

# ProCon – Proteomics Conversion Tool

ProCon – Proteomics Conversion Tool.....	1
1 Introduction .....	2
2 Installation .....	2
3 Start ProCon .....	4
4 Workflow 1: Sequest.out / comet.out Import and mzIdentML export.....	5
4.1 Sequest-specific Configuration .....	5
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: .....	5
·Header.txt only for SEQUEST.out; not required for comet.out conversions .....	5
·sequest.log only for SEQUEST.out; not required for comet.out conversions .....	5
·sequest.params <b>resp.</b> comet.params.....	5
·<folder-name>.log, e.g. if your folder is named 'test_folder' then the file should be named 'test_folder.log'.....	5
4.1.1 File ./config/SEQUEST_massvalues.txt .....	7
4.1.2 File ./config/SEQUESTMOD.obo .....	7
4.1.3 File ./config/unimod.obo .....	8
4.1.4 File ./config/Sequest.properties .....	8
4.1.5 File ./config/mzidAuditCollection_1.1.xml .....	8
5 Workflow 2: ProteinScape <sup>®</sup> 1.3 Import and PRIDE XML export.....	10
5.1 Configuration .....	10
The following configuration files are text files and can be edited with any text editor:.....	10
File ./config/ProCon.properties: .....	10
File ./config/log4j.properties: .....	10
5.2 Prerequisites .....	10
Connection to your local ProteinScape <sup>®</sup> database .....	10
Data for First Test: .....	11
Connection to Ontology Lookup Service [9]PRIDE XML.....	11
:.....	11
5.3 ProteinScape <sup>®</sup> -specific Configuration.....	11
File ./config/PAG-PS.obo .....	11
File ProteinScape.properties.....	11

5.4 ProteinScape <sup>®</sup> Data Generation .....	12
5.5 Testing / Initializing the Database Connection .....	12
5.6 Converting a Search Event .....	13
5.7 Converting Gel Data.....	13
5.8 Instrument information and References .....	14
5.9 Missing Information .....	14
6 Workflow 3: Proteome Discoverer <sup>®</sup> to mzIdentML conversion.....	16
Configuration (file ./config/log4j.properties):.....	16
7 Workflow 4: ProteinScape <sup>®</sup> 2.1 to mzIdentML conversion .....	20
8 Workflow 5: Spectral Counts to mzQuantML conversion.....	22
9 Tools menu .....	23
10 Command Line Arguments for batch mode .....	23
11 KNIME integration .....	27
12 Versioning Information and Release Notes .....	27
13 How to cite .....	30
14 Known Bugs .....	30
15 Planned future functionality .....	31
16 Acknowledgements .....	32
17 References .....	33

## 1 Introduction

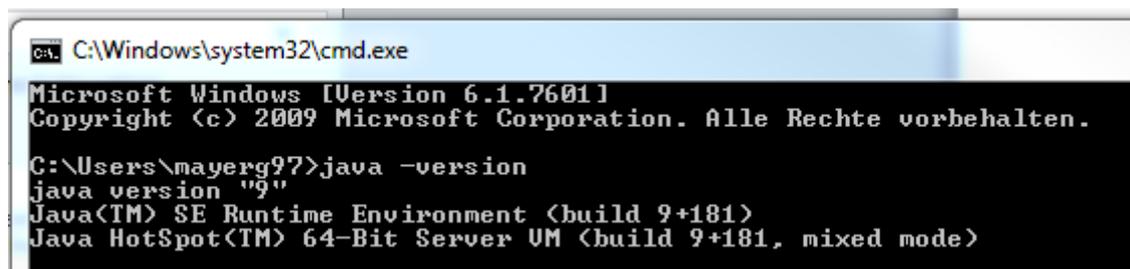
With ProCon you can convert proteomics identification and quantification results into HUPO standard formats [1, 2], which can be used to upload your results into public repositories.

## 2 Installation

Install Java SE (Standard Edition) (JRE): ProCon 0.9.806 was developed and tested with JOpenJDK 13.0.1, 64 Bit. For Java 8, 9 and 10 you should use the old version 0.9.723 instead. The latest Java Runtime Environment (JRE) is downloadable from <http://www.oracle.com/technetwork/java/javase/downloads/index.html> resp from <https://jdk.java.net/13/>.

You can inspect, which Java version is installed on your PC by executing the following command in the Windows command prompt (cmd.exe):

```
java -version
```



```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Alle Rechte vorbehalten.

C:\Users\mayerg97>java -version
java version "9"
Java(TM) SE Runtime Environment (build 9+181)
Java HotSpot(TM) 64-Bit Server VM (build 9+181, mixed mode)
```

### **! 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<sup>®</sup> 1.3 Import ...” sections).

ProCon since version 0.9.723 only supports access to the SQLite database (.msf file) of ProteomeDiscoverer<sup>®</sup> files via JDBC, which utilizes the driver from <https://bitbucket.org/xerial/sqlite-jdbc>.

For starting ProCon you must simply right click (or double click) on the ProCon.jar file.

### 3 Start ProCon

ProCon 0.9.806 runs with Java 13 and higher. For Java 8, 9 and 10 you can use the old version 0.9.723.

Start the ProCon.jar by double-clicking the ProCon.jar file.

Any firewall message or any question of a protection tool about “an application start” should be confirmed / allowed. In case ProCon does not start, you can 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<sup>®</sup> and Sequest classes and **pride.log** for pride classes).

