Agilent G2721AA Spectrum Mill MS Proteomics Workbench

Application Guide

Agilent Technologies

Notices

© Agilent Technologies, Inc. 2007

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

G2721-90029

Edition

Fifth edition, June 2007

Printed in USA

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

If you have any comments about this guide, please send an e-mail to feedback_lcms@agilent.com.

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 the basics to analyze protein and peptide data with the Spectrum Mill MS Proteomics Workbench. In this guide you will learn:

- How to process MS/MS and MS-only data sets
- Details about the advanced Spectrum Mill workbench data review and validation capabilities
- How to perform Sherenga *de novo* sequencing
- How to process ICAT, iTRAQ, and other data for differential expression quantitation
- How to use the many protein/peptide analysis tools and utilities that are included with the Spectrum Mill workbench
- System administration details

If you have any comments about this guide, please send an e-mail to feedback_lcms@agilent.com.

1 Basics for Processing MS/MS Data

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

2 MS/MS Data Review and Validation

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

3 Sherenga *de novo* Sequencing

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

4 Processing Ion Trap - ETD Data

Learn how to process electron transfer dissociation (ETD) data from an Agilent ion trap.

5 Processing ICAT, iTRAQ, and Other Data for Differential Expression Quantitation

Learn details to process quantitative data for differential profiling studies.

6 Basics for Processing MS-Only Data

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 Basics for Processing MS/MS Data

In this chapter you learn the basics to process multiple MS/MS data files with the Spectrum Mill MS Proteomics Workbench. This chapter guides you through a typical workflow using the software. You use this workflow for any MS/MS data, whether it originated with LC/MS/MS or MALDI MS/MS. While the steps described here are representative of what many users want to do, there are certainly other paths through the software. As you gain experience, you may want to modify your process to best suit your experiments and needs. This manual and the online help give you the details you need to customize this workflow.

This chapter assumes that the Spectrum Mill workbench has already been installed on your server, databases have been downloaded and indexed, and the system is ready to go. If this has not yet been done, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

If you want more details about the Protein/Peptide Summary page described in this chapter, see Chapter 2, “MS/MS Data Review and Validation”. For more details about Sherenga *de novo* sequencing, see Chapter 3, “Sherenga de novo Sequencing”. To learn more about processing ICAT and similar data, see Chapter 5, “Processing ICAT, iTRAQ, and Other Data for Differential Expression Quantitation”.

Iterative search strategy

Introduction

The Spectrum Mill workbench allows you to process your data through multiple rounds of database search and results validation. Figure 1 shows a roadmap for processing MS/MS data. How much of this roadmap you complete depends upon how important it is to extract every last bit of information from your sample.

For some samples, it is sufficient to process only through the basic workflow indicated by the darker (blue) shading. These steps are covered in "First-pass data processing" on page 17. For others, you take advantage of the optional processing indicated by the lighter (yellow) shading. These additional steps are covered in "Optional data processing to extract additional information" on page 49.

As you process data with the Spectrum Mill workbench, you may iterate through multiple rounds of database search and results validation, with the goal of identifying as many spectra as possible. 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. Results that are not validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that are validated can be summarized in a results table.

CAUTION The default settings for the Spectrum Mill workbench have been carefully optimized. In general, you should change settings only for the settings that are highlighted in red text. If you change other settings, be sure you are confident in your choices. If you wish to restore default settings, see the online help for the Spectrum Mill page of interest.

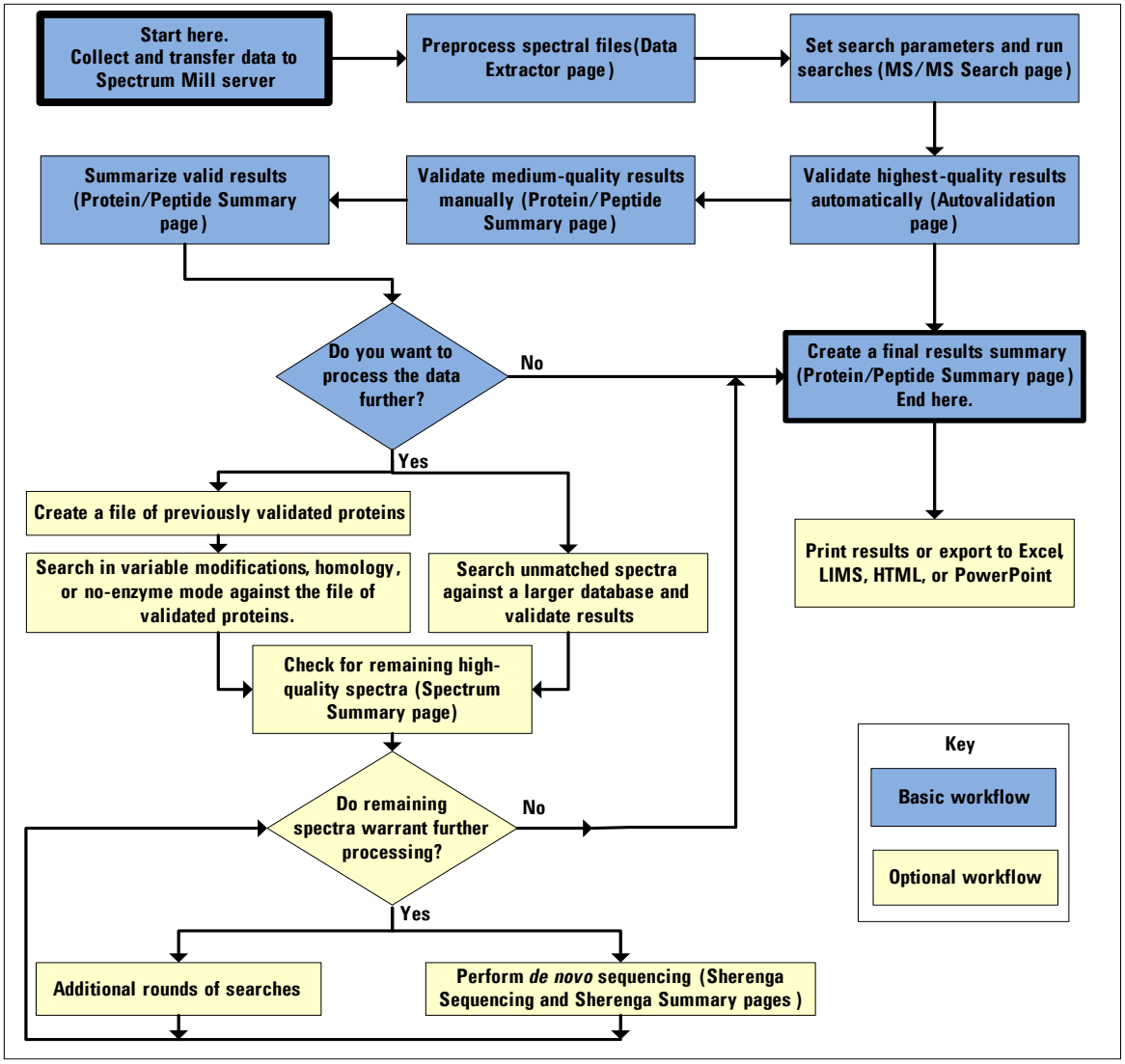


Figure 1 Roadmap to process MS/MS data

Procedure summary

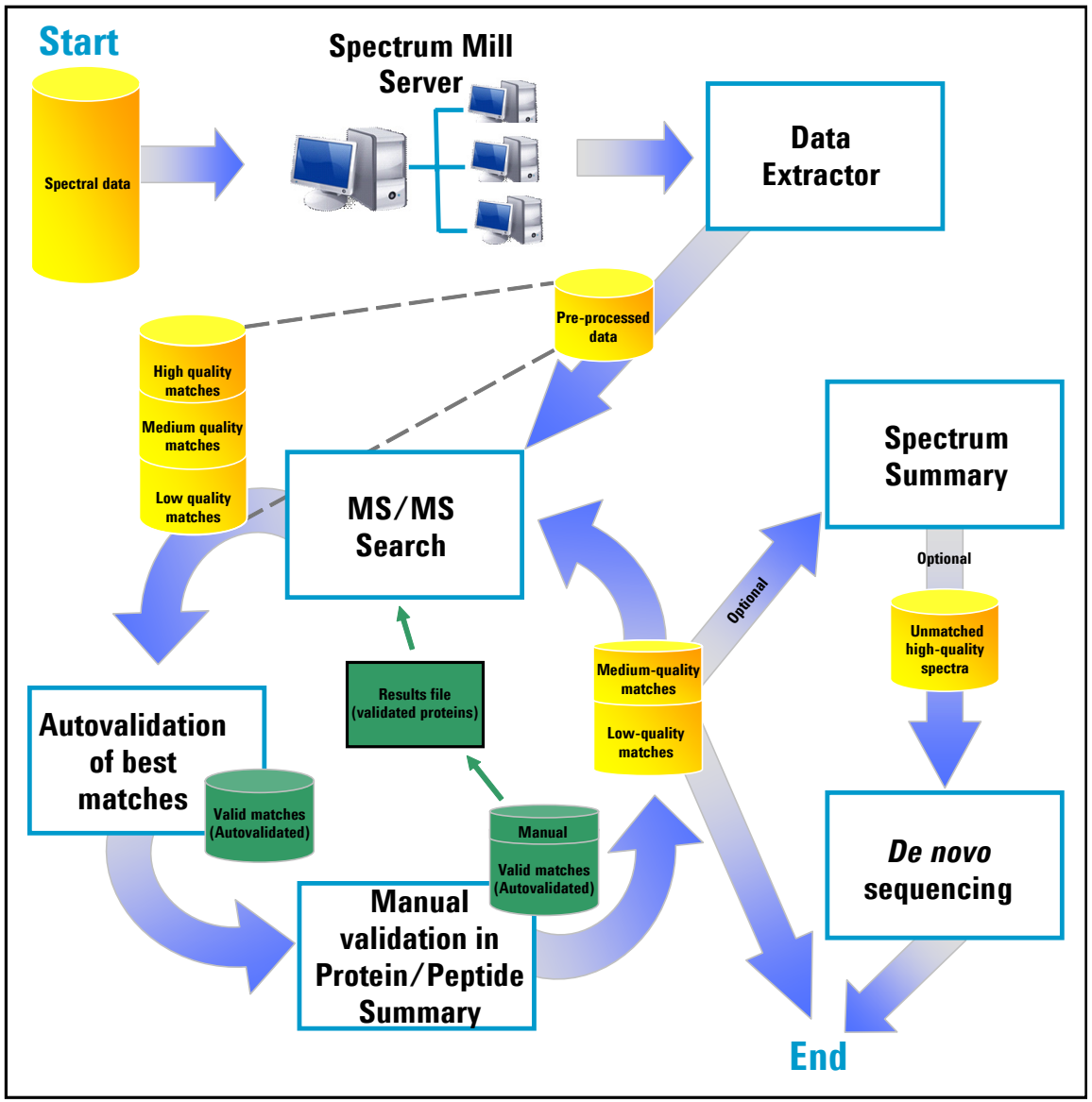


Figure 2 Summary of iterative processing with the Spectrum Mill workbench

Figure 2 and the summary below describe the iterative MS/MS search strategy with the Spectrum Mill workbench. The strategy assumes that the goal is to identify as many proteins as possible. If your study does not require you to identify so many proteins, you may omit the steps between step 5 and step 13. While the following summarizes the overall strategy, the step-by-step procedures are given in the rest of this chapter.

1 Copy or move the raw MS/MS data files to the Spectrum Mill server.
   - Be sure to set up a directory structure on the server that makes it easy to summarize and compare your results. See "To set up a useful directory structure" on page 19.

2 Extract the data.
   - Set the correct cysteine modification.
   - For **Merge scans with same precursor m/z**, if necessary change the time range to be compatible with your chromatographic data. A good starting point is the default of 15 seconds with a +/- 1.4 *m/z* window.

3 Search a database in identity mode, preferably a species subset database for the first search.

4 Autovalidate the results with the highest scores, first in **Protein details** mode, then (optionally) in **Peptide** mode. Use default settings.
   - Validation means that you accept the proposed database match as the correct match for the MS/MS spectrum.

5 Manually review the medium-quality spectra and validate additional results (for ion trap data down to score of 6, SPI 60, and for Agilent Q-TOF data down to score of 5, SPI 60).
   - Depending on the size of the data set, it may be easiest to do this in batches (i.e., first score > 8, then score >6. etc.).
   - For guidelines, see "To manually validate results" on page 91.
   - If you display the forward – reversed and rank 1 – rank 2 scores, differences of at least 1 and preferably 2 can indicate a good match, especially for the lower-scoring hits.
   - If your goal is to maximize the number of identified proteins, do a thorough job at this step because your validated results will be the basis for further searches.
   - If you do not need to maximize the number of identified proteins, skip this step because it can be time-consuming and may not produce a large number of identifications.

6 Use Tool Belt to create a saved results file of validated protein hits (***.res** file).
   - Be sure to indicate the database you searched since this maps to accession numbers.

7 Search the spectra that have not been validated. Search them in variable modifications mode against previous validated results. Search as follows with autovalidation and manual validation between each set of conditions:
   - **a** Combine the following common modifications into a single search: oxidized methionine (methionine sulfoxide) and pyroglutamic acid (N-terminal).
   - **b** If you suspect phosphorylation, combine the following modifications into a single search: phosphorylated S, phosphorylated T, and phosphorylated Y.
   - **c** Search for any other modifications you suspect for your sample.

8 Search in no enzyme mode against previous validated results to find non-specific cleavages for proteins you have already identified. (Set **Digest:** to **No enzyme**.) Repeat autovalidation and manual validation.

9 When you think you are done, list sequence-not-validated spectra in Protein Summary Details mode and look for proteins with multiple peptides. These may represent legitimate proteins at low levels. Re-examine the spectra to confirm.

10 Optional: Search again using a larger database (entire database or larger subset). This is most important when the original subspecies is not well-represented in the database. Autovalidate and manually validate.

11 Check statistics in Tool Belt. If there is a significant number of unmatched filtered spectra, continue searching.

NOTE If the Tool Belt statistics show that only a small percentage of your collected MS/MS spectra were filtered, check your data file to see if you collected a large number of MS/MS spectra on background. If so, your MS/MS data acquisition threshold may be set too low.

12 Use Spectrum Summary to check for sequence-not-validated spectra with sequence tags greater than 6 or 7. Review these and mark as **Good Spectrum** as appropriate.

13 Subject the good spectra to *de novo* sequencing.

14 When you have gained enough information from your data, summarize the results.

First-pass data processing

This following sections describe the steps to perform the initial Spectrum Mill processing for MS/MS data. After you are familiar with the steps in this section and have a good understanding of the software, we recommend you try the Easy MS/MS Search form. This form streamlines the data processing steps by consolidating the Data Extractor, MS/MS Search, Autovalidation, and Protein/Peptide Summary pages onto a single form that shows only the most important parameters. For more details on Easy MS/MS Search, see the online help for that form.

Step 1. Access the Spectrum Mill server

1. Double-click the desktop icon to launch the Spectrum Mill workbench.
2. If you do not see this icon, start your Internet Explorer web browser on the client PC.
3. In your browser window, type the address of the Spectrum Mill home page. (If you don't know the URL, ask the person who installed the software.)
4. Ensure that you see the Spectrum Mill home page, as shown in Figure 3.

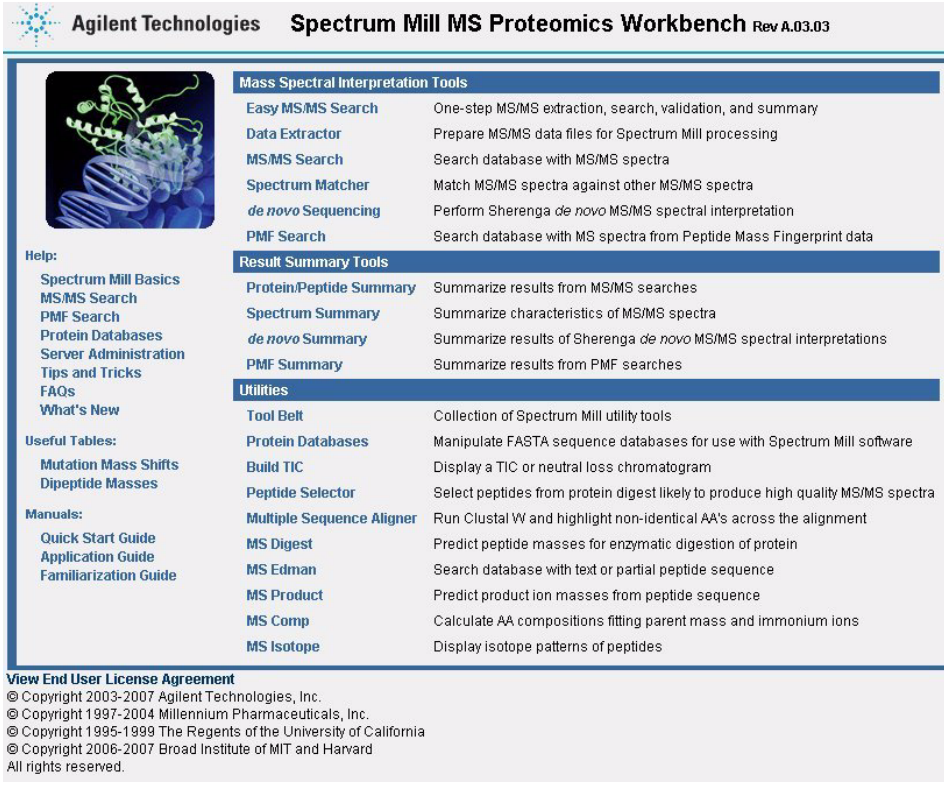


Figure 3 Spectrum Mill home page

Step 2. Prepare spectral files and transfer to Spectrum Mill server

Agilent ion trap and Q-TOF *.d, Thermo Fisher Scientific ion trap *.raw, and Applied Biosystems/MDS Sciex QSTAR *.wiff data files require no special preparation; you merely move or copy the files into the appropriate directory on the Spectrum Mill server, as described below. Some other types of data files (e.g., Waters Micromass Q-Tof) must be preprocessed on the manufacturer's data system to generate generic peak list files. You then copy or move peak list files as described below.

To set up a useful directory structure

To make it easy to compare data sets, it is important that you set up the appropriate directory structure for your files on the Spectrum Mill server. Whenever you want to compare samples in a set, you need to set up a subdirectory for each sample. This sample directory may contain data files from multiple sample fractions or gel slices. Here are some examples:

- If you perform a two-dimensional LC/MS/MS analysis of a single sample, transfer all the files to a single directory on the server.
- If you run the same two-dimensional LC/MS/MS analysis on a second sample, or if you repeat the run on the first sample, transfer all these files to a second directory.
- For differential expression quantitation experiments where you use isotopic labels, transfer all the files to a single directory on the server. This is because all of the versions of the peptide (light, medium, heavy) are found in the same file.
- If you conduct a differential expression study where you use the absolute abundances to perform the quantitation, transfer samples from one cell state into one directory and the second cell state into a second directory.

To transfer Agilent ion trap and Q-TOF, Thermo Fisher Scientific ion trap, or Applied Biosystems/MDS Sciex QSTAR data

1. In your Spectrum Mill file system on the server, navigate to the directory **msdataSM**. (If you don't know how to find your Spectrum Mill file system, ask the person who installed your software.)
2. In the directory **msdataSM**, create a new subdirectory for your sample files. Name it anything you want, but do not use spaces or parentheses in the name.

If you have multiple samples you wish to compare, remember to set up a subdirectory for each sample. Ensure that each subdirectory contains all fractions for a single sample.

You may add up to ten directory levels between **msdataSM** and your sample files, but shorter path lengths reduce memory usage and speed processing of large data sets.

3 Copy or move your *.d, *.raw, or *.wiff files to your new subdirectory or subdirectories.

4 After you transfer data files to the server, remove any spaces or parentheses in the data file names. Delete them or replace them with other characters, such as underscores.

5 Make sure you have both read and write permissions for each subdirectory that contains your sample files. (To check, right-click the folder and select **Properties**.) Make sure all user groups have full permissions. (Check the **Security** tab.)

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

1 Prepare a generic peak list file.

- For Waters Micromass Q-Tof, export centroided data with the Waters Micromass ProteinLynx package. The results are individual *.pkl or *.dta spectral files, or appended *.pkl files that contain multiple spectra.
- For MALDI-TOF-TOF, export *.mgf files.

2 In your Spectrum Mill file system on the server, navigate to the directory **msdataSM**. (If you don't know how to find your Spectrum Mill file system, ask the person who installed your software.)

3 In the directory **msdataSM**, create a new subdirectory for your sample files. Name it anything you want, but do not use spaces or parentheses in the name.

If you have multiple samples you wish to compare, remember to set up a subdirectory for each sample. Ensure that each subdirectory contains all fractions for a single sample.

You may add up to ten directory levels between **msdataSM** and your sample files, but shorter path lengths reduce memory usage and speed processing of large data sets.

4 If you have *.mgf files or appended *.pkl files, where each contain multiple spectra, copy or move them to your new subdirectory or subdirectories. Proceed to step 6. If you have individual *.pkl or *.dta spectral files, proceed to step 5.

5 *Under* your new subdirectory, create another subdirectory called **cpick_in** and copy or move your files there. Note that you *must* use this subdirectory name!

6 After you transfer data files to the server, remove any spaces or parentheses in the data file names. Delete them or replace them with other characters, such as underscores.

7 Make sure you have both read and write permissions for each subdirectory that contains your sample files. (To check, right-click the folder and select **Properties**.) Make sure all user groups have full permissions. (Check the **Security** tab.)

Step 3. Run the Data Extractor

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

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 *.d ion trap and Thermo Fisher Scientific *.raw ion trap data, the extractors optionally merge MS^2 and MS^3 scans from the same precursor. The extractors assign precursor charges where possible, 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.

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

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

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

1. Navigate to the Data Extractor page, as shown in Figure 4.
2. In the **Data Directory** section, click the **Select...** button to select the folder that contains your files.
3. Fill in boxes similar to what is shown in Figure 4. Click the blue divider bars for more information in the online help.

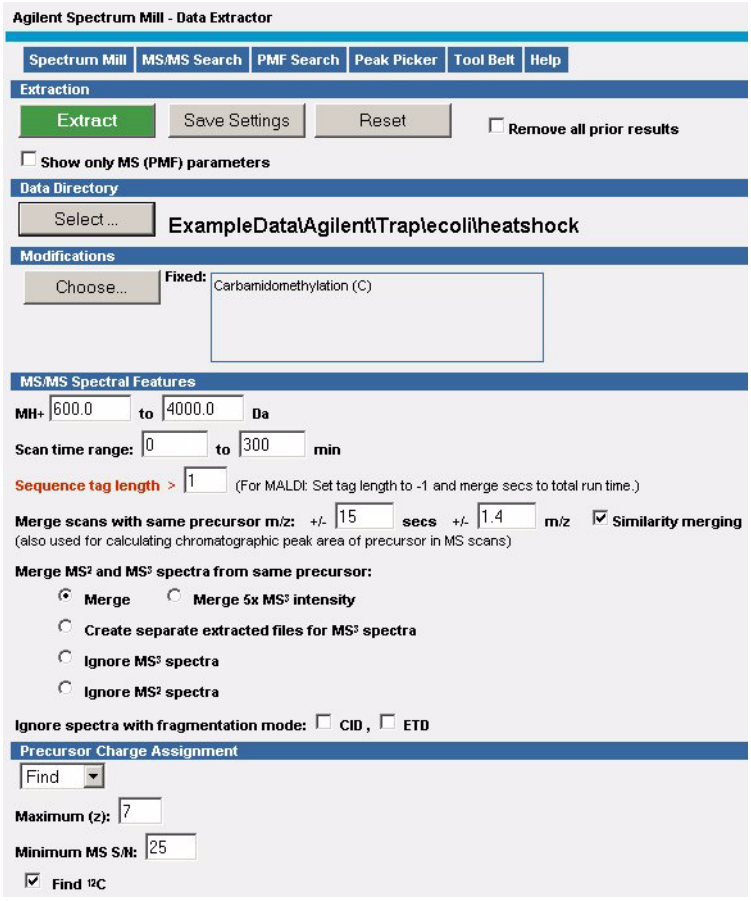


Figure 4 Data Extractor for raw files, shown with typical settings for Agilent ion trap data

Extract Click this button after you have set all the parameters. In addition to initiating extraction, this button saves your settings until you close your web browser.

Save Settings Click this button to save current extraction settings (except data directory). Otherwise, your current settings are retained only until you close your web browser. If you cannot save settings, make sure cookies are enabled for your web browser.

Reset Click this button to reset options to those that were selected the last time you clicked the **Extract** button or the **Save Settings** button on this page.

Data Directory Click the **Select...** button to select the folder that contains your data files. Once you select a data directory, other related forms assume the same selection. To keep this selection even after you have closed your web browser, mark the **Make Default** check box in the **Select Data Directory – Web Page Dialog**. If you cannot save settings, make sure cookies are enabled for your web browser. If you fail to see some or all of your data directories, see "Tips and Tricks" in the online help.

Modifications 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.

Sequence tag length This parameter filters out spectra that are not peptide spectra or that have little information content. The sequence tag length represents the longest sequence of amino acids that can be located in the spectrum. The default of **>1** removes most of the noisy spectra without removing good data. Except for MALDI spectra, keep the default since the MS/MS Search page provides an opportunity to further restrict spectral quality. For MALDI MS/MS spectra, set the value to **-1** so that no filtering is performed.

Merge scans with same precursor m/z If necessary, change the time range to be compatible with your chromatographic data. When you extract MALDI MS/MS data, change the time range to 2000 sec, or to the spot analysis time. Since MALDI data is not chromatographic, you can expand the time range to ensure that you merge all instances of the same precursor.

Similarity merging For complex samples, keep the default setting, with the check box marked. For simple samples (for example, a single protein), clear the check box to improve coverage and to detect more low-level peptides.

Merge MS^2 and MS^3 spectra from same precursor Select one of the options. Note that the Spectrum Mill workbench supports merging of MS^2 and MS^3 files for both*.d and *.raw data files.

4 Click the **Extract** button.

5 As the Data Extractor runs, check that you see a display similar to that shown in Figure 5. 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.

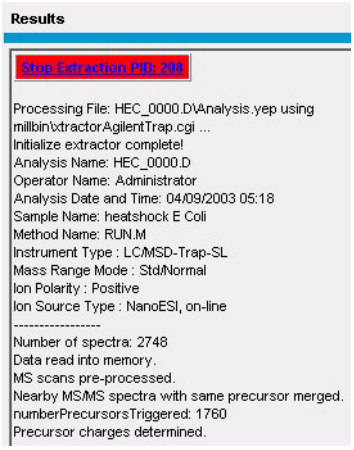


Figure 5 Display while the raw file Data Extractor runs

NOTE While the Data Extractor runs, you may use your client PC for other tasks.

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

1 Navigate to the Data Extractor page.

2 In the Data Directory section, click the **Select...** button to select the folder that contains your files. The Data Extractor form changes to match Figure 6.

3 Fill in boxes similar to what is shown in Figure 6. In general, you should change only the data directory and the parameters highlighted in red text. Click the blue divider bars for more information in the online help.

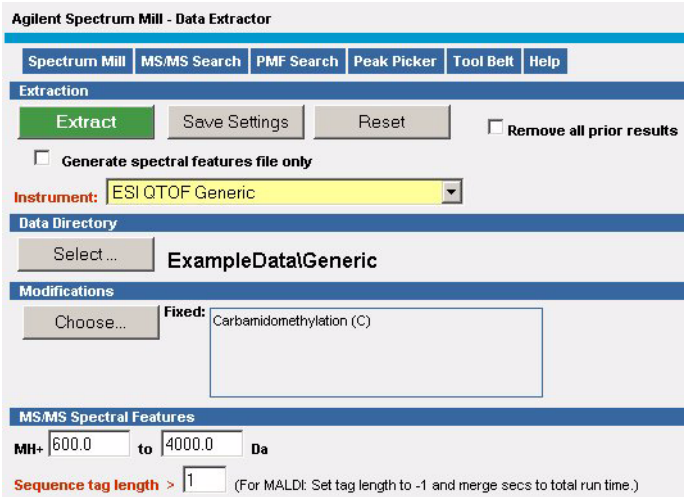


Figure 6 Data Extractor for generic (peak list) files

Extract Click this button after you have set all the parameters.

Save Settings Click this button to save current page settings. (Otherwise, your current settings are retained only until you close your web browser.) If you cannot save settings, make sure cookies are enabled for your web browser.

Reset Click this button to reset options to those that were selected the last time you clicked the **Extract** button or the **Save Settings** button on this page.

Generate spectral features file only Mark this check box 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.)

Instrument Select the instrument type you used for the analyses.

Data Directory Click the **Select...** button to select the folder that contains your data files. Once you select a data directory, other related forms assume the same selection. To keep this selection even after you have closed your web browser, mark the **Make Default** check box in the **Select Data Directory – Web Page**

Dialog. If you cannot save settings, make sure cookies are enabled for your web browser. If you fail to see some or all of your data directories, see "Tips and Tricks" in the online help.

Modifications 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.

Sequence tag length This parameter filters out spectra that are not peptide spectra or that have little information content. The sequence tag length represents the longest sequence of amino acids that can be located in the spectrum. The default of **>1** removes most of the noisy spectra without removing good data. Except for MALDI spectra, it is recommended to keep the default since the MS/MS Search page provides an opportunity to further restrict spectral quality. For MALDI MS/MS spectra, set the value to **-1** so that no filtering is performed.

4 Click the **Extract** button.

5 As the Data Extractor runs, check that you see a display similar to that shown in Figure 7. 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.

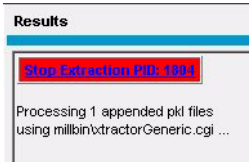


Figure 7 Display while the generic (peak list) Data Extractor runs

NOTE While the Data Extractor runs, you may use your client PC for other tasks.

Step 4. Run database searches

After you have extracted the high-quality MS/MS spectra from your data files, you are ready to search the spectra. The example shown here uses the demonstration data from the **ecoli\heatshock** data directory. If you wish to duplicate this example, then please extract the files as indicated in Figure 4 on page 23.

1 Navigate to the MS/MS Search page, shown in Figure 8.

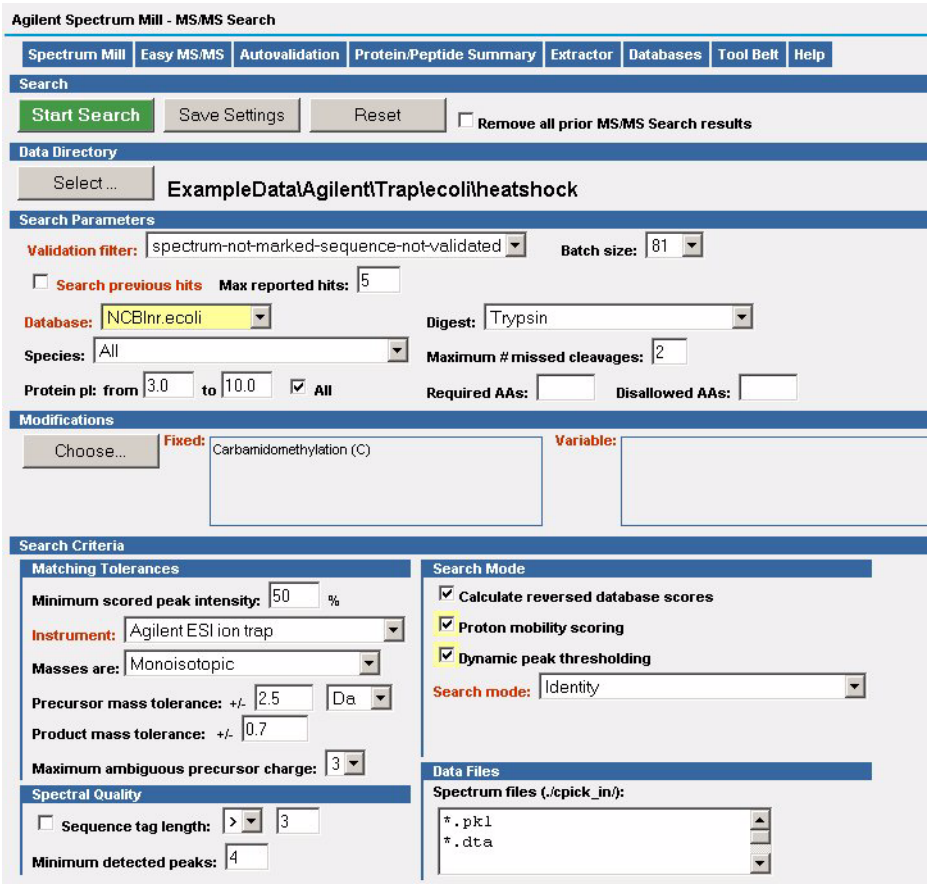


Figure 8 MS/MS Search page

2 If necessary, click the **Select...** button to select the folder that contains your data files. (If you have just performed data extraction, the directory is already set for you.)

3 Select a database, preferably a species subset database for the first search. See "To create a species and protein molecular weight subset database" on page 202.

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 Select the correct instrument. Note that the fragmentation mode is part of the instrument type, so be sure to select the appropriate fragmentation mode within the instrument setting.

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.

10 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.
- If you have an Agilent ion trap data file that contains spectra from both CID and electron transfer dissociation (ETD), the software selectively applies proton mobility scoring to CID spectra and does not apply it to ETD spectra.

11 Mark the check box for **Dynamic peak thresholding**, which is a scoring enhancement that enables identification of more low- abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.

12 Set the other parameters. In general, you should use default values, and change only the parameters highlighted in red text. Click the blue divider bars for more information in the online help.

13 Click the **Start Search** button.

14 As the search runs, check that you see a display similar to that shown in Figure 9. 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.

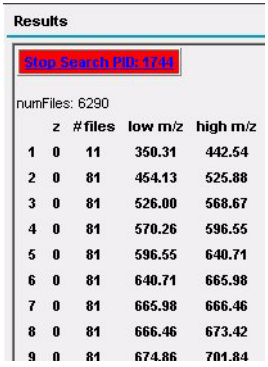


Figure 9 Display while MS/MS Search runs

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 172.

