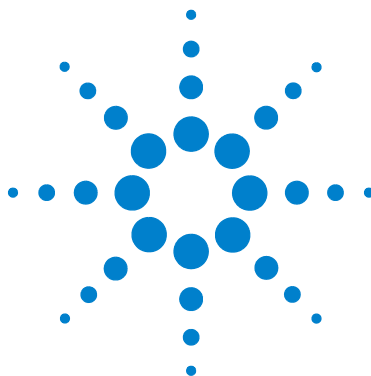




**Agilent
G2721AA/G2733AA
Spectrum Mill MS
Proteomics Workbench**

Application Guide



Agilent Technologies

Notices

© Agilent Technologies, Inc. 2012

No part of this manual may be reproduced in any form or by any means (including electronic storage and retrieval or translation into a foreign language) without prior agreement and written consent from Agilent Technologies, Inc. as governed by United States and international copyright laws.

Manual Part Number

G2721-90037

Edition

Revision A, January 2012

Printed in USA

Agilent Technologies, Inc.
3501 Stevens Creek Blvd.
Santa Clara, CA USA 95052

Software Revision

This guide is valid for the B.04.00 revision or higher of the Agilent G2721AA/G2733AA Spectrum Mill MS Proteomics Workbench software, until superseded.

Windows[®] is a U.S. registered trademark of Microsoft Corporation.

UNIX[®] is a registered trademark of the Open Group.

Java[™] is a U.S. trademark of Sun Microsystems, Inc.

Warranty

The material contained in this document is provided “as is,” and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Technology Licenses

The hardware and/or software described in this document are furnished under a license and may be used or copied only in accordance with the terms of such license.

Restricted Rights Legend

U.S. Government Restricted Rights. Software and technical data rights granted to the federal government include only those rights customarily provided to end user customers. Agilent provides this customary commercial license in Software and technical data pursuant to FAR 12.211 (Technical Data) and 12.212 (Computer Software) and, for the Department of Defense, DFARS 252.227-7015 (Technical Data - Commercial Items) and DFARS 227.7202-3 (Rights in Commercial Computer Software or Computer Software Documentation).

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A **WARNING** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a **WARNING** notice until the indicated conditions are fully understood and met.

In This Guide...

The *Application Guide* presents instructions to analyze protein and peptide data with the Spectrum Mill MS Proteomics Workbench.

This guide assumes that the Spectrum Mill workbench has already been installed on your server, databases have been downloaded and indexed, and the server is ready to go (*Installation Guide*).

It also assumes you have set up your PC client, transferred files to the server and started the software (*Quick Start Guide*). Refer to the *Quick Start Guide* for a diagram of the workflows.

1 Processing MS/MS Data Interactively

Learn basic step-by-step procedures to process MS/MS data using a basic workflow and an iterative workflow.

2 Reviewing and Validating MS/MS Data

Learn details to help you customize MS/MS data review and validation and to do manual validation and review.

3 Automating Workflows

Learn how to set up and execute workflows containing sequential Spectrum Mill tasks (parameter files), e.g., extraction, search, autovalidation and summary.

4 Sherenga *de novo* Sequencing

Learn details to help you understand and use the software module for Sherenga *de novo* sequencing.

5 Processing for Differential Expression Quantitation

Learn details to process quantitative data for differential profiling studies.

6 Processing MS-Only Data Interactively

Learn basic step-by-step procedures to process MS-only data.

7 Using the Tool Belt

Learn how to use the tools on the Tool Belt page.

8 Using Spectrum Mill Utilities

Learn how to use the many useful protein/peptide utilities included with the Spectrum Mill workbench.

9 System Administration

Learn common system administration tasks, including how to install and update databases.

10 Files Created during Spectrum Mill Data Processing

Learn about the files the Spectrum Mill workbench creates. This information is useful to troubleshoot data processing, to selectively remove some of the data processing, and to make decisions about data archives.

Contents

1	Processing MS/MS Data Interactively	11
	Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively	13
	Step 1. Run the Data Extractor	13
	Step 2. Search in variable modifications mode	19
	Step 3. Validate high quality results automatically	23
	Step 4. Examine the validated results for variable modifications mode	27
	Iterative Workflow – Alternative Interactive Processing	32
	Overview	32
	Step 1. Run an identity search	36
	Step 2. Autovalidate the data	39
	Step 3. Create a file of previously validated results	41
	Step 4. Search in variable modifications mode	42
	Step 5. Autovalidate variable modifications data	47
	Step 6. Check for remaining high-quality spectra	47
	Step 7. Make a decision regarding further processing	51
	Step 8. Summarize valid results	52
	To search in no enzyme mode	58
	To search a larger database	58
	To search in homology modes, or do a mass gap search	59
	To use Sherenga <i>de novo</i> sequencing	60
	To review Sherenga <i>de novo</i> sequencing results	63
	Additional Tools	67
	Spectrum Matcher	67
	Peak Picker	69
	Build TIC	70
2	Reviewing and Validating MS/MS Data	71
	Setting Up Autovalidation	73
	Step 1. Select Data Directory and/or parameter file	73

Step 2. Select the Strategy and Mode to autovalidate results	74
Step 3. Save parameters and validate data	80
Setting Up the Protein/Peptide Summary Page	81
Step 1. Set data directory and mode	81
Step 2. Set filtering, sorting, and validation parameters	83
Step 3. Choose review fields	86
Step 4. Select Protein Quantitation Options	87
Step 5. Save parameters and summarize data	88
Reviewing and Validating Results Interactively	89
To display results in preparation for manual validation	89
To manually review and validate results	93
To use the Spectrum Viewer	96
Protein/Peptide Summary Displays	104
Peptide	105
Protein Summary	106
Protein Summary Details	108
Protein-Single Peptide ID	110
Protein-Protein Comparison Columns	111
Protein-Protein Comparison Redundant	114
Protein-Sample Centric Rows	115
Protein-Sample Centric Rows Details	116
Protein-Peptide Distribution Columns	117
Protein-Peptide Comparison Columns	118
3 Automating Workflows	119
Setting Up Automated Workflows	120
To create parameter files or workflow tasks	120
To view already existing workflows and their tasks	120
To create/edit an automated workflow	122
Executing and Monitoring Workflows	124
To execute an automated workflow	124

	To monitor results	125
	To view workflow progress	126
	To check the Completion Log	128
4	Sherenga <i>de novo</i> Sequencing	129
	To set parameters and run the <i>de novo</i> sequencing algorithm	131
	To generate a Sherenga report	134
	To view detailed Sherenga results	136
	To compare Sherenga results with MS/MS Search results	138
5	Processing Data for Differential Expression Quantitation	139
	To use the Data Extractor for a differential profiling study	141
	To use MS/MS Search for a differential profiling study	143
	To calculate DEQ ratios for isotopic labels using the Protein/Peptide Summary page	144
	To interpret DEQ results for isotopic labels in peptide mode	145
	To interpret DEQ results for isotopic labels in protein modes	147
	To calculate iTRAQ and TMT ratios using the Protein/Peptide Summary page	148
	To interpret results for iTRAQ labels in peptide mode	149
	To interpret results for iTRAQ labels in protein mode	150
	To view light/heavy results on the Spectrum Summary page	151
6	Processing MS-Only Data Interactively	153
	Acquiring Agilent TOF Data for Use with the Spectrum Mill Workbench	154
	MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively	155
	Step 1. Run the Data Extractor	156
	Step 2. Run database searches	162
	Step 3. Summarize PMF Search results	164
	Step 4. Manually review results	165
	Step 5. Summarize and print results	167
	Single-Spectrum Processing	169

Step 1. Find the compounds of interest in Mass Hunter Qualitative Analysis 169

Step 2. Copy and paste compound masses into PMF Search 170

Step 3. Search the MS-only data and display results 171

Step 4. Examine the results 174

7 Using the Tool Belt 175

To terminate a process 176

To create a saved results file so you can search previous hits 178

To create an MS/MS Search summary file if search terminated abnormally 180

To create a summary table of previously-used parameters 181

To create a summary table of MS/MS identification statistics 183

To report FDR and search statistics 186

To copy spectra to your collections directory 190

To list details about amino acid modifications 191

To create a file of iTRAQ correction factors 193

To apply a file of iTRAQ correction factors 194

To calculate discriminant scoring coefficients 195

To export PepXML files 197

To convert spectra 199

To archive data 200

8 Using Spectrum Mill Utilities 203

To identify digest peptides likely to meet specific experimental goals (Peptide Selector) 204

To build MRM lists for triple quadrupole instruments 207

To align sequences (Multiple Sequence Aligner) 209

To list peptides that correspond to a theoretical protein digest (MS Digest) 210

To retrieve database entries using text searches (MS Edman) 212

To list theoretical fragment ion masses for peptides (MS Product) 214

To list amino acid compositions that fit precursor mass and partial composition (MS Comp) 216

To show isotope patterns of peptides (MS Isotope) 218

To calculate the masses and formulas for peptides 219

9 System Administration 221

Manipulating sequence databases 222

To install or update databases 222

To create database indices 224

To create a species and protein molecular weight subset database 226

To create a subset database from saved hits 227

To create a user (proprietary) database or add sequences to one 228

To generate a database summary report 229

To compare two databases 230

To calculate statistics 231

To make a subset FASTA file from Accession Numbers 232

Other system administration tasks 233

To add custom amino acid modifications 233

To change the URLs of HTML links in the search results 234

To enable the HTML link to BLAST search 234

To add/change options related to biology/chemistry 235

To add instrument types 235

To maintain server performance 236

To use server administration scripts 236

To avoid problems with connection time-outs 236

To configure server with drives other than the default configuration 237

To remove and reinstall the Spectrum Mill workbench 238

To limit the number of parallel workflow processes to less than the CPU count 239

To start and stop the Spectrum Mill Workflow Manager Service 240

10 Files Created during Spectrum Mill Data Processing 241

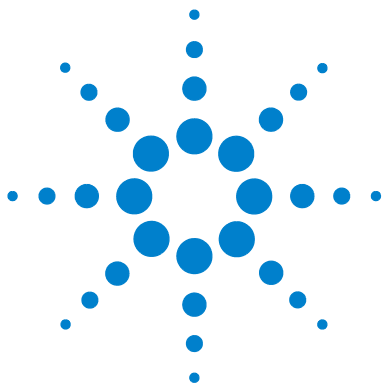
Architecture Overview 242

Data Extractor (MS/MS raw data) 244

Data Extractor (generic data) 246

Contents

Data Extractor (MS-only raw data)	248
MS/MS Search	249
Protein/Peptide Summary, Spectrum Summary, and Autovalidation	251
Tool Belt	253
Sherenga <i>de novo</i> Sequencing	255
PMF Search	256
MS Edman	257
Protein Databases	258



1

Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks
Interactively [13](#)

Iterative Workflow – Alternative Interactive Processing [32](#)

Additional Tools [67](#)

This chapter gives more details for the steps in basic and iterative MS/MS workflows ([Figure 1](#)) and shows you how to set up parameter files for automating workflows, how to automate tasks interactively and how to execute some single tasks immediately.

The basic workflow is a one-pass search and validation workflow that lets you identify most spectra in your sample. If you want to gather more information to characterize identified proteins, use an iterative workflow.

First read the *Quick Start Guide* to learn this basic information:

- Features new to the current version of software
- How to find other information for using Spectrum Mill successfully
- How to transfer spectral files to the server
- How to configure the PC client and start the software

Both this chapter and the *Quick Start Guide* assume that the Spectrum Mill workbench has already been installed on your server and databases have been downloaded and indexed. If this has not yet been done, see the *Installation Guide*.

- Descriptions of the Spectrum Mill workflows that Agilent recommends for analyzing protein/peptide data
- How to process data automatically and interactively



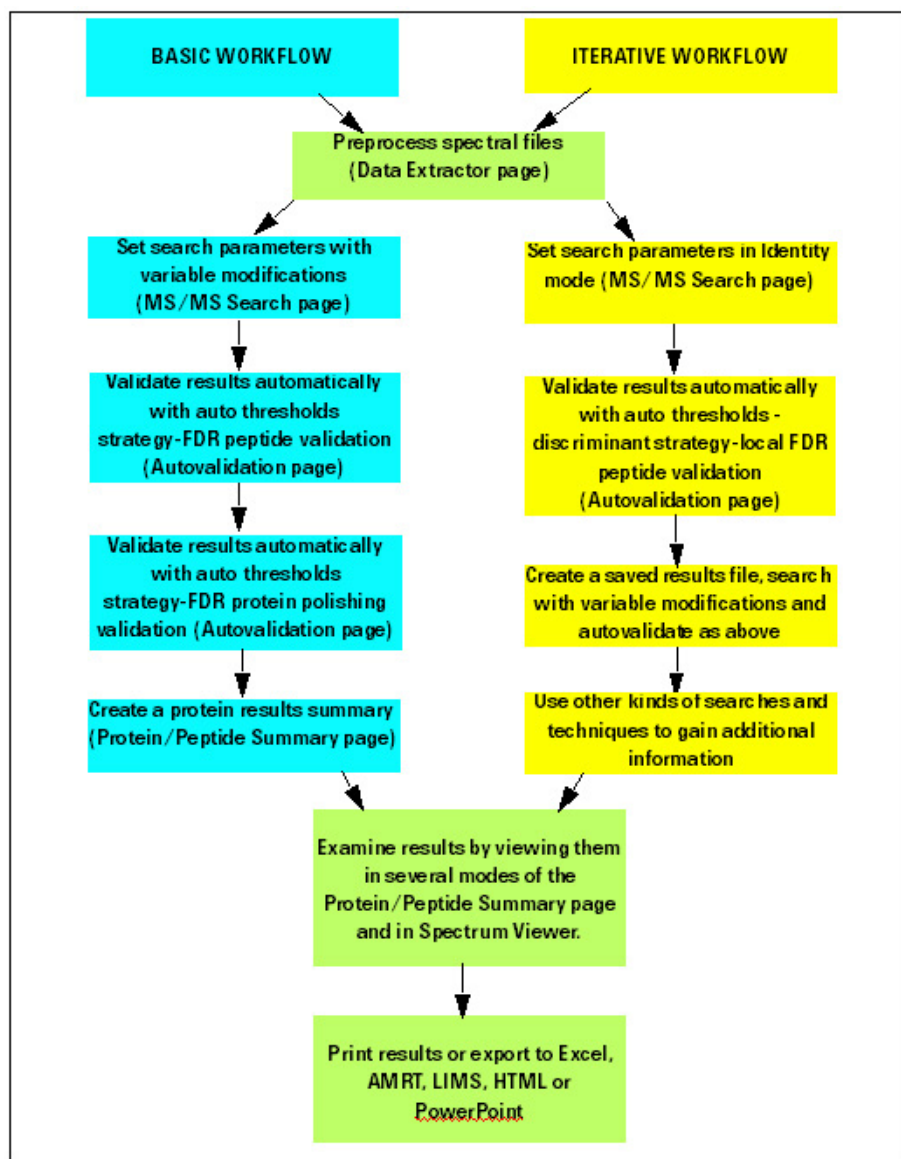


Figure 1 Basic and Iterative Workflows

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

This section gives you the instructions for setting up the parameter files needed to create automated basic workflows and to run the workflow tasks interactively. See Chapter 3 for instructions on creating automated workflows from the parameter files you set up here.

When started interactively, extraction and search tasks are automatically submitted to the workflow manager, or queue, to be executed. You have the option to select a Queue request for autovalidation and P/P Summary, but Autovalidation cannot start (Queue request unmarked) before searching is complete. While extraction and searching are being executed, start autovalidation and P/P Summary after marking the Request queue check boxes. This is called *interactive automation*.

Step 1. Run the Data Extractor

The Spectrum Mill Data Extractor software has two fundamental types of MS/MS data extractors: raw and generic.

The Spectrum Mill workbench recognizes the type of MS/MS data file and chooses the appropriate data extractor. The Data Extractor page changes accordingly.

To preprocess raw data files (Agilent Q-TOF or ion trap, Thermo Fisher Scientific ion trap, or Applied Biosystems/MDS Sciex QSTAR data)

The raw data extractors operate on *.d, *.raw and *.wiff files. They extract and merge nearby MS/MS spectra from the same precursor ion. They optionally apply MS/MS similarity criteria prior to merging scans, to avoid merging closely eluting or co-eluting isobaric peptides.

For Agilent ion trap data, and Thermo Fisher Scientific *.raw ion trap data, the extractors optionally merge MS² and MS³ scans from the same precursor. The extractors assign precursor charges where possible,

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

centroid the MS/MS spectra, calculate spectral features, filter MS/MS spectra by quality, and calculate extracted ion chromatograms (EICs) for the intervening MS precursor scans. The latter are used for quantitation.

- 1 Navigate to the Data Extractor page.
- 2 In the **Data Directories** section, click the **Select...** button to select the folder or folders that contain your files.
- 3 For Agilent Q-TOF data and most extractions, fill in the boxes as shown in [Figure 2](#).

The screenshot displays the 'Agilent Spectrum Mill - Data Extractor' window. The title bar includes a placeholder file name '- YourName\extract-y1'. The interface is organized into several sections with blue headers:

- Extraction:** Contains buttons for 'Extract' (highlighted in green), 'Save As...', and 'Load...'. There is a checked checkbox for 'Remove all prior results' and an unchecked checkbox for 'Show only MS (PMF) parameters'.
- Data Directories:** Features a 'Select...' button and a list box containing 'ExampleData\Agilent\Q-TOF\yeast\y1' with a checked selection box.
- Modifications:** Includes a 'Choose...' button and a 'Fixed:' label followed by a text box containing 'Carbamidomethylation (C)'.
- MS/MS Spectral Feature Filtering:** This section contains several input fields and checkboxes:
 - 'MH+' is set to 600.0 and 'Da' is set to 4000.0.
 - 'Scan time range:' is set from 0 to 300 minutes.
 - 'Sequence tag length >' is set to 1, with a note: '(For MALDI: Set tag length to -1 and merge secs to total run time.)'
 - 'Ignore spectra with dissociation mode:' has checkboxes for 'CID' (unchecked) and 'ETD' (unchecked).
 - 'Merge nearby MSⁿ scans with same precursor m/z' is checked.
 - 'Retention time & m/z tolerance:' is set to +/- 30 seconds and +/- 1.4 m/z, with a note: '(also used for calculating chromatographic peak area of precursor in MS scans)'.
 - 'General MS/MS Merging Constraints:' is set to 'Spectral Similarity & RT & m/z' via a dropdown menu.
 - 'Merge MS² and MS³ spectra from same precursor:' is set to 'Merge' via a dropdown menu.
- Precursor m/z & Charge Assignment:** This section includes:
 - 'Precursor Charge:' set to 'Find' via a dropdown.
 - 'Maximum (z):' set to 7.
 - 'Minimum MS1 S/N:' set to 25.
 - A checked checkbox for 'Find ¹²C precursor m/z'.
 - 'MS Noise threshold:' set to 100 counts, with a note: '(only applies to Agilent Q-TOF)'.

Figure 2 Data Extractor for .d files, shown with typical settings for Agilent Q-TOF

- 4 If you have a different data type, or want to change other parameters, choose from the optional steps below.

To find explanations of all the fields in a particular category, click the blue divider bar to bring up that category in the online help.

- Click the **Choose...** button to select the modifications that match the chemistry for your samples.
 - To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog.
 - For more information about choosing modifications, click the blue bar labeled **Modifications** to access the online help.
 - Note that your system administrator can add custom modifications.
- If you are extracting MALDI MS/MS data, change the **Sequence tag length** to -1.

Otherwise, leave the default value. The sequence tag length represents the longest sequence of amino acids that can be located in the spectrum. The default of >1 for non-MALDI data removes most of the noisy spectra without removing good data.

- If necessary, change the **Retention Time & m/z tolerance** time range to be compatible with your chromatographic data.

For MALDI MS/MS data, change the time range to 2000 sec, or to the spot analysis time.

- For complex samples, keep the default selection for **General MS/MS Merging Constraints**. See the online help for descriptions of the other options.

For Agilent Q-TOF data, you may want to increase the **MS Noise threshold**, especially if the extractor reports an out-of-memory error.

- Set any other parameters. In general, you should use default values, and change only the parameters highlighted in red text.

- 5 Click **Save As** to save the parameter file, enter a name and click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 6 Click **Extract**.

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

The extraction is submitted to the workflow manager to be executed.

- 7 Click on the **Monitor Results** link to monitor the extractor results, or the **View Request Queue** link to view and monitor tasks.

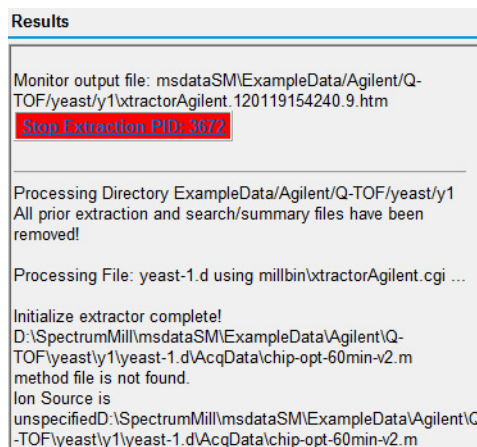


Figure 3 Display while the raw file Data Extractor runs

The exact display depends upon the type of file you extract. The Data Extractor extracts all files in the subdirectory. Extraction time varies depending on the number and size of the files. A message informs you when extraction is complete.

If, after extracting, you put other single tasks in the queue, one right after the other, such as MS/MS Search, Autovalidation and Protein/Peptide Summary, this process is called *interactive automation*.

To preprocess generic peak list files (e.g., Waters Micromass Q-Tof or MALDI-TOF-TOF data)

The generic data extractor processes peak list files, such as centroided peaks from Waters Micromass (Q-Tof) *.pkl and *.dta files, and most *.mgf files.

Since this data extractor operates on exported peak list files rather than raw files, it does not have access to chromatographic time information and mass information contained in MS precursor scans. This limits quantitation capabilities. Like the raw data extractor, the generic data extractor filters MS/MS spectra by quality and calculates spectral features.

- 1 Navigate to the Data Extractor page, shown in Figure 4.

Figure 4 Data Extractor for generic (peak list) files

- 2 In the **Data Directories** section, click the **Select...** button to select the folder or folders that contain your files. The Data Extractor page changes to match Figure 4.
- 3 Mark the **Generate spectral features file only** to generate the file **SpecFeatures.#.tsv**, without actually generating the extracted spectra.

This option appears when you select a directory that contains peak list files but no raw data file. When you have *.dta files, or *.pkl files that represent individual spectra, you put your files in the **cpick_in** folder, and then you must mark this check box. (When you have *.mgf files or appended *.pkl files, where each file contains multiple spectra, then you put your file in the root sample directory and you do not mark the check box.)

- 4 Select the instrument type you used for the analyses.

This option only appears when you select a data directory with generic peak list files.

- 5 For most extractions, fill in the boxes as shown in Figure 4.

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

- 6 If you want to change other parameters, choose from the optional steps below.

To find explanations of all the fields in a particular category, click the blue divider bar to bring up that category in the online help.

- Click the **Choose...** button to select the modifications that match the chemistry for your samples.
 - To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog.
 - For more information about choosing modifications, click the blue bar labeled **Modifications** to access the online help.
 - Note that your system administrator can add custom modifications.
- If you are extracting MALDI MS/MS data, change the **Sequence tag length** to -1.

Otherwise, leave the default value. The sequence tag length represents the longest sequence of amino acids that can be located in the spectrum. The default of >1 for non-MALDI data removes most of the noisy spectra without removing good data.

- Set the other parameters. In general, you should use default values, and change only the parameters highlighted in red text.

- 7 Click **Save As** to save the parameter file, enter a name and click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 8 Click **Extract**.

The extraction is submitted to the workflow manager to be executed.

- 9 Click on the **Monitor Results** link to monitor the extractor results, or the **View Request Queue** link to view and monitor tasks.

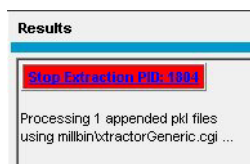


Figure 5 Display while monitoring the results for a generic (peak list) extraction

The Data Extractor extracts all files in the subdirectory. Extraction time varies depending on the number and size of the files. When monitoring results, a message informs you when extraction is complete.

If, after extracting, you put other single tasks in the queue, one right after the other, such as MS/MS Search, Autovalidation and Protein/Peptide Summary, this process is called *interactive automation*.

Step 2. Search in variable modifications mode

- 1 Navigate to the MS/MS Search page, as shown in [Figure 6](#).
- 2 Click the **Select...** button to choose the **Data Directories** where your files reside.
- 3 To load settings from an existing parameter file, click **Load**.
- 4 Set the **Validation filter** to **spectrum- not- marked- sequence- not- validated**.
- 5 Select a **Database**.

If you want to search a species-specific subset database, select the database name for a subset database you previously created.

If you have not yet created a species subset database, see “[To create a species and protein molecular weight subset database](#)” on page 226.

- 6 To restrict your search to peptides that contain specific amino acids (e.g., cysteine-containing peptides for ICAT experiments), type the single-letter abbreviation(s) for the amino acid(s) in the box for **Required AAs**. See the online help for details.
- 7 Click the **Choose...** button to select **Fixed** or **Mix Modifications**.
- 8 Add any **Variable Modifications** that you wish to search, and click **OK**.
 - Note that when you select any variable modifications, the Search mode changes to **Variable modifications**.
 - To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog.
 - For more information about choosing modifications, click the blue bar labeled **Modifications** to access the online help.
 - Note that your system administrator can add custom modifications.

NOTE

If you are not sure what variable modifications to choose, do a mass gap search to find the unknown modifications. See “To search in homology modes, or do a mass gap search” on page 59.

In a basic workflow you combine multiple variable modifications into a single search, but as the number of permutations increases, the search time increases and the number of hits decreases (as a result of FDR-based validation to limit the number of false positives). So the best strategy for a basic workflow is to use the variable modifications you think are present as long as the search speed is acceptable. Remember that the goal of a basic workflow is to obtain as much information as possible in a single round of processing.

You can combine the following common modifications into a single search: oxidized methionine (methionine sulfoxide) and pyroglutamic acid.

If you suspect phosphorylation, combine the following modifications into a single search: phosphorylated S, phosphorylated T, and phosphorylated Y.

Note that you can search only a single type of variable modification for a given amino acid. For example, you cannot select both guanidination and carbamylation of lysine as variable modifications within the same search. Metabolic modifications such as SILAC and ¹⁵N-mix are the exception, where you can select an additional fixed or variable modification that applies to the same amino acid. For example, if the SILAC modifies K, you can still select carbamylated lysine.

-
- 9 From the **Instrument** list, select the correct instrument.
 - 10 To help limit false positives and allow FDR calculations when you review the data, verify that the check box for **Calculate reversed database scores is marked**.
 - 11 If you have ion trap data, decide whether to use **Proton mobility scoring**.
 - In most situations, you mark this check box because this scoring refinement for ion trap data provides better results.
 - But do not mark the check box if you have peptides that are modified by iTRAQ, lysine mass tagging, guanidination or phosphorylation, because these modifications change the expected fragmentation pattern.
 - 12 Mark the check box for **Dynamic peak thresholding**, which is a scoring enhancement that enables identification of more low-abundance and short-chain peptides.

13 To search the variable modification(s) that you chose in [step 8](#) set the **Precursor mass shift range** to a value that is appropriate for the types and numbers of modification(s) that you are searching. Larger values may produce longer searches.

14 Under **Data Files**, select the **Fragmentation** mode to match the one within the instrument settings.

For CID-only instruments, be sure to select **All**.

15 Set the other parameters. See [Figure 6](#).

In general, you should keep the defaults except for the settings that are highlighted in red. Click the blue divider bars for more information in the online help.

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

Agilent Spectrum Mill - MS/MS Search - (none)

Spectrum Mill | Autovalidation | Protein/Peptide Summary | Extractor | Databases | Workflows | Tool Belt | Help

Search

Start Search | Save As... | Load... | ☒ Remove all prior MS/MS Search results

Data Directories

Select... | ☒ ExampleData/Agilent/Q-TOF/yeast/y1

Search Parameters

Validation filter: spectrum-not-marked-sequence-not-validated | Batch size: 81

☐ Search previous hits | Max reported hits: 5

Database: NCBI/nr-yeast | Digest: Trypsin

Species: All | Maximum # missed cleavages: 2

Protein pI: from 3.0 to 10.0 | ☒ All | Required AAs: | Disallowed AAs:

Modifications

Choose... | Fixed: Carbamidomethylation (C) | Variable: Oxidized methionine (M), Pyroglutamic acid (N-termQ), Phosphorylated S (S), Phosphorylated T (T)

Search Criteria

Matching Tolerances

Minimum matched peak intensity: 50 %

Instrument: Agilent ESI Q-TOF

Masses are: Monoisotopic

Precursor mass tolerance: +/- 20 ppm

Product mass tolerance: +/- 50 ppm

Maximum ambiguous precursor charge: 3

Spectral Quality

☐ Sequence tag length: > 3

Minimum detected peaks: 4

Search Mode

☒ Calculate reversed database scores

☐ Proton mobility scoring

☒ Dynamic peak thresholding

Discriminant scoring: Disable (same as Score)

Search mode: Variable modifications

Precursor mass shift range: -18.0 to 177.0 Da

Data Files

Fragmentation mode: All

Spectrum files (./cpick_in/): *.pk1

Figure 6 MS/MS Search page, set up for variable modifications mode search

- 16** Click **Save As** to save the settings to the existing parameter file or a new one, enter a name and click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 17** Click **Start Search**.

The search is submitted to the workflow manager to be executed.

- 18 Click on the **Monitor Results** link to monitor the search results, or the **View Request Queue** link to view and monitor tasks.
- 19 When the search is complete, autovalidate high-quality results in the next step.

Step 3. Validate high quality results automatically

Validation means you accept that the database match is good. The Spectrum Mill workbench provides two ways to validate MS/MS search results. One way uses the Autovalidation page, and is totally automated. You use this method to validate high quality results. The other method uses the Protein/Peptide Summary page for manual review and validation. See [“Reviewing and Validating Results Interactively”](#) on page 89.

- 1 Navigate to the **MS/MS Autovalidation** page, as shown in [Figure 7](#).

You navigate to this page from the Spectrum Mill Home page, the MS/MS Search page or the Protein/Peptide Summary page.

When the page first appears, you see the Auto threshold strategy and the Peptide mode.

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

Agilent Spectrum Mill - MS/MS Autovalidation - (none)

Help

Validate Files ☐ Queue request Undo Last Clear All Save As... Load...

Data Directories

Select... Fragmentation mode: All

(none selected) Search result files:

*.spo

Validation Parameters

Strategy: ☐ Fixed thresholds ☒ Auto thresholds ☐ Auto thresholds - discriminant

Mode: Peptide Auto determine using score, delta R1-R2 thresholds to reach a target FDR

Optimize score & R1-R2 score thresholds for each run with max FDR: 1.2 %

Precursor charge range: 2 to 6 Required AAs: any Disallowed AAs: none

Filtering	Automatic variable range for each run	Fixed range for all runs
☒ None (ppm)	☐ Auto precursor mass error	☐ Fixed precursor mass error Low -1.0 High 30.0 ppm
☒ None (SC/pl)	☐ Auto SCX Solution Charge, pH3	☐ Fixed Solution Charge Low -2 High 6
	☐ Auto OGE/IEF peptide pl	☐ Fixed peptide pl Low 3.0 High 10.0

Figure 7 Autovalidation of MS/MS Search results – Auto thresholds strategy/Peptide mode

- 2 Select the **Data Directories** for which you want to validate results.
- 3 If you want to use the settings from an existing parameter file, click **Load**.

If you are not sure of a field's meaning, click the category divider bar for the section containing the field to see the online help explanation for the field.

- 4 Choose the **Strategy** for autovalidation and the **Mode** that suits your study best. (See [Table 1.](#))

Table 1 Choose the Autovalidation Strategy and Mode

Select this Strategy:	Select this Mode:	What happens:	Pay attention to this:
Auto thresholds	Peptide	- Automatically optimizes the score and R1-R2 score thresholds until they reach a specified target FDR (false discovery rate) - See the online help to learn more about the false discovery rate (FDR)	- Use this mode before the Protein Polishing mode. - You may want to review the results before autovalidation again in the Protein polishing mode. 1.2% is set as the Target FDR instead of 1% because, in practice, a target FDR of 1.2% results in an actual FDR of 1%.
	Protein polishing (optional)	- Part 1: Minimum protein score and "... raise minimum protein score to yield maximum protein FDR %" - Controls the maximum allowed protein FDR. - Part 2: Peptide FDR for validated proteins % - Recalculates the peptide FDR based on validated proteins only. This recalculation usually increases the number of validated peptides, thereby increasing the sequence coverage of the validated proteins but leaving their number unchanged.	Part 1 is accomplished by "invalidating" peptides from the lowest-scoring proteins. Part 1 enables you to be aggressive with peptide level FDR thresholds in peptide mode and then remove suspicious proteins with Protein polishing.
Auto thresholds - Discriminant	Peptide	Automatically optimizes the discriminant score thresholds until they reach a specified global or local FDR	- In combining results from multiple searches (from successive iterations of the iterative workflow), it is preferable to use the local FDR rather than the global FDR. - The local FDR measures the error rate of an individual validated spectrum, while the global FDR measures the error rate of an entire set of validated spectra. - Use this mode before the Protein Polishing mode.
	Protein polishing (optional)	See the "What happens" description for protein polishing above.	See the "Pay attention to this" description for protein polishing above.

Table 1 Choose the Autovalidation Strategy and Mode

Select this Strategy:	Select this Mode:	What happens:	Pay attention to this:
Fixed thresholds	Protein details	Summarizes results by protein and considers all the peptides that belong to a given protein by entering fixed score thresholds beyond which the matches are not validated.	- Individual peptides typically have scores > 6 to 12; the cumulative protein score is typically > 20. - You must use this mode before the peptide mode; Protein details is step 1 and Peptide mode is step 2. - You can use the fixed thresholds strategy in an iterative workflow.
	Peptide	Summarizes results by peptide by entering fixed score thresholds beyond which the matches are not validated.	Using the default scoring, individual peptides typically have scores greater than 11 to 15 (depending on charge state). If the score is high enough, this mode validates the search results even if it finds only a single peptide corresponding to a protein.

- 5 After selecting a strategy and mode, change any parameter values you choose.

Agilent recommends that for Agilent Q-TOF .d data you select the Auto strategy and validate in the peptide mode first, then the protein polishing mode (Figure 8). See Chapter 1 of the *Familiarization Guide* for the recommended parameter values. You may want to try other strategies with other instrument data.

NOTE

Click the blue divider bar for Validation Parameters to learn more about the parameters for each strategy and mode.

- 6 Click **Save As** to save settings to the original or a different parameter file, then click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 7 If you want to see results immediately, make sure the **Queue Request** check box is clear. If you want to queue the task as part of an *interactive automation*, mark the check box.

8 Click **Validate Files**.

If the Queue Request check box is clear, you quickly see a validation summary that lists the hits and spectra that have been validated. If it is marked, the task is submitted to the workflow manager to be executed.

Figure 8 Autovalidation of MS/MS Search results – Auto thresholds/Peptide mode

Step 4. Examine the validated results for variable modifications mode

After autovalidating your results, you now have a choice of viewing different types of protein and peptide summary reports. See [Table 2](#) on page 53. For a basic workflow review, however, we want to see a peptide report showing the site locations of the variable modifications.

- 1 Navigate to the Protein/Peptide Summary page.
- 2 Set **Mode** to **Peptide**.
- 3 Click the **Select...** button to select the folder or folders that contain your data files.

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

- 4 If you want to use the settings from an existing Peptide parameter file, click **Load**.

If you are not sure of a field's meaning, click the category divider bar for the section containing the field to see the online help explanation for the field.

- 5 Set **Validation and Sorting** parameters as shown in [Figure 9](#), or click **Load** to load a parameter file with these settings. These settings filter and present all the validated data. Note that only the settings highlighted in yellow have been changed from defaults.
- 6 Make sure the **Filter Results by** is set to **valid** and the **Validation preset** is set to **none**.
- 7 To see the variable modifications grouped at the top and/or bottom of the results table, for **Sort peptides by**, select **Delta Mass Shift**.
- 8 If you want to see the variable modifications and/or amino acid substitutions, additionally mark the check boxes for **Variable mod sites**, **VML score**, **Delta mass**, and **Modification Names**, as shown in [Figure 9](#).
 - If the variable modifications occur on the N-terminus, C-terminus, or cysteines, also mark the appropriate check box(es) in the last column of the **Review Fields**.
 - **VML score** assigns probabilities when there are multiple possible variable sites.
 - To use the pI filter for modified peptides, ask your server administrator to first verify that the pK of the modified amino acid is specified in **smconfig.std.xml** or **smconfig.custom.xml**.
 - To see a b/y map instead of a regular sequence, clear **Sequence** and mark the **b/y** check box.

Figure 9 Settings to filter and keep valid results

- 9** Click **Save As** to save settings to the original or a different parameter file, then click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 10** If you want to see results immediately, make sure the **Queue Request** check box is clear. If you want to queue the task as part of an *interactive automation*, mark the check box.

- 11** Click **Summarize**.

If the Queue Request check box is clear, you quickly see a protein/peptide summary table that lists the hits and spectra that have been validated. Queuing a summary is useful when you want the results saved to an html file that can be accessed via the Completion Log.

NOTE

Once you click the **Summarize** button, the button is disabled until the results appear. If you need to re-enable the button, click the **Summary Settings** button to reload the Protein/Peptide Summary page.

- 12** After this task is complete, make sure that you see a display similar to that shown in [Figure 10](#).

Note that the filenames are in the format Data_File_Name.aaaa.bbbb.c, where

- aaaa = first merged scan

1 Processing MS/MS Data Interactively

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

- bbbb = last merged scan
- c = assigned precursor charge (0 means charge was ambiguous)

NOTE

The filenames are somewhat different for Applied Biosystems/MDS Sciex data. Instead of using filename.(firstscanindex).(lastscanindex).pkl, the Sciex extracted files use filename.(1/10 x retention time in seconds).(lastscanindex).pkl. Since Analyst software does not show scan numbers, this format allows you to examine the data in Analyst and find the scan by the retention time.

#	Filename	z	Score	SPI (%)	Spectrum Intensity	Score VML	# STY	# sty Sites	# Loc sty Sites	# Ambig sty Sites	Variable Sites
1	yeast-1.18983.19173.3	3	4.27	59.7	1.89e+006	-0.10	2	1	0	1	T928t
2	yeast-1.19137.19137.2	2	6.03	71.0	9.25e+005	3.03	2	1	1	0	Y145y
3	yeast-1.19554.19554.3	3	19.49	91.3	9.27e+005	99	5	0	0	0	M173m
4	yeast-1.20291.20291.2	2	8.79	62.6	1.23e+004	-1	3	0	0	0	
5	yeast-1.12773.12773.2	2	16.76	89.7	7.89e+005	-1	4	0	0	0	
6	yeast-1.11334.11334.2	2	12.56	69.8	4.63e+004	-1	2	0	0	0	
7	yeast-1.14658.14658.2	2	10.80	70.5	3.56e+005	-1	1	0	0	0	
8	yeast-1.16280.16280.2	2	14.32	73.6	7.12e+004	-1	1	0	0	0	
9	yeast-1.11288.11288.2	2	13.42	81.2	5.38e+005	-1	2	0	0	0	
10	yeast-1.5613.5613.4	4	18.34	83.2	1.74e+005	-1	7	0	0	0	

Figure 10 Protein/Peptide Summary report - Peptide mode - first part

Note that the resulting data display includes the additional headings **Variable Sites**, **Modifications Map**, **MH⁺ mass shift (Da)**, and **MH⁺ error (ppm)**, as shown in [Figure 11](#). These columns summarize the variable modifications (and for homology modes the amino acid substitutions). The variable modifications are indicated as lower-case letters under **Variable Sites**.

Basic Workflow – Setting Up Parameter Files and Processing Tasks Interactively

Sequence Map	Modifications Map	MH ⁺ Matched (Da)	MH ⁺ Mass Shift (Da)	MH ⁺ Error (ppm)
(R)V T M/G V I t G D I L V N G I P R(D) VT(0.0)MGVIT(0.0)GDILVNGIPR	t:Phosphorylated T	1754.983	79.9975	17.0
(K)I Y G G D y F V K P I K(N) IY(0.0)GGDY(0.99)FVKPIK	y:Phosphorylated Y	1399.762	79.9555	-7.3
(K)V I N D A F G I/E/E G L/m T T V H S L T A T Q K(I) VINDAFGIEEGLM(1.0)TTVHSLTATQK	m:Oxidized methionine	2575.307	15.9961	0.5
(R)T M L E L L/N/Q/L D/G F/S S D D R(V)		1953.922	0.0383	19.6
(K)A Y T Y F/G E/Q/S N L P K(I)		1517.727	0.0220	14.5
(K)L/D N/T D/F/S/G L R(L)		1137.554	0.0180	15.9
(R)D E S L L/A/D E L/R(F)		1160.579	0.0128	11.0
(R)E G E/Q F A/N/A/F/F/D T E K(K)		1632.718	0.0105	6.4
(R)G D\ V E E/Y/Q/Y L R(S)		1271.590	0.0099	7.8
(K)G/S D Y/D/Y/N/N/S/T H S A E H T P R(H)		2050.849	0.0096	4.7

Figure 11 Additional columns identifying and locating the sites of selected variable modifications**NOTE**

If you click the SPI (%) value for any hit, the standalone Spectrum Viewer appears. You must have marked the SPI (%) check box under Review Fields for this to be possible.

Iterative Workflow – Alternative Interactive Processing

With an iterative workflow, you do a series of searches and autovalidations after extraction and then summarize all the valid results. You use the basic workflow to identify most of the MS/MS spectra in one round. To obtain as much information as possible to help characterize the identified proteins, use an iterative workflow.

Overview

The Spectrum Mill workbench retains a cumulative summary of valid results, so if you elect to further process and validate the data, the software keeps track of all validation states for you. At the end, you can generate a report that contains everything you have validated through multiple interpretation/review/validation steps.

To segregate the valid search results, the software keeps track of both the validation states of the spectra and their interpretations in a coordinated way. Furthermore, it permits spectra to be segregated according to quality (via Spectrum Summary) without regard to the validation states of their database interpretations. Because the software keeps track of which spectra you have segregated as “good” and which database interpretations you have designated as “valid,” you can intelligently apply processing steps to segregated groups of spectra. For example, you can:

- Restrict further database searches to spectra that do not yet have a valid interpretation
- Restrict further database searches to only “good” spectra that do not yet have a valid interpretation
- Summarize only the valid results

Steps for iterative processing

The iterative data processing steps are generalized in [Figure 1](#). Which steps you choose depend upon your sample, the available information in the databases you search, and how important it is to extract all available information from a sample set.

Some possibilities are:

- Search previous hits in variable modifications mode for oxidized methionine and pyroglutamic acid and review/validate results. See [“Step 4. Search in variable modifications mode”](#) on page 42.
- If you suspect additional modifications (e.g., phosphorylation or carbamylation), search previous hits in variable modifications mode for those modifications and review/validate results. (Note: You may break the modifications into groups and repeat this step as often as necessary to cover all the modifications you suspect.)

NOTE

Choosing variable modifications against which to search is a trade-off. If you do not search with a modification, you cannot find it. If you search with too many modifications, the search runs too slowly.

To avoid having the search run too slowly you can take one or both of two actions:

- Search a smaller database.
- Choose fewer variable modifications for each iterative search.

-
- Search previous hits with a no-enzyme or half-enzyme (trypsin nonspecific C term or trypsin nonspecific N term) digest and review/validate results. See [“To search in no enzyme mode”](#) on page 58.
 - Search a larger database in identity mode and review/validate results. See [“To search a larger database”](#) on page 58.
 - Search previous hits in homology mode and review/validate results. See [“To search in homology modes, or do a mass gap search”](#) on page 59
 - Check for remaining high-quality spectra. See [“Step 6. Check for remaining high-quality spectra”](#) on page 47.
 - Create a final results summary. See [“Step 8. Summarize valid results”](#) on page 52.

- Use Sherenga *de novo* sequencing on remaining high quality spectra and review results. See [“To use Sherenga de novo sequencing”](#) on page 60.

You can also manually interpret, review and validate the data if you wish to mine the data more deeply or to confirm autovalidation results. See [“Reviewing and Validating Results Interactively”](#) on page 89.

Search modes

The following modes are available for MS/MS Search:

- Identity: Matches MS/MS spectra to sequences directly created from the databases (no modifications)
- Variable modifications: Searches for variable modifications you have selected. This mode tests all permutations of the combined set of modifications, within the limits of the precursor mass shift that you specify.
- Homology - All mutations: Searches for matches that are consistent with a single amino acid substitution. The substitution does not need to be one that would result from a point mutation. If you have selected variable modifications for an amino acid, they are also tested. The search also matches to sequences of unmodified peptides.
- Homology - Single base pair mutations: Searches for matches that are consistent with a single amino acid substitution that would result from a point mutation (single base substitution within a codon). If you have selected variable modifications for an amino acid, they are also tested. The search also matches to sequences of unmodified peptides.
- For the two homology modes, you can enable a search that allows an **Unassigned single mass gap**. You can use this type of search to identify unknown or unexpected modifications. This search also matches to sequences of unmodified peptides. If you enable this mode, you cannot search for variable modifications in the same search.

Types of databases

You can run a variable modifications or homology mode search against a full database, a species-specific subset database, or a small database of proteins that have already been identified in your sample. To create a species-specific subset database, see [“To create a species and protein molecular weight subset database”](#) on page 226. To create a small

database of proteins that have already been identified, see “[Step 3. Create a file of previously validated results](#)” on page 41. The broader the database search, the longer the search time. A search of previous hits is the most efficient.

If you search species-subset databases, you may want to run more than one variable modifications or homology mode search. For example, if you study proteins from a rare species, you may want to search first a species-specific subset database and second, a more general subset database.

Step 1. Run an identity search

After you have extracted the high-quality MS/MS spectra from your data files (see “[Step 1. Run the Data Extractor](#)” on page 13), you are ready to search the spectra. The example shown here uses the demonstration data from the **QTOF\yeast\y1** data directory. If you wish to setup and run this example, then please follow the exercises in Chapter 2 of the *Familiarization Guide*.

- 1 Navigate to the MS/MS Search page, shown in [Figure 12](#).

Figure 12 MS/MS Search page

- 2 If necessary, click the **Select...** button to select the folder or folders that contain your data files. (If you have just completed data extraction, the directories are already set for you.)

- 3 Select a **Database**, preferably a species subset database for the first search.

The yeast and *ecoli* species databases, from which you can create your own subset database, should already be installed and indexed. See [“To create a species and protein molecular weight subset database”](#) on page 226.

- 4 To restrict your search to peptides that contain specific amino acids (e.g., cysteine-containing peptides for ICAT experiments), type the single-letter abbreviation(s) for the amino acid(s) in the box for **Required AAs**. See the online help for details.

- 5 Click the **Choose...** button to select the **Fixed** or **Mix Modifications** that match the chemistry for your samples. You will choose **Variable Modifications** in a later search.

- To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog.
- For more information about choosing modifications, click the blue bar labeled **Modifications** to access the online help.
- Note that your system administrator can add custom modifications.

- 6 From the **Instrument** list, select the correct instrument.

- 7 For **Maximum ambiguous precursor charge**, select the maximum charge state that you want MS/MS Search to use when it encounters an extracted spectrum for which the charge state is unknown.

While the default of +3 is optimized for collision-induced dissociation (CID) spectra and trypsin digestion, it may be advantageous to select higher charge states for electron transfer dissociation (ETD) spectra or when the proteolytic enzyme produces longer peptides.

- 8 Make sure the **Search mode** is set to **Identity** for this round of processing.
- 9 Verify that the check box for **Calculate reversed database scores** is marked.

This setting triggers a database search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, there may be an incorrect assignment. Such a search helps to rule out false positives. This must be marked if False Discovery Rate (FDR) results are to be reported.

- 10** Select a **Discriminant Scoring** set of coefficients based on your instrument.

If you are not using an Agilent instrument, use the Tool Belt utility to create your own set of coefficients, which will then appear in this list.

The **discriminant score** is calculated as a weighted linear combination of a number of metrics (such as score, SPI, etc.), where the weights determine the relative importance of each of the component metrics.

Agilent recommends that you turn discriminant scoring on for an iterative search in order to calculate the most accurate local FDR %, which is used in the iterative autovalidation.

- 11** Under **Data Files**, select the **Fragmentation** mode to match the one within the instrument settings.

For CID-only instruments, be sure to select **All**.

- 12** Set the other parameters. In general, you should use default values, and change only the parameters highlighted in red text.

To find explanations of the fields in a blue divider category, click the blue divider bar to access the online help for that category's fields.

- 13** Click **Save As** to save the parameter file, enter a name and click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 14** Click **Start Search**.

The search is submitted to the workflow manager to be executed.

- 15** Click on the **Monitor Results** link to monitor the search results, or the **View Request Queue** link to view and monitor tasks.

MS/MS Search processes all files in the subdirectory. Search time varies depending on the size of the database searched. A message informs you when search is complete (Figure 13).

Results				
Monitor output file: msdataSM\ExampleData\Agilent\Q-TOF\yeast\y1 \\mstag.120120080745.10.htm				
MS/MS Search Done!				
Directories to Process Directory: ExampleData\Agilent\Q-TOF\yeast\y1				
Processing Directory ExampleData\Agilent\Q-TOF\yeast\y1 Removing all prior MS/MS results All prior MS/MS results removed numFiles: 15307				
	z	# files	low m/z	high m/z
1	1	76	600.35	692.36
2	1	34	704.52	781.19
3	1	4	792.42	809.40
4	1	1	923.47	923.47
5	1	2	998.47	1004.50

Figure 13 Display after the MS/MS Search is complete

You can also use the MS/MS Search task for interactive automation, where you place the task in the queue interactively after data extraction and before validation in the workflow.

NOTE

If your MS/MS Search terminates abnormally (for example, if you abort the search), see [“To create an MS/MS Search summary file if search terminated abnormally”](#) on page 180.

CAUTION

If you accidentally run MS/MS Search with incorrect settings, mark the check box to **Remove all prior MS/MS results**, or do so if you or someone else has run your data file previously. Do not mark this check box if you do iterative searches and wish to retain all results for a final summary.

Step 2. Autovalidate the data

After the Identity search, autovalidate the results using the Auto thresholds - discriminant strategy and Peptide mode.

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

- 1 Navigate to the Autovalidation page from the MS/MS Search page or from the Spectrum Mill home page.
- 2 To remove any previous autovalidation results for your data file, click **Clear All**.
- 3 Click **Auto thresholds - discriminant**.
- 4 From the **Mode** list, select **Peptide**, if necessary.
- 5 Click **Local FDR** and keep the target percentage the same (Figure 14).
 - In combining results from multiple searches (from successive iterations of the iterative workflow), it is preferable to use the local FDR rather than the global FDR.
 - The local FDR measures the error rate of an individual validated spectrum, while the global FDR measures the error rate of an entire set of validated spectra.
- 6 Click **Validate Files**.

