AB SCIEX MS Data Converter User Guide

July 2011

AB SCIEX

71 Four Valley Dr., Concord, Ontario, Canada. L4K 4V8.

AB SCIEX LP is ISO 9001 registered.

Contents

This page intentionally left blank.

CRITICAL INFORMATION ABOUT THIS BETA SOFTWARE – PLEASE READ THIS!

Whether you are new to this tool or you have used a previous beta version, please be sure to read all text in red in this document carefully as it gives information about the beta aspects of this software.

Introduction

This document describes how to use the AB SCIEX MS Data Converter to convert mass spectral data from any AB SCIEX instrument into open data formats. Data can be converted into two different formats – mzML or MGF. You can also control what kind of information is written. You can choose to get an exact translation of what the instrument recorded or choose to convert to a processed version, reducing the data down to peak lists.

About the mzML Format and Conversion to It

The mzML format is the single XML standard mass spectrometry format that was created by the merger of the older mzXML and mzData formats. This converter encodes according to the version 1.1 specification for the mzML format. More information on mzML is available here:

http://www.psidev.info/index.php?q=wiki/Mass_Spectrometry

If you still require data in one of mzML's retired parent formats, you can use this tool to first convert data to mzML and then use publicly available tools to convert from mzML to a variety of other formats, including mzXML and mzData.

Be warned that, because XML is inherently verbose, the conversion of data in unreduced profile mode will result in mzML output files that are many times larger than the original raw data. It takes much more file space to convey the same amount of information with XML. Compression functions are used in this converter by default to reduce this effect, but the files will still be very large.

What about mzXML?

If you require mzXML format rather than mzML, we recommend you first convert to mzML using this converter and then use a public tool such as msconvert in the ProteoWizard package to convert mzML to mzXML. This tool is available here:

http://proteowizard.sourceforge.net/downloads.shtml

You may also be able to do direct conversion from wiff to mzXML using ProteoWizard, but be sure to compare the performance of mzXML files made by each of these two or other routes for your application of interest. ProteoWizard is currently not providing the same processing in ProteinPilot™ Software as is available in this tool. This difference may

be very important for proteomics applications and of lesser use to non-proteomics applications.

About the MGF Format and Conversion to It

The Mascot Generic Format (MGF) was created by Matrix Science as input for the Mascot search engine, but it has since become arguably the single most used open format for proteomics applications. More information on MGF is available here:

http://www.matrixscience.com/help/data_file_help.html

If you require a different peak list format, you can use this tool to convert to MGF and then use one of several publicly available tools to convert to other formats. However, if you require MS1 level data, you should convert to mzML as an intermediate because it can contain the survey level data. The MGF format produced by this tool only includes MS2 level data.

Improvements and Bug Fixes since the Revision 404 Beta Release

This release is still a beta version but there have been many improvements since the last beta release, revision 404:

- **Integration with PeakView™ Software Version 1.1** – If you install this tool properly, it will enable conversion of a wiff file to MGF format from a menu in PeakView Software. This support is currently limited to TripleTOF™ 5600 data and no other types of wiff data. See installation instructions.

- **Explicit file paths no longer required** – The previous beta would not work if the full file path was not included in the command line text. Normal command line tool behavior is to assume the local folder if file names are indicated without full paths. This now functions normally with one key caveat that Windows 7 SP1 prohibits writing to the Program Files or Program Files (x86) folder or any of its subfolders. Running a batch file from the installed path on this OS will result in an 'access denied' error message, since it is a subfolder of Program Files. Implied path assumption behavior is described in the relevant sections of the Command Line Syntax section.

- **Spectral indices now match between TOF/TOF profile and TOF/TOF peak list data** – If you converted the same TOF/TOF data to mzML format in both profile and either of the two peak list options using the previous beta version, the spectral indices assigned for the same spectra will not be the same for these two conversion approaches, meaning you were not be able to use these indices to map a centroided version of a spectrum to its unprocessed profile spectrum parent across these two files. This problem is now resolved, but be sure to note the software requirements for TOF/TOF usage.

