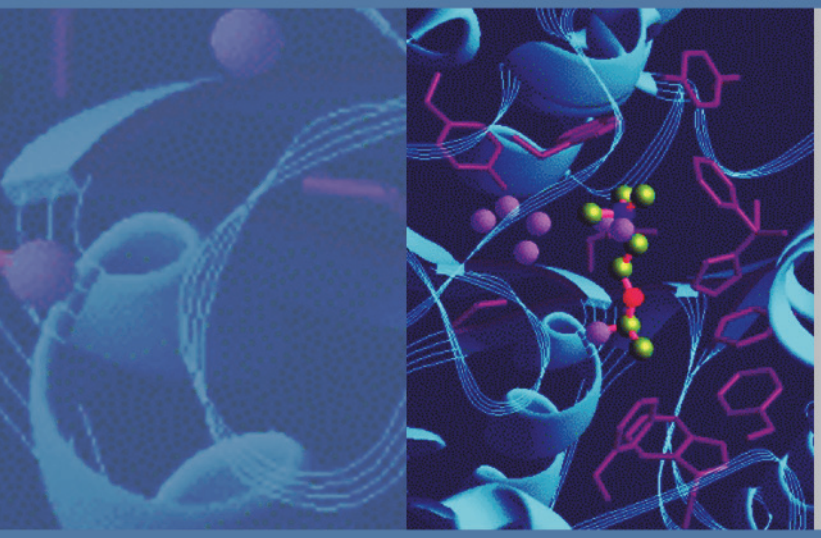


Proteome Discoverer

Version 1.4

User Guide

XCALI-97506 Revision A December 2012

DOCUMENTATION SURVEY

Part of Thermo Fisher Scientific

Thermo SCIENTIFIC

Release history: Release A, December 2012

- Software version: Thermo Proteome Discoverer version 1.4, Microsoft Windows XP 32/64 Professional (English version), Microsoft Windows 7 32/64 Professional (English version), Mascot Server 2.1

For Research Use Only. Not for use in diagnostic procedures.

Contents

Preface

This guide describes how to use the Proteome Discoverer™ 1.4 application for peptide and protein mass spectrometry analyses.

Contents

To provide us with comments about this document, click the link below. Thank you in advance for your help.

Related Documentation

The Proteome Discoverer application includes Help and these manuals as PDF files:

- *Proteome Discoverer User Guide*
- *Proteome Discoverer Installation Guide*

❖ **To view product manuals**

- *Proteome Discoverer User Guide*: Go to **Start > Programs > Thermo Proteome Discoverer 1.4 > Proteome Discoverer 1.4 User Guide**.
- *Proteome Discoverer Installation Guide*: Go to **Start > Programs > Thermo Proteome Discoverer 1.4 > Proteome Discoverer 1.4 Installation Guide**.

❖ **To open Help**

- From the main Proteome Discoverer window, choose **Help > Help Contents**.
- If available for a specific window or view, click **Help** or press F1 for information about setting parameters.

For more information, visit www.thermo.com. You can find application notes at www.thermo.com/appnotes.

System Requirements

The Proteome Discoverer application requires a license. In addition, your system must meet the following minimum requirements.

System	Requirements
Hardware	• 2 GHz processor with 2 GB RAM • DVD/R-ROM drive • Video card and monitor capable of 1280 × 1024 resolution (XGA) • Screen resolution of 96 dpi • 75 GB available on the C: drive • NTFS format
Software	• Microsoft™ Windows™ XP 32/64 Professional (English version) with latest service pack installed • Microsoft Windows 7 32/64 Professional (English version)
Mascot℠ Server	• Mascot Server 2.1 – Mascot servers running version 2.1 should be usable, but retrieving the result files (protein sequences) from the servers can be a lengthy process because you can only retrieve the protein sequences one at a time. – Mascot servers running version 2.1 should have all available updates, patches, or both from Matrix Science installed. In particular, you must install a patch that enables MIME format for the result files; otherwise, the Proteome Discoverer application cannot receive the search results from the Mascot server. • Mascot Server 2.2: Proteome Discoverer 1.4 does not support error-tolerant searches. • Mascot Server 2.3: Proteome Discoverer 1.4 does not support error-tolerant searches, Percolator-based scoring, or searches against multiple-sequence databases.

Note Ensure that port 28199 is not blocked by firewalls.

Note Ensure that the Windows operating system first has the latest Microsoft .NET Framework and Windows updates installed before installing the Proteome Discoverer application.

Special Notices

Make sure you follow the precautionary statements presented in this guide. Special notices appear in boxes.

IMPORTANT Highlights information necessary to prevent damage to software, loss of data, or invalid test results; or might contain information that is critical for optimal performance of the system.

Note Highlights information of general interest.

Tip Highlights helpful information that can make a task easier.

Contacting Us

There are several ways to contact Thermo Fisher Scientific for the information you need.

❖ **To contact Technical Support**

Phone	800-532-4752
Fax	561-688-8736
E-mail	us.techsupport.analyze@thermofisher.com
Knowledge base	www.thermokb.com

Find software updates and utilities to download at mssupport.thermo.com.

❖ **To contact Customer Service for ordering information**

Phone	800-532-4752
Fax	561-688-8731
E-mail	us.customer-support.analyze@thermofisher.com
Web site	www.thermo.com/ms

❖ **To get local contact information for sales or service**

Go to www.thermoscientific.com/wps/portal/ts/contactus.

❖ To copy manuals from the Internet

Go to mssupport.thermo.com, agree to the Terms and Conditions, and then click **Customer Manuals** in the left margin of the window.

❖ To suggest changes to documentation or to Help

- Fill out a reader survey online at www.surveymonkey.com/s/PQM6P62.
- Send an e-mail message to the Technical Publications Editor at techpubs-lcms@thermofisher.com.

Introduction

This chapter introduces you to the Proteome Discoverer application and describes its features and functionality.

Contents

Features

The Proteome Discoverer application identifies proteins from the mass spectra of digested fragmented peptides. It compares the raw data from mass spectrometry to the information from the selected FASTA database. You can use this application to analyze spectral data from all Thermo Scientific and other mass spectrometers. Specifically, the Proteome Discoverer application does the following:

- Works with peak-finding search engines such as Sequest™ and Mascot to process all data types collected from low- and high-mass-accuracy mass spectrometry (MS) instruments. The peak-finding algorithm searches the raw mass spectrometry data and generates a peak list and relative abundances. The peaks represent the fragments of peptides for a given mass and charge.
- Produces complementary data from a variety of dissociation methods and data-dependent stages of tandem mass spectrometry.
- Combines, filters, and annotates results from several database search engines and from multiple analysis iterations. The search engines correlate the uninterrupted tandem mass spectra of peptides with databases, such as FASTA. See "Using FASTA Databases" on page 101.

The Proteome Discoverer application includes the following features:

- Support for the Sequest HT, SEQUEST, and Mascot search engines.

  The Sequest HT and Mascot search engines are available as wizards or as nodes in the Workflow Editor. "Search Engines" on page 3 describes these search engines.

  Note This document refers to the algorithm and general capabilities of SEQUEST and Sequest HT collectively as Sequest. It refers to the nodes implementing Sequest's features as SEQUEST or Sequest HT.

- The Workflow Editor for searching with multiple algorithms and merging results from multiple dissociation techniques. See "Starting a New Search by Using the Workflow Editor" on page 42.

- Support for both precursor ion quantification (for example, SILAC), reporter ion quantification (for example, iTRAQ™ and Tandem Mass Tag™ [TMT]), and peak area calculation quantification. For details, see "Performing Precursor Ion Quantification" on page 243, "Performing Reporter Ion Quantification" on page 249, and "Performing Peak Area Calculation Quantification" on page 259, respectively.

- Access to annotation information from ProteinCenter, including information from the Gene Ontology (GO) database, Protein Family (Pfam) database from the Wellcome Trust Sanger Institute, and gene identifications from the Entrez gene database maintained by the National Center for Biotechnology Information (NCBI). You can use this information to annotate the proteins in your results report (Magellan storage file, or MSF). ProteinCenter is a Web-based application that you can use to download biologically enriched annotation information for a single protein, such as molecular functions, cellular components, and biological processes, from the GO database. For information, see "Protein Annotation" on page 201. You can also upload search results directly from the Proteome Discoverer application to ProteinCenter.

- Proteome Discoverer Daemon, which can perform multiple searches on multiple raw files at any given time. You can use it to perform searches on multiple raw files taken from multiple samples or replicates from the same sample. See "Using the Proteome Discoverer Daemon Utility" on page 69.

- A number of graphical views that contain detailed information about the selected peptides and proteins. You can display more than one view to perform a comparative analysis of your selected peptide or proteins. For more information, refer to the Help.

- The presentation of database search results available from multiple raw files in a single protein or peptide report. For more information, refer to the Help.

- Support for FASTA databases and indexes. See "Using FASTA Databases" on page 101.

- The ability to import protein and peptide reports in standard spectrum data formats, such as MZDATA, MZXML, MZML, and MGF. See "Importing Raw Data Files in Other Formats into a Workflow" on page 65.

- The ability to export protein and peptide reports in standard spectrum data formats, such as MZDATA, DTA, MZML, and MGF. You can also export search results to XML and tab-delimited TXT files. In addition, you can export annotated spectra for selected peptides into a ZIP file that includes an HTML page with peptide information and links to spectrum images. The Help describes how to export your data to these and other formats.
- The ability to merge filtered or unfiltered search results. For information, refer to the Help.
- A number of protein and peptide filtering and grouping options to help you sort and filter your data. For information on Proteome Discoverer's filtering capabilities, see "Filtering Data" on page 153. For information on grouping, see "Grouping Proteins" on page 174 and "Grouping Peptides" on page 185.

Search Engines

The Proteome Discoverer application includes the Sequest HT, SEQUEST, and Mascot search engines; each produces complementary data. The Sequest HT and SEQUEST search engines are distributed by Thermo Fisher Scientific. Mascot is a protein identification search engine created by Matrix Science.

The Mascot search engine uses mass spectrometry data to identify proteins from primary sequence databases. The Sequest HT and SEQUEST search engines can analyze different data types:

- Electron-transfer dissociation (ETD)
- Electron-capture dissociation (ECD)
- Collision-induced dissociation (CID)
- High-energy collision-induced dissociation (HCD)
- Pulsed collision-induced dissociation (PQD)
- ETD and ECD generate primarily c and z fragment ions with preferences for precursor ion charge states of +3 or higher.
- CID and HCD generates primarily b and y fragment ions with preferences for precursor ion charge states of +3 or lower.
- PQD and HCD do not exhibit a low-mass cutoff and are good for reporter-ion experiments.

Frequently, peptides identified by CID, PQD, or HCD are not observed with ETD or ECD, and vice versa, so that combining results from, for example, CID and ETD can enhance sequence coverage. Many times CID and ETD identify the same peptides, often with different precursor ion charge states. Combining ETD and CID results improves confidence in identifications.

SEQUEST Search Engine

The SEQUEST search engine is specifically developed and optimized to evaluate both high-mass-accuracy and low-mass-accuracy ETD, ECD, CID, HCD, and PQD data. You can use Sequest combined with automated LC-MS/MS and intelligent data acquisition tools to ensure the routine identification of low-abundance proteins in complex mixtures.

The Proteome Discoverer application extracts relevant MS/MS spectra from the raw file and determines the precursor charge state and the quality of the fragmentation spectrum.

The Sequest search algorithm correlates experimental MS/MS spectra through comparisons to theoretical in-silico peptide candidates derived from protein databases. The proprietary cross-correlation identification algorithm at the core of Sequest uses a sophisticated scoring system to help assess results. Sequest looks for characteristic spectral patterns and then critically evaluates the equivalence of experimental and theoretical MS/MS spectra. The identification algorithm extracts information and correctly identifies proteins even when protein sample sizes are limited and the signal-to-noise ratio of spectra is low.

You can extract specific information from your results through the interactive data summary screens. With a click, you can examine a fully annotated MS/MS spectrum, or view the percent peptide coverage of an identified protein.

Sequest provides excellent search results on data acquired with Thermo Scientific ion trap mass spectrometers. Using accurate mass windows decreases the search time, increases the accuracy of the result, and decreases the false positive rate.

The Proteome Discoverer probability-based scoring system rates the relevance of the best matches found by the Sequest algorithm. With this probability-based scoring, the application can independently rank the peptides and proteins and increase the confidence in protein identification. Additionally, this scoring system minimizes the time needed for data interrogation or results review, increasing the overall throughput of the analysis.

You can also automatically determine false discovery rates by comparing the results of forward and reversed databases, which provides an additional means of increasing confidence in protein identification.

Sequest HT Search Engine

The Sequest HT search engine calculates XCorr scores for peptide matches and provides the peptide matches having the best XCorr score for each spectrum. It is similar to the SEQUEST search node, which calculates a preliminary SpScore score and uses it to filter peptide candidates. It calculates XCorr values for PSMs only if they pass the SpScore filter. The Sequest HT node calculates the XCorr value for every peptide candidate. It can therefore take longer than the SEQUEST node, especially when the number of peptide candidates is large and the processing uses several dynamic modifications. In most cases, however, multiple-thread searching is faster with Sequest HT.

Mascot Search Engine

Mascot uses mass spectrometry data to identify proteins from primary sequence databases.

For more details on the Mascot search engine, visit http://www.matrixscience.com.

Wizards and Workflow Editor

You can use the Proteome Discoverer application's search wizards or its Workflow Editor to conduct data analysis searches of your spectra.

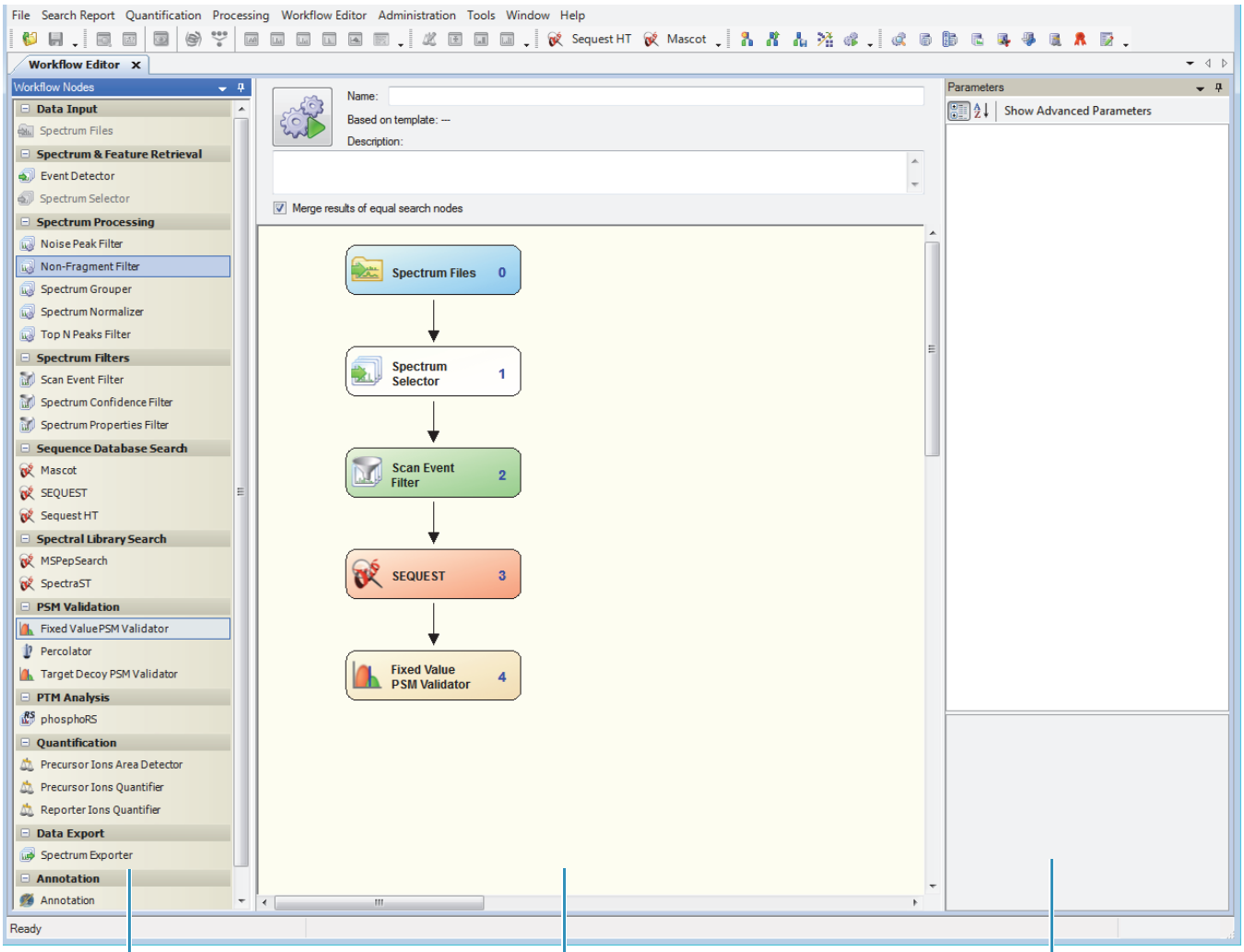
The search wizards are predefined to enable you to quickly set your search parameters and obtain results. The Proteome Discoverer application includes a wizard for the Sequest HT and Mascot search engines.

For information about how to use the wizards, see "Starting a New Search by Using the Search Wizards" on page 29.

The Workflow Editor provides greater flexibility in creating custom search results. Use its three-pane display to create a custom workflow. The Workflow Nodes pane of the application's interface contains seven categories of workflow choices. A typical workflow uses three or more options from these categories, as shown in Figure 1. To start a new workflow, begin with a node from the Data Input category. For more information, see "Starting a New Search by Using the Workflow Editor" on page 42.

When you activate any node from the Workflow Nodes pane, the parameters appear in the Parameters pane.

Figure 1. Workflow Editor workspace



Quantification

The Proteome Discoverer application offers both isotopically labeled precursor ion quantification and isobarically labeled reporter ion quantification methods, which you can also edit.

SILAC is an isotopically labeled quantification method that uses in-vivo metabolic labeling to detect differences in the abundance of proteins in multiple samples. SILAC uses the Precursor Ions Quantifier node in the Workflow Editor.

iTRAQ and TMT are very similar isobarically labeled quantification methods that use external reagents, or tags, to chemically label proteins and peptides to detect differences in abundances. TMT quantification offers default 2plex and 6plex quantification methods, and iTRAQ offers 4plex and 8plex quantification methods. You can use these methods to create your own quantification templates.

iTRAQ and TMT use the Reporter Ions quantifier node in the Workflow Editor.

For detailed information about isobarically and isotopically labeled quantification, see "Performing Reporter Ion Quantification" on page 249 and "Performing Precursor Ion Quantification" on page 243.

The Proteome Discoverer application also offers peak area calculation quantification, which you can use to determine the area for any quantified peptide. This type of quantification uses the Precursor Ions Area Detector node. For more information about peak area calculation quantification, see "Performing Peak Area Calculation Quantification" on page 259.

The Qual Browser Application

With the Qual Browser application, you can view the entire ion chromatogram and browse individual precursor and MS^n data. You can filter the results in a variety of ways, for example, to produce a selected ion chromatogram. When you select a peptide and choose Tools > Open QualBrowser, the Proteome Discoverer application passes the currently active raw file for Qual Browser operations. For more information about the Qual Browser application, see "Using the Qual Browser Application" on page 149.

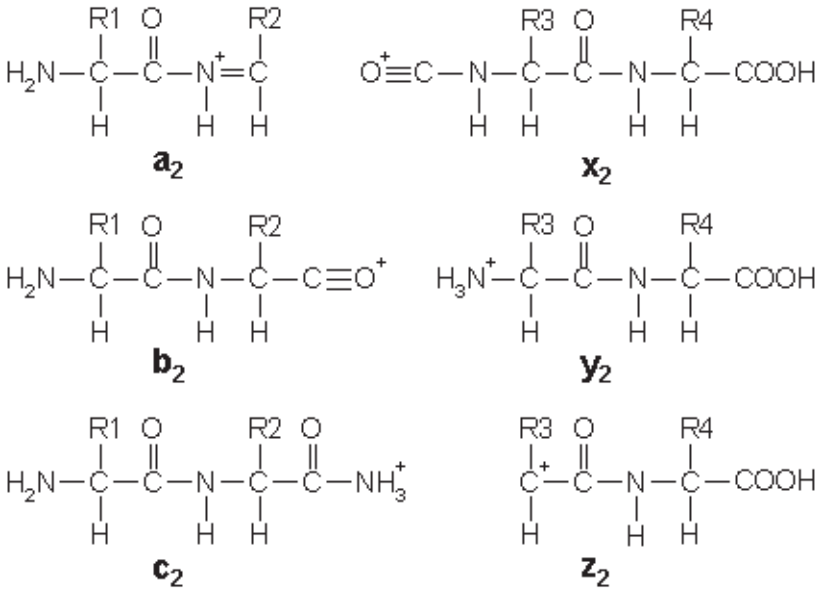
Peptides and Fragment Ions

The types of fragment ions observed in an MS/MS spectrum depend on several factors, such as the primary sequence, the energy source, and the charge state.

Fragment ions of peptides are produced by a collision-induced dissociation (CID) process in which a peptide ion is fragmented in a collision cell. Low-energy CID spectra are generated by MS/MS and ESI, and are sequence-specific. The fragment ion spectra contain peaks of the fragment ions formed by cleavage of the peptide bond and are used to determine the amino acid sequence. A fragment must have at least one charge for it to be detected. If this charge is retained on the N terminal fragment, the ion is classed as a, b, or c. If the charge is retained on the C terminal fragment, the ion type is x, y, or z. A subscript indicates the number of residues in the fragment.

In addition to the proton carrying the charge, c ions and y ions abstract an additional proton from the precursor peptide, as shown in Figure 2.

Figure 2. Structures of six singly charged sequence ions



Fragmentation Methods

The Proteome Discoverer application supports the following fragmentation types:

- CID – Uses the collision-induced dissociation (CID) method of fragmentation, where molecular ions are accelerated to high kinetic energy in the vacuum of a mass spectrometer and then allowed to collide with neutral gas molecules such as helium, nitrogen, or argon. The collision breaks the bonds and fragments the molecular ions into smaller pieces.
- ECD – Uses the electron capture dissociation (ECD) method of fragmentation, where multiply protonated molecules are introduced to low-energy free electrons. Capture of the electrons releases electric potential energy and reduces the charge state of the ions by producing odd-electron ions, which easily fragment.
- HCD – Uses the high-energy collision-induced dissociation (HCD) method of fragmentation, where the projectile ion has laboratory-frame translation energy higher than 1 keV. HCD produces a highly abundant series of reporter ions for TMT and iTRAQ quantification.
- ETD – Uses the electron transfer dissociation (ETD) method of fragmentation, where singly charged reagent anions transfer an electron to multiply protonated peptides within an ion trap mass analyzer to induce fragmentation. ETD cleaves along the peptide backbone while side chains and modifications such as phosphorylation are left intact. This method is used to fragment peptides and proteins.

- IRMPD – With the infrared multi-photon dissociation (IRMPD) method of fragmentation, an infrared laser is directed at the ions in the vacuum of the mass spectrometer. The target ions absorb multiple infrared photons until they reach more energetic states and begin to break bonds, resulting in fragmentation.
- PQD – Uses the pulsed Q collision-induced dissociation (PQD) method of fragmentation, where precursor ions are activated at a high value, a parameter that determines the stability of an ion's trajectory in an ion trap mass analyzer. Then, a time delay occurs to allow the precursor to fragment, and then a rapid pulse is applied to a low value where all fragment ions are trapped. The product ions can then be scanned out of the ion trap and detected. PQD fragmentation produces precise, reproducible fragmentation and has been used for iTRAQ peptide quantification on the LTQ™ mass spectrometer using both electrospray and MALDI source ionization.

MudPIT Experiments

Multidimensional Protein Identification Technology (MudPIT) experiments investigate complex proteomes by applying multidimensional chromatography to the samples before acquisition in the mass spectrometer. Typically, this process results in several dozen or even a few hundred fractions that are separately analyzed by LC-MS, resulting in one raw file per sample fraction. Analyzing gel slices or performing in-depth follow-up acquisitions also results in multiple fractions. Because all these fractions belong to the same sample, the Proteome Discoverer application can process all raw files from these fractions as one contiguous input file and generate a single result file. For detailed information about processing MudPIT samples, see "Using the Proteome Discoverer Daemon Utility" on page 69.

Workflow

Through settings that you specify in the Proteome Discoverer application, you can search, filter, and sort raw files with the Sequest and Mascot algorithms. In addition to creating reports from the analyzed data, the application extracts relevant MS/MS spectra from the raw file and determines the precursor charge state. Filters in the application remove false positives and other irrelevant information with a variety of user-specified methods.

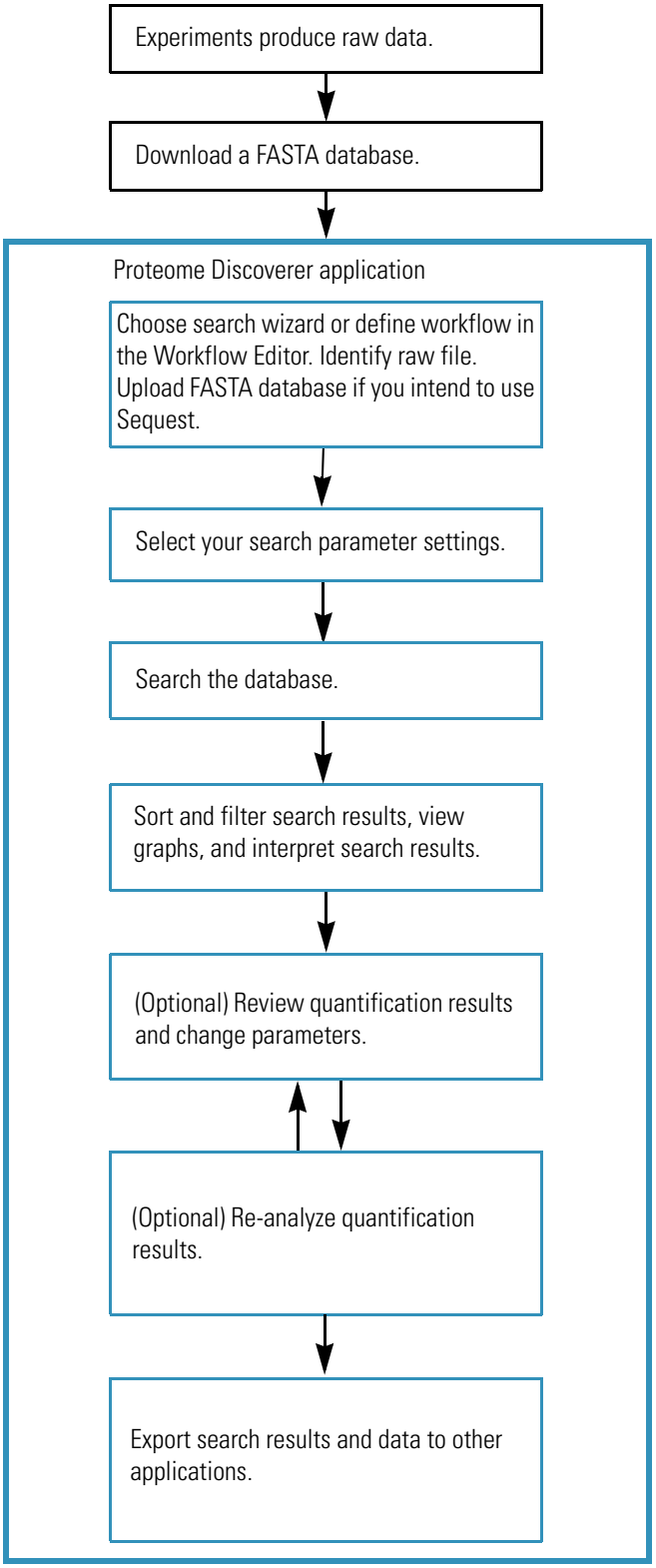
Note You can filter data according to false discovery rates that you define through the use of decoy databases that you specify in the workflow.

Using the standard Proteome Discoverer workflow involves the following steps when you process, analyze, and interpret mass spectrometry data. These steps are shown graphically in Figure 3.

1. Upload a FASTA database, if necessary, to use Sequest.
2. Choose a search wizard or create a workflow in Workflow Editor. Identify the raw file.
3. Select parameter settings in the search wizard or the nodes of the Workflow Editor.

4. Begin a search of the raw data. The Proteome Discoverer application initiates a search against a FASTA database.
5. Sort and filter the search report, generate graphs and views, and interpret the search results.
6. (Optional) Review the quantification results and change parameters.
7. Reanalyze the quantification results.

Figure 3. The Proteome Discoverer workflow



Inputs and Outputs

The Proteome Discoverer application can accept several different file formats as input and can export data in several formats.

FASTA Databases

The Proteome Discoverer application includes FASTA databases, including multiple example FASTA databases and example raw files. Use these files when exploring and learning how to use the application. For a detailed description of the different types of FASTA databases and their purpose, see "Using FASTA Databases" on page 101.

Inputs

The Proteome Discoverer application accepts the following file types as input:

- Xcalibur raw files contain raw data collected from a mass spectrometer.
- Mascot Generic Format (MGF) files are mass spectral files produced during Mascot analysis. They contain a list of precursor ions, their fragments, and the masses of the fragments.
- Extensible Markup Language (XML) files contain workflow templates.
- MZXML files are standard 2.*x* mass spectrometer data format files, developed at the Seattle Proteome Center at the Institute for Systems Biology (ISB), that contain a list of precursor ions, their fragments, and the masses of the fragment.
- MZDATA files are common data format files developed by the Human Proteome Organization (HUPO) for proteomics mass spectrometry data. These files are in version 1.05 format. They are exported with XML indentation enabled so that the different XML tags are broken into multiple lines instead of merged into one line.
- MZML files are a combination of .mzData and .mzXML formats developed by the Human Proteome Organization Standard Initiative (HUPO-PSI) and the Seattle Proteome Center at the Institute for Systems Biology (ISB). The Proteome Discoverer application supports version 1.1.0 of the MZML format.
- Magellan Storage (MSF) files contain the results of the searches conducted by the search wizards or the Workflow Editor.

Outputs

The Proteome Discoverer application creates the following file types as output:

- DTA Archive (DTA) files are files containing MS^n data for single or grouped scans.
- Mascot Generic Format (MGF) files are mass spectral files produced during Mascot analysis. They contain a list of precursor ions, their fragments, and the masses of the fragments.
- MZDATA files are common data format files developed by the Human Proteome Organization Standard Initiative (HUPO-PSI) for proteomics mass spectrometry data. These files are in version 1.05 format. They are exported with XML indentation enabled so that the different XML tags are broken into multiple lines instead of merged into one line.
- Magellan storage (MSF) files contain the results of the searches conducted by the search wizards or the Workflow Editor.
- Extensible Markup Language (XML) files contain workflow templates.
- MZXML files are standard 2.*x* mass spectrometer data format files developed at the Seattle Proteome Center at the Institute for Systems Biology (ISB) that contain a list of precursor ions, their fragments, and the masses of the fragment.
- MZML files are a combination of MZDATA and MZXML formats developed by the Human Proteome Organization Standard Initiative (HUPO-PSI) and the Seattle Proteome Center at the Institute for Systems Biology (ISB). The Proteome Discoverer application supports version 1.1.0 of the MZML format.
- ProtXML files contain protein identifications from MS/MS-derived peptide sequence data. They are created by the File > Export > To ProtXML command.
- PepXML files contain peptides that are included in the results of searches performed by the Sequest HT, SEQUEST, and Mascot search engines. They are in PepXML format version 1.14, which is an open data format developed by SPC/Institute for Systems Biology for storing, exchanging, and processing peptide sequence assignments from MS/MS scans. PepXML files are created by the File > Export > pepXML command.

  The Proteome Discoverer application supports version 1.14.
- Tab-delimited TXT files are in a simple text format that stores tabular data and is widely used to exchange data between different computer programs.

Limitations

This release of the Proteome Discoverer application has the following limitations:

- The spectra count is not directly available in the application results report. However, the number of identified peptides is displayed for each protein. This number should be similar to the spectra count for that protein.
- The Proteome Discoverer application supports peptide quantification methods that use reporter ions. Examples of these methods are TMT and iTRAQ. The application also supports peptide quantification methods that measure precursor ion abundances. Examples of these methods are SILAC, ICPL, ^{18}O, ^{15}N, and label-free methods.

New Features in This Release

The Proteome Discoverer application version 1.4 adds the following new features.

Sequest HT Search Engine

The new Sequest HT search engine is a reimplementation of the Sequest algorithm that increases overall performance by using modern multicore and multiprocessor systems. It also uses multiple search threads. It does not use the SpScore filter; instead, it calculates XCorr for every candidate. The scores from the Sequest HT and SEQUEST search engines are not identical, because the Sequest HT search engine uses a slightly changed cross-correlation and exact mass differences for the flanking ions of peaks in the theoretical spectra.

Spectrum Library Searching

The Proteome Discoverer application offers the ability to search large spectrum libraries, which are libraries of measured (consensus) spectra from actual previous experiments. Two new spectral library search nodes, SpectraST and MSPepSearch, use spectral libraries. These search engines identify peptides by comparing the spectra to the reference spectra in the library. You can search spectrum libraries downloaded from the National Institute of Standards and Technology (NIST™) and the PeptideAtlas home page.

MSPepSearch Node

The MSPepSearch node searches spectrum libraries downloaded from NIST. It is faster than SpectraST, but there are no decoy spectral libraries available that are required to estimate the false discovery rate (FDR) by using a target decoy false discovery rate calculation or by using Percolator.

SpectraST Node

The SpectraST node searches spectrum libraries downloaded from NIST and the PeptideAtlas home page. It searches more slowly than the MSPepSearch node but automatically generates decoy libraries when you register a library. You can therefore calculate the false discovery rate by using the Target Decoy PSM Validator node or the Percolator node.

Spectral Library Administration

The new Spectrum Libraries view on the Administration page lists all the spectrum libraries that you downloaded from NIST or the Peptide Atlas home page.

Mirror Plots

In the Peptide Details Identification view, you can display a mirror plot for PSMs identified by a spectral library search to visually verify matches between measured spectra from your experiment and the reference spectra in the spectrum library.

New Workflow Editor Nodes

Proteome Discoverer version 1.4 divides the Peptide Validator node of the 1.3 release into the Fixed Value PSM Validator node and the Target Decoy PSM Validator node.

Fixed Value PSM Validator Node

The Fixed Value PSM Validator node assigns confidence levels according to the fixed score thresholds that you chose in preceding searches.

You can only connect search nodes that do not perform decoy searches, such as MSPepSearch, to the Fixed Value PSM Validator node.

The Fixed Value PSM Validator node has no parameters.

Target Decoy PSM Validator Node

The Target Decoy PSM Validator node automatically calculates confidence levels according to the outcome (score distribution) of the target-decoy search that preceded it.

Phospho*RS* 3.0 Node

The phospho*RS* 3.0 node updates the preliminary version of the phospho-site localization algorithm that was distributed with the 1.3 Proteome Discoverer application. The new features of this update are the following:

- Improved performance: The updated phospho-site localization algorithm performs parallel calculations using multiple processor cores, if available.

- Individual peak depth approach: The algorithm determines the optimal number of peaks (that is, the best peak depth) considered for localization of phosphorylation sites for each *m/z* window individually, which increases the sensitivity of site localization for CID data.
- Optimized scoring parameters: Depending on the applied fragmentation technique, the algorithm uses different fragment ion types for scoring to provide the highest possible sensitivity. For CID data, it scores only singly and doubly charged b and y ions. For analysis of HCD spectra, the algorithm also considers neutral loss ions. In contrast, when localizing phosphorylation sites in ETD spectra, the algorithm considers only singly charged c, z, and y+H ions.
- Additional node parameters: The phospho*RS* 3.0 node adds new parameters. For example, you can specify whether the Proteome Discoverer application should consider neutral loss peaks for scoring. Moreover, you can set the maximum number of phospho-isoforms and PTMs per peptide that the application considers. If a certain peptide exceeds this cutoff, the application does not analyze it.
- Changed output column headings: The phospho*RS* 3.0 output appears in three columns in the MSF file: phospho*RS* Site Probabilities, Binomial Peptide Score, and Isoform Confidence Probability. The Site Probabilities column appears by default, but you must choose the other two columns with the Column Chooser.

New Protein Annotations

The Proteome Discoverer application has added new features to its retrieval of protein annotations.

Entrez Gene IDs

The Proteome Discoverer application can retrieve Entrez gene identifications from ProteinCenter. The Entrez gene identification is a unique identification assigned to the genes in the Entrez database maintained by the National Center for Biotechnology Information (NCBI). The database assigns an identifier to all proteins transcribed from the corresponding gene. The Proteins page of the results report displays these identifications in the Gene IDs column. You can use this information to group or cluster together biologically meaningful proteins.

Hierarchical GO Terms

Gene ontology (GO) terms are related in hierarchical graphs. These graphs contain all the ancestor terms of the term associated with a protein. You can display the annotated GO term and all its hierarchical terms in the new GO Terms column in the output MSF file. For more information on this feature, see "Displaying GO Accessions" on page 212.

Mascot Quantification Mode

When you use the Mascot node on the Mascot server as the search engine in a quantification workflow, you can set up to nine dynamic and static modifications as parameters. However, if you want to set more modifications as parameters, you can use the Mascot node to configure quantification methods on the Mascot server. Modifications in a quantification method are organized into groups classified as fixed, variable, or exclusive. You can use the node's From Quan Method parameter to select the dynamic modifications to search for rather than manually specifying each modification with a Dynamic Modifications parameter.

For detailed information on this capability, see "Searching for Quantification Modifications with Mascot" on page 261.

Getting Started

This chapter describes how to use Proteome Discoverer search wizards and the Workflow Editor to define your search parameters. The search wizards are the quickest way to start using the Proteome Discoverer application.

Contents

Starting the Proteome Discoverer Application

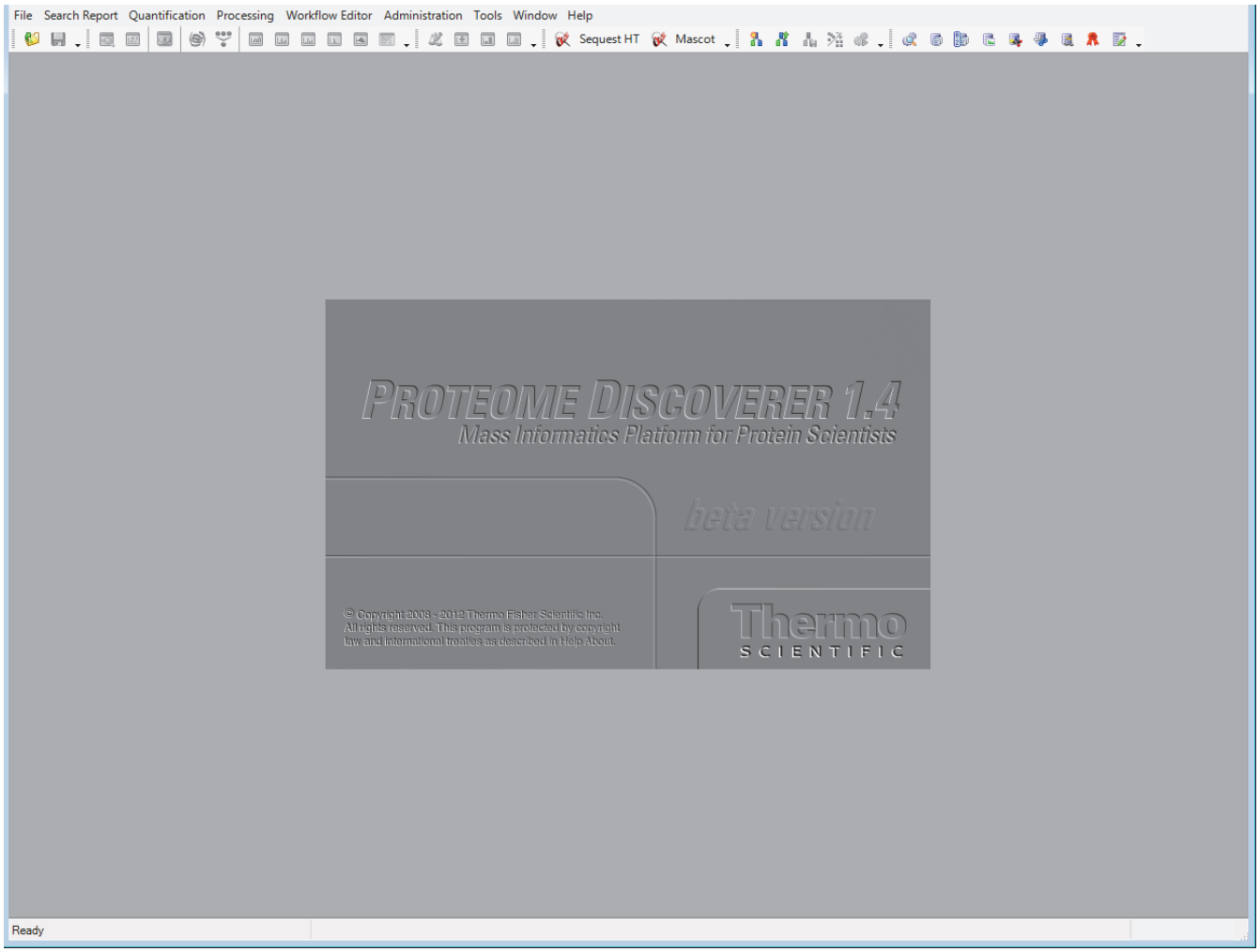
Open the Proteome Discoverer application by choosing a Start menu command or clicking a desktop icon.

❖ **To start the Proteome Discoverer application**

- From the Start menu, choose **Programs > Thermo Proteome Discoverer** or click the **Proteome Discoverer** icon,  , on your desktop.

  The Proteome Discoverer main window opens, as shown in Figure 4.

Figure 4. Proteome Discoverer main window



For information on the features of this window and how to customize them, refer to the Help. For instructions on opening an MSF file, refer to the Help.

Closing the Proteome Discoverer Application

Save your changes before you exit the Proteome Discoverer application, because it does not prompt you.

❖ **To close the Proteome Discoverer application**

- Choose **File > Exit**.

  The Proteome Discoverer application closes.

Configuring Search Engine Parameters

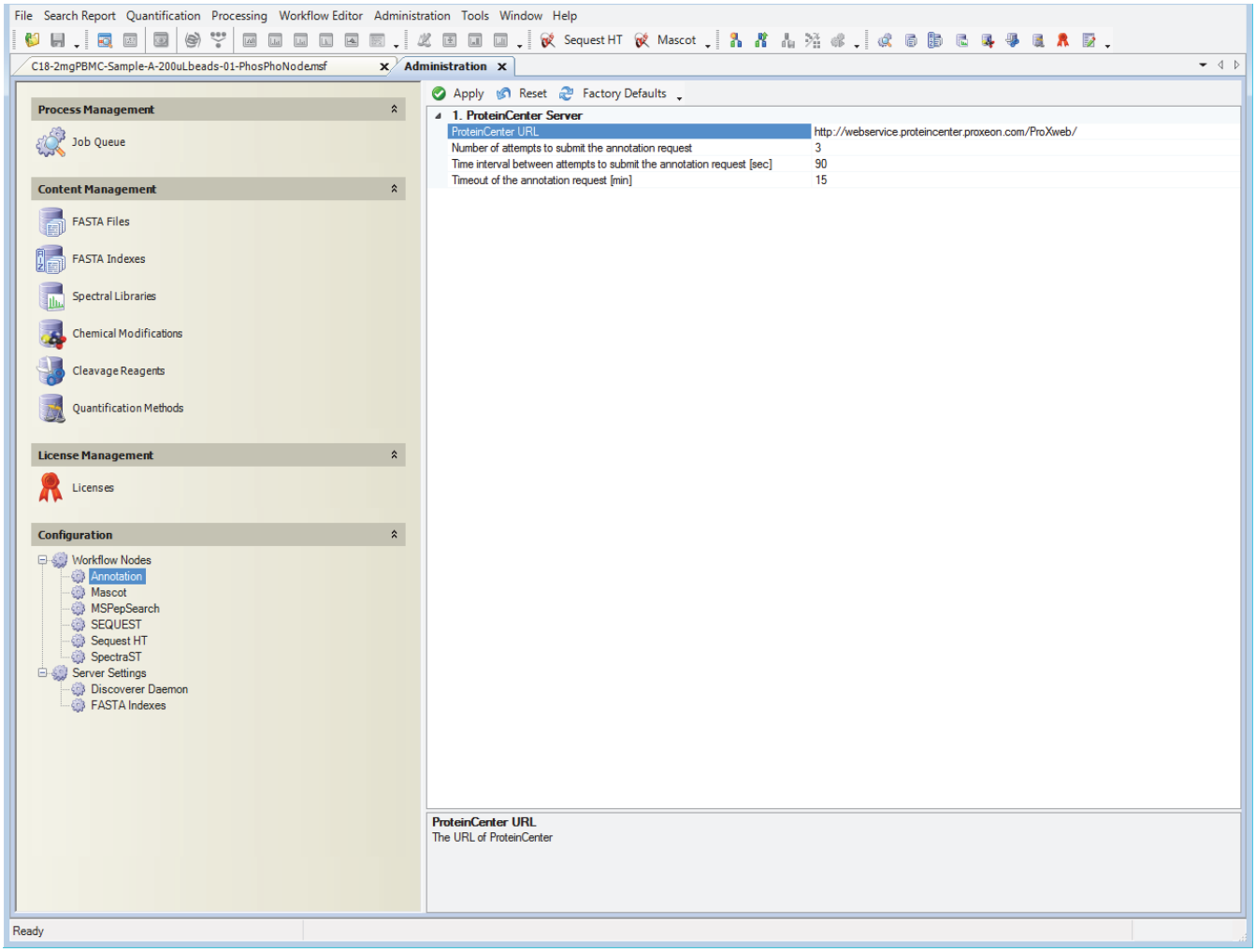
Before you execute the search, you can configure certain search parameters for the Sequest HT, SEQUEST, and Mascot search engines.

❖ To configure search parameters

1. Choose **Administration > Configuration**, or click the **Edit Configuration** icon, .

   The Administration page changes to the Configuration view, shown in Figure 5.

Figure 5. Configuration view of the Administration page



2. Follow these procedures:

   - Configuring the SEQUEST Search Engine
   - Configuring the Mascot Search Engine
   - Configuring the Sequest HT Search Engine

Configuring the Sequest HT Search Engine

Follow these steps to configure the Sequest HT search engine.

❖ To configure the Sequest HT search engine

1. On the Administration page, click **Sequest HT** under Workflow Nodes in the Configuration section.
2. In the Automatic box, specify whether you want the Proteome Discoverer application to automatically estimate the workload level.

   The default is True, which means that the application automatically estimates the workload level.
3. (Optional) If you set the Automatic parameter to False, do the following:
   a. In the Number of Spectra Processed At Once box, specify the maximum number of spectra that the Sequest HT search engine can process at once.

      The minimum value is 1000, and there is no maximum. The default is 3000.

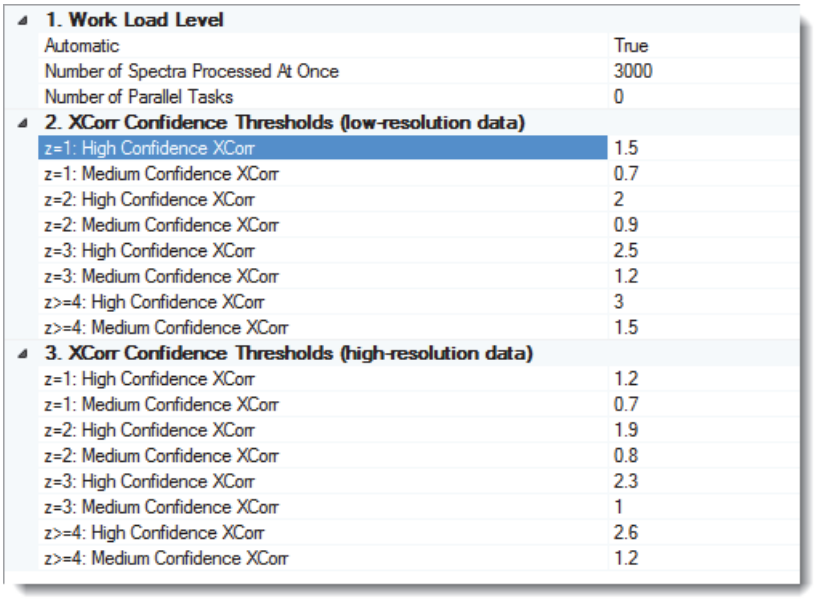
      The larger the value, the more memory is required.
   b. In the Number of Parallel Tasks box, specify the number of search tasks that Sequest HT can perform at the same time.

      The minimum value is 0, and there is no maximum. The default is 0.

      If you set this parameter to 0, this search engine performs as many parallel tasks as the number of available CPUs can handle.
4. If you are using the Sequest HT search engine to search low-resolution data, set the XCorr confidence thresholds under the XCorr Confidence Thresholds (low-resolution data) parameter.

   The default values appear in Figure 6.

Figure 6. Sequest HT configuration parameters



For information on these parameters, refer to the Help.

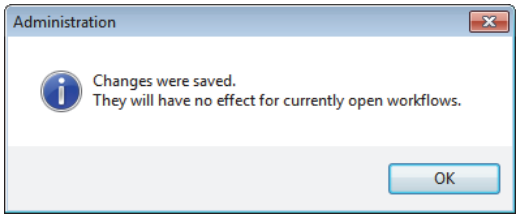
5. If you are using the Sequest HT search engine to search high-resolution data, set the XCorr confidence thresholds under the XCorr Confidence Thresholds (high-resolution data) parameter.

   The default values appear in Figure 6.

6. If you changed any settings, click Apply.

   The message box shown in Figure 7 appears:

Figure 7. Administration message box



7. Click **OK**.

Note Click Reset to return to the default values.

8. Restart your machine.

Configuring the SEQUEST Search Engine

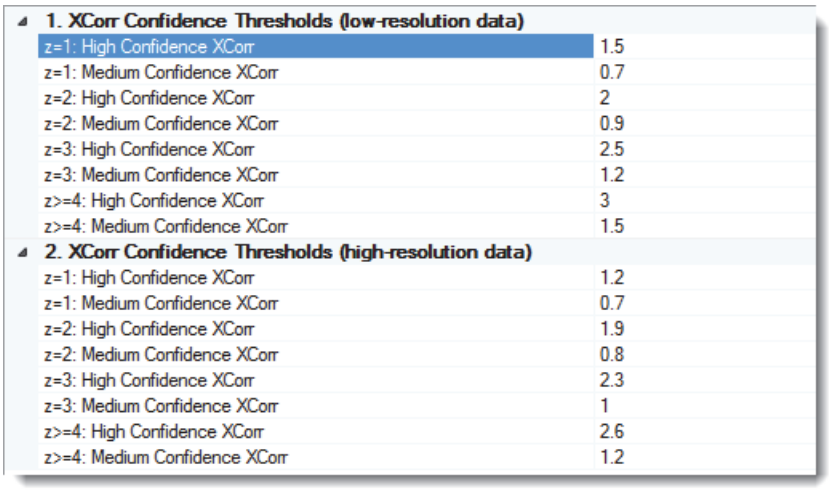
For searches with the SEQUEST search engine, specify how to display the peptide confidence by default. The SEQUEST Search engine scores the number of fragment ions that are common to two different peptides with the same precursor mass and calculates the cross-correlation score for all candidate peptides queried from the database. By default, it sorts the resulting XCorr values in descending order.

❖ To configure the SEQUEST search engine

1. On the Administration page, click **SEQUEST** under Workflow Nodes in the Configuration section.
2. If you are using the SEQUEST search engine to search low-resolution data, set the XCorr confidence thresholds under the XCorr Confidence Thresholds (low-resolution data) parameter.

   The default values appear in Figure 8.

   Figure 8. XCorr confidence thresholds for the SEQUEST search engine

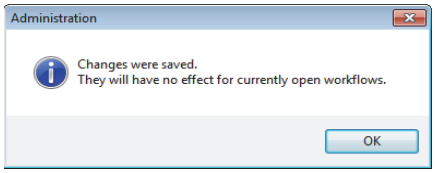
   

   For information on these parameters, refer to the Help.
3. If you are using the SEQUEST search engine to search high-resolution data, set the XCorr confidence thresholds under the XCorr Confidence Thresholds (high-resolution data) parameter.

   The default values appear in Figure 8.
4. If you changed any settings, click Apply.

The message box shown in Figure 9 appears:

Figure 9. Administration message box



5. Click **OK**.

Note Click Reset to return to the default values.

Configuring the Mascot Search Engine

Before using the Mascot search engine, you must direct the Proteome Discoverer application to the location of the Mascot server and configure the parameters that control access to the Mascot server. If your Mascot search fails, the following procedure can help you check for server problems.

- Directing the Proteome Discoverer Application to the Mascot Server Location
- Configuring Mascot Parameters
- Troubleshooting Failed Mascot Searches

Directing the Proteome Discoverer Application to the Mascot Server Location

To connect to a Mascot server, refer to the "How to Connect to a Mascot Server" section of the Proteome Discoverer release notes included on every Proteome Discoverer installation DVD. To test the connection between the Proteome Discoverer application and the Mascot server, refer to "Testing the Connection to the Mascot Server," in the *Proteome Discoverer Installation Guide*.

❖ **To direct Proteome Discoverer to the Mascot server location**

1. Open a Web browser and try to access the Mascot server through its URL.

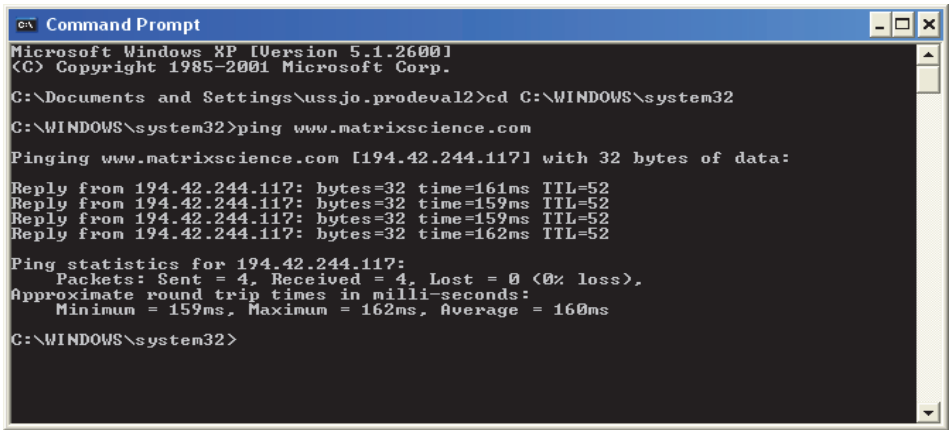
   If you cannot access the Mascot server, it might not be running, or the URL might not be correct. In this case, contact your system administrator to assist you.

2. If you can obtain Web access to the Mascot server, test to see if the ping command, which is used to reach the sever, is blocked. Do the following:

   - Open a command shell and type **`ping`** ***`Mascot_server_name`***.

     If the ping command is successful, the output should resemble that shown in Figure 10.

Figure 10. Output of a successful ping command



If the pin command is unsuccessful, a firewall on your computer or on the Mascot server computer or a bad network connection might be blocking the ping command. Contact your system administrator to assist you in resolving this problem.

If you can obtain Web access to the Mascot server and the ping test is successful but the same URL is not accepted in the Proteome Discoverer application, a type of user authentication restriction might be active. In this case, the error message issued by the Proteome Discoverer application should provide information about missing authentication. If it does not, send an error report.

Configuring Mascot Parameters

Before using the Mascot search engine, set the parameters that govern access to the Mascot server.

❖ To configure the Mascot search engine

1. On the Administration page, click **Mascot** under Workflow Nodes in the Configuration section.

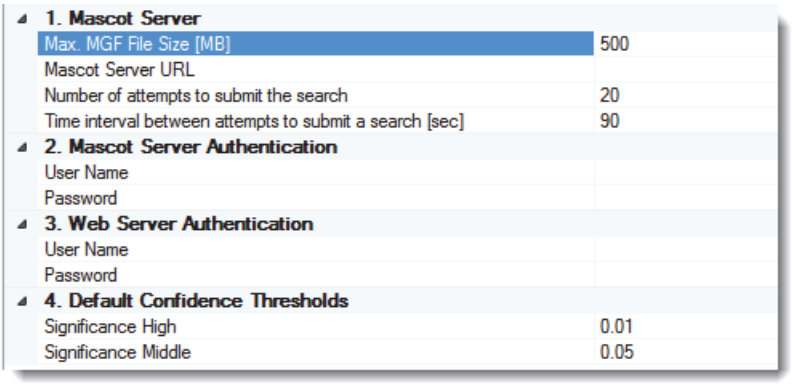
   The Proteome Discoverer application generates an MGF file that contains the search settings and all mass spectral information. It submits this file to the Mascot server through a Web server, which might have a file size limitation. A search that generates large amounts of data—for example, a search with multiple raw files—could create an MGF file that exceeds this limitation. The Max. MGF File Size parameter avoids this limitation by performing several separate Mascot searches and merging the results.

2. To split the MGF file and avoid any potential file-size limitations on the Web server, enter the maximum size, in megabytes, that the MGF file can be in the Max. MGF File Size [MB] box as shown in Figure 11.

   This size should be less than the file size permitted by the Web server.

   The minimum file size is 20, and there is no maximum. The default file size is 500 megabytes.

Figure 11. Maximum MGF file size on the Mascot server



For information on these parameters, refer to the Help.

3. In the Number of Attempts to Submit the Search box, specify the number of times that the Proteome Discoverer application tries to submit the search when the Mascot server is busy.

   The minimum value is 0, and there is no maximum value. The default is 20.

4. In the Time Interval between Attempts to Submit a Search [sec] box, specify the interval of time, in seconds, that elapses between attempts to submit a search when the Mascot server is busy.

   The minimum value is 20, and there is no maximum value. The default is 90 seconds.

5. If you are accessing a Mascot server through your own network and security for that server is enforced, enter your user name and password in the boxes beneath the Mascot Server Authentication parameter.

6. If you are accessing a Mascot server through the Web and security for that server is enforced, enter your user name and password in the boxes beneath the Web Server Authentication parameter.

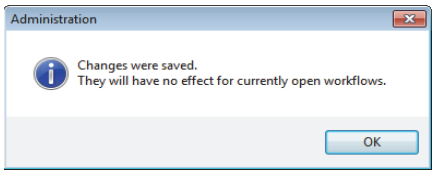
7. Set the Default Confidence Thresholds parameters:

   - Significance High: Calculates the thresholds for high -confidence peptides. The Proteome Discoverer application automatically sets this value to the calculated relaxed significance when it performs a decoy search. The minimum value is 0.0, and the maximum value is 1.0. The default is 0.01.
   - Significance Middle: Calculates the thresholds for medium-confidence peptides. The Proteome Discoverer application automatically sets this value to the calculated relaxed significance when it performs a decoy search. The minimum value is 0.0, and the maximum value is 1.0. The default is 0.05.

8. If you changed any settings, click Apply.

   The message box shown in Figure 12 appears.

Figure 12. Administration message box



9. Click **OK**.

Note Click Reset to return to the default values.

Troubleshooting Failed Mascot Searches

If all your searches with Mascot fail, follow these instructions to locate the problem.

❖ To troubleshoot failed Mascot searches

1. Verify that the Mascot server is running and accessible from the computer that is running the Proteome Discoverer application. For details on how to do this, see "Directing the Proteome Discoverer Application to the Mascot Server Location" on page 25.
2. With the Mascot server is running, verify that it is operating properly by submitting a simple search from the Mascot Web interface. Do one of the following:
   - If the search from the Mascot Web interface is successful, go to step 3.
   - If the search fails, contact your system administrator. There might be a problem with the Mascot server itself.
3. If your Mascot server is operating properly and you can access it from the Proteome Discoverer application, try to perform a very simple search using the Mascot wizard. Do one of the following as applicable:
   - If simple searching fails, there might be a general problem in the interaction between the Proteome Discoverer application and the Mascot server. In this case, file an error report.
   - If you can perform simple Mascot searches, investigate your failing searches more closely:

     Does the search finish successfully on the Mascot server according to the Mascot search log?

     Do the process messages sent to the job queue during the search indicate the problem?
4. If the search problems persist after you take these measures, file an error report.

Starting a New Search by Using the Search Wizards

As mentioned earlier, the quickest way to begin using the Proteome Discoverer application is to define your search parameters using the search wizards. You can access the Sequest HT and Mascot search wizards from the Proteome Discoverer application interface. Use these search wizards to perform basic functions such as setting the search parameters, selecting a database and a search engine, and selecting the chemical modifications that you will use to conduct your search.

To perform the application's more sophisticated operations, such as quantification or using decoy searches to estimate the number of incorrect PSMs that exceed a given threshold, you must use the nodes available in the Workflow Editor. The SEQUEST search engine is only available as a node in the Workflow Editor. You can also access the Sequest HT and Mascot search engines through nodes in the Workflow Editor.

For detailed information about the wizards, see "Search Engines" on page 3.

❖ To prepare to use the search wizards

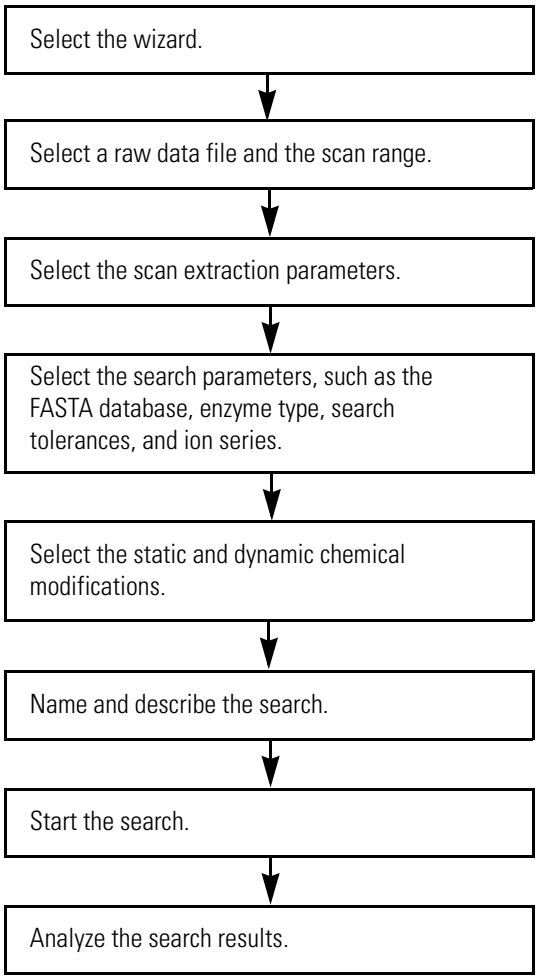
1. Configure the search parameters for Sequest HT or Mascot. See "Configuring the Sequest HT Search Engine" on page 22 and "Configuring the Mascot Search Engine" on page 25, respectively.
2. Download a FASTA file, if necessary, if you have not already done so. See "Adding FASTA Files" on page 128.
3. Make spectrum source files available as RAW, MGF, MZDATA, MZXML, or MZML files.

   The search wizards do not support multiple-spectrum source files. To process multiple-spectrum source files, you must use the Workflow Editor. For detailed information about this process, see "Starting a New Search by Using the Workflow Editor" on page 42.
4. Start the appropriate search wizard. See "Starting a New Search by Using the Search Wizards" on page 29.

You can also set dynamic and static chemical modifications.

Figure 13 shows the general procedure for using the search engine wizards.

Figure 13. The Proteome Discoverer search wizard process



Starting a New Search

The following procedure describes how to search your data by using a search wizard, using Sequest HT as an example. The procedure is very similar for Mascot searches; differences between the two procedures are noted where appropriate.

Note Although the basic procedure for using the Mascot wizard and the Sequest HT wizard is the same, see "Configuring the Mascot Search Engine" on page 25 for information about the unique aspects of conducting Mascot searches.

If you have not selected a FASTA database to search, you must add one before you start a search wizard. For instructions on adding a FASTA file, see "Adding FASTA Files" on page 104.

Note The available FASTA files are registered and available through the Proteome Discoverer application. See "Using FASTA Databases" on page 101.

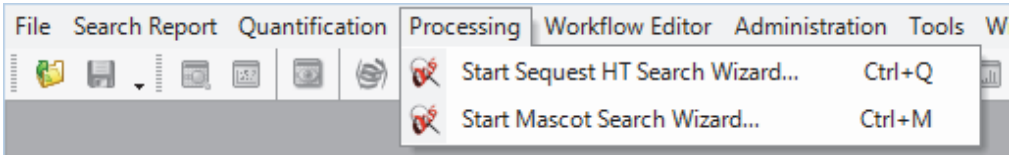
❖ To start a new search using a search wizard

1. (Optional) Open the job queue by choosing **Administration > Show Job Queue** or clicking the **Show Job Queue** icon, .

   You can find more information about the job queue in the Help.

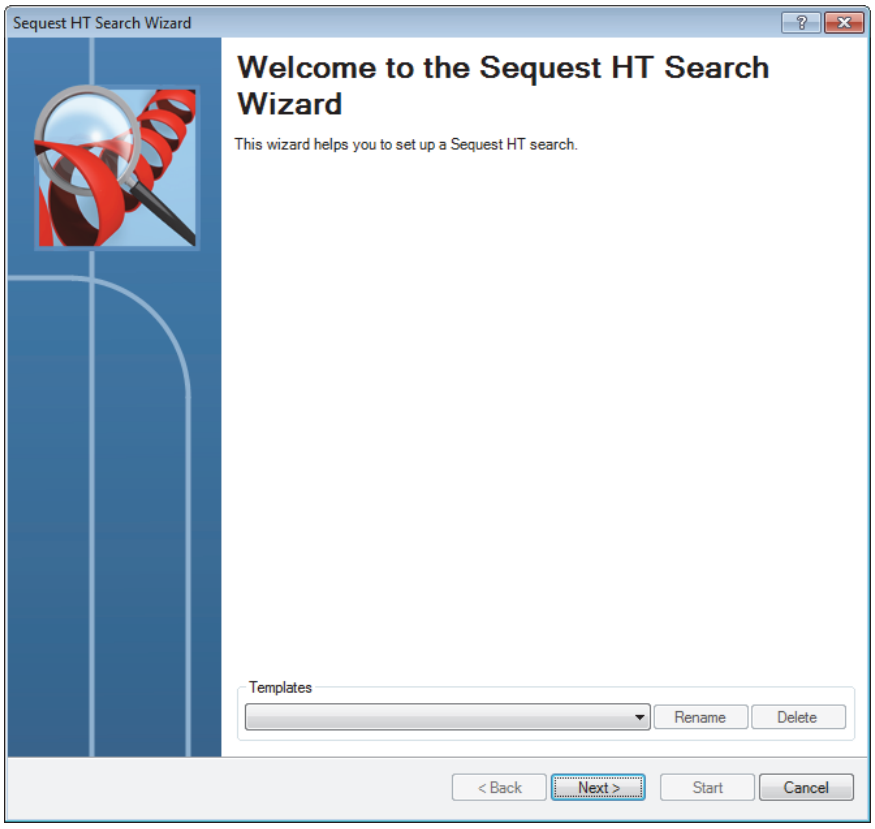
2. Choose **Processing > Start *Wizard_name* Search Wizard**, as shown in Figure 14, or click the appropriate wizard icon in the toolbar: Sequest HT or Mascot .

   Figure 14. Two wizard options in the Processing menu

   The Welcome to the *Wizard_name* Search Wizard page appears, as shown in Figure 15.

   Figure 15. Welcome to the *Wizard_name* Search Wizard page

3. To use a template from a previous search, select it from the Templates list.

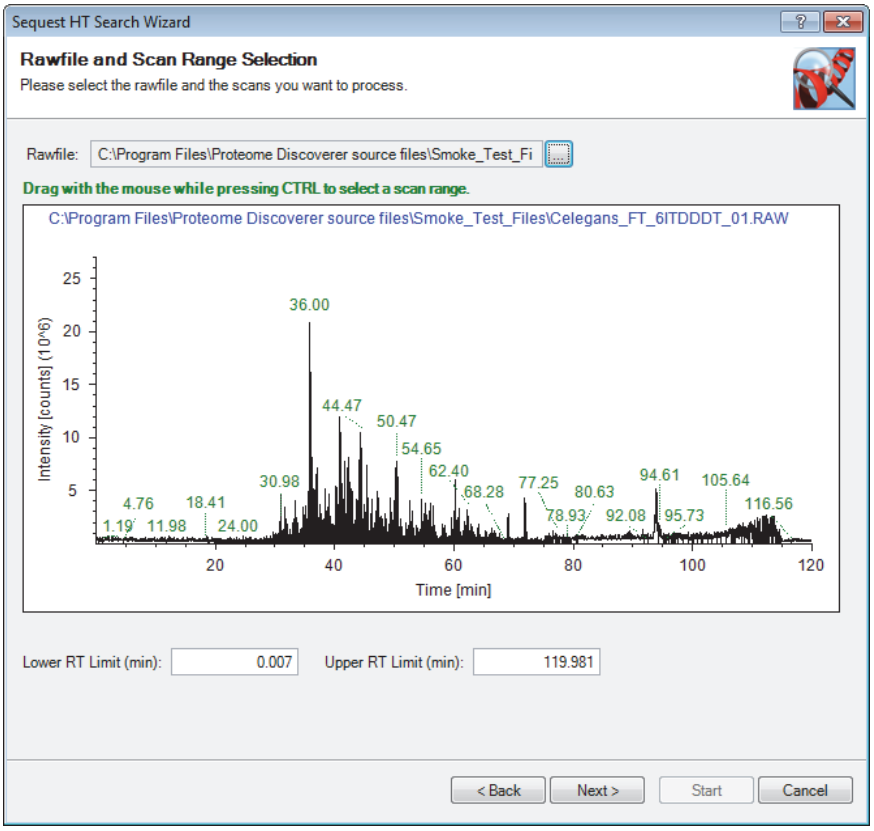
   To give the selected template a new name, click **Rename**, and in the Renaming Template dialog box, type the new name in the New Name box and click **OK**.

   To delete the selected template, click **Delete** and in the confirmation box, click **OK**.

4. Click **Next**.

   The Rawfile and Scan Range Selection page of the wizard opens, as shown in Figure 16.

   Figure 16. Rawfile and Scan Range Selection page



5. Set the basic search parameters:

   a. In the Rawfile box, click the **Browse** button (...) to search for the raw file in the Open Analysis File(s) dialog box.

   > **Note** The Workflow Editor can accept multiple input raw data files, but the search wizards cannot. For information about creating a workflow for multiple input raw data files, see "Starting a New Search by Using the Workflow Editor" on page 42.

   A base peak chromatogram for the raw data file appears on the page, as shown in Figure 16.

   b. Select the range of data to use by choosing either of these methods:

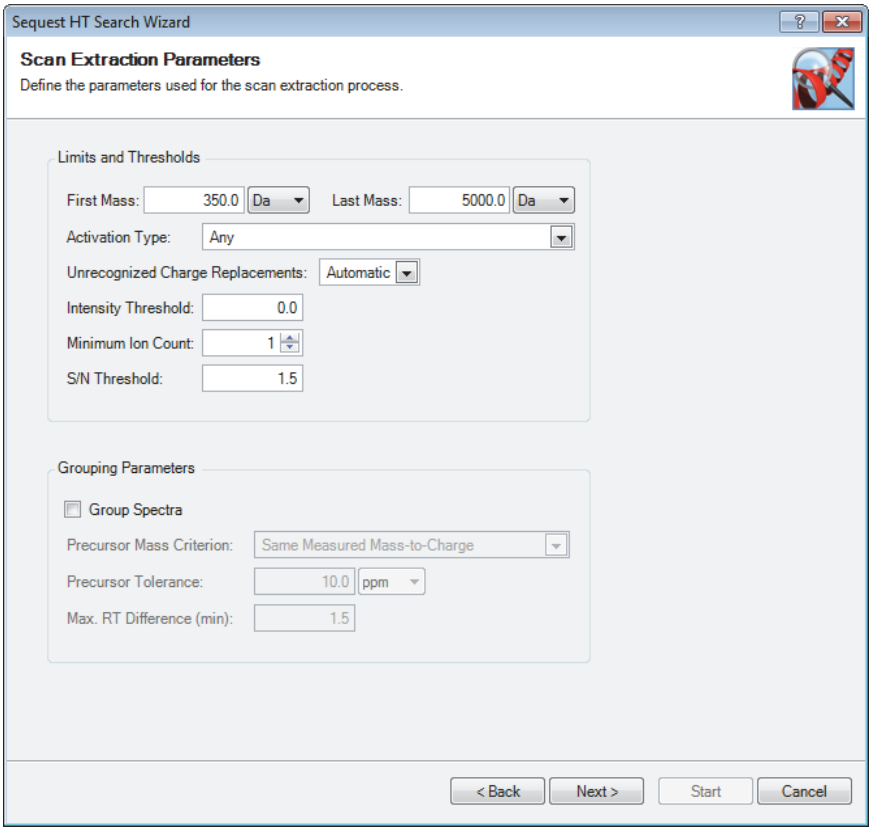
   - Hold down the CTRL key and drag the cursor over the range.
   - Enter the beginning of the range in the Lower RT Limit (min) box. Enter the end of the range in the Upper RT Limit (min) box.

You might want to exclude the first few minutes of collected data in the raw data file because they contain no peptides or exclude the last few minutes because of cleanup at the end of the data collection.

6. Click **Next**.

The Scan Extraction Parameters page appears, as shown in Figure 17.

Figure 17. Scan Extraction Parameters page



7. Set the scan extraction parameters:

a. In the First Mass box, type the mass of the first precursor ion, in daltons. In the Last Mass box, type the mass of the last precursor ion, in daltons.

These two parameters define the range of ion fragments to search for in the database.

b. From the Activation Type list, select the fragmentation method to use to activate the scan:

- CID (Collision-Induced Dissociation)
- MPD (Multi-Photon Dissociation)
- ECD (Electron Capture Dissociation)
- PQD (Pulsed Q Collision-Induced Dissociation)
- ETD (Electron Transfer Dissociation)

- HCD (High-Energy Collision Dissociation)
- Any Activation Type

See "Fragmentation Methods" on page 8 for descriptions of these methods.

The default is Any Activation Type.

c. In the Unrecognized Charge Replacements list, select the charge number of the precursor ions.

From the data in the raw file, the Proteome Discoverer application evaluates the spectrum and uses an algorithm to determine the charge state of the spectrum. It cannot calculate the mass without knowing the charge state of the spectrum. If the algorithm cannot determine the charge state of the evaluated spectrum, the application assigns the charge state that you select to the spectrum. You can assign the following charge number:

- Automatic: Assigns a charge number of +2 and +3 to the spectrum.
- 1 through 8: Assigns a charge number of from 1 through 8 to the spectrum.

The default is Automatic.

d. In the Intensity Threshold box, enter an intensity value below which to filter out ions.

The Proteome Discoverer application filters out low-intensity ions, which are ions that are most likely chemical noise and serve only to slow down the analysis without improving the results.

The default is 0.0.

e. In the Minimum Ion Count box, enter a value for the minimum ion count or use the increment or decrement buttons.

The minimum ion count is the minimum number of ions that must be present in an MS/MS spectrum for it to be included in a search.

The default is 1.

f. In the S/N Threshold box, enter a value for the signal-to-noise threshold setting.

This setting specifies the intensity of the signal to the intensity of the background noise. It filters out low-intensity ions that function as noise.

The default is 3.0.

g. (Optional) Select the **Group Spectra** check box.

The rest of the boxes in the Grouping Parameters area become available.

In the Grouping Parameters area, you can set grouping parameters to group similar spectra in the raw data file into a single spectrum.

Grouping spectra speeds up the analysis. The application evaluates an ion only once rather than every time it is observed within the given retention-time limits.

h. In the Precursor Mass Criterion list, select the criteria for grouping. You can select either of these settings:

- Same Measured Mass-to-Charge: Groups spectra according to the mass-to-charge ratio (*m/z*) of the precursor ion.
- Same Singly Charged Mass: Groups all charge states with the same singly charged precursor mass. For example, this option groups +2 and +3 ions for the same peptide because they have the same singly charged parent.

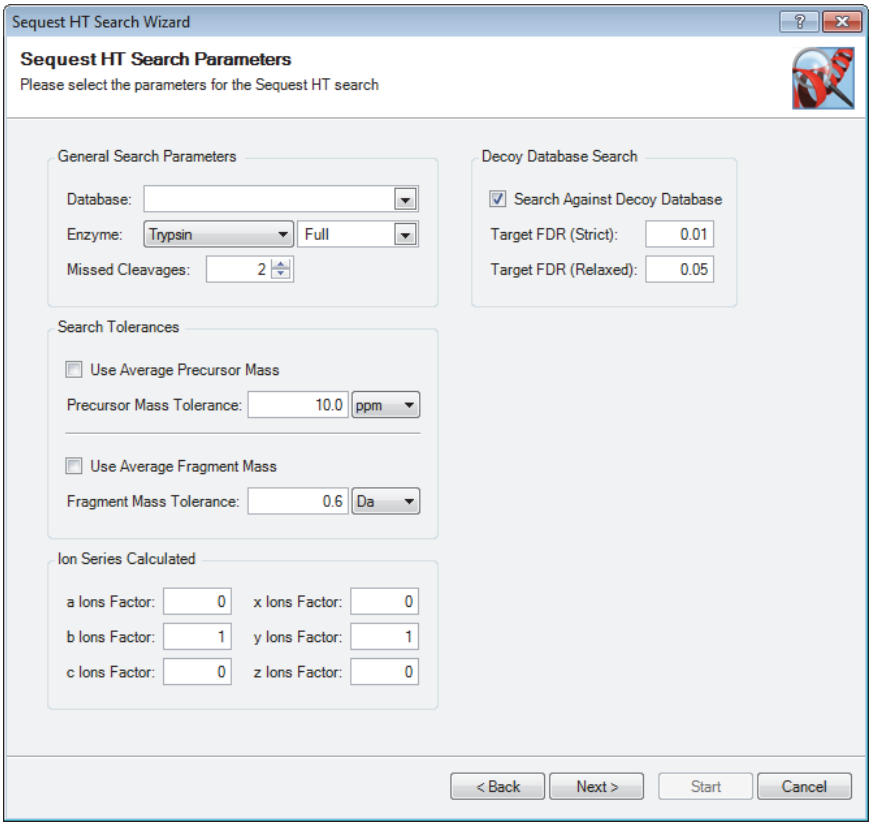
i. In the Precursor Tolerance box, type the range of the precursor tolerance, in daltons (Da), milli-mass units (mmu), or parts per million (ppm). For example, if the mass-to-charge ratio of a spectrum is 100.0001 Da and the tolerance is 2 Da, all the spectra with masses in the range of 100.0001 plus or minus 2 Da are valid mass candidates.

j. In the Max. RT Difference (min) box, enter the maximum retention time, in minutes. Retention time is the time in the mass chromatogram when any particular precursor ion is observed. This parameter limits the maximum retention-time difference between scans to be considered for grouping. In general, if the precursor masses of spectra are within the tolerance and the maximum retention time window, they are grouped into a single spectrum. The default is 1.5.

8. Click **Next**.

The Sequest HT Search Parameters page appears, as shown in Figure 18.

Figure 18. Sequest HT Search Parameters page



9. Set the Sequest HT search parameters:
   a. In the Database list in the General Search Parameters area, select one of the FASTA databases that you registered.
   b. In the Enzyme list, select the enzyme used for digestion and indicate whether the cleavage is full or partial.

      The default enzyme is trypsin, and the default cleavage is Full.
   c. In the Missed Cleavages box, use the increment and decrement buttons to specify the maximum number of internal cleavage sites per peptide fragment that is acceptable for an enzyme to miss when cleaving peptides during digestion.

      Normally, the digestion time is too short to enable the enzyme to cleave the protein at all allowed positions, so you must specify the number of missed positions in one resulting peptide fragment where the enzyme could cleave but did not. The minimum value is 0, and the maximum value is 12. The default is 2.

Note The following parameters are also available in the General Search parameters in Mascot:

- Instrument: Specifies the instrument used to process the data in the raw data file.
- Taxonomy: Specifies the category of organism in the Linnaean biological classification system from which the sample was drawn.

In the Search Tolerances area, specify the precursor mass search tolerance.

d. Select the **Use Average Precursor Mass** option to use the average mass for matching the precursor.

e. In the Precursor Mass Tolerance box, specify the precursor mass tolerance value used for finding peptide candidates, in daltons (Da), milli-mass units (mmu), or parts per million (ppm).

- For daltons, the minimum value is 0.0001 and the maximum value is 5.0.
- For milli-mass units, the minimum value is 0.1, and the maximum value is 5000.
- For parts per million, the minimum value is 0.01, and the maximum value is 5000. The default is 10.0.

In the Search Tolerances area, specify the fragment mass search tolerance.

f. Select the **Use Average Fragment Masses** option to use the average mass for matching the fragments.

g. In the Fragment Mass Tolerance box, specify the mass tolerance value used for matching fragment peaks, in daltons (Da) or milli-mass units (mmu).

- For daltons, the minimum value is 0.0001, and the maximum is 2.0. The default is 0.8.
- For milli-mass units, the minimum value is 0.1, and the maximum value is 2000.

h. In the Ion Series Calculated area, specify the ion factors for a, b, c, x, y, and z ions for your experiment type.

You can use a range of 0 through 1.0 for all ion factors. For CID, HCD, and PQD activation types, use b and y ion factors. For ETD and ECD activation types, use c, y, and z ion factors.

Note The Ion Series Calculated area does not appear in the Mascot wizard.

i. (Optional) Set up a decoy database by selecting the **Search Against Decoy Database** check box and setting the false discovery rate (FDR) parameters. For detailed information about this procedure, see "Calculating False Discovery Rates" on page 186.

A decoy database gives a probability value to identifiers and the percentage of false discoveries that you can expect, typically 1 percent.

Note You must select the Search Against Decoy Database check box to see peptide confidence determined by FDR.

- To specify a strict target false discovery rate for peptide matches with high confidence, type a value of **0.0** through **1.0** in the Target FDR (Strict) box.

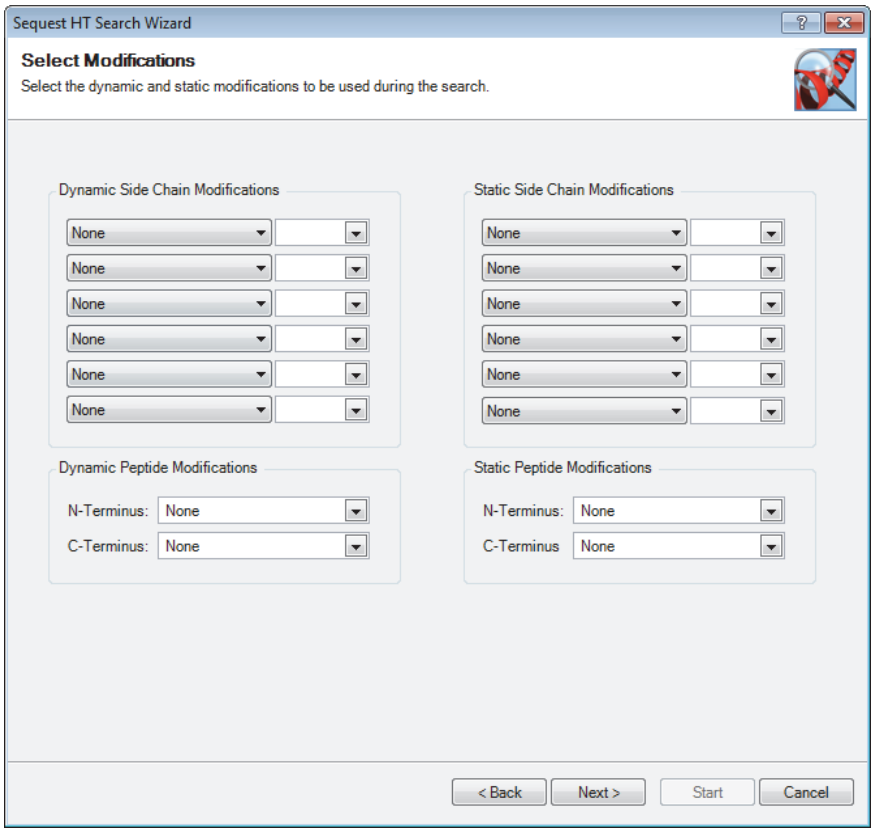
  The default is 0.01 (1 percent FDR).

- To specify a relaxed target false discovery rate for peptide matches with moderate confidence, type a value of **0.0** through **1.0** in the Target FDR (Relaxed) box. The default is 0.05 (5 percent FDR).

j. Click **Next**.

The Select Modifications page appears, as shown in Figure 19.

Figure 19. Select Modifications page



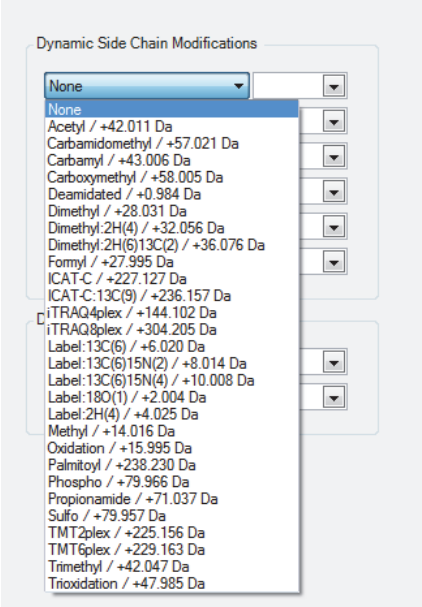
10. Specify which modifications you want the search algorithm to include during its in-silico digestion of the protein database.

    For a description of static and dynamic modifications, see "Updating Chemical Modifications" on page 141.

a. If you are searching for dynamic modifications, select the modifications and the amino acids on which they can occur in the Dynamic Side Chain Modifications area. In the boxes on the left, select the modifications. In the boxes on the right, select the amino acids on which the modifications occur.

In the Sequest HT wizard, delta masses appear next to the names of the modifications in the modification lists to clearly identify the modification, as shown in Figure 20.

Figure 20. Modifications with identifying delta masses



Note In the Mascot wizard, the Dynamic Modifications area replaces both the Dynamic Side Chain Modifications and Dynamic Peptide Modifications areas. You set these modifications on the Mascot server.

The Mascot wizard does not identify by delta masses the modifications that appear on the modification lists as the Sequest HT wizard does.

b. If you are searching for static modifications, select the modifications and the amino acids on which they can occur in the Static Side Chain Modifications area. In the boxes on the left, select the modifications. In the boxes on the right, select the amino acids on which the modifications occur.

Note In the Mascot wizard, the Static Modifications area replaces both the Static Side Chain Modifications and Static Peptide Modifications areas. You set these modifications on the Mascot server.

The modifications that appear on the modification lists in the Mascot wizard are not identified by delta masses as they are in the Sequest HT wizard.

c. In the N-Terminus list in the Dynamic Peptide Modifications area, select the dynamic modification that occurs on the N terminus of the peptide.

d. In the C-Terminus list in the Dynamic Peptide Modifications area, select the dynamic modification that occurs on the C terminus of the peptide.

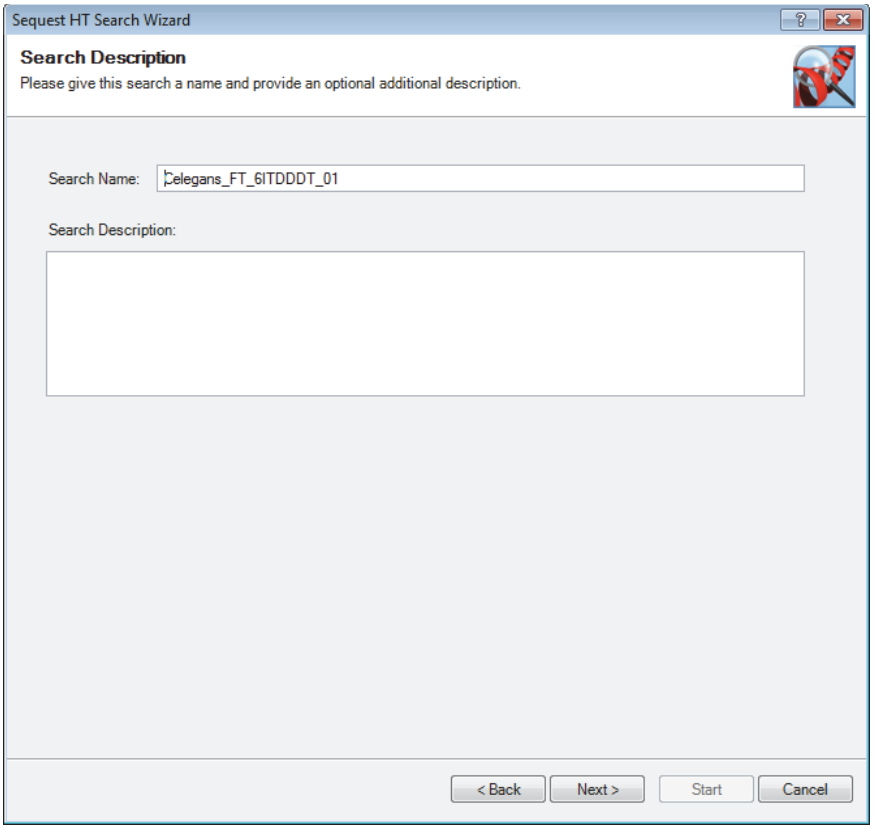
e. In the N-Terminus list in the Static Peptide Modifications area, select the static modification that occurs on the N terminus of the peptide.

f. In the C-Terminus list in the Static Peptide Modifications area, select the static modification that occurs on the C terminus of the peptide.

g. Click **Next**.

The Search Description page opens, as shown in Figure 21.

Figure 21. Search Description page



11. Give your search a name and a brief description:

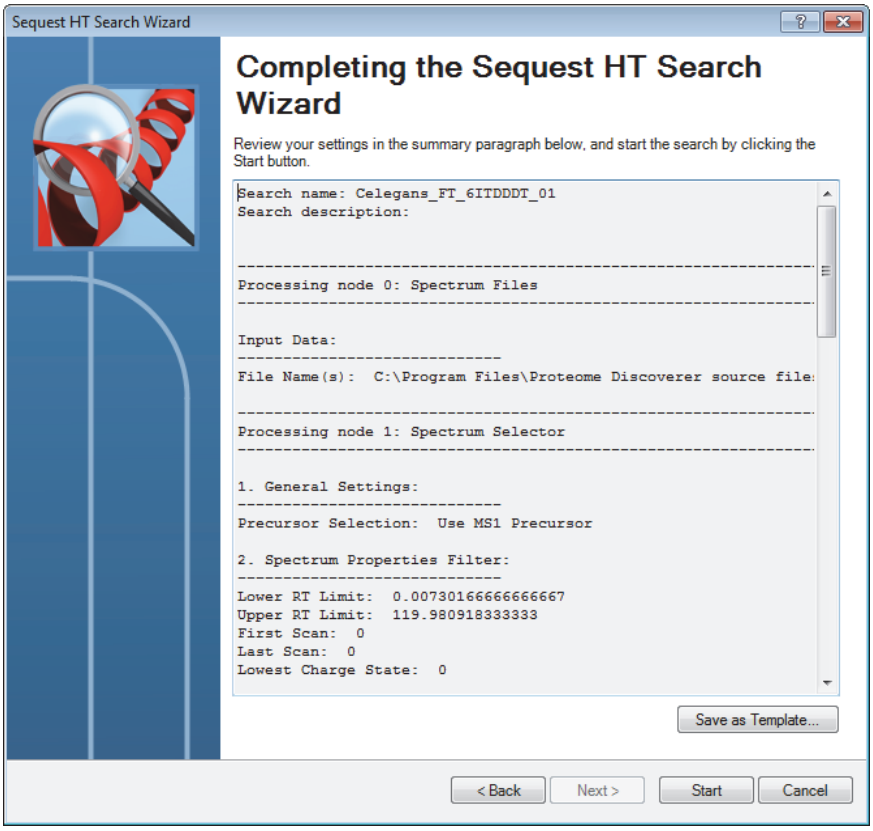
a. In the Search Name box, type a name for your search.

b. In the Search Description box, type a brief description of the search.

c. Click **Next**.

The Completing the *Wizard_name* Search Wizard page appears, as shown in Figure 22.

Figure 22. Completing the *Wizard_name* Search Wizard page

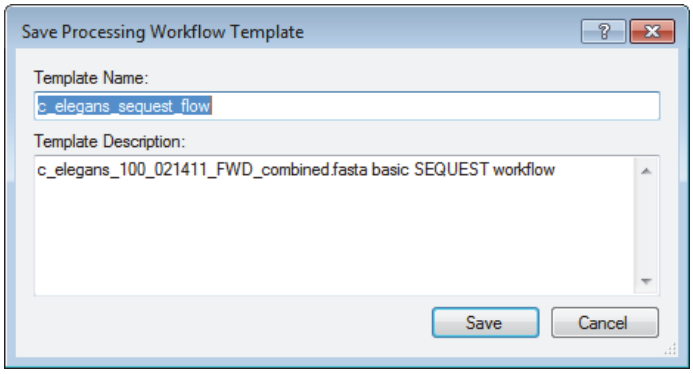


12. (Optional) Save the search parameters as a template that you can use in the future:

    a. Click **Save as Template**.

    The Save Processing Workflow Template dialog box appears, as shown in Figure 23.

    Figure 23. Save Processing Workflow Template dialog box



    b. In the Template Name box, give the search workflow a name.

    The Template Description box reflects the description that you entered on the Search Description page, shown in Figure 21 on page 40.

    c. Click **Save**.

13. Click **Finish** on the Completing the *Wizard_name* Search Wizard page to start the search.

    You can monitor the progress of the search in the job queue. Refer to the Help.

14. Choose **File > Open Report** to display your search results. Refer to the Help.
    a. Filter and sort your results. See "Filtering the Search Results" on page 154.
    b. Use different views to aid in your analysis. Refer to the Help.

Starting a New Search by Using the Workflow Editor

You can create a customized search by using the Proteome Discoverer Workflow Editor instead of the search wizards. The Workflow Editor is a flexible and complex tool that you can use to create customized data-processing workflows. Instead of using the standard wizards available through the Processing menu, you can develop a workflow specific to your needs. The Workflow Editor searches with multiple algorithms and merges results from multiple fragmentation methods. It also provides great flexibility in creating custom search results. Unlike the search wizards, the Workflow Editor can accept multiple input raw files.

You can create a reusable processing workflow template by saving your design to load and use at another time. A unique workflow gives you the ability to set parameters that are normally static settings in the wizard or use a function that would not normally be available, such as deconvoluting the precursor ions for all high-mass-accuracy data or exporting a spectrum.

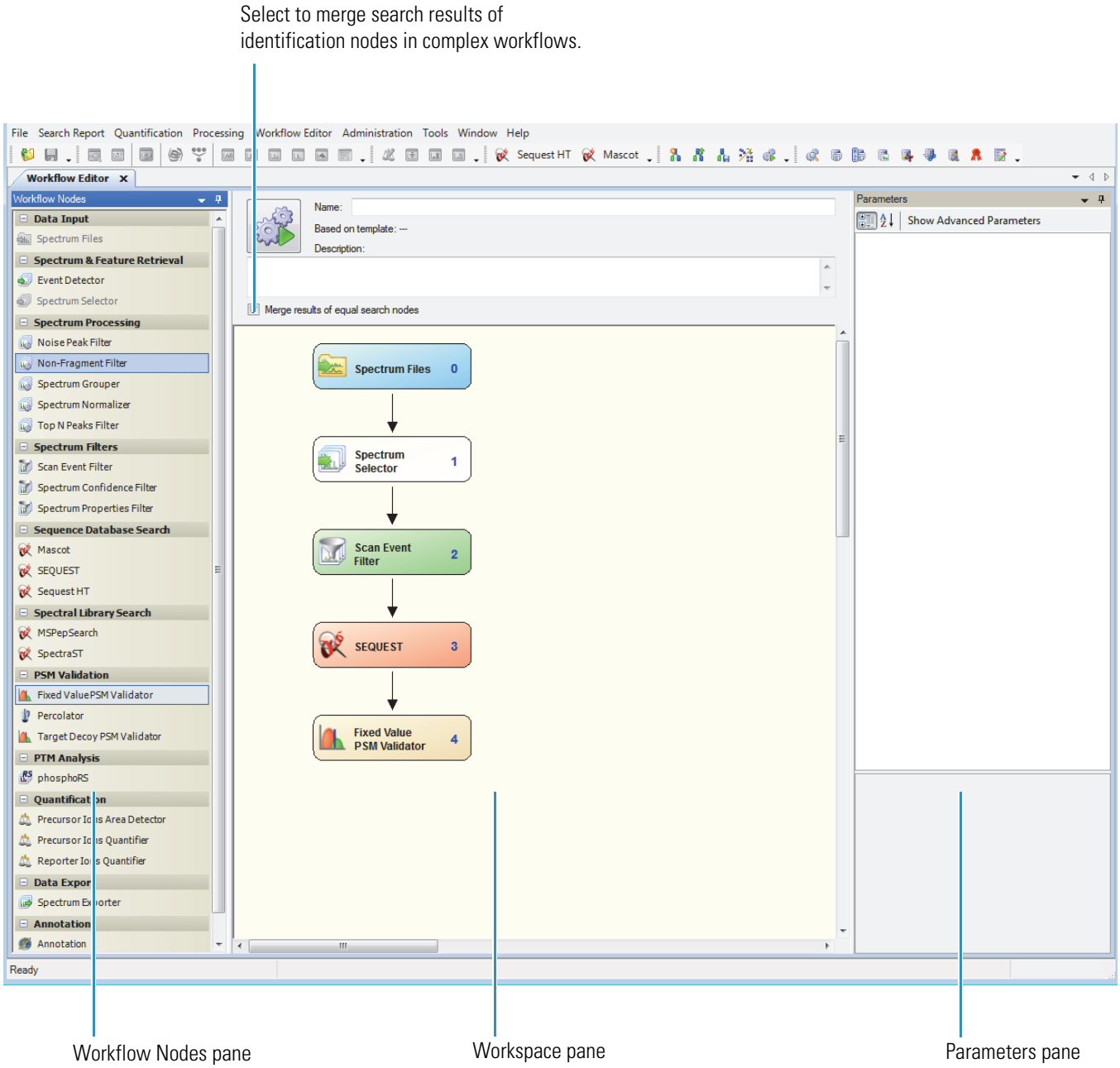
The workflow is the layout of processing nodes, or workflow steps, which you then submit to process your data. The nodes are like building blocks that you can use to create a unique search sequence. You can use them to define your own search parameter tolerances and criteria.

WARNING As a prerequisite to using the Proteome Discoverer Workflow Editor, you must know how each workflow node functions. If you do not understand the function (or interconnectivity) of these nodes, you can potentially build a sequence that creates bad results and makes no analytical sense. For a detailed description of these nodes, refer to the Help.

You can access the Workflow Editor through the Workflow Editor menu in the Proteome Discoverer application or through the Workflow Editor icons on the main toolbar. After you choose a menu command or click an icon, the application opens a Workflow Editor page in the main window.

The three-pane layout of the Workflow Editor page provides a pane for node selections, a workspace for placing the nodes, and a pane where you can choose parameters for each node, as shown in Figure 24.

Figure 24. Workflow Editor workspace



To create a workflow, see "Creating a Search Workflow" on page 44.

Before Creating a Workflow

As with the search engines, follow these steps before using the Workflow Editor to create a workflow:

- Download a FASTA file, if necessary, if you have not already done so. See "Adding FASTA Files" on page 104.
- Make spectrum source files available as RAW, MGF, MZDATA, MZXML, or MZML files.

Creating a Search Workflow

You can use the following procedure to process one raw file from one sample, multiple raw files from one sample, or multiple raw files from multiple samples. For additional details on creating a workflow for multiple raw files from one sample, see "Creating a Search Workflow for Multiple Raw Files from the Same Sample" on page 53.

For a demonstration showing how to create a new workflow, see "Demonstrating How to Create a Workflow" on page 51.

❖ To create a new workflow

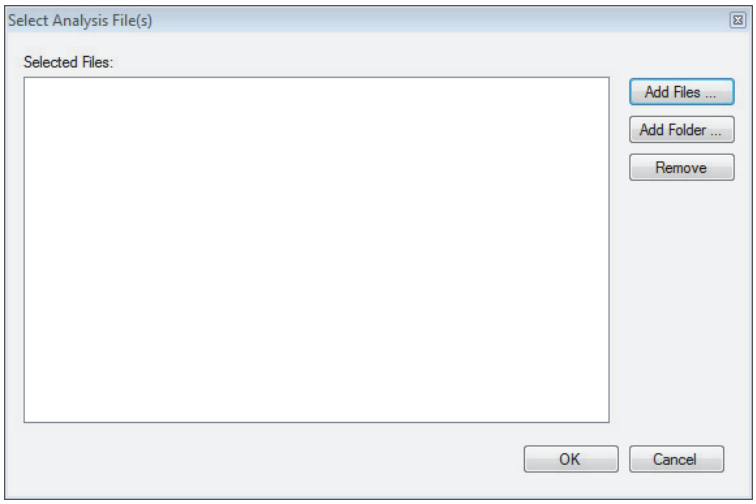
1. Choose **Workflow Editor > New Workflow** or click the **New Workflow** icon,  .

   The Workflow Editor opens, as shown in Figure 24 on page 43.

2. In the Name box in the workspace pane, type a name for the workflow.
3. (Optional) In the Description box, type a description of the workflow.
4. To perform two searches using the same search engine node and then merge the search results in the output MSF file, select the **Merge Results of Equal Search Nodes** check box.
5. From the Data Input area of the Workflow Nodes pane, drag the **Spectrum Files** node to the workspace pane.
6. Select the **Spectrum Files** node if it is not already selected.
7. Select the data input file:
   a. In the Input Data section at the top right of the Parameters pane, click the File Name(s) row (see Figure 24 on page 43).
   b. Click the **Browse** button (...) in that row.

      The Select Analysis File(s) dialog box appears, as shown in Figure 25.

Figure 25. Select Analysis File(s) dialog box



c. Click **Add Files** to open the Add Analysis File(s) dialog box.

d. Browse to the location of the data input file, select the file, and click **Open**.

e. Click **OK** to close the Select Analysis File dialog box.

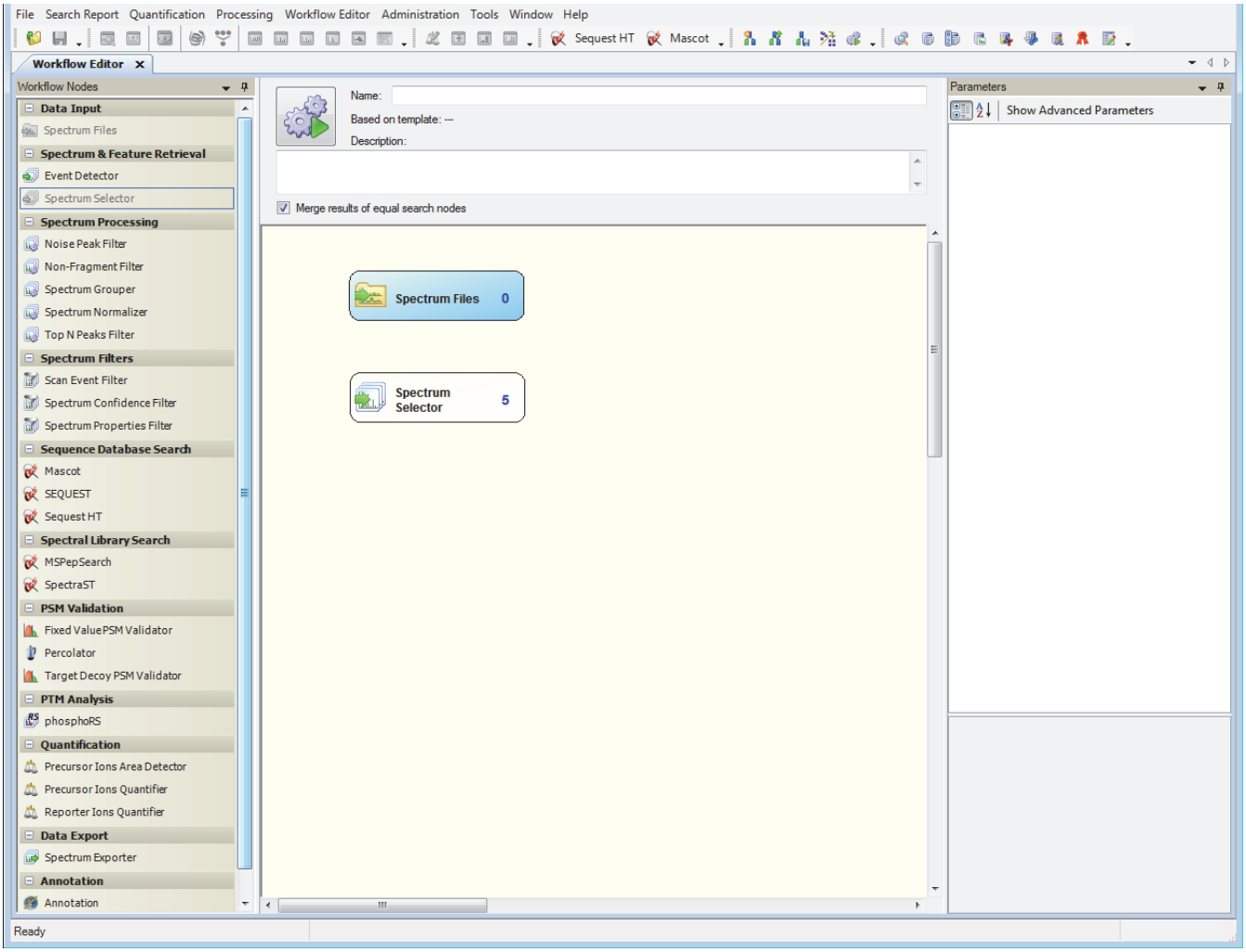
8. If you selected the Spectrum Files node in step 5, drag the **Spectrum Selector** node to the workspace and place it beneath the Spectrum Files node.

   Figure 26 shows the addition of the Spectrum Files and Spectrum Selector nodes to the workspace. Selecting the Spectrum Selector node in the workspace pane displays the available parameters for that node in the right pane.

   The numbers that appear on each workflow node indicate the order in which the Proteome Discoverer application processes the nodes.

   Note You can set the Spectrum Selector node to select which precursor mass to use for a given MS^n scan, such as choosing the precursor from the parent scan.

Figure 26. Spectrum Files and Spectrum Selector nodes added to a workflow



9. Depending on your data needs, drag the appropriate nodes from the Workflow Nodes pane to the workspace pane.

   For a description of the nodes that you can select, refer to the Help. The nodes in each section of the Workflow Nodes pane appear in unique colors; for example, the Data Input nodes are blue, the quantification nodes are pink, and the Spectrum Processing nodes are yellow.

   When you use any of the search engine nodes in the workflow, you must attach the Fixed Value PSM Validator or the Percolator node to it.

   You can also add third-party nodes that are in your installation that are not documented in this manual. For further information on those nodes, consult the third-party documentation.

You cannot drag workflow nodes into the workspace pane that cannot logically be added at that point. For example, if you add the Target Decoy PSM Validator node, you cannot connect it to the Percolator node.

10. Organize the nodes to reflect a procedural order from top to bottom so that the Spectrum Files node remains on top as the root node.

    Delete a node by selecting the node in the workspace pane and pressing DELETE or by right-clicking the node and choosing **Cut** (or **CTRL+X**) from the shortcut menu.

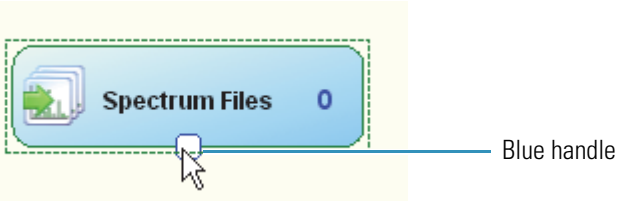
    You can use the Cut command and the Paste (or CTRL+P) command on the shortcut menu to move a node to another place in the workspace or use the Copy (or CTRL+C) and Paste commands to duplicate a node in the workspace.

    You can paste copied or cut nodes into other workflows.

11. Connect the nodes:

    a. Click the top node so that a blue handle is activated at the bottom center of the node, as shown in Figure 27.

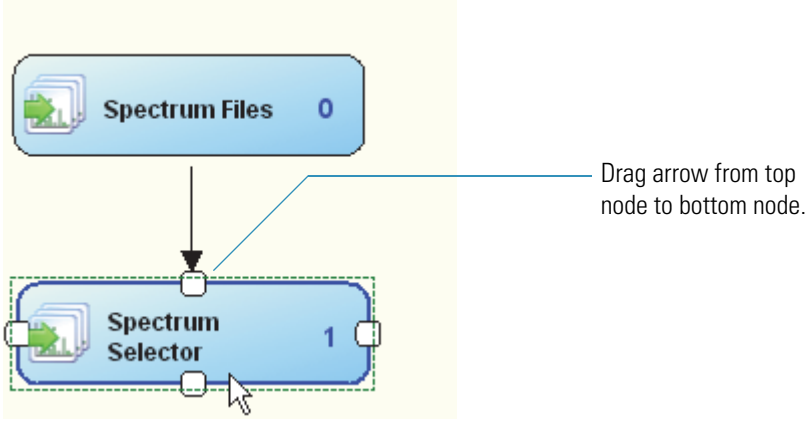
    Figure 27. Activated node example

    Joining the nodes together creates a sequence of steps for the Proteome Discoverer application to follow.

    b. Drag the blue handle down to the top-center of the node below it, as shown in Figure 28.

    Figure 28. Joining two nodes

IMPORTANT If the next node appears with a red edge at this point, you cannot connect to the previous node.

If the Workflow Editor prevents you from connecting two nodes, the workflow is erroneous.

c. Link all the nodes to develop a workflow.

12. After you join all your chosen nodes, align them by choosing **Workflow Editor > Auto Layout**, or clicking the **Auto Layout** icon ( ), or right-clicking a node and choosing **Auto Layout** from the shortcut menu.

13. (Optional) You can renumber the workflow nodes in the workflow in consecutive order by choosing **Workflow Editor > Auto Number**.

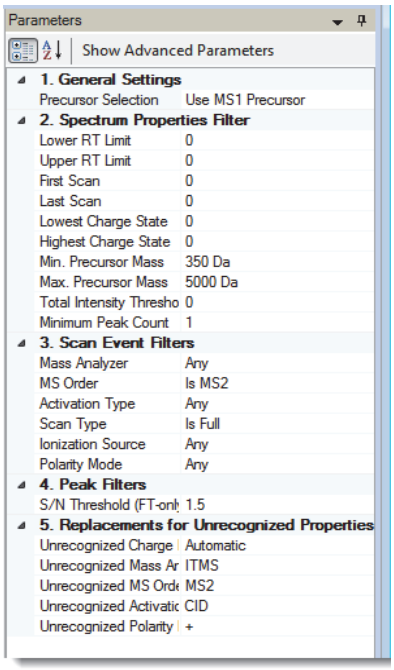
14. Set the parameters for each node in the workspace pane:

a. Click the node to activate its functions.

The available parameters for the node appear in the Parameters pane, as shown in the example for the Spectrum Selector node in Figure 29.

Note The same options are available in the search wizards.

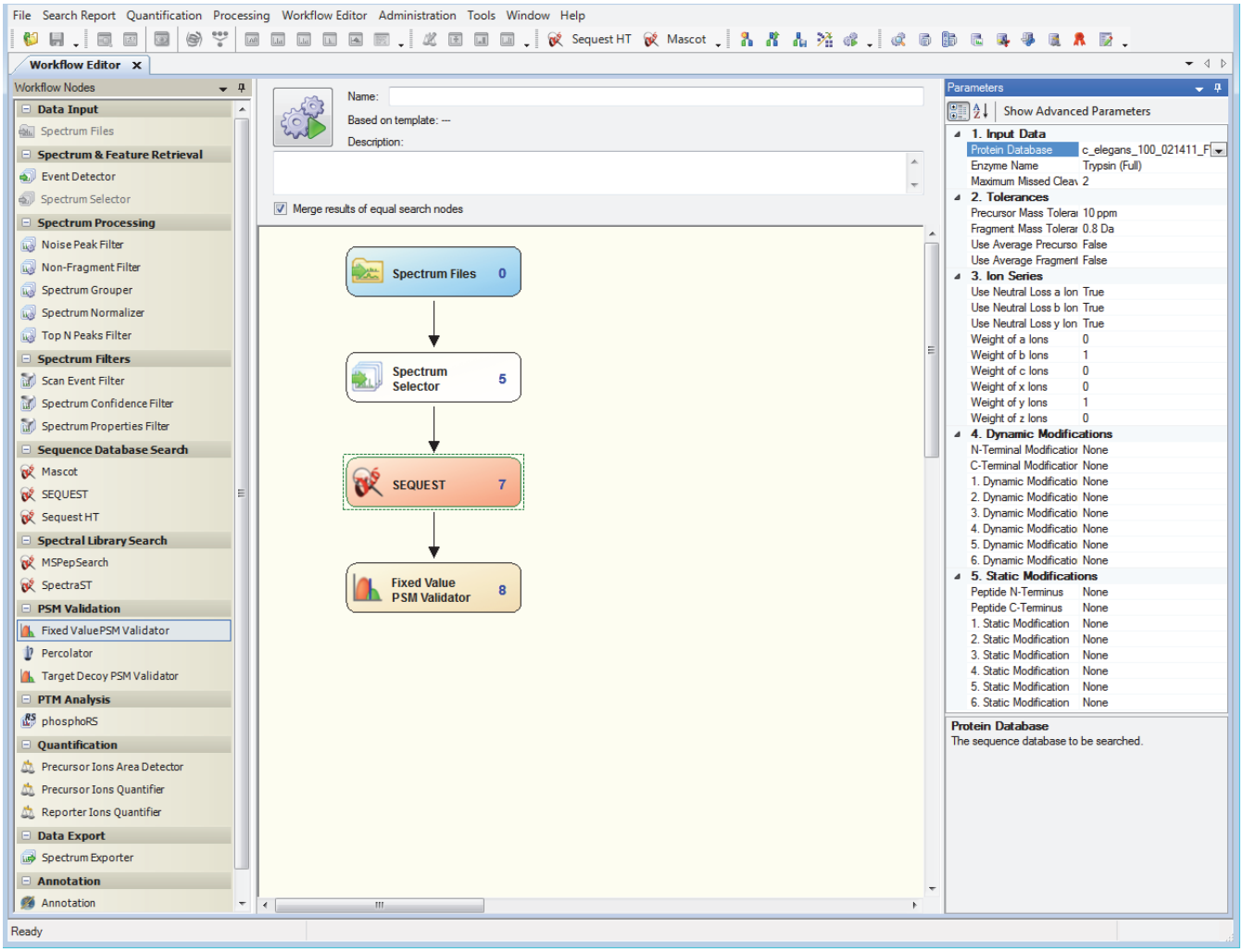
Figure 29. Spectrum Selector node parameters in the Parameters pane



b. Set the node's parameters. Complete this step for each node that you select.