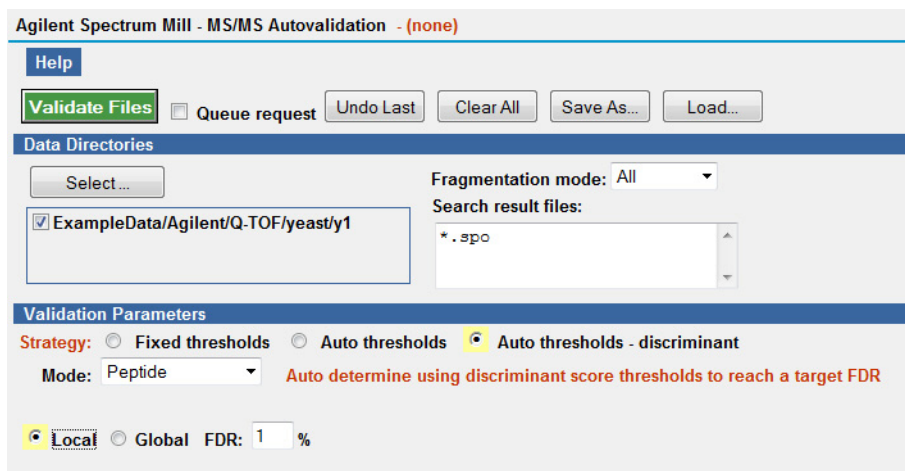


Figure 14 Local FDR selection for the Auto thresholds - discriminant strategy, Peptide mode

Step 3. Create a file of previously validated results

You can follow the steps below to create a file of previously validated results, but this is not necessary if you search previous hits in MS/MS search. The search page automatically creates one for you if you have marked “Search previous hits”. (Figure 17) On the page, be sure to select the database that you searched previously.)

- 1 Navigate to the Tool Belt page and click **Create saved results file**.

Here, you create a file of validated hits from your previous search results. You later search this hit list in variable modifications or homology mode, or in no-enzyme mode.

- 2 Select your data directory and the database you searched. See Figure 15.

The screenshot shows the 'Agilent Spectrum Mill - Tool Belt Utilities' window. At the top, there are tabs: 'Spectrum Mill', 'Extractor', 'MS/MS Search', 'Protein/Peptide Summary', 'Spectrum Summary', and 'Autovalidation'. Below the tabs, under 'Select a Tool:', there are several radio button options. The 'Create saved results file' option is selected. Other options include 'Stop process', 'Spectral collector', 'Export PepXML', 'List modifications details', 'Convert spectra', 'Create MS/MS search summary file', 'Create iTraQ correction factors', 'Archive data', 'View parameter table', 'Apply iTraQ correction factors', 'Report FDR and search statistics', and 'Calculate discriminant scoring coefficients'. Below the tool selection, there is a section titled 'Create Saved Results File' with a description: 'Create a saved results file from validated rank 1 hits (small subset to quickly search previous hits with MS/MS Search)'. A green 'Create File' button is present. Below this, a 'Database:' dropdown menu is set to 'NCBI nr.yeast'. At the bottom, there is a 'Data Directory' section with a 'Select...' button and the path 'ExampleData\Agilent\Q-TOF\yeasty1'.

Figure 15 Utility to create a saved results file

- 3 Click the **Create File** button.
- 4 Check that you see a display similar to that shown in Figure 16.

```
database: NCBIhr.ecoli input program: MS-Tag  
  
OutputResFilename: NCBIhr.ecoli.res  
numValidHits: 850  
numIndices: 850  
numNonRedundantIndices: 126  
numAccessionNumbers: 850  
numNonRedundantAccessionNumbers: 126  
created file F:\SpectrumMill\msdata\SM\ExampleData\Agilent\Trap\ecoli\heatshock\results_mstag\NCBIhr.ecoli.res
```

Figure 16 Display during creation of saved results file

Step 4. Search in variable modifications mode

- 1 Navigate to the MS/MS Search page, as shown in [Figure 12](#).
- 2 Click the **Select...** button to choose the **Data Directories** where your files reside.
- 3 To load settings from an existing parameter file, click **Load**.
- 4 Set the **Validation filter** to **spectrum- not- marked- sequence- not- validated**.
- 5 If you want to search the file of validated results that you created in the previous section, mark the check box next to **Search previous hits**.
- 6 Select a **Database**.

If you want to search a species-specific subset database, select the database name for a subset database you previously created.

If you have not yet created a species subset database, see [“To create a species and protein molecular weight subset database”](#) on page 226.

NOTE

If you attempt a variable modifications search against a large database, the search will take a longer time than using a smaller database, and the chance of random matches will increase. If at all possible, do a variable modifications search against a file of previously validated hits. See [“Step 3. Create a file of previously validated results”](#) on page 41. If that is not possible, reduce the **Maximum # missed cleavages** to 1.

- 7 To restrict your search to peptides that contain specific amino acids (e.g., cysteine-containing peptides for ICAT experiments), type the single-letter abbreviation(s) for the amino acid(s) in the box for **Required AAs**. See the online help for details.

- 8 Click the **Choose...** button to select the same **Fixed** or **Mix Modifications** that you selected for the previous identity mode search.
- 9 Add any **Variable Modifications** that you wish to search, and click **OK**.
 - Note that when you select any variable modifications, the Search mode changes to **Variable modifications**.
 - Unlike in a basic workflow, where you want to add as many modifications as you think are present without slowing down the search significantly, you add just a few for the first variable modifications search in an iterative workflow, and then can do additional searches with other modifications later.
 - To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog.
 - For more information about choosing modifications, click the blue bar labeled **Modifications** to access the online help.
 - Note that your system administrator can add custom modifications.

NOTE

If you are not sure what variable modifications to choose, do a mass gap search to find the unknown modifications. See “[To search in homology modes, or do a mass gap search](#)” on page 59.

As you add more variable modifications, the number of permutations increases, which means the search time increases and the number of hits decreases (as a result of FDR-based validation to decrease the number of false positives). The best strategy for an iterative workflow is to use no more than two or three modifications per search. Remember that the goal of an iterative workflow is to obtain as much information as possible during an iterative process of searches and validations.

Because you have already validated results from an identity-mode search in an iterative workflow, create a file of validated hits (see [page 41](#)), and search against this file. This file is like a mini-database of validated protein hits. It allows you to increase the complexity of the search without significantly increasing the search time. It also greatly reduces the chance of false positive matches, which is important when using threshold-based autovalidation.

You can combine the following common modifications into a single search: oxidized methionine (methionine sulfoxide) and pyroglutamic acid.

If you suspect phosphorylation, combine the following modifications into a single search: phosphorylated S, phosphorylated T, and phosphorylated Y.

Note that you can search only a single type of variable modification for a given amino acid. For example, you cannot select both guanidination and carbamylation of lysine as variable modifications within the same search. Metabolic modifications such as SILAC and ¹⁵N-mix are the exception, where you can select an additional fixed or variable modification that applies to the same amino acid. For example, if the SILAC modifies K, you can still select carbamylated lysine.

10 From the **Instrument** list, select the correct instrument.

11 To help limit false positives and allow FDR calculations when you review data, verify that the check box for **Calculate reversed database scores is marked**.

12 If you have ion trap data, decide whether to use **Proton mobility scoring**.

- In most situations, you mark this check box because this scoring refinement for ion trap data provides better results.

- But do not mark the check box if you have peptides that are modified by iTRAQ, lysine mass tagging, guanidination or phosphorylation, because these modifications change the expected fragmentation pattern.

13 Mark the check box for **Dynamic peak thresholding**, which is a scoring enhancement that enables identification of more low- abundance and short-chain peptides.

14 Select a set of coefficients for **Discriminant scoring**.

If you are not using an Agilent instrument, you may create your own set of coefficients with the Tool Belt utility. The new set of coefficients will then appear as a selection in this list.

15 To search the variable modification(s) that you chose in [step 9](#), set the **Precursor mass shift range** to a value that is appropriate for the types and numbers of modification(s) that you are searching. Larger values may produce longer searches.

16 Under **Data Files**, select the **Fragmentation** mode to match the one within the instrument settings.

For CID-only instruments, be sure to select **All**.

17 Set the other parameters. See [Figure 17](#).

In general, you should keep the defaults except for the settings that are highlighted in red. Click the blue divider bars for more information in the online help.

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

Agilent Spectrum Mill - MS/MS Search - (none)

Spectrum Mill Autovalidation Protein/Peptide Summary Extractor Databases Workflows Tool Belt Help

Search

Start Search Save As... Load... ☐ Remove all prior MS/MS Search results

Data Directories

Select... ☒ ExampleData/Agilent/Q-TOF/yeast/y1

Search Parameters

Validation filter: spectrum-not-marked-sequence-not-validated Batch size: 81

☐ Search previous hits Max reported hits: 5

Database: NCBInr.yeast Digest: Trypsin

Species: All Maximum # missed cleavages: 2

Protein pI: from 3.0 to 10.0 ☒ All Required AAs: Disallowed AAs:

Modifications

Choose... Fixed: Carbamidomethylation (C) Variable: Oxidized methionine (M)
Pyroglutamic acid (N-termQ)

Search Criteria

Matching Tolerances

Minimum matched peak intensity: 50 %

Instrument: Agilent ESI Q-TOF

Masses are: Monoisotopic

Precursor mass tolerance: +/- 20 ppm

Product mass tolerance: +/- 50 ppm

Maximum ambiguous precursor charge: 3

Spectral Quality

☐ Sequence tag length: > 3

Minimum detected peaks: 4

Search Mode

☒ Calculate reversed database scores

☐ Proton mobility scoring

☒ Dynamic peak thresholding

Discriminant scoring: Agilent QTOF standard

Search mode: Variable modifications

Precursor mass shift range: -18.0 to 177.0 Da

Data Files

Fragmentation mode: All

Spectrum files (.cpick_in/): *.pkl

Figure 17 MS/MS Search page, set up for variable modifications mode search

- 18** Click **Save As** to save the settings to the existing parameter file or a new one, enter a name and click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 19** Click **Start Search**.

The search is submitted to the workflow manager to be executed.

20 Click on the **Monitor Results** link to monitor the search results, or the **View Request Queue** link to view and monitor tasks.

21 When the search is complete, follow the instructions for Step 5 below.

Step 5. Autovalidate variable modifications data

In this step, autovalidate again using the same strategy and mode as in step 2 but do not remove the previous autovalidation results.

- 1 Navigate to the Autovalidation page from the MS/MS Search page or from the Spectrum Mill home page.
- 2 Click **Auto thresholds - discriminant**.
- 3 From the **Mode** list, select **Peptide**, if necessary.
- 4 Click **Local FDR** and keep the target percentage the same (Figure 14).
- 5 Click **Validate Files**.

NOTE

You can also use the Fixed thresholds strategies in an iterative workflow.

Step 6. Check for remaining high-quality spectra

- 1 Navigate to the Spectrum Summary page.

This page provides a convenient means to find good spectra that have not yet been interpreted.

One of the most useful measures of spectral quality is the maximum sequence tag length, which is the longest sequence of amino acids that can be located in the spectrum. The Spectrum Summary results show spectra annotated with the amino acid sequences that correspond with the maximum sequence tag lengths. These are not intended as *de novo* interpretations, but rather show how much sequence information *could* be contained in the spectra.

- 2 Set parameters similar to those shown in Figure 18, or click Load to load an existing parameter file.

The steps below discuss important parameters.

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

Figure 18 Spectrum Summary page

- 3 In the **Data Directories** section, click the **Select...** button to select the folder or folders that contain your data files.
- 4 Set the **Spectrum validation filter** as follows:
 - If this is the first round of processing using the Spectrum Summary page, set to **spectrum-not-marked-sequence-not-validated**.
 - For later rounds of processing (after you have already designated spectra as “good”), select one of the options that start with **good-spectrum**.
- 5 Choose one of the three choices for **Validation preset**:
 - If you set filtering parameters such that poor spectra are likely to be excluded (for example, **Maximum tag length** > 7), select **good-spectrum**.
 - Otherwise, select **bad-spectrum** or **reset**.

As you review the spectra later, you can select **reset** to undo a **good-spectrum** or **bad-spectrum** designation.
- 6 Select the **Filter by:** parameter to filter out spectra that are not peptide spectra or that have little information content.

The **Maximum tag length** selection represents the longest sequence of amino acids that can be located in the spectrum. Higher numbers filter out more spectra, keeping those of the best quality.
- 7 Mark the **MS L/H EIC intensity** check box for additional information on ICAT and other isotopically labeled data. For more information, see [“To view light/heavy results on the Spectrum Summary page”](#) on page 151 of [Chapter 5](#).
- 8 Click **Save Settings** if you want to save the parameters.

9 Click **Summarize**.

NOTE

Once you click the **Summarize** button, the button is disabled until the results appear. If you need to re-enable the button, click the **Summary Settings** button to reload the Spectrum Summary page.

10 Check that you see a display similar to the one shown in [Figure 19](#).

Note that the filenames are in the format Data_File_Name.aaaa.bbbb.c, where

- aaaa = first merged scan
- bbbb = last merged scan
- c = assigned precursor charge (0 means charge was ambiguous)

If you have QSTAR files, see the note on [page 30](#).

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

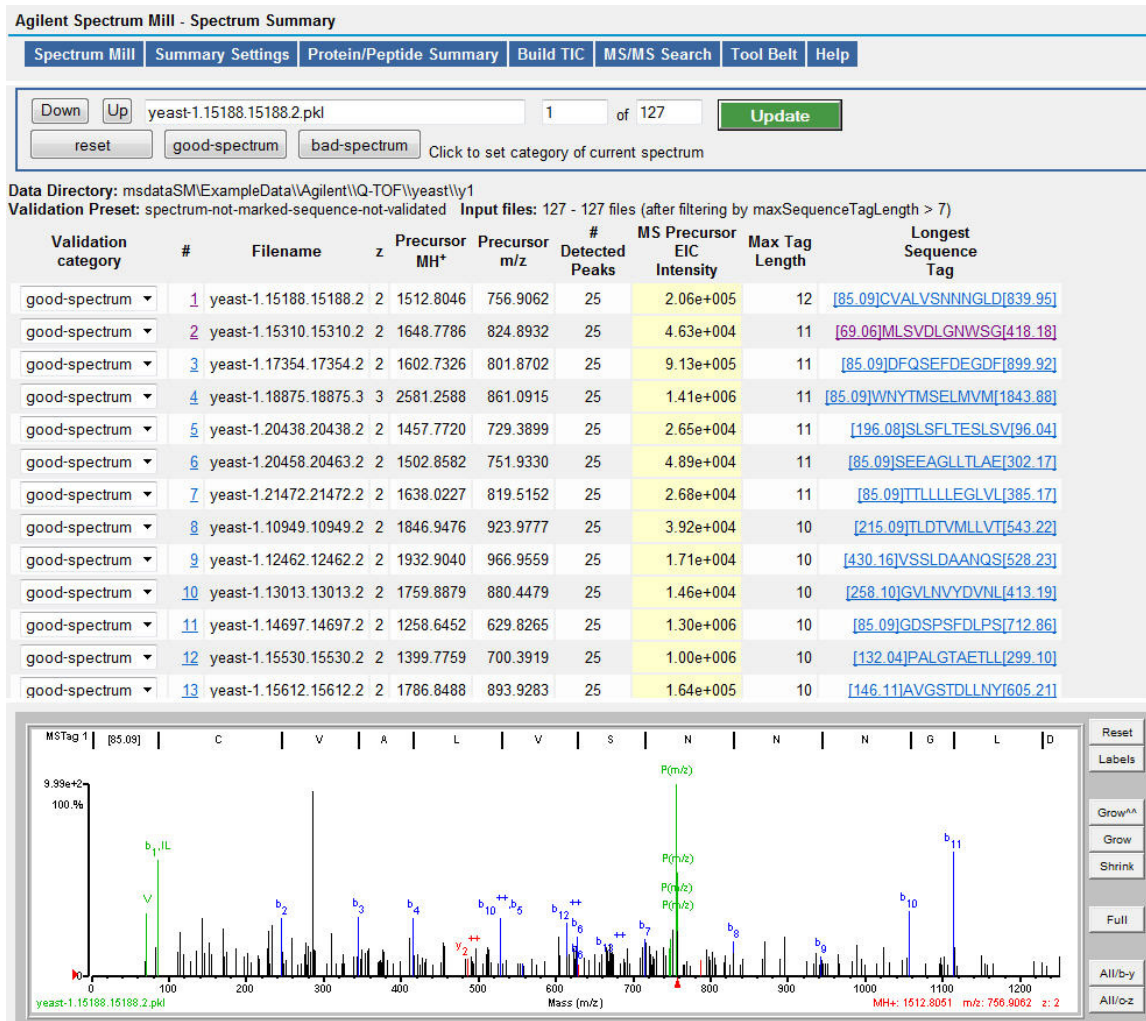


Figure 19 Spectrum Summary results

11 Now examine the spectra. The Spectrum Viewer, shown at the bottom of Figure 19, automatically displays the first spectrum. To display others, click a link under the # or Longest Sequence Tag headers, or click the Up or Down button at the top of the page.

- If you click a link under the # header or the Up or Down button, the spectrum is displayed at the bottom of the page.

- If you click a link under the **Longest Sequence Tag** header, the spectrum is displayed in a new window.
 - For more details about manipulating the spectral display, see “[To use the Spectrum Viewer](#)” on page 96 of [Chapter 2](#).
- 12** After you review each spectrum, designate each one as “good” or “bad,” or “reset.” Use either the drop-down list under the **Validation category** heading, or the **good-spectrum**, **bad-spectrum**, and **reset** buttons near the top of the page.
- 13** Click **Update** to save your designations for future reference.

NOTE

If you find you have designated spectra as “good” or “bad” in error, and you have already clicked the **Update** button, you can undo the action. Change the designation(s) to **Reset** and click the **Update** button again.

If at any point you wish to change parameters and create a new summary table, click the **Summary Settings** button to start over.

Step 7. Make a decision regarding further processing

- 1** After you use the Spectrum Summary page, assess whether you still have a lot of good spectra that lack valid interpretations and decide whether you want to further process the data.

If you do not wish to further process the data, go to step 4.
- 2** If you do wish to further process the data, decide how much you wish to do. The possible next steps are:
 - a** Search for additional variable modifications
 - b** Search in no enzyme mode against previous protein hits. (See “[To search in no enzyme mode](#)” on page 58.)
 - c** Search in one of the homology modes against previous protein hits (See “[To search in homology modes, or do a mass gap search](#)” on page 59.)
 - d** Search a larger database. (See “[To search a larger database](#)” on page 58.)
- 3** Process the data with one or more of the choices above.

- 4 If you do not wish to further process the data with the above choices, create a final results summary. See [“Step 8. Summarize valid results”](#) on page 52.
- 5 Continue with manual review and validation, if you choose. (See [“Reviewing and Validating Results Interactively”](#) on page 89.)

NOTE

If you choose to continue processing, two Spectrum Mill tools are useful to keep track of processing parameters and identification statistics. See [“To create a summary table of previously-used parameters”](#) on page 181 and [“To create a summary table of MS/MS identification statistics”](#) on page 183. The Build TIC page, accessible from Protein/Peptide Summary, also helps you visualize how much of your sample has been interpreted. See [“Build TIC”](#) on page 70.

Step 8. Summarize valid results

At this point you summarize all the valid results.

- 1 If you have been using links to examine the data, click the **Summary Settings** button to display the settings at the bottom of the page.
- 2 Make sure that your data folder is still selected. (Note that if the **Validation preset** is set to **none**, some data display modes permit selection of multiple directories.)
- 3 Select the data display mode that organizes data in the best way for your study, or click **Load** to load an already existing parameter file that includes the display mode and other parameters you want.

Use the following table as a guide. Note that display modes determine which other options are available.

Table 2 Protein/Peptide Summary Modes

If you want to:	And you want results organized by:	Then use this mode:	Example application
Validate results	Peptides	Peptide	Manual review and validation of MS/MS search results, organized by peptide
	Proteins, then peptides	Protein Summary Details	Manual review and validation of MS/MS search results, organized by protein
	Samples, then proteins, then peptides	Protein-Sample Centric Rows Details	Manual review and validation of MS/MS search results, organized by sample
Summarize results by proteins	Proteins only	Protein Summary	List of all proteins identified in the data
	Proteins, then samples	Protein-Protein Comparison Columns	Compare two or more samples, each of which may contain multiple fractions. Each sample (with all fractions) is organized in a separate directory.
	Proteins, then redundant hits	Protein-Protein Comparison Redundant	Same as immediately above, with additional detail on isoforms of proteins
	Proteins, then peptides	Protein Summary Details	View proteins, with supporting peptide details
Summarize results by samples	Samples, then proteins	Protein-Sample Centric Rows	View proteins from multiple 2D gel spots organized in a single directory
	Samples, then proteins, then peptides	Protein-Sample Centric Rows Details	Same as immediately above, with supporting peptide details
Summarize results by peptides	Peptides only	Peptide	List of all peptides identified in the data
	Peptides, then samples	Protein-Peptide Distribution Columns	Method development (evaluation of 2D LC/MS/MS or other fractionation scheme)
	Peptides, then samples	Protein-Peptide Comparison Columns	Evaluation of fractionation scheme (provides more information and easier export to Excel)
View a list of proteins identified via a single peptide	Peptides only	Protein-Single Peptide ID	Examination of results where a single peptide was used to generate a protein identification

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

- 4 In some modes, you have the option to display results by file or directory. Decide how you want to display data and select the **Group results by** option accordingly.
 - If you want to summarize all data files in the directory as a single sample, select **Directory**. This is useful when the samples in your directory are salt fractions from a 2D LC/MS/MS analysis or spots from a 2D gel, and you want to summarize fractions or spots as a single sample.
 - If you want to display results individually for each sample in the directory, select **File**.
- 5 Mark **DEQ ratios** to display ratios for differential expression quantitation, such as light/heavy ratios for ICAT reagents or other reagents that use isotopic labels. See [Chapter 5](#) for details.
- 6 Set the remaining parameters as shown in [Figure 20](#), or use your customized settings. Note that in [Figure 20](#), only the settings highlighted in yellow have been changed from defaults.

See the online help for descriptions of all the fields.

Figure 20 Example of settings to summarize all valid results

- 7 Click **Save As** to save settings to the same or different parameter file, then click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See [Chapter 3](#), “Automating Workflows”.

8 If you want to see results immediately, make sure the **Queue Request** check box is clear. If you want to queue the task as part of an *interactive automation*, mark the check box.

9 Click **Summarize**.

If the Queue Request check box is clear, you quickly see a protein/peptide summary table that lists the hits and spectra that have been validated. Queuing a summary is useful when you want the results saved to an html file that can be accessed via the Completion Log.

10 Check that you see a summary display.

The display shown in [Figure 21](#) was generated with the settings shown in [Figure 20](#). The score and % SPI filters were set to zero to capture all data marked **valid**.

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

ExampleData/Agilent/Q-TOF/yeast/y2 # spectra total intensity	ExampleData/Agilent/Q-TOF/yeast/y1 # spectra total intensity	Protein MW (Da)	Database Accession #	%AA Coverage	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	Group #	Subgroup #	Protein Name
105 1.03e+008	47 3.24e+008	46872.3	6321693	76.4	41	665.66	1	1.1	2 X enolase I; Eno1p
82 7.72e+008	51 2.75e+008	46970.4	6321968	77.5	37	586.40		1.2	enolase; Eno2p
54 1.49e+008	50 1.09e+008	69827.1	417149	71.8	42	587.55	2	2.1	5 X HEAT SHOCK PROTEIN SSA1 (HEAT SHOC
43 1.20e+008	39 9.12e+007	69639.9	6323004	60.4	36	472.10		2.2	member of 70 kDa heat shock protein family; Ssa
18 3.37e+007	16 3.03e+007	70659.6	6319396	23.8	16	212.83		2.3	heat-inducible cytosolic member of the 70 kDa he
10 3.92e+006	8 3.21e+006	74523.7	6322426	19.9	12	141.96		2.4	Involved in translocation of nascent polypeptides
58 4.66e+008	57 3.75e+008	44794.6	10383781	80.5	33	514.41	3	3.1	3-phosphoglycerate kinase; Pgk1p
75 7.02e+008	48 3.43e+008	35845.2	3720	89.7	30	503.30	4	4.1	5 X reading frame
40 4.50e+008	39 2.62e+008	35960.3	6322468	80.4	26	401.83		4.2	glyceraldehyde 3-phosphate dehydrogenase; Tc
32 3.88e+008	38 2.46e+008	35863.4	6322409	84.6	26	368.25		4.3	Glyceraldehyde-3-phosphate dehydrogenase 1:
11 2.03e+008	13 1.24e+008	35609.1	1362244	30.5	9	133.49		4.4	glyceraldehyde-3-phosphate dehydrogenase (ph
45 7.51e+007	33 3.43e+007	93743.9	6320593	54.2	37	491.72	5	5.1	translation elongation factor 2 (EF-2); Eft2p
30 1.20e+007	28 9.44e+006	229542.8	6322666	23.6	37	475.07	6	6.1	pentafunctional enzyme consisting of the following
40 1.55e+008	38 8.84e+007	54943.0	6319279	70.8	33	467.03	7	7.1	Required for START A in the cell cycle and sporu
35 5.27e+007	31 2.25e+007	86029.5	6320936	50	33	450.14	8	8.1	cobalamin-independent methionine synthase; Me
42 1.37e+008	45 7.11e+007	61722.5	6323073	64.6	30	441.60	9	9.1	4 X pyruvate decarboxylase; Pdc1p
9 4.11e+007	11 2.38e+007	62139.1	6323163	19.8	10	120.81		9.2	pyruvate decarboxylase; Pdc5p

Figure 21 Summary of valid results- comparison of two samples

The example in [Figure 21](#) includes two yeast samples. The summary results allow you to compare the proteins in the two samples, to see which are up-regulated and which are down-regulated in the cells.

The color-code in the left columns represents the relative protein concentrations. Dark red (not shown) is highest, orange is intermediate, and yellow is lowest. The top number in each cell is the number of peptides detected for each protein, while the bottom number is the mean peptide intensity. The mean peptide intensity is an average of the intensities for all the peptides assigned to that protein. The peptide intensities are calculated from extracted ion chromatograms from the precursor ions.

These intensity results are sufficient for studies where you are interested in sample differences of two-fold or more.

11 To import data into Excel or LIMS:

- a** Mark the **Excel Export** check box.

You can add this change to the parameter file, or not.

- b** Click **Summarize**.

- c** Check that you see a display with two buttons, as in [Figure 22](#). At this point, a new file with extension ***.ssv** has been created. You can import this file into Excel or upload it to a LIMS system (if configured by your system administrator), or display it on your screen.

- d** Do one of the following:

- To import the data into Excel, import as semicolon-delimited data.
- To upload to LIMS, see the Server Administration online help.

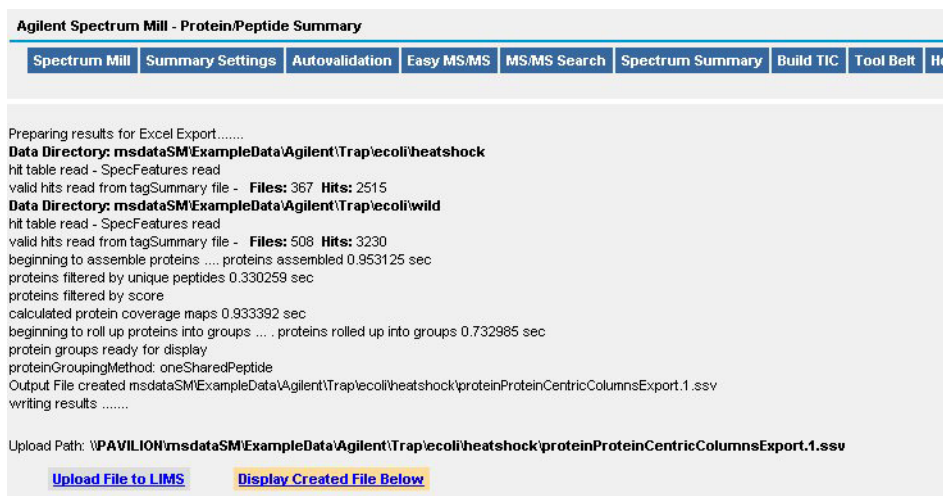


Figure 22 Display when you mark the Excel export check box

12 To print results, use your browser's print function:

- a** Click **File > Page Setup**, and mark the check boxes for **Print background colors and images** and for **Landscape**.
- b** Click **OK**.

- c** Click in the results frame.
- d** Select **File > Print Preview**.
- e** At the top of the Print Preview window, select **Only the selected frame**.
- f** Click the **Print...** button.

NOTE

To export results into HTML or PowerPoint files, see the Frequently Asked Questions section of the online help.

To search in no enzyme mode

- 1** Navigate to the MS/MS Search page.
- 2** Fill in the search page as described in “[Step 1. Run an identity search](#)” on page 36, except:
 - Mark the check box to **Search previous hits**.
 - Set **Digest** to **No enzyme**.
- 3** Run the search.
- 4** Use the Autovalidation page and the Protein/Peptide Summary page to review and validate results. Follow instructions on [page 23](#) through [page 52](#).

To search a larger database

- 1** Navigate to the MS/MS Search page.
- 2** Fill in the search page as described in “[Step 1. Run an identity search](#)” on page 36, except select a different database this time.
 - For example, if you first searched against the SwissProt database, try the larger NCBI nr database.
 - If you first searched a species subset database (e.g., human), search a larger subset (e.g., mammals). This is most important when the original subspecies is not well-represented in the database.

- If you cannot find hits in the standard protein databases, try searching dbEST. It is possible that there is a known expressed sequence tag that contains part of your protein. (Note that dbEST searches take longer.)
- 3 Run the search.
 - 4 Use the Autovalidation page and the Protein/Peptide Summary page to review and validate results. Follow instructions on [page 23](#) through [page 52](#).

NOTE

To search your own custom library, see “[Spectrum Matcher](#)” on page 67.

To search in homology modes, or do a mass gap search

- 1 Follow the instructions in “[Step 4. Search in variable modifications mode](#)” on page 42, but make these substitutions depending on what you want to do:
 - To search for matches that are consistent with a single amino acid substitution that would result from a point mutation (single base substitution within a codon), select **Homology - Single base pair mutations**. This mode also searches for any variable modifications you have selected.
 - To search for matches that are consistent with a single amino acid substitution, where the substitution does not need to be one that would result from a point mutation, select **Homology - All mutations**. This mode also searches for any variable modifications you have selected.
 - To search for an unknown or unexpected modification, select one of the homology modes described above, then click **Unassigned single mass gap**. This search looks for an unexpected modification (a mass gap) or an amino acid substitution (but not both within the same peptide). Note that you cannot simultaneously search variable modifications in this mode.
- 2 If you do a homology mode search, autovalidate and review results as described in “[Step 5. Autovalidate variable modifications data](#)” on page 47 and in “[Step 8. Summarize valid results](#)” on page 52

Or you can create a peptide report with modifications fields selected; see “[Step 4. Examine the validated results for variable modifications mode](#)” on page 27.

Autovalidation is not recommended for the homology modes with the Fixed threshold strategy because of the increased possibility of false positives when so many permutations are considered. But using the Auto threshold strategies (including discriminant) is fine because the FDR automatically adjusts thresholds as the search space gets larger and larger.

NOTE

If you attempt a homology search against a large database, the search will take a very long time and the chance of random matches will increase. If at all possible, conduct a homology search against a file of previously validated hits. See “[Step 3. Create a file of previously validated results](#)” on page 41. Or select **Search previous hits** in the MS/MS Search page. If a file of previously validated results has not yet been created, the program will automatically create the file when the search starts.

If that is not possible, reduce the **Maximum # missed cleavages** to 1.

To use Sherenga *de novo* sequencing

Sherenga *de novo* sequencing is useful as a first step in manual spectral interpretation, or as a confirmation of MS/MS Search results. This section describes briefly how to use the Sherenga *de novo* sequencing software. For additional details and background information, see Chapter 5.

- 1 Navigate from the Spectrum Mill home page to the Sherenga *de novo* Sequencing page, shown in [Figure 23](#).

Agilent Spectrum Mill - Sherenga *de novo* Sequencing - (none)

Spectrum Mill	<i>de novo</i> Summary	Protein/Peptide Summary	Spectrum Summary	Workflows	Tool Belt	Help
---------------	------------------------	-------------------------	------------------	-----------	-----------	------

Sherenga *de novo* Sequencing

Sequence ☒ Remove all prior Sherenga results

Maximum reported hits: 10

Validation filter: good-spectrum-sequence-not-validated

Show Equivalent Masses

I/L: ☒ I ☐ L

K/Q: ☐ K ☒ Q ☐ Both

Data Directory

ExampleData\Agilent\Q-TOF\yeast1

Modifications

Fixed: Carbamidomethylation (C) Variable: Oxidized methionine (M)

Sequencing Parameters

Scoring: Agilent ESI Q-TOF Min. vertex score: 2 ☒ Sequence tag length: > 7

Data Files

Spectrum files (./cpick_in/):

- *.2.pk1
- *.0.pk1
- *.1.pk1
- *.3.pk1

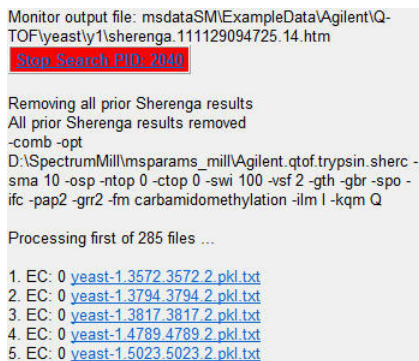
Figure 23 Sherenga *de novo* Sequencing page

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

- 2 To use an already existing parameter file, click **Load**.
- 3 Set the **Validation filter**:
 - If you took the time to designate your spectra as “good” or “bad,” as described in “[Step 6. Check for remaining high-quality spectra](#)” on page 47, set the **Validation filter** to **good-spectrum-sequence-not-validated**. This ensures that only the uninterpreted spectra you designated as “good” are sequenced.
 - Otherwise, set the **Validation filter** to **spectrum-not-marked-sequence-not-validated**.
- 4 Click the **Select...** button to select the data directory where your files reside.
- 5 Click the **Choose...** button to select the modifications that match the chemistry for your samples.
- 6 Under **Sequencing Parameters**, for **Scoring**, select your instrument.
- 7 Set the other parameters. In general, you should keep the defaults except for settings that are highlighted in red text. Click the blue divider bars for more information in the online help.
- 8 To save the settings to a parameter file click **Save As**.

You can use these parameter files for workflow automation.
- 9 Click **Sequence**.
- 10 After clicking **Monitor Results**, check that you see a display similar to that shown in [Figure 24](#). The sequencing program processes the entire directory. A message alerts you when processing is complete.



Monitor output file: msdataSM\ExampleData\Agilent\Q-TOF\yeasty1\sherenga.111129094725.14.htm

Stop Search PID: 2040

Removing all prior Sherenga results
All prior Sherenga results removed
-comb -opt
D:\SpectrumMill\msparams_mill\Agilent.qtof.trypsin.sherc -
sma 10 -osp -ntop 0 -ctop 0 -swi 100 -vsf 2 -gth -gbr -spo -
ifc -pap2 -gr2 -fm carbamidomethylation -ilm 1 -kqm Q

Processing first of 285 files ...

1. EC: 0 [yeast-1.3572.3572.2.pkl.txt](#)
2. EC: 0 [yeast-1.3794.3794.2.pkl.txt](#)
3. EC: 0 [yeast-1.3817.3817.2.pkl.txt](#)
4. EC: 0 [yeast-1.4789.4789.2.pkl.txt](#)
5. EC: 0 [yeast-1.5023.5023.2.pkl.txt](#)

Figure 24 Display while Sherenga runs

To review Sherenga *de novo* sequencing results

- 1 Click the ***de novo* Summary** button to navigate from the Sherenga *de novo* Sequencing page to the Sherenga *de novo* Summary page.
- 2 Check that you see the page shown in [Figure 25](#).

Figure 25 Sherenga *de novo* Summary page

- 3 You should generally keep the default settings, but if you want to load a set of different parameters, click **Load** to load a parameter file.
- 4 If you change any settings, click **Save As** to save the new settings to an existing parameter file or a new one.
- 5 Click **Summarize**.
- 6 Check that you see a display like that shown in [Figure 26](#).

Note that the filenames are in the format
 Data_File_Name.aaaa.bbbb.c.pkl.txt, where

- aaaa = first merged scan
- bbbb = last merged scan
- c = assigned precursor charge (0 means charge was ambiguous)

If you have QSTAR files, see the note on [page 30](#).

1 Processing MS/MS Data Interactively

Iterative Workflow – Alternative Interactive Processing

SS	TL	TS	MS/MS Search Delta MH*	Filename	Sherenga Sequence Tag	MS/MS Search Result
87	8	22.1	0.0	yeast-1.17489.17882.2.pkl.txt	VPTVDVSVDITVQ	VPTVDVSVDLTVK
76	13	22.5	0.0	yeast-1.16410.16787.2.pkl.txt	[213.12]IDTDCQYAYITIGR	VNLDTCQYAYLTGIR
73	11	20.0	0.0	yeast-1.19422.19515.2.pkl.txt	[255.12]MADAINAI[301.13][255.13]	SSVLADALNAINNAEK
67	13	23.7	0.0	yeast-1.19496.19511.2.pkl.txt	[200.12]IWWIEQAY[247.07][154.11]	TVIGEVLLQAYGGMR
61	10	8.9	0.0	yeast-1.16412.16777.3.pkl.txt	[142.08]AM[214.12]YAY[288.07]DP[184.12]A[172.10]	VNLDTCQYAYLTGIR
61	9			yeast-1.17135.17361.2.pkl.txt	[257.11]VYmDVEA[172.07]Q[210.12][213.13]	
59	9			yeast-1.16078.16100.2.pkl.txt	[228.12]SEASVD[270.12][238.03][225.15]	
58	8	19.3	0.0	yeast-1.11439.11888.2.pkl.txt	[171.07]PTVEVP[172.10]VV[234.08]	GNPTVEVELTTEK
58	11	20.5	-0.0	yeast-1.14343.14406.2.pkl.txt	[214.13]ADYDAIDIA[270.15]	LTADYDALDIANR
58	11	23.4	0.0	yeast-1.20456.20456.2.pkl.txt	[214.13]AETIAEm[208.11][185.09][199.13]	TIAETLAEELINAAK
57	8	13.1	-0.0	yeast-1.10000.10416.2.pkl.txt	[241.15]EGVGS[217.15]PVQ	QLENVSSNIVK
57	11			yeast-1.20458.20463.2.pkl.txt	[200.10]mIQEII[226.07]EQ	
56	9	17.9	0.0	yeast-1.17897.18106.2.pkl.txt	[234.11]TmQAIRR[170.10]	T SFFQALGVPTK
55	8			yeast-1.17013.17262.2.pkl.txt	[213.12]IDTDCQYAYIT[326.20]	
53	9	20.1	-0.0	yeast-1.20174.20174.2.pkl.txt	[699.33]AIAAIIQIED[768.36]	IEAALSDAALQIEDPSADELR
52	8			yeast-1.21507.21573.3.pkl.txt	[404.20][265.14]V[273.09]W[345.13]VEIISHII[225.07]	
51	8	18.6	0.0	yeast-1.18520.18763.2.pkl.txt	[170.11]D[262.10]G[200.13]VI[430.15]	AVDDFLLSLDGTANK
50	9			yeast-1.13123.13130.2.pkl.txt	[213.11]EIGT[248.13]AI[225.08]	
49	8			yeast-1.14137.14362.2.pkl.txt	ISVQD[228.11][241.18]	
49	8			yeast-1.18517.18517.2.pkl.txt	[602.29]MIT[228.13]ATAD[566.27]	
49	10	17.6	0.0	yeast-1.5860.5860.2.pkl.txt	[276.10]N[316.11]AEWA[210.16]Am[226.06]	KAEASGEAADEADEE
48	9	16.7	0.0	yeast-1.17912.18171.2.pkl.txt	[200.12]GVVV[260.15]DmTQ	SIGGEVFIIDFTK
48	11			yeast-1.20438.20438.2.pkl.txt	PVSISETIE[236.19][247.10]	
48	9	18.5	0.0	yeast-1.9018.9018.2.pkl.txt	[202.10]EEININ[202.10]PDR	TTEEQVENAVDR
47	9	19.2	0.0	yeast-1.10534.10534.2.pkl.txt	[416.13]SVT[226.10][158.08][246.12][168.05]	ADDDSVTIISAGNDK
47	10			yeast-1.12462.12462.2.pkl.txt	W[343.15]SQNAWIQ[241.14][316.13]	
47	8	16.8	0.0	yeast-1.13991.14234.3.pkl.txt	[200.11][487.28]SA[376.17]SE[185.12]	ISHVSTGGGASLELEGK
47	8			yeast-1.17464.17464.2.pkl.txt	[229.11][170.09][241.14]V[361.12][154.10]EA[170.11][268.11]	
47	9			yeast-1.22296.22413.3.pkl.txt	[288.11]NINMDI[221.53][184.11]Q[184.10][226.23]NI[242.13]	
46	10			yeast-1.21302.21302.2.pkl.txt	[1021.48][227.13]IAIAAY[200.11][499.19]	
46	9	15.4	0.0	yeast-1.7543.7888.2.pkl.txt	[241.14]ETNI[213.11][142.04]V[225.15]	LQETNPEEVPK
46	9	18.4	-0.0	yeast-1.9716.9716.2.pkl.txt	[229.07]VDVI[185.08][255.09][299.15]	NDVDVIANPDGER
45	11	16.3	0.0	yeast-1.14895.14895.2.pkl.txt	[953.51]YWVSII[224.09]	NQILVSGEIPSTLNEESK
45	10	13.7	-0.0	yeast-1.6780.7004.2.pkl.txt	[142.08]SWIHNA[255.12]	AASDAIPASPCK
45	9	21.8	-0.0	yeast-1.9857.9857.2.pkl.txt	[242.13]SSVESQm[371.20]	IESSVESQFSAGR
44	11	18.6	-0.0	yeast-1.16408.16408.2.pkl.txt	[349.13]ANGVI[214.16]TVI[225.15]	ADYANGVLTITVPMK

Figure 26 Sherenga top-level report

7 Review Sherenga results:

- Click a link under the **Filename** header. Check that you see the Spectrum Viewer as shown in [Figure 27](#) and detailed Sherenga results as shown in [Figure 28](#). The detailed results list multiple Sherenga interpretations for the same spectrum. The most probable interpretation (Sherenga 1) is initially displayed at the top of the Spectrum Viewer.

- b** Click the **Rank** arrow buttons shown in the lower right-hand corner of [Figure 27](#) to navigate through the *de novo* possibilities. See “[To use the Spectrum Viewer](#)” on page 96 of [Chapter 2](#) for more details.
 - If you see an **MStag 1** sequence at the top of the Spectrum Viewer, first click the **Sherenga 1** sequence and make sure the Sherenga sequence turns from gray to white background. Then click the **Rank** arrow buttons.
- c** To display a user-defined sequence above the spectrum, type the sequence in the box to the left and then click the **Go** button.
- d** Scroll below the display shown in [Figure 28](#) to view details of the Sherenga interpretation.

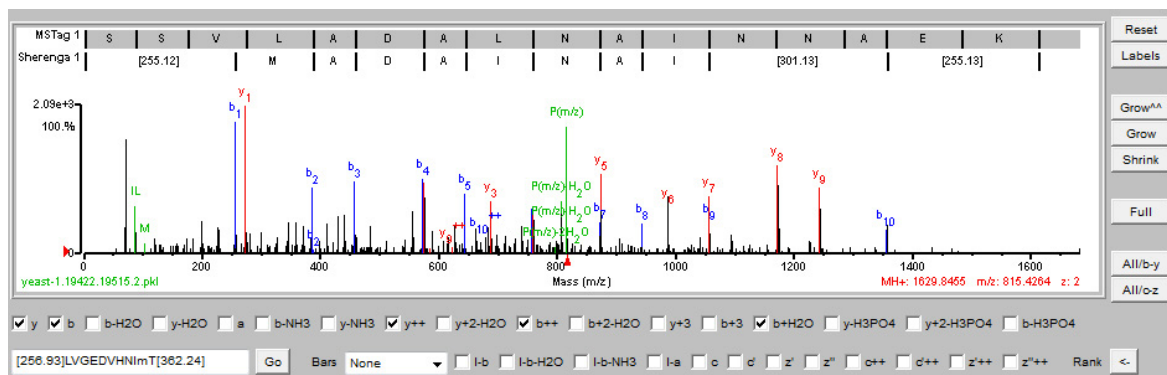


Figure 27 Spectrum Viewer with Sherenga results

1 Processing MS/MS Data Interactively

Example Data		Agilent Q-TOF		yeast1		yeast-1.19422.19515.2.pkl.txt										yeast-1.19422.19515.2.pkl.spo															
Rank	Orig. Score	Score	MS-Tag Sequence																		Sherenga Sequence Tag	Matched Precursor MH+									
20.02			SSVLADALNAINNAEK																		[255.12][MADAINAI][301.13][255.13]										1629.8442
			S	S	V		L	A	D	A	L	N	A	I	N	N	A	E	K												
1.	74.6	74.6	[255.12]				M	A	D	A	I	N	A	I	[301.13]			[255.13]			1629.85										
2.	73.7	73.7	[255.12]				M	A	D	A	I	N	A	I	[188.05]			I	[255.13]												
3.	72.4	72.4	[272.14]				N	A	D	A	I	N	A	I	[301.13]			[255.13]													
4.	71.9	71.9	[255.12]			[202.13]		D	A	I	N	A	I	[188.05]			I	[255.13]													
5.	71.5	71.5	[272.14]				N	A	D	A	I	N	A	I	[188.05]			I	[255.13]												
6.	71.3	71.3	[273.14]				I	A	D	A	I	N	A	I				[566.25]													
7.	70.7	70.7	[255.12]				M	A	D	A	I	N	A	I				[566.25]													
8.	70.7	70.7	[386.22]				A	D	A	I	N	A	I	[301.13]				[255.13]													
9.	70.4	70.4	[273.14]				I	A	D	A	I	N	A	I	[188.05]				[368.21]												
10.	70.3	70.3	[273.14]				I	A	D	A	I	N	A	I	[216.04]				[340.21]												

Figure 28 Detailed Sherenga results

8 Compare Sherenga results with MS/MS Search results. When database search results are available, the software displays them along with the Sherenga interpretations, so you can compare the Sherenga and MS/MS Search sequences. When MS/MS Search results exist, the database sequence (labeled **MSTag 1**) is displayed just above the Sherenga sequence (labeled **Sherenga 1**). The latter is shown at the top of [Figure 27](#).

Additional Tools

Spectrum Matcher

You use the Spectrum Matcher to search unknown spectra against another set of library spectra on the Spectrum Mill server. The library spectra may reside in a **Collections** directory, or any other data directory on the server.

To create a library in your collections directory, see [“To copy spectra to your collections directory”](#) on page 190.

To search the spectra

- 1 On the Spectrum Mill home page, click the **Spectrum Matcher** link.
- 2 Check that you see the page shown in [Figure 29](#).
- 3 Click **Select** to select the **Query Set** data directory; then click **Select** to select the **Library Set** data directory.
- 4 Fill in the options. They are self-explanatory or are explained in the online help.
- 5 Click the **Match** button.
- 6 If you have a large data set and you see an error that the CGI input string is too long, use the **Search result files** box under **Query Set** to search a subset of the file names.

For example, you can search just a single data file, typed into the Search result files box.

- 7 After time for processing, check that you see results like those shown in [Figure 30](#).

Note that file names are in the format Data_File_Name.aaaa.bbbb.c.pkl, where

- aaaa = first merged scan
- bbbb = last merged scan
- c = assigned precursor charge (0 means charge was ambiguous)

If you have QSTAR files, see the note on [page 30](#).

1 Processing MS/MS Data Interactively

Additional Tools

Agilent Spectrum Mill - Spectrum Matcher

Spectrum Mill | **Match Settings** | **Spectrum Summary** | **Protein/Peptide Summary** | *de novo* | **Help**

Match Spectra

Match | **Save Settings** | **Reset**

Search Criteria

Spectral Quality

☒ Sequence tag length: > 3

Minimum # of peaks: 4

Search Mode

Search mode: Identity

Matching Tolerances

Minimum matched peak intensity: 50 %

Precursor m/z: +/- 1.5 Da

Product m/z: +/- 0.3

MS/MS Peak Detection

Minimum S/N: 8 < 10 Peaks (most intense ¹²C)

Data Sets

Query Set

Select... ExampleData/Agilent/Trap/ecoli/wild

Validation state: spectrum-not-marked-sequence-not-validated

Search result files:

NEC_0020*.pk1

Library Set

Select... ExampleData/Agilent/Trap/ecoli/heatshock

Validation state: valid

Search result files:

*.pk1

Figure 29 Spectrum Matcher

Agilent Spectrum Mill - Spectrum Matcher

Spectrum Mill

Match Settings

Spectrum Summary

Protein/Peptide Summary

de novo

Help

	Query	Library
Directory:	ExampleData/Agilent/Trap/ecoli/wild	ExampleData/Agilent/Trap/ecoli/heatshock
Validation State:	spectrum-not-marked-sequence-not-validated	valid
# spectra:	254	266

Matched Spectra

Score	SPI (%)	# Query Spectrum Peaks	Query num	query file	Query Parent m/z	Library Parent m/z	Delta Parent m/z	Delta Parent MH+	Library num	library file	
1	6.8	80.4	10	3	NEC_0020.0967.0967.2.pkl	523.5	523.5	0.0	0.0	2	HEC_0020.0944.0952.0.pkl
2	5.6	60.0	10	3	NEC_0020.0967.0967.2.pkl	523.5	523.5	-0.0	-0.1	3	HEC_0020.0923.0925.0.pkl
3	6.7	72.2	10	20	NEC_0020.0993.0993.0.pkl	669.4	669.5	-0.2	-0.4	62	HEC_0020.1003.1003.2.pkl
4	7.8	82.0	10	20	NEC_0020.0993.0993.0.pkl	669.4	669.6	-0.3	-0.5	63	HEC_0020.0976.0979.0.pkl
5	5.6	61.9	10	29	NEC_0020.0958.0964.2.pkl	723.3	723.0	0.3	0.6	84	HEC_0020.0937.0943.2.pkl

Figure 30 Spectrum Matcher results

To examine the results

Click the links in the report to view and compare the spectra:

- Click a number link on the left side of the report to view both the query spectrum and the library spectrum in side-by-side Spectrum Viewer windows at the bottom of the page.
- Click a **query file** link to view the query spectrum in the Spectrum Viewer at the bottom left side of the page.
- Click a **library file** link to view the library spectrum in the Spectrum Viewer at the bottom right side of the page.

You cannot view these results with the Protein/Peptide Summary page.

Peak Picker

You use the Peak Picker (menu button on the Data Extractor page) to independently troubleshoot the peak detection done in MS/MS Search. The output of Data Extractor is used as input to both the Peak Picker and MS/MS Search. The Peak Picker displays a comparison of raw spectra (top) and peak-detected spectra (bottom). After determining revised parameters with Peak Picker, your system administrator can edit the peak detection parameters for the appropriate instrument in the configuration file `mparams_mill\instrument.txt`.

CAUTION

Before you modify any files, be sure to archive the originals so they can be restored if necessary.

Build TIC

Build TIC constructs a graph that shows interpreted spectra (red) and uninterpreted spectra (black), as a function of scan number. It also provides statistics on the interpretation percentage of the spectra or the ion current.

With Build TIC, you can select a neutral loss to locate spectra exhibiting that loss. This is useful to locate peptides that have a phosphorylation site or that readily lose water. See [Figure 31](#) as an example.

