The checkbox "**Report ProCon**" determines, if the converter ProCon is listed as AnalysisSoftware in the generated .mzid file.

The "**Peptide filter criteria**" can be set to ALL, RELAXED or STRICT and filters according the peptide scores and the thresholds set in the ProteomeDiscoverer® workflow. The following threshold values are used:

- **STRICT**
  - for XCorr (SEQUEST): CutOffStdCharge1High, e.g. 1.35
    CutOffStdCharge2High, e.g. 2.455
    CutOffStdCharge3High, e.g. 2.87
    CutOffStdCharge4High, e.g. 2.875
  - for IonScore (Mascot): DefaultStrictScoreThreshold, e.g. 2.3
  - for AmandaScore (MS Amanda): AmandaScoreHighConfidenceThreshold, e.g. 120.0
- **RELAXED**
  - for XCorr (SEQUEST): CutOffStdCharge1Middle, e.g. 1.35
    CutOffStdCharge2Middle, e.g. 2.25
    CutOffStdCharge3Middle, e.g. 2.87
    CutOffStdCharge4Middle, e.g. 2.875
  - for IonScore (Mascot): DefaultRelaxedScoreThreshold, e.g. 1.5
  - for AmandaScore (MS Amanda): AmandaScoreMiddleConfidenceThreshold, e.g. 90.0
- **ALL**
  Here all score thresholds are set to 0.0, so that no filtering using peptide scores takes place.

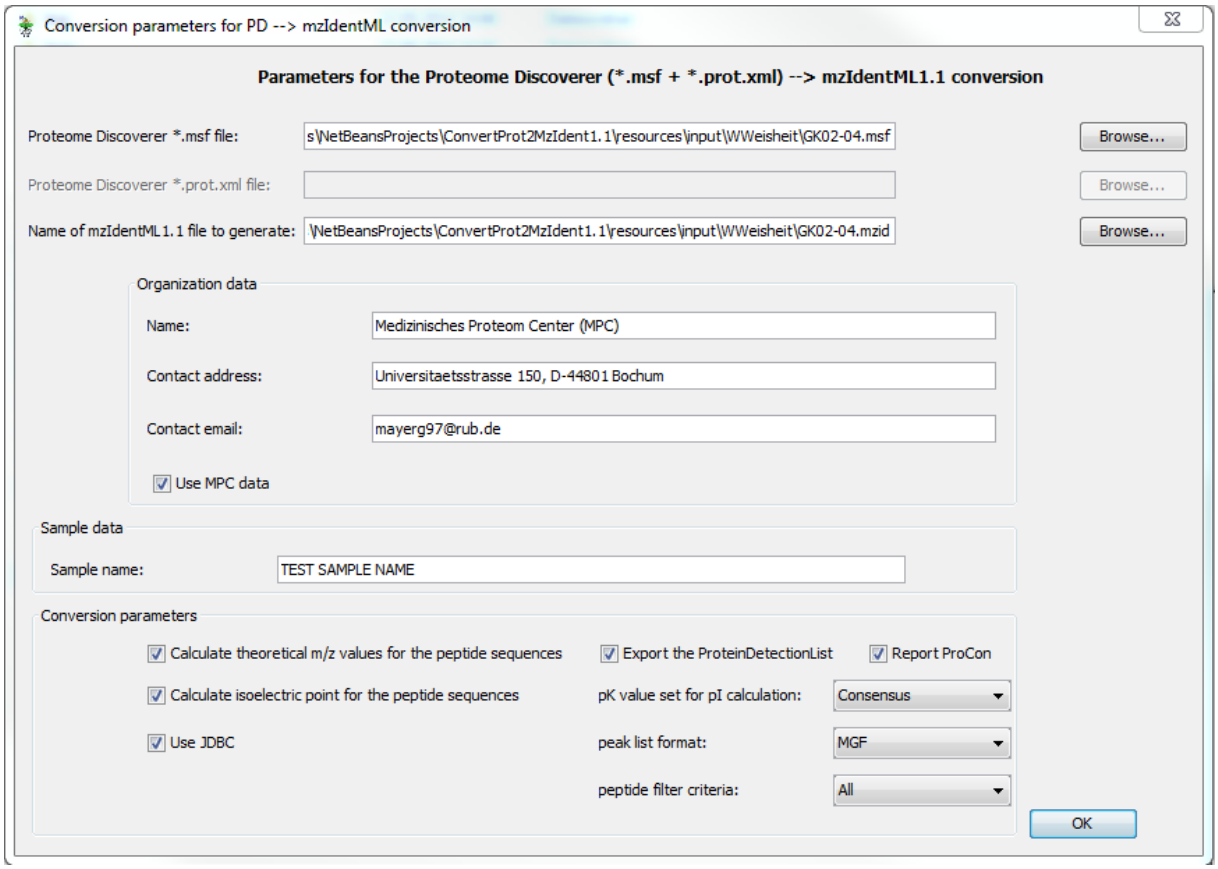
The specification of **User-defined** score thresholds as they can be afterwards set in the ProteomeDiscoverer® "ResultFilter" tab is in preparation.