There are five converters you can follow in the current version:

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

## 4 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) [6]

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>.

### 4.1 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 - xxxxxxxxxxxxxxxxxxx ... // (xxxxxxxxxxxxxxxx = 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 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx ...
// (xxxxxxxxxxxxxxxxxxxxxxxxxxxxx = 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 [7] 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:

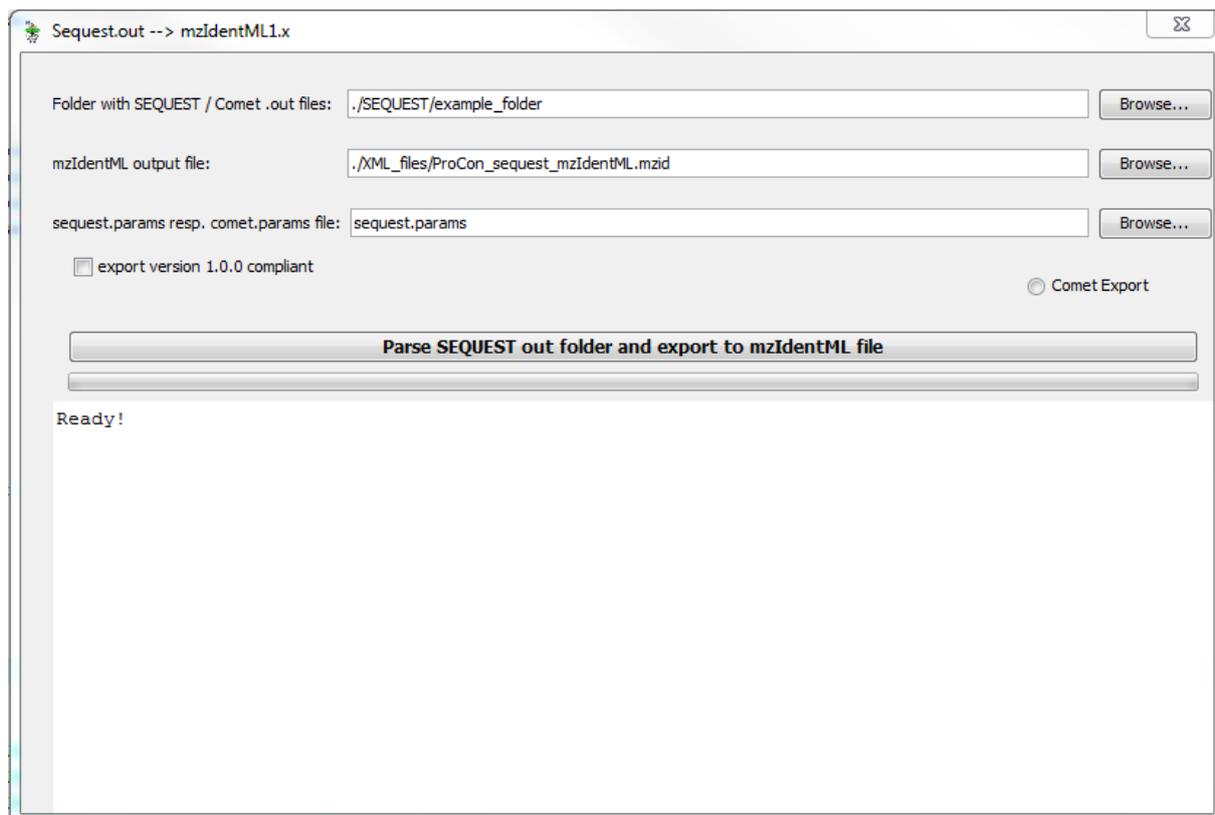
```
file://www.medizinisches-proteom-
center.de/sequestmaster/work/Datenbank/StdCry_nr.fasta
```

#### 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 (Functional Genomics Experiment) [8]). Please overwrite with your details!

You should not change the sequence of comment and content lines. If you accidentally 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.



## 5 Workflow 2: ProteinScape<sup>®</sup> 1.3 Import and PRIDE XML export

### 5.1 Configuration

The following configuration files are text files and can be edited with any text editor:

#### 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 “=”).

#### File `./config/log4j.properties`:

The line `log4j.appender.pride_core_file.File` should contain `\\` as path separator for Windows and `/` for Unix-based operating systems.

### 5.2 Prerequisites

#### Connection to your local ProteinScape<sup>®</sup> 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<sup>®</sup> 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 [9]PRIDE XML:

During import of ProteinScape<sup>®</sup> 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 can establish an online connection at runtime. You should configure firewall rules appropriately or answer firewall questions with Yes.

## **5.3 ProteinScape<sup>®</sup>-specific Configuration**

### File ./config/PAG-PS.obo

In this file the mapping from ProteinScape<sup>®</sup> modifications to PSI-MOD [10] is configured. As it is “name-based” there may exist differences in your ProteinScape<sup>®</sup> 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<sup>®</sup> 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
```

### 5.4 ProteinScape<sup>®</sup> Data Generation

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

- Fill in all fields for to describe project, sample, separation and spot/band (nearly all fields are exported, if not as CVPParam [11, 12], then as userParam);
  - o AVOID empty fields or “default” or “not specified” fields;
  - o 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.

### 5.5 Testing / Initializing the Database Connection

A default database connection string is given on the **ProteinScape<sup>®</sup> 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=iuser;password=iuser
```

Before you can import ProteinScape<sup>®</sup> 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<sup>®</sup> 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<sup>®</sup> 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>.