- **Fragment peaks converted to +1 equivalent in MGF format** – Mascot prefers peaks lists be written as +1 equivalent m/z values, rather than the actual observed charge state in the case of multiply charged fragments. The exporting behavior has been changed for writing MGF files accordingly. The charge value in the third column still indicates the actual observed charge state.

- **Support of Windows 7 SP1** – SP1 previously caused failed conversions.

- **Multiple sample wiff files are supported correctly for mzML export** – Wiff files that contain multiple samples are exported as separate mzML files. A bug remains for MGF export where these files are exported as a single MGF file.

- **TOF/TOF mzML export now includes survey** – Only product spectra were included in the previous release without including their associated survey spectra.

Remaining Limitations and Known Bugs

This is beta software, and there are several important limitations and known issues you should be aware of:

- **MRM data are currently not converted correctly** – Do not use this tool to convert MRM data.

- **Centroid mode failures** – Centroid mode is not working correctly for wiff data. Regardless of this issue, we strongly advise using the higher quality ProteinPilot mode for peak list conversions for proteomics applications, rather than centroid.

- **TOF/TOF spot-based data not fully supported –** Only MS/MS data will be exported and exporting fidelity has not been verified. These limitations will be addressed in a future release.

- **PMF and other survey level data cannot be written to MGF –** Related to shortcomings in the previous point. This will be addressed in a future release.

- **TOF/TOF profile mode limit for reprocessed job runs –** Exporting reprocessed job runs to mzML will not include survey level data.

- **TOF/TOF profile data are not available for reprocessed job runs –** This is true for both survey and product spectra.

- **TOF/TOF parent runs with multiple product job runs cannot be exported to one file –** This is a limitation of the current syntax.

- **MGF currently can only be written to single precision –** The double precision argument will be ignored for writing to MGF format.

Software Requirements

Operating System and Supporting Components

This software is tested on Dell computers running Microsoft Windows XP with SP3, Windows XP Professional with SP2 64-bit system, and Windows 7 SP1 64-bit system. ProteinPilot software does not need to be installed. However, Analyst® Software, PeakView Software, and BioAnalyst® Software do not officially support Windows 7, and they are not currently known to work on Windows 7. If you want to use this converter on the Windows 7 64-bit operating system, you will also need to install the Microsoft *Visual C++ 2008 SP1 Redistributable Package (x86)* download from the Microsoft Download Center:

http://www.microsoft.com/downloads/en/details.aspx?FamilyID=a5c84275-3b97-4ab7-a40d-3802b2af5fc2

This same component should also allow function on Windows Server 2008.

You must have Microsoft .NET version 3.5 SP1 (or higher) installed. If you are using this software on a system already running ProteinPilot software version 4.0, you should already have this .NET version, as it installs if not already present. This converter's installer does not install it if not already present. If you do not have this .NET version, you can download it from the Microsoft website here:

http://www.microsoft.com/downloads/en/details.aspx?FamilyID=d0e5dea7-ac26-4ad7-b68c-fe5076bba986

PeakView Software version 1.1 is required to use the plug in function provided by this converter.

Installation

To install the AB SCIEX MS Data Converter, first do all necessary component installations and updates described in the previous section.

1. If you used a previous beta version of the converter, you must uninstall the previous version first. Do this using the normal control panel for adding and removing programs.

2. If you are using PeakView Software version 1.1 and want to use MGF conversion option enabled by integration with this converter, update PeakView Software to version 1.1, if using an older version. Close PeakView Software if it is running.

3. Install the new converter by opening the installer folder, finding the setup.exe, and double clicking it to launch the installer.

Converting to MGF Format Using Integration with PeakView Software

When you install this converter, it will install a plug in to PeakView Software version 1.1 that enables you to convert wiff files to the MGF format. To convert a file:

1. Open the wiff file of interest in PeakView Software.

2. Select *Process -> Create MGF File* to initiate file conversion.

3. The software will open an MS DOS window while the conversion proceeds and will close when the conversion is done.

4. The resulting MGF file will be written in the same path as the parent wiff file and will be automatically named via the convention <original file name.MGF>.

