Employing ProteoWizard to Convert Raw Mass Spectrometry Data

Jerry D. Holman,[1] David L. Tabb,[1] and Parag Mallick[2]

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee
[2]Department of Radiology, Stanford School of Medicine, Palo Alto, California

ABSTRACT

After raw data have been captured by mass spectrometers in biological LC-MS/MS experiments, they must be converted from vendor-specific binary files to open-format files for manipulation by most software. This protocol details the use of ProteoWizard software for this conversion, taking format features, coding options, and vendor particularities into account. This protocol will aid researchers in preparing their data for analysis by database search engines and other bioinformatics tools. *Curr. Protoc. Bioinform.* 46:13.24.1-13.24.9. © 2014 by John Wiley & Sons, Inc.

Keywords: LC-MS/MS • mzML • database search • raw files • proteomics

INTRODUCTION

Tandem mass spectrometry data sets are captured to binary files or databases by the software controlling the instruments. The ProteoWizard library and tools are designed to extract data from these proprietary formats for export in community standard formats or for direct access via its API (Kessner et al., 2008). It is distinctive for its support of instruments from many instrument vendors (Chambers et al., 2012).

The three protocols included here are intended to complement each other. Basic Protocol 1 is intended for first-time users of ProteoWizard who feel most comfortable with graphical user interfaces, while the Alternate Protocol will assist researchers who are comfortable in a command-line environment. Both of these protocols address the process of conversion from instrument vendor-specific raw data formats (Table 13.24.1) to mzML (Martens et al., 2011), mzXML (Pedrioli et al., 2004), mz5 (Wilhelm et al., 2012), or MGF (*http://www.matrixscience.com/help/data_file_help.html*) format. Basic Protocol 2 is intended to assist researchers who need to convey their data to search engines requiring simpler text formats.

TRANSCODING MS DATA FROM RAW FORMAT VIA MSConvert GUI

BASIC PROTOCOL 1

LC-MS/MS instruments collect data into binary files that are not easily read or manipulated. Most instrument vendors, however, have provided software libraries that allow access to data within these files on Microsoft Windows operating systems. ProteoWizard includes the MSConvert GUI tool to enable easy translation of these raw data to a variety of formats. Most users will opt to perform some level of data filtering at this step, such as reporting peak lists rather than peak profiles in the resulting files.

Necessary Resources

Software

Researchers will need to install ProteoWizard on their computers prior to attempting this protocol, which was developed using version 3.0.5471. The

Current Protocols in Bioinformatics 13.24.1-13.24.9, June 2014
Published online June 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi1324s46

Table 13.24.1 MSConvert-Supported Data Formats

Vendor/creator	Format
AB SCIEX	WIFF; T2D (with DataExplorer)
Agilent	MassHunter (.d directories)
Bruker	FID; .d directories; XMASS XML
Thermo	RAW
Waters	raw directories
HUPO PSI	mzML
ISB Seattle Proteome Center	mzXML
Matrix Science	MGF
Yates/MacCoss Laboratories	MS2/CMS2/BMS2
Steen & Steen Laboratory	mz5

software is available from *http://proteowizard.sourceforge.net/*. Conversion from native vendor formats requires the use of Microsoft Windows. The GUI version of msConvert demonstrated here can only be run under Microsoft Windows. While many of the software libraries for data access employ 32-bit code, one may deploy ProteoWizard on 64-bit versions of Microsoft operating systems. If converting from one open format to another (e.g., mzXML to mzML) one may use the command line version described below on Microsoft Windows, Linux, or Macintosh platforms.