## 5.6 Converting a Search Event

- Specify a SearchEventID on the **ProteinScape<sup>®</sup> Source** tab and click the **Import ProteinScape<sup>®</sup> 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<sup>®</sup> 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 [13] file specified in the text field of this tab.

## 5.7 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<sup>®</sup> 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<sup>®</sup>, 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<sup>®</sup> imports are added to the internally assembled PRIDE data set and can be flushed out together using **Export to File**.

## 5.8 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.

## 5.9 Missing Information

Depending on the information or type of results you want to export, ProteinScape<sup>®</sup> 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 ProteinScope<sup>®</sup> (ProteinScope<sup>®</sup> 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 ProteinScope1.3 --> mzIdentML1.1

Conversion ProteinScope 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/ProteinScope1\_0;user=sa;password=dchbpb

Logging messages

```
Initializing DB connection...
Some project names:
demo_2D-Gel
demo_LC-MALDI
demo_Quantification
DB connection successfully initialized.
[Import ProteinScope 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 ProteinScope 1.3 content to convert and export the selected ProteinScope 1.3 content into PRIDE XML

Select Separation: PROFILE.CCC | CCC | 2701 up (gelID:721)

OR

Specify SearchEventID: <SearchEventId>

Select PRIDE XML output file: ./XML\_files/ProCon\_PRIDEXML.xml Browse...

Import ProteinScope Separation and export to PRIDE XML

OR

Import ProteinScope SearchEvent and export to PRIDE XML

PMF (MS)

PFF (MS/MS)

## 6 Workflow 3: Proteome Discoverer<sup>®</sup> to mzIdentML conversion

### Configuration (file `./config/log4j.properties`):

The line `log4j.appender.ProCon_file.File` should contain `\\` as path separator for Windows and `/` for Unix-based operating systems.

The output of the Mascot [14], SEQUEST [15], ZCore and MS Amanda [16] search engines of Proteome Discoverer<sup>®</sup> can be converted into the standard mzIdentML1.1 format. For the conversion of ProteomeDiscoverer<sup>®</sup> 1.3 and 1.4 results, only the `*.msf` (Mass Spec Format, Thermo) file is needed. For ProteomeDiscoverer<sup>®</sup> 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 1:** Version 2.0 and newer of ProteomeDiscoverer<sup>®</sup> are not supported, since beginning with version 2.2 ProteomeDiscoverer<sup>®</sup> has an integrated mzIdentML exporter.