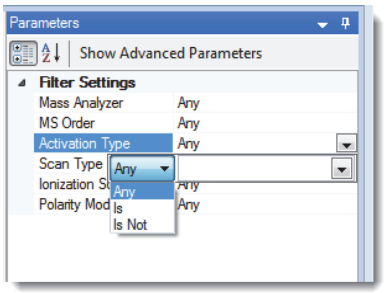
Figure 30 shows the parameters set for the SEQUEST node.

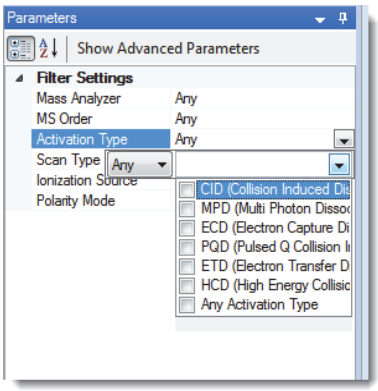
Figure 30. Setting parameters for the workflow



When you click some parameters, two lists appear, as shown in Figure 31.

Figure 31. Settings and filters





The list on the right gives the activation types available. You can apply a filter option on the left to the setting that you select in the list on the right. The list on the left consists of three options:

- Is: Applies the setting selected in the list on the right. In the example in Figure 31, "Is" means that the workflow processes data from the CID activation type.
- Is Not: Applies all settings in the list on the right except the selected setting. In the example in Figure 31, "Is Not" means that the workflow processes data from all activation types except CID.
- Any: Applies all settings available for the parameter in the list on the right. In the example in Figure 31, "Any" means that the workflow processes data from any activation type available in the list on the right.

Any is the default.

You can filter input data before searching the database to remove lower-quality spectral peak lists from your analysis. This step might help to decrease search times and false positive identifications. The Spectrum Filters area of the Workflow Nodes pane provides three types of spectrum filters to use for your search. Use these pre-analysis filters to streamline your search results. For information about these nodes, refer to the Help.

Use the Scan Event Filter node for high-mass-accuracy data, such as Mascot analysis and Sequest analysis of mixed fragmentation-mode-type data (CID and ETD). It can filter information according to fragmentation type, mass analyzer identity, and other parameters. Refer to the Help for information about the Scan Event Filter node.

❖ To save the workflow as a template

1. Choose **Workflow Editor > Save as Template** or click the **Save As Template** icon,  .

   (To save the workflow in XML format, see "Saving a Workflow as an XML Template" on page 66.)

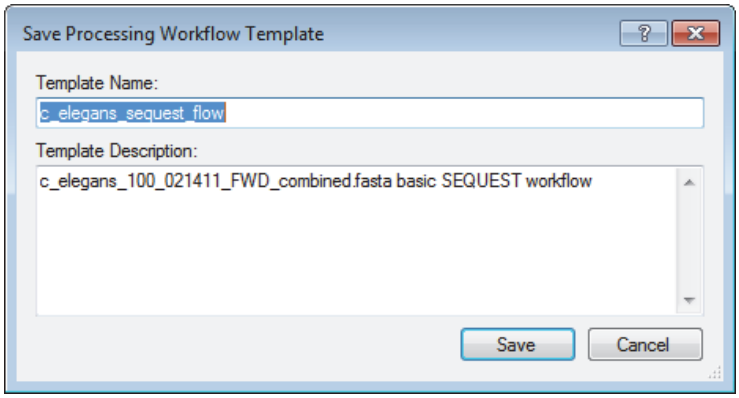
2. In the Save Processing Workflow Template dialog box, shown in Figure 32, do the following:

   a. Type a template name in the Template Name box.

   b. Type a description in the Template Description box.

   c. Click **Save**.

Figure 32. Save Processing Workflow Template dialog box



❖ **To perform the search**

1. Choose **Workflow Editor > Start Workflow** or click the **Start Workflow** icon, .

   The job queue appears, showing the status of your search.

2. Use the job queue to check the status of your search as the search progresses.

   For information about the job queue, refer to the Help.

3. Choose **File > Open Report** to display your search results. Refer to the Help.

   a. Filter and sort your results. See "Filtering Data" on page 153.

   b. Use different views to aid in your analysis. Refer to the Help.

Demonstrating How to Create a Workflow

The following demonstration shows you how to set up a workflow. In this example, a sample containing a trypsin digest of Caenorhabditis elegans, a nematodal worm, was submitted to an LTQ Orbitrap XL mass spectrometer at a resolution of 60 000 for MS/MS processing, using both the ETD and CID fragmentation methods for better confidence. The example searches a FASTA database to determine how the worm's proteins are expressed.

❖ Click the button below to view the demonstration.

Creating a Search Workflow for Multiple Raw Files from the Same Sample

Multidimensional Protein Identification Technology (MudPIT) experiments investigate complex proteomes by applying multidimensional chromatography to the samples before acquisition in the mass spectrometer. Typically, this process results in several dozen or even a few hundred fractions that are separately analyzed by LC-MS, resulting in one raw file per sample fraction. Analyzing gel slices or performing in-depth follow-up acquisitions also results in multiple fractions. Because all these fractions belong to the same sample, the Proteome Discoverer application can process all raw files from these fractions as one contiguous input file and generates only one result file.

You have two ways to search for sample fractions:

- Search the sample fractions one at a time and open them in a multiconsensus report.

  This method is appropriate for searching multiple samples. When you open a multiconsensus report from several searches, the Proteome Discoverer application does not calculate a combined protein score, and it orders the proteins by their coverage.

- Search the fractions all at one time in MudPIT.

  To search the fractions of only one sample, use MudPIT. In this mode, the Proteome Discoverer application searches all fractions as one logical sample and creates a single MSF result file. It automatically merges all identified peptides and proteins from all fractions and creates a single combined score for every protein that includes all peptides identified from the different fractions.

  Opening a MudPIT report is faster and consumes less memory than combining separate reports into a multiconsensus report. For example, if the Proteome Discoverer application identifies a protein in every fraction and opens all fractions into a multiconsensus report, a copy of the same protein resides in memory for every fraction that the protein was identified in. It must merge the proteins into an additional protein instance that it displays in the multiconsensus report, slowing performance and consuming memory unnecessarily. However, if you searched the fractions in MudPIT mode, the proteins are already merged from the different fractions, and the Proteome Discoverer application only needs to load the identified merged proteins.

The following procedure describes how to create a workflow for multiple raw files from the same sample. This workflow is basically the same as that given in "Creating a Search Workflow" on page 44, except that you select multiple files to load with the File Name(s) parameter of the Spectrum Files node.

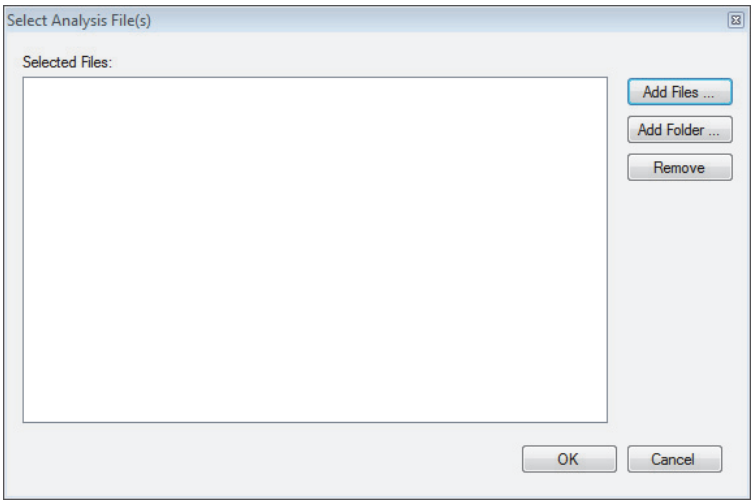
Note The following method is not appropriate for batch-processing different sample data files because the process generates a single result file.

❖ To load multiple raw files from the same sample

1. In the Workflow Editor, drag the **Spectrum Files** node from the Data Input section of the Workflow Nodes pane to the workspace pane.
2. Select the **Spectrum Files** node.
3. In the Parameters pane, click **File Name(s)**, and click the **Browse** button (...).

   The Select Analysis File(s) dialog box appears, as shown in Figure 33.

   Figure 33. Select Analysis File(s) dialog box



4. To add new input files, click **Add Files**, and in the Add Analysis File(s) dialog box, select the raw data files to load and click **Open**.

   –or–

   To add all the raw data files in a specific folder, click **Add Folder**, and in the Browse for File dialog box, click **OK**.

   To remove a file or folder from the Selected Files area of the dialog box, select the file and click **Remove**.

5. In the Select Analysis File(s) dialog box, click **OK**.
6. Drag the Spectrum Selector node to the workspace pane beneath the Spectrum Files node, and continue with the process of creating a workflow, as described in "Creating a Search Workflow" on page 44.
7. Choose **Workflow Editor > Start Workflow** to start the workflow.

You can use the Proteome Discoverer Daemon utility to monitor multiple searches on multiple raw data files. For information about this tool, see "Using the Proteome Discoverer Daemon Utility" on page 69.

Creating a Quantification Workflow

To perform quantification, you must run a quantification workflow. A quantification workflow is a search workflow that includes one of three quantification nodes found in the quantification section of the Workflow Nodes pane of the Workflow Editor. Table 1 lists these nodes and where you can obtain information about creating a quantification workflow for each.

Table 1. Quantification nodes

Quantification node	Use	For more information
Precursor Ions Quantifier node	For precursor ion quantification (for example, SILAC)	See "Performing Precursor Ion Quantification" on page 243.
Reporter Ions Quantifier node	For reporter ion quantification (for example, iTRAQ and TMT)	See "Performing Reporter Ion Quantification" on page 249.
Precursor Ions Area Detector node	For peak area calculation quantification	See "Performing Peak Area Calculation Quantification" on page 259.

You must attach the selected quantification node directly to the Event Detector node. For information about the parameters that you can set for the quantification nodes, see "General Configuration Parameters" on page 597.

Creating an Annotation Workflow

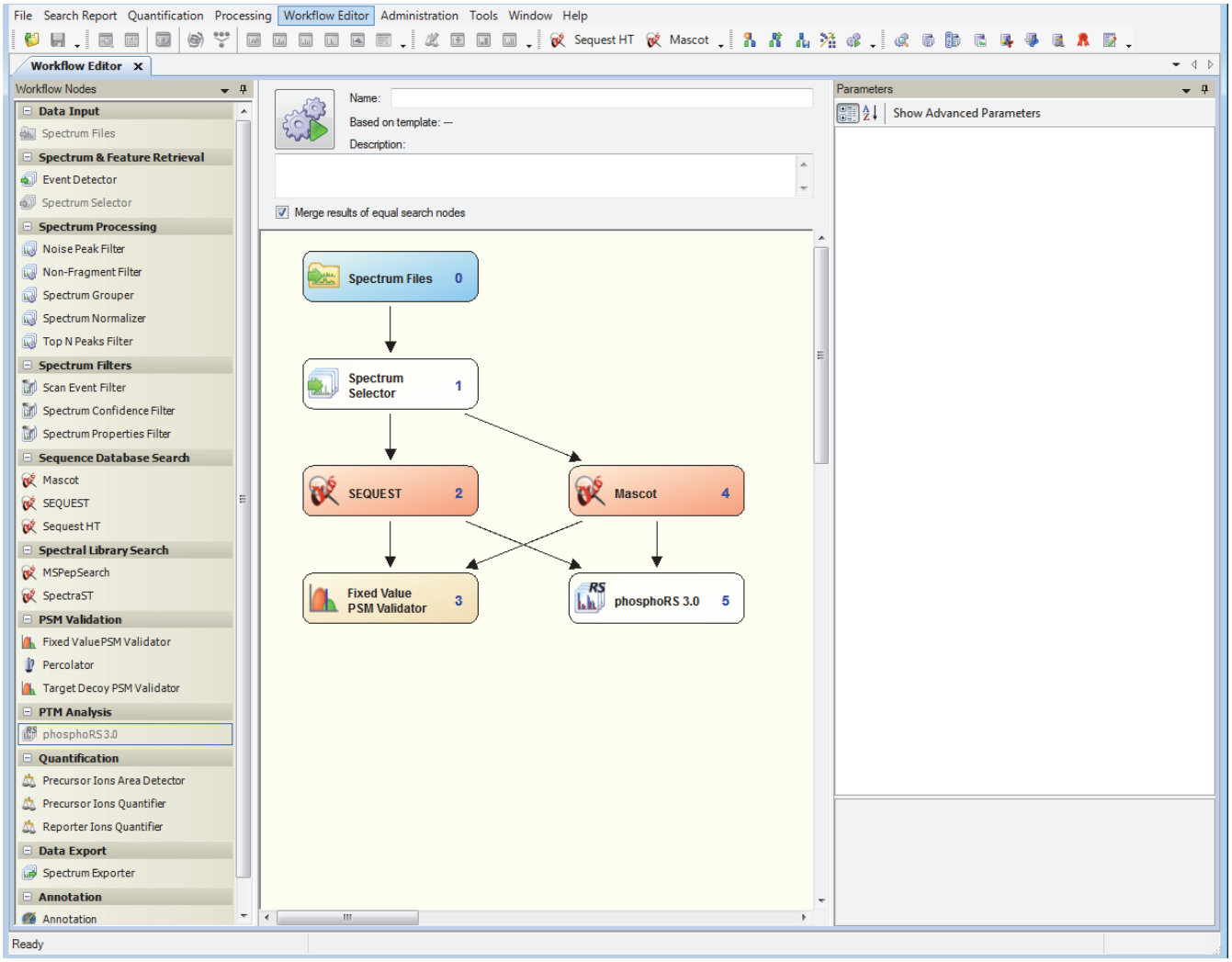
To create a workflow that uses the Annotation node to retrieve GO, Pfam, Entrez, and UniProt database information from ProteinCenter and install it in the Proteome Discoverer results files, see "Creating a Protein Annotation Workflow" on page 206.

Creating a PTM Analysis Workflow

If you want to focus on studying the biologically relevant post-translational modifications of proteins, you can create a workflow that includes the phospho*RS* node (refer to the Help). This node calculates PTM site localization scores for phosphorylation and makes them available in the Proteins Identification Details view when you choose Search Report > Protein ID Details View. This view color-codes the found phosphorylation modification above the amino acid sequences to indicate the probability of the modification being found on those portions of the amino acid. The PTM Site Probabilities area to the left of the sequence table displays a legend explaining the color-coding. For more information o this view, refer to the Help.

You can use only one phospho*RS* node in a workflow. Connect it to all search nodes whose results you want to submit to phosphorylation site localization scoring. Figure 34 gives an example of a workflow with two different search nodes attached to the phospho*RS* node.

Figure 34. Workflow with two different search nodes attached to the phospho*RS* node

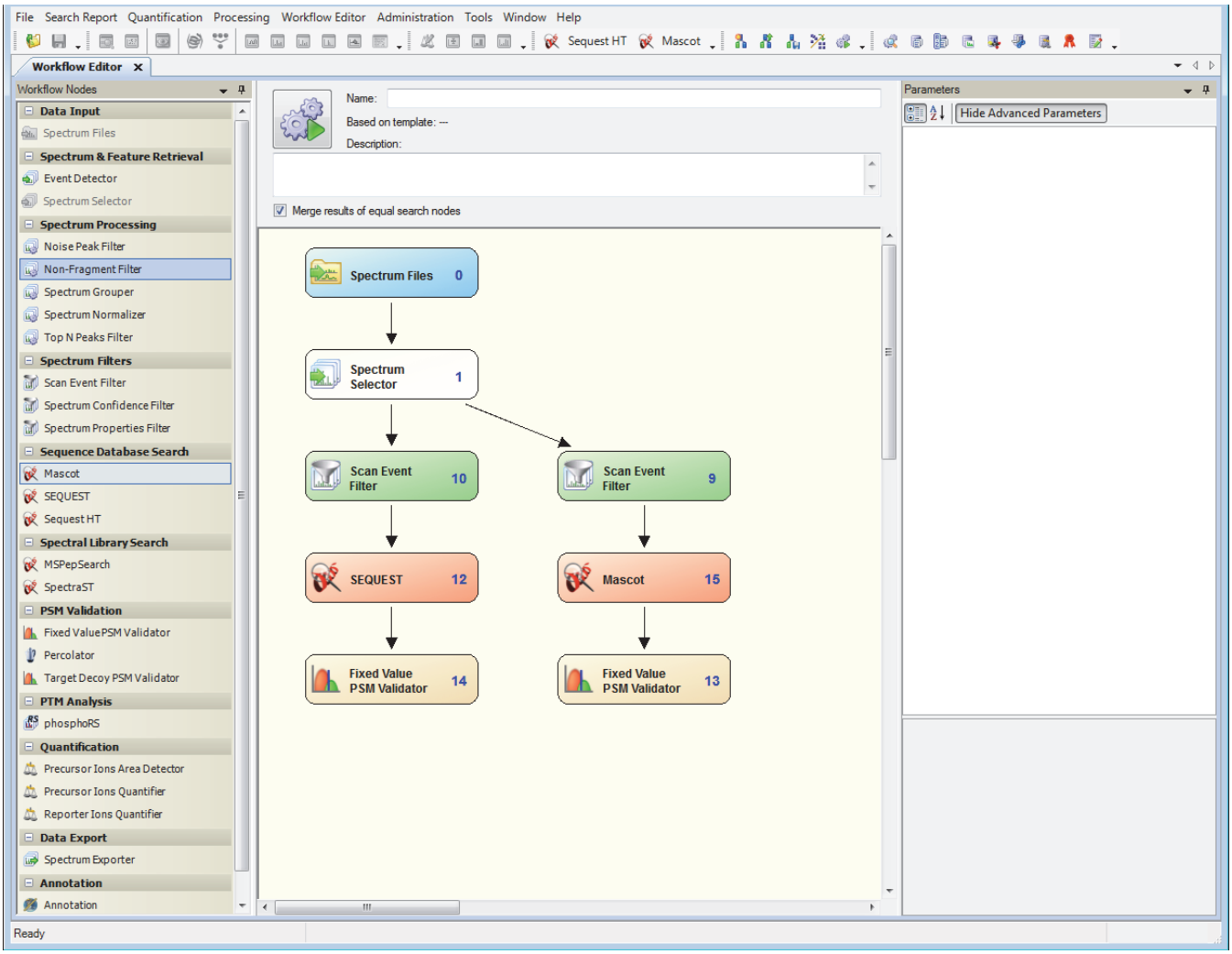


The phospho*RS* node retrieves the phosphorylation sites that were searched and the mass tolerance used for matching fragment ions directly from the attached search nodes. It has two additional parameters for choosing a specific mass tolerance to use when matching fragment ions (refer to the Help). With these parameters, you can overwrite the default mass tolerance setting used in the search node.

Creating Parallel Workflows

Parallel workflows are workflows that search the same raw data file and the same part of the spectrum but specify different criteria, different search nodes for the search, or both. They resemble the example workflow shown in Figure 35. You can use parallel workflows to conduct two or more searches using two or more search engines on the same raw data and to compare the results of these two searches at the same time. For example, you may want to search both CID and ETD data from the same raw data file to increase the chances of finding a match. CID data contains b and y ions, and ETD data contains b, c, and z ions, so the two types of data are complementary. You can also use a parallel workflow for quantification.

Figure 35. Parallel workflow



The following instructions show you how to create the simple parallel workflow shown in Figure 35.

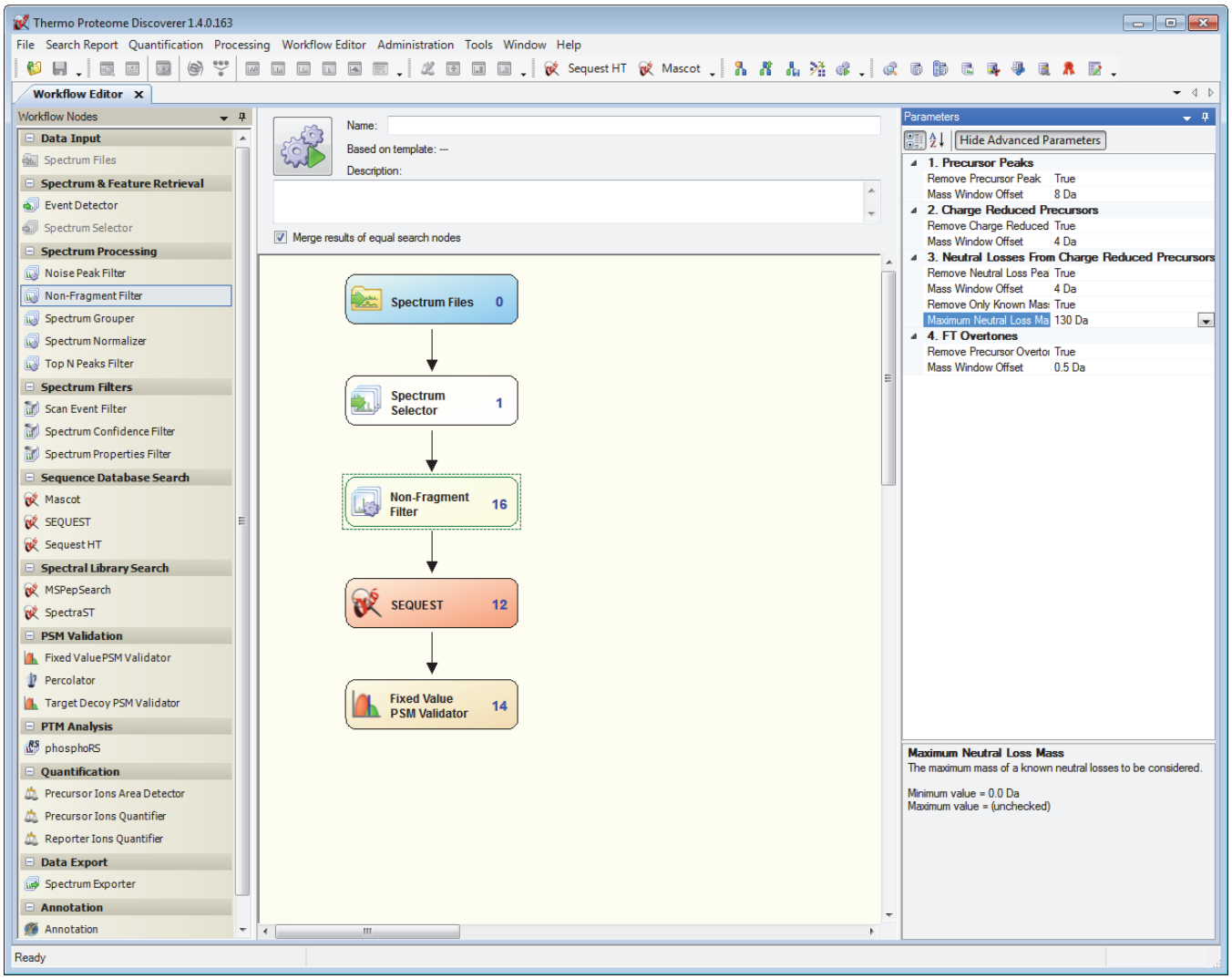
❖ To create a parallel workflow

1. Drag the Spectrum Files node to the workspace pane, and specify the name and path of the raw data file in the Parameters pane.
2. Drag the Spectrum Selector node to the workspace pane and place it directly under the Spectrum Files node. Set the parameters.
3. Drag two Scan Event Filter nodes to the workspace pane and place them side by side beneath the Spectrum Selector node. In the Parameters pane, set the Activation Type parameter to CID for one node and to ETD for the other node.
4. Drag the SEQUEST node to the workspace pane and place it beneath the Scan Event Filter node set to the CID activation type.
5. Drag the Mascot node to the workspace pane and place it beneath the Scan Event Filter node set to the ETD activation type.
6. Drag two Fixed Value PSM Validator nodes to the workspace pane and place one beneath the SEQUEST node and one beneath the Mascot node.
7. Connect the nodes as shown in Figure 35.
8. Choose **Workflow Editor > Start Workflow** to start the parallel workflow.

Adding a Non-Fragment Filter Node for High-Resolution Data

The main purpose of the Non-Fragment Filter node is to remove precursor peaks from the spectra that are not related to peptide fragments and could therefore increase the risk of the search engines making false positive matches. If you add a Non-Fragment Filter node to the workflow for processing data taken from Orbitrap instruments, Thermo Fisher Scientific recommends that you remove most of the precursor peaks. Setting the window to a smaller width increases the risk of leaving some of the precursor peaks or their side bands in the spectrum. Figure 36 shows the recommended settings with wider tolerances.

Figure 36. Non-Fragment Filter node settings for data taken from LTQ Orbitrap instruments



Peaks arising from overtones are rarely seen within Orbitrap spectra but are prominent peaks in spectra from the LT FT instruments. The range in which neutral loss peaks from the charge-reduced precursor peaks are removed is scaled by the charge of the charge-reduced peak. Therefore, if you specify a value of 130 Da, as in Figure 36, the Proteome Discoverer application removes neutral loss peaks within a 130-Da range for +1 peaks, a 65-Da range for +2 peaks, and so forth. To remove neutral losses, you can remove either every peak within the specified range or only those peaks from an internal table of known neutral loss masses from charge-reduced precursor ions, such as those shown in Table 2.

Table 2. Mass of known neutral losses from charge-reduced precursor ions

Mass	Neutral loss
17.027	NH3
18.011	H2O
27.995 Da	CO
32.026 Da	CH3OH
34.053 Da	N2H6 (2xNH3)
35.037 Da	H4NO
36.021 Da	H4O2 (2xH20)
44.037 Da	CH4N2
45.021 Da	CH3NO
46.006 Da	CH202
46.042 Da	C2H6O
59.037 Da	C2H5NO
59.048 Da	CH5N3
73.089 Da	C4H11N
74.019 Da	C3H6S
82.053 Da	C4H6N2
86.072 Da	C3H8N3
99.068 Da	C4H9N3
101.095 Da	C4H11N3
108.58 Da	C7H8O
131.074 Da	C9H9N

Opening an Existing Workflow

You can open an existing workflow from a template that you saved, or you can open it from an MSF or XML file. See the following:

- Opening an Existing Workflow from a Template
- Opening an Existing Workflow from an XML or MSF File

Opening an Existing Workflow from a Template

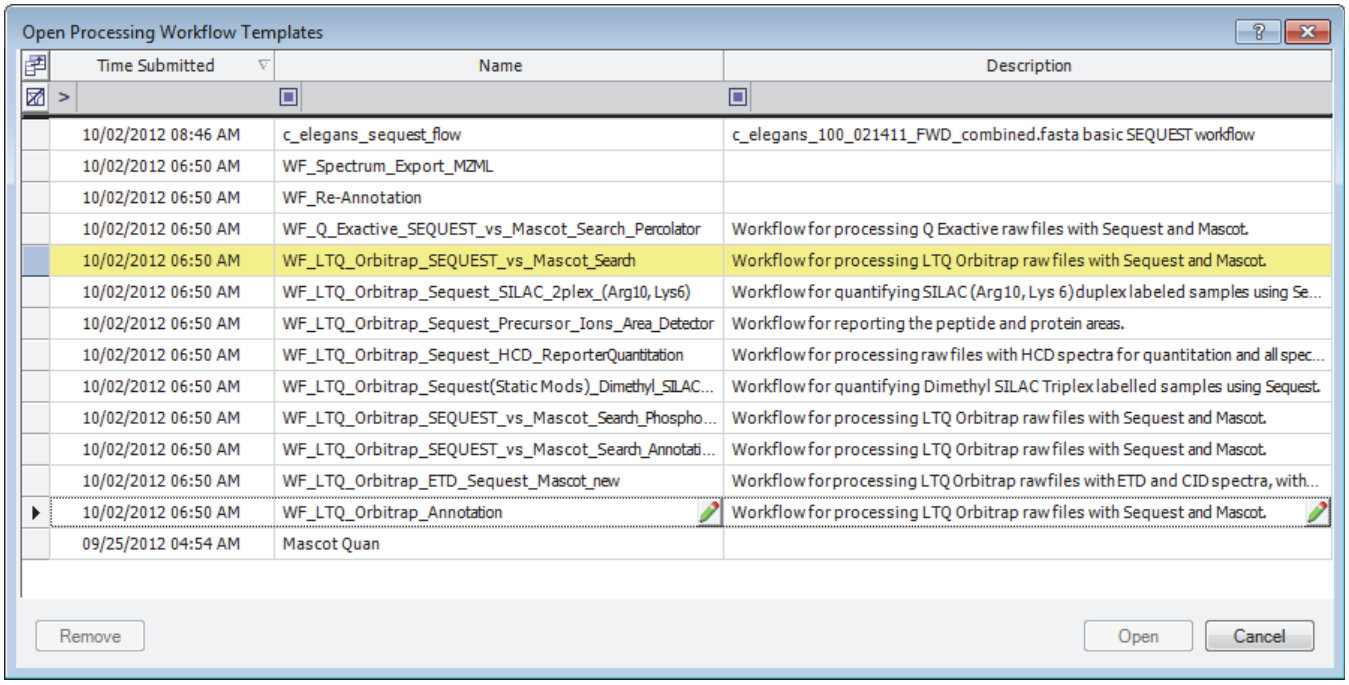
You can open an existing workflow that you previously saved when you chose Workflow Editor > Save As Template.

❖ To open an existing workflow from a template

1. Choose **Workflow Editor > Open From Template** or click the **Open From Template** icon,  .

   The Open Processing Workflow Templates dialog box appears, as shown in the example in Figure 37, listing the available workflow templates.

Figure 37. Open Processing Workflow Templates dialog box

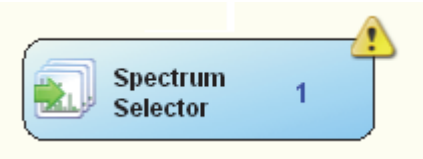


2. Select a workflow from the list.
3. Click **Open**.

   The Workflow Editor window opens, displaying the selected workflow. The Based on Template area now displays the name of the template that you chose.

When you open an existing workflow template, some of the nodes in the workspace pane might exhibit a yellow warning symbol, as shown in the example in Figure 38. This symbol indicates that the version of the node used when the template was created has been superseded by a later version in the current Proteome Discoverer application. Delete the node from the workflow, and drag the node with the same name from the Workflow Nodes pane to the workspace pane.

Figure 38. Warning symbol indicating an outdated node version

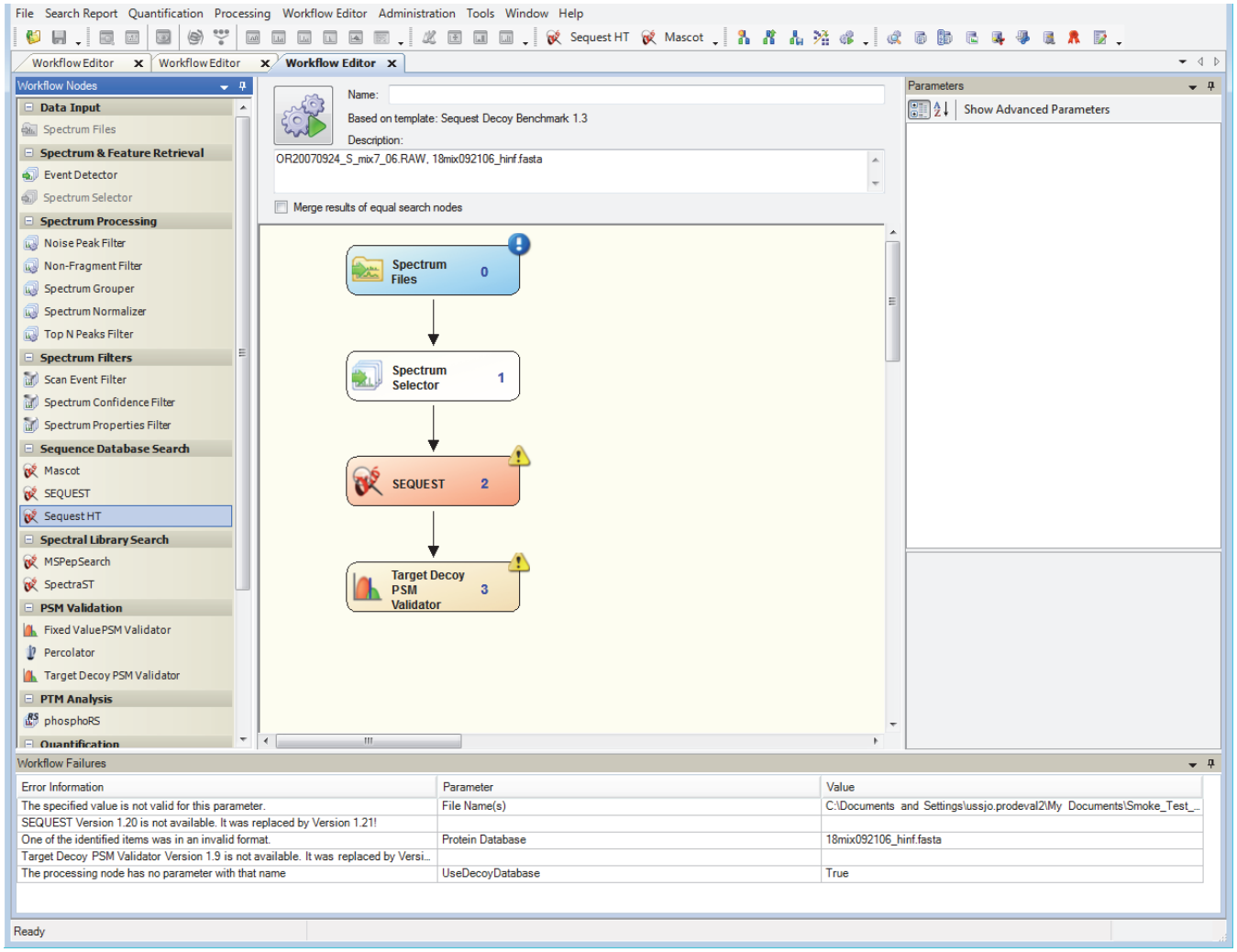


A round blue warning symbol containing an exclamation point, as shown in Figure 39, indicates that one or more of the parameter settings for the node are incorrect or outdated. Click on the node and reset the parameters in the Parameters pane.

Figure 39. Warning symbol indicating incorrect parameter settings

When you use a node that is outdated or has incorrect parameter settings, a Workflow Failures pane opens beneath the Workflow Nodes pane, as shown in Figure 40.

Figure 40. Workflow Failures pane



The Workflow Failures pane contains three columns:

- Error Information: Displays information about the problem that the application encountered in the workflow.
- Parameter: Displays the name of the node parameter that has an erroneous setting.
- Value: Displays the erroneous setting of the node parameter.

When a warning symbol is attached to a node, the Proteome Discoverer application automatically updates the node with the correct version, preserving the previous parameter values in the updated node. It does not include any node parameters that are no longer available and adds any new parameters set to their defaults.

If the Parameter and Value columns indicate a problem with the parameter settings, enter the correct parameter settings in the Parameters pane of the Workflow Editor.

Opening an Existing Workflow from an XML or MSF File

You can open a workflow from an existing MSF or XML file.

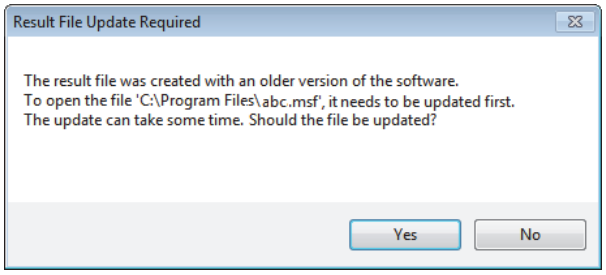
❖ **To open an existing workflow from an XML or MSF file**

1. Choose **Workflow Editor > Import Workflow**.
2. In the Import Workflow dialog box, browse to the XML or MSF file containing the workflow to import, and click **Open**.

   The selected workflow now opens in the Workflow Editor. The Proteome Discoverer application validates parameter settings and uses warning symbols to indicate outdated nodes. It displays error information in the Workflow Failures pane, as shown in Figure 40 on page 63.

   If you selected an MSF file and this file was created with an older version of the Proteome Discoverer application, the message box shown in Figure 41 appears.

   Figure 41. Message box

3. Click **Yes** to update to the current version.

The Proteome Discoverer application validates parameter settings and displays the selected workflow in the Workflow Editor, using warning symbols to indicate outdated nodes and displaying error information in the Workflow Failures pane.

If the Proteome Discoverer application cannot load the selected MSF file, it displays a message box with information about the issue. It cannot load files that are read-only or invalid, could not be updated, or were created with a newer version of the Proteome Discoverer application.

Deleting an Existing Workflow Template

You can delete an existing workflow template.

❖ **To delete an existing template**

1. Choose **Workflow Editor > Open From Template**.

   The Open Processing Workflow Templates dialog box appears, as shown in the example in Figure 37 on page 61, listing the available workflow templates.

2. Click the row displaying the name of the template that you want to delete.

3. Click **Remove**.
4. In the Confirm Deletion dialog box, click **Yes**.
5. Click **Remove** again.

Changing the Name and Description of a Workflow Template

You can change the name and the description of a workflow template.

❖ To change the name and description of a workflow template

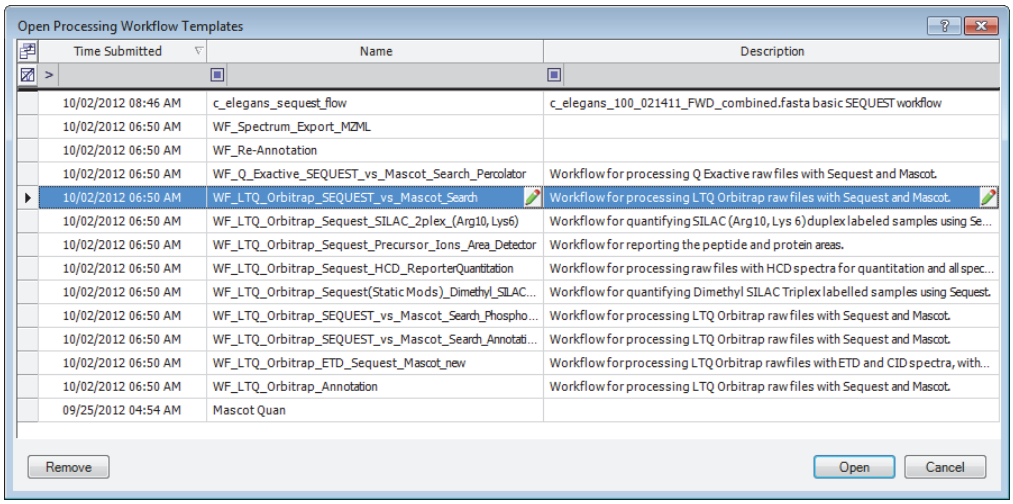
1. Choose **Workflow Editor > Open From Template**.

   The Open Processing Workflow Templates dialog box appears, as shown in the example in Figure 37 on page 61, listing the available workflow templates.

2. Click the row displaying the name of the template that you want to change.

   A Pen icon,  , now appears to the right of the template name and to the right of the template description, as shown in Figure 42.

   Figure 42. Pen icons in the Open Processing Workflow Templates dialog box



3. Click the **Pen** icon,  , and type the new name or the new description.

Importing Raw Data Files in Other Formats into a Workflow

You can import raw data files that were saved as MGF, MZDATA, MZXML, or MZML files into a workflow.

❖ To import raw data as MGF, MZDATA, MZXML, or MZML files

1. In the Workflow Editor, drag the Spectrum Files node to the workspace pane and select it.
2. In the Parameters pane, click the **Browse** button (...) next to the File Name(s) box.

3. In the Select Analysis File(s) dialog box, click **Add Files**.
4. Browse to the location of the MGF, MZDATA, MZXML, or MZML file and select it.
5. Click **Open**.
6. In the Select Analysis File(s) dialog box, click **OK**.
7. Continue with constructing the workflow according to the instructions in "Creating a Search Workflow" on page 44.

Saving a Workflow as an XML Template

To avoid losing any changes, you might want to save a workflow file as an XML template if you intend to transfer it to another computer, another software version, or another person.

❖ To save a search workflow as an XML template

1. Choose **Workflow Editor > Export Workflow to XML**.
2. In the Export Workflow Template dialog box, browse to the location where you would like to save the template, type a file name in the File Name box, and click **Save**.

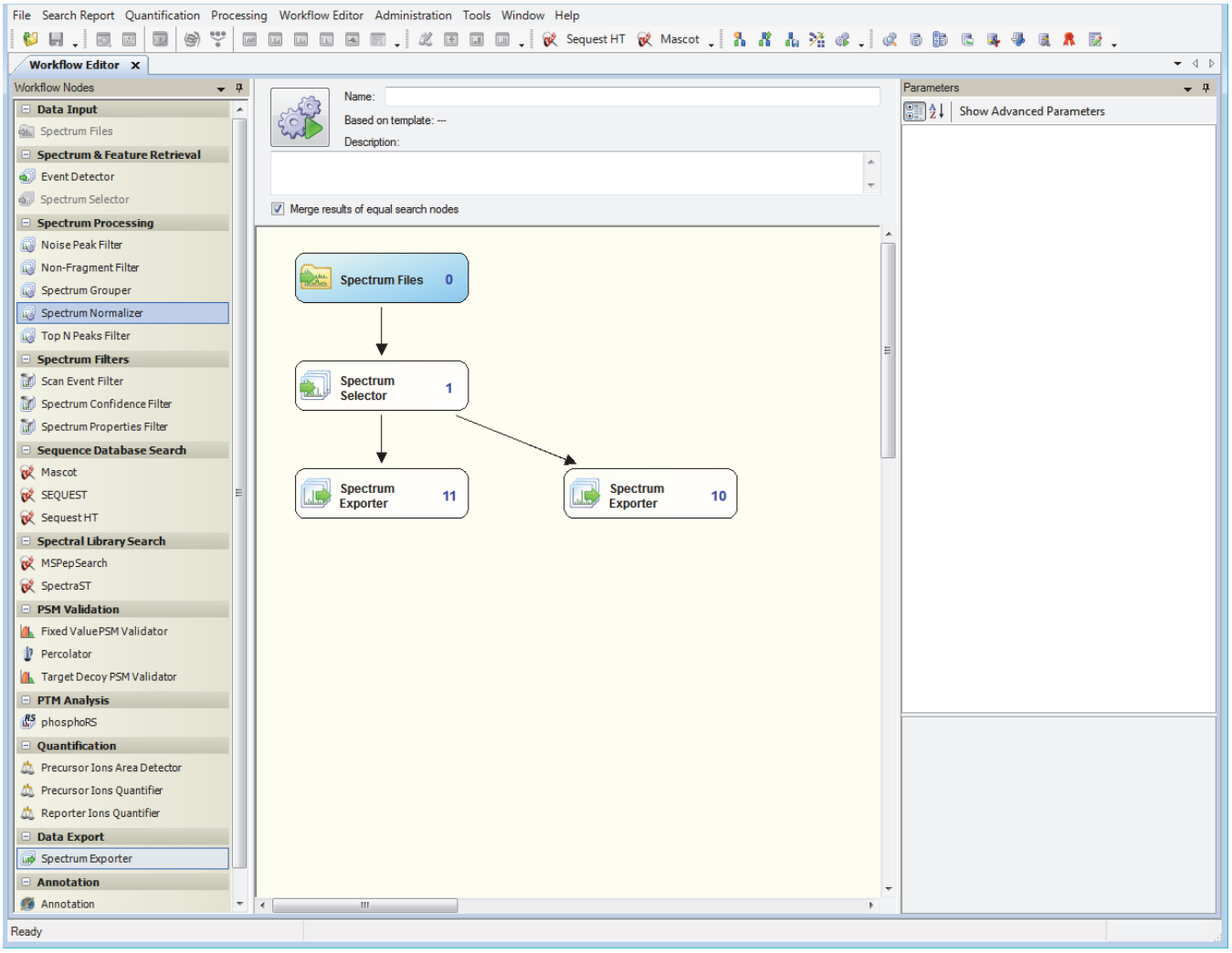
Exporting Spectra

By using the Spectrum Exporter node in your workflow, you can export spectra in the following standard formats:

- Data Archive (DTA): Places the exported spectra into DTA zip files, which are files containing MS^n data for single or grouped scans.
- Mascot Generic Format (MGF): Places the exported spectra into MGF files, which are mass spectral files produced during Mascot analysis. They contain a list of precursor ions, their fragments, and the masses of the fragments.
- MZDATA: Places the exported spectra into MZDATA files, which are common data format files developed by the Human Proteome Organization (HUPO) for proteomics mass spectrometry data. These files are in version 1.05 format. They are exported with XML indentation enabled so that the different XML tags are broken into multiple lines instead of merged into one line.
- MZXML: These files are standard 2.*x* mass spectrometer data format files developed at the Seattle Proteome Center at the Institute for Systems Biology (ISB) that contain a list of precursor ions, their fragments, and the masses of the fragment.
- MZML: These files are a combination of MZDATA and MZXML formats developed by the Human Proteome Organization Standard Initiative (HUPO-PSI) and the Seattle Proteome Center at the Institute for Systems Biology (ISB). The Proteome Discoverer application supports version 1.1.0.

You can select only one format for each Spectrum Exporter node. To export to multiple formats in a single workflow, you must add more than one Spectrum Exporter node to your workflow, as shown in Figure 43. Set the Export Format parameter on the first Spectrum Exporter node to one format and the Export Format parameter on the next node to another format, and so forth.

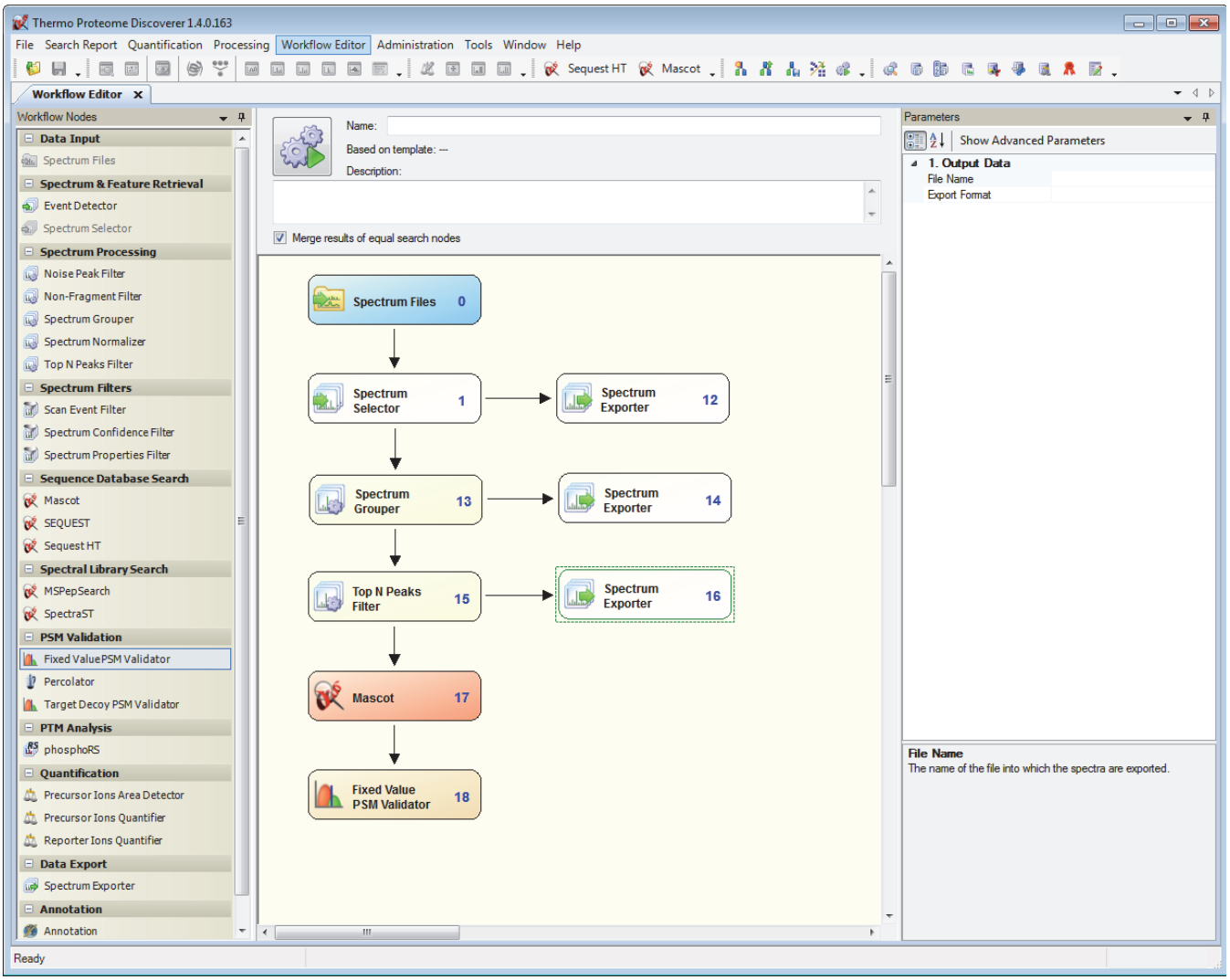
Figure 43. Workflow set to export data in two different formats



After starting the export process, the workflow starts like any other workflow processing job. After the application has finished processing the workflow, you can find the output of the Spectrum Exporter node in the same folder as the raw file. The Spectrum Files node specifies the location of the raw file.

You can also attach the Spectrum Exporter node to every node that creates, modifies, or outputs spectra, as shown in Figure 44. For example, you can add the Spectrum Exporter node to the Spectrum Selector node, the Spectrum Filter node, and the Spectrum Processing node. You can use this type of process flow to more closely inspect different spectrum processing steps in a workflow.

Figure 44. Using the Spectrum Exporter node to export spectra from different steps of the workflow



Using the Proteome Discoverer Daemon Utility

This chapter describes the Proteome Discoverer Daemon utility, which you can use to monitor job execution, perform batch processing, and process Multidimensional Protein Identification Technology (MudPIT) samples. You can select a server to connect to, start workflows, and monitor the execution of jobs on the configured server. Unlike the search wizards, which can only perform searches on one raw data file at a time, the Proteome Discoverer Daemon application can perform multiple searches on multiple raw data files at any given time. It can perform searches on multiple raw data files taken from multiple samples or from one sample. You can run the Proteome Discoverer Daemon application on the command line or in a window interface.

Contents

For information about MudPIT and creating a MudPIT workflow, see "Creating a Search Workflow for Multiple Raw Files from the Same Sample" on page 53.

Starting the Proteome Discoverer Daemon Application in a Window

You can start the Proteome Discoverer Daemon application on the command line or in a window. To run it on the command line, see "Running the Proteome Discoverer Daemon Application on the Command Line" on page 97.

❖ To start the Proteome Discoverer Daemon application in a window

1. Start the Proteome Discoverer Daemon application in Windows by choosing **Start > Programs > Thermo Proteome Discoverer *release_number* > Proteome Discoverer Daemon *release_number*** or by clicking the **Daemon** icon,  , on your desktop.
2. After the Proteome Discoverer Daemon application window appears, connect to a computer that is running the Proteome Discoverer application.

Selecting the Server

The Proteome Discoverer Daemon application can connect to a remote server so that you can perform searches on multiple raw data files from multiple samples or one sample on a remote computer. It can also connect to a local server.

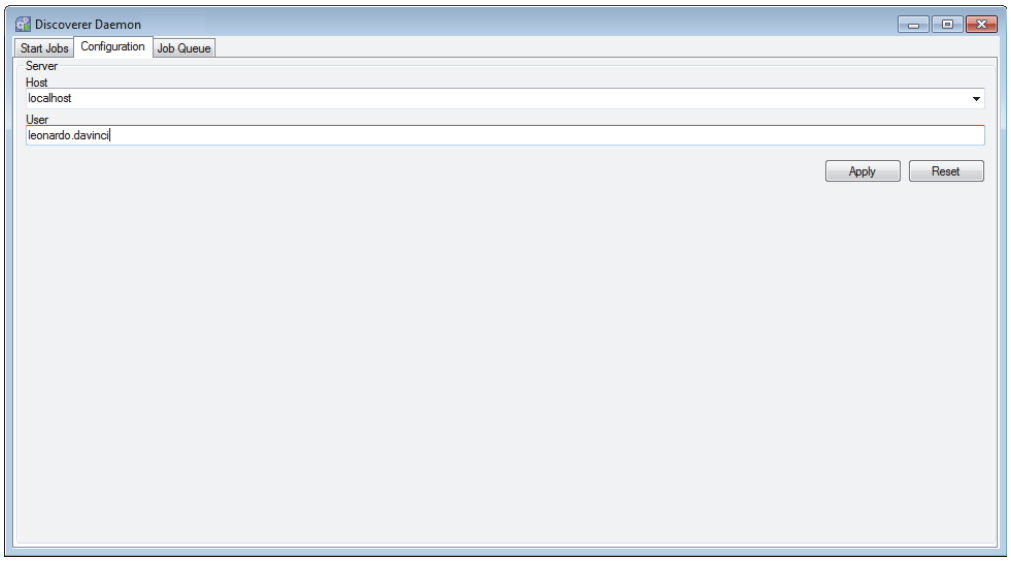
❖ To specify the server to connect to

1. Click the **Configuration** tab in the Proteome Discoverer Daemon application window.
2. From the Host list, select the name of the server that you want to use, or type the server name.

   You must connect the Proteome Discoverer Daemon application to a computer running the Magellan server. Your local host is the default server, that is, the computer that you are working on. To connect to a remote server, see "Logging On to a Remote Server" on page 76.
3. In the User box, type the login name of the server.

   The Configuration page now resembles Figure 45.

Figure 45. Configuration page of the Proteome Discoverer Daemon application



4. Click **Apply** to activate the newly entered settings.
5. To return to the previous settings, click **Reset**.

Starting a Workflow

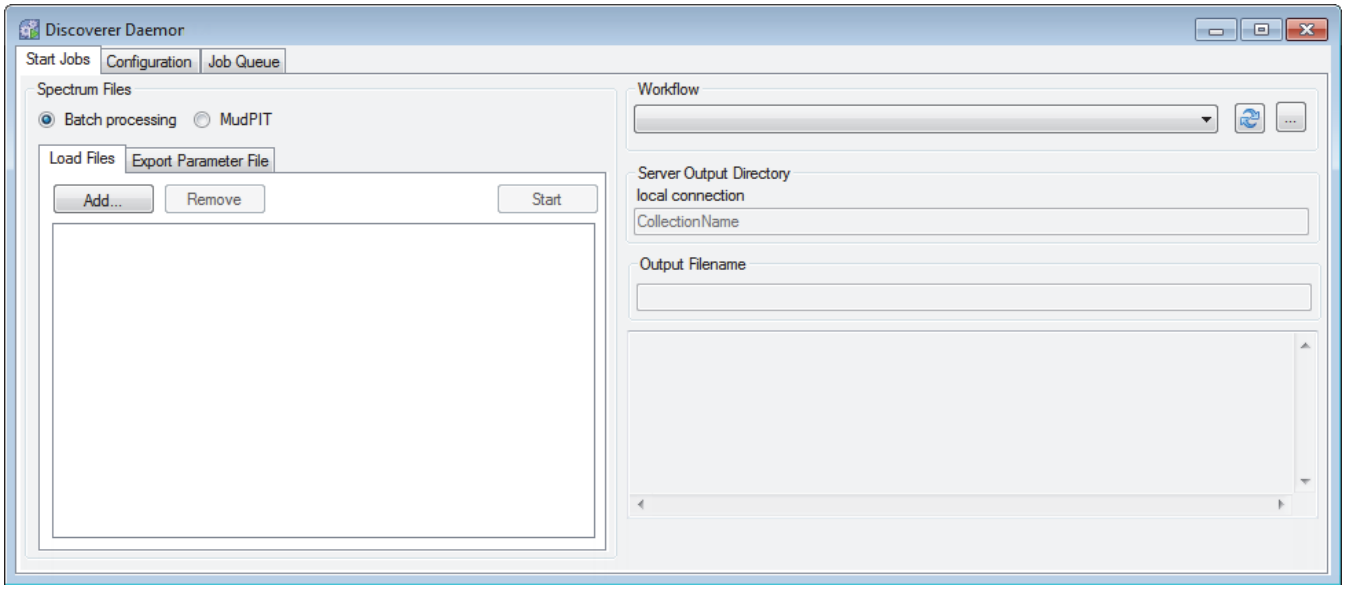
You can start a workflow for batch processing or MudPIT processing.

❖ To start a workflow

1. Click the **Start Jobs** tab.

   The Start Jobs page appears, as shown in Figure 46.

Figure 46. Start Jobs page of the Proteome Discoverer Daemon application



2. Click the **Load Files** tab, if it is not already selected.
3. Click **Add**.
4. In the Open dialog box, locate the file folder containing your raw data, select the spectrum (raw) file or files that you want to load, and click **Open**.

   The selected spectrum file or files appear on the Load Files page.

   To remove a file from the Load Files page, select the file and click **Remove**.
5. To specify the type of processing, select the **Batch Processing** or **MudPIT** option.
   - Batch processing (the default): Executes the workflow once for each spectrum file.
   - MudPIT: Feeds all spectrum files into one workflow.

   When you select the MudPIT option, the Output Filename box becomes available.
6. In the Workflow list, select the workflow template that you want to import.
   - Select the workflow from the Workflow list if it resides on the server that the Proteome Discoverer Daemon application is connected to.

     This workflow must be the one that was saved with the search parameters to be used with the given searches. You cannot modify parameters from the Proteome Discoverer Daemon application itself. Workflow templates that are missing more than the Spectrum File Names parameter do not appear in the Workflow list because the Proteome Discoverer Daemon application cannot complete them.

     –or–

- Select a valid workflow by clicking the **Browse** button (...) to select the workflow from your local machine.

  If you add workflow templates to the Proteome Discoverer application while the Proteome Discoverer Daemon application is running, click the **Refresh** icon,  , to display the workflow.

7. Connect to the server:

   If you have a local connection, the Proteome Discoverer application disables the Server Output Directory box and displays `local connection`. Then it places the output files are placed beneath the input files.

   If you connect to a remote server, in the Server Output Directory box, type the name of the directory where you want the original output files placed on the server.

   By default, the Proteome Discoverer Daemon application places this directory under the following directories:

   - Windows 7: c:\ProgramData\Thermo\Discsoverer <*release_number*> PublicFiles
   - Windows XP: c:\Documents and Settings\All Users\...\DiscovererDaemon\SpectrumFiles

   If you choose this directory, you must type a file folder name in the Server Output Directory box. You can specify a different directory by choosing Administration > Configuration in the Proteome Discoverer application, clicking Discoverer Daemon in the Server Settings section, and browsing for the location in the New Directory box.

8. If you selected the MudPIT option in the Spectrum Files area, in the Output Filename box, type the name of the output file that you want to store the results of the search in.

   The Start Jobs page should now resemble Figure 47 for batch processing or Figure 48 for MudPIT processing.

Figure 47. Start Jobs page of the Proteome Discoverer Daemon application for batch processing

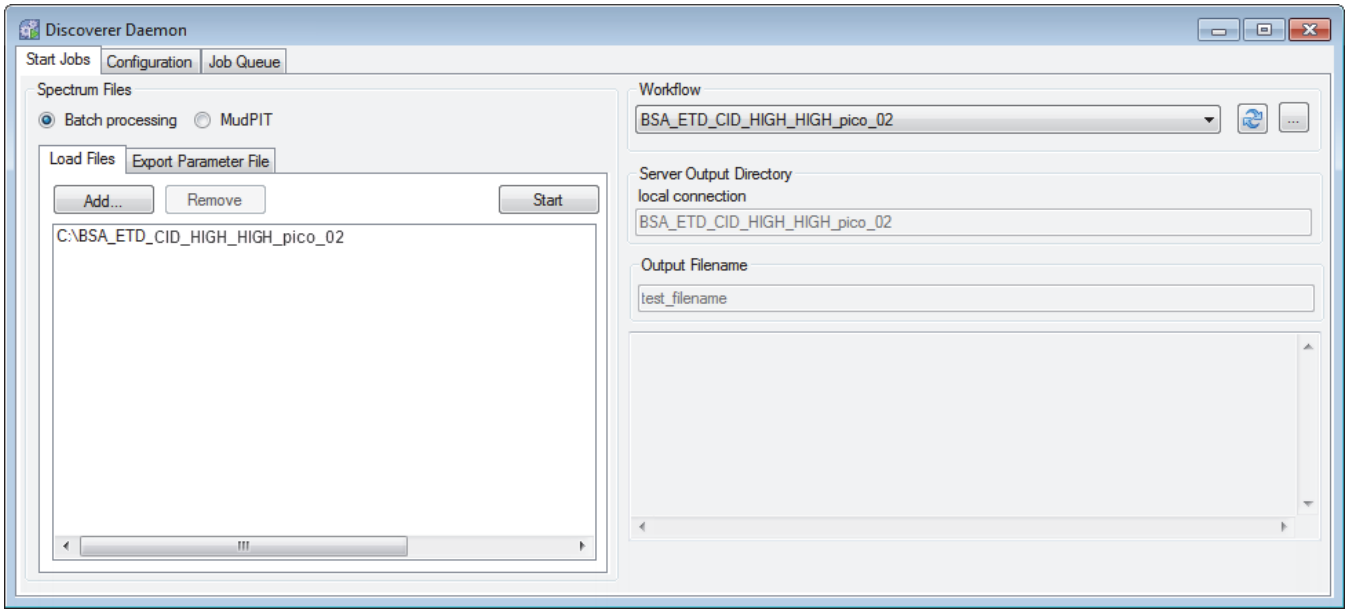
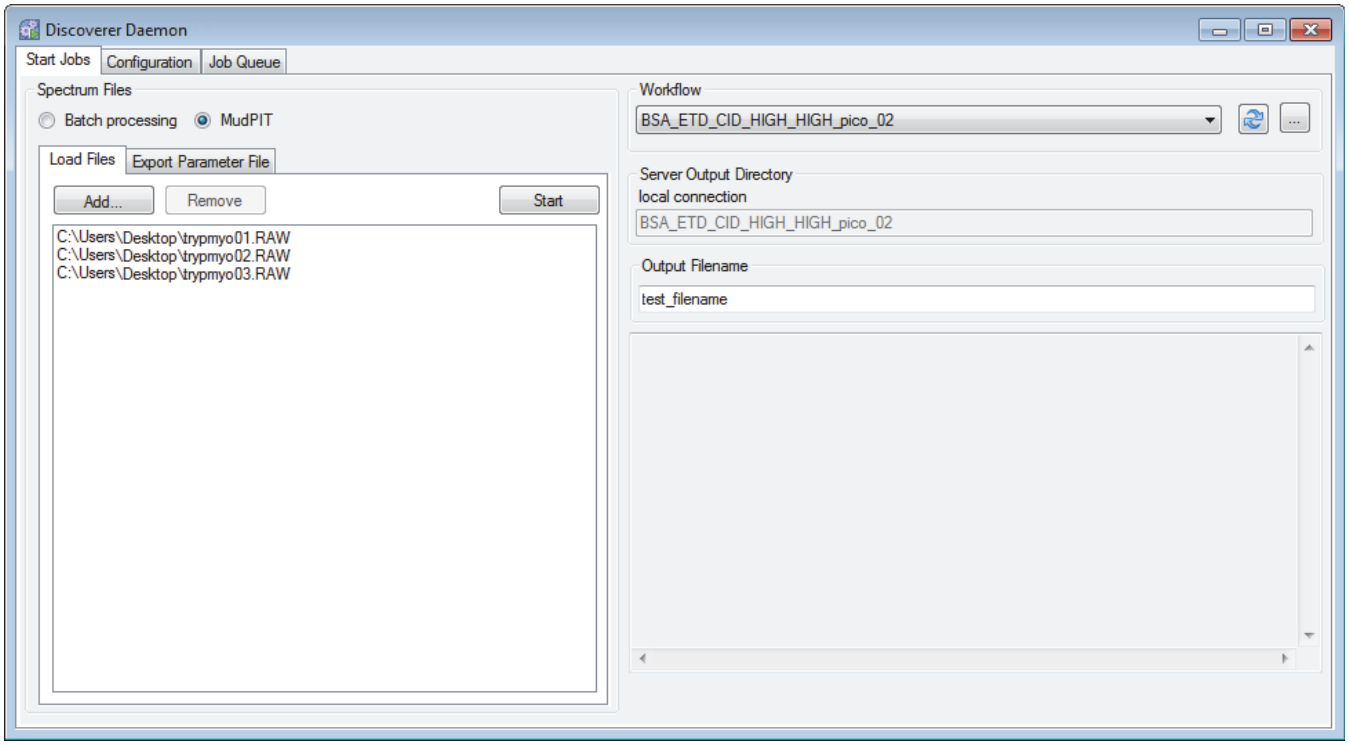


Figure 48. Start Jobs page of the Proteome Discoverer Daemon application for MudPIT processing



9. Click **Start** to execute the job.

Monitoring Job Execution in the Proteome Discoverer Daemon Application

You can use the Job Queue page in the Proteome Discoverer Daemon application window to monitor the execution of the jobs that you submit. It performs the same function as the job queue in the Proteome Discoverer interface. For information about the features of the job queue in the Proteome Discoverer interface, refer to the Help.

A progress bar displays the progress of the overall batch processing. This progress bar is only visible if you have started batch jobs.

❖ To monitor the job execution

- Click the **Job Queue** tab of the Proteome Discoverer Daemon application window.

Figure 49 shows the completed job for batch processing, and Figure 50 shows the completed job for MudPIT processing.

Figure 49. Job Queue page of the Proteome Discoverer Daemon application for batch processing

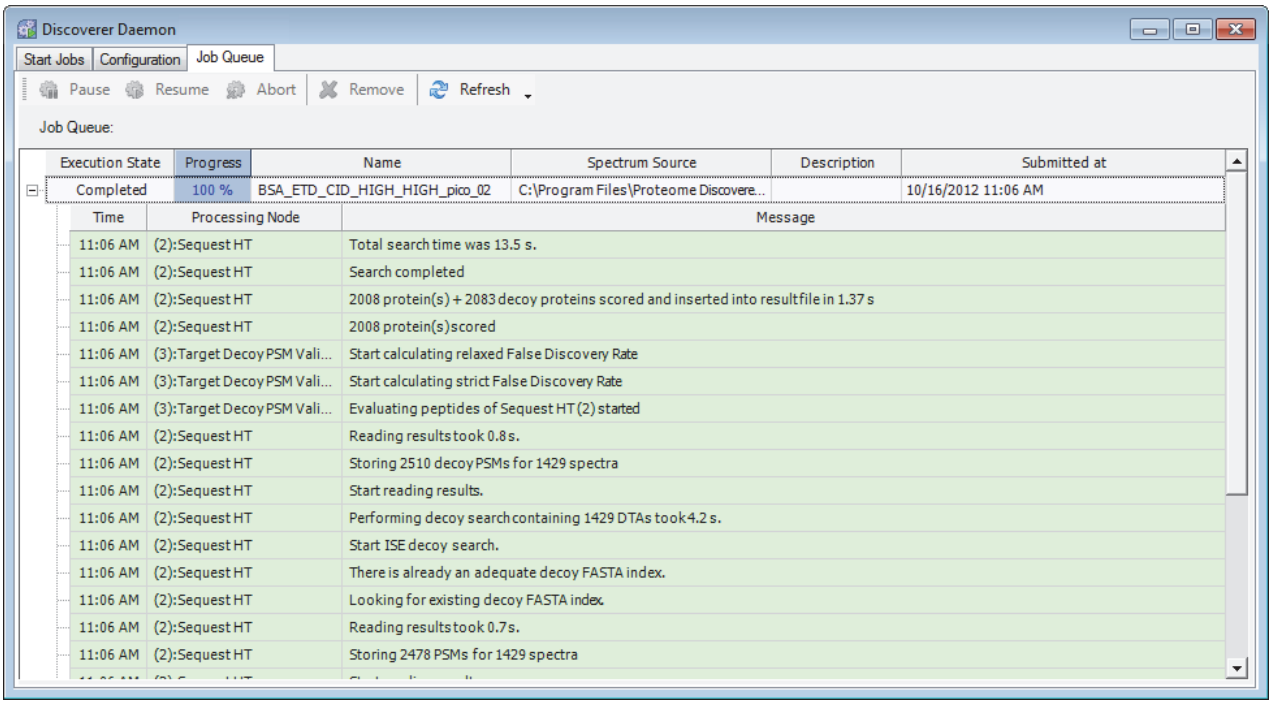
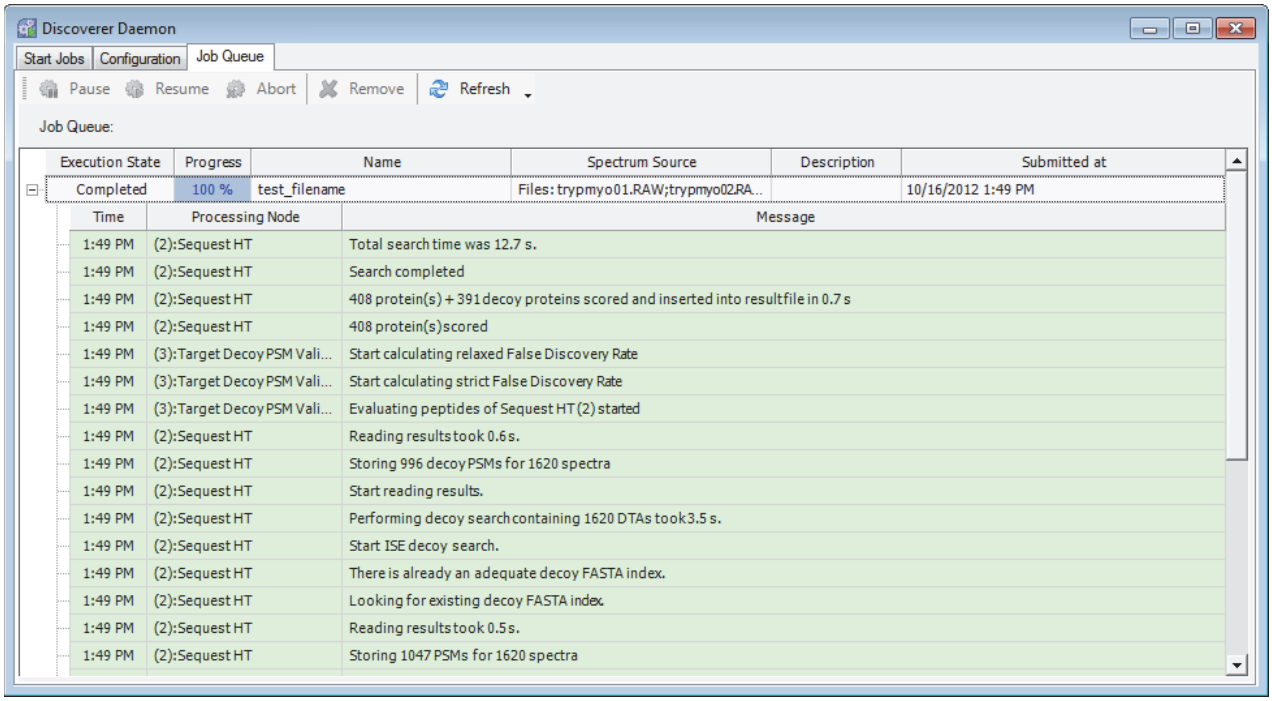


Figure 50. Job Queue page of the Proteome Discoverer Daemon application for MudPIT processing



Logging On to a Remote Server

The searches started by the Proteome Discoverer application consume memory and can potentially cause the data-acquiring computer to crash and lose the sample in the mass spectrometer. To avoid this outcome, Thermo Fisher Scientific recommends that you connect the Proteome Discoverer Daemon application to a remote computer running the Magellan server before data acquisition.

❖ To log on to a remote server

1. Start the Proteome Discoverer application on the remote machine.
2. If you want to store the output files in a location other than the default, do the following:
   a. Choose **Administration > Configuration > Server Settings > Discoverer Daemon**.

      The PublicFiles folder is the default file displayed in the Current File Directory box, as shown in Figure 51.

   b. In the New Directory box, browse to the location of the user-named folder in the PublicFiles folder on the server where you want to store the output files.

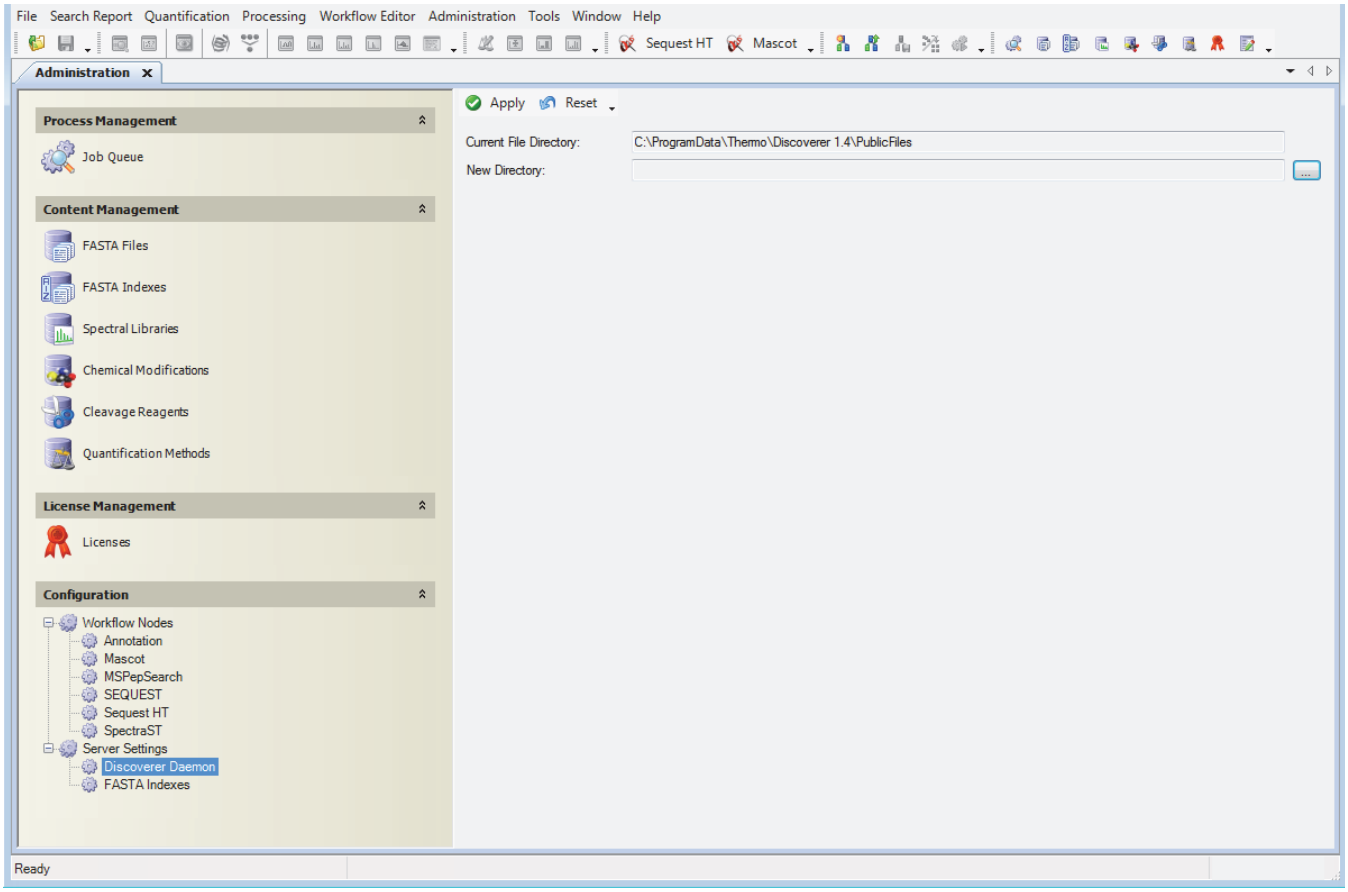
   c. Click Apply.

   If the directory already exists, it automatically appends the date and an incremental index number to the name.

If you attempt to create a file other than in the PublicFiles folder in the Current File Directory box, Discoverer Daemon issues a message informing you that the Proteome Discoverer application will apply the change the next time that you start it.

To return to the default directory, click Reset.

Figure 51. Discoverer Daemon area of the Configuration view



3. Start the Proteome Discoverer Daemon application on the local machine.

   A message box informs you that the Proteome Discoverer Daemon application cannot connect to the server.

4. Click **OK** in the message box.

   The Proteome Discoverer Daemon application opens with the Configuration page selected.

5. In the Host box, type the name of the remote computer.

6. In the User box, type the login name of the remote server.

7. Click **Apply**.

Running the Proteome Discoverer Daemon Application from the Xcalibur Data System

You can use the parameter file created in the Proteome Discoverer Daemon application to call the application from the Xcalibur data system.

For the Xcalibur 2.0.7 data system, you can start the Discoverer Daemon application in two ways:

- You can add a parameter file that calls the Discoverer Daemon application to the processing method specified in the Xcalibur injection sequence.
- You can select a parameter file for post-acquisition processing in the Programs area of the Run Sequence dialog when you start a sequence run.

For the Xcalibur 2.1.0 or later data system, you can start the Discoverer Daemon application only by adding a parameter file to the processing method specified in the Xcalibur injection sequence.

These topics describe how to run the Discoverer Daemon application from the Xcalibur data system:

Before You Start

Before you start running the Proteome Discoverer Daemon application from the Xcalibur data system, perform the following steps to ensure that the interface between the Proteome Discoverer Daemon application and the Xcalibur data system is optimal.

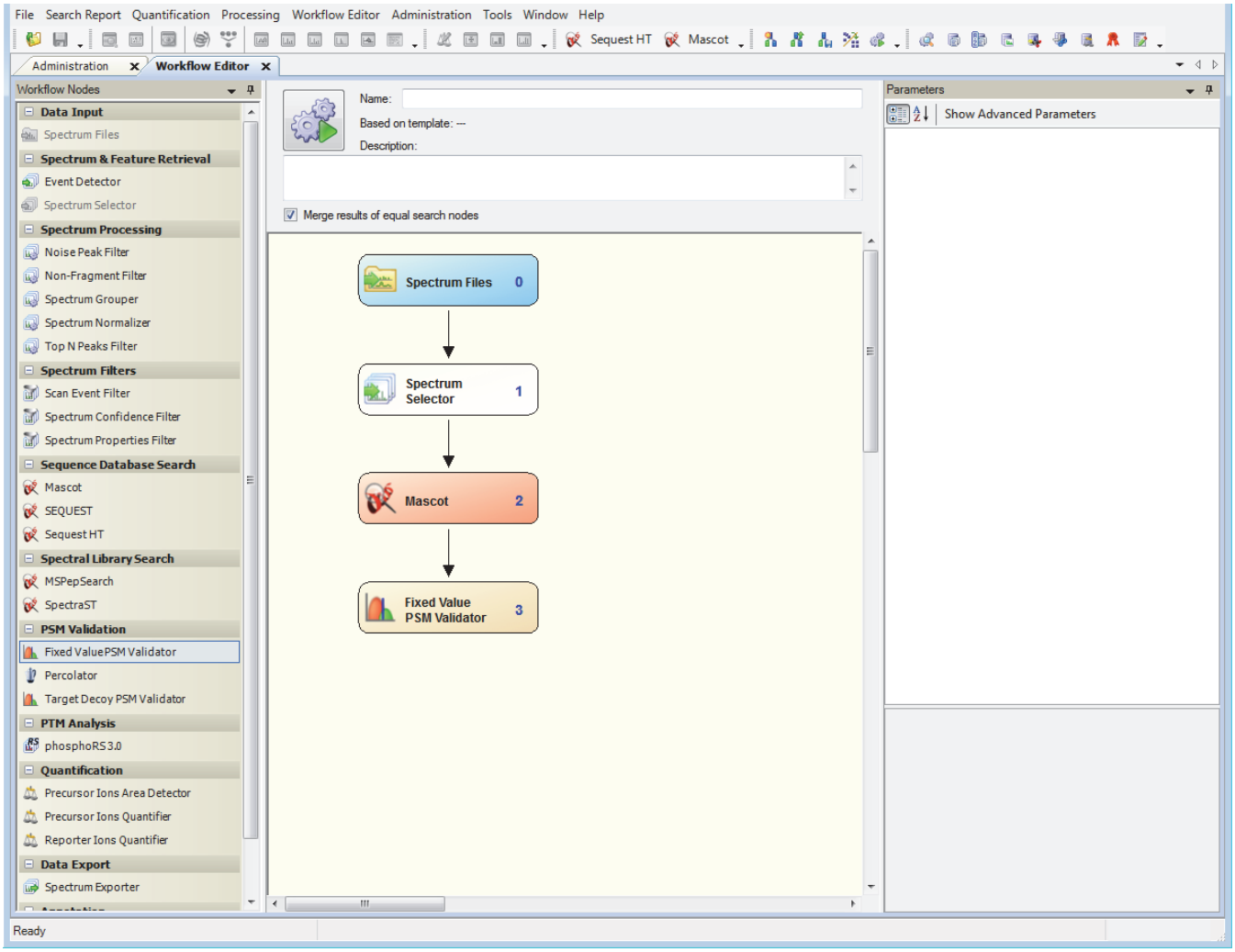
❖ To prepare to run the Proteome Discoverer Daemon application from the Xcalibur data system

1. Before you start the Proteome Discoverer Daemon application, install the Proteome Discoverer application on a remote computer to decouple data processing from data acquisition.

Thermo Scientific strongly recommends that you perform data analysis and data acquisition on two different computers to avoid disturbing the data acquisition by resource-consuming data processing.

2. Start the Proteome Discoverer application.
3. Install the Proteome Discoverer Daemon application on the same computer that the Xcalibur data system is running on.
4. In the Proteome Discoverer application, prepare the workflow to be used by the Proteome Discoverer Daemon application, as shown in Figure 52. Save this workflow.

Figure 52. Simple workflow used for the samples



After you install the Proteome Discoverer Daemon application, the Proteome Discoverer application places the directory where it saves the raw files and stores the results in the following files.

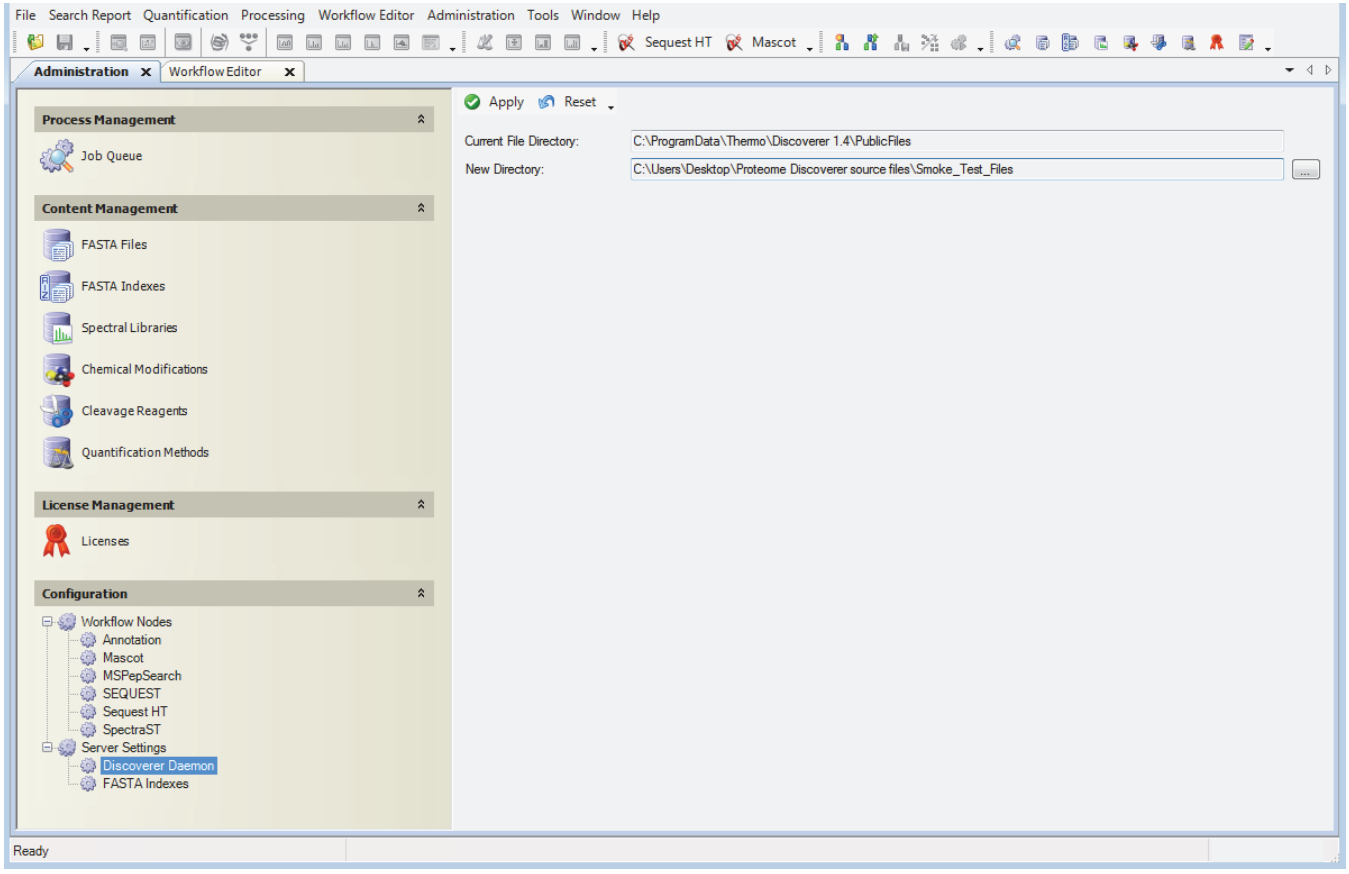
- Windows 7: c:\ProgramData\Thermo\Discsoverer <*release_number*> PublicFiles
- Windows XP: c:\Documents and Settings\All Users\Application data\ Thermo\Discoverer\Public Files.

This directory might be invisible to you because the C:\Documents and Settings\All Users\Application data directory is hidden. To display hidden directories, choose **Tools > Folder Options > View > Hidden files and folders > Show hidden files and folders** in Windows Explorer.

5. (Optional) To change this directory for easier data access, open the Proteome Discoverer application, choose **Administration > Configuration**, click **Discoverer Daemon** beneath Server Settings in the Configuration area on the left side of the Administration view, and change the directory in the New Directory box, shown in Figure 53.

   The settings are applied after you restart the Proteome Discoverer application.

Figure 53. Changing the destination directory where results from the Proteome Discoverer Daemon application are stored



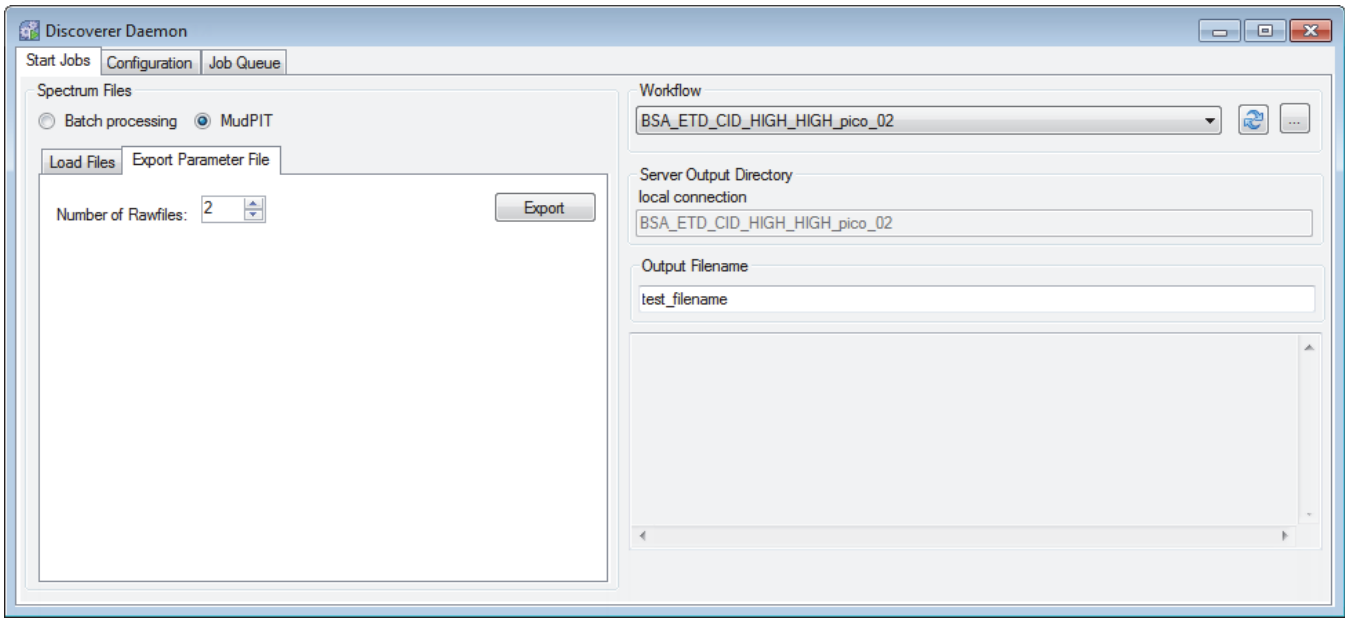
Creating a Parameter File That the Discoverer Daemon Application Uses

In the Proteome Discoverer Daemon application, you can create a parameter file that you can use to call the application from the Xcalibur data system. The application automatically translates the options that you set in the Proteome Discoverer Daemon application interface and in the workflow used for the search into text commands in the parameter file.

❖ To create a parameter file that calls the Discoverer Daemon application

1. Set up the search according to the instructions in "Starting a Workflow" on page 71. However, you do not have to have files loaded to create a parameter file.
2. Click the **Export Parameter File** tab, shown in Figure 54, on the Start Jobs page.

Figure 54. Export Parameter File page



3. In the Number of Rawfiles box for a MudPIT search, select the number of files that will appear in the Xcalibur Sequence Setup dialog box.

   The Number of Rawfiles option is not available when you select batch processing.

4. Click **Export**.

   The Save a Parameter File dialog box appears.

5. Specify the path and name of the parameter file, and click **Save**.

   The Proteome Discoverer application writes the parameter file in .xml format to the specified directory.

To call the Proteome Discoverer Daemon application through the parameter file, see "Running the Proteome Discoverer Daemon Application from the Xcalibur Data System."

Creating a Processing Method That Calls the Discoverer Daemon Application

The following procedure describes how to create a processing method that calls the Daemon application. It assumes that you have already created an appropriate processing method for your raw data files. Processing methods have a .pmd file extension.

❖ To add a processing method that calls the Discoverer Daemon application to a processing method

1. Choose **Start > All Programs > Thermo Xcalibur > Xcalibur** to start the Xcalibur data system.

   The Roadmap view of the Xcalibur Home Page window opens.

2. In the Roadmap view, do one of the following:

   - Choose **GoTo > Processing Setup**.

   –or–

   - Click the **Processing Setup** icon,  .

   The Processing Setup window opens.

3. Open the processing method that you want to modify as follows:

   a. Choose **File > Open**.

   b. Browse to the location of the processing method file and select the file.

   c. Click **Open**.

   The selected processing method opens in the Processing Setup window.

4. Open the Programs view of the Processing Setup window as follows:

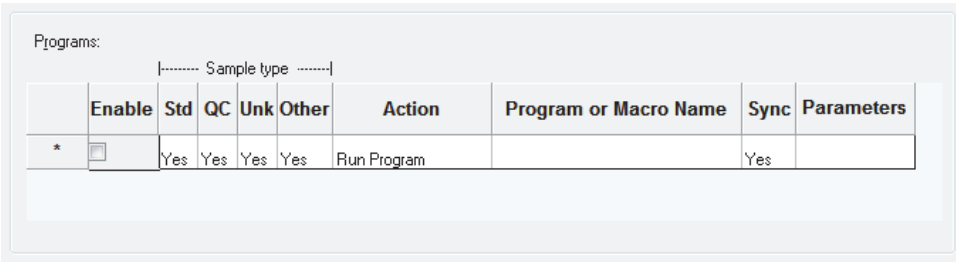
   a. Choose **View > View Bar**.

   The view bar appears on the left side of the dialog box.

   b. On the view bar, click the **Programs** icon,  .

   The Programs view of the Processing Setup window opens, as shown in Figure 55.

   Figure 55. Programs view with an empty table



5. If the Programs view contains an empty table, right-click the table and choose **Insert Row** from the shortcut menu.

   A new row appears above the placeholder row, as shown in Figure 56. An asterisk to the left side of a table row defines the row as a placeholder row.

   Figure 56. Programs view with an unedited table row

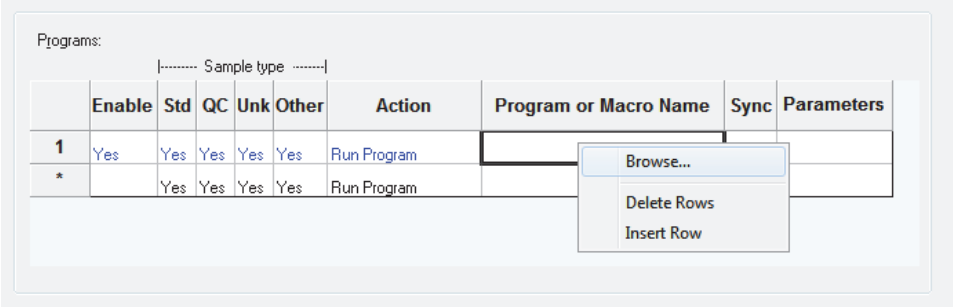
6. In the added table row, specify the name and location of the parameter file as follows:

   a. In the Enable column, select the check box.

   b. In the Action list column, select **Run Program**.

   c. Right-click the **Program or Macro Name** column and choose **Browse** from the shortcut menu, as shown in Figure 57.

      Figure 57. Programs view with the shortcut menu displayed



      The Browse for Program dialog box opens.

   d. Browse to the following executable, and click **Open**:

      C:\Program Files\Thermo\Discoverer\System\Release\DiscovererDaemon.exe

> **Note** If the following warning appears, click **OK**:
>
> `The file 'DiscovererDaemon' does not exist on this computer.`

e. In the Parameters column, type the location of the parameter file containing the commands that will execute the Proteome Discoverer Daemon application:

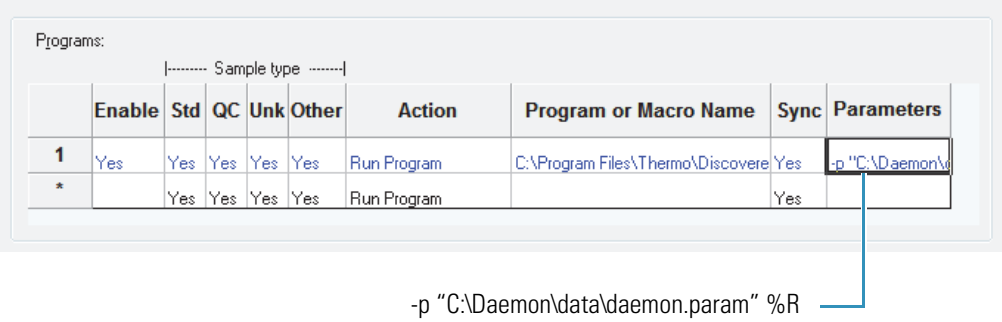
-p *path_to_parameter_file\parameter_filename* **%R**

> **IMPORTANT** If the name of the parameter file contains a space, you must enclose the name in quotation marks, as in this example:
>
> -p "C:\Xcalibur\methods\batch processing.param" %R

7. In the Std, QC, Unk, Other, and Sync columns, accept the default settings or modify them according to your requirements. For information about setting the sample types to be sent to the Discoverer Daemon application, see "To specify the sample types to be sent to the Discoverer Daemon application."

   To send all sample types to the Discoverer Daemon application, make sure that all of the sample type columns are set to **Yes**, as shown in Figure 58.

Figure 58. Program table with a call to the Daemon application



8. Click **OK** to save the changes to the processing method.
9. Choose **File > Save**.

❖ To specify the sample types to be sent to the Discoverer Daemon application

1. If the processing method that you want to modify is not open, open it and make sure that the parameter file and its location are specified as described in "To add a processing method that calls the Discoverer Daemon application to a processing method," on page 82.

2. In the Std, QC, Unk, and Other columns, do the following:
   - To send a sample to the Daemon application, make sure that "Yes" appears in the column for its sample type.
   - To avoid processing a sample with the Discoverer Daemon application, clear the column for its sample type.

   Tip Use the Other column for the Blank sample type. For example, if you do not want to send blank samples to the Discoverer Daemon application for further processing, clear the Other column.
3. Save the processing method.

Batch Processing with a Processing Method That Calls the Discoverer Daemon Application

To inject samples and to acquire and process data files with the Xcalibur data system, you must create one or more instrument methods, one or more processing methods, and a sequence that defines the sample injection set.

For information about creating an instrument method for your LC/MS system, refer to the Help for the LC devices and the Help for the mass spectrometer. For information about creating processing methods and sequences, refer to the Xcalibur Help.

Tip For a typical LC/MS experiment, an autosampler automates the sample injection process, and the position nomenclature depends on the autosampler tray type.

For information about specifying the autosampler tray type and the position nomenclature for the specified tray type, refer to the Help for the autosampler.

For some autosamplers, you can change the tray type from the Sequence Setup view by choosing Change > Tray Name, and then selecting a different tray type.

To start the Discoverer Daemon application from the Xcalibur data system version 2.10 or later, you must add a processing method that calls the Discover Daemon application to the sequence.

❖ To set up and run an injection sequence with a processing method that starts the Discoverer Daemon application

1. From the Home Page window of the Xcalibur data system, do one of the following:
   - Click the **Sequence View** icon,  , on the Home Page window toolbar.

   –or–
   - Click the **Sequence Setup** icon,  , on the Roadmap view.

The Sequence Setup view opens with an empty sequence table. Refer to the Xcalibur – Sequence Setup view Help for information about filling out the sequence table.

2. In the Proc Meth column, select a processing method with a parameter file that calls the Daemon application as follows:

   - Type the file location and name of the processing method.

   –or–

   - Double-click the column to open the Select Processing Method dialog box, where you can browse to and select the processing method.

   You can now start the sequence without first saving it or you can save the sequence for later use.

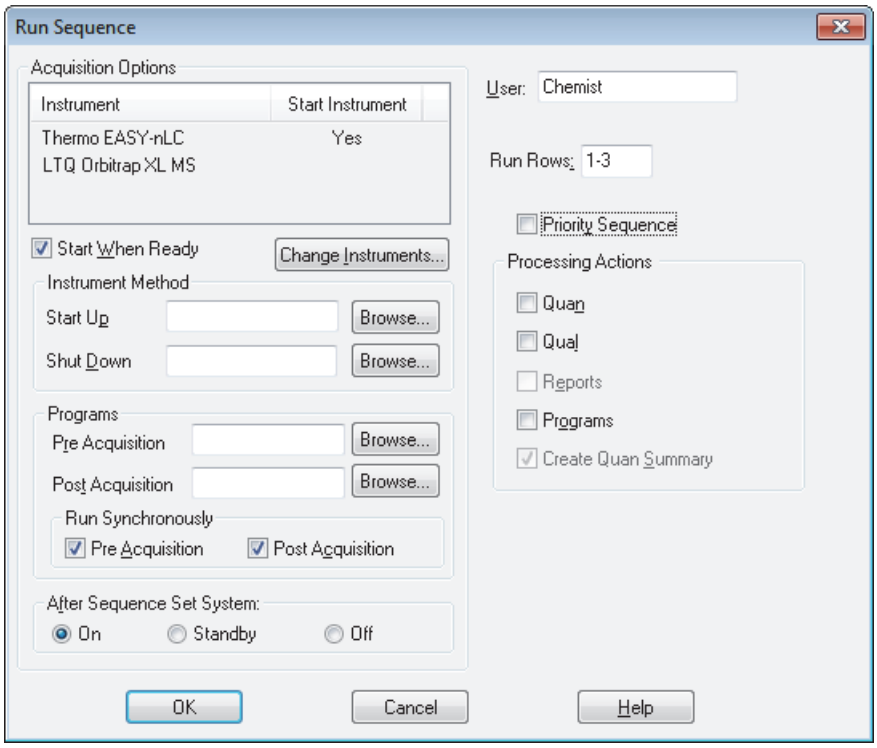
3. In the sequence table, select the row or rows that you want to run.

4. Choose **Actions > Run Sequence** or click the **Run Sequence** icon,  .

   If you have changed the instrument configuration in Foundation platform after the previous sequence run, the Change Instruments In Use dialog box opens. Otherwise, the Run Sequence dialog box opens, as shown in Figure 59.

   For an LC/MS system, the autosampler (or device with an autosampler) is specified as the start instrument. When the autosampler makes an injection, it triggers the mass spectrometer to begin data acquisition.

Figure 59. Run Sequence dialog box



5. Click **OK**.

   If you have not already saved the sequence, the File Summary Information dialog box opens.

6. Save the sequence as follows:

   a. In the File Summary Information box, click **OK**.

   b. In the File Name box, type a unique name for the sequence.

   c. In the Save In list, select the appropriate folder location for the sequence.

   d. Click **Save**.

   The Xcalibur data system adds the sequence to the acquisition queue.

For each sequence row, after the data system acquires a raw file, it sends the processing method and the raw data file to the Proteome Discoverer application, which stores the raw file and the MSF file in the server output directory specified in the Server Output Directory box of the Export Parameter File page of the Start Jobs page. All the search results of the batch processing are stored in the same directory. If the same directory name is used for the results of another batch process, the date and an index number that increments are appended to the folder name.

Batch Processing with Multiple Processing Methods

In some cases, you might need to use more than one processing method in the sequence. For example, the sequest.pmd method runs the Proteome Discoverer Daemon application with a parameter file containing a simple Sequest workflow, and the export.pmd method runs the Proteome Discoverer Daemon application with an export workflow.

❖ To use more than one processing method in a sequence

1. In the Sequence Setup view, choose **File > New**.

   The New Sequence Template dialog box opens.

2. Enter the appropriate values in each of the boxes.

3. In the Bracket Type area, select the **None** option, as shown in Figure 60.

   With this bracket type, you can change the processing methods individually for each sample.

Figure 60. New Sequence Template with the selection of None for the bracket type

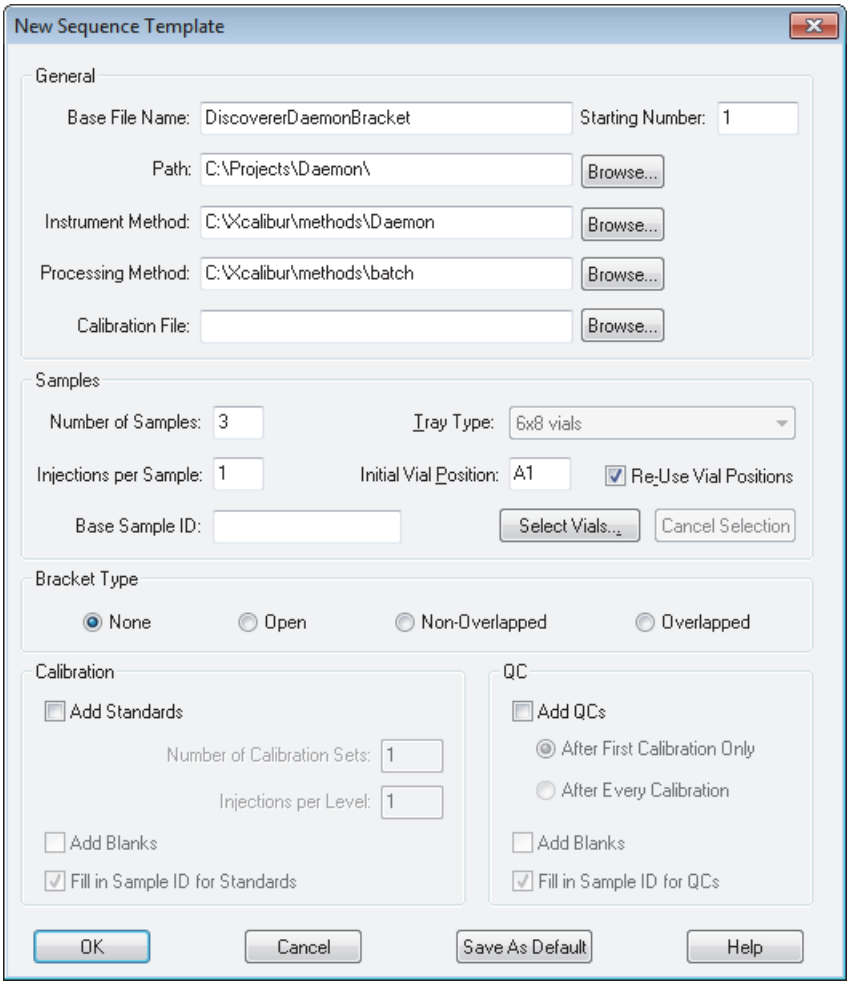


Figure 61 shows a sequence using two different processing methods.

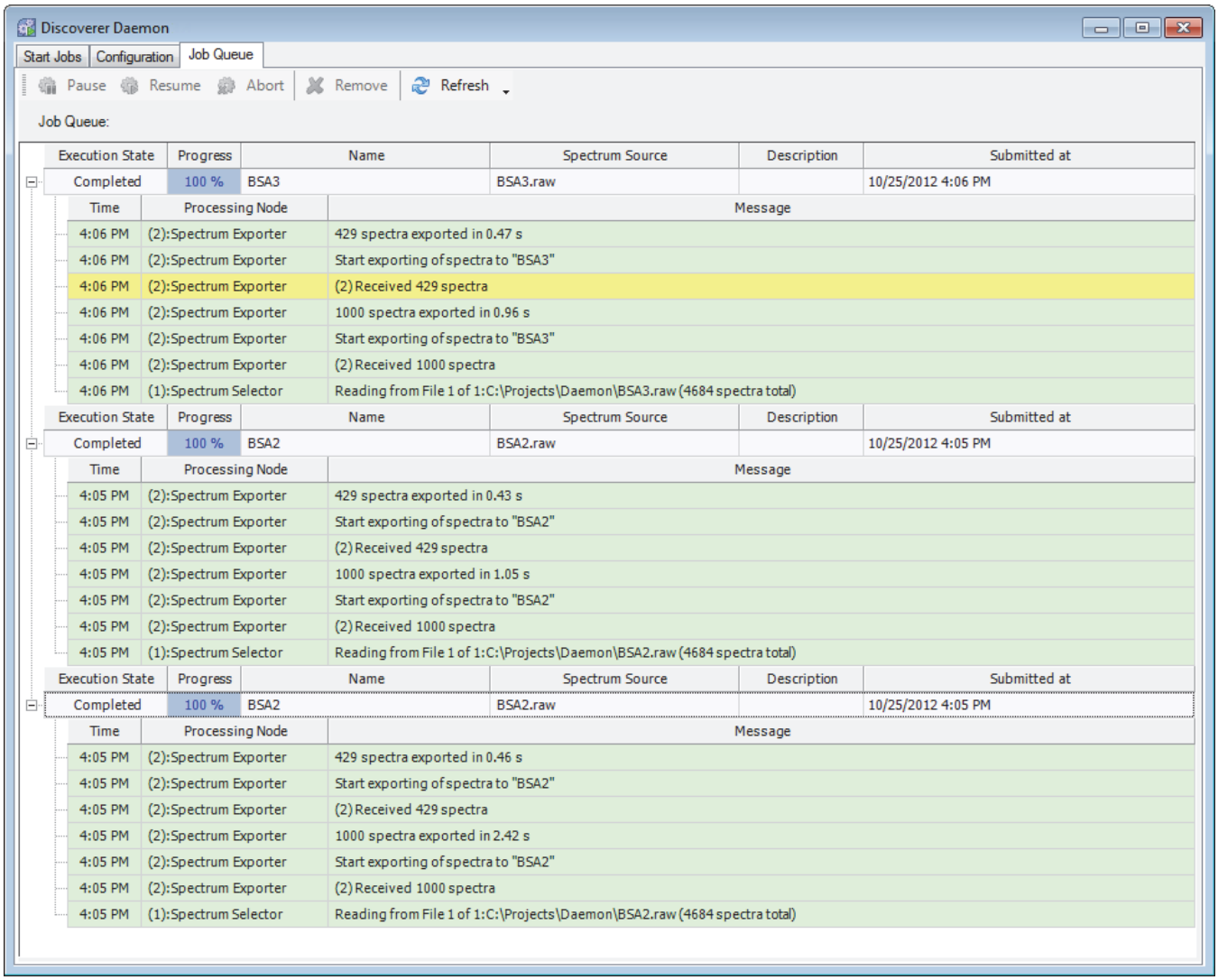
Figure 61. Sequence with two different processing methods

	Sample Type	File Name	Sample ID	Path	Inst Meth	Proc Meth	Position	Inj Vol
1	Unknown	BSA1	1	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\batch	A1	10.00
2	Unknown	BSA2	2	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\batch_export	A2	10.00
3	Unknown	BSA3	3	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\batch	A3	10.00
*								0.00

4. Click **OK**.

In this example, the Xcalibur data system starts two different workflows (performing a Sequest search and exporting a raw file) for the recorded raw data files in the Proteome Discoverer application, as shown in Figure 62.

Figure 62. Two workflows in the job queue started by two different processing methods



Batch Processing by Using a Post-Acquisition Method (Xcalibur Data System 2.0.7 Only)

You can perform batch processing by using different processing methods for different samples. However, editing the processing method is complicated. For quick synchronous processing of the same workflow, you can use the Proteome Discoverer Daemon application as a post-acquisition method in the Run Sequence dialog box.

Note Using the post-acquisition method with the Proteome Discoverer Daemon application does not work with the Xcalibur data system 2.1.0. It only works with the Xcalibur data system 2.0.7, which runs on Windows XP.

To use the Proteome Discoverer Daemon application in the Run Sequence dialog box, you do not need a processing method. Figure 63 shows the sequence setup without a processing method.

Figure 63. Sequence used to start batch processing in the Run Sequence dialog box

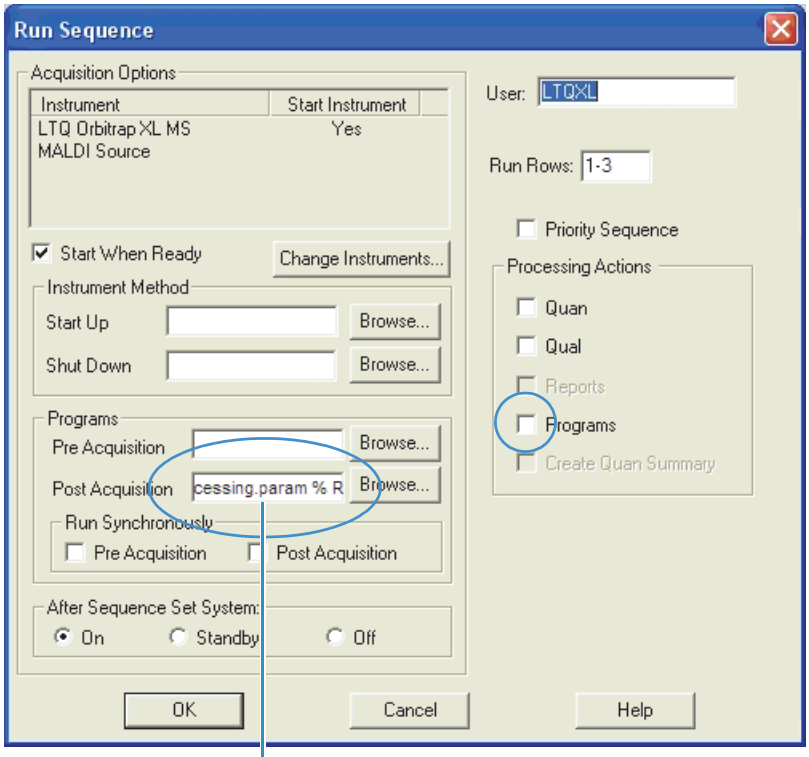
	File Name	Path	Inst Meth	Position	Proc Meth	Sample Type
1	BatchSample1	C:\Xcalibur\Data	C:\Xcalibur\methods\top50plu	H6		Unknown
2	BatchSample2	C:\Xcalibur\Data	C:\Xcalibur\methods\top50plu	H7		Unknown
3	BatchSample3	C:\Xcalibur\Data	C:\Xcalibur\methods\top50plu	H8		Unknown
*						

❖ To perform batch processing by using the Run Sequence dialog box

1. To start the sequence, click the **Run Sequence** icon,  .
2. In the Run Sequence dialog box, shown in Figure 64, enter the following in the Post Acquisition box:

   C:\Program Files\Thermo\Discoverer\System\Release\discovererdaemon.exe -p C:\Xcalibur\methods\BatchProcessing.param %R

   Figure 64. Using the Proteome Discoverer Daemon application in the Run Sequence dialog box (Windows XP only)



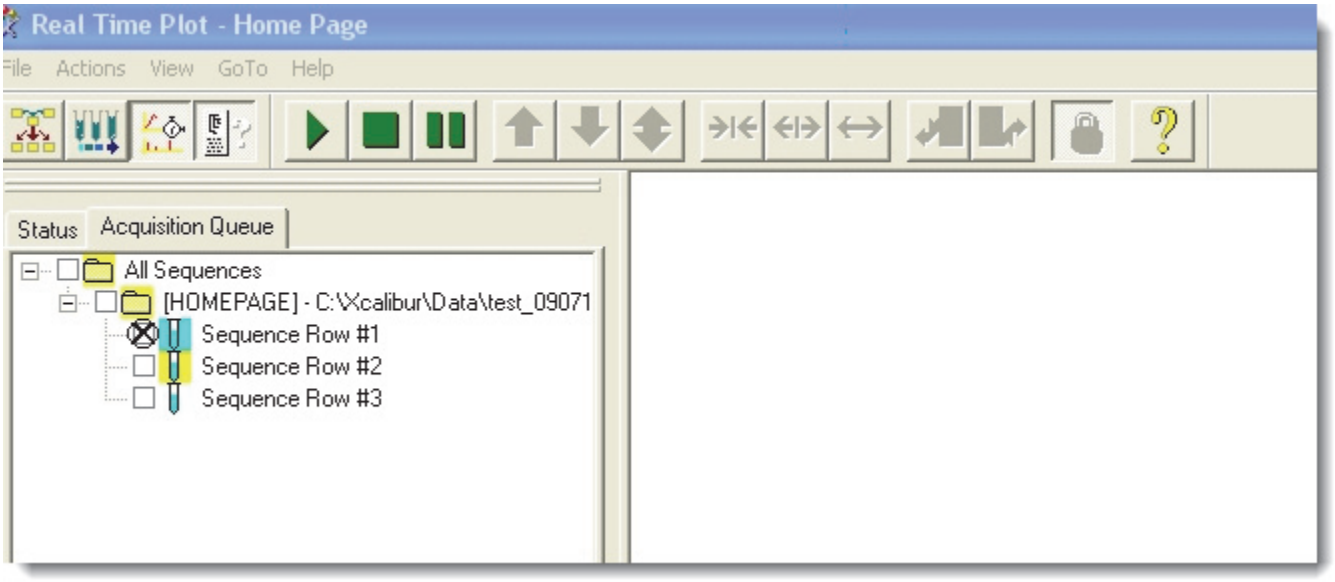
C:\Program Files\Thermo\Discoverer\System\Release\discovererdaemon.exe -p C:\Xcalibur\methods\BatchProcessing.param %R

3. If the **Programs** check box in the Processing Actions area on the right is selected, clear it.

The Xcalibur data system sends the acquired raw data files synchronously to the Proteome Discoverer application, as shown in Figure 65.