The conversion to MGF uses the -proteinpilot peak list option described in the Command Line Arguments section. Multiple conversions can be conducted at the same time. If you have a multi-core system, each conversion will run on a separate core, providing a speed advantage. Once you initiate a conversion, you can close PeakView Software and the conversion process will continue.

Converting MS Data to mzML or MGF Formats by Command Line

The following section describes how to use this tool by constructing a batch file. If you are unfamiliar with command line tools, this approach is recommended. If you are familiar with this type of tool, you will know where you can deviate from this procedure.

1. After installation is complete, find the folder where the software was installed, which is:

   C:\Program Files\AB SCIEX\MS Data Converter

   Or on a 64bit system:

   C:\Program Files (x86)\AB SCIEX\MS Data Converter

   Or an analogous path, depending on your choice during installation.

2. Create a new text file in this folder and name it "**Example batch file.bat**". Your computer may be set so that file extensions are not shown for known file types. Set the folder to *View -> Details* and be sure the *Type* for the file you created is listed as *MS-DOS Batch File*, not *Text Document*. If the latter is shown, you need to change the folder options so file extensions are not hidden and change the extension. In Windows XP, do this by going to *Tools -> Folder Options -> View* and unchecking the *Hide*

extensions for known file types option. You will see the full file name is **Example batch file.bat.txt**. Remove the .txt extension to make the real extension .bat.

3. Open the batch file in a text editor such as Notepad or WordPad to enter the command line instructions as detailed in the command line syntax section. Adding a *Pause* command on the next line of the batch file is recommended. This will keep the command window open, which will allow you to see error messages in the event of incorrect syntax.

4. Once the correct syntax is constructed, save the batch file, close it, and then run the conversion by double clicking on the closed file.

For example, a batch file with the following content:

```
AB_SCIEX_MS_Converter  WIFF "D:\Data\File A.wiff" -proteinpilot MGF "D:\Data\File A.mgf"

Pause
```

will convert this wiff file into an MGF file by the same name other than extension, where the peak list written will be the exact peak list that would be searched by ProteinPilot software (same peak list for Mascot or Paragon™ algorithm search). See the next section for details on all command line arguments and options.

Conversion of a Set of Multiple Files

In the example in the previous section, a single file is converted using a DOS batch file (a specific technical use of the term not to be confused with the more generic meaning of *batch*, a set of things). If you want to do multiple conversions, you can also construct a single batch file to do this, and, you can get a speed gain on a multi-core computer if you construct the batch to start each conversion operation as a separate instance of the converter tool. This will allow each conversion to run in parallel on a separate core. For example, this would allow you to process 8 conversions on an 8 core computer in roughly the same time it would take to do a single conversion. Single conversions currently cannot be distributed over multiple cores. To convert multiple files in parallel, you need to make only one small change to the batch file syntax, as shown below:

```
start /i AB_SCIEX_MS_Converter WIFF "D:\File A.wiff" -proteinpilot MGF "D:\File A.mgf"

Sleep 10


start /i AB_SCIEX_MS_Converter WIFF "D:\File B.wiff" -proteinpilot MGF "D:\File B.mgf"

Sleep 10


start /i AB_SCIEX_MS_Converter WIFF "D:\File C.wiff" -proteinpilot MGF "D:\File C.mgf"

Sleep 10


start /i AB_SCIEX_MS_Converter WIFF "D:\File D.wiff" -proteinpilot MGF "D:\File D.mgf"

Sleep 10


Pause
```

This batch file will initiate four separate conversions with a 10 second gap between starting each subsequent operation. Each conversion will run on a separate core if the computer has at least 4 cores.

Command Line Syntax

The generic command line syntax to use the tool is:

> **AB_SCIEX_MS_Converter <input format> <input data> <output content type> <output format> <output file> [data compression setting] [data precision setting] [create index flag]**

Square brackets indicate optional elements of the syntax. Full file paths are required to indicate any location outside the folder the batch file is run from. Use quotes to enclose any filename or full path having any spaces, as shown in the example above. The details on all other arguments in the syntax are described below.

INPUT FORMAT ARGUMENT

There are only two options.

WIFF	Use for data from all instruments producing any kind of wiff file. (all AB SCIEX quadrupole-containing instruments)
TOFTOF	Use for all data from AB SCIEX TOF/TOF instruments.