CAUTION If you accidentally run MS/MS Search with incorrect settings, mark the check box to **Remove all prior MS/MS results**. Do not mark this check box if you do iterative searches and wish to retain all results for a final summary.

Step 5. Validate the highest-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 the highest-scoring results — those that do not require manual review. The other method uses the Protein/Peptide Summary page for manual review and validation.

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. Results that are *not* validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that *are* validated can be summarized in a results table.

1 Navigate to the **MS/MS Autovalidation** page, as shown in Figure 10. You navigate to this page from the MS/MS Search page or the Protein/Peptide Summary page.

2 Select the **Data Directory** for which you want to validate results. You may validate twice — first in **Protein details** mode and second in **Peptide** mode.

3 Select the **Protein details** mode.

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

4 If the data files from a single sample occupy multiple directories, mark the check box for **Group proteins across all directories**.

5 Click the **Validate Files** button.

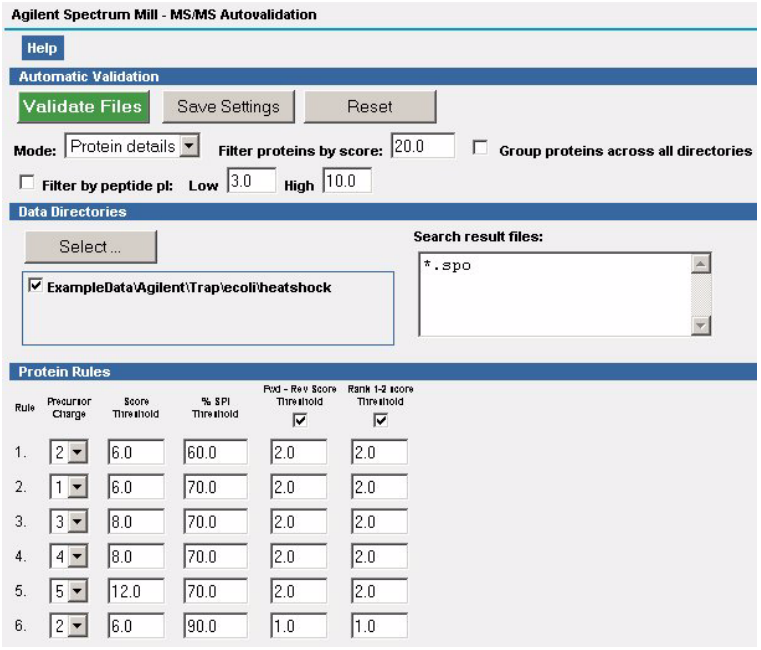


Figure 10 Autovalidation of MS/MS Search results

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

7 Now select the **Peptide** mode.

- In this mode, the software summarizes results by peptide. Even if it finds only a single peptide corresponding to a protein, it validates the corresponding search results provided that the peptide score is high enough.
- Using the default scoring, individual peptides must 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.

8 If you wish to autovalidate only the peptides in a given pI range, mark the check box for **Filter by 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 autovalidate the specific file(s) for which that pI range applies.

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**.

9 Click the **Validate Files** button.

10 You quickly see a validation summary that lists the new hits and spectra that have been validated, along with those validated previously.

Step 6. (Optional) Examine the validated results

This section is optional and is included to show you how to convince yourself that the Autovalidation page did its job correctly.

1 Navigate to the Protein/Peptide Summary page.

2 Set **Mode** to **Peptide**.

3 Click the **Select...** button to select the folder that contains your data files.

4 Set **Validation and Sorting** parameters as shown in Figure 11. These settings filter and present all the validated data. Note that only the settings highlighted in yellow have been changed from defaults.

5 Make sure the **Validation preset** is set to **valid**.

Always set the **Validation preset** to the setting most likely to be accurate for the majority of the data. This minimizes manual changes later. Since the score thresholds were set on the Autovalidation page to select the highest-quality results, the **valid** setting is best for this round of processing.

6 Mark check boxes for the following **Review Fields**: **Fwd-Rev score** (marked by default), **Rank 1-2 score**, **# Cleavages, Unmatched ions**. All are helpful to evaluate data quality.

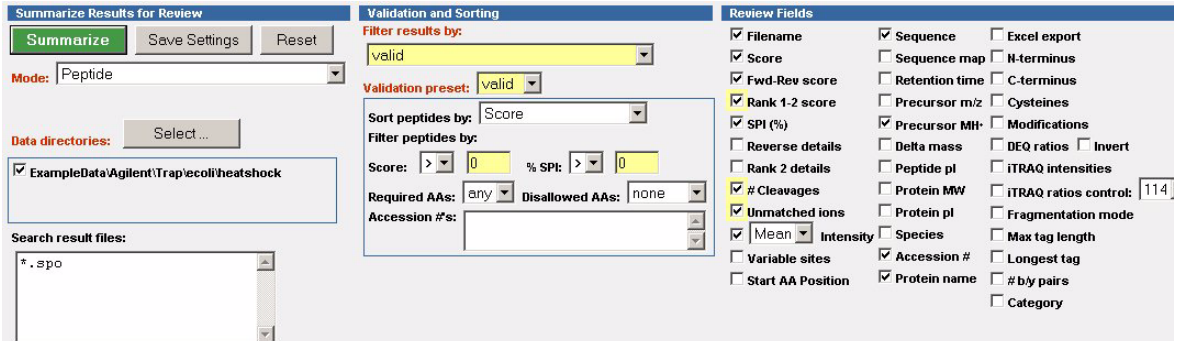


Figure 11 Settings to filter and keep valid results

7 Click the **Summarize** button.

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 form.

8 Check that you see a display similar to that shown in Figure 12.

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)

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.

Perform Validation **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:** 366 **Hits:** 366

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_0040.1867.1879.0	3	26.91	26.91	26.91	98.4	17	1 / 25	3.51e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
V R	2	HEC_0040.2138.2138.0	3	25.56	25.56	25.56	86.6	15	2 / 24	1.50e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
V R	3	HEC_0020.1468.1480.2	2	24.85	23.17	24.85	100.0	12	0 / 25	4.31e+007	(R)SGETEDATIADLAVGTAAGQIK(T)	2118.056	563868	enolase; phosphopyruvate hydrat
V R	4	HEC_0040.1927.1927.0	3	24.63	24.63	24.63	95.0	14	4 / 25	1.20e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
V R	5	HEC_0020.1546.1555.2	2	24.27	19.10	16.44	98.9	10	1 / 24	6.15e+007	(R)TILVPIDISDSELTQR(V)	1799.975	15801753	hypothetical protein Z2335
V R	6	HEC_0010.1397.1405.2	2	23.73	20.23	23.73	97.4	11	2 / 24	9.49e+006	(K)DVTTGDTLCDPDAPIILER(M)	2101.012	15803853	elongation factor EF-2
V R	7	HEC_0010.1547.1562.2	2	23.51	23.51	23.51	94.0	11	3 / 25	8.29e+006	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237	15803763	stringent starvation protein A
V R	8	HEC_0020.1307.1319.0	2	23.37	23.37	23.37	93.6	13	3 / 25	2.75e+007	(K)DTTTIIDGVGEEAAIQGR(V)	1845.919	15804735	chaperonin GroEL
V R	9	HEC_0250.1283.1286.0	3	23.12	23.12	23.12	96.2	12	2 / 25	3.90e+006	(K)VLESAIANAEHNDGADIDDLK(V)	2210.057	15803842	50S ribosomal protein L22
V R	10	HEC_0020.1259.1259.0	2	22.95	22.95	21.43	94.7	12	2 / 25	2.70e+007	(K)DTTTIIDGVGEEAAIQGR(V)	1845.919	15804735	chaperonin GroEL
V R	11	HEC_0010.1372.1384.2	2	22.19	22.19	22.19	95.1	11	2 / 23	1.71e+007	(K)DVTTGDTLCDPDAPIILER(M)	2101.012	15803853	elongation factor EF-2
V R	12	HEC_0020.0995.1009.2	2	22.11	15.97	16.32	98.6	11	1 / 25	1.14e+008	(K)AIAQVGTISANSDETVGK(L)	1760.902	18028150	GroEL
V R	13	HEC_0100.1323.1332.0	3	21.97	21.97	19.82	95.6	13	2 / 24	6.53e+006	(K)ACEEAAEGQVVSPVNFNSPGQVVIAGHK(E)	2894.410	16129055	malonyl-CoA-[acyl-carrier-protein]
V R	14	HEC_0020.0886.0898.0	2	21.78	14.59	13.57	96.1	9	4 / 24	1.71e+008	(K)TGEVPADVAAQAR(Q)	1284.654	15804834	hypothetical protein Z5854
V R	15	HEC_0020.1390.1399.2	2	21.68	18.39	21.68	95.6	9	2 / 23	4.49e+007	(R)GDLGVEIGDPELVGIQK(A)	1738.922	16129807	pyruvate kinase II

Figure 12 Protein/Peptide Summary table

9 To convince yourself that the results are valid, follow procedures described in "Step 8. Manually review and validate results" on page 39. Change the validation preset (under the **Validation category** header in the summary table) back to **reset** (**R**) if you find results that do not meet your quality standards.

10 If you change any **Validation category** back to **reset**, click the **Perform Validation** button at the end of your data review session. This makes your changes a permanent part of the data record.

Step 7. Display medium-quality 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 directory is not selected, click the **Select...** button to select the folder that contains your data files.
4. Set parameters for **Validation and Sorting** as shown in Figure 13. These settings locate medium-to-lower quality 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.

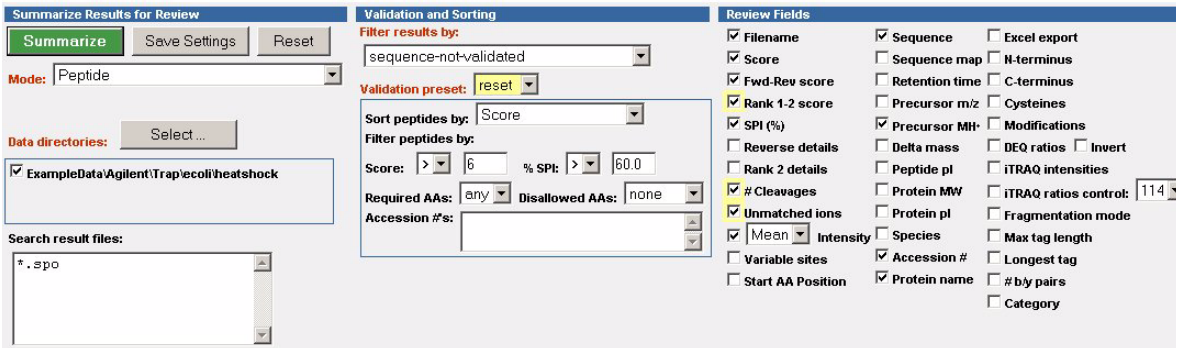


Figure 13 Settings for medium-quality results

5. Use the following guidelines for setting the **Filter peptides by** parameters:
   - If you have ion trap data, for the **Score** settings, see the guidelines in Table 1.
   - 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.

- **% SPI** (Scored Peak Intensity) is the percentage of the MS/MS peak-detected spectral ion current explained by the search interpretation. For ion trap spectra, **% SPI** below 70 (or 60 for doubly-charged precursor ions) are more likely to represent poor interpretations, so do not set lower values. For Agilent Q-TOF spectra, **% SPI** below 60 may 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.

Table 1 Guidelines for validating **ion trap** results based on MS/MS Search scores

Peptide score	Quality	Peptide fragmentation	Likelihood of valid interpretation
13-25	Excellent	Extensive	When combined with **% SPI** of 70 or greater, very likely to be valid
9-13	Good	Substantial	When combined with **% SPI** of 70 or greater, likely to be valid
6-9	Mediocre	Moderate	Review results to determine whether interpretation is valid
3-6	Weak	Sparse	Not likely to be valid

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

- 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**.
- 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**.

7 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 **# 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**.

8 Click the **Summarize** button.

NOTE

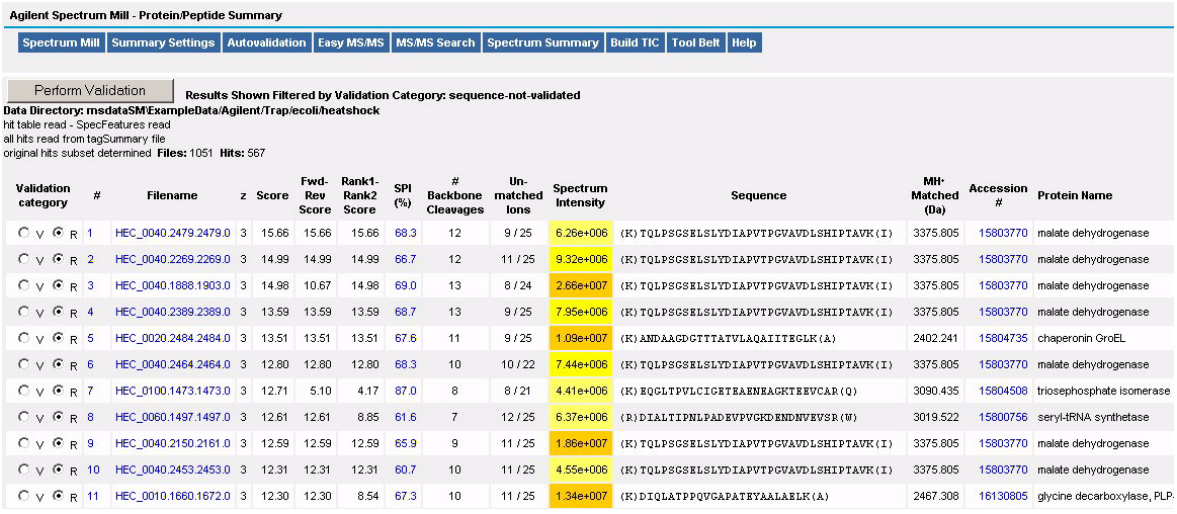
Once you click the **Summarize** button, the button is disabled until the results appear. If you need to re-enable the button, click the **Summary Settings** button to reload the Protein/Peptide Summary form.

9 Check that you see a display similar to that shown in Figure 14 on page 39. 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.

10 Manually review the results, as described in the next section.

Step 8. 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 14.



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: msdataSM\ExampleData/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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)ANDAAGDGTTTATVLAQAIITEGLK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)EQGLTPVLCIGETAAENEAGKTEEVCAR(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)DIALTIPNLPADEVPVGKDENDNVEVSR(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)DIQLATPPQVGAPATEYAALAELK(A)	2467.308	16130805	glycine decarboxylase, PLP-

Figure 14 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 manually validate results" on page 91.

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 15), 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 87 of Chapter 2.

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

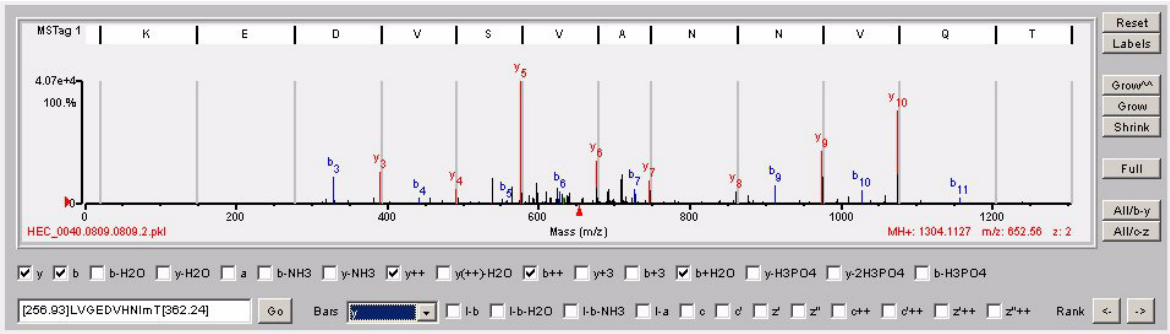


Figure 15 Spectrum Viewer, with y-ions highlighted using the **Bars** feature

To view MS/MS Search results

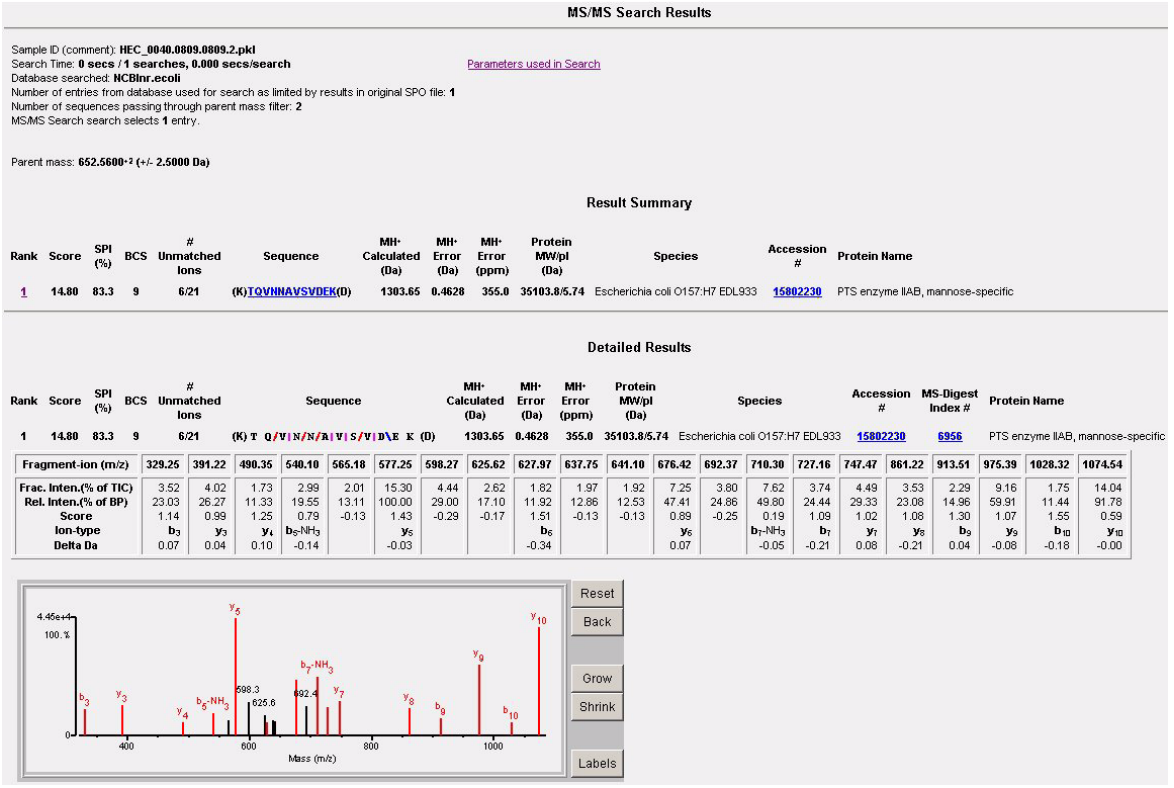
If you want more detailed information than is presented in the Spectrum Viewer, you view the MS/MS search details, as shown in Figure 16.

1 To view these details, click a link under the **Filename** header shown in Figure 14. 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.



MS/MS Search Results

Sample ID (comment): **HEC_0040.0809.0809.2.pkl**
Search Time: **0 secs / 1 searches, 0.000 secs/search** Parameters used in Search
Database searched: **NCBInr.ecoli**
Number of entries from database used for search as limited by results in original SPO file: **1**
Number of sequences passing through parent mass filter: **2**
MS/MS Search search selects **1** entry.

Parent mass: **652.5600+2 (+/- 2.5000 Da)**

Result Summary

Rank	Score	SPI (%)	BCS	# Unmatched Ions	Sequence	MH+ Calculated (Da)	MH+ Error (Da)	MH+ Error (ppm)	Protein MW/pI (Da)	Species	Accession #	Protein Name
1	14.80	83.3	9	6/21	(K)TQVNNAVSVDEK(D)	1303.65	0.4628	355.0	35103.8/5.74	Escherichia coli O157:H7 EDL933	15802230	PTS enzyme IIAB, mannose-specific

Detailed Results

Rank	Score	SPI (%)	BCS	# Unmatched Ions	Sequence	MH+ Calculated (Da)	MH+ Error (Da)	MH+ Error (ppm)	Protein MW/pI (Da)	Species	Accession #	MS-Digest Index #	Protein Name
1	14.80	83.3	9	6/21	(K) T Q/V|N/N/A|V|S/V|D\E K (D)	1303.65	0.4628	355.0	35103.8/5.74	Escherichia coli O157:H7 EDL933	15802230	6956	PTS enzyme IIAB, mannose-specific

Fragment-ion (m/z)	329.25	391.22	490.35	540.10	565.18	577.25	598.27	625.62	627.97	637.75	641.10	676.42	692.37	710.30	727.16	747.47	861.22	913.51	975.39	1028.32	1074.54
Frac. Inten.(% of TIC)	3.52	4.02	1.73	2.99	2.01	15.30	4.44	2.62	1.82	1.97	1.92	7.25	3.80	7.62	3.74	4.49	3.53	2.29	9.16	1.75	14.04
Rel. Inten.(% of BP)	23.03	26.27	11.33	19.55	13.11	100.00	29.00	17.10	11.92	12.86	12.53	47.41	24.86	49.80	24.44	29.33	23.08	14.96	59.91	11.44	91.78
Score	1.14	0.99	1.25	0.79	-0.13	1.43	-0.29	-0.17	1.51	-0.13	-0.13	0.89	-0.25	0.19	1.09	1.02	1.08	1.30	1.07	1.55	0.59
Ion-type	b_3	y_3	y_4	b_5-NH_3		y_5			b_6			y_6		b_7-NH_3	b_7	y_7	y_8	b_9	y_9	b_{10}	y_{10}
Delta Da	0.07	0.04	0.10	-0.14		-0.03			-0.34			0.07		-0.05	-0.21	0.08	-0.21	0.04	-0.08	-0.18	-0.00

Figure 16 MS/MS Search results

4 Click additional links shown in the MS/MS Search results in Figure 16:

- 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.

Step 9. 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. 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

Table 2 Protein/Peptide Summary Modes

If you want to:	And you want results organized by:	Then use this mode:	Example application
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

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 Set the remaining parameters as shown in Figure 17, or use your customized settings. See the text below for additional information. Note that in Figure 17, only the settings highlighted in yellow have been changed from defaults.

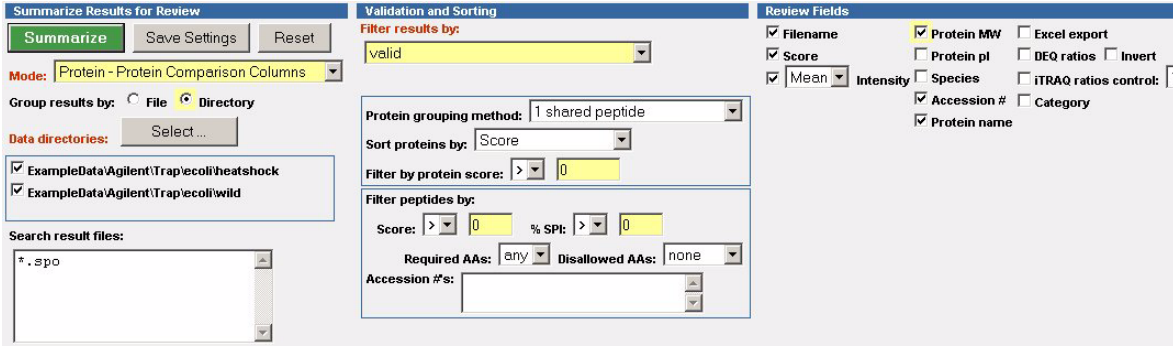


Figure 17 Example of settings to summarize all valid results

Intensity Mark this check box to calculate and include protein intensities. The software calculates the intensities from the extracted ion chromatograms of the precursor ions for each peptide. Depending upon your selection, the software then calculates either a mean or total intensity for the peptides that make up the protein. Studies have shown that the total intensity provides more accurate quantitation than the mean intensity.

Excel export Mark this check box to export data to Excel or LIMS.

DEQ ratios Mark to 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 134 of Chapter 5.

6 Click the **Summarize** button.

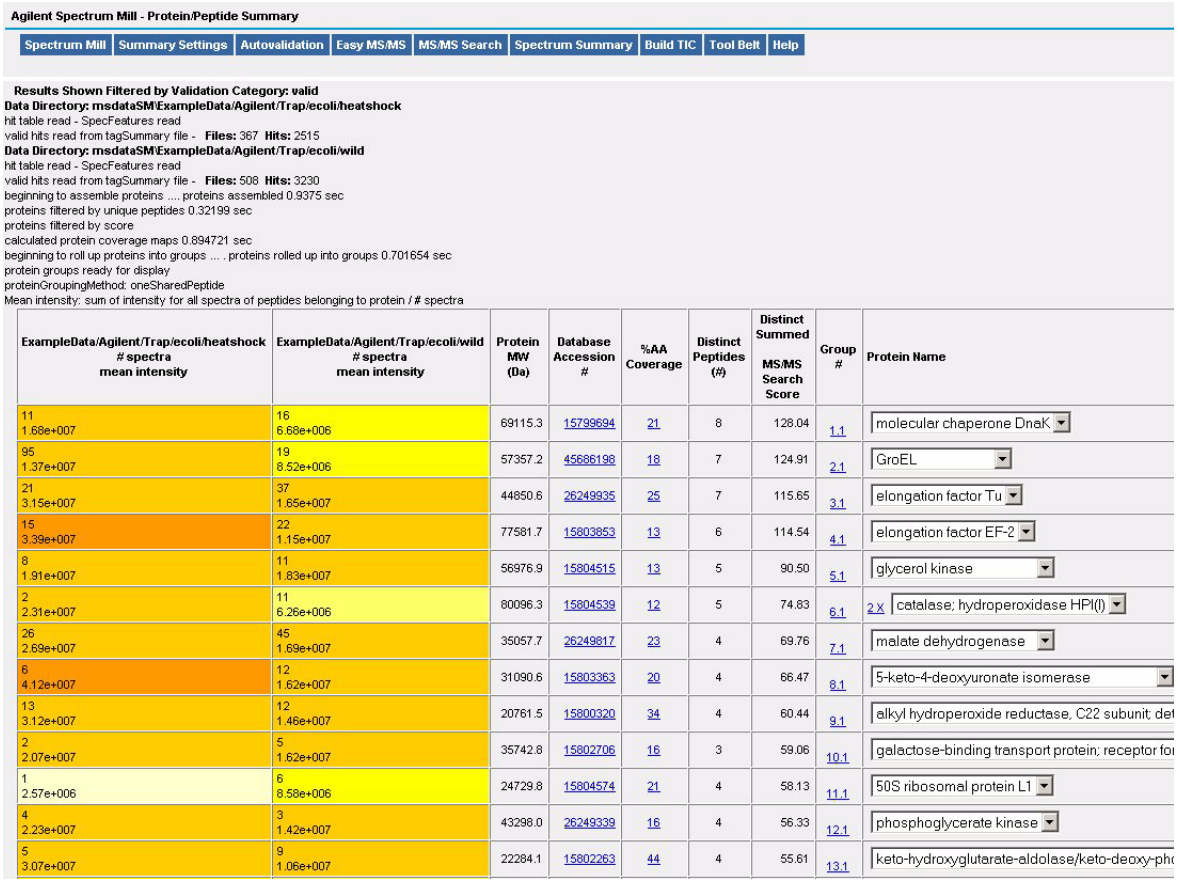
7 Check that you see a summary display.

The display shown in Figure 18 was generated with the settings shown in Figure 17. The score and % SPI filters were set to zero to capture all data marked **valid**.

The example in Figure 18 includes two *E. coli* samples — a "heatshock" (stressed) sample and a "wild" (control) sample. 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 stressed 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.



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/Agilent/Trap/ecoli/heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 367 **Hits:** 2515
Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/wild
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 508 **Hits:** 3230
beginning to assemble proteins .... proteins assembled 0.9375 sec
proteins filtered by unique peptides 0.32199 sec
proteins filtered by score
calculated protein coverage maps 0.894721 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.701654 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	ExampleData/Agilent/Trap/ecoli/wild # spectra mean intensity	Protein MW (Da)	Database Accession #	%AA Coverage	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	Group #	Protein Name
11 1.68e+007	16 6.68e+006	69115.3	15799694	21	8	128.04	1.1	molecular chaperone DnaK
95 1.37e+007	19 8.52e+006	57357.2	45686198	18	7	124.91	2.1	GroEL
21 3.15e+007	37 1.65e+007	44850.6	26249935	25	7	115.65	3.1	elongation factor Tu
15 3.39e+007	22 1.15e+007	77581.7	15803853	13	6	114.54	4.1	elongation factor EF-2
8 1.91e+007	11 1.83e+007	56976.9	15804515	13	5	90.50	5.1	glycerol kinase
2 2.31e+007	11 6.26e+006	80096.3	15804539	12	5	74.83	6.1	2 X catalase; hydroperoxidase HPI(I)
26 2.69e+007	45 1.69e+007	35057.7	26249817	23	4	69.76	7.1	malate dehydrogenase
6 4.12e+007	12 1.62e+007	31090.6	15803363	20	4	66.47	8.1	5-keto-4-deoxyuronate isomerase
13 3.12e+007	12 1.46e+007	20761.5	15800320	34	4	60.44	9.1	alkyl hydroperoxide reductase, C22 subunit; det
2 2.07e+007	5 1.62e+007	35742.8	15802706	16	3	59.06	10.1	galactose-binding transport protein; receptor fo
1 2.57e+006	6 8.58e+006	24729.8	15804574	21	4	58.13	11.1	50S ribosomal protein L1
4 2.23e+007	3 1.42e+007	43298.0	26249339	16	4	56.33	12.1	phosphoglycerate kinase
5 3.07e+007	9 1.06e+007	22284.1	15802263	44	4	55.61	13.1	keto-hydroxyglutarate-aldolase/keto-deoxy-pho

Figure 18 Summary of valid results

8 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 19. 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.

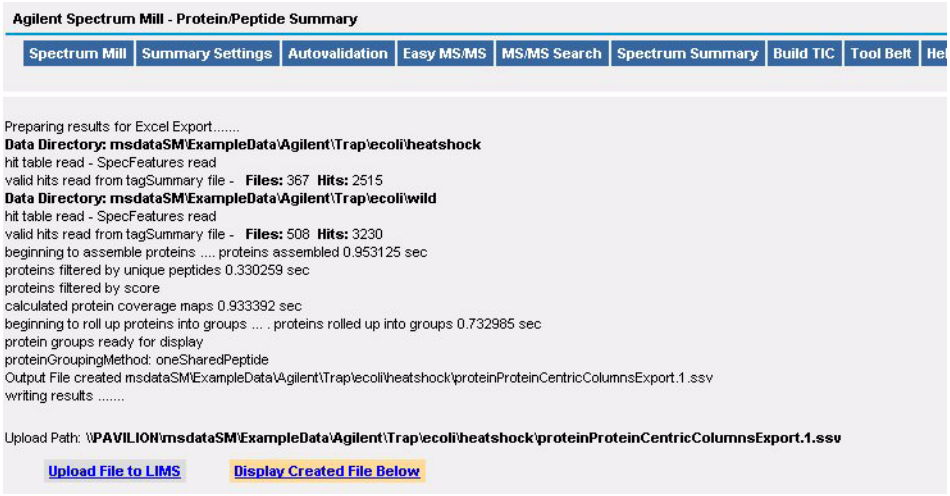


Figure 19 Display when you mark the Excel export check box

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

a If you have not already done so, enable **Print background colors and images** in Internet Explorer. Select **Tools > Internet Options...**. Click the **Advanced** tab and mark the check box for **Print background colors and images**.

b Do one of the following:

- In Internet Explorer (IE) 6.0, select **File > Page Setup...** and set the page to landscape mode.

- In IE 7.0, click the arrow to the right of the printer icon, select **Page Setup...** and set the page to landscape mode.

c Click in the frame you wish to print.

d Do one of the following:

- In IE 6.0, select **File > Print Preview...**.
- In IE 7.0, click the arrow to the right of the printer icon and select **Print Preview...**.

e At the top of the **Print Preview** window, select **Only the selected frame**.

f Click the **Print...** button.

10 If you wish to further process your data, continue to the next section.

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

Optional data processing to extract additional information

After you complete the first-pass processing described in the previous section, you may carry out additional processing steps. If you are doing the additional processing for the first time, read the overview that follows. If you are already familiar with the overview, skip to "To create a file of previously validated results" on page 52.

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 optional processing

The optional data processing steps are generalized in Figure 1, “Roadmap to process MS/MS data,” on page 13. Which steps you choose depends 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. The possibilities are:

- Search previous hits in variable modifications mode for oxidized methionine and pyroglutamic acid and review/validate results.
- 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.)
- Search previous hits in no-enzyme mode and review/validate results.
- Search a larger database in identity mode and review/validate results.
- Search previous hits in homology mode and review/validate results.
- Check for remaining high-quality spectra.
- Create a final results summary.
- Perform Sherenga *de novo* sequencing on remaining high quality spectra and review results.

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 perform 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 202. To create a small database of proteins that have already been identified, see "To create a file of previously validated results" on page 52. 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 perform 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.

To create a file of previously validated results

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

Here, you will create a file of validated hits from your previous search results. You will 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 20.