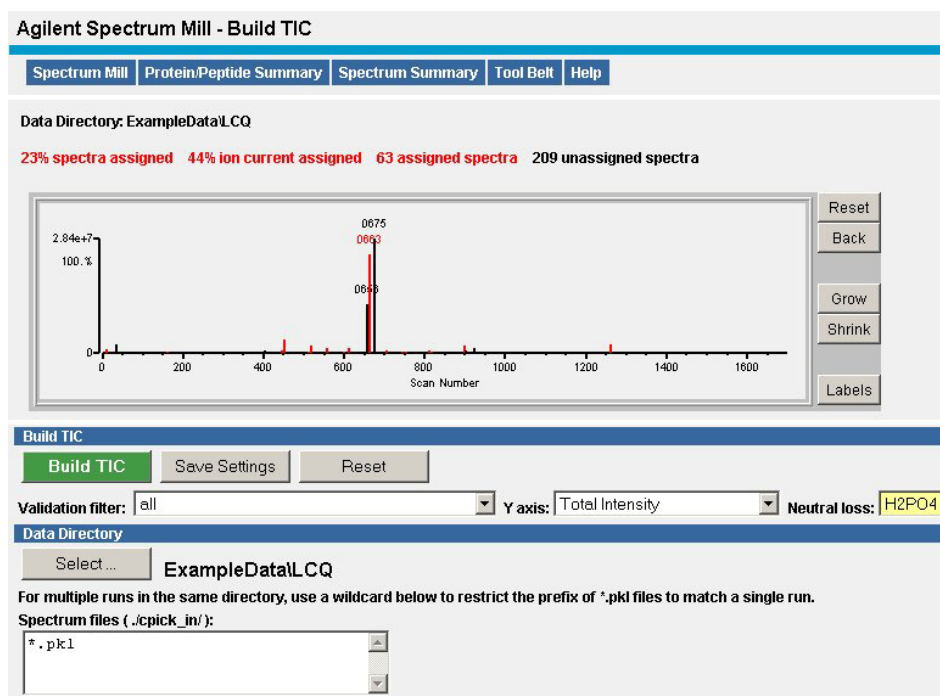
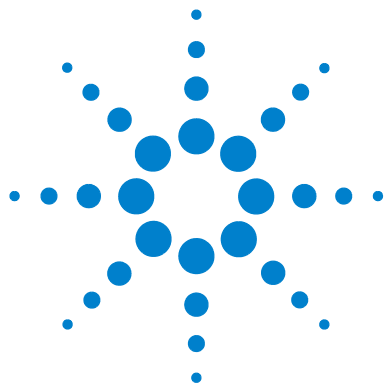


Figure 31 Use of Build TIC to locate scan numbers of spectra that indicate potential phosphorylation sites



2 Reviewing and Validating MS/MS Data

Setting Up Autovalidation	73
Setting Up the Protein/Peptide Summary Page	81
Reviewing and Validating Results Interactively	89
Protein/Peptide Summary Displays	104

In this chapter you learn in detail how to review and validate database search results from MS/MS analyses of peptides and protein digests. Validation means that you accept the database search results for a particular MS/MS spectrum or group of spectra.

The Spectrum Mill workbench provides a means to segregate search results that contain a valid interpretation of an MS/MS spectrum from those that do not. You can then subject results that are *not* validated to subsequent rounds of searches (against other databases or in variable modifications mode, for example). You can summarize results that *are* validated in a results table.

To segregate the valid search results, the software keeps track of both the spectra and their interpretations in a coordinated way. Furthermore, the software permits spectra to be segregated according to quality (via Spectrum Summary) without regard to the validation states of their database interpretations. Because the software keeps track of which spectra you have segregated as “good” and which database interpretations you have designated as “valid,” you can intelligently apply processing steps to individual groups of spectra.

In this chapter, you learn how to use the MS/MS Autovalidation page to quickly validate the search results with high scores. You learn how to use the Protein/Peptide Summary page to manually review and validate search results that have not been autovalidated. You learn how the various Protein/Peptide Summary data displays are organized to provide insight



into different types of scientific problems. You also learn how to manipulate results in the Spectrum Viewer, an indispensable tool to visualize the sequence information contained in MS/MS spectra.

If you want a quick introduction to this subject matter, without a lot of details, see the relevant sections of [Chapter 1](#), “Processing MS/MS Data Interactively”.

Setting Up Autovalidation

Automatic results validation saves time and lets you use different threshold strategies to meet the data quality needs of your experiment.

Step 1. Select Data Directory and/or parameter file

- 1 Navigate to the **MS/MS Autovalidation** page, as shown in [Figure 32](#).

You navigate to this page from the Spectrum Mill Home page, the MS/MS Search page or the Protein/Peptide Summary page.

Agilent Spectrum Mill - MS/MS Autovalidation - (none)

Help

Validate Files Queue request Undo Last Clear All Save As... Load...

Data Directories

Select... (none selected)

Fragmentation mode: All

Search result files: *.apo

Validation Parameters

Strategy: ☐ Fixed thresholds ☒ Auto thresholds ☐ Auto thresholds - discriminant

Mode: Peptide Auto determine using score, delta R1-R2 thresholds to reach a target FDR

Optimize score & R1-R2 score thresholds for each run with max FDR: 1.2 %

Precursor charge range: 2 to 6 Required AAs: any Disallowed AAs: none

Filtering

Automatic variable range for each run

Fixed range for all runs

None (ppm) Auto precursor mass error Fixed precursor mass error Low -1.0 High 30.0 ppm

None (SC/pl) Auto SCX Solution Charge, pH3 Fixed Solution Charge Low -2 High 6

Auto OGE/IEF peptide pl Fixed peptide pl Low 3.0 High 10.0

Figure 32 Default Autovalidation page: Auto thresholds strategy - Peptide mode

- 2 Select the **Data Directory** for which you want to validate results.
- 3 To use an already existing set of parameters, click **Load**.

Step 2. Select the Strategy and Mode to autovalidate results

For more detailed explanations of the strategies/modes and false discovery rate (FDR), click the Validation Parameters blue bar.

To use the Auto thresholds strategy

You can use this strategy in a basic workflow, but not an iterative one. After validating in peptide mode first (always first), you have the option of refining the validation with the Protein polishing mode.

Peptide mode This strategy optimizes the score threshold and the delta R1-R2 threshold to reach a target FDR.

- 1 Select the **Peptide** mode, if necessary. (Figure 32)
- 2 Type in a new **max FDR** if necessary.
1.2% was chosen because it typically produces an actual FDR of 1%.
- 3 Select a filter of interest.
- 4 Move to “[Step 3. Save parameters and validate data](#)” on page 80.

Protein polishing mode (optional) This strategy optimizes the score threshold and the delta R1-R2 threshold to reach a target FDR. Use this mode to remove suspicious proteins only after you have autovalidated in Peptide mode.

- 1 Select the **Protein Polishing** mode. (Figure 33)
- 2 Change the parameters as you intend.

A **Minimum protein score** of 0 and a **maximum protein FDR** of 1.0% is typically used for Agilent Q-TOF data.

See the Online Help for explanations of each field.

- 3 Move to “[Step 3. Save parameters and validate data](#)” on page 80.

The screenshot shows the 'Validation Parameters' dialog box. Under the 'Strategy' section, 'Auto thresholds' is selected. The 'Mode' is set to 'Protein polishing'. A red text label indicates 'Auto determine using score, delta R1-R2 thresholds to reach a target FDR'. Below this, there are input fields for 'Minimum protein score' (20.0), a checked box for 'Group proteins across all directories', 'Automatically raise minimum protein score to yield maximum protein FDR' (0.0 %), and a checked box for 'Peptide FDR for validated proteins' (1.0 %).

Figure 33 Auto thresholds strategy - Protein polishing mode

To use the Auto thresholds - discriminant strategy

You can use this strategy in either a basic workflow or an iterative one. It is typically used in an iterative workflow. After validating in peptide mode first (always first), you have the option of refining the validation with the Protein polishing mode.

Peptide mode This strategy optimizes the discriminant score thresholds until a target FDR is reached.

- 1 Select the **Peptide** mode. (Figure 34)
- 2 Click **Local**, and type an **FDR %**.

The Auto thresholds - discriminant strategy with Local FDR % marked in the Peptide mode is the only Auto thresholds strategy that can be used in iterative workflows. You cannot use this strategy with Global % marked in an iterative workflow.

- 3 Move on to “[Step 3. Save parameters and validate data](#)” on page 80.

The screenshot shows the 'Validation Parameters' dialog box. Under the 'Strategy' section, 'Auto thresholds - discriminant' is selected. The 'Mode' is set to 'Peptide'. A red text label indicates 'Auto determine using discriminant score thresholds to reach a target FDR'. Below this, the 'Local' radio button is selected, and the 'FDR' is set to 1 %.

Figure 34 Auto thresholds - discriminant strategy - Peptide mode

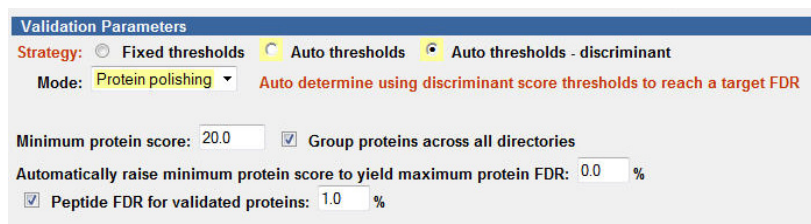
Protein polishing mode (optional) This strategy optimizes the discriminant score threshold and the delta R1-R2 threshold to reach a target FDR. Use this mode to remove suspicious proteins only after you have autovalidated in Peptide mode.

1 Select **Protein Polishing** mode. (Figure 35)

2 Change the parameters as you intend.

A **Minimum protein score** of 0 and a **maximum protein FDR** of 1.0% is typically used for Agilent Q-TOF data.

3 Move on to “[Step 3. Save parameters and validate data](#)” on page 80.



The screenshot shows the 'Validation Parameters' dialog box. The 'Strategy' section has three radio buttons: 'Fixed thresholds' (unselected), 'Auto thresholds' (unselected), and 'Auto thresholds - discriminant' (selected). Below this, the 'Mode' is set to 'Protein polishing' in a dropdown menu. A red text label reads 'Auto determine using discriminant score thresholds to reach a target FDR'. The 'Minimum protein score' is set to 20.0. There is a checked checkbox for 'Group proteins across all directories'. The 'Automatically raise minimum protein score to yield maximum protein FDR' is set to 0.0%. There is a checked checkbox for 'Peptide FDR for validated proteins' set to 1.0%.

Figure 35 Auto thresholds - discriminant strategy - Protein polishing mode

To use the Fixed thresholds strategy

This strategy summarizes results by protein and peptide, uses fixed thresholds for validation and can calculate a false discovery rate (FDR) using reversed hits. If you want the program to calculate an FDR using reversed hits, you must have marked the **Calculate reversed database scores** check box on the MS/MS Search page.

The two modes for this strategy are Protein details and Peptide. For high quality results you autovalidate first with Protein details mode and then with Peptide mode. These are not optional steps.

You can use both of these modes in an iterative workflow.

Protein details mode Using this mode is the first step in a complete validation.

- In the Protein details mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein.
- Using the default scoring, individual peptides typically have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score typically is greater than 20.

1 Select **Protein details** mode. (Figure 36)

2 Change the **Minimum protein score** if necessary.

- 3 If the data files from a single sample occupy multiple directories, mark the check box for **Group proteins across all directories**.
- 4 If you intend to filter the data, click **Fixed precursor mass error** and type in the limits.
- 5 Change the **Protein Rules** thresholds, if necessary.
- 6 Move on to [“Step 3. Save parameters and validate data”](#) on page 80.

Peptide mode For this strategy, using Peptide mode is the second step for a complete autovalidation. You can also use this mode in an iterative workflow.

- In this mode, the software summarizes results by peptide. Even if it finds only a single peptide that corresponds to a protein, it validates the search results if the peptide score is high enough.
 - Using the default scoring, individual peptides typically have scores greater than 11 to 15 (depending on charge state). This score threshold is higher than in the **Protein details** mode, where you have the additional assurance of knowing you have identified more than one peptide per protein.
- 1 Select the **Peptide** mode ([Figure 37](#)).
 - 2 Enter abbreviations for the amino acids you intend to allow or disallow.
 - 3 Select filters you intend to use.

If you wish to autovalidate only the peptides in a given pI range, select **Auto OGE/IEF Peptide pI**, or select **Filter Peptide pI** and type in a pI range.

If you ran an OFFGEL fractionation, and the data files from all pI fractions are in the same directory, then use the **Search result files** filter to select the specific file(s) for which that pI range applies.

- 4 Change the **Peptide Rules**, if necessary.
- 5 Move on to [“Step 3. Save parameters and validate data”](#) on page 80.

2 Reviewing and Validating MS/MS Data

Setting Up Autovalidation

NOTE

If you wish to use the pI filter for modified peptides, ask your server administrator to first verify that the pK of the modified amino acid is specified in **smconfig.std.xml** or **smconfig.custom.xml**.

Agilent Spectrum Mill - MS/MS Autovalidation - (none)

Help

Validate Files ☐ Queue request **Undo Last** **Clear All** **Save As...** **Load...**

Data Directories

Select...

☐ ExampleData/Agilent/Q-TOF/OGE

Fragmentation mode: All

Search result files:

*.spo

Validation Parameters

Strategy: ☒ Fixed thresholds ☐ Auto thresholds ☐ Auto thresholds - discriminant

Mode: Protein details **Fixed score thresholds, calculate an FDR**

Minimum protein score: 20.0 ☐ Group proteins across all directories

☒ Calculate FDR using reversed hits

Filtering **Fixed range for all runs**

☒ None (ppm) ☐ Fixed precursor mass error Low -1.0 High 30.0 ppm

Protein Rules

Rule	Precursor Charge	Score Threshold	% SPI Threshold	Fwd - Rev Score Threshold	Rank 1-2 score Threshold
1.	2	6.0	60.0	☐	☒
2.	1	6.0	70.0		2.0
3.	3	8.0	70.0		2.0
4.	4	8.0	70.0		2.0
5.	5	12.0	70.0		2.0
6.	2	6.0	90.0		1.0

Figure 36 Fixed thresholds strategy - Protein details mode

Agilent Spectrum Mill - MS/MS Autovalidation - (none)

Help

Validate Files ☐ Queue request

Data Directories

Select... ☐ ExampleData/Agilent/Q-TOF/OGE

Fragmentation mode: All

Search result files: *.spo

Validation Parameters

Strategy: ☒ Fixed thresholds ☐ Auto thresholds ☐ Auto thresholds - discriminant

Mode: Peptide Fixed score thresholds, calculate an FDR

☒ Calculate FDR using reversed hits Required AAs: any Disallowed AAs: none

Filtering **Automatic variable range for each run** **Fixed range for all runs**

☒ None (ppm) ☐ Auto precursor mass error ☐ Fixed precursor mass error Low -1.0 High 30.0 ppm

☒ None (SC/pl) ☐ Auto SCX Solution Charge, pH3 ☐ Fixed Solution Charge Low -2 High 6

☐ Auto OGE/IEF peptide pl ☐ Fixed peptide pl Low 3.0 High 10.0

Peptide Rules

Rule	Precursor Charge	Score Threshold	% SPI Threshold	Fwd - Rev Score Threshold ☐	Rank 1-2 score Threshold ☒
1.	2	11.0	60.0		2.0
2.	1	13.0	70.0		2.0
3.	3	13.0	70.0		2.0
4.	4	13.0	70.0		2.0
5.	5	15.0	70.0		2.0

Figure 37 Fixed thresholds strategy - Peptide mode

Step 3. Save parameters and validate data

Before you validate data, you can choose to save any parameters you changed to the same or different parameter file and you can choose to place this command in the queue.

- 1 To save the settings you changed to a parameter file, click **Save As**, name the file, then click **Save**.

This file can now be used as a task in workflow automation.

- 2 To delay execution of the command or have it be part of the interactive automation process, mark **Queue request** to place the command in the queue.

To see results right away, do not mark Queue request.

- 3 Click **Validate Files**.

You see a validation summary that lists the hits and spectra that have been validated.

- 4 If you have completed autovalidation, move on to [“Setting Up the Protein/Peptide Summary Page”](#) on page 81.

If you intend to autovalidate with the second mode in a strategy, repeat [“Step 2. Select the Strategy and Mode to autovalidate results”](#) on page 74 with the second mode.

- Auto thresholds strategy - [“Protein polishing mode \(optional\)”](#) on page 74
- Auto thresholds - discriminant strategy - [“Protein polishing mode \(optional\)”](#) on page 75
- Fixed thresholds strategy - [“Peptide mode”](#) on page 77

Setting Up the Protein/Peptide Summary Page

To set up the Protein/Peptide Summary page, you select options in three categories:

- **Summarize Results for Review** - Select data directory and results display mode.
- **Validation and Sorting** - Select validation, sorting and filtering options.
- **Review Fields** - Mark check boxes to choose columns to include in the report.

Then you click the **Summarize** button to summarize results.

Step 1. Set data directory and mode

- 1 Navigate to the Protein/Peptide Summary page.
- 2 Click the **Select...** button to select the data directory where your files reside.
- 3 To upload a set of existing parameter values, click **Load**.
- 4 Click the **Mode** drop-down list and select one of the summary modes shown in [Figure 38](#). Use [Table 3](#) to guide you.

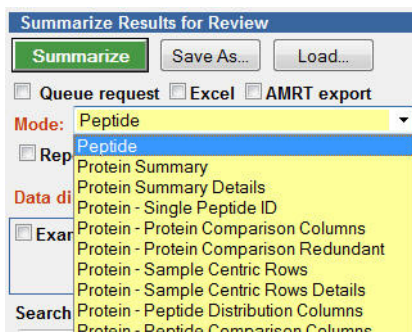


Figure 38 Selecting the data display mode

2 Reviewing and Validating MS/MS Data

Setting Up the Protein/Peptide Summary Page

Table 3 Protein/Peptide Summary Modes

If you want to:	And you want results organized by:	Then use this mode:	Example application
Validate results	Peptides	Peptide	Manual review and validation of MS/MS search results, organized by peptide
	Proteins, then peptides	Protein Summary Details	Manual review and validation of MS/MS search results, organized by protein
	Samples, then proteins, then peptides	Protein-Sample Centric Rows Details	Manual review and validation of MS/MS search results, organized by sample
Summarize results by proteins	Proteins only	Protein Summary	List of all proteins identified in the data
	Proteins, then samples	Protein-Protein Comparison Columns	Compare two or more samples, each of which may contain multiple fractions. Each sample (with all fractions) is organized in a separate directory.
	Proteins, then redundant hits	Protein-Protein Comparison Redundant	Same as immediately above, with additional detail on isoforms of proteins
	Proteins, then peptides	Protein Summary Details	View proteins, with supporting peptide details
Summarize results by samples	Samples, then proteins	Protein-Sample Centric Rows	View proteins from multiple 2D gel spots organized in a single directory
	Samples, then proteins, then peptides	Protein-Sample Centric Rows Details	Same as immediately above, with supporting peptide details
Summarize results by peptides	Peptides only	Peptide	List of all peptides identified in the data
	Peptides, then samples	Protein-Peptide Distribution Columns	Method development (evaluation of 2D LC/MS/MS or other fractionation scheme)
	Peptides, then samples	Protein-Peptide Comparison Columns	Evaluation of fractionation scheme (provides more information and easier export to Excel)
View a list of proteins identified via a single peptide	Peptides only	Protein-Single Peptide ID	Examination of results where a single peptide was used to generate a protein identification

For more details on the displays, see [“Protein/Peptide Summary Displays”](#) on page 104.

Step 2. Set filtering, sorting, and validation parameters

In this step, you determine which data are displayed, how the data are sorted, and the initial validation state of the data. As you set parameters in this section, it is useful to keep in mind the following:

- Filters limit the data that is viewed. For example, they can be used to show only results where search scores exceed a chosen value.
- Sorting parameters determines the order of the data that is viewed.
- Validation presets determine the initial settings of the validation states for the data that is viewed. You can always change these presets after you review the data. They are not made a permanent part of the data record at this point.
- A validation state represents your judgment as to whether a particular data interpretation (e.g., a search result) is satisfactory.

Figure 39 Validation, sorting and filtering parameters - Protein Summary Details mode

- Set the validation, sorting, and filtering parameters (Figure 39).

To set these parameters, refer to the instructions below and the *Online Help* by clicking the Validation and Sorting blue bar.

Note that the specific options that are available depend upon the mode you selected earlier.

- Select the **Filter results by** validation filter to determine which data are displayed: sequence-not-validated, valid, good-spectrum-sequence-not-validated or all.
- Select the **Validation preset** to determine how you categorize your data upon initial display: none, valid, reset or status.
- Select the **Protein-grouping method** to determine how proteins are grouped in certain protein summary modes: 1 shared peptide or 1 shared peptide, expand subgroups.
- From the **Sort proteins by** list, select a data sorting method for your report.

Choices include score, number of peptides, protein molecular weight, category and function.

- Set a **Filter by protein score**.

Scores >15-18 almost always represent valid results for Q-TOF data.

- From the **Sort peptides by** list, select a sorting method.

Note that when you sort by accession number, the sort is alphabetical rather than numerical. This is because some databases do not have strictly numeric accession numbers.

- Set a **Filter by peptide score**.

For ion trap data, peptide scores greater than 13 almost always represent valid results. Scores below 6 generally represent poor results. For Agilent Q-TOF data, you search with a narrower mass tolerance; so there is a better chance that lower-scoring results are valid. It is not unusual for a score of 5 to represent a valid result, but only if the peptide is short or in low abundance.

- Set a %SPI, or filter for scored peak intensity.

For ion trap data, values greater than 70 (or 60% for doubly-charged precursor ions) generally represent high-quality results. For Agilent Q-TOF spectra, % SPI below 60 may represent a poor interpretation. The lower the value, the greater the likelihood that the MS/MS spectrum contains significant peaks not explained by the database search results.

- From the **Required AAs** list select the amino acid(s) that must be present in the peptides that are shown. To disable, select **any**.

- From the **Disallowed AAs** list select the amino acid(s) that cannot be present in the peptides that are shown. To disable, select **none**.
- Type a range of peptide pI values. Mark **All** to ignore the range.

This field appears only when you have marked **Peptide pI** in the Review Fields.

- Type or paste a list of accession numbers in various formats (space-separated, separated by '|', comma-separated, etc.). To disable, leave the box blank.

NOTE

Filters that are more stringent produce higher confidence results. Set the score filters in a way that makes sense for your study. For some studies, it may be better to provide fewer results, albeit with greater probability of being correct. For others, it may be better to provide more results and take the risk that some false positives could be reported.

Your system administrator can configure the options for **Required AAs** and **Disallowed AAs**.

If you wish to use the pI filter for modified peptides, ask your server administrator to first verify that the pK of the modified amino acid is specified in **smconfig.std.xml** or **smconfig.custom.xml**.

Step 3. Choose review fields

In this step you mark the check boxes for data you wish to display.

Click the Review Fields blue bar to see the descriptions of all the review fields for every data display mode.

Note that some fields are unavailable in certain data display modes. For example, fields that pertain only to peptides are available only in the modes that display peptide data.

The 'Review Fields' dialog box is divided into three columns. The first column contains: ☒ Filename, ☒ Score, ☐ FDR (Discriminant), ☐ Fwd-Rev score, ☐ Rank 1-2 score, ☒ SPI (%), ☐ Unmatched ions, ☐ Var mod sites, ☐ VML score (with a dropdown menu showing 's|t|y'), ☐ Solution charge, and ☐ Start AA position. The second column contains: ☒ Sequence (with a 'b/y map' checkbox), ☐ Rev Sequence, ☐ Rank 2, ☐ Var mod sequence, ☐ Retention time, width, ☐ Precursor m/z, ☒ MH⁺, ☐ Delta mass, ☐ Peptide pl, ☐ Protein MW, ☐ pl, ☐ Species, ☒ Accession #, and ☒ Protein name. The third column contains: ☒ Intensity, ☐ DEQ ratios, ☐ Invert, ☐ Reporter Ratios (with a dropdown menu showing 'iTRAQ 4'), ☐ Control (with a dropdown menu showing '114'), ☐ Intensities, ☐ Modification names, ☐ N-term, ☐ C-term, ☐ Cysteines, ☐ Fragmentation mode, ☐ Max tag length, ☐ Longest tag, ☐ # Backbone Cleavages, ☐ # b/y pairs, and ☐ Category.

Figure 40 Fields for display in Peptide page

- 1 Choose from items shown in [Figure 40](#).

The marked check boxes are ones you would typically use.

- 2 Mark any others that suit your experiment.

Here are instructions for changing a few.

- Mark **Fwd-Rev score** to see the difference between scores for top hits from forward and reversed database searches.

If you obtain similar scores for both searches, there is an increased likelihood of an incorrect assignment.

- Mark **Rank 1-2** score to see the difference between rank 1 and rank 2 database search scores.

- Mark **Var mod sites**, **VML score** and/or **Var mod sequence** to display sites of variable modifications and amino acid substitutions, the sequences around the sites, and the scores.
- Mark **Peptide pI** to display peptide pI and enable filtering by peptide pI.
- Mark **DEQ ratios** to calculate and display ratios for differential expression quantitation, such as light/heavy ratios for ICAT reagents or other reagents that use isotopic labels.

See “[To calculate DEQ ratios for isotopic labels using the Protein/Peptide Summary page](#)” on page 144 of [Chapter 5](#).

- Mark **Reporter Ratios** and from the iTRAQ/TMT list select the tag type used for quantitation – iTRAQ4, iTRAQ8, TMT2 (tandem market tag) or TMT6 – and mark intensities to display intensities for the marker ion masses.
- Mark **Ratios control**, then from the Ratios control list, select the iTRAQ mass for the denominator for ratio calculations
- Mark **Fragmentation mode** to show which fragmentation mode, for example, collision-induced dissociation (CID) or electron transfer dissociation (ETD), was used for searching the results.

Step 4. Select Protein Quantitation Options

- Select one or more of the options in [Figure 41](#) in order to filter out data for protein quantitation.

Note that the third option appears only when you mark the DEQ ratios check box in the Review Fields panel.

For complete descriptions of each of these options, click the Protein Quantitation Options blue bar.

Protein Quantitation Options	
☐	Exclude poor isotope quality Precursor XIC's (< 0.85 Chi2 vs. Average)
☐	Exclude poor Precursor Isolation Purity: < 75 %
☐	Exclude outlier DEQ Ratios (> 2 std dev from mean)

Figure 41 Protein Quantitation Options when DEQ ratios are marked in Review Fields

Step 5. Save parameters and summarize data

Before you summarize data, you can save the changed parameters, place the command in the queue and/or choose to export the summarized data to an Excel spreadsheet or indirectly to a MassHunter database.

- 1 To save the settings you changed to a parameter file, click **Save As**, name the file, then click **Save**.

This file can now be used as a task in workflow automation.

- 2 To delay execution of the Summarize command or to place the command in a queue to execute later as part of interactive automation, mark **Queue request**.

- 3 To export protein or peptide data to an Excel spreadsheet, mark **Excel export**, or

To export peptide data indirectly (through a CSV file) to an AMRT (Accurate Mass and Retention Time) database, mark **AMRT export**.

Which option is available depends on the data display mode. In Peptide mode both options are available, but you can choose only one.

- 4 Click **Summarize** to display results in tabular format.

Reviewing and Validating Results Interactively

After you have completed a workflow of searches and autovalidations you can manually check to see if lower quality results are, in fact, valid.

To display results in preparation for manual validation

- 1 Navigate to the Protein/Peptide Summary page. If this page is already open, click the **Summary Settings** button to display the parameter settings at the bottom of the page.
- 2 Make sure **Mode** is set to **Peptide**.
- 3 If your **Data Directories** are not selected, click the **Select...** button to select the folder or folders that contain your data files.
- 4 Set parameters for **Validation and Sorting** as shown in [Figure 42](#). These settings locate results that have not yet been validated. Note that only the settings that are highlighted in yellow have been changed from defaults. See the following steps for more details.

The screenshot displays the 'Summarize Results for Review' interface with three main sections:

- Summarize Results for Review:** Includes buttons for 'Summarize', 'Save As...', and 'Load...'. It has checkboxes for 'Queue request', 'Excel', and 'AMRT export'. The 'Mode' is set to 'Peptide'. There is a 'Report only distinct peptides' checkbox and a 'Data directories' section with a 'Select...' button. A list of search result files is shown, including 'ExampleData/Agilent/Q-TOF/yeast/y1'.
- Validation and Sorting:** The 'Filter results by:' dropdown is set to 'sequence-not-validated'. The 'Validation preset' is 'reset'. The 'Sort peptides by:' dropdown is 'Score'. The 'Filter peptides by:' section shows 'Score: > 5' and '% SPI: > 60' highlighted in yellow. Other options include 'Required AAs', 'Disallowed AAs', and 'Accession #s'.
- Review Fields:** A grid of checkboxes for various fields. Checked fields include: 'Filename', 'Score', 'FDR (Discriminant)', 'Fwd-Rev score', 'Rank 1-2 score', 'SPI (%)', 'Unmatched ions', 'VML score', 'Solution charge', 'Start AA position', 'Sequence', 'Rev Sequence', 'Var mod sequence', 'Retention time, width', 'Precursor m/z', 'Delta mass', 'Peptide pl', 'Protein MW', 'Species', 'Accession #', 'Protein name', 'Intensity', 'DEQ ratios', 'Reporter Ratios', 'Modification names', 'N-term', 'C-term', 'Cysteines', 'Fragmentation mode', 'Max tag length', 'Longest tag', 'Backbone Cleavages', 'b/y pairs', and 'Category'.

Figure 42 Settings to prepare for a manual validation

- 5 From the **Filter results by** list, select **sequence-not-validated**.

- 6 Use the following guidelines for setting the **Filter peptides by** parameters:

Q-TOF:

Peptide score	Quality	Peptide fragmentation	Likelihood of valid interpretation
> 15	Outstanding, well-assigned	Thorough	When combined with % SPI of 60 or greater, very likely to be valid
> 9	Good	Substantial	When combined with % SPI of 60 or greater, likely to be valid
> 5	Mixed Quality	Moderate	Review results to determine whether interpretation is valid; more likely to be valid than same scores for ion trap data
< 5	Generally poor	Less	Not likely to be valid

Because the mass accuracy is better than with ion trap data, you search with a narrower mass tolerance, so there is a better chance that lower-scoring results are valid.

Ion Trap:

Peptide score	Quality	Peptide fragmentation	Likelihood of valid interpretation
> 15	Outstanding, well-assigned	Thorough	When combined with % SPI of 70 or greater, very likely to be valid
> 10	Good	Substantial	When combined with % SPI of 70 or greater, likely to be valid
> 5	Mediocre	Few ions, look for mostly b's or y's	Review results to determine whether interpretation is valid; less likely to be valid than same scores for Q-TOF data
< 5	Poor	Little to none	Not likely to be valid

- If you have Agilent Q-TOF data, set the **Score** to 5.

Because the mass accuracy is better than with ion trap data, you search with a narrower mass tolerance, so there is a better chance that lower-scoring results are valid.

- If you have ion trap data, for the **Score** settings, see the guidelines above.
- **% SPI** (Scored Peak Intensity) is the percentage of the MS/MS peak-detected spectral ion current explained by the search interpretation. For Agilent Q-TOF spectra, **% SPI** below 60 may represent poor interpretations; so do not set lower values. For ion trap spectra, **% SPI** below 70 (or 60 for doubly-charged precursor ions) are more likely to represent poor interpretations..
- If you want to limit your data review, set the **Filter peptides by** parameters to higher values to filter out the poorer-quality results.

7 Set the **Validation preset** in a way that minimizes manual changes later:

- If you have Agilent Q-TOF data and you set **Filter peptides by** to a score of 7 or greater, select **valid**.
- If you have Agilent Q-TOF data and you set **Filter peptides by** to a value less than 7, select **reset**.
- If you have ion trap data and you set **Filter peptides by** to a score of 9 or greater, select **valid**.
- If you have ion trap data and you set **Filter peptides by** to a value less than 9, select **reset**.

8 In the **Review Fields**:

- a** Mark the check box for **Rank 1-2 score**. Verify that the check box for **Fwd-Rev score** is also marked. These parameters help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification.
- b** Mark the check box for **# Backbone Cleavages**, which is the number of cleavages of the amino acid backbone that are represented in the spectrum for the assigned sequence. For CID spectra, these cleavages include the b- and y-fragments, and for ETD spectra, they include the c- and z-fragments. Other ions, such as neutral losses from b and y ions, are not included.
- c** If you used **Dynamic peak thresholding** in the search, mark the check box for **Unmatched ions**. These results tell you both the

number of unmatched ions and the number of ions that were included in the search score that you see in the report. If the latter number is less than 15, check that the peptide is either short or in low abundance. If not, be suspicious of the results.

NOTE

In the **Review Fields**, if you mark the check box for **Peptide pI**, then under **Validation and Sorting** you have the option to display only peptides that meet your pI criteria. This feature is useful if you need to validate peptides from OFFGEL fractionation.

If you wish to use the pI filter for modified peptides, ask your server administrator to first verify that the pK of the modified amino acid is specified in **smconfig.std.xml** or **smconfig.custom.xml**.

- 9 Click **Save As** to save settings to the original or a different parameter file, then click **Save**.

You can now use this parameter file as a “task” in the list of tasks for an automated workflow. See Chapter 3, “Automating Workflows”.

- 10 If you want to see results immediately, make sure the **Queue Request** check box is clear. If you want to queue the task as part of an *interactive automation*, mark the check box.

- 11 Click **Summarize**.

If the Queue Request check box is clear, you quickly see a protein/peptide summary table that lists the hits and spectra that have been validated. Queuing a summary is useful when you want the results saved to an html file that can be accessed via the Completion Log.

NOTE

Once you click the **Summarize** button, the button is disabled until the results appear. If you need to re-enable the button, click the **Summary Settings** button to reload the Protein/Peptide Summary page.

- 12 After this task is complete, make sure that you see a display similar to that shown in [Figure 43](#) on page 93. Note that in this example, where the score threshold was set to only 6, results have been categorized as **R** for **reset**. These results require manual review to avoid the risk of false positives.

- 13 Manually review the results, as described in the next section.

To manually review and validate results

- 1 Depending on how you set validation and filtering parameters, you will want to manually review some or all of the results. Start by examining the overall results, as shown in [Figure 43](#).

Agilent Spectrum Mill - Protein/Peptide Summary																	
Spectrum Mill		Summary Settings		Autovalidation		Easy MS/MS		MS/MS Search		Spectrum Summary		Build TIC		Tool Belt		Help	
Perform Validation		Results Shown Filtered by Validation Category: sequence-not-validated															
Data Directory: msdata\SMExampleData\Agilent\Trap\ecoli\heatshock																	
hit table read - SpecFeatures read																	
all hits read from tagSummary file																	
original hits subset determined Files: 1051 Hits: 567																	
Validation category	#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	# Backbone Cleavages	Un-matched Ions	Spectrum Intensity	Sequence	MH-Matched (Da)	Accession #	Protein Name			
☐ V ☒ R	1	HEC_0040.2479.2479.0	3	15.66	15.66	15.66	68.3	12	9 / 25	6.26e+006	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	2	HEC_0040.2269.2269.0	3	14.99	14.99	14.99	66.7	12	11 / 25	9.32e+006	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	3	HEC_0040.1888.1903.0	3	14.98	10.67	14.98	69.0	13	8 / 24	2.66e+007	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	4	HEC_0040.2389.2389.0	3	13.59	13.59	13.59	68.7	13	9 / 25	7.95e+006	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	5	HEC_0020.2484.2484.0	3	13.51	13.51	13.51	67.6	11	9 / 25	1.09e+007	(K) ANDAAGDCITTTATVLAQAIIIEGLK(A)	2402.241	15804735	chaperonin GroEL			
☐ V ☒ R	6	HEC_0040.2464.2464.0	3	12.80	12.80	12.80	68.3	10	10 / 22	7.44e+006	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	7	HEC_0100.1473.1473.0	3	12.71	5.10	4.17	87.0	8	8 / 21	4.41e+006	(K) EQGLTPVLCIGETRAENEGKTEVCAK(Q)	3090.435	15804508	triosephosphate isomerase			
☐ V ☒ R	8	HEC_0060.1497.1497.0	3	12.61	12.61	8.85	61.6	7	12 / 25	6.37e+006	(R) DIALITPTNLPADVFVPGHENDNVFVSR(W)	3019.522	15800756	seryl-tRNA synthetase			
☐ V ☒ R	9	HEC_0040.2150.2161.0	3	12.59	12.59	12.59	65.9	9	11 / 25	1.86e+007	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	10	HEC_0040.2453.2453.0	3	12.31	12.31	12.31	60.7	10	11 / 25	4.55e+006	(K) TQLPSCGSELSLYDIAPVTGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase			
☐ V ☒ R	11	HEC_0010.1660.1672.0	3	12.30	12.30	8.54	67.3	10	11 / 25	1.34e+007	(K) DIQLATPPQVGAPATFYAALAEIK(A)	2467.308	16130805	glycine decarboxylase, PLP-			

Figure 43 Peptide mode display

- 2 To review spectra, click links as described on the next few pages. As you encounter valid interpretations, select the **V** under the **Validation category** heading to change the validation state to **valid**. For guidelines on how to decide whether a database match is reasonable, see [“To use the Spectrum Viewer”](#) on page 96 and [Table 4](#), “Guidelines for manual results validation,” on page 101.
- 3 When you are satisfied that the validation states are set correctly, click the **Perform Validation** button. This saves your validation states for further reference.
- 4 Note that you quickly see lists of validated hits and spectra. These are cumulative and include both the new hits and spectra you just validated, as well as those you validated previously via autovalidation or manual validation.

NOTE

If you find you have designated valid results in error, and you have already clicked the **Perform Validation** button, you can reverse the action. Change the validation state(s) and click the **Perform Validation** button again.

To compare theoretical and experimental spectra

Your primary review tool is the Spectrum Viewer. In many cases, this gives you all the information you need to evaluate database search results. To display the Spectrum Viewer (shown in [Figure 44](#)), click a link under the # or **SPI** headers.

- If you click a link under the # header, the spectrum is displayed at the bottom of the page.
- If you click a link under the **SPI** header, the spectrum is displayed in a new window.
- For details on use of the Spectrum Viewer, see [“To use the Spectrum Viewer”](#) on page 96.

The Spectrum Viewer shows your extracted spectrum annotated with theoretical fragments from the top database search result.

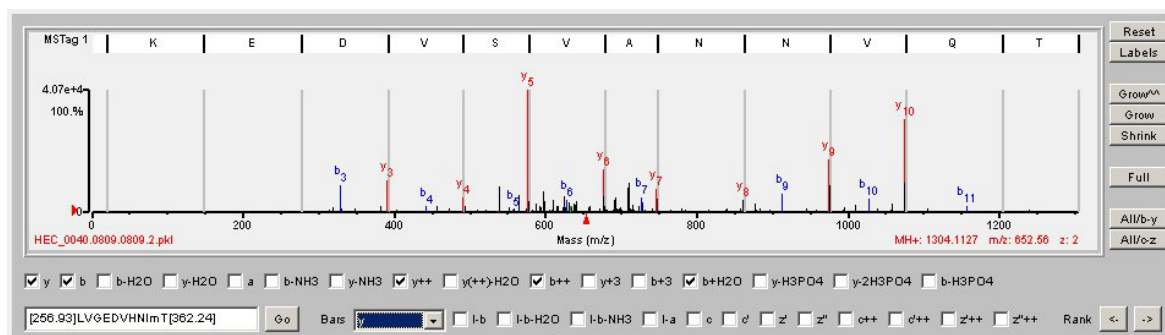


Figure 44 Spectrum Viewer, with y-ions highlighted using the **Bars** feature

To view MS/MS Search results

If you want more detailed information than is presented in the Spectrum Viewer, you view the MS/MS search details, as shown in [Figure 45](#).

- 1 To view these details, click a link under the **Filename** header shown in [Figure 43](#). Note that you may need to use your mouse to move the white page divider to see more than the spectrum.
- 2 Examine the MS/MS Search results, where you see up to five database hits with their scores, as well as the species for each protein. (While five is the default, you can increase the number of displayed hits with the **Max reported hits** setting on the MS/MS Search page.)
- 3 Scroll through the **Detailed Results** to see how the ions were assigned for each of the database hits.
 - The final **Score** from the database search is the sum of the scores for each of the ions. (The unassigned ions subtract from the score.)
 - The % **SPI** is the sum of the fractional intensities — **Frac.Inten. (% of TIC)** — for each of the assigned fragment ions.
 - For phosphorylated peptides, the **Rank 2**, **Rank 3**, etc. hits may show other forms of the phosphopeptide where the modification is on other sites. Examine the ions that constitute the evidence for those locations to judge which is the best interpretation. If there is ambiguity in the location of the modification, the various hits have similar scores.

NOTE

If you enable discriminant scoring, you also see the Discriminant Score in the Detailed Results.

2 Reviewing and Validating MS/MS Data

Reviewing and Validating Results Interactively

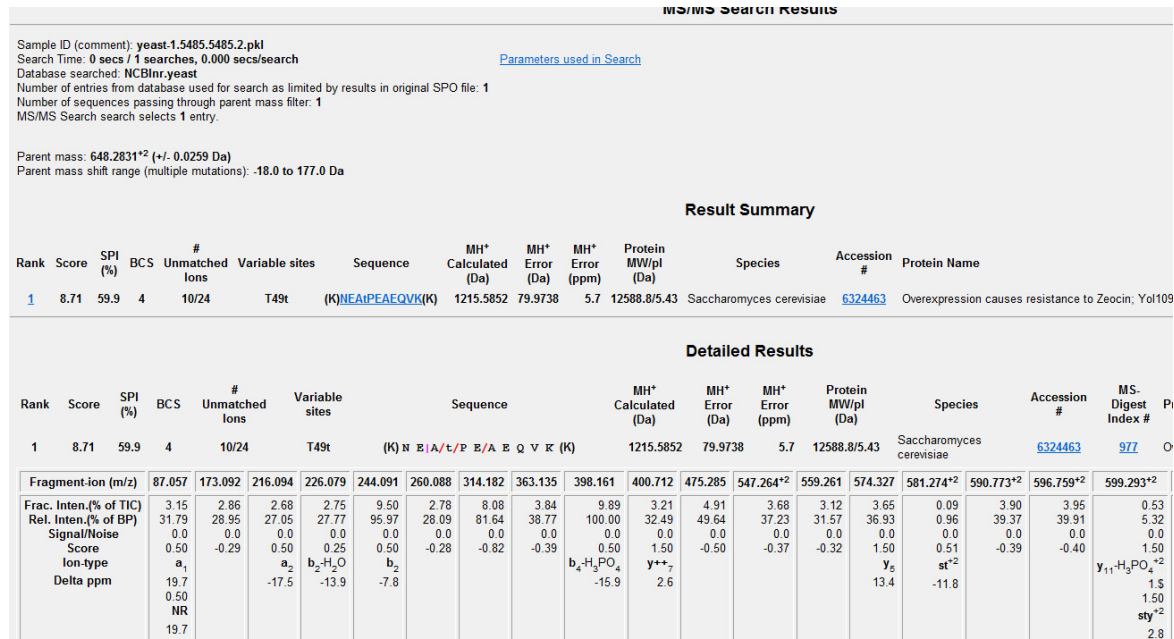


Figure 45 MS/MS Search results

4 Click additional links shown in the MS/MS Search results in [Figure 45](#):

- To view an MS Product listing that shows all theoretical fragment ions, click an amino acid sequence.
- To view information in the database you searched, click an accession number in the MS/MS Search report.

To view MS Digest results that show peptide masses from a theoretical digestion, click an MS Digest index number in the MS/MS Search report.

To use the Spectrum Viewer

[Figure 46](#) shows the Spectrum Viewer, a tool to visualize the sequence information contained in MS/MS spectra and to evaluate spectral interpretations from MS/MS database searches or Sherenga *de novo* sequencing. To manipulate the Spectrum Viewer, see the descriptions below.

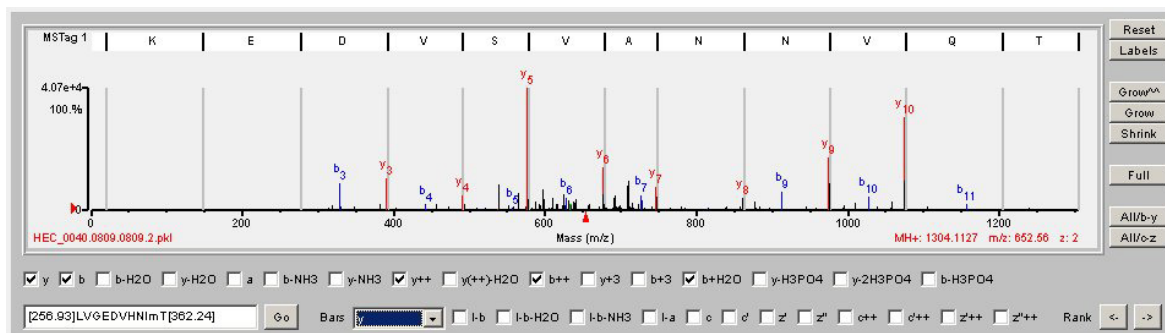


Figure 46 Spectrum Viewer, with y-ions highlighted using the **Bars** feature

Color-coding

In the Spectrum Viewer, the software color-codes the ions as follows:

- Red: C-terminal fragments (y-ions and z-ions)
- Blue: N-terminal fragments (b-ions and c-ions)
- Green: Precursor ions, neutral losses from precursor ions, immonium ions

To use the buttons

- Use the buttons on the right side of the Spectrum Viewer to manipulate the display. The buttons do the following:

Reset Resets the spectrum to the original x- and y-axis values.

Labels Toggles among different peak labeling options. These are the default of interpreted peaks (b, y, etc.), interpreted peaks plus mass labels of all peaks, interpreted peaks plus mass labels of interpreted peaks, and no labels.

To see the spectrum without the peak interpretations, click the red file name under the sequence at the top of the Spectrum Viewer. Now click the **Labels** button to turn the mass labels on and off. Note that you need to re-display the spectrum to see the interpretations again.

Grow^^ Expands the spectrum by ten times in the vertical axis

Grow Expands the spectrum in the vertical axis

Shrink Reduces the spectrum in the vertical axis

- Full** Displays the spectrum with the full x-axis values. If you see a black arrow at the lower right-hand corner of the spectrum, this means that some of the spectrum is not displayed because there were no significant peaks in that region. Click the **Full** button to display the full x-axis range of the spectrum.
- All/b-y** Toggles marking of check boxes for b- and y-ions in the first row under the spectrum and labels peaks accordingly. The toggle either marks all the check boxes or resets to defaults.
- All/c-z** Toggles marking of check boxes for c- and z-ions in the second row under the spectrum and labels peaks accordingly. The toggle either marks all the check boxes or resets to a default subset. In addition, the **All/c-z** button labels precursor ions that have reduced charges, as are typically observed in ETD spectra. When you analyze ETD data, you must click the **All/c-z** button to see the appropriate labels for the fragment ions.

To use the features above the spectrum

- Spectra are annotated with interpretations based on the sequence shown in white. When both **MSTag** and **Sherenga** sequence bars are displayed, click the sequence for which you wish to see annotations.
- To see the spectrum without the peak interpretations, click the red file name under the sequence at the top of the Spectrum Viewer. Now click the **Labels** button to turn the mass labels on and off. Note that you need to re-display the spectrum to display the interpretations again.

To use the check boxes in the first row below the spectrum

- Mark check boxes for the ion types you want labeled in the spectrum. If you mark check boxes but do not see any labels, click the **Labels** button to turn the labels back on.

To use the features in the second row below the spectrum

- Use these features to manipulate the display. These features do the following:
- Go** To display a sequence above the spectrum, type the sequence in the box to the left and then click the **Go** button.

The sequence in the box is initially set to a default sequence. Mass gaps shown in brackets indicate portions of the spectrum where there was insufficient fragmentation to provide an amino acid sequence. You can enter mass gaps in the middle of the sequence as well as at the ends.

Note that in addition to single-letter capitalized abbreviations for the 20 amino acids, you can type the following lower-case abbreviations for modified amino acids:

Symbol	Modified amino acid
k	Carbamylated lysine
m	Oxidized methionine
q	Pyroglutamic acid (only at N-terminus of peptide)
s	Phosphorylated serine
t	Phosphorylated threonine
y	Phosphorylated tyrosine

The variable modifications **kmqsty** are defined by default for the Spectrum Viewer. But if in MS/MS Search, you specified a different variable modification for K, M, Q, S, T, or Y (for example, guanidination of K), then that modification is used instead.

Use the **Rank** arrow buttons (<- and ->) to go from the sequences that were identified by MS/MS Search or Sherenga *de novo* Sequencing to the sequence that you typed. For MS/MS Search, the arrow buttons cycle between the peptide from the highest-scoring MS/MS search result and the sequence you typed. For Sherenga, the arrow buttons cycle through all the Sherenga result sequences. If you add a custom sequence, the software appends it to the list of Sherenga sequences that can be cycled.

Bars Select an ion type to highlight. As shown in [Figure 46](#), this allows you to more easily visualize how the ions align with the amino acid sequence displayed at the top of the Spectrum Viewer.

Check boxes labeled I-x Mark these if you want to label peaks with amino acid sequences combined with common losses.

2 Reviewing and Validating MS/MS Data

Reviewing and Validating Results Interactively

Check boxes for c- and z-ions Mark check boxes for the ion types you want to see labeled in the spectrum. If you mark check boxes but do not see any labels, click the **Labels** button to turn the labels back on.

Rank Click arrows to display additional Sherenga results. Note that when both MS-Tag and Sherenga sequence bars are displayed in the Spectrum Viewer, you must first click the Sherenga sequence bar and make sure it turns from gray to white before you click the **Rank** arrows.

To expand the x-axis

- Use the cursor to expand a portion of the spectrum in the x-axis. Move your mouse over the spectrum. When a crosshair is displayed, select the portion of the spectrum you wish to expand.
- Double-click the spectrum or click the **Reset** button to return to the original display.

To change the threshold for mass labeling

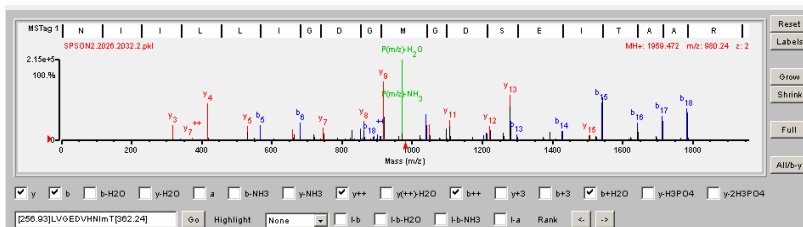
- The red triangle to the left of the y-axis indicates the threshold for peak labeling. Click anywhere on the y-axis to change this threshold.