1. From the ProteoWizard folder, execute MSConvertGUI. The graphical user interface should appear (see Fig. 13.24.1).
2. In the top-left corner click the Browse button to bring up a source selection dialog. This may take a few seconds to load depending on the size of the initial directory.
3. Using the left panel of the dialog (see Fig. 13.24.2), navigate to the directory containing the source (raw files or directories) you wish to convert. Currently msConvert supports the formats listed in Table 13.24.1. The software automatically scans the selected directory for any files or folders which can be read as sources. It is possible to show only sources of a certain type by filtering from the drop down menu at the bottom of the Open Data Source dialog. Select multiple files as you would in Windows Explorer by holding down the Ctrl key while clicking, holding down Shift while clicking the ends of a range of files, or simply dragging a box over the files you want to select.
4. Once all sources have been selected, click "Open." If only one source was selected, it will appear in the text box beside the browse button and must be added to the list by clicking "Add" right underneath. If multiple sources were selected, they will be added to the list automatically.
5. Once at least one source has been added to the list, the "Output Directory" box will automatically be filled in if it had been left blank. If output files are to be separate from the original source location, click the browse button and navigate to the desired output folder.
6. Select the output format for the converted files. By default files will be written using the format's standard extension; however, the utilized extension can be changed using the text box next to the format selection box. This will not change the contents of the output file but will simply change the extension used. This feature can be useful for