If you choose the checkbox "**Use JDBC**", which is strongly recommended, then the access to the SQLite database (.msf file) is done via the general JDBC interface. Otherwise the native sqlite4java binding is used. Note that if you decide not to use JDBC, then you have to install the respective .dll (Dynamic Link Library) (Windows), .dylib (Mac OS X) resp. .so (shared object, Linux) files first to your computer (see section 1: Installation) and that sqlite3.dll currently only works for 32 Bit Windows.

The **isoelectric point** [8-10] is calculated by calculating the zero point of the Henderson-Hasselbach equation (http://isoelectric.ovh.org/files/isoelectric-point-theory.html) in an iterative way. An optimized algorithm is used, so that mostly only between 7 and 9 iterations are needed for convergence. Because the result depends of the underlying pK values, one can choose from the combo box, which pK value set for the calculation should be used – by default a consensus of the results of all pK value sets (with exception of the Patrickios [9]

value set) is used. The Patrickios value set is left out from the consensus calculation, because it uses no pK values for the residues Cys, His and Tyr and therefore the results of the Patrickios value set calculation often differ significantly from the results got by using the other value sets, which use pK values for all charged residues (i.e. the terminal -NH2, -COOH, Cys, Asp, Glu, His, Lys, Arg and Tyr).

The list box "**peak list format**" allows one to specify the peak list file format, since this is information is not stored in the .msf file. We recommend to use .mgf as peak list format, which can be easily exported from ProteomeDiscoverer®.



After export of the .mzid file you can check it together with the peak list files by using the PRIDE Inspector [11] and the mzIdentML validator [12] software.

8 Workflow 4: ProteinScape® 2.1 to mzIdentML conversion

The SearchEvent results of ProteinScape® 2.1 can be converted into the standard mzIdentML1.1 format.

In the panel “Connection data” (Figure 2) one must first select the server to use and must specify the user name and the password. Also the database owner should be changed if it’s not “dbo”. If you don’t know your database owner, you can check it with the Microsoft SQLServer® 2012 Management Studio (http://www.microsoft.com/en-us/download/details.aspx?id=29062). There are empty entries selectable from the comboboxes, which are editable and allow you to specify your own server. After pressing the “Connect ...” button one can choose in the tables the desired project, sample and single SearchEvent, for which the results should be converted into mzIdentML 1.1. Then in the text field an output file name is automatically proposed, but it can be changed by pressing the “Browse ...” button. After pressing the “Convert ...” button the conversion process is started and a progress bar shows the status of the conversion.
Alternatively one can click on the radio button “Convert Gels”. Then all the gels for a given project – sample combination are shown. If you select a gel, then all search events for the whole gel are shown automatically in the table for search events and you can again start the conversion process by pressing the “Convert ...” button.
After finishing the conversion a message box informs the user and the connection to the ProteinScape 2.1 database is automatically closed.