Table 4 Guidelines for manual results validation

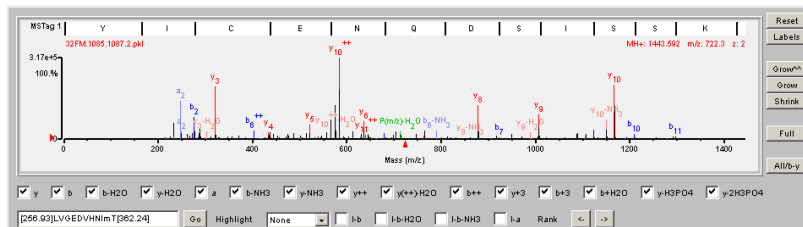
Guideline

Example

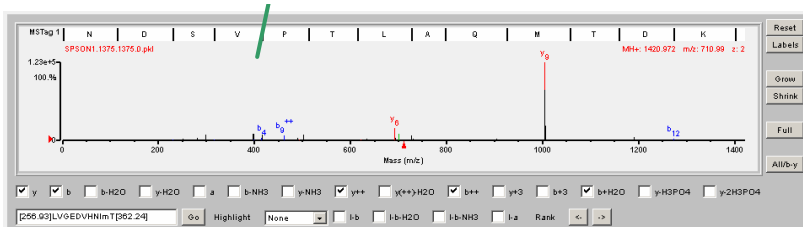
The best spectra are generally found by autovalidation, so do not expect to see spectra like the one to the right.



- Loss of NH_3 occurs from R, K, Q and N residues.
- Loss of H_2O occurs from S, T, E and D.
- R, H, K and N are charge-bearing residues and increase the maximum charge state allowed for a peptide fragment.
- Enhanced fragmentation may be observed at the following bonds:
 - His – Xaa
 - Xaa – Gly
 - Xaa – Ser



The peptide bond on the N-terminal side of Pro is particularly labile (high local proton affinity) and usually results in an MS/MS spectrum dominated by the y-ion ending in Pro.*



It is typical to lack fragmentation between the first two N-terminal amino acids (i.e., no b_1 or y_{x-1} ions).

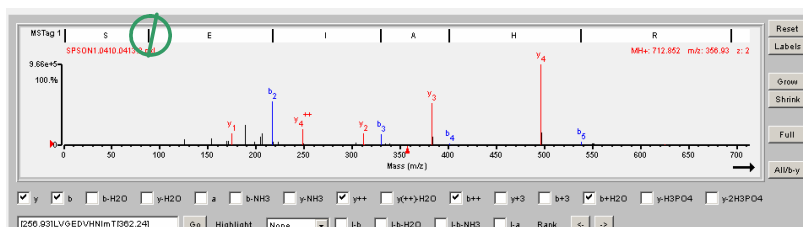


Table 4 Guidelines for manual results validation

Guideline

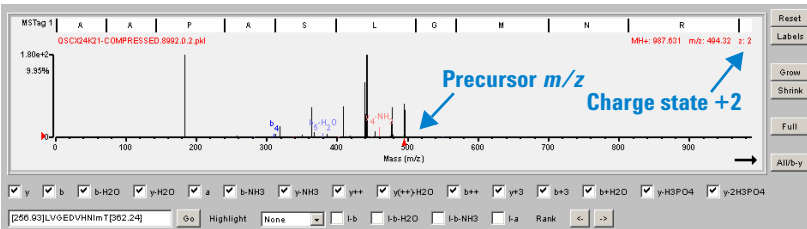
Look for coverage over a significant portion of the backbone. A series of consecutive b or y fragments is better than a random distribution.

Example

The peptide bond on the C-terminal side of acidic residues (D, E) may show enhanced cleavage.[†]

Be suspicious of a clumped pattern that occurs frequently throughout the LC run (usually a background ion).

Table 4 Guidelines for manual results validation

Guideline	Example
Be suspicious if no ions appear above the precursor ion (when precursor charge state is >1).	 MS/MS 1 Q5C20421-COMPRESSIONED.0992.0.2.pH 1.80e+2 9.95% Mass (m/z) Precursor m/z Charge state +2 Reset Labels Grow Shrink Full All/b-y ☒ y ☒ b ☒ b-H2O ☒ y-H2O ☒ a ☒ b-NH3 ☒ y-NH3 ☒ y++ ☒ y(++)H2O ☒ b++ ☒ y+3 ☒ b+3 ☒ b+H2O ☒ y-H3PO4 ☒ y-2H3PO4 [206.93]LV9EDVHNimT[62.24] Highlight None ☐ i-b ☐ i-b-H2O ☐ i-b-NH3 ☐ i-a Rank < >

Remember - When in doubt, *don't* validate.

- The goal is *no* false positives.
- Validating a questionable hit removes it from subsequent searches.

* Proton mobility scoring accounts for enhanced fragmentation around P.

† Proton mobility scoring accounts for enhanced fragmentation around D and E.

Protein/Peptide Summary Displays

The data display modes on the Protein/Peptide Summary page allow you to view data in a number of ways, so you can select the mode that best fits your experiment. The list of display modes is shown in [Figure 47](#). This section shows examples of data displayed in each mode. The modes are discussed in the order that they appear in the drop-down list.

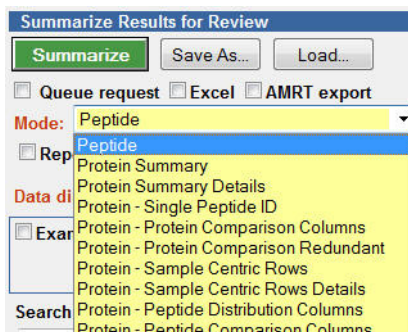


Figure 47 Data display modes

Peptide

Select this mode to display results by peptide. This mode is particularly useful to validate the database identifications from each MS/MS spectrum. The output looks like that in [Figure 48](#).

Note that file names are in the format Data_File_Name.aaaa.bbbb.c, where

- aaaa = first merged scan
- bbbb = last merged scan
- c = assigned precursor charge (0 means charge was ambiguous)

If you have QSTAR files, see the note on [page 30](#).

#	Filename	z	Score	SPI (%)	Spectrum Intensity	Sequence	MH ⁺ Matched (Da)	Accession #	Protein Name
1	yeast-1.17757.17757.2	2	24.17	91.1	1.10e+005	(K) EYFQDVYDLTGSEPIK (E)	1903.896	4243	FAS1
2	yeast-1.19496.19511.2	2	23.72	100.0	7.86e+005	(K) TVIGEVLLQAYGGMR (G)	1735.905	6324328	citrate synthase, Nu
3	yeast-1.20693.20693.2	2	23.44	98.4	1.54e+004	(K) ETNPGTDVTVSSVESVLAHL (-)	2055.024	6323138	thioredoxin peroxida
4	yeast-1.20456.20456.2	2	23.41	95.3	9.17e+005	(K) TIAETLAEELINAAK (G)	1586.864	6322583	ribosomal small sub
5	yeast-1.20290.20290.3	3	23.27	97.7	1.18e+005	(K) AVGDNDPIDVLEIGETIAYTGQVK (Q)	2517.272	124875	Inorganic pyrophosp
6	yeast-1.16410.16787.2	2	22.47	100.0	1.52e+007	(K) VNLDTDCQYAYLTGIR (D)	1901.906	6322790	Low Temperature-re
7	yeast-1.18085.18085.2	2	22.46	90.8	5.54e+004	(K) DIFSNDLLSDAYDAK (L)	1815.828	6322794	Hypothetical ORF; \
8	yeast-1.6725.6725.3	3	22.34	82.2	2.33e+004	(K) SSYGSNNDDSYGSNNDDSYGSSNK (K)	2533.946	6323826	DNA damage induci
9	yeast-1.17489.17882.2	2	22.07	97.7	4.59e+007	(R) VPTVDVSVVDLTVK (L)	1470.841	6322468	glyceraldehyde 3-ph
10	yeast-1.20082.20082.2	2	22.03	96.4	3.08e+005	(K) MAIFDELLQMPETK (L)	1665.823	6322565	adenosine kinase; A
11	yeast-1.9857.9857.2	2	21.81	97.5	4.81e+005	(K) IESSVESQFSAGR (L)	1396.670	6319399	Homology to mamr
12	yeast-1.17526.17526.2	2	21.70	89.0	3.40e+005	(R) EETYYQESAGVADLIITCAGGR (N)	2391.077	6320181	glycerol-3-phosphat
13	yeast-1.19253.19253.2	2	21.64	94.1	1.50e+005	(R) AEVSDVGNAILDGADCVMLSGETAK (G)	2522.175	6319279	Required for START

Figure 48 Peptide display(partial)

Protein Summary

Select this mode to display results organized by protein. This mode is useful when your sample fractions are organized in a single data directory, and you want the peptides from a single protein organized together even if they originate from different data files.

If you want to include a column for protein intensities, mark the **Intensity** check box in the **Review Fields** prior to clicking the **Summarize** button. The output looks like that in [Figure 49](#).

The color code in the **Distinct Peptides** column indicates the relative number of peptides detected for each protein. You can see from the corresponding numbers that darker colors represent larger numbers of peptides.

The color code in the **Mean Peptide Spectral Intensity** column indicates relative protein concentrations. Again, darker colors represent higher concentrations. The number in each cell is the mean peptide spectral intensity, which is an average of the intensities for all the peptides assigned to that protein. The peptide intensities are calculated from extracted ion chromatograms from the precursor ions.

These intensity results are sufficient for studies where you are interested in sample differences of two-fold or more.

Agilent Spectrum Mill - Protein/Peptide Summary								
Spectrum Mill	Summary Settings	Autovalidation	Easy MS/MS	MS/MS Search	Spectrum Summary	Build TIC	Tool B	
Results Shown Filtered by Validation Category: valid Data Directory: msdataSM\ExampleData\ICAT hit table read - SpecFeatures read valid hits read from tagSummary file - Files: 36 Hits: 35 beginning to assemble proteins proteins assembled 0.00445 sec proteins filtered by unique peptides 0.003495 sec proteins filtered by score calculated protein coverage maps 0.047986 sec beginning to roll up proteins into groups proteins rolled up into groups 0.002542 sec protein groups ready for display proteinGroupingMethod: oneSharedPeptide 6 Proteins listed								
Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Protein MW (Da)	Database Accession #	Protein Name
1	20	12	159.67	21	3.29e+007	77050.4	4557871	transferrin ▾
2	10	8	93.73	19	2.80e+007	69293.9	1351907	bovine serum albumin ▾
3	1	1	17.20	2	4.43e+007	53354.2	71826	fibrinogen beta chain ▾
4	2	1	16.34	3	4.52e+007	42881.5	129293	ovalbumin ▾
5	1	1	14.17	13	8.13e+006	10902.8	115114	Aprotinin ▾
6	1	1	11.20	1	1.00e+007	116483.5	114939	beta-galactosidase ▾
Totals:	35	24						

Figure 49 Protein Summary display

Protein Summary Details

Select this mode to see results organized by proteins, with all the supporting peptide information. This mode is useful when your sample fractions are organized in a single data directory, and you want the peptides from a single protein organized together even if they originate from different data files.

The top of the summary shows the proteins, as in [Figure 50](#), while the bottom of the summary shows the individual peptides for each protein, as in [Figure 51](#). This mode is very useful for manual review and validation of MS/MS search results. If your sample has multiple fractions, you sometimes see the same peptide sequences in multiple data files. You also see this if you chose not to merge spectra during extraction.

Group (#)	Subgroup (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Total Protein Spectral Intensity	Protein MW (Da)	Database Accession #	Protein Name
1	1.1	61	30	478.66	87.6	3.49e+008	35845.2	3720	reading frame
1	1.2	39	24	343.65	67.7	2.62e+008	35960.3	6322468	glyceraldehyde 3-ph
1	1.3	38	22	310.32	68.3	2.46e+008	35863.4	6322409	Glyceraldehyde-3-ph
2	2.1	58	33	473.79	80.5	3.75e+008	44794.6	10383781	3-phosphoglycerate
3	3.1	50	36	471.17	66.8	1.09e+008	69827.1	417149	HEAT SHOCK PROT
3	3.2	39	29	362.14	49.9	9.12e+007	69639.9	6323004	member of 70 kDa he
3	3.3	16	11	146.12	17.8	3.03e+007	70659.6	6319396	heat-inducible cytos
3	3.4	8	8	84.53	14.9	3.21e+006	74523.7	6322426	Involved in translocat
4	4.1	53	27	423.29	71.3	2.76e+008	46970.4	6321968	enolase: Eno2p
4	4.2	49	27	415.87	71.8	3.25e+008	46872.3	6321693	enolase I: Eno1p
5	5.1	38	30	399.62	68.8	8.84e+007	54943.0	6319279	Required for START
6	6.1	50	27	379.42	59.8	7.05e+007	61722.5	6323073	pyruvate decarboxyl:
6	6.2	14	10	120.52	20.4	2.43e+007	62139.1	6323163	pyruvate decarboxyl:
6	6.3	9	6	83.70	14	1.50e+007	61636.3	6321524	Third, minor isozyme
7	7.1	31	27	360.47	44.1	2.25e+007	86029.5	6320936	cobalamin-independ
8	8.1	39	26	356.58	63.3	3.71e+007	57007.9	6324950	Glucose repressed. I
9	9.1	35	26	343.66	52.6	4.69e+007	66707.5	6324120	stress-seventy subfa
9	9.2	35	26	338.02	51.3	4.73e+007	66771.5	6319972	Stress-Seventy Subf
10	10.1	33	27	343.27	43.2	3.43e+007	93743.9	6320593	translation elongator

Figure 50 Protein Summary Details display - top

Group #	Subgroup #	Spectra #	Distinct Peptides #	Distinct Summed MS/MS Search Score	% AA Coverage	Total Protein Spectral Intensity	Database Accession #	Protein Name
6	6.1	47	26	358.27	57.9	5.89e+007	6323073	pyruvate decarboxylase; Pdc1p

#	Filename	z	Score	Discriminant	Local FDR (%)	Global FDR (%)	FDR Search #	SPI (%)	Spectrum Intensity	Variable Sites	Sequence	Modifications Map
1	yeast-1.13773.13773.2	2	21.07	23.4401	<0.1%	<0.1%	1	85.5	8.43e+005		(R) WAGNANELNAAYAADGYAR (I)	
2	yeast-1.20640.20640.3	3	20.80	21.3927	<0.1%	<0.1%	1	97.7	3.00e+005		(R) MIEIMLPVFDAPQNLVEQAK (L)	
3	yeast-1.19425.19521.3	3	20.62	20.3754	<0.1%	<0.1%	1	93.6	2.32e+006		(K) AQYNEIQGDHLSLLPTFGAK (D)	
4	yeast-1.20437.20437.3	3	19.75	14.9302	<0.1%	<0.1%	1	96.2	2.16e+004		(K) EVIDTILALVK (D)	
5	yeast-1.20148.20148.3	3	19.23	19.4799	<0.1%	<0.1%	1	92.6	1.21e+006		(K) LIDLTFQFAFVTIMGK (G)	
6	yeast-1.20426.20426.2	2	19.10	15.2844	<0.1%	<0.1%	1	91.1	1.95e+005		(K) EVIDTILALVK (D)	
7	yeast-1.19203.19342.3	3	18.07	19.5553	<0.1%	<0.1%	1	94.1	4.55e+006		(R) ITTYVTQRFVYLGLEPANLVDLNVPAR (L)	
8	yeast-1.19655.19655.3	3	18.04	19.0360	<0.1%	<0.1%	2	94.5	2.76e+005	M247m	(K) LIDLTFQFAFVTIMGK (G)	m:Oxidized methionine
9	yeast-1.20635.20635.2	2	16.94	14.7662	<0.1%	<0.1%	1	81.0	5.16e+005		(R) MIEIMLPVFDAPQNLVEQAK (L)	
10	yeast-1.11568.11568.3	3	16.09	16.5193	<0.1%	<0.1%	1	80.5	6.96e+006		(R) YGVVYVGTLSKPEVK (E)	
11	yeast-1.19886.19886.2	2	15.72	19.8115	<0.1%	<0.1%	1	79.3	7.20e+005		(R) MSANISSETTAMITDIATAPAEIDR (C)	
12	yeast-1.19513.19513.3	3	15.17	17.1574	<0.1%	<0.1%	1	100.0	7.22e+005		(K) KLIDLTFQFAFVTIMGK (G)	
13	yeast-1.10516.10516.3	3	15.05	15.2542	<0.1%	<0.1%	1	77.8	1.47e+006		(K) NPVILADACCSR (H)	C:Carbamidomethylation
14	yeast-1.2691.2806.2	2	14.92	12.8055	<0.1%	<0.1%	1	88.9	1.62e+006		(K) SFNDNSK (I)	
15	yeast-1.19906.19906.3	3	14.64	16.5382	<0.1%	<0.1%	1	71.2	6.86e+005		(R) MSANISSETTAMITDIATAPAEIDR (C)	
16	yeast-1.19209.19396.4	4	14.51	16.1476	<0.1%	<0.1%	1	96.1	1.05e+006		(R) ITTYVTQRFVYLGLEPANLVDLNVPAR (L)	
17	yeast-1.9669.9734.2	2	14.42	13.9073	<0.1%	<0.1%	1	92.2	5.02e+006		(K) IYEVGNR (W)	
18	yeast-1.21057.21057.2	2	14.26	16.2047	<0.1%	<0.1%	2	92.6	4.14e+004	Q16q	(K) qVNVNIVFGLPGDFNLSLLDK (I)	q:Pyroglutamic acid
19	yeast-1.9482.9482.2	2	14.04	16.4979	<0.1%	<0.1%	1	85.8	4.65e+004		(R) TPANAAVFASFTPLK (Q)	
20	yeast-1.19631.19631.2	2	13.52	15.4284	<0.1%	<0.1%	2	77.1	1.26e+005	M247m	(K) LIDLTFQFAFVTIMGK (G)	m:Oxidized methionine

Figure 51 Protein Summary Details display - bottom

To view sequence coverage, click a link under the % AA Coverage heading.

Protein-Single Peptide ID

Select this mode to display proteins that have been identified via a single peptide. This mode enables review of the annotated MS/MS spectra that led to these identifications. At least one journal (*Molecular & Cellular Proteomics*) requires that these annotated spectra be submitted with the manuscript.

An example Protein-Single Peptide ID report is shown in Figure 52. Each row corresponds to the highest scoring instance of the single peptide used to identify the protein, with a link to the corresponding spectrum. The fourth column shows how many spectra correspond to that peptide. (The other spectra arise from additional charge states, same peptide from one of the other raw files in the directory, etc.)

Group #	Distinct Summed MS/MS Search Score	Num Unique Peps	Num Spectra	#	Filename	z	Score	SPI (%)	Spectrum Intensity	Sequence	MH ⁺ Matched (Da)	Accession #	Protein Name
1	23.44	1	1	1	yeast-1.20693.20693.2	2	23.44	98.4	1.54e+004	(K) ETNPGIDVTVSSVESVLAHL (-)	2055.024	6323138	thioredoxin peroxidase; Ahp1p ▾
2	22.34	1	1	1	yeast-1.6725.6725.3	3	22.34	82.2	2.33e+004	(K) SSYGSNNDDSYGSNNDDSYGSNNK (K)	2533.946	6323826	DNA damage inducible; implicated in the p
3	22.03	1	1	1	yeast-1.20082.20082.2	2	22.03	96.4	3.08e+005	(K) MAIFDELLQMPETK (L)	1665.823	6322565	adenosine kinase; Ado1p ▾
4	20.52	1	1	1	yeast-1.14343.14406.2	2	20.52	90.5	2.28e+006	(R) LTADYDALDIANR (I)	1450.717	320176	ribosomal protein L23a.e
5	20.38	1	1	1	yeast-1.18109.18109.2	2	20.38	93.4	5.04e+005	(K) EFDEAYDLFEVQR (V)	1660.749	6321842	subunit VI of cytochrome c oxidase; Cox6p
6	20.21	1	1	1	yeast-1.20704.20704.3	3	20.21	90.0	6.63e+004	(K) NAPAIIFIDEIDSIAPK (R)	1826.990	6320077	Microsomal protein of CDC48/PAS1/SEC1
7	20.19	1	1	1	yeast-1.17335.17335.2	2	20.19	85.9	1.70e+006	(R) SGVAVADESLTAFNDLK (L)	1736.870	6322978	yeast cortical cytoskeleton component; ms
8	20.13	1	1	1	yeast-1.20174.20174.2	2	20.13	91.4	3.82e+005	(K) IEAALS DALAALQIEDPSADELR (K)	2411.230	6319972	Stress-Seventy Subfamily B; involved in tr
9	20.02	1	1	1	yeast-1.19422.19515.2	2	20.02	97.8	2.30e+006	(R) SSVLADALNAINNAEK (T)	1629.844	6322271	Homology to rat S15a; Rps22ap ▾
10	19.60	1	1	1	yeast-1.9833.9833.3	3	19.60	91.3	5.57e+005	(K) HGENSELLSGTSQDSLRL (E)	1787.852	6321679	phosphofructokinase alpha subunit Pfk1p

Figure 52 Protein-Single Peptide ID display

Protein-Protein Comparison Columns

Select this mode to organize your results by proteins, especially if you want to compare proteins across multiple samples, as in a differential expression study. This mode is useful when your sample fractions are organized in a single data directory, and you want the peptides from a single protein organized together even if they originate from different data files.

If you have multiple fractions for each sample, and each sample has its own subdirectory on the Spectrum Mill server, then you can use the **Group results by** parameter to control whether the individual fractions or the overall sample results (with all fractions consolidated) are displayed.

[Figure 53](#) and [Figure 54](#) show examples of each type of display. [Figure 53](#) shows the distribution of proteins across a 2D LC/MS/MS run (individual fractions), while [Figure 54](#) combines the results of all salt fractions into a single sample.

For an example showing a comparison of two samples (directories), each of which contain multiple salt fractions, see [Figure 21](#) on page 56.

2 Reviewing and Validating MS/MS Data Protein/Peptide Summary Displays

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum MillSummary SettingsAutovalidationEasy MS/MSMS/MS SearchSpectrum SummaryBuild TICTool BeltHelp

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSMExampleDataAgilentTrap/ecoli/heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - Files: 376 Hits: 2557
beginning to assemble proteins proteins assembled 0.380508 sec
proteins filtered by unique peptides 0.149504 sec
proteins filtered by score
calculated protein coverage maps 0.448646 sec
beginning to roll up proteins into groups ... proteins rolled up into groups 0.187058 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide
Mean intensity: sum of intensity for all spectra of peptides belonging to protein / # spectra

HEC_0000 # spectra mean intensity	HEC_0010 # spectra mean intensity	HEC_0020 # spectra mean intensity	HEC_0040 # spectra mean intensity	HEC_0060 # spectra mean intensity	HEC_0080 # spectra mean intensity	HEC_0100 # spectra mean intensity	HEC_0150 # spectra mean intensity	HEC_0250 # spectra mean intensity	HEC_0500 # spectra mean intensity	HEC_1000 # spectra mean intensity	Database Accession #	%AA Coverage	Distinct Peptides (#)	Distin Summ MS/M Searc Score
0 0.00e+000	2 1.61e+006	31 2.05e+007	2 1.20e+007	43 1.27e+007	8 8.86e+006	5 3.14e+006	0 0.00e+000	1 3.10e+006	3 2.33e+006	3 2.06e+006	45686198	18	7	125
0 0.00e+000	0 0.00e+000	4 3.14e+007	1 4.87e+007	4 6.67e+006	3 3.75e+006	4 6.89e+006	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	15799694	19	7	104
0 0.00e+000	2 1.33e+007	9 4.23e+007	3 2.71e+007	1 1.96e+007	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	15803853	9	4	76
0 0.00e+000	3 4.72e+006	2 1.95e+007	0 0.00e+000	0 0.00e+000	2 1.02e+007	2 7.18e+006	1 4.59e+006	1 1.94e+006	0 0.00e+000	0 0.00e+000	15800756	9	3	49
0 0.00e+000	0 0.00e+000	4 5.42e+007	2 1.53e+007	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	15803363	15	3	49
0 0.00e+000	4 6.79e+006	6 5.75e+007	2 2.44e+007	1 6.36e+006	2 2.67e+006	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	15800320	21	3	46
0 0.00e+000	0 0.00e+000	6 1.34e+007	2 3.63e+007	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	15804515	8	3	46
0 0.00e+000	0 0.00e+000	2 5.63e+007	21 2.62e+007	2 1.73e+007	0 0.00e+000	0 0.00e+000	1 2.51e+006	0 0.00e+000	0 0.00e+000	0 0.00e+000	26249817	13	2	45
0 0.00e+000	0 0.00e+000	4 1.66e+007	1 2.22e+007	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	0 0.00e+000	15803459	10	2	34

Figure 53 Protein-Protein Comparison Columns display mode, **Group results by File**

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum MillSummary SettingsAutovalidationEasy MS/MSMS/MS SearchSpectrum SummaryBuild TICTool BeltHelp

Results Shown Filtered by Validation Category: valid

Data Directory: msdataSMExampleData\Agilent\Trap\ecoli\heatshock

hit table read - SpecFeatures read

valid hits read from tagSummary file - Files: 376 Hits: 2557

beginning to assemble proteins proteins assembled 0.377028 sec

proteins filtered by unique peptides 0.137336 sec

proteins filtered by score

calculated protein coverage maps 0.456765 sec

beginning to roll up proteins into groups proteins rolled up into groups 0.195292 sec

protein groups ready for display

proteinGroupingMethod: oneSharedPeptide

Mean intensity: sum of intensity for all spectra of peptides belonging to protein / # spectra

ExampleData\Agilent\Trap\ecoli\heatshock # spectra mean intensity	Database Accession #	%AA Coverage	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	Group #	Protein Name
98 1.34e+007	45686198	18	7	125.18	1.1	GroEL
16 1.50e+007	15799694	19	7	104.63	2.1	molecular chaperone DnaK
15 3.39e+007	15803853	9	4	76.96	3.1	elongation factor EF-2
11 8.59e+006	15800756	9	3	49.13	4.1	seryl-tRNA synthetase
6 4.12e+007	15803363	15	3	49.12	5.1	5-keto-4-deoxyuronate isomerase
13 3.12e+007	15800320	21	3	46.66	6.1	alkyl hydroperoxide reductase, C22 subunit; detoxification
8 1.91e+007	15804515	8	3	46.09	7.1	glycerol kinase
26 2.69e+007	26249817	13	2	45.59	8.1	malate dehydrogenase
5 1.77e+007	15803459	10	2	34.78	9.1	fructose-bisphosphate aldolase
21 3.15e+007	26249935	6	2	34.68	10.1	elongation factor Tu
4 1.65e+007	26246730	25	2	34.55	11.1	Hypothetical protein yfcZ

Figure 54 Protein-Protein Comparison Columns display mode, **Group results by Directory**

If you want the software to align and compare amino acid sequences of proteins within a protein group, use one of these modes:

- Protein-Protein Comparison Columns
- Protein-Protein Comparison Redundant (discussed on [page 114](#))

To view the aligned sequences, click a link for **Group #** or **Subgroup #** in the report. The software highlights the detected peptides within the sequences. For more information, see the online help about the multiple sequence alignment tool within Protein/Peptide Summary (in **Spectrum Mill Basics**).

Protein-Protein Comparison Redundant

This mode is equivalent to Protein-Protein Comparison Columns; however, it expands each protein group to show all the individual redundant protein entries that were consolidated in the former mode because of shared peptides. An example is shown in [Figure 55](#). The redundant hits share a common **Group #**. Redundant hits arise from duplicate database entries, homologous hits, and variants of the protein. You can pick out duplicate database entries because they exhibit the same score (**Distinct Summed MS/MS Search Score**) and amino acid coverage (**%AA Coverage**).

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum MillSummary SettingsAutovalidationEasy MS/MSMS/MS SearchSpectrum SummaryBuild TICTool Belt

Display Peptides by Protein SubGroup

Results Shown Filtered by Validation Category: valid

Data Directory: msdataSM\ExampleData\Agilent\Trap/ecoli/heatshock

hit table read - SpecFeatures read

valid hits read from tagSummary file - Files: 376 Hits: 2557

beginning to assemble proteins proteins assembled 0.374028 sec

proteins filtered by unique peptides 0.15597 sec

proteins filtered by score

calculated protein coverage maps 0.450265 sec

beginning to roll up proteins into groups proteins rolled up into groups 0.188672 sec

protein groups ready for display

proteinGroupingMethod: oneSharedPeptide

Mean intensity: sum of intensity for all spectra of peptides belonging to protein / # spectra

ExampleData\Agilent\Trap/ecoli/heatshock # spectra mean intensity	Database Accession #	%AA Coverage	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	Group #	Protein Name
98 1.34e+007	45686198	18	7	125.18	☐ 1.1	GroEL
98 1.34e+007	38491462	18	7	125.18	☐	GroEL
98 1.34e+007	38491468	18	7	125.18	☐	GroEL
98 1.34e+007	38491464	18	7	125.18	☐	GroEL
98 1.34e+007	38491478	18	7	125.18	☐	GroEL
98 1.34e+007	18028150	18	7	125.18	☐	GroEL
98 1.34e+007	18028158	20	7	125.18	☐	GroEL
97 1.34e+007	38491476	18	6	113.74	☐	GroEL
97 1.30e+007	37624680	22	6	104.87	☐	GroEL protein
97 1.30e+007	37624678	22	6	104.87	☐	GroEL protein

Figure 55 Protein-Protein Comparison Redundant mode, Group results by Directory

Protein-Sample Centric Rows

Select this mode to organize your results by data file. This mode is useful when your sample fractions (e.g., 2D gel spots) are organized into a single directory (e.g., spots belonging to the same gel), and you want to organize your results by fraction. Unlike the Protein-Protein Comparison Columns and Protein Summary Details modes, this mode requires the one-to-one correspondence of a protein to a particular data file. The same protein found in separate data files is listed once for each data file.

This display is useful to view all the proteins in each gel spot or band. The output looks like that in [Figure 56](#).

Run #	Run Name	Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Database Accession #	Protein Name
1	HEC_0000	1	1	1	17.90	2	2.44e+007	15804578	DNA-directed RNA polymerase beta' subunit
2	HEC_0010	1	2	1	23.73	2	1.33e+007	15803853	elongation factor EF-2
2	HEC_0010	2	6	1	23.51	10	9.37e+006	15803763	stringent starvation protein A
2	HEC_0010	3	2	1	20.75	2	8.54e+006	16130805	glycine decarboxylase, PLP-dependent, subunit (protein P) of glycine c
2	HEC_0010	4	3	1	16.31	4	4.72e+006	15800756	seryl-tRNA synthetase
2	HEC_0010	5	2	2	14.61	5	1.61e+006	45686198	GroEL
2	HEC_0010	6	1	1	13.61	4	7.77e+006	15802125	phenylalanyl-tRNA synthetase beta subunit
3	HEC_0020	1	31	5	96.28	16	2.05e+007	45686198	GroEL
3	HEC_0020	2	9	4	73.29	9	4.23e+007	15803853	elongation factor EF-2
3	HEC_0020	3	4	4	69.77	11	3.14e+007	15799694	molecular chaperone DnaK
3	HEC_0020	4	4	2	38.34	11	5.42e+007	15803363	5-keto-4-deoxyuronate isomerase
3	HEC_0020	5	2	2	34.44	5	9.55e+006	48783	ProRS
3	HEC_0020	6	2	2	33.67	8	2.00e+007	15804766	adenylosuccinate synthetase
3	HEC_0020	7	3	2	31.49	4	2.25e+007	16128108	pyruvate dehydrogenase, dihydrolipoyltransacetylase component E2

Figure 56 Protein-Sample Centric Rows display mode, Group results by File

Protein-Sample Centric Rows Details

This mode is similar to the Protein-Sample Centric Rows mode, except that the supporting peptide information also is displayed for each protein. You can use this mode to validate database identifications. The output looks like that in [Figure 57](#).

Note that file names are in the format Data_File_Name.aaaa.bbbb.c, where

- aaaa = first merged scan
- bbbb = last merged scan
- c = assigned precursor charge (0 means charge was ambiguous)

If you have QSTAR files, see the note on [page 30](#).

Run #	Run Name	Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Database Accession #	Protein Name
1	HEC_0000	1	1	1	17.90	2	2.44e+007	15804578	DNA-directed RNA polymerase be
#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence	MH-Matched (Da)	
1	HEC_0000.1948.1958.0	3	17.90	17.90	94.7	2.44e+007	(R)AAGEAPAAPQVTAEDASASLAELLNAGLCGSDNE (-)	3167.497	
2	HEC_0010	1	2	1	23.73	2	1.33e+007	15803853	elongation factor EF-2
#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence	MH-Matched (Da)	
1	HEC_0010.1397.1405.2	2	23.73	20.23	97.4	9.49e+006	(K)DVTTGDTLDCPDAPILLER (M)	2101.012	
2	HEC_0010.1372.1384.2	2	22.19	22.19	95.1	1.71e+007	(K)DVTTGDTLDCPDAPILLER (M)	2101.012	
2	HEC_0010	2	6	1	23.51	10	9.37e+006	15803763	stringent starvation protein A
#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence	MH-Matched (Da)	
1	HEC_0010.1547.1562.2	2	23.51	23.51	94.0	8.29e+006	(K)DNPPQDLIDLNPQSVPTLVDR (E)	2460.237	
2	HEC_0010.1561.1570.0	3	18.47	12.28	95.7	2.01e+007	(K)DNPPQDLIDLNPQSVPTLVDR (E)	2460.237	
3	HEC_0010.1594.1609.0	3	17.54	12.26	87.3	8.24e+006	(K)DNPPQDLIDLNPQSVPTLVDR (E)	2460.237	
4	HEC_0010.1534.1543.0	3	15.85	9.56	93.0	1.37e+007	(K)DNPPQDLIDLNPQSVPTLVDR (E)	2460.237	
5	HEC_0010.1616.1616.0	3	9.74	6.50	93.3	2.86e+006	(K)DNPPQDLIDLNPQSVPTLVDR (E)	2460.237	
6	HEC_0010.1568.1568.2	2	9.39	9.39	95.8	3.11e+006	(K)DNPPQDLIDLNPQSVPTLVDR (E)	2460.237	

Figure 57 Protein-Sample Centric Rows Details display mode

Protein-Peptide Distribution Columns

Select this mode to troubleshoot your fractionation method. This display mode, shown in [Figure 58](#), allows you to look at individual analyses that comprise your sample. The individual analyses could be 2D LC fractions, gel slices, or other types of fractions. This display gives insight into whether you need to adjust analysis parameters for better separation.

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum MillSummary SettingsAutovalidationEasy MS/MSMS/MS SearchSpectrum SummaryBuild TICTool BeltHelp

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData\Agilent\Trap\ecoli\heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - Files: 376 Hits: 2557
beginning to assemble proteins proteins assembled 0.381551 sec
proteins filtered by unique peptides 0.164088 sec
proteins filtered by score
calculated protein coverage maps 0.458243 sec
beginning to roll up proteins into groups ... proteins rolled up into groups 0.179455 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Database Accession #	Protein Name								
1	98	7	125.18	18	1.34e+007	45686198	GroEL								
#	Sequence	MH-Matched (Da)	HEC_0000 Intensity	HEC_0010 Intensity	HEC_0020 Intensity	HEC_0040 Intensity	HEC_0060 Intensity	HEC_0080 Intensity	HEC_0100 Intensity	HEC_0150 Intensity	HEC_0250 Intensity	HEC_0500 Intensity	HEC_1000 Intensity		
1	(K)AIAQVGTISANSDTVGK(L)	1760.9025			3.03e+008	1.55e+007									
2	(K)ANDAAGDGTITATVLAQAITEGLK(A)	2402.2409			1.12e+008	8.60e+006	5.30e+008	6.66e+007	1.26e+007		3.10e+006	6.98e+006	6.19e+006		
3	(K)ATLEDLGQAQK(R)	1045.5524			4.39e+007										
4	(K)DTTIDGVGEEAAIQGR(V)	1845.9189		1.50e+006	1.23e+008										
5	(R)QVLNCGEEPSVYANTVK(G)	1957.0059			5.48e+007										
6	(R)QQIEEATSDYDR(E)	1454.6394					7.04e+006	4.22e+006	3.10e+006						
7	(R)qQIEEATSDYDR(E)	1454.6394		1.73e+006											
8	(R)qQIEEATSDYDREK(L)	1711.7769					8.74e+006								

Group (#)Spectra (#)Distinct Peptides (#)Distinct Summed MS/MS Search Score% AA CoverageMean Peptide Spectral IntensityDatabase Accession #Protein Name

2167104.63191.50e+00715799694molecular chaperone DnaK

#	Sequence	MH-Matched (Da)	HEC_0000 Intensity	HEC_0010 Intensity	HEC_0020 Intensity	HEC_0040 Intensity	HEC_0060 Intensity	HEC_0080 Intensity	HEC_0100 Intensity	HEC_0150 Intensity	HEC_0250 Intensity	HEC_0500 Intensity	HEC_1000 Intensity
1	(K)ASSGLNEDEIGK(M)	1290.6172			6.93e+007	4.87e+007	4.40e+006						
2	(K)DDDVYDAEFEEVK(D)	1509.6591					1.33e+007	8.59e+006	2.35e+006				
3	(K)DQGIDLRNDPLAMQRLK(E)	1983.0440							2.18e+007				
4	(K)IELSSAQQTVDYNLPYITADATGPK(H)	2532.2828			1.12e+007								
5	(K)QVEEAQDKLPADDK(T)	1514.7333					3.73e+006	2.66e+006					
6	(K)qVEEAQDKLPADDK(T)	1514.7333			2.34e+007		5.25e+006						

Figure 58 Protein-Peptide Distribution Columns display mode, **Group results by File**

Protein-Peptide Comparison Columns

This display mode, shown in [Figure 59](#), is similar to Protein-Peptide Distribution Columns, but allows filename, score, SPI, and intensity to be reported together, and enables easier Excel export. This mode ensures reporting of ambiguous locations of modifications (for example, phosphorylated serine presence not distinguishable between two or more possible serine residues in a peptide).

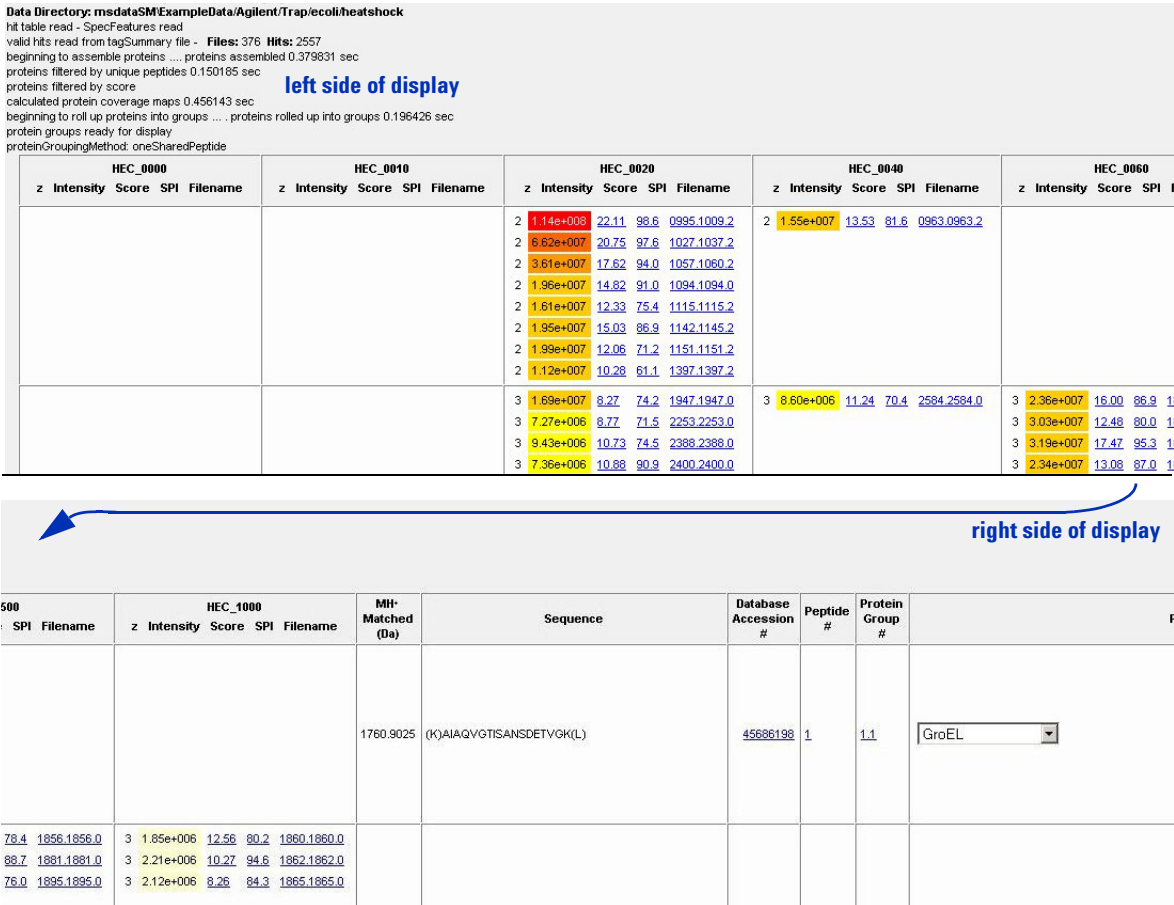
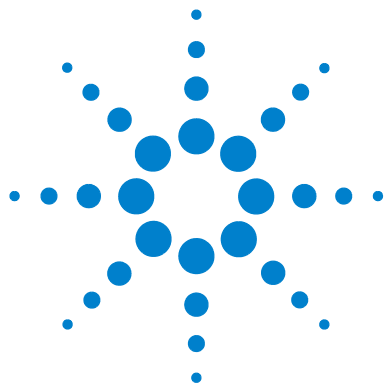


Figure 59 Protein-Peptide Comparison Columns display mode, **Group Results by File**



3 Automating Workflows

Setting Up Automated Workflows [120](#)

Executing and Monitoring Workflows [124](#)

Workflow automation lets you place a series of tasks in either a basic workflow or iterative workflow into a queue; then Spectrum Mill workbench executes them in the proper order. For example, you can automate a typical basic workflow for MS/MS data files from protein digests:

- Spectral extraction
- MS/MS search
- Autovalidation
- Protein/Peptide Summary

You can also automate an MS data analysis:

- Spectral extraction
- PMF search
- PMF Summary



Setting Up Automated Workflows

The first step to set up an automated workflow is to save the settings you intend to use in parameter files. The names of these parameter files become the “tasks” that make up the “workflow”.

You then place the workflow tasks in the order you intend. If you need to view a task before placing it in the sequence of the workflow, you can.

After creating the workflow you then execute it, monitoring the results to see the outcome of a task execution, viewing the queue to know which task is being executed at any time, and checking the completion log to view those tasks and workflows that have completed.

NOTE

For help on any section of a Workflows page, click the blue bar for that section.

To create parameter files or workflow tasks

Parameter files and workflow tasks are one in the same. You create parameter files, or edit them, in each individual Spectrum Mill page.

See [Chapter 1](#), “Processing MS/MS Data Interactively” and [Chapter 6](#), “Processing MS-Only Data Interactively” for details on how to create parameter files.

To view already existing workflows and their tasks

- 1 On the Spectrum Mill home page, click **Workflows**.
- 2 From Data Directories, click **Select**, then select a file of files and click **Open**.
- 3 Select a workflow in the list on the left to see a list of workflow tasks on the right that comprise the workflow ([Figure 60](#)).

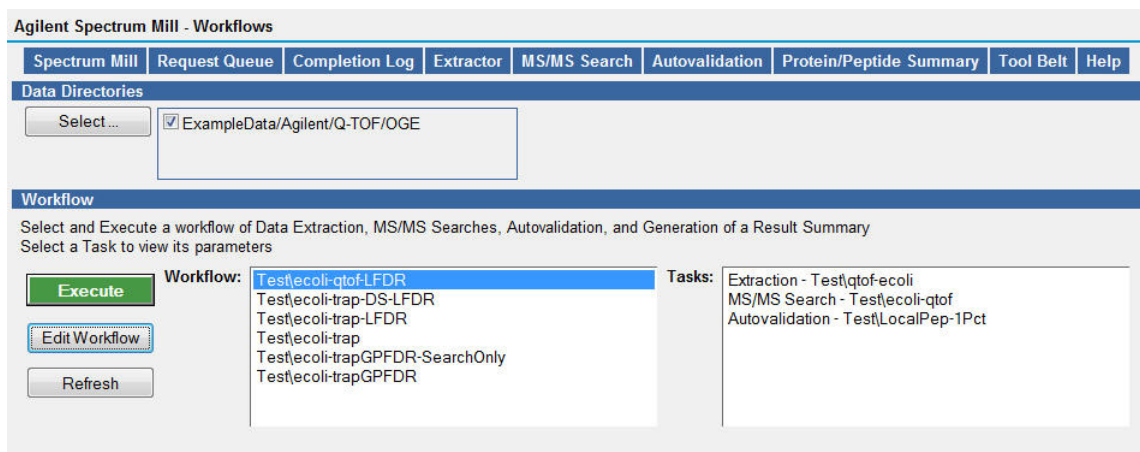


Figure 60 Workflow page with selected workflow

- 4 To make sure the parameters you want are in a task, click the task of interest to view a Read-only version of the page ([Figure 61](#)).

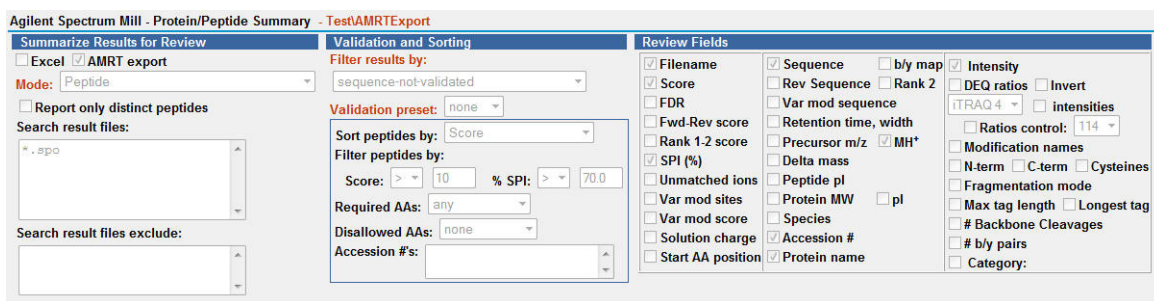


Figure 61 View of parameter values for selected workflow task

To create/edit an automated workflow

- 1 From the Workflows page, select a workflow, and click **Edit Workflow**.

Note that the name of the selected workflow to edit is in the title bar in red. (Figure 62)

If no workflows appear in the list, click **Edit Workflow** and create a new one with steps 2-6.

- 2 To filter the list of available tasks to show only a portion of them, select a category from the **Available tasks** list.

Click **Refresh** to make sure all the available tasks are in the list.

- 3 To make sure the parameter values for the selected task are the ones you want, select the task on the left and click **View** (Figure 61).
- 4 To add a task to a workflow, select the task in the list on the left, and click **Add**.

If the added task is in the wrong place in the order, select the task, then click **Up** or **Down** to move its order in the sequence.

To remove a task from the Workflow tasks list, click **Remove**.

To remove all the tasks, click **Clear All**.

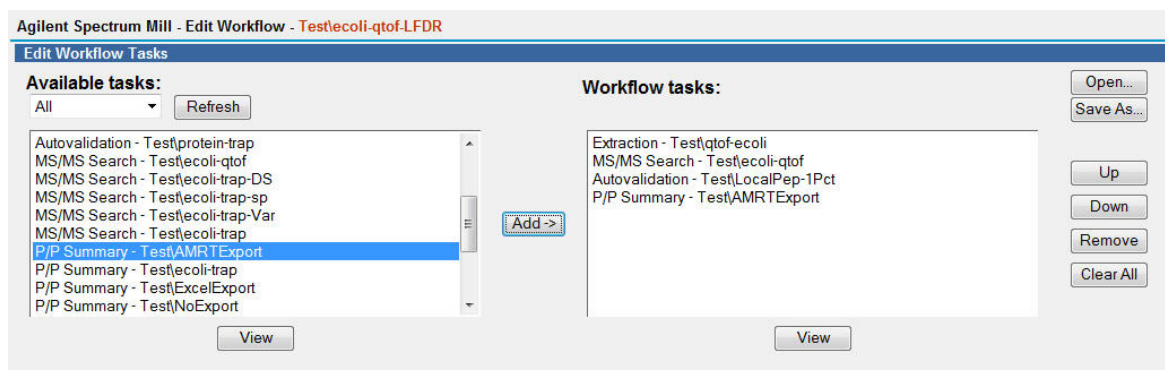


Figure 62 Edit Workflow page with new task added to the workflow

- 5 To save the edited workflow to the same name or a different one, click **Save As**, enter the name of the workflow and click **Save** (Figure 63).

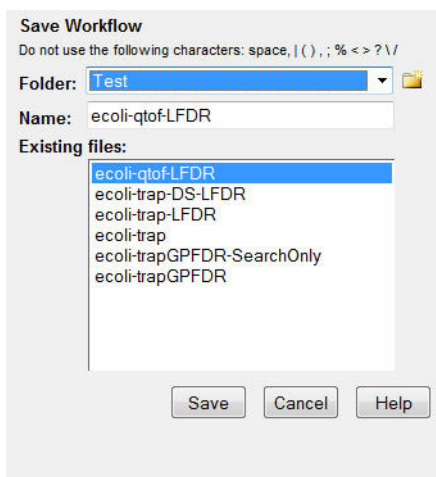


Figure 63 Save Workflow dialog box

- 6 Close the Edit Workflow window, and click **Refresh**.

The new workflow appears in the list on the left of the Workflows page.

To create/edit another workflow from within the Edit Workflow window

- 1 In the Edit Workflow window, click **Open**.

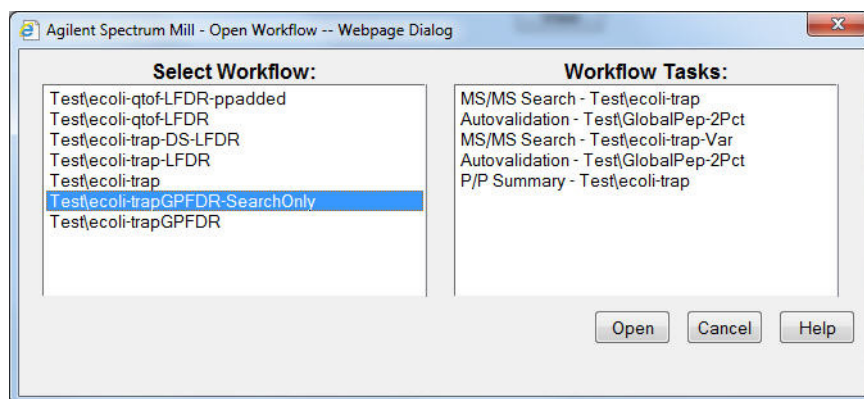


Figure 64 Open window accessed from Edit Workflow window

- 2 Select a workflow, then click **Open** and follow steps 4-6 above.

Executing and Monitoring Workflows

To execute an automated workflow

- 1 If you are not on the Workflows page, from the Spectrum Mill home page click **Workflows**.
- 2 Click **Select** to select one or more directories containing the data for executing the workflow.
- 3 Select a workflow from the list on the left, and click **Execute**.

A request submissions section for each task appears below the workflows list.