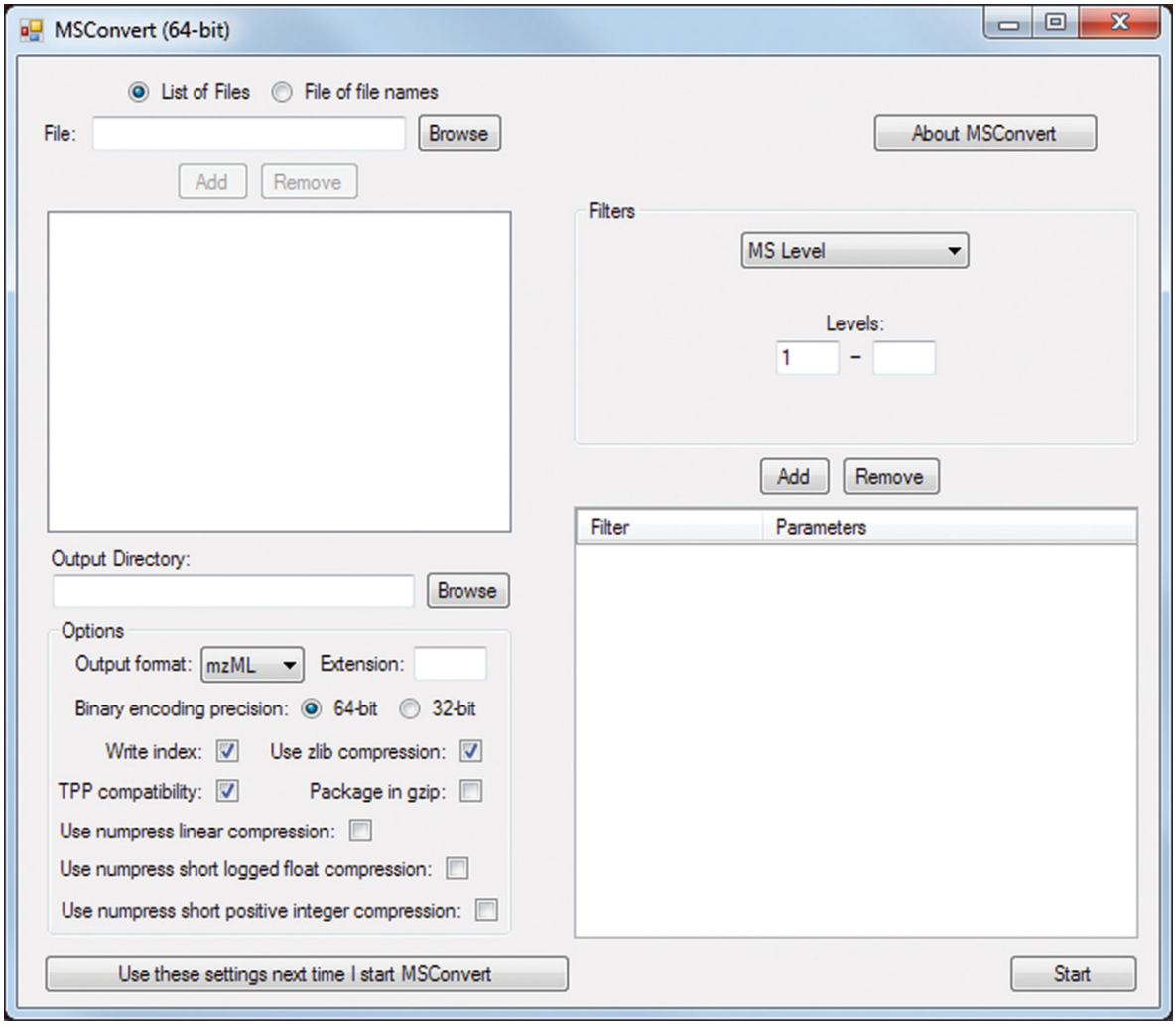

Figure 13.24.1 The main MSConvert graphical user interface.

keeping track of filters used in multiple conversions of the same set of sources within the same output folder.

7. MSConvertGUI is capable of using a number of different compression methods when converting files. The default options should be sufficient for normal use; however, if a specific type of compression or writing method is required, it can be enabled using the checkboxes located below the format selection box.

8. The GUI for MSConvert allows users to apply a subset of available filters while converting files (see Table 13.24.2). These can be accessed by selecting the desired filter from the drop-down box in the filter area. Once selected, a number of options will appear. Fill in the options as desired and click the button below the filter area labeled "Add" to add it to the list. Filters can be removed from the list simply by selecting the unwanted filter and clicking "Remove." For example, to employ peak picking of all MS levels, one can select "Peak Picking" from the drop-down box and then click "Add."

   Data filters in Table 13.24.2 enable users to control which data appear in translated data files and how they are recorded there. Filters marked with an asterisk are available in the GUI. All other filters are available through the command line version of MSConvert