**Note 2:** For upload to the PRIDE repository [17] via ProteomeXchange [18-21] (<http://www.proteomexchange.org/submission-proteomexchange-pride>) you need also the peak list files (`.mgf` files). Because in the `.msf` files only the names of the `.raw` files are stored, there is the requirement that for each `.raw` file there is **exactly one** `.mgf` file and that all exported `.mgf` files have the same name (except the file type suffix of course) as the `.raw` files, e.g. `myfile.raw` → `myfile.mgf`. In case you have more `.mgf` files for a `.raw` file, you must merge the `.mgf` files (see for instance <http://www.proteinmetrics.com/merging-multiple-mgf-files/> or <http://www.uni-muenster.de/hippler/proteomatic/files/scripts-merge-mgf-files.xhtml>). Please also note that ProteomeDiscoverer<sup>®</sup> offers to export either "search inputs" or "peptide groups". Note that one has to use the "search inputs" option for the `.mgf` file export.

The tab for the ProteomeDiscoverer<sup>®</sup> 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 [3]

In the dialog box for entering the conversion parameters (Figure 1) one can choose the **input files** (the `*.msf` and for ProteomeDiscoverer<sup>®</sup> 1.1 and 1.2 also the `*.prot.xml` output files) of the ProteomeDiscoverer<sup>®</sup> 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 own protein inference algorithm, as e.g. the Protein Inference Algorithms PIA [22] (<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 to the peptide scores and the thresholds set in the ProteomeDiscoverer<sup>®</sup> workflow. The following threshold values are used:

- **STRICT**

- for XCorr (SEQUEST):
  - CutOffStdCharge1High
  - CutOffStdCharge2High
  - CutOffStdCharge3High
  - CutOffStdCharge4High
- for IonScore (Mascot): DefaultStrictScoreThreshold
- for AmandaScore (MS Amanda): AmandaScoreHighConfidenceThreshold

- **RELAXED**

- for XCorr (SEQUEST):
  - CutOffStdCharge1Middle
  - CutOffStdCharge2Middle
  - CutOffStdCharge3Middle
  - CutOffStdCharge4Middle
- for IonScore (Mascot): DefaultRelaxedScoreThreshold
- for AmandaScore (MS Amanda): AmandaScoreMiddleConfidenceThreshold

- **ALL**

Here all score thresholds are set to 0.0, so that no filtering using peptide scores takes place, which makes the resulting .mzid files very big

If you are in doubt, which filtering to use, we recommend the RELAXED filtering version.

The specification of **User-defined** score thresholds as they can be afterwards set in the ProteomeDiscoverer<sup>®</sup> “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.

The **isoelectric point** [23-25] 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 [23] 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 achieved by using the other value sets, which use pK values for all charged residues (i.e. the terminal -NH<sub>2</sub>, -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 information is not stored in the .msf file. We recommend to use .mgf as peak list format, which can be easily exported from ProteomeDiscoverer<sup>®</sup>.

If the CheckBox “Allow tracking” is checked, counting information is sent to us via Piwik (<https://piwik.org>), which helps us documenting the usage of this converter, in order to apply for funding for the further development of ProCon.