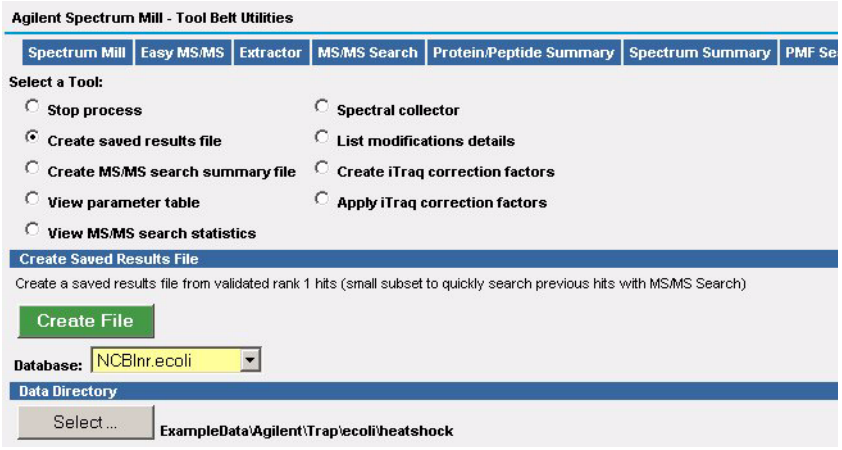


Figure 20 Utility to create a saved results file

3 Click the **Create File** button.

4 Check that you see a display similar to that shown in Figure 21.

```
database: NCBInr.ecoli input program: MS-Tag

OutputResFilename: NCBInr.ecoli.res
numValidHits: 850
numIndicies: 850
numNonRedundantIndicies: 126
numAccessionNumbers: 850
numNonRedundantAccessionNumbers: 126
created file F:\SpectrumMill\msdataSM\ExampleData\Agilent\Trap\ecoli\heatshock\results_mstag\NCBInr.ecoli.res
```

Figure 21 Display during creation of saved results file

Another way to create the file of previously validated results

The above steps show how to create a file of previously validated hits by using the Tool Belt page. Alternatively, you can create the file automatically from the MS/MS Search page. When you mark the check box to **Search previous hits** in MS/MS Search, if you have not created a saved results file, then the search page automatically creates one for you. (On the form, be sure to select the database that you searched previously.)

To search in variable modifications or homology modes

1 Navigate to the MS/MS Search page, as shown in Figure 22.

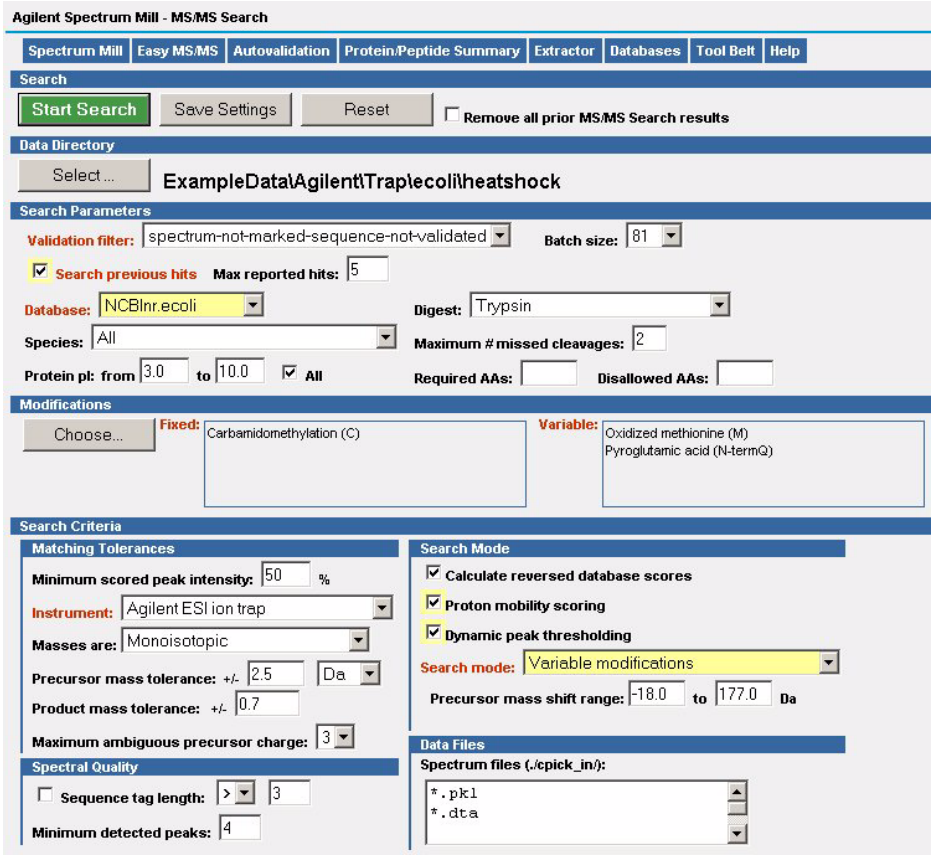


Figure 22 MS/MS Search form, set up for variable modifications mode search

2 Click the **Select...** button to choose the data directory where your files reside.

3 Set the **Validation filter** to **spectrum-not-marked-sequence-not-validated.**

4 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**.

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 202.

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 the same **Fixed** or **Mix Modifications** that you selected for the previous identity mode search. Add any **Variable Modifications** that you wish to 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.

NOTE

You can combine multiple variable modifications into a single search, but as the number of permutations increases, the chance of false positives and the search time increases as well, so it is best to limit the number of variable modification to two to three per search.

However, if you have already validated results from an identity-mode search, create a file of validated hits (see page 52), 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.

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 N15-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.

8 Select the correct instrument. Note that the fragmentation mode is part of the instrument type, so be sure to select the appropriate fragmentation mode within the instrument setting.

9 For **Search mode**, select the **Variable modifications** mode or one of the homology modes for this round of processing. Use the following as a guide:

- To search the variable modification(s) that you chose in step 7, select the **Variable modifications** mode. 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 and increase the possibility of false positives.
- 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.

NOTE

If you attempt a homology or variable modifications 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 and variable modifications search against a file of previously validated hits. See "To create a file of previously validated results" on page 52. If that is not possible, reduce the **Maximum # missed cleavages** to **1**.

10 To help to rule out false positives when you perform data review, 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. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.

13 Set the other parameters. 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.

14 Click the **Start Search** button.

15 When the search is complete, do one of the following:

- If you performed a variable modifications search that was limited to a few variable modifications, first automatically validate results on the MS/MS Autovalidation page. See "Step 5. Validate the highest-quality results automatically" on page 31. Then optionally review and validate additional results on the Protein/Peptide Summary page, as described in the next section.
- If you performed a homology mode search, review and validate results on the Protein/Peptide Summary page, as described in the next section. Autovalidation is not recommended for the homology modes because of the increased possibility of false positives when so many permutations are considered.

To review/validate variable modifications or homology mode results

To use the Protein/Peptide Summary page to review and validate results, follow instructions on page 36 through page 43, with the following changes:

- If you want to see the variable modifications and/or amino acid substitutions, additionally mark the check boxes for **Variable sites**, **Delta mass**, and **Modifications**, as shown in Figure 23. 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**.

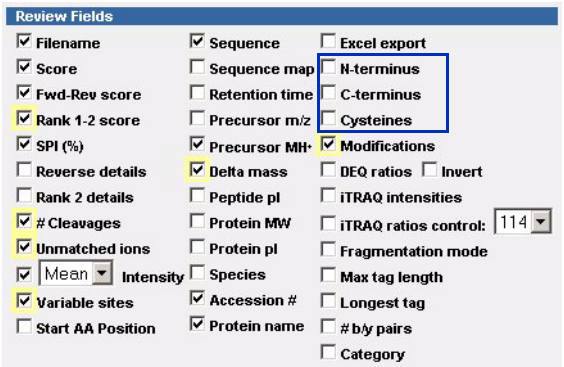
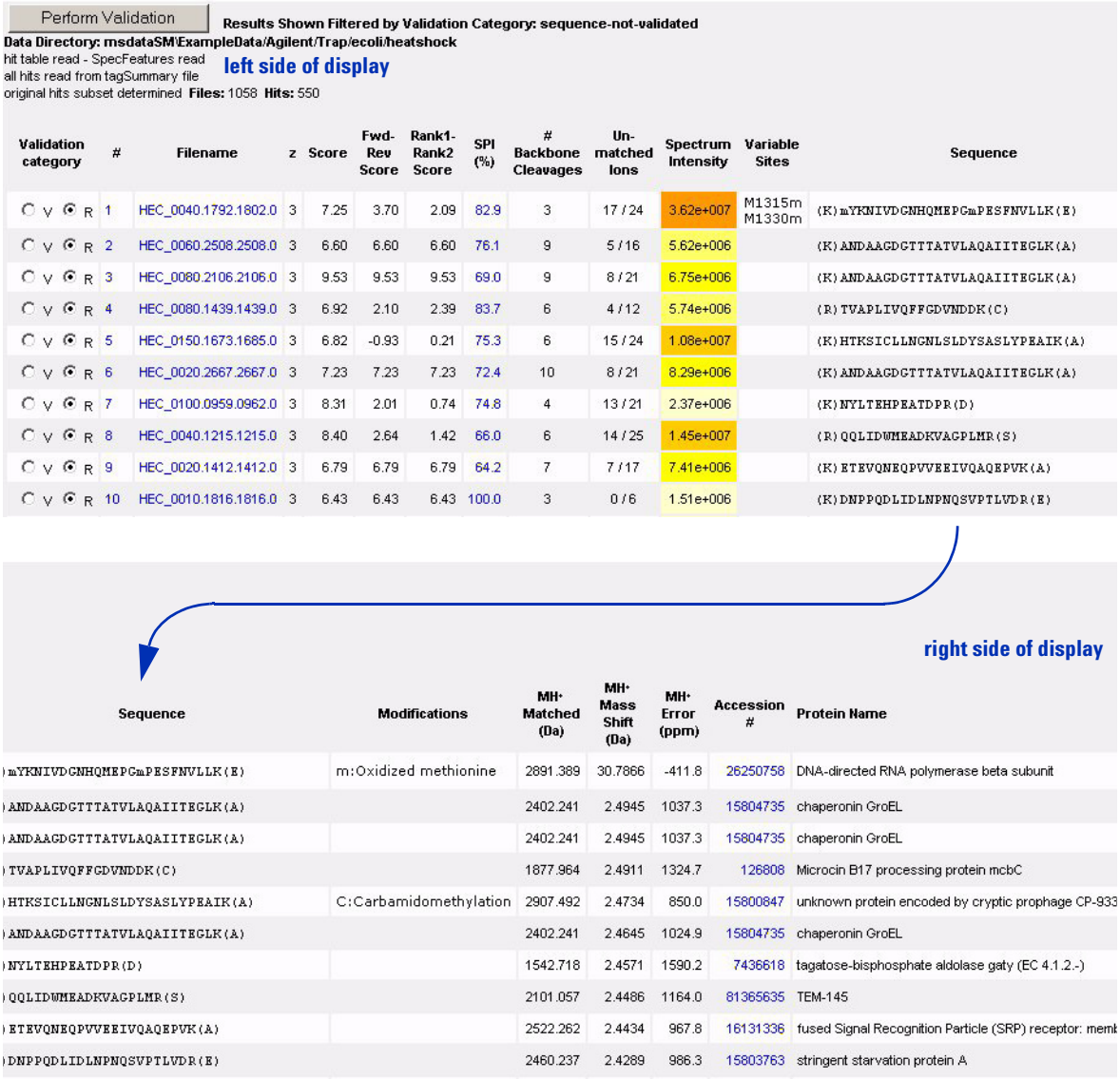


Figure 23 Check boxes to mark for review of results from variable modifications or homology mode searches

- 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**.
- If you want to see the variable modifications grouped at the top and/or bottom of the results table, for **Sort peptides by**, select **Delta Mass Shift**.
- Note that the resulting data display includes the additional headings **Variable Sites**, **Modifications**, **MH^+ mass shift (Da)**, and **MH^+ error (ppm)**, as shown in Figure 24. 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**.



Validation category	#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	# Backbone Cleavages	Un-matched Ions	Spectrum Intensity	Variable Sites	Sequence
V R	1	HEC_0040.1792.1802.0	3	7.25	3.70	2.09	82.9	3	17 / 24	3.62e+007	M1315m M1330m	(K)mYKNIVDGNHQMEPGmPESFNVLLK(E)
V R	2	HEC_0060.2508.2508.0	3	6.60	6.60	6.60	76.1	9	5 / 16	5.62e+006		(K)ANDAAGDGTTTATVLAQAIITEGLK(A)
V R	3	HEC_0080.2106.2106.0	3	9.53	9.53	9.53	69.0	9	8 / 21	6.75e+006		(K)ANDAAGDGTTTATVLAQAIITEGLK(A)
V R	4	HEC_0080.1439.1439.0	3	6.92	2.10	2.39	83.7	6	4 / 12	5.74e+006		(R)TVAPLIVQFFGDVNDDK(C)
V R	5	HEC_0150.1673.1685.0	3	6.82	-0.93	0.21	75.3	6	15 / 24	1.08e+007		(K)HTKSICLLNGNLSLDYSASLYPEAIK(A)
V R	6	HEC_0020.2667.2667.0	3	7.23	7.23	7.23	72.4	10	8 / 21	8.29e+006		(K)ANDAAGDGTTTATVLAQAIITEGLK(A)
V R	7	HEC_0100.0959.0962.0	3	8.31	2.01	0.74	74.8	4	13 / 21	2.37e+006		(K)NYLTEHPEATDPR(D)
V R	8	HEC_0040.1215.1215.0	3	8.40	2.64	1.42	66.0	6	14 / 25	1.45e+007		(R)QQLIDWMEADKVAGPLMR(S)
V R	9	HEC_0020.1412.1412.0	3	6.79	6.79	6.79	64.2	7	7 / 17	7.41e+006		(K)ETEVQNEQPVVEEIVQAQEPVK(A)
V R	10	HEC_0010.1816.1816.0	3	6.43	6.43	6.43	100.0	3	0 / 6	1.51e+006		(K)DNPPQDLIDLNPNQSVPTLVDR(E)

Sequence	Modifications	MH+ Matched (Da)	MH+ Mass Shift (Da)	MH+ Error (ppm)	Accession #	Protein Name
)mYKNIVDGNHQMEPGmPESFNVLLK(E)	m:Oxidized methionine	2891.389	30.7866	-411.8	26250758	DNA-directed RNA polymerase beta subunit
)ANDAAGDGTTTATVLAQAIITEGLK(A)		2402.241	2.4945	1037.3	15804735	chaperonin GroEL
)ANDAAGDGTTTATVLAQAIITEGLK(A)		2402.241	2.4945	1037.3	15804735	chaperonin GroEL
)TVAPLIVQFFGDVNDDK(C)		1877.964	2.4911	1324.7	126808	Microcin B17 processing protein mcbC
)HTKSICLLNGNLSLDYSASLYPEAIK(A)	C:Carbamidomethylation	2907.492	2.4734	850.0	15800847	unknown protein encoded by cryptic prophage CP-933
)ANDAAGDGTTTATVLAQAIITEGLK(A)		2402.241	2.4645	1024.9	15804735	chaperonin GroEL
)NYLTEHPEATDPR(D)		1542.718	2.4571	1590.2	7436618	tagatose-bisphosphate aldolase gaty (EC 4.1.2.-)
)QQLIDWMEADKVAGPLMR(S)		2101.057	2.4486	1164.0	81365635	TEM-145
)ETEVQNEQPVVEEIVQAQEPVK(A)		2522.262	2.4434	967.8	16131336	fused Signal Recognition Particle (SRP) receptor: memb
)DNPPQDLIDLNPNQSVPTLVDR(E)		2460.237	2.4289	986.3	15803763	stringent starvation protein A

Figure 24 Combined results from identity and variable modifications searches (partial list)

To search in no enzyme mode

1 Navigate to the MS/MS Search page.

2 Fill in the search form as described in "Step 4. Run database searches" on page 28, 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 31 through page 43.

To search a larger database

1 Navigate to the MS/MS Search page.

2 Fill in the search form as described in "Step 4. Run database searches" on page 28, except select a different database this time.

- For example, if you first searched against the SwissProt database, try the larger NCBInr 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 31 through page 43.

NOTE To search your own custom library, see "Spectrum Matcher" on page 71.

To 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 25. The notes below discuss important parameters.

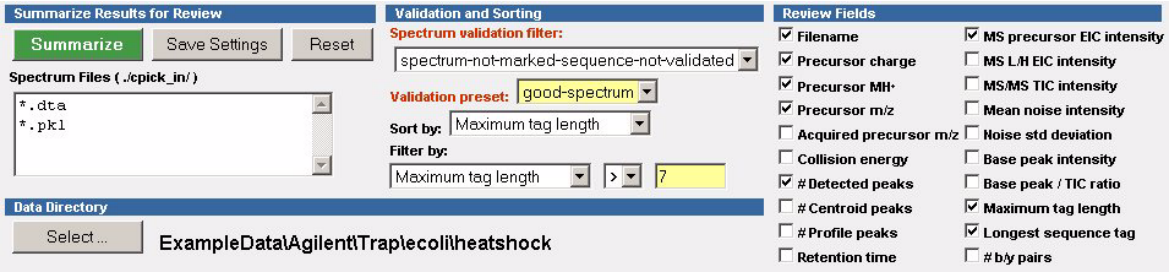


Figure 25 Spectrum Summary form

Data Directory Click the **Select...** button to select the folder that contains your data files.

Spectrum validation filter Set 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**.

Validation preset You have three choices:

- 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.

Filter by: Set this parameter to filter out spectra that are not peptide spectra or that have little information content. The maximum tag length 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.

MS L/H EIC intensity Mark this 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 141 of Chapter 5.

3 Click the **Summarize** button.

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 form.

4 Check that you see a display similar to the one shown in Figure 26.

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 34.

Figure 26 Spectrum Summary results

5 Now examine the spectra. The Spectrum Viewer, shown at the bottom of Figure 26, 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 87 of Chapter 2.

6 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.

7 Click the **Update** button 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.

To 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.

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

   c Search in one of the homology modes against previous protein hits

   d Search a larger database.

   e Perform Sherenga *de novo* sequencing

3 If you do not wish to further process the data, create a final results summary. See "Step 9. Summarize valid results" on page 43.

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 173 and "To create a summary table of MS/MS identification statistics" on page 175. 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 74.

To perform 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 3.

1 Navigate from the Spectrum Mill home page to the Sherenga *de novo* Sequencing page, shown in Figure 27.

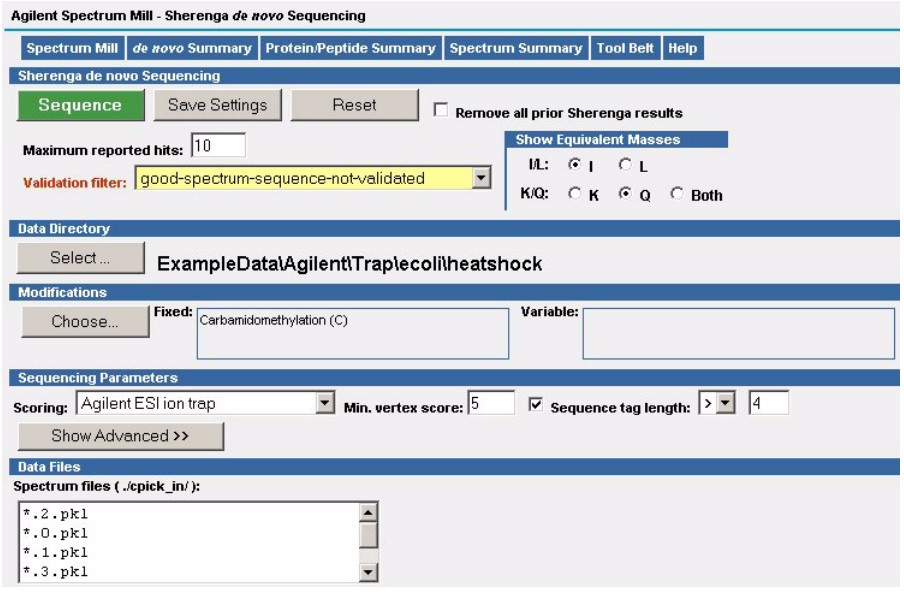


Figure 27 Sherenga *de novo* Sequencing form

2 Set the **Validation filter**:

- If you took the time to designate your spectra as “good” or “bad,” as described in “To check for remaining high-quality spectra” on page 62, 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**.

3 Click the **Select...** button to select the data directory where your files reside.

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

5 Under **Sequencing Parameters**, for **Scoring**, select your instrument.

6 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.

7 Click the **Sequence** button.

8 As the sequencing program runs, check that you see a display similar to that shown in Figure 28. The sequencing program processes the entire directory. A message alerts you when processing is complete.

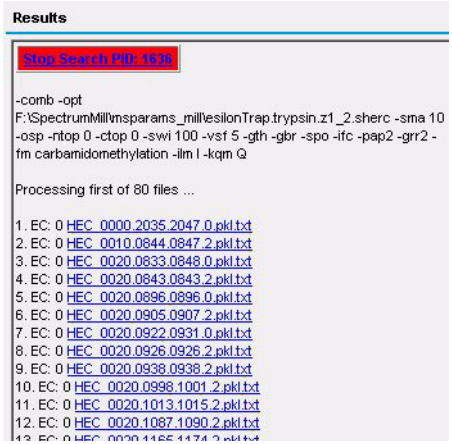


Figure 28 Display while Sherenga runs

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 form shown in Figure 29. You should generally keep the default settings.

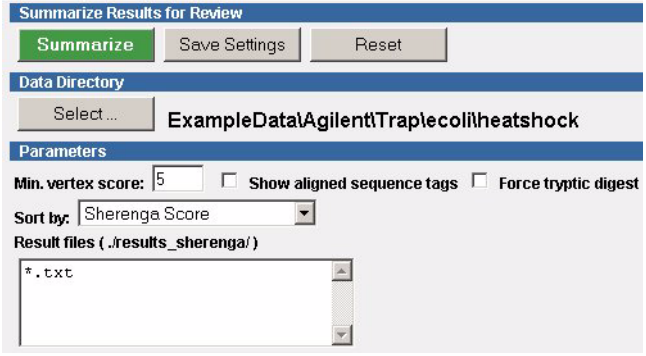


Figure 29 Sherenga *de novo* Summary form

3 Click the **Summarize** button.

4 Check that you see a display like that shown in Figure 30.

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 34.

Results

SS	TL	TS	MS/MS Search 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]NINVVDAGGAGQFISA[215.27]	MGLVIKAALGALVVLLIGVLAK
292	10			HEC_0020.1766.1780.2.pkl.txt	[251.35]YNINVVDAGGVASFGG[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]IFSVAAGTAVGVGIN[396.32]	
280	10	5.1	0.6	HEC_0020.1537.1549.2.pkl.txt	[414.38]NGGNVVDAGGEQFI[157.54][215.86]	MGLVIKAALGALVVLLIGVLAK
278	10	6.5	0.7	HEC_0020.1387.1391.2.pkl.txt	[327.27]SGGIGGVVPGSMGQAATS[303.99]	MGLVIKAALGALVVLLIGVLAK
275	10			HEC_0020.1294.1298.2.pkl.txt	[414.45]NGGNVVEGGAGQFI[373.48]	
274	9			HEC_0020.1510.1510.2.pkl.txt	[324.48]ITIIATGGTIAGGGD[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]NINVVADGGEGAFI[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]GGNAGGVGGADVVNHI[157.85][215.56]	
262	9	6.7	1.1	HEC_0020.1583.1588.0.pkl.txt	[414.39]GGINVVADNEQFI[373.44]	MGLVIKAALGALVVLLIGVLAK
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]	

Figure 30 Sherenga top-level report

5 Review Sherenga results:

a Click a link under the **Filename** header. Check that you see the Spectrum Viewer as shown in Figure 31 and detailed Sherenga results as shown in Figure 32. 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 31 to navigate through the *de novo* possibilities. See "To use the Spectrum Viewer" on page 87 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 32 to view details of the Sherenga interpretation.

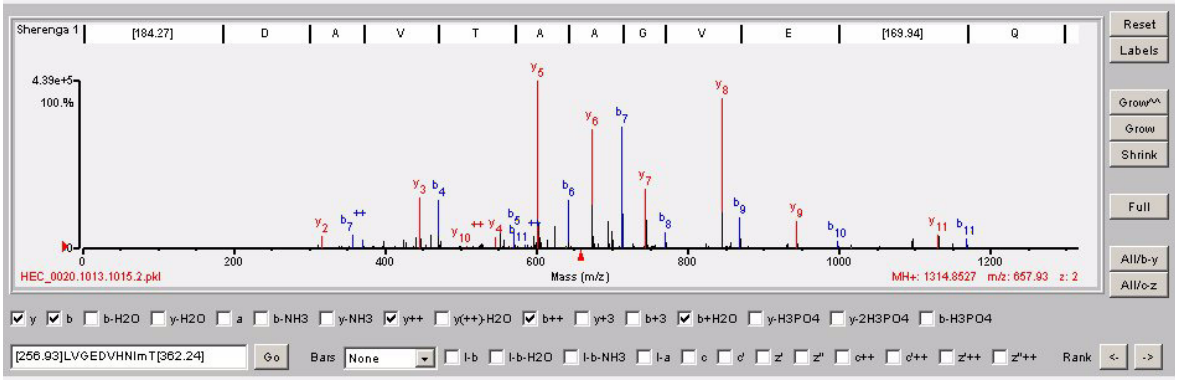


Figure 31 Spectrum Viewer with Sherenga results

ExampleData\Agilent\Trap\ecoli\heatshock HEC_0020.1013.1015.2.pkl.txt

Rank	Orig. Score	Score	MS-Tag Sequence	Sherenga Sequence Tag	Matched Precursor MH+	Delta Precursor MH+	Measured Precursor m/z
				[184.27]DAVTAAGVE[169.94]Q			
1.	179.2	179.2	[184.27] D A V T A A G V E [169.94] Q		1314.85		657.93
2.	176.3	176.3	[184.27] D A V T A A G V [227.53] A Q				
3.	174.0	174.0	[184.27] W V T A A G V E [169.94] Q				
4.	173.6	173.6	[184.27] D A A E A A G V E [169.94] Q				
5.	171.8	171.8	[184.17] D A A E A A G V E [169.99] Q				
6.	171.1	171.1	[299.52] A V T A A G V E [169.94] Q				
7.	170.7	170.7	[184.27] D A A E A A G V [227.53] A Q				
8.	169.7	169.7	[184.27] D A V T A A G V E [298.35]				
9.	169.0	169.0	[184.27] W V T A A G V [227.53] A Q				
10.	169.0	169.0	[184.17] D A A E A A G V [227.59] A Q				

Figure 32 Detailed Sherenga results

6 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 31.

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 177.

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 33.
3. Fill in the options. They are self-explanatory or are explained in the online help.
4. Click the **Match** button.
5. 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. See the lower left hand corner of Figure 33 as an example. Here, the extracted spectra were searched from only a single data file.
6. After time for processing, check that you see results like those shown in Figure 34.

   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 34.

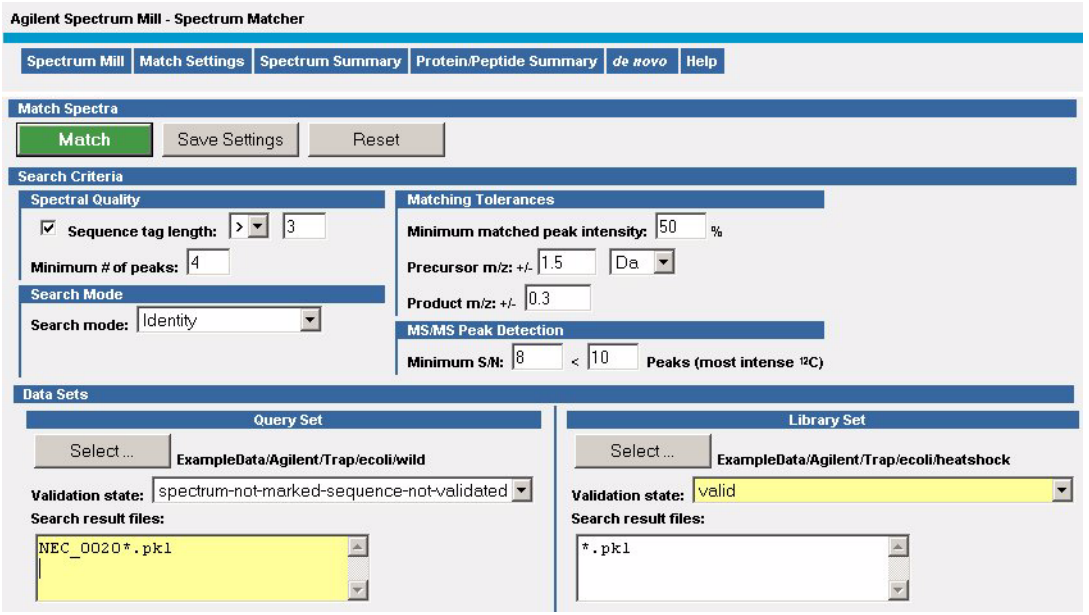
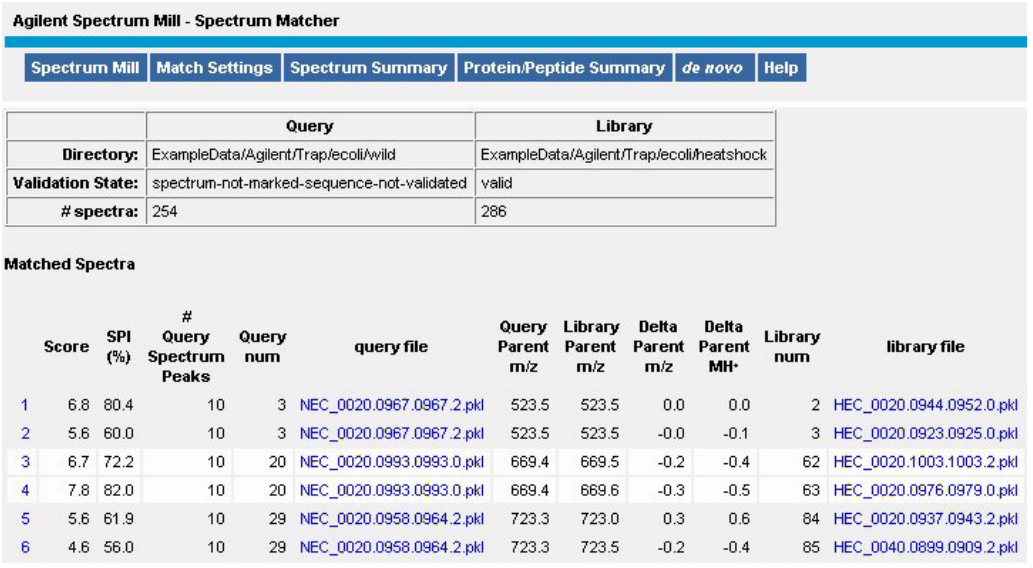


Figure 33 Spectrum Matcher



	Query	Library
Directory:	ExampleData/Agilent/Trap/ecoli/wild	ExampleData/Agilent/Trap/ecoli/heatshock
Validation State:	spectrum-not-marked-sequence-not-validated	valid
# spectra:	254	286

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
6	4.6	56.0	10	29	NEC_0020.0958.0964.2.pkl	723.3	723.5	-0.2	-0.4	85	HEC_0040.0899.0909.2.pkl

Figure 34 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 **msparams_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 35 as an example.

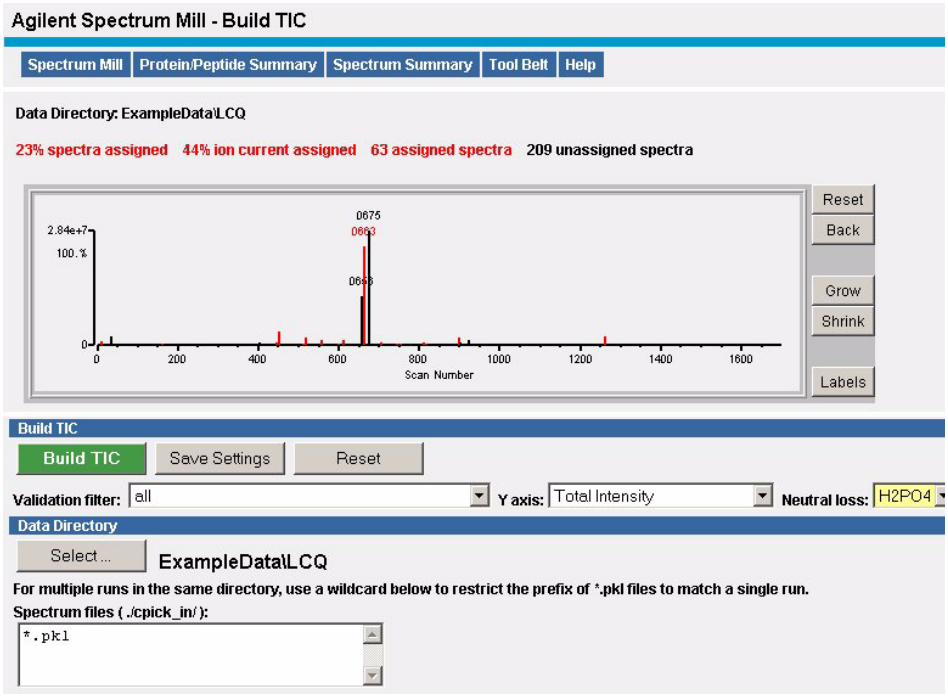


Figure 35 Use of Build TIC to locate scan numbers of spectra that indicate potential phosphorylation sites.

2 MS/MS Data Review and Validation

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 the highest scores. You learn how to use the Protein/Peptide Summary page to manually review and validate search results with medium scores. You learn how the various Protein/Peptide Summary data display modes 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, “Basics for Processing MS/MS Data”.

Autovalidation

Automatic results validation saves time and allows you to set score thresholds that meet the data quality needs of your experiment.

1 Navigate to the **MS/MS Autovalidation** page, as shown in Figure 36. You navigate to this page from the MS/MS Search page or the Protein/Peptide Summary page.

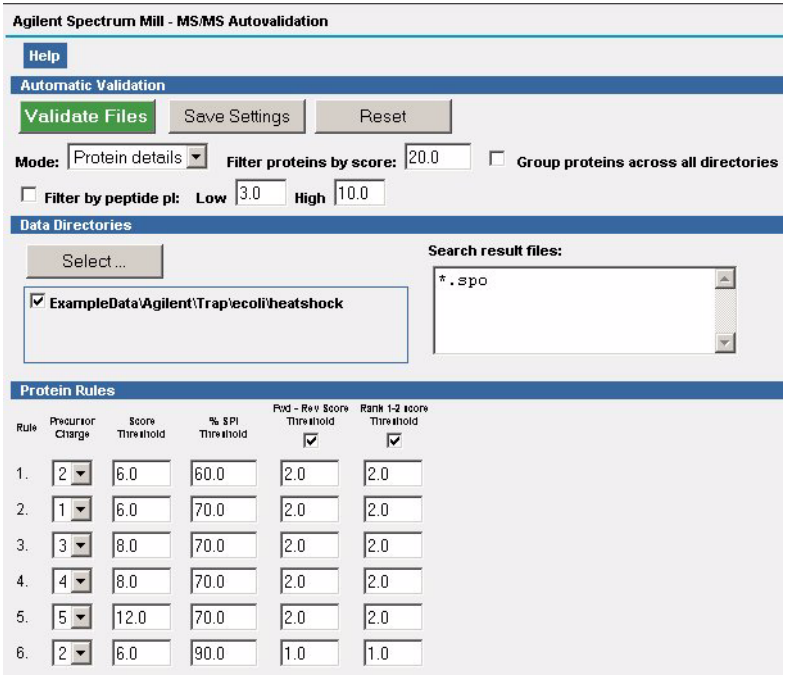


Figure 36 Automatic validation of MS/MS search results

2 Select the **Data Directory** for which you want to validate results. You may validate twice – first in **Protein details mode** and second in **Peptide mode**.

3 Select the **Protein details** mode.

- In this mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein.

- Using the default scoring, individual peptides must have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score must be greater than 20.

4 If the data files from a single sample occupy multiple directories, mark the check box for **Group proteins across all directories**.

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**.

5 Click the **Validate Files** button.

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

7 Now select the **Peptide** mode.

- 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 must 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.

8 If you wish to autovalidate only the peptides in a given pI range, mark the check box for **Filter by 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.

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**.

9 Click the **Validate Files** button.

10 You quickly see a validation summary that lists the new hits and spectra that have been validated, along with those validated previously.

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 Click the **Mode** drop-down list and select one of the summary modes shown in Figure 37. Use Table 3 to guide you. For more details on the display modes, see "Data display modes on the Protein/Peptide Summary page" on page 94.

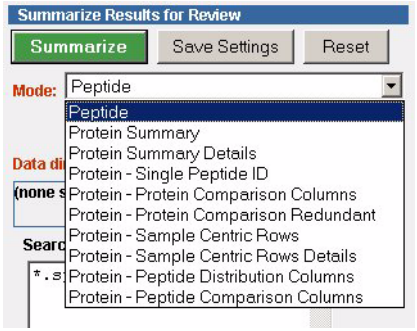


Figure 37 Selecting the data display mode

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

Step 2. Set filtering, sorting, and validation parameters

- Set the validation, sorting, and filtering parameters shown in Figure 38. To set these parameters, refer to the guidelines below. Note that the specific options that are available depend upon the mode you selected earlier.

  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 determine 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.

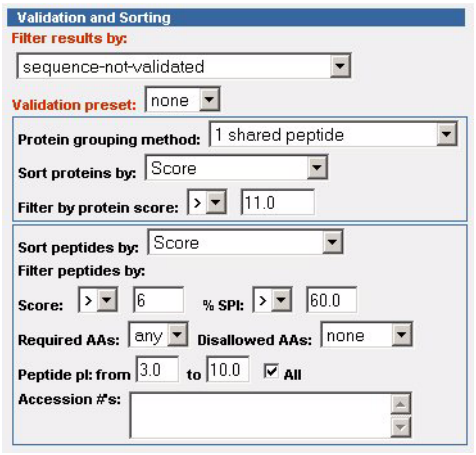


Figure 38 Validation, sorting and filtering parameters - Protein Summary Details mode, with check box for **Peptide pI** marked under **Review Fields**

Filter results by The validation filter determines which data are displayed. Select one of the following options:

- **sequence-not-validated**: Displays only search results that have *not* yet been categorized as "valid" with the Autovalidation page or the Protein/Peptide Summary page.
- **valid**: Displays only search results that *have* been categorized as "valid" with the Autovalidation page or the Protein/Peptide Summary page.
- **good-spectrum-sequence-not-validated**: Displays only data where the spectrum has been categorized as "good" with the Spectrum Summary page, but the search results have not been categorized as "valid" with the Autovalidation page or the Protein/Peptide Summary page.
- **all**: Displays all data, regardless of current validation state.

Validation preset The validation preset determines how you categorize your data upon initial display. Select one of the following options:

- **none**: Select if you do not want to validate data right now. Choose this setting if you want only a summary report. Also choose this setting if you want to compare data in two or more data directories.
- **valid**: Select if you want to initially categorize all data in your report as "valid." If you filter the data for only search results that have a high probability of being correct (described below), select this setting to minimize manual changes during data review. You can later change the setting as you review the data, but before you exit the form, be sure to click the **Perform Validation** button to save the new validation state.
- **reset**: Select if you *do not* want to initially categorize all data in your report as "valid." The "reset" setting has two uses: either the data have not yet been reviewed and validated, or the data have been reviewed but the results have not been satisfactory to this point. You can change validation states to "valid" as you review the data, but before you exit the form, be sure to click the **Perform Validation** button to save the new validation state.
- **status**: Select if you want to see all peptides that belong to a protein, regardless of validation state. To generate such a display, set **Filter results by:** to **all** and set **Mode** to one of the display modes that show both proteins and peptides. When the data are displayed, look under **Validation category** to see if a particular peptide was validated (**V**) or not validated (**R**). (**R** stands for **reset**.) You may also change the validation state, but before you exit the form, be sure to click the **Perform Validation** button to save the new validation state.

Protein grouping method Determines how proteins are grouped in certain protein summary modes. Select one of the following options:

- **1 shared peptide**: When a peptide sequence >8 residues long is contained in multiple protein entries in the sequence database, the software groups the proteins together and then reports the highest-scoring one and its accession number.
- **1 shared peptide, expand subgroups**: The software initially groups the proteins as described for **1 shared peptide**. In some cases when the protein sequences are grouped in this manner, there are distinct peptides that uniquely represent a lower-scoring member of the group (isoforms and family members). When you choose **1 shared peptide, expand subgroups**, more than one member of the group is reported and counted towards the total number of proteins. The proteins are given related protein subgroup numbers (for example, 11.1 and 11.2 for protein group 11).

Sort proteins by Select a data sorting method for your report. Choices include score, number of peptides, protein molecular weight, category and function.

Filter by protein score Set a filter for protein score. Scores greater than 25 almost always represent valid results.

Sort peptides by 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.

Filter peptides by Score Set a filter for peptide score. For ion trap data, peptide scores greater than 13 almost always represent valid results. Scores below 6 generally represent poor results. For more details, see Table 1 on page 37. 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.

%SPI Set a filter for scored peak intensity. This is the percentage of the MS/MS peak-detected ion current explained by theoretical fragments from the database hit. 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.

Required AAs Filters search results so that peptides are shown only if they contain the required amino acid(s). To disable, select **any**.

Disallowed AAs Filters search results so that peptides are *not* shown if they contain disallowed amino acid(s). To disable, select **none**.

Peptide pI Filters by a range of peptide pI values

Accession #'s Filters search results by accession numbers. You can 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**.

To enable the capability to filter by peptide pI, mark the check box for **Peptide pI** under **Review Fields** and then look for the pI filter under **Validation and Sorting**.

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

- Mark the check boxes for data you wish to display. Choose from items shown in Figure 39. See the text below for additional information.

  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.

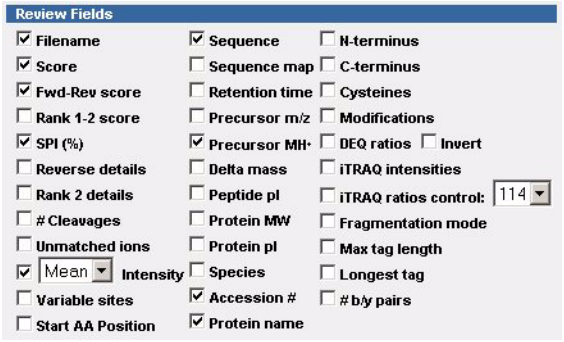


Figure 39 Fields for display in Protein/Peptide Summary form

Fwd-Rev score 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.

Rank 1-2 score Difference between rank 1 and rank 2 database search scores

Intensity Calculates and includes peptide intensities. The software calculates the intensities from the extracted ion chromatograms of the precursor ions for each peptide. The software then calculates either a mean or total intensity for the peptides that make up the protein. Studies have shown that the total intensity provides more accurate quantitation than the mean intensity.

Variable sites Displays sites of variable modifications and amino acid substitutions

Sequence map Displays the amino acid sequence of the matched peptide from the database search, annotated with the following:

- Red forward-slashes for locations of y-ions
- Blue backslashes for locations of b-ions
- Magenta pipes (vertical lines) for locations of both b- and y-ions

	The display allows you to assess b- and y-ion coverage without inspecting the spectrum, and helps to identify the site of a post-translational modification by highlighting existence of surrounding b/y ions.
peptide pI	Displays peptide pI, and enables filtering by peptide pI
Excel export	Exports data to Excel or LIMS
DEQ ratios	Calculates and displays 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 134 of Chapter 5.
iTRAQ intensities	Displays intensities for iTRAQ marker ion masses
iTRAQ ratios control	Selects the iTRAQ mass for the denominator for ratio calculations
Fragmentation mode	Shows the fragmentation mode, for example, collision-induced dissociation (CID) or electron transfer dissociation (ETD)

Step 4. Summarize data

- Click the **Summarize** button to display results in tabular format.

Reviewing and validating results using the Protein/Peptide Summary page

To perform the overall procedure

- See "Step 8. Manually review and validate results" on page 39 of Chapter 1.

To use the Spectrum Viewer

Figure 40 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.

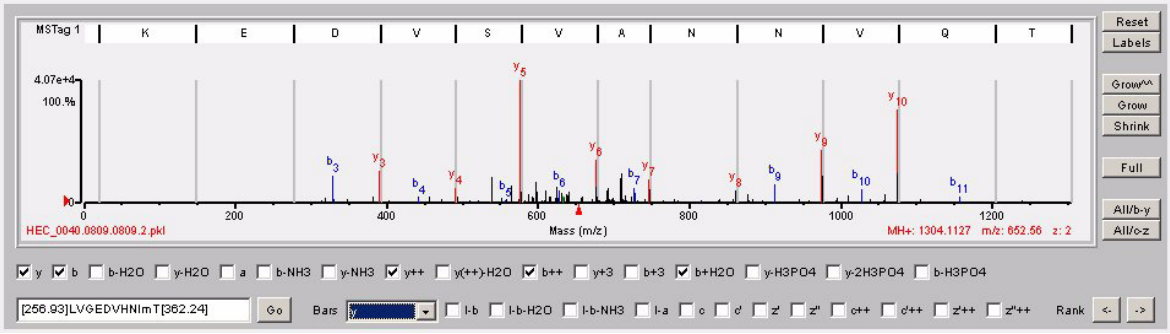


Figure 40 Spectrum Viewer, with y-ions highlighted using the **Bars** feature

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 40, 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 l-x Mark these if you want to label peaks with amino acid sequences combined with common losses.

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.

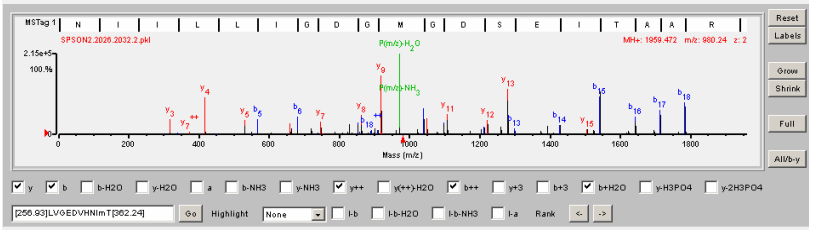
To produce final results summaries

- See "Step 9. Summarize valid results" on page 43 of Chapter 1.

To manually validate results

For MS/MS spectra from collision-induced dissociation (CID), see guidelines in Table 4. For MS/MS spectra from electron transfer dissociation (ETD), see "Interpreting ETD spectra" on page 126.

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.	

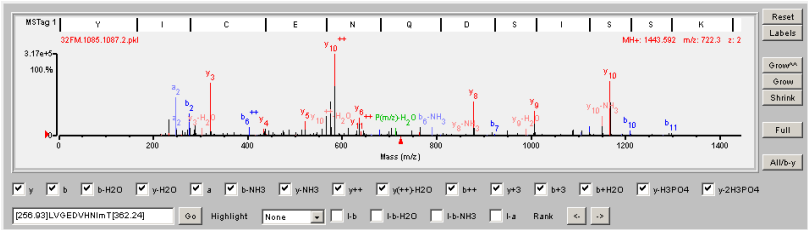
• Loss of NH_3 occurs from R, K, Q and N residues.
• Loss of H_2O occurs from S, T, E and D.
• R, H, K and N are charge-bearing residues and increase the maximum charge state allowed for a peptide fragment.
• Enhanced fragmentation may be observed at the following bonds:
  ◦ His – Xaa
  ◦ Xaa – Gly
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

Table 4 Guidelines for manual results validation

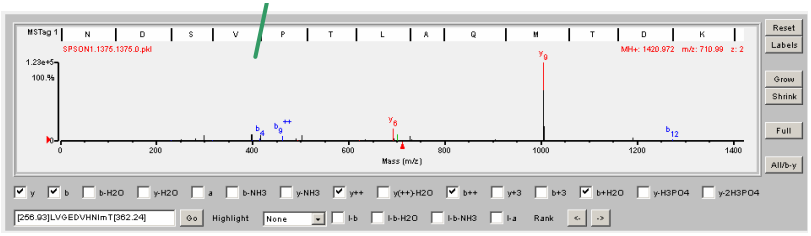
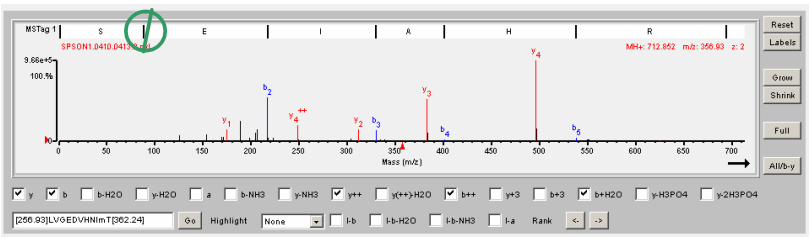
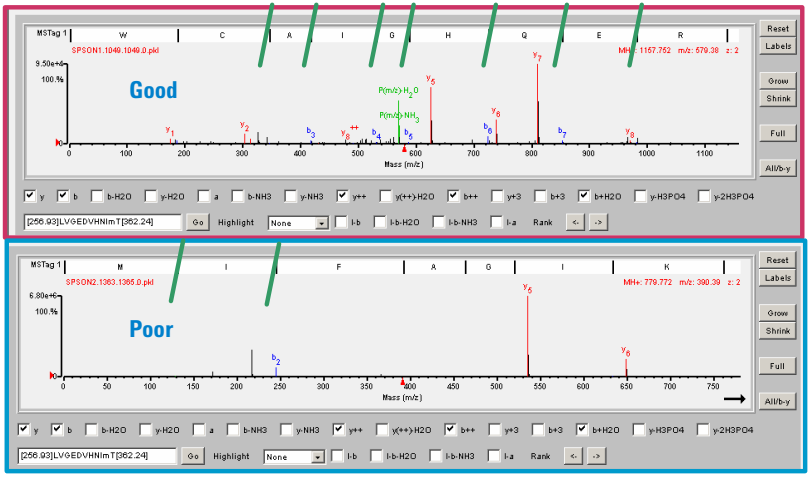
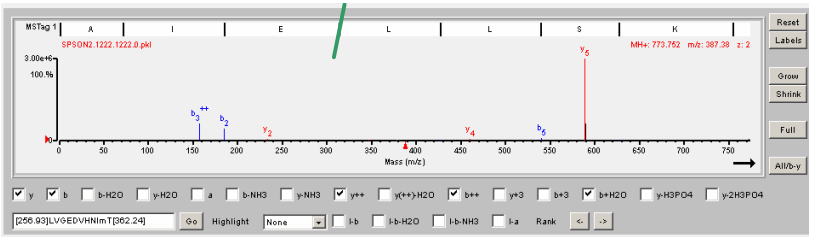
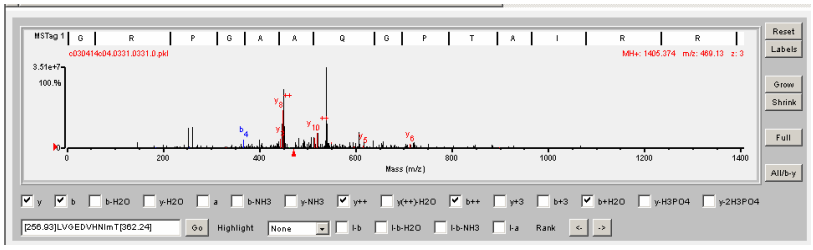
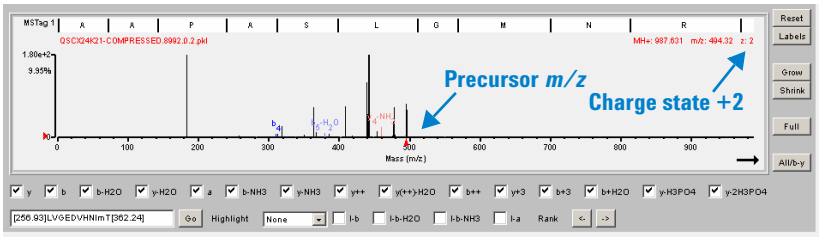
Guideline	Example
The peptide bond on the N-terminal side of Pro is particularly labile (high local proton affinity) and usually results in an MS/MS spectrum dominated by the y-ion ending in Pro.*	
It is typical to lack fragmentation between the first two N-terminal amino acids (i.e., no b_1 or y_{x-1} ions).	
Look for coverage over a significant portion of the backbone. A series of consecutive b or y fragments is better than a random distribution.	

Table 4 Guidelines for manual results validation

Guideline	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).	
Be suspicious if no ions appear above the precursor ion (when precursor charge state is >1).	
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.

Data display modes on the Protein/Peptide Summary page

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 41. 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.

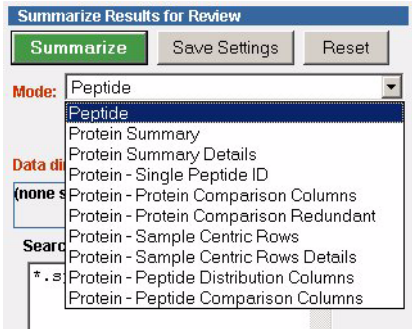


Figure 41 Data display modes

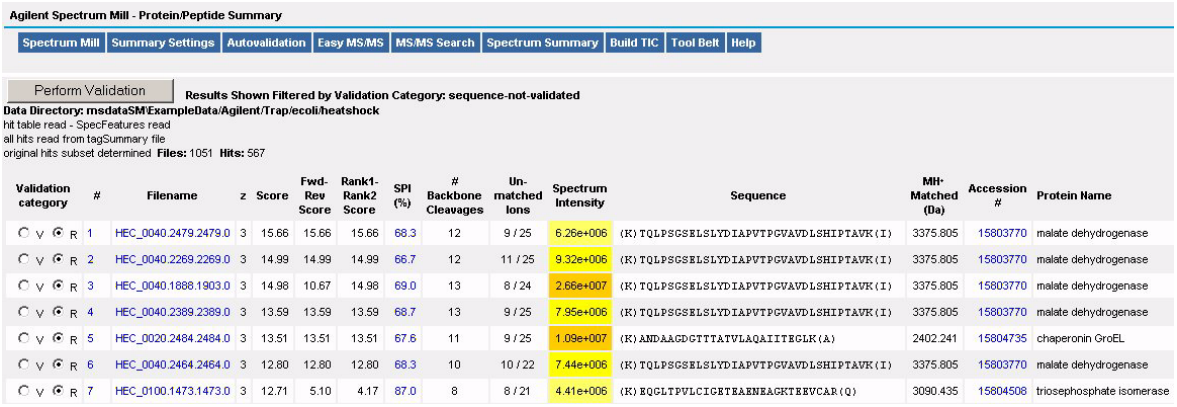
Peptide

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 42.

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 34.



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: msdataSM\ExampleData/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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)ANDAAGDGTTTATVLAQAIITEGLK(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)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(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)EQGLTPVLCIGETEAENEAGKTEEVCAR(Q)	3090.435	15804508	triosephosphate isomerase

Figure 42 Peptide display mode

Protein Summary

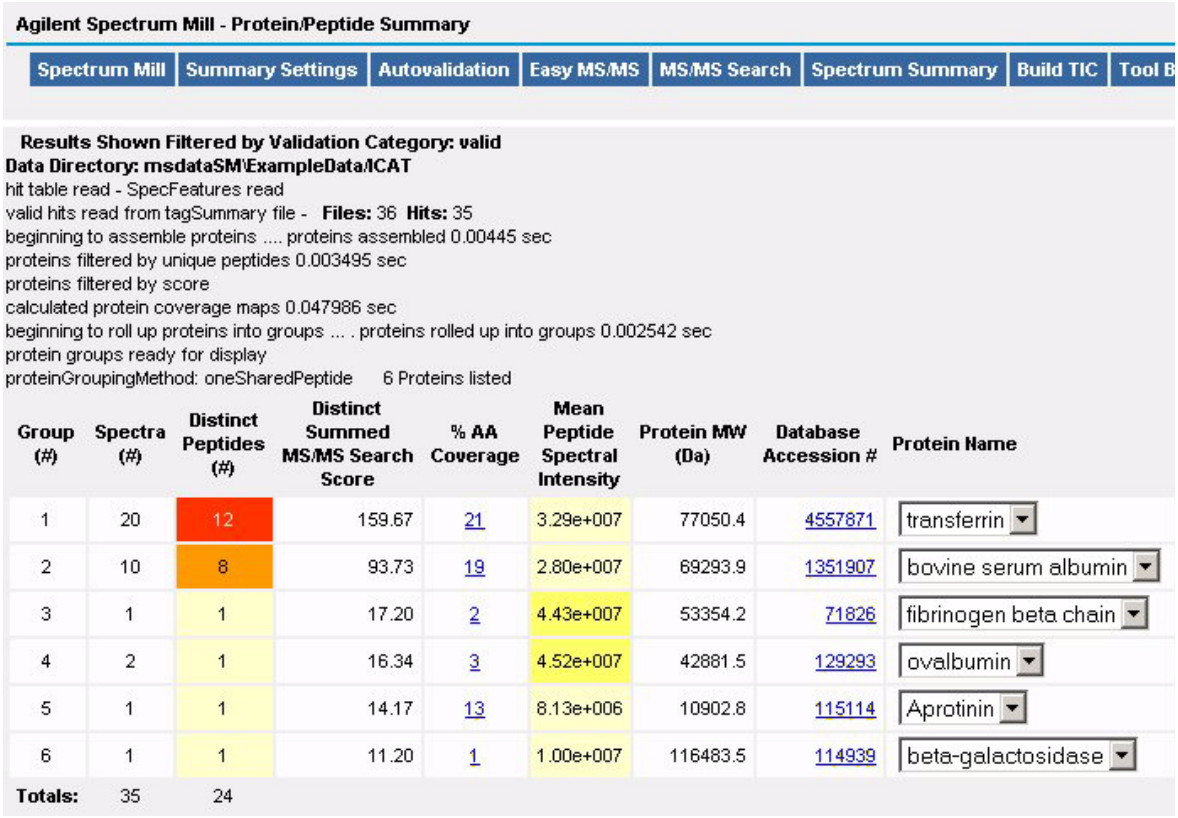
Select this mode to display results organized by protein. This mode is useful when your sample fractions are organized in a single data directory, and you want the peptides from a single protein organized together even if they originate from different data files.

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 43.

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 43 Protein Summary display mode

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 44, while the bottom of the summary shows the individual peptides for each protein, as in Figure 45. 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.

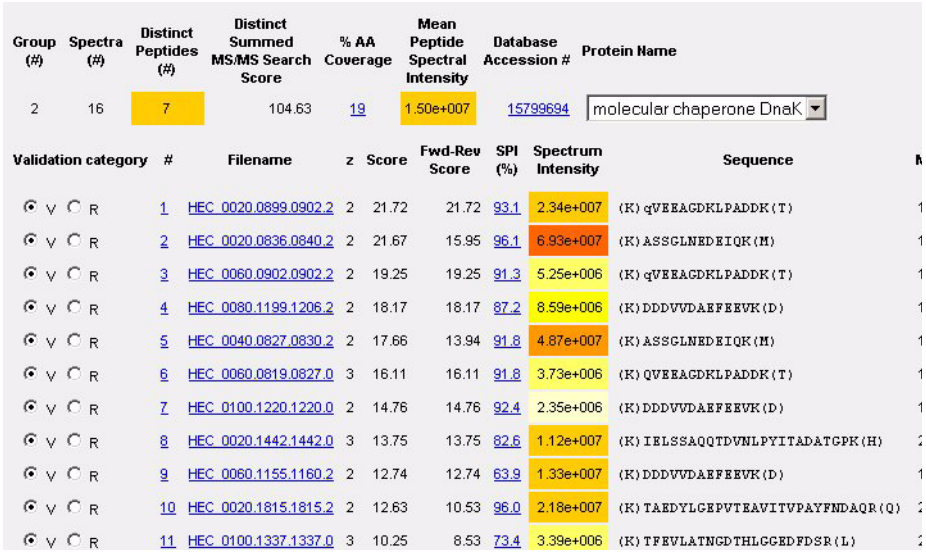
Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build

Perform Validation **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 0.388199 sec
proteins filtered by unique peptides 0.141813 sec
proteins filtered by score
calculated protein coverage maps 0.454553 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.184269 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide 64 Proteins listed

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
1	98	7	125.18	18	1.34e+007	38491462	GroEL
1	98	7	125.18	18	1.34e+007	38491468	GroEL
1	98	7	125.18	18	1.34e+007	38491464	GroEL
1	98	7	125.18	18	1.34e+007	38491478	GroEL
1	98	7	125.18	18	1.34e+007	18028150	GroEL
1	98	7	125.18	20	1.34e+007	18028158	GroEL
1	97	6	113.74	18	1.34e+007	38491476	GroEL
1	97	6	104.87	22	1.30e+007	37624680	GroEL protein
1	97	6	104.87	22	1.30e+007	37624678	GroEL protein
1	89	6	103.07	15	1.12e+007	15804735	chaperonin GroEL
1	91	6	101.81	15	1.31e+007	38491472	GroEL
2	16	7	104.63	19	1.50e+007	15799694	molecular chaperone DnaK
2	3	1	18.17	22	8.08e+006	40891648	DnaK
3	15	4	76.96	9	3.39e+007	15803853	elongation factor EF-2
3	1	1	15.25	72	4.07e+007	147896	elongation factor G
4	11	3	49.13	9	8.59e+006	15800756	seryl-tRNA synthetase

Figure 44 Protein Summary Details display mode - top

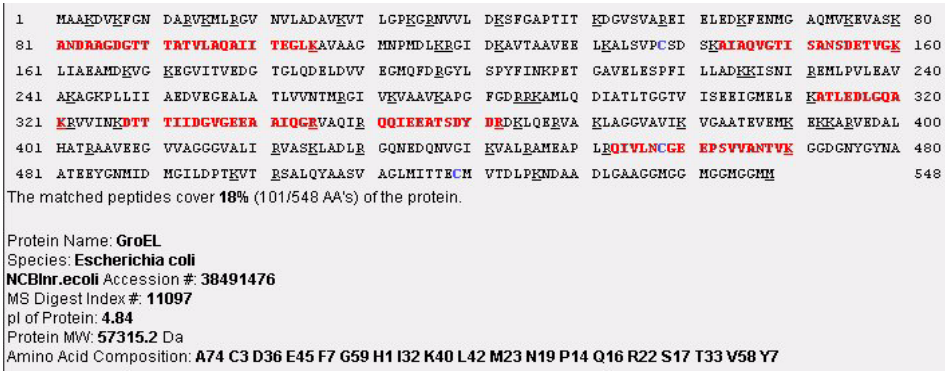


Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Database Accession #	Protein Name
2	16	7	104.63	19	1.50e+007	15799694	molecular chaperone DnaK

Validation category	#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence
V R	1	HEC_0020.0899.0902.2	2	21.72	21.72	93.1	2.34e+007	(K)qVEEAGDKLPADDK(T)
V R	2	HEC_0020.0836.0840.2	2	21.67	15.95	96.1	6.93e+007	(K)ASSGLNEDEIQK(M)
V R	3	HEC_0060.0902.0902.2	2	19.25	19.25	91.3	5.25e+006	(K)qVEEAGDKLPADDK(T)
V R	4	HEC_0080.1199.1206.2	2	18.17	18.17	87.2	8.59e+006	(K)DDDVVDAEFEEVK(D)
V R	5	HEC_0040.0827.0830.2	2	17.66	13.94	91.8	4.87e+007	(K)ASSGLNEDEIQK(M)
V R	6	HEC_0060.0819.0827.0	3	16.11	16.11	91.8	3.73e+006	(K)QVEEAGDKLPADDK(T)
V R	7	HEC_0100.1220.1220.0	2	14.76	14.76	92.4	2.35e+006	(K)DDDVVDAEFEEVK(D)
V R	8	HEC_0020.1442.1442.0	3	13.75	13.75	82.6	1.12e+007	(K)IELSSAQQTDVNLPYITADATGPK(H)
V R	9	HEC_0060.1155.1160.2	2	12.74	12.74	63.9	1.33e+007	(K)DDDVVDAEFEEVK(D)
V R	10	HEC_0020.1815.1815.2	2	12.63	10.53	96.0	2.18e+007	(K)TAEDYLGEPVTEAVITVPAYFNDAQR(Q)
V R	11	HEC_0100.1337.1337.0	3	10.25	8.53	73.4	3.39e+006	(K)TFEVLATNGDTHLGGEDFDSR(L)

Figure 45 Protein Summary Details display mode - bottom

To view sequence coverage, click a link under the **% AA Coverage** heading. Check that you see a display like that in Figure 46. Other protein display modes have links to sequence coverage as well.



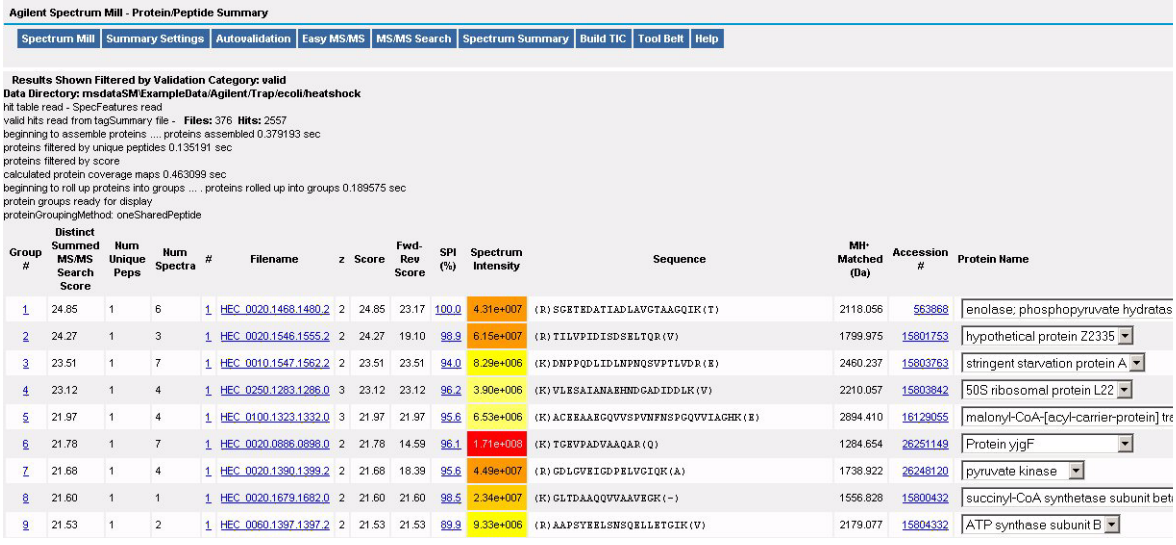
```
1   MAAKDVKFGN DARVKMLRGV NVLADAVKVT LGPKGRNVVL DKSFGAPTIT KDGVSVAREI ELEDKFENMG AQMVKEVASK 80
81  ANDAAGDGTT TATVLAQAII TEGLKAVAAG MNPMDLKRGI DKAVTAAVEE LKALSVPCSD SKAIAQVGTI SANSDETVGK 160
161 LIAEAMDKVG KEGVITVEDG TGLQDELDVV EGMQFDRGYL SPYFINKPET GAVELESPFI LLADKKISNI REMLPVLEAV 240
241 AKAGKPLLII AEDVEGEALA TLVVNTMRGI VKVAAVKAPG FGDRRKAMLQ DIATLTGGTV ISEEIGMELE KATLEDLGQA 320
321 KRVVINKDTT TIIDGVGEEA AIQGRVAQIR QQIEEATSDY DREKLQERVA KLAGGVAVIK VGAATEVEMK EKKARVEDAL 400
401 HATRAAVEEG VVAGGGVALI RVASKLADLR GQNEDQNVGI KVALRAMEAP LRQIVLNCGE EPSVVANTVK GGDGNYGYNA 480
481 ATEEYGNMID MGILDPTKVT RSALQYAASV AGLMITTECM VTDLPKNDAA DLGAAGGMGG MGGMGGMM             548
The matched peptides cover 18% (101/548 AA's) of the protein.

Protein Name: GroEL
Species: Escherichia coli
NCBInr.ecoli Accession #: 38491476
MS Digest Index #: 11097
pI of Protein: 4.84
Protein MW: 57315.2 Da
Amino Acid Composition: A74 C3 D36 E45 F7 G59 H1 I32 K40 L42 M23 N19 P14 Q16 R22 S17 T33 V58 Y7
```

Figure 46 Sequence coverage

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 47. 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.)



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/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 0.379193 sec
proteins filtered by unique peptides 0.135191 sec
proteins filtered by score
calculated protein coverage maps 0.463099 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.189575 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide

Group #	Distinct Summed MS/MS Search Score	Num Unique Peps	Num Spectra	#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence	MH+ Matched (Da)	Accession #	Protein Name
1	24.85	1	6	1	HEC_0020.1468.1480.2	2	24.85	23.17	100.0	4.31e+007	(R)SGETEDATIADLAVGTAAGQIK(T)	2118.056	563868	enolase; phosphopyruvate hydratas
2	24.27	1	3	1	HEC_0020.1546.1555.2	2	24.27	19.10	98.9	6.15e+007	(R)TILVPIDISDSELTQR(V)	1799.975	15801753	hypothetical protein Z2335
3	23.51	1	7	1	HEC_0010.1547.1562.2	2	23.51	23.51	94.0	8.29e+006	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237	15803763	stringent starvation protein A
4	23.12	1	4	1	HEC_0250.1283.1286.0	3	23.12	23.12	96.2	3.90e+006	(K)VLESAIANAEHNDGADIDDLK(V)	2210.057	15803842	50S ribosomal protein L22
5	21.97	1	4	1	HEC_0100.1323.1332.0	3	21.97	21.97	95.6	6.53e+006	(K)ACEEAAEGQVVSPVNFNSPGQVVIAGHK(E)	2894.410	16129055	malonyl-CoA-[acyl-carrier-protein] tr
6	21.78	1	7	1	HEC_0020.0886.0898.0	2	21.78	14.59	96.1	1.71e+008	(K)TGEVPADVAAQAR(Q)	1284.654	26251149	Protein yjgF
7	21.68	1	4	1	HEC_0020.1390.1399.2	2	21.68	18.39	95.6	4.49e+007	(R)GDLGVEIGDPELVGIQK(A)	1738.922	26248120	pyruvate kinase
8	21.60	1	1	1	HEC_0020.1679.1682.0	2	21.60	21.60	98.5	2.34e+007	(K)GLTDAAQQVVAAVEGK(-)	1556.828	15800432	succinyl-CoA synthetase subunit bet
9	21.53	1	2	1	HEC_0060.1397.1397.2	2	21.53	21.53	89.9	9.33e+006	(R)AAPSYEELSNSQELLETGIK(V)	2179.077	15804332	ATP synthase subunit B

Figure 47 Protein-Single Peptide ID display mode

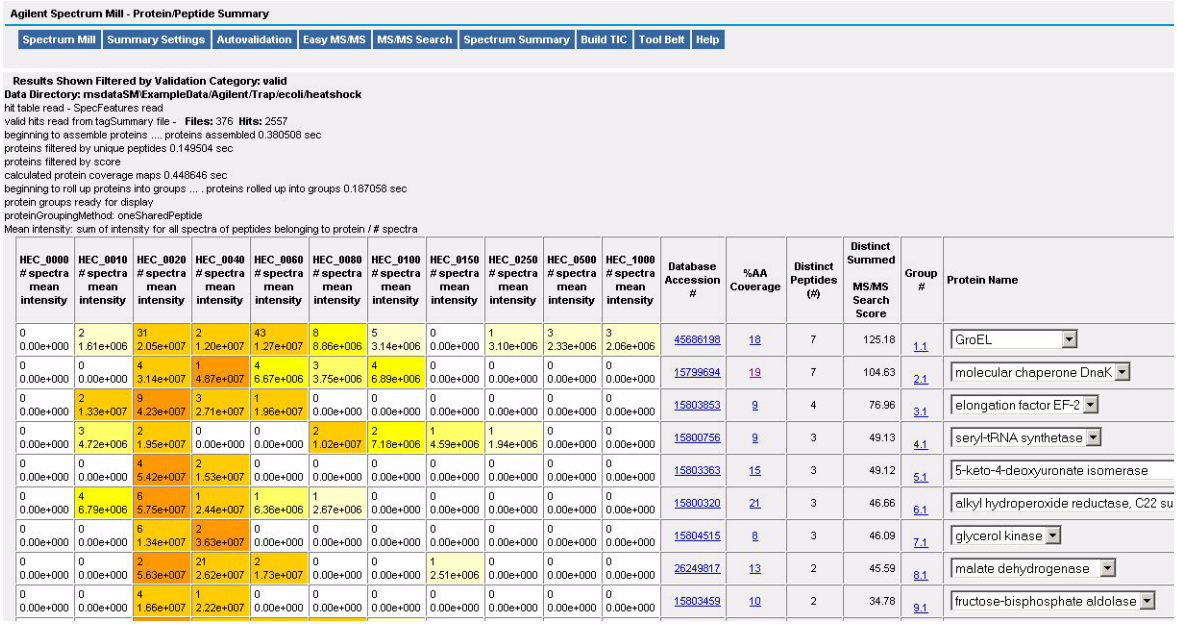
Protein-Protein Comparison Columns

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 48 and Figure 49 show examples of each type of display. Figure 48 shows the distribution of proteins across a 2D LC/MS/MS run (individual fractions), while Figure 49 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 18 on page 46.



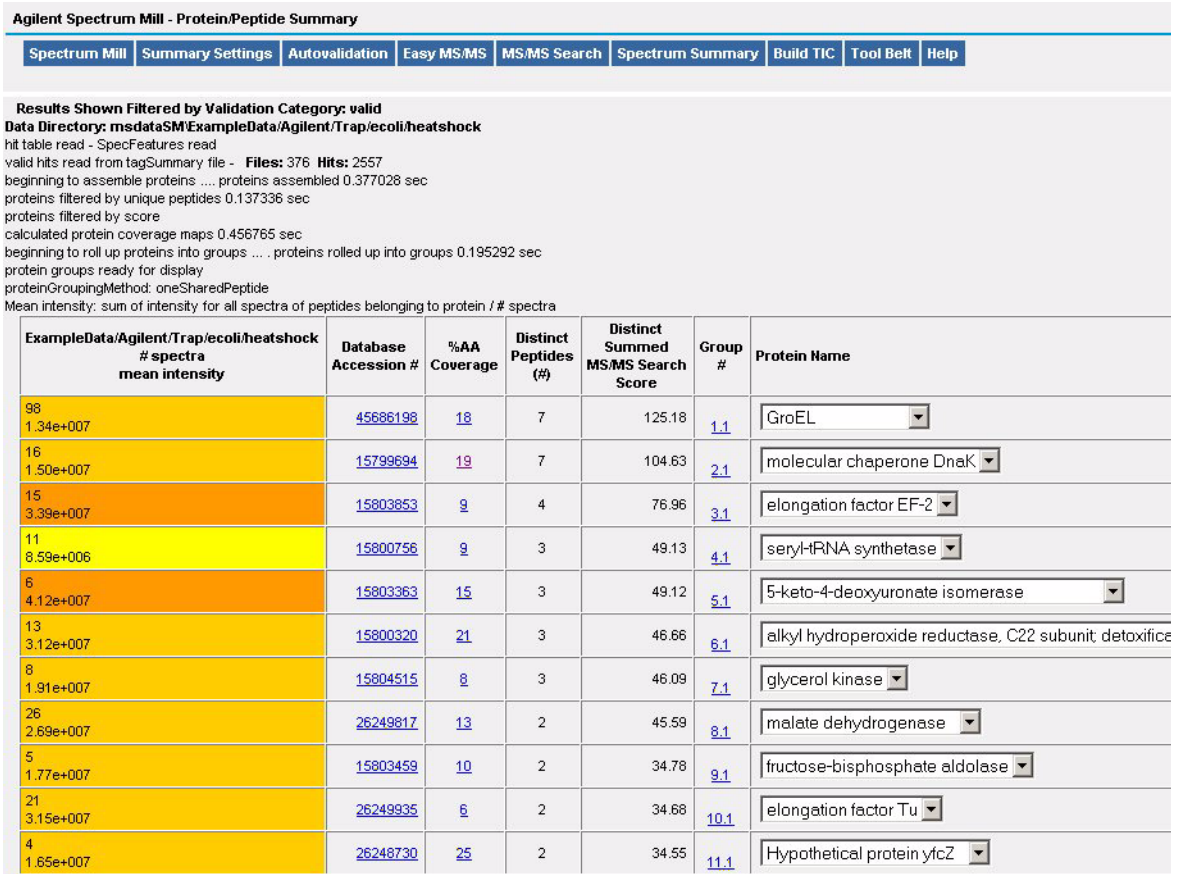
Agilent Spectrum Mill - Protein/Peptide Summary

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/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 0.380508 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 (#)	Distinct Summed MS/MS Search Score	Group #	Protein Name
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	125.18	1.1	GroEL
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	104.63	2.1	molecular chaperone DnaK
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	76.96	3.1	elongation factor EF-2
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	49.13	4.1	seryl-tRNA synthetase
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	49.12	5.1	5-keto-4-deoxyuronate isomerase
0 0.00e+000	4 6.79e+006	6 5.75e+007	1 2.44e+007	1 6.36e+006	1 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	46.66	6.1	alkyl hydroperoxide reductase, C22 su
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	46.09	7.1	glycerol kinase
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	45.59	8.1	malate dehydrogenase
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	34.78	9.1	fructose-bisphosphate aldolase

Figure 48 Protein-Protein Comparison Columns display mode, **Group results by File**



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/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; detoxifice
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	26248730	25	2	34.55	11.1	Hypothetical protein yfcZ

Figure 49 Protein-Protein Comparison Columns display mode, **Group results by Directory**

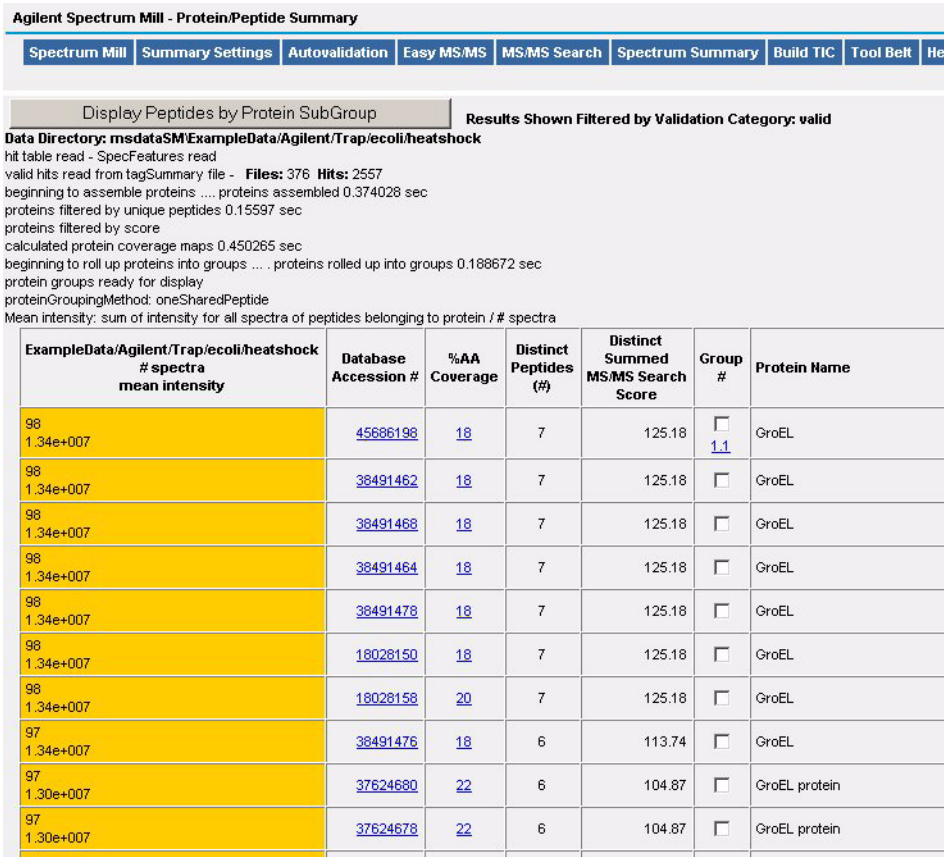
If you want the software to align and compare amino acid sequences of proteins within a protein group, use one of these modes:

- Protein-Protein Comparison Columns
- Protein-Protein Comparison Redundant (discussed on page 103)

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 50. 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 Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | He

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 50 Protein-Protein Comparison Redundant display 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. If the same protein as been found in separate data files, it 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 51.

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 cl
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 51 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 52.

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 34.

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)AAGEAPAAPQVTAEDASASLAELLNAGLGGSDNE(-)	3167.497

Run #	Run Name	Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Database Accession #	Protein Name
2	HEC_0010	1	2	1	23.73	2	1.33e+007	15803653	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)DVTTGDTLCDPDAPIILER(M)	2101.012
2	HEC_0010.1372.1384.2	2	22.19	22.19	95.1	1.71e+007	(K)DVTTGDTLCDPDAPIILER(M)	2101.012

Run #	Run Name	Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Database Accession #	Protein Name
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)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237
2	HEC_0010.1561.1570.0	3	18.47	12.28	95.7	2.01e+007	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237
3	HEC_0010.1594.1609.0	3	17.54	12.26	87.3	8.24e+006	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237
4	HEC_0010.1534.1543.0	3	15.85	9.56	93.0	1.37e+007	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237
5	HEC_0010.1616.1616.0	3	9.74	6.50	93.3	2.86e+006	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237
6	HEC_0010.1568.1568.2	2	9.39	9.39	95.8	3.11e+006	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237

Figure 52 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 53, 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.

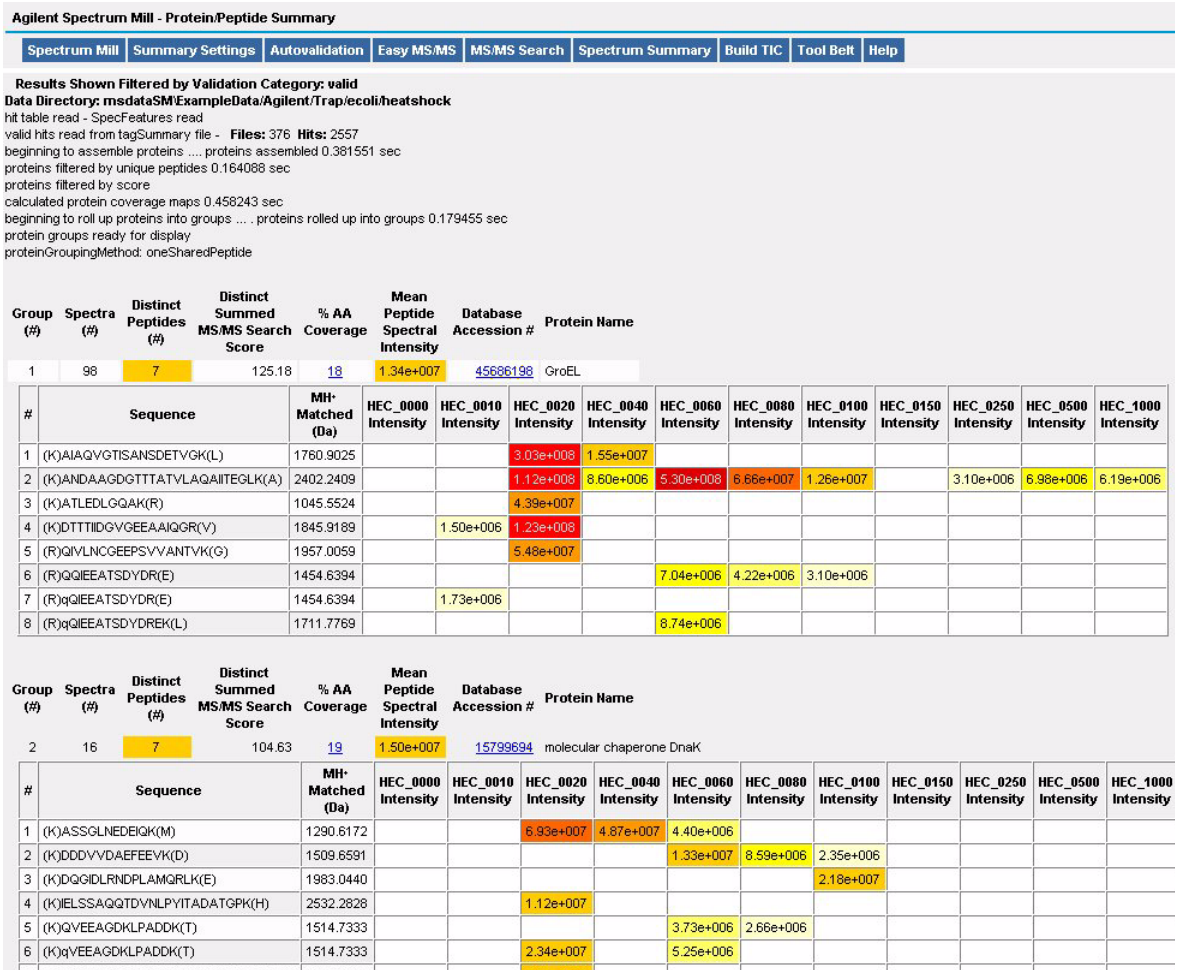


Figure 53 Protein-Peptide Distribution Columns display mode, **Group results by File**

Protein-Peptide Comparison Columns

This display mode, shown in Figure 54, is similar to Protein-Peptide Distribution Columns, but allows filename, score, SPI, and intensity to all 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).

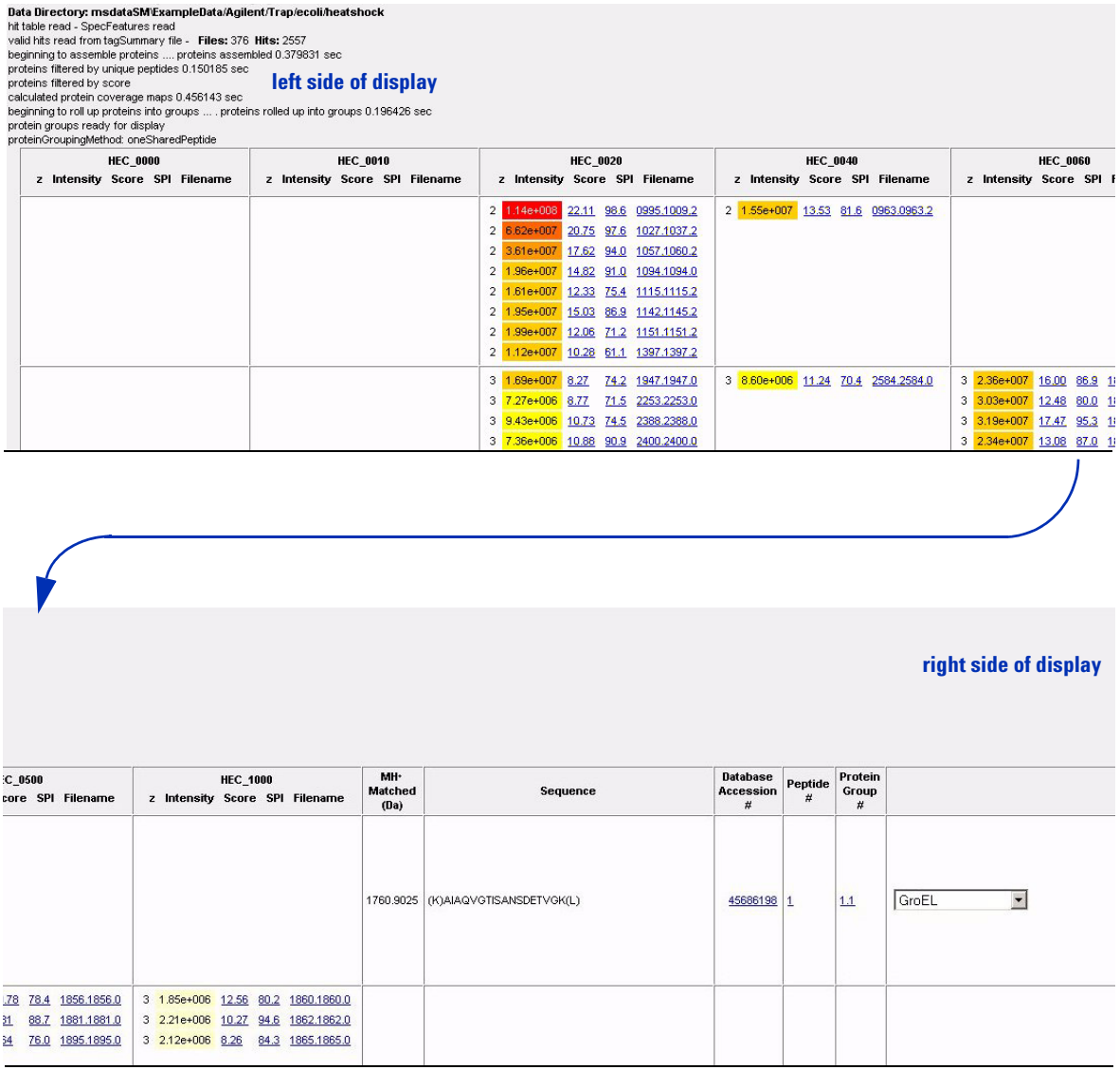


Figure 54 Protein-Peptide Comparison Columns display mode, **Group Results by File**

3
Sherenga *de novo* Sequencing

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 which 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

* 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.

either the N-terminus or the C-terminus, all peaks (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 55.

2 Use the information below to set parameters. In general, you should keep the default values, except for the data directory.

3 Click the **Sequence** button.

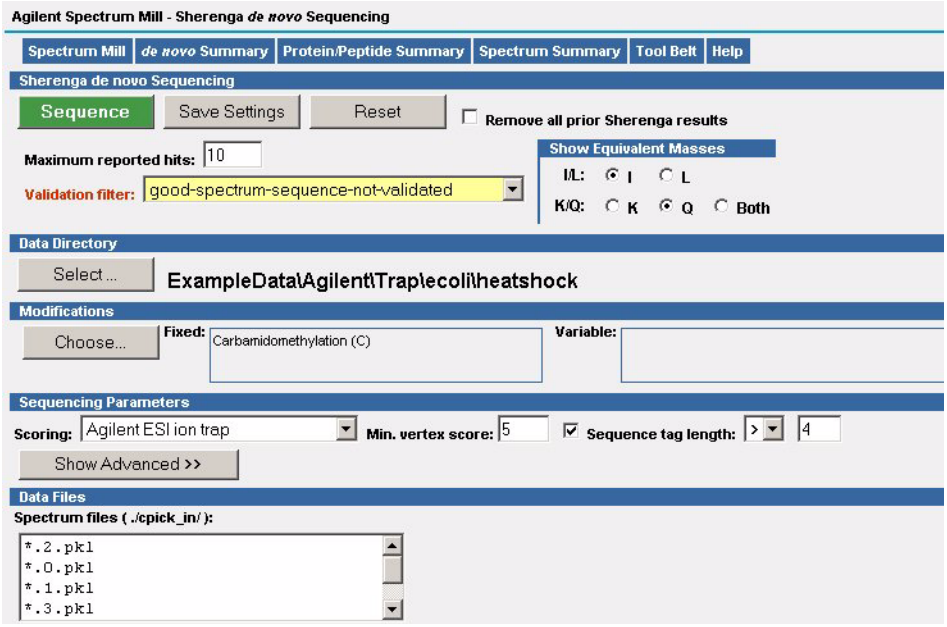


Figure 55 Sherenga *de novo* Sequencing page

Maximum reported hits Set to the maximum number of Sherenga interpretations you would like for each spectrum.

Validation filter Set in one of the following ways:

- If you took the time to designate your spectra as “good” or “bad,” as described in “To check for remaining high-quality spectra” on page 62, 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**.

Show Equivalent Masses

For amino acids that have the same nominal mass, select how you would like to display the results. For **K/Q**, choose **Both** if your instrument has sufficient mass accuracy to distinguish between these two.

Data Directory

Click the **Select...** button to choose the folder where your files reside.

Modifications

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.

Scoring

Choose the appropriate file for Sherenga's learned parameters. Since the learned parameters are instrument-dependent, select your instrument.

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.

Sequence tag length

Mark this option 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.

Show Advanced

Click this button to display advanced **Advanced Parameters**. These are described in the online help.

To generate a Sherenga report

1 While Sherenga runs, check that you see a display like that shown in Figure 56.

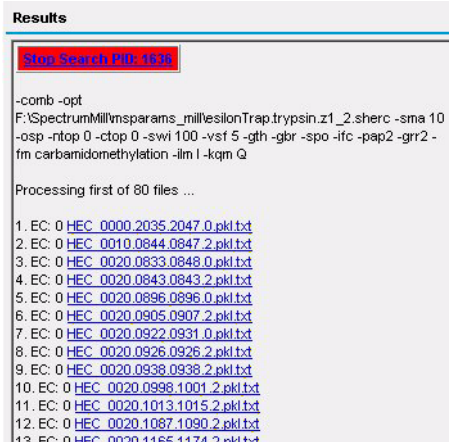


Figure 56 Display while Sherenga *de novo* Sequencing runs

2 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.

3 When the Sherenga algorithm is complete, navigate from the Sherenga *de novo* Sequencing page to the Sherenga *de novo* Summary page.

4 Check that you see the form shown in Figure 57.

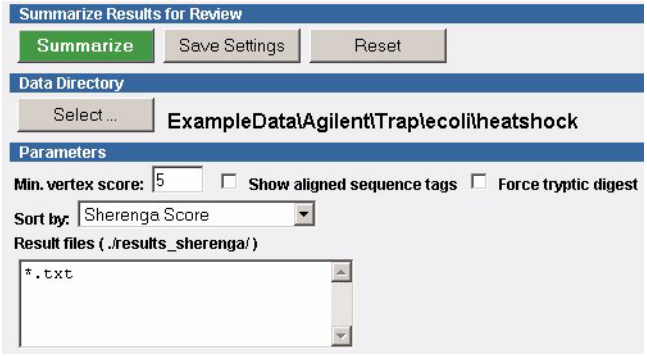


Figure 57 Sherenga *de novo* Summary form

5 If it is not already selected, select the directory where your data resides.

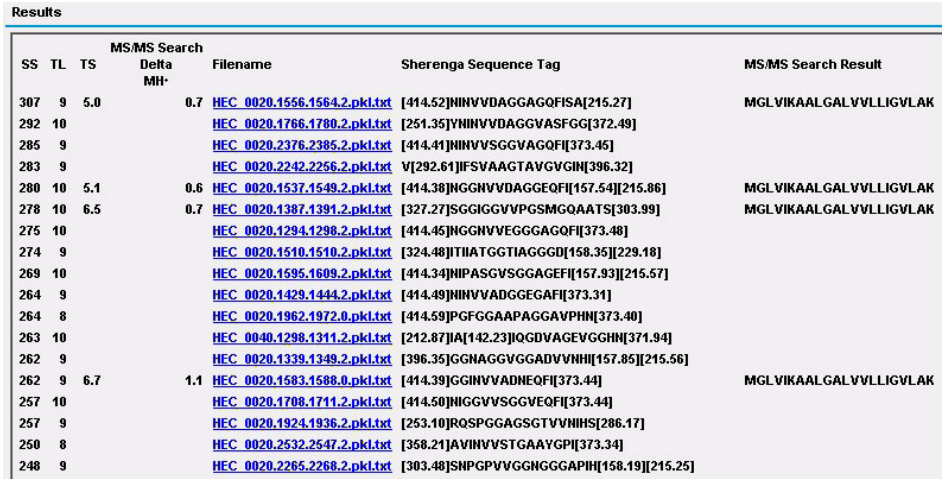
6 Set the **Min. vertex score**. If you do not want to include interpretations made on smaller peaks, raise this threshold. Amino acids interpreted based on smaller peaks are then replaced by their mass gaps. In the detailed results that follow (see bottom blue highlight in Figure 59 on page 117), **Orig. 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.

7 Set other parameters. (The defaults are good choices.)

8 If you wish to consider only tryptic peptides in the results, mark the check box to **Force tryptic digest**.

9 Click the **Summarize** button.

10 Check that you see the Sherenga top-level report, as shown in Figure 58. The components of the report are described below.



Results

SS	TL	TS	MS/MS Search 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]NINVVDAGGAGQFISA[215.27]	MGLVIKAALGALVVLLIGVLAK
292	10			HEC_0020.1766.1780.2.pkl.txt	[251.35]YNINVVDAGGVASFGG[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]IFSVAAGTAVGVGIN[396.32]	
280	10	5.1	0.6	HEC_0020.1537.1549.2.pkl.txt	[414.38]NGGNVVDAGGEQFI[157.54][215.86]	MGLVIKAALGALVVLLIGVLAK
278	10	6.5	0.7	HEC_0020.1387.1391.2.pkl.txt	[327.27]SGGIGGVVPGSMGQAATS[303.99]	MGLVIKAALGALVVLLIGVLAK
275	10			HEC_0020.1294.1298.2.pkl.txt	[414.45]NGGNVVEGGGAGQFI[373.48]	
274	9			HEC_0020.1510.1510.2.pkl.txt	[324.48]ITIIATGGTIAGGGD[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]NINVVADGGEGAFI[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]GGNAGGVGGADVVNHI[157.85][215.56]	
262	9	6.7	1.1	HEC_0020.1583.1588.0.pkl.txt	[414.39]GGINVVADNEQFI[373.44]	MGLVIKAALGALVVLLIGVLAK
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]SNPGPVVGGNGGGAPIH[158.19][215.25]	

Figure 58 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 34.

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 59, and the Sherenga results list, as shown at the bottom of Figure 59.

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 87 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.

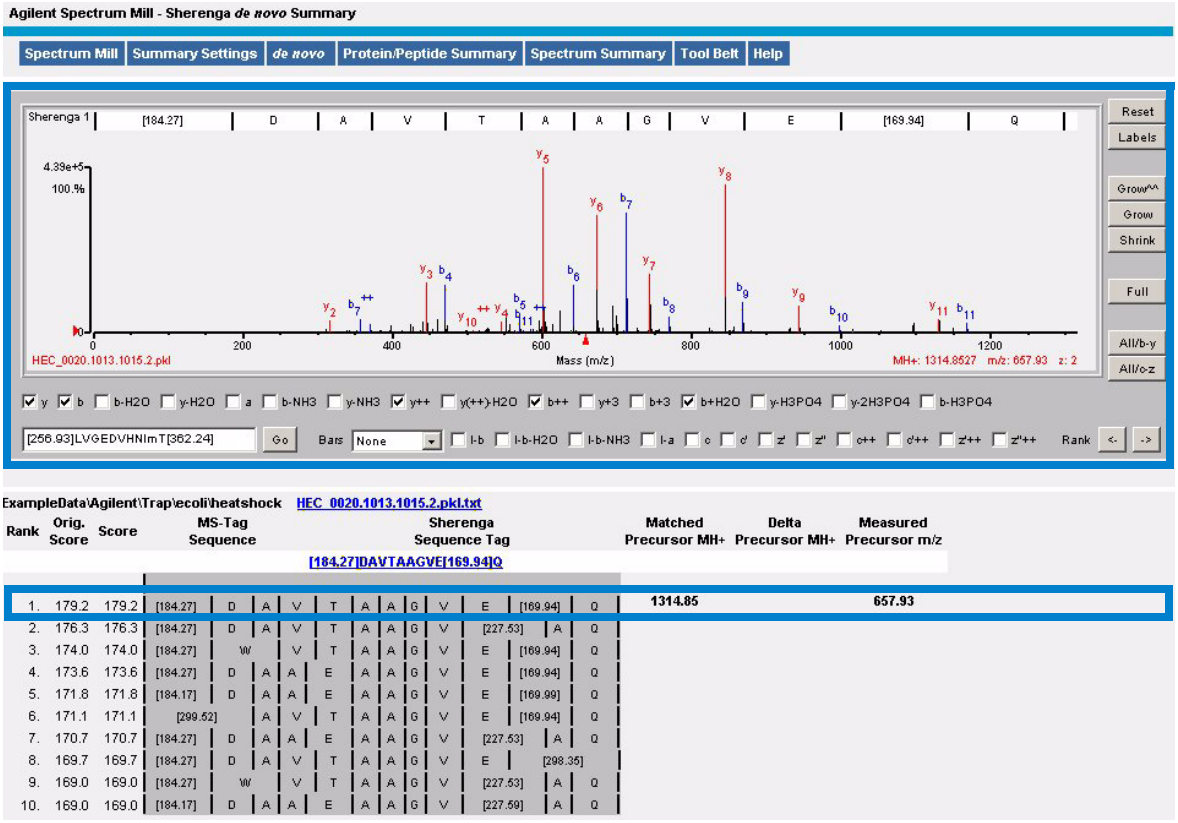


Figure 59 First Sherenga interpretation outlined in rectangles in both the Spectrum Viewer (top) and Sherenga results list (bottom)

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 59. It is also displayed above the Sherenga results list shown at the bottom of Figure 59.

4
Processing Ion Trap - ETD Data

This chapter describes use of the Spectrum Mill workbench to process electron transfer dissociation (ETD) data from Agilent ion traps. The data files can contain ETD spectra alone or in combination with collision-induced dissociation (CID) spectra.

ETD 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, “Basics for Processing MS/MS Data”. This chapter discusses the unique aspects of ETD data, and includes the following:

- Several sections that describe specific Spectrum Mill settings for ETD data
- Two sections that provide general information about ETD data
  - Overview of electron transfer dissociation
  - Interpreting ETD spectra

Please see the *Familiarization Guide* for a related exercise on Spectrum Mill processing of ETD data.

To set up Agilent ion trap ETD analyses for use with the Spectrum Mill workbench

When you set up data acquisition parameters on the Agilent ion trap, it is critical that you establish analysis settings that are compatible with Spectrum Mill processing. Please keep in mind the following:

1 Do not generate CID-ETD MS^3 data. That data type is not supported with the Spectrum Mill workbench.

2 If you wish to generate ETD-only data, in the Auto MS(n) tab within the ETD dialog box, be sure to *clear* the **CID-ETD** check box. See Figure 60.

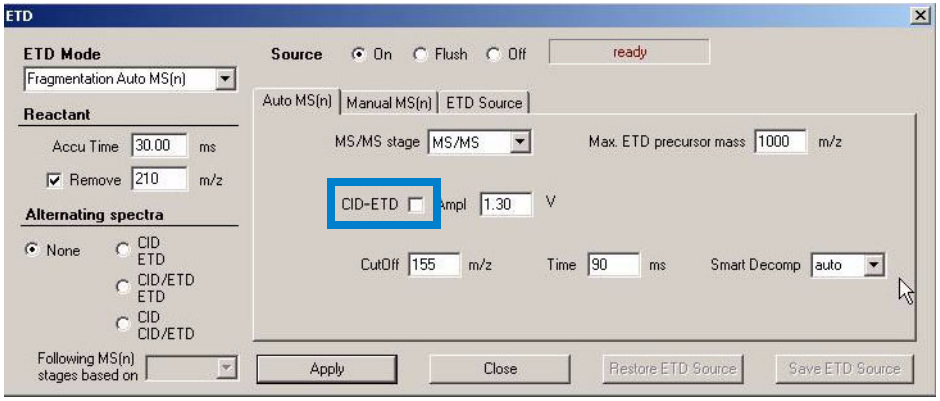


Figure 60 For ETD-only data, do *not* mark the check box that is highlighted.

To use the Data Extractor for ETD data

- Use the Data Extractor for ETD data exactly the same as you would for any other Agilent ion trap data. See Chapter 1, “Basics for Processing MS/MS Data,” starting on page 11.

The Data Extractor automatically determines whether the spectra are from CID or ETD experiments, and it merges spectra of like data types only. This means that a given data file can contain either or both types of spectra. The Data Extractor also notes in a file (**SpecFeatures.1.tsv**) the dissociation method (CID or ETD) for each extracted spectrum, so that MS/MS Search can apply the correct scoring.

To use MS/MS Search for ETD data

- Specify the correct instrument, either:
  - Agilent ESI ion trap ETD
  - Agilent ESI ion trap MIX-CID-ETD

  If you accidentally set the instrument to **Agilent ESI ion trap**, the software searches each spectrum using the correct parameters, but the file that saves the search settings (**mstag.#.params**) may indicate the wrong instrument type.
- If you have spectra with higher charge states, change the setting for **Maximum ambiguous precursor charge** to a larger number.
  - For enzymes such as trypsin that produce shorter peptides, it is best to try **3** first, then see if you get more valid hits by setting to **4** or **5**.
  - For enzymes such as LysC that produce longer peptides, it is best to try **4** first, then try **5** in a subsequent search.

Other than the specifics described in the bullets above, you run MS/MS Search the same as you would for other ion trap data. MS/MS Search automatically detects the extracted ETD and CID spectra and applies the appropriate search conditions and scoring. For example, if you enable proton mobility scoring, the software applies it to CID spectra but not to ETD spectra. The software searches ETD and CID spectra in separate batches.

To use Protein/Peptide Summary for ETD data

When you summarize ETD results in peptide mode:

- If you have mixed CID/ETD data, under **Review Fields**, mark the check box for **Fragmentation mode** so you can properly judge each spectral interpretation.
- In the Spectrum Viewer, click the appropriate button for each fragmentation type:
  - **All/c-z** for ETD
  - **All/b-y** for CID
- In the Spectrum Viewer, choose the appropriate setting for **Bars**:
  - **c** or **z'** for ETD
  - **b** or **y** for CID

When you summarize ETD results in one of the protein modes, do the same as for any other ion trap MS/MS data processing. See Chapter 1 and Chapter 2 of this manual.

To use Sherenga *de novo* Sequencing for ETD data

- Specify the correct instrument, either:
  - Agilent ESI ion trap ETD
  - Agilent ESI ion trap CID-ETD-mix

Other than changing the instrument setting, you run *de novo* sequencing the same as with other MS/MS data.

Overview of electron transfer dissociation

Electron transfer dissociation (ETD) is a method of inducing MS/MS fragmentation by reacting peptide cations with negatively charged ions (anions). In the case of the Agilent ion trap, the anions are fluoranthene anions that are introduced into the ion optics just prior to the second octopole. These radical anions react with pre-selected peptide precursor ions that are multiply charged. The anions transfer electrons to the precursor ions, which then rapidly undergo fragmentation.

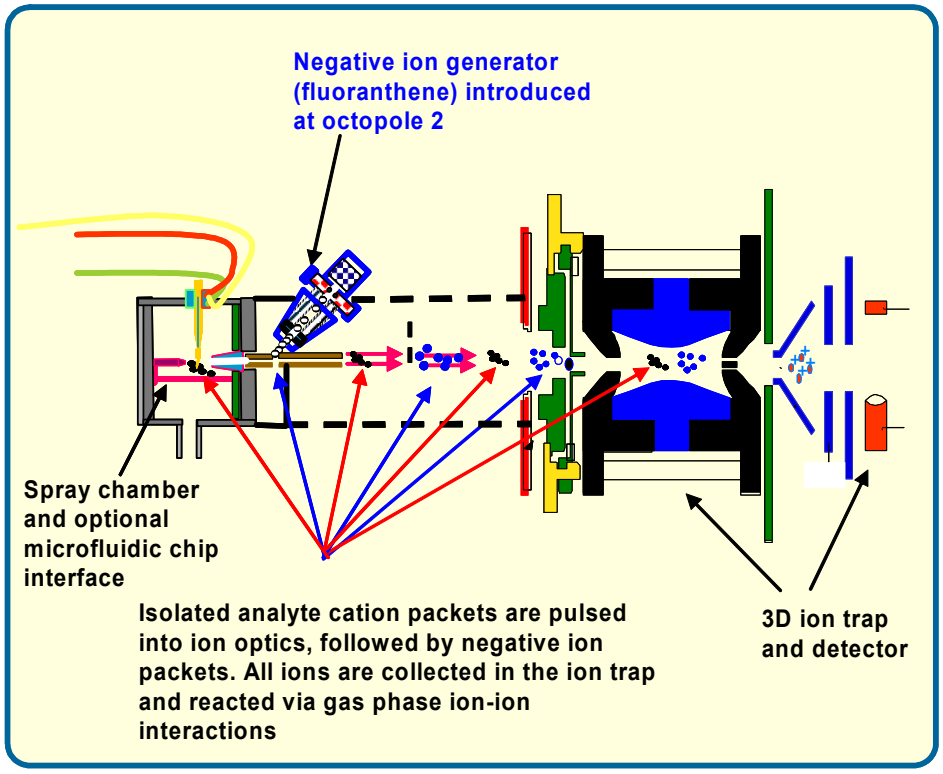


Figure 61 Overview of electron transfer dissociation on the Agilent ion trap

Relative to CID, the primary advantages of ETD are more complete fragmentation and enhanced capabilities to characterize post-translational modifications.

- ETD generally affords a fairly complete series of c- and z-ions. (See Figure 62.) Relative to tandem MS by CID, ETD fragmentation is less sensitive to the amino acid makeup of the peptide sequence, the peptide length, and the presence of post-translational modifications.
- ETD works better as the precursor charge state increases. CID generally produces better results at 2+, but ETD usually produces more sequence information for charge states of 3+ and higher.
- ETD spectra are complimentary to CID spectra, which typically exhibit b- and y-ions.
- With ETD spectra, database matching and *de novo* sequencing are enhanced due to more complete fragmentation.
- In the ETD process, phosphorylation and glycosylation sites remain intact, which enables confident localization of these post-translational modifications within the peptide amino acid sequence.

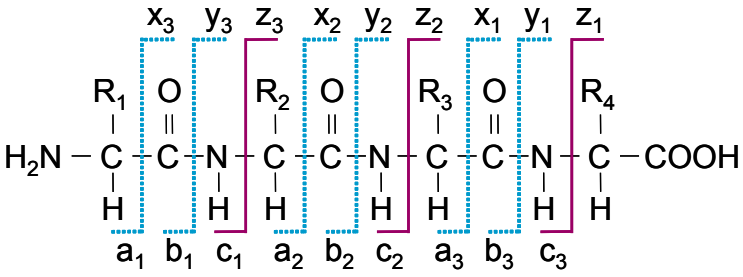


Figure 62 ETD generates primarily c- and z-ions

For further reading

Coon, J.J.; Syka, J.E.P.; Shabanowitz, J.; Hunt, D.F. “Tandem Mass Spectrometry for Peptide and Protein Sequence Analysis”, *BioTechniques* **2005**, vol. 38, no. 4, 519–523.

Syka, J.E.P.; Coon, J.J.; Schroeder, M.J.; Shabanowitz, J.; Hunt, D.F. “Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry”, *Proc. Natl. Acad. Sci.* **2004**, vol. 101, no. 26, 9528–9533.

Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. “Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process”, *J. Am. Chem. Soc.* **1998**, vol. 120, no. 13, 3265–3266.

Interpreting ETD spectra

To learn what to expect with ETD spectra, examine the spectra in Table 5.

Table 5 Comparison of ETD and CID spectra of peptides

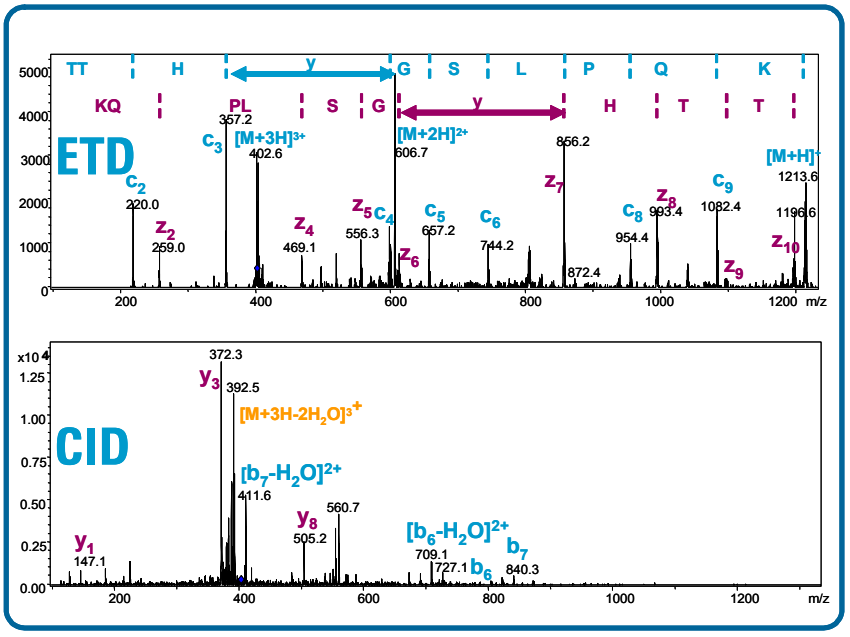
Comparison	Example
Comparison of ETD and CID for the triply-charged phosphopeptide TTHyGSLPQK (*m/z* = 404.7): • ETD shows c- and z-ions, while CID shows b- and y-ions. • ETD shows an essentially complete fragment series, while CID shows only a few of the fragments. • ETD fragments clearly show the phosphotyrosine residue, while CID shows no indication of the phosphate group.	

Table 5 Comparison of ETD and CID spectra of peptides

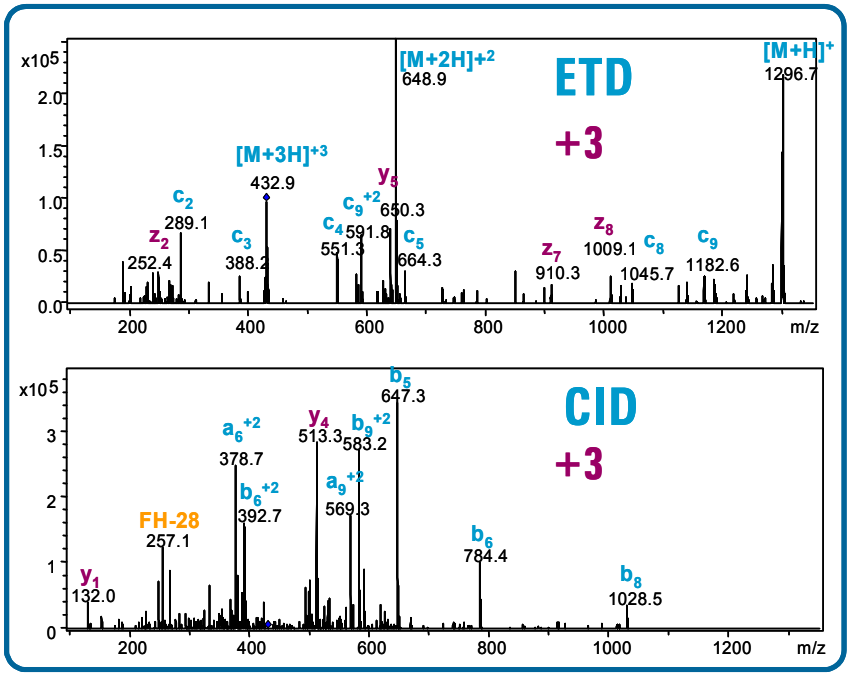
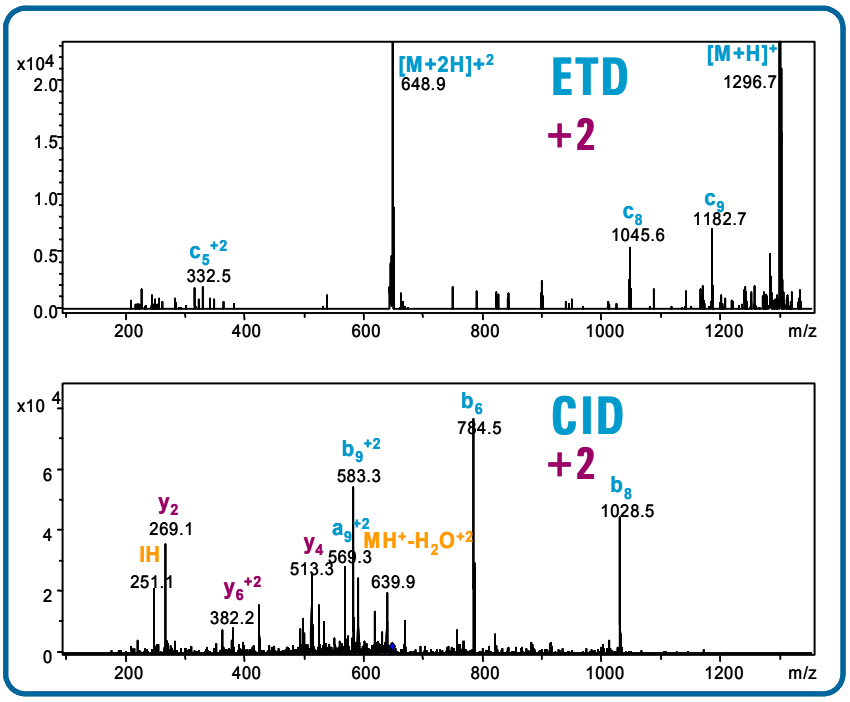
Comparison	Example
Comparison of ETD and CID for triply-charged angiotensin I (*m/z* = 432.9): • There are more fragments in the ETD spectrum than in the CID spectrum. • For precursor ions with higher charge states (+3 and up), ETD spectra are generally rich in fragmentation.	

Table 5 Comparison of ETD and CID spectra of peptides

Comparison	Example
Comparison of ETD and CID of doubly-charged angiotensin I (*m/z* = 648.8). • The doubly-charged precursor is more difficult to fragment using ETD. • For +2 charge states, ETD spectra generally show limited fragmentation.	
In ETD spectra, proline never yields c- and z-ions because their formation would involve cleavage of a N-C bond in the ring system.	Note the absence of the z_3 and c_7 ions from the spectrum on page 126.

5 Processing ICAT, iTRAQ, and Other Data for Differential Expression Quantitation

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 process data from iTRAQ studies, which use isobaric tags with low-mass MS/MS signature ions rather than isotopic labels.

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, “Basics for Processing MS/MS Data”. As you process the data, you set specific parameters in the Data Extractor, MS/MS Search, Protein/Peptide Summary, and Spectrum Summary forms to enable differential expression quantitation. This chapter describes those settings.

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

- ICAT D_0/D_8

- cICAT $^{12}C/^{13}C$
- C-terminal methyl ester D_0/D_3
- N-terminal propionyl D_0/D_5
- C-terminal $^{16}O/^{18}O$
- iTRAQ reagents
- Metabolic labels:
  - SILAC 2 (Arg 0-6Da, Lys 0-8Da)-mix
  - SILAC 3 (Arg 0-6-10Da)-mix
  - ^{15}N and $^{14}N/^{15}N$ 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 $^{14}N/^{15}N$ and iTRAQ 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 6 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 6 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	N-terminus and K	iTRAQ or iTRAQ Partial-mix (the latter if you suspect incomplete labeling)
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 and $^{14}N/^{15}N$ 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 light/heavy reagents are used, Data Extractor calculates additional EICs to ensure that both the heavy and light precursor ions are quantified, even if only one of these was actually 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. Taking ICAT as

an example, the reason for this 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, and it does not know 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

The iTRAQ intensity calculations do not require EICs. The abundances of *m/z* 114 through 118 are calculated from the MS/MS data.

You can calculate iTRAQ ratios for data that originates with the generic Data Extractor.

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

For **$^{14}N/^{15}N$ 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 performed 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 Data Extractor provides quantitative information for all ions for which MS/MS is performed, and the ratios are calculated in Protein/Peptide Summary based on this information.

You can calculate $^{14}N/^{15}N$ ratios with data from the generic Data Extractor.

CAUTION If you wish to perform 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 6 on page 131 for details.

2 Set other parameters as described in Chapter 1, except if you have iTRAQ data, do *not* mark the check box for **Proton mobility scoring**.

With the exception of iTRAQ, 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, only a single search is necessary. Since the four tags are isobaric, all four versions of the iTRAQ label are simultaneously fragmented during MS/MS. Further, all four iTRAQ labels produce the same MS/MS fragments for a given precursor peptide. Therefore, the iTRAQ labels do not have to be searched as a mix. However, the four tags produce different peaks in the 114 to 117 mass range 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

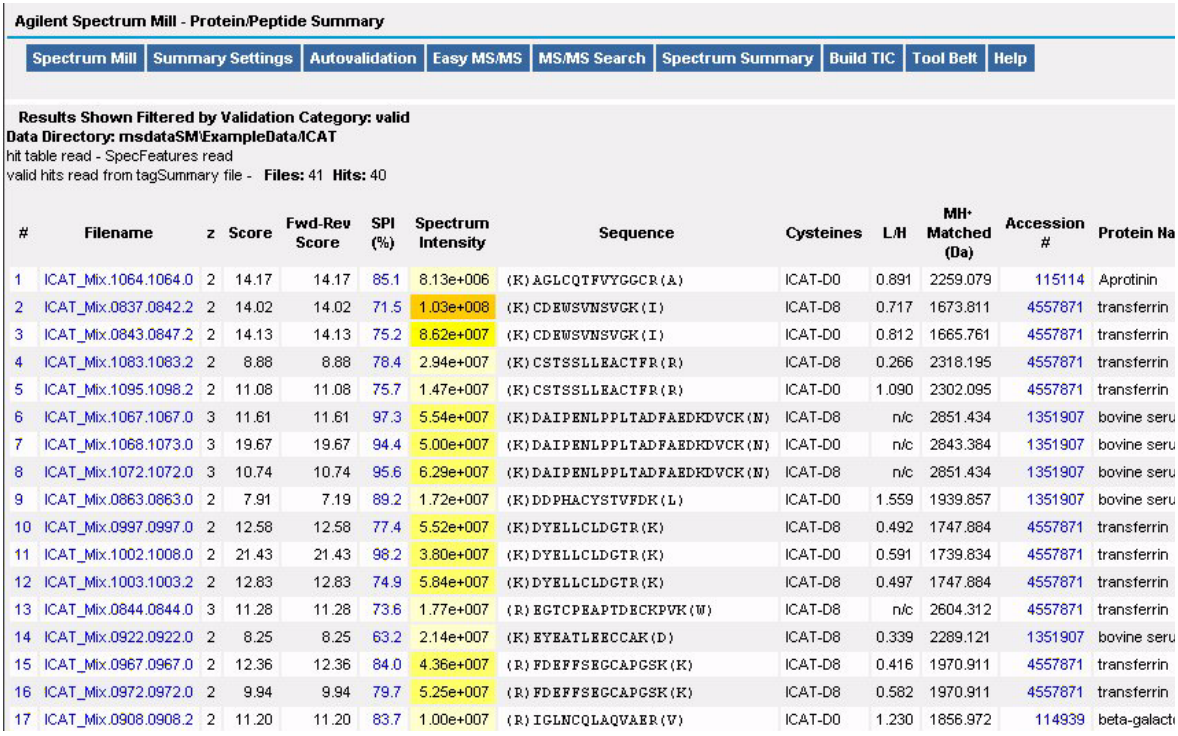
Use the Protein/Peptide Summary page to calculate DEQ ratios.

1. Mark the **DEQ ratios** (differential expression quantitation ratios) check box.
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 **Modifications** 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 the **Summarize** button.

To interpret DEQ results for isotopic labels in peptide mode

- Check that your results resemble those in Figure 63, 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 | 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

#	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)CDEWSVNSVGK(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)CDEWSVNSVGK(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)DAIPENLPPLTADFAEDKDVCK(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)DAIPENLPPLTADFAEDKDVCK(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)DAIPENLPPLTADFAEDKDVCK(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)DDPHACYSTVFDK(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)DYELLCLDGTR(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)DYELLCLDGTR(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)DYELLCLDGTR(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)EYEATLEECCAK(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)IGLNCQLAQVAER(V)	ICAT-D0	1.230	1856.972	114939	beta-galact

Figure 63 ICAT results shown in peptide mode

In the peptide mode, you may see a calculation discrepancy in cases where both members of a D_8/D_0 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 D_8- and D_0-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 64, 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 64 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 ratios using the Protein/Peptide Summary page

Use the Protein/Peptide Summary page to calculate iTRAQ ratios. In the 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 180.

   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 181.

2 Do the following on the Protein/Peptide Summary page:

   a Mark the **iTRAQ intensities** check box (for modes that show peptides).

   b Mark the **iTRAQ ratios control** check box and select the reporter mass you wish to use as the control (denominator) for the ratio calculations.

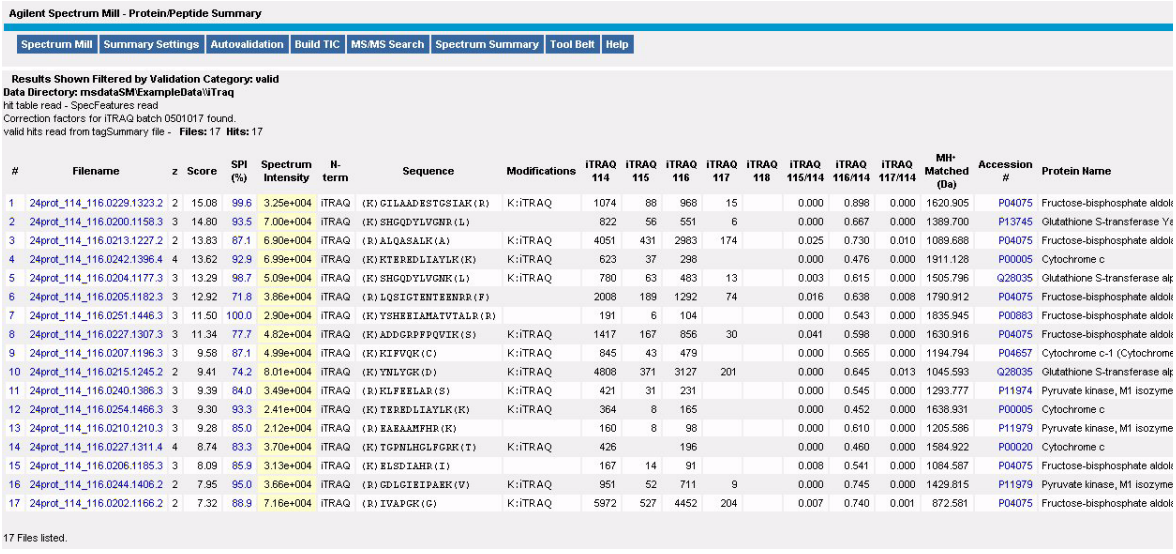
   c If you want to see the iTRAQ modification in a report that shows peptides, mark check boxes for both **N-terminus** and **Modifications**, since the reagents react at both the N-terminus and lysines.

   d Set other options.

   e Click the **Summarize** button.

To interpret results for iTRAQ labels in peptide mode

- Check that your results resemble those in Figure 65, which shows iTRAQ 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.
- The 118 intensity is included in the report so the software can apply the iTRAQ correction factors.



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: msdataSM\ExampleData\iTraq
hit table read - SpecFeatures read
Correction factors for iTRAQ batch 0501017 found.
valid hits read from tagSummary file - **Files:** 17 **Hits:** 17

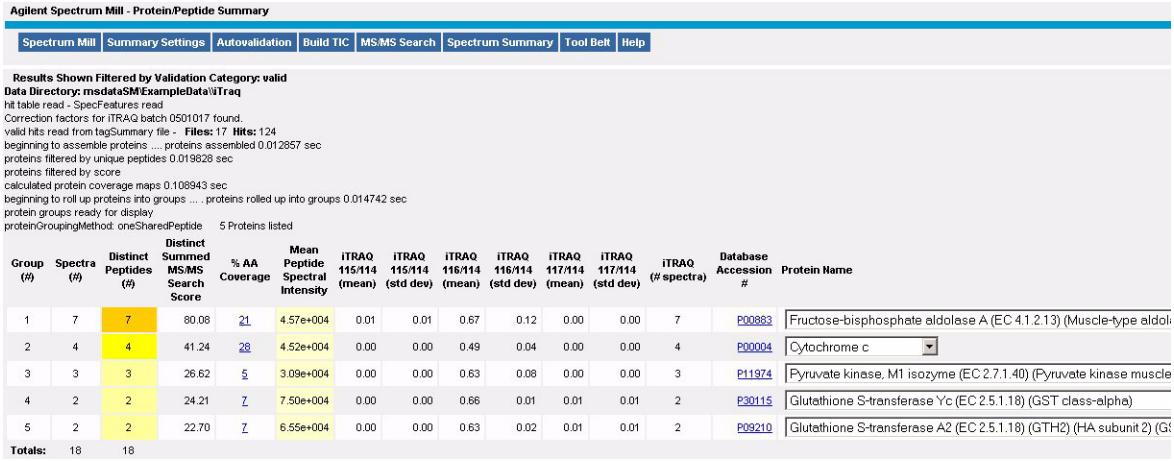
#	Filename	z	Score	SPI (%)	Spectrum Intensity	N-term	Sequence	Modifications	iTRAQ 114	iTRAQ 115	iTRAQ 116	iTRAQ 117	iTRAQ 118	iTRAQ 115/114	iTRAQ 116/114	iTRAQ 117/114	MH+ Matched (Da)	Accession #	Protein Name
1	24prot_114_116.0229.1323.2	2	15.08	99.6	3.25e+004	iTRAQ	(K)GILAADESTGSIAK(R)	K:iTRAQ	1074	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)SHGQDYLVGNR(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_116.0242.1396.4	4	13.62	92.9	6.99e+004	iTRAQ	(K)KTEREDLIAYLK(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)SHGQDYLVGNK(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)LQSIGTENTEENNR(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)YSHEEIAMATVTALR(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)ADDGRPFPQVIK(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)KIFVQK(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)TEREDLIAYLK(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)EAEAAMFHR(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)TGPNLHGLFGRK(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)GDLGIEIPAEK(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)IVAPGK(G)	K:iTRAQ	5972	527	4452	204		0.007	0.740	0.001	872.581	P04075	Fructose-bisphosphate aldol

17 Files listed.

Figure 65 iTRAQ results shown in peptide mode

To interpret results for iTRAQ labels in protein mode

- Check that your results resemble those in Figure 66, which shows iTRAQ 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

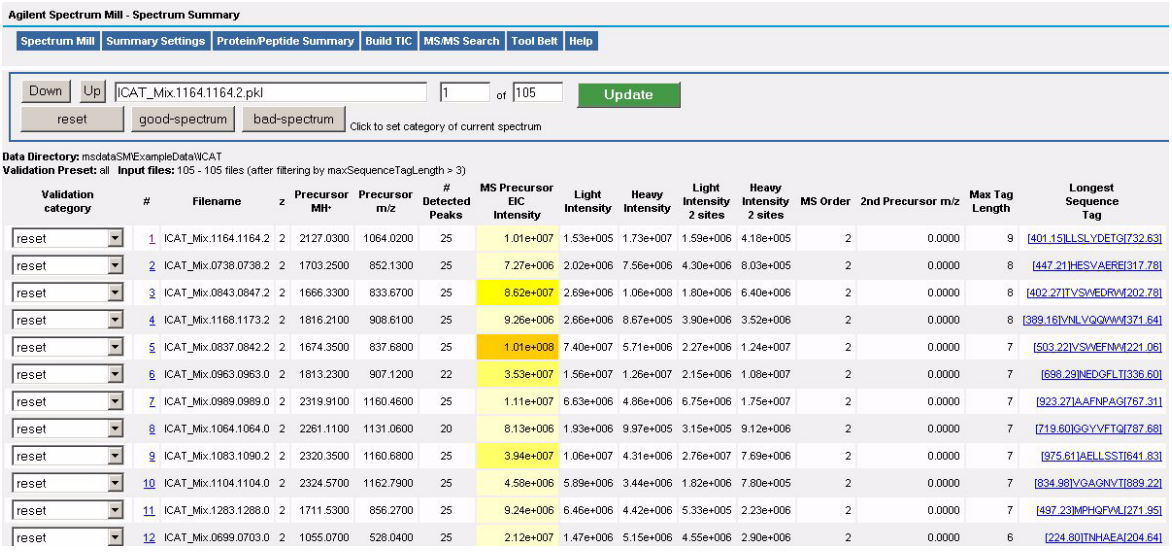
Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData\iTraq
hit table read - SpecFeatures read
Correction factors for iTRAQ batch 0501017 found.
valid hits read from tagSummary file - **Files:** 17 **Hits:** 124
beginning to assemble proteins ... proteins assembled 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 aldol
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 66 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 67.



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]LLSLYDETG[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]HESVAERE[317.78]
reset	3	ICAT_Mix.0843.0847.2	2	1666.3300	833.6700	25	8.62e+007	2.69e+006	1.06e+008	1.80e+006	6.40e+006	2	0.0000	8	[402.27]TVSWEDRW[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]VNLVQQWWM[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]VSWEFNWM[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]NEDGFLT[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]AAFNPAG[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]GGYVFTQ[787.68]
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	[975.61]AELLSSTI[641.83]
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]VGAGNVTI[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]MPHQFWL[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]TNHAEA[204.64]

Figure 67 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 which 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 $^{14}N/^{15}N$ mixes.

6
Basics for Processing MS-Only Data

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

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.

This chapter assumes that the Spectrum Mill workbench has already been installed on your server, databases have been downloaded and indexed, and the system is ready to go. If this has not yet been done, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

If you only want to process a single spectrum and want to paste in the peak list, see the description of Manual PMF Search on page 163.

Acquiring Agilent TOF data for use with the Spectrum Mill workbench

When you set up data acquisition parameters on the Agilent TOF, 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**.

First-pass data processing

Figure 68 shows a roadmap for processing MS-only data. How much of this roadmap you complete depends upon how important it is to extract every last bit of information from your sample.

For some samples, it is sufficient to process only through the basic workflow indicated by the darker (blue) shading. These steps are covered in this section, "First-pass data processing." For others, you take advantage of the optional processing indicated by the lighter (yellow) shading. These additional steps are covered in the next major section of this chapter, "Optional data processing to extract additional information" on page 162.

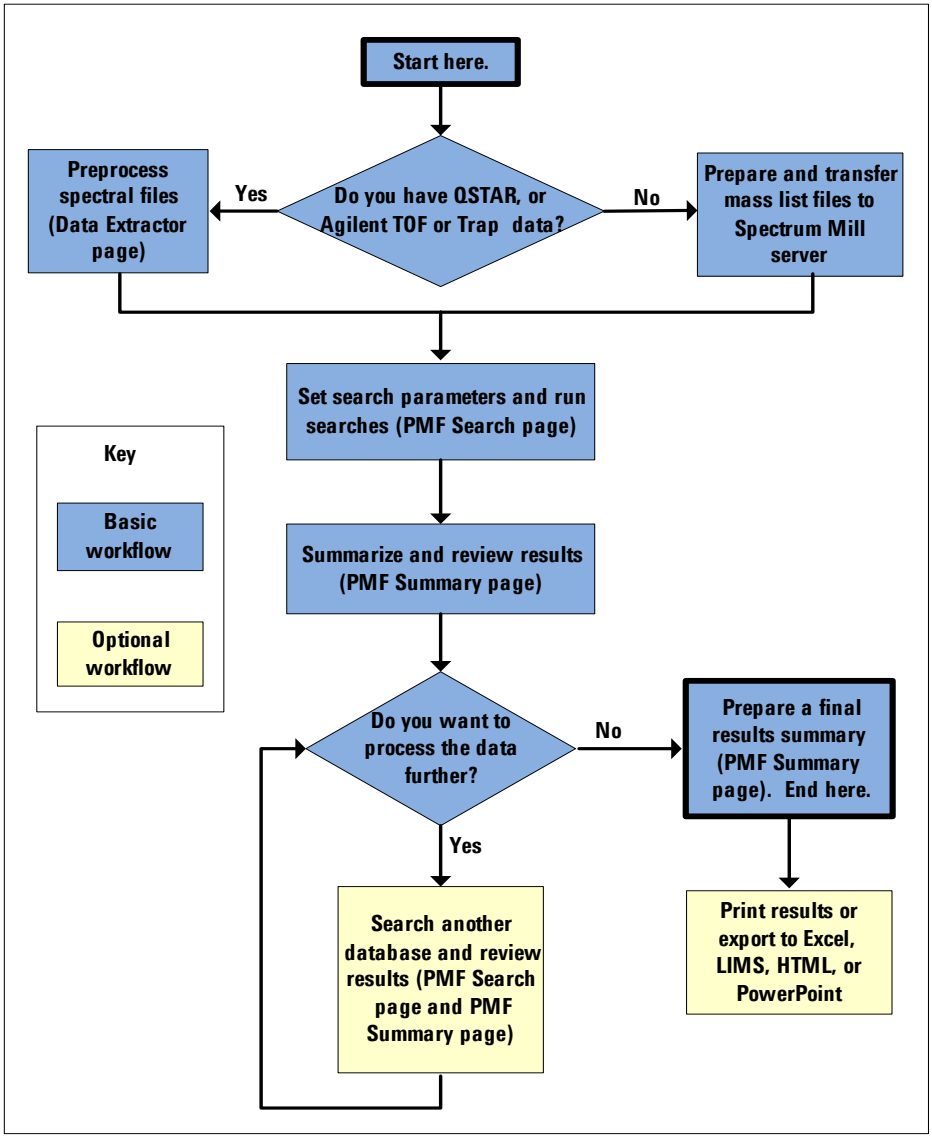


Figure 68 Roadmap to process MS-only data

Step 1. Transfer files to Spectrum Mill server

Agilent ion trap *.d, Agilent TOF *.wiff, and Applied Biosystems/MDS Sciex QSTAR o-MALDI *.wiff data files require no special preparation; you merely move or copy the files into the appropriate directory on the Spectrum Mill server, as described below. You then use the Spectrum Mill Data Extractor to extract the data.

Other types of data files must be preprocessed on the manufacturer's data system to generate generic peak list files. You then copy or move peak list files as described below.

NOTE For Agilent Q-TOF data in MS-only mode, you first generate a mass list with the Molecular Feature Extraction algorithm within the MassHunter Workstation qualitative analysis software, then copy the mass list from the compound table into the Spectrum Mill Manual PMF form, described on page 163. You do *not* follow the workflow described below for other Agilent instruments.

To transfer Agilent ion trap, Agilent TOF, or Applied Biosystems/MDS Sciex QSTAR o-MALDI data

1. In your Spectrum Mill file system on the server, navigate to the directory **msdataSM**. (If you don't know how to find your Spectrum Mill file system, ask the person who installed your software.)
2. In the directory **msdataSM**, create a new subdirectory for your sample files. Name it anything you want, but do not use spaces or parentheses in the name.

   If you have multiple samples you wish to compare, remember to set up a subdirectory for each sample. Ensure that each subdirectory contains all fractions for a single sample.

   You may add up to ten directory levels between **msdataSM** and your sample files, but shorter path lengths reduce memory usage and speed processing of large data sets.
3. Copy or move your *.d or *.wiff files to your new subdirectory or subdirectories.
4. After you transfer data files to the server, remove any spaces or parentheses in the data file names. Delete them or replace them with other characters, such as underscores.

5 Make sure you have both read and write permissions for each subdirectory that contains your sample files. (To check, right-click the folder and select **Properties**.) Make sure all user groups have full permissions. (Check the **Security** tab.)

6 Make sure that *.wiff files are *not* marked read-only.

To prepare and transfer other data types

1 Use your instrument data system to export calibrated, centroided, de-isotoped data as peak list files. Next you will transfer the peak lists to the Spectrum Mill server.

2 In your Spectrum Mill file system, navigate to the directory **msdataSM**. (If you don't know how to find your Spectrum Mill file system, ask the person who installed your software.)

3 In the directory **msdataSM**, create a new subdirectory for your sample files. Name it anything you want, but do not use spaces or parentheses in the name.

You may add up to ten directory levels between **msdataSM** and your sample files, but shorter path lengths reduce memory usage and speed processing of large data sets.

4 *Under* your new subdirectory, create another subdirectory called **fit_batch_in** and copy or move your files there. Note that you *must* use this subdirectory name!

5 After you transfer data files to the server, remove any spaces or parentheses in the data file names. Delete them or replace them with other characters, such as underscores.

6 Make sure you have both read and write permissions for each subdirectory that contains your sample files. (To check, right-click the folder and select **Properties**.) Make sure all user groups have full permissions. (Check the **Security** tab.)

Step 2. Run the Data Extractor

This step applies only for Agilent ion trap MS-only data, Agilent TOF data, 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 3. Run database searches" on page 155. If you want to extract Agilent ion trap MS/MS data, see "Step 3. Run the Data Extractor" on page 22 of Chapter 1, "Basics for Processing MS/MS Data".

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 performs 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 69.
2. Mark the check box for **Show only MS (PMF) parameters**.
3. In the Data Directory section, click the **Select...** button to select the folder that contains your files.
4. 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.

5 Make sure that the check box for **Agilent ESI TOF data** is cleared.

6 Check that your form resembles the one in Figure 69.

7 Click the **Extract** button.

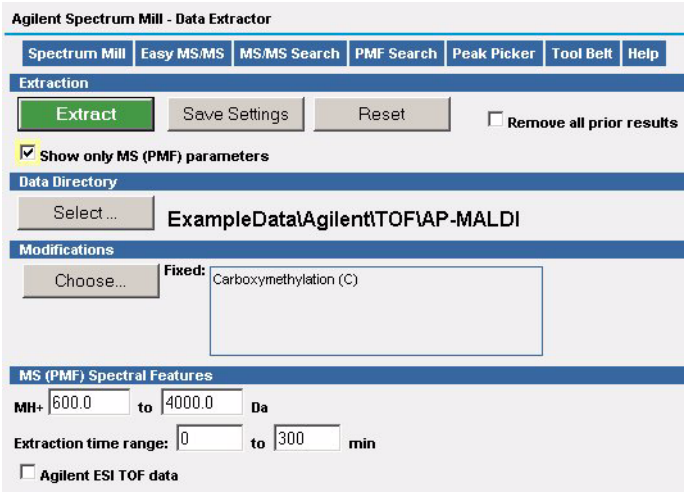


Figure 69 MS-only Data Extractor filled out for MALDI data

To run the MS-only Data Extractor for Agilent LC-ESI-TOF data

1 Navigate to the Data Extractor page, as shown in Figure 70 on page 152.

2 If this is the first time you have done this, read the background information the follows Figure 70.

3 Mark the check box for **Show only MS (PMF) parameters**.

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 Set the **Extraction time range** to the range you want to extract.

7 Mark the check box for **Agilent ESI TOF data**.

8 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.

9 Click the **LC** option.

10 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.

11 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.

12 Check that your form resembles the one in Figure 70.

13 Click the **Extract** button.

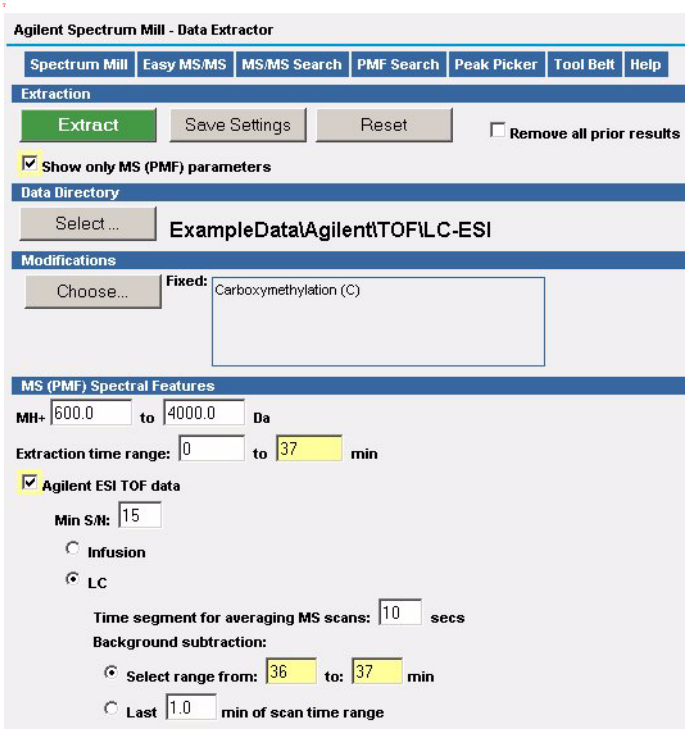


Figure 70 MS-only Data Extractor filled out for chromatographic data

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 71 on page 154.

2 If this is the first time you have done this, read the background information the follows Figure 71.

3 Mark the check box for **Show only MS (PMF) parameters**.

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 Set the **Extraction time range** to the range you want to extract.

7 Mark the check box for **Agilent ESI TOF data**.

8 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.

9 Click the **Infusion** option.

10 Check that your form resembles the one in Figure 71.

11 Click the **Extract** button.

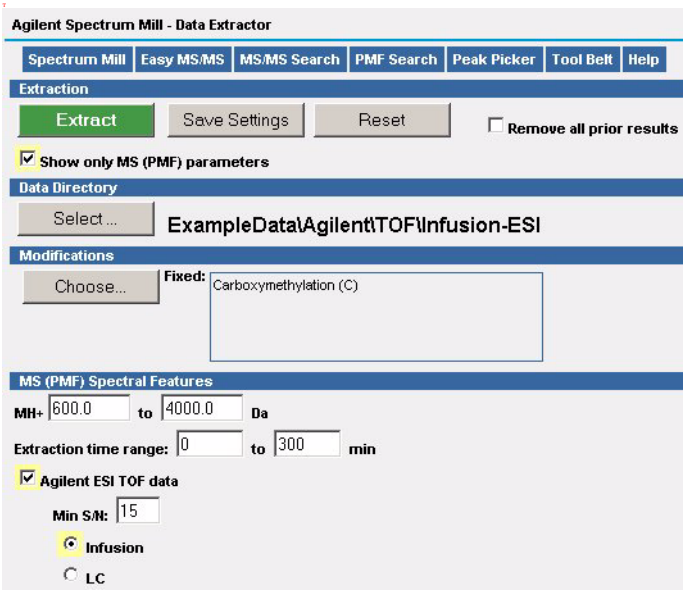


Figure 71 MS-only Data Extractor filled out for infusion-ESI-TOF data

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 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 3. Run database searches

1 Navigate to the PMF Search page, shown in Figure 72 on page 156.

2 Select the data directory where your files reside.

3 If you previously executed a search with incorrect settings, mark the check box for **Remove all prior PMF results**.

4 If the sample contains peptides from a mixture of proteins, mark the check box for **Mixture scoring**.

5 Set the **Database**.

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.

7 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**.

8 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.

9 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.

- 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.

10 Set **Report Details** to reflect the number of hits and detailed results you wish to view.

11 Set other parameters. Click the blue divider bars for more information in the online help.

12 Click the **Start Search** button.

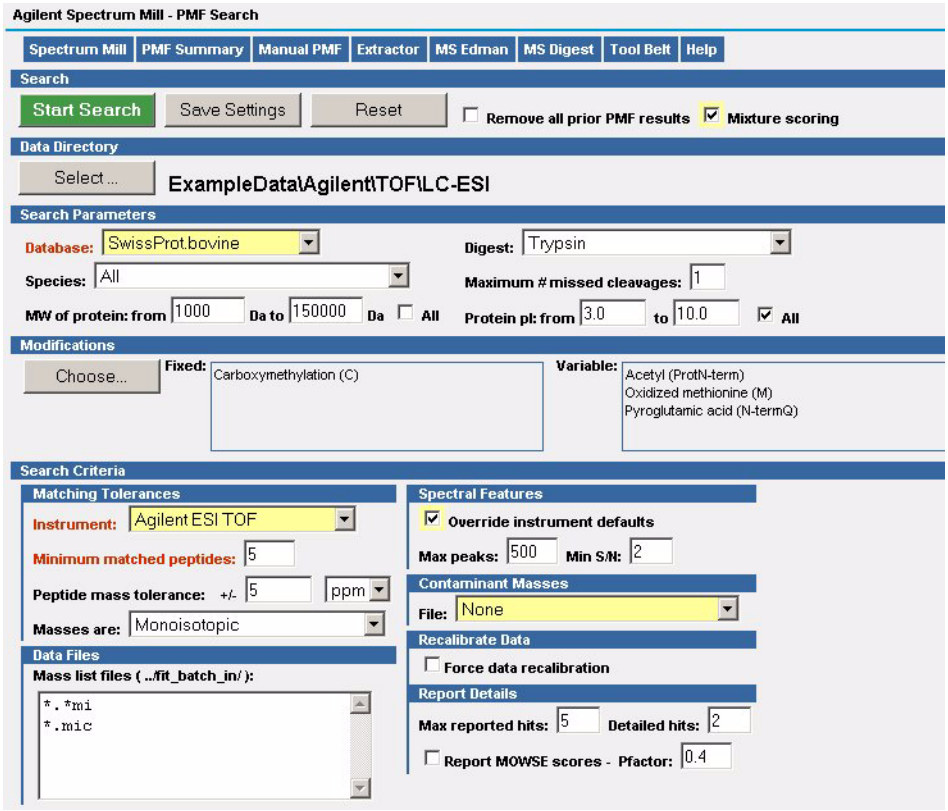


Figure 72 PMF Search page

Step 4. Summarize PMF Search results

1 Navigate to the PMF Summary page.

2 Select the data directory where your files reside.

3 Set parameters similar to those shown in Figure 73. In general, you do not need to change the defaults.

4 Click the **Summarize** button.

5 Check that you see a display similar to that shown in Figure 74.

NOTE

The setting for **Filter hits by score** is counterintuitive if you are used to MS/MS Search scores. With MS/MS Search scores, larger numbers represent better results. With PMF Search scores, smaller numbers represent better results. The probability scores represent the likelihood that the protein match occurs by chance. 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.

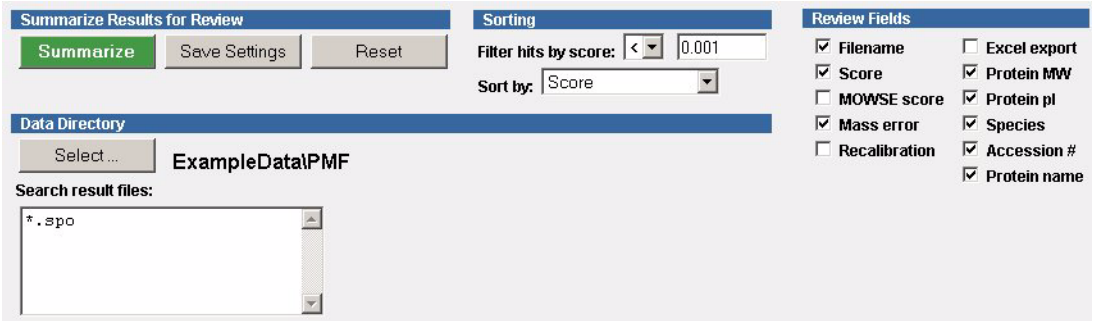
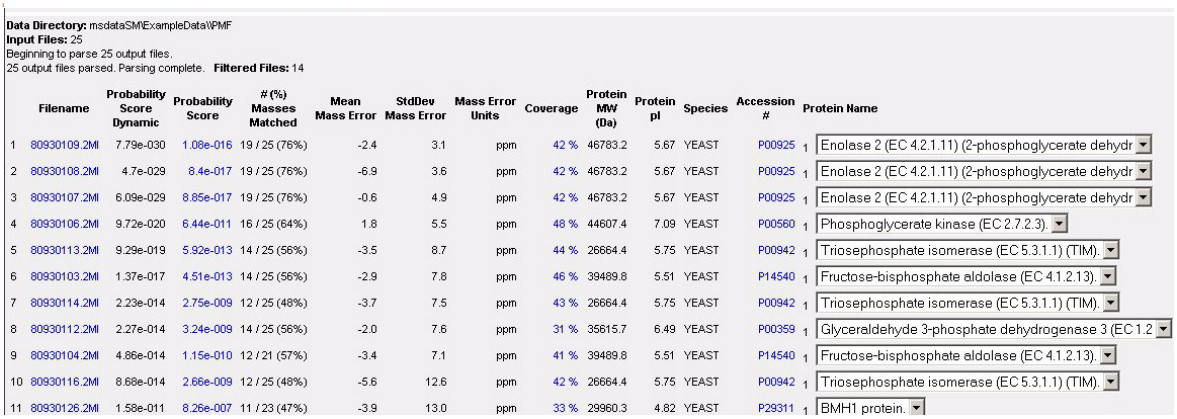


Figure 73 Settings to summarize PMF Search results



Data Directory: msdataSM\ExampleData\\PMF
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 74 PMF Search Summary table

Step 5. Manually review results

1 Manually review some or all of the results. First examine the overall results, as shown in Figure 74. 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 75, click a link under the **Filename** header and then scroll up. Scroll down or click links under the **Rank** heading shown in Figure 75 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 75, scored results would show a value of **1**, while non-scored results would show a value of **0**.

PMF Search Results

Sample ID (comment): **80930109.2MI**
Database searched: **SwissProt** Parameters used in Search
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.

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 75 PMF Search results

3 To continue examining data, click additional links shown in the PMF Search results in Figure 74 and Figure 75:

- To view information in the database you searched, click a link under **Accession #**.
- To view amino acid coverage, as shown in Figure 76, 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

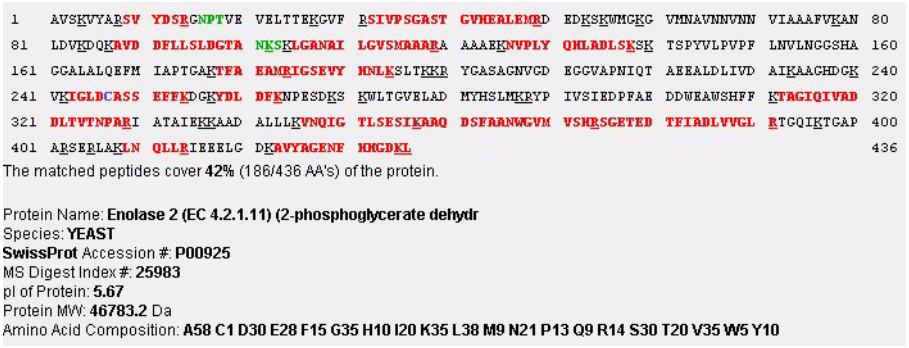


Figure 76 Amino acid coverage from PMF Search

Step 6. 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 4. Summarize PMF Search results" on page 157.

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 77. 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 77 Display when you mark the Excel export check box

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

a If you have not already done so, enable **Print background colors and images** in Internet Explorer. Select **Tools > Internet Options...**. Click the **Advanced** tab and mark the check box for **Print background colors and images**.

b Do one of the following:

- In Internet Explorer (IE) 6.0, select **File > Page Setup...** and set the page to landscape mode.
- In IE 7.0, click the arrow to the right of the printer icon, select **Page Setup...** and set the page to landscape mode.

c Click in the frame you wish to print.

d Do one of the following:

- In IE 6.0, select **File > Print Preview...**.
- In IE 7.0, click the arrow to the right of the printer icon and select **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, see the next section.

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

Optional data processing to extract additional information

The optional data processing steps are shown in Figure 68 on page 146, and are described in detail below.

Step 1. Search another database

1. Navigate to the PMF Search page.
2. Fill in the search form as described in "Step 3. Run database searches" on page 155, except specify another database.
3. Run the search.
4. Summarize and review results in the PMF Summary page, as described on page 157 through page 160.

Step 2. Create a final results summary

- See "Step 6. Summarize and print results" on page 160.

Single-spectrum processing

To search the spectrum

You use the Manual PMF page to process a single spectrum that you paste into the program.

1 Navigate from the PMF Search page to the Manual PMF Search page, shown in Figure 78.

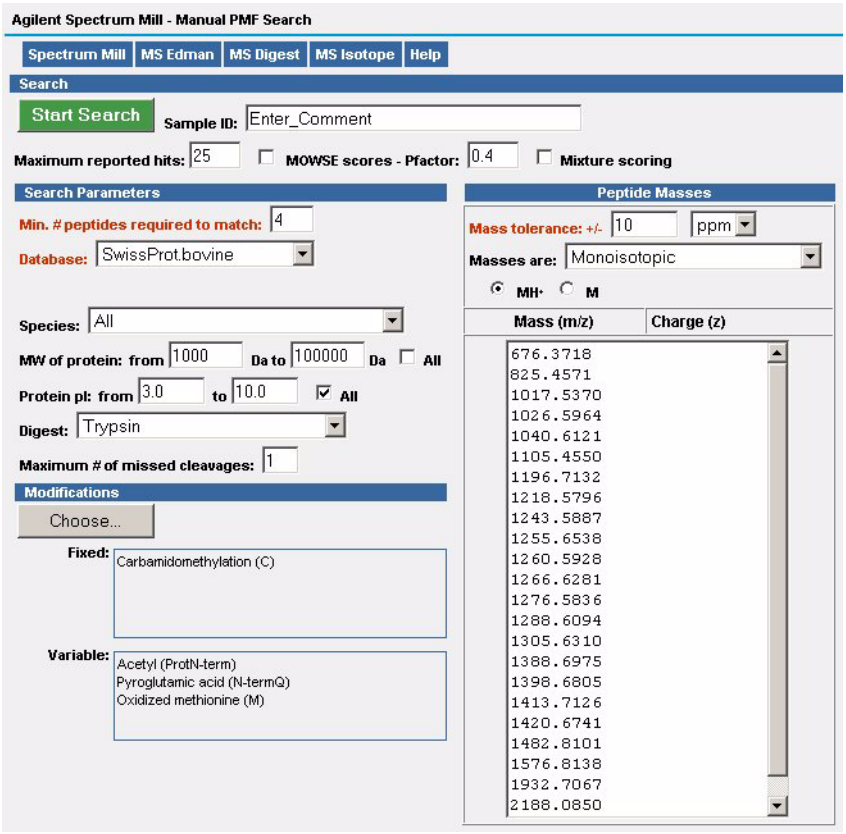


Figure 78 Manual PMF Search form

2 Paste your mass list into the **Peptide Masses** box.

3 If your masses represent protonated species, click **MH$^+$**. If your masses represent neutral species, click **M**.

4 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.

5 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.

6 Set the other parameters. Click the blue divider bars for more information in the online help.

7 Click the **Start Search** button.

8 When the search is complete, scroll down and check that you see PMF Search Results, as shown in Figure 79 on page 165.

PMF Search Results

Sample ID (comment): **Enter_Comment**
Database searched: **SwissProt** Parameters used in Search
Molecular weight search (**1000** - **100000 Da**) selects **253250** entries.
Full pI range: **267354** entries.
Combined molecular weight and pI searches select **253250** entries.
PMF search selects **2** entries.

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	2.92e-026	2.92e-026	16/23 (69%)	-3.6 (4.4)	55%	30276.5/5.71	BOVIN	P15497	Apolipoprotein A-I precursor (Apo-AI) (ApoA-I) - Bos taurus (Bovine)
2	0.0153	0.0153	4/23 (17%)	-0.5 (2.7)	11%	19316.4/5.26	RICFE	Q4UM09	NADH-quinone oxidoreductase subunit E (EC 1.6.99.5) (NADH dehydroge

Detailed Results

1. 16/23 matches (69%). BOVIN. Apolipoprotein A-I precursor (Apo-AI) (ApoA-I) - Bos taurus (Bovine) (30276.5 Da) pI = 5.71

m/z submitted	MH+ matched	Delta ppm	Score Counted	Tolerance Bin Index (0-1)	start	end	Peptide Sequence	Modifications
825.4571	825.4577	-0.8	1	**0**	177	183	**(R)AHVETLR(Q)**	
1017.5370	1017.5364	0.6	1	**1**	142	150	**(K)VAPLGEEFR(E)**	
1026.5964	1026.5942	2.1	1	**1**	164	172	**(K)LSPLAQELR(D)**	
1040.6121	1040.6099	2.1	1	**1**	229	237	**(K)AKPVLEDLR(Q)**	
1196.7132	1196.7249	-9.8	1	**0**	238	248	**(R)QGLLPVLESLK(V)**	
1218.5796	1218.5749	3.8	1	**1**	206	217	**(K)EGGGSLAEYHAK(A)**	
1255.6538	1255.6569	-2.5	1	**0**	36	46	**(K)DFATVYVEAIK(D)**	
1260.5928	1260.6008	-6.3	1	**0**	131	139	**(K)WHEEVEIYR(Q)**	
1266.6281	1266.6365	-6.6	1	**0**	120	129	**(K)VQPYLDEFQK(K)**	
1288.6094	1288.6168	-5.7	1	**0**	184	194	**(R)QQLAPYSDDLR(Q)**	**pryoGlu**
1305.6310	1305.6434	-9.5	1	**0**	184	194	**(R)QQLAPYSDDLR(Q)**	
1388.6975	1388.6957	1.3	1	**1**	130	139	**(K)KWHEEVEIYR(Q)**	
1398.6805	1398.6900	-6.8	1	**0**	51	63	**(R)DYVAQFEASALGK(Q)**	
1482.8101	1482.8203	-6.9	1	**0**	34	46	**(R)VKDFATVYVEAIK(D)**	
1576.8138	1576.8217	-5.0	1	**0**	69	82	**(K)LLDNWDTLASTLSK(V)**	
2188.0850	2188.1033	-8.4	1	**0**	83	100	**(K)VREQLGPVTQEFWDNLEK(E)**	
	Mean:	-3.6						
	Std dev:	4.4						

Figure 79 Manual PMF Search results

To examine the results

- To see more search details, scroll down or click links under the **Rank** heading shown in Figure 79. 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

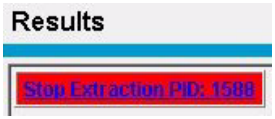
This chapter describes helpful tools that you can use when you process data with the Spectrum Mill workbench. These tools help you accomplish a range of tasks, from terminating processes to summarizing identification statistics. You access these tools from the Tool Belt page.

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 80. (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.

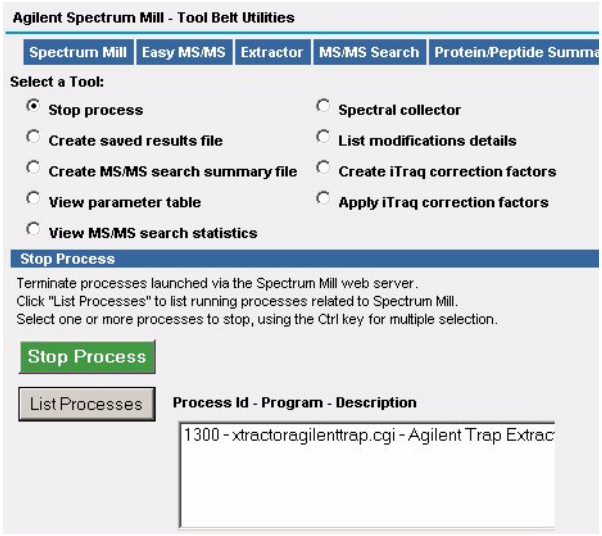


Figure 80 Stop process

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**.

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.

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 81.
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 You can instead create a saved results file from the MS/MS Search or Easy MS/MS Search page. When you mark the check box to **Search previous hits**, if you have not created a saved results file, then the search page automatically creates one for you.

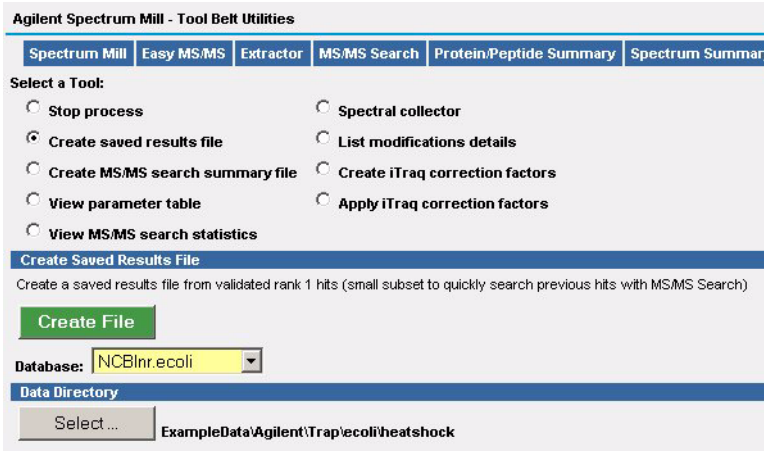


Figure 81 Create saved results file

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 82.

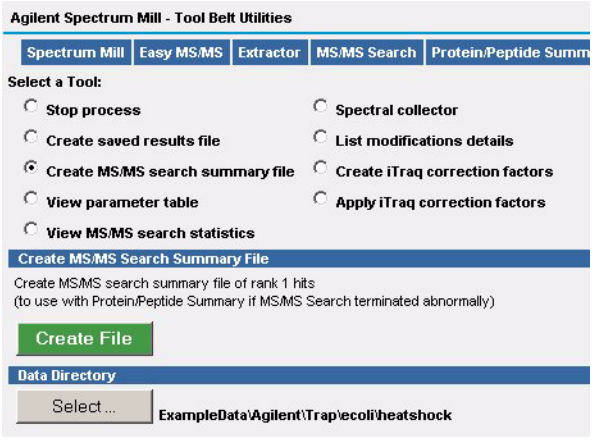


Figure 82 Create MS/MS search summary file

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. To view previously-used parameters:

1 Navigate to the Tool Belt page and click the **View parameter table** option, as shown in Figure 83.

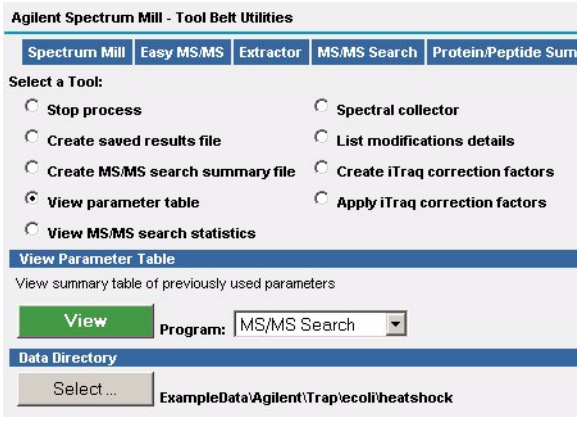


Figure 83 View parameter table

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 84. 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	NCBInr.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	NCBInr.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	-
msparams_dir	msparams_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\msdataSM\ExampleData\Agilent\Trap\ecoli\heatshock\results_mstag	-
results_from_file	-	1
results_to_file	-	-
search_type	identity	variable

Figure 84 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. To use this capability:

1 Navigate to the Tool Belt page and click the **View MS/MS search statistics** option, as shown in Figure 85.

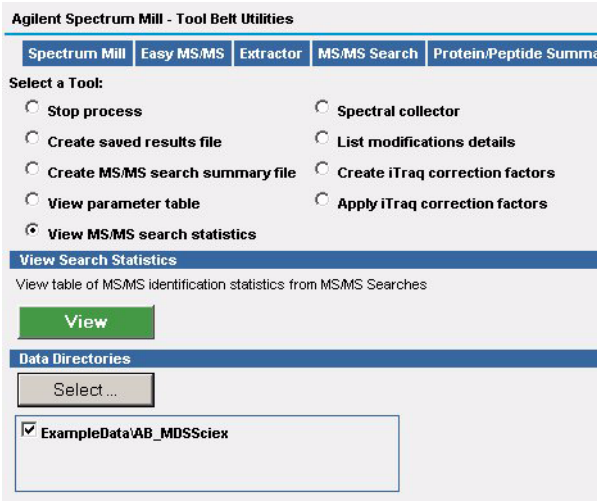
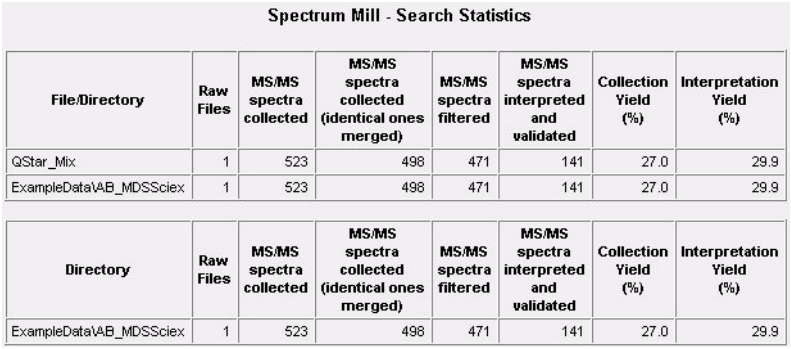


Figure 85 View search statistics

2 Select the **Data Directories** for which you want to summarize statistics.

3 Click the **View** button.

4 Check that you see a report like that shown in Figure 86. See the discussion below for explanations of some of the headers.



Spectrum Mill - Search Statistics

File/Directory	Raw Files	MS/MS spectra collected	MS/MS spectra collected (identical ones merged)	MS/MS spectra filtered	MS/MS spectra interpreted and validated	Collection Yield (%)	Interpretation Yield (%)
QStar_Mix	1	523	498	471	141	27.0	29.9
ExampleData\AB_MDSSciex	1	523	498	471	141	27.0	29.9

Directory	Raw Files	MS/MS spectra collected	MS/MS spectra collected (identical ones merged)	MS/MS spectra filtered	MS/MS spectra interpreted and validated	Collection Yield (%)	Interpretation Yield (%)
ExampleData\AB_MDSSciex	1	523	498	471	141	27.0	29.9

Figure 86 View search statistics report

MS/MS spectra collected Number of MS/MS spectra in the raw data file

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 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 87.

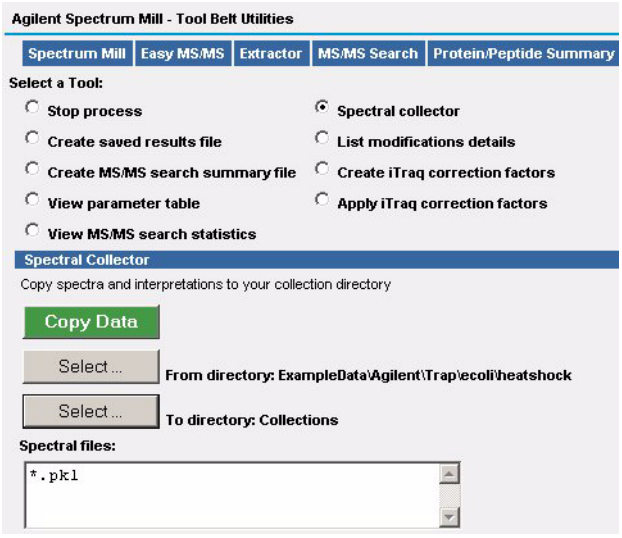


Figure 87 Spectral collector

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 88.