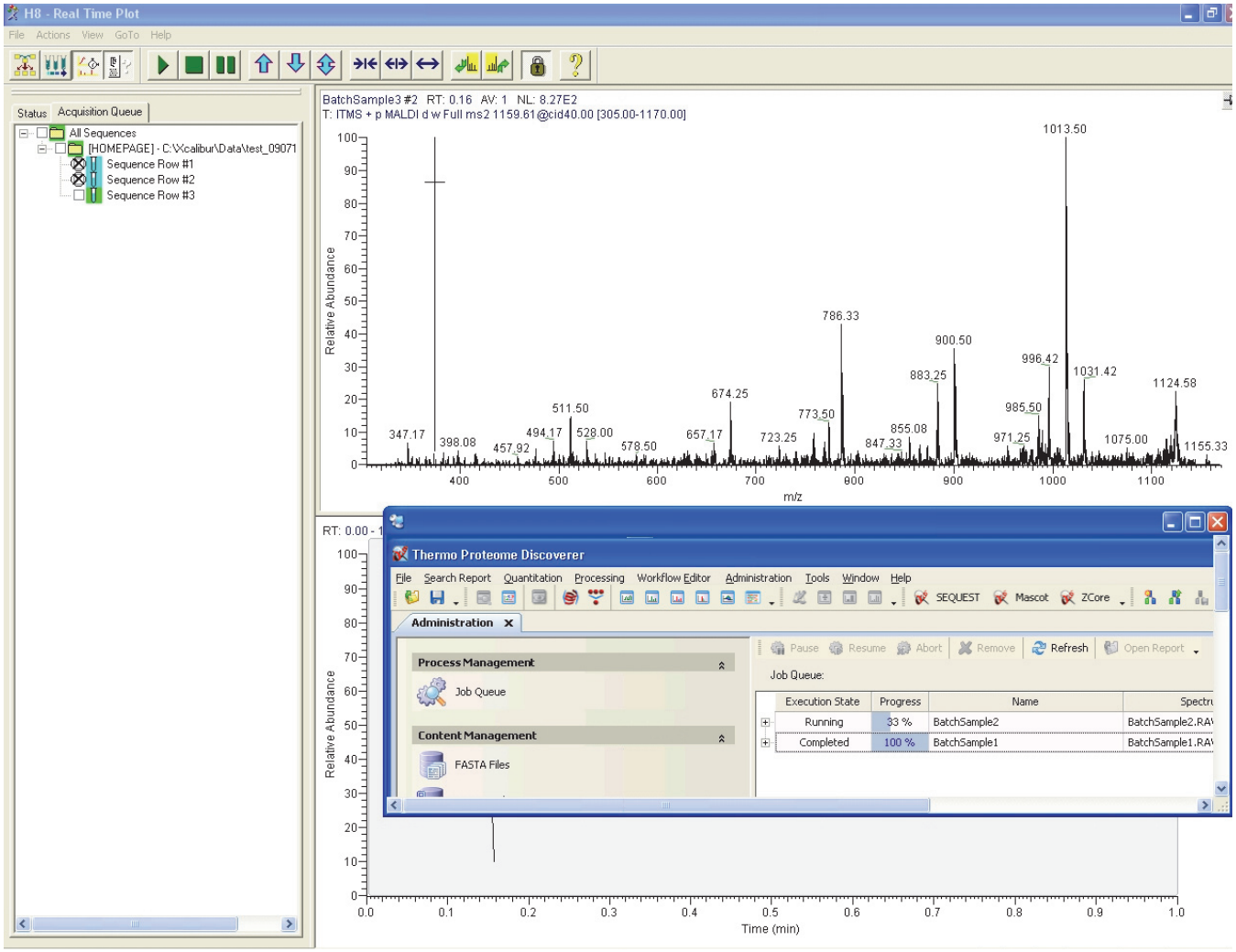
Note Only the Xcalibur 2.0.7 data system sends the acquired raw data to the Proteome Discoverer application. This functionality is not available in version 2.1.0.

Figure 65. Sending the raw data files synchronously to the Proteome Discoverer application after the first sample is finished



The Proteome Discoverer application synchronously processes the raw files on the remote host, as shown in Figure 66.

Figure 66. Processing the raw files synchronously on the remote host



In this example, the Proteome Discoverer application processes all three raw data files and places them in the directory that you set for the Discoverer Daemon application on the computer running the Proteome Discoverer application, as shown in Figure 67 and Figure 68.

Figure 67. Completed data processing

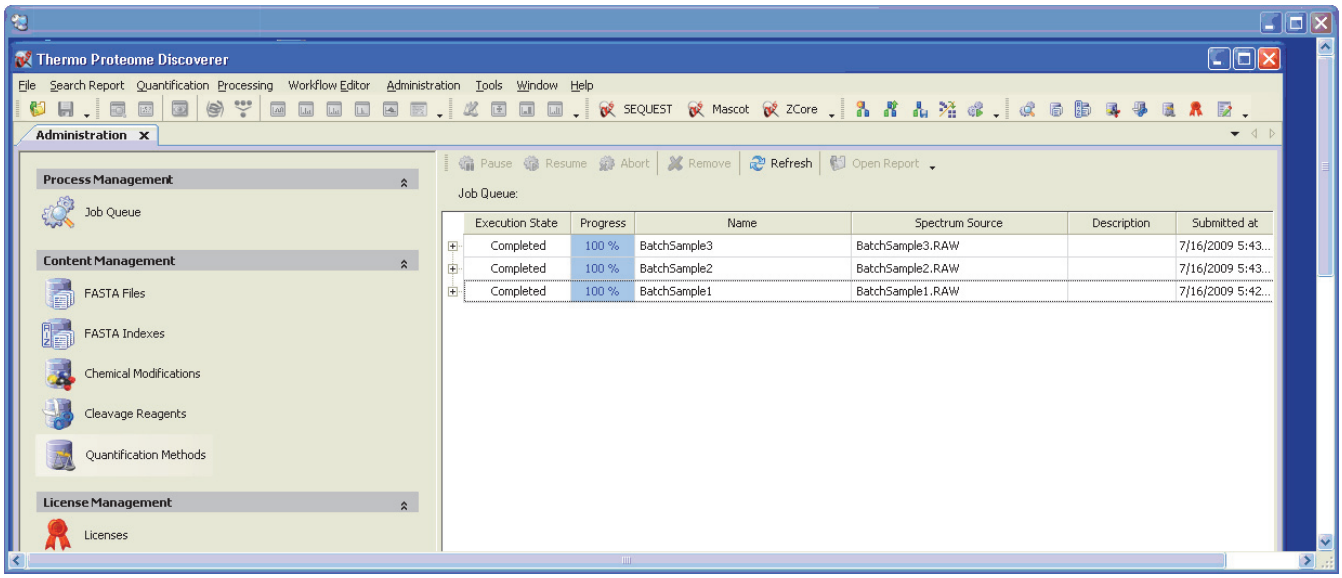
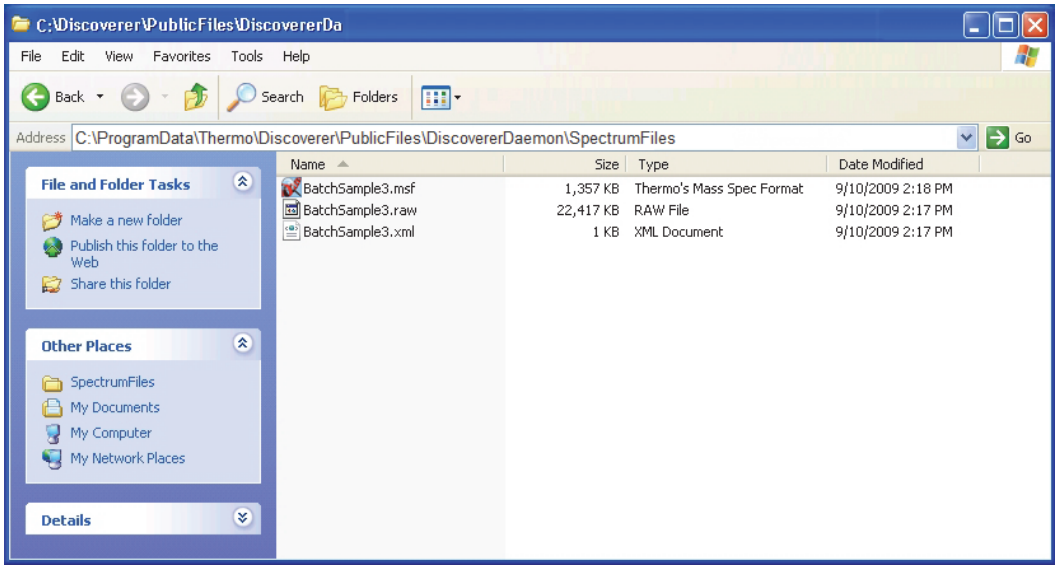


Figure 68. Storing the data in the Public Files directory



Processing MudPIT Samples by Using a Processing Method

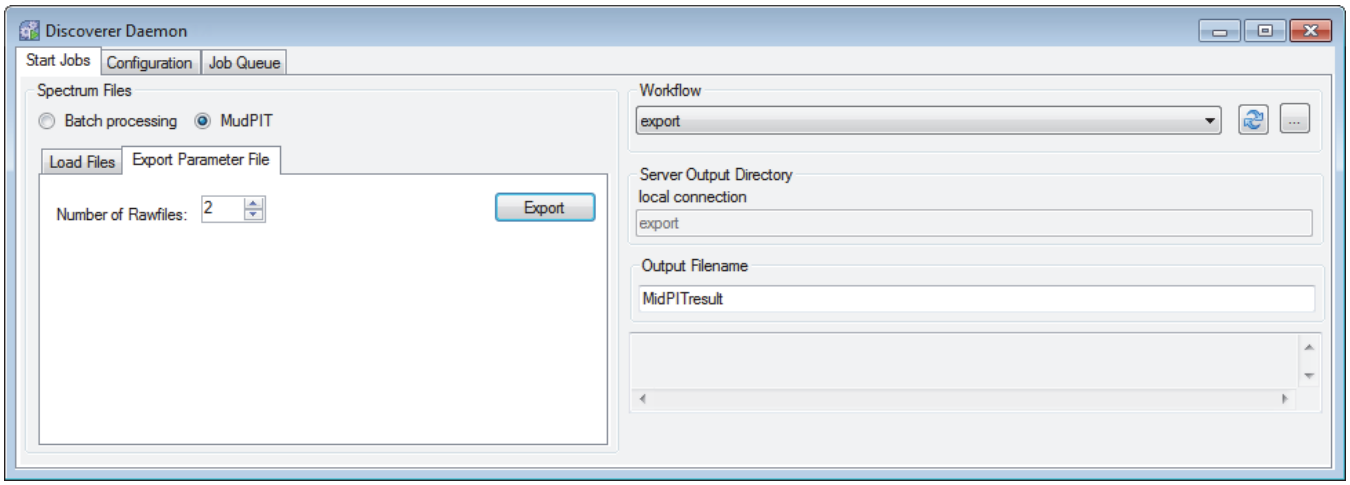
You can process MudPIT samples by using the Quantification Method Editor.

❖ To process MudPIT samples

1. Start the Proteome Discoverer Daemon application and export a parameter file for MudPIT processing. For information about exporting a parameter file, see "Creating a Parameter File That the Discoverer Daemon Application Uses" on page 81.

   Figure 69 shows how to configure the Export Parameter File page in the Proteome Discoverer Daemon application to export a parameter file. In the following example, the parameter file is saved in C:\Xcalibur\methods.

Figure 69. Selecting MudPIT processing on the Start Jobs page



   This example features two MudPIT samples, and each one is composed of two raw data files (for a total of four raw data files).

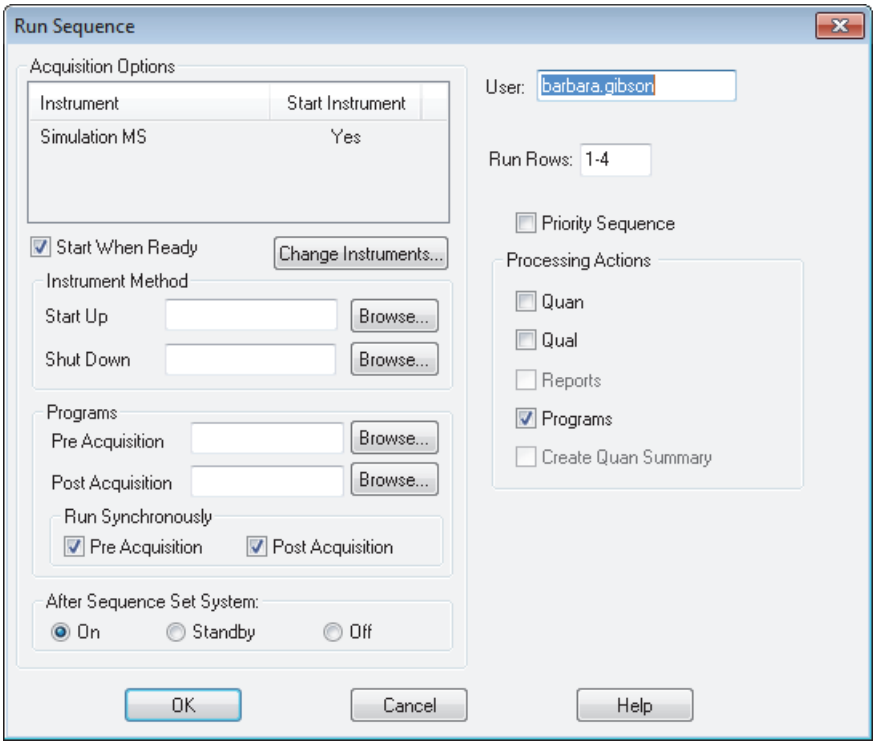
2. Define a processing method (see "Creating a Processing Method That Calls the Discoverer Daemon Application" on page 82) using the parameter file exported in step 1, and select the method as the processing method in the Proc Meth column, as shown in Figure 70.

Figure 70. Sequence used for MudPIT processing

	Sample Type	File Name	Sample ID	Path	Inst Meth	Proc Meth	Position	Inj Vol	Level
1	Unknown	BSA1	1	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\mudPIT	1	10.00	
2	Unknown	BSA2	2	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\mudPIT	2	10.00	
3	Unknown	BSA3	3	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\mudPIT	3	10.00	
4	Unknown	BSA4	3	C:\Projects\Daemon	C:\Xcalibur\methods\Daemon	C:\Xcalibur\methods\mudPIT	3	10.00	
*								0.00	

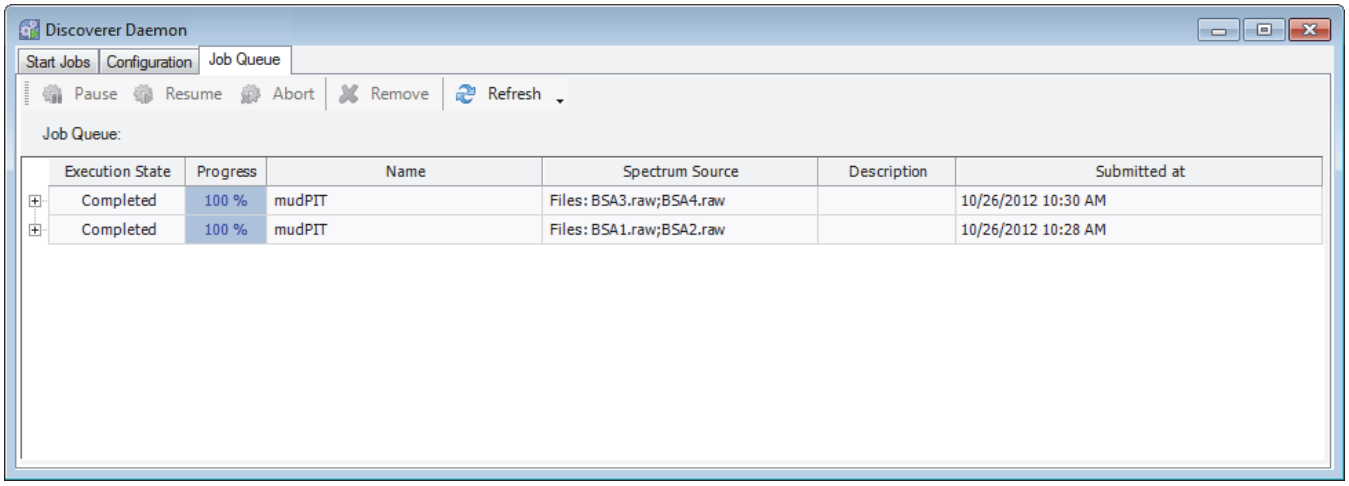
3. Start processing the MudPIT samples in the Run Sequence dialog box, as shown in Figure 71.

Figure 71. Starting the processing of the MudPIT samples



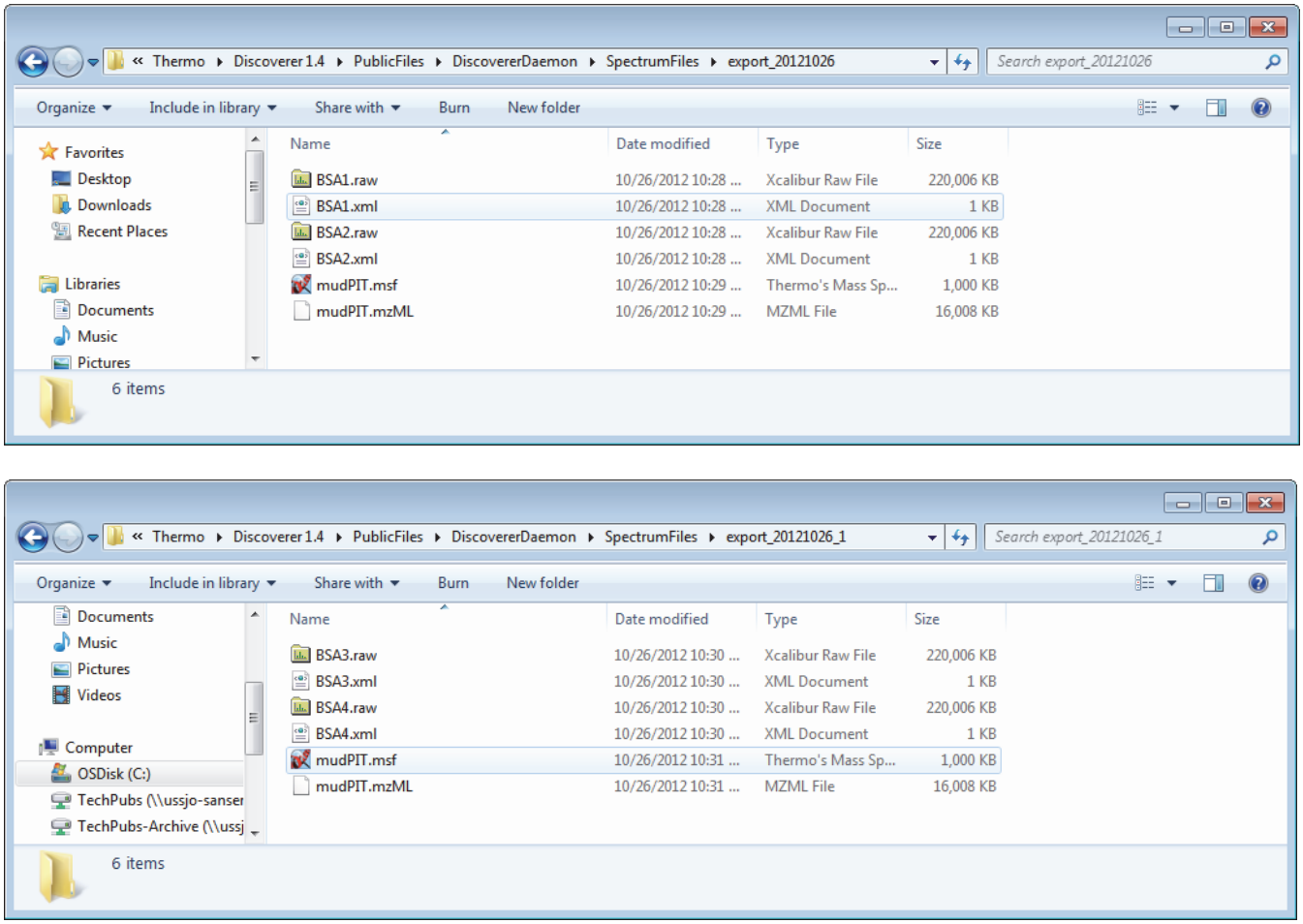
The Proteome Discoverer application processes the two samples as MudPIT, as shown in Figure 72.

Figure 72. Processing two MudPIT samples in the Proteome Discoverer application



The Proteome Discoverer application saves the data in the two MudPIT samples in two directories, each one containing the raw data files of one MudPIT sample (in this example, two raw data files), as shown in Figure 73.

Figure 73. Saving the raw data files of each MudPIT group in two directories



MudPIT Processing Using the Run Sequence Dialog Box

Running MudPIT samples using the Run Sequence dialog box is similar to the batch processing described in "Batch Processing by Using a Post-Acquisition Method (Xcalibur Data System 2.0.7 Only)" on page 89. Replace the batchprocessing.param file with a parameter file for MudPIT.

You can use the Proteome Discoverer Daemon application to export raw files to MGF, MZDATA, DTA, MZXML, and MZML files. To export files, use a workflow that includes the Spectrum Files, Spectrum Selector, and Spectrum Exporter nodes. Set the appropriate file type in the Spectrum Exporter node. In batch processing, the Proteome Discoverer Daemon application exports all the raw files with the file name of the spectrum.

Running the Proteome Discoverer Daemon Application on the Command Line

You can run the Proteome Discoverer Daemon application on the command line or in an interface window.

❖ **To run the Proteome Discoverer Daemon application on the command line**

1. Open a command shell and use the cd command to move to **Program Files > Thermo > Discoverer > System > Release**.
2. Type **DiscovererDaemon** and any of the following options on the command line:

```
DiscovererDaemon
    [-e foldername FileCount Workflow ParameterAssignment]
    [-c foldername]
    [-a foldername SpectrumFile]
    [-h]
    [-l serverName userName]
    [-r outputFilename]
    [-p parameterFile rawFile]
    [-f foldername]
```

Syntax

The Discoverer Daemon command-line syntax includes the following parameters:

- `[-e foldername FileCount Workflow ParameterAssignment]`

  Executes the workflow on the server using these specified parameters:

  – *foldername* : Specifies the location where the raw files are stored. You can give it any name, for example, RawFiles or Fractions.

  – *FileCount*: Specifies the number of spectrum files that must be included before the workflow is executed. This parameter is intended to be used with MudPIT experiments and acquisition on several machines. If the workflow should be executed regardless of the number of files contained in the file collection, use ANY instead of a number.

  – *Workflow*: Specifies the name of the template file containing the workflow in .xml format. You must have created this workflow template file in the Proteome Discoverer application by choosing Workflow Editor > Export Workflow to XML.

  – *ParameterAssignment*: Specifies the name and value of a parameter in the format of *parameter=value*. Some examples follow.

    This example sets the FASTA database for any node to equine.fasta:

    ```
    FastaDatabase=equine.fasta
    ```

The next example sets the FASTA database for all Mascot nodes to equine.fasta:

```
Mascot.FastaDatabase=equine.fasta
```

The last example sets the FASTA database for Mascot nodes having 4 as the processing node number to equine.fasta. It is equivalent to [4].FastaDatabase=equine.fasta because the processing node numbers are unique.

```
Mascot[4].FastaDatabase=equine.fasta
```

- `[-c foldername]`
  - Remote server: Creates a user-named folder in the PublicFiles folder on the server where you store output files. The PublicFiles folder is the default file in the Current File Directory box in the view displayed in the Proteome Discoverer application when you select Administration > Configuration > Server Settings > Discoverer Daemon. The -c option automatically appends the date and, if the directory already exists, an incremental index number to the name.

    You can only create a folder in the directory configured in the view opened by the Administration > Configuration > Server Settings > Discoverer Daemon command on the remote server. If you attempt to create a file other than in the PublicFiles folder in the Current File Directory box, Discoverer Daemon issues a message informing you that the Proteome Discoverer application will apply the change the next time that you start it.

    This option performs the same function as the -f *foldername* option, except that you can use the name of the folder more than once. When you use the name more than once, the Proteome Discoverer application appends the date and an incremental index number to the name.
  - Local server: Does nothing.
- `[-a foldername SpectrumFile]`
  - Remote server: Uploads the spectrum file to the location specified on the configured server. *SpectrumFile* is the name of the spectrum file.
  - Local server: Does nothing.
- `[-h]`: Lists the options available with the Thermo.Magellan.DiscovererDaemon command.
- `[-l serverName userName]`: Connects Discover Daemon to the specified local or remote host machine.
  - *serverName*: Specifies the name of the local or remote host.
  - *userName*: Specifies the name to log on.
- `[-r outputFilename]`: Specifies the name of the output file. You must use this option with the `-e` option, as in this example:

  ```
  DiscovererDaemon -e sfcid any mascot3.xml -r silac1noMT_AS4DE.msf.
  ```

- [-p *parameterFile rawFile*]: Processes the specified raw data file with all the parameters given in the parameter file, including the connection to the server.
  - *parameterFile*: Specifies the name of the parameter file.
  - *rawFile*: Specifies the name of the raw file.

  In the following example of the -p syntax, the Proteome Discoverer Daemon application processes the 9mix_LysC_monolith.raw file with the parameters given in the parameter file called c:\Xcalibur\methods\batchprocessing.param.

  ```
  DiscovererDaemon -p C:\Xcalibur\methods\batchprocessing.param
  9mix_LysC_monolith.raw
  ```

- [-f *foldername*]: On a remote server, this option creates a user-named folder in the PublicFiles folder of the server where the local version of the raw file and the result files are stored. If the directory already exists, the Proteome Discoverer Daemon application issues an error message, and the process returns with exit code -1 (standard exit code 0).

  If you attempt to create a file other than in the PublicFiles folder in the Current File Directory box, Discoverer Daemon issues a message informing you that the Proteome Discoverer application will apply the change the next time that you start it.

  This option performs the same function as the -c *foldername* option, except that you cannot use the name of the folder more than once.

  On a local server, this option does nothing.

Examples

The following are some examples of the Proteome Discoverer Daemon command-line syntax.

This example constructs the spectrum file collection called Rawfiles, adds the TrypMyo.raw file to the collection, and executes the SequestEquine workflow using the raw file in the Rawfiles directory:

```
DiscovererDaemon -c Rawfiles -a Rawfiles
c:\Rawfiles\TrypMyo.raw -e Rawfiles ANY c:\Workflows\SequestEquine.xml
```

In the following example, the Proteome Discoverer Daemon application evaluates several fractions in a single workflow:

```
DiscovererDaemon -c Fractions
DiscovererDaemon -a Fractions c:\rawfiles\fraction1.raw
DiscovererDaemon -a Fractions c:\rawfiles\fraction2.raw
DiscovererDaemon -a Fractions c:\rawfiles\fractionN.raw
DiscovererDaemon -e Fractions ANY c:\wfs\fractions.xml
```

The next example demonstrates that you can start several workflows with one invocation of the Proteome Discoverer Daemon application.

```
DiscovererDaemon
```

```
-c RawFile
-a RawFile c:\Rawfiles\TrypMyo.raw
-e RawFile ANY c:\wfs\SequestEquine.xml
-c RawFile
-a RawFile c:\Rawfiles\BSADigest.raw
-e RawFile ANY c:\Workflows\SequestEquine.xml
```

The following example runs the Proteome Discoverer Daemon application on a remote host called protlab2, uploads the iTRA_BSA_3ITMS2_3HCD.raw spectrum file to the server, executes the workflow in c:\Workflows\MascotEcoli.xml:

```
DiscovererDaemon -l protlab2 leo_davinci -c sfcid -a sfcid
iTRA_BSA_3ITMS2_3HCD.raw -e sfcid any c:\Workflows\MascotEcoli.xml
```

The following sequence of commands submits multiple raw files for processing on a remote server:

```
DiscovererDaemon.exe -c AllTrypMyo
DiscovererDaemon.exe -a AllTrypMyo_020110303 C:\DaemonTest\mudpit4\Tryp_Myo.raw
DiscovererDaemon.exe -a AllTrypMyo_020110303 C:\DaemonTest\mudpit4\Tryp_Myo_1.raw
DiscovererDaemon.exe -a AllTrypMyo_020110303 C:\DaemonTest\mudpit4\Tryp_Myo_2.raw
DiscovererDaemon.exe -e AllTrypMyo_020110303 3 C:\DaemonTest\mudpit4\wf_sequest.xml
```

The next sequence of commands submits multiple raw files for processing on a local server:

```
DiscovererDaemon.exe -a AllTrypMyo C:\DaemonTest\mudpit4\Tryp_Myo.raw
DiscovererDaemon.exe -a AllTrypMyo C:\DaemonTest\mudpit4\Tryp_Myo_1.raw
DiscovererDaemon.exe -a AllTrypMyo C:\DaemonTest\mudpit4\Tryp_Myo_2.raw
DiscovererDaemon.exe -e AllTrypMyo 3 C:\DaemonTest\mudpit4\wf_sequest.xml
```

The Discoverer Daemon appends a time stamp to each file when it processes the files on a remote server.

Searching for Data

This chapter describes the features that you can use when searching for and analyzing data in the Proteome Discoverer application.

Contents

Using FASTA Databases

You can use the FASTA database utilities to add, delete, and find protein references and sequences. You can also extract information from an existing FASTA file, place it into a new FASTA file, and compile it for availability in the Proteome Discoverer application.

For more information about FASTA databases, see "FASTA Reference" on page 339.

Displaying FASTA Files

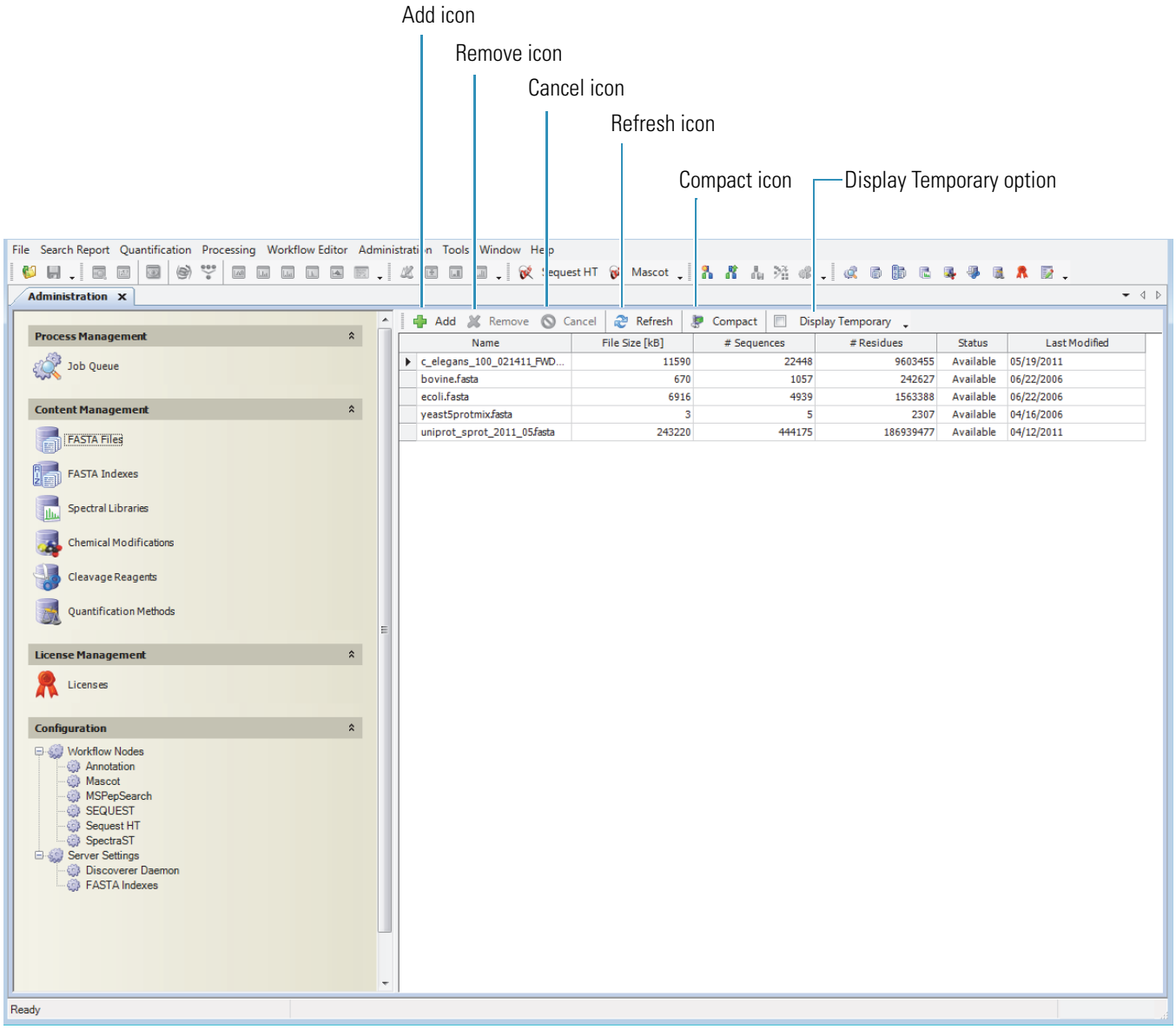
You can list all the FASTA files that you have downloaded from other sources onto your hard drive and registered.

❖ **To list the available FASTA files**

- Choose **Administration > Maintain FASTA Files** or click the **Maintain FASTA Files** icon,  , either in the toolbar or on the Administration page.

The FASTA files view shown in Figure 74 appears. It lists all the FASTA files that you have downloaded from other sources and registered. It displays the processed FASTA file properties, such as the file name, file size, and the number of proteins stored. The Proteome Discoverer application analyzes each protein entry to determine if the FASTA file meets the application requirements for use in a spectra search. It processes the FASTA file and makes it available for use.

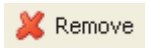
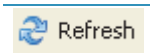
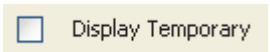
Figure 74. FASTA files view



FASTA Files View Parameters

Table 3 describes the options and columns in the FASTA files view in the Proteome Discoverer application.

Table 3. Options and columns in the FASTA files view

Parameter	Description
	Activates the Open dialog box, so you can choose the FASTA database to import.
	Deletes a FASTA database from the FASTA files view.
	Cancels the addition or removal of a FASTA file.
	Redisplays the view on the screen.
	Releases the storage space previously occupied by proteins that were imported from FASTA files and inserted during a Mascot search but subsequently deleted.
	Displays FASTA files that contain the proteins found by a Mascot search. The Proteome Discoverer application temporarily imports these FASTA files, which are not available for Sequest searches.
Name	Displays the name of the FASTA file.
Size [kB]	Displays the current size of the FASTA file.
#Sequences	Displays the number of sequences found in the FASTA file during processing.
#Residues	Displays the number of amino acids found in the FASTA file during processing.
Status	Displays the current status of the FASTA file: • Imported: Indicates that the FASTA file has been downloaded from a source and registered. • Available: Indicates that the FASTA file is available for Sequest searches. • Processing: Indicates that the FASTA file is in the process of being registered.
Last Modified	Displays the date when the FASTA file was last modified or created.

Adding FASTA Files

You must add a FASTA file to the Proteome Discoverer application before you can conduct a search with Sequest.

❖ To add a FASTA file

1. Choose **Administration > Maintain FASTA Files** or click the **Maintain FASTA Files** icon, .

   The Administration page appears with the FASTA files view, shown in Figure 74 on page 102.

2. Click Add.

3. In the Open dialog box that appears, browse for and select the FASTA file that you want to process, and then click **Open**.

   The FASTA file that you selected appears as a job in the job queue. To cancel the addition of this file, click Abort.

   When you see the Completed in the Execution State column, the database has finished downloading.

4. To add another FASTA file, wait until the Execution State column indicates that the addition of the FASTA file is completed, click **FASTA Files** in the left pane of the Administration page under Content Management, and then click **Add** to add the next file.

The amount of time that it takes to process a FASTA file depends on the file size. When a FASTA file finishes processing, the Status column displays the Available status. The FASTA file is now available to use for a protein or peptide search with the Proteome Discoverer application.

Deleting FASTA Files

You can delete a FASTA file from the application.

❖ To delete a FASTA file

1. Choose **Administration > Maintain FASTA file**.

   The Administration page appears with the FASTA files view, shown in Figure 74 on page 102.

2. Click at the beginning of a row to select the row.

3. Click Remove.

4. In the Remove FASTA databases dialog box, click **OK**.

The FASTA file that you selected appears as a job in the job queue. After you start the deletion of the file, you cannot cancel the deletion. You can remove the completed job from the job queue by clicking Remove and then clicking OK in the Delete Jobs dialog box.

Compressing a Protein Database

A protein database contains the proteins of imported FASTA files. It also contains proteins found during a Mascot search that are inserted into the database. When you remove a FASTA file from the database by using the FASTA file manager, it automatically deletes protein entries but does not make the storage space available. Although following this next procedure can explicitly make the storage space available, it can be time-consuming for large databases.

❖ To compress a protein database

1. Choose **Administration > Maintain FASTA file**.

   The Administration page appears with the FASTA files view.

2. Click Compact .

   A message informs you that compressing the protein database can take a long time.

3. To continue with the database compression, click **OK** in the message box.

   A job starts and appears in the job queue. Before you start the job, you can remove it if necessary. However, you cannot cancel the job, and it will restart automatically if you shut down the Proteome Discoverer application during job execution.

Displaying Temporary FASTA Files

The Proteome Discoverer application temporarily imports FASTA files that contain the proteins found by a Mascot search, but these files are not available for Sequest searches. You can optionally display these files in the FASTA files view.

❖ To display temporary FASTA files

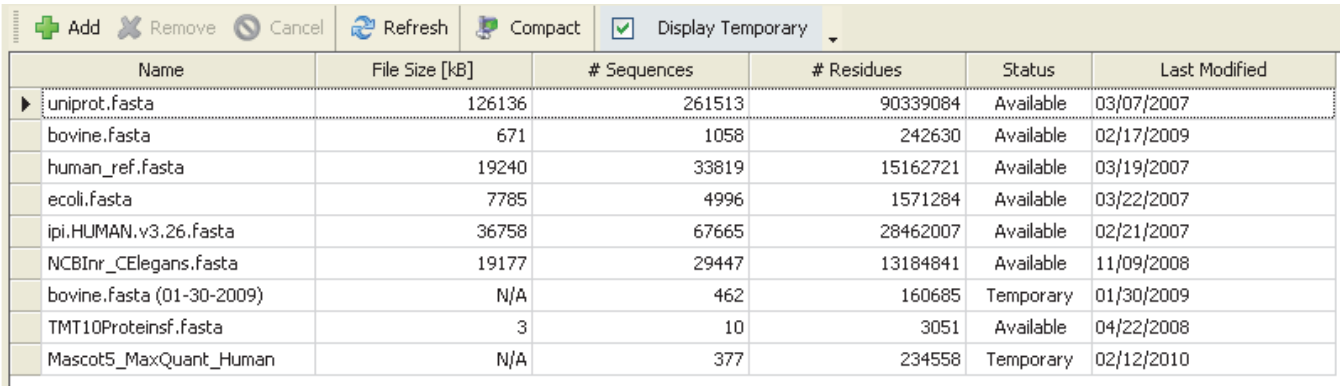
1. Choose **Administration > Maintain FASTA file**.

   The Administration page appears with the FASTA files view, shown in Figure 74 on page 102.

2. Select the **Display Temporary** check box, Display Temporary .

   You now see any temporary FASTA files; for example, Figure 75 shows Temporary for two files in the Status column.

Figure 75. Displaying temporary FASTA files



Name	File Size [kB]	# Sequences	# Residues	Status	Last Modified
uniprot.fasta	126136	261513	90339084	Available	03/07/2007
bovine.fasta	671	1058	242630	Available	02/17/2009
human_ref.fasta	19240	33819	15162721	Available	03/19/2007
ecoli.fasta	7785	4996	1571284	Available	03/22/2007
ipi.HUMAN.v3.26.fasta	36758	67665	28462007	Available	02/21/2007
NCBInr_CElegans.fasta	19177	29447	13184841	Available	11/09/2008
bovine.fasta (01-30-2009)	N/A	462	160685	Temporary	01/30/2009
TMT10Proteinsf.fasta	3	10	3051	Available	04/22/2008
Mascot5_MaxQuant_Human	N/A	377	234558	Temporary	02/12/2010

Adding a Protein Sequence and Reference to a FASTA Database File

You can add a protein sequence and a protein reference to a registered FASTA database file. The protein sequence refers to the sequence of amino acids that constitute the protein, and the protein reference refers to the name or reference of the protein.

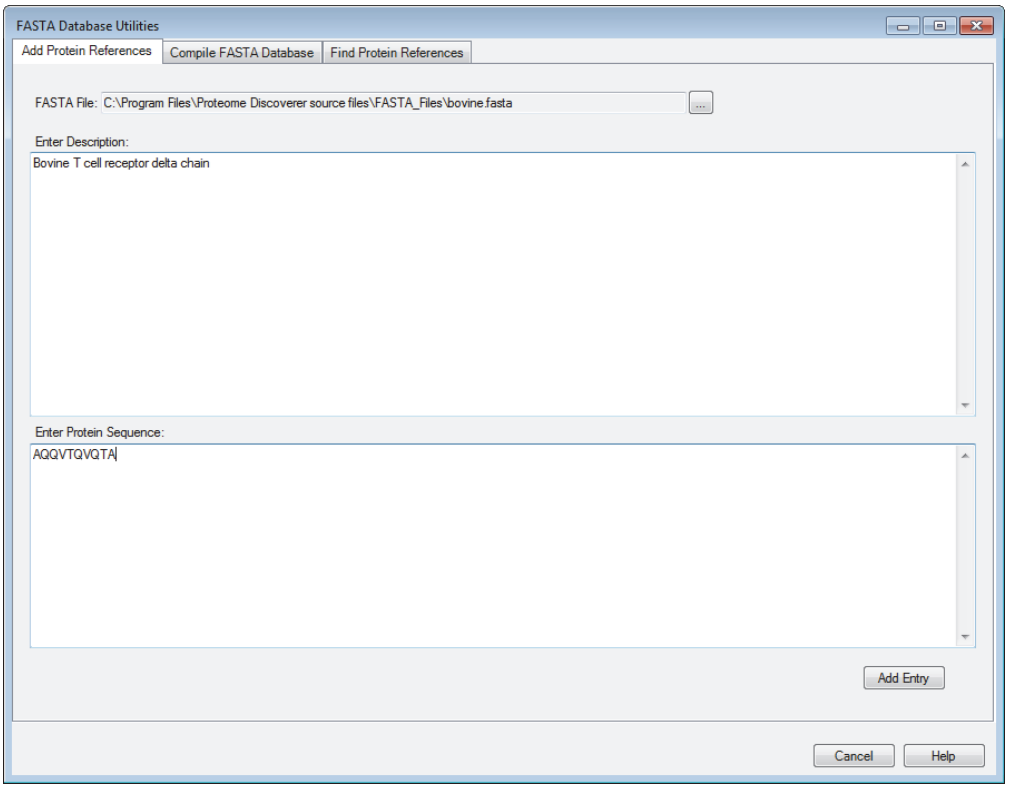
❖ To add a protein sequence and reference

1. Choose **Tools > FASTA Database Utilities**.
2. In the FASTA Database Utilities dialog box, click the **Add Protein References** tab.

   The Add Protein References page of the dialog box appears.
3. Click the **Browse** button (...) next to the FASTA File box.
4. In the Save/Add to FASTA File dialog box, select the FASTA database that you want to add the protein sequence and reference to, and click **Save**.
5. In the Enter Description box of the FASTA Database Utilities dialog box, type a description of the protein sequence that you are adding.
6. In the Enter Protein Sequence box, type the protein sequence that you want to add to the FASTA database.

   The Add Protein References page should resemble the illustration in Figure 76.

Figure 76. Add Protein References page of the FASTA Database Utilities dialog box



7. Click **Add Entry** to add the protein sequence.

Finding Protein Sequences and References

You can find a protein sequence or reference in an existing FASTA database file.

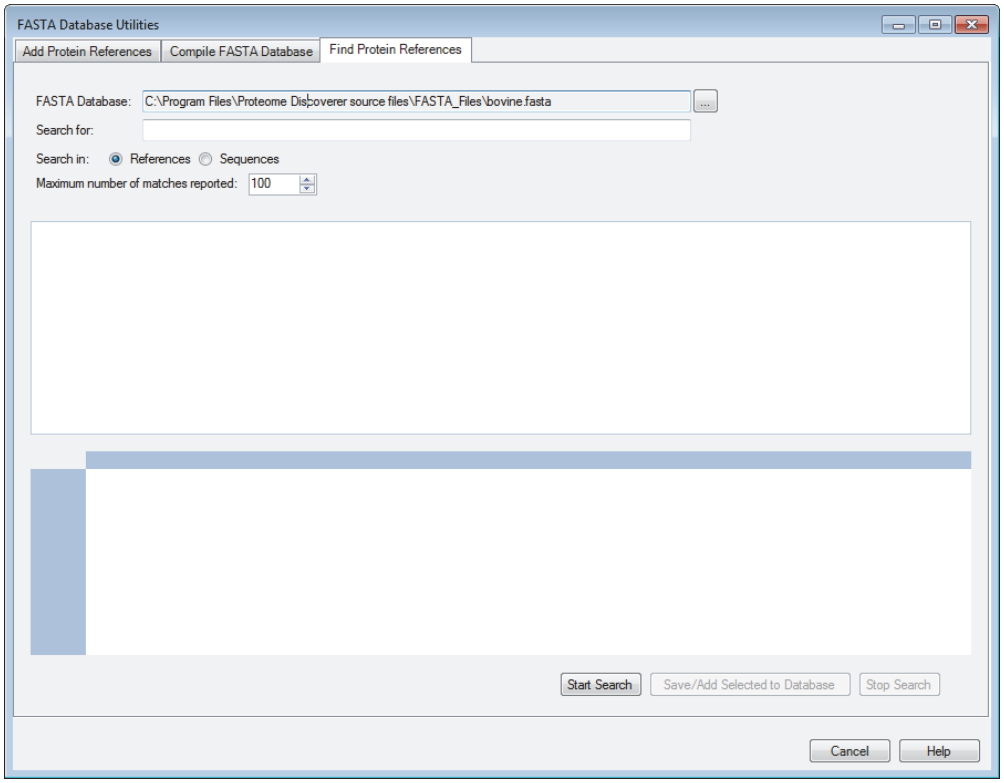
- To find a protein sequence or reference
- To filter a protein reference search
- To refine a filtered protein reference search
- To delete conditions in filtered protein reference searches

❖ To find a protein sequence or reference

1. Choose **Tools > FASTA Database Utilities**.
2. In the FASTA Database Utilities dialog box, click the **Find Protein References** tab.

   The Find Protein References page appears, as shown in Figure 77.

Figure 77. Find Protein References page of the FASTA Database Utilities dialog box



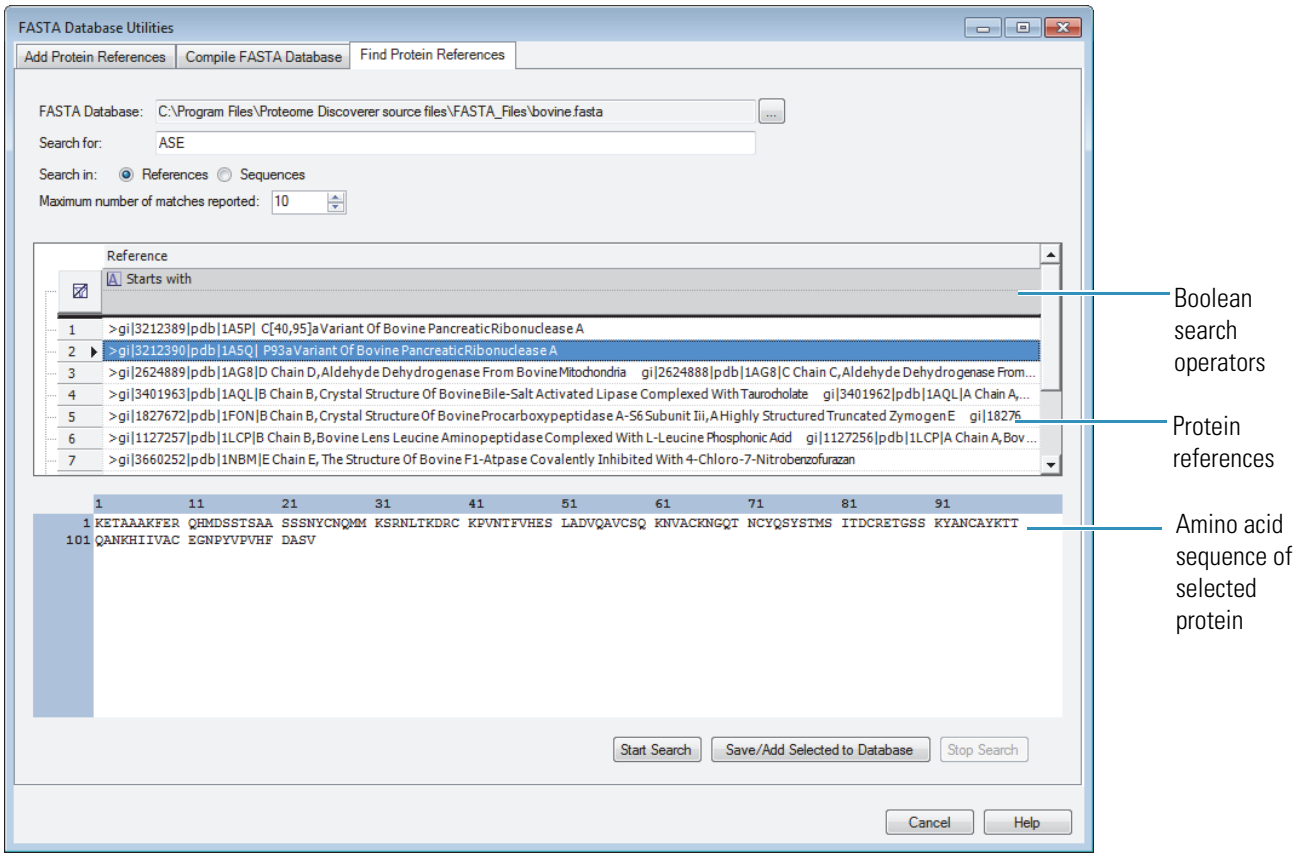
3. Click the **Browse** button (...) next to the FASTA Database box to locate the FASTA file of interest.
4. In the Please Select a FASTA Database dialog box, select the FASTA file, and click **Open**.
5. In the Search For box of the Find Protein References page, type an amino acid sequence or a protein reference search string.
6. In the Search In area, specify whether the Proteome Discoverer application should search for the search string in the protein references or sequences.
   - References: Searches for the search string in the protein references.
   - Sequences: Searches for the specified amino acid sequence within the protein sequences.

   You can further refine the results by using filters either before or after you run the search. For instructions on filtering, see "To filter a protein reference search" on page 109.
7. In the Maximum Number of Matches Reported box, select the maximum number of references or sequences to report.
8. Click **Start Search**.

   Results appear if the search parameters match the data, as shown in Figure 78. Click a protein row to see the amino acid sequences that constitute that protein.

9. To suspend the search, click **Stop Search**.

Figure 78. Find Protein References page in the FASTA Database Utilities dialog box



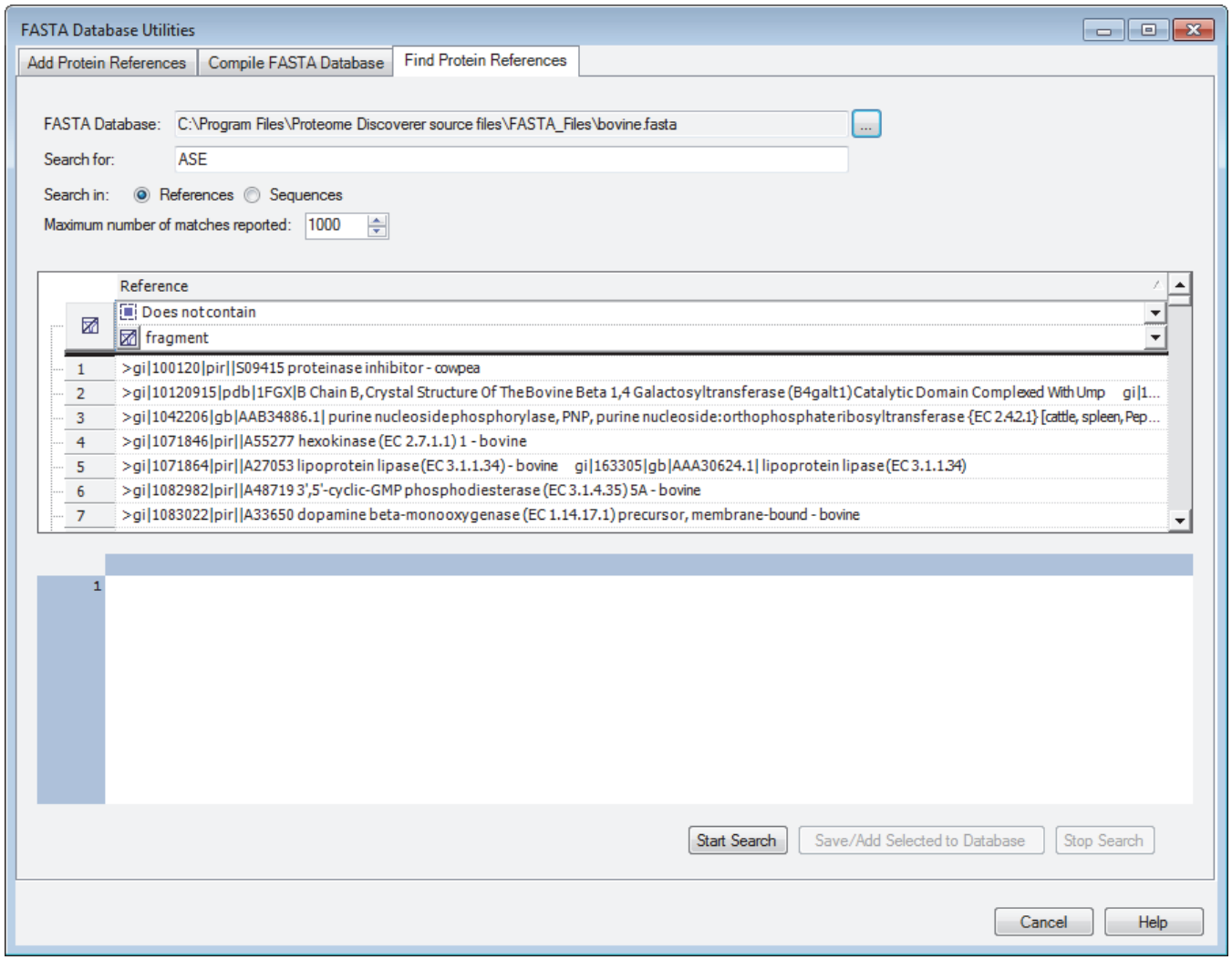
10. (Optional) To save a protein result row in another FASTA database, select the protein row, click **Save/Add Selected to Database**, select the database in the Save/Add to FASTA File dialog box, and click **Save**.

❖ To filter a protein reference search

1. On the Find Protein References page of the FASTA Database Utilities dialog box, click the line below "Reference" in the middle of the page to access a list of operators that you can use to filter the references. (The default operator is "Starts with.") For a list of all operators, refer to the Help.
2. In the line below the operator that you selected, type the search string or condition that you want the operator to apply to.

   The example in Figure 79 filters out those protein references that contain "fragment."

Figure 79. Filtering out protein references containing “fragment”

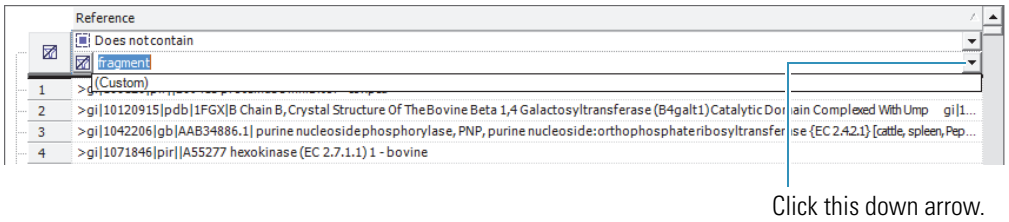


❖ To refine a filtered protein reference search

1. Select the **Custom** option from the list in the line below the search operator.

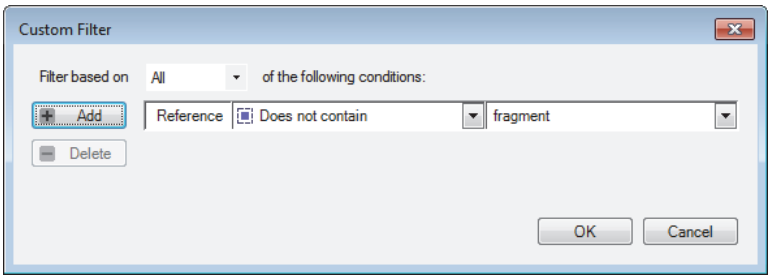
   To make the Custom option available, click the down arrow in the line below the operator, as shown in Figure 80.

   Figure 80. Selecting the Custom option



The Custom option opens the Custom Filter dialog box, shown in Figure 81, so you can add multiple conditions.

Figure 81. Custom Filter dialog box



2. Click **Add**.

   A new line appears in the Operator (left) and Operand (right) lists.

3. Select an operator from the Operator list.

4. Type an operand on the line in the Operand column.

5. In the Filter Based On list, do one of the following:

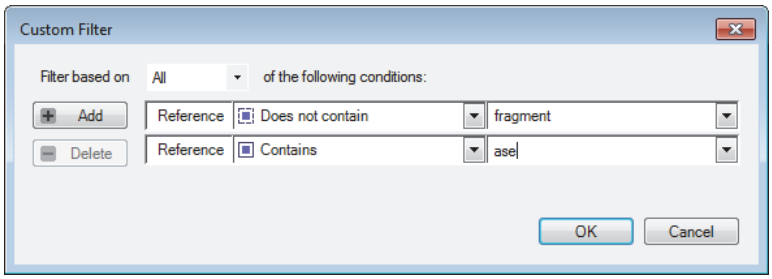
   Select the **All** option to indicate whether the search algorithm should search for protein references that meet both conditions.

–or–

   Select the **Any** option to indicate whether the search algorithm should search for protein references that meet only one of the conditions.

   Figure 82 gives an example of a search for protein references that meet both of the conditions.

Figure 82. Specifying two conditions



6. Click **OK**.

❖ To delete conditions in filtered protein reference searches

- To delete a condition in the Custom Filter dialog box, select the check box to the left of the appropriate condition in the Operator column, and click **Delete**.
- To delete the condition in the Reference area on the Find Proteins References page, click the **Clear Reference Filter Criteria** icon,  , in the line below the operator.

- To delete all conditions in both the Custom Filter dialog box and the Reference area on the Find Proteins References page, click the **Clear All Filter Criteria** icon,  , in the box to the left of the filters.

Compiling a FASTA Database

You can extract information from an existing FASTA file and place it into a new FASTA file, replace an existing FASTA file, or append it to an existing FASTA file. Then you must compile the new or changed FASTA file to make it available in the Proteome Discoverer application.

❖ To compile a FASTA database

1. Choose **Tools > FASTA Database Utilities**.
2. In the FASTA Database Utilities dialog box, click the **Compile FASTA Database** tab.

   The Compile FASTA Database page appears.
3. In the Original box, browse for the FASTA file that you are taking the information from, or type its path and name.
4. In the Please Select a FASTA Database dialog box, click **Open**.
5. In the Target box, browse for the FASTA file that you are placing the extracted information into, or type its path and name.
6. In the Save/Add to FASTA File dialog box, select the file, verify that the file extension is .fasta, and click **Save**.
7. In the Target Database Options area, select one of the following options to indicate what you want to do with the extracted information:
   - Create/Replace: Creates a new FASTA file for storing the information or overwriting an existing FASTA file. This option is the default.
   - Append: Adds the extracted information to an existing FASTA file.
8. In the Search In area, specify whether the Proteome Discoverer application should search for the search string in the protein references or sequences.
   - References: Searches for the search string in the protein references.
   - Sequences: Searches for the specified amino acid sequence within the protein sequences.
9. To disregard the case of the information to be extracted, select the **Ignore Case of Reference Strings** check box.

10. Specify the information to be extracted:
    a. Click  above the Step 1: String(s) to Include box.

       A line enabling you to specify the first set of conditions appears in the box.

    b. Click the first line in the Select Operator column, and select the operator to apply to the information to be extracted. You can select from the following:

       - Starts With: Extracts information that begins with this string.
       - Does Not Start With: Extracts information that does not begin with this string.
       - Ends With: Extracts information that ends with this string.
       - Does Not End With: Extracts information that does not end with this string.
       - Contains: Extracts information that includes this string.
       - Does Not Contain: Extracts information that does not includes this string.

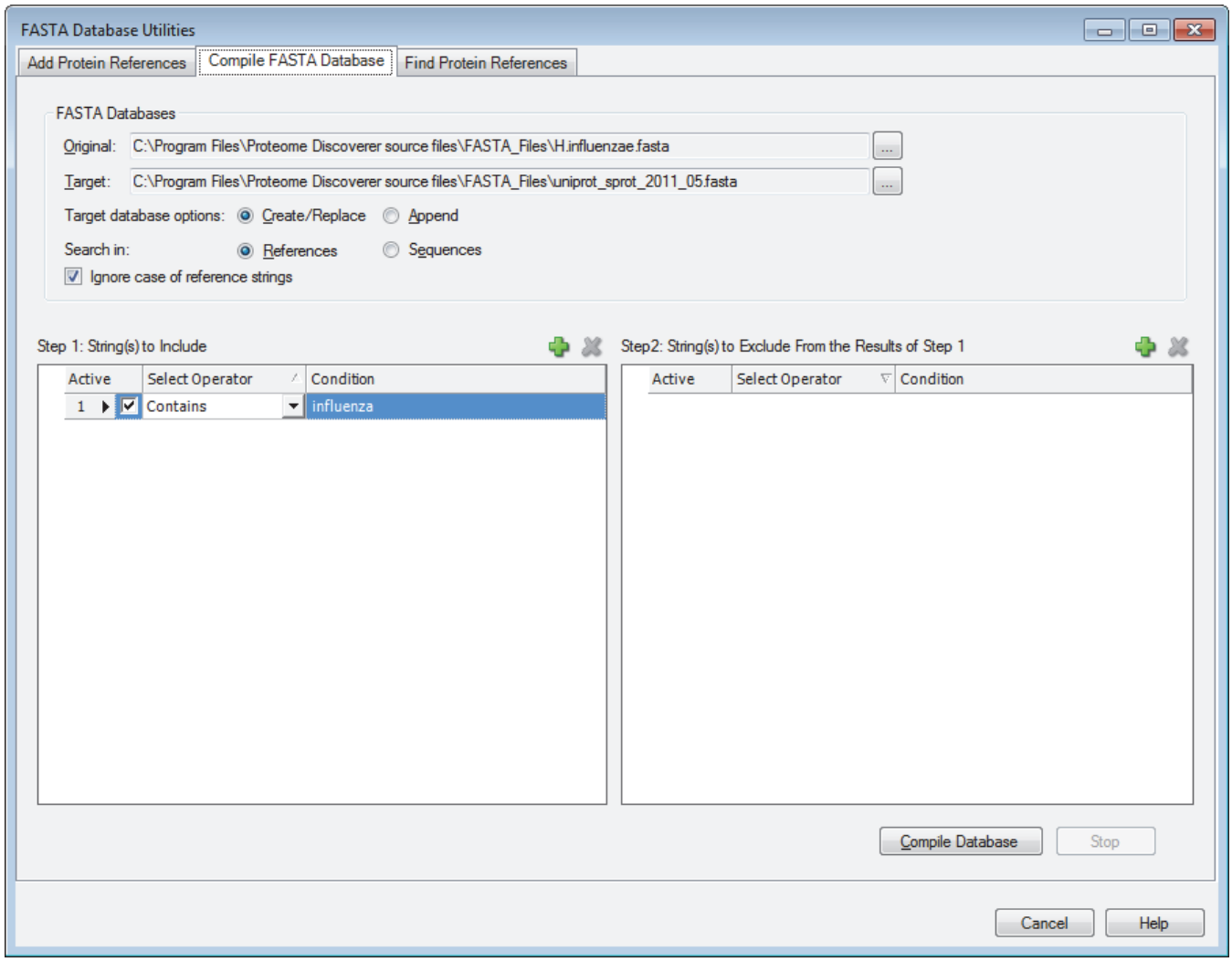
    c. Click the first line in the Condition column, and type the condition that the information must meet in order to be extracted.

    d. Repeat step a through step c to add more sets of conditions for the information to be extracted.

    e. To delete a set of conditions, in the Active column select the line that you want to delete and click  .

    The Compile FASTA Database page should now resemble the example in Figure 83.

Figure 83. Compile FASTA Database page of the FASTA Database Utilities dialog box



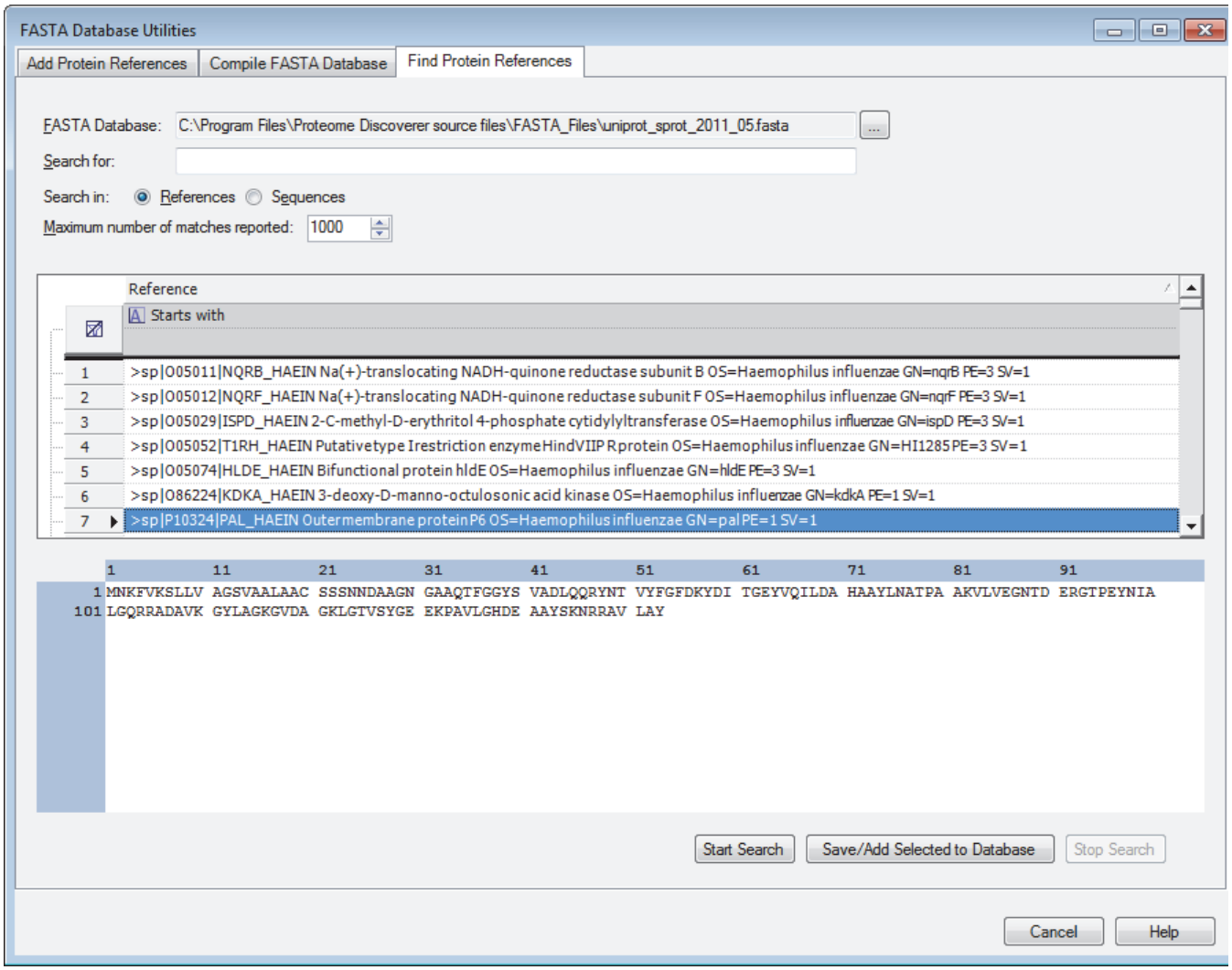
11. Click **Compile Database**.

    Click **Stop** if you want to halt the compilation.

12. After the compilation, click **Start Search** on the Find Protein References page to view the results of the extraction, as shown in the example in Figure 84.

    You do not have to enter information into the Search For box.

Figure 84. Results of search



13. (Optional) To specify any information that you want to exclude from the extracted results, follow these steps:

    a. Click above the Step 2: String(s) to Exclude From the Results of Step 1 box on the Compile FASTA Database page.

    A line enabling you to specify the first set of conditions now appears in the box.

    b. Click the first line in the Select Operator column, and select the operator to apply to the information from the list. You can choose from the following:

    - Starts With: Excludes information that begins with this string.
    - Does Not Start With: Excludes information that does not begin with this string.
    - Ends With: Excludes information that ends with this string.
    - Does Not End With: Excludes information that does not end with this string.

- Contains: Excludes information that includes this string.
- Does Not Contain: Excludes information that does not include this string.

c. Click the first line in the Condition column, and type the condition that the information must meet in order to be excluded.

d. Repeat step a through step c to add more sets of conditions for the information that you want to exclude.

e. To delete a set of conditions, in the Active column select the line that you want to delete and click  .

14. Click **Compile Database**.

15. Click **Start Search** on the Find Protein References page to view the results of the extraction, as shown in the example in Figure 84 on page 115.

    You do not have to enter information into the Search For box.

Excluding Individual Protein References and Sequences from a FASTA Database

You can exclude individual entries from a FASTA file.

❖ **To exclude individual protein references and sequences from a FASTA file**

1. Choose **Tools > FASTA Database Utilities**.
2. In the FASTA Database Utilities dialog box, click the **Compile FASTA Database** tab.
3. In the Original box, browse for the FASTA database that contains the protein that you want to remove, or type its path and name. In the Please Select a FASTA Database dialog box, click **Open**.
4. In the Target box, browse for the output FASTA file or type its path and name. In the Save/Add to FASTA File dialog box, select the file, verify that the file extension is .fasta, and click **Save**.
5. Select the **Ignore Case of References Strings** check box.
6. Click  above the Step 1: String(s) to Include box.

   A line enabling you to specify the first set of conditions now appears in the box.
7. Click the first line in the Select Operator column, and select **Contains**, if it is not already selected. Leave the first line in the Condition column blank.
8. Click  above the Step 2: String(s) to Exclude From the Results of Step 1 box.

   A line enabling you to specify the first set of conditions now appears in the box.
9. Click the first line in the Select Operator column, and select **Contains**.

10. In the first line of the Condition column, type the protein reference or sequence that you want to remove.
11. Click **Compile Database**.

    The compiling process creates the target FASTA file that excludes protein entries that match the condition.

Managing FASTA Indexes

A FASTA index is a type of lookup table containing masses, theoretical peptide sequences, and associated proteins, which minimizes search time. The index lists all possible amino acid sequences that can be produced when an enzyme digests a protein or peptide. The peptide fragments are listed by molecular weight. The index stores information about every nominal mass, every peptide that has that mass, every protein that contains this peptide, and the location of its protein description in the FASTA file. Rather than read all protein sequences from the FASTA file, digest them in silico with the specified enzyme, calculate the mass of each peptide, and compare it to the given precursor mass, the Proteome Discoverer application looks for the specific mass in the FASTA index and uses it to find the peptides that have this mass and the associated proteins that contain the peptides.

For full enzymatic searches, the Proteome Discoverer application automatically creates FASTA indexes as they are needed. It does not automatically create FASTA indexes during semi-enzymatic or no-enzyme searches because these searches usually consume a large amount of space on a computer's hard disk. However, you can manually create FASTA indexes for these types of searches.

- Specifying the Location and Number of FASTA Indexes Stored
- Displaying the FASTA Indexes View
- Specifying the Columns to Display
- Automatically Creating a FASTA Index
- Manually Creating FASTA Indexes
- Controlling Automatic FASTA Index Removal
- Deleting a FASTA Index
- Changing Number and Location of Stored FASTA Indexes
- Removing FASTA Indexes When a FASTA File Is Deleted

Specifying the Location and Number of FASTA Indexes Stored

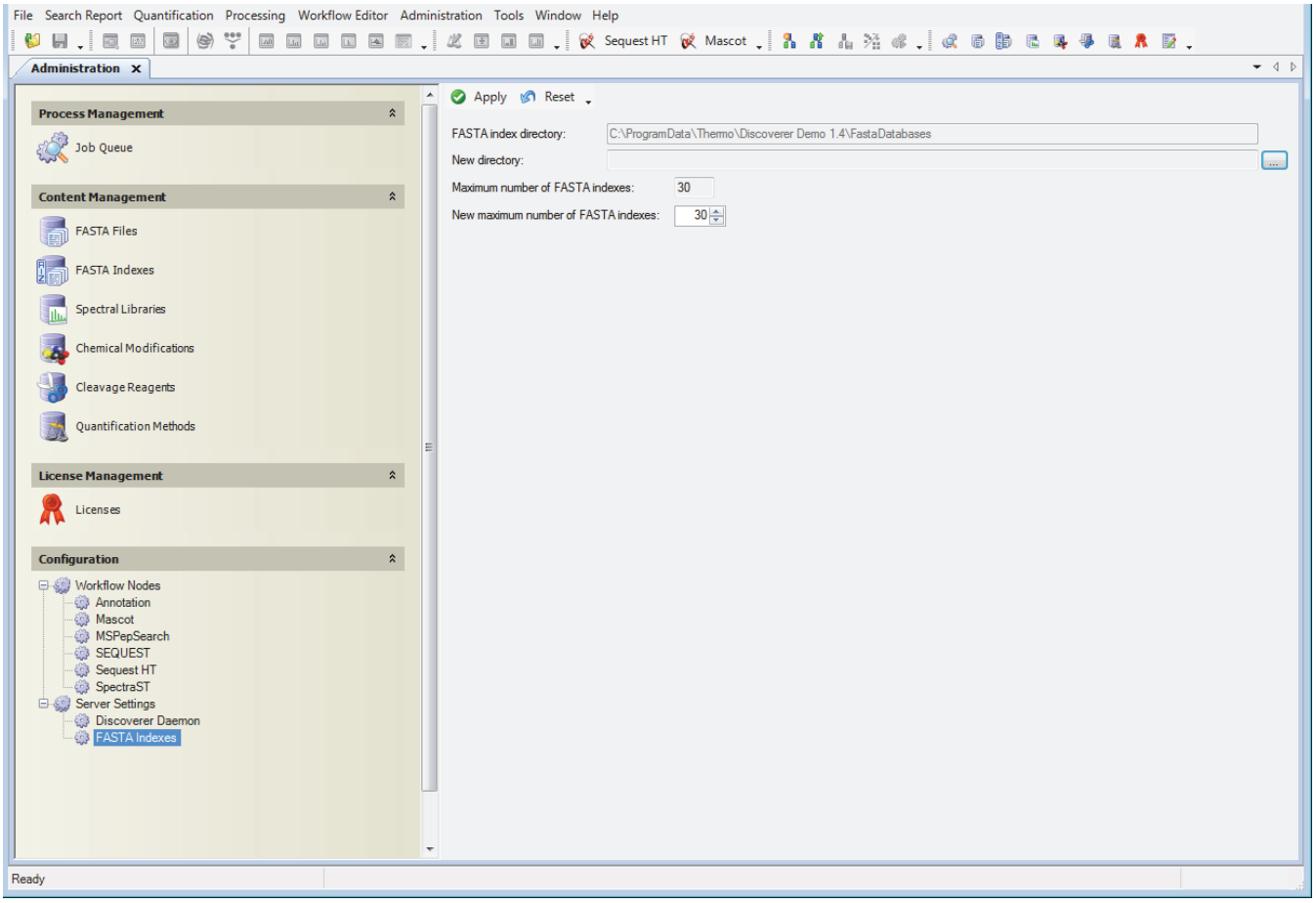
If you do not want to store the FASTA indexes in the default directory shown in Figure 85, you can specify an alternate directory in the FASTA Indexes configuration view. You can also change the maximum number of FASTA indexes stored.

❖ To specify the location and number of the FASTA indexes stored

1. Choose **Administration > Server Settings > FASTA Indexes**.

   The configuration view shown in Figure 85 appears.

Figure 85. FASTA Indexes configuration view



2. In the New Directory box, browse to the location of the folder to store the FASTA indexes in.

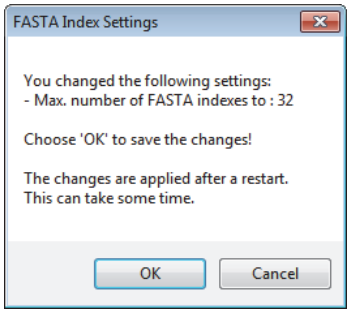
3. In the New Maximum Number of FASTA Indexes, box, select the maximum number of FASTA indexes to store.

   If you generate more FASTA indexes than the number to store in the New Maximum Number of FASTA Indexes box, the Proteome Discoverer application discards the difference from the oldest FASTA indexes the next time that you restart the application.

4. If you changed any settings, click Apply.

   A FASTA message box similar to that shown in Figure 86 appears.

Figure 86. Administration message box



5. Click **OK**.

Note Click Reset to return to the default values.

6. Restart your machine.

Displaying the FASTA Indexes View

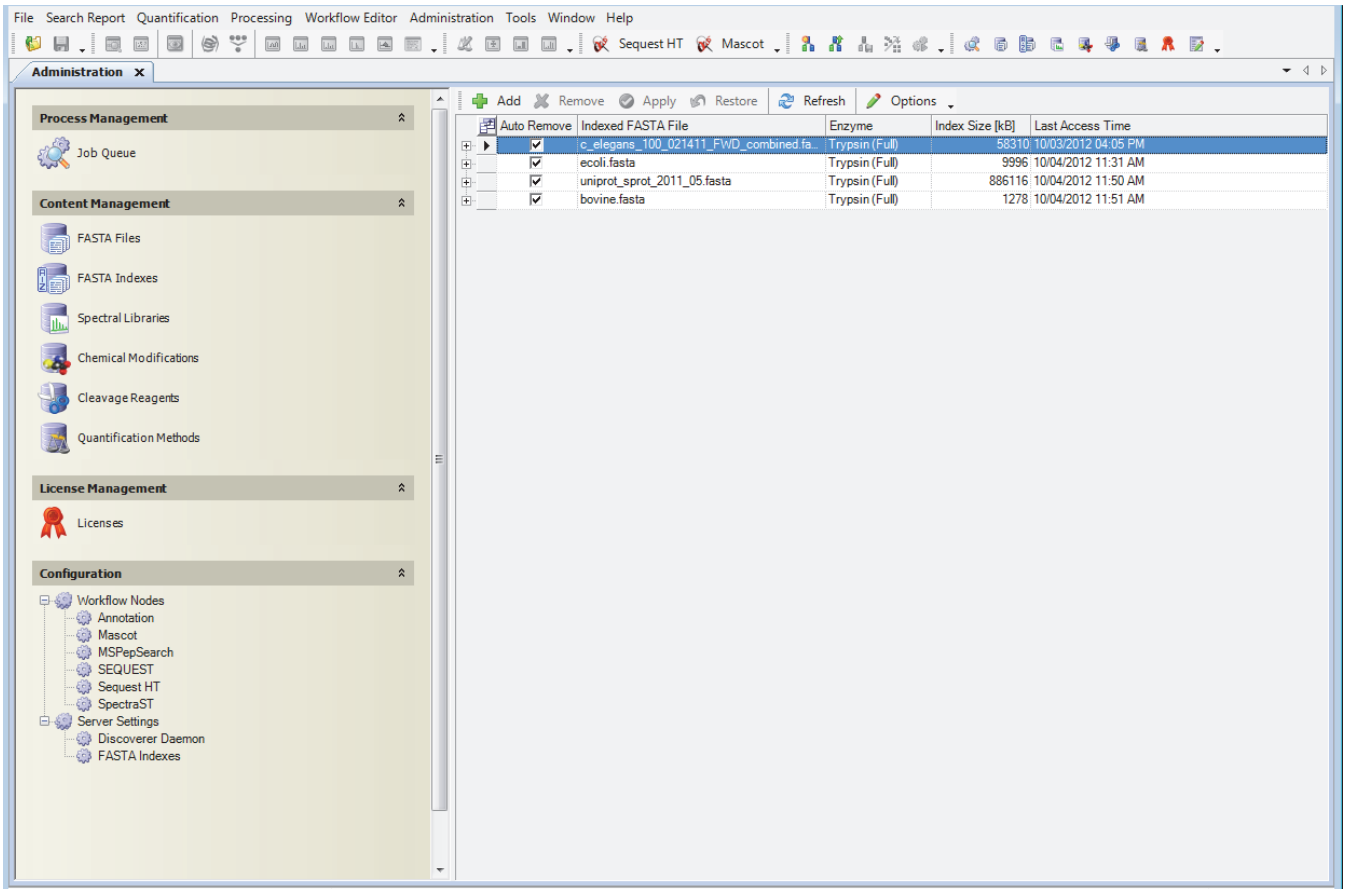
You can access FASTA indexes through the FASTA Indexes view.

❖ To display the FASTA Indexes view

1. Choose **Administration > Maintain FASTA Indexes** or click the **Maintain FASTA Indexes** icon, .

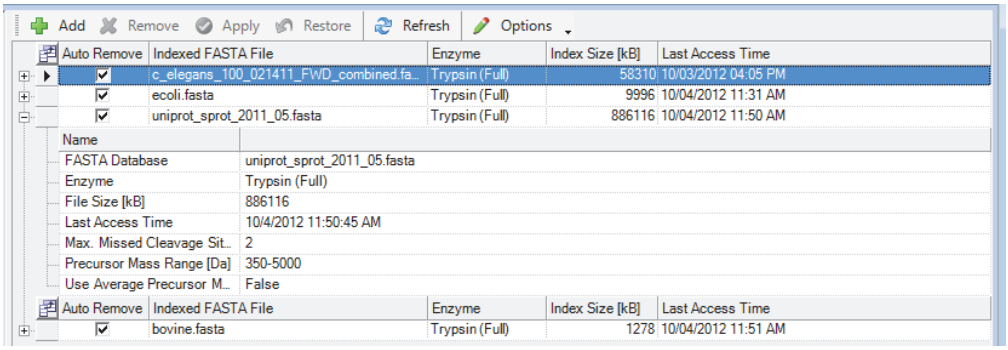
   The FASTA Indexes view appears, as shown in Figure 87.

Figure 87. FASTA Indexes view



2. Click the plus (+) sign to the left of a database name to vertically display the settings for that database, as shown for the uniprot.fasta database in Figure 88.

Figure 88. Database settings in the FASTA Indexes view



Specifying the Columns to Display

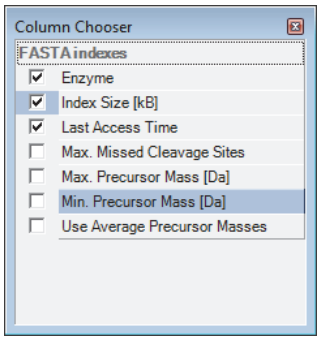
Use the Column Chooser to specify the columns that you want to display.

❖ To set the columns that you want to display

1. Click the **Column Chooser** icon, .
2. In the Column Chooser dialog box, shown in Figure 89, select the check boxes corresponding to the columns that you want to display in the FASTA Indexes view.

   The Proteome Discoverer application instantly makes the selected columns visible and the cleared columns invisible. For a description of these columns, refer to the Help.

Figure 89. Column Chooser dialog box in the FASTA Indexes view



Automatically Creating a FASTA Index

The Proteome Discoverer application automatically creates FASTA indexes for a full enzymatic digestion during a Sequest search, if an adequate FASTA index does not already exist. You can manually create a FASTA index for a semi-enzymatic or non-specific digestion (see "Manually Creating FASTA Indexes" on page 125).

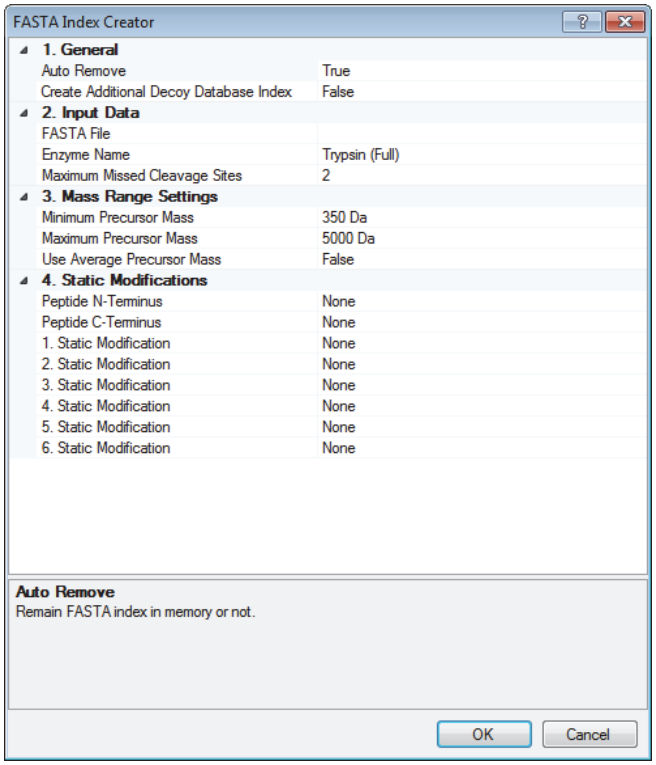
You can only create a specific FASTA index once.

❖ To automatically create a FASTA index

1. Choose **Administration > Maintain FASTA Indexes** or click the **Maintain FASTA Indexes** icon, .
2. Click the **Add** icon, Add .

   The FASTA Index Creator dialog box appears, as shown in Figure 90.

Figure 90. FASTA Index Creator dialog box



3. In the General section, specify whether the available FASTA indexes will be removed from memory after the number of indexes reaches the specified maximum.

   - (Default) True: Automatically removes the FASTA indexes from memory.
   - False: Keeps the FASTA indexes in memory.

   For information about how the Proteome Discoverer application removes FASTA indexes after the maximum has been reached, see "Manually Creating FASTA Indexes" on page 125. For instructions on specifying the maximum number of indexes, see "Changing Number and Location of Stored FASTA Indexes" on page 128.

4. In the Input Data section, specify the basic information that the Proteome Discoverer application needs to create the index:

   - FASTA File: Select the FASTA database to be indexed from the list.
   - Enzyme Name: Select the enzyme used in the digestion from the list on the left (the enzymes on this list are set in the Cleavage Reagents window) and the type of digestion from the list on the right:
     - Full: Specifies a full enzymatic digestion.
     - Semi: Specifies semi-enzymatic digestion.
     - Unspecific: Specifies a non-specific digestion.
     - No Cleavages: Specifies that no cleavages occur.

- Maximum Missed Cleavage Sites: Specifies the maximum number of internal cleavage sites per peptide fragment that is acceptable for an enzyme to miss when cleaving peptides during digestion. Normally the digestion time is too short to enable the enzyme to cleave the peptide at all positions, so you must specify the number of missed positions in one resulting peptide fragment where the enzyme could cleave but did not.

  The minimum value is 0, and the maximum value is 12. The default is 2.

5. In the Mass Range Settings section, set the limits of the mass range of the singly charged precursor ion to be processed:

   - Minimum Precursor Mass: Specifies the minimum mass of the precursor ion. The minimum value is 0.0 Da, and the maximum value is 10000.0 Da. The default is 350 Da.
   - Maximum Precursor Mass: Specifies the maximum mass of the precursor ion. The minimum value is 0.0 Da, and the maximum value is 10000.0 Da. The default is 5000 Da.
   - Use Average Precursor Mass: Determines whether the average mass is used to match the precursor ion.
     - True: Uses the average mass to match the precursor ion.
     - False (Default): Uses the monoisotopic mass to match the precursor ion, which is the mass of the most abundant isotope of the protein, peptide, or fragment ion.

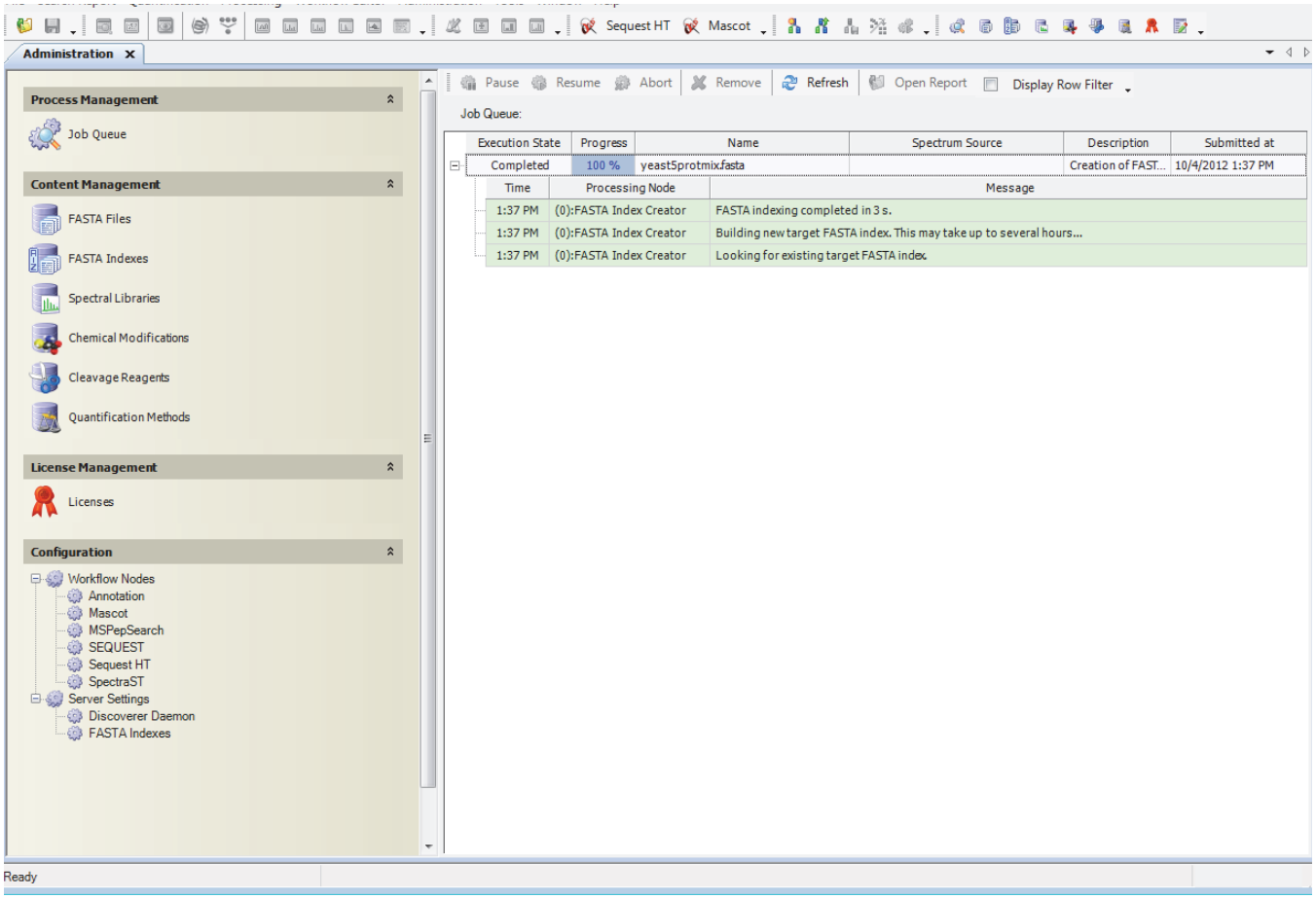
6. In the Static Modifications area, specify the static modifications that occur on the amino acid:

   - Peptide N-Terminus: Select the static modification that occurs on the N terminus of the peptide.
   - Peptide C-Terminus: Select the static modification that occurs on the C terminus of the peptide.
   - Static Modification: Select the static modification that occurs on the amino acid side chain.

7. Click **OK**.

   The Proteome Discoverer application starts creating the FASTA index, and the job queue appears, as shown in Figure 91.

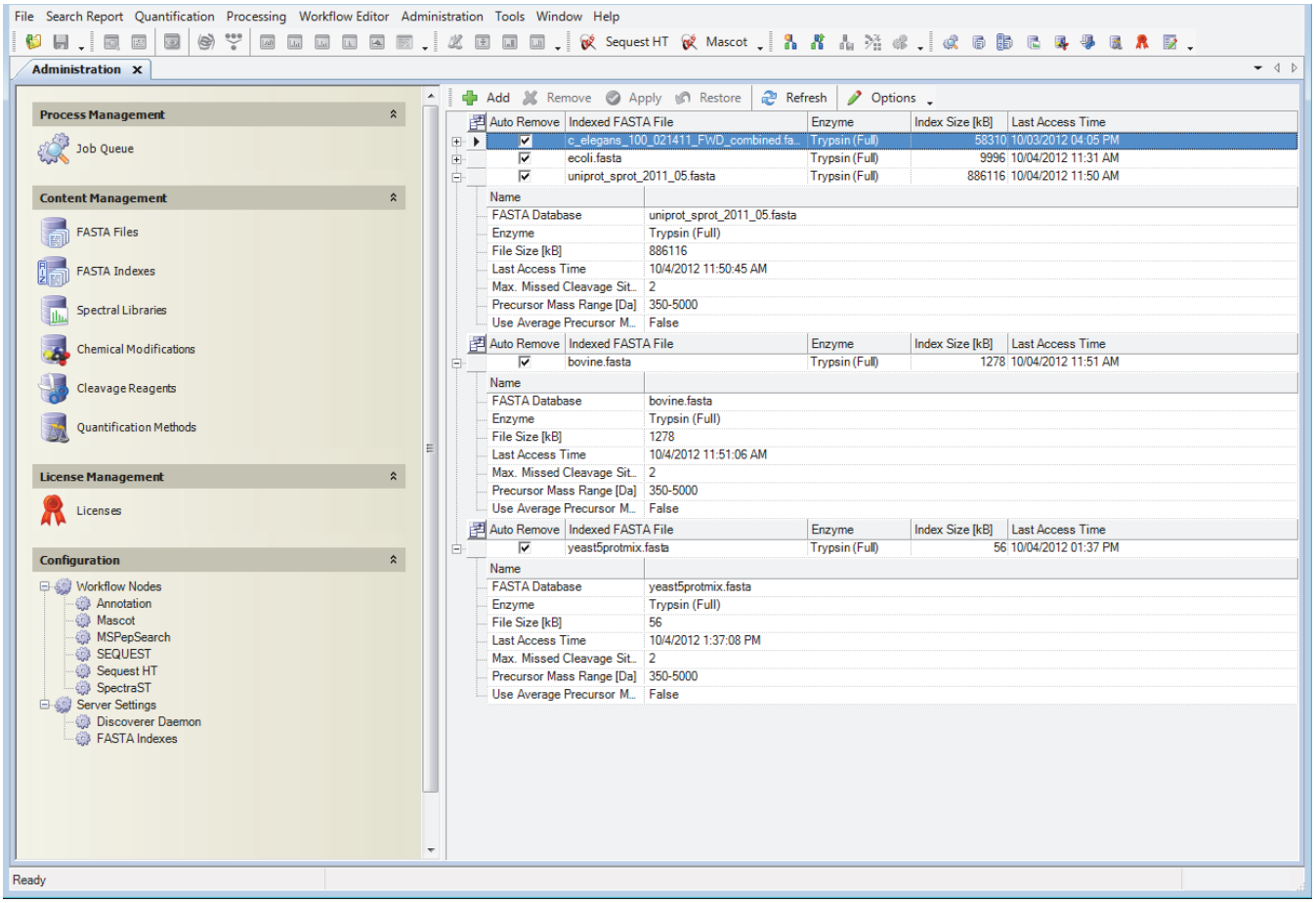
Figure 91. Creating a FASTA index



8. When the job finishes, choose **Administration > Maintain FASTA Indexes** or click the **Maintain FASTA Indexes** icon, , to display the FASTA Indexes view.
9. In the FASTA Indexes view, click the **Refresh** icon, Refresh.

   The new FASTA index appears in the FASTA Indexes view on the Administration page, as shown in Figure 92.

Figure 92. FASTA Indexes view



Manually Creating FASTA Indexes

As noted earlier, you can manually create FASTA indexes for semi-enzymatic or no-enzyme searches.

❖ To manually create a FASTA index

1. Follow the procedure in "Automatically Creating a FASTA Index" on page 121. Also, set the Create Additional Decoy Database Index parameter in the FASTA Index Creator dialog box to **True**, as shown in Figure 90 on page 122.

   The Proteome Discoverer application starts creating the FASTA index, and the job queue appears.

2. When the job finishes, choose **Administration > Maintain FASTA Indexes** or click the **Maintain FASTA Indexes** icon, , to display the FASTA Indexes view.

3. In the FASTA Indexes view, click the **Refresh** icon, Refresh.

The new FASTA index appears in the FASTA Indexes view on the Administration page. The Proteome Discoverer application creates an index for the specified FASTA file and the decoy version of the FASTA file.

Controlling Automatic FASTA Index Removal

After the number of FASTA indexes reaches the specified maximum, the Proteome Discoverer application automatically removes from memory the number of FASTA indexes over the maximum. It first removes the oldest indexes (that is, the ones with the earliest access time). However, you can mark specific FASTA indexes so that they will not be removed from memory, even after the maximum is reached.

- To deactivate automatic FASTA index removal
- To activate automatic FASTA index removal

❖ To deactivate automatic FASTA index removal

1. In the FASTA Indexes view on the Administration page, clear the **Auto Remove** check box.

   The Apply icon now becomes available.

2. Click the **Apply** icon,  .
3. In the Remove FASTA indexes confirmation box, click **OK**.

❖ To activate automatic FASTA index removal

1. Select the **Auto Remove** check box.
2. Click the **Apply** icon,  .
3. In the Remove FASTA indexes confirmation box, click **OK**.

Deleting a FASTA Index

You can only delete FASTA indexes that have an Auto Remove check box selected.

- To delete a FASTA index
- To restore a deleted FASTA index

❖ To delete a FASTA index

1. Be sure that the **Auto Remove** check box is selected for the index that you want to delete.
2. Select the index that you want to delete by clicking the first cell to the right of the plus (+) sign.

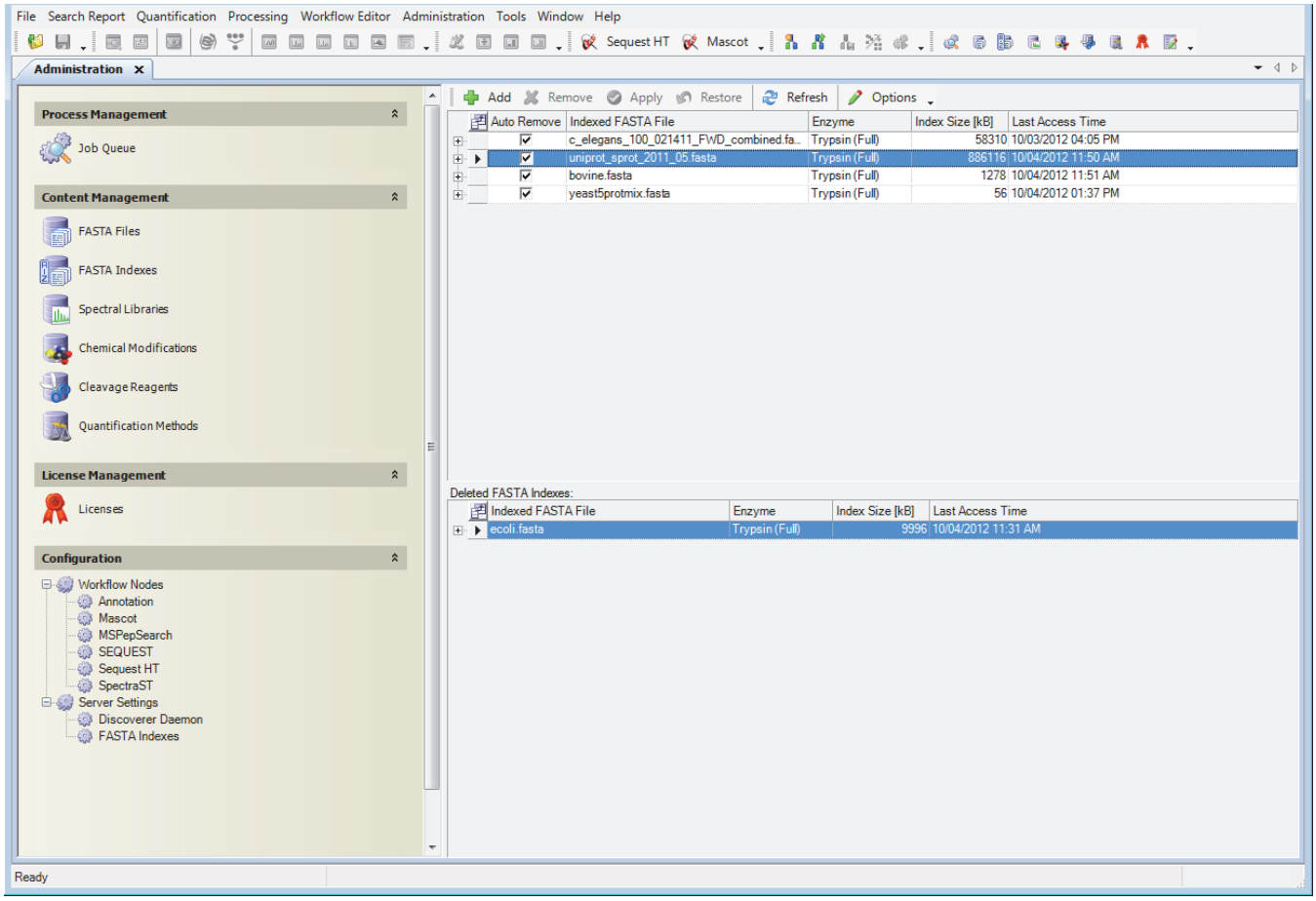
   The cell now changes to the Right Arrow icon,  .

3. Click the **Right Arrow** icon,  .

4. Click the **Remove** icon, Remove.
5. Click **OK** in the Remove FASTA Indexes confirmation box.

The name of the deleted index disappears from the FASTA Indexes table and reappears in a separate table called Deleted FASTA Indexes, as shown in Figure 93. It no longer appears in the FASTA Indexes table. However, because the FASTA index might be used in some calculations, its removal from the application only takes place the next time that the server starts.

Figure 93. Deleted FASTA Indexes table



❖ To restore a deleted FASTA index

1. In the Deleted FASTA Indexes table, select the deleted index by clicking the **Right Arrow** icon, .
2. Click the **Restore** icon, Restore.
3. In the Restore FASTA indexes confirmation box, click **OK**.

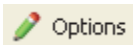
   The restored index appears in the FASTA Indexes table and disappears from the Deleted FASTA Indexes table.

Changing Number and Location of Stored FASTA Indexes

You can specify a new directory for storing the FASTA indexes and change the maximum number of FASTA indexes stored. The Proteome Discoverer application counts all FASTA indexes, even the indexes that cannot be automatically removed with the Auto Remove option.

- To change the number and location of stored FASTA indexes
- To reset the changes made in a previous FASTA index session

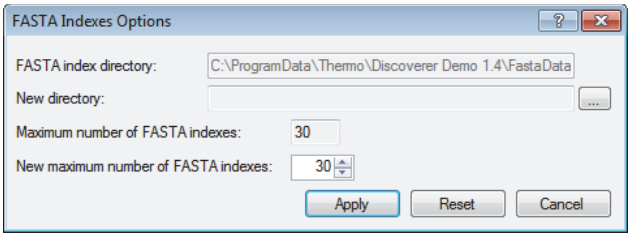
❖ To change the number and location of stored FASTA indexes

1. Click the **Options** icon, 

.

   The FASTA Indexes Options dialog box appears, as shown in Figure 94.

   Figure 94. FASTA Indexes Options dialog box



   Note Another way to access these options is to choose Administration > Configuration and click FASTA Indexes in the Server Settings area.

   The FASTA Indexes Options dialog box contains two read-only parameters:

   - The FASTA Index Directory box displays the name of the current directory where the FASTA indexes are saved.
   - The Maximum Number of FASTA Indexes box displays the current maximum number of FASTA indexes allowed.

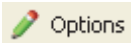
2. In the New Directory box, browse to the directory where you want to store the FASTA indexes.

   You can change the directory only if the server runs on the local machine.

3. In the New Maximum Number of FASTA Indexes box, type the new maximum number of FASTA indexes allowed.
4. Click **OK**.
5. In the FASTA index settings confirmation box, click **OK**.

After you confirm the changes, the Proteome Discoverer application saves them, but the changes are only executed the next time that the server starts. You can undo the changes made since the last time that the server started and before the next time that the server starts, even though you clicked OK in the FASTA Indexes Options dialog box and closed it. For example, when you change the location of the directory in the FASTA Indexes Options dialog box, click OK, and close the dialog box, the server moves all FASTA indexes to the new target directory when the server restarts. But if you reinvoke the dialog box and click Reset before restarting the server, the changes that you made previously are deleted, and the directory reverts to its previous location.

❖ **To reset the changes made in a previous FASTA index session**

1. Click the **Options** icon, 

.

   The FASTA Indexes Options dialog box appears, as shown in Figure 94 on page 128.

2. Click **Reset**.

Removing FASTA Indexes When a FASTA File Is Deleted

When you delete a FASTA file in the Proteome Discoverer application, it removes the FASTA indexes belonging to the deleted FASTA file the next time that the server starts.

Searching Spectrum Libraries

Spectrum library search is a different search approach from the sequence database search ubiquitously used in shotgun proteomics. The main difference between a database search and a spectrum library search is in the origin of the spectra that the measured spectra from your experiments are compared to. Sequence database searches use theoretical spectra generated from peptide sequences, but spectrum libraries are libraries of measured (consensus) spectra from actual previous experiments. Using a library of already well-identified peptides avoids identifying already known peptides over and over again by a time-consuming database search. Restricting the library to previously identified peptides also drastically reduces the search space and therefore the search time. In addition, comparisons that use consensus spectra consider the measured peak intensities, increasing the selectivity and making the identification more accurate.

You can use the SpectraST and the MSPepSearch nodes to search large spectrum libraries downloaded from the NIST or the PeptideAtlas home page.

All currently available libraries are for collision-induced dissociation (CID) or quadrupole time-of-flight (QTOF) data. The QTOF libraries also work for high-energy collision-induced dissociation (HCD) data.

Displaying Spectrum Libraries

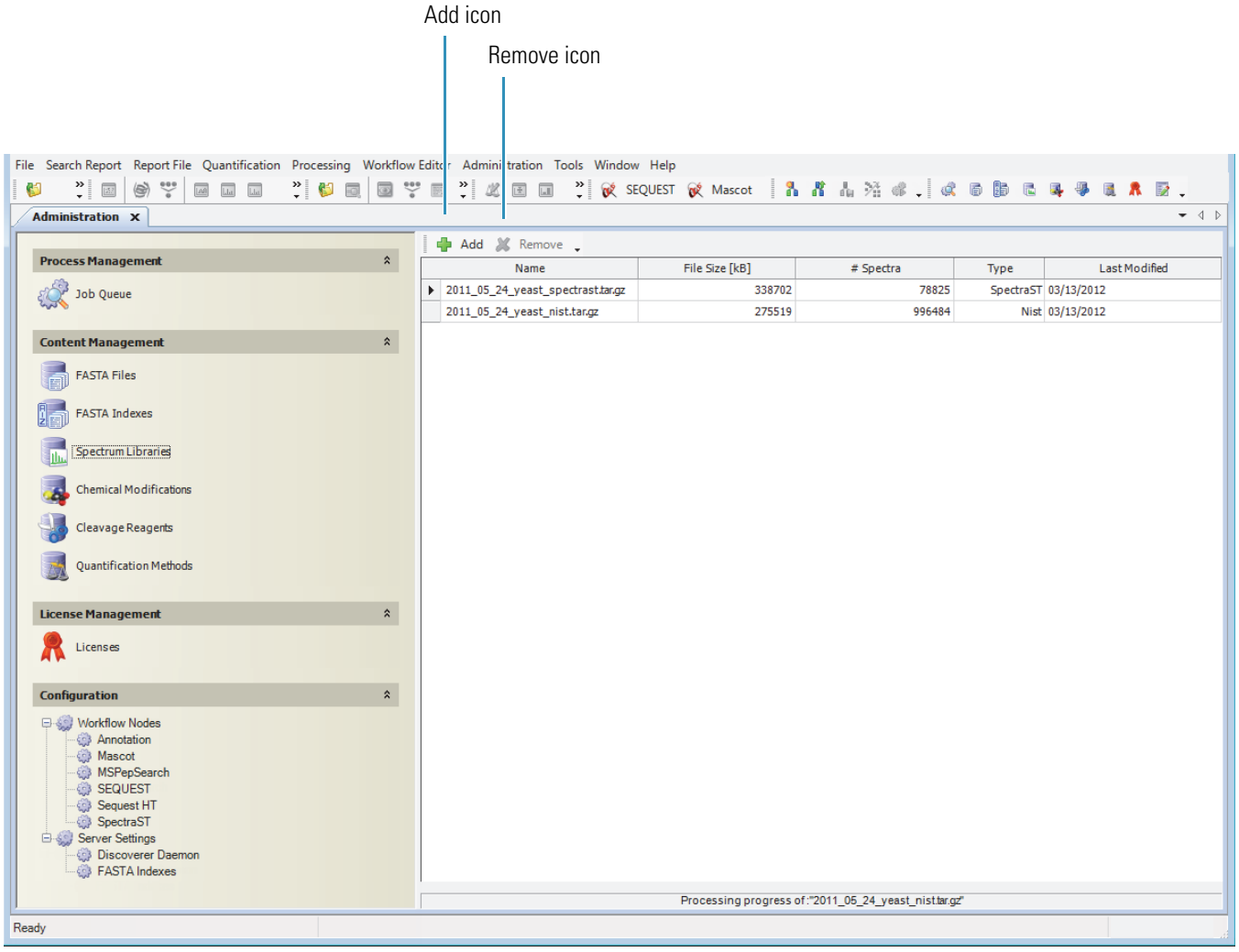
You can display a list of all the spectrum libraries that you registered in the Proteome Discoverer application.

❖ To list the available spectrum libraries

- Choose **Administration > Maintain Spectrum Libraries,** or on the Administration page, click the **Maintain Spectrum Libraries** icon,  , on the toolbar or in the Content Management area.

  The Spectrum Libraries view shown in Figure 95 appears. It lists all the spectrum libraries that you downloaded from NIST or the Peptide Atlas home page and registered. It displays the processed spectrum library properties, such as the file name, file size, the number of proteins stored, and the library type, which determines the search node to use. The Proteome Discoverer application processes the spectrum library and makes it available for use.

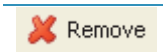
Figure 95. Spectrum libraries view



Spectrum Libraries View Parameters

Table 4 describes the options and columns in the Spectrum Libraries view in the Proteome Discoverer application.

Table 4. Options and columns in the Spectrum Libraries view

Parameter	Description
	Activates the Select a Spectrum Library dialog box, so you can choose the spectrum library to import.
	Deletes a spectrum library.
Name	Displays the name of the spectrum library.
File Size [kB]	Displays the current size of the spectrum library.
# Spectra	Displays the number of spectra found in the spectrum library during processing.
Type	Displays the type of spectrum library downloaded, either SpectraST, which are spectrum libraries that you can use with the SpectraST node, or NIST, which are spectrum libraries that you can use with the MSPepSearch node.
Last Modified	Displays the date when the spectrum library was last modified or created.

Adding a Spectrum Library

You must add a spectrum library to the Proteome Discoverer application before you can conduct a search with the SpectraST or MSPepSearch node. In the registration process, the Proteome Discoverer application automatically recognizes the type of the spectral library. The type determines the search node that you can use the library with. Adding the spectrum libraries is similar to the procedure for adding FASTA files.

- To add a spectrum library for searching with the SpectraST node
- To add a spectrum library for searching with the MSPepSearch node

❖ To add a spectrum library for searching with the SpectraST node

1. Download the appropriate spectrum libraries from the NIST at http://peptide.nist.gov or from Peptide Atlas at http://www.peptideatlas.org/speclib.

   The Proteome Discoverer application recognizes the following file formats for searching spectrum libraries with the SpectraST node:

   - *.msp files, which you can find in the * _consensus_final_true_lib.tar.gz file on the library download site at NIST or on the PeptideAtlas home page. You will need an unpacking tool, such as 7-Zip or WinRAR™, to unpack the downloaded *.gz file before you can add the *.msp file to the Proteome Discoverer application.

- *.zip/*.gz files from the NIST or PeptideAtlas. You can find these files, named *_spectrast.tar.gz or *_splib.zip, on the library download site at NIST or on the PeptideAtlas home page. The *.zip file must contain four files with suffixes *.splib, *.sptxt, *.pepidx, and *.spidx. If one of these files is missing, the file is not added to the Proteome Discoverer application.