The screenshot shows a window titled "Conversion parameters for PD --> mzIdentML conversion". The main heading is "Parameters for the Proteome Discoverer (\*.msf + \*.prot.xml) --> mzIdentML1.1 conversion".

Fields for file selection:

- Proteome Discoverer \*.msf file: C:\Users\mayerg97\Downloads\MiKo\GK02-04.msf (Browse...)
- Proteome Discoverer \*.prot.xml file: (Browse...)
- Name of mzIdentML 1.1 file to generate: C:\Users\mayerg97\Downloads\MiKo\GK02-04.mzid (Browse...)

Organization data:

- Name: Medizinisches Proteom Center (MPC)
- Contact address: Universitaetsstrasse 150, D-44801 Bochum
- Contact email: mayerg97@rub.de
- Use MPC data

Sample data:

- Sample name: TEST SAMPLE

Conversion parameters:

- Calculate theoretical m/z values for the peptide sequences
- Calculate isoelectric point for the peptide sequences
- Use JDBC
- Allow tracking
- Export the ProteinDetectionList
- Report ProCon
- pK value set for pI calculation: Consensus
- peak list format: .mgf
- peptide filter criteria: RELAXED

OK button is present at the bottom right.

After export of the .mzid file you can check it together with the peak list files by using the PRIDE Inspector [26] and the mzIdentML validator [27] software.

## 7 Workflow 4: ProteinScape<sup>®</sup> 2.1 to mzIdentML conversion

The SearchEvent results of ProteinScape<sup>®</sup> 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<sup>®</sup> 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.

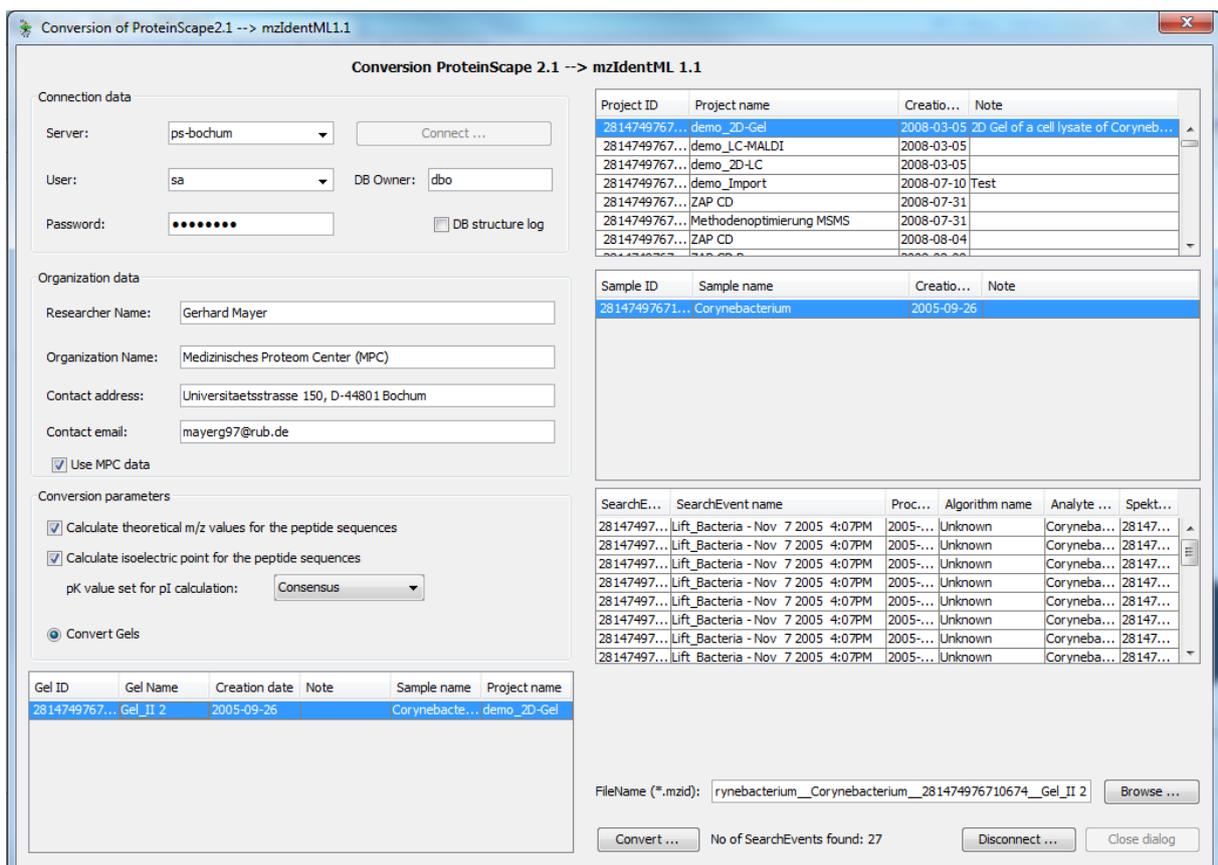
In the panel “Conversion parameters” one can choose if the theoretical m/z values and isoelectric points for the peptide sequences should be calculated.

The isoelectric point [23-25] 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 [23] 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 -NH<sub>2</sub>, -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 ProteinScope 2 databases (proteinscape, gum, lcc, processingkernel).



## 8 Workflow 5: Spectral Counts to mzQuantML conversion

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

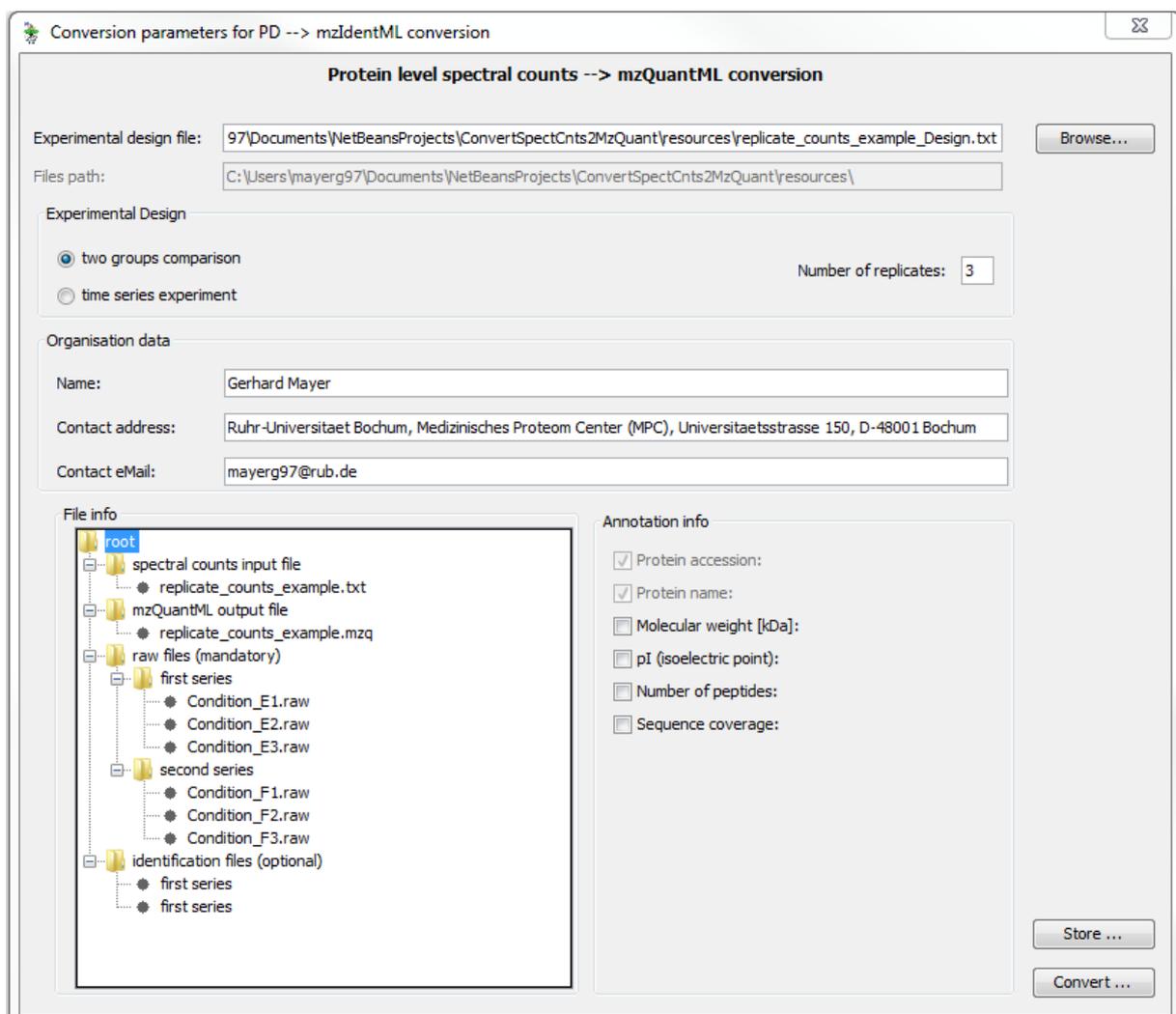
For that the user must define a so called experimental design file (of type .txt), which contains all parameters for the spectral counts conversion. After this design file is chosen all other fields of the GUI are set. By pressing the Convert... button the conversion into mzQuantML is started.

Currently two designs are 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.



## 9 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<sup>®</sup> MS SQLServer<sup>®</sup> backend database. If you cannot get access to the database, you must check your network and firewall configuration.

In addition it contains links to:

- the Medizinisches Proteom-Center (MPC), Ruhr-University Bochum
- the German Network for Bioinformatics Infrastructure (de.NBI)
- the service center “Bioinformatics for Proteomics” of de.NBI (BioInfra.Prot)
- the ProCon download page

## 10 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\jdk-13.0.1\bin"
```