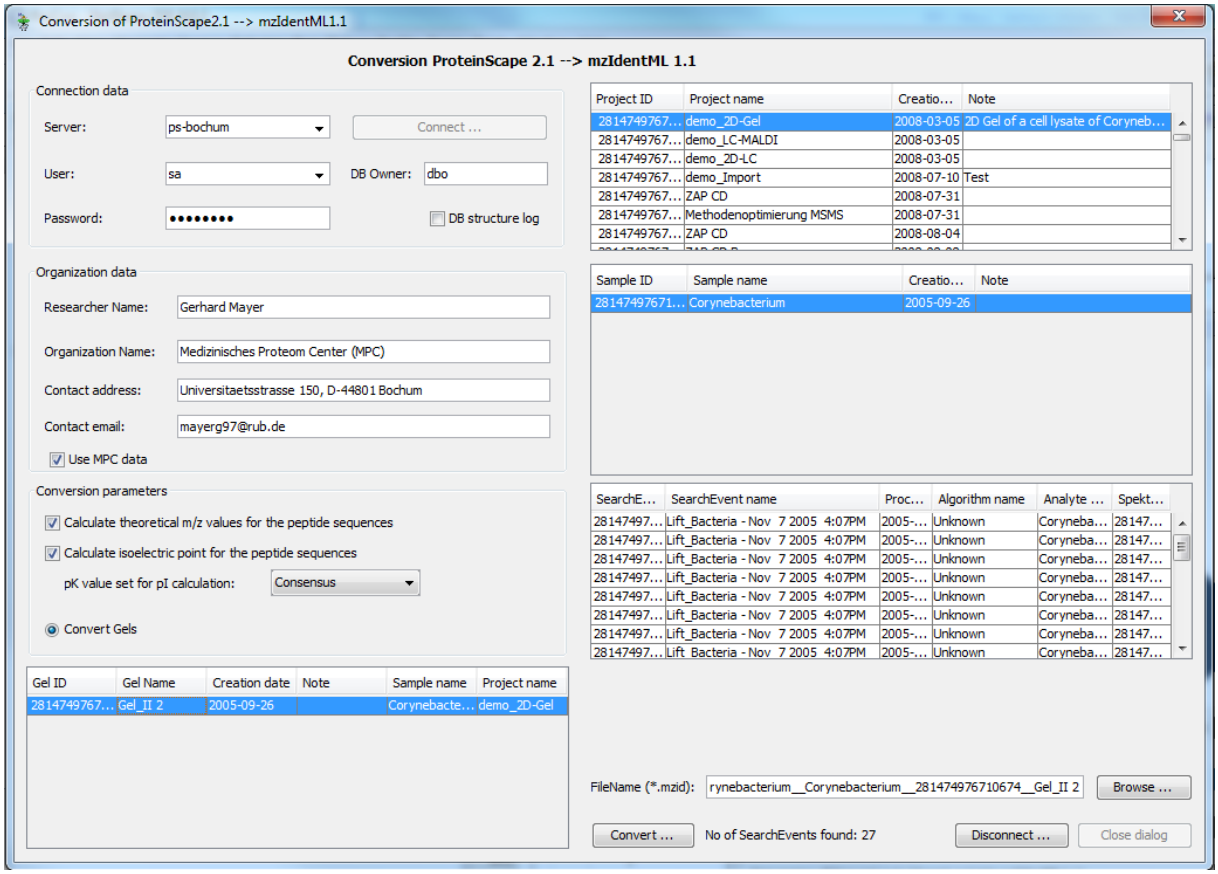
In the panel “Organization data” one can enter the name and contact details. If one clicks the checkbox “Use MPC data” then these contact fields are filled in with the data of the MPC (Medical Proteome Center) in Bochum.

In the panel “Conversion parameters” one can choose if the theoretical m/z values and isoelectric points for the peptide sequences should be calculated.

The isoelectric point [8-10] is calculated by calculating the zero point of the Henderson-Hasselbach equation (http://isoelectric.ovh.org/files/isoelectric-point-theory.html) in an iterative way. An optimized algorithm is used, so that mostly only between 7 and 9 iterations are needed for convergence. Because the result depends of the underlying pK values, one can choose from the combo box, which pK value set for the calculation should be used – by

default a consensus of the results of all pK value sets (with exception of the Patrickios [9] value set) is used. The Patrickios value set is left out from the consensus calculation, because it uses no pK values for the residues Cys, His and Tyr and therefore the results of the Patrickios value set calculation often differ significantly from the results got by using the other value sets, which use pK values for all charged residues (i.e. the terminal -NH2, -COOH, Cys, Asp, Glu, His, Lys, Arg and Tyr).