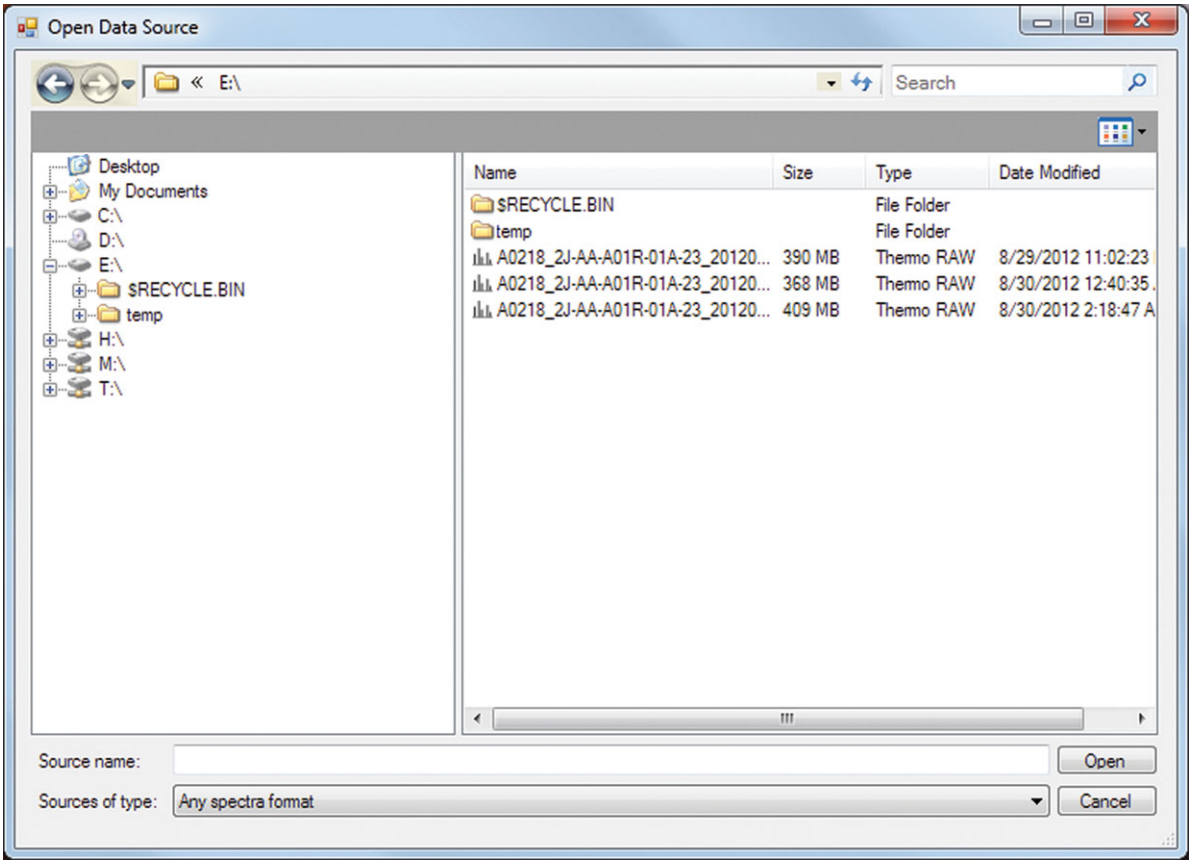

Figure 13.24.2 The data source selection dialog.

(described in the Alternate Protocol below). To view a full list of options and filters for the command line version of MSConvert, execute the program with the argument `"--help"`.

9. Once all options have been filled, click "Start." A new window will appear with a list of files to be converted and current progress. Details on the conversion of the currently selected file will be shown in the text box at the bottom of the new menu. Once all files have finished converting, it is safe to close the progress window and use the resulting files.

ALTERNATE PROTOCOL

TRANSCODING MS DATA FROM RAW FORMAT VIA MSConvert

Command line tools may be more convenient for researchers who want to streamline raw file management by creating batch scripts. This protocol illustrates how to accomplish this with the fully featured "msConvert" executable included in ProteoWizard.

Necessary Resources

Software

The same necessities found in Basic Protocol 1 apply to this protocol, as well. Notably, some attempts to run command line ProteoWizard conversion under WINE (*http://www.winehq.org*) emulation have been successful, enabling users to convert vendor files within Linux environments.

1. Locate the path to ProteoWizard executable software. This can either be added to the "PATH" variable for command line operations, or one can simply write the explicit path to the msConvert binary, e.g., `"C:\Program Files (x86)\ProteoWizard\ProteoWizard 3.0.5471\msconvert.exe"`.

Table 13.24.2 Data Filters Available Within ProteoWizard

Filter	Parameters	Description
index	<index_value_set>	Selects spectra by index—an index value 0—based numerical order in which the spectrum appears in the input.
*msLevel	<mslevels>	Selects only spectra with the indicated MS levels.
*chargeState	<charge_states>	Keeps spectra that match the listed charge state. Use 0 to include spectra with no charge state at all.
precursorRecalculation		Recalculates the precursor *m/z* and charge for MS2 spectra based on the MS1 data. Only works on orbitrap and FT data.
precursorRefine		Recalculates the precursor *m/z* and charge for MS2 spectra based on the MS1 data. Works on Orbitrap, FT, and TOF data.
*peakPicking	<prefer_vendor>, <ms_levels>	Performs centroiding on spectra with the selected MS levels. If used with other filters, this must be set first.
*scanNumber	<scan_numbers>	Selects spectra by scan number.
scanEvent	<scan_event_set>	Selects spectra by scan event.
*scanTime	<scan_time_range>	Selects spectra within a given time range.
sortByScanTime		Reorders spectra, sorting them by ascending scan start time.
stripIT		Rejects ion trap data spectra with MS level 1.
metadataFixer		Add or replace a spectra's TIC/BPI metadata, usually after peak picking where the change from profile to centroided data may make the TIC and BPI values inconsistent with the revised scan data.
titleMaker	<format_string>	Adds or replaces spectrum titles according to format string given.
*threshold	<type>, <threshold>, <orientation>, [<mslevels>]	Keeps data whose values meet various threshold criteria. Details for the different criteria can be found in the GUI tooltips.
*mzWindow	<mzrange>	Keeps *m/z* intensity pairs whose *m/z* values fall within the specified range.
mzPrecursors	<precursor_mz_list>	Retains spectra with specific precursor *m/z* values.
defaultArrayLength	<peak_count_range>	Keeps only spectra with peak counts within a given range.
*zeroSamples	<mode>, [<MS_levels>]	Deals with zero values in spectra—either removing them, or adding them where they are missing.
mzPresent	<tolerance>, <type>, <threshold>, <orientation>, <mz_list>, [<include_or_exclude>]	Keeps data whose values meet various threshold criteria. Contains a wider array of options than the "threshold" filter.
MS2Denoise	[<peaks_in_window>, [<window_width_Da>, [multi-charge_fragment_relaxation]]]	Noise peak removal for spectra with precursor ions.