2 Select an **smconfig** file:

- **smconfig.xml** - includes all the modifications that are currently displayed in the Spectrum Mill forms. 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.

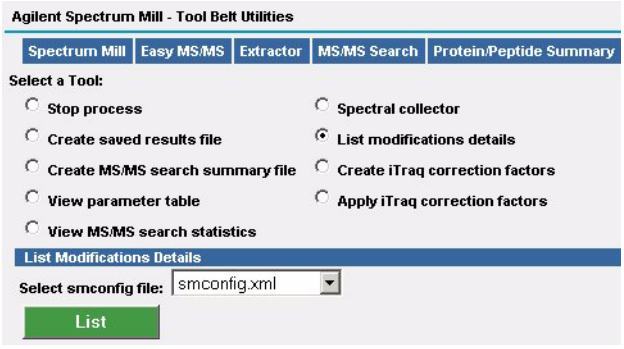


Figure 88 List modifications details

The listed details include:

- **Modification** - the name you see in the Spectrum Mill forms
- **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. 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 form. See Figure 89 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 181.

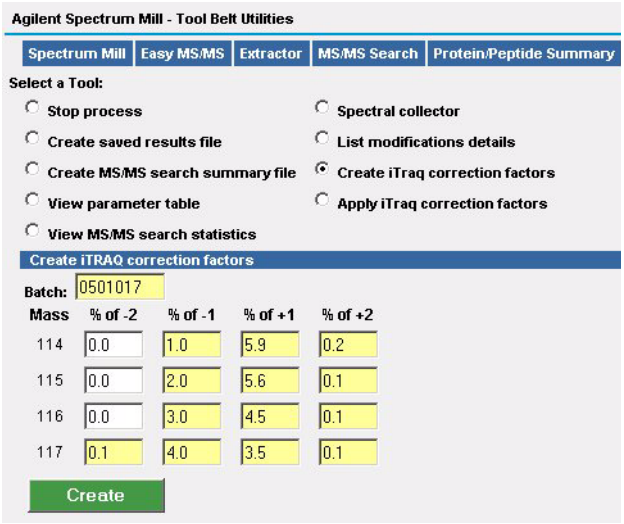


Figure 89 Create iTRAQ correction factors, shown with example correction factors

To apply a file of iTRAQ correction factors

After you have stored iTRAQ correction factors in a file (see page 180), 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.

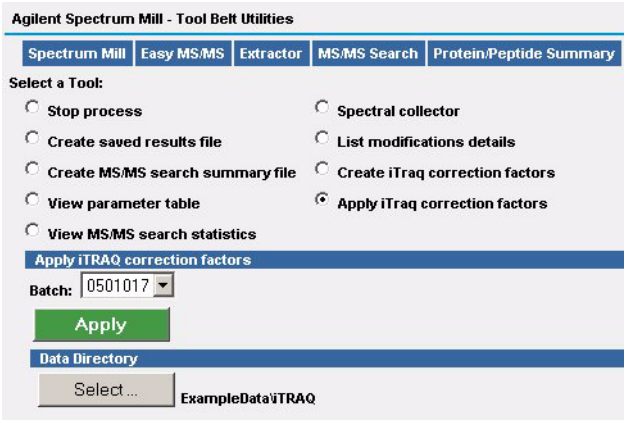


Figure 90 Apply iTRAQ correction factors

When the software performs 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.

8
Using Spectrum Mill Utilities

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 performs 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 187. Both Peptide Selector and MS Digest perform 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 91.
3. Set **Digest Parameters** and **Product Ion Parameters**. See the explanations in the online help.
4. 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.
5. Click the **Choose...** button to select the modifications that match the chemistry for your samples.
6. 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.

7 Set **Search Mode** parameters as described in the online help.

8 Click the **Select** button.

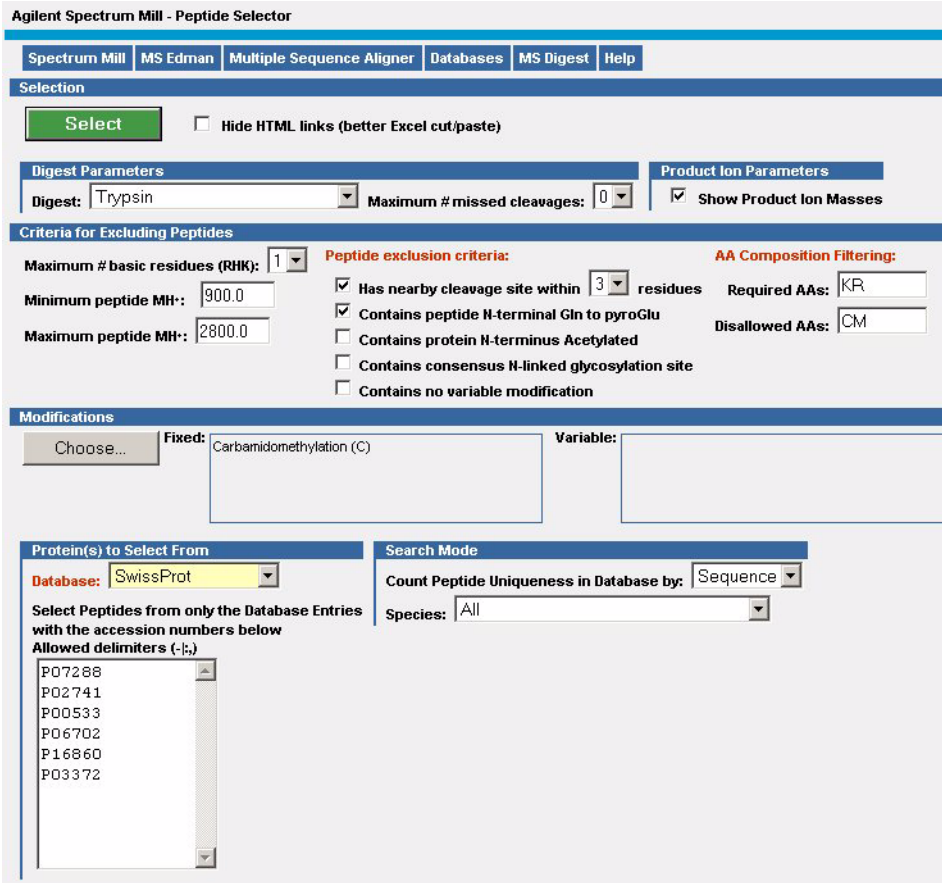


Figure 91 Peptide Selector

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 tactility 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 92.

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.

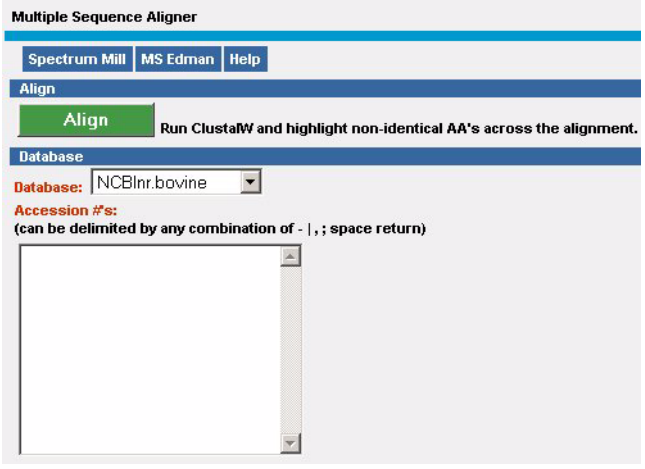


Figure 92 Multiple Sequence Aligner

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

MS Digest performs 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 93.
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 the **Digest** button.

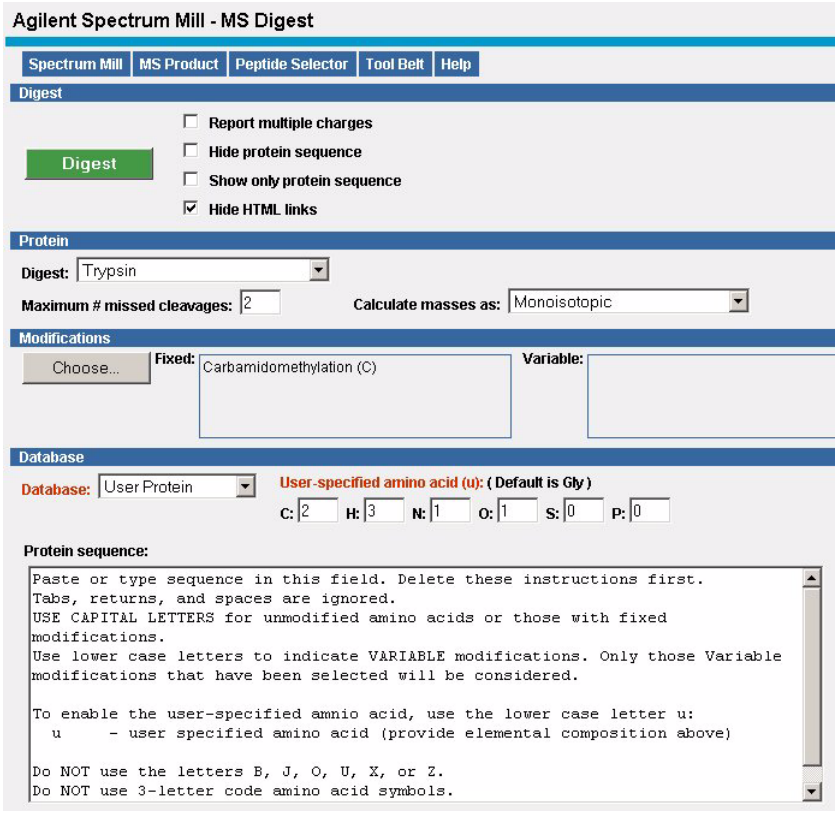


Figure 93 MS Digest

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 form, 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 203.

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 94.
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 94, 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.

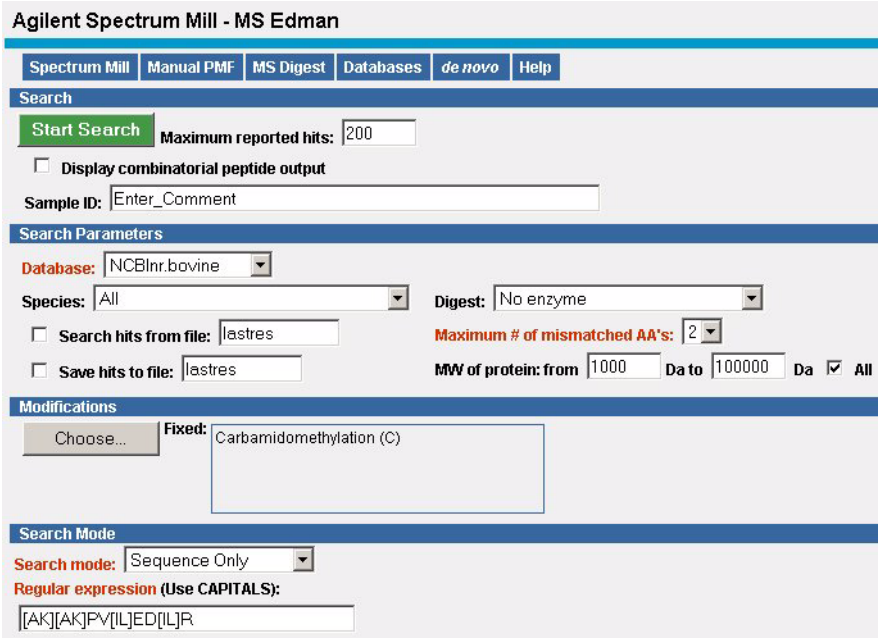


Figure 94 MS Edman

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 95.

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 form.

5 Click the **Fragment** button.

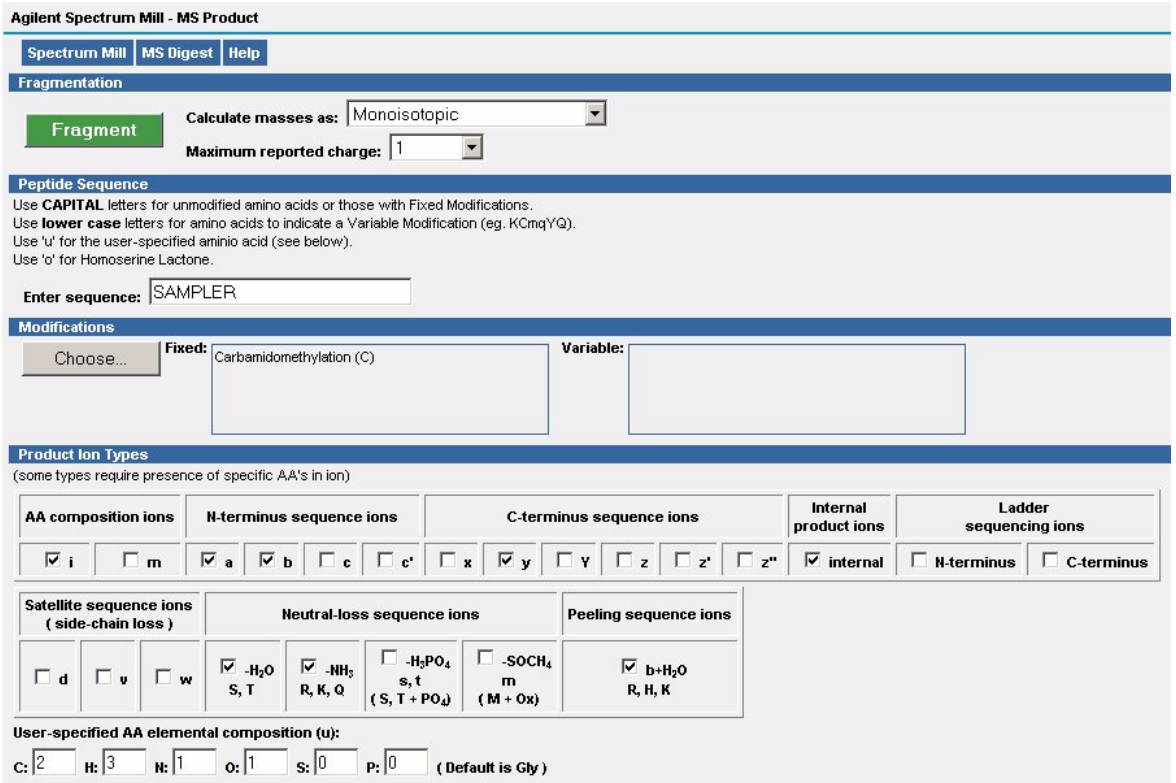


Figure 95 MS Product

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 96.
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.

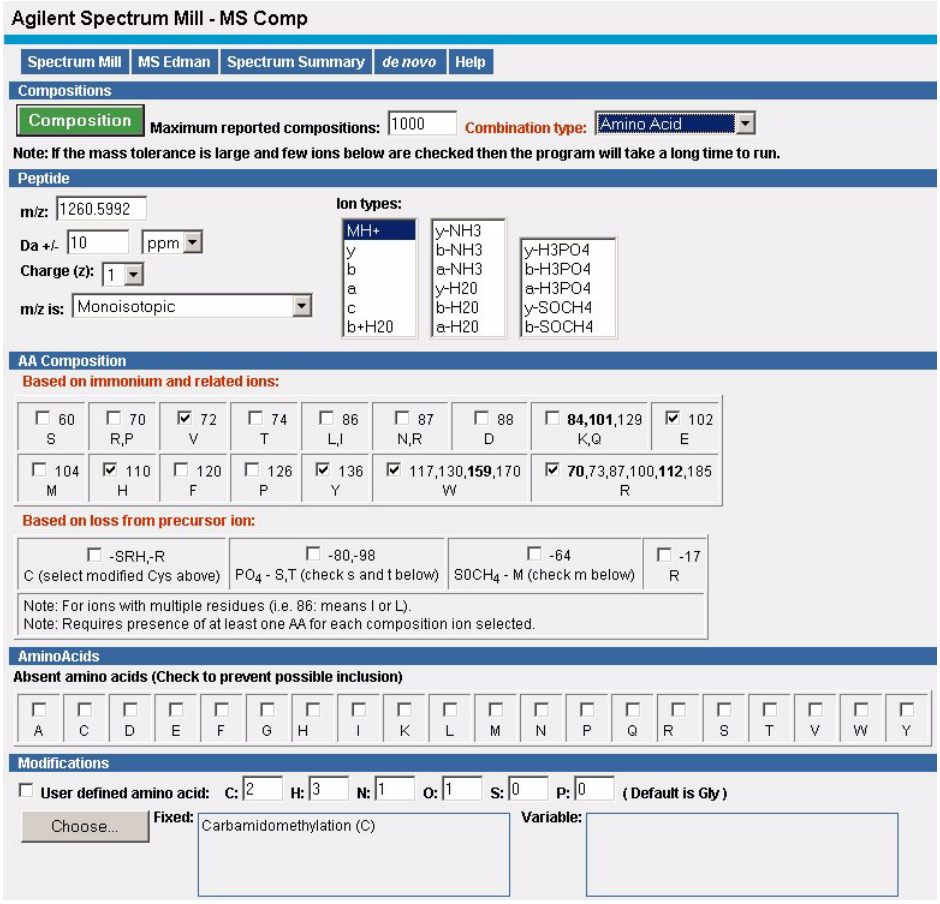


Figure 96 MS Comp

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 97.

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 the **Calculate** button.

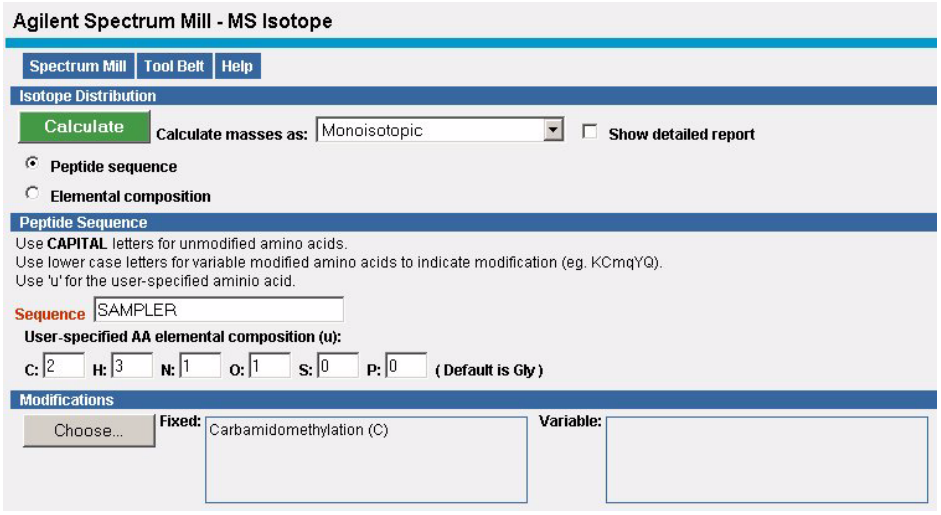


Figure 97 MS Isotope

9
System Administration

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 208.

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 form shown in Figure 98.

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 99.

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.

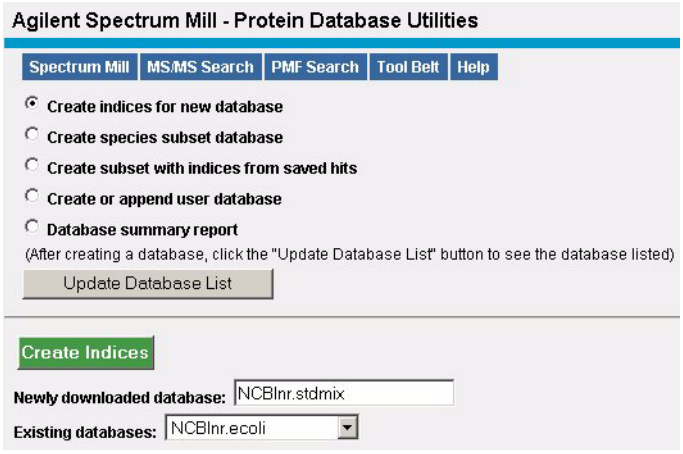


Figure 98 Create indices for new database

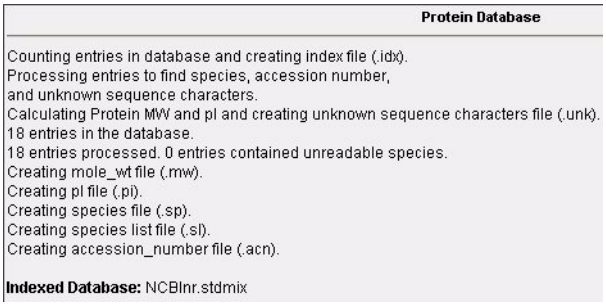


Figure 99 Display while database indices are created

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 202.
- Searches of results saved from other search programs. See "To create a saved results file so you can search previous hits" on page 170 and "To create a subset database from saved hits" on page 203.
- Creation of custom databases. See "To create a user (proprietary) database" on page 204.

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 form shown in Figure 100.
4. Fill in the form. 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**.

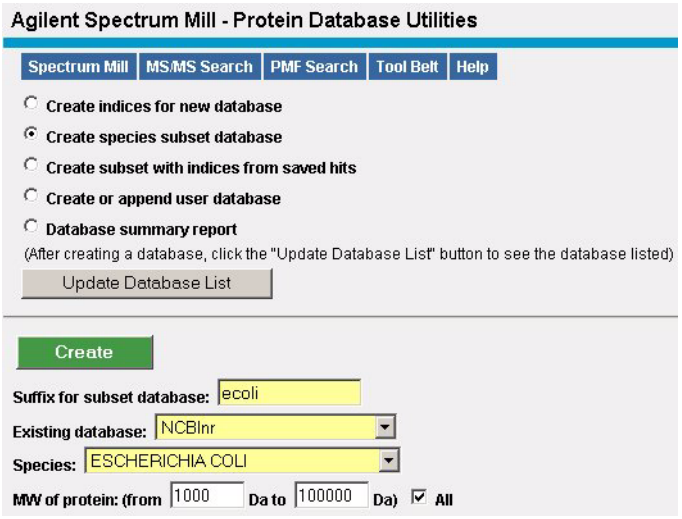


Figure 100 Create species and protein molecular weight subset database

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 form shown in Figure 101.

5 Fill in the form. Be sure to use a unique suffix so you don't overwrite previously-created databases.

6 Click the **Create** button.

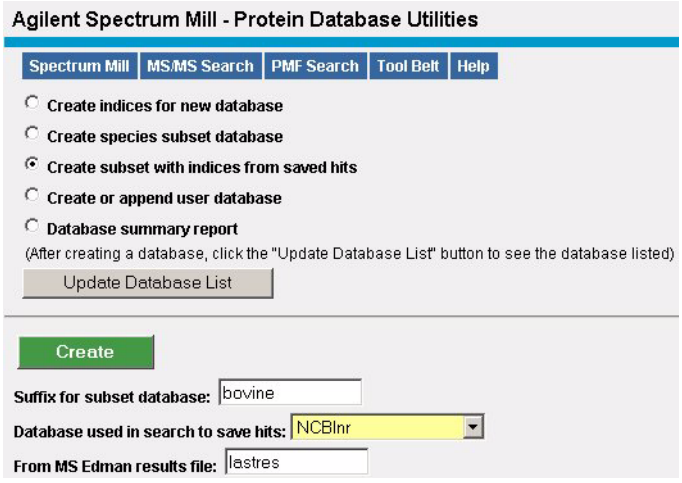


Figure 101 Create subset database with indices from saved hits

To create a user (proprietary) database

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 form shown in Figure 102.
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.)

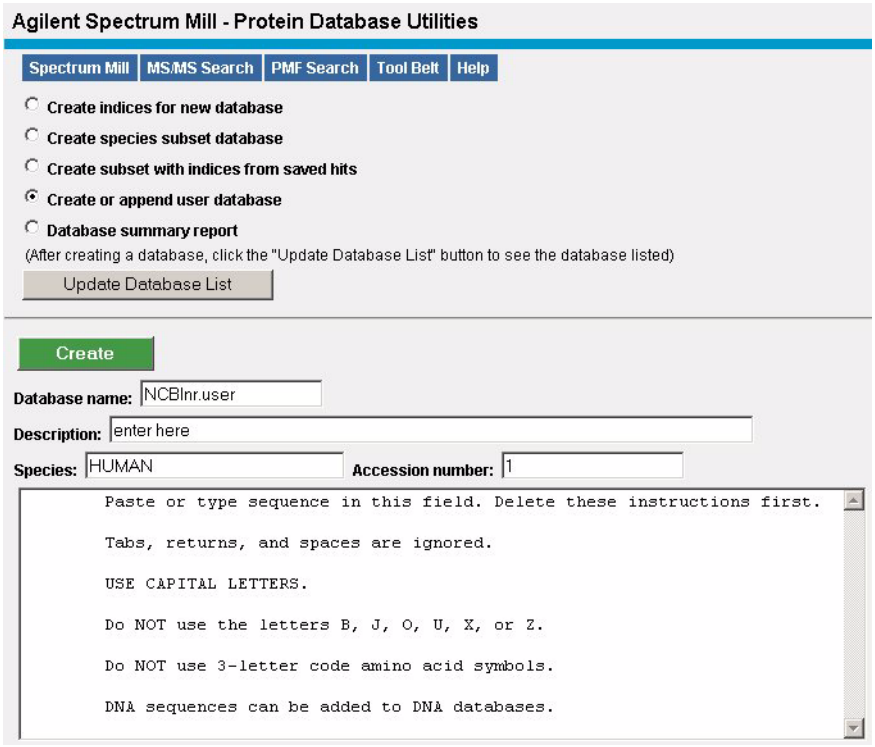


Figure 102 Create or append to user database

To add user sequences to an existing database

- See "To create a user (proprietary) database" on page 204.

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 form shown in Figure 103.

4 Fill in the form.

5 Click the **Summarize** button.

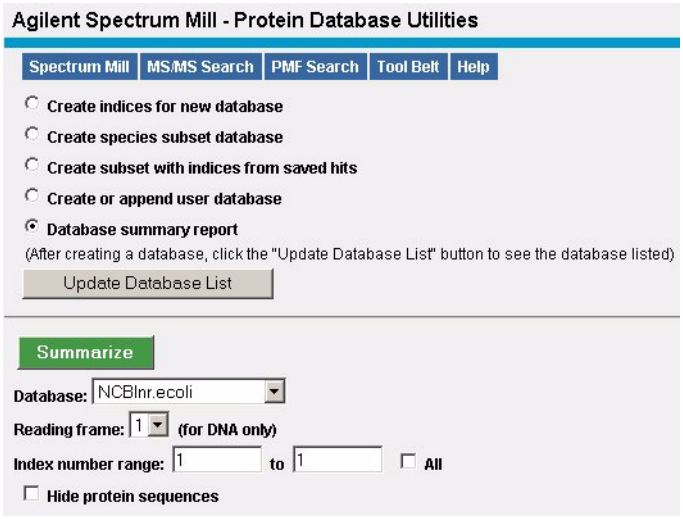


Figure 103 Database summary report

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 forms.
   - 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	msparams_mill\ urlsAccessionNum.txt
MS Digest index number	MS Digest results for database sequence	msparams_mill\ urlsIndexNum.txt
Amino acid sequence (in individual MS/MS or PMF search results)	MS Product listing for that sequence	msparams_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 **millscripts\libSpecMill\lsmGlobals.pm**.

1 Make a backup of the **millscripts\libSpecMill\lsmGlobals.pm** file.

2 Open the **millscripts\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 207.

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.
- 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**), nor the protein databases (**seqDB**).
- 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**
  - **msparams_mill** (various *.txt files, including contaminant mass list files)

10 Files Created during Spectrum Mill Data Processing

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.

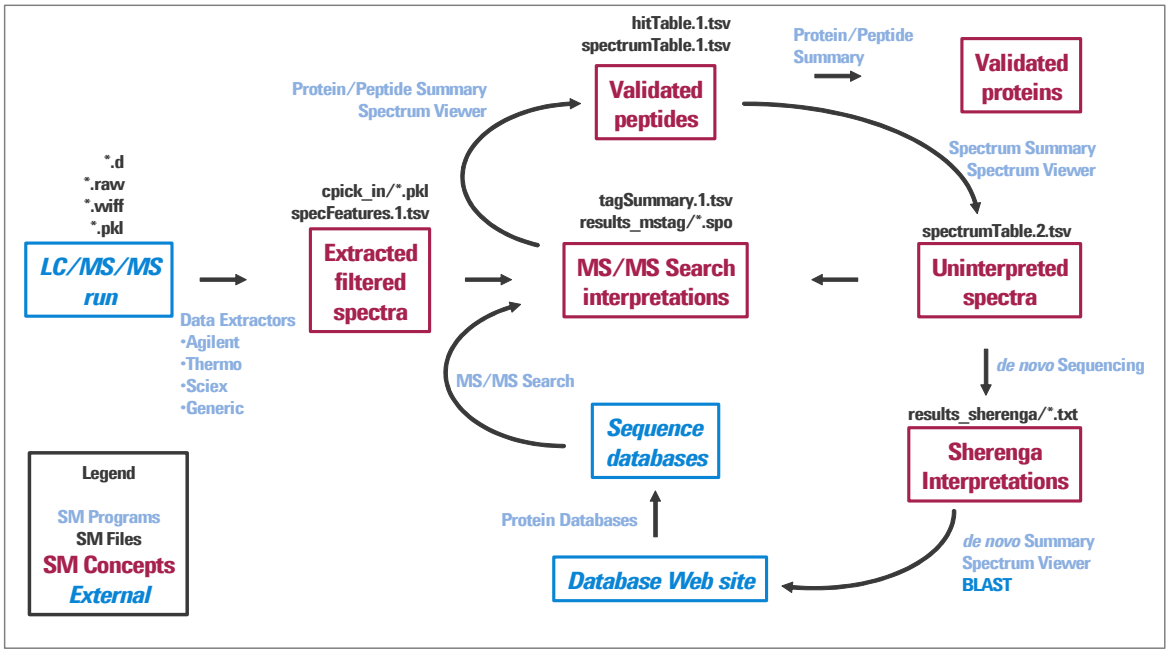


Figure 104 Architecture for MS/MS workflow

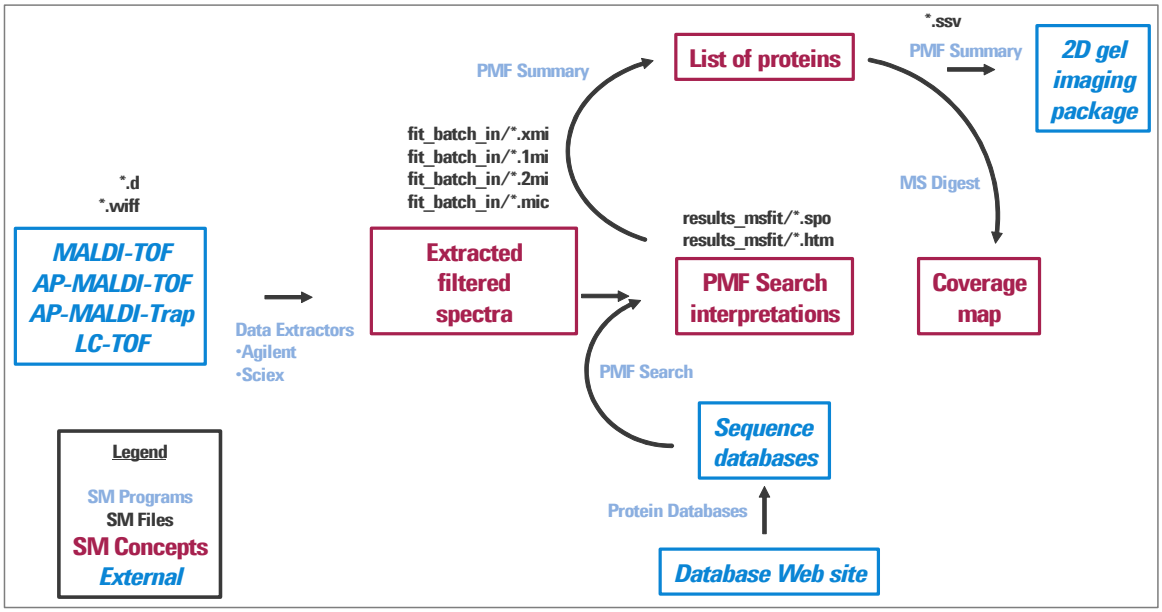


Figure 105 Architecture for MS-only workflow

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	• Series of files - one for each Agilent (*.d) raw file extracted • File name in format: Data_File_Name.aph	• These are placeholder files that correlate extracted spectra with original raw data file. • These are created only for Agilent files.

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.

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 221.
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 **msparams_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
iTRAQbatches.js	• iTRAQ 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 **millhtml\SM_js** directory • File is appended every time you type correction factors for a new batch number
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 form, 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 *de novo* Sequencing search parameters	• New file created each time you run Sherenga *de novo* 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 *de novo* 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 form.	• Contents are overwritten unless you change the file name in the MS Edman form.
lastres.spo	• Empty file • File name is **lastres** or whatever you have chosen in the MS Edman form.	
lastres.par	• Search parameters in text format • File name is **lastres** or whatever you have chosen in the MS Edman form.	• Contents are overwritten unless you change the file name in the MS Edman form.
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 form.	• Contents are overwritten unless you change the file name in the MS Edman form.

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 form.	• 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 form.	• 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 form.	• Created when you select the **Create or append user database** option and then click the **Create** button
*.idx *.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

N

O

P

Q

R

S

T

U

V

www.agilent.com

In this Book

The *Application Guide* presents the basics to get started with the Spectrum Mill MS Proteomics Workbench. In this guide you learn:

- How to process MS/MS and MS-only data sets
- Details about the software's advanced data review and validation capabilities
- 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. 2007

Printed in USA
Fifth edition, June 2007

G2721-90029

Agilent Technologies