2. In the Proteome Discoverer application, choose **Administration > Maintain Spectrum Libraries** or click the **Maintain Spectrum Libraries** icon,  , on the toolbar.
3. Click **Add**.
4. In the Select a Spectrum Library dialog box, do the following:

   If you want to add an .msp file to the Proteome Discoverer application:

   a. In the list box in the lower right corner of the Select a Spectrum Library dialog box, select **All Spectrum Library Files (*.gz, *.msp, *.zip)** or **msp files (*.msp)**.

   b. Browse to the location of the spectrum library where you downloaded and unpacked the *_consensus_final_true_lib.tar.gz file.

   c. Select the ***filename*.msp** file.

   d. Click **Open**.

   If you want to add a *.gz or .zip file to the Proteome Discoverer application:

   a. Browse to the location of the spectrum library where you downloaded the *_spectra.tar.gz file.

   b. Select the ***filename*_spectra.tar.gz** file.

   c. In the list box in the lower right corner of the Select a Spectrum Library dialog box, select **All Spectrum Library Files (*.msp, *.gz, *.zip)** or **Zip archives (*.gz; *.zip)**.

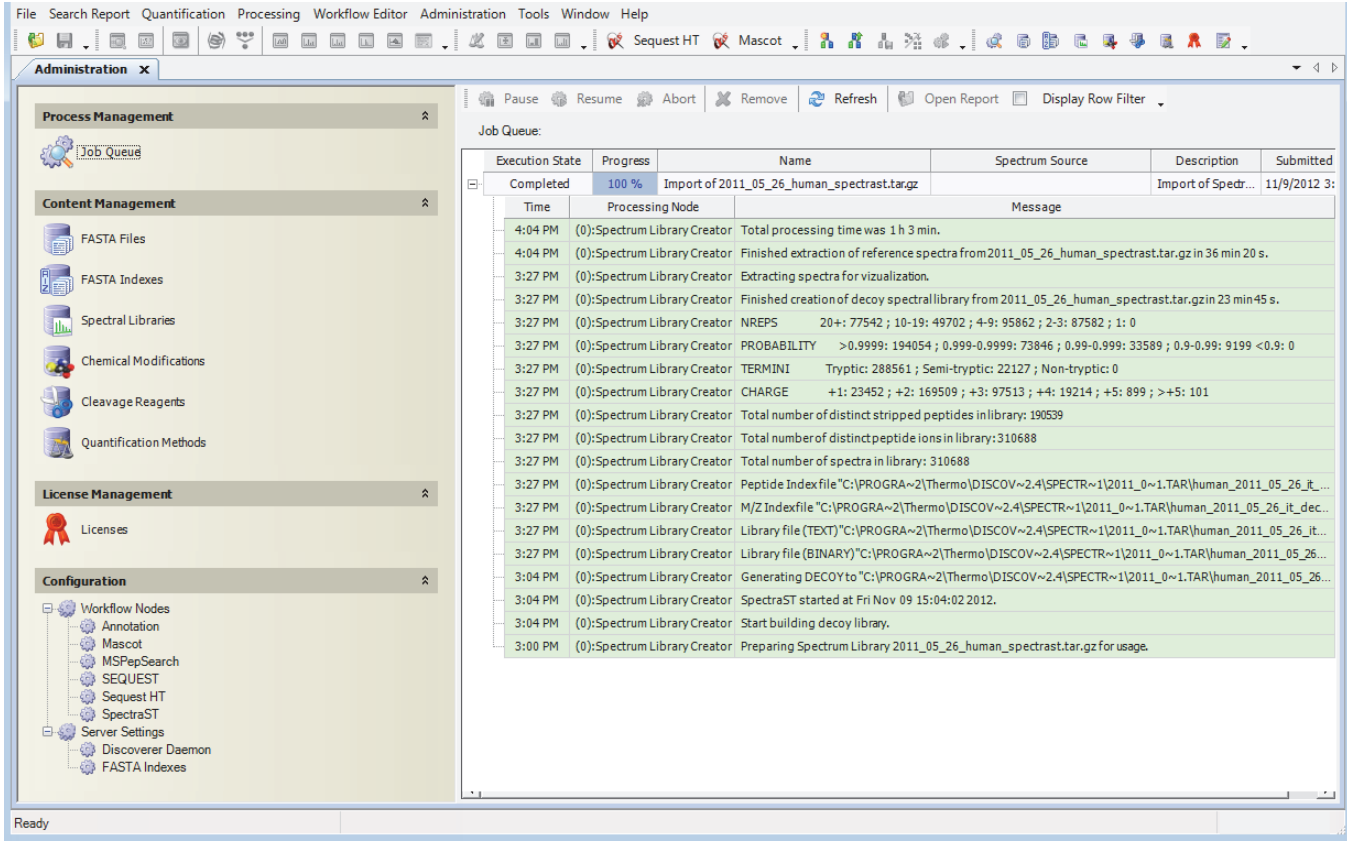
   d. Click **Open**.

   When you add a spectrum library file, the Proteome Discoverer application takes the following steps:

   - Constructs the library from the *filename*.msp file or extracts the archive file.
   - Creates a decoy spectrum library and other files needed for the actual search.
   - Extracts spectra for visualization.

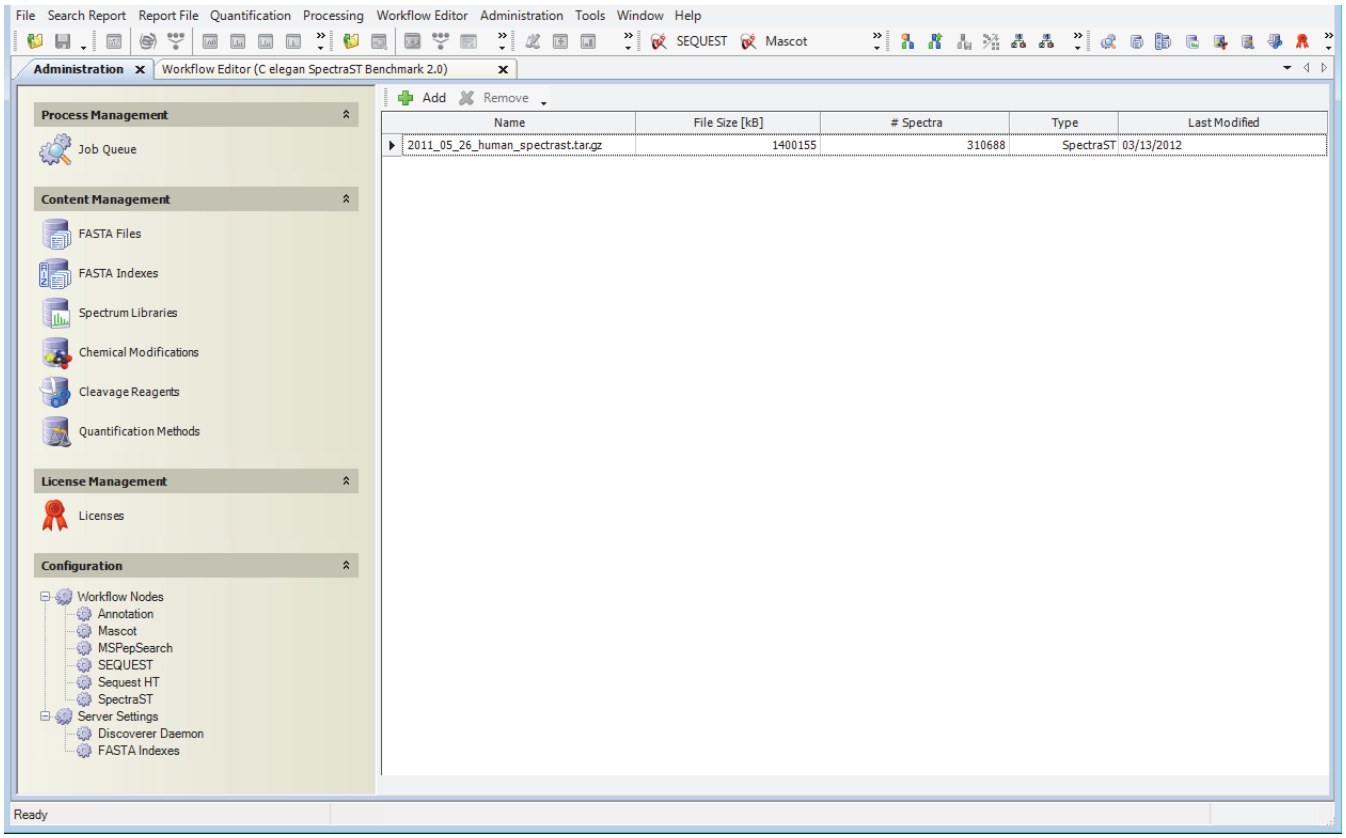
   During library creation, the job queue in the Administration view displays each step, as shown in Figure 96.

Figure 96. Adding a spectrum library for searching with the SpectraST node



When the Proteome Discoverer application finishes adding the spectrum library, the file name and the spectrum library properties appear in the Spectrum Libraries view, as shown in Figure 97.

Figure 97. Added .tar.gz file and the spectrum library properties in the Spectrum Libraries view



Now you are ready to search the spectrum library. For more information on the SpectraST node, refer to the Help. To search with the SpectraST node, see "Searching Spectrum Libraries with the SpectraST Node" on page 137. For more information on the SpectraST node, refer to the Help.

❖ To add a spectrum library for searching with the MSPepSearch node

1. Download the appropriate spectrum libraries from the NIST at http://peptide.nist.gov or from Peptide Atlas at http://www.peptideatlas.org/speclib.

   The Proteome Discoverer application recognizes the following file formats for searching spectrum libraries with the MSPepSearch node:

   - *.zip/*.gz files from NIST or PeptideAtlas. You can find these files in the *_nist.tar.gz file on the library download site at NIST or the *_nist.zip file on the PeptideAtlas home page. The file must contain a complete spectrum library in MSPepSearch. If files are missing, the Proteome Discoverer application does not add the library.

2. In the Proteome Discoverer application, choose **Administration > Maintain Spectrum Libraries** or click the **Maintain Spectrum Libraries** icon,  , on the toolbar.

3. Click **Add**.

4. In the Select a Spectrum Library dialog box, do the following:

   a. In the list box in the lower right corner of the Select a Spectrum Library dialog box, select **All Spectrum Library Files** (*.gz, *.msp, *.zip) or Zip archives (*.gz, *.zip).

   b. Browse to the location of the spectrum library where you downloaded and unpacked the *_nist.tar.gz file.

   c. Select the ***filename*.gz** file.

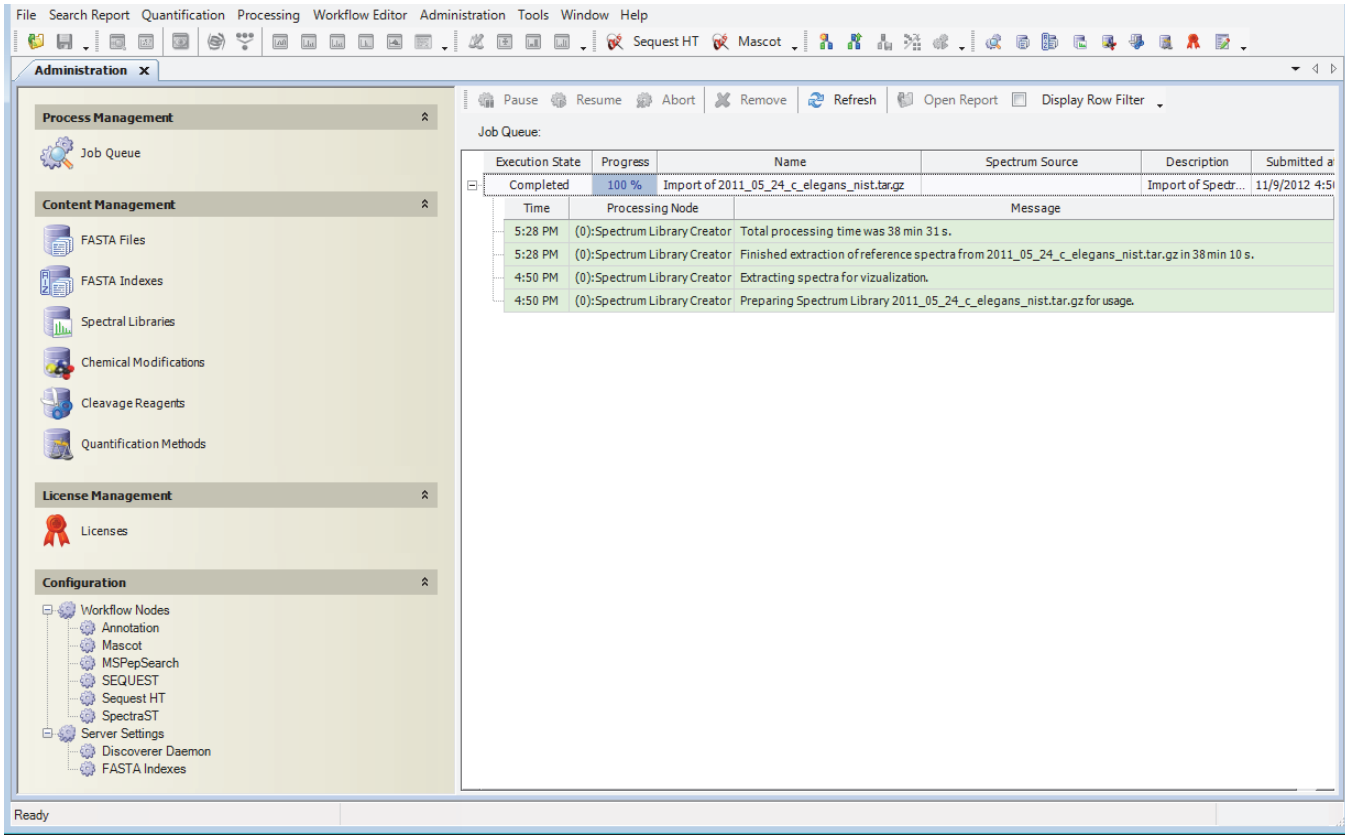
   d. Click **Open**.

   When you add a spectrum library file, the Proteome Discoverer application takes the following steps:

   - Extracts the archive file.
   - Extracts spectra for visualization.

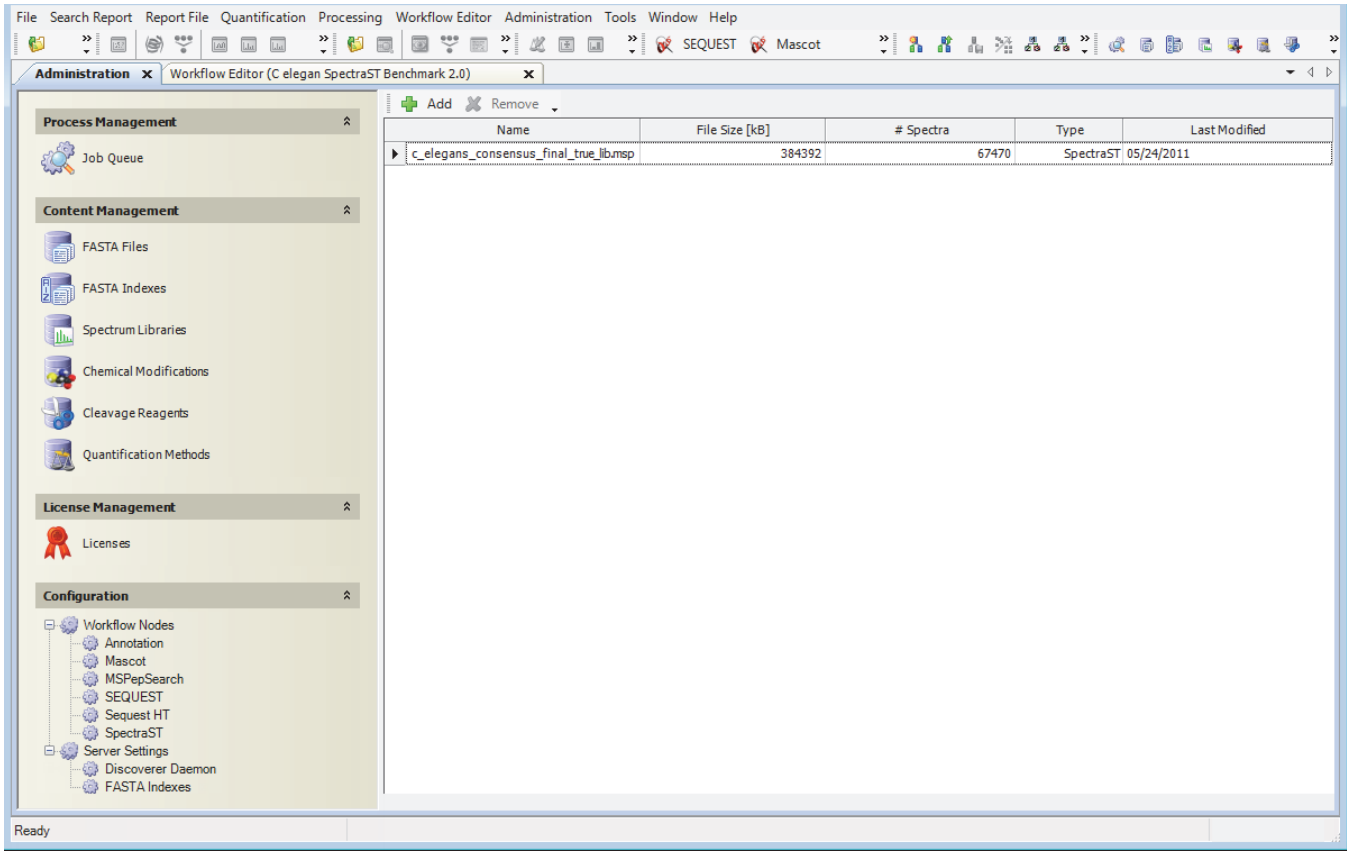
   During library creation, the job queue in the Administration view displays each step, as shown in Figure 98.

Figure 98. Adding a spectrum library for searching with the MSPepSearch node



When the Proteome Discoverer application finishes adding the spectrum library, the spectrum library file appears in the Spectrum Libraries view, as shown in Figure 99.

Figure 99. Added NIST spectrum library in the Spectrum Libraries view



Now you are ready to search the spectrum library. To search with the MSPepSearch node, see "Searching Spectrum Libraries with the MSPepSearch Node" on page 139. For more information on the MSPepSearch node, refer to the Help.

Deleting a Spectrum Library

You can delete a spectrum library from the application.

❖ To delete a spectrum library

1. Choose **Administration > Maintain Spectrum Libraries**.

   The Administration page appears with the Spectrum Libraries view.

2. Click ▶ at the beginning of a row to select the row.

3. Click Remove.

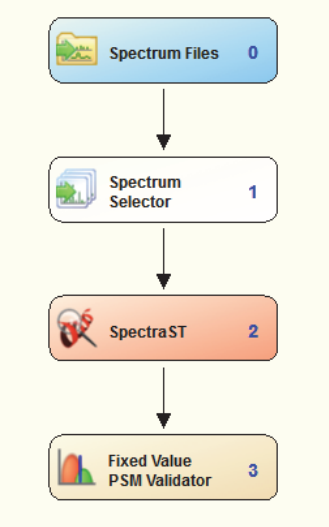
4. In the Remove Spectrum Libraries Databases dialog box, click **OK**.

The Spectrum Libraries file that you selected appears as a job in the job queue. After you start the deletion of the file, you cannot cancel the deletion. You can remove the completed job from the job queue by clicking Remove and then clicking OK in the Delete Jobs dialog box.

Searching Spectrum Libraries with the SpectraST Node

Figure 100 shows the basic workflow for searching spectrum libraries with the SpectraST node. You can use this node as an alternative to a search node such as SEQUEST.

Figure 100. Workflow using SpectraST to search spectrum libraries



For a description of the parameters available in the SpectraST node, refer to the Help.

The spectrum library search reports the three scores shown in Table 5.[1] The dot score and the dot bias are secondary scores, and their values are not shown by default.

Table 5. Scores generated by the SpectraST search node (Sheet 1 of 2)

Score	Description
F-value	Specifies the discriminant scoring function that the Proteome Discoverer application calculates from the dot score, dot bias, and the normalized difference between the best and second-best hit (ΔD). The application uses the f-value for FDR calculation. For more information on the f-value, see "F Value."

[1] Lam, Henry, et al. *Proteomics 7*, **2001**, 655-667.

Table 5. Scores generated by the SpectraST search node (Sheet 2 of 2)

Score	Description
Dot score	Specifies the spectral dot product as the primary similarity score. For more information on the dot score, see "Dot Score."
Dot bias score	Measures how much the dot score is dominated by only a few peaks, which might indicate false positive hits. For more information on the dot bias, see "Dot Bias Score."

Dot Score

The dot score is the primary score from the spectral library search. To calculate the dot score, the Proteome Discoverer application splits the reference spectrum into equal bins. It then adds the product of the normalized intensities of each bin up to the dot score, as shown in the following formula:

$$D = \sum_{j} \hat{I}_{\text{library,j}} \hat{I}_{\text{query,j}}$$

where Îlibrary,j and Îquery,j are normalized intensities of the j^{th} bin of the spectra. D is the dot score.

The application reports the dot score together with the dot bias.

Dot Bias Score

The application calculates the dot bias score as follows:

$$DB = \frac{\sqrt{\sum_{j} \hat{I}^{2}_{\text{library,j}} \hat{I}^{2}_{\text{query,j}}}}{D}$$

where Îlibrary,j and Îquery,j are normalized intensities of the j^{th} bin of the spectra. D is the dot score. A high dot bias (DB) value indicates that the dot score results from only a few peaks.

F Value

The Proteome Discoverer application calculates the ΔD value in the F value formula as follows:

$$\Delta D = \frac{D_1 - D_2}{D_1}$$

The application calculates the F value (F) as follows:

$$F = 0.6\mathrm{D} + 0.4\Delta D - b$$

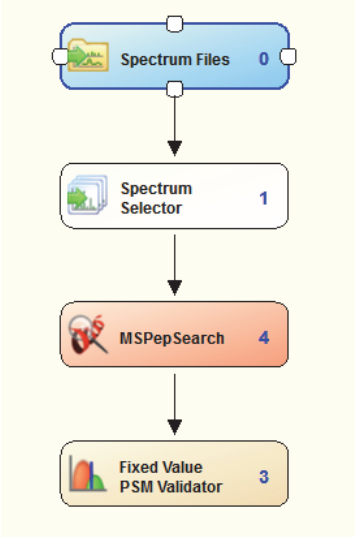
where D is the dot score, and b is the following (DB is the dot bias):

$$b = \begin{cases} 0.12 & \text{if } DB < 0.1 \\ 0.12 & \text{if } 0.35 < DB \leq 0.4 \\ 0.18 & \text{if } 0.4 < DB \leq 0.45 \\ 0.24 & \text{if } DB > 0.45 \\ 0 & \text{for all other values of } DB \end{cases}$$

Searching Spectrum Libraries with the MSPepSearch Node

Figure 101 shows the basic workflow for searching spectrum libraries with the MSPepSearch node. You can use this node as an alternative to a search node such as SEQUEST.

Figure 101. Workflow using the MSPepSearch node to search spectrum libraries



The Fixed Value PSM Validator is the only possible peptide validator for the MSPepSearch node. It is impossible to perform a decoy search because there is no proper decoy spectrum library.

For a description of the parameters available in the MSPepSearch node, refer to the Help.

The spectrum library search reports the three scores shown in Table 6. Dot score and reversed dot score are secondary scores, and their values are not shown by default.

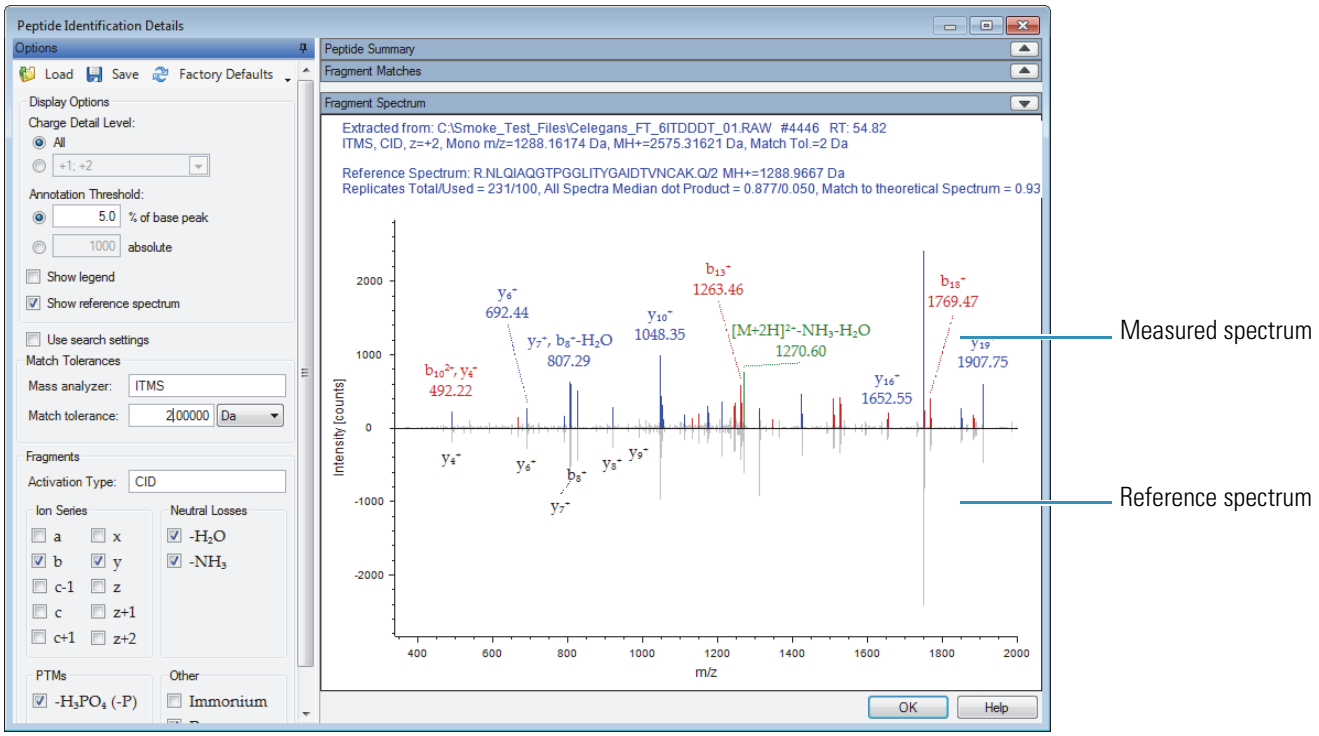
Table 6. Scores generated by the MSPepSearch node

Score	Description
MSPepSearch	Is the main score of MSPepSearch.
Dot score	Is the score from a cross-correlation computed between two spectra.
Reverse dot score	Is the reversed spectral dot product.

Visually Verifying Spectrum Library Matches

You can visually verify matches between measured spectra from your experiment and the reference spectra in the spectrum library for peptides identified with the SpectraST or the MSPepSearch node. In the Peptide Identification Details view, you can display a mirror plot of the matching peptides, as shown in Figure 102. You can use the reference spectrum with the fragment match settings (refer to the Help).

Figure 102. Mirror plot in the Peptide Identification Details view



The Proteome Discoverer application displays the reference spectrum using intensities multiplied by –1 in the same plot as the measured spectrum. In the reference spectrum, it also labels peaks of the a, b, c, ion series and the x, y, and z ion series, as well as the peaks from the precursor peptide. It does not display labels for all fragments with a mass difference, isotope peaks, and "?" peaks in the spectrum library.

❖ To generate a mirror plot

1. Open the MSF file for the results of the spectrum library search performed with the SpectraST node or the MSPepSearch node.
2. If you used spectrum library nodes and other search nodes in the workflow, ungroup the peptides by right-clicking and clearing the **Show Peptide Groups** check box.

   Ungrouping peptides is not necessary if you used only spectrum library search nodes in the workflow.

3. Follow the instructions for generating a Peptide Identification Details view given in "Interpreting Your Results with the Peptide Identification Details View" on page 276.

Updating Chemical Modifications

You can update the chemical modifications that you use to conduct a peptide identification search. The available modifications are defined in the Chemical Modifications view on the Administration page that is opened by choosing Administration > Maintain Chemical Modifications. Use this view to customize the chemical modifications that you use to do your search. You can import a new list or the latest UNIMOD list. You can also modify the chemical modification list by adding amino acids to the modifications, creating new modifications, or activating or deactivating existing modifications.

Note A modification must be active to be usable during a search.

The Proteome Discoverer application offers two types of modifications, dynamic and static.

Dynamic Modifications

Dynamic modifications, also known as variable amino acid modifications, are modifications that might or might not be present. They are mainly used for determining post-translational modifications (PTMs). For example, some phosphorylated peptide serines are modified, and some are not modified.

You can set the parameters for a dynamic search on the Select Modifications page of the Mascot and Sequest HT search wizards. For instructions on setting these parameters in the wizards, see Figure 19 on page 38 and the steps that follow it.

Static Modifications

Static modifications apply the same specific mass to all occurrences of that named amino acid, as in an exhaustive chemical modification.

A static modification might result from derivatization or isotopic labeling of an amino acid. For example, a carboxymethylated cysteine has a delta mass of 58.005479, which is added to each cysteine residue appearing in a protein.

In static searches, the Proteome Discoverer application assumes that every amino acid residue will be modified in that way. Constant mass is changed. The search wizards perform static modification searches by adding the specified constant value to the mass of the specified amino acid.

You can set the parameters for a static search on the Select Modifications page of the Mascot and Sequest HT search wizards. For instructions on setting these parameters in the search wizards, see Figure 19 on page 38 and the steps that follow it.

Opening the Chemical Modifications View

The Chemical Modifications view is an advanced feature of the Proteome Discoverer application. You use it to build and maintain the static and dynamic modifications data that is available when you define your search settings.

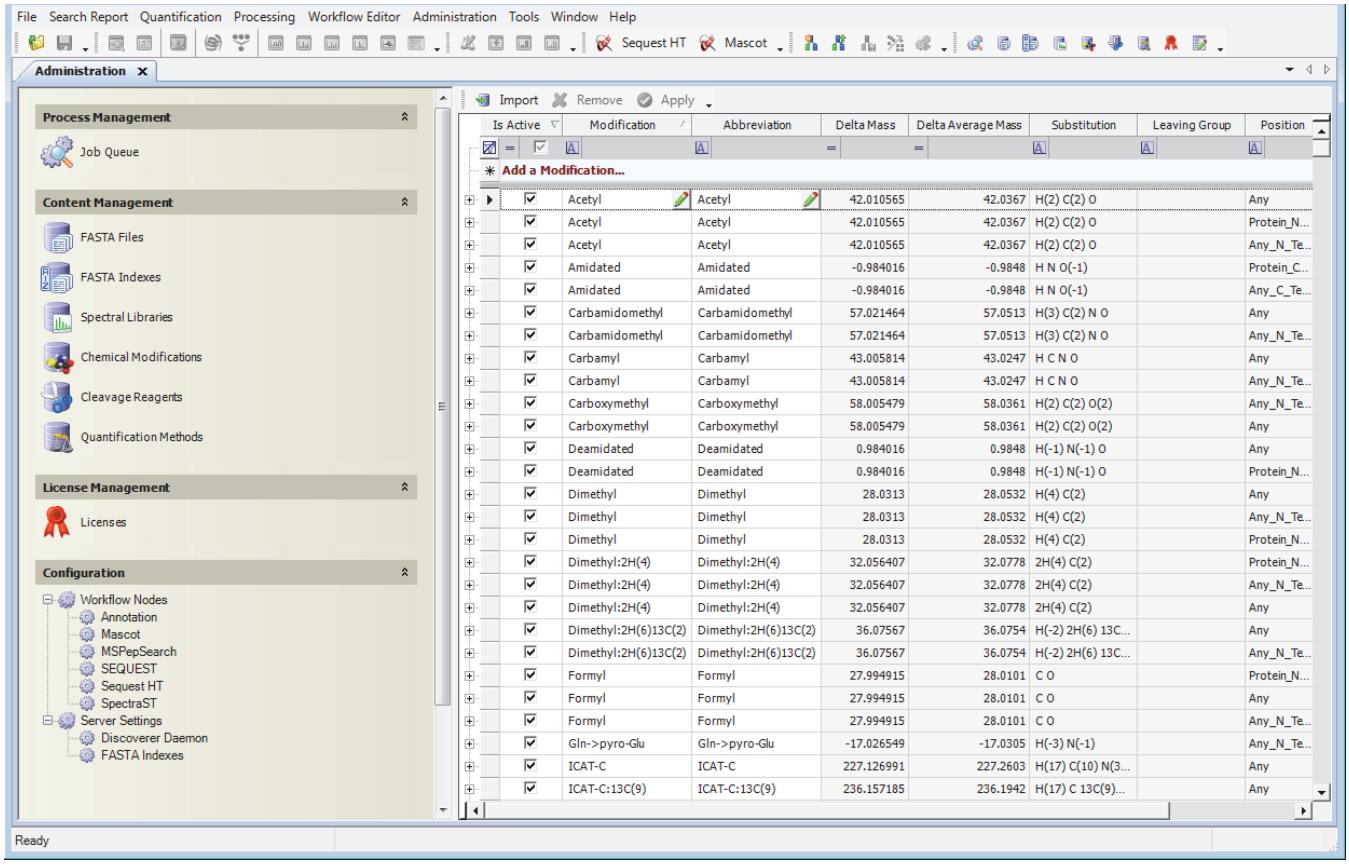
In the Chemical Modifications view, you can explore the default types of modifications and their corresponding amino acids. It contains the modification's delta mass, amino acids, and substitutions. By using the Chemical Modifications view, you can add amino acids to existing modifications and create new modifications.

❖ To open the Chemical Modifications view

1. Choose **Administration > Maintain Chemical Modifications**, or click the **Maintain Chemical Modifications** icon,  , either on the toolbar or on the Administration page.

   The Chemical Modifications view appears on the Administration page, as shown in Figure 103. The amino acids listed are those where the modifications can appear.

Figure 103. Chemical Modifications view



2. Click + to the left of each modification row to see the amino acids that the modification is found on, the letter abbreviation of this amino acid, and the modification type or

category. Figure 104 shows an example of the information given for the Acetyl modification. Table 7 lists the available modification categories.

Figure 104. Displaying modification information for acetyl

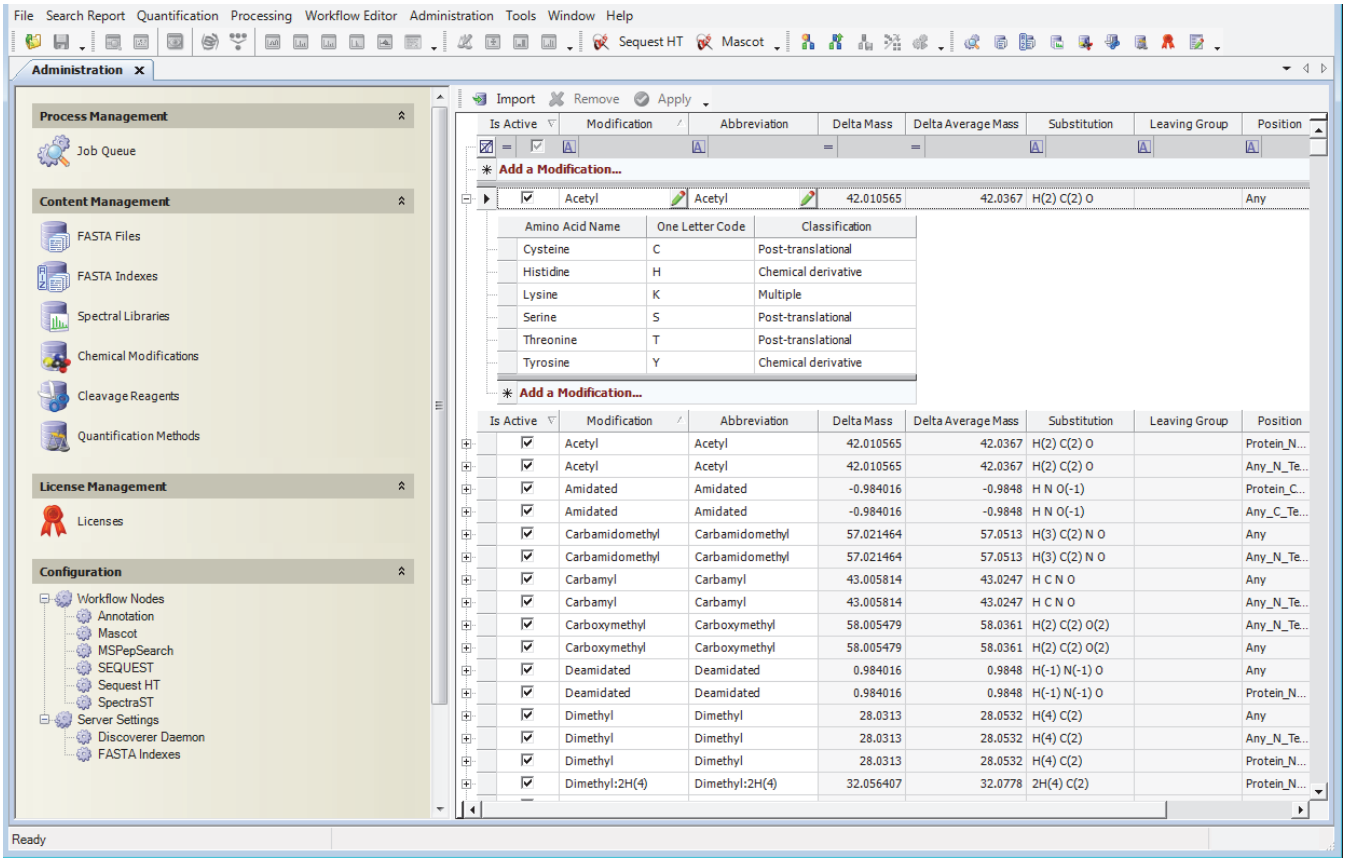


Table 7. Available modification categories (Sheet 1 of 2)

Classification	Description
Post-translational	Protein modification after translation (*in vivo*)
Co-translational	Amino acid modified in translation (for example, myristyl glycine)
Pre-translational	Amino acid modified before integration into a protein (for example, formyl methionine)
Chemical derivative	Chemically induced modification (for example, during sample preparation)
Artifact	Modification made during sample preparation
N-linked glycosylation	Glycosylation (in vivo)
O-linked glycosylation	Glycosylation (in vivo)
Other glycosylation	Glycosylation (in vivo)

Table 7. Available modification categories (Sheet 2 of 2)

Classification	Description
Synthetic peptide protection group	Protection group used in chemical peptide synthesis (for example, trityl (triphenylmethyl))
Isotopic label	Label for quantification
Non-standard residue	Amino acid derivative like selenomethionine
Multiple	More than one classification possible
AA substitution	Amino acid replaced by another amino acid (mutation)
Other	Modification not fitting into another category

The Proteome Discoverer application automatically imports the classifications from unimod.org, the protein modifications online database for mass spectrometry applications. You can also manually define your own classifications.

Adding Chemical Modifications

You can create new chemical modifications and add them to the Chemical Modifications view. For example, you might have a new or experimental label that you want to add to the list of chemical modifications.

- To add a new chemical modification
- To update an existing chemical modification

❖ To add a new chemical modification

1. Choose **Administration > Maintain Chemical Modifications**.

   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. Click the **Add a Modification** heading.

   An empty row appears, as shown in Figure 105.

Figure 105. Adding a row in the Chemical Modifications view

Is Active	Modification	Abbreviation	Delta Mass	Delta Average Mass	Substitution	Leaving Group	Position	Unimod Accession No.
			1	1			Any	0
☑	Acetyl	Acetyl	42.010565	42.0367	H(2) C(2) O		Any	1
☑	Acetyl	Acetyl	42.010565	42.0367	H(2) C(2) O		Protein_N...	1
☑	Acetyl	Acetyl	42.010565	42.0367	H(2) C(2) O		Any_N_Te...	1
☑	Amidated	Amidated	-0.984016	-0.9848	H N O(-1)		Protein_C...	2
☑	Amidated	Amidated	-0.984016	-0.9848	H N O(-1)		Any_C_Te...	2
☑	Carbamidomethyl	Carbamidomethyl	57.021464	57.0513	H(3) C(2) N O		Any	4
☑	Carbamidomethyl	Carbamidomethyl	57.021464	57.0513	H(3) C(2) N O		Any_N_Te...	4
☑	Carbamyl	Carbamyl	43.005814	43.0247	H C N O		Any	5
☑	Carbamyl	Carbamyl	43.005814	43.0247	H C N O		Any_N_Te...	5

3. In the empty row, enter the name of the modification, the delta masses, the chemical substitution, the chemical group that is leaving, the position, and the abbreviations of the modifications.

   If you select Any in the Position column, a message box opens to inform you that you must specify which amino acids (target amino acids) will possibly have the modification. For instructions on this procedure, see "Adding Amino Acids" on page 145.

4. To accept the new modifications, click the **Apply** icon, 

.

5. Add an amino acid to the modifications. See "Adding Amino Acids" on page 145.

❖ **To update an existing chemical modification**

1. Choose **Administration > Maintain Chemical Modifications**.

   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. In the Modification column, click the cell that you want to update**.**

3. Type your changes for the delta masses, the substitution, the group that it is leaving, the position, or the abbreviations of the modifications.

   For chemical modifications that you add yourself, you can edit any column except the Unimod Accession No. column. The Unimod Accession No. column identifies these modifications by a zero. For chemical modifications that you import from UNIMOD, you can edit only the Modification and Abbreviation columns. UNIMOD chemical modifications are identified by a number greater than zero in the Unimod Accession No. column.

   Columns that you can edit activate an edit button when you click them. Columns that you cannot edit display a gray background.

4. To accept the changes, click the **Apply** icon, 

.

Adding Amino Acids

You can add amino acids to a modification that has been set up for any position.

❖ **To add an amino acid to a modification**

1. Choose **Administration > Maintain Chemical Modifications**.

   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. Click **+** to the left of the modification row that you want to update.

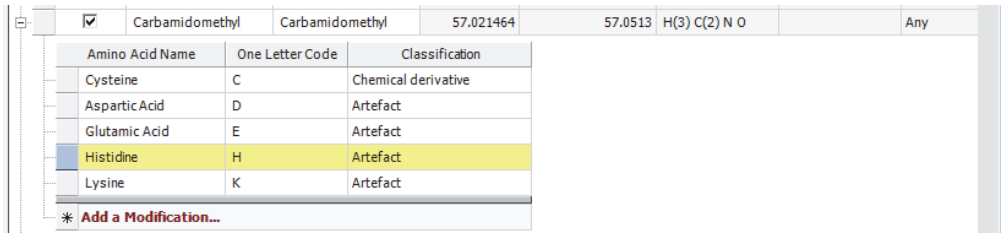
   The row must display **Any** in the Position column.

   The list of classifications now appears, as shown in Figure 104 on page 143.

3. Click the **Add a Modification** line below the list of amino acids.

Figure 106 shows this line.

Figure 106. Adding an amino acid to a modification



An empty row appears.

4. In the empty row, select the amino acid from the list in the Amino Acid Name column.

   The amino acid and the one letter abbreviation appear.

5. From the list in the Classification column, select the type of modification.

6. To save the modifications, click the **Apply** icon, Apply.

When you reimport data from unimod.org, the Proteome Discoverer application retains the modification that you added. However, if you want to change the classification of an amino acid, you must do so before reimporting the Unimod data. After you import the Unimod data, the only way to change the classification is to delete the amino acid and re-add it with another classification.

Deleting Chemical Modifications

You can remove chemical modifications from the Chemical Modifications view.

❖ To delete a modification

1. Choose **Administration > Maintain Chemical Modifications**.

   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. Select the row of the modification that you want to delete.

3. Click the **Remove** icon, Remove.

4. In the Delete Row dialog box, click **Yes**.

   The row is removed from the chemical modifications table.

Importing Chemical Modifications

You can import chemical modifications from a local file or obtain an updated version from unimod.org, a public domain database.

When you install the Proteome Discoverer application, it automatically imports accessions from unimod.org as chemical modifications.

- To import chemical modifications from a local file
- To import chemical modifications from unimod.org

❖ To import chemical modifications from a local file

1. Choose **Administration > Maintain Chemical Modifications**.

   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. Click the **Import** icon, 

.

3. In the Import From list of the Import Modifications dialog box, select **Local File**.

4. In the adjacent box, click the **Browse** button (…) to browse for your file, or type the name and path of the file in the box.

5. To overwrite an existing upload, select the **Overwrite Existing** check box.

6. Click **Import**.

   A status message appears.

7. When the upload is complete, click **Close**.

❖ To import chemical modifications from unimod.org

1. Choose **Administration > Maintain Chemical Modifications**.

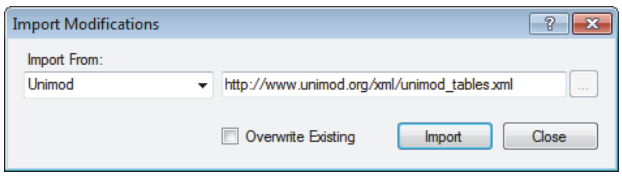
   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. Click the **Import** icon, 

.

   The Import Modifications dialog box appears, as shown in Figure 107.

   Figure 107. Import Modifications dialog box

3. In the Import From list, select **Unimod**.

   The UNIMOD URL appears in the adjacent box.

4. To overwrite an existing upload, select the **Overwrite Existing** check box.

5. Click **Import**.

   A status message appears.

6. When the upload is complete, click **Close**.

For chemical modifications imported from unimod.org, you can only edit the Is Active, Modification, and Abbreviation columns. You do not have access to the Delta Mass, Delta Average Mass, Substitution, Leaving Group, Position, and UNIMOD Accession No. columns. Chemical modifications imported from unimod.org have a number greater than zero in the Unimod Access No. column.

If you select the Overwrite Existing check box, the Proteome Discoverer application does the following when it imports chemical modifications from unimod.org:

- Updates the columns that are inaccessible to you.
- Updates the names and the abbreviations of the modifications.
- Adds any new amino acids found in unimod.org.
- Adds any amino acids that you removed if they are defined in unimod.org.
- Removes any amino acids that you added if they are defined in unimod.org.

If you do not select the Overwrite Existing check box, the Proteome Discoverer application performs the same tasks as it does during installation:

- Updates the columns that are inaccessible to you.
- Leaves the modification name and abbreviation unchanged.
- Adds any new amino acids found in unimod.org.
- Adds any amino acids that you removed if they are defined in unimod.org.
- Leaves unchanged any amino acids that you added.

Deleting Amino Acids

You can also delete amino acids from chemical modifications.

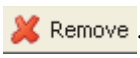
❖ To delete an amino acid from a chemical modification

1. Choose **Administration > Maintain Chemical Modifications**.

   The Chemical Modifications view appears, as shown in Figure 103 on page 142.

2. Click **+** to the left of the modification row that you want to delete.

   The row expands and the associated amino acids appear.

3. Select the amino acid row that you want to delete.
4. Click the **Remove** icon, 

.
5. In the Delete Row dialog box, click **Yes**.

   The row is removed from the chemical modifications table.

Using the Qual Browser Application

The Proteome Discoverer application includes the Qual Browser application, which you can use to examine spectra and chromatograms in detail. With the Qual Browser application, you can view the entire ion chromatogram and browse individual precursor and MS^n data. You can filter the results in a variety of ways, for example, to produce a selected ion chromatogram. The Qual Browser application automatically displays the elemental composition, theoretical mass, delta values, and ring and double-bond (RDB) equivalents for your high-resolution data. (RDB equivalents measure the number of unsaturated bonds in a compound and limit the calculated formulas to only those that make sense chemically.)

You must have the Xcalibur data system installed to use the Qual Browser application. For information about using the Qual Browser application, refer to the *Thermo Xcalibur Qualitative Analysis User Guide*.

You must also have a search results file open and a specific peptide or search input row selected before the Qual Browser application becomes available. If you are viewing the Administration page, the Qual Browser application does not open a raw file.

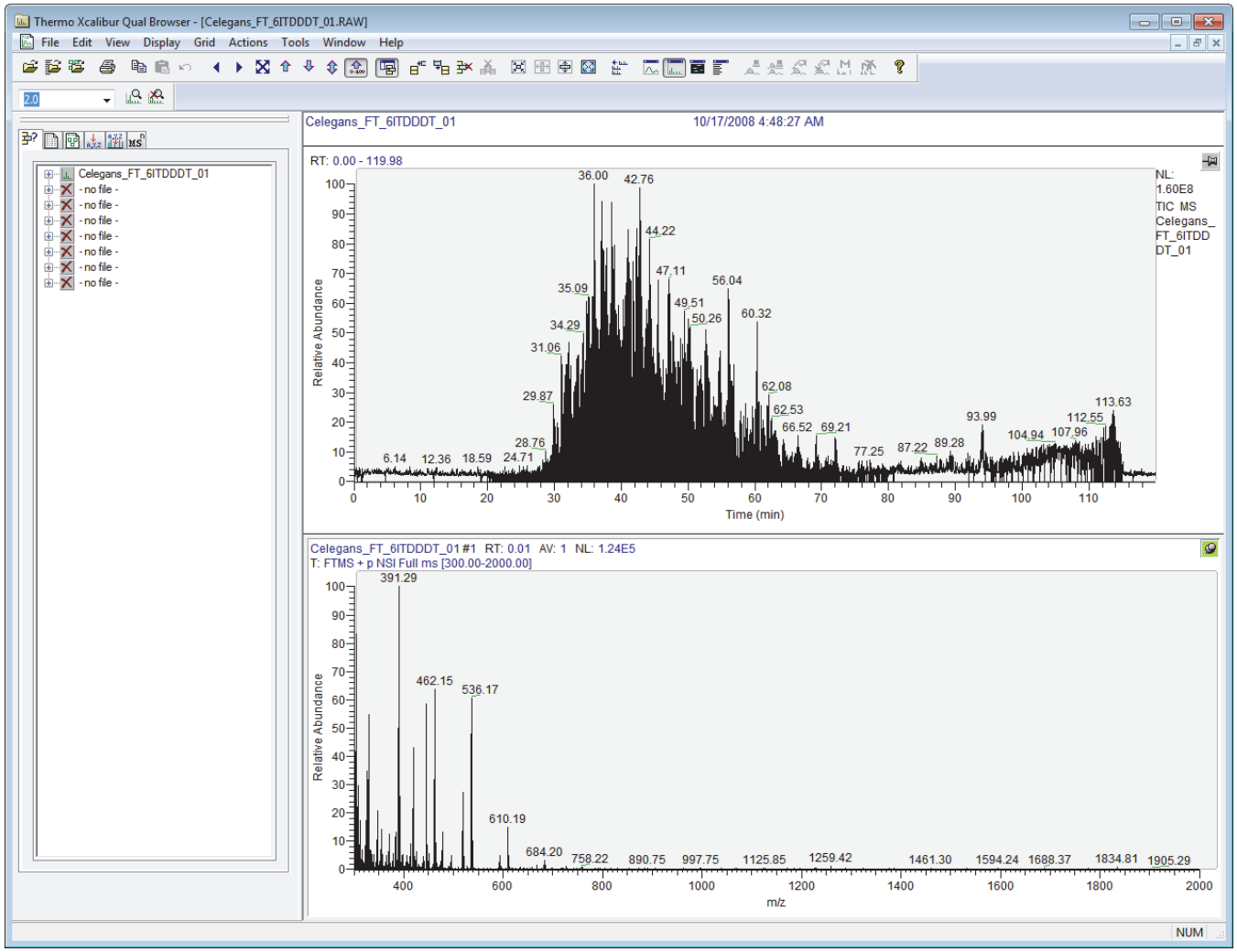
❖ To open the Qual Browser application

1. In the Proteome Discoverer application, choose **Tools > Open QualBrowser**, or click the **Qual Browser** icon,  , or press CTRL+SHIFT+B to open the Spectrum window.

   Note You must have a search results (MSF) file open and selected before the Open QualBrowser command becomes available on the Tools menu. In addition, the Open QualBrowser command is available only when peptides are ungrouped and you select at least a single peptide or a search input item first. You cannot use QualBrowser if the original raw file or files are missing. The MSF file and the raw file must reside in the same directory.

   The Qual Browser application opens, as shown in Figure 108.

Figure 108. The Qual Browser application window



2. Right-click the lower pane and choose **Display Options** from the shortcut menu.
3. To automatically annotate your peaks with the elemental composition, theoretical mass, RDB equivalent, or mass delta, click the **Composition** tab and select the labels for display.

Customizing Cleavage Reagents

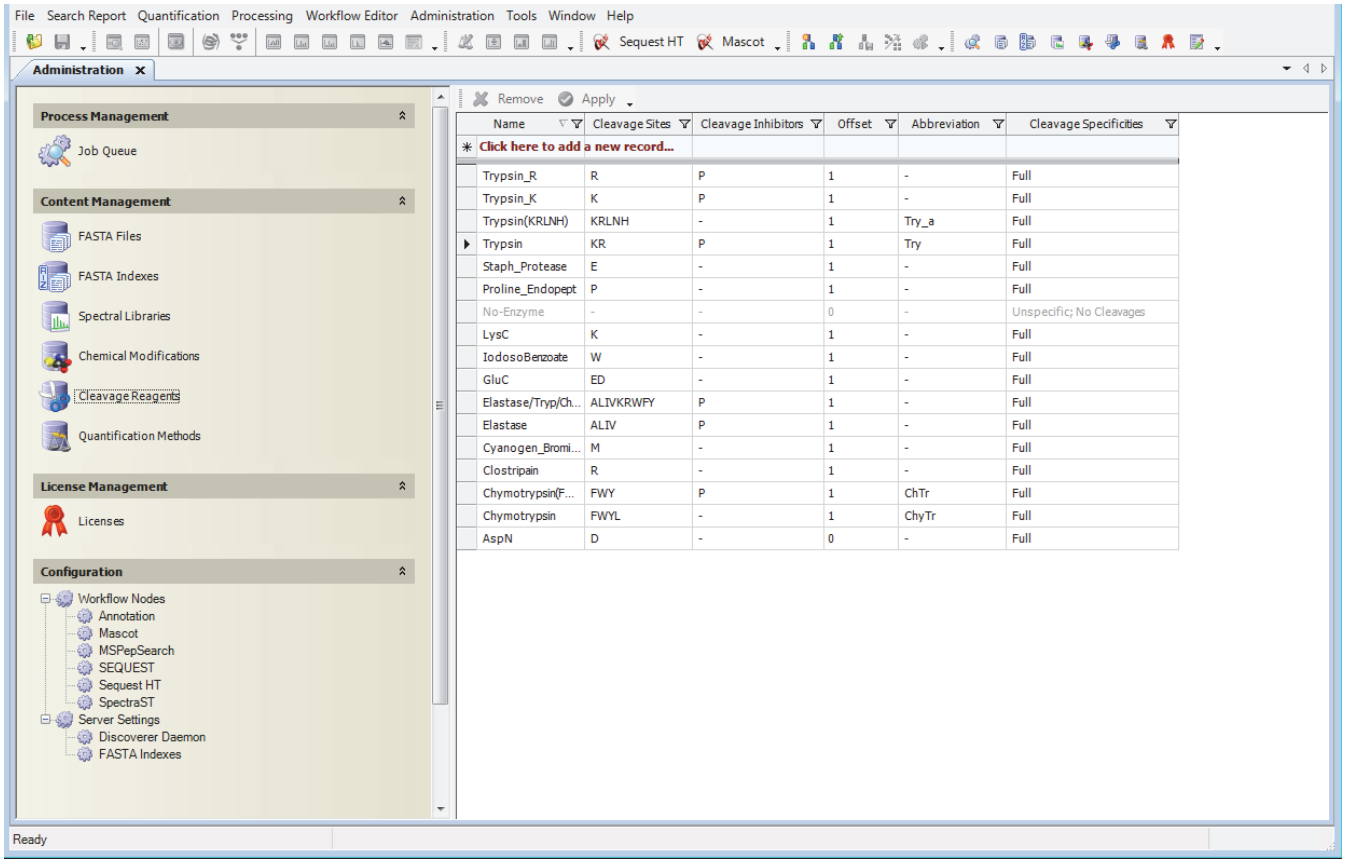
In the Cleavage Reagents view, you can explore the default types of reagents and their corresponding settings. You can also add, remove, and modify the reagents and their corresponding settings. The Cleavage Reagents view contains the cleavage sites, cleavage inhibitors, abbreviations, and cleavage specificities.

❖ **To display the Cleavage Reagents view**

- Choose **Administration > Maintain Cleavage Reagents**, or click the **Maintain Cleavage Reagents** icon, , on the toolbar or on the Administration page.

  The Cleavage Reagents view appears, as shown in Figure 109.

Figure 109. Cleavage Reagents view



Adding a Cleavage Reagent

❖ **To add a new cleavage reagent**

1. Click the Name column cell and click **Click Here To Add a New Record**.
2. Modify the default values in the row of that new reagent.
3. Click **Apply**.

Deleting a Cleavage Reagent

❖ To delete a cleavage reagent

1. Click the box in the * column next to the row that you want to delete.
2. Click **Delete**.
3. Click **Yes** in the confirmation box that appears.

Modifying a Cleavage Reagent

❖ To modify a cleavage reagent

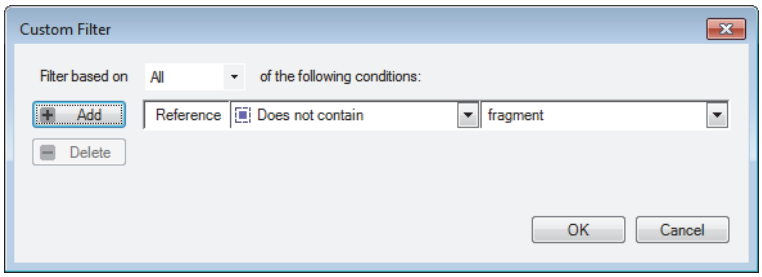
1. Click in the column for the reagent you want to modify, select the current contents, and enter the new information.
2. Click **Apply**.

Filtering Cleavage Reagent Data

❖ To filter cleavage reagent data

1. Click the **Funnel** icon,  , next to the header of the column.
2. Select one of the following:
   - All: Returns the filtered search results to the results that were first loaded.
   - Custom: Opens the Custom Filter dialog box, shown in Figure 110.

     Figure 110. Custom Filter dialog box

     For information about using this type of dialog box, see "Filtering Results with Row Filters" on page 167.
   - Blanks: Filters out rows that have data-filled cells in the column whose funnel icon you clicked.
   - NonBlanks: Filters out rows that have empty cells in the column whose funnel icon you clicked.

Filtering Data

The single or multiconsensus MSF report displays a list of matching peptides and proteins identified by the search engine that you specify. This chapter explains how to sort and filter the data from your Proteome Discoverer results report.

Contents

Result Filters Page

On the Result Filters page, shown in Figure 111, you can select the proteins and peptides to filter out of the search results. Refining your search results in this way can make your analysis quicker. By using filters, you can sort and filter your results by charge state, modifications, or even peptide probability. You can also create and apply more than one filter to your search results.

In addition to the Result Filters page, you can filter the data while opening your MSF file by setting filters on the Result Filters page that appears when you choose File > Open Report (refer to the Help). These filters are identical to the filters on the Result Filters page for an already opened MSF file, except that you can only set protein filters on the Result Filters page for an already opened MSF file.

Protein scores give some indication of the relevance of a protein. They are calculated from a list of peptides identified for a particular protein and can be expected to change as soon as the peptides are removed by the application of result filters. The Proteome Discoverer application recalculates the protein scores after you apply peptide filters or change the score thresholds on the Peptide Confidence page. For information on how the application calculates protein scores, refer to the Help.

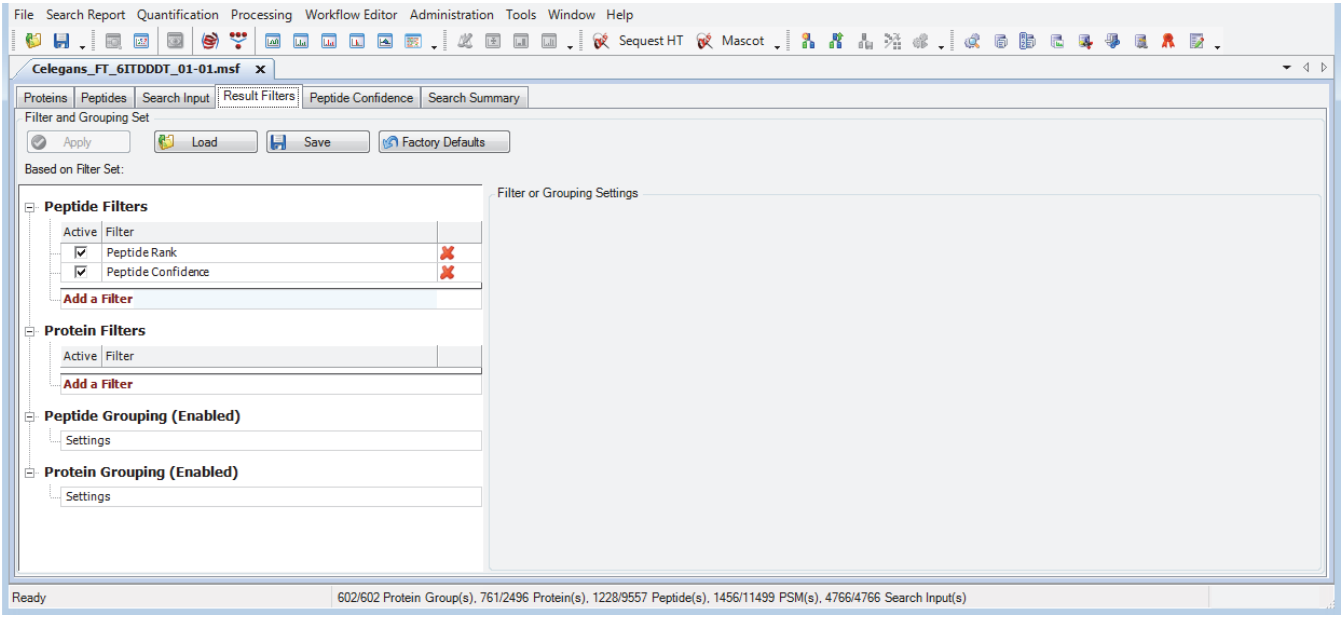
To filter the number of proteins and peptides visible on the Proteins and Peptides pages in an MSF file that is already open, use the Result Filters page, shown in Figure 111. For information about filtering MSF files while opening them, refer to the Help.

❖ To display the Result Filters page in an open MSF file

- In an open report, click the **Result Filters** tab.

  The Result Filters page of your results report appears, as shown in Figure 111.

Figure 111. Result Filters page



Filtering the Search Results

You can use Proteome Discoverer application filters to selectively hide and sort the visible results of the matched search results. You have two methods of filtering your search results data:

- Results filters on the Result Filters page exclude peptides and proteins from the results on the Proteins and Peptides pages. Applying these filters to filter out peptides, does the following:
  - Changes the number of identified peptides and the percentages shown in the Coverage column of the Proteins page.
  - Affects the numbers of filtered peptides and proteins versus the total number of peptides and proteins displayed in the Result Items Per File area at the bottom of the Input Files/Result Filters page.
  - Affects the quantification results of proteins.

For information about filtering with the Result Filters page, see "Filtering Results with the Filters on the Result Filters Page" on page 155.

- Row filters on the shortcut menu of the Proteins, Peptides, and Search Input pages display filters only. Use these filters with the filters on the Result Filters page to narrow your search results even further. When you display the filtered-out rows, the affected lines for both filters are seen as unavailable rows. Excluding peptides by setting row filters does not change the number of identified peptides and the percentage coverage values of the proteins. For information about filtering with row filters, see "Filtering Results with Row Filters" on page 167.

If you save your report, you can save the filters that you set on the Result Filters page with your results report. You cannot save the filters that you set with the row filters with your results report. The row filters only work on the visible rows in the report. However, you can save the row filters in a saved layout. For information about saving layouts, refer to the Help.

Filtering Results with the Filters on the Result Filters Page

The following procedures describe how to filter your results using the result filters on the Result Filters page.

- Filtering Search Results with Protein Filters
- Filtering Search Results with Peptide Filters
- Filtering Peptides by Rank
- Filtering Peptides by the Delta Cn Value
- Filtering Results by the Original Rank Assigned by the Search Engine

Filtering Search Results with Protein Filters

Follow this procedure to apply protein filters to your search results.

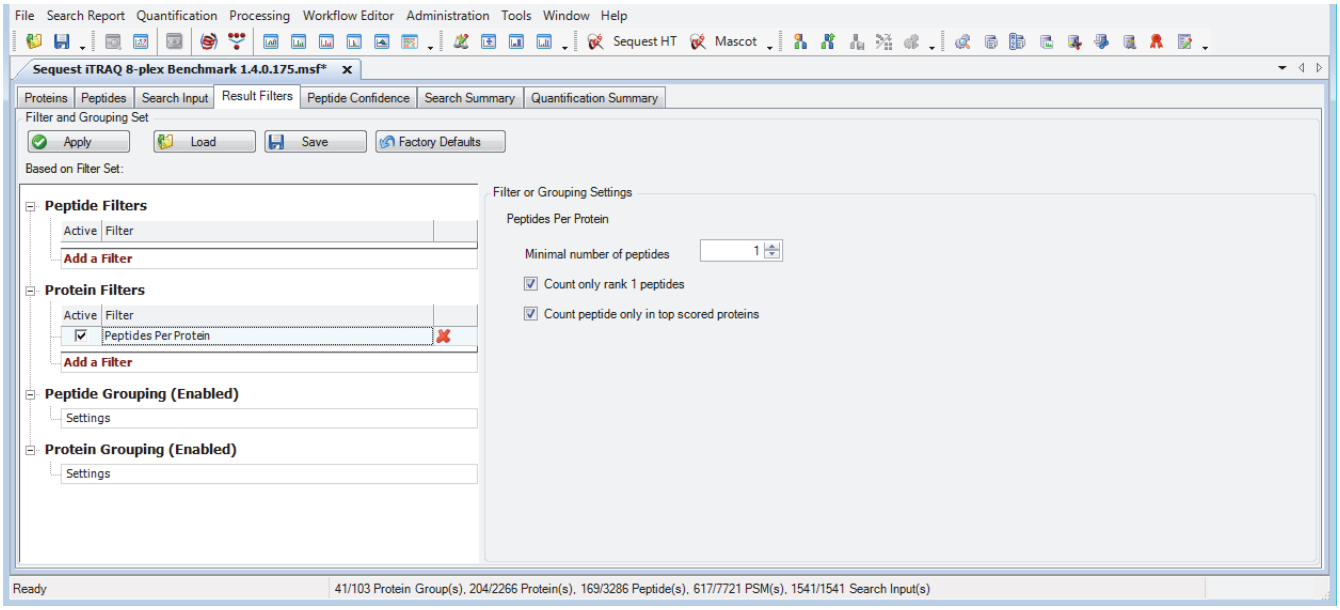
❖ **To filter your search results with protein filters**

1. Open your search results. Refer to the Help.
2. Click the **Result Filters** tab, which is shown in Figure 111 on page 154.
3. Click **Add a Filter** in the Protein Filters area.

   A list of filters appears. For a description of the available filters, refer to the Help.
4. Select the filter to apply from the list of filters.

   Settings pertaining to the selected filter appear in the Filter or Grouping Settings area on the right, as shown in Figure 112. For a description of the available settings, refer to the Help.

Figure 112. Protein filter options



5. Set the options pertaining to the selected filter in the Filter or Grouping Settings area. For example, in Figure 112, you can set the Minimal Number of Peptides and also select the Count Only Rank 1 Peptides and the Count Peptide Only in Top Scored Proteins options.
6. If it is not already selected, select the check box in the Active column. (The check box is selected by default.)
7. To remove a filter before you apply it, click .
8. To update the search results, click Apply in the Filter and Grouping Set area.

Note The Proteome Discoverer application might take several seconds to display the filtered data.

Filtering Search Results with Peptide Filters

Follow this procedure to apply peptide filters to your search results.

❖ **To filter your search results with peptide filters**

1. Open your search results. Refer to the Help.
2. Click the **Result Filters** tab, which is shown in Figure 111 on page 154.
3. Click **Add a Filter** in the Peptide Filters area.

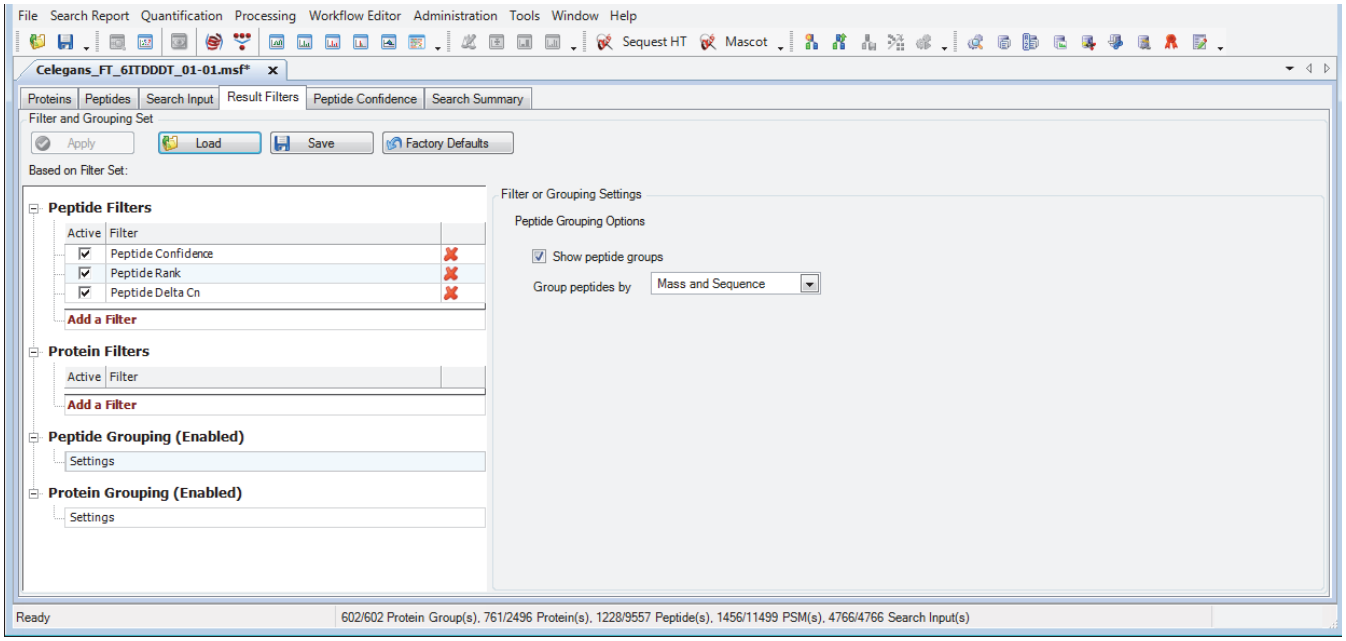
   A list of filters now appears. For a description of the filters available, refer to the Help.

   The Peptide Rank and Peptide Confidence filters are selected by default.

4. Select the filter to apply from the list of filters.

   Options pertaining to the selected filter now appear in the Filter or Grouping Settings area, as shown for the Peptide Score filter in Figure 113. The Help describes these options.

Figure 113. Peptide filter options



5. Set the options pertaining to the selected filter in the Filter or Grouping Settings area. For example, in Figure 113, you can set the Show Peptide Groups option and the Group Peptides By option.

6. If it is not already selected, select the check box in the Active column (it is selected by default).

7. To remove a filter before you apply it, click .

8. To update the search results, click Apply in the Filter and Grouping Set area.

Note The Proteome Discoverer application might take several seconds to display the filtered data.

Filtering Peptides by Rank

From the acquired MS/MS spectra, search engines like Sequest HT or Mascot create a list of possible peptides whose masses match the measured mass of the precursor ions of the MS/MS spectrum and whose fragmentation patterns match the peaks detected in the MS/MS spectrum. The better the match, the better the score of every peptide candidate considered. The Proteome Discoverer application ranks all considered peptide candidates by their scores and reports a user-specified number of peptide candidates per spectrum. The default is

usually 10. The rank of a peptide is its position in the reported list of identified peptide candidates per spectrum that is ordered from better to worse scores. Peptides with a top ranking (for example, 1 or 2) are more likely to be the correct peptide than peptides with a lower ranking (for example, less than 2).

The Proteome Discoverer application does not store the peptide rank in the results file but calculates it after loading the results file. Only loaded peptides affect the peptide rank. The Proteome Discoverer application loads peptides that pass all other peptide filters before applying the Peptide Rank filter. It rejects those peptides that do not pass the Peptide Rank filter.

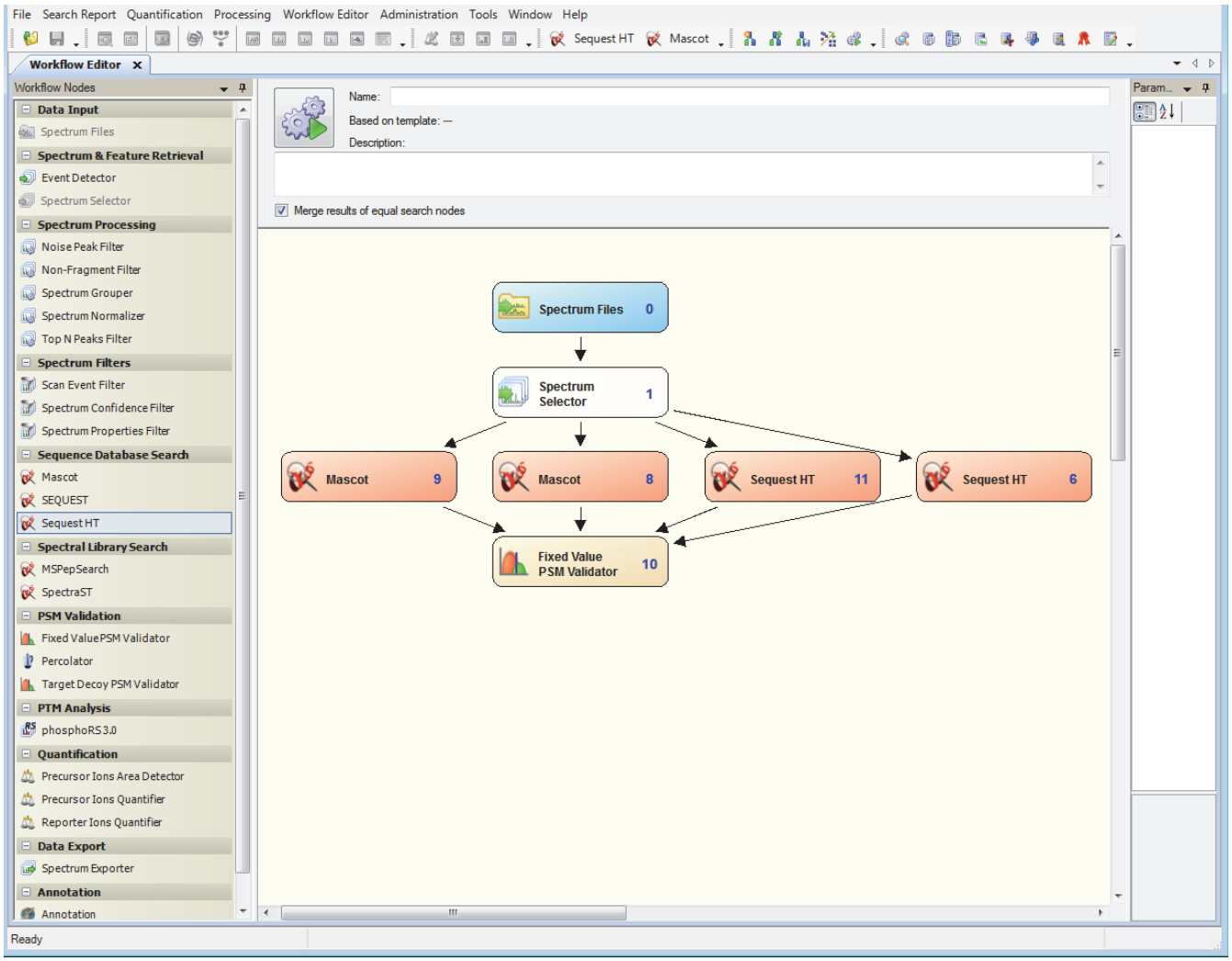
You can use the Peptide Rank filter to filter out peptides with a rank higher than the maximum rank that you specify with the Maximum Peptide Rank option.

Calculating Peptide Rank

The Merge Results of Equal Search Nodes option in the Workflow Editor determines whether peptides and proteins identified by the same type of search engine are merged together. If you select this option, the Proteome Discoverer application ranks the peptides identified by the same search engine together. Only one peptide can have rank 1 for each spectrum and search engine. If you do not select the Merge Results of Equal Search Nodes option, the Proteome Discoverer application ranks peptides identified by one search engine independently from the peptides identified by another search engine. Therefore, there can be multiple peptides having rank 1 for each spectrum.

For example, consider the workflow with two SEQUEST nodes and two MASCOT nodes shown in Figure 114.

Figure 114. Workflow with two Mascot nodes and two SEQUEST nodes



The search engines find the peptides shown in Table 8 for spectrum 10:

Table 8. Peptides found for spectrum 10

Sequest (2)	Sequest (3)	Mascot (4)	Mascot (5)
Peptide 2.1 (XCorr = 20)	Peptide 3.1 (XCorr = 12)	Peptide 4.1 (IonScore = 33)	Peptide 5.1 (IonScore = 34)
Peptide 2.2 (XCorr = 8)	Peptide 3.1 (XCorr =12)		

If you selected the Merge Results of Equal Search Nodes option, peptides 4.1 and 4.2, which Mascot identified, are ranked together. Peptides 2.1, 2.2, 3.1, and 3.2, which Sequest identified, are ranked together.

If you did not select the Merge Results of Equal Search Nodes option, peptides 4.1 and 4.2, which Mascot identified, are ranked independently. Sequest-identified peptides 2.1 and 2.2 are ranked together, and peptides 3.1 and 3.2 are ranked together.

To calculate the rank, the Proteome Discoverer application sorts all peptides belonging together by their main score. For Sequest, the main score is XCorr. For Mascot, the main score is IonScore. Peptides with the same main score have the same rank.

For example, if you selected the Merge Results of Equal Search Nodes option, the Proteome Discoverer application ranks the peptides shown in Table 8 as follows:

Sequest:

- Peptide 2.1 (XCorr = 20): Rank 1
- Peptide 3.1 (XCorr = 12): Rank 2
- Peptide 3.2 (XCorr = 12): Rank 2
- Peptide 2.2 (XCorr = 8): Rank 4

Mascot:

- Peptide 5.1 (IonScore =n 34): Rank 1
- Peptide 4.1 (IonScore = 33): Rank 2

If you did not select the Merge Results of Equal Search Nodes option, the Proteome Discoverer application ranks the peptides shown in Table 8 on page 159 as follows:

Sequest (2):

- Peptide 2.1 (XCorr = 20): Rank 1
- Peptide 2.2 (XCorr = 8): Rank 2

Sequest (3):

- Peptide 3.1 (XCorr = 12): Rank 1
- Peptide 3.2 (XCorr = n12): Rank 1

Mascot (4):

- Peptide 4.1 (IonScore = 33): Rank 1

Mascot (5):

- Peptide 5.1 (IonScore = 34): Rank 1

Recalculating Peptide Rank

The Proteome Discoverer application does not consider filtered-out peptides in calculating peptide ranks. Filtered-out peptides have a rank of infinite. If you apply filters to an open MSF report, the application recalculates the peptide ranks. It also recalculates the delta score values each time that the peptide ranks change.

Using the Peptide Rank Filter

If you use the Peptide Rank filter when you open a report, the Proteome Discoverer application reads the peptides twice. In the first step, it collects identifications and the main scores of all peptides passing the peptide filters except the Peptide Rank filter. Then it calculates the ranks for these peptides and loads all peptides having a higher rank than the maximum allowed rank. It loads the remaining peptides in the second step.

If you apply the Peptide Rank filter to an open report, the application filters out those peptides that do not pass the peptide filters except the Peptide Rank filter. It calculates the ranks for the remaining peptides. Finally, it applies the Peptide Rank filter.

❖ To filter peptides by rank

1. Open the MSF file. Refer to the Help.
2. Click the **Result Filters** tab.
3. Select **Peptide Rank** in the Peptide Filters area of the Result Filters page, if it is not already selected.

   The Maximum Peptide Rank option appears in the middle of the Result Filters page.
4. (Optional) In the Maximum Peptide Rank box, set the maximum rank that a peptide must have to avoid being filtered out.

   The minimum value is 1, and there is no maximum value. The default value is 1.

Filtering Peptides by the Delta Cn Value

Search engines often provide multiple possible matching peptides as explanations for the same spectrum. Most of the time you can clearly distinguish the top-scoring match from the other PSMs, but sometimes, especially in the presence of dynamic modifications, the best-scoring matches of the same spectrum have very similar scores. In this case, you can filter the results to select the best-scoring PSMs and the matches that have very similar scores by using the Δ Cn peptide filter.

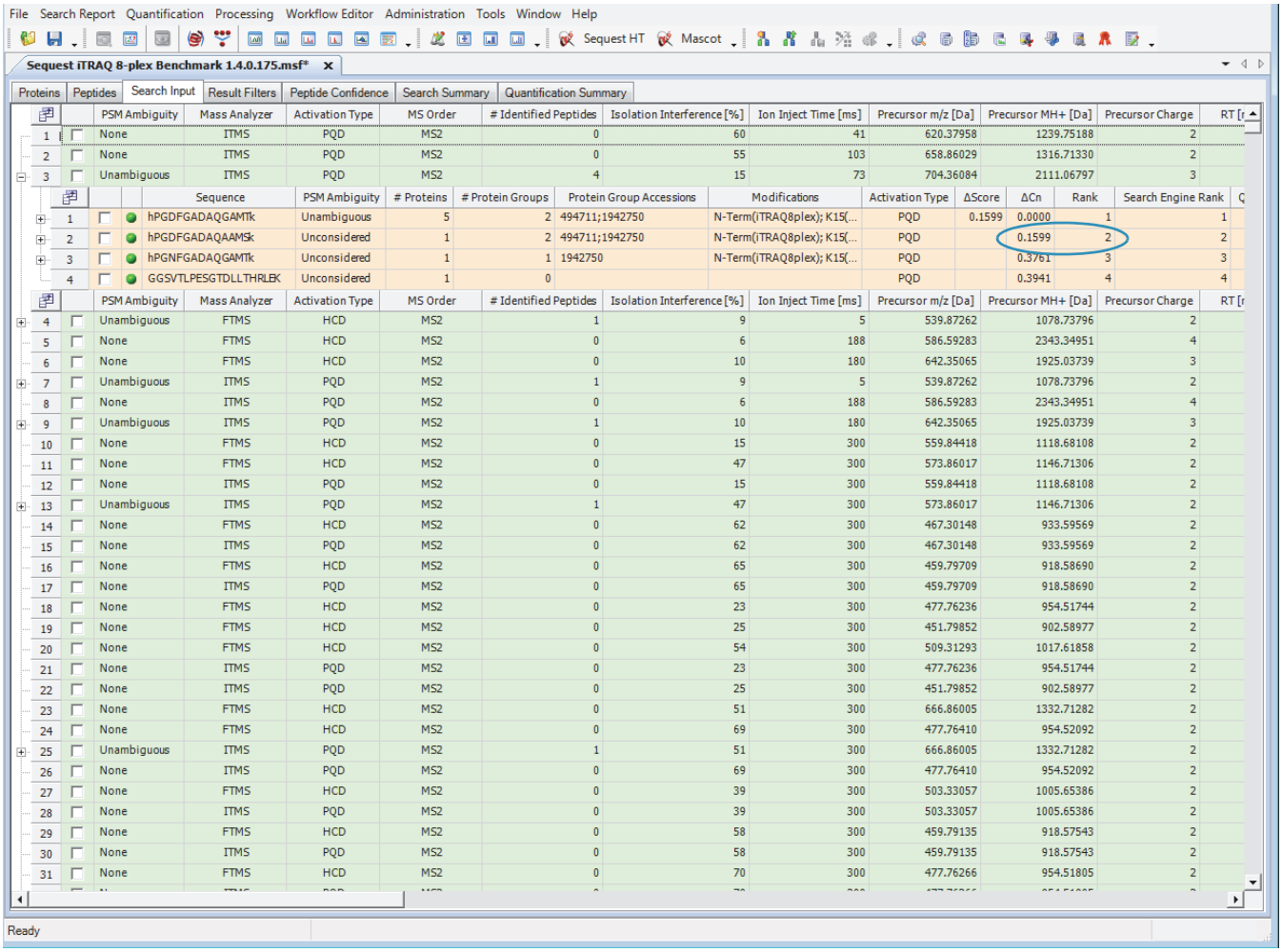
The Δ Cn value displays the normalized score difference between the currently selected PSM and the highest-scoring PSM for that spectrum:

$$\Delta\mathrm{Cn}(rank\ i) = \frac{score_{rank1} - score_{rank1}}{score_{rank1}}$$

The Δ Cn peptide filters out all PSMs with a Δ Cn score larger than the specified value.

On the Peptides page or the peptides sections of the Proteins and Search Input pages, the Δ Cn column displays the Δ Cn values. For example, Figure 115 shows how the score of a peptide ranked 2 compares to other multiple high-confidence peptides from the same spectrum.

Figure 115. Δ Cn scores for multiple high-confidence peptides from the same spectrum



❖ To filter peptides by the Δ Cn value

1. Open the MSF file. Refer to the Help.
2. Click the **Result Filters** tab.
3. Select **Peptide Delta Cn** in the Peptide Filters area of the Result Filters page.

   The Peptide Delta Cn option appears in the middle of the Result Filters page.

4. (Optional) In the Maximum Delta Cn box, specify a Δ Cn threshold that will filter out all PSMs with a Δ Cn larger than this value.

   The minimum value is 0.0, and the maximum value is 1.0.

Filtering Results by the Original Rank Assigned by the Search Engine

If you apply PSM-level result filters, the Proteome Discoverer application dynamically recalculates the displayed ranks, delta scores, and Δ Cn values. However, you can also view the original rank assigned by the search engine for all PSMs and peptide groups by displaying the Search Engine Rank column on the Peptides page. In addition, you can filter by this rank.

For example, you might find this feature helpful when you know that your raw data has a true mass accuracy below 5 ppm. If you search this data with a precursor tolerance of 5 ppm and validate it by calculating FDRs, you obtain false positive matches within this mass deviation tolerance. You could find some of these incorrect matches if you searched the data with a larger precursor tolerance, such as 50 ppm. This step increases the chance of replacing incorrect matches with a mass deviation below 5 ppm by incorrect matches with a higher mass deviation. When you review the results, you can set a mass deviation filter of more than 5 ppm to remove all matches that have a mass deviation outside the true mass accuracy. You can now find many of the remaining incorrect matches. They have a Search Engine Rank worse than rank 1, because they were initially replaced by incorrect matches with a larger mass deviation.

Using Filter Sets

You can save your selected filter settings as a group for future use. You can also save your protein and peptide grouping settings as a set. You can make this set the default or assign it a name. These sets are saved in and loaded from external files so that you can export filter sets from one instance of the Proteome Discoverer application and import them into another instance. The filter sets have an extension of .filters.

If you want to use a filter set from one installation of the Proteome Discoverer application in another installation of the Proteome Discoverer application, you must copy the filter set from the root directory of the first installation to the root directory of the other installation.

You can create these filter setting groups on the Result Filters page that appears during report loading or on the Result Filters page that appears after the report has already been opened.

You can load a previously stored filter set. Loading a filter set replaces the currently set peptide and protein filters and the settings for the protein grouping with the filters and settings stored in the loaded filter set, unless the filters were loaded before the MSF file was opened.

- To create and save a filter set
- To load a filter set
- To delete a filter set

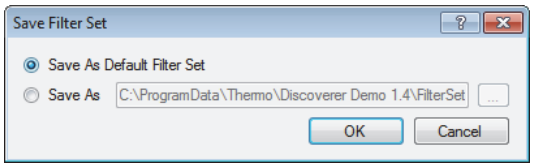
- To clear the default filter set
- To restore the default filter set in effect after installing the Proteome Discoverer application

❖ To create and save a filter set

1. For filters, select the appropriate protein and peptide filters, as described in "To filter your search results with peptide filters" on page 156 and "To filter your search results with protein filters" on page 155, and click Apply.
2. In the Filter and Grouping Set area, click Save.

   The Save Filter Set dialog box appears, as shown in Figure 116.

   Figure 116. Save Filter Set dialog box

3. In the Save Filter Set dialog box, do one of the following:

   - To save the filter set or set of protein grouping settings as the default filter set, select the **Save As Default Filter Set** option.

     The Proteome Discoverer application automatically applies this filter set to the opened MSF results file.

     –or–

     To save the filter set in a file, select the **Save As**. option. Click the **Browse** button (...) and browse to the file to save it in. You can also type the name of a new file in the box next to Save As.

   - Click **OK** in the Save Filter Set dialog box.

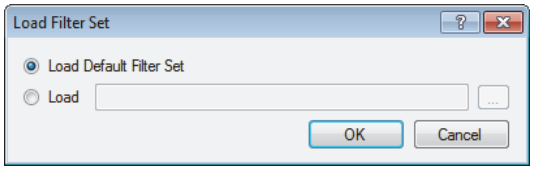
   The saved filter set appears in the list in the Filter and Grouping Set area. The default set is named "Default" in this list.

❖ To load a filter set

1. In the Filter and Grouping Set area, click Load.

   The Load Filter Set dialog box appears, as shown in Figure 117.

   Figure 117. Load Filter Set dialog box

2. In the Load Filter Set dialog box, do the following:

   a. To load the default filter set, select the **Load Default Filter Set** option.

      –or–

      To load another filter set, click the **Browse** button (...), and select the file containing the filter set that you want to load. You can also type the name and path of the file to load in the box next to Load.

   b. Click **OK** in the Load Filter Set dialog box.

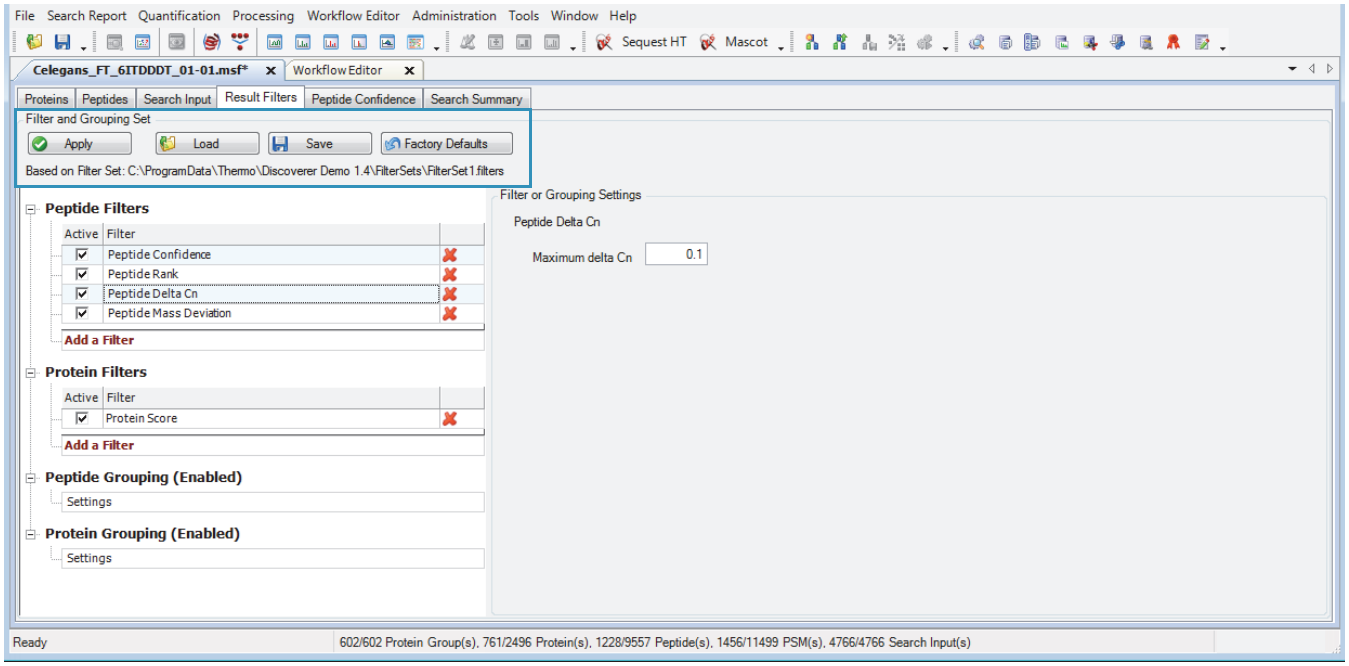
   A Loading Filter Set confirmation box appears if you have already selected other filter settings.

3. If the Loading Filter Set confirmation box appears, click **OK**.

4. If you are loading a filter set on the Results Filter page in an open MSF file, click Apply. If you are loading a filter set on the Results Filter page during report loading, the Proteome Discoverer application automatically applies the filters or sets.

   The name and path of the selected filter set appear in the Filter and Grouping Set area of the page, as shown in Figure 118.

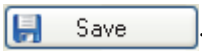
Figure 118. Loaded filter set



❖ To delete a filter set

- Click ✖ next to the peptide filters, protein filters, or grouping sets that compose the set.

❖ To clear the default filter set

1. Remove all the peptide and protein filters from the Result Filters page by clicking  next to the peptide filters, protein filters, or grouping sets that compose the set.
2. Click 

.
3. In the Save Filter Set dialog box, shown in Figure 116 on page 164, select the **Save As Default Filter Set** option.
4. Click **OK**.

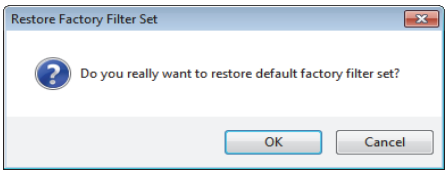
❖ To restore the default filter set in effect after installing the Proteome Discoverer application

1. Click 

.

   The confirmation box shown in Figure 119 appears.

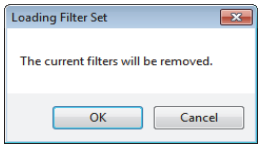
   Figure 119. Restore Factory Filter Set confirmation box

2. Click **OK**.

   The confirmation box shown in Figure 120 appears.

   Figure 120. Loading Filter Set confirmation box

3. Click **OK**.

Removing and Deactivating Filters

You can remove or deactivate filters to alter the search results.

- To remove a filter
- To deactivate a filter

❖ To remove a filter

1. Open your search results.
2. Click the **Result Filters** tab.
3. Select the filter in the list of filters in the Peptide Filters or Protein Filters area.

4. Click  .

   The filter is removed from the list of filters.

5. Click **Apply** to update the Proteins, Peptides, or Search Input page.

❖ **To deactivate a filter**

1. Open your search results.
2. Click the **Result Filters** tab.
3. Clear the check box in the Active column.
4. To update the Proteins, Peptides, or Search Input page, click **Apply**.

   The filter is deactivated but not removed from the Result Filters page.

Filtering Results with Row Filters

The following procedures describe how to set and clear basic row filters, display filtered-out rows, use row filters to filter precursor masses, and filter peptides and proteins by site localization scores from phospho*RS*.

- Setting and Clearing Row Filters
- Displaying Filtered-Out Rows
- Filtering Precursor Masses
- Filtering PSMs and Peptides for Site Localization Scores from phosphoRS
- Grouping Proteins

Setting and Clearing Row Filters

You can use row filters on the Proteins, Peptides, and Search Input pages to set up simple filter criteria that only consist of a single filter statement, such as "*number* is greater than 5," or "*text* contains kinase."

❖ **To filter your search results using row filters**

1. Open your search results.
2. Select the **Proteins**, **Peptides**, or **Search Input** page.
3. Ungroup the peptides by right-clicking and choosing **Show Peptide Groups**.
4. Right-click to access the shortcut menu and choose **Enable Row Filters**.

A filter row appears beneath the column header that contains the icons shown in Figure 121. For a description of these icons, see the Help. You can select an operator, enter the filter value, clear the currently set filter, or open the Enter filter criteria for *header_name* dialog box for more complex transactions.

Figure 121. Row filter icons

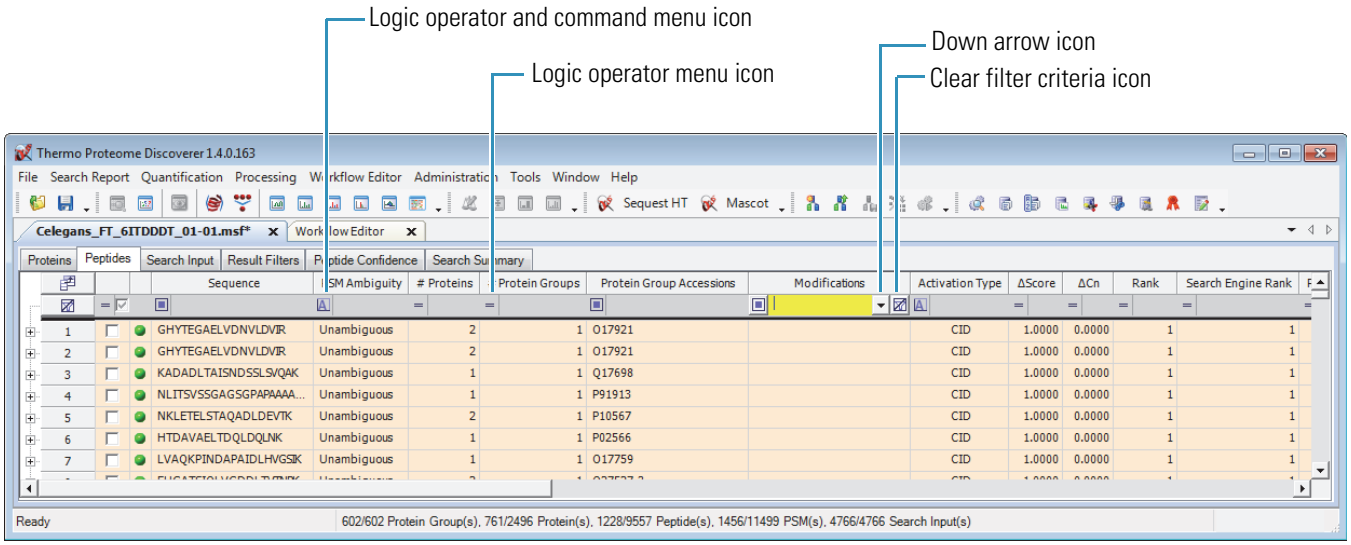


Figure 122. Logic operator and command menu

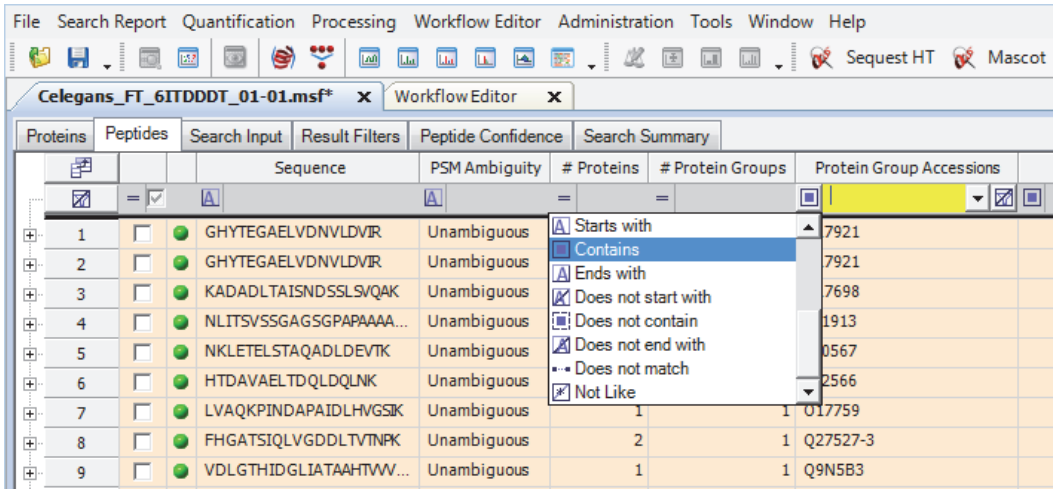


Figure 123. Logic operator menu

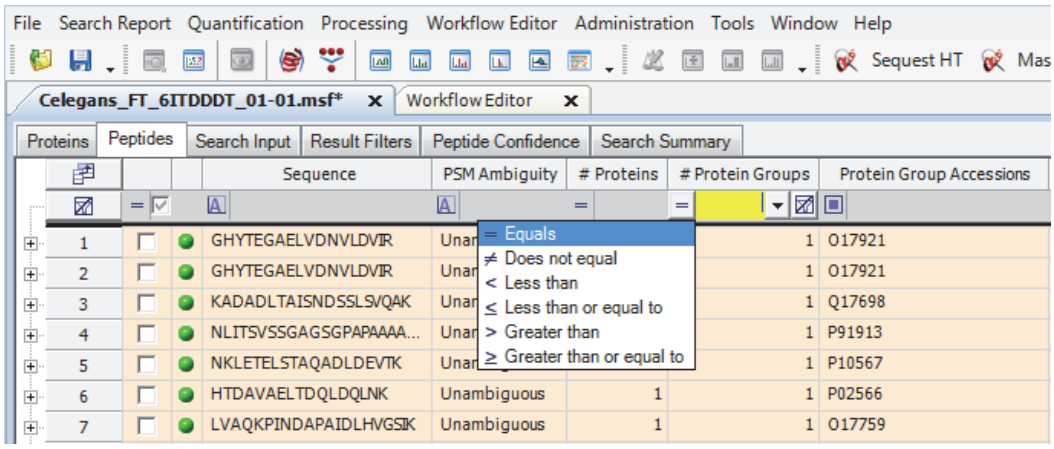
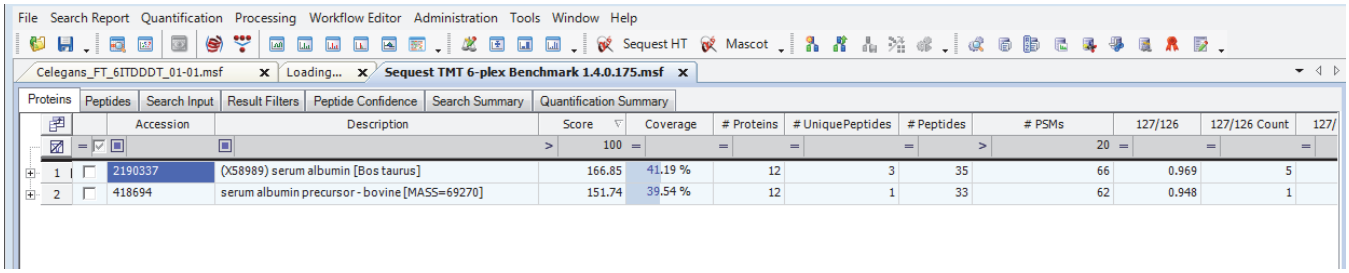


Figure 124 gives an example of simple filter criteria being entered in the row filter line. In this example, Score is set to be greater than 100, and # PSMs is set to be greater than 20.

Figure 124. Setting row filter criteria



The following example shows how to use the row filter menu opened by the down arrow icon in the MSF report columns, , and the Enter Filter Criteria for *Header_name* dialog box. This example sets a precursor mass filter.

❖ To clear all filter conditions set by the row filter menu

- Click the **Clear Filter Criteria** icon, , if you want to clear all filter criteria set by the commands on the row filter menu (opened by clicking).

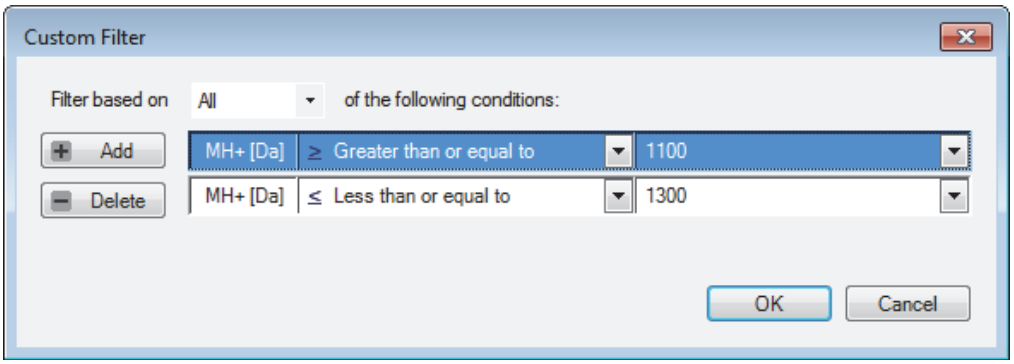
❖ To clear an individual filter set by the row filter menu

1. In the appropriate column, move your cursor over the row with the filter set by the commands on the row filter menu (opened by clicking).
2. Click and choose **Custom**.

   The Custom Filter dialog box appears. For information on the parameters in the Custom Filter dialog box, see the Help.
3. In the dialog box, click in the first column in the row of interest.

   The condition is activated, as shown in Figure 125.

Figure 125. Deleting filter condition



4. Click **Delete**.
5. Click **OK**, or if you are deleting all filters, click **No Filters**, which appears instead of the OK button.

Displaying Filtered-Out Rows

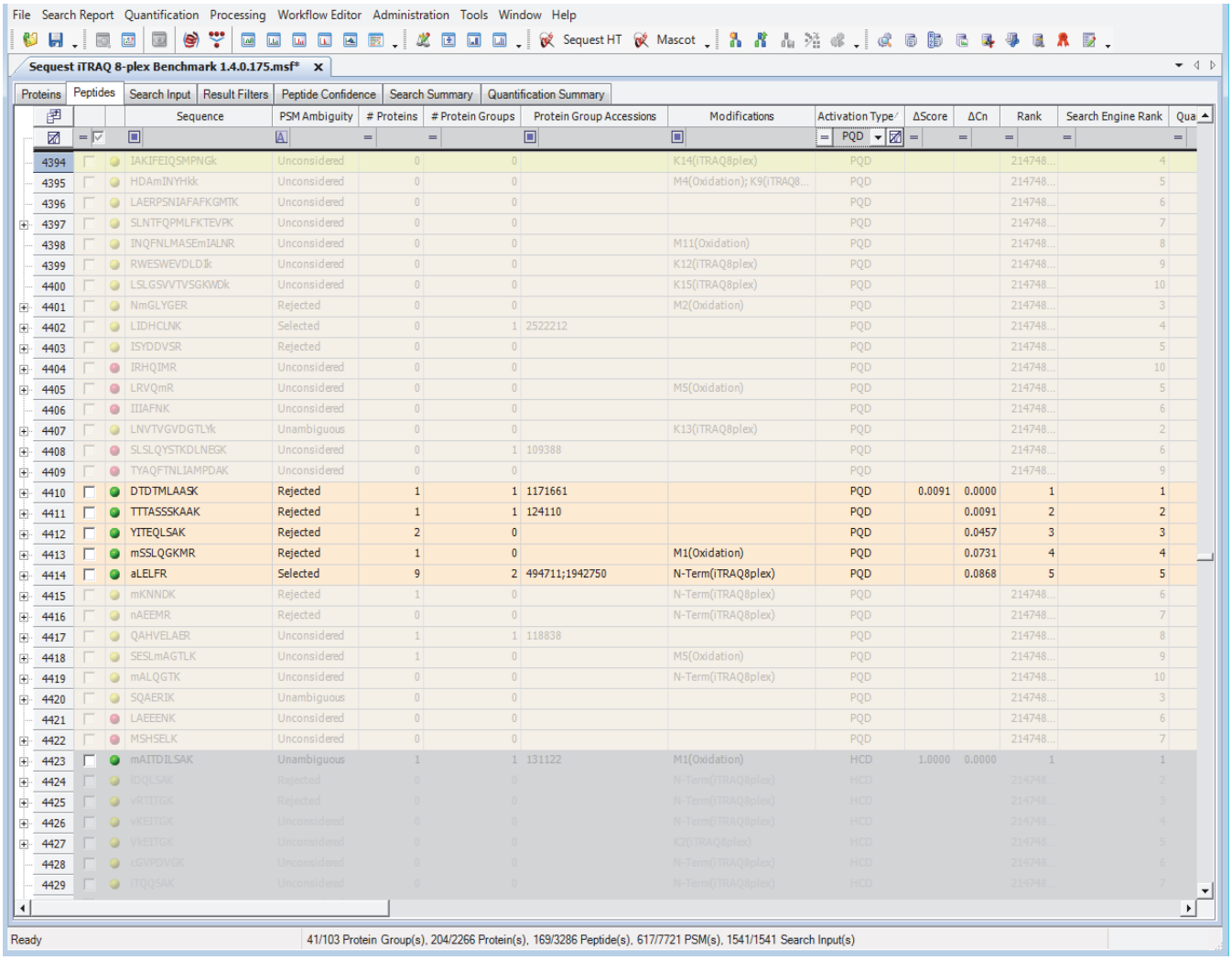
If you choose Enable Row Filters in the shortcut menu, the Proteome Discoverer application hides the filtered-out rows on the Proteins, Peptides, or Search Input page so that you can easily view your results. However, you can still display these filtered-out rows to perform a comparative analysis.

❖ To display filtered-out rows

1. Follow the procedure given in "Filtering Results with Row Filters" on page 167 to set any row filters.
2. Right-click to display the shortcut menu, as shown in Figure 126, and choose **Show Filtered Out Rows**.

   The application now displays both the filtered-out and unfiltered rows. The peptides or proteins filtered out by filters set on the Result Filters page appear in light gray rows. The peptides or proteins filtered out by row filters appear in darker gray rows. Figure 126 shows both types of filtered-out rows.

Figure 126. Displaying filtered-out rows



3. To hide the filtered-out rows after you have displayed them, right-click the page and again choose **Show Filtered Out Rows**.

Filtering Precursor Masses

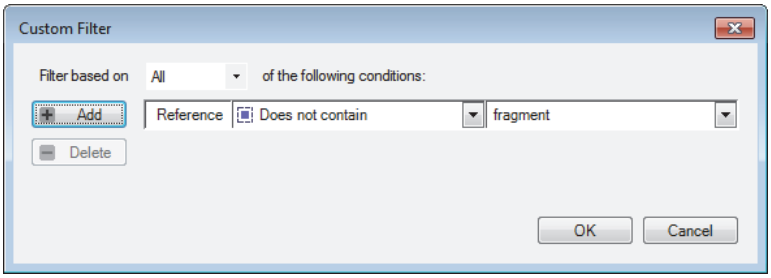
You can set filter criteria to display peptides that have precursor masses between certain specified values.

❖ To set a precursor mass filter by using the row filter menu

1. Click and choose **Custom** from the menu.

   The Custom Filter dialog box appears, as shown in Figure 127.

Figure 127. Custom Filter dialog box



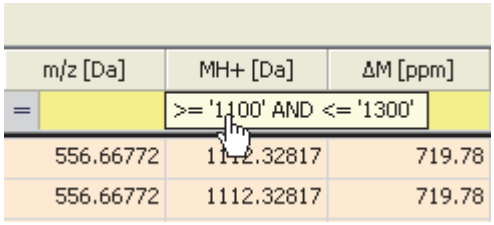
2. From the list in the center, select the logic operator value, for example, ≥ **Greater Than or Equal To**.
3. In the box to the right, type a value, for example, **1100**.
4. To open another row in the Custom Filter dialog box, click **Add**.
5. From the list in the center, select the logic operator value, for example, ≤ **Less Than or Equal To**.
6. In the box to the right, type a value, for example, **1300**.

   The Custom Filter dialog box should look like the example in Figure 127.
7. Click **OK** to accept the filter settings.

   In this example, only peptides that have a precursor MH+ mass between 1100 and 1300 are displayed.

   The filter conditions that you set appear when you move the cursor over the filters row, as shown in Figure 128.

Figure 128. Displaying filter conditions



Filtering PSMs and Peptides for Site Localization Scores from phospho*RS*

You can set a row filter that allows you to filter for the following:

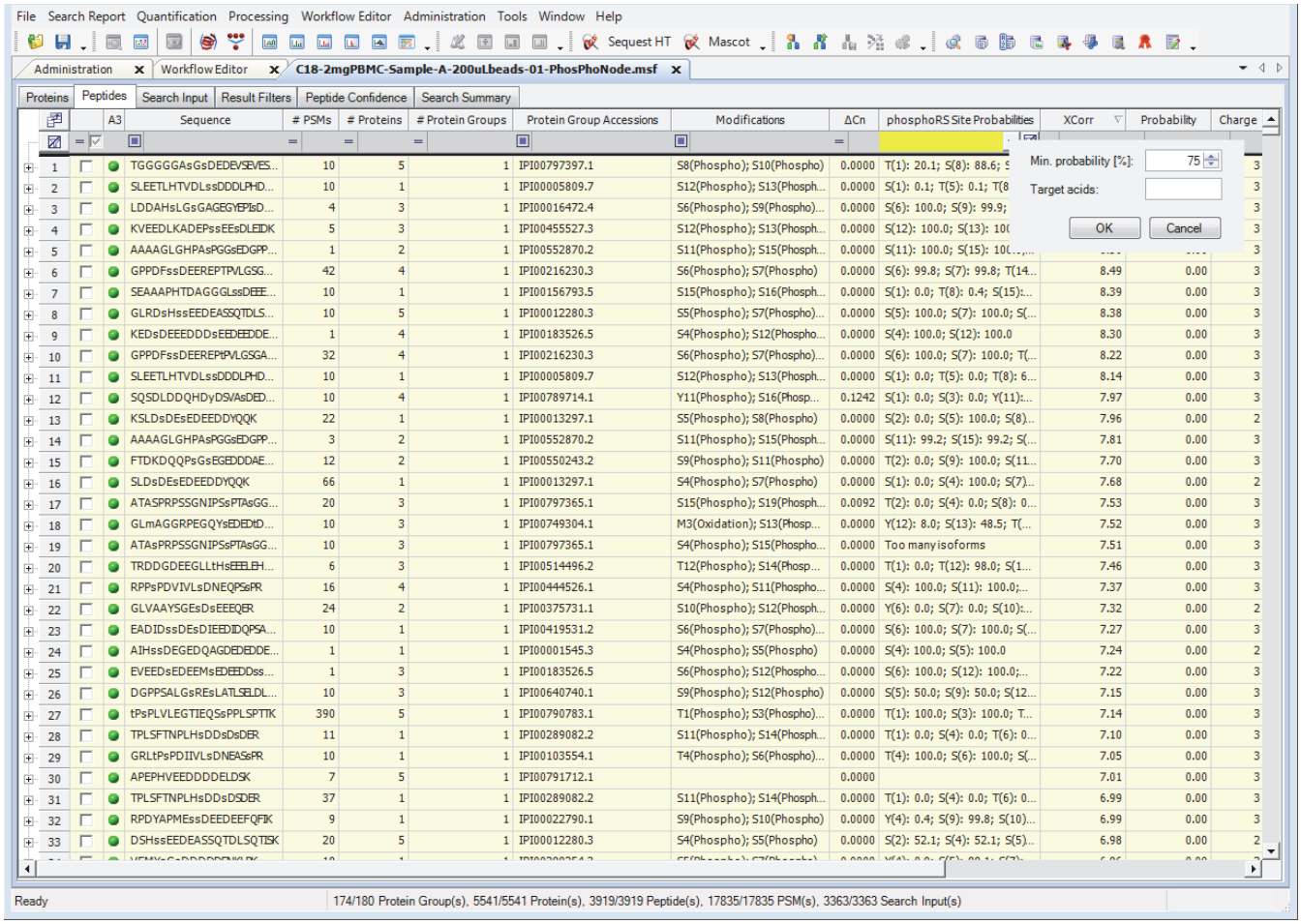
- At least one site with a localization probability equal to or above the specified value
- At least one site of the specified type (such as S, T, or Y) with a localization probability equal to or above the specified value

❖ To filter PSMs and peptides for site localization scores from phospho*RS*

1. On the Peptides page of the MSF file, right-click and choose **Enable Row Filters** to turn on the row filters.
2. In the phospho*RS* Site Probabilities column of an MSF file containing results from a phospho*RS* search, click the down arrow icon, .