continued

Table 13.24.2 Data Filters Available Within ProteoWizard, *continued*

Filter	Parameters	Description
MS2Deisotope	[<hi_res>, [<mz_tolerance>]]	Deisotopes MS level 2 spectra using Markey method.
*ETDFilter	[<removePrecursor>, [<removeChargeReduced>, [<removeNeutralLoss>, [<blanketRemoval>, [<matchingTolerance>]]]]]	Filters ETD MSn spectrum data points, removing unreacted precursors, charge-reduced precursors, and neutral losses.
chargeStatePredictor	[<overrideExistingCharge>, [<maxMultipleCharge>, [<minMultipleCharge>, [<singleChargeFractionTIC>, [<algorithmMakeMS2>]]]]]	Predicts MSn spectrum precursors to be singly or multiply charged depending on the ratio of intensity above and below the precursor *m/z*.
*activation	<precursor_activation_type>	Keeps only spectra whose precursors have the specified activation type.
analyzer	<analyzer>	Keeps only spectra with the indicated mass analyzer type.
polarity	<polarity>	Keeps only spectra with scan of the selected polarity.

2. For each filter to be included, add text structured like the following: `--filter "peakPicking true 1-"`.

   The first piece signals that a new data filter is being added. The section in double quotes supplies the name of the filter (taken from Table 13.24.2) and the configuration options for it.

3. Specify the output format options, such as specifying the overall format by using one of the following flags: `"--mzML", "--mzXML", "--mz5", "--mgf"`. If subsequent software allows it, specify `"zip"` encoding for data within the files through use of the `"-z"` option. The full list of these options can be found by running the msConvert executable without any parameters. When a complex set of parameters needs to be applied to a conversion, enumerate the options in a text file, invoking those options with the `-c` option, as in `"msconvert data.RAW -c config.txt"`, where `config.txt` contains options like this:

```
# example config.txt file

mzXML=true

zlib=true

filter="index [3,7]"

"filter=precursorRecalculation"
```

4. Specify which files are to be converted. Both absolute and relative paths are permitted, and wildcard characters such as `'*'` and `'?'` can be used to process multiple files in a single pass.

5. Combine these elements into a single command line, such as the following:

```
"C:\Program Files (x86)\ProteoWizard\ProteoWizard
3.0.5471\msconvert.exe" --filter "peakPicking true 1-"
--mz5
-z *.raw
```

CONVERTING mzML DATA TO SIMPLE TEXT FORMATS FOR SEARCH ENGINES

Once data have been translated to an open format, an additional step may be necessary to prepare data for handling by some database search engines. In particular, older versions of Sequest (Eng et al., 1994) and of Spectrum Mill (*http://www.chem.agilent.com/en-US/products-services/Software-Informatics/Spectrum-Mill/Pages/default.aspx*) may need MS/MS scans to be presented as individual text files in DTA or PKL format. This protocol details the process for this conversion step.