INPUT DATA ARGUMENT

Input data is specified in two different ways depending on whether it is wiff data or TOF/TOF origin data. The syntax used is the same as used to script searches in ProteinPilot Software.

For wiff data:	For wiff input data the filename and extension are indicated, and the path is generally included. For example: "C:\Data\datafilename.wiff" The path may be omitted if the input file is in the same folder as the executable. For wiff data types having multiple file components (.scan and .mtd, depending on version), only indicate the .wiff file. The other file components must be in the same folder, but they need not be indicated via the command line syntax. Conversion will fail if all necessary file components are not in the same folder.
For TOF/TOF data:	For TOF/TOF input data, you must indicate the location in the Oracle database via the following syntax: "server\project\subproject\...\spotset_name\MSMS_job_run\ chromatogram_name\MS_job_run" See Figure 1 and the example below using the Add TOF/TOF Data dialog box in ProteinPilot Software to help construct the required information. Each chromatogram is a separate DATA element in the parameter file. ***Note:*** MSMS and MS should contain only the number, without the parentheses or time.

For the two TOF/TOF data examples shown in Figure 1 below, the input data would be indicated as follows:

"4800\ProteinPilot Getting Started Guide\iTRAQ Reagents - 8Plex 8 Protein Mix\3\1\2"

"4800\Don\Project Work\mTRAQ samples 080731\10\mTRAQ Fr4\7"

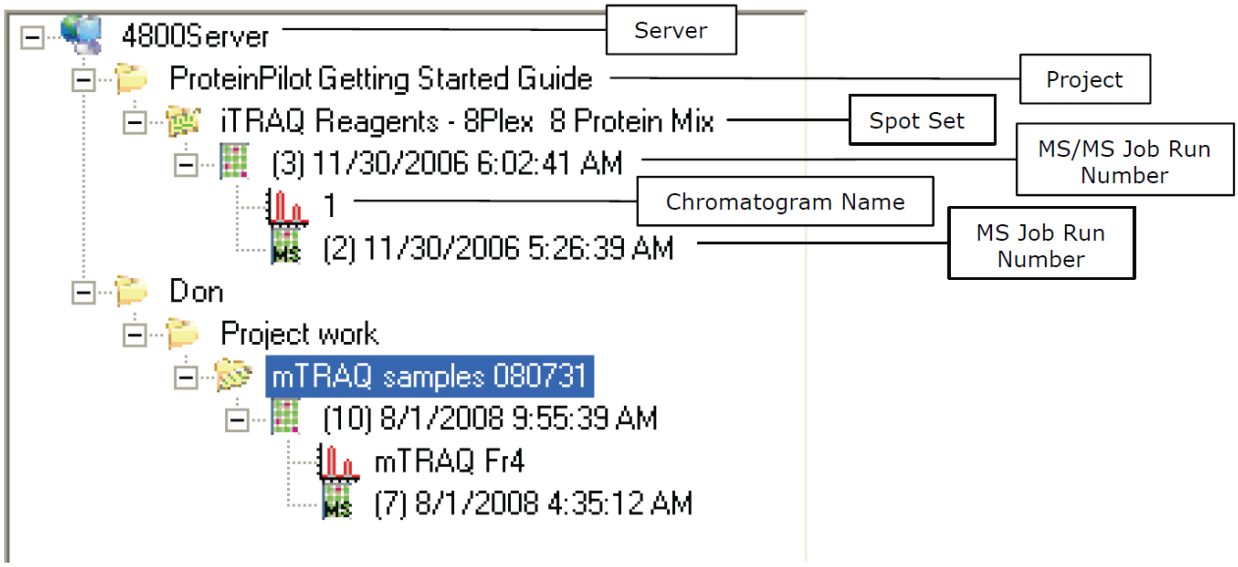

Figure 1 – Selected Data list in the Add TOF/TOF dialog box in ProteinPilot Software

OUTPUT CONTENT TYPE ARGUMENT

There are three options. Be sure to note that a dash must precede this argument.