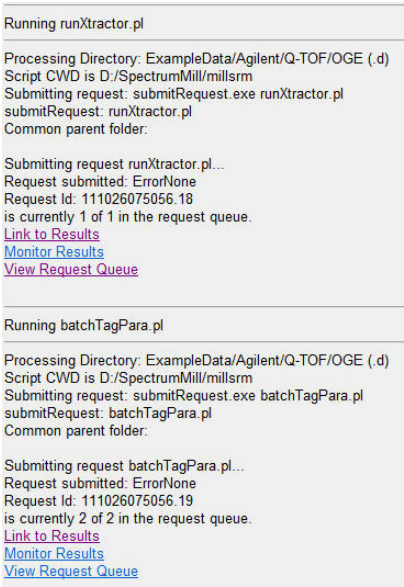


Figure 65 Request submissions list

- 4 Monitor results and view workflow progress. See below.

If you start the workflow with multiple data folders, the Spectrum Mill workbench does the Spectral Extraction and MS/MS Search on all of them first. (If you have a dual-core or quad-core CPU, it processes them in parallel.)

After search is complete, autovalidation is done independently for each folder on the same CPU unless "Group proteins across ..." is marked. After autovalidation, results for all folders for P/P Summary are summarized together.

If you want to process each of multiple folders independently, you must make sure only one is marked before execution. After completion, mark the next data folder, and execute the workflow again. Do this until all data folders have been executed.

To monitor results

To monitor the results for each task, click one of three links from the request submissions list:

Link to Results Click this link to see the results for the task in a separate window. The results in this window will not update until you close the window and click the link once more.

After you return to the request submissions list for all tasks, you can then link to the results of other tasks.

Monitor Results Click this link to see the results for the task on the Workflows page. With this choice you are able to view the updated execution messages for that task in real time but are unable to return to the request submissions list to see the links to other tasks' results.

View Request Queue Click this link to open the Request Queue viewer and select to monitor the results from that viewer. See ["To view workflow progress"](#) on page 126. With this choice you can monitor each task in real time and choose another task to monitor after closing the window.

3 Automating Workflows

Executing and Monitoring Workflows

```
Monitor output file: msdataSM\ExampleData\Agilent\Q-TOF\OGE\xttractorAgilent.111026131207.50.htm
Stop Extraction PID: 3020

Processing Directory ExampleData\Agilent\Q-TOF\OGE
All prior extraction and search/summary files have been removed!

Processing File: ec-oge-2.d using millbin\xttractorAgilent.cgi ...
Initialize extractor complete! D:\SpectrumMill\msdataSM\ExampleData\Agilent\Q-TOF\OGE\ec-oge-2.d\AcqData\oge-45.m method file is not found.
Ion Source Type: ESI D:\SpectrumMill\msdataSM\ExampleData\Agilent\Q-TOF\OGE\ec-oge-2.d\AcqData\oge-45.m method file is not found.

-----
Number of spectra: 5210
1000 scans read.
2000 scans read.
3000 scans read.
4000 scans read.
5000 scans read.
Read Scans: InitTime=0.31 secs : ReadScans=11.61 secs

-----
Data read into memory.
```

Figure 66 Results for the Extraction task (click **Monitor** in Request Queue)

To view workflow progress

- From the menu bar, click **Request Queue**,
- or
- From the request submissions list, click **View Request Queue**.

Agilent Spectrum Mill - Request Queue

Request Queue Completion Log Help

There are 4 requests in the queue.

X	#	Task Id	Status	Task Type	Data Directory	Dependencies	User
	1	111026080041.22 Monitor	Running	xttractorAgilent	ExampleData\Agilent\Q-TOF\OGE	none	Anonymous scsvpn245216
☐	2	111026080042.23 Monitor	Queued	mstag	ExampleData\Agilent\Q-TOF\OGE	111026080041.22	Anonymous scsvpn245216
☐	3	111026080042.24 Monitor	Queued	validate	ExampleData\Agilent\Q-TOF\OGE	111026080041.22	Anonymous scsvpn245216
☐	4	111026080043.25 Monitor	Queued	PPSummary	ExampleData\Agilent\Q-TOF\OGE	111026080041.22	Anonymous scsvpn245216

Figure 67 Request Queue while executing a workflow

- To monitor the individual task results from the Request Queue, click **Monitor** under the Task Id (Figure 66).
- To delete a task after the workflow starts running, mark the check box next to the task and click **Remove**.

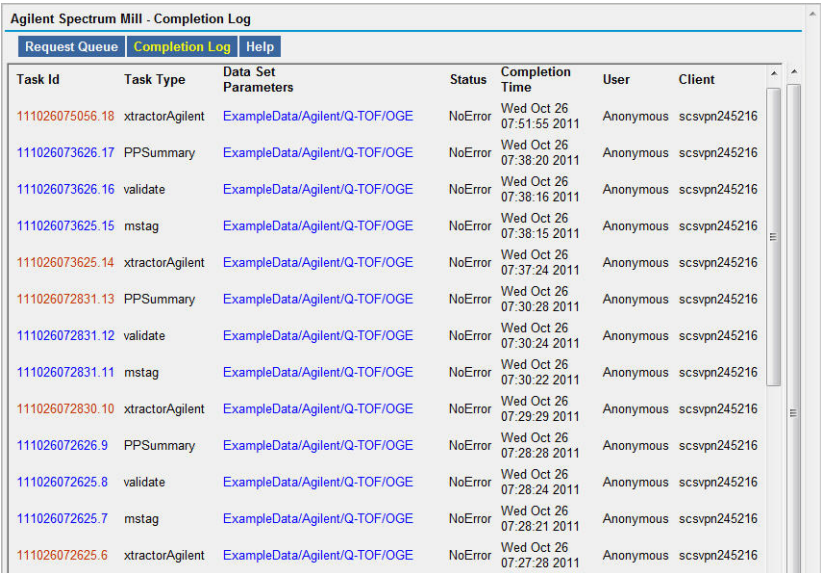
When a task is complete, it is automatically removed from the Request Queue list and appears in the Completion Log.

NOTE

You do not see any changes to the Request Queue Viewer unless you click Request Queue again to refresh the screen.

To check the Completion Log

1 In the top menu bar, click **Completion Log**.



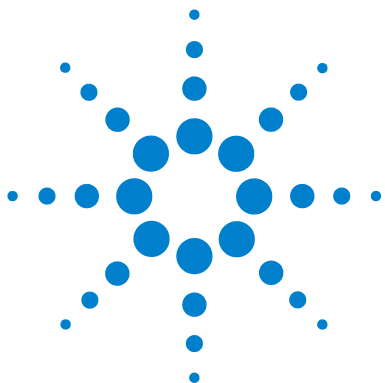
Task Id	Task Type	Data Set Parameters	Status	Completion Time	User	Client
111026075056.18	xtractorAgilent	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:51:55 2011	Anonymous	scsvpn245216
111026073626.17	PPSummary	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:38:20 2011	Anonymous	scsvpn245216
111026073626.16	validate	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:38:16 2011	Anonymous	scsvpn245216
111026073625.15	mstag	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:38:15 2011	Anonymous	scsvpn245216
111026073625.14	xtractorAgilent	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:37:24 2011	Anonymous	scsvpn245216
111026072831.13	PPSummary	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:30:28 2011	Anonymous	scsvpn245216
111026072831.12	validate	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:30:24 2011	Anonymous	scsvpn245216
111026072831.11	mstag	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:30:22 2011	Anonymous	scsvpn245216
111026072830.10	xtractorAgilent	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:29:29 2011	Anonymous	scsvpn245216
111026072626.9	PPSummary	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:28:28 2011	Anonymous	scsvpn245216
111026072625.8	validate	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:28:24 2011	Anonymous	scsvpn245216
111026072625.7	mstag	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:28:21 2011	Anonymous	scsvpn245216
111026072625.6	xtractorAgilent	ExampleData/Agilent/Q-TOF/OGF	NoError	Wed Oct 26 07:27:28 2011	Anonymous	scsvpn245216

Figure 68 Completion Log

2 To see the final result html file, click **Task Id**.

The Completion Log lets you view this information:

- Task Id - a number that identifies the workflow and the task within it.
- Task Type - extraction, search, autovalidation, summary, etc.
- Data Set Parameters - folder whose selected data files are submitted for workflow execution. Click to see a parameters table for the task.
- Status - no error, error and abort messages
- Completion Time - Date and time task completed
- User - name of user logged onto server
- Client - name of PC client being used



4 Sherenga *de novo* Sequencing

- To set parameters and run the *de novo* sequencing algorithm [131](#)
- To generate a Sherenga report [134](#)
- To view detailed Sherenga results [136](#)
- To compare Sherenga results with MS/MS Search results [138](#)

The Spectrum Mill workbench includes an advanced algorithm, called Sherenga, for *de novo* sequencing. This algorithm helps to sequence new peptides and to confirm results of database searches.

Historically, *de novo* sequencing algorithms have been difficult to develop. The major challenge in sequencing peptides from MS/MS spectra is that peptides do not usually fragment between each amino acid. While incomplete fragmentation information complicates any *de novo* program, Sherenga makes reasonable attempts to fill in the missing information.

The Sherenga algorithm uses graph theory for *de novo* peptide sequencing and sequence scoring.* Graph theory is a branch of mathematics which describes how networks (such as transportation networks) can be represented and their properties measured. A graph is an abstract symbolic representation of a network that consists of a set of vertices connected by edges. To simplify, one may use the analogy that the edges are like streets and the vertices are like intersections.

When graph theory is applied to spectra, the mass peaks are considered vertices and the differences in mass peaks are considered edges. These edges have lengths equal to the masses of the amino acids (or sometimes di- or tri-peptides). Each peak in a spectrum is assigned several possible fragment ion interpretations (vertices). Although peaks can arise from fragmentation at either the N-terminus or the C-terminus, all peaks

* Dancik, V.; Clauser, K. R.; Addona, T.A.; Vath, J. E.; Pevzner, P.A. "De novo Peptide Sequencing via Tandem Mass Spectrometry", J. Comp. Biol. 1999, 6, 327-342.



(vertices) are converted to N-terminal equivalents. Then the Sherenga algorithm attempts to find the longest possible path through as many interpretations (vertices) as possible.

To set parameters and run the *de novo* sequencing algorithm

- 1 Navigate from the Spectrum Mill home page to the Sherenga *de novo* Sequencing page, as shown in Figure 69.

Figure 69 Sherenga *de novo* Sequencing page

- 2 If you need to change the Data Directory, click **Select**, choose one, and click **OK**.
- 3 To use an existing parameter file to set parameter values, click **Load**.
- 4 Change parameters by using the instructions below and the Online Help. (Click the blue bar for each section to find explanations.)

In general, you should keep the default values, except for the data directory.

- Set the **Maximum reported hits** to the maximum number of Sherenga interpretations you would like for each spectrum.
- Set the **Validation filter** in one of the following ways:

If you took the time to designate your spectra as “good” or “bad,” as described in “[Step 6. Check for remaining high-quality spectra](#)” on page 47, set the **Validation filter** to **good-spectrum-sequence-not-validated**. This ensures that the software processes only the uninterpreted spectra you designated as “good”.

Otherwise, set the **Validation filter** to **spectrum-not-marked-sequence-not-validated**.

- For amino acids that have the same nominal mass, select how you would like to display the results under **Show Equivalent Masses**.

For **K/Q**, choose **Both** if your instrument has sufficient mass accuracy to distinguish between these two.

- Click the **Choose...** button to select the modifications that match the chemistry for your samples.

For details, click the blue bar labeled **Modifications** to access the online help.

- Select the instrument on which to base the **Scoring**.

The scoring is based on learned parameters for each type of tandem mass spectrometer. The algorithm has “learned” rules regarding ion types and intensity thresholds from sample spectra of known amino acid sequence.

- Mark **Sequence tag length** to prevent *de novo* sequencing of spectra that contain little sequence information. Do not set the tag length to less than 4.

The sequence tag length represents the longest sequence of amino acids that can be located in the spectrum. Larger numbers filter out more spectra, keeping those of the best quality.

- Click **Advanced Parameters** to display them.

These are described in the online help.

- 5 To save any changed parameter values, click **Save As**.

Once you save parameter files for this page, you can use them for workflow automation.

- 6 Click **Sequence**.

- 7 While Sherenga runs, check that you see a display like that shown in [Figure 70](#).

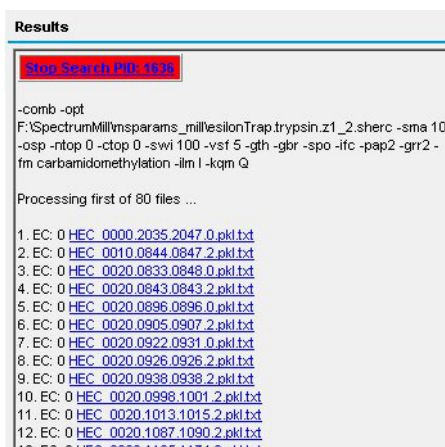


Figure 70 Display while Sherenga *de novo* Sequencing runs

- 8 If you want to view results while Sherenga runs, click a file name link. After you see the Sherenga report at the bottom of your screen, click an amino acid sequence link to display the Spectrum Viewer.

To generate a Sherenga report

- 1 When the Sherenga algorithm is complete, navigate from the Sherenga *de novo* Sequencing page to the Sherenga *de novo* Summary page.
- 2 Check that you see the page shown in [Figure 71](#).

Figure 71 Sherenga *de novo* Summary page

- 3 If it is not already selected, select the directory where your data resides.
- 4 If you want to use an already existing parameter file, click **Load**, select the file, and click **Load**.
- 5 Set the **Min. vertex score**.

If you do not want to include interpretations made on smaller peaks, raise this threshold. Amino acids interpreted based on smaller peaks are then replaced by their mass gaps.

In the detailed results that follow (see bottom blue highlight in [Figure 73](#) on page 137), **Orig. Score** represents the Sherenga score calculated with a minimum vertex score of 5, while **Score** is the Sherenga score calculated with the value you specify here.

- 6 Set other parameters. (The defaults are good choices.)
- 7 If you wish to consider only tryptic peptides in the results, mark the check box to **Force tryptic digest**.
- 8 To save the newly changed parameters, click **Save As**, name the parameter file and click **Save**.

Although parameter files can be saved, you cannot use them with workflow automation.

- 9 Click **Summarize**.
- 10 Check that you see the Sherenga top-level report, as shown in [Figure 72](#). The components of the report are described below.

Results						
MS/MS Search						
SS	TL	TS	Delta MH ⁺	Filename	Sherenga Sequence Tag	MS/MS Search Result
307	9	5.0	0.7	HEC_0020.1556.1564.2.pkl.txt	[414.52]NINVVVDAGGAGQFISA[215.27]	MGLVIKAALGALVLLIGVLAK
292	10			HEC_0020.1766.1780.2.pkl.txt	[251.35]YHINVVVDAGGVASFEG[372.49]	
285	9			HEC_0020.2376.2385.2.pkl.txt	[414.41]NINVVSGGVAGQFI[373.45]	
283	9			HEC_0020.2242.2256.2.pkl.txt	V[292.61]IFSVAAGTAVGVGN[396.32]	
280	10	5.1	0.6	HEC_0020.1537.1549.2.pkl.txt	[414.38]HGGHVVVDAGGEQFI[157.54][215.86]	MGLVIKAALGALVLLIGVLAK
278	10	6.5	0.7	HEC_0020.1387.1391.2.pkl.txt	[327.27]SGGIGGVVPGSMGQAATS[303.99]	MGLVIKAALGALVLLIGVLAK
275	10			HEC_0020.1294.1298.2.pkl.txt	[414.45]HGGHVVVEGGGAGQFI[373.48]	
274	9			HEC_0020.1510.1510.2.pkl.txt	[324.48]TIATGGTIAGGDI[158.35][229.18]	
269	10			HEC_0020.1595.1609.2.pkl.txt	[414.34]NIPASGVSGGAGEFI[157.93][215.57]	
264	9			HEC_0020.1429.1444.2.pkl.txt	[414.49]NINVVADGGGAFI[373.31]	
264	8			HEC_0020.1962.1972.0.pkl.txt	[414.59]PGFGGAAPAGGAVPHN[373.40]	
263	10			HEC_0040.1298.1311.2.pkl.txt	[212.87]IA[142.23]IQGDVAGEVGGHN[371.94]	
262	9			HEC_0020.1339.1349.2.pkl.txt	[396.35]GGNAGGVGGADVNNH[157.85][215.56]	
262	9	6.7	1.1	HEC_0020.1583.1588.0.pkl.txt	[414.39]GGHVVADNEQFI[373.44]	MGLVIKAALGALVLLIGVLAK
257	10			HEC_0020.1708.1711.2.pkl.txt	[414.50]NIGGVVSGGVEQFI[373.44]	
257	9			HEC_0020.1924.1936.2.pkl.txt	[253.10]RQSPGGAGSGTVVNIHS[286.17]	
250	8			HEC_0020.2532.2547.2.pkl.txt	[358.21]AVINVVSTGAAYGPI[373.34]	
248	9			HEC_0020.2265.2268.2.pkl.txt	[303.48]SNPGPVVGGHGGGAPIH[158.19][215.25]	

Figure 72 Sherenga top-level report

- SS** Sherenga Score for the most likely interpretation
- TL** Tag Length - sequence tag length
- TS** Tag Score - peptide score for MS/MS Search results
- MS/MS Search Delta MH⁺** Difference in mass between experimental MH⁺ and MH⁺ of database match
- Filename** Filename for detailed Sherenga results. These are in the format Data_File_Name.aaaa.bbbb.c.pkl.txt, where
 - aaaa = first merged scan
 - bbbb = last merged scan
 - c = assigned precursor charge (0 means charge was ambiguous)

If you have QSTAR files, see the note on [page 30](#).

Sherenga Sequence Tag	Amino acid sequence for Sherenga interpretation with highest score. A bracketed number in the sequence indicates a mass segment that could not be reliably interpreted because there were insufficient cleavages to assign amino acids or because there was an unusual modification of an amino acid.
MS/MS Search Result	Amino acid sequence for best MS/MS Search interpretation

To view detailed Sherenga results

- 1 Click a link under the **Filename** header.
- 2 Check that you see the Spectrum Viewer, as shown at the top of [Figure 73](#), and the Sherenga results list, as shown at the bottom of [Figure 73](#).

The Sherenga results list is comprised of multiple Sherenga interpretations for the same spectrum. The most probable interpretation (**Sherenga 1**) is initially displayed in the Spectrum Viewer.

- 3 Click the **Rank** arrow buttons in the lower right-hand corner of the Spectrum Viewer to display the various interpretations in the Sherenga results list. See [“To use the Spectrum Viewer”](#) on page 96 of [Chapter 2](#) for more details.

NOTE

If you see an **MStag 1** sequence at the top of the Spectrum Viewer, first click the **Sherenga 1** sequence and make sure the background for the Sherenga sequence turns from gray to white. Then click the **Rank** arrow buttons.

- 4 Scroll below the Sherenga results list to view details on the interpretation.
- 5 If you have additional information about the sequence, you can evaluate your own sequence versus the spectrum.
 - a Be sure that you have selected the Sherenga sequence as described in the note just above. That way your sequence will append to the Sherenga list.
 - b Type your sequence in the box to the left of the **Go** button in the Spectrum Viewer. Where you have missing information, enter mass

- gaps in brackets, as illustrated with the default sequence shown in the box. You can enter these mass gaps anywhere in the sequence, including the middle or ends.
- c Click the **Go** button.
 - d Select **Bars** options to see how the theoretical fragment ions from your sequence line up with the spectrum.
 - e Use the **Rank** arrow buttons (<- and ->) to go from the sequences that were identified by Sherenga *de novo* Sequencing to the sequence that you typed.

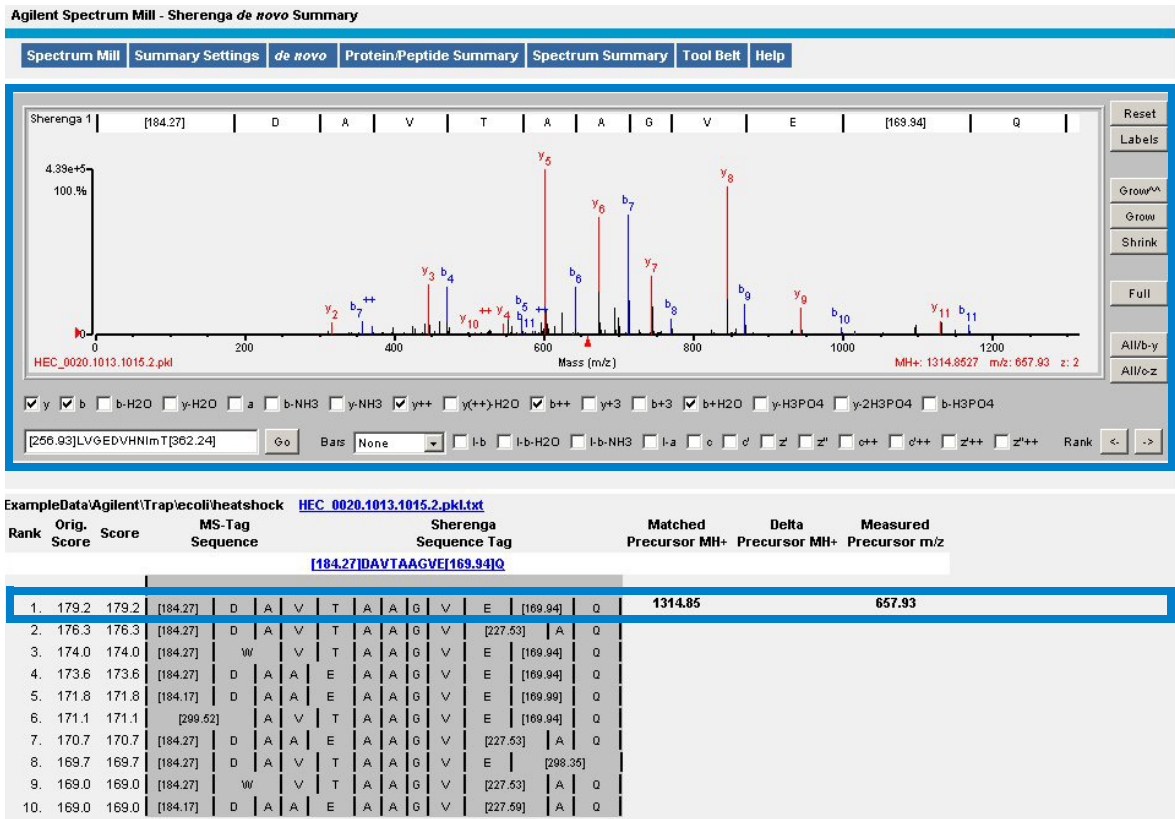


Figure 73 First Sherenga interpretation outlined in rectangles in both the Spectrum Viewer (top) and Sherenga results list (bottom)

To compare Sherenga results with MS/MS Search results

- In the detailed Sherenga results, compare the *de novo* sequencing results with any available MS/MS Search results.

When database search results are available, the software displays them along with the Sherenga interpretations. The database sequence (labeled **MSTag 1**) is displayed just above the Sherenga sequence shown at the top of [Figure 73](#). It is also displayed above the Sherenga results list shown at the bottom of [Figure 73](#).



5 Processing Data for Differential Expression Quantitation

- To use the Data Extractor for a differential profiling study [141](#)
- To use MS/MS Search for a differential profiling study [143](#)
- To calculate DEQ ratios for isotopic labels using the Protein/Peptide Summary page [144](#)
- To interpret DEQ results for isotopic labels in peptide mode [145](#)
- To interpret DEQ results for isotopic labels in protein modes [147](#)
- To calculate iTRAQ and TMT ratios using the Protein/Peptide Summary page [148](#)
- To interpret results for iTRAQ labels in peptide mode [149](#)
- To interpret results for iTRAQ labels in protein mode [150](#)
- To view light/heavy results on the Spectrum Summary page [151](#)

In this chapter, you learn in detail how to process data for differential expression quantitation (DEQ). You learn how to process isotope-coded affinity tag (ICAT) and other isotopically labeled data to compute light/heavy ratios. You also learn how to do quantitation with isobaric modifications, which use isobaric tags with low-mass MS/MS signature ions rather than isotopic labels (iTRAQ and TMT).

Labeled data is in general processed the same way as any other MS/MS data, so you follow the step-by-step procedures described in [Chapter 1](#), “Processing MS/MS Data Interactively”. As you process the data, you set specific parameters in the Data Extractor, MS/MS Search, Protein/Peptide Summary, and Spectrum Summary pages to enable differential expression quantitation. This chapter describes those settings.

At installation, the software supports the following modifications for differential expression quantitation:



- ICAT D₀/D₈
- cICAT ¹²C/¹³C
- C-terminal methyl ester D₀/D₃
- N-terminal propionyl D₀/D₅
- C-terminal ¹⁶O/¹⁸O
- iTRAQ and TMT reagents
- Metabolic labels:
 - SILAC 2 (Arg 0-6Da, Lys 0-8Da)-mix
 - SILAC 3 (Arg 0-6-10Da)-mix
 - ¹⁵N and ¹⁴N/¹⁵N mix

The software allows you to add custom modifications to support quantitation studies. See the online help for system administrators.

Note that with the exception of ¹⁴N/¹⁵N, iTRAQ and TMT calculations, differential expression calculations are not supported for data that uses the generic Data Extractor. Because the generic extractor processes exported rather than raw files, it retains less functionality than the raw file extractors.

To use the Data Extractor for a differential profiling study

- 1 Click the **Choose...** button to set the appropriate **Fixed/Mix Modification**. [Table 5](#) lists the supported labels and the corresponding options you must select. If your sample contains a mixture of labels and you want the software to calculate the ratio, select one of the **mix** options.
- 2 Set other parameters as described in [Chapter 1](#).

Table 5 Reagents for differential expression quantitation and corresponding Spectrum Mill settings

Reagent	Fixed/Mix Modification listed under:	Option to select
ICAT D ₀ /D ₈ (original ICAT reagent)	Cysteine	ICAT-D0, ICAT-D8, or ICAT-mix
cICAT ¹² C/ ¹³ C (newer, cleavable ICAT reagent)	Cysteine	cICAT-C12, cICAT-C13, or cICAT-mix
N-terminal propionyl D ₀ /D ₅	N-terminus	Propionyl-D0, Propionyl-D5, or Propionyl-mix
iTRAQ/TMT	N-terminus and K	iTRAQ or iTRAQ Partial-mix (the latter if you suspect incomplete labeling) or TMT
C-terminal methyl ester D ₀ /D ₃	C-terminus, D, and E	Methyl_Ester, Methyl_Ester-D3, or Methyl_Ester-mix
C-terminal ¹⁶ O/ ¹⁸ O	C-terminus	O18 Free Acid or O16/O18 Free Acid-mix
¹⁵ N and ¹⁴ N/ ¹⁵ N mix	Metabolic	N15 or N14/N15 mix
SILAC	Metabolic	Any of the SILAC options (All are mixes.)

Extractor details for all -mix modifications except iTRAQ, TMT and ¹⁴N/¹⁵N mix

The **mix** options ensure that all the necessary chromatogram peak areas are calculated as the data are extracted. Data Extractor always calculates peak areas from the extracted ion chromatograms (EICs) of each precursor ion subjected to MS/MS.

When the modification is set to a mixture of light/heavy labels, Data Extractor calculates peak areas for a number of additional parallel EICs, one of which represents the other member of the light/heavy pair. More EICs are calculated than are actually necessary for the light/heavy computation. The reason for this in the case of ICAT quantitation is that

Data Extractor does not know whether it has extracted a cysteine-containing peptide, a light member of the pair, or a heavy member of the pair. It also does not know how many cysteines are present, nor the charge state of the precursor ions.

All the parallel EIC areas are stored until the Spectrum Mill workbench's interpretation is complete. Then the correct ones are retrieved for ICAT calculations. As a result, EICs are calculated for both the light and heavy labels, even if only one is subjected to MS/MS. A similar process occurs for SILAC, which can include light/medium/heavy labels.

The generic Data Extractor (the one that processes peak lists) does not have access to the raw mass spectral files, so it cannot calculate the EICs that are needed for ICAT-like quantitation.

Extractor details for iTRAQ and TMT

The iTRAQ and TMT intensity calculations do not require EICs. The abundances for the reporter ions are calculated from the MS/MS data.

You can calculate iTRAQ and TMT ratios for generic extractor data.

Extractor details for $^{14}\text{N}/^{15}\text{N}$ mix

For $^{14}\text{N}/^{15}\text{N}$ mix, quantitation occurs at the report level rather than at the extractor level. The Data Extractor does not calculate parallel EICs; so quantitation is possible only when MS/MS is run on both the ^{14}N and the ^{15}N members of a peptide pair, and the MS/MS Search results indicate the same sequence and charge. The calculations assume complete incorporation of ^{15}N .

The Data Extractor provides quantitative information for all ions for which MS/MS is run, and the ratios are calculated in Protein/Peptide Summary based on this information. You can calculate $^{14}\text{N}/^{15}\text{N}$ ratios with generic extractor data.

CAUTION

If you wish to do quantitation for labels with small mass differences (~4 Da) between the light and heavy versions, change **Merge scans with same precursor m/z** to **+/-1.0 m/z** or lower.

To use MS/MS Search for a differential profiling study

- 1 Click the **Choose...** button to set the appropriate **Fixed/Mix Modification**. See [Table 5](#) on page 141 for details.
- 2 Set other parameters as described in [Chapter 1](#), but only mark **Proton mobility scoring** if the MS/MS spectra are acquired on an ion trap mass spectrometer.

With the exception of iTRAQ and TMT, if you select one of the variations that ends in **mix**, each spectrum is searched multiple times – once for each possible label. The results are merged as a single output.

For iTRAQ and TMT, only a single search is necessary. Since the tags are isobaric, all versions of the iTRAQ or TMT label, whether iTRAQ4, iTRAQ8, TMT2 or TMT6, are simultaneously fragmented during MS/MS. Further, all iTRAQ labels produce the same MS/MS fragments for a given precursor peptide, as do TMT labels. Therefore, neither the iTRAQ nor TMT labels have to be searched as a mix. However, the tags produce different peaks depending on the mass associated with the tag, and the abundances of these peaks are used by the Spectrum Mill workbench for relative quantitation.

To calculate DEQ ratios for isotopic labels using the Protein/Peptide Summary page

Use the Protein/Peptide Summary page to calculate DEQ ratios.

- 1 Mark the **DEQ ratios** (differential expression quantitation ratios) check box.

NOTE

Only for protein modes: When you mark the DEQ ratios check box, you can select the Protein Quantitation Option to exclude outlier DEQ ratios (>2 std dev from mean).

- 2 If you want to see your modification in a report that shows peptides:
 - Mark the check box for **N-terminus**, **C-terminus**, or **Cysteines**, if your label reacts at one of those sites. (For example, mark **Cysteines** for ICAT reagents.)
 - Mark the check box for **Modification Names** if your label reacts at an amino acid other than cysteine.
- 3 For ICAT and similar labels, when you want to see the pairs together in a peptide report, set **Sort peptides by** to **Sequence**.
- 4 (Optional) Use the **Required AAs** box to filter results so that only the isotopically labeled peptides are shown. For example, select **C** in the **Required AAs** box if the isotopic label modifies cysteine.
- 5 Set other options.
- 6 Click **Summarize**.

To interpret DEQ results for isotopic labels in peptide mode

- Check that your results resemble those in [Figure 74](#), which shows ICAT results displayed in peptide mode on the Protein/Peptide Summary page.
- Note the columns labeled **Cysteines** and **L/H**. **L/H** means light/heavy. For SILAC reagents with three labels, you also see **L/M**, for light/medium.
- If **n/c** appears in the report, it means not calculated. If the Data Extractor cannot determine a charge for a peptide, it assumes a charge of +2 for determining the mass shifts for quantitation, and it looks for up to two modification sites in the peptide (e.g., two cysteines at most). When the actual charge is not +2, or when there are more than two modification sites in the peptide, the ratio is not calculated, and is reported as **n/c**.

Ratios are also reported as **n/c** when the peptide does not contain the amino acid that reacts with the labeling reagent.

Agilent Spectrum Mill - Protein/Peptide Summary														
Spectrum Mill		Summary Settings		Autovallation		Easy MS/MS		MS/MS Search		Spectrum Summary		Build TIC	Tool Belt	Help
Results Shown Filtered by Validation Category: valid														
Data Directory: msdataSM\ExampleData\ICAT														
hit table read - SpecFeatures read														
valid hits read from tagSummary file - Files: 41 Hits: 40														
#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence	Cysteines	L/H	MH+ Matched (Da)	Accession #	Protein Na		
1	ICAT_Mix.1064.1064.0	2	14.17	14.17	85.1	8.13e+006	(K)AGLCQTFVYGGCR(A)	ICAT-D0	0.891	2259.079	115114	Aprotinin		
2	ICAT_Mix.0837.0842.2	2	14.02	14.02	71.5	1.03e+008	(K)CDEWSVNSVCK(I)	ICAT-D8	0.717	1673.811	4557871	transferrin		
3	ICAT_Mix.0843.0847.2	2	14.13	14.13	75.2	8.62e+007	(K)CDEWSVNSVCK(I)	ICAT-D0	0.812	1665.761	4557871	transferrin		
4	ICAT_Mix.1083.1083.2	2	8.88	8.88	78.4	2.94e+007	(K)CSTSSLLEACTFR(R)	ICAT-D8	0.266	2318.195	4557871	transferrin		
5	ICAT_Mix.1095.1098.2	2	11.08	11.08	75.7	1.47e+007	(K)CSTSSLLEACTFR(R)	ICAT-D0	1.090	2302.095	4557871	transferrin		
6	ICAT_Mix.1067.1067.0	3	11.61	11.61	97.3	5.54e+007	(K)DAIPENLPPLTADFADKDVCK(N)	ICAT-D8	n/c	2851.434	1351907	bovine seru		
7	ICAT_Mix.1068.1073.0	3	19.67	19.67	94.4	5.00e+007	(K)DAIPENLPPLTADFADKDVCK(N)	ICAT-D0	n/c	2843.384	1351907	bovine seru		
8	ICAT_Mix.1072.1072.0	3	10.74	10.74	95.6	6.29e+007	(K)DAIPENLPPLTADFADKDVCK(N)	ICAT-D8	n/c	2851.434	1351907	bovine seru		
9	ICAT_Mix.0863.0863.0	2	7.91	7.19	89.2	1.72e+007	(K)DDPHACYSTVFCK(L)	ICAT-D0	1.559	1939.857	1351907	bovine seru		
10	ICAT_Mix.0997.0997.0	2	12.58	12.58	77.4	5.52e+007	(K)DYELLCLDCTR(K)	ICAT-D8	0.492	1747.884	4557871	transferrin		
11	ICAT_Mix.1002.1008.0	2	21.43	21.43	98.2	3.80e+007	(K)DYELLCLDCTR(K)	ICAT-D0	0.591	1739.834	4557871	transferrin		
12	ICAT_Mix.1003.1003.2	2	12.83	12.83	74.9	5.84e+007	(K)DYELLCLDCTR(K)	ICAT-D8	0.497	1747.884	4557871	transferrin		
13	ICAT_Mix.0844.0844.0	3	11.28	11.28	73.6	1.77e+007	(R)EGTCPEAPTDECKPVK(W)	ICAT-D8	n/c	2604.312	4557871	transferrin		
14	ICAT_Mix.0922.0922.0	2	8.25	8.25	63.2	2.14e+007	(K)EYEATLECCAK(D)	ICAT-D8	0.339	2289.121	1351907	bovine seru		
15	ICAT_Mix.0967.0967.0	2	12.36	12.36	84.0	4.36e+007	(R)FDEFFSEGCAPGSK(K)	ICAT-D8	0.416	1970.911	4557871	transferrin		
16	ICAT_Mix.0972.0972.0	2	9.94	9.94	79.7	5.25e+007	(R)FDEFFSEGCAPGSK(K)	ICAT-D8	0.582	1970.911	4557871	transferrin		
17	ICAT_Mix.0908.0908.2	2	11.20	11.20	83.7	1.00e+007	(R)IGLNCLQAQVAER(V)	ICAT-D0	1.230	1856.972	114939	beta-galact		

Figure 74 ICAT results shown in peptide mode

In the peptide mode, you may see a calculation discrepancy in cases where both members of a D₈/D₀ ICAT pair have been subjected to MS/MS. This is because the heavy- and light-labeled peptides often do not quite co-elute, so the EICs used for the calculations are slightly offset. (The EIC windows are determined by the Data Extractor setting for the time window for merging scans.)

You do not see this discrepancy in the protein mode, *provided* that both the D₈- and D₀-labeled precursor ions were subjected to MS/MS and that these results were of sufficient quality to be interpreted and included in the final results summary. In the protein modes, the peptide ratios are combined to calculate a ratio for the protein. In these modes, the ratio for the ICAT pair is recalculated directly from the EICs of each precursor, rather than the parallel EICs from the calculated *m/z* shifts from the precursor.

To interpret DEQ results for isotopic labels in protein modes

- Check that your results resemble those in [Figure 75](#), which shows ICAT results displayed in one of the protein modes on the Protein/Peptide Summary page.
- Note the special light/heavy columns labeled **L/H (mean)**, **L/H (std dev)**, and **L/H (# pairs)**. For SILAC reagents with three labels, you also see columns for **L/M (mean)** and **L/M (std dev)**.

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill

Summary Settings

Autovalidation

Easy MS/MS

MS/MS Search

Spectrum Summary

Build TIC

Tool Belt

Help

Results Shown Filtered by Validation Category: valid

Data Directory: msdataSM\ExampleData\ICAT

hit table read - SpecFeatures read

valid hits read from tagSummary file - Files: 41 Hits: 40

beginning to assemble proteins proteins assembled 0.005125 sec

proteins filtered by unique peptides 0.003677 sec

proteins filtered by score

calculated protein coverage maps 0.039664 sec

beginning to roll up proteins into groups proteins rolled up into groups 0.002366 sec

protein groups ready for display

proteinGroupingMethod: oneSharedPeptide 6 Proteins listed

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	L/H (mean)	L/H (std dev)	L/H (# pairs)	Protein MW (Da)	Database Accession #	Protein Name
1	24	10	148.36	19	3.15e+007	0.72	0.23	11	77050.4	4557871	transferrin
2	10	7	83.55	16	2.67e+007	0.91	0.47	5	69293.9	1351907	bovine serum albumin
3	2	1	17.20	2	4.40e+007	0.84	0.00	1	53354.2	71826	fibrinogen beta chain
4	2	1	16.34	3	4.52e+007	0.96	0.00	1	42881.5	129293	ovalbumin
5	1	1	14.17	13	8.13e+006	0.89	0.00	1	10902.8	115114	Aprotinin
6	1	1	11.20	1	1.00e+007	1.23	0.00	1	116483.5	114939	beta-galactosidase
Totals:	40	21									

Figure 75 ICAT results shown in Protein Summary mode

The protein modes all work similarly. First, the interpreted spectra for peptides are grouped because they correspond to a single protein. Then a light/heavy ratio for the protein is calculated from the mean of the values for the peptides. The protein modes report the mean, standard deviation, and number of values that contribute to the mean. Note that the standard deviations are larger for low-level, noisy data.

To calculate iTRAQ and TMT ratios using the Protein/Peptide Summary page

Use the Protein/Peptide Summary page to calculate iTRAQ and TMT ratios. In the iTRAQ calculations, the Spectrum Mill workbench incorporates the correction factors from your iTRAQ certificate of analysis. You must first use the Tool Belt page to enter the correction factors and apply them to your sample.

- 1 Do the following on the Tool Belt page:
 - a Enter the correction factors from the iTRAQ certificate of analysis that you received for the reagents that you used to process your sample. See [“To create a file of iTRAQ correction factors”](#) on page 193.
 - b Apply the correction factors from the iTRAQ certificate of analysis to the data directory for your sample. See [“To apply a file of iTRAQ correction factors”](#) on page 194.
- 2 Do the following on the Protein/Peptide Summary page:
 - a Select the iTRAQ or TMT (Tandem Mass Tag) option you are using for your quantitation: iTRAQ4, iTRAQ8, TMT2 or TMT6.
 - a Mark the **intensities** check box (for modes that show peptides).
 - b Mark the **Reporter Ratios** check box and select the reporter mass you wish to use as the control (denominator) for the ratio calculations.
 - c If you want to see the iTRAQ modification in a report that shows peptides, mark check boxes for both **N-terminus** and **Modification Names**, since the reagents react at both the N-terminus and lysines.
 - d Set other options.
 - e Click **Summarize**.

To interpret results for iTRAQ labels in peptide mode

- Check that your results resemble those in [Figure 76](#), which shows iTRAQn results displayed in peptide mode on the Protein/Peptide Summary page.
- Confirm that you see a message near the top of the report that says the correction factors were found, and confirm that the batch number is correct.
- Note the columns that show the iTRAQ intensities and ratios.

Agilent Spectrum Mill - Protein/Peptide Summary																			
Spectrum Mill		Summary Settings		Autovallation		Build TIC		MS/MS Search		Spectrum Summary		Tool Belt		Help					
Results Shown Filtered by Validation Category: valid Data Directory: msdata\SMExampleData\iTraQ hit table read - SpecFeatures read Correction factors for iTRAQ batch 0501017 found. valid hits read from tagSummary file - Files: 17 Hits: 17																			
#	Filename	z	Score	SPI (%)	Spectrum Intensity	H-term	Sequence	Modifications	iTRAQ 114	iTRAQ 115	iTRAQ 116	iTRAQ 117	iTRAQ 118	iTRAQ 115/114	iTRAQ 116/114	iTRAQ 117/114	MH-Matched (Da)	Accession #	Protein Name
1	24prot_114_116.0229.1323.2	2	15.08	99.6	3.25e+004	iTRAQ	(K) GILAADESTGSIK(R)	K:iTRAQ	1074	88	968	15		0.000	0.898	0.000	1620.905	P04075	Fructose-bisphosphate aldol
2	24prot_114_116.0200.1158.3	3	14.80	93.5	7.00e+004	iTRAQ	(K) SHQDQVLYGNR(L)		822	56	551	6		0.000	0.667	0.000	1389.700	P13745	Glutathione S-transferase Ya
3	24prot_114_116.0213.1227.2	2	13.83	87.1	6.90e+004	iTRAQ	(R) ALQASALK(A)	K:iTRAQ	4051	431	2983	174		0.025	0.730	0.010	1089.688	P04075	Fructose-bisphosphate aldol
4	24prot_114_116.0242.1386.4	4	13.62	92.9	6.99e+004	iTRAQ	(K) RTEREDLIATLK(K)	K:iTRAQ	623	37	298			0.000	0.476	0.000	1911.128	P00005	Cytochrome c
5	24prot_114_116.0204.1177.3	3	13.29	98.7	5.09e+004	iTRAQ	(K) SHQDQVLYGNR(L)	K:iTRAQ	780	63	483	13		0.003	0.615	0.000	1505.796	Q28035	Glutathione S-transferase alp
6	24prot_114_116.0205.1182.3	3	12.92	71.8	3.86e+004	iTRAQ	(R) LQSIGTENTENRR(F)		2008	189	1292	74		0.016	0.638	0.008	1790.912	P04075	Fructose-bisphosphate aldol
7	24prot_114_116.0251.1446.3	3	11.50	100.0	2.90e+004	iTRAQ	(K) YSHEETIAMATVTLR(R)		191	6	104			0.000	0.543	0.000	1835.945	P00883	Fructose-bisphosphate aldol
8	24prot_114_116.0227.1307.3	3	11.34	77.7	4.82e+004	iTRAQ	(K) ADDGPPFPQVIK(S)	K:iTRAQ	1417	167	856	30		0.041	0.598	0.000	1630.916	P04075	Fructose-bisphosphate aldol
9	24prot_114_116.0207.1196.3	3	9.58	87.1	4.99e+004	iTRAQ	(K) KLFVQK(C)	K:iTRAQ	845	43	479			0.000	0.565	0.000	1194.794	P04657	Cytochrome c-1 (Cytochrome
10	24prot_114_116.0215.1245.2	2	9.41	74.2	8.01e+004	iTRAQ	(K) YNLYGK(D)	K:iTRAQ	4808	371	3127	201		0.000	0.645	0.013	1045.593	Q28035	Glutathione S-transferase alp
11	24prot_114_116.0240.1386.3	3	9.39	84.0	3.49e+004	iTRAQ	(R) KLFEELAR(S)	K:iTRAQ	421	31	231			0.000	0.545	0.000	1293.777	P11974	Pyruvate kinase, M1 isozyme
12	24prot_114_116.0254.1466.3	3	9.30	93.3	2.41e+004	iTRAQ	(K) TEREDLIATLK(K)	K:iTRAQ	364	8	165			0.000	0.452	0.000	1638.931	P00005	Cytochrome c
13	24prot_114_116.0210.1210.3	3	9.28	85.0	2.12e+004	iTRAQ	(R) EAAAMFHR(K)		160	8	98			0.000	0.610	0.000	1205.586	P11979	Pyruvate kinase, M1 isozyme
14	24prot_114_116.0227.1311.4	4	8.74	83.3	3.70e+004	iTRAQ	(K) TQPNLHGLFGRK(T)	K:iTRAQ	426		196			0.000	0.460	0.000	1584.922	P00020	Cytochrome c
15	24prot_114_116.0206.1185.3	3	8.09	85.9	3.13e+004	iTRAQ	(K) ELSDIAHR(I)		167	14	91			0.008	0.541	0.000	1084.587	P04075	Fructose-bisphosphate aldol
16	24prot_114_116.0244.1406.2	2	7.95	95.0	3.66e+004	iTRAQ	(R) GDLCIEIIPAEK(V)	K:iTRAQ	951	52	711	9		0.000	0.745	0.000	1429.815	P11979	Pyruvate kinase, M1 isozyme
17	24prot_114_116.0202.1166.2	2	7.32	88.9	7.16e+004	iTRAQ	(R) IVAPGR(G)	K:iTRAQ	5972	527	4452	204		0.007	0.740	0.001	872.581	P04075	Fructose-bisphosphate aldol
17 Files listed.																			

17 Files listed.

Figure 76 iTRAQ results shown in peptide mode

To interpret results for iTRAQ labels in protein mode

- Check that your results resemble those in [Figure 77](#), which shows iTRAQ4 results displayed in one of the protein modes on the Protein/Peptide Summary page.
- Confirm that you see a message near the top of the report that says the correction factors were found, and confirm that the batch number is correct.
- Note the columns that show the means and standard deviations of iTRAQ ratios, as well as the number of iTRAQ pairs. For the proteins, the software calculates the means and standard deviations of the iTRAQ ratios from the component peptides.

Agilent Spectrum Mill - Protein/Peptide Summary														
Spectrum Mill Summary Settings Autovalidation Build TIC MS/MS Search Spectrum Summary Tool Belt Help														
Results Shown Filtered by Validation Category: valid Data Directory: metadata\SampleData\iTRAQ Hit table read - SpecFeatures read Correction factors for iTRAQ batch 0501017 found. valid hits read from tagSummary file - Files: 17 Hits: 124 beginning to assemble proteins ... proteins assembled 0.012857 sec proteins filtered by unique peptides 0.019828 sec proteins filtered by score calculated protein coverage maps 0.108943 sec beginning to roll up proteins into groups ... proteins rolled up into groups 0.014742 sec protein groups ready for display proteinGroupingMethod: oneSharedPeptide 5 Proteins listed														
Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	iTRAQ 115/114 (mean)	iTRAQ 115/114 (std dev)	iTRAQ 116/114 (mean)	iTRAQ 116/114 (std dev)	iTRAQ 117/114 (mean)	iTRAQ 117/114 (std dev)	iTRAQ (# spectra)	Database Accession #	Protein Name
1	7	7	80.08	21	4.57e+004	0.01	0.01	0.67	0.12	0.00	0.00	7	P00883	Fructose-bisphosphate aldolase A (EC 4.1.2.13) (Muscle-type aldolase)
2	4	4	41.24	28	4.52e+004	0.00	0.00	0.49	0.04	0.00	0.00	4	P00004	Cytochrome c
3	3	3	26.62	5	3.09e+004	0.00	0.00	0.63	0.08	0.00	0.00	3	P11974	Pyruvate kinase, M1 isozyme (EC 2.7.1.40) (Pyruvate kinase muscle)
4	2	2	24.21	7	7.50e+004	0.00	0.00	0.66	0.01	0.01	0.01	2	P30115	Glutathione S-transferase Yc (EC 2.5.1.18) (GST class-alpha)
5	2	2	22.70	7	6.55e+004	0.00	0.00	0.63	0.02	0.01	0.01	2	P09210	Glutathione S-transferase A2 (EC 2.5.1.18) (GTH2) (HA subunit 2) (G)
Totals:	18	18												

Figure 77 iTRAQ results shown in Protein Summary mode

To view light/heavy results on the Spectrum Summary page

- To view ICAT or other light/heavy intensities on the Spectrum Summary page, mark the check box for **MS L/H EIC intensity**. Check that the results are similar to those in [Figure 78](#).

Agilent Spectrum Mill - Spectrum Summary

Spectrum Mill Summary Settings Protein/Peptide Summary Build TIC MS/MS Search Tool Belt Help

Down Up ICAT_Mix.1164.1164.2.pkl 1 of 105 Update

reset good-spectrum bad-spectrum Click to set category of current spectrum

Data Directory: msdata\SM\ExampleData\ICAT
Validation Preset: all Input files: 105 - 105 files (after filtering by maxSequenceTagLength > 3)

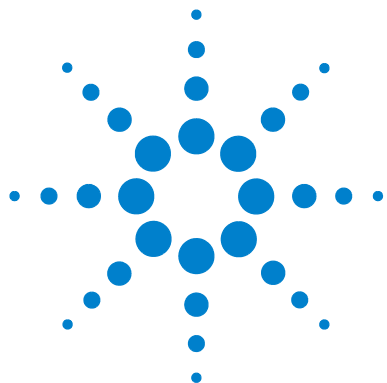
Validation category	#	Filename	z	Precursor MH-	Precursor m/z	# Detected Peaks	MS Precursor EIC Intensity	Light Intensity	Heavy Intensity	Light Intensity 2 sites	Heavy Intensity 2 sites	MS Order	2nd Precursor m/z	Max Tag Length	Longest Sequence Tag
reset	1	ICAT_Mix.1164.1164.2	2	2127.0300	1064.0200	25	1.01e+007	1.53e+005	1.73e+007	1.59e+006	4.18e+005	2	0.0000	9	[401.15]LLSLYDETGI[732.63]
reset	2	ICAT_Mix.0738.0738.2	2	1703.2500	852.1300	25	7.27e+006	2.02e+006	7.56e+006	4.30e+006	8.03e+005	2	0.0000	8	[447.21]HESVAEREI[317.78]
reset	3	ICAT_Mix.0843.0847.2	2	1666.3300	833.6700	25	6.62e+007	2.69e+006	1.06e+008	1.80e+006	6.40e+006	2	0.0000	8	[402.27]TVSIEDRM[202.78]
reset	4	ICAT_Mix.1168.1173.2	2	1816.2100	908.6100	25	9.26e+006	2.66e+006	8.67e+005	3.90e+006	3.52e+006	2	0.0000	8	[389.16]VNLVGGWMM[371.64]
reset	5	ICAT_Mix.0837.0842.2	2	1674.3500	837.6800	25	1.01e+008	7.40e+007	5.71e+006	2.27e+006	1.24e+007	2	0.0000	7	[503.22]VSWEFNM[221.06]
reset	6	ICAT_Mix.0963.0963.0	2	1813.2300	907.1200	22	3.53e+007	1.56e+007	1.26e+007	2.15e+006	1.08e+007	2	0.0000	7	[698.29]NEDGFLTI[336.60]
reset	7	ICAT_Mix.0989.0989.0	2	2319.9100	1160.4600	25	1.11e+007	6.63e+006	4.86e+006	6.75e+006	1.75e+007	2	0.0000	7	[923.27]IAAFNPAGI[767.31]
reset	8	ICAT_Mix.1064.1064.0	2	2261.1100	1131.0600	20	8.13e+006	1.93e+006	9.97e+005	3.15e+005	9.12e+006	2	0.0000	7	[719.60]GGYVFTGI[767.68]
reset	9	ICAT_Mix.1083.1090.2	2	2320.3500	1160.6800	25	3.94e+007	1.06e+007	4.31e+006	2.76e+007	7.69e+006	2	0.0000	7	[875.61]AELLSSIT[641.63]
reset	10	ICAT_Mix.1104.1104.0	2	2324.5700	1162.7900	25	4.58e+006	5.89e+006	3.44e+006	1.82e+006	7.80e+005	2	0.0000	7	[834.98]VAGAGNVIT[889.22]
reset	11	ICAT_Mix.1283.1288.0	2	1711.5300	856.2700	25	9.24e+006	6.46e+006	4.42e+006	5.33e+005	2.23e+006	2	0.0000	7	[497.23]MPHQPNL[271.95]
reset	12	ICAT_Mix.0699.0703.0	2	1055.0700	528.0400	25	2.12e+007	1.47e+006	5.15e+006	4.55e+006	2.90e+006	2	0.0000	6	[224.80]TNHAEAI[204.64]

Figure 78 ICAT intensities shown on the Spectrum Summary page

NOTE

The Spectrum Summary page shows the intensities of the parallel light/heavy EICs calculated during data extraction. Since you typically use the Spectrum Summary page to display spectra that have not been interpreted, you do not see the light/heavy ratios calculated in this table, nor should you attempt to calculate them from these data. Instead, use the calculations from the Protein/Peptide Summary page, as described earlier in this chapter.