   The filters shown in Figure 129 appear.

Figure 129. Row filters in the Phospho*RS* Site Probabilities column



3. In the Min. Probability [%] box, select the probability that a modification will be found on the specified amino acid.

   You can select values between 1% and 100%. The default is 75%.
4. In the Target Acids box, type the symbol or name of the amino acid.

   You can use any lowercase or uppercase letters.

If you select a target amino acid, all rows having a site probability for a target amino acid of at least the minimum value pass the filter. If you do not select any target acids, all rows containing a site probability of at least the defined minimum probability pass the filter.

5. Click **OK**.

Grouping Proteins

Although MS/MS-based proteomics studies are centered around peptides, you can also explore what proteins are present in a sample and their associations through related peptides. Deducing protein identities from a set of identified peptides becomes difficult because of sequence redundancy, such as the presence of proteins that have shared peptides. These redundant proteins are automatically grouped and are not initially displayed in the search results report.

In the results report, you can turn protein grouping on or off with the Enable Protein Grouping command on the shortcut menu or with the settings in the Protein Grouping (Enabled) area on the Result Filters page. The latter method enables you to select more options in grouping. Grouping is turned on by default. For information about the grouping mechanism that the Proteome Discoverer application uses to group proteins, see "Protein Grouping Algorithm" on page 179.

The proteins within a group are ranked according to the number of peptide sequences, the number of PSMs, their protein scores, and the sequence coverage. The top-ranking protein of a group becomes the master protein of that group. By default, the Proteins page displays only the master proteins.

Proteins are grouped according to the peptide sequences identified for the proteins. A protein group consists of the following:

- One master protein that is identified by a set of peptides that are not included (all together) in any other protein group
- All proteins that are identified by the same set or a subset of those peptides

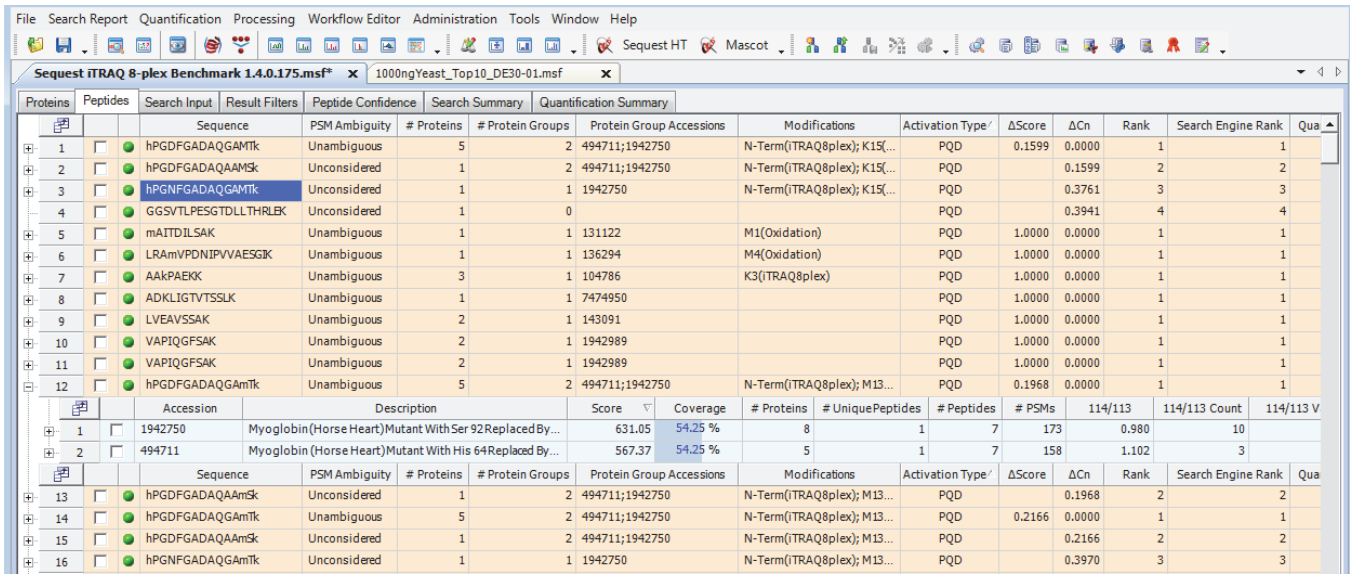
The # Proteins column on the Proteins and Peptides pages of the results report displays the number of identified proteins in the protein group of a master protein. It should match the number of proteins that are displayed in the Protein Group Members view when you choose Search Report > Show Protein Group Members (see Figure 132 on page 178).

Protein groups can overlap because proteins might be included in several master proteins. Each of two compared master proteins must have at least one peptide that is not contained in the other master protein. However, if you do not select the Apply Strict Maximum Parsimony Principle option in the Protein Grouping area of the Result Filters page, the peptides that distinguish these two master proteins could be contained in other master proteins. A master

protein does not have to contain a unique peptide, unless you select the Apply Strict Maximum Parsimony Principle option. A unique peptide is only contained in the proteins of one protein group. In the results report, the # Unique Peptides column on the Proteins page displays the number of distinct peptide sequences for a protein group.

When you expand an identified peptide, as shown in Figure 130, the Peptides page shows only the master proteins of all protein groups that contain the peptide. To display all the proteins that belong to any of the protein groups, choose **Search Report > Show Protein Group Members**, which opens the Protein Group Members view (see Figure 132 on page 178). To display all proteins that contain the peptide, choose **Search Report > Show Protein References**, which opens the Protein References of a Peptide view (see the Help). The # Unique Peptides column on the Proteins page displays the number of peptide sequences unique to a protein group.

Figure 130. Expanding an identified peptide



Go to the following sections:

- To group the proteins in your search results and set grouping options
- To display other proteins belonging to the same protein group
- To turn off protein grouping

❖ To group the proteins in your search results

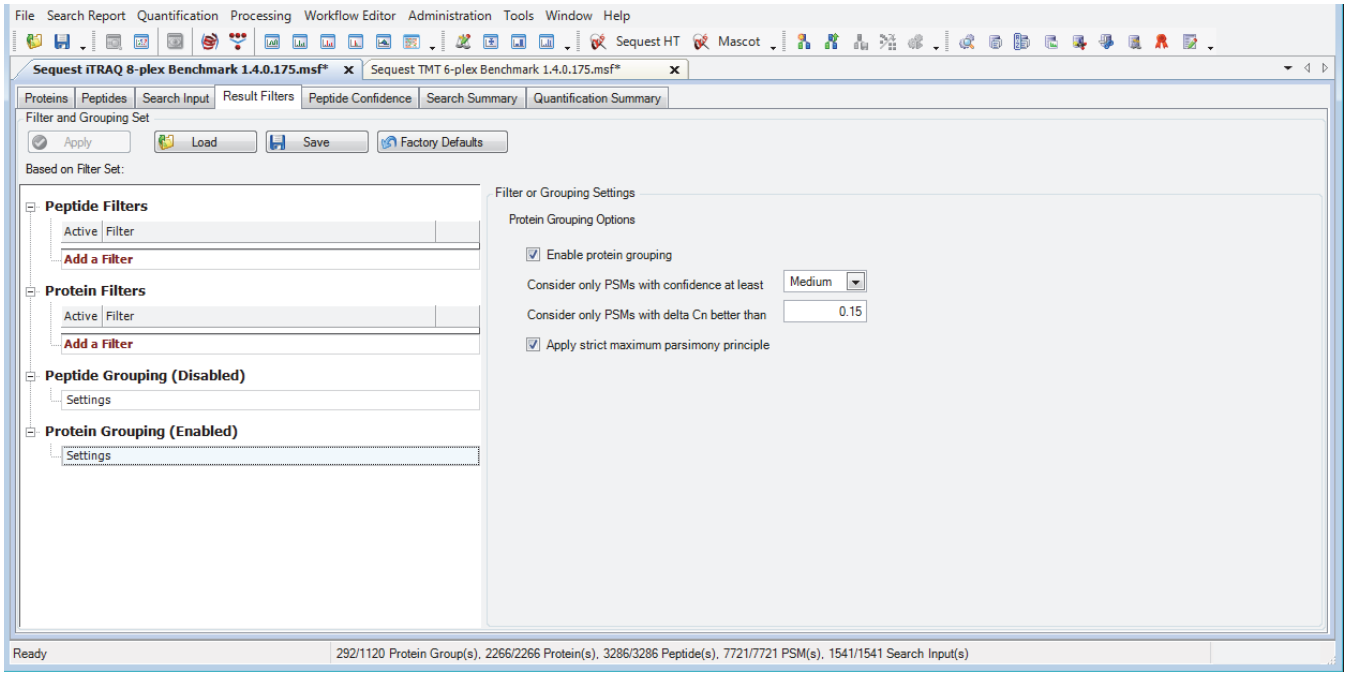
1. Open the MSF file.
2. On the Peptides or Proteins page of the MSF file, right-click a protein grid cell or row to access the shortcut menu, and choose **Enable Protein Grouping**.

❖ To group the proteins in your search results and set grouping options

1. Open the MSF file.
2. Click the **Result Filters** tab.
3. On the Results Filters page, click **Settings** beneath Protein Grouping.

   Protein grouping options appear in the Filter or Grouping Settings area, as shown in Figure 131.

Figure 131. Protein grouping options



4. If you want to group homologous proteins, select the **Enable Protein Grouping** check box, if it is not already selected by default.
5. To specify the type of PSMs that the Proteome Discoverer application considers for inclusion in protein grouping, set the **Consider Only PSMs with Confidence at Least** parameter to the desired setting:
   - Low: Considers all (low-, medium-, and high-confidence) PSMs for inclusion in protein grouping.
   - (Default) Medium: Considers medium- and high-confidence PSMs for inclusion in protein grouping.
   - High: Considers high-confidence PSMs for inclusion in protein grouping.
6. If you want the Proteome Discoverer application to consider only PSMs with values lower than or equal to a specified value for inclusion in the protein grouping process, specify a value in the **Consider Only PSMs with Delta Cn Better Than** box.

The default Δ Cn value is 0.15. To have the Proteome Discoverer application consider all PSMs, set the value to 1.0.

7. If you want to remove all protein groups that are not necessary to explain the found peptides, select the **Apply Strict Maximum Parsimony Principle** check box.

   The Apply Strict Maximum Parsimony Principle option ensures that only one PSM per spectrum is used for protein grouping. If the Δ Cn range of the spectrum includes more than one PSM, the Proteome Discoverer application selects the "best" PSM and rejects the others for grouping and quantification.

8. Click 

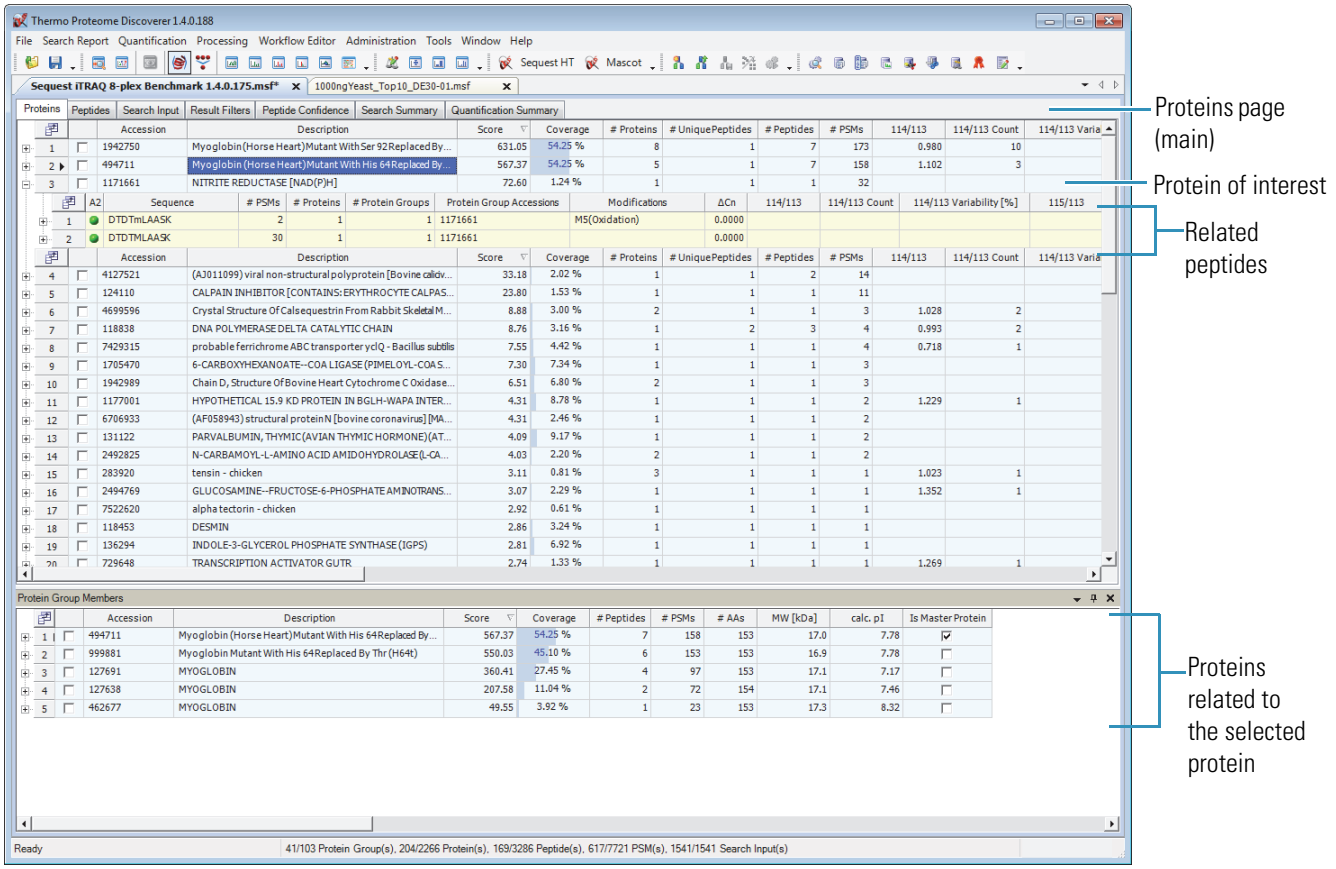
.

❖ To display other proteins belonging to the same protein group

1. Open the MSF file.
2. On the Proteins page, click anywhere in a protein row.
3. Choose **Search Report > Show Protein Group Members**, or click the **Show Protein Group Members View** icon,  .

   The Protein Group Members view appears below the Proteins page, as shown in Figure 132.

Figure 132. Proteins in the same group



The Is Master Protein column in the Protein Group Members view indicates whether the protein is the master protein of a protein group. For some peptides, a list of proteins might contain this peptide sequence, but none of them is a master protein. This situation can occur if the peptide contains isoleucine at a position where the master protein has leucine or vice versa.

❖ To turn off protein grouping

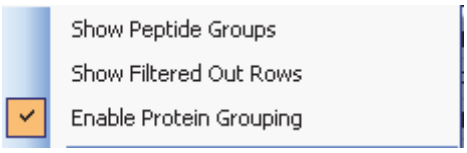
1. On the Result Filters page, click **Settings** below Protein Grouping (Enabled), and clear the **Enable Protein Grouping** check box.

   –or–

   On the Proteins or Peptides page, right-click a protein grid cell or row to access the shortcut menu, and clear the check mark for **Enable Protein Grouping**, shown in Figure 133.

   The proteins are no longer grouped.

Figure 133. Enable Protein Grouping command on the Proteins page shortcut menu



2. To regroup proteins, reselect the **Enable Protein Grouping** check box on the Result Filters page.

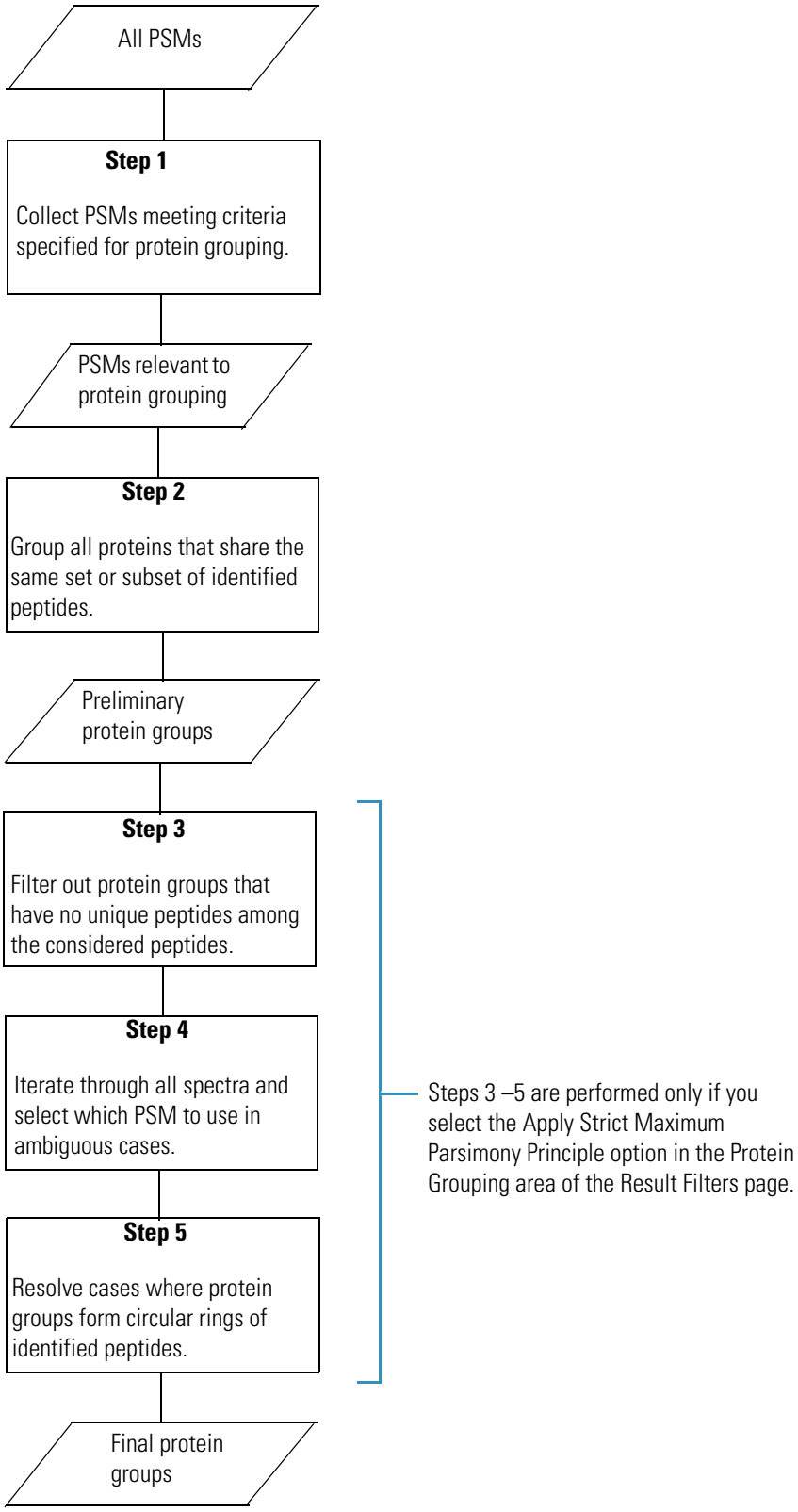
   –or–

   Right-click a protein grid cell or row in the Proteins or Peptides page and choose **Enable Protein Grouping** from the shortcut menu.

Protein Grouping Algorithm

The Proteome Discoverer application uses a protein grouping inference process to group proteins. Figure 134 shows the steps involved in this process.

Figure 134. Protein grouping inference process in the Proteome Discoverer application



1. In the first step, the application collects all peptide spectrum matches (PSMs) that meet the selection criteria that you specified through the settings of the parameters in the Protein Grouping (Enabled) area on the Result Filters page (see Figure 131 on page 176). The Help explains these parameters. You can use these settings to specify which PSMs to consider for the inference of the protein groups. For example, if you set the Consider Only PSMs with Confidence at Least parameter to Medium, the Proteome Discoverer application considers only PSMs with a medium- or high-identification confidence when it creates the protein groups and ignores PSMs with a low-identification confidence. You can further use the Consider Only PSMs with Delta Cn Better Than parameter to filter out PSMs over a normalized score and consider the remaining PSMs for inclusion in the protein group inference process if their confidence levels fit.

   Note Setting the Consider Only PSMs with Confidence at Least parameter to Low and the Consider Only PSMs with Delta Cn Better Than parameter to 1 and leaving the Apply Strict Maximum Parsimony Principle option unselected creates the same protein groups as the previous release of the Proteome Discoverer application.

   This first step prevents protein groups from including low-scoring, low-confidence PSMs. Even if the Proteome Discoverer application loads all PSMs initially identified by the search engines without applying further result filters, it considers only those PSMs meeting the specified criteria when inferring protein groups. If the set result filters filter out PSMs, the application does not consider them for the protein grouping process, even if they would otherwise fit the set grouping criteria.

2. In the second step, the application creates preliminary protein groups from the PSMs collected in the first step. It combines all proteins into one protein group that contains the same subset of peptides.

   The Proteome Discoverer application takes the next steps in the protein grouping process if you select the Apply Strict Maximum Parsimony Principle parameter in the Result Filters page.

3. In the third step, the application removes all protein groups that have no unique peptides among the peptides that it considers for the protein grouping process. If a protein group does not contain at least one unique peptide, all of its peptides are also included by other protein groups, so there is no supporting evidence for the existence of this protein group. At this point, the application explicitly retains all protein groups that form circular rings of overlapping shared peptides. For example, suppose a circular ring is composed of the protein groups:

   - ABCD (identified by peptides a, b, c, and d)
   - CDEF (identified by peptides c, d, e, and f)
   - EFAB (identified by peptides e, f, a, and b)

   To explain all identified peptides, only two of the three protein groups are needed, but at this point it is not clear which to take and which to reject. The application postpones the resolution of this issue until step 5.

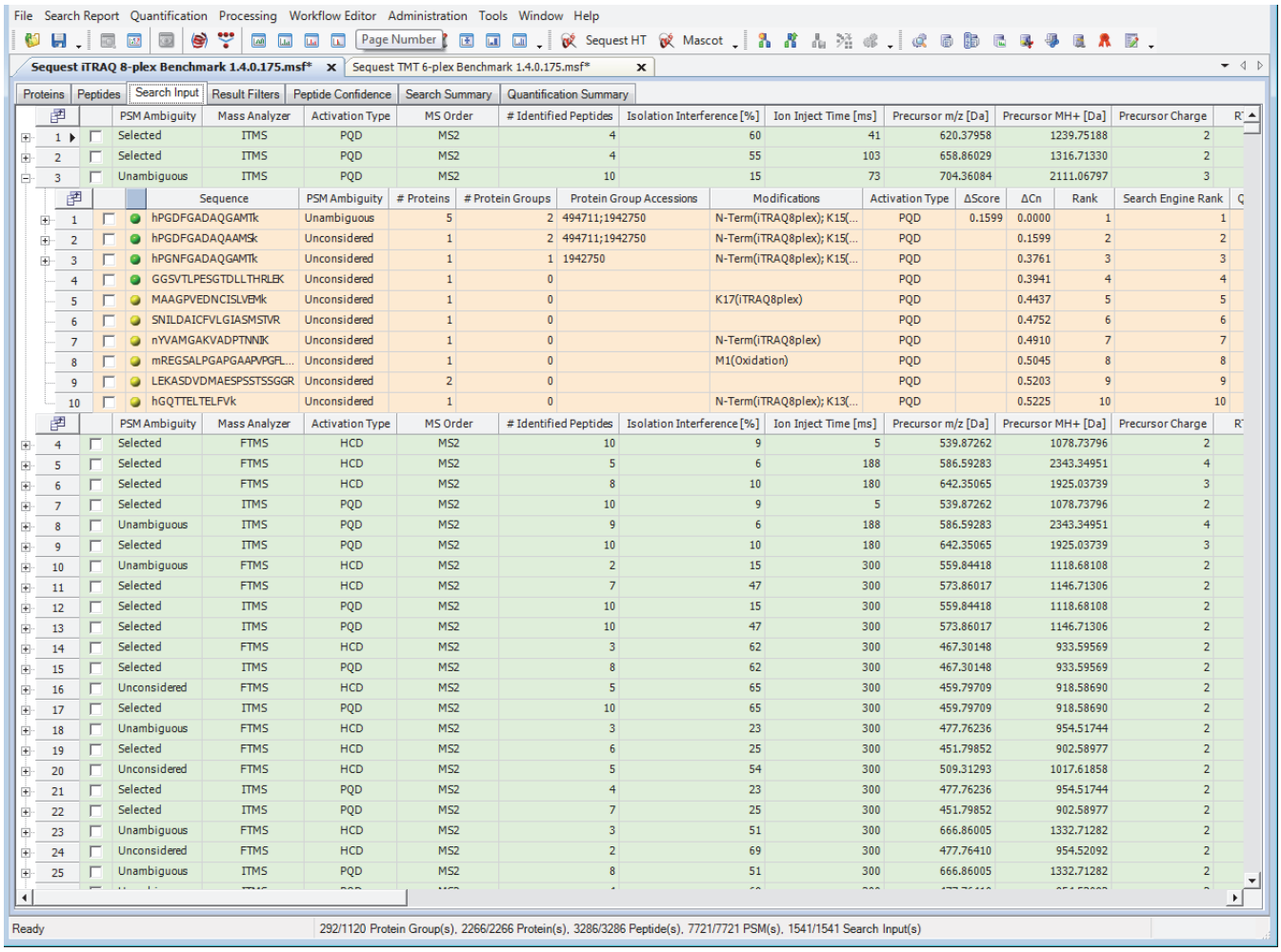
4. In the fourth step, the application first collects all spectra with more than one peptide match to consider for the protein grouping process. It then resolves these ambiguous cases and selects one of the PSMs to use for the protein grouping process while rejecting the remaining peptide matches of a spectrum. In cases where more than one PSM is considered for a spectrum, it resolves this ambiguity by selecting the PSM that is connected to the "best" protein group and rejecting the other PSMs. The "best" protein group is the group with the highest number of unambiguous and unique peptides and the highest protein score.
5. In the fifth step, the application resolves the cases where protein groups form circular rings of overlapping identified peptides. This step is the last step of the protein group inference process, resulting in the final list of protein groups that are reported in the Proteins page of the MSF file.

The PSM Ambiguity column on the Peptides and Search Input pages can help you understand the process of selecting PSMs for the protein group. This column is available for every PSM, every search input entry (representing the searched spectra), and every peptide group. For the search input entries and the peptide groups, this column displays the best PSM ambiguity from all connected PSMs. Refer to the Help for a description of the categories of ambiguity in this column.

Note If you want to investigate the protein grouping mechanism in detail, set the Group Peptides By option in the Peptide Grouping (Enabled) area of the Result Filters page to Sequence and not to Mass and Sequence. This way, the peptide groups created are similar to the protein groups created, which are always based on peptide sequences.

Consider the example shown in Figure 135, where 10 different PSMs are identified for search input 3. The four PSMs ranked 1 through 4 all meet the specified protein grouping criteria. They are of high confidence, and their Δ Cn values are below the threshold of 0.4, so the protein group inference algorithm considers all three PSMs for grouping. It does not consider the remaining PSMs of the spectrum, which are ranked 3 through 10 and are of medium confidence, when creating protein groups.

Figure 135. PSMs shown for search input



Proteins Containing Peptides with Sequences Not Belonging to a Master Protein

Because the Proteome Discoverer application considers for inclusion in the protein grouping process only PSMs that meet the criteria set in the Protein Grouping (Enabled) area of the Results Filters page, a protein group might contain proteins that have identified peptides whose sequences are not all contained in the master protein of the protein group. For example, if you specify that the protein grouping inference process consider only PSMs that have at least medium confidence, a protein group might include a protein with a low-confidence peptide that does not belong to a master protein.

Protein Groups in the Status Bar

The status bar shows the actual number of protein groups versus the total number of protein groups (refer to the Help). The difference is the number of protein groups that the application removed to comply with the selection of the Apply Strict Maximum Parsimony Principle option on the Results Filters page. By enabling the display of filtered-out protein groups, you can investigate the protein groups that were removed during this process.

Proteins Grouped by the Grouping Algorithm in Previous Releases

The Proteome Discoverer application removes some protein groups that the protein grouping mechanism created in previous versions of the application. The previous algorithm might have created these groups from only low-confidence peptides, or the application removed them to comply with the selection of the Apply Strict Maximum Parsimony Principle option on the Results Filters page. Therefore, some peptides might not belong to any protein group. To investigate these cases, right-click the Proteins page and choose **Show Filtered Out Rows** to display the filtered-out peptides in the results file. You can also use the Protein References of a Peptide view, opened by choosing Search Report > Show Protein References, to help you.

Number of Unique Peptides Column on the Proteins Page

The value in the # Unique Peptides column on the Proteins page that is listed for each protein group is the number of peptides that are only contained in this protein group. The Proteome Discoverer application counts only peptides that display a status of Selected or Unambiguous in the PSM Ambiguity column, because assessing the uniqueness of peptides that were not used to form protein groups has no relevance.

PSMs Identified by Multiple Workflow Nodes

In search results where the application identifies PSMs by multiple search nodes within a single workflow, the protein grouping algorithm selects one of the PSMs identified for the same spectrum for building the protein groups.

In search results where PSMs are identified by multiple search nodes from multiple workflows (multiconsensus report), the application treats PSMs and spectra from the different workflows as separate, even if it searched the same raw data files and therefore the same spectra. In this case, determining whether the application searched the exact same spectra is difficult, because they might have changed in the different workflows.

Grouping Peptides

In the results report, you can turn peptide grouping on or off with the Show Peptide Groups command on the shortcut menu or with the settings in the Peptide Grouping (Enabled) area on the Result Filters page. Using the latter method, you can select more options in grouping. Grouping is turned on by default.

In the Peptide Grouping area of the Result Filters page, you can specify whether you want to group peptides only by sequence or by mass and sequence. The Mass and Sequence setting of the Group Peptides By option separates the differently modified forms of a peptide into different peptide groups. This setting is the default.

The number of peptides displayed in the status bar is always the number of distinct sequences. The number of peptide groups, on the other hand, depends on the peptide grouping settings. If you group peptides by sequence only, the two numbers are the same. If you group peptides by sequence and mass, the number of peptide groups is normally larger than the number of peptides displayed in the status bar, unless the peptides have no modifications.

❖ To group the peptides in your search results

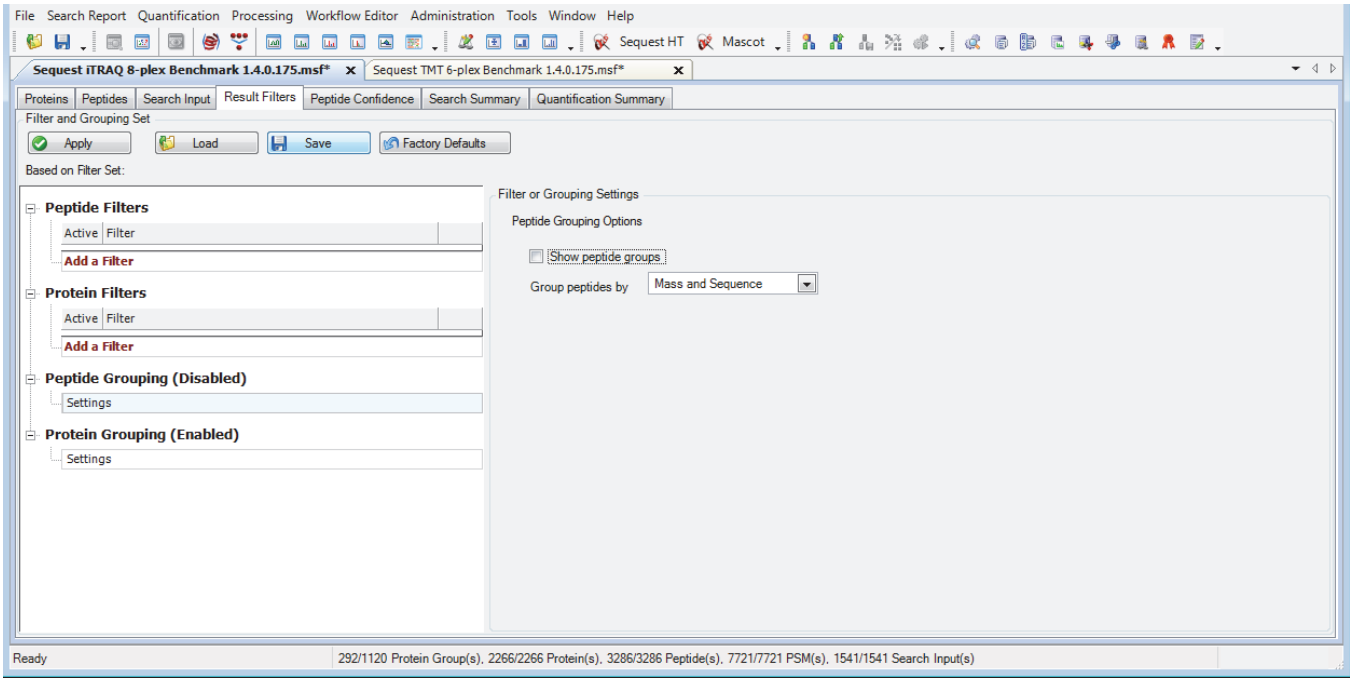
1. Open the MSF file.
2. On the Peptides or Proteins page of the MSF file, right-click a peptide grid cell or row to access the shortcut menu, and choose **Show Peptide Groups**.

❖ To group the peptides in your search results and set grouping options

1. Open the MSF file.
2. Click the **Results Filters** tab.
3. On the Results Filters page, click **Settings** beneath Peptide Grouping.

   Peptide grouping options appear in the Filter or Grouping Settings area, as shown in Figure 136.

Figure 136. Peptide grouping options



4. If you want peptides to be grouped on the Peptides page of the results report, select the **Show Peptide Groups** check box.
5. Select the method of grouping peptides from the Group Peptides By list:
   - **Sequence**: Groups peptides by sequence.
   - **Mass and Sequence**: Groups peptides by mass and sequence.
6. Click Apply.

Calculating False Discovery Rates

The false discovery rate (FDR), or the false positive rate, is a statistical value that estimates the number of false positive identifications among all identifications found by a peptide identification search. It is a measure of the certainty of the identification. You can use the Proteome Discoverer decoy database search feature to determine FDRs.

You can use FDRs to validate MS/MS searches of large data sets, but they are not effective on searches of a small number of spectra or searches against a small number of protein sequences, because the number of matches will likely be too small to give a statistically meaningful estimate.

A decoy database gives a probability value to identifiers and the percentage of false discoveries that you can expect. A one percent FDR is a typical target for searches.

A good decoy database should contain entries that look like real proteins but do not contain genuine peptide sequences. The simplest approach to achieving such a decoy database is to reverse all protein sequences, which is the scheme that the Proteome Discoverer application currently uses. It is a suitable approach for enzymatic MS/MS searches.

IMPORTANT Reversing the database is not suitable for peptide mass fingerprinting or no-enzyme MS/MS searches, especially for dynamic modifications. You might see mass shifts at each end of a peptide sequence that transform a genuine y series match into a false b series match or vice versa.

You can perform the decoy database search in two ways:

- Perform two separate searches, one against the non-decoy database and one against the decoy database. Then count the number of matches from both searches to determine the FDRs. This approach is the more conservative approach.
- Create a concatenated database from the non-decoy and the decoy database and then perform the search against this concatenated database.

The difference between the two approaches becomes clear in the case where you find two significant matches for a given spectrum. The first match is from the non-decoy database, and the second one is from the decoy database. Because the Proteome Discoverer application considers only the top matches when calculating the FDRs, finding two significant matches for a given spectrum is not considered a false positive in the concatenated database approach, but it counts in the separate databases approach. The latter case is considered the more conservative one and is the approach that the application currently uses.

To calculate the FDR, the application counts the matches that pass a given set of filter thresholds from the decoy database and from the non-decoy database. It counts only the top match per spectrum, assuming that for any given spectrum only one peptide can be the correct match.

Target FDRs

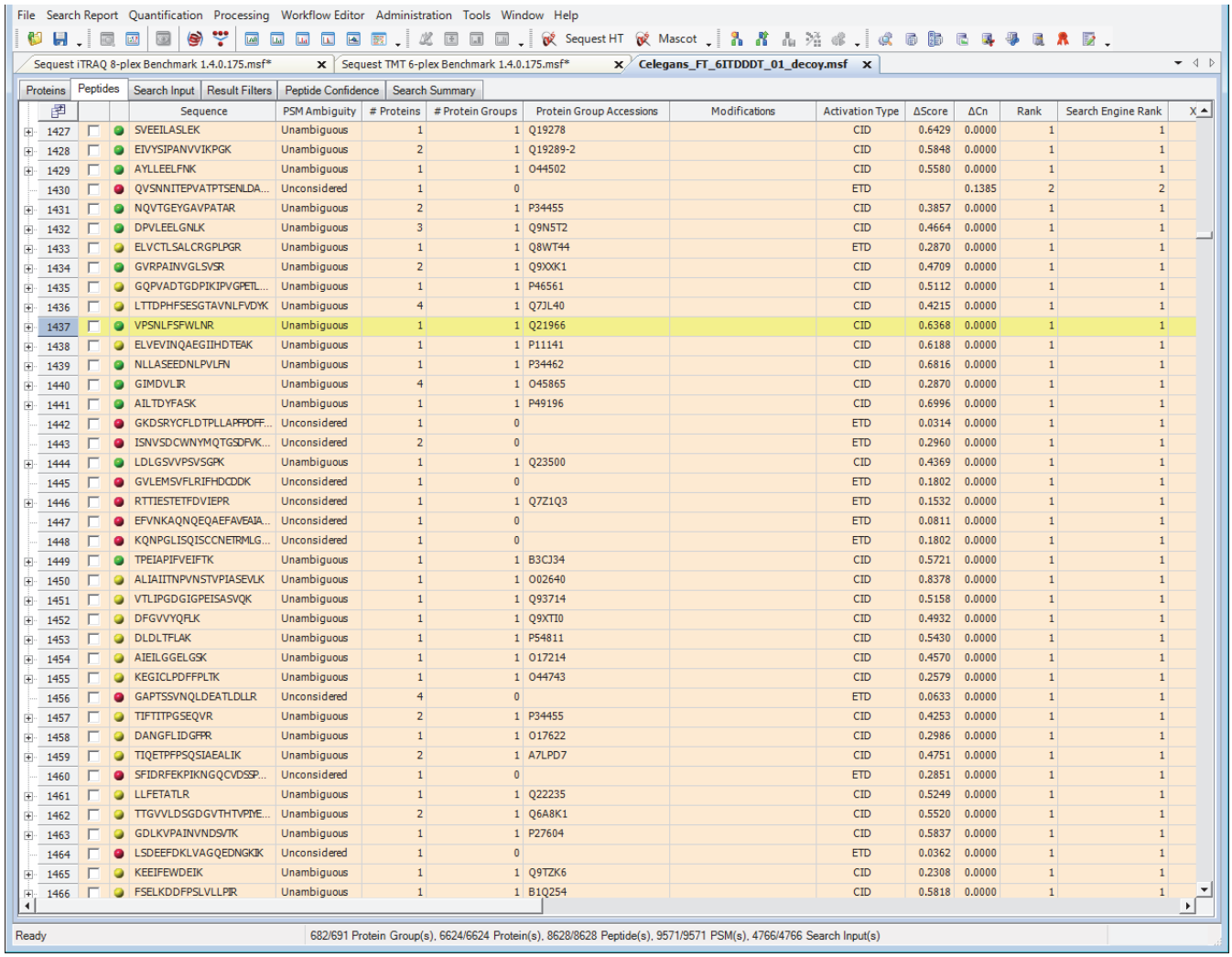
If you set an FDR target value for a decoy database search, the application determines and applies filter thresholds to identified matches so that the resulting FDR is not higher than the set target value. The confidence indicators applied to each peptide match are distributed according to these calculated filter thresholds (see Figure 142 on page 195).

You must specify two target values for a decoy database search: a strict target FDR and a more relaxed FDR. Figure 139 on page 191 shows the decoy search setting with target FDRs of one percent and five percent, respectively. After completing the search, the system automatically determines two sets of filter settings so that the resulting separate FDRs do not exceed their corresponding target value.

Peptide Confidence Indicators

The filter settings that determine FDRs are used to distribute the confidence indicators for the peptide matches (these are the green, yellow, and red circles attached to each peptide match). Whenever you perform a decoy database search during the database search and apply filter settings to achieve the specified target FDRs, the same filters are used to distribute the confidence indicators. Peptide matches that pass the filter associated with the strict FDR are assigned a green confidence indicator, peptide matches that pass the filter associated with the relaxed FDR are assigned a yellow confidence indicator, and all other peptide matches receive a red indicator of low confidence. Figure 137 gives an example of these confidence indicators.

Figure 137. Decoy search results



	Sequence	PSM Ambiguity	# Proteins	# Protein Groups	Protein Group Accessions	Modifications	Activation Type	ΔScore	ΔCn	Rank	Search Engine Rank
1427	SVEEILASLEK	Unambiguous	1	1	Q19278		CID	0.6429	0.0000	1	1
1428	EIVYSIPANVVIKPGK	Unambiguous	2	1	Q19289-2		CID	0.5848	0.0000	1	1
1429	AYLLEELFNK	Unambiguous	1	1	O44502		CID	0.5580	0.0000	1	1
1430	QVSNNITEPVATPTSENLDA...	Unconsidered	1	0			ETD		0.1385	2	2
1431	NQVTGEYGAVPATAR	Unambiguous	2	1	P34455		CID	0.3857	0.0000	1	1
1432	DPVLEELGNLK	Unambiguous	3	1	Q9N5T2		CID	0.4664	0.0000	1	1
1433	ELVCTLSALCRGPLPGR	Unambiguous	1	1	Q8WT44		ETD	0.2870	0.0000	1	1
1434	GVRPAINVGLSVSR	Unambiguous	2	1	Q9XXK1		CID	0.4709	0.0000	1	1
1435	GQPVADTGDPIKIPVGPETL...	Unambiguous	1	1	P46561		CID	0.5112	0.0000	1	1
1436	LTTDPHFSESGTAVNLFVDYK	Unambiguous	4	1	Q7JL40		CID	0.4215	0.0000	1	1
1437	VPSNLFSFWLNR	Unambiguous	1	1	Q21966		CID	0.6368	0.0000	1	1
1438	ELVEVINQAEGIIHDTEAK	Unambiguous	1	1	P11141		CID	0.6188	0.0000	1	1
1439	NLLASEEDNLPVLFN	Unambiguous	1	1	P34462		CID	0.6816	0.0000	1	1
1440	GIMDVLIR	Unambiguous	4	1	O45865		CID	0.2870	0.0000	1	1
1441	AILTDYFASK	Unambiguous	1	1	P49196		CID	0.6996	0.0000	1	1
1442	GKDSRYCFLDTPLLAPFPDFF...	Unconsidered	1	0			ETD	0.0314	0.0000	1	1
1443	ISNVSDCWNYMQTGSDFVK...	Unconsidered	2	0			ETD	0.2960	0.0000	1	1
1444	LDLGSVVPSVSGPK	Unambiguous	1	1	Q23500		CID	0.4369	0.0000	1	1
1445	GVLEMSVFLRIFHDCDDK	Unconsidered	1	0			ETD	0.1802	0.0000	1	1
1446	RTTIESTETFDVIEPR	Unconsidered	1	1	Q7Z1Q3		ETD	0.1532	0.0000	1	1
1447	EFVNKAQNQEQAEFAVEAIA...	Unconsidered	1	0			ETD	0.0811	0.0000	1	1
1448	KQNPGLISQISCCNETRMLG...	Unconsidered	1	0			ETD	0.1802	0.0000	1	1
1449	TPEIAPIFVEIFTK	Unambiguous	1	1	B3CJ34		CID	0.5721	0.0000	1	1
1450	ALIAIITNPVNSTVPIASEVLK	Unambiguous	1	1	O02640		CID	0.8378	0.0000	1	1
1451	VTLIPGDGIGPEISASVQK	Unambiguous	1	1	Q93714		CID	0.5158	0.0000	1	1
1452	DFGVVYQFLK	Unambiguous	1	1	Q9XTI0		CID	0.4932	0.0000	1	1
1453	DLDLTFLAK	Unambiguous	1	1	P54811		CID	0.5430	0.0000	1	1
1454	AIEILGGELGSK	Unambiguous	1	1	O17214		CID	0.4570	0.0000	1	1
1455	KEGICLPDFFPLTK	Unambiguous	1	1	O44743		CID	0.2579	0.0000	1	1
1456	GAPTSSVNQLDEATLDLLR	Unconsidered	4	0			ETD	0.0633	0.0000	1	1
1457	TIFTITPGSEQVR	Unambiguous	2	1	P34455		CID	0.4253	0.0000	1	1
1458	DANGFLIDGFPR	Unambiguous	1	1	O17622		CID	0.2986	0.0000	1	1
1459	TIQETPFPSQSIAEALIK	Unambiguous	2	1	A7LPD7		CID	0.4751	0.0000	1	1
1460	SFIDRFEKPIKNGQCVDSSP...	Unconsidered	1	0			ETD	0.2851	0.0000	1	1
1461	LLFETATLR	Unambiguous	1	1	Q22235		CID	0.5249	0.0000	1	1
1462	TTGVVLDSGDGVTHTVPIYE...	Unambiguous	2	1	Q6A8K1		CID	0.5520	0.0000	1	1
1463	GDLKVPAINVNDSVTK	Unambiguous	1	1	P27604		CID	0.5837	0.0000	1	1
1464	LSDEEFDKLVAGQEDNGKIK	Unconsidered	1	0			ETD	0.0362	0.0000	1	1
1465	KEEIFEWDEIK	Unambiguous	1	1	Q9TZK6		CID	0.2308	0.0000	1	1
1466	FSELKDDFPSLVLLPIR	Unambiguous	1	1	B1Q254		CID	0.5818	0.0000	1	1

Note You can change the default confidence levels to alternative values on the Peptide Confidence page.

Setting Up FDRs in Search Wizards and the Workflow Editor

You can set up FDRs in both the search wizards and the Workflow Editor.

- Setting Up FDRs in the Search Wizards
- Setting Up FDRs in the Workflow Editor

Setting Up FDRs in the Search Wizards

You can set the strict and relaxed FDRs for every available search wizard.

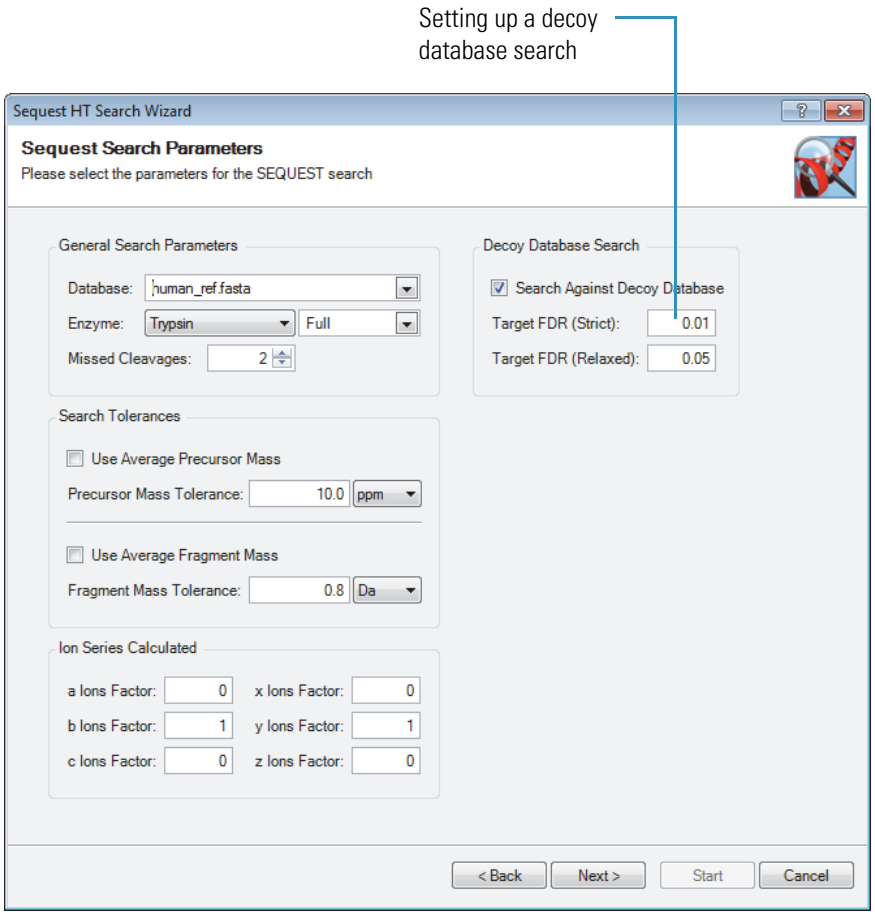
❖ To set up FDRs in a search wizard

1. Start your search by using the search wizards.

   For information about using the search wizards, see "Starting a New Search by Using the Search Wizards" on page 29.

2. On the <*Wizard_name*> Search Parameters page, select the **Search Against Decoy Database** option, as shown in Figure 138.

Figure 138. Setting up a decoy database search in a search wizard



3. In the Target FDR (Strict) box, set the target FDR for high-confidence peptide hits.
4. In the Target FDR (Relaxed) box, set the target FDR for medium-confidence peptide hits.
5. Click **Next**.

Setting Up FDRs in the Workflow Editor

You can set up FDRs through the Target Decoy PSM Validator node or the Percolator node in the workflow. For information about the Target Decoy PSM Validator node, refer to the Help. For detailed information about the Percolator node and its processing, refer to the Help.

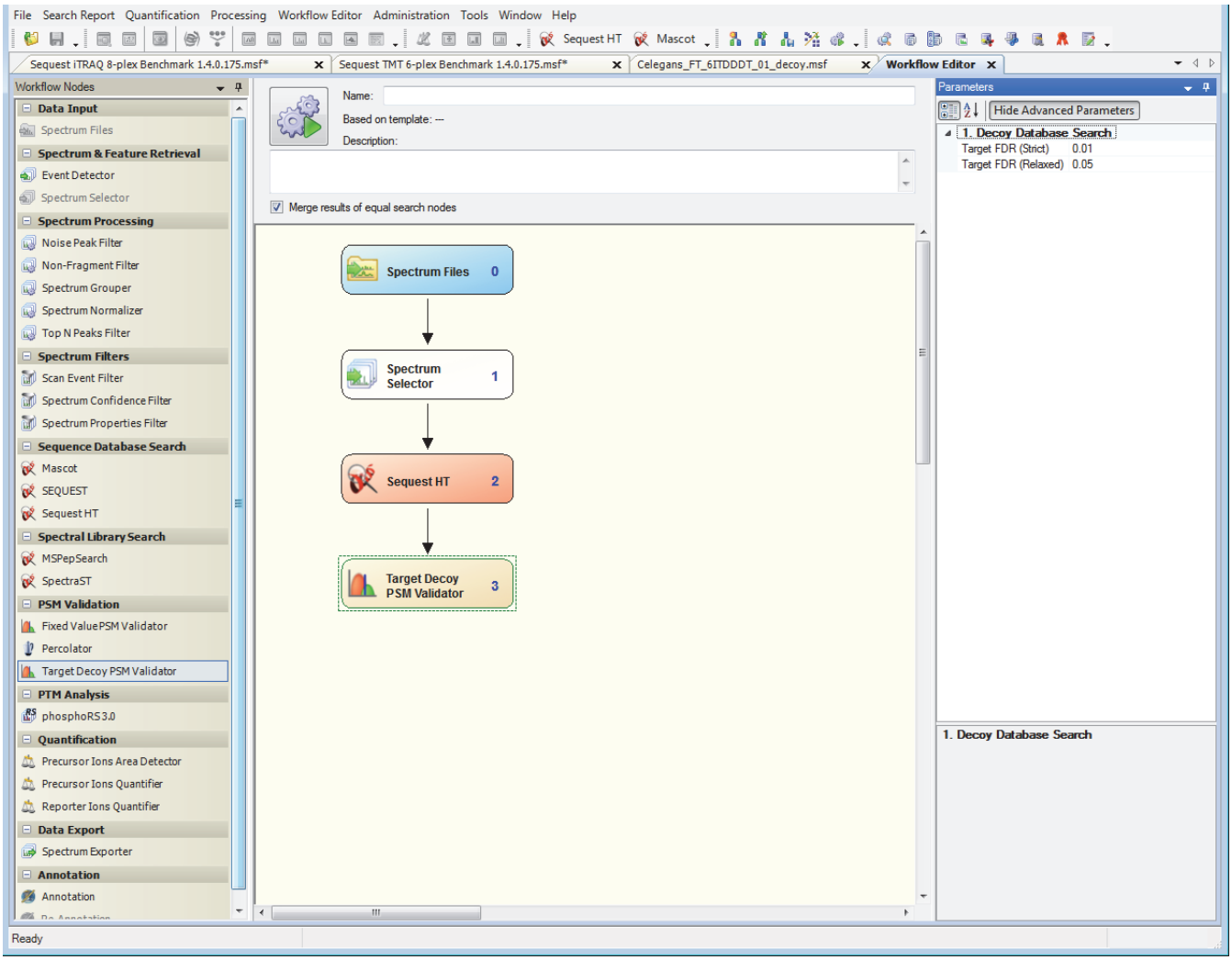
❖ To set up FDRs by using the Target Decoy PSM Validator node

1. Create a search workflow that includes at least one of the search engine nodes (SEQUEST, Mascot, or Sequest HT) and the Target Decoy PSM Validator node.

For information about creating a workflow, see "Creating a Search Workflow" on page 44.

2. Click the **Target Decoy PSM Validator** node, as shown in Figure 139.

Figure 139. Setting up a decoy database search in the Workflow Editor



3. In the Target FDR (Strict) box, set the target FDR for high-confidence peptide hits.
4. In the Target FDR (Relaxed) box, set the target FDR for Peptides medium-confidence peptide hits.
5. Choose **Workflow Editor > Start Workflow**, or click the **Start Workflow** icon, .

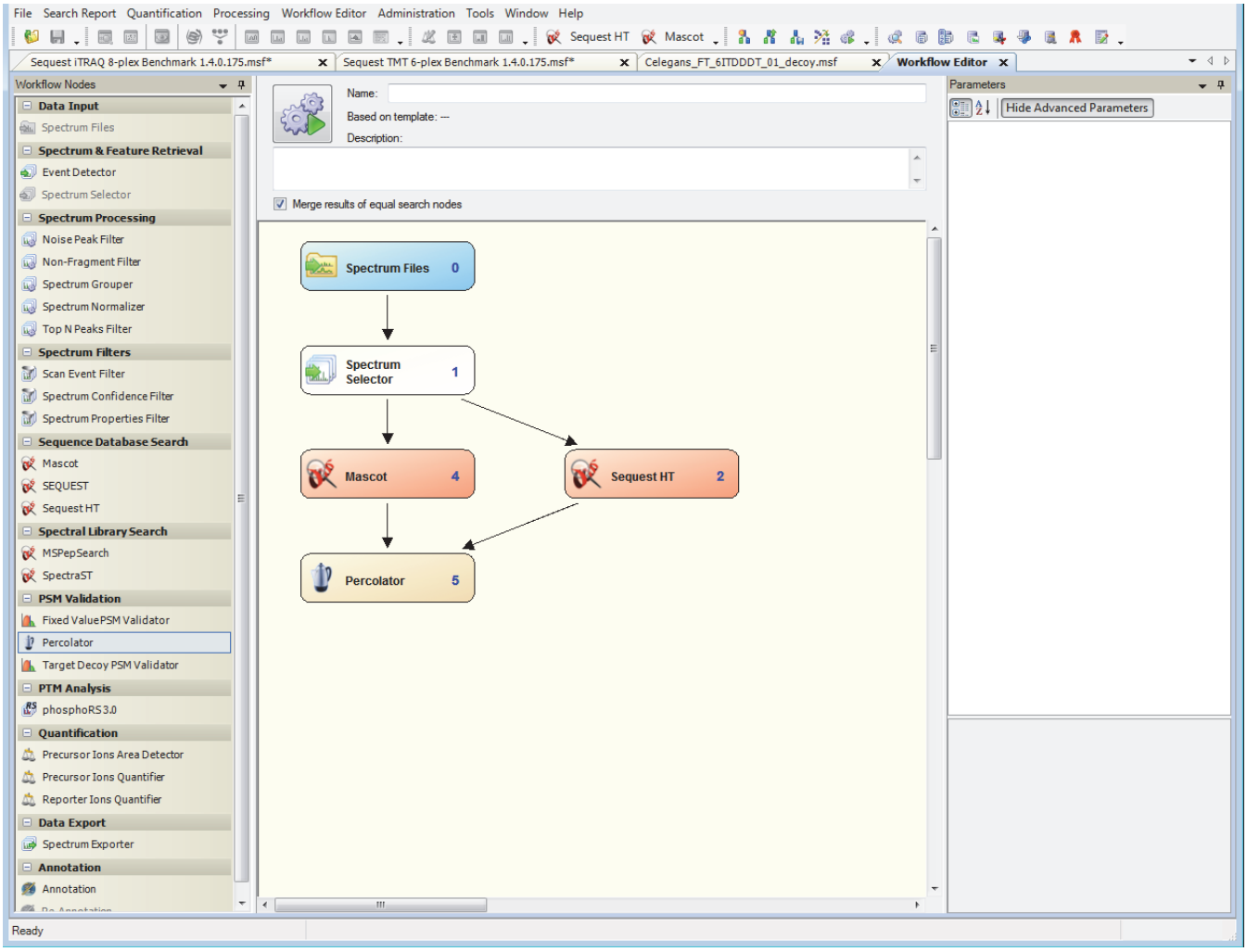
❖ To set up FDRs by using the Percolator node

1. Create a search workflow that includes at least one of the search engine nodes (SEQUEST, Mascot, or Sequest HT) and the Percolator node.
2. For information about creating a workflow, see "Creating a Search Workflow" on page 44.
3. Connect all search nodes whose results you want to submit for validation to the Percolator node.

   Figure 140 gives an example of such a workflow.

> **Note** To work properly, Percolator needs a sufficient number of PSMs from the target and the decoy search. If the search identified fewer than 200 target or decoy PSMs, or if fewer than 20 percent decoy PSMs are available compared to the number of target matches, Percolator rejects them for processing and displays an appropriate message in the Proteome Discoverer job queue or in the Search Summary of an open report.

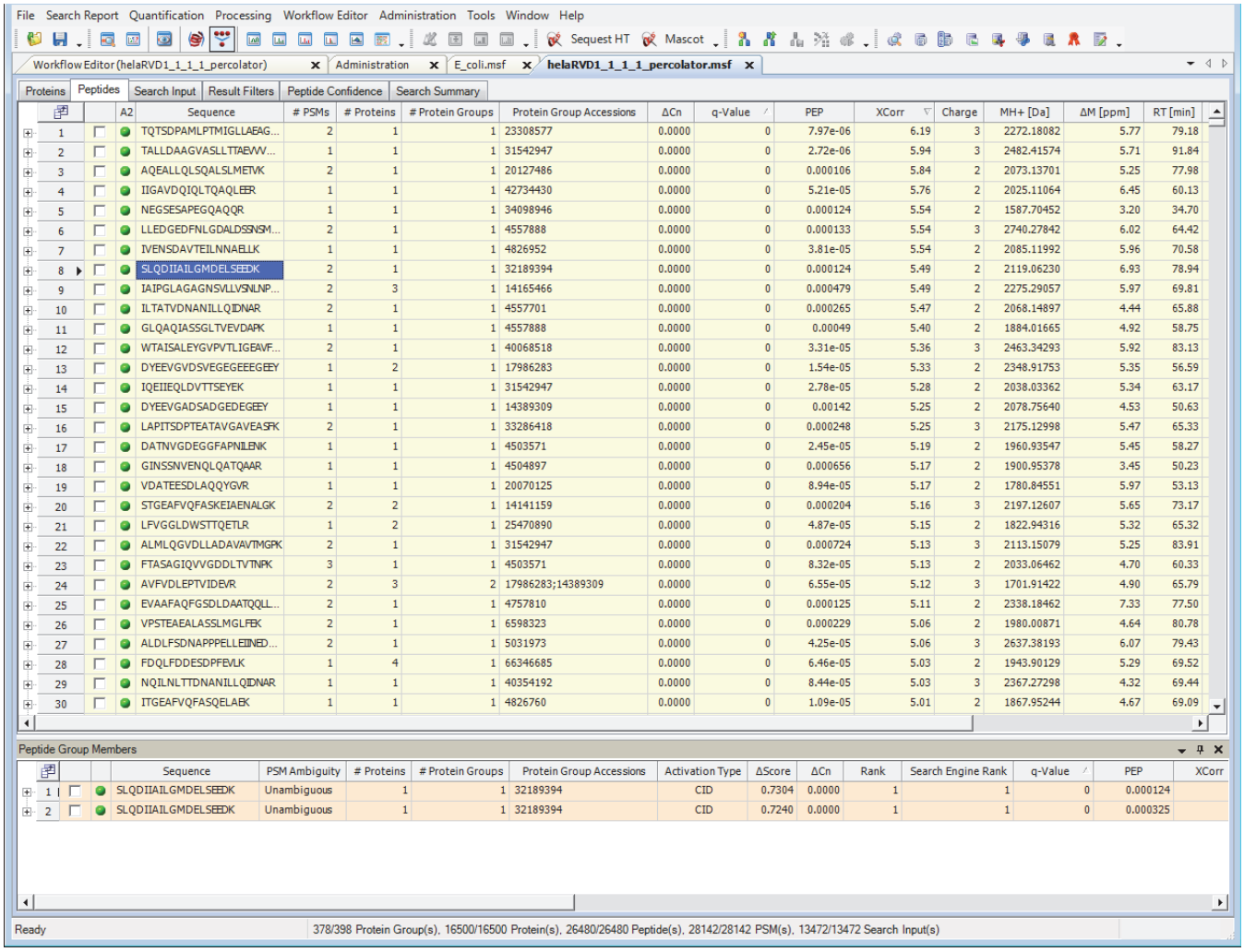
Figure 140. Workflow with Percolator attached to two different search nodes



4. Click the **Percolator** node.
5. In the Maximum Delta Cn box in the parameters list, specify the Δ Cn value. For information on this parameter, see "Filtering Peptides by the Delta Cn Value" on page 161.
6. In the Target FDR (Strict) box, set the target FDR for high-confidence peptide hits.
7. In the Target FDR (Relaxed) box, set the target FDR for medium-confidence peptide hits.
8. In the Validation Based On box, select either **q-Value** or **PEP** (posterior error probability) to assign to the target and decoy PSMs. For more information on these options, refer to the Help.
9. Choose **Workflow Editor > Start Workflow**, or click the **Start Workflow** icon,  .

When you open results processed with the Percolator node, each PSM and peptide group has two additional scores on the Peptides page, a q-value score and a posterior error probability (PEP) value, as shown in Figure 141.

Figure 141. PEP and q-Value columns on the Peptides page of results processed with Percolator



Viewing the Results on the Peptide Confidence Page

After the Proteome Discoverer application completes the search, open the results (MSF) file and view the decoy database search results on the Peptide Confidence page. This page shows the relaxed and strict FDRs with their corresponding filter settings listed above them.

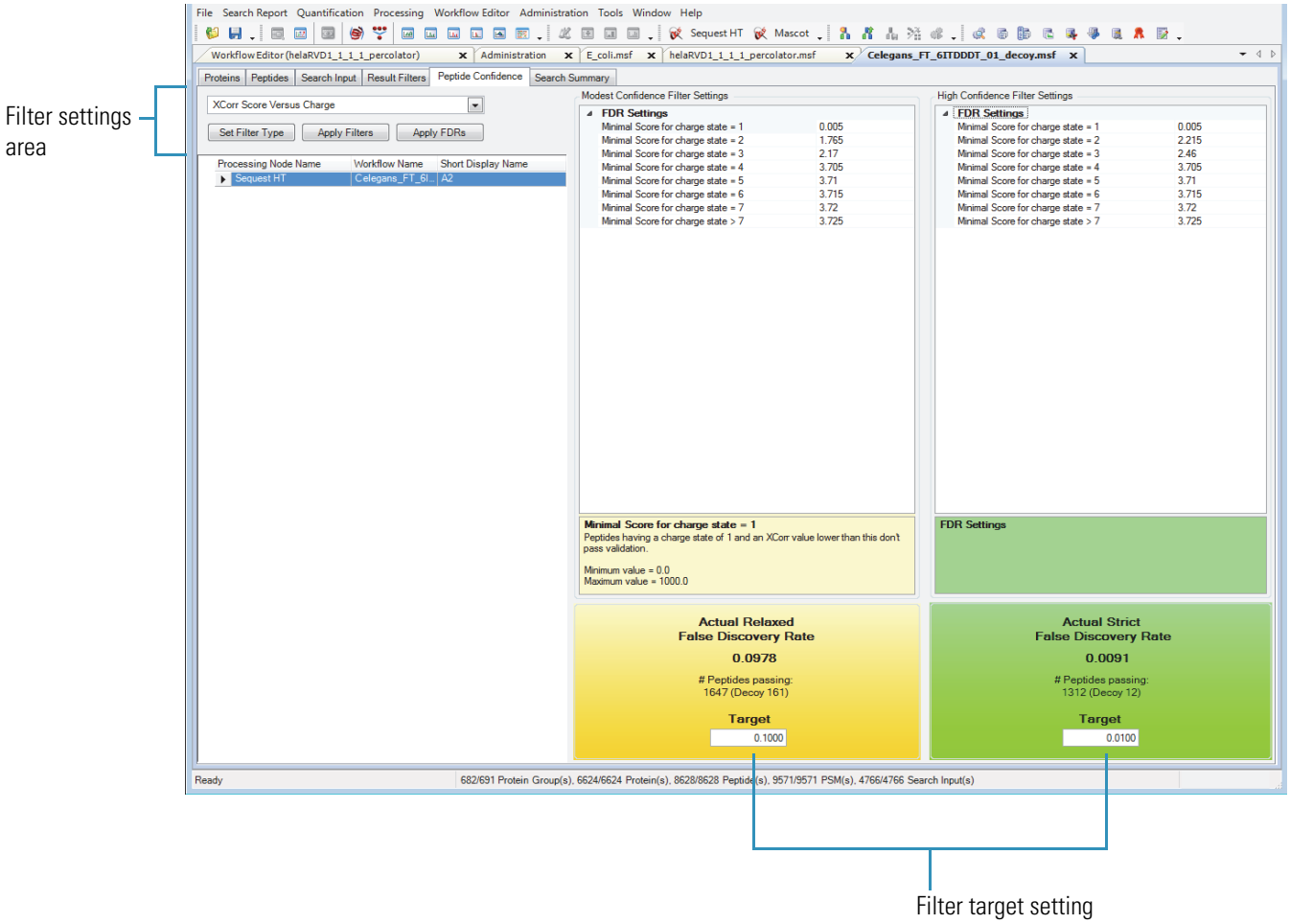
❖ To display the Peptide Confidence page

- In an open report, click the **Peptide Confidence** tab.

  The Peptide Confidence page of your search report appears, as shown in Figure 142. It filters out peptides to two predefined FDRs and sets the confidence levels for database searches.

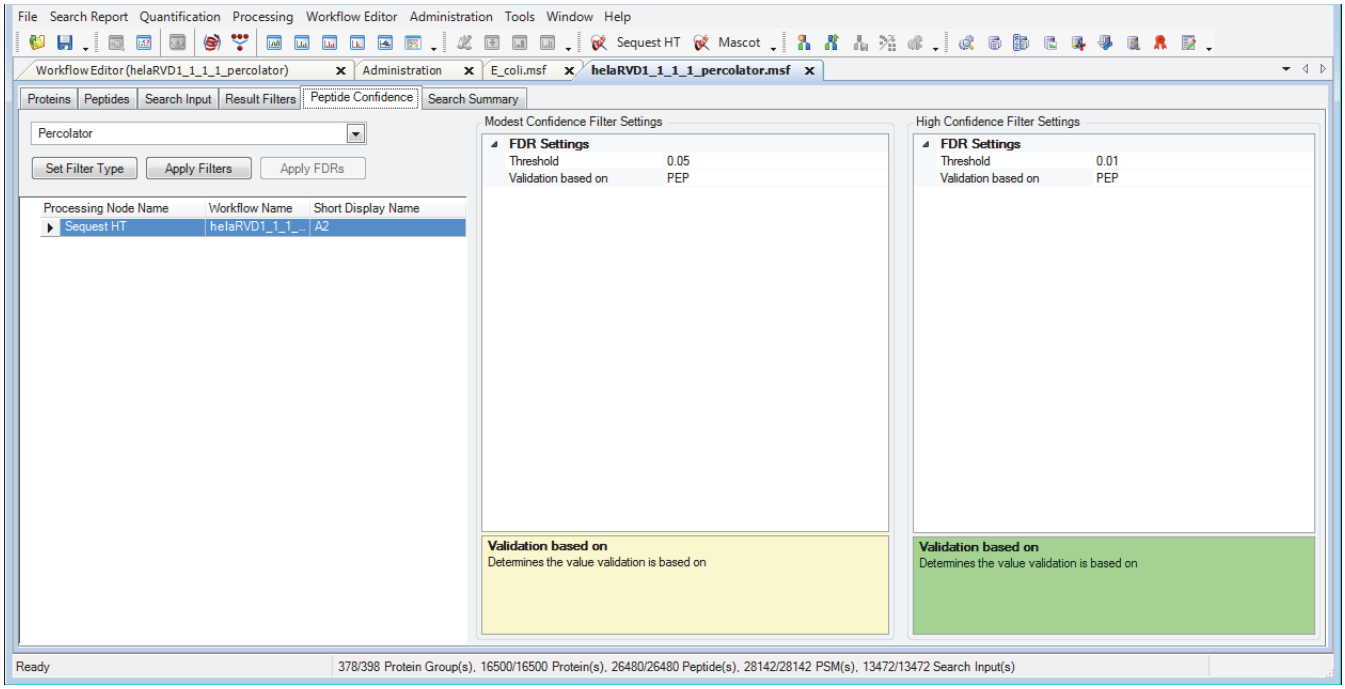
  Use the splitter bar to separate the two columns in the FDR Settings panes.

Figure 142. Peptide Confidence page with the actual relaxed and strict FDRs



If you used the Percolator node in the workflow, you can set thresholds for the Percolator scores to separate PSMs of high confidence, medium confidence, and low confidence, as shown in Figure 143.

Figure 143. Setting thresholds for Percolator scores



In the box in the upper left of the Peptide confidence page, you can switch between validation based on Percolator and validation based on the calculation of target- and decoy-estimated FDRs from the search engine scores. This choice is always available, even if Percolator refused to process the data because it did not meet one of the requirements for the number of target and decoy matches.

Use the Peptide Confidence page to do the following:

- Set new filters and recalculate new FDRs based on these new filter criteria.
- Set new target FDRs and then recalculate new filter settings that, when applied, lead to FDRs no higher than the new target.

Note If you filter on peptide confidence during the loading of the report, all of the options on the Peptide Confidence page are unavailable because you can no longer adjust the settings.

Recalculating the FDRs

You can recalculate the false discovery rate on the Peptide Confidence page.

❖ To recalculate the FDRs

1. Open an MSF file, and click the **Peptide Confidence** tab.
2. In the filter list, select the filter for determining the peptide confidence. The available options are different for each search engine:
   - Sequest:
     - (Default) XCorr Score Versus Charge: Uses this filter to calculate the FDR for determining peptide confidence.
     - Peptide Score: Uses this filter to calculate the FDR for determining peptide confidence.
   - Mascot:
     - (Default) Mascot Significance Threshold: Uses this filter to calculate the FDR for determining peptide confidence.
     - Peptide Score: Uses this filter to calculate the FDR for determining peptide confidence.
3. Click **Set Filter Type** to apply the option that you selected in the Filter list to the settings in the Modest Confidence Filter Settings and the High Confidence Filter Settings panes.

Changing the Target Rate and Filter Settings

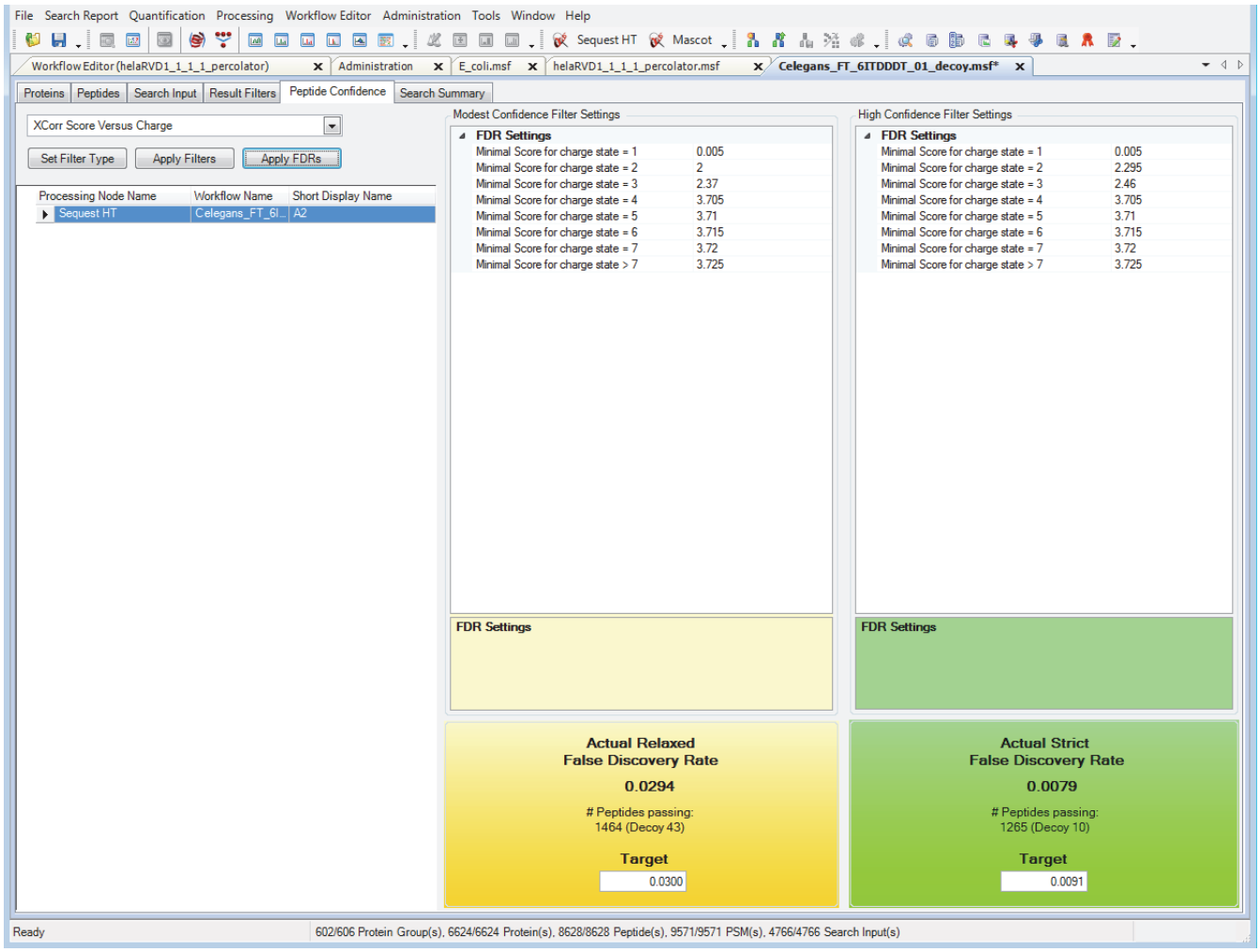
You can change the filter settings on the Peptide Confidence page by changing the target rate or changing the filter settings.

If you change the target rate or the filter settings, the application finds the actual relaxed FDR, the strict FDR, or both that come the closest to your target rate. It displays this number under Actual Relaxed False Discovery Rate or Actual Strict False Discovery Rate. It also displays the number of peptides and decoy peptides that pass the filters set in the Filter Settings area and changes the filter settings in the Filter Settings area.

Whether you change the target rate or the filter settings, the Proteome Discoverer application updates the peptide confidence indicators in the MSF report.

As an example, Figure 144 shows the results of entering a new target rate of 0.030 in the Target box of the Actual Relaxed False Discovery Rate area of the Peptide Confidence page shown in Figure 142 on page 195.

Figure 144. Results of new relaxed target rate



Go to the following sections:

- To change the target rate
- To change the filter settings
- To save the peptide confidence and FDR settings on the Result Filters page

❖ To change the target rate

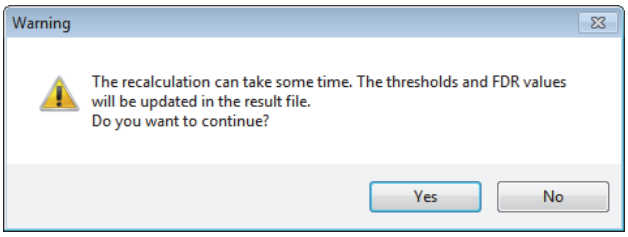
1. Change the value in the Target box of the Actual Relaxed False Discovery Rate area for medium confidence, the Actual Strict False Discovery Rate area for high confidence, or both.
2. Click **Apply FDRs**.

❖ To change the filter settings

1. Select the filter settings that you want to change in the Filter Settings area in the upper left corner of the Peptide Confidence page, and enter the new values in the FDR Settings area. The Minimal Score for Charge State values in the FDR Settings area specify the charge state above which peptides are filtered out. Charge state values can range from 0 to 20.
2. Click **Apply Filters**.

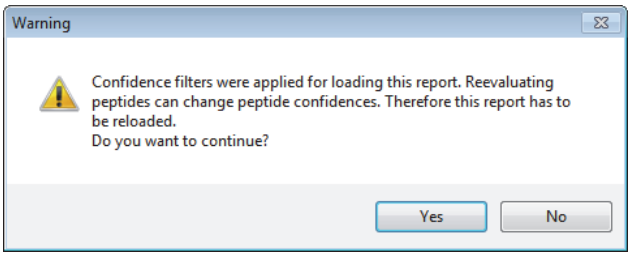
   If you set any filters except the Peptide Confidence filter on the Result Filters page when you loaded the report, the warning shown in Figure 145 appears.

   Figure 145. FDR recalculation message box for all filters except Peptide Confidence

   If you set the Peptide Confidence filter on the Result Filters page when you loaded the report, the warning shown in Figure 146 appears.

   Figure 146. FDR recalculation message box for Peptide Confidence filters

3. In either box, click **Yes**.

❖ To save the peptide confidence and FDR settings on the Result Filters page

- Choose **File > Save Report**.

Protein Annotation

This chapter explains how the Proteome Discoverer application retrieves annotation information from ProteinCenter, including GO (Gene Ontology) annotations, Pfam (Protein Families) annotations, Entrez gene annotations, and information about post-translational modifications (PTMs) from UniProt.

Contents

ProteinCenter

ProteinCenter is a Web-based application that you can use to download biologically enriched annotation information for a single protein, such as molecular functions, cellular components, and biological processes from the GO database; annotation information for protein families from the Pfam database; gene identifications from the Entrez database; and post-translational modification information from the UniProt database. The data in ProteinCenter is updated biweekly.

The Proteome Discoverer application gives you access to ProteinCenter in two ways:

- The Annotation node used in a search workflow retrieves GO, Pfam, Entrez, and UniProt database information from ProteinCenter and stores it in the Proteome Discoverer results files. This information is displayed in columns on the Proteins page of the MSF file. For information on setting up an Annotation workflow to achieve these results, see "Configuring the Proteome Discoverer Application for Protein Annotation" on page 204 and "Creating a Protein Annotation Workflow" on page 206.
- The ProteinCard available for each protein displays the annotation data available in ProteinCenter and displays it on a page of the Protein Identification Details dialog box (see "Accessing ProteinCards" on page 221). You can display this information for the following proteins:
  - Proteins on the Proteins page of the MSF file
  - Proteins associated with identified peptides
  - Proteins shown in the Protein Group Members view

  You can access the ProteinCard for each protein by double-clicking its row in the MSF report or clicking its row and choosing Search Report > Show Protein ID Details and then clicking the ProteinCard tab of the Protein Identification Details dialog box. The ProteinCard itself is split into separate tabs representing different aspects of that protein: General, Keys, Features, Molecular Functions, Cellular Components, Biological Processes, Diseases, and External Links. You can display a ProteinCard for every identified protein whose accession is tracked in ProteinCenter. For information on ProteinCard, see "Accessing ProteinCards" on page 221 and "ProteinCard Parameters" on page 222.

  You can also upload protein results directly from the Proteome Discoverer application to ProteinCenter. For information, see "Uploading Results to ProteinCenter" on page 218.

Gene Ontology (GO) Annotation

The Gene Ontology (GO) database is a collaborative effort, incorporating community input from database and genome annotation groups to address the need for consistent descriptions of gene products in different databases. The GO project has developed three structured, controlled vocabularies (ontologies) that describe gene products in a species-independent manner.

biological processes

cellular components

molecular functions

Each gene ontology is divided into categories and subcategories called GO terms, which define the protein in more specific terms. For example, chloroplast, a term in the cellular component ontology, is subdivided as follows.

chloroplast

[p] *chloroplast* envelope

[p] *chloroplast* membrane

[i] *chloroplast* inner membrane

[i] *chloroplast* outer membrane

You can obtain more information on the GO Ontology Web site at www.geneontology.org/.

Pfam Annotation

In addition to GO annotations, you can also retrieve from ProteinCenter Pfam annotations from the Pfam database at the Wellcome Trust Sanger Institute (//pfam.sanger.ac.uk). These are annotations of protein families, which are proteins with similar sequences and similar biological functions. A special sequence comparison algorithm called the Hidden Markov Model groups proteins into the families by comparing the sequences. Each family has its own ID number that starts with Pf … . The Proteins page of the MSF file displays this number in the Pfam IDs column. You can use the Pfam identification number to go to the Pfam database to obtain more details about the protein family. You can also activate the ProteinCard for each protein by double-clicking the Pfam identification number.

The Pfam annotation system is an alternative to GO annotations. You might want to use the Pfam system to filter your proteins when you want the results to be traceable, scored, and uniformly grouped. You might also consider its computationally based data more reliable. However, it might be easier to use the hierarchy and grouping of the GO system to help you interpret results.

Table 9 compares the features of the GO and Pfam databases.

Table 9. Comparison of GO and Pfam features

GO features	Pfam features
Proteins grouped in biologically meaningful categories	Proteins grouped by similarity
Deep hierarchical order of terms	Few hierarchies
Data input by experts with different confidence levels and differing opinions	Computational data input with no human influence or expert knowledge

Entrez Gene Database Annotation

The Proteome Discoverer application can retrieve the Entrez gene identifications from ProteinCenter. The Entrez gene identification is a unique identification assigned to the genes in the Entrez database maintained by the National Center for Biotechnology Information (NCBI). The database assigns an identifier to all proteins transcribed from the corresponding gene. The Proteins page of the results report displays these identifications in the Gene IDs column. You can use this information to group or cluster together the proteins that are biologically meaningful.

Because not all genes are stored in the Entrez gene database, some proteins do not have a valid gene identification. In this case, the value displayed in the Gene IDs column on the Proteins page of the results file is 0.

UniProt Database Annotation

From ProteinCenter, you can retrieve information on known PTMs from the UniProt database and compare it with information on found PTMs. For details on this feature, refer to the Help.

Configuring the Proteome Discoverer Application for Protein Annotation

Before you can start a search that includes protein annotation in the results or display ProteinCards for proteins, you must configure the Proteome Discoverer application for protein annotation.

❖ **To configure the Proteome Discoverer application for protein annotation**

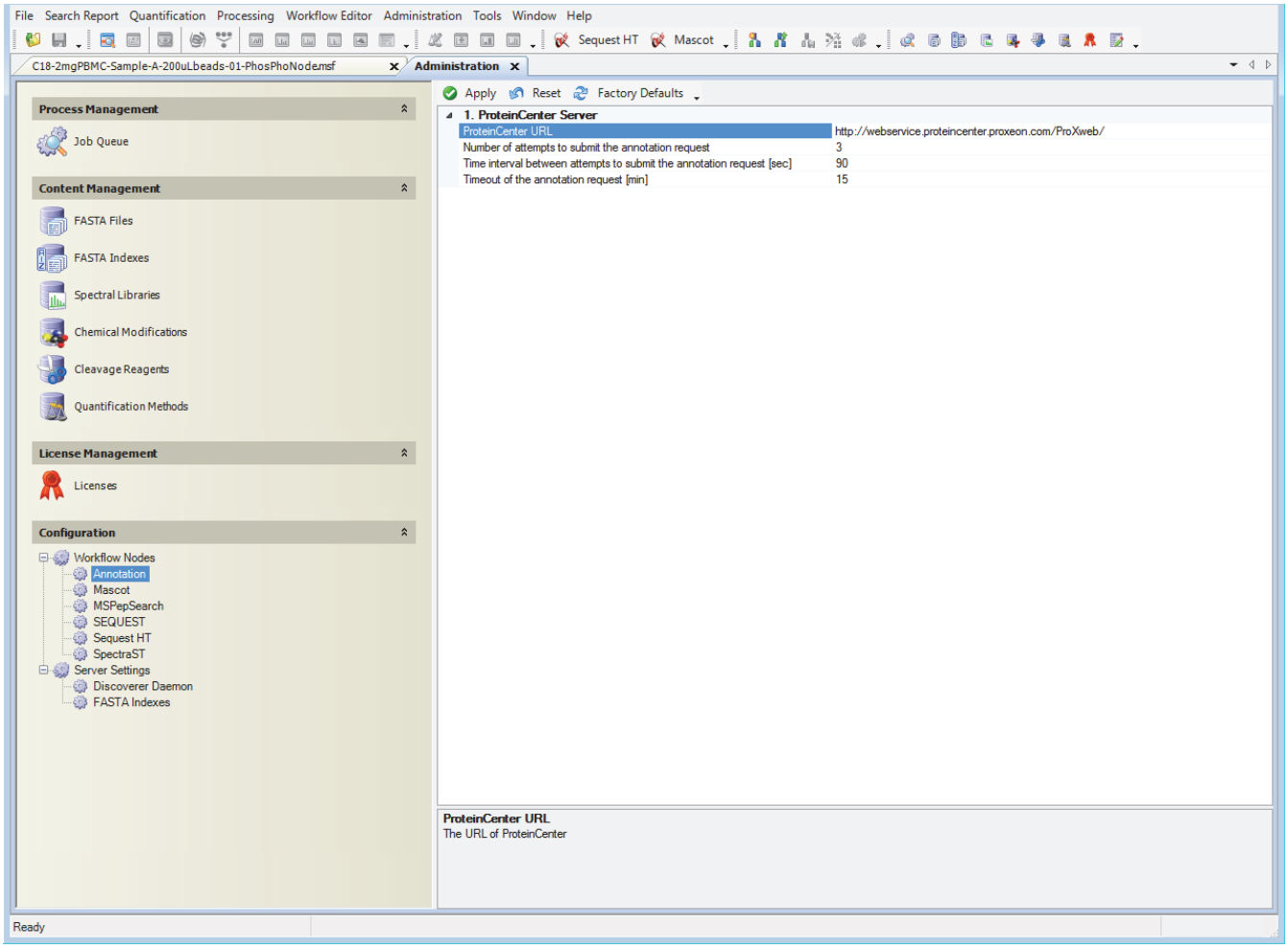
1. Choose **Administration > Configuration** or click the **Edit Configuration** icon,  .

   The Administration page changes to the Configuration view.

2. Under Workflow Nodes in the Configuration section of the left pane, click **Annotation**, if it is not already selected.

   The Annotation view appears, as shown in Figure 147.

Figure 147. Annotation view



3. In the ProteinCenter URL box, type the path and name of the ProteinCenter Web server.

   Thermo Fisher Scientific gives you this URL, a user name, and a password when you subscribe to ProteinCenter.

   Changes in the URL take effect after you restart the Proteome Discoverer application. If you entered an incorrect URL, the ProteinCard tab of the Protein Identification Details dialog box displays an error message.

4. In the Number of Attempts to Submit the Annotation Request box, specify the number of times that the Proteome Discoverer application should try to obtain the requested annotations if the ProteinCenter Web service issues an error.

   The default is 3.

5. In the Time Interval Between Attempts to Submit the Annotation Request [sec] box, specify the amount of time, in seconds, that the Proteome Discoverer application should

wait between tries to obtain the requested annotations if the ProteinCenter Web service issues an error.

The default is 90 seconds.

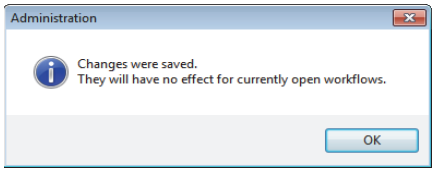
6. In the Timeout of the Annotation Request [min] box, specify the amount of time, in minutes, that the Proteome Discoverer application should continue to try to access the ProteinCenter Web service.

   The default is 15 minutes.

7. If you changed any settings, click Apply.

   The message box shown in Figure 148 appears:

   Figure 148. Administration message box



8. Click **OK**.

Tip Click Reset to return to the previous values. Click Factory Defaults to return to the values set when you first installed the Proteome Discoverer application.

9. Restart your machine.

Creating a Protein Annotation Workflow

You can retrieve annotations of all identified proteins from ProteinCenter by using the Annotation node in a workflow. This node can retrieve the following information:

- Gene Ontology (GO) annotations, which are displayed in the GO Accessions column of the Proteins page of the MSF file.
- GO Slim annotations, which are displayed in the Molecular Function, Cellular Component, and Biological Process columns of the Proteins page of the MSF file. In addition, you can define your own categories of GO Slim annotations.
- Gene identifications from the Entrez gene database, which are displayed in the Gene IDs column of the Proteins page of the MSF file.
- Protein family (Pfam) annotations, which are displayed in the Pfam IDs column of the Proteins page of the MSF file.
- UniProt PTM modifications documented in the UniProt database, which are displayed on the Proteins Identification Details view in the Proteins page of the MSF file.

The Proteome Discoverer application retrieves the annotation data after all the search nodes have finished processing.

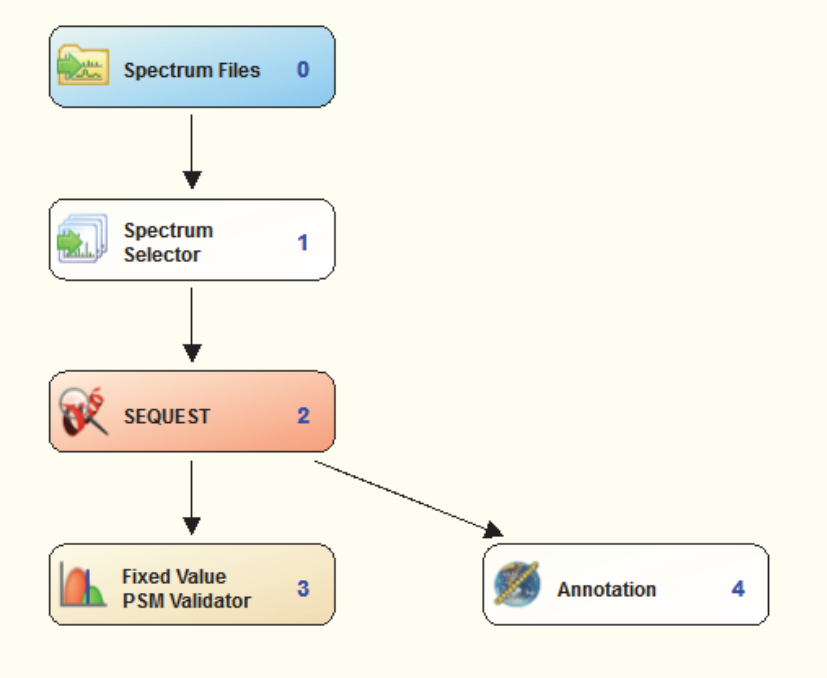
❖ To create an annotation workflow

1. Choose **Workflow Editor > New Workflow**.
2. Set up your workflow by following the instructions in "Starting a New Search by Using the Workflow Editor" on page 42.
3. In the Annotation area of the Workflow Nodes pane, select the **Annotation** node and drag it to the Workspace pane.

   The Annotation node automatically connects to the other nodes in the workflow.
4. (Optional) After you join all your chosen nodes, align them by choosing **Workflow Editor > Auto Layout** or clicking the **Auto Layout** icon ( ) or right-clicking a node and choosing **Auto Layout** from the shortcut menu.
5. (Optional) Renumber the workflow nodes in the workflow in consecutive order by choosing **Workflow Editor > Auto Number**.

   Figure 149 shows the basic protein annotation workflow.

   Figure 149. Protein Annotation workflow



6. Choose **Workflow Editor > Start Workflow** or click the **Start Workflow** icon, .

Displaying the Annotated Protein Results

The Proteome Discoverer application retrieves GO, Pfam, Entrez gene, and UniProt PTM annotation data from ProteinCenter when it finishes processing all search nodes. You can display the annotated protein results in the MSF file. For GO annotations, the application can filter the list of identified proteins by selected Go Slim categories.

Note The Proteome Discoverer application cannot retrieve annotations from searches conducted in the UniRef FASTA database because of the prefix appended to the accession number.

Displaying GO Protein Annotation Results

Follow these procedures to display GO protein categories in the MSF file.

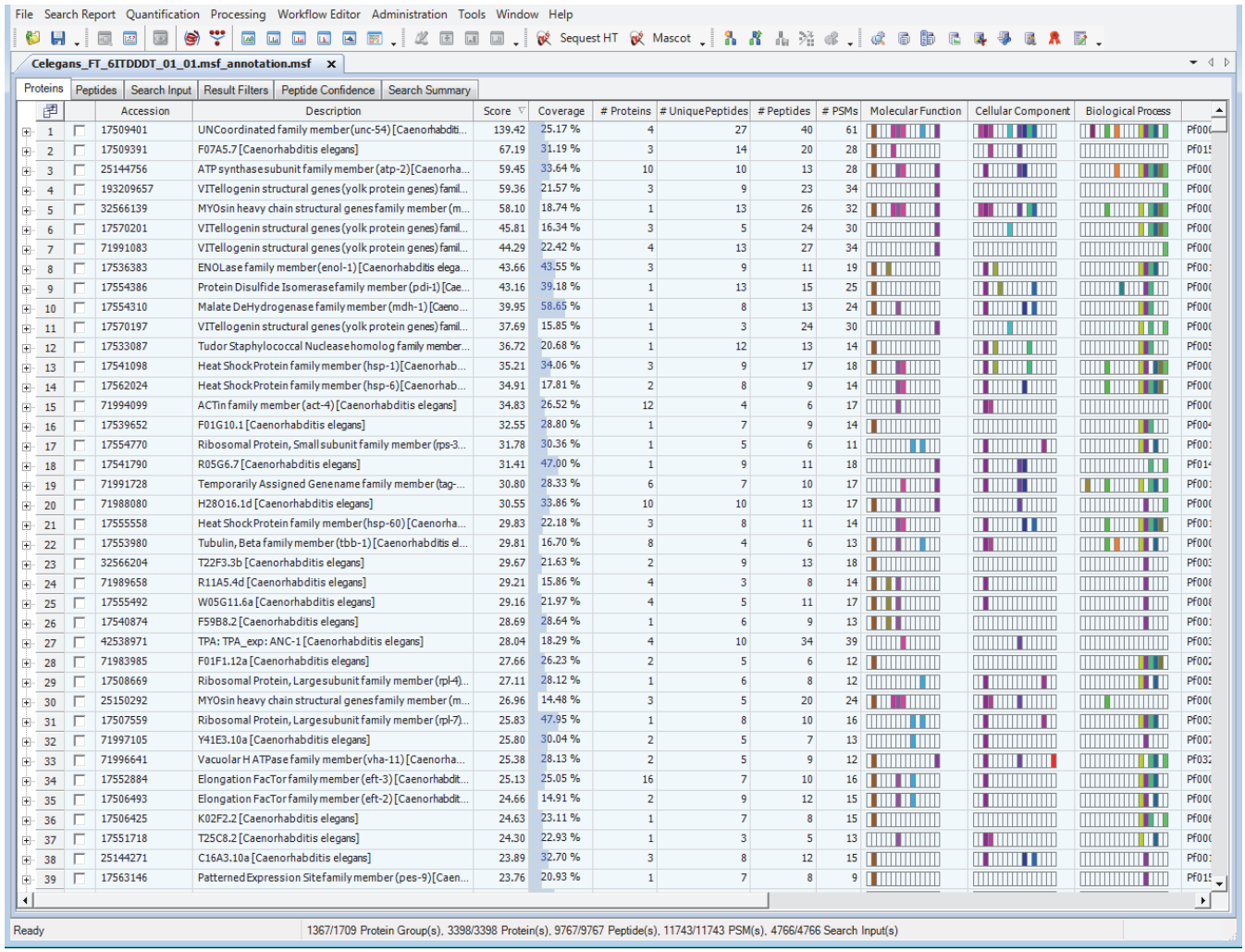
❖ To display the GO protein annotation results

1. Open the generated MSF file by following the instructions in the Help.
2. In the Column Chooser dialog box of the proteins page, select the Molecular Function, Cellular Component, and Biological Processes columns.

   For information on the Column Chooser dialog box, refer to the Help.

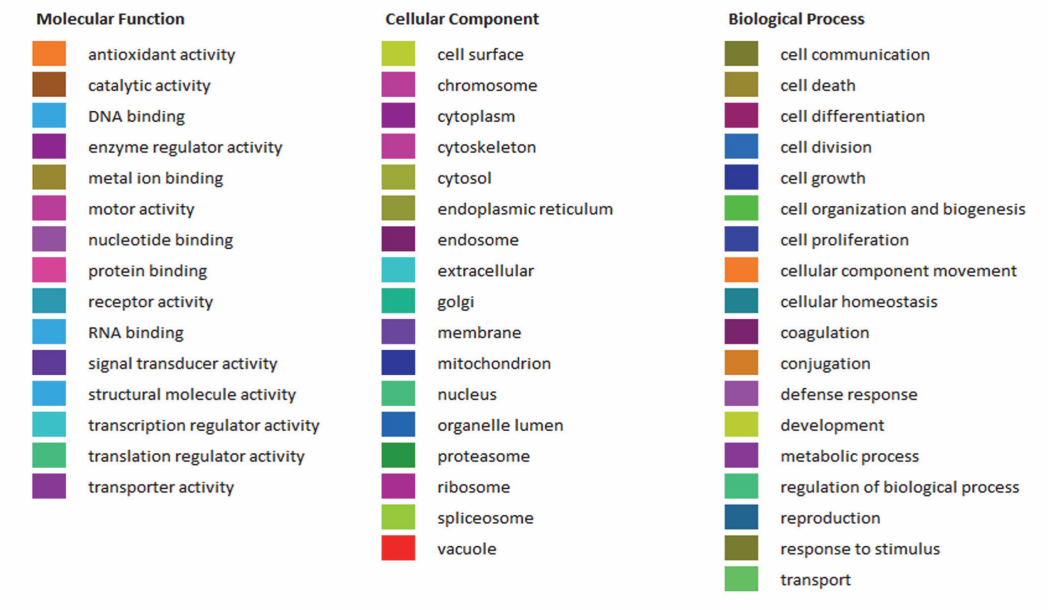
   The Proteome Discoverer application displays the results on the Proteins page of the MSF report as colored boxes similar to those shown in ProteinCenter. Figure 150 gives an example. If the application does not find the requested protein in ProteinCenter, it displays a “protein not found” message in the annotation columns. If the annotation retrieval failed because of issues with the Web request, you see an error message in the annotation columns.

Figure 150. GO Slim category boxes for the protein groups shown in the results of an annotation search



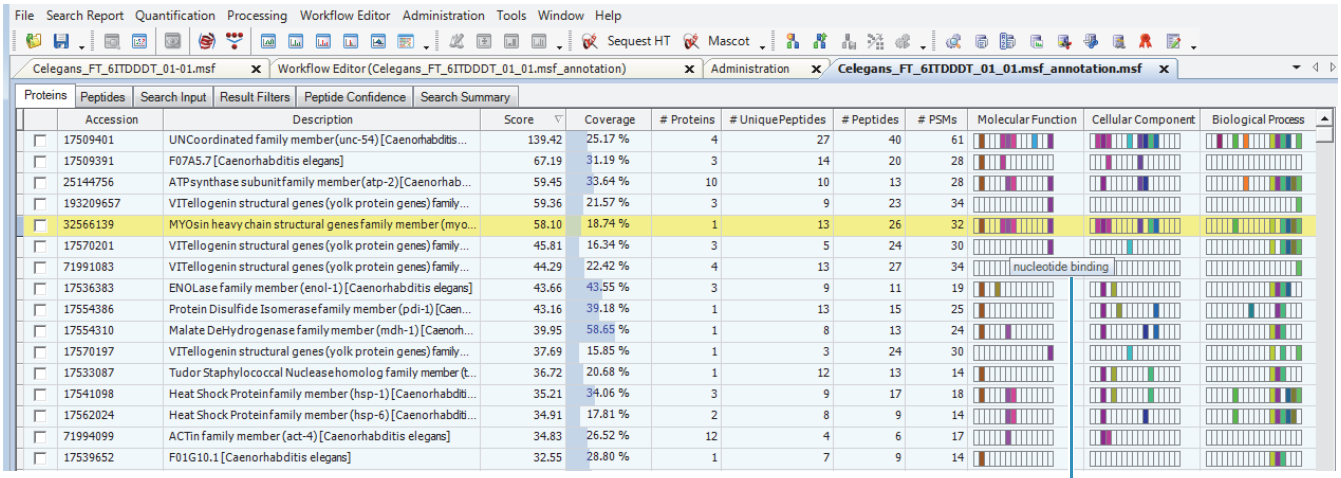
Each aspect of the annotation (biological processes, cellular components, and molecular functions) is represented in a separate column. Each box represents a GO Slim category, which is a selected subset of the Gene Ontology annotations. If the protein annotation is included in one of these subsets, the corresponding box is highlighted by a color specific to this GO Slim category. Figure 151 provides the column names and shows the meaning of the GO Slim category colors.

Figure 151. GO Slim category colors



When you hold the cursor over the GO Slim category box, the category name appears in a ToolTip, as shown in the Molecular Function column in Figure 152.

Figure 152. ToolTip identifying the annotation category



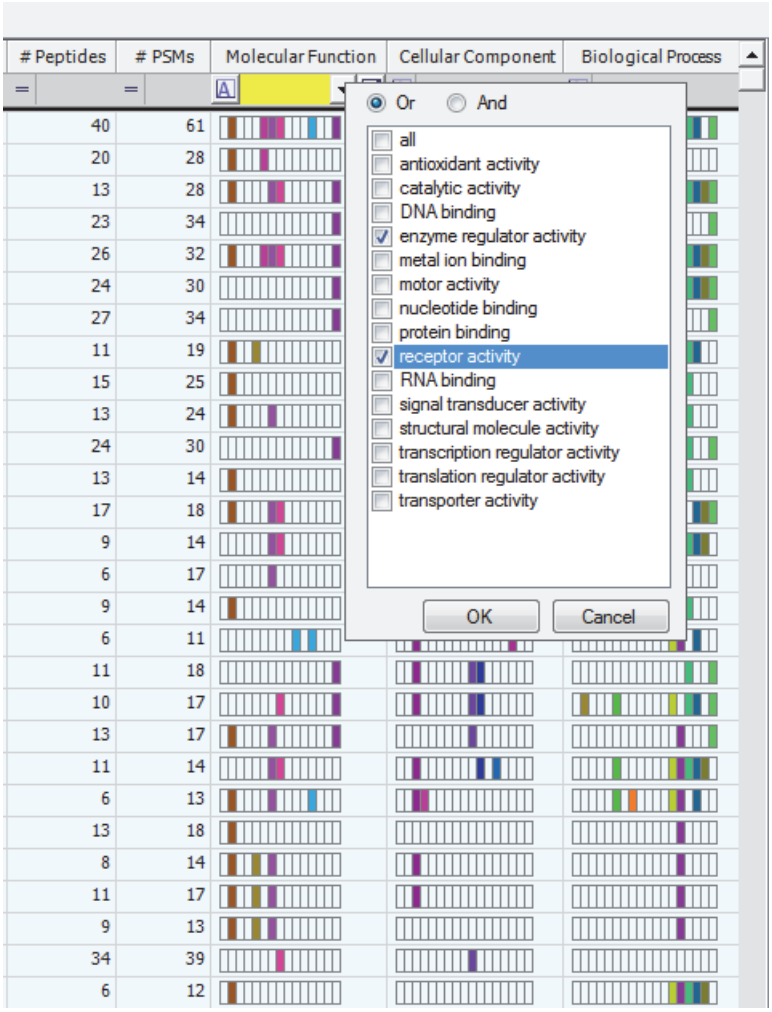
In multiconsensus reports, the protein information is displayed for the master protein of a protein group.

❖ To filter the identified proteins by GO Slim categories

1. In the MSF report, right-click the Proteins page and choose **Enable Row Filters**.
2. Click in the filter row that appears beneath the column headers in one of the GO columns, for example, Molecular Function.
3. Click  in this row.

   A dialog box appears that lists the GO Slim categories that you can filter by, as shown in Figure 153.

   Figure 153. Filtering by GO Slim category



4. Select one or more of the GO Slim categories.
5. If you selected more than one GO Slim category, select the logical **And** option at the top of the dialog box to indicate that the Proteome Discoverer application should filter by the combined categories, or select the logical **Or** option to indicate that it should filter by only one category.

6. Click **OK**.

The Proteome Discoverer application displays the identified proteins belonging to the selected categories. The names of the categories selected appear in the filter row when you expand the width of the column, as shown in Figure 154.

Figure 154. List of proteins filtered by Go Slim category

Displaying GO Accessions

Gene ontology terms are related in hierarchical graphs called GO accessions. The GO term annotated to a special protein is always part of a complex directed graph. All ancestor elements—that is, the elements between the annotated GO term and one of the three top-level terms (molecular functions, cellular components, and biological processes)—are additional less-specific descriptions of the annotated value. For example, the "iron ion binding (GO:0005506)" term contains in its graph the "metal ion binding (GO:0046872)" value, which is less specific. All GO terms contained in the graph of the annotated GO term of the protein are represented in the GO Terms column on the Proteins page.

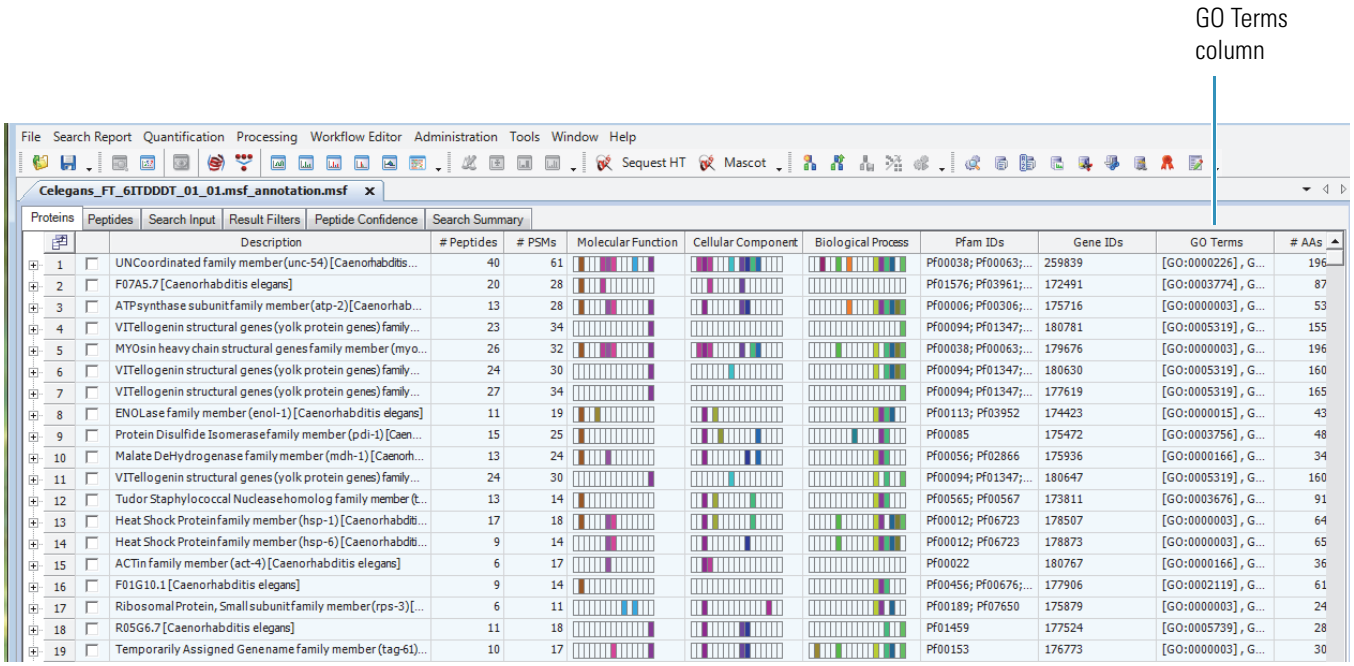
❖ To display GO accessions

1. Open the generated MSF file by following the instructions in "Opening the Results Report" on page 195.
2. In the Column Chooser dialog box of the Proteins page, select the GO Terms column.

   For information on the Column Chooser dialog box, see "Selecting the Columns to Display" on page 197.

   The Proteome Discoverer application displays the protein's GO terms contained in the graph of the annotated GO term on the Proteins page of the MSF report in the GO Terms column, as shown in Figure 155.

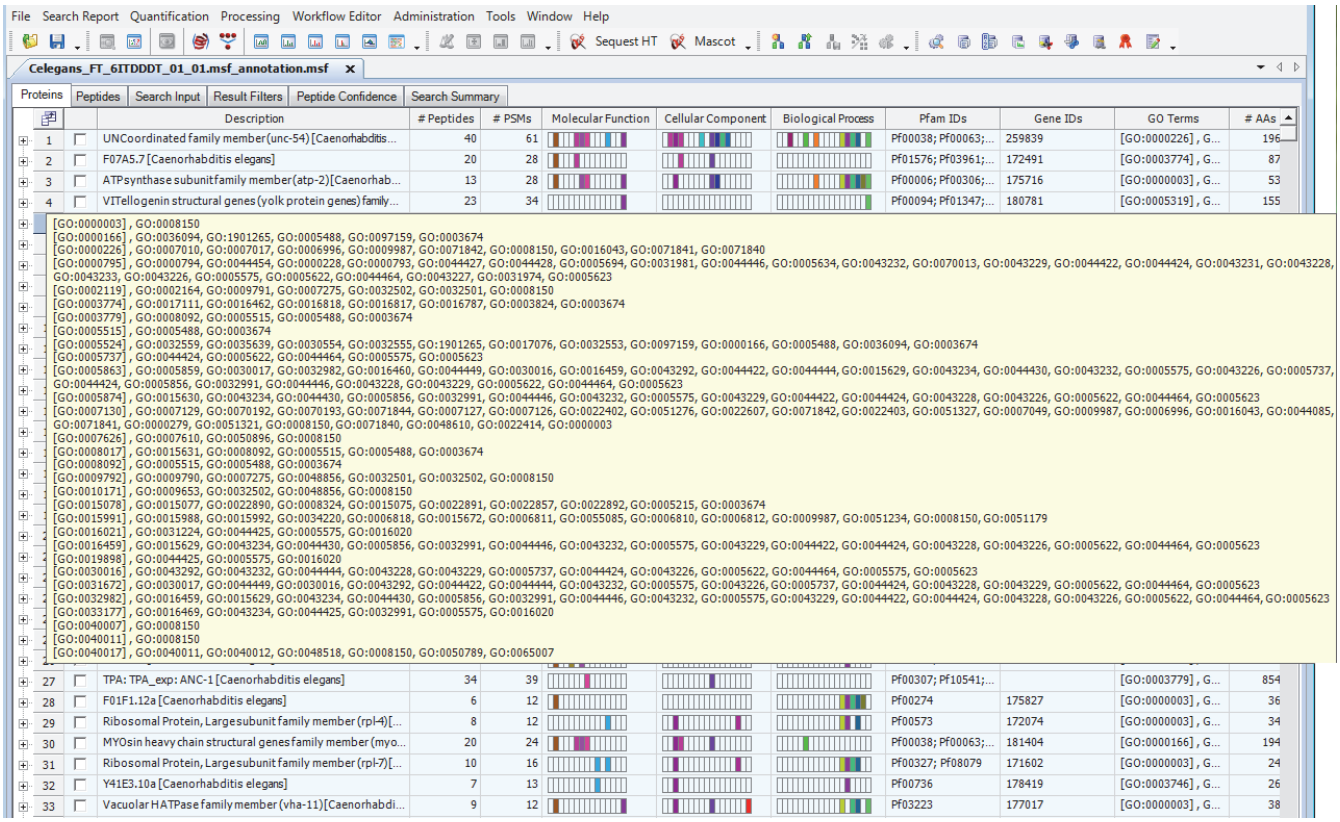
Figure 155. GO Terms column in results report



3. Move the cursor over the GO Terms column.

   The application displays the annotated GO term and all ancestor terms associated with a protein, as shown in Figure 156. It shows the term annotated to the protein in brackets, followed by their ancestor terms. Each annotated GO term starts on a new line. If you want all proteins to have a higher-level annotation that is not provided by the Molecular Function, Cellular Component, and Biological Process annotation columns, you can filter for the GO term in this column.

Figure 156. The complete list of GO terms associated with a protein



Displaying Protein Family (Pfam) Annotation Results

As noted in "Pfam Annotation" on page 203, you can retrieve Pfam annotations from the Pfam database as an alternative to GO annotations.