-profile	Gives the full x-y trace recorded by the instrument without any reduction in information content. This mode cannot be used to convert to the MGF format.
-centroid	This option gives the centroided peak list as determined by the instrument's software in real time during acquisition.
-proteinpilot	This option gives the exact peak list that would be searched by ProteinPilot Software. In the case of TOF/TOF data, the centroid data is what is searched by this software, so they are identical. In the case of wiff data, the raw data are processed by a slower but higher quality signal processing approach that produces better results for protein identification applications.

OUTPUT FORMAT ARGUMENT

There are two options for this argument.

MGF	Converts to the Mascot Generic Format. Only MS/MS spectra are converted. The *-profile* mode may not be used to convert to the MGF format, as this is a peak list format.
MZML	Converts to the mzML format. All data levels are converted (MS1, MS2… MSn). Conversion of MRM data is also supported.

OUTPUT FILE ARGUMENT

Normally the full file path, file name, and file extension are indicated. For example:

"C:\Data\ouputfilename.mgf" or "C:\Data\datafilename.XML"

The path can be omitted, and the output will go to the same folder as the executable. However, beware that Windows 7 SP1 prohibits writing to any subfolder in Program Files, which would generally include the installation folder for this tool. Thus, you cannot use implied path for output files with this operating system and service pack. Presumably, updates from Microsoft will also have this limitation.

OPTIONAL MODIFIERS

Up to three optional modifiers may be used, providing explicit control over the use of compression in the data, the encoding precision of the data, and whether or not an index is created.

/nocompression	Stores the binary arrays without using any compression. This option is mutually exclusive with the */zlib* option. If specified for MGF format export, this will be ignored.
/zlib	Compresses the binary arrays using zlib algorithm. This option is the default behavior if nothing is indicated. It is mutually exclusive with the */nocompression* option. If specified for MGF format export, this will be ignored.
/singleprecision	Outputs binary data using 32-bit float single precision. This option is mutually exclusive with the */doubleprecision* option.
/doubleprecision	Outputs binary data using 64-bit float double precision. This option is the default behavior if nothing indicated. It is mutually exclusive with the */singleprecision* option.
/index	Writes the index for mzML data. Please, refer to

	http://www.peptideatlas.org/tmp/mzML1.1.0_idx.html and http://psidev.cvs.sourceforge.net/*checkout*/psidev/psi/psi-ms/mzML/schema/mzML1.1.1_idx.xsd. This option only applies to the mzML format. If specified for MGF format export, this will be ignored.

Fine Details on Conversion Output

There are a number of fine details in exactly what data are written to the files produced by this tool. This section attempts to capture these fine points that may matter greatly to some users.

MGF Output

The data written during conversion to MGF format obeys the MGF specification as much as possible. There are few fine details not clearly prescribed by this specification. When writing the MS/MS fragments for a spectrum, there are three columns of data provided by this tool to describe fragment peaks:

Column 1	The m/z value of the +1 charge state of a fragment, regardless of what charge state is actually observed. Mascot prefers searching fragments in their +1 equivalent m/z, rather than the observed m/z. If the charge state is unknown, the observed m/z is listed.
Column 2	The intensity of the fragment peak measured as the apex intensity of the monoisotopic peak. These are real measured intensity, not normalized intensity measurements.
Column 3	The charge state at which the fragment was actually observed. Fragment ions are listed as their corresponding +1 charge state m/z and this charge value allows for the reconstruction of the observed spectrum. If a given fragment is observed at multiple charge states in a spectrum, there will be multiple rows with very similar m/z values and different charge states. If the charge cannot be determined as is the case with low resolution data, a value of 0 is listed.

mzML Output

There are many fine points on the nature of the data written to mzML, and since this converter offers three modes of conversion rather than the usual profile vs. centroid, it is important to be clear about what data are included in each case.

Profile mode	Profile data are written as they are measured by the instrument, and for

	TOF analyzers, this is not regularly spaced.
Centroid mode	Unlike in the MGF export where fragment peaks are converted to their +1 charge equivalent m/z, the m/z values are written as the actual observed m/z.
ProteinPilot peak list mode	Unlike in the MGF export where fragment peaks are converted to their +1 charge equivalent m/z, the m/z values are written as the actual observed m/z.