Note that the values are reported as 0's for "metabolic" modifications, such as SILAC and $^{14}\text{N}/^{15}\text{N}$ mixes.



6

Processing MS-Only Data Interactively

Acquiring Agilent TOF Data for Use with the Spectrum Mill
Workbench 154

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks
Interactively 155

Single-Spectrum Processing 169

In this chapter you learn the basics to process multiple MALDI-TOF, ESI-TOF or Q-TOF MS-only data files with the Spectrum Mill workbench. This chapter guides you through a typical workflow with the software.

For Agilent TOF .wiff files, as well as other .wiff files, you use PMF search and the instructions in “[MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively](#)” on page 155.

PMF Search is limited to the analysis of digests from simple protein mixtures (usually three to five proteins). When there are peptides from too many different proteins in your spectrum, you do not achieve statistically significant scores because the more complex the protein mixture, the more non-matching (noise) peptides for any given protein. Use of the **Mixture scoring** option helps to overcome this limitation.

For Agilent TOF and Q-TOF .d data, you must copy masses from MassHunter Qualitative Analysis and paste them into Manual PMF search in order to use Spectrum Mill for that data type. See the instructions in “[Single-Spectrum Processing](#)” on page 169.



Acquiring Agilent TOF Data for Use with the Spectrum Mill Workbench

This section, as well as the next two, applies only to instruments that acquire .wiff data.

When you set up data acquisition parameters on the Agilent TOF (.wiff data only), it is critical that you establish analysis settings that are compatible with Spectrum Mill processing.

- For MALDI-TOF and infusion-ESI-TOF, make sure to acquire data only during the first period. The Spectrum Mill workbench processes data only from the first period.
- For LC-ESI-TOF, you may set up the analysis with multiple periods, but the Spectrum Mill workbench processes only the first period that contains profile data.
- For all Agilent TOF data, set the acquisition parameters such that the first experiment contains the data to be used with the Spectrum Mill workbench. The Data Extractor processes only Experiment 1 data.

For more details see “Acquisition and Extraction of Agilent TOF Data” in the online help file **Spectrum Mill Basics**.

NOTE

For help on any section of a page, click the blue bar at the top of the section.

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively

This section assumes you have already transferred your spectral files to the Spectrum Mill server and started the software. See the *Quick Start Guide* for details.

The instructions given in this section apply to Agilent TOF .wiff data, as well as to other non-Agilent .wiff files. They do not apply to Agilent TOF and Q-TOF .d files. See [“Single-Spectrum Processing”](#) on page 169 for processing instructions for Agilent TOF and Q-TOF .d files.

You create automated workflows from parameter files/tasks that you set up in each individual Spectrum Mill page. This section gives you the instructions for setting up the parameter files and for running the workflow tasks interactively. See Chapter 3 for instructions on creating automated workflows from the parameter files you set up here.

Step 1. Run the Data Extractor

This step applies only for Agilent ion trap MS-only data, Agilent TOF data (.wiff files), and Applied Biosystems/MDS Sciex QSTAR o-MALDI data. If you have MS-only data from another instrument and it is already in peak list format, skip to “[Step 2. Run database searches](#)” on page 162. The MS-only data extractor is customized to the data to be extracted.

For some data types, such as PDF-MALDI TOF data, the extractor averages and centroids the data from raw data files, but leaves signal-to-noise for PMF Search. For other data types, such as LC-ESI-TOF and infusion-ESI-TOF, the extractor does the averaging, centroiding and signal-to-noise filtering, but you can also filter by signal-to-noise in PMF Search.

NOTE

Statements about PDF-MALDI data apply to AP-MALDI data as well.

To run the MS-only Data Extractor for MALDI data

These instructions apply for:

- Agilent ion trap PDF-MALDI data
- Agilent TOF PDF-MALDI data
- Applied Biosystems/MDS Sciex QSTAR o-MALDI data.

- 1 Navigate to the Data Extractor page, as shown in [Figure 79](#).
- 2 Mark the check box for **Show only MS (PMF) parameters**.
- 3 If you intend to use an already existing parameters file, click **Load**.
- 4 In the Data Directory section, click the **Select...** button to select the folder that contains your files.
- 5 Click the **Choose...** button to select the modifications that match the chemistry for your samples. For details, click the blue bar labeled **Modifications** to access the online help.

NOTE

The **Mix Modifications** and associated ratio calculations are not supported for MS-only data.

- 6 Make sure that the check box for **Agilent ESI TOF data** is cleared.
- 7 Check that your page resembles the one in [Figure 79](#).
- 8 Click **Save As** to save the new parameters to the same or a new parameter file, enter the file name, and click **Save**.
- 9 Click the **Extract** button.

A request for execution is automatically placed in the queue.

Figure 79 MS-only Data Extractor filled out for MALDI data

To run the MS-only Data Extractor for Agilent LC-ESI-TOF data

- 1 Navigate to the Data Extractor page, as shown in [Figure 80](#) on page 159.
- 2 If this is the first time you have done this, read the background information that follows [Figure 80](#).
- 3 Mark the check box for **Show only MS (PMF) parameters**.
- 4 If you intend to use an already existing parameters file, click **Load**.
- 5 In the Data Directory section, click the **Select...** button to select the folder that contains your files.

- 6 Click the **Choose...** button to select the modifications that match the chemistry for your samples. For details, click the blue bar labeled **Modifications** to access the online help.

NOTE

The **Mix Modifications** and associated ratio calculations are not supported for MS-only data.

- 7 Set the **Extraction time range** to the range you want to extract.
- 8 Mark the check box for **Agilent ESI TOF data**.
- 9 In the **Min S/N** box, set the minimum signal-to-noise for which you want to extract peaks.
 - Try the default of 15 first.
 - If you think you did not identify all the proteins in the sample, first increase **Max peaks** in PMF Search and search again.
 - If you have increased **Max peaks** to include all the peaks that were extracted, then lower the **Min S/N** described here, re-extract, and search again with all the extracted peaks.
- 10 Click the **LC** option.
- 11 Set the **Time segment for averaging MS scans**.
 - If sequence coverage is not critical, use a segment value on the order of one-half a typical chromatographic peak width.
 - If coverage is critical, use a smaller value.
 - This value is used to sample the base peak chromatogram for mass spectral peaks in “bins” of the time segment width. The extraction time is proportional to this value.
 - If the segment value you specify results in fewer than four scans, the extractor selects a value that gives at least four scans, and outputs a warning in the results window.
- 12 For **Background subtraction**, select one of the options and fill in appropriate boxes:
 - **Select range from:** Generates a background peak list from the time range that you specify. It is important to select a region of the chromatogram with a relatively flat baseline.

- **Last __ min. of scan time range:** Generates a background peak list from the last minute(s) of the acquired data. (Be sure this area has a relatively flat baseline.)

NOTE

The time range for background subtraction does not need to be within the **Extraction time range**. For example, you can select 0 to 1 minute for background subtraction, but then extract from 10 to 40 minutes.

13 Check that your page resembles the one in [Figure 80](#).

14 Click **Save As** to save the new parameters to the same or a new parameter file, enter the file name, and click **Save**.

15 Click the **Extract** button.

A request for execution is automatically placed in the queue.

Figure 80 MS-only Data Extractor filled out for chromatographic data

For LC-ESI-TOF data, the extractor does the following:

- Generates a base peak chromatogram (BPC)

- Generates a background peak list by signal-to-noise filtering (but not de-isotoping) the specified background range of the BPC
- Uses the **time segment** value to sample the BPC in “bins” of that width. Each “bin's” MS peak list is background subtracted, peak-picked (signal-to-noise filtering, de-isotoping) and charge assigned. The resulting peaks are merged with the prior running accumulation of peaks.
- After the file has been processed, the results are written to a *.MIC (Mass-Intensity-Charge) file.

This extractor filters the spectra so that all peaks are of sufficient quality to be submitted to PMF Search.

To run the MS-only Data Extractor for Agilent infusion-ESI-TOF data

- 1 Navigate to the Data Extractor page, as shown in [Figure 81](#) on page 161.
- 2 If this is the first time you have done this, read the background information that follows [Figure 81](#).
- 3 Mark the check box for **Show only MS (PMF) parameters**.
- 4 If you intend to use an already existing parameters file, click **Load**.
- 5 In the Data Directory section, click the **Select...** button to select the folder that contains your files.
- 6 Click the **Choose...** button to select the modifications that match the chemistry for your samples. For details, click the blue bar labeled **Modifications** to access the online help.

NOTE

The **Mix Modifications** and associated ratio calculations are not supported for MS-only data.

- 7 Set the **Extraction time range** to the range you want to extract.
- 8 Mark the check box for **Agilent ESI TOF data**.
- 9 In the **Min S/N** box, set the minimum signal-to-noise for which you want to extract peaks. The default of 15 is a good choice.
- 10 Click the **Infusion** option.
- 11 Check that your page resembles the one in [Figure 81](#).

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively

12 Click **Save As** to save the new parameters to the same or a new parameter file, enter the file name, and click **Save**.

13 Click the **Extract** button.

A request for execution is automatically placed in the queue.

Figure 81 MS-only Data Extractor filled out for infusion-ESI-TOF data

The Data Extractor processes infusion-ESI-TOF data a little differently from LC-ESI-TOF data. Since infusion data has no chromatographic peaks, the extractor does not generate a base peak chromatogram. It does not subtract a background range; this is done in PMF Search. As with LC-ESI-TOF data, this extractor filters by signal-to-noise, detects mass spectral peaks, de-isotopes, and assigns charge. The results are written to a *.MIC file.

Step 2. Run database searches

- 1 Navigate to the PMF Search page, shown in [Figure 82](#) on page 163.
- 2 If you intend to use an already existing parameters file, click **Load**.
- 3 Select the data directory where your files reside.
- 4 If you previously executed a search with incorrect settings, mark the check box for **Remove all prior PMF results**.
- 5 If the sample contains peptides from a mixture of proteins, mark the check box for **Mixture scoring**.
- 6 Set the **Database**.
- 7 Click the **Choose...** button to select the modifications that match the chemistry for your samples. For details, click the blue bar labeled **Modifications** to access the online help.
- 8 Set **Instrument** to the instrument on which you acquired the data.

NOTE

If the instrument was not properly calibrated and mass shifts were larger than expected, then increase the **Peptide mass tolerance**.

- 9 Set the **Spectral features**.
 - If you have a mixture and the data was acquired with PDF-MALDI, mark the check box for **Override instrument defaults**, and make sure that **Max peaks** is set to **100**. For more complex samples, or for better sequence coverage, set **Max peaks** to a larger number.
 - If you have ESI-TOF data, mark the check box for **Override instrument defaults** and check that **Max peaks** is set to **500**. For more complex samples, or for better sequence coverage, set **Max peaks** to a larger number.
 - For PDF-MALDI data, check that the **Min S/N** is set to **2**. A lower S/N may result in search of matrix ions.
- 10 Under **Contaminant Masses**, select the appropriate file.
 - For PDF-MALDI analyses, select **Porcine Trypsin-Keratin**. This file includes trypsin autolysis products and keratin.
 - For infusion analyses on the Agilent TOF, select **ESI-Cal + Porcine Trypsin-Keratin**. This file includes the Agilent ESI TOF calibrants, as well as trypsin autolysis products and keratin.

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively

- For LC-ESI-TOF analyses, select **None** because the background subtraction was already done during data extraction.
 - To create a custom contaminant file, see the online help.
- 11 Set **Report Details** to reflect the number of hits and detailed results you wish to view.
 - 12 Set other parameters. Click the blue divider bars for more information in the online help.
 - 13 Click **Save As** to save the new parameters to the same or a new parameter file, enter the file name, and click **Save**.
 - 14 Click the **Start Search** button.

A request for execution automatically is placed in the queue.

Figure 82 PMF Search page

Step 3. Summarize PMF Search results

- 1 Navigate to the PMF Summary page.
- 2 If you intend to use an already existing parameters file, click **Load**.
- 3 Select the data directory where your files reside.
- 4 Set parameters similar to those shown in [Figure 83](#). In general, you do not need to change the defaults.
- 5 Click **Save As** to save the new parameters to the same or a new parameter file, enter the file name, and click **Save**.
- 6 Click the **Summarize** button.

To summarize results on the spot, make sure the **Request Queue** check box is clear before clicking this button.

To place the command in the queue, mark the check box.

- 7 Check that you see a display similar to that shown in [Figure 84](#).

NOTE

The setting for **Filter hits by score** is counterintuitive if you are used to MS/MS Search scores, for which larger numbers represent better results. With PMF Search scores, smaller numbers represent better results. A score of 0.5 means the match has a 50% chance of occurring randomly. A score of 1e-6 means the match has a one-in-a-million chance of occurring randomly. The probability distribution is calculated after counting the occurrences in the database of each mass submitted within the specified mass tolerance. Consequently, the score for the same set of masses changes if the mass tolerance, the enzyme, the number of missed cleavages, or the database changes. Also note that modified amino acids such as met-sulfoxide do not contribute to the score.

The screenshot shows the 'Summarize Results for Review' dialog box. It has three main panels. The left panel, 'Summarize Results for Review', contains buttons for 'Summarize' (highlighted in green), 'Save As...', and 'Load...', along with an unchecked 'Queue request' checkbox. The middle panel, 'Sorting', shows 'Filter hits by score' set to '<' and '0.001', and 'Sort by' set to 'Score'. The right panel, 'Review Fields', lists various fields with checkboxes: 'Filename' (checked), 'Score' (checked), 'MOWSE score' (unchecked), 'Mass error' (checked), 'Recalibration' (unchecked), 'Excel export' (unchecked), 'Protein MW' (checked), 'Protein pl' (checked), 'Species' (checked), 'Accession #' (checked), and 'Protein name' (checked). Below these panels, the 'Data Directory' section shows a 'Select...' button and the path 'ExampleData\Agilent\TOF\LC-ESI'. The 'Search result files:' section shows a list box with '*.spo' selected.

Figure 83 Settings to summarize PMF Search results

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively

Data Directory: msdata\SM\ExampleData\VPMF
Input Files: 25
Beginning to parse 25 output files.
25 output files parsed. Parsing complete. Filtered Files: 14

	Filename	Probability Score Dynamic	Probability Score	# (%) Masses Matched	Mean Mass Error	StdDev Mass Error	Mass Error Units	Coverage	Protein MW (Da)	Protein pI	Species	Accession #	Protein Name
1	80930109.2MI	7.79e-030	1.08e-016	19 / 25 (76%)	-2.4	3.1	ppm	42 %	46783.2	5.67	YEAST	P00925	1 Enolase 2 (EC 4.2.1.11) (2-phosphoglycerate dehydr
2	80930108.2MI	4.7e-029	8.4e-017	19 / 25 (76%)	-6.9	3.6	ppm	42 %	46783.2	5.67	YEAST	P00925	1 Enolase 2 (EC 4.2.1.11) (2-phosphoglycerate dehydr
3	80930107.2MI	6.09e-029	8.85e-017	19 / 25 (76%)	-0.6	4.9	ppm	42 %	46783.2	5.67	YEAST	P00925	1 Enolase 2 (EC 4.2.1.11) (2-phosphoglycerate dehydr
4	80930106.2MI	9.72e-020	6.44e-011	16 / 25 (64%)	1.8	5.5	ppm	48 %	44607.4	7.09	YEAST	P00560	1 Phosphoglycerate kinase (EC 2.7.2.3).
5	80930113.2MI	9.29e-019	5.92e-013	14 / 25 (56%)	-3.5	8.7	ppm	44 %	26664.4	5.75	YEAST	P00942	1 Triosephosphate isomerase (EC 5.3.1.1) (TIM).
6	80930103.2MI	1.37e-017	4.51e-013	14 / 25 (56%)	-2.9	7.8	ppm	46 %	39489.8	5.51	YEAST	P14540	1 Fructose-bisphosphate aldolase (EC 4.1.2.13).
7	80930114.2MI	2.23e-014	2.75e-009	12 / 25 (48%)	-3.7	7.5	ppm	43 %	26664.4	5.75	YEAST	P00942	1 Triosephosphate isomerase (EC 5.3.1.1) (TIM).
8	80930112.2MI	2.27e-014	3.24e-009	14 / 25 (56%)	-2.0	7.6	ppm	31 %	35615.7	6.49	YEAST	P00359	1 Glyceraldehyde 3-phosphate dehydrogenase 3 (EC 1.2
9	80930104.2MI	4.86e-014	1.15e-010	12 / 21 (57%)	-3.4	7.1	ppm	41 %	39489.8	5.51	YEAST	P14540	1 Fructose-bisphosphate aldolase (EC 4.1.2.13).
10	80930116.2MI	8.68e-014	2.66e-009	12 / 25 (48%)	-5.6	12.6	ppm	42 %	26664.4	5.75	YEAST	P00942	1 Triosephosphate isomerase (EC 5.3.1.1) (TIM).
11	80930126.2MI	1.58e-011	8.26e-007	11 / 23 (47%)	-3.9	13.0	ppm	33 %	29960.3	4.82	YEAST	P29311	1 BMH1 protein.

Figure 84 PMF Search Summary table

Step 4. Manually review results

- 1 Manually review some or all of the results. First examine the overall results, as shown in Figure 84. Note the following terminology:
 - **Probability Score Dynamic** - database search score based on the peptide mass tolerance determined by the Spectrum Mill software from the actual data
 - **Probability Score** - database search score based on the **Peptide mass tolerance** you set in PMF Search
- 2 To view individual PMF Search results, as shown in Figure 85, click a link under the **Filename** header and then scroll up. Scroll down or click links under the **Rank** heading shown in Figure 85 to see more search details. You may want to resize the page section to better view these results.

NOTE

If this were ESI-TOF data, there could be duplicate PMF matches from singly- and multiply-charged peptides. In this case, only one match would be scored, but all would be displayed. The scored match would be the one for which **m/z submitted** (after adjusting for charge) was closest to **MH⁺ matched**. In the **Score Counted** column shown in Figure 85, scored results would show a value of **1**, while non-scored results would show a value of **0**.

6 Processing MS-Only Data Interactively

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively

PMF Search Results

Sample ID (comment): 80930109.2MI
Database searched: SwissProt
Molecular weight search (1000 - 150000 Da) selects 116917 entries.
Full pI range: 119029 entries.
Combined molecular weight and pI searches select 116917 entries.
PMF search selects 5 entries.

[Parameters used in Search](#)

Rank	Dynamic Probability Score	Static Probability Score	# (%) Masses Matched	Mass Error Mean (Std Dev) (ppm)	Protein Coverage	Protein MW (Da)/pI	Species	Accession #	Protein Name
1	7.79e-030	1.08e-016	19/25 (76%)	-2.4 (3.1)	42%	46783.2/5.67	YEAST	P00925	Enolase 2 (EC 4.2.1.11) (2-phosphoglycerate dehydr
2	2.05e-008	0.000586	10/25 (40%)	-5.3 (7.3)	26%	46671.1/6.17	YEAST	P00924	Enolase 1 (EC 4.2.1.11) (2-phosphoglycerate dehydr
3	0.0678	5.84	5/25 (20%)	-39.1 (13.7)	9%	21851.2/5.29	MYCLE	P54884	Sensory transduction protein regX3.
4	0.15	63.4	5/25 (20%)	-10.0 (6.9)	8%	47518.9/5.25	CLAHE	P42040	Enolase (EC 4.2.1.11) (2-phosphoglycerate dehydrat
5	0.886	0.886	5/25 (20%)	-14.7 (58.0)	32%	13646.9/9.73	THEAC	Q9HLA6	30S ribosomal protein S6e.

Detailed Results

1. 19/25 matches (76%). YEAST. Enolase 2 (EC 4.2.1.11) (2-phosphoglycerate dehydr (46783.2 Da) pI = 5.67

m/z submitted	MH ⁺ matched	Delta ppm	Score Counted	Tolerance Bin Index (9-10)	start	end	Peptide Sequence (counted in score)	Modifications
726.3360	726.3422	-8.6	1	9	9	14	(R)SVYDSR(G)	
756.4690	756.4732	-5.5	1	9	409	414	(K)LNQLLR(I)	
800.3830	800.3830	-0.1	1	9	258	263	(K)YDLDFK(N)	
825.3930	825.3929	0.1	1	10	178	184	(K)TFAEAMR(I)	
841.3850	841.3878	-3.3	1	9	178	184	(K)TFAEAMR(I)	1Met-ox
1159.6090	1159.6111	-1.9	1	9	185	194	(R)IGSEVYHNLK(S)	

Figure 85 PMF Search results

- To continue examining data, click additional links shown in the PMF Search results in [Figure 84](#) and [Figure 85](#):
 - To view information in the database you searched, click a link under **Accession #**.
 - To view amino acid coverage, as shown in [Figure 86](#), click a link under the **Coverage** header. In the coverage map:
 - Red = matched
 - Black = not matched
 - Blue = modified amino acid (for example, cysteine)
 - Green = consensus N-linked glycosylation site

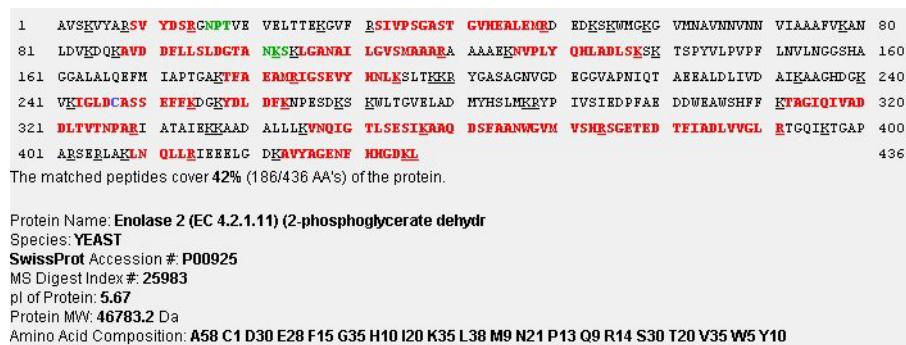


Figure 86 Amino acid coverage from PMF Search

Step 5. Summarize and print results

After you have reviewed results, you summarize and print them.

- 1 To summarize results, click the **Summary Settings** button to display the summary input form at the bottom of the page. Then repeat the procedure in “[Step 3. Summarize PMF Search results](#)” on page 164.
- 2 To import data into Excel or LIMS:
 - a Mark the **Excel Export** check box.
 - b Click the **Summarize** button.
 - c Check that you see a display with two buttons, as in [Figure 87](#). At this point, a new file with extension ***.ssv** has been created. You can import this file into Excel or upload it to a LIMS system (if configured by your system administrator), or display it on your screen.
 - d Do one of the following:
 - To import the data into Excel, import as semicolon-delimited data.
 - To upload to LIMS, see the Server Administration online help.

6 Processing MS-Only Data Interactively

MS-Only Workflow – Setting Up Parameter Files and Processing Tasks Interactively



Figure 87 Display when you mark the Excel export check box

- 3 To print results, use your browser's print function:
 - a Click **File > Page Setup**, and mark the check boxes for **Print background colors and images** and for **Landscape**.
 - b Click **OK**.
 - c Click in the results frame.
 - d Select **File > Print Preview**.
 - e At the top of the Print Preview window, select **Only the selected frame**.
 - f Click the **Print...** button.
- 4 If you wish to further process your data, search against a different database and create a final results summary.

NOTE

To export results into HTML or PowerPoint files, see the Frequently Asked Questions section of the online help.

Single-Spectrum Processing

To process MS-only data for Agilent TOF and Q-TOF .d files, you must use the instructions in this section.

See Chapter 5 of the *Familiarization Guide* to learn how to process a single spectrum for a specific Agilent Q-TOF .d file.

Step 1. Find the compounds of interest in Mass Hunter Qualitative Analysis

The MFE settings in MassHunter Qualitative Analysis may need to be optimized for a given sample set or analytical condition.

- 1 Open **MassHunter Qualitative Analysis**.
- 2 Under the **Data** folder, select an Agilent TOF or Q-TOF .d file, and click **Open**.
- 3 In Method Explorer, click **Find by Molecular Feature**.
- 4 Under the **Extraction** tab and from the **Target data type** list, select **small molecules (chr)**.
- 5 Click **Ion Species** and clear all check boxes except for **+H** and **-H**.
- 6 Click **Charge State** and from the **Isotope Model** list, select **peptides**.
- 7 Mark **Limit assigned charge states to a maximum of** and type 7.
- 8 Click **Compound Filters**, and clear **Relative height**, then specify an **Absolute height** estimated from an examination of the 5-10 compounds of lowest height/abundance.
- 9 Click **Results** and clear all check boxes in the tab except for the first one.
- 10 Click **Find Compounds by Molecular Feature**.

The compound list should resemble the one in [Figure 88](#).

6 Processing MS-Only Data Interactively

Single-Spectrum Processing

The screenshot shows the Agilent MassHunter Qualitative Analysis 8.04.00 - Default.m interface. The 'Compound List' table is displayed with the following data:

Show/Hide	Cpd	Label	Mass	Avg Mass	Std Dev	Base Peak	m/z	Polarity
☒	1	Cpd 1: 1.917	648.3247	648.5725		325.1694	325.1694	Posit
☒	2	Cpd 2: 1.970	648.3236	648.7793	0.0051	325.1692	325.1692	Posit
☒	3	Cpd 3: 2.003	648.3242	648.8177		325.1694	325.1694	Posit
☒	4	Cpd 4: 2.185	858.3254	858.6022		430.1703	430.1703	Posit
☒	5	Cpd 5: 2.187	1406.5426	1407.4199	0.0008	469.855	469.855	Posit
☒	6	Cpd 6: 2.212	1149.4439	1150.2097		575.7292	575.7292	Posit
☒	7	Cpd 7: 2.227	1464.5494	1465.5009	0.0026	489.1914	489.1914	Posit
☒	8	Cpd 8: 2.228	658.3713	658.6258		330.1929	330.1929	Posit
☒	9	Cpd 9: 2.271	422.0193	422.2034		423.0262	423.0262	Posit
☒	10	Cpd 10: 2.288	404.0075	404.158		405.0137	405.0137	Posit

Figure 88 MassHunter compound list

Step 2. Copy and paste compound masses into PMF Search

In this step you copy only the masses from all the compound features and paste them into the Manual PMF Search page.

- 1 Right-click the **Masses** heading, and select **Copy Column to Clipboard** > **Using Newline separator**.
- 2 Open the **Spectrum Mill** software.
- 3 Navigate to the **PMF Search** page from the Spectrum Mill home page.
- 4 Click the **Manual PMF** button at the top of the page.
- 5 Paste your mass list into the **Peptide Masses** box.

See [Figure 89](#) on page 172.

Step 3. Search the MS-only data and display results

In this step you enter the remaining parameters and start a search.

- 1 If your masses represent protonated species, click **MH⁺**. If your masses represent neutral species, click **M**.
- 2 If necessary, type charges for the masses in the **Peptide Masses** box.
 - If you clicked **MH⁺**, the software assumes a charge of +1.
 - If you clicked **M**, the software assumes a charge of 0.
 - A charge that you type into the box overrides the **MH⁺/M** settings.
- 3 For **Masses are**:
 - You generally select **Monoisotopic**. With **Monoisotopic**, for H⁺, the mass calculation uses the mass of a proton rather than the mass of a hydrogen atom.
 - Select **Monoisotopic no e- correction** if your data is from an instrument such as MALDI QSTAR where the mass calibration is done in such a way that there is no correction for the loss of the electron for protonated species.
- 4 Set the other parameters. Click the blue divider bars for more information in the online help.
- 5 Click **Start Search**.
- 6 When the search is complete, scroll down and check that you see PMF Search Results, as shown in [Figure 90](#) on page 173.

6 Processing MS-Only Data Interactively

Single-Spectrum Processing

Agilent Spectrum Mill - Manual PMF Search

Spectrum Mill | **MS Edman** | **MS Digest** | **MS Isotope** | **Help**

Search

Start Search Sample ID:

Maximum reported hits: ☐ MOWSE scores - Pfactor: ☐ Mixture scoring

Search Parameters

Min. # peptides required to match:

Database:

Species:

MW of protein: from Da to Da ☐ All

Protein pI: from to ☒ All

Digest:

Maximum # of missed cleavages:

Modifications

Choose...

Fixed: Carboxymethylation (C)

Variable: Oxidized methionine (M)
Pyroglutamic acid (N-termQ)

Peptide Masses

Mass tolerance: +/- ppm

Masses are:

☐ MH⁺ ☒ M

Mass (m/z)	Charge (z)
648.3247	
648.3236	
648.3242	
858.3254	
1406.5426	
1149.4439	
1464.5494	
658.3713	
422.0193	
404.0075	
422.019	
404.0072	
422.0182	
422.0184	
1616.5428	
846.4944	
1072.4838	
885.4068	
664.3672	
664.3676	
1051.3914	
536.2729	
973.4474	

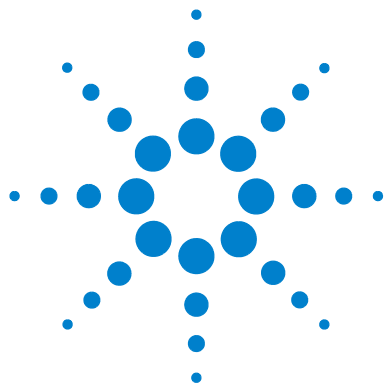
Figure 89 Manual PMF Search page

PMF Search Results									
Sample ID (comment): Enter_Comment Database searched: NCBIInr.stdmix Parameters used in Search Molecular weight search (1000 - 100000 Da) selects 17 entries. Full pl range: 18 entries. Combined molecular weight and pl searches select 17 entries. PMF search selects 1 entry.									
Rank	Dynamic Probability Score	Static Probability Score	# (%) Masses Matched	Mass Error Mean (Std Dev) (ppm)	Protein Coverage	Protein MW (Da)/pl	Species	Accession #	Protein Name
1	0	0	50/90 (55%)	-1.0 (2.8)	59%	71323.5/5.82	bovine	1351907	bovine serum albumin
Detailed Results									
1. 50/90 matches (55%). bovine. bovine serum albumin (71323.5 Da) pl = 5.82									
m/z submitted	MH ⁺ matched	Delta ppm	Score Counted	Tolerance Bin Index (0-0)	start	end	Peptide Sequence	Modifications	
537.2802	537.2820	-3.4	1	0	157	160	(K)FWGK(Y)		
649.3315	649.3338	-3.5	1	0	205	209	(K)IETMR(E)		
649.3309	649.3338	-4.5	1	0	205	209	(K)IETMR(E)		
649.3320	649.3338	-2.8	1	0	205	209	(K)IETMR(E)		
665.3745	665.3770	-3.7	1	0	156	160	(K)KFWGK(Y)		
665.3749	665.3770	-3.1	1	0	156	160	(K)KFWGK(Y)		
689.3712	689.3729	-2.6	1	0	236	241	(K)AWSVAR(L)		
759.4058	759.4069	-1.5	1	0	198	204	(K)GACLLPK(I)		
759.4052	759.4069	-2.3	1	0	198	204	(K)GACLLPK(I)		
759.4048	759.4069	-2.8	1	0	198	204	(K)GACLLPK(I)		
789.4697	789.4716	-2.5	1	0	257	263	(K)LVTDLT(K)		
789.4719	789.4716	0.3	1	1	257	263	(K)LVTDLT(K)		
847.5017	847.5036	-2.3	1	0	242	248	(R)LSQKFPK(A)		
886.4141	886.4153	-1.3	1	0	131	138	(K)DDSPDLPK(L)		
899.4629	899.4655	-2.9	1	0	483	489	(R)LCVLHEK(T)		
922.4870	922.4880	-1.1	1	0	249	256	(K)AEFVEVTK(L)		
927.4934	927.4934	-0.1	1	0	161	167	(K)YLYEIAR(R)		
974.4547	974.4578	-3.2	1	0	37	44	(K)DLGEEHFK(G)		
997.5963	997.5928	3.5	1	1	549	557	(K)QTALVELLK(H)	pyroGlu	
1002.5826	1002.5830	-0.4	1	0	598	607	(K)LVVSTQTALA(-)		

Figure 90 Manual PMF Search results

Step 4. Examine the results

- To see more search details, scroll down or click links under the **Rank** heading shown in [Figure 90](#). Note that you see detailed results for only the top two hits.
- To view information in the database you searched, click a link under the **Accession #** header.
- To view amino acid coverage, click a link under the **Protein Coverage** header. In the coverage map:
 - Red = matched
 - Black = not matched
 - Blue = modified amino acid, e.g., cysteine
 - Green = consensus N-linked glycosylation site



7 Using the Tool Belt

- To terminate a process [176](#)
- To create a saved results file so you can search previous hits [178](#)
- To create an MS/MS Search summary file if search terminated abnormally [180](#)
- To create a summary table of previously-used parameters [181](#)
- To create a summary table of MS/MS identification statistics [183](#)
- To report FDR and search statistics [186](#)
- To copy spectra to your collections directory [190](#)
- To list details about amino acid modifications [191](#)
- To create a file of iTRAQ correction factors [193](#)
- To apply a file of iTRAQ correction factors [194](#)
- To calculate discriminant scoring coefficients [195](#)
- To export PepXML files [197](#)
- To convert spectra [199](#)
- To archive data [200](#)

This chapter describes helpful tools that you can use when you process data with the Spectrum Mill workbench. These tools help you accomplish a range of tasks, from terminating processes to summarizing identification statistics. You access these tools from the Tool Belt page.

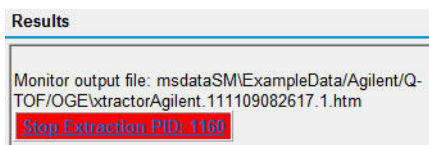


To terminate a process

The Spectrum Mill workbench performs functions like Data Extractor, MS/MS Search, and Sherenga *de novo* Sequencing by launching a parent script, which then repetitively launches child programs. The **Stop** button for a program aborts the parent script and prevents new child programs from starting. However, it does not abort the child program that is currently running. The child program runs until its normal conclusion unless it is aborted separately through the **Stop Process** button on the Tool Belt page.

To use the Stop button on the main program page

Processes (such as Data Extractor, MS/MS Search, and Sherenga *de novo* Sequencing) that take longer than a few seconds have a red **Stop** process link that is displayed in the **Results** window where the process runs.



- 1 To abort a process, click the **Stop** process link.
- 2 Wait a few seconds until a message is displayed to indicate that the process has stopped.
- 3 Stop child programs with the **Stop Process** button on the Tool Belt page. See the next section.

To use the Stop Process button on the Tool Belt page

The **Stop** button described above terminates the parent script, but not child programs. Go to the Tool Belt to terminate these child programs:

- 1 Navigate to the Tool Belt page and select the **Stop process** option, as shown in Figure 91. (This is the default.)
- 2 Click the **List Processes** button to list Spectrum Mill processes that are currently running on the server.
- 3 Select the processes you wish to stop.
- 4 Click the **Stop Process** button.
- 5 Click **List Processes** again to make sure the stopped processes no longer appear in the list.

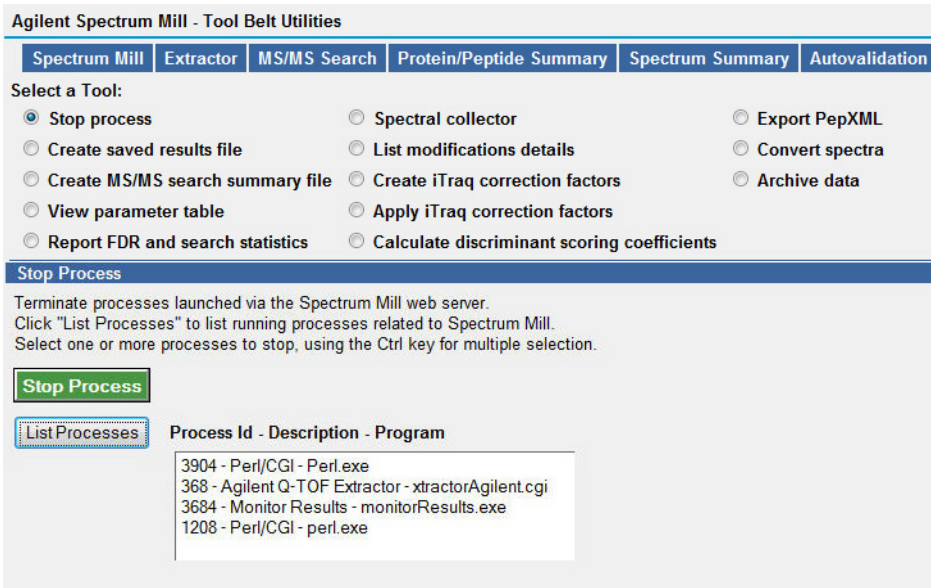


Figure 91 Stop process

NOTE

It is also possible to terminate parent processes from the Tool Belt page. This may be more convenient if you have navigated to another Spectrum Mill page and you no longer see the red **Stop** process link. Use the same instructions as for child processes.

CAUTION

Avoid terminating processes initiated from other clients. You may be able to identify your processes by name. For example, the data extractor for Agilent Q-TOF files is called **xtractorAgilent.cgi**, while the data extractor for Agilent ion trap files is called **xtractorAgilentTrap.cgi**.

To create a saved results file so you can search previous hits

The Spectrum Mill workbench allows you to save valid hits (protein identifications) and then search them again. For example, you can save a set of hits from an identity mode MS/MS Search and then search this set with a variable modifications or homology mode MS/MS Search. This saves time because the modified peptides you find in variable modifications or homology modes are often derived from the proteins you already identified in identity mode. The computation-intensive homology mode search is much faster when you search a smaller list.

When you create a saved results file, Tool Belt creates a ***.res** file (list of accession numbers) of valid hits. When you search this list, MS/MS Search reads this ***.res** file and gets entries from the database file. This requires you to both save and use the ***.res** file from the same data directory.

To save hits from MS/MS Search:

- 1 Make sure you already have a set of search results you wish to save. The saved results file will contain only the search results for which you designated the protein identifications as “valid.”
- 2 Navigate to the Tool Belt page and click the option to **Create saved results file**, as shown in [Figure 92](#).
- 3 Select the **Database** you previously searched.
- 4 Select the **Data Directory** where the search results reside. This is the same data directory where the data files reside.
- 5 Click the **Create File** button.

NOTE

If you mark **Search previous hits** in the MS/MS Search page, it automatically creates a saved results file if it doesn't exist.

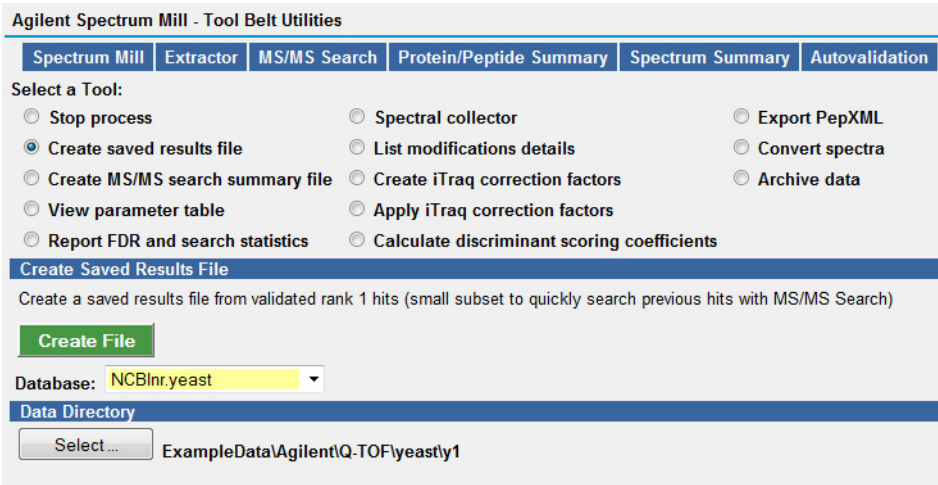


Figure 92 Create saved results file

To create an MS/MS Search summary file if search terminated abnormally

If your MS/MS Search terminated abnormally (e.g., because you aborted it), you can still create an MS/MS Search Summary File so that you can review partial results on the Protein/Peptide Summary page. To create this file:

- 1 Navigate to the Tool Belt page and click the option to **Create MS/MS search summary file**, as shown in [Figure 93](#).

Agilent Spectrum Mill - Tool Belt Utilities

Spectrum Mill	Extractor	MS/MS Search	Protein/Peptide Summary	Spectrum Summary	Autovalidation
---------------	-----------	--------------	-------------------------	------------------	----------------

Select a Tool:

- ☐ Stop process
- ☐ Create saved results file
- ☒ **Create MS/MS search summary file**
- ☐ View parameter table
- ☐ Report FDR and search statistics
- ☐ Spectral collector
- ☐ List modifications details
- ☐ Create iTRAQ correction factors
- ☐ Apply iTRAQ correction factors
- ☐ Calculate discriminant scoring coefficients
- ☐ Export PepXML
- ☐ Convert spectra
- ☐ Archive data

Create MS/MS Search Summary File

Create MS/MS search summary file of rank 1 hits
(to use with Protein/Peptide Summary if MS/MS Search terminated abnormally)

Create File

Data Directory

Select ... ExampleData\Agilent\Q-TOF\yeasty1

Figure 93 Create MS/MS search summary file

- 2 Select the **Data Directory** for which you want to create a summary file.
- 3 Click the **Create File** button.

To create a summary table of previously-used parameters

The Spectrum Mill workbench keeps a cumulative record of the parameters you set in programs such as Data Extractor, MS/MS Search, and Sherenga *de novo* Sequencing. This option can be helpful to diagnose problems because the settings are the internal Spectrum Mill variables, not those displayed in the Browser.

To view previously-used parameters:

- 1 Navigate to the Tool Belt page and click the **View parameter table** option, as shown in [Figure 94](#).

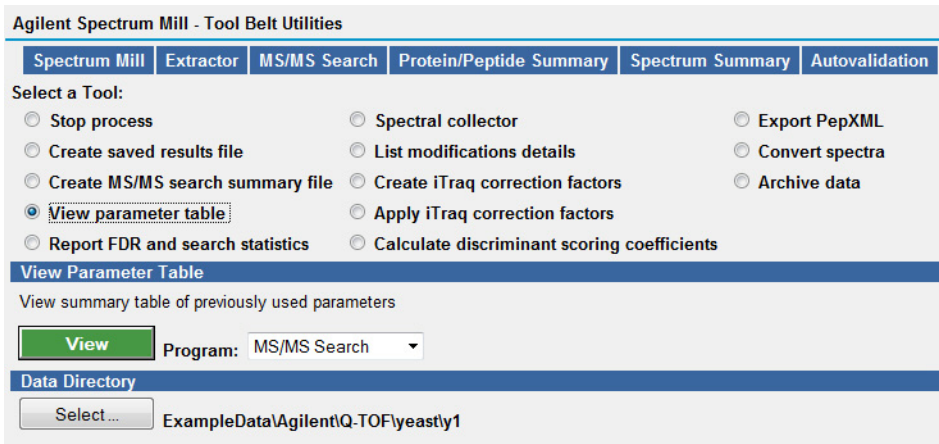


Figure 94 View parameter table

- 2 Select the **Program** for which you want to summarize parameters.
- 3 Select the **Data Directory** for which you want to summarize parameters.
- 4 Click the **View** button.
- 5 Check that you see a report like that shown in [Figure 95](#). In the table, a dash indicates that the setting was the same as used previously.

Spectrum Mill - Parameter Table		
	mstag.1	mstag.2
batchSize	81	-
composition_exclude	-	-
composition_include	-	-
dataDir	ExampleData\Agilent\Trap\ecoli\heatshock	-
database	NCBI\nc_ecoli	-
dna_frame_translation	6	-
enable_multiple_mods	1	-
enzyme	Trypsin	-
filterByFeature	maxSequenceTagLength	-
fixedMods	carbamidomethylation	-
fragment_mass_tolerance	0.7	-
full_mw_range	1	-
full_pi_range	-	-
hide_html_links	-	-
high_pi	10.0	-
homology_gap_mode	-	-
input_filename	NCBI\nc_ecoli	-
input_program_name	mstag	-
instrument_name	ESI-ION-TRAP-AGILENT	-
instrument_nameMSMS	-	-
low_pi	3.0	-
maxSequenceTagLengthFilter	3	-
maxSequenceTagLengthFilterDirection	>	-
max_ms_prod_charge	3	-
max_num_unused_ions	50	-
max_reported_hits	5	-
minMatchedPercent	50	-
min_ions_to_search	4	-
missed_cleavages	2	-
mparams_dir	mparams_mill/	-
mutation_matrix_off	-	-
parent_mass_convert	monoisotopic	-
parent_mass_tolerance	2.5	-
parent_shift	130.0	-
parent_shift_type	+/-	-
proton_mobility_scoring	1	-
requestScript	batchTagPara.pl	-
res_mstag_dir	F:\SpectrumMill\msdata\SM\ExampleData\Agilent\Trap\ecoli\heatshock\results_mstag	-
results_from_file	-	1
results_to_file	-	-
search_type	identity	variable

Figure 95 Parameter Table that shows settings used for an identity mode search and a subsequent variable modifications search

To create a summary table of MS/MS identification statistics

The Spectrum Mill workbench compiles identification statistics, such as the percentage of MS/MS spectra that have been interpreted for a set of data files. This is the same set of statistics and report that you accessed previously using the View MS/MS search statistics option but now produce using the Report FDR and search statistics option.

To create the summary table for basic search statistics:

- 1 Navigate to the Tool Belt page and click the **Report FDR and search statistics** option, as shown in [Figure 96](#).

The screenshot shows the 'Agilent Spectrum Mill - Tool Belt Utilities' window. The 'Spectrum Mill' tab is active. Under 'Select a Tool:', the 'Report FDR and search statistics' option is selected. Below this, the 'Report FDR and Search Statistics' section is visible, which includes a 'Report' button and various checkboxes for statistics to include in the report. The 'Data Directories' section at the bottom shows '(none selected)'.

Agilent Spectrum Mill - Tool Belt Utilities

Spectrum Mill | Extractor | MS/MS Search | Protein/Peptide Summary | Spectrum Summary | Autovalidation | PMF Search | Workflows | Help

Select a Tool:

- ☐ Stop process
- ☐ Create saved results file
- ☐ Create MS/MS search summary file
- ☐ View parameter table
- ☒ Report FDR and search statistics
- ☐ Spectral collector
- ☐ List modifications details
- ☐ Create iTRAQ correction factors
- ☐ Apply iTRAQ correction factors
- ☐ Calculate discriminant scoring coefficients
- ☐ Export PepXML
- ☐ Convert spectra
- ☐ Archive data

Report FDR and Search Statistics

Creates table of statistics using search results: FDR, mass accuracy, pI, and chromatography

Report

- ☐ FDR at the peptide & spectra level (from valid hits)
- ☐ FDR at the protein level
- ☐ Group proteins across all directories
- Grouping method: 1 shared, expand subgroups
- ☐ Fragmentation Mode
- ☐ Precursor mass error mean (ppm)
- ☐ Precursor charge count (from valid spectra)
- ☐ Peptide pI median for each run
- ☐ Chromatography stats for each run
- Middle RT portion of validated spectra in LC run: 80 (%)
- ☐ Distinct Peptide Fraction Overlap Table
- Same sequences with different mods are: Distinct
- ☐ Precursor Acquisition Uncertainty: m/z and z.
- ☐ MS1 intensity stats for each run.

Data Directories

Select ...

(none selected)

Figure 96 Report FDR and search statistics option in Tool Belt - default view

- 2 Do NOT mark any of the check boxes.
You see MS/MS identification statistics by just clicking the Report button without marking any other check boxes. See [“To report FDR and search statistics”](#) on page 186 to learn how to report FDR calculations and other statistics in addition to the basic set described here.
- 3 Select the **Data Directories** for which you want to summarize statistics.
- 4 Click **Report**.
- 5 Check that you see a report like that shown in [Figure 97](#).

Data Directory: ExampleData/Agilent/Q-TOF/yeast/y1							
Data Directory: ExampleData/Agilent/Q-TOF/yeast/y2							
File/Directory	Raw Files	MS/MS spectra collected	MS/MS spectra merged	MS/MS spectra filtered	MS/MS spectra valid	Collection Yield V/C (%)	Validation Yield V/F (%)
yeast-1	1	19276	17632	15307	4190	21.7	27.4
ExampleData/Agilent/Q-TOF/yeast/y1	1	19276	17632	15307	4190	21.7	27.4
yeast-2	1	22102	19599	16740	4595	20.8	27.4
ExampleData/Agilent/Q-TOF/yeast/y2	1	22102	19599	16740	4595	20.8	27.4

Directory	Raw Files	MS/MS spectra collected	MS/MS spectra merged	MS/MS spectra filtered	MS/MS spectra valid	Collection Yield V/C (%)	Validation Yield V/F (%)
ExampleData/Agilent/Q-TOF/yeast/y1	1	19276	17632	15307	4190	21.7	27.4
ExampleData/Agilent/Q-TOF/yeast/y2	1	22102	19599	16740	4595	20.8	27.4
all	2	41378	37231	32047	8785	21.2	27.4

Figure 97 View search statistics report

Note that under File/Directory the stats are given for the file first, then the sum of the files under the directory (in this case, there is only one file under each of two directories) and that under Directory appear the stats for each directory and then the sum of the directories.

See below for explanations of some of the headers.