Necessary Resources

Software

This protocol depends upon mzXML2Search, a tool available as part of the Trans-Proteomic Pipeline (*http://sourceforge.net/projects/sashimi/*). The step employing mzXML2Search may be executed under Windows or Linux, as the XML-based formats do not require the instrument vendor libraries. Despite the name of the software, mzXML2Search may be applied to both mzML files and mzXML files.

1. Determine whether PKL, MGF, or DTA files are required by the search engine to be employed. Correspondingly, replace `"[option]"` in the command line below with `"-pkl"`, `"-mgf"`, or `"-dta"`.

   NOTE: *MGF format can be exported directly by msConvert via Basic Protocol 1.*

   Users of DTA and PKL file formats should keep in mind that exporting each spectrum as a separate text file can take up considerable space on a file system. Newer branches of classic database search algorithms (such as Comet, from the University of Washington; Eng et al., 2013) can work directly with mzML format or concatenated files rather than requiring the creation of thousands of text files.

2. Next, determine which source format is available; the software is applicable to mzML or mzXML. In the command line below, `"[format]"` should be replaced by `"*.mzML"` or `"*.mzXML"`.

3. Execute mzXML2Search with this command line: `"mzXML2Search [option] [format]"`.

GUIDELINES FOR UNDERSTANDING RESULTS

In most cases, file transcoding is a relatively rapid step. Frequently, these conversions can be conducted in <1 min for each LC-MS/MS experiment. The run time of conversion, however, may be lengthened by some of the available filters, particularly calculation-heavy options.

In some cases, raw data may have become garbled. This is generally true of "mis-injections," which produce raw file stubs. Garbles may also occur as a file is transmitted from computer to computer. For some types of errors, msConvert is able to return an error message that indicates the nature of the fault. When converting a large set of raw data, it is generally advisable to determine whether or not the output files are equal in number to the input files.

In most cases, one can expect that the files generated by conversion will be larger than the raw data used as sources. In the case of XML-based formats, this is generally due to the explicit labeling of each element in the file; where a vendor-format binary file would store a double-precision floating point number for a peptide *m/z* value, e.g., the XML file might store the number plus a label for what that number represents. The use of zip compression and efficient database formats can greatly reduce expected file sizes (Wilhelm et al., 2012).

COMMENTARY

Best practices by instrument vendor

The ProteoWizard project has attempted to support all biological mass spectrometer vendors equally, but the completeness of implementation for each vendor can vary. Support for Thermo and Agilent instruments is quite comprehensive. Bruker support is quite robust, but producing peak profiles rather than peak lists is currently problematic for some instrument models. AB SCIEX produces two challenges; the software library does not allow msConvert to infer precursor charges, and only the basic peak centroider is available for peak-listing. Researchers may instead opt to use the free AB SCIEX MS Data Converter (see Internet Resources), which can address both these challenges. Waters instruments have posed long-standing challenges for data access. At present, the data access layer implemented in ProteoWizard cannot offer access to Waters peak-listing, precursor charge state inference, or spectral summation functions. A project nearing completion in the Tabb Laboratory should soon address these challenges by the implementation of open-source alternatives to this functionality in ProteoWizard. Many users may find that export to database search engines is feasible through the use of the Waters Protein Lynx Global Server.

Metadata content among file formats

Though files in open formats are often required for the vast majority of downstream operations, it is important to archive vendor data files. Typically, vendor files contain extensive metadata about instrument operating parameters. For example, Thermo RAW files contain several hundred parameters (voltages, pressures, flow rates, etc.) that are not typically extracted. Notably, the ProteoWizard programs ThermoRawMetaDump and msAccess are able to export these parameters. These metadata may be used in conjunction with the mass spectra themselves to evaluate instrument performance more comprehensively.

Importantly, the mzML data standard was designed to fully hold nearly any piece of metadata. However, the data access APIs used by msConvert are not able to access all of it for all of the vendor files. Consequently, mzML is not a complete transformation of a vendor file. Beyond mzML, formats described in Basic Protocol 2 contain almost no metadata and should not be considered archival formats.