If you have connection problems to your SQLServer you can use the test program TestSQLServerAccess from the tools menu, which allows you to check your connection parameters.
Maybe you must create a new user, which not have to go through windows authentication. One must give that user read privileges for all ProteinScape 2 databases (proteinscape, gum, lcc, processingkernel).



9 Workflow 5: Spectral Counts to mzQuantML conversion

ProCon supports also the conversion from spectral count result files into mzQuantML:

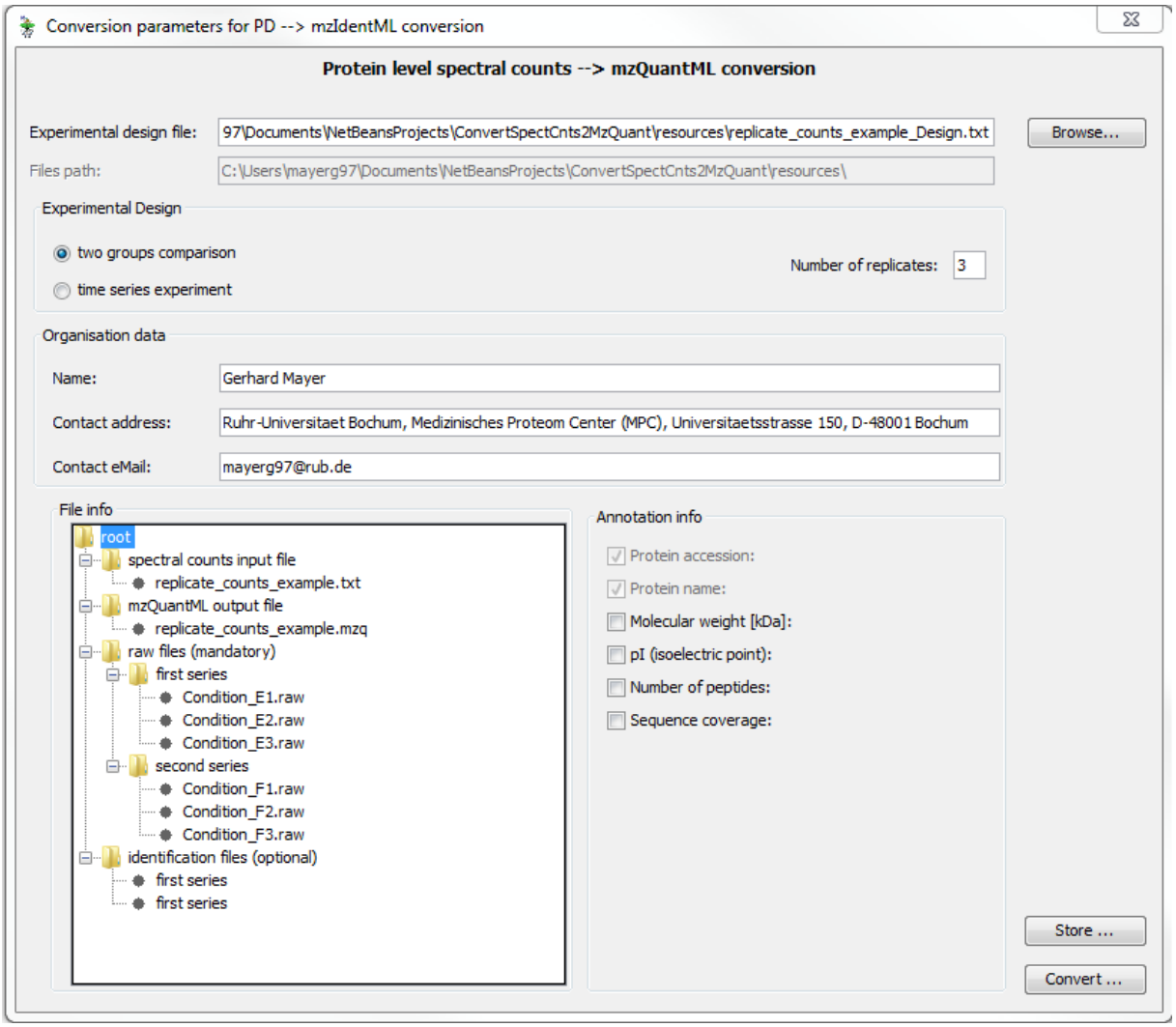
For that the user must define a so called experimental design file (of type .txt), which contains all parameters for the spectral counts conversion. After this design file is chosen all other fields of the GUI are set. By pressing the Convert… button the conversion into mzQuantML is started.