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 /	Default
--------	-------------	-------------	---------

		<b>Optional</b>	<b>value</b>
-piwik	Use Piwik tracking	optional	false
-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
-jdbc	Use JDBC driver (recommended)	optional	true
-pdl	Export of ProteinDetectionList (recommended)	optional	true
-rpc	Report ProCon as AnalysisSoftware in .mzid file	optional	false
-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	---
-samppname	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 -piwik -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.msff -mzid=D:/ProteomeDiscoverer/Oscar/Test2/2012\_310\_MCH\_Banda2\_03.mzid -samppname="Test Sample" -plff=MGF -peptf=ALL -h

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

<b>Option</b>	<b>Description</b>	<b>Mandatory / Optional</b>	<b>Default value</b>
-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) [28] 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	
-seid	SearchEvent ID	is mandatory	
-pmf	PMF (peptide mass fingerprint) flag	at least one of	
-pff	PFF (peptide fragment fingerprint) flag	them is mandatory	
-inp_sm	search method	mandatory	---
-inp_dbv	search database version	mandatory	---
-inp_prot	protocol name	mandatory	---
-inp_swn	instrument software name	mandatory	---
-inp_svw	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 dcchbpb -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\_svw="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,	optional	Consensus

	('Consensus' recommended)		
-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

**Remark:** If you want to execute several conversions in a row, just use a command line for each data file, with the corresponding file names set.

## 11 KNIME integration

The integration into the KNIME [29, 30] workflow system is not yet supported, but planned for a future release of ProCon.

## 12 Versioning Information and Release Notes

- **0.9.806 (29<sup>th</sup> January 2019) Adapted to OpenJDK13**
- 0.9.726 (14<sup>th</sup> August 2019) Now checks, if the .msf file version is from PD 1.x (x=1,2,3,4)
- 0.9.725 (07<sup>th</sup> July 2019) Adaptions to Java 12
- 0.9.724 (13<sup>th</sup> March 2019) Adaptions to Java 11
- 0.9.723 (12<sup>th</sup> September 2018) Some small error corrections
- 0.9.720 (11<sup>th</sup> May 2018) Added links to de.NBI and BioInfra.Prot in the Help menu
- 0.9.718 (02<sup>nd</sup> February 2018) Added parameter for Piwik tracking to batch mode
- 0.9.716 (09<sup>th</sup> January 2018) Introduced Piwik-tracking; adapted to Java 8 style; some performance optimizations; library updates
- 0.9.708 (12<sup>th</sup> October 2017) Folder Demo\_Data contains now all demo data sets; now runs also under Java9

- 0.9.706 (18<sup>th</sup> September 2017) Some small usability improvements
- 0.9.705 (05<sup>th</sup> September 2017) Corrected error with enzyme name “No-Enzyme (Unspecific)”, added missing modification form Unimod and PSI-MOD, and done library updates
- 0.9.703 (22<sup>nd</sup> August 2017) Corrected error with enzyme name “Trypsin/P (Full)” and library updates
- 0.9.654 (07<sup>th</sup> June 2017) Corrected selection of peptide filter = ALL; resources and library updates
- 0.9.652 (29<sup>th</sup> March 2017) Corrected a bug in reading fasta lines
- 0.9.647 (23<sup>rd</sup> November 2016) Some minor GUI
- 0.9.646 (02<sup>nd</sup> November 2016) Updated to use the latest ontology and taxonomy files
- 0.9.641 (06<sup>th</sup> April 2016) Filenames in <SpectraData> are now of type anyURL and it is ensured that they are unique
- 0.9.640 (15<sup>th</sup> March 2016) corrected NullPointerException in case of STRICT and RELAXED peptide filtering
- 0.9.639 (14<sup>th</sup> March 2016) command-line parameter `-rpc` is false by default in PD1.x conversion
- 0.9.638 (03<sup>rd</sup> March 2016) some further performance optimizations
- 0.9.637 (26<sup>th</sup> February 2016) performance optimization and corrected an error in peptide filtering when started from batch mode
- 0.9.636 (11<sup>th</sup> February 2016) performance optimization and a further `accession=""` problem corrected; corrected errors in parameter handling and in batch options parsing
- 0.9.631 (02<sup>nd</sup> February 2016) Corrected problems with MappingL, “use flanking ions” and `accession=""` under <ProteinDetectionHypothesis>
- 0.9.630 (07<sup>th</sup> January 2016) Corrected problem with MappingI
- 0.9.628 (17<sup>th</sup> December 2015) Changed handling of FastaLines with `>sw|...`
- 0.9.627 (14<sup>th</sup> December 2015) Corrected error with <PeptideEvidence> references
- 0.9.625 (03<sup>th</sup> December 2015) Build in the filtering of peptides (ALL, RELAXED, STRICT); added the flag `-peptf` for peptide filtering to the command line options
- 0.9.623 (18<sup>th</sup> November 2015) Made some preparations for planned source code switch to Java 8.
- 0.9.620 (16<sup>th</sup> November 2015) Location of <SpectraData> element works now for lower case and upper case peak list files.

- 0.9.619 (13<sup>th</sup> November 2015) - Added flag `-rpc` for reporting of ProCon as `<AnalysisSoftware>` in the `.mzid` file
- 0.9.618 (10<sup>th</sup> November 2015) Handling of Fasta lines without accession (i.e. containing only protein name)
- 0.9.616 (10<sup>th</sup> September 2015) Integrated the spectral counts converter; corrected error in Sequest / Comet conversion
- 0.9.610 (26<sup>th</sup> August 2015) Solved now the Spectrum-ID 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 [31]:

**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

- ProteomeDiscoverer<sup>®</sup> 1.x conversions of .msf files which reference .mzML peak list files produce output which is not validated by the ProteomeXchange [18, 19] / PRIDE [17] validation procedure.
- Support for the search Engine MSFit (<http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msfitstandard>) in ProteinScape<sup>®</sup> 1.3 converter is not yet implemented → an exception is 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 of the ProteinScape<sup>®</sup> 1.3 converter could be to let the user select the SepAndDig protocol, for which he wants to convert results.