MS/MS spectra collected	Number of MS/MS spectra in the raw data file
MS/MS spectra merged	Number of MS/MS spectra that result from merging by the Data Extractor
MS/MS spectra filtered	Number of MS/MS spectra exported by the Data Extractor program after filtering by spectral quality
Collection Yield (%)	Number of MS/MS spectra interpreted and validated divided by number of MS/MS spectra collected, expressed as percentage
Interpretation Yield (%)	Number of MS/MS spectra interpreted and validated divided by number of MS/MS spectra filtered, expressed as percentage

It is typical that not all spectra are interpreted. The obvious reason is that some spectra could not be matched to database entries, but there are other reasons as well. For example, to avoid missing any good spectra, you

typically set Data Extractor to remove most, but not all, of the noisy spectra. The remaining noisy spectra are not peptides and are not interpreted.

If your **Collection Yield** seems particularly low, there may have been an unusually high number of noisy spectra in your analysis. Perhaps you used a low threshold for data acquisition, or maybe there was a high instrument background. In these cases, the relative number of spectra that are picked by the Data Extractor is low.

Both the **Collection Yield** and the **Interpretation Yield** reflect to some degree how much time you spent processing the data, via variable modifications searches, broader databases, etc. In general, processing is complete when sufficient information has been extracted from the data to meet the experimental goals.

To report FDR and search statistics

With the Tool Belt you can now report additional metrics in addition to the basic search statistics already calculated by the program. (See “To create a summary table of MS/MS identification statistics” on page 183.)

- 1 Navigate to the Tool Belt page and click the **Report FDR and search statistics** option, as shown in Figure 96.
- 2 Mark each check box whose calculation or statistic you want the program to report.

Click the Report FDR and Search Statistics blue bar to see explanations for each option and why you would want to see that calculation or statistic.

Agilent Spectrum Mill - Tool Belt Utilities

Spectrum Mill

Extractor

MS/MS Search

Protein/Peptide Summary

Spectrum Summary

AutovValidation

PMF Search

Workflows

Help

Select a Tool:

☐ Stop process
 ☐ Create saved results file
 ☐ Create MS/MS search summary file
 ☐ View parameter table
 ☒ Report FDR and search statistics

☐ Spectral collector
 ☐ List modifications details
 ☐ Create iTRAQ correction factors
 ☐ Apply iTRAQ correction factors
 ☐ Calculate discriminant scoring coefficients

☐ Export PepXML
 ☐ Convert spectra
 ☐ Archive data

Report FDR and Search Statistics

Creates table of statistics using search results: FDR, mass accuracy, pl, and chromatography

Report

☐ FDR at the peptide & spectra level (from valid hits)
 ☒ FDR at the protein level
 ☐ Group proteins across all directories

Grouping method: 1 shared, expand subgroups

☐ Fragmentation Mode

☐ Precursor mass error mean (ppm)
 ☒ Precursor charge count (from valid spectra)
 ☐ Peptide pl median for each run
 ☒ Chromatography stats for each run

Middle RT portion of validated spectra in LC run: 80 (%)

☐ Distinct Peptide Fraction Overlap Table

Same sequences with different mods are: Distinct

☐ Precursor Acquisition Uncertainty: m/z and z.
 ☐ MS1 intensity stats for each run.

Data Directories

Select...

☒ ExampleData/Agilent/Q-TOF/yeast/y1

Figure 98 Report FDR and search statistics option in Tool Belt - check boxes marked

- 3 Select the **Data Directories** for which you want to report the FDR or statistics.

The Tool Belt should now look like Figure 98.

- 4 Click **Report**.

The first part of the report you see will be like that shown in [Figure 97](#). Added to that report are columns containing the statistics you chose to have the program report. Refer to [Table 6](#) to view the column headers that appear for each option.

Table 6 Results After Reporting FDR and Search Statistics

Report this statistic:	To view these report column headers:	Because you want to know:
FDR at the peptide & spectra level (from valid hits)	FDR Spectra (%) # Spectra False FDR Distinct Peptide (%) # Distinct Peptides # Distinct Peptides False	The FDR at both the spectra and peptide levels. The distinct peptide FDR can only be reported with the Tool Belt. The spectral level FDR may be seen in other places depending on the autovalidation strategy.
FDR at the protein level	All of the above, plus: FDR Distinct Protein (%) # Protein Groups # Distinct Proteins # Distinct Proteins False	The FDR at the spectra, peptide and proteins levels. This can only be reported with the Tool Belt.
Precursor mass error mean (ppm)	Mean Precursor Mass Error (ppm) Std Dev Precursor Mass Error (ppm) Mean is the average of all the values. Standard deviation measures the spread in the values away from the mean.	The ppm difference between the experimentally measured mass and the theoretical mass calculated from the amino acid sequence. A ppm error larger than expected can help identify instrumentation or data quality issues.
Precursor charge count (from valid spectra)	# spectra z2-z7 z2 (%) - z7 (%)	The number of spectra for each charge state from +2 to +7 and the percentage of spectra in each charge state from +2 to +7
Peptide pI median for each run	# Distinct Peptides pI Median Peptide pI Std Dev Peptide pI	Distribution of peptide pI in each run; typically used when pI-based fractionation has been done.

Table 6 Results After Reporting FDR and Search Statistics

Report this statistic:	To view these report column headers:	Because you want to know:
Chromatography stats for each run	Start time mid xx% matched spectra in run (min) End time mid xx% matched spectra in run (min) Time span mid xx% matched spectra in run (min) Median MS1 peak width mid xx% matched spectra (sec) Total precursor XIC mid xx% matched spectra in run Median MS1 intensity Trigger Apex mid 50% matched spectra (%)	How good the chromatography was For only the matched spectra between +/- one-half xx%, the program reports each of these statistics. For example, if mid xx% is specified as mid 60% and 10,000 spectra gave IDs in the run, then the statistics specified at mid 60% are reported for spectra #2000 to #8000 (+/- 30%). See the Tool Belt Online Help for descriptions of each of these statistics. To learn more about these metrics, refer to Rudnick PA, Clauser KR, Kilpatrick LE, et. al., "Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses", Mol Cell Proteomics. 2010 Feb;9(2):225-41 (http://www.ncbi.nlm.nih.gov/pubmed/19837981).
Distinct Peptide Fraction Overlap Table - Distinct	Distinct Peptide Fraction Overlap Table	The generated table shows for each peptide in each data file, the number of data files that peptide is found in. This is useful for seeing how good your pre-fractionation was. If everything is found in only one fraction and is also evenly distributed across the data files, your pre-fractionation worked perfectly. This is not likely, of course.
Distinct Peptide Fraction Overlap Table - Not Distinct	Distinct Peptide Fraction Overlap Table	The "best" peptide for each primary sequence, no matter the modification.

Table 6 Results After Reporting FDR and Search Statistics

Report this statistic:	To view these report column headers:	Because you want to know:
Precursor Acquisition Uncertainty: m/z and z	MS/MS spectra collected acquisition MI m/z = 0.0 MS/MS spectra collected acquisition z = 0 MS/MS spectra filtered acquisition MI m/z = 0.0 MS/MS spectra filtered acquisition z = 0 MS/MS spectra valid precursors adjusted	Spectral quality through these indicators; abnormally higher values for samples of the same complexity indicate poorer quality mass spectral data.
MS1 intensity stats for each run	# MS Scans Total TIC Total BPI Mean TIC Mean BPI	If something might be wrong, e.g., if the injection of the standard material shows an abnormally low value for these

To copy spectra to your collections directory

Use this tool to create a custom set of library spectra that you can search with the Spectrum Matcher. This tool copies spectra from one directory to another.

- 1 Navigate to the Tool Belt page and click the **Spectral collector** option, as shown in [Figure 99](#).

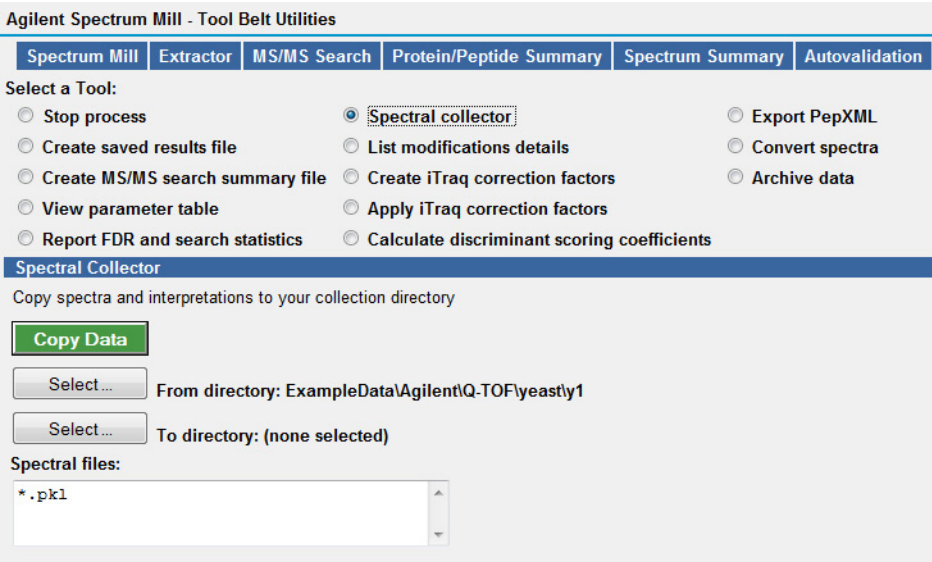


Figure 99 Spectral collector

- 2 Select the **From directory**.
- 3 Select the **To directory**
- 4 Click the **Copy Data** button.

NOTE

If you create a new directory for your spectral collections, you must copy a data file or placeholder file (e.g., *.aph) to the directory so that the software recognizes the directory.

To list details about amino acid modifications

- 1 Navigate to the Tool Belt page and click the **List modifications details** option, as shown in [Figure 100](#).
- 2 Select an **smconfig** file:
 - **smconfig.xml** - includes all the modifications that are currently displayed in the Spectrum Mill pages. The file **smconfig.xml** is created automatically by merging **smconfig.std.xml** and **smconfig.custom.xml** (which is a configuration file your server administrator optionally creates).
 - **smconfig.std.xml** - includes all of the modifications that are available when you first install the Spectrum Mill workbench
 - **smconfig.misc.xml** - includes additional modification definitions that Agilent provides “as is”. They are provided as a convenience and can be copied into **smconfig.custom.xml**, but they have not been rigorously tested at Agilent.
- 3 Click the **List** button.

Agilent Spectrum Mill - Tool Belt Utilities

Spectrum Mill	Extractor	MS/MS Search	Protein/Peptide Summary	Spectrum Summary	Autovalidation
---------------	-----------	--------------	-------------------------	------------------	----------------

Select a Tool:

☐ Stop process

☐ Spectral collector

☐ Export PepXML

☐ Create saved results file

☒ List modifications details

☐ Convert spectra

☐ Create MS/MS search summary file

☐ Create iTraQ correction factors

☐ Archive data

☐ View parameter table

☐ Apply iTraQ correction factors

☐ Report FDR and search statistics

☐ Calculate discriminant scoring coefficients

List Modifications Details

List

Select smconfig file: smconfig.xml

Figure 100 List modifications details

The listed details include:

- **Modification** - the name you see in the Spectrum Mill pages
- **Internal ID** - the name used by the Spectrum Mill programs
- **Type** - the modification type (fixed, variable, or cyclic)

- **Site(s)** - the amino acid or peptide-terminal location(s) for the modification
- **Delta formula** - the difference in chemical composition between the modified and unmodified site
- **Delta mass (Da)** - the monoisotopic mass difference between the modified and unmodified site

To create a file of iTRAQ correction factors

The Tool Belt provides the capability to incorporate the correction factors from an iTRAQ certificate of analysis into a file within the Spectrum Mill workbench. Corrections factors apply only to iTRAQ4 quantitation.

Once you have created the file with the correction factors, you can apply it to iTRAQ calculations for multiple samples. The file can store correction factors from multiple batches of iTRAQ reagents and is appended every time you type correction factors for a new batch number.

To access the capability to store correction factors:

- 1 Navigate to the Tool Belt page.
- 2 Click **Create iTRAQ correction factors**.
- 3 Refer to your iTRAQ certificate of analysis to fill in the page. See [Figure 101](#) as an example.
- 4 Click the **Create** button.
- 5 To apply the correction factors from a particular iTRAQ batch to the calculations for your sample, see [page 194](#).

Agilent Spectrum Mill - Tool Belt Utilities

Spectrum Mill

Extractor

MS/MS Search

Protein/Peptide Summary

Spectrum Summary

Autovalidation

Select a Tool:

☐ Stop process

☐ Create saved results file

☐ Create MS/MS search summary file

☐ View parameter table

☐ Report FDR and search statistics

☐ Spectral collector

☐ List modifications details

☒ Create iTraQ correction factors

☐ Apply iTraQ correction factors

☐ Calculate discriminant scoring coefficients

☐ Export PepXML

☐ Convert spectra

☐ Archive data

Create iTRAQ correction factors

(Only applies to 4-plex)

Create

Batch: 12345

Mass	% of -2	% of -1	% of +1	% of +2
114	0.0	1.0	5.9	0.2
115	0.0	2.0	5.6	0.1
116	0.0	3.0	4.5	0.1
117	0.1	4.0	3.5	0.1

Figure 101 Create iTRAQ correction factors, shown with example correction factors

To apply a file of iTRAQ correction factors

After you have stored iTRAQ correction factors in a file (see [page 193](#)), then you can apply them to your data file so that the Spectrum Mill workbench can do all the final iTRAQ calculations. (You do not need to export results to Excel to finalize the iTRAQ calculations.)

- 1 Navigate to the Tool Belt page.
- 2 Click **Apply iTRAQ correction factors**.
- 3 For **Batch**, select the iTRAQ batch number for the reagents you used with the samples.
- 4 Select the **Data Directory** that contains your iTRAQ data files.
- 5 Click the **Apply** button.

The screenshot shows the 'Agilent Spectrum Mill - Tool Belt Utilities' window. At the top is a tabbed interface with tabs for 'Spectrum Mill', 'Extractor', 'MS/MS Search', 'Protein/Peptide Summary', 'Spectrum Summary', and 'Autovalidation'. The 'Protein/Peptide Summary' tab is active. Below the tabs, under 'Select a Tool:', there are three columns of radio button options. The option 'Apply iTRAQ correction factors' is selected. Below this, the text '(Only applies to 4-plex)' is displayed. A green 'Apply' button is present. Below the button, the 'Batch:' field is set to '123'. The 'Data Directory' section shows a 'Select...' button and the path 'ExampleData\Agilent\Q-TOF\yeasty1'.

Figure 102 Apply iTRAQ correction factors

When the software does iTRAQ calculations in Protein/Peptide Summary, it automatically applies the correction factors. Be sure to check the report for a message about the batch number of the correction factors that the software used.

NOTE

If you copy or move your data files to a new directory, copy **iTRAQ.correction.txt** as well. This file must be present in the data directory so that future Spectrum Mill calculations can use the iTRAQ correction factors.

To calculate discriminant scoring coefficients

To have MS/MS Search calculate discriminant scores, you select among sets of standard Agilent discriminant scoring coefficients, either Q-TOF or ion trap. With this Tool Belt option you can create and update your own set of coefficients, but you must remember to use only validated hits.

To create your own set of coefficients:

- 1 Navigate to the Tool Belt page and click **Calculate discriminant scoring coefficients**.

Agilent Spectrum Mill - Tool Belt Utilities

Spectrum Mill

Extractor

MS/MS Search

Protein/Peptide Summary

Spectrum Summary

Autovalidation

Select a Tool:

☐ Stop process

☐ Spectral collector

☐ Export PepXML

☐ Create saved results file

☐ List modifications details

☐ Convert spectra

☐ Create MS/MS search summary file

☐ Create iTraQ correction factors

☐ Archive data

☐ View parameter table

☐ Apply iTraQ correction factors

☒ Calculate discriminant scoring coefficients

☐ Report FDR and search statistics

Calculate Discriminant Scoring Coefficients

Create

Name: Existing coefficient sets:

Disable (same as Score)

☐ Update

Data Directory

Select ...

ExampleData\Agilent\Q-TOF\yeasty1

Figure 103 Calculate discriminant scoring coefficients fields - Existing coefficient sets

- 2 In the Name field, type a name for the new set of coefficients.
The name appears in the **Name** field.
- 3 Under **Data Directory**, click **Select** to choose the directory containing valid data results.
- 4 Click **Create**.

At the bottom of the page you see the calculations as they are done. After they have completed, the name of the new set of coefficients appears in the Discriminant Scoring selection list on the MS/MS Search page.

To update your coefficients:

- 1** Navigate to the **Tool Belt** page and click **Calculate discriminant scoring coefficients**.
- 2** In the **Existing Coefficient Sets** list, select the coefficients you intend to update.
- 3** Mark the Update check box.
- 4** Under **Data Directory**, click Select to choose the directory containing valid data results.
- 5** Click **Create**.

To export PepXML files

PepXML format is a results exchange format that is supported by the Trans-Proteomic Pipeline from the Institute for Systems Biology. This format lets you import Spectrum Mill results into other software that supports the pepXML format, such as Skyline, an MRM data analysis package from the University of Washington.

To convert search results to this format:

- 1 Navigate to the Tool Belt page and click **Export PepXML**.

The screenshot shows the 'Agilent Spectrum Mill - Tool Belt Utilities' window. It has a tabbed interface with tabs for 'Spectrum Mill', 'Extractor', 'MS/MS Search', 'Protein/Peptide Summary', 'Spectrum Summary', and 'Autovalidation'. The 'MS/MS Search' tab is active. Under 'Select a Tool:', there are several radio button options, with 'Export PepXML' selected. Below this is the 'Export Spectrum Mill MS/MS Search Results to PepXML' section, which includes an 'Export' button, a description 'Convert MS/MS search summary file to multiple PepXML files (runName.pepXML)', and two checked checkboxes: 'Integrate rank 1 reversed hits as pseudo <search_hit> entries' and 'Omit protein names'. The 'Validation filter' is set to 'all'. The 'Data Directories' section has a 'Select...' button and a list containing 'ExampleData/Agilent/Q-TOF/yeast/y1' which is checked. At the bottom, there is another 'Select...' button and the text 'Export directory: (none selected)'.

Figure 104 Export PepXML fields

- 2 For **Validation filter**, select **Valid** or **All**.
- 3 If you want to include the top-ranked reversed hits, mark the check box for **Integrate rank 1 reversed hits as pseudo <search hit> entries**.
- 4 If you want to exclude the protein names, mark the check box for **Omit protein names**.
- 5 Under **Data Directories** click **Select** to choose the data directory that contains the summary files you want to export. You may select one or more data directories.

6 Click **Select** to choose the Export directory to which you intend to export the PepXML-formatted data.

7 Click **Export**.

Note that the program converts the search summary file(s) to PepXML format, and puts the output file(s) in the selected Export directory for which they apply. The output files are of the form DataFileName.pep.xml.

8 To access these files, do one of the following:

- Click the link(s) from the output frame, under **Spectrum Mill - Convert to PepXML**.
- Navigate to the data folder and click to open the pep.xml file.

To convert spectra

This option lets you convert the data extractor pkl files into mzXML files so you can import Spectrum Mill results into other software programs that support the mzXML format, such as Skyline, an MRM data analysis package from University of Washington. Or, you can use these files to contain the labeled MS/MS spectra required to identify post-translational modifications for papers published in MCP.

- 1 Navigate to the Tool Belt page and click **Convert Spectra**.

The screenshot shows the 'Agilent Spectrum Mill - Tool Belt Utilities' window. At the top, there are tabs for 'Spectrum Mill', 'Extractor', 'MS/MS Search', 'Protein/Peptide Summary', 'Spectrum Summary', and 'Autovalidation'. The 'Extractor' tab is active. Below the tabs, there is a section titled 'Select a Tool:' with a grid of radio button options. The 'Convert spectra' option is selected and highlighted with a dashed border. Other options include 'Stop process', 'Spectral collector', 'Export PepXML', 'Create saved results file', 'List modifications details', 'Archive data', 'Create MS/MS search summary file', 'Create iTraQ correction factors', 'View parameter table', 'Apply iTraQ correction factors', 'Report FDR and search statistics', and 'Calculate discriminant scoring coefficients'. Below the 'Select a Tool:' section, there is a blue header for 'Convert Spectra'. Under this header, it says 'Convert MS/MS spectral files ./cpick_in/runName.*.pkl to ./runName.mzXML'. There is a green 'Convert' button and a dropdown menu for 'Subset of spectra to convert:' set to 'all'. Below this is a section titled 'Data Directories'. It contains a 'Select...' button, a text field showing 'Data directory: ExampleData/Agilent/Q-TOF/yeast/y1' with a checkmark, and another 'Select...' button for 'Export directory: (none selected)'.

Figure 105 Convert spectra option

- 2 Select **All** or **Valid** as the **Subset of spectra to convert**.
- 3 Under **Data Directories**, click the top **Select** button to choose the directory containing the spectra you intend to convert.
- 4 Click the bottom **Select** button to choose the **Export** directory to which the mzXML files are exported when you click Convert.
- 5 Click **Convert**.

To archive data

This option lets you archive instrument-created files, search result files, spectral files and data directories, and lets you unzip search result files and spectral files.

- 1 Navigate to the Tool Belt page and click **Archive Data**. (Figure 106)

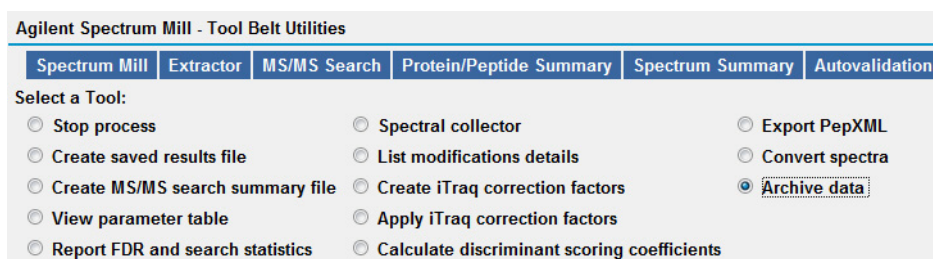


Figure 106 Tool Belt - Archive data option

- 2 Under **Data Directories**, click **Select** and choose the directory whose files and/or folder you intend to zip, move or unzip.

You can select only one at a time.

- 3 Select the general categories of files or directory you intend to archive and how you want to archive them. (Figure 107)

Below the options in each category, please read the description of what happens when you select an option.

- 4 Click **Archive**.

Archive Data

Archive

Instrument created files

☒ Ignore instrument data files
☐ Delete data files after making placeholder
☐ Move data files after making placeholder

Creates a placeholder file for each raw data file. Then deletes the raw data files or moves them to the temporary archive location on the Spectrum Mill server. Afterwards, the previously extracted data may be searched and summarized, but the raw data files cannot be re-extracted.

Spectrum Mill search result files

☒ Ignore
☐ Zip results_mstag/*.spo to spo.zip
☐ Unzip spo.zip to results_mstag/*.spo

After zipping to create spo.zip deletes results_mstag/*.spo. After unzipping to create results_mstag/*.spo deletes spo.zip.

Spectrum Mill spectral files

☒ Ignore
☐ Zip cpick_in/*.pkl to pkl.zip
☐ Unzip pkl.zip to cpick_in/*.pkl

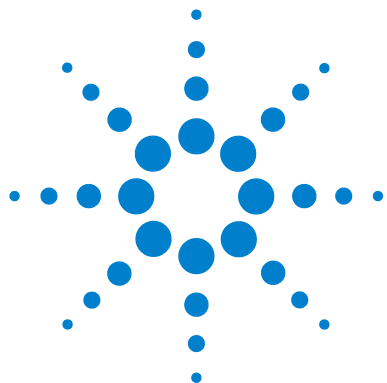
Zippping prevents further searching and spectral visualization. After zipping to create pkl.zip deletes cpick_in/*.pkl. After unzipping to create cpick_in/*.pkl deletes pkl.zip.

Spectrum Mill data directories

☒ Ignore
☐ Zip dataDir/*. * to dataDir.zip

Zippping prevents further use in SM. After zipping to create dataDir.zip deletes dataDir/*. *. After unzipping to create dataDir/*. * deletes dataDir.zip.

Figure 107 Archive Data selections



8

Using Spectrum Mill Utilities

To identify digest peptides likely to meet specific experimental goals
(Peptide Selector) [204](#)

To build MRM lists for triple quadrupole instruments [207](#)

To align sequences (Multiple Sequence Aligner) [209](#)

To list peptides that correspond to a theoretical protein digest (MS
Digest) [210](#)

To retrieve database entries using text searches (MS Edman) [212](#)

To list theoretical fragment ion masses for peptides (MS Product) [214](#)

To list amino acid compositions that fit precursor mass and partial
composition (MS Comp) [216](#)

To show isotope patterns of peptides (MS Isotope) [218](#)

To calculate the masses and formulas for peptides [219](#)

This chapter describes Spectrum Mill workbench peptide and protein utilities. These utilities help you accomplish a range of tasks, from retrieving database entries using text searches to simulating protein digestions. You access these utilities from the Spectrum Mill home page.



To identify digest peptides likely to meet specific experimental goals (Peptide Selector)

Peptide Selector runs theoretical digestions and then automatically selects from the theoretical peptides those that fit specific criteria. The most common uses are:

- List only those peptides suited to be synthesized with a stable isotope label and used for quantitation via a multiple reaction monitoring (MRM) experiment.
- List only those peptides suited for incorporation into an accurate mass inclusion list to be used in a data-dependent MS/MS experiment.
- List only those peptides expected to give doubly-charged ESI spectra.
- List only the likely detectable forms of all peptides that contain a possible phosphorylation site.
- List only those peptides that contain cysteines.

In many ways, Peptide Selector is similar to MS Digest, described on [page 210](#). Both Peptide Selector and MS Digest run automatic protein digestions. The difference is that Peptide Selector additionally creates limited lists based on specific criteria.

- 1 On the Spectrum Mill home page, click the **Peptide Selector** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 108](#).
- 3 To use an existing parameter file, click **Load**.
- 4 Set **Digest Parameters** and **Product Ion Parameters**. See the explanations in the online help.
- 5 Set **Saved File Parameters**. To see all the options for this section as shown in [Figure 108](#) on page 206, mark the **Generate inclusion or MRM list file** check box. See the explanations in the online help.
- 6 Set **Criteria for Excluding Peptides**. These criteria allow you to set the rules to discard peptides that you do not want to include in the list. The excluded peptides do not appear in the final report.
- 7 Click the **Choose...** button to select the modifications that match the chemistry for your samples.
- 8 Under **Protein(s) to Select From**, set **Database** options as follows:

- To use a sequence from a database, select the appropriate database name.
- To supply your own sequence, select **User Protein** and type or paste a sequence into the **User-supplied sequence** box. If you have multiple sequences, you can use a variety of delimiters, as found in HTML- and Excel-formatted Spectrum Mill reports.

9 Set **Search Mode** parameters as described in the online help.

10 Set the **Scoring** parameters.

If you do not get enough peptides to pass the filter, you can apply a penalty rather than excluding the peptides. When you mark the **Penalize rather than exclude** check box, the settings for which you can apply penalties change to green font. (Figure 108) These settings no longer cause peptides to be excluded. See the *Online Help* for details.

11 To save the changed parameters, click **Save As**.

12 Click **Select**.

See the *Online Help* to learn how to transfer the inclusion list to the Agilent MassHunter Data Acquisition software.

Agilent Spectrum Mill - Peptide Selector - (none)

Spectrum Mill | MS Edman | Multiple Sequence Aligner | Databases | MS Digest | Help

Selection

Select Save As... Load ... ☐ Hide HTML links (better Excel cut/paste)

Saved File Parameters

☒ Generate inclusion or MRM list file MassHunter QQQ MRM list

Filename: lastPeptideSelectorResult.txt Run time(min): 89.9 List Type: m/z

Precursor Charge Min.: 2 Max.: 4 Precursor m/z Min.: 300.0 Max.: 2000.0

Digest Parameters

Digest: Trypsin Maximum # missed cleavages: 0

Product Ion Parameters

☒ Show Product Ion Masses

Criteria for Excluding Peptides

Max. # basic residues (RHK): 2

Peptide MH⁺ Min.: 900.0 Max.: 2800.0

AA Composition Filtering:

AAs Required : KR Disallowed: CM

Peptide exclusion criteria:

☒ Has nearby cleavage site within 3 residues

☒ Contains peptide N-terminal Gln to pyroGlu

☐ Contains protein N-terminus Acetyltable

☐ Contains consensus N-linked glycosylation site

☐ Contains no variable modification

Protein Position Filtering:

Each peptide must contain at least 1 of the following positions in the protein (enter 1 position/line)

Modifications

Choose... Fixed: Variable:

Protein(s) to Select From

Database: IPL_HUMAN

Select Peptides from only the Database Entries with the accession numbers below

Allowed delimiters (-,:)

P07288
P02741
P00533
P06702
P16860
P03372

Search Mode

Count Peptide Uniqueness in Database by: Sequence

Species: All

Scoring

☒ Penalize rather than exclude

Max reported peptides/protein: 5

Penalty

Excess basic residues 0.1

Nearby cleavage site 0.09

Disallowed AA composition 0.3

Lower relative uniqueness 0.1

Figure 108 Peptide Selector

To build MRM lists for triple quadrupole instruments

The MRM Selector uses Spectrum Mill MS/MS Search results to create multiple reaction monitoring (MRM) lists for triple quadrupole instruments. You select data that has been searched, then filter the results to include only those peptides that meet certain requirements. The filters are similar to those in Protein/Peptide Summary.

- 1 On the Spectrum Mill home page, click the **MRM Selector** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 109](#).
- 3 Set the **Format** and choose the **Data Directories**.
- 4 Set the **Validation and Sorting** parameters. They are very similar to the ones in Protein/Peptide Summary.
- 5 Set the **MRM Parameters**, which determine the contents of the MRM list. These specific settings depend on the **Format** you selected.
 - The Agilent Q-TOF Data Extractor outputs the collision energy (CE), peak apex, and chromatographic peak width in the **specFeatures.tsv** file in order to use these values in MRM and dynamic MRM (DMRM) generation.
 - You may instead choose to let the program calculate the collision energies based on an equation.
 - You have the option to type a chromatographic peak width that will apply for all peaks in the DMRM analysis. This peak width, which corresponds to the **delta RT** setting in the MassHunter Data Acquisition software, is the retention time window for which the MRM transitions are monitored. For example, if you have a peak apex at 2.5 min and a delta RT of 1.0 min., the MassHunter Data Acquisition software monitors the MRM transitions for that peak from 2.0 min until 3.0 min.
- 6 Click **Select MRMs**.
- 7 See the online help to learn how to transfer the MRM list into MassHunter Data Acquisition software.

Select Results for MRM Selection

Select MRMs

Save As...

Load...

Format: MassHunter QQQ DMRM

Data directories: Select...

☐ ExampleData/Agilent/Q-TOF/OGF

Search result files:

*.spo

Validation and Sorting

Filter results by: valid

Protein grouping method: 1 shared, expand subgroups

Sort proteins by: Score

Filter by protein score: > 13.0

Top n peptides for MRM: Limit to 5

Rank peptides by: Score

Sort MRM List by: Precursor m/z

Filter peptides by:

Score: > 10 % SPI: > 70.0

Required AAs: any Disallowed AAs: none

Peptide pI: from 3.0 to 10.0 ☒ All

Accession #'s:

MRM Parameters

☐ Text file export

Top n Transitions: 5 ≥ 100 % of precursor ☐ y-ions only

Trigger threshold: 500 for 0 secondary transitions

Z options: Observed precursor/fragment charge only

Dwell time (ms): 10

☐ Use peak width of: 1 min

Collision Energy:

(CE= slope*m/z + intercept)

☒ Use actual CE if available

z-state: slope: intercept:

2	3.6	-4.8
3	3.6	-4.8
4	3.6	-4.8

Figure 109 MRM Selector

To align sequences (Multiple Sequence Aligner)

The Multiple Sequence Aligner enables alignment and comparison of the amino acid sequences of proteins that are present in a database. This stand-alone utility within the Spectrum Mill software highlights the amino acids that differ among the sequences. (When you access multiple sequence alignment from Protein/Peptide Summary, the software highlights the identified peptides within the sequence.)

- 1 On the Spectrum Mill home page, click the **Multiple Sequence Aligner** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 110](#).
- 3 Select the name of the database that contains the proteins you wish to align.
- 4 Type the accession numbers of the proteins that you wish to compare.
- 5 Click the **Align** button.

Figure 110 Multiple Sequence Aligner

NOTE

If the database is too large, the alignment does not work properly. In that case, create a subset database before you do the alignment.

To list peptides that correspond to a theoretical protein digest (MS Digest)

MS Digest runs theoretical digestions and calculates masses of peptides that result. The program accepts both user proteins and sequences from databases.

- 1 On the Spectrum Mill home page, click the **MS Digest** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 111](#).
- 3 Click the **Choose...** button to select the modifications that match the chemistry for your samples. For details, click the blue bar labeled **Modifications** to access the online help.
- 4 Fill in the other options. See the explanations below and in the online help. In general, you should keep the default values, except for settings that are highlighted in red.
- 5 If you want to include a **User-specified amino acid**, indicate it with a lower-case **u** in the **Protein sequence** box, and then specify an elemental composition.
- 6 Click **Digest**.

Figure 111 MS Digest

Database To select a sequence from a database, select the appropriate database name. To supply your own sequence, select **User Protein** and supply a sequence in the **Protein sequence** box.

Retrieve database entry by:

When you select a database name, you then have two choices:

- Select **Accession Number** and supply the correct **Database accession number** in the box to the right.
- Select **Index Number** and supply the correct **MS Digest index number** in the box to the right.

The **Accession Number** and **Index Number** entries apply only if you have selected a database name as the **Database** option. If you have selected **User Protein**, these entries are ignored.

To retrieve database entries using text searches (MS Edman)

MS Edman allows you to search text fields (such as sequence, name, accession number or species) in protein databases.

MS Edman can help identify a protein if you know only the molecular weight of a tryptic fragment and some of the sequence. For example, if you know a peptide mass and some sequence information from *de novo* sequencing, then MS Edman is a useful search tool.

MS Edman is also the first step when you want to create a specialized subset database of entries that match your text search criteria. For example, you could use MS Edman to find all proteins that contain a certain amino acid sequence. If you marked the **Save hits to file** check box on the MS Edman page, you could then use Protein Databases to create a subset database from these saved results. See [“To create a subset database from saved hits”](#) on page 227.

- 1 On the Spectrum Mill home page, click the **MS Edman** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 112](#).
- 3 Set options for **Search**, **Search Parameters**, and **Modifications**. See explanations in the online help.
- 4 Select a **Search Mode**. Your choice determines which additional options appear. There are three types of search modes:
 - **Sequence Only**: Search for an amino acid sequence. Type the sequence as a **Regular expression**. (See the next step.)
 - **Sequence and Mass**: This is the same as the **Sequence Only** mode, with the addition that search results are filtered by peptide mass. Type a sequence and mass(es).
 - **Name, Accession Number or Species**: Type one of these in the box below.
- 5 To type the sequence in the box under **Regular expression**, use regular expressions that the UNIX® **grep** command understands. Think of regular expressions as very intelligent wild cards. Use the following guidelines:
 - **WVTF** (no brackets) means the sequence is exactly WVTF. Use this type of designation for parts of the sequence that are known with certainty.

- [AK] means the amino acid is *either* A or K. The program searches for either one.
- [^EF] means the amino acid is *neither* E nor F. The program searches for all other possibilities.
- . (period) means a single amino acid is unknown. The program searches for all possibilities.
- .* (dot star) means a section of the sequence is unknown. The length of that section is not specified. For example, FMQ.*K means that the sequence begins with FMQ and ends with K, but there is a section of unknown sequence of unspecified length in the middle.

In the example shown in [Figure 112](#), the sequence is: either A or K, followed by either A or K, followed by PV, followed by either I or L, followed by ED, followed by either I or L, followed by R.

For more information on the grep command, go to a UNIX® system and type **man grep**, or search the internet for **grep**.

6 Click the **Start Search** button.

Agilent Spectrum Mill - MS Edman

Spectrum Mill Manual PMF MS Digest Databases *de novo* Help

Search

Start Search Maximum reported hits: 200

☐ Display combinatorial peptide output

Sample ID: Enter_Comment

Search Parameters

Database: NCBI nr.bovine

Species: All

Digest: No enzyme

☐ Search hits from file: lastres

☐ Save hits to file: lastres

Maximum # of mismatched AA's: 2

MW of protein: from 1000 Da to 100000 Da ☒ All

Modifications

Choose... Fixed: Carbamidomethylation (C)

Search Mode

Search mode: Sequence Only

Regular expression (Use CAPITALS):

[AK][AK]PV[I]L[ED][I]R

Figure 112 MS Edman

To list theoretical fragment ion masses for peptides (MS Product)

MS Product calculates theoretical ion masses from peptides that undergo dissociation via post-source decay or high- or low-energy collision-induced dissociation.

- 1 On the Spectrum Mill home page, click the **MS Product** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 113](#).
- 3 Fill in the options. See the online help for explanations. Note that under **Product Ion Types**, you need to select the ion types that are appropriate for your instrument.
- 4 If you want to include a user-specified amino acid, indicate it with a lower-case **u** in the **Sequence** box, and then specify an elemental composition at the bottom of the page.
- 5 Click the **Fragment** button.

Agilent Spectrum Mill - MS Product

Spectrum Mill **MS Digest** **Help**

Fragmentation

Fragment Calculate masses as: Maximum reported charge:

Peptide Sequence

Use **CAPITAL** letters for unmodified amino acids or those with Fixed Modifications.
 Use **lower case** letters for amino acids to indicate a Variable Modification (eg. KCmqYQ).
 Use 'u' for the user-specified amino acid (see below).
 Use 'o' for Homoserine Lactone.

Enter sequence:

Modifications

Choose... Fixed: Variable:

Product Ion Types
 (some types require presence of specific AA's in ion)

AA composition ions	N-terminus sequence ions	C-terminus sequence ions	Internal product ions	Ladder sequencing ions
☒ i ☐ m	☒ a ☒ b ☐ c ☐ c' ☐ x	☒ y ☐ Y ☐ z ☐ z' ☐ z''	☒ internal	☐ N-terminus ☐ C-terminus

Satellite sequence ions (side-chain loss)	Neutral-loss sequence ions	Peeling sequence ions
☐ d ☐ v ☐ w	☒ -H ₂ O S, T ☒ -NH ₃ R, K, Q ☐ -H ₃ PO ₄ s, t (S, T + PO ₄) ☐ -SOCH ₃ m (M + 0x)	☒ b+H ₂ O R, H, K

User-specified AA elemental composition (u):
 C: H: N: O: S: P: (Default is Gly)

Figure 113 MS Product

To list amino acid compositions that fit precursor mass and partial composition (MS Comp)

MS Comp fills in possible amino acid compositions for a peptide, given a peptide mass and partial composition determined from immonium ions present in MS/MS spectra.

- 1 On the Spectrum Mill home page, click the **MS Comp** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 114](#).
- 3 Select the **Combination type**. You have three choices:
 - Select **Amino Acid** to get a list of amino acids.
 - Select **Peptide Elemental** to get a list of elemental compositions corresponding to these amino acids.
 - Select **Elemental** to get a list of elemental compositions without regard to whether or not they could be peptides. This option speeds searches at higher masses.
- 4 In the **Peptide** section, type an **m/z** value and select possible **Ion types** that correspond to this *m/z* value.
- 5 Fill in the other options. See the explanations in the online help.
- 6 Click the **Composition** button.

Agilent Spectrum Mill - MS Comp

Spectrum Mill | **MS Edman** | **Spectrum Summary** | *de novo* | **Help**

Compositions

Composition Maximum reported compositions: 1000 Combination type: Amino Acid

Note: If the mass tolerance is large and few ions below are checked then the program will take a long time to run.

Peptide

m/z: 1260.5992
 Da +/-: 10 ppm
 Charge (z): 1
 m/z is: Monoisotopic

Ion types:

MH+	y-NH3	y-H3PO4
y	b-NH3	b-H3PO4
b	a-NH3	a-H3PO4
a	y-H2O	y-SOCH4
c	b-H2O	b-SOCH4
b+H2O	a-H2O	

AA Composition

Based on immonium and related ions:

☐ 60 S	☐ 70 R,P	☒ 72 V	☐ 74 T	☐ 86 L,I	☐ 87 N,R	☐ 88 D	☐ 84,101,129 K,Q	☒ 102 E
☐ 104 M	☒ 110 H	☐ 120 F	☐ 126 P	☒ 136 Y	☒ 117,130,159,170 W	☒ 70,73,87,100,112,185 R		

Based on loss from precursor ion:

☐ -SRH-R C (select modified Cys above)	☐ -80,-98 PO ₄ - S,T (check s and t below)	☐ -64 SOCH ₄ - M (check m below)	☐ -17 R
--	---	---	-----------------------------------

Note: For ions with multiple residues (i.e. 86: means I or L).
 Note: Requires presence of at least one AA for each composition ion selected.

AminoAcids

Absent amino acids (Check to prevent possible inclusion)

☐ A	☐ C	☐ D	☐ E	☐ F	☐ G	☐ H	☐ I	☐ K	☐ L	☐ M	☐ N	☐ P	☐ Q	☐ R	☐ S	☐ T	☐ V	☐ W	☐ Y
----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------	----------------------------

Modifications

☐ User defined amino acid: C: 2 H: 3 N: 1 O: 1 S: 0 P: 0 (Default is Gly)

Choose... Fixed: Carbamidomethylation (C) Variable:

Figure 114 MS Comp

To show isotope patterns of peptides (MS Isotope)

MS Isotope calculates and displays isotope patterns of peptides.

- 1 On the Spectrum Mill home page, click the **MS Isotope** link in the **Utilities** section.
- 2 Check that you see the page shown in [Figure 115](#).
- 3 Select **Peptide sequence** to display the isotope pattern for a peptide sequence, or select **Elemental composition** to display the isotope pattern for an elemental composition. The other options change accordingly.
- 4 Fill in the options. For explanations, see the online help.
- 5 Click **Calculate**.

Agilent Spectrum Mill - MS Isotope

Spectrum Mill Tool Belt Help

Isotope Distribution

Calculate Calculate masses as: Monoisotopic ☐ Show detailed report

☒ Peptide sequence
☐ Elemental composition

Peptide Sequence

Use **CAPITAL** letters for unmodified amino acids.
 Use lower case letters for variable modified amino acids to indicate modification (eg. KCmqYQ).
 Use 'u' for the user-specified amino acid.

Sequence:

User-specified AA elemental composition (u):

C: H: N: O: S: P: (Default is Gly)

Modifications

Choose... Fixed: Carbamidomethylation (C) Variable:

Figure 115 MS Isotope

To calculate the masses and formulas for peptides

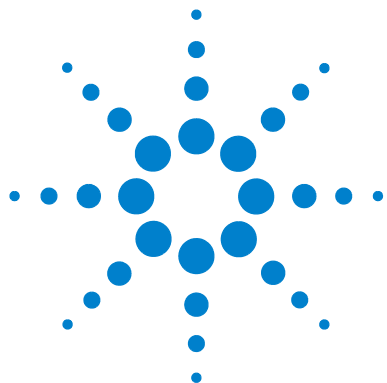
With the Peptide List to Masses utility, you can have the program calculate the masses and formulas for a set of peptides you specify.

- 1 On the Spectrum Mill Home page, under Utilities, click **Peptide List to Masses**.

The screenshot shows the 'Agilent Spectrum Mill - Peptide List to Masses' utility window. It features a navigation bar with 'Spectrum Mill', 'MS Digest', 'Peptide Selector', 'MS Product', 'Tool Belt', and 'Help'. The main section is titled 'Calculate Masses' and includes a 'Calculate' button, a dropdown for 'Calculate masses as:' (set to 'Monoisotopic'), and input fields for 'Reported precursor charge Min:' (2) and 'Max:' (3). Below these are radio buttons for 'Limit actual charge by:' with options 'RKH present' (selected), 'RKHQN present', and 'Above min/max'. A 'Modifications' section has a 'Choose...' button, a 'Fixed:' field containing 'Carbamidomethylation (C)', and an empty 'Variable:' field. The 'Peptide Sequences' section provides instructions on using capital letters for unmodified amino acids and lower case for variable modifications, followed by an 'Enter peptide sequences:' label and a list box containing: 'SAMPLER', 'SAMcPLER', 'SAMsPLEK', 'SAMyPLEK', 'SQAMyPLEK', and 'SNAMyPLEK'.

Figure 116 Peptide List to Masses utility

- 2 From the **Calculate masses as:** list, select how to calculate masses.
- 3 Type a minimum and maximum charge for the **Reported precursor charge**.
- 4 Next to **Limit actual charge by:**, select to limit the charge by the arginine, histidine and lysine present (**RKH present**), by those three amino acids plus glutamine and asparagine (**RKHQN present**), or by the Min and Max specified in the previous line (**Above min/max**).
- 5 Click **Choose** to select the modifications.
- 6 Type the **Peptide Sequences** whose masses and formulas you want to know.
- 7 Click **Calculate**.



9

System Administration

Manipulating sequence databases [222](#)

Other system administration tasks [233](#)

This chapter describes system administration tasks to install, maintain, and customize sequence databases, and to customize various aspects of the Spectrum Mill workbench. This chapter covers the highlights; you can find additional information in the online help. See **millhtml\SM_instruct\faman.htm** and **millhtml\SM_instruct\servadmn.htm**. You can link to these from any Spectrum Mill help page. The **Protein Databases** link takes you to **faman.htm**, while the **Server Administration** link takes you to **servadmn.htm**.



Manipulating sequence databases

To install or update databases

A number of public protein and DNA sequence databases are available via the Internet. Here is how you install them:

- 1 Establish an internet connection to a site that has a database available for download.
 - The easiest way to connect is to click the appropriate link in the “Updating Databases” section of the online help file **faman.htm** (help for the Protein Databases page). If the links have become obsolete, check the *Agilent Software Status Bulletin and Patches* Web site for a file update. See the *Quick Start Guide* for details to connect to this Web site.
 - If you want to automate this process, you may want to use an ftp utility.
- 2 To download the database file, drag the file (zip format) from the browser window to an open directory window on your Spectrum Mill server. For performance reasons, it is best to:
 - Locate the databases in a separate drive from the rest of the Spectrum Mill files.
 - Locate the databases on the Spectrum Mill server PC rather than on a remote computer.
- 3 Use a ZIP file utility program to extract the database file into the **seqDB** directory in your Spectrum Mill workbench installation.
- 4 Rename the database file with the required name prefix, as described in the online help. Use capital letters *exactly* as described in the online help. This prefix is critical for the Spectrum Mill workbench to parse the header for each protein to retrieve protein name and species information. Use of a file extension (such as **.fas**) is optional.

NOTE

When you update a database, use the same file name. Do not append a new suffix. Otherwise, you will have problems when you review older data.

You *must* create database indices each time you update a database.

- 5 If you download a second database of the same FASTA format, and you want to link to different URLs from the search results, you rename the second database with a similar lower case database prefix. See the “Updating Databases” section of the online help file **faman.htm**. See also [“To change the URLs of HTML links in the search results”](#) on page 234.
- 6 If you want to search a user-created database, it is important to know the FASTA format of the database and to follow Spectrum Mill naming conventions. See the “The Spectrum Mill Workbench File Naming Conventions for Proprietary/Generic FASTA Databases” in the online help file **faman.htm**.
- 7 Follow directions in the next section to create database indices.

To create database indices

- 1 If you have not done this before, read the explanation at the end of the task.
- 2 If you wish to use the command line version of **Protein Databases** for automation purposes, skip the rest of this section and see the online help document **faman.htm**. Read the section entitled “The Command Line Version of Protein Databases.” Otherwise, continue to the next step.
- 3 To use a browser window to create indices, click the link to the **Protein Databases** page from the **Utilities** section of the Spectrum Mill home page.
- 4 Select the **Create indices for new database** option.
- 5 Check that you see the page shown in [Figure 117](#).
- 6 Type the *exact* name of the new database. Be sure to use capital and lower-case letters *exactly* as in the database name.
- 7 Click the **Create Indices** button.
- 8 As the indices are created, check that you see a display like the one in [Figure 118](#).
- 9 Click the **Back** arrow to return to the **Protein Databases** page.
- 10 Click **Update Database List** and verify that the database you just indexed appears in the list of databases.
- 11 To maintain server performance, always defragment the database hard drive after you install a new database.

Agilent Spectrum Mill - Protein Database Utilities

Spectrum Mill MS/MS Search PMF Search Tool Belt Help

☒ Create indices for new database
 ☐ Compare two databases

☐ Create species subset database
 ☐ Calculate statistics

☐ Create subset with indices from saved hits
 ☐ Make subset FASTA file from Accession Numbers

☐ Create or append user database

☐ Database summary report

Update Database List (After a re-install, or if you don't see the database listed, click the "Update Database List" button.)

Create Indices

Newly downloaded database: Enter_filename_here

Existing databases: IPI_HUMAN

Figure 117 Create indices for new database

Protein Database
Counting entries in database and creating index file (.idx). Processing entries to find species, accession number, and unknown sequence characters. Calculating Protein MW and pI and creating unknown sequence characters file (.unk). 18 entries in the database. 18 entries processed. 0 entries contained unreadable species. Creating mole_wt file (.mw). Creating pI file (.pI). Creating species file (.sp). Creating species list file (.sl). Creating accession_number file (.acn). Indexed Database: NCBIInr.stdmix

Figure 118 Display while database indices are created

Indices serve a number of useful purposes in the Spectrum Mill workbench, including:

- Faster output generation when you review results
- Accelerated searches when you use molecular weight or species filters. See [“To create a species and protein molecular weight subset database”](#) on page 226.
- Searches of results saved from other search programs. See [“To create a saved results file so you can search previous hits”](#) on page 178 and [“To create a subset database from saved hits”](#) on page 227.
- Creation of custom databases. See [“To create a user \(proprietary\) database or add sequences to one”](#) on page 228.

NOTE

Spectrum Mill indices bear no relationship to indices used in another popular protein database search program.

To create a species and protein molecular weight subset database

For faster searches, you create subset databases based on species and molecular weight filters.

- 1 To create a subset database, click the link to the **Protein Databases** page from the **Utilities** section of the Spectrum Mill home page.
- 2 Select the **Create species subset database** option.
- 3 Check that you see the page shown in [Figure 119](#).
- 4 Fill in the page. Choose a descriptive suffix for the database, such as “human” or “yeast.”
- 5 Click the **Create** button.
- 6 In this example, the created subset database is **NCBInr.ecoli**.

Agilent Spectrum Mill - Protein Database Utilities

Spectrum Mill | MS/MS Search | PMF Search | Tool Belt | Help

☐ Create indices for new database

☒ Create species subset database

☐ Create subset with indices from saved hits

☐ Create or append user database

☐ Database summary report

☐ Compare two databases

☐ Calculate statistics

☐ Make subset FASTA file from Accession Numbers

Update Database List (After a re-install, or if you don't see the database listed, click the "Update Database List" button.)

Create

Suffix for subset database: sub

Existing database: IPI_HUMAN

Species: All

MW of protein: (from 1000 Da to 100000 Da) ☒ All

Figure 119 Create species and protein molecular weight subset database

NOTE

Because of inconsistencies in the way species information is organized in different databases, the Spectrum Mill workbench cannot read about 1 to 2% of the species designations in NCBInr, and cannot read any of the species information in trEMBL. To create a molecular weight subset database from trEMBL, set the **Species** filter to **All**.