Currently two designs are currently supported:

- the comparison of two groups with technical replicates
- the comparison of two time series data sets (without technical replicates)

Example design and data files you can find under the SpectCnt folder.

The spectral counts input file, the .mzq (mzQuantML) output file and their paths are already defined in the experimental design files.



10 Tools menu

The tools menu contains the following two small tools:

1) PeptidePropertyCalculator: this tool allows you to enter a peptide sequence and to calculate the pI value and the molecular weight for this peptide.

2) Test MS SQLServer access: This tool allows you to test the access to your ProteinScape MS SQLServer backend database. If you cannot get access to the database, you must check your network and firewall configuration.

11 Command Line Arguments for batch mode

ProCon can also be started from the command line by specifying all the arguments, which are normally specified via the GUI (Graphical User Interface). One must specify all mandatory arguments. If an argument of an option contains blank characters, one must use quotation marks. If an optional argument is not specified, then the given default is used (for flag options false means not set).

Make sure that the variables JAVA_HOME and JAVA_OPTS are set properly in your batch file, e.g. for the Windows environment:

set JAVA_HOME="C:\Program Files\Java\jre1.8.0_66\bin\java.exe"

set JAVA_OPTS=-d64 -server -XX:+AggressiveOpts –Xmn1g –Xms2g –Xmx4g -XX:+UseParNewGC -XX:+UseConcMarkSweepGC -XX:ParallelGCThreads=20

The following two options are common to all five converters:

Option	**Description**	**Mandatory / Optional**	**Default value**
-h	Print help screen for command line arguments	optional	false
-conv	name of the … --> .mzid converter program (PD1x, PS13, PS21, SEQO)	mandatory	---

Example: %JAVA_HOME% %JAVA_OPTS% -jar ProCon.jar –conv -h

a) For the ProteomeDiscoverer 1.x (x=1,..,4) converter the following options are defined:

Option	**Description**	**Mandatory / Optional**	**Default value**
-pks	pK value set for isoelectric Point (IP) calculation, (‘Consensus’ recommended)	optional	Consensus
-mz	Calculate theoretical mass/charge values for the	optional	true

	peptide sequences (recommended)		
-ip	Calculate isoelectric points for the peptide sequences (recommended)	optional	true
-jdbc	Use JDBC driver (recommended)	optional	true
-pdl	Export of ProteinDetectionList (recommended)	optional	true
-rpc	Report ProCon as AnalysisSoftware in .mzid file	optional	true
-affname	Contact information - affiliation name	mandatory	---
-affaddr	Contact information - affiliation address	mandatory	---
-umail	Contact information - user email	mandatory	---
-msf	ProteomeDiscoverer 1.x (x=2,…,4) *.msf input file	mandatory	---
-prot	ProteomeDiscoverer 1.1 / 1.2 *.prot.xml input file	mandatory only for PD1.x	---
-sampname	Sample name	mandatory	
-plff	Peak list file format, e.g. MGF, PKL, mzML, …	optional	MGF
-peptf	Peptide filtering (ALL, RELAXED or STRICT)	optional	All
-mzid	mzIdentML .mzid output file	mandatory	---

Example: %JAVA_HOME% %JAVA_OPTS% -jar ProCon.jar –conv=PD1x –pks=Consensus –mz –ip –jdbc –pdl –rpc –affname="Medizinisches Proteom Center (MPC)" –affaddr="Universitätsstraße 150, D-44801 Bochum" –umail=mayerg97@rub.de –msf=D:/ProteomeDiscoverer/Oscar/Test2/2012_310_MCH_Banda2_03.msf –mzid=D:/ProteomeDiscoverer/Oscar/Test2/2012_310_MCH_Banda2_03.mzid –sampname="Test Sample" –plff=MGF – peptf=ALL -h

b) For the ProteinScape 1.3 converter the following options are defined:

Option	**Description**	**Mandatory / Optional**	**Default value**
-sname	Connection data - server name or IP adress	mandatory	---
-uname	Connection data - user name	mandatory	---
-pw	Connection data - password	mandatory	---
-instrf	mzData instruments file	optional	none
-brf	bibliographic references file	optional	none
-smf	SLoMo (Site LOcalization of MOdifications) [13] file	optional	none

-smthr	SLoMo (Site LOcalization of MOdifications) threshold	optional	0.0
-pride	PRIDE XML output file	mandatory	---
-sep	separation name	one of them is mandatory	
-seid	SearchEvent ID		
-pmf	PMF (peptide mass fingerprint) flag	at least one of them is mandatory	
-pff	PFF (peptide fragment fingerprint) flag		
-inp_sm	search method	mandatory	---
-inp_dbv	search database version	mandatory	---
-inp_prot	protocol name	mandatory	---
-inp_swn	instrument software name	mandatory	---
-inp_swv	instrument software version	mandatory	---
-inp_swc	instrument software comments	mandatory	---
-inp_si	instrument	optional	
-inp_ord	String indicating the order of digestion and separation; either “Dig_before_Sep” or “Sep_before_Dig”	one of them is mandatory	"Dig_before_Sep" or "Sep_before_Dig"
-dig_first	flag indicating “first digestion, then separation”		
-sep_first	flag indicating “first separation, then digestion”		

Example: %JAVA_HOME% %JAVA_OPTS% -jar ProCon.jar –conv=PS13 –sname=ZAPPS –uname=sa –pw dcchbpp –pride=./XML_files/ProCon_PRIDEXML.xml –sep="Profile CCC | CCC | 2701 up (gelID:721)" –pmf -inp_sm="Search machine" -inp_dbv="3.84" -inp_prot="Protocol name" -inp_swn="SW name" -inp_swv="SW version" -inp_swc="SW comments" -dig_first-h

c) For the ProteinScape 2.1 converter the following options are defined:

Option	**Description**	**Mandatory / Optional**	**Default value**
-pks	pK value set for isoelectric Point (IP) calculation, (‘Consensus’ recommended)	optional	Consensus
-mz	Calculate theoretical mass/charge values for the peptide sequences (recommended)	optional	true
-ip	Calculate isoelectric points for the peptide sequences (recommended)	optional	true

-dblog	Switch on DataBase structure logging (not recommended - off as default)	optional	false
-sname	Connection data - server name or IP address (‘maldiraumserver’ as default)	mandatory	
-uname	Connection data - user name (‘sa’ as default)	mandatory	sa
-pw	Connection data - password	mandatory	
-dbo	Connection data - database owner (‘dbo’ as default)	optional	dbo
-resname	Contact information - researcher name	mandatory	
-orgname	Contact information - organization name	mandatory	
-affaddr	Contact information - affiliation address	mandatory	
-umail	Contact information - user email	mandatory	
-mzid	mzIdentML .mzid output file	mandatory	
-projID	Project ID	mandatory	
-sampID	Sample ID	mandatory	
-seid	SearchEvent ID	mandatory	

Example: %JAVA_HOME% %JAVA_OPTS% -jar ProCon.jar –conv=PS21 –sname=maldiraumserver –uname=sa –pw=bruker2008 –resname="Gerhard Mayer" –orgname="Medizinisches Proteom Center (MPC)" –affaddr="Universitaetsstrasse 150, D-44801 Bochum" –umail=mayerg97@rub.de –mzid="C:/Users/Gerhard/101217_sAPP_first_562949953421517_562949953425925_562949953431912.mzid" -pks=Consensus –projID="562949953421517" -sampID="562949953425925" -seid="562949953431912"

d) For the SEQUEST.out converter the following options are defined:

Option	**Description**	**Mandatory / Optional**	**Default value**
-seqout	SEQUEST.out / comet.out (input) file folder	mandatory	
-mzid	mzIdentML .mzid (output) file	mandatory	
-vers10	Convert to mzIdentML version 1.0 (not recommended – off as default)	optional	false
-comet	Conversion of comet.out folder	optional	false

Example: %JAVA_HOME% %JAVA_OPTS% -jar ProCon.jar -conv=SEQO –seqout=./SEQUEST/example_folder –mzid=./XML_files/ProCon_mzIdentML.mzid –vers10 -mod –h

e) For the spectral counts converter the following options are defined:

Option	Description	Mandatory / Optional	Default value
-df	spectral counts design file (input) file	mandatory	

Example: %JAVA_HOME% %JAVA_OPTS% -jar ProCon.jar -conv=SC -df=./SpectCnt/replicate_counts_example_Design.txt -h

12 Versioning Information and Release Notes

- **0.9.627 (14th December 2015) Corrected error with PeptideEvidence references**
- 0.9.625 (03th December 2015) Build in the filtering of peptides (ALL, RELAXED, STRICT); added the flag –peptf for peptide filtering to the cmd-line options
- 0.9.624 (30th November 2015) Withdraw the reporting of unique peptides.
- 0.9.623 (18th November 2015) Made some preparations for planned source code switch to Java 8.
- 0.9.620 (16th November 2015) Location of <SpectraData> element works now for lower case and upper case peak list files.
- 0.9.619 (13th November 2015) - Added flag –rpc for reporting of ProCon as AnalysisSoftware in the .mzid file
- 0.9.618 (10th November 2015) Handling of Fasta lines without accession (i.e. containing only protein name)
- 0.9.617 (22th September 2015) Some small errors corrected in PD converter
- 0.9.616 (10th September 2015) Integrated the spectral counts converter; corrected error in Sequest / Comet conversion
- 0.9.610 (26th August 2015) Solved now the spectrumID problem also for merged .msf files (see http://www.ebi.ac.uk/mzidentml-documentation-developers)

13 How to cite

If you want to cite or acknowledge ProCon, you can cite the following paper [14]:

ProCon - PROteomics CONversion tool.
Mayer G, Stephan C, Meyer HE, Kohl M, Marcus K, Eisenacher M.
J Proteomics. 2015 Jul 13. S1874-3919(15)30053-1.
doi: 10.1016/j.jprot.2015.06.015.
PMID: 26182917

14 Known Bugs

-
- PD conversions of .msf files which reference .mzML peak list files produce output which is not validated by the ProteomeXchange [15] / PRIDE [16] validation procedure.
- Search Engine MSFit (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msfitstandard) not implemented, Exception raised
- If more than one separation and digestion protocol is used for one gel (regarding tables: MaldiPreparationProtocol, EnzymeLot, CleavageEnzyme, IEF_Protocol, StainProtocol, PAA_Protocol, DigestionProtocol, SpotTools), than a "more than one rows in result set" exception is thrown (PRIDE (http://www.ebi.ac.uk/pride) allows only one protocol). Future implementation could be to let the user select the SepAndDig protocol, for which he wants to convert results.

If you find further bugs please sent an email with the following log files:

- For ProteomeDiscoverer conversions:
  - ConvertProt2MzIdent1.1.txt
  - Memory_Properties_Log.txt
  - ProCon.txt
- For ProteinScape 2.1 conversions:
  - ConvertPS2MzIdent1.1.txt
  - SQLServerDBLogFile.txt
  - ProCon.txt
- For ProteinScape 1.3 conversions:
  - pride.txt

from the /log resp. /logs directories to mayerg97@rub.de or martin.eisenacher@rub.de .

Please specify also:

- your exact operating system (Windows or Linux), 32 or 64 bit
- the versions of your Java runtime (type *java –version* on your command interpreter cmd.exe)
- the version of ProCon (see Help-About menu from ProCon)
- in case of ProteomeDiscoverer conversions the .msf file or at least the .msf file size in kBytes.