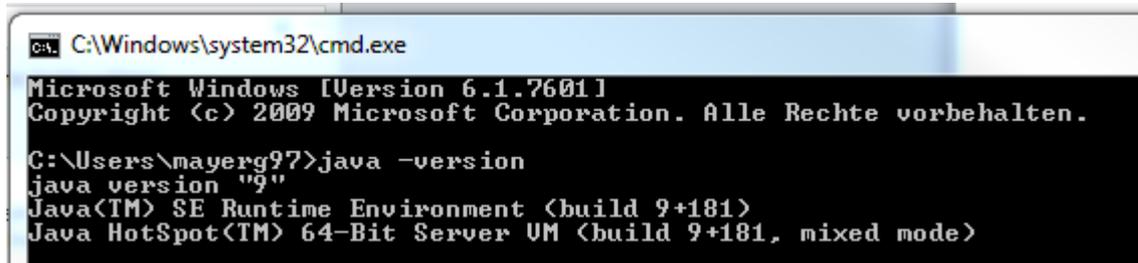
If you find further bugs please sent an email with the following log files:

- 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](mailto:mayerg97@rub.de) or [martin.eisenacher@rub.de](mailto:martin.eisenacher@rub.de) .

Please specify also:

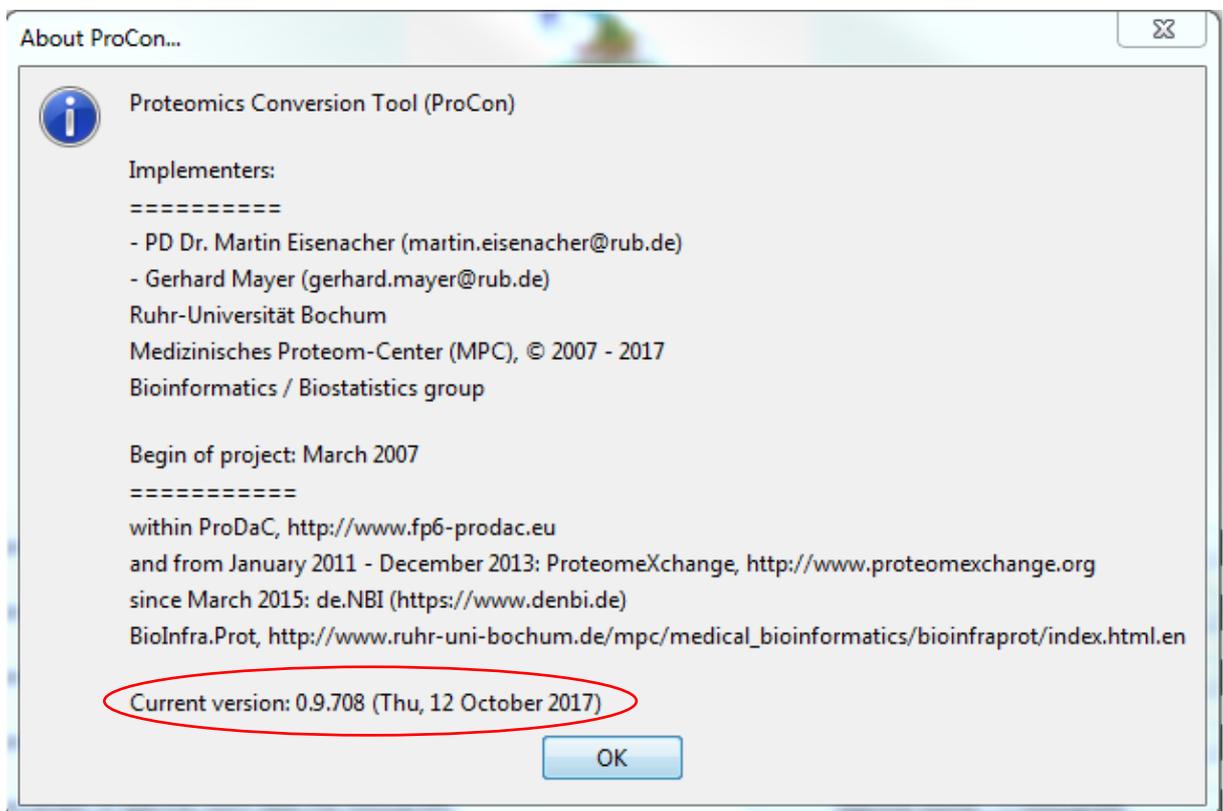
- your exact operating system (Windows or Linux), 32 or 64 bit
- the versions of your Java runtime (type `java -version` on your command interpreter `cmd.exe`)



```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Alle Rechte vorbehalten.

C:\Users\mayerg97>java -version
java version "9"
Java(TM) SE Runtime Environment (build 9+181)
Java HotSpot(TM) 64-Bit Server VM (build 9+181, mixed mode)
```

- the version of ProCon (see Help-About menu from ProCon)



- in case of ProteomeDiscoverer<sup>®</sup> 1.x 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)
- Integration into the KNIME [29, 30] workflow system

- Generating mzIdentML 1.2 [32, 33] output

Remark: Integration of ProteomeDiscoverer<sup>®</sup> 2.x support is not planned. The reason is that beginning with version 2.2 of ProteomeDiscoverer<sup>®</sup> offers an exporter for mzIdentML (a 60 day demo-version of PD2.2 can be downloaded from <https://portal.thermo-brims.com>).