❖ To display Protein Family (Pfam) annotation results

1. Open the MSF file by following the instructions in the Help.
2. In the Column Chooser dialog box of the proteins page, select the Pfam IDs column.

   For information on the Column Chooser dialog box, refer to the Help.

   Figure 157 shows the Pfam IDs column on the Proteins page.

Displaying Entrez Gene Identifications

Entrez gene identifications are unique identifications assigned to all genes stored in the Entrez gene database, NCBI's database of gene-specific information. The Proteome Discoverer application displays these identifications in the Gene IDs column on the Proteins page, as shown in Figure 157. All proteins derived from the same gene have the same gene ID. You can

use this information to group or cluster biologically meaningful proteins together. Because not all genes are stored in the Entrez gene database, some proteins do not have a valid gene identification. In this case, the column is empty. For more information on the Entrez gene identifications, see "Entrez Gene Database Annotation" on page 204.

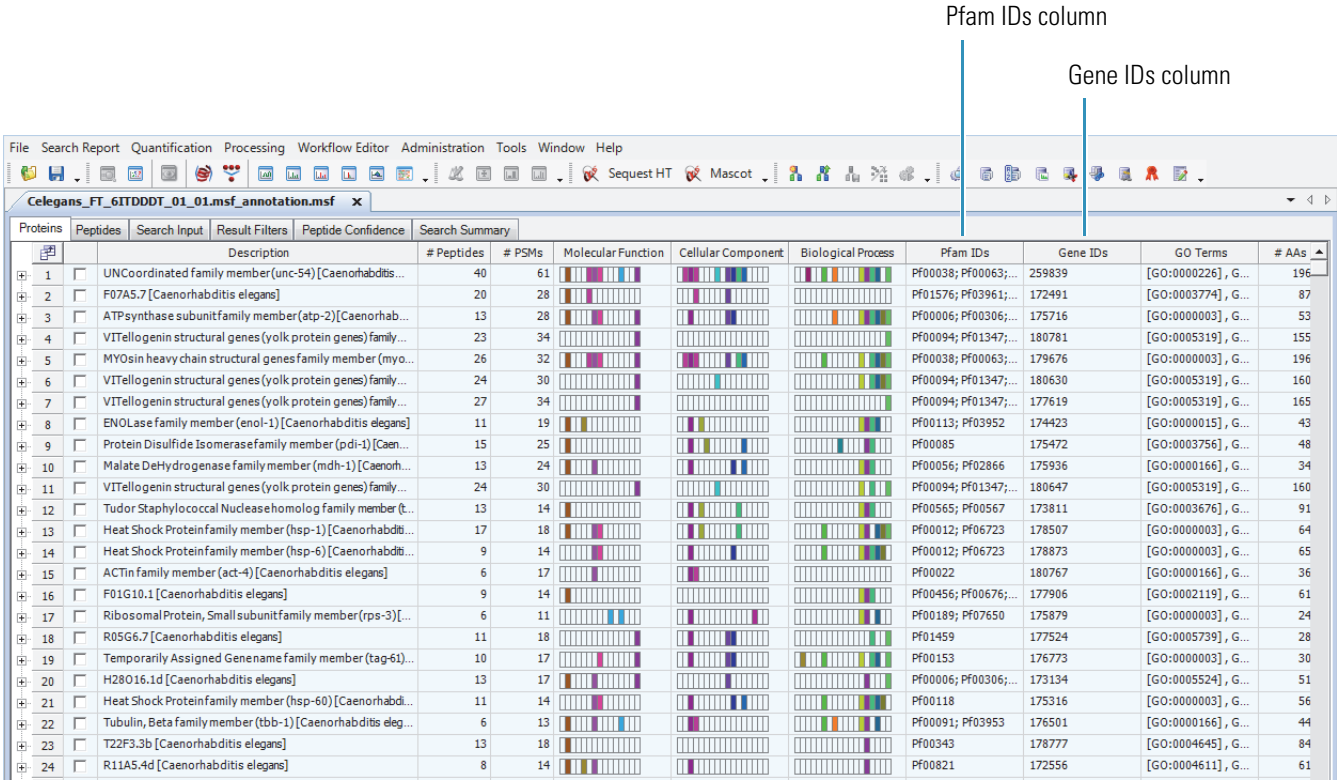
❖ To display Entrez gene identifications

1. Open the MSF file by following the instructions in the Help.
2. In the Column Chooser dialog box of the Proteins page, select the Gene IDs column.

   For information on the Column Chooser dialog box, refer to the Help.

   The Proteome Discoverer application displays the gene identifications on the Proteins page of the MSF report in the Gene IDs column, as shown in Figure 157.

Figure 157. Gene IDs column and Pfam IDs column on the Proteins page



	Description	# Peptides	# PSMs	Molecular Function	Cellular Component	Biological Process	Pfam IDs	Gene IDs	GO Terms	# AAs
1	UNCoordinated family member(unc-54) [Caenorhabditis...	40	61				Pf00038; Pf00063;...	259839	[GO:0000226], G...	196
2	F07A5.7 [Caenorhabditis elegans]	20	28				Pf01576; Pf03961;...	172491	[GO:0003774], G...	87
3	ATPsynthase subunit family member(atp-2)[Caenorhab...	13	28				Pf00006; Pf00306;...	175716	[GO:0000003], G...	53
4	VITellogenin structural genes (yolk protein genes) family...	23	34				Pf00094; Pf01347;...	180781	[GO:0005319], G...	155
5	MYOsin heavy chain structural genes family member (myo...	26	32				Pf00038; Pf00063;...	179676	[GO:0000003], G...	196
6	VITellogenin structural genes (yolk protein genes) family...	24	30				Pf00094; Pf01347;...	180630	[GO:0005319], G...	160
7	VITellogenin structural genes (yolk protein genes) family...	27	34				Pf00094; Pf01347;...	177619	[GO:0005319], G...	165
8	ENOLase family member (enol-1) [Caenorhabditis elegans]	11	19				Pf00113; Pf03952	174423	[GO:0000015], G...	43
9	Protein Disulfide Isomerase family member (pdi-1) [Caen...	15	25				Pf00085	175472	[GO:0003756], G...	48
10	Malate DeHydrogenase family member (mdh-1) [Caenorh...	13	24				Pf00056; Pf02866	175936	[GO:0000166], G...	34
11	VITellogenin structural genes (yolk protein genes) family...	24	30				Pf00094; Pf01347;...	180647	[GO:0005319], G...	160
12	Tudor Staphylococcal Nuclease homolog family member (t...	13	14				Pf00565; Pf00567	173811	[GO:0003676], G...	91
13	Heat Shock Protein family member (hsp-1) [Caenorhabditi...	17	18				Pf00012; Pf06723	178507	[GO:0000003], G...	64
14	Heat Shock Protein family member (hsp-6) [Caenorhabditi...	9	14				Pf00012; Pf06723	178873	[GO:0000003], G...	65
15	ACTin family member (act-4) [Caenorhabditis elegans]	6	17				Pf00022	180767	[GO:0000166], G...	36
16	F01G10.1 [Caenorhabditis elegans]	9	14				Pf00456; Pf00676;...	177906	[GO:0002119], G...	61
17	Ribosomal Protein, Small subunit family member (rps-3)[...	6	11				Pf00189; Pf07650	175879	[GO:0000003], G...	24
18	R05G6.7 [Caenorhabditis elegans]	11	18				Pf01459	177524	[GO:0005739], G...	28
19	Temporarily Assigned Gene name family member (tag-61)...	10	17				Pf00153	176773	[GO:0000003], G...	30
20	H28O16.1d [Caenorhabditis elegans]	13	17				Pf00006; Pf00306;...	173134	[GO:0005524], G...	51
21	Heat Shock Protein family member (hsp-60) [Caenorhabdi...	11	14				Pf00118	175316	[GO:0000003], G...	56
22	Tubulin, Beta family member (tbb-1) [Caenorhabditis eleg...	6	13				Pf00091; Pf03953	176501	[GO:0000166], G...	44
23	T22F3.3b [Caenorhabditis elegans]	13	18				Pf00343	178777	[GO:0004645], G...	84
24	R11A5.4d [Caenorhabditis elegans]	8	14				Pf00821	172556	[GO:0004611], G...	61

Displaying UniProt Annotation Data

For information on displaying UniProt PTM annotation data, refer to the Help.

Reannotating MSF Files

You can use the Re-Annotation node in the Workflow Editor or the batch processing function in Discoverer Daemon to update existing annotations or annotate existing MSF files that do not yet include annotations.

Use the Re-Annotation node in the Workflow Editor to reannotate a single file. The Re-Annotation node must be the only node in a workflow. It takes an existing MSF file as input, retrieves up-to-date annotations for the proteins contained in the MSF file, and stores them in the same MSF file.

Note If you used a previous version of the Proteome Discoverer application to create the MSF file to reannotate, the application first updates the file to comply with the current results file schema.

Use the batch processing function in Discoverer Daemon to reannotate multiple files.

- To reannotate an MSF file in the Workflow Editor
- To reannotate an MSF file in Proteome Discoverer Daemon

❖ To reannotate an MSF file in the Workflow Editor

1. Choose **Workflow Editor > New Workflow**.
2. In the Annotation area of the Workflow Nodes pane, select only the **Re-Annotation** node and drag it to the Workspace pane.
3. Select the **Re-Annotation** node.
4. Click the MSF File Path box, and then click the **Browse** button (...) to open the Select Analysis File dialog box.
5. Browse to the MSF file to save the new annotations in, or type the path and name of the file in the File Name box, and click **Open**.

   The name of the MSF file appears in the Name box in the Workflow Editor.
6. Choose **Workflow Editor > Start Workflow** or click the **Start Workflow** icon,  .

The Proteome Discoverer application submits the workflow to standard workflow processing and displays the reannotation progress in the job queue.

Note If you created the MSF file that you want to reannotate with a previous version of the Proteome Discoverer application, the application updates the file first to comply with the current result file schema.

❖ To reannotate an MSF file in Proteome Discoverer Daemon

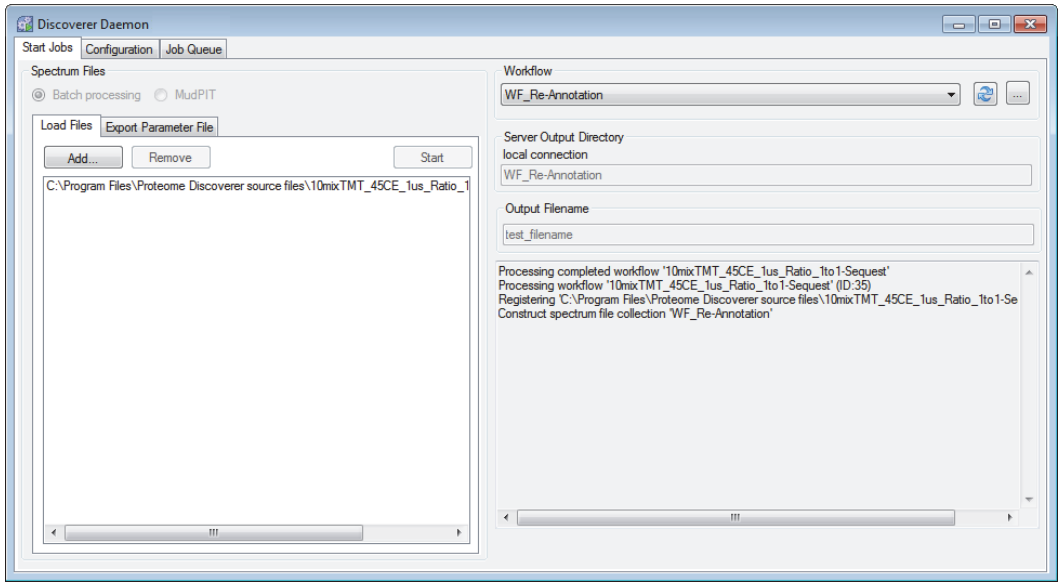
1. Create a reannotation workflow in the Workflow Editor according to the instructions in "To reannotate an MSF file in the Workflow Editor" on page 216.
2. Save the workflow as a new workflow template:
   a. Choose **Workflow Editor > Save As Template**.
   b. In the Save Processing Workflow Template dialog box, type the name of the template in the Template Name box.
   c. Give a brief description of the template in the Template Description box.
   d. Click **Save**.

   This newly created workflow template is now available in Discoverer Daemon.
3. To start Discoverer Daemon, follow the instructions in "Starting the Proteome Discoverer Daemon Application in a Window" on page 70.
4. To select the server, follow the instructions in "Selecting the Server" on page 70.
5. Click the **Start Jobs** tab if it is not already selected.
6. Click the **Load Files** tab if it is not already selected.
7. Click **Add**.
8. In the Open dialog box, select **Result Files (*.msf)** from the list next to File Name.
9. Browse to the MSF file that you want to save the new annotations in, or type the name of the file in the File Name box, and click **Open**.
10. Repeat step 8 to add the names of multiple MSF files to reannotate.
11. In the Spectrum Files area, click **Batch Processing**.
12. From the menu in the Workflow box, select the reannotation workflow template that you saved in the Workflow Editor.
13. Start the batch processing:
    - If you are connected to an instance of the Proteome Discoverer application running on the same computer, click **Start** in Discoverer Daemon.
    - If you are connected to an instance of the Proteome Discoverer application running on a remote machine, specify in the Server Output Directory box the name of the folder where you want the original output files placed on the server, and then click **Start**.

By default, the Proteome Discoverer Daemon application places this folder in the c:\Documents and Settings\All Users\...\DiscovererDaemon\SpectrumFiles\ directory. You can specify a different folder by choosing Administration > Configuration in the Proteome Discoverer application, clicking Discoverer Daemon in the Server Settings section, and browsing to the location in the New Directory box.

Figure 158 shows MSF files being processing in batch mode in Discoverer Daemon.

Figure 158. Reannotating MSF files in batch mode in Discoverer Daemon



For more information about processing files with Discoverer Daemon, see "Using the Proteome Discoverer Daemon Utility" on page 69.

Uploading Results to ProteinCenter

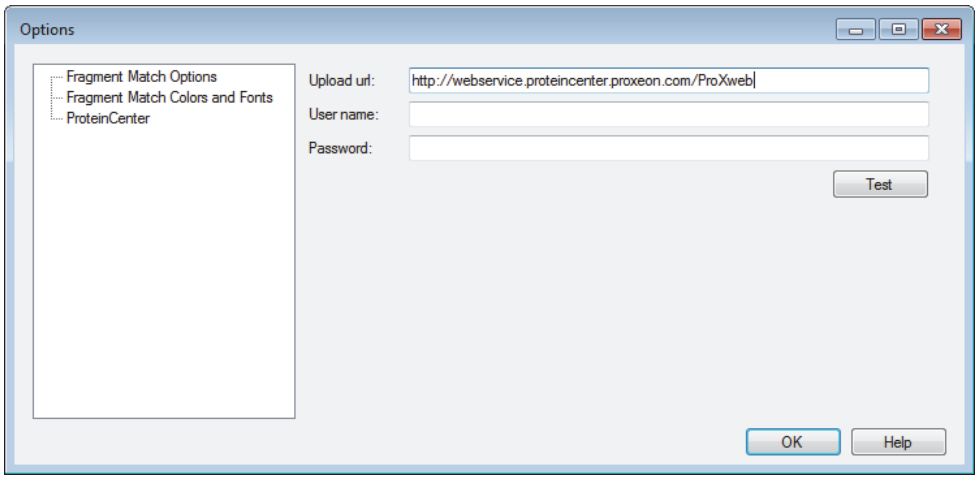
If you have a user account on a ProteinCenter server, you can upload search results directly from the Proteome Discoverer application to ProteinCenter.

❖ To upload search results to ProteinCenter

1. Open an MSF file and be sure that it is selected.
2. Choose **Tools > Options**.
3. In the Options dialog box, click **ProteinCenter**.

   The ProteinCenter page opens, as shown in Figure 159.

Figure 159. ProteinCenter page of the Options dialog box



a. In the URL box, type the URL of the ProteinCenter server to use.

b. In the User Name box, type the user name of your ProteinCenter user account.

c. In the Password box, type the password of your ProteinCenter user account.

d. Click **OK**.

A message box appears with the following message:

Settings of Protein Center changed. Do you want to save your changes?

4. Click **Yes**.

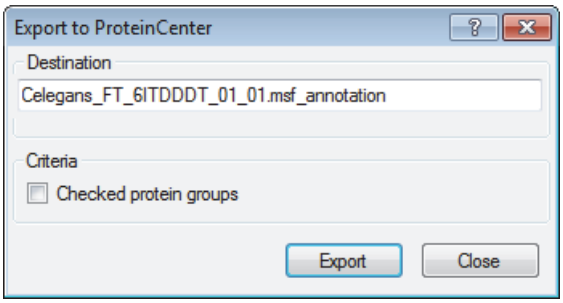
5. Open an MSF file in the Proteome Discoverer application. Refer to the Help.

6. Choose **Tools > Export to ProteinCenter**.

   The Export to ProteinCenter dialog box opens.

7. In the Destination box, specify the name of the data set to upload to ProteinCenter, as shown in Figure 160.

Figure 160. Export to ProteinCenter dialog box



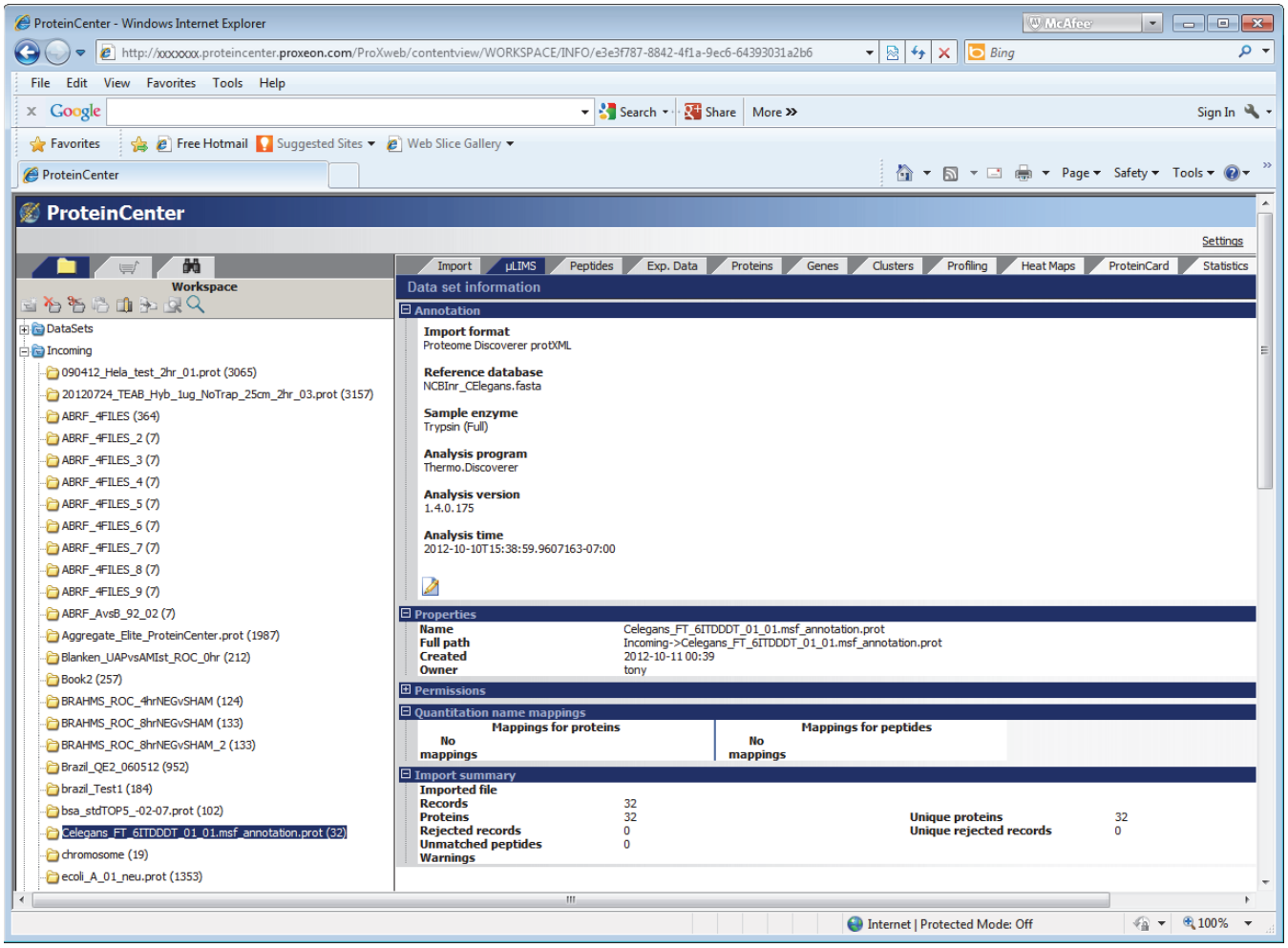
8. If you want to export only the result data from selected protein groups, select the **Checked Protein Groups** check box.

If you do not select Checked Protein Groups, the Proteome Discoverer application exports the result data of all protein groups.

9. Click **Export**.

After the Proteome Discoverer application exports the data set to ProteinCenter, you can log in to your ProteinCenter account. The uploaded data set appears under the Incoming node in the ProteinCenter window, as shown in Figure 161.

Figure 161. Uploaded data set in the ProteinCenter window



ProteinCenter Page Parameters

Table 10 lists the parameters on the ProteinCenter page of the Options dialog box.

Table 10. ProteinCenter page parameters

Command or Option	Description
Upload URL	Specifies the URL of the ProteinCenter server to use to upload your search results.
User Name	Specifies the user name of your ProteinCenter user account.
Password	Specifies the password of your ProteinCenter user account.
Test	Verifies that the URL that you specified in the URL box is valid. However, it does not verify that the user name and password are valid.

Accessing ProteinCards

You can access the data in ProteinCenter through the ProteinCard for each protein. In ProteinCard, a protein is considered a specific amino acid sequence in a given species.

❖ **To access the data in ProteinCenter**

1. Double-click a grid cell on the Proteins page of the MSF file, or select a cell and choose **Search Report > Show Protein ID Details**, or click the **Show Protein/Peptide ID Details** icon,  .

   You might experience a short delay as the Proteome Discoverer application accesses the URL.

2. In the Protein Identification Details dialog box, click the **ProteinCard** tab.

   After loading data from the ProteinCenter server, the Proteome Discoverer application displays the data in the ProteinCard tab. By default, it shows the General tab, shown in Figure 163 on page 223.

3. Click the tab of the page containing the information that you are seeking:

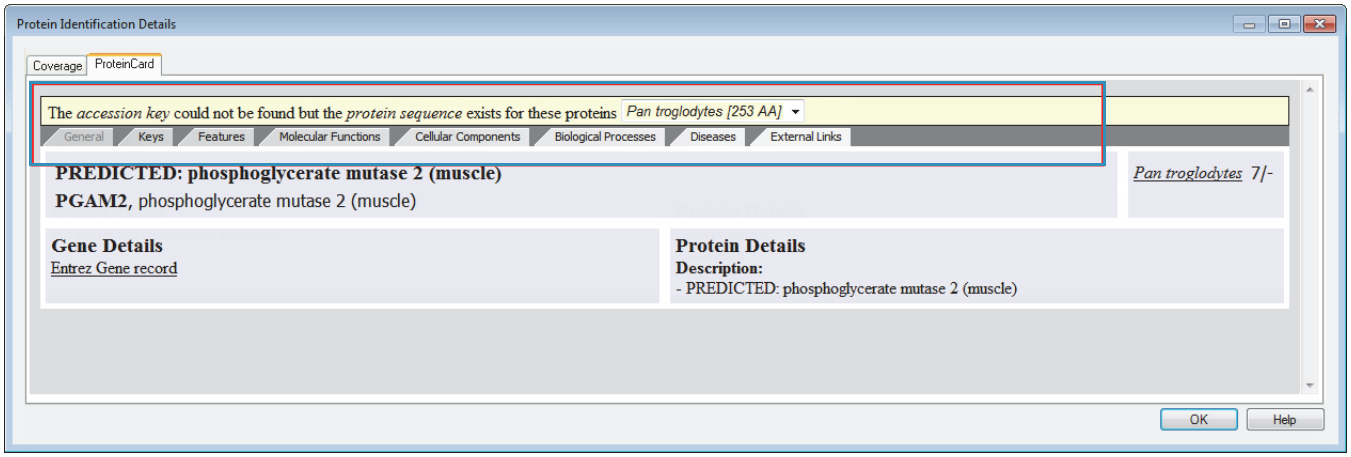
   - General Page
   - Keys Page
   - Features Page
   - Molecular Functions Page
   - Cellular Components Page

- Biological Processes Page
- Diseases Page
- External Links Page

4. Click **OK** to close the Protein Identification Details dialog box.

If the entire protein is not found in ProteinCenter but a protein with the same sequence exists, the ProteinCard displays a warning that the displayed information is from a protein with different accession, as shown in Figure 162. If there is more than one protein with the same sequence but from different organisms, an additional list box appears so that you can select the correct species.

Figure 162. Warning displayed for protein with different accession



ProteinCard Parameters

The ProteinCard page of the Protein Identification Details dialog box contains the following pages.

- General Page
- Keys Page
- Features Page
- Molecular Functions Page
- Cellular Components Page
- Biological Processes Page
- Diseases Page
- External Links Page

General Page

The General page of the ProteinCard, shown in Figure 163, displays information about the protein: its name, its description, its function, the keywords that produce it in a database search, and the gene that ultimately directs the protein's synthesis through RNA.

Figure 163. General page of the ProteinCard

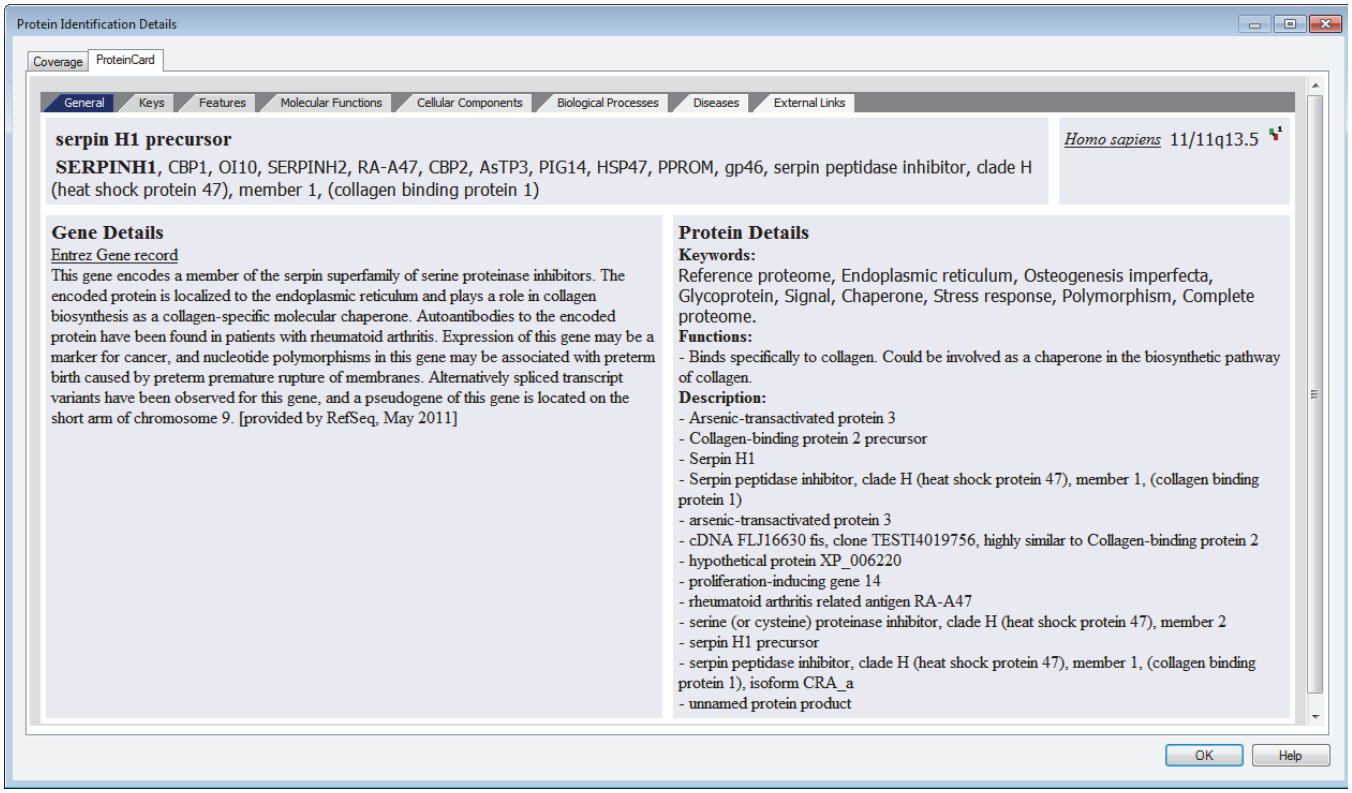


Table 12 lists the parameters on the General page of the ProteinCard page.

Table 11. Parameters on the General page of the ProteinCard page (Sheet 1 of 2)

Command	Description
Top area	Displays the protein name in bold font on the first line. The second line in bold font is the official symbol of the gene that ultimately directs the synthesis of the protein through RNA, and the text following it is the alternative name or names of the gene.
Top right area	Displays the name of the species that contains the gene that ultimately directs the synthesis of this protein through RNA, the number of the chromosome that the gene resides on, and the location of the chromosome that the gene resides on. The name of the species is linked to the National Center for Biotechnology Information (NCBI) taxonomy browser.

Table 11. Parameters on the General page of the ProteinCard page (Sheet 2 of 2)

Command	Description
Gene Details area	Displays information about the gene that directs the synthesis of the protein. If no information about the gene is available, a link to the Entrez database Web site is given.
Protein Details area	Lists the keywords that produce this protein in a database search, the functions of the protein, and a description of the protein.

Keys Page

The Keys page of the ProteinCard, shown in Figure 164, lists all the accession keys for a given protein.

Figure 164. Keys page of the ProteinCard page

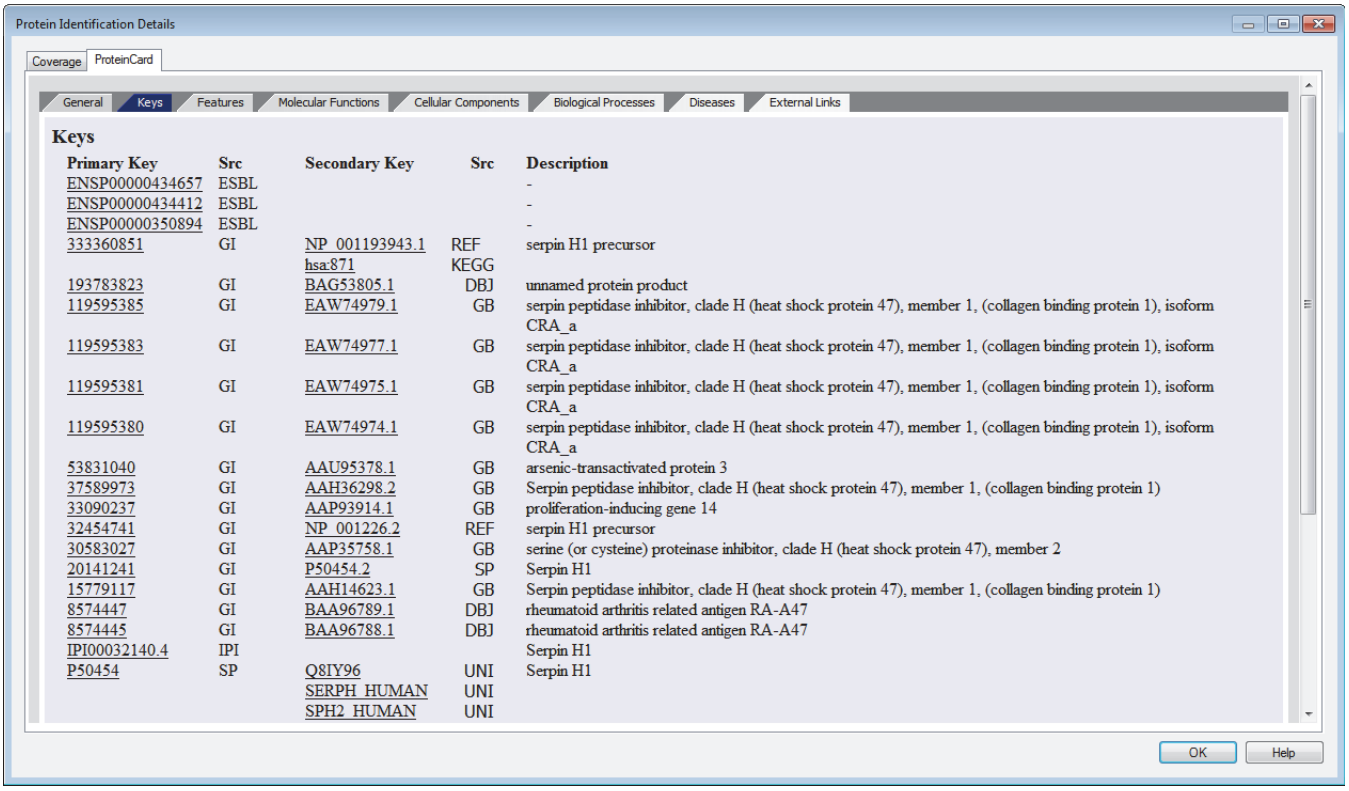


Table 12 lists the parameters on the Keys page of the ProteinCard page.

Table 12. Parameters on the Keys page of the ProteinCard page

Command	Description
Primary Key	Lists the accession key of the database that the sequence was imported from. It is linked to the original database records in the source database, such as Ensembl, SGD, NRDB, IPI, or UniProt. The preferred type of accession is emphasized.
Src	Specifies the abbreviation of the primary source database.
Secondary Key	Lists the secondary accession key which is either an alternative key used in the source database or the key of the original database.
Src	Specifies the abbreviation of the secondary source database.
Description	Displays the original description for the original database entry.

An exclamation mark flags outdated protein keys, and the keys are linked to the outdating history in their respective source database.

Features Page

The Features page of the ProteinCard page, shown in Figure 165, includes a selection of sequence features from UniProt, from various conserved domain predictions, and from the computational enrichment undertaken by ProteinCenter. (Computational enrichment refers to information that has no experimental evidence but was found by using a computer prediction program.) The features are sorted according to their start positions in the protein sequence.

Figure 165. Features page of the ProteinCard page

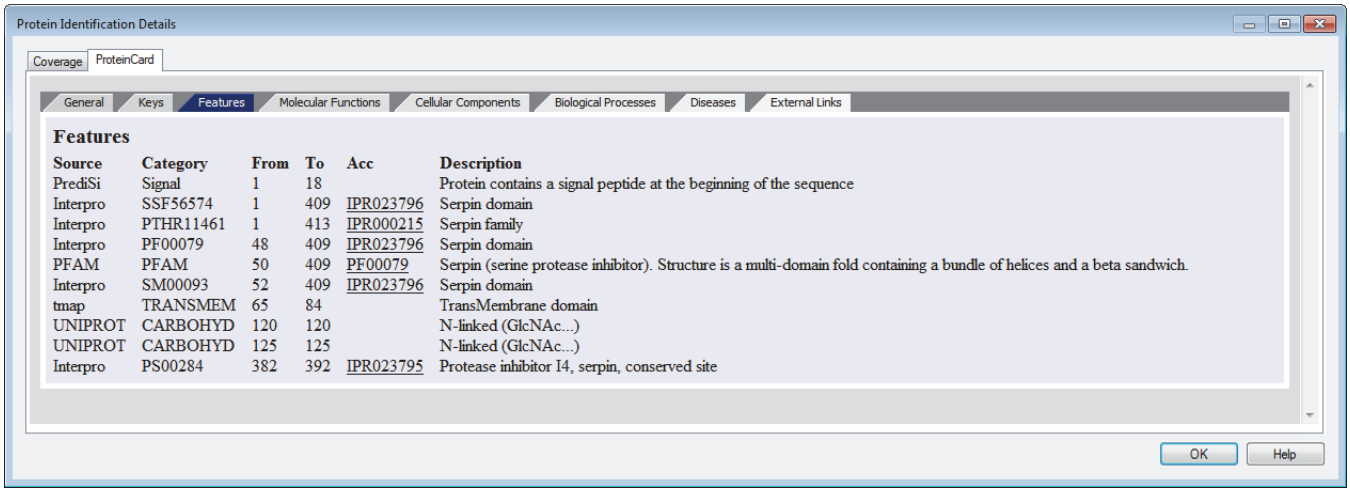


Table 13 lists the parameters on the Features page of the ProteinCard page.

Table 13. Parameters on the Features page of the ProteinCard page

Command	Description
Source	Specifies the name of the database that the information about the feature was taken from: • InterPro • Tmap (computational enrichment) • PrediSi (computational enrichment) • Pfam (computational enrichment) • UniProt
Category	Displays the type of information that UniProt, InterPro, and Tmap include for each row. For example, UniProt might include “CARBOHYD” as one of its types of information, and InterPro might include “SSF57184” as one of its types of information.
From	Specifies the start position of the amino acid.
To	Specifies the end position of the amino acid.
Acc	Specifies the accession identifier for the domain linked to InterPro or Pfam.
Description	Describes the feature.

Molecular Functions Page

The Molecular Functions page of the ProteinCard page, shown in Figure 166, summarizes information about the function of the protein. It consolidates GO data and Enzyme Category (EC) information. The EC designation indicates whether a protein has been categorized with a certain enzyme function.

Figure 166. Molecular Functions page of the ProteinCard page

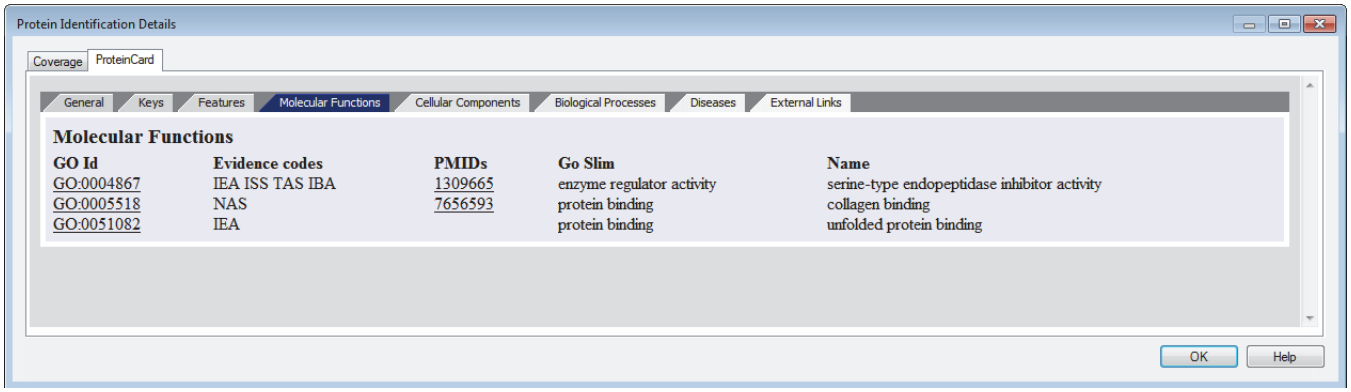


Table 14 lists the parameters on the Molecular Functions page of the ProteinCard page.

Table 14. Parameters on the Molecular Functions page of the ProteinCard page (Sheet 1 of 2)

Command	Description
GO Id	Lists the GO code for each of the protein's molecular functions. Each code is linked to the QuickGO browser of the European Bioinformatics Institute (EBI), which hosts several databases and services.
Evidence Codes	Lists the evidence codes for each of the protein's molecular functions for GO annotation. Evidence codes describe how the GO information was proven—for example, by computer prediction or by experiment.
PMIDs	Lists the molecular function codes in the PubMed database, which is maintained by the U.S. National Library of Medicine (NLM) and the National Institutes of Health (NIH). Each code is linked to the PubMed browser.

Table 14. Parameters on the Molecular Functions page of the ProteinCard page (Sheet 2 of 2)

Command	Description
Go Slim	Specifies the basic GO Slim category for the GO term. GO Slim categories are reduced versions of the GO ontologies containing a subset of the terms in the entire GO database. They give a broad overview of the ontology content without the detail of the specific fine-grained terms. Table 17 on page 233 provides the Go Slim categories for molecular functions.
Name	Describes the molecular function for a GO term. This description is created by the GO consortium.

Enzymes with an EC number for IUBMB Enzyme Nomenclature are displayed with links to detailed information at the International Union of Biochemistry and Molecular Biology.

Cellular Components Page

The Cellular Components page of the ProteinCard page, shown in Figure 167, summarizes information about where the protein carries out its function in the cell.

Figure 167. Cellular Components page of the ProteinCard page

Table 15 lists the parameters on the Cellular Components page of the ProteinCard page.

Table 15. Parameters on the Cellular Components page of the ProteinCard page

Command	Description
GO Id	Lists the GO code for each of the protein's molecular functions. Each code is linked to the QuickGO browser of the EBI, which hosts a number of databases and services.
Evidence Codes	Lists the evidence codes for each of the protein's cellular components for GO annotation. Evidence codes describe how the GO information was proven—for example, by computer prediction or by experiment.
PMIDs	Lists the cellular component codes in the PubMed database, which is maintained by the NLM and the NIH. Each code is linked to the PubMed browser.
Go Slim	Specifies the basic GO Slim category for the GO term. GO Slim categories are reduced versions of the GO ontologies containing a subset of the terms in the entire GO database. They give a broad overview of the ontology content without the detail of the specific fine-grained terms. Table 18 provides the Go Slim categories for cellular components.
Name	Describes the cellular component for a GO term. This description is created by the GO consortium.

Enzymes with an EC number for IUBMB Enzyme Nomenclature are displayed with links to detailed information at the International Union of Biochemistry and Molecular Biology.

Biological Processes Page

The Biological Processes page of the ProteinCard page, shown in Figure 168, summarizes information about the biological processes that the protein is a part of.

Figure 168. Biological Processes page of the ProteinCard page

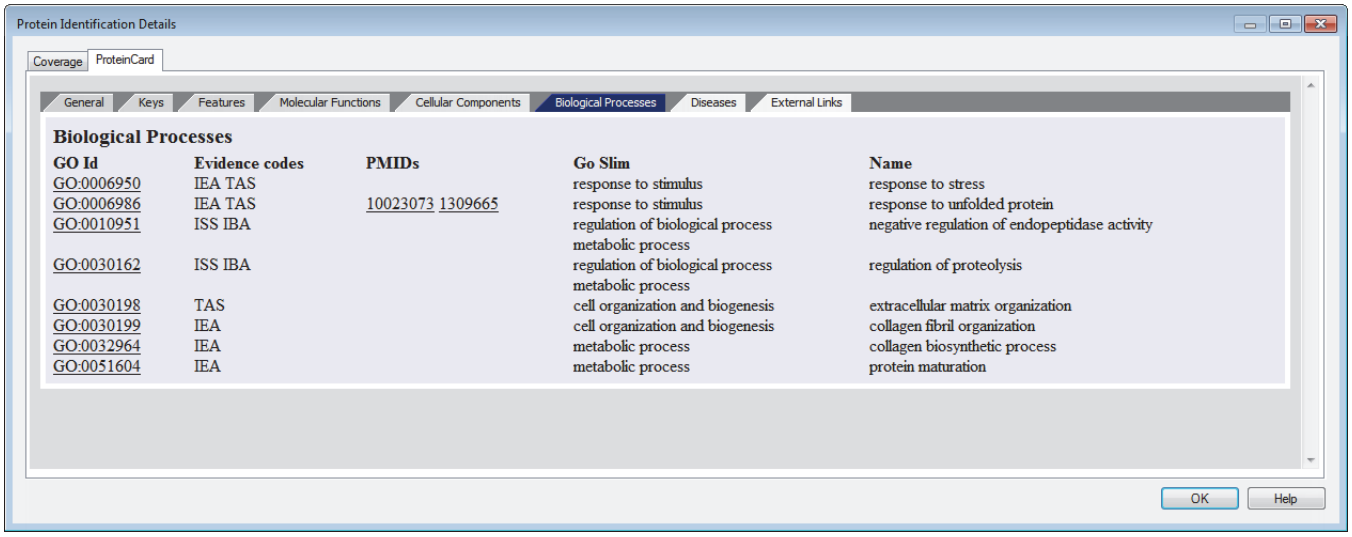


Table 16 lists the parameters on the Biological Processes page of the ProteinCard page.

Table 16. Parameters on the Biological Processes page of the ProteinCard page

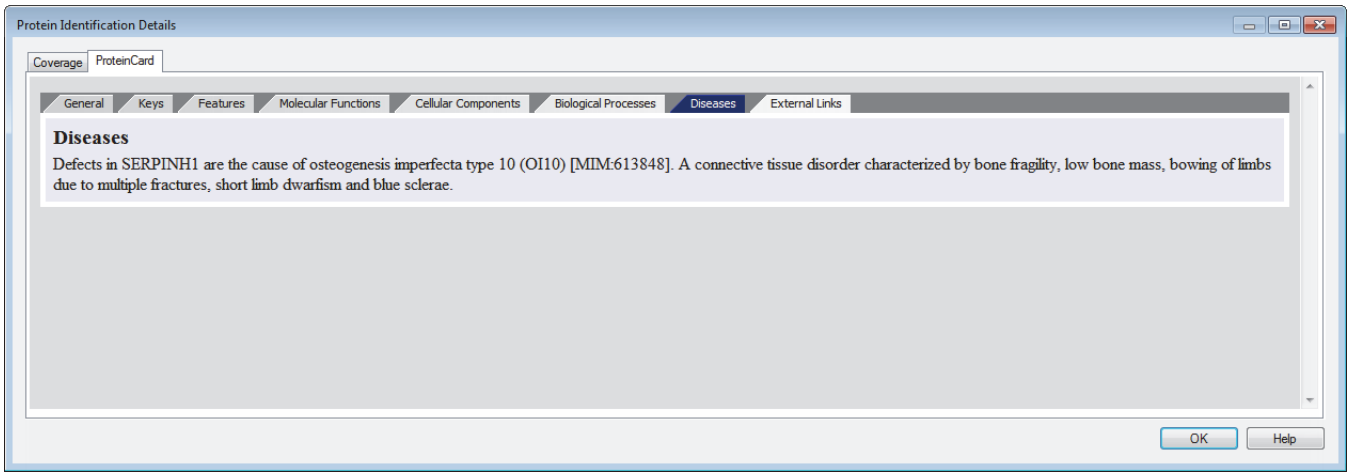
Command	Description
GO Id	Lists the GO code for each of the protein's molecular functions. Each code is linked to the QuickGO browser of the EBI, which hosts a number of databases and services.
Evidence Codes	Lists the evidence codes for each of the protein's biological processes for GO annotation. Evidence codes describe how the GO information was proven—for example, by computer prediction or by experiment.
PMIDs	Lists the biological process codes in the PubMed database, which is maintained by the NLM and the NIH. Each code is linked to the PubMed browser.
Go Slim	Specifies the basic GO Slim category for the GO term. GO Slim categories are reduced versions of the GO ontologies containing a subset of the terms in the entire GO database. They give a broad overview of the ontology content without the detail of the specific fine-grained terms. Table 19 provides the Go Slim categories for biological components.
Name	Describes the biological process for a GO term. This description is created by the GO consortium.

Enzymes with an EC number for IUBMB Enzyme Nomenclature are displayed with links to detailed information at the International Union of Biochemistry and Molecular Biology.

Diseases Page

The Diseases page of the ProteinCard page, shown in Figure 169, lists the diseases that the selected protein is associated with.

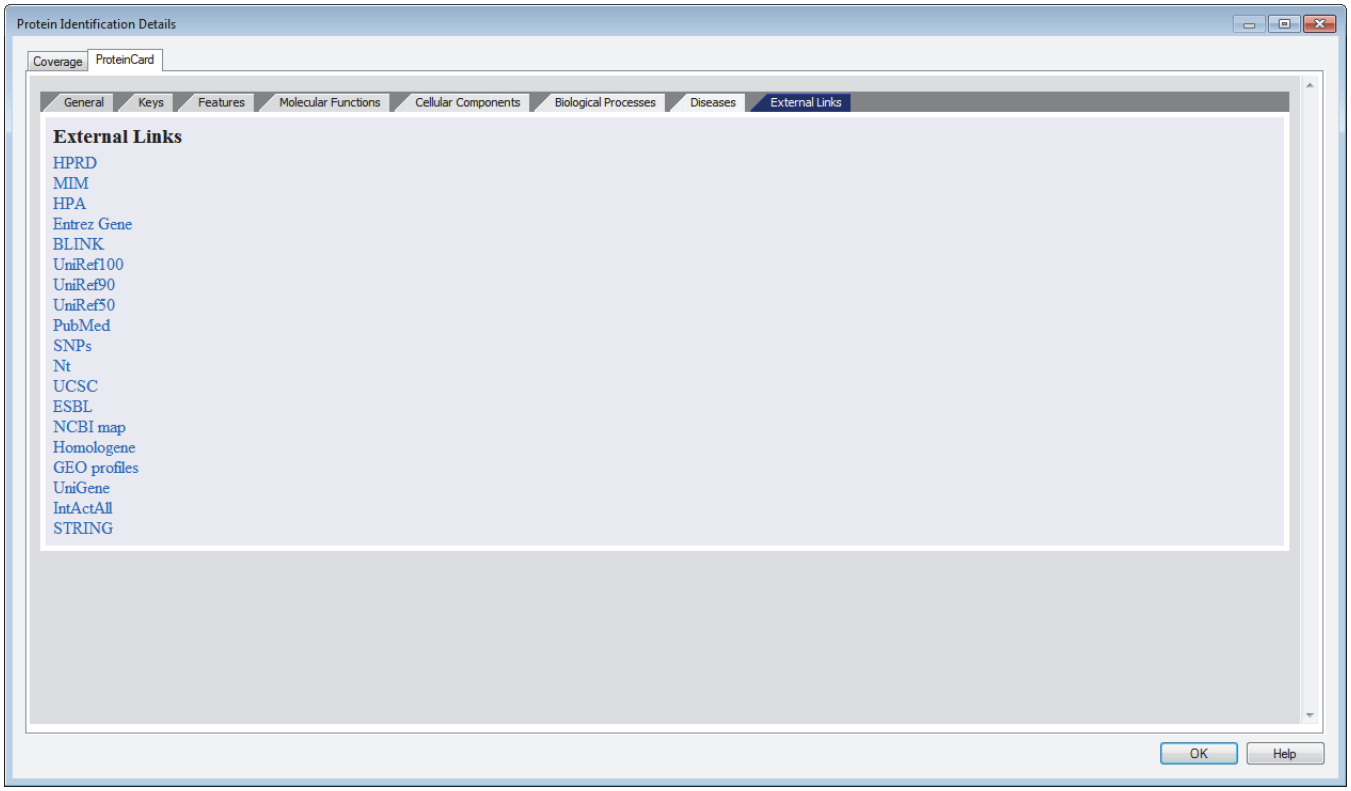
Figure 169. Diseases page of the ProteinCard page



External Links Page

The External Links page of the ProteinCard page, shown in Figure 170, lists the Web links to resources containing information about the protein.

Figure 170. External Links page of the ProteinCard page



Click the appropriate link to open the browser for the database. The external links contains links to resources containing information about the respective protein.

GO Slim Categories

This section defines the GO Slim terms for molecular functions, cellular components, and biological processes.

GO Slim Categories for Molecular Functions

Table 17 describes the GO Slim categories for molecular functions.

Table 17. GO Slim categories for molecular functions (Sheet 1 of 2)

GO Slim molecular function	Description
Antioxidant activity	Inhibition of the reactions brought about by dioxygen (O2) or peroxides. Usually the antioxidant is effective because it can be more easily oxidized than the substance protected. The term is often applied to components that can trap free radicals, breaking the chain reaction that normally leads to extensive biological damage.
Catalytic activity	Catalysis of a biochemical reaction at physiological temperatures. In biologically catalyzed reactions, the reactants are known as substrates, and the catalysts are naturally occurring macromolecular substances known as enzymes. Enzymes possess specific binding sites for substrates and are usually composed wholly or largely of protein.
DNA binding	Selective interaction with DNA (deoxyribonucleic acid).
Enzyme regulator activity	Modulation of an enzyme.
Metal ion binding	Selective interaction with any metal ion.
Motor activity	Catalysis of movement along a polymeric molecule such as a microfilament or microtubule, coupled to the hydrolysis of a nucleoside triphosphate.
Nucleotide binding	Selective interaction with a nucleotide, which is any compound consisting of a nucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on the ribose or deoxyribose moiety.
Protein binding	Selective interaction with any protein or protein complex (a complex of two or more proteins that may include other nonprotein molecules).
Receptor activity	The mediation by protein or gene products of a signal from the extracellular environment to a intracellular messenger.

Table 17. GO Slim categories for molecular functions (Sheet 2 of 2)

GO Slim molecular function	Description
RNA binding	Selective interaction with an RNA molecule or a portion of it.
Signal transducer activity	Mediation of the transfer of a signal from the outside to the inside of a cell by means other than the introduction of the signal molecule itself into the cell.
Structural molecule activity	The action of a molecule that contributes to the structural integrity of a complex or assembly within or outside a cell.
Transcription regulator activity	Activity that plays a role in regulating transcription; it might bind a promoter or enhancer DNA sequence or interact with a DNA-binding transcription factor.
Translation regulator activity	The initiation, activation, perpetuation, repression, or termination of polypeptide synthesis at the ribosome.
Transporter activity	Activity that enables the directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within, or between cells.

GO Slim Categories for Cellular Components

Table 18 describes the GO Slim categories for cellular components.

Table 18. GO Slim categories for cellular components (Sheet 1 of 4)

GO Slim cellular component	Description
Cell surface	Proteins that are attached to the external part of the cell wall, cell membrane, or both.
Chromosome	A structure composed of a very long molecule of DNA and associated proteins (for example, histones) that carry hereditary information.
Cytoplasm	All of the contents of a cell excluding the plasma membrane and nucleus but including other subcellular structures.

Table 18. GO Slim categories for cellular components (Sheet 2 of 4)

GO Slim cellular component	Description
Cytoskeleton	Any of the various filamentous elements that form the internal framework of cells and that typically remain after treatment of the cells with mild detergent to remove membrane constituents and soluble components of the cytoplasm. The term embraces intermediate filaments, microfilaments, microtubules, the microtrabecular lattice, and other structures characterized by a polymeric filamentous nature and long-range order within the cell. The various elements of the cytoskeleton not only serve in the maintenance of cellular shape but also have roles in other cellular functions, including cellular movement, cell division, endocytosis, and movement of organelles.
Cytosol	That part of the cytoplasm that does not contain membranous or particulate subcellular components.
Endosome	A membrane-bound organelle that carries materials newly ingested by endocytosis. It passes many of the materials to lysosomes for degradation.
Endoplasmatic reticulum	The irregular network of unit membranes, visible only by electron microscopy, that occurs in the cytoplasm of many eukaryotic cells. The membranes form a complex meshwork of tubular channels, which are often expanded into slit-like cavities called cisternae. The endoplasmatic reticulum takes two forms, rough (or granular), with ribosomes adhering to the outer surface, and smooth, with no ribosomes attached.
Extracellular	The space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures, this term refers to the space outside of the plasma membrane. It only applies to proteins that are not attached to the cell surface. It covers the host cell environment outside an intracellular parasite.

Table 18. GO Slim categories for cellular components (Sheet 3 of 4)

GO Slim cellular component	Description
Golgi	A compound membranous cytoplasmic organelle of eukaryotic cells consisting of flattened, ribosome-free vesicles arranged in a more or less regular stack. The Golgi apparatus differs from the endoplasmic reticulum in often having slightly thicker membranes, appearing in sections as a characteristic shallow semicircle so that the convex side (cis or entry face) abuts the endoplasmic reticulum, secretory vesicles emerging from the concave side (trans or exit face). In vertebrate cells, there is usually one such organelle, but in invertebrates and plants, where they are known usually as dictyosomes, there may be several scattered in the cytoplasm. The Golgi apparatus processes proteins produced on the ribosomes of the rough endoplasmic reticulum. Such processing includes modification of the core oligosaccharides of glycoproteins and the sorting and packaging of proteins for transport to a variety of cellular locations.
Membrane	Double layer of lipid molecules that encloses all cells, and, in eukaryotic cells, many organelles. The membrane can be a single or double lipid bilayer. It also includes associated proteins. **Note** This term is not restricted to the plasma membrane but applies to all types of membranes present in the cell, that is, nuclear membranes and mitochondrial membranes.
Mitochondrion	A semiautonomous, self-replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.
Nucleus	A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes and is the site of RNA synthesis and processing. In some species or in specialized cell types, RNA metabolism or DNA replication might be absent.
Spliceosome	A ribonucleoprotein complex containing RNA and small nuclear ribonucleoproteins (snRNPs), which is assembled during the splicing of messenger RNA primary transcript to excise an intron.

Table 18. GO Slim categories for cellular components (Sheet 4 of 4)

GO Slim cellular component	Description
Protein complex	Any protein group composed of two or more subunits, which may or may not be identical. Protein complexes might have other associated non-protein prosthetic groups, such as nucleic acids, metal ions, or carbohydrate groups.
Ribosome	An intracellular organelle, about 200 Angstroms in diameter, consisting of RNA and protein. It is the site of protein biosynthesis resulting from translation of messenger RNA (mRNA).
Vacuole	A closed structure found only in eukaryotic cells, completely surrounded by unit membrane and containing liquid material. Cells contain one or several vacuoles that might have different functions from each other. Vacuoles have a diverse array of functions. They can act as a storage organelle for nutrients or waste products, as a degradative compartment, as a cost-effective way of increasing cell size, and as a homeostatic regulator controlling both the turgor pressure and the pH of the cytosol.
Organelle lumen	The volume enclosed by the membranes of a particular organelle, for example, endoplasmic reticulum lumen or the space between the two lipid bilayers of a double membrane surrounding an organelle (for example, nuclear membrane lumen).

GO Slim Categories for Biological Processes

Table 19 describes the GO Slim categories for biological processes.

Table 19. GO Slim categories for biological processes (Sheet 1 of 3)

Go Slim biological process	Description
Cell communication	Any process that mediates interactions between a cell and its surroundings. Cell communication encompasses interactions such as signaling or attachment between one cell and another cell, between a cell and an extracellular matrix, or between a cell and any other aspect of its environment.
Cell death	The specific activation or halting of processes within a cell so that its vital functions markedly cease, rather than simply deteriorating gradually over time, which culminates in cell death.

Table 19. GO Slim categories for biological processes (Sheet 2 of 3)

Go Slim biological process	Description
Cell differentiation	The process in which relatively unspecialized cells—for example, embryonic or regenerative cells—acquire specialized structural features, functional features, or both that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history. Differentiation includes the processes involved in commitment of a cell to a specific fate.
Cell division	The processes resulting in the physical partitioning and separation of a cell into daughter cells.
Cell growth	The process by which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present.
Cell homeostasis	The processes involved in the maintenance of an internal equilibrium at the level of the cell.
Cell motility	Any process involved in the controlled movement of a cell.
Cell organization and biogenesis	A process that is carried out at the cellular level and that results in the formation, arrangement of constituent parts, or disassembly of a cellular component. The process includes the plasma membrane and any external encapsulating structures, such as the cell wall and cell envelope.
Cell proliferation	The multiplication or reproduction of cells, resulting in the rapid expansion of a cell population.
Coagulation	The process by which a fluid solution, or part of it, changes into a solid or semisolid mass.
Conjugation	The union or introduction of genetic information from compatible mating types that results in a genetically different individual. Conjugation requires direct cellular contact between the organisms.
Defense response	Reactions triggered in response to the presence of a foreign body or the occurrence of an injury, which result in restriction of damage to the organism attacked or prevention and recovery from the infection caused by the attack.

Table 19. GO Slim categories for biological processes (Sheet 3 of 3)

Go Slim biological process	Description
Development	The biological process whose specific outcome is the progression of an organism over time from an initial condition (for example, a zygote or a young adult) to a later condition (for example, a multicellular animal or an aged adult).
Metabolic process	Processes that cause many of the chemical changes in living organisms, including anabolism and catabolism. Metabolic processes typically transform small molecules but also include macromolecular processes such as DNA repair and replication, and protein synthesis and degradation.
Regulation of biological process	Any process that modulates the frequency, rate, or extent of a biological process. Biological processes are regulated by many means, for example, control of gene expression, protein modification, or interaction with a protein or substrate molecule.
Reproduction	The production by an organism of new individuals that contain some portion of their genetic material inherited from that organism.
Response to stimulus	A change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, and so forth) as a result of a stimulus.
Transport	The directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within, or between cells.

Quantification

This chapter describes how to perform precursor-, reporter-, and peak area-based quantification in the Proteome Discoverer application.

Contents

Contents - continued

Activating the Quantification Menu

In the Proteome Discoverer application, the Quantification menu becomes available when you open an MSF file generated by a workflow in the Workflow Editor that includes the Reporter Ions Quantifier node, the Precursor Ions Quantifier node, or the Precursor Ions Area Detector node.

❖ **To activate the Quantification menu**

- Choose **File > Open Report** and follow the procedure in the Help to open an MSF file containing quantification results.

  The commands on the Quantification menu become available.

  If you do not have an MSF file containing quantification results, see "Performing Precursor Ion Quantification," "Performing Reporter Ion Quantification" on page 249, or "Performing Peak Area Calculation Quantification" on page 259 for instructions on creating one.

Proteins Included in the Quantification

To determine the proteins to include in the quantification, the Proteome Discoverer application first creates protein groups from the identified PSMs. When the search results include quantification data, it then performs quantification on all protein groups.

The application calculates the ratio for each of the defined quantification ratios for the protein group as the median of all PSMs belonging to the protein group that are marked as being usable. Whether the application considers a PSM usable is determined by the settings of the Quantification Method Editor dialog box, including two options on the Protein Quantification page, Use Only Unique Peptides and Consider Proteins Groups for Peptide Uniqueness. The Use Only Unique Peptides option includes in the quantification peptides that do not occur in other proteins. The Proteins Groups for Peptide Uniqueness option defines peptide uniqueness on the basis of protein groups rather than individual proteins.

When it determines peptide uniqueness for classification in the PSM Ambiguity column on the Peptides page, the application only considers the PSMs that it considered when creating the protein groups, if you select the Use Only Unique Peptides option. For example, it does not use for quantification a PSM of low confidence that it did not use to create the protein groups.

Performing Precursor Ion Quantification

In precursor ion quantification, also called isotopically labeled quantification, **p**rotein abundance is determined from the relative MS signal intensities of an isotopically labeled sample and an unlabeled control sample. Stable-isotope labeling by amino acid in cell culture (SILAC) is a proteomics identification and quantification technique that uses in-vivo metabolic labeling to detect differences in the abundance of proteins in multiple samples. It is a type of isotopically labeled quantification, which uses stable (nonradioactive) heavy isotopes as labels. You can also introduce the stable isotopes by chemical labeling at the protein or peptide level with the isotopomeric tags (for example, dimethyl labeling).

The following default quantification methods are available for precursor ion (isotopically labeled) quantification:

- SILAC 2plex (Arg10, Lys6): Uses arginine 10 and lysine 6.
- SILAC 2plex (Arg10, Lys8): Uses arginine 10 and lysine 8.
- SILAC 2plex (Ile6): Uses isoleucine 6.
- SILAC 3plex (Arg6, Lys4|Arg10, Lys8): Uses arginine 10 and lysine 8 for “heavy” labels and arginine 6 and lysine 4 for “medium” labels.
- SILAC 3plex (Arg6, Lys6|Arg10, Lys8): Uses arginine 10 and lysine 8 for “heavy” labels and arginine 6 and lysine 6 for “medium” labels.
- Dimethylation 3plex: Chemically adds isotopically labeled dimethyl groups to the N-terminus and to the ε-amino group of lysine.
- ^{18}O labeling: Introduces 2 or 4 Da mass tags through the enzyme-catalyzed exchange reaction of C-terminal oxygen atoms with ^{18}O.

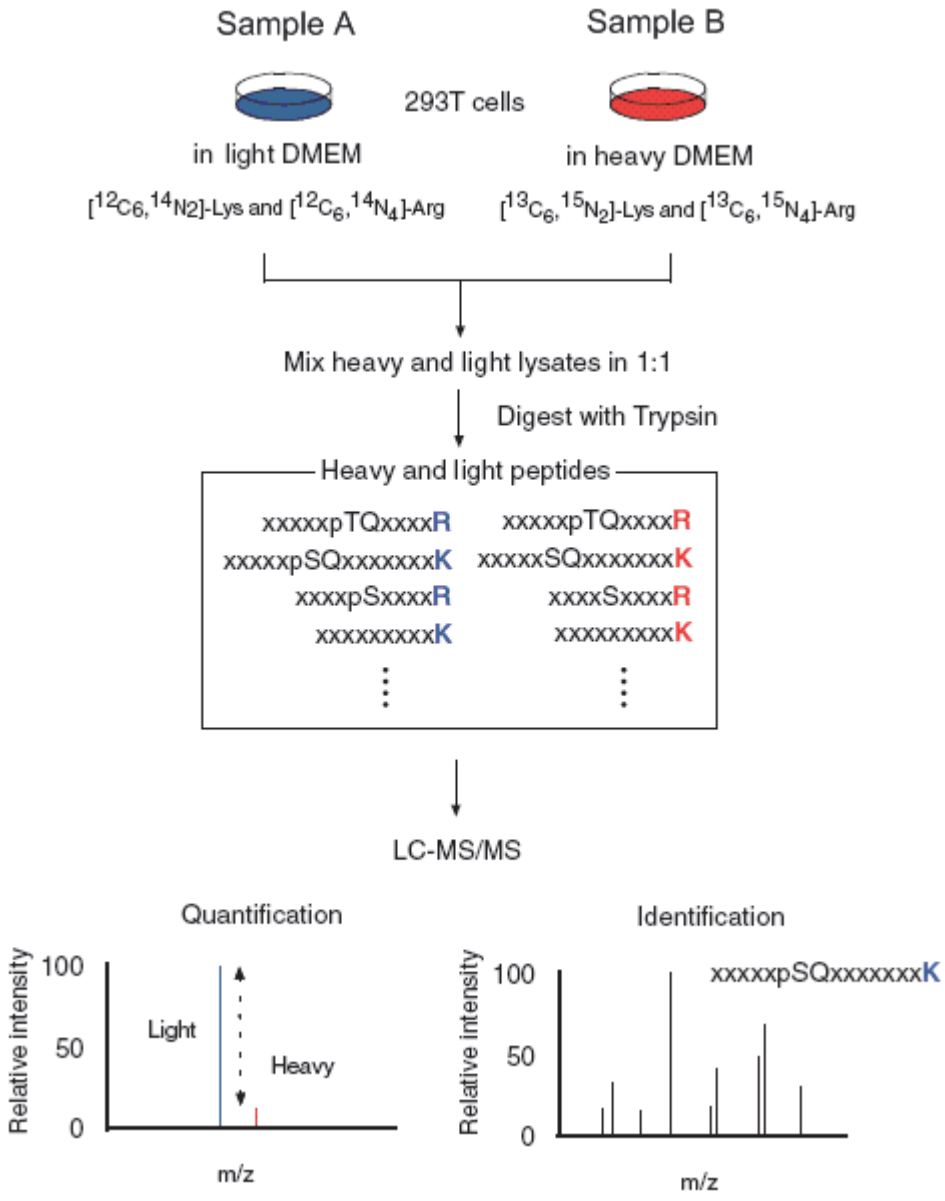
SILAC 2plex Methods

In a typical SILAC quantification experiment, two cell populations grow in media that are deficient in lysine and arginine. One population grows in a medium containing normal (“light”) amino acids, such as lysine ($^{12}C_6{}^{14}N_2$). The other population grows in a medium containing amino acids where stable heavy isotopes, such as lysine 6 ($^{13}C_6{}^{14}N_2$) or lysine 8 ($^{13}C_6{}^{15}N_2$), have been substituted for normal atoms. SILAC quantification usually uses

"heavy" arginine and lysine, because these are the cleavage sites for the generally used trypsin protease. Both populations incorporate these amino acids into proteins through natural cellular protein synthesis. The cells growing in the medium with the heavy isotopes incorporate these isotopes into all of their proteins.

After altering the proteome in one sample through chemical treatment or genetic manipulation, you then combine equal amounts of protein from both cell populations and digest with trypsin before MS analysis. Because peptides labeled with "heavy" and "light" amino acids are chemically identical, they co-elute during reverse-phase chromatographic separation. This means they are detected simultaneously during MS analysis. To determine the average change in protein abundance in the treated sample, you use the relative peak intensities of multiple isotopically distinct peptides from each protein, as shown in Figure 171.

Figure 171. Schematic workflow for SILAC-based peptide and protein quantification



SILAC can differentiate peptides in single MS mode without requiring you to perform tandem mass spectrometry. However, SILAC cannot identify peptides, so you must use tandem mass spectrometry for that purpose.

You can use several SILAC 2plex methods, for example (Arg10, Lys6) and (Arg10, Lys8), to compare two samples.

SILAC 3plex Methods

SILAC 3plex methods are similar to SILAC 2plex methods except, in addition to a "heavy" sample (containing, for example, Arg10 and Lys8), they also use a "medium" sample (containing, for example, Arg6 and Lys4). Protein abundance is determined from the relative MS signal intensities of the heavy sample, medium sample, and a control sample containing "light" (^{12}C and ^{14}N) arginine and lysine.

Dimethylation 3plex Method

The Proteome Discoverer application also includes the dimethylation 3plex method. It is not metabolomic labeling in cell culture but is a form of peptide chemical labeling. This method uses formaldehyde and sodium cyanoborohydride to add dimethyl groups $(CH_3)_2$ to the N-terminus and to the ε-amino group of lysine. By choosing the isotopomers of formaldehyde and sodium cyanoborohydride, you can create light, medium, and heavy labels. For the light label, the (natural-isotope) dimethyl group is $^{12}C_2{}^1H_6$. For the medium label, the dimethyl group is $^{12}C_2{}^2H_4{}^1H_2$, which is 4 Da more massive. For the heavy label, the dimethyl group is $^{13}C_2{}^2H_6$, which is an additional 4 Da more massive.

You can use the dimethylation 3plex method to compare up to three samples. You cannot apply labels to the C terminus, nor to arginine.

^{18}O Labeling Method

The ^{18}O labeling method introduces 2 or 4 Da mass labels through the enzyme-catalyzed exchange reaction of C-terminal oxygen atoms with ^{18}O.

Creating a Workflow for Precursor Ion Quantification

To use a precursor ion quantification method, you must use a workflow that includes the Precursor Ions Quantifier node.

❖ To create a workflow for precursor ion quantification

Note This procedure uses a SILAC 2plex example.

1. Choose **Workflow Editor > New Workflow**.

   For instructions on creating a workflow with the Workflow Editor, see "Starting a New Search by Using the Workflow Editor" on page 42.

2. In the Workflow Editor, drag the **Spectrum Files** node to the workspace.

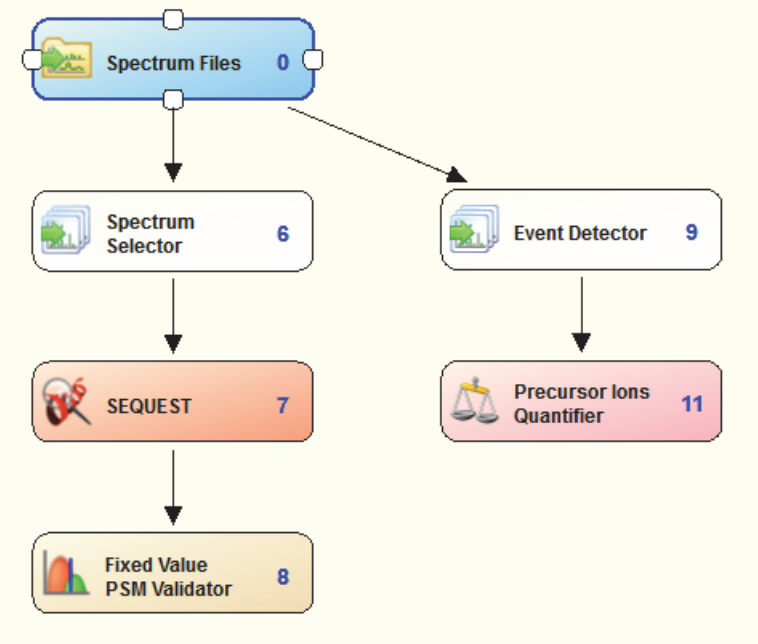
3. If you selected the Spectrum Files node as your input, do the following:
   a. Drag the **Spectrum Selector** node and the **Event Detector** node to the workspace.
   b. Connect the Spectrum Selector node and the Event Detector node to the Spectrum Files node.
4. Drag the **Precursor Ions Quantifier** node to the workspace pane and attach it directly to the Event Detector node.

   The Precursor Ions Quantifier node performs quantification for isotopically labeled amino acids.

   Note You cannot use the Precursor Ions Quantifier node and the Precursor Ions Area Detector node in the same workflow. You cannot use the Reporter Ions Quantifier node in the same workflow with either of these two nodes.
5. Drag the appropriate search engine node (for example, **SEQUEST**) to the workspace pane and attach it to the Spectrum Selector node.
6. Drag the **Fixed Value PSM Validator** or the **Percolator** node to the workspace pane and attach it to the search engine node.

   Figure 172 illustrates the workflow up to this point.

   Figure 172. Beginning of the workflow for precursor ion quantification



7. Add any other nodes that you want and connect them. For general information about creating a workflow in the Workflow Editor, see "Starting a New Search by Using the Workflow Editor" on page 42.
8. In the Parameters pane of the Workflow Editor, click **Show Advanced Parameters**.

9. Click the **Spectrum Files** node and specify the raw file(s) in the Parameters pane.
10. Click the **Event Detector** node and set the parameters for it in the Parameters pane:
    a. In the Mass Precision box, specify the expected standard deviation of the mass precision.

       Three times the standard deviation is used to create extracted ion chromatograms. The minimum value is 1 ppm. The maximum value is 4 ppm. The default is 2 ppm.

    b. In the S/N Threshold box, specify a threshold signal-to-noise value that determines whether the Proteome Discoverer application removes peaks from the spectrum. It removes peaks with a signal-to-noise value below this threshold.

       The minimum value is 1.0, and there is no maximum value. The default is 1.
11. Click the **Spectrum Selector** node, and set the parameters for it in the Parameters pane:
    a. Change the setting in the Max. Precursor Mass box to an appropriate setting. For example, for SILAC 2plex (Arg10, Lys6) quantification, set this option to **6500**.
    b. Change the setting in the S/N Threshold box to an appropriate setting. For example, for SILAC 2plex (Arg10, Lys6) quantification, set this option to **1.5**.

    For other parameters that you can optionally set for the Spectrum Selector node, refer to the Help.
12. Click the search engine node (for example, **SEQUEST**), and set the parameters for it in the Parameters pane:
    a. In the Protein Database box, select the **FASTA** database.
    b. In the Dynamic Modifications area, select the dynamic modifications.

       For example, for SILAC 2plex (Arg10, Lys6) quantification, you might select the following two dynamic modifications:

       - **13C(6)/ +6.020 Da (K)**
       - **13C(6)/15N(4)/+10.008 Da (R)**

       If you do not find these labels, you can enable them by following the instructions in "Updating Chemical Modifications" on page 141.
    c. In the Static Modifications area, select the static modifications. For example, for SILAC 2plex (Arg10, Lys6) quantification, select **Carbamidomethyl/ +57.021 Da (C)** in the Static Modification box.
    d. Set any other parameters that you prefer.
13. Set the parameters for all other nodes in the Parameters pane. For information about all the parameters that you can set for each node, refer to the Help. For information on the parameters that you can set for the Precursor Ions Quantifier node, see step 14 of this procedure.

14. Click the **Precursor Ions Quantifier** node and set the parameters for it in the Parameters pane:
    a. Set up the quantification method. Click the **Quantification Method** parameter, and follow the procedure in "Setting Up the Quantification Method" on page 264 to specify the quantification method.
    b. Set the parameters that identify the isotope patterns:
        i. In the RT Tolerance of Isotope Pattern Multiplet [min] box, specify the maximum retention-time tolerance of the A0 peak in the isotope pattern of a quantification multiplet, in minutes. The default is 0.2 minutes.
        ii. In the Single-Peak/Missing Channels Allowed box, specify the maximum number of single-peak or missing quantification channels that are allowed for a valid peptide quantification result. A single-peak quantification channel is a channel that is identified with just a single peak. The maximum number used will not exceed the number of specified channels. The minimum value is 0 (the default). This value indicates that there are at least two peaks in the quantification channel used for quantification.
15. Choose **Workflow Editor > Start Workflow** or click the **Start Workflow** icon,  .

Performing Reporter Ion Quantification

In contrast to the metabolic labeling used by isotopically labeled precursor ion quantification methods such as SILAC, isobarically labeled reporter ion quantification methods use external reagents, or tags, to enzymatically or chemically label proteins and peptides. Reporter ion quantification uses tags that have the same mass. (A reporter ion is a fragment ion with a tag.) The Proteome Discoverer application supports reporter ion quantification for Tandem Mass Tag (TMT) and Isobaric Tag for Relative and Absolute Quantification (iTRAQ) and any user-defined tags. Identification and quantification with both TMT and iTRAQ are performed in the MS/MS scan.

You can quantify all isobarically labeled samples. For iTRAQ, 4plex and 8plex default methods are available. For TMT, 2plex and 6plex default methods are available. You can also add new methods.

TMT Quantification

TMT quantification is a reproducible, highly accurate quantification method that provides both comparative and absolute MS/MS-based quantification of proteins and peptides in biological samples. TMT tagging produces data to calculate the relative abundances of proteins. You can evaluate differential protein expression in one to six samples in a single experiment.

Each sample is labeled with chemically identical tags before mixing the samples, and a single MS run generates a single peak for each peptide, irrespective of which tag it has been given. Between the normalizer and reporter is a cleavable linker, which breaks during MS/MS. The mass reporter ion is split off and measured by the mass spectrometer.

Only MS/MS fragmentation can differentiate the tagged proteins. The reporter ion, measured by the mass spectrometer, generates a different peak. As a result, the peak height/peak integral for each reporter denotes the relative amount of protein originating from each of the labeled samples.

With the quantification functions in the application, you can set filters to see only unique peptides so that every protein associated with the same peptide is not counted, producing a best-results list of peptides. Filtering the number of proteins can give you a more robust final analysis of your experimental set.

Quantification with TMT tags is no different from quantification with iTRAQ (described in "iTRAQ Quantification" on page 252), except that it uses the following default mass tags by Proteome Sciences PLC:

- TMT 2plex
- TMT 6plex
- iodo TMT 6plex
- TMTe 6plex
- TMT 10plex

Note If you are installing the Proteome Discoverer application for the first time, the TMT 6plex quantification method is no longer available. The TMTe 6plex method replaces it.

You can use these default methods to create your own quantification templates. For information on adding quantification methods, see "Changing a Quantification Method" on page 288.

Table 20 lists the masses of the reporter ions of the tags available in the different TMT kits. The masses for the original TMT reagents, which but no longer available, are included for reference.

Table 20. Monoisotopic masses of the reporter ions after CID or HCD fragmentation of the tags in the different TMT kits (Sheet 1 of 2)

TMT 2plex		TMT 6plex (Original)		TMTe 6plex (Current)		TMT 10plex		iodo TMT 6plex	
Tag	Mass	Tag	Mass	Tag	Mass	Tag	Mass	Tag	Mass
126	126.127725	126	1216.127725	126	126.127725	126	126.127725	126	126.127725
127	127.131079	127	127.131079	127	127.124760	127_N 127_C	127.124760 127.131079	127	127.124760

Table 20. Monoisotopic masses of the reporter ions after CID or HCD fragmentation of the tags in the different TMT kits (Sheet 2 of 2)

TMT 2plex		TMT 6plex (Original)		TMTe 6plex (Current)		TMT 10plex		iodo TMT 6plex	
Tag	**Mass**	**Tag**	**Mass**	**Tag**	**Mass**	**Tag**	**Mass**	**Tag**	**Mass**
		128	128.134433	128	128.134433	128_N 128_C	128.128115 128.134433	128	128.134433
		129	129.13779	129	129.131468	129_N 129_C	129.131468 129.13779	129	129.1311468
		130	130.141141	130	130.141141	130_N 130_C	130.134825 130.141141	130	130.141141
		131	131.138176	131	131.138176	131	131.138176	131	131.138176

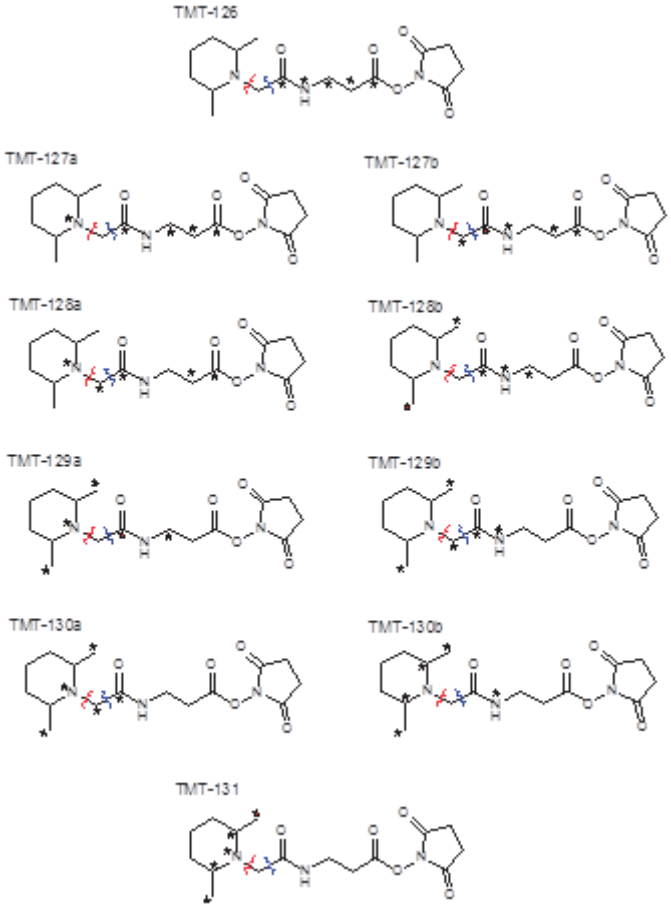
The iodo TMT 6plex includes cysteine reactive TMT reagents.

The TMT 10plex leverages the high resolution of recent mass spectrometers to routinely differentiate the ^{13}C isotopes from the ^{15}N isotopes[1, 2]. For the 127, 128, 129, and 130 tags, the TMT 10plex contains two reagents, the ^{13}C and the ^{15}N reagent. For the monoisotopic masses of the different reporter ions after CID or HCD fragmentation, see Table 20. Figure 173 shows the position of the ^{13}C and ^{15}N atoms in the different reagents. In this illustration, the stars indicate the positions of the ^{13}C and the ^{15}N substitutions, the red lines indicate the position of the ETD fragmentation sites, and the blue lines indicate the position of the CID fragmentation sites.

[1] McAlister G. C., Huttlin E. L., Haas W., Ting L., Jedrychowski M. P., Rogers J. C., Kuhn K., Pike I., Grothe R. A., Blethrow J. D., and Blethrow G. S. P., "Increasing the Multiplexing Capacity of TMTs Using Reporter Ion Isotopologues with Isobaric Masses," *Analytical Chemistry*, **2012**, *Volume 84*: 7469–7478.

[2] Werner T., Becher I. Sweetman G., Doce C., Savitski M. M., and Savitski B. M., "High-Resolution Enabled TMT 8-plexing," *Analytical Chemistry*, 2012, *Volume 84*: 7188–7194.

Figure 173. Structures of the TNT reagents contained in the TMT 10plex quantification method



Recent research concludes that avoiding the application of any correction for isotopic impurities improves quantification results for the TMTe 6plex, TMT 10plex, and iodo TMT 6plex kits, so the default methods for these kits turn off the purity correction.

iTRAQ Quantification

iTRAQ is a protein quantification technique that uses isobaric amine-specific, stable isotope reagents to label all peptides in up to eight different samples simultaneously. The labeled peptides from each sample are combined, separated by two-dimensional liquid chromatography, and analyzed with tandem mass spectrometry (MS/MS). The same peptide from each sample appears as a single peak in the MS spectrum. In single MS mode, the differentially labeled versions of a peptide are indistinguishable. In tandem MS mode, which isolates and fragments peptides, each tag generates a unique reporter ion. Protein quantification compares the peak intensity of the reporter ions in the MS/MS spectra to assess the relative abundance of the peptides and therefore the proteins that they are derived from.

iTRAQ includes two default mass tags available from Applied Biosystems (ABI) that you can use to label all peptides:

- iTRAQ 4plex, which is standard
- iTRAQ 8plex

The Proteome Discoverer application includes default quantification methods for processing data from iTRAQ 4plex- and iTRAQ 8plex-labeled samples. You can use these methods to create your own workflow templates. For information on adding quantification methods, see "Changing a Quantification Method" on page 288.

iTRAQ quantification works exactly the same as TMT quantification, except that TMT quantification offers 2plex, 6plex, and 10plex quantification methods, and iTRAQ offers 4plex and 8plex quantification methods.

Creating a Workflow for Reporter Ion Quantification

To use an isobarically labeled reporter ion quantification method, you must open an MSF file generated from a workflow that includes the Reporter Ions Quantifier node.

Setting up the workflow for TMT and iTRAQ quantification is basically the same.

❖ To create a workflow for reporter ion quantification

1. Choose **Workflow Editor > New Workflow**.

   For instructions on creating a workflow with the Workflow Editor, see "Starting a New Search by Using the Workflow Editor" on page 42.

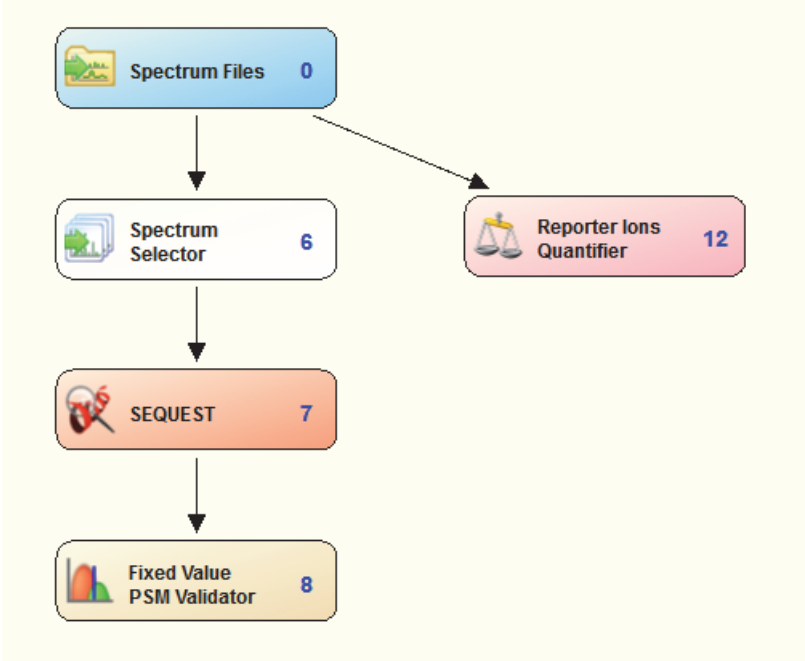
2. In the Workflow Editor, drag the **Spectrum Files** node to the workspace.
3. If you selected the Spectrum Files node as your input, drag the **Spectrum Selector** node to the workspace, and attach the Spectrum Files node to the Spectrum Selector node.
4. Drag the **Reporter Ions Quantifier** node to the workspace pane, and attach the Spectrum Files node to the Reporter Ions Quantifier node.

   Note You cannot use the Reporter Ions Quantifier node in the same workflow with either the Precursor Ions Quantifier node or the Precursor Ions Area Detector node.

5. Drag the search engine node that you want (for example, **SEQUEST**) to the workspace pane, and attach the Spectrum Selector node to the search engine node.
6. Drag the **Fixed Value PSM Validator** or the **Percolator** node to the workspace pane and attach it to the search engine node.

   Figure 174 illustrates the workflow up to this point.

Figure 174. Beginning of the workflow for reporter ion quantification



7. Add any other nodes that you would like and connect them.

   For general information about creating a workflow in the Workflow Editor, see "Starting a New Search by Using the Workflow Editor" on page 42.

8. Click the **Spectrum Files** node and specify the raw file in the Parameters pane.
9. Click the **Spectrum Selector** node, and set the parameters for it in the Parameters pane:
   a. Change the setting in the Total Intensity Threshold box to an appropriate setting.

      For example, for TMTe 6plex quantification, you could set this option to **20 000**.

   b. Change the setting in the Minimum Peak Count box to an appropriate setting.

      For example, for TMTe 6plex quantification, you could set this option to **200**.

   For other parameters that you can optionally set for the Spectrum Selector node, refer to the Help.
10. Click the search engine node (for example, **SEQUEST**), and set the parameters for it in the Parameters pane:
    a. In the Protein Database box, select the **FASTA** database.
    b. In the Dynamic Modifications area, select the dynamic modifications.

       Use the following modifications for a Sequest HT search:

       - TMT 2plex (seldom used):
         – TMT 2plex for lysine and N-terminal (you can use these as static or dynamic modifications)

  - Dynamic TMT 2plex for threonine
- TMTe 6plex or TMT 6plex:
  - TMT 6plex for lysine and N-terminal (you can use these as static or dynamic modifications)
  - Dynamic TMT 6plex for threonine
- TMT 10plex: the same modifications as for TMT 6plex
- iodo TMT 6plex: iodo TMT 6plex for cysteine (you can use these as static or dynamic modifications)

For example, for TMTe 6plex quantification, you would select a dynamic modification of **TMT6plex / +229.163 Da (K)**. If you do not find this label, you can enable it by following the instructions in "Updating Chemical Modifications" on page 141.

c. In the Static Modifications area, select the static modifications. For example, for TMTe 6plex quantification, you would select **TMT6plex / +229.163 Da (K)** in the Peptide N-Terminus box.

d. Set any other parameters that you prefer.

11. Set the parameters for all other nodes in the Parameters pane.

    For information about all the parameters that you can set for each node, refer to the Help. For information on the parameters that you can set for the Reporter Ions Quantifier node, see step 12 of this procedure.

12. Click the **Reporter Ions Quantifier** node and set the parameters for it in the Parameters pane:

    a. Set up the quantification method. Click the **Quantification Method** parameter, and follow the procedure in "Setting Up the Quantification Method" on page 264 to specify the quantification method.

    b. Set the parameters that specify the peak integration:

       i. In the Integration Tolerance box, specify the mass-to-charge (*m/z*) window that enables you to look for the reporter peaks. The default is 20 ppm.

       ii. In the Integration Method box, select which peak to choose when more than one peak is found inside the integration window.

– (Default) Most Confident Centroid: Lays a Gaussian curve around the target peak (the tag mass) with a sigma value equal to the mass accuracy or integration window. Then the Gaussian curve normalizes all peaks in the window, and the largest is considered to be the most confident peak. This method is also used by the Spectrum Selector node in the Workflow Editor to pick the monoisotopic peak from the survey scan. The only difference is that the Spectrum Selector uses a 3-sigma interval, but Most Confident Centroid uses only a 1-sigma interval. This means the Most Confident Centroid is almost always the largest peak inside the integration window because of the small inclination of the Gaussian curve in the 1-sigma interval.

– Most Intense Centroid: Selects the highest peak.

– Centroid With Smallest Delta Mass: Selects the peak with the smallest deviation from the theoretical mass.

– Centroid Sum: Sums the intensity of all the peaks in the window.

c. Specify the scan event filters:

i. In the Mass Analyzer box, select the type of mass spectrometer used in the acquisition of the spectrum:

– Ion Trap (ITMS)

– (Default) Fourier Transform (FTMS)

– Time of Flight (TOFMS)

– Single Quad (SQMS)

– Triple Quad (TQMS)

– Sector Field (SectorMS)

ii. In the MS Order box, specify the level of tandem mass spectrum to be processed, for example, **MS2** or **MS3**. The default is MS2.

iii. In the Activation Type list, specify the fragmentation method used to activate the scan.

Note You cannot perform TMT quantification on both PQD and HCD scans. You can choose only one activation type.

– CID (Collision-Induced Dissociation)

– ECD (Electron Capture Dissociation)

– ETD (Electron Transfer Dissociation)

– (Default) HCD (High-Energy Collision Dissociation)

– MPD (Multi-Photon Dissociation)

– PQD (Pulsed Q Collision-Induced Dissociation)

For a description of these fragmentation types, see "Fragmentation Methods" on page 8.

13. Choose **Workflow Editor > Start Workflow**, or click the **Start Workflow** icon,  .

Performing TMT Quantification on HCD and CID Scans

If a raw file contains both CID scans for identification and HCD scans for quantification, you can use the following workflow to both quantify the HCD scans and identify peptides in the CID scans, the HCD scans, or both.

❖ To perform TMT Quantification on HCD and CID scans

1. Drag the **Reporter Ions Quantifier** node to the workspace pane and connect it to the workflow.
2. Set the Activation Type parameter for the Reporter Ions Quantifier node to **HCD**.
3. Set the Activation Type parameter for the Spectrum Selector node to **Any**, **Is CID**, **HCD**, or **Is CID**, depending on your method setup and identification strategy.
4. Set all other parameters—modifications, tolerances, FASTA files, and so forth—and choose **Workflow Editor > Start Workflow**, or click the **Start Workflow** icon,  .

Demonstrating How to Create a Workflow for Reporter Ion Quantification

The following demonstration shows you how to set up a workflow for reporter ion quantification and how to specify the quantification method.

❖ Click the button below to view the demonstration.

Performing Peak Area Calculation Quantification

If you want to determine the area for any quantified peptide, you can use peak area calculation quantification. You might want to use this quantification method to obtain an idea of the relative quantities of all peptides in a sample.

If the Proteome Discoverer application calculates peptide areas during processing, it uses them to automatically calculate protein areas for the proteins in the MSF report. It calculates the area of any given protein as the average of the three most abundant distinct peptides identified for the protein.[3] The peptides must have different sequences to be considered distinct. Peptides with different charge states or modification variants of the same sequence are considered the same peptide. If you apply result filters, the application recalculates the protein areas.

❖ To create a workflow for peak area calculation quantification

1. In the Workflow Editor, set up a quantification workflow.

   For instructions on creating a workflow with the Workflow Editor, see "Starting a New Search by Using the Workflow Editor" on page 42.

2. Choose **Workflow Editor > New Workflow**.
3. In the Workflow Editor, drag the **Spectrum Files** node to the workspace.
4. Drag the **Spectrum Selector** node and the **Event Detector** node to the workspace.
5. Connect the Spectrum Selector node and the Event Detector node to the Spectrum Files node.
6. Drag the **Precursor Ions Area Detector** node to the workspace pane and attach it directly to the Event Detector node.

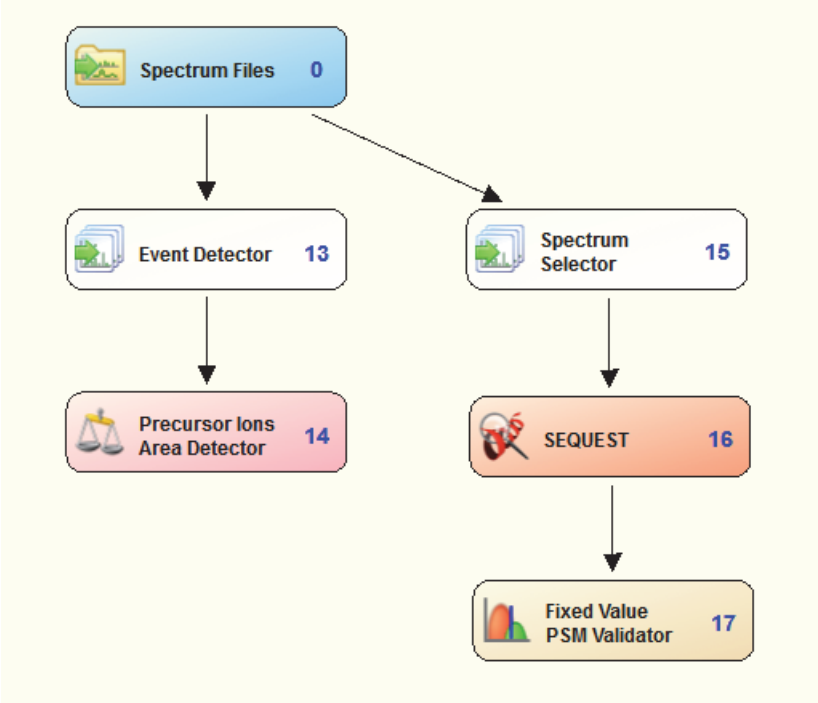
   Note You cannot use the Precursor Ions Area Detector node in the same workflow with the Precursor Ions Quantifier node or the Reporter Ions Quantifier node.

7. Drag the search engine node that you prefer (for example, **SEQUEST**) to the workspace pane and attach it to the Spectrum Selector node.
8. Drag the **Fixed Value PSM Validator** node or the **Percolator** node to the workspace pane and attach it to the search engine node.

   Figure 175 illustrates the workflow up to this point.

[3] Silva, J.C.; Gorenstein, M.V.; Li, G.-Z.; Vissers, J.P. C.; and Geromanos, S.J. Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel MS Acquisition. *Molecular & Cellular Proteomics*, **2006**, *5*, 144-156

Figure 175. Beginning of the workflow for area calculation quantification



9. Add any other nodes that you would like and connect them.

   For general information about creating a workflow in the Workflow Editor, see "Starting a New Search by Using the Workflow Editor" on page 42.

10. Click the **Spectrum Files** node and specify the raw file in the Parameters pane.
11. Click the **Event Detector** node and set the parameters for it in the Parameters pane:
    a. In the Mass Precision box, specify the expected standard deviation of the mass precision.

       To create extracted ion chromatograms, use three times the standard deviation. The minimum value is 1 ppm. The maximum value is 4 ppm. The default is 2 ppm.

    b. In the S/N Threshold box, specify a threshold signal-to-noise value that determines whether the Proteome Discoverer application removes peaks from the spectrum.

       The application removes peaks with a signal-to-noise value below this threshold. The minimum value is 0.0, and there is no maximum value. The default is 1.

12. Click the **Spectrum Selector** node, and set the parameters for it in the Parameters pane:
    a. Change the setting in the Max. Precursor Mass box to an appropriate setting. For example, you could set this option to **6500**.
    b. Change the setting in the S/N Threshold box to an appropriate setting. For example, you could set this option to **1.5**.

For other parameters that you can optionally set for the Spectrum Selector node, refer to the Help.

13. Click the search engine node (for example, **SEQUEST**), and set the parameters for it in the Parameters pane:
    a. In the Protein Database box, select an appropriate **FASTA** database.
    b. In the Dynamic Modifications area, select the dynamic modifications.

       For example, you might select the Oxidation+15.995 Da (M) dynamic modification. If you do not find this label, you can enable it by following the instructions in "Updating Chemical Modifications" on page 141.
    c. In the Static Modifications area, select the static modifications.

       For example, you might select Carbamidomethyl / +57.021 Da (C) in the Static Modification box.
    d. Set any other parameters as needed.
14. Set the parameters for all other nodes in the Parameters pane.

    For information about all the parameters that you can set for each node, refer to the Help. For information on the parameters that you can set for the Precursor Ions Quantifier node, see step 14 of "Creating a Workflow for Precursor Ion Quantification."
15. Choose **Workflow Editor > Start Workflow**, or click the **Start Workflow** icon,  .

Searching for Quantification Modifications with Mascot

When you use the Mascot node on the Mascot server as the search engine in a quantification workflow, you can set the dynamic and static modifications as parameters. For samples with isotopic labels and several PTMs, you might need to specify several dynamic modifications usable within a single search, but the current number that you can specify is limited to nine.

To avoid this limitation, you can configure quantification methods on the Mascot server. In a quantification method, modifications are organized into groups classified as fixed, variable, or exclusive. You can define modification groups as variable or exclusive at the component level, where they usually characterize the component. You can also define them at the method level, but only as fixed or variable. Defining modifications at the method level is convenient for modifications that are important to the method and saves having to choose them in the Workflow Editor. Exclusive groups are effectively a choice of fixed modifications, so the restrictions that apply to fixed modifications also apply to them.

With the Mascot node, you can use the modification groups specified as part of a quantification method on the Mascot server. You can use the node's From Quan Method parameter in the Parameters pane to select the dynamic modifications to search for rather than manually specifying each modification with a Dynamic Modifications parameter.

In the editor in the Mascot server window, you can specify that these groups be variable, fixed, or exclusive. You can also define them directly for the method in report ion quantification or for each component in precursor ion quantification.

❖ To specify the quantification modifications to search for

1. Choose **Administration > Configuration > Mascot**, and configure the Mascot search engine by following the instructions in "Configuring the Mascot Search Engine" on page 25. Be sure that in the Mascot Server URL box, you enter the URL of the Mascot server to be used for Mascot searches.
2. Set up a workflow that includes, at a minimum, the nodes shown in Figure 172 on page 247 for precursor ion quantification, Figure 174 on page 254 for reporter ion quantification, or Figure 175 on page 260 for Precursor Ions Area Detector quantification.
3. Click the **Mascot** node.
4. Select the dynamic modifications to search for:
   - Select a dynamic modification from the list in each Dynamic Modification parameter.

     You can select up to nine modifications.

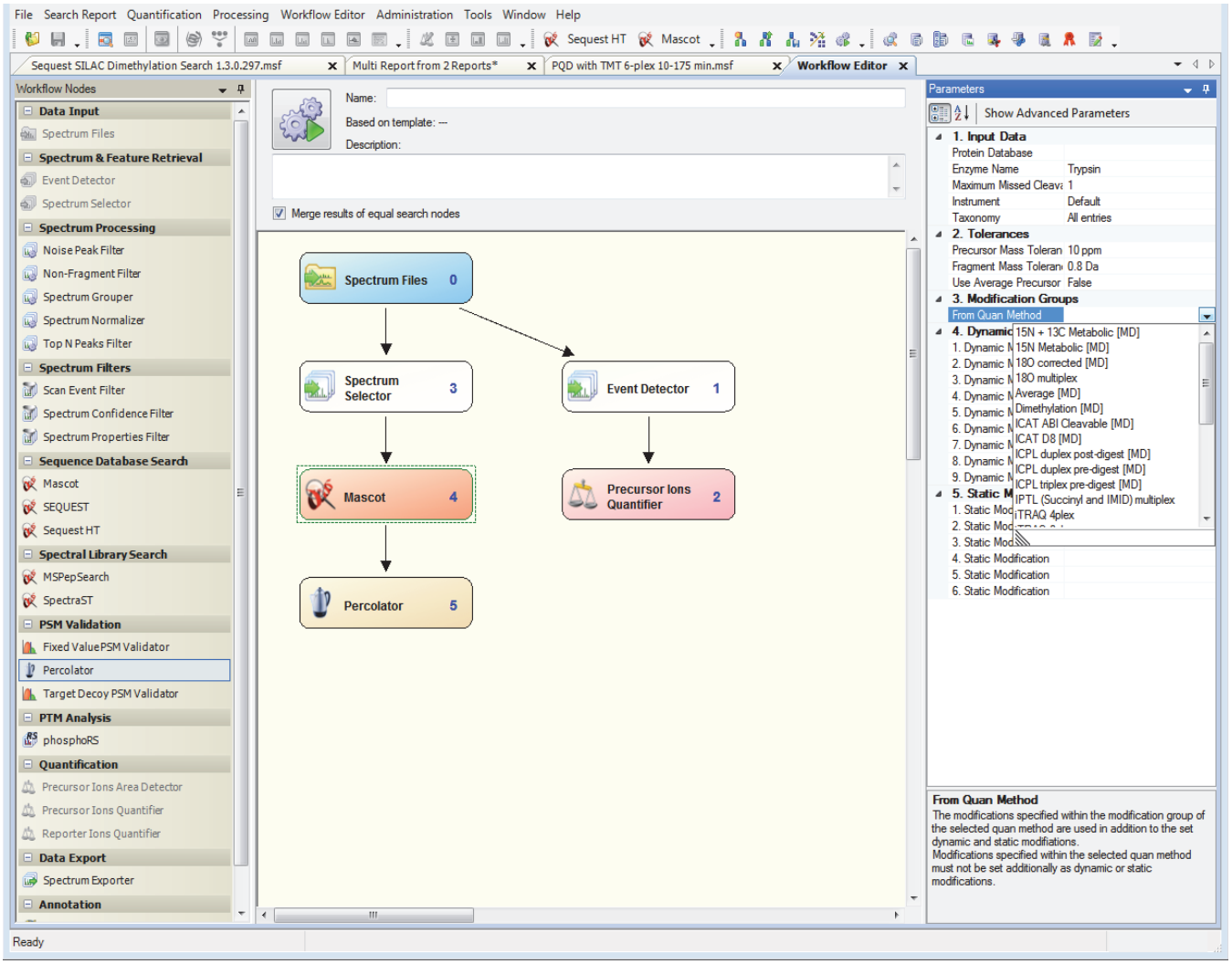
   –or–

   - Click the **From Quan Method** parameter in the Parameters pane under Modification Groups, and from the list (see Figure 176 for an example), select the modifications that you want to search for.

     You can select more than nine modifications.

     Note Do not use the modifications that you specify as part of the modification groups in the selected quantification method as additional dynamic or static modifications.

Figure 176. From Quan Method list



5. (Optional) If you want to group these modifications, go to the editor in the Mascot server window and choose **Configuration Editor > Quantitation**.

   Once you group the modifications, you can define them as fixed, variable, or exclusive. You can also define them directly for the method in reporter ion quantification or for each component in precursor ion quantification. Refer to the Mascot documentation for information on grouping modifications and defining the groups.

For the final search results, it does not matter whether you explicitly specified a modification as either a dynamic or a static modification or indirectly specified a modification from the chosen quantification method. As an exception, when you select an exclusive modification group, the Mascot search engine modifies all or none of the affected residues of a peptide sequence. Peptide matches with inconsistent labeling therefore no longer occur.

Note Using a Mascot quantification method to retrieve the modification groups to use does not affect how the Proteome Discoverer application performs the quantification. The application itself exclusively performs the quantification. You must specify in the application's methods any quantification labels used for the quantification.

Setting Up the Quantification Method

Setting up the quantification method is similar for both precursor ion quantification and reporter ion quantification. Both methods use values called quantification (quan) channels as the basis for the ratio reporting. You do not need to set up a quantification method for peak area calculation quantification.

For reporter ion quantification, a quantification channel is one of several masses, states, or tags (depending on which quantification method you use) for which you measure a quantification value. The Proteome Discoverer application calculates the reported quantification ratios from the quantification values of the different quantification channels. For example, for iTRAQ 4plex, the different reporter tags (114, 115, 116, 117) are the four quantification channels of the iTRAQ 4plex method. The application calculates the ratios from the detected quantification values of the four quantification channels.

For precursor ion quantification, a quantification channel is one of the different possible labeling states of a peptide corresponding to the different heavy amino acids used in the cell cultures. For example, the SILAC 2plex methods are normally used with two quantification channels named "light" and "heavy." The light quantification channel uses the natural isotopes of lysine ($^{12}C_6{}^{14}N_2$) and arginine ($^{12}C_6{}^{14}N_4$). In the heavy quantification channel, arginine 10 ($^{13}C_6{}^{15}N_4$) replaces all arginines, and either lysine 6 ($^{13}C_6{}^{14}N_2$) or lysine 8 ($^{13}C_6{}^{15}N_2$) replaces all lysines.

❖ To set up the quantification method

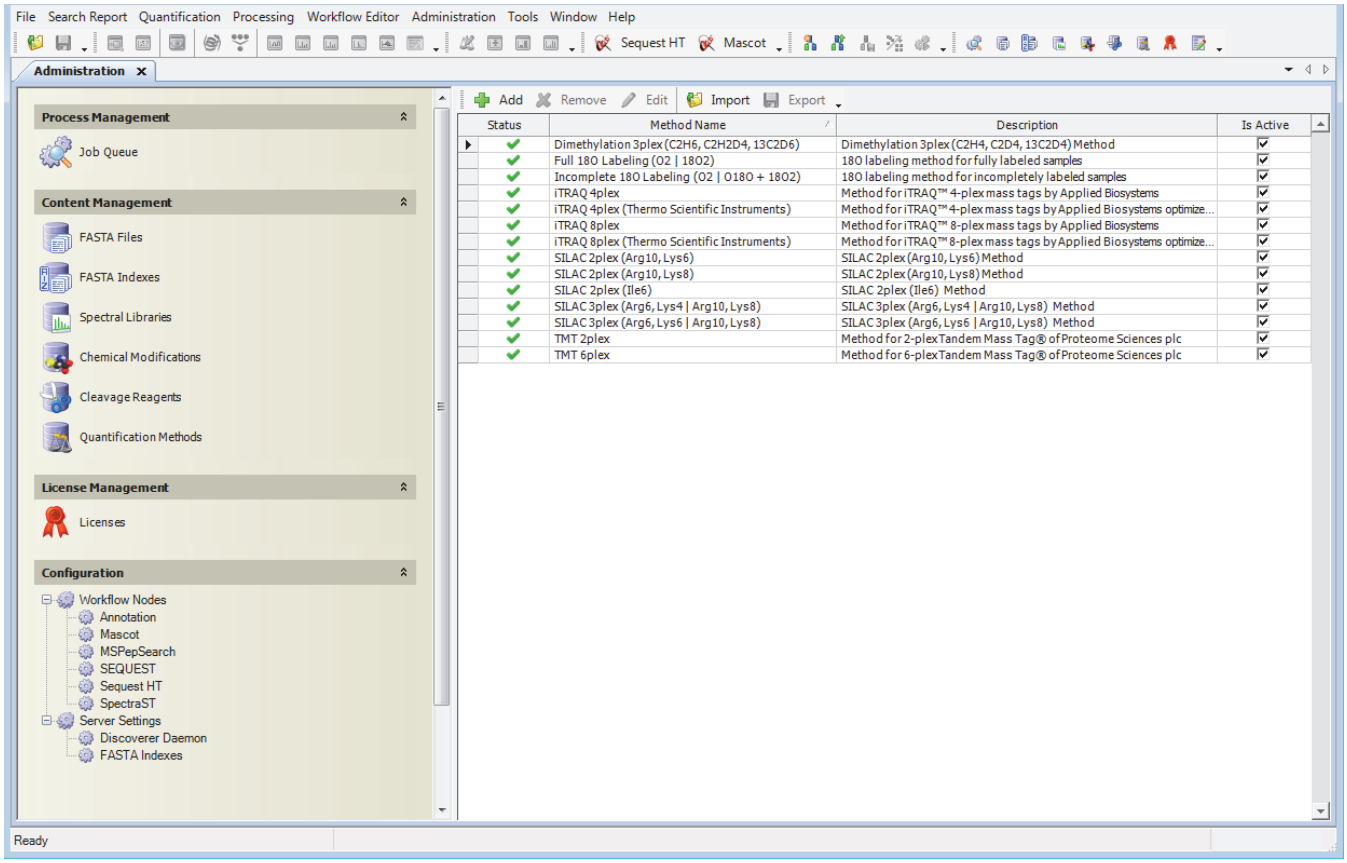
1. Set up a search by following the instructions in "Starting a New Search by Using the Workflow Editor" on page 42.
2. In the workspace pane of the Workflow Editor, add the **Precursor Ions Quantifier** node for precursor ion quantification or the **Reporter Ions Quantifier** node for reporter ion quantification.
3. Click the **Precursor Ions Quantifier** node or the **Reporter Ions Quantifier** node, and in the Quantification Method box, click the **Browse** button (...) that appears.

   The Quantification Method Editor dialog box opens to the Quan Channels page. Use this dialog box to set up the quantification method.

You can also access the Quantification Method Editor dialog box by choosing Administration > Maintain Quantification Methods or by clicking the Maintain Quantification Methods icon, , to open the Quantification Methods view, shown in Figure 177. This view lists all of the available methods for both precursor ion and reporter ion quantification.

Double-click the appropriate method in the Method Name column.

Figure 177. Quantification Methods view



4. From the list at the top of the Quantification Method Editor dialog box, select the quantification method.

   For precursor ion quantification, you can choose from the following methods when you initially set up a workflow and first access the Quantification Method Editor dialog box:

   - SILAC 2plex (Arg10, Lys6): Uses arginine 10 and lysine 6.
   - SILAC 2plex (Arg10, Lys8): Uses arginine 10 and lysine 8.
   - SILAC 2plex (Ile6): Uses isoleucine 6.
   - SILAC 3plex (Arg6, Lys4|Arg10, Lys8): Uses arginine 10 and lysine 8 for “heavy” labels and arginine 6 and lysine 4 for “medium” labels.

- SILAC 3plex (Arg6, Lys6|Arg10, Lys8): Uses arginine 10 and lysine 8 for "heavy" labels and arginine 6 and lysine 6 for "medium" labels.
- Dimethylation 3plex: Chemically adds isotopically labeled dimethyl groups to the N-terminus and to the ε-amino group of lysine.
- ^{18}O labeling: Introduces 2 or 4 Da mass tags through the enzyme-catalyzed exchange reaction of C-terminal oxygen atoms with ^{18}O.

For more information on these methods, see "Performing Precursor Ion Quantification" on page 243.

For reporter ion quantification, you can choose from the following methods when you initially set up a workflow and first access the Quantification Method Editor dialog box:

Note If you are installing the Proteome Discoverer application for the first time, the TMT 6plex quantification method is no longer available. The TMTe 6plex method replaces it.

- iTRAQ 4plex
- iTRAQ 4plex (Thermo Scientific Instruments)
- iTRAQ 8plex
- iTRAQ 8plex (Thermo Scientific Instruments)
- TMT 2plex
- TMT 6plex
- iodo TMT 6plex
- TMTe 6plex
- TMT 10plex

The two methods labeled "Thermo Scientific Instruments" have purity corrections optimized for the way Thermo Scientific mass spectrometers process samples and produce data.

For more information on these methods, see "Performing Reporter Ion Quantification" on page 249.

Specifying the Quantification Channels

The first step in setting up the quantification is to specify the quantification channels to use. This process includes a validation step. For precursor ion quantification, the validation step ensures that each peptide is in a valid labeling state according to the labels for the different channels, as defined in the quantification method. For reporter ion quantification, the validation step ensures that only peptides that have one of the specified reporter labels as a modification are considered for protein quantification.

The process of specifying label modifications is similar for precursor ion quantification and reporter ion quantification, but it also has some differences:

- For precursor ion quantification, you specify the label modifications for each quantification channel. For reporter ion quantification, you set the label modifications for the whole method.
- For precursor ion quantification, specifying the label modifications for quantification channels other than the unlabeled channel is mandatory. For reporter ion quantification, specifying the label modifications is optional because the information about the modification of the peptides is not necessary for processing the data. It is only used to verify the peptides when the Proteome Discoverer application loads the reports.