15 Planned future functionality

- Support of filtering afterwards (Tab “ResultFilter” in ProteomeDiscoverer)
- Generating mzIdentML 1.2 output
- Integration into the KNIME [17] workflow system

16 Acknowledgements

The ProCon development was funded by the following projects: The ProDaC (http://www.fp6-prodac.eu) Coordination Action, 6th EU framework programme, project number LSHG-CT-2006-036814, the ProteomeXchange (http://www.proteomexchange.org) consortium, EU FP7 grant 'ProteomeXchange' [grant number 260558] and de.NBI (http://www.denbi.de) project funded by the German Federal Ministry of Education and 414 Research (BMBF), grant number FKZ 031 A 534A, the Deutsche Gesetzliche 411 Unfallversicherung (DGUV, http://www.dguv.de/de/Forschung/Aktuelle-(laufende)-Forschungsprojekte/index.jsp) project DGUV-Lunge (617.0 FP 339A) and P.U.R.E. 412 (http://www.pure.rub.de), a project of Nordrhein-Westfalen, a federal state of Germany.

Thanks to all people who contributed test files and reported errors, especially Óscar Gallardo, Laboratori de Proteòmica CSIC/UAB, Barcelona, Spain and Gorka Prieto Agujeta, University of the Basque Country (UPV/EHU), Bilbao, Spain.

17 References

[1] Deutsch EW, Albar JP, Binz PA *et al.* Development of data representation standards by the human proteome organization proteomics standards initiative. *J Am Med Inform Assn*, 22(3), 495-506 (2015).

[2] Jones AR, Eisenacher M, Mayer G *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*, 11(7), M111 014381 (2012).

[3] Walzer M, Qi D, Mayer G *et al.* The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol Cell Proteomics*, 12(8), 2332-2340 (2013).

[4] Mayer G, Jones AR, Binz PA *et al.* Controlled vocabularies and ontologies in proteomics: overview, principles and practice. *Biochimica et biophysica acta*, 1844(1 Pt A), 98-107 (2014).

[5] Mayer G, Montecchi-Palazzi L, Ovelleiro D *et al.* The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. *Database-Oxford*, (2013).

[6] Dorfer V, Pichler P, Stranzl T *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res*, 13(8), 3679-3684 (2014).

[7] Uszkoreit J, Maerkens A, Perez-Riverol Y *et al.* PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J Proteome Res*, 14(7), 2988-2997 (2015).

[8] G.R. Grimsley, J.M. Scholtz, C.N. Pace, A summary of the measured pK values of the ionizable groups in folded proteins, *Protein Sci*, 18 (2009) 247-251.

[9] C.S. Patrickios, E.N. Yamasaki, Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory, *Analytical biochemistry*, 231 (1995) 82-91.

[10] A. Sillero, J.M. Ribeiro, Isoelectric points of proteins: theoretical determination, Anal Biochem, 179 (1989) 319-325.

[11] Perez-Riverol Y, Xu QW, Wang R *et al.* PRIDE Inspector Toolsuite: moving towards a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. *Mol Cell Proteomics*, (2015).

[12] Ghali F, Krishna R, Lukasse P *et al.* Tools (Viewer, Library and Validator) that Facilitate Use of the Peptide and Protein Identification Standard Format, Termed mzIdentML. *Mol Cell Proteomics*, 12(11), 3026-3035 (2013).

[13] C.M. Bailey, S.M. Sweet, D.L. Cunningham, M. Zeller, J.K. Heath, H.J. Cooper, SLoMo: automated site localization of modifications from ETD/ECD mass spectra, *J Proteome Res*, 8 (2009) 1965-1971.

[14] G. Mayer, C. Stephan, H.E. Meyer, M. Kohl, K. Marcus, M. Eisenacher, ProCon - PROteomics CONversion tool, *J Proteomics* (2015), 2015 Jul 13. pii: S1874-3919(15)30053-1. doi: 10.1016/j.jprot.2015.06.015.

[15] Vizcaíno JA, Deutsch EW, Wang R *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3), 223-226 (2014).

[16] Vizcaíno JA, Csordas A, Del-Toro N *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic acids research*, (2015).

[17] Aiche S, Sachsenberg T, Kenar E *et al.* Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics*, 15(8), 1443-1447 (2015).