Options for file size reduction

As noted above, conversion from a vendor format to an open format may lead to an increase in file size. ProteoWizard provides several filters which may be used to reduce file size. These options may alter your data, thus impacting downstream processing, and should be used carefully. The most common size-reduction technique employed is centroiding (peak-picking). High-resolution MS data often sample each peak multiple times. Consequently, a given peak will be represented by several data points. This sampling provides detailed information about peak shape and can be useful for de-convolving overlapping peaks. However, these data can consume considerable space. Centroiding operations attempt to determine the center of a peak and to then compress the data by including only a single measurement for each peak, instead of several data points. Centroiding should be performed carefully, as it may inadvertently shift the inferred *m/z* of a peak. In addition, it may impact quantification. As discussed above, the "peakPicking" option of msConvert gives you the option of using either vendor algorithms for this purpose (when available) or the ProteoWizard internal function. Centroiding is a lossy operation.

In addition to centroiding, other space savings may come from writing files using 32-bit numbers and zlib compression for *m/z* and intensity values. Though most software can handle zlib-compressed data, some software packages lack this ability. Consequently, one should test one's pipeline with zlib-compressed data before batch converting a large number of files. These two operations are not lossy.

To use these options on the command line, one would execute:

`msconvert <input_file> --32 --zlib` msConvert also supports the Levander lab's msNumPress Library, which can be executed using command line flags such as `–numpressAll`. Compression with numpress is lossy.

Additional options for altering your data to decrease file size include the `--filter "zeroSamples removeExtra"` option. When `<mode>` is `"removeExtra"`, consecutive zero intensity peaks are removed from spectra. For example, a peak list "100.1,1000 100.2,0 100.3,0 100.4,0 100.5,0 100.6,1030" would become "100.1,1000 100.2,0 100.5,0 100.6,1030" and a peak list "100.1,0 100.2,0

100.3,0 100.4,0 100.5,0 100.6,1030 100.7,0 100.8,1020 100.9,0 101.0,0" would become "100.5,0 100.6,1030 100.7,0 100.8,1020 100.9,0." This particular data processing is relatively innocuous for most applications. More aggressive filter options include denoising and thresholding. These operations explicitly discard data and should be used cautiously.

While this protocol addresses the use of msConvert and its GUI counterpart for file format handling, incorporating ProteoWizard directly in software as a reader of mass spectrometry formats is very powerful. Software that adopts this strategy will gain the ability to read any format that ProteoWizard can, thus limiting the need to convert formats to those cases where operating system requirements (such as processing on a Linux cluster) are necessary. Simply employing msConvert and msConvertGUI as described in this protocol, however, greatly reduces the challenges for handling mass spectrometry data.

Acknowledgments

D.L.T. was supported by U01 CA152647, while J.D.H. was supported by U24 CA159988. P.M. was supported by CCNET U54 CA151459, PSOC-MCSTART U54 CA143907, and by the Canary Foundation.

Literature Cited

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., and Mallick, P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 30:918-920.

Eng, J.K., McCormack, A.L., and Yates, J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5:976-989.

Eng, J.K., Jahan, T.A., and Hoopmann, M.R. 2013. Comet: An open-source MS/MS sequence database search tool. *Proteomics* 13:22-24.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. 2008. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* 24:2534-2536.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.A., and Deutsch, E.W. 2011. mzML–A community standard for mass spectrometry data. *Mol. Cell. Proteomics* 10:R110.000133.

Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., and Aebersold, R. 2004. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 22:1459-1466.

Wilhelm, M., Kirchner, M., Steen, J.A.J., and Steen, H. 2012. mz5: Space- and time-efficient storage of mass spectrometry data sets. *Mol. Cell. Proteomics* 11:O111.011379.

Key Reference

Chambers et al., 2012. See above.
Reporting updates during the first 5 years of the ProteoWizard project, this manuscript broadly summarizes the capabilities of these tools and libraries

Internet Resources

http://proteowizard.sourceforge.net/
ProteoWizard is available from this Web site.

http://www.absciex.com/downloads/software-downloads
AB SCIEX provides the AB SCIEX MS Data Converter.