To create a subset database from saved hits

The Spectrum Mill workbench allows you to create a subset database from a *.res file (list of accession numbers) you save in MS Edman. This is useful because MS Edman allows you to do a text search on a database and then save the results. The text search could find, for example, all proteins with a given name, or all proteins that contain a certain partial amino acid sequence. You use Protein Databases to read the *.res file and create a new subset database. You can then use any of the Spectrum Mill database search programs to search this database from any data directory.

- 1 Make sure you have already run MS Edman and saved hits to a file.
- 2 To create a subset database, click the link to the **Protein Databases** page from the **Utilities** section of the Spectrum Mill home page.
- 3 Select the **Create subset with indices from saved hits** option.
- 4 Check that you see the page shown in [Figure 120](#).
- 5 Fill in the page. Be sure to use a unique suffix so you don't overwrite previously-created databases.
- 6 Click the **Create** button.

Agilent Spectrum Mill - Protein Database Utilities

Spectrum Mill | MS/MS Search | PMF Search | Tool Belt | Help

☐ Create indices for new database
 ☐ Compare two databases
☐ Create species subset database
 ☐ Calculate statistics
☒ Create subset with indices from saved hits
 ☐ Make subset FASTA file from Accession Numbers
☐ Create or append user database
☐ Database summary report

Update Database List (After a re-install, or if you don't see the database listed, click the "Update Database List" button.)

Create

Suffix for subset database: lastres

Database used in search to save hits: IPL_HUMAN

From MS Edman results file: lastres

Figure 120 Create subset database with indices from saved hits

To create a user (proprietary) database or add sequences to one

The Spectrum Mill workbench allows you to create and search your own FASTA-format databases.

- 1 To create a user database, click the link to the **Protein Databases** page from the **Utilities** section of the Spectrum Mill home page.
- 2 Select the **Create or append user database** option.
- 3 Check that you see the page shown in [Figure 121](#).
- 4 Follow procedures in the online help document **faman.htm**. It is best if you read the entire document. As a minimum, read the following two sections:
 - “Creating or Appending to a Database Containing User Supplied Protein or DNA Sequences”
 - “The Spectrum Mill Workbench File Naming Conventions for Proprietary/Generic FASTA Databases.” (The database will not work if you fail to name it correctly.)

Agilent Spectrum Mill - Protein Database Utilities

Spectrum Mill | MS/MS Search | PMF Search | Tool Belt | Help

☐ Create indices for new database

☐ Create species subset database

☐ Create subset with indices from saved hits

☒ Create or append user database

☐ Database summary report

(After a re-install, or if you don't see the database listed, click the "Update Database List" button.)

Database name:

Description:

Species: Accession number:

Paste or type sequence in this field. Delete these instructions first.

Tabs, returns, and spaces are ignored.

USE CAPITAL LETTERS.

Do NOT use the letters B, J, O, U, X, or Z.

Do NOT use 3-letter code amino acid symbols.

DNA sequences can be added to DNA databases.

Figure 121 Create or append to user database

To generate a database summary report

You use a database summary report to view a range of database entries. This report is useful for troubleshooting.

- 1 To create a summary report, click the link to the **Protein Databases** page from the **Utilities** section of the Spectrum Mill home page.
- 2 Select the **Database summary report** option.
- 3 Check that you see the page shown in [Figure 122](#).
- 4 Fill in the page.
- 5 Click the **Summarize** button.

Agilent Spectrum Mill - Protein Database Utilities

Spectrum Mill MS/MS Search PMF Search Tool Belt Help

☐ Create indices for new database
 ☐ Compare two databases
☐ Create species subset database
 ☐ Calculate statistics
☐ Create subset with indices from saved hits
 ☐ Make subset FASTA file from Accession Numbers
☐ Create or append user database
☒ Database summary report

Update Database List (After a re-install, or if you don't see the database listed, click the "Update Database List" button.)

Summarize

Database: IPI_HUMAN

Reading frame: 1 (for DNA only)

Index number range: 1 to 1 ☐ All

☐ Hide protein sequences

Figure 122 Database summary report

To compare two databases

With this utility you can compare the listings of two databases to view their similarities and differences.

- Longest and shortest proteins in the two selected databases
- # of entries of exactly matching sequences
- # of matching entries with different accession #s
- # of entries in one database with no matching sequences in the other
- Sequence Length Histograms for both databases

- 1 From the Utilities section of the Spectrum Mill Home page, click **Protein Databases**.
- 2 Click **Compare two databases**.
- 3 Select a database from the **Database 1** list.
- 4 Select a database from the **Database 2** list.

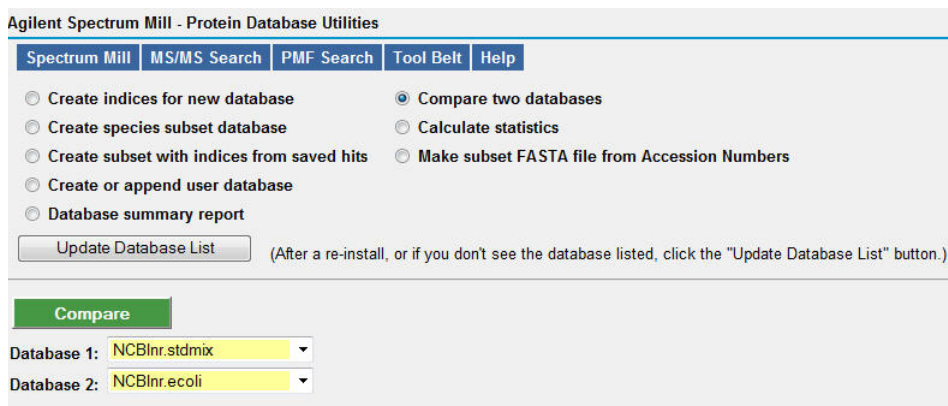


Figure 123 Compare two databases option

- 5 Click **Compare**.

To calculate statistics

If you choose this option, the program calculates these statistics:

- Number of entries with leading methionine
- Number of distinct peptides of length 8-40
- Number of non-distinct peptides of length 8-40
- Most repeated peptide
- Distinct Peptide Redundancy Histogram
- Sequence Length Histogram

- 1 From the Utilities section of the Spectrum Mill Home page, click **Protein Databases**.
- 2 Click **Calculate Statistics**.

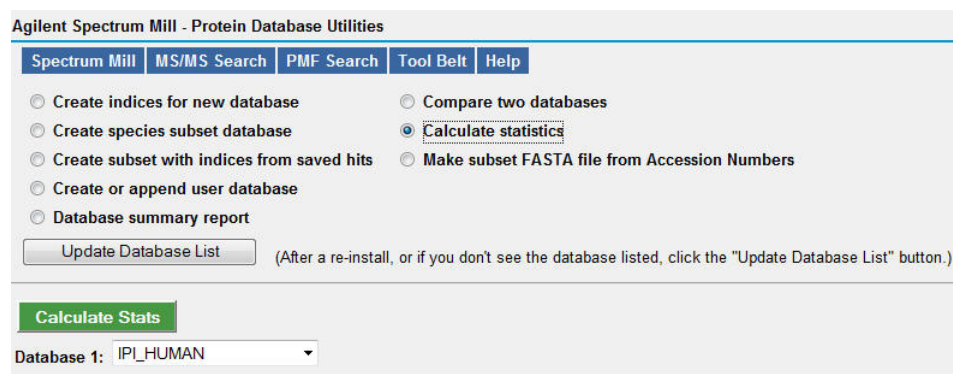


Figure 124 Calculate statistics option

- 3 From the **Database 1** list, select the database for which you want the program to calculate statistics.
- 4 Click **Calculate Stats**.
- 5 Click **Update Database List**.

To make a subset FASTA file from Accession Numbers

This option creates a subset FASTA file from accession numbers that you type. It is useful for limiting searches to the set of proteins of particular interest.

- 1 Under the Utilities section of the Spectrum Mill Home page, click **Protein Databases**.
- 2 Click the **Make subset FASTA file from Accession Numbers** option.

Agilent Spectrum Mill - Protein Database Utilities

Spectrum Mill | MS/MS Search | PMF Search | Tool Belt | Help

☐ Create indices for new database
☐ Create species subset database
☐ Create subset with indices from saved hits
☐ Create or append user database
☐ Database summary report

☐ Compare two databases
☐ Calculate statistics
☒ **Make subset FASTA file from Accession Numbers**

(After a re-install, or if you don't see the database listed, click the "Update Database List" button.)

Make Subset

Suffix for subset database:

Database:

Accession #'s:

Figure 125 Make subset FASTA file from Accession Numbers option

- 3 In the **Suffix for subset database** field, type the name of the suffix for your database.
- 4 From the **Database** list, choose a database.
- 5 Type the accession numbers you want to include, separated by a semi-colon (;).
- 6 Click Make Subset.
- 7 On the Spectrum Mill server, navigate to the folder where your databases are stored (for example, D:\seqdb).

Note the new files created there.

- 8 Create indices for the new subset database. See [“To create database indices”](#) on page 224.

Other system administration tasks

To add custom amino acid modifications

To set up custom amino acid modifications, start with the file `\msparams_mill\smconfig.custom.empty`.

CAUTION

Do *not* modify **smconfig.std.xml** or **smconfig.xml**.

The file **smconfig.misc.xml** contains definitions that are provided as a convenience but that have not been thoroughly tested at Agilent. Please verify the definitions before you use them.

- 1 Copy **smconfig.custom.empty** to **smconfig.custom.xml**.
- 2 Create definitions for modifications and/or mappings for homology search modes.
 - To make the task easier, examine the files **smconfig.std.xml** or **smconfig.misc.xml** for modification definitions that are similar to the ones you need. You may copy and paste these definitions into your **smconfig.custom.xml** file, and then modify the definitions.
 - See the online help document **SMCustomModifications.doc** for details and syntax for adding modifications.
- 3 To make your custom modifications available in the software, run the perl script `\millscripts\updateModsFile.pl` or click the **Choose...** button in one of the Spectrum Mill pages.
 - The Spectrum Mill workbench automatically combines the **smconfig.custom.xml** file and the **smconfig.std.xml** file to produce the **smconfig.xml** file that is used by the Spectrum Mill programs.
 - When the software creates the **smconfig.xml** file, definitions in **smconfig.custom.xml** override any definitions in **smconfig.std.xml**.

NOTE

The **smconfig.custom.xml** file is *not* overwritten during a re-install of the Spectrum Mill workbench.

To change the URLs of HTML links in the search results

By default, search results use the HTML links shown in [Table 7](#). You edit text files listed in the last column of [Table 7](#) to change these links.

- 1 Make a backup of each file before you modify it.
- 2 Edit the file. See the online help document **servadmn.htm** for details.

Table 7 HTML links in search results

Click this	To go to this	Text file to edit link
Accession number	Remote database entry	mparams_mill\ urlsAccessionNum.txt
MS Digest index number	MS Digest results for database sequence	mparams_mill\ urlsIndexNum.txt
Amino acid sequence (in individual MS/MS or PMF search results)	MS Product listing for that sequence	mparams_mill\urlsSequence.txt

To enable the HTML link to BLAST search

If you wish to have the amino acid sequence in Protein/Peptide Summary link to BLAST (Basic Local Alignment Search Tool), edit the file **millscripsts\libSpecMill\lsmGlobals.pm**.

- 1 Make a backup of the **millscripsts\libSpecMill\lsmGlobals.pm** file.
- 2 Open the **millscripsts\libSpecMill\lsmGlobals.pm** file in a plain text editor (e.g., Notepad) to edit it.
- 3 Change the following line (near line 36):
 - From: **\$GLOBAL_ADD_BLAST_LINK = 0;**
 - To: **\$GLOBAL_ADD_BLAST_LINK = 1;**
- 4 Save the file.

If you enable the BLAST link, be aware that it will take longer to load large data sets in Protein/Peptide Summary.

To add/change options related to biology/chemistry

The Spectrum Mill workbench allows you to modify a number of parameters related to the biology or chemistry of your study. To customize these options, you modify text files in the folder **msparams_mill**.

- 1 Archive each file before you modify it.
- 2 Edit the file(s). See the online help document **servadmn.htm** for details. For edits that involve **smconfig.custom.xml**, see also [“To add custom amino acid modifications”](#) on page 233.

The biology/chemistry options that you can change include:

- Species filter (**species.txt**)
- Cysteine modification options (**smconfig.custom.xml**)
- N- or C-terminal groups (**smconfig.custom.xml**)
- Other custom amino acid modifications, including those used for differential expression quantitation (**smconfig.custom.xml**)
- Elements (**smconfig.custom.xml**)
- Enzyme cleavage rules (**enzyme.txt**)
- Categories of proteins (**categories.#.tsv**)

To add instrument types

You can add instrument types and configure MS/MS Search scoring and peak detection parameters for these instruments.

See the online help document **servadmn.htm** for details.

- 1 Archive these files before you modify them:
 - **msparams_mill\instrument.txt**
 - **millhtml\SM_js\instrument.js**
- 2 Edit the files.

To maintain server performance

To maintain server performance:

- Periodically archive data that is no longer needed; remove it from the server and defragment the data drive.

Use the Tool Belt feature for archiving data to zip up files in folders that are not necessary for your current projects. You can unzip them if you need to look at or reprocess the data later.

- Always defragment the database drive after installing a new database.

Spectrum Mill data directories contain many very small files. Over time, disk fragmentation from accumulated files causes newly-added raw data files to become heavily fragmented. If you fail to defragment the data drive, Data Extractor speed declines.

If you fail to defragment the database drive, search speeds decline.

To use server administration scripts

The **milladmin** folder provides scripts that system administrators may use to configure the Spectrum Mill installation. These scripts:

- Configure IIS for the Spectrum Mill workbench
- Set file permissions for critical Spectrum Mill files
- Set file access permissions for **msdataSM** and subfolders
- Create data file placeholders to allow you to remove raw data files

See the online help document **servadmn.htm** for details.

To avoid problems with connection time-outs

The default Internet Information Server (IIS) settings applied by the Spectrum Mill installation may not be suitable when you process large data sets with long extraction and search times. To avoid having your connections time out, change IIS settings on the Spectrum Mill server. For new settings, see the online help for server administration (**servadmn.htm**).

To configure server with drives other than the default configuration

- 1 Read the discussion below.
- 2 If the hard drive configuration changes (e.g. additional hard drive installed), then you need to uninstall and reinstall the Spectrum Mill software to see new partitions. If you need to move the collected data to a new data storage path, you must do this manually.

The Spectrum Mill Install Shield program examines your current drive configuration and allows you to select from the partitions available where to install the Spectrum Mill program and example data. The default choice is the largest partition found. This path is highly recommended since all data used with the Spectrum Mill program must reside in the same partition as the program, and because the data will require the most space for storage.

The software installation also allows you to designate the location for the database(s). By default, the second largest partition is selected, but as with the program and data, you can change this designation. The second largest partition is recommended for installation of the databases because they are typically a few hundred megabytes in size. You should assign the Windows® Paging File location to the same partition since few disk writes will occur that could lead to fragmentation of the Paging File.

To remove and reinstall the Spectrum Mill workbench

If you wish to remove and reinstall the Spectrum Mill workbench, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

If you remove the Spectrum Mill workbench and intend to reinstall it, note the following:

- Software removal does not delete the data folders (**msdataSM**), the protein databases (**seqDB**), or the workflow-related files (**millauto**).
- Be sure to archive and restore any configuration files that have been customized for your site. These might include:
 - **millhtml\sm_js** (**instrument.js**, **massOpt.js**, **protChem.js**, **species.js**, **iTRAQbatches.js**)
 - **millscripts\libSpecMill\lsmGlobals.pm**
 - **mparams_mill** (various *.txt files, including contaminant mass list files)

To limit the number of parallel workflow processes to less than the CPU count

Before you modify any files, be sure to archive the originals so they can be restored if necessary!

When you execute a workflow, its tasks are added to the Spectrum Mill Request Queue, and executed in order for a particular data folder. If you select multiple data folders, the data folders are processed in parallel (extraction and search tasks) if you have more than one CPU. The number of parallel tasks is limited by the number of (logical) processors on the server. If you want to limit the maximum number to less than the CPU count, edit the file:

E:\SpectrumMill\millsrc\SMSRM.config

- 1** Find the line that contains **maxConcurrentTasks**. The line of text is similar to this example: **<provider hostname="localhost" available="true" maxConcurrentTasks="2">**
- 2** Change the number after **maxConcurrentTasks** and resave the file.
- 3** Stop and restart the SRM service for the new setting to take effect.

See [“To start and stop the Spectrum Mill Workflow Manager Service”](#) below.

To start and stop the Spectrum Mill Workflow Manager Service

Spectrum Mill uses a Windows Service Request Manager (SRM) service to process workflows. The "Spectrum Mill Workflow Manager Service" is configured to automatically start when Windows starts.

To stop the service:

- 1 Wait until the Request Queue is empty. (Anything in the queue will be lost.)
- 2 Navigate to **E:\SpectrumMill\millsrm**.
- 3 Double-click **stopSRM.bat**.

If the service continues to run, as a last resort, stop it with Windows Task Manager. In the **Processes** tab of Task Manager, click **SRMHostSvc.exe**, and then click **End Process**.

To start the service:

- 1 Navigate to **E:\SpectrumMill\millsrm**.
- 2 Double-click **startSRM.bat**.



10 Files Created during Spectrum Mill Data Processing

Architecture Overview	242
Data Extractor (MS/MS raw data)	244
Data Extractor (generic data)	246
Data Extractor (MS-only raw data)	248
MS/MS Search	249
Protein/Peptide Summary, Spectrum Summary, and Autovalidation	251
Tool Belt	253
Sherenga de novo Sequencing	255
PMF Search	256
MS Edman	257
Protein Databases	258

This chapter begins with an architecture overview and then lists the files the Spectrum Mill workbench creates as you process results from analyses of protein digests. The file lists are organized by data processing module. Since the Spectrum Mill modules automatically create and call these files as needed, you don't need to know about them to use the software. They are listed as an aid to troubleshoot data processing, to selectively remove parts of the processing, and to decide which files to archive.



Architecture Overview

The following figures show an overview of the Spectrum Mill architecture, including programs, scripts, and some of the files that are produced. For more details on the files, see the tables that follow.

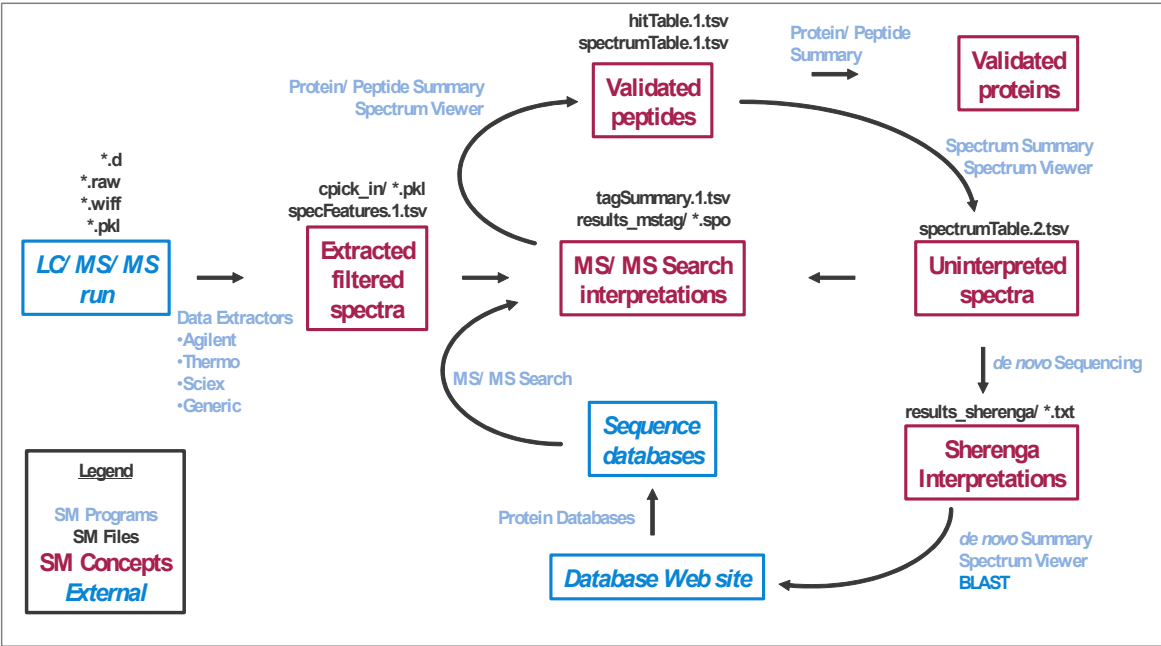


Figure 126 Architecture for MS/MS workflow

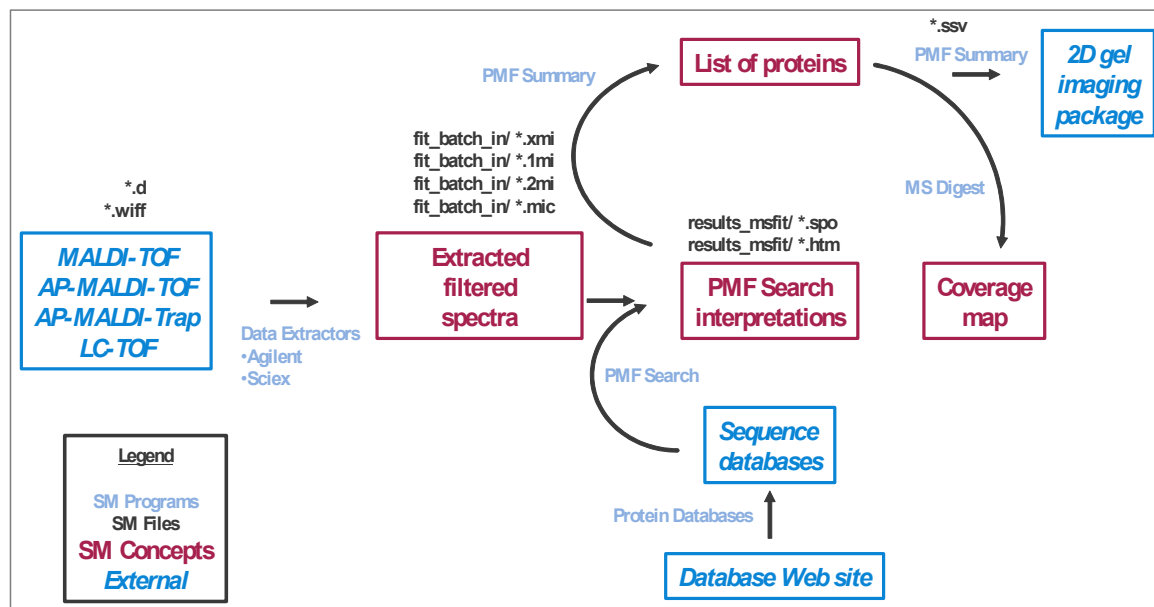


Figure 127 Architecture for MS-only workflow

Data Extractor (MS/MS raw data)

Table 8 lists the files created when you run the MS/MS raw data extractor. These are all created under the **msdataSM** directory, in the subdirectory that contains your raw data files. All are text files; you can open them with Notepad.

Table 8 Files created by MS/MS raw data extractor

Files and Folders Created	Contents	Notes
parentChargeCheck.txt	• Scan number in format Data_File_Name.#####, where ##### is the scan number • Charge assignment for precursor ion • Precursor ion • Number of b/y pairs found	
xtractorRejectReasons.txt	• Scan number in format Data_File_Name.##### for single scan or Data_File_Name.aaaa.bbbb for merged scans, where • aaaa = first merged scan • bbbb = last merged scan • Reject reasons. Examples are: • Scans merged based on precursor ion and time window tolerances • Minimum sequence tag length requirement not met • Precursor ion out of range specified for MH⁺ • <4 peaks after picking	• Useful to troubleshoot data extraction problems • For Applied Biosystems/MDS Sciex instruments, aaaa = (1/10 x retention time in seconds).
runStats.txt	• Various statistics for the extractor run, including numOutputFiles, which is the number of extracted spectral files entered in the subdirectory cpick_in	
SpecFeatures.1.tsv	• Spectral features calculated for each spectrum in the subdirectory cpick_in	
xtractor.1.params	• Record of data extractor parameters	

Table 8 Files created by MS/MS raw data extractor

Files and Folders Created	Contents	Notes
cpick_in*.pkl	- Series of files - one for each spectrum extracted - File name in format: Data_File_Name.aaaa.bbbb.c.pkl, where - aaaa = first merged scan - bbbb = last merged scan - c = assigned precursor charge (0 means charge was ambiguous)	- If you rerun the Data Extractor using a new set of extraction parameters, mark the check box to Remove all prior results. - For Applied Biosystems/MDS Sciex instruments, aaaa = (1/10 x retention time in seconds).
Series of folders in the subdirectory that contains your data file(s). Folders are listed in next column.	- cpick_in - fit_batch_in - results_msedman - results_msfit - results_mstag - results_sherenga	
*.aph for .d (Agilent) files *.rph for .raw (Thermo) files *.wph for .wiff files *.gph for .mgf (generic extractor) *.lnk for a linked archive file	These are placeholder files that correlate extracted spectra with original raw data file.	File name in format: Data_File_Name.xxx, where "xxx" is the file suffix.

Data Extractor (generic data)

Table 9 lists the files created when you run the generic data extractor. These are all created under the **msdataSM** directory, in the subdirectory where you placed your appended *.pkl spectral files, (or one level above the **cpick_in** subdirectory where you placed your individual *.pkl or *.dta spectral files). All are text or html files.

Table 9 Files created by generic data extractor

Files and Folders Created	Contents	Notes
parentChargeCheck.txt	• Headers for scan number and charge assignment • No other data	
xtractorRejectReasons.txt	• Spectral file name • Reject reasons	• Useful to troubleshoot data extraction problems
runStats.txt, runStats.htm	• Empty files	
SpecFeatures.1.tsv	• Spectral features calculated for each spectrum in the subdirectory cpick_in	
qtofXtractor.1.params	• Record of data extractor parameters	

Table 9 Files created by generic data extractor

Files and Folders Created	Contents	Notes
cpick_in*.pkl	- Series of files - one for each spectrum extracted - created when you extract an appended *.pkl file that contains multiple spectra - File name in format: Data_File_Name.Scan_Number.0. Precursor_Charge.pkl, where - Data_File_Name = prefix of appended *.pkl file - Scan_Number = consecutive order of the scan in the appended *.pkl file 0 = placeholder where function number would be if created by Micromass ProteinLynx software - Precursor_Charge = charge of the precursor ion	If you rerun the Data Extractor using a new set of extraction parameters, mark the check box to Remove all prior results.
Series of folders in the subdirectory that contains your data file(s). Folders are listed in next column.	- cpick_in - fit_batch_in - results_msedman - results_msfit - results_mstag - results_sherenga	

Data Extractor (MS-only raw data)

Table 10 lists the files created when you run the MS-only raw data extractor. These are all created under the **msdataSM** directory, in the subdirectory that contains your raw data files. All are text files; you can open them with Notepad.

Table 10 Files created by MS-only raw data extractor

Files and Folders Created	Contents	Notes
parentChargeCheck.txt	• Headers for scan number and charge assignment • No other data	
xtractorRejectReasons.txt	• Headers for spectrumName, numScansAfterParent, and RejectReason • No other data	
runStats.txt	• Information such as fileName and sampleName	
SpecFeatures.1.tsv	• Various headers • No other data	
xtractor.1.params	• Record of data extractor parameters	
fit_batch_in*.mi or fit_batch_in*.mic	• Extracted file • File name in format: Data_File_Name.*mi or Data_File_Name.mic	• If you rerun the Data Extractor using a new set of extraction parameters, mark the check box to Remove all prior results.
Series of folders in the subdirectory that contains your data file(s). Folders are listed in next column.	• cpick_in • fit_batch_in • results_msedman • results_msfit • results_mstag • results_sherenga	

MS/MS Search

Table 11 lists the files created when you run MS/MS Search. These are all created under the **msdataSM** directory, in the subdirectory that contains your raw data files. All are either text or html files.

If you accidentally run a search with incorrect search settings, you should remove the associated files from your system, especially the files in the **results_mstag** subdirectory. Otherwise, they will be included in future results reviews, validations, and summaries. If you have run multiple searches, be sure to check the time and date stamps on the files so that you do not mistakenly remove good search results.

Table 11 Files created by MS/MS Search

Files Created	Contents	Notes
tagSummary.#.tsv	Summarizes all search results	- New file created each time you run MS/MS Search. Files are numbered tagSummary.1.tsv, tagSummary.2.tsv, etc. - You also create this file when you click the Create MS/MS search summary file button on the Tool Belt page.
mstag.#.params	Record of MS/MS Search parameters	- New file created each time you run MS/MS Search. Files are numbered mstag.1.params, mstag.2.params, etc. - When you search an ion trap data file that contains both CID and ETD spectra, then two additional files are created because there are two sets of search conditions. The additional files are of the formats mstag.1a.params and mstag1b.params.
mstagSpeedLog.tsv	Speed statistics for MS/MS Search run	File is not overwritten unless you mark the check box for Remove all prior MS/MS Search results. For iterative searches that use the same data set, the new results are appended to the end of the file.

Table 11 Files created by MS/MS Search

Files Created	Contents	Notes
results_mstag*.pkl.spo	- Series of text search results files - one for each extracted spectrum searched. (Some extracted spectra do not meet quality criteria for searching.) - File name in format: Data_File_Name.aaaa.bbbb.c.pkl.spo, where - aaaa = first merged scan - bbbb = last merged scan - c = assigned precursor charge (0 means charge was ambiguous)	- Contents of results_mstag subdirectory are updated each time you run MS/MS Search on the same file set. - For Applied Biosystems/MDS Sciex instruments, aaaa = (1/10 x retention time in seconds).
results_mstag\junk.#.htm	Scratch file of html output before redirection to browser	New file created each time you run MS/MS Search. Files are numbered junk.1.htm, junk.2.htm, etc.
fdr"type".#.tsv, where "type" is the name extension for the .tsv file; e.g., fdrSummary.1.tsv	A series of FDR files created in MS/MS search and used by discriminant-based FDR	

Protein/Peptide Summary, Spectrum Summary, and Autovalidation

Table 12 lists the files created when you validate data with the Protein/Peptide Summary page, the Spectrum Summary page, or the Autovalidation page. These are all created under the **msdataSM** directory, in the subdirectory that contains your raw data files. All are text files.

Note that correlated files from a given validation step have the same time and data stamp, but not necessarily the same number in the file sequence. For example, **autovalidation.1.params** may be correlated with **spectrumTable.3.tsv** and **hitTable.2.tsv**. This is because multiple modules write to subsets of these files.

Table 12 Files created by Protein/Peptide Summary, Spectrum Summary, and Autovalidation

Files Created	Contents	Notes
hitTable.#.tsv	Record of MS/MS Search hit validation states	- Created when you click the Perform Validation button on the Protein/Peptide summary page or the Validate Files button on the Autovalidation page - New file created each time you click the button. Files are numbered hitTable.1.tsv, hitTable.2.tsv, etc. - Results are cumulative; each new file starts with the previous file and adds information.

Table 12 Files created by Protein/Peptide Summary, Spectrum Summary, and Autovalidation

Files Created	Contents	Notes
spectrumTable.#.tsv	- Record of spectrum validation states, including: - whether the spectrum has been designated good or bad - whether the spectrum has a validated database hit	- Created when you click the Update button on the Spectrum Summary page or the Perform Validation button on the Protein/Peptide summary page or the Validate Files button on the Autovalidation page - New file created each time you click the button. Files are numbered spectrumTable.1.tsv, spectrumTable.2.tsv, etc. - Results are cumulative; each new file starts with the previous file and adds information.
autovalidation.#.params	Record of autovalidation parameters	- Created when you click the Validate Files button on the Autovalidation page - New file created each time you click the button. Files are numbered autovalidation.1.params, autovalidation.2.params, etc.

Tool Belt

[Table 13](#) lists the files created with the Tool Belt page. All are text files.

Table 13 File created by Tool Belt

File Created	Contents	Notes
*.res	- Results file of validated rank one hits - File name in format: Database_Name_with_any_extensions.res	- Created when you select the Create saved results file option and click the Create File button - Created in the results_mstag folder within the subdirectory that contains your raw data files - Overwritten each time you create a saved results file for that database and data directory
tagSummary.#.tsv	Summarizes all search results	- Created when you select the Create MS/MS search summary file option and click the Create File button - Created in the subdirectory that contains your raw data files - The MS/MS Search module also writes to this file. See Table 11 on page 249.
iTRAQ.corrections	iTRAQ correction factors and batch numbers	- Created when you select the Create iTraQ correction factors option, type a batch number and correction factors, and click the Create button - Created under the mparams_mill directory - File is appended every time you type correction factors for a new batch number - Software will not allow you to overwrite existing correction factors for a batch

Table 13 File created by Tool Belt

File Created	Contents	Notes
Discriminant Coefficients.txt	Includes predefined sets for Agilent discriminant coefficients and is updated when new coefficients are calculated.	
iTRAQ.corrections.txt	iTRAQ correction factors that apply to a particular sample or set of samples	- Created when you select the Apply iTRAQ correction factors option, set up the page, and click the Apply button - Created under the msdataSM directory, in the subdirectory that contains your raw data file(s)

Sherenga *de novo* Sequencing

Table 14 lists the files created when you run Sherenga *de novo* sequencing. These are all created under the **msdataSM** directory, in the subdirectory that contains your raw data files. All are text files.

Table 14 Files created by MS/MS Search

Files Created	Contents	Notes
sherenga.#.params	Record of Sherenga <i>de novo</i> Sequencing search parameters	New file created each time you run Sherenga <i>de novo</i> Sequencing. Files are numbered sherenga.1.params, sherenga.2.params, etc.
results_sherenga\#.pkl.txt	- Series of text Sherenga results files - one for each extracted spectrum processed. - File name in format: Data_File_Name.aaaa.bbbb.c.pkl.txt, where - aaaa = first merged scan - bbbb = last merged scan - c = assigned precursor charge (0 means charge was ambiguous)	- Contents of results_sherenga subdirectory are updated each time you run <i>de novo</i> sequencing on the same file set. - For Applied Biosystems/MDS Sciex instruments, aaaa = (1/10 x retention time in seconds).

PMF Search

Table 15 lists the files created when you run PMF Searches. These are all created in the subdirectory that contains your spectral files, which is under the **msdataSM** directory. All are either text or html files.

Table 15 Files created by PMF Search

Files and Folders Created	Contents	Notes
msfit.#.params	Record of PMF Search parameters	New file created each time you run PMF Search. Files are numbered msfit.1.params, msfit.2.params, etc.
results_msfit*.MI.htm or results_msfit*.MIC.htm	- Series of html search results files - one for each extracted spectrum searched - File name formats: - Peak_File_Name.MI.htm - Data_File_Name.MI.htm - Data_File_Name.MIC.htm	Contents of results_msfit subdirectory are updated each time you run PMF Search on the same file set.
results_msfit*.MI.spo or results_msfit*.MIC.spo	- Series of text search results files - one for each extracted spectrum searched - File name formats: - Peak_File_Name.MI.spo - Data_File_Name.MI.spo - Data_File_Name.MIC.spo	Contents of results_msfit subdirectory are updated each time you run PMF Search on the same file set.
Series of folders in the subdirectory that contains the data file. Folders are listed in next column.	- cpick_in - fit_batch_in - results_msedman - results_msfit - results_mstag - results_sherenga	
results_msfit\junk.#.htm	Diagnostic html file created when PMF Search runs	New file created each time you run PMF Search. Files are numbered junk.1, junk.2, etc.

MS Edman

Table 16 lists the files created when you run MS Edman *and* you mark the check box to **Save hits to file**. These files are created under the **results_msedman** subdirectory directly under your Spectrum Mill installation. (They are not under a data file subdirectory.)

Table 16 Files created by MS Edman

Files Created	Contents	Notes
lastres.htm	- Search results in html format - File name is lastres or whatever you have chosen in the MS Edman page.	Contents are overwritten unless you change the file name in the MS Edman page.
lastres.spo	- Empty file - File name is lastres or whatever you have chosen in the MS Edman page.	
lastres.par	- Search parameters in text format - File name is lastres or whatever you have chosen in the MS Edman page.	Contents are overwritten unless you change the file name in the MS Edman page.
lastres.res	- List of accession numbers that correspond with MS Edman search hits - File name is lastres or whatever you have chosen in the MS Edman page.	Contents are overwritten unless you change the file name in the MS Edman page.

Protein Databases

Protein Databases creates files in the directory where you store your sequence databases (usually **X:/seqDB**, where X is the disc drive letter).

Table 17 Files created by Protein Databases

Files Created	Contents	Notes
*.sub	• Subset database created from species and/or molecular weight information contained in the original database • File name includes sub or whatever suffix you have chosen in the Protein Databases page.	• Created when you select the Create species subset database option and click the Create button
*.lastres	• Subset database created from hits saved in MS Edman • File name includes lastres or whatever extension you have chosen in the Protein Databases page.	• Created when you select the Create subset with indices from saved hits option and then click the Create button
*.user	• User database • File name includes user or whatever extension you have chosen in the Protein Databases page.	• Created when you select the Create or append user database option and then click the Create button
*.idx, idx2 (for databases > 4.2 Gb) *.unk *.mw *.pi *.sp *.sl *.usp *.acc or *.acn	• Index files	• Created when you click the Create button after you select any of these options on the Protein Databases page: • Create indices for new database • Create species subset database • Create subset with indices from saved hits • Create or append user database • See the online help document faman.htm for details regarding these files.

Index

A

abort process, [176](#)
 Agilent Q-TOF
 Data Extractor, [13 to 16](#)
 manual data validation, [84, 90 to 91](#)
 Agilent TOF
 acquiring data (wiff files only), [154](#)
 amino acid filtering, [19, 37, 42](#)
 amino acid modifications
 customizing, [233](#)
 list details, [191](#)
 AMRT export, [88](#)
 architecture overview, Spectrum Mill, [242](#)
 archive data, [200](#)
 automated workflow
 completion log, [128](#)
 create/edit, [122](#)
 execute, [124](#)
 view request queue, [126](#)
 automated workflows, [119 to 128](#)
 automation
 interactive, [13](#)
 Autovalidation, [23 to 27, 73 to 80](#)
 auto thresholds - discriminant strategy, [75](#)
 auto thresholds strategy, [74](#)
 auto thresholds-discriminant strategy, [40, 47](#)
 false discovery rate (FDR), [25](#)
 fixed thresholds strategy, [76](#)
 local FDR, [40, 47](#)
 strategy/mode table, [25](#)
 autovalidation
 global FDR, [40, 47](#)

B

basic workflow, [12, 13 to 31](#)
 Build TIC, [70](#)

C

Choose button, [15, 18, 19, 37, 43, 62, 132, 141, 143, 156, 158, 160, 162, 204, 210, 233](#)
 collection of spectra and interpretations, [190](#)
 comparing two samples, [56](#)
 completion log, [128](#)
 connection time-outs, resolving, [236](#)
 contaminant mass list files, [162, 238](#)
 convert spectra
 pkl files into mzXML files, [199](#)
 copy spectra to collections directory, [190](#)
 create a saved results file, [178](#)
 create discriminant scoring coefficients, [195](#)
 create library for Spectrum Matcher, [190](#)
 create MS/MS search summary file, [180](#)
 create saved results file, [41](#)
 create subset database from saved hits, [227](#)
 create/edit an automated workflow, [122](#)
 custom amino acid modifications, [233](#)

D

data display mode, [81, 104 to 118](#)
 Data Extractor
 generic peak list files, [16 to 19](#)
 MS/MS raw files, [13 to 16](#)
 MS-only raw files, [156 to 161](#)
 data file placeholders, [236](#)
 database
 previous hits, [178](#)
 saved results, [178](#)

database search

 homology mode, [33 to 34, 59 to 60](#)
 identity mode, [36 to 39](#)
 manual PMF, [169 to 174](#)
 PMF, [162 to 163](#)
 unknown modification, [34, 59](#)
 variable modifications mode, [19 to 23, 33 to 34, 42 to 47](#)

database search modes, [34](#)

database summary report, [229](#)

databases

 creating indices, [224](#)
 installing, [222](#)
 molecular weight, [226](#)
 species, [226](#)
 subset, [227](#)
 updating, [222](#)
 user-created, [223, 228](#)

de novo sequencing, [60 to 66, 129 to 138](#)

DEQ data

 Data Extractor, [141](#)
 MS/MS Search, [143](#)
 peptide mode, [145](#)
 protein mode, [147](#)
 Protein/Peptide Summary, [144](#)

differential expression quantitation, [87, 139 to 151, 235](#)

discriminant scoring

 coefficients, [38, 195](#)

dynamic peak thresholding, [20, 45, 91](#)

E

evaluating results

 MS/MS Search, [89 to 103](#)
 PMF Search, [164 to 167](#)

Excel, [57, 118, 167, 194](#)

export

 AMRT, [88](#)
 Excel, [88](#)
 pepXML files, [197](#)

F

false discovery rate (FDR), [74](#)
 autovalidation, [25](#)
 global FDR, [40](#)
 local FDR, [40](#)
 Tool Belt search statistics, [186 to 189](#)
 file
 permissions, [236](#)
 placeholders, [236](#)
 files created
 Autovalidation, [251](#)
 Data Extractor (generic data), [246](#)
 Data Extractor (MS/MS raw data), [244](#)
 Data Extractor (MS-only data), [248](#)
 MS Edman, [257](#)
 MS/MS Search, [249](#)
 PMF Search, [256](#)
 Protein Databases, [258](#)
 Protein/Peptide Summary, [251](#)
 Sherenga *de novo* sequencing, [255](#)
 Spectrum Summary, [251](#)
 Tool Belt, [253](#)
 filtering search results, [83 to 85](#)
 filtering spectra, [47 to 51](#)
 folder selection, [14, 17, 19, 27, 37, 42, 48, 62, 89, 156, 157, 160](#)

G

global FDR, [40, 47](#)
 global files, [238](#)
 graph theory, [129](#)
 guidelines
 manual results validation, [89 to 103](#)

H

homology mode searches, [33 to 34, 59 to 60](#)
 HTML link to BLAST search, [234](#)
 HTML links to search results, [234](#)

I

ICAT data
 Data Extractor, [141](#)
 general discussion, [139 to 151](#)
 MS/MS Search, [143](#)
 peptide mode, [145](#)
 protein mode, [147](#)
 Protein/Peptide Summary, [144](#)
 Spectrum Summary, [151](#)
 identification statistics
 Build TIC, [70](#)
 keeping track, [52](#)
 summary table, [183](#)
 identity mode searches, [36 to 39](#)
 IIS configuration, [236](#)
 indices, [224](#)
 installation, [238](#)
 installing databases, [222](#)
 interactive automation, [13](#)
 isotope-coded affinity tag data. *See* ICAT data.
 isotopically labeled data, [139 to 151](#)
 iterative MS/MS data processing, [32 to 66](#)
 iterative workflow, [12, 32 to 66](#)
 iTRAQ data
 correction factors, [148 to 150, 193 to 194](#)
 Data Extractor, [141](#)
 MS/MS Search, [143](#)
 overview, [139 to 140](#)
 peptide mode, [149](#)
 protein mode, [150](#)
 Protein/Peptide Summary, [148](#)

J

journal submission, [110](#)

L

light/heavy isotope ratios, [139 to 151](#)
 LIMS, [57, 167](#)
 link to results, [125](#)

list details of amino acid
 modifications, [191](#)
 local FDR, [40, 47](#)

M

MALDI-TOF-TOF
 Data Extractor, [16](#)
 manual PMF searches, [169 to 174](#)
 manual results validation
 Agilent Q-TOF, [84, 90 to 91](#)
 ion trap CID data, [89 to 103](#)
 manual validation, [89 to 94](#)
 manuscript for journal, [110](#)
 mass gap search, [20, 34, 59](#)
 Maximum ambiguous precursor
 charge, [37](#)
 maximum sequence tag length, [47, 48](#)
 mixture scoring, [162](#)
 mode
 Peptide, [105](#)
 Protein - Single Peptide ID, [110](#)
 Protein Summary, [106](#)
 Protein Summary Details, [108](#)
 Protein-Peptide Comparison
 Columns, [118](#)
 Protein-Peptide Distribution
 Columns, [117](#)
 Protein-Protein Comparison
 Columns, [111](#)
 Protein-Protein Comparison
 Redundant, [114](#)
 Protein-Sample Centric Rows, [115](#)
 Protein-Sample Centric Rows
 Details, [116](#)
 modes, data display
 table, [82](#)
 modifications, customizing, [233](#)
 molecular weight subset databases, [226](#)
 monitor results, [18, 125](#)
 MRM lists for triple quadrupole
 instruments, [207](#)
 MRM Selector, [207](#)
 MS Comp, [216](#)
 MS Digest, [210](#)

MS Edman, 212
 MS Isotope, 218
 MS Product, 214
 MS/MS data processing
 iterative, 32 to 66
 MS/MS search
 discriminant scoring, 38
 homology mode, 33 to 34, 59 to 60
 identity mode, 36 to 39
 larger database, 58
 no enzyme mode, 58
 search modes, 34
 summary file, 180
 unknown modification, 34, 59
 variable modifications mode, 19 to 23, 33 to 34, 42 to 47
 MS-only data processing
 Agilent TOF and Q-TOF .d files, 169 to 174
 wiff files, 155 to 168
 MS-only search
 automated, 162 to 163
 manual, 169 to 174
 Multiple Sequence Aligner, 209
 mzXML files, 199

N

neutral loss, 70, 91
 no enzyme, 51
 no enzyme mode search, 58

O

offgel fractionation, 92

P

parameter files, 47, 120
 MS/MS, 13
 MS-only, 155
 parameters
 keeping track, 52
 summary table, 181
 Peak Picker, 69
 peptide charge, 101, 103

peptide list to masses utility, 219
 Peptide mode, 105
 peptide pl filtering, 92
 Peptide Selector, 204
 pepXML files, 197
 phosphorylation, 20, 33, 44, 70
 pl, 28, 77, 92
 placeholders, data files, 236
 PMF searches, 162 to 163
 PowerPoint, 58, 168
 print results, 57, 168
 proprietary databases, 223, 228
 Protein - Single Peptide ID mode, 110
 protein and peptide utilities, 204 to 219
 protein grouping, 84
 protein quantitation options, 87
 Protein Summary Details mode, 108
 Protein Summary mode, 106
 Protein/Peptide Summary, 81 to 88
 protein quantitation options, 87
 Protein/Peptide Summary displays, 104 to 118
 table, 82
 Protein/Peptide Summary modes
 table, 82
 Protein-Peptide Comparison Columns
 mode, 118
 Protein-Peptide Distribution Columns
 mode, 117
 Protein-Protein Comparison Columns
 mode, 111
 Protein-Protein Comparison Redundant
 mode, 114
 Protein-Sample Centric Rows Details
 mode, 116
 Protein-Sample Centric Rows mode, 115

Q

Q-TOF
 Data Extractor, 13 to 16
 manual data validation, 84, 90 to 91

R

rank1 - rank 2 scores, 91
 report FDR and search statistics, 186 to 189
 request queue, 18
 results
 monitor, 18
 results validation
 guidelines, 89 to 103
 parameters, 83 to 85
 reversed database scores, 20, 37, 44, 91
 review fields
 PMF Summary, 164
 Protein/Peptide Summary, 86
 Spectrum Summary, 47

S

scripts, server administration, 236
 search modes, 34, 212
 search previous hits, 41, 178, 227
 search results
 changing HTML links, 234
 enabling HTML link to BLAST, 234
 Select button, 14, 17, 19, 27, 37, 42, 48, 62, 89, 156, 157, 160
 sequence alignment, 209
 sequence tag length, 47, 48
 sequencing - *de novo*, 60 to 66, 129 to 138
 server
 administration scripts, 236
 drives, 237
 performance, 236
 Sherenga *de novo* sequencing, 60 to 66, 129 to 138
 SILAC, 140 to 147
 smconfig.xml, 233
 sorting search results, 83 to 85
 sorting spectra, 47 to 51
 species subset databases, 226
 spectral validation parameters, 47 to 51
 Spectrum Matcher, 67
 library, 190

Index

Spectrum Mill server
 configuring drives, 237
 maintaining performance, 236
Spectrum Summary, 47 to 51
Spectrum Viewer, 96 to 100
stop process, 176
subset database, 37
subset database for variable modifications
 or homology search, 178, 227
Summarize valid results, 27
summarize valid results, 52 to 58, 164 to
 165, 167 to 168
summary
 database, 229
 identification statistics, 183
 parameters used, 181
summary mode, 81, 104 to 118

system administration
 adding instrument types, 235
 calculating statistics, 231
 changing biology/chemistry
 options, 235
 changing categories of proteins, 235
 changing cysteine modification
 options, 235
 changing elements, 235
 changing enzyme cleavage rules, 235
 changing HTML links from search
 results, 234
 changing N- or C-terminal groups, 235
 changing species filter, 235
 comparing two databases, 230
 configuring server drives, 237
 connection time-outs, 236
 creating indices for databases, 224
 creating species or molecular weight
 subset databases, 226
 creating subset databases from saved
 hits, 227
 creating user (proprietary)
 databases, 223, 228
 customizing amino acid
 modifications, 233
 enabling HTML link from search results
 to BLAST, 234
 generating database summary
 report, 229
 installing or updating databases, 222
 limiting number of parallel workflow
 processes, 239
 maintaining server performance, 236
 making a subset FASTA file from
 Accession Numbers, 232
 reinstalling Spectrum Mill
 workbench, 238
 server administration scripts, 236
 settings for large data sets, 236
 starting and stopping the Spectrum Mill
 Workflow Manager Service, 240

T

terminate process, 176

time-outs, connection, 236
TMT data
 Data Extractor, 141
 MS/MS Search, 143
 overview, 139 to 140
 Protein/Peptide Summary, 148
TOF
 acquiring data (wiff files only), 154
Tool Belt utilities, 175 to 200

U

unknown modification search, 34
unzip files, 200
updating databases, 222
user-created databases, 223, 228
utilities
 protein and peptide, 204 to 219
 Tool Belt, 175 to 200

V

valid results
 print, 57, 168
 summarize, 52 to 58, 164 to 165, 167
 to 168
validation
 automatic, 23 to 27, 73 to 80
 manual, 89 to 94
 spectra, 47 to 51
variable modifications search, 19 to 23, 33
 to 34, 42 to 47
Variable modifications sites, 30
view request queue, 18, 125, 126

W

workflow automation, 119 to 128
 completion log, 128
 link to results, 125
 monitor results, 125
 view request queue, 125, 126
 Workflow Manager Service, 240
workflow tasks, 120

workflows

- automation, [119 to 128](#)
- basic, [13 to 31](#)
- basic and iterative, [12](#)
- iterative, [32 to 66](#)
- parameter files, [120](#)
- tasks, [120](#)

Z

- zip files, [200](#)

In this Book

The *Application Guide* presents basic instructions to use the Spectrum Mill MS Proteomics Workbench. In this guide you learn:

- How to process MS/MS and MS-only data sets
- How to automate Spectrum Mill workflows
- How to perform Sherenga *de novo* sequencing
- How to process data for differential expression quantitation
- How to use the software's protein/peptide analysis tools and utilities
- System administration details

© Agilent Technologies, Inc. 2012

Printed in USA
Revision A, January 2012



G2721-90037



Agilent Technologies