## 16 Acknowledgements

The ProCon development was funded by the following projects: The **ProDaC** (<http://www.fp6-prodac.eu>) Coordination Action, 6<sup>th</sup> 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](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. The further maintenance is now funded by the **de.NBI** (<https://www.denbi.de>) project of the German Federal Ministry of Education and Research (BMBF) [grant number FKZ 031 A 534A].

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 Assoc* 2015;22:495-506.
2. Deutsch EW, Orchard S, Binz PA et al. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work, *J Proteome Res* 2017.
3. Jones AR, Eisenacher M, Mayer G et al. The mzIdentML data standard for mass spectrometry-based proteomics results, *Mol Cell Proteomics* 2012;11:M111.014381.
4. Blüggel M, Bailey S, Körting G et al. Towards data management of the HUPO Human Brain Proteome Project pilot phase, *Proteomics* 2004;4:2361-2362.
5. Walzer M, Qi D, Mayer G et al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics, *Mol Cell Proteomics* 2013;12:2332-2340.
6. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool, *Proteomics* 2013;13:22-24.
7. Creasy DM, Cottrell JS. Unimod: Protein modifications for mass spectrometry, *Proteomics* 2004;4:1534-1536.
8. Jones AR, Lister AL, Hermida L et al. Modeling and managing experimental data using FuGE, *OMICS* 2009;13:239-251.
9. Côté R, Reisinger F, Martens L et al. The Ontology Lookup Service: bigger and better, *Nucleic Acids Res* 2010;38:W155-160.
10. Montecchi-Palazzi L, Beavis R, Binz PA et al. The PSI-MOD community standard for representation of protein modification data, *Nat Biotechnol* 2008;26:864-866.
11. Mayer G, Jones AR, Binz PA et al. Controlled vocabularies and ontologies in proteomics: overview, principles and practice, *Biochim Biophys Acta* 2014;1844:98-107.
12. Mayer G, Montecchi-Palazzi L, Ovelheiro D et al. The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary, *Database (Oxford)* 2013;2013:bat009.
13. Jones P, Côté RG, Martens L et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community, *Nucleic Acids Res* 2006;34:D659-663.
14. Perkins DN, Pappin DJ, Creasy DM et al. Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis* 1999;20:3551-3567.
15. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J Am Soc Mass Spectrom* 1994;5:976-989.
16. Dorfer V, Pichler P, Stranzl T et al. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra, *J Proteome Res* 2014;13:3679-3684.
17. Vizcaino JA, Csordas A, Del-Toro N et al. 2016 update of the PRIDE database and its related tools, *Nucleic Acids Res* 2015;44.
18. Deutsch EW, Csordas A, Sun Z et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition, *Nucleic Acids Res* 2017;45:D1100-D1106.
19. Vizcaíno JA, Deutsch EW, Wang R et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination, *Nat Biotechnol* 2014;32:223-226.
20. Jarnuczak AF, Vizcaíno JA. Using the PRIDE Database and ProteomeXchange for Submitting and Accessing Public Proteomics Datasets, *Curr Protoc Bioinformatics* 2017;59:13.31.11-13.31.12.
21. Ternent T, Csordas A, Qi D et al. How to submit MS proteomics data to ProteomeXchange via the PRIDE database, *Proteomics* 2014;14:2233-2241.
22. Uszkoreit J, Maerkens A, Perez-Riverol Y et al. PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface, *J Proteome Res* 2015;14:2988-2997.

23. Patrickios CS, Yamasaki EN. Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory, *Anal Biochem* 1995;231:82-91.
24. A S. Isoelectric points of proteins: theoretical determination, *Anal Biochem*. 1989;179:319-325.
25. Grimsley GR, Scholtz JM, Pace CN. A summary of the measured pK values of the ionizable groups in folded proteins, *Protein Sci* 2009;18:247-251.
26. Perez-Riverol Y, Xu QW, Wang R et al. PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets, *Mol Cell Proteomics* 2016;15:305-317.
27. 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* 2013;12:3026-3035.
28. Bailey CM, Sweet SM, Cunningham DL et al. SLoMo: automated site localization of modifications from ETD/ECD mass spectra, *J Proteome Res* 2009;8:1965-1971.
29. Aiche S, Sachsenberg T, Kenar E et al. Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry, *Proteomics* 2015;15:1443-1447.
30. de la Garza L, Veit J, Szolek A et al. From the desktop to the grid: scalable bioinformatics via workflow conversion, *BMC Bioinformatics* 2016;17:127.
31. Mayer G, Stephan C, Meyer HE et al. ProCon - PROteomics CONversion tool, *J Proteomics* 2015;129:56-62.
32. Vizcaíno JA, Mayer G, Perkins S et al. The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics, *Mol Cell Proteomics* 2017;16:1275-1285.
33. Seymour SL, Farrah T, Binz PA et al. A standardized framing for reporting protein identifications in mzIdentML 1.2, *Proteomics* 2014;14:2389-2399.