When you specify at least one of the label modifications in the quantification method, the Proteome Discoverer application verifies that each identified peptide has at least one of the specified modifications. It does not matter if only one terminal or only one residue is modified with the specified label modification.

- When the application identifies a peptide with none of the specified label modifications, this peptide cannot be the source of reporter peaks in the MS/MS spectra. As a result, the application marks the peptide “No Quan Labels” in the MSF report. It does not use these peptides when it calculates the protein quantification values from the peptides.
- When the application finds a peptide that does not have an iTRAQ or TMT label as a modification, even though reporter ions were present, it leaves the *Ratio* columns blank.

When you install the Proteome Discoverer application, the default methods for TMT and iTRAQ include the correct label modification. The application does not automatically update already existing reporter methods; you must manually specify the label modifications.

When you open old MSF files that contain reporter quantification data, the label modifications of the quantification method of the MSF file appear as None on the Quan Channels page of the Quantification Method Editor dialog box. You can manually specify the label modification, which then triggers the validation of the peptides, and save the change in the quantification method in the MSF file.

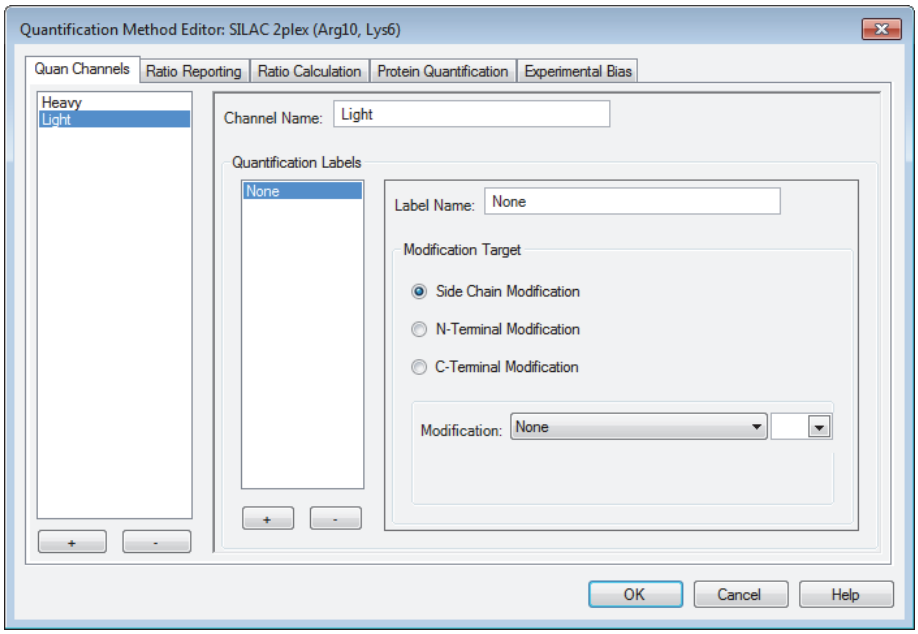
When you do not set the label modifications on the Quan Channels page, the Proteome Discoverer application does not perform the validation.

The process of specifying quantification channels for precursor ion quantification is slightly different from the process of specifying label modifications for reporter ion quantification.

❖ To specify quantification channels for precursor ion quantification

1. Click the **Quan Channels** tab of the Quantification Method Editor dialog box, shown in Figure 178, if it is not already selected.

   Figure 178. Quan Channels page of the Quantification Method Editor dialog box for precursor ion quantification



2. In the top list, select the name of the labeling method to use.

   When you create a new workflow and first access the Quantification Method Editor dialog box to set up a quantification method, the default methods available in the top list of the Quan Channels page are as follows:

   - SILAC 2plex (Arg10, Lys6)
   - SILAC 2plex (Arg10, Lys8)
   - SILAC 2plex (Ile6)
   - SILAC 3plex (Arg6, Lys4|Arg10, Lys8)
   - SILAC 3plex (Arg6, Lys6|Arg10, Lys8)
   - Dimethylation 3plex (C2H6, C2D4H2, 13C2D6)
   - ^{18}O labeling

   For a description of these methods, see "Performing Precursor Ion Quantification" on page 243.

   However, after you have chosen a method or set up your own method, only that method appears in the top list of the Quan Channels page after you execute the workflow.

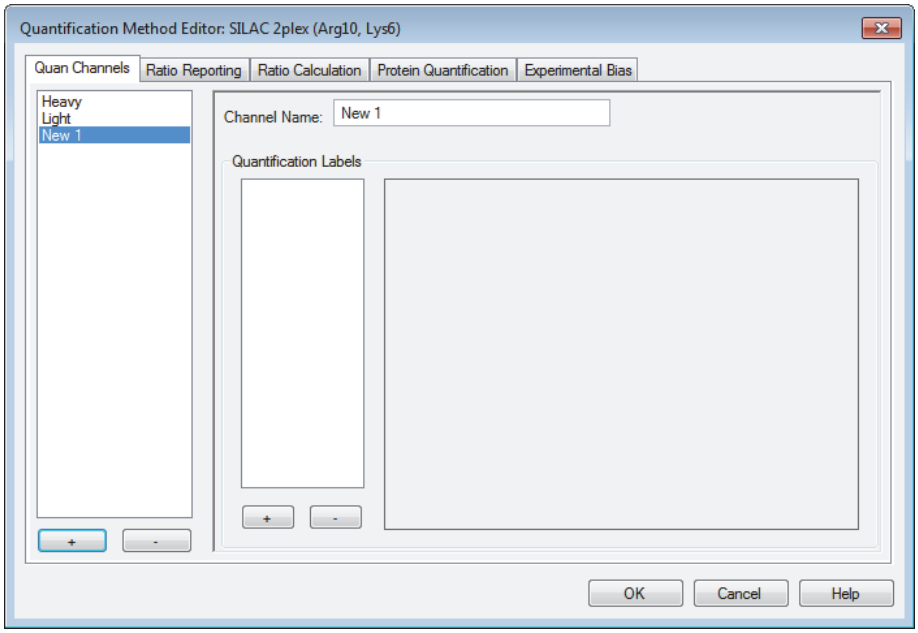
The left box of the Quan Channels page displays two types of default isotopic labels:

- Heavy: Refers to amino acid labels that use heavy isotopes, for example Arg10 and Lys8.
- Medium (3plex methods only): Refers to amino acid labels that use less massive isotopes, for example Arg6 and Lys4.
- Light: Refers to amino acid labels that use normal isotopes.

3. To add a quantification channel, click **+** beneath the list of quantification channels in the box on the left.

   The default name of New *number* now appears in the list of quantification channels and in the Channel Name box, as shown in Figure 179.

Figure 179. New quantification channel on the Quan Channels page



To remove a quantification channel, select the quantification channel in the list of quantification channels and click **–** beneath the list.

4. To specify a name for the new quantification channel, backspace over the default name in the Channel Name box and type the new name. The example in Figure 180 uses Medium.

   The new name now appears in the quantification channel (left) box.

5. To specify a quantification label to assign to a quantification channel, click **+** beneath the Quantification Labels box.

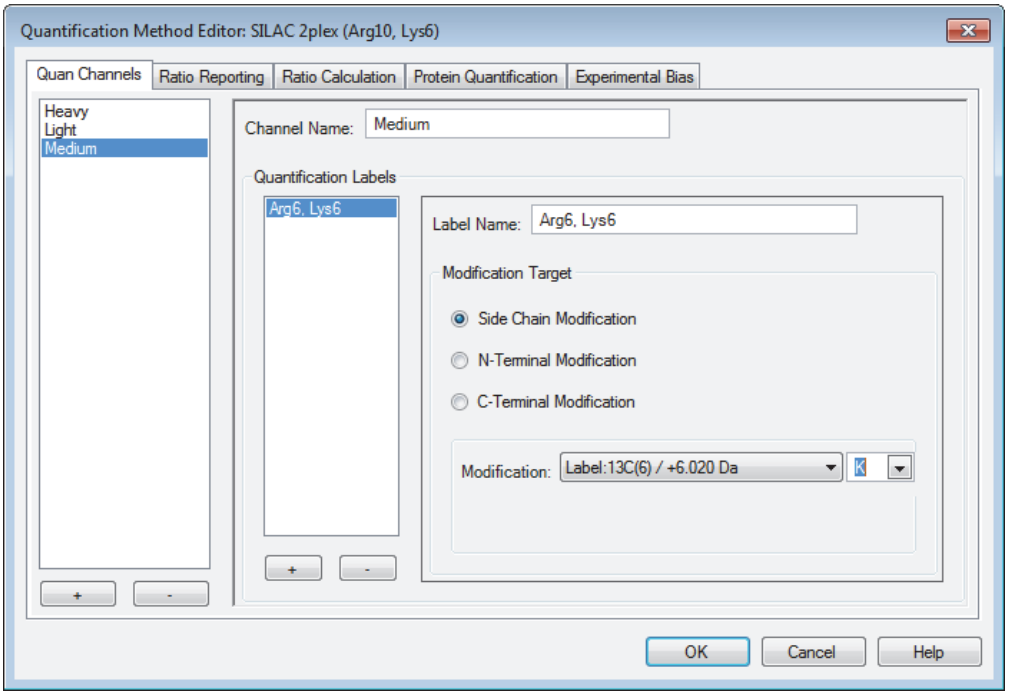
   A default quantification label of New *number* now appears in the Quantification Labels box and the Label Name box.

To remove an existing quantification label, select the label in the Quantification Labels box and click – beneath the box.

6. To change the default quantification channel name, backspace over the name in the Label Name box and type the new name. The example in Figure 180 uses Arg6, Lys6.
7. In the Modification Target area, select the location of the label on the peptide:
   - Side Chain Modification: Indicates that the label occurs on a side chain.
   - N-Terminal Modification: Indicates that the label occurs on the N terminus.
   - C-Terminal Modification: Indicates that the label occurs on the C terminus.
8. From the Modification list, select the modification to label the amino acid with. This example shows Label:13C(6) / +6.020 Da.
9. From the list adjacent to the Modification list, select the abbreviation of the amino acid selected in the Quantification Labels box on which the modification should occur. In this example, K is selected.

   The completed Quan Channels page will resemble Figure 180.

Figure 180. Completed Quan Channels page



10. Continue setting up the quantification method by following the instructions in "Setting Up Quantification Channels for Ratio Reporting" on page 273.

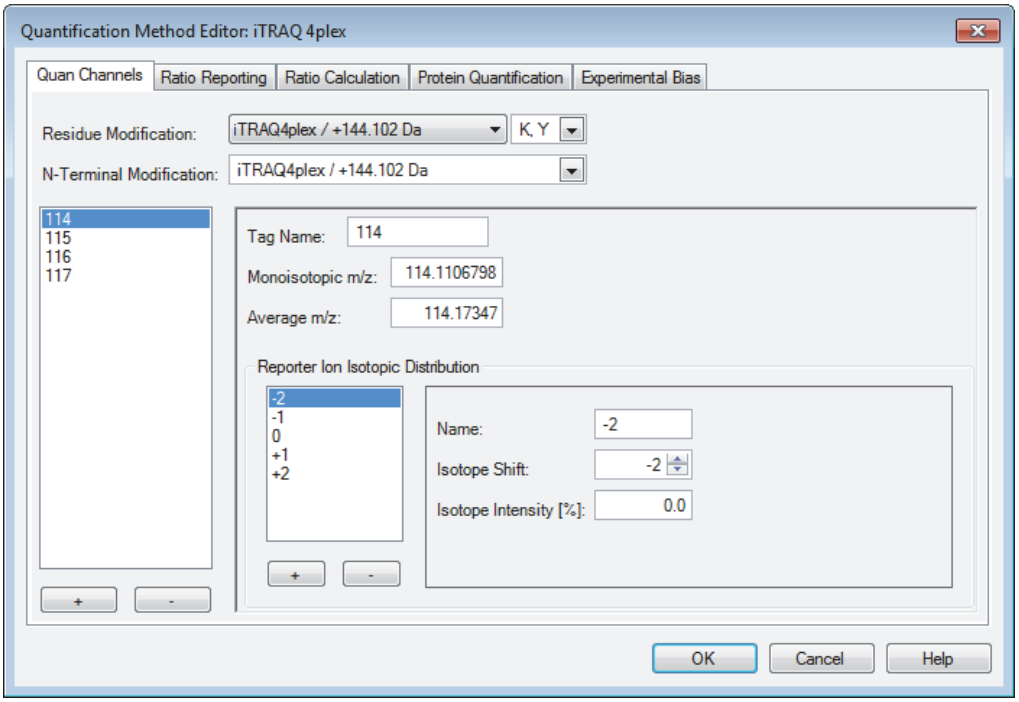
❖ To specify label modifications for reporter ion quantification

1. Click the **Quan Channels** tab of the Quantification Method Editor dialog box if it is not already selected (see Figure 181).
2. In the top list, select the name of the method to use. For reporter ion quantification, you can select the following default methods when you initially set up a workflow and first access the Quantification Method Editor dialog box.

   Note If you are installing the Proteome Discoverer application for the first time, the TMT 6plex quantification method does not appear in the application. The TMTe 6plex method replaces it.

   - iTRAQ 4plex
   - iTRAQ 4plex (Thermo Scientific Instruments)
   - iTRAQ 8plex
   - iTRAQ 8plex (Thermo Scientific Instruments)
   - TMT 2plex
   - TMT 6plex
   - iodo TMT 6plex
   - TMTe 6plex
   - TMT 10plex

Figure 181. Quan Channels page of the Quantification Method Editor dialog box for reporter ion quantification



3. From the Residue Modification list, select the label modification that would be found on the target amino acid residue. From the adjacent list, select the appropriate letter to indicate that the modification should occur on the indicated residue and will have an increased mass.

4. From the N-Terminal Modification list, select the label modification that would be found on the N terminus of each peptide.

   The left box of the Quan Channels page displays a list of mass tags, which are the fragmented labels.

5. To add a mass tag, click **+** beneath the list of mass tags in the box on the left.

   To remove a mass tag, select the mass tag you want to remove and click **–** beneath the list of mass tags.

6. When you add a mass tag or change the settings of an existing mass tag, do the following:

   a. In the Tag Name box, enter the name of the new mass tag if you do not want to use the default name.

   b. In the Monoisotopic m/z box, enter the monoisotopic mass-to-charge ratio of the new mass tag.

   c. In the Average m/z box, enter the average mass-to-charge ratio of the new mass tag.

   d. In the Reporter Ion Isotopic Distribution area, select the correction factor for the mass tags. Click **+** and **–** to add and delete correction factors.

For information on these correction factors, see "Using Reporter Ion Isotopic Distribution Values To Correct for Impurities" on page 308.

You must correct the purity of mass tags because of impurities in the tags themselves.

7. If you add a correction factor, do the following:
   a. In the Name box to the right of the list of correction factors, enter the name of the new correction factor.

      For the name, Thermo Fisher Scientific recommends that you use a plus (+) or a minus (–) symbol and the preferred shift number.

   b. In the Isotope Shift box, enter the isotope shift of the new correction factor.

      Isotope shift is a change in the spectral lines caused by different isotopes in an element. It often reflects impurities in the sample, and you must remove its corresponding mass-to-charge ratio from the calculations.

   c. In the Isotope Intensity [%] box, enter the isotope intensity of the new correction factor as a percentage.

      Isotope intensity is the intensity of the different isotopes in an element, often from impurities in the sample.

      Note The sum of the isotope intensities for each tag should add up to 100.

8. Continue setting up the quantification method by following the instructions in "Setting Up Quantification Channels for Ratio Reporting" on page 273.

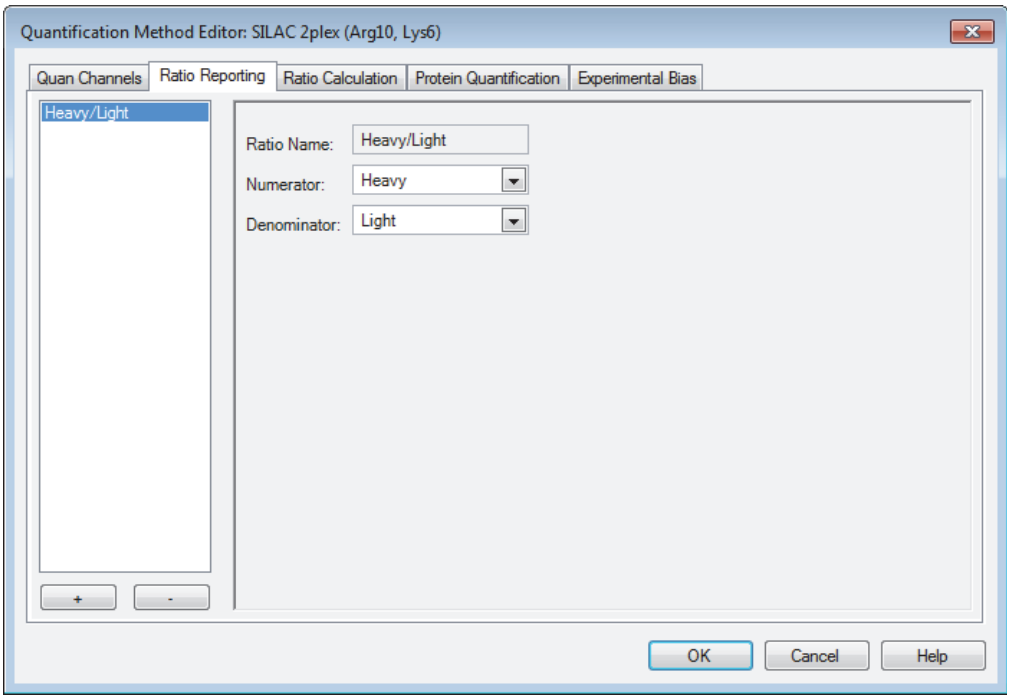
Setting Up Quantification Channels for Ratio Reporting

The Ratio Reporting page of the Quantification Method Editor dialog box specifies the names of the quantification channels (for precursor ion quantification) or mass tags (for reporter ion quantification) for the reporting of ratios that appear in the *Ratio* columns of the Proteins and Peptides pages.

❖ To set up the quantification channels for ratio reporting

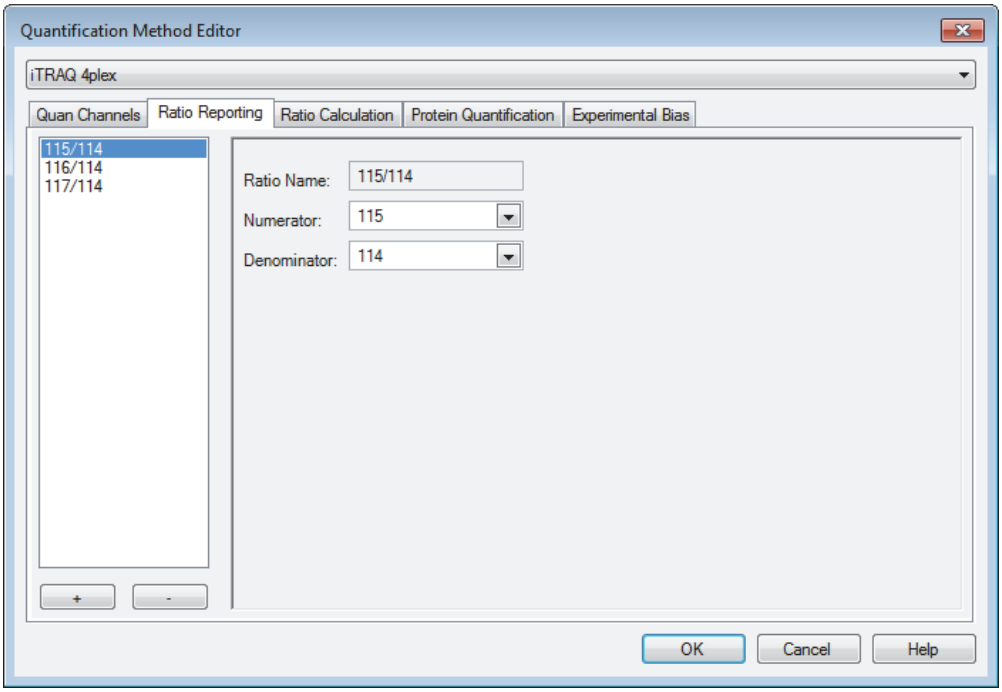
1. Click the **Ratio Reporting** tab, shown in Figure 182 for precursor ion quantification and in Figure 183 for reporter ion quantification.

Figure 182. Ratio Reporting page of the Quantification Method Editor dialog box for precursor ion quantification



In precursor ion quantification, the quantification ratios (left) box displays the ratio of the amino acids using heavy isotopes to the amino acids using normal isotopes.

Figure 183. Ratio Reporting page of the Quantification Method Editor dialog box for reporter ion quantification



In reporter ion quantification, the quantification ratios box to the left displays the name of the fragmented mass tag of a sample over the name of the mass tag of the reference sample.

2. To add any new quantification ratios, click **+** beneath the quantification ratios box.

   A new quantification ratio with the default name of New *number* appears in the quantification ratios pane.

   To remove the quantification ratio, select a quantification ratio and click **–** beneath the label type box.

3. If you added a quantification ratio, follow these steps.

   For precursor ion quantification:

   a. In the Numerator list, select the **Light** or **Heavy** label.

   b. In the Denominator list, select the **Light** or **Heavy** label that you did not select in the Numerator box.

   For reporter ion quantification:

   a. In the Numerator list, select the fragmented mass tag of the sample.

   b. In the Denominator list, select the name of the mass tag of the reference sample.

   You now see the specified numerator and denominator in the Ratio Name box, which is read-only.

4. Continue setting up the quantification method by following the instructions in "Setting Up the Ratio Calculation" on page 275.

Setting Up the Ratio Calculation

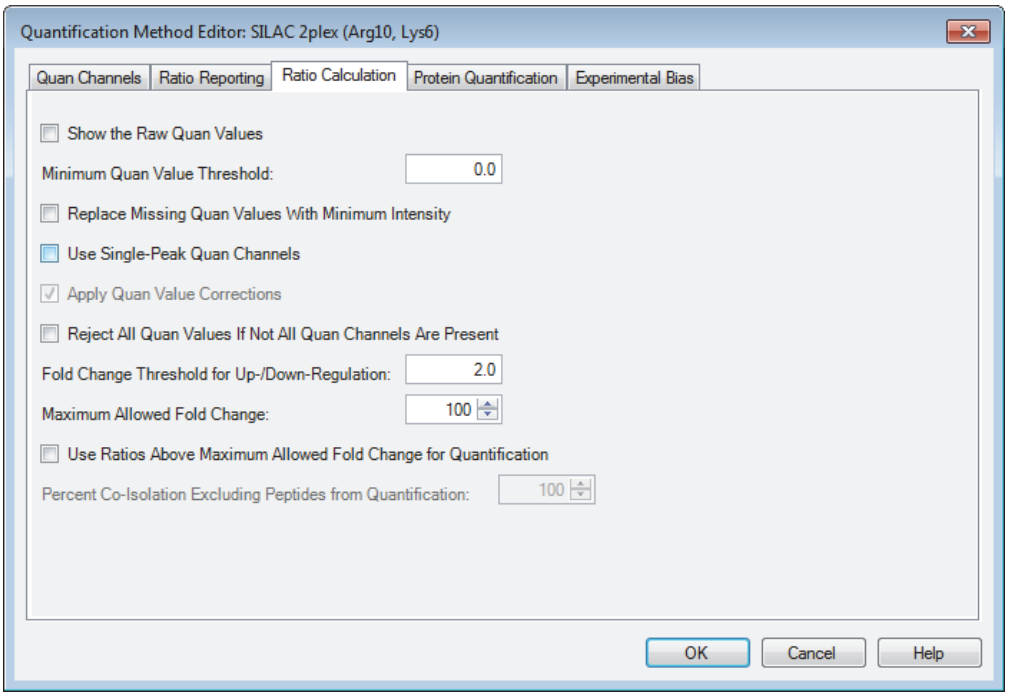
The Ratio Calculation page of the Quantification Method Editor dialog box controls how peptide and protein ratios are calculated from the raw quantification values of each quantification channel and how they are displayed on the Proteins and Peptides pages. For background information on the options available on this page, see "Missing Reporter Peaks in the Quantification Spectrum" on page 300.

❖ To set up the ratio calculation

1. Click the **Ratio Calculation** tab, shown in Figure 184.

   This page is the same for both precursor ion and reporter ion quantification.

Figure 184. Ratio Calculation page of the Quantification Method Editor dialog box



2. To create additional columns in the results report that display the reporter ion intensities (or the corrected reporter ion intensities when you selected Apply Quan Value Corrections) for every peptide, select the **Show the Raw Quan Values** check box.

   By default, this option is clear.

3. To set all quantification values whose intensity falls below a specified threshold to zero, type the threshold in the Minimum Quan Value Threshold box.

   The default threshold value is 0.0.

4. When the ratio of the ion intensity of the peptide in a sample to the ion intensity of the peptide in the control sample is missing or is 0 and you want to replace it with the minimum ion intensity detected, select the **Replace Missing Quan Values With Minimum Intensity** check box.

   The Proteome Discoverer application searches for the minimum ion intensity that is detected on all quantification channels and uses it as a best guess for the detection limit. It then uses this minimum value instead of the missing quantification values. When you specify a value higher than the detected minimum value, the application uses the value that you specify instead. The Quantification Summary page lists the minimum quantification value detected and the value actually used for the calculations. For information on the Quantification Summary page, see "Summarizing the Quantification" on page 292.

   By default, this check box is clear.

5. When you are performing precursor ion quantification and want the Proteome Discoverer application to consider missing quantification channels or quantification channels with just one peak as valid quantification results in the ratio calculation, do the following:

   a. Select the **Use Single-Peak Quan Channels** check box on the Ratio Calculation page.

   b. Set the Single-Peak/Missing Channels Allowed parameter of the Precursor Ions Quantifier node to **1**.

      For more information on this parameter, refer to the Help.

   By default, missing quantification channels or quantification channels with just one peak are not used for protein quantification. On the Peptides page, these peptides are marked "Excluded by Method" in the Quan Info column.

6. To apply the purity correction for the detected quantification values, select the **Apply Quan Value Corrections** check box.

   For reporter ion quantification, this option applies the correction for isotopic impurities. No such correction is currently available for precursor ion quantification. The application applies this purity correction after applying other settings that potentially change the quantification values.

   This option is selected by default.

7. To avoid using quantification values from any of the channels when one or more of the quantification channels has a detected intensity of zero, select the **Reject All Quan Values If Not All Quan Channels Are Present** check box.

   By default, this check box is clear.

8. To highlight a change in the ion intensity ratio (that is, the ratio of the ion intensity of the peptide in an experimental sample to the ion intensity of the peptide in the control sample) larger than n or smaller than $1/n$ in the results, specify n in the Fold Change Threshold for Up-/Down-Regulation box.

   The default is 2.0.

   For example, if you select 2 in the Fold Change Threshold for Up-/Down-Regulation box, the Proteome Discoverer application highlights those experimental results that are greater than twice as large (up-regulation) or less than half as large (down-regulation) as the control.

9. To exclude a peptide ion intensity ratio (that is, the ratio of the ion intensity of the peptide in a sample to the ion intensity of the peptide in the control sample) that exceeds a certain maximum, enter this maximum number in the Maximum Allowed Fold Change box.

   The minimum value is 1, and the maximum value is 100 000.

The default is 100. With the default setting, calculated ratios above 100 are set to 100, and calculated ratios below 0.01 are set to 0.01.

For example, if you set Maximum Allowed Fold Change to 10, the Proteome Discoverer application excludes any peptide ratios showing a greater than a ten-fold change in ion intensity for an experiment compared to the control.

10. To report larger ratios than you have indicated in the Maximum Allowed Fold Change box, select the **Use Ratios Above Maximum Allowed Fold Change for Quantification** check box.

    This option reports the quantification ratios based on the maximum values. Values greater than the value selected in the Maximum Allowed Fold Change box are replaced by the maximum or minimum value.

    By default, this check box is clear.

11. Continue setting up the quantification method by following the instructions in "Setting Peptide Parameters Used to Calculate Protein Ratios" on page 278.

The settings of the options on the Ratio Calculation page govern the appearance of the experimental results in the columns in the MSF report. The data can appear in the following colors:

- Pink: The experimental results are down-regulated.
- Blue: The experimental results are up-regulated.
- Red: The experimental results exceed the setting in the Maximum Allowed Fold Change box. These results are not used in calculations unless you select the Use Ratios Above Maximum Allowed Fold Change for Quantification option.

Setting Peptide Parameters Used to Calculate Protein Ratios

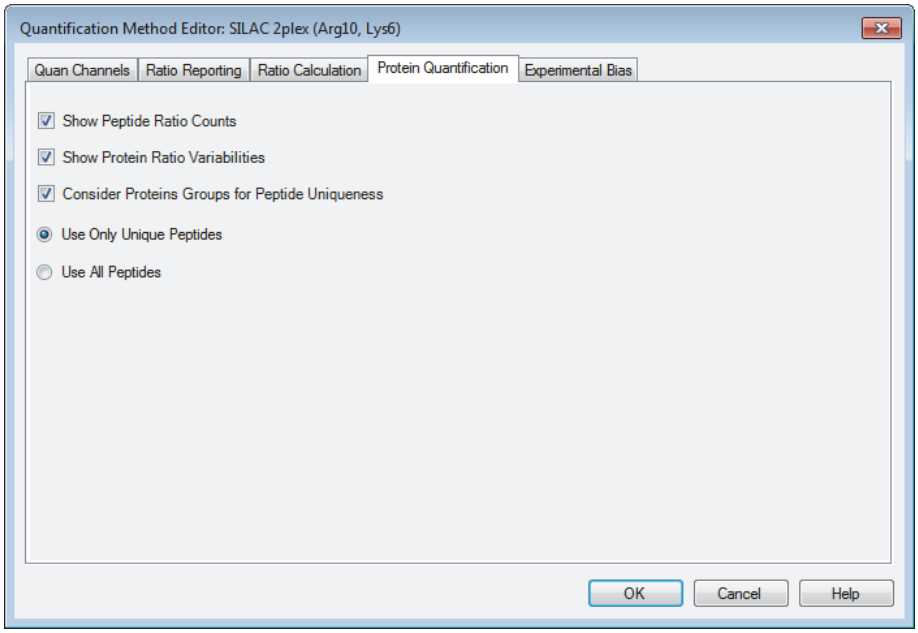
Use the Protein Quantification page of the Quantification Method Editor dialog box to set the peptide parameters for calculating protein ratios.

❖ To set the peptide parameters used to calculate protein ratios

1. Click the **Protein Quantification** tab, shown in Figure 185.

   This page is the same for both precursor ion and reporter ion quantification.

Figure 185. Protein Quantification page of the Quantification Method Editor dialog box



2. If you want to display the number of peptide ratios that are used to calculate a protein ratio, select the **Show Peptide Ratio Counts** check box.

   The results appear in the Heavy/Light Count column of the Proteins page of the MSF report for precursor ion quantification and in the *Ratio* Count columns of the Proteins page for reporter ion quantification. For more information on ratio counts, see "Ratio Count" on page 324.

   This option is selected by default.

3. If you want to show the variability of the peptide ratios used to calculate the protein ratios, select the **Show Protein Ratio Variabilities** check box.

   The results appear in the Heavy/Light Variability [%] column of the Proteins page of the MSF report for precursor ion quantification and in the *Ratio* Variability [%] columns of the Proteins page for reporter ion quantification. For more information on protein variability, see "Ratio Variability" on page 324.

   This option is selected by default.

4. If you want to define peptide uniqueness on the basis of protein groups rather than on individual proteins, select the **Consider Proteins Groups for Peptide Uniqueness** check box.

   This option is selected by default.

5. Choose the type of peptides for the Proteome Discoverer application to use in the quantification:

   - (Default) Use Only Unique Peptides: Includes peptides that do not occur in other proteins.

- Use All Peptides: Includes all detected peptides, whether or not they also occur in other proteins.

6. Continue setting up the quantification method by following the instructions in the next section, "Correcting Experimental Bias."

Correcting Experimental Bias

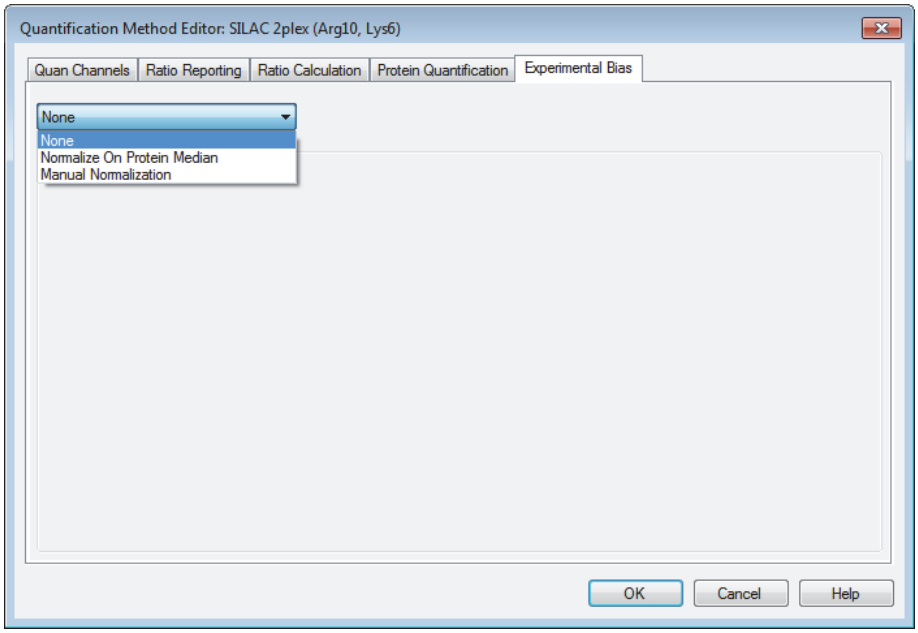
The purpose of the Experimental Bias page of the Quantification Method Editor dialog box is to correct experimental bias, which is the difference in the total observed protein abundance between two or more samples. Assuming that in real samples most of the proteins are not regulated, the intensity of the median protein in sample *x* should be the same as the intensity of the median protein in sample *y*. If it is not, it may indicate experimental bias caused by, for example, errors in pipetting or the determination of protein concentration in the mixed samples. You must correct for the difference. For best results, always enter a small normalization factor.

❖ To correct experimental bias

1. Click the **Experimental Bias** tab, shown in Figure 186.

   This page is the same for both precursor ion and reporter ion quantification.

   Figure 186. Experimental Bias page of the Quantification Method Editor dialog box



2. Select the normalization factor to apply from the list at the top of the page:

   - (Default) None: Performs no normalization.
   - Normalize on Protein Median: Normalizes all peptide ratios by the median protein ratio. The median protein ratio should be 1 after the normalization.

- Manual Normalization: Specifies a user-defined normalization number.

When you select the Normalize on Protein Median setting, the Minimum Protein Count box appears.

When you select the Manual Normalization setting, the Normalization Factor box appears.

3. Do one of the following:

   - In the Minimum Protein Count box, which appears when you selected Normalization on Protein Median, enter the minimum number of proteins that must be observed to allow normalization.

   –or–

   - In the Normalization Factor box, which appears when you select Manual Normalization, enter the normalization factor.

   The default for the Minimum Protein Count option is 20, and the default for the Normalization Factor option is 1.0.

   Normalization cannot work if there are too few proteins in a sample.

4. Click **OK**.

Checking the Quantification Method

The Proteome Discoverer application checks the parameters that you have set for the quantification method. For reporter ion quantification, it verifies that the method has at least two channels. For precursor ion quantification, it checks for the following:

- At least one quantification channel
- At least one label for each quantification channel
- Unique label names in a channel
- The modification of each label applied to at least one amino acid, unless you chose None for a modification
- Each amino acid labeled only once in a channel. Labels must have an elemental composition defined.
- Each label mass used only once (label masses vary by at least 1.0 Da)

You cannot apply changes to a quantification method unless the method meets all these criteria.

Restoring Quantification Method Template Defaults

If you have altered one of the quantification method templates listed at the beginning of "Setting Up the Quantification Method" on page 264, you can restore the original template.

❖ To restore the original template

1. Choose **Administration > Maintain Quantification Methods**, or click on the **Maintain Quantification Methods** icon, .

   The Quantification Methods view opens, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

2. To open the Quantification Method Editor dialog box, click **Add** in the Quantification Methods view.

   The Create Quantification Method dialog box opens, as shown in Figure 192 on page 287.

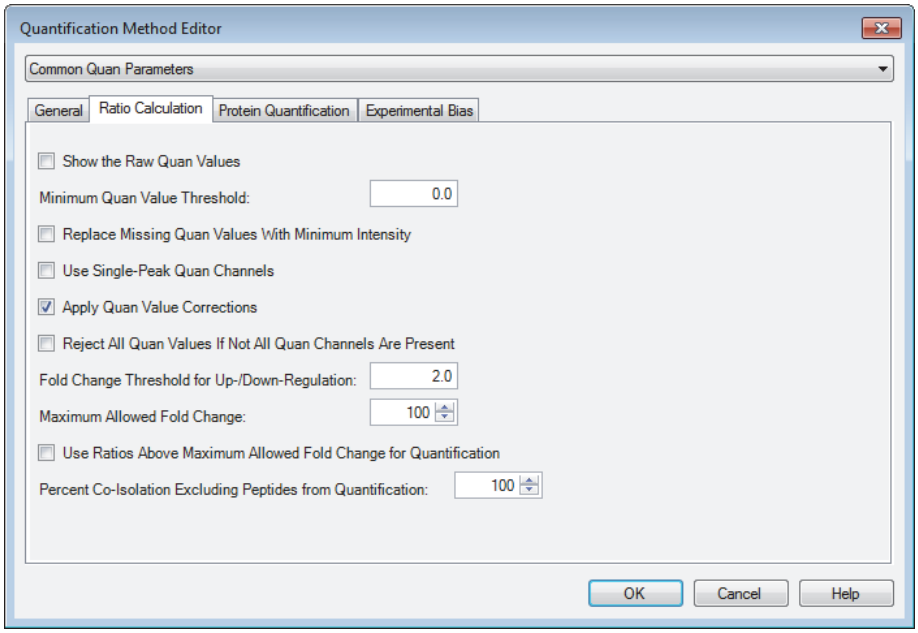
3. Select the appropriate template from the Create from Factory Defaults list.

4. Set up the quantification method according to the instructions in "Setting Up the Quantification Method" on page 264.

Setting Up the Quantification Method for Multiple Input Files

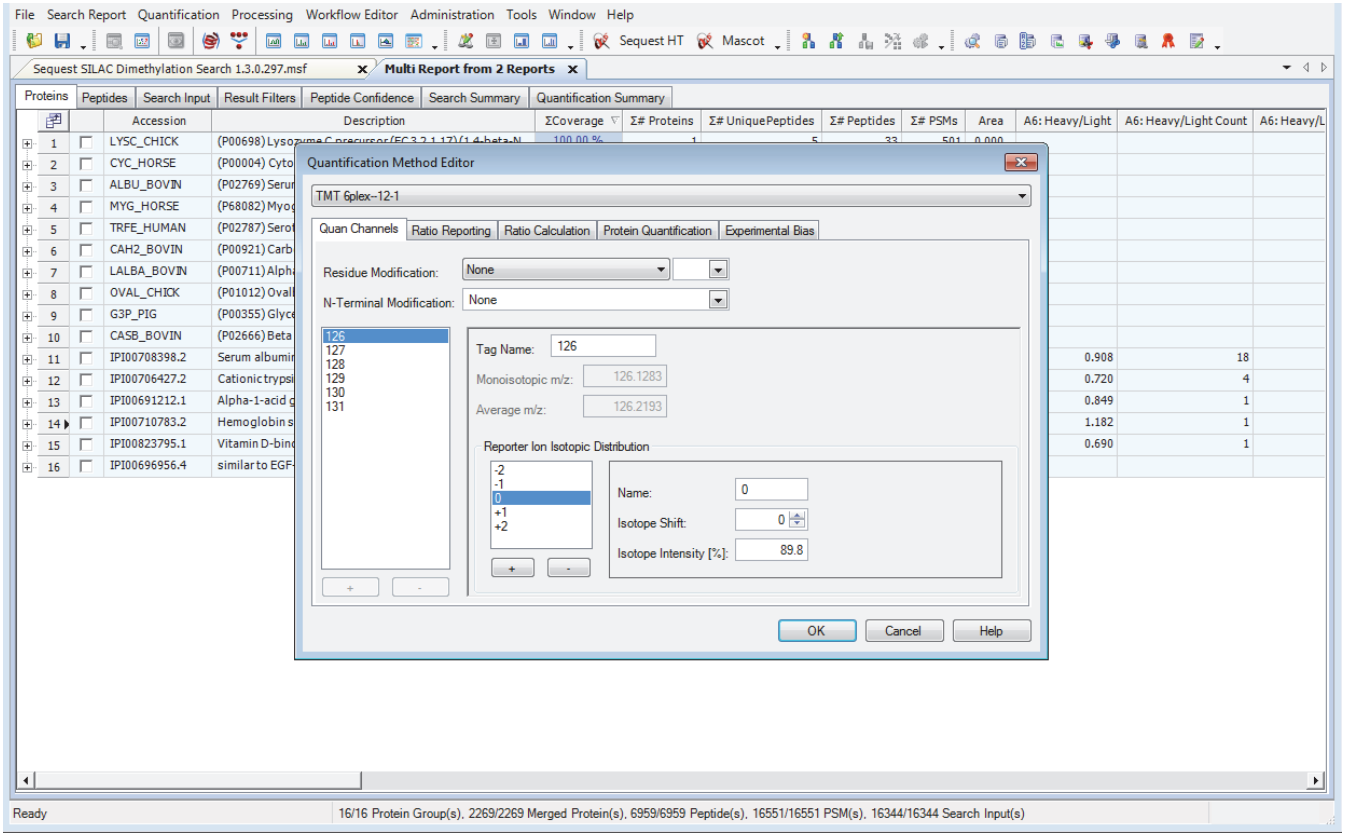
When you load multiple MSF files, you can apply the settings of the Ratio Calculation, Protein Quantification, and Experimental Bias pages of the Quantification Method Editor dialog box to all the loaded input files by selecting Common Quan Parameters from the list at the top of the dialog box, as shown in Figure 187. These pages contain the same options as those for single-file processing.

Figure 187. Quantification Method Editor dialog box for multiple input files



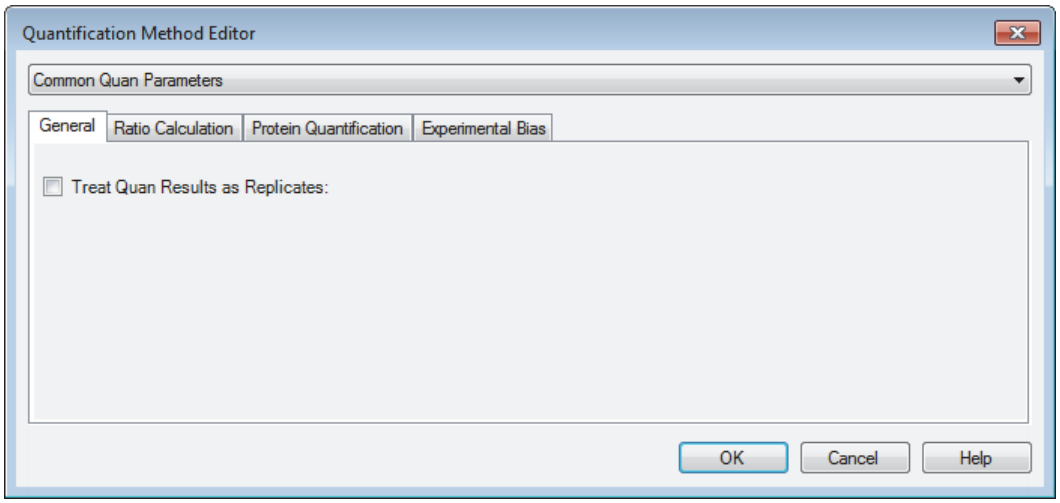
Although you cannot apply the settings of the Quan Channels and Ratio Reporting pages to multiple loaded MSF files, you can apply them to individual MSF files, as shown in Figure 188. You can access these two pages by selecting the individual MSF file from the list at the top of the dialog box.

Figure 188. Applying Quan Channel page settings to an individual MSF file when multiple MSF files are loaded



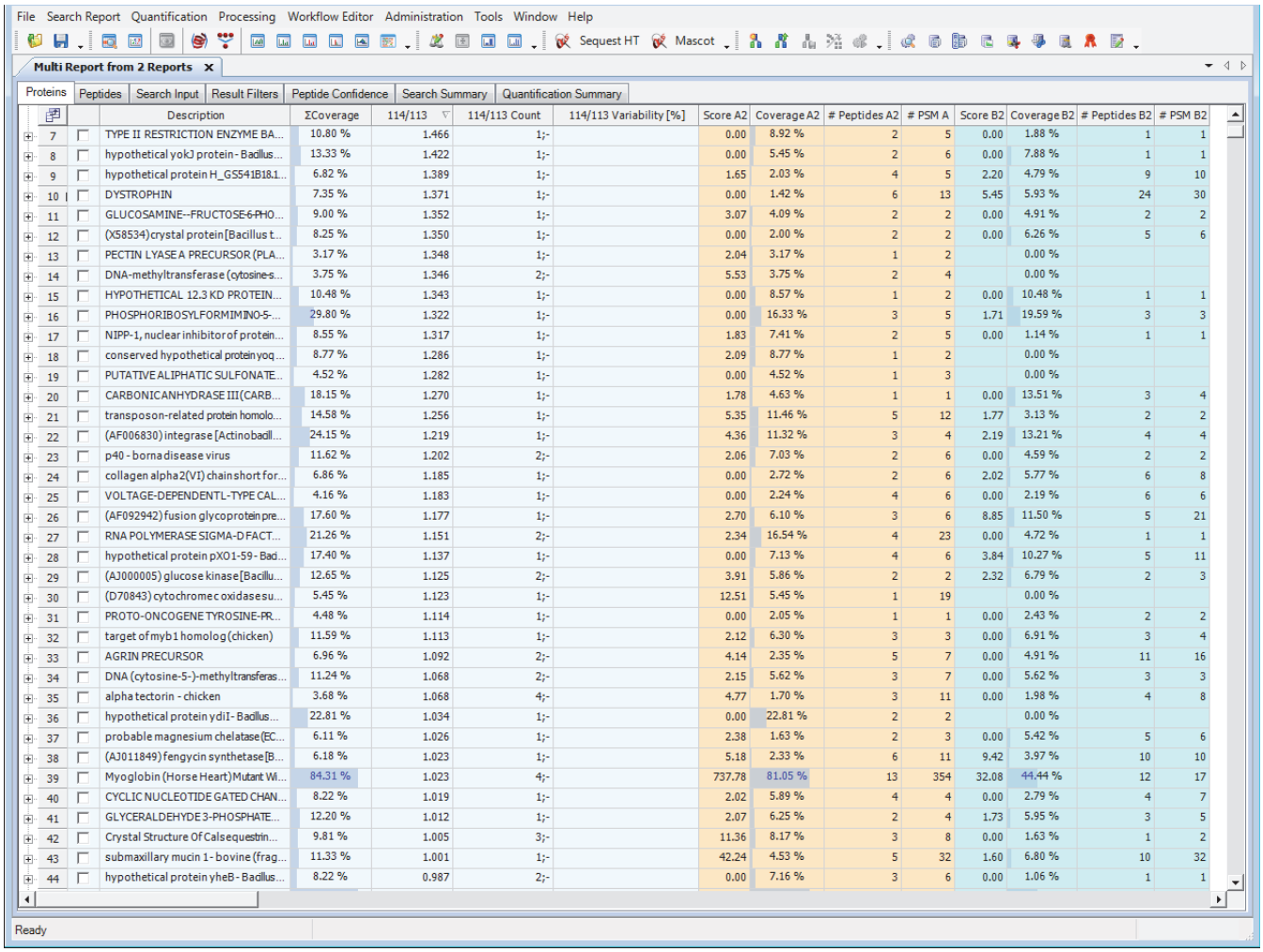
The Quantification Method Editor dialog box also includes a General page when multiple MSF files are loaded at the same time. It contains one option, Treat Quan Results as Replicates, as shown in Figure 189. This option treats protein-level quantification values with the same ratio names and the same quantification method as replicates (that is, the protein ratios of the individual files are averaged into a replicate ratio).

Figure 189. General page of the Quantification Method Editor dialog box



When you select Treat Quan Results as Replicates and click OK, the protein quantification data looks like the data in Figure 190.

Figure 190. Protein quantification data in replicate mode



		Description	ΣCoverage	114/113	114/113 Count	114/113 Variability [%]	Score A2	Coverage A2	# Peptides A2	# PSM A	Score B2	Coverage B2	# Peptides B2	# PSM B2
7		TYPE II RESTRICTION ENZYME BA...	10.80 %	1.466	1;-		0.00	8.92 %	2	5	0.00	1.88 %	1	1
8		hypothetical yokJ protein - Bacillus...	13.33 %	1.422	1;-		0.00	5.45 %	2	6	0.00	7.88 %	1	1
9		hypothetical protein H_GS541B18.1...	6.82 %	1.389	1;-		1.65	2.03 %	4	5	2.20	4.79 %	9	10
10		DYSTROPHIN	7.35 %	1.371	1;-		0.00	1.42 %	6	13	5.45	5.93 %	24	30
11		GLUCOSAMINE--FRUCTOSE-6-PHO...	9.00 %	1.352	1;-		3.07	4.09 %	2	2	0.00	4.91 %	2	2
12		(X58534) crystal protein [Bacillus t...	8.25 %	1.350	1;-		0.00	2.00 %	2	2	0.00	6.26 %	5	6
13		PECTIN LYASE A PRECURSOR (PLA...	3.17 %	1.348	1;-		2.04	3.17 %	1	2		0.00 %		
14		DNA-methyltransferase (cytosine-s...	3.75 %	1.346	2;-		5.53	3.75 %	2	4		0.00 %		
15		HYPOTHETICAL 12.3 KD PROTEIN...	10.48 %	1.343	1;-		0.00	8.57 %	1	2	0.00	10.48 %	1	1
16		PHOSPHORIBOSYLFORMIMINO-5-...	29.80 %	1.322	1;-		0.00	16.33 %	3	5	1.71	19.59 %	3	3
17		NIPP-1, nuclear inhibitor of protein...	8.55 %	1.317	1;-		1.83	7.41 %	2	5	0.00	1.14 %	1	1
18		conserved hypothetical protein yoq...	8.77 %	1.286	1;-		2.09	8.77 %	1	2		0.00 %		
19		PUTATIVE ALIPHATIC SULFONATE...	4.52 %	1.282	1;-		0.00	4.52 %	1	3		0.00 %		
20		CARBONICANHYDRASE III (CARB...	18.15 %	1.270	1;-		1.78	4.63 %	1	1	0.00	13.51 %	3	4
21		transposon-related protein homolo...	14.58 %	1.256	1;-		5.35	11.46 %	5	12	1.77	3.13 %	2	2
22		(AF006830) integrase [Actinobacill...	24.15 %	1.219	1;-		4.36	11.32 %	3	4	2.19	13.21 %	4	4
23		p40 - borna disease virus	11.62 %	1.202	2;-		2.06	7.03 %	2	6	0.00	4.59 %	2	2
24		collagen alpha2(VI) chain short for...	6.86 %	1.185	1;-		0.00	2.72 %	2	6	2.02	5.77 %	6	8
25		VOLTAGE-DEPENDENTL-TYPE CAL...	4.16 %	1.183	1;-		0.00	2.24 %	4	6	0.00	2.19 %	6	6
26		(AF092942) fusion glycoprotein pre...	17.60 %	1.177	1;-		2.70	6.10 %	3	6	8.85	11.50 %	5	21
27		RNA POLYMERASE SIGMA-D FACT...	21.26 %	1.151	2;-		2.34	16.54 %	4	23	0.00	4.72 %	1	1
28		hypothetical protein pXO1-59 - Bad...	17.40 %	1.137	1;-		0.00	7.13 %	4	6	3.84	10.27 %	5	11
29		(AJ000005) glucose kinase [Bacillu...	12.65 %	1.125	2;-		3.91	5.86 %	2	2	2.32	6.79 %	2	3
30		(D70843) cytochrome c oxidase su...	5.45 %	1.123	1;-		12.51	5.45 %	1	19		0.00 %		
31		PROTO-ONCOGENE TYROSINE-PR...	4.48 %	1.114	1;-		0.00	2.05 %	1	1	0.00	2.43 %	2	2
32		target of myb1 homolog (chicken)	11.59 %	1.113	1;-		2.12	6.30 %	3	3	0.00	6.91 %	3	4
33		AGRIN PRECURSOR	6.96 %	1.092	2;-		4.14	2.35 %	5	7	0.00	4.91 %	11	16
34		DNA (cytosine-5-)-methyltransferas...	11.24 %	1.068	2;-		2.15	5.62 %	3	7	0.00	5.62 %	3	3
35		alpha tectorin - chicken	3.68 %	1.068	4;-		4.77	1.70 %	3	11	0.00	1.98 %	4	8
36		hypothetical protein ydiI - Bacillus...	22.81 %	1.034	1;-		0.00	22.81 %	2	2		0.00 %		
37		probable magnesium chelatase (EC...	6.11 %	1.026	1;-		2.38	1.63 %	2	3	0.00	5.42 %	5	6
38		(AJ011849) fengycin synthetase [B...	6.18 %	1.023	1;-		5.18	2.33 %	6	11	9.42	3.97 %	10	10
39		Myoglobin (Horse Heart) Mutant Wi...	84.31 %	1.023	4;-		737.78	81.05 %	13	354	32.08	44.44 %	12	17
40		CYCLIC NUCLEOTIDE GATED CHAN...	8.22 %	1.019	1;-		2.02	5.89 %	4	4	0.00	2.79 %	4	7
41		GLYCERALDEHYDE 3-PHOSPHATE...	12.20 %	1.012	1;-		2.07	6.25 %	2	4	1.73	5.95 %	3	5
42		Crystal Structure Of Calsequestrin...	9.81 %	1.005	3;-		11.36	8.17 %	3	8	0.00	1.63 %	1	2
43		submaxillary mucin 1 - bovine (frag...	11.33 %	1.001	1;-		42.24	4.53 %	5	32	1.60	6.80 %	10	32
44		hypothetical protein yheB - Bacillus...	8.22 %	0.987	2;-		0.00	7.16 %	3	6	0.00	1.06 %	1	1

There is only one ratio column for each specified ratio, and the *Ratio* Counts columns show the number of peptides used from every single MSF file for calculating the individual protein ratios for each file. These individual protein ratios are then averaged, and the average is displayed in the ratio columns for the proteins.

Adding a Quantification Method

You can use the following procedure to add a quantification method. You can also use it to access the quantification methods without loading an MSF file.

❖ To add a quantification method

1. Choose **Administration > Maintain Quantification Methods**, or click the **Maintain Quantification Methods** icon,  .

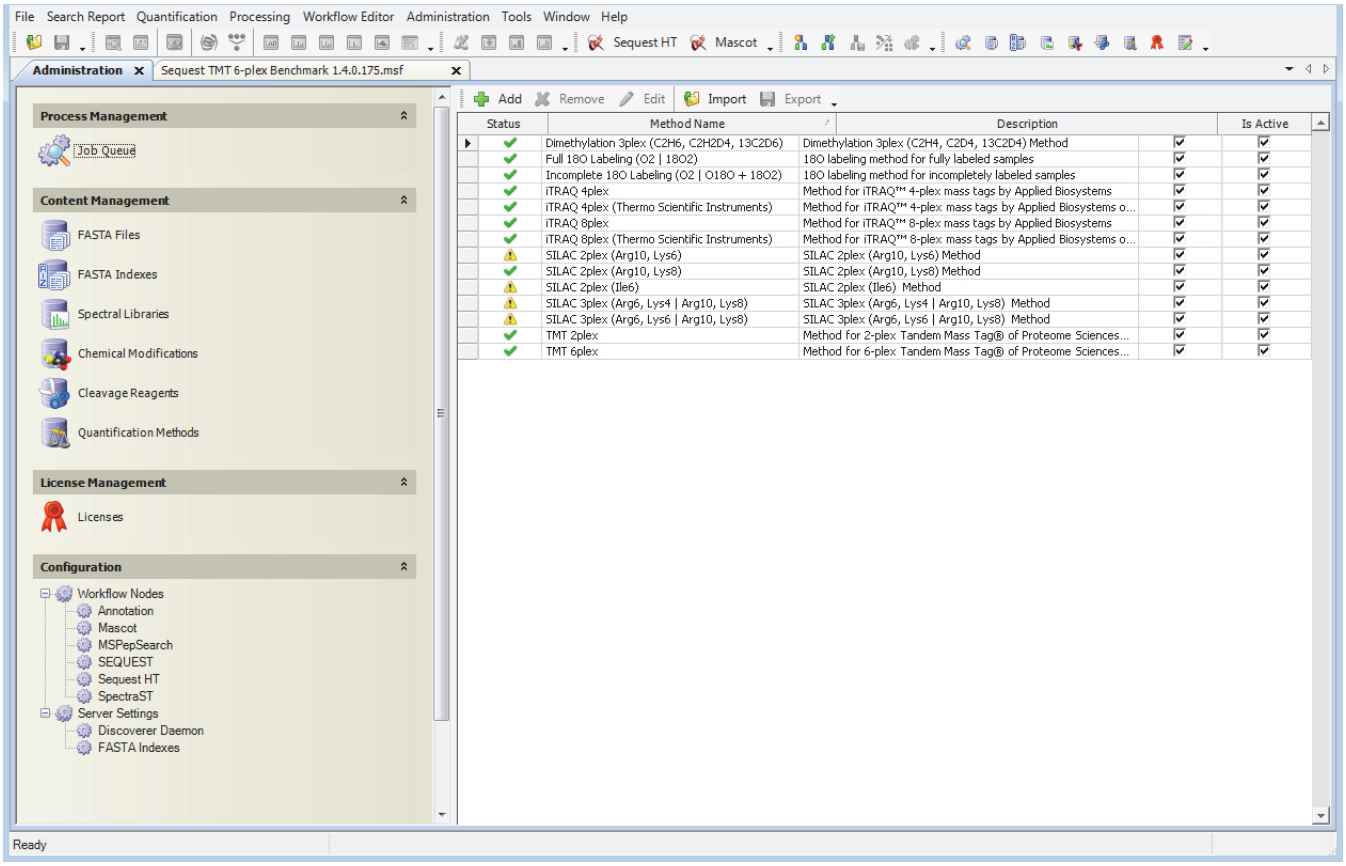
The Quantification Methods view opens, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

The Status column indicates whether the quantification method is valid for use in quantification:

- A green check mark means that the quantification method is valid and can be used for quantification.
- An exclamation point in a yellow triangle means that the quantification method is not valid. Double-click this mark to view a message that describes the error and provides information on how to fix it.

Figure 191 provides examples of these symbols in the Status column.

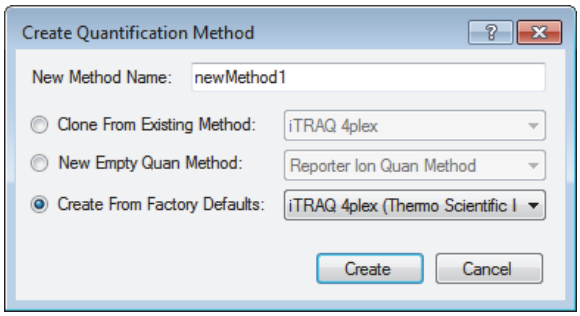
Figure 191. Method validity symbols in the Quantification Methods view



2. Click **Add**.

   The Create Quantification Method dialog box now appears, as shown in Figure 192.

   Figure 192. Create Quantification Method dialog box



3. In the New Method Name box, type the name of the quantification method that you want to create.

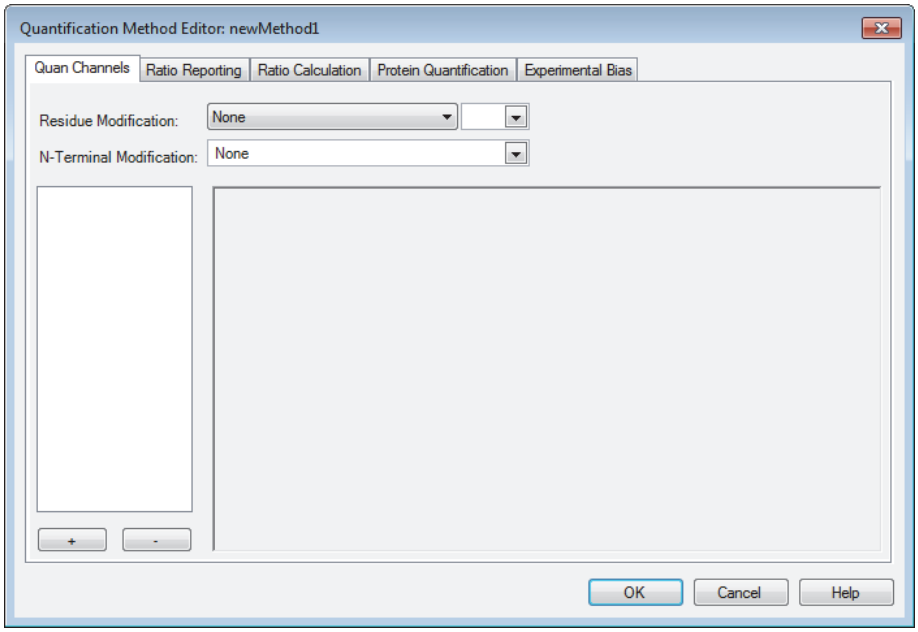
4. Select one of the following methods of creating a quantification method:

   - Clone From Existing Method: Uses the same settings as those of the existing quantification method that you select from the list. The list of methods is the same as that given at the beginning of "Setting Up the Quantification Method" on page 264.
   - New Empty Quan Method: Uses one of the following templates so that you can build a new processing method from scratch:
     - Reporter Ion Quan Method: Provides a template for reporter ion quantification.
     - Precursor Ion Quan Method: Provides a template for precursor ion quantification.
   - (Default) Create From Factory Defaults: Creates a new method using the same settings from one of the default settings that appear when the Proteome Discoverer application is newly installed.

5. Click **Create**.

   The Quantification Method Editor dialog box appears, as shown in Figure 181 on page 272 through Figure 186 on page 280. The Quan Channels page and the Ratio Reporting page are blank if you selected the New Empty Quan Method option. In this case, the Quan Channels page resembles Figure 193.

Figure 193. Empty quantification method template



6. To specify the parameters of the new quantification method, follow the procedure given in "Setting Up the Quantification Method" on page 264.

Changing a Quantification Method

After you perform quantification, you can change the quantification method of the current report. You can add new quantification methods by copying an existing method and editing it. You can also activate and deactivate methods that you want visible or hidden when setting up a quantification workflow. However, you cannot define mass tags or labels as you can when setting up the initial quantification method, because they have already been measured.

You can access the quantification methods without loading an MSF file by using Administration > Maintain Quantification Methods (see "Adding a Quantification Method" on page 285), but if you want to save any changes to the quantification method in a report, you must first open that report and use Quantification > Edit Quantification Method.

❖ To change a quantification method

1. Select the MSF report whose quantification method you want to change.
2. Do one of the following:
   - Choose **Quantification > Edit Quantification Method**, or click the **Edit Quantification Method** icon, , either on the toolbar or on the Administration page.

     Note To access the Edit Quantification Method command, you must first run a workflow that uses the Reporter Ions Quantifier node or the Precursor Ions Quantifier node.

     The Quantification Method Editor dialog box appears, as shown in Figure 181 on page 272 through Figure 186 on page 280.

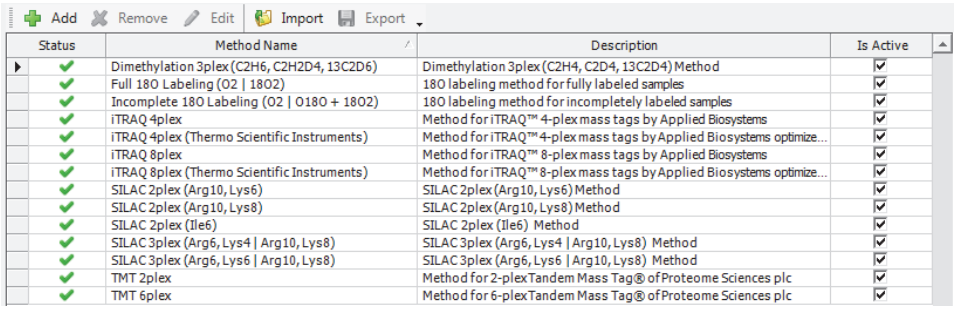
   –or–

   - Choose **Administration > Maintain Quantification Methods**, or click the **Maintain Quantification Methods** icon, .

     The Quantification Methods view appears, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

     Then, either double-click the row for the appropriate method in the Method Name or Description column, or click the column to the left of Method Name for the method, as shown in Figure 194, and click **Edit**.

     Figure 194. Selecting the method to edit

     The Quantification Method Editor dialog box appears, as shown in Figure 181 on page 272 through Figure 186 on page 280.

3. Follow the procedure in "Setting Up the Quantification Method" on page 264.

The Proteome Discoverer application checks the parameters that you have changed to be sure that they conform to the guidelines given in "Checking the Quantification Method" on page 281. It does not apply the changes to a quantification method unless the method meets all these criteria.

The changes that you make to a quantification method only affect the method in the selected results report.

Removing a Quantification Method

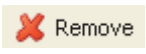
You can delete a quantification method if it is no longer useful, or make a quantification method temporarily unavailable to new workflows.

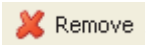
❖ To remove a quantification method

1. Choose **Administration > Maintain Quantification Methods**, or click the **Maintain Quantification Methods** icon,  .

   The Quantification Methods view opens, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

2. Click the box to the left of the method that you want to remove.

   The Remove icon,  , now becomes available.

3. Click  .

4. In the Delete Methods dialog box, click **OK**.

❖ To deactivate a quantification method

1. Choose **Administration > Maintain Quantification Methods**, or click the **Maintain Quantification Methods** icon,  .

   The Quantification Methods view opens, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

2. Clear the check box in the Is Active column on the line containing the quantification method that you want to render inactive.

To make the quantification method active again, select the same check box.

Importing a Quantification Method

You can import a new quantification method from another computer.

❖ **To import a quantification method**

1. Choose **Administration > Maintain Quantification Methods** or click the **Maintain Quantification Methods** icon, , either on the toolbar or on the Administration page.

   The Quantification Methods view opens, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

2. Click Import .

3. In the Import Quan Method dialog box, select the .method file containing the method that you want to import, and click **Open**.

   - If the new method is valid, the Quantification Method Editor dialog box opens, showing the new method.
   - If the new method is not valid, a message box appears that describes the error.

4. If the new method is valid, click **OK** in the Quantification Method Editor dialog box.

5. Change the name of the imported quantification method by changing it in the table of the Quan Method Manager.

Exporting a Quantification Method

You can save a quantification method to use on another computer.

❖ **To export a quantification method**

1. Choose **Administration > Maintain Quantification Methods** or click the **Maintain Quantification Method**s icon, , either on the toolbar or on the Administration page.

   The Quantification Methods view opens, as shown in Figure 177 on page 265. It lists all of the available methods for both precursor ion and reporter ion quantification.

2. Select the method that you want to export in the Quantification Methods view by clicking in the leftmost column.

3. Click Export .

4. In the Export Quan Method dialog box, select the name of the .method file containing the quantification method to be exported, and click **Save**.

Summarizing the Quantification

The Quantification Summary page summarizes the settings that you chose for the Precursor Ions Quantifier node or the Reporter Ions Quantifier node in the parameters pane of the Workflow Editor. It also shows the settings that you chose on the pages of the Quantification Method Editor for precursor ion and reporter ion quantification.

You must conduct a search with a workflow that includes a quantification node for this page to appear.

❖ To display the Quantification Summary page

- In an open MSF file, click the **Quantification Summary** tab.

  Figure 195 shows the Quantification Summary for precursor ion quantification, and Figure 196 shows the Quantification Summary for reporter ion quantification.

Figure 195. Quantification Summary page for precursor ion quantification

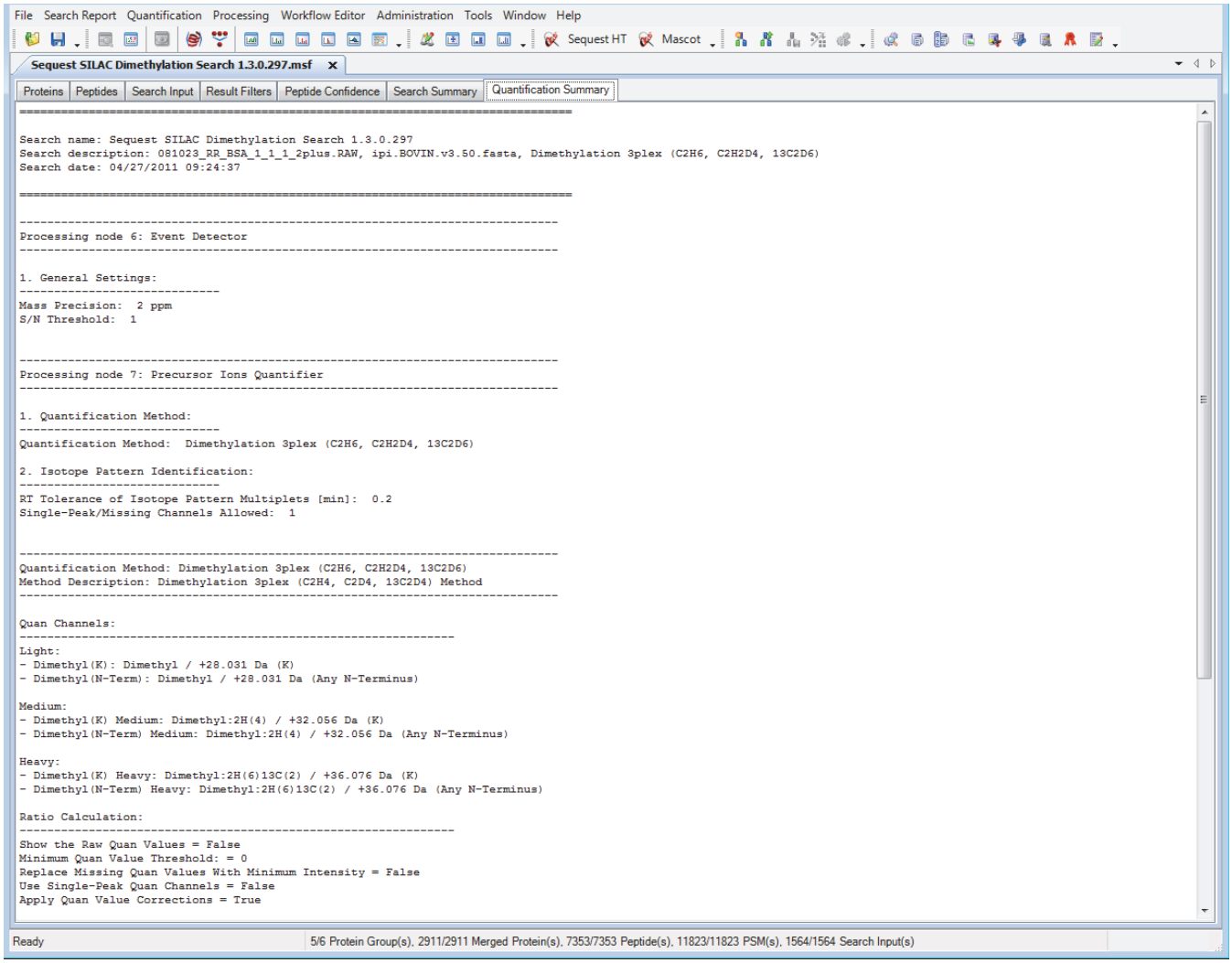
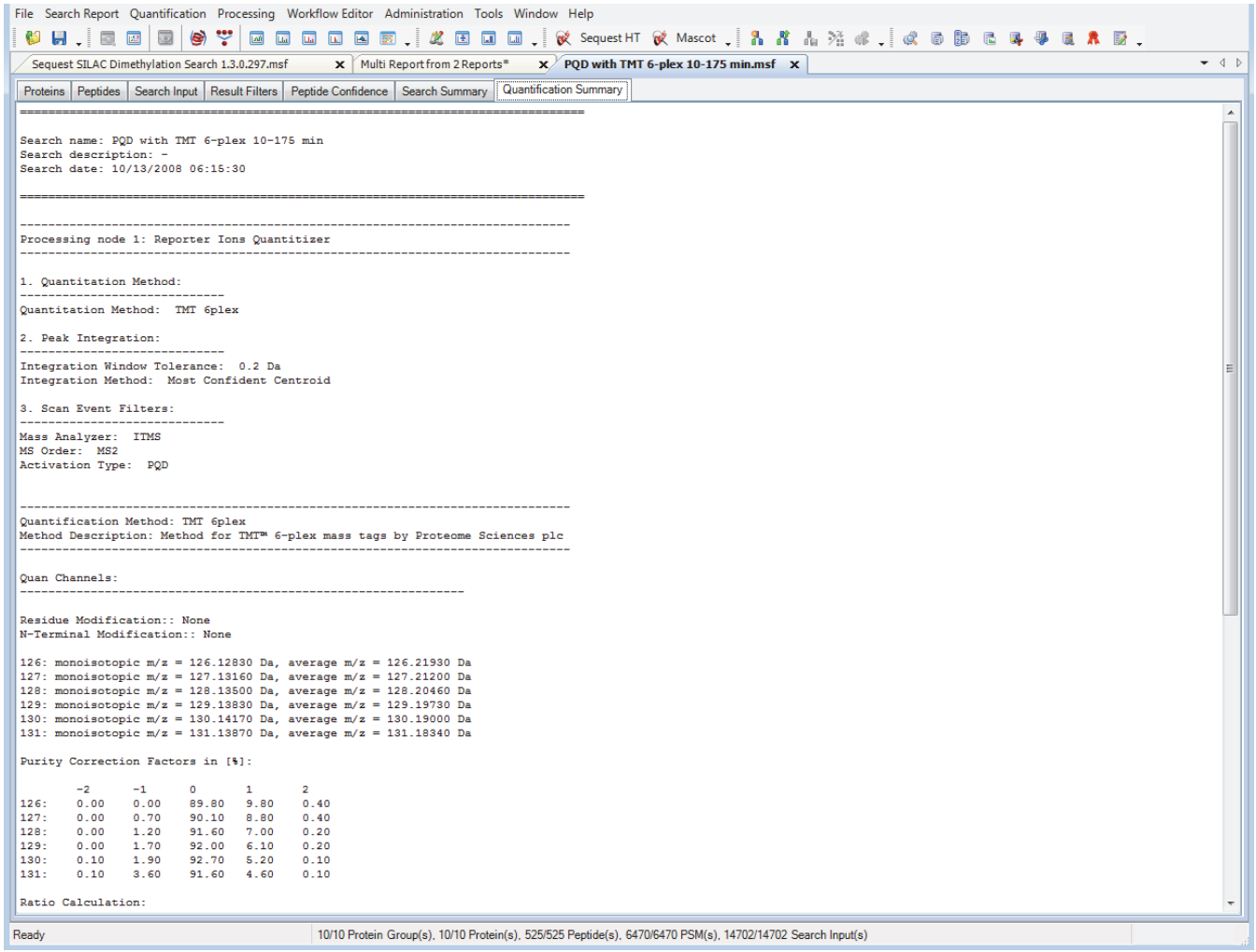


Figure 196. Quantification Summary page for reporter ion quantification



Displaying Quantification Spectra

After you perform reporter ion quantification, you can display the Quan Spectra page. This page displays the TMT intensities and ratios for all spectra in reporter ion quantification, regardless of whether they have been identified.

❖ To display the Quan Spectra page

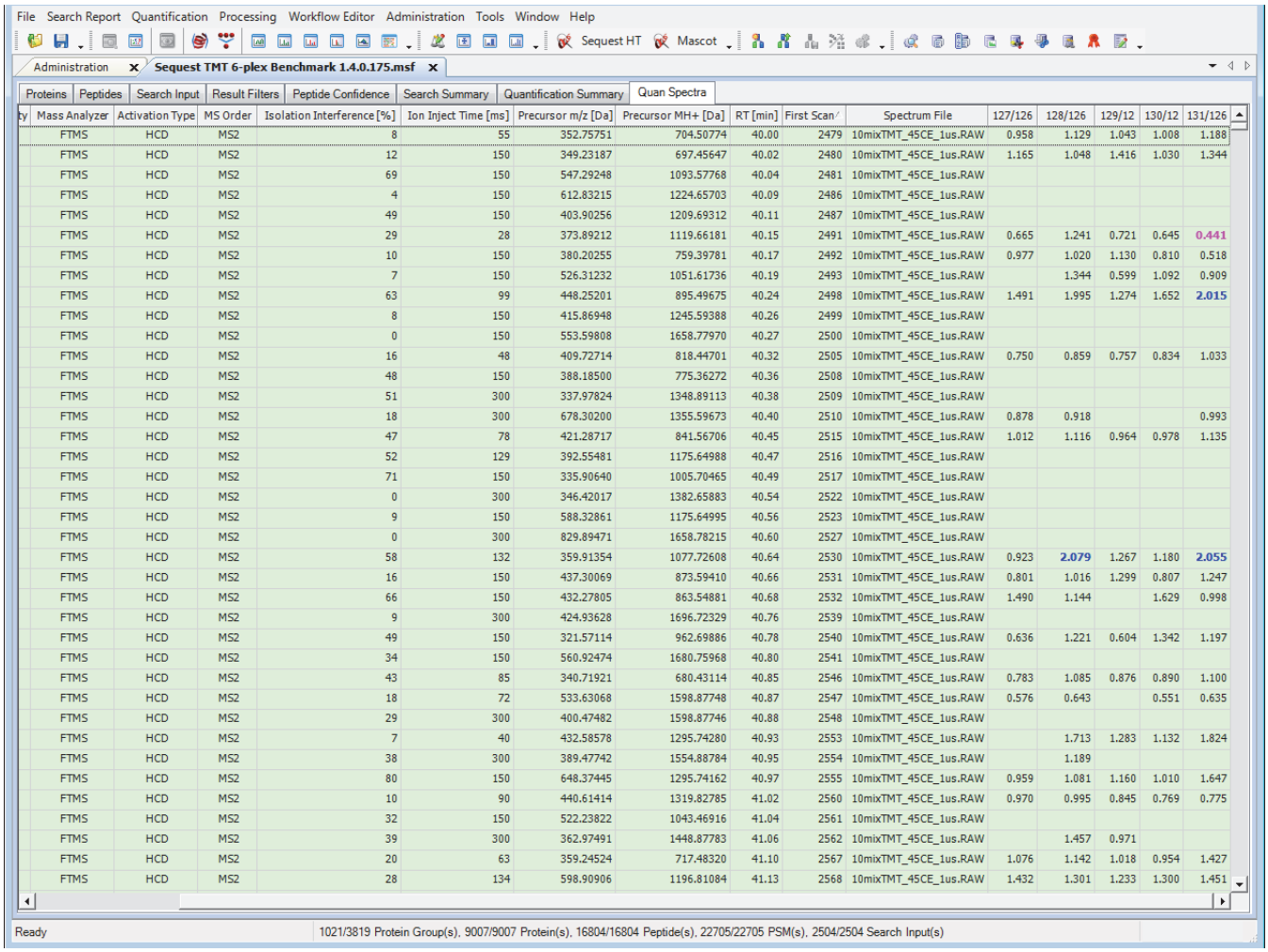
1. Perform reporter ion quantification.
2. Choose **File > Open Report** to open the resulting MSF file.
3. On the Input files page, click **Add**.
4. In the Add Analysis File(s) dialog box, select the file to open, and click **Open**.
5. Select the **Show Quan Spectra on Separate Tab** check box.

This option generates the Quan Spectra page in the MSF report only if you included a Reporter Ion Quantification node in your workflow.

6. Click **Open**.

Figure 197 gives an example of the Quan Spectra page.

Figure 197. Quan Spectra page



Mass Analyzer	Activation Type	MS Order	Isolation Interference [%]	Ion Inject Time [ms]	Precursor m/z [Da]	Precursor MH+ [Da]	RT [min]	First Scan	Spectrum File	127/126	128/126	129/12	130/12	131/126
FTMS	HCD	MS2	8	55	352.75751	704.50774	40.00	2479	10mixTMT_45CE_1us.RAW	0.958	1.129	1.043	1.008	1.188
FTMS	HCD	MS2	12	150	349.23187	697.45647	40.02	2480	10mixTMT_45CE_1us.RAW	1.165	1.048	1.416	1.030	1.344
FTMS	HCD	MS2	69	150	547.29248	1093.57768	40.04	2481	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	4	150	612.83215	1224.65703	40.09	2486	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	49	150	403.90256	1209.69312	40.11	2487	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	29	28	373.89212	1119.66181	40.15	2491	10mixTMT_45CE_1us.RAW	0.665	1.241	0.721	0.645	0.441
FTMS	HCD	MS2	10	150	380.20255	759.39781	40.17	2492	10mixTMT_45CE_1us.RAW	0.977	1.020	1.130	0.810	0.518
FTMS	HCD	MS2	7	150	526.31232	1051.61736	40.19	2493	10mixTMT_45CE_1us.RAW		1.344	0.599	1.092	0.909
FTMS	HCD	MS2	63	99	448.25201	895.49675	40.24	2498	10mixTMT_45CE_1us.RAW	1.491	1.995	1.274	1.652	2.015
FTMS	HCD	MS2	8	150	415.86948	1245.59388	40.26	2499	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	0	150	553.59808	1658.77970	40.27	2500	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	16	48	409.72714	818.44701	40.32	2505	10mixTMT_45CE_1us.RAW	0.750	0.859	0.757	0.834	1.033
FTMS	HCD	MS2	48	150	388.18500	775.36272	40.36	2508	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	51	300	337.97824	1348.89113	40.38	2509	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	18	300	678.30200	1355.59673	40.40	2510	10mixTMT_45CE_1us.RAW	0.878	0.918			0.993
FTMS	HCD	MS2	47	78	421.28717	841.56706	40.45	2515	10mixTMT_45CE_1us.RAW	1.012	1.116	0.964	0.978	1.135
FTMS	HCD	MS2	52	129	392.55481	1175.64988	40.47	2516	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	71	150	335.90640	1005.70465	40.49	2517	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	0	300	346.42017	1382.65883	40.54	2522	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	9	150	588.32861	1175.64995	40.56	2523	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	0	300	829.89471	1658.78215	40.60	2527	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	58	132	359.91354	1077.72608	40.64	2530	10mixTMT_45CE_1us.RAW	0.923	2.079	1.267	1.180	2.055
FTMS	HCD	MS2	16	150	437.30069	873.59410	40.66	2531	10mixTMT_45CE_1us.RAW	0.801	1.016	1.299	0.807	1.247
FTMS	HCD	MS2	66	150	432.27805	863.54881	40.68	2532	10mixTMT_45CE_1us.RAW	1.490	1.144		1.629	0.998
FTMS	HCD	MS2	9	300	424.93628	1696.72329	40.76	2539	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	49	150	321.57114	962.69886	40.78	2540	10mixTMT_45CE_1us.RAW	0.636	1.221	0.604	1.342	1.197
FTMS	HCD	MS2	34	150	560.92474	1680.75968	40.80	2541	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	43	85	340.71921	680.43114	40.85	2546	10mixTMT_45CE_1us.RAW	0.783	1.085	0.876	0.890	1.100
FTMS	HCD	MS2	18	72	533.63068	1598.87748	40.87	2547	10mixTMT_45CE_1us.RAW	0.576	0.643		0.551	0.635
FTMS	HCD	MS2	29	300	400.47482	1598.87746	40.88	2548	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	7	40	432.58578	1295.74280	40.93	2553	10mixTMT_45CE_1us.RAW		1.713	1.283	1.132	1.824
FTMS	HCD	MS2	38	300	389.47742	1554.88784	40.95	2554	10mixTMT_45CE_1us.RAW		1.189			
FTMS	HCD	MS2	80	150	648.37445	1295.74162	40.97	2555	10mixTMT_45CE_1us.RAW	0.959	1.081	1.160	1.010	1.647
FTMS	HCD	MS2	10	90	440.61414	1319.82785	41.02	2560	10mixTMT_45CE_1us.RAW	0.970	0.995	0.845	0.769	0.775
FTMS	HCD	MS2	32	150	522.23822	1043.46916	41.04	2561	10mixTMT_45CE_1us.RAW					
FTMS	HCD	MS2	39	300	362.97491	1448.87783	41.06	2562	10mixTMT_45CE_1us.RAW		1.457	0.971		
FTMS	HCD	MS2	20	63	359.24524	717.48320	41.10	2567	10mixTMT_45CE_1us.RAW	1.076	1.142	1.018	0.954	1.427
FTMS	HCD	MS2	28	134	598.90906	1196.81084	41.13	2568	10mixTMT_45CE_1us.RAW	1.432	1.301	1.233	1.300	1.451

Quan Spectra Page Parameters

The parameters on the Quan Spectra page are basically the same as those on the Search Input page (refer to the Help). However, they also include reporter ion quantification ratio columns that display the corrected ratio of the intensity of the fragmented tag in a sample to the intensity of the fragmented tag in the control sample for all spectra, regardless of whether they have been identified.

The Proteome Discoverer application generates the Quan Spectra page only if you included a Reporter Ion Quantification node in your workflow and select the Show Quan Spectra on Separate Tab check box on the Result Filters page when you open an MSF file. For more information on generating this page, refer to the Help.

Displaying the Quantification Channel Values Chart

You can generate a chart that displays the absolute intensity (for reporter ion quantification) or the area (for precursor ion quantification) of the quantification values detected for the available quantification channels.

❖ **To display the quantification channel values chart**

1. Click the row of the peptide that interests you.

   To obtain meaningful results, "Used" must appear in the Quan Info column of the report.

2. Choose **Quantification > Show Quan Channel Values**, or click the **Show Quan Channel Values** icon,  .

To see the results, see the following sections:

- Displaying Quantification Channel Values for Reporter Ion Quantification
- Displaying Quantification Channel Values for Precursor Ion Quantification

Displaying Quantification Channel Values for Reporter Ion Quantification

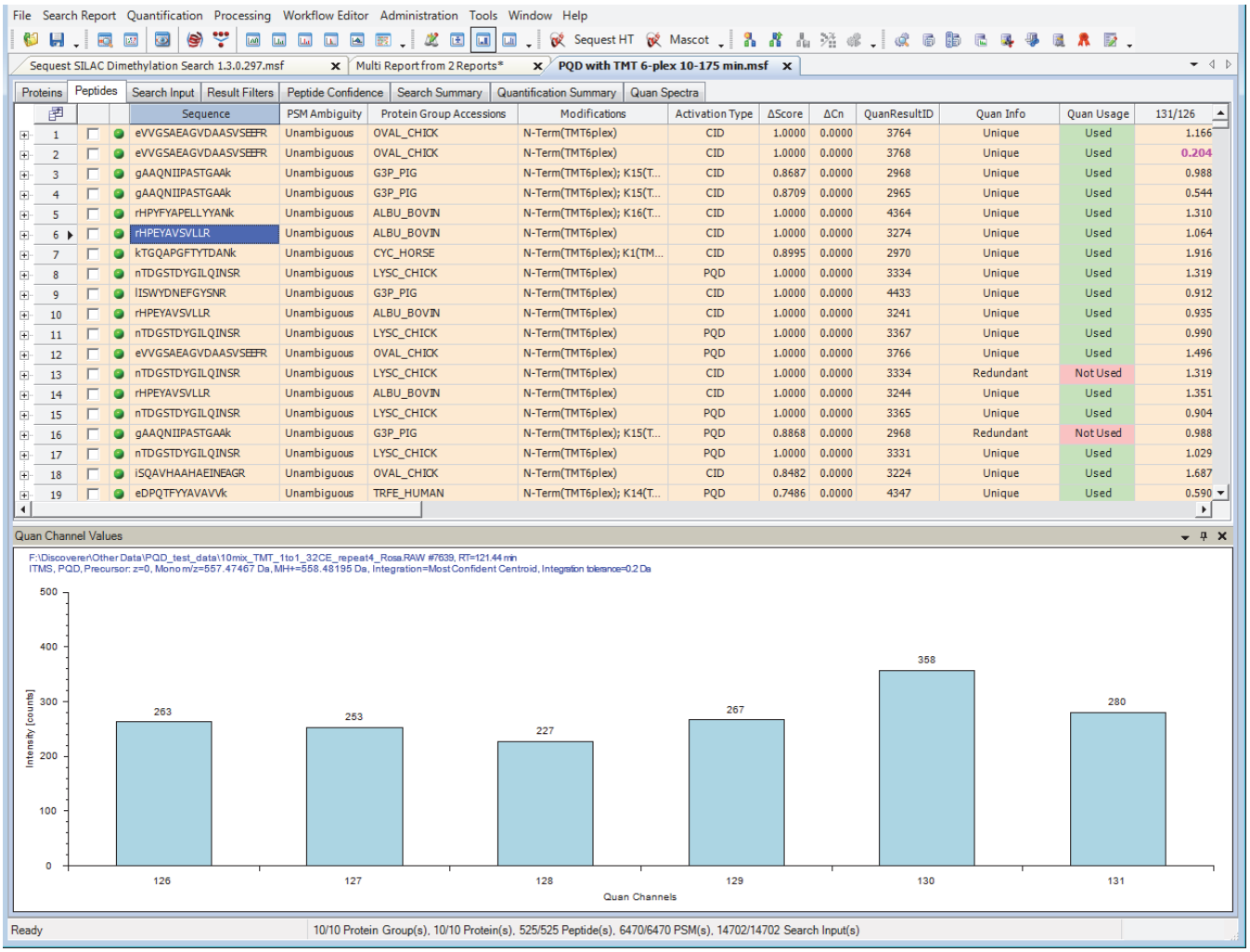
For reporter ion quantification, you can generate a Quan Channel Values chart that displays the absolute intensity of the reporter ions detected for the available quantification channels.

Reporter ions, or reporters, are the labels affixed to peptide samples in reporter ion quantification. They fragment in the MS/MS process. You can use the quantification value intensity to calculate the relative ratio of a peptide. You might also want to view the absolute quantification value intensity to verify that the peptide ratio calculation is correct.

The *x* axis of the chart shows the names of the quantification channels, and the *y* axis shows the intensity of the reporter ions, in counts.

The 4plex quantification method in iTRAQ has four reporter ions. Suppose that they are used to label four biological samples: 114, 115, 116, and 117. Figure 198 shows the Quan Channel Values chart created by the Show Quan Channel Values command for these samples. It shows the relative intensities of the samples labeled with the 114, 115, 116, and 117 reporter ions. Clearly, the sample labeled 115 is the sample with the greatest reporter ion intensity.

Figure 198. Quan Channel Values chart for reporter ion quantification



Displaying Quantification Channel Values for Precursor Ion Quantification

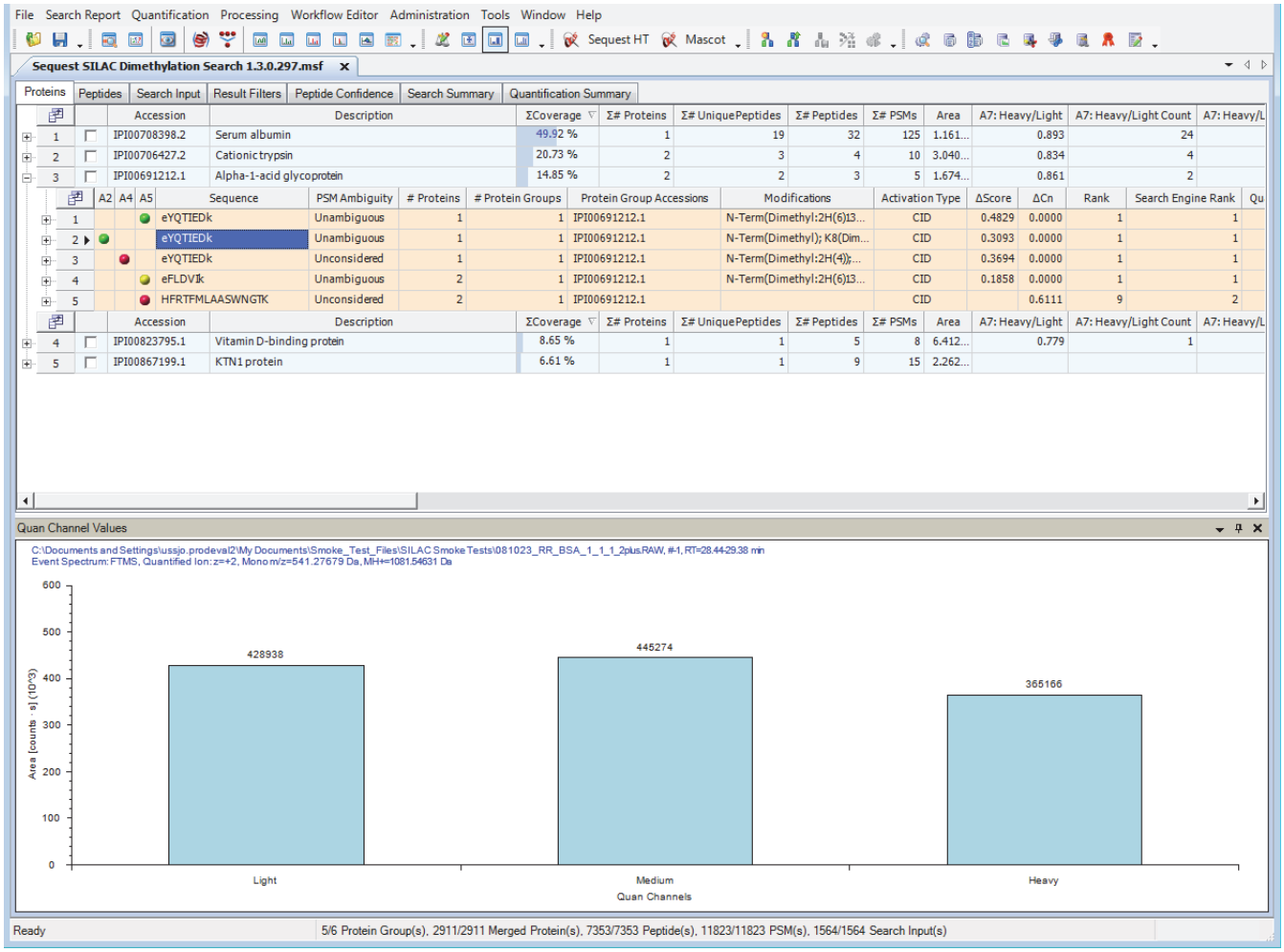
For precursor ion quantification, you can generate a Quan Channel Values chart that displays the area of the isotopes detected for the available quantification channels.

Heavy isotopes are incorporated into proteins in precursor ion quantification. You can use the quantification value area to calculate the relative ratio of a peptide. You might also want to view the quantification value area to verify that the peptide ratio calculation is correct.

The *x* axis of the chart shows the quantification channels, and the *y* axis shows the detected area for the given quantification channel, defined by counts per minute.

The dimethylation 3plex quantification method in SILAC has a sample labeled with the Light isotope, a sample labeled with the Medium isotope, and a sample labeled with the Heavy isotope. Figure 199 shows the chart created by the Show Quan Channel Values command for these samples. It shows the relative area of the samples labeled with the Light, Medium, and Heavy isotopes. The sample labeled Medium is the sample with the greatest area.

Figure 199. Quan Channel Values chart for precursor ion quantification



Displaying the Quantification Spectrum Chart

You can generate a chart showing the spectrum used for quantification. This chart is available for every peptide with an associated quantification result.

❖ To display the Quantification Spectrum chart

1. Select the peptide of interest. If Show Peptide Groups is already selected, you might need to ungroup the peptides first by right-clicking and choosing **Show Peptide Groups**.

   The peptide must be labeled “Used” in the Quan Usage column of the Peptides page.

2. Choose **Quantification > Show Quantification Spectrum**, or click the **Show Quantification Spectrum** icon,  .

To see the results, see the following sections:

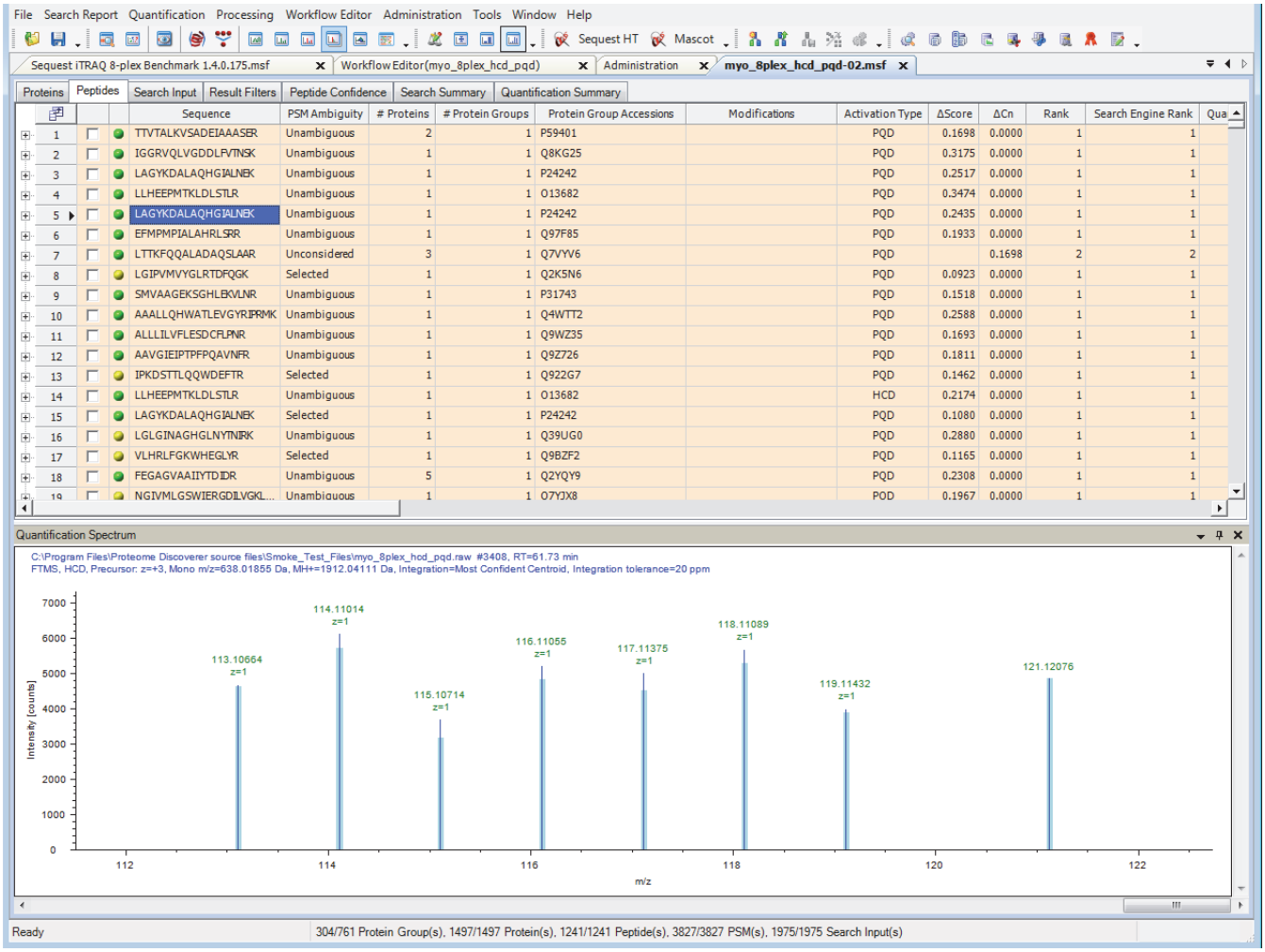
- Displaying the Quantification Spectrum Chart for Reporter Ion Quantification
- Displaying the Quantification Spectrum Chart for Precursor Ion Quantification

Displaying the Quantification Spectrum Chart for Reporter Ion Quantification

For reporter ion quantification, the Quantification Spectrum chart displays the intensity of the reporter ions, in counts. It shows a spectrum for each peptide, except for those peptides labeled “No Quan Values.”

Figure 200 shows an example of a quantification spectrum from an iTRAQ 8plex sample quantified with an Integration Tolerance setting (in the Reporter Ions Quantifier node) of 0.3 Da for extracting the reporter peaks from the quantification spectrum.

Figure 200. Quantification Spectrum chart for an iTRAQ 8plex sample using Proteome Discoverer scans



The Quantification Spectrum chart includes the following features:

- The light blue boxes represent the integration windows for the reporter tags. The boxes are centered on the masses of the reporter tags, as specified in the quantification method. The width of the boxes is the integration window used for extracting the reporter tags. It is ±0.3 Da, as specified by the settings of the parameters in the Reporter Ions Quantifier node (you can look up all these values on the Quantification Summary page). The height of the line in the box represents the actual tag intensity used for calculating the peptide ratios. The height of the box represents the corrected tag intensity. The height depends on the setting of the Integration Method parameter specified in the Reporter Ions Quantifier node. It is always the value that results from correction for isotopic impurities, as specified in the Reporter Ion Isotopic Distribution area of the Quan Channels page of the Quantification Method Editor dialog box, shown in Figure 181 on page 272.
- To calculate the actual intensity of a particular tag, the Proteome Discoverer application chooses the blue fragment peaks from the spectrum, and considers only peaks in the integration window.

- The black fragment peaks represent peaks that are present in the spectrum but that are not chosen for calculating the tag intensities. They might not be chosen because the peaks lie outside of any integration window, or because the setting of the Integration Method parameter specified in the Reporter Ions Quantifier node determined that only one peak per integration window should be chosen from any integration window. A different peak was picked for this integration window according to the criterion specified by the Integration Method setting.

Missing Reporter Peaks in the Quantification Spectrum

If reporter ions are missing in the quantification spectra, you can use settings on the Ratio Calculation page of the Quantification Method Editor dialog box to influence how the Proteome Discoverer application handles this problem. For example, if all six intensities of an TMTe 6plex are missing, or if the reference ion is missing (for example, in the TMTe 6plex method shown in Figure 201, the 126 ion is missing), the corresponding spectrum is always excluded from the protein quantification. In the Quan Info column of the Peptides page, these peptides are marked “No Quan Values” as shown in Figure 201. The protein ratios were calculated according to the settings displayed in Figure 202 on page 302.

Figure 201. Quantification results with missing reporter ions

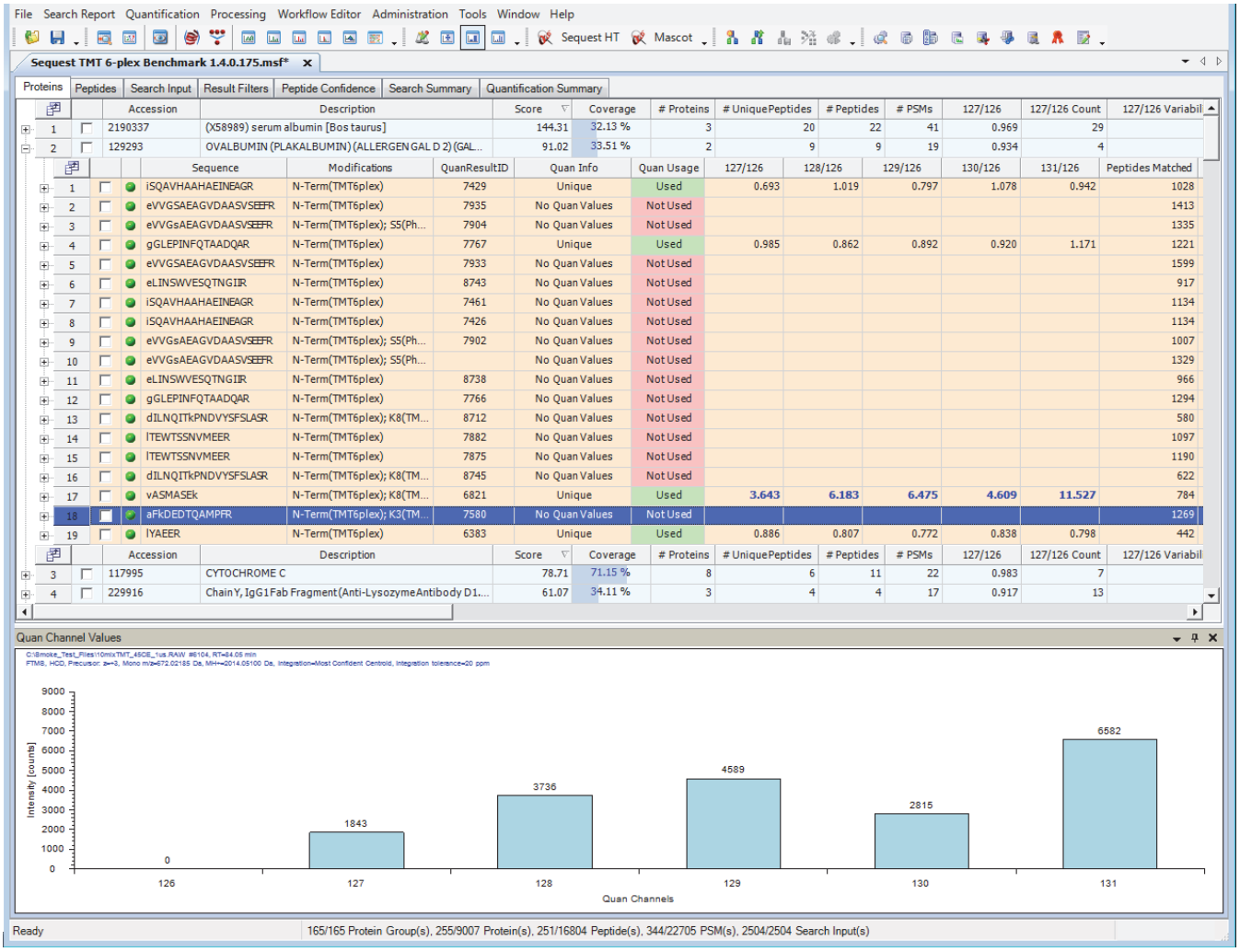
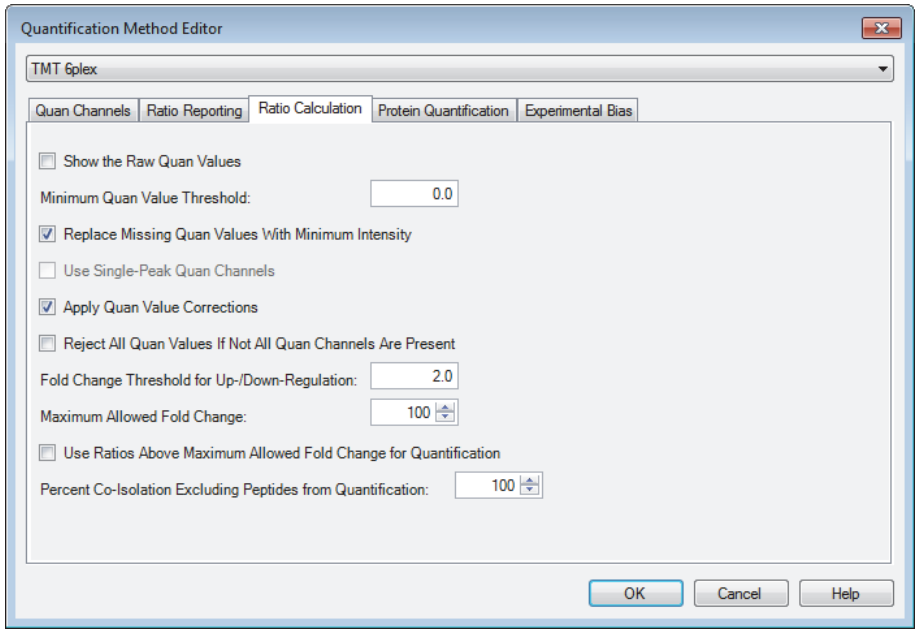


Figure 202. Ratio Calculation page of the Quantification Method Editor dialog box



For information about the options on the Ratio Calculation page of the Quantification Method Editor dialog box, see "Setting Up the Ratio Calculation" on page 275 or refer to the Help.

If one or more of the reporter, or mass, tags are missing in the quantification spectrum, the calculated ratios are either zero or infinity, depending on which tag intensity is the numerator and which is the denominator. Even if all tags are present, the calculated ratios might be very high or very low. You can use the Maximum Allowed Fold Change option on the Ratio Calculation page of the Quantification Method Editor dialog box to replace such extremely high or extremely low ratios with the maximum allowed number of times that the ratios can be multiplied. In the example in Figure 202, the maximum allowed number of times that the ratio can be multiplied is 100. That is, extremely high ratios are replaced by 100, and extremely low ratios are replaced by 0.01. You can expect an inherent dynamic range to be valid or detectable with the given instrumentation and method.

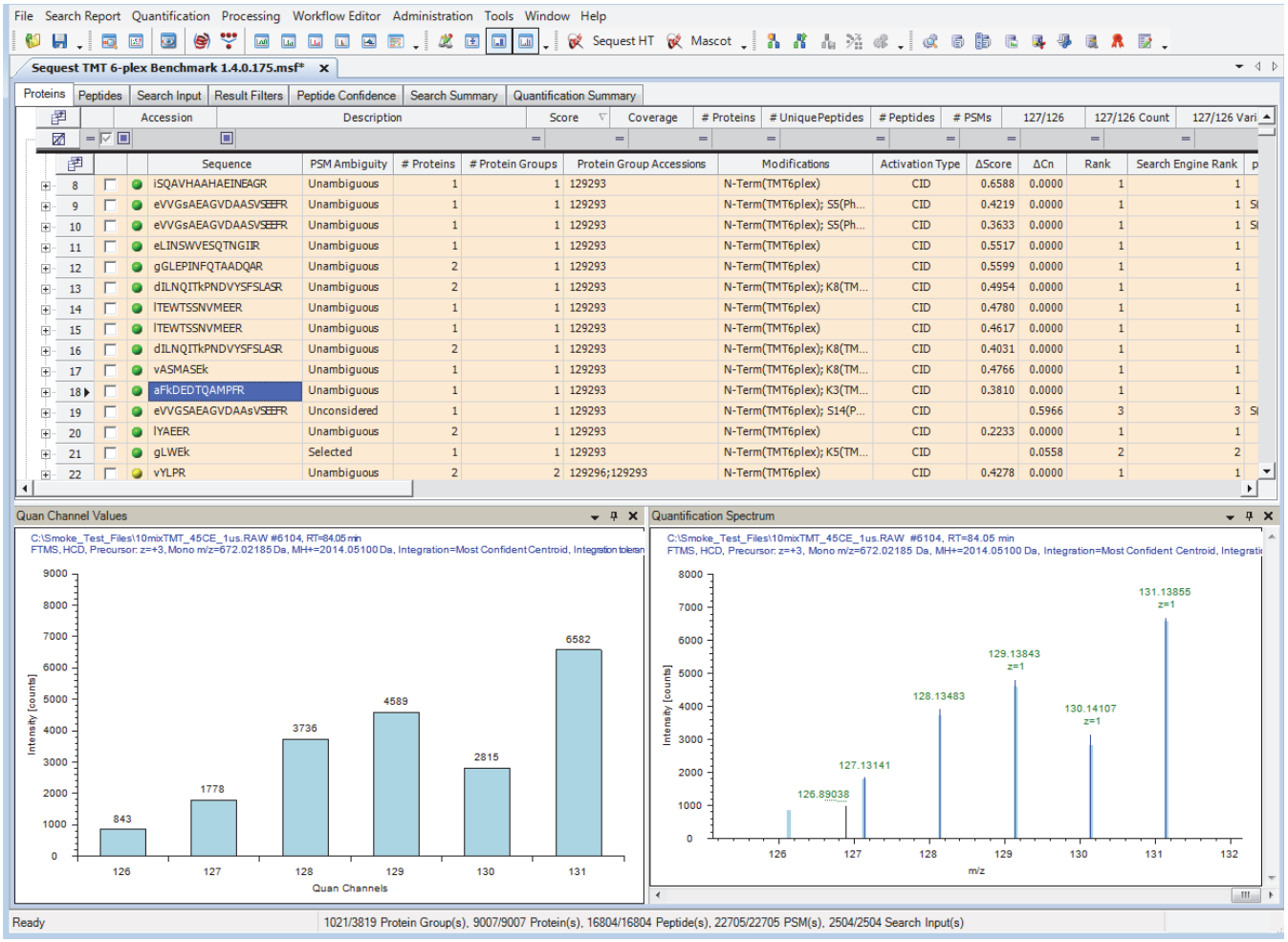
The Use Ratios Above Maximum Allowed Fold Change for Quantification option in the Ratio Calculation dialog box specifies whether such maximum calculated ratios should be considered when the Proteome Discoverer application calculates the protein ratios. You can use this option to automatically include extreme values when the application calculates the protein ratios. Since the protein ratios are calculated as the median, outlier protein ratios are likely to occur only if you have a sufficient number of peptides to use for protein quantification.

In the example in Figure 201 on page 301, the protein is identified with 30 peptides. For 19 of the peptides, the corresponding quantification spectra show no reporter tag at all. These peptides are never considered in calculating the protein ratios and are marked as “No Quan Values” in the Quan Info column of the Peptides page. Two additional spectra are missing individual reporter ions. Although two peptides are marked as “Used” for quantification in the Quan Info column, their extreme ratios are not considered in the protein ratio calculation (with the settings in Figure 202 on page 302).

If at least one of the reporter intensities is present (see Figure 201), you can use the Replace Missing Reporter Intensities With Minimum Intensity option on the Ratio Calculation page of the Quantification Method Editor dialog box to replace the missing intensities with the minimum intensity detected among all spectra on all reporter channels. Reporter intensities are missing because they fall under the detection limit, so replacing them with an intensity estimate that is close to the detection limit might make sense.

Figure 203 shows the same protein as in Figure 201 on page 301 after the selection of the Replace Missing Reporter Intensities with Minimum Intensity option. In the example, the 126 reporter ion has been replaced with a minimum intensity value. This is not exactly the true value, but it is better than having no estimates for the ratios of this protein. Whether this option gives valuable results for you depends on your experimental design and quantification strategy.

Figure 203. Quantification results after applying the Replace Missing Reporter Intensities with the Minimum Intensity option

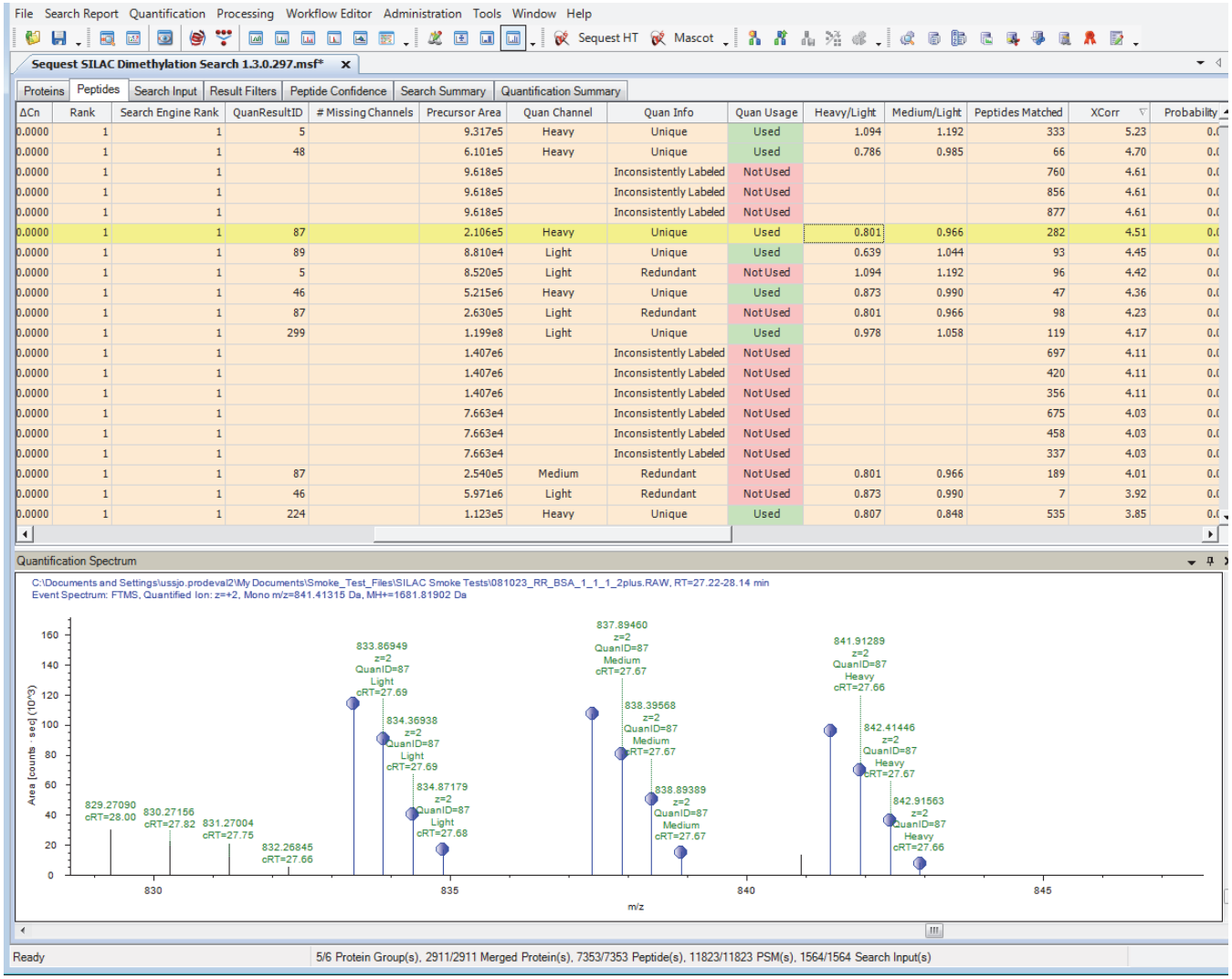


You exclude spectra with one or more missing reporter peaks from the protein ratio calculation by selecting the Reject All Quan Values If Not All Quan Channels Are Present option on the Ratio Calculation page.

Displaying the Quantification Spectrum Chart for Precursor Ion Quantification

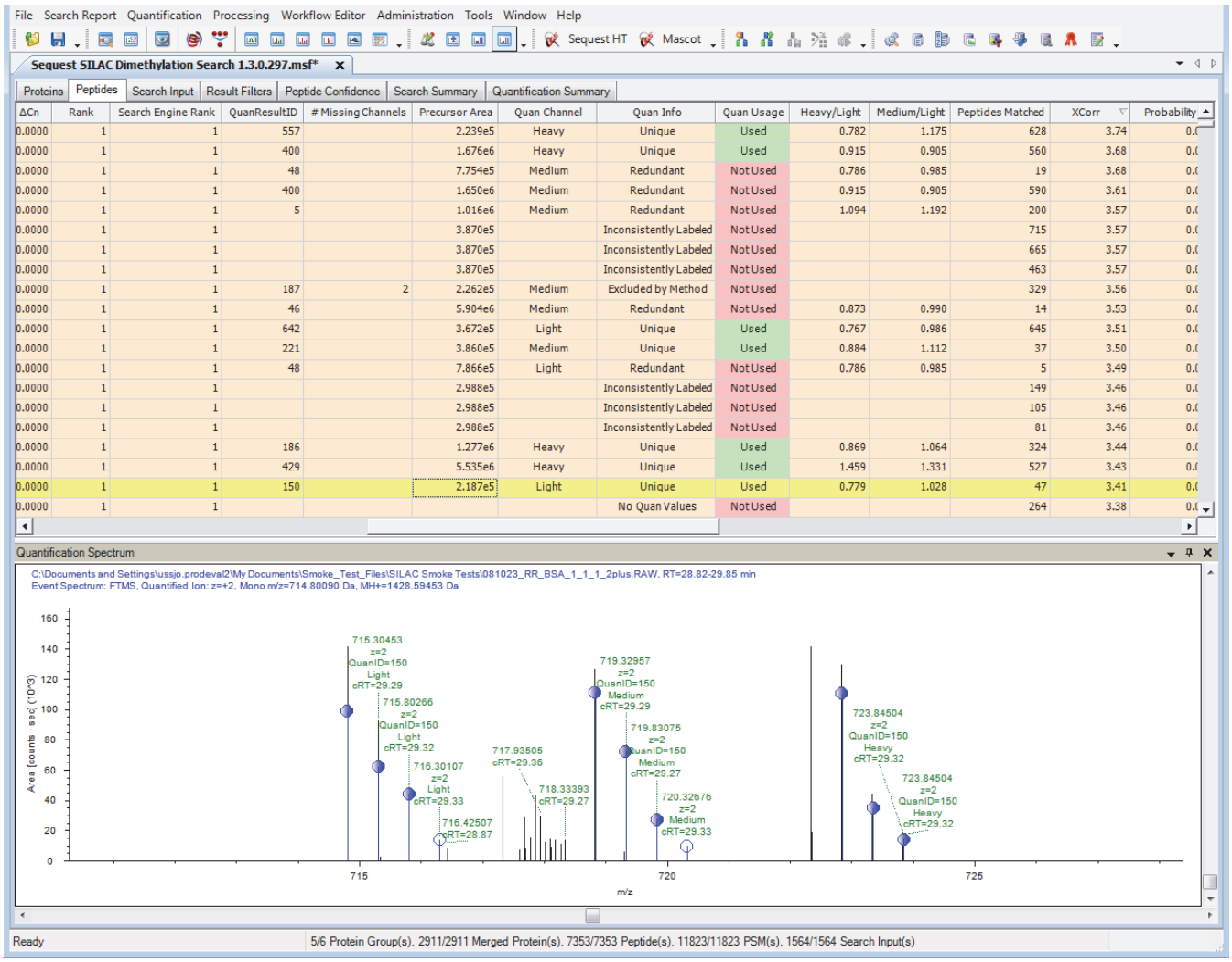
For precursor ion quantification, the Quantification Spectrum chart displays a quantification spectrum for each peptide. It also displays the different abundances of the identified Light, Medium, and Heavy isotopic peak patterns used to quantify a peptide. The abundances are measured by calculating the area of the extracted ion chromatogram of each isotope of a pattern. The chart highlights the corresponding isotope pattern peaks and labels them with the quantification channel names, as shown in Figure 204. It also includes any peaks that are not part of an isotope pattern.

Figure 204. Quantification Spectrum chart for precursor ion quantification



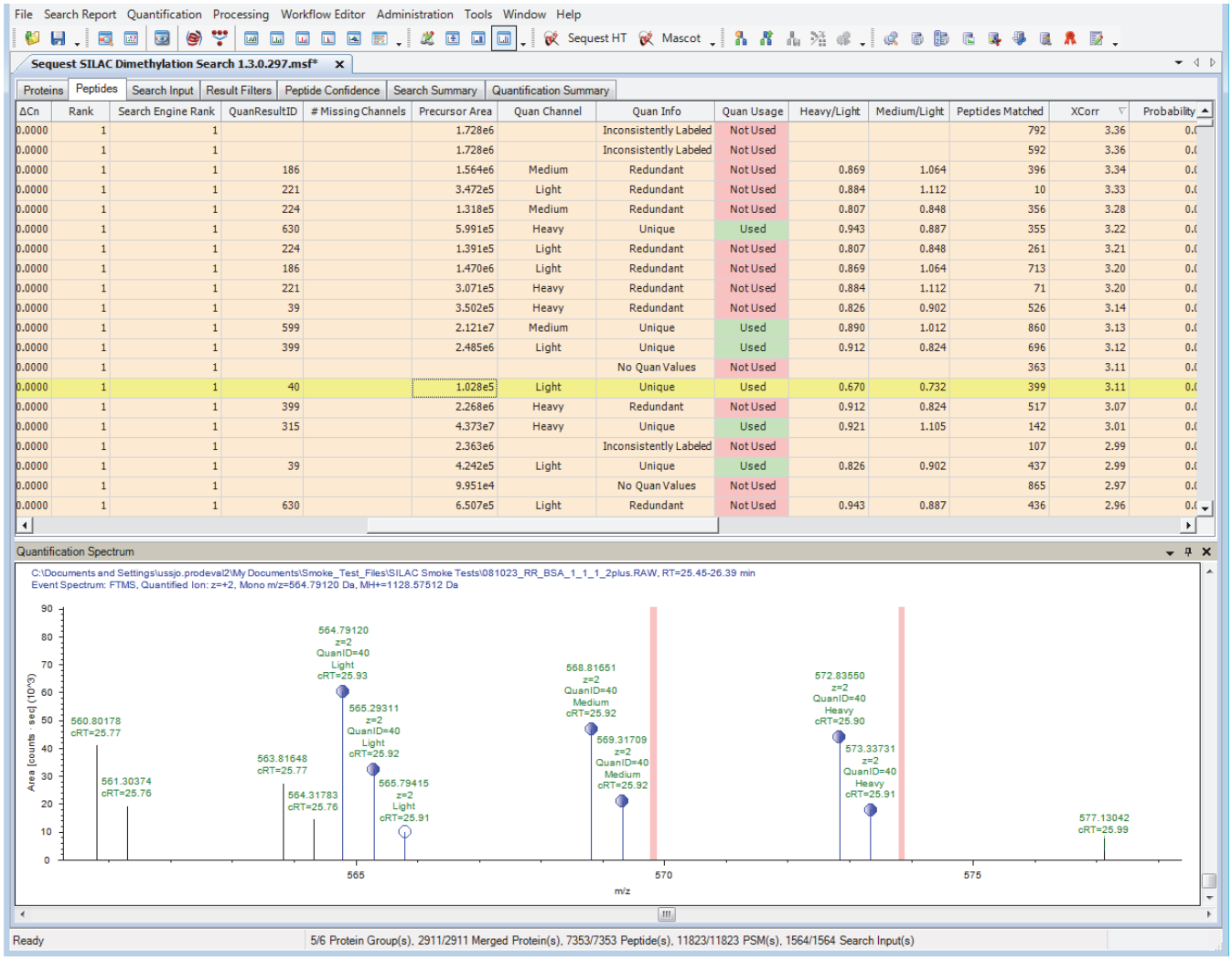
The *x* axis of the chart displays the mass-to-charge ratio of the isotopes, and the *y* axis displays the area of the extracted ion chromatogram for the isotopes. Filled blue circles mark the isotope pattern peaks that were used for calculating the quantification values for the different quantification channels. Unfilled blue circles mark the isotope pattern peaks that were identified but not used. The Quantification Spectrum chart always compares the exact same isotopic pattern peaks for each label. For example, the chart in Figure 205 compares the first three isotopic pattern peaks among all three types: Light, Medium, and Heavy. But the chart also contains an additional Light isotopic pattern peak and an additional Heavy isotopic pattern peak that are not used, so these two peaks are represented by unfilled circles.

Figure 205. Extra isotopic pattern peaks represented by unfilled circles in the Quantification Spectrum chart



The Quantification spectrum chart can also indicate whether an expected quantification pattern peak is absent. Regions in pink indicate where a quantification pattern peak was expected but is absent; Figure 206 shows these regions. This ion pattern peak is not used in calculating the quantification values for the different quantification channels.

Figure 206. Expected but absent peak in the Quantification Spectrum chart



Regions in blue, shown in Figure 207, indicate where a quantification pattern peak was expected but is unsuitable. Pattern peaks might be unsuitable because of the wrong centroid retention time, a range out of the delta mass, the wrong intensity, or a peak that has been used by another isotopic pattern. This ion pattern peak is not used in calculating the quantification values for the different quantification channels.

Figure 207. Expected but unsuitable peaks in the Quantification Spectrum chart

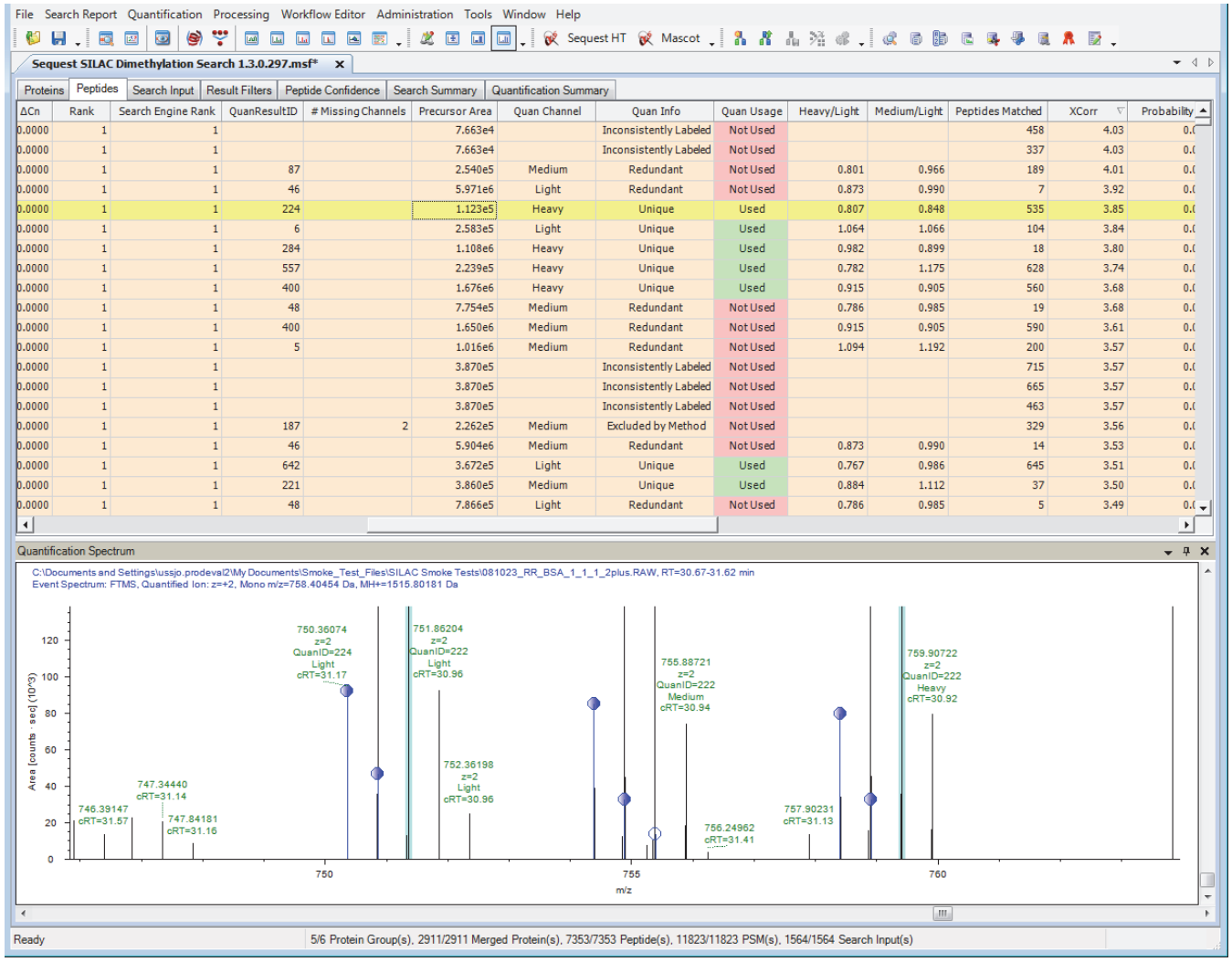


Table 22 shows what the various colors mean on the Quantification Spectrum charts in Figure 206 on page 306 and Figure 207.

Table 22. The meaning of colors in the Quantification Spectrum chart (Sheet 1 of 2)

Color	Meaning
Filled blue circle	Indicates the isotope pattern peaks that are used in calculating the quantification values for the different quantification channels.
Unfilled blue circle	Indicates the isotope pattern peaks that are not used in calculating the quantification values for the different quantification channels.

Table 22. The meaning of colors in the Quantification Spectrum chart (Sheet 2 of 2)

Color	Meaning
Yellow box	Indicates that the pattern includes peaks from only one channel. This ion pattern peak is not used in calculating the quantification values for the different quantification channels.
Pink bar	Indicates that a quantification pattern peak is expected but is missing. This ion pattern peak is not used in calculating the quantification values for the different quantification channels.
Blue bar	Indicates that a quantification pattern peak is present but is unsuitable because of errors in peptide labeling or because of the wrong centroid retention time, a range out of the delta mass, the wrong intensity, or a peak that has been used by another isotopic pattern. This ion pattern peak is not used in calculating the quantification values for the different quantification channels.

Using Reporter Ion Isotopic Distribution Values To Correct for Impurities

iTRAQ and TMT kits consist of labels that contain different numbers of ^{13}C atoms, ^{15}N atoms, or both. For simplicity, assume that a 4plex kit yields peaks at 114, 115, 116, and 117 *m/z*, which correspond to $^{13}C1$, $^{13}C2$, $^{13}C3$, and $^{13}C4$, respectively. Because the label substances are not 100 percent isotopically pure, each label contains a certain number of other atoms. For example, the 116 label would not consist only of label molecules having three ^{13}C atoms but might also contain label molecules with only one or two ^{13}C atoms or even four or five ^{13}C atoms. As a result, these impurities lead to an observed peak of 116 *m/z*, which is smaller than might be expected if the tag were 100 percent isotopically pure, and to additional peaks at positions –2, –1, +1, +2 Da apart from 116 *m/z*. The intensities of the latter peaks are proportional to the amount of the described isotopic impurities. When the 116 label and the 114, 115, and 117 labels are used, these latter three labels contribute to the peak at 116 *m/z* because of their isotopic impurities.

The intensity of the peak at 116 *m/z* effectively includes the following contributions:

(*observed intensity 116*) = (*true intensity 116*) – (*intensity loss because of 116 impurities*) + (*intensity gain because of other label impurities*)

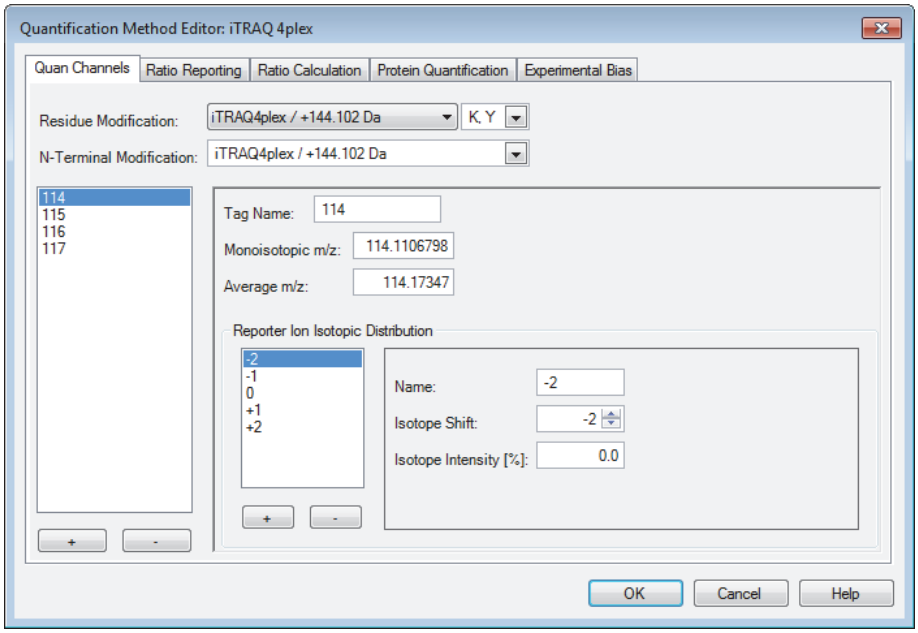
To obtain the true intensity value of the 116 label—that is, the amount of the substance initially labeled with the 116 tag—you must correct the experimentally observed peak for the impurity of the labels.

For a 4plex sample, there are four formulas that use the equation just given for each of the labels, and the proper correction would consider both contributions in the formula by solving the system of coupled linear equations:

(*intensity_of_loss_because_of_116_impurity* and *intensity_of_gain_because_of_other_label_impurities*)

For this correction, you must enter the isotopic distribution of each of the labels used in the quantification method, as shown in Figure 208. The values are part of each of the iTRAQ or TMT label kits used.

Figure 208. Entering values for the isotopic distribution of a specific reporter tag



You can also deconvolute the overlapping labels using other methods. Compatible with the Mascot search engine, the Proteome Discoverer application uses a first-order approximation to the solution. The error made is small when the intensities of all possible contributing labels are of similar height, and it becomes larger if the intensity differences become larger.

Excluding Peptides from the Protein Quantification Results

You can manually exclude and include certain peptides from the protein quantification results. You can also return excluded or included peptides to their default status.

You cannot include peptides if No Quan Values, Inconsistently Labeled, or Excluded by Method appears in the Quan Info column.

❖ To exclude a peptide from the quantification results

1. On the Proteins page, click the plus sign (+) next to the protein of interest to display its constituent peptides.

2. Right-click the peptide of interest, which must display "Used" in the Quan Usage column, and choose **Include/Exclude Peptide(s) from Protein Quantification > Exclude** from the shortcut menu.

   "Not Used (Excluded)" now appears in the Quan Usage column.

3. To save the information in the MSF file resulting from this setting, choose **File > Save Report**.

❖ To include an excluded peptide in the quantification results

1. On the Proteins page, click the plus sign (+) next to the protein of interest to display its constituent peptides.
2. Right-click the peptide of interest, which must display "Not Used (Excluded)," "Redundant," or "Not Unique" in the Quan Info column, and choose **Include/Exclude Peptide(s) from Protein Quantification > Include** from the shortcut menu.

   "Used (Included)" now appears in the Quan Usage column.

3. To save the information in the MSF file resulting from this setting, choose **File > Save Report**.

❖ To return an included or excluded peptide to its default status

1. On the Proteins page, click the plus sign (+) next to the protein of interest to display its constituent peptides.
2. Right-click the peptide of interest and choose **Include/Exclude Peptide(s) from Protein Quantification > Default** from the shortcut menu.

   The Quan Usage column now displays the peptide's usage status when the MSF file was first opened.

Excluding Peptides with High Levels of Co-Isolation

To create a fragment spectrum, you select a precursor mass for isolation, isolate and fragment the ions within a mass window that you define, and record the product ion masses created.

Ideally, you would isolate and fragment only the precursor ions of a single selected component. However, in practice you isolate the precursor ions within a user-specified window—typically 1 or 2 daltons around the isolation mass. Co-eluting components with a mass falling into this isolation window are also isolated and fragmented. This process is called co-isolation. The co-isolating components are likely to be peptides whose fragments are observed in the created fragment spectra. The co-isolation can exacerbate the identification of the selected peptide and lower the identification confidence.

Co-isolation is an issue in reporter ion quantification. In this type of quantification, the peptides from different charges of the same sample—for example, different treatment states—are modified with special isobaric labels. The isobaric labels disaggregate during precursor ions fragmentation and create reporter tags that appear in the low-mass region of the fragment spectra. You use the intensity ratio of the observed fragment tags for relative quantification of the peptides from the different sample charges.

The co-isolating peptides also create reporter tags that superimpose on the reporter tags of the selected peptide. Because most of the proteins in a real sample are unregulated, the co-isolated peptides often create reporter tags with equal intensity. If these superimpose on the reporter tags of a selected peptide of a regulated protein, the observed ratios of the reporter tags in the fragment spectra can be false. Furthermore, the perturbed ratios of the selected peptides that are greatly affected by co-isolation can also adversely affect the ratios that the Proteome Discoverer application calculates for the proteins that include these peptides.

Determining the extent to which the real reporter tag ratios of the selected peptides are perturbed is difficult. It depends on the level of co-isolation and the isolation characteristics of the instrument. The Proteome Discoverer application flags PSMs with a high level of co-isolation. For newly generated MSF files, it calculates and displays the percentage of interference within the precursor isolation window. This percentage is the relative amount of ion current within the isolation window that is not attributed to the precursor itself:

$$\%\_isolation\_interference = 100 \times \left[1 - \left(\frac{precursor\_intensity\_in\_isolation\_window}{total\_intensity\_in\_isolation\_window}\right)\right]$$

The application displays the calculated interference value in the % Isolation Interference column on the Peptides and Search Input pages. For reporter ion quantification, a high isolation interference value could indicate that a calculated peptide ratio is skewed by the presence of co-isolated peptide species.

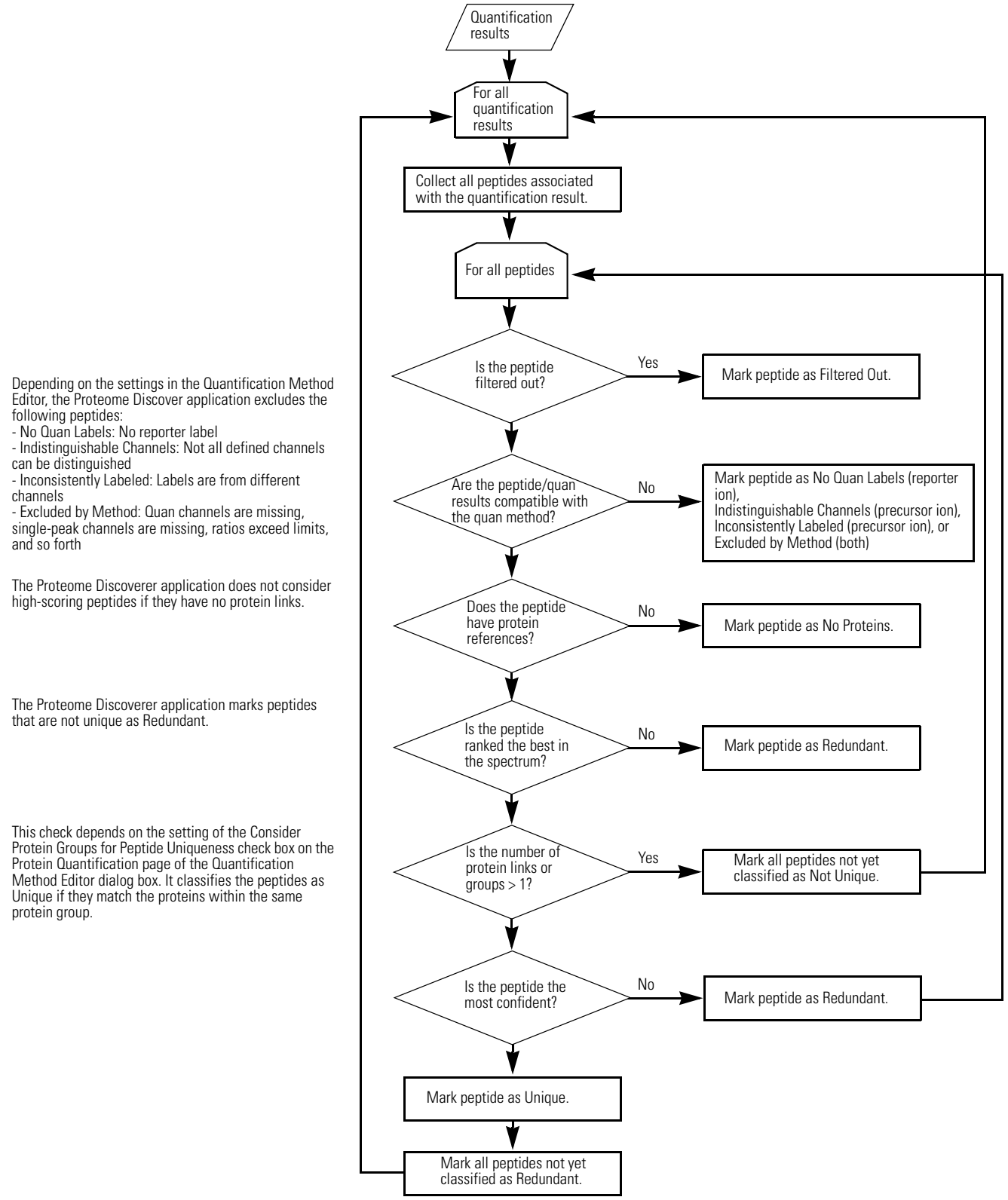
Note The Proteome Discoverer application only calculates the % Isolation Interference value if the precursor scans are high-resolution, high-mass-accuracy scans.

You can use the Percentage Co-Isolation Excluding Peptides from Quantification parameter on the Ratio Calculation page of the Quantification Method dialog box (shown in Figure 202 on page 302) to specify a threshold of between 0 and 100 percent for the allowed co-isolation interference. The default value is 100 percent, which means that no PSM is excluded. This parameter is only available for reporter ion quantification.

Classifying Peptides

The flowchart in Figure 209 shows how the Proteome Discoverer application classifies peptides for protein quantification. It displays this classification in the Quan Info column of the results report. Refer to the Help for descriptions of these classifications.

Figure 209. Classifying peptides for protein quantification



Depending on the settings in the Quantification Method Editor, the Proteome Discover application excludes the following peptides:
- No Quan Labels: No reporter label
- Indistinguishable Channels: Not all defined channels can be distinguished
- Inconsistently Labeled: Labels are from different channels
- Excluded by Method: Quan channels are missing, single-peak channels are missing, ratios exceed limits, and so forth

The Proteome Discoverer application does not consider high-scoring peptides if they have no protein links.

The Proteome Discoverer application marks peptides that are not unique as Redundant.

This check depends on the setting of the Consider Protein Groups for Peptide Uniqueness check box on the Protein Quantification page of the Quantification Method Editor dialog box. It classifies the peptides as Unique if they match the proteins within the same protein group.

Calculating Peptide Ratios

For both precursor ion and reporter ion quantification, the Proteome Discoverer application calculates protein ratios as the median, not the mean, of all peptide hits belonging to a protein that is marked “Used” in the Quan Usage column of the report. It chooses the median to calculate the protein ratios because it is relatively robust in the presence of outliers. In principle, the Proteome Discoverer application uses only the peptides in the filtered results for protein ratio calculation when the result filters are applied to the search result. These result filters are what you want to apply to quantification. For example, protein ratios that change because you filter peptides having a specific sequence tag will skew the results.

Protein ratios are the median of the peptides of the protein. If you want to recalculate the peptide ratio, you must ensure that all peptides are displayed. By default, the application considers only unique peptides in the calculation so that only peptides that have no other protein references are considered.

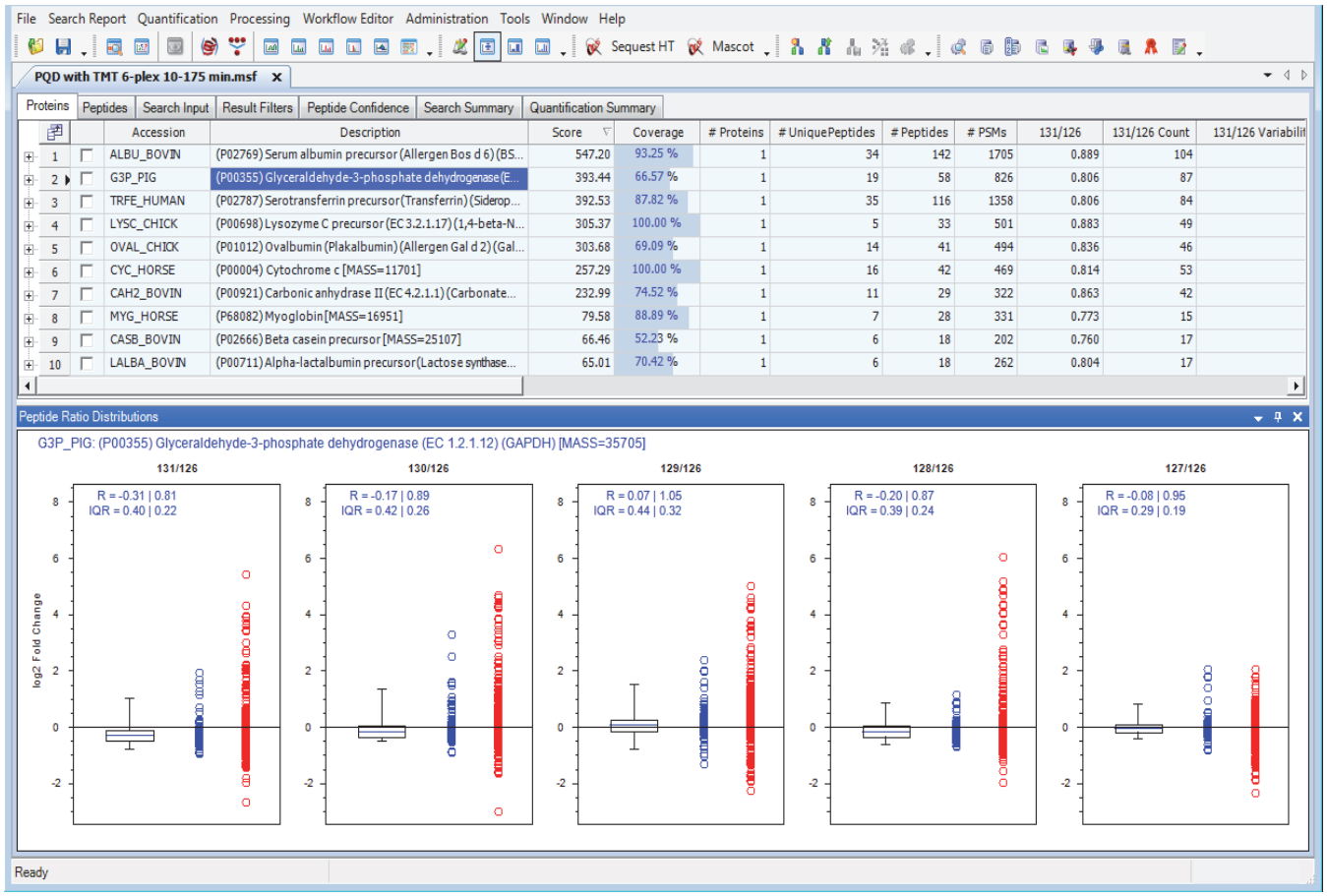
You can activate a chart of the peptide ratios. This graph shows the distribution of peptide ratios for the selected protein, displaying the ratios of the peptides associated with the selected protein as a log2-fold change.

❖ To calculate peptide ratios

1. Click the row of the peptide or protein that you are interested in.
2. Choose **Quantification > Show Peptide Ratios**, or click the **Show Peptide Ratios** icon,  .

The Peptide Ratio Distributions chart shown in Figure 210 appears. The following sections describe the pages available in this view.

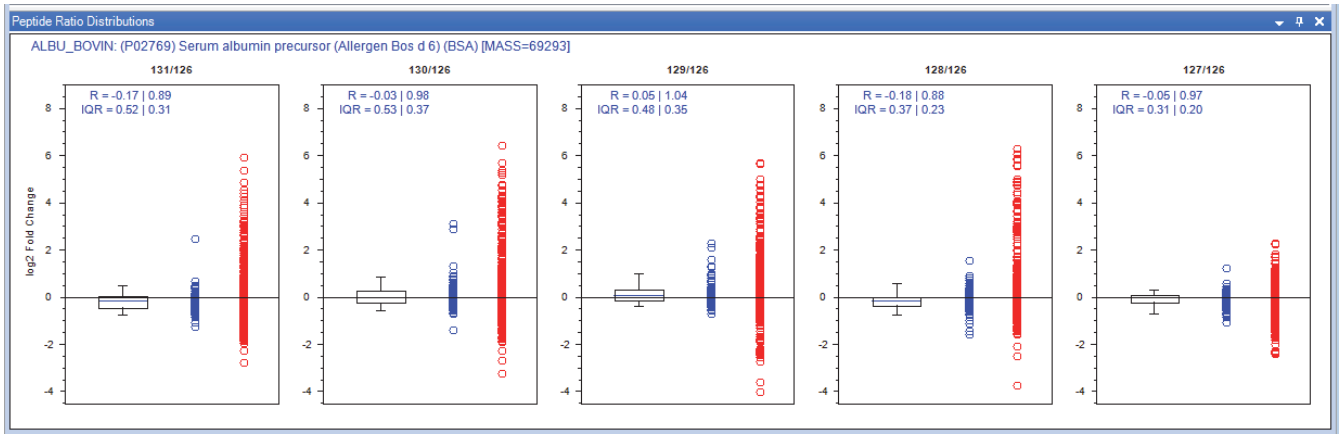
Figure 210. Peptide Ratio Distributions chart



Understanding the Peptide Ratio Distributions Chart

The Peptide Ratio Distributions chart shows the distribution and spread of the ratios of all peptides belonging to a particular protein. Figure 211 shows an example for the albumin protein.

Figure 211. Peptide Ratio Distributions chart

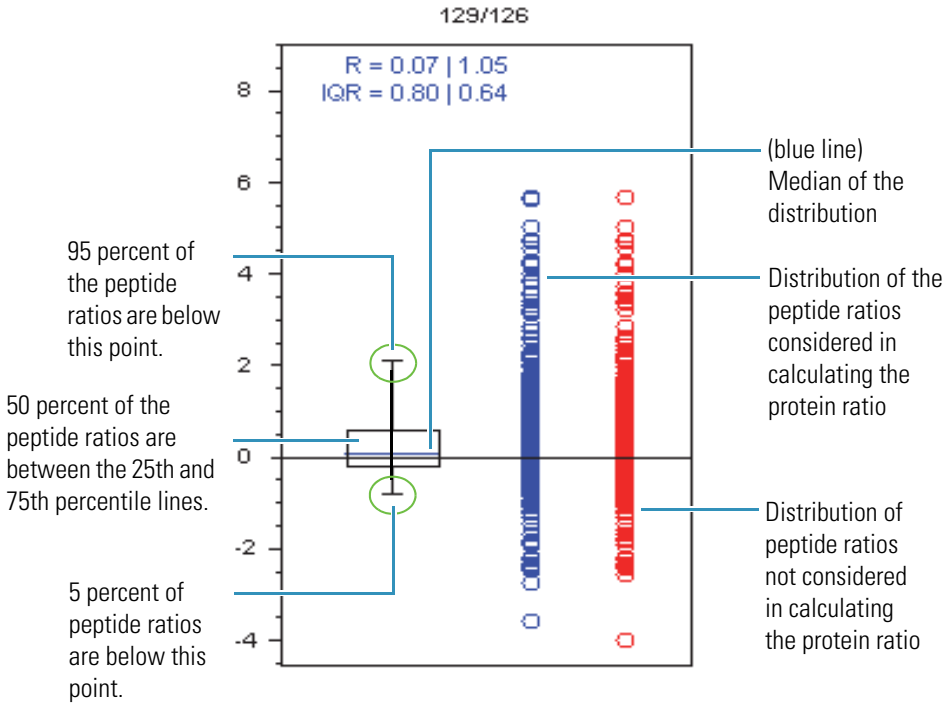


The chart shows the distribution of peptide ratios for each of the ratios reported, as defined in the quantification method for this search. Each of the ratio distribution charts displays the peptide ratios as the binary logarithm. The logarithmic form is common for such displays, because it provides a reasonable display, even when there is a large spread of the displayed values. In binary logarithmic form, a value of 1 means a two-fold increase, a value of 2 means a four-fold increase, a value of 3 means an eight-fold increase, and so forth. Each of the separate distribution charts displays the peptide ratios in three sections. The chart legend explains the meaning of these sections. You can access the chart legend by right-clicking the chart and choosing Show Legend.

The Peptide Ratio Distributions charts contain the three sections illustrated in Figure 212:

- The first section displays the distribution of the ratios of all peptides considered for calculating the ratio of this protein as a box-and-whisker plot. A box-and-whisker plot is a convenient way of graphically depicting groups of numerical data through a five-number summary: 5 percent lower bound, lower quartile, median, upper quartile, 95 percent upper bound. The range between the lower and upper quartile (this is the range of the box) is also known as the inter-quartile range (IQR) and, like the standard deviation for normally distributed data, is a measure of the spread of the data.
  - The box represents the peptide ratios between the 25th and the 75th percentiles.
  - The error bars represent the peptide ratios below the 5th and the 95th percentiles.
  - The blue lines inside the horizontal bar represent the median of the distribution.
- The second section (blue circles) displays the distribution of the ratios of all peptides considered in calculating the protein ratio.
- The third section (red circles) displays the distribution of the ratios of all peptides that were not considered in calculating the protein ratio (for example, the peptide ratio was considered too extreme, or this peptide is not unique to this protein or this protein group) according to the rules defined in the quantification method.

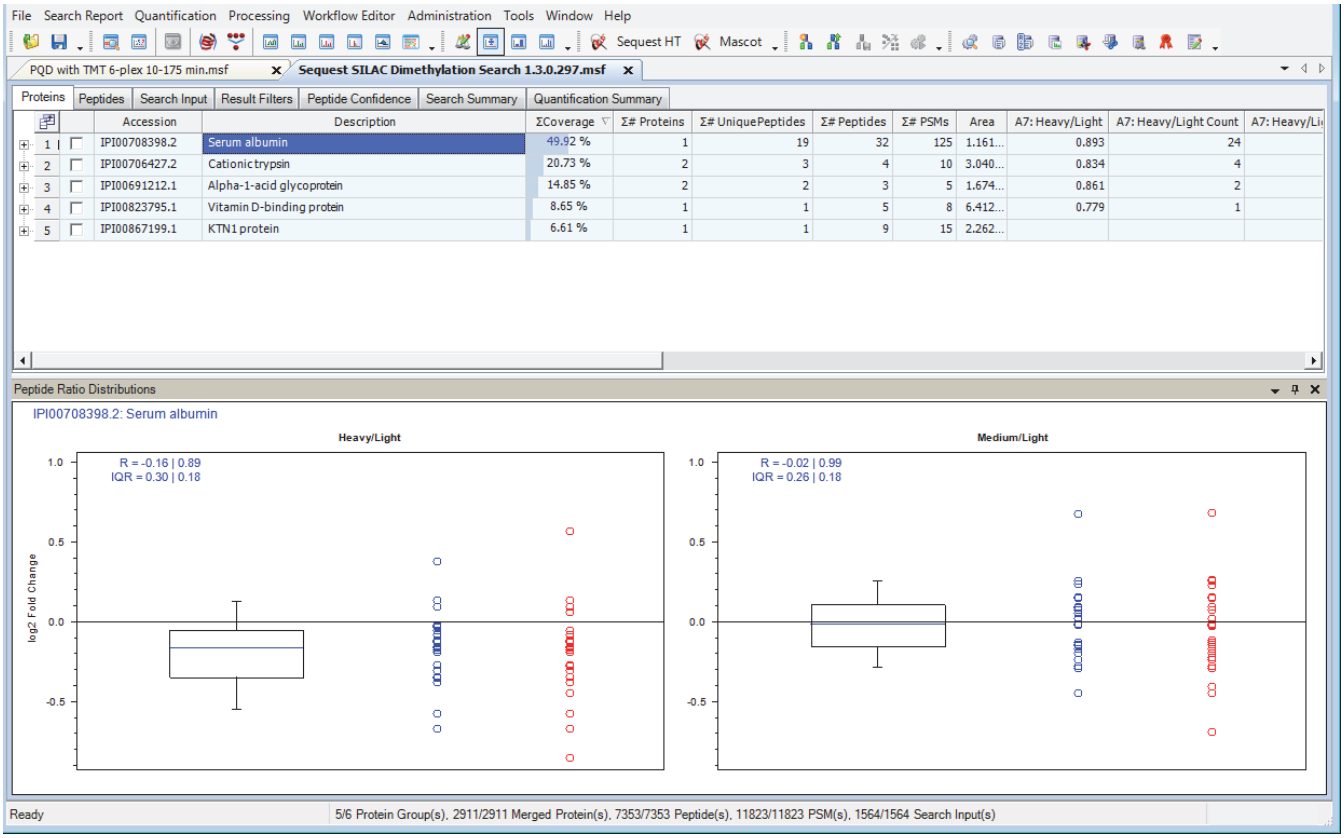
Figure 212. Peptide Ratio Distributions chart for reporter ion quantification



In addition, each chart displays the median ratio (R) and the inter-quartile range (IQR) in linear and logarithmic format. The header of the chart identifies the protein that the peptide belongs to. Right-click the chart and choose **Show Legend** for the identity of other notations on the chart.

Figure 213 shows the Peptide Ratio Distributions chart for precursor ion quantification.

Figure 213. Peptide Ratio Distributions Chart for precursor ion quantification



Handling Missing and Extreme Values in Calculating Peptide Ratios

Table 23 and Table 24 on page 319 list some of the different circumstances that can arise in calculating quantification ratios for peptides from the selected quantification values. A quantification value is the intensity or area detected for a given quantification channel. For reporter ion quantification, a quantification channel is one of the mass or reporter tags, and for precursor ion quantification, it is one of the different possible labeling states of a peptide corresponding to the different heavy amino acids used in the cell cultures. “Intensity” refers to both the intensity of the reporter peaks in reporter ion quantification and to the areas detected in precursor ion quantification.

When the Proteome Discoverer application detects the quantification values for the different quantification channels, some of the quantification values might be missing, probably because they fell below the detection limit. In addition, some channels might show very low or very high intensities, leading to the calculation of very high or very low ratios. Major changes might indicate exceptional cases, which you can exclude from the calculation of the protein ratios by using the settings on the Ratio Calculation page of the Quantification Method Editor dialog box.

Table 23 and Table 24 do not include cases that arise as a result of peptide uniqueness and protein grouping. They focus on cases that arise where one or both of the quantification channels that are used for calculating peptide ratios are zero. In these cases, the application detects nothing on a channel because the spectrum does not contain one of the reporter peaks, the heavy or light isotope pattern is missing, a quantification value falls below a specified minimum threshold, or the calculated ratios are very high or very low.

Table 23 and Table 24 list the different possible cases exemplified by arbitrary values. The values in the tables have [counts] as units if the cases are presented for reporter ion quantification. For precursor ion quantification, 114 and 115 are replaced by Light and Heavy, and the quantification values have [counts × min] as units.

In addition to the options listed in the tables, the handling of quantification values is also affected by the Apply Quan Value Corrections option on the Ratio Calculation page of the Quantification Method Editor dialog box and by the options on the Experimental Bias page of the same dialog box. For reporter ion quantification, the Apply Quan Value Corrections option determines whether to apply the purity correction for the detected quantification values. The Proteome Discoverer application applies the purity correction after it applies the other settings that potentially change the quantification values. It applies the experimental bias correction after the first time that it calculates all peptide and protein ratios. The application then determines the bias correction factor and applies it to every peptide and protein ratio.

Table 23. Calculating peptide ratios when quantification values are missing (Sheet 1 of 2)

Case	Minimum detected quan. value	Minimum Quan Value Threshold setting	Replace Missing Quan Values with Minimum Intensity setting	Reject All Quan Values If Not All Quan Channels Are Present setting	Detected quantification values			Displayed/used quantification values		
					114	115	116	114	115	116
All quan. values detected	33	0	No	Irrelevant	100	50	300	100	50	300
	33	0	Yes	Irrelevant	100	50	300	100	50	300
	33	75	No	Irrelevant	100	50	300	100	0	300
	33	75	Yes	Irrelevant	100	50	300	100	75	300

Table 23. Calculating peptide ratios when quantification values are missing (Sheet 2 of 2)

Case	Minimum detected quan. value	Minimum Quan Value Threshold setting	Replace Missing Quan Values with Minimum Intensity setting	Reject All Quan Values If Not All Quan Channels Are Present setting	Detected quantification values			Displayed/used quantification values		
					114	115	116	114	115	116
Quan. value missing for a quan. channel	33	0	No	No	100	0	300	100	0	300
	33	0	Yes	No	100	0	300	100	33	300
	33	75	No	No	100	0	300	100	0	300
	33	75	Yes	No	100	0	300	100	75	300
	33	0	No	Yes	100	0	300	0	0	0
	33	0	Yes	Yes	100	0	300	100	33	300
	33	75	No	Yes	100	0	300	0	0	0
	33	75	Yes	Yes	100	0	300	100	75	300
Quan. value missing for all quan. channels	33	0	No	Irrelevant	0	0	0	0	0	0
	33	0	Yes	Irrelevant	0	0	0	0	0	0
	33	75	No	Irrelevant	0	0	0	0	0	0
	33	75	Yes	Irrelevant	0	0	0	0	0	0

Table 24. Calculating peptide ratios when values are very high or low (Sheet 1 of 2)

Case	Maximum Allowed Fold Change setting	Use Ratios Above Maximum Allowed Fold Change for Quantification setting	Calculated ratios		Displayed ratios	
			115/114	116/114	115/114	116/114
Ratio is within the limits	100	Irrelevant	2.000	3.000	2.000	3.000
	100	Irrelevant	0.500	0.250	0.500	0.250
Ratio is 0 or ∞ because one quan. channel value is missing	100	No	0	3.000	0.000	3.000
	100	Yes	∞	0.250	100.000	0.250

Table 24. Calculating peptide ratios when values are very high or low (Sheet 2 of 2)

Case	Maximum Allowed Fold Change setting	Use Ratios Above Maximum Allowed Fold Change for Quantification setting	Calculated ratios		Displayed ratios	
			115/114	116/114	115/114	116/114
A ratio exceeds the limits	100	No	2000.000	3.000	100.000	3.000
	100	No	0.300	0.002	0.300	0.010
	100	Yes	2000.000	3.000	100.000	3.000
	100	Yes	0.300	0.002	0.300	0.010
All ratios exceed the limits	100	No	2000.000	0.002	100.000	0.010
	100	Yes	2000.00	0.002	100.000	0.010

Calculating Protein Ratios from Peptide Ratios

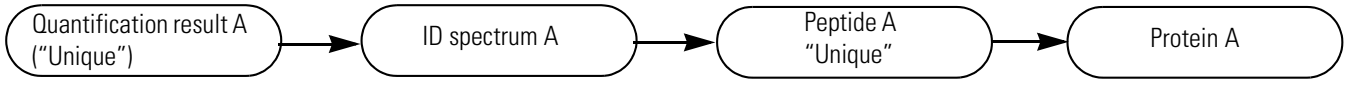
This section describes seven different scenarios that can occur when you derive protein quantification ratios from peptide quantification ratios. These cases show how the validity of using a given quantification result for the quantification of a certain protein depends on whether this particular quantification result is unique or shared among other peptides.

The peptide quantification ratios are taken from the associated quantification results. The term *quantification result* in this section refers to MS/MS reporter intensities taken from the same scan as the identification (for example, ID-CID) or from a separate quantification scan (for example, Quan-HCD). The term also refers to intensities derived from the precursor scans in precursor ion quantification. A quantification result here is a general quantity associated with one or more peptides that are, in turn, associated with one or more proteins.

Case 1: Quantification Result Associated with One Spectrum, One Peptide, and One Protein

Case 1, shown in Figure 214, is the simplest case. The quantification result is associated with one identification spectrum—whether the quantification results come from the same identification spectrum, from a different quantification spectrum, or from the precursor ion—and one peptide that is contained in one protein. The quantification result is unique for this protein. The Proteome Discoverer application can mark peptide A "Unique" in the Quan Info column of the Peptides page if the quantification result meets other criteria.

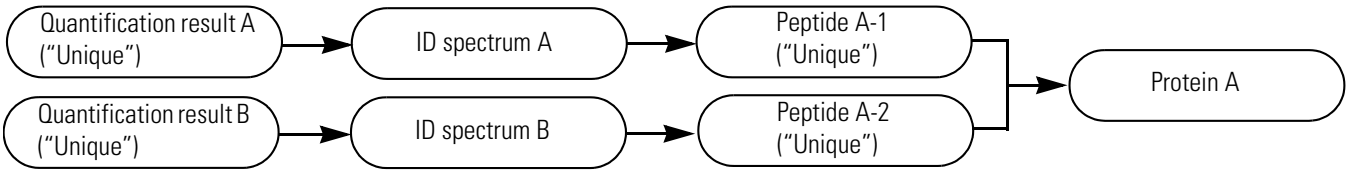
Figure 214. Case 1: Quantification result associated with one identification spectrum, one peptide, and one protein



Case 2: Two Quantification Results Associated with Two Spectra, One Peptide, and One Protein

Case 2, shown in Figure 215, is a variant of case 1. Each of two different quantification results is associated with a different identification spectrum. Both identification spectra identify peptide A, which is a peptide with the same sequence. Peptide A is only contained in one protein. Each of the two different quantification results is unique for just one protein. The peptides are redundantly identified and quantified, and you could use both for the quantification of protein A.

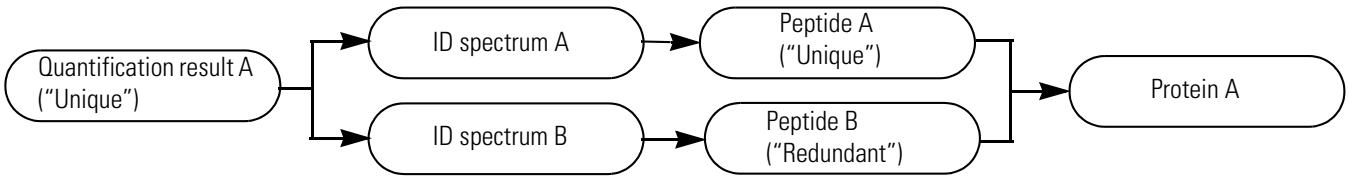
Figure 215. Case 2: Two different quantification results associated with two identification spectra, one peptide, and one protein



Case 3: Quantification Result Associated with Two Spectra, Two Peptides, and One Protein

Case 3, shown in Figure 216, is similar to case 2 but varies from it in a slight but important way. In case 3, the two identification spectra are associated with the same quantification result rather than with two different quantification results. For example, you might obtain these results if you trigger the same precursor two times for MS/MS. It does not matter whether peptide A and peptide B are the same peptides (redundantly identified) or different peptides that are accidentally contained in the same protein. It also does not matter whether they are identified by the same search engine or by two different search engines, for example, a CID spectrum and an ETD spectrum. The quantification result is still unique for just one protein. However, you cannot use the quantification ratio of both peptides A and B to calculate the quantification ratio of protein A, because it is the same quantification result, and you do not want to use the same quantification result multiple times for the same protein. In this case, the Proteome Discoverer application marks peptide A, the peptide with the better identification, as "Unique" and the other peptide as "Redundant" for quantification (rather than redundant for identification).

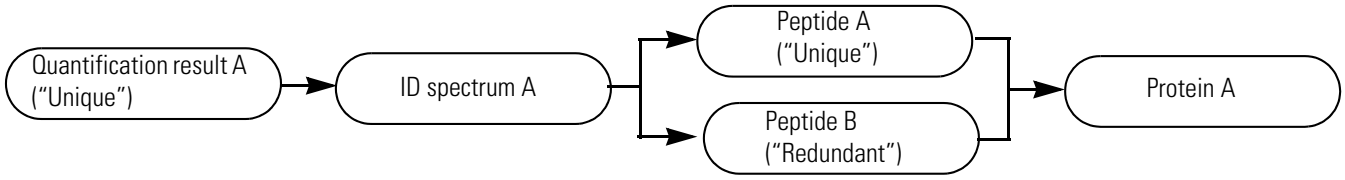
Figure 216. Case 3: Quantification result associated with two identification spectra, two peptides, and one protein



Case 4: Quantification Result Associated with One Spectrum, Two Peptides, and One Protein

In case 4, shown in Figure 217, the two peptides could be identified by the same search engine and have different ranks, or they could be identified by different search engines and both have rank 1. It does not matter whether peptide A and B have the same sequence with different PTM states or different sequences. The quantification result is unique for protein A. You can use it to calculate the protein ratio, but you must only use it once. The Proteome Discoverer application marks the "better" peptide as "Unique" and the other as "Redundant" for quantification.

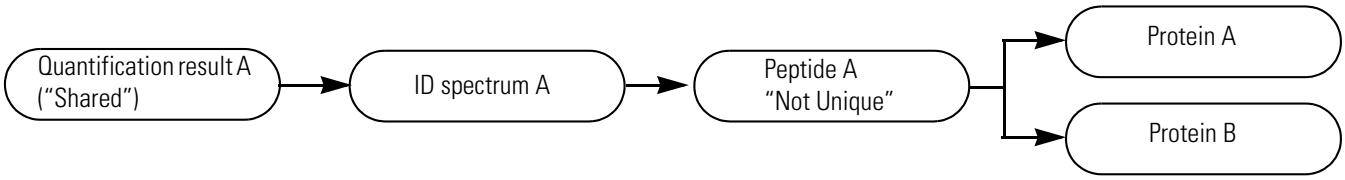
Figure 217. Case 4: Quantification result associated with one identification spectrum, two peptides, and one protein



Case 5: Quantification Result Associated with One Spectrum, One Peptide, Two Proteins

In case 5, shown in Figure 218, the quantification result is associated with one identification spectrum and one peptide, but this peptide is contained in more than one protein. The quantification result is potentially shared between these proteins, and you do not know how to share it. If the quantification method specifies using only unique peptides for protein quantification, you would not use peptide A in this case. If the quantification method specifies using all peptides for protein quantification, the quantification result of peptide A would be divided equally between both proteins.

Figure 218. Case 5: Quantification result associated with one identification spectrum, one peptide, and two proteins

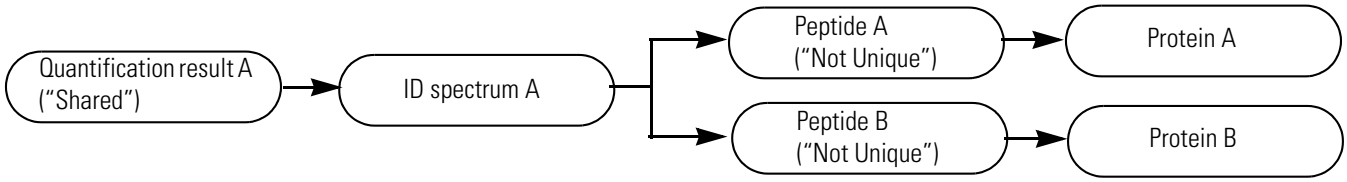


Case 6: Quantification Result Associated with One Spectrum, Two Peptides, and Two Proteins

In case 6, shown in Figure 219, the quantification result is associated with one identification spectrum from which two different peptides are identified either by the same search engine as different ranks or by different search engines. The two different peptides are contained in two different proteins. The two different peptides are both unique to just one protein. Nevertheless, the associated quantification result is the same, and you do not want to use it for

the calculation of the protein ratios if you specified in the quantification method to use only unique peptides. Only if you specify using all peptides can you use them for protein quantification. This case illustrates the discrepancy between the uniqueness of peptides and the uniqueness of the quantification results.

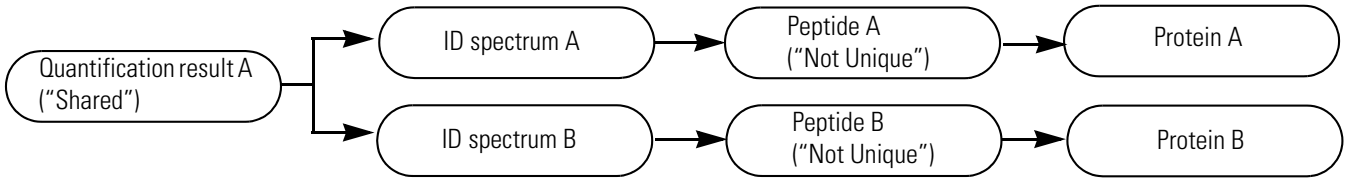
Figure 219. Case 6: Quantification result associated with one identification spectrum and two peptides unique to one protein



Case 7: Quantification Result Associated with Two Spectra, Two Peptides, and Two Proteins

Case 7, shown in Figure 220, is a variant of case 6. Either the same search engine or different search engines identify different identification spectra, for example, CID and ETD spectra. As in case 6, the peptides are unique, but the quantification result is not. The result depends on whether you specified in the quantification method to use only unique peptides or all peptides.

Figure 220. Case 7: Quantification result associated with two identification spectra and two different peptides unique to one protein



Calculating Ratio Count and Variability

The Proteins page of search reports with precursor ion quantification results displays columns called Heavy/Light Variability and Heavy/Light Count. Similarly, the Proteins page of search reports containing reporter ion quantification results displays columns called *Ratio* Variability [%] (for example, 114/113 Variability [%]) and *Ratio* Count (for example, 114/113 Count). The way the Proteome Discoverer application calculates and displays the values in these columns depends on whether you want the results treated as replicates or as treatments.

Replicates

Replicates are repeated measurements of the same sample. You repeat measurements to obtain better statistics. Without replicates, you cannot be sure that something that you observed is real—that is, statistically significant—and not a result of an error in the sample preparation, the liquid chromatography, the acquisition, and so forth. To generate replicates, you can

repeat the sample preparation or use the same sample and measure it multiple times. This data highlights the variance within the different steps. For example, if you measure a difference of 17 percent between two samples representing different treatments with a new drug, but the variance between the different replicates of the same treatment is already 28 percent, the observed difference might not be significant.

Treatments

Treatments are samples that are brought to different states. For example, they might be different samples representing different exposure levels to a certain drug or cultures of the same cells exposed to different levels of stress, such as radiation, salts, or heat.

Ratio Count

The *Ratio* Count or the Heavy/Light Count column displays the number of peptide ratios that were used to calculate a particular protein ratio. If only one ratio was reported (for example, the Heavy/Light ratio for SILAC data), the displayed count is the number of peptides marked "Used" for this protein. If more than one ratio was used (for example, the ratios in iTRAQ or TMT data), the count for a particular protein ratio might be smaller than the number of peptides marked "Used." The count could be lower than the number of peptides marked "Used" in cases where some of the ratios are excluded by the different settings or thresholds defined by the quantification method.

For replicates, the *Ratio* Count columns display a list of the separate counts for each replicate. If a protein was not identified in one of the replicates or no peptide usable for calculating the protein was identified for this replicate, a "-" appears in the *Ratio* Count cell. If none of the replicates provide a usable peptide, the *Ratio* Count cell is empty.

Ratio Variability

The *Ratio* Variability [%] columns show the variability of the peptide ratios that are used to calculate a particular protein ratio. They are similar to a coefficient of variation for the calculated protein ratios as a normalized measure of the peptide ratio spread used for calculating the protein ratio. The Proteome Discoverer application calculates the displayed variability differently for single search reports, multiconsensus reports that are treated as treatments, and multiconsensus reports that are treated as replicates.

Single Search Reports

For single search reports, the protein ratio variability is calculated as a coefficient-of-variation for log-normal distributed data ($CV_{log\text{-}normal}$). In this case, the protein ratio variability is calculated from the used peptide ratios $r_1 \ldots r_n$ as follows:

$$CV_{log\_normal} = \sqrt{\exp([StdDev(\log(r_1)\ldots\log(r_n))]^2) - 1}$$

where $StdDev\ (\log(r_1) \ldots \log(r_n) = 1.483 \times MAD(\log(r_1) \ldots \log(r_n))$.

$$CV_{log\_normal} = \sqrt{\exp([1.483 \times MAD(\log(r_1)\ldots\log(r_n))]^2) - 1}$$

$$variability(r_1\ldots r_n) = 100 \times CV_{log\text{-}normal}$$

where $MAD(r)$ is the median absolute deviation (MAD) of the peptide ratios $r_1 \ldots r_n$. In statistics, the median absolute deviation is a robust measure of the variability of a univariate sample of quantitative data.

$$MAD(\log(r_1\ldots r_n)) = median[|\log(r_1) - median(\log(r_1\ldots r_n))|]$$

Starting with the residuals (deviations) from the data's median, the median absolute deviation is the median of their absolute values. The 1.483 constant ensures consistency for the r_i distributed normally as $N(\mu, \sigma 2)$ and large N:

$$E(1.483 \times MAD(r_1\ldots r_n)) = \sigma$$

The Proteome Discoverer application uses these statistics because they are more robust in the presence of outliers as a classical coefficient of variation (CV). It also uses them to calculate the protein ratio as the median of the used peptide ratios.

Calculating Variability in Multiconsensus Reports Treated as Treatments

For multiconsensus reports that treat quantification data as different treatments, the results of the single searches are simply displayed side by side, and the variabilities are the same as those of the single reports. For more information on how the Proteome Discoverer application calculates protein ratios when treating quantification results in multiconsensus reports as treatments, see "Calculating Protein Ratios in Multiconsensus Reports Treated as Treatments" on page 328.

Calculating Variability in Multiconsensus Reports Treated as Replicates

For multiconsensus reports that treat quantification data as replicates, the Proteome Discoverer application calculates the protein ratios for single searches and then calculates a classical coefficient variation for these ratios. It calculates the variability of the protein ratio calculated from N replicates from the protein ratio $r_1 \ldots r_n$ of the single searches:

$$variability(r_1\ldots r_n) = CV = 100 \times \frac{\text{std. dev.}(r_1\ldots r_n)}{\text{arith. mean}(r_1\ldots r_n)}$$

Using the protein ratios rather than their logarithms is reasonable because in contrast to the peptide ratios, which are (at least approximately) log-normally distributed, the protein ratios of the single searches should be normally distributed—at least for larger values of n. For more information on how the Proteome Discoverer application calculates protein ratios when treating quantification results in multiconsensus reports as replicates, see "Calculating Protein Ratios in Multiconsensus Reports Treated as Replicates" on page 328.

Calculating and Displaying Protein Ratios for Multiconsensus Reports

The Proteome Discoverer application can treat the different single quantification results in multiconsensus reports as replicates of the same sample or as different treatments of a sample.

In real-world studies of quantitative responses of a sample to certain treatments, such as a particular change in environmental condition or an administration of a drug, you might be interested in the quantitative difference of the sample before and after the treatment, or between different treatment states of the sample (for example, different points in time after application of a certain drug or application of different amounts of a certain drug). Quantitative studies could also investigate the quantitative difference between samples in different states, for example, between similar samples from healthy and different disease states.

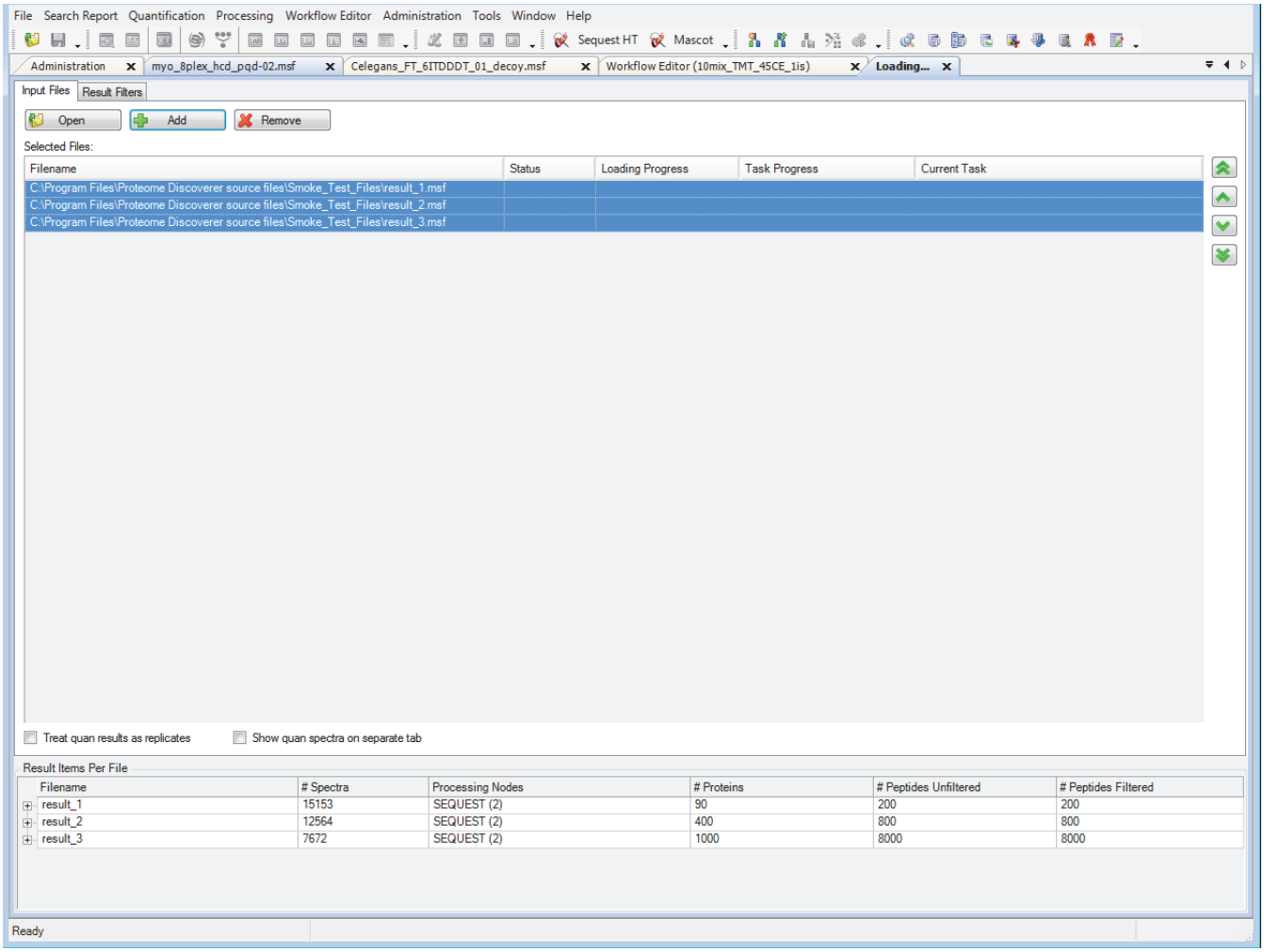
Such experimental investigations must assess the variability inherent in the different stages of the experiment. For example, samples from different animals or patients can vary significantly in their expression level for certain proteins or in the amount of proteins and peptides with PTMs. Other sources of variability are differences in sample preparation, differences in chromatographic separation, or differences in measurement in the mass spectrometer. When you examine the quantitative differences between two measurements, all these single factors combine to create an overall variability of the quantitative values under investigation, for example, the expression levels of certain proteins.

This overall variability can be quite significant. To minimize the variability when comparing two samples, such as different treatments or disease states, and to calculate a statistical measure of the inherent variability, you must measure replicates. In this process, you repeatedly measure a sample multiple times and calculate the average values for the quantitative values under investigation. You perform these measurements for all states of the sample and then compare the calculated average values. You can then calculate whether a detected difference between two states of a sample is statistically significant or is only due to the inherent variability of the sample.

In the Proteome Discoverer application, you can load multiple result files containing quantification results and treat the single results as replicates of the same sample or as different treatments of a sample. You determine whether the single results of an open multiconsensus report should be treated as replicates or treatments, and you can change them from replicates to treatments and vice-versa.

As an example, assume that you have three result files from measuring and processing a yeast sample: result_1.msf, result_2.msf, and result_3.msf. Assume that the samples were prepared with the iTRAQ 4plex quantification method, giving quantifiable reporter peaks at 114, 115, 116, and 117 *m/z*. When you open these three MSF files, the Proteome Discoverer application adds the files to the Input Files page, as shown in Figure 221.

Figure 221. Loading three single result files containing quantification data



By default, the Proteome Discoverer application initially treats the quantification results in the single MSF files as if they were from different treatments of a sample. You can change this treatment by selecting the Treat Quan Results as Replicates option on the Input Files page. The application then initially treats the quantification data as if it were replicates of the same sample when it creates and opens the multiconsensus report.

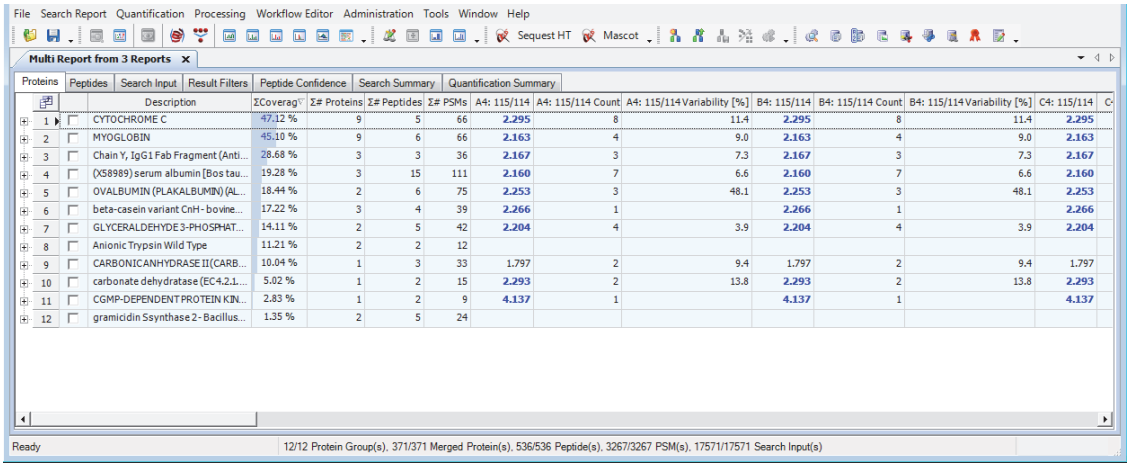
If you do not select the Treat Quan Results as Replicates option and click Open on the Input Files page, the application creates the multiconsensus report and calculates the quantification results (the quantification ratios as specified in the quantification method) for each of the three single results files separately. In this example, all three result files are iTRAQ 4plex files, so the application usually calculates ratios such as 115/114, 116/114, and 117/114 for each of the files. Because you did not select Treat Quan Results as Replicates, the application reports them as if the three files represented different treatment states of a sample.

Calculating Protein Ratios in Multiconsensus Reports Treated as Treatments

In the case of treatments, the Proteome Discoverer application reports all calculated protein ratios of the single result files side by side for the multiconsensus report. It prefixes each file by a single letter identifier of the particular report and the number of the quantification node in the processing workflow. In the example given in Figure 222, the protein and corresponding variability and ratio count columns are A4: 115/114, B4: 115/114, C4: 115/114, A4: 116/114, B4: 116/114, C4: 116/114, A4: 117/114, B4: 117/114, and C4: 117/114. At the peptide level, there is no difference between treatments or replicates in multiconsensus reports and single reports.

Figure 222 shows the protein ratios in a multiconsensus report when the quantification results of the single result files are treated as different treatments of the sample.

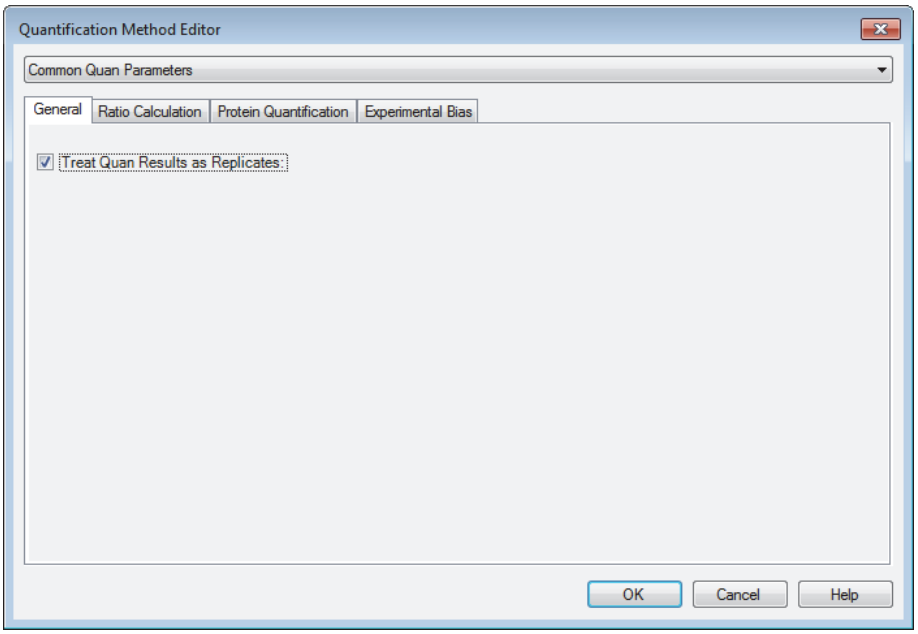
Figure 222. Protein ratios when single quantification result files are treated as different treatments of the sample



Calculating Protein Ratios in Multiconsensus Reports Treated as Replicates

With replicates, the application treats the quantification results like replicates of the same sample. You can specify that quantification results be treated as replicates by selecting the Treat Quan Results as Replicates option on the Input Files page or by using the Quantification Method Editor dialog box (opened by choosing Quantification > Edit Quantification Method) when the multiconsensus report is open. For multiconsensus reports, the Quantification Method Editor dialog box features a Common Quan Parameters box (shown in Figure 187 on page 282) so that you can set common quantification parameters for all contained result files at once. On the General page of the dialog box, shown in Figure 223, you can switch between treatment and replicate mode.

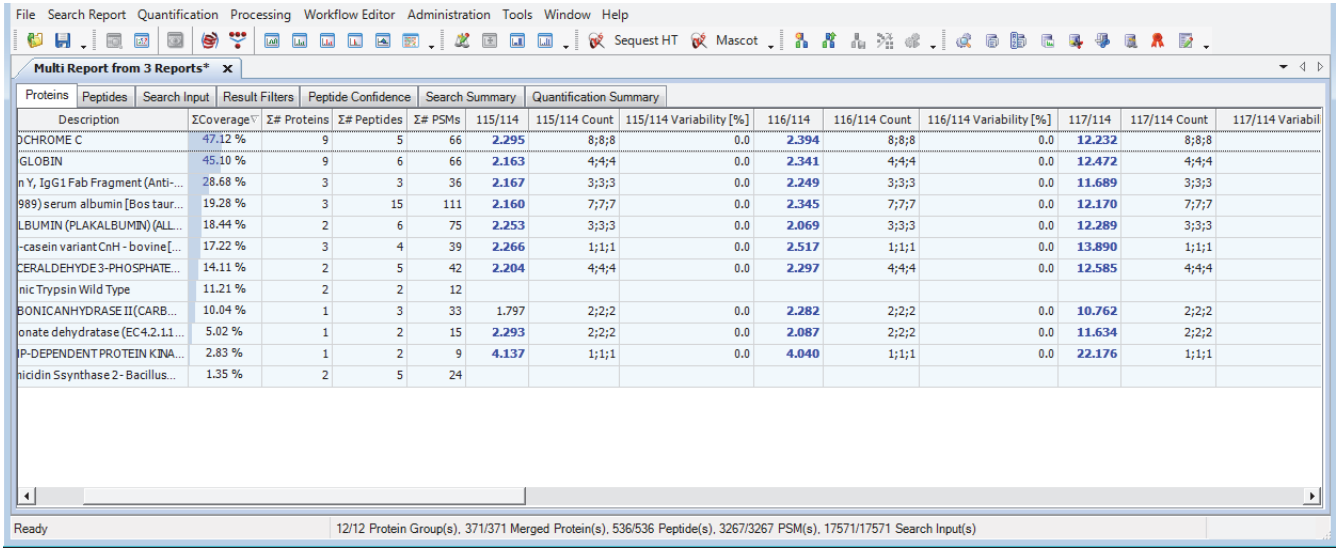
Figure 223. Switching between treatment and replicate mode by editing the common quantification parameters



As in treatment mode, multiconsensus reports are no different from single reports for replicates at the peptide level. At the protein level, the Proteome Discoverer application combines the protein ratios of the single result files into averaged protein ratios, as shown in Figure 224. It calculates the combined protein ratio as the arithmetic mean of the protein ratios of the single reports (and calculates the protein ratios as the median of the "used" peptide ratios of the particular result file). See "Calculating Ratio Count and Variability" on page 323 for information on how the application calculates and displays the values in the Ratio Count and Variability columns for multiconsensus reports.

Figure 224 shows protein ratios in a multiconsensus report when the quantification results of the single result files are treated as replicates of the same sample.

Figure 224. Protein ratios when single quantification result files are treated as replicates of the same sample



The Proteome Discoverer application combines only protein quantification ratios from the same type of quantification—that is, either precursor-ion- or reporter-ion-based quantification—into replicate ratios. The names of the protein ratios must be the same to be combined into replicate ratios. For example, the ratios to combine into replicates must all be from reporter-ion-based quantification, and they must all be identically named (such as 115/114) in the result files to be combined. The application reports ratios from different types of quantification or ratios with different names as if they were treatments—that is, side by side on the protein level of the multiconsensus report.

Mixed Mode

You can also mix replicate and treatment mode. For example, you can load three result files from an iTRAQ 4plex experiment and two files from a SILAC experiment, and specify treating the quantification results as replicates. In this case, the Proteome Discoverer application tries to treat all defined protein quantification ratios as replicates, if possible. It reports everything else side by side at the protein level of the multiconsensus report. In this example, it calculates the combined averaged ratios from the three iTRAQ 4plex files and the two SILAC 2plex files, and reports the iTRAQ and SILAC ratios side by side, as shown in Figure 225 and Figure 226 on page 332. In this way, the application can mimic complex experimental setups.

Figure 225. Opening a multiconsensus report from three iTRAQ and two SILAC files in replicate mode

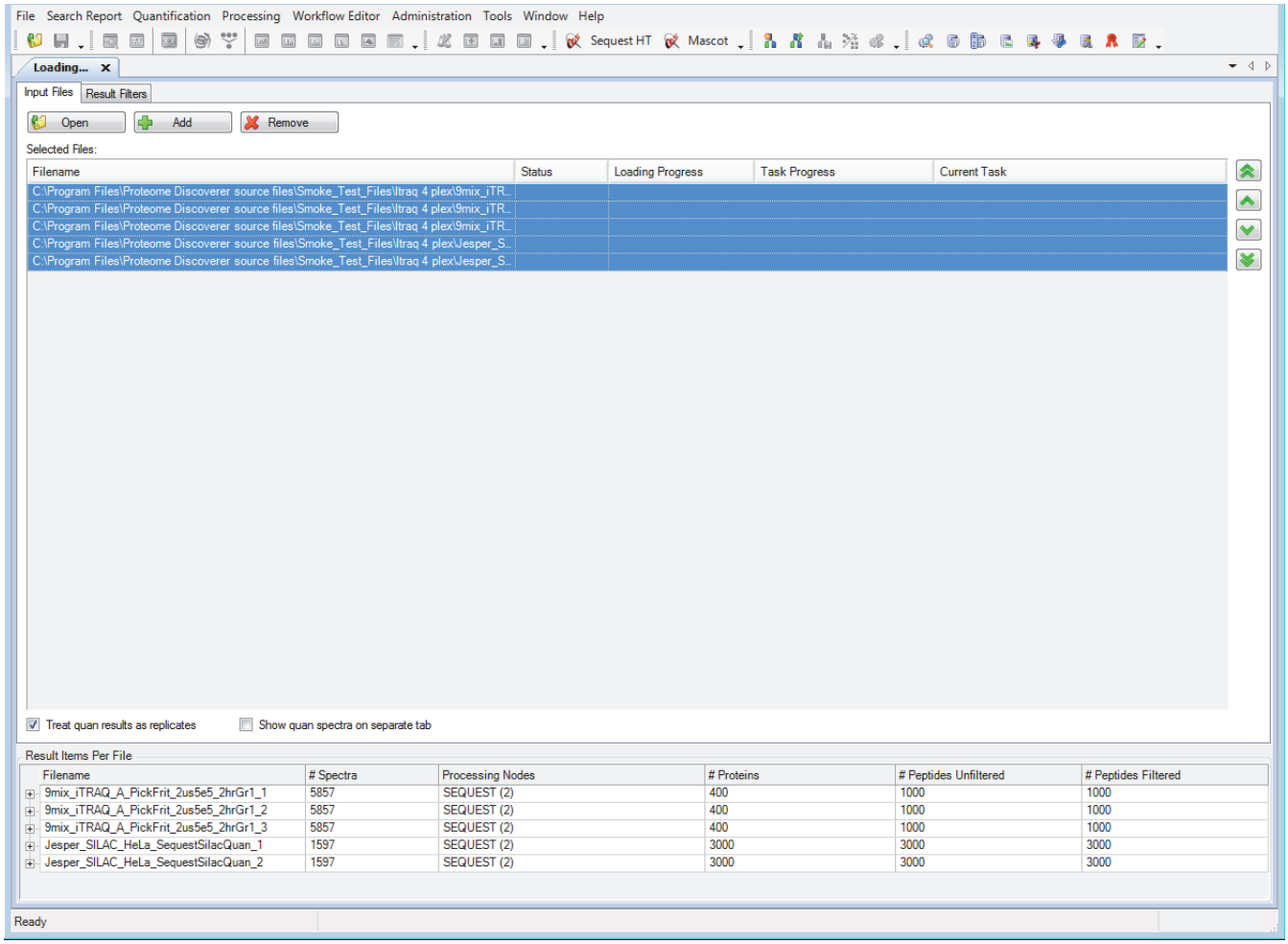
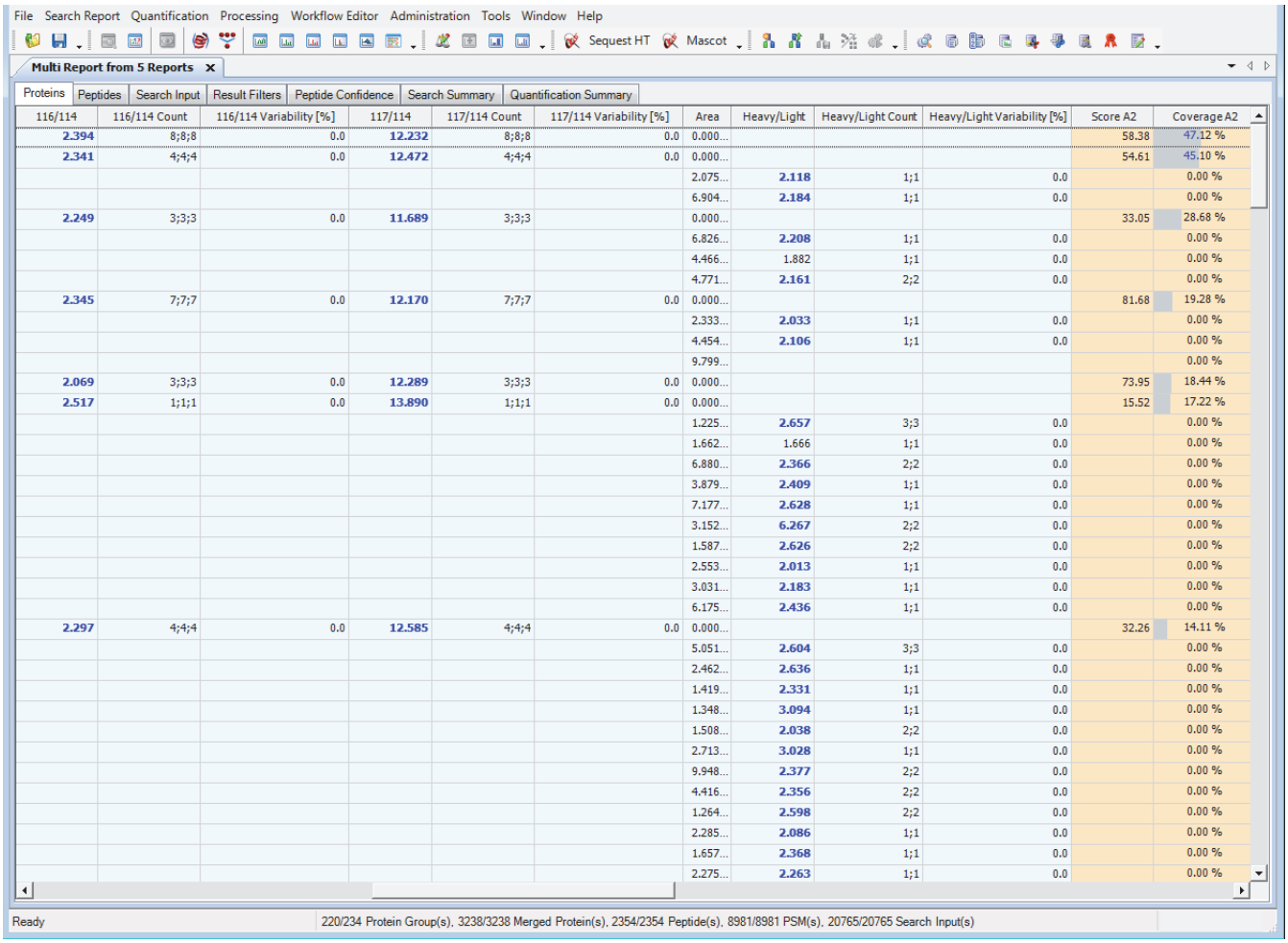


Figure 226 shows the opened multiconsensus report loaded in Figure 225. The combined ratios from the iTRAQ and the SILAC quantification are displayed side by side. In this example, the two types of searches are from different samples, and the two different types of quantification share no proteins.

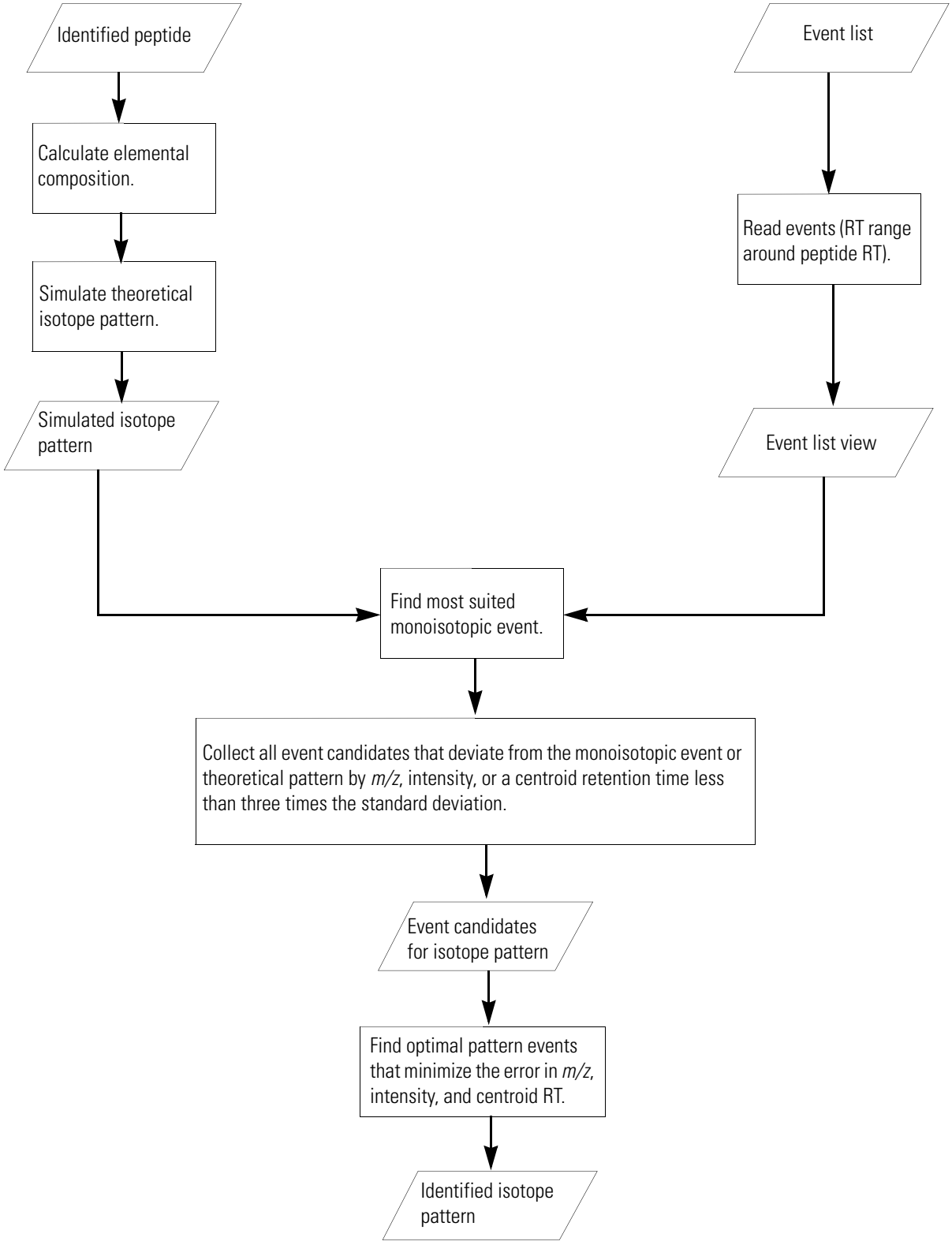
Figure 226. Opened multiconsensus report from three iTRAQ and two SILAC files in replicate mode



Identifying Isotope Patterns in Precursor Ion Quantification

The quantification spectra on the pages of the MSF report show the isotope pattern used for quantifying the peptides. The algorithm used in precursor ion quantification finds isotope patterns by identifying target components—that is, known elemental compositions from event lists. It identifies the peptides and searches in the event lists for the isotope patterns of these identified peptides. After peptide identification, the algorithm follows the steps shown in Figure 227 to identify the isotope patterns.

Figure 227. Identifying isotope patterns



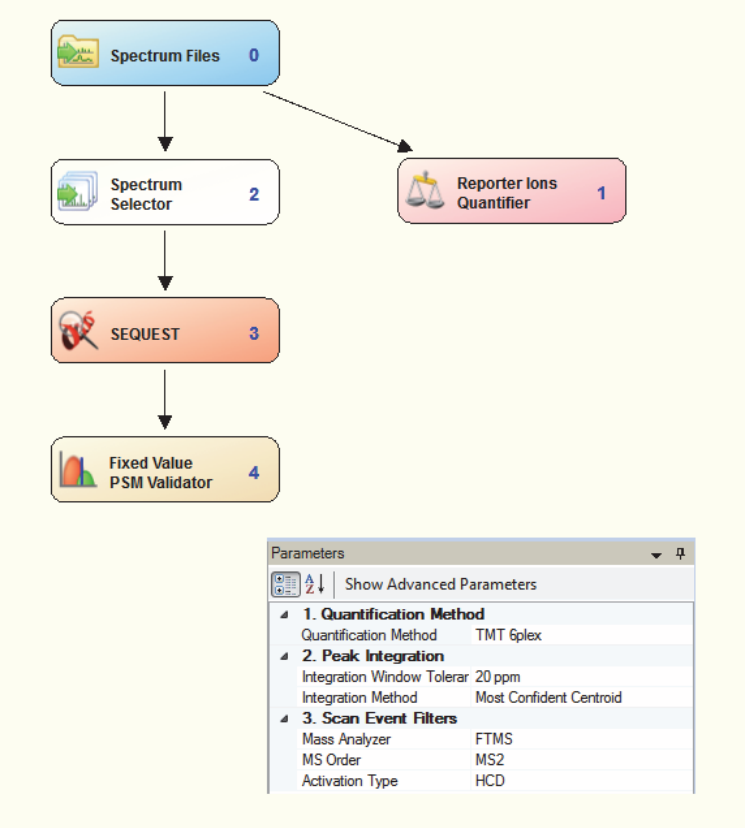
Troubleshooting Quantification

The following procedures can help you obtain optimal results when performing quantification.

❖ To troubleshoot reporter ion quantification

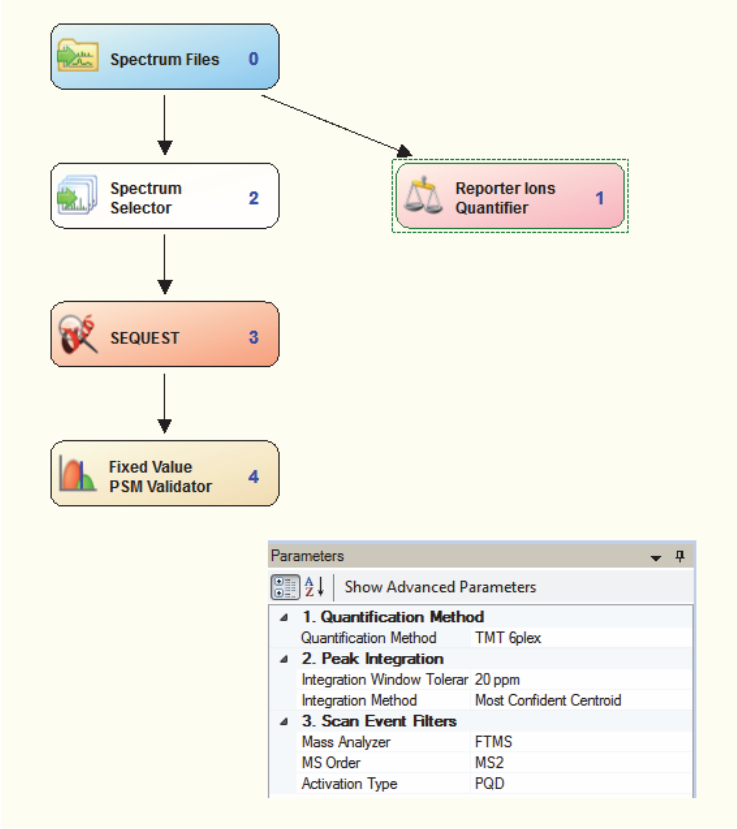
- If you obtain unexpected quantification results, verify that all settings of the nodes in your processing workflow are reasonable.
  - Make sure that the Integration Tolerance parameter of the Reporter Ions Quantifier node fits the data that you are processing. The default is 20 ppm, which is too low if you are processing PQD data from an ion trap.
  - Make sure that the settings of the Mass Analyzer, MS Order, and Activation Type parameters of the Reporter Ions Quantifier node are correct for the data that you are processing. Figure 228 shows the typical settings to use if you want to quantify HCD scans from the Orbitrap.

Figure 228. Typical settings for quantifying iTRAQ or TMT tags from HCD scans



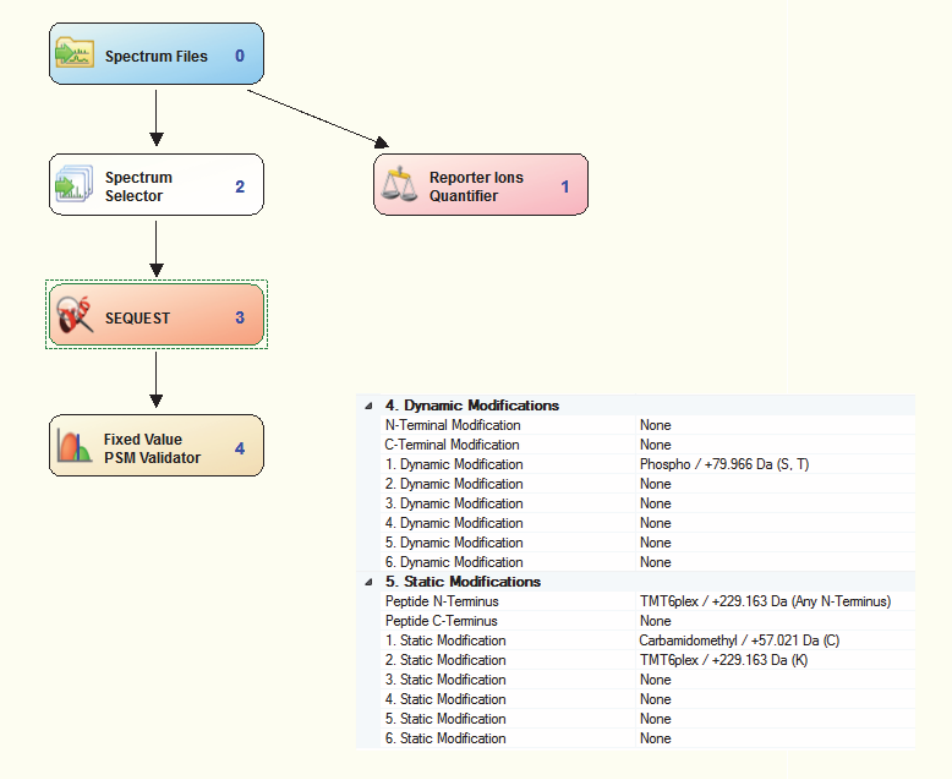
To quantify PQD scans from an ion trap, use the typical settings shown in Figure 229.

Figure 229. Typical settings for quantifying iTRAQ or TMT tags from the ion trap PQD scans



– Make sure that you have used the correct set of static and dynamic modifications for the search engine. For example, if you are searching TMT 6plex data with SEQUEST, check that your settings resemble those in Figure 230.

Figure 230. Modifications required for searching TMT 6plex samples



❖ To troubleshoot precursor ion quantification

- If you obtain unexpected precursor ion quantification results, verify that all settings of your processing workflow are reasonable.
  - Check the dynamic modification parameters in the Sequest HT, SEQUEST, or Mascot search engines.

  These should match your isotope labeling sample.
  - Check the node parameters that you set before performing the quantification to see if they are appropriate for your sample.

  See "Performing Reporter Ion Quantification" on page 249 for more information.
  - Verify that your isotopic labeling is one of the following options in the protein ID/search node (either Sequest HT, SEQUEST, or Mascot):
    - SILAC 2plex (Arg10, Lys6): Uses arginine 10 and lysine 6.
    - SILAC 2plex (Arg10, Lys8): Uses arginine 10 and lysine 8.
    - SILAC 2plex (Ile6): Uses isoleucine 6.
    - SILAC 3plex (Arg6, Lys4|Arg10, Lys8): Uses arginine 10 and lysine 8 for "heavy" labels and arginine 6 and lysine 4 for "medium" labels.

- SILAC 3plex (Arg6, Lys6|Arg10, Lys8): Uses arginine 10 and lysine 8 for "heavy" labels and arginine 6 and lysine 6 for "medium" labels.
- Dimethylation 3plex: Chemically adds isotopically labeled dimethyl groups to the N-terminus and to the ε-amino group of lysine.
- ^{18}O labeling: Introduces 2 or 4 Da mass tags through the enzyme-catalyzed exchange reaction of C-terminal oxygen atoms with ^{18}O.

Note Low-mass accuracy cannot be used for precursor ion quantification or precursor ion area detection.

- Check your tolerance window. If you get too many results, decrease the size of the window. For too few results, increase the size of the window.
- Make sure you chose the right database.
- Check the species listed to make sure the samples came from that species.
- Verify that the activation type used is correct.
- Verify that the instrument type in the Mascot search engine is correct.
- Use only the ETD Spectrum Charger node for low-mass resolution ETD data.

FASTA Reference

This appendix lists the most important FASTA databases and parsing rules that the Proteome Discoverer application uses to obtain protein sequences, accession numbers, and descriptions.

Contents

FASTA Databases

These are the most important FASTA databases that the Proteome Discoverer application uses.

- NCBI
- MSIPI
- IPI
- UniRef100
- SwissProt and TrEMBL
- MSDB

Follow the links given for each database if you would like to download the database and save it to your local machine. Some databases are more time-consuming to load than others.

NCBI

NCBI is a non-redundant database compiled by the NCBI (National Center for Biotechnology Information) as a protein database for Blast searches. It contains nonidentical sequences from GenBank CDS translations, Protein Data Bank (PDB), SwissProt, Protein Information Resource (PIR), and Protein Research Foundation (PRF).

http://blast.ncbi.nlm.nih.gov/Blast.cgi

ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz

A typical NCBI title line follows:

```
>gi|70561|pir||MYHO myoglobin - horse_i|418678|pir||MYHOZ myoglobin -
common zebra (tentative sequence) [MASS=16950]
```

FASTA ID:

- Accession#:gi70561
- Description:myoglobin - horse_i

MSIPI

MSIPI is a database derived from IPI that contains additional information about cSNPs, N-terminal peptides, and known variants in a format suitable for mass spectrometry search engines. MSIPI is produced by the Max Planck Institute for Biochemistry at Martinsried and the University of Southern Denmark. It is distributed by the European Bioinformatics Institute (EBI).

ftp://ftp.ebi.ac.uk/pub/databases/IPI/msipi/current/

A typical MSIPI title line follows:

```
>MSIPI:IPI00000001.2| Gene_Symbol=STAU1 Isoform Long of
Double-stranded RNA-bin ding protein Staufen homolog 1 lng=577 #
CON[595,R,359,A] #
```

FASTA ID:

- Accession#:IPI00000001.2
- Description:Isoform Long of Double-stranded RNA-bin ding protein Staufen homolog 1 lng=577 # CON[595,R,359,A] #

IPI

The International Protein Index (IPI) is compiled by the European Bioinformatics Institute (EBI) to provide a top-level guide to the main databases that describe the human and mouse proteomes: SwissProt, TrEMBL, NCBI RefSeq, and Ensembl.

http://www.ebi.ac.uk/IPI/

ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/

A typical IPI title line follows:

```
>IPI:IPI00685094.1|SWISS-PROT:2KIJ2|ENSEMBL:ENSBTAP00000028878|REFSE:
NP_001073825;XP_593190 Tax_Id=9913 Gene_Symbol=MGC137286;LOC515210
Uncharacterized protein C1orf156 homolog
```

FASTA ID:

- Accession#:IPI00685094.1
- Description:Uncharacterized protein C1orf156 homolog

UniRef100

UniRef, also known as UniProt NREF, is a set of comprehensive protein databases curated by the Universal Protein Resource consortium. UniRef100 contains only nonidentical sequences, and UniRef90, and UniRef50 are non-redundant at a sequence similarity level of 90 percent and 50 percent, respectively.

http://www.ebi.ac.uk/uniref/

ftp://ftp.uniprot.org/pub/databases/uniprot/*current_release*/uniref/uniref100/

A typical UniRef100 title line follows:

```
>UniRef100_4U9M9 Cluster: 104 kDa microneme-rhoptry antigen precursor;
n=1; Theileria annulata|Rep: 104 kDa microneme-rhoptry antigen
precursor - Theileria annulata
```

FASTA ID:

- Accession#:4U9M9
- Description:Cluster: 104 kDa microneme-rhoptry antigen precursor; n=1; Theileria annulata|Rep: 104 kDa microneme-rhoptry antigen precursor - Theileria annulata

SwissProt and TrEMBL

The SwissProt database is developed by the SwissProt groups at the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

TrEMBL is a computer-annotated supplement of SwissProt that contains all the translations of EMBL nucleotide sequence entries not yet integrated into SwissProt.

http://www.expasy.org/sprot/

ftp://ftp.expasy.org/databases/uniprot/knowledgebase/uniprot_sprot.fasta.gz

ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/uniprot_trembl.fasta.gz

A typical SwissProt title line follows:

```
>43495|108_SOLLC Protein 108 precursor - Solanum lycopersicum (Tomato)
(Lycopersicon esculentum)
```

FASTA ID:108_SOLLC

- Accession#:43495
- Description:Protein 108 precursor - Solanum lycopersicum (Tomato) (Lycopersicon esculentum)

MSDB

The Mass Spectrometry Protein Sequence Database (MSDB) is compiled by the Clinical and Biomedical Proteomics group at the University of Leeds, using the PIR, TrEMBL, GenBank, SwissProt, and NRL3D source databases.

http://proteomics.leeds.ac.uk/bioinf/msdb.html

ftp://ftp.ncbi.nih.gov/repository/MSDB/

A typical MSDB title line follows:

```
>CBMS Ubiuinol-cytochrome-c reductase (EC 1.10.2.2) cytochrome b -
mouse mitochondrion
```

FASTA ID:

- Accession#:CBMS
- Description:Ubiuinol-cytochrome-c reductase (EC 1.10.2.2) cytochrome b - mouse mitochondrion

Custom Database Support

The Proteome Discoverer application has three "general" parsing rules to support custom sequence database formats. The generic parsing rules are applied only if no other parsing rule matches the given FASTA title line.

- Custom Parsing Rule A
- Custom Parsing Rule B
- Custom Parsing Rule C

Custom Parsing Rule A

The application uses custom parsing rule A if the FASTA ID, the accession number, and the description are separated by a pipe (|) symbol. A typical FASTA title line that matches this parsing rule would look like this one:

```
>tr|18FC3|18FC3_HALWD IS1341-type transposase - Halouadratum walsbyi
(strain DSM 16790).
```

FASTA ID:18FC3_HALWD

- Accession#:18FC3
- Description:IS1341-type transposase - Halouadratum walsbyi (strain DSM 16790).

Custom Parsing Rule B

The application uses custom parsing rule B if the accession number and the description are separated by a space. A typical FASTA title line that matches this parsing rule would look like this one:

```
>HP0001 hypothetical protein {Helicobacter pylori 26695}
```

FASTA ID:

- Accession#:HP0001
- Description:hypothetical protein {Helicobacter pylori 26695}

Custom Parsing Rule C

The application uses custom parsing rule C if the FASTA title line only contains the accession number. A typical FASTA title line that matches this parsing rule would look like this one:

```
>143B_HUMAN
```

FASTA ID:

- Accession#:143B_HUMAN
- Description:143B_HUMAN

Chemistry References

The tables in this appendix list amino acid symbols and mass values, enzyme cleavage properties, and the fragment ions used in the Proteome Discoverer application.

Contents

Amino Acid Mass Values

The Proteome Discoverer application uses the amino acid symbols and mass values listed in Table 25 and Table 26.

Table 25. Amino acid mass values (Sheet 1 of 2)

Amino acid	One-letter code	Three-letter code	Monoisotopic mass	Average mass	Sum formula
Glycine	G	Gly	57.02147	57.0517	C_2H_3NO
Alanine	A	Ala	71.03712	71.0787	C_3H_5NO
Serine	S	Ser	87.03203	87.078	$C_3H_5NO_2$
Proline	P	Pro	97.05277	97.1168	C_5H_7NO
Valine	V	Val	99.06842	99.1328	C_5H_9NO
Threonine	T	Thr	101.04768	101.1051	$C_4H_7NO_2$
Cysteine	C	Cys	103.00919	103.145	C_3H_5NOS
Isoleucine	I	Ile	113.08407	113.1598	$C_6H_{11}NO$
Leucine	L	Leu	113.08407	113.1598	$C_6H_{11}NO$
Asparagine	N	Asn	114.04293	114.1039	$C_4H_6N_2O_2$
Aspartic Acid	D	Asp	115.02695	115.0885	$C_4H_5NO_3$
Glutamine	Q	Gln	128.05858	128.13091	$C_5H_8N_2O_2$
Lysine	K	Lys	128.09497	128.1745	$C_6H_{12}N_2O$

Table 25. Amino acid mass values (Sheet 2 of 2)

Amino acid	One-letter code	Three-letter code	Monoisotopic mass	Average mass	Sum formula
Glutamic Acid	E	Glu	129.0426	129.1156	$C_5H_7NO_3$
Methionine	M	Met	131.0405	131.1994	C_5H_9NOS
Histidine	H	His	137.05891	137.1414	$C_6H_7N_3O$
Phenylalanine	F	Phe	147.06842	147.1772	C_9H_9NO
Arginine	R	Arg	156.10112	156.188	$C_6H_{12}N_4O$
Tyrosine	Y	Tyr	163.06332	163.17661	$C_9H_9NO_2$
Tryptophan	W	Trp	186.07932	186.2141	$C_{11}H_{10}N_2O$

Table 26. Special amino acids

Amino acid	One-letter code	Three-letter code	Monoisotopic mass	Average mass	Sum formula
Avrg. N/D	B	Bnd	114.53494	114.5962	$C_4H_5NO_3$
Avrg. /E	Z	Ze	128.55059	128.62326	$C_5H_7NO_3$
Unknown acid (X)	X	Xxx	110	110	N/A
Pyrrolysine	O	Pyl	237.14772	237.29874	$C_{12}H_{19}N_3O_2$
Seleno Cysteine	U	Sec	150.95309	150.0369	C_3H_5NOSe

Enzyme Cleavage Properties

Table 27 lists the enzymes and reagents with cleavage properties.

Table 27. Cleavage properties of enzymes and reagents (Sheet 1 of 2)

Enzymes/Reagents	Cleaves after	Cleaves before	Except when
Enzymes for digestion			
AspN		D	
Chymotrypsin	F, W, Y, or L		
Chymotrypsin (FWY)	F, W, or Y		P is after F, W, or Y
Clostripain	R		
Elastase	A, L, I, or V		P is after A, L, I, or V
Elastase/Tryp/Chymo	A, L, I, V, K, R, W, F, or Y		P is after Al, L, I, V, K, R, W, F, or Y

Table 27. Cleavage properties of enzymes and reagents (Sheet 2 of 2)

Enzymes/Reagents	Cleaves after	Cleaves before	Except when
GluC	E or D		
LysC	K		
Proline_Endopept	P		
Staph_protease	E		
Trypsin	K or R		P is after K or R
Trypsin (KRLNH)	K, R, L, N, or H		
Trypsin_K	K		P is after K
Trypsin_R	R		P is after R
Chemicals for degradation			
Cyanogen bromide	M		
Iodobenzoate	W		

Fragment Ions

Fragment ions of peptides are produced by several different fragmentation techniques, such as ECD, ETD, CID, higher-energy C-trap dissociation (HCD), and infrared multi-photon dissociation (IRMPD).

As an example, low-energy CID spectra, which are sequence-specific, are generated by MS/MS and ESI. The fragment ion spectra contain peaks of the fragment ions formed by the cleavage of the peptide bond and are used to determine amino acid sequences. A fragment must have at least one charge for it to be detected.

The fragment ions produced are identified according to where they are fragmented in the peptide. A, b, and c fragment ions have a charge on the N-terminal side, and x, y, and z fragment ions have a charge on the C-terminal side. Fragment ions a*, b*, and y* are ions that have lost ammonia (–17 Da), and fragment ions a^o, b^o, and c^o are ions that have lost water (–18 Da). The subscript next to the letter indicates the number of residues in the fragment ion.[1]

Table 28 summarizes the fragment ions used in the Proteome Discoverer application.

[1] For more information about fragment ions and nomenclature, see Roepstorff, P. and Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed. Mass Spectrum*. **1984**, *11* (11) 601.

Table 28. Fragment ions

Ions	Description
a	A ion with charge on the N-terminal side
b	B ion with charge on the N-terminal side
c	C ion with charge on the N-terminal side
y	Y ion with charge on the C-terminal side
z	Z ion with charge on the C-terminal side
a*	A ion that has lost ammonia (–17 Da)
b*	B ion that has lost ammonia (–17 Da)
y*	Y ion that has lost ammonia (–17 Da)
a^o	A ion that has lost water (–18 Da)
b^o	B ion that has lost water (–18 Da)
c^o	C ion that has lost water (–18 Da)